STAT 231 Statistics

University of Waterloo - Fall 2024

12th December 2024

Instructor: Michael Wallace

LaTeX : Xing Liu

Contents

1 Introduction to Statistical Sciences

1.1 Empirical Studies and Statistical Sciences

Lecture 1

Statistical Science is concerned with all aspects of **empirical studies**.

Definition 1.1 (**Empirical Study**). An **empirical study** is one in which we learn by observation or experimentation.

Empirical study involves **uncertainty**.

Definition 1.2 (**Unit**). A **unit** is an individual person, place, or thing about which we can take some measurement(s).

Definition 1.3 (**Population**). A **population** is a collection of units.

Remark. It is essential to be precise with all statistical terminology. Lack of precision in definitions/terminology is a very common mistake in exams!

Definition 1.4 (**Process**). A **process** is also a collection of units, but those units are 'produced' over time.

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

Compare:

- **Population:** All current UW undergraduate students.
- **Process:** All UW undergraduate students for the next 10 years.

1.2 Data Collection

Definition 1.5 (**Variate**)**.** A **variate** is a characteristic of a unit.

Types of variates:

- Continuous: can be measured to an infinite degree of accuracy in theory (height, weight, etc.).
- Discrete: can only take a finite or countably infinite number of values (# of car accidents, # of cups of tea consumed, etc.).
- Categorical: units fall into a (non-numeric) category (hair color, university program, etc.).
- Ordinal: an ordering is implied, but not necessarily through a numeric measure (strongly disagree, disagree, neutral in surveys, etc.).
- Complex: more unusual, and include open-ended responses to a survey question, or an image.

Lecture 2

Definition 1.6 (**Attribute**)**.** An **attribute** of a population or process is a function of the variates, which is defined for all units in the population or process.

Example. In the STAT 231 assignments example, we might be interested in knowing:

- The modal number of completed assignments.
- The proportion of assignments submitted in the final 24 hours.

Broad types of empirical study:

- Sample surveys.
- Observational studies.
- Experimental studies.

Definition 1.7 (**Sample Survey**)**.** An **sample survey** is where information is obtained about a finite population by selecting a 'representative' sample of units from that population and determining the variates of interest for each unit in the sample.

Example. A poll to predict who will win an election. A polling company will select a number of people at random and ask them questions about their voting preferences.

Example. Student course perception surveys. A (non-random) sample of students complete these surveys to help an instructor learn about their course.

Definition 1.8 (**Observational Study**). An **observational study** is where information about a population or process is collected without any attempt to change one or more variates for the sampled units.

Definition 1.9 (**Experimental Study**). An **experimental study** is one in which the experimenter intervenes and changes or sets the values of one or more variates for the units in the study.

1.3 Data Summaries

1.3.1 Numerical Summaries

Measures of location

Sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Definition 1.10 (**Order Statistic**). Let $\{y_{(1)}, y_{(2)}, \dots, y_{(n)}\}$ where $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ be the **order statistic** for the data set $\{y_1, y_2, \dots, y_n\}$.

For odd number of observations:

$$\text{sample median} = \hat{m} = y_{\left(\frac{n+1}{2}\right)}.$$

For even number of observations:

$$\text{sample median} = \hat{m} = \frac{1}{2}\left[y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}\right]$$

Sample mode: the most frequent value in a set of data.

Measures of dispersion or variability

- Sample variance:

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

  and the sample standard deviation is $s = \sqrt{s^2}$.

  Example. Suppose we have a sample of data from a Gaussian distribution. Then,
  - Approximately 68% of the sample lie in $[\bar{y} - s, \bar{y} + s]$.
  - Approximately 95% of the sample lie in $[\bar{y} - 2s, \bar{y} + 2s]$.

- Range:

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

  where $y_{(n)} = \max(y_1, y_2, \dots, y_n)$ and $y_{(1)} = \min(y_1, y_2, \dots, y_n)$.

- IQR:

  Definition 1.11 (Interquartile Range). $\text{IQR} = q(0.75) - q(0.25)$.

Definition 1.12 (Quantiles). The p^{th} **quantile** (or $100p^{\text{th}}$ percentile) is the value such that a fraction p of the data is at or below that value. One way to define it is the value, denoted $q(p)$ with $0 < p < 1$, as follows:

- Let $m = (n+1)p$, where n is the sample size.
- If m is an integer, and $1 \leq m \leq n$, then take the m^{th} smallest value $q(p) = y_{(m)}$.
- If m is not an integer, but $1 < m < n$, then determin the closest integer j such that $j < m < j+1$ and take $q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right]$.

Note. The 0.5^{th} quantile (or 50^{th} percentile) is the median. Also, the quantiles $q(0.25)$ and $q(0.75)$ are called the lower or first quartile, and the upper or third quartile, respectively.

Measure of shape

- Sample skewness:

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

  Note. The denominator is equal to the sample standard deviation cubed, except we replace the $n - 1$ in the denominator with n.

  - If data are symmetric, then the skewness is close to 0.
  - If skewness is positive, then the data are right-skewed.
  - If skewness is negative, then the data are left-skewed.

  Note. Skewness is negative if the numerator is negative in above equation.

- Sample kurtosis:

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

  Note. The denominator is equal to the sample standard deviation to the power of 4, except we replace the $n - 1$ in the denominator with n.

  - Data that look Gaussian have a kurtosis close to 3.
  - Data with large tails (more extreme values) have a sample kurtosis greater than 3.
  - Data with shorter tails (more points near the mean) have a sample kurtosis less than 3.
  - Data that look uniform have a sample kortosis close to 1.8.

Lecture 3

A very important principle

We never 'prove' an assumption is true. Instead, we see if we can find evidence against an assumption. Never use definitive statements usch as "the assumption is true/false".
Words such as 'reasonably' and 'approximately' are your friends!

Definition 1.13 (**Five Number Summary**). The **five number summary** of a data set consists of: $y_{(1)}$, $q(0.25)$, $q(0.5)$, $q(0.75)$, and $y_{(n)}$.

1.3.2 Graphical Summaries

Some examples we'll consider:

- Histograms.
- Empirical cumulative distribution functions (e.c.d.f.).
- Boxplots.
- Run charts.
- Scatterplots.
- Bar charts, pie charts.

Histograms

The idea is to create a graphical summary of our data that we can use to compare with a p.d.f. for a continuous random variable, or a p.f. for a discrete variable.

Constructing histograms:

1. Our observed data are denoted by $\{y_1, \dots, y_n\}$.
2. Partition the range of y into k non-overlapping intervals $I_j = [a_{j-1}, a_j)$ for $j = 1, 2, \dots, k$.
3. Let $f_j =$ the number of values from $\{y_1, \dots, y_n\}$ that are in I_j. The f_j are called the observed frequencies.
4. Draw a rectangle above each of the intervals with height proportional to the observed frequency or relative frequency.

In a **standard histogram**, the intervals are of equal width and the heights are equal to the frequencies (if the intervals are of width 1) or the relative frequencies (if the intervals are not of width 1). The sum of the areas of the bars will equal the sample size n.

In a **relative frequency histogram**, the height of the rectangle is chosen so that the area of the rectangle equals $\frac{f_j}{n}$, that is

$$\text{height} = \frac{\frac{f_j}{n}}{a_j - a_{j-1}}.$$

In this case, the sum of the areas of the rectangles equals 1.

Histograms allow us to compare the distribution of a dataset with a p.d.f. (or p.f.)!

Empirical c.d.f.

An empirical c.d.f., in contrast, lets us compare the distribution of a dataset with a c.d.f. of a random variable.

Definition 1.14 (**Empirical C.D.F.**).
For a data set $\{y_1, \dots, y_n\}$, the **empirical c.d.f.** is defined by

$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, \dots, y_n\} \text{ which are} \leq y}{n} \quad \text{for all } y \in \mathbb{R}.$$

The empirical c.d.f. is an estimate, based on the data of the population c.d.f.

Histograms and empirical c.d.f.s give us a graphical way to check the fit of distributions. The principle is to compare what we'd expect to see if our data were from a normal (or some other) distribution, with what we actually observe.

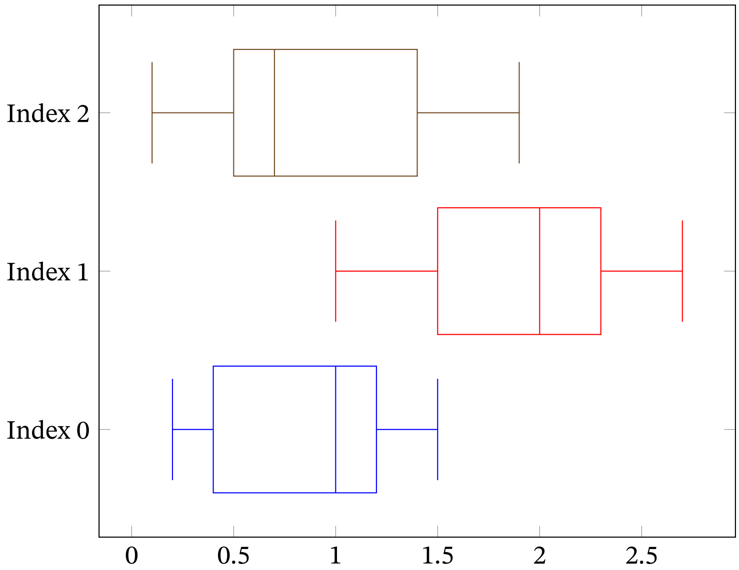
Boxplots

A boxplot gives a graphical summary of the shape if a dataset's distribution in a very similar way to the five number summary.

General approach to drawing a boxplot:

1. Draw a box with ends at $q(0.25)$ and $q(0.75)$ so the box height = IQR.
2. Draw a line in the box at $q(0.5)$ = median.
3. Draw short horizontal lines at the smallest observation that is larger than $q(0.25) - 1.5 \times \text{IQR}$ and at the largest observation that is smaller than $q(0.75) + 1.5 \times \text{IQR}$.
4. Draw two lines or 'whiskers' extending up and down from the box to the lines in step 3.
5. Plot any additional poitns beyond these lines individually using a special symbol like '$+$' or '$*$'. These points are called **outliers**.

Boxplots look like this (should be vertical by the way):



- It can help to visualize a boxplot alongside a histogram of the same data.
- Boxplots can also help demonstrate skewness.
- Boxplots can also be used to compare the values of variates in two or more groups.

Run Chart

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

Scatterplots

So far we've only considered univariate datasets. We often have bivariate data, of the form $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i and y_i are real numbers both observed on unit i. A scatterplot is simply the plot of the points (x_i, y_i), $i = 1, \dots, n$.

Correlation

Recall that for random variables X and Y with expectations μ_X and μ_Y and standard deviations σ_X and σ_Y, the **correlation** is defined as

$$\rho = \mathrm{corr}(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\,[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$$

The **sample correlation** gives us a numerical summary of a bivariate dataset.

Definition 1.15 (**Sample Correlation**).
For data $\{(x_1, y_1), \dots, (x_n, y_n)\}$, the **sample correlation** is defined as

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

where

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2, \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2, \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \end{aligned}$$

The sample correlation takes values between -1 and 1. It is a measure of the **linear** relationship between x and y.

- If the valur of r is close to 1, we say there is a strong positive linear relationship between the two variates.
- If r is close to -1, we say there is a strong negative linear relationship.
- If r is close to 0, we say there is no linear relationship.

Very important points:

- A strong linear relationship is not necessarily a **causal relationship**, that is, just because $r \approx 1$ does not mean that x causes changes in y.

  (Correlation does not necessarily imply causation!)

- Just because $r \approx 0$ does not mean that x and y are unrelated, that is, it is possible to construct examples where x and y have a strong functional relationship, but where $r = 0$.

  (No correlation does not necessarily imply no causation!)

Definition 1.16 (**Relative Risk**). For bivariate categorical data in the form of a table below, the relative risk of event A in group B as compared to group $\overline{B}$ is

$$\text{relative risk} = \frac{\frac{y_{11}}{(y_{11}+y_{12})}}{\frac{y_{21}}{(y_{21}+y_{22})}}.$$

The table below summarizes the relationship between two categorical variates A and B.

	A	$\bar{A}$	Total
B	y_{11}	y_{12}	$y_{11} + y_{12}$
$\bar{B}$	y_{21}	y_{22}	$y_{21} + y_{22}$
Total	$y_{11} + y_{21}$	$y_{12} + y_{22}$	n

Bar Charts and Pie Charts

Pie charts are bad! A sis-by-side bar chart is much better.

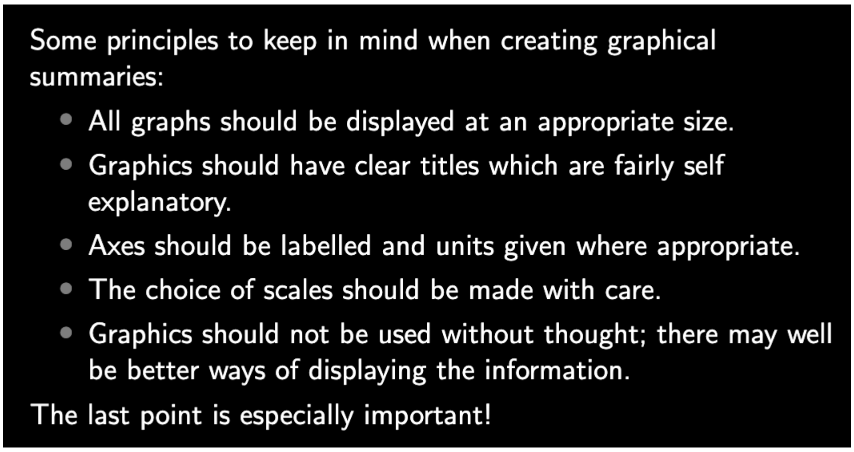


Figure 1: Principles to keep in mind when creating graphical summaries.

1.4 Probability Distributions and Statistical Models

Skip for now?

Lecture 4

1.5 Data Analysis and Statistical Inference

Two broad aspects of the analysis and interpretation of data:

- Descriptive Statistics
- Statistical Inference

Definition 1.17 (**Descriptive Statistics**). **Descriptive statistics** are portrays of the data, or parts of the data, in numerical and graphical ways to show features of interest.

Example. The numerical and graphical summaries we have examined.

Definition 1.18 (**Statistical Inference**). When data obtained in the study of a population or process are used to draw general conclusions about the population or process itself, we call this process **statistical inference**.

Example. "Based on my sample, I expect 90% of assignments this term to be submitted within the final 24 hours before the deadline."

- **Inductive reasoning**: we reason from the specific (the observed data on a sample of units) to the general (the target population or process).
- **Deductive reasoning**: when we use general results (axioms) to prove theorems.

The methods of statistical inference will be used to examine 3 main types of problems:

- **Estimation problems**: we are interested in estimating one or more attributes of a process or population.
- **Hypothesis testing problems**: we use the data to assess the truth of some question or hypothesis.
- **Prediction problems**: we use the data to predict a future value of a variate for a unit to be selected from the population or process.

2 Statistical Models and Maximum Likelihood Estimation

2.1 Choosing a Statistical Model

By proposing a statistical model (probability distribution) for our data, we can use our knowledge of that distribution's theoretical properties to answer questions about our study.

Key concept: a statistical model is a mathematical model that incorporates probability.

Once we propose a probability model, we can then use data to help us learn about its parameter(s), and other features such as variance.

We write the p.f. or p.d.f. of a random variable Y as

$$f(y; \theta) \quad \text{for } y \in A = \text{range}(Y)$$

to emphasize the dependence of the model on the parameter θ.

- θ: the truth value.
- $\hat{\theta}$: the estimate value.

Steps of choosing a model:

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

Note. In STAT 231, we will focus on settings in which the models are not too complicated (mostly the models from STAT 230). At this point, need to know Binomial, Poisson, Gaussian, and Exponential distributions.

2.2 Point Estimates and Maximum Likelihood Estimation

Definition 2.1 (**Point Estimate**). A **point estimate** of a parameter θ is the value of a function of the observed data $\mathbf{y}$ and other known quantities such as the sample size n.

Note. Most often the data are of the form $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. The estimate is denoted by $\hat{\theta} = \hat{\theta}(\mathbf{y})$.

Example. $\mathrm{G}(\mu, \sigma)$: we estimate μ by $\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

The estimate in above example seems reasonable given what we know about the Gaussian distribution. Thus, we need a method of estimation which has a mathematical justifiacation and which can be used when a reasonable estimate is not obvious.

Example. Let $Y \sim \mathrm{Bin}(25, \theta)$ model the number of heads observed from flipping a coin 25 times. Suppose in one experiment we observe $y = 10$ heads. Based on this information, the following statement is valid:

- We don't know what $P(\text{head}) = \theta$ is equal to, but $\theta = 0.4$ seems a reasonable guess.

If $\theta = 0.4$, we have observed an event that occurs with probability 0.161, which seems reasonable.

What is the value of θ which make the observed data $y = 10$ heads in 25 flips, most probable? Since $Y \sim \mathrm{Bin}(25, \theta)$, then

$$P(Y = 10; \theta) = \binom{25}{10}\theta^{10}(1-\theta)^{15} \quad \text{for } 0 < \theta < 1.$$

To find the value of θ that maximizes this expression, we will differentiate with respect to θ and set it to 0. We have

$$\frac{\mathrm{d}P(Y = 10; \theta)}{\mathrm{d}\theta} = \binom{25}{10}\theta^{9}(10 - 25\theta)(1-\theta)^{14} = 0$$

from where we can show that $\theta = \frac{10}{25}$ is a solution.

Remark. If the value of θ is in the domain, we can assume that it is the max value. For example, here we have $\theta = \frac{10}{25} \in (0, 1)$. We only need to verify when there isn't a solution in the domain or at the boundaries.

For the example above, we can also see (e.g.) $P(Y = 10; \theta = 0.42)$ is larger than $P(Y = 10; \theta = 0.1)$: the data are more likely if $\theta = 0.42$ than if $\theta = 0.1$. The data are **most likely** when $\theta = 0.4$.

Here, we have used the **Method of Maximum Likelihood** to estimate an unknown parameter θ in an assumed model for the observed data y.

Let the discrete random variable Y represent potential data that will be used to estimate θ and let $\mathbf{y}$ represent the actual observed data.

Definition 2.2 (Likelihood Function for Discrete Distributions).
The **likelihood function** for θ is defined as

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

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

We decide on how reasonable (plausible) a value of θ is by looking at how probable it makes the observed data:

- Values of θ which make the observed data probable are considered to be more reasonable than values of θ which make the observed data improbable.
- In other words, values of θ for which the likelihood function $L(\theta)$ is larger are more consistent with the observed data $\mathbf{y}$.

Definition 2.3 (Maximum Likelihood Estimate).
The value of θ that maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the **maximum likelihood estimate** (MLE) of θ. It is the value which maximizes the probability of observing the data $\mathbf{y}$. This value is denoted by $\hat{\theta}$.

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

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

and we can solve for the maximum likelihood estimate of θ by setting the derivative to 0. We obtain the estimate $\hat{\theta} = \frac{y}{n}$, and it is also called the sample proportion.

Example. In our coin flip example, we flipped a coin $n = 25$ times and saw $y = 10$ heads. We can calculate:

$$P(Y = 10; \theta = 0.4) = 0.161$$
$$P(Y = 10; \theta = 0.25) = 0.042$$

So the data are approximately $\frac{0.161}{0.042} \approx 4$ times more likely if $\theta = 0.4$ than if $\theta = 0.25$.

Remark. We care about the relative likelihood of the data for these two possible values of θ, we don't actually care about the absolute likelihood for each one!

The ratio $\frac{L(\theta_1)}{L(\theta_2)}$ indicates how much less consistent the data are with the value $\theta = \theta_1$ as compared to the value $\theta = \theta_2$.

Definition 2.4 (Relateive Likelihood Function).
The **relative likelihood function** is defined as

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

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

Example. For binomial data

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \quad \text{for } 0 < \theta < 1.$$

The relative likelihood function is

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

Note. For the binomial data above, we can simply write $L(\theta) \propto \theta^y(1-\theta)^{n-y}$.

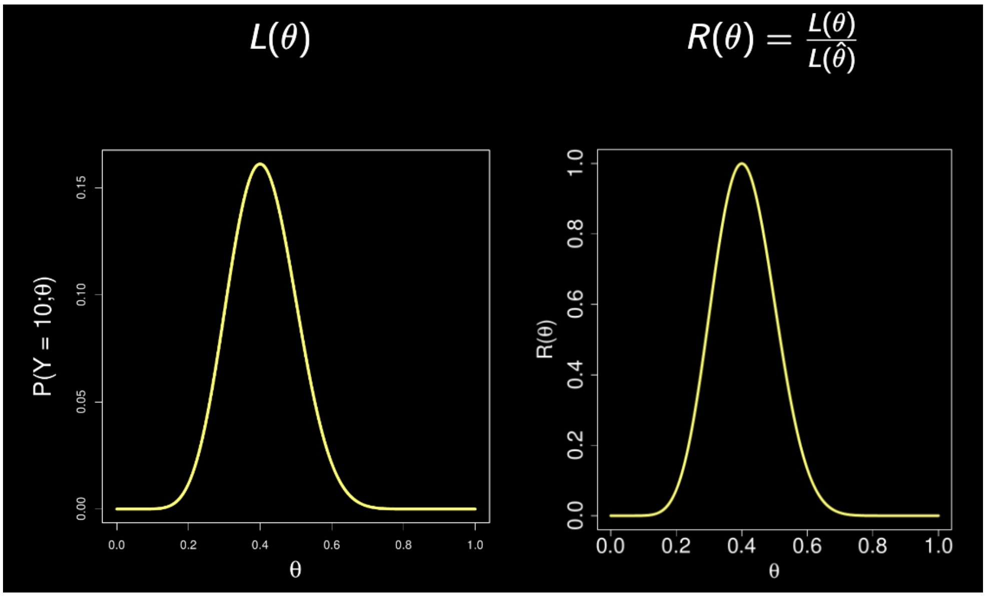


Figure 2: $L(\theta)$ and $R(\theta)$ for the coin example have the same shape!

Definition 2.5 (**Log Likelihood Function**).
The **log likelihood function** is defined as

$$\ell(\theta) = \log(L(\theta)) \quad \text{for } \theta \in \Omega.$$

Note. $\log = \ln$. And the log likelihood function is maximized for the same value of θ as the regular likelihood function.

Lecture 5

Example (Binomial log likelihood).

Recall the binomial likelihood function

$$L(\theta) = \theta^y (1-\theta)^{n-y} \quad \text{for } 0 < \theta < 1.$$

The binomial log likelihood function is

$$\begin{aligned}\ell(\theta) &= \log(L(\theta)) \\ &= \log\left(\theta^y (1-\theta)^{n-y}\right) \\ &= y\log(\theta) + (n-y)\log(1-\theta) \quad \text{for } 0 < \theta < 1.\end{aligned}$$

The graph of the log likelihood function is typically quadratic in shape. Often it is easier to maximize the log likelihood function rather than $L(\theta)$, as the sum rule for differentiation is easier to use than the product rule.

Likelihood Function for Independent Experiments

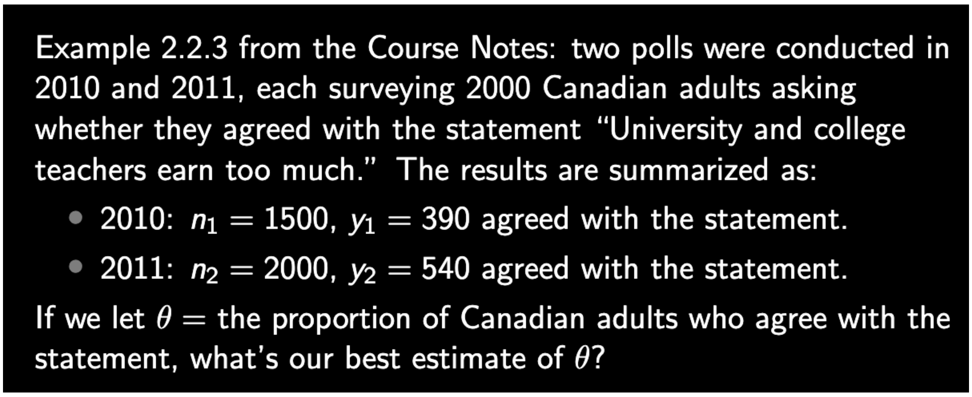


Figure 3: Example 2.2.3 from the Course Notes.

Intuition: a total of 3500 people were surveyed, of whome 930 agreed with the statement, so is our best guess $\frac{930}{3500}$?

Let Y_1, Y_2 correspond to Binomial$(1500, \theta_1)$ and Binomial$(2000, \theta_2)$, respectively. Then, for each survey we can estimate the probability/likelihood as:

$$P(Y_1 = y_1; \theta) = \binom{1500}{390} \theta_1^{390}(1-\theta_1)^{1110}$$
$$P(Y_2 = y_2; \theta) = \binom{2000}{540} \theta_2^{540}(1-\theta_2)^{1460}$$

If we combine the sueveys, we are interested in the joint probability

$$P(Y_1 = y_1, Y_2 = y_2; \theta).$$

If we assume the surveys are independent, we have

$$\begin{aligned} L(\theta) &= P(Y_1 = y_1, Y_2 = y_2; \theta) \\ &= P(Y_1 = y_1; \theta) P(Y_2 = y_2; \theta) \\ &= \binom{1500}{390} \theta^{390}(1-\theta)^{1110} \binom{2000}{540} \theta^{540}(1-\theta)^{1460} \end{aligned}$$

from which we can find $\hat{\theta} = \frac{930}{3500} = 0.266$.

Suppose we have two independent datasets $\mathbf{y}_1$ and $\mathbf{y}_2$ corresponding to independent random variables $\mathbf{Y}_1$ and $\mathbf{Y_2}$. Since $P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2; \theta) = P(\mathbf{Y}_1 = \mathbf{y}_1; \theta) P(\mathbf{Y}_2 = \mathbf{y}_2; \theta)$, the combined likelihood function is

$$L(\theta) = L_1(\theta) L_2(\theta) \quad \text{for } \theta \in \Omega$$

where $L_i(\theta) = P(\mathbf{Y}_i = \mathbf{y}_i; \theta)$ for $i = 1, 2$. By the same argument, if we have n independent datasets, then

$$L(\theta) = \prod_{i=1}^{n} P(\mathbf{Y}_i = \mathbf{y}_i; \theta) \quad \text{for } \theta \in \Omega.$$

Example (**Likelihood Function for Poisson Data**).

Suppose we observe data $y_1, y_2, \dots, y_n$. Assume that these data represent a set of independent and identically distributed observations from a Poisson(θ) model. We want to find the maximum likelihood estimate of θ on these data. For Poisson data $y_1, y_2, \dots, y_n$,

$$P(Y_i = y_i; \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}.$$

The likelihood function for θ is

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

The y_i term doesn't depend on θ, so we can ignore it:

$$L(\theta) = \theta^{n\bar{y}} e^{-n\theta} \quad \text{for } \theta > 0.$$

Let's find the MLE of θ. Using differentiation:

$$\begin{aligned} L(\theta) &= \theta^{n\bar{y}} e^{-n\theta} \\ \implies \frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta) &= (\bar{y} - \theta)\theta^{n\bar{y}-1} n e^{-n\theta}. \end{aligned}$$

If $\frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta) = 0$, then $\hat{\theta} = \bar{y}$, the sample mean.

This is a good example where the log likelihood function makes life easier!

$$\begin{aligned} \ell(\theta) &= n\bar{y}\log(\theta) - n\theta \\ \implies \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta) &= \frac{n\bar{y}}{\theta} - n, \end{aligned}$$

and so $\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta) = 0$ leads to $\hat{\theta} = \bar{y}$. And so

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\hat{\theta}^{n\bar{y}} e^{-n\hat{\theta}}} \quad \text{for } \theta > 0.$$

Definition 2.6 (**Random Sample**). Suppose $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are independent and identically distributed (i.i.d.) ramdom variables with probability function

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

We say $Y_1, Y_2, \ldots, Y_n$ is a **random sample**. And

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

2.3 Likelihood Functions for Continuous Distributions

For discrete random variables, the likelihood function is equal to the probability of observing the data, that is

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

Definition 2.7 (**Likelihood Function for Continuous Distributions**).
If $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ represents a realization of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, a random sample from a continuous distribution with p.d.f. $f(y; \theta)$ for $\theta \in \Omega$, then the likelihood function for θ based on the observed data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is

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

Example (Exponential Likelihood Function).

Suppose $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are observed data from $Y \sim \text{Exponential}(\theta)$. The p.d.f. of Y is $f(y;\theta) = \frac{1}{\theta}e^{\frac{-y}{\theta}}$. The likelihood function for θ is

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

The log likelihood function is

$$\ell(\theta) = \log(L(\theta)) = -n\log(\theta) - \frac{n\bar{y}}{\theta}.$$

To find the MLE of θ, we differentiate $\ell(\theta)$ with respect to θ and set it to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta) = -\frac{n}{\theta} + \frac{n\bar{y}}{\theta^2} = \frac{n}{\theta^2}(\bar{y} - \theta) = 0.$$

Therefore, $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate of θ.

Note. The following example on the next page is from the Course Notes.

Also, the MLE of σ^2 is

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

The sample variance is:

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

Example 2.3.2: Likelihood function for Gaussian distribution

As an example involving more than one parameter, suppose that $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(\mu, \sigma)$ distribution The likelihood function for

$\theta = (\mu, \sigma)$ is

$$\begin{aligned} L(\theta) = L(\mu, \sigma) &= \prod_{i=1}^{n} f(y_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \text{ for } \mu \in \mathbb{R} \text{ and } \sigma > 0 \end{aligned}$$

or more simply (ignoring constants with respect to μ and σ)

$$L(\theta) = L(\mu, \sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \text{ for } \mu \in \mathbb{R} \text{ and } \sigma > 0$$

Since

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$$

and

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \mu)\sum_{i=1}^{n}(y_i - \bar{y}) + \sum_{i=1}^{n}(\bar{y} - \mu)^2 \tag{2.3.3}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \tag{2.3.4}$$

we can write the likelihood function as

$$L(\mu, \sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right] \exp\left[-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right]$$

The log likelihood function for $\theta = (\mu, \sigma)$ is

$$l(\theta) = l(\mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{n(\bar{y} - \mu)^2}{2\sigma^2} \text{ for } \mu \in \mathbb{R} \text{ and } \sigma > 0$$

To maximize $l(\mu, \sigma)$ with respect to both parameters μ and σ we solve [3] the two equations[4]

$$\frac{\partial l}{\partial \mu} = \frac{n}{\sigma^2}(\bar{y} - \mu) = 0 \text{ and } \frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0$$

simultaneously. We find that the maximum likelihood estimate of θ is $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$, where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \text{ and } \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}$$

Summary of Maximum Likelihood Method for Named Distributions

Named Distribution	Observed Data	Maximum Likelihood Estimate	Maximum Likelihood Estimator	Relative Likelihood Function
Binomial(n, θ)	y	$\hat{\theta} = \frac{y}{n}$	$\tilde{\theta} = \frac{Y}{n}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{y} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y}$ $0 < \theta < 1$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{n\hat{\theta}} e^{n\left(\hat{\theta}-\theta\right)}$ $\theta > 0$
Geometric(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{1}{1+\bar{y}}$	$\tilde{\theta} = \frac{1}{1+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{n} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Negative Binomial(k, θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{k}{k+\bar{y}}$	$\tilde{\theta} = \frac{k}{k+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{nk} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\hat{\theta}}{\theta}\right)^{n} e^{n\left(1-\hat{\theta}/\theta\right)}$ $\theta > 0$

2.4 Likelihood Functions for Multinomial Models

Skip for now, come back in Chapter 7.

2.5 Invariance Property of Maximum Likelihood Estimate

One reason the method of maximum likelihood is so popular is the **invariance property**.

Theorem 2.1. If $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ is the maximum likelihood estimate of $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, then $g(\hat{\theta})$ is the maximum likelihood estimate of $g(\theta)$.

Note. $g(\theta)$ is a function of θ.

Example. If $Y \sim \text{Poisson}(\theta)$ and $\hat{\theta} = 3$. Find the MLE of $P(Y \geq 3)$.

We can write the probability as

$$P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - \sum_{y=0}^{2} P(Y = y) = 1 - \sum_{y=0}^{2} \frac{\theta^y e^{-\theta}}{y!}.$$

Notice that this is just a function of θ. By the invaraince property, the MLE of $P(Y \geq 3)$ is

$$1 - \sum_{y=0}^{2} \frac{\hat{\theta}^y e^{-\hat{\theta}}}{y!} = 1 - \sum_{y=0}^{2} \frac{3^y e^{-3}}{y!} = 0.647.$$

2.6 Checking the Model

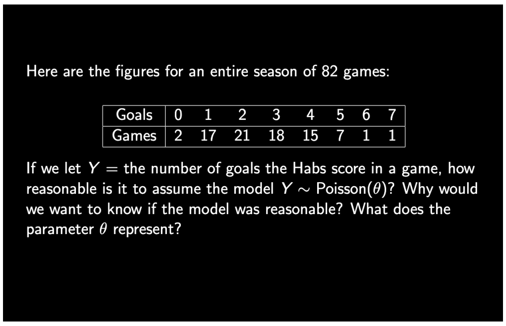
We've already seen two methods of checking model fit:

- Compare a relative frequency histogram of the observed data with a superimposed graph of the p.d.f. of the assumed model (continuous).
- Compare a graph of the ecdf with a superimposed graph of the cdf of the assumed model (continuous).

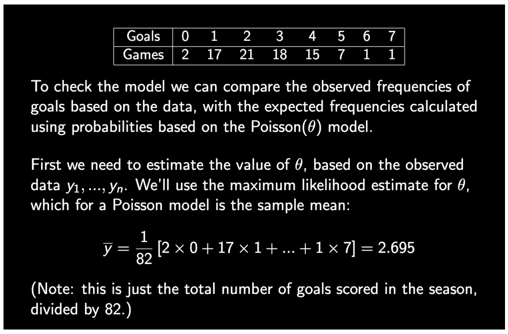
In this section, we will discuss those methods in more detail, as well as two other methods:

- Compare observed frequencies with expected frequencies calculated using the assumed model (discrete & continuous).
- Examine a Gaussian qqplot.

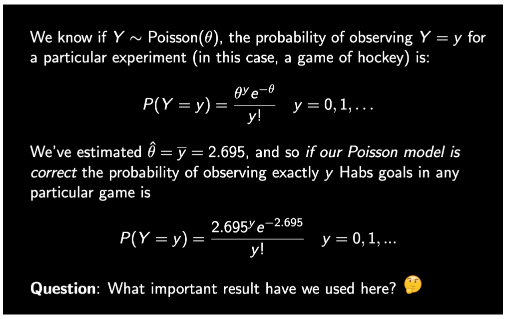
Hockey goals example from Prof. Wallace's slides:



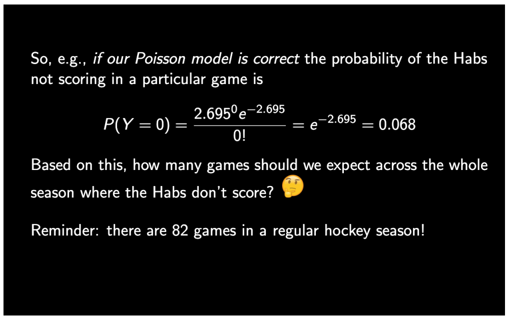
(1)



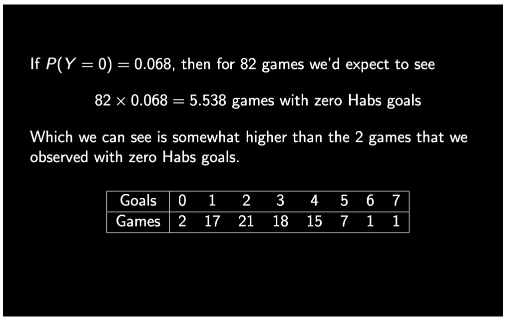
(2)



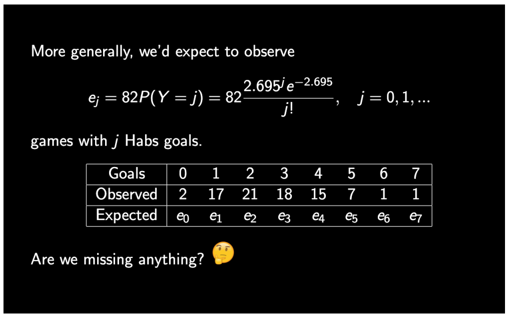
(3)



(4)



(5)



(6)

Often, we should group some values together:

Goals	0	1	2	3	4	5	6	≥ 7
Observed	2	17	21	18	15	7	1	1
Expected	e_0	e_1	e_2	e_3	e_4	e_5	e_6	e_{7+}

This accounts for the fact that (for example) the range of the Poisson distribution is all non-negative integers.

But how do we compute e_{7+}? 🤔

(7)

We know that the expected number of games with 7 or more goals is 82 times the probability a random game has 7 or more goals:

$$e_{7+} = 82P(Y \geq 7)$$

and since our probabilities must sum to 1:

$$e_{7+} = 82P(Y \geq 7) = 82\,[1 - P(Y \leq 6)] = 82\left[1 - \sum_{j=0}^{6} P(Y = j)\right]$$

(8)

So if e_{7+} is the expected number of games in which 7 or more goals are scored, we have

$$e_{7+} = 82\left[1 - \sum_{j=0}^{6} P(Y = j)\right] = 82 - \sum_{j=0}^{6} 82P(Y = j)$$

But don't we already know something about $82P(Y = j)$? 🤔

(9)

Because $e_j = 82P(Y = j)$ we can simplify this further to

$$e_{7+} = 82 - \sum_{j=0}^{6} 82P(Y = j) = 82 - \sum_{j=0}^{6} e_j$$

We can find the expected number of games with 7 or more goals by taking the total of 82, and subtracting the expected number of games for $j = 0, 1, \ldots, 6$ which we have already calculated.

For more examples of this, see Problems 11-13 at the end of Chapter 2!

(10)

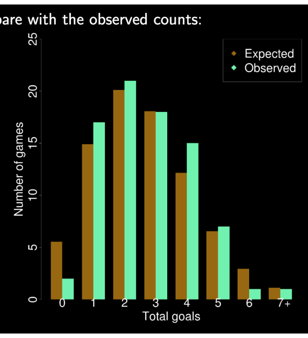
Here's how our observed and expected counts work out:

Goals	0	1	2	3	4	5	6	≥ 7
Observed	2	17	21	18	15	7	1	1
Expected	5.54	14.93	20.11	18.07	12.17	6.56	2.95	1.67

How do these compare?

(11)

And compare with the observed counts:



(12)

Lecture 6

Example (Continuous Data).

Let $Y \sim G(\mu, \sigma)$. By default, we estimate the parameters by taking $\mu = \bar{y}$ (the MLE) and $\sigma = s$ (the sample standard deviation).

For some specific example data, we have $Y \sim G(159.77, 6.03)$, where Y = height of a randomly selected woman.

Question: How do we compute the expected number of women with height in [160, 162] (a randomly selected interval for illustration)?

The probability is $P(160 \leq Y \leq 162)$. We can calculate this using R:

```
> pnorm(162, 159.77, 6.03) - pnorm(160, 159.77, 6.03)
[1] 0.1290278
```

Let n be the number of women with height in [160, 162]. The expected number is then $n \times 0.1290278$.

Q-Q Plot (Gaussian)

Suppose a set of data (see Prof's slide) follow $Y \sim G(50, 10)$, we can check the theoretical $G(50, 10)$ distribution via histogram/ecdf. For numerical summaries, the median of a $G(50, 10)$ random variable is 50. We extend this idea to calculate other quantiles.

For example: What is the 25^{th} percentile of $Y \sim G(50, 10)$? What is the 75^{th} percentile?

We can calculate these using R:

```
> qnorm(0.25, 50, 10)
[1] 43.2551
> qnorm(0.75, 50, 10)
[1] 56.7449
```

Then, we know that $P(Y < 43.26) = 0.25$ and $P(Y < 56.74) = 0.75$.

In general, suppose our data are from a population with distribution $G(\mu, \sigma)$.

We know that the median of our sample should be close to the median of a $G(\mu, \sigma)$ distribution (which is also the mean).

More generally, the lower quartile $q(0.25)$ of our sample should be close to the *theoretical* lower quartile of the $G(\mu, \sigma)$ distribution. The upper quartile $q(0.75)$ should be close to the *theoretical* upper quartile of the $G(\mu, \sigma)$ distribution.

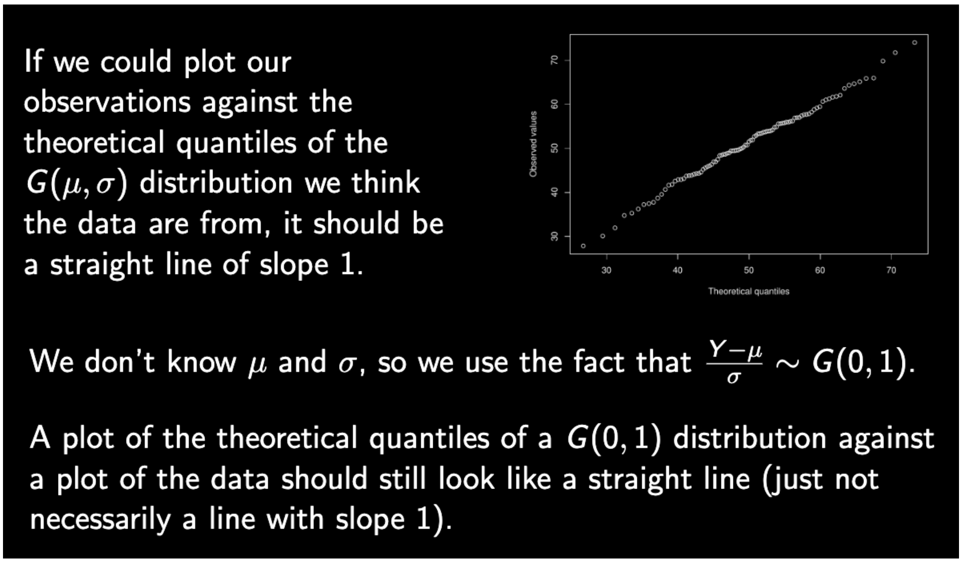


Figure 6: Q-Q Plot for Data before Standardization.

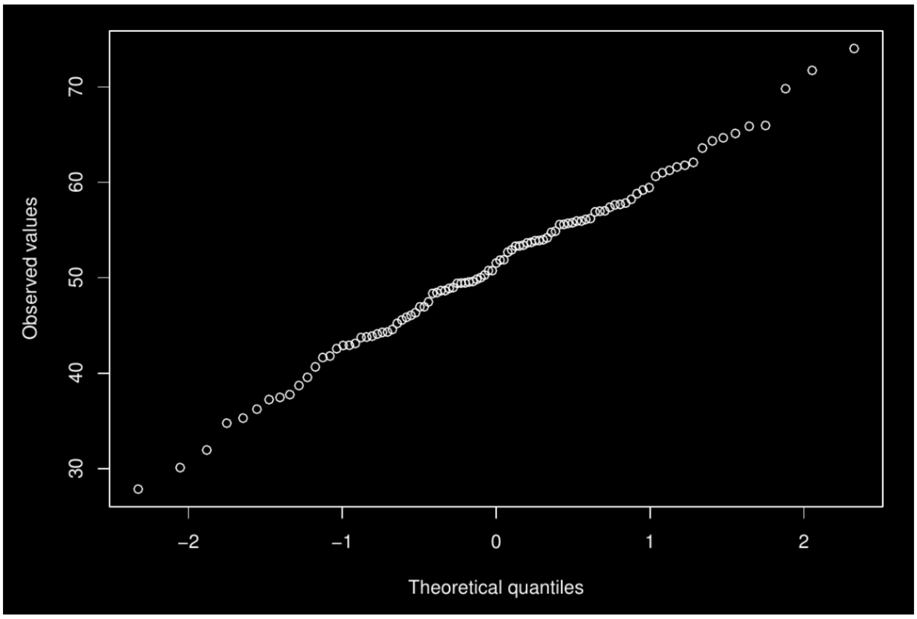


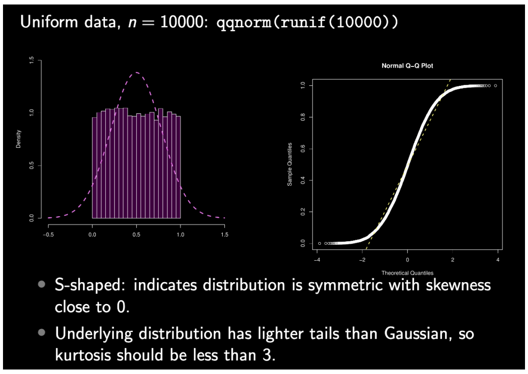
Figure 7: Q-Q Plot after Standardization.

Therefore, if the points appear to lie reasonably along a straight line, then they are consistent with the Gaussian distribution. Q-Q plots can also identify other features of our sample:

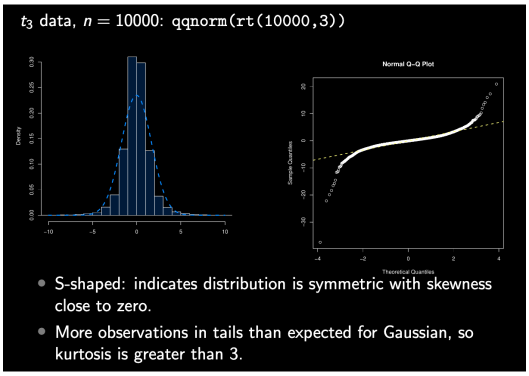
- Gaussian or non-Gaussian.
- Symmetric or skewed.
- High or low kurtosis.

Remark. Generate random data from a Gaussian distribution and plot a Q-Q plot:

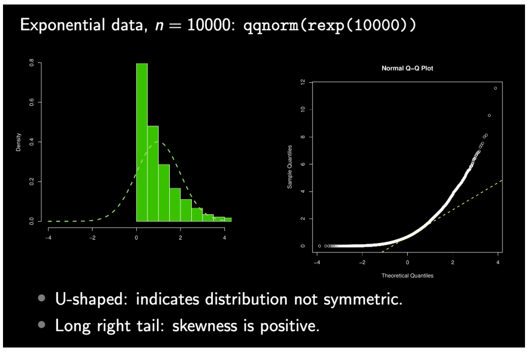
```
qqnorm(rnorm(100))
```



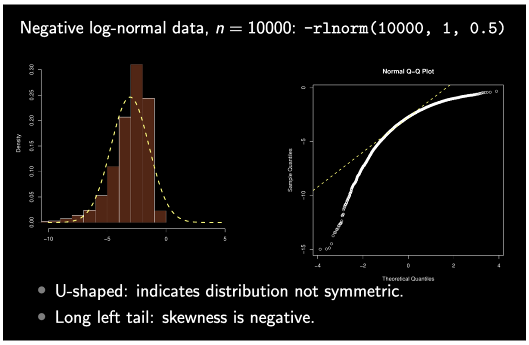
(1)



(2)



(3)



(4)

Note. The x-axis of the histogram will become the y-axis of the Q-Q plot.

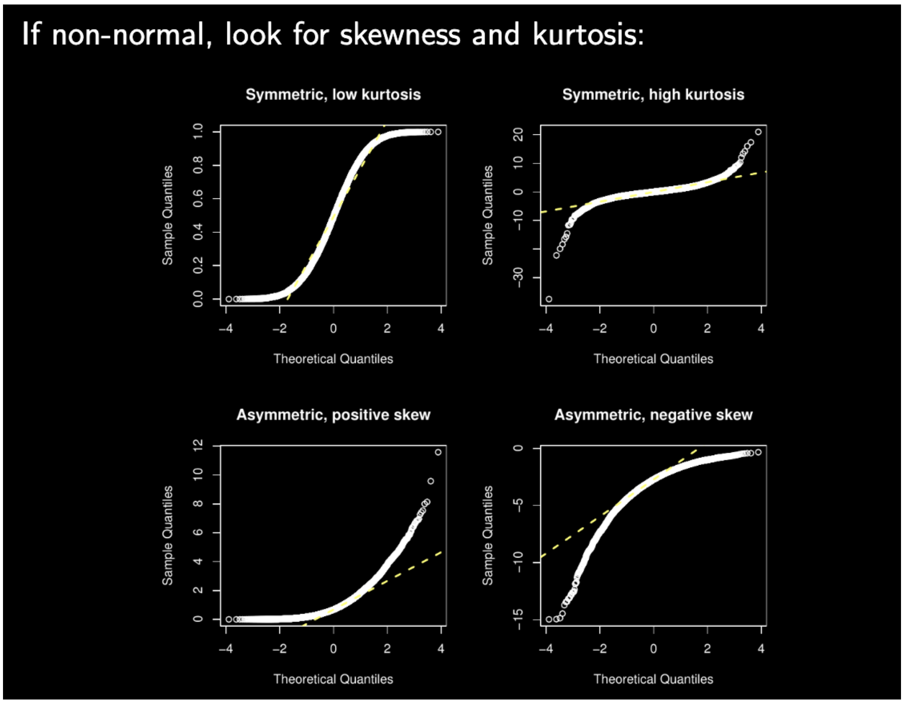


Figure 9: Summary for Non-Gaussian.

- A straight line indicates normality.
- If not, are they S-shaped? This indicates symmetry and low kurtosis.
- If not, are they U-shaped? This indicates asymmetry.
- If asymmetric, are there more points in the left of right tails? This tells us about skewness.
- If symmetric, are there more/fewer observations in the tails than we'd expect? This tells us about kurtosis.

3 Planning and Conducting Empirical Studies

3.1 Empirical Studies

The goal of an empirical study is to collect data in order to learn about a population or process.

PPDAC emphasizes the statistical aspects of designing an empirical study.

- **Problem**: a clear statement of the study's objectives.
- **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, accounting for considerations in the Problem and the Plan.
- **Conclusion**: the conclusions that are drawn about the Problem and any limitations of the study.

Lecture 7

3.2 The Steps of PPDAC

Problem

The Problem step address questions starting with 'What'.

Definition 3.1 (**Target Population/Process**).
The **target population** or **target process** is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.

Definition 3.2 (**Variate**). A **variate** is a characteristic of a unit.

Note. To determine the variates, look at what is measured or recorded on each unit.

Definition 3.3 (**Attribute**). An **attribute** is a function of the variates over a population.

Types of Problems:

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

Usually, we cannot answer causative problems from observational studies or sample surveys.

Plan

The purpose of the Plan step is to decide what units are available for study, what units will be examined, and what variates will be collected and how.

Definition 3.4 (**Study Population/Process**).
The **study population** or **study process** is the collection of units available to be included in the study.

Remark. 'Available to be included' means that set of units that **could** be included in the study. Often, the study population is a strict subset of the target population, but not always.

Definition 3.5 (**Study Error**). If the attributes in the study population differ from the attributes in the target population, then the difference is called **study error**.

Example. If we use online survey to estimate the most common favorite colour in our population, we would not expect younger people to be more/less likely to have a particular favourite colour than older people, and so this would **not** be an example of study error!

Definition 3.6 (**Sampling Protocal, Sample Size**).
The **sampling protocol** is the procedure used to select a sample of units from the study population. The number of units sampled is called the **sample size**.

Definition 3.7 (**Sample Error**). If the attributes in the sample differ from the attributes in the study population, then the difference is called **sample error**.

Note. The sample is only a subset of the units in the study population. Different sampling protocols may lead to different sample errors. Statistical models are used to quantify the size of this error.

Definition 3.8 (**Measurement Error**). If the measured value and the true value of a variate are not identical, the difference is called **measurement error**.

Example. In a survey, participants may not tell the truth. This leads to measurement error.

Lecture 8

Data

The purpose is to collect data according to the Plan.

In many studies, the units must be tracked and measured over a long period of time (longitudinal data). When data are recorded over time or in different locations, the time and place for each measurement should be recorded. Also, departures from the Plan may arise over time, and these should be recorded.

Analysis

The Analysis step consists of the analyses of the data collected. This should include numerical and graphical summaries of the data, selecting an appropriate model and checking the fit of model.

Conclusion

The questions posed in the Problem are answered to the extent permitted by the data: the Conclusion step is directed by the Problem.

Potential study, sample or measurement errors, as described in the Plan step, should be discussed and quantified if possible. Departures from the Plan that affect the Analysis must be addressed. Also, the limitations of the study must be discussed.

4 Estimation

4.1 Statistical Models and Estimation

In choosing a model for data collected in an empirical study in the Analysis step of PPDAC, we need two models:

(1) A model for variation in the (target) population being studied which includes the attributes which are to be estimated.

(2) A model which takes into account how the data (study population) were collected and which is constructed in conjunction with the model in (1).

In this course, we usually assume that the data arise as a random sample from the study population and that the variates are measured without error. This means we are only able to estimate attributes of interest in the study population (not the target population).

Thus, we either limit our conclusions to the study population or we make clear any assumptions we make about whether the results translate to the target population.

4.2 Estimators and Sampling Distributions

If we could take repeated samples, each sample mean would probably be different. Each sample mean is a realization of a random variable.

If we sample from $Y_i \sim \mathrm{G}(\mu, \sigma)$, we know that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathrm{G}(\mu, \frac{\sigma}{\sqrt{n}})$.

Definition 4.1 (Point Estimate). A **point estimate** of θ is a function $\hat{\theta} = g(y_1, \ldots, y_n)$ of the observed data used to estimate the unknown parameter θ.

Since estimates vary as we take repeated samples, we associate a random variable with these estimates.

Definition 4.2 (Point Estimator, Sampling Distribution).
A **ponit estimator** is a random variable which is a function $\tilde{\theta} = g(Y_1, \ldots, Y_n)$ of the random variables $Y_1, \ldots, Y_n$. The distribution of $\tilde{\theta}$ is called the **sampling distribution** of the estimator.

Note. Since $\tilde{\theta}$ is a random variable, it has a distribution. In other words, it has a p.f or a p.d.f.. Note that $\hat{\theta}$ is an estimate (numerical value) and $\tilde{\theta}$ is the corresponding estimator (random variable).

Lecture 9

Example (Gaussian with Known σ).
Suppose $Y_i \sim G(\mu, \sigma)$ for $i = 1, \dots, n$. Then the estimator of μ is

$$\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{\sqrt{n}}).$$

Note that the probability we draw a sample that has an estimate $\hat{\mu}$ that is close to μ

- increases as n increases.
- decreases as σ increases.
- does not change with μ (probability not depending on μ).

Suppose μ is unknown, and say we want to find

$$P(\mu - 1 \leq \bar{Y} \leq \mu + 1) \text{ where } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G(\mu, \frac{\sigma}{\sqrt{n}}).$$

We do not know the distribution of $\bar{Y}$, but we know that μ does not affect the probability. We can standardize $\bar{Y}$:

$$P(\mu - 1 \leq \bar{Y} \leq \mu + 1) = P\left(\frac{\mu - 1 - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu + 1 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(-\frac{\sqrt{n}}{\sigma} \leq Z \leq \frac{\sqrt{n}}{\sigma}\right).$$

Example (Non-Gaussian). Suppose we have a Poisson model. Let $Y_1, \dots Y_n$ be i.i.d. with $Y_i \sim$ Poisson(θ). Let $y_1, \dots, y_n$ be the observed data. The sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the corresponding estimator is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Question: What is the probability my sample will result in a point estimate within 0.5 of the true population mean?

Equivalently, we want to find

$$P(\theta - 0.5 \leq \bar{Y} \leq \theta + 0.5).$$

Sometimes the sampling distribution must be determined approximately using the **Central Limit Theorem** (CLT). Let $Y_1, \ldots, Y_n$ be i.i.d. with $\mathbb{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$. Define

$$Z_n = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then for sufficiently large n, Z_n has an approximate $\mathrm{G}(0,1)$ distribution. In other words, if our sample size is large enough, we can take observations from any probability distribution and transform them into a standard normal!

For $Y_i \sim \text{Poisson}(\theta)$, $i = 1, \ldots, n$, independently and if n is large, then by CLT,

$$\frac{\bar{Y} - \theta}{\sqrt{\frac{\theta}{n}}} \sim \mathrm{G}(0,1) \text{ approximately.}$$

We can rearrange to find

$$\bar{Y} \sim \mathrm{G}\left(\theta, \sqrt{\frac{\theta}{n}}\right) \text{ approximately.}$$

In other words, the sampling distribution of the sample mean for Poisson data will be approximately Gaussian.

Theorem 4.1 (**Summary of Gaussian Approximation**).

- **Poisson**: $\bar{Y} \sim \mathrm{G}\left(\theta, \sqrt{\frac{\theta}{n}}\right)$ approximately.
- **Exponential**: $\bar{Y} \sim \mathrm{G}\left(\theta, \frac{\theta}{\sqrt{n}}\right)$ approximately.
- **Binomial**: $\frac{Y}{n} \sim \mathrm{G}\left(\theta, \sqrt{\frac{\theta(1-\theta)}{n}}\right)$ approximately (sum of Bernoulli distributions).

Note. Be careful, for Poisson data as an example, the true mean θ does affect the standard deviation of the sampling distribution. Also, the shape of our population distribution will affect how many of our sample estimates will be close to the true mean μ.

4.3 Interval Estimation Using the Likelihood Function

Definition 4.3 (**Relative Likelihood Function**).

Given a likelihood function $L(\theta)$. The **relative likelihood function** $R(\theta)$ is defined as:

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

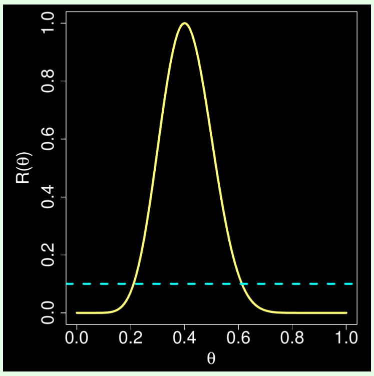
Note: $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$ and $R(\hat{\theta}) = 1$.

Definition 4.4 (**Likelihood Interval**).

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

Remark. So for $p = 0.1$, we define a 10% likelihood interval as the set of values of θ where $R(\theta) \geq 0.1$.

Example (**Coin Example**). Let $n = 25$ and 10 heads. We have 10% LI $\approx (0.2, 0.6)$.



Lecture 10

Important: the likelihood interval is a function of the data!

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

Figure 10: Guidelines for Interpreting Likelihood Intervals.

As the sample size n increases, the graph of $R(\theta)$ becomes narrorer. But, a larger n does not guarantee a narrower likelihood interval!

Definition 4.5 (**Log Relative Likelihood Function**).
The **log relative likelihood function** is

$$r(\theta) = \log R(\theta) = \ell(\theta) - \ell(\hat{\theta}). \ \theta \in \Omega,$$

where $\ell(\theta) = \log L(\theta)$ is the log likelihood function.

Remark. $r(\theta)$ is often easier to compute than $R(\theta)$. Also, the maximum value of $r(\theta)$ is 0.

Note.

- If $R(\theta)$ is unimodal then $r(\theta)$ is also unimodal, and both graphs get maximum value at the MLE of θ.
- $R(\theta)$ often looks like bell-shaped while $r(\theta)$ looks like a quadratic function of θ.
- Since $R(\theta) \geq p \iff r(\theta) \geq \log p$, we can find a $100p\%$ likelihood interval by drawing a line at $r(\theta) = \log p$.

4.4 Confidence Intervals and Pivotal Quantities

Definition 4.6 (**Coverage Probability**).
Suppose $[L(\mathbf{Y}), U(\mathbf{Y})]$ is an interval estimator (a rule) which can be used to construct an interval of plausible values for the unknown parameter θ. The value

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

is the **coverage probability** for the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$.

The **coverage probability** is the probability that the random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$ contains the true (unknown) value of θ.

Note. Remember that $L(\mathbf{Y})$ and $U(\mathbf{Y})$ are both random variables.

As the likelihood level increases, the likelihood intervals become narrower, and the coverage decreases!

- Higher coverage seems preferable: more likely to have an interval that covers the true value.
- Narrower intervals seem preferable: more precise estimate.

Definition 4.7 (**Confidence Interval, Confidence Coefficient**).
A $100p\%$ **confidence interval** for a parameter is an interval estimate $[L(y), U(y)]$ for which

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

The value p is called the **confidence coefficient** for the confidence interval.

Example. Suppose $0.95 = P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})])$ and 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, this equation tells us that we should expect 95% of these constructed intervals to contain the true but unknown value of θ.

Remark. 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 just a constant (is whether in or not in the interval).

Definition 4.8 (**Pivotal Quantity**).

A **pivotal quantity** $Q = Q(\mathbf{Y}; \theta)$ is a function of the data $\mathbf{Y}$ and the unknown parameter θ such that the distribution of the random variable Q is completely known. That is, probability statements such as $P(Q \le b)$ and $P(Q \ge a)$ depend on a and b but not on θ or any other unkown information.

Example. $\bar{Y} \sim \mathrm{G}(\mu, \frac{1}{\sqrt{16}})$,

$$\frac{\bar{Y} - \mu}{\frac{1}{\sqrt{16}}} \sim \mathrm{G}(0, 1),$$

where μ is the unknown parameter, $\bar{Y}$ is the data, and $\mathrm{G}(0, 1)$ is completely known.

In general, we can use a pivotal quantity to construct a $100p\%$ confidence interval as follows:

1. Determine numbers a and b such that $P(a \le Q(\mathbf{Y}; \theta) \le b) = p$.

   There is an infinite pairs of a and b, but since Gaussian is symmetric, we can choose a and $-a$ for some a to give the narrowest confidence interval.

2. Solve the probability for θ: $P(L(\mathbf{Y}) \le \theta \le U(\mathbf{Y})) = p$.

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

Example. Suppose that $Q(\mathbf{Y}; \mu) = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{G}(0, 1)$ and $p = 0.95$.

1. Find a and b such that $P(a \le \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \le b) = 0.95$. We choose $a = -1.96$ and $b = 1.96$ which give the narrowest confidence interval.
2. Solve for μ:

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

3. A 95% confidence interval for μ based on the observed data $y_1, \dots y_n$ is therefore $\left(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$. Note: this is the numerical interval estimate.

Remark. If we calculate an interval with a different confident level for the same data, only a changes for our confidence interval of the form $\bar{y} \pm a\frac{\sigma}{\sqrt{n}}$. Greater confidence means wider confidence intervals. If we have greater confidence, then we are more certain that the true value of μ lies in the confidence interval.

Lecture 11

A $100p\%$ confidence interval for μ is of the form:

$$\text{point estimate} \pm (\text{distribution quantile}) \times (\text{sd}(\text{estimator})).$$

Such an interval is called a twosided, equal-tailed confidence interval.

Note. The width of our confidence interval is $2a\frac{\sigma}{\sqrt{n}}$.

- A larger n will result in a narrower CI.
- A larger σ will result in a wider CI.

Example. Suppose $Y \sim \text{Binomial}(n, \theta)$. We know that $\tilde{\theta} = \frac{Y}{n}$. How can we construct a pivotal quantity in this case (non-Gaussian)?

By CLT and for large n, $\tilde{\theta} \sim \mathrm{G}\left(\theta, \sqrt{\frac{\theta(1-\theta)}{n}}\right)$ and so $\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim \mathrm{G}(0,1)$ approximately.

A useful result: suppose $\tilde{\theta}$ is a point estimator for an unknown parameter θ, and that the CLT can be used to obtain

$$\frac{\tilde{\theta} - \theta}{\frac{g(\theta)}{\sqrt{n}}} \sim \mathrm{G}(0,1)$$

approximately for large n, where $\mathbb{E}\left[\tilde{\theta}\right] = \theta$ and $\text{sd}(\tilde{\theta}) = \frac{g(\theta)}{\sqrt{n}}$. Then

$$\frac{\tilde{\theta} - \theta}{\frac{g(\tilde{\theta})}{\sqrt{n}}} \sim \mathrm{G}(0,1)$$

approximately for large n.

For Binomial data, we have $\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim \mathrm{G}(0,1)$ approximately for large n too. This is an example of an approximate pivotal quantity.

Note. For Binomial, an approximate CI for θ is $\hat{\theta} \pm a\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}$ with $P(Z \leq a) = \frac{1+p}{2}$.

Definition 4.9 (**Asymptotic/Approximate Pivotal Quantity**).
We often find random variables $Q_n = Q_n(\mathbf{Y};\theta)$ such that as $n \to \infty$, the distribution of Q_n stops to depend on θ or other unknown information. We call Q_n an **asymptotic** or **approximate pivotal quantity**.

4.5 The Chi-Squared and t Distributions

Definition 4.10 (**Chi-Squared Distribution**).
The **chi-squared distribution** with k degrees of freedom, denoted χ^2_k or $\chi^2(k)$, has probability density function

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2},\ x > 0$$

where $\Gamma(z) = \displaystyle\int_0^\infty t^{z-1}e^{-x}\,dx$ and $k = 1, 2, 3, \ldots$.

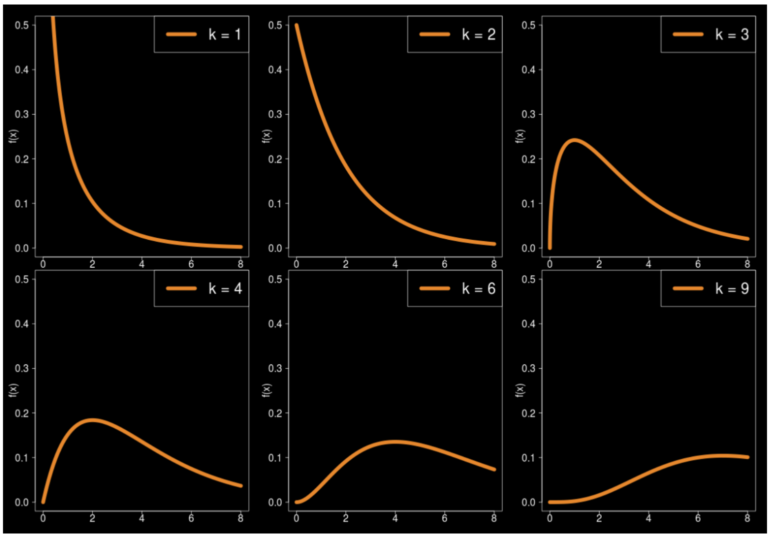


Figure 11: Chi-Squared Distribution p.d.f.s for different k.

Note. $k = 1, 2$ are special cases. As k increases, the distribution becomes more symmetric.

Remark.

- The command `pchisq(w, df)` will return $P(W \leq w)$ where $W \sim \chi^2_{df}$.
- The command `qchisq(q, df)` will return a value w s.t. $P(W \leq w) = q$ where $W \sim \chi^2_{df}$.

Proposition 4.2. Suppose $W \sim \chi^2_k$, then $\mathbb{E}[W] = k$ and $\text{Var}(W) = 2k$.

Note. Recommend knowing how to derive these if you are going into higher level stat courses. We can show these results by first showing that $\mathbb{E}\left[X^j\right] = 2^j \dfrac{\Gamma(\frac{k}{2} + j)}{\Gamma(\frac{k}{2})}$ for $j = 1, 2, \ldots$.

Theorem 4.3. 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 4.4. If $Z \sim \text{G}(0, 1)$, then the distribution of $W = Z^2$ is χ^2_1.

Remark (**Useful Results**).

1. If $W \sim \chi^2_1$, then $P(W \leq w) = P(Z^2 \leq w) = P(-\sqrt{w} \leq Z \leq \sqrt{w}) = 2P(Z \leq \sqrt{w}) - 1$.
2. If $W \sim \chi^2_2$, then $W \sim \text{Exponential}(2)$ and $P(W \leq w) = 1 - e^{-w/2}$.

Corollary 4.5. If $Z_1, Z_2, \ldots, Z_n \sim \text{G}(0, 1)$ independently, then

$$S = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

We skip t distributions for now and come back in 4.7!

4.6 Likelihood-Based Confidence Intervals

We've now looked at two types of interval estimation:

- **Likelihood intervals**: values of θ such that $R(\theta) \geq p$.
- **Confidence intervals**: values of θ such that $P(L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})) = q$.

Theorem 4.6. If $L(\theta)$ is based on $\mathbf{Y} = (Y_1, \dots, Y_n)$, a random sample of size n, and if θ is the true value of the parameter, then

$$\Lambda(\theta) = -2\log\left(\frac{L(\theta)}{L(\tilde{\theta})}\right) \sim \chi_1^2 \text{ as } n \to \infty$$

where $\tilde{\theta} = \tilde{\theta}(\mathbf{Y})$ is the maximum likelihood estimator of θ.

$\Lambda(\theta)$ is therefore a random variable that depends on $\mathbf{Y}$ and we call it the **likelihood ratio statistic**.

Note. For large n, $\Lambda(\theta)$ is an approximate pivotal quantity that can be used to obtain approximate confidence intervals for θ.

Lecture 12

Example (Relationship between CI and LI). We want to explore the relationship between likelihood intervals and confidence intervals. Remember, we can use an <u>approximate</u> pivital quantity to construct a $100q\%$ CI as follows:

1. Determine a, b such that $P(a \leq Q(\mathbf{Y};\theta) \leq b) \approx q$.
2. Solve for θ: $P(L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})) \approx q$, so the coverage probability is approximately q.
3. For observed data y, the interval $[L(y), U(y)]$ is a $100q\%$ CI for θ.

Now, look at step 1. We want to find a, b such that

$$P\left(a \leq -2\log\left(\frac{L(\theta);\mathbf{Y}}{L(\tilde{\theta});\mathbf{Y}}\right) \leq b\right) \approx q$$

where $-2\log\left(\frac{L(\theta);\mathbf{Y}}{L(\tilde{\theta});\mathbf{Y}}\right) \sim \chi^2_1$ approximately. Equivalently, we want to find c such that

$$P\left(-2\log\left(\frac{L(\theta);\mathbf{Y}}{L(\tilde{\theta});\mathbf{Y}}\right) \leq c\right) \approx q.$$

For example, find c such that

$$P\left(-2\log\left(\frac{L(\theta);\mathbf{Y}}{L(\tilde{\theta});\mathbf{Y}}\right) \leq c\right) = 0.95.$$

This is equivalent to finding c such that $P(W \leq c) = 0.95$ where $W \sim \chi^2_1$. By a result, $W = Z^2$ where $Z \sim \mathrm{G}(0,1)$. Equivalently, $P(-\sqrt{c} \leq Z \leq \sqrt{c}) = 0.95 \implies c = 1.96^2$.

It is difficult to isolate θ so we just skip to step 3. We can see that an approximate 95% CI for a specific sample can be written as

$$\{\theta : L(y) \leq \theta \leq U(y)\} = \{\theta : -2\log\left(\frac{L(\theta;y)}{L(\tilde{\theta};y)}\right) \leq 1.96^2\}.$$

Note. A CI is, by definition, just a set of values of θ which satisfy some constraint.
If we look at

$$\{\theta : L(y) \leq \theta \leq U(y)\} = \{\theta : -2\log\left(\frac{L(\theta;y)}{L(\tilde{\theta};y)}\right) \leq c\}.$$

We have

$$\begin{aligned}\{\theta : L(y) \leq \theta \leq U(y)\} &= \{\theta : -2\log(R(\theta)) \leq c\} \\ &= \{\theta : R(\theta) \geq e^{-c/2}\}.\end{aligned}$$

where $P(W \leq c) = q$ and $W \sim \chi^2_1$. Thus, our approximate CI is just a set of values of θ such that $R(\theta) \geq$ some value.
Note. A $100p\%$ likelihood interval is $\{\theta : R(\theta) \geq p\}$. Thus, A CI and a LI are (approximately) the same thing!
In particular, we can say that a $100q\%$ CI is approximately a $100p\%$ likelihood interval with $p = e^{-c/2}$ and $P(W \leq c) = q$, $W \sim \chi^2_1$.

Example (From CI to LI).

Question: A 95% confidence interval is approximately what level of likelihood interval?

- Step 1: Find c such that $P(W \leq c) = 0.95$ where $W \sim \chi^2_1$. This gives us $c = 1.96^2$.
- Step 2: The likelihood level is $p = e^{-c/2}$.

  Plugging c into this expression gives $p = e^{-1.96^2/2} = 0.146$.

Thus, a 95% confidence interval is approximately a 15% likelihood interval!

Example. If we toss a coin n times and see y heads, then if $\theta = P(\text{heads})$, we have $Y \sim \text{Binomial}(n, \theta)$ and relative likelihood function

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}$$

where $\hat{\theta} = \frac{y}{n}$. Suppose $n = 100$ and $y = 40$, let's calculate a 15% likelihood interval for θ as well as an apprxoimate 95% confidence interval.

15% likelihood interval based on $R(\theta)$:

$$\{\theta : \frac{\theta^{40}(1-\theta)^{60}}{\hat{\theta}^{40}(1-\hat{\theta})^{60}} \geq 0.15\} \approx [0.308, 0.497].$$

Approximate 95% confidence interval based on CLT:

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.4 \pm 1.96\sqrt{\frac{0.4 \times 0.6}{100}} \approx [0.304, 0.496].$$

The intervals are very similar!

Remark. This result tells us that if we took repeated samples and calculated a 15% likelihood interval for each sample, then approximately 95% of samples would result in an interval that contained the true value.

Example (From LI to CI).

Question: What confidence level would you expect for a 10% likelihood level?

Let's restate the question in terms of notation: An approximate 100%q CI is:

$$\{\theta : R(\theta) \geq e^{-c/2}\},\ P(W \leq c) = q,\ W \sim \chi_1^2.$$

A 100p% LI is:

$$\{\theta : R(\theta) \geq p\}.$$

Question: Given p, what is q?

Given p, this tells us that $c = -2\log(p)$. Then,

$$q = P(W \leq c) = P(W \leq -2\log(p)).$$

For a 10% LI we take $p = 0.1$. So

$$c = 4.60517 \implies q = P(W \leq 4.60517) = 0.9681243731.$$

Therefore, a 10% LI is approximately a 97% CI!

Remark. The likelihood ratio statistic shows how CI and LI relate:

- LIs have an associated coverage probability.
- CIs contain values of θ which are 'more plausible' given the data.

Theorem 4.7 (LI to CI). A 100p% LI is an apprxoimate 100q% CI where $q = P(\Lambda(\theta) \leq -2\log(p))$ and $\Lambda(\theta) \sim \chi_1^2$. This implies:

$$q = 2P\left(Z \leq \sqrt{-2\log(p)}\right) - 1,\ Z \sim \mathrm{G}(0,1).$$

Theorem 4.8 (CI to LI). Given confidence level q, the LI level p is found by:

1. Find c such that $q = P(W \leq c) = P(|Z| \leq \sqrt{c})$ where $W \sim \chi_1^2$.
2. The LI is given by $\{\theta : R(\theta) \geq e^{-c/2}\}$.
3. The corresponding LI level is $p = e^{-c/2}$.

Remark. These intervals are only approximately equivalent - not identical! For the above example, we had a $n = 100$ and $\tilde{\theta} = 0.4$, which is somewhat symmetric. Let's take $n = 30$ and $y = 3$ so that $\tilde{\theta} = 0.1$. Then 95% CI $= [-0.007, 0.207]$, whereas 15% LI $= [0.026, 0.238]$. The intervals are not very similar!

In general, if $\tilde{\theta} = 0.5$ or n is large enough, the intervals will be very similar. If $\tilde{\theta}$ is close to 0 or 1 and n is small, the intervals will not be similar (the LI will not be symmetric about $\tilde{\theta}$).

4.7 Confidence Intervals for Parameters in the Gaussian Model

Suppose $Y \sim \mathrm{G}(\mu, \sigma)$. We showed that if we knew σ, then for a sample with mean $\bar{y}$ a $100p\%$ CI for μ is

$$\left(\bar{y} - a\frac{\sigma}{\sqrt{n}}, \bar{y} + a\frac{\sigma}{\sqrt{n}}\right)$$

where a is such that $P(Z \leq a) = \frac{1+p}{2}$. What if we don't know σ?

Example (Gaussian with Unknown σ).

Suppose $Y_1, \ldots, Y_n$ is a random sample with $Y_i \sim \mathrm{G}(\mu, \sigma)$ where both μ and σ are unknown. Recall that the maximum likelihood estimator of μ is

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

Two estimators for σ^2:

1. Maximum likelihood estimator: $\tilde{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
2. Sample variance: $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

We need to decide which to use: we prefer S^2 because it is an unbiased estimator, that is, $\mathbb{E}\left[S^2\right] = \sigma^2$.

Reminder: if we know σ, we can derive a $100p\%$ CI for μ via the pivotal quantity

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{G}(0, 1).$$

If we don't know σ, we need a new result!

Idea: since the above pivotal quantity uses $\bar{Y}$ as an estimator of μ, what happens if we replace σ with its estimator S? Is

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

a pivotal quantity?

We need a new distribution: the **Student's t distribution**.

Definition 4.11 (Student's t Distribution).
A random variable T has a **Student's t distribution** if its p.d.f. is

$$f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2} \quad \text{for } t \in \mathbb{R} \text{ and } k = 1, 2, \ldots$$

where the constant c_k is given by

$$c_k = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})},$$

where $\Gamma(z) = \displaystyle\int_0^\infty x^{z-1} e^{-x}\, dx$.

Note. t distribution does not have real-world interpretation. It is constructed for doing more statistics... The parameter k is called the **degrees of freedom**, and we write $T \sim t_k$ or $T \sim t(k)$.

- For small k, the t distribution has larger 'tails' (more area in the tails).
- As k increases, t distribution approaches $\mathrm{G}(0, 1)$.

Theorem 4.9. Suppose $Z \sim \mathrm{G}(0, 1)$ and $U \sim \chi^2_k$ independently. Let

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

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

Example (Continued from above example).
If $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, then we can show that

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

Note that $\mathbb{E}[V] = \frac{(n-1)\mathbb{E}[S^2]}{\sigma^2} = n-1$ which aligns with the degrees of freedom. Putting this all together:

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{V/(n-1)}} \sim t_{n-1}.$$

So the distribution is completely known, and is therefore a pivotal quantity. We can use this to construct CI for μ without knowing σ!

A comparison of the processes:

σ known	σ unknown
Pivotal quantity: $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Pivotal quantity: $\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
1. Quantiles: $P\left(-a \leq \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \leq a\right) = p$	1. Quantiles: $P\left(-a \leq \frac{\overline{Y}-\mu}{S/\sqrt{n}} \leq a\right) = p$
2. Rearrange: $P\left(\overline{Y} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\overline{Y} - a\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + a\frac{S}{\sqrt{n}}\right) = p$
3. CI: $\left[\overline{y} - a\frac{\sigma}{\sqrt{n}}, \overline{y} + a\frac{\sigma}{\sqrt{n}}\right]$ $P(-a \leq Z \leq a) = p, Z \sim G(0,1)$	3. CI: $\left[\overline{y} - a\frac{s}{\sqrt{n}}, \overline{y} + a\frac{s}{\sqrt{n}}\right]$ $P(-a \leq T \leq a) = p, T \sim t_{n-1}$

Note. Differences: interchange σ with s and Z with T.

Example. If $P(-a \leq Z \leq a) = p$, then $P(Z \leq a) = \frac{1+p}{2}$. So, R commands for 90% and 95% CIs:

```
> qnorm(0.95) # 90% CI
> qnorm(0.975) # 95% CI
```

For t distributions, we must also specify the degrees of freedom. R commands for 90% and

95% CIs when $n = 30$:

```
> qt(0.95, 29) # 90% CI
> qt(0.975, 29) # 95% CI
```

Note.

- When sample size is small, the quantile from t distribution is larger than that from the Gaussian distribution.
- As sample size increases, the quantile from t distribution approaches that from the Gaussian distribution.

Lecture 13

Note. The width of a CI conveys a measure of uncertainty. Assuming all other parts of the problem remain unchanged:

- Increasing confidence level - wider.
- Increasing sample size - narrower.
- Decreasing standard deviation - narrower.
- Decreasing sample mean - same width.

Example (**Sample size calculation: Binomial data**).

Note. This should be included right before section 4.5.

From an earlier example, we have an approximate 95% CI for θ based on Binomial data:

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

If we want a 95% CI of width $\leq 2\ell$, then

$$2 \times \left(1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right) \leq 2\ell \implies n \geq \left(\frac{1.96}{\ell}\right)^2 \hat{\theta}(1-\hat{\theta}).$$

But we don't know $\hat{\theta}$ before we collect the data... So we take $\hat{\theta} = 0.5$ (make RHS the largest).

Example (Sample size calculation: Gaussian data).

For Gaussian data, we have two CI formulas, depending on whether or not we know σ:

$$\bar{y} \pm a\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{y} \pm a\frac{s}{\sqrt{n}}$$

where a is found from either G(0, 1) or t_{n-1} distribution. However, s depends on our sample, so we don't know it before we take our sample. Also, there is no 'worst case' value for s. Thus, we assume σ is known.

If we want our 95% CI $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ to have width $\leq 2\ell$, then we should choose n such that

$$1.96\frac{\sigma}{\sqrt{n}} \approx \ell \iff n \approx \left(\frac{1.96\sigma}{\ell}\right)^2.$$

In practice, since we usually don't know σ, we choose n larger than RHS and always round up to next integer (otherwise, CI will be wider).

Example (CI for σ^2). Suppose a random sample $Y_1, \dots Y_n$ from a G(μ, σ) distribution where μ, σ are both unknown, and we want to estimate σ. We established an estimator for σ^2:

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

and seen that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Since the distrubition is completely known, we have a pivotal quantity. We can use it to construct CIs for σ^2.

1. Find a, b such that

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = P(a \leq W \leq b) = p \quad \text{where } W \sim \chi^2_{n-1}.$$

   But W is not symmetric, so we can find a, b such that

$$P(W \leq a) = \frac{1-p}{2} \quad \text{and} \quad P(W > b) = \frac{1-p}{2} \iff P(W \leq b) = \frac{1+p}{2}.$$

2. Rearrange to get $P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) = p.$

3. For observed data y, the interval $[L(y), U(y)] = \left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$ is a $100p\%$ CI for σ^2, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

A summary for a $100p\%$ CI:

CI for μ, σ known	CI for σ^2
Pivotal quantity: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	Pivotal quantity: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
1. Quantiles: $P\left(-a \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le a\right) = p$	1. Quantiles: $P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = p$
2. Rearrange: $P\left(\bar{Y} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + a\frac{\sigma}{\sqrt{n}}\right) = p$	2. Rearrange: $P\left(\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right) = p$
3. CI: $\left[\bar{y} - a\frac{\sigma}{\sqrt{n}}, \bar{y} + a\frac{\sigma}{\sqrt{n}}\right]$ $P(-a \le Z \le a) = p$, $Z \sim G(0,1)$	3. CI: $\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right)$ $P(a \le W) = P(b \ge W) = (1-p)/2$, $W \sim \chi^2_{n-1}$

Note. The CI is not symmetric about s^2.

If we want a CI for σ instead, taking square roots to give $\left[\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right]$.

Approximate Confidence Intervals for Named Distributions based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Observed Data	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Asymptotic Gaussian Pivotal Quantity	Approximate $100p\%$ Confidence Interval
Binomial(n,θ)	y	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\bar{Y}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}}{n}}$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\bar{Y}$	$\frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}}$	$\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$

Note: The value a is given by $P(Z \le a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$. In R, $a =$ `qnorm`$(\frac{1+p}{2})$

Table 4.8.2
Confidence/Prediction Intervals for Gaussian and Exponential Models

Model	Unknown Quantity	Pivotal Quantity	$100p\%$ Confidence/Prediction Interval
$G(\mu,\sigma)$ σ known	μ	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	$\bar{y} \pm a\sigma/\sqrt{n}$
$G(\mu,\sigma)$ σ unknown	μ	$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{y} \pm bs/\sqrt{n}$
$G(\mu,\sigma)$ μ unknown σ unknown	Y	$\frac{Y-\overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$	$100p\%$ Prediction Interval $\bar{y} \pm bs\sqrt{1+\frac{1}{n}}$
$G(\mu,\sigma)$ μ unknown	σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)s^2}{d}, \frac{(n-1)s^2}{c}\right]$
$G(\mu,\sigma)$ μ unknown	σ	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}\right]$
Exponential(θ)	θ	$\frac{2n\overline{Y}}{\theta} \sim \chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1}, \frac{2n\bar{y}}{c_1}\right]$

Notes: (1) The value a is given by $P(Z \le a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$. In R, $a = \texttt{qnorm}(\frac{1+p}{2})$.

(2) The value b is given by $P(T \le b) = \frac{1+p}{2}$ where $T \sim t(n-1)$. In R, $b = \texttt{qt}(\frac{1+p}{2}, n-1)$.

(3) The values c and d are given by $P(W \le c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n-1)$. In R, $c = \texttt{qchisq}(\frac{1-p}{2}, n-1)$ and $d = \texttt{qchisq}(\frac{1+p}{2}, n-1)$.

(4) The values c_1 and d_1 are given by $P(W \le c_1) = \frac{1-p}{2} = P(W > d_1)$ where $W \sim \chi^2(2n)$. In R, $c_1 = \texttt{qchisq}(\frac{1-p}{2}, 2n)$ and $d_1 = \texttt{qchisq}(\frac{1+p}{2}, 2n)$.

5 Hypothesis Testing

5.1 Introduction

Null hypothesis H_0: A single 'default' hypothesis.

Alternative hypothesis H_A: Alternative to null (in most cases H_A means that H_0 is not true).

Example. A professor claims that he randomizes the option labels so it's equally likely that a question's answer will be A or B in an exam. But a student thinks the prof is lying. In one of the quizzes with 25 questions, 10 had the answer A and 15 had the answer B.
Let Y = the number of questions with the answer 'A' out of a randomly chosen 25 questions, so $Y \sim \text{Binomial}(25, 0.5)$. Then

- What should the null hypothesis be?
  H_0: $\theta = 0.5$.
- What observed values y are consistent with H_0?
  Values close to $\mathbb{E}[Y] = 12.5$.
- What observed values y provide evidence against H_0?
  Values close to 0 or 25.
- How do we measure the strength of the evidence against H_0?
  The evidence against H_0 increases as y gets further from 12.5.

We need a way to quantify the strength of evidence against H_0.

Definition 5.1 (Test Statistic/Discrepancy Measure).
A **test statistic** or **discrepancy measure** is a function of the data $D = g(Y)$ that is constructed to measure the degree of 'agreement' between the data Y and the null hypothesis H_0.

Remark. In other words: how much does what we observed align with what we would expect if H_0 was true?

Note. D is a function of Y, so is also a random variable. Once we observe $Y = y$, then the observed value of D is $d = g(y)$. Larger values of d indicate poorer agreement between the data and H_0 (i.e. stronger evidence against H_0).

Example (Continued from above). We define the discrepancy measure

$$D(Y) = |Y - 12.5|.$$

Lecture 14

Definition 5.2 (***p*-value**)**.** 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** of the test of hypothesis H_0 using test statistic D is $P(D \geq d; H_0)$.

Note. In other wods, the p-value is the probability of observing a value of the test statistic greater than or equal to the observed value of the test statistic assuming H_0 is true. It can help to think of p-values as a measure of 'surprise' assuming H_0 is true.
A small p-value indicates that if H_0 were true, it would be unlikely to have observed data **at least as surprising** as the data we actually observed (strong evidence against H_0).

Steps of a statistical test of a hypothesis:

1. Specify H_0 to be tested using data Y.
2. Define a test statistic $D(Y)$, for which larger values of D are less consistent with H_0. Let $d = D(y)$ be the corresponding observed value of D.
3. Calculate p-value $= P(D \geq d; H_0)$.
4. Draw a conclusion based on the p-value.

If $d = D(y)$ is large, and consequently the p-value is small, then one of the following must be true:

1. H_0 is true but we observed an event that does not happen very often when H_0 is true.
2. H_0 is false.

$p - value$	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_0based 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.

Figure 12: Guidelines for interpreting p-values.

Example (Test of hypothesis for Binomial for large n**).**

Suppose the prof's final exam has 100 questions and 40 questions had the answer 'A'. The p-value for the 25-question case was 0.424, what about the 100-question case?

Following the steps:

1. Let Y = the number of 'A' answers, assuming $Y \sim \text{Binomial}(n, \theta)$, then $H_0 : \theta = 0.5$.
2. Define $D(Y) = |Y - 50|$.
3. $d = |40 - 50| = 10$ here, so $P(D \geq 10; H_0) = P(|Y - 50| \geq 10) = P(Y \leq 40) + P(Y \geq 60)$. Using R for exact calculation:

```
sum(dbinom(c(0:40), 100, 0.5)) + sum(dbinom(c(60:100), 100, 0.5))
```

We get p-value $= 0.0569$ (exact hypothesis test). We can also use the Normal approximation to the Binomial. Remember: suppose $Y \sim \text{Binomial}(n, \theta)$, if n is large then by CLT, $\frac{Y - n\theta}{\sqrt{n\theta(1-\theta)}} \sim \text{G}(0.1)$ approximately.

$$
\begin{aligned}
p\text{-value} &= P(|Y - 50| \geq 10) \\
&= P\left(\frac{|Y - 50|}{\sqrt{100(0.5)(0.5)}} \geq \frac{10}{\sqrt{100(0.5)(0.5)}}\right) \\
&\approx 2P(Z \geq 2) = 2\left(1 - P(Z \leq 2)\right) = 0.04550.
\end{aligned}
$$

4. Note that both exact and approximated p-values are very close, so they provide some evidence against H_0 (nothing special about 0.05).

A larger sample with the same sample proportion led to a smaller p-value, and thus stronger evidence against H_0.

5.2 Hypothesis Testing for Parameters in the Gaussian Model

As an example, a 'large' tea at Tim Hortons should have 590ml of tea. Suppose the company wanted to check a store was serving the correct amount. What model might we use, what hypothesis might we test?

Suppose $Y \sim G(\mu, \sigma)$ denotes the volume of tea in a random chosen cup. We might test the null hypothesis $H_0 : \mu = 590$.

Intuition: suppose the mean tea per cup acrooss our sample was $\bar{y}$.

If $|\bar{y} - 590|$ is large, that would indicate our sample was inconsistent with H_0. This suggests a discrepancy measure $D(Y) = |\bar{Y} - 590|$. Suppose we observe a sample of size n cups with sample mean $\bar{y} = 595$. Then $d = |\bar{y} - 590| = 5$. Then

$$p\text{-value} = P(D \geq 5; H_0) = P(|\bar{Y} - 590| \geq 5) = P(\bar{Y} \geq 595) + P(\bar{Y} \leq 585).$$

If $Y_i \sim G(\mu, \sigma)$, then $\bar{Y} \sim G(\mu, \sigma/\sqrt{n})$. So now we have $\bar{Y} \sim G(590, \sigma/\sqrt{n})$. More generally, for testing $H_0 : \mu = \mu_0$, we use $D = |\bar{Y} - \mu_0|$ as the discrepancy measure, and the p-value is

$$P(|\bar{Y} - \mu_0| \geq |\bar{y} - \mu_0|).$$

Example (Testing $H_0 : \mu = \mu_0$, σ unknown).
Suppose $Y_i \sim G(\mu, \sigma)$ for $i = 1, \ldots, n$. Recall the pivotal quantity

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

which was used to construct CI for μ without knowing σ. To test $H_0 : \mu = \mu_0$, we use the test statistic:

$$D = \frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}}.$$

If we observe $D = d = \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}$, then

$$p\text{-value} = P(D \geq d; H_0) = P\left(\frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \geq d\right) = P(|T| \geq d) = 2(1 - P(T \leq d))$$

where $T \sim t_{n-1}$.

Example. Male quolls have a mean weight of 1kg. An ecologist took a sample of 10 quolls to see if this mean weight appears accurate. Let $y_1, \dots, y_{10}$ denote the 10 measurements.

1.026	0.998	1.017	1.045	0.978
1.004	1.018	0.965	1.010	1.000

We assume $Y_i \sim G(\mu, \sigma)$ independently. The hypothesis of interest is $H_0 : \mu = 1$. For these data

$$\bar{y} = 1.0061 \quad \mu_0 = 1 \quad s = 0.023 \quad n = 10.$$

Then $d = \frac{|\bar{y}-\mu_0|}{s/\sqrt{n}} = \frac{|1.0061-1|}{0.023/\sqrt{10}} = 0.8378$. Our p-value is

$$P(D \geq 0.8378) = P(|T| \geq 0.8378) = 2(1 - P(T \leq 0.8378))$$

where $T \sim t_9$. Using R:

```
> pt(0.8378, 9)
[1] 0.7880836
> 2*(1 - pt(0.8378, 9))
[1] 0.4238328
```

We can also use t.test() in R to compute this:

```
> y <- c(1.026, 0.998, 1.017, 1.045, 0.978,
1.004, 1.018, 0.965, 1.010, 1.000)
> t.test(y, mu = 1)
t = 0.83782, df = 9, p-value = 0.4238
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
0.9896297 1.0225703
sample estimates:
mean of x
1.0061
```

Lecture 15

The p-value just tells us how surprised we'd be by the data if H_0 were true (does not imply how wrong H_0 is). A CI indicates the magnitude and direction of the departure form H_0.

Statistical significance: is used to describe when a hypothesis test returns a small p-value.

Practical significance: is used to describe when our results have important 'real-world' implications, such as future decision-making.

Example (Relationship between Hypothesis Testing and Confidence Intervals). Suppose we test $H_0 : \mu = \mu_0$ for $G(\mu, \sigma)$, then

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

In other words:

- If p-value ≥ 0.05, then μ_0 is in side the 95% CI for μ.
- If μ_0 is inside a 95% CI for μ, then p-value ≥ 0.05.

Remark. More generally, suppose we use the same pivotal quantity to construct a CI for θ and a test of the hypothesis $H_0 : \theta = \theta_0$. Then $\theta = \theta_0$ is inside a $100q\%$ CI for $\theta \iff$ the p-value of the test of H_0 is $\geq 1 - q$. Also, this result only approximately holds if we use different pivotal quantities.

Example (Testing $H_0 : \sigma^2 = \sigma_0^2$, μ is unknown).

Suppose $Y_i \sim G(\mu, \sigma)$. Recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. To test $H_0 : \sigma^2 = \sigma_0^2$, we use the test statistic

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

If H_0 is true, then $U \sim \chi^2_{n-1}$. To test H_0:

1. Draw a sample of size n with sample variance s^2.
2. Compute test statistic $u = \frac{(n-1)s^2}{\sigma_0^2}$.
3. Compute $P(U \leq u)$ for $U \sim \chi^2_{n-1}$.
4. If $P(U \leq u) < 0.5$, then p-value $= 2P(U \leq u)$.
5. If $P(U \geq u) < 0.5$, then p-value $= 2P(U \geq u)$.

5.3 Likelihood Ratio Test of Hypothesis

We can use the likelihood ratio statistic $\Lambda(\theta) = -2\log\left(\frac{L(\theta)}{L(\tilde{\theta})}\right)$ to derive a more general method for conducting hypothesis tests.

Principles of constructing a test statistic:

1. If H_0 is true, we know the (approximate) distribution of the test statistic.

   For a general test of $H_0 : \theta = \theta_0$, we might consider

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

   because if H_0 is true, then $\Lambda(\theta_0) \sim \chi^2_1$.

2. If the data are inconsistent with H_0, the observed value of test statistic will be large.

   For a specific example, we can calculate the observed value

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

Note. The test statistic $\lambda(\theta_0)$ is larger for data that are more surprising if $\theta = \theta_0$. Then

$$\begin{aligned} p\text{-value} &\approx P(W \geq \lambda(\theta_0)) \quad \text{where } W \sim \chi^2_1 \\ &= P\left(|Z| \geq \sqrt{\lambda(\theta_0)}\right) = 2\left[1 - P\left(Z \leq \sqrt{\lambda(\theta_0)}\right)\right] \quad \text{where } Z \sim G(0,1). \end{aligned}$$

In summary, process for using the likelihood ration statistic for a test of $H_0 : \theta = \theta_0$:

1. Propose a model for the data and from $L(\theta)$. Use this to derive an expression for $\hat{\theta}$.
2. Gather data and calculate $\hat{\theta}$ for the observed data.
3. Compute the observed value of the test statistic $\lambda(\theta_0) = -2\log(R(\theta_0))$.
4. The p-value $\approx P(W \geq \lambda(\theta_0))$ where $W \sim \chi^2_1$.

Example. Too lazy to include an example here.

Table 5.5.1
Hypothesis Tests for Named Distributions based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Test Statistic for $H_0 : \theta = \theta_0$	Approximate $p - value$ based on Gaussian approximation
Binomial(n, θ)	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$ $Z \sim G(0,1)$
Poisson(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}\right)$ $Z \sim G(0,1)$
Exponential(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}\right)$ $Z \sim G(0,1)$

Note: Note: To find $2P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $2 * (1-$ `pnorm`$(d))$.

Table 5.5.2
Hypothesis Tests for Gaussian and Exponential Models

Model	Hypothesis	Test Statistic	Exact $p-value$
$G(\mu,\sigma)$ σ known	$H_0: \mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{\sigma/\sqrt{n}}$	$2P\left(Z \ge \frac{\lvert\bar{y}-\mu_0\rvert}{\sigma/\sqrt{n}}\right)$ $Z \sim G(0,1)$
$G(\mu,\sigma)$ σ unknown	$H_0: \mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{S/\sqrt{n}}$	$2P\left(T \ge \frac{\lvert\bar{y}-\mu_0\rvert}{s/\sqrt{n}}\right)$ $T \sim t(n-1)$
$G(\mu,\sigma)$ μ unknown	$H_0: \sigma=\sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$	$\min(2P\left(W \le \frac{(n-1)s^2}{\sigma_0^2}\right),$ $2P\left(W \ge \frac{(n-1)s^2}{\sigma_0^2}\right))$ $W \sim \chi^2(n-1)$
Exponential(θ)	$H_0: \theta=\theta_0$	$\frac{2n\bar{Y}}{\theta_0}$	$\min(2P\left(W \le \frac{2n\bar{y}}{\theta_0}\right),$ $2P\left(W \ge \frac{2n\bar{y}}{\theta_0}\right))$ $W \sim \chi^2(2n)$

Notes: (1) To find $P(Z \ge d)$ where $Z \sim G(0,1)$ in R, use $1 - \texttt{pnorm}(d)$.

(2) To find $P(T \ge d)$ where $T \sim t(k)$ in R, use $1 - \texttt{pt}(d,k)$.

(3) To find $P(W \le d)$ where $W \sim \chi^2(k)$ in R, use $\texttt{pchisq}(d,k)$.

6 Gaussian Response Models

Lecture 16

6.1 Introduction

We are often interested in studying **bivariate** data, where two variates are measured per unit. Recall from chapter 1:

Sample correlation: $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ with $-1 \le r \le 1$, where

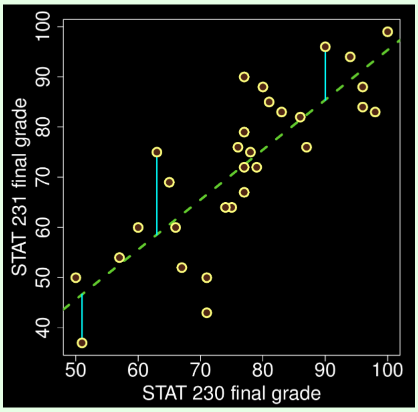
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Correlation: $\operatorname{corr}(X,Y) = \dfrac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} = \dfrac{\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y}$.

Example (**STAT 230/231 Final Grades**).

For final grades of 30 students who took both STAT 230 and STAT 231, we find $r = 0.82$ (strong positive linear relationship), see data in Figure 16. We want to fit a stright line to these data on a scatterplot.

Consider the vertical distances between a fitted line and the data. These are called **residuals**.



We want to understand the relationship between STAT 230 and STAT 231 grades in terms of how much of the variation in STAT 231 grades is ‘explained’ by variation in STAT 230 grades.

6.2 Simple Linear Regression

Example (**Continued from above**).

Goal: find the fitted line $y = \alpha + \beta x$ which minimizes some function of the distances between the observed points and the fitted line.

Note. There are many options for what function to choose!

It is conventional to find $y = \alpha + \beta x$ which minimizes the sum of squares of the residuals.

Least Squares Estimates

Estimates α and β (denoted $\hat{\alpha}$ and $\hat{\beta}$) are called the **least squares estimates**.

We can define the residual for each pair of points (x_i, y_i) as $r_i = y_i - (\alpha + \beta x_i)$. To find the least estimates $\hat{\alpha}$ and $\hat{\beta}$, we minimize the sum of squares of residuals:

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

We need to solve the following two equations after differentiation to find $\hat{\alpha}$ and $\hat{\beta}$:

$$\frac{\partial g}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \tag{6.1}$$

$$\frac{\partial g}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0. \tag{6.2}$$

We can rewrite (6.1) as: $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$ and (6.2) as: $\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0$. Substitute $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ into (6.2):

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

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

We can rearrange to get:

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

Note. $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) - \bar{x}\sum(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) - \bar{x}\underbrace{[n\bar{y} - n\bar{y}]}_{=0}$.

Thus, our least squares estimates of α and β are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

Relationship between sample correlation and the slope of the least squares regression line:

$$\hat{\beta} = r\sqrt{\frac{S_{yy}}{S_{xx}}}.$$

Note.

- $\hat{\beta}$ and r have the same sign.
- If $S_{yy} >> S_{xx}$, then $\hat{\beta}$ is much larger than r, vice-versa.
- $S_{yy} = (n-1)\operatorname{Var}(y)$ where $\operatorname{Var}(y)$ is the sample variance of y. Similarly, $S_{xx} = (n-1)\operatorname{Var}(x)$.

Example (Finding Best Fit in R).

Continued from STAT 230/231 example above.

```
> lm(y ~ x)
```

Note. $y \sim x$ means y is the response variate and x is the explanatory variate. The output is below:

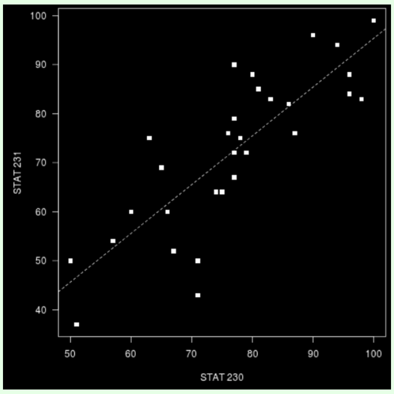
```
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
    -4.0667     0.9944
```

Commonly, we create a model object and use `summary()`:

```
> mod <- lm(y ~ x)
> summary(mod)
Call:
lm(formula = y ~ x)
Residuals:
Min 1Q Median 3Q Max
-23.5324 -6.2104 0.9704 4.5820 17.5015
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.0667 10.2785 -0.396 0.695
x             0.9944 0.1320 7.530 3.34e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.463 on 28 degrees of freedom
Multiple R-squared: 0.6695, Adjusted R-squared: 0.6576
F-statistic: 56.71 on 1 and 28 DF, p-value: 3.342e-08
```

We can plot the data and the fitted line:

```
> plot(x, y, xlab = 'STAT 230', ylab = 'STAT 231', pch = 15)
> abline(coef(mod), lty = 2)
```



*** Our least squares line is $y = -4.0667 + 0.9944x$ in the form of $y = \hat{\alpha} + \hat{\beta}x$.

How to measure uncertainty? Perhaps an likelihood extimate!

Generally, we need a model that models the variability in final grades for each STAT 230 final grade x. Let Y_x denote the STAT 231 grade of a randomly chosen student who got a grade of x in STAT 230, and assume $Y_x \sim \mathrm{G}(\mu_x, \sigma_x)$. Then for each STAT 230 grade x in our sample, use that subset of students to estimate μ_x, σ_x using techniques of estimation.

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

Note. We are NOT using the linear relationship that we established in the assumed model.

We assumed a model in which μ_x = mean STAT 231 final grade for students in the study population who obtained a final grade of x in STAT 230. It takes a linear form in x:

$$\mu_x = \alpha + \beta x.$$

We also assume σ_x is the same for all x, and denote it by σ. For data $(x_i, y_i), i = 1, 2, \dots, n$ we therefore assume $Y_i \sim \mathrm{G}(\alpha + \beta x_i, \sigma)$. This model is called a **simple linear regression**.

Interpretation of α, β, σ

α: mean value of the response variate in the study population of individuals for whom the explanatory variate is zero.

β: increase in the mean value of the response variate in the study population for a one unit increase in the value of the explanatory variate.

σ: the variability in the response variate Y in the study population, and does not vary with x.

Example.

α: the mean STAT 231 final grade in the study population with a STAT 230 final grade 0.

Note. Parameters may not always have real-world interpretations. In this case, students who get a 0 in STAT 230 cannot take STAT 231.

β: the increase in the mean STAT 231 final grade in the study population for a one mark increase in STAT 230 final grade.

Note. This is the same regardless of the value of x.

σ: the variability in STAT 231 final grades for the study population, and does not vary with x.

Likelihood Function for α and β

We have $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently, where x_i's are constants. Then,

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2}.$$

Assume σ is known, then the term $\dfrac{1}{\sqrt{2\pi}\sigma}$ is a constant, and simply write:

$$L(\alpha, \beta) = e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}.$$

To obtain MLE (maximize $L(\alpha, \beta)$), we need to minimize $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$, which is the sum of squared residuals. Then

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

Lecture 17

Interval Estimation

For our linear model, we have a sample $(x_1, y_1), \ldots, (x_n, y_n)$ and assume $Y_i \sim G(\alpha + \beta x_i, \sigma)$. We have point estimates $\hat{\beta}$ and $\hat{\alpha}$ for β and α. For β:

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

A corresponding estimator is $\tilde{\beta} = \underbrace{\frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})}_{\text{constant}} Y_i = \sum_{i=1}^{n} a_i Y_i$ where $a_i = \dfrac{x_i - \bar{x}}{S_{xx}}$. This is a linear combination of Gaussian random variable Y_i.

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\beta}\right] = \sum_{i=1}^{n} a_i \mathbb{E}\left[Y_i\right] = \sum_{i=1}^{n} a_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} a_i + \beta \sum_{i=1}^{n} a_i x_i = \beta \sum_{i=1}^{n} a_i x_i = \beta. \\
&\operatorname{Var}\left(\tilde{\beta}\right) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 = \frac{\sigma^2}{S_{xx}}. \\
&\implies \tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).
\end{aligned}$$

Proposition 6.1 (**Distribution of $\tilde{\beta}$**). If $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, \ldots, n$ independently, where x_i's are constants. Then the least squares estimator of β has the following distribution:

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).$$

Confidence Interval for β

Since $\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$, then $\frac{\tilde{\beta} - \beta}{\sigma/\sqrt{S_{xx}}} \sim G(0,1)$ is a pivotal quantity if σ is known. We know that if $Y \sim G(\mu, \sigma)$ and σ is unknown, we can use the pivotal quantity $\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$. Could we do something similar here for σ?

Example (**Estimate σ^2 in Simple Linear Regression**).

We estimate σ^2 using the following:

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

Note. The quantity $\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$ is called the **sum of squared errors** and s_e^2 is called the **mean squared error**.

We define the estimator $S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} x_i)^2$, where $\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$ and $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$. Then, $\mathbb{E}\left[S_e^2\right] = \sigma^2$. If S_e^2 is an estimator of the mean squared error s_e^2, then

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

Since $\frac{\tilde{\beta} - \beta}{\sigma/\sqrt{S_{xx}}} \sim G(0,1)$ and $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$ independently, then by Theorem 4.9, we have

$$\frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{S_{xx}}}} \sim t_{n-2}.$$

This pivotal quantity will be used to construct CIs for β and tests of hypotheses about β.

1. Quantiles: we can find a such that $P(-a \leq T \leq a) = p$ for $T \sim t_{n-2}$, so

$$p = P\left(-a \leq \frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{S_{xx}}}} \leq a\right).$$

2. Rearrange: isolate β

$$p = P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \leq \beta \leq \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right).$$

3. CI: a $100p\%$ CI for β is

$$\left[\hat{\beta} - a\frac{s_e}{\sqrt{S_{xx}}}, \hat{\beta} + a\frac{s_e}{\sqrt{S_{xx}}}\right]$$

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

- If s_e increases, then the CI will be wider. In other words, greater variability in our response variates leads to greater uncertainty in our estimates.
- If S_{xx} increases, then the CI will be narrower. This means that if we have larger variability in x, then our estimate of β is based on a more variable set of values of x. This reduces uncertainty in our estimate of β as it is based on a broader range of x values

Example (95% **CI for** β **in STAT 230/231 Grades Data**).

We need $\hat{\beta} \pm a\frac{s_e}{\sqrt{S_{xx}}}$ and a such that $P(T \leq a) = 0.975$, where $T \sim t_{28}$. Using R:

```
> qt(0.975, 28)
[1] 2.048407
```

Thus, a 95% CI for β is $[0.724, 1.265]$. We are 95% confident that β is in this interval.

Example (**Where do the degrees of freedom come from?**).

For $Y_1, \ldots, Y_n \sim \mathrm{G}(\mu, \sigma)$ data, we used

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

and a pivotal quantity with $n-1$ degrees of freedom.

For our linear model, we used

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

and a pivotal quantity with $n-2$ degrees of freedom.

Degrees of freedom arise from constraints placed on our estimation process.

Note. Constraint is when you let something equal to something else.

For example, if we solve $\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0$, we have a constraint and we lose a degree of freedom ($n-1$ now).

Another example: in our linear regression model, in estimating α, β, we had two constraints:

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

These 2 equations in 2 unknowns means two constraints (so $n-2$ degrees of freedom).

We will discuss degress of freedom in more details in Chapter 7!

Example (Hypotheses Tests about β).

Referring to Example 5.2, we can draw a similar conclusion about β. The p-value for testing $H_0 : \beta = \beta_0$ is

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

Note. Make sure you can derive this from scratch.

Hypothesis of No Relationship

Since $\mu_x = \alpha + \beta x$, a test of $H_0 : \beta = 0$ is a test of the hypothesis that μ_x does not (linearly) depend on x. This hypothesis is referred to as the **hypothesis of no linear relationship** between the variates Y and x.

Example (STAT 230/231 Grades Example Continued).

For our STAT 230/231 data, a test of $H_0 : \beta = 0$ is a test of no linear relationship between STAT 230 and STAT 231 grades. We've seen that a 95% CI for β is $[0.724, 1.265]$, so we already know p-value < 0.05.

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

Note that

```
> pt(7.5304, 28)
[1] 1
```

Important: we cannot say p-value $= 0$, this is almost never the case! A p-value of 0 means the observed data are impossible under H_0. We write $p \approx 0$ and conclude a very strong evidence against the hypothesis of no linear relationship.

We can see the test of $H_0 : \beta = 0$ in our R output:

```
> mod <- lm(y ~ x)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.0667    10.2785  -0.396    0.695
x             0.9944     0.1320   7.530 3.34e-08 ***
----
Residual standard error: 9.463 on 28 degrees of freedom
```

And we can get the confidence intervals from R too!

```
> confint(mod, level = 0.95)
                  2.5 %    97.5 %
(Intercept) -25.1211911 16.987750
x             0.7238725  1.264834
```

Note that $s_e = 9.463$, the '(Intercept)' row corresponds to α and the 'x' row corresponds to β, so $\hat{\alpha} = -4.0667$ and $\hat{\beta} = 0.9944$. The p-value for $H_0 : \beta = 0$ is 3.34×10^{-8}. Also, the p-value for $H_0 : \alpha = 0$ is 0.695. Our test statistic is 7.530 (7.5304 previously due to rounding). Note that the "Estimate" column divided by the "Std. Error" column gives the test statistic (t value), thus standard error $= \frac{s_e}{\sqrt{S_{xx}}}$ corresponds to the 'x' row.

Example (**Exercise**). We may also be interested in testing whether β is some other value. Suppose we want to test $H_0 : \beta = 1$. We know the formula for p-value is

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

1. A 95% CI for β was $[0.7239, 1.2648]$: what does this tell us about the p-value resulting from such a hypothesis test?

   Ans: we know p-value > 0.05.
2. What R command can we use to help us find the p-value?

   Ans: $2 * (1 - \text{pt}(c, 28))$, where $c = \frac{|\hat{\beta}-\beta_0|}{s_e/\sqrt{S_{xx}}} = 0.0424$.
3. What is the resulting p-value for our sample?

   Ans: about 0.966.
4. What is our conclusion about the hypothesis?

   Ans: there is no evidence against the hypothesis that $\beta = 1$.

Confidence Intervals for mean response $\mu_x = \alpha + \beta x$

We have a point estimate for μ_x:

$$\hat{\mu}_x = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$$

and so this has a corresponding estimator: $\tilde{\mu}_x = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x})$. We need the distribution of $\tilde{\mu}_x$ to construct a CI for the mean response $\mu_x = \alpha + \beta x$.

Since $\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$ and $Y_i \sim \text{G}(\alpha + \beta x_i, \sigma)$, we have

$$\begin{aligned}
\tilde{\mu}_x &= \bar{Y} + \tilde{\beta}(x - \bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})(x - \bar{x})Y_i \\
&= \sum_{i=1}^{n} \underbrace{\left(\frac{1}{n} + (x - \bar{x})\frac{x_i - \bar{x}}{S_{xx}}\right)}_{\text{constant}} Y_i
\end{aligned}$$

We write $\tilde{\mu}_x = \sum_{i=1}^{n} b_i Y_i$ where $b_i = \frac{1}{n} + (x - \bar{x})\frac{x_i - \bar{x}}{S_{xx}}$. Then,

$$\mathbb{E}\left[\tilde{\mu}_x\right] = \sum_{i=1}^{n} b_i \mathbb{E}\left[Y_i\right] = \sum_{i=1}^{n} b_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} b_i + \beta \sum_{i=1}^{n} b_i x_i.$$

Note that

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

Also,

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

Note that $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i(x_i - \bar{x})$ is a constant (not depending on the i here). Thus, $\mathbb{E}\left[\tilde{\mu}_x\right] = \alpha + \beta x = \mu_x$. For the variance,

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\mu}_x\right) = \sum_{i=1}^{n} b_i^2 \,\mathrm{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} b_i^2 &= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} + (x - \bar{x})\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{2(x - \bar{x})}{n} \cdot \frac{x_i - \bar{x}}{S_{xx}} + \left[(x - \bar{x})\frac{x_i - \bar{x}}{S_{xx}}\right]^2\right) \\
&= \sigma^2 \left(\frac{1}{n} + 0 + \frac{(x - \bar{x})^2}{S_{xx}^2} \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{S_{xx}}\right) \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right].
\end{aligned}$$

Proposition 6.2 (Distribution of $\tilde{\mu}_x$).
For $\tilde{\mu}_x = \tilde{\alpha} + \tilde{\beta} x = \sum_{i=1}^{n} \left(\frac{1}{n} + (x - \bar{x})\frac{x_i - \bar{x}}{S_{xx}}\right) Y_i$, where $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, \ldots, n$,

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

where $\mu_x = \alpha + \beta x$.

Equivalently, $\dfrac{\tilde{\mu}_x - \mu_x}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0, 1)$. By Theorem 4.9, we have

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

We can compare with the CI for β to obtain a $100p\%$ CI for $\mu_x = \alpha + \beta x$:

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

where $\hat{\mu}_x = \hat{\alpha} + \hat{\beta}x$, a is such that $P(T \leq a) = \frac{1+p}{2}$ and $T \sim t_{n-2}$.

Remark. Consider factors that change the width of the CI for μ_x:

- n: width decreases as n increases.
- x: width increases as $(x - \bar{x})^2$ increases.
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$: width decreases as S_{xx} increases.

Note that $\alpha = \mu(0)$, a $100p\%$ CI for α is:

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

Remark. If $\bar{x}$ is large (meaning x_i are typically large), then the CI for α will be very wide. However, it is somehow meaningless if $x = 0$.

Example. After calculation, a 95% CI for the mean STAT 231 grade of the population of students who score a 75 in STAT 230 is $[66.9, 74.1]$.
One way to interpret this: We are 95% confident that, in the population of students who receive a grade of 75 in STAT 230, the mean grade in STAT 231 will be between 66.9 and 74.1.
Important: the word"population" is key here. We are estimating the mean response for a population of students, not just one individual!

We can use R to compute the above CI:

```
> predict(mod, data.frame(x = 75), interval = 'confidence')
       fit      lwr      upr
1 70.50979 66.93985 74.07973
```

where 70.50979 is equal to $\hat{\mu}_{75}$, 66.93985 is the lower bound, and 74.07973 is the upper bound.

Remark. The `predict()` function can calculate more than one type of interval.

Lecture 18

Confidence Interval for an Individual Response Y at x

The previous result is for the mean STAT 231 final grade for the population of students who scored a 75 in STAT 230. Now, we consider the CI for $Y =$ the STAT 231 final grade for a one student who scored a 75 in STAT 230.

Let $Y =$ potential observation for given value of x. We have

$$Y = \mu_x + R, \quad \text{where } R \sim \mathrm{G}(0, \sigma)$$

independent of $Y_1, \dots, Y_n$. Note that $Y \sim \mathrm{G}(\mu_x, \sigma)$. We now want the distribution of $Y - \tilde{\mu}_x$, the error in the point estimator of Y:

$$Y - \tilde{\mu}_x = Y - \mu_x + \mu_x - \tilde{\mu}_x = R + [\mu_x - \tilde{\mu}_x].$$

Remark. R is independent of $\tilde{\mu}_x$ because it is not connected to the existing sample. Thus, the above is a sum of independent, normally distributed random variables.

Note that

$$\begin{aligned}
\mathbb{E}[Y - \tilde{\mu}_x] &= \mathbb{E}[R] + \mathbb{E}[\mu_x - \tilde{\mu}_x] = 0 + \mu_x - \mu_x = 0. \\
\operatorname{Var}(Y - \tilde{\mu}_x) &= \operatorname{Var}(Y) + \operatorname{Var}(\tilde{\mu}_x) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right].
\end{aligned}$$

Proposition 6.3 (**Distribution of** $Y - \tilde{\mu}_x$).

For $Y - \tilde{\mu}_x = R + [\mu_x - \tilde{\mu}_x]$, where $Y = \mu_x + R$ and $R \sim \mathrm{G}(0, \sigma)$, the error

$$Y - \tilde{\mu}_x \sim \mathrm{G}\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right).$$

Equivalently, $\dfrac{Y - \tilde{\mu}_x}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim \mathrm{G}(0, 1)$. Since we don't know σ, by Theorem 4.9, we have

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

The resulting interval estimate 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 a is such that $P(T \leq a) = \frac{1+p}{2}$ and $T \sim t_{n-2}$. This interval is called a $100p\%$ **prediction interval** for Y inestead of a CI, because here Y is not a parameter but a "future" observation.

To recap we first found a $100p\%$ confidence interval for μ_x:

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

And now we a $100p\%$ prediction interval for future observation Y:

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

The only difference is that extra '1' inside the square root.

Note. For above CI, if we have all the data, there will be no uncertainty, i.e. as $n \to \infty$, width of CI $\to 0$. For PI, there will always be uncertainty, i.e. as $n \to \infty$, width of PI $\not\to 0$.

Example. Previously, we saw that a 95% CI for the mean STAT 231 grade for the population of students who scored a 75 in STAT 230 was [66.9, 74.1]. The corresponding prediction interval is [50.8, 90.2], which is much wider than the CI. This intuitively makes sense, we have all the uncertainty in trying to estimate the mean population, plus the additional uncertainty in that prediction step.

We can use `predict()` again to compute this PI:

```
> predict(mod, data.frame(x = 75), interval = 'prediction')
       fit      lwr      upr
1 70.50979 50.79979 90.21979
```

Exercise: Show that a 95% PI for someone who got a 60 in STAT 230 is [35.4, 75.8].

Table 6.2.2
Hypothesis Tests for
Simple Linear Regression Model

Hypothesis	Test Statistic	$p - value$
$H_0 : \beta = \beta_0$	$\frac{\lvert\tilde{\beta}-\beta_0\rvert}{S_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\lvert\hat{\beta}-\beta_0\rvert}{s_e/\sqrt{S_{xx}}}\right)$ where $T \sim t(n-2)$
$H_0 : \alpha = \alpha_0$	$\frac{\lvert\tilde{\alpha}-\alpha_0\rvert}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$ where $T \sim t(n-2)$
$H_0 : \sigma = \sigma_0$	$\frac{(n-2)S_e^2}{\sigma_0^2}$	$\min\left(2P\left(W \leq \frac{(n-2)s_e^2}{\sigma_0^2}\right), 2P\left(W \geq \frac{(n-2)s_e^2}{\sigma_0^2}\right)\right)$ $W \sim \chi^2(n-2)$

Table 6.2.1

Confidence/Prediction Intervals for Simple Linear Regression Model

Unknown Quantity	Estimate	Estimator	Pivotal Quantity	100p% Confidence/ Prediction Interval
β	$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$	$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$	$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$	$\hat{\beta} \pm a s_e/\sqrt{S_{xx}}$
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	$\tilde{\alpha} = \overline{Y} - \tilde{\beta}\bar{x}$	$\frac{\tilde{\alpha}-\alpha}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\alpha} \pm a s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}$
$\mu(x) = \alpha + \beta x$	$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$	$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$	$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\mu}(x) \pm a s_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$s_e^2 = \frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}$	$S_e^2 = \frac{\sum_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)^2}{n-2}$	$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$	$\left[\frac{(n-2)s_e^2}{c}, \frac{(n-2)s_e^2}{b}\right]$
Y	$\hat{Y} = \hat{\alpha} + \hat{\beta}x$		$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	Prediction Interval $\hat{\mu}(x) \pm a s_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$

Notes: The value a is given by $P(T \le a) = \frac{1+p}{2}$ where $T \sim t(n-2)$.
The values b and c are given by $P(W \le b) = \frac{1-p}{2} = P(W > c)$ where $W \sim \chi^2(n-2)$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat{\alpha}$	$s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}$	$\frac{\hat{\alpha}-\alpha_0}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$
x	$\hat{\beta}$	$s_e/\sqrt{S_{xx}}$	$\frac{\hat{\beta}-\beta_0}{s_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\lvert\hat{\beta}-\beta_0\rvert}{s_e/\sqrt{S_{xx}}}\right)$

Note. This tables gives parts of the output from `summary(lm(y ~ x))`.

6.3 Checking the Model

Definition 6.1 (**General Gaussian Response Model**).

A **general Gaussian response model** of the form

$$Y_i \sim G(\mu(\mathbf{x}_i), \sigma), \quad \text{for } i = 1, \ldots, n$$

independently, where $\mathbf{x}_i$ is a linear combination of any number of explanatory covariates.

The Gaussian response model can also be written as $Y_i = \mu(\mathbf{x}_i) + R_i$, where $R_i \sim G(0, \sigma)$, $i = 1, \ldots, n$ independently. In this case, Y_i is the sum of two components"

- $\mu(\mathbf{x}_i)$: a deterministic component (i.e. not a random variable).
- R_i: a random component (i.e. a random variable).

In many examples the deterministic component takes the form

$$\mathbb{E}[Y_i] = \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

so $\mathbb{E}[Y_i]$ is a linear function of a vector of k explanatory variates for unit i and the unknown parameters $\beta_0, \ldots, \beta_k$. Note that we are estimating $k+1$ parameters in the model. If we want to fit ths model, when we find $\beta_0, \ldots, \beta_k$ that minimize $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2$.

Example (Multiple Linear Regression & Interpreting $\hat{\beta}_j$).

If we have a model that $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, then in R:

```
> mod <- lm(y ~ x1 + x2)
> summary(mod)
```

More generally, $\hat{\beta}_j$ can be interpreted as the amount of increase in response y when x_j increases by one unit when the other predictors $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k$ are held constant.

In our $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ case, we say: "For every extra unit in x_1, the average y inproved by $\hat{\beta}_1$ units, assuming x_2 is held constant."

Example (Hypothesis Tests). In the simple linear regression model $y = \alpha + \beta x$, we considered testing $H_0 : \beta = 0$ using the test statistic $\frac{\hat{\beta}}{s_e/\sqrt{S_{xx}}}$, which we can think of the test statistic as being of the form $\frac{\text{estimate}}{\text{standard error}}$. If H_0 is true, then it follows a t-distribution with $n-2$ degrees of freedom.

In the more genral model $y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$, we can test $H_0 : \beta_j = 0$ using the test statistic

$$t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} = \frac{\text{estimate}}{\text{standard error}}.$$

If H_0 is true, then the corresponding estimator $T_j \sim t_{n-k-1}$ as we are estimating $k+1$ parameters.

For the STAT 230/231 & ENGL 119 example in class:

We can see these tests in our R output:

```
> mod1 <- lm(y ~ x1 + x2)
> summary(mod1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01375    5.01527  -0.202  0.84133
x1           0.73142    0.07664   9.544 3.83e-10 ***
x2           0.28225    0.09850   2.866  0.00797 **
```

Note: You do *not* need to be able to calculate these test statistics yourself, out you should be able to use R output such as the above to conduct hypothesis tests in these extended models.

Checking the Model

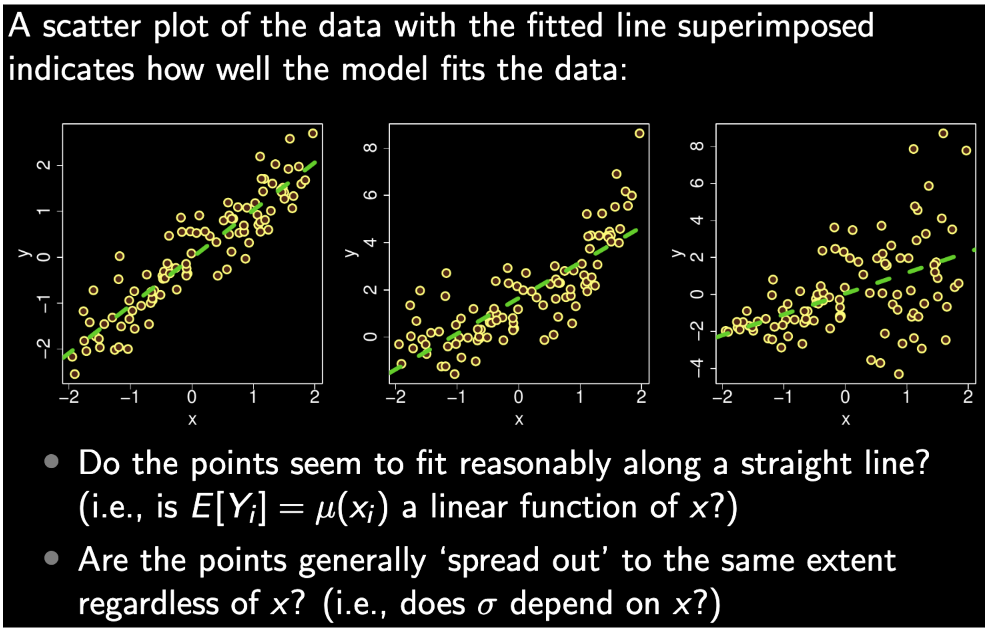
We can now consider how to check our model assumptions for Gaussian response models. There are two main assumptions (with (1) split into two parts):

(1a) **Gaussian**: Y_i (given covariates $\mathbf{x}_i$) has a Gaussian distribution.

(1b) **Constant variance**: That distribution has standard deviation σ which does not depend on the covariates.

(2) **Linearity**: $E[Y_i] = \mu(\mathbf{x}_i)$ is a linear combination of known covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and the unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

Model checking method 1



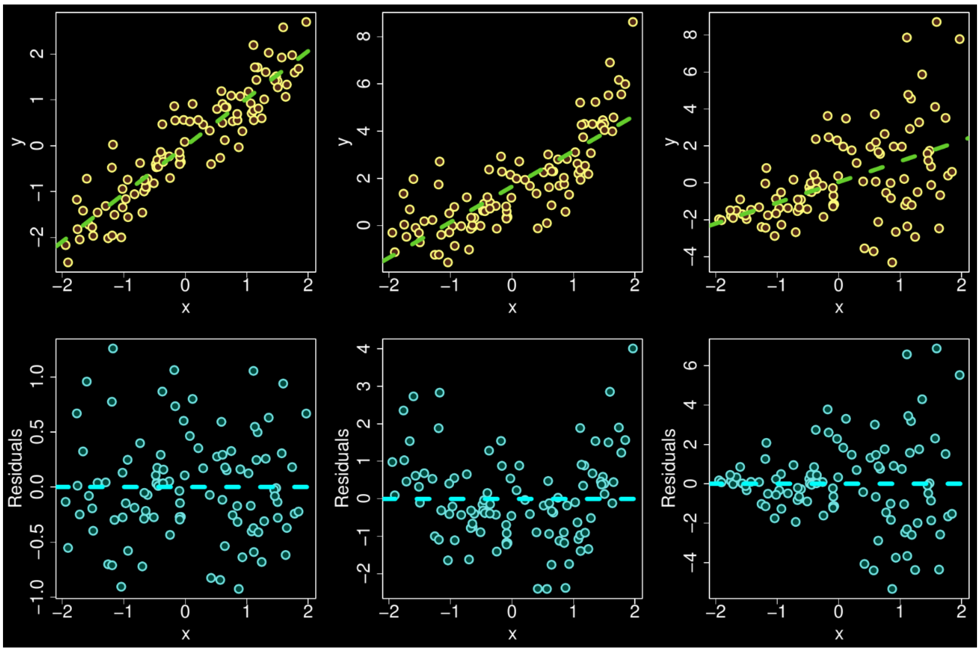
Model checking method 2: residuals

Scatterplots can sometimes be hard to read, so another mehod is **residual plot**. For simple linear regression model, let $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$ (this is called the **fitted response**), and then let $\hat{r}_i = y_i - \hat{\mu}_i$. The $\hat{r}_i$ are called the **residuals** since it represents what is 'left over' after the model has been fitted to the data.

Idea: $\hat{r}_i$ can be thought as observed values of R_i in the model $Y_i = \mu_i + R_i$ where $R_i \sim \mathrm{G}(0, \sigma)$, $i = 1, \ldots, n$ independently. If the model is correct, the $\hat{r}_i$ should act like a random sample from $\mathrm{G}(0, \sigma)$.

Exercise: Show that our least squares estimates of regression parameters imply $\sum_{i=1}^{n} \hat{r}_i = 0$. In other words, the average of our residuals is always zero.

If the model assumptions hold, then a plot of the points $(x_i, \hat{r}_i)$ should look randomly scattered about a horizontal line around $\hat{r}_i = 0$. There are three examples of such plots on the following figure.



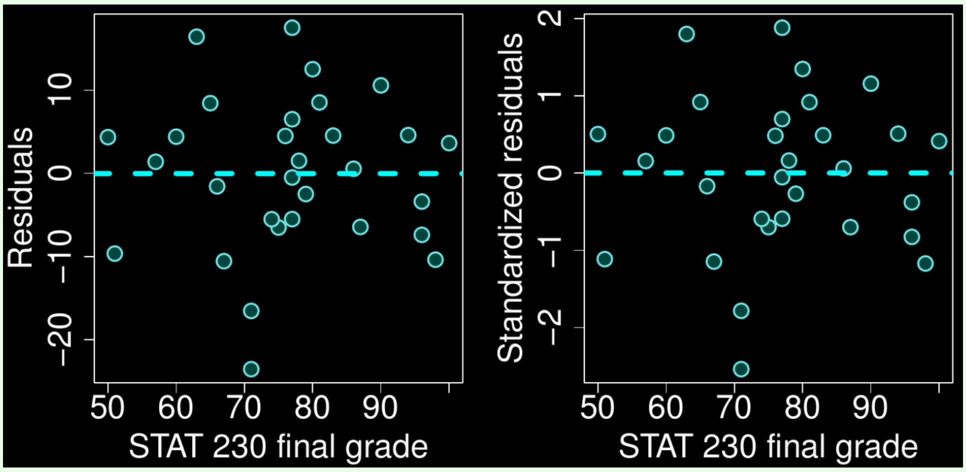
Remark. The bottom plots are the corresponding residual plots. For the pattern in the top right plot, we call it **heteroscedasticity**. For the top left plot, we call it **homoscedasticity**.

The variance in our residuals depends on σ, and so different datasets will result in more/less variable residuals! An approach: **standardized residuals**.

Definition 6.2 (Standardized Residuals).
We define $\hat{r}_i^* = \dfrac{\hat{r}_i}{s_e} = \dfrac{y_i - \hat{\mu}_i}{s_e}$, $i = 1, \ldots, n$. These are called the **standardized residuals**.

If we plot $(x_i, \hat{r}_i^*)$ instead of $(x_i, \hat{r}_i)$, the plot will look the same, but be scaled.
In fact, the $\hat{r}_i^*$ values should (almost entirely) lie in $(-3, 3)$ as they are approximately $\mathrm{G}(0, 1)$.

Example (**STAT 230/231 Grades Example**).



Non-standardized and standardized residuals plot look almost identical except for the scale of the y-axis. Also, we expect to see most of points in $(-3, 3)$.

Remark. If you want to apply the residual plot method to more complex models, you need to do a plot for each of the explanatory variates. Instead, we can plot the fitted values

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

against the residuals! Just as in the case of simple linear regression: we look to see if the points appear to be randomly scattered around a horizontal line at 0.

Key differences from the scatterplot method:

- More general: they can be used when we have more than one covariate.
- Make visualization easier: assessing whether the points lie along a horizontal line rather than an angled line.

When we check the assumptions:

- Are the points generally 'spead out' to the same extent regardless of x? (i.e. does σ depend on x?)
- Do the points seem to fit reasonably along a stright line? (i.e. is $\mathbb{E}[Y_i] = \mu(\mathbf{x}_i)$ a linear function of x?)

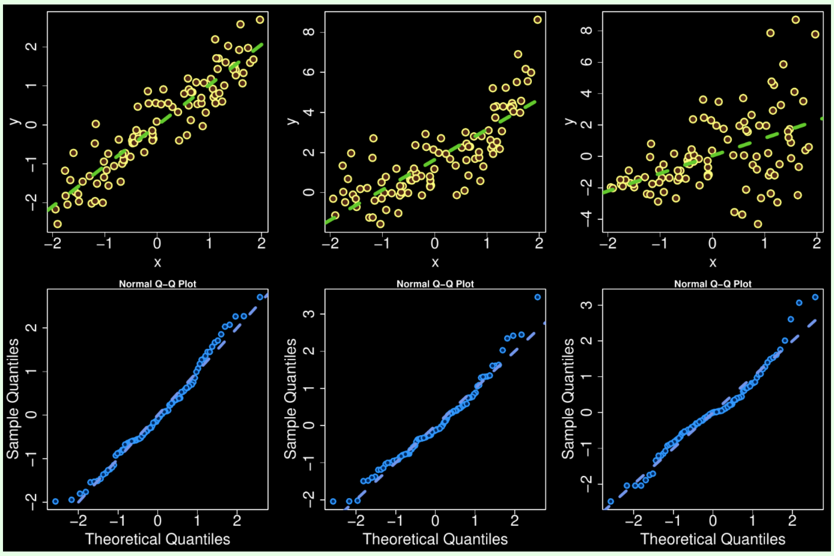
Q-Q Plot of Residuals

To check the Gaussian assumption, we use a Q-Q plot of the standardized residuals $\hat{r}_i^*$. Since our assumed model (for the standardized residuals) is

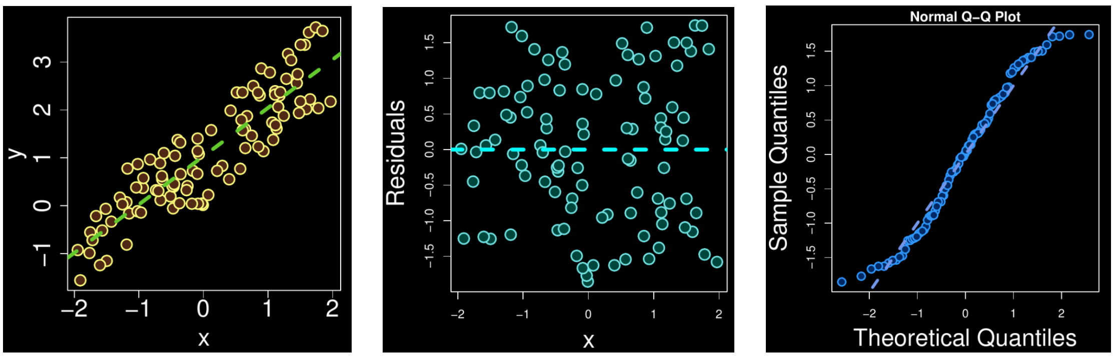
$$\frac{R_i}{\sigma} = \frac{Y_i - \mu_i}{\sigma} \sim \mathrm{G}(0,1),$$

then a Q-Q plot of $\hat{r}_i^*$ should give approximately a stright line if the model assumptions hold.

Example. Below are the corresponding Q-Q plots for our previous examples.



Important: we look carefully to find evidence a model assumption is being violated!

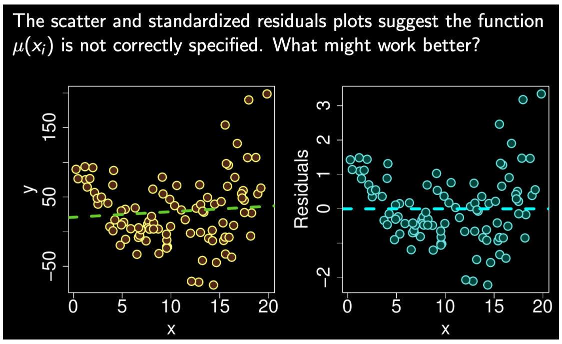


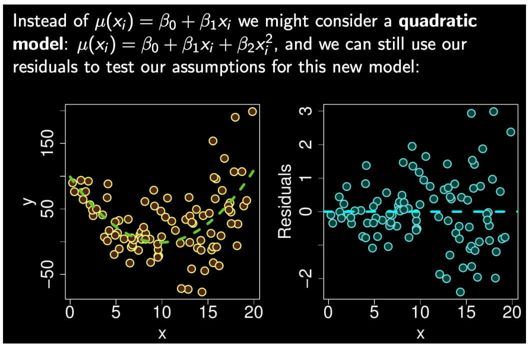
(a) Scatterplot (b) Residual (c) Q-Q of Residuals

Note. The first two plots look fine, but the Q-Q plot shows that the model assumption is violated!

Remark. We should try not to read too much into plots especially if the plots are based on a small number of points. Practice!

Other Types of Relationship





Predicting beyond the range of the covariates

Note that our assumptions are checked based on our observed data! We can only make predictions for what we have observed. Two key problems:

- Our model assumptions may no longer hold, and we have no way to check them.
- Our predictions may not make sense.

Lecture 19

6.4 Comparison of Two Population Means

In Chapter 4, we estimated the average length of 'good' and 'bad' movies based on their ratings.

Question: Is the average movie length for good movies the same as that for bad movies? (Or it's just we took a sample with longer good movies and shorter bad movies?)

To answer our question, we need a model:

Let Y_{1i} be the length of the i^{th} bad movie ($i = 1, \dots, 42$), and Y_{2i} be the length of the i^{th} good movie ($i = 1, \dots, 58$). It seems reasonable to assume Gaussian models for Y_{1i} and Y_{2i}:

$$\begin{aligned} Y_{1i} &\sim \mathrm{G}(\mu_1, \sigma) \quad \text{for } i = 1, \dots, 42 \\ Y_{2i} &\sim \mathrm{G}(\mu_2, \sigma) \quad \text{for } i = 1, \dots, 58 \end{aligned}$$

We also assumed both populations have the standard deviation σ. Interpretation:

- μ_1: average length in minutes of all bad movies in the study population.
- μ_2: average length in minutes of all good movies in the study population.
- σ: standard deviations of movie lengths in the study population (all movies).

Generally, for sample sizes n_1 and n_2,

$$\begin{aligned} Y_{1i} &\sim \mathrm{G}(\mu_1, \sigma) \quad \text{for } i = 1, \dots, n_1 \\ Y_{2i} &\sim \mathrm{G}(\mu_2, \sigma) \quad \text{for } i = 1, \dots, n_2 \end{aligned}$$

We call this a **two sample Gaussian problem**. This is a special case of the Gaussian response model.

Question: What is our H_0 for the movies example?

Answer: Good movies and bad movies are the same length on average. We have

$$H_0 : \mu_1 = \mu_2 \iff H_0 : \mu_1 - \mu_2 = 0.$$

Example (Reminder: Steps to Test $H_0 : \mu = \mu_0$ in $\mathrm{G}(\mu, \sigma)$).

1. Found a pivotal quantity for μ: $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.
2. Formed $100p\%$ CI: $\bar{y} \pm a\frac{s}{\sqrt{n}}$, $P(T \leq a) = \frac{1+p}{2}$, $T \sim t_{n-1}$.
3. Identified a test statistic: $D = \frac{\bar{Y}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$.
4. Calculated p-value: $2\left[1 - P(T \leq d)\right]$, $T \sim t_{n-1}$.

Confidence Intervals for $\mu_1 - \mu_2$

A point estimator of $\mu_1 - \mu_2$ is $\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2$. Since

$$\tilde{\mu}_1 = \bar{Y}_1 \sim \mathrm{G}\left(\mu_1, \frac{\sigma}{\sqrt{n_1}}\right) \quad \text{and} \quad \tilde{\mu}_2 = \bar{Y}_2 \sim \mathrm{G}\left(\mu_2, \frac{\sigma}{\sqrt{n_2}}\right)$$

independently, then

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2 \sim \mathrm{G}\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right).$$

Remark. Recall, if Y_1 and Y_2 are independent, then $\mathrm{Var}\left(Y_1 + Y_2\right) = \mathrm{Var}\left(Y_1\right) + \mathrm{Var}\left(Y_2\right) = \mathrm{Var}\left(Y_1 - Y_2\right)$.

Equivalently, $\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{G}(0, 1)$. To estimate σ, we first define

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2$$

which is the point estimator of σ^2 based on only the Y_1 and Y_2. A point estimator of σ^2 is then

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

which is called the **pooled estimator of variance**.

Remark. S_p^2 is NOT the MLE of σ^2, we use it because it is unbiased, i.e. $\mathbb{E}\left[S_p^2\right] = \sigma^2$. Also, s_p^2 is a weighted average of the sample variances s_j^2 with weights equal to $w_j = n_j - 1$.

We have

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

Since $\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{G}(0, 1)$ and $\dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$ independently, by Theorem 4.9, we have

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

Theorem 6.4. If $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from $\mathrm{G}(\mu_1, \sigma)$ and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from $\mathrm{G}(\mu_2, \sigma)$, 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 $\dfrac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$.

We can use these pivotal quantities to form a $100p\%$ CI and hypothesis tests for $\mu_1 - \mu_2$.

A $100p\%$ CI for $\mu_1 - \mu_2$ is

$$\left[\bar{y}_1 - \bar{y}_2 - as_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{y}_1 - \bar{y}_2 + as_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

where $P(T \leq a) = \frac{1+p}{2}$ and $T \sim t_{n_1+n_2-2}$.

Note.

- As n_1 and n_2 increases, CI narrower.
- As s_p increases, CI wider.

We test $H_0 : \mu_1 - \mu_2 = 0$, and use the test statistic $D = \frac{|\bar{Y}_1 - \bar{Y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with the observed value $d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, which is an observation from a $t_{n_1+n_2-2}$ distribution. So, large values of d would be surprising if H_0 is true. The p-value is then $2\left[1 - P(T \leq d)\right]$, $T \sim t_{n_1+n_2-2}$.

Example (Movie Length Example).

- Bad movies: $n_1 = 42, \hat{\mu}_1 = 86.190, s_1^2 = 264.1092, s_1 = 16.251$.
- Good movies: $n_2 = 58, \hat{\mu}_2 = 97.707, s_2^2 = 332.281, s_2 = 18.229$.

Therefore, $\hat{\mu}_1 - \hat{\mu}_2 = -11.517$, and $s_p = 17.42872$ (note: this is between s_1 and s_2). Also, s_p is closer to s_2 because $n_2 > n_1$ and s_p is a weighted average of s_1 and s_2.

For a 95% CI:

```
> qt(0.975, 42 + 58 - 2)
[1] 1.984467
```

So, the CI is $[-18.525, -4.509]$.

Remark. Here, $\hat{\mu}_1 - \hat{\mu}_2 < 0$ means that bad movies are shorter.

Since $0 \notin [-18.525, -4.509]$, we know that p-value for testing H_0 is < 0.05. We can use $d = 3.261463$ to find the p-value:

```
> pt(3.261463, 42 + 58 - 2)
[1] 0.9992372
> 2*(1-pt(3.261463, 98))
[1] 0.00152551
```

Since p-value < 0.01, there is a strong evidence against H_0 based on the data.

```
> t.test(bad, good, var.equal = TRUE)

 Two Sample t-test

data:  bad and good
t = -3.2613, df = 98, p-value = 0.001526
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.524043  -4.508797
sample estimates:
mean of x mean of y
 86.19048  97.70690
```

Warning: R calculates $d = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = -3.261463$, and not its absolute value!

Note. We tell R that we made the assumption that the variances are equal by setting `var.equal = TRUE`.

Two Gaussian Populations with Unequal Variances

Assume that

$$Y_{1i} \sim \mathrm{G}(\mu_1, \sigma_1) \quad \text{for } i = 1, \dots, n_1$$

$$Y_{2i} \sim \mathrm{G}(\mu_2, \sigma_2) \quad \text{for } i = 1, \dots, n_2$$

independently with $\sigma_1 \neq \sigma_2$. If n_1 and n_2 are large, then we can use the approximate pivotal quantity

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathrm{G}(0,1) \quad \text{approximately}$$

to construct CIs and hypothesis tests for $\mu_1 - \mu_2$.

Example (Above Continued). Assuming equal variances, our 95% CI for $\mu_1 - \mu_2$ was $[-18.525, -4.509]$. If we assume unequal variances, we would get $[-18.312, -4.722]$. Our intervals are almost the same, so our conclusions would be basically the same for both assumptions.

Comparison of Means Using Paired Data

Example. An experiment was designed to compare two methods: long multiplication and box multiplication. A class of 20 students were first taught to use long multiplication and then wrote a 30-question test where they were told to use long multiplication only. The following week, they were taught to use box multiplication and then wrote a similar test using box multiplication only. The results are shown below:

Test 1	18, 21, 16, 22, 19, 24, 17, 21, 23, 18, 14, 16, 16, 19, 18, 20, 12, 22, 15, 17	Mean: 18.40
Test 2	22, 25, 17, 24, 16, 29, 20, 23, 19, 20, 15, 15, 18, 26, 18, 24, 18, 25, 19, 16	Mean: 20.45

Note. We could test $H_0 : \mu_1 = \mu_2$ to find p-value $= 0.0824$, suggesting weak evidence of a difference between the scores on the two tests. Also, there are limitations on the experiment, e.g. students do better in the second test because they improved.

Our observations for scores on two tests are not independent. It makes sense to pair up our observations, so that the score of student i on test 1 is paired with the students's score on test 2. For example, student 1 got 4 marks better...

This makes intuitive sense: we are eliminating some factors (such as the student's math ability, study habits, and so on) which might otherwise affect conclusion about the parameter of interest, which is the mean difference $\mu_1 - \mu_2$.

This is an example of **paired data** and we might describe this as a **paired experiment**.

Let's now look at the mathematical "intuition". For a paired experiment, if $\mathrm{Var}(Y_{1i}) = \sigma_1^2$ and $\mathrm{Var}(Y_{2i}) = \sigma_2^2$, then

$$\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\,\mathrm{Cov}\left(\bar{Y}_1, \bar{Y}_2\right).$$

If $\mathrm{Cov}\left(\bar{Y}_1, \bar{Y}_2\right) > 0$, then $\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right)$ is smaller than for an unpaired experiment, where $\mathrm{Var}\left(\bar{Y}_1 - \bar{Y}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ since $\bar{Y}_1$ and $\bar{Y}_2$ are independent. To make inferences about $\mu_1 - \mu_2$, we analyze the within-pair differences

$$Y_i = Y_{1i} - Y_{2i} \quad \text{for } i = 1, \dots, n$$

by assuming $Y_i \sim \mathrm{G}(\mu_1 - \mu_2, \sigma)$ independently.

Note. Think about what this means in terms of how the fact that data are paired gives us information about the experiment!

Example (Continued).

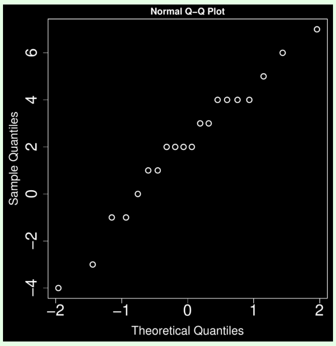
For the grades data, we have

$$y_i = \text{Score on test 1} - \text{Score on test 2}$$

which looks like this:

Difference	-4, -4, -1, -2, 3, -5, -3, -2, 4, -2, -1, 1, -2, -7, 0, -4, -6, -3, -4, 1	Mean: -2.05

We will check the Gaussian assumption and the independence assumption. The dataset is small to use a histogram to assess normality, but we can use a Q-Q plot.



For $Y \sim \mathrm{G}(\mu, \sigma)$ with μ and σ unknown, we test $H_0 : \mu = \mu_0$ using the test statistic

$$D = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

For our data, we have $\bar{y} = -2.05, s = 2.837252, n = 20$. To test $H_0 : \mu = \mu_1 - \mu_2 = 0$, we have $d = 3.231$. Using R to get p-value $= 0.00439746 < 0.01$, there is a strong evidence against H_0. Recall that we had p-value $= 0.0824$ when we analyzed the test scores assuming independence between the groups.

A handy summary 👀 for tests of $H_0 : \mu_1 = \mu_2$ where $Y_{1i} \sim G(\mu_1, \sigma)$ and $Y_{2i} \sim G(\mu_2, \sigma)$.

- Unpaired:

$$d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad p = 2[1 - P(T \le d)],\ T \sim t_{n_1+n_2-2}$$

- Paired: define $Y_i = Y_{1i} - Y_{2i} \sim G(\mu, \sigma)$ and test $H_0 : \mu = 0$

$$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}} \qquad p = 2[1 - P(T \le d)],\ T \sim t_{n-1}.$$

Table 6.4.1

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 6.4.2
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	$H_0 : \mu_1 = \mu_2$	$\frac{\left\vert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\right\vert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$2P\left(Z \geq \frac{\vert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\vert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$ $Z \sim G(0,1)$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ σ unknown	$H_0 : \mu_1 = \mu_2$	$\frac{\left\vert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\right\vert}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$2P\left(T \geq \frac{\vert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\vert}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)$ $T \sim t(n_1 + n_2 - 2)$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ 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	$H_0 : \mu_1 = \mu_2$	$\frac{\left\vert\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)\right\vert}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	approximate $p - value$ $2P\left(Z \geq \frac{\vert\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)\vert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right)$ $Z \sim G(0,1)$

7 Multinomial Models and Goodness of Fit Tests

Lecture 20

7.1 Likelihood Ratio Test for the Multinomial Model

Suppose $Y_1, \ldots Y_k$ has a Multinomial distribution with joint probability function

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k; \theta) = f(y_1, y_2, \ldots, y_k; \theta) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

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

Note. The experiment is repeated n times with k distinct outcomes.

The likelihood function is

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

Thus, we can write $L(\theta) = \prod_{j=1}^{k} \theta_j^{y_j}$. We can show (Lagrange multipliers) that the MLE of θ_j is

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

while the maximum likelihood estimator of θ is $\tilde{\theta}_j = \frac{Y_j}{n}$, $j = 1, 2, \ldots, k$.

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

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

Note. Both $L(\theta_0)$ and $L(\tilde{\theta})$ are random variables here. Also,

- $\theta_0 = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$.
- $\tilde{\theta} = \left(\frac{Y_1}{n}, \frac{Y_2}{n}, \ldots, \frac{Y_k}{n}\right)$.
- $L(\theta_1, \ldots, \theta_k) = \prod_{j=1}^{k} \theta_j^{y_j}$.

Note that, we have

$$L(\theta_0) = \prod_{j=1}^{k} \left(\frac{1}{k}\right)^{Y_j} \quad \text{and} \quad L(\tilde{\theta}) = \prod_{j=1}^{k} \left(\frac{Y_j}{n}\right)^{Y_j}$$

Then, $\dfrac{L(\theta_0)}{L(\tilde{\theta})} = \prod_{j=1}^{k} \left(\dfrac{n/k}{Y_k}\right)^{Y_j}$, where Y_j is the random variable corresponding to the observed frequency in category j. If $H_0 : \theta_j = \frac{1}{k}$ is true, then the expected frequency in category j is $\frac{n}{k}$, so we can write

$$\frac{L(\theta_0)}{L(\tilde{\theta})} = \prod_{j=1}^{k} \left(\frac{E_j}{Y_j}\right)^{Y_j}$$

where $E_j = \frac{n}{k}$, note that E_js might not be the same in other examples. The likelihood ratio statistic is:

$$\Lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right) = -2\log\left[\prod_{j=1}^{k} \left(\frac{E_j}{Y_j}\right)^{Y_j}\right] = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$

with observed value $\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\dfrac{y_j}{e_j}\right)$.

Note. What makes $\lambda(\theta_0)$ large/small?

- If $y_j = e_j$, then catogory j will not affect the statistic.
- If $y_j \geq e_j$, then category j will increase the statistic.
- If $y_j \leq e_j$, then category j will decrease the statistic.

Remember: our categories are NOT independent. If $y_j > e_j$ in one category, we must have $y_j < e_j$ in another category. Also, $\lambda(\theta_0)$ gets larger as the data deviates more from H_0.

Note. y_j is the observed count in category j, and e_j is the expected count in category j under H_0.

Proposition 7.1. If n is large and if H_0 is true, then

$$\Lambda(\theta_0) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right) \sim \chi^2_{k-1-p} \quad \text{approximately}$$

where k is the number of categories and p is the number of parameters estimated under H_0.

Note. A guideline for a large enough sample is $e_j \geq 5$ for all j. 4.9, 5.1 are basically equivalent to 5.

Remark (**Degrees of Freedom**). Values that are "free to move". For example, in a sample with n

observations and k categories, we have $k-1$ degrees of freedom. The last category is determined by the other $k-1$ categories.

The approximate p-value of a test of H_0 is;

$$p\text{-value} = P(W \geq \lambda(\theta_0)), \quad \text{where } W \sim \chi^2_{k-1-p}$$

Alternative Test Statistic: Person goodness of fit statistic

$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j}$, and $D \sim \chi^2_{k-1-p}$ approximately for large n. Observed value $d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$.

7.2 Goodness of Fit Tests

Example. Referring to Figure 57, the hockey example in Chapter 2, we assumed the Poisson model, but we can now assess it using multinomial!

Goals	0	1	2	3	4	5	6	≥ 7
Games	2	17	21	18	15	7	1	1

Recall: Y = number of goals in a randomly chosen game, assume $Y \sim$ Poisson(θ), and θ is the average number of goals per game.

Each event (a hockey game) falls into one of the eight categories. Let θ_i represent the probability a game has i goals (note: θ_7 includes 7 or more goals). Let Y_j = number of games from a sample of n we would expect to fall into category j:

$$Y_j \sim \text{Multinomial}(n; \theta_0, \ldots, \theta_7)$$

Note that $P(Y = j) = \frac{e^{-\theta}\theta^j}{j!}$, where $j = 0, 1, 2, \ldots$. Thus, our null hypothesis is

$$H_0 := \begin{cases} \theta_j = \frac{e^{-\theta}\theta^j}{j!}, & j = 0, 1, \ldots, 6, \\ \theta_j = \sum_{y=7}^{\infty} \frac{\theta^y e^{-\theta}}{y!}, & j = 7 \end{cases}$$

The expeted counts:

Goals	0	1	2	3	4	5	6	≥ 7
Observed	2	17	21	18	15	7	1	1
Expected	5.54	14.93	20.11	18.07	12.17	6.56	2.95	1.67

What are the degress of freedom? Three ways to look at it.

- Note that $k = 8$. In forming our expected counts in Chapter 2, we estimated θ from the data, so $p = 1$. Thus, the degrees of freedom is $k - 1 - p = 8 - 1 - 1 = 6$.
- From the data, we know that sample size is 82 and $\hat{\theta} = 2.695$, then there are $82 \times 2.695 = 221$ expected goals. If the first six observed counts are known (they add up to 208 goals), then only possible combination for the last two categories is 1 and 1. Thus, the degrees of freedom is 6.
- We have two constraints: $\sum_{j=0}^{6} y_j = 82$ and $\hat{\theta} = 2.695$. Thus, the degrees of freedom is 6.

Note that we need $e_j \geq 5$ for all j, we can collapse the last two categories into the $j = 5$ category. Then

Goals	0	1	2	3	4	≥ 5
Observed	2	17	21	18	15	9
Expected	5.538	14.925	20.112	18.068	12.174	10.645

Warning: the degrees of freedom is now 4!

Using R, p-value $\approx P(W \geq 5.270) = 0.2606985 > 0.1$ and there is no evidence against H_0. Poisson model is reasonable here!

Exercise: Show that the p-value from using the Pearson test statistic is approximately 0.478.

Lecture 21

7.3 Two-Way (Contigency) Tables

Example. Below is a 2×2 contigency table.

	Canadian hometown	**Non-Canadian hometown**	**Total**
Hockey :)	$y_{11} = 33$	$y_{12} = 9$	42
Hockey :(	$y_{21} = 22$	$y_{22} = 43$	65
Total	55	52	107

The relative risk of liking hockey among those with a Canadian hometow $= \frac{(33/55)}{(9/52)} = 3.467$.
Meaning: students with a Canadian hometown are over 3 times as likely to like hockey!
We may wish to test whether hometown and hockey love are **independent**. If they are independent, we would expect to see the following.

	Canadian hometown	**Non-Canadian hometown**	**Total**
Hockey :)	$y_{11} = 22$	$y_{12} = 20$	42
Hockey :(	$y_{21} = 33$	$y_{22} = 32$	65
Total	55	52	107

Note. Intuition: $\frac{42}{107}$ (39%) of the sample liked hockey (if hockey and hometown were unrelated). The relative risk for these data would be 1.
More generally, suppose n individuals are classified according to two different variates which have two possible values. Then,

	B	$\overline{B}$	Total
A	y_{11}	y_{12}	$r_1 = y_{11} + y_{12}$
$\overline{A}$	y_{21}	y_{22}	$r_2 = y_{21} + y_{22}$
Total	$c_1 = y_{11} + y_{21}$	$c_2 = y_{12} + y_{22}$	n

We can therefore define random variables:

$$
\begin{aligned}
Y_{11} &= \# A \cap B \text{ outcomes} & Y_{12} &= \# A \cap \overline{B} \text{ outcomes} \\
Y_{21} &= \# \overline{A} \cap B \text{ outcomes} & Y_{22} &= \# \overline{A} \cap \overline{B} \text{ outcomes}
\end{aligned}
$$

A suitable model: $(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}(n; \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ where

$$
\theta_{11} = P(A \cap B) \quad \theta_{12} = P(A \cap \overline{B}) \quad \theta_{21} = P(\overline{A} \cap B) \quad \theta_{22} = P(\overline{A} \cap \overline{B})
$$

The null hypothesis is that A and B are independent:

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

Let $P(A) = \alpha$ and $P(B) = \beta$, then $H_0 : \theta_{11} = \alpha\beta$. This is equivalent to the following:

- $\theta_{12} = P(A \cap \overline{B}) = \alpha(1 - \beta)$
- $\theta_{21} = P(\overline{A} \cap B) = (1 - \alpha)\beta$
- $\theta_{22} = P(\overline{A} \cap \overline{B}) = (1 - \alpha)(1 - \beta)$

The likelihood function (ignoring constants) is

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

The MLEs are $\hat{\theta}_{ij} = \frac{y_{ij}}{n}, i, j = 1, 2$ with corresponding estimators $\tilde{\theta}_{ij} = \frac{Y_{ij}}{n}, i, j = 1, 2$. Then

$$
L(\tilde{\theta}) = \tilde{\theta}_{11}^{Y_{11}} \tilde{\theta}_{12}^{Y_{12}} \tilde{\theta}_{21}^{Y_{21}} \tilde{\theta}_{22}^{Y_{22}} = \left(\frac{Y_{11}}{n}\right)^{Y_{11}} \left(\frac{Y_{12}}{n}\right)^{Y_{12}} \left(\frac{Y_{21}}{n}\right)^{Y_{21}} \left(\frac{Y_{22}}{n}\right)^{Y_{22}}.
$$

We only need $L(\theta_0)$ now for $\Lambda(\theta_0)$. If $H_0 : \theta_{11} = \alpha\beta$ is true, then

$$
\begin{aligned}
L(\theta) &= \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}} \\
L(\alpha, \beta) &= (\alpha\beta)^{y_{11}} (\alpha(1 - \beta))^{y_{12}} ((1 - \alpha)\beta)^{y_{21}} ((1 - \alpha)(1 - \beta))^{y_{22}} \\
&= \alpha^{y_{11} + y_{12}} (1 - \alpha)^{y_{21} + y_{22}} \beta^{y_{11} + y_{21}} (1 - \beta)^{y_{12} + y_{22}}
\end{aligned}
$$

with $0 < \alpha < 1$ and $0 < \beta < 1$.

Then, we require the following random variable to form our $\Lambda(\theta_0)$:

$$L(\tilde{\alpha}, \tilde{\beta}) = \tilde{\alpha}^{Y_{11}+Y_{12}}(1-\tilde{\alpha})^{Y_{21}+Y_{22}}\tilde{\beta}^{Y_{11}+Y_{21}}(1-\tilde{\beta})^{Y_{12}+Y_{22}}.$$

Note that we have

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

with corresponding maximum likelihood estimators

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

Now, we can form the likelihood ratio statistic for testing H_0:

$$\begin{aligned}\Lambda &= -2\log\left(\frac{L(\tilde{\alpha}, \tilde{\beta})}{L(\tilde{\theta}_{11}, \tilde{\theta}_{12}, \tilde{\theta}_{21}, \tilde{\theta}_{22})}\right) \\ &= -2\log\left(\frac{\tilde{\alpha}^{Y_{11}+Y_{12}}(1-\tilde{\alpha})^{Y_{21}+Y_{22}}\tilde{\beta}^{Y_{11}+Y_{21}}(1-\tilde{\beta})^{Y_{12}+Y_{22}}}{\tilde{\theta}_{11}^{Y_{11}}\tilde{\theta}_{12}^{Y_{12}}\tilde{\theta}_{21}^{Y_{21}}\tilde{\theta}_{22}^{Y_{22}}}\right)\end{aligned}$$

We can show that the likelihood ratio statistic for our independence test can also be rearranged to this form:

$$\Lambda = 2\left[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)\right],$$

where $E_{11} = n\tilde{\alpha}\tilde{\beta}$, $E_{12} = n\tilde{\alpha}(1-\tilde{\beta})$, ... In our sample,

$$e_{11} = n\hat{\alpha}\hat{\beta} = 55 \times \frac{33}{107} \times \frac{55}{107} = \frac{42 \times 55}{107} = 21.589.$$

Note. We don't need to calculate all of the expected values, we can use the row and column totals to find the rest.

In general, we will compare the observed frequencies:

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

With the frequencies if the null hypothesis were true:

	B	$\overline{B}$	Total
A	$e_{11} = \frac{r_1 c_1}{n}$	$e_{12} = r_1 - e_{11}$	$r_1 = y_{11} + y_{12}$
$\overline{A}$	$e_{21} = c_1 - e_{11}$	$e_{22} = r_2 - e_{21}$	$n - r_1$
Total	$c_1 = y_{11} + y_{21}$	$n - c_1$	n

For our example,

	Canadian hometown	Non-Canadian hometown	Total
Hockey :)	33	9	42
Hockey :(	22	43	65
Total	55	52	107

Expected counts are:

	Canadian hometown	Non-Canadian hometown	Total
Hockey :)	21.589	20.411	42
Hockey :(	33.411	31.589	65
Total	55	52	107

We can compute our likelihood ratio test statistic λ where to get our p-value we use the χ^2_{k-1-p} distribution (and $e_j \geq 5$). In our $H_0 : \theta_{11} = \alpha\beta$, we estimated 2 parameters. So the degrees of freedom is $4 - 1 - 2 = 1$. This makes sense because we only need need to know one entry in the table to determine the rest. In our example,

$$\begin{aligned}\lambda &= 2\left[y_{11}\log\left(\frac{y_{11}}{e_{11}}\right) + y_{12}\log\left(\frac{y_{12}}{e_{12}}\right) + y_{21}\log\left(\frac{y_{21}}{e_{21}}\right) + y_{22}\log\left(\frac{y_{22}}{e_{22}}\right)\right] \\ &= 21.4034.\end{aligned}$$

The p-value is approximately $P(W \geq 21.4034) = 2.721107 \times 10^{-6} < 0.01$ where $W \sim \chi^2_1$. There is a strong evidence against H_0.

Lecture 22

Let A and B denote the two categorical variates:

- Let A have a cotegories: $A_1, A_2, \dots, A_a$.
- Let B have b categories: $B_1, B_2, \dots, B_b$.

Example (Roll Up to Win).

For the Roll Up data, we have

- A: outcome, with A_1 = lose and A_2 = win.
- B: day, with B_1 = Monday, … B_4 = Thursday.

	Day				
	Mon	**Tue**	**Wed**	**Thu**	**Total**
Lose	$y_{11} = 8$	$y_{12} = 9$	$y_{13} = 10$	$y_{14} = 8$	$r_1 = 35$
Win	$y_{21} = 4$	$y_{22} = 4$	$y_{23} = 6$	$y_{24} = 7$	$r_2 = 21$
Total	$c_1 = 12$	$c_2 = 13$	$c_3 = 16$	$c_4 = 15$	$n = 56$

- Y_{ij} = number of units in category A_i and category B_j in a random sample of size n.
- θ_{ij} = probability of a randomly selected unit is in category A_i and B_j (for example. θ_{11} = the probability of a roll on Monday loses).
- $(Y_{11}, Y_{12}, \dots, Y_{ab}) \sim \text{Multinomial}(n; \theta_{11}, \theta_{12}, \dots, \theta_{ab})$.

Question: How do we form our H_0 in terms of the unknown parameters?

Answer: We have $H_0 : \theta_{ij} = P(A_i \cap B_j) = P(A_i)P(B_j)$ (since our null hypothesis is that A_i and B_j are independent). Let $\alpha_i = P(\text{unit is type } A_i)$ and $\beta_j = P(\text{unit is type } B_j)$. Then, to test whether A and B are independent variates, we test

$$H_0 : \theta_{ij} = \alpha_i \beta_j, \quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, b.$$

Let us estimate α_1, β_1 and e_{11}:

- $\hat{\alpha}_1 = \frac{r_1}{n} = \frac{35}{56}$.
- $\hat{\beta}_1 = \frac{c_1}{n} = \frac{12}{56}$.
- Expected number of Monday losers: $e_{11} = n \times P(A_1 \cap B_1) = n \cdot \frac{r_1}{n} \cdot \frac{c_1}{n} = \frac{35 \times 12}{56}$.

In general, we have $\hat{\alpha}_i = \frac{r_i}{n}$, $\hat{\beta}_j = \frac{c_j}{n}$, and $e_{ij} = \frac{r_i c_j}{n}$. Our expected counts are:

	Day				
	Mon	**Tue**	**Wed**	**Thu**	**Total**
Lose	7.500	8.125	10.000	9.375	35
Win	4.500	4.875	6.000	5.625	21
Total	12	13	16	15	56

The likelihood ratio statistic for an $a \times b$ contigency table:

$$\begin{aligned}\Lambda &= 2\left[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)\right] \\ &= 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij}\log\left(\frac{Y_{ij}}{E_{ij}}\right) \quad \text{(general version)}\end{aligned}$$

Note. In the previous hockey example, we can write $\Lambda = 2\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}\log\left(\frac{Y_{ij}}{E_{ij}}\right)$.

For large n and under H_0, $\Lambda \sim \chi^2_{k-1-p}$ approximately, where $k = ab$ and $p = (a-1)+(b-1)$.

Note. For p, we can use the row and column totals to determine the last entry.

Now, we have

$$k - 1 - p = (ab) - 1 - (a - 1 + b - 1) = ab - a - b + 1 = (a-1)(b-1).$$

degrees of freedom. Thus, $\Lambda \sim \chi^2_{(a-1)(b-1)}$ approximately. For our data, we have 3 degrees of freedom. This makes sense because we only need to know 3 entries in the table to determine the rest (3 entries are free to move).

Using R, we find $\lambda = 0.8727337$ and use χ^2_3 to find p-value $= 0.8320023 > 0.1$. Thus, there is no evidence based on the data against the null hypothesis that winning probability and day of week are independent.

Note. Not all of our expected counts were at least 5, so we would have concerns about the validity of our chi-squared approximation. However, they are very close to the threshold 5, we only need to increase the sample size slightly to satisfy this condition. Also, we can combine the categories to satisfy it. Thus, we have very little concern about the validity of our test. We might want to revisit that if we have a very small p-value or combining the categories will dramatically change the results.

Using R: Test of Equal Proportions

Example. We will compare the use of hashtags by two twitter accounts.

User	No Hashtags	Hashtags	Total
@AnishNation	121	69	190
@TheTorontoZoo	98	92	190
Total	219	161	380

Under a null hypothesis of equal proportions of tweets using hashtags, we have the following expected counts.

User	No Hashtags	Hashtags
@AnishNation	121	69
Expected	109.5	80.5
@TheTorontoZoo	98	92
Expected	109.5	80.5

Note. The expected count of values in a cell with row total r and column total c is $\frac{rc}{n}$.
We have 1 degree of freedom for this test. We can compute the likelihood ratio and Pearson test statistics as usual:

```
# Likelihood ratio test statistic:
> 2*sum(observed*log(observed/expected))
[1] 5.716968
# p-value:
> 1 - pchisq(5.716968, 1)
[1] 0.01680172
# Pearson test statistic:
> sum((observed - expected)^2/expected)
[1] 5.701239
# p-value:
> 1 - pchisq(5.701239, 1)
[1] 0.01695294
```

We can also use `chisq.test()`:

```
# Pearson test statistic:
> sum((observed - expected)^2/expected)
[1] 5.701239
# p-value:
> 1 - pchisq(5.701239, 1)
[1] 0.01695294
# Pearson test:
> chisq.test(observed, correct = FALSE)
    Pearson's Chi-squared test
data: observed
X-squared = 5.7012, df = 1, p-value = 0.01695
```

Note. We have specified "correct = FALSE" to indicate that we do not want to apply a Yates' continuity correction. This is not required for STAT 231.

8 Casual Relationships

8.1 Establishing Causations

Lecture 23

Definition 8.1 (**Casuation 1**). Let y be a response variate and let x be an explanatory variate associated with units in a population or process. Then, if all other factors that affect y are held constant, let us change x (or observe different values of x) and see if y changes. If it does, we say that x has a **casual effect** on y.

Remark. Causation is difficult to define. Above definition is not broad enough, for example, a change in x may only lead to a change in the distribution of y.

Definition 8.2 (**Causation 2**). x has a **casual effect** on Y if, when all other factors that affect Y are held constant, a change in x induces a change in a property of the distribution of Y.

Note. This is an improved definition. However, this definition is impractical since we cannot hold all other factors that affect y constant (we may not even know what all the factors are). The definition is idealized and should be used to conduct studies in order to show that a casual relationship exists.

Example (**COVID-19 Vaccination Example**).
Let Y = the number of people in a random sample of n people who have been hospitalized due to COVID-19. Assume $Y \sim \text{Binomial}(n, \theta)$.
We could say that vaccination has a casual effect on hospitalization due to COVID-19 if the value of θ was different for people who had been vaccinated and those who had not, assuming everything else that affects Y is held constant.

Reasons Two Variates can be Related

1. The explanatory variate is the direct cause of the response variate.
2. The response variate is causing a change in the explanatory variate.
3. The explanatory variate is a contributing but not sole cause of the response variate.
4. Both variates are changing with time.
5. The association may be due to coincidence.
6. Both variates may result from a common cause.

Example (Reason 1: Example).
If I'm thirsty, drink a cup of tea, and am then not thirsty, we would probably agree that the act of drinking tea caused my thirst to decrease.
Note. Here, a change in the explanatory variate is the direct cause of a change in the response variate.

Example (Reason 1: Example).
Even if one variate is the direct cause of another, we may not see a strong association.
For example, playing the lottery is the direct cause of winning the lottery. But the relationship between playing the lottery and winning the lottery is not srong; most decisions to play the lottery do not result in winning it.

Example (Reason 2: Reverse Causality).
Many video game players engage in what is known as 'toxic' behaviour. Suppose a study to investigate the effect of 'toxic' behaviour of players on their performance.

- Explanatory variate = player toxicity level.
- Response variate = porportion of games won.

We might expect toxic behaviour to cause players to lose more often (such as teammates quitting). However, sometimes the casual conncetion is the opposite of what we might expect. For example, perhaps players are more likely to engage in toxic behaviour as a result of losing matches. This is an example of reverse causality.

Example (Reason 3: Example).

The presence of a specific gene mutation is required in order for certain cancers to occur. However, presence of the mutation alone does not guarantee someone will develop that form of cancer.

Example (Reason 4: Example).

Nonsensical associations often result from correlating two variates that are both changing over time. For example, there is a strong (negative) correlation between global average temperature, and the number of pirates. There is no casual relationship between them.

Example (Reason 5: Example).

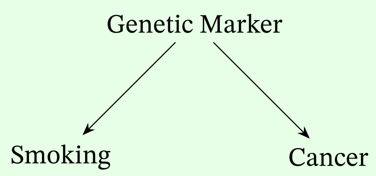
Thomas Jefferson, third president of the United States, died shortly after noon on July 4, 1826. A few hours later, John Adams, second president of the United States, also died. There is no casual relationship here, it's just a coincidence (randomness).

For reason 6, an association between two variates may be observed because both variates are responding to changes in some unobserved variate(s). These variates are called **confounding or lurking variates**.

Note. Lurking variates can lead to seemingly counterintuitive results.

Example (Reason 6: Example).

A much used explanation by the tobacco companies for the association between smoking and lung cancer is that smoking behaviour and lung cancer are both responses to a genetic predisposition.



Example (Simpson's Paradox Example).

A study into two treatments for kidney stones involved 700 patients, with following results:

- Treatment A: 273/350 successful treatments, or 78%.

- Treatment B: 289/350 successful treatments, or 83%.

However, treatments weren't assigned at random! If the kidney stones are categorized as 'small' or 'large', we see the following:

	Treatment A	**Treatment B**
Small	93% (81/87)	87% (234/270)
Large	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

Treatment A is better in both groups! Treatment A is assigned with harder jobs... This is an example of Simpson's Paradox.

8.2 Experimental Studies

Randomization: One way to establish causality from observational data. Referring to above example, if treatments had been randomized, we might expect a more even distribution of treatments across the two sizes of kidney stones. This helps eliminate the effect of confounding variates.
Note. Randomized studies can help identify casual relationships but they are not always possible, they might have unethical, impractical, or resources constraints.

8.3 Observational Studies

In observational studies, controlling variates and using randomization is not possible. Establishing causation here is then much more difficult and requires at least the following four features:

1. The association between the two variates must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects.
2. The association must continue to hold when the effects of plausible confounding variates are taken into account.
3. There must be a plausible scientific explanation for the direct influence of one variate on the other variate, so that a causal link does not depend on the observed association alone.
4. There must be a consistent response, that is, one variate always increases (decreases) as the other variate increases.

Example. We can see that each of the above features appears to be satisfied in the case of smoking and lung cancer:

1. A strong association between smoking and lung cancer has been observed in numerous studies in many countries.
2. Many possible sources of confounding variates have been examined in these studies and have not been found to explain the association. For example, data about nonsmokers who are exposed to secondhand smoke contradicts the genetic hypothesis.
3. Animal experiments have demonstrated conclusively that tobacco smoke contains substances that cause cancerous tumors. Therefore, there is a known pathway by which smoking causes lung cancer.
4. The lung cancer rates for ex-smokers decrease over time since smoking cessation.

The evidence for causation here is about as strong as non-experimental evidence can be!

A Tutorials

A.1 Tutorial 1

Most Important Topics from STAT 230

1. Random variables: basic concepts and properties.
2. Discrete and continuous probability distributions.
3. Multivariate distributions and linear combinations of random variables.
4. Central Limit Theorem.

Random Variables

- Upper case letters to denote random variables and lower case letters to denote observed values.
- Be very familiar with the Discrete Uniform, Binomial, Geometric, and Poisson distributions.
- When choosing a statistical model for real-world data, we will often choose a continuous distribution even though our real-world measurements are discrete.

Gaussian Distribution

- In STAT 231, we refer to Normal distribution as **Gaussian distribution**, and write $Y \sim \mathrm{G}(\mu, \sigma)$.
- $P(Y = a) = 0$ at each point.
- Standard normal: $Z \sim \mathrm{G}(0, 1)$. We will want to transform non-standard normal into a standard normal (standardization).
- $Y \sim \mathrm{G}(\mu, \sigma)$. If $X = a + Y$, then $X \sim \mathrm{G}(a + \mu, \sigma)$. If $X = bY$, then $X \sim \mathrm{G}(b\mu, |b|\sigma)$.
- General R syntax to find $P(Y < y)$ for $Y \sim \mathrm{G}(m, s)$ is

  ```
  > pnorm(y, m, s)
  ```
- We do not always need to standardize a Gaussian random variable to calculate probabilities - don't standardize unnecessarily.
- Make sure you can use R to compute the probabilities. Need to know R commands for exams.

Central Limit Theorem

- We will revisit "Repeated Sampling" in chapter 4.
- $\overline{Y}$ is also a random variable. It has a distribution, an expected value, and a variance. We can use the Central Limit Theorem to think about the distribution of $\overline{Y}$.
- CLT tells us if $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables with mean μ and standard deviation σ, then for large n,

$$\frac{\overline{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{G}(0, 1)$$

  approximately, where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- To apply CLT, we need
  - The number n of observations.
  - The mean μ of Y.
  - The standard deviation σ of Y.
- CLT will be very important in chapter 4. Must know how to answer the two bonus questions in the last slide of Tutorial 1 on LEARN.
- Continuity corrections.

A.2 Tutorial 2

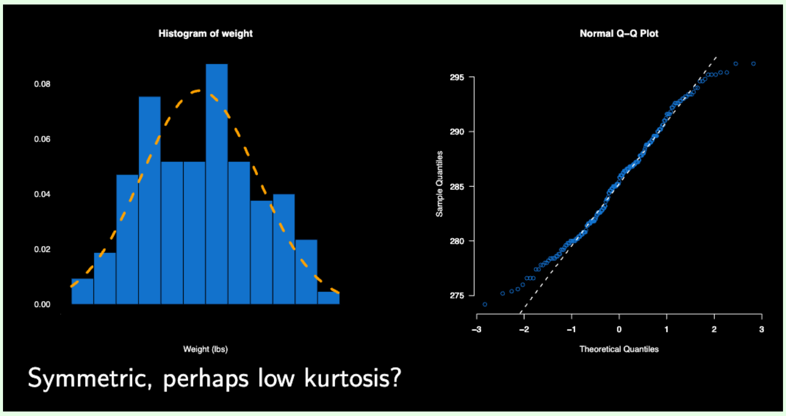
Talked about techniques for midterm 1 and worked through some practice problems.

A.3 Tutorial 3

Reviewed some important examples covering ideas in Chapter 2, 3 and 4. Key concepts:

- Q-Q plots of real-world data.
- Study, sample, and measurement errors: make sure you know exactly what they are!
- Use of change of variables to demonstrate a radom variable is a pivotal quantity, and form a confidence interval.

Example (**Q-Q Plot Example**).



- Symmetric, low kurtosis.
- y-axis of Q-Q plot matches x-axis of histogram.
- x-axis of Q-Q plot follows $G(0, 1)$.
- The sample median is the y value at $x = 0$ on the Q-Q plot (285.6 in this case).

Example (**PPDAC Example**).

Three important types of error:

1. **Study error**: if attributes in the study population differ from attributes in the target population.
2. **Sample error**: if attributes in the sample differ from attributes in the study population.
3. **Measurement error**: if measured value and true value of a variate are not the same.

A statistics professor was interested in learning about how current and future University of Waterloo undergraduates in the math faculty feel about flexible grading schemes. In particular, what proportion are in favour of flexible grading schemes.

The professor decided to survey students enrolled in STAT 231 in Fall 2018 to investigate this question. The students were asked questions in two surveys (one midway through the course, one just before the end of term), and only those who gave permission would have their responses used in the subsequent research and analysis.

Question: What are the target population, study population, sample, and measurement?

- Target population: current and future UW math faculty undergraduate students.
- Study population: students enrolled in STAT 231 in Fall 2018.
- Sample: Students who gave permission for their responses to be used.
- Measured variate values: the responses to the survey.

Question: What are possible sources of study, sample, and measurement error?

Hint: consider the attribute.

- Study error: more recently, students may be more in favour of flexible grading schemes.
- Sample error: those who don't give permission may be less favourable about flexible grading schemes.
- Measurement error: students lie or forget.

Example (**Chapter 4: Problem 27**). Important example!

A.4 Tutorial 4

Went through two examples and techniques for Midterm 2, focusing on Chapter 4 and 5.

A.5 Tutorial 5

If we write $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, we can define the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{r}_i^2.$$

The smaller this is, the less 'error' in our model fit.

This motivates the R^2 statistic to check overall model adequacy:

$$R^2 = 1 - \frac{SSE}{S_{yy}} = \frac{S_{yy} - SSE}{S_{yy}}.$$

We can interpret the R^2 as

$$R^2 = \frac{\text{Variability explained by the regression model}}{\text{Total variability in the data}}.$$

R^2 takes values between 0 (the regression explains none of the variation in our response) and 1 (the regression explains all of the variation in our response).

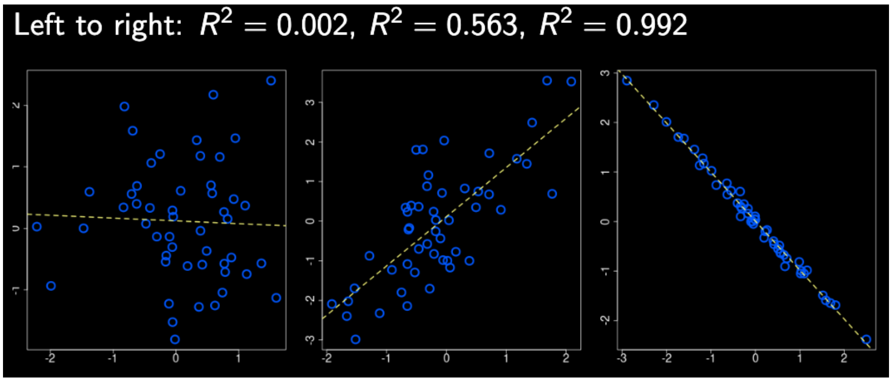


Figure 15: Examples of R^2 values.

In class, we showed that

$$s_e^2 = \frac{1}{n-2} SSE = \frac{1}{n-2}(S_{yy} - \hat{\beta} S_{xy}).$$

Therefore, we have

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} = \hat{\beta}\frac{S_{xy}}{S_{yy}}$$

where $SSE = S_{yy} - \hat{\beta} S_{xy}$ and $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$. Then,

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = r^2$$

where r^2 is the sample correlation squared.

The R^2 statistic can also be applied to more general linear models. If we write $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$, then we have $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Remark. Adding new explanatory will always appear to 'improve' our model by increasing R^2, even if the new variate is unrelated to our response. This leads to a problem of "overfitting". Thus, we consider the adjusted R^2:

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{S_{yy}/(n-1)}$$

where k = the number of explanatory variables in the model.

For the rest of the tutorial, we talked about an example from Chapter 7.

We introduced Benford's Law: numbers occurring with a first digit j occur with probability

$$P(j) = \log_{10}\left(1 + \frac{1}{j}\right).$$

B Real Appendix

ID	S230	S231	ID	S230	S231	ID	S230	S231
1	76	76	11	87	76	21	98	83
2	77	79	12	71	50	22	80	88
3	57	54	13	63	75	23	67	52
4	75	64	14	77	72	24	78	75
5	74	64	15	96	84	25	100	99
6	60	60	16	65	69	26	94	94
7	81	85	17	71	43	27	83	83
8	86	82	18	66	60	28	51	37
9	96	88	19	90	96	29	77	90
10	79	72	20	50	50	30	77	67

Figure 16: Final grades for 30 students who took both STAT 230 and STAT 231.