DATA SCIENCE INDONESIA

Melek For Member (MFM) 2020

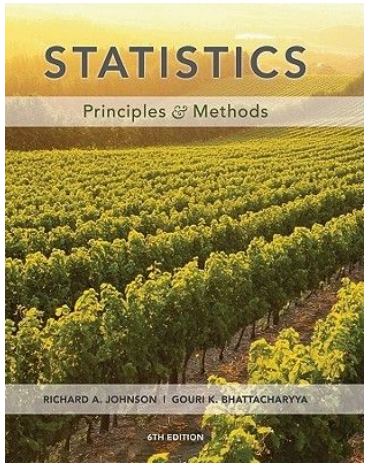
Statistics 2

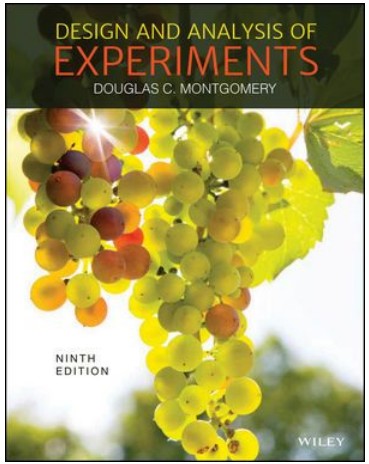
Novri Suhermi, GradStat
PhD Candidate at Lancaster University UK

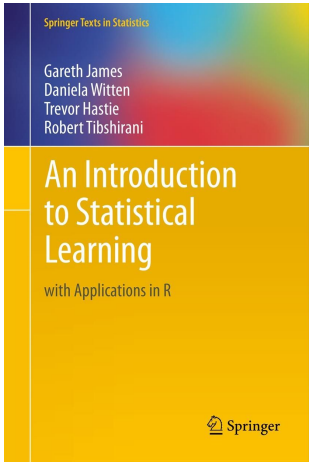
Outline

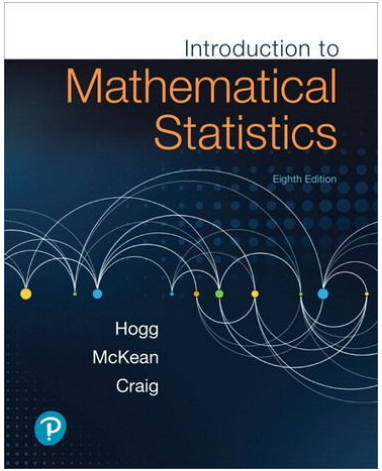
- Sampling Distribution
- Central Limit Theorem
- Parameter Estimation
- Confidence Interval
- Hypothesis Testing
- Analysis of Variance (ANOVA)
- Regression Analysis
- Kernel Density Estimation

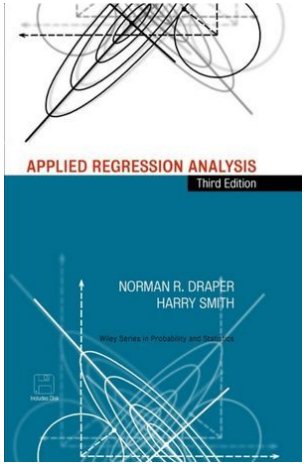
Books to read





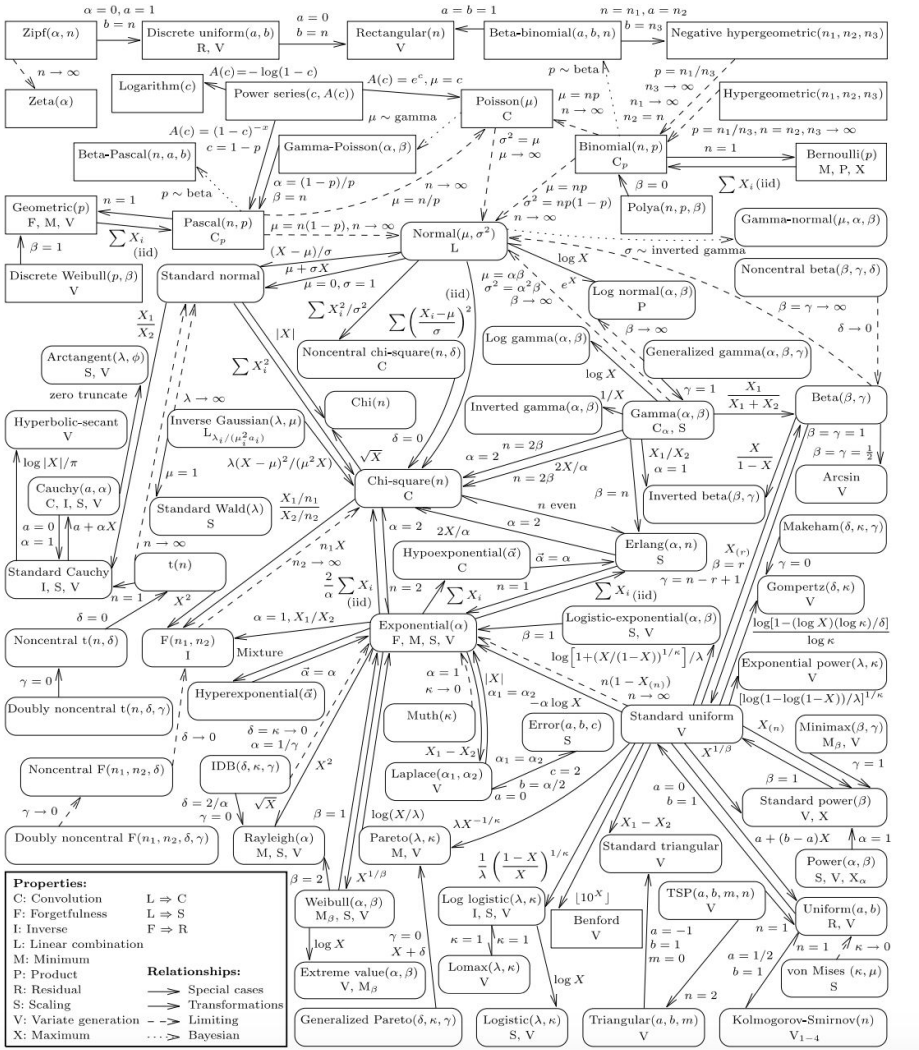






and many more...

Univariate Distribution Relationships



Source: Leemis & McQueston (2008)

Basic Definitions

DATA SCIENCE INDONESIA

Sample

- A sample is a set of observable random variables, $X_1, X_2, \ldots, X_n$. The number n is called the sample size.

Random sample

- A random sample of size n from a population is a set of n independent and identically distributed (iid) observable random variables $X_1, X_2, \ldots, X_n$.

Statistic

- Statistic is a function of observable random variables, $X_1, X_2, \ldots, X_n$ that does not depend on any unknown parameter.

Sampling distribution

Sampling distribution is a probability distribution of a statistic

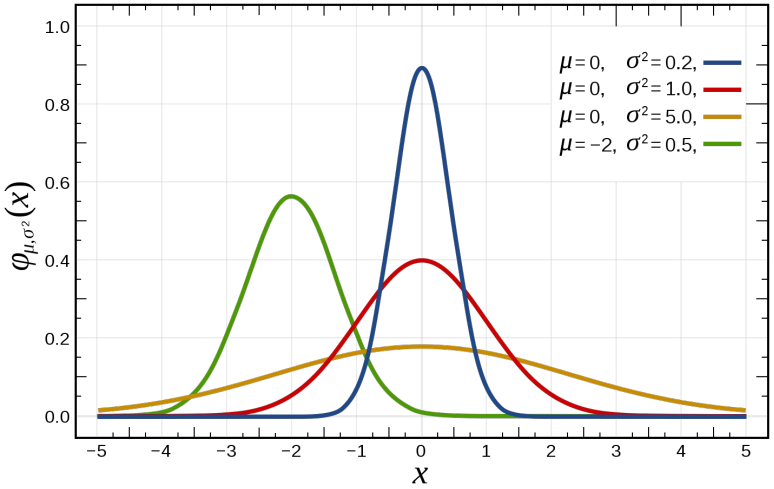
Statistic(s)

- Mean (central measurement / location)

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- Variance (spread measurement / shape)

$$S^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{n-1}$$



Sampling distribution

Let $X_1, X_2, ..., X_n$ be random samples of size n such that

$$\mathrm{E}(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2, i = 1, 2, ..., n$$

Mean, Variance and Standard Deviation of $\bar{X}$

$E(\bar{X}) = \mu$ (=Population mean)

$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ (=Population variance / sample size)

$\mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (=Population variance / sqrt(sample size))

Central Limit Theorem (CLT)

***Whatever the population*, *the distribution of* $\bar{X}$ *is approximately normal when n is large*.**

In random sampling from an arbitrary population with mean μ and standard deviation σ, when *n* is large, the distribution of $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Consequently,

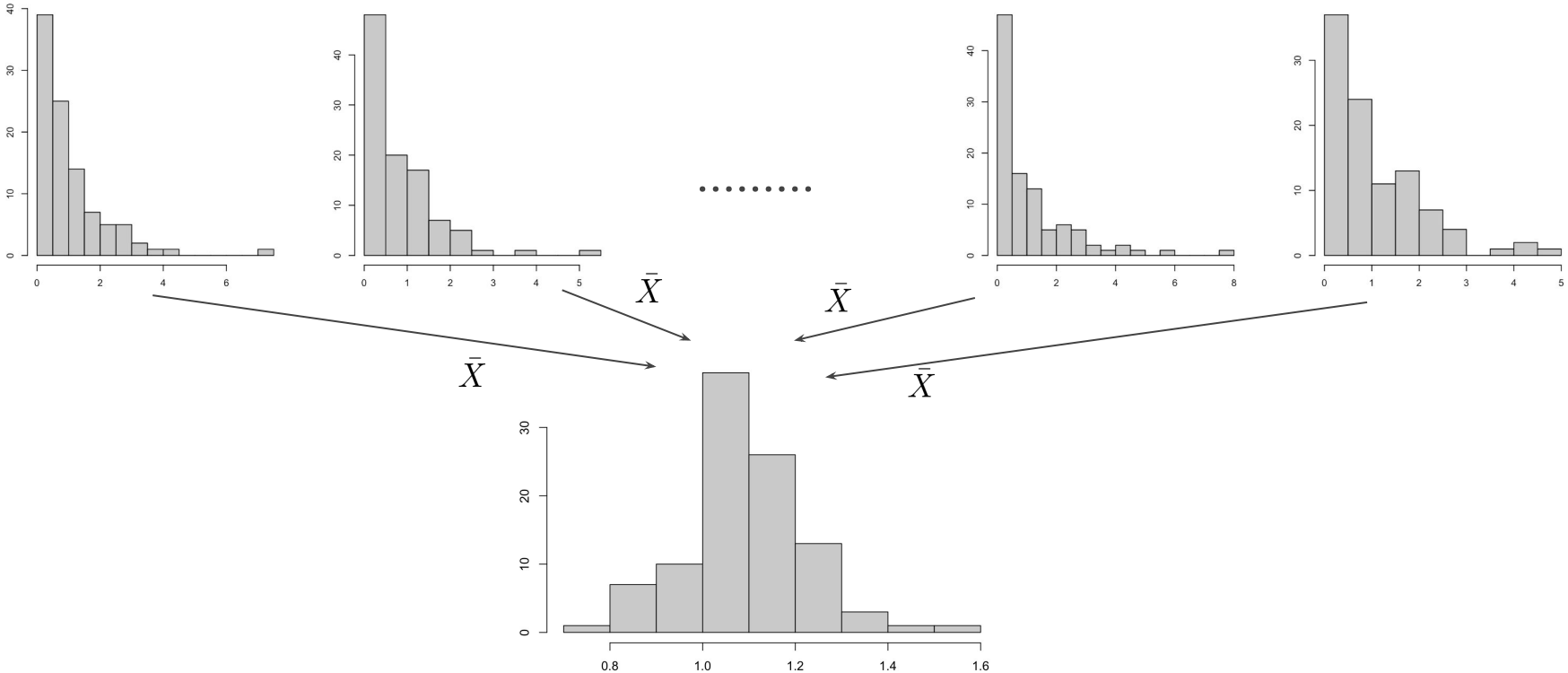
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

with

$$\lim_{n \to \infty} P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

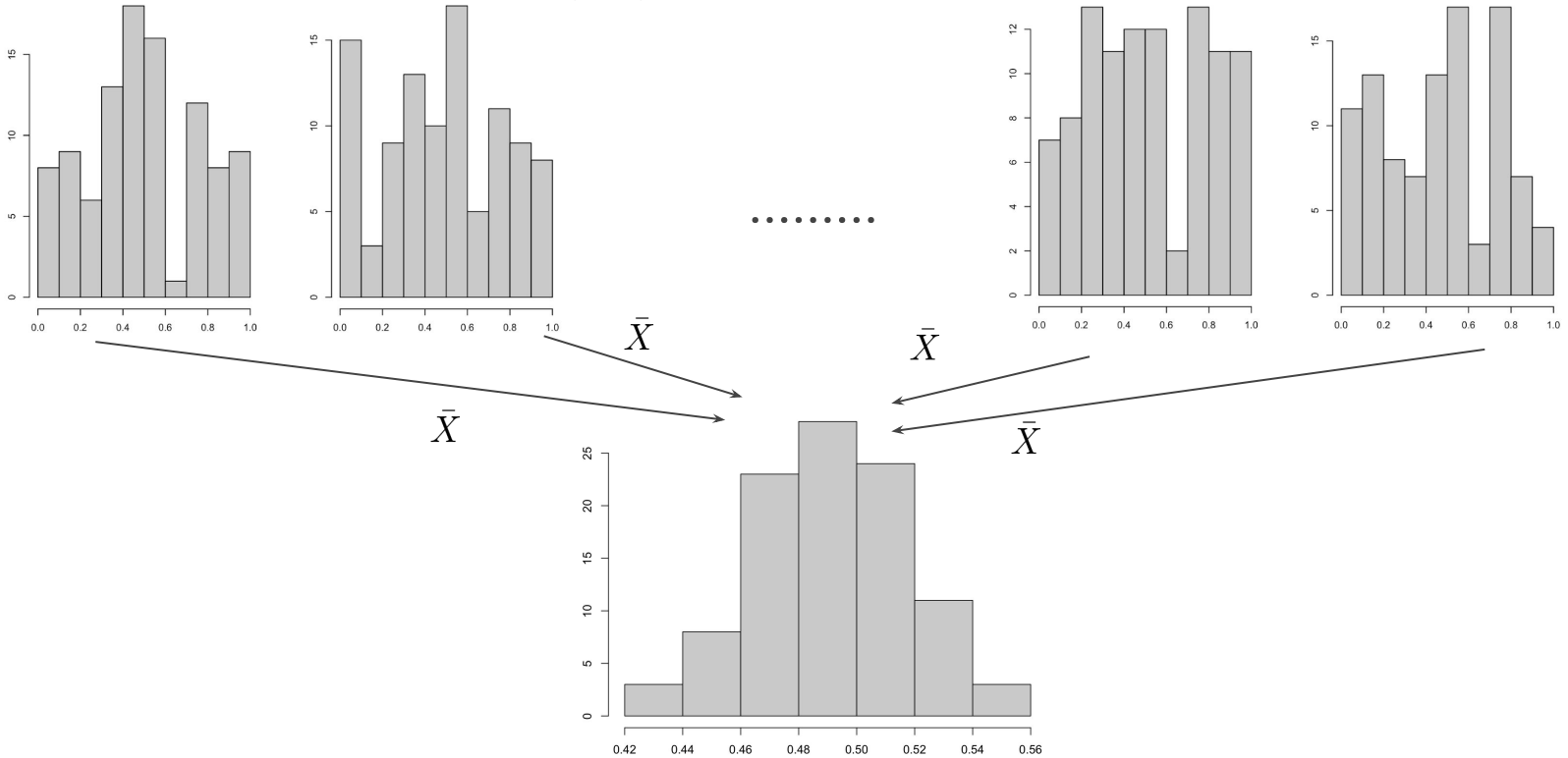
Examples of CLT (1)

Exponential Distribution: X ~ Exp(1)

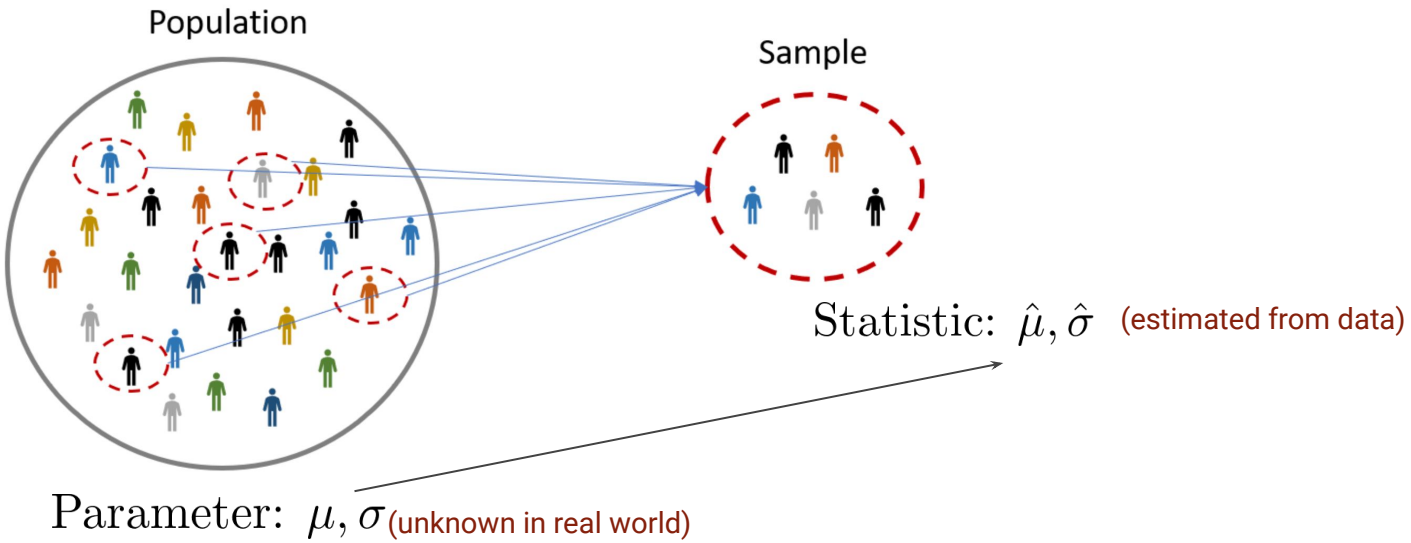


Examples of CLT (2)

Uniform Distribution: X ~ Unif(0,1)



Parameter Estimation



Question: How to calculate $\hat{\mu}, \hat{\sigma}$?

Image source: omniconvert.com

Parameter Estimation

Let $\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n; \theta), \theta \in \Theta \subseteq \mathbb{R}^{\mathbf{k}}$ be the joint probability (or density) function of *n* random variables $X_1, X_2, \ldots, X_n$ with sample values $x_1, x_2, \ldots, x_n$. The **likelihood function** of the sample is given by

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta)$$

Maximum Likelihood Estimation (MLE)

MLE is a method of estimating the parameters of a probability distribution by maximizing a likelihood function

$$L\left(\widehat{\theta}; x_1 \ldots, x_n\right) = \max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

Parameter Estimation

Procedure to find the MLE estimator

1. Define the Likelihood function, $L(\theta)$
2. Take the natural logarithm (ln) of $L(\theta)$
3. Differentiate ln $L(\theta)$ with respect to θ, and then equate the derivative to zero
4. Solve the parameter for θ, and we will obtain $\hat{\theta}$
5. Check whether it is a maximizer or a global maximizer

Parameter Estimation

Example: Normal Distribution case

If $X_1, X_2, \ldots, X_n$ be $N(\mu,\sigma^2)$ where μ and σ^2 are both unknown, find the MLE for μ and σ^2.

To avoid notational confusion, let $\theta = \sigma^2$. Then the likelihood function is:

$$L(\mu, \theta) = (2\pi\theta)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\theta}\right)$$

$$\ln L(\mu, \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\theta}$$

Parameter Estimation

Example: Normal Distribution case

We need to differentiate with respect to both μ and θ individually:

$$\frac{\partial \ln L(\mu,\theta)}{\partial \mu} = \frac{2\sum_{i=1}^{n}(x_i-\mu)}{2\theta}$$

$$\frac{\partial \ln L(\mu,\theta)}{\partial \theta} = \frac{-n}{2\theta} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta^2}$$

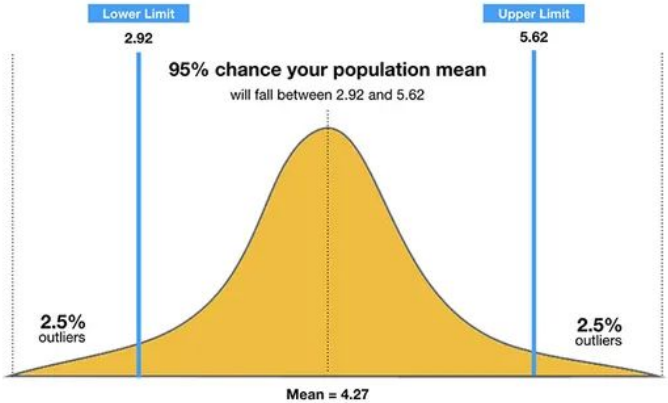
Setting the derivatives equal to zero and solving simultaneously, we obtain:

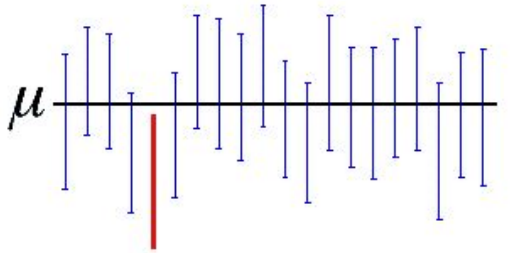
$$\hat{\mu} = \bar{X}$$

$$\widehat{\sigma}^2 = \hat{\theta} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n} = S^2$$

Confidence Interval

A confidence interval is a range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter.





A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

Image source: simplypsychology.org, blog.minitab.com

Confidence Interval

Large-sample confidence intervals

If the sample size is large, then by the CLT, certain sampling distributions can be assumed to be approximately normal. That is, if θ is an unknown parameter (such as μ, p, (μ_1-μ_2), (p_1-p_2)), then for large samples, by the CLT, the z-transform:

$$z = \frac{\widehat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim \mathcal{N}(0, 1)$$

Rule of thumb:

- for $\theta = \mu$, $n \geq 30$ is considered large
- for the binomial parameter p, n, is considered large if np and $n(1-p)$ are both ≥ 5

Confidence Interval

Procedure to calculate large-sample confidence interval

- Find an estimator of θ, say $\hat{\theta}$ and obtain standard error $\sigma_{\hat{\theta}}$
- Calculate the z-transform
- Find two tail values -$z_{\alpha/2}$ and $z_{\alpha/2}$
- An approximate (1-α)100% confidence intervals (CI) for θ is given by

$$P\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \widehat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

- Conclusion: we are (1-α)100% confident that the true parameter θ lies in the interval

$$\left(\widehat{\theta} - Z_{\alpha/2}\sigma_{\hat{\theta}}, \widehat{\theta} + Z_{\alpha/2}\sigma_{\hat{\theta}}\right)$$

Confidence Interval

Confidence intervals for mean

$$P\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Example

Two statistics professors want to estimate average scores for an elementary statistics course that has two sections. Each professor teaches one section and each section has a large number of students. A random sample of 50 scores from each section produced the following results

Section I: $\bar{x}_1 = 77.01, s_1 = 10.32$
Section II: $\bar{x}_2 = 72.22, s_2 = 11.02$

For α = 0.05, we obtain $z_{\alpha/2}$ = 1.96. The 95% CIs are

$$\bar{x}_1 \pm z_{\alpha/2}\frac{s_1}{\sqrt{n}} = 77.01 \pm 1.96\left(\frac{10.32}{\sqrt{50}}\right)$$

95% CI (74.149, 79.871)

$$\bar{x}_2 \pm z_{\alpha/2}\frac{s_2}{\sqrt{n}} = 72.22 \pm 1.96\left(\frac{11.02}{\sqrt{50}}\right)$$

95% CI (69.165, 75.275)

Confidence Interval

Confidence intervals for proportion

$$P\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1 - \alpha$$

Margin of error and sample size

Margin of error is a common summary of sampling error that quantifies uncertainty about a survey result. In order to achieve a certain level of margin of error, say *d*, we require the minimum sample of size *n* where

$$n = \frac{z^2_{\alpha/2}}{4d^2}$$

Example: Suppose that a local TV station in a city wants to conduct a survey to estimate support for the president's policies on the economy within 3% error with 95% confidence. The minimum sample size we require is

$$n = \frac{z^2_{\alpha/2}}{4d^2} = \frac{(1.96)^2}{4(0.03)^2} = 1067.1 \approx 1068$$

Confidence Interval

Small-sample confidence intervals for μ (n < 30)

Let $X_1, X_2, \ldots, X_n$ be random sample from $N(\mu,\sigma^2)$. We have

$$T = \frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{(n-1)S^2/\left[\sigma^2(n-1)\right]}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

where *T* has a *t* distribution with (n-1) degree of freedom.

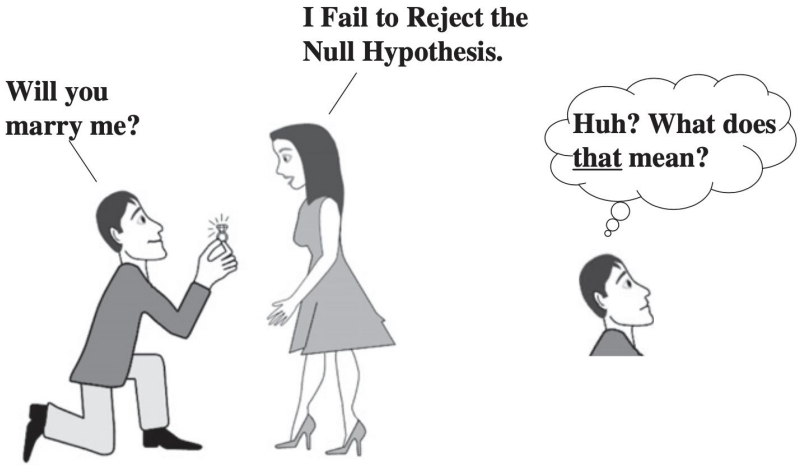
(1-α)100% CI for mean

$$\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

Hypothesis Testing

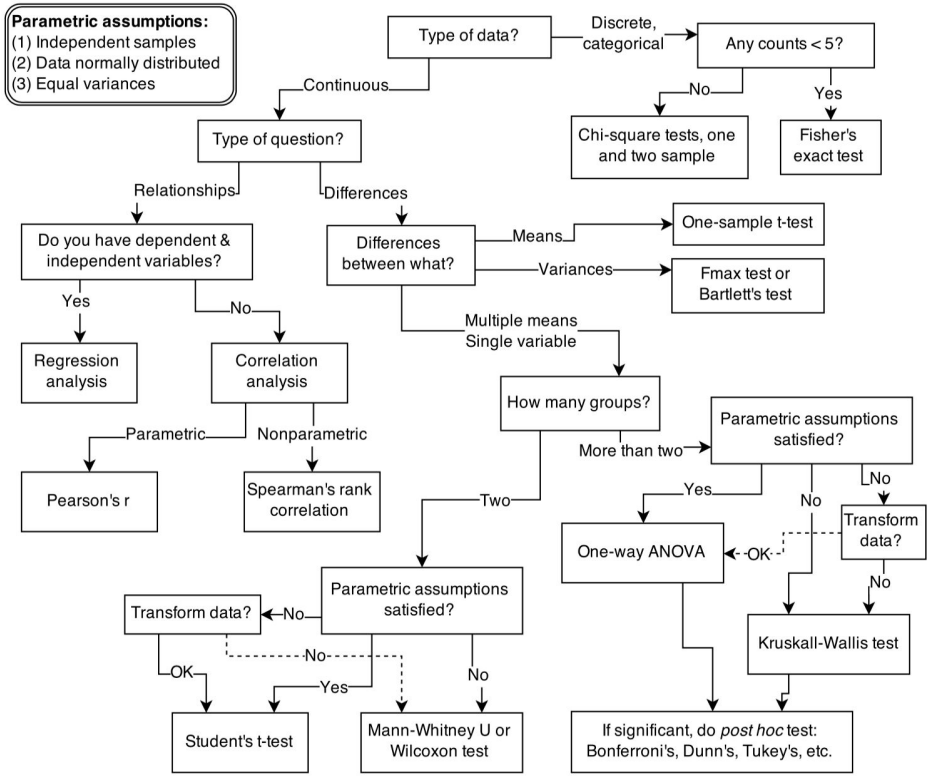
A statistical hypothesis is a statement concerning the probability distribution of a random variable or population parameters that are inherent in a probability distribution.

Hypothesis testing is the procedure that enable us to decide whether to reject hypotheses or to determine whether observed samples differ significantly from expected results



Source: Statistics from A to Z - Confusing Concepts Clarified (2016)

Statistical Procedures



Source: Statistical Rethinking (2020)

Hypothesis Testing

Elements of statistical hypothesis

01	Null hypothesis, H_0	The nullification of a claim. Unless evidence from the data indicates otherwise, the null hypothesis is assumed to be true.
02	Alternative hypothesis, H_1	The opposite of H_0
03	Test statistic	A function of the sample measurements upon which the statistical decision, to reject or not to reject the null hypothesis, will be based
04	Rejection region	The region that specifies the values of the observed test statistic for which the null hypothesis will be rejected
05	Conclusion	• If the value of test statistic falls in rejection region, the null hypothesis is rejected • Otherwise, we cannot reject the null hypothesis

Hypothesis Testing

Formulation of H_0 and H_1

When our goal is to establish an assertion with substantive support obtained from the sample, the negation of the assertion is taken to be the null hypothesis H_0 and the assertion itself is taken to be the alternative hypothesis H_1.

Hypothesis Testing

	Court Trial	Hypothesis testing
Requires strong evidence to establish	Guilt	Conjecture (research hypothesis)
Null hypothesis (H_0)	Not guilty	Conjecture is false
Alternative hypothesis (H_1)	Guilty	Conjecture is true
Attitude	Uphold “not guilty” unless there is a strong evidence of guilt	Retain H_0 unless it makes the sample data very unlikely to happen

False rejection of H_0 is a more serious error than failing to reject H_0 when H_1 is true

Hypothesis Testing

Two Types of Error

Decision based on sample	Unknown True Situation	
	H_0 is true	H_0 is false
Reject H_0	Type I error (α)	Correct decision
Retain H_0	Correct decision	Type II error (β)

P(rejecting H_0 | H_0 is true) = α

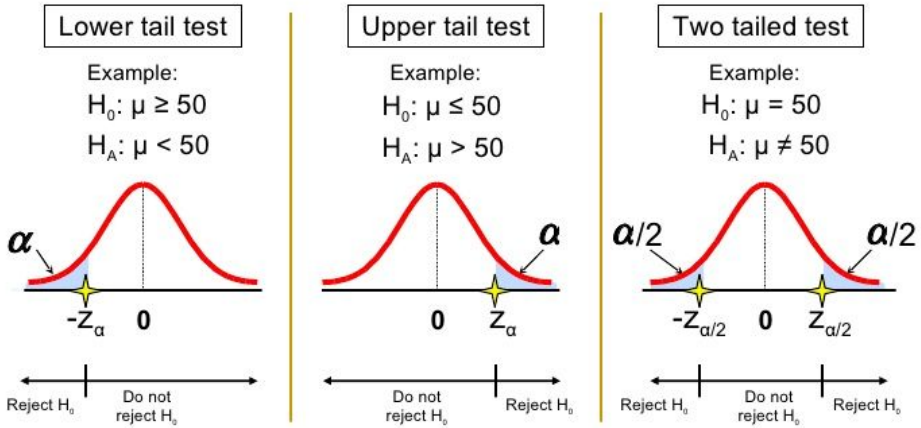
P(not rejecting H_0 | H_0 is false) = β

Power of test = 1 - β

Hypothesis Testing

Level of significance and rejection region

Level of significance = α



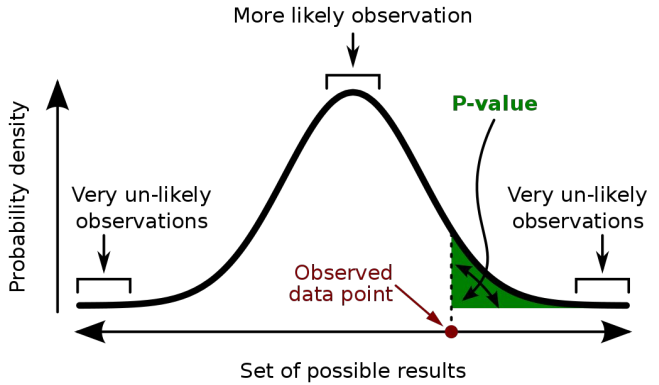
Courtesy: Orhan Erdem

Hypothesis Testing

P-value

P-value is the probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis *H* is correct. The p-value is given by

$$\begin{aligned} &\mathrm{P}(T \geq t \mid H) && \text{for a right tail test} \\ &\mathrm{P}(T \leq t \mid H) && \text{for a left tail test} \\ &2\,\mathrm{P}(T \geq |t| \mid H) && \text{for a two tail test} \end{aligned}$$



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Hypothesis Testing

Hypothesis tests for mean (μ)

Large sample ($n \geq 30$)

$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$ **or** $\mu < \mu_0$ **or** $\mu \neq \mu_0$

Test statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Replace σ with S, if σ is unknown.

Rejection region:
$$\begin{cases} z < z_\alpha, & \text{upper tail} \\ z < -z_\alpha, & \text{lower tail} \\ |z| > z_{\alpha/2}, & \text{two tail} \end{cases}$$

Assumption: $n \geq 30$ and $\sigma^2 < \infty$

Small sample ($n < 30$)

$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$ **or** $\mu < \mu_0$ **or** $\mu \neq \mu_0$

Test statistic: $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Rejection region:
$$\begin{cases} t < t_{\alpha,n-1}, & \text{upper tail} \\ t < -t_{\alpha,n-1}, & \text{lower tail} \\ |t| > t_{\alpha/2,(n-1)}, & \text{two tail} \end{cases}$$

Assumption: Random sample comes from normal population

Hypothesis Testing

Hypothesis test for $\mu_1 - \mu_2$ for large samples (n_1 and $n_2 \geq 30$)

Large sample

$H_0: \mu_1 - \mu_2 = D_0$

$H_1: \mu_1 - \mu_2 > D_0$ or $\mu_1 - \mu_2 < D_0$ or $\mu_1 - \mu_2 \neq D_0$

Test statistic: $$Z = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Replace σ with S, if σ is unknown.

$$\text{Rejection region: } \begin{cases} z < z_\alpha, & \text{upper tail} \\ z < -z_\alpha, & \text{lower tail} \\ |z| > z_{\alpha/2}, & \text{two tail} \end{cases}$$

Assumption: n_1 and $n_2 \geq 30$ and $\sigma^2 < \infty$

Hypothesis Testing

Hypothesis test for μ_1 - μ_2 for small samples

$H_0: \mu_1 - \mu_2 = D_0$

$H_1: \mu_1 - \mu_2 > D_0$ **or** $\mu_1 - \mu_2 < D_0$ **or** $\mu_1 - \mu_2 \neq D_0$

Test statistic: $T = \dfrac{\bar{X}_1 - \bar{X}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ (equal variance)

$T_v = \dfrac{\bar{X}_1 - \bar{X}_2 - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ (unequal variance)

$$S_p^2 = \frac{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}{n_1 + n_2 - 2}$$

$$\text{Rejection Region: } \begin{cases} t > t_\alpha, & \text{upper tail} \\ t < -t_\alpha, & \text{lower tail} \\ |t| > t_{\alpha/2}, & \text{two tail} \end{cases}$$

Degree of freedom for equal variance = (n_1 + n_2 - 2)

Degree of freedom for unequal variance = $\dfrac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$

Hypothesis Testing

Example: bacteria in water

An environmental scientist wants to test if the average number of bacteria per unit of water volume in the river is still below the safe threshold of 200. Then, the scientist collects 10 samples of water and find the number of bacteria. Conduct a statistical test with significance level α = 5%.

$H_0: \mu = 200$
$H_1: \mu < 200$

No	1	2	3	4	5	6	7	8	9	10
Number of bacteria	175	190	215	198	184	207	210	193	196	180

Input:

```
# Implementation in R

data <- c(175, 190, 215, 198, 184, 207, 210, 193, 196, 180)
xbar <- mean(data)
stdev <- sd(data)
test_stat <- (xbar-200)/(stdev/sqrt(length(data)))
pval <- pt(test_stat, df=length(data)-1)
cat('p-value = ', pval)
```

Output:

```
p-value =  0.1211388
```

Hypothesis Testing

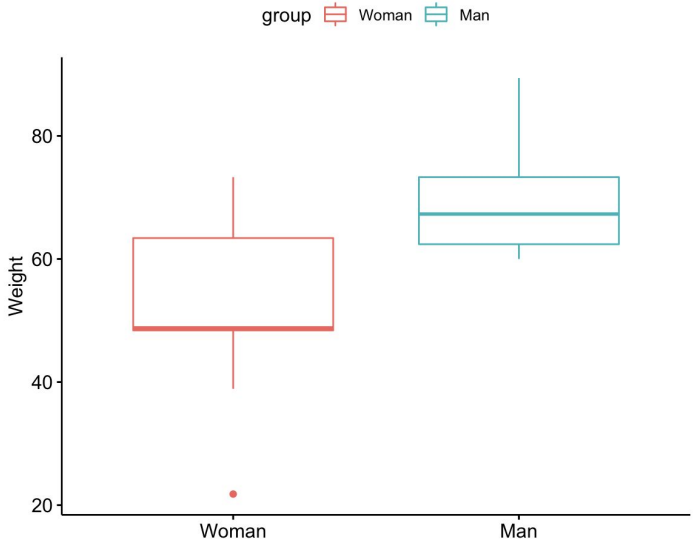
Example: men weight vs women weight

Input:

```
library(ggpubr)
women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
data <- data.frame(
  group = rep(c("Woman", "Man"), each = 9),
  weight = c(women_weight,  men_weight)
)

ggboxplot(data, x = "group", y = "weight", color = "group",
          ylab = "Weight", xlab = "")

results <- t.test(weight ~ group, data = data, var.equal = TRUE)
print(results)
```

Output:

```
Two Sample t-test

data:  weight by group
t = 2.7842, df = 16, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.029759 29.748019
sample estimates:
  mean in group Man mean in group Woman
           68.98889            52.10000
```



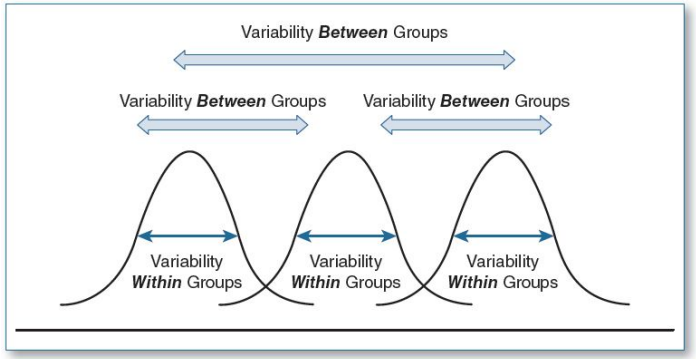
Analysis of Variance (ANOVA)

ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. The null hypothesis is given by

$$H_0: \mu_1 = \mu_2 = \ \ldots = \mu_k$$

In the context of comparing *k* group means, the two sources of variation are

1. differences between group
2. differences within group



Analysis of Variance (ANOVA)

DATA SCIENCE INDONESIA

Linear Model (One-way ANOVA)

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n_i \end{cases}$$

$$N = n_1 + n_2 + ... + n_k$$

Decomposition of the Total Sum of Squares

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = n_i \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

SS Total = SS Treatments + SSE

Analysis of Variance (ANOVA)

One-way ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$\text{SS Treatments} = n_i \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2$	k - 1	SS Treatments / (k-1)	$F_0 = \frac{\text{MS Treatments}}{\text{MSE}}$
Error (within treatments)	$\text{SSE} = \text{SS}_T - \text{SS Treatments}$	N - k	SSE / (N-k)	
Total	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$	N - 1		

F_0 has an F distribution with degree of freedom $v_1 = k - 1$ and $v_2 = N - k$

Analysis of Variance (ANOVA)

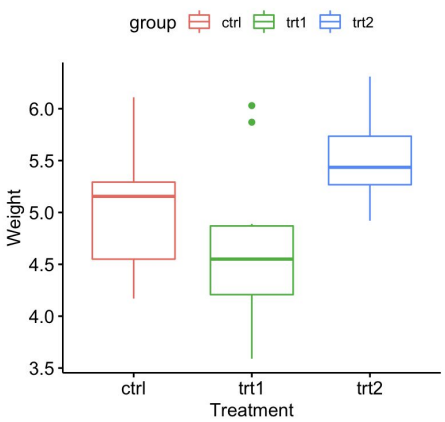
Example: `PlantGrowth` dataset

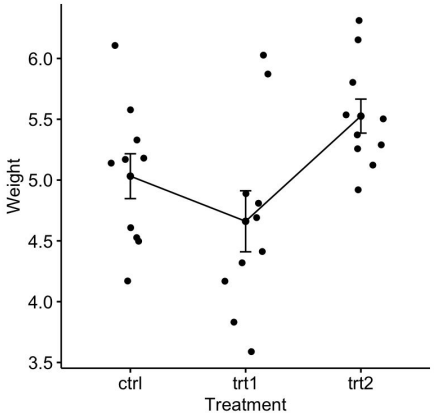
Input:

```
data <- PlantGrowth
ggboxplot(data, x = "group", y = "weight", color = "group",
          ylab = "Weight", xlab = "Treatment")

ggline(data, x = "group", y = "weight",
       add = c("mean_se", "jitter"),
      order = c("ctrl", "trt1", "trt2"),
       ylab = "Weight", xlab = "Treatment")

fit <- aov(weight ~ group, data=data)
summary(fit)
TukeyHSD(fit)
```

Output:

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ group, data = data)

$group
            diff        lwr       upr     p adj
trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```





Regression Analysis

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').

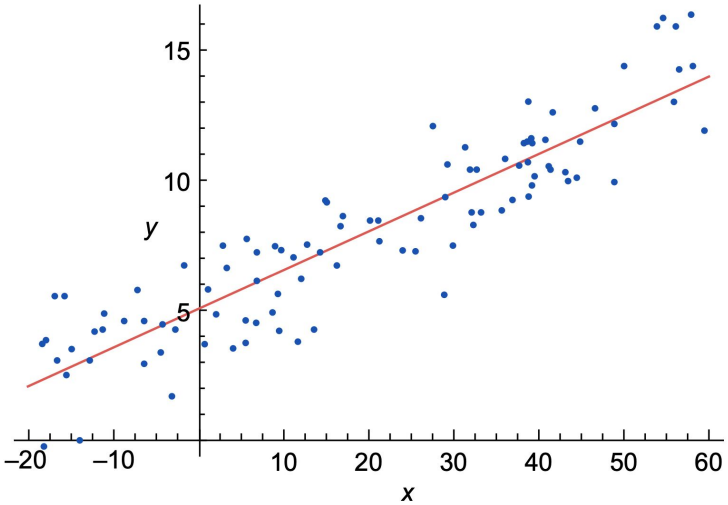
Simple linear regression

$Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim N(0,\sigma^2)$

Multiple linear regression

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (matrix form)



Regression Analysis

Parameter Estimation using Ordinary Least Square (OLS)

Minimizing the sum square error

$$\begin{aligned}\text{SSE} = \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k \right) \right]^2 \\ &= (y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= y'y - y'X\widehat{\beta} - (X\widehat{\beta})'y + (\widehat{\beta}X)'X\widehat{\beta}\end{aligned}$$

$$\frac{\partial}{\partial \beta} \left(y'y - y'X'\beta - \beta'X'y + X'\beta'X\beta \right) = 0$$

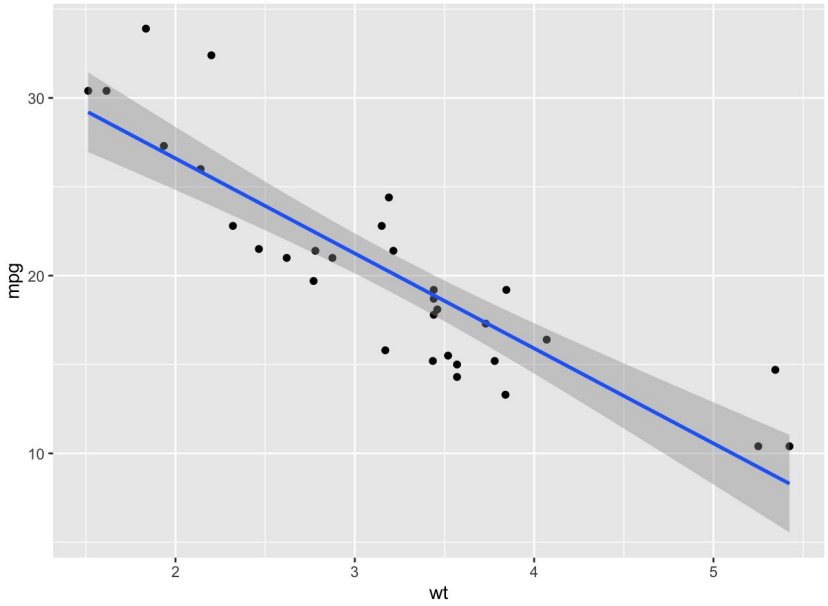
Yielding

$$\widehat{\beta} = \left(X'X \right)^{-1} X'Y$$

Simple Linear Regression

Example: `mtcars` dataset

Output: scatter plot and regression line



Input

```
# Implementation in R

m1 <- lm(mpg ~ wt, data=data)
summary(m1)

ggplot(data, aes(x=wt,y=mpg)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x)
```

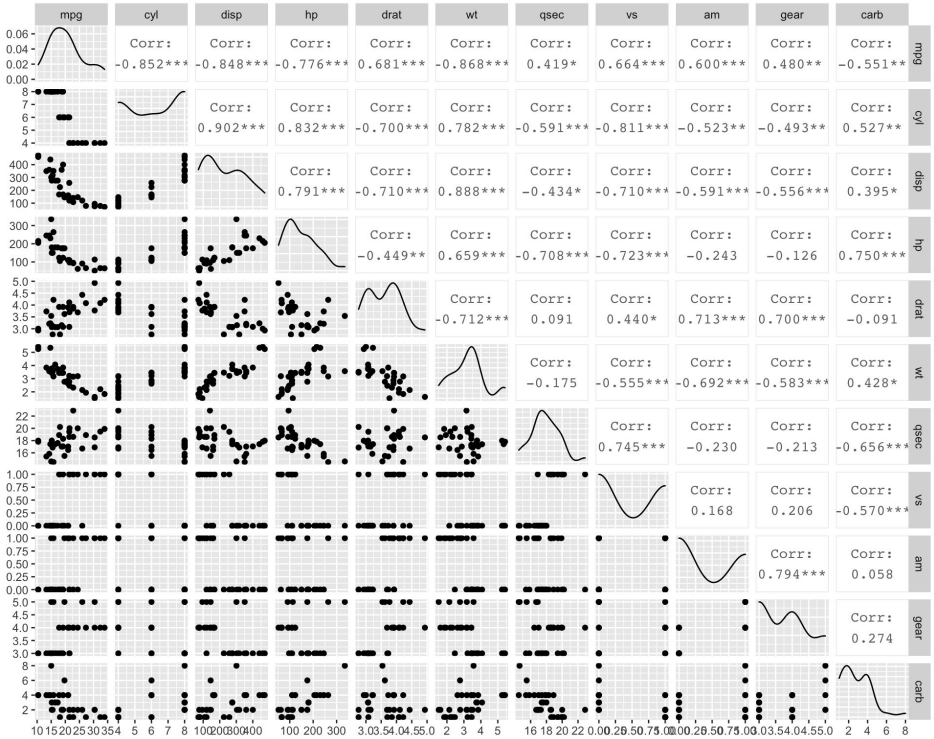
Output

```
Call:
lm(formula = mpg ~ wt, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,  Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Multiple Linear Regression

Example: `mtcars` dataset

Output: Scatter plot matrix



Input

```
# Implementation in R

ggpairs(data)
m2 <- lm(mpg~cyl+disp+hp+drat+wt+qsec,data=data)
summary(m2)
```

Output

```
Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9682 -1.5795 -0.4353  1.1662  5.5272

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.30736   14.62994   1.798  0.08424 .
cyl         -0.81856    0.81156  -1.009  0.32282
disp         0.01320    0.01204   1.097  0.28307
hp          -0.01793    0.01551  -1.156  0.25846
drat         1.32041    1.47948   0.892  0.38065
wt          -4.19083    1.25791  -3.332  0.00269 **
qsec         0.40146    0.51658   0.777  0.44436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.557 on 25 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:   0.82
F-statistic: 24.53 on 6 and 25 DF,  p-value: 2.45e-09
```

Regularization and Feature Selection

DATA SCIENCE INDONESIA

L_1 Regularization (LASSO)

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

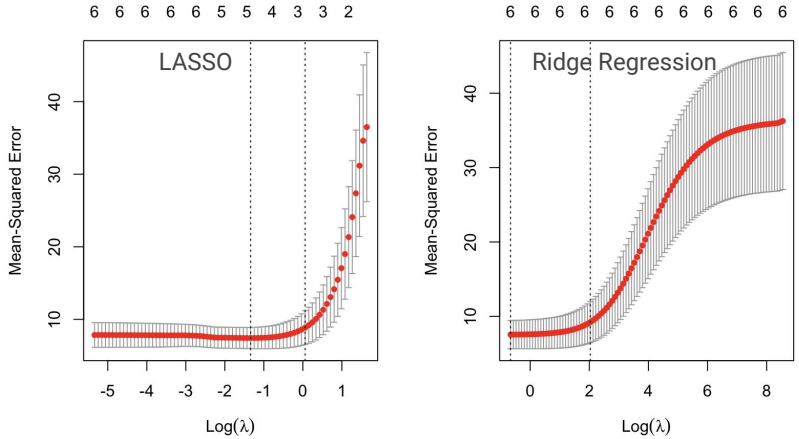
L_2 Regularization (Ridge Regression)

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

Regularization and Feature Selection

Input:

```
library(glmnet)
y <- as.matrix(data %>% select(mpg))
x <- as.matrix(data %>% select(cyl,disp,hp,drat,wt,qsec))

m1 <- glmnet(y=y,x=x)
cvfit1 <- cv.glmnet(x, y)
m2 <- glmnet(y=y,x=x)
cvfit2 <- cv.glmnet(x, y, alpha=0)

par(mfrow=c(1,2))
plot(cvfit1)
plot(cvfit2)
```



Output:

```
> cvfit1$lambda.min
[1] 0.2621943

> coef(cvfit1, s = "lambda.min")
7 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 34.73902220
cyl         -0.80127658
disp         .
hp          -0.01701741
drat         0.47576157
wt          -2.94688035
qsec         0.03224244

> cvfit2$lambda.min
[1] 0.5146981

> coef(cvfit2, s = "lambda.min")
7 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) 26.7941865992
cyl         -0.6462829827
disp        -0.0004221221
hp          -0.0176449845
drat         1.3398924005
wt          -2.7303872625
qsec         0.2210988151
```

Kernel Density Estimation (KDE)

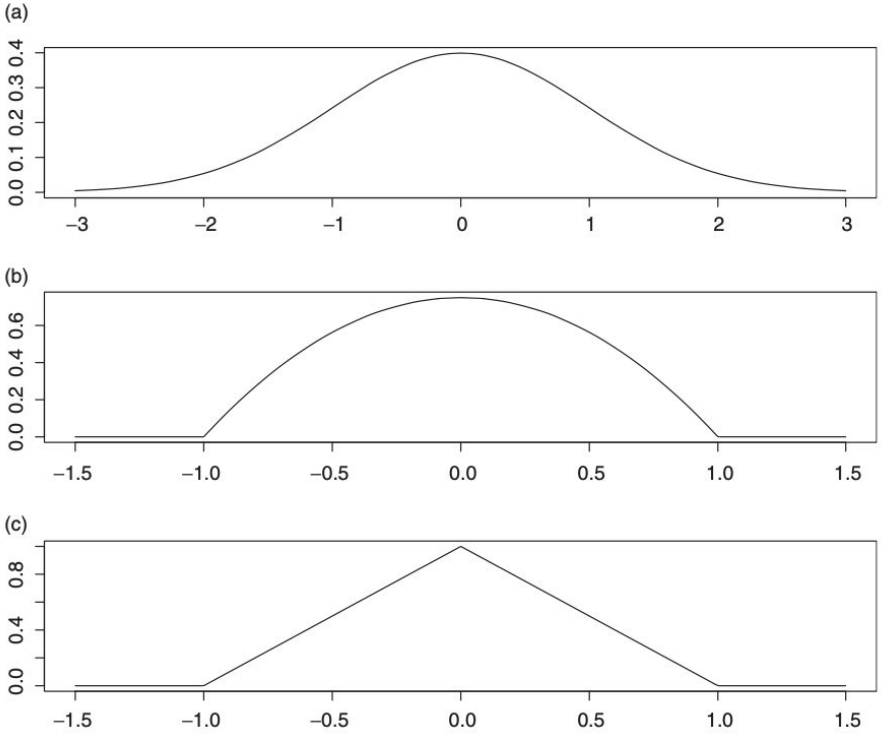
KDE is a nonparametric way to estimate the **probability density function** of a random variable.

Let (x_1, x_2, …, x_n) be a univariate independent and identically distributed sample drawn from some distribution with an unknown density ***f***. We are interested in estimating the shape of this function ***f***. Its kernel density estimator is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where $K(.)$ is a kernel function and $h > 0$ is a smoothing parameter called the bandwidth.

Kernel Functions



(a) Gaussian, (b) Epanechnikov, (c) Triangular

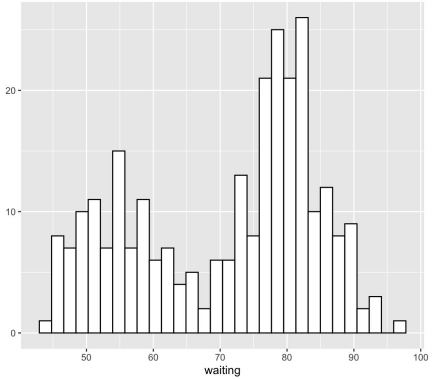
$$\text{Gaussian: } K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

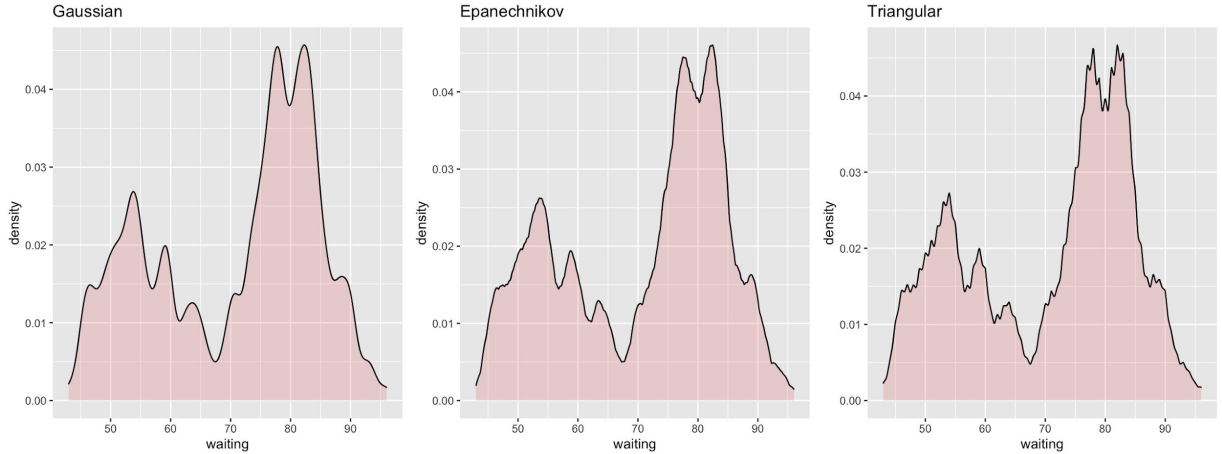
$$\text{Epanechnikov: } K(x) = \frac{3}{4} \cdot \max\left\{1 - x^2, 0\right\}$$

$$\text{Triangular: } K(x) = \max\left\{1 - |x|, 0\right\}$$

Kernel Density Estimation (KDE)

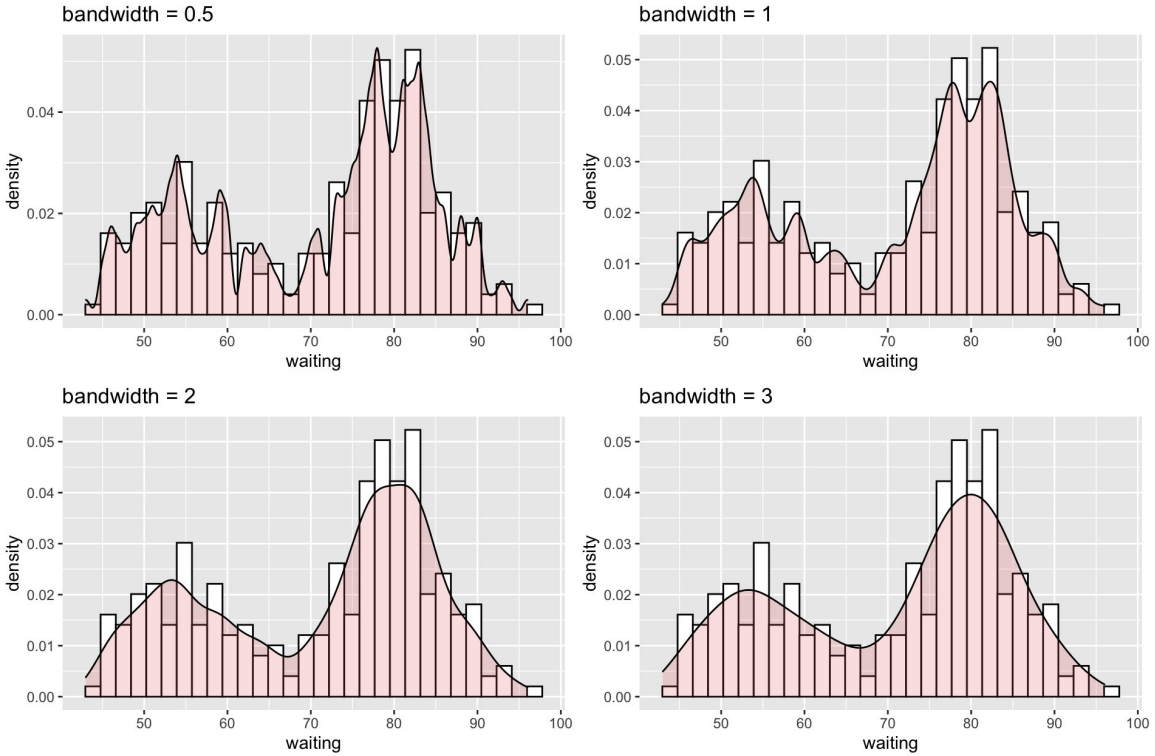
Example: Faithful Dataset





Kernel Density Estimation (KDE)

Bandwidth selection



Bandwidth controls the level of smoothness

We want the one that is not too wiggly nor too smooth

End of slide