STAT 231 - Statistics

Cameron Roopnarine

Last updated: March 31, 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. $Y_1, \ldots Y_n$ iid $N(\mu, \sigma^2)$
- $\sigma^2 =$ known
- $\mu =$ unknown
- Sample: $\{y_1, \ldots, y_n\}$
- $\overline{y} =$ sample mean
- H_0: $\mu = \mu_0$
- H_1: $\mu \neq \mu_0$

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

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

$$\begin{aligned} p\text{-value} &= P(D \geqslant d) && \text{assuming } H_0 \text{ is true} \\ &= P(|Z| \geqslant d) && Z \sim N(0,1) \end{aligned}$$

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.

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

$$\mu_0 \leqslant \overline{y} + 1.96 \frac{\sigma}{\sqrt{n}}$$
$$\mu_0 \geqslant \overline{y} - 1.96 \frac{\sigma}{\sqrt{n}}$$

These two equations yield

$$d = \left| \frac{\overline{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \leqslant 1.96$$
$$P(|Z| \geqslant d) > 0.05$$

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 $(y_1, \ldots, y_n)$
- H_0: The scale is unbiased
- H_1: The scale is biased

If the scale was unbiased,

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

If the scale was biased,

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

- 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:

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

Compute d.

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

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

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

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

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

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

If n is large, then

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

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

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

where Λ satisfies all the properties of D. Also,

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

and

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

2020-03-16

Roadmap:

(i) General info

(ii) Testing for variance for Normal

(iii) An example

The general problem: $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ 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:

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

can we use this as our test statistic?

EXAMPLE 0.0.4.

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

What is the p-value? We know

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

Compute U:

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

χ^2_{19}

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

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

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

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

We calculated the p-value:

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

If

- $U >$ median $\chi^2_{n-1} \implies p$-value $= 2P(U \geqslant u)$
- $U <$ median $\chi^2_{n-1} \implies p$-value $= 2P(U \leqslant u)$

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)

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

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

Step 1: Test statistic:

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

If H_0 is true:

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

Step 2: Calculate λ

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

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

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

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

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

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

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

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

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: $Y_1, \ldots, Y_n \sim$ Normal

Before: $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ iid. with $\mu, \sigma^2 =$ 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 $\sigma^2 =$ 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:
- $E(Y) = \mu(x) = \alpha + \beta x$
- $Y \sim$ Normal

- $\sigma^2 =$ 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,

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

$$\frac{\partial \ell}{\partial \alpha} = 0 \implies \hat{\alpha} = \overline{y} - \hat{B}\overline{x}$$

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

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

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.

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

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

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

Regression:

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

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

$x_{15} = 75 \to y_{15}$ = 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.

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

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

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

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

Some Algebra

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

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})x_i$$

Thus,

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

where $a_i = \frac{x_i - \overline{x}}{S_{xx}}$. 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

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

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 $t^* = 2.05$.

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

Coverage interval:

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

Confidence interval:

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

$$\implies [0.72, 1.26]$$

(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 $d = 7.53$.

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

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