Ignatius Novianto Hariwibowo

ANALISA DATA DENGAN MS EXCEL: PENDEKATAN STATISTIK

Analisa Data Dengan Ms Excel: Pendekatan Statistik

Ignatius Novianto Hariwibowo

Universitas Atma Jaya Yogyakarta

Analisa Data Dengan Ms Excel: Pendekatan Statistik

Penulis:
Ignatius Novianto Hariwibowo

Hak Cipta © 2025, pada penulis

Hak publikasi pada Penerbit Universitas Atma Jaya Yogyakarta

Dilarang memperbanyak, memperbanyak sebagian atau seluruh isi dari buku ini dalam bentuk apapun, tanpa izin tertulis dari penerbit.

Cetakan ke-	05	04	03	02	01
Tahun	21	22	23	24	25

Diterbitkan oleh
UNIVERSITAS ATMA JAYA YOGYAKARTA
Jl. Babarsari No. 5-6 Yogyakarta 55281
Telp. +62 274 487711
E-mail: lib.publisher@uajy.ac.id

ISBN: 978-623-10-8343-2 (PDF)

Analisa Data Dengan Ms Excel: Pendekatan Statistik

Kata Pengantar

Puji syukur saya panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga buku ini, "Analisa Data dengan Ms Excel: Pendekatan Statistik," dapat terselesaikan dengan baik. Buku ini disusun sebagai panduan praktis bagi para pembaca yang ingin memahami dan mengaplikasikan teknik-teknik analisis data menggunakan Microsoft Excel dengan pendekatan statistik.

Dalam era digital saat ini, kemampuan untuk menganalisis data secara efektif menjadi sangat penting. Microsoft Excel, sebagai salah satu perangkat lunak pengolah data yang paling populer, menawarkan berbagai fitur yang dapat memudahkan proses analisis data. Buku ini dirancang untuk membantu pembaca, baik yang masih pemula maupun yang sudah berpengalaman, dalam menguasai teknik-teknik analisis data statistik dengan menggunakan Excel.

Saya menyadari bahwa buku ini masih jauh dari sempurna. Oleh karena itu, saya sangat mengharapkan kritik dan saran yang membangun dari para pembaca untuk perbaikan di masa mendatang. Akhir kata, saya berharap buku ini dapat memberikan manfaat dan kontribusi positif bagi perkembangan ilmu pengetahuan dan keterampilan analisis data di Indonesia.

Selamat membaca.
Yogyakarta, Desember 2024
Penulis

Daftar Isi

Kata Pengantar ..... **v**
Daftar Isi ..... **vii**

BAB 1
POPULASI DAN SAMPEL **1**
1.1. Populasi vs Sampel 3
1.2. Penentuan Sampel 5

BAB 2
STATISTIK DESKRIPTIF **7**
2.1. Jenis Data 9
2.2. Tingkat Pengukuran (*Level of Measurement*) 11
2.3. Variabel Kategorial (Teknik visualisasi) 13
2.4. Variabel Numerik (Tabel Distribusi Frekuensi) 17
2.5. Histogram 20
2.6. *Cross Table* dan *Scaterplot* 22
2.7. Mengukur Karakter Sampel atau Populasi 26
2.7.1. *Mean*, Median, Modus 26
2.7.2. Kecondongan (*Skewness*) 29
2.8. Mengukur Keragaman atau Variabilitas 32
2.8.1. Varians (*Variance*) 32
2.8.2. Standar Deviasi (*Standard Deviation*) 34
2.9. Mengukur Hubungan Dua Variabel 37
2.9.1. Kovarians (*Covariance*) 37
2.9.2. Koefisien korelasi (*Correlation Coefficient*) 39

Appendix Bab 2: Statistik Deskriptif 42
A.2.1.Excel untuk Mengukur Data Observasi 42
A.2.2.Excel untuk mengukur karakteristik sampel atau populasi 51
A.2.3.Excel untuk mengukur keragaman data 52
A.2.4.Excel untuk mengukur hubungan dua variabel ... 52
Latihan 53

BAB 3
DASAR STATISTIK INFERENSIAL..... 61
3.1. Pengertian Distribusi 63
3.2. Distribusi Normal (*Normal Distribution*)..... 67
3.3. Standarisasi Distribusi Normal 70
3.4. *Central Limit Theorem* (CLT) 70
3.5. *Standard error* 72
3.6. *Estimators* dan *Estimates* 73
Latihan 75

BAB 4
STATISTIK INFERENSIAL:
CONFIDENCE INTERVALS 77
4.1. Interval Kepercayaan (*Confidence Intervals*) 79
4.2. Klarifikasi *Confidence interval* 80
4.3. Distribusi-t (*Student's t-Distribution*)..... 83
4.4. *Confidence Intervals: Varians* 84
4.4.1. Populasi Diketahui: z-score 84
4.4.2. Populasi Tidak Diketahui: *t-score* 87
4.5. *Margin of Error* 88
4.6. Interfal keyakinan: dua rata-rata (*Confidence intervals: Two means*) 90
4.6.1. Sampel terikat (*Dependent samplesi*) 90
4.6.2. Sampel bebas (*Independent samples*) 93

Appendix Bab 4:
Excel *Inferential Statistics - Confidence Intervals*..... 101
A.4.1.Menggunakan Distribusi Normal Tabel Z..... 101
A.4.2.Menggunakan Distribusi Normal Tabel t (Distribusi-t)..... **105**
A.4.3.Confidence intervals: Two means (*Dependent samples*)..... 109
A.4.4.Confidence intervals: Two means (*Independent samples*)..... 110
Latihan 113

BAB 5
UJI HIPOTESIS..... 117
5.1. Pengertian Hipotesis..... 119
5.2. Uji hipotesis dan tingkat signifikansi..... 121
5.3. Error tipe I dan error tipe II..... 126
5.4. Uji Mean (populasi variance diketahui)..... 126
5.5. Nilai-P (P-*value*)..... 129
5.6. Uji Rata-rata: Varians Populasi Tidak Diketahui..... 133
5.7. Uji Rata-rata: Sampel Dependen..... 135
5.8. Uji rata-rata: Sampel independen..... 140
5.8.1. Sampel independent dengan varian diketahui..... 140
5.8.2. Uji rata-rata: Sampel independent dengan varian tidak diketahui diasumsikan sama..... 142

Appendix Bab 5:
Excel uji hipotesis..... 147
A.5.1.Kalkulator statistik..... 147
A.5.2.P-value untuk Z-score..... 148
A.5.3.P-value untuk t score..... 149
Latihan 151

Appendix 6:
Excel Untuk Analisa Data 156
A.6.1.Statistik Deskriptif 158
A.6.2 Histogram..... 162
A.6.3.Korelasi (*Correlation*) 165
A.6.4.Regresi linear (*Linear Regression*)..... 166

Daftar Pustaka..... 171

BAB 1

POPULASI DAN SAMPEL

Bab pertama dari buku ini, yang berjudul "Populasi dan Sampel," merupakan fondasi penting dalam memahami analisis statistik. Dalam bab ini, kita akan belajar konsep dasar yang menjadi landasan bagi seluruh proses analisis data, yaitu populasi dan sampel.

Populasi merujuk pada keseluruhan objek atau individu yang menjadi subjek penelitian, sementara sampel adalah bagian dari populasi yang dipilih untuk dianalisis. Pemahaman yang mendalam tentang kedua konsep ini sangat penting karena kesalahan dalam pemilihan sampel dapat mengakibatkan hasil analisis yang tidak akurat dan menyesatkan. Bab ini dirancang untuk memberikan pemahaman yang mendasar tentang bagaimana menentukan populasi dan memilih sampel yang representatif.

1.1. Populasi vs Sampel

Sebelum menghitung angka dan membuat keputusan, kita harus mengenal beberapa definisi kunci. Langkah pertama dari setiap analisis statistik yang kita lakukan adalah menentukan apakah data yang kita hadapi adalah populasi atau sampel. **Populasi** adalah kumpulan semua item yang menarik untuk penelitian kita, dan biasanya dilambangkan dengan huruf besar. Angka yang kita peroleh saat menggunakan populasi disebut parameter. **Sampel** adalah bagian dari populasi dan dilambangkan dengan huruf kecil n, sedangkan populasi dengan huruf besar N, dan angka yang diperoleh dengan sampel disebut statistik. Inilah bidang yang kita pelajari sekarang, yaitu statistik [1].

Contoh, kita ingin melakukan survei prospek pekerjaan para siswa yang belajar di suatu universitas NYU, berapa populasi yang ada di universitas tersebut? Populasi siswa yang akan disurvei tidak hanya mencakup siswa di kampus tetapi juga yang di rumah, yang mengikuti pertukaran di luar negeri, siswa pendidikan jarak

jauh, siswa paruh waktu, bahkan yang siswa yang mendaftar tetapi masih di sekolah menengah.

Meskipun populasi lebih lengkap, populasi sulit untuk didefinisikan dan sulit untuk diamati dalam kehidupan nyata, sedangkan sampel jauh lebih mudah untuk dikumpulkan. Menggunakan sampel memerlukan lebih sedikit waktu dan sumber daya yang lebih murah. Degan demikian hal ini adalah alasan utama kita lebih suka menggambar sampel dibandingkan dengan menganalisis seluruh populasi.

Jadi mari menentukan sampel, kita bisa pergi ke kampus NYU. Selanjutnya mari kita masuk ke kantin karena kita tahu itu akan penuh dengan orang. Kita kemudian dapat mewawancarai 50 dari mereka. Ini adalah sampel yang diambil dari populasi mahasiswa NYU. Populasi yang baik, lebih sulit untuk diamati. Itulah mengapa tes statistik dirancang untuk bekerja dengan data yang tidak lengkap. Kita akan hampir selalu bekerja dengan data sampel dan membuat keputusan serta kesimpulan berdasarkan data tersebut.

Karena uji tesis biasanya didasarkan pada sampel, data sampel adalah kunci untuk wawasan tesis yang akurat. Sampel memiliki dua karakteristik yang mendasar: ***Randomness and representativeness*** (keacakan dan keterwakilan). Sampel haruslah acak dan representatif agar suatu wawasan menjadi tepat. Sampel acak dikumpulkan ketika setiap anggota sampel dipilih dari populasi secara kebetulan. Sampel yang representatif adalah bagian dari populasi yang secara akurat mencerminkan anggota seluruh populasi [2].

Mari kita kembali ke contoh sampel yang baru saja kita bahas, yaitu 50 mahasiswa dari kantin NYU yang masuk ke kantin universitas namun melanggar kedua syarat tersebut. Orang-orang itu tidak dipilih secara kebetulan. Mereka adalah sekelompok mahasiswa NYU yang berada di sana untuk makan siang. Sebagian besar anggota bahkan tidak mendapat kesempatan untuk dipilih

karena mereka tidak berada di kantin. Jadi kita menyimpulkan bahwa sampelnya tidak acak tetapi cukup representatif. Sampel tersebut mewakili sekelompok orang, tetapi tentu saja tidak mewakili semua mahasiswa di universitas. Sampel ini mewakili orang-orang yang makan siang di kantin universitas. Sekitainya survei kita adalah tentang prospek kerja mahasiswa NYU yang makan di kantin universitas, maka kita akan melakukannya dengan baik.

Kita pasti bertanya-tanya bagaimana cara menarik sampel yang acak dan representatif. Cara yang paling aman adalah dengan mendapatkan akses ke database mahasiswa dan menghubungi individu secara acak. Namun survei semacam itu hampir tidak mungkin dilakukan tanpa bantuan dari universitas. Sepanjang kursus ini kita akan mengeksplorasi statistik sampel dan populasi. Setelah menyelesaikan kursus ini, sampel dan populasi akan menjadi sangat mudah bagi kita.

1.2. Penentuan Sampel

Penentuan sampel memiliki peran penting dalam statistik. Ada beberapa alasan untuk menggunakan sampel, yaitu: mendapatkan semua populasi adalah hal yang sulit, adanya keterbatasan dana untuk mendapatkan populasi, dan populasi yang sulit dijangkau. Oleh karena itu, sampel menjadi solusi atas keterbatasan tersebut. Berikut ini adalah beberapa pendekatan yang bisa dilakukan untuk menentukan sampel menurut Lind et al. [3], supaya dapat mencerminkan populasi.

1. ***Simple random sampling*** (sampel acak sederhana). Sampel acak sederhana adalah pendekatan sampel yang memungkinkan setiap item atau orang dalam populasi memiliki peluang yang sama untuk dimasukkan. Pada pendekatan ini, kita dapat dengan bebas memilih secara acak sampel yang akan digunakan, namun semua perlu mendapatkan kesempatan yang sama.

2. ***Systematic Random Sampling*** (Pengambilan Sampel Acak Sistematis). Pendekatan ini dilakukan dengan menentukan titik awal acak dipilih, dan kemudian setiap anggota ke-k populasi dipilih. Pada pendekatan ini, kita dapat menentukan terlebih dahulu interval (k) dengan cara populasi dibagi dengan sampel yang kita perlukan. Contohnya adalah 1000 populasi dibagi dengna 100 sampel yang diperlukan, hasilnya 10. Maka pemilihan sampel dapat dilakukan dengan menetapkan setiap 10 observasi secara berurutan, sampai jumlahnya 100.
3. ***Stratified Random Sampling*** (Pengambilan Sampel Acak Bertingkat). Populasi dibagi menjadi subkelompok, yang disebut strata, dan sampel dipilih secara acak dari setiap strata.
4. ***Cluster Sampling*** (Pengambilan Sampel Cluster). Dengan pendekatan ini, populasi dibagi menjadi beberapa kelompok menggunakan batas geografis, misalnya, yang terjadi secara alami atau batas lainnya yang lebih sesuai dengan kondisi pengamatan. Kemudian, cluster dipilih secara acak dan sampel dikumpulkan dengan memilih secara acak dari setiap cluster.

Pada akhirnya kita dapat memilih salah satu dari keempat pendekatan di atas, yang paling sesuai dengan kondisi saat akan melakukan observasi untuk mendapatkan gambaran populasi.

BAB 2

STATISTIK DESKRIPTIF

Bab kedua dari buku ini, yang berjudul "Statistik Deskriptif," bertujuan untuk memberikan pemahaman mendalam tentang cara meringkas dan menggambarkan data secara efektif. Statistik deskriptif adalah langkah awal yang sangat penting dalam analisis data, karena memungkinkan kita untuk memahami karakteristik dasar dari data yang kita miliki sebelum melangkah ke analisis yang lebih kompleks.

Dalam bagian ini, kita akan belajar berbagai teknik statistik deskriptif, termasuk ukuran pemusatan seperti mean, median, dan modus, serta ukuran penyebaran seperti range, varians, dan standar deviasi. Selain itu, kita juga akan memperkenalkan konsep distribusi data dan cara visualisasi data menggunakan grafik dan tabel. Pemahaman yang baik tentang statistik deskriptif akan membantu pembaca dalam menginterpretasikan data dengan lebih akurat dan membuat keputusan yang lebih baik berdasarkan data tersebut.

2.1. Jenis Data

Kita mempelajari statistik yang sesuai untuk melakukan tes yang berbeda. Mungkin kita dapat menggunakan pengetahuan ini sebagai batu loncatan untuk berkarir di bidang *data science*. Bagaimanapun juga, sebelum kita dapat memulai pengujian statistik lebih lanjut, kita harus berkenalan dengan jenis variabel yang biasanya kita temui; berbagai jenis variabel memerlukan berbagai jenis pendekatan statistik dan visualisasi. Oleh karena itu, mengklasifikasikan data adalah poin penting yang dipelajari.

Kita dapat mengklasifikasikan data dalam dua cara utama berdasarkan jenisnya dan pada tingkat pengukurannya. Mari kita mulai dari jenis-jenis data yang bisa kita miliki. Jenis data ada **data kategorikal dan numerik** [4]**.** Berikut penjelasannya**: Data kategorikal** menjelaskan kategori atau grup. Salah satu contohnya adalah merek mobil seperti Mercedes, BMW dan Audi. Mereka menunjukkan kategori atau kelompok yang berbeda. Contoh lain

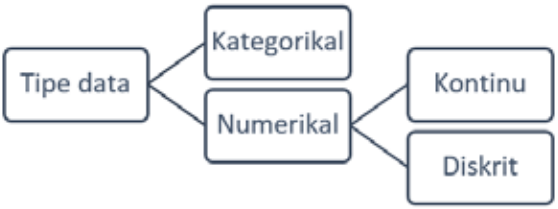
adalah jawaban untuk pertanyaan ya dan tidak. Jika kita mengajukan pertanyaan seperti, apakah kita saat ini terdaftar di universitas atau apakah kita memiliki mobil, maka jawaban ya dan tidak akan menjadi dua kelompok jawaban yang dapat diperoleh, ini adalah data kategorikal.

Data numerik merupakan data yang mewakili angka. Data ini dibagi lagi menjadi dua himpunan bagian, yaitu: **data diskrit dan kontinu** yang biasanya dapat dihitung dalam materi terbatas. Contoh data numerik yang baik adalah jumlah anak atau kendaraan yang ingin kita miliki. Bahkan jika kita tidak tahu persis berapa banyak yang kita inginkan, kita yakin bahwa nilainya akan menjadi bilangan bulat, seperti 0, 1, 2, atau bahkan 10. Contoh lain adalah nilai pada ujian. Kita mungkin mendapatkan 1.000, 1.560, atau 2.400 atau lebih. Yang penting bagi variabel untuk didefinisikan sebagai **data diskrit** adalah, kita dapat membayangkan setiap anggota kumpulan data dengan jelas. Semua skor atau nilai yang mungkin dapat diperoleh adalah kunci dari data diskrit. Contoh lainnya adalah nilai ujian dan jumlah obyek.

Data kontinu tidak terbatas dan tidak mungkin dihitung. Misalnya berat badan, yang bisa memiliki rentang dalam suatu nilai. Mari kita gali lebih dalam tentang hal ini. Kita menggunakan timbangan dan layar menunjukkan 150 pon atau 68,9 kilogram. Tapi ini hanyalah perkiraan. Jika berat badan bertambah 0,01 pon, angka pada timbangan tidak akan berubah, tapi berat badan yang baru adalah 150.01 (seratus lima puluh koma nol satu) pon atau 68,034 kilogram. Sekarang pikirkan bahwa setiap tetes keringat yang keluar akan mengurangi berat badan kita sebesar berat keringat tersebut. Namun sekali lagi, timbangan tidak mungkin dapat menangkap perubahan berat badan kita, yang merupakan contoh dari variabel atau data kontinu. Data kontinu dapat memiliki nilai yang tak terbatas, tidak peduli berapa banyak digit yang ada setelah titik. Singkatnya, berat badan kita dapat bervariasi dalam jumlah yang

sangat kecil dan bersifat kontinu, sementara jumlah anak yang ingin kita miliki dapat langsung dimengerti dan bersifat diskrit.

Selain berat badan, pengukuran lain yang juga bersifat kontinu, contohnya adalah tinggi badan, luas daerah, jarak dan waktu. Semua data ini dapat bervariasi dengan jumlah yang sangat kecil. Waktu manusia pada jam adalah diskrit tetapi waktu secara umum adalah kontinu. Nilai ini bisa berupa seperti 72,123456 detik. Pengukuran kontinu dibatasi pada berat, tinggi, jarak, dan waktu yang dapat dihitung atau diukur. Namun secara umum mereka dapat mengambil nilai berapa pun. Contoh lain, uang dapat masuk sebagai kedua jenis data tersebut, karena memiliki nilai yang kontinu, namun secara fisik bisa dihitung.



Gambar 1.1. Tipe Data

2.2. Tingkat Pengukuran (*Level of Measurement*)

Setelah kita bisa membedakan data kategorikal dan numerik, selanjutnya kita pahami bahwa data numerik dapat berupa data diskrit dan kontinu. Kita perlu memahami juga pengukuran klasifikasi lainnya. Tingkat pengukuran dapat dibagi menjadi dua kelompok: **data kualitatif dan kuantitatif** [5].

Data kualitatif sangat intuitif sehingga bisa dibagi menjadi **data nominal** atau **data ordinal**. **Data nominal** dapat dipahami seperti pada data kategorikal yang baru saja disampaikan sebelumnya, yaitu contohnya: Mercedes, BMW atau Audi, atau seperti empat musim: Musim dingin, musim semi, musim panas dan musim gugur, mereka bukan angka dan tidak dapat dipesan.

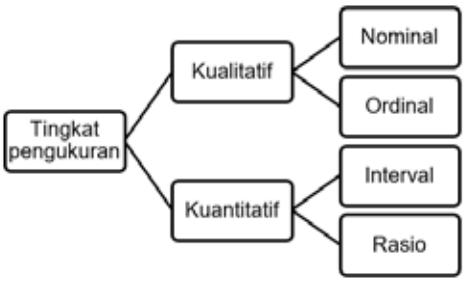
Sedangkan **data ordinal**, di sisi lain, terdiri dari kelompok-kelompok dalam kategori yang mengikuti urutan yang ketat. Bayangkan jika kita telah diminta untuk menilai makan siang dengan pilihannya antara lain: buruk, tidak menggugah selera, netral, enak, dan lezat. Meskipun kita memiliki kata-kata dan bukan angka, jelas bahwa preferensi ini diurutkan dari negatif ke positif. Dengan demikian tingkat pengukuran tersebut adalah **ordinal kualitatif**.

Jadi bagaimana dengan **data kuantitatif**, seperti yang mungkin sudah disampaikan, data kuantitatif juga dibagi menjadi dua kelompok, yaitu **data interval dan data rasio.** Interval dan rasio keduanya diwakili oleh angka, tetapi memiliki satu perbedaan utama yaitu: rasio memiliki nol sejati. Contoh data rasio, jika anda memiliki dua apel dan saya memiliki enam apel, maka saya akan memiliki tiga kali lebih banyak dari pada yang anda miliki. Sedangkan suhu adalah contoh paling umum dari data interval karena tidak memiliki nilai nol sejati.

Berikut adalah penjelasannya. Biasanya, suhu dinyatakan dalam Celsius atau Fahrenheit. Keduanya adalah variabel interval, katakanlah hari ini adalah 5 derajat Celsius atau 41 derajat Fahrenheit, dan kemarin adalah 10 derajat Celsius atau 50 derajat Fahrenheit. Dalam hal nilai Celsius, tampaknya hari ini dua kali lebih dingin (dari 10 derajat menjadi 5 derajat Celsius), tetapi dalam hal Fahrenheit tidak turun senilai dua kalinya. Masalahnya adalah berasal dari fakta bahwa nol derajat Celsius dan nol derajat Fahrenheit bukanlah nol sejati. Keduanya memiliki ukuran berbeda.

Namun demikian ada skala lain yang disebut Kelvin yang memiliki nol sejati. Nol derajat Kelvin adalah suhu di mana atom berhenti bergerak dan tidak ada yang bisa lebih dingin dari nol derajat Kelvin. Nilai nol derajat Kelvin sama dengan -273,15 derajat Celsius atau -459,67 derajat Fahrenheit. Data yang ditunjukkan dalam Kelvin adalah rasio karena memiliki nilai nol sejati, sedangkan Celsius dan Fahrenheit tidak memiliki nol sejati dan merupakan interval.

Akhirnya, angka-angka seperti dua, tiga, sepuluh, sepuluh, lima, dan seterusnya, dapat berupa interval atau rasio. Tetapi kita harus berhati-hati dengan konteks atau obyek yang sedang diobservasi atau diteliti.



Gambar 1.2. Tingkat Pengukuran

2.3. Variabel Kategorial (Teknik visualisasi)

Sekarang kita telah melihat berbagai jenis data dan tingkat pengukuran. Kita membuat grafik dan tabel yang berbeda yang akan memungkinkan kita untuk, secara visual, mewakili data yang dikerjakan. Visualisasi data merupakan cara paling intuitif atau mudah untuk menafsirkan atau memahami data. Untuk itu, jauh lebih mudah untuk memvisualisasikan data jika kita tahu jenis dan tingkat pengukurannya.

Ada dua jenis variabel atau data, yaitu: kategorikal dan numerik. Kali ini, kita akan fokus pada variabel kategorikal terlebih dahulu. Beberapa cara paling umum untuk memvisualisasikan data atau variabel kategorikal adalah dengan:

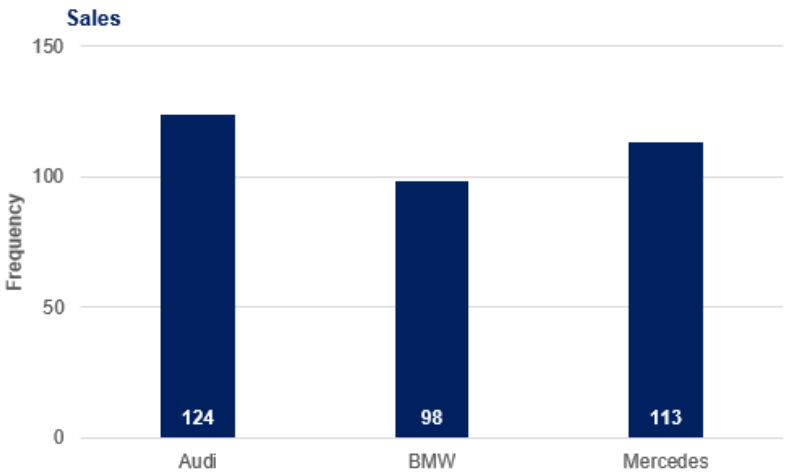
1. Tabel distribusi frekuensi,
2. Diagram batang,
3. Diagram pie atau diagram lingkaran
4. Diagram presto.

Pertama mari kita lihat seperti apa tabel distribusi frekuensi. Tabel ini memiliki dua kolom, yaitu kategori itu sendiri dan frekuensi (banyaknya data) seperti contohnya toko mobil, dan kita hanya menjual mobil Jerman. Tabel di bawah ini menunjukkan kategori mobil.

Tabel 2.1. Distribusi

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

Jumlah unit yang terjual menentukan atau menunjukkan data yang kita miliki. Dengan cara ini Kita dapat membandingkan merek yang berbeda, dan melihat bahwa Audi telah terjual paling banyak. Selain itu, kita dapat membuat diagram batang yang juga dikenal sebagai bagan kolom, sumbu vertikal menunjukkan jumlah unit yang terjual, sementara setiap batang mewakili kategori berbeda yang ditunjukkan pada sumbu horizontal. Dengan cara ini, jauh lebih jelas bahwa Audi adalah merek terlaris toko.



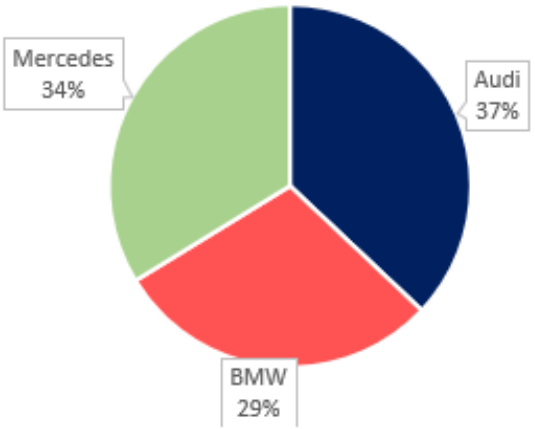
Gambar 2.2. Diagram batang

Mari kita mewakili data yang sama dengan diagram lingkaran. Untuk membuatnya, kita perlu menghitung berapa persentase dari total yang diwakili setiap merek dalam statistik. Pendekatan ini dikenal sebagai frekuensi relatif. Secara alami, semua frekuensi relatif menambahkan hingga 100 persen, diagram lingkaran sangat berguna ketika kita ingin tidak hanya membandingkan item antara satu sama lain, tetapi juga melihat bagian mereka dari total.

Tabel 2.2. Frekuensi relative *(Relative frequency)*

	Frekuensi	*Relative frequency*
Audi	124	37%
BMW	98	29%
Mercedes	113	34%
Total	335	100%

Dari tabel tersebut dapat kita buat diagram pie sebagai berikut:



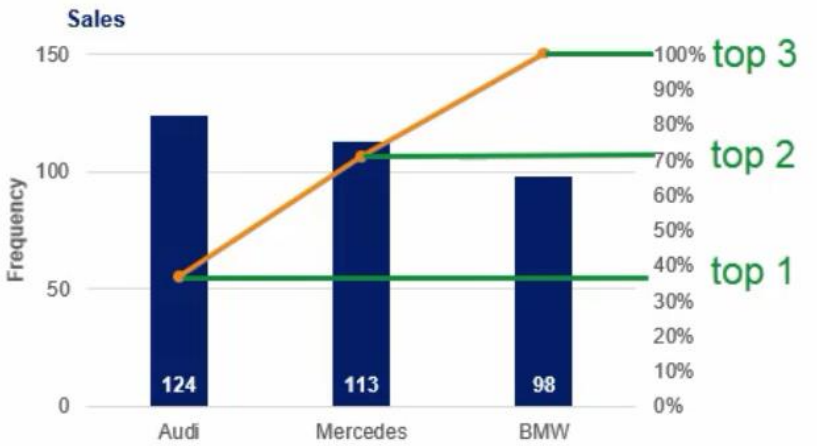
Gambar 2.3. Diagram pie atau diagram lingkaran

Diagram pie ini dapat digunakan untuk menunjukkan pangsa pasar. Pangsa pasar sangat didominasi oleh diagram lingkaran sehingga, jika kita mencari pangsa pasar pada Google,

kita akan mendapatkan diagram lingkaran. Bayangkan bahwa data dan tabel tersebut mewakili penjualan Audi, BMW, dan Mercedes di suatu daerah, maka bagan akan menunjukkan kepada kita pangsa pasar yang dimiliki masing-masing merek tersebut.

Terakhir kita memiliki diagram pareto. Faktanya, diagram pareto tidak lebih dari jenis diagram batang khusus di mana kategori ditampilkan dalam urutan frekuensi yang menurun menurut frekuensi, yang berarti jumlah kemunculan setiap item tersebut. Seperti yang dikatakan sebelumnya dalam contoh, jumlah tersebut menunjukkan unit yang terjual.

Mari kita kembali ke tabel distribusi frekuensi kita dan memesan merek berdasarkan frekuensi. Sekarang kita dapat membuat diagram batang berdasarkan tabel penyusunan ulang (diurutkan dari tertinggi sampai terendah), maka kita dapat membuat diagram pareto dengan menjadikannya satu. Kurva pada grafik yang sama menunjukkan frekuensi kumulatif. Sedangkan, frekuensi kumulatif adalah jumlah dari frekuensi relatif, yang dimulai dari frekuensi merek pertama dan kemudian kita menambahkan yang kedua, ketiga, dan seterusnya hingga selesai pada 100 persen.



Gambar 2.4. Diagram pareto

Pada gambar 2.4. garis kurva tersebut diukur dengan sumbu vertikal yang berbeda di sebelah kanan grafik. Pada setiap simpulnya menunjukkan subtotal kategori di sebelah kirinya. Kita dapat melihat bagaimana diagram pareto menggabungkan sisi kuat batang dan diagram lingkaran. Sangat mudah untuk membandingkan data baik antara kategori dan sebagai bagian dari total. Selanjutnya, jika ini adalah grafik pangsa pasar, maka kita dapat dengan mudah melihat pangsa pasar dari dua atau lima perusahaan teratas.

Kita mungkin pernah mendengar ide lain adalah prinsip Pareto yang juga dikenal sebagai aturan 80-20. Ide ini menyatakan bahwa 80 persen dari efek berasal dari 20 persen dari penyebab. Contoh nyata adalah pernyataan oleh Microsoft bahwa dengan memperbaiki 20 persen bug perangkat lunaknya, mereka berhasil menyelesaikan 80 persen masalah yang dialami pelanggan. Diagram paradigma dapat mengungkapkan informasi seperti tersebut yang menunjukkan perubahan total dengan perubahan setiap kategori tambahan, dan memberi kita pemahaman yang lebih baik tentang data. Ini adalah cara utama di mana kita dapat secara visual mewakili data kategorikal.

2.4. Variabel Numerik (Tabel Distribusi Frekuensi)

Kita sudah tahu cara membuat grafik dan tabel untuk variabel kategorikal. Kita akan melakukan hal yang sama untuk variabel numerik, mengingat bahwa data numerik adalah fokus utama dari pembalajaran ini, Jadi seperti yang kita lakukan dengan variabel kategorikal. Mari kita mulai dengan membuat tabel distribusi frekuensi.

Berikut daftar 20 nomor yang berbeda. Jika kita mengaturnya dalam tabel frekuensi seperti yang kita gunakan untuk variabel kategorikal, maka kita akan mendapatkan tabel dengan 20 baris yang masing-masing mewakili satu angka dengan frekuensi satu yang sesuai karena setiap angka terjadi tepat satu kali. Tabel ini

tidak praktis digunakan untuk analisis apa pun (seperti gambar 2.5). Ketika kita berurusan dengan variable atau data numerik, jauh lebih masuk akal untuk mengelompokkan data ke dalam interval dan kemudian menemukan frekuensi yang sesuai. Dengan cara ini kita membuat ringkasan data yang memungkinkan representasi visual yang bermakna.

Dataset	Frequency
1	1
9	1
22	1
24	1
32	1
41	1
44	1
48	1
57	1
66	1
70	1
73	1
75	1
76	1
79	1
82	1
87	1
89	1
95	1
100	1
Total	20

Gambar 2.5. Tabel frekuensi numerik

Bagaimana kita memilih interval ini. Umumnya data analis lebih suka bekerja dengan kelompok data yang berisi 5 hingga 20 interval. Namun, ini bervariasi untuk setiap kasus, dan pilihan interval yang benar sangat tergantung pada jumlah data yang kita kerjakan. Dalam contoh ini, kita akan membagi data menjadi lima interval dengan panjang yang sama. Rumus sederhana yang kita gunakan adalah sebagai berikut:

$$\text{Lebar interal} = \frac{\text{(Jumlah terbesar} - \text{angka terkecil)}}{\text{Jumlah interval yang diinginkan}}$$

Interval yang sama dengan angka terbesar dikurangi angka terkecil dibagi dengan jumlah interval yang diinginkan. Dalam kasus kita, panjang interval harus (100 – 1) / 5. Hasilnya adalah 19,8 yang dapat dibulatkan menjadi 20. Sehingga interval adalah per 20. Oleh karena itu, interval kita adalah sebagai berikut. 1 hingga 21, 21 hingga 41, 41 hingga 61, 61 hingga 81, dan 81 hingga 101.

Tabel 2.1. Frekuensi numerik

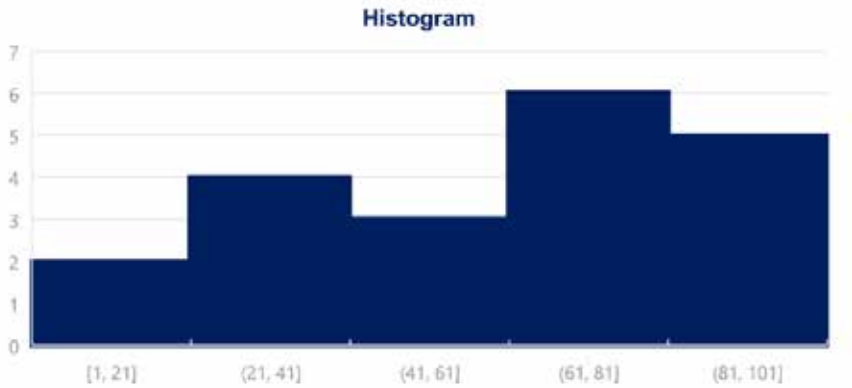
No	Interval awal	Interval akhir	Frekuensi	Frekuensi relatif
1	1	21	2	0.10
2	21	41	4	0.20
3	41	61	3	0.15
4	61	81	6	0.30
5	81	101	5	0.25

Setiap interval memiliki lebar 20. Mari kita coba untuk membangun tabel distribusi frekuensi. Angka yang masuk dalam interval tertentu adalah jika angka itu lebih besar dari batas terendah dan sama dengan atau kurang dari jumlah terbesar, seperti yang dapat kita lihat dari tabel 2.1. Ada nilai 2 dalam interval pertama, 4 untuk interval kedua, 3 untuk interval ketiga, 6 untuk interval keempat, dan 5 di interval kelima. Untuk banyak analisis, penting untuk menghitung frekuensi relatif dari titik data di setiap interval. Seperti yang sampaikan sebelumnya, frekuensi relatif adalah frekuensi interval tertentu sebagai bagian dari total.

Kita dapat tambahkan kolom lain ke tabel 2.1. dengan nama **frekuensi relatif**. Jadi interval dari 1 sampai 21 memiliki frekuensi absolut 2 tetapi frekuensi relatif dua dibagi dengan jumlah 20 angka yang memberi kita 10 persen dan seterusnya sampai kita mengisi tabel. Selanjutnya, beginilah cara kita menghitung frekuensi relatif.

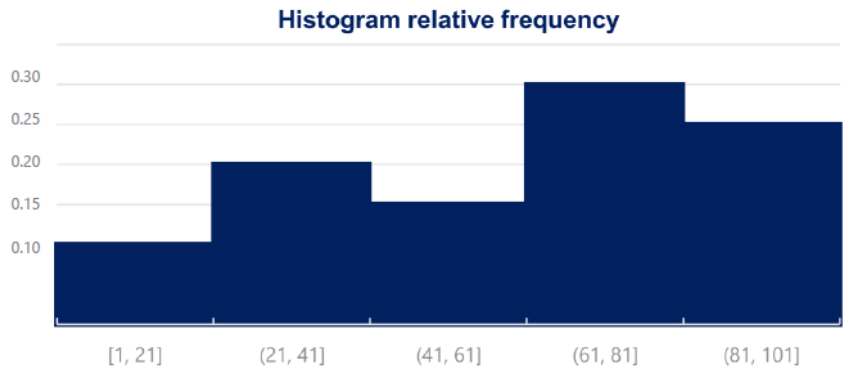
2.5. Histogram

Grafik yang paling umum digunakan untuk mewakili data numerik adalah histogram, pertama kita akan belajar cara membuatnya, dan kemudian kita akan memberikan deskripsi tentang cara data dapat diwakili. Kita akan menggunakan tabel distribusi frekuensi dari contoh sebelumnya (tabel 2.1) untuk membantu kita. Histogram tampak seperti diagram batang tetapi sebenarnya menyampaikan informasi yang sangat berbeda. Sumbu vertikal adalah tipe numerik dan menunjukkan frekuensi absolut.



Gambar 2.6. Diagram histogram

Kali ini, meskipun sumbu horizontal adalah numerik, setiap bar memiliki interval yang sama dan tinggi yang sama dengan frekuensi. Perhatikan bagaimana bar yang berbeda saling menyentuh atau bersambung. Hal itu menunjukkan bahwa ada kelanjutan antara interval, setiap interval berakhir di mana yang berikutnya dimulai di diagram batang, batang batang yang berbeda mewakili kategori yang berbeda. Jadi jeruji itu benar-benar terpisah.



Gambar 2.7. Diagram histogram frekuensi relatif

Kadang-kadang berguna untuk menentukan interval terhadap relatif daripada frekuensi absolut karena histogram dapat terlihat sama secara visual tetapi memberikan informasi yang berbeda kepada audiens, ingat frekuensi relatif terdiri dari persentase.

Kita dapat membuat histogram dengan interval yang sama. Contohnya adalah merancang kelompok umur dalam suatu survei. Jawaban yang mungkin adalah 18 hingga 25, lalu 26 hingga 30, 31 hingga 35, dan seterusnya hingga 60. Interval tersebut mencerminkan kelompok fokus yang berbeda, sehingga memiliki penjelasan bahwa mungkin orang dewasa muda di bawah 18 tidak mampu membeli produk sementara orang dewasa di atas 60 tidak tertarik pada produk.

Bagaimanapun juga, kita harus cukup memahami data untuk merancang dan menafsirkan kelompok-kelompok tersebut secara akurat. Dalam hal ini, pemahaman teori dan konteks menjadi sangat penting. Sangat disarankan agar kita tetap dengan interval yang sama sampai kita mendapatkan pengalaman yang cukup. Ini adalah bagaimana kita dapat membangun histogram untuk mewakili data numerik.

2.6. *Cross Table* dan *Scaterplot*

Sejauh ini kita telah membahas grafik yang hanya mewakili satu variable atau data tetapi bagaimana kita mewakili hubungan antara dua variabel. Selanjutnya, kita akan menjelajahi tabel silang (*cross table*) dan *scatterplot*. Sekali lagi kita memiliki pembagian antara variabel kategorikal dan numerik. Mari kita mulai dengan variabel kategorikal.

Cara paling umum untuk mewakili variable kategorikal adalah menggunakan tabel silang atau tabel kontingensi. Bayangkan kita adalah seorang manajer investasi dan mengelola saham, obligasi, dan investasi real estat untuk tiga investor berbeda. Masing-masing dari mereka memiliki gagasan risiko yang berbeda dan karenanya uang mereka dialokasikan dengan cara yang berbeda di antara tiga kelas aset di seluruh tabel yang mewakili semua data terlihat dengan cara berikut.

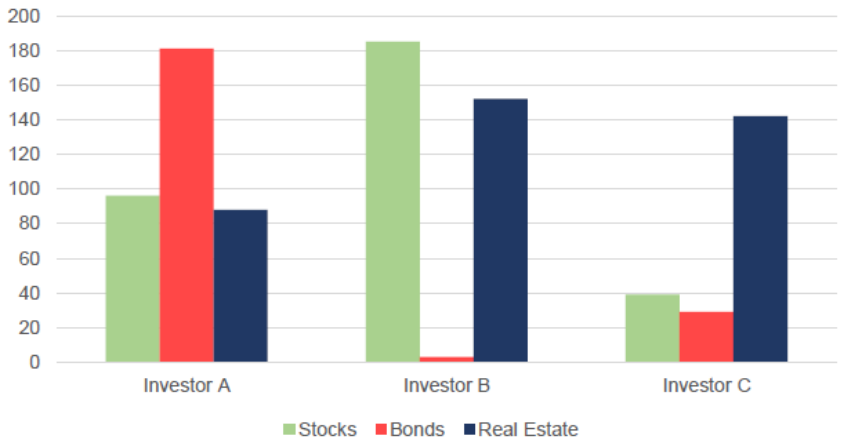
Kita dapat dengan jelas melihat baris yang menunjukkan jenis investasi yang telah dilakukan dan kolom dengan alokasi masing-masing investor. Ini adalah praktik yang baik untuk menghitung total setiap baris dan kolom karena sering berguna dalam analisis lebih lanjut.

Tabel 2.2. Data Investasi

Tipe investasi	Investor A	Investor B	Investor C	Total
Saham/*stock*	96	185	39	320
Obligasi/*bond*	181	3	29	213
Real estate	88	152	142	382
Total	365	340	210	915

Perhatikan bahwa subtotal baris memberi kita total investasi dalam saham, obligasi, dan real estate. Di sisi lain, subtotal kolom memberi kita kepemilikan masing-masing investor, setelah kita membuat tabel silang, kita dapat melanjutkan dengan

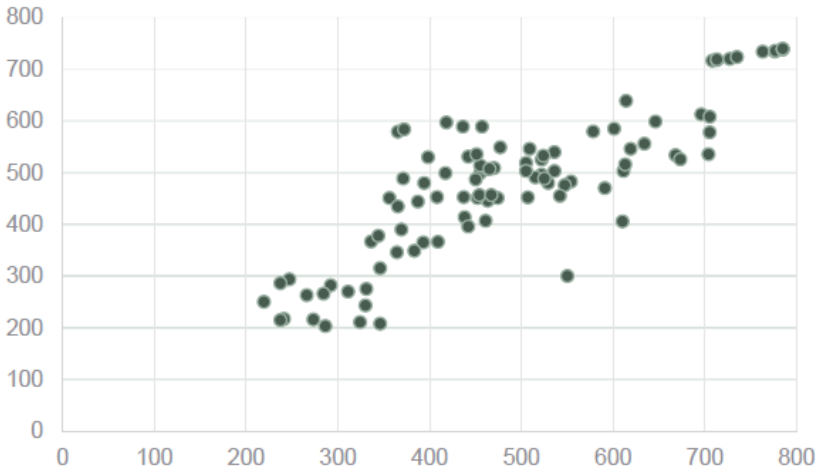
memvisualisasikan data. Grafik yang sangat berguna dalam kasus seperti itu adalah variasi dari diagram batang yang disebut s*ide by side bar chart.*



Gambar 2.8. Diagram batang *side by side bar chart*

Gambar 2.8. tersebut mewakili kepemilikan masing-masing investor dalam berbagai jenis aset, saham dalam obligasi hijau atau merah. Dan *real estate* berwarna biru. Nama jenis grafik ini berasal dari fakta bahwa untuk setiap investor kategori aset diwakili berdampingan dengan cara ini. Kita dapat dengan mudah membandingkan kepemilikan aset untuk investor tertentu atau di antara investor. Semua grafik sangat mudah dibuat dan dibaca setelah kita mengidentifikasi jenis data yang kita hadapi, dan memutuskan cara terbaik untuk memvisualisasikannya. Akhirnya kita dapat menyimpulkan bahwa grafik menjadi sangat penting.

Scatterplot digunakan dalam mewakili dua variabel numerik. Untuk contoh ini kita telah mengumpulkan skor membaca dan menulis S.A.T. dari 100 individu. Mungkin kita perlu melihat grafik terlebih dahulu sebelum menganalisisnya lebih lanjut.



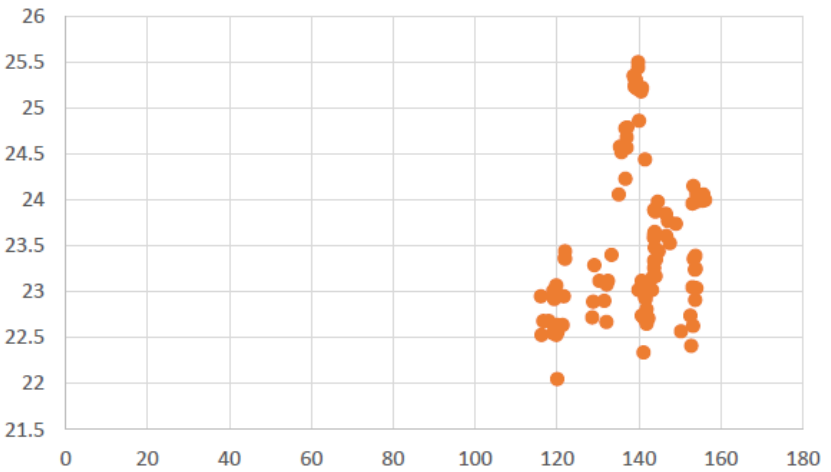
Gambar 2.9. Diagram scatterplot hubungan skor membaca dan menulis

Skor S.A.T. pertama berdasarkan rentang komponen dari 200 hingga 800 poin, dan itulah sebabnya data kita dibatasi dalam kisaran 200 hingga 800 poin. Sumbu vertikal menunjukkan skor menulis, sedangkan sumbu horizontal berisi skor membaca. Dalam observasi ada 100 siswa, dan hasilnya sesuai dengan titik tertentu pada grafik. Setiap poin atau titik memberi kita informasi tentang kinerja siswa tertentu, contohnya siswa dengan nama Jane. Dia mencetak 300 pada menulis tetapi 550 pada bagian membaca, maka dapat digambarkan dengan satu titik atau poin perpotongan dari dua nilai tersebut.

Scatterplot biasanya mewakili banyak pengamatan, oleh karena itu, saat menafsirkan scatterplot, kita tidak diharapkan untuk melihat ke dalam titik data tunggal. Kita perlu lebih tertarik untuk mendapatkan ide utama tentang bagaimana data didistribusikan. Hal **pertama** yang kita lihat adalah bahwa ada tren naik yang jelas. Hal ini dikarenakan nilai menulis yang lebih rendah biasanya diperoleh oleh siswa dengan nilai membaca yang lebih rendah dan nilai menulis yang lebih tinggi telah dicapai oleh siswa dengan nilai

membaca yang lebih tinggi. Hal ini adalah hal yang logis. Siswa lebih cenderung melakukan keduanya dengan baik karena kedua tugas tersebut terkait erat. Kondisi tersebut (gambar 2.9) menunjukkan ada pola hubungan positif antara kemampuan membaca dan menulis, karena peningkatan satu faktor diikuti dengan peningkatan yang faktor yang lain. Sedangkan jika gambar tidak menunjukkan pola maka kondisi tersebut menunjukkan tidak adanya hubungan kedua faktor bidang yang diobservasi (gambar 2.10)

Hal **kedua** adalah, kita melihat konsentrasi siswa di tengah grafik dengan skor di wilayah 450 hingga 550 pada membaca dan menulis. Ingat kita mengatakan bahwa skor bisa di mana saja antara 200 dan 800. Scor 500 adalah skor rata-rata yang bisa didapat seseorang. Jadi masuk akal bahwa banyak orang jatuh ke daerah itu.



Gambar 2.10. Diagram scatterplot tidak ada hubungan

Hal **ketiga**, ada kelompok orang dengan nilai menulis dan membaca yang sangat tinggi. Siswa luar biasa cenderung sangat baik di kedua komponen. Kondisi ini kurang benar untuk siswa yang buruk karena kinerja mereka cenderung menyimpang ketika melakukan tugas yang berbeda. Akhirnya kita memiliki informasi

bahwa Jane, jauh dari setiap pengamatan lain karena dia mendapat nilai di atas rata-rata dalam membaca tetapi buruk dalam menulis. Pengamatan ini disebut **outlier** karena bertentangan dengan logika seluruh kumpulan data.

Jadi kita telah melalui dasar-dasarnya. Kita telah membahas jenis-jenis variabel sampel populasi, grafik dan tabel, dan sekarang saatnya untuk menyelami inti statistik deskriptif, pengukuran kecenderungan dan variabilitas sentral.

2.7. Mengukur Karakter Sampel atau Populasi

Pengukuran karakter sampel atau populasi dilakukan dalam rangka untuk menilai apakah sampel atau populasi cenderung homogen atau heterogen.

2.7.1. *Mean*, Median, Modus

Bagian ini kita akan membahas tiga ukuran kecenderungan sentral dari data set atau observasi. Ukuran pertama yang akan kita pelajari adalah ***mean***, yang disebut juga sebagai rata-rata sederhana (*simple average*) yang dilambangkan dengan huruf Yunani μ (miu) untuk populasi, dan simbol $\overline{x}$ untuk sampel [6]. Kita dapat menemukan rata-rata kumpulan data dengan menjumlahkan semua komponennya dan kemudian membaginya dengan jumlahnya. Rumusnya adalah sebagai berikut (n merupakan jumlah data yang diobservasi):

$$\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{atau} \quad \frac{x_1 + x_2 + x_3 + \cdots + x_{N-1} + x_N}{N}$$

Rata-rata adalah ukuran paling umum dari kecenderungan sentral tetapi memiliki kelemahan besar, yaitu hasilnya mudah dipengaruhi nilai ekstrim.

Berikut adalah contohnya.

Tabel 2.2. Contoh data outlier pada harga pizza

Posisi	New York City	Los Angeles
1	$ 1,00	$ 1,00
2	$ 2,00	$ 2,00
3	$ 3,00	$ 3,00
4	$ 3,00	$ 4,00
5	$ 5,00	$ 5,00
6	$ 6,00	$ 6,00
7	$ 7,00	$ 7,00
8	$ 8,00	$ 8,00
9	$ 9,00	$ 9,00
10	$ 11,00	$ 10,00
11	$ 66,00	

Tabel 2.3. di atas adalah harga pizza di 11 lokasi berbeda di New York City (NYC) dan 10 lokasi berbeda di LA. Mari kita hitung rata-rata dari dua kumpulan data menggunakan rumus untuk rata-rata. Di NYC kita mendapatkan $ 11 sedangkan untuk LA hanya $ 5.5, yang artinya rata-rata pizza di New York bisa dua kali lebih mahal daripada di LA. Masalahnya adalah bahwa dalam sampel tersebut, telah terdapat satu tempat mewah di New York di mana mereka mengenakan biaya $ 66 untuk pizza. Nilai tersebut mengakibatkan peningkatan nilai rata-rata. Dari contoh ini dapat diketahui bahwa rata-rata tidak cukup untuk membuat kesimpulan yang pasti. Jadi bagaimana kita bisa mencegahi dari masalah tersebut. Kita dapat menghitung ukuran kedua yaitu: **Median** [6].

Median pada dasarnya adalah angka tengah dalam himpunan data yang diurutkan. Median adalah titik tengah dari himpunan data yang dipesan. Ini tidak sepopuler rata-rata tetapi sering digunakan dalam akademisi dan ilmu data. Itu karena tidak terpengaruh oleh outlier. Dalam himpunan data yang diurutkan, median adalah angka pada posisi. Rumusnya adalah (n + 1) / 2, (n

adalah jumlah data yang diobservasi). Jika posisi ini bukan bilangan bulat, median adalah rata-rata sederhana dari dua angka pada posisi yang paling dekat dengan nilai yang dihitung.

Mari kita lihat cara kerjanya, misalnya, untuk menghitung median kita harus mengurutkan data kita dalam urutan berurutan. Median kumpulan data adalah angka pada posisi dan ditambah 1 dibagi 2 dalam daftar terurut di mana n adalah jumlah pengamatan, oleh karena itu median untuk NYC berada di posisi keenam atau enam dolar, jauh lebih dekat dengan harga yang diamati daripada rata-rata $ 11. Bagaimana dengan L.A. Kita hanya memiliki 10 pengamatan di L.A. Menurut rumus kita, median adalah posisi 5.5. Dalam kasus seperti ini, median adalah rata-rata sederhana dari angka pada posisi 5 dan 6. Oleh karena itu pada median harga LA adalah $ 5,5. Kita telah melihat median tidak terpengaruh oleh harga ekstrim.

Posisi	New York City	Los Angeles
1	$ 1,00	$ 1,00
2	$ 2,00	$ 2,00
3	$ 3,00	$ 3,00
4	$ 3,00	$ 4,00
5	$ 5,00	$ 5,00
6	$ 6,00	$ 6,00
7	$ 7,00	$ 7,00
8	$ 8,00	$ 8,00
9	$ 9,00	$ 9,00
10	$ 11,00	$ 10,00
11	$ 66,00	

Kita perlu mengetahui ukuran lain, yaitu modus. **Modus** (mode) adalah nilai yang paling sering terjadi. Ini dapat digunakan untuk data numerik dan kategorikal tetapi kita akan tetap berpegang pada contoh numerik. Setelah menghitung frekuensi setiap nilai, kita menemukan bahwa modus harga pizza The New York adalah

$ 3. Hal ini menarik karena harga pizza yang paling umum di NYC hanya $ 3 tetapi rata-rata dan median membuat kita percaya itu jauh lebih mahal. Jadi mari kita lakukan hal yang sama dan temukan mode harga pizza LA dari setiap harga hanya muncul sekali. Bagaimana kita menemukan yang paling banyak? Kita dapat mengatakan bahwa tidak ada nilai modus karena tidak ada nilai yang muncul lebih dari satu kali.

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50
Mode	$ 3.00	-

Secara umum, kita sering memiliki beberapa modus, biasanya dua atau tiga modus dapat ditoleransi, tetapi jika lebih dari itu akan mengalahkan tujuan menemukan modus. Ada satu pertanyaan terakhir yang belum kita jawab ukuran mana yang terbaik: contoh menunjukkan kepada kita bahwa ukuran kecenderungan sentral harus digunakan bersama dari pada secara independent atau terpisah. Karena itu tidak ada yang terbaik. Tetapi hanya menggunakan satu jelas merupakan yang terburuk.

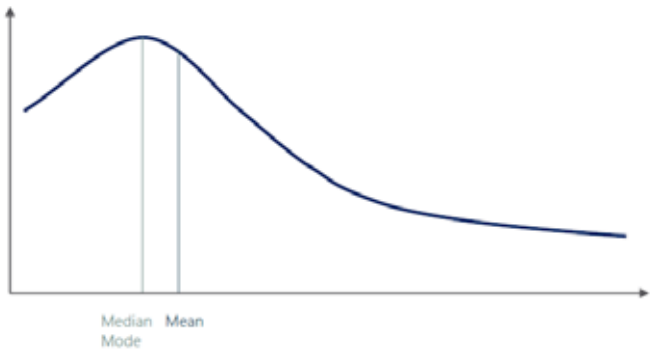
2.7.2. Kecondongan (*Skewness*)

Setelah menjelajahi ukuran kecenderungan sentral, saat ini kita beralih ke ukuran asimetri. Alat yang paling umum digunakan untuk mengukur simetri adalah kecondongan atau ***skewness*** [6]. Pendekatan ini adalah rumus yang hampir selalu digunakan perangkat lunak yang melakukan perhitungan deskriptif.

Rumusnya adalah:

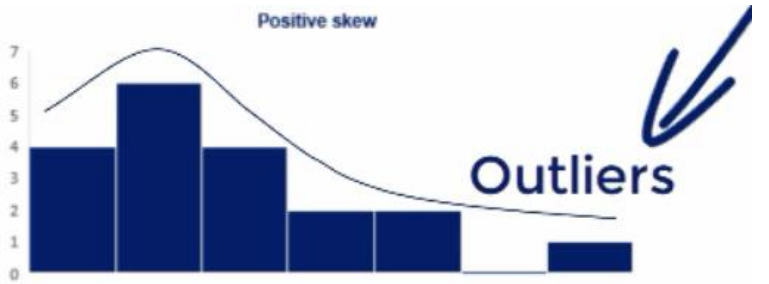
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^3}$$

Namun dalam pembahasan ini kita tidak akan masuk ke perhitungan melainkan arti dari kecondongan, sehingga kecondongan menunjukkan apakah pengamatan dalam dataset terkonsentrasi pada satu sisi. Kecondongan kanan (positif) terlihat seperti yang ada di grafik. Kondisi ini berarti bahwa outlier berada di sebelah kanan (*long tail* atau ekor ke kanan). Kecondongan kiri (negatif) berarti *outlier* berada di sebelah kiri.



Gambar 2.11. Skewness

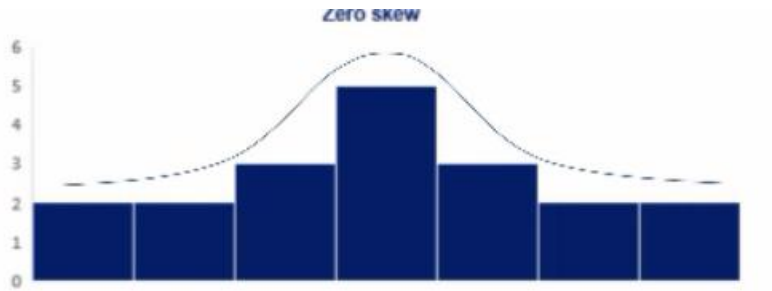
Jadi contohnya ada di tempat mengingat tabel distribusi frekuensi dari materi sebelumnya. Di sini kita memiliki tiga set data dan distribusi frekuensi masing-masing. Kita juga telah menghitung median dan mode rata-rata. Dataset pertama memiliki *mean* 2,79 dan median 2, maka rata-rata lebih besar dari median. Kita mengatakan bahwa ini adalah kemiringan positif atau kanan dari grafik.



Gambar 2.12. Skewness positif

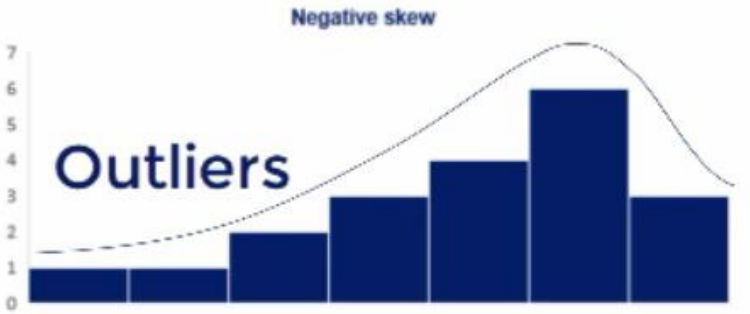
Kita dapat dengan jelas melihat bahwa titik data terkonsentrasi di sisi kiri. Perhatikan bahwa arah kemiringan berlawanan dengan intuisi. Itu tidak tergantung pada sisi mana garis itu mengarah tetapi lebih ke sisi mana ekornya condong. Jadi kecondongan kanan berarti bahwa outlier berada di sebelah kanan.

Pada gambar 2.13, kita telah memplot kumpulan data yang memiliki median dan modus rata-rata yang sama, frekuensi kemunculannya benar-benar simetris, dan kita menyebutnya nol atau tidak ada kemiringan, paling sering Kita akan mendengar orang mengatakan bahwa distribusinya simetris untuk kumpulan data. Pada kondisi tersebut nilai *mean* sama dengan nilai median dan modus.



Gambar 2.14. Skewness nol

Kita memiliki nilai 4,9, median 5 dan modus 6. Dengan demikian, karena rata-rata lebih rendah dari median, kita dapat mengatakan bahwa ada kemiringan negatif atau kiri. Sekali lagi titik tertinggi ditentukan oleh modus. Mengapa disebut condong kiri adalah benar, karena outlier ada di sebelah kiri.



Gambar 2.14. Skewness negatif

Mengapa kecondongan penting, kecondongan memberi tahu kita banyak hal tentang di mana data berada. Seperti yang kita sebutkan dalam pelajaran sebelumnya, median, rata-rata dan modus harus digunakan bersama untuk mendapatkan pemahaman yang baik tentang dataset. Ukuran asimetri seperti kecondongan atau hubungan antara ukuran kecenderungan sentral dan teori probabilitas yang pada akhirnya memungkinkan kita untuk mendapatkan pemahaman yang lebih lengkap tentang data yang sedang kita kerjakan.

2.8. Mengukur Keragaman atau Variabilitas

Selanjutnya kita membahas ukuran variabilitas. Ada banyak cara untuk mengukur variabilitas. Namun, kita akan fokus pada yang paling umum: varians, standar deviasi, dan koefisien variasi di bidang statistik, yang biasanya akan menggunakan rumus yang berbeda ketika bekerja dengan data populasi dan data sampel [6].

2.8.1. Varians (*Variance*)

Ketika kita memiliki seluruh populasi, dan setiap titik data populasi tersebut diketahui, maka kita dapat 100% yakin dengan ukuran yang kita hitung. Ketika kita mengambil sampel populasi ini dengan menghitung statistik sampel, maka kondisi ini

ditafsirkan sebagai perkiraan parameter populasi. Selain itu, jika kita mengekstrak 10 sampel berbeda dari populasi yang sama, kita akan mendapatkan 10 ukuran berbeda. Untuk itu, statistik memecahkan masalah tersebut dengan menyesuaikan rumus aljabar. Oleh karena itu kita akan mengeksplorasi rumus populasi dan sampel karena keduanya digunakan.

Mungkin muncul pertanyaan dari sini, mengapa ada rumus unik untuk median dan modus, dan rata-rata (mean). Sebenarnya rata-rata sampel adalah rata-rata dari titik data sampel. Sedangkan rata-rata populasi adalah rata-rata dari titik data populasi. Jadi secara teknis ada dua rumus yang berbeda tetapi mereka dihitung dengan cara yang sama.

Sekarang, setelah mengingat kembali rata-rata sampel dan populasi, saatnya untuk beralih ke varian. **Varians mengukur *dispersi* atau sebaran satu set titik data di sekitar nilai rata-rata** [1], [6]. Semakin besar nilai varians maka kondisi tersebut menunjukkan data observasi yang semakin beragam, dan sebaliknya. Varians populasi yang dilambangkan dengan Sigma kuadrat, sama dengan jumlah perbedaan kuadrat antara nilai yang diamati dan rata-rata populasi dibagi dengan jumlah total pengamatan.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

Varians sampel di sisi lain dilambangkan dengan s kuadrat dan sama dengan jumlah perbedaan kuadrat antara nilai sampel yang diamati dan rata-rata sampel dibagi dengan jumlah pengamatan sampel dikurangi satu.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

Mungkin kita akan kesulitan untuk memahami semuanya dengan segera. Karena itu, mari kita memeriksa formula untuk populasi dan mencoba mengklarifikasi maknanya. Bagian utama dari rumus ini adalah pembilangnya.

Jumlah perbedaan antara pengamatan dan rata-rata kuadrat mereka. Jadi semakin dekat angka dengan rata-rata, semakin rendah nilai varian yang akan kita peroleh, dan semakin jauh dari rata-rata dari nilai rata-rata maka nilai varian semakin besar. Tetapi mengapa kita perlu melakukan pangkat 2 atau kuadrat? Mengkuadratkan perbedaan memiliki tujuan, yaitu: dengan mengkuadratkan angka, kita selalu mendapatkan perhitungan non-negatif. Tanpa masuk terlalu jauh ke dalam matematika itu intuitif bahwa disperse (jarak/ sebaran) tidak bisa negatif, dispersi adalah tentang jarak dan jarak tidak bisa negatif.

2.8.2. Standar Deviasi (*Standard Deviation*)

Varians adalah ukuran umum dari dispersi data. Dalam kebanyakan kasus, angka yang akan kita peroleh cukuo besar dan sulit dibandingkan karena unit pengukuran dikuadratkan. Perbaikan yang mudah adalah menghitung akar kuadratnya dan mendapatkan statistik yang dikenal sebagai standar deviasi [6]. Dalam kebanyakan analisis yang kita lakukan, standar deviasi akan jauh lebih bermakna daripada varians.

Ada langkah-langkah yang berbeda untuk populasi dan sampel, akibatnya adalah adanya standar deviasi populasi dan sampel. Rumusnya adalah akar kuadrat dari varians populasi dan akar kuadrat dari varians sampel masing-masing. Rumus untuk standar deviasi sampel adalah sebagai berikut:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Sedangkan rumus untuk standar deviasi populasi adalah:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Ukuran lain yang masih harus kita perkenalkan adalah koefisien variasi. Ukuran ini sama dengan standar deviasi dibagi dengan rata-rata, nama lain untuk istilah ini adalah standar deviasi relatif. Sekali lagi, standar deviasi adalah ukuran variabilitas yang paling umum untuk satu dataset. Tetapi mengapa kita membutuhkan ukuran lain, seperti koefisien variasi, karena membandingkan standar deviasi dari dua kumpulan data yang berbeda tidak ada artinya. Namun membandingkan dua koeffisien variasi dapat memberikan arti yang lebih baik.

Berikut adalah contoh perbandingan antara standar deviasi. Mari kita ambil harga pizza di 10 tempat berbeda di New York. Harganya berkisar dari 1 Dollar hingga 11 Dollar. Sekarang bayangkan bahwa Kita hanya memiliki Peso Meksiko dan bagi Kita harganya lebih mirip 18,81 Peso hingga 206,91 Peso.

NY Dollars		Pesos	
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

Mengingat nilai tukar 18,81 Peso seharga 1 Dolar. Mari gabungkan apa yang sudah dipelajari sejauh ini dan temukan standar deviasi dan koefisien variasi dari dua set data ini. Pertama kita harus melihat apakah objek ini adalah sampel atau populasi. Apakah hanya ada 11 restoran di New York. Tentu saja tidak sehingga objek observasi ini jelas merupakan sampel yang diambil dari semua restoran di kota. Kemudian kita harus menggunakan rumus untuk ukuran sampel variabilitas. Kedua, kita harus menemukan rata-rata: rata-rata dalam dolar sama dengan 5,5 dan rata-rata dalam peso menjadi 100 dalam 3,46.

	Dollars		Pesos
Mean	$ 5.50	MXN	103.46
Sample variance	$^2 10.72	MXN^2	3793.69
Sample standard deviation	$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60

Langkah ketiga dari proses ini adalah menemukan varians sampel. Mengikuti rumus yang ditunjukkan sebelumnya, kita dapat memperoleh varians sebesar 10,72 Dollar dan 3793,69 Peso, standar deviasi sampel masing-masing adalah 3,27 Dollar dan 61,59 Peso. Mari kita lakukan beberapa pengamatan. Varian pertama memberikan hasil dalam satuan kuadrat sedangkan simpangan baku atau standar deviasi dalam satuan asli. Hal ini adalah alasan utama mengapa para profesional lebih suka menggunakan standar deviasi sebagai ukuran utama variabilitas. Kedua, kita mendapat standar deviasi 3,27, dan 61,59 untuk pizza yang sama di sebelas restoran yang sama di New York City. Sepertinya dua nilai tersebut sangat berbeda. Oleh karena itu, sudah waktunya untuk menggunakan alat terakhir. Koefisien variasi membagi standar deviasi dengan cara masing-masing kita mendapatkan dua koefisien variasi. Hasilnya sama 0,60, yang berarti keduanya memiliki varians yang sama. Perhatikan bahwa itu bukan Dollar atau Peso, bukan Dollar kuadrat

atau Peso kuadrat, hanya nilai 0,60. Hasil ini menunjukkan kepada kita arti yang diberikan koefisien variasi.

Sekarang kita dapat dengan yakin mengatakan bahwa kedua kumpulan data memiliki variabilitas yang sama. Mari kita rekap apa yang telah kita pelajari sejauh ini. Ada tiga ukuran utama variabilitas: varians, standar deviasi, dan koefisien variasi. Masing-masing memiliki kekuatan dan aplikasi yang berbeda.

2.9. Mengukur Hubungan Dua Variabel

Kita telah membahas semua tindakan univariat. Sekarang saatnya untuk melihat ukuran yang digunakan ketika kita bekerja dengan lebih dari satu variabel dalam dua pelajaran berikutnya. Kita akan mengeksplorasi langkah-langkah yang dapat membantu dalam hubungan antar variabel. Fokus kita adalah pada kovarians dan koefisien korelasi linier [6].

2.9.1. Kovarians (*Covariance*)

Kovarians adalah ukuran variabilitas bersama dari dua variabel. Dalam hubungan ini, ada beberapa pengertian dasar yang perlu dipahami terlebih dahulu, yaitu:

1. Kovarians positif berarti bahwa kedua variabel bergerak bersama atau searah.
2. Kovarians 0 berarti bahwa kedua variabel independen.
3. Kovarians negatif berarti bahwa kedua variabel bergerak berlawanan arah.

Berikut adalah rumus kovarians untuk sampel:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})*(y_i-\bar{y})}{n-1}$$

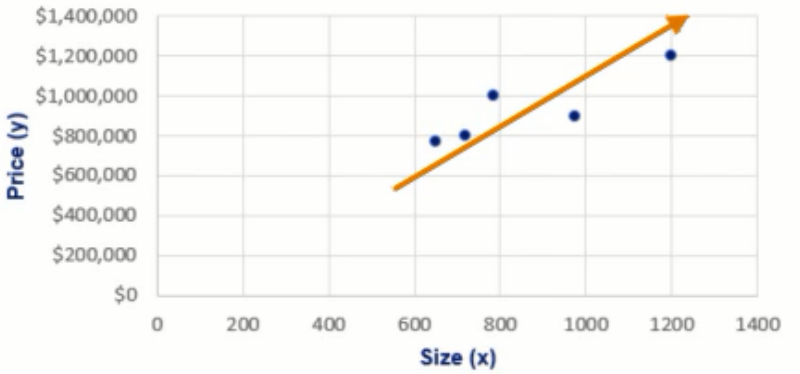
Dan berikut rumus kovarians untuk populasi:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)*(y_i-\mu_y)}{N}$$

Untuk dapat memahami kovarians, kita dapat menggunakan contoh real estate, yang merupakan salah satu faktor utama yang menentukan harga rumah. Ukurannya jelas, biasanya rumah yang lebih besar lebih mahal. Tabel data harga real estate yang dapat kita lihat di bawah menunjukkan kepada kita data tentang beberapa rumah di sisi kiri adalah ukuran setiap rumah dan di sebelah kanan adalah harga yang tercantum di koran lokal.

	x	y
	Size (ft.)	Price ($)
	650	772,000
	785	998,000
	1200	1,200,000
	720	800,000
	975	895,000
Mean	866	933,000

Data di atas dapat menyajikan titik data ini dalam scatterplot. Sumbu X akan menampilkan ukuran rumah dan sumbu Y akan memberikan informasi tentang harganya. Kita pasti bisa melihat sebuah pola bahwa ada hubungan yang jelas antara variabel-variabel ini. Kita mengatakan bahwa kedua variabel berkorelasi dan statistik utama untuk mengukur korelasi ini disebut kovarians.



Gambar 2.15. Scaterplot hubungan positif ukuran rumah dan harga

Kita dapat juga mengacu pada rumus kovarians yang memungkinkan kita menghitung kovarians antara dua variabel. Karena ini jelas data sampel, kita harus menggunakan rumus kovarians sampel. Mari kita terapkan dalam praktik untuk contoh yang kita lihat sebelumnya. Sumbu x merupakan nilai ukuran (*size*) dan y menunjukkan dari harga (*price*) rumah. Kita perlu menghitung ukuran rata-rata dan harga rata-rata, dan menghitung standar deviasi sampel. Singkatnya kita akan mendapat nilai kovarian sampel sebesar 33,49. Nilai ini lebih besar dari 0, atau memiliki nilai positif. Hal ini menunjukkan kedua data bergerak bersama atau searah.

2.9.2. Koefisien korelasi (*Correlation Coefficient*)

Korelasi adalah ukuran variabilitas dari dua variabel. Tidak seperti kovarians, korelasi dapat dianggap sebagai ukuran standar [4]. Dibutuhkan nilai antara -1 dan 1, sehingga mudah bagi kita untuk menafsirkan hasilnya.

1. Korelasi 1, yang dikenal sebagai korelasi positif sempurna, berarti bahwa satu variabel dijelaskan dengan sempurna oleh yang lain.
2. Korelasi 0 berarti bahwa variabel independen.

3. Korelasi dari -1, yang dikenal sebagai korelasi negatif sempurna, berarti bahwa satu variabel menjelaskan yang lain dengan sempurna, tetapi mereka bergerak ke arah yang berlawanan.

Rumus koefisien korelasi pada sampel adalah sebagai berikut:

$$r = \frac{s_{xy}}{s_x s_y}$$

Rumus koefisien korelasi pada populasi adalah sebagai berikut:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Korelasi menyesuaikan kovarians sehingga hubungan antara dua variabel menjadi mudah untuk menafsirkan rumus untuk koefisien korelasi adalah kovarians dibagi dengan produk dari standar deviasi dari dua variabel.

Kita sudah memiliki standar deviasi dari dua dataset. Sekarang kita akan menggunakan rumus untuk menemukan koefisien korelasi sampel. Secara matematis tidak ada cara untuk mendapatkan nilai korelasi lebih besar dari satu atau kurang dari minus satu. Dengan kata lain, nilai koefisien korelasi berada pada rentang nilai -1 sampai 1.

Dengan menggunakan data ukuran rumah dan harga rumah, kemudian dihitung dengan rumus koefisien korelasi sampel, kita mendapatkan nilai koefisien korelasi sampel 0,87. Jadi ada hubungan yang kuat antara kedua nilai tersebut. Korelasi satu juga dikenal sebagai korelasi positif sempurna yang berarti bahwa seluruh variabilitas satu variabel dijelaskan oleh variabel lainnya. Namun, secara logis kita tahu bahwa ukuran menentukan harga,

yaitu semakin besar rumah yang kita bangun, semakin mahal harganya. Hubungan ini hanya berjalan setelah rumah dibangun.

Kita mengharapkan korelasi nol antara harga kopi di Brasil dan harga rumah di London. Kedua variabel tidak memiliki kesamaan. Akhirnya kita dapat memiliki koefisien korelasi negatif. Ini bisa menjadi korelasi negatif sempurna minus satu atau lebih mungkin korelasi negatif yang tidak sempurna dari nilai antara minus 1 dan nol.

Pikirkan contoh berikut. Sebuah perusahaan yang memproduksi es krim dan perusahaan yang menjual payung. Es krim cenderung dijual lebih banyak saat cuaca sangat bagus dan orang membeli payung saat hujan. Jelas ada korelasi negatif antara keduanya dan karenanya ketika salah satu perusahaan menghasilkan lebih banyak uang, yang lain tidak.

Selanjutnya, sebelum kita melanjutkan, kita harus mencatat bahwa korelasi antara dua variabel: x dan y sama dengan korelasi antara y dan x. Rumusnya benar-benar simetris sehubungan dengan kedua variabel. Oleh karena itu korelasi harga dan ukuran sama dengan ukuran dan harga. Ini membawa kita ke kausalitas. Sangat penting bagi setiap analis atau peneliti untuk memahami arah hubungan sebab akibat dalam bisnis perumahan.

Appendix Bab 2: Statistik Deskriptif

Kita sudah melihat pengolahan data sederhana pada meteri deskriptif analisis. Pada bagian ini, kita akan mempelajari bagaimana mengolah data yang ditampilan pada materi Statistik Deskriptif dapat dikerjakan dengan Ms Excel. Dengan demikian pada bagian akan fokus pada fungsi Ms Excel yang dapat digunakan untuk mengolah data secara statistik menjadi lebih mudah. Tentu saja, Dalam Excel kita dapat mengtikan sendiri rumus seperti yang juga ditampilkan pada materi ini. Namun menggunakan fungsi yang ada pada MS Excel akan membuat pekerjaan kita lebih mudah dan cepat, serta menghidari kesalahan saat mengetikan rumus.

A.2.1. Excel untuk Mengukur Data Observasi

A. Data Kategori

Berikut adalah cara menampilkan data kategori

1. Menampilan kategorial data atau data kategorikal dalam tabel

   Pada materi kita menggunakan contoh data tiga produk mobil, yaitu Audi, BMW, dan Mercedes.

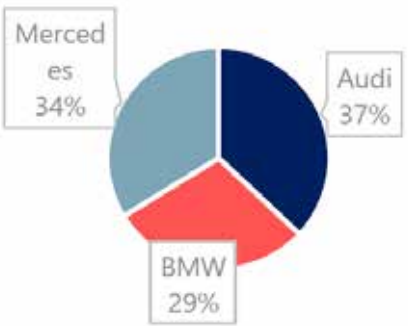
	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

   Di Excel, kode frekuensi atau hitung dengan memilih fungsi hitung (*count function*), maka akan muncul nanti: Total formula: =SUM()

2. Menampilan kategorial data atau data kategorikal dalam diagram batang

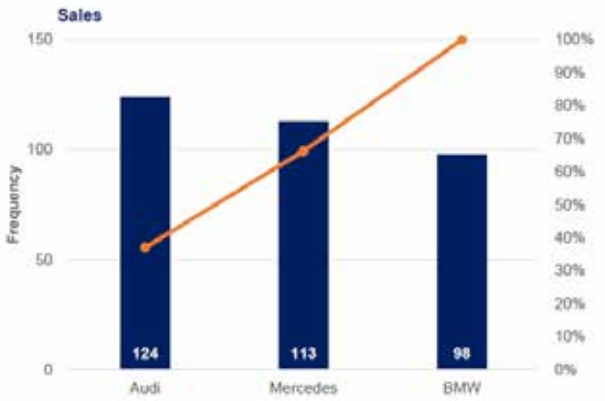
Bagan batang juga disebut c*lustered column charts* di Excel. Pilih semua data Kita, kemudian pilih **menu Insert -> Charts -> Clustered column or Bar chart**.

3. Menampilan kategorial data atau data kategorikal: dalam diagram pie atau lingkaran



Bagan pei dibuat dengan cara berikut: Pilih semua data, kemudian pilih **menu Insert ->Charts -> Pie chart**

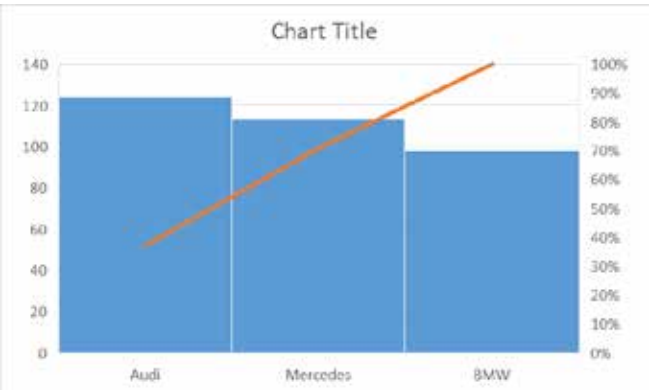
4. Menampilan kategorial data atau data kategorikal dalam diagram pareto



Pada Ms Excel terbaru, terdapat fungsi yang dapat menampilkan diagram pareto secara langsung. Kita masih menggunakan data penjualan mobil, langkahnya adalah sebagai berikut:

- Pilih area data nama mobil dan frequensi.
- Pilih menu insert > chart,

- Pada diagram histogram terdapat pilihan, diantaranya diagram pareto. Kita dapat memilihnya untuk menampilkan secara otomatis.



- Tampilan tersebut masih perlu disesuaikan. Double klik diagram yang sudah muncul, sehingga muncul format data series pada kiri layar Excel, seperti berikut ini

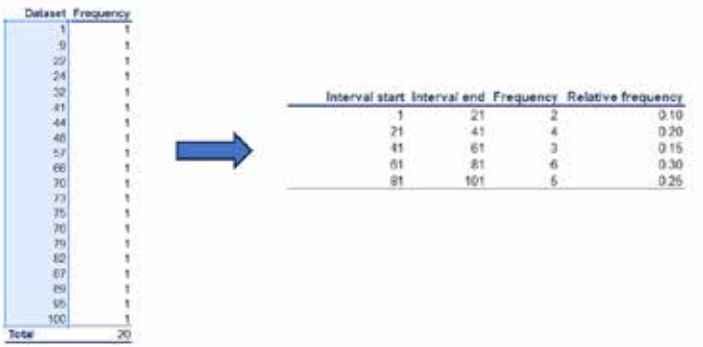
- Pastikan pilih series, dan pilih ikon bar, kemudian pilih gap width di 90%. Kita dapat mengatur gap width, semakin besar maka jarak antar bar semakin besar juga.

B. Data Numerik

Berikut ini adalah cara menampilkan data numerik

1. Menampilan data numerik dalam tabel.

   Pada Latihan ini, kita akan menampilkan data menjadi tabel untuk menentukan frakuensi dan frekuensi relatif seperti pada gambar di bawah ini.



Dataset	Frequency
1	1
9	1
22	1
24	1
32	1
41	1
44	1
46	1
57	1
66	1
70	1
73	1
75	1
76	1
79	1
82	1
87	1
89	1
95	1
100	1
Total	20

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

Membuat tabel distribusi frekuensi di Excel

- Buatlah tabel yang memuat kolom: *interval start, interval end, frequency*, dan *relative frequency* seperti di atas.
- Tentukan jumlah interval yang ingin kita gunakan. Pada umumnya interval yang bias digunakan adalah 5 sampai 20. Karena data yang ada adalah 20 maka kita pilih 5 sebagai interval. Untuk menentukan lebar interval berikut adalah caranya:

  → Temukan lebar interval menggunakan rumus di atas = Nilai terbesar dikurangi nilai terkecil dibagi dengan jumlah interval yang dikehendaki, yaitu (100 – 1) / 5 maka hasilnya adalah 19.8, yang dapat dibulatkan menjadi 20. Jadi kita buat rentang interval 20 dari nilai data.

  → Mulai interval pertama pada nilai terendah dalam himpunan data.

  → Selesaikan interval pertama pada nilai terendah ditambah lebar interval (start_interval_cell + interval_width_cell) yaitu 1 + 20, sehingga interval pertama adalah 1 sampai 21, dan seterusnya.

  → Lanjutkan dengan cara ini sampai kita telah membuat jumlah interval yang diinginkan

- Hitung **frekuensi** absolut menggunakan yang berikut ini COUNTIF formula **=COUNTIF(dataset_range,">="&interval start) - COUNTIF(dataset_range,">"&interval end)**

  Atau dengan **COUNTIFS formula**

  =COUNTIFS(dataset_range;">="&interval start; dataset_range;"<="&interval end)

  Keterangan:
  - *dataset_range* diisi dengan blok data yang akan diolah.
  - *Interval start* dan *interval end* adalah nilai sel pada kolom *interval star* dan *interval end*

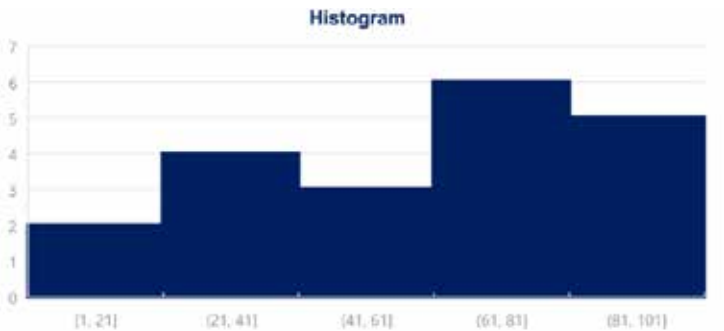
- Untuk menghitung **frekuensi relatif**, gunakan rumus berikut *absolute_frequency_cell number_of_observations* untuk menghitung frekuensi kumulatif

  → Frekuensi kumulatif pertama sama dengan frekuensi relatif

  → Setiap frekuensi kumulatif berurutan frekuensi kumulatif sebelumnya frekuensi relatif masing-masing. Rumus excel adalah:

  =cell_nilai_frekuensi/COUNT(dataset_range)

2. Menampilan data numerik dalam histogram

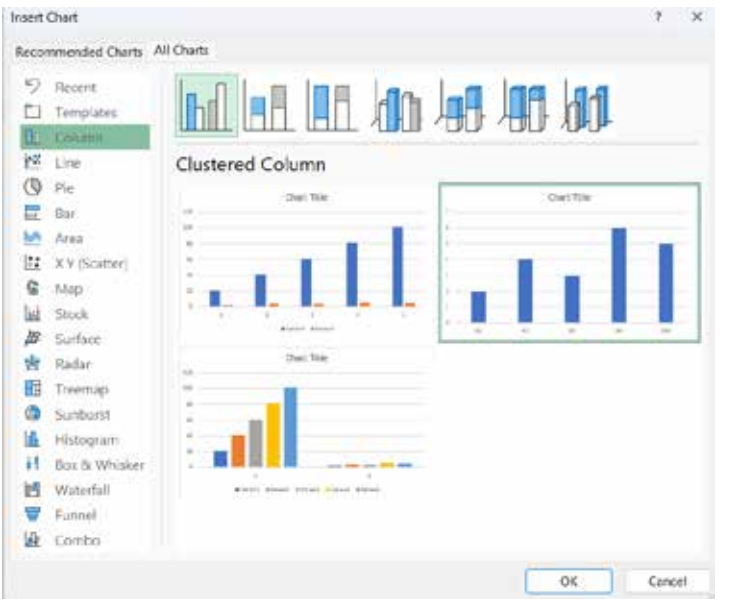


Kita masih menggunakan data frequensi dalam histogram ini. Membuat histogram di Excel:

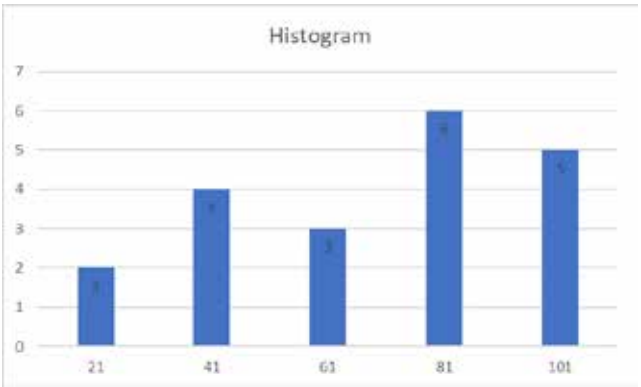
- Pilih data dengan blok range dataset.
- lalu pilih **menu Insert > Charts > Histogram.** Histogram dapat digunakan untuk nilai distribusi atau frekuensi relatif. Untuk mengubah jumlah biner (interval)
- Pilih sumbu x (Hasil mungkin akan berbeda dengan dengan gambar karena Excel secara otomatis menghitung ulang).

Cara lain adalah:

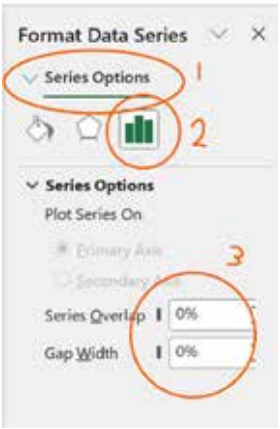
- Pilih data dari tabel, yaitu blok data interval end dan frekuensi
- lalu pilih **menu Insert > Charts > Coloumn.** Pilih yang sesuai dengna data, seperti gambar berikut:



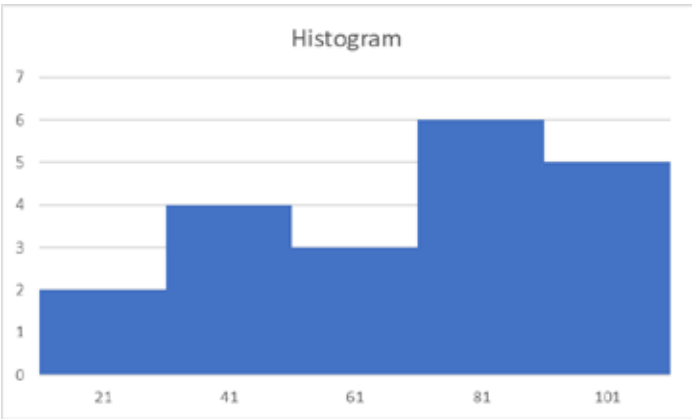
- Setelah gambar muncul, maka hasilnya seperti diagram bar.



- Untuk merubah menjadi histogram, maka bisa klik pada diagram kemudian pada format bisa dipilih seperti pada gambar berikut:



Maka hasilnya

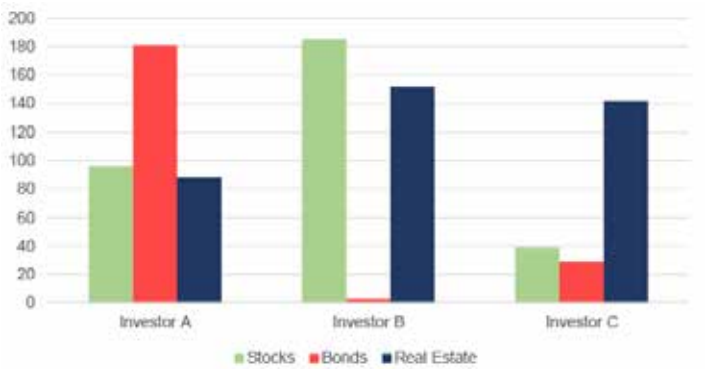


Walaupun mirip dengan diagram batang, namun histogram memiliki informasi yang berbeda.

3. Menampilan data numerik: dalam *side by side chart* dan *scatterplot*.

Untuk visualisasi ini, kita akan menggunakan data berikut ini:

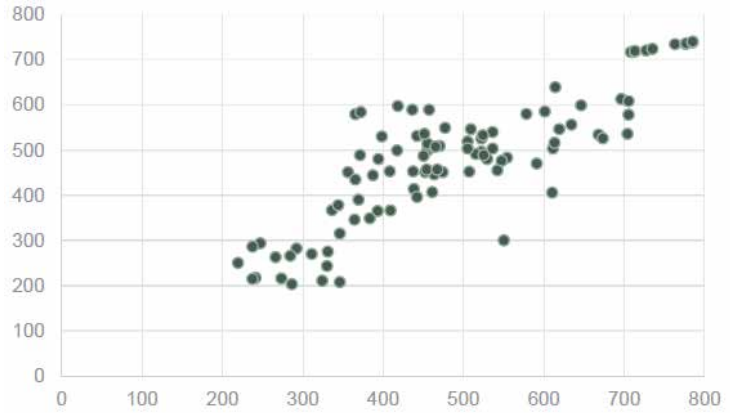
Tipe investasi	Investor A	Investor B	Investor C	Total
Saham/*stock*	96	185	39	320
Obligasi/*bond*	181	3	29	213
Real estate	88	152	142	382
Total	365	340	210	915



Dengan contoh data di atas, membuat bagan berdampingan di Excel dapat dilakukan dengan: Pilih data kita kemudian pilih menu **Insert > Charts > Clustered Column**

Memilih lebih dari satu grup seri data akan secara otomatis meminta Excel untuk membuat bagan *side by side bar* (colomn)

4. Menampilan data numerik: dalam *scaterplot*



Data dari dua variabel juga dapat digambarkan dengan scaterplot. Dengan data nilai membaca dan menulis siswa, maka kita dapat membuat scatter plot dalam Excel, caranya adalah pilih **menu Insert > Charts > Scatte**r

A.2.2. Excel untuk mengukur karakteristik sampel atau populasi

Pada materi ini poin pokok yang dapat digunakan statistik untuk mengukur karakteristik sampel adalah mean, median, dan modus. Fungsi excel untuk mean, median, modus adalah sebagai berikut:

1. Mean. Di Excel, rata-rata dihitung dengan rumus **=AVERAGE(dataset_range)**
2. Median. Di Excel, median dihitung dengan rumus **=MEDIAN(dataset_range)**
3. Modus. Di Excel, mode dihitung dengan rumus **=MODE.SNGL(dataset_range)** untuk mengembalikan satu mode, dan rumus **=MODE.MULT(dataset_range)** untuk mengembalikan array dengan mode ini digunakan ketika kita memiliki lebih dari 1 mode

4. Skewness. Menghitung kecondongan di Excel dapat dengan rumus **=SKEW(dataset_range)**

Keterangan: dataset_range merupakan rentang data dalam kolom excel yang akan diolah atau dianalisa

A.2.3. Excel untuk mengukur keragaman data

Untuk mengukur keragaman data, materi ini menggunakan varians dan standar deviasi. Keduanya dapat dilakukan pada sampel atau populasi. Rumusnya adalah sebagai berikut

- Varians sampel: **=VAR.S(dataset_range)**
- Varians populasi: **=VAR.P(dataset_range)**
- Standar deviasi (Simpangan baku) sampel: **STDEV.S(dataset_range)**
- Standar deviasi (Simpangan baku) populasi: **=STDEV.P(dataset_range)**

Keterangan: dataset_range merupakan rentang data dalam kolom excel yang akan diolah atau dianalisa

A.2.4. Excel untuk mengukur hubungan dua variabel

Pada materi ini kita akan menggunakan kovarian dan korelasi (covariance dan correlation). Pada analisa kovarian, olah data dapat dialukan pada tingkat sampel atau populasi. Sedangkan pada korelasi, perhitungan akan sama antara sampel dan populasi. Rumusnya adalah

- Kovarians sampel**: =COVARIANCE.S(dataset_range)**
- Kovarians populasi**: =COVARIANCE.P(dataset_range)**
- Di Excel, korelasi dihitung dengan: **=CORREL(dataset_range)**

Keterangan: dataset_range merupakan rentang data dalam kolom excel yang akan diolah atau dianalisa

Latihan

Untuk mempermudah latihan, file excel untuk latihan dapat di akses pada link berikut: https://bit.ly/Data_Latihan_Excel . Pilih pada folder yang sesuai dan file yang sesuai. Latihan diambil dari 365 Business Intelegence

Latihan 2.1. Variabel kategori

Ada toko es krim, yang beroperasi di Canbera, Sedney, dan Hobart, dengan masing-masing penjualan adalah 12.300 es krim di Canbera; 17.110 di Sedney dan 19.900 di Hobart.

Tugas:

1. Urutkan data dalam tabel distribusi frekuensi.
2. Buat kolom baru di tabel Anda, yang mewakili frekuensi relatif input. Anda dapat memilih untuk mengungkapkannya dalam persentase atau sebagai desimal.
3. Buat diagram lingkaran, yang mewakili pangsa setiap kota terhadap penjualan perusahaan Anda.
4. Urutkan tabel berdasarkan frekuensi.
5. Buat bagan batang (kolom) yang mewakili data yang diurutkan.
6. Di kolom baru, hitung frekuensi kumulatif data.
7. Pada sumbu kedua dalam bagan yang sama, representasikan frekuensi kumulatif data.

Latihan 2.2. Variabel numerik

Dataset:

8	30	30	54	86	94	102	110	169
170	176	236	240	241	242	255	262	276
279	282							

Tugas

1. Dari dataset, bagilah angka-angka menjadi 6 interval dengan lebar yang sama, hitung lebar interval itu (dalam pengerjaan, data set dapat disusun horizontal atau ke bawah untuk mempermudah). Pembulatan ke bilangan bulat terdekat, lebih besar dari hasil yang diperoleh.
2. Membuat tabel distribusi frekuensi yang menunjukkan :
   a. Interval
   b. Frekuensi absolut setiap interval
   c. Frekuensi relatif dari setiap interval
3. Ulangi tugas 1 dan 2, tetapi kali ini, gunakan lebar interval yang tepat. Jangan membulatkan ke bilangan bulat terdekat.

Latihan 2.3. Histogram

Dataset

13	68	165	193	216	228	361	470	500
529	544	602	647	692	696	699	809	892
899	936							

Tugas:

1. Dari dataset di atas, buat tabel distribusi frekuensi (dalam pengerjaan, data set dapat disusun horizontal atau ke bawah untuk mempermudah).
2. Catatan: Buka lembar berikutnya jika Anda ingin melewati bagian ini.
3. Buat histogram dengan 10 interval, berdasarkan himpunan data Anda.

Latihan 2.4. Cross table

Berikut ini adalah data ketenagakerjaan tentang negara X. Buatlah tabel silang (*cross table*) yang menunjukkan data berikut

60% dari usia 18 hingga 25 tahun, statusnya bekerja

85% dari usia 25 hingga 35 tahun, statusnya bekerja

5% dari usia 35 hingga 45 tahun, statusnya menganggur

3% dari usia 45 hingga 55 tahun, statusnya menganggur

3% dari usia 55 hingga 65 tahun, statusnya menganggur

Semua 65+, statusnya bekerja.

Catatan: definisi pengangguran adalah: tanpa pekerjaan, tetapi aktif mencarinya.

Tugas

1. Buat tabel silang yang merangkum data yang telah diberikan tersebut di atas.
2. Buat bagan batang berdampingan (disebut bagan kolom berkluster di Excel), untuk menyempurnakan ringkasan Anda secara visual.

Latihan 2.5. Scaterplot

Anda diberikan harga penutupan saham Apple, Alphabet Inc. (Google) dan Bank of America pada tanggal yang berbeda untuk jangka waktu 2 bulan.

Date	Apple (AAPL)	Alphabet (GOOGL)	Bank of America (BAC)
1/3/2017	116.15	808.01	22.53
1/4/2017	116.02	807.77	22.95
1/5/2017	116.61	813.02	22.68

1/6/2017	117.91	825.21	22.68
1/9/2017	118.99	827.18	22.55
1/10/2017	119.11	826.01	22.94
1/11/2017	119.75	829.86	23.07
1/12/2017	119.25	829.53	22.92
1/13/2017	119.04	830.94	23.01
1/17/2017	120	827.46	22.05
1/18/2017	119.99	829.02	22.63
1/19/2017	119.78	824.37	22.53
1/20/2017	120	828.17	22.64
1/23/2017	120.08	844.43	22.56
1/24/2017	119.97	849.53	22.95
1/25/2017	121.88	858.45	23.37
1/26/2017	121.94	856.98	23.44
1/27/2017	121.95	845.03	23.36
1/30/2017	121.63	823.83	22.95
1/31/2017	121.35	820.19	22.64
2/1/2017	128.75	815.24	22.89
2/2/2017	128.53	818.26	22.72
2/3/2017	129.08	820.13	23.29
2/6/2017	130.29	821.62	23.12
2/7/2017	131.53	829.23	22.9

Tugas:

1. Buat plot sebaran dengan harga saham Apple dan Alphabet.
2. Membuat plot sebar dengan harga saham Apple dan Bank of America
3. Jelajahi apakah ada hubungan antara kedua variabel.

(soal diambl dari Latihan 365 Business Intelligence)

Latihan 2.6. Mean, median, modus

Berikut sampel 11 orang dan pendapatan tahunan pribadi mereka.

$62,000.00	$64,000.00	$49,000.00	$324,000.00
$1,264,000.00	$54,330.00	$64,000.00	$51,000.00
$55,000.00	$48,000.00	$53,000.00	

Tugas:

1. Hitung rata-rata, median, dan modus
2. Cobalah untuk menafsirkan angka-angka yang Anda dapatkan

Latihan 2.7. Skewness

Dataset

212	869	220	654	511	624	420	121	428
865	799	405	230	670	870	366	99	55
489	312	493	163	221	84	144	48	375
86	168	100						

Tugas:

Identifikasi kemiringan kumpulan data di atas. Anda dapat menggunakan rumus dari pelajaran, rumus kemiringan di excel (=SKEW) atau dapat memplotnya pada grafik.

Latihan 2.8. Varians

Dataset pendapatan tahunan, 11 orang di AS:

$62,000.00	$64,000.00	$49,000.00	$324,000.00
$1,264,000.00	$54,330.00	$64,000.00	$51,000.00
$55,000.00	$48,000.00	$53,000.00	

Anda memiliki penghasilan pribadi tahunan 11 orang.

Tugas

1. Putuskan apakah anda harus menggunakan rumus sampel atau populasi untuk varians
2. Hitung varians penghasilan mereka
3. Secara umum, apa yang dikatakan angka ini kepada Anda?

Latihan 2.9. deviasi standar dan coefisien

Pendapatan tahunan Denmark

462,852.37 kr	500,179.17 kr
470,317.73 kr	589,763.50 kr
567,367.42 kr	515,109.89 kr
492,713.81 kr	425,525.56 kr
507,644.53 kr	522,575.25 kr
500,179.17 kr	

Penghasilan tahunan AS

$62,000.00	$64,000.00
$64,000.00	$51,000.00
$49,000.00	$55,000.00
$324,000.00	$48,000.00
$1,264,000.00	$53,000.00
$54,330.00	

Anda memiliki data penghasilan pribadi tahunan 11 orang dari AS dan 11 orang dari Denmark. Anda memiliki pendapatan rata-rata untuk AS dari latihan sebelumnya (latihan 2.8.)

Tugas:

1. Putuskan apakah Anda harus menggunakan rumus sampel atau populasi untuk standar deviasi dan koefisien variasi.
2. Hitung standar deviasi pendapatan di AS dan di Denmark. (Petunjuk: Anda dapat mulai dengan menghitung rata-rata dan varians)
3. Hitung koefisien variasi pendapatan di AS dan di Denmark. (Cobalah untuk menafsirkan angka yang Anda dapatkan.)

Latihan 2.9. Covarians

Dataset:

Menulis	Membaca
344	378
383	349
611	503
713	719
536	503

Data di atas adalah tentang skor membaca dan menulis dari beberapa siswa

Tugas:

1. Tentukan apakah ini sampel atau populasi
2. Hitung kovarian dari dua himpunan data
3. Plot data pada plot sebar dan menggunakan komentar pengetahuan Anda sebelumnya tentang apakah ada hubungan yang nyata antara kedua variabel.
4. Hitung koefisien korelasi dari dua kumpulan data.
5. Komentari kekuatan korelasi antara kedua kumpulan data

BAB 3

DASAR STATISTIK INFERENSIAL

Bab ketiga dari buku ini, yang berjudul "Dasar Statistik Inferensial," dirancang untuk memperkenalkan konsep-konsep dasar yang menjadi inti dari analisis statistik yang lebih mendalam. Statistik inferensial memungkinkan kita untuk membuat kesimpulan tentang populasi berdasarkan data sampel, serta menguji hipotesis dan membuat prediksi.

Dalam bab ini, kita akan membahas berbagai konsep penting seperti distribusi sampling, estimasi parameter, dan uji hipotesis. Pada bagian ini kuta akan belajar teknik-teknik dasar yang digunakan untuk membuat inferensi statistik, termasuk interval kepercayaan dan uji signifikan. Kita juga akan menjelaskan bagaimana menggunakan Microsoft Excel untuk melakukan analisis statistik inferensial ini.

Kita telah membahas dasar-dasar statistik deskriptif, sekarang saatnya untuk beralih ke statistik inferensial, statistik inferensial mengacu pada metode yang mengkitalkan teori probabilitas dan distribusi, khususnya untuk memprediksi nilai populasi berdasarkan data sampel. Materi Ini akan membawa kita ke perkiraan titik dan kita akan menyimpulkan bagian dengan interval kepercayaan. Kita akan memahami pengembangan tentang proses statistik. Pemahaman ini akan sangat membantu kita, jika kita akan memutuskan untuk masuk ke ilmu data (*data science*). Pemahaman ini akan membantu kita dalam memahami dasar-dasar penelitian kuantitatif dan pengambilan keputusan berbasis data.

3.1. Pengertian Distribusi

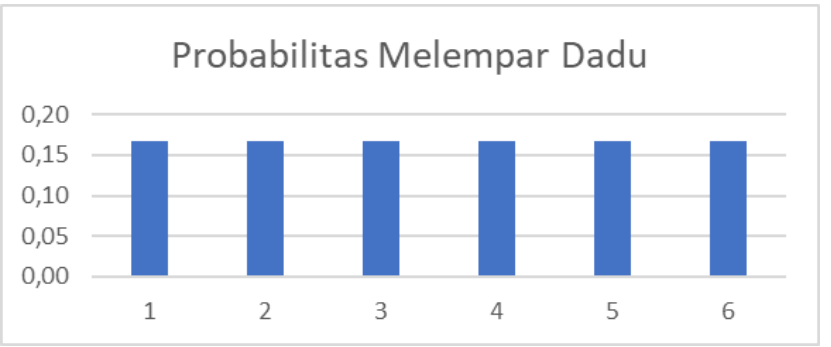
Sebelum kita dapat berbicara tentang pengujian dalam statistik, kita harus mempelajari apa itu distribusi. Dalam statistik, ketika kita menggunakan istilah distribusi, kita biasanya merujuk pada distribusi probabilitas, contohnya distribusi normal, distribusi T, distribusi binomial dan distribusi seragam. **Distribusi adalah fungsi yang menunjukkan nilai yang mungkin untuk suatu variabel dan seberapa sering nilai tersebut muncul** [4].

Contohnya adalah dadu yang memiliki 6 sisi bernomor dari 1 hingga 6. Jika kita melempar dadu, berapa probabilitas mendapatkan angka 1, yaitu 1 dari 6 atau 1/6 (0.17). Berapa probabilitas mendapatkan 2, yaitu 1/6 (0,17). Hal yang sama berlaku untuk 3, 4, 5 dan 6. Kita memiliki kesempatan yang sama untuk mendapatkan masing-masing dari 6 hasil. Sekarang berapa probabilitas mendapatkan 7. Tidak mungkin mendapatkan 7 saat melempar satu dadu. Oleh karena itu, probabilitasnya adalah nol. Jadi kita dapat generalisasi bahwa distribusi suatu peristiwa tidak hanya terdiri dari nilai input yang dapat diamati, tetapi terdiri dari semua nilai yang mungkin. Jadi distribusi melempar dadu ditunjukkan oleh tabel 3.1. berikut.

Tabel 3.1. Probabilitas dadu

Hasil	Probabilitas terjadinya
1	0.17
2	0.17
3	0.17
4	0.17
5	0.17
6	0.17
yang lain	0.00
	1.00

Kondisi pada contoh tersebut di atas dapat disebut distribusi seragam diskrit, yaitu semua hasil memiliki peluang yang sama untuk terjadi. Setiap distribusi probabilitas memiliki representasi visual. Berikut ini adalah grafik yang menggambarkan kemungkinan terjadinya setiap peristiwa. Berikut grafik untuk contohnya.



Gambar 3.1 Grafik probabilitas melempar dadu

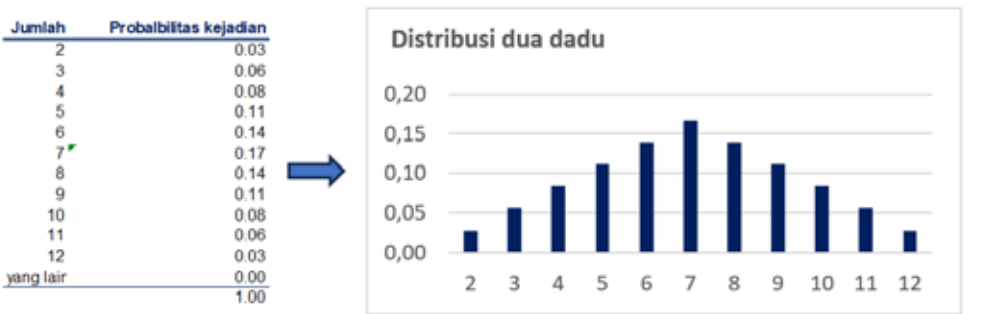
Dalam hal ini menjadi sangat penting untuk memahami bahwa grafik hanyalah representasi visual dari distribusi, sehingga seringkali ketika kita berbicara tentang distribusi, kita menggunakan grafik. Itu sebabnya banyak orang percaya bahwa distribusi adalah grafik itu sendiri. Namun, hal ini tidak benar. Distribusi didasari oleh probabilitas yang mendasarinya dan bukan grafik. Grafik hanyalah representasi visual.

Setelah klarifikasi singkat ini, mari kita jelajahi contoh yang berbeda. Pikirkan tentang melempar dua dadu. Apa kemungkinan hasil 1 banding 1, kemudian 2 dan 1, kemudian 1 dan 2 dan seterusnya. Berikut adalah tabel dengan semua kemungkinan kombinasi.

(1,1)	**(2,1)**	**(3,1)**	**(4,1)**	**(5,1)**	**(6,1)**
(1,2)	**(2,2)**	**(3,2)**	**(4,2)**	**(5,2)**	**(6,2)**
(1,3)	**(2,3)**	**(3,3)**	**(4,3)**	**(5,3)**	**(6,3)**
(1,4)	**(2,4)**	**(3,4)**	**(4,4)**	**(5,4)**	**(6,4)**
(1,5)	**(2,5)**	**(3,5)**	**(4,5)**	**(5,5)**	**(6,5)**
(1,6)	**(2,6)**	**(3,6)**	**(4,6)**	**(5,6)**	**(6,6)**

Dari kombinasi tersebut dapat dihitung, terdapat 36 kombinasi. Katakanlah kita sedang memainkan permainan di mana kita mencoba menebak jumlah dari dua dadu. Berapa probabilitas

mendapatkan jumlah 1:0 (satu dan nol). Hal ini tidak mungkin karena jumlah minimum yang bisa kita dapatkan minimal adalah 2. Jadi berapa probabilitas mendapatkan jumlah 2. Hanya ada satu kombinasi yang akan memberi kita jumlah 2 ketika kedua dadu sama memberikan angka 1. Jadi, nilai jumlah 2 memiliki total kemungkinan 0,03, yaitu 1 kemungkinan dari 36 kemungkinan yang ada. Demikian pula, probabilitas mendapatkan jumlah 3 diberikan oleh jumlah kombinasi yang memberikan jumlah 3 kedua dadu, yaitu ada 2 kemungkinan dari 36 kemungkinan. Oleh karena itu angka 2 dibagi dengan 36 hasilnya 0,06, kita dapat melanjutkan dengan cara ini sampai kita memiliki distribusi probabilitas penuh. Mari kita lihat grafik yang terkait dengannya, maka kita dapat dengan mudah memahami bahwa ketika melempar dua dadu, kemungkinan mendapatkan jumlah 7 adalah kombinasi dan kemungkinan yang tertinggi. Selain itu, kita juga dapat membandingkan hasil yang berbeda seperti probabilitas mendapatkan 10 dan probabilitas mendapatkan 5.



Gambar 3.2. Contoh distribusi probabilitas

Jelas bahwa kecil kemungkinan kita akan mendapatkan 10 dibanding 7. Contoh tersebut adalah variabel diskrit. Selanjutnya, kita akan fokus pada distribusi berkelanjutan karena mereka lebih umum dalam kesimpulan dalam beberapa pelajaran berikutnya. Kita akan memeriksa beberapa jenis utama distribusi kontinu yang dimulai dengan distribusi normal.

3.2. Distribusi Normal (*Normal Distribution*)

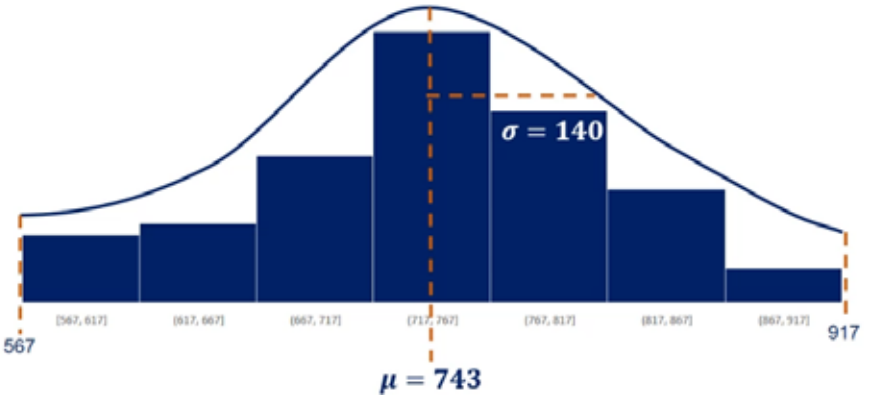
Saat ini, kita belajar bahwa distribusi kumpulan data menunjukkan kepada kita frekuensi di mana nilai-nilai yang mungkin terjadi dalam suatu interval. Dalam materi ini, kita lebih fokus pada distribusi normal dan distribusi T karena alasan berikut, yaitu: mereka mendekati berbagai variabel acak. Distribusi sampel berarti bahwa ukuran sampel yang cukup besar dapat didekati dengan normal. Semua statistik yang dapat dihitung, atau keputusan yang dibuat berdasarkan distribusi normal akan memiliki rekam jejak yang baik.

Berikut adalah representasi visual dari distribusi normal. Kita pasti terlihat dalam distribusi normal sebelumnya karena ini adalah yang paling umum. Istilah statistik untuk itu adalah distribusi Gaussian tetapi banyak orang menyebutnya kurva lonceng karena berbentuk seperti lonceng.

Gambar 3.3. A. Distribusi normal, B. Distribusi T (*Student T distribution*)

Grafik tersebut menunjukkan simetris yang berarti bahwa median dan mode rata-ratanya sama. Jika Kita ingat pelajaran tentang kecondongan, dari gambar 3.3. dapat dipahami bahwa grafik tersebut tidak memiliki kemiringan. Pada kondisi tersebut, data sangat berpusat di sekitar rata-rata. Dengan demikian gambar tersebut menunjukkan normal. Grafik tersebut dapat dipahami bahwa titik tertinggi terletak di rata-rata karena bertepatan dengan mode. Penyebaran grafik tersebut juga ditentukan oleh standar deviasi.

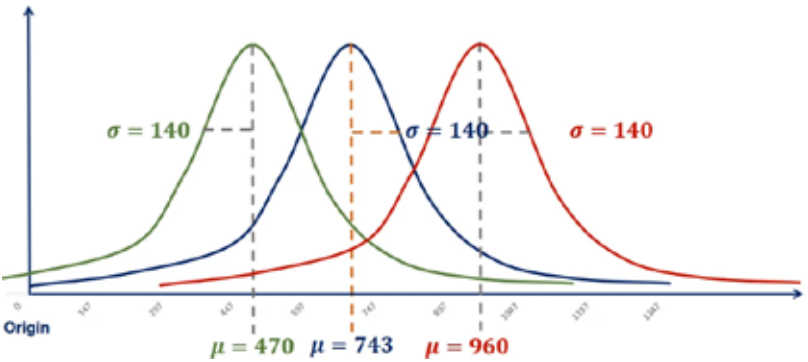
Kita coba memahami distribusi normal dengan contoh berikut. Kita lihat histogram yang kira-kira terdistribusi normal ini seperti pada gambar 3.4. Adanya konsentrasi data pengamatan di sekitar rata-rata sangat baik karena sama dengan atau mendekati modus.



Gambar 3.4 Contoh grafik distribusi normal

Gambar tersebut simetris di kedua sisi rata-rata. Contoh tersebut menggunakan 80 pengamatan untuk membuat histogram tersebut, maka, dengan mean 743, maka kita dapatkan standar deviasinya adalah 140.

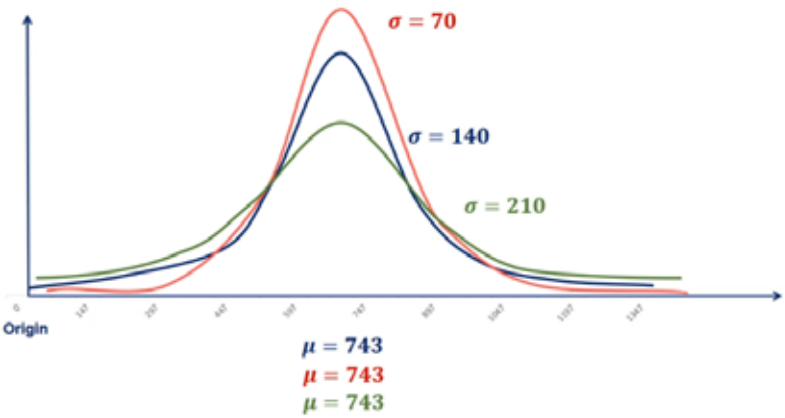
Grafik tersebut menunjukkan kondisi yang baik, namun bagaimana jika rata-rata lebih kecil atau lebih besar. Jika kita perkecil dengan menggunakan grafik yang sama, berasal dari titik nol, maka menambahkan grafik apa pun memberikan perspektif. Menjaga standar deviasi tetap atau mengendalikan standar deviasi dengan rata-rata yang lebih rendah akan menghasilkan bentuk distribusi yang sama.



Gambar 3.5 Distribusi normal dengan kontrol standar deviasi

Rata-rata yang lebih besar akan memindahkan grafik ke kanan. Dalam contoh kita, ini menghasilkan dua distribusi baru, satu dengan rata-rata 470 dan standar deviasi 140, dan satu dengan rata-rata 960 dan standar deviasi 140.

Kita dapat lakukan yang sebaliknya. Mengontrol rata-rata kita dapat mengubah standar deviasi dan melihat apa yang terjadi. Kali ini grafik tidak bergerak melainkan membentuk kembali standar deviasi yang lebih rendah menghasilkan dispersi yang lebih rendah. Jadi lebih banyak data di tengah dan ekor yang lebih tipis, di sisi lain, standar deviasi yang lebih tinggi akan menyebabkan grafik mendatar dengan lebih sedikit titik di tengah dan lebih ke akhir.



Gambar 3.5 Distribusi normal dengan kontrol rata-rata (*mean*)

3.3. Standarisasi Distribusi Normal

Setiap distribusi dapat distandarisasi, standardisasi adalah proses mengubah variabel ini menjadi satu dengan nol rata-rata dan standar deviasi satu, rumus sederhana berikut ini:

$$z = \frac{x - \mu}{\sigma}$$

Keterangannya sebagai berikut: Z adalah distribusi normal terstandar, μ (miu) adalah mean, σ (sigma) adalah standar deviasi, X adalah nilai data. Distribusi normal standar adalah kasus khusus dari distribusi normal, nilai ini memiliki rata-rata 0 dan standar deviasi 1. Secara logis, distribusi normal juga dapat distandarisasi. Hasilnya disebut distribusi normal standar. Mengapa kita standarisasi dilakukan? Standardisasi memungkinkan kita untuk:

1. Bandingkan kumpulan data terdistribusi normal yang berbeda;
2. mendeteksi normalitas;
3. Mendeteksi pencilan;
4. Buat interval kepercayaan;
5. Uji hipotesis;
6. Melakukan analisis regresi.

3.4. *Central Limit Theorem* (CLT)

Sebelum kita melanjutkan, mari kita perkenalkan konsep: distribusi sampel, contohnya jika kita memiliki populasi mobil bekas di toko mobil. Kita ingin menganalisis harga mobil dan dapat membuat beberapa prediksi tentang harganya. Parameter populasi yang mungkin menarik adalah rata-rata harga mobil, standar deviasi harga, kovarians dan sebagainya secara normal.

Dalam statistik kita tidak akan memiliki data tentang seluruh populasi melainkan hanya sampel, mari kita menarik sampel dari penjualan mobil. Misalkan, dari sampel pertama rata-rata harga mobil adalah 2.617,23 Dollars. Masalah muncul saat kita mengambil sampel lain, kita mungkin mendapatkan rata-rata yang sama sekali berbeda, misalnya adalah 3.201,34 Dollars. Kemudian yang ketiga adalah dengan rata-rata 2.844,33 Dollars. Seperti yang kita lihat, rata-rata sampel tergantung pada sampel itu sendiri. Jadi mengambil nilai tunggal seperti yang kita lakukan dalam statistik deskriptif jelas kurang optimal. Yang bisa kita lakukan adalah mengambil banyak sampel dan membuat dataset baru yang terdiri dari sampel rata-rata nilai yang didistribusikan dalam beberapa cara. Jadi kita memiliki distribusi baru ketika kita mengacu pada distribusi yang dibentuk oleh berbagai sampel. Kita dapat menggunakan istilah distribusi sampling untuk kasus ini, dan hasilnya kita bisa lebih tepat. Sekarang jika kita memeriksa nilai-nilai ini dengan cermat, kita akan menyadari bahwa mereka berbeda tetapi terkonsentrasi di sekitar nilai tertentu.

Semakin besar ukuran sampel, semakin rendah variansnya, atau dengan kata lain, semakin dekat perkiraan yang kita dapatkan. Jadi, jika kita dapat menarik sampel yang lebih besar, hasil statistik akan lebih akurat. Biasanya untuk menerapkan *Central Limit Theorem* (CLT), kita membutuhkan ukuran sampel setidaknya 30 pengamatan.

$$N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

Dengan penjelasan tersebut, dapat dipahami bahwa CLT adalah salah satu wawasan statistik terbesar. Konsep ini menyatakan bahwa tidak peduli distribusi yang mendasari suatu dataset, distribusi sampling dari nilai rata-rata akan mendekati distribusi

normal [3]. Selain itu, rata-rata distribusi sampling akan sama dengan rata-rata distribusi asli dan varians akan n kali lebih kecil, di mana n adalah ukuran sampel. CLT berlaku setiap kali kita memiliki jumlah atau rata-rata dari banyak variabel (misalnya jumlah angka yang digulung saat melempar dadu).

CLT memungkinkan kita untuk mengasumsikan normalitas untuk banyak variabel yang berbeda. Itu sangat berguna untuk interval kepercayaan, pengujian hipotesis, dan analisis regresi. Infact, distribusi Normal begitu dominan diamati di sekitar kita karena fakta bahwa setelah CLT, banyak variabel menyatu ke Normal. CLT memungkinkan kita untuk melakukan tes, memecahkan masalah dan membuat kesimpulan menggunakan distribusi normal bahkan ketika populasi tidak terdistribusi normal.

3.5. *Standard error*

Materi ini akan sangat singkat dan memiliki tujuan tunggal untuk mendefinisikan apa itu *standard error* (kesalahan standar). *Standard error* adalah standar deviasi distribusi yang dibentuk oleh sampel, atau dengan kata lain disebut juga standar deviasi distribusi sampling[5]. *Standard error* menunjukkan variabilitas, yang dalam hal ini adalah variabilitas sampel yang berbeda. Nilai tersebut juga menunjukkan ukuran seberapa jauh rata-rata dari sampel kita diperkirakan dari rata-rata populasi sebenarnya.

Bayangkan kita ingin mengetahui tinggi rata-rata siswa di sekolah kita. Jika kita hanya mengukur tinggi dari beberapa siswa saja, maka rata-rata tinggi dari sampel kita mungkin tidak sempurna mewakili rata-rata tinggi di seluruh sekolah. *Standard error* dalam hal ini akan memberi tahu kita sejauh mana rata-rata tinggi dari sampel kita bisa berbeda dari rata-rata tinggi sebenarnya di seluruh sekolah. Jika *standard error* rendah, itu seperti mengukur tinggi dari sejumlah siswa yang cukup banyak dan mendapatkan rata-rata yang sangat dekat dengan tinggi rata-rata di sekolah. Namun, jika

standard error tinggi, itu mungkin karena kita hanya mengukur tinggi dari sedikit siswa, dan rata-rata tinggi dari sampel kita bisa jauh dari rata-rata tinggi sebenarnya di seluruh sekolah.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Dengan memahami *standard error*, kita dapat menilai sejauh mana kita dapat mengkaitkan data kita untuk membuat kesimpulan tentang populasi secara keseluruhan. Beberapa makna dan manfaat dari *standard error* meliputi:

1. Ketidakpastian: *Standard error* membantu kita memahami sejauh mana kita dapat yakin bahwa rata-rata sampel kita mewakili rata-rata sebenarnya di populasi. Semakin rendah standard error, semakin dekat rata-rata sampel dengan rata-rata populasi, dan sebaliknya.
2. Akurasi Pengukuran: *Standard error* membantu kita mengukur akurasi estimasi rata-rata sampel. Dengan kata lain, seberapa baik sampel kita merepresentasikan keseluruhan populasi.

3.6. *Estimators* dan *Estimates*

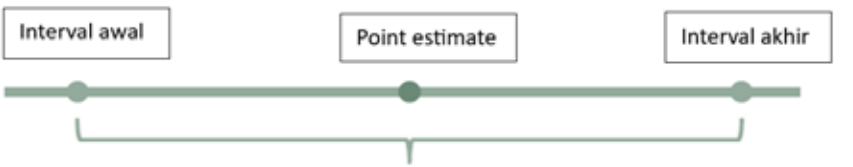
Kita lanjutkan dengan konsep *estimator* (penaksir) parameter populasi. Konsep ini adalah perkiraan yang hanya bergantung pada informasi sementara, nilai tertentu disebut perkiraan. Secara umum, *estimator* adalah fungsi matematika yang mendekati parameter populasi yang hanya bergantung pada informasi sampe[3]l. Contoh *estimator* (penaksir) dan parameter yang sesuai:

Istilah (*Term*)	*Estimator*	Parameter
Mean	$\overline{x}$	$\boldsymbol{\mu}$
Variance	S^2	σ^2
Correlation	r	$\boldsymbol{\rho}$

Estimator (penaksir) memiliki dua sifat penting:

1. **Bias**. Nilai yang diharapkan dari *estimator* yang tidak bias adalah parameter populasi. Bias dalam hal ini adalah 0. Jika nilai yang diharapkan dari estimator adalah nilai parameter + b, maka biasnya adalah b.
2. **Efisiensi**. Penaksir yang paling efisien adalah estimator dengan varians terkecil.

Estimates (perkiraan) adalah hasil yang kita dapatkan dari *estimator* (saat kita menerapkan rumus). Ada dua jenis *estimates* (perkiraan): perkiraan titik (*point estimates*) dan perkiraan interval kepercayaan (*confidence intervals*). Perkiraan titik adalah angka tunggal, contohnya nilai angka 1; 5; 122,27; 0,32. Sedangkan, interval kepercayaan secara alami adalah interval, contohnya nilai (1;5), (12;33), (-0.77;0.11). Kedua jenis perkiraan tersebut terkait erat. Perkiraan titik terletak tepat di tengah interval kepercayaan.



Namun demikian, interval kepercayaan memberikan lebih banyak informasi dan lebih disukai saat membuat kesimpulan. Pembahasan tentang interval kepercayaan akan dibahas pada materi berikutnya.

Kita tidak membahas topik ini secara detail karena topik ini bisa luas. Namun, setidaknya kita harus memiliki pemahaman tentang konsep *estimator* yang seperti hakim. Kita perlu selalu mencari estimator tidak bias dan yang paling efisien. Estimator yang tidak bias memiliki nilai yang diharapkan sama dengan parameter populasi.

Mari kita pikirkan *estimator* bias untuk menjelaskan hal itu. Misalkan seseorang memberi tahu kita tentang tinggi rata-rata orang Amerika dengan mengambil sampel, kemudian menemukan rata-ratanya, dan kemudian menambahkan 1 meter ke hasil itu. Jadi, rumusnya adalah $\overline{x}$ ditambah 1 meter. Nilai tersebut memberi kita *estimator* (penaksir) tetapi bias. Jauh lebih masuk akal bahwa tinggi rata-rata orang Amerika diperkirakan hanya dengan sampel rata-rata saja. Maka, kita dapat mengatakan bahwa bias penaksir ini adalah 1 meter.

Sekarang kita beralih ke efisiensi. Penaksir yang paling efisien adalah yang memiliki variabilitas hasil paling kecil. Dalam kasus ini, sulit untuk mencari contohnya. Namun cukup untuk mengetahui bahwa yang paling efisien berarti penaksir yang tidak bias dengan varians terkecil.

Latihan

Untuk mempermudah latihan, file excel untuk latihan dapat di akses pada link berikut: https://bit.ly/Data_Latihan_Excel . Pilih pada folder yang sesuai dan file yang sesuai. Latihan diambil dari 365 Business Intelegence

Dataset: 80 data numerik

567,45	705,78	748,2	799,7
572,45	705,86	748,28	800
572,45	708,11	748,53	810,9
589,11	711,03	750,03	811,5
613,86	714,03	752,12	813,6
615,78	716,03	754,7	814

628,45	722,28	755,03	814,8
644,86	728,11	758,37	817,9
650,45	728,7	760,54	818,9
652,2	729,03	764,03	820,7
656,86	730,12	769,28	821,1
661,45	731,95	775,45	825,6
666,45	735,03	781,2	828,6
667,7	736,95	781,7	841,5
668,95	737,37	785,62	842
675,28	738,28	792,78	842,9
675,78	739,78	793,37	849,6
685,53	740,62	795,28	874,7
694,28	743,62	797,62	878,8
697,61	747,2	798,95	897,5

Tugas:

1. Hitung rata-rata dan standar deviasi kumpulan data
2. Buatlah standarisasi himpunan data baru
3. Buat Plot data pada grafik untuk melihat perubahan sebelum standarisasi dan setelahnya.

BAB 4

STATISTIK INFERENSIAL: CONFIDENCE INTERVALS

Bab keempat dari buku ini, yang berjudul "Statistik Inferensial: *Confidence Intervals*," bertujuan untuk memberikan pemahaman mendalam tentang salah satu konsep penting dalam statistik inferensial. *Confidence intervals* atau interval kepercayaan adalah alat yang sangat berguna dalam membuat estimasi tentang parameter populasi berdasarkan data sampel.

Pemahaman yang baik tentang confidence intervals sangat penting karena mereka memberikan informasi tentang tingkat ketidakpastian dalam estimasi kita dan membantu dalam pengambilan keputusan yang lebih baik. Diharapkan bab ini dapat memberikan wawasan yang berguna dan menjadi dasar yang kuat untuk melakukan analisis statistik inferensial.

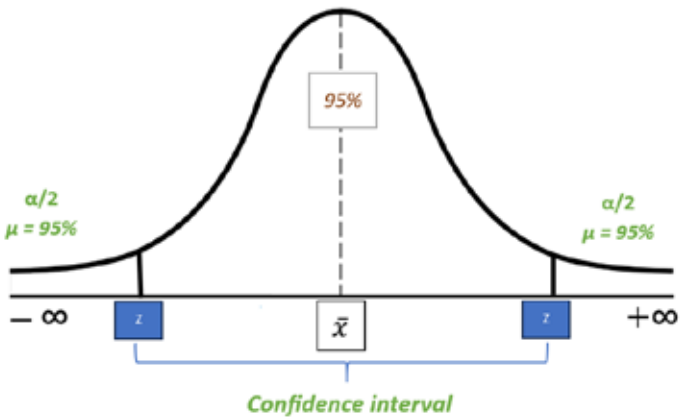
Pada bagian ini kita akan membahas lebih lanjut tentang statistik inferensial berdasarkan pada tingkat keyakinan. Jika pada materi sebelumnya kita mempelajari statisktik inferensial pada konteks probabilitas, maka kita akan melihat bagaimana penerapan statistik inferesial pada suatu tingkat *Confidence Intervals* (interval kepercayaan) tertentu [4], [6].

4.1. Interval Kepercayaan (*Confidence Intervals*)

Dalam pelajaran sebelumnya kita belajar tentang nilai estimator. Namun, nilai estimator tidak terlalu bisa diandalkan. Bayangkan jika kita mengunjungi 5% dari restoran di London, dan mengatakan bahwa harga makanan rata-rata bernilai 22,50 pound. Kita mungkin benar, tetapi kemungkinan besar nilai sebenarnya bisa dalam antara 20 sampai 25 pound. Jauh lebih aman untuk mengatakan bahwa makanan rata-rata di London adalah antara 20 dan 25 pound. Cara ini menunjukkan bahwa kita telah menciptakan interval kepercayaan (*Confidence Intervals*) di sekitar perkiraan 20 hingga 25 pound. Interval kepercayaan adalah representasi realitas yang jauh lebih akurat atau mendekati kondisi yang ada. Namun, masih ada beberapa ketidakpastian yang ada.

Oleh karena itu kita dapat mengukur dalam tingkat kepercayaan. Jika kembali ke contoh di atas, kita dapat mengatakan bahwa kita memiliki 95% keyakinan bahwa parameter populasi terletak antara 20 dan 25 pond. Perlu diingat bahwa kita tidak akan pernah bisa mengatakan 100% yakin, kecuali kita menilai seluruh populasi. Namun demikian, sekalipun kita menggunakan populasi, tetap ada kemungkinan 5% parameter populasi sebenarnya berada di luar kisaran 20 hingga 25 pound.

Tingkat kepercayaan (*confidence level*) itu dilambangkan dengan satu minus Alpha (1 – α) dan nilainya antara 0 dan 1 [6]. Misalnya, jika kita ingin 95% yakin bahwa parameternya ada di dalam interval, maka Alpha adalah 5%. Jika kita menginginkan tingkat kompetensi atau akurasi yang lebih tinggi, katakanlah sembilan puluh sembilan persen, Alpha akan menjadi 1%.

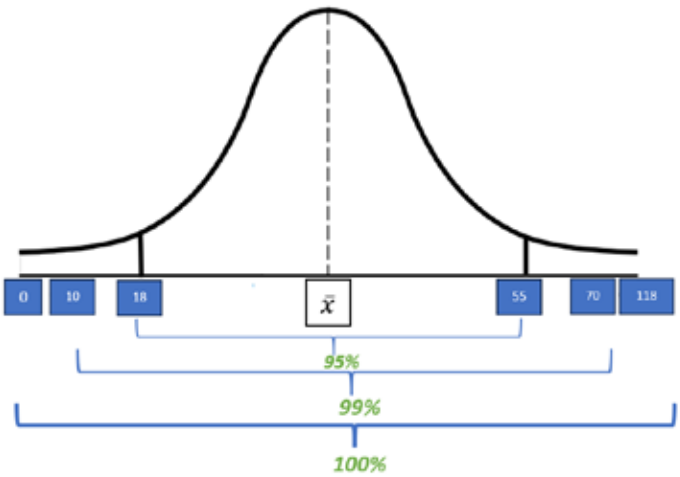


Gambar 4.1. *Confidence Interval*

4.2. Klarifikasi *Confidence interval*

Setelah mendapatkan gambaran tentang interval kepercayaan, kita diminta untuk menggunakan data yang sama untuk menemukan interval kepercayaan lainnya. Mari kita mundur selangkah dan mencoba memahami interval kepercayaan diri sedikit

lebih baik. Berdasarkan grafik distribusi normal, dapat diketahui di mana rata-rata sampel berada, di elati grafik. Sekarang, jika kita tahu bahwa suatu variabel berdistribusi normal, pada dasarnya kita membuat pernyataan bahwa sebagian besar pengamatan akan berada di sekitar rata-rata dan sisanya jauh darinya. Dalam interval kepercayaan, ada batas bawah dan batas atas. Seperti pada gambar 4.1, interval kepercayaan 95 akan menyiratkan bahwa kita 95 yakin bahwa rata-rata populasi sebenarnya jatuh dalam interval ini, sehingga ada 2,5 kemungkinan berada di sebelah kiri batas bawah dan 2,5 persen peluang berada di sebelah kanan. Secara keseluruhan, ada kemungkinan 5 persen bahwa interval kepercayaan kita tidak mengandung rata-rata populasi yang sebenarnya. Jadi elati Alpha adalah 0,05 atau 5 persen, kita memiliki Alpha dibagi dengan 2 hasilnya adalah 2,5 persen peluang rata-rata sebenarnya ada di sebelah kiri interval, dan 2,5 persen di sebelah kanan. Bagaimana jika kita melihat interval kepercayaan 90. Dalam hal ini, intervalnya ada kemungkinan 10 rata-rata sebenarnya ada di luar, dengan interval sebenarnya lima persen di setiap sisi. Hal ini menyebabkan interval kepercayaan menyusut, jadi kepercayaan diri kita juga menurun dengan interval lebih kecil. Demikian pula, untuk interval kepercayaan 99. Kita akan memiliki kepercayaan diri yang lebih tinggi tetapi dengan interval kepercayaan yang lebih besar.



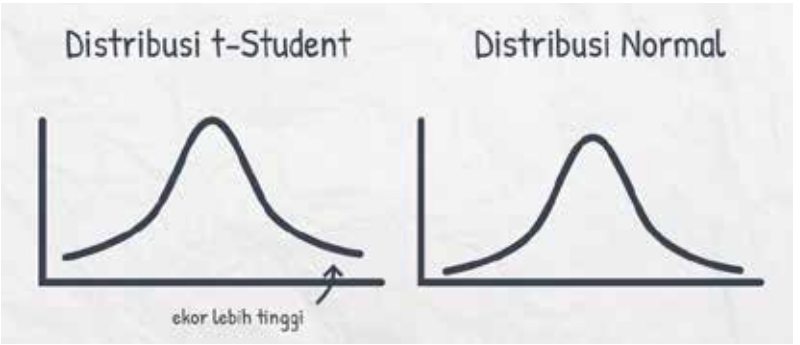
Gambar 4.2. Klarifikasi *Confidence interval*

Gambar 4.2. adalah contoh usia orang atau siswa yang mengambil kuliah (S1, S2, S3). Dalam contoh ini kita akan menebak usia orang yang ikut kuliah. Kita bisa 95% yakin bahwa kemungkinan besar, orang berusia antara 18 dan 55 tahun berdasarkan pada kenyataan data online. Namun jika kita tidak memiliki informasi tentang usia salah satu siswa maka interval bisa menjadi luas. Jadi saya 95% percaya diri, siswa kuliah berusia antara 18 dan 55 tahun. Kita juga dapat 99% yakin bahwa siswa berusia antara 10 dan 70 tahun. Namun saya 100% yakin bahwa siswa kuliah berusia antara nol dan 118 tahun, yang merupakan usia orang tertua yang masih hidup. Dalam hal ini yang kita lihat, ada *trade-off* antara tingkat kepercayaan dan rentang interval. Namun demikian, interval kepercayaan 100% sama sekali tidak berguna karena kita harus memasukkan semua usia manusia untuk mendapatkan kepercayaan diri 100%. Keyakinan 99 memberi kita rentang yang lebih sempit tetapi masih belum cukup pemahaman untuk masalah khusus ini. Jika kita menebak usia 25 tahun atau dengan tingkat keyakinan 5%, maka perkiraan yang cukup berguna karena kita menyebutkan angka pasti. Namun tingkat kepercayaan 5% terlalu kecil untuk

kita gunakan dalam analisis yang berarti atau penting. Jadi selalu ada *trade-off*, yang tergantung pada masalah yang sedang dihadapi atau diteliti. Dalam hal ini, tingkat keyakinan 95% adalah **norma** yang diterima umum ketika kita tidak terlalu banyak berkompromi dengan akurasi, tetapi masih mendapatkan interval yang relatif sempit.

4.3. Distribusi-t (*Student's t-Distribution*)

Mari kita memahami lebih lanjut tentang distirbusi T atau disebut juga *student's t-distribution*. Poin penting dari distribusi-t adalah kemampuan untuk mengekstrak sampel kecil tetapi masih menghasilkan prediksi yang baik. Berikut adalah gambar dari distribusi-t.



Gambar 4.3. Perbandingan distribusi T dengan distibusi normal
(Sumber: https://www.rachmatwahid.com/2020/06/distribusi-t-student.html)

Pada gambar 4.3. dapat dipahami bahwa perbedaan distribusi-t dengan distribusi normal adalah terletak pada ekor yang lebih rata. Distribusi-t merupakan distribus sampel, sedang distribusi normal pada dasarnya adalah distribusi pada populasi. Oleh karena itu distribusi-t menggunakan derajat kebebasan (*degree of freedom*) yang disingkat dengan df, sedangkan distribusi normal menggunakan parameter.

Rumus yang memungkinkan kita untuk menghitungnya distribusi-t dengan dan minus 1 derajat kebebasan (*degrees of freedom*) dan tingkat *alpha* yang signifikan sama dengan rata-rata sampel dikurangi rata-rata populasi dibagi dengan kesalahan standar sampel. Mirip dengan rumus Z, karena bagaimanapun juga, distribusi-t adalah perkiraan dari distribusi normal. Karakteristik terakhir dari distribusi-t adalah bahwa ada derajat kebebasan biasanya untuk sampel n yang kita miliki dan minus 1 derajat kebebasan. Contohnya jika sampel dari 20 pengamatan, maka tingkat kebebasan adalah 19.

$$t_{n-1,\alpha} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(d.f. adalah n-1)

Harap dicatat bahwa setelah baris ketiga puluh jumlahnya tidak terlalu bervariasi. Sebenarnya, setelah 30 derajat kebebasan, tabel distribusi-t menjadi hampir sama dengan statistik Z, karena derajat kebebasan bergantung pada sampel. Intinya, semakin besar sampel, semakin dekat kita dengan angka aktual, aturan umum praktis adalah bahwa untuk sampel yang berisi lebih dari 50 pengamatan kita menggunakan tabel Z bukan tabel distribusi-t. Bagaimana penggunaan tabel distribusi T akan disampaikan lebih lanjut.

4.4. *Confidence Intervals: Varians*

4.4.1. Populasi Diketahui: z-score

Interval kepercayaan adalah rentang parameter populasi dan estimasi yang diharapkan. Ada dua situasi utama ketika kita menghitung interval kepercayaan untuk suatu populasi adalah (1) ketika varians populasi diketahui (z-score), dan (2) ketika tidak

diketahui (t-score). Keduanya digunakan pada saat ditentukan pada situasi mana kita berada kita akan menggunakan metode perhitungan yang berbeda.

Dalam praktik statistik yang ada, kita hampir tidak pernah memiliki data populasi bahkan jika kita memiliki data populasi sekalipun. Kita mungkin tidak dapat menganalisis populasi, mungkin sangat banyak sehingga tidak masuk akal untuk digunakan sekaligus. Sebagai contoh adalah orang-orang yang menggunakan internet. Data yang dimiliki Google dapat dikatakan mendekati data populasi. Namun, bahkan mereka tidak menggunakannya, Google sadar bahwa ada orang yang bukan bagian dari ekosistem Google dengan cara apa pun. Di luar Google, orang dapat menggunakan browser lain seperti Opera, Safari, dan atau mesin pencari lain seperti Bing, Duck Go atau penyedia video yang berbeda dari YouTube. Selain itu, mereka juga dapat menelusuri internet dengan mode penyamaran. Orang-orang ini adalah bagian dari populasi orang yang menggunakan Internet. Tetapi Google tidak memiliki banyak data tentang mereka.

Seperti yang kita lihat, bahkan perusahaan yang memiliki data paling banyak belum tentu memiliki data populasi. Jadi, jika Google ingin menggunakan metode statistik untuk menargetkan pengguna dengan iklan, Google dasarnya akan menggunakan data sampel untuk dibentuk dengan varians populasi yang tidak diketahui untuk menebak preferensi mereka.

Dalam materi ini, kita akan mengeksplorasi interval kepercayaan untuk rata-rata populasi dengan varians yang diketahui. Asumsi penting dalam perhitungan ini adalah bahwa populasi didistribusikan secara normal, bahkan jika tidak. Kita harus menggunakan sampel besar dan mengunakan *Center limit Theorem* (CLT) untuk melakukan normalisasi. Ingat jika kita bekerja dengan sampel yang cukup besar, kita dapat mengasumsikan normalitas sampel berarti baik-baik saja. Hasil dari *confidence*

interval (interval keyakinan) dapat diperoleh dengan rumus:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Nilai z pada rumus tersebut merupakan z-score yang dapat dilihat pada gambar 4.1. Detail bagaimana rumus tersebut digunakan akan dibahas lebih lanjut pada appendix bab 4 ini. Namun demikian, sampai pada saat ini, kita bisa mendapat gambaran bagaimana *confidence interval* melalui contoh 30 data gaji analis data.

$117,313	$104,002	$113,038	$101,936
$84,560	$113,136	$80,740	$100,536
$105,052	$87,201	$91,986	$94,868
$90,745	$102,848	$85,927	$112,276
$108,637	$96,818	$92,307	$114,564
$109,714	$108,833	$115,295	$89,279
$81,720	$89,344	$114,426	$90,410
$95,118	$113,382		

Dari contoh data tersebut kita akan mencari tahu berapa rata-rata nilai gaji seorang data analis. Dengan contoh tersebut maka kita dapat gunakan contoh nilai rata-rata, standar deviasi, dan standar eror, sebagai berikut:

Sample mean	$ 100,200
Population std	$ 15,000
Standard error	$ 2,739

Perhitungan selanjutnya adalah menentukan nilai *confidence interval* dengan tabel Z. Dengan menggunakan tabel tersebut, maka secara ringkas hasilnya adalah

95% CI, $z_{0.025}$	1.96
99% CI, $z_{0.005}$	2.58

Jika kita menggunakan tingkat kepercayaan 95% maka hasilnya adalah 1.96, sedangkan jika kita menggunakan tingkat kepercayaan 99% maka hasilnya adalah 2.58. Dengan menggunakan rumus yang ada kita akan mendapatkan nilai rata-rata sebagai berikut:

Z	CI rendah	CI tinggi
95%	$ 94,833	$ 105,568
99%	$ 93,135	$ 107,266

Hasil tersebut menunjukkan jika kita menggunakan tingkat kepercayaan (CI) 95% maka nilai rata-rata gaji seorang data analis ada pada rentang nilai $94,833 sampai $ 105,568. Jika kita menggunakan tingkat kepercayaan (CI) 99% maka nilai rata-rata gaji data analis adalah dalam rentang nilai $ 93,135 sampai $ 107,266. Dari kedua hasil tersebut kita dapat melihat bahwa jika kita menggunakan tingkat kepercayaan yang lebih tinggi maka rentang nilai akan jadi lebih besar, dan juga sebaliknya.

4.4.2. Populasi Tidak Diketahui: *t-score*

Jadi kita telah belajar bahwa interval kepercayaan berdasarkan sampel kecil dari populasi yang didistribusikan secara normal dihitung dengan statistik t. Mari kita periksa contoh serupa dengan yang kita lihat sebelumnya. Pada contoh kali ini, kita akan mencoba mencari tahu barapa nilai rata-rata gaji seorang data analis. Sebagai contoh, kita memiliki data sampel sebanyak sembilan data sebagai berikut: $78,000; $90,000; $75,000; $117,000; $105,000; $96,000; $89,500; $102,300; $80,000. Dari data tersebut maka kita dapat mencari rata-rata (sample mean) sebesar: $92,533; standar deviasi sampel sebesar $ 13,932; dan standar eror $ 4,644 (bagaimana mendapatkan angkanya dengan excel akan ditunjukkan dalam appendix bab 4 ini). Dari semua informasi tersebut kita tidak mendapatkan informasi kunci yaitu, varians populasi. Oleh karena

sampel yang kecil dan tidak ada informasi varians populasi maka kita akan menggunakan distribusi t. Rumusnya adalah:

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Jika diperhatikan maka rumus tersebut mirip dengan z-score. Namun sekali lagi n-1 merupakan d.f. untuk varians yang tidak diketahui. Dengan contoh di atas, dengan nilai sampel 9 maka d.f. adalah 8, dan dengan nilai alpha 5% maka nilai yang diginakan adalah α/2 maka hasilnya adalah 0,025 (0.05/2). Dari hasil perhitungan tersebut maka dapat dicari nilai t-statistik adalah sebesar 2,306, atau 2,31.

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055

Dengan rumus yang ada maka hasil yang diperoleh adalah, dengan tingkat kepercayaan 95% maka gaji adalah $81,806 sampai $103,261.

4.5. *Margin of Error*

Dalam beberapa pelajaran terakhir kita berkonsentrasi pada interval kepercayaan diri. Kita belajar tentang dua situasi

ketika varians populasi diketahui yang sangat jarang terjadi dalam praktik dan ketika varians tidak diketahui maka dapat mengikuti distribusi-t (*Student t*). Dengan rumus yang ada memungkinkan kita menemukan interval kepercayaan seperti yang sudah kita pelajari sebelumnya.

Nilai dalam rumus pada bagian setelah penjumlahan atau pengurangan rata-rata sampel (±), merupakan nilai yang menentukan rentang interval kepercayaan, yang menunjukkan *margin of eror*. Dengan demikian, rata-rata populasi sebenarnya berada dalam interval yang ditentukan oleh rata-rata sampel ditambah atau dikurangi *margin of eror* (margin kesalahan). Sekarang kita sampai pada inti dari pelajaran ini. Mendapatkan margin kesalahan yang lebih kecil berarti interval kepercayaan akan lebih sempit karena kita menginginkan prediksi yang lebih baik. Dalam hal ini adalah penting untuk memiliki interval kepercayaan sesempit mungkin. Bagian terbaiknya adalah kita dapat mengontrol margin kesalahan.

Dengan demikian, poin penting dari *margin of error* yang dapat dipahami adalah, semakin besar *confidence interval* atau tingkat kepercayaan, maka nilai margin of error juga semakin besar. Tingkat kepercayaan yang lebih rendah menghasilkan interval yang lebih sempit. Kondisi ini memberikan margin of error yang lebih kecil. Demikian juga dengan standar deviasi. Standar deviasi yang lebih rendah menunjukkan bahwa dataset lebih terkonsentrasi di sekitar *mean*. Kondisi ini membawa pada margin of error yang lebih kecil. Terakhir, jika kita memiliki ukuran sampel yang lebih tinggi, maka kita akan mengurangi *margin of error* (margin kesalahan). Semakin banyak pengamatan yang kita miliki dalam sampel, maka kita semakin yakin dalam prediksi. Dengan demikian semakin banyak pengamatan yang ada dalam sampel, semakin tinggi peluang untuk mendapatkan nilai rata-rata sebenarnya dari seluruh populasi.

4.6. Interfal keyakinan: dua rata-rata (*Confidence intervals: Two means*)

4.6.1. Sampel terikat (*Dependent samplesi*)

Sejauh ini kita telah berbicara tentang *confidence interval* dengan varians populasi yang diketahui atau tidak diketahui. Namun sampai saat ini, kita hanya mempertimbangkan satu populasi. Pada saat ini kita akan mengeksplorasi interval kompetensi ke dalam dua populasi. Pemahaman ini penting dalam aplikasi di dunia nyata. Dalam beberapa kasus, sampel yang telah kita ambil dari dua populasi akan bergantung satu sama lain dan dalam kasus lain mereka akan menjadi *dependent samples* (sampel bergantung/ terikat).

Kita bisa mendapatkan contoh *dependent samples* secara langsung yang dapat terjadi dalam beberapa situasi. Pertama, ketika kita meneliti subjek yang sama, dari waktu ke waktu. Contohnya adalah orang dengan penurunan berat badan dan uji sampel darah. Pada dasarnya kita akan melihat orang yang sama pada kondisi sebelum dan sesudah orang tersebut mendapatkan *treatment* atau perlakuan, misalnya olah raga tertentu. Dalam hal ini, kita juga dapat melihat hubungan sebab akibat dari suatu *treatment* yang sedang diobservasi, yaitu sebelum dan sesudah *treatment*.

Tes statistik *dependent samples* sering digunakan saat mengembangkan obat. Dalam analisa ini, distribusi dapat dianggap normal. Contohnya jika ilmuan yang mengembangkan pil yang dapat meningkatkan konsentrasi magnesium dalam darah. Obat ini tampak sangat menjanjikan tetapi tidak ada data untuk mendukung klaim tersebut. Setelah menguji obat di laboratorium, sekarang saatnya untuk melihat efek sebenarnya pada orang-orang. Apa yang dapat dilakukan adalah mengambil sampel 10 orang, kemudian tes atau kadar magnesium sebelum dan sesudah minum pil tersebut.

Dua sampel yang bergantung adalah kadar magnesium sebelum dan kadar magnesium setelahnya, dan jelas bahwa sampel

tersebut adalah orang yang sama yang diuji karena sampelnya tergantung/dependen. Setiap kali Kita melakukan tes darah, kadar magnesium dinyatakan dalam miligram per desiliter dan orang sehat biasanya memiliki antara 1,7 dan 2,2 miligram magnesium per desiliter. Berikut adalah tabel yang berisi sampel 10 orang dan ada kadar magnesium sebelum dan sesudah minum pil untuk beberapa waktu:

Tabel 4.1. Contoh kasus observasi pasien

Pasien	Sebelum	Sesudah	Selisih
1	2.00	1.70	-0.30
2	1.40	1.70	0.30
3	1.30	1.80	0.50
4	1.10	1.30	0.20
5	1.80	1.70	-0.10
6	1.60	1.50	-0.10
7	1.50	1.60	0.10
8	0.70	1.70	1.00
9	0.90	1.70	0.80
10	1.50	2.40	0.90
Rata-rata selisih			0.33

Dari data tersebut terlihat seperti populasi tunggal, maka mari kita hitung rata-rata dan standar deviasi dari perbedaan. rata-ratanya adalah 0.33 untuk ukuran sampel 10. Rumus yang memungkinkan kita menghitung interval kepercayaan adalah sebagai berikut, dimana $\bar{d}$ adalah rata-rata selisih obyek 1 dan 2, t_{n-1} adalah nilai tabel, α adalah tingkat kesalahan, S_d adalah standar deviasi, n adalah jumlah sampel atau pengamatan.

$$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

Populasi didistribusikan secara normal tetapi sampel yang ada hanya berisi 10 pengamatan. Oleh karena itu distribusi menggunakan distribusi T-*student* dan statistik yang sesuai adalah T. Secara tidak langsung, kita dapat melihat bahwa rumus tersebut sama dengan yang untuk satu populasi dengan varians yang tidak diketahui. Mari kita gunakan tingkat kepercayaan dan masukkan angka-angka pada tingkat kepercayaan 95% yang paling umum dipakai. Dari data tersebut juga dapat dihitung, nilai rata-rata dan standar deviasi, sebagai berikut:

Mean	0.33
St. deviation	0.45

Dengan nilai d.f. adalah 9 (derajat kebebasan yaitu n-1, maka 10-1=9) untuk interval kepercayaan 95% (tingkat alpha 5% : 2 = 0.025) adalah 2,262.

d.f. / α	0.1	0.05	[illegible]	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
[illegible]					
9	1.383	1.833	2.262	2.821	3.250
[illegible]					
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921

Sekarang, dengan data yang ada, kita dapat menghitung interval kepercayaan. Kita dapat memasukan semua angka pada rumus tersebut dan berikut adalah hasilnya:

$$0.33 \pm 2.26 \frac{0.45}{\sqrt{10}} = (0.01, 0.65)$$

Dengan menggunakan *confidence interval*, maka hasilnya terletak di kisaran antara 0.01 sampai 0.65. Kemudian bagaimana kita menafsirkan hasil ini dengan baik, (1) dalam tingkat kepercayaan 95%, maka hasil nilai rata-rata sebenarnya akan berada dalam interval tersebut (lihat tabel 4.1. nilai rata-rata 0.33 berada dalam rentang interval 0.01 dan 0.65). (2) Selain itu, seluruh interval bernilai positif, hal ini menunjukkan bahwa arti sebenarnya dari perbedaan tersebut pasti positif. (3) Oleh karena itu, dengan kepastian 95% kita dapat mengatakan bahwa kadar magnesium dalam darah subjek uji menjadi lebih tinggi. Tujuan dari tes ini adalah untuk menentukan apakah obat tersebut efektif berdasarkan sampel kecil, dan hasilnya menunjukkan ada peningkatan magnesium dalam darah.

4.6.2. Sampel bebas (*Independent samples*)

Setelah mempelajari *dependent samples*, kita akan mempelajari *confidence interval* untuk *independent sample* (sampel bebas) seperti yang telah disampaikan sebelumnya. Dalam hal ini ada tiga subkasus untuk *independent sample*, yaitu:

1. kasus yang diketahui varians populasi,
2. kasus yang varians populasi yang tidak diketahui tetapi diasumsikan sama.
3. kasus yang varians populasi yang tidak diketahui yang diasumsikan berbeda.

Walaupun terdapat 3 kasus *independent sample,* namun kita kana mempelajari nomer 1 dan 2 saja.

1. Varians populasi diketahui

Pertama-tama, pada pembelajaran ini kita akan fokus pada sampel independen dengan varians populasi yang diketahui. Kita akan menggunakan contoh yang memungkinkan kita untuk memahami konsep ini lebih baik. Contoh kali ini adalah kita ingin menguji nilai dari dua departemen pada suatu universitas. Nilai

dalam universitas tersebut dinyatakan dalam persentase, sampel kita gunakan adalah dari departemen teknik dan manajemen. Berikut adalah ringkasannya:

Tabel 4.2. Contoh kasus nilai rata-rata dua departemen

	Teknik	Manajemen
Size (pengamatan)	100	70
Sample mean	58	65
Population standard error	10	5

Terdapat sampel 100 mahasiswa teknik dan nilai rata-rata adalah 58%, sedangkan nilai standar deviasi populasi adalah 10 poin persentase dengan varians diketahui. Pada kelompok yang lain juga diuji sampel 70 siswa dari departemen manajemen, yang memiliki nilai rata-rata 65%, sedangkan diketahui bahwa standar deviasi populasi adalah 5.

Sebelum melakukan analisa lanjut, ada beberapa hal yang perlu diperhatikan dalam contoh ini. Pertama, distribusi populasi diasumsikan normal. Kedua, varians populasi diketahui, dan ketiga, ukuran sampel berbeda. Ukuran sampel yang berbeda adalah kejadian umum di dunia nyata pada pengujian *independent sample*. Pengamatan pada kasus ini benar-benar berbeda, mereka adalah siswa dari departemen yang berbeda dengan guru yang berbeda memperoleh nilai yang berbeda saat mengikuti ujian yang berbeda. Disamping itu, seorang siswa atau pengajar dari teknik sama sekali tidak mempengaruhi apa pun dari seorang siswa atau pengajar pada departemen manajemen. Intinya, kedua sampel tersebut benar-benar independen.

Dari contoh ini, apa yang akan kita uji? Kita ingin menemukan perbedaan antara nilai siswa dari teknik dan manajemen dengan tingkat interval kepercayaan 95%. Untuk prosesnya, pertama-tama informasi yang diperoleh adalah sampel statistik atau varians populasi yang diketahui, dan populasi diasumsikan

mengikuti distribusi normal. Semua informasi ini mengarahkan kita ke uji statistik Z, bukan T. Point yang perlu diperhatikan juga adalah varians, varians diperoleh dari selisih antara kedua rata-rata tersebut sama dengan varians nilai yang diterima oleh mahasiswa teknik dibagi dengan ukuran sampel mahasiswa teknik ditambah varians nilai yang diperoleh mahasiswa manajemen dibagi dengan ukuran sampel mahasiswa manajemen. Secara ringkas rumus perbedaan varians adalah sebagai berikut, dimana σ adalah varians.

$$\sigma^2_{diff} = \frac{\sigma^2_e}{n_e} + \frac{\sigma^2_m}{n_m}$$

Logika yang mendasari rumus tersebut adalah bahwa dispersi bersifat aditif. Lebih banyak variabel berarti variabilitas yang lebih tinggi atau sama. Berikut adalah hasilnya jika nilai angka dalam tabel 4.2. dimasukan dalam rumus.

$$\sigma^2_{diff} = \frac{10^2}{100} + \frac{5^2}{70} = 1.36$$

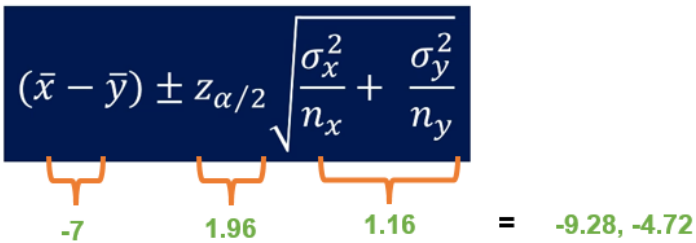
Dari hasil tersebut, kita dapat menghitung standar error adalah akar kuadrat dari 1.36, yaitu 1.16. Dengan semua informasi ini maka kita dapat masukkan pada tabel 4.3.

Tabel 4.3. Perhitungan kasus nilai rata-rata dua departemen

	Teknik (x)	**Manajemen (Y)**	**Selisih (x-y)**
Size (pengamatan)	100	70	?
Sample mean	58	65	-7
Population standard error	10	5	1.16

Nilai yang dicari adalah berselisih pengamatan. Yang menunjukkan ada tidaknya perbedaan nilai dari dua departemen

yang diuji. Dengan tingkat kepercayaan 95% maka nilai tabelnya adalah 1.96. Maka kita akan menggunakan rumus *confidence interval*:



Dengan demikian maka dengan sumus yang ada hasil confidence interval adalah -9.28, -4.72. Dari hasil tersebut kita akan mengartikannya, yaitu: (1) dengan tingkat kepercayaan 95% kita dapat yakin bahwa perbedaan rata-rata sebenarnya antara nilai teknik dan manajemen berada dalam interval ini. (2) hasil perhitungan menunjukkan bahwa seluruh interval negatif. Hal ini karena kita menghitung nilai Teknik dikurangai nilai manajemen. Hal ini dapat dipahami juga bahwa para insinyur secara konsisten mendapatkan nilai yang lebih rendah. Kondisi sebaliknya akan terjadi jika rata-rata manajemen dikurangi nilai rata-rata teknik, maka hasil akan positif (4.72,9.28) yang menunjukkan nilai rata-rata manajemen consisten lebih tinggi dari pada teknik.

2. Varians populasi yang tidak diketahui tetapi diasumsikan sama

Sekarang kita akan masuk pada pembahasan tentang sampel independen dengan varians yang tidak diketahui tetapi diasumsikan sama. Untuk mempermudah, kita akan langsung menggunakan contoh pada materi ini. Pada kasus kali ini, kita akan menghitung selisih harga apel di New York dan LA, misalnya. Kita mendapat sepuluh toko di New York dan delapan di Los Angles (LA). Dari hasil pengamatan kita dapat membuat tabel dan bertanya bagaimana kita dapat membuat interval kepercayaan yang menunjukkan perbedaan

antara harga apel di New York dan LA. Dalam hal ini, kita tidak tahu berapa varians populasi apel di New York atau LA, tetapi kita berasumsi sama untuk NY dan LA.

Tabel 4.4. Sampel harga apel di New York dan Los Angles

NY	LA
$ 3,80	$ 3,02
$ 3,76	$ 3,22
$ 3,87	$ 3,24
$ 3,99	$ 3,02
$ 4,02	$ 3,06
$ 4,25	$ 3,15
$ 4,13	$ 3,81
$ 3,98	$ 3,44
$ 3,99	
$ 3,62	

Dari contoh tabel 4.4. dapat dihitung rata-rata, dan standard deviasi sebagai berikut:

Tabel 4.5. Deskripsi data harga apel di New York dan Los Angles

	NY (X)	LA (Y)
Sample mean	$ 3,94	$ 3,25
Sample std	$ 0,18	$ 0,27
Sample size	10	8

Setelah informasi tersebut diperoleh maka kita perlu memperkirakan varians sampel gabungan (*pooled variance*) dengan sebagai berikut, dengan keteragnan S_p^2 adalah *pooled variance*, n_x

adalah jumlah sampel x, n_y adalah jumlah sampel y, S_x^2 dan S_y^2 masing-masing adalah standar deviasi x dan y.

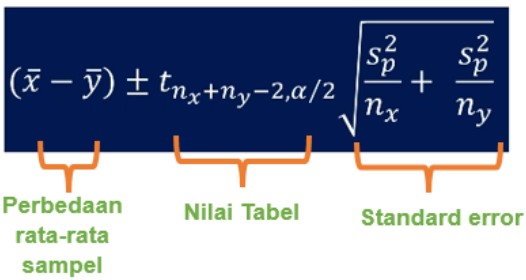
$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

Jika kita masukan semua angka yang ada maka hasil *pooled variance* adalah berikut:

$$\frac{(10 - 1)0.18^2 + (8 - 1)0.27^2}{10 + 8 - 2} = 0.05$$

Hasil perhitungan menunjukkan sampel ***pooled variance*** adalah 0,05 maka untuk mencari standar deviasi ***pooled*** (gabungan) adalah akar kuadrat dari 0.05 yaitu $ 0.22 (22 cent).

Untuk menentukan *confidence interval*, kita perlu melihat kriteria informasi yang ada yaitu: (1) populasi varians yang tidak diketahui, dan (2) sampel kecil. Atas kedua informasi ini maka kita dapat menggunakan uji statistik T. Berikut rumus untuk interval kepercayaan untuk *independent sampel* dengan varians populasi yang tidak diketahui.



Dalam rumus tersebut terdapat hal yang berbeda dalam perhitungan *degree of freedom* untuk mencari nilai tabel. yaitu rumusnya (data menggunakan tabel 4.5):

$$n_x + n_y - 2 = 10 + 8 - 2 = 16$$

Derajat kebebasan sama dengan total ukuran sampel dikurangi jumlah variabel yaitu dua karena ada dua kelompok yang diuji. Sejauh ini kita telah melihat t dengan n minus 1 derajat kebebasan karena kita memiliki satu ukuran sampel dan hanya satu variabel. Kali ini kita memiliki dua ukuran sampel, satu dari sepuluh pengamatan dan satu dari delapan pengamatan dari variabel harga apel di New York dan di LA. Oleh karena itu kita memiliki 16 derajat kebebasan (d.f.). Dalam tabel kita dapat mencarinya dengan hasil 2.120 sebagai berikut:

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861

Kita dapat meringkas kembali semua data yang dibutuhkan sebagai berikut:

	NY	LA	Selisih
Sample mean	$ 3,94	$ 3,25	$ 0.69
Sample std	$ 0,18	$ 0,27	
Sample size	10	8	
Pooled variance	0,05		
Pooled std	0,22		
95% CI, $t_{16,0.025}$	2,12		

Dari semua informasi yang diperoleh maka kita dapat masukan kedalam rumus confidence interval sebagai berikut:

$$(3.94 - 3.25) \pm 2.12 \sqrt{\frac{0.05}{10} + \frac{0.05}{8}}$$

Hasilnya adalah rentang nilai $CI_{95\%}$ adalah 0.47, 0.92. Apa arti nilai tersebut, kita dapat yakin bahwa 95% perbedaan sebenarnya antara harga apel di New York dan di LA adalah antara 0.47 dan 0.92. Dalam hal ini nilai selisih rata-rata adalah $ 0.69 berada di antara 0.47 dan 0.92, sehingga jelas bahwa apel di New York lebih mahal daripada di LA.

Appendix Bab 4:
Excel *Inferential Statistics - Confidence Intervals*

Pada bagian ini, kita akan mengulang kembali apa yang telah dipelajari pada bab 4 tentang *Inferential Statistics: Confidence Intervals*. Fokus pada bagian ini adalah penggunaan rumus Ms Excel pada materi yang sudah dibahas pada bab 4. Dengan demikian, kita dapat terbantu pada saat menggunakan rumus statistika yang sudah dipelajari.

A.4.1. Menggunakan Distribusi Normal Tabel Z

Berikut adalah Tabel Z yang akan digunakan menentukan nilai *confidence intervals*

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Contoh dari penggunaan tabel Z tersebut adalah sebagai berikut:

Contoh 1. *Confidence intervals* (dan tes hipotesis dua sisi)

Pertanyaan : berapa nilai kritis untuk interval kepercayaan 95%?

Langkah 1 : 95% = 0.95 = (1-α)
α = 0.05

Langkah 2 : Karena kita menggunakan interval kepercayaan (atau uji dua sisi) kita tertarik pada α/2, atau lebih baik: (1-α/2) cara menghitung (1-α/2)
(1-α/2) = (1 - 0.025) = 0.975

Langkah 3 : Cari nilai terdekat dengan 0,975 (atau yang tepat) dalam tabel. Dalam hal ini adalah: 0,975

Langkah 4 : Nilai kritis (Z α/2), diberikan oleh jumlah header baris dan kolom yang sesuai 1.9+0.06 = 1.96. **Jawaban: $z_{0.025} = 1.96$**

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

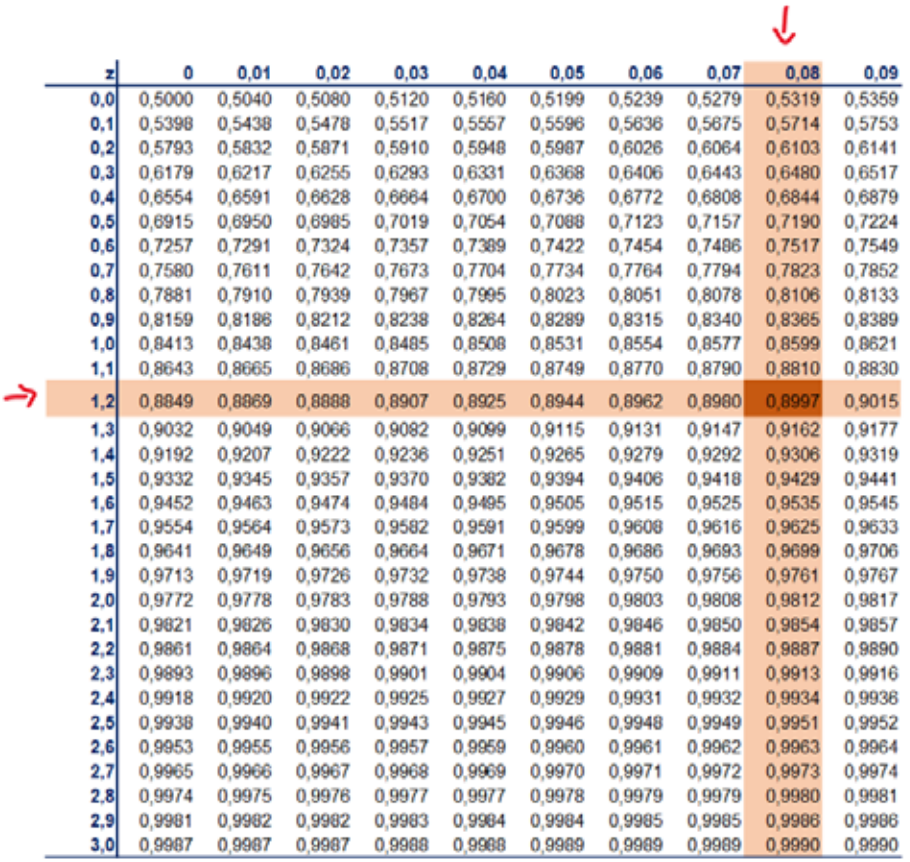
Contoh 2: Kasus umum (pengujian hipotesis satu sisi).

Pertanyaan: berapa nilai kritis untuk α = 0,1

Langkah 1 : Temukan (1-α)
1-α = (1-0.1) = 0.9

Langkah 2 : Cari nilai terdekat dengan 0,9 dalam tabel.
Dalam hal ini adalah: 0,8997

Langkah 3 : Nilai kritis (z), diberikan oleh jumlah header baris dan kolom yang sesuai, yaitu
1.2+0.08 = 1.28, **Jawaban: $z_{0.1}$ = 1.28**



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Dalam contoh kasus penggunaan tabel Z untuk populasi yang diketahui, yang digunakan dalam materi, kita mendapatkan contoh 30 data harga gaji data analis.

$117,313	$104,002	$113,038	$101,936
$84,560	$113,136	$80,740	$100,536
$105,052	$87,201	$91,986	$94,868
$90,745	$102,848	$85,927	$112,276
$108,637	$96,818	$92,307	$114,564
$109,714	$108,833	$115,295	$89,279
$81,720	$89,344	$114,426	$90,410
$95,118	$113,382		

Dari contoh data tersebut kita akan mencari nilai interval rata-rata gaji data analis. Maka data yang diperlukan adalah rata-rata sampel, standar deviasi, dan standar eror. Dari contoh di atas maka nilai rata-rata dapat dihitung dengan menggunakan rumus excel **=AVERAGE(….. : ….)** dan diperoleh nilai $100.200. Sedangkan nilai standar deviasi pada contoh tersebut telah diketahui atau ditentukan yaitu sebesar $ 15.000. Dari data tersebut maka untuk mencari nilai *standard error* adalah dengan rumus excel: **= nilai standar deviasi populasi /SQRT(30) (15.000/SQRT(30))**. hasilnya adalah $ 2.739. Maka rekap hasil adalah

Sample mean	$ 100,200
Population std	$ 15,000
Standard error	$ 2,739

Dengan tingkat keyakinan 95% maka dapat diperoleh nilai tabel Z adalah $\mathbf{Z}_{0.025} = \mathbf{1.96.}$

Nilai *Confidence interval* rendah diperoleh dengan rumus =

rata-rata sampel – (standar error x nilai $\mathbf{Z}_{0.025}$)

yaitu: 100.200 – (2.739 x 1.96) = 94.833

Nilai *Confidence interval* tinggi diperoleh dengan rumus = rata-rata sampel + (*standard error* x nilai $\mathbf{Z}_{0.025}$) **yaitu: 100.200 + (2.739 x 1.96) = 105.568**

Z	CI low	CI high
	$	$
95%	94.833	105.568

A.4.2. Menggunakan Distribusi Normal Tabel t (Distribusi-t)

Setelah kita mempelajari tabel Z untuk distribusi populasi yang diketahui, maka sekarang kita akan mempelajari penggunaan tabel t untuk populasi yang tidak diketahui.

Berikut adalah Tabel t.

d.f. / α	0,1	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
35	1,306	1,690	2,030	2,438	2,724
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
inf.	1,282	1,645	1,960	2,326	2,576
CI*	80%	90%	95%	98%	99%

*CI stands for confidence intervals

Contoh penggunaan dari tabel t adalah sebagai berikut:

Pertanyaan : berapa nilai kritis untuk α = 0,05 (tingkat keyakinan 95%) dengan d.f. (*degrees of freedom*) = 10

Langkah 1 : Cari header kolom yang menunjukkan α yang sesuai

Langkah 2 : Cari header baris yang menunjukkan d.f. yang sesuai.

Langkah 3 : Nilai kritis (t), diberikan dalam sel yang merupakan persimpangan baris dan kolom yang telah Kita temukan sebelumnya
Jawab: $\mathbf{t}_{10,\,0.05} = \mathbf{1.812}$

Berikut cara menentukan nilai dalam tabel

d.f. / α	0,1	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
35	1,306	1,690	2,030	2,438	2,724
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
inf.	1,282	1,645	1,960	2,326	2,576
CI*	80%	90%	95%	98%	99%

Berikut kita akan menggunakan kembali contoh kasus pada materi untuk mengetahui bagaimana perhitungannya dalam MS Excel. Kasus pada tabel t dengan populasi yang tidak ditahui adalah,

kita ingin mengetahui rata-rata nilai gaji data analis dari sampel yang ada. Berikut adalah datanya:

Dataset	
$	78.000
$	90.000
$	75.000
$	117.000
$	105.000
$	96.000
$	89.500
$	102.300
$	80.000

Dari 9 data sampel tersebut maka kita akan mencari nilai *Mean, Standar Deviasi, Standard error*, nilai t tabel pada tingkat keyakinan 95%. Rumus Excel adalah sebagai berikut, misalkan kita letakan data pada kolom excel B5 sampai B13.

Mean = AVERAGE(B5:B13)
St. Deviation = STDEV.S(B5:B13)
Standard error = nilai standar deviasi / √9

Maka hasil adalah:

Mean	$92.533
St. deviation	$13.932
Standard error	$4.644

Nilai t tabel adalah d.f. 9-1 = 8, dengan 5% dibagi 2 = 2.5% (8, 0.025), maka nilai tabel t adalah sebesar 2.306 atau 2.31.

Nilai *Confidence interval* rendah diperoleh dengan rumus =

rata-rata sampel – (standar error x nilai t $_{8,0.025}$)

yaitu: 92.533 – (4.644 x 2.31) = 81.806

Nilai *Confidence interval* tinggi diperoleh dengan rumus =

rata-rata sampel – (standar error x nilai t $_{8,0.025}$)

yaitu: 92.533 + (4.644 x 2.31) = 103.261

	T	CI low	CI high
95%	$	81.806	$ 103.261

A.4.3. Confidence intervals: Two means (*Dependent samples*)

Selanjutnya kita akan mempelajari penggunaan Excel untuk Dependent samples. Kita akan menggunakan tabel t pada pada kasus ini. Kenapa tabel t, kita sudah membahasnya pada materi sebelumnya. Secara singkat kita akan menggunakan kembali data pasien setelah minum obat. Berikut datanya, dengan contoh selisih sebelum dan sesudah kita letakan pada kolom E5 sampai E14.

Pasien	Sebelum	Sesudah	Selisih
1	2.00	1.70	-0.30
2	1.40	1.70	0.30
3	1.30	1.80	0.50
4	1.10	1.30	0.20
5	1.80	1.70	-0.10
6	1.60	1.50	-0.10
7	1.50	1.60	0.10
8	0.70	1.70	1.00
9	0.90	1.70	0.80
10	1.50	2.40	0.90
	Rata-rata selisih		0.33

Mean (rata-rata selisih) **=AVERAGE(E5:E14)** =0,33

St. deviation (selisih) **=STDEV.S(E5:E14)** =0,45

Sedangkan nilai tabel untuk keyakinan 95%, α=5% (dengan nilai yang digunakan adalah 5% dibagi 2 = 2.5%), d.f. (10-1=9), maka nilai untuk CI adalah d.f., α/2, yaitu $\mathbf{t_{9,0.025}}$ **= 2,26**

Nilai *Confidence interval* rendah diperoleh dengan rumus =

rata-rata selisih – (standar deviasi x nilai t $_{8,0.025}$ **/** $\sqrt{n}$ **),** dimana n adalah 10

yaitu: 0.33 – (0.45 x 2.26/SQRT(10)) = 0.01

Nilai *Confidence interval* tinggi diperoleh dengan rumus =

rata-rata selisih + (standar deviasi x nilai t $_{8,0.025}$ **/** $\sqrt{10}$ **),**

yaitu: 0.33 + (0.45 x 2.26/SQRT(10)) = 0.65

	T	CI low	CI high
95%		0,01	0,65

A.4.4. *Confidence intervals: Two means* (*Independent samples*)

Dengan data yang sama dengan materi, maka kita akan sedikit mengulang perhitungan ini. Pada contoh kali ini, kita akan mkitapatkan data yang sudah dioleh terlebih dahulu, yaitu rata-rata sampel dan standar eror. Kita dapat menggunakan rumus excel yang sudah dipelajari sebelumnya. Maka, secara singkat kita mendapatkan data yang siap dicari perbedaannya pata tingkat keyakinan 95%, kita akan menggunakan tabel z. alasan penggunakan tabel z dapat dilihat lagi pada materi. Misalkan dalam Excel kita letakan data teknik pada kolom C7 sampai C9, dan data manajemen pada kolom D7 sampai D9, sedangkan selisih ada kolom E7 sampai E9.

		Teknik (x)	Manajemen (Y)	Selisih (x-y)
A	**B**	**C**	**D**	**E**
7	*Size (*pengamatan*)*	100	70	?
8	*Sample mean*	58	65	-7
9	*Population standard error*	10	5	1.16

Perhitungan standar error selisih dapat dihitung dengan rumus excel

=SQRT((C9*C9/C7+D9*D9/D7)) = SQRT((10*10/100+5*5/70)).

95% CI, z0.025 = 1,96 (silahkan lihat cara penggunaan tabel z)

Nilai *Confidence interval* rendah diperoleh dengan rumus =

rata-rata sampel – (standar error x nilai $Z_{0.025}$)

yaitu: -7 – (1.16 x 1.96) = -9.28

Nilai *Confidence interval* tinggi diperoleh dengan rumus =

rata-rata sampel + (standar error x nilai $Z_{0.025}$)

yaitu: -7 + (1.16 x 1.96) = -4.72

Latihan

Untuk mempermudah latihan, file excel untuk latihan dapat di akses pada link berikut: https://bit.ly/Data_Latihan_Excel. Pilih pada folder yang sesuai dan file yang sesuai. Latihan diambil dari 365 Business Intelegence

Latihan 4.1. Varians populasi diketahui (nilai Z)

Dataset: data gaji 30 data analisi

$	117.313	$	112.276
$	104.002	$	108.637
$	113.038	$	96.818
$	101.936	$	92.307
$	84.560	$	114.564
$	113.136	$	109.714
$	80.740	$	108.833
$	100.536	$	115.295
$	105.052	$	89.279
$	87.201	$	81.720
$	91.986	$	89.344
$	94.868	$	114.426
$	90.745	$	90.410
$	102.848	$	95.118
$	85.927	$	113.382

Dari himpunan data di atas, Standar deviasi populasi diketahui sebesar $ 15.000.

Tugas:

1. Hitung rata-rata dan kesalahan standar
2. Temukan skor-z yang sesuai untuk menghitung interval kepercayaan 90%
3. Temukan interval kepercayaan 90%

Latihan 4.2. Varians populasi tidak diketahui (nilai t)

Dataset: data gaji 9 karyawan

$	78.000
$	90.000

$	75.000
$	117.000
$	105.000
$	96.000
$	89.500
$	102.300
$	80.000

Dari dataset tersebut

Tugas

1. Hitung rata-rata dan kesalahan standar kumpulan data
2. Menentukan statistik mana yang akan digunakan untuk inferensi
3. Temukan statistik yang sesuai, dengan mempertimbangkan derajat kebebasan (jika ada) untuk kepercayaan 99%
4. Temukan interval kepercayaan 99%

Latihan 4.3. *Confidence interval two means: Dependent samples*

Dataset: berat badan sebelum dan sesudah program

Subyek	Berat sebelum (kg)	Berat sesudah (kg)
1	103,68	92,87
2	110,68	101,58
3	119,05	105,66
4	101,75	96,18
5	91,69	86,97
6	112,03	105,90
7	88,84	80,56
8	105,18	97,00
9	110,37	99,27
10	120,99	107,44

Berikut adalah data program olahraga untuk menurunkan berat badan terdapat sampel 10 orang yang telah menyelesaikan program 12 minggu.

Tugas:

1. Hitung rata-rata dan standar deviasi kumpulan data
2. Tentukan statistik yang sesuai untuk digunakan
3. Hitung interval kepercayaan 95%
4. Menafsirkan hasilnya
5. Anda dapat mencoba menghitung interval kepercayaan 90% dan 99% untuk melihat perbedaannya. Tidak ada solusi yang disediakan untuk kasus-kasus ini.

Latihan 4.4. *Confidence interval two means: Independent samples*

Dataset: perbandingan nilai mahasiswa dua departemen

	Engineering	Management
Size	100	70
Sample mean	58	65
Population std	10	5

Tugas

1. Anda memiliki himpunan data dari pelajaran.
2. Hitung interval kepercayaan 99%
3. Bandingkan dengan interval kepercayaan 95% dari pelajaran

Latihan 4.5. *Confidence interval for difference of two means; independent samples, variances unknown but assumed to be equal*

Dataset: Perbandingan harga apel di dua kota berbeda

N		A	
$	3,80	$	3,02

$	3,76	$	3,22
$	3,87	$	3,24
$	3,99	$	3,02
$	4,02	$	3,06
$	4,25	$	3,15
$	4,13	$	3,81
$	3,98	$	3,44
$	3,99		
$	3,62		

Tugas:

1. Anda memiliki himpunan data yang sama dari pelajaran.
2. Hitung interval kepercayaan 90%
3. Bandingkan hasilnya dengan interval kepercayaan 95% dari pelajaran

BAB 5

UJI HIPOTESIS

Bab kelima dari buku ini, yang berjudul "Uji Hipotesis," dirancang untuk memperkenalkan salah satu aspek paling penting dalam statistik inferensial. Uji hipotesis adalah metode yang digunakan untuk membuat keputusan berdasarkan data sampel, dan merupakan alat yang sangat kuat dalam penelitian ilmiah atau pengambilan keputusan rasional.

Dalam bab ini, kita akan membahas konsep dasar dari uji hipotesis, termasuk formulasi hipotesis nol dan hipotesis alternatif, serta langkah-langkah dalam melakukan uji hipotesis. Pembaca akan diperkenalkan pada berbagai jenis uji hipotesis, seperti uji t dan uji z, serta bagaimana menentukan tingkat signifikansi dan p-value. Kita juga akan belajar menggunakan Microsoft Excel untuk melakukan uji hipotesis ini.

Pemahaman yang baik tentang uji hipotesis sangat penting karena mereka memungkinkan kita untuk membuat kesimpulan yang valid dan dapat diandalkan dari data yang kita miliki. Bab ini dapat memberikan pengetahuan yang diperlukan untuk melakukan analisis data yang lebih kompleks dan membuat keputusan yang lebih baik berdasarkan data. Pada materi ini, pemahaman terkait dengan materi sebelumnya menjadi sangat penting, antara lain terkait dengan *confidence interval* (interval keyakinan), serta pemahaman distribusi tabel Z dan T.

5.1. Pengertian Hipotesis

Dalam membuat keputusan, mungkin kita membutuhkan jawaban ya atau tidak. Pendekatan yang benar dalam hal ini adalah dengan menggunakan tes atau uji, seperti di bagian ini kita pelajari dalam topik ini, Dimana melakukan uji hipotesis merupakan salah satu tugas mendasar dan statistik pengujian hipotesis. Setikdanya terdapat empat langkah dalam pengambilan keputusan berbasis data [1], yaitu:

1. Merumuskan hipotesis,

2. Menemukan tes yang tepat untuk hipotesis tersebut.
3. Menjalankan tes, dan
4. Membuat keputusan berdasarkan hasilnya.

Mari kita mulai pembahasannya dengan penjelasan, apa itu hipotesis. Meskipun ada banyak cara untuk mendefinisikannya, secara mudah hipotesis dapat dipahami sebagai ide yang dapat diuji. Pengertian tersebut bukan definisi formal, tetapi kita dapat memahaminya sebagai berikut. Jika kita mendapatkan informasi bahwa apel di New York mahal, maka ini adalah ide atau pernyataan tetapi tidak dapat diuji sampai kita memiliki sesuatu (informasi atau data) untuk membandingkannya. Misalnya, jika saya mendefinisikan mahal sebagai harga yang lebih tinggi dari $ 1.75 per pon, maka itu segera menjadi hipotesis.

Bagaimana sesuatu tidak bisa menjadi hipotesis. Contohnya adalah, apakah Amerika Serikat akan lebih baik atau lebih buruk di bawah pemerintahan Presiden Clinton dibandingkan dengan pemerintahan Presiden Biden, yang saat ini sedang menjabat, secara statistik. Pernyataan tersebut adalah ide tetapi tidak ada data untuk mengujinya. Oleh karena itu tidak bisa menjadi hipotesis tes statistik. Sebenarnya, pernyataan tersebut lebih mungkin menjadi topik disiplin lain dalam statistik. Kita dapat membandingkan kepresidenan AS yang berbeda yang telah diselesaikan seperti pemerintahan Obama dan pemerintahan Bush karena kita memiliki data tentang keduanya.

Mari kita keluar dari politik dan masuk ke hipotesis, berikut topik sederhana yang bisa diuji, menurut Glassdoor yang merupakan situs web informasi gaji populer, gaji data scientist utama di AS adalah $ 113.000 dolar. Jadi jika kita ingin menguji perkiraan mereka benar, ada dua hipotesis yang dibuat. Hipotesis nol dilambangkan H0 dan hipotesis alternatif dilambangkan H1 atau HA. Hipotesis nol adalah yang akan diuji dan alternatifnya adalah yang lain. Dalam contoh kita, hipotesis nol adalah gaji ilmuwan

data rata-rata adalah $ 113.000, sedangkan alternatifnya adalah gaji ilmuwan data rata-rata bukan $ 113.000. Langkah selanjutnya adalah kita dapat memeriksa apakah $ 113.000 cukup dekat dengan rata-rata sebenarnya yang diprediksi oleh sampel yang diobservasi. Kita berasumsi bahwa gaji rata-rata adalah $ 113.000 dan kita mencoba membuktikan sebaliknya. Dalam hal ini, tes dapat disebut sebagai tes dua sisi atau dua ekor. Jika ditulis dalam notasi matematika adalah H0 : μ0 = 130.000, HA : μA ≠ 130.000

Kita juga dapat membentuk tes satu sisi atau satu ekor. Contohnya sebagai berikut: Paul memberi tahu kita bahwa dia pikir ilmuwan data menghasilkan lebih dari $ 125.000 per tahun, Karena kita meragukannya, maka kita merancang tes untuk melihat siapa yang benar, hipotesis nol (H0) dari tes ini adalah rata-rata gaji ilmuwan data lebih atau sama dengan $ 125.000. Sedangkan hipotesis alternatifnya (HA) adalah gaji ilmuwan data rata-rata kurang dari $ 125.000 dolar. Jika pernyataan tersebut ditulisa dalam notasi matematika maka H0 : μ0 => 125.000, dan HA : μA < 125.000. Dalam hal ini, penting untuk diketahui bahwa hasil tes mengacu pada parameter populasi dari pada statistik sampel.

Pada umumnya peneliti mencoba menolak hipotesis nol. Coba kita pikirkan tentang hipotesis nol sebagai *status quo* dan hipotesis alternatif sebagai perubahan atau inovasi yang menantang status quo. Dengan demikian hipotesis nol adalah pernyataan yang coba kita tolak. Oleh karena itu, hipotesis nol adalah keadaan saat ini, sedangkan alternatifnya adalah pendapat pribadi kita yang itu pasti berlawanan dengan kondisi pada awalnya.

5.2. Uji hipotesis dan tingkat signifikansi

Kita telah mempelajari apa itu hipotesis dan kita memiliki gagasan tentang bagaimana membentuk hipotesis nol dan alternatif. Selanjutnya kita perlu memahami alasan mengapa pengujian hipotesis berhasil. Pertama, kita harus mendefinisikan istilah

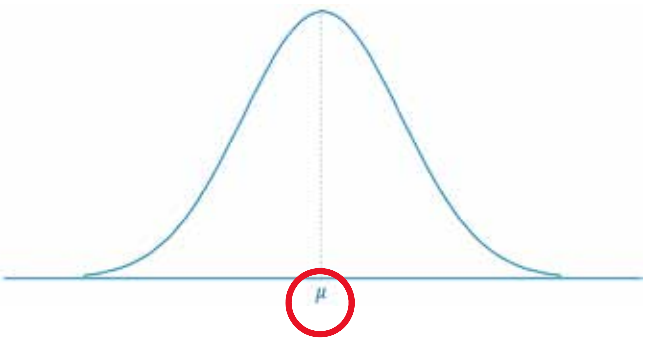
tingkat signifikansi. Biasanya uji hipotesis bertujuan untuk menolak hipotesis nol. Namun, seperti halnya tes apa pun, ada kemungkinan bahwa kita bisa salah dan menolak hipotesis nol yang benar.

Tingkat signifikansi dilambangkan dengan *alpha* (α) dan merupakan probabilitas menolak hipotesis nol, jika itu benar. Jadi probabilitas membuat kesalahan ini nilai tipikal untuk Alpha adalah 0.01 (1%), 0.05 (5%), dan 0.1 (10%). Nilai tersebut adalah nilai yang ditentukan berdasarkan kepastian yang dibutuhkan sesuai dengan kasusnya. Pilihan tingkat Alpha ditentukan oleh konteks tempat kita melakukan observasi beroperasi, tetapi 0.05 adalah nilai yang paling umum digunakan.

Contohnya, misalnya kita perlu menguji mesin bekerja dengan baik, maka kita akan mengharapkan hasil tes yang dilakukan akan membuat sedikit atau tidak ada kesalahan karena kita ingin hasilnya sangat tepat. Oleh karena itu kita harus memilih tingkat signifikansi rendah seperti 0.01. Botol kaca Coca-Cola yang terkenal adalah 12 ons. Jika mesin menuangkan 12,1 ons maka sebagian cairan akan tumpah dan label juga akan rusak. Jadi dalam situasi tertentu kita harus seakurat mungkin, namun jika kita menganalisis manusia atau perusahaan, kita akan mengharapkan perilaku yang lebih acak atau setidaknya tidak pasti dan karenanya tingkat kesalahan yang lebih tinggi. Di sisi lain, jika kita ingin memprediksi berapa banyak Coca-Cola yang diminum rata-rata konsumennya, perbedaan antara 12 ons dan 12,1 ons tidak akan terlalu penting. Jadi kita bisa memilih tingkat signifikansi yang lebih tinggi seperti 0.05 atau 0.1.

Sekarang setelah kita memiliki gambaran tentang tingkat signifikansi, mari kita masuk ke mekanisme pengujian hipotesis, bayangkan jika kita sedang berkonsultasi dengan universitas dan ingin melakukan analisis tentang bagaimana kinerja siswa rata-rata, Dekan Universitas percaya bahwa rata-rata siswa memiliki IPK 70 persen. Sebagai peneliti berbasis data, kita tidak bisa begitu saja setuju dengan pendapatnya.

Jadi kita mulai menguji hipotesis nol adalah nilai rata-rata populasi adalah 70 persen. Nilai ini adalah nilai hipotesis dan kita menunjukkannya dengan hipotesis nol, sedangkan hipotesis alternatif adalah nilai rata-rata populasi bukan 70 persen. Dengan asumsi bahwa populasi nilai didistribusikan secara normal, semua nilai yang diterima oleh siswa harus terlihat seperti ini.



Dimana nilai μ adalah rata-rata populasi yang sebenarnya. Uji yang bisa kita lakukan adalah uji z, karena populasinya diketahui. Rumus tes Z adalah sebagai berikut:

$$z = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

Keterangan:

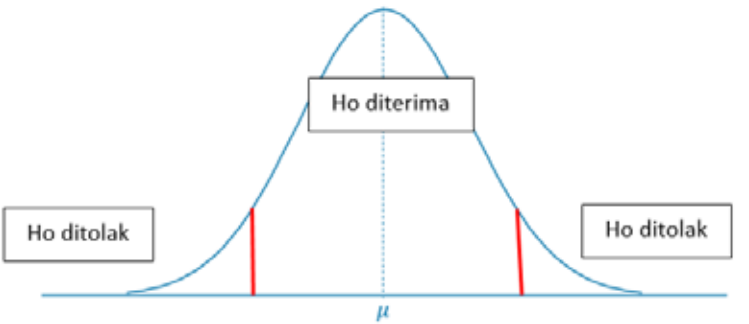
$\bar{x}$ = rata-rata sampel (sample mean)

μ = rata-rata hipotesis

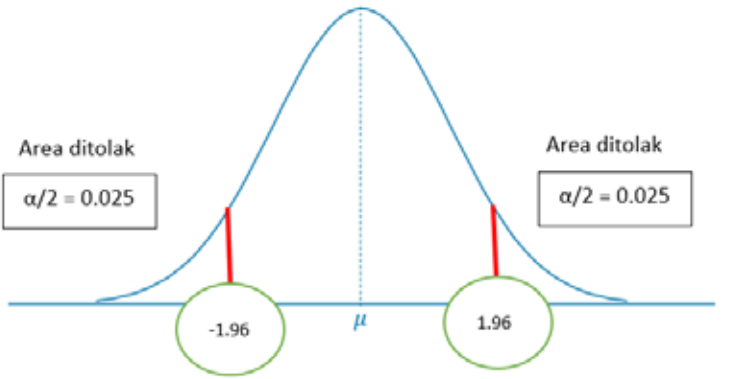
S/√n = rumus standar error (S = standar deviasi, n = jumlah sampel)

Ide dari rumus tersebut adalah kita menstandarisasi atau menskalakan rata-rata sampel yang kita dapatkan. Jika rata-rata sampel cukup dekat dengan rata-rata yang dihipotesiskan, maka Z akan mendekati nol, dan sebaliknya. Secara alami, jika rata-rata sampel persis sama dengan rata-rata yang dihipotesiskan, Z akan menjadi nol dalam semua kasus ini, kita akan menerima hipotesis nol.

Pertanyaannya di sini adalah sebagai berikut, seberapa besar nilai Z untuk menolak hipotesis nol? Untuk itu ada garis batas karena kita melakukan tes dua sisi atau dua ekor. Ada dua garis *cutoff* satu di setiap sisi.

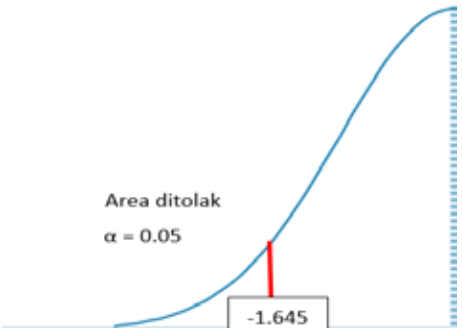


Ketika kita menghitung Z, kita akan mendapatkan nilai. Jika nilai ini berada pada bagian dalam garis merah maka kita menerima hipotesis nol. Jika berada di luar wilayah yang digaris maka kita dapat menolak hipotesis nol. Itulah sebabnya bagian terluar dari kurva tersebut adalah daerah penolakan. Area yang tersebut sebenarnya tergantung pada tingkat signifikansi atau tingkat *alpha* yang ditentukan di awal. Tingkat signifikansi atau *alpha* adalah 0.05 (5%), maka kemudian kita memiliki *alpha* dibagi dengan dua atau titik 0.025 di sisi kiri dan titik 0.025 di sisi kanan.



Nilai-nilai tersebut adalah nilai-nilai yang dapat kita periksa dari tabel z statistik dengan *alpha* 0.05. Maka kita dapatkan nilai z adalah 1.96, dengan pembagian 1.96 di sisi kanan dan -1.96 di sisi kiri. Oleh karena itu, jika nilai yang kita dapatkan dari tes lebih rendah dari -1.96 atau lebih tinggi dari 1.96, maka kita akan menolak hipotesis nol, yang berarti kita menerima hipotesis alternatif. Demikian, kurang lebih bagaimana pengujian hipotesis bekerja, kita membuat skala rata-rata sampel sehubungan dengan nilai hipotesis, jika nilai Z *score* mendekati nol maka kita tidak dapat menolak hipotesis nol, jika jauh dari nol maka kita menolak hipotesis nol.

Selanjutnya bagaimana dengan tes satu sisi? Kita dapat mengambil contoh yang sudah kita gunakan sebelumnya, Paul mengatakan bahwa ilmuwan data menghasilkan lebih dari $125.000 dolar, jadi hipotesis nol (H0) dari tes ini adalah rata-rata gaji ilmuwan data lebih atau sama dengan $ 125.000. Sedangkan hipotesis alternatifnya (HA) adalah gaji ilmuwan data rata-rata kurang dari $ 125.000 dolar. Kali ini seluruh wilayah penolakan ada di sebelah kiri. Jadi wilayah penolakan memiliki area *alpha*, melihat tabel z yang sesuai dengan skor z, yaitu 1.645 dan karena berada di sebelah kiri dengan tanda minus.



Sekarang kita menghitung statistik tes z, jika kita mendapatkan nilai yang lebih rendah dari -1.645, kita akan menolak hipotesis nol karena kita memiliki bukti bahwa gaji ilmuwan data

kurang dari $125.000, jika tidak, kita akan menerima Ho. Mari kita jelajahi tes satu ekor lainnya, katakanlah Dekan universitas memberi tahu kita bahwa rata-rata IPK siswa lebih rendah dari 70%. Dalam hal ini hipotesis nol adalah IPK lebih rendah atau sama dengan 70% sedangkan hipotesis alternatif adalah lebih besar dari 70%. Dalam situasi ini wilayah penolakan berada di sisi kanan atau positif, jika nilai tes lebih besar dari skor Z *cutoff*.

5.3. Error tipe I dan error tipe II

Pada uji ini kita akan belajar tentang kesalahan yang dapat dilakukan dalam pengujian hipotesis. Secara umum, kita dapat memiliki dua jenis kesalahan, kesalahan tipe 1 dan tipe 2. Kesalahan tipe 1 dapat terjadi saat kita menolak hipotesis nol yang sebenarnya. Sedangkan kesalahan tipe 2, kita menerima hipotesis nol yang salah. Kesalahan ini dapat terjadi pada saat kita menentukan tingkat signifikansi atau *alpha* pada tingkat tertentu. Tanggung jawab untuk membuat kesalahan ini semata-mata terletak pada kita yang menentukan tingkat kesalahan atau *alpha*. Yang pasti, kita tidak benar-benar ingin membuat salah satu dari dua kesalahan tersebut, tetapi hal itu kadang-kadang terjadi. Kita juga harus menyadari bahwa statistik sangat berguna tetapi tidak sempurna.

5.4. Uji Mean (populasi variance diketahui)

Setelah kita membahas teori yang diperlukan, sekarang saatnya untuk beberapa pengujian. Kita akan mengeksplorasi dua jenis tes (1) yang diambil dari satu populasi dan (2) diambil dari beberapa populasi. Tes ini sangat mirip dengan interval kepercayaan untuk satu populasi dan interval kepercayaan untuk dua populasi yang kita bahas sebelumnya (bab 4).

Kita akan menjalankan tes untuk varian yang diketahui dan varian yang tidak diketahui. Mari kita mulai dengan tes di mana varians diketahui untuk pengujian ini. Kita akan menggunakan

kembali contoh gaji ilmuwan data. Berikut kumpulan datanya untuk 30 sampel:

Tabel 5.1. Contoh data gaji ilmuan data

$ 117.313	$ 91.986	$ 109.714
$ 104.002	$ 94.868	$ 108.833
$ 113.038	$ 90.745	$ 115.295
$ 101.936	$ 102.848	$ 89.279
$ 84.560	$ 85.927	$ 81.720
$ 113.136	$ 112.276	$ 89.344
$ 80.740	$ 108.637	$ 114.426
$ 100.536	$ 96.818	$ 90.410
$ 105.052	$ 92.307	$ 95.118
$ 87.201	$ 114.564	$ 113.382

Dari data tersebut kita dapat menghitung:

Rata-rata sampel tersebut $ 100.200.

Varians populasi diketahui dan standar deviasinya $ 15.000,

sedangkan standar error $ 2.739.

Namun kita melihat bahwa menurut *situs web glassdoor* yang merupakan informasi gaji popular. Gaji rata-rata ilmuwan data adalah $130.000. Dalam kasus ini kita ingin melihat apakah nilai tersebut benar. Kita membutuhkan tes dua sisi karena kita ingin untuk mengetahui bahwa gajinya secara signifikan lebih rendah dari $ 130.000 atau secara signifikan lebih dari $ 130.000. Hipotesis nol adalah gaji rata-rata populasi adalah $ 130.000, maka H0 : $\mu 0 = \$$ 130.000. Hipotesis alternatifnya adalah bahwa populasi berarti gaji berbeda dari $130.000, maka Ha : $\mu 0 \neq \$$ 130.000.

Pengujian dilakukan dengan menstandarisasi variabel yang ada dan membandingkannya dengan nilai Z yang mengikuti distribusi normal standar. Kita melakukan standarisasi variabel dengan mengurangi rata-rata dan membagi dengan standar deviasi.

Dengan data yang digunakan adalah sampel, maka kita akan menggunakan *standard error*. Dengan demikian rumus standarisasi menjadi huruf Z kapital (Z-*score*), yaitu sama dengan rata-rata sampel dikurangi nilai dari hipotesis nol dibagi dengan standard error.

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \longrightarrow \textit{Standard error}$$

Dengan cara ini kita mendapatkan distribusi dengan rata-rata nol dan standar deviasi 1. Nilai Z besar ini tidak boleh disalahartikan dengan huruf z kecil. **Huruf Z (besar) adalah variabel yang terkait dengan tes, dan akan disebut skor Z atau Z-score. Sedangkan, huruf z (kecil) adalah nilai yang berasal dari tabel (nilai tabel) yang telah kita bicarakan sebelumnya dan selanjutnya akan disebut sebagai nilai kritis.**

Bagaimana kita akan melakukan pengujian, nilai z (kecil) yang didistribusikan secara normal dengan rata-rata dan standar deviasi 1. Sedangkan huruf Z (besar) didistribusikan secara normal dengan rata-rata $\bar{x}$ (dibaca X bar) dikurangi μ_0. Jika selisih antara $\bar{x}$ (X bar) dan μ_0 semakin mendekati nol, maka semakin dekat skor Z tersebut dengan nol. Kondisi ini menunjukkan peluang yang lebih tinggi untuk menerima hipotesis nol. Mari kita kembali ke contoh. Kita memasukkan angka-angka yang kita miliki dari contoh. Hasilnya sebagai berikut:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{100200 - 113000}{2739} = -4.67$$

Apa yang kita dapatkan adalah skor Z adalah -4.67. Sekarang kita akan membandingkan nilai absolut -4.67 dengan z (huruf kecil) dari *alpha* dibagi 2 ($z_{\alpha/2}$) di mana *alpha* adalah tingkat signifikansi. Kita dapat menggunakan nilai absolut karena jauh lebih mudah

untuk selalu membandingkan Z (huruf besar) positif dengan z (huruf kecil) positif. Selain itu, nilai tabel z tidak terdapat nilai negatif. Kita harus mengerti bahwa dua kondisi, yaitu bahwa: –4.67 < nilai z negatif, memiliki arti sama dengan: 4.67 > nilai z positif. Dengan demikian, kita dapat menggunakan nilai absolut dari skor Z (Z-*score*) harus lebih tinggi dari nilai absolut dari nilai z kritis (z tabel). (keputusan adalah hipotesis 0 ditolak jika nilai absolut Z-*score* jika lebih besar dari nilai kritis atau nilai tabel (nilai z)).

Pada tingkat signifikansi 5%, atau *alpha* adalah 0,05, maka uji dua sisi, kita akan memeriksa tabel z pada baris 0.025 ($z_{0.025}$), nilai yang sesuai adalah 1.96. Selanjutnya yang kita lakukan adalah membandingkan nilai hitung kita dengan nilai kritis/tabel. Jika nilai Z (Z-*score*) lebih tinggi dari 1,96, maka kita akan menolak hipotesis nol, namun jika lebih rendah maka kita akan menerimanya. Hasilnya adalah 4.67 lebih tinggi dari 1.96 (Z (besar) = 4.67 > z (kecil) = 1.96), sehingga kita menolak hipotesis nol. Dengan demikian, pada tingkat signifikansi 5% kita telah menolak hipotesis nol atau pada signifikansi 5 persen. Hasil ini dapat dimaknai bahwa tidak ada bukti statistik bahwa gaji rata-rata ilmuan data adalah $ 130.000.

Z-score > z table = Hipotesis nol ditolak

Z-score < z table = Hipotesis nol diterima

Bagaimana jika kita memiliki tingkat signifikansi yang berbeda, yaitu menggunakan signifikansi 1%. Dengan demikian, kita memiliki *alpha* 0.01 dengan *alpha* dibagi 2 maka nilai $z_{0.005}$ = 2,58. Sekali lagi skor Z atau Z-score yang diperoleh yaitu 4,67 lebih tinggi dari nilai kritis atau nilai z tabel yaitu 2,58. Dengan demikian kita akan menolak hipotesis nol bahkan pada signifikansi 1%.

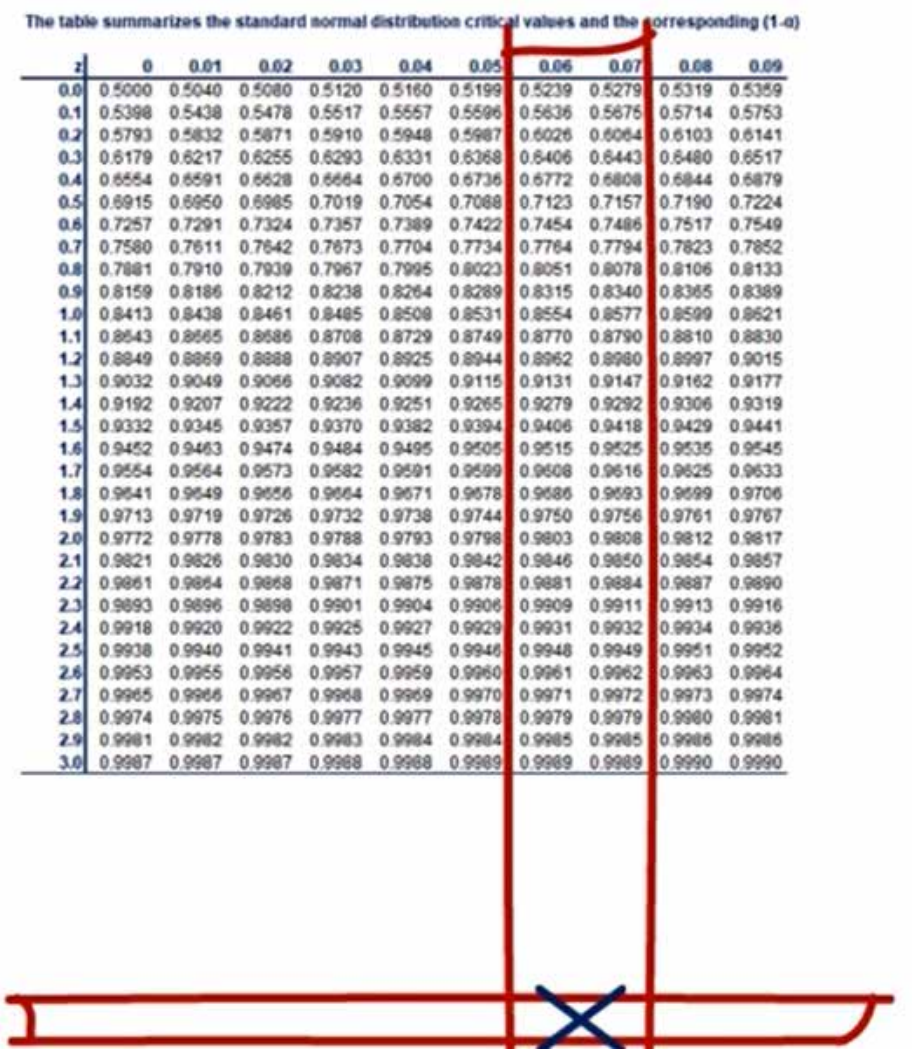
5.5. Nilai-P (P-*value*)

Kita telah belajar bagaimana menguji hipotesis dan bagaimana menolaknya. Poin yang perlu kita pahami adalah kita

menolak hipotesis nol pada berbagai tingkat signifikansi, tetapi kita tidak dapat menemukan tingkat signifikansi (karena keterbatasan tabel). Oleh karena itu ini adalah waktu untuk belajar ukuran yang disebut nilai-P. Cara ini adalah cara paling umum untuk menguji hipotesis seperti menguji hipotesis pada tingkat signifikansi yang telah dipelajari sebelumnya. Nilai-P atau P-*value* merupakan tingkat signifikansi terkecil namun kita masih dapat menolak hipotesis nol karena sampel yang diamati.

Dengan contoh data yang sama, kita menggunakan tes pada gaji ilmuwan data. Berikut adalah deskripsinya informasi yang sudah diketahui. Kita memiliki kesalahan standar (*standard error*) adalah 2.739, sedangkan standar deviasi populasi yang diketahui adalah 15.000, populasi yang terdistribusi normal, ukuran sampel 30, dan skor Z yang sesuai adalah -4,67.

Dari contoh kasus di atas, kita dapat menolak hipotesis nol pada tingkat signifikan 0.05 dan 0.01, tetapi kita coba lihat nilai tersebut dalam tabel. Kita dapat memeriksa tabel z untuk 4.67 yang memberi kita hasil yang sama dengan -4.67. Permasalahannya adalah di sebagian besar tabel statistik, kita tidak akan menemukan nilai ini karena nilainya sangat besar (lihat ilustrasi pada gambar 4.1).



The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Gambar 4.1. Ilustrasi bahwa nilai diluar tabel statistik.

Dengan demikian, kita dapat membulatkan ke nilai terdekat yang tersedia dan mendapatkan titik P-*value* sebesar 0.001 (**P-*value* = 1 – (nilai tabel),** yaitu: 1 - 0.999 = 0.001).

Sekarang bagaimana jika kita benar-benar menguji hipotesis. Setelah menentukan tingkat signifikan alpha, maka kita dapat membandingkan nilai-P dengan nilai alpha. Kita harus menolak

hipotesis nol (menerima hipotesis alternaitf) jika nilai P-*value* lebih rendah dari tingkat signifikansi alpha.

P-value > α, maka Ho diterima

P-value < α, maka Ho ditolak

Dari ilustrasi contoh sebelumnya, berikut ini adalah kondisi yang menunjukkan hipotesis nol (H0) ditolak

1. Uji tingkat signifikansi 90%: 0.0001 < 0.1
2. Uji tingkat signifikansi 95%: 0.0001 < 0.05
3. Uji tingkat signifikansi 99%: 0.0001 < 0.01

Oleh karena itu kita dapat dengan aman mengatakan bahwa hasil tersebut sangat signifikan pada tingkat pengukuran signifikansi apa pun.

Mari kita lihat contoh lain, jika kita memiliki skor Z (Z-*score*) adalah sebesar 2.12, kita akan menerima hipotesis nol pada alpha 5% dan 10%, tetapi menolaknya pada signifikansi 1%. Mengapa demikian? Sekarang kita dapat melihat tabel dan kemudian menemukan nilai-P yang kita cari nilai yang sesuai dengan nilai 2.12 dan menemukan bahwa itu adalah 0,983. Nilai-P dapat dihitung secara manual untuk uji satu sisi adalah dengan rumus berikut:

P-*value* = 1 – nilai dari tabel

Dengan rumus tersebut, maka nilai-P adalah 1 – 0.983 yaitu 0.017. Untuk menghitung nilai-P pada uji dua sisi, dapat dicari dengan angka yang kita lihat dalam tabel dikalikan dengan dua.

P-*value* = (1 – nilai dari tabel) x 2

Dengan rumus tersebut maka nilai-P adalah (1 – 0.983) x 2, hasilnya adalah 0,034. Hasil ini juga merupakan jawaban atas pertanyaan contoh di atas.

Jadi nilai-P merupakan nilai yang digunakan sebagian besar paket perangkat lunak statistik, untuk menjalankan pengujian dan kemudian memberi kita serangkaian hasil. Kemudian terserah peneliti untuk memutuskan apakah variabel tersebut signifikan secara statistik atau tidak. Semakin dekat dengan 0, maka nilai-P akan semakin signifikan dengan hasil yang diperoleh. Pertimbangan terakhir adalah bahwa nilai-P adalah ukuran yang sangat kuat karena berfungsi untuk semua distribusi. Tidak peduli apakah kita berurusan dengan distribusi siswa normal, distribusi t, distribusi binomial atau seragam. Apa pun uji hipotesis yang dilakukan, alasan penggunaan nilai-P dapat digunakan, yaitu jika nilai-P lebih rendah dari tingkat signifikansi, maka kita menolak hipotesis nol.

Cara menghitung nilai-P secara online akan disampaikan lebih lanjut pada bagian Appendix. Hal ini akan membantu kita bekerja lebih mudah.

5.6. Uji Rata-rata: Varians Populasi Tidak Diketahui

Setelah kita tahu apa itu nilai-P dan bagaimana menggunakannya, sekarang kita akan mempelajari bersama kasus yang lebih sering terjadi yaitu varians populasi yang tidak diketahui. Contoh, misalkan kita adalah analis pemasaran sebuah perusahaan maka kita diminta untuk memperkirakan tingkat keterbukaan email dari salah satu pesaing perusahaan di atas perusahaan tempat kita bekerja. Tingkat keterbukaan email adalah ukuran berapa banyak orang di daftar email yang benar-benar membuka email yang mereka terima pada awalnya. Perusahaan kita memiliki tingkat keterbukaan 40%. Dalam hal ini, kita berusaha untuk mencari tahu bagaimana mendapatkan informasi spesifik tentang perusahaan pesaing. Mari kita nyatakan hipotesis dari kondisi tersebut: hipotesis nol, rata-rata tingkat keterbukaan lebih rendah atau sama dengan 40% ($Ho:\mu_{or} \leq 40$), hipotesis alternatif rata-rata tingkat keterbukaan lebih tinggi dari 40% ($Ha:\mu_{or} > 40$).

Perhatikan bahwa dalam pengujian hipotesis kita bertujuan untuk menolak hipotesis nol ketika kita ingin menguji apakah tingkat terbuka lebih tinggi dari 40%. Rumusan hipotesis tersebut juga menunjukkan bahwa pengujian hipotesis adalah satu sisi. Dalam kasus ini kita dapat menggunakan tingkat alpha 5% atau 0.05.

Tabel 5.1. Contoh data tingkat keterbukaan email

26%	23%	42%	49%	23%	59%	29%	29%	57%	40%

Dari data tersebut maka dapat dicari:

Rata-rata sampel (*Sample mean*)	37.70%
Standar deviasi sampel (*Sample standard dev)*	13.74%
Standard error	4.34%

Hasil perhitungan nilai tersebut akan ditunjukkan pada bagian appendix. Kita dapat berasumsi bahwa populasi tingkat keterbukaan email tersebut didistribusikan secara normal seperti interval kepercayaan dengan varians tidak diketahui dan sampel kecil. Oleh karena itu, statistik yang benar untuk digunakan dalam kasus ini adalah statistik t (t-statistic), jika kita tidak tahu varians dan sampelnya tidak cukup besar. Artinya, variabel mengikuti distribusi t-*student*, sehingga kita harus menggunakan statistik t.

Kita menghitung nilai T (T-*score*) dengan cara yang sama seperti kita menghitung skor Z. Nilai T dapat dicari dengan rumus:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Keterangan dari rumus tersebut yaitu: $\bar{x}$ adalah rata-rata sampel, μ adalah nilai rata-rata yang dihipotesiskan, sedangkan $s / \sqrt{n}$ adalah rumus untuk *standard error* (s adalah standard deviasi, dan n adalah jumlah sampel). Hasilnya adalah:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{37.70\% - 40\%}{4.34\%} = -0.53$$

Hasilnya adalah -0.53. Tentu kita akan lebih mudah untuk bekerja dengan bilangan positif. Jadi kita harus membandingkan nilai absolut -0.53 dengan nilai t-tabel, yang dicari dengan cara *degree of freedom* (df) = n - 1 = 9 (sampel (n) dikurangi satu). Dengan nilai signifikansi 0.05 satu sisi, maka nilai t tabel adalah: 1.833. Kesimpulan yang diperoleh:

T t
0.53 < 1.83

Ingat aturan keputusan, jika nilai absolut skor T lebih rendah dari statistik dari tabel, kita tidak dapat menolak hipotesis nol. Oleh karena itu kita harus menerima hipotesis nol.

T-score > t table = Hipotesis nol ditolak

T-score < t table = Hipotesis nol diterima

Dengan kembali pada kasus yang diangkat, maka kita dapat memberi tahu pimpinan perusahaan, bahwa, pada tingkat signifikansi ini, secara statistik kita tidak dapat mengatakan bahwa tingkat keterbukaan email pesaing kita lebih tinggi dari 40%.

Bagaimana jika kita menggunakan nilai-P. Nilai-P dari statistik ini adalah 0.304, karena nilai-P lebih besar dari tingkat signifikansi 0.05, maka kita sampai pada kesimpulan yang sama. Kita menerima hipotesis nol (nilai-P = 0.34 > α = 0.05, Ho diterima). Jika kita mengganti tingkat signifikansi menjadi 0.01, maka nilai-P masih akan lebih tinggi, sehingga kita menerima hipotesis nol.

5.7. Uji Rata-rata: Sampel Dependen

Kita dapat mengingat kembali pada materi sebelumnya, populasi tunggal dan multi populasi. Kini kita mulai dengan interval

keyakinan untuk rata-rata populasi tunggal dan kemudian beralih ke contoh yang mempertimbangkan beberapa populasi.

Kali ini, kita juga akan menilai populasi tunggal sebagai permulaan. Kita akan mulai dengan sampel dependen. Contoh sampel dependen yang pernah kita pakai sebelumnya, seperti penurunan berat badan dan tes darah, dimana sampel diambil dari data penurunan berat badan atau data konsentrasi nutrisi dengan subjek yang diminati adalah orang yang sama sebelum dan sesudahnya.

Kita akan mengingat kembali dengan contoh pada meteri sebelumnya, yaitu tentang kadar magnesium dalam darah seseorang. Ada perusahaan obat yang mengembangkan pil baru yang seharusnya meningkatkan kadar penerima magnesium. Ada 10 orang yang terlibat dalam penelitian yang mengonsumsi obat tersebut selama beberapa waktu dan kita menghitung interval keyakinan yang membantu mempelajari efek obat tersebut. Mereka menunjukkan kisaran nilai yang masuk akal untuk rata-rata populasi saat ini. Kita ingin sampai pada satu kesimpulan pasti tentang efektivitas obat tersebut.

Selanjutnya, mari kita nyatakan hipotesis nol (Ho) dari kasus tersebut, bahwa rata-rata populasi sebelum lebih besar atau sama dengan rata-rata populasi sesudahnya (Ho : $\mu_b \geq \mu_a$). Sedangkan hipotesis alternatifnya (Ha) adalah rata-rata populasi sebelum lebih rendah daripada rata-rata populasi sesudahnya (Ho : $\mu_b < \mu_a$). Kita ingin melihat apakah obat bekerja efektif atau tidak. Kita ingin tahu apakah kadar magnesium lebih tinggi, kita buat hipotesis nol dan alternatif sedemikian rupa sehingga kita bertujuan untuk menolak hipotesis nol.

Kita perkirakan kadarnya akan lebih tinggi, jadi ketika hipotesis nol yang kita nyatakan akan lebih rendah atau sama, Selanjutnya, mari kita susun ulang rumusan hipotesis nol (Ho) sebagai berikut: Ho : $\mu_b - \mu_a \geq 0$. Rumusan tersebut bisa kita sesuai

menjadi Ho : Do ≥ 0, sedangkan hipotesis alternatif dapat dilakukan sebagai berikut: Ha : Do < 0. Kita dapat menggantinya dengan huruf kapital D nol, ini adalah perbedaan rata-rata populasi yang dihipotesiskan.

Berikut adalah contoh kumpulan data. Kita memiliki 10 pengamatan yang telah dicatat orang sebelum dan sesudahnya secara alami seperti pada tabel berikut.

Sebelum	Sesudah	Selisih(B - A)
2	1,7	0,3
1,4	1,7	-0,3
1,3	1,8	-0,5
1,1	1,3	-0,2
1,8	1,7	0,1
1,6	1,5	0,1
1,5	1,6	-0,1
0,7	1,7	-1
0,9	1,7	-0,8
1,5	2,4	-0,9

Dari data tersebut maka kita dapat menghitung:

Sample mean	-0,33
Standard deviation	0,45
Standard error	0,14
T-*score*	-2,29
p-*value*	0,024

Berikut adalah penjelasannya. Statistik yang tepat untuk digunakan di sini adalah statistik t, karena (1) kita memiliki sampel kecil, (2) kita mengasumsikan distribusi normal populasi, dan (3) kita tidak mengetahui variansnya sehingga skor T sama dengan

ekspresi berikut.

$$T = \frac{\bar{d} - \mu_0}{St.error} = \frac{-0.33 - 0}{0.14} = -2.29$$

Sekarang kita dapat melakukan perhitungan dan menemukan bahwa nilainya adalah –2.29 karena kita tidak ingin memilih tingkat signifikansi. Jika kita lihat pada tabel adalah sebagai berikut:

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf	1.282	1.645	1.960	2.326	2.576

Kita dapat masuk ke tabel dan melihatnya berada di antara 0.01 dan 0.025. Karena hasilnya tidak ada yang tepat, maka hal ini akan membuat keputusan tidak mudah. Mari kita selesaikan masalah ini dengan nilai-P untuk menemukan nilai-P dari uji satu sisi ini. Seperti yang kita pelajari sebelumnya menggunakan perangkat lunak jauh lebih mudah. Jadi setelah menggunakan kalkulator nilai-P online. Kita dapat mengetahui bahwa nilainya tepat 0.024.

Aturannya adalah jika nilai-P lebih besar dari tingkat alpha, maka Ho diterima, sedangkan nilai-P lebih kecil dari tingkat alpha maka Ha diterima. Jadi jika tingkat signifikansinya adalah 0.05 dan nilai-P lebih rendah, maka kita akan dapat menolak hipotesis nol (Ho) pada tingkat signifikan 5% ($0.024 < 0.05$). Jika tingkat signifikansinya adalah 0.01, dengan nilai-P lebih tinggi, maka kita menerima hipotesis nol (Ho) pada tingkat signifikansi 1% ($0.024 < 0.01$).

Dalam hal ini, terserah peneliti untuk memilih tingkat signifikansi dalam kasus pil magnesium. Kita berharap peneliti akan sangat berhati-hati karena ia ingin tahu apakah ini pil yang efektif yang benar-benar dapat membantu orang. Jika kita tidak dapat mengatakan bahwa pil tersebut bekerja pada tingkat signifikansi 1%, mungkin lebih baik membawanya kembali ke laboratorium. Alternatifnya adalah menguji lagi dan meningkatkan ukuran sampel untuk hasil yang lebih baik. Sampel sebanyak 100 orang akan meningkatkan tingkat presisi secara signifikan. Selanjutnya, jadi kita telah melakukan beberapa pengujian hipotesis lagi dan telah mengeksplorasi beberapa faktor yang membantu Anda menentukan tingkat signifikansi pengujian.

5.8. Uji rata-rata: Sampel independen

5.8.1. Sampel independent dengan varian diketahui

Saat ini kita akan belajar tentang sampel independen dengan varian yang diketahui. Kita akan menggunakan contoh dan data pada topik sampel independent yang sudah kita pelajari sebelumnya. Pada contoh kali ini, kita akan menguji nilai rata-rata siswa dari dua departemen berbeda di universitas Inggris. Kita perlu mengetahui dahulu bahwa di Inggris nilai dinyatakan dalam persentase. Kita akan membandingkan dua departemen di suatu universitas, yaitu teknik dan manajemen. Sebagai ilustrasi cerita, kita mendapatkan informasi oleh Dekan bahwa teknik adalah disiplin yang lebih keras dan orang cenderung mendapatkan nilai yang lebih rendah. Dia percaya bahwa rata-rata mahasiswa manajemen mengungguli mahasiswa teknik dengan empat poin persentase.

Sekarang tugas kita untuk memverifikasi pernyataan tersebut. Dalam kasus ini kita dapat menyatakan dua hipotesis nol, yang pertama adalah perbedaan antara rata-rata kedua populasi sama dengan negatif 4%, yang dapat kita sampaikan dalam bentuk persamaan Ho : μ_e-μ_m = -4%. Kedua adalah dinyatakan dalam bentuk nilai positif, yaitu Ho : μ_e-μ_m = 4%. Bagaimana dengan hipotesis alternatif. Kita juga dapat menggunakan nilai negatif atau positif, yaitu Ha : μ_e-$\mu_m \neq$ -4% atau Ha : μ_e-$\mu_m \neq$ 4%.

Kedua cara tersebut dapat digunakan, maka pertanyaan penelitian yang akan kita jawab bukan untuk menemukan perbedaannya tetapi untuk memeriksa apakah pernyataan tersebut benar. Kita perlu mengingat bahwa pengujian yang dilakukan adalah untuk dua sisi. Berikut adalah data yang akan kita uji.

	Engineering(e)	Management(m)	Difference
Sample	100	70	?
Mean	58%	65%	-7,00%
Population std	10%	6%	1,23%

Jika Anda ingat ketika populasi diketahui untuk sampel independen, maka perbedaan atau selisih *standard error* adalah sama dengan akar kuadrat dari jumlah varians teknik dibagi dengan ukuran sampel dan varians manajemen. Berikut adalah rumus untuk mencari perbedaan nilai standar deviasi populasi yang hasilnya 1,23%.

$$\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}$$

Karena kita memiliki sampel besar dan varians yang diketahui, maka kita dapat menggunakan statistik Z. Berikut ini akan membantu kita dalam menentukan uji satistik:

1. Jika Sampel kecil dan varians yang tidak diketahui dapat menggunakan uji statistik t,
2. Jika sampel besar dan varians yang diketahui berarti menggunakan uji statistik z.
3. Namun bagaimana jika, ketika kita memiliki sampel besar dan varians yang tidak diketahui. Dalam hal ini peneliti bebas memilih, tetapi umumnya tidak apa-apa untuk menggunakan uji z dalam kasus itu juga.

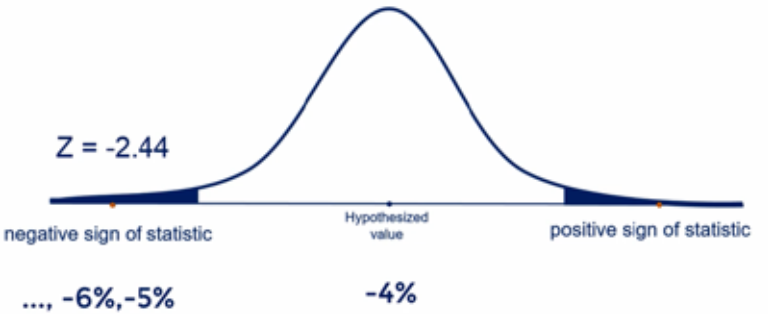
Berikut adalah rumus untuk uji rata-rata perbedaan sampel statistik dikurangi rata-rata perbedaan yang dihipotesiskan dibagi dengan kesalahan standar.

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m}}} = \frac{(-7\%) - (-4\%)}{1.23\%} = -2.44$$

Kita dapat memasukkan angka-angka yang ada dan mendapatkan skor z adalah -2.44. Untuk menghitung nilai p dengan menggunakan perangkat lunak (yang ada pada Appendix), dan hasilnya nilai-P dari uji dua sisi adalah 0,015.

Apa yang dapat kita simpulkan adalah bahwa pada signifikansi 5 persen yang umum untuk penelitian, nilai-P adalah 0,015 lebih rendah dari 0,05. Dengan demikian kita dapat menolak hipotesis nol (menerima hipotesis alternatif). Dengan demikian, ada cukup bukti statistik bahwa perbedaan kedua rata-rata tidak 4 persen.

Bagaimana jika kita ingin mengetahui apakah perbedaannya lebih tinggi atau lebih rendah dari pada tes yang diuji. Berikut ini mungkin bisa membantu untuk memahaminya. Karena kita menggunakan nilai z yaitu -2.44, maka nilai mungkin dapat lebih rendah dari -4%. Berikut adalah ilustrasinya.



Dalam kasus ini dapat diterjemahkan ke dalam rata-rata sebenarnya kemungkinan lebih rendah dari -4%, mungkin adalah -5, -6 dan seterusnya.

5.8.2. Uji rata-rata: Sampel independent dengan varian tidak diketahui diasumsikan sama

Kasus terakhir yang akan kita pelajari adalah kasus dengan sampel independen dan varians yang tidak diketahui namun diasumsikan sama. Kita akan kembali menggunakan contoh kasus yang sudah kita gunakan sebelumnya, yaitu perbandingan harga apel di New York dan Los Angles. Pada materi interval kepercayaan yang kita telah coba untuk melihat apakah apel di New York sama mahalnya dengan yang ada di LA. Kita menggunakan 10 toko di

New York dan 8 toko di L.A. Setelah kita mendapatkan semua harga dan memasukkannya ke dalam tabel, dengan variasi populasi harga apel yang diasumsikan sama untuk New York dan L.A. Dari kondisi tersebut, maka kita dapat nyatakan hipotesis nol dan alternatif. yaitu:

$$H0 : \mu_{ny} = \mu_{LA} \text{ atau } \mu_{ny} - \mu_{LA} = 0$$
$$Ha : \mu_{ny} \neq \mu_{LA} \text{ atau } \mu_{ny} - \mu_{LA} \neq 0$$

Berikut adalah datanya:

NY apples	LA apples
$ 3,80	$ 3,02
$ 3,76	$ 3,22
$ 3,87	$ 3,24
$ 3,99	$ 3,02
$ 4,02	$ 3,06
$ 4,25	$ 3,15
$ 4,13	$ 3,81
$ 3,98	$ 3,44
$ 3,99	
$ 3,62	

Dari data tersebut dapat diketahui:

	NY	LA
Mean	$ 3,94	$ 3,25
Std. deviation	$ 0,18	$ 0,27
Sample size	10	8

Dari data yang diketahui tersebut maka dapat dihitung: *pooled varians* dengan rumus:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(10-1)0.0324+(8-1)0.0729}{10+8-2} = 0.05$$

Selanjutnya, kita dapat mencari pooled standar deviasi dengan rumus:

$$\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} = \sqrt{\frac{0.05}{10} + \frac{0.05}{8}} = 0.11$$

Karena data yang digunakan adalah sampel kecil dengan varians yang tidak diketahui maka kita akan menggunakan uji statistik-t, dengan df (*degree of freedom*) adalah jumlah sampel dikurangi 2, yaitu: 10 + 8 – 2 = 16. Nilai T-score dapat dicari dengan rumus berikut ini:

$$T = \frac{\bar{d} - \mu_0}{St.error} = \frac{0.69-0}{0.11} = 6.53$$

Nilai d adalah selisih rata-rata kedua sampel yaitu: 3.94 - 3.25 = yaitu 0.69. Maka kita akan mendapatkan semua data sebagai berikut

Pooled variance	0,05
Pooled std	0,11
T-score	6,53
p-value	0,000

Jika kita akan mencari nilai t-statistik maka kita dapat melihat dalam tabel t:

d.f. / α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
inf.	1.282	1.645	1.960	2.326	2.576

Dari tabel t tersebut, diketahui bahwa untuk setiap nilai pada tingkat df 16 memiliki nilai lebih kecil dari T-score (6.53). Jika nilai T-score lebih tinggi dari 2 maka kita akan menolak Ho. (t-tabel < T-score). Dengan data yang sama kita juga dapat mencari nilai-P dengan menggunakan kalkulator web, dan hasilnya adalah 0.000. Dengan demikian dapat diartikan bahwa kita menolak Ho pada semua tingkatan signifikansi.

Dan dengan standar deviasi yang kecil dan dekat sekitar 0.2, kita dapat dengan mudah mengatakan bahwa harganya berbeda, maka tidak ada pengujian yang diperlukan. Pertanyaan yang jauh lebih menarik adalah apakah harga apel di New York 20 persen lebih tinggi daripada di LA. Saya akan meninggalkan Anda latihan ini untuk pekerjaan rumah. Selanjutnya kita selesai dengan pengujian hipotesis.

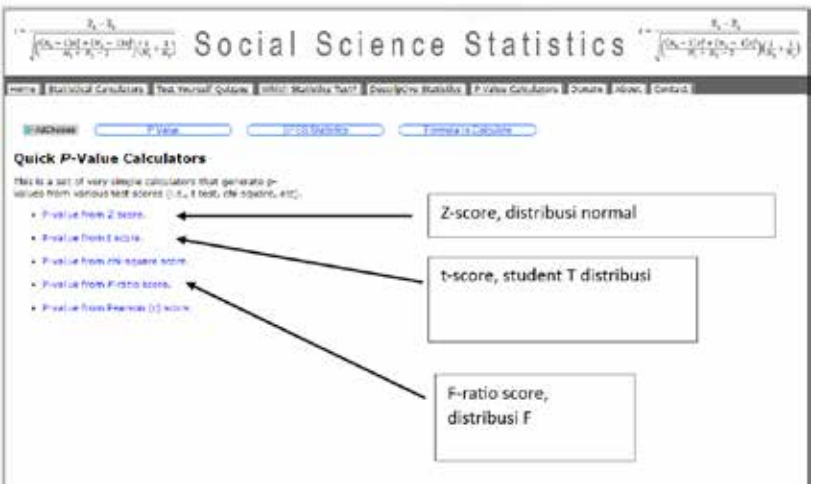
Appendix Bab 5:
Excel uji hipotesis

A.5.1. Kalkulator statistik

Pada bagian ini kita akan melihat kembali bagaimana nilai dalam hitungan yang sudah diperoleh pada bab 5 dalam Ms Excel. Dalam materi bab 5 tentang uji hipotesis, kita banyak menggunakan contoh kasus dari materi-materi sebelumnya, namun kita menguji dalam hipotesis. Dengan demikian kita dapat merujuk kembali materi sebelumnya untuk melihat bagaimana nilai diperoleh dari Ms Excel. Oleh karena itu kita akan fokus untuk mengetahui bagaimana penggunaan kalkulator statistik untuk mengetahui nilai-P, yang merupakan poin materi penting pada bab 5 ini.

Berikut adalah cara untuk menghitung nilai-P lebih mudah:

1. Kita dapat masuk pada alamat web berikut: http://www.socscistatistics.com/pvalues/
2. Pilih tes yang berlaku untuk masalah Anda dan klik tautannya.



A.5.2. P-value untuk Z-score

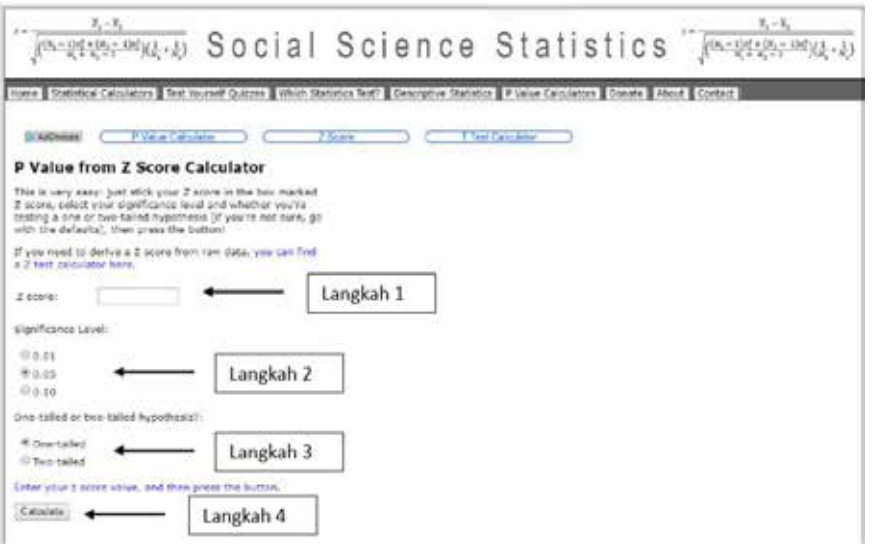
Berikut ini adalah penggunaan kalkulator statistik untuk menghitung nilai-P pada Z-score

Langkah 1: Ketik skor Z yang didapatkan dari tes.

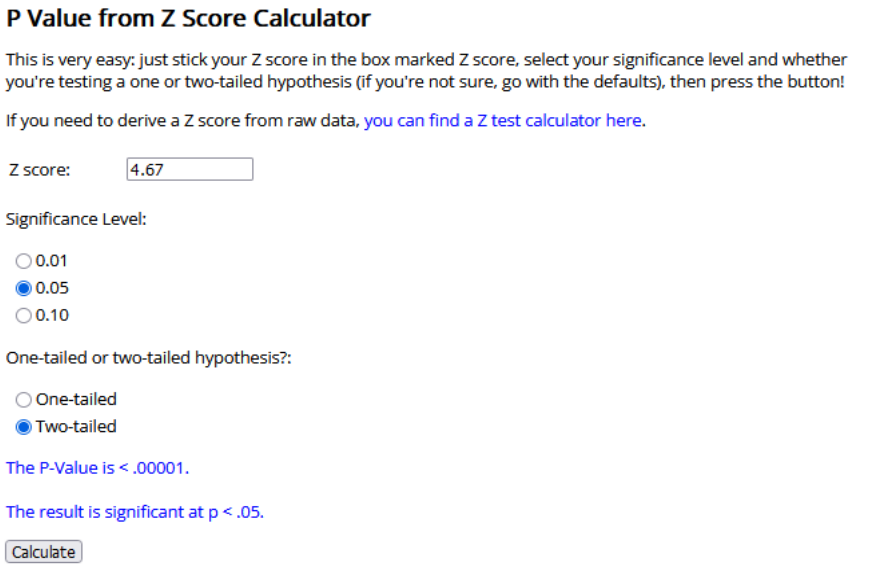
Langkah 2 (opsional): Pilih tingkat signifikansi.

Langkah 3: Pilih apakah ini tes satu ekor atau dua ekor.

Langkah 4: Klik Hitung.



Untuk lebih jelas, maka kita dapat menggunakan contoh kasus bab 5 tentang Uji Mean (5.4. populasi variance diketahui). Dalam contoh kasus tersebut hasil Z-score adalah -4.67. Karena kita menggunakan dua sisi maka kita dapat menggunakan nilai absolut maka kita dapat menggunakan calculator nilai-P sebagai berikut:



Dari hasil tersebut dapat dilihat bahwa kalkulator tersebut telah menunjukkan langsung hasil nilai $p < 0.05$, yang artinya Ho ditolak.

A.5.3. P-value untuk t score

Berikut ini adalah penggunaan kalkulator statistik untuk menghitung nilai-P pada t-score

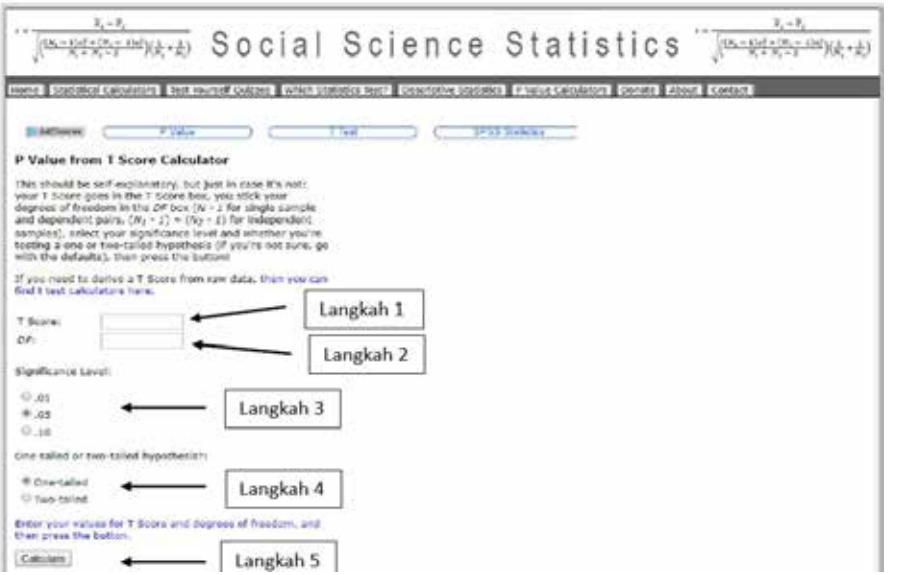
Langkah 1: Ketik skor t yang didapatkan dari tes Anda.

Langkah 2: Ketik derajat kebebasan (df) yang terkait dengan tes Anda.

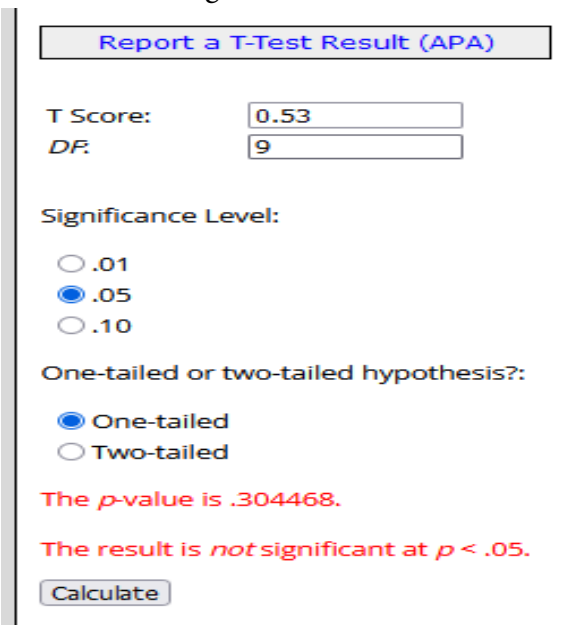
Langkah 3 (opsional): Pilih tingkat signifikansi.

Langkah 4: Pilih apakah ini tes satu ekor atau dua bekor.

Langkah 5: Klik hitung.



Kita akan menggunakan contoh kasus pada bab 5 pada bagian 5.6. Uji Rata-rata: Varians Populasi Tidak Diketahui. Kita dapat menggunakan data pada tabel 5.1. dengan nilai T-score yang sudah dihitung adalah 0.53, dengan df 9. Maka nilai dapat dimasukan sebagai berikut:



Hasil tersebut menunjukkan bahwa nilai-P adalah 0.304, dengan demikian Ho diterima. Pada web tersebut juga terdapat beberapa alat hitung statistik yang bisa dicoba.

Latihan

Untuk mempermudah latihan, file excel untuk latihan dapat di akses pada link berikut: https://bit.ly/Data_Latihan_Excel. Pilih pada folder yang sesuai dan file yang sesuai. Latihan diambil dari 365 Business Intelegence

Latihan 5.1. Uji rata-rata untuk varians diketahui

Dataset: gaji data analis, 30 data.

$ 117.313	$ 112.276
$ 104.002	$ 108.637
$ 113.038	$ 96.818
$ 101.936	$ 92.307
$ 84.560	$ 114.564
$ 113.136	$ 109.714
$ 80.740	$ 108.833
$ 100.536	$ 115.295
$ 105.052	$ 89.279
$ 87.201	$ 81.720
$ 91.986	$ 89.344
$ 94.868	$ 114.426
$ 90.745	$ 90.410
$ 102.848	$ 95.118
$ 85.927	$ 113.382

Tugas:

1. Dari gaji data analis di atas, buatlah rumusan hipotesis jika kita ingin mengetahui apakah nilai gaji tersebut sama dengan $ 113.000
2. Uji hipotesis nol pada signifikansi 10%

Latihan 5.2. Uji rata-rata untuk varians populasi tidak diketahui

Dataset: data rata-rata keterbukaan email.

Rata-rata keterbukaan
26%
23%
42%
49%
23%
59%
29%
29%
57%
40%

Tugas: Dari data keterbukaan email tersebut kerjakanlah:

1. Bagaimana jika pertanyaannya adalah: apakah tingkat terbuka pesaing TEPAT 40%. Buatlah rumusan hipotesisnya.
2. Uji pada signifikansi 5%. Komentari keputusan dengan jargon statistik yang sesuai.
3. Uji pada signifikansi 1%. Komentari keputusan dengan jargon statistik yang sesuai.
4. Hitung nilai-p dari test. (Selesaikan tugas no 2 and 3 using the p-value)
5. Apa yang akan menjadi keputusan akhir dari hasil terhitungan di atas?

Latihan 5.3. : Uji rata-rata untuk sampel dependen

Dataset: berat badan sebelum dan setelah program

Subyek	Berat sebelum (kg)	Berat sesudah (kg)
1	103,68	92,87
2	110,68	101,58
3	119,05	105,66
4	101,75	96,18
5	91,69	86,97
6	112,03	105,90
7	88,84	80,56
8	105,18	97,00
9	110,37	99,27
10	120,99	107,44

Beberapa orang di internet merancang program penurunan berat badan. Kita akan menguji apakah program tersebut berhasil atau tidak.

Tugas

1. Hitung perbedaan antara sebelum dan sesudah
2. Nyatakan hipotesis nol.
3. Hitung statistik yang sesuai
4. Putuskan apakah ini tes satu sisi atau dua sisi. Hitung nilai-p dari statistik ini
5. Berdasarkan nilai-p, putuskan pada signifikansi 1%, 5% dan 10%, apakah program berfungsi.

Latihan 5.4. : Uji rata-rata untuk sampel independen (varians populasi diketahui)

Ajax merilis deterjen piring baru, yang disebut Extreme Ajax+, yang mereka iklankan untuk dapat membersihkan lebih banyak piring dengan satu botol.

Anda ingin melihat apakah itu benar dan telah membeli 25 botol Extreme Ajax+ dan 25 botol Ajax dan mengumpulkan data tentang berapa banyak piring kotor yang dapat Anda cuci dengan satu botol. (Varians populasi diasumsikan diketahui.)

Dataset: perbandingan dua deterjen

	Extreme Ajax+	Ajax
Size	25	25
Mean	115	100
Variance	625	400

Tugas:

1. Temukan standar deviasi dari kedua sampel
2. Nyatakan Hipotesis Nol
3. Tentukan Skor-Z
4. Temukan nilai-p
5. Menafsirkan hasilnya

Latihan 5.5. : Uji rata-rata untuk sampel independen (varians populasi tidak diketahui diasumsikan sama)

Dataset: perbandingan data belanja

	Monday	Saturday
Mean	1.078,00	908,20
Std. deviation	633,00	469,80
Sample size	24	21

Anda memiliki data tentang berapa kali orang mengklik add-up pop-up pada 24 hari Senin dan 21 hari Sabtu di platform e-learning selama beberapa tahun. Sampel ditarik secara independen.

Tugas:

Secara statistik, apakah ada bukti kuat bahwa jumlah klik yang dicatat ditambahkan pada hari Senin lebih tinggi daripada jumlah klik pada hari Sabtu?

Appendix 6:
Excel Untuk Analisa Data

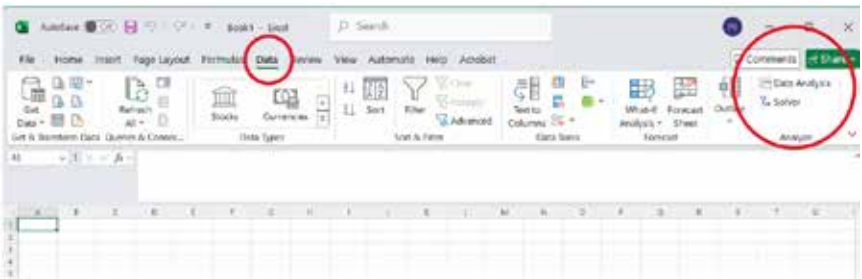
Ms Excel telah secara mendasar memberikan fitur atau layanan untuk dapat mengolah/mempersiapkan data dan menganalisa data. Mengolah data merupakan tahapan yang dasar yang perlu disiapkan untuk dapat menganalisa data lebih lanjut. Hanya dengan mengolah untuk persiapan data dari Ms Excel, kita bisa mendapatkan informasi yang dasar yang cukup mengenai data set yang akan dianalisa. Beberapa fitur yang dapat digunakan dalam melakukan olah data dalam rangka mempersiapkan data antara lain:

1. Validasi data,
2. Mengelompokan data dengan kriteria tertentu,
3. Menemukan data dengan kriteria tertentu,
4. Menghapus data ganda atau duplikat,
5. Memisahkan atau menyatukan data,
6. Menyusun data menggunakan pivot.

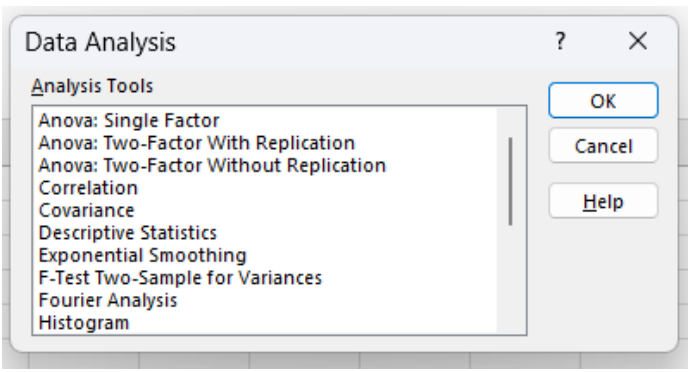
Semua tahapan tersebut merupakan tahapan dasar mengolah data dengan excel untuk dapat dianalisa lebih lanjut. Hanya dengan melakukan olah data yang dasar tersebut, kita bisa mendapatkan informasi dasar dari data yang ada. Namun seperti pada judul dan tujuan dari buku ini, maka pembahasan dan contoh pada buku ini akan lebih fokus pada analisa data dengan pendekatan statistic. Pendekatan statistik tentu dapat memberikan hasil olah data yang lebih akurat terutama jika akan digunakan untuk membuat keputusan, terutama jika kita akan memprediksi suatu kejadian. Secara khusus pada bagian Appendix 6 ini, kita akan melihat fitur yang sudah diberikan Ms Excel untuk analisa data dan belajar mengunakan beberapa dari fitur atau layanan tersebut.

Dalam analisa data, selain rumus dan fungsi Ms Excel yang sudah disampaikan pada materi sebelumnya, Ms Excel memiliki

fitur atau fungsi khusus yang dapat digunakan untuk melakukan analisa data menggunakan pendekatan statistik. Seperti pada fungsi statistik dan matematika yang ada, Ms Excel juga menyediakan tool khusus untuk analisa data yang bisa langsung digunakan. Tools tersebut pada menu data – *data analysis*.



Jika kita klik pada Data Analysis, maka kita akan mendapatkan tampilan berikut ini.



Kita lihat pada tools untuk data analysis yang disediakan oleh Ms Excel, tools tersebut telah mengarahkan langsung pada fungsi statistik. Secara tidak langsung, beberapa fungsi dan pengertian telah kita bahas pada materi di bab 1 sampai 5. Materi tersebut akan membantu kita untuk memahami konsep dasar analisa data dengan menggunakan pendekatan statistik. Diharapkan setelah kita memahami dasar analisa data dengan pendekatan statistika, kita dapat menggunakan fungsi statistik yang tersedia dalam Ms Excel dengan lebih terarah sesuai dengan kebutuhan analisis kita. Pada

appendix 6, kita tidak akan membahas satu per satu fungsi yang ada pada tools data analysis tersebut. Untuk mempermudah latihan, file excel untuk latihan dapat di akses pada link berikut: https://bit.ly/Data_Latihan_Excel. Pilih pada folder yang sesuai dan file yang sesuai. Latihan diambil dari 365 Business Intelegence

A.6.1. Statistik Deskriptif

Contoh pertama penggunaan tools data analysis adalah untuk melihat gambaran data melalui statistik deskriptif. Detail penjelasan terkait dengan statistik deskriptif dapat dilihat pada Bab 2, dan kita akan langsung masuk pada contoh pada bagian ini.

Kasus yang digunakan pada bagian ini adalah data tentang harga rumah dan luas rumah. Pertanyaan dasar dari data tersebut adalah kita ingin mengetahui hubungan dari luas rumah dengan harga rumah. Berikut adalah contoh datanya:

Area (ft.) of Property	Price of Property
717.05	$ 193,660.62
747.50	$ 237,060.15
1121.95	$ 372,001.70
1121.95	$ 290,031.26
827.87	$ 238,811.06
747.50	$ 199,054.20
1608.84	$ 496,266.41
1132.06	$ 346,906.89
1383.84	$ 376,964.62
927.83	$ 315,733.15
669.16	$ 188,273.73
928.16	$ 253,831.02
798.50	$ 278,575.87
1305.62	$ 402,081.80
1121.95	$ 310,832.59
785.48	$ 257,183.48
927.08	$ 326,885.34
1109.25	$ 344,568.74

Terdapat 200 data dalam data set ini. Dari data ini kita akan menggunakan tools data analysis. Berikut langkahnya:

1. Dalam contoh adalah pada kolom A dan B. Pada data tersebut, kita akan menempatkan harga sebagai faktor atau variabel yang dipengaruhi atau dependen/terikat, sedangkan luas rumah adalah variabel yang mempengaruhi atau independen/bebas. Kemudian, kita akan lihat bagaimana pola hubungan luas rumah dengan harga rumah dengan menggunakan visualisasi data yaitu scater plot.

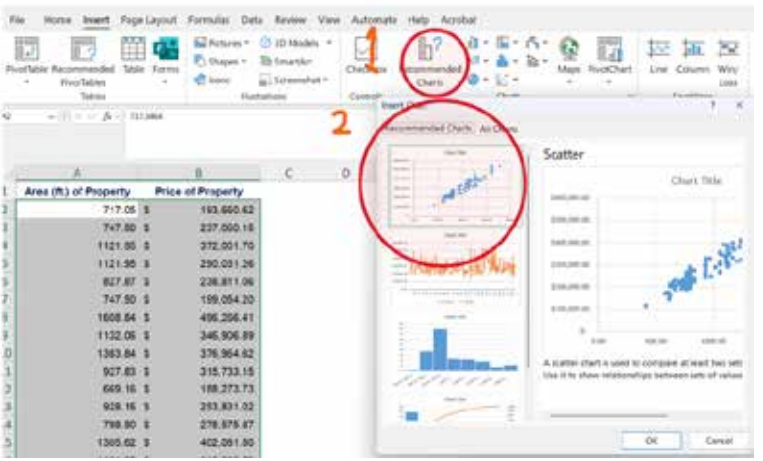
   Untuk melihat hubungan tersebut langkahnya adalah:

   → Blok data pada kolom A dan B.

   → Pilih menu insert

   → Pilih Chart

   → Pilih Scatter



Hasilnya adalah sebagai berikut:

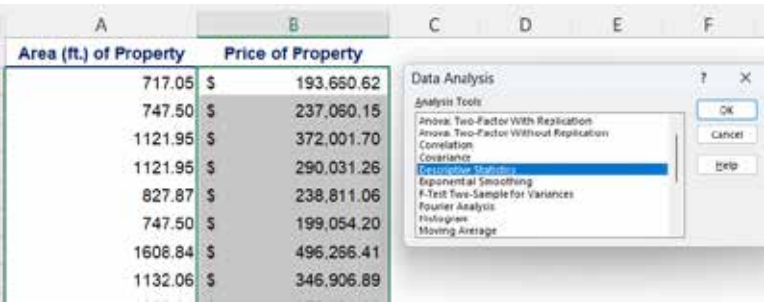
Dari gambar tersebut maka kita dapatkan informasi bahwa hubungan antara luas rumah dengan harga rumah adalah linear /lurus dan sifatnya positif, yang artinya semakin luas area rumah maka harga semakin tinggi.

2. Selanjutnya kita akan melihat deskripsi data menggunakan tools data analysis. Langkahnya adalah sebagai berikut:

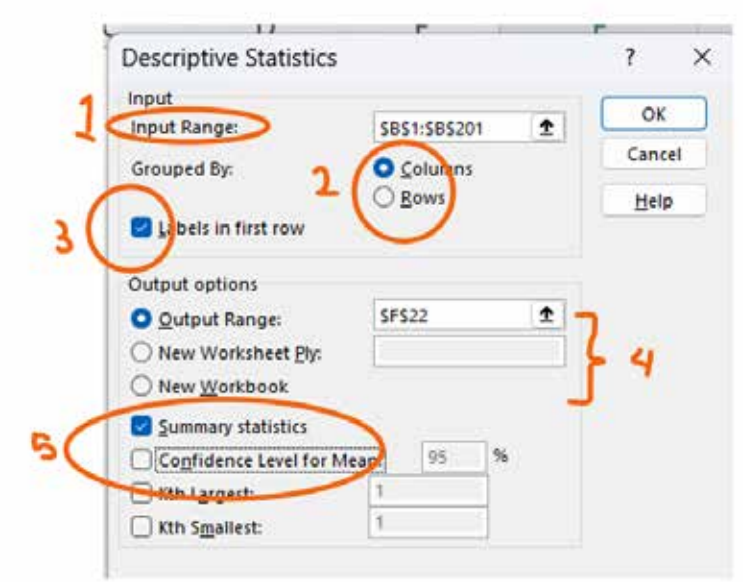
   → Pilih menu **data**

   → Pilih **Data Analysis** pada ribbon

   → Pilih **Descriptive Statistic,** seperti pada gambar di bawah



Setelah memilih descriptive statistics maka akan muncul box seperti berikut:



Keterangan:

1. Pada bagian 1, *input range* adalah area yang ingin kita lihat deskripsi datanya. Dalam contoh kasus ini, kita akan melihat statistik deskriptif dari faktor atau variabel harga rumah. Maka pada kolom input range, kita bisa masukan range data harga rumah mulai dari judul sampai data terakhir (B1:B201)
2. Pada bagian 2, kita pilih *columns* karena data harga rumah ada pada satu kolom B.
3. Pada bagian 3 kita centang karena baris pertama pada kolom data adalah judul harga rumah.
4. Pada bagian 4, output options. Kita bebas memilih dimana hasil statistik deskriptif akan ditampilkan.
5. Pada bagian 5, kita dapat memilih *summary statistic* untuk menampilkan hasil statistik deskriptif. Setelah itu, klik ok

Hasilnya adalah sebagai berikut:

Price of Property	
Mean	282,956
Standard Error	6,285
Median	251,672
Mode	#N/A
Standard Deviation	88,881
Sample Variance	7,899,904,641
Kurtosis	1
Skewness	1
Range	420,708
Minimum	117,564
Maximum	538,272
Sum	56,591,208
Count	200

Jika melihat kembali pada appendix Bab 2 maka hasil tersebut dapat dicari dengan rumus berikut ini:

Price of Property		Rumus Fungsi
Mean	282,956	AVERAGE
Standard Error	6,285	STDEV/SQRT(COUNT)
Median	251,672	MEDIAN
Mode	#N/A	MODE
Standard Deviation	88,881	STDEV
Sample Variance	7,899,904,641	VAR
Kurtosis	1	KURT
Skewness	1	SKEW
Range	420,708	MAX-MIN
Minimum	117,564	MIN
Maximum	538,272	MAX
Sum	56,591,208	SUM
Count	200	COUNT

A.6.2 Histogram

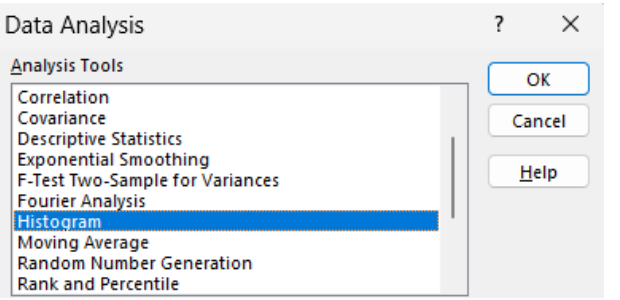
Histogram merupakan bentuk visualisasi data untuk menampilkan data frekuensi atau banyaknya data. Fitur pada tools data analytic juga tersedia fitur untuk menampilkan histogram.

Dengan data yang sama, maka kita akan menampilkan histogram pada data harga rumah. Langkahnya adalah sebagai berikut:

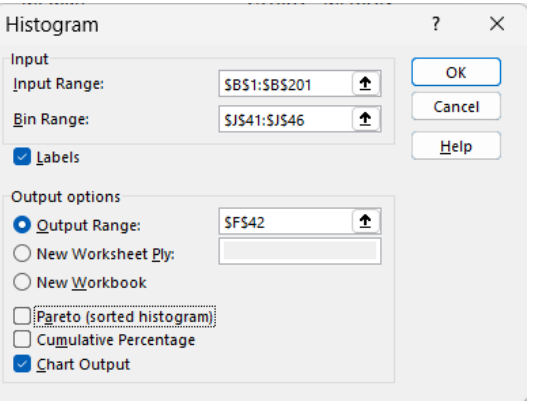
1. Kita perlu menentukan jangkauan nilai data yang diinginkan. Range data kita sebut juga dengan bins. Pada sel yang kosong kita dapat mengetikan bins dengan rentang 50.000 dari nilai 200.000. Berikut contohnya:

Bins
200,000
250,000
300,000
350,000
400,000

2. Setelah itu, kita kembali memilih menu data, kemudian data analysis, lalu memilih histogram.



Kemudian akan muncul box seperti berikut:

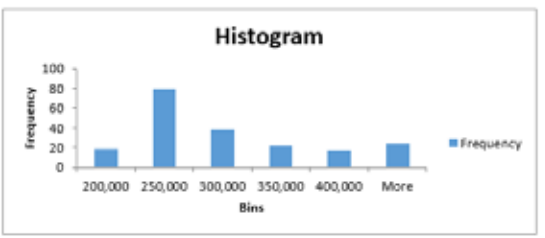


Keterangan:

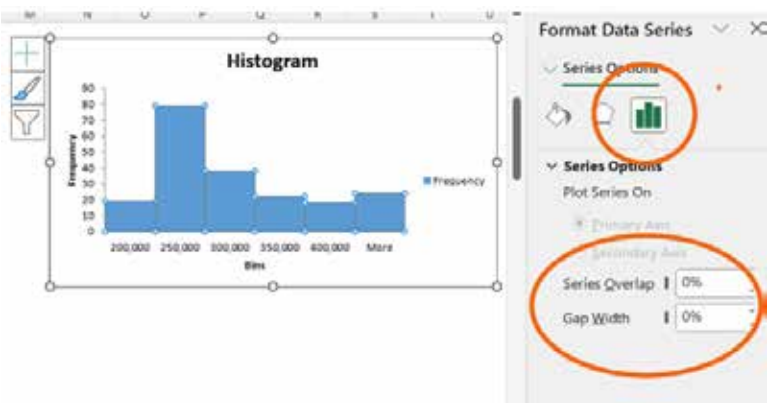
1. *Input range* merupakan data yang akan kita buat visualisasinya, yaitu data pada kolom harga rumah (B1-B201)
2. Bin range merupakan data bins yang tadi sudah kita buat.
3. Centang *labels* karena dalam kolom data kita termasuk judul pada baris pertama kolom.
4. Pada *output option* kita dapat memilih tempat hasil histogram dimunculkan.
5. Kemudian terakhir kita dapat memilih *chart output*.

Hasilnya adalah sebagai berikut:

Bins	Frequency
200,000	19
250,000	79
300,000	38
350,000	22
400,000	18
More	24



Kita dapat mengedit gambar batang dengan klik pada bar, maka akan muncul seperti ini:

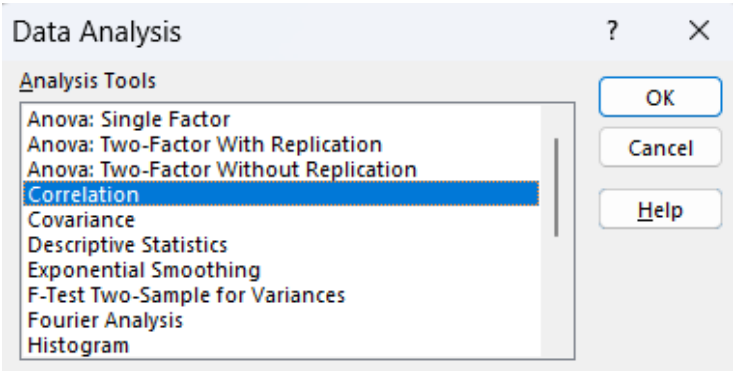


Kita dapat memilih ikon pada gambar diagram, kemudian pada series dan gap kita buat 0%.

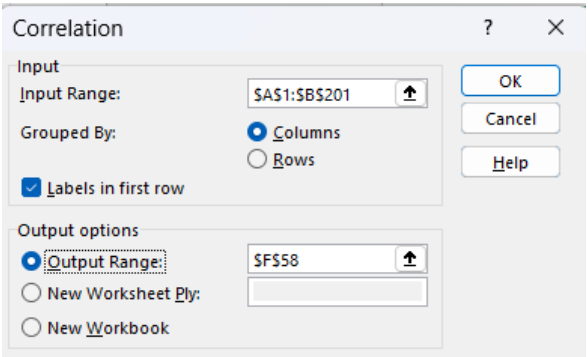
A.6.3. Korelasi (*Correlation*)

Uji korelasi dilakukan untuk menilai apakah ada hubungan antara dua variable/faktor dan menilai seberapa kuat hubungan tersebut. Nilai korelasi ditunjukkan melalui nilai 0 sampai 1, dimana nilai 1 (satu) menunjukkan hubungan yang kuat. Untuk melihat adanya hubungan dua variabel, kita dapat melihat visualisasi data scatter plot. Jika ada pola garis miring, maka kondisi tersebut dapat dipahami sebagai adanya hubungan 2 variabel. Namun kita perlu menguji nilai kekuatan hubungan tersebut. Kita kembali dapat menggunakan fitur yang ada pada tools data analysis. Kita kembali menggunakan data yang sama. Dengan langkah yang sama maka kita akan mempersingkat langkahnya:

1. Kita pilih kembali *data analysis* dan pilih *correlation*



2. Setelah memilih *correlation* maka, akan muncul box seperti berikut ini:



Keterangan:

1. Pada bagian *input range*, kita dapat memilih semua data dengan blok semua data area rumah dan harga rumah (A1-B201).
2. Kita pilih columns dan centang *labels in first row*, karena data kita termasuk judul pada baris pertama kolom.
3. Terakhir kita pilih tempat untuk menampilkan hasil. Berikut adalah hasilnya

	Area (ft.) of Property	*Price of Property*
Area (ft.) of Property	1	
Price of Property	0.955733006	1

Hasil dari hubungan atau nilai korelasi antar luas rumah dengan harga rumah adalah 0.9557. Nilai tersebut menunjukkan nilai yang mendekati 1 dan bernilai positif. Hal ini menunjukkan bahwa luas area rumah dan harga rumah memiliki hubungan yang positif dan kuat karena nilai mendekati 1.

A.6.4. Regresi linear (*Linear Regression*)

Regresi linear merupakan pendekatan dalam statistik yang digunakan untuk memprediksi. Regresi linear memiliki ciri khusus yaitu adanya faktor/variabel bebas atau indenpenden dan adanya variabel terikat atau dependen. Oleh karena itu, dalam setiap data set yang akan dianalisa, kita perlu menentukan dahulu mana yang menjadi variabel independen dan mana yang menjadi variabel dependen. Terdapat beberapa dasar yang dapat digunakan untuk menentukan variabel independen dan dependen, antara lain adalah teori. Selain itu, dalam analisa akademis teori menjadi dasar yang mutlak untuk menentukan variable indenpenden dan dependen. Namun dalam analisa industri atau bisnis, penentuan dalam variabel dapat dilakukan dengan lebih fleksibel, yaitu dengan dengan eksplorasi data secara visual, atau berdasarkan kasus atau kebutuhan. Dalam buku ini, kita tidak membahas detail tentang regresi, namun contoh ini diharapkan untuk memberikan

pemahaman pemanfaatan Ms Excel untuk melakukan olah data, terutama dalam analisa industri atau bisnis.

Regresi yang ditunjukkan pada contoh ini adalah regresi sederhana dengan menggunakan satu variabel indenpeden, yang memiliki rumus fungsi:

$$Y = a + bX$$

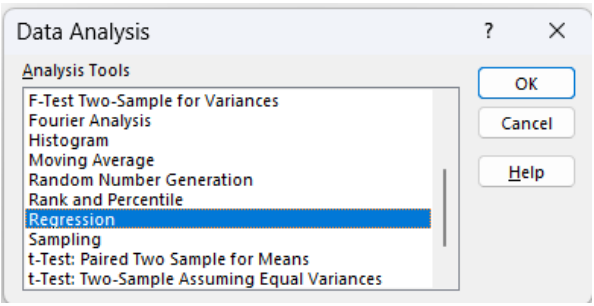
Dimana Y adalah hasil atau variabel dependen, a adalah nilai constanta, b adalah nilai *coeffsien*, dan X adalah nilai variabel independen.

Dalam contoh ini, kita akan menggunakan kasus luas rumah dan harga rumah. Langkahnya adalah sebagai berikut:

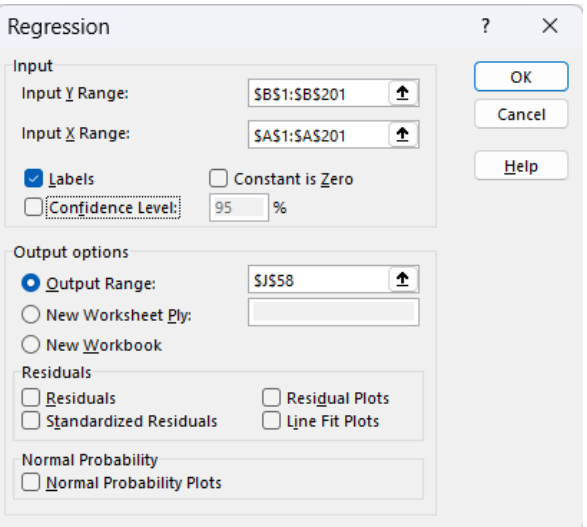
1. Menentukan variabel independen dan dependen.

   Pada tahap ini kita dapat melihat kembali hasil scatter gambar tersebut menunjukkan adanya hubungan, sehingga kita bisa melanjutkan regresi. Kita tentukan bahwa luas rumah merupakan variabel independen dan harga rumah merupakan dependen. Secara dasar, harga merupakan faktor yang muncul dari nilai barang atau ditentukan dari nilai barang, sehingga harga dijadikan sebagai faktor/variabel dependen.

2. Masuk kembali pada *data analysis*, kemudian pilih *regression*.



Setelah kita pilih *regression* maka akan muncul box berikut:



Keterangan:

1. Pada bagian *input Y range*, kita dapat blok kolom harga rumah sebagai variabel dependen.
2. Pada bagian *input X range*, kita dapat blok kolom luas rumah sebagai variabel independen.
3. Centang *label*, karena dalam kolom data kita juga terdapat judul faktor/variabel.
4. Kemudian pilih output tempat hasil akan dimunculkan. Hasilnya sebagai berikut:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.955733006
R Square	0.913425579
Adjusted R Square	0.912988334
Standard Error	26218.00644
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.43598E+12	1.43598E+12	2089.049655	3.7505E-107
Residual	198	1.36102E+11	687383861.7		
Total	199	1.57208E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-4147.869999	6549.382344	-0.633322317	0.527254022	-17063.36645	8767.626451	-17063.36645	8767.626451
Area (ft.) of Property	304.4861942	6.661825069	45.70612273	3.75E-107	291.3489587	317.6234297	291.3489587	317.6234297

Kita dapat melihat bahwa hasil yang diberikan cukup lengkap. Berikut adalah beberapa poin yang dapat diperhatikan dalam memahami hasil output regresi tersebut.

1. Nilai *R-Square* sebesar 0.913 menunjukkan bahwa variabel *independent* dapat menjelaskan pengaruh terhadap variabel dependen sebesar 0.913 atau 91.3%.
2. Selanjutnya kita dapat melihat nilai *Signifincance* F dan nilai P-value yang memiliki nilai 3.75E-107. Nilai ini menunjukkan bahwa tingkat signifikasi yang lebih rendah dari tingkat alpha yang 1%, 5%, dan 10%. Hal ini dapat diartikan juga bahwa hasil *coeffisients* dapat diterima secara statistik.
3. Nilai *coefficients* sebesar 304.49. Nilai ini menunjukkan nilai positif, yang artinya peningkatan variabel independen akan meningkatkan variabel dependen, dengan nilai peningkatan variabel indpenden sebesar 304.49.
4. Nilai *constanta* sebesar -4147.87. Nilai ini menunjukkan nilai tetap yang mempengaruhi nilai Y

Dengan nilai yang sudah kita dapatkan maka kita dapat masukan dalam rumus fungsi untuk memperkirakan harga rumah.

$$Y = a + bX \sim Y = -4147.87 + 304.49(X)$$

Dengan rumus di atas, jika kita memilih informasi luas rumah adalah sebesar 700 maka rumusnya adalah

$$Y = -4147.87 + 304.49(700)$$

Hasilnya adalah 208,992.47. Hasil ini dapat membantu memberikan gambaran jika kita akan membeli rumah dengan luas 700 m^2 maka harga sekitar 208,992.47. Dengan demikian dapat menggunakan pendekatan untuk dapat membuat keputusan yang baik.

Dalam analisa regresi, kita dapat juga membangun hipotesis seperti pada materi Bab 5. Melalui hasil data analysis excel ini, maka kita mendapatkan semua nilai yang diperlukan untuk merumuskan hipotesis dan ujinya, khususnya dalam penentuan nilai-P (P-*value*).

Daftar Pustaka

[1] B. Illowsky and S. Dean, *Introductory Statistics (2nd ed.)*, 2nd ed. OpenStax, 2024.

[2] D. Diez, M. Çetinkaya-Rundel, and C. Barr, *OpenIntro Statistics: Fourth Edition*, 4th ed. openintro.org/os., 2019.

[3] D. A. Lind, W. C. Marchal, and S. A. Wathe, *Basic Statistics For Business and Economics*. McGraw-Hil, 2006.

[4] Mario F. Triola, *Elementary Statistics, 14th edition*, 14th ed. Pearson, 2022.

[5] D. M. McEvoy, *A Guide to Business Statistics*. Wiley Global Research (STMS), 2018.

[6] J. T. McClave and T. Sincich, *Statistics 12th Edition*. Pearson Education, 2012.

ANALISA DATA DENGAN MS EXCEL: PENDEKATAN STATISTIK

Buku Analisa Data Dengan Ms Excel: Pendekatan Statistik dirancang khusus bagi mahasiswa yang ingin mempelajari cara mengolah dan menganalisis data menggunakan Microsoft Excel secara sederhana dan efektif. Dengan bahasa yang mudah dipahami dan pendekatan praktis, buku ini menggabungkan konsep dasar statistika dengan kemampuan Excel sebagai alat bantu utama.

Buku ini mencakup berbagai topik penting, seperti:

- Pengantar Microsoft Excel dan dasar-dasar pengolahan data.
- Statistika Deskriptif, termasuk perhitungan mean, median, modus, deviasi standar, dan visualisasi data dengan grafik.
- Analisis Data menggunakan tabel pivot, fungsi statistik, dan formula Excel.
- Penerapan Statistika Inferensial sederhana, seperti uji hipotesis dan regresi linear dasar.
- Studi Kasus Praktis, yang membantu mahasiswa memahami aplikasi konsep statistik dalam kehidupan nyata menggunakan Excel.

Dengan kombinasi penjelasan teori yang ringkas dan latihan-latihan praktis, buku ini cocok digunakan sebagai bahan ajar maupun panduan belajar mandiri. Mahasiswa akan dibekali keterampilan dasar dalam analisis data yang berguna untuk studi akademis maupun dunia kerja. Bagi mereka yang belum memiliki latar belakang statistik atau mahir menggunakan Excel, buku ini memberikan langkah-langkah yang sistematis untuk memahami dan menerapkan analisis data secara bertahap

Universitas Atma Jaya Yogyakarta
Jl. Babarsari No. 5-6 Yogyakarta 55281
Telp. +62 274 487711
E-mail: lib.publisher@uajy.ac.id

ISBN 978-623-10-8343-2 (PDF)
9 786231 083432