STAT 231 - Statistics

Cameron Roopnarine

Last updated: June 6, 2020

Contents

Chapter 1

Introduction to Statistical Sciences

2020-01-06 to 2020-01-17

Lectures $[1, 6]$ have been excluded from these notes since Surya Banerjee was away.

2020-01-20

Roadmap:

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

EXAMPLE 1.0.1 (STAT 230)**.** A **fair** die is rolled 60 times. What is the probability that 12 of them are sixes?
Solution. Let $X =$ the number of successes (sixes), then $X \sim \text{Binomial}(60, 1/6)$.

$$P(X = 12) = \binom{60}{12} \left(\frac{1}{6}\right)^{12} \left(1 - \frac{1}{6}\right)^{60-12} \approx 0.11$$

EXAMPLE 1.0.2 (STAT 231)**.** A die is rolled 60 times and 12 of them were sixes. What can we say about the "fairness" of the die?
Solution. We will solve this answer later.

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.

1.1 Data Summaries

We summarize our data into two categories:

(1) Numerical

(2) Graphical

1.1.1 Numerical Summaries

- Location: mean, median, and mode
- Variability: variance and standard deviation
- Shape: skewness and kurtosis
- Quantiles and Percentiles

Location

- Sample mean $(\bar{y})$:

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

- Sample median $(\hat{m})$:
  - Odd n:

$$\hat{m} = y_{(n/2)}$$

  - Even n:

$$\hat{m} = \frac{1}{2}\left(y_{(\lfloor n/2 \rfloor)} + y_{(\lceil n/2 \rceil)}\right)$$

- Sample mode: value of y which appears in the sample with the highest frequency (not necessarily unique)

The sample mean, median and mode describe the "center" of the distribution of variate values in a data set. Since the median is less affected by a few extreme observations, it is a more robust measure of location.

Variability

- Sample variance $\left(s^2\right)$:

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

- Sample standard deviation (s):

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

- Range:

$$\text{Range} = y_{(n)} - y_{(1)}$$

- IQR:

$$\text{IQR} = q_{(0.75)} - q_{(0.25)}$$

The sample variance and standard deviation measure the variability or spread of the variate values in a data set. Since the interquartile range is less affected by a few extreme observations, it is a more robust measure of variability.

Shape

- Sample skewness (g_1):

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}}$$

- Sample kurtosis (g_2):

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{2}}$$

Measures of shape generally indicate how the data, in terms of a relative frequency histogram, differ from the Normal bell-shaped curve. The sample skewness is a measure of (lack of) symmetry in the data. If the relative frequency histogram of the data has a long right tail, then the sample skewness will be positive. The sample kurtosis measures the heaviness of the tails and the peakedness of the data relative to to data that are Normally distributed. If the sample kurtosis is greater than 3, then this indicates heavier tails (and a more peaked center) than data that are Normally distributed. For data that arise from a model with no tails, for example the Uniform distribution, the sample kurtosis will be less than 3.

EXAMPLE 1.1.1. Suppose we have 20 observations and the following data is given.
- $\bar{y} = 50$
- $s^2 = 5000$

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

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

Sample Quantiles and Percentiles

DEFINITION 1.1.2. Let $\{y_{(1)}, \ldots, y_{(n)}\}$ where $y_{(1)} \leqslant \cdots y_{(n)}$ be the ***order statistic*** for the data set $\{y_1, \ldots, y_n\}$. For $0 < p < 1$, the p^{th} sample quantile (also called the $100p^{\text{th}}$ sample percentile), is a value, call it $q(p)$, determined as follows:
- Let $m = (n+1)p$ where n is the sample size.
- 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 $q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right]$.

DEFINITION 1.1.3. The quantiles $q(0.25)$, $q(0.5)$ and $q(0.75)$ are called the ***lower (first) quartile***, the ***median***, and the ***upper (third) quartile*** respectively.

DEFINITION 1.1.4. The ***interquartile range*** is $\text{IQR} = q(0.75) - q(0.25)$.

DEFINITION 1.1.5. The ***five number summary*** of a data set consist of the smallest observation, the lower quartile, the median, the upper quartile, and the largest value, that is, the five values: $y_{(1)},\ q(0.25),\ q(0.5) = \hat{m},\ q(0.75),\ y_{(n)}$.

1.1.2 Graphical Summaries

- Frequency histogram
- Empirical cumulative distribution function
- Box plots
- Scatter plot
- Run charts

DEFINITION 1.1.6. For a data set $\{y_1, \ldots, y_n\}$, the ***empirical cumulative distribution function*** (e.c.d.f) is defined by

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

for all $y \in \mathbb{R}$. The e.c.d.f is an estimate, based on the data, of the population cumulative distribution function.

Chapter 2

Statistical Models and Maximum Likelihood Estimation

2020-01-22

2.1 Choosing a Statistical Model

STAT 231: Characteristics of the population are unknown.

Data Summary:

- Extract important properties
- Fit the right model

Disappearance of the 400 hitter

- Batting average $\stackrel{?}{=}$ 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

All these arguments are incorrect.

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 2.1.1. 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.

2.2 Maximum Likelihood Estimation

EXAMPLE 2.2.1. 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" θ. 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 2.2.2. 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^{th} Canadian appeared on.

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

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

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

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

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.

2.3 2020-01-24

Roadmap:

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

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 θ using your data. Based on your data set, what is the "likely" value of θ?

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

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

Notation:

- Population parameters are denoted with greek letter such as: θ, μ, σ^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 2.3.1. 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 2.3.2. Suppose we are interested in the average number of texts a UW math student receives every half hour and n students were interviewed.
Let μ be the population average of texts received by a UW student.

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

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

DEFINITION 2.3.3. 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 θ.

DEFINITION 2.3.4. The ***likelihood function*** for θ is defined as

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

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

DEFINITION 2.3.5. The value of θ which maximizes $L(\theta)$ for given data $\boldsymbol{y}$ is called the ***maximum likelihood estimate*** (m.l. estimate) of θ. It is the value of θ which maximizes the probability of observing the data $\boldsymbol{y}$. This value is denoted $\hat{\theta}$.

EXAMPLE 2.3.6. A coin is tossed 100 times and we get $y = 40$ heads. Let θ be the probability of heads. Find the MLE of θ.

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

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

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

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

We can generalize this further.

2.4 2020-01-27

Roadmap:

- Statistical Models
- Likelihood and the MLE for discrete
  - Binomial
  - Poisson
  - Geometric
- Invariance property of the MLE
- Relative likelihood function

DEFINITION 2.4.1. 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 2.4.2. 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 2.4.3. *If $Y \sim \text{Binomial}(n, \theta)$ with y successes, then the maximum likelihood estimate for θ 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 $0 < \theta < 1$.

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

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

□

2.5 2020-01-29

Roadmap:

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

DEFINITION 2.5.1. 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 2.5.2. *Suppose the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is independently drawn from a $\text{Poisson}(\theta)$ distribution, where θ is unknown. The maximum likelihood estimate for θ is given by*

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

Proof. The likelihood function is

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

or more simply

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

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

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

for $\theta > 0$.

$$\begin{aligned} \frac{d\ell}{d\theta} = n\left(\frac{\bar{y}}{\theta} - 1\right) = \frac{n}{\theta}\left(\bar{y} - \theta\right) &:= 0 \\ \implies \hat{\theta} = \bar{y} \end{aligned}$$

□

EXAMPLE 2.5.3.

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

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

DEFINITION 2.5.4. 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 2.5.5. *Suppose the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is independently drawn from a $\text{Exponential}(\theta)$ distribution, where θ is unknown. The maximum likelihood estimate for θ is given by*

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

Proof. The likelihood function is

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

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

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

for $\theta > 0$.

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

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

□

EXAMPLE 2.5.6.
- $\mu =$ 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^{th} student

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

PROPOSITION 2.5.7. *Suppose the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is independently drawn from a $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown. The maximum likelihood estimate for the pair (μ, σ^2) is given by*

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

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

THEOREM 2.5.8. *If $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is the maximum likelihood estimate of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, then $g(\hat{\boldsymbol{\theta}})$ is the maximum likelihood estimate of $g(\boldsymbol{\theta})$.*

EXAMPLE 2.5.9. Suppose $Y_1, \ldots, Y_{25} \sim \text{Poisson}(\mu)$ with $\bar{y} = 5$. Find the MLE for $P(Y = 1)$.
Solution.

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

2.6 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 2.6.1. The MLE of θ for

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

is $\hat{\theta} = \bar{y}$. Find the corresponding MLE for λ for

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

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

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

by the invariance property. Thus, the MLE for λ is

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

EXAMPLE 2.6.2. Suppose 4 people (A, B, C, D) run a 100 meter race every week. Let θ_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: $Y \sim \text{Multinomial}(n, \theta_A, \ldots, \theta_D)$

Questions:

(a) What is the likelihood function?

(b) What are the MLEs?

The likelihood function is given by

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

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

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

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

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

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 $i = 1, \ldots, k$, then the following result holds.

PROPOSITION 2.6.3. *Suppose* $Y \sim \text{Multinomial}(n, \theta_1, \ldots, \theta_k)$*, then the MLE for* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ *is*

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

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

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

EXAMPLE 2.6.4. 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 θ 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 θ.
Solution.

$$\frac{d\ell}{d\theta} = \frac{1}{\theta} - \frac{1}{1-\theta} = \frac{1-2\theta}{\theta(1-\theta)} := 0$$

$$\implies \hat{\theta} = \frac{1}{2} = 0.5$$

EXAMPLE 2.6.5 (Using the likelihood functions to test models)**.** Suppose $W_1, \ldots, W_n$ are iid. We collect data $\boldsymbol{w} = (w_1, \ldots, w_n)$.

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 $\hat{\theta} = \bar{y}$. 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.

2.7 2020-02-03

Roadmap:

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

EXAMPLE 2.7.1. 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 2.7.2. 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^{th} student
- $y_i =$ STAT 231 score of the i^{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 $x_i > \bar{x}$ and $y_i < \bar{y}$, then

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

Sample Correlation

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

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

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

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

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

2.8 2020-02-05

Roadmap:

- Two examples
  - Likelihood and the MLE for $\mathcal{U}[0,\theta]$
  - Discrete example
- PPDAC
  - Example and definitions

EXAMPLE 2.8.1. $Y_1,\ldots,Y_n$ are iid random variables with $\mathcal{U}(0,\theta)$ where θ = 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 θ.

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 2.8.2.
- Students come out of a classroom with equal probability
- There are N students in the class identified as $\{1, \dots, N\}$, 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 2.8.1.

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

- P: Problem
- P: Plan
- D: Data
- A: Analysis
- C: Conclusion

EXAMPLE 2.8.3. 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.

2.9 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 $\rightarrow$ Placebo (sugar pill)
- Experimental group $\rightarrow$ 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 θ, given by data set
- To quantify the "reasonableness" of your constructed interval

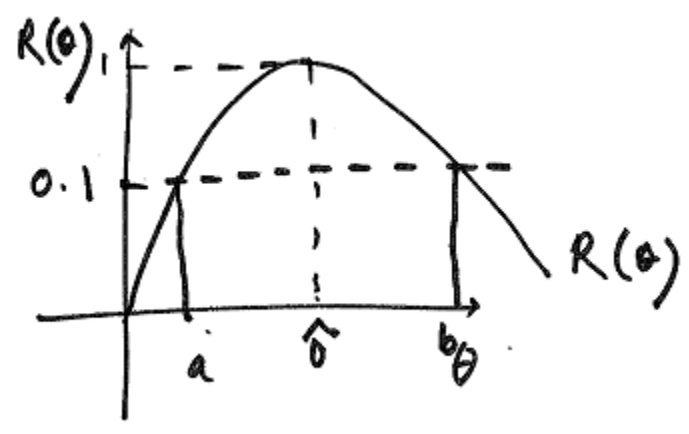
Method 1: Through the likelihood function (likelihood interval)

DEFINITION 2.9.1. The $100p\%$ likelihood interval where $p \in [0, 1]$, is given by

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

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

EXAMPLE 2.9.2. Find the 10% likelihood interval given the figure below.



Guidelines for Interpreting Likelihood Intervals

Values of θ inside a 50% likelihood interval are very plausible in light of the observed data.
Values of θ inside a 10% likelihood interval are plausible in light of the observed data.
Values of θ outside a 10% likelihood interval are implausible in light of the observed data.
Values of θ 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 θ is in the $p\%$ likelihood interval, it has to be in the $q\%$ likelihood interval if $q > p$.

(a) True

(b) **False**

2.10 2020-02-10

Roadmap:

- Interval Estimation
  - Likelihood Estimation
  - Confidence Intervals: Coverage probabilities, Pivotal Quantities

EXAMPLE 2.10.1. The approval rating of Trump is 49% (49% is the most "likely" value of θ) 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 θ.

Method 1: Through the relative likelihood function.

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

DEFINITION 2.10.2. A $100p\%$ likelihood interval for θ 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 θ 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 2.10.3. A coin is tossed 200 times and we observe 120 heads. Let $\theta = P(\text{H})$. Is $\theta = 0.5$ plausible?
Solution. Find the 10% likelihood interval for θ.

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

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

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

Thus,

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

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

EXAMPLE 2.10.4. 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 θ with that probability.

EXAMPLE 2.10.5. 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

$$\bar{y} = 75$$

Find the 95% confidence interval for μ.

Sampling Distributions

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

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

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

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

2.11 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 θ is in every likelihood interval for all $p \in [0, 1]$

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

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

(iv) Plausibility

$$\begin{array}{lcll} R(\theta) & \geqslant & 0.5 & \implies \text{very plausible} \\ & \vdots & & \\ R(\theta) & < & 0.01 & \implies \text{very implausible} \end{array}$$

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

Confidence Interval

EXAMPLE 2.11.1. The STAT 231 scores are $N(\mu, 64)$. A sample of 25 students are taken
- $\bar{y} = 75$
- $s^2 = 81$

Given this data, find the 95% confidence interval for μ.

Central Limit Theorem

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

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

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

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

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

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

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

EXAMPLE 2.11.2. $Y_1, \ldots, Y_n \sim \text{Exponential}(100)$ with $n = 50$.

$$P\left(\bar{Y} > 102\right)$$

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

EXAMPLE 2.11.3. $Y \sim \text{Binomial}(n, \theta)$. If n is large, then

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

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

EXAMPLE 2.11.4. 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 $\bar{Y}_n \sim N(\mu, \sigma^2/n)$ 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 2.11.5. $Y_1, \ldots, Y_n \sim N(\mu, 64)$ with

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

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

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

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

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

Step 3:

$$P\left(-1.96 \leqslant Z \leqslant 1.96\right) = 0.95$$

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

Step 4: The confidence interval is:

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

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

(a) True

(b) **False**

2.12 2020-02-14 ♡

Roadmap:

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

DEFINITION 2.12.1. A $100p\%$ confidence interval for θ 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 μ.

Step 1: Identify the sampling distribution of the estimator

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

$$\bar{Y} \sim N\left(\mu, \sigma^2/n\right)$$

Step 2: Construct the pivotal quantity Q

DEFINITION 2.12.2. 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 θ is

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

can be calculated without knowing θ.

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

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

and the pivotal distribution is Z.

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

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

Step 4: Estimate the coverage interval using your data.

Confidence Interval:

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

Notes:

(i) Interpretation of a confidence interval.

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

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

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

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

(iii) Unrealistic example since σ 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.

Proof.

$$\begin{aligned}
&P(Y \leqslant y_{(\alpha)}) = \alpha \\
&P\left(\frac{Y-\mu}{\sigma} \leqslant \frac{y_{(\alpha)}-\mu}{\sigma}\right) = \alpha \\
&P(Z \leqslant W) = \alpha \\
&F(W) = \alpha \\
&\implies W = z_{(\alpha)} \\
&\implies \frac{y_{(\alpha)}-\mu}{\sigma} = z_{(\alpha)} \\
&\implies y_{(\alpha)} = \mu + \sigma z_{(\alpha)}
\end{aligned}$$

□

Clicker Question:

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

We want to half the length of the interval.

$$\bar{y} \pm a \to \bar{y} \pm \frac{a}{2}$$

How many more sample points do you need.

(a) 100

(b) 300

2.13 2020-02-24

Midterm review session.

2.14 2020-02-26

Roadmap:

- Recap
- Confidence interval for the Binomial problem
- How to choose the "right" sample size?

Confidence Intervals

- $\theta =$ unknown parameter
- $Y_i \sim f(y_i; \theta)$ for $i = 1, \ldots, n$ with Y_i's independent
  - $f =$ distribution (density) function
- Data set: $\{y_1, \ldots, y_n\}$
- $[a, b]$ which is an estimate of the random interval $[A, B]$ which contain θ with the given probability

Step 1: Estimate $\theta \longleftrightarrow \hat{\theta}$ and find the sampling distribution of $\tilde{\theta}$.

- $\hat{\theta} =$ estimate
- $\tilde{\theta} =$ estimator
- $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ where σ^2 is known
- $\hat{\mu} = \bar{y} \longleftrightarrow \hat{\theta}$
- $\bar{Y} \longleftrightarrow \tilde{\theta}$
- Sampling distribution: $\bar{Y} \sim N(\mu, \sigma^2/n)$

Step 2: Construct the pivotal quantity.

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = Z$$

where Z is the pivotal distribution.

Step 3: Construct the coverage interval. The 95% coverage interval is given by

$$\begin{aligned}
&P\left(-1.96 \leqslant Z \leqslant 1.96\right) = 0.95 \\
&P\left(-1.96 \leqslant \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant 1.96\right) = 0.95 \\
&P\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95
\end{aligned}$$

Step 4: Construct the confidence interval.

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

For the normal problem with $\sigma =$ known, the confidence interval is given by

$$\bar{y} \pm z^* \frac{\sigma}{\sqrt{n}}$$

EXAMPLE 2.14.1 (Binomial Distribution)**.** In the 2020 US election, CNN does an exit poll in Wisconsin of 1200 voters.

- 56% voted for Trump
- 44% voted for Bernie Sanders

Find the 95% confidence interval for $\theta =$ proportion of votes that Trump gets.
Model: $Y \sim \text{Binomial}(1200, \theta)$ with $\theta =$ probability of voting for Trump.
Solution.
Step 1: $\hat{\theta} = y/n = 0.56$, $\tilde{\theta} = Y/n$ where $Y \sim N(n\theta, n\theta(1-\theta))$

$$\frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} = Z$$

$$\implies \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} = Z$$

However, if we use this pivotal quantity separating θ could be problematic. Thus, using version 2 of CLT we get

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} = Z$$

is a better pivotal quantity. Thus, the general confidence interval for Binomial is

$$\hat{\theta} \pm z^* \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

$$\implies 0.56 \pm 1.96\sqrt{\frac{0.56 \times 0.44}{1200}} \iff [0.53, 0.59]$$

Even in the worst case scenario, Trump wins (call the election for CNN).
Note that

$$z^* \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$$

is called the ***margin of error***.
Suppose we want the margin of error to be $\leqslant 0.03$ for a 95% interval, then

$$z^* \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}} \leqslant 0.03 \iff n \geqslant \left(\frac{z^*}{0.03}\right)^2 \hat{\theta}\left(1-\hat{\theta}\right)$$

We note that $\hat{\theta} = 0.5$ is the maximum, so we choose n such that

$$n \geqslant \left(\frac{z^*}{0.03}\right)^2 (0.5)(0.5)$$

Thus, for the 95% confidence interval we get

$$n \geqslant \left(\frac{1.96}{0.03}\right)^2 (0.5)(0.5) \approx 1048$$

2.15 2020-02-28

- 5 min recap
- The Chi-squared and the T-distribution
- Normal problem with unknown variance
- Clicker questions

Confidence intervals

Case I: Confidence interval for the mean for normal when σ is known

$$\bar{y} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- $\bar{y} =$ sample mean
- $\sigma =$ population standard deviation
- $n =$ sample size
- z^* depends on the level of confidence
  - $z^* = 1.96$ if confidence level is 95% for every n

Case II: Binomial Confidence

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

- $\theta =$ probability of success (unknown)

Confidence interval is given by

$$\hat{\theta} \pm \underbrace{z^* \sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}}_{\text{margin of error}}$$

- $\hat{\theta} =$ sample proportion
- $n =$ sample size

If we want the margin of error to be $\leqslant \ell$, then

$$n \geqslant \left(\frac{z^*}{\ell}\right)^2 \left(\frac{1}{4}\right)$$

The Chi-Squared Distriubtion

DEFINITION 2.15.1. W is a continuous random variable taking all non-negative values. W is said to follow a ***Chi-Squared*** distribution with n degrees of freedom (d.f), denoted $W \sim \chi^2(n)$, if

$$W = Z_1^2 + \cdots + Z_n^2$$

where $Z_i \sim N(0,1)$ with Z_i's independent.

Properties of the Chi-Squared

(i) $n =$ d.f. = parameter of the Chi-squared. Once n is specified, the d.f. is known

(ii) Density function looks like a gaussian distribution as $df \to \infty$

(iii) If $W \sim \chi^2_n$, then $E(W) = n$ and $Var(W) = 2n$

Cases:

- Case I: $n = 1$, then $W = Z^2$
- Case II: $n = 2$, then $W \sim \text{Exponential}(2)$
- Case III: n is "large", then $W \sim N(n, 2n)$ approximately
- Case IV: n is intermediate, then we use the table

Let (X, Y) be a random point on a Cartesian plane. Assuming X and Y have independent $G(0, 1)$ distributions, the probability that a point is greater than 1.96 away from the origin is

Hint: The distance formula is $x^2 + y^2 = d^2$.

(A) **less than 40**$\%$

(B) at least 40% but less than 60%

(C) at least 60% but less than 80%

(D) at least 80%

Why? We know that $D^2 \sim \text{Exponential}(2)$, then we compute the following.

$$P(D \geqslant 1.96) = 1 - F(1.96) = 1 - \left(1 - \frac{1}{2}e^{-1.96/2}\right) \approx 0.19 = 19\%$$

The Student's T-distribution

DEFINITION 2.15.2. T is said to follow a ***Student's T-distribution*** with n degrees of freedom, denoted $T \sim t(n)$, if

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

where $Z \sim N(0, 1)$ and $W \sim \chi^2(n)$.

Properties

(i) T can take all possible values

(ii) T is symmetric around zero

(iii) Similar to Z, but with flatter tails

(iv) As $n \to +\infty$, then $T \to Z$

Clicker Question:

- $Z \sim N(0, 4)$
- $T \sim t(15)$
- $W \sim \chi^2(3)$
- $Z,\ T,\ W$ are all independent

$\mathbf{E}\left[W + T + \left(\frac{Z}{2}\right)^2\right] =$

(A) 3

(B) **4**

(C) 5

(D) None of the above.

Why?

- $\mathbf{E}[W] = 3$
- $\mathbf{E}[T] = 0$ since T is symmetric around zero for $n > 1$
- Let $Y = \frac{Z}{2}$. Then,

$$\mathbf{E}\left[Y^2\right] = \mathbf{Var}[Y] + \mathbf{E}[Y]^2 = \left(\frac{1}{2}\right)^2 \mathbf{Var}[Z] + \frac{1}{2}\mathbf{E}[Z] = \frac{1}{4}(4) + 0 = 1$$

Thus, $\mathbf{E}[W + T + Y] = 3 + 0 + 1 = 4$.

2.16 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

THEOREM 2.16.1. *Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$ where μ and σ are unknown. Let*

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

and

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

Then,

(i) The pivotal quantity for μ is:

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

(ii) The pivotal quantity for σ^2 is:

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

REMARK 2.16.2. (i) Shows that if we replace σ by its estimator S, then it follows a T-distribution with $(n-1)$ degrees of freedom.

EXAMPLE 2.16.3. An independent sample of 25 students are taken and STAT 231 scores are recorded.

- $\bar{y} = 75$
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 64$

(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

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

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,

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

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

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

Thus, the 99% confidence interval for μ is:

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

(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.

$$Y_{26} \sim N(\mu, \sigma^2)$$

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

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

Therefore, the pivotal quantity is:

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

we replace σ by its estimator and get

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

Thus,

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

yields the general 99% prediction interval:

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

We make the following remark:

REMARK 2.16.4. Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$. Then,

(i) The general confidence interval for μ 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 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 $b - a =$ RHS.

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

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

THEOREM 2.16.5. *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 2.16.6. 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:

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

2.17 2020-03-04

DEFINITION 2.17.1. An estimator $\tilde{\theta}$ is called ***unbiased*** for θ if

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

EXAMPLE 2.17.2. Let $W = \frac{(n-1)S^2}{\sigma^2}$. Prove S^2 is an unbiased estimator for σ^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 σ^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 2.17.3. *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 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

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

where a and b are from the χ^2 table.

THEOREM 2.17.4. *If we have a $p\%$ coverage interval with Z as a pivot, and n is large, then the corresponding likelihood is given by*

$$\exp\left[-(z^*)^2/2\right]$$

EXAMPLE 2.17.5. If $p = 0.95$ and $z^* = 1.96$, then the corresponding likelihood is:

$$\exp\left[-(1.96)^2/2\right] \approx 0.15$$

2.18 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 2.18.1. 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 2.18.2. 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 2.18.3. 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**

2.19 2020-03-09

Roadmap:

(i) Binomial testing

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

DEFINITION 2.19.1. p-value: Probability of observing as extreme an observation of your data, given the null hypothesis is true.

DEFINITION 2.19.2. 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 2.19.3. 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.

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

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

EXAMPLE 2.20.1. The STAT 231 scores are normally distributed with mean μ and variance $\sigma^2 = 49$.
- H_0: $\mu = 75$
- H_1: $\mu \neq 75$

A random sample of 25 students are taken $\bar{y} = 72$. Find the p-value.
Solution. From Chapter 4 we know that

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

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

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

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

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

Evidence against H_0.

EXAMPLE 2.20.2. UW brochure claims that the average starting salary of UW graduates is \$60000/year. We assume normality. We want to test this claim. Let $\bar{y} = 58000$ and $s = 5000$. What should you conclude?
Solution.
- H_0: $\mu = 60000$
- H_1: $\mu \neq 60000$

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

where all the properties of D are satisfied.

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

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

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

2.21 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 2.21.1. $Y_1, \dots Y_n$ iid $N(\mu, \sigma^2)$

- $\sigma^2 =$ known
- $\mu =$ unknown
- Sample: $\{y_1, \dots, y_n\}$
- $\bar{y} =$ sample mean
- H_0: $\mu = \mu_0$ where μ_0 is given
- H_1: $\mu \neq \mu_0$

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

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

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

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.

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

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

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

These two equations yield

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

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

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

- $\bar{y} = 10.5$
- $s = 0.4$
- 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:

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

Compute d.

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

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

Very strong evidence against H_0.

EXAMPLE 2.21.3 (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: $\mu < 3000$ ($\mu \neq 3000$)

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

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

If n is large, then

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

- 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,

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

and

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

Chapter 3

Online Lectures

3.1 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 3.1.1.

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

- 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, $2P(5\%\text{and }\%10) = 10\%$ and20%, which implies $p > 0.1$ and we conclude there is no evidence against null-hypothesis.

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

Exercise For 3.1.1,

- 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:

$$\Lambda(\theta_0) = -2 \ln \left[\frac{L(\theta_0)}{L(\tilde{\theta})} \right] \sim \chi^2_1$$

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

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

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

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

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

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

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

A few final points:

(i) Careful about the previous example:

- $n = 20$ is not large

(ii) λ and the relationship with R:

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

(iii) Next video

3.3 Gaussian Response Models

3.3.1 Intro

EXAMPLE 3.3.1 (STAT 230 and 231 Final Grades)**.**

No.	S230	S231
1	76	76
2	77	79
3	57	54
4	75	64
5	74	64
6	60	60
7	81	85
8	86	82
9	96	88
10	79	72

No.	S230	S231
11	87	76
12	71	50
13	63	75
14	77	72
15	96	84
16	65	69
17	71	43
18	66	60
19	90	96
20	50	50

No.	S230	S231
21	98	83
22	80	88
23	67	52
24	78	75
25	100	99
26	94	94
27	83	83
28	51	37
29	77	90
30	77	67

- Why might we be interested in collecting data such as these?
- What might be a reasonable choice for the target and study population?
- What are the variates? What type are they?
- What is the explanatory variate? What is the response variate?
- How do we summarize these data numerically and graphically?
- What model could we use to analyse these data?

3.3.2 Sample Correlation

Recall that the sample correlation is a numerical measure of the linear relationship between two variates. It is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

Recall that $-1 \leqslant r \leqslant 1$

EXAMPLE 3.3.2 (Sample Correlation for STAT 230/231 Final Grades)**.** Let x be the STAT 230 final grade, and y be the STAT 231 final grade.

For these data, we have

$$S_{xx} = 5135.8667 \quad S_{xy} = 5106.8667 \quad S_{yy} = 7585.3667$$

Thus,

$$r = \frac{5106.8667}{\sqrt{(5135.8667)(7585.3667)}} = 0.82$$

Since r is close to 1, we would say that there is a strong positive linear relationship between STAT 230 and STAT 231 final grades.

3.3.3 Least Squares Estimates

Fitting a Straight Line: Least Squares Approach

To determine the fitted line $y = \alpha + \beta x$, which minimizes the sum of the squares of the distances between the observed points and the fitted line.

We need to find the values of α and β which minimize

$$g(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

These values are determined by simultaneously solving the equations

$$\frac{\partial g}{\partial \alpha} = \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2 (y_i - \alpha - \beta x_i) (-1) = 0$$
$$\frac{\partial g}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2 (y_i - \alpha - \beta x_i) (-x_i) = 0$$

These equations can be written as

$$\bar{y} - \alpha - \beta \bar{x} = 0 \tag{1}$$

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (x_i) = 0 \tag{2}$$

From equation (1), we obtain $\alpha = \bar{y} - \beta \bar{x}$ which we can substitute into equation (2) to obtain

$$\sum_{i=1}^{n} x_i \left[y_i - \bar{y} - \beta (x_i - \bar{x}) \right] = 0$$

or

$$\beta = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Therefore, the least squares estimates are

$$\alpha = \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \quad \text{and} \quad \beta = \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

And, the equation of the fitted line is

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

3.3.4 STAT 231 Versus STAT 230 Final Grades

EXAMPLE 3.3.3 (Fitted Line for STAT 230/231 Final Grades)**.** For the STAT 230/231 data, we have the following

$$\bar{x} = \frac{2302}{30} = 76.7333 \quad \bar{y} = \frac{2167}{30} = 72.2333$$

$$S_{xx} = 5135.8667 \quad S_{xy} = 5106.8667 \quad S_{yy} = 7585.3667$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{5106.8667}{5135.8667} = 0.9944$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 72.2333 - (0.9944)(76.7333) = -4.0667$$

The fitted line is

$$y = -4.0667 + 0.9944x$$

Predicting STAT 231 Final Grade

Given your STAT 230 Final Grade what is the best estimate of your STAT 231 final grade based on these data?

If your final grade in STAT 230 was $x = 75$, then the least squares estimate of your STAT 231 final grade is

$$\hat{y} = -4.0667 + 0.9944(75) = 70.51$$

What can we say about the say about the uncertainty in this estimate?

We need a statistical model in order to obtain an interval estimate of your mark, in particular we would like to construct a confidence interval.

We need a model which captures the fact that not everyone with a final grade of 75 in STAT 230 gets a final grade of 70.51 in STAT 231.

Determining a Model for STAT 230/231 Final Grades

Let's begin by considering the population of students who obtained a final grade of $x = 75$ in STAT 230.

Let $Y =$ STAT 231 final grade of a student drawn at random from this population.

What distribution might we assume for Y?

Since frequency histograms for marks often exhibit a bell shape, we might assume $Y \sim G(\mu, \sigma)$ where μ represents the mean STAT 231 final grade for students in the study population who obtained a final grade of $x = 75$ in STAT 230.

We could then use this model along with the observed data to obtain interval estimates for the mean μ.

Linear Relationship Between STAT 230/231 Final Grades

In our sample of 30 students, we only observed one student with a final grade of 75 in STAT 230.

Does it make sense to do estimation with only one observation?

What to do? We do have 29 other observations.

Since the other 29 students had different STAT 230 final grades, they are observations drawn from populations which have different means (and possibly different variances).

From the scatterplot however, the relationship between STAT 230 and STAT 231 final grades look very linear.

Linear Relationship Between STAT 230/231 Final Grades: Note

Note we have assumed that the standard deviation σ does not depend on x_i.

We will look at ways of assessing whether this assumption is reasonable.

Model for STAT 230/231 Grades

It seems reasonable to assume a model in which the mean STAT 231 final grade for students in the study population who obtained a final grade of x in STAT 230 takes the form:

$$\mu(x) = \alpha + \beta x$$

For the data (x_i, y_i), $i = 1, 2, \ldots, n$, we assume the model

$$Y_i \sim G(\alpha + \beta x_i, \sigma)$$

for $i = 1, 2, \ldots, n$ independently, where x_i is assumed to be a known constant.

This model is usually referred to as a **simple linear regression model**.

3.3.5 Simple Linear Regression Model

In the model, defined by

$$Y_i \sim G(\alpha + \beta x_i, \sigma)$$

for $i = 1, \ldots, n$ independently where x_i is assumed to be a known constant, there are three unknown parameters: α, β, and σ.

For the STAT 230/231 data, the parameter $\mu(x) = \alpha + \beta x$ represents the mean STAT 231 final grade in the study population of students with a STAT 230 final grade equal to x.

Simple Linear Regression Model: Parameters α and β

What does the parameter β represent?

The parameter β represents the change in the mean $\mu(x) = \alpha + \beta x$ for a unit increase in x.

What does the parameter α represent?

The parameter α represents the mean in the study population of students with a STAT 230 final grade equal to 0. (Note: In this example, the parameter α is not of interest since students with a final grade of 0 cannot take STAT 231!)

3.3.6 Maximum Likelihood Estimates

Since our assumed model is

$$Y_i \sim G(\alpha + \beta x_i, \sigma), \quad \text{for } i = 1, \ldots, n \text{ independently}$$

where x_i is assumed to be a known constant, the likelihood function for α and β (assuming for the moment that σ is known) is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right] \end{aligned}$$

where we have ignored terms not involving α and β.

To obtain the maximum likelihood estimates of α and β, we would maximize

$$L(\alpha, \beta) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right], \alpha \in \mathbb{R}, \beta \in \mathbb{R}$$

with respect to α and β, or equivalently minimize

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

with respect to α and β.

This is a problem we've already solved!

Fitting a Straight Line: Maximum Likelihood Approach

Important Result

For the model

$$Y_i \sim G(\alpha + \beta x_i, \sigma), \text{for } i = 1, 2, \ldots, n \text{ independently}$$

where x_i is assumed to be a known constant, the maximum likelihood estimates of α and β (usually called the **regression parameters**) are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

which are also the least square estimates of α and β.

So if the least squares estimates and the maximum likelihood estimates are the same, why do we actually assume a probability model? Why complicate things?

We need a probability model to model the variability in the data. Remember, not every student with a 75 in STAT 230 obtain the same STAT 231 final grade.

3.3.7 Distribution of the Maximum Likelihood Estimator of the Slope, Beta

In order to construct a confidence interval for your STAT 231 mark, we need to derive some distributional results.

Important Result

The first of these is

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

where

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

The distribution of $\tilde{\beta}$ is determined by the assumption

$$Y_i \sim G(\alpha + \beta x_i, \sigma), \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

where x_i, $i = 1, 2, \ldots, n$, are assumed to be known constants.

Since

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

is a linear combination of independent normal random variables.

Thus, by a theorem learned in STAT 230, $\tilde{\beta}$ has a normal distribution.

We only need to find $\mathbf{E}\left[\tilde{\beta}\right]$ and $\mathbf{Var}\left[\tilde{\beta}\right]$.

$$\begin{aligned}
\mathbf{E}\left[\tilde{\beta}\right] &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \mathbf{E}\left[Y_i\right] \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} (\alpha + \beta x_i) \qquad \text{since} \mathbf{E}\left[Y_i\right] = \alpha + \beta x_i \\
&= \beta \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} x_i \qquad \text{since} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \alpha = \frac{\alpha}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \\
&= \beta \frac{S_{xx}}{S_{xx}} \\
&= \beta
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbf{Var}\left[\tilde{\beta}\right] &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(S_{xx})^2} \mathbf{Var}\left[Y_i\right] \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(S_{xx})^2} \sigma^2 \qquad \text{since} \mathbf{Var}\left[Y_i\right] = \sigma^2 \\
&= \frac{\sigma^2}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{(S_{xx})^2} S_{xx} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Thus,

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

3.3.8 Distribution of the Variance Estimator

Estimate of the Variance in Simple Linear Regression

Since σ is usually unknown, we estimate it using

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 = \frac{1}{n-2} \left(S_{yy} - \hat{\beta} S_{xy}\right)$$

Note: $\sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$ is usually called the **sum of the squared errors** and s_e^2, the unbiased estimator, is called the **mean squared error**. s_e^2 is not the maximum likelihood estimate of σ^2. The maximum likelihood estimate is similar, but with a denominator of n rather than $n-2$.

We use S_e^2 as the estimator of σ^2 since it can be shown that $\mathbf{E}\left[S_e^2\right] = \sigma^2$ where

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \tilde{\alpha} - \tilde{\beta} x_i\right)^2 \quad \tilde{\beta} = \frac{S_{xy}}{S_{xx}} \quad \tilde{\alpha} = \bar{Y} - \tilde{\beta} \bar{x}$$

Distribution of the Estimator of the Variance

It can be shown that

$$\frac{(n-2) S_e^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \tilde{\alpha} - \tilde{\beta} x_i\right)^2 \sim \chi^2(n-2)$$

The proof is beyond the scope of this course, but it is usually proved in a third year linear regression course.

Note that there are $n-2$ degrees of freedom due to the two restrictions:

- $\sum_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)=0$
- $\sum_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)x_i=0$

Recall that these 2 equations 2 unknowns determined the estimators $\tilde{\alpha}$ and $\tilde{\beta}$.

3.3.9 Constructing the Confidence Interval for β

Since

$$\frac{\tilde{\beta}-\beta}{\sigma/\sqrt{S_{xx}}}\sim G(0,1) \quad \text{and} \quad \frac{(n-2)S_e^2}{\sigma^2}\sim\chi^2(n-2)$$

independently, it follows that

$$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}}\sim t(n-2)$$

This t pivotal quantity can be used to construct confidence intervals and tests of hypotheses for β.

$100p\%$ Confidence Interval and p-Value for β

A $100p\%$ confidence interval for β is given by

$$\hat{\beta}\pm a\frac{s_e}{\sqrt{S_{xx}}}$$

where $P(T\leqslant a)=\frac{(1+p)}{2}$ and $T\sim t(n-2)$.

For testing $H_0:\beta=\beta_0$,

$$p\text{-value } =2\left[1-P\left(T\leq\frac{\left|\hat{\beta}-\beta_0\right|}{s_e/\sqrt{S_{xx}}}\right)\right], \text{where} T\sim t(n-2)$$

3.3.10 Hypothesis of No Relationship

Since $\mu(x)=\alpha+\beta x$, a test of $H_0:\beta=0$ is a test of the Hypothesis that the mean $\mu(x)$ does not depend on x.

This hypothesis is usually referred to as the hypothesis of no relationship between variates Y and x.

EXAMPLE 3.3.4 (Estimates of β and σ for STAT 230/231 Final Grades)**.**

$$\bar{x}=\frac{2302}{30}=76.7333 \quad \bar{y}=\frac{2167}{30}=72.2333$$

$$S_{xx}=5135.8667 \quad S_{xy}=5106.8667 \quad S_{yy}=7585.3667$$

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}}=\frac{5106.8667}{5135.8667}=0.9944$$

$$s_e=\sqrt{\frac{1}{n-2}\left(S_{yy}-\hat{\beta}S_{xy}\right)}=\sqrt{\frac{1}{28}[7585.3667-(0.9944)(5106.8667)]}=9.4630$$

3.3.11 Interferences for the Slope

EXAMPLE 3.3.5 (95% Confidence Interval for β for STAT 230/231 Final Grades)**.** Since

$$P(T \leq 2.0484) = \frac{1+0.95}{2} = 0.975 \quad \text{where} T \sim t(28)$$

a 95% confidence interval for β is

$$\hat{\beta} \pm 2.0484 \frac{s_e}{\sqrt{S_{xx}}} = 0.9944 \pm 2.0484 \frac{9.4630}{\sqrt{5135.8667}}$$

or

$$[0.7239, 1.2648]$$

EXAMPLE 3.3.6 (Testing $H_0 : \beta = 0$ for STAT 230/231 Final Grades)**.** Since the 95% confidence interval, $[0.7239, 1.2648]$, does not contain the value $\beta = 0$, the p-value for testing $H_0 : \beta = 0$ is smaller than 0.05.
This means there is evidence against the hypothesis of no relationship between STAT 231 final grades and STAT 230 final grades.

EXAMPLE 3.3.7 (p-Value for STAT 230/231 Final Grades)**.** The actual p-value for testing $H_0 : \beta = 0$ is

$$\begin{aligned} p\text{-value} &= 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - 0|}{s_e/\sqrt{S_{xx}}}\right)\right] \quad \text{where} T \sim t(28) \\ &= 2\left[1 - P\left(T \leq \frac{0.9944}{9.4630/\sqrt{5135.8667}}\right)\right] \\ &= 2[1 - P(T \leq 7.5304)] \\ &\approx 0 \end{aligned}$$

Therefore, there is very strong evidence against the hypothesis of no relationship between STAT 230 and STAT 231 final grades.

What does the hypothesis $H_0 : \beta = 1$ represent?

Recall, that the slope β represents the change in STAT 231 final grade for a unit change of 1 mark in the STAT 230 final grade in the study population.

EXAMPLE 3.3.8 (Testing $H_0 : \beta = 1$ for STAT 230/231 Final Grades)**.** Since the 95% confidence interval, $[0.7239, 1.2648]$, does contain the value $\beta = 1$, the p-value for testing $H_0 : \beta = 1$ is larger than 0.05 and there is no evidence against the hypothesis $H_0 : \beta = 1$.
The actual p-value for testing $H_0 : \beta = 1$ is

$$\begin{aligned} p\text{-value} &= 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - 1|}{s_e/\sqrt{S_{xx}}}\right)\right] \quad \text{where} T \sim t(28) \\ &= 2\left[1 - P\left(T \leq \frac{|0.9944 - 1|}{9.4630/\sqrt{5135.8667}}\right)\right] \\ &= 2[1 - P(T \leq 0.0428)] \end{aligned}$$

Since $P(T \leqslant 0.2558) = 0.6$,

$$p\text{-value} \geqslant 2(1 - 0.6) = 0.8$$

3.3.12 Interferences for the Mean Response at x

Suppose we wanted a confidence interval for the mean STAT 231 final grade for students who obtained a final grade of 75 in STAT 230; that is, we want a confidence for $\mu(75) = \alpha + \beta(75)$.

More generally we are often interested in a confidence interval for the mean response $\mu(x) = \alpha + \beta x$ for a specified value of x.

The maximum likelihood estimator of $\mu(x)$ is obtained by replacing the unknown values of α and β by their maximum likelihood estimators, which gives

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x})$$

since $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$.

We will now show that

$$\begin{aligned}\tilde{\mu}(x) &= \tilde{\alpha} + \tilde{\beta}x \\ &= \bar{Y} + \tilde{\beta}\,(x - \bar{x}) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)\end{aligned}$$

or

$$\tilde{\alpha} + \tilde{\beta}x \sim G\left(\alpha + \beta x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

We notice that the standard deviation is larger for values of x which are not close to the center or mean of the x data.

3.3.13 Distribution of the Estimator of the Mean Response at x

Recall our model assumption

$$Y_i \sim G\left(\alpha + \beta x_i, \sigma\right) \quad \text{for } i = 1, \ldots, n \text{ independently}$$

where x_i is assumed to be a known constant, and the fact that

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

Then,

$$\tilde{\mu}(x) = \bar{Y} + \tilde{\beta}(x - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i + (x - \bar{x}) \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i = \sum_{i=1}^{n} a_i Y_i$$

where

$$a_i = \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}$$

Since $\tilde{\mu}(x)$ is a linear combination of Gaussian random variables, it has a Gaussian distribution. We only need to find $\mathbf{E}\left[\tilde{\mu}(x)\right]$ and $\mathbf{Var}\left[\tilde{\mu}(x)\right]$. Facts:

$$\sum_{i=1}^{n} a_i = 1$$

$$\sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} \frac{1}{n} + \sum_{i=1}^{n} x_i\,(x - \bar{x})\,\frac{(x_i - \bar{x})}{S_{xx}} = \bar{x} + (x - \bar{x}) \sum_{i=1}^{n} x_i \frac{(x_i - \bar{x})}{S_{xx}} = x$$

Thus,

$$\begin{aligned}\mathbf{E}\left[\tilde{\mu}(x)\right] &= \sum_{i=1}^{n} a_i \mathbf{E}\left[Y_i\right] \\ &= \sum_{i=1}^{n} a_i \left(\alpha + \beta x_i\right) \\ &= \alpha \sum_{i=1}^{n} a_i + \beta \left(\sum_{i=1}^{n} a_i x_i\right) \\ &= \alpha + \beta x \\ &= \mu(x)\end{aligned}$$

$$\begin{aligned}&\mathbf{Var}\left[\tilde{\mu}(x)\right] \\ &= \sum_{i=1}^{n} a_i \mathbf{Var}\left[Y_i\right] \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + \frac{2}{n}\frac{\left(x-\bar{x}\right)\left(x_i-\bar{x}\right)}{S_{xx}} + \frac{\left(x-\bar{x}\right)^2\left(x_i-\bar{x}\right)^2}{\left(S_{xx}\right)^2}\right] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\left(x-\bar{x}\right)^2}{S_{xx}}\right]\end{aligned}$$

Therefore, we have that

$$\mu(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

3.3.14 Interferences for the Mean

Since

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1) \quad \text{and} \quad \frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$$

independently, it follows that

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

This pivotal quantity can be used to construct confidence intervals for $\mu(x)$ and to test hypotheses about $\mu(x)$.

Confidence Interval for the Mean Response at x

A $100p\%$ confidence interval for $\mu(x) = \alpha + \beta x$ is given by

$$\hat{\mu}(x) \pm a s_e \sqrt{\frac{1}{n} + \frac{\left(x-\bar{x}\right)^2}{S_{xx}}} = \hat{\alpha} + \hat{\beta}x \pm a s_e \sqrt{\frac{1}{n} + \frac{\left(x-\bar{x}\right)^2}{S_{xx}}}$$

where $P(T \leqslant a) = \frac{1+p}{2}$ and $T \sim t(n-2)$.

EXAMPLE 3.3.9 (95% Confidence Interval for the Mean STAT 231 Final Grade)**.** Since

$$P(T \leqslant 2.0484) = \frac{1+0.95}{2} = 0.975 \quad \text{where} \quad T \sim t(28)$$

a 95% confidence interval for the mean STAT 231 final grade for students who obtained a final grade of 75 in STAT 230 is

$$\begin{aligned}\hat{\alpha} + \hat{\beta}x \pm as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} &= -4.0667 + 0.9944(75) \pm 2.0484(9.4630)\sqrt{\frac{1}{30} + \frac{(75-76.7333)^2}{5135.8667}} \\ &= 70.51 \pm 3.5699\end{aligned}$$

or

$$[66.9, 74.1]$$

Confidence Interval for the y-Intercept, α

Since $\mu(0) = \alpha + \beta(0) = \alpha$, a $100p\%$ confidence interval for α, is given by

$$\hat{\alpha} \pm as_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

If $\bar{x}$ is large in magnitude (which means the average x_i is large), then the confidence interval for α will be very wide.

This would be disturbing if the value $x = 0$ is a value of interest, but it often not.

Confidence Interval for the Variance

The pivotal quantity

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

can be used to construct intervals for σ^2.

A $100p\%$ confidence interval for σ^2 is given by

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

where

$$P(U \leqslant a) = \frac{1-p}{2} = P(U > b), \quad \text{where} \quad U \sim \chi^2(n-2)$$

A $100p\%$ confidence interval for σ is given by

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

3.3.15 Interference for an Individual Response Y at x

Suppose we wanted an interval for $Y =$ STAT 231 final grade for one student who obtained a final grade of $x = 75$ in STAT 230.

Now

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x \sim G\left(\alpha + \beta x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

and

$$Y \sim G(\alpha + \beta x, \sigma)$$

independently.

To obtain a confidence interval for Y, we first obtain the distribution of the random variable $Y - \tilde{\mu}(x)$ using its mean and variance.

Since

$$\mathbf{E}\left[Y - \tilde{\mu}(x)\right] = 0$$

and

$$\begin{aligned}\mathbf{Var}\left[Y - \tilde{\mu}(x)\right] &= \mathbf{Var}\left[Y\right] + \mathbf{Var}\left[\tilde{\mu}(x)\right] \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})}{S_{xx}}\right]\end{aligned}$$

we have

$$Y - \tilde{\mu}(x) \sim G\left(0, \sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}\right)$$

Since

$$Y - \tilde{\mu}(x) \sim G\left(0, \sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}\right)$$

and

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

independently, we have

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

This pivotal quantity can be used to construct a prediction interval for Y.

3.3.16 A $100p\%$ Prediction Interval for a Future Response Y

The corresponding interval

$$\hat{\mu}(x) \pm as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} = \hat{\alpha} + \hat{\beta}x \pm as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where a is the value from the t-table such that

$$P(T \leq a) = \frac{1+p}{2}, \quad \text{where} T \sim t(n-2)$$

We usually call such an interval a $100p\%$ prediction interval instead of a confidence interval, since Y is not a parameter but a random variable.

EXAMPLE 3.3.10 (Prediction Interval Example)**.** A 95% prediction interval for the STAT 231 final grade

for a student who obtained a final grade of 75 in STAT 230 is

$$\hat{\alpha}+\hat{\beta}x \pm as_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}} = -4.0667+0.9944(75)\pm 2.0484(9.4630)\sqrt{1+\frac{1}{30}+\frac{(75-76.7333)^2}{5135.8667}}$$
$$= 70.51 \pm 19.7100$$

or

$$[50.8, 90.2]$$

3.3.17 Gaussian Response Models

The simple linear regression model we have just considered,

$$Y_i \sim G(\alpha+\beta x_i, \sigma), \quad \text{for } i=1,2,\ldots,n \text{ independently}$$

where x_i is assumed to be a known constant, is a member of a larger family of models called **Gaussian response models**.

The general form of a Gaussian response model is

$$Y_i \sim G\left(\mu\left(x_i\right), \sigma\right), \quad \text{for } i=1,2,\ldots,n \text{ independently}$$

where x_i is assumed to be a known constant, and where x_i can be a vector of explanatory variates also called **covariates**.

Note that the mean of Y_i depends on x_i, which may be a vector of variates (also called covariates in the linear regression model) or a scalar. However, the standard deviation σ does not depend on x_i.

The Gaussian response model

$$Y_i \sim G\left(\mu\left(x_i\right), \sigma\right), \quad \text{for } i=1,2,\ldots,n \text{ independently}$$

can also be written in the form

$$Y_i = \mu\left(x_i\right)+R_i, \quad \text{where } R_i \sim G(0,\sigma), i=1,2,\ldots,n \text{ independently}$$

Y_i is the sum of two componenets.

The first component, $\mu(x_i)$, is a deterministic component (not a random variable), and the second component R_i is a random component or random variable.

Linear Regression Models

In many examples,

$$\mu(x_i) = \beta_0 + \sum_{j=1}^{k} B_j x_{ij}$$

so the mean of Y_i is a linear function of $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, the vector of covariates for unit i and the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

These models are called **linear regression models**, and the B_j's are called the **regression parameters.**

3.3.18 Model Checking

There are two main assumptions for Gaussian linear response models:

1. Y_i (given covariates x_i) is Gaussian with standard deviation σ which does not depend on covariates.

2. $\mathbf{E}[Y_i] = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.

MODEL ASSUMPTIONS SHOULD ALWAYS BE CHECKED!!!

We will use graphical methods to do this.

Model Checking Method 1: Scatterplot with Fitted Line

In a simple linear regression, a scatter plot of the data with the fitted line superimposed shows us how well the model fits.

Model Checking Method 2: Residual Plots

Residual plots are very useful for model checking when there are two or more explanatory variates.

DEFINITION 3.3.11. For the simple linear regression model, let

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$$

(often called the "fitted" response), and let

$$\hat{r}_i = y_i - \hat{\mu}_i, \quad i = 1, 2, \ldots, n$$

The $\hat{r}_i$'s are **residuals** since $\hat{r}_i$ represents what is "left" after the model has been "fitted" to the data.

The $\hat{r}_i$'s can often be though of as "observed" R_i's in the model $Y_i = \mu_i + R_i$, where $R_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently.

This isn't exactly correct since we are using $\hat{\mu}_i$ instead of μ_i, but if the model is correct, then the $\hat{r}_i$'s should behave roughly like a random sample from the $G(0, \sigma)$ distribution.

Recall $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, which implies that $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$ or

$$0 = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y}_i - \hat{\alpha} - \hat{\beta}\bar{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i$$

A plot of the points $(x_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$ should lie more or less within a horizontal band around the line $\hat{r}_i = 0$.

Standardized Residual Plots

DEFINITION 3.3.12. Define the **standardized residuals** as

$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e}, \quad i = 1, 2, \ldots, n$$

What is the only difference between a plot of points $(x_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$ and a plot of points $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$.

If the model is correct, then the $\hat{r}_i^*$ values will lie in the range of $(-3, 3)$. Why is this?

Model Checking Method 3: Plot of Residuals Versus Expected

Another type of residual plot can be used to check the assumption about the form of the mean. It consists of plotting the points $(\hat{\mu}_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$.

For the simple linear regression model, we are checking whether the assumed mean

$$\mathbf{E}[Y_i] = \mu(x_i) = \alpha + \beta x_i$$

is reasonable.

if the assumed mean is reasonable, we should see approximately a horizontal band around the line $\hat{r}_i = 0$.

A plot of $(\hat{\mu}_i, \hat{r}_i^*)$ looks like the plot of $(\hat{\mu}_i, \hat{r}_i)$. The only difference is that the vertical axis is rescaled.

Model Checking Method 4: Qqplot of Standardized Residuals

To check the normality assumption, we use a qqplot of the standardized residuals.

Since our assumed model is

$$\frac{Y_i - \mu_i}{\sigma} \sim G(0,1)$$

the $\hat{r}_i^*$'s should represent a sample (not quite random since $\sum_{i=1}^{n} \hat{r}_i^* = 0$) from the $G(0,1)$ distribution.

Therefore, a qqplot of the $\hat{r}_i^*$ terms should give approximately a straight line if the normality assumption holds.

Intepreting Residual Plots

If a plot of the points $(x_i, \hat{r}_i)$ or $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$ shows a distinctive pattern, then this suggests the assumed form for $\mu(x_i)$ may be inappropriate.

If a plot of the points $(\hat{\mu}_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$ indicates that the variability in the $\hat{r}_i$'s is bigger for large values of $\hat{\mu}_i$ than for small values of $\hat{\mu}_i$ (or vice versa), thn there is evidence to suggest that the assumption of constant variance $\mathbf{Var}[Y_i] = \mathbf{Var}[R_i] = \sigma^2$, $i = 1, 2, \ldots, n$ does not hold.

Intepreting Residual Plots: Warning

Reading these plots takes practice. You should try not to read too much into plots especially if the plots are based on a small number of points.

The following plots exhibit patterns.

3.4 Comparing the Means of Two Populations

EXAMPLE 3.4.1 (Hand Span Example)**.** Suppose we wanted to answer the question: Are the hand spans of females enrolled in STAT 231 in Winter 2015 different on average from the hand spans of males enrolled in STAT 231 in Winter 2015?
To do this we test the hypothesis that there is no difference in mean hand spans between males and females enrolled in STAT 231 in Winter 2015.
Let Y_{1i} = the hand span of the ith male, $i = 1, 2, \ldots, 78$, and let Y_{2i} = the hand span of the ith female, $i = 1, 2, \ldots, 64$.
Based on these observations, a Gaussian model seems reasonable for both the Y_{1i}'s and Y_{2i}'s.
Assume Y_{1i}, $i = 1, 2, \ldots, 78$ is a random sample from a $G(\mu_1, \sigma)$ distribution, and independently Y_{2i}, $i = 1, 2, \ldots, 64$ is a random sample from a $G(\mu_2, \sigma)$ distribution.
We call this a two sample Normal or Gaussian problem.
Note that we have assumed both Gaussian populations have the same standard deviation, σ.
There are three unknown parameters in the model: μ_1, μ_2, and σ.
The parameter μ_1 represents the mean hand span in centimeters for males enrolled in STAT 231 in Winter 2015 (the study population). Note that we are assuming there is no bias in the measurements due to the measurement system.
The parameter μ_2 represents the mean hand span in centimeters for females enrolled in STAT 231 in Winter 2015 (the study population).
The hypothesis of interest is $H_0 : \mu_1 = \mu_2$, or $H_0 : \mu_1 - \mu_2 = 0$.

3.4.1 Special Case of the Gaussian Response Model

By letting

$$\begin{aligned} Y_i = Y_{1i}, i = 1, 2, \ldots, n_1 \quad \text{and} Y_{n_1+i} = Y_{2i}, \quad i = 1, 2, \ldots, n_2 \\ \mathbf{E}[Y_i] = \mu_1, i = 1, 2, \ldots, n_1 \quad \text{and} EY_{n_1+i} = \mu_2, \quad i = 1, 2, \ldots, n_2 \end{aligned}$$

and

$$\mathbf{Var}[Y_i] = \sigma^2, \quad i = 1, 2, \ldots, n_1 + n_2$$

we can see that this model is just a special case of the Gaussian response model.

$$Y_i \sim G(\mu(x_i, \sigma)), \quad i = 1, 2, \ldots, n \text{ independently}$$

where $\mu(x_i) = \mu_1$, $i = 1, 2, \ldots, n_1$ and $\mu(x_i) = \mu_2$, $i = n_1 + 1, 2, \ldots, n_1 + n_2$.

Likelihood Function for Random Samples from Two Gaussian Populations

The likelihood function for μ_1, μ_2, and σ is

$$L(\mu_1, \mu_2, \sigma) = \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_{1i} - \mu_1}{\sigma}\right)^2\right] \prod_{i=1}^{n_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_{2i} - \mu_2}{\sigma}\right)^2\right]$$

or more simply

$$L(\mu_1, \mu_2, \sigma) = \sigma^{-(n_1+n_2)} \exp\left[-\frac{1}{2}\sum_{i=1}^{n_1}\left(\frac{y_{1i} - \mu_1}{\sigma}\right)^2\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n_2}\left(\frac{y_{2i} - \mu_2}{\sigma}\right)^2\right]$$

$\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, and $\sigma > 0$.

Maximum Likelihood Estimators

The log likelihood function is

$$\ell(\mu_1, \mu_2, \sigma) = -(n_1 + n_2)\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_{1i} - \mu_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(y_{2i} - \mu_2)^2$$

Maximization of this function gives the maximum likelihood estimators

$$\tilde{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} = \bar{Y}_1$$

$$\tilde{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_1} Y_{2i} = \bar{Y}_2$$

$$\tilde{\sigma}^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}\left(Y_{1i} - \bar{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \bar{Y}_2\right)^2\right]$$

3.4.2 Pooled Estimator of Variance

Define

$$S_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}\left(Y_{1i} - \bar{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \bar{Y}_2\right)^2\right]$$

Note that

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

where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_{1i} - \bar{Y}_1\right)^2$$

is the sample variance for the data from population 1, and

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_{2i} - \bar{Y}_2\right)^2$$

is the sample variance for the data from population 2.

DEFINITION 3.4.2.

$$\begin{aligned} S_p^2 &= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} \left(Y_{1i} - \bar{Y}_1\right)^2 + \sum_{i=1}^{n_2} \left(Y_{2i} - \bar{Y}_2\right)^2 \right] \\ &= \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} \end{aligned}$$

S_p^2 is called the **pooled estimator of variance**, since it is obtained by "pooling" the estimators S_1^2 and S_2^2 of σ^2 from the two samples.

Why does this estimator make sense?

The degrees of freedom are $n_1 + n_2 - 2$ because of the two restrictions

$$\sum_{i=1}^{n_1} \left(Y_{1i} - \bar{Y}_1\right) = 0 \quad \text{and} \quad \sum_{i=1}^{n_2} \left(Y_{2i} - \bar{Y}_2\right) = 0$$

It cal also be shown that

$$\frac{(n_1 + n_2 - 2) S_p^2}{\sigma^2} = \frac{\sum_{i=1}^{n_1} \left(Y_{1i} - \bar{Y}_1\right)^2 + \sum_{i=1}^{n_2} \left(Y_{2i} - \bar{Y}_2\right)^2}{\sigma^2} \sim \chi^2 (n_1 + n_2 - 2)$$

3.4.3 Inferences for the Difference Between the Means

To make inferences about the mean difference $\mu_1 - \mu_2$, we note that $\tilde{\mu}_1 = \bar{Y}_1$ is a point estimator of μ_1 and $\tilde{\mu}_2 = \bar{Y}_2$ is a point estimator of μ_2, so that $\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2$ is a point estimator of $\mu_1 - \mu_2$.

Since

$$\bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \quad \text{and} \quad \bar{Y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) \quad \text{independently}$$

we have

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

Pivotal Quantity for Confidence Interval for Mean Difference

Since

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0, 1)$$

and

$$\frac{(n_1 + n_2 - 2) S_p^2}{\sigma^2} \sim \chi^2 (n_1 + n_2 - 2)$$

independently, then

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

and this is the pivotal quantity that we can use to make inferences about the mean difference $\mu_1 - \mu_2$ when σ is unknown.

Confidence Interval for Difference in Means

Since

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

is a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is given by

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

where

$$P(T \leq a) = \frac{1+p}{2} \quad \text{and} \quad T \sim t\,(n_1 + n_2 - 2)$$

3.4.4 Test of Hypothesis for No Difference in Means

Since

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

to test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$, we use the test statistic

$$D = \frac{|\bar{Y}_1 - \bar{Y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Let the observed value be

$$d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

then the p-value is

$$p\text{-value } = 2[1 - P(T \leq d)], \quad \text{where} T \sim t\,(n_1 + n_2 - 2)$$

EXAMPLE 3.4.3 (Hand Span Example: Test of Hypothesis for No Difference in Means)**.** For the males:

$$\hat{\mu}_1 = \bar{y}_1 = 21.50 \quad \text{and} \quad s_1^2 = 3.4309$$

For the females:

$$\hat{\mu}_2 = \bar{y}_2 = 19.37 \quad \text{and} \quad s_2^2 = 2.055$$

with

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2 = 21.50 - 19.37 = 2.14$$

and

$$s_p^2 = \frac{(77)(3.4309) + (63)(2.055)}{78 + 64 - 2} = 2.8117 \quad \text{and} \quad s_p = \sqrt{2.8117} = 1.6768$$

EXAMPLE 3.4.4 (Hand Span Example: 95% Confidence Interval for $\mu_1 - \mu_2$)**.** Using R, we obtain

$P(T \leqslant 1.97705) = 0.975$, where $T \sim t(140)$. A 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned}\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= 2.14 \pm (1.97705)(1.6768)\sqrt{\frac{1}{78} + \frac{1}{64}} \\ &= 2.14 \pm 0.5591 \\ &= [1.58, 2.70]\end{aligned}$$

Since $\mu_1 - \mu_2 = 0$ is not contained in this 95% confidence interval, we already know that the p-value for testing $H_0 : \mu_1 - \mu_2 = 0$ is less than 0.05.

Hand Span Example: p-Value

Since

$$d = \frac{|2.14 - 0|}{(1.67687)\sqrt{\frac{1}{78} + \frac{1}{64}}} = 7.56$$

the actual p-value is

$$\begin{aligned}p\text{-value} &= 2[1 - P(T \leq 7.56)], \quad \text{where} T \sim t(140) \\ &\approx 0\end{aligned}$$

Therefore, there is very strong evidence to contradict the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ based on the data. The difference is statistically significant. Is the difference of practical significance?

3.4.5 Comparison of Two Means, Unequal Variances

Recall that the previous analysis depends on the assumptions

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from $G(\mu_1, \sigma_1)$
- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from $G(\mu_2, \sigma_2)$
- the two samples are independent
- $\sigma_1 = \sigma_2$

What if $\sigma_1 = \sigma_2$ is not a reasonable assumption?

Note: $H_0 : \sigma_1 = \sigma_2$ could be tested using a likelihood ratio test.

Approximate Pivotal Quantity, Unequal Unknown Variances

If n_1 and n_2 are both large, then the approximate pivotal quantity is

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim G(0,1) \text{approximately}$$

This pivotal quantity can be used to construct confidence intervals and test hypotheses for the mean difference $\mu_1 - \mu_2$.

EXAMPLE 3.4.5. For example, an approximate 95% confidence interval for $\mu_1 - \mu_2$ would be given by

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

EXAMPLE 3.4.6 (Hand Span Example: 95% Confidence Interval for $\mu_1 - \mu_2$)**.** For the males we have the following:

$$n_1 = 78, \quad \hat{\mu}_1 = \bar{y}_1 = 21.50, \quad \text{and} s_1^2 = 3.4309$$

For the females we have the following:

$$n_2 = 64, \quad \hat{\mu}_2 = \bar{y}_2 = 19.37, \quad \text{and} s_2^2 = 2.055$$

An approximate 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 21.50 - 19.37 \pm 1.96\sqrt{\frac{3.4309}{78} + \frac{2.055}{64}} = [1.60, 2.68]$$

as compared to $[1.58, 2.70]$.

3.5 Gaussian Response Models

3.5.1 Bean Experiment

In my Winter 2015 STAT 231 class, I conducted the following experiment.

Each student was given two small paper cups. One cup contained 30 black beans, and the other cup was empty.

Each student was also given a piece of paper on which two circles (the size of the paper cup bottom) were drawn.

Students were asked to place one paper cup on each circle so the cups were the same distance apart for each student, and then to use one hand to hold the cup containing the beans in place. Using the other hand, they were asked to move as many black beans as possible to the empty cup, one at a time, in a timed 15 -second interval. The students then performed the same task using the opposite hands.

Students were randomized with respect to whether they moved the beans with their dominant or nondominant hand first.

Bean Experiment Data

Difference in the Number of Beans Moved by the Dominant and Non-Dominant Hands

Difference	Frequency
-4	2
-3	1
-2	1
-1	4
0	13
1	15
2	14
3	5
4	2
Total	57

- Is there a difference in the mean number of beans moved in 15 seconds between the dominant and non-dominant hands?
- How do we analyze these data? What model should we assume? Is it a two sample problem?
- The assumptions for the two sample model are:

  $$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \text{is a random sample from} G(\mu_1, \sigma)$$

  and independently

  $$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \text{is a random sample from} G(\mu_2, \sigma)$$

Bean Experiment: Correlation in the Data

Let Y_{1i} = number of beans moved using the dominant hand, and let Y_{2i} = number of beans moved using the non-dominant hand for the ith student.

Does it seem reasonable to assume the Y_{1i}'s are independent of the Y_{2i}'s?

We could expect the observations on the i th student (Y_{1i}, Y_{2i}) to be positively correlated. That is, we would expect $\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right] > 0$.

3.5.2 Paired Experiment

In fact, the observations from the bean experiment have been deliberately paired to eliminate some factors (e.g., finger size, agility, competitive spirit, etc.) which might otherwise affect conclusions about the parameter of interest, which is the mean difference $\mu_1 - \mu_2$

The bean experiment is an example of a **paired experiment**.

For a paired experiment (can you show this?)

$$\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2}{n}\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right]$$

If $\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right] > 0$, then $\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right]$ is smaller than for an unpaired experiment.

Paired Experiment: Making Inferences about the Difference $\mu = \mu_1 - \mu_2$

To make inferences about $\mu = \mu_1 - \mu_2$, we analyze the within-pair differences

$$Y_i = Y_{1i} - Y_{2i}, \quad i = 1, 2, \ldots, n$$

We assume

$$Y_i = Y_{1i} - Y_{2i} \sim G\left(\mu_1 - \mu_2, \sigma\right), \quad i = 1, 2, \ldots, n$$

independently.

We can now use the one-sample analysis that we used previously for analysing a random sample from a $G(\mu, \sigma)$ distribution, with $\mu = \mu_1 - \mu_2$.

Bean Experiment: Testing the Null Hypothesis

For the bean data,

$$\bar{y} = 0.86 \quad \text{and} \quad s = \left[\frac{1}{56}\sum_{i=1}^{57}\left(y_i - \bar{y}\right)^2\right]^{1/2} = 1.66$$

To test $H_0 : \mu = 0$, we use the test statistic

$$D = \frac{|\bar{Y} - 0|}{S/\sqrt{n}}$$

with the observed value

$$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}} = \frac{|0.86 - 0|}{1.66/\sqrt{57}} = 3.90$$

and p-value

$$p\text{-value } = 2[1 - P(T \leq 3.90)] \approx 0, \quad \text{where} T \sim t(56)$$

and there is strong evidence against $H_0 : \mu = 0$ based on the observed data.

Bean Experiment: Practical Versus Statistical Significance

The difference is statistically significant.

Is the difference of practical significance?

3.5.3 Examples of Experiments on Differences Between Means

Examples in which the parameter of interest is the mean difference $\mu_1 - \mu_2$

1. Test for the difference in execution time between two algorithms A and B with randomly generated data.
2. Test for a difference in the error rates or speeds of two algorithm designed for image resolution or character/speech recognition on many different scenarios/problems.
3. Test whether one numerical method for nonlinear optimization is faster than another on a large population of potential test functions.
4. Artificial intelligence: test whether one learning algorithm learns a task faster than another.

Experimental Design I

Generate or select n_1 random "problems" (for example, data sets to be sorted, functions to be minimized, images to be resolved), and compute the mean execution time $\bar{Y}_1$ of algorithm A.

Generate or select another n_2 ($n_1 = n_2$ possibly) random "problems", and compute the mean execution time $\bar{Y}_2$ of algorithm B.

Estimate the difference as $\bar{Y}_1 - \bar{Y}_2$.

In this case, $\bar{Y}_1$ and $\bar{Y}_2$ are independent random variables, thus

$$\mathbf{E}\left[\bar{Y}_1 - \bar{Y}_2\right] = \mu_1 - \mu_2$$

and

$$\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right] = \mathbf{Var}\left[\bar{Y}_1\right] + \mathbf{Var}\left[\bar{Y}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Experimental Design II

Generate or select n random "problems" (for example, data sets to be sorted, functions to be minimized, images to be resolved), and compute the mean execution time $\bar{Y}_1$ of algorithm A.

Compute the mean execution time $\bar{Y}_2$ of algorithm B on the same set of n problems.

Estimate the difference as $\bar{Y}_1 - \bar{Y}_2$.

In this case $\bar{Y}_1, \bar{Y}_2$ are dependent random variables, thus

$$\mathbf{E}\left[\bar{Y}_1 - \bar{Y}_2\right] = \mu_1 - \mu_2$$

and

$$\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2}{n}\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right]$$

If $\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right] > 0$, then $\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right]$ is smaller for Design II.

3.5.4 Pairing as a Design Choice

We expect that the covariance between the execution times of algorithms on the same problem to be positively co rrelated (harder problems have longer execution times).

A sample of dependent pairs (Y_{1i}, Y_{2i}) is better than two independent random samples (one of Y_{1i}'s and one of Y_{2i}'s) for estimating $\mu_1 - \mu_2$, since the difference $\mu_1 - \mu_2$, can be estimated more accurately (shorter confidence intervals) if $\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right] > 0$.

Note: If $\mathbf{Cov}\left[Y_{1i}, Y_{2i}\right] < 0$, then pairing is a bad idea since it increases the value of $\mathbf{Var}\left[\bar{Y}_1 - \bar{Y}_2\right]$.

In a paired experiment, we do not assume that Y_{1i} and Y_{2i} are independent random variables. We do, however, assume the differences $Y_i = Y_{1i} - Y_{2i}, i = 1, 2, \ldots, n$ are independent (all different problems).

Paired Versus Unpaired

When you see data from a comparative study (i.e., one whose objective is to compare two distributions, often through their means), you have to determine whether it involves paired data or not.

Of course, a sample of Y_{1i}'s and Y_{2i}'s cannot be from a paired study unless there are equal numbers of each, but if there are equal numbers, the study might be either "paired" or "unpaired".

3.6 Multinomial Models and Goodness of Fit

3.6.1 Multinomial Models and Goodness of Fit

Is the distribution of colours uniform?

Smarties Data

Colour	Observed Number	Expected Number
Red	80	$610\left(\frac{1}{8}\right) = 76.25$
Green	73	76.25
Yellow	77	76.25
Blue	107	76.25
Purple	61	76.25
Brown	73	76.25
Orange	72	76.25
Pink	77	76.25
Total	610	610

How do we conduct a formal test of the hypothesis that the distribution of different colours is uniform?

Model and Hypothesis

$$\begin{aligned} P\left(\text{observing the data } y_1, \ldots, y_k; \theta_1, \ldots, \theta_k\right) &= P\left(Y_1 = y_1, \ldots, Y_k = y_k; \theta_1, \ldots, \theta_k\right) \\ &= \frac{n!}{y_1! \cdots y_k!} \theta_1^{y_1} \ldots \theta_k^{y_k} \end{aligned}$$

where $0 < \theta_j < 1, \sum_{j=1}^{k} \theta_j = 1, y_j = 0, 1, \ldots$, and $\sum_{j=1}^{k} y_j = n$.

For our example, $Y_j =$ number of Smarties of colour $j, j = 1, 2, \ldots, 8$ and $k = 8$.

We want to test the hypothesis

$$H_0 : \theta_j = \frac{1}{k}, \quad \text{for all} j = 1, 2, \ldots, k$$

3.6.2 Multinomial Likelihood Function

The multinomial likelihood function is

$$L\left(\theta_1, \theta_2, \ldots, \theta_k\right) = \frac{n}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

or more simply

$$L\left(\theta_1, \theta_2, \ldots, \theta_k\right) = \theta_1^{y_1} \theta_2^{y_2} \ldots \theta_k^{y_k} = \prod_{j=1}^{k} \theta_j^{y_j}$$

The maximum likelihood estimates are

$$\hat{\boldsymbol{\theta}}_j = \frac{\boldsymbol{y}_j}{n}, \quad j = 1, 2, \ldots, k$$

and the maximum likelihood estimators are

$$\tilde{\theta}_j = \frac{Y_j}{n}, \quad j = 1, 2, \ldots, k$$

Multinomial Likelihood Ratio Test Statistic

For testing $H_0 : \theta = \theta_0 = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$, we use

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

where

$$\tilde{\theta} = \left(\frac{Y_1}{n}, \frac{Y_2}{n}, \ldots, \frac{Y_k}{n}\right)$$

Now

$$\begin{aligned}\frac{L\left(\theta_0\right)}{L(\tilde{\theta})} &= \left[\prod_{j=1}^{k}\left(\frac{1}{k}\right)^{Y_j}\right] \div \left[\prod_{j=1}^{k}\left(\frac{Y_j}{n}\right)^{Y_j}\right] \\ &= \prod_{j=1}^{k}\left(\frac{n/k}{Y_j}\right)^{Y_j} = \prod_{j=1}^{k}\left(\frac{E_j}{Y_j}\right)^{Y_j}, \quad \text{where} E_j = n\left(\frac{1}{k}\right) = \frac{n}{k}\end{aligned}$$

Note that Y_j is the observed number, and E_j is the expected number of observations in category j if H_0 is true.

Therefore, the likelihood ratio test statistic for testing $H_0 : \theta = \theta_0 = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$ is

$$\begin{aligned}\Lambda\left(\theta_0\right) &= -2\log\left[\frac{L\left(\theta_0\right)}{L(\tilde{\theta})}\right] = -2\log\left[\prod_{j=1}^{k}\left(\frac{E_j}{Y_j}\right)^{Y_j}\right] \\ &= 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right) \\ &= 2\sum_{j=1}^{k}(\text{observed frequency }) \times \log\left(\frac{\text{observed frequency}}{\text{expected frequency}}\right)\end{aligned}$$

Why does this test statistic make sense? When does it take on large values?

When does it take on small values?

Behaviour of the Multinomial Test Statistic

If all of the observed frequencies (the Y_j's) equal the expected frequencies under H_0 (the E_j's), then $\log\left(\frac{Y_j}{E_j}\right) = 0$, and $\Lambda\left(\theta_0\right) = 0$; otherwise, $\Lambda\left(\theta_0\right) > 0$.

So large values of $\Lambda\left(\theta_0\right)$ give evidence against H_0.

3.6.3 Distribution of the Multinomial Likelihood Ratio Test Statistic

Multinomial Likelihood Ratio Test Statistic

The likelihood ratio test statistic for testing $H_0 : \theta_j = \frac{1}{8}, j = 1, 2, \ldots, 8$ is

$$\begin{aligned}\Lambda(\theta_0) &= -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] \\ &= 2\sum_{j=1}^{8} Y_j \log\left(\frac{Y_j}{E_j}\right) \\ &= 2\sum_{j=1}^{8}(\text{observed frequency}) \times \log\left(\frac{\text{observed frequency}}{\text{expected frequency}}\right)\end{aligned}$$

where Y_j is the observed number and E_j is the expected number of observations in category j if H_0 is true.

Distribution of the Multinomial Likelihood Ratio Test Statistic

Recall that if θ is a scalar, then $\Lambda(\theta_0)$ has approximately a $\chi^2(1)$ distribution for large n if $H_0 : \theta = \theta_0$ is true.

If θ is a vector, then $\Lambda(\theta_0)$ still has approximately a χ^2 distribution for large n if $H_0 : \theta = \theta_0$ is true, but the degrees of freedom change.

The degrees of freedom in the multiparameter case depend on both how many parameters are unknown in the original model, and how many parameters must be estimated under the null hypothesis.

The multinomial likelihood function $L(\theta_1, \theta_2, \ldots, \theta_8)$ is a function of $8 - 1 = 7$ parameters.

Under the hypothesis $H_0 : \theta_j = \frac{1}{8}, j = 1, 2, \ldots, 8$, there are no unspecified parameters and no parameters needed to be estimated.

Therefore, $\Lambda(\theta_0)$ has approximately a $\chi^2(7)$ distribution if $H_0 : \theta_j = \frac{1}{8}$, $j = 1, 2, \ldots, 8$ is true.

Approximate p-Value for Smarties Data

The observed value of the likelihood ratio test statistic is

$$\begin{aligned}\lambda(\theta_0) &= 2\sum_{j=1}^{8} y_j \log\left(\frac{y_j}{e_j}\right), \quad \text{where} e_j = 610\left(\frac{1}{8}\right) = 76.25 \\ &= 2\left[80\log\left(\frac{80}{76.25}\right) + \cdots + 77\log\left(\frac{77}{76.25}\right)\right] = 35.00\end{aligned}$$

The approximate p -value is

$$\begin{aligned}p\text{-value} &\approx P(W \geq 35), \quad \text{where} W \sim \chi^2(7) \\ &\approx 0\end{aligned}$$

and there is very strong evidence based on the observed data against the hypothesis of an equal number of each colour.

Note: The Chi-squared approximation is good when n is large and the expected frequencies under H_0 are all at least five.

3.6.4 Pearson's Chi-Squared Goodness of Fit Statistic

An alternative test statistic that was developed historically before the likelihood ratio test statistic is the Pearson Goodness of Fit Statistic:

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}$$

Note that when $Y_j = E_j$ for all j, then $D = 0$. Otherwise, $D \geq 0$.

For large n, D and Λ are asymptotically equivalent and have the same asymptotic Chi-squared distribution.

The Pearson Goodness of Fit Test Statistic is more popular than the Likelihood Ratio Test.

3.7 Goodness of Fit Examples

3.7.1 Checking the Fit of the Model

In order to study a data set, we typically assume a model in which $(Y_1, Y_2, \ldots, Y_n)$ is a random sample from a distribution which is a member of the family of models

$$f(y;\theta), \quad \text{for} \theta \in \Omega$$

It is important to check that the model adequately represents the variability in the data. This can be done by comparing the observed data $(y_1, y_2, \ldots, y_n)$ with what we would expect to get using the model.

One way to do this is to compare the observed frequencies based on the data with the expected frequencies calculated using probabilities from the assumed model If the model is suitable, then the observed and expected frequencies should be "close".

Discrete Data Example: Alpha-Particle Emissions

Recall a previous example regarding the data collected by the scientists, Rutherford and Geiger on the number of alpha-particles emitted during a fixed time interval.

We examined the fit of the Poisson model to these data by comparing the observed frequencies with the expected frequencies calculated assuming a Poisson model with mean equal to the sample mean.

Observed and Expected Frequencies Under Assumed Poisson Model

Number of Alpha-Particles Detected	Freq. f_j	Expected Freq. e_j
0	57	54.42
1	203	210.42
2	383	407.43
3	525	525.54
4	532	508.41
5	408	393.47
6	273	253.77
7	139	140.28
8	45	67.86
9	27	29.18
10	10	11.29
11+	6	5.77
Total	2608	2607.99

We decided based on the frequencies in this table that a Poisson model fits these data reasonably well.

The drawback of this method is we don't have a good way of deciding how "close" is good enough.

We will now see how to do a formal test of the hypothesis that the Poisson model fits these data.

Alpha-Particle Emissions: Test of Fit of Poisson Model

Our model is Multinomial $(n; \theta_0, \theta_1, \ldots, \theta_{11})$, and $\sum_{j=0}^{11} \theta_j = 1$, which is a function of 11 parameters. H_0: Data fit a Poisson model or, more specifically,

$$H_0 : \theta_j = \frac{\theta^j e^{-\theta}}{j!}, \quad j = 0, 1, \ldots, 10$$

Under H_0, there is one unknown parameter θ which must be estimated. Therefore, the degrees of freedom for the Chi-squared approximation for the likelihood ratio test statistic are $11 - 1 = 10$

Note that the expected frequencies are all at least five, so we can use the Chi-squared approximation to obtain the p -value. The observed value of the likelihood ratio statistic is

$$\begin{aligned} 2\sum_{j=0}^{11} f_j \log\left(\frac{f_j}{e_j}\right) &= 2\left[57\log\left(\frac{57}{54.42}\right) + \cdots + 6\log\left(\frac{6}{5.77}\right)\right] \\ &= 14.01 \end{aligned}$$

with

$$p\text{-value } \approx P(W \geq 14.01) = 0.17, \text{ where} W \sim \chi^2(10)$$

and there is no evidence against the Poisson model.

The observed value of the Pearson Goodness of Fit Statistic is

$$\sum_{j=0}^{11} \frac{(f_j - e_j)^2}{e_j} = 12.98$$

with

$$p\text{-value } \approx P(W \geq 12.98) = 0.22, \quad \text{where} W \sim \chi^2(10)$$

and there is no evidence against the Poisson model.

What to Do if Expected Frequencies Are Not All at Least 5?

Suppose we have the following table of observed frequencies and expected frequencies calculated under the null hypothesis.

Category	Observed Number	Expected Number
1	53	50
2	21	25
3	11	12.5
4	8	6.25
5	4	3.125
6	3	3.125
Total	100	100

The last two categories have expected frequencies less than five, so it may not be appropriate to use the Chi-squared approximation.

Usually we collapse two or more adjacent categories with the smallest expected frequencies.

Category	Observed Number	Expected Number
1	53	50
2	21	25
3	11	12.5
4	8	6.25
$\geqslant 5$	7	6.25
Total	100	100

3.7.2 Two-Way Tables and Testing for Independence of Two Variates

Program/ Hometown	Canadian Hometown	Non-Canadian Hometown	Total
Computer Science	33	14	47
Non-Computer Science	39	46	85
Total	72	60	132

Is there a relationship between a student's program and their hometown?

Relative Risk

Previously, we summarized these data using relative risk as a numerical summary. Proportion of Computer Science students with Canadian hometown:

$$\frac{33}{47} = 0.7021$$

Proportion of non-Computer Science students with Canadian hometown:

$$\frac{39}{85} = 0.4588$$

Relative risk $= \frac{0.7021}{0.4588} = 1.53$ Students who have a Canadian hometown are about one and a half times more likely to be in a Computer Science program.

How could we test the hypothesis that a student's program is independent of their hometown?

Two-Way Tables and Testing for Independence of Two Variates

Suppose n individuals are classified according to two different variates which have two possible values.

	B	$\overline{B}$	Total
A	y_{11}	y_{12}	$r_1 = y_{11} + y_{12}$
$\overline{A}$	y_{21}	y_{22}	$n - r_1$
Total	$c_1 = y_{11} + y_{21}$	$n - c_1$	n

Let Y_{11} = number of $A \cap B$ outcomes, Y_{12} = number of $A \cap \overline{B}$ outcomes Y_{21} = number of $\overline{A} \cap B$ outcomes, and Y_{22} = number of $\overline{A} \cap \overline{B}$ outcomes. Then

$$(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}\,(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

The null hypothesis is that the variates A and B are independent, or

$$H_0 : P(A \cap B) = P(A)P(B)$$

Let $P(A) = \alpha$ and $P(B) = \beta$, then the null hypothesis may be written as

$$H_0 : \theta_{11} = \alpha\beta$$

Likelihood and Maximum Likelihood Estimators

Since $(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim$ Multinomial $(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, the likelihood function is

$$L\left(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}\right) = \frac{n}{y_{11}!y_{12}!y_{21}!y_{22}!}\theta_{11}^{y_{11}} \cdot \theta_{12}^{y_{12}} \cdot \theta_{21}^{y_{21}} \cdot \theta_{22}^{y_{22}}$$

or more simply

$$L\left(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}\right) = \theta_{11}^{y_{11}} \cdot \theta_{12}^{y_{12}} \cdot \theta_{21}^{y_{21}} \cdot \theta_{22}^{y_{22}}$$

The maximum likelihood estimates are

$$\hat{\theta}_{ij} = \frac{y_{ij}}{n}, \quad i = 1, 2 \text{and} j = 1, 2$$

and the maximum likelihood estimators are

$$\tilde{\theta}_{ij} = \frac{Y_{ij}}{n}, \quad i = 1, 2 \text{and} j = 1, 2$$

3.7.3 Parameter Estimation under the Null Hypothesis

If $H_0 : \theta_{11} = \alpha\beta$ is true, then the likelihood function under H_0 is

$$\begin{aligned} L(\alpha, \beta) &= (\alpha\beta)^{y_{11}}[\alpha(1-\beta)]^{y_{12}}[(1-\alpha)\beta]^{y_{21}}[(1-\alpha)(1-\beta)]^{y_{22}} \\ &= \alpha^{y_{11}+y_{12}}(1-\alpha)^{y_{21}+y_{22}} \cdot \beta^{y_{11}+y_{21}}(1-\beta)^{y_{12}+y_{22}}, \quad 0 < \alpha < 1, \quad 0 < \beta < 1 \end{aligned}$$

which is maximized for

$$\hat{\alpha} = \frac{y_{11} + y_{12}}{n} = \frac{r_1}{n} \quad \text{and} \quad \hat{\beta} = \frac{y_{11} + y_{21}}{n} = \frac{c_1}{n}$$

The maximum likelihood estimators for α and β are

$$\tilde{\alpha} = \frac{Y_{11} + Y_{12}}{n} \quad \text{and} \quad \tilde{\beta} = \frac{Y_{11} + Y_{21}}{n}$$

Why do these estimates make sense?

Likelihood Ratio Test Statistic

$$\begin{aligned} &-2\log\left[\frac{L(\tilde{\alpha},\tilde{\beta})}{L(\tilde{\theta}_{11},\tilde{\theta}_{12},\tilde{\theta}_{21},\tilde{\theta}_{22})}\right] \\ &= 2\left[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)\right] \end{aligned}$$

which is of the form

$$2\sum_{j=1}^{k}(\text{observed frequency}) \times \log\left(\frac{\text{observed frequency}}{\text{expected frequency}}\right)$$

Observed Likelihood Ratio Test Statistic

The observed value of the likelihood ratio test statistic is

$$\lambda = 2\left[y_{11}\log\left(\frac{y_{11}}{e_{11}}\right) + y_{12}\log\left(\frac{y_{12}}{e_{12}}\right) + y_{21}\log\left(\frac{y_{21}}{e_{21}}\right) + y_{22}\log\left(\frac{y_{22}}{e_{22}}\right)\right]$$

Note that

$$e_{11} = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \frac{r_1 c_1}{n}$$

and the other expected frequencies can be obtained by subtraction.

Two-Way Table: Observed and [Expected]

	B	$\tilde{B}$	Total
A	y_{11} $\left[e_{11} = \frac{r_1 c_1}{n}\right]$	y_{12} $[e_{12} = r_1 - e_{11}]$	$r_1 = y_{11} + y_{12}$
$\tilde{A}$	y_{21} $[e_{21} = c_1 - e_{11}]$	y_{22} $[e_{22} = r_2 - e_{21}]$	$n - r_1$
Total	$c_1 = y_{11} + y_{21}$	$n - c_1$	n

Degrees of Freedom for the Chi-Squared Approximation

What are the degrees of freedom for the Chi-squared approximation? How many parameters were there in the original model?

$$(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}\,(n; \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

How many parameters under $H_0 : \theta_{11} = \alpha\beta$? Why do the degrees of freedom make sense?

Approximate p-Value

$$\begin{aligned} p\text{-value} &\approx P(W \geq \lambda), \quad \text{where} W \sim \chi^2(1) \\ &= 2[1 - P(Z \leq \sqrt{\lambda})], \quad \text{where} Z \sim G(0,1) \end{aligned}$$

EXAMPLE 3.7.1 (Program/Hometown Example: Two-Way Table)**.**

Program / Hometown	Canadian Hometown	Non-Canadian Hometown	Total
Computer Science	$e_{11} = \frac{47 \times 72}{132} = 25.64$	$e_{12} = 47 - 25.64 = 21.36$	47
Non-Computer Science	$e_{21} = 72 - 25.64 = 46.36$	$e_{22} = 60 - 21.36 = 38.64$	85
Total	72	60	132

Program/Hometown Example: p-Value

$$\lambda = 2\left[33\log\left(\frac{33}{25.64}\right) + 39\log\left(\frac{39}{46.36}\right) + 14\log\left(\frac{14}{21.36}\right) + 46\log\left(\frac{46}{38.64}\right)\right] = 7.38$$

$$p\text{-value} = 2[1 - P(Z \leq \sqrt{7.38})] = 2[1 - P(Z \leq 2.72)] \approx 0.0066$$

There is strong evidence based on these observed data against the hypothesis that a student's program is independent of their hometown.