STAT 231 (Winter 2012 - 1121)

Honours Statistics

Prof. R. Metzger
University of Waterloo

LaTeXer: W. Kong
http://wwkong.github.io

Last Revision: April 30, 2014

Table of Contents

These notes are currently a work in progress, and as such may be incomplete or contain errors.

List of Figures

ACKNOWLEDGMENTS:

Special thanks to *Michael Baker* and his LaTeX formatted notes. They were the inspiration for the structure of these notes.

Abstract

The purpose of these notes is to provide a guide to the second year honours statistics course. The contents of this course are designed to satisfy the following objectives:

- To provide students with a basic understanding of probability, the role of variation in empirical problem solving and statistical concepts (to be able to critically evaluate, understand and interpret statistical studies reported in newspapers, internet and scientific articles).
- To provide students with the statistical concepts and techniques necessary to carry out an empirical study to answer relevant questions in any given area of interest.

The recommended prerequisites are Calculus I, II (one of Math 137,147 and one of Math 138, 148), and Probability (Stat 230). Readers should have a basic understanding of single-variable differential and integral calculus as well as a good understanding of probability theory.

1 PPDAC

PPDAC is a process or recipe used to solve statistical problems. It stands for:

Problem / **P**lan / **D**ata / **A**nalysis / **C**onclusion

1.1 Problem

The problem step's job is to clearly define the

1. Goal or Aspect of the study
2. Target Population and Units
3. Unit's Variates
4. Attributes and Parameters

Definition 1.1. The **target population** is the set of animals, people or things about which you wish to draw conclusions. A **unit** is a singleton of the target population.

Definition 1.2. The **sample population** is a specified subset of the target population. A **sample** is a singleton of the sample population and a unit of the study population.

Example 1.1. If I were interested in the average age of all students taking STAT 231, then:

Unit = student of STAT 231
Target Population (T.P.)=students of STAT 231

Definition 1.3. A **variate** is a characteristic of a single unit in a target population and is usually one of the following:

1. **Response variates** - interest in the study
2. **Explanatory variate** - why responses vary from unit to unit
   (a) Known - variates that are know to cause the responses
      i. Focal - known variates that divide the target population into subsets
   (b) Unknown - variates that cannot be explained in the that cause responses

Using Ex. 1.1. as a guide, we can think of the response variate as the age of of a student and the explanatory variates as factors such as the age of the student, when and where they were born, their familial situations, educational background and intelligence. For an example of a focal variate, we could think of something along the lines of domestic vs. international students or male vs. female.

Definition 1.4. An **attribute** is a characteristic of a population which is usually denoted by a function of the response variate. It can have two other names, depending on the population studied:

- **Parameter** is used when studying populations
- **Statistic** is used when studying samples
- Attribute can be used interchangeably with the above

Definition 1.5. The **aspect** is the goal of the study and is generally one of the following

1. Descriptive - describing or determining the value of an attribute
2. Comparative - comparing the attribute of two (or more) groups
3. Causative - trying to determine whether a particular explanatory variate causes a response to change
4. Predictive - to predict the value of a response variate using your explanatory variate

1.2 Plan

The job of the plan step is to accomplish the following:

1. Define the **Study Protocol** - this is the population that is actually studied and is NOT always a subset of the T.P.
2. Define the **Sampling Protocol** - the sampling protocol is used to draw a sample from the study population
   (a) Some types of the sampling protocol include
      i. Random sampling (ad verbatim)
      ii. Judgment sampling (e.g. gender ratios)
      iii. Volunteer sampling (ad verbatim)
      iv. Representative sampling
         A. a sample that matches the sample population (S.P.) in all important characteristics (i.e. the proportion of a certain important characteristic in the S.P. is the same in the sample)
   (b) Generally, statisticians prefer Random sampling
3. Define the Sample - ad verbatim
4. Define the **measurement system** - this defines the tools/methods used

A visualization of the relationship between the populations is below. In the diagram, T.P. is the target population and S.P. is the sample population.

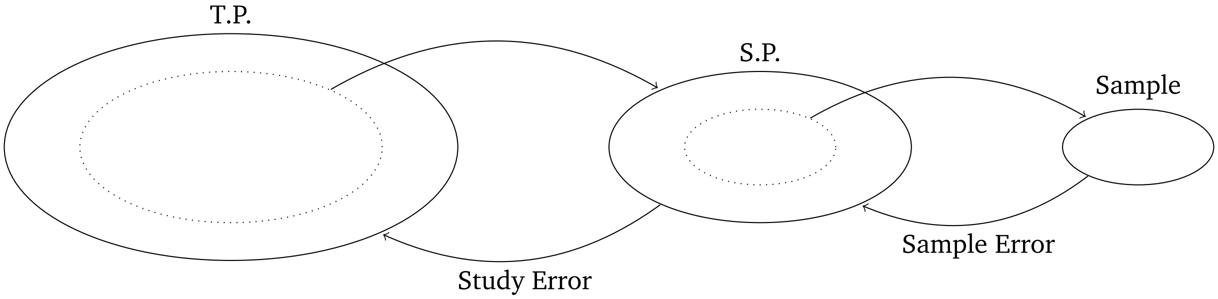

Figure 1.1: Population Relationships

Example 1.2. Suppose that we wanted to compare the common sense between maths and arts students at the University of Waterloo. An experiment that we could do is take 50 arts students and 33 maths students from the University of Waterloo taking an introductory statistics course this term and have them write a statistics test (non-mathematical) and average the results for each group. In this situation we have:

- T.P. = All arts and maths students
- S.P. = Arts and maths students from the University of Waterloo taking an introductory statistics course this term
- Sample = 50 arts students and 33 maths students from the University of Waterloo taking an introductory statistics course this term
- Aspect = Comparative study
- Attribute(s): Average grade of arts and maths students (for T.P., S.P., and sample)
  – Parameter for T.P. and S.P.
  – Statistic for Sample

There are, however some issues:

1. Sample units differ from S.P.
2. S.P. units differ from T.P. units
3. We want to measure common sense, but the test is measuring statistical knowledge
4. Is it fair to put arts students in a maths course?

Example 1.3. Suppose that we want to investigate the effect of cigarettes on the incidence of lung cancer in humans. We can do this by purchasing online mice, randomly selecting 43 mice and letting them smoke 20 cigarettes a day. We then conduct an autopsy at the time of death to check if they have lung cancer. In this situation, we have:

- T.P. = All people who smoke
- S.P. = Mice bought online
- Sample = 43 selected online mice
- Aspect = Not Comparative
- Attribute:
  - T.P. - proportion of smoking people with lung cancer
  - S.P. - proportion of mice bought online with lung cancer
  - Sample - proportion of sample with lung cancer

Like the previous example, there are some issues:

1. Mice and humans may react differently to cigarettes
2. We do not have a baseline (i.e. what does X% mean?)
3. Is have the mice smoke 20 cigarettes a day realistic?

Definition 1.6. Let $a(x)$ be defined as an attribute as a function of some population or sample x. We define the **study error** as

$$a(T.P.) - a(S.P.).$$

Unfortunately, there is no way to directly calculate this error and so its value must be argued. In our previous examples, one could say that the study error in Ex. 1.3. is higher than that of Ex. 1.2. due to the S.P. in 1.3. being drastically different than the T.P..

Definition 1.7. Similar to above, we define the **sample error** as

$$a(S.P.) - a(sample).$$

Although we may be able to calculate $a(sample)$, $a(S.P.)$ is not computable and like the above, this value must be argued.

Remark 1.1. Note that if we use a random sample, we hope it is representative and that the above errors are minimized.

1.3 Data

This step involves the collecting and organizing of data and statistics.

Data Types

- **Discrete Data***:* Simply put, there are “holes” between the numbers
- **Continuous** (CTS) **Data***:* We **assume** that there are no “holes”
- **Nominal Data***:* No order in the data
- **Ordinal Data***:* There is some order in the data
- **Binary Data***:* e.g. Success/failure, true/false, yes/no
- **Counting Data***:* Used for counting the number of events

1.4 Analysis

This step involves analyzing our data set and making well-informed observations and analyses.

Data Quality

There are 3 factors that we look at:

1. **Outliers**: Data that is more extreme than their counterparts.
   - (a) Reasons for outliers?
     - i. Typos or data recording errors
     - ii. Measurement errors
     - iii. Valid outliers
   - (b) Without having been involved in the study from the start, it is difficult to tell which is which
2. Missing Data Points
3. Measurement Issues

Characteristic of a Data Set

Outliers could be found in any data set but these 3 always are:

1. **Shape**
   - (a) Numerical methods: **Skewness** and **Kurtosis** (not covered in this course)
   - (b) Empirical methods:
     - i. Bell-shaped: Symmetrical about a mean
     - ii. Skewed left (negative): densest area on the right
     - iii. Skewed right (positive): densest area on the left
     - iv. Uniform: even all around; straight line
2. **Center** (location)
   - (a) The “middle” of our data
     - i. **Mode**: statistic that asks which value occurs the most frequently
     - ii. **Median** ($Q2$): The middle data value
       - A. The definition in this course is an algorithm

B. We denote n data values by $x_1, x_2, ..., x_n$ and the sorted data by $x_{(1)}, x_{(2)}, ..., x_{(n)}$ where

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

If n is odd then $Q2 = x_{\left(\frac{n+1}{2}\right)}$ and if n is even, then $Q2 = \frac{x_{\left(\frac{n}{2}\right)} - x_{\left(\frac{n+1}{2}\right)}}{2}$

iii. **Mean**: The sample mean is $\bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$

A. The sample mean moves in the the direction of the outlier if an outlier is added (median as well but less of an effect)

(b) **Robustness**: The median is less affected by outliers and is thus *robust*.

3. **Spread** (variability)

(a) **Range**: By definition, this is $x_{(n)} - x_{(1)}$

(b) **IQR** (Interquartile range): The middle half of your data

i. We first define $posn(a)$ as the function which returns the index of the statistic a in a data set. The value of $Q1$ is defined as the median in the data set

$$x_{(1)}, x_{(2)}, ..., x_{(posn(Q2)-1)}$$

and $Q2$ is the median of the data set

$$x_{(posn(Q2)+1)}, x_{(posn(Q2)+2)}, ..., x_{(n)}$$

ii. The IQR of set is defined to be the difference between $Q3$ and $Q1$. That is,

$$IQR = Q3 - Q1$$

iii. A **box plot** is visual representation of this. The whiskers of a box plot represent values that are some value below or above the data set, according to the upper and lower fences, which are the upper and lower bounds of the whiskers respectively. The lower fence, LL, is bounded by a value of

$$LL = Q_1 - 1.5(IQR)$$

and the upper fence, UL, a value of

$$UL = Q_3 + 1.5(IQR)$$

We say that any value that is less than LL or greater than UL is an **outlier**.

Here, we have a visualization of a box plot, courtesy of Wikipedia:

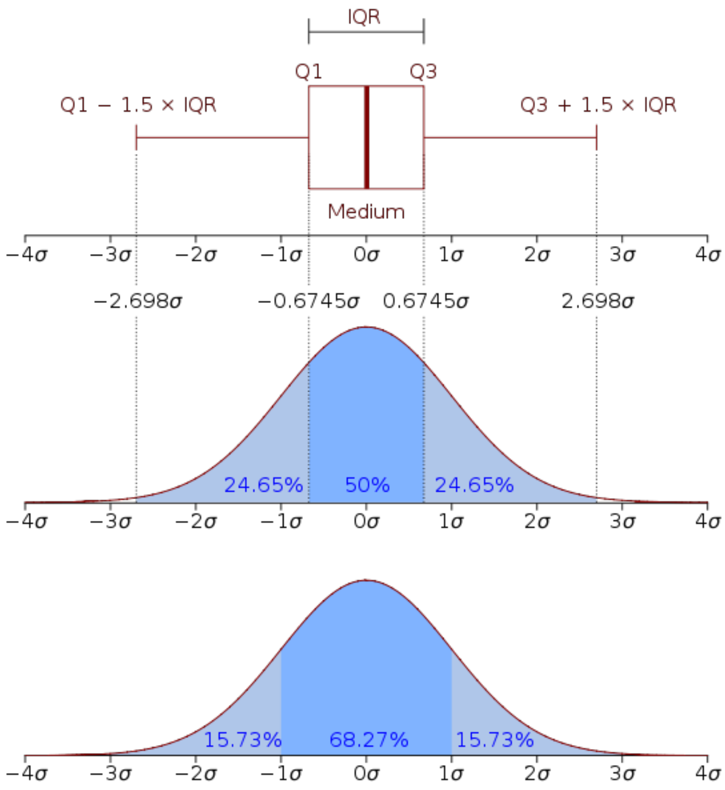

Figure 1.2: Box Plots (from Wikipedia)

(c) *Variance*: For a sample population, the variance is defined to be

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

which is called the **sample variance**.

(d) **Standard Deviation**: For a sample population, the standard deviation is just the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

which is called the **sample standard deviation**.

1.5 Conclusion

In the conclusion, there are only two aspects of the study that you need to be concerned about:

1. Did you answer your problem
2. Talk about **limitations** (i.e. study errors, samples errors)

2 Measurement Analysis

The goal of measures is to explain how far our data is spread out and the relationship of data points.

2.1 Measurements of Spread

The goal of the standard deviation is to approximate the average distance a point is from the mean. Here are some other methods that we could use for standard deviation and why they fail:

1. $\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{n}$ does not work because it is always equal to 0
2. $\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$ works but we cannot do anything mathematically significant to it
3. $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$ is a good guess, but note that $\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 \leq \sum\limits_{i=1}^{n}|x_i-\bar{x}|$
   (a) To fix this, we use $n-1$ instead of n (proof comes later on)

Proposition 2.1. *An interesting identity is the following:*

$$s = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i^2-n\bar{x}^2)}{n-1}}$$

Proof. Exercise for the reader. □

Definition 2.1. Coefficient of Variation (CV)

This measure provides a unit-less measurement of spread:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

2.2 Measurements of Association

Here, we examine a few interesting measures which test the relationship between two random variables X any Y.

1. **Covariance**: In theory (a population), the covariance is defined as

   $$\text{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$$

   but in practice (in samples) it is defined as

   $$s_{XY} = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}.$$

   Note that $\text{Cov}(X,Y), s_{XY} \in \mathbb{R}$ and both give us an idea of the direction of the relationship but not the magnitude.

2. **Correlation***:* In theory (a population), the correlation is defined as

   $$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

but in practice (in samples) it is defined as

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

Note that $-1 \leq \rho_{XY}, r_{XY} \leq 1$ and both give us an idea of the direction of the relationship AND the magnitude.

(a) An interpretation of the values is as follows:

i. $|r_{XY}| \approx 1 \implies$ strong relationship
ii. $|r_{XY}| = 1 \implies$ perfectly linear relationship
iii. $|r_{XY}| > 1 \implies$ positive relationship
iv. $|r_{XY}| < 1 \implies$ negative relationship
v. $|r_{XY}| \approx 0 \implies$ weak relationship

3. **Relative-risk**: From $STAT230$, this the probability of something happening under a condition relative to this same thing happening if the condition is note met. Formally, for two events A and B, it is defined as

$$RR = \frac{P(A|B)}{P(A|\bar{B})}.$$

An interesting property is that if $RR = 1$ then $A \perp B$ and vice versa.

4. **Slope**: This will be covered later on.

3 Probability Theory

All of this content was covered in $STAT230$ so I will not be typesetting it. Check out any probability textbook just to review the concepts and properties of expectation and variance. The only important change was a notational one, specifically that instead of writing $X \sim Bin(n, p)$, we write $X \sim Bin(n, \Pi)$ where $p = \Pi$ still.

4 Statistical Models

Recall that the goal of statistics is to guess the value of a population parameter on the basis of a (or more) sample statistic.

4.1 Generalities

We make our measurements on our sample units. The data values that are collected are:

- The response variate
- The explanatory variate(s)

The response variate is a characteristic of the unit that helps us answer the problem. It will be denoted by Y and will be assumed to be random with a random component ϵ.

Every model is relating the population parameter $(\mu, \sigma, \pi, \rho, ...)$ to the sample values (units). We will use at least one subscript representing the value of unit i. Note that a realization, y_i, is the response that is a achieved by a response variate Y_i.

Example 4.1. Here is a situation involving coin flips:

$$\begin{aligned} Y_i &= \text{flipping a coin that has yet to land} \\ y_i &= \text{coin lands and realizes its potential (H/T)} \end{aligned}$$

In every model we assume that sampling was done randomly. This allows us to assume that $\epsilon_i \perp \epsilon_j$ for $i \neq j$.

4.2 Types of Models

Goal of **statistical models**: explain the relationship between a parameter and a response variate.

The following are the different types of statistical models that we will be examining :

1. **Discrete** (Binary) **Model -** either the population data is within parameters or it is not.

   (a) Binomial: $Y_i = \epsilon_i$, $\epsilon_i \sim Bin(1, \Pi)$

   (b) Poisson: $Y_i = \epsilon_i$, $\epsilon_i \sim Pois(\mu)$

2. **Response Model -** these model the response and *at most* use the explanatory variate implicitly as a focal explanatory variate.

   (a) $Y_i = \mu + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

   (b) $Y_{ij} = \mu_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2)$

   (c) $Y_{ij} = \mu + \tau_i + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2)$

   (d) $Y_i = \mu + \delta + \epsilon_i$

   Where Y_j, Y_{ij} are the responses of unit j [in group i], μ the overall average, μ_i the average in the i^{th} group, τ_i the difference between the overall average and the average in the i^{th} group, and δ equal to some bias.

3. **Regression Model -** these create a function that relates the response and the explanatory variate (attribute or parameter); note here that we assume $Y_i = Y_i | X$.

   (a) $Y_i = \alpha + \beta x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

   (b) $Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

   (c) $Y_i = \beta x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

   (d) $Y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2)$

   (e) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$

   Where Y_j, Y_{ij} are the response of unit j [in group i], α, β_0 are the intercepts, β is the slope, x_j, x_{ij} are the explanatory variates of unit j [in group i], and β_1,β_2 the slopes for two explanatory variates (indicating that Y_i in (e) is a function of two explanatory variates). Note that in the above models, we assume that we can control our explanatory variate and so we treat the $x_i^\prime s$ constant.

Theorem 4.1. *Linear Combinations of Normal Random Variables*

Let $X_i \sim N(\mu_i, \sigma_i^2)$, $X_j \perp X_i$ for all $i \neq j$, $k_i \in \mathbb{R}$. Then,

$$T = k_0 + \sum_{i=1}^{n} k_i X_i \implies T \sim N\left(k_0 + \sum_{i=1}^{n} k_i \mu_i, \sum_{i=1}^{n} k_i^2 \sigma_i^2\right)$$

Proof. Do as an exercise. □

5 Estimates and Estimators

In this section, we continue to develop the relationship between our population and sample.

5.1 Motivation

Suppose that I flip a coin 5 times, with the number of heads $Y \sim \text{Bin}(5, \Pi = \frac{1}{2})$. Note that Π is given. What's the value of y that maximizes $f(y)$? Unfortunately, since y is discrete, the best that we can do is draw a histogram and look for the maximum. This is a very boring problem.

The **maximum likelyhood test**, however, asks the question in reverse. That is, we find the optimal parameter for Π , given y, such that $f(y)$ is at its maximum. From the formula of f, given by

$$f(y) = \binom{5}{y}\Pi^y(1-\Pi)^{5-y}$$

we can see that if we consider f as a function Π, it makes it a continuous function. To compute the maximum, it is just a matter of using logarithmic differentiation,

$$\ln f(y) = \ln\binom{5}{y} + y\ln\Pi + (5-y)\ln(1-\Pi)$$

finding the partials,

$$\frac{\partial \ln f(y)}{\partial \Pi} = \frac{y}{\Pi} - \frac{5-y}{1-\Pi}$$

and setting them to zero to find the maximum

$$\begin{aligned}\frac{\partial \ln f(y)}{\partial \Pi} = 0 \quad &\implies \quad 0 = \frac{y}{\Pi} - \frac{5-y}{1-\Pi} \\ &\implies \quad y = 5\Pi \\ &\implies \quad \Pi = \frac{5}{y}\end{aligned}$$

and so if we take $\Pi = \frac{y}{5}$, then $f(y)$ is maximized.

5.2 Maximum Likelihood Estimation (MLE) Algorithm

Here, we will formally describe the **maximum likelihood estimation** (MLE) algorithm whose main goal is to know what estimate for a study population parameter θ should be used, given a set of data points such that probability of the data set being chosen is at its maximum.

Suppose that we posit that a certain population follows a given statistical model with unknown parameters $\theta_1, \theta_2, ..., \theta_m$. We then draw n data points, $y_1, y_2, ..., y_n$ randomly (this is important) and using the data, we try to estimate the unknown parameters. The algorithm goes as follows:

1. Define $L = f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i)$ where we call L a **likelihood function**. Simplify if possible. Note that $f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i)$ because we are assuming random sampling, implying that $y_i \perp y_j, \forall i \neq j$.
2. Define $l = \ln(L)$. Simplify l using logarithmic laws.
3. Find $\frac{\partial l}{\partial \theta_1}, \frac{\partial l}{\partial \theta_2}, ..., \frac{\partial l}{\partial \theta_n}$, set each of the partials to zero, and solve for each θ_i, $i = 1, ..., n$. The solved $\theta_i' s$ are called the **estimates** of f and we add a hat, $\hat{\theta}_i$, to indicate this.

To illustrate the algorithm, we give two examples.

Example 5.1. Suppose $Y_i = \epsilon_i,\ \epsilon_i \sim \text{Exp}(\theta)$ and θ is a rate. What is the optimal $\hat{\theta}$ according to MLE?

Solution.

1. $L = \prod_{i=1}^{n} \theta \exp(-y_i\theta) = \theta^n \exp(-\sum_{i=1}^{n} y_i\theta)$
2. $l = \ln(L) = n\ln\theta - \sum_{i=1}^{n} y_i\theta$
3. $\frac{\partial l}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} y_i = \frac{n}{\theta} - n\bar{y}$, $\frac{\partial l}{\partial\theta} = 0 \implies \frac{n}{\hat{\theta}} - n\bar{y} = 0 \implies \hat{\theta} = \frac{1}{\bar{y}}$.

So the maximum l, and consequently maximum L is obtained, when $\theta = \frac{1}{\bar{y}}$.

Example 5.2. Suppose $Y_i = \alpha + \beta x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$. Use MLE to estimate α and β.

Solution.

First, take note that $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

1. We first simplify L as follows.

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-(y_i - \alpha - \beta x_i)^2/2\sigma^2) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{\sigma^n} \exp(\sum_{i=1}^{n} -(y_i - \alpha - \beta x_i)^2/2\sigma^2) \\ &= K \cdot \frac{1}{\sigma^n} \exp(\sum_{i=1}^{n} -(y_i - \alpha - \beta x_i)^2/2\sigma^2) \end{aligned}$$

   where $K = \frac{1}{(2\pi)^{\frac{n}{2}}}$.

2. By direct evaluation,

$$l = \ln K - n\ln\sigma - \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2/2\sigma^2$$

3. Computing the partials, we get

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

   and

$$\frac{\partial l}{\partial\alpha} = \sum_{i=1}^{n} (x_i)(y_i - \alpha - \beta x_i)/\sigma^2.$$

   So first solving for α we get

$$\begin{aligned} \frac{\partial l}{\partial\alpha} = 0 \quad &\implies \quad \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \\ &\implies \quad n\bar{y} - n\hat{\alpha} - n\hat{\beta}\bar{x} = 0 \\ &\implies \quad \hat{\alpha} = \bar{y} - \hat{B}\bar{x} \end{aligned} \tag{5.1}$$

and extending to β we get

$$\begin{aligned}
\frac{\partial l}{\partial \beta} = 0 &\implies \sum_{i=1}^{n}(x_i)(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\
&\implies \sum_{i=1}^{n} x_i y_i - n\bar{x}\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i^2 \\
&\overset{(5.1)}{\implies} \sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0 \\
&\implies \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 0 \\
&\implies \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{\hat{\beta}\left(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{n-1} \\
&\implies \hat{\beta}s_x^2 = s_{xy} \\
&\implies \hat{\beta} = \frac{s_{xy}}{s_x^2}
\end{aligned}$$

Before we head off into the next section, we should recall a very important theorem from $STAT230$.

Theorem 5.1. *(Central Limit Theorem)*

Let $X_1, X_2, ..., X_n$ be i.i.d.[1] *random variables with distribution F, $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $\forall i$, and $X_i \perp X_j$, $\forall i \neq j$. Then as $n \to \infty$, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, $\sum\limits_{i=1}^{n} X_i \sim N(n\mu, \sigma^2 n)$.*

Proof. Beyond the scope of this course. □

5.3 Estimators

One of the pitfalls that you may notice with the MLE Algorithm is that the estimate is relative to the sample data that we collect from the study population. For example, suppose that we have two samples drawn from our population, $S_1 = \{y_{i1}\}$ and $S_2 = \{y_{i2}\}$, of the same size, taken randomly with replacement and suppose that we model the data in our population with the model $Y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$. It could be the case that when we apply MLE to both of them to get the following two models, $y_{i1} = \hat{\alpha}_1 + \hat{\beta}_1 x_i + \epsilon_i$ and $y_{i2} = \hat{\alpha}_2 + \hat{\beta}_2 x_i + \epsilon_i$, that the estimates may differ. That is, it is possible that $\hat{\alpha}_1 \neq \hat{\alpha}_2$ and/or $\hat{\beta}_1 \neq \hat{\beta}_2$.

This is because we are doing random sampling, which motivates us to believe that if we take n random samples, $S_j = \{y_{ij}\}$, from the study population of equal size, then the estimates, θ_j, for a parameter θ of the samples should follow a distribution.

We call the random variable representing distribution an **estimator** and denote it as the parameter in question with a tilde on the top, $\tilde{\theta}$. Note that $\tilde{\theta}$ describes the distribution of θ for the population and not the sample. In other words, the relationship is that $\hat{\theta}$ is a specific realization of $\tilde{\theta}$ through the sampled data points. Now the obvious question to ask here is what exactly the distribution of an estimator is, given the model that we are using to model the population. This is actually dependent on what we get as the estimate, which we will see in the following examples.

Example 5.3. Given a model $Y_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, if one were to compute the best estimate for the parameter μ using MLE, one would obtain $\hat{\mu} = \bar{y}$. Because the estimator that we want to examine describes the population and not the sample, and because we are doing random sampling, we capitalize the y to show this (we move from the sample, up a level of abstraction, into the study population). That is, $\tilde{\mu} = \bar{Y}$. So what is the distribution of $\tilde{\mu}$ in this case? Recall from Thm 4.1. that a linear combination of normally distributed random variables is normal. Since

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

[1] i.i.d = independent and identically distributed

$\tilde{\alpha}$ is normal with $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \frac{\sigma^2}{n}$. So $\tilde{\alpha} \sim N(\mu, \frac{\sigma^2}{n})$.

Example 5.4. Let's try a slightly more difficult example. Suppose we want to find the distribution of $\tilde{\alpha}$ and $\tilde{\beta}$ for the model $Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$. Through similar methods shown in Ex. 5.2., using MLE, one should be able to obtain the estimates $\hat{\alpha} = \bar{y}$ and $\hat{\beta} = \frac{s_{xy}}{s_x^2}$. Now, from Ex. 5.3., we already computed the distribution for $\tilde{\alpha}$ as $\tilde{\alpha} \sim N(\mu, \frac{\sigma^2}{n})$, so what we are more interested in is $\tilde{\beta}$. Through a simple re-arrangement of the definition of $\tilde{\beta}$,

$$\tilde{B} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(x_i - \bar{x})^2} Y_i$$

since $\bar{Y} \sum\limits_{i=1}^{n}(x_i - \bar{x}) = 0$. So $\tilde{\beta}$ is actually a linear combination of normal random variables, making it normal as well. Computing the expectation, we get

$$\begin{aligned} E(\tilde{B}) &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(x_i - \bar{x})^2} E(Y_i) \\ &= \sum_{i=1}^{n} \left[\frac{1}{(x_i - \bar{x})^2} \right] (x_i - \bar{x})(\alpha + \beta(x_i - \bar{x})) \\ &= \frac{1}{(n-1)s_x^2} \left[\alpha \overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{=0} + \beta \overbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}^{=s_x^2(n-1)} \right] \\ &= \beta \left(\frac{s_x^2}{s_x^2} \right) \\ &= \beta. \end{aligned}$$

We can similarly compute the variance using this method, although this was not done in lectures. I will leave it to the reader as an exercise (I posit that variance should be 0). Thus, $\tilde{\beta} \sim N(\beta, Var(\tilde{B}) \overset{?}{=} 0)$.

5.4 Biases in Statistics

In this section we take a look at one of the problems of using estimators.

Definition 5.1. We say that for a given estimator, $\tilde{\theta}$, of an estimate for a model is **unbiased** if the following holds

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

Otherwise, we say that our estimator is **biased**.

One way to intuitively look at this definition is that we say an estimator is unbiased if we take n samples of equal size with estimators $\tilde{\theta}_i$ for each sample, and on average $\frac{\sum\limits_{i=1}^{n}\tilde{\theta}_i}{n} \approx \theta$ for large enough n, or as $n \to \infty$, $\frac{\sum\limits_{i=1}^{n}\tilde{\theta}_i}{n} \to \theta$. Since all the the examples that we've seen were so far unbiased, let's take a look at an unbiased one.

Example 5.5. Consider the model $Y_i = \mu + \epsilon_i,\ \epsilon_i \sim N(\mu, \sigma^2)$. If one were to compute the best estimate for σ^2 using MLE, the result would be

$$\hat{\sigma}^2 = \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}{n} \implies \tilde{\sigma}^2 = \frac{\sum\limits_{i=1}^{n}(Y_i - \bar{Y})^2}{n} = \frac{\left(\sum\limits_{i=1}^{n} Y_i^2\right) - n\bar{Y}^2}{n}.$$

From here, let's try to compute the expectation. However, before that, it looks that finding $E(\bar{Y}^2)$ and $E(Y^2)$would be helpful in this situation. We can find it directly through the variance of $\bar{Y}$ and Y.

$$\begin{aligned} Var(\bar{Y}) = E(\bar{Y}^2) - \left[E(\bar{Y})\right]^2 \quad &\implies \quad \frac{\sigma^2}{n} = E(\bar{Y}^2) - \mu^2 \\ &\implies \quad E(\bar{Y}^2) = \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

and similarly

$$E(Y^2) = \sigma^2 + \mu^2$$

So computing expectation, we get

$$\begin{aligned} E(\tilde{\sigma}^2) &= E\left(\frac{\left(\sum_{i=1}^{n} Y_i^2\right) - n\bar{Y}^2}{n}\right) \\ &= \frac{1}{n}\left[\left(\sum_{i=1}^{n} E(Y_i^2)\right) - nE(\bar{Y}^2)\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\ &= \frac{1}{n}\left[n\sigma^2 + n\mu^2 - \sigma^2 + n\mu^2\right] \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

showing that our best estimate, and consequently estimator, for σ^2 is biased! We have to correct for this by changing our estimator to

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

and consequently the estimate that naturally comes out of this is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - E(\hat{Y}_i))^2}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}.$$

Theorem 5.2. *Given a model in the form $Y_{ji} = B_0 + \sum_{j=1}^{q} B_{ji}x_i + \epsilon_i$, the best estimate for σ^2 is*

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

where there are $q+1$ non-sigma parameters.

Proof. Exercise for the reader. □

6 Distribution Theory

In this section we will examine a couple new distributions using information about currently known distributions so far. Recall from $STAT230$ that

- If $X_i \sim \text{Bin}(1, \Pi)$ then $\sum\limits_{i=1}^{n} X_i \sim \text{Bin}(n, \Pi)$.
- If $X_i's$ are i.i.d with $X_i \sim N(\mu, \sigma^2)$ for all i, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

6.1 Student t-Distribution

Now we introduce the following new distributions.

- If $X \sim N(0,1)$ then $X^2 \sim \chi^2_1$ which we call a **Chi-squared** (pronounced "Kai-Squared") **distribution** on one degree of freedom
- Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$. Then $X + Y \sim \chi^2_{n+m}$ which is a Chi-squared on $n+m$ degrees of freedom
- Let $N \sim N(0,1)$, $X \sim \chi^2_v$, $X \perp N$. Then $\frac{N}{\sqrt{\frac{X}{v}}} \sim t_v$ which we call a **student's t-distribution** on v degrees of freedom

You can see the density functions for the above two in the Appendix, although they will not be necessary for this course.

The Chi-squared is not motivated by anything in particular, other than as a means to describe the student's t. But where does the motivation for the student t come from you ask? It comes from trying to find a distribution for $\tilde{\sigma}^2$ from the last section!

There is going to be a lot of hand-waving in this "proof" because most of the rigourous theory is beyond the scope of this course so just believe most of what I am going to do. We'll start off with the simplest response model $Y_i = \mu + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$. Recall that our unbiased estimator for σ^2 was

$$\tilde{\sigma}^2 = \frac{\sum\limits_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$$

and doing some rearranging, using the fact that $\bar{\epsilon} = \mu - \bar{Y}$ and $\epsilon_i = \mu - Y_i$, this becomes

$$\begin{aligned}
\tilde{\sigma}^2 &= \frac{\sum\limits_{i=1}^{n} (Y_i - \mu + \mu - \bar{Y})^2}{n-1} \\
&= \frac{\sum\limits_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2}{n-1} \\
&= \frac{\sum\limits_{i=1}^{n} (\epsilon_i^2 - 2\epsilon_i\bar{\epsilon} + \bar{\epsilon})}{n-1}
\end{aligned}$$

and multiplying both sides by $n-1$ and dividing through by σ^2 yields

$$\begin{aligned}
\frac{(n-1)\tilde{\sigma}^2}{\sigma^2} &= \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma^2}\right)^2 - \frac{2\bar{\epsilon}\sum\limits_{i=1}^{n}\epsilon_i}{\sigma^2} + \frac{n\bar{\epsilon^2}}{\sigma^2} \\
&= \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma^2}\right)^2 - \frac{n\bar{\epsilon}^2}{\sigma^2} \\
&= \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma^2}\right)^2 - \left(\frac{\bar{\epsilon}}{\frac{\sigma}{\sqrt{n}}}\right)^2
\end{aligned}$$

where the terms in the round brackets are Chi-squared with one degree of freedom (you can prove this as an exercise). Thus, based on one of the properties of the Chi-squared, $X = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$, although we cannot show this explicitly since the terms in the round brackets are not independent (hence the hand-waving). Next, we let $Z = \frac{\bar{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ and consider

$T = \frac{Z}{\sqrt{\frac{X}{n-1}}} \sim t_{n-1}$:

$$T = \frac{\left(\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\frac{\left(\frac{(n-1)\tilde{\sigma}^2}{\sigma^2}\right)}{n-1}}} = \frac{\bar{Y}-\mu}{\frac{\tilde{\sigma}^2}{\sqrt{n}}} \sim t_{n-1}$$

and in general

$$\frac{(n-q)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-q} \implies \frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}} \sim t_{n-q}$$

for some non-sigma parameter θ and sufficient large n, where q is the number of non-sigma parameters, $\tilde{\sigma}^2$ the estimator of σ^2 and the σ^2 the population variance.

Properties of the Student's t-Distribution

- This distribution is symmetric
- For distribution $T \sim t_v$, when $v > 30$, the student's t is almost identical to the normal distribution with mean 0 and variance 1
- For $v \ll 30$, T is very close to a uniform distribution with thick tails and very even, unpronounced center

(From here on out, the content will be supplemental to the course notes and will only serve as a review)

6.2 Least Squares Method

There are two ways to use this method. First, for a given model Y and parameter θ, suppose that we get a best fit $\hat{y}$ and define $\hat{\epsilon}_i = |\hat{y} - y_i|$. The least squares approach is through any of the two

1. (Algebraic) Define $W = \sum\limits_{i=1}^{n} \hat{\epsilon}_i^2$. Calculate and minimize $\frac{\partial W}{\partial \theta}$ to determine θ.
2. (Geometric) Define $W = \sum\limits_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\epsilon}^t \hat{\epsilon}$. Note that $W \perp \text{span}\{\overrightarrow{1}, \overrightarrow{x}\}$ and so $\hat{\epsilon}^t \overrightarrow{1} = 0$ and $\hat{\epsilon}^t \overrightarrow{x} = 0$. Use these equations to determine θ.

7 Intervals

Here, we deviate from the order of lectures and focus on the various types of constructed intervals.

7.1 Confidence Intervals

Suppose that $\tilde{\theta} \sim N(\theta, Var(\tilde{\theta}))$. Find L, U, equidistant from $\tilde{\theta}$, such that $P(L < \tilde{\theta} < U) = 1 - \alpha$ where α is known as our **margin of error**. Usually this value is 0.05 by convention, unless otherwise specified. Normalizing $\tilde{\theta}$, we get that

$$(L, U) = \left(\theta - c\sqrt{Var(\tilde{\theta})}, \theta + c\sqrt{Var(\tilde{\theta})}\right)$$

where c is the value such that $1 - \alpha = P(-c < Z < c)$ and $Z \sim N(0,1)$. We call this interval, $\theta \pm c\sqrt{Var(\tilde{\theta})}$, a **probability interval**.

Usually, though, we don't know θ, so we replace it with our best estimate $\hat{\theta}$ to get $\hat{\theta} \pm c\sqrt{Var(\tilde{\theta})}$, which is called a $(1-\alpha)\%$ **confidence interval**. If $Var(\tilde{\theta})$ is known, then $C \sim N(0,1)$ and if it is unknown, we replace $Var(\tilde{\theta})$ with $\hat{Var}(\tilde{\ }\theta)$ and

$C \sim t_{n-q}$. Another compact notation for confidence intervals is $EST \pm cSE$. Note that the confidence interval says that we are 95% confident in our result, but does not say anything significant. When we say that we are 95% confident, it means that given 100 randomly selected trials, we expect that in 95 of the 100 confidence intervals constructed from the samples, the population parameter will be in those intervals.

7.2 Predicting Intervals

A **predicting interval** is an extension of a model Y, by adding the subscript p. The model is usually of the form $Y_p = f(\tilde{\theta}) + \epsilon_p$ and the model of the form $EST \pm cSE = f(\hat{\theta}) \pm \sqrt{Var(Y_p)}$. Note that Y_p is different from a standard model $Y_i = f(\theta) + \epsilon_i$ in that the first component is random.

7.3 Likelyhood Intervals

Recall the likelyhood function $L(\theta, y_i) = \prod_{i=1}^{n} f(y_i, \theta)$ and note that $L(\hat{\theta})$ is the largest value of the likelyhood function by MLE. We define the **relative likelyhood function** as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, 0 \leq R(\theta) \leq 1.$$

From a theorem in a future statistics course (STAT330), if $R(\theta) \approx 0.1$, then solving for θ (usually using the quadratic formula) will form an approximate 95% confidence interval. This interval is called the **likelyhood interval** and one of its main advantages to the confidence interval is that it does not require that the model be normal.

8 Hypothesis Testing

While our confidence interval does not tell us in a yes or no way whether or not a statistical estimate is true, a hypothesis test does. Here are the steps:

1. State the hypothesis, $H_0 : \theta = \theta_0$ (this is only an example), called the null hypothesis (H_1 is called the alternative hypothesis and states a statement contrary to the null hypothesis).
2. Calculate the discrepancy (also called the test statistic), denoted by

$$d = \frac{\hat{\theta} - \theta_0}{\sqrt{Var(\tilde{\theta})}} = \frac{estimate - H_0\, value}{SE}$$

   assuming that $\tilde{\theta}$ is unbiased and the realization of d, denoted by D, is $N(0,1)$ if $Var(\tilde{\theta})$ is known and t_{n-q} otherwise. Note that d is the number of standard deviations θ_0 is from $\hat{\theta}$.
3. Calculate a $p-$value given by $p = 2P(D > |d|)$. It is also the probability that one sees a value worse than $\hat{\theta}$, given that the null hypothesis is true. The greater the $p-$value, the more evidence against the model in order to reject.
4. Reject or not reject (note that we do not “accept” the model)

The following the table that subjectively describes interpretations for $p-$values:

P value	Interpretation
p-value<1%	A ton of evidence against H_0
1%≤p-value<5%	A lot of evidence against H_0
5%≤p-value≤10%	Some evidence against H_0
p-value>10%	There is virtually no evidence against H_0

Note that one model for which $D \sim N(0,1)$ is $Y_i = \epsilon_i$ where $\epsilon_i \sim$ Bin$(1, \Pi)$ since $\sqrt{Var(\tilde{\theta})} = \sqrt{\frac{\hat{\Pi}_0(1-\hat{\Pi}_0)}{n}}$ by our null hypothesis and central limit theorem.

9 Comparative Models

The goal of a comparative model is to compare the mean of two groups and determine if there is a causation relationship between one and the other.

Definition 9.1. If x causes y and there is some variate z that is common between the two, then we say z is a **confounding variable** because it gives the illusion that z causes y. It is also sometimes called a **lurking variable**.

There are two main models that help determine if one variate causes another and they are the following.

Experimental Study

1. For every unit in the T.P. set the F.E.V. (focal explanatory variate) to level 1
2. We measure the attribute of interest
3. Repeat 1 and 2 but with set the F.E.V. to level 2
4. Only the F.E.V. changes and every other explanatory variate is fixed
5. If the attribute changes between steps 1 and 4, then causation occurs

Problems?

- We cannot sample the whole T.P.
- It is not possible to keep all explanatory variates fixed
- The attributes change (on average)

Observational Study

1. First, observe an association between x and y in many places, settings, types of studies, etc.
2. There must be a reason for why x causes y (either scientifically or logically)
3. There must be a consistent dose relationship
4. The association has to hold when other possible variates are held fixed

10 Experimental Design

There are three main tools that are used by statisticians to improve experimental design.

1. Replication
   (a) Simply put, we increase the sample size
      i. This is to decrease the variance of confidence intervals, which improves accuracy
2. Randomization
   (a) We select units in a random matter (i.e. if there are 2+ groups. we try to randomly assign units into the groups)
      i. This is to reduce bias and create a more representative sample
      ii. It allows us to assume independence between $Y_i's$
      iii. It reduces the chance of confounding variates by unknown explanatory variates

3. Pairing
   (a) In an experimental study, we call it blocking and in an observational study, we call it matching
   (b) The actual process is just matching units by their explanatory variates and matched units are called twins
      i. For example in a group of 500 twins, grouped by gender and ages, used to test a vaccine, one of the twins in each group will take the vaccine and another will take a placebo
      ii. We do this in order to reduce the chance of confounding due to known explanatory variates
   (c) Pairing also allows us to perform subtraction between twins to compare certain attributes of the population
      i. Note that taking differences does not change the variability of the difference distribution

11 Model Assessment

We usually want the following four assumptions to be true, when constructing a model $Y_i = f(\theta) + \epsilon_i$ to fit a sample. We also use certain statistical tools to measure how well our model fits these conditions. Note that these tools/tests require subjective observation.

- $\epsilon_i \sim N(0, \sigma^2)$
  - Why? Because Y_i is not normal if ϵ_i is not normal. However, $\tilde{\theta}$ is still likely to be normal by CLT.
  - Tests:
    * Histogram of residuals $\hat{\epsilon_i} = |\hat{y} - y_i|$ (should be bell-shaped)
    * QQ Plot, which is the plot of theoretical quartiles versus sample quartiles (should be linear with intercept ∼0)
      · Usually a little variability at the tails of the line is okay
- $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2, \epsilon_i's$ are independent
  - Why? All of our models, tests and estimates depend on this.
  - Tests:
    * Scatter plot of residuals (y-axis) versus fitted values (x-axis)
      · We hope that it is centered on 0, has no visible pattern and that the data is bounded by two parallel lines (constant variance)
      · If there is not a constant variance, such as a funnel (funnel effect), we usually transform the fitted values (e.g. $y \to \ln y$)
      · If the plot seems periodic, we will need a new model (STAT 371/372)
    * Scatter plot of fitted values (y-axis) versus explanatory variates (x-axis)
      · This is used mainly in regression models
      · We hope to see the same conditions in the previous scatter plot

12 Chi-Squared Test

The purpose of a Chi-squared test is to determine if there is an association between two random variables X,Y, given that they both contain only counting data. The following are the steps

1. State the null hypothesis as H_0 : X and Y are not associated.
2. If there are m possible observations for X and n possible observations for Y, then define

$$d = \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(expected - observed)^2}{expected} = \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(e_{ij} - o_{ij})^2}{e_{ij}}$$

   where $e_{ij} = P(X = x_i) \cdot P(Y = y_j)$, $o_{ij} = P(X = x_i, Y = y_j)$, for $i = 1, ..., m$ and $j = 1, ..., n$.

3. Assume that $D \sim \chi^2_{(m-1)(n-1)}$.
4. Calculate the p-value which in this case is $Pr(D > d)$ since $d \geq 0$, which means we are conducting a one-tailed hypothesis.
5. Interpret it as always (see the table in Section 8)

Appendix A

Chi-squared c.d.f. and p.d.f. on k degrees of freedom (χ_k^2)[2]:

$$f(x,k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

$$F(x,k) = \frac{1}{\Gamma(\frac{k}{2})}\gamma(\frac{k}{2},\frac{x}{2})$$

Student's t c.d.f. and p.d.f. on v degrees of freedom (t_v)[3]:

$$f(x,v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

$$F(x,v) = \frac{1}{2} + x\Gamma\left(\frac{v+1}{2}\right)\cdot\frac{{}_2F_1\left(\frac{1}{2},\frac{v+1}{2};\frac{3}{2};-\frac{x^2}{v}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}$$

[2] $\Gamma(x)$ denotes the regular gamma functionand $\gamma(x)$ is the lower incomplete gamma function.

[3] ${}_2F_1$ denotes the hypergeometric function.

Index