# STAT 231 - Statistics

Cameron Roopnarine

Last updated: April 23, 2020

# Contents


**1 Lectures** **2**
1.1 2020-01-20 . . . . . 2
1.2 2020-01-22 . . . . . 4
1.3 2020-01-24 . . . . . 5
1.4 2020-01-27 . . . . . 6
1.5 2020-01-29 . . . . . 8
1.6 2020-01-31 . . . . . 10
1.7 2020-02-03 . . . . . 13
1.8 2020-02-05 . . . . . 15
1.9 2020-02-07 . . . . . 16
1.10 2020-02-10 . . . . . 18
1.11 2020-02-12 . . . . . 20
1.12 2020-02-14 ♡ . . . . . 22
1.13 2020-03-02 . . . . . 24
1.14 2020-03-04 . . . . . 27
1.15 2020-03-06 . . . . . 29
1.16 2020-03-09 . . . . . 30
1.17 2020-03-11 . . . . . 31
1.18 2020-03-13 . . . . . 32

**2 Online Lectures** **35**
2.1 2020-03-16: Testing for Variances . . . . . 35
2.2 2020-03-18: Likelihood Ratio Test Statistic Example . . . . . 36
2.3 2020-03-20: Intro to Gaussian Response Models . . . . . 38
2.4 2020-03-23: MLE Regression . . . . . 39
2.5 2020-03-23: Beta Properties and a Look Ahead . . . . . 40
2.6 2020-03-25: Interval Estimation and Hypothesis for Beta . . . . . 41
2.7 2020-03-26: Pivotal Distribution for Beta and Confidence for the Mean . . . . . 43
2.8 2020-03-28: Prediction Interval and Intro to Model Checking . . . . . 45
2.9 2020-03-29: Model Checking and Final Points . . . . . 46
2.10 2020-03-30: Two Population Case I Equal Variance . . . . . 47
2.11 2020-04-01: Large Samples and Paired Data . . . . . 48
2.12 2020-03-02: The Big Picture–Take 2 . . . . . 50
2.13 2020-03-02: Goodness of Fit . . . . . 52
2.14 2020-03-02: Contingency Tables . . . . . 53

# Chapter 1

# Lectures

## 1.1 2020-01-20

Roadmap:

- Intro
- Big picture of STAT 230 and STAT 231
- Quiz Recap

**EXAMPLE 1.1.1** (STAT 230)**.** A fair die is rolled $60$ times. What is the probability that $12$ of them are sixes? Let $X =$ the number of successes, thus $X \sim \text{Binomial}(60, 1/6)$. Then, we want $P(X = 12)$.

**EXAMPLE 1.1.2** (STAT 231)**.** A die is rolled $60$ times, $12$ of them were sixes. What can we say about the "fairness" of the die?

1. STAT 230: Population $\to$ Sample
2. STAT 231: Sample $\to$ Population

Think of STAT 231 as the "reverse" of STAT 230.

Errors are inevitable Data collection is extremely important. Why do we summarize data?

(a) To identify the "model".

(b) To extract important properties.

How can we summarize data? There are two categories

(1) Numerical: Discrete "count" & Continuous "measure"

(2) Categorical "ordinal": Underlying order

Summary

(a) Numerical

(b) Graphical

Numerical

- Location: mean, median, and mode

- Variability: variance and standard deviation
- Skewness: right-tailed or left-tailed
- Kurtosis: how frequent extreme observations are

Location

- Mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Variability

- Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right] = \frac{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}{n-1}$$

- Standard deviation

$$s = \sqrt{s^2}$$

**EXAMPLE 1.1.3.** Suppose we have 20 observations and the following data is given.

- $\overline{y} = 50$
- $s^2 = 5000$

Suppose one observation is unreliable, say $y_i = 60$. Calculate the new mean.

**Solution.**

$$\begin{aligned}
\overline{y}_{\text{new}} &= \frac{\text{New Total}}{19} \\
&= \frac{\text{Old Total } -60}{19} \\
&= \frac{50 \times 20 - 60}{19} \\
&= \frac{940}{19} \\
&\approx 49.47
\end{aligned}$$

5 Number Summary Let $\{y_{(1)}, \ldots, y_{(n)}\}$ be the sorted data set of $\{y_1, \ldots, y_n\}$ where $y_{(1)}$ is the smallest number, and $y_{(n)}$ is the largest number.

(1) min

(2) $q(0.25)$

(3) $q(0.5)$

(4) $q(0.75)$

(5) max

You can use the rule below to determine the location of $q(p)$ in the sorted list

$$m = (n+1)p$$

- If $m$ is an integer and $1 \leqslant m \leqslant n$, then $q(p) = y_{(m)}$.
- If $m$ is not an integer, but $1 < m < n$, then we determine the closest integer $j$ such that $j < m < j+1$ and then $q(p) = \frac{1}{2}\left(y_{(j)} + y_{(j+1)}\right)$.

Graphical

- Histogram
- Empirical CDF
- Box Plot

The empirical cumulative distribution function is

$$F(y) = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \text{ which are } \leqslant y}{n}$$

## 1.2 2020-01-22

STAT 231: Characteristics of the population are unknown.

Data summary:

- Extract important properties
- Fit the right model

Disappearance of the 400 hitter

- Batting average ?= proportion of successes
- Battling champion = person with the highest batting average
- Before $1950$: 3 champions $\geqslant 400$
- Since $1953$: 0

Question: Why?

Arguments

- Absolute
- Relative
- Better pitchers: Relief
- Better fielding: Glove sizes
- Better managing

The average points of the generic batter is roughly the same over time, but the standard deviation decreases by a lot. Thus, we have a tighter Gaussian distribution for the model today compared to back then since the average player is pretty good (before there was huge variability).

“The median isn’t the message’–Stephen Jay Gould

**DEFINITION 1.2.1.** A ***statistical model*** is a specification of the distribution from which the data set is drawn, where the attribute of interest is a parameter of that distribution.

**EXAMPLE 1.2.2.** A coin is tossed $200$ times with $y = 110$ heads. What can we say about the “fairness” of the coin?
The attribute of interest is

$$P(\text{H}) = \text{probability of heads} = \theta = \text{unknown}$$

Based on our sample, we try to “estimate” $\theta$. Let $Y$ be the number of heads when we toss a coin $200$ times, then our statistical model is: $Y \sim \text{Binomial}(200, \theta)$ with $y = 110$.

**EXAMPLE 1.2.3.** How good are Canadians on Jeopardy? Let $\{y_1, \ldots, y_{10}\}$ be our data set where $y_i$ is the number of shows that the $i^{\text{th}}$ Canadian appeared on.

$$\theta = P(\text{Canadian wins Jeopardy})$$

Is $\hat{\theta} >> 1/3$?

$$\{y_1 = 2, y_2 = 3, y_3 = 1, y_4 = 5\}$$

- $y_1 = \theta(1-\theta)$
- $y_4 = \theta^4(1-\theta)$

Then, our statistical model is $Y_i \sim \text{Geometric}(1-\theta)$ for $i = 1, \ldots, 10$.

Objective: The average salary of a UW co-op student is $\$10000$ per term. Is this claim true? Suppose $\{y_1, \ldots, y_{100}\}$ is given and

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

where each $i \in [1, 100]$ are independent. We will answer this question later in the course.

## 1.3 2020-01-24

Roadmap:

- Statistical models
- Notations and Definitions
- Likelihood function for discrete data
- MLE (Maximum Likelihood Estimate)

**DEFINITION 1.3.1.** A ***model*** is a specification of the experiment (random variable) from which your data set are outcomes.

A coin is tossed $100$ times with $y = 40$ heads. What can we say about the fairness of the coin?

Step 1: Identify the attribute of interest.

$$\begin{aligned}\theta &= P(\text{H}) \\ &= \text{population proportion of heads} \\ &= \text{population parameter} \\ &= \text{unknown constant}\end{aligned}$$

Step 2: Estimate $\theta$ using your data. Based on your data set, what is the "likely" value of $\theta$?

$$\begin{aligned}\hat{\theta}(y_1, \ldots, y_n) &= \text{number that can be calculated using our data set} \\ &= \text{point estimate of } \theta\end{aligned}$$

Step 3: Given $\hat{\theta}$, is $\theta = 0.5$ "reasonable"?

Notation:

- Population parameters are denoted with greek letter such as: $\theta$, $\mu$, $\sigma^2$, $\tilde{n}$
- Data sets are denoted with English letter such as: $y$, $y_1, \ldots, y_n$ when the data set is unknown or $\hat{\theta}$, $\hat{\mu}$ if your data set is known.
- Random variables are denoted with upper case English letters such as: $Y_1, \ldots, Y_n$, $Y$, $Z$

- $y = 40$ heads where $y$ is an outcome of a Binomial experiment. Model:

$$Y \sim \text{Binomial}(100, \theta)$$

**EXAMPLE 1.3.2.** Question: Will trump win Wisconsin in 2020? A sample of 500 people are picked up and 200 of them said that they will vote for Trump. Based on this data will Trump win in 2020?
Let $\theta =$ proportion of the population that vote for Trump

$$Y \sim \text{Binomial}(500, \theta)$$

**EXAMPLE 1.3.3.** Suppose we are interested in the average number of texts a UW math student receives every half hour and $n$ students were interviewed.
Let $\mu$ be the population average of texts received by a UW student.

$$Y_i \sim \text{Poisson}(\mu)$$

for $i = 1, \ldots, n$.

**DEFINITION 1.3.4.** A ***point estimate*** of a parameter is the value of a function of the observed data $y_1, \ldots, y_n$ and other known quantities such as the sample size $n$. We use $\hat{\theta}$ to denote an estimate of the parameter $\theta$.

**DEFINITION 1.3.5.** The ***likelihood function*** for $\theta$ is defined as

$$L(\theta) = L(\theta; \boldsymbol{y}) = P(\boldsymbol{Y} = \boldsymbol{y})$$

for $\theta \in \Omega$ where the ***parameter space*** $\Omega$ is the set of all possible values for $\theta$.

**DEFINITION 1.3.6.** The value of $\theta$ which maximizes $L(\theta)$ for given data $\boldsymbol{y}$ is called the ***maximum likelihood estimate*** (MLE) of $\theta$. It is the value of $\theta$ which maximizes the probability of observing the data $\boldsymbol{y}$. This value is denoted $\hat{\theta}$.

**EXAMPLE 1.3.7.** A coin is tossed 100 times and we get $y = 40$ heads. Let $\theta$ be the probability of heads. Find the MLE of $\theta$.

$$L(\theta) = \binom{100}{40} \theta^{40} (1-\theta)^{60}$$

$$\ell(\theta) = \ln\left[\binom{100}{40}\right] + 40\ln(\theta) + 60\ln(1-\theta)$$

$$\frac{d\ell}{d\theta} = \frac{40}{\theta} - \frac{60}{1-\theta} := 0$$

$$\implies \hat{\theta} = 0.4$$

We can generalize this further.

## 1.4 2020-01-27

<u>Roadmap</u>:

- Statistical Models
- Likelihood and the MLE for discrete

    Binomial

    Poisson

    Geometric
- Invariance property of the MLE
- Relative likelihood function

**DEFINITION 1.4.1.** The ***relative likelihood function*** is defined as

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

for $\theta \in \Omega$. Note that $0 \leqslant R(\theta) \leqslant 1$ for all $\theta \in \Omega$.

**DEFINITION 1.4.2.** The ***log likelihood function*** is defined as

$$\ell(\theta) = \ln\left[L(\theta)\right]$$

for $\theta \in \Omega$.

† Why does maximizing $\ell(\theta)$ also maximize $L(\theta)$? Answer: $\ln(\cdot)$ is an increasing function, in fact it will work for all increasing functions.

Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing monotonic function; that is $t > s \iff g(t) > g(s)$. Suppose $f(\hat{x})$ is maximum for $\hat{x}$. That means $f(\hat{x}) > f(x)$ for all $x$. Thus,

$$g(f(\hat{x})) > g(f(x))$$

Let $t = f(\hat{x})$ and $s = f(x)$. The result now follows.

**PROPOSITION 1.4.3.** *If $Y \sim \mathrm{Binomial}(n, \theta)$ with $y$ successes, then the maximum likelihood estimate for $\theta$ is given by*

$$\hat{\theta} = \frac{y}{n}$$

*Proof.* If $y = 0$, then

$$L(\theta) = P(Y = 0; \theta) = \binom{n}{0}\theta^0(1-\theta)^n = (1-\theta)^{n-0}$$

for $0 \leqslant \theta \leqslant 1$. $L(\theta)$ is a decreasing function for $\theta \in [0, 1]$ and its maximum on the interval $[0, 1]$ occurs at the endpoint $\theta = 0$ and so $\hat{\theta} = 0 = \frac{0}{n}$.

If $y = n$, then

$$L(\theta) = P(Y = n; \theta) = \binom{n}{n}\theta^n(1-\theta)^{n-n} = \theta^n$$

for $0 \leqslant \theta \leqslant 1$. $L(\theta)$ is an increasing function for $\theta \in [0, 1]$ and its maximum on the interval $[0, 1]$ occurs at the endpoint $\theta = 1$ and so $\hat{\theta} = 1 = \frac{n}{n}$.

If $y \neq 0$ and $y \neq n$, then

$$L(\theta) = P(Y = y; \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

for $0 \leqslant \theta \leqslant 1$. Then,

$$\ell(\theta) = \ln\left[\binom{n}{y}\right] + y\ln(\theta) + (n-y)\ln(1-\theta)$$

for $0 < \theta < 1$.

$$\frac{d\ell}{d\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta} = \frac{y-n\theta}{\theta(1-\theta)} := 0$$

$$\implies \hat{\theta} = \frac{y}{n}$$

□

## 1.5 2020-01-29

Roadmap:

- 5 min recap
- Likelihood and the MLE for continuous distributions
- Invariance property of the MLE
- Parameter, Estimate, and Estimator

**DEFINITION 1.5.1.** In many applications, the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are independent and identically distributed (iid) random variables each with probability function $f(y;\theta)$ for $\theta \in \Omega$. We refer to $\boldsymbol{Y}$ as a random sample from the distribution $f(y;\theta)$. In this case, the observed data are $\boldsymbol{y} = (y_1, \ldots, y_n)$ and

$$L(\theta) = L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i;\theta)$$

for $\theta \in \Omega$. Recall that if $Y_1, \ldots, Y_n$ are independent random variables, then their joint probability function is the product of their individual probability functions.

**PROPOSITION 1.5.2.** *Suppose the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is independently drawn from a* $\text{Poisson}(\theta)$ *distribution, where $\theta$ is unknown. The maximum likelihood estimate for $\theta$ is given by*

$$\hat{\theta} = \overline{y}$$

*Proof.* The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) \\ &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \theta^{\sum\limits_{i=1}^{n} y_i} e^{-n\theta} \end{aligned}$$

or more simply

$$L(\theta) = \theta^{n\overline{y}} e^{-n\theta}$$

for $\theta \geqslant 0$. The log likelihood function is

$$\ell(\theta) = n\left[\overline{y}\ln(\theta) - \theta\right]$$

for $\theta > 0$.

$$\frac{d\ell}{d\theta} = n\left(\frac{\overline{y}}{\theta} - 1\right) = \frac{n}{\theta}\left(\overline{y} - \theta\right) := 0$$

$$\implies \hat{\theta} = \overline{y}$$

□

**EXAMPLE 1.5.3.**
- $\mu =$ average time between two volcanic eruptions
- $\boldsymbol{y} = (y_1, \ldots, y_n)$
- $y_i =$ waiting time for the $i^{\text{th}}$ eruption

Model: $Y_i \sim \text{Exponential}(\theta)$ iid

**DEFINITION 1.5.4.** If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are the observed values of a random sample from a distribution with probability distribution function $f(y; \theta)$, then the ***likelihood function*** is defined as

$$L(\theta) = L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i; \theta)$$

for $\theta \in \Omega$.

**PROPOSITION 1.5.5.** *Suppose the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is independently drawn from a* $\text{Exponential}(\theta)$ *distribution, where $\theta$ is unknown. The maximum likelihood estimate for $\theta$ is given by*

$$\hat{\theta} = \overline{y}$$

*Proof.* The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} \\ &= \frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} y_i/\theta\right) \\ &= \theta^{-n} e^{-n\overline{y}/\theta} \end{aligned}$$

for $\theta > 0$. The log likelihood function is

$$\ell(\theta) = -n\left(\ln(\theta) + \frac{\overline{y}}{\theta}\right)$$

for $\theta > 0$.

$$\frac{d\ell}{d\theta} = -n\left(\frac{1}{\theta} - \frac{\overline{y}}{\theta^2}\right) = \frac{n}{\theta^2}\left(\overline{y} - \theta\right) := 0$$

$$\implies \hat{\theta} = \overline{y}$$

□

**EXAMPLE 1.5.6.**
- $\mu =$ average score in STAT 231
- $\sigma^2 =$ variance in STAT 231 scores
- $\boldsymbol{y} = (y_1, \ldots, y_n)$

- $y_i$ = STAT 231 score of the $i^{\text{th}}$ student

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

**PROPOSITION 1.5.7.** *Suppose the data $\boldsymbol{y} = (y_1, \dots, y_n)$ is independently drawn from a $N(\mu, \sigma^2)$ distribution, where $\mu$ and $\sigma$ are unknown. The maximum likelihood estimate for the pair $(\mu, \sigma^2)$ is given by*

$$\hat{\mu} = \overline{y},$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

**THEOREM 1.5.8.** *If $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ is the maximum likelihood estimate of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, then $g(\hat{\boldsymbol{\theta}})$ is the maximum likelihood estimate of $g(\boldsymbol{\theta})$.*

**EXAMPLE 1.5.9.** Suppose $Y_1, \dots, Y_{25} \sim \text{Poisson}(\mu)$ with $\overline{y} = 5$. Find the MLE for $P(Y = 1)$.
**Solution.**

$$P(Y = 1) = \frac{e^{-\mu}\mu^y}{y!} = \frac{e^{-5}5^1}{1!} = \frac{5}{e^5}$$

## 1.6 2020-01-31

Roadmap:

- 5 min recap
- Likelihood function for multinomial
- Testing for the model

  Observed vs Expected frequencies
- Likelihood function and the MLE for the uniform distribution

**EXAMPLE 1.6.1.** The MLE of $\theta$ for

$$f(y; \theta) = \frac{1}{\theta} e^{-y/\theta}$$

is $\hat{\theta} = \overline{y}$. Find the corresponding MLE for $\lambda$ for

$$f(y; \lambda) = \lambda e^{-\lambda y}.$$

**Solution.** Since $\lambda = \frac{1}{\theta}$, we have

$$\hat{\theta} = \overline{y} \implies \frac{1}{\lambda} = \overline{y}$$

by the invariance property. Thus, the MLE for $\lambda$ is

$$\hat{\lambda} = \frac{1}{\overline{y}}.$$

**EXAMPLE 1.6.2.** Suppose $4$ people $(A, B, C, D)$ run a 100 meter race every week. Let $\theta_i$ be the probability person $i$ wins a race for $i \in \{A, B, C, D\}$. Suppose also the following data is given to us.

- $n = 20$
- $y_A = 8$
- $y_B = 6$
- $y_C = 4$
- $y_D = 2$

Model: $Y \sim \text{Multinomial}(n, \theta_A, \dots, \theta_D)$
**Questions:**
(a) What is the likelihood function?
(b) What are the MLEs?
The likelihood function is given by

$$L(\theta_A, \dots, \theta_D) = \frac{20!}{8!6!4!2!}\theta_A^8\theta_B^6\theta_C^4\theta_D^2$$

Intuitively, the MLEs are given by

- $\hat{\theta}_A = \frac{8}{20}$
- $\hat{\theta}_B = \frac{6}{20}$
- $\hat{\theta}_C = \frac{4}{20}$
- $\hat{\theta}_D = \frac{2}{20}$

The Multinomial joint probability function is

$$f(y_1, \dots, y_k; \boldsymbol{\theta}) = \frac{n!}{y_1! \cdots y_k!}\prod_{i=1}^{k}\theta_i^{y_i}$$

for $y_i = 0, 1, \dots$ where $\sum_{i=1}^{k} y_i = n$. The likelihood function for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ based on data $y_1, \dots, y_k$ is given by

$$L(\boldsymbol{\theta}) = L(\theta_1, \dots, \theta_k) = \frac{n!}{y_1! \cdots y_k!}\prod_{i=1}^{k}\theta_i^{y_i}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{k}\theta_i^{y_i}$$

The log likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k}[y_i \ln(\theta_i)]$$

If $y_i$ represents the number of times outcome $i$ occurred in the $n$ "trials" for $i = 1, \dots, k$, then the following result holds.

**PROPOSITION 1.6.3.** *Suppose* $Y \sim \text{Multinomial}(n, \theta_1, \dots, \theta_k)$*, then the MLE for* $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ *is*

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

*for* $i = 1, \dots, k$.

*Proof.* Use Lagrange multiplier method for $\ell(\boldsymbol{\theta})$ satisfying the linear constraint $\sum_{i=1}^{k}\theta_i = 1$. □

**EXAMPLE 1.6.4.** Let $Y$ be a discrete random variable taking values in $\{0, 1, 2, 3\}$ and

$$P(Y=0)=\theta^3,\ P(Y=1)=3\theta(1-\theta)^2,\ P(Y=2)=3\theta^2(1-\theta),\ P(Y=3)=(1-\theta)^3$$

where $\theta$ is an unknown parameter, with $0<\theta<1$. We make a table of $80$ independent observations from the distribution above.

| $Y$ | Observed Frequency |
|---|---|
| 0 | 10 |
| 1 | 30 |
| 2 | 30 |
| 3 | 10 |

(a) Determine the likelihood function, $L(\theta)$.
**Solution.**

$$\begin{aligned} L(\theta) &= \left(\theta^3\right)^{10}\left[3\theta(1-\theta)^2\right]^{30}\left[3\theta^2(1-\theta)\right]^{30}\left[(1-\theta)^3\right]^{10} \\ &= 3^{30}3^{30}\theta^{30}\theta^{30}\theta^{60}(1-\theta)^{60}(1-\theta)^{30}(1-\theta)^{30} \\ &= 3^{30}3^{30}\theta^{120}(1-\theta)^{120} \end{aligned}$$

or more simply

$$L(\theta)=\theta^{120}(1-\theta)^{120}$$

(b) Determine the log likelihood function, $\ell(\theta)$.
**Solution.**

$$\ell(\theta)=120\ln(\theta)+120\ln(1-\theta)$$

or more simply

$$\ell(\theta)=\ln(\theta)+\ln(1-\theta)$$

(c) Using the function $\ell(\theta)$ in (b) in order to derive the maximum likelihood estimate of $\theta$.
**Solution.**

$$\begin{aligned} \frac{d\ell}{d\theta} = \frac{1}{\theta}-\frac{1}{1-\theta} = \frac{1-2\theta}{\theta(1-\theta)} := 0 \\ \implies \hat{\theta}=\frac{1}{2}=0.5 \end{aligned}$$

**EXAMPLE 1.6.5** (Using the likelihood functions to test models)**.** Suppose $W_1,\dots,W_n$ are iid. We collect data $\boldsymbol{w}=(w_1,\dots,w_n)$.
Model: $W_i\sim \text{Poisson}(\theta)$

| $W$ | Observed Frequency | Expected Frequency |
|---|---|---|
| 0 | $y_0$ | $e_1$ |
| 1 | $y_1$ | $e_2$ |
| 2 | $y_2$ | $e_3$ |
| 3 | $y_3$ | $e_4$ |
| 4 | $y_4$ | $e_5$ |
| $\geqslant 5$ | $y_5$ | $e_6$ |

To calculate the expected $e_i$'s we use the formula

$$e_i=n\cdot p_i$$

where

$$p_i=P(Y=i).$$

for $i \in [0,4]$ where $n$ is the total number of observations (observed frequencies summed). For example, $e_i$ would be the following.

$$e_i = n \cdot \left( \frac{e^{-\hat{\theta}} \cdot \hat{\theta}^i}{i!} \right)$$

for $j \in [0,4]$. Note that $\hat{\theta} = \overline{y}$. To estimate $e_5$, we write

$$e_5 = n \cdot P(Y \geqslant 5) = n \cdot \left( 1 - \sum_{i=0}^{4} P(Y = i) \right)$$

Then, we compare the observed frequencies to the expected frequencies.

## 1.7 2020-02-03

Roadmap:

- Review for the midterm
- Likelihood and the MLE for Uniform distribution

**EXAMPLE 1.7.1.** The average number of typos in an academic journal. A random sample of $100$ pages are taken. Let $y_1, \ldots, y_{100}$ be the observed data where $y_i$ is the number of typos in page $i$.

**EXAMPLE 1.7.2.** Average score in STAT 231 and whether STAT 231 scores are correlated with STAT 230 scores. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observed data where

- $x_i =$ STAT 230 score of the $i^{\text{th}}$ student
- $y_i =$ STAT 231 score of the $i^{\text{th}}$ student

Step 1: Identify the population, the parameter of interest, the type of study, variates, attributes (function of the variates), etc.

Step 2: Collect data

- Observational: None of the variables are controlled
- Experimental: Some variables are under the control of the person doing the experiment

Types of problems

- Estimation: We are trying to estimate a population attribute
- Hypothesis testing: Testing a claim made about the population
- Prediction: Predict the “future” value of a variate

Step 3: Summarize data (to identify the model)

- Numerical
- Graphical
- Test whether the model is appropriate

    Compare the CDF to the ECDF

    Compare the theoretical properties

    Compare the observed vs expected frequencies

Step 4: Do the statistical analysis based on your final model

- Parameter: Unknown constant, e.g. $\theta =$ population mean
- Estimate: A number that can be computed from the data set, e.g. $\hat{\theta} =$ (sample mean)
- Estimator: The random variable from which $\hat{\theta}$ is drawn, denoted $\tilde{\theta}$.

Likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

where $f =$ distribution/density function.

$$\ell(\theta) = \ln\left[L(\theta)\right]$$

$\hat{\theta}$ is the MLE of $\hat{\theta}$ that maximizes $L(\theta)$

Measures of Association

- Data set: $(x_1, y_1), \ldots, (x_n, y_n)$

  $x_i =$ number of bears you drink per week

  $y_i =$ STAT 231 score in MT 1

If $x_i > \overline{x}$ and $y_i < \overline{y}$, then

$$(x_i - \overline{x})(y_i - \overline{y}) < 0$$

Sample Correlation

$$r_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 \sum\limits_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Note that we always have $-1 \leqslant r_{xy} \leqslant 1$.

- If $|r_{xy}| \approx 1$, then there is evidence of a strong linear relationship
- If $|r_{xy}| \approx 0$, then there is no evidence of a linear relationship

Note that

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

| | Rich | Poor |
|---|---|---|
| Smoker | $\underbrace{20}_{n_{11}}$ | $\underbrace{80}_{n_{12}}$ |
| Non-smoker | $\underbrace{50}_{n_{21}}$ | $\underbrace{50}_{n_{22}}$ |

$$\begin{aligned}\text{Relative Risk} &= \frac{\frac{20}{20+80}}{\frac{50}{50+50}} \\ &= \frac{\frac{n_{11}}{n_{11}+n_{12}}}{\frac{n_{21}}{n_{21}+n_{22}}}\end{aligned}$$

## 1.8 2020-02-05

Roadmap:

- Two examples

    Likelihood and the MLE for $\text{Uniform}[0, \theta]$

    Discrete example

- PPDAC

    Example and definitions

**EXAMPLE 1.8.1.** $Y_1, \ldots, Y_n$ are iid random variables with $\text{Uniform}[0, \theta]$ where $\theta =$ unknown parameter (attribute) of interest.

- Data set: $(y_1, \ldots, y_n)$ where $y_i > 0$ for each $i \in [1, n]$

What is the MLE for $\theta$.

**Solution.**

$$f(y_i; \theta) = \text{density function}$$

$$f(y_i; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leqslant y_i \leqslant \theta \quad \forall i \in [1, n] \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the likelihood function is

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & 0 \leqslant y_i \leqslant \theta \quad \forall i \in [1, n] \\ 0 & \text{otherwise} \end{cases}$$

Note that $0 \leqslant y_i \leqslant \theta \quad \forall i \in [1, n] \iff \theta > \max\{y_1, \ldots, y_n\}$, thus

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & \theta > \max\{y_1, \ldots, y_n\} \\ 0 & \text{otherwise} \end{cases}$$

Thus, the MLE is

$$\hat{\theta} = \max(y_1, \ldots, y_n)$$

**EXAMPLE 1.8.2.**

- Students come out of a classroom with equal probability
- There are $N$ students in the class identified as $\{1, \ldots, N\}$, where $N$ is unknown
- We observe $3$ students come out $(1, 2, 7)$

What is $\hat{N}$ given your data?

**Solution.**

$$L(N; (1, 2, 7)) = \begin{cases} 0 & N < 7 \\ \binom{N}{3} & N \geqslant 7 \end{cases}$$

Given this likelihood,

$$\hat{N} = 7$$

can be thought of as a discrete version of 1.8.1.

PPDAC A step-by-step, algorithmic approach to a statistical question.

- P: Problem
- P: Plan

- D: Data
- A: Analysis
- C: Conclusion

**EXAMPLE 1.8.3.** We are interested in the attitude of Canadian residents to climate change (whether or not climate change is the number one issue facing the world).
The area of Kitchener-Waterloo and Wellington County were selected and $200$ people were randomly selected and interviewed.
$126$ of them agreed that climate change is the number one issue.

Problem

- What question are we trying to answer?
- Types of problems:

  Descriptive: Estimating attributes of the population

  Causative: Check whether there is a relationship between $x$ and $y$

  Predictive: Predicting (forecasting) future values of a variate
- Target population: The population of interest

  All Canadian residents
- Variate: The property of the unit of the population we are interested in

$$y_i = \begin{cases} 0 & \text{climate change is not the number one issue} \\ 1 & \text{otherwise} \end{cases}$$

- Attribute: A function of the variate

  $\theta =$ proportion of Canadians who believe climate change is the number one issue

Plan

- Study population: The population from which the sample is drawn

  The study population is *usually* a subset of the target population, but **does not** have to be, e.g. medical tests on mice.

## 1.9 2020-02-07

Roadmap:

- PPDAC example
- Interval estimation

  Intervals using the likelihood function

  Confidence intervals

PPDAC

- Problem
- Plan
- Data

- Analysis
- Conclusion

Problem

- What kind of study is this?

  Observational

  Experimental

- What kind of problem is this?

  Descriptive

  Causative

  Predictive

- What is the target population?

  Target population: Population of interest

- What are the variates and attributes of interest?

  Attribute = function of the variate of interest

  $\theta$ = proportion of Canadians who believe climate change is the number one issue

- What is the study population?

  Study population: The act of observing from which the sample is drawn

- What is the sampling protocol?

  How is the sample collected?

- What could be a source of study error?
- What could be a source of sampling error?

Analysis

Data: Try to avoid **bias** where bias is systematic error.

Blind study: Medical tests

- Control group $\to$ Placebo (sugar pill)
- Experimental group $\to$ Actual drug
- The patient does not know.

Double blind study: the doctors do not know

Types of errors

- Study errors: the difference in the value of the attribute between the target population and the study population

  $\phi$ = proportion of people in Kitchener-Waterloo area who believe climate change is the number one issue: $\theta - \phi$

- Sampling errors: the difference in value of the attribute between the study population and the sample: $\phi - \hat{\pi}$ where $\hat{\pi}$ = sample proportion
- Measurement errors: the value of the variate vs what is actually recorded in the data

Conclusion: Non-mathematical discussion of the final result

Interval estimation

Objective:

- To find the "reasonable" values of $\theta$, given by data set
- To quantify the "reasonableness" of your constructed interval

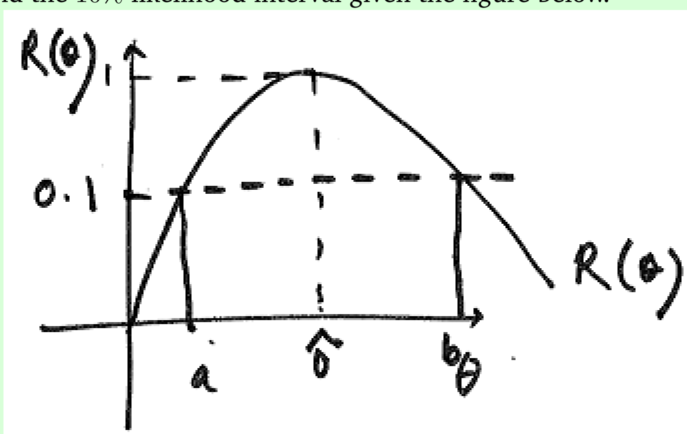
Method 1: Through the likelihood function (likelihood interval)

**DEFINITION 1.9.1.** The $100p\%$ likelihood interval where $p \in [0, 1]$, is given by

$$\{\theta : R(\theta) \geqslant p\}$$

where $R(\theta) =$ relative likelihood function.

**EXAMPLE 1.9.2.** Find the $10\%$ likelihood interval given the figure below.




Guidelines for Interpreting Likelihood Intervals

| |
|---|
| Values of $\theta$ inside a $50\%$ likelihood interval are very plausible in light of the observed data. |
| Values of $\theta$ inside a $10\%$ likelihood interval are plausible in light of the observed data. |
| Values of $\theta$ outside a $10\%$ likelihood interval are implausible in light of the observed data. |
| Values of $\theta$ outside a $1\%$ likelihood interval are very implausible in light of the observed data. |

Clicker Question 1: THE MLE $\hat{\theta}$ is in every likelihood interval for all $p \in [0, 1]$.

(a) **True**

(b) False

Clicker Question 2: If $\theta$ is in the $p\%$ likelihood interval, it has to be in the $q\%$ likelihood interval if $q > p$.

(a) True

(b) **False**

## 1.10 2020-02-10

Roadmap:

- Interval Estimation

Likelihood Estimation

Confidence Intervals: Coverage probabilities, Pivotal Quantities

**EXAMPLE 1.10.1.** The approval rating of Trump is $49\%$ ($49\%$ is the most "likely" value of $\theta$) where $\theta =$ population approval rating.

- What is the "Margin of Error"?
  How does one calculate it?

Setup $Y_1, \ldots, Y_n$ are iid random variables with distribution (density) $f(y;\theta)$ where $\theta =$ unknown attribute.

Objective: Based on our data $\{y_1, \ldots, y_n\}$, we would construct an interval $[a, b]$

$$a(y_1, \ldots, y_n), b(y_1, \ldots, y_n)$$

which are the "reasonable" values of $\theta$.

Method 1: Through the relative likelihood function.

Intuition: $\theta$ is "reasonable" of $L(\theta)$ is "close" to $L(\hat{\theta})$, where $\theta =$ MLE.

**DEFINITION 1.10.2.** A $100p\%$ likelihood interval for $\theta$ where $p \in [0, 1]$

$$\{\theta : R(\theta) \geqslant p\}$$

Take $p = 0.5$, we get that $R(\theta) \geqslant 0.5$, so

$$\implies L(\theta) \geqslant 0.5 L(\hat{\theta})$$

The value of the likelihood at $\theta$ is at least $50\%$ of the value of the likelihood evaluated at the MLE.

Convention

- $R(\theta) \geqslant 0.5 \implies \theta$ is very plausible
- $0.1 \leqslant R(\theta) < 0.5 \implies \theta$ is plausible
- $0.01 \leqslant R(\theta) < 0.1 \implies \theta$ is implausible
- $R(\theta) < 0.01 \implies \theta$ is very implausible

**EXAMPLE 1.10.3.** A coin is tossed $200$ times and we observe $120$ heads. Let $\theta = P(\mathrm{H})$. Is $\theta = 0.5$ plausible?
**Solution.** Find the $10\%$ likelihood interval for $\theta$.

$$L(\theta) = \binom{200}{120} \theta^{120} (1-\theta)^{80}$$

We are given that $\hat{\theta} = 0.6$.

$$\left\{\theta : \frac{\theta^{120}(1-\theta)^{8}0}{0.6^{120}(0.4)^{80}} \geqslant 0.1\right\}$$

Thus,

$$R(\theta) = \frac{\theta^{120}(1-\theta)^{8}0}{0.6^{120}(0.4)^{80}}$$

Is $\theta = 0.5$ plausible? Plug in $\theta = 0.5$ and check if $R(0.5) \geqslant 0.1$.

**EXAMPLE 1.10.4.** Two Binomial experiments.
- $n_1 = 1000$, $y_1 = 200$
- $n_2 = 100$, $y_2 = 20$
- $y =$ number of successes
- $n =$ number of trials

Which $10\%$ likelihood interval is wider?
**Solution.** We have $\hat{\theta} = 0.2$. $n = 100$ yields a wider interval.

Method 2: Confidence intervals.

Setup: There is a pre-specified probability (coverage probability), say $95\%$ or $99\%$ for example.

Objective: Based on your data, we want to estimate the (random) interval which would contain $\theta$ with that probability.

**EXAMPLE 1.10.5.** The STAT 231 scores of UW Math students is normally distributed independently

$$Y_i \sim N(\mu, 64)$$

A sample of $25$ students are collected

$$\overline{y} = 75$$

Find the $95\%$ confidence interval for $\mu$.

Sampling Distributions

Idea: All the data summaries are also outcomes of some random experiment.

$$Y_1, \ldots, Y_n \sim N(\mu, \sigma^2) \quad \text{iid}$$

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

Our sample mean $\overline{y}$ is an outcome of this experiment.

## 1.11 2020-02-12

Roadmap:

- 5 min recap
- Recap of STAT 230

  (Strong) Law of Large #'s

  CLT
- Confidence Interval for the Normal problem with known variance

$$Y_i \sim f(y_i; \theta)$$

$i = 1, \ldots, n$ and $Y_i$'s independent with $\theta =$ unknown parameter.

Likelihood Interval A $10\%$ likelihood interval:

$$\{\theta : R(\theta) \geqslant 0.1\}$$

Notes

(i) The MLE $\theta$ is in every likelihood interval for all $p \in [0, 1]$

(ii) Suppose $\theta$ belongs to the $100p\%$ likelihood interval, then $\theta$ belongs to the $100q\%$ likelihood interval, where $q < p$.

(iii) As $n$ becomes large, the intervals become narrower, for given $p$.

(iv) Plausibility

$$R(\theta) \geqslant 0.5 \implies \text{very plausible}$$

$$\vdots$$

$$R(\theta) < 0.01 \implies \text{very implausible}$$

(v) $\{\theta : R(\theta) \geqslant p\} \iff \{\theta : r(\theta) \geqslant \ln(p)\}$, where $r(\theta) =$ log relative likelihood function

Confidence Interval

**EXAMPLE 1.11.1.** The STAT 231 scores are $N(\mu, 64)$. A sample of 25 students are taken
- $\overline{y} = 75$
- $s^2 = 81$

Given this data, find the $95\%$ confidence interval for $\mu$.

Central Limit Theorem

Law of Large Numbers: $Y_1, \ldots, Y_n$ are iid random variables with mean $\mu$ and variance $\sigma^2$.

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

Then, $\overline{Y}_n \to \mu$ as $n \to \infty$.

CLT: If $Y_1, \ldots, Y_n$ are iid random variables with mean $\mu$ and variance $\sigma^2$, and

$$S_n = \sum_{i=1}^{n} Y_i$$

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

Then, $S_n \sim N(n\mu, n\sigma^2)$ and $\overline{Y}_n \sim N(\mu, \sigma^2/n)$ approximately as $n \to \infty$.

**EXAMPLE 1.11.2.** $Y_1, \ldots, Y_n \sim \text{Exponential}(100)$ with $n = 50$.

$$P(\overline{Y} > 102)$$

$$\overline{Y} \sim N(100, 100^2/50)$$

**EXAMPLE 1.11.3.** $Y \sim \text{Binomial}(n, \theta)$. If $n$ is large, then

$$Y \sim N(n\theta, n\theta(1-\theta))$$

where $Y = Y_1 + \cdots + Y_n$ where $Y_i \sim \text{Bernoulli}(p)$.

**EXAMPLE 1.11.4.** For any iid Normal variables, the result is true for any $n$ (not just large). $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$, then $S_n \sim N(n\mu, n\sigma^2)$ and $\overline{Y}_n \sim N(\mu, \sigma^2/n)$ for all $n$.

Back to the Confidence Interval problem:

Steps

Step 1: Identify the sampling distribution of your estimator.

Step 2: Construct the Pivotal Quantity.

Step 3: Use the pivot to construct the coverage interval.

Step 4: Estimate this interval using your data (confidence interval).

**EXAMPLE 1.11.5.** $Y_1, \ldots, Y_n \sim N(\mu, 64)$ with

- $n = 25$
- $\overline{y} = 75$
- $s^2 = 81$

Objective: To construct a $95\%$ confidence interval.
Step 1: $\hat{\mu} = \overline{y} = 75$, then

$$\overline{Y} \sim N(\mu, 64/25)$$

where $\overline{Y}$ is the sampling distribution of the sample mean.
Step 2: The pivotal quantity is given by

$$\frac{\overline{Y} - \mu}{8/5} = Z \sim N(0,1)$$

Step 3:

$$P\left(-1.96 \leqslant Z \leqslant 1.96\right) = 0.95$$
$$\implies P\left(\overline{Y} - 1.96 \times \frac{8}{5} \leqslant \mu \leqslant \overline{Y} + 1.96 \times \frac{8}{5}\right) = 0.95$$

Step 4: The confidence interval is:

$$\left[\overline{y} - 1.96 \times \frac{8}{5},\ \overline{y} + 1.96 \times \frac{8}{5}\right]$$

Clicker Question: The sample population is always a subset of the target population.

(a) True

(b) **False**

## 1.12 2020-02-14 ♡

Roadmap:

- Confidence interval for a Normal problem with known variance
- The Q-Q-plot, and how to interpret it?

**DEFINITION 1.12.1.** A $100p\%$ confidence interval for $\theta$ is an interval $[\ell, u]$ where $\ell = \ell(y_1, \ldots, y_n)$ and $u = u(y_1, \ldots, y_n)$ which is an estimate of the random interval (coverage interval)

$$[L(Y_1, \ldots, Y_n), U(Y_1, \ldots, Y_n)]$$

such that

$$P\left(L(Y_1, \ldots, Y_n) \leqslant \theta \leqslant U(Y_1, \ldots, Y_n)\right) = p$$

where $p$ is the coverage probability.

Problem: $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$

- $\sigma^2 =$ known
- $\mu =$ unknown parameter of interest
- a probability is pre-specified
- Sample: $\{y_1, \ldots, y_n\}$

Objective: To construct a $95\%$ confidence interval for $\mu$.

Step 1: Identify the sampling distribution of the estimator

- $\mu =$ attribute
- $\overline{y} =$ sample mean $=$ estimate
- $\overline{Y} =$ estimator $= \tilde{\mu}$
- If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, then
$$\overline{Y} \sim N\left(\mu, \sigma^2/n\right)$$

Step 2: Construct the pivotal quantity $Q$

**DEFINITION 1.12.2.** A ***pivotal quantity*** $Q((Y_1, \ldots, Y_n); \theta)$ is a function of $(Y_1, \ldots, Y_n; \theta)$ (a random variable) whose probabilities can be calculated without knowing what $\theta$ is

$$P(Q \geqslant a) \; P(Q \leqslant b)$$

can be calculated without knowing $\theta$.

For example, if $\overline{Y} \sim N(\mu, \sigma^2/n)$, then the pivotal quantity is

$$\frac{\overline{Y} - \mu}{\sigma/n}$$

and the pivotal distribution is $Z$.

Step 3: Find the coverage interval using the pivotal distribution. For $95\%$ we got

$$\left[\overline{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Step 4: Estimate the coverage interval using your data.

Confidence Interval:

$$\left[\overline{y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Notes:

(i) Interpretation of a confidence interval.

Coverage: $\left[\overline{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

Confidence: $\left[\overline{y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

If we did this experiment many times, approximately $95\%$ of the intervals will contain $\mu$.

(ii) As the confidence level increases, the interval is wider.

(iii) Unrealistic example since $\sigma$ is known

(iv) Can we choose the length of the interval? Yes.

The Q-Q-Plot

Model Selection

The Q-Q plot is given by $(y_{(\alpha)}, z_{(\alpha)})$ where

- $y_{(\alpha)} = \alpha^{\text{th}}$ quantile of your data set
- $z_{(\alpha)} = \alpha^{\text{th}}$ quantile of $Z \sim N(0,1)$

If the Q-Q plot is linear, then there is evidence of normality.

Let $Y \sim N(\mu, \sigma^2)$. Show that the Q-Q plot is a straight line.

Clicker Question:

- $n = 100$
- Confidence level: $95\%$

We want to half the length of the interval.

$$\overline{y} \pm a \rightarrow \overline{y} \pm \frac{a}{2}$$

How many more sample points do you need.

(a) $100$

(b) $300$

## 1.13 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 1.13.1.** *Let $Y_1, \dots, Y_n$ be iid $N(\mu, \sigma^2)$ where $\mu$ and $\sigma$ 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 $\mu$ is:*

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

*(ii) The pivotal quantity for $\sigma^2$ is:*

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

**REMARK 1.13.2.** (i) Shows that if we replace $\sigma$ by its estimator $S$, then it follows a $T$-distribution with $(n-1)$ degrees of freedom.

**EXAMPLE 1.13.3.** 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 $\mu$.
(b) Find the $95\%$ confidence interval for $\sigma^2$.
(c) Find the $99\%$ prediction interval for $Y_{26}$.

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

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

We want a $t^*$ such that

$$P(|T_{24}| \leqslant t^*) = 0.99 \iff 2F(t^*) - 1 = 0.99 \iff p = 0.995 = F(t^*)$$

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 for $\mu$ 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 $\sigma^2$ 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.

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

Therefore, the pivotal quantity is:

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

we replace $\sigma$ 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 1.13.4.** Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$. Then,

(i) The general confidence interval for $\mu$ is:

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

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

(ii) The general confidence interval for $\sigma^2$ is:

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

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

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

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

**THEOREM 1.13.5.** *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 1.13.6.** Suppose $n$ is large, and we have a $10\%$ likelihood interval. What is the corresponding coverage probability?
**Solution.** $10\%$ likelihood interval $\implies R(\theta) \geqslant 0.1$

$$\implies \frac{L(\theta)}{L(\hat{\theta})} \geqslant 0.1$$

$$\implies -2\ln\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leqslant -2\ln(0.1)$$

$$\implies \lambda(\theta) \leqslant -2\ln(0.1)$$

Thus, the corresponding coverage:

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

## 1.14 2020-03-04

**DEFINITION 1.14.1.** An estimator $\tilde{\theta}$ is called ***unbiased*** for $\theta$ if

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

**EXAMPLE 1.14.2.** Let $W = \frac{(n-1)S^2}{\sigma^2}$. Prove $S^2$ is an unbiased estimator for $\sigma^2$.
**Solution.**

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

Thus, $S^2$ is an unbiased estimator for $\sigma^2$ by definition.

Other Confidence Intervals

Poisson Suppose $Y_1, \ldots, Y_n \sim \text{Poisson}(\mu)$ 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 1.14.3.** *If $Y \sim \text{Exponential}(\theta)$, then*

$$\frac{2Y}{\theta} \sim \text{Exponential}(2)$$

*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,

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

□

Using this theorem, we can find the confidence interval for $\theta$.

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

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 $\chi^2$ table.

**THEOREM 1.14.4.** *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 1.14.5.** If $p = 0.95$ and $z^* = 1.96$, then the corresponding likelihood is:

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

## 1.15 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 1.15.1.** 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 1.15.2.** 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 1.15.3.** 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**

## 1.16 2020-03-09

Roadmap:

(i) Binomial testing

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

**DEFINITION 1.16.1.** $p$-value: Probability of observing as extreme an observation of your data, given the null hypothesis is true.

**DEFINITION 1.16.2.** 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 1.16.3.** 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: $Y \sim \text{Binomial}(100, \theta)$.

$$D = |Y - 50|$$

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.

## 1.17 2020-03-11

Roadmap:

(i) Testing for normal problems

(ii) How to test for a "bias" of a scale

(iii) One-sided tests

(iv) Relationship between C.I and H.T

(v) Other distributions

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

- $H_0$: $\mu = \mu_0$
- $H_1$: $\mu \neq \mu_0$

Steps involved:

(i) Construct the Discrepancy measure $D$ (satisfying the properties), this measures how much the data disagrees with $H_0$

(ii) Calculate the value of $D$ from your sample ($d$)

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

(iv) Draw appropriate conclusions based on your $p$-value

**EXAMPLE 1.17.1.** The STAT 231 scores are normally distributed with mean $\mu$ and variance $\sigma^2 = 49$.

- $H_0$: $\mu = 75$
- $H_1$: $\mu \neq 75$

A random sample of $25$ students are taken $\overline{y} = 72$. Find the $p$-value.

**Solution.** From Chapter $4$ we know that

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

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

where we can see that $D$ is a legitimate test statistic as it satisfies all the required properties since:

1. $D \geqslant 0$ for all $d$
2. $D = 0 \implies$ best news for $H_0$
3. High values of $D \implies$ bad news for $H_0$
4. Probabilities can be calculated if $H_0$ is true

Thus, we have

$$d = \left|\frac{\overline{y}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| = \left|\frac{72-75}{\frac{7}{\sqrt{5}}}\right| = \frac{15}{7} = 2.14$$

$$\begin{aligned} p\text{-value} &= P(D \geqslant d) \\ &= P(|Z| \geqslant 2.14) \\ &< 0.05 \end{aligned}$$

Evidence against $H_0$.

**EXAMPLE 1.17.2.** UW brochure claims that the average starting salary of UW graduates is \$60000/year. We assume normality. We want to test this claim. Let $\overline{y} = 58000$ and $s = 5000$. What should you conclude?

**Solution.**

- $H_0$: $\mu = 60000$
- $H_1$: $\mu \neq 60000$

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

where all the properties of $D$ are satisfied.

$$d = \left|\frac{\overline{y}-60000}{\frac{5000}{\sqrt{25}}}\right| = 2$$

$$\begin{aligned} p\text{-value} &= P(D \geqslant d) \\ &= P(|T_{24}| \geqslant 2) \end{aligned}$$

The $p$-value for this test is between $5\%$ and $10\%$. Weak evidence against $H_0$.

## 1.18 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 1.18.1.** $Y_1, \ldots Y_n$ iid $N(\mu, \sigma^2)$

- $\sigma^2 =$ known
- $\mu =$ unknown
- Sample: $\{y_1, \ldots, y_n\}$
- $\overline{y} =$ sample mean
- $H_0$: $\mu = \mu_0$ where $\mu_0$ is given
- $H_1$: $\mu \neq \mu_0$

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

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

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

<u>Question</u>: Suppose the $p$-value for the test $> 0.05$ if and only if $\mu_0$ belongs in the $95\%$ confidence interval for $\mu$?

YES.

Suppose $\mu_0$ is in the $95\%$ confidence interval for $\mu$, 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

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

$$p\text{-value} = P(|Z| \geqslant d) > 0.05$$

<u>General result</u> (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 $\theta_0$ belongs to the $100(1-q)\%$ confidence interval and vice versa.

**EXAMPLE 1.18.2** (Bias)**.** A $10$ kg weight is weighed $20$ times $(y_1, \ldots, y_n)$.

- $\overline{y} = 10.5$
- $s = 0.4$
- $H_0$: 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 $d$.

$$d = \left| \frac{\overline{y} - 10}{\frac{s}{\sqrt{n}}} \right| = \left| \frac{10.5 - 10}{\frac{0.4}{\sqrt{20}}} \right| = 5.59017$$

$$\begin{aligned} p\text{-value} &= P(D \geqslant d) \\ &= P(|T_{19}| \geqslant 5.59) \\ &= 1 - P(|T_{19}| \leqslant 5.59) \\ &= 1 - [2P(T_{19} \leqslant 5.59) - 1] \\ &\approx 1 - (2 - 1) \\ &= 0 \end{aligned}$$

Very strong evidence against $H_0$.

**EXAMPLE 1.18.3** (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 $\Lambda$ 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 $\mu$ and $\sigma$ 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 $\chi^2_{n-1}$:

- 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

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

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 $\chi^2_{19}$. We know it will be to the right since the mean of $\chi^2_{19}$ is 19. $\chi^2$ 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 $\sigma^2$.
- 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).

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

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

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

A few final points:

(i) Careful about the previous example:

$n = 20$ is not large

(ii) $\lambda$ and the relationship with $R$:

high values of $\lambda \implies$ low values of $R(\theta_0)$

(iii) Next video

## 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)
- $\mu$ ***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 $\alpha$ and $\beta$.

## 2.4 2020-03-23: MLE Regression

Roadmap:

(i) 5 min recap

(ii) MLE for $\alpha$, $\beta$, $\sigma$

(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.

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

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

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

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

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 $\beta$
- $H_0$: $\beta = 0$ ($x$ 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 $\beta$

(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 $\sigma^2$.

<u>Some Algebra</u>

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

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

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, $\sigma$ 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 $\beta$. We can use*

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

*as a pivotal quantity for $\sigma^2$.*

**EXAMPLE 2.6.3.** Continuation of 2.6.1.
(i) Find the 95% Confidence Interval for $\beta$.
(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 $\beta$

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

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

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

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 $\alpha$? Yes.

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

## 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 $Y_{\text{new}}$ when $x = x_{\text{new}}$.

Difference:

- $\mu$ was constant (stationary target)
- $Y_{\text{new}}$ is a random variable with mean $\mu$ (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: $\sum\limits_{i=1}^{n} r_i = 0$ (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 $(E(Y) = \text{non-linear})$: STAT 4xx

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

Two population problems

Roadmap: Gaussian mean problem with equal variances

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

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

But $\sigma$ 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 $\sigma^2$. 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:

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

$$d = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{10.4 - 9.0}{1.5480\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.0223$$

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 $< 0.05$.

Paired Data: Natural $1-1$ map between the units of the population.

(i) Examples

(ii) Idea of Pivotal Quantity

(iii) Example

(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 $\sigma^2$ (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
- $\ell$ is the number of components of $\theta$ 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 $\theta_i$'s.

$$\begin{aligned} p\text{-value} &= P(\Lambda \geqslant \lambda) \\ &= P(\chi_5^2 \geqslant \lambda) \end{aligned}$$

**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 $i^{\text{th}}$ 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 $\mu$ and $\sigma$. If we were given $\sigma$, 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