

STAT 231 - Statistics

Cameron Roopnarine

Last updated: April 3, 2020

Contents

2020-03-13

Roadmap:

- (i) Recap and the relationship between Confidence and Hypothesis
- (ii) Example: Bias Testing
- (iii) Testing for variance (Normal)
- (iv) What if we don't know how to construct a Test-Statistic?

EXAMPLE 0.0.1. Y_1, \dots, Y_n iid $N(\mu, \sigma^2)$

- $\sigma^2 = \text{known}$
- $\mu = \text{unknown}$
- Sample: $\{y_1, \dots, y_n\}$
- $\bar{y} = \text{sample mean}$
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$

$$D = \left| \frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \rightarrow \text{Test-Statistic (r.v.)}$$

$$d = \left| \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \rightarrow \text{Value of the Test-Statistic}$$

$$p\text{-value} = P(D \geq d) \quad \text{assuming } H_0 \text{ is true}$$

$$= P(|Z| \geq d) \quad Z \sim N(0, 1)$$

Question: Suppose the p -value for the test > 0.05 if and only if μ_0 belongs in the 95% confidence interval for μ ?

YES.

Suppose μ_0 is in the 95% confidence interval for μ , i.e.

$$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\mu_0 \leq \bar{y} + 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\mu_0 \geq \bar{y} - 1.96 \frac{\sigma}{\sqrt{n}}$$

These two equations yield

$$d = \left| \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \leq 1.96$$

$$P(|Z| \geq d) > 0.05$$

General result (assuming same pivot)

p -value of a test $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$ is more than $q\%$, then θ_0 belongs to the $100(1 - q)\%$ confidence interval and vice versa.

EXAMPLE 0.0.2 (Bias). A 10 kg weighted 20 times (y_1, \dots, y_n)

- H_0 : The scale is unbiased
- H_1 : The scale is biased

If the scale was unbiased,

$$Y_1, \dots, Y_n \sim N(10, \sigma^2)$$

If the scale was biased,

$$Y_1, \dots, Y_n \sim N(10 + \delta, \sigma^2)$$

- $H_0: \delta = 0$ (unbiased)
- $H_1: \delta \neq 0$ (biased)

is equivalent to

- $H_0: \mu = 10$
- $H_1: \mu \neq 10$

Test-statistic:

$$D = \left| \frac{\bar{Y} - 10}{\frac{s}{\sqrt{n}}} \right|$$

Compute d .

$$d = \left| \frac{\bar{y} - 10}{\frac{s}{\sqrt{n}}} \right|$$

$$\begin{aligned} p\text{-value} &= P(D \geq d) \\ &= P(|T_{19}| \geq d) \end{aligned}$$

EXAMPLE 0.0.3 (Draw Conclusions). $Y_1, \dots, Y_n = \text{co-op salaries}$. $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$

- $H_0: \mu = 3000$
- $H_1: \mu < 3000$ ($\mu \neq 3000$)

$$D = \left| \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \right|$$

$$D = \begin{cases} 0 & \bar{Y} > \mu_0 \\ \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} & \bar{Y} < \mu_0 \end{cases}$$

If n is large, then

$$Y_1, \dots, Y_n \sim f(y_i; \theta)$$

- $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0$

$$\Lambda(\theta) = -2 \ln \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right]$$

where Λ satisfies all the properties of D . Also,

$$\lambda(\theta) = -2 \ln \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right]$$

and

$$p\text{-value} = P(\Lambda \geq \lambda) = P(Z^2 \geq \lambda)$$

2020-03-16

Roadmap:

- (i) General info

(ii) Testing for variance for Normal

(iii) An example

The general problem: $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ iid where μ and σ^2 are both unknown. $H_0: \sigma^2 = \sigma_0^2$ vs two sided alternative.

(i) Test statistic? Problem

(ii) Convention?

The pivot is:

$$U = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

can we use this as our test statistic?

EXAMPLE 0.0.4.

- Normal population: $\{y_1, \dots, y_n\}$
- $n = 20$, $\sum y_i = 888.1$, $\sum y_i^2 = 39545.03$
- $H_0: \sigma = 2$
- $H_1: \sigma \neq 2$

What is the p -value? We know

$$s^2 = \frac{1}{n-1} \left[\sum y_i^2 - n\bar{y}^2 \right] = 5.7342$$

Compute U :

$$U = \frac{(n-1)s^2}{\sigma_0^2} = 27.24$$

χ_{19}^2

$$\begin{aligned} p\text{-value} &= 2P(U \geq 27.24) \\ &= 2P(\chi_{19}^2 \geq 27.24) \\ &= 10\% \text{ and } 20\% \end{aligned}$$

so, $p > 0.1$ means there is no evidence against null-hypothesis.

2020-03-18

Roadmap:

(i) 5 min recap

(ii) LTRS for large n

(iii) An example

Y_1, \dots, Y_n iid $\sim N(\mu, \sigma^2)$

- $H_0: \sigma^2 = \sigma_0^2$
- $U = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$

We calculated the p -value:

$$U = \frac{(n-1)s^2}{\sigma_0^2}$$

If

- $U > \text{median } \chi_{n-1}^2 \implies p\text{-value} = 2P(U \geq u)$
- $U < \text{median } \chi_{n-1}^2 \implies p\text{-value} = 2P(U \leq u)$

Exercise: Construct the 95% confidence interval for σ^2 . Then, check if $\sigma_0^2(4) \in 95\%$ confidence interval.

- $H_0: \sigma^2 = 4$ (more than 10%, so it is in the 95% confidence interval)

Likelihood Ratio Test Statistic (one parameter)

Y_1, \dots, Y_n iid $f(y_i; \theta)$ with n large.

- Sample: $\{y_1, \dots, y_n\}$
- $\theta =$ unknown parameter
- $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0$

Step 1: Test statistic:

$$\Lambda = -2 \ln \left[\frac{L(\theta)}{L(\hat{\theta})} \right]$$

If H_0 is true:

$$\Lambda = -2 \ln \left[\frac{L(\theta)}{L(\hat{\theta})} \right] \sim \chi_1^2$$

Step 2: Calculate λ

$$\lambda = -2 \ln \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right] = -2 \ln [R(\theta_0)]$$

$$\begin{aligned} p\text{-value} &= P(\Lambda \geq \lambda) \\ &= P(Z^2 \geq \lambda) \\ &= 1 - P(|Z| \leq \lambda) \end{aligned}$$

EXAMPLE 0.0.5. Suppose $Y_1, \dots, Y_n \sim f(y_i; \theta)$ iid. where

$$f(y, \theta) = \frac{2y}{\theta} e^{-y^2/\theta}$$

Data: $n = 20, \sum y_i^2 = 72$

We want to test $H_0: \theta = 5$ (two sided alternative).

- $\hat{\theta} = \frac{1}{n} \sum y_i^2 = 3.6$
- $R(\theta_0) = \frac{\hat{\theta}}{\theta_0} e^{(1-\hat{\theta}/\theta_0)^n}$
- $\lambda(\theta_0) = \dots$

We know $\lambda = -2 \ln [R(\theta_0)] = 1.9402$ and so

$$R(\theta_0) = \frac{L(\theta_0)}{L(\hat{\theta})} = 0.3791$$

also $\theta_0 = 5$. Lastly, calculate the p -value.

$$\begin{aligned} p\text{-value} &= P(\Lambda \geq \lambda) \\ &= P(Z^2 \geq 1.9402) \\ &\approx 16.5\% \end{aligned}$$

Thus, no evidence against null-hypothesis (H_0).

A few final points:

- (i) Careful about the previous example.
- (ii) λ and the relationship with R
- (iii) Next video
 - $n = 20$ is not large
 - $\lambda = -2 \ln [R(\theta_0)]$: high values of $\lambda \implies$ low values of $R(\theta_0)$

2020-03-20

Roadmap:

(a) Housekeeping

Modified Syllabus + Incentives

Extra materials

Dropbox link + Mathsoc

(b) Gaussian Response Model: An introduction

Gaussian Response Models

Assumption: $Y_1, \dots, Y_n \sim \text{Normal}$

Before: $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ iid. with $\mu, \sigma^2 = \text{unknown}$.

$$Y_i = \mu + R_i$$

where $R_i \sim N(0, \sigma^2)$ and R_i 's independent for each $i \in [1, n]$. We call:

- Y_i **response** variable
- μ **systematic part**
- R **random part**

Now:

- x = explanatory variable
- $\mu = \mu(x)$
- $\sigma^2 = \sigma^2(x)$

For example,

$$Y_i \sim N(\mu(x), \sigma^2(x_i))$$

Simple Linear Regression: $\mu = \alpha + \beta x$ and $\sigma^2 = \text{constant}$.

EXAMPLE 0.0.6.

- Response: Y_i = STAT 231 score of student i
- Explanatory (Covariate): x_i = STAT 230 score of student i (given)

Can Y be explained by x ?

Simple Linear Regression Model

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

for each $i \in [1, n]$ independent.

Our assumptions are:

- $E(Y) = \mu(x) = \alpha + \beta x$
- $Y \sim \text{Normal}$

- $\sigma^2 = \text{constant}$ (independent of x)
 - independent
- We want to estimate α and β .

2020-03-23

Roadmap:

- (i) 5 min recap
- (ii) MLE for α, β, σ
- (iii) Least Squares
- (iv) Example

Recap:

General: $Y \sim N(\mu(x), r(x))$

Assumptions for the Simple Linear Regression Model (Gauss Markov Assumptions)

- (i) One covariate (for the time being)
- (ii) Normality: Y_i 's are Normal
- (iii) Linearity: $E(Y) = \alpha + \beta x$
- (iv) Independence: Y_i 's are all independent
- (v) Homoscedasticity: $\sigma^2 = \sigma^2(x) = \sigma^2$ for all x

We call it a Simple since x is the only explanatory variate. If we used more than one explanatory variate, we call it a multi-variable regression (not covered in this course).

MLE Calculation

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

for each $i \in [1, n]$ independent. We can also write

$$Y_i = (\alpha + \beta x_i) + R_i$$

where $R_i \sim N(0, \sigma^2)$ and R_i 's independent.

$$f(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - (\alpha + \beta x_i))^2}$$

$$L(\alpha, \beta, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum [y_i - (\alpha + \beta x_i)]^2}$$

so,

$$\ell(\alpha, \beta, \sigma) = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{1}{2\sigma^2} \sum [y_i - (\alpha + \beta x_i)]^2$$

$$\frac{\partial \ell}{\partial \alpha} = 0 \implies \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

$$\frac{\partial \ell}{\partial \beta} = 0 \implies \hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\frac{\partial \ell}{\partial \sigma} = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$$

Roadmap:

- (i) Interpretation of SLRM and Recap
- (ii) An example
- (iii) Possible Questions

What we know so far:

- Y_i = response variate = R.V where $i = 1, \dots, n$
- x_i = explanatory variable = given (known numbers)

Examples:

- Y_i = STAT 231, x = STAT 230
- Y_i = stock price in month i , $x = P/E$
- Y_i = wage of UW graduate, x = major

Model: $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ $i \in [1, n]$ independent.

$$Y_i = \alpha + \beta x_i + R_i$$

R_i = residuals and $R_i \sim N(0, \sigma^2)$.

Goal: Extract the relationship between x and Y .

Interpretation:

$$E(Y_i) = \alpha + \beta x_i + 0$$

β = change in $E(Y)$ if x changes by 1 unit

Suppose $x = 0$, then $Y_i = \alpha + R_i$. So $E(Y_i) = \alpha$.

EXAMPLE 0.0.7.

- $n = 30$
- $\bar{x} = 76.733$
- $\bar{y} = 72.233$
- $S_{yy} = 7585.3667$
- $S_{xx} = 5135.8667$
- $S_{xy} = 5106.8667$

What is $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$?

- $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = -4.0677$
- $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.9944$

Regression:

$$Y = \underbrace{-4.0677}_{\hat{\alpha}} + \underbrace{0.9944}_{\hat{\beta}} x$$

$$(x_1, y_1), \dots, (x_{30}, y_{30})$$

$x_{15} = 75 \rightarrow y_{15}$ = predicted with the regression. However, it may or may not lie on the line.

Suppose $\beta = 0$, this means that x has no effect on Y_i since

$$Y_i \sim N(\alpha, \sigma^2)$$

Exercise: $\hat{\beta} = 0 \iff r_{xy} = 0$?

We could also figure out the following (next lecture):

- $H_0: \beta = 0$
- $H_1: \beta \neq 0$
- Confidence interval for β .

2020-03-25

Roadmap:

- (i) Confidence Interval for β
- (ii) Testing for $H_0: \beta = 0$ – Test for correlation for X and Y

EXAMPLE 0.0.8.

- $n = 30$
- $\bar{x} = 76.733$
- $\bar{y} = 72.233$
- $S_{yy} = 7585.3667$
- $S_{xx} = 5135.8667$
- $S_{xy} = 5106.8667$

Regression (Least Squared Equation): $y = -4.0677 + 0.9944x$

- $\hat{\alpha} = -4.0677$
- $\hat{\beta} = 0.9944$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2$
- $s_e^2 = \frac{1}{n-2} \sum_{i=1}^n [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2$
- $s_e = \text{standard error} = 9.4630$ (sqrt of s_e^2)

A look ahead: s_e^2 is an unbiased estimator for σ^2 .

Some Algebra

$$\begin{aligned}
 S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i \\
 &= \sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y} \\
 S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i
 \end{aligned}$$

Thus,

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{S_{xx}} = \sum_{i=1}^n a_i y_i$$

where $a_i = \frac{x_i - \bar{x}}{S_{xx}}$. Also,

$$\tilde{\beta} = \sum_{i=1}^n a_i Y_i$$

Result:

$$\tilde{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

Therefore,

$$\frac{\tilde{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0, 1)$$

but, σ is unknown, so

$$\frac{\tilde{\beta} - \beta}{\frac{s_e}{\sqrt{S_{xx}}}} \sim T_{n-2}$$

THEOREM 0.0.9. We can use

$$\frac{\tilde{\beta} - \beta}{\frac{s_e}{\sqrt{S_{xx}}}} \sim T_{n-2}$$

as a pivotal quantity for β . We can use

$$\frac{(n-2)s_e^2}{\sigma^2} \sim \chi_{n-2}^2$$

as a pivotal quantity for σ^2 .

EXAMPLE 0.0.10.

- (i) Find the 95% Confidence Interval for β .
- (ii) Test whether $\beta = 0$
- (i) The pivot is:

$$\frac{\tilde{\beta} - \beta}{\frac{s_e}{\sqrt{S_{xx}}}} \sim T_{28}$$

Step 1: Critical points $t^* = 2.05$.

$$P(-2.05 \leq \frac{\tilde{\beta} - \beta}{\frac{s_e}{\sqrt{S_{xx}}}} \leq 2.05) = 0.95$$

Coverage interval:

$$\tilde{\beta} \pm t^* \frac{s_e}{\sqrt{S_{xx}}}$$

Confidence interval:

$$\tilde{\beta} \pm t^* \frac{s_e}{\sqrt{S_{xx}}} \\ \Rightarrow [0.72, 1.26]$$

- (ii) We know $\beta = [0.72, 1.26]$. We want to test $\beta = 0$ (we can already see it's not within this interval).

- $H_0: \beta = 0$
- $H_1: \beta \neq 0$

$$D = \left| \frac{\tilde{\beta}}{\frac{s_e}{\sqrt{S_{xx}}}} \right|$$

Value of the test $d = 7.53$.

$$\begin{aligned} p\text{-value} &= P(D \geq d) \\ &= P(|T_{28}| \geq 7.53) \\ &\approx 0 \end{aligned}$$

There is very strong evidence against H_0 . We could also test for any $\beta = \beta_0 \in \mathbb{R}$.

2020-03-27

Roadmap:

- (i) A look back: Pivot for β
- (ii) A look ahead: Confidence interval for $\mu(x) = \text{mean response}$

STAT 230: If $X \sim N(\mu_1, \sigma^2)$, Y is $N(\mu_2, \sigma^2)$, X and Y independent, then

$$aX + bY \sim N(a\mu_1 + b\mu_2, \sigma^2(a^2 + b^2))$$

General result: If $X_i \sim N(\mu_i, \sigma^2)$ with $i = 1, \dots, n$ independent, then

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sigma^2 \sum_{i=1}^n a_i^2\right)$$

We know

$$\hat{\beta} = \sum_{i=1}^n y_i \quad \tilde{\beta} = \sum_{i=1}^n a_i Y_i \quad Y_i \sim N(\underbrace{\alpha + \beta x_i}_{\mu_i}, \sigma^2)$$

$$\tilde{\beta} = \left(\sum_{i=1}^n a_i (\alpha + \beta x_i), \sigma^2 \sum_{i=1}^n a_i^2 \right)$$

Recall:

$$a_i = \frac{x_i - \bar{x}}{s_{xx}}$$

1. $\sum_{i=1}^n a_i = 0$
2. $\sum_{i=1}^n a_i x_i = 1$
3. $\sum_{i=1}^n \frac{1}{s_{xx}}$

So, the mean is

$$\begin{aligned} &= \sum_{i=1}^n a_i \alpha + \sum_{i=1}^n a_i \beta x_i \\ &= \alpha \sum_{i=1}^n a_i + \beta \sum_{i=1}^n a_i x_i \\ &= \beta \end{aligned}$$

the result now follows. \square

Now, we fix x were

- $Y = \text{STAT 231}$
- $X = \text{STAT 230}$

Confidence interval for $\mu(x) = \alpha + \beta x$.

(Average STAT 231 score for all students with a 75 in STAT 230).

$$\begin{aligned} \mu(x) &= \alpha + \beta 75 \\ \hat{\mu}(x) &= \hat{\alpha} + \hat{\beta} x \\ \tilde{\mu}(x) &= \tilde{\alpha} + \tilde{\beta} x \end{aligned}$$

We know $\tilde{\beta}$ is normal, and we can show $\tilde{\alpha}$ is normal. So,

$$\tilde{\mu} \sim N\left(\mu(x), \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}\right)\right)$$

(proof beyond the scope of this course) Thus, the corresponding pivot is

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e \sqrt{\frac{1}{n} - \frac{(x - \bar{x})^2}{S_{xx}}}} \sim T_{n-2}$$

Therefore, the confidence interval (exercise) for $\mu(x)$ is:

$$\left[\hat{\alpha} + \hat{\beta}x \right] \pm t^* S_e \sqrt{\frac{1}{n} - \frac{(x - \bar{x})^2}{S_{xx}}}$$

Can we find the confidence interval for α ? Yes.

Recall, $\alpha = \mu(0)$, so we can just plug in 0 and we get the confidence interval for α .

2020-03-28

Roadmap:

- (i) Prediction Interval for Y given $x = x_{\text{new}}$
- (ii) Model Checking

Problem: $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ $i = 1, \dots, n$ independent. Find the 95% Prediction Interval for Y_{new} when $x = x_{\text{new}}$.

Difference:

- μ was constant (stationary target)
- Y_{new} is a random variable with mean μ (moving target)

EXAMPLE 0.0.11. $x = x_{\text{new}}$

Problem 1: Find the 95% Confidence Interval for $\mu = \alpha + \beta(75)$. Done last lecture.

Problem 2: Find the 95% Prediction Interval for Y when $x_{\text{new}} = 75$.

$$Y \sim N(\alpha + \beta(75), \sigma^2) \tag{1}$$

$$\tilde{\mu}(75) \sim N\left(\mu(75), \sigma^2 \left(\frac{1}{n} + \frac{(75 - \bar{x})^2}{S_{xx}}\right)\right) \tag{2}$$

Subtracting (1) from (2), we get

$$Y - \tilde{\mu}(75) \sim N\left(0, 1 + \frac{1}{n} + \frac{(75 - \bar{x})^2}{S_{xx}}\right)$$

Thus,

$$\frac{Y - \tilde{\mu}(75)}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(75 - \bar{x})^2}{S_{xx}}}} = Z \sim N(0, 1)$$

we replace S_e , then we get

$$\frac{Y - \tilde{\mu}(75)}{S_e \sqrt{1 + \frac{1}{n} + \frac{(75 - \bar{x})^2}{S_{xx}}}} = T_{n-2}$$

Finally, the Prediction Interval is:

$$\hat{\mu}(x_{\text{new}}) \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$$

$$\hat{\mu}(x_{\text{new}}) = \hat{\alpha} + \hat{\beta}x_{\text{new}}$$

Checking the assumptionsMain assumptions

- (i) Normality, with constant variance
- (ii) Linearity: $E(Y) = \alpha + \beta x$
- (iii) Independence

Checking

- (i) Warning
- (ii) The Least Squares line
- (iii) The residual plots

Estimated residuals $= r_i = y_i - \underbrace{(\hat{\alpha} + \hat{\beta}x_i)}_{\hat{y}_i}$. The r_i 's should behave like independent outcomes of $N(0, \sigma^2)$.

Some questions to think about:

- (1) (r_i, x_i)
- (2) (r_i, \hat{y}_i)
- (3) Q-Q plot of r_i 's

2020-03-29

Roadmap:

- (i) Model Checking
- (ii) Final points

SLRM: $Y_i = \alpha + \beta x_i$, $R_i \sim N(0, \sigma^2)$

Residuals: $r_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$.

(a) If the model is correct, how should r_i 's behave?

$$\hat{r}_i = r_i / s_e = \text{standardized residuals} \sim N(0, 1)$$

(b) How should \hat{r}_i 's behave?

Note: $\sum_{i=1}^n r_i = 0$ (check)

Graphical methods

- (i) Residual plots

$$(r_i, x_i)$$

$$(r_i, \hat{y}_i)$$

Q-Q plot of r_i 's

$$\hat{r}_i?$$

- (ii) Warning signs

Final points

- Extensions

Multivariate (x_1, x_2, \dots, x_R) : STAT 3xx

Time Series $(Y_{t-1}, Y_{t-2}, \dots, Y_{t-k})$: STAT 443 (Forecasting)

Non-linearity $(E(Y) = \text{non-linear})$: STAT 4xx

2020-03-30

Two population problems

Roadmap: Gaussian mean problem with equal variances

Problem: $Y_{11}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma^2)$ and $Y_{21}, \dots, Y_{2n_2} \sim N(\mu_2, \sigma^2)$

Question:

- (i) Test $H_0: \mu_1 = \mu_2$ (Two sided alternative)
- (ii) Equivalently, find the confidence interval for $(\mu_1 - \mu_2)$

EXAMPLE 0.0.12.

- CS vs FARM (STAT 231 score)
- Constant variance assumption

Idea:

$$\begin{aligned}
 Y_{1i} &\sim N(\mu_1, \sigma^2) \implies \bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \\
 Y_{2j} &\sim N(\mu_2, \sigma^2) \implies \bar{Y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) \\
 \implies \bar{Y}_1 - \bar{Y}_2 &\sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)
 \end{aligned}$$

Therefore,

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = Z$$

But σ is unknown, so we can say

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2}$$

for some S_p , we need to find this.

The calculation of the MLE

$$\begin{aligned}
 \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\
 \hat{\mu}_2 &= \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} \\
 \hat{\sigma}^2 &= \frac{1}{n_1 + n_2} \left[\sum (y_{1i} - \bar{y}_1)^2 + \sum (y_{2j} - \bar{y}_2)^2 \right]
 \end{aligned}$$

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

Check $E(S_p^2) = \sigma^2$.

EXAMPLE 0.0.13. Assume equal variances hold.

- $n_1 = 10$
- $n_2 = 10$
- $\bar{y}_1 = 10.4$
- $\bar{y}_2 = 9.0$
- $s_1 = 1.1314$
- $s_2 = 1.8742$

Test whether $H_0: \mu_1 = \mu_2$ vs the two sided alternative.

Test statistic:

$$D = \left| \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| = \left| \frac{(\bar{Y}_1 - \bar{Y}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

which we get $s_p = 1.548$, and $d = 2.2215$.

$$p\text{-value} < 5\%$$

reject H_0 .

Final points:

- Relationship with SLRM?
- A look ahead

2020-04-01

Roadmap:

- (i) Independent population, unequal variance
- (ii) Paired Data
- (iii) Housekeeping: evaluate.uwaterloo.ca
- (iv) Recap

Recap: Equal variances: