STAT 231 - Statistics

Cameron Roopnarine

Last updated: April 8, 2020

Contents

Chapter 1

Lectures

1.1 2020-03-02

Roadmap:

(i) 5 min recap

(ii) Confidence for Normal with unknown variance

(iii) Prediction Intervals

(iv) Relationship between likelihood intervals and confidence intervals

$$W \sim \chi^2_n \iff W = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

where each $Z_i \sim N(0,1)$ and Z_i's independent. We know $E(W) = n$ and $Var(W) = 2n$.

Let $W_1 \sim \chi^2_{n_1}$ and $W_2 \sim \chi^2_{n_2}$ be independent, then

$$W_1 + W_2 \sim \chi^2_{n_1+n_2}$$

Student's T-distribution

We say $T \sim T_n$ if

$$T = \frac{Z}{\sqrt{W/n}}$$

where $Z \sim N(0,1)$ and $W \sim \chi^2_n$ are independent. Note that $E(T) = 0$ and T is symmetric. Also, as $n \to \infty$, then $T \to Z \sim N(0,1)$.

THEOREM 1.1.1. *Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$ where μ and σ are unknown. Let*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Then,

(i) The pivotal quantity for μ is:

$$\frac{\overline{Y} - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1}$$

(ii) The pivotal quantity for σ^2 is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

REMARK 1.1.2. (i) Shows that if we replace σ by its estimator S, then it follows a T-distribution with $(n-1)$ degrees of freedom.

EXAMPLE 1.1.3. An independent sample of 25 students are taken and STAT 231 scores are recorded.

- $\overline{y} = 75$
- $s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2 = 64$

(a) Find the 99% confidence interval for μ.
(b) Find the 95% confidence interval for σ^2.
(c) Find the 99% prediction interval for Y_{26}.

Solution. We know $Y_1, \ldots, Y_{25} \sim N(\mu, \sigma^2)$ where $Y_i =$ STAT 231 score of the i^{th} student.
(a) We know

$$\frac{\overline{Y} - \mu}{\frac{S}{\sqrt{n}}} \sim T_{24}$$

We want a t such that

$$P(|T_{24}| \leqslant t) = 0.99$$

Using the table we see that $t = 2.80$. Now,

$$P(-2.8 \leqslant T_{24} \leqslant 2.8) = 0.99$$

$$\implies P\left(-2.8 \leqslant \frac{\overline{Y} - \mu}{\frac{S}{\sqrt{n}}} \leqslant 2.8\right) = 0.99$$

$$\implies P\left(\overline{Y} - 2.8\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \overline{Y} + 2.8\frac{S}{\sqrt{n}}\right) = 0.99$$

Thus, the 99% confidence interval is:

$$\overline{y} \pm 2.8\frac{s}{\sqrt{n}} \implies [62.2, 87.8]$$

(b) We know

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{24}$$

We want any value a and b such that

$$P(a \leqslant \chi^2_{24} \leqslant b) = 0.95$$

We choose the symmetric solution with $a = 0.025 \to 13.120$ and $b = 0.975 \to 40.646$. Now,

$$P\left(13.120 \leqslant \chi^2_{24} \leqslant 40.646\right) = 0.95$$

$$\implies P\left(13.120 \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant 40.646\right) = 0.95$$

$$\implies P\left(\frac{(n-1)S^2}{40.646} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{13.120}\right) = 0.95$$

Thus, the 95% confidence interval for σ^2 is:

$$\left[\frac{(n-1)s^2}{40.646}, \frac{(n-1)s^2}{13.120}\right] \implies [37.79, 117.07]$$

(c) Prediction interval.

$$Y_{26} \sim N(\mu, \sigma^2)$$
$$\overline{Y} \sim N(\mu, {}^{\sigma^2}/_{n})$$
$$\implies Y_{26} - \overline{Y} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$$

Therefore, the pivotal quantity is:

$$\frac{Y_{26} - \overline{Y}}{\sigma\sqrt{1 + \frac{1}{n}}} = Z \sim N(0,1)$$

we replace σ by its estimator and get

$$\frac{Y_{26} - \overline{Y}}{S\sqrt{1 + \frac{1}{n}}} \sim T_{24}$$

Thus,

$$P(|T_{24}| \leqslant 2.8) = 0.99$$

yields the general 99% prediction interval:

$$\overline{y} \pm t^* s\sqrt{1 + \frac{1}{n}}$$

We make the following remark:

REMARK 1.1.4. Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$. Then,

(i) The general confidence interval for μ is:

$$\overline{y} \pm z^* \frac{\sigma}{\sqrt{n}} \quad \text{if } \sigma \text{ is known}$$

$$\overline{y} \pm t^* \frac{s}{\sqrt{n}} \quad \text{if } \sigma \text{ is unknown}$$

(ii) The general confidence interval for σ^2 is:

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$$

where a and b come from the χ^2_{n-1} table and $b - a = \text{RHS}$.

(iii) The general prediction interval for Y_{n+1} is:

$$\overline{y} \pm t^* s\sqrt{1 + \frac{1}{n}}$$

THEOREM 1.1.5. *As $n \to \infty$,*

$$\Lambda(\theta) = -2\ln\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \sim \chi^2_1$$

where $\tilde{\theta}$ is the maximum likelihood estimator. We call the random variable $\Lambda(\theta)$ the likelihood ratio statistic.

EXAMPLE 1.1.6. Suppose n is large, and we have a 10% likelihood interval. What is the corresponding coverage probability?
Solution. 10% likelihood interval $\implies R(\theta) \geqslant 0.1$

$$\begin{aligned}
&\implies \frac{L(\theta)}{L(\hat{\theta})} \geqslant 0.1 \\
&\implies -2\ln\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leqslant -2\ln(0.1) \\
&\implies \lambda(\theta) \leqslant -2\ln(0.1)
\end{aligned}$$

Thus, the corresponding coverage:

$$\begin{aligned}
P(\Lambda(\theta) \leqslant -2\ln(0.1)) &= P(Z^2 \leqslant -2\ln(0.1)) \\
&= P(|Z| \leqslant \sqrt{-2\ln(0.1)}) \\
&\approx 97\%
\end{aligned}$$

1.2 2020-03-04

DEFINITION 1.2.1. An estimator $\tilde{\theta}$ is called ***unbiased*** for θ if

$$E(\tilde{\theta}) = \theta$$

EXAMPLE 1.2.2. Let $W = \frac{(n-1)S^2}{\sigma^2}$. Find $E(S^2)$.
Solution.

$$\begin{aligned}
&E(W) = n - 1 \\
&\implies E\left(\frac{(n-1)S^2}{\sigma^2}\right) = n - 1 \\
&\implies \frac{n-1}{\sigma^2}E(S^2) = n - 1 \\
&\implies E(S^2) = \sigma^2
\end{aligned}$$

Therefore S^2 is an unbiased estimator for σ^2, and this is why we divide by $(n-1)$.

Other Confidence Intervals

Poisson Suppose $Y_1, \ldots, Y_n \sim \text{Poi}(\mu)$ are independent and n is large. Find the 95% confidence interval.

$$\overline{Y} \sim N(\mu, \sigma^2 = {}^{\mu}\!/_{n})$$

Find the pivotal quantity now.

Exponential Suppose $Y_1, \ldots, Y_n \sim \exp(\theta)$ are independent and n is small.

THEOREM 1.2.3. *If $Y \sim \exp(\theta)$, then*

$$\frac{2Y}{\theta} \sim \exp(2)$$

If $W_i = {}^{2Y_i}\!/_{\theta}$, then

$$\sum_{i=1}^{n} W_i \sim \chi^2_{2n}$$

Proof. Let $F_W(w)$ be the cumulative distribution function of W. Then,

$$\begin{aligned} F_W(w) &= P(W \leqslant w) \\ &= P\left(\frac{2Y}{\theta} \leqslant w\right) \\ &= P\left(Y \leqslant \frac{w\theta}{2}\right) \\ &= 1 - e^{-\frac{w\theta/2}{\theta}} \\ &= 1 - e^{-w/2} \end{aligned}$$

Therefore,

$$f(w) = \frac{1}{2}e^{-w/2}$$

$\square$

Using this theorem, we can find the confidence interval for θ.

$$\begin{aligned} &P\left(a \leqslant \chi^2_{2n} \leqslant b\right) = 0.95 \\ \implies\ &P\left(a \leqslant \sum_{i=1}^{n} W_i \leqslant b\right) = 0.95 \\ \implies\ &P\left(a \leqslant \sum_{i=1}^{n} \frac{2Y_i}{\theta} \leqslant b\right) = 0.95 \\ \implies\ &P\left(a \leqslant \frac{2}{\theta}\sum_{i=1}^{n} Y_i \leqslant b\right) = 0.95 \end{aligned}$$

yields

$$\left[\frac{2\sum\limits_{i=1}^{n} Y_i}{b}, \frac{2\sum\limits_{i=1}^{n} Y_i}{a}\right]$$

where a and b are from the χ^2 table.

THEOREM 1.2.4. *If we have a $p\%$ coverage interval with Z as a pivot, and n is large, then the corresponding likelihood is given by*

$$e^{-(z^*)^2/2}$$

EXAMPLE 1.2.5. If $p = 0.95$ and $z^* = 1.96$, then the corresponding likelihood is:

$$e^{-(1.96)^2/2} \approx 0.15$$

1.3 2020-03-06

Roadmap:

(i) Recap (excluded from these notes)

(ii) Testing of hypotheses (Null vs Alternate) and (Two-sided vs One-sided tests)

(iii) Clicker

Hypothesis Testing

DEFINITION 1.3.1. A hypothesis is a statement about the (parameters of) population. There are two (competing) hypotheses.
Null Hypothesis H_0: current belief, conventional wisdom
Alternate Hypothesis H_1: challenger to the conventional wisdom

EXAMPLE 1.3.2. Suppose we want to test whether a coin is biased. We flip the coin 100 times and get 52 heads. Let $\theta = P(H)$

- H_0: $\theta = \frac{1}{2}$
- H_1: $\theta \neq \frac{1}{2}$

Approach p-value approach.

DEFINITION 1.3.3. The p-value: is the probability of observing my evidence (or worse) under the assumption that H_0 is true. The lower the p-value, the strong is the evidence against H_0.

Notes:

- H_0 and H_1 are not treated symmetrically.
- Unless there is overwhelming evidence ("beyond a reasonable doubt") against $H - 0$, we stick with it. The burden is on the challenger.

	H_0 is true	H_1 is true
Reject H_0 (convict)	X_1	✓
Do not reject H_0	✓	X_2

where X_1 is a Type I error and X_2 is a Type II error.

Two-sided vs One-sided tests:

- H_0: $\theta = \frac{1}{6}$
- H_1: $\theta < \frac{1}{6}$

Clicker Question The p-value $= P(H_0$ is true$)$.

(a) True

(b) **False**

1.4 2020-03-09

Roadmap:

(i) Binomial testing

(ii) Review for the midterm (excluded from these notes)

DEFINITION 1.4.1. p-value: Probability of observing as extreme an observation of your data, given the null hypothesis is true.

DEFINITION 1.4.2. A test statistic (discrepancy measure) is a random variable that measures the level of disagreement of your data with the null hypothesis. Typically, it satisfies the following properties:

(i) $D \geqslant 0$

(ii) $D = 0 \implies$ best news for H_0

(iii) High values of $D \implies$ bad news for H_0

(iv) Probabilities can be calculated if H_0 is true

Steps for a Statistical test

Step 1: Construct the test-statistic D

EXAMPLE 1.4.3. Test whether a coin is fair (against the two sided alternative). Let $n = 100$ and $y = 52$ heads.

- H_0: $\theta = \frac{1}{2}$
- H_1: $\theta \neq \frac{1}{2}$

where $\theta = P(H)$.

Model: $Y \sim \text{Bin}(100, \theta)$.

$$D = |Y - 50|$$

as it satisfies (i)-(iv).

Step 2: Find d from your data set.

$$p\text{-value} = P(D \geqslant d; H_0 \text{ is true})$$

Step 3: Make conclusions based on your p-value

For our Binomial problem,

$$D = |Y - 50| \implies d = |52 - 50| = 2$$

Thus,

$$p\text{-value} = P(|Y - 50| \geqslant 2)$$

but this is difficult to calculate. For n large enough, we can use

$$D = \left| \frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \right|$$

as a possible test statistic.

1.5 2020-03-11

TODO.

1.6 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 1.6.1. $Y_1, \dots Y_n$ iid $N(\mu, \sigma^2)$

- $\sigma^2 =$ known
- $\mu =$ unknown
- Sample: $\{y_1, \dots, y_n\}$
- $\overline{y} =$ sample mean
- H_0: $\mu = \mu_0$ where μ_0 is given
- 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

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

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 1.6.2 (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 1.6.3 (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,

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

and

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

Chapter 2

Online Lectures

2.1 2020-03-16: Testing for Variances

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 are both unknown.
- Sample: $\{y_1, \ldots, y_n\}$
- 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? We will calculate

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

We want to compare u to the median of χ^2_{n-1}:

- If $u >$ median, then p-value $= 2P(U \geqslant u)$.
- If $u <$ median, then p-value $= 2P(U \leqslant u)$.

EXAMPLE 2.1.1.

- Normal population: $\{y_1, \ldots, y_n\}$
- $n = 20$
- $\sum\limits_{i=1}^{n} y_i = 888.1$
- $\sum\limits_{i=1}^{n} y_i^2 = 39545.03$

- H_0: $\sigma = \sigma_0 = 2 \iff \sigma^2 = \sigma_0^2 = 4$
- H_1: $\sigma \neq \sigma_0 = 2 \iff \sigma^2 \neq \sigma_0^2 = 4$

What is the p-value? We know

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] = \frac{1}{19}\left[(39545.03) - (20)\left(\frac{888.1}{20}\right)\right] = 5.7342$$

Compute u:

$$u = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(19)(5.7342)}{4} = 27.24$$

We need to determine if u is to the right or left of the median χ^2_{19}. We know it will be to the right since the mean of χ^2_{19} is 19. χ^2 is right-skewed, so the mean must be bigger than the median, thus the median must be less than 19. Therefore, $u >$ median. Alternatively, we can use the table and look at $p = 0.5$, $df = 19 \to 18.338 < u$.

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

We see that 27.24 falls between $p = 0.9$ and $p = 0.95$. The area to the right of $p = 0.9$ is 10% and the area to the right of $p = 0.95$ is 5%. Thus, $2P(5\%$ and $\%10) = 10\%$ and 20%, which implies $p > 0.1$ and we conclude there is no evidence against null-hypothesis.

2.2 2020-03-18: Likelihood Ratio Test Statistic Example

Roadmap:

(i) 5 min recap

(ii) LTRS for large n

(iii) An example

(i) 5 min recap

$Y_1, \ldots, Y_n$ 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 $= 2P(U \geqslant u)$ (twice right tail)
- If $u <$ median $\chi^2_{n-1} \implies p$-value $= 2P(U \leqslant u)$ (twice left tail)

Exercise For 2.1.1,

- Construct the 95% confidence interval for σ^2.
- Check if $\sigma_0^2(4) \in 95\%$ confidence interval.

We already know that H_0: $\sigma^2 = 4$ yields a p-value > 0.1, so it should be in the 90% confidence interval $\implies$ it's in the 95% confidence interval.

(ii) LTRS for large n

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

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

Step 1: Test statistic:

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

If H_0 is true:

$$\Lambda(\theta_0) = -2\ln\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] \sim \chi_1^2$$

Step 2: Calculate $\lambda(\theta_0)$

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

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

(iii) An example

EXAMPLE 2.2.1. Suppose $Y_1, \ldots, Y_n \sim f(y_i;\theta)$ iid where

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

- $n = 20$
- $\sum\limits_{i=1}^{n} y_i^2 = 72$

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

- $\hat{\theta} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i = \frac{1}{20}(72) = 3.6$
- $R(\theta_0) = \left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{\left(1-\frac{\hat{\theta}}{\theta_0}\right)n} = 0.379052$
- $\lambda(\theta_0) = -2\ln\left[R(\theta_0)\right] = 1.94016$

$$\begin{aligned} p\text{-value} &= P(\Lambda \geqslant \lambda) \\ &= P(Z^2 \geqslant 1.94016) \\ &= 1 - \left[2P(Z \leqslant \sqrt{1.94016}) - 1\right] \\ &= 1 - [2(0.97381) - 1] \\ &= 0.16452 \\ &\approx 16.5\% \end{aligned}$$

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
- High values of $\lambda \implies$ low values of $R(\theta_0)$

2.3 2020-03-20: Intro to Gaussian Response Models

Roadmap:

(i) Housekeeping

Modified Syllabus + Incentives

Extra materials

Dropbox link + MathSoc

(ii) 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. Equivalently,

$$Y_i = \mu + R_i$$

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

- Y_i ***response*** variate (dependent variable)
- μ ***systematic part***
- R ***random part***

Now:

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

The general gaussian response model is:

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

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

EXAMPLE 2.3.1.

- Response variable: $Y_i =$ STAT 231 score of student i
- Explanatory variable: $x_i =$ 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

Goal: We want to estimate α and β.

2.4 2020-03-23: MLE Regression

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. We say $\alpha + \beta x_i$ is the systematic part, and R_i is the random part.

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

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

so,

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

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

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

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

EXAMPLE 2.4.1 (Numerical Example).

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
1	2	-4	-4	16	16	16
3	3	-2	-3	4	9	6
5	7	0	1	0	1	0
7	9	2	3	4	9	6
9	9	4	3	16	16	12
		0	0	$S_{xx}=40$	S_{yy}	$S_{xy}=40$

- $\overline{x}=5$
- $\overline{y}=6$

Find the regression equation.
Solution.

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}}=40/40=1$$

$$\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}=6-(1)(5)=1$$

Thus, the regression equation is:

$$y=\hat{\alpha}+\hat{\beta}x=1+x$$

Method of Least Squares

$$\text{minimize} \quad \sum_{i=1}^{n}\Big[y_i-(\hat{\alpha}+\hat{\beta}x_i)\Big]^2$$

This is exactly the same as what we did previously. Sometimes we call $\hat{\alpha}$ and $\hat{\beta}$ least square estimates.

2.5 2020-03-23: Beta Properties and a Look Ahead

Roadmap:

(i) Interpretation of SLRM and Recap

(ii) An example

(iii) Possible Questions

What we know so far:

- Y_i = response variate = random variable 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:

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

$\beta =$ 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 2.5.1.

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

Find the regression equation.

Solution.

- $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{5106.8667}{5135.8667} = 0.9944$
- $\hat{\alpha} = \overline{y} - \hat{B}\overline{x} = 72.233 - (0.9944)(76.733) = -4.0677$

Thus, the regression equation is:

$$y = -4.0677 + 0.9944x$$

Note: Suppose we have the data set $\{(x_1, y_1), \ldots, (x_{30}, y_{30})\}$. If $x_{15} = 75$, we can predict y_{15} using the regression equation. However, it may or may not lie on the line.

Given $y = -4.0677 + 0.9944x.$, 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):

- Confidence interval for β
- H_0: $\beta = 0$ (x is uncorrelated to Y)
- H_1: $\beta \neq 0$

2.6 2020-03-25: Interval Estimation and Hypothesis for Beta

Roadmap:

(i) Confidence Interval for β

(ii) Testing for H_0: $\beta = 0$ (Test for correlation for x and Y)

EXAMPLE 2.6.1. Last class we found the least square equation using the following data.

- $n = 30$
- $\overline{x} = 76.733$
- $\overline{y} = 72.233$
- $S_{yy} = 7585.3667$
- $S_{xx} = 5135.8667$
- $S_{xy} = 5106.8667$
- $\hat{\alpha} = -4.0677$
- $\hat{\beta} = 0.9944$

$$y = -4.0677 + 0.9944x$$

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

We now introduce the standard error, denoted s_e, where we divide by $(n-2)$ instead of $(n-1)$ in our sample standard variance.

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - (\hat{\alpha} + \hat{\beta}x_i)\right]^2$$

In our example, $s_e = 9.4630$. Don't forget to square root 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,

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

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 2.6.2. *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 2.6.3. Continuation of 2.6.1.
(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 using table with $p = 0.975$, $df = 28 \to t^* = 2.05$.

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

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:

$$d = \frac{\hat{\beta}}{\frac{s_e}{s_{xx}}} = \frac{0.9944}{\frac{9.4630}{\sqrt{5135.8667}}} = 7.53$$

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

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

2.7 2020-03-26: Pivotal Distribution for Beta and Confidence for the Mean

Roadmap:

(i) A look back: Pivot for β

(ii) A look ahead: Confidence interval for $\mu(x) =$ mean response

STAT 230: If $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$, 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 $i = 1, \ldots, n$ 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

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

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

Recall:

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

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

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

- $Y =$ STAT 231
- $x =$ 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,

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

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

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

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

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

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.

2.8 2020-03-28: Prediction Interval and Intro to Model Checking

Roadmap:

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

(ii) Model Checking

Problem: $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ $i = 1, \ldots, n$ 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 2.8.1. $x = x_{\text{new}}$
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 $x_{\text{new}} = 75$.

$$Y \sim N(\alpha + \beta(75), \sigma^2) \tag{2.1}$$

$$\tilde{\mu}(75) \sim N\left(\mu(75), \sigma^2\left(\frac{1}{n} + \frac{(75-\overline{x})^2}{S_{xx}}\right)\right) \tag{2.2}$$

Subtracting (1) from (2), we get

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

Thus,

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

we replace S_e, then we get

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

Finally, the Prediction Interval is:

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

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

Checking the assumptions

Main 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

2.9 2020-03-29: Model Checking and Final Points

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 Linear Regression $(x_1, x_2, \ldots, x_k)$: STAT 3xx

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

  Non-linearity ($E(Y)$ = non-linear): STAT 4xx

2.10 2020-03-30: Two Population Case I Equal Variance

Two population problems

Roadmap: Gaussian mean problem with equal variances

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

Question:

(i) Test H_0: $\mu_1 = \mu_2$ (Two sided alternative)

(ii) Equivalently, find the confidence interval for $(\mu_1 - \mu_2)$

EXAMPLE 2.10.1.
- CS vs FARM (STAT 231 score)
- Constant variance assumption

Idea:

$$Y_{1i} \sim N(\mu_1, \sigma^2) \implies \overline{Y}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

$$Y_{2j} \sim N(\mu_2, \sigma^2) \implies \overline{Y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

$$\implies \overline{Y}_1 - \overline{Y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Therefore,

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

But σ is unknown, so we can say

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

for some S_p, we need to find this.

The calculation of the MLE

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j}$$

$$\hat{\sigma}^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(y_{1i} + \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2\right] = \frac{1}{n_1+n_2}\left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right]$$

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

Check $E(S_p^2) = \sigma^2$; that is, S_p^2 is an unbiased estimator for σ^2. Hint: We already know $E(S_1^2) = E(S_2^2) = \sigma^2$

EXAMPLE 2.10.2. Assume equal variances hold.

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

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

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

$$d = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_2} + \frac{1}{n_2}}} = \frac{10.4 - 9.0}{1.5480\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.0223$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(10-1)(1.1314)^2 + (10-1)(1.8742)^2}{10 + 10 - 2} = 2.3963458$$

Thus, $s_p = 1.5480$ and $d = 2.0223$. Look in the table with $df = 18$, $t = 2.10 \to p = 0.975$.

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

reject H_0.

Final points:

- Relationship with SLRM?
- A look ahead

2.11 2020-04-01: Large Samples and Paired Data

Roadmap:

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

The following are equivalent:

- H_1: $\mu_1 = \mu_2$
- Confidence interval: $\mu_1 - \mu_2 = 0$

Recap: Equal variances:

$$Y_{1i} \sim N(\mu_1, \sigma^2),\ Y_{2j} \sim N(\mu_2, \sigma^2)$$

Pivotal Quantity:

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2} \implies (\overline{y}_1 + \overline{y}_2) \pm t^* s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

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

Test statistic is the absolute value of above.

Unequal variances, large samples, independent population

$$Y_{1i} \sim N(\mu, \sigma_1^2),\ Y_{2j} \sim N(\mu_2, \sigma_2^2)$$

where $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$.

THEOREM 2.11.1. *If n_1 and n_2 are large, then*

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim Z$$

The 95% confidence interval; that is, we solve $P(-1.96 \leqslant Z \leqslant 1.96) = 0.95$ where Z is defined as in the theorem is:

$$(\overline{y}_1 - \overline{y}_2) \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $z^* = 1.96$. To test H_0: $\mu_1 = \mu_2$, check if 0 is within the interval.

EXAMPLE 2.11.2.

- $n_1 = 278$
- $n_2 = 345$
- $\overline{y}_1 = 60.2$
- $\overline{y}_2 = 58.1$
- $s_1 = 10.16$
- $s_2 = 9.02$

Find the 95% confidence interval for $\mu_1 - \mu_2$.

Solution.

$$(\overline{y}_1 - \overline{y}_2) \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

yields

$$[0.57, 3.63]$$

Suppose we are given H_0: $\mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$ at 5%, is this reasonable? No, since 0 is not within the interval above $\implies$ p-value < 0.05.

<u>Paired Data</u>: Natural $1-1$ map between the units of the population.

(i) <u>Examples</u>

(ii) <u>Idea of Pivotal Quantity</u>

(iii) <u>Example</u>

(i)

- Before and after
- Same car, same driver, number of miles travelled between fuel A and fuel B (not independent)

$$\begin{pmatrix} b_1 \\ a_1 \end{pmatrix}, \ldots, \begin{pmatrix} b_n \\ a_n \end{pmatrix}$$

where each b_i are before data and each a_i are after data.

$$B_i \sim N(\mu_1, \sigma_1^2)$$

$$A_i \sim N(\mu_2, \sigma_2^2)$$

these are pairs, so let's subtract them

$$(B_i - A_i) = Y_i \sim N(\mu_1 - \mu_2, \sigma^2)$$

for some σ^2 (there will be covariance within there). We are testing H_0: $\mu = 0$. Population of differences (B_i's vs A_i's)

EXAMPLE 2.11.3. See Table 6.3 in the course notes for the data. Step 1: Construct $y_i = b_i - a_i$ for each $i \in [1, n]$.

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

and test H_0: $\mu = 0$.

- $\overline{y} = -0.020$
- $s = 0.411$
- $d = \frac{\overline{y}}{s/\sqrt{n}} \sim T_{n-1}$ where $n - 1 = 19$
- Confidence interval: $[-0.212, 0.172]$
  $\overline{y} + t^* s/\sqrt{n}$, t^* = column 19, row 0.975.

0 falls within the confidence interval, so the p-value is less than 5%.

Final points

(i) Case I: Equal variance, independent samples

(ii) Case II: Unequal variance, independent samples, large sample sizes

(iii) Case III: Paired data

We ignored one case: small sample sizes, unequal variances (we don't worry about it in this course).

Typically, in paired data the two variables are not independent, but positively correlated, however the variance is $\sigma_1^2 + \sigma_2^2 - 2\text{Cov}(b_i, a_i)$ where $\text{Cov}(b_i, a_i) > 0$ if the variance is lower, the variances are more accurate. We should always go for the paired method iff the covariance is positively correlated.

2.12 2020-03-02: The Big Picture–Take 2

Roadmap

(i) The big picture

(ii) Two examples

Example 1: Check whether a die is fair

- $\theta_i = P(i^{\text{th}}\text{ face})$ where $i = 1, \ldots, 6$
- H_0: $\theta_1 = \theta_2 = \cdots = \theta_6 = \frac{1}{6}$
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$

If H_0 was true, then the expected frequency would be close to the observed frequency.

	Observed Frequency	Expected Frequency
1	48	50
2	72	50
3	60	50
4	40	50
5	40	50
6	40	50

The question we want to answer is how close is close enough?

Example 2: $W_1, \ldots, W_n \sim Poi(\mu)$. H_0: $W_i \sim Poi(\mu)$.

	Observed Frequency	Expected Frequency
0	y_0	e_0
1	y_1	e_1
2	y_2	e_2
3	y_3	e_3
$\geqslant 4$	y_4	e_4

where

$$e_i = n \times \frac{e^{-\hat{\mu}}\hat{\mu}^i}{i!}$$

Multinomial

- Extension to the Binomial
- Distribution function
- Likelihood function
- MLE
- LRTS

Distribution function and likelihood function:

$$\frac{n!}{x_1! \cdots x_k!}\theta_1^{x_1} \cdots \theta_k^{x_k}$$

where $x_1 + \cdots + x_k = n$.

The MLE is

$$\hat{\theta}_i = \frac{x_i}{n}$$

for each $i \in [1, k]$.

LRTS: If n is large, we can construct a LRTS to test H_0.

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

The particular form is,

$$\Lambda = 2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{Y_i}{E_i}\right)\right] \sim \chi^2_{k-\ell-1}$$

where

- Y_i is the observed frequency,
- E_i is the expected frequency if H_0 was true,

- k is the number of categories, and
- ℓ is the number of components of θ we need to estimate under H_0.

EXAMPLE 2.12.1. H_0: $\theta_1 = \cdots = \theta_6 = \frac{1}{6}$.

	Observed Frequency	Expected Frequency
1	48	50
2	72	50
3	60	50
4	40	50
5	40	50
6	40	50

Calculate the p-value.
Solution.

$$\lambda = 2\sum_{i=1}^{6}\left[y_i \ln\left(\frac{y_i}{e_i}\right)\right]$$

Then, let n the number of categories and k be the number of parameters we estimate under H_0. So the degrees of freedom in our case is $6 - k - 1 = 5$ where $k = 0$ since we are given all of the θ_i's.

$$\begin{aligned} p\text{-value} &= P(\Lambda \geqslant \lambda) \\ &= P(\chi^2_5 \geqslant \lambda) \end{aligned}$$

REMARK 2.12.2. In the example we have different letters for the degrees of freedom compared to our derivation to match the course notes.

2.13 2020-03-02: Goodness of Fit

Roadmap:

(i) Recap

(ii) Goodness of fit

Discrete $\to$ Poisson

Continuous $\to$ Exponential

These results will only hold for large n. Also, the observed frequencies should be at least 5.

EXAMPLE 2.13.1 (Poisson)**.** Let W_i be the number of service interruptions on the i^{th} day over 200 days.

Number of interruptions	0	1	2	3	4	5	$\geqslant 5$
Observed Frequency (y_i)	64	71	42	18	4	1	0
Expected Frequency (e_j)	63.3	72.8	41.8	16.0	4.6	1.3	$\cdots$

Is the Poisson model appropriate? H_0: $W \sim Poi(\theta)$. We must calculate the expected frequencies (done above, formula below).

- We estimate: $\hat{\theta} = \frac{1}{n}\sum\limits_{i=1}^{n} w_i = 1.15$
- $e_j = n \times \frac{e^{-\hat{\theta}}\hat{\theta}^i}{i!}$

$$\lambda_j = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{e_i}\right)\right] = 0.43$$

$$\begin{aligned} p\text{-value} &= P(\Lambda \geqslant \lambda) \\ &= P(\chi^2_{5-1-1} \geqslant \lambda) \\ &= P(\chi^2_3 \geqslant 0.43) \\ &\geqslant 0.9 \end{aligned}$$

No evidence against H_0, so Poisson is a good model.

EXAMPLE 2.13.2 (Exponential).

Interval	$[0, 100]$	$(100, 200]$	$(200, 300]$	$(300, 400]$	$(400, 600]$	$(600, 800]$	> 800
y_j	29	22	12	10	10	9	8
e_j	27.6	20	14.4	$\cdots$			

H_0: $W \sim \exp(\theta)$. $\hat{\theta} = \overline{w} = 310$

$$e_1 = n \times P\left[W \in [100, 200]\right] = n \times [F(200) - F(100)] = n \times \left(1 - e^{-\frac{200}{310}} - \left(1 - e^{-\frac{100}{310}}\right)\right)$$

$$\Lambda \sim \chi^2_{7-1-1}$$

Final points:

(a) In all our problems above, we always try to convert to a multinomial.

(b) Suppose we are given $W \sim N(\mu, \sigma^2)$ with 5 intervals. Our LRTS will have $df = 5 - 2 - 1 = 2$ where we subtract by 2 since we estimate μ and σ. If we were given σ, we would have $df = 5 - 1 - 1 = 3$.

(c) Final answer (p-value) will depend on how we divide our data into categories.

2.14 2020-03-02: Contingency Tables

Roadmap:

(i) Independence of categorical variables

(ii) Equality of proportions