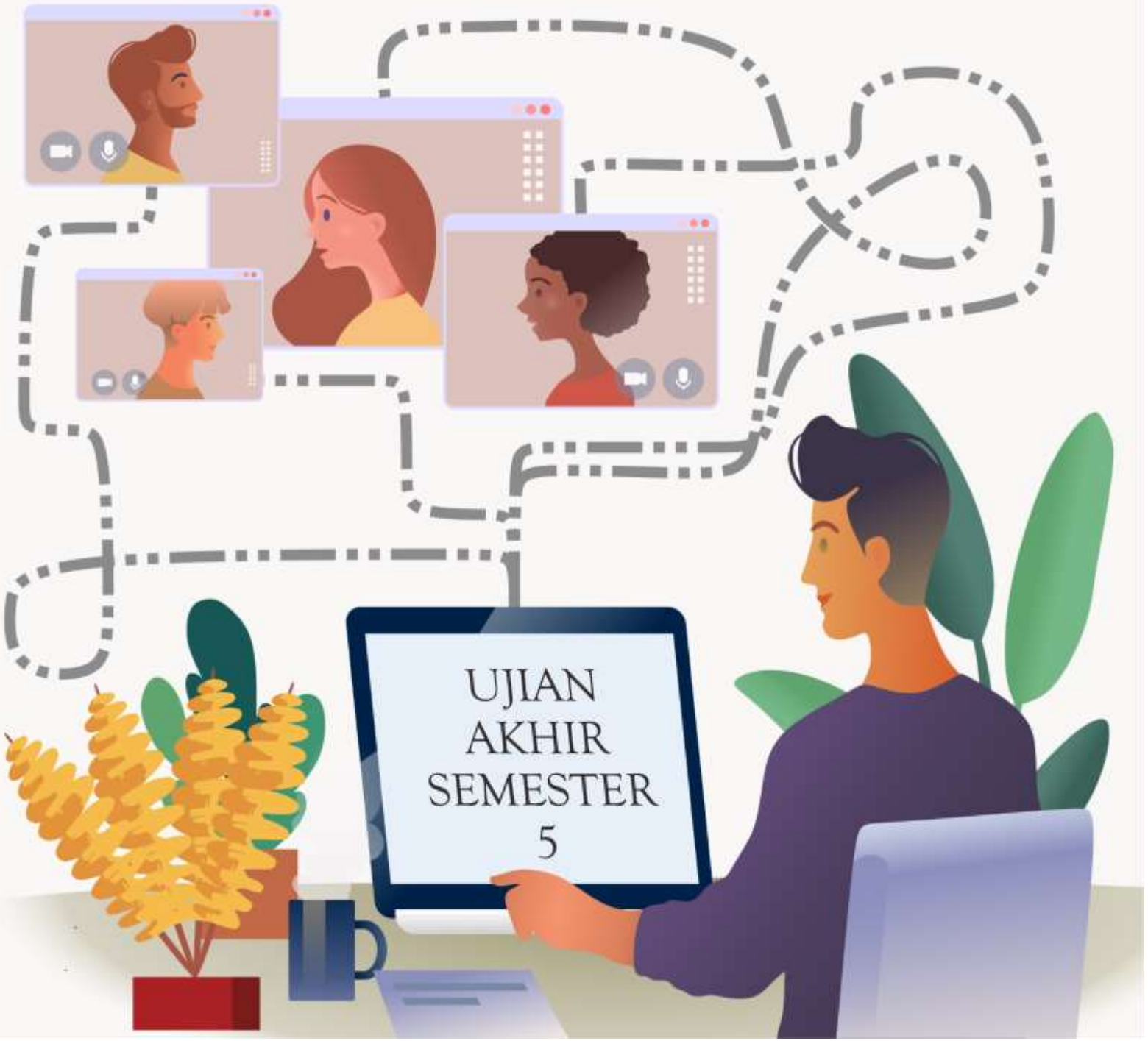
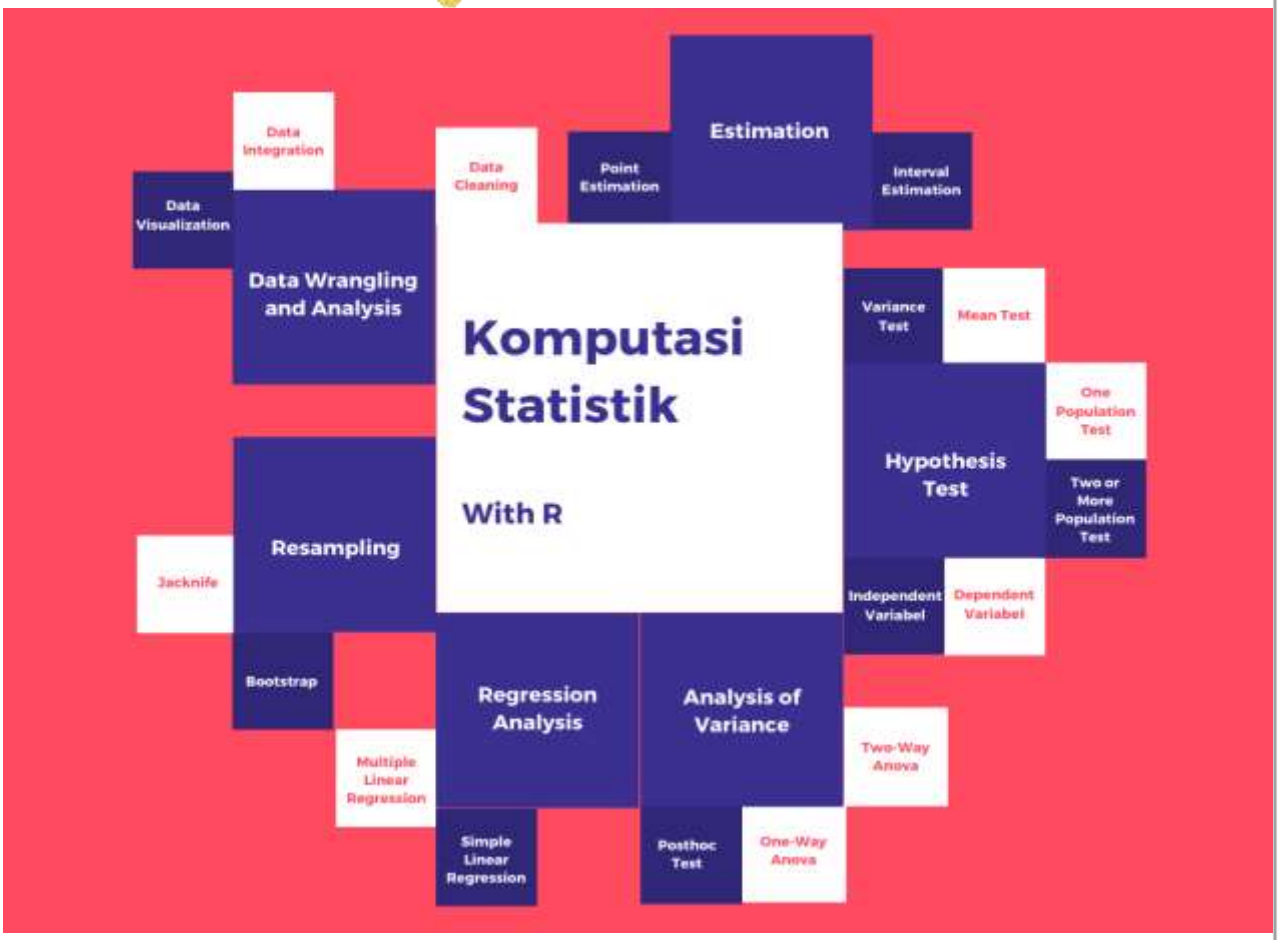
KOMPUTASI STATISTIK





Link Source Code

https://github.com/modul60stis/komstat-uas

Contoh kodingan untuk tiap bab tidak tersedia didalam modul ini, **semuanya ada di github**. Pembahasan latihan soal per-bab dan pembahasan soal tahun lalu juga **semuanya ada di github**. Beberapa latihan soal memakai `data_dummy_komstat.csv`, data tersebut dapat diakses disini https://raw.githubusercontent.com/modul60stis/data/main/data_dummy_komstat.csv

"Anakku, selamat berjuang. Hidup sekali, **hiduplah yang berarti.**"

— Ahmad Fuadi, Negeri 5 Menara

STATISTIK SEDERHANA

Statistik	Rumus	Fungsi	Package
Minimal	-	`min(x)`	base (Bawaan R)
		`which.min(x)`	base (Bawaan R)
Median	-	`median(x)`	stats (Bawaan R)
Maksimal	-	`max(x)`	base (Bawaan R)
		`which.max(x)`	base (Bawaan R)
Mean	$\underline{x} = \frac{\sum x_i}{n}$	`mean(x)`	base (Bawaan R)
Modus	-	-	-
Quartil		`quantile(x)`	stats (Bawaan R)
Range	$R = x_{max} - x_{min}$	-	-
Interquartil Range	$IQR = Q_3 - Q_1$	-	-
Sample Variance	$s^2 = \frac{\sum (x_i - \underline{x})^2}{n-1}$	`var(x)`	stats (Bawaan R)
Standard Deviasi	$s = \sqrt{\frac{\sum (x_i - \underline{x})^2}{n-1}}$	`sd(x)`	stats (Bawaan R)
Standard Error	$SE = \sqrt{\frac{s^2}{n}}$	`sd(x) / sqrt(n)`	stats (Bawaan R)
Jumlah	$sum = \sum x_i$	`sum(x)`	base (Bawaan R)
Skewness	$\frac{\frac{1}{n}\sum (x_i - \underline{x})^3}{\left(\frac{1}{n}\sum (x_i - \underline{x})^2\right)^{\frac{3}{2}}}$	`skewness(x)`	moments
Kurtosis	$\frac{\frac{1}{n}\sum (x_i - \underline{x})^4}{\left(\frac{1}{n}\sum (x_i - \underline{x})^2\right)^{2}}$	`kurtosis(x)`	moments

LATIHAN SOAL

1) Buatlah statistik sederhana (yang bisa diterapkan) dari variabel pada `data_dummy_komstat.csv`.

"Justru karena ini hal kecil. **Jangan sampai** dia **meremehkan** suatu hal, sekecil apapun."

— Ahmad Fuadi, Negeri 5 Menara

DISTRIBUSI PELUANG

Semua fungsi distribusi peluang di R tersedia pada package `stats`, yang merupakan package bawaan dari R.

Sintaks	Output
"d"	Probability Density Function (PDF)
"p"	Cumulative Density Function (CDF)
"q"	Inverse Cumulative Density Function
"r"	Random Generated Data

No	Distribution	Function	No	Distribution	Function
1	Normal	`dnorm, pnorm, qnorm, rnorm`	9	Hypergeometric	`dhyper, phyper, qhyper, rhyper`
2	Chi-Squared	`dchisq, pchisq, qchisq, rchisq`	10	Uniform	`dunif, punif, qunif, runif`
3	Student's t	`dt, pt, qt, rt`	11	Beta	`dbeta, pbeta, qbeta, rbeta`
4	Binomial	`dbinom, pbinom, qbinom, rbinom,`	12	Cauchy	`dcauchy, pcauchy, qcauchy, rcauchy`
5	Poisson	`dpois, ppois, dpois, rpois`	13	Geometric	`dgeom, pgeom, qgeom, rgeom`
6	F	`df, pf, qf, rf`	14	Logistic	`dlogis, plogis, qlogis, rlogis`
7	Gamma	`dgamma, pgamma, qgamma, rgamma`	15	Weibeull	`dweibull, pweibull, qweibull, rweibull`
8	Exponential	`dexp, pexp, qexp, rexp`	16	Wicoxon	`dwilcox, pwilcox, qwilcox, rwilcox`
9	Multinomial	`dmultinom, pmultinom, qmultinom, rmultinom`			

LATIHAN SOAL

1) Diasumsikan seorang telemarketer pada suatu hari berhasil menjual 20 dari 100 panggilan ($p = 0.2$). Jika ia menelpon 12 orang hari ini, berapakah peluang
   a) Tidak ada penjualan?
   b) Tepat 2 penjualan
   c) Paling banyak 2 penjualan
   d) Minimum 4 penjualan?
2) Setiap lot sebanyak 40 komponen dikatakan tidak lolos jika ditemukan produk cacat sebanyak 3 atau lebih. Suatu rencana sampling dilakukan dengan memilih 5 komponen secara acak dan menolak lot tersebut jika 1 produk cacat ditemukan. Berapakah peluang tepat 1 cacat ditemukan dalam sampel jika terdapat 3 produk cacat dikeseluruhan lot?
3) Seorang karyawan administrasi bertugas memasukkan 75 kata per menit dengan 6 error/kesalahan per jam. Berapakah peluang ia membuat 0-1 kesalahan dalam 255 kata yang dibuat?
4) Dibagian pengendalian kualitas usai bola lampu diasumsikan berdistribusi normal dengan $\mu = 2000$ jam dan $\sigma = 200$ jam. Berapakah peluang suatu bola lampu menyala selama
   a) Antara 2000 dan 2400 jam?
   b) Kurang dari 1470 jam?

ESTIMASI INTERVAL

Estimasi interval menunjukkan pada interval berapa suatu parameter populasi akan berada. Estimasi ini dibatasi oleh dua nilai, disebut sebagai **batas atas** dan **batas bawah**, yang masing-masing mempunyai simpangan d dari estimatornya. Besarnya d akan tergantung kepada ukuran sampel acak yang digunakan tingkat keyakinan (*level of confidence*), dan distribusi probabilitas untuk estimated value yang digunakan.

1. Estimasi Interval Rata-Rata

Varians	Populasi	Rumus
Diketahui	Terbatas	$\underline{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \underline{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
	Tidak Terbatas	$\underline{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \underline{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
Tidak Diketahui	Terbatas	$\underline{x} - t_{(\frac{\alpha}{2},df)} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \underline{x} + t_{(\frac{\alpha}{2},df)} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
	Tidak Terbatas	$\underline{x} - t_{(\frac{\alpha}{2},df)} \frac{s}{\sqrt{n}} < \mu < \underline{x} + t_{(\frac{\alpha}{2},df)} \frac{s}{\sqrt{n}}$

2. Estimasi Interval Proporsi

Varians	Populasi	Rumus
Diketahui	Terbatas	$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}}$

	Tidak Terbatas	$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

3. Ukuran Sample dari Rata-Rata Populasi dan Proporsi

Kualitas dari survei contoh bisa ditingkatkan dengan **menambah contoh (sample).** Di bawah ini formula untuk menghitung ukuran sample yang diperlukan untuk mengestimasi interval rata-rata populasi dengan selang kepercayaan $1 - \propto$, margin of error E, dan varians populasi σ^2.

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \sigma^2}{E^2}$$

Untuk proporsi

$$n = p(1-p)\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2$$

LATIHAN SOAL

1) An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours, find a 96% confidence interval for the population mean of all bulbs produced by this firm.
2) Hitunglah 92% confident interval untuk rata-rata 2 variabel continues pada `data_dummy_komstat.csv`
3) Sebuah percobaan dilakukan untuk mengetahui proporsi orang dewasa yang mengalami sindrom kelelahan kronis. Dalam percobaan ini, 4000 orang dipilih secara acak untuk melakukan survey. Suvey dilakukan dengan bertanya: apakah mereka mengalami kelelahan yang tidak normal (kronis) sehingga mengganggu aktivitas pekerjaan kantor atau pekerjaan rumah tangga mereka selama 6 bulan terakhir ? Dari 3066 yang menjawab survey, ada 590 orang yang mengalami kelelahn kronis Tentukan
   a) Berapakah point estimate untuk proporsi populasi orang dewasa yang mengalami kelelahan kronis?
   b) Buatlah confidence interval sebasar 95% untuk mengestimasi proporsi orang dewasa yang megalami kelelahan
4) Hitunglah 92% confident interval untuk proporsi mahasiswa yang merasa puas dengan metode pengajaran yang digunakana pada `data_dummy_komstat.csv.`
5) A marketing agency wishes to determine the average time, in days, that it takes to sell a product in various stores in a city. How large a sample will they need in order to be 98% confident that their sample mean will be within 2 days of the true mean? Assume that σ = 5 days.

UJI KESESUAIAN SEBARAN

1. Uji Shapiro Wilk

Uji Shapiro-Wilk digunakan sebagai uji kenormalan data. Uji ini cocok digunakan sampel kecil ($n < 30$). **Formula Shapiro-Wilk**:

$$W = \frac{b^2}{SS};\ b^2 = \sum_{i=1}^{m} a_i(x_{n+1-i} - x_i);\ SS = \sum_{i=1}^{n} (x_i - \underline{x})^2$$

Dengan

n adalah jumlah sampel atau banyaknya data yang akan diuji

$m = \frac{n}{2}$ bila n genap

$m = \frac{(n-1)}{2}$ bila n ganjil

a_i adalah nilai penimbang yang didapat dari tabel Shapiro Wilk

Hipotesis

H_0: Data berdistribusi normal

H_1: Data tidak berdistribusi normal

Dari nilai W akan didapat *p-value* yang dirujuk berdasarkan tabel2 Shapiro-Wilk. Jika *p- value* $<$ α maka keputusannya adalah menolak H_0 yang berarti data tidak berdistribusi normal. Jika *p-value* $\geq$ α maka keputusannya adalah menerima H_0 yang berarti data berdistribusi normal.

2. Uji Liliefors/Kolmogorov-Smirnov

Uji ini sama dengan uji kolmogorov-smirnov untuk 1 sampel, bedanya **kolmogorov menggunakan** varians dan rata-rata dari populasi. **Lilliefors menggunakan** varians dan rata-rata dari data. Hipotesisnya sebagai berikut.

H_0 : Data berdistribusi normal

H_1 : Data tidak berdistribusi normal

Dalam hal ini kita menggunakan distribusi kumulatif sampel (*sample cumulative distribution*) $S(x)$ dengan probabilita kumulatif normal (*normal cumulative probability*) $F(x)$. Jika H_0 benar maka $S(x)$ harus serupa (similar) dengan $F(x)$. $S(x)$ didefinisikan sebagai proporsi dari nilai sampel yang lebih kecil atau sama dengan x.

Statistik Uji

$$D = max|F(x) - S(x)|$$

Keputusan : Tolak Ho jika

D > nilai tabel D (Tabel liliefors pada Keller)

D > nilai tabel E (Tabel Kolmogorov pada Sidney)

3. Uji Chi-Square (Goodness of fit)

Pada uji ini data dibagi dalam beberapa interval, kemudian dihitung probabilita masing-masing interval. Probabilita dihitung dengan menggunakan distribusi normal dengan rata-rata (X) dan standar deviasi (S) sebagai estimator dari μ dan σ. **Hipotesisnya** sebagai berikut.

H_0 : Data berdistribusi normal

H_1 : Data tidak berdistribusi normal

Statistik Uji

$$\chi^2_{hitung} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Wilayah Kritis. $\chi^2_{hitung} > \chi^2_{(\alpha,k-3)}$

Fungsi	Kegunaan	Package
`shapiro.test(x)`	Uji Shapiro Wilk	stats (Bawaan R)

`ks.test(x, pnorm, mean = mu0, sd = sd0)`	Uji Kolmogorov Smirnov	stats (Bawaan R)
`lillie.test(x)`	Uji Liliefors	nortest
`chisq.test(x, p)`	Goodness of Fit	stats (Bawaan R)

LATIHAN SOAL

1) Ujilah apakah dua variabel continue pada `data_dummy_komstat.csv` berdistribusi normal?
2) Jika variabel sebelum dan sesudah pada `data_dummy_komstat.csv` dibagi berdasarkan metode apakah berdistribusi normal?

UJI RAGAM/VARIANS

Pengujian hipotesis mengenai variansi populasi atau simpangan baku berarti kita ingin menguji hipotesis mengenai keseragaman suatu populasi ataupun membandingkan keseragaman suatu populasi dengan populasi lainnya.

1. Uji Varians Satu Populasi

Hipotesis

$H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2 \; atau \; \sigma^2 < \sigma_0^2 \; atau \; \sigma^2 > \sigma_0^2$

Statistik Uji

$$\chi^2_{hitung} = \frac{(n-1)s^2}{\sigma_0^2} \; ; df = n - 1$$

Wilayah Kritis

- Dua arah, $\chi^2_{hitung} > \chi^2_{(\frac{\alpha}{2},df)}$ atau $\chi^2_{hitung} < \chi^2_{(1-\frac{\alpha}{2},df)}$
- Satu arah, $\chi^2_{hitung} > \chi^2_{(\alpha,df)}$ untuk $H_1: \sigma^2 > \sigma_0^2$ atau $\chi^2_{hitung} < \chi^2_{(1-\alpha,df)}$ untuk $H_1: \sigma^2 < \sigma_0^2$

2. Uji Varians Dua Populasi

Uji yang digunakan adalah uji F. Uji F **sangat sensitif terhadap asumsi kenormalan**, oleh karena sebaiknya data perlu memenuhi asumsi kenormalan.

Hipotesis

$H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2 \; atau \; \sigma_1^2 < \sigma_2^2 \; atau \; \sigma_1^2 > \sigma_2^2$

Statistik Uji

$$F_{hitung} = \frac{s_1^2}{s_2^2}; df_1 = n_1 - 1; \; df_2 = n_2 - 1$$

Wilayah Kritis

- Dua arah, $F_{hitung} > F_{(\frac{\alpha}{2},df_1,df_2)}$ atau $F_{hitung} < F_{(1-\frac{\alpha}{2},df_1,df_2)}$
- Satu arah, $F_{hitung} > F_{(\alpha,df_1,df_2)}$ untuk $H_1: \sigma_1^2 > \sigma_2^2$ atau $F_{hitung} < F_{(1-\alpha,df_1,df_2)}$untuk $H_1: \sigma_1^2 < \sigma_2^2$

3. Uji Varians Tiga atau Lebih Populasi

- **Uji Bartlet**

Uji ini digunakan untuk melihat kesamaan varians dari beberapa populasi yang **berdistribusi normal.** Hipotesa yang digunakan pada uji Bartlett adalah sebagai berikut:

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
$H_1: Setidaknya\ ada\ satu\ \sigma_i^2\ yang\ tidak\ sama$

Statistik Uji

$$B = \frac{1}{C}\left[(N-K)\ ln\ ln\ (MSE) - \sum_{i=1}^{k}(n_i - 1)ln(s_i^2)\right]$$

Dengan

$$C = 1 + \frac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\frac{1}{n_i - 1}\right) - \frac{1}{n-k}\right]$$

$$MSE = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k}$$

Wilayah Kritis. $B > \chi^2_{(\alpha,k-1)}$

- **Uji Levene**

Uji Levene merupakan metode pengujian homogenitas varians yang hampir sama dengan uji Bartlett**. Perbedaan uji Levene dengan uji Bartlett** yaitu bahwa data yang diuji dengan uji Levene **tidak harus berdistribusi normal, namun harus kontinu**. Hipotesis

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
$H_1: Setidaknya\ ada\ satu\ \sigma_i^2\ yang\ tidak\ sama$

Statistik Uji

$$W = \frac{(n-k)\sum_{i=1}^{k}(\underline{Z}_{i.} - \underline{Z}_{..})}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \underline{Z}_{i.})}$$

Dengan

$$Z_{ij} = \left|X_{ij} - \underline{X}_i\right|$$

$\underline{Z}_i$ adalah rata-rata dari kelompok Z_i
$\underline{Z}_{..}$ adalah rata-rata menyeluruh

Wilayah Kritis. $W > F_{(\alpha,k-1,n-k)}$

Fungsi	Kegunaan	Package
`var.test(x1, x2)`	Uji varians dua populasi	stats (Bawaan R)
`bartlett.test(formula, data)`	Uji Bartlett	stats (Bawaan R)
`leveneTest(formula, data)`	Uji Levene	car

LATIHAN SOAL

Dengan menggunakan `data_dummy_komstat.csv` dan $\alpha = 0.05$

1) Ujilah apakah variabel sebelum memiliki $\sigma = 10$?
2) Ujilah apakah variabel sebelum dan sesudah memiliki varians yang sama?
3) Jika variabel sebelum dan sesudah masing-masing dikelompokkan berdasarkan metode apakah memiliki varians yang sama?

Link Source Code

https://github.com/modul60stis/komstat-uas

"Kenapa aku mengharapkan dunia yang berubah? **Seharusnya akulah yang menyesuaikan** dan dengan begitu bisa mengubah duniaku."

— Ahmad Fuadi, Ranah 3 Warna

UJI HIPOTESIS RATA-RATA

Uji beda T (Uji T) adalah salah satu teknik analisis dalam ilmu statistika yang **digunakan untuk** mengetahui signifikansi perbedaan dan membuat kesimpulan tentang suatu populasi berdasarkan data dari sampel yang diambil dari populasi itu. Teknik uji beda t dilakukan atas **data rasio atau interval**. Teknik yang dilakukan dengan **membandingkan nilai mean**. Statistik uji ini digunakan dalam pengujian hipotesis. **Asumsi/syarat** uji –t :

- Data berdistribusi normal
- Skala rasio/Interval
- Sampel independen

1. Uji Hipotesisi 1 Populasi

Hipotesis

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0 \; atau \; \mu < \mu_0 \; atau \; \mu > \mu_0$

- **Varians Tidak Diketahui**

  Statistik Uji

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}; \; df = n - 1$$

  Wilayah Kritis

  - Dua arah, $|t| > t_{(\frac{\alpha}{2},df)}$
  - Satu arah, $t < -t_{(\alpha,df)}$ untuk $H_1: \mu < \mu_0$ atau $t > t_{(\alpha,df)}$ untuk $H_1: \mu > \mu_0$

- **Varians Diketahui**

  Statistik Uji

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  Wilayah kritis

  - Dua arah, $|z| > Z_{(\frac{\alpha}{2},)}$
  - Satu arah, $z < -Z_{(\alpha)}$ untuk $H_1: \mu < \mu_0$ atau $z > Z_{(\alpha)}$ untuk $H_1: \mu > \mu_0$

2. Uji Hipotesisi 2 Populasi Independent

Hipotesis

$H_0: \mu_1 - \mu_2 = d_0$

$H_1: \mu_1 - \mu_2 \neq d_0$ *atau* $\mu_1 - \mu_2 < d_0$ *atau* $\mu_1 - \mu_2 > d_0$

- σ_1 **dan** σ_2 **Diketahui**
  Statistik Uji

$$z = \frac{(\underline{x}_1 - \underline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  Wilayah Kritis
  - ❖ Dua arah, $|z| > Z_{(\frac{\alpha}{2})}$
  - ❖ Satu arah, $z < -Z_{(\alpha)}$ untuk $H_1: \mu_1 - \mu_2 < d_0$ atau $z > Z_{(\alpha)}$ untuk $H_1: \mu_1 - \mu_2 > d_0$

- σ_1 **dan** σ_2 **Tidak Diketahui dan Diasumsikan** $\sigma_1 = \sigma_2$
  Statistik Uji

$$t = \frac{(\underline{x}_1 - \underline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}};\ df = n_1 + n_2 - 2;\ s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

  Wilayah Kritis
  - ❖ Dua arah, $|t| > t_{(\frac{\alpha}{2},df)}$
  - ❖ Satu arah, $t < -t_{(\alpha,df)}$ untuk $H_1: \mu_1 - \mu_2 < d_0$ atau $t > t_{(\alpha,df)}$ untuk $H_1: \mu_1 - \mu_2 > d_0$

- σ_1 **dan** σ_2 **Tidak Diketahui dan Diasumsikan** $\sigma_1 \neq \sigma_2$
  Statistik Uji

$$t = \frac{(\underline{x}_1 - \underline{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}};\ df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

  Wilayah Kritis
  - ❖ Dua arah, $|t| > t_{(\frac{\alpha}{2},df)}$
  - ❖ Satu arah, $t < -t_{(\alpha,df)}$ untuk $H_1: \mu_1 - \mu_2 < d_0$ atau $t > t_{(\alpha,df)}$ untuk $H_1: \mu_1 - \mu_2 > d_0$

3. Uji Hipotesisi 2 Populasi Berpasangan

$H_0: \mu_d = \mu_0$
$H_1: \mu_d \neq \mu_0$ *atau* $\mu_d < \mu_0$ *atau* $\mu_d > \mu_0$

- **Varians Diketahui**
  Statistik Uji

$$z = \frac{x_d - \mu_0}{\frac{\sigma_d}{\sqrt{n_d}}}$$

  Wilayah Kritis
  - ❖ Dua arah, $|z| > Z_{(\frac{\alpha}{2})}$
  - ❖ Satu arah, $z < -Z_{(\alpha,df)}$ untuk $H_1: \mu_d < \mu_0$ atau $z > Z_{(\alpha,df)}$ untuk $H_1: \mu_d > \mu_0$

- **Varians Tidak Diketahui**

Statistik Uji

$$t = \frac{\bar{x}_d - \mu_0}{s_d/\sqrt{n_d}};\ df = n_d - 1$$

Wilayah Kritis

- Dua arah, $|t| > t_{(\frac{\alpha}{2},df)}$
- Satu arah, $t < -t_{(\alpha,df)}$ untuk $H_1: \mu_d < \mu_0$ atau $t > t_{(\alpha,df)}$ untuk $H_1: \mu_d > \mu_0$

Fungsi	Kegunaan	Package
`t.test(x, mu = mu0)`	Uji hipotesis rata-rata satu populasi	stats (Bawaan R)
`t.test(x, y, mu = mu0, paired = FALSE, var.equal = FALSE)`	Uji hipotesis rata-rata dua populasi independent dengan varians berbeda	stats (bawaan R)
`t.test(x, y, mu = mu0, paired = FALSE, var.equal = TRUE)`	Uji hipotesis rata-rata dua populasi independent dengan varians sama	stats (bawaan R)
`t.test(x, y, mu = mu0, paired = TRUE)`	Uji hipotesis rata-rata dua populasi dependent	stats (bawaan R)

LATIHAN SOAL

Dengan menggunakan `data_dummy_komstat.csv`

1) Dengan $\alpha = 0.05$, ujilah apakah variabel sebelum memiliki $\mu > 65$, jika diketahui $\sigma = 10$?
2) Dengan $\alpha = 0.08$, Ujilah apakah variabel sesudah memiliki $\mu < 65$, jika σ tidak diketahui?
3) Dengan $\alpha = 0.10$, ujilah apakah variabel sesudah dengan metode A memiliki rata-rata yang sama dengan variabel sesudah dengan metode C?
4) Dengan $\alpha = 0.05$, ujilah apakah variabel sesudah memiliki rata-rata yang lebih besar daripada variabel sebelum?
5) Dengan $\alpha = 0.01$, ujilah apakah selisih niliai sesudah dan sebelum jika dikelompokkan berdasarkan jenis kelamin memiliki nilai rata-rata yang sama?

UJI PROPORSI

1. Uji Satu Proporsi

Hipotesis

- $H_0: p = p_0;\ H_1: p \neq p_0$
- $H_0: p \leq p_0;\ H_1: p > p_0$
- $H_0: p \geq p_0;\ H_1: p < p_0$

Statistik Uji

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Wilayah Kritis

- Dua arah, $|z| > Z_{\left(\frac{\alpha}{2}\right)}$
- Satu arah, $z < -Z_{(\alpha)}$ untuk $H_1: p < p_0$ atau $z > Z_{(\alpha)}$ untuk $H_1: p > p_0$

2. Uji Dua Proporsi

Hipotesis

- $H_0: p_1 = p_2;\ H_1: p_1 \neq p_2$
- $H_0: p_1 \leq p_2;\ H_1: p_1 > p_2$
- $H_0: p_1 \geq p_2;\ H_1: p_1 < p_2$

Statistik Uji

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}};\ \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Wilayah Kritis

- Dua arah, $|z| > Z_{\left(\frac{\alpha}{2}\right)}$
- Satu arah, $z < -Z_{(\alpha)}$ untuk $H_1: p_1 < p_2$ atau $z > Z_{(\alpha)}$ untuk $H_1: p_1 > p_2$

Fungsi	Kegunaan	Package
`prop.test(x, n, p = p0, alternative = "two.sided")`	Uji hipotesis proporsi satu atau dua populasi	stats (Bawaan R)

LATIHAN SOAL

Dengan menggunakan `data_dummy_komstat.csv` dan tingkat signifikansi 5%.

1) Jika dilihat berdasarkan nilai sesudah apakah proporsi mahasiswa yang puas lebih dari 0.6?
2) Jika dilihat berdasarkan nilai sesudah apakah mahasiswa yang puas pada setiap metode pengajaran memiliki proporsi yang sama?

ANOVA

1. One-Way Anova

Pengujian ini **digunakan** pada data kuantitatif **untuk menentukan** apakah ada perbedaan diantara beberapa populasi rata-ratanya. Nama dari pengujian yaitu ANOVA (*analysis of variances*) diambil berdasarkan cara perhitungan yang digunakan yaitu suatu teknik yang menganalisa variasi data untuk menentukan apakah kita dapat menyatakan ada perbedaan diantara rata-rata populasi yang kita teliti. **Asumsi pada uji Anova** ini **adalah data bersifat independent, berdistribusi normal**, dan **memiliki varians yang sama. Jika asumsi** kenormalan dan varians **tidak terpenuhi** maka uji ANOVA **dapat diganti** dengan **uji Kruskal-Wallis** (uji non-parametrik). **Hipotesa** untuk pengujian k populasi ditulis sebagai berikut:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$
$H_1: Setidaknya\ ada\ satu\ \mu_i\ yang\ tidak\ sama$

Statistik Uji

$$F = \frac{MSB}{MSW}$$

Source	SS	df	MS	F
Between	SSB	k-1	$MSB = \frac{SSB}{k-1}$	$\frac{MSB}{MSW}$
Within	SSW	n-k	$MSW = \frac{SSW}{n-k}$	
Total	SSTO	n-1		

$$SSB = \sum_{i=1}^{k} n_i(\underline{x}_i - \bar{\bar{x}})^2$$

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(x_{ij} - \underline{x}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_j^2$$

$$SSTO = SSB + SSW$$

Wilayah Kritis. $F > F_{(\alpha,k-1,n-k)}$

2. Two-Way Anova

Source	SS	df	Mean Squre	F
Faktor A	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Faktor B	SSB	$b-1$	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Interaksi	SSAB	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	SSE	$N-ab$	$MSE = \frac{SSE}{N-ab}$	
Total	SSTO	$N-1$		

$$SSA = bn\sum_{i=1}^{a}(\underline{x}_{i..} - \underline{x}_{...})^2 \;\; ; \;\; SSB = an\sum_{i=1}^{b}(\underline{x}_{j.} - \underline{x}_{...})^2 ; \;\; SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(x_{ijk} - \underline{x}_{ij.})$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(x_{ijk} - \underline{x}_{...}) \;\; ; \;\; SSAB = SSTO - SSA - SSB - SSE$$

- **Cek Efek Faktor A**

  H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_a$

  H_1: *Setidaknya ada satu* α_i *yang tidak sama*

Statistik Uji.

$$F_A = \frac{MSA}{MSE}$$

Wilayah Kritis. $F_A > F_{(\alpha,a-1,N-ab)}$

- **Cek Efek Faktor B**

$H_0: \beta_1 = \beta_2 = \cdots = \beta_b$
$H_1: Setidaknya\ ada\ satu\ \beta_i\ yang\ tidak\ sama$

Statistik Uji.

$$F_B = \frac{MSB}{MSE}$$

Wilayah Kritis. $F_B > F_{(\alpha,b-1,N-ab)}$

- **Cek Efek Interaksi**

$H_0: \alpha\beta_{11} = \alpha\beta_{12} = \cdots = \beta_{ab}$
$H_1: Setidaknya\ ada\ satu\ \beta_{ab}\ yang\ tidak\ sama$

Statistik Uji.

$$F_{AB} = \frac{MSAB}{MSE}$$

Wilayah Kritis. $F_{AB} > F_{(\alpha,(a-1)(b-1),N-ab)}$

3. Uji Perbandingan Ganda (*Posthoc Test*)

Jika dalam pengujian ANOVA H_0 **ditolak**, maka untuk mengetahui seberapa besar pengaruhnya, maka **dilakukan uji menggunakan analisis perbandingan ganda**. **Syarat** dapat dilakukannya pengujian perbandingan ganda ini adalah **jumlah level faktornya (perlakuan) lebih dari dua**. Salah satu uji perbandingan ganda adalah **uji Tukey**.

Uji Tukey atau disebut juga dengan Tukey Honestly Significant Difference (HSD) merupakan pengujian perbandingan berbagai kelompok rata-rata. Uji ini biasanya digunakan pada sampel besar. Uji Tukey HSD **menggunakan** statistik range studentized untuk membuat semua perbandingan berpasangan antar goup dan menentukan tingkat kesalahan kelompok percobaan untuk membuat perbandingan berpasangan. **Nilai Tukey biasa diberi simbol** ω **(omega).**

$$\omega = q_\alpha(k, n-k)\sqrt{\frac{MSE}{n_g}}$$

Dengan
$q_\alpha(k, n-k)$ adalah nilai kritis dari studentized range dengan k adalah katagori dan $n-k$ adalah derajat bebas
n_g adalah jumlah observasi setiap k katagori.
Jika jumlah observasi setiap katagori sama, maka $n_g = n_1 = n_2 = \cdots = n_k$
Jika tidak sama, maka digunakan rata-rata harmonic

$$n_g = \frac{k}{\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k}}$$

Kita dapat menyimpulkan **bahwa** μ_i **dan** μ_j **berbeda jika** $|\underline{x}_i - \underline{x}_j| > \omega$

Fungsi	Kegunaan	Package
`aov(formula, data)`	Uji anova	stats (Bawaan R)
`TukeyHSD(aov.model)`	Uji perbandingan ganda Tukey	stats (Bawaan R)

LATIHAN SOAL

Dengan `data_dummy_komstat.csv` dan $\alpha = 0.025$

1) Apakah ada pengaruh yang berbeda dari metode pengajaran terhadap nilai yang dipeorleh (variabel sesudah)?
2) Apakah metode pengajaran dan jenis kelamin mempengaruhi nilai yang diperoleh? Liat juga interaksi antara metode pengajaran dan jenis kelamin!

KORELASI

1. Korelasi Pearson (*Pearson Correlation Coefficient*)

Koefisien korelasi Pearson dapat digunakan untuk menyatakan besar hubungan linier antara dua variabel ketika data adalah **data kuantitatif** (data berskala interval atau rasio) dan kedua variabel adalah bivariat yang **berdistribusi normal**. Formula korelasi Pearson:

$$r = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}}$$

2. Uji Hipotesis Korelasi Pearson

Hipotesis

H_0: Tidak terdapat korelasi antar dua variabel ($\rho = 0$)

H_1: Terdapat korelasi antar dua variabel ($\rho \neq 0$)

Statistik Uji

$$t_{hitung} = \frac{r}{\sqrt{1-r}}\sqrt{n-2}\,; dengan\ df = n-2$$

Wilayah Kritis

$$|t_{hitung}| > t_{(\frac{\alpha}{2},\ df)}$$

3. Korelasi Spearman (*Spearman Rank Correlation Coefficient*)

Statistik ini merupakan suatu ukuran asosiasi atau hubungan yang dapat digunakan pada kondisi satu atau kedua variabel yang diukur adalah **skala ordinal** (berbentuk ranking) **atau** kedua variabel adalah **kuantitatif namun kondisi normal tidak terpenuhi**. Formula korelasi Spearman:

$$r_s = 1 - \frac{6\sum d_i^2}{N^3 - N}$$

$$r_s = \frac{2\left(\frac{N^3-N}{12}\right) - \sum T_1 - \sum T_2 - \sum d_i^2}{2\sqrt{\left(\frac{N^3-N}{12} - \sum T_1\right)\left(\frac{N^3-N}{12} - \sum T_2\right)}}\ dimana\ T = \frac{t^3 - t}{12}$$

4. Uji Hipotesis Korelasi Spearman

Hipotesis

H_0: Tidak terdapat korelasi antar dua variabel ($\rho = 0$)

H_1: Terdapat korelasi antar dua variabel ($\rho \neq 0$)

Statistik Uji

$$Z_{hitung} = r_s\sqrt{n-1}; dengan\ df = n-2$$

Wilayah Kritis

$$|Z_{hitung}| > Z_{(\frac{\alpha}{2},)}$$

Fungsi	Kegunaan	Package
`cor(x, y, method = "pearson")`	Menghitung korelasi pearson x dan y	stats (bawaan R)
`cor(x, y, method = "spearman")`	Menghitung korelasi spearman x dan y	stats (bawaan R)
`cor.test(x, y, method = "pearson")`	Melakukan uji korelasi pearson antara x dan y	stats (bawaan R)
`cor.test(x, y, method = "spearman")`	Melakukan uji korelasi spearman antara x dan y	Stats (bawaan R)

LATIHAN SOAL

Dengan menggunakan `data_dummy_komstat.csv` dan $\alpha = 0.05$

1) Hitung dan ujilah, korelasi dari variabel sebelum dan sesudah.
2) Jika ternyata ada ketambahan 6 sampel baru seperti tabel dibawah, hitung dan uji kembali korelasi dari variabel sebelum dan sesudah.

Sebelum	Sesudah	Jenis Kelamin	Metode	Puas
91	100	Laki-Laki	B	Ya
95	100	Perempuan	D	Ya
97	100	Laki-Laki	A	Ya
98	100	Laki-Laki	C	Ya
98	100	Perempuan	A	Ya
100	100	Perempuan	C	Tidak

ANALISIS REGRESI

Regresi linear adalah sebuah pendekatan untuk **memodelkan hubungan** antara **variable terikat Y** dan satu atau lebih **variable bebas yang disebut X**. Salah satu kegunaan dari regresi linear adalah untuk **melakukan prediksi** berdasarkan data-data yang telah dimiliki sebelumnya. Terdapat **beberapa asumsi** untuk melakukan **analisis regresi**.

- Homoscedastis, suatu kondisi dimana residual bersifat konstan
- Non-Multikolineritas, keadaan dimana setiap pasang variabel x tidak saling berkorelasi yang berarti harus saling independent.
- Normalitas. Uji kenormalan bukan dilakukan pada data x atau y akan tetapi pada residual. Residual dari model yang dibuat harus berdistribusi normal.
- Non-Autokorelasi

Berikut model regresi secara umum

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

Berikut rumus regresi secara umum dengan menggunakan pendekatan matrix.

$$b = [X^T X]^{-1} X^T Y$$

$$b = [b_0\ b_1\ \dots\ b_{p-1}]\ ;\ p = banyaknya\ paramete;\ Y = [y_1\ y_2\ \dots\ y_n]$$

$$X = [1\ x_{11}\ \dots\ x_{1k}\ 1\ x_{21}\ \dots\ x_{2k}\ \dots\ \dots\ \dots\ \dots\ 1\ x_{n1}\ \dots\ x_{nk}]\ ;\ n = banyak\ sampel\ ;\ k = p - 1$$

Tabel Anova

Source	SS	df	MS
Regression	SSR	$k-1$	$MSR = \frac{SSR}{k-1}$
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$
Total	SSTO	$n-1$	

$$SSTO = \sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2\ ;\ SSE = Y^T Y - b^T X^T Y\ ;\ SSR = SSTO - SSE$$

1. Uji Koefisien Regresi Secara Simultan

$H_0: \beta_1 = \beta_2 = \dots = \beta_b$

$H_1: Setidaknya\ ada\ satu\ \beta_i\ yang\ tidak\ sama$

Statistik Uji

$$F = \frac{MSR}{MSE}$$

Wilayah Kritis. $F > F_{(\alpha,k-1,n-k)}$

Jika dalam uji koefisien regresi secara simultan **keputusannya adalah menolak** H_0, maka uji **koefisien regresi secara parsial perlu dilakukan** untuk melihat variabel bebas mana saja yang mempengaruhi variasi dari variabel respon. Hasil dari pengujian secara parsial akan memberikan kesimpulan sedikitnya ada satu variabel bebas yang mempunyai hubungan linier dengan variabel respon. **Jika** dalam uji koefisien regresi secara simultan keputusannya adalah **menerima** H_0, maka **uji secara parsial tidak perlu dilakukan**.

2. Uji Koefisien Regresi Parsial

$H_0: \beta_k = 0$

$H_1: \beta_k \neq 0$

Statistik Uji

$$t = \frac{b_k}{s(b_k)}\ \ ;\ \ s^2(b) = MSE(X^T X)^{-1}$$

$$= [s^2(b_0)\ cov\ \dots\ cov\ cov\ s^2(b_0)\ \dots\ cov\ \dots\ \dots\ \dots\ \dots\ cov\ cov\ \dots\ s^2(b_k)]$$

Wilayah Kritis. $|t| > t_{(\frac{\alpha}{2},n-k)}$

Fungsi	Kegunaan	Package
`lm(formula, data)`	Menghitung membuat model regresi	stats (bawaan R)

`ols_regress(formula, data)`		olsrr
`ols_step_all_possible(model)`	All possible model	olsrr
`bptest(model)`	Breusch-Pagan test untuk cek heteroscedasticity	lmtest
`dwtest(model)`	Durbin-Watson test untuk cek autocorrelation	lmtest
`ols_vif_tol(model)`	Variance Inflation Factors untuk cek multikolineritas	olsrr
`vif(model)`		car

LATIHAN SOAL

Dengan menggunakan data `marketing` dari package `datarium` dan tingkat signifikansi 5%

1) Buatlah model untuk memprediksi nilai sales berdasarkan budged investasi iklan di youtube! Evaluasi model yang dibuat!
2) Buatlah model untuk memprediksi nilai sales berdasarkan bidged investasi iklan di youtube, fb dan koran, kemudian evaluasi model yang dibuat!

METODE RESAMPLING

Sering kita dihadapkan pada permasalahan, hanya mendapatkan jumlah sampel yang kecil dalam suatu pemodelan dan dikhawatirkan parameter yang diperoleh bias, underestimate atau overestimate. Salah satu cara mengatasinya adalah dengan teknik resampling Bootstrap dan Jackknife. **Bootstrap dan Jackknife** adalah **teknik nonparametrik dan resampling** yang **bertujuan untuk** menaksir standar eror dan confidence interval parameter populasi, seperti : mean, median, proporsi, koefisien korelasi dan regresi, dengan tidak selalu memperhatikan asumsi distribusi. Jackknife sendiri adalah alternatif dari bootstrap.

1. Bootstrap

Prosedur Bootstrap :

1. Menarik n-sampel acak dari suatu populasi sebanyak satu kali. (Mengambil sample x)
2. Dari sampel yang didapat, lakukan resample dengan pengembalian (non-parametrik) atau ambil sub-sampel acak dari distribusi teoritis populasi yang diasumsikan dengan estimasi parameter dari data sampel yang didapat (parametrik)
3. Dari sub-sampel yang ada, estimasikan nilai parameter yang ingin didapat, dinotasikan dengan $\hat{\theta}_{b1}, \hat{\theta}_{b2}, \ldots, \hat{\theta}_{bB}$, bentuk umumnya $\hat{\theta}_{bi}$.
4. Ulangi langkah 2 dan 3 sebanyak B kali yang mungkin (direkomendasikan $B = 1000$ atau lebih untuk mengestimasi confidence Interval)
5. Dari semua estimasi parameter sub-sampel ($\hat{\theta}_{b1}, \hat{\theta}_{b2}, \ldots, \hat{\theta}_{bB}$), hitung rata – ratanya untuk mendapatkan estimasi parameter dan standard deviasinya untuk nilai standard error

2. Jacknife

Pada prinsipnya prosedur Jackknife adalah melakukan sampling dari sampel awal x (berukuran n) secara berulang dengan cara menghilangkan pengamatan ke-i, $i = 1, 2, \ldots, n$. Dari prinsip ini menghasilkan sampel-sampel Jackknife : $x(i) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Prosedur Jacknife sebagai berikut.

1. Menarik n-sampel acak dari suatu populasi sebanyak satu kali. (Mengambil sample x)
2. Resample dengan mengeluarkan elemen sampel ke i ($i = 1, 2, \dots, n$). Didapatkan resample ke-i ($i = 1, 2, \dots, n$). Setiap hasil resample akan berukuran (n-1)
3. Perhitungan penaksir setiap hasil resample, didapatkan : $\hat{\theta}_{j1}, \hat{\theta}_{j2}, \dots, \hat{\theta}_{jn}$; bentuk umumnya $\hat{\theta}_{ji}$.
4. Perhitungan penaksir jackknife : $\hat{\theta}_{jacknife} = \hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{ji}$. Dengan kata lain, rata-ratakan penaksir yang didapatkan dari setiap hasil resample.

LATIHAN SOAL

1. Gunakan prosedur bootstrap untuk menduga parameter pada model regresi dari data marketing package datarium
2. Dengan menggunakan `data_dummy_komstat.csv`. Gunakan prosedur Jacknife untuk menduga rata-rata dari variabel sesudah dan sebelum

Link Source Code

https://github.com/modul60stis/komstat-uas

Link `data_dummy_komstat.csv`

https://raw.githubusercontent.com/modul60stis/data/main/data_dummy_komstat.csv

"Jangan gampang terbuai **keamanan dan kemapanan**. Hidup itu kadang **perlu beradu, bergejolak, bergesekan**. Dari gesekan dan kesulitanlah, sebuah pribadi akan terbentuk matang. **Banyak profesi di luar sana**, usahakanlah untuk memilih yang **paling mendewasakan** dan yang **paling bermanfaat buat sesama**. (Kiai Rais)"

— Ahmad Fuadi, Rantau 1 Muara

DATA WRANGLING AND ANALYSIS CHALLENGE

Data wrangling adalah kegiatan penyeragaman data / pembersihan data data "kotor" menjadi data bersih yang siap digunakan. Setelah data bersih maka dapat digunakan untuk melakukan analisis agar dapat memperoleh informasi dari data tersebut.

Pada challenge kali ini, terdapat dua dataset yang berbeda yaitu **data covid-19** dan **playlist lagu spotify**. SIlahkan **pilih salah satu** atau **boleh dua-duanya**, kemudian **ekplorasi** lebih jauh tentang data tersebut. Buatkan **visualisasi**, jika memungkinkan gunakanan juga berbagai proses **inferensia dan uji statistik** kemudian **jelaskan dengan bahasa yang mudah dipahami** apa informasi yang dapat kamu peroleh dari dataset yang kamu analisis. Salah satu poin penting dalam analisis data adalah bagaimana kita menyampaikan hasil analisis kita ke orang lain. Ketika ingin bercerita tentang data posisikan dirimu sebagai orang yang tidak mengerti statistik dan data tersebut.

Oh ya, kamu dapat **mengupload hasil analisis kamu di github angkatan**, untuk teknis uploadnya silahkan baca disini https://github.com/modul60stis/komstat-uas/tree/main/data-wrangling%20-challenge. **Apa keuntungannya?** Tentu saja kamu dapat berbagi teknik analisismu keteman-teman yang lainnya dan mungkin saja ada teman-teman yang lain yang terbantu dengan melihat analisismu. Selain itu kamu juga dapat melatih ketajaman menganalisis data dan mengasah kemampuan bercerita tentang data.

1) **Covid-19 Dataset**

   Dataset ini dapat kamu akses disini https://github.com/modul60stis/data/tree/main/covid-19, dataset yang ada di github angkatan hanya dataset sampai tanggal 24 Desember 2020. Untuk data real time dapat kamu akses disini https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series. Dataset ini terdiri dari 5 file, 3 file data covid-19 secara global dan 2 file data covid-19 di US. Analisis yang mungkin dapat kamu lakukan adalah melihat perkemabangan jumlah kasus, jumlah meninggal, dan jumlah yang sembuh dari hari-kehari, mungkin dapat dibuatkan plot timeseriesnya. Atau mau lebih luar biasa lagi, kamu bisa buat model regresi time-series untuk meramal jumlah kasus. Jika dataset terasa kurang cukup silahkan cari saja dataset lain dan gabungkan dengan dataset ini.

2) **Dataset of Song in Spotify**

   Dataset ini dapat kamu akses disini https://github.com/modul60stis/data/tree/main/spotify-playlist. Dataset ini diperoleh dari Kaggle, https://www.kaggle.com/mrmorj/dataset-of-songs-in-spotify. Dataset ini terdiri dari dua file yaitu file genre_v2.csv dan playlist.csv. File genre_v2.csv terdiri dari 42.305 baris. Analisis yang dapat kamu lakukan mungkin melihat bagaimana karakteristik dari lagu-lagu yang ada di Spotify dan mungkin juga dapat melihat genre apa aja yang ada di Spotif. Jika dataset terasa kurang cukup silahkan cari saja dataset lain dan gabungkan dengan dataset ini.

"Aku hanya orang biasa, tetapi aku **bekerja lebih keras** daripada orang biasa."

— Ahmad Fuadi, Berjalan Menembus Batas

POLITEKNIK STATISTIKA STIS
JAKARTA

UJIAN AKHIR SEMESTER GASAL
TAHUN AKADEMIK 2017/2018

Mata Kuliah	:	Komputasi Statistik
Tingkat	:	II (Dua)
Dosen	:	Tim Dosen
Hari/Tanggal	:	Selasa, 14 Februari 2017
Waktu Ujian	:	120 Menit
Sistem Ujian	:	Buka buku dan boleh memakai leptop, tidak boleh menggunakan handphone\gadget

Kerjakan soal berikut dengan menggunakan R! tuliskan syntax serta output/jawaban pada lembar jawaban yang tersedia. Kumpulkan file R code yang digunakan dalam file *.R kepada dosen melalu PJ paling lambat 1 jam setelah ujian selesai.

Untuk soal no 1, 2, dan 3 gunakan data `SusenasUAS.csv` yang telah dikirimkan sebelumnya.

1) Terdapat opini bahwa rumah tangga dengan ART bekerja pada lapangan usaha industry memiliki rata-rata pengeluaran sebulan lebih besar dibandingkan RT dengan lapangan usaha pertanian. Tetapkan H_0 dan H_1 nya, serta lakukan uji yang sesuai, kemudian interpretasikan hasil yang diperoleh. **(15)**

2) Lakukan uji apakah terdapat perbedaan rata-rata Rata-Rata Pengeluaran sebulan antara status pekerjaan yang berbeda. Interpretasikan hasil yang didapat kemudian lakukan uji lanjutan (posthoc test) sesuai dengan hasil uji tersebut. **(20)**

3) Seorang peneliti berhipotesis bahwa proporsi jenis lantai terluas berbeda antara daerah perkotaan dan pedesaan. Tetapkan H_0 dan H_1 serta lakukan uji yang sesuai untuk menjawab pertanyaan tersebut. **(15)**

4) .
   a) Buatlah sebuah fungsi untuk uji beda proporsi dengan input: Proporsi grup 1 ($P1$), besar sampel grup 1 ($n1$), Proporsi grup 2 ($P2$), besar sampel grup 2 ($n2$), dan taraf signifikansi (α). Fungsi tersebut akan menghasilkan nilai statistik uji, p-value, serta selang kepercayaan. **(25)**

   b) Implementasikan fungsi pada soal 4a pada kasus berikut, serta implementasikan hasilnya.
   Dua orang pegawai A dan B masing-masing telah bekerja sebagai petugas entry selama 10 tahun dan 5 tahun. Pimpinan perusahaan beranggapan bahwa persentase kesalahan entry kedua pegawai tersebut tidak sama yang kemungkinan dikarenakan lama masa kerja yang berbeda. Untuk menguji hipotesis tersebut diambil sampel hasil entry dari 50 kuisioner yang dilakukan oleh masing-masing petugas A dan Dari sampel tersebut petugas A membuat 10% kesalahan sedangkan petugas B 12%. Ujilah hipotesis di atas dengan tingkat signifikansi 1%. **(15)**

5) Tuliskan perbedaan prosedur dan algoritma dari metode resampling Jacknife dan bootstrap. Jelaskan kegunaan kedua metode tersebut. **(10)**

Link Pembahasan Soal 2017/2018

https://github.com/modul60stis/komstat-uas/tree/main/pembahasan-2017-2018

"Kita berdua mungkin punya persamaan. Kita sedang berlari. Aku berlari **menuju sesuatu**. Kamu berlari **menjauhi sesuatu**."

— Ahmad Fuadi, Rantau 1 Muara

"Saya menyangsikan kalimat plesetan **'takkan lari jodoh dikejar'**. Gunung memang tidak akan lari. Tapi jodoh? **Dia punya kaki dan keinginan**, dia bisa berlari-lari kesana-kemari. kemana dia suka. **Bahkan dia bisa hilang**, seperti lenyap ditelan bumi. Atau **dia jatuh ketangan orang lain**."

— Ahmad Fuadi, Rantau 1 Muara

UJIAN AKHIR SEMESTER GASAL
TAHUN AKADEMIK 2018/2019

Mata Kuliah : Komputasi Statistik
Tingkat : II (Dua)
Dosen : Tim Dosen
Hari/Tanggal : Selasa, 14 Februari 2018
Waktu Ujian : 120 Menit
Sistem Ujian : Buka buku dan boleh memakai leptop, tidak boleh menggunakan handphone\gadget dan wifi

Kerjakan soal berikut dengan menggunakan R! tuliskan syntax serta output/jawaban pada lembar jawaban yang tersedia. Kumpulkan file R code yang digunakan dalam file *.R kepada dosen melalu PJ paling lambat 1 jam setelah ujian selesai.

1) Gunakanlah dataset dari R "OrchardSprays". Data tersebut berasal dari sebuah studi yang ingin mengetahui pengaruh dari beberapa perlakuan (treatment) terhadap pengurangan jumlah lebah (*decrease*) yang mendekati sebuah kebun buah.
   Dengan mengasumsikan bahwa variabel pengurangan jumlah lebah (*decrease*) untuk tiap kategorik treatment memiliki varians yang sama dan berdistribusi normal, kerjakanlah hal-hal berikut (dengan menggunakan taraf uji 5%):
   a) Ujilah apakah rata-rata pengurangan jumlah lebah dari treatment A lebih kecil 2.8 dibandingkan rata-rata decrease dari treatment B.
   b) Ujilah apakah treatment C,D,F, dan H memberikan pengaruh yang berbeda terhadap besarnya pengurangan jumlah lebah? Lakukan analisis lanjutan jika diperlukan.

2) Buat fungsi untuk menghitung estimasi Bootstrap dan Jecknife dari korelasi beserta 90% Confidence Intervalnya. Gunakan fungsi tersebut pada data mtcars untuk mendapatkan nilai estimasi korelasi antara variabel mile per gallon (mpg) dan horse power (hp).

3) Sebuah lembaga penelitian melakukan survei ke Rumah Tangga (ruta) pertanian di sebuah kabupaten di Pulau Jawa yang hasilnya dapat ditampilkan pada tabel di bawah ini.
   Berdasarkan tabel jumlah rumah tangga berdasarkan status kemiskinan multidimensi, status pekerjaan KRT, Lapangan Usaha sumber pendapatan dan Kuintil Pendapatan, sesunlah data tabulasi menjadi sebuah data frame (cantumkan syntaxnya), yang nantinya akan memudahkan analisis lebih lanjut. Setelah dibuat data framenya, lakukan pengujian dengan tingkat signifikansinya 5%.
   a) Bagaimana proporsi Rumah Tangga dengan Status Kemiskinan multidimensinya . apakah berbeda-beda atau tidak. Jelaskan secara inferensia
   b) Apakah ada perbedaan proporsi ruta berdasarkan kategori kuintil pendapatan?

						Kuintil Pendapatan		
						40% Terbawah	40% Menengah	20% Teratas
Status Kemiskinan	Tidak Miskin MD	Status Pekerjaan	Bukan Petani	Sumber Pendapatan Utama	Bukan Pertanian	100	434	79
					Pertanian	151	0	0

					Pendapatan lainnya	71	10	0
			Petani	Sumber Pendapatan Utama	Bukan Pertanian	3776	7335	2088
					Pertanian	5070	2726	1253
					Pendapatan Lainnya	2901	224	19
	Miskin MD	Status Pekerjaan	Bukan Petani	Sumber Pendapatan Utama	Bukan Pertanian	85	73	51
					Pertanian	0	30	0
					Pendapatan Lainnya	35	0	0
			Pertanian	Sumber Pendapatan Utama	Bukan Pertanian	270	326	194
					Pertanian	266	502	164
					Pendapatan Lainnya	339	0	0

Link Pembahasan Soal 2018/2019

https://github.com/modul60stis/komstat-uas/tree/main/pembahasan-2018-2019

"Inilah masalahnya, berlagak cuek, merasa tidak cocok, tapi terus penasaran."

— Ahmad Fuadi, Rantau 1 Muara

POLITEKNIK STATISTIKA STIS
JAKARTA

UJIAN AKHIR SEMESTER GENAP
TAHUN AKADEMIK 2019/2020

Mata Kuliah	:	Komputasi Statistik
Tingkat	:	3 KS (Dua)
Dosen	:	Tim Dosen
Hari/Tanggal	:	Jumat, 13 Desember 2019
Waktu Ujian	:	120 Menit
Sistem Ujian	:	GUNAKAN LEPTOP

Kerjakan soal berikut dengan menggunakan R! tuliskan syntax serta output/jawaban pada lembar jawaban yang tersedia. Kumpulkan file R code yang digunakan dalam file *.R kepada dosen melalu PJ paling lambat 1 jam setelah ujian selesai.

1) Dari library(nycflights13), gunakan dataset flights, airports, airlines, planes, dan weather. Kemudian lakukan
   a) Buat bar-chart untuk urutan 10 maskapai penerbangan airlines (lengkap dengan nama maskapainya) dengan rata-rata waktu keterlambatan kedatangan (departure delay) paling lama
   b) Uji apakah terdapat beda rata-rata jarak yang ditempuh dari pesawat yang dibuat sebelum tahun 2000 dan setelah (termasuk) tahun 2000
   c) Uji apakah terdapat beda rata-rata waktu keterlambatan keberangkatan dari empat musim (winter: oct-jan, spring: feb-may, summer: june-sept). Lakukan uji lanjutan jika diperlukan.
   d) Berdasarkan data diatas, maskapai mana yang paling bagus kinerjanya serta berikan alasannya

2) Konsumsi listrik setiap bulan oleh sebuah pabrik kimia dianggap berkorelasi dengan rata-rata suhu sekitar (x_1), jumlah hari dalam bulan (x_2), kemurnian produk rata-rata (x_3) dan jumlah produk yang diproduksi dalam ton (x_4). Data histori pada tahun sebelumnya disajikan dalam tabel dibawah ini:

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105

288	50	25	90	100
261	38	23	89	98

a) Buatlah model regresi linear sederhana berganda yang fit kedata tersebut!
b) Hitung penduga dan selang kepercayaan bootstrap dari model linear berganda pada soal a.
c) Interpertasikan hasil yang dihasilkan

3) Sebuah perusahaan komputer membuat dua buah software antivirus yakni jenis A dan B. Untuk keperluan penelitian, maka 1000 virus diinstal pada dua buah komputer. Setelah kedua komputer diisi dengan virus tersebut, kemudian diinstal antivirus jenis A untuk komputer I dan antivirus jenis B untuk komputer II. Beberapa saat kemudian, diketahui dalam komputer I terdapat 825 virus yang dapat dinonaktifkan dan pada komputer II terdapat 760 virus yang berhasil dinonaktifkan. Tentukan selang kepercayaan 95% bagi beda proporsi kematian virus oleh antivirus jenis A dan B.

Link Pembahasan Soal 2019/2020

https://github.com/modul60stis/komstat-uas/tree/main/pembahasan-2019-2020

"Saya harap ini **bukan** sebuah **'good bye' tapi** cukuplah sebuah **'see you'**."

— Ahmad Fuadi, Rantau 1 Muara

Catatan: