University of Waterloo

STAT 231 - Statistics

Personal Course Notes

Brandon Zhou

Spring 2023

Contents

DISCLAIMER

While I've made every effort to ensure the accuracy of the content, there's potential for errors or outdated information. It's important to approach these notes as a supplementary reference and not as a primary source. Should you come across any uncertainties or ambiguities in the material, I strongly recommend consulting with your course instructors or the course staff for a clearer understanding. I apologize in advance for any potential discrepancies or oversights. Additionally, any changes made to this notebook after its initial creation are not endorsed or recognized by me. Always cross-reference with trusted resources when in doubt.

1 INTRODUCTION TO STATISTICAL SCIENCES

An *empirical study* is one in which knowledge is gained by observation or by experiment. Empirical studies deal with *populations* and *processes* which are collections of individual units. A key feature of an empirical study is that it involves **uncertainty**!

A *unit* is an individual person, place or thing about which we can take some measurements. A *population* is a collection of *units.* A *process* is a system by which units are produced.

A key feature of processes is that they usually occur over time whereas populations are often static (defined at one moment in time).

Definition 1.1 A variate is a characteristic of a unit.

Variates can be of different types:

continuous those that can be measured - at least in theory - to an infinite degree of accuracy

discrete can only take a certain number of values

categorical units fall into a (non-numeric) category such as marital status or hair color. Categorical variates can be redefined to discrete variate. A variate only takes on values 0 or 1 such a variate is often referred to as a *binary* variate.

ordinal categorical variates where an ordering is implied (i.e. small, medium, large for sizes)

complex more unusual variates such open-ended responses to a survey question, or an image. Anything that doesn't fit the above categories is complex(by default).

Cutoff: Lecture 1, May 8 2023

Definition 1.2 An attribute of a population or process is a function of the variates over the population or process.

Example 1.1 Here are some examples of attributes:

- proportion of adults in Ontario who own a smartphone
- average drop in blood pressure due to a treatment for individuals with hypertension.

  population Everyone who gets the treatment

  Possible population Everyone who WILL/SHOULD get this treatment OR everyone with hypertension

Sample Studies Asking; In this case information about the population may be obtained by selecting a "representative" sample of units from the population and determining the variates of interest for each unit in the sample.

Observational Studies Looking; An observational study is one in which data are collected about a population or process without any attempt to change the value of one or more variates for the sampled units. Just measuring(didn't warn in advance). An advantage is this can be done with datasets that happened long in the past. However, one disadvantage is limitations on the inferences you can make. You CANNOT make a CAUSATIVE study statement from an observation study.

Experimental Studies Active, intentional changes; (In contrast to observational studies,) An experimental study is one in which the experimenter (that is, the person conducting the study) intervenes and changes or sets the values of one or more variates for the units in the sample. An advanatge is you can make CAUSATIVE statement, which means you can say something like "IF this happens, THEN this other thing hapens", however, this method is expensive, not scalable, and with ethical concerns.

Both the last two studies need to have *randomization* or else their results are irrelevent.
A distinction between a sample survey and an observational study is that for observational studies the population of interest is usually infinite or conceptual.
Now let's summarize numerical data, there are 3 types of numrical measures:

1. Measures of location (sample mean, median and more)
2. Measures of variabeility or dispersion (sample variance, sample standard, deviation, range, and IQR)
3. Measures of shape (sample skewness and sample kurtosis)

These summaries are used when the variate is either discrete or continuous.
Measures of Central Tendency or Location
Let the data be represented as $y_1, y_2, \ldots, y_n$ where y_i is a real number.
Numerical measures of the "center" of the data:

Sample mean or average $\bar{y} = \frac{y_1+y_2+\ldots+y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Sample median $\hat{m}(q(0.5)$, or $50^{th})$

Mode Most frequent observed observation(s). It is the most common value in the set of data. If the values are all unique then the model does not exist. And the mode is *not necessarily unique.* For frequency or grouped data the group or class with the highest frequency is called the modal class.

The units for mean, median and mode (e.g. centimeters, degrees Celsius, etc.) are the same as for the original variate.
We denote the **ordered sample** (called the order statistic) as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ where $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ and $y_1 = min(y_1, \ldots, y_n), y_n = max(y_1, \ldots, y_n)$.
For an **odd** number of observations: $median = \hat{m} = y_{(\frac{n+1}{2})}$. For an **even** number of observations: $median = \hat{m} = \frac{1}{2}(y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)})$
The p^{th} quantile (or $100 * p^{th}$ percentile) is a value, call it q(p) determined as follows: let $m = (n+1) * p$,

- If $m \in 1, 2, \ldots, n$ then we take the m^{th} smallest value, q(p) $= y_m$, where $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ denotes the ***ordered sample values***.
- If $m \notin 1, 2, \ldots, n$, but $1 < m < n$ then determine the closest integer j such that $j < m < j+1$ and take q(p)$=\frac{1}{2}[y_{(j)} + y_{(j+1)}]$

Cutoff: Lecture 2, May 10 2023

Which of mean, median and mode should we use?

- The mode is least useful.
- If you have well-behaved data, mean is preferred.
- If you have outliers, or "strange observations", consider median. It is said to be **robust**. (the median is less affected by a few extreme observations)

If the mean is greater than the median, then the distribution is said to be positively skewed, or skewed to the right, as the distribution has a longer tail to the right. In contrast, if a distribution is more symmetric, the mean, median and mode will be similar.

Measures of Dispersion of Spread

Let the data be represented as $y_1, y_2, \ldots, y_n$ where y_i is a real number.

The **sample variance** is defined as

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

The sample standard deviation is s, the square root of s^2.

If the data are unimodal and roughly symmetric then:

- Approximately 68% of the data will lie within one standard deviation of the mean.
- Approximately 95% of the data will lie within two standard deviations of the mean.
- Approximately 99.7% of the data will lie within three standard deviations of the mean.

The **range** of the data set is given by the difference between the largest observation and the lowest observation $= y_{(n)} - y_{(1)}$. Although it is easy to calculate, it is a function of only two observations in the data set. It is common for data to be divided into quartiles. There are $25th$, $50th$ and $75th$ percentile.

Sample Quantiles and Percentiles

For $0 < p < 1$, the pth *quantile* (also called the 100pth *percentile*) is a value such that approximately a fraction p of the y values in the data set are less than $q(p)$ and approximately $1 - p$ are greater than $q(p)$.

Definition 1.3 Let $\{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\}$ where $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be the order statistics for the dataset $\{y_1, y_2, \ldots, y_n\}$. For $0 < p < 1$, the pth (sample) quantile (also called the $100p$ (sample) percentile), is a value, call it $q(p)$, determined as follows:

- Let $k = (n+1)p$ where n is a sample size.
- If k is an integer and $1 \leq k \leq n$, then $q(p) = y_{(k)}$.
- If k is not an integer but $1 < k < n$ then determine the closet integer j such that $j < k < j+1$ and the $q(p) = \frac{1}{2}[y_{(j)} + y_{(j+1)}]$.

Note that we use $n+1$ instead of 1 since $Q(1) = \infty$.

Definition 1.4 The quantiles $q(0.25)$, $q(0.5)$ and $q(0.75)$ are called the lower or first quartile, the median, and the upper or third quartile respectively.

- $q(0.25)$, or the $25th$ percentile is known as the lower or first quartile.
- $q(0.75)$, or the $75th$ percentile is known as the upper or third quartile.

Definition 1.5 The interquartile range is $IQR = q(0.75) - q(0.25)$.

Interquartile Range (IQR) is the difference between the upper and lower quartiles $= q(0.75) - q(0.25)$. The IQR measures the spread of the middle half (or 50%) of the distribution / dataset. The IQR is often used as a measure of spread instead of the crude range. It is more 'robust' measure of variability, as it is less affected by extreme outliers.

Definition 1.6 The five number summary of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile and the largest value, that is, the five values: $y_{(1)}$, $q(0.25)$, $q(0.5)$, $q(0.75)$, $y_{(n)}$.

The units for standard deviation, range, and interquartile range (e.g. centimeters, degrees Celsius, etc.) are the same as for the original variate.
Since the interquartile range is less affected by a few extreme observations, it is a more robust measure of variability.
Measures of shape
Skewness measures the degree of asymmetry of a distribution. It can either be positively skewed, negatively skewed and symmetric (not skewed).

- If the mean is less than the median, then the distribution is said to be negatively skewed, or skewed to the left, as the distribution has a longer tail to the left. If the relative frequency histogram of the data had a long left tail then the negative values of $\sum_{i=1}^{n}(y_i - \bar{y})^3$ dominate the positive values in the sum and the value of the skewness will be negative.

- If the mean is greater than the median, then the distribution is said to be positively skewed, or skewed to the right, as the distribution has a longer tail to the right. If the relative frequency histogram of the data has a long right tail, then the positive values of $\sum_{i=1}^{n}(y_i - \bar{y})^3$ dominate the negative values in the sum and the value of the skewness will be positive.

- If the distribution is symmetric, then it is not skewed, and would have a skewness of 0.

The formula for sample skewness is

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

Note: The numerator dictates the sign for the skewness coefficient. The denominator will always be positive. The more skewed the distribution is, the higher the absolute value of $g1$ will be. *It is unitless, actually, sample skewness and sample kurtosis have no units.*

When the relative frequency histogram of the data is approximately symmetric then there is an approximately equal balance between the positive and negative values in the sum $\sum_{i=1}^{n}(y_i - \bar{y})^3$ and this results in a value for the sample skewness that is approximately zero.

Cutoff: Lecture 3, May 12 2023

---

unimodal - one peak, roughly symmetric.

Bimodal - two peaks.

- The sample skewness

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

- The sample kurtosis

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

Measures of shape generally indicate how the data, in terms of a relative frequency histogram, differ from the Normal bell-shaped curve, for example whether one tail of the relative frequency histogram is substantially larger than the other so the histogram is asymmetric, or whether both tails of the relative frequency histogram are large so the data are more prone to extreme values than data from a Normal distribution.

WARNING: Sample skewness and sample kurtosis have no units.

The **sample skewness** is a measure of the (lack of) symmetry in the data. When the relative frequency histogram of the data is approximately symmetric then there is an approximately equal balance between the positive and negative values in the sum $\sum_{i=1}^{n}(y_i - \bar{y})^3$ and this results in a value for the sample skewness that is approximately *zero*.

The **sample kurtosis** measures the heaviness of the tails of the data relative to data that are Normally distributed. Since the term $\sum_{i=1}^{n}(y_i - \bar{y})^4$ is always positive, the kurtosis is always positive. If the sample kurtosis is greater than 3 then this indicates heavier tails 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.

Definition 1.7 The sample correlation, denoted by r, for data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is

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

where

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

Numerical summary for bivariate categorical data: Recall that values for a categorical variate are category names that do not necessarily have any ordering. If two variates of interest in a study are categorical variates then the sample correlation cannot be used as a measure of the relationship between the two variates.

Cutoff: Lecture 4, May 15 2023

Graphical Summaries

- A **histogram** is a graph in which a rectangle is constructed above each interval $I_1, I_2, \ldots, I_n$. The height of the rectangle for interval I_j is chosen so that the area of the rectangle is proportional to f_j, which is the number of values from $y_1, y_2, \ldots, y_n$ that are in I_j.
- A **boxplot** gives a graphical summary of the shape of a dataset's distribution in a very similar way to the five number summary.
- For categorical data, a **bar graph** or bar chart is a useful graphical summary. A bar graph has a bar for each of the possible values of the categorical variate with height equal to the frequency or relative frequency of that category. Usually the order of the di§erent possible categories is not important. The width of the bar is also not important. Gaps are left between the bars to emphasize that the data are categorical.
- **Run charts** are useful when examining the behaviour of data over time. Convention: time goes from left to right on x-axis. A run chart is another type of two dimensional plot which is used when we are interested in a graphical summary which illustrates how a single variate is changing over time.
- In **Scatter Plots**, anything with missing values in either variable can't be plotted, so we ignore it.

Sample correlation is an index number such that $-1 \leq r \leq 1$. It gives an indication of the strength of the linear relationship between x and y. The stronger the linear relationship, the closer the absolute value of r will be to 1. The sample correlation, r, has no units. The correlation between x and y is the same as the correlation between y and x.
Note:

- We sometimes distinguish between the response variate y and the explanatory variate x, which partially explains or determines the distribution of Y.

- A strong linear relationship does not necessarily imply a casual relationship, that is, $r \approx 1$ does not necessarily mean that x causes change in y.
- Similarly, $r \approx 0$ does not necessarily mean that x and y are unrelated, merely that they are uncorrelated.

Definition 1.8 For a data set of numerical values $\{y_1, \ldots, y_n\}$, the empirical cumulative distribution function or e.c.d.f. is defined by

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

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

Cutoff: Lecture 5, May 17 2023

2 STATISTICAL MODELS AND MAXIMUM LIKELIHOOD ESTIMATION

For Normal/Gaussian distribution, in G form, the second parameter is the standard deviation.
A note about notation: for Normal/Gaussian distribution, we can write $Y\ G(\mu, \sigma)$ or $Y \sim N(\mu, \sigma^2)$.
Additionally, μ is the true mean value, which we'll never know; $\hat{\mu}$ is our estimate, which we hope is close to μ.

Example 2.1 Suppose that the random variable $Y \sim G(\mu, \sigma)$ adequately models the height of a randomly chosen female in some population and that we are interested in estimating the unknown quantity (parameter) $E(Y) = \mu$.

Some important things to know:

- μ is not necessarily equal to the sample mean (in fact, it almost certainly isn't!)
- Different draws of the sample $y_1, y_2, \ldots, y_n$ will result in different values of the sample mean and therefore different estimates of μ

Three levels - Concepts and Notations:

Parametr μ is an UNKNOWN, fixed value. Think of it as a target.

Estimator $\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. It represents the function (rule or operation) that will allow us to obtain estimates, based on the data (observations).

Estimate $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. It represents an actual realized value of $\bar{Y}$, based on data (observations).

An (point) estimate of a parameter θ is the value of a function of the observed data y, denoted by $\hat{\theta} = \hat{\theta}(y)$.

Definition 2.1 A point estimate of a parameter is the value of a function of the observed data $y_1, y_2, \ldots, y_n$ and other known quantities such as the sample size n. We use $\hat{\theta}$ to denote an estimate of the parameter θ.

Cutoff: Lecture 6, May 19 2023

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

Likelihood Function for Discrete Distributions

Let the discrete (vector) random variable Y represent potential data that will be used to estimate θ, and let y represent the actual observed data that are obtained in a specific application.

Definition 2.2 The likelihood function for θ is defined as

$$L(\theta) = L(\theta; y) = P(Y = y; \theta) \text{ for } \theta \in \Omega$$

where the parameter space Ω is the set of possible values for θ.

Definition 2.3 The value of θ which maximizes $L(\theta)$ for given data 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 y. This value is denoted $\hat{\theta}$.

WARNING: The shape of the likelihood function and the value of θ at which it is maximized do not change if $L(\theta)$ is multiplied by a constant.

Definition 2.4 The relative likelihood function is defined as $R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$ Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$.

Sometimes it is easier to work with the $log(log = ln)$ of the likelihood function.

Definition 2.5 The log likelihood function is defined as $l(\theta) = lnL(\theta) = logL(\theta)$ for $\theta \in \Omega$.

A local max for $log(f(x))$ is also a local max for $f(x)$.

Because functions are often (but not always!) maximized by setting their derivatives equal to zero, we can usually obtain $\hat{\theta}$ by solving the equation

$$\frac{d}{d\theta} l(\theta) = 0$$

Likelihood function for a random sample

In many applications the data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent and identically distributed (i.i.d.) random variables each with probability function $f(y; \theta), \theta \in \Omega$. We refer to $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as a random sample from the distribution $f(y; \theta)$. In this case the observed data are $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and

$$L(\theta) = L(\mathbf{y}; \theta) = \prod f(y_i; \theta) \text{ for } \theta \in \Omega$$

Cutoff: Lecture 7, May 23 2023

A typical model has probability density function $f(y;\theta)$ if the variate Y is continuous, or probability function $f(y;\theta)$ if Y is discrete, where θ is (possibly) a vector of parameter values.

Expected frequency is calculated by $e_j = n \times \hat{p}(Y = j)$ where n is the sample size. Note that the sum of the expected values may not always equal the sample size, simply due to rounding.

Qqplot

The advantage of a qqplot is that the unknown parameters μ and σ do not need to be estimated.

The line in the qqplot is the line joining the lower and upper quartiles of the empirical and Gaussian distributions.

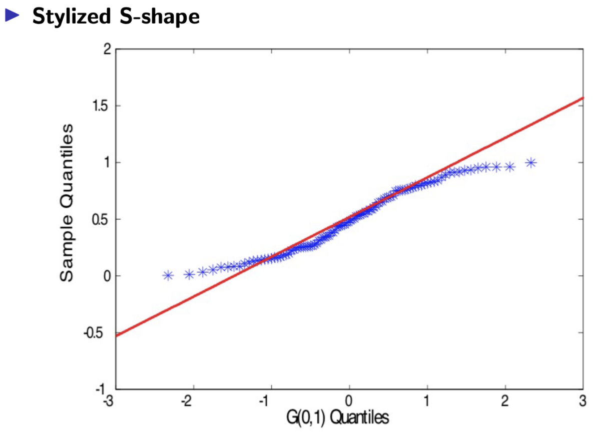
We do not expect the points to lie exactly along a straight line since the sample quantiles are based on the observed data which in general will be different every time the experiment is conducted. We only expect $Q\left(\frac{i}{n+1}\right)$ to be close in value to the sample quantile $q\left(\frac{i}{n+1}\right)$ for a reasonbly large dataset. As well the points at both ends of the line can be expected to lie further from the line since the quantiles of the Gaussian distribution change in value more rapidly in the tails of the distribution. In general if a dataset has a relative frequency histogram with a long right tail then the qqplot will exhibit this U-shape behaviour. If the data points form an S-shape. This is typical of data which are best modeled by a Uniform distribution. In general if a dataset has a relative frequency histogram which is quite symmetric and with short tails then the qqplot will exhibit this S-shape behaviour.

(Please see course notes Page 94 - 97 to get a overview of Summary of Model Checks for Named Distributions and Logarithmic Rules, Summation Notation and Product Notation) The line in the qqplot is the line joining the lower and upper quartiles of the empirical and Gaussian distributions, that is, the line joining $(Q_Z(0.25), q(0.25))$ and $(Q_Z(0.75), q(0.75))$.

Important: Do not expect a perfectly straight plot even with normal data!

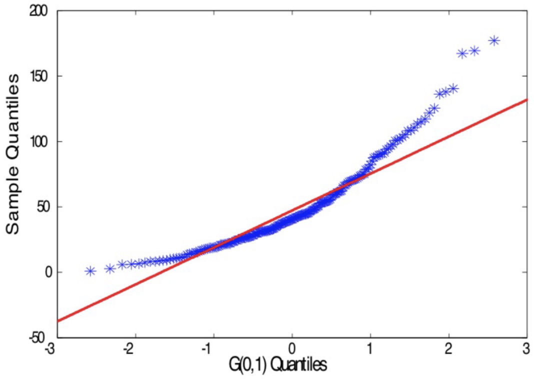
There are several patterns used to pick up specific characteristics of the observed data, such as:

Stylized S-shape Data with thin tails, best modeled by a Uniform distribution ...



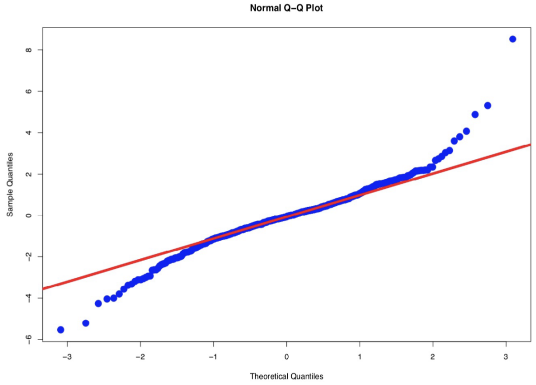
U-shape Positively/right skewed data, best modeled by an Exponential distribution ...

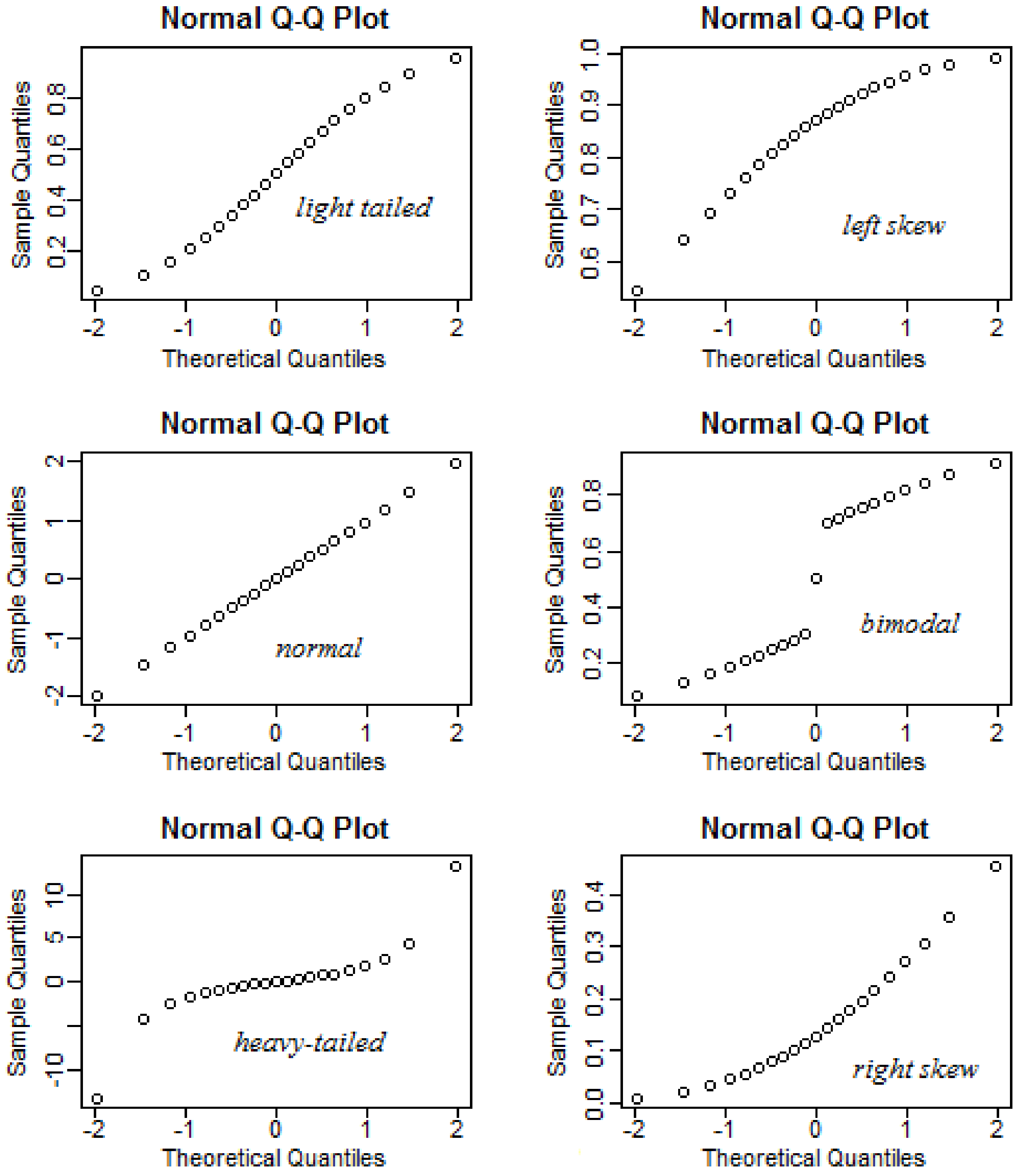
▶ **U-shape**



Chair shape Data with thicker tails, for instance, t-distribution ...

▶ **Chair shape**





Retrieved from https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot.

3 PLANNING AND CONDUCTING EMPIRICAL STUDIES

An empirical study is one which is carried out to learn about a population or process by collecting data. What is **PPDAC**? Here you go:

Problem a clear statement of the studyís objectives, usually involving one or more questions

Plan the procedures used to carry out the study including how the data will be collected

Data the physical collection of the data, as described in the Plan

Analysis the analysis of the data collected in light of the Problem and the Plan

Conclusion The conclusions that are drawn about the Problem and their limitations

If you see the phrase "evidence based decision" or "evidence based management", look for an empirical study.

The Steps of PPDAC

Problem The Problem step describes what the experimenters are trying to learn or what questions they want to answer. Often this can be done using questions starting with "what". There are 3 common types of statistical problems that are encountered are described below.

Descriptive The problem is to determine a particular attribute of a population or process. Much of the function of official statistical agencies such as Statistics Canada involves problems of this type. For example, the government needs to know the national unemployment rate and whether it has increased or decreased over the past month.

Causative The problem is to determine the existence or non-existence of a causal relationship between two variates.

Predictive The problem is to predict a future value for a variate of a unit to be selected from the process or population. This is often the case in finance or in economics. For example, financial institutions need to predict the price of a stock or interest rates in a week or a month because this effects the value of their investments.

In a causative problem, the experimenter is interested in whether one variate x tends to cause an increase or a decrease in another variate Y. Where possible this is conducted in a controlled experiment in which x is increased or decreased while holding everything else in the experiment constant and we observe the changes in Y. **An observational study in which the experimenter is not in control of the explanatory variates cannot usually be used to investigate a causative problem.**

Definition 3.1 The target population or target process is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

Recall following definition from Chapter 1:

Definition 3.2 A variate is a characteristic associated with each unit.

Definition 3.3 An attribute is a function of the variates over a population or process.

Plan The Plan step depends on the questions posed in the Problem step. The Plan step includes a description of the population or process of units from which units will be selected, what variates will be collected for the units selected, and how the variates will be measured.

In most cases, the attributes of interest for the target population/process cannot be estimated since only units from a subset of the target population/process can be considered for study or only units from another population completely can be considered for study.

Definition 3.4 The study population or study process is the collection of units available to be included in the study.

WARNING: The study population is often but not always a subset of the target population.

Definition 3.5 If the attributes in the study population/process differ from the attributes in the target population/process then the difference is called study error.

Study error cannot be quantified since the values of the target population/process attributes and the study population/process attributes, are unknown. (If these attributes were known then an empirical study would not be necessary!)

Definition 3.6 The sampling protocol is the procedure used to select a sample of units from the study population/process. The number of units sampled is called the sample size.

Definition 3.7 If the attributes in the sample differ from the attributes in the study population/process the difference is called sample error.

Sample error cannot be quantified since the values of the study population/process attributes are unknown. Different random sampling protocols can produce different sample errors. Sample error should be suspected in all surveys in which the participants are **volunteers**.

Definition 3.8 If the measured value and the true value of a variate are not identical the difference is called measurement error.

Measurement errors are unknown since the true value of the variate is unknown. (If we knew the true value we would not need to measure it!) Measurement error should always be suspected when variates are measured by self-reporting.

Data The goal of the Data step is to collect the data according to the Plan. Any deviations from the Plan should be noted. The data must be stored in a way that facilitates the Analysis.

Analysis The Analysis step includes both simple and complex calculations to process the data into information. A key component of the Analysis step is the selection of an appropriate model that describes the data and how the data were collected.

In the Problem step, the problems of interest were stated in terms of the attributes of interest. These attributes need to be described in terms of the parameters and properties of the model.

Conclusion The purpose of the Conclusion step is to address the questions posed in the Problem. The conclusions can only be made in relation to the study population. For example, it is not reasonable to make a conclusion about humans if the study population only consisted of laboratory animals. An attempt should also be made to quantify (or at least discuss) potential errors as described in the Plan step. Limitations to the conclusions should be discussed.

If the target population and study population are different, then we are only able to estimate the attributes of interest in the study population.

Important Note: The parameters in the model are always related to attributes of interest in the study population not in the sample.

4 ESTIMATION

RECALL CLT

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

Imagining an experiment may help you to understand sampling distributions:

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

Fortunately, you don't need to actually repeatedly sample a population to know the shape of the sampling distribution. The parameters of the sampling distribution of the mean are determined by the parameters of the population:

- The mean of the sampling distribution is the mean of the population.

$$\mu_{\bar{X}} = \mu$$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means
- $\sim$ means "follows the distribution"
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

Theorem 39 [39] ***Central Limit Theorem*** *If $X_1, X_2, \ldots, X_n$ are independent random variables all having the same distribution, with mean μ and variance σ^2, then as $n \to \infty$, the cumulative distribution function of the random variable*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

approaches the $N(0,1)$ cumulative distribution function. Similarly, the cumulative distribution function of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches the $N(0,1)$ cumulative distribution function.

Retrieved from STAT 230 Couse Notes.
A **rule of thumb** to apply CLT: sample size ≥ 30. **Remember**, if the parent population is Normal (Gaussian), then the distribution of the average will be normal as we have a linear combination of Normal random variables. Also, the CLT states that if the parent population is not Normal (Gaussian), the distribution of the average will still be approximately Normal (for large n!)

Definition 4.1 A (point) estimator $\tilde{\theta}$ is a random variable which is a function $\tilde{\theta} = g(Y_1, Y_2, Y_3, \ldots, Y_n)$ of the random variables $Y_1, Y_2, \ldots, Y_n$. The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.

The number of simulations k only affects how good the approximation is.

Definition 4.2 Suppose θ is scalar and that some observed data (say a random sample $y_1, y_2, \ldots, y_n$ have given a likelihood function $L(\theta)$. The relative likelihood function $R(\theta)$ is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \text{ for } \theta \in \Omega$$

where $\hat{\theta}$ is the maximum likelihood estimate and Ω is the parameter sapce.

Note:

$$0 \leq R(\theta) \leq 1 \text{ for all } \theta \in \Omega$$

Definition 4.3 A $100p\%$ likelihood interval for θ is the set $\{\theta : R(\theta) \geq p\}$.

The set $\{\theta : R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal.
Guidelines for Interpreting Likelihood Intervals

1. Values of θ inside a 50% likelihood interval are very plausible in light of the observed data.
2. Values of θ inside a 10% likelihood interval are plausible in light of the observed data.
3. Values of θ outside a 10% likelihood interval are implausible in light of the observed data.
4. Values of θ outside a 1% likelihood interval are very implausible in light of the observed data.

A drawback of likelihood intervals is that we never know whether the interval obtained contains the true value of the parameter or not.
Important: Sample size effect, i.e., likelihood intervals become narrower as the sample size increases. The dependence of the interval $\{\theta : R(\theta) \geq p\}$ on the data y is somewhat hidden but remember that $R(\theta) = R(\theta; y)$ is a function of the data y so the endpoints of the interval will depend on the data.

Definition 4.4 The log relative likelihood function is

$$r(\theta) = logR(\theta) = log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] = l(\theta) - l(\hat{\theta}) \text{ for } \theta \in \Omega$$

where $l(\theta) = logL(\theta)$ is the log likelihood function.

The log relative likelihood function $r(\theta)$ can also be used to compute a $100p\%$ likelihood interval since

$$R(\theta) \geq p \leftrightarrows r(\theta) \geq \log p$$

If the likelihood function $R(\theta)$ is unimodal then the log likelihood function $r(\theta)$ is also unimodal. Both $R(\theta)$ and $r(\theta)$ obtain a maximum value at $\theta = \hat{\theta}$. Note however that $R(\hat{\theta}) = 1$ while $r(\hat{\theta}) = 0$.
The log relative likelihood function can also be used to obtain a $100p\%$ likelihood interval since $R(\theta) \geq p$ if and only if $r(\theta) \geq log(p)$.
$C(\theta) = P(\theta \in [L(Y), U(Y)]) = P(L(Y) \leq \theta \leq U(Y))$ is called the coverage probability for the interval estimator, $[L(Y), U(Y)]$.
A $100p\%$ confidence interval for a parameter is an interval estimate $[L(Y), U(Y)]$ for which

$$P(\theta \in [L(Y), U(Y)]) = P(L(Y) \leq \theta \leq U(Y)) = p$$

The value p is called the confidence coefficient.

Definition 4.5 Suppose the interval estimator $[L(Y), U(Y)]$ has the property that

$$P(\theta \in [L(Y), U(Y)]) = P(L(Y) \leq \theta \leq U(Y)) = p$$

Suppose the interval estimate $[L(Y), U(Y)]$ is constructed for the parameter θ based on observed data y. The interval esitimate $[L(Y), U(Y)]$ is called $100p\%$ confidence interval for θ and p is called the confidence coefficient.

Interpretation

- The value θ is an unknown constant associated with the population. It is NOT a random variable and therefore does not have a distribution
- For an observed set of data y, $L(y)$ and $U(y)$ are all numerical values.
- It is not valid to say that the probability that θ lies in the interval $L(y), U(y)$ is equal to p since θ is a constant
- Now we usually only have one data set and one interval and for this one interval we do not know whether it contains the true value of θ or not. We can only say that we are 95% confident that our interval contains the true value of θ. In other words, we hope we were one of the lucky 95% who constructed an interval containing the true value of θ. From a larger context, suppose that we draw repeated independent random samples from the same population and each time we construct the interval $[L(y), U(y)]$ based on the observed data y, then we say that we should expect $100p\%$ of these constructed intervals to contain the true but unknown values of θ. We also expect that $(1 - 100p\%)$ of these constructed intervals will not contain the true but unknown values of θ.

Important: $P(\theta \in [L(y), U(y)]) = p$ is an incorrect statement. The parameter θ is a constant, not a random variable.
ChatGPT says: the correct interpretation of a 95% confidence interval is that if we were to take many, many samples from the population and calculate a 95% confidence interval from each one, we would expect about 95% of those intervals to contain the true population mean. In other words, the “95% confidence” refers to the method of calculation and its long-term performance if we were to repeat the sampling process many times, not to any single interval; the confidence interval is about where we estimate the population mean (which is a fixed but unknown value) to be.
Confidence intervals become narrower as the size of the sample on which they are based increases =¿ sample size effect.

Definition 4.6 A pivotal quantity $Q = g(Y; \theta)$ is a function of data Y and the unknown parameter θ such that the distribution of the random variable Q is fully known. That is, probability statements such as $P(Q \leq a)$ and $P(Q \geq b)$ depend on a and b but not θ or any other unknown information.

Hmmmmm so how do we construct a confidence interval using a pivotal quantity? A general idea is:

1. Determine numbers a and b such that

$$P[a \leq Q(Y; \theta) \leq b] = p$$

2. Re-express the inequality $a \leq Q(Y;\theta) \leq b$ in the form $L(Y) \leq \theta \leq U(Y)$, so that

$$p = P[a \leq Q(Y;\theta) \leq b] = P[L(Y) \leq \theta U(Y)]$$

3. For observed data y, the interval $[L(y), U(y)]$ is a $100p\%$ confidence interval for θ.

95% Confidence Interval for the Gaussian Mean with KNOWN Variance

$$[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}] = [L(y), U(y)]$$

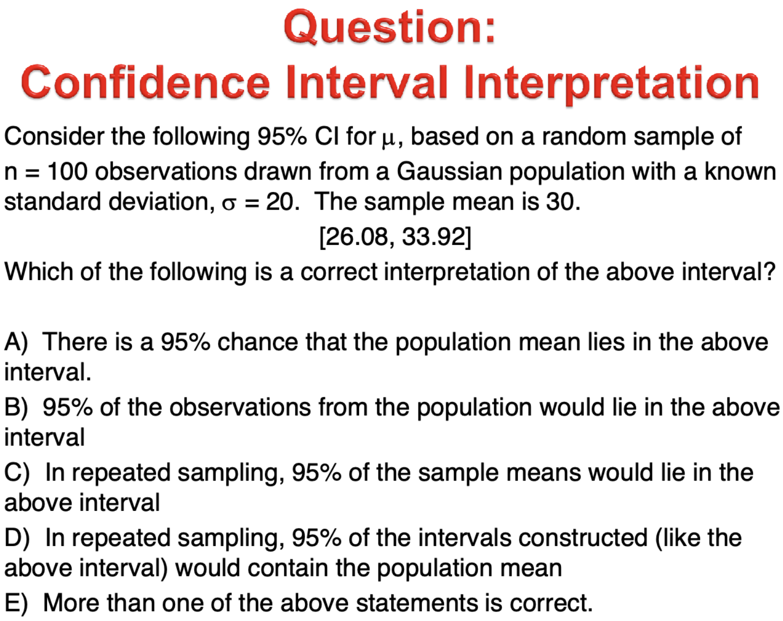
represents the 95% confidence interval for μ based on the data, $y = (y_1, y_2, \ldots, y_n)$.
Notice that the above interval is symmetric about the sample mean, $\bar{y}$. It is also the narrowest possible interval.

$$[\bar{Y} - c\frac{\sigma}{\sqrt{n}}, \bar{Y} + c\frac{\sigma}{\sqrt{n}}] = [L(y), U(y)]$$

represents a confidence interval for μ based on the data, $y = (y_1, y_2, \ldots, y_n)$
Point of interest

- The above interval is symmetric about the sample mean, $\bar{y}$
- It is also the narrowest possible interval
- The interval becomes larger / wider as the confidence level p increases
- The interval becomes smaller / narrower as the sample size n increases
- The interval becomes wider as σ increases

Two-sided Confidence Intervals
point estimate $\pm c*$standard deviation of the estimator, where $c*$standard deviation of the estimator is referred to as the margin of error / sampling error / sampling allowance.
Here comes to an exmaple question, the solution is D.



Question:
Confidence Interval Interpretation

Consider the following 95% CI for μ, based on a random sample of n = 100 observations drawn from a Gaussian population with a known standard deviation, σ = 20. The sample mean is 30.

[26.08, 33.92]

Which of the following is a correct interpretation of the above interval?

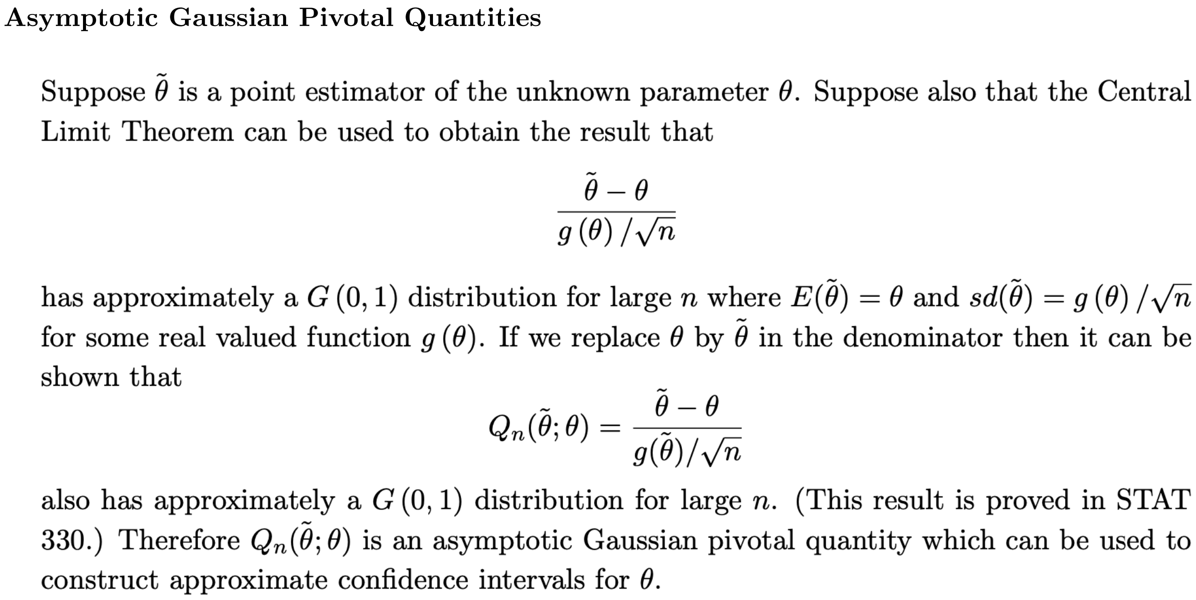
A) There is a 95% chance that the population mean lies in the above interval.

B) 95% of the observations from the population would lie in the above interval

C) In repeated sampling, 95% of the sample means would lie in the above interval

D) In repeated sampling, 95% of the intervals constructed (like the above interval) would contain the population mean

E) More than one of the above statements is correct.



Asymptotic Gaussian Pivotal Quantities

Suppose $\tilde{\theta}$ is a point estimator of the unknown parameter θ. Suppose also that the Central Limit Theorem can be used to obtain the result that

$$\frac{\tilde{\theta}-\theta}{g\left(\theta\right)/\sqrt{n}}$$

has approximately a $G\left(0,1\right)$ distribution for large n where $E(\tilde{\theta})=\theta$ and $sd(\tilde{\theta})=g\left(\theta\right)/\sqrt{n}$ for some real valued function $g\left(\theta\right)$. If we replace θ by $\tilde{\theta}$ in the denominator then it can be shown that

$$Q_n(\tilde{\theta};\theta)=\frac{\tilde{\theta}-\theta}{g(\tilde{\theta})/\sqrt{n}}$$

also has approximately a $G\left(0,1\right)$ distribution for large n. (This result is proved in STAT 330.) Therefore $Q_n(\tilde{\theta};\theta)$ is an asymptotic Gaussian pivotal quantity which can be used to construct approximate confidence intervals for θ.

which means that we can find random variables $Q_n = g(Y_1, Y_2, \ldots, Y_n; \theta)$ such that as $n \to \infty$, the distribution of Q_n ceases to depend on θ or other unknown information. Here, we say that Q_n is Asymptotically pivotal. In practice, we treat Q_n as a pivotal quantity for sufficiently large values of n. We call Q_n an approximate pivotal quantity. The above image retrieved from STAT 231 Course Notes.

To choose the sample size of an experiement using confidence interval(see course notes on page 160), you can only round up the integer, don't round down!

Cutoff: Lecture 13, June 5 2023

Remember that the sample size is under a square root. (e.g., it takes four times as much data to get twice as much precision)

Gamma function:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy \text{ for } \alpha > 0$$

Some properties of the Gamme function:

1. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
2. $\Gamma(\alpha) = (\alpha-1)!$, for $\alpha = 1, 2, 3, \ldots$
3. $\Gamma(0.5) = \sqrt{\pi}$

For **Chi-Squared distribution**, we write $X \sim \chi^2(k)$ and read as X as it has a Chi-squared distribution on k degrees of freedom (d.f.).

If $Z \sim N(0,1)$, then $W = Z^2 \sim \chi^2(1)$

If $X \sim \chi^2(2)$, which means $k = 2$, then $X \sim$ Exponential(2)

For $k > 2$, the probability density function is unimodal with maximum value at $x = k-2$. For values of $k \geq 30$, the probability density function resembles that of a $N(k, 2k)$ proabbility density function.

If $X \sim \chi^2(k)$, then

$$E(X) = k, Var(X) = 2k$$

Theorem 4.1 *Let $W_1, W_2, \ldots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$. Then $S = \sum_{i=1}^n W_i \sim \chi^2(\sum_{i=1}^n k_i)$*

Theorem 4.2 *If $Z \sim G(0,1)$, then the distribution of $W = Z^2$ is $\chi^2(1)$*

Corollary 4.1. *If $Z_1, Z_2, \ldots, Z_n$ are mutually independent $G(0,1)$ random variables and $S = \sum_{i=1}^n Z_i^2$, then $S \sim \chi^2(n)$*

Useful Results:

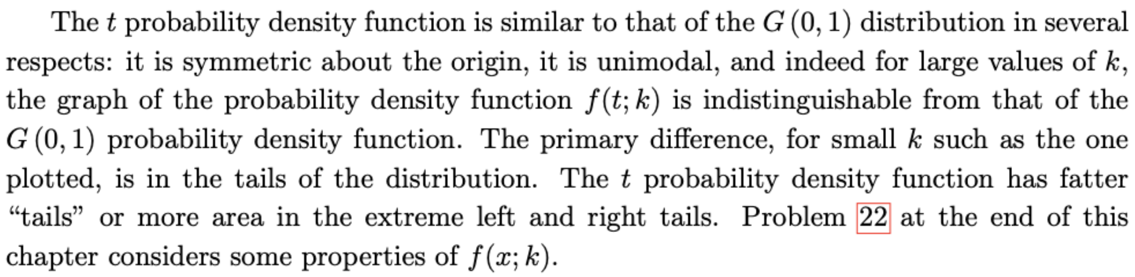
1. If $W \sim \chi^2(1)$ then $P(W \geq w) = 2[1 - P(Z \leq \sqrt{w})]$ where $Z \sim G(0,1)$
2. If $W \sim \chi^2(2)$ then $W \sim$ Exponential(2) and $P(W \geq w) = e^{-w/2}$

Student's t Distribution

Studentís t distribution (or more simply the t distribution) has probability density function

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

for $t \in \mathcal{R}$ and $k = 1, 2, 3, \ldots$.
The parameter k is called the degrees of freedom.



The t probability density function is similar to that of the $G(0,1)$ distribution in several respects: it is symmetric about the origin, it is unimodal, and indeed for large values of k, the graph of the probability density function $f(t;k)$ is indistinguishable from that of the $G(0,1)$ probability density function. The primary difference, for small k such as the one plotted, is in the tails of the distribution. The t probability density function has fatter "tails" or more area in the extreme left and right tails. Problem 22 at the end of this chapter considers some properties of $f(x;k)$.

Retrieved from STAT 231 Course Notes.
The middle 95% of t is bigger than that of the normal, depends on degrees of freedom(more d.f. is closer to normal).
It's bigger because it's reflecting the ADDITIOANL, UNCERTAINTY, for not knowing sigmas. Degree represents how much info(at least with the t distribution).

Theorem 4.3 *Suppose $Z \sim G(0,1)$ and $U \sim \chi^2(k)$ independently. Let*

$$T = \frac{Z}{\sqrt{U/k}}$$

Then T has a ***Student's t distribution with k degrees of freedom****.*

Cutoff: Lecture 14, June 7 2023

You should replace the lower bound of your confidence interval with 0 because you KNOW θ cannot be negative.

Definition 4.7 The random variable

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

where $\tilde{\theta}$ is the maximum likelihood estimator.

The random variable $\Lambda(\theta)$ is called the likelihood ratio statistic. The following theorem implies that $\Lambda(\theta)$ is an asymptotic pivotal quantity.

Theorem 4.4 *If $L(\theta)$ is based on $Y = (Y_1, \ldots, Y_n)$, a random variable $\Lambda(\theta)$ of size n, and if θ is the* *true value of the scalar parameter, then (under mild mathematical conditions) the distribution of $\Lambda(\theta)$ converge to a $\chi^2(1)$ distribution as $n \to \infty$.*

This theorem means that $L(\theta)$ can be used as a pivotal quantity for sufficiently large n in order to obtain approximate confidence intervals for θ. More importantly we can use this result to show that the likelihood intervals discussed in Section 4.3 are also approximate confidence intervals.

Theorem 4.5 *A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where*

$$q = 2P(Z \leq \sqrt{-2\log p}) - 1$$

and $Z \sim N(0,1)$.

Theorem 4.6 *If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval*

$$\{\theta : R(\theta) \geq e^{-a^2/2}\}$$

is an approximate $100p\%$ confidence interval.

Important: A smaller % for the likelihood interval $\leftrightarrows$ A higher % for the confidence interval.

Sometimes you may find that the approximate $100q\%$ confidence interval is quite different than the likelihood based approximate confidence interval and which also contains negative values for θ. Of course θ can only take on values between 0 and 1. This happens because the confidence interval is always symmetric about $\hat{\theta}$ and if $\hat{\theta}$ is close to 0 or 1 and n is not very large then the interval can contain values less than 0 or bigger than 1. The graph of the likelihood interval in Figure 4.8 is not symmetric about θ. In this case the 15% likelihood interval is a better summary of the values which are supported by the data.

More generally, if $\hat{\theta}$ is close to 0.5 or n is large then the likelihood interval will be fairly symmetric about $\hat{\theta}$ and there will be little difference in the two approximate confidence intervals. If $\hat{\theta}$ is close to 0 or 1 and n is not large then the likelihood interval will not be symmetric about $\hat{\theta}$ and the two approximate confidence intervals will not be similar.

Cutoff: Lecture 15, June 9 2023

We will estimate the population variance using the sample variance, and use a t-distribution quantity with d.f. $= n-1$ instead of a z-distribution.

Theorem 4.7 *Suppose $Y_1, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample mean $\bar{Y}$ and sample variance S^2. Then*

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Note: The random variable T is a pivotal quantity since it is a function of the data $Y_1, \ldots, Y_n$ and the unknown parameter μ whose distribution $t(n-1)$ is completely known.

Theorem 4.8 *Suppose $Y_1, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample variance S^2.*

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

Note: The random variable U is a pivotal quantity since it is a function of the data $Y_1, \ldots, Y_n$ and the unknown parameter σ^2 whose distribution $\chi^2(n-1)$ is completely known.
Note that the confidence interval for σ^2 is not symmetric about s^2, the estimate of σ^2. This happens of course because the $\chi^2(n-1)$ distribution is not a symmetric distribution.

Confidence intervals for parameters in the $G(\mu, \sigma)$ model

If we are interested in an upper bound on σ, we take $b = \infty$ and find an one-sided $100p\%$ confidence interval for σ to be

$$\left[0, s\sqrt{\frac{n-1}{a}}\right]$$

with $P(U \geq a) = p$ or $P(U \leq a) = 1 - p$.

Prediction interval for a future observation

For a confidence coefficient p, $100p\%$ prediction interval

$$\left[\bar{y} - as\sqrt{1 + \frac{1}{n}}, \bar{y} + as\sqrt{1 + \frac{1}{n}}\right]$$

is an interval of values for the future observation Y. Here the constant a should be chosen such that

$$P(-a \leq T \leq a) = p \text{ or } P(T \leq a) = \frac{1+p}{2}$$

where $T \sim t(n-1)$

Cutoff: Lecture 12, June 9 2023

---

5 HYPOTHESIS TESTING

Definition 5.1 A test statistic or discrepancy measure D is a function of the data Y that is constructed to measure the degree of “agreement” between the data Y and the null hypothesis H_0.

Definition 5.2 Suppose we use the test statistic $D = D(Y)$ to test the hypothesis H_0. Suppose also that $d = D(y)$ is the observed value of D. The p-value or observed significance level of the test of hypothesis H_0 using test statistic D is

$$p - value = P(D \geq d; H_0)$$

In other words, the $p - value$ is the probability, calculated assuming H_0 is true, of observing a value of the test statistic greater than or equal to the observed value of the test statistic. If d, the observed value of D, is large and consequently the $p - value$ is small then one of the following two statements is correct:

- H_0 is true but by chance we have observed an outcome that does not happen very often when H_0 is true
- H_0 is false

Note that :

- If the evidence against H_0 is statistically significant, the size of the p-value does not imply how "wrong" H_0 is.
- A confidence interval however does indicate the magnitude and direction of the departure from H_0.
- If strong evidence against H_0 is found in a particular direction then this might suggest conducting further experiments to investigate this evidence.

Remarks

(1) Note that the $p - value$ is defined as $P(D \geq d; H_0)$ and not $P(D = d; H_0)$ even though the event that has been observed is $D = d$. If D is a continuous random variable then $P(D = d; H_0)$ is always equal to zero which is not very useful. If D is a discrete random variable with many possible values then $P(D = d; H_0)$ will be small which is also not very useful. Therefore to determine how unusual the observed result is we compare it to all the other results which are as unusual or more unusual than what has been observed.

(2) The $p - value$ is NOT the probability that H_0 is true. This is a common misinterpretation.

The following table gives a rough guideline for interpreting $p - values$. *These are only guidelines for this course. The interpretation of* $p - values$ *must always be made in the context of a given study.*

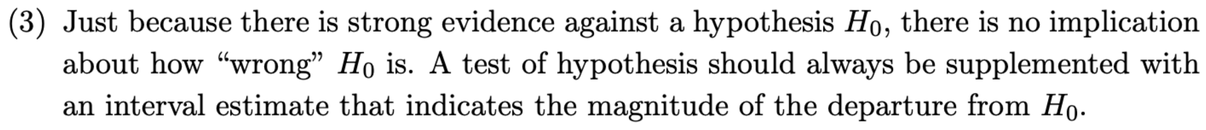
Table 5.1: Guidelines for interpreting $p - values$

$p - value$	Interpretation
$p - value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p - value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p - value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p - value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p - value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

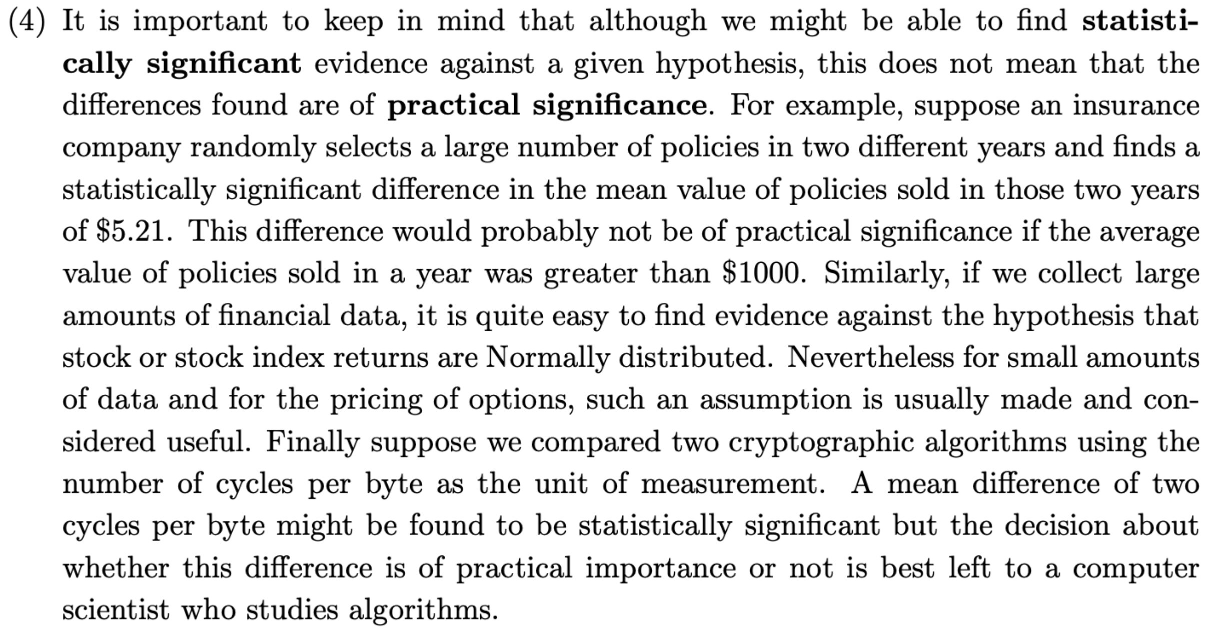
The approach to testing a hypothesis described above is very general and straightforward, but a few points should be stressed:

(1) If the $p-value$ is very small then we would conclude that there is **strong evidence against H_0 in light of the observed data**; this is often termed "statistically significant" evidence against H_0. We believe that statistical evidence is best measured when we interpret $p-values$ as in Table 5.1. However, it is still common in some areas of research to adopt a threshold $p-value$ such as 0.05 and **"reject H_0" whenever the p-value is below this threshold.** This may be necessary when there are only two possible decisions from which to choose. For example in a trial, a person is either convicted or acquitted of a crime. In the examples in these Course Notes we report the $p-value$ and use the guidelines in Table 5.1 to make a conclusion about whether there is evidence against H_0 or not. We emphasize the point that any decisions which are made after determining the $p-value$ for a given hypothesis H_0 must be made in the context of the empirical study.

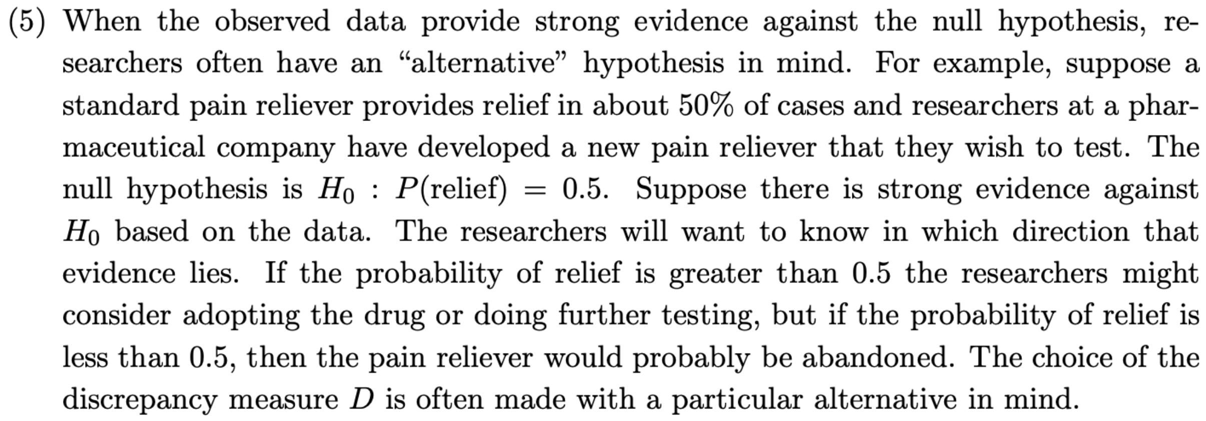
(2) If the $p-value$ is not small, we **do not conclude that H_0 is true**. We simply say there is **no evidence against H_0 in light of the observed data**. The reason for this "hedging" is that in most settings a hypothesis may never be strictly "true". For example, one might argue when testing $H_0 : \theta = 1/6$ in Example 5.1.2 that no real die ever has a probability of exactly $1/6$ for side 1. Hypotheses can be "disproved" (with a small degree of possible error) but not proved.



(3) Just because there is strong evidence against a hypothesis H_0, there is no implication about how "wrong" H_0 is. A test of hypothesis should always be supplemented with an interval estimate that indicates the magnitude of the departure from H_0.



(4) It is important to keep in mind that although we might be able to find **statistically significant** evidence against a given hypothesis, this does not mean that the differences found are of **practical significance**. For example, suppose an insurance company randomly selects a large number of policies in two different years and finds a statistically significant difference in the mean value of policies sold in those two years of \$5.21. This difference would probably not be of practical significance if the average value of policies sold in a year was greater than \$1000. Similarly, if we collect large amounts of financial data, it is quite easy to find evidence against the hypothesis that stock or stock index returns are Normally distributed. Nevertheless for small amounts of data and for the pricing of options, such an assumption is usually made and considered useful. Finally suppose we compared two cryptographic algorithms using the number of cycles per byte as the unit of measurement. A mean difference of two cycles per byte might be found to be statistically significant but the decision about whether this difference is of practical importance or not is best left to a computer scientist who studies algorithms.

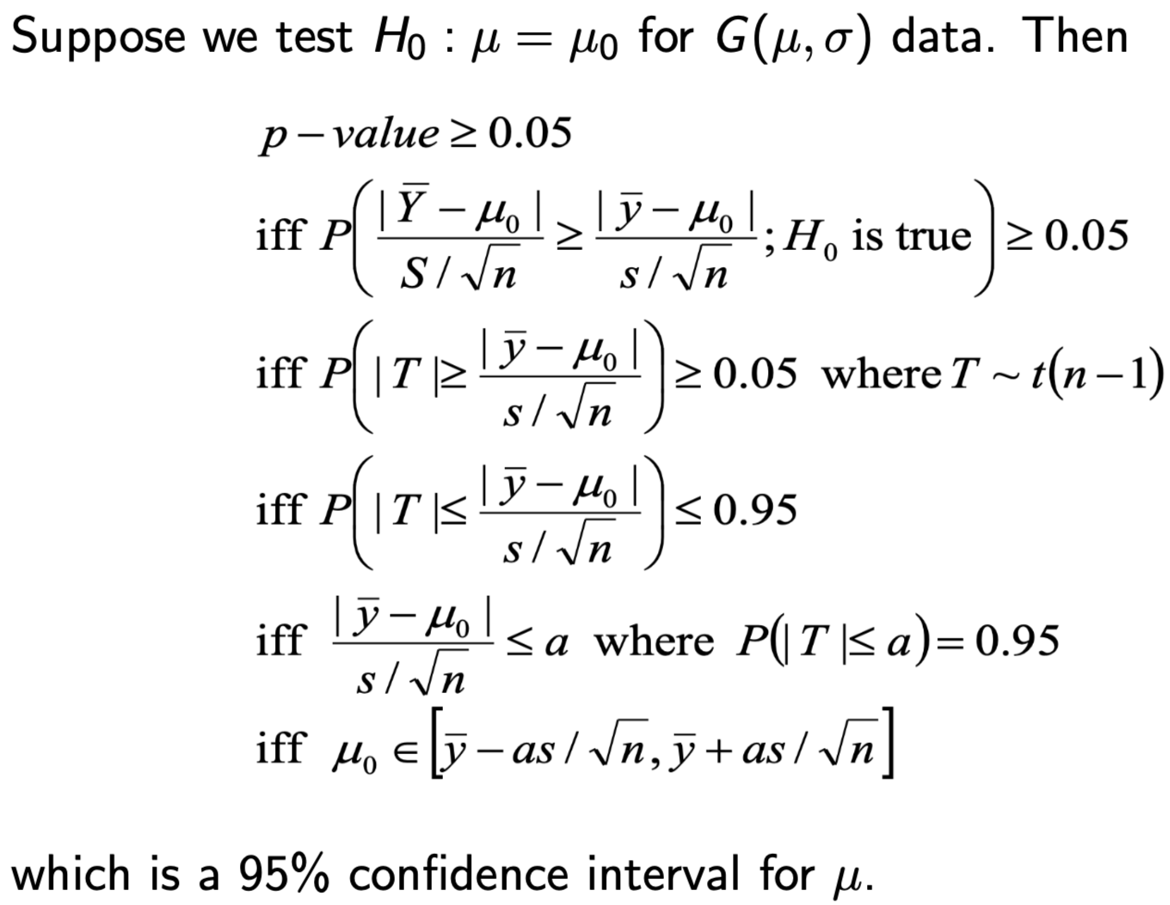


(5) When the observed data provide strong evidence against the null hypothesis, researchers often have an "alternative" hypothesis in mind. For example, suppose a standard pain reliever provides relief in about 50% of cases and researchers at a pharmaceutical company have developed a new pain reliever that they wish to test. The null hypothesis is $H_0 : P(\text{relief}) = 0.5$. Suppose there is strong evidence against H_0 based on the data. The researchers will want to know in which direction that evidence lies. If the probability of relief is greater than 0.5 the researchers might consider adopting the drug or doing further testing, but if the probability of relief is less than 0.5, then the pain reliever would probably be abandoned. The choice of the discrepancy measure D is often made with a particular alternative in mind.

Above 3 images retrieved from STAT 231 Course Notes.

Relationship Between Tests of Hypothesis and Confidence Intervals

Suppose we have data y and a model $f(y;\theta)$. Suppose also that we use the same pivotal quantity to construct the (approximate) confidence interval for θ and to test the hypothesis $H_0 : \theta = \theta_0$. Then the parameter value $\theta = \theta_0$ is an element of the $100q\%$ (approximate) confidence interval for θ if and only if the $p - value$ for testing $H_0 : \theta = \theta_0$ is greater than or equal to $1 - q$.



Suppose we test $H_0 : \mu = \mu_0$ for $G(\mu, \sigma)$ data. Then

$$\begin{aligned}
&p - value \geq 0.05 \\
&\text{iff } P\left(\frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}; H_0 \text{ is true} \right) \geq 0.05 \\
&\text{iff } P\left(|T| \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \right) \geq 0.05 \ \text{ where } T \sim t(n-1) \\
&\text{iff } P\left(|T| \leq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \right) \leq 0.95 \\
&\text{iff } \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \leq a \ \text{ where } P(|T| \leq a) = 0.95 \\
&\text{iff } \mu_0 \in \left[\bar{y} - as/\sqrt{n}, \bar{y} + as/\sqrt{n} \right]
\end{aligned}$$

which is a 95% confidence interval for μ.

Test hypothesis for σ or σ^2

We use the test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

where $U \sim \chi^2(n-1)$. We calculatet he p-value using the following procedure in order to obtain p-value:

1. Calculate the observed value
$$u = (n-1)s^2/\sigma_0^2$$
of the test statistic.
2. If $P(U \leq u) > \frac{1}{2}$, compute p-value $= 2P(U \geq u)$.
3. If $P(U \leq u) < \frac{1}{2}$, compute p-value $= 2P(U \leq u)$.

Note: only one of the two values $2P(U \geq u)$ and $2P(U \leq u)$ will be less than one and this one is the desired p-value.

Likelihood Ratio Test of Hypothesis - One Parameter

Consider

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

We choose this particular function because if $H_0 : \theta = \theta_0$ is true, then $\Lambda(\theta_0)$ has approximately $\chi^2(1)$ distribution.
To determine the p-value we first calculate the observed value of $\Lambda(\theta_0)$, denoted by $\lambda(\theta_0)$ and given by

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

where $R(\theta_0)$ is the relative likelihood function evaluated at $\theta = \theta_0$. The approximate p-value is then

$$\begin{aligned} p - value &\approx P[W \geq \lambda(\theta_0)] \text{ where } W \sim \chi^2(1) \\ &= P(|Z| \geq \sqrt{\lambda(\theta_0)}) \text{ where } Z \sim G(0,1) \\ &= 2[1 - P(Z \leq \sqrt{\lambda(\theta_0)})] \end{aligned}$$

Note that small values of $R(\theta_0)$ correspond to large observed values of $\Lambda\left(\theta_0\right)$ and therefore large observed value of $\Lambda\left(\theta_0\right)$ indicate evidence against the hypothesis $H_0 : \theta = \theta_0$. We illustrate this in Figure 5.2. Notice that the more plausible values of the parameter θ correspond to larger values of $R(\theta)$ or equivalently, in the bottom panel, to small values of $\Lambda(\theta) = -2 \log \left[R(\theta)\right]$. The particular value displayed θ_0 is around 0.3 and it appears that $\Lambda\left(\theta_0\right) = -2 \log \left[R(\theta_0)\right]$ is quite large, in this case around 9. To know whether this is too large to be consistent with H_0, we need to compute the $p - value$.

Cutoff: Lecture 19, June 19 2023

6 GAUSSIAN RESPONSE MODELS

Definition 6.1 A Gaussian response model is one for which the distribution of the response variate Y, given the associated vector of covariates $x = (x_1, \ldots, x_k)$ for an individual unit, is of the form

$$Y \sim G(\mu(x), \sigma(x))$$

If observations are made on n randomly selected units we write the model as

$$Y_i \sim G(\mu(x_i), \sigma(x_i))$$

for $i = 1, 2, \ldots, n$ independently.

Sometimes the model is written as

$$Y_i = \mu(x_i) + R_i$$

where $R_I \sim G(0, \sigma)$

Simple Linear Regression

Consider the model with independent Y_i's such that

$$Y_i \sim G(\mu(x_i), \sigma)$$

where

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

We have maximum likelihood estimates

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_x x}\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}(S_{yy} - \hat{\beta}S_{xy})\end{aligned}$$

In summary we have that the least squares estimates and the maximum likelihood estimates obtained assuming the model are the same estimates.

Note that $y = \hat{\alpha} + \hat{\beta}x$ is often called the fitted regression line for y on x or more simply the fitted line.

Distribution of the estimator $\tilde{\beta}$

$$\begin{aligned}E(\tilde{\beta}) &= \beta\\ Var(\tilde{\beta}) &= \frac{\sigma^2}{S_{xx}}\end{aligned}$$

In summary,

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

Confidence interval for β

Although the maximum likelihood estimate σ^2 is

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

We will esitimate σ^2 using

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

since $E(S_e^2) = \sigma^2$ where

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

The pivotal quantity

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

can be used to obtain confidence interval for β and to construct tests of hypothesis about β.

Using `R` we find the constant a such that $P(-a \leq T \leq a) - p$ where $T \sim t(n-2)$.

$$\begin{aligned} P &= P(-a \leq T \leq a) \\ &= P(-a \leq \frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \leq a) \\ &= P(\tilde{\beta} - aS_e/\sqrt{S_{xx}} \leq \beta \leq \tilde{\beta} + aS_e/\sqrt{S_{xx}}) \end{aligned}$$

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

$$[\tilde{\beta} - aS_e/\sqrt{S_{xx}}, \tilde{\beta} + aS_e/\sqrt{S_{xx}}] = \tilde{\beta} \pm aS_e/\sqrt{S_{xx}}$$

To test the hypothesis $H_0 : \beta = \beta_0$ the p-value is given by

$$\begin{aligned} p - value &= P\left(|T| \geq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{S_{xx}}}\right) \\ &= 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{S_{xx}}}\right)\right] \end{aligned}$$

where $T \sim t(n-2)$. Note that $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$ can be used to obtain confidence intevrals for σ^2 or σ. A $100p\%$ confidence interval for σ^2 is

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

and a $100p\%$ confidence interval for σ is

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

where

$$P(U \leq a) = P(U > b) = \frac{1-p}{2} \text{ and } U \sim \chi^2(n-2)$$

Note that

$$P(U \leq b) = \frac{1+p}{2}$$

Confidence interval for the mean response $\mu(x) = \alpha + \beta x$

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

In summary,

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

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

$$\left[\hat{\mu}(x) - as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \hat{\mu}(x) + as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right]$$

where $p = P(-a \leq T \leq a)$ and $T \sim t(n-2)$ and $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$

Since $\alpha = \mu(0)$, a $100p\%$ confidence interval for α is given by $\left[\hat{\mu}(x) - as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \hat{\mu}(x) + as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right]$ with $x = 0$ which gives

$$\left[\hat{\alpha} - as_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}, \hat{\alpha} + as_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}\right]$$

Prediction Interval for Future Response

$$E[Y - \tilde{\mu}(x)] = 0$$
$$Var[Y - \tilde{\mu}(x)] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]$$

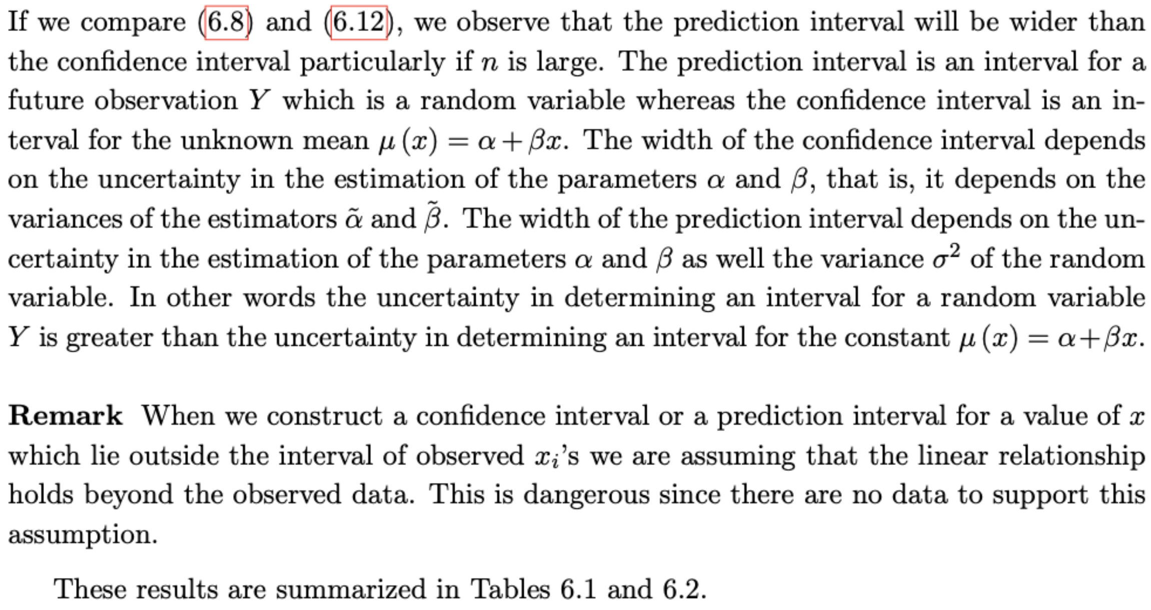
In summary,

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

A $100p\%$ prediction interval is

$$\left[\hat{\mu}(x) - as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \hat{\mu}(x) + as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right]$$

where $p = P(-a \leq T \leq a)$ and $T \sim t(n-2)$



If we compare (6.8) and (6.12), we observe that the prediction interval will be wider than the confidence interval particularly if n is large. The prediction interval is an interval for a future observation Y which is a random variable whereas the confidence interval is an interval for the unknown mean $\mu(x) = \alpha + \beta x$. The width of the confidence interval depends on the uncertainty in the estimation of the parameters α and β, that is, it depends on the variances of the estimators $\tilde{\alpha}$ and $\tilde{\beta}$. The width of the prediction interval depends on the uncertainty in the estimation of the parameters α and β as well the variance σ^2 of the random variable. In other words the uncertainty in determining an interval for a random variable Y is greater than the uncertainty in determining an interval for the constant $\mu(x) = \alpha + \beta x$.

Remark When we construct a confidence interval or a prediction interval for a value of x which lie outside the interval of observed x_i's we are assuming that the linear relationship holds beyond the observed data. This is dangerous since there are no data to support this assumption.

These results are summarized in Tables 6.1 and 6.2.

Retrieved from STAT 231 Course Notes.

Comparison of Two Population Means

An estimate of the varianc e σ^2 called the *pooled estimate of variance* is

$$\begin{aligned} s_p^2 &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2\right] \\ &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \end{aligned}$$

Theorem 6.1 *If $Y_{11},\ldots,Y_{1n_1}$ is a random sample from a $G(\mu_1,\sigma)$ distribution and independently $Y_{21},\ldots,Y_{2n_2}$ is a random sample from a $G(\mu_2,\sigma)$ distribution 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

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji}-\bar{Y}_j)^2 \sim \chi^2(n_1+n_2-2)$$

Table 6.3
Confidence Intervals for
Two Sample Gaussian Model

Model	Parameter	Pivotal Quantity	100p% Confidence Interval
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ σ_1, σ_2 known	$\mu_1-\mu_2$	$\frac{\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ $\sim G(0,1)$	$\bar{y}_1-\bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1=\sigma_2=\sigma$ σ unknown	$\mu_1-\mu_2$	$\frac{\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $\sim t(n_1+n_2-2)$	$\bar{y}_1-\bar{y}_2 \pm bs_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ μ_1, μ_2 unknown	σ^2	$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ $\sim \chi^2(n_1+n_2-2)$	$\left[\frac{(n_1+n_2-2)s_p^2}{d}, \frac{(n_1+n_2-2)s_p^2}{c}\right]$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1\neq\sigma_2$ σ_1, σ_2 unknown	$\mu_1-\mu_2$	asymptotic Gaussian pivotal quantity $\frac{\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$ for large n_1, n_2	approximate 100p% confidence interval $\bar{y}_1-\bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

Notes:
The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n_1+n_2-2)$.
The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n_1+n_2-2)$.

Table 6.4
Hypothesis Tests for
Two Sample Gaussian Model

Model	Hypothesis	Test Statistic	$p - value$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ σ_1, σ_2 known	$H_0 : \mu_1 = \mu_2$	$\frac{\lvert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$2P\left(Z \geq \frac{\lvert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$ $Z \sim G(0, 1)$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ σ unknown	$H_0 : \mu_1 = \mu_2$	$\frac{\lvert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\rvert}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$2P\left(T \geq \frac{\lvert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\rvert}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)$ $T \sim t(n_1 + n_2 - 2)$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ μ_1, μ_2 unknown	$H_0 : \sigma = \sigma_0$	$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma_0^2}$	$\min(2P\left(W \leq \frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right),$ $2P\left(W \geq \frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right))$ $W \sim \chi^2(n_1 + n_2 - 2)$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 \neq \sigma_2$ σ_1, σ_2 unknown	$H_0 : \mu_1 = \mu_2$	$\frac{\lvert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\rvert}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	approximate $p - value$ $2P\left(Z \geq \frac{\lvert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right)$ $Z \sim G(0, 1)$

Retrieved from STAT 231 Course Notes.

7 MULTINOMIAL MODELS AND GOODNESS OF FIT TESTS

Likelihood Ratio Test for the Multinomial Model
For Multinomial model, we can rewrite Λ as

$$\Lambda = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$

where E_j is expected frequency and $E_j = n\theta_j(\tilde{\alpha})$ for $j = 1, 2, \ldots, k$. The observed value of Λ is

$$\lambda = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right)$$

where $e_j = n\theta_j(\hat{\alpha}), j = 1, \ldots, k$. Note that the value of λ will be close to 0 if the observed values $y_1, \ldots, y_k$ are close to the expected values $e_1, \ldots, e_k$ and that the value of λ will be large if the y_j's and e_j's differ greatly.
If n is large and H_0 is true then the distribution of Λ is approximately $\chi^2(k-1-p)$. This enables us to compute p-value from observed data by using approximation

$$p - value = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \Lambda)$$

where $W \sim \chi^2(k-1-p)$. p is equal to the number of parameters that need to be estimated in the model assuming the null hypothesis. This approximation is accurate when n is large and none of the θ_j's is too small. In particular, the expected frequencies determined assuming H_0 is true should all be at least 5 to use the Chi-squared approximation.
An alternative test statistic that was developed historiacally before the likelihood ratio test statistic is the Pearson goodness of fit statistic

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

with observed value

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$$

The Pearson goodness of fit statistic has similar properties to Λ, that is, d takes on small values if the y_j's and e_j's are close in value and d takes on large values if the y_j's and e_j's differ greatly. It also turns out that, like Λ, the test statistic D has a limiting $\chi^2(k-1-p)$ distribution when H_0 is true.
Goodness of Fit Tests
The expected frequency is $e_j = np_j$ where p_j is the probability density function.
As indicated there we did not know how close the observed and expected frequencies needed to be to conclude that the model was adequate. It is possible to test the appropriateness of a model by using the Multinomial model.
A goodness of fitt test has some arbitrary elements, since we could have used different intervals and a different number of intervals. Theory has been developed on how best to choose the intervals. For this course we only give rough guidelines which are: chose $4-10$intervals, so that the observed expected frequencies under H_0

are at least 5.

Two-Way (Contingency) Tables

Often we want to assess whether two factors or variates appear to be related. One tool for doing this is to test the hypothesis that the factors are independent and thus statistically unrelated. We will consider this in the case where both variates are discrete, and take on a fairly small number of possible values. This turns out to cover a great many important settings.

Two types of studies give rise to data that can be used to test independence, and in both cases the data can be arranged as frequencies in a two-way table. These tables are also called contingency tables.

8 FINAL REVIEW SUMMARY

In Simple Linear Regression, we have

$$E(\tilde{\beta}) = \beta, Var(\tilde{\beta}) = \frac{\sigma^2}{S_{xx}}$$

$$S_e^2 = \frac{S_{yy} - \hat{\beta}S_{xy}}{n-2}$$

In Comparison of Two Population Means, we have the pooled estimate of variance is

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

where

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2}{n_1-1}, s_2^2 = \frac{\sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2}{n_2-1}$$