STAT 231 Chapter 1, 2, 3

Statistics

David Duan, 2019 Fall

Contents

1 Introduction to Statistical Science

1.1 Terminology

A **unit** is an individual object about which we can take *measurements*.

Populations and **processes** are *collections* of units. The key difference is that processes usually *occur over time* whereas populations are *static*.

Variates are *characteristics* of units which are usually represented by lowercase letters. There are five categories: **continuous**, **discrete**, **categorical**, **ordinal**, and **complex**.

An **attribute** of a population or process is a *function* of a variate which is defined for all units in the population or process.

An **empirical study** is one which we learn by observation or experimentation. There are three types of empirical study: *sample surveys*, *observational studies*, and *experimental studies*.

1.2 Data Summary

We are interested in measuring *location*, *variability/dispersion*, and *shape* of a data set.

The pth **quantile** $(0 < p < 1)$ is the value $q(p)$ such that $100p$ percent of data fall at or below this value. Let $i = (n+1)p$. If $i \in \mathbb{Z}$, return $A[i]$. Otherwise, return $\frac{1}{2}(A[\lfloor i \rfloor] + A[\lceil i \rceil])$.

Sample **skewness** measures the degree of *asymmetry* from the symmetrical bell curve.

- 0 = symmetric | positive = skewed to the right | negative = skewed to the left.
- Values between -1 and $+1$ are considered to be "close to 0" for Gaussian data.

Sample **kurtosis** measures whether data are concentrated in the center or tail.

- 3 = bell-shaped | > 3 = large tail, less at center | < 3 = short tail, more at center.
- Values between 2 and 4 are considered to be "close to 3" for Gaussian data.

Five number summary: minimum, lower quartile, median, upper quartile, maximum.

In order to assume a *Gaussian* model is reasonable for a given data set,

1. The sample mean and median should be approximately equal.
2. The sample skewness should be close to 0, i.e., between -1 and $+1$.
3. The sample kurtosis should be close to 3, i.e., between 2 and 4.
4. Approximately 95% of the observation should lie in the interval $[\overline{y} - 2s, \overline{y} + 2s]$.

In statistics, we never "prove" an assumption is true; instead we see if we can find evidence *against* an assumption. We should *never* use definitive statements such as "the assumption is true/false". The correct statement is "the data are (not) consistent with the assumption".

An **empirical cumulative distribution function** for a r.v. Y is a function giving $P(Y \leq y)$.

A **run chart** gives a graphical summary of data which are varying over time.

The **sample correlation** $-1 \leq r \leq 1$ measures the *linear* relationship between two variables.

The **explanatory variate** x is in the study to partially explain or determine the distribution of Y, the random variable representing the **response variate**.

The **relative risk** is the ratio of the *probability* of an outcome in an *exposed* group to the probability of an outcome in an *unexposed* group, i.e., $P(A \mid B)/P(A \mid B^c)$. If we hypothetically find that 17% of smokers develop lung cancer and 1% of non-smokers develop lung cancer, then we can calculate the relative risk of lung cancer in smokers versus non-smokers as $17\%/1\% = 17$. Thus, it is 17 times more likely for smokers to develop lung cancer than non-smokers.

1.3 Statistics

Two broad aspects of the analysis and interpretation of data:

- **Descriptive statistics** is the portrayal of the data in numerical and graphical ways to show features of interest.
- When the data obtained in the study of a population or process are use to draw general conclusion about the population or process itself we call this process **statistical inference**.

Three problems we study in this course:

- In an **estimation problem** (Chapter 4) we are interested in estimating one or more attributes of a process or population.
- In a **hypothesis testing problem** (Chapter 5) we use data to assess the truth of a question or hypothesis.
- In a **prediction problem** (Chapter 6) we use the data to predict a future value of a variate for a unit to be selected from the population or process.

1.4 Problems

Think: How do mean, median, standard deviation, variance, IQR, range, skewness, and kurtosis change if we transform data by $f(y) = a + by$? How do they change if we have an extreme value added to the data set?

2 Statistical Models and Maximum Likelihood

A **statistical model** is a mathematical model that incorporates probability. A **random variable** is used to represent a characteristic or variate of a randomly selected unit from the population or process. Drawing conclusions from data involves some degree of uncertainty; statistical models can be used to quantify this uncertainty.

2.1 Probability Models

How we could choose a probability model:

- Background knowledge or assumption
- Past experience with similar data
- Mathematical convenience
- A real data set against which the model can be accessed

The steps to choose a probability model:

1. Collect and examine the data.
2. Propose a model.
3. Fit the model.
4. Check the model.
5. If required, propose a revised model and return to 2.
6. Draw conclusions using the chosen model and the observed data.

2.1.1 Probability Distributions

$Y \sim \textit{Binomial}(n, \theta)$: n Bernoulli trials, each with θ probability of success. Y is the number of successes. Example: coin toss.

$Y \sim \textit{Poisson}(\theta)$: θ occurrences on average per unit time. Y is the number of event occurrences. Example: number of houses sold per day; number of cars crossing an intersection.

$Y \sim \textit{Exponential}(\theta)$: θ occurrences on average per unit time. Y is the time between event occurrences. Example: time between two houses being sold; time between two website accesses.

$Y \sim \textit{Gaussian}(\theta)$, $\theta = (\mu, \sigma)$: μ average, σ standard deviation. Y is the value of the quantity. Example: height, weight, or sum of statistical models (central limit theorem).

2.1.2 Probability Density Function

The p.d.f of a random variable is $f(y; \theta) = P(Y = y)$ (for discrete models) or $f(y; \theta) = \frac{d}{dy} P(Y \leq y)$ (for continuous models) where $y \in \mathsf{range}(Y)$. We emphasize the dependence of the model on the parameter θ.

For example, suppose $Y \sim \mathit{Binomial}(n, \theta)$. Then the probability function of Y is

$$f(y; \theta) = P(Y = y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \quad y \in \{0, 1 \ldots, n\}; 0 < \theta < 1.$$

For each value of θ, we get a different p.d.f. In other words, we have a *family* of models.

2.2 Maximum Likelihood Estimation

Suppose that the random variable $Y \sim G(\mu, \sigma)$ adequately models how long it takes me to walk to class from home and I'm interested in estimating the parameter μ here. I could randomly select n days, measure the length of each trip $y_1, y_2, \ldots, y_n$, and estimate μ using $\overline{y}$, the sample mean. We might write $\hat{\mu} = \overline{y}$.

Note that μ is not necessarily equal to the sample mean (in fact, it almost certainly isn't!). The estimate of μ is a function of the observed data $y_1, y_2, \ldots, y_n$. Thus, different draws of the sample will result in different values of the sample mean and therefore different estimates of μ.

2.2.1 Point Estimate

A **point estimate** of parameter θ is an estimate $\hat{\theta} = \hat{\theta}(\mathbf{y})$ that is the result of a function (called a *point estimator*, see Chapter 4) of observed data $\mathbf{y} = (y_1, \ldots, y_n)$.

For example, we could estimate the mean μ of a Gaussian distribution $G(\mu, \sigma)$ by

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

(the sample mean) or estimate the probability of success θ of a binomial distribution $\mathit{Bin}(n, \theta)$ by $\hat{\theta} = y/n$, the ratio of success in sample.

Besides guessing with intuition, is there a formal way to do this?

2.2.2 Method of Maximum Likelihood

The method of **maximum likelihood** is the most widely used method of estimation.

Let the random variable Y represent potential data that will be used to estimate θ and let y represent the actual observed data. To start with, we assume Y is a discrete random variable. The **likelihood function** for θ is defined as

$$L(\theta) = L(\theta; \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}; \theta), \quad \theta \in \Omega$$

where Ω is the parameter space - the set of possible values of θ. Intuitively, $L(\theta)$ is the *probability we observe the data* $\mathbf{y}$ as a function of θ. For convenience, we usually write just $L(\theta)$, but don't forget that it is a function of both θ and $\mathbf{y}$!

As a remark, the shape of $L(\theta)$ and its maximum value are not affected if we multiple $L(\theta)$ by a positive constant. (See example below.)

Example. Let Y be the number of success in n Bernoulli trials with $P(1) = \theta$. Then $Y \sim \textit{Binomial}(n, \theta)$. Suppose a Binomial experiment is conducted and y successes are observed. The likelihood function for θ based on the observed data is

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

and the maximum likelihood estimate of θ is $\hat{\theta} = y/n$ (check this by taking its first derivative and set to zero). If $y = 10$ and $n = 25$, our most reasonable guess is that $\hat{\theta} = y/n = 10/25 = 0.4$. Note that $\binom{n}{y}$ does not affect what value of θ maximizes $L(\theta)$. Therefore, the maximum likelihood estimate of θ also maximizes $k\theta^y(1 - \theta)^{n-y}$ for any $k \in \mathbb{R}^+$. ■

2.2.3 Relative Likelihood

The **relative likelihood function** is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \quad \theta \in \Omega$$

Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$, and that $R(\hat{\theta}) = 1$.

The maximum likelihood estimate of θ, denoted $\hat{\theta}$, is of particular interest. We know that the likelihood is maximized when $\theta = \hat{\theta}$, therefore, for any value of θ that is not the maximum likelihood estimate, the relative likelihood will be less than 1.

Example. Using the example above, where $y = 10$ and $n = 25$. We see $P(Y = 10; \theta = 0.4) = 0.161$ and $P(Y = 10; \theta = 0.25) = 0.042$. Thus, the data are approximately $0.161/0.042 \approx 4$ times more likely to appear if $\theta = 0.4$ than if $\theta = 0.25$. Note that we care about the *relative* likelihood, not the *absolute* likelihood for each one. ■

Example. For our example above, the relative likelihood function is

$$R(\theta) = \frac{\theta^{10}(1 - \theta)^{15}}{0.4^{10} 0.6^{15}}.$$

It is easy to check that $\theta = 0.4$ maximizes $R(\theta)$. ■

2.2.4 Log Likelihood

The **log likelihood function** is defined as

$$\ell(\theta) = \ln L(\theta), \quad \theta \in \Omega.$$

The log likelihood function is maximized for the same value of θ as the regular likelihood function, with the additional benefit that algebra often becomes easier.

2.2.5 Summary of Three Functions

Let $\hat{\theta}$ denote the maximum likelihood estimate of a population parameter θ for a particular dataset, and let $L(\theta)$, $R(\theta)$, and $\ell(\theta)$ denote the likelihood, relative likelihood, and log likelihood functions, respectively. Then $\hat{\theta}$, our "best guess" of θ based on the observed data, maximizes all three functions.

In one short sentence, ***the maximum likelihood estimate $\hat{\theta}$ for the parameter θ maximizes the likelihood that you observed the things you observed.***

2.2.6 Likelihood Function for Independent Experiments

If we observe data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ that are independent and identically distributed each with probability function $P(Y_i = y_i; \theta)$, then

$$L(\theta) = \prod_{i=1}^{n} P(Y_i = y_i; \theta) \quad \theta \in \Omega$$

Example. For Poisson data $y_1, \ldots, y_n$, we can derive the likelihood as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(Y_i = y_i; \theta) \\ &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \left[\prod_{i=1}^{n} \frac{1}{y_i!}\right] \left[\prod_{i=1}^{n} \theta^{y_i}\right] \left[\prod_{i=1}^{n} e^{-\theta}\right] \\ &= \left[\prod_{i=1}^{n} \frac{1}{y_i!}\right] \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} && \theta > 0 \\ &\sim \theta^{n\overline{y}} e^{-n\theta} && \theta > 0 \end{aligned}$$

Two remarks:

1. Be comfortable with algebraic manipulations like $\left[\prod_{i=1}^{n} \theta^{y_i}\right] = \theta^{\sum_{i=1}^{n} y_i}$.
2. The constants $\left[\prod_{i=1}^{n} \frac{1}{y_i!}\right]$ does not affect θ, thus we can ignore it.

To find the value of θ that maximizes $L(\theta)$, we set the first derivative to zero. Note that $\ell(\theta)$ is easier to work with than $L(\theta)$:

$$\begin{aligned} L(\theta) = \theta^{n\overline{y}} e^{-n\theta} &\implies \ell(\theta) = n\overline{y}\log(\theta) - n\theta \\ &\implies \frac{d}{d\theta}\ell(\theta) = \frac{n\overline{y}}{\theta} - n \\ \frac{d}{d\theta}\ell(\hat{\theta}) = 0 &\implies \hat{\theta} = \overline{y}. \end{aligned}$$

2.2.7 Likelihood Function for a Random Sample

Suppose we have data of the form $\mathbf{y} = (y_1, \ldots, y_n)$ where $(y_1, \ldots, y_n)$ is assumed to be a realization of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

Suppose also that the Y_i's are assumed to be independent and identically distributed random variables with probability function $P(Y = y; \theta) = f(y; \theta)$ for $\theta \in \Omega$. Then $Y_1, Y_2, \ldots, Y_n$ is called a **random sample**.

The likelihood function for θ based on the observed sample $y_1, y_2, \ldots, y_n$ is

$$\begin{aligned} L(\theta) &= P(\text{Observing the data } y_1, \ldots, y_n \text{ given } \theta) \\ &= P(Y_1 = y_1, \ldots, Y_n = y_n; \theta) \\ &= P(Y_1 = y_1; \theta) \cdots P(Y_n = y_n; \theta) && Y_i\text{'s are independent.} \\ &= \prod_{i=1}^{n} P(Y_i = y_i; \theta) \quad \theta \in \Omega \end{aligned}$$

2.3 Likelihood for Continuous Random Variables

A continuous distribution has a probability density function $f(y, \theta)$ defined as

$$P(a \leq Y \leq b) = \int_a^b f(y; \theta)\, dy.$$

For continuous distributions, $L(\theta; \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}; \theta)$ for $\theta \in \Omega$ is an unsuitable definition, since $P(\mathbf{Y} = \mathbf{y}; \theta) = 0$.

2.3.1 Likelihood Function

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a random sample from a continuous distribution with PDF $f(y; \theta)$ for $\theta \in \Omega$, and $\mathbf{y} = (y_1, \ldots, y_n)$ represents a realization of $\mathbf{Y}$. We define the **likelihood function** for θ based on the observed data $\mathbf{y}$ as

$$L(\theta) = L(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta) \quad \theta \in \Omega.$$

2.3.2 Exponential Likelihood Function

- p.d.f. for *exponential*(θ): $f(y; \theta) = \frac{1}{\theta} e^{-y/\theta}$.
- Likelihood function: $L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} = \theta^{-n} e^{-n\bar{y}/\theta}$.
- Log likelihood function: $\ell(\theta) = \log[\theta^{-n} e^{-n\bar{y}/\theta}] = -n\log(\theta) - \frac{n\bar{y}}{\theta} = -n[\log(\theta) + \frac{\bar{y}}{\theta}]$.
- Maximum (log) likelihood estimate: $\frac{d}{d\theta}\ell(\theta) = -n(\frac{1}{\theta} - \frac{\bar{y}}{\theta^2}) = \frac{n}{\theta^2}(\bar{y} - \theta)$.
- We can show that $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate of θ.

2.3.3 Gaussian Likelihood Function

- p.d.f. for $G(\mu,\sigma)$: $f(y;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2\sigma^2}(y-\mu)^2]$. Note this is 2-dimensional.
- Likelihood function: $L(\theta)=\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2\sigma^2}(y-\mu)^2]\sim\sigma^{-n}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2]$.
- Log likelihood function: $\ell(\theta)=-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2$, $\mu\in\mathbb{R}$, $\sigma>0$.
- Maximum (log) likelihood estimate (we compute two partial derivatives and set to zero):

$$\frac{\partial\ell}{\partial\mu}=\frac{n}{\sigma^2}(\overline{y}-\mu),\ \frac{\partial\ell}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n(y_i-\mu)^2$$

- We can show that $\hat{\mu}=\overline{y}$ and $\hat{\sigma}=[\frac{1}{n}\sum_{i=1}^n(y_i-\overline{y})^2]^{1/2}$ is the maximum likelihood estimate.

2.4 Invariance Property of Maximum Likelihood Estimates

2.4.1 The Invariance Property

If $\hat{\theta}$ is the maximum likelihood estimate of θ then $g(\hat{\theta})$ is the maximum likelihood estimate of $g(\theta)$
.

2.4.2 Example

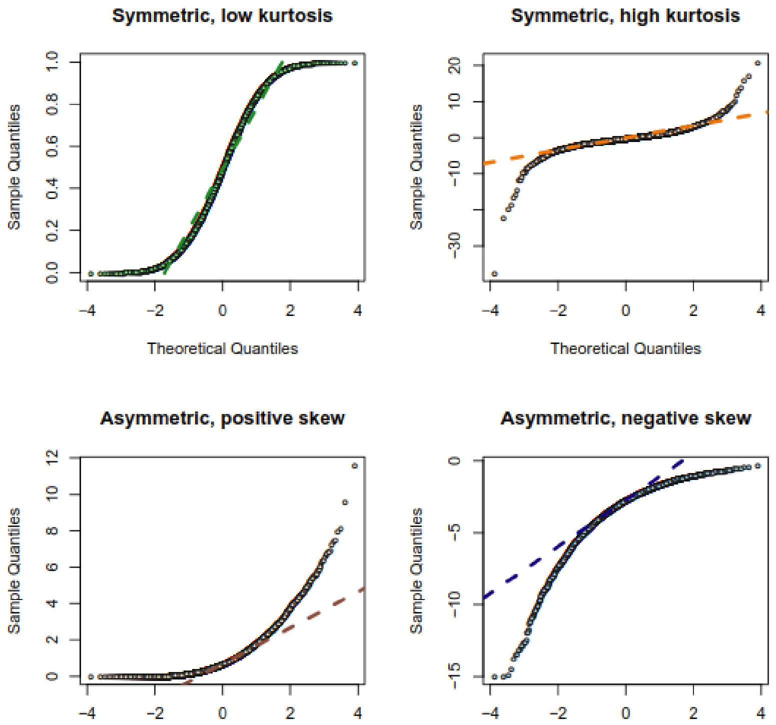
Suppose consumers are given a choice between two products, A and B. The probability of a random chosen consumer will pick A is θ. We estimate θ by surveying 100 consumers and see which product they prefer. Let $Y\sim Bin(100,\theta)$. We are interested in $\hat{\theta}$ and $Var(Y)$.

For binomial data with y successes observed from n Bernoulli trials, the maximum likelihood estimate of $P(success)=\theta$ is $\hat{\theta}=y/n$. We also know that $Var(Y)=n\theta(1-\theta)$. Thus, we can apply the invariance theorem to find the maximum likelihood estimate of $Var(Y)$:

$$n\hat{\theta}(1-\hat{\theta})=n\left(\frac{y}{n}\right)\left(1-\frac{y}{n}\right)=y\left(1-\frac{y}{n}\right).$$

2.5 Checking Model Fit

A (relatively) straight line indicates a Gaussian model is a good fit. Otherwise:



Symmetric, low kurtosis

Symmetric, high kurtosis

Asymmetric, positive skew

Asymmetric, negative skew

3 PPDAC

PPDAC: Problem, Plan, Data, Analysis, Conclusion.

3.1 Terminology

The *target population* or *target process* is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

A *variate* is a characteristic of every unit.

An *attribute* is a function of the variates over a population.

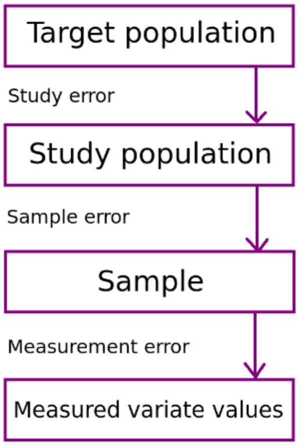
The *study population* or *study process* is the collection of units available to be included in the study. Often the study population is a strict subset of the target population.

The *sampling protocol* is the procedure used to select a sample of units from the study population. The number of units sampled is called the *sample size*. Sample size is usually a compromise between cost, availability and desired precision as calculated using a model.

The types of problems that an empirical study are designed to study can be categorized as one of three types:

- *Descriptive*: to determine a particular attribute of the population.
- *Causative*: to determine the (non-)existence of a causal relationship between two variates.
- *Predictive*: to predict the response of a variate for a given unit.

3.2 Error



3.2.1 Study Error

If the attributes in the study population differ from the attributes in the target population, then the difference is called *study error*.

Since we do not know the values of the target population attributes or the study population attributes, the study error cannot be quantified. Instead, we rely on expertise from other sources to determine whether conclusions derived from the study population may apply to the target population.

3.2.2 Sample Error

If the attributes in the sample differ from the attributes in the study population, then the difference is called *sample error*.

Different sampling protocols are likely to produce different sample errors. Some protocols tend to have smaller errors than others. Since the values of the study population attributes are unknown, the sample error is unknown. We can use statistical models to quantify how large this error might be.

3.2.3 Measurement Error

If the measured value and the true value of a variate are not identical, then the difference is called *measurement error*.

4 Appendix: Problem Solving Strategies

4.1 Quiz 1 [Chapter 1]

Just get the definitions correct and find the appropriate formula to use. A couple useful results to remember are:

1. The five number summary contains the minimum, lower quartile, median, upper quartile, and maximum.
2. The correlation measures the linear relationship between two variates and is between -1 and 1.
3. The top and bottom bars in a bar plot are $q1 - 1.5IQR$ and $q3 + 1.5IQR$; anything not in this range is considered an outlier.
4. The distribution is positively skewed if it has a long right tail and negatively skewed if it has a long left tail.
5. A data set is not unimodal when the maximum value is achieved by more than one category.
6. To determine something from empirical CDF, first determine the corresponding PDF, i.e., determine the number and/or proportion for each category.

4.2 Test 1 [Section 2.5]

4.2.1 Probability Distributions

Assumptions for a Poisson model:

- Independence: Events must be independent (e.g. the number of goals scored by a team should not make the number of goals scored by another team more or less likely.)
- Homogeneity: The mean number of goals scored is assumed to be the same for all teams.
- Time period (or space) must be fixed.

4.2.2 MLE

Given a PDF, derive the MLE for θ:

1. Determine the likelihood function: $L(\theta; y) = P(Y = y; \theta)$ if discrete and $L(\theta) = \prod f(y; \theta)$ if continuous.
    a. Most important transformation rule: $\prod_{i=1}^{n} \theta^{y_i} = \theta^{\sum_{i=1}^{n} y_i}$.
    b. It might be helpful if you expand $\prod$ and explicitly write out every term.
    c. Don't forget to define the parameter space, i.e., the set of possible value for variables, e.g., $\theta > 0$.
2. Determine the log likelihood function: $l(\theta) = \ln L(\theta)$.

   a. $\log x^n = n \log x$.
   b. $\log xy = \log x + \log y$, $\log \sum y_i = \prod \log y_i$.
   c. Don't forget to define the parameter space, i.e., the set of possible value for variables, e.g., $\theta > 0$.
3. Differentiate it wrt θ and set to 0: $\frac{d}{d\theta} l(\theta) = 0$.
   a. Before differentiation, you can ignore the irrelevant constants.
   b. $\frac{d}{d\theta} \log \theta = \theta^{-1}$.
   c. $\frac{d}{d\theta} \theta^{-1} = -\theta^{-2}$.
4. Solve for θ.

Determine $R(\theta)$: write out $L(\theta)/L(\hat{\theta})$ and plug in $\hat{\theta}$ (expression or number).

Determine XX using the invariance property of MLE: find the appropriate formula containing θ, then plug in $\hat{\theta}$ (expression or number).

4.3 Quiz 2 [Section 4.3]

4.3.1 PPDAC

1. What type of study is this and why?

Possible solutions include *experimental study*, *observational study*, and *survey*, depends on whether researchers intervened or not. The last one is very unlikely.

- This study is an experimental study since the researchers are in control of which schools received the regular curriculum and which schools are using the JUMP program.
- This is an observational study because the researchers did not attempt to change or control any of the variates for the sampled units.

2. Define the problem of this study.

Use full sentence. Don't miss details.

- The problem is to compare the performance in math students at Ontario schools using the current provincial curriculum as compared to the performance in math of students at Ontario schools using the JUMP math problem.

3. What type of problem is this and why?

Possible solutions include *causative problem*, *descriptive problem*, and *predictive problem*. The last one is very unlikely.

- This is a causative problem since the researchers are interested in whether the JUMP program causes better student performance in math.

- This is a descriptive problem since the aim of the study is to determine the attributes for a population of eligible voters.

4. Define a suitable target population/process for this study.

Think about the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

- The target population is all elementary students in Ontario publish schools at the time of the study (and into the future).
- Since this study recruited new participants to the study every year since 1999, it would be best to define a target process (recall population is static and process is dynamic). A suitable target process for this study is the set of adults living in a Mediterranean country when the study began and into the future.

5. Define a suitable study population/process for this study.

Think about the collection of units available to be included in the study. Often the study population is a strict subset of the target population.

- The study population is all Ontario elementary students in Grades 2 and 5 in publish schools at the time of the study.
- A suitable study process for this is the the of Spanish university graduates when the study began and into the future.

6. Define the variate(s) of interest in this problem and specify the type of each.

Think about what can be measured in the study. Possible answers for type include *continuous*, *categorical*, *discrete*, *ordinal*, and *complex*. The last one is very unlikely.

7. What is the sampling protocol?

Think about how researchers select samples.

- The sampling protocol was to select the schools in one school board in Ontario. The researchers did not indicate how this school board was chosen.

8. What is a possible source of study error?

Think about how the study population/process may misrepresent the target population/process.

- A possible source of study error is that the polling firms only called eligible voters in urban areas. Urban eligible voters may have different views that rural eligible voters. This is a difference between the target and study population. Also, eligible voters with phones may have different views than those without.

9. What is a possible source of sample error?

Think about how the sample selected may misrepresent the study population/process.

- A possible source of sample error is that many of the people called refused to participate in the survey. These people who refused to participate may have different voting preferences as compared to people who participate. For example, they may also be less likely to vote.

10. What is a possible source of measurement error?

Think about where can go wrong in the process of taking measurements. We usually care less about this type of errors in this course.

- People who were asked about how many cups of coffee they drink a day may give inaccurate answers because they might not track this number very carefully.

11. Why was it important to keep a control group (or using other similar strategies)?

The answer is usually to ensure that the difference in the outcome is only due to the factor of interest, and not due to other potential confounders.

- Randomization ensures that the difference in the learning outcome is only due to different teaching programs, and not due to other potential confounders, e.g., class size, parent's education level and socio-economic status, etc.

12. Estimate the attribute of interest for the study population based on the given data.

This is usually some obvious calculation. For example, use sample mean as the estimate of population mean.

4.3.2 Gaussian Approximation of Binomial Distribution

Consider the random variable Y and the estimator $\tilde{\theta} = Y/n$.

- *Determine* $P(-0.03 \leq \tilde{\theta} - \theta \leq 0.03)$, *if* $n = 1000$ *and* $\theta = 0.5$ *using the Normal approximation to the Binomial.*

$$\begin{aligned} P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) &= P\left(-0.03 \leq \frac{Y}{1000} - 0.5 \leq 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \leq \frac{\frac{Y}{1000} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \leq \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}}\right) \\ &\approx P\left(-1.90 \leq Z \leq 1.90\right) \quad \text{where } Z \sim N(0,1) \\ &= 2P\left(Z \leq 1.90\right) - 1 = 2\left(0.97128\right) - 1 \\ &= 0.94256 \end{aligned}$$

- *If* $\theta = 0.5$, *determine how large* n *should be to ensure that* $P(-0.03 \leq \tilde{\theta} - \theta \leq 0.03) = P(|\tilde{\theta} - \theta| \leq 0.03) \geq 0.95$.

$$\begin{aligned}
P\left(-0.03 \le \tilde{\theta} - \theta \le 0.03\right) &= P\left(-0.03 \le \frac{Y}{n} - 0.5 \le 0.03\right) \\
&= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}} \le \frac{\frac{Y}{n} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{n}}} \le \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}}\right) \\
&\approx P\left(-0.06\sqrt{n} \le Z \le 0.06\sqrt{n}\right)
\end{aligned}$$

where $Z \sim N(0,1)$. Since $P(-1.96 \le Z \le 1.96) = 0.95$, we need $0.06\sqrt{n} \ge 1.96$ or $n \ge (1.96/0.06)^2 = 1067.1$. Therefore n should be at least 1068.

- *If θ is unknown, determine how large n should be to ensure that* $P(-0.03 \le \tilde{\theta} - \theta \le 0.03) = P(|\tilde{\theta} - \theta| \le 0.03) \ge 0.95$ *for all* $\theta \in [0,1]$.

$$\begin{aligned}
P\left(-0.03 \le \tilde{\theta} - \theta \le 0.03\right) &= P\left(-0.03 \le \frac{Y}{n} - \theta \le 0.03\right) \\
&= P\left(\frac{-0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le \frac{\frac{Y}{n} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le \frac{0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right) \\
&\approx P\left(-\frac{0.03\sqrt{n}}{\sqrt{\theta(1-\theta)}} \le Z \le \frac{0.03\sqrt{n}}{\sqrt{\theta(1-\theta)}}\right)
\end{aligned}$$

where $Z \sim N(0,1)$. Since $P(-1.96 \le Z \le 1.96) = 0.95$, we need

$$\frac{0.03\sqrt{n}}{\sqrt{\theta(1-\theta)}} \ge 1.96$$

or

$$n \ge \left(\frac{1.96}{0.03}\right)^2 \theta(1-\theta)$$

Since θ is unknown we take $\theta = 0.5$ so the inequality is true for all $0 < \theta < 1$. Thus

$$n \ge \left(\frac{1.96}{0.03}\right)^2 (0.5)^2 = 1067.1$$

and n should be at least 1068.

4.3.3 Z-Table

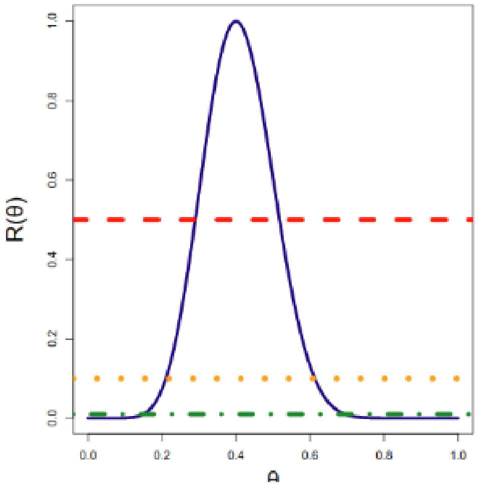
Let $Y \sim G(\mu, \sigma)$. Suppose $\sigma = 5$. Based on a random sample of 64 observations, what is $P(|\bar{Y} - \mu| \le 1)$?

$$\begin{aligned}
P(|\bar{Y} - \mu| \le 1) &= P(-1 \le \bar{Y} - \mu \le 1) \\
&= P\left(\frac{-1}{\sigma/\sqrt{n}} \le \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \le \frac{1}{\sigma/\sqrt{n}}\right) \\
&= P(-1.6 \le Z \le 1.6) \\
&= 2P(Z \le 1.6) - 1 \\
&= 2 * 0.94520 - 1 \\
&= 0.8904.
\end{aligned}$$

4.3.4 Determine Relative Likelihood Interval

Draw a horizontal line at $y = 100p$.

For example, the following are 1%, 10%, and 50% likelihood intervals for the coin example given $n = 25$ and $y = 10$ heads:



R(θ)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
θ