An Introduction to Statistics

Demetris Athienitis
Department of Statistics,
University of Florida

Contents

Part I

Modules 1-2

Module 1

Descriptive Statistics

1.1 Concept

Definition 1.1. *Population parameters* are a numerical summary concerning the complete collection of subjects, i.e. the population.

The population parameters are notated by Greek symbols such as population mean μ.

Definition 1.2. *Sample statistics* are a numerical summary concerning a subset of the population, i.e. the sample, from which we try to draw inference about the population parameter.

Sample statistics are notated by the "hat" symbol over the population parameter such as the sample mean $\hat{\mu}$, or sometimes for convenience a symbol from the English alphabet. For the sample mean $\hat{\mu} \equiv \bar{x}$.

1.2 Summary Statistics

Let $x_1, \ldots, x_n$ denote n observations/numbers.

1.2.1 Location

- The **mode** is the most frequently encountered observation.
- The **mean** is the arithmetic average of the observations. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- The **p$^{\text{th}}$ percentile** value divides the ordered data such that p% of the data are less than that value and (100-p)% greater than it. It is located at (p/100)(n+1) position of the ordered data. If the position value is not an integer then take a weighted average of the values at $\lfloor (p/100)(n+1) \rfloor$ and $\lceil (p/100)(n+1) \rceil$. **The median is actually the 50$^{\text{th}}$ percentile**.

- The $\alpha\%$ **trimmed mean** is the mean of the data with the smallest $\alpha\% \times n$ observations and the largest $\alpha\% \times n$ observations truncated from the data.

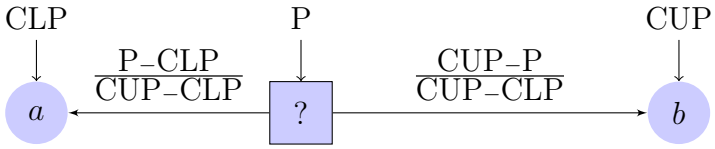
Example 1.1. The following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- There are three modes, 80, 179 and 232.
- Hence, $\sum_{i=1}^{24} x_i = 30 + 75 + \cdots + 384 + 470 = 4690$ and thus $\bar{x} = 4690/24 = 195.4167$.
- The median is located at the $0.5(24+1) = 12.5$ position. Hence, we average of the observations at the 12^{th} and 13^{th} position of the ordered data, i.e. $\tilde{x} = (191 + 223)/2 = 207$.
- The 25^{th} percentile (a.k.a. 1^{st} Quartile) is located at $(25/100)(24+1) = 6.25$ position. So take a weighted average of the values at 6^{th} and 7^{th} position, i.e.
$$0.75(105) + 0.25(126) = 110.25$$
- To compute the 5% trimmed mean we need to remove $0.05(24) = 1.2 \approx 1$ observations from the lower and upper side of the data. Hence remove 30 and 470 and recalculate the average of those 22 observations. That is 190.45.

http://www.stat.ufl.edu/~athienit/IntroStat/loc_stats.R

Remark 1.1. To calculate a weighted average when a percentile is located at position P that is between two observed position Closest Lower Position to P (CLP) and the Closest Upper Position to P (CUP)



And the weighted average is going to gice less weight to CUP with corresponding value b, as it's further away, than to CLP with value a.

$$? = b\left(\frac{\text{P} - \text{CLP}}{\text{CUP} - \text{CLP}}\right) + a\left(\frac{\text{CUP} - \text{P}}{\text{CUP} - \text{CLP}}\right) \tag{1.1}$$

The first weight goes to the second value (the largest) and the second weight goes to the first value (the smallest).

Remark 1.2. Note that the mean is more sensitive to outliers-observations that do not fall in the general pattern of the rest of the data-than the median. Assume we have values

$$2, 3, 5.$$

The mean is 3.33 and the median is 3. Now assume we add a value and now have

$$2, 3, 5, 112.$$

The mean is 30.5 but the median is now 4.

1.2.2 Spread

- The **variance** is a measure of spread of the individual observations from their center (as indicated by the mean).

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\left[\sum_{i=1}^{n} x_i^2\right] - n\bar{x}^2\right)$$

  The **standard deviation** is simply the square root of the variance in order to return to the original units of measurement.

- The **range** is the maximum observation - minimum observation.
- The **interquartile range** (IQR) is 75^{th} percentile - 25^{th} percentile (or $Q_3 - Q_1$).

Example 1.2. Continuing from Example 1.1

- we have $\sum x_i^2 = 1152494$, and hence $s^2 = \frac{1}{23}(1152494 - 24(195.4167)^2) = 10260.43$ and $s = \sqrt{10260.43} = 101.2938$.
- The range is $470 - 30 = 440$.
- The IQR is $244.25 - 110.25 = 134$.

http://www.stat.ufl.edu/~athienit/IntroStat/loc_stats.R

1.2.3 Effect of shifting and scaling measurements

As we know measurements can be made in different scales, e.g. cm, m, km, etc and even different units of measurements, e.g. Kelvin, Celsius, Fahrenheit. Let us see how shifting and rescaling influence the mean and variance. Let $x_1, \ldots, x_n$ denote the data and define $y_i = ax_i + b$, where a and b are some constants. Then,

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum ax_i + b = \frac{1}{n}\sum(ax_i + b) = \frac{1}{n}\left(nb + a\sum x_i\right) = a\bar{x} + b,$$

and,

$$s_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\sum(ax_i + b - a\bar{x} - b)^2 = \frac{1}{n-1}a^2\sum(x_i - \bar{x})^2 = a^2 s_x^2$$

1.3 Graphical Summaries

1.3.1 Dot Plot

Stack each observation on a horizontal line to create a dot plot that gives an idea of the "shape" of the data. Some rounding of data values is allowed in order to stack.

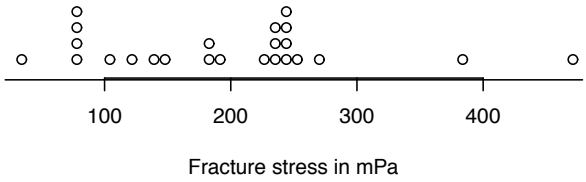


Figure 1.1: Dot plot of data from Example 1.1

1.3.2 Histogram

1. Create class intervals (by choosing boundary points) in which to place the data.
2. Construct a Frequency Table.
3. Draw a rectangle for each class.

It is up to the researcher to decide how many class intervals to create. As a guideline one creates about $2n^{1/3}$ classes. For Example 1.1 that is 5.75 so we can either go with 5 or 6 classes.

Class Interval	Freq.	Relative Freq.	Density
0 -<100	5	5/24=0.208	0.208/100=0.00208
100 -<200	7	7/24=0.292	0.292/100=0.00292
200 -<300	10	0.417	0.00417
300 -<400	1	0.0417	0.000417
400 -<500	1	0.0417	0.000417

Table 1.1: Frequency Table for Example 1.1

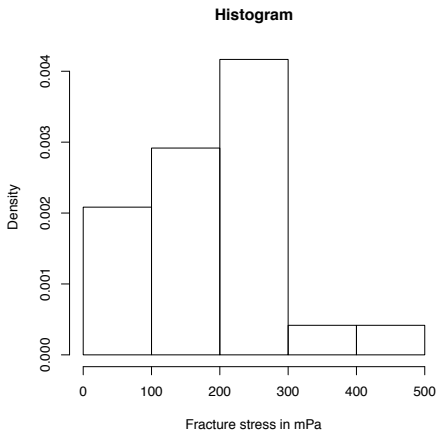


Figure 1.2: Histogram of data from Example 1.1

Remark 1.3. May use Frequency, Relative Frequency or Density as the vertical axis when class widths are equal. However, class widths are not necessarily equal; usually done to create smoother graphics if not mandated by the situation at hand. If this is the case then we must use Density that accounts for the width because large classes may have unrepresentative large frequencies.

http://www.stat.ufl.edu/~athienit/IntroStat/hist1_boxplot1.R

1.3.3 Box-Plot

Box-Plot is a graphic that only uses quartiles. A box is created with Q_1, Q_2, and Q_3. A lower whisker is drawn from Q_1 down to the smallest data point that is within 1.5 IQR of Q_1. Hence from $Q_1 = 110.25$ down to $Q_1 - 1.5IQR = 110.25 - 1.5(134) = -90.75$, but we stop at the smallest point within than which is 30. Similarly the upper whisker is drawn from $Q_3 = 244.25$ to $Q_3 + 1.5IQR = 445.25$ but we stop at the largest point within which is 384.

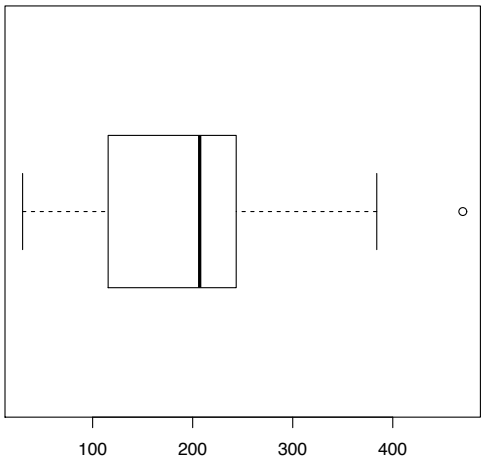


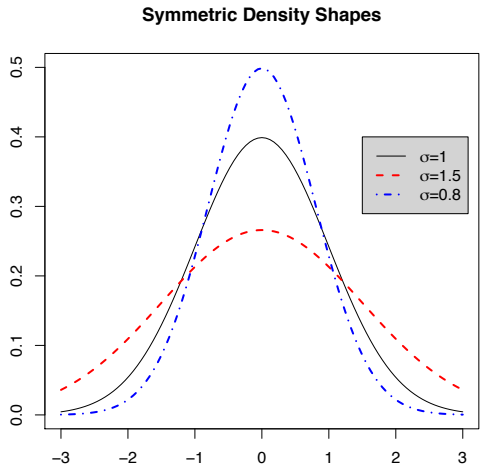
Figure 1.3: Box-Plot of data from Example 1.1

Remark 1.4. Any point beyond the whiskers is classified as an outlier and any point beyond $3IQR$ from either Q_1 or Q_3 is classified as an extreme outlier.

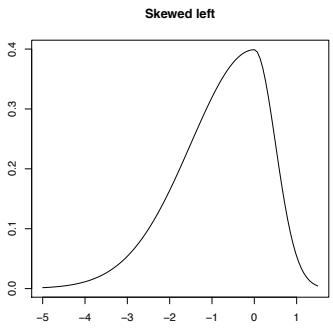
http://www.stat.ufl.edu/~athienit/IntroStat/hist1_boxplot1.R

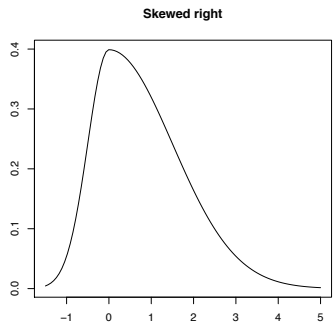
These densities have shapes that can be described as:

- Symmetric

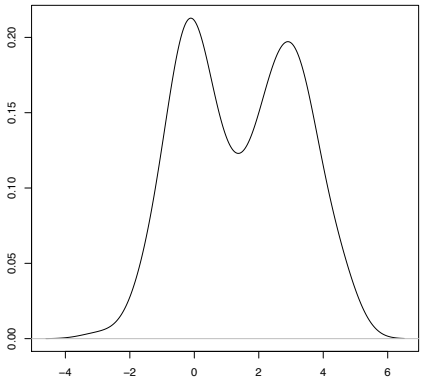


- Skewed Left and Right





- Bi-Modal (or more than two modes)



1.3.4 Pie chart

A pie or circle has 360 degrees. For each category of a variable, the size of the slice is determined by the fraction of 360 that corresponds to that category.

Example 1.3. There is a total of 337,297,000 native English speakers of the world, categorizes as

Country	Pop. (1000)	% of Total	% of pie
USA	226,710	67.21	$0.6721(360) = 241.97°$
UK	56,990	16.90	$60.83°$
Canada	19,700	5.84	$21.02°$
Australia	15,316	4.54	$16.35°$
Other	18,581	5.51	$19.83°$
Total	337,297	100	$360°$

Table 1.2: Frequency table for native English speakers of 1997

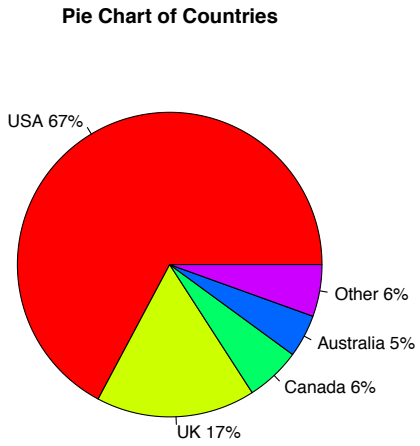


Figure 1.4: Pie chart of English speaking countries

http://www.stat.ufl.edu/~athienit/IntroStat/pie.R

1.3.5 Scatterplot

It is used to plot the raw 2-D points of two variables in an attempt to discern a relationship.

Example 1.4. A small study with 7 subjects on the pharmacodynamics of LSD on how LSD tissue concentration affects the subjects math scores yielded the following data.

Score	78.93	58.20	67.47	37.47	45.65	32.92	29.97
Conc.	1.17	2.97	3.26	4.69	5.83	6.00	6.41

Table 1.3: Math score with LSD tissue concentration

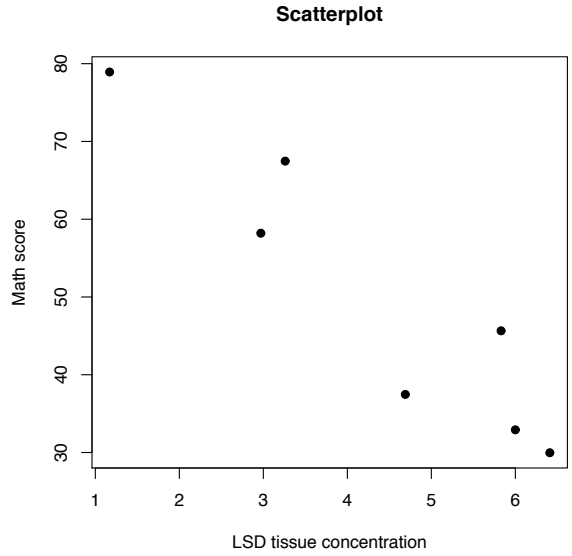


Figure 1.5: Scatterplot of Math score vs. LSD tissue concentration

http://www.stat.ufl.edu/~athienit/IntroStat/scatterplot.R

Module 2

Probability and Random Variables

The study of probability began in the 17^{th} century when gamblers starting hiring mathematicians to calculate the odds of winning for different types of games.

2.1 Sample Space and Events

2.1.1 Basic concepts

Definition 2.1. The set of all possible outcomes of an experiment is called the *sample space* ($\mathcal{S}$) for the experiment.

Example 2.1. Here are some basic examples:

- Rolling a die. Then $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
- Tossing a quarter and a penny. $\mathcal{S} = \{$Hh, Ht, Th, Tt$\}$
- Counting the number of flaws in my personality. $\mathcal{S} = \{1, 2, \ldots\}$
- Machine cuts rods of certain length (in cm). $\mathcal{S} = \{x | 5.6 < x < 6.4\}$

Remark 2.1. Elements in $\mathcal{S}$ may not be equally weighted.

Definition 2.2. A subset of a sample space is called an *event*.

For instance the empty set $\emptyset = \{\}$ and the entire sample space $\mathcal{S}$ are also events.

Example 2.2. Let A be the event of an even outcome when rolling a die. Then, $A = \{2, 4, 6\} \subset \mathcal{S}$.

2.1.2 Relating events

When we are concerned with multiple events within the sample space, Venn Diagrams are useful to help explain some of the relationships. Let's illustrate this via an example.

Example 2.3. Let,

$$\begin{aligned}\mathcal{S} &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \\ A &= \{1, 3, 5, 7, 9\} \\ B &= \{6, 7, 8, 9, 10\}\end{aligned}$$

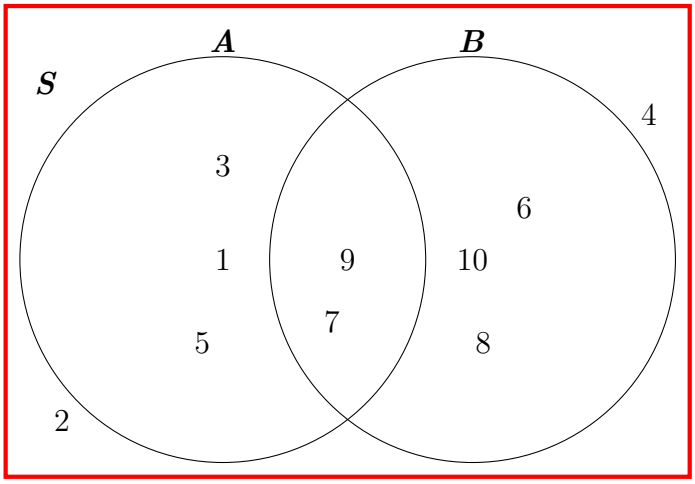


Figure 2.1: Venn Diagram

Combining events implies combining the elements of the events. For example,

$$A \cup B = \{1, 3, 5, 6, 7, 8, 9, 10\}.$$

Intersecting events implies only listing the elements that the events have in common. For example,

$$A \cap B = \{7, 9\}.$$

The **complement of an event** implies listing all the elements in the sample space that are not in that event. For example,

$$A^c = \{2, 4, 6, 8, 10\} \qquad (A \cup B)^c = \{2, 4\}.$$

Remark 2.2. Other useful properties

- De Morgan's law: For any sets $A_1, A_2, \ldots, A_n$, we have
  - $(A_1 \cup A_2 \cup \cdots A_n)^c = A_1^c \cap A_2^c \cap \cdots A_n^c$
  - $(A_1 \cap A_2 \cap \cdots A_n)^c = A_1^c \cup A_2^c \cup \cdots A_n^c$

- Distributive law: For any sets A, B, and C we have
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Definition 2.3. A collection of events $A_1, A_2, \ldots$ is *mutually exclusive* if no two of them have any outcomes in common. That is, $A_i \cap A_j = \emptyset, \forall i, j$

In terms of the Venn Diagram, there is no overlapping between them.

Example 2.4. Let,

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \quad A = \{1, 5\} \quad B = \{6, 8, 10\}$$

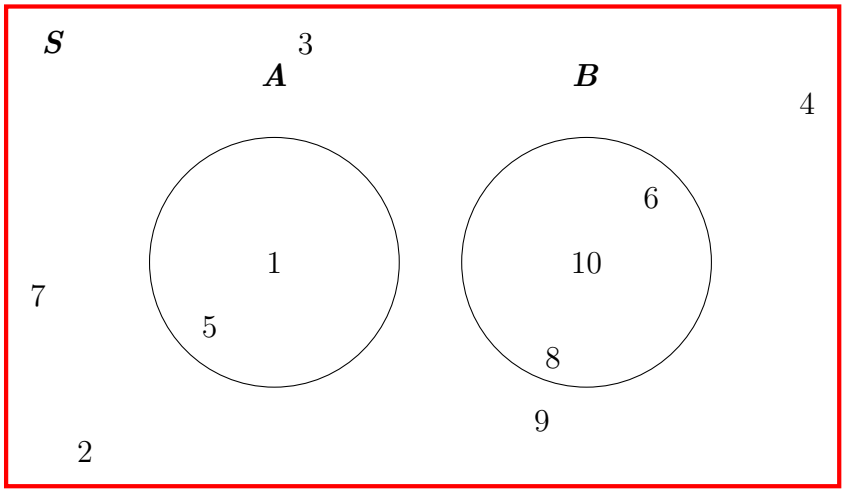


Figure 2.2: Venn Diagram

Example 2.5. A set is always mutually exclusive with its complement. Let,

$$\mathcal{S} = \{1, 2\} \quad A = \{1\}$$

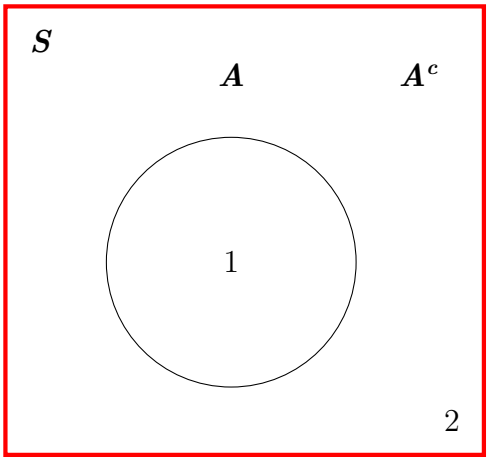


Figure 2.3: Venn Diagram

2.2 Probability

Notation: Let $P(A)$ denote the probability that the event A occurs. It is the proportion of times that the event A would occur in the long run.

Axioms of Probability:

- $P(\mathcal{S}) = 1$
- $0 \leq P(A) \leq 1$, since $A \subseteq \mathcal{S}$
- If $A_1, A_2, \ldots$ are mutually exclusive, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

As a result of the axioms we have that $P(A) = 1 - P(A^c)$ and that $P(\emptyset) = 0$.

Example 2.6. In a computer lab there are 4 computers and once a day a technician inspects them and counts the number of computer crashes. Hence, $\mathcal{S} = \{0, 1, 2, 3, 4\}$ and

Crashes	Probability
0	0.60
1	0.30
2	0.05
3	0.04
4	0.01

Table 2.1: Probabilities for computer crashes

Let A be the event that at least one crash occurs on a given day.

$$\begin{aligned} P(A) &= 0.30 + 0.05 + 0.04 + 0.01 \\ &= 0.4 \end{aligned}$$

or

$$\begin{aligned} &= 1 - P(A^c) \\ &= 1 - 0.60 \\ &= 0.4 \end{aligned}$$

If $\mathcal{S}$ contains N equally likely outcomes/elements and the event A contains $k(\leq N)$ outcomes then,

$$P(A) = \frac{k}{N}$$

Example 2.7. The experiment consists of rolling a die. There are 6 outcomes in the sample space, all of which are equally likely (assuming a fair die). Then, if A is the event of an outcome of a roll being even, $A = \{2, 4, 6\}$ with 3 elements so, $P(A) = 3/6 = 0.5$

The axioms provide a way of finding the probability of a union of two events but only if they are mutually exclusive. In Example 2.3 we have seen that

$$A \cup B = \{1, 3, 5, 6, 7, 8, 9, 10\} \quad \text{and} \quad A \cap B = \{7, 9\}.$$

So, to find the $P(A \cup B)$ we can either

- Break up the set into 3 mutually exclusive sets

  1. $A \cap B^c = \{1, 3, 5\}$
  2. $A \cap B = \{7, 9\}$
  3. $A^c \cap B = \{6, 8, 10\}$

  and by using the axioms

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$$

- Add $P(A) + P(B)$ but we need to subtract the probability of the intersection as that probability was double counted since it is included within A and within B, leading to

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example 2.8. A sample of 1000 aluminum rods is taken and a quality check was performed on each rod's diameter and length.

	Diameter			
Length	Too Thin	OK	Too Thick	Sum
Too Short	10	3	5	18
OK	38	900	4	942
Too Long	2	25	13	40
Sum	50	928	22	1000

- $P(\text{Diameter OK}) = 928/1000 = 0.928$
- $P(\text{Length OK}) = 942/1000 = 0.942$
- $P(\text{Diameter OK} \cap \text{Length OK}) = 900/1000 = 0.9$
- $P(\text{Diameter OK} \cup \text{Length OK}) = 0.928 + 0.942 - 0.9 = 0.97$

2.3 Counting Methods

Proposition 2.1. Product Rule: If the first task of an experiment can result in n_1 possible outcomes and for each such outcome, the second task can result in n_2 possible outcomes, and so forth up to k tasks, then the total number of ways to perform the sequence of k tasks is $\prod_{i=1}^{k} n_i$ □

Example 2.9. When buying a certain type and brand of car a buyer has the following number of choices:

(2) Engine

(5) Color

(4) Interior

Then, the total number of car choices, i.e. number of unique cars, is

$$2 \times 5 \times 4 = 40.$$

Notation: In mathematics, the factorial of a non-negative integer n, denoted by $n!$, is the product of all positive integers less than or equal to n. For example,

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

with $0! = 1$.

2.3.1 Permutations

Definition 2.4. *Permutation* is the number of *ordered* arrangements, or permutations, of r objects selected from n distinct objects ($r \leq n$) *without replacement*. It is given by,

$$P_r^n = \frac{n!}{(n-r)!}.$$

The number of permutations of n objects is $n!$.

Example 2.10. Take the letters A, B, C. Then there are $3! = 6$ possible permutations. These are:

$$ABC, ACB, BAC, BCA, CAB, CBA$$

In effect we have three "slots" in which to place the letters. There are 3 options for the first slot, 2 for the second and 1 for the third. Hence we have $3 \times 2 \times 1 = 6 = P_3^3$.

Example 2.11. Let $n = 10$ with $A, B, C, D, E, F, G, H, I, J$ and we wish to select 5 letters at random, where order is important. Then there are 5 slots with 10 choices for the first slot, 9 for the second, 8 for the third, 7 for the fourth and 6 for the fifth. That is,

$$10 \times 9 \times 8 \times 7 \times 6 = \frac{10!}{5!} = P_5^{10} = 30240.$$

Example 2.12. An election is held for a board consisting of 3 individuals out of a pool of 40 candidates. There will be 3 positions of:

1. President
2. Vice-President
3. Secretary

How many possible boards are there?

Since order is important there are

$$P_3^{40} = \frac{40!}{37!} = 59280$$

That is 40 choices for president, then, 39 choices for vice-president, and finally 38 choices for secretary, i.e. $(40)(39)(38) = 59280$

2.3.2 Combinations

Definition 2.5. *Combination* is the number of *unordered* arrangements, or combinations, of r objects selected from n distinct objects ($r \leq n$) *without replacement*. It is given by,

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{n}{n-r}.$$

The way to think about combinations is that they are a special case or permutations. When selecting r objects from n we know that there are P_r^n permutations but also there are $r!$ different orderings, which for combinations cannot be considered different. Hence,

$$\frac{P_r^n}{r!} = C_r^n.$$

Example 2.13. Referring back to Example 2.10 where there were $P_3^3 = 6$ permutations of the letters. However, there are also $3! = 6$ different orderings. Consequently, there is only 6/6=1 combination.

Example 2.14. Continuing from Example 2.12, assume instead of a board we are interested in the number of possible *committees* where each member has equal power.

Then, there are

$$C_3^{40} = \binom{40}{3} = 9880$$

With combinations we have only seen two groups. Those items chosen and those that were not. A generalization of combinations to more that two groups states that the number of ways of partitioning n distinct objects into k categories containing $n_1, \ldots, n_k$ objects respectively, with $n_1 + \cdots + n_k = n$ is

$$\frac{n!}{n_1! \cdots n_k!}$$

Application of Combinations to Probability Problems. Knowing the total number of combinations of a certain set and knowing the number of combinations for a certain subset we, are able to calculate the probability of an event assuming each possible outcome is equally likely. This is done by dividing the number of ways a certain outcome occurs over the total number of ways.

Example 2.15. A lot of 100 articles contains 10 defective. If a sample of 5 articles is chose at random, the probability that it contains exactly 3 defective is

$$\frac{\binom{10}{3}\binom{90}{2}}{\binom{100}{5}} = \frac{480600}{75287520} = 0.0064$$

- $\binom{10}{3}$ is the number of ways of selecting 3 defective out of 10.
- $\binom{90}{2}$ is the number of ways of selecting 2 non-defective out of 90.
- $\binom{100}{5}$ is the total number of ways of selecting 5 articles out of a 100.

2.4 Conditional Probability and Independence

Definition 2.6. A probability that is based upon the entire sample space is called an *unconditional probability*, but when it is based upon a subset of the sample space it is a *conditional (on the subset) probability*.

Definition 2.7. Let A and B be two events with $P(B) \neq 0$. Then the conditional probability of A given B (has occurred) is

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

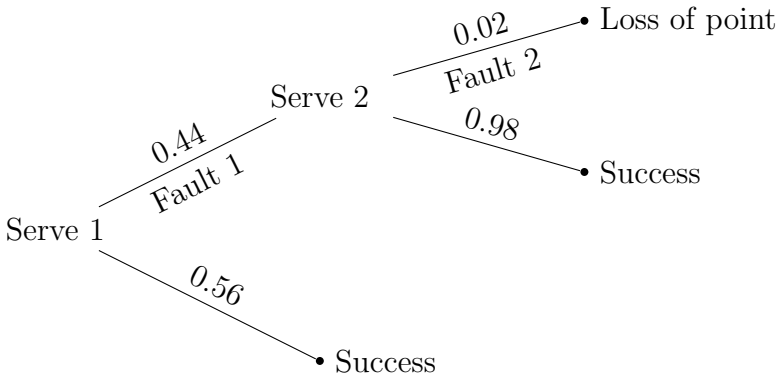
The reason that we divide by the probability of given said occurrence, i.e. $P(B)$ is to re-standardize the sample space. We update the sample space to be just B, i.e. $\mathcal{S} = B$ and hence $P(B|B) = 1$. The only part of event A that occurs within this new $\mathcal{S} = B$ is $P(A \cap B)$.

Proposition 2.2. Rule of Multiplication:

- If $P(A) \neq 0$, then $P(A \cap B) = P(B|A)P(A)$.
- If $P(B) \neq 0$, then $P(A \cap B) = P(A|B)P(B)$.

□

Example 2.16. A player serving at tennis is only allowed one fault. At a double fault the server loses a point/other player gains a point. Given the following information:



What is the probability that the server loses a point, i.e. $P(\text{Fault 1 and Fault 2})$?

$P(\text{Fault 1 and Fault 2}) = P(\text{Fault 2}|\text{Fault 1})P(\text{Fault 1}) = (0.02)(0.44) = 0.009$

Example 2.17. Referring to Example 2.8. Given that the length of a rod is too long, what is the probability that the diameter is okay, i.e.

$$\begin{aligned} P(\text{Diam. OK} \mid \text{Length too long}) &= \frac{P(\text{Diam. OK} \cap \text{Length too long})}{P(\text{ Length too long})} \\ &= \frac{25/1000}{40/1000} \\ &= \frac{25}{40} = 0.625. \end{aligned}$$

2.4.1 Independent Events

When the given occurrence of one event does not influence the probability of a potential outcome of another event, then the two events are said to be independent.

Definition 2.8. Two events A and B are *independent* if the probability of each remains the same, whether or not the other has occurred. If $P(A) \neq 0, P(B) \neq 0$, then

$$P(B|A) = P(B) \Leftrightarrow P(A|B) = P(A).$$

If either $P(A) = 0$, or $P(B) = 0$, then the two events are independent.

Definition 2.9. (Generalization) The events $A_1, \ldots, A_n$ are independent if for each A_i and each collection $A_{j1}, \ldots A_{jm}$ of events with $P(A_{j1} \cap \cdots \cap A_{jm}) \neq 0$,

$$P(A_i|A_{j1} \cap \cdots \cap A_{jm}) = P(A_i)$$

As a consequence of independence, the rule of multiplication then says

$$P(A \cap B) = P(A|B)P(B) \overset{\text{ind.}}{=} P(A)P(B),$$

and in the general case

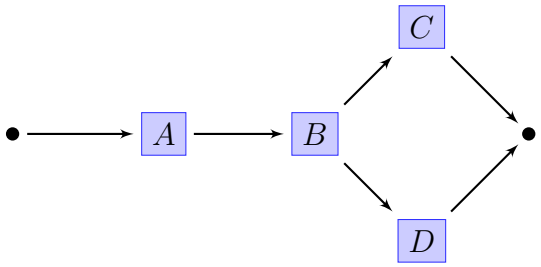
$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i) \quad 0 < k$$

Example 2.18. Of the microprocessors manufactured by a certain process, 20% of them are defective. Assume they function independently. Five microprocessors are chosen at random. What is the probability that they will all work?

Let A_i denote the event that the i^{th} microprocessor works, for $i = 1, 2, 3, 4, 5$. Then,

$$\begin{aligned} P(\text{all work}) &= P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) \\ &= P(A_1)P(A_2)P(A_3)P(A_4)P(A_5) \\ &= 0.8^5 \\ &= 0.328 \end{aligned}$$

Example 2.19. A system consists of four components, connected as shown. Suppose that the components function independently, and that the probabilities of failure are 0.05 for A, 0.03 for B, 0.07 for C, and 0.014 for D. Find the probability that the system functions.



Denote the probability of failure for component A as follows, $P(A) = 0.05$. Hence,

$$\begin{aligned} P(\text{System fncs}) &= P\left(A^c \cap B^c \cap (C^c \cup D^c)\right) \\ &= P(A^c)P(B^c)P(C^c \cup D^c) && \text{by ind.} \\ &= P(A^c)P(B^c)\left[P(C^c) + P(D^c) - P(C^c)P(D^c)\right] && \text{by ind.} \\ &= (0.95)(0.97)\left[(0.93) + (0.986) - (0.93)(0.986)\right] \\ &= 0.9205969 \end{aligned}$$

2.4.2 Law of Total Probability

Recall that the sequence of events $A_1, \ldots, A_n$ is mutually exclusive if no two pairs have any elements in common, i.e. $A_i \cap A_j = \emptyset, \forall\, i, j$. We also say that the sequence is *exhaustive* if the union of all the events is the sample space, i.e. $\cup_{i=1}^{n} A_i = \mathcal{S}$.

Proposition 2.3. Law of Total Probability
If $A_1, \ldots A_n$ are mutually exclusive and exhaustive events, and B is any event, then,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = P(A_1 \cap B) + \cdots + P(A_n \cap B).$$

Equivalently, if $P(A_i) \neq 0$ for each A_i,

$$P(B) = \sum_{i=1}^{n} \underbrace{P(B|A_i)P(A_i)}_{P(A_i \cap B)}.$$

□

To better illustrate this proposition let $n = 4$ and look at Figure 2.4.

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)$$

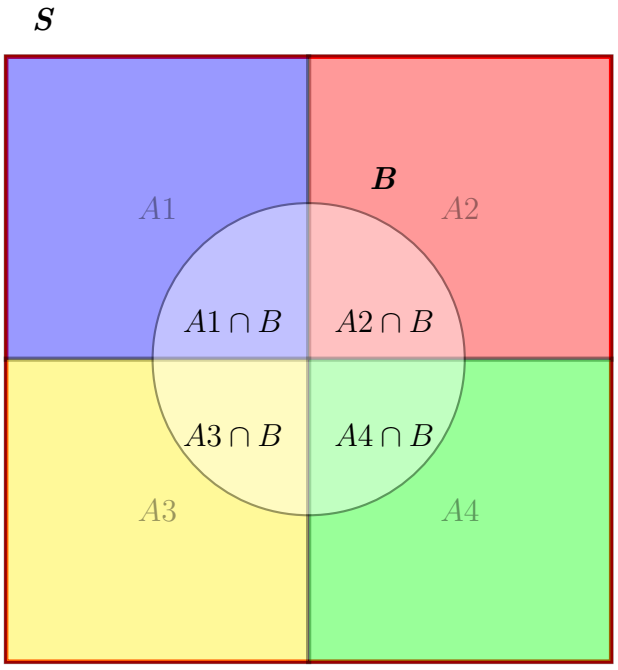


Figure 2.4: Venn Diagram illustrating Law of Total Probability

Example 2.20. Customers can purchase a car with three options for engine sizes

- Small 45% sold
- Medium 35% sold
- Large 20% sold

Of the cars with the small engine 10% fail an emissions test within 10 years of purchase, while 12% fail of the medium and 15% of the large.

What is the probability that a randomly chosen car will fail the emissions test within 10 years?

What we have is:

- $P(S) = 0.45, P(M) = 0.35, P(L) = 0.20$
- $P(F|S) = 0.1, P(F|M) = 0.12, P(F|L) = 0.15$

Therefore

$$P(F) = P(F|S)P(S) + P(F|M)P(M) + P(F|L)P(L) = 0.117$$

2.4.3 Bayes' Rule

In most cases $P(B|A) \neq P(A|B)$. Baye's rule provides a method to calculate one conditional probability if we know the other one. It uses the rule of multiplication in conjunction with the law of total probability.

Proposition 2.4. Bayes's Rule
Special Case: Let A and B be two events with $P(A) \neq 0, P(A^c) \neq 0$, and $P(B) \neq 0$. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

General Case: Let $A_1, \ldots, A_n$ be mutually exclusive and exhaustive events with $P(A_i) \neq 0$ for each $i = 1, \ldots, n$. Let B be any event with $P(B) \neq 0$. Then,

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

□

Example 2.21. In a telegraph signal a dot or dash is sent. Assume that

$$P(\bullet \text{ sent}) = \frac{3}{7}, \quad P(-\text{sent}) = \frac{4}{7}$$

Suppose that there is some interference and with probability 1/8 a dot is mistakenly received on the other end as a dash, and vice versa. Find $P(\bullet \text{ sent}| - \text{received})$.

$$\begin{aligned} P(\bullet \text{ sent}| - \text{received}) &= \frac{P(-\text{received}| \bullet \text{ sent})P(\bullet \text{ sent})}{P(-\text{received}| \bullet \text{ sent})P(\bullet \text{ sent}) + P(- \text{ received}| - \text{sent})P(-\text{sent})} \\ &= \frac{(1/8)(3/7)}{(1/8)(3/7) + (7/8)(4/7)} \\ &= 0.09677 \end{aligned}$$

2.5 Random Variables and Probability Distributions

Definition 2.10. A *random variable* is a function that assigns a numerical value to each outcome of an experiment. It is a measurable function from a probability space into a measurable space known as the state space.

It is an outcome characteristic that is unknown prior to the experiment.

For example, an experiment may consist of tossing two dice. One potential random variable could be the sum of the outcome of the two dice, i.e. $X=$ sum of two dice. Then, X is a random variable that maps an outcome of the experiment (36 of them) into a numerical value (11 of them), the sum in this case.

$$\begin{aligned}(1,1) &\overset{X}{\longmapsto} 2\\ (1,2) &\overset{X}{\longmapsto} 3\\ (1,3) &\overset{X}{\longmapsto} 4\\ &\vdots\\ (6,6) &\overset{X}{\longmapsto} 12\end{aligned}$$

Some times the function is just the identity function, that is, the numerical value assigned is the outcome value. For example, if you are measuring the height of trees (in an agricultural experiment) then the random variable might simply be the height of the tree.

Quantitative random random variables can either be **discrete**, by which they have a countable set of possible values, or **continuous** which have uncountably infinite.

Notation: For a discrete random variable (r.v.) X, the probability distribution is the probability of a certain outcome occurring, called the probability mass function (p.m.f.)

$$P(X = x) = p_X(x).$$

Notation: For a continuous random variable (r.v.) X, the probability density function (p.d.f.), denoted by $f_X(x)$, models the relative frequency of X. Since there are infinitely many outcomes within an interval, the probability evaluated at a singularity is always zero, e.g. $P(X = x) = 0, \forall x$, X being a continuous r.v.

Conditions for a function to be:

- p.m.f. $0 \leq p(x) \leq 1$ and $\sum_{\forall x} p(x) = 1$
- p.d.f. $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$

Example 2.22. (Discrete) Suppose a storage tray contains 10 circuit boards, of which 6 are type A and 4 are type B, but they both appear similar. An inspector selects 2 boards for inspection. He is interested in X = number of type A boards. What is the probability distribution of X?

Since,

$$\begin{aligned}
\text{(A,A)} &\xmapsto{X} 2 \\
\text{(A,B)} &\xmapsto{X} 1 \\
\text{(B,A)} &\xmapsto{X} 1 \\
\text{(B,B)} &\xmapsto{X} 0
\end{aligned}$$

the sample space of X is $\{0, 1, 2\}$. We can calculate the following:

$$\begin{aligned}
p(2) &= P(\text{A on first})P(\text{A on second}|\text{A on first}) \\
&= (6/10)(5/9) = 0.3333
\end{aligned}$$

$$\begin{aligned}
p(1) &= P(\text{A on first})P(\text{B on second}|\text{A on first}) \\
&\quad + P(\text{B on first})P(\text{A on second}|\text{B on first}) \\
&= (6/10)(4/9) + (4/10)(6/9) = 0.5333
\end{aligned}$$

$$\begin{aligned}
p(0) &= P(\text{B on first})P(\text{B on second}|\text{B on first}) \\
&= (4/10)(3/9) = 0.1334
\end{aligned}$$

Consequently,

$X = x$	$p(x)$
0	0.1334
1	0.5333
2	0.3333
Total	1.0

Table 2.2: Probability Distribution of X

Example 2.23. (Continuous) The lifetime of a certain battery has a distribution that can be approximated by $f(x) = 0.5e^{-0.5x}$, $x > 0$. Note that this is the short way of writing the piecewise function

$$f(x) = \begin{cases} 0.5e^{-0.5x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

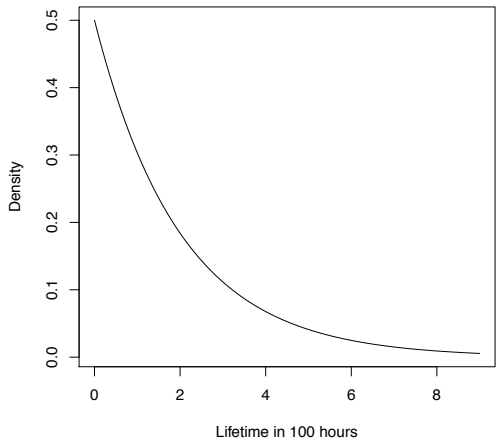


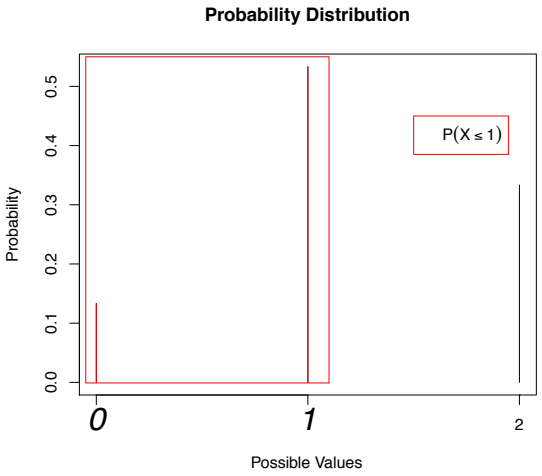
Figure 2.5: Probability density function of battery lifetime.

Notation: You may recall that $\int f(t)dt$ is contrived from $\lim \sum f(t_i)\Delta_i$. Hence for the following definitions and expressions we will only be using notation for continuous variables and wherever you see "$\int$" simply replace it with "$\sum$".

Definition 2.11. The cumulative distribution function (c.d.f.) of a r.v. X is denoted by $F_X(x)$ and defined as

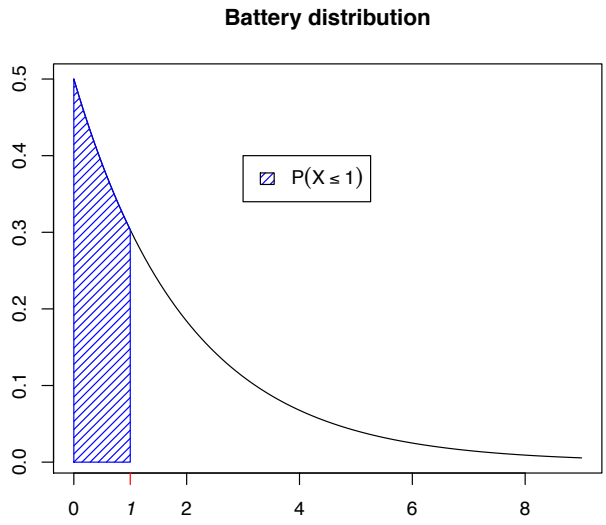
$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \left(\overset{discrete}{=} \sum_{t \leq x} p(t) \right)$$

Example 2.24. Example 2.22 continued.
Find $F(1)$. That is,



$$\begin{aligned} F(1) &= P(X \leq 1) \\ &= P(X = 0) + P(X = 1) \\ &= 0.1334 + 0.5333 = 0.6667 \end{aligned}$$

Example 2.25. Example 2.23 continued.
Find $F(1)$. That is,



$$\begin{aligned} F(1) &= \int_{-\infty}^{1} f(x)dx \\ &= \int_{-\infty}^{0} 0dx + \int_{0}^{1} 0.5e^{-0.5x}dx \\ &= 0 + (-e^{-0.5x})|_0^1 = 0.3935 \end{aligned}$$

or in software such as Wolfram Alpha, input:

```
integrate 0.5*e^(-0.5*x) dx from 0 to 1
```

2.5.1 Expected Value And Variance

The expected value of a r.v. is thought of as the long term average for that variable. Similarly, the variance is thought of as the long term average of values of the r.v. to the expected value.

Definition 2.12. The *expected value* (or mean) of a r.v. X is

$$\mu_X := E(X) = \int_{-\infty}^{\infty} xf(x)dx \left(\overset{discrete}{=} \sum_{\forall x} xp(x) \right).$$

In actuality, this definition is a special case of a much broader statement.

Definition 2.13. The expected value (or mean) of function $h(\cdot)$ of a r.v. X is

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx.$$

Due to this last definition, if the function h performs a simple linear transformation, such as $h(t) = at + b$, for constants a and b, then

$$E(aX + b) = \int (ax + b)f(x)dx = a \int xf(x)dx + b \int f(x)dx = aE(X) + b$$

Example 2.26. Referring back to Example 2.22, the expected value of the number of type A boards (X) is

$$E(X) = \sum_{\forall x} xp(x) = 0(0.1334) + 1(0.5333) + 2(0.3333) = 1.1999.$$

We can also calculate the expected value of (i) $5X + 3$ and (ii) $3X^2$.

(i) $5(1.1999) + 3 = 8.995$ since $E(5X + 3) = 5E(X) + 3$

(ii) $3(0^2)(0.1334) + 3(1^2)(0.5333) + 3(2^2)(0.3333) = 5.5995$

Definition 2.14. The *variance* of a r.v. X is

$$\begin{aligned}
\sigma_X^2 := V(X) &= E\left[(X - \mu_X)^2\right] \\
&= \int (x - \mu_X)^2 f(x)dx \\
&= \int (x^2 - 2x\mu_X + \mu_X^2) f(x)dx \\
&= \int x^2 f(x)dx - 2\mu_X \int xf(x)dx + \mu_X^2 \int f(x)dx \\
&= E(X^2) - 2E^2(X) + E^2(X) \\
&= E(X^2) - E^2(X)
\end{aligned}$$

Example 2.27. This refers to Example 2.22. We know that $E(X) = 1.1999$ and $E(X^2) = 0^2(0.1334) + 1^2(0.5333) + 2^2(0.3333) = 1.8665$. Thus,

$$\begin{aligned} V(X) &= E(X^2) - E^2(X) \\ &= 1.8665 - 1.1999^2 \\ &= 0.42674 \end{aligned}$$

Example 2.28. This refers to example 2.23. If we were to do this by hand we would need to do integration by parts (multiple times). However we can use software such as Wolfram Alpha.

1. Find $E(X)$, so in Wolfram Alpha simply input:

   ```
   integrate x*0.5*e^(-0.5*x) dx from 0 to infinity
   ```

   So $E(X) = 2$.

2. Find $E(X^2)$, so input:

   ```
   integrate x^2*0.5*e^(-0.5*x) dx from 0 to infinity
   ```

   So, $E(X^2) = 8$.

3. $V(X) = E(X^2) - E^2(X) = 8 - 2^2 = 4$.

Definition 2.15. The variance of a function h of a r.v. X is

$$\begin{aligned} V(h(X)) &= \int [h(x) - E(h(X))]^2 f(x) dx \\ &= E(h^2(X)) - E^2(h(X)) \end{aligned}$$

Notice that if h stands for a linear transformation function then,

$$\begin{aligned} V(aX + b) &= E\left[\{aX + b - E(aX + b)\}^2\right] \\ &= a^2 E\left[\{X - E(X)\}^2\right] \\ &= a^2 V(X) \end{aligned}$$

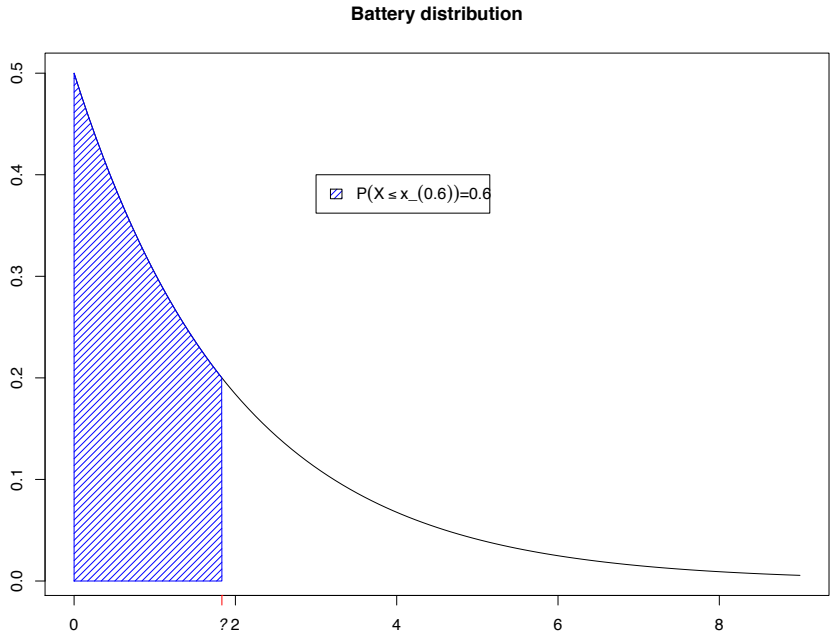
2.5.2 Population Percentiles

Let X be a continuous r.v. with p.d.f. f and c.d.f. F. The population p^{th} percentile, x_p is found by solving the following equation for x_p

$$F(x_p) = \int_{-\infty}^{x_p} f(t)dt = \frac{p}{100}.$$

Example 2.29. Let r.v. X have p.d.f. $f(x) = 0.5e^{-0.5x}, x > 0$. The 60th percentile of X is found by solving for x_m in

$$F(x_{0.6}) = \int_0^{x_{0.6}} 0.5e^{-0.5t}dt = 0.6.$$



$$\begin{aligned} F(x) = \int_0^x 0.5e^{-0.5t}dt &= \frac{0.5}{-0.5}e^{-0.5t}|_0^x \\ &= -e^{-0.5x} + 1. \end{aligned}$$

Hence, we need to solve

$$F(x_{0.6}) = -e^{-0.5x_{0.6}} + 1 = 0.6 \Rightarrow x_{0.6} = 1.83258$$

Or once again use Wolfram Alpha

```
(integrate 0.5*e^(-0.5*x) dx from 0 to p) equals 0.6
```

Example 2.30. Refer back to Example 2.22.

$X = x$	$p(x)$	$F(x)$
0	0.1334	0.1334
1	0.5333	0.6667
2	0.3333	1.0000
Total	1.0000	

Table 2.3: Probability Distribution of X

Hence, the median, $\tilde{\mu}$, is between 0 and 1. Using a weighted average

$$\frac{0.6667 - 0.5}{0.6667 - 0.1334}(0) + \frac{0.5 - 0.1334}{0.6667 - 0.1334}(1) = 0.687418$$

2.5.3 Chebyshev's inequality

This is useful concept when the distribution of a r.v. is unknown.

Proposition 2.5. Let X be a random variable with $E(X) = \mu$ and $V(X) = \sigma^2$. Then,

$$P\left(|X - \mu| < k\sigma\right) \geq 1 - \frac{1}{k^2} \Leftrightarrow P\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

□

This proposition implies (through the second term) that the probability that a random variable differs from its mean by standard deviations or more is never greater than $1/k^2$.

Example 2.31. A widget manufacturer is interested in the largest proportion of days when production is not in the optimal range of 100 to 140 widgets. The manufacturing process is known to produce on average 120 widgets with a standard deviation of 10.

$$\begin{aligned} P(X \leq 99 \text{ or } X \geq 141) &= P(|X - 120| \geq 21) \\ &= P(|X - 120| \geq 2.1(10)) \\ &\leq \frac{1}{2.1^2} = 0.2268 \end{aligned}$$

2.5.4 Jointly distributed random variables

When two or more random variables are associated with each item in a population, the random variables are said to be *jointly distributed*.

Example 2.32. Let X denote the length and Y denote the width of a box in mm.

	y		
x	15	16	$p_X(x)$
129	0.12	0.08	0.20
130	0.42	0.28	0.70
131	0.06	0.04	0.10
$p_Y(y)$	0.60	0.40	1.00

For example, the probability that a box has length 15mm and width 129mm is

$$P(X = 129 \cap Y = 15) = 0.12.$$

We also implement the law of total probability to find marginal probabilities such as

$$\begin{aligned} P(X = 129) &= \underbrace{P(X = 129 \cap Y = 15)}_{0.12} + \underbrace{P(X = 129 \cap Y = 16)}_{0.08} \\ &= 0.20 \end{aligned}$$

Finding marginal probabilities or one r.v. involves summing out or integrating out the other(s).

$$f_X(x) = \int f(x,y)dy \qquad f_Y(y) = \int f(x,y)dx$$

where $f(x,y)$ denotes the joint p.d.f.

In the continuous case, we have seen that finding probabilities involves finding the area under the p.d.f. Hence,

$$P(a < X < b \text{ and } c < Y < d) = \int_a^b \int_c^d f(x,y)dydx$$

Example 2.33. For a certain type of washer, both the thickness and the hole diameter vary from item to item. Let X denote the thickness and Y the diameter, both in mm. Assume that the joint p.d.f. is given by

$$f(x,y) = \frac{1}{6}(x+y) \qquad x \in [1,2],\ y \in [4,5]$$

Find the probability that a randomly chosen washer has a thickness between 1.0 and 1.5mm and a hole diameter between 4.5 and 5mm.

What we need to find then is

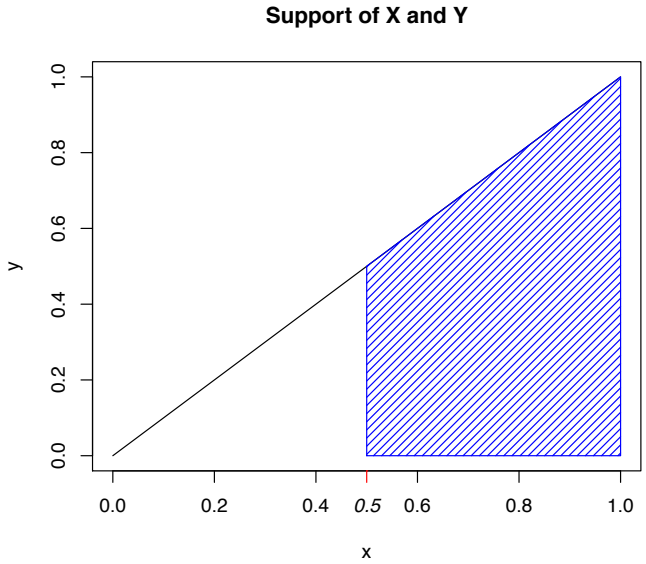
$$\begin{aligned} P(1 \le X \le 1.5 \text{ and } 4.5 \le Y \le 5) &= \int_1^{1.5} \int_{4.5}^{5} (1/6)(x+y)\ dydx \\ &= \int_1^{1.5} \left(\frac{x}{12} + \frac{19}{48}\right)\ dx \\ &= 1/4. \end{aligned}$$

Example 2.34. Let X and Y be random variables with p.d.f.

$$f(x,y) = 8xy \quad x \in [0,1],\ y \in [0,x]$$

We wish to find

- $P(X > 0.5 \text{ and } Y < X)$. All we need to do is integrate the p.d.f over the joint support



$$\begin{aligned} P(X > 0.5 \text{ and } Y < X) &= \int_{0.5}^{1}\int_{0}^{x} 8xy\ dydx \\ &= \int_{0.5}^{1} 4x^3\ dx \\ &= 0.9375. \end{aligned}$$

- The marginals
  - of X

$$\begin{aligned} f_X(x) &= \int_0^x 8xy\ dy \\ &= 4x^3 \qquad 0 \le x \le 1 \end{aligned}$$

  - of Y

$$\begin{aligned} f_Y(y) &= \int_y^1 8xy\ dx \\ &= 4y(1-y^2) \qquad 0 \le y \le 1 \end{aligned}$$

Expanding definition 2.13 we have

Definition 2.16. The expected value of a function $h(\cdot)$ of r.vs X and Y is

$$E\left[h(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)\ dxdy$$

2.5.5 Conditional distributions

Definition 2.17. Let X and Y be jointly continuous r.vs with joint p.d.f. $f(x, y)$. Let $f_X(x)$ denote the marginal p.d.f. of X and let x be any number for which $f_X(x) > 0$. The conditional p.d.f. of $Y|X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Example 2.35. Continuing from Example 2.33, it is easy to show that the marginal distribution of X is $(1/6)(x + 4.5)$ for $x \in [1, 2]$. To find the p.d.f. of Y given that $X = 1.2$

$$\begin{aligned} f_{Y|X}(y|1.2) &= \frac{\cancel{(1/6)}(1.2 + y)}{\cancel{(1/6)}(1.2 + 4.5)} \\ &= \frac{1.2 + y}{5.7} \qquad y \in [4, 5] \end{aligned}$$

We can use this distribution to find probabilities such as the probability $P(Y \leq 4.8|X = 1.2)$. Do as an exercise.

Knowing the conditional distribution of one r.v., Y, given another, X, allows us to find the *conditional expectation* of the r.v. $Y|X = x$ which is

$$E(Y|X = x) = \int y f_{Y|X}(y|x)\, dy$$

Example 2.36. In the previous example we had that $f_{Y|X}(y|1.2) = (1.2 + y)/5.7$ if $y \in [4, 5]$. Hence,

$$\begin{aligned} E(Y|X = 1.2) &= \int_4^5 y \frac{1.2 + y}{5.7} dy \\ &= \frac{1}{5.7} \left[0.6y^2 + \frac{y^3}{3} \right]_4^5 \\ &= 4.51462 \end{aligned}$$

2.5.6 Independent random variables

Definition 2.18. Let X and Y be two continuous r.vs. They are independent if

$$f(x,y) = f_X(x)f_Y(y)$$

This provides us with a method to determine from a joint p.d.f. if two r.vs are independent. We need to be able to partition the joint p.d.f. into a product of the two marginal p.d.fs. This may not be a trivial task because we have to make sure that the two partitioned functions are actually p.d.fs, that is that they are greater than or equal to 0 and that they integrate to 1 when integrated over the whole support. Fortunately the following lemma makes matters easier.

Lemma 2.6. *Two r.vs X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ strictly either nonnegative or nonpositive such that for every $x, y \in \mathbb{R}$*

$$f(x,y) = g(x)h(y)$$

Proof. ($\Rightarrow$) If X and Y are independent then by definition $f(x,y) = f_X(x)f_Y(y)$. Let $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.

($\Leftarrow$) Define $c := \int_{-\infty}^{\infty} g(x)dx$ and $d := \int_{-\infty}^{\infty} h(y)dy$. Note that

$$\begin{aligned} cd &= \int\int g(x)h(y)dxdy \\ &= \int\int f(x,y)dxdy \\ &= 1 \end{aligned}$$

The marginal p.d.f. of X is given by

$$f_X(x) = \int f(x,y)dy = \int g(x)h(y)dy = g(x)\int h(y)dy = g(x)d.$$

Similarly $f_Y(y) = h(y)c$ and therefore

$$f(x,y) = g(x)h(y) = g(x)h(y)cd = [g(x)d][h(y)c] = f_X(x)f_Y(y).$$

□

Example 2.37. Consider

$$f(x,y) = \frac{1}{384}x^2y^4e^{-y-x/2} \qquad x > 0, y > 0.$$

Since,

$$f(x,y) = \underbrace{\frac{1}{384}x^2e^{-x/2}I(x>0)}_{g(x)}\underbrace{e^{-y}y^4I(y>0)}_{h(y)}$$

where $I(\cdot)$ denotes the indicator (0,1) function. It is clear that X and Y independent.

Example 2.38. Back in Example 2.34 we had p.d.f. $8xy$ for $x \in [0,1]$ and $y \in [0,x]$ which can be expressed as

$$f(x,y) = 8xI(x \in [0,1])yI(y \in [0,x]).$$

However, $I(y \in [0,x])$ cannot be further decomposed into a function simply of x and one simply of y. Hence, X and Y are not independent.

2.5.7 Covariance

The population covariance is a measure of strength of a linear relationship among two variables. It is not a measure of the slope of the linear relationship but how close points lie to a straight line. See example 2.41

Definition 2.19. Let X and Y be two r.vs. The population *covariance* of X and Y is

$$\begin{aligned}\text{Cov}(X,Y) &= E\left[(X - E(X))\,(Y - E(Y))\right] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

Remark 2.3. If X and Y are *independent*, then

$$\begin{aligned}E(XY) &= \int\int xyf(x,y)dxdy \\ &\overset{\text{ind.}}{=} \int\int xyf_X(x)f_Y(y)dxdy \\ &= \int xf_X(x)dx \int yf_Y(y)dy \\ &= E(X)E(Y)\end{aligned}$$

and consequently $\text{Cov}(X,Y) = 0$. This is because under independence $f(x,y) = f_X(x)f_Y(y)$. However, the converse is not true. Think of a circle such as $\sin^2 X + \cos^2 Y = 1$. Obviously, X and Y are dependent but they have no linear relationship. Hence, $\text{Cov}(X,Y) = 0$.

The covariance is not unitless so a measure called the population *correlation* is used to describe the strength of the linear relationship that is

- unitless
- ranges from -1 to 1

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}},$$

A negative relationship implies a negative covariance and consequently a negative correlation.

Example 2.39. Recall that in Example 2.34 we had two r.vs X and Y with joint p.d.f.

$$f(x,y) = 8xy \qquad x \in [0,1],\ y \in [0,x].$$

To find $\text{Cov}(X,Y)$ we will have to find the three terms $E(XY), E(X)$ and $E(Y)$.

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^x xy(8xy)dydx \\ &= \int_0^1 (8x^5)/3dx \\ &= 4/9 \end{aligned}$$

We had also shown that the marginal p.d.fs where

- $f_X(x) = 4x^3 \qquad 0 \leq x \leq 1$
- $f_Y(y) = 4y - 4y^3 \qquad 0 \leq y \leq 1$

and therefore, $E(X) = 4/5$ and $E(Y) = 8/15$. Thus,

$$\text{Cov}(X,Y) = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = 0.01778.$$

To find the correlation, it can be shown that $V(X) = 0.02667$ and $V(Y) = 0.04889$. Therefore,

$$\rho_{XY} = \frac{0.01778}{\sqrt{0.02667}\sqrt{0.04889}} = 0.49239$$

Moving away from the population parameters, to estimate the sample statistic of the covariance and the correlation we need

$$\begin{aligned} \hat{\sigma}_{XY} := \widehat{\text{Cov}(X,Y)} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}\right] \end{aligned}$$

Therefore,

$$r_{XY} := \hat{\rho}_{XY} = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{(n-1)s_X s_Y}.$$

Example 2.40. Let's assume that we want to look at the relationship between two variables, height (in inches) and self esteem for 20 individuals.

Height	68	71	62	75	58	60	67	68	71	69
Esteem	4.1	4.6	3.8	4.4	3.2	3.1	3.8	4.1	4.3	3.7
	68	67	63	62	60	63	65	67	63	61
	3.5	3.2	3.7	3.3	3.4	4.0	4.1	3.8	3.4	3.6

Table 2.4: Height to self esteem data

Hence,

$$r_{XY} = \frac{4937.6 - 20(65.4)(3.755)}{19(4.406)(0.426)} = 0.731$$

there is a moderate to strong positive linear relationship.

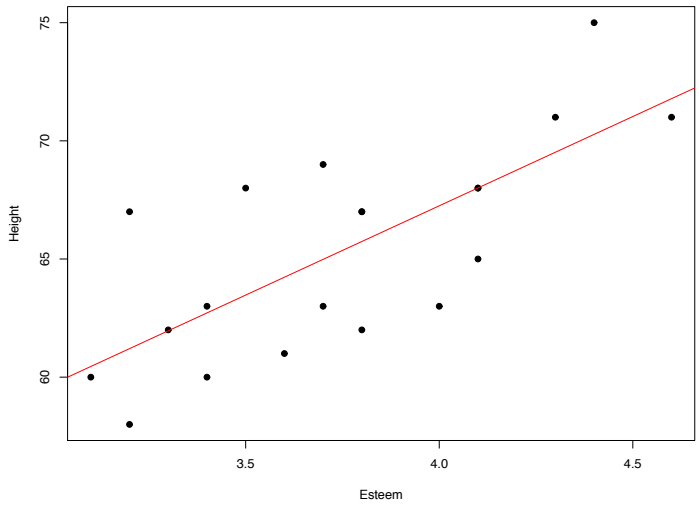


Figure 2.6: Scatterplot with linear relationship

http://www.stat.ufl.edu/~athienit/IntroStat/esteem.R

Example 2.41. Here are some examples that illustrate different sample correlations. Again we note that correlation measures the strength of the linear relationship and not the slope.

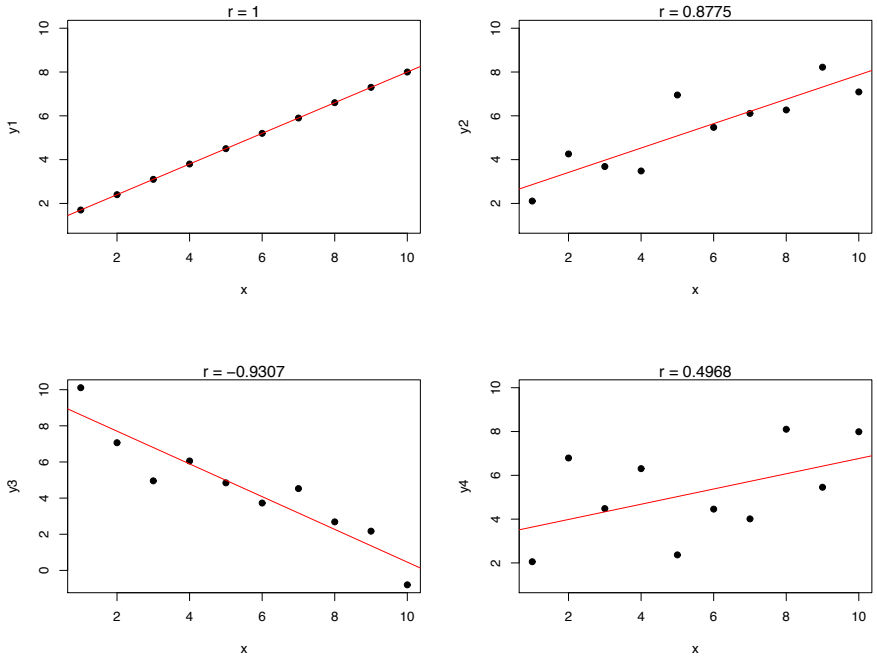


Figure 2.7: Scatterplot with linear relationship

2.5.8 Mean and variance of linear combinations

Let X and Y be two r.vs, for $(aX+b)+(cY+d)$ for constants a, b, c and d,

$$E(aX+b+cY+d) = aE(X)+cE(Y)+b+d$$
$$V(aX+b+cY+d) = \underbrace{\text{Cov}(aX,aX)}_{a^2V(X)} + \underbrace{\text{Cov}(cY,cY)}_{c^2V(Y)} + \underbrace{\text{Cov}(aX,cY)+\text{Cov}(cY,aX)}_{2ac\text{Cov}(X,Y)}$$

Example 2.42. Let X be a r.v. with $E(X)=3$ and $V(X)=2$, and Y be another r.v. independent of X with $E(Y)=-5$ and $V(Y)=6$. Then,

$$E(X-2Y) = E(X)-2E(Y) = 3-2(-5) = 13$$

and

$$V(X-2Y) = (1)^2V(X)+(-2)^2V(Y)+2(1)(-2)\cancelto{0}{\text{Cov}(X,Y)} = 2+4(6) = 26$$

Now we extend these two concepts to more than two r.vs. Let $X_1,\ldots,X_n$ be a sequence of r.vs and $a_1,\ldots,a_n$ a sequence of constants. Then the r.v. $\sum_{i=1}^n a_iX_i$ has mean and variance

$$E\left(\sum_{i=1}^n a_iX_i\right) = \sum_{i=1}^n a_iE(X_i)$$

and

$$V\left(\sum_{i=1}^n a_iX_i\right) = \sum_{i=1}^n\sum_{j=1}^n a_ia_j\text{Cov}(X_i,X_j) \tag{2.1}$$
$$= \sum_{i=1}^n a_i^2V(X_i) + 2\sum\sum_{i<j} a_ia_j\text{Cov}(X_i,X_j) \tag{2.2}$$

Example 2.43. Assume the random sample, i.e. independent identically distributed (i.i.d.) r.vs, $X_1,\ldots,X_n$ are to be obtained and of interest will be the specific linear combination corresponding to the sample mean $\bar{X} = (1/n)\sum_{i=1}^n X_i$. Since the r.vs are i.i.d., let $E(X_i)=\mu$ and $V(X_i)=\sigma^2$ $\forall i=1,\ldots,n$. Then,

$$E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}n\mu = \mu$$

and

$$V\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \overset{\text{ind.}}{=} \frac{1}{n^2}\sum_{i=1}^n V(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Remark 2.4. As the sample size increases, the variance of the sample mean decreases with $\lim_{n\to\infty} V(\bar{X}) = 0$. That is, there is no uncertainty in the sample mean, which is the estimate of the population mean. With $n\to\infty$ the sample mean is the population mean.

2.5.9 Common Discrete Distributions

In the following sections we will be reviewing some of the most frequently used discrete distributions. Probability calculations can be software once a p.m.f. is specified for any distribution, however for these common ones software have built-in p.m.f., c.d.f, quantile (inverse function of c.d.f.) function and so on.

Bernoulli

Imagine an experiment where the r.v. X can take only two possible outcomes,

- success $(X = 1)$ with some probability p
- failure $(X = 0)$ with probability $1 - p$.

The p.m.f. of X is

$$p(x) = p^x(1-p)^{1-x} \qquad x = 0, 1 \quad 0 \leq p \leq 1$$

and we denote this by stating $X \sim \text{Bernoulli}(p)$. The mean of X is

$$E(X) = \sum_{\forall x} xp(x) = 0p(0) + 1p(1) = p,$$

and the variance is

$$V(X) = E(X^2) - E^2(X) = [0^2p(0) + 1^2p(1)] - p^2 = p - p^2 = p(1-p).$$

Example 2.44. A die is rolled and we are interested in whether the outcome is a 6 or not. Let,

$$X = \begin{cases} 1 & \text{if outcome is 6} \\ 0 & \text{otherwise} \end{cases}$$

Then, $X \sim \text{Bernoulli}(1/6)$ with mean $1/6$ and variance $5/36$.

Binomial

If $X_1, \ldots, X_n$ correspond to n Bernoulli trials conducted where

- the trials are independent
- each trial has identical probability of success p
- the r.v. X is the total number of successes

then $X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$. The the intuition behind the form of the p.m.f. can be motivated by the following example.

Example 2.45. A fair coin is tossed 10 times and $X =$ the number of heads is recorded. What is the probability that $X = 3$?

One possible outcome is

$$\text{(H) (H) (H) (T) (T) (T) (T) (T) (T) (T)}$$

The probability of this outcome occurring in exactly this order is $p^3(1-p)^7$. However there are $\binom{10}{3} = \frac{10!}{3!7!} = 120$ possible ways of 3 Heads and 7 Tails since order is not important.

Consequently, the p.m.f. of $X \sim \text{Bin}(n, p)$ is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

with $E(X) = np$ and $V(X) = np(1-p)$.

Another variable of interest concerning experiments with binary outcomes is the proportion of successes $\hat{p} = X/n$. Note that $\hat{p}$ is simply the r.v. X multiplied by a constant, $1/n$. Hence,

$$E(\hat{p}) = E(X/n) = \frac{np}{n} = p$$

and

$$V(\hat{p}) = V(X/n) = \frac{1}{n^2} V(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Example 2.46. A die is rolled 4 times and the number of 6s is observed. Find the probability that there is at least one 6.

Let, X be the number of 6s which implies $X \sim \text{Bin}(4, 1/6)$.

$$\begin{aligned} P(X \geq 1) &= \sum_{i=1}^{4} \binom{4}{i} \left(\frac{1}{6}\right)^i \left(1 - \frac{1}{6}\right)^{4-i} \\ &= 1 - P(X < 1) \\ &= 1 - P(X = 0) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^4 \\ &= 0.518 \end{aligned}$$

In `R`, one would simple use the c.d.f. function `pbinom()` in

```
1-pbinom(0,4,1/6)
```

Also, $E(X) = 4(1/6) = 2/3$ and $V(X) = 4(1/6)(5/6) = 5/9$.

The expected value of the proportion of 6s which is $E(\hat{p}) = 1/6$ and has variance $V(\hat{p}) = (5/36)/4 = 5/144$.

Geometric

Assume that a sequence of i.i.d. Bernoulli trials is conducted and of interest is the number of trials necessary to achieve the first success. Let, X denote the total number of trials up to and including the first success. Then, $X \sim \text{Geom}(p)$ with p.m.f.

$$p(x) = p(1-p)^{x-1} \qquad x = 1, 2, \ldots$$

and $E(X) = 1/p$ and $V(X) = (1-p)/p^2$.

Example 2.47. In an experiment, such as tossing a fair coin, what is the probability that it will take the experimenter exactly 5 attempts to land the coin heads up, i.e. $P(X = 5)$?

This implies that the outcome Head (H) is preceded by 4 Tails.

$$\text{(T) (T) (T) (T) (H)}$$

The probability of this outcome is $(1-p)^4 p$, where p denotes the probability of head (success) on each try.

Remark 2.5. In `R`, we would use the function `dgeom(4,p)` since the function in `R` counts the number of failures before the success. In this case there are 4 failures before the first success. This is simply an alternate form. Please look up "geometric" on wikipedia or in `R` help files.

Negative Binomial

A r.v. with a negative binomial distribution is simple an extension of the geometric. It is the number of trials up to and including the r^{th} success. So if $X \sim \text{NB}(r, p)$ then it has p.m.f.

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, \ldots$$

with $E(X) = r/p$ and $V(X) = r(1-p)/p^2$.

Example 2.48. Continuing the coin toss scenario assume we have 8 tosses until the 3^{rd} Head with the following outcome:

$$\text{(T) (T) (H) (T) (H) (T) (T) (H)}$$

The probability of this occurring is $p^3(1-p)^5$. This is one possibility that 3 Heads occur in 8 tosses with the restriction that the last Head occurs on the 8^{th} trial. Consequently, there are 7 prior positions to which to place 2 Heads, so there are $\binom{7}{2}$ such ways. Hence the probability of 8 trials necessary to achieve the 3^{rd} Head is $\binom{7}{2} p^3 (1-p)^5$.

Remark 2.6. In `R`, we would use the function `dnbinom(5,3,p)` since the function in `R` counts the number of failures before r^{th} success. In this case there are 5 failures before the 3^{rd} success. This is simply an alternate form. Please look up "negative binomial" on wikipedia or in `R` help files.

Remark 2.7. A negative binomial is a sum of geometric r.vs, in the same way a binomial is a sum of Bernoullis. In our example

$$\underbrace{\text{(T) (T) (H)}}_{\text{Geometric}}\underbrace{\text{(T) (H)}}_{\text{Geometric}}\underbrace{\text{(T) (T) (H)}}_{\text{Geometric}}$$

is a sum of 3 geometric r.vs. Hence, if $X \sim \text{NB}(r, p)$ then

$$X = \sum_{i=1}^{r} Y_i \qquad \text{where } Y_i \overset{ind.}{\sim} \text{Geom}(p)$$

Poisson

The Poisson distribution occurs when we count the number of occurrences of an event over a given interval of time and/or space. These occurrences are assumed to occur with a fixed rate, for example the number of particles that decay in a radioactive process.

If $X \sim \text{Poisson}(\lambda)$ then it has p.m.f.

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, \ldots \quad \lambda > 0$$

with $E(X) = V(X) = \lambda$.

Example 2.49. Assume that the number of hits a website receives during regular business hours occurs with a mean rate of 5 hits per minute. Find the probability that there will be exactly 17 hits in the next 3 minutes.

Denote X to be the number of hits in the next 3 minutes. Hence, $X \sim \text{Poisson}(5 \times 3)$ and

$$P(X = 17) = \frac{15^{17} e^{-15}}{17!} = 0.0847$$

and in `R`, `dpois(17,15)`

In a sample setting, where we collect data, assume a Poisson distribution (with unknown rate) we can estimate λ, similar to estimating p for the binomial. If $X \sim \text{Poisson}(\lambda \times t)$, then

- estimate $\hat{\lambda} = X/t$
- with (estimated) uncertainty

$$\hat{\sigma}_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{t}}$$

2.5.10 Common Continuous Distributions

Uniform

A continuous r.v. that places equal weight to all values within its support, $[a, b]$, $a \leq b$, is said to be a uniform r.v. It has p.d.f.

$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

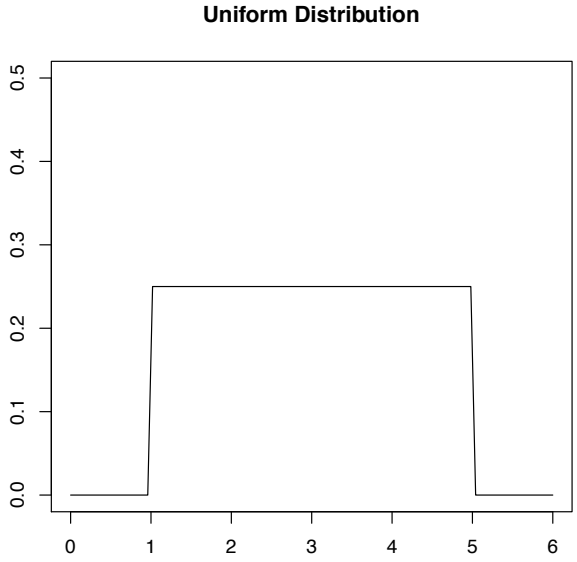


Figure 2.8: Density function of Uniform$[1, 5]$.

Hence if $X \sim$ Uniform$[a, b]$ then $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$.

Example 2.50. Waiting time for the delivery of a part from the warehouse to certain destination is said to have a uniform distribution from 1 to 5 days. What is the probability that the delivery time is two or more days?

Let $X \sim$ Uniform$[1, 5]$. Then, $f(x) = 0.25$ for $1 \leq x \leq 5$ and hence

$$P(X \geq 2) = \int_2^5 0.25 \, dt = 0.75.$$

In `R`, one could simply use

```
1-punif(2,1,5)
```

Normal

The normal distribution (Gaussian distribution) is by far the most important distribution in statistics. The normal distribution is identified by a location parameter μ and a scale parameter $\sigma^2(>0)$. A normal r.v. X is denoted as $X \sim N(\mu, \sigma^2)$ with p.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$

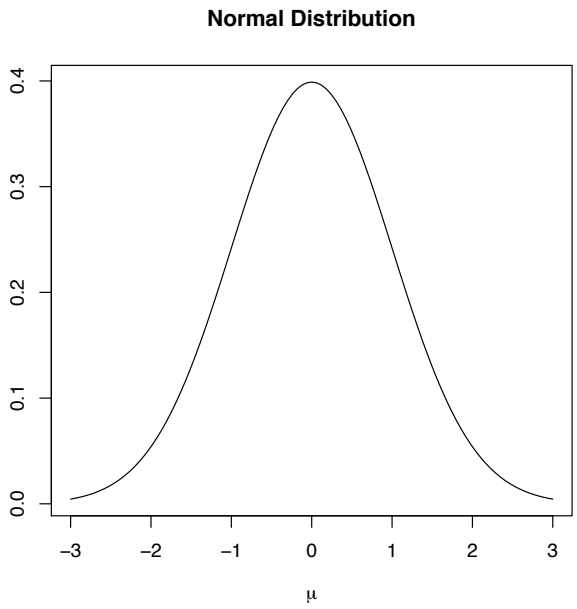


Figure 2.9: Density function of $N(0, 1)$.

It is symmetric, unimodal, bell shaped with $E(X) = \mu$ and $V(X) = \sigma^2$.

Notation: A normal random variable with mean 0 and variance 1 is called a *standard normal* r.v. It is usually denoted by $Z \sim N(0, 1)$. The c.d.f. of a standard normal is readily provided in software (and other resources) so we will always try to solve a problem using c.d.f. (and not p.d.f.).

Example 2.51. Find $P(-2.34 < Z < -1)$. From the relevant remark,

$$\begin{aligned} P(-2.34 < Z < -1) &= P(Z < -1) - P(Z \leq -2.34) \\ &= F_Z(-1) - F_Z(-2.34) \\ &= 0.1587 - 0.0096 \\ &= 0.1491 \end{aligned}$$

In `R`: `pnorm(-1)-pnorm(-2.34)`

If Z is standard normal then it has mean 0 and variance 1. Now if we take a linear transformation of Z, say $X = aZ + b$, for constants $a \neq 0$ and b, then

$$E(X) = E(aZ + b) = a\cancelto{0}{E(Z)} + b = b$$

and

$$V(X) = V(aZ + b) = a^2\cancelto{1}{V(Z)} = a^2.$$

This fact together with the following proposition allows us to express any normal r.v. as a linear transformation of the standard normal r.v. Z by setting $a = \sigma$ and $b = \mu$.

Proposition 2.7. The r.v. X that is expressed as the linear transformation $\sigma Z + \mu$, is a also a normal r.v. with $E(X) = \mu$ and $V(X) = \sigma^2$. □

Proof. In statistics it is known that two r.vs with the same c.d.f. implies that they have the same p.d.f. Our goal is to show that X has the p.d.f. as in equation (2.5.10)

$$\begin{aligned}
P(X \leq x) &= P(\sigma Z + \mu \leq x) \\
&= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-(1/2)z^2} dz \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-1/2\left(\frac{t-\mu}{\sigma}\right)^2} dt \qquad \text{by substitution } t = \sigma z + \mu
\end{aligned}$$

which by definition is the c.d.f. of a normal r.v. with mean μ and variance σ^2. Hence, $X \sim N(\mu, \sigma^2)$. □

Linear transformations are completely reversible, so given a normal r.v. X with mean μ and variance σ^2 we can revert back to a standard normal by

$$Z = \frac{X - \mu}{\sigma}.$$

As a consequence any probability statements made about an arbitrary normal r.v. can be reverted to statements about a standard normal r.v.

Example 2.52. Let $X \sim N(15, 7)$. Find $P(13.4 < X < 19.0)$.

We begin by noting

$$\begin{aligned} P(13.4 < X < 19.0) &= P\left(\frac{13.4-15}{\sqrt{7}} < \frac{X-15}{\sqrt{7}} < \frac{19.0-15}{\sqrt{7}}\right) \\ &= P(-0.6047 < Z < 1.5119) \\ &= F_Z(1.5119) - F_Z(-0.6047) \\ &= P(Z < 1.5119) - P(Z \leq -0.6047) \\ &= 0.6620312 \end{aligned}$$

If one is using a computer there is no need to revert back and forth from a standard normal, but it is always useful to standardize concepts. You could find the answer by using
`pnorm(1.5119)-pnorm(-0.6047)`, or
`pnorm(19,15,sqrt(7))-pnorm(13.4,15,sqrt(7))`
which standardizes for you.

Example 2.53. The height of males in inches is assumed to be normally distributed with mean of 69.1 and standard deviation 2.6. Let $X \sim N(69.1, 2.6^2)$. Find the 90^{th} percentile for the height of males.

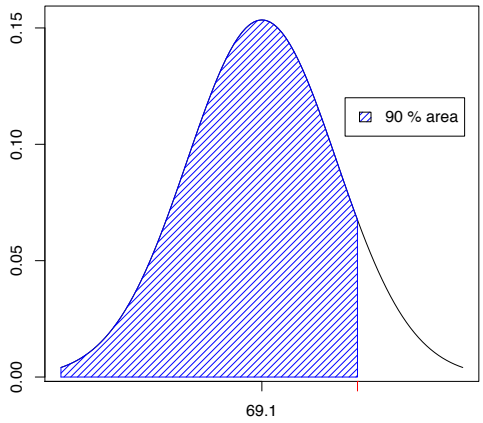


Figure 2.10: $N(69.1, 2.6^2)$ distribution

First we find the 90^{th} percentile of the standard normal which is `qnorm(0.9)`$=$ 1.281552. Then we transform to

$$2.6(1.281552) + 69.1 = 72.43204.$$

Or, just input into `R`: `qnorm(0.9,69.1,2.6)`.

A very useful theorem (whose proof is beyond the scope of this class is the following.

Proposition 2.8. A linear combination of (independent) normal random variables is a normal random variable. □

2.6 Central Limit Theorem

The *Central Limit Theorem* (C.L.T.) is a powerful statement concerning the mean of a random sample. There are three versions, the classical, the Lyapunov and the Linderberg but in effect they all make the same statement that the asymptotic distribution of the sample mean $\bar{X}$ is normal, irrespective of the distribution of the individual r.vs. $X_1, \ldots, X_n$.

Proposition 2.9. (Central Limit Theorem)
Let $X_1, \ldots, X_n$ be a random sample, i.e. i.i.d., with $E(X_i) = \mu < \infty$ and $V(X_i) = \sigma^2 < \infty$. Then, for $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{d} N(0,1)$$

as by example 2.43, $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$ □

Although the central limit theorem is an asymptotic statement, i.e. as the sample size goes to infinity, we can in practice implement it for sufficiently large sample sizes $n > 30$ as the distribution of $\bar{X}$ will be approximately normal.

$$\bar{X} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Remark 2.8. The following additional conditions/guidelines are needed to $n > 30$ for the following underlying distributions in order to implement C.L.T.

- Binomial: $np > 5$ and $n(1-p) > 5$
- Poisson: $\lambda > 10$

Example 2.54. At a university the mean age of students is 22.3 and the standard deviation is 4. A random sample of 64 students is to be drawn. What is the probability that the average age of the sample will be greater than 23?

By the C.L.T.

$$\bar{X} \overset{\text{approx.}}{\sim} N\left(22.3, \frac{4^2}{64}\right).$$

So we need to find

$$\begin{aligned} P(\bar{X} > 23) &= P\left(\frac{\bar{X} - 22.3}{4/\sqrt{(64)}} > \frac{23 - 22.3}{4/\sqrt{(64)}}\right) \\ &= P(Z > 1.4) \\ &= 1 - P(Z \leq 1.4) \\ &= 0.0808 \end{aligned}$$

or in `R: 1-pnorm(23,22.3,4/sqrt(64))`

Example 2.55. At a university assume it is known that 25% of students are over 21. In a sample of 400 what is the probability that more than 110 of them are over 21?
Exact solution: Let X be the number of students over 21. Then $X \sim \text{Bin}(400, 0.25)$.

$$\begin{aligned} P(X > 110) &= 1 - P(X \leq 110) \\ &= 1 - 0.8865 \\ &= 0.1135 \end{aligned}$$

in `R:1-pbinom(110,400,0.25)`
Approximation via C.L.T.: Let $\hat{p} = X/n$, then we have shown that $E(\hat{p}) = p = 0.25$ and $V(\hat{p}) = p(1-p)/n = 0.0004688$. Since, $\hat{p}$ is in fact an average we can use the C.L.T. to find $P(\hat{p} > 110/400)$. To reiterate

$$\hat{p} \overset{\text{approx.}}{\sim} N(0.25, 0.0004688)$$

Hence,

$$\begin{aligned} P(\hat{p} > 110/400) &= P(Z > 1.155) \\ &= 1 - P(Z \leq 1.155) \\ &= 0.1240 \end{aligned}$$

in `R: 1-pnorm(110/400,0.25,sqrt(0.0004688)`

Remark 2.9. We could have actually solved this using $X = n\hat{p}$, since if $\hat{p}$ is normal via the C.L.T. then $n\hat{p}$ is simply a linear transformation on a normal. In general same applies for $\sum_{i=1}^{n} X_i = n\bar{X}$.

So recall that if we can assume $\bar{X}$ is normal (or at least approxiamte normal) then so is any linear transformation of it...such as the sum.

2.7 Normal Probability Plot

The C.L.T. provides us with the tools to understand and make inference on the sampling distribution of the sample mean $\bar{X}$. When the sample size is small we cannot implement the C.L.T. However, by proposition 2.8, if the data are normally distributed then $\bar{X}$ is guaranteed to have a normal distribution. This is where normal probability plots are useful.

A probability plot is a graphical technique for comparing two data sets, either two sets of empirical observations, one empirical set against a theoretical set.

Definition 2.20. The empirical distribution function, or empirical c.d.f., is the cumulative distribution function associated with the empirical measure of the sample. This c.d.f. is a step function that jumps up by $1/n$ at each of the n data points.

$$\hat{F}_n(x) = \frac{\text{number of elements} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{x_i \leq x\}$$

Example 2.56. Consider the sample: 1, 5, 7, 8. The empirical c.d.f. is

$$\hat{F}_4(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.25 & \text{if } 1 \leq x < 5 \\ 0.50 & \text{if } 5 \leq x < 7 \\ 0.75 & \text{if } 7 \leq x < 8 \\ 1 & \text{if } x \geq 8 \end{cases}$$

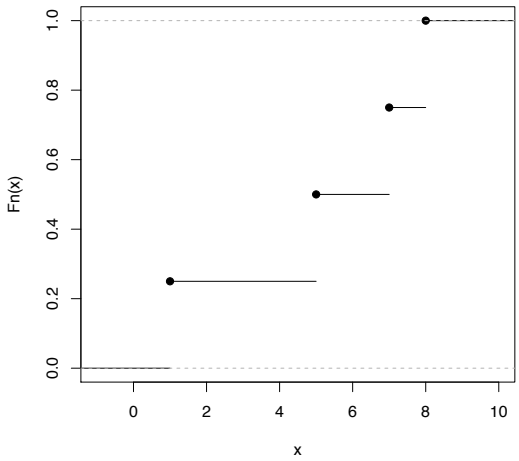


Figure 2.11: Empirical c.d.f.

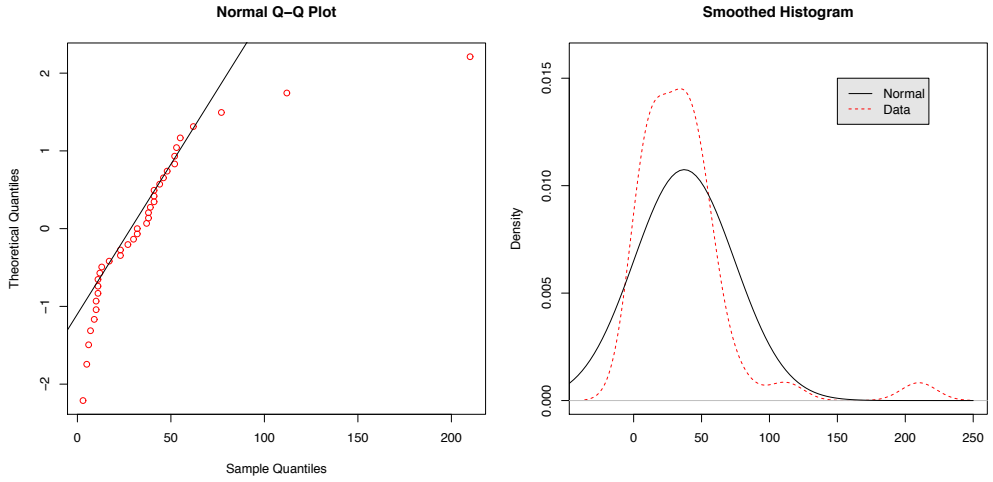
The normal probability plot is a graphical technique for normality testing by assessing whether or not a data set is approximately normally distributed.

The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

There are two types of plots commonly used to plot the empirical c.d.f. to the normal theoretical one ($G(\cdot)$).

- P-P plot that plots $(\hat{F}_n(x), G(x))$ (with scaled changed to look linear),
- Q-Q plot which plots the quantile functions $(\hat{F}_n^{-1}(x), G^{-1}(x))$.

Example 2.57. An experiment of lead concentrations (mg/kg dry weight) from 37 stations, yielded 37 observations. Of interest is to determine if the data are normally distributed (of more practical use if sample sizes are small, e.g. < 30).



Note that the data appears to be skewed right, with a lighter tail on the left and a heavier tail on the right (as compared to the normal).
http://www.stat.ufl.edu/~athienit/IntroStat/QQ.R

For interpretation of Q-Q plots, please watch the relevant podcasts. With the vertical axis being the theoretical quantiles, and the horizontal axis being the sample quantiles the interpretation of P-P plots and Q-Q plots is equivalent. Compared to straight line that corresponds to the distribution you wish to compare your data, here is a quick guideline of how the tails are

	Left tail	Right tail
Above line	Heavier	Lighter
Below line	Lighter	Heavier

Part II

Modules 3-4

Module 3

Inference for One Population

3.1 Inference for Population Mean

When a population parameter is estimated by a sample statistic such as $\hat{\mu} = \bar{x}$, the sample statistic is a point estimate of the parameter. Due to sampling variability the point estimate will vary from sample to sample. The fact that the sample estimate is not 100% accurate has to be taken into account. We have actually been able to model how the sample mean $\bar{X}$ should behave...exactly/approximately like a normal $\bar{X} \sim N(\mu, \sigma^2/n)$, by the following:

- $X_1, \ldots, X_n$ be i.i.d. from a normal distribution, so that Proposition 2.8 is invoked,
- $n > 30$ and the C.L.T. is invoked.

At the end of this module we will address the case where neither of these points can be used and we will also take a look at the sample variance too.

3.1.1 Confidence intervals

An alternative or complementary approach is to report an interval of plausible values based on the point estimate sample statistic and its standard deviation (a.k.a. standard error). A *confidence interval* (C.I.) is calculated by first selecting the *confidence level*, the degree of reliability of the interval. A $100(1-\alpha)\%$ C.I. means that the method by which the interval is calculated will contain the true population parameter $100(1-\alpha)\%$ of the time. That is, if a sample is replicated multiple times, the proportion of times that the C.I. will not contain the population parameter is α.

For example, assume that we know the (in practice unknown) population parameter μ is 0 and from multiple samples, multiple C.Is are created.

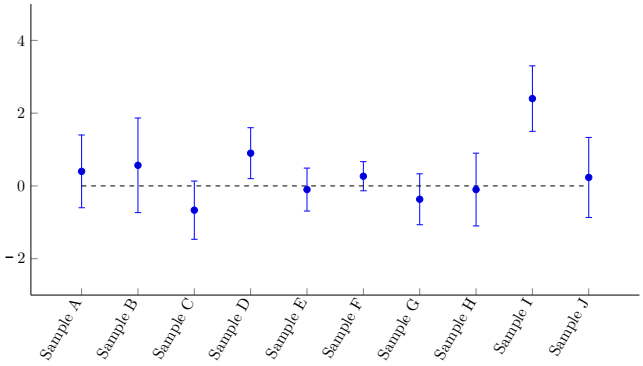
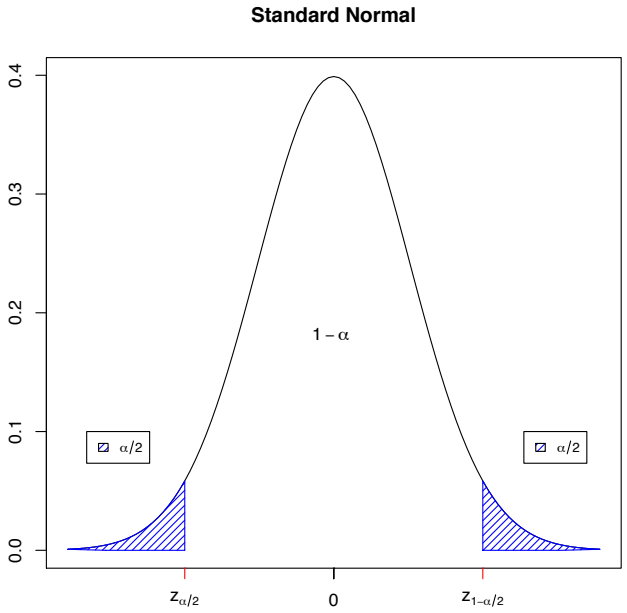


Figure 3.1: Multiple confidence intervals from different samples

Known population variance

Let $X_1, \ldots, X_n$ be i.i.d. from some distribution with finite unknown mean μ and known variance σ^2. The methodology will require that $\bar{X} \sim N(\mu, \sigma^2/n)$.

Let z_c stand for the value of $Z \sim N(0,1)$ such that $P(Z \leq z_c) = c$. Hence, the proportion of C.Is containing the population parameter is,



Due to the symmetry of the normal distribution, $z_{1-\alpha/2} = |z_{\alpha/2}|$ and $z_{\alpha/2} = -z_{1-\alpha/2}$.

Note: Some books may define z_c such that $P(Z > z_c) = c$, i.e. c referring to

the area to the right.

$$1-\alpha = P\left(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \tag{3.1}$$
$$= P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

and the probability that (on the long run) the random C.I. interval,

$$\bar{X} \mp z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

contains the true value of μ is $1-\alpha$. When a C.I. is constructed from a single sample we can no longer talk about a probability as there is no long run temporal concept but we can say that we are $100(1-\alpha)\%$ confident that the methodology by which the interval was contrived will contain the true population parameter.

Example 3.1. A forester wishes to estimate the average number of count trees per acre on a plantation. The variance is assumed to be known as 12.1.

A random sample of $n = 50$ one acre plots yields a sample mean of 27.3. A 95% C.I. for the true mean is then

$$27.3 \mp \underbrace{z_{1-0.025}}_{1.96}\sqrt{\frac{12.1}{50}} \to (26.33581, 28.26419)$$

Unknown population variance

In practice the population variance is unknown, that is **σ is unknown**. A large sample size implies that the sample variance s^2 is a good estimate for σ^2 and you will find that many simply replace it in the C.I. calculation. However, there is a technically "correct" procedure for when variance is unknown.

Note that s^2 is calculated from data, so just like $\bar{x}$, there is a corresponding random variable S^2 to denote the theoretical properties of the sample variance. In higher level statistics the distribution of S^2 is found, as once again, it is a statistic that depends on the random variables $X_1, \ldots, X_n$. It is shown that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \tag{3.2}$$

where t_{n-1} stands for Student's-t distribution with parameter degrees of freedom $\nu = n-1$. A Student's-t distribution is "similar" to the standard normal except that it places more "weight" to extreme values as seen in Figure 3.2.

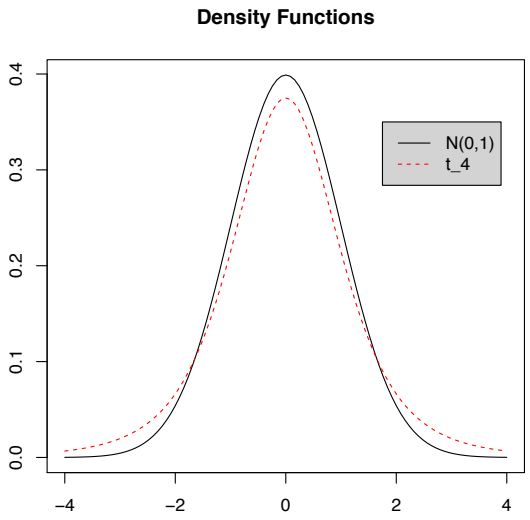


Figure 3.2: Standard normal and t_4 probability density functions

It is important to note that Student's-t is not just "similar" to the standard normal but asymptotically (as $n \to \infty$) is the standard normal. One just needs to view the t-table to see that under infinite degrees of freedom the values in the table are exactly the same as the ones found for the standard normal. Intuitively then, using Student's-t when σ^2 is unknown makes sense as it adds more probability to extreme values due to the uncertainty placed by estimating σ^2.

The $100(1-\alpha)\%$ C.I. for μ is then

$$\bar{x} \mp t_{(1-\alpha/2,n-1)} \frac{s}{\sqrt{n}}. \tag{3.3}$$

Example 3.2. In a packaging plant, the sample mean and standard deviation for the fill weight of 100 boxes are $\bar{x} = 12.05$ and $s = 0.1$. The 95% C.I. for the mean fill weight of the boxes is (using `qt(0.975,99)` for $t_{(1-0.025,99)}$)

$$12.05 \mp \underbrace{t_{(1-0.025,99)}}_{1.984} \frac{0.1}{\sqrt{100}} \to (12.03016, 12.06984), \tag{3.4}$$

Remark 3.1. If we wanted to perform a 90% we would simply replace $t_{(1-0.05/2,99)}$ with $t_{(1-0.10/2,99)} = 1.660$, which would lead to CI of $(12.0334, 12.0666)$ that is a narrower interval. Thus, as $\alpha \uparrow$ then $100(1-\alpha) \downarrow$ which implies a narrower interval.

Example 3.3. Suppose that a sample of 36 resistors is taken with $\bar{x} = 10$ and $s^2 = 0.7$. A 95% C.I. for μ is

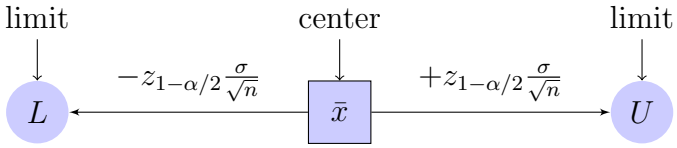
$$10 \mp \underbrace{t_{(1-0.025,35)}}_{2.03} \sqrt{\frac{0.7}{36}} \to (9.71693, 10.28307)$$

Remark 3.2. So far we have only discussed two-sided confidence intervals. In equation (3.1) However, **one-sided confidence intervals** might be more appropriate in certain circumstances. For example, when one is interested in the minimum breaking strength, or the maximum current in a circuit. In these instances we are not interested in an upper and lower limit respectively but only in a lower or only in an upper limit, respectively. Then we simply replace $z_{1-\alpha/2}$ or $t_{(1-\alpha/2,n-1)}$ by $z_{1-\alpha}$ or $t_{1-\alpha,n-1}$, e.g. a $100(1-\alpha)\%$ C.I. for μ

$$\left(\bar{x} - t_{(1-\alpha,n-1)}\frac{s}{\sqrt{n}}, \infty\right) \qquad \text{or} \qquad \left(-\infty, \bar{x} + t_{(1-\alpha,n-1)}\frac{s}{\sqrt{n}}\right)$$

Sample size for a C.I. of fixed level and width

The price paid for a higher confidence level, for the same sample statistics, is a wider interval - try this at home using different α values. We know that as the sample size n increases the standard deviation of $\bar{X}$, $\sigma/\sqrt{n}$ decreases and consequently so does the margin of error. Thus, knowing some preliminary information such as a rough estimate for σ can help us determine the sample size needed to obtain a fixed margin of error.



If we assume σ is known then the width of the interval is twice the margin of error

$$\text{width} = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Thus,

$$\sqrt{n} = 2z_{1-\alpha/2}\frac{\sigma}{\text{width}} \quad \Rightarrow \quad n \geq \left\lceil \left(2z_{1-\alpha/2}\frac{\sigma}{\text{width}}\right)^2 \right\rceil.$$

However, an estimate of σ is required, which can be from similar studies, pilot studies or some times just rough guesses such as the (range)/4.

Example 3.4. In Example 3.2 we had that $\bar{x} = 12.05$ and $s = 0.1$ for the 100 boxes, leading to a 95% C.I. for the true mean width 0.0392 or ± 0.0196 (see (3.4)). Boss man requires a narrower 95% C.I. of ± 0.0120.

So,

$$\left(\frac{2(1.96)}{2(0.0120)}\frac{0.1}{}\right)^2 = 266.7778$$

and we round up to $n \geq 267$. In practice we should try and round up to as far as our resources allow.

3.1.2 Hypothesis tests

A statistical hypothesis is a claim about a population characteristic (and on occasion more than one). An example of a hypothesis is the claim that the population is some value, e.g. $\mu = 75$.

Definition 3.1. The *null hypothesis*, denoted by H_0, is the hypothesis that is initially assumed to be true.

The *alternative hypothesis*, denoted by H_a or H_1, is the complementary assertion to H_0 and is usually the hypothesis, the new statement that we wish to test.

A test procedure is created under the assumption of H_0 and then it is determined how likely that assumption is compared to its complement H_a. The decision will be based on

- **Test statistic**, a function of the sampled data.
- **Rejection region/criteria**, the set of all test statistic values for which H_0 will be rejected.

The basis for choosing a particular rejection region lies in an understanding of the errors that can be made.

Definition 3.2. A *type I* error consists of rejecting H_0 when it is actually true.

A *type II* error consists of failing to reject H_0 when in actuality H_0 is false.

The type I error is generally considered to be the most serious one, and due to limitations, we can only control for one, so the rejection region is chosen based upon the maximum $P(\text{type I error}) = \alpha$ that a researcher is willing to accept.

Known population variance

We motivate the test procedure by an example whereby the drying time of a certain type of paint, under fixed environmental conditions, is known to be normally distributed with mean 75 min. and standard deviation 9 min. Chemists have added a new additive that is believed to decrease drying time and have obtained a sample of 35 drying times and wish to test their assertion. Hence,

$H_0 : \mu \geq 75$ (or $\mu = 75$ since it is sufficient to focus on the boundary point)
$H_a : \mu < 75$

Since we wish to control for the type I error, we set $P(\text{type I error}) = \alpha$. The default value of α is usually taken to be 5%.

An obvious candidate for a test statistic, that is an unbiased estimator of the population mean, is $\bar{X}$ which is normally distributed. If the data were not known to be normally distributed the normality of $\bar{X}$ can also be confirmed by the C.L.T. Thus, under the null assumption H_0

$$\bar{X} \overset{H_0}{\sim} N\left(75, \frac{9^2}{35}\right),$$

or equivalently

$$\frac{\bar{X} - 75}{\frac{9}{\sqrt{35}}} \overset{H_0}{\sim} N(0,1).$$

The test statistic will be

$$T.S. = \frac{\bar{x} - 75}{\frac{9}{\sqrt{35}}},$$

and assuming that $\bar{x} = 70.8$ from the 35 samples, then, $T.S. = -2.76$. This implies that 70.8 is 2.76 standard deviations below 75. Although this appears to be far, we need to use the *p-value* to reach a formal conclusion.

Definition 3.3. The *p-value* of a hypothesis test is the probability of observing the specific value of the test statistic, $T.S.$, or a more extreme value, under the null hypothesis. The direction of the extreme values is indicated by the alternative hypothesis.

Therefore, in this example values more extreme than -2.76 are

$$\{x | x \leq -2.76\},$$

as indicated by the alternative, $H_a : \mu < 75$. Thus,

$$\text{p-value} = P(Z \leq -2.76) = 0.0029.$$

The criterion for rejecting the null is **p-value** $< \boldsymbol{\alpha}$, the null hypothesis is rejected in favor of the alternative hypothesis as the probability of observing the test statistic value of -2.76 or more extreme (as indicated by H_a) is smaller than the probability of the type I error we are willing to undertake.

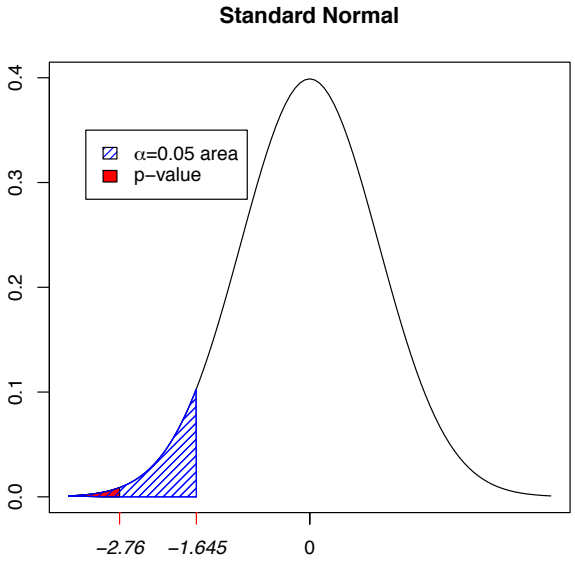


Figure 3.3: Rejection region and p-value.

If we can assume that $\bar{X}$ is normally distributed and $\boldsymbol{\sigma^2}$ **is known** then, to test

(i) $\mathrm{H}_0 : \mu \leq \mu_0$ vs $\mathrm{H}_a : \mu > \mu_0$

(ii) $\mathrm{H}_0 : \mu \geq \mu_0$ vs $\mathrm{H}_a : \mu < \mu_0$

(iii) $\mathrm{H}_0 : \mu = \mu_0$ vs $\mathrm{H}_a : \mu \neq \mu_0$

at the α significance level, compute the test statistic

$$T.S. = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}. \tag{3.5}$$

Reject the null if the p-value $< \alpha$, i.e.

(i) $P(Z \geq T.S.) < \alpha$ (area to the right of $T.S. < \alpha$)

(ii) $P(Z \leq T.S.) < \alpha$ (area to the left of $T.S. < \alpha$)

(iii) $P(|Z| \geq |T.S.|) < \alpha$ (area to the right of $|T.S.|$ plus area to the left of $-|T.S.| < \alpha$)

Example 3.5. A scale is to be calibrated by weighing a 1000g weight 60 times. From the sample we obtain $\bar{x} = 1000.6$. Assume $\sigma = 2$. Test whether the scale is calibrated correctly.

$\mathrm{H}_0 : \mu = 1000$ vs $\mathrm{H}_a : \mu \neq 1000$

$$T.S. = \frac{1000.6 - 1000}{2/\sqrt{60}} = 2.32379$$

Hence, the p-value is 0.02013675 and we reject the null hypothesis and conclude that the true mean is not 1000.

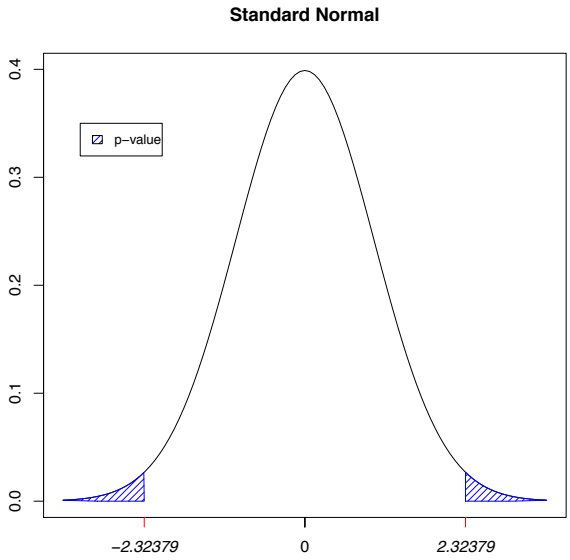


Figure 3.4: p-value.

Since 1000.6 is 2.32379 standard deviations greater than 1000, we can conclude that not only is the true mean not a 1000 but it is *greater than* 1000.

Example 3.6. A company representative claims that the number of calls arriving at their center is no more than 15/week. To investigate the claim, 36 random weeks were selected from the company's records with a sample mean of 17. Assume $\sigma = 3$. Do the sample data contradict this statement?

First we begin by stating the hypotheses of

$$\mathrm{H}_0 : \mu \leq 15 \qquad \text{vs} \qquad \mathrm{H}_a : \mu > 15$$

The test statistic is

$$T.S. = \frac{17 - 15}{3/\sqrt{36}} = 4$$

The conclusion is that there is significance evidence to reject H_0 as the p-value (the area to the right of 4 under the standard normal) is very close to 0.

Unknown population variance

If $\boldsymbol{\sigma}$ **is unknown**, which is usually the case, we replace it by its sample estimate s. Consequently,

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{\text{H}_0}{\sim} t_{n-1},$$

and the for an observed value $\bar{X} = \bar{x}$, the test statistic becomes

$$T.S. = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

At the α significance level, for the same hypothesis tests as before, we reject H_0 if

(i) p-value= $P(t_{n-1} \geq T.S.) < \alpha$

(ii) p-value= $P(t_{n-1} \leq T.S.) < \alpha$

(iii) p-value= $P(|t_{n-1}| \geq |T.S.|) < \alpha$

Example 3.7. In an ergonomic study, 5 subjects were chosen to study the maximin weight of lift (MAWL) for a frequency of 4 lifts/min. Assuming the MAWL values are normally distributed, do the following data suggest that the population mean of MAWL exceeds 25?

$$25.8,\ 36.6,\ 26.3,\ 21.8,\ 27.2$$

$\text{H}_0 : \mu \leq 25$ vs $\text{H}_a : \mu > 25$

$$T.S. = \frac{27.54 - 25}{5.47/\sqrt{5}} = 1.03832$$

The p-value is the area to the right of 1.03832 under the t_4 distribution, which is 0.1788813. Hence, we fail to reject the null hypothesis. In `R` input:

```
> t.test(c(25.8, 36.6, 26.3, 21.8, 27.2),mu=25,alternative="greater")

One Sample t-test

data:  c(25.8, 36.6, 26.3, 21.8, 27.2)
t = 1.0382, df = 4, p-value = 0.1789
alternative hypothesis: true mean is greater than 25
95 percent confidence interval:
 22.32433      Inf
sample estimates:
mean of x
    27.54
```

Remark 3.3. The values contained within a two-sided $100(1-\alpha)\%$ C.I. are precisely those values (that when used in the null hypothesis) will result in the p-value of a two sided hypothesis test to be greater than α.

For the one sided case, an interval that only uses the

- upper limit, contains precisely those values for which the p-value of a one-sided hypothesis test, with alternative *less than*, will be greater than α.
- lower limit, contains precisely those values for which the p-value of a one-sided hypothesis test, with alternative *greater than*, will be greater than α. (As in example 3.7)

Example 3.8. The lifetime of single cell organism is believed to be on average 257 hours. A small preliminary study was conducted to test whether the average lifetime was different when the organism was placed in a certain medium. The measurements are assumed to be normally distributed and turned out to be 253, 261, 258, 255, and 256. The hypothesis test is

$$\mathrm{H}_0 : \mu = 257 \text{ vs. } \mathrm{H}_a : \mu \neq 257$$

With $\bar{x} = 256.6$ and $s = 3.05$, the test statistic value is

$$T.S. = \frac{256.6 - 257}{3.05/\sqrt{5}} = -0.293.$$

The p-value is $P(t_4 < -0.293) + P(t_4 > 0.293) = 0.7839$. Hence, since the p-value is large (> 0.05) we fail to reject H_0 and conclude that population mean is not statistically different from 257.

Instead of a hypothesis test if a two sided 95% was constructed by

$$256.6 \mp \underbrace{t_{(1-0.025,4)}}_{2.776} \frac{3.05}{\sqrt{5}} \quad \rightarrow \quad (252.81, 260.39),$$

it clear that the null hypothesis value of $\mu = 257$ is a plausible value and consequently H_0 is plausible, so it is not rejected.

3.2 Inference for Population Proportion

3.2.1 Large sample confidence interval

In the binomial setting, experiments had binary outcomes and of interest was the number of successes out of the total number of trials. Let X be the total number of successes, then $X \sim \text{Bin}(n, p)$. Once an experiment is conducted and data obtained an estimate for p can be obtained,

$$\hat{p} = \frac{x}{n}$$

which is an average. It is the total number of successes over the total number of trials. As such, if the number of successes and number of failures are greater than 5, the C.L.T. tells us that

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

A $100(1-\alpha)\%$ C.I. can be created as before,

$$\hat{p} \mp z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This is the classical approach for when the sample size is large. There does exist an interval similar to classical version that works relatively well for small sample sizes (not too small) and is equivalent for large sample sizes. It is called the *Agresti-Coull* $100(1-\alpha)\%$ C.I.,

$$\tilde{p} \mp z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

where $\tilde{n} := n + 4$, and $\tilde{p} := (x+2)/\tilde{n}$.

Example 3.9. A map and GPS application for a smartphone was tested for accuracy. The experiment yielded 26 error out of the 74 trials. Find the 90% C.I. for the proportion of errors.

Since $n = 74$ and $x = 26$, then $\tilde{n} = 74 + 4$ and $\tilde{p} = (26+2)/78 = 0.359$. Hence the 90% C.I. for p is

$$0.359 \mp \underbrace{z_{1-0.05}}_{1.645}\sqrt{\frac{0.359(1-0.359)}{78}} \quad \to \quad (0.2696337, 0.4483151)$$

or in `R`:

```
> library(binom)
> binom.agresti.coull(26,74,conf.level=0.90)
         method  x  n      mean     lower     upper
1 agresti-coull 26 74 0.3513514 0.2666357 0.4465532
```

Note that the answers are slightly different because in R the function states: "this method does not use the concept of adding 2 successes and 2 failures,"but rather uses the formulas explicitly described in [the paper]". Hence we **recommend and encourage the use of software**.

3.2.2 Large sample hypothesis test

Let X be the number of successes in n Bernoulli trials with probability of success p, then $X \sim \text{Bin}(n, p)$. We know by the the C.L.T. that under certain regularity conditions, then

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

To test

(i) $\text{H}_0 : p \leq p_0$ vs $\text{H}_a : p > p_0$

(ii) $\text{H}_0 : p \geq p_0$ vs $\text{H}_a : p < p_0$

(iii) $\text{H}_0 : p = p_0$ vs $\text{H}_a : p \neq p_0$

The test statistic equivalent to the *Agresti-Coull* method is

$$T.S. = \frac{\tilde{p} - p_0}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}} \overset{\text{H}_0}{\sim} N(0,1)$$

Reject the null if

(i) p-value= $P(Z \geq T.S.) < \alpha$

(ii) p-value= $P(Z \leq T.S.) < \alpha$

(iii) p-value= $P(|Z| \geq |T.S.|) < \alpha$

Example 3.10. In example 3.9, if we wished to test whether the proportion of errors is less than half the time then, $\text{H}_a : p < 0.5$.

$$T.S. = \frac{28/78 - 0.5}{\sqrt{\frac{28/78(1-28/78)}{78}}} = -2.596426$$

with p-value $= 0.00470996 < \alpha = 0.10$, so reject the null. In a way, we kind of knew from the previous C.I. since the upper limit of the interval was 0.4483 which is less than than 0.5.

3.3 Inference for Population Variance

The sample statistic s^2 is widely used as the point estimate for the population variance σ^2, and similar to the sample mean it varies from sample to sample and has a sampling distribution.

Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s. We already have some tools that help us determine the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, a function of the r.v.'s, and hence $\bar{X}$ is a r.v. itself and once a sample is collected a realization $\bar{X} = \bar{x}$ is observed. Similarly, let

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

be a function of the r.v.'s $X_1, \ldots, X_n$ and hence is a r.v. itself. A realization of this r.v. is the sample variance s^2. If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ then

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

where χ^2 denotes a chi-square distribution with $(n-1)$ degrees of freedom.

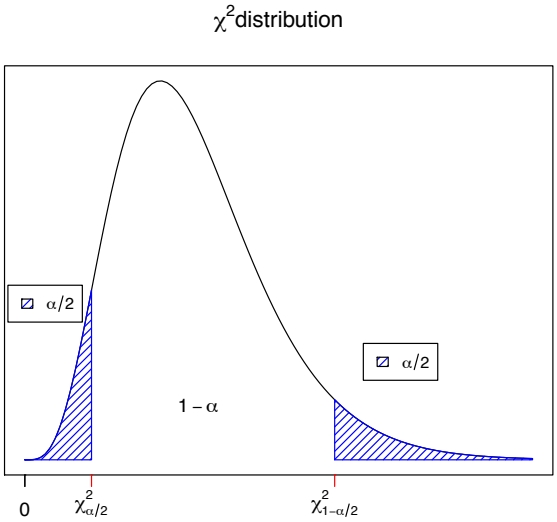


Figure 3.5: χ^2 distribution and critical value.

It is worth mentioning that a χ^2_{n-1} has mean $n-1$ and variance $2(n-1)$.

3.3.1 Confidence interval

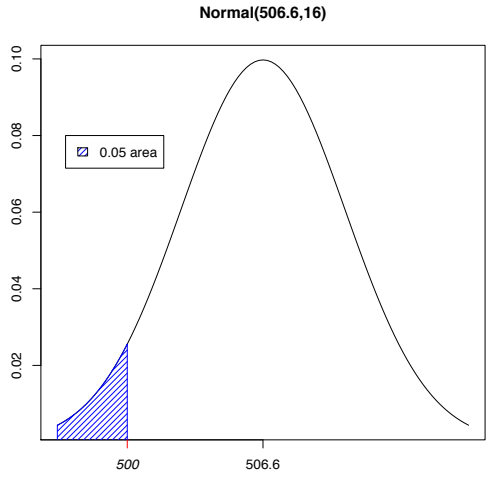
Consequently,

$$1-\alpha = P\left(\chi^2_{(\alpha/2,n-1)} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(1-\alpha/2,n-1)}\right)$$
$$= P\left(\frac{(n-1)S^2}{\chi^2_{(1-\alpha/2,n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(\alpha/2,n-1)}}\right)$$

which implies that on the long run this interval will contain the true population variance parameter $100(1-\alpha)\%$ of the time. Thus, the $100(1-\alpha)\%$ C.I. for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}, \frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}}\right).$$

Example 3.11. At a coffee plant a machine fills 500g coffee containers. Ideally, the amount of coffee in a container should vary only slightly about the 500g nominal value. The machine is designed to have a mean to dispense coffee amounts that have a normal distribution with mean 506.6g and standard deviation of 4g. This implies that only 5% of containers weigh less than 500g.



A quality control engineer samples 30 containers every hour. A particular sample yields

$$\bar{x} = 506.2046, \quad s = 3.9302.$$

We have already seen how to create a C.I. for μ, so we skip ahead to constructing a 95% C.I. for σ^2 (or σ). Assuming that the data are normally distributed, by creating a normal probability plot, we have

$$\frac{29(3.9302^2)}{45.722} < \sigma^2 < \frac{29(3.9302^2)}{16.047}$$
$$9.7971 < \sigma^2 < 27.9144$$

as a 95% C.I. for σ^2, or equivalently, by taking the square root, $(3.1300, 5.2834)$ as a 95% C.I. for σ.

In `R` we can find the critical values using the `qchisq` function as in `qchisq(0.025,29)`. If we have the raw data we use the `varTest` function to calculate the CI. See
http://www.stat.ufl.edu/~athienit/IntroStat/vartest.R

Remark 3.4. In this example, of more interest is the upper limit of the variance. A smaller variance for a $N(506.6, 16)$ means that the area to the left of 500 will be smaller and a larger variance that the area to the left of 500 will be larger. To construct a one-sided C.I. simply replace $\alpha/2$ in the formula by α.

3.3.2 Hypothesis test

To test

(i) $\text{H}_0 : \sigma^2 \leq \sigma_0^2$ vs. $\text{H}_a : \sigma^2 > \sigma_0^2$

(ii) $\text{H}_0 : \sigma^2 \geq \sigma_0^2$ vs. $\text{H}_a : \sigma^2 < \sigma_0^2$

(iii) $\text{H}_0 : \sigma^2 = \sigma_0^2$ vs. $\text{H}_a : \sigma^2 \neq \sigma_0^2$

$$T.S. = \frac{(n-1)S^2}{\sigma_0^2} \overset{\text{H}_0}{\sim} \chi^2_{n-1}$$

The null hypothesis is rejected, as always, when the p-value$< \alpha$, where p-value is calculated as

(i) $P(\chi^2_{n-1} \geq T.S.)$ (area to the right of test statistic)

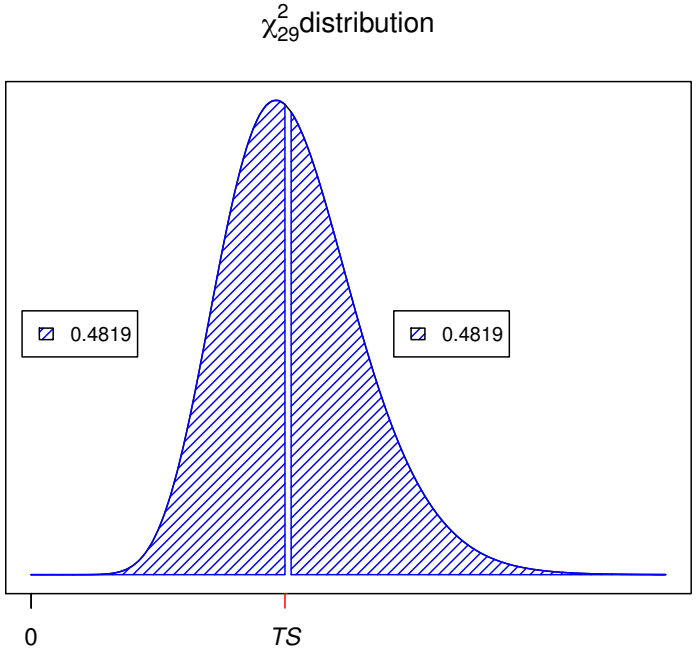
(ii) $P(\chi^2_{n-1} \leq T.S.)$ (area to the left of test statistic)

(iii) $\begin{cases} 2P(\chi^2_{n-1} \geq T.S.) & \text{if } P(\chi^2_{n-1} > T.S.) \leq P(\chi^2_{n-1} < T.S.) \\ 2P(\chi^2_{n-1} \leq T.S.) & \text{if } P(\chi^2_{n-1} > T.S.) > P(\chi^2_{n-1} < T.S.) \end{cases}$

where (iii) simply stated is twice the smallest of the two probabilities/areas.

Example 3.12. Let us find p-value for $\text{H}_0 : \sigma^2 = 4^2$ in the coffee example. This is a two-sided test and is comparable to the two sided C.I. we just did. Obviously $4^2 = 16$ is in the C.I. for the variance so we will fail to reject the null, i.e. p-value> 0.05 (as the C.I. was done at the 100(1-0.05)% level).

$$T.S. = \frac{29(3.9302^2)}{16} = 27.99649$$



To find the p-value we need to find the area in the two tails. First order of business is to determine is 27.99649 is in the right or left tail. Since the area to the left of 27.99649, `pchisq(27.99649,29)=0.4818993`, is smaller than the area to the right, it means our T.S. is in the left tail. As can be seen from the figure.

When we worked with the normal or Student-t distributions it always easy to find the other tail due to symmetry. For, example, if the T.S.= -2.5 then it means we are in the left tail and the right tail is +2.5. In any case, the two sided p-value was just twice the probability of the tail which is what we will do here as well. Multiply 2 to get the two tail equivalent p-value of 0.9637986. Note this is greater than 0.05. Or use software, see
http://www.stat.ufl.edu/~athienit/IntroStat/vartest.R
Where we get

```
> varTest(x,alternative="two.sided",sigma.squared=16,conf.level=0.95)

Null Hypothesis:                 variance = 16
Alternative Hypothesis:          True variance is not equal to 16
Test Name:                       Chi-Squared Test on Variance
Estimated Parameter(s):          variance = 15.44634
Data:                            x
Test Statistic:                  Chi-Squared = 27.99649
Test Statistic Parameter:        df = 29
P-value:                         0.9637987
95% Confidence Interval:         LCL =  9.797057
                                 UCL = 27.914367
```

3.4 Distribution Free Inference

When the sample size is small and we cannot assume that the data are normally distributed we need must use exact nonparametric procedures to perform inference on population cental values. Instead of means we will be referring to medians ($\tilde{\mu}$) and other location concepts as they are less influenced by outliers which can have a drastic impact (especially) on estimates from small samples.

3.4.1 Sign test

Recall that for p^{th} percentile, expect $(1-p)\%$ of the data to fall above that value. Let B be the number of observations that are **strictly greater** than the p^{th} percentile. (This will be the test statistic irrespective of the type of hypothesis test). By definition, we expect a $p\%$ of the points to below and $(1-p)\%$ to be above the specified null hypothesis percentile. Therefore, $B \sim \text{Bin}(n, 1-p)$. Let $\tilde{\mu}_p$ denote the population p^{th} percentile. To test the hypotheses

(i) $\text{H}_0 : \tilde{\mu}_p \leq \tilde{\mu}_{p0}$ vs $\text{H}_a : \tilde{\mu}_p > \tilde{\mu}_{p0}$

(ii) $\text{H}_0 : \tilde{\mu}_p \geq \tilde{\mu}_{p0}$ vs $\text{H}_a : \tilde{\mu}_p < \tilde{\mu}_{p0}$

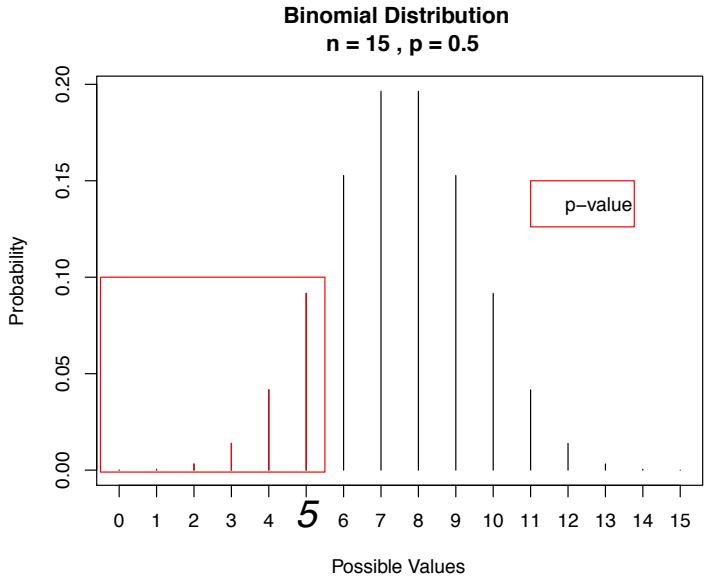
(iii) $\text{H}_0 : \tilde{\mu}_p = \tilde{\mu}_{p0}$ vs $\text{H}_a : \tilde{\mu}_p \neq \tilde{\mu}_{p0}$

we reject H_0 if the p-value $< \alpha$. We illustrate the calculation of the p-value with the following example.

Example 3.13. Pulse rates for a sample of 15 students were:

60, 62, 72, 60, 63, 75, 64, 68, 63, 60, 52, 64, 82, 68, 64

To test whether the median is less than 65 the hypotheses are: $\text{H}_0 : \tilde{\mu} \geq 65$ vs $\text{H}_a : \tilde{\mu} < 65$ we have $B = 5$. The p-value, (i.e. the probability of observing the test statistic or a value more extreme) is



$$\begin{aligned}\text{p-value} &= P(B \leq 5 | B \sim \text{Bin}(15, 0.5)) \\ &= P(B = 0) + \ldots + P(B = 5) \\ &= \sum_{i=0}^{5} \binom{15}{i} 0.5^i 0.5^{15-i} \\ &= 0.1509.\end{aligned}$$

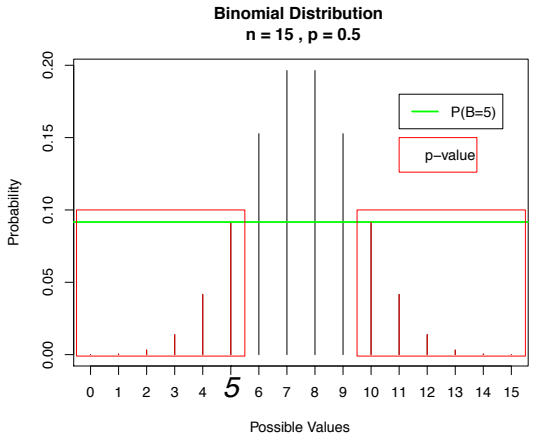
Hence, we fail to reject H_0. In R we would simply run `pbinom(5,15,0.5)` or even create a C.I. for said percentile

```
> library(EnvStats) #first install package
> x=c(60,62,72,60,63,75,64,68,63,60,52,64,82,68,64)
> eqnpar(x,p=0.5,type=6,ci=T,ci.method="exact",ci.type="upper",
+  approx.conf.level = 0.95)
Confidence Interval Type: upper
Confidence Level: 98.24219%
Confidence Interval: LCL = -Inf, UCL =   68
```

which yields the one sided 98.24219% C.I. $(-\infty, 68)$. This implies that since 65 is in the interval so we fail to reject the null.

Remark 3.5. Because we are working with a discrete distribution it is not always possible to find the α (one sided) or $\alpha/2$ and $1 - \alpha/2$ percentiles so we find the next ones that will give us a true α less than what was specified.

Assume we wished to perform a two-sided test, then the p-value would correspond to the sum of the probabilities in the two tails *as long as the probabilities in the other tail are less than or equal to the probability of the test statistic*, i.e. $\leq P(B = 5)$. If the left tail is 0 to 5 the the right tail is 10 to 15.

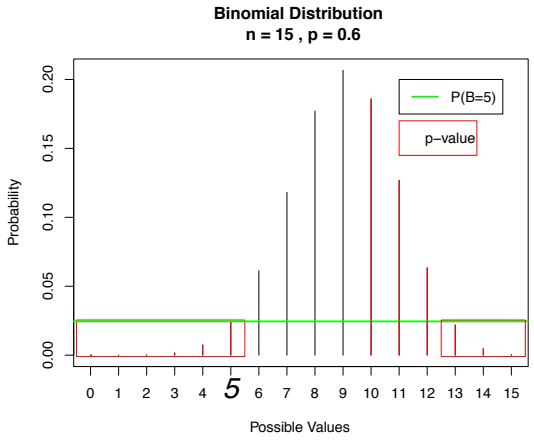


So it would be 2(0.1509) so we cannot reject the null. In R to get the two-sided C.I.

```
eqnpar(x,p=0.5,type=6,ci=T,ci.method="exact",ci.type="two-sided",
+  approx.conf.level = 0.95)
```

to get the 96.48438% C.I. (60,68) where 65 is in the interval.

Example 3.14. With the same data to test $H_a : \tilde{\mu}_{40} \neq 65$, then under the null $B \sim \text{Bin}(15, 0.6)$ and using a two-sided test



find that the p-value is $P(B \leq 5, B \geq 13) = 0.0609473$ or run

```
eqnpar(x,p=0.4,type=6,ci=T,ci.method="exact",ci.type="two-sided",
+  approx.conf.level = 0.95)
```

to get the 96.09947% C.I. for the 40th percentile (60,64). Note a slight discrepancy with the test, but that is because the α is different in the two cases.

3.4.2 Wilcoxon signed-rank test

This procedure does not require the distributional assumption of normality. However, it require the assumption of a continuous symmetric p.d.f. Assume that $X_1, \ldots, X_n$ are i.i.d. from some c.d.f. $F(x)$ that meets these assumptions. The null hypothesis, H_0 is that the distribution is centrally located around some value μ_0 which is tested against

(i) X's tend to be larger than μ_0.

(ii) X's tend to be smaller than μ_0.

(iii) X's tend to be different than μ_0.

To conduct the test

1. Calculate the differences $d_i = x_i - \mu_0$ for each observation, i.e. center data according to H_0.

2. Discard any $d_i = 0$ (as long as they are not more that 10% of the data, otherwise research "adjusted" methods).

3. Rank $|d_i|$ (i.e. rank ignoring the sign).

4. Calculate the test statistic S_+ = sum of the ranks corresponding to positive d_i's. Note that

$$S_+ + S_- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

   where n is the number of non-zero $d_i's$. Hence, it might be more convenient (in certain cases to calculate) S_-, and the figure out S_+. No issue when issuing software though.

The sampling distribution of the test statistic, denoted here as W, has been determined and is available in textbooks and software for calculation of p-values.

$$S_+ \overset{H_0}{\sim} W_n$$

The p-value calculation for the three alternatives will be done by software (as we have not discussed the W_n distribution) is

(i) $P(W_n \geq S_+)$

(ii) $P(W_n \leq S_+)$

(iii) Find the two tails and add probabilities in similar fashion as the sign test.

Example 3.15. Take for example the a dataset with values and we wish to test H_a: center is greater than 16.

x	$d = x - 16$	Rank
13.9	-2.1	1 (-)
11.0	-5.0	2 (-)
21.7	5.7	3 (+)
9.3	-6.7	4 (-)
5.0	-11.0	5 (-)
0.9	-15.1	6 (-)

The test statistic is $S_+ = 3$ and the p-value is 0.9531.

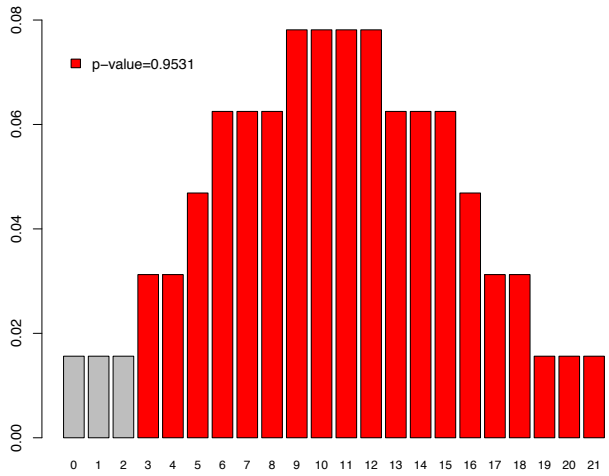


Figure 3.6: Distribution of Wilcoxon signed-rank test for $n = 6$

```
> x=c(13.9, 11.0,  5.0, 21.7,  9.3,  0.9)
> wilcox.test(x,mu=16,alternative="greater",conf.int=TRUE)

Wilcoxon signed rank test

data:  x
V = 3, p-value = 0.9531
alternative hypothesis: true location is greater than 16
95 percent confidence interval:
   5 Inf
sample estimates:
(pseudo)median
         10.15
```

Notice, the built in `R` function can also provide us with a C.I. for the (pseudo)median; and that 16 is in the interval.

If we wished to perform a two-sided test, then the p-value=0.1563

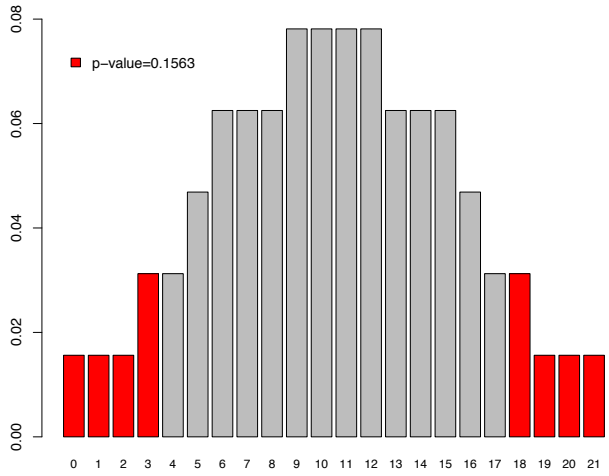


Figure 3.7: Distribution of Wilcoxon signed-rank test for $n = 6$

http://www.stat.ufl.edu/~athienit/IntroStat/wilcox_1sample.R

Remark 3.6. If there are **ties** in the data then the values that are tied get the average of the ranks that they would have gotten if not tied. For example, the rank of the data

Values	0.3	0.5	0.5	0.7
Ranks	1	2.5	2.5	4

The values 0.5 should have gotten ranks 2 and 3 if they were slightly different. Now consider a potential three-way tie.

Values	0.3	0.5	0.5	0.5	0.7
Ranks	1	3	3	3	5

Module 4

Inference for Two Populations

4.1 Inference for Population Means

4.1.1 Confidence intervals

There are instances when a C.I. for the difference between two means is of interest when one wishes to compare the sample mean from one population to the sample mean of another.

Known population variances

Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ represent two *independent* random samples with means μ_X, μ_Y and variances σ_X^2, σ_Y^2 respectively. Once again the methodology will require $\bar{X}$ and $\bar{Y}$ to be normally distributed. This can occur by:

- $X_1, \ldots, X_n$ be i.i.d. from a normal distribution, so that by Proposition 2.8, $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$
- $n_X > 40$ and the C.L.T. is invoked.

Similarly for $\bar{Y}$. Note that if the C.L.T. is to be invoked we require a more conservative criterion of $n_X > 40, n_Y > 40$ as we are using the theorem (and hence an approximation twice).

To compare two populations means μ_X and μ_Y we find it easier to work with a new parameter the difference $\mu_K := \mu_X - \mu_Y$. Let $K := \bar{X} - \bar{Y}$ is a normal random variable (by Proposition 2.8) with

$$E(K) = E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y = \mu_K,$$

and

$$V(K) = V(\bar{X} - \bar{Y}) \overset{\text{ind.}}{=} \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}.$$

Therefore,

$$K := \bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right),$$

and hence a $100(1-\alpha)\%$ C.I. for the difference of $\mu_K = \mu_X - \mu_Y$ is

$$\bar{x} - \bar{y} \mp z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

Example 4.1. In an experiment, 50 observations of soil NO_3 concentration (mg/L) were taken at each of two (independent) locations X and Y. We have that $\bar{x} = 88.5$, $\sigma_X = 49.4$, $\bar{y} = 110.6$ and $\sigma_Y = 51.5$. Construct a 95% C.I. for the difference in means and interpret.

$$88.5 - 110.6 \mp 1.96\sqrt{\frac{49.4^2}{50} + \frac{51.5^2}{50}} \to (-41.880683, -2.319317)$$

Note that 0 is not in the interval as a plausible value. This implies that $\mu_X - \mu_Y < 0$ is plausible. In fact μ_X is less than μ_Y by at least 2.32 units and at most 41.88.

Unknown population variances

As in equation (3.2)

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim t_\nu$$

where

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}}. \tag{4.1}$$

Hence the $100(1-\alpha)\%$ for $\mu_X - \mu_Y$ is

$$\bar{x} - \bar{y} \mp t_{1-\alpha/2,\nu}\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}.$$

Example 4.2. Two methods are considered standard practice for surface hardening. For Method A there were 15 specimens with a mean of 400.9 (N/mm^2) and standard deviation 10.6. For Method B there were also 15 specimens with a mean of 367.2 and standard deviation 6.1. Assuming the samples are independent and from a normal distribution the 98% C.I. for $\mu_A - \mu_B$ is

$$400.9 - 367.2 \mp t_{0.99,\nu}\sqrt{\frac{10.6^2}{15} + \frac{6.1^2}{15}}$$

where

$$\nu = \frac{\left(\frac{10.6^2}{15} + \frac{6.1^2}{15}\right)^2}{\frac{(10.6^2/15)^2}{14} + \frac{(6.1^2/15)^2}{14}} = 22.36$$

and hence $t_{0.99,22.36} = 2.5052$ giving a 98% C.I. for the difference $\mu_A - \mu_B$ of (25.7892 41.6108).

Notice that 0 is not in the interval so we can conclude that the two means are different. In fact the interval is purely positive so we can conclude that μ_A is at least 25.7892 N/mm^2 larger than μ_B and at most 41.6108 N/mm^2.

Remark 4.1. When population variances are believed to be equal, i.e. $\sigma_X^2 \equiv \sigma_Y^2$ we can improve on the estimate of variance by using a pooled or weighted average estimate. If in addition to the regular assumptions, if we can assume equality of variances then the $100(1-\alpha)\%$ C.I. for $\mu_X - \mu_Y$ is

$$\bar{x} - \bar{y} \mp t_{1-\alpha/2, n_X+n_Y-2}\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}},$$

with

$$s_p^2 = \left(\frac{(n_X-1)}{n_X+n_Y-2}\right)s_X^2 + \left(\frac{(n_Y-1)}{n_X+n_Y-2}\right)s_Y^2.$$

The assumption that the variances are equal must be made a priori and not used simply because the two variances may be close in magnitude.

Example 4.3. Consider Example 4.2 but now assume that $\sigma_X^2 \equiv \sigma_Y^2$. A 98% C.I. for the difference of $\mu_X - \mu_Y$ constructed with

$$s_p^2 = \frac{14(10.6^2) + 14(6.1^2)}{28} = 8.648^2$$

is

$$400.9 - 367.2 \mp \underbrace{t_{0.99,28}}_{2.467}(8.648)\sqrt{\frac{2}{15}} \quad \rightarrow \quad (25.9097, 41.4903)$$

How is this interval different from the one in Example 4.2?

Large sample C.I. for two population proportions

A simple extension of Section 3.2.1 to the two sample framework yields the $100(1-\alpha)\%$ C.I. for the difference of two population proportions. Let $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$ be two independent binomial r.vs. Define $\tilde{n}_X = n_X + 2$ and $\tilde{p}_X = (x+1)/\tilde{n}_X$, similarly for Y. Then the $100(1-\alpha)\%$ C.I. for $p_X - p_Y$ is

$$\tilde{p}_X - \tilde{p}_Y \mp z_{1-\alpha/2}\sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}}.$$

Intuitively, since proportions are between 0 and 1, the difference of two proportions must lie between -1 and 1. Hence if the bounds of a C.I. are outside the intuitive ones, they should be replaced by the intuitive bounds.

Example 4.4. In a clinical trial for a pain medication, 394 subjects were blindly administered the drug, while an independent group of 380 were given a placebo. From the drug group, 360 showed an improvement. From the placebo group 304 showed improvement. Construct a 95% C.I. for the difference and interpret.

Let D stand for drug and P for placebo, then $\tilde{p}_D = 361/396$ and $\tilde{p}_P = 305/382$

$$\tilde{p}_D - \tilde{p}_P \mp 1.96\sqrt{\frac{\tilde{p}_D(1-\tilde{p}_D)}{396} + \frac{\tilde{p}_P(1-\tilde{p}_P)}{382}} \to (0.0642, 0.1622)$$

Hence the proportion of subjects that showed substantial improvement under the drug treatment was at least 6.42% and at most 16.22% greater than under the placebo.

Paired data

There are instances when two samples are not independent, when a relationship exists between the two. For example, before treatment and after treatment measurements made on the same experimental subject are dependent on eachother through the experimental subject. This is a common event in clinical studies where the effectiveness of a treatment, that may be quantified by the difference in the before and after measurements, is dependent upon the individual undergoing the treatment. Then, the data is said to be *paired*.

Consider the data in the form of the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. We note that the pairs, i.e. two dimensional vectors, are independent as the experimental subjects are assumed to be independent with marginal expectations $E(X_i) = \mu_X$ and $E(Y_i) = \mu_Y$ for all $i = 1, \ldots, n$. By defining,

$$\begin{aligned} D_1 &= X_1 - Y_1 \\ D_2 &= X_2 - Y_2 \\ &\vdots \\ D_n &= X_n - Y_n \end{aligned}$$

a two sample problem has been reduced to a one sample problem. Inference for $\mu_X - \mu_Y$ is equivalent to one sample inference on μ_D as was done in Module 3. This holds since,

$$\mu_D := E(\bar{D}) = E\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - Y_i\right) = E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y.$$

In addition we note that the variance of $\bar{D}$ does incorporate the covariance between the two samples and does not have to be calculated separately as

$$\sigma_D^2 := V(\bar{D}) = V\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(D_i) = \frac{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}{n}.$$

Example 4.5. A *new* and *old* type of rubber compound can be used in tires. A researcher is interested in a compound/type that does not wear easily. Ten random cars were chosen at random that would go around a track a predetermined number of times. Each car did this twice, once for each tire type and the depth of the tread was then measured.

	Car									
	1	2	3	4	5	6	7	8	9	10
New	4.35	5.00	4.21	5.03	5.71	4.61	4.70	6.03	3.80	4.70
Old	4.19	4.62	4.04	4.72	5.52	4.26	4.27	6.24	3.46	4.50
d	0.16	0.38	0.17	0.31	0.19	0.35	0.43	-0.21	0.34	0.20

With $\bar{d} = 0.232$ and $s_D = 0.183$. Assuming that the data are normally distributed, a 95% C.I. for $\mu_{\text{new}} - \mu_{\text{old}} = \mu_D$ is

$$0.232 \mp \underbrace{t_{0.975,9}}_{2.262} \frac{0.183}{\sqrt{10}} \quad \rightarrow \quad (0.101, 0.363)$$

and we note that the interval is strictly greater than 0, implying that that the difference is positive, i.e. that $\mu_{\text{new}} > \mu_{\text{old}}$. In fact we can conclude that μ_{new} is larger than μ_{old} by at least 0.101 units and at most 0.363 units.

Remark 4.2. It is important to note, that any one sample/population procedure can be applied to paired data. By taking the differences, the problem becomes a one sample/population problem and any conclusions made are made on the differences. For example we could even do a sign test on the differences or as we will see later the Wilcoxon signed-rank test but on differences.

4.1.2 Hypothesis tests

Known variance

Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ represent two *independent* random large samples with $n_X > 40, n_Y > 40$ with means μ_X, μ_Y and variances σ_X^2, σ_Y^2 respectively. We have seen in Section 4.1.1 that if can assume that $\bar{X}$ and $\bar{Y}$ are normally distributed, that

$$\bar{X} - \bar{Y} \sim N\left(\underbrace{\mu_X - \mu_Y}_{\overset{\mathrm{H}_0}{=}\Delta_0}, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} \right).$$

To test

(i) $\mathrm{H}_0 : \mu_X - \mu_Y \leq \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y > \Delta_0$

(ii) $\mathrm{H}_0 : \mu_X - \mu_Y \geq \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y < \Delta_0$

(iii) $\mathrm{H}_0 : \mu_X - \mu_Y = \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y \neq \Delta_0$

we assume that the variances are known and the test statistic is

$$T.S. = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}}.$$

The r.v. corresponding to the test statistic has a standard normal distribution under the null hypothesis H_0, that $\mu_X - \mu_Y = \Delta_0$. Reject the null if

(i) p-value=$P(Z \geq T.S.) < \alpha$

(ii) p-value=$P(Z \leq T.S.) < \alpha$

(iii) p-value=$P(|Z| \geq |T.S.|) < \alpha$

Unknown variance test for difference of two means

Usually the **variances are unknown** and have to be estimated, then the test statistic is

$$T.S. = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}},$$

which has a t_ν distribution under H_0, where the degrees of freedom ν are given by equation (4.1).

Remark 4.3. As in Remark 4.1, when population variances are believed to be equal, i.e. $\sigma_X^2 \equiv \sigma_Y^2$ we can improve on the estimate of variance, and hence obtain a more powerful test, by using a pooled estimate of the variance. If in

addition to the regular assumptions, if we can assume equality of variances then set the sample variances of X and Y to

$$s_p^2 := \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2},$$

and the degrees of freedom for the t distribution by $n_X + n_Y - 2$.

Large sample test for two population proportions

Let $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$ represent two independent binomial r.vs from two Bernoulli trial experiments. To test

(i) $\text{H}_0 : p_X - p_Y \leq \Delta_0$ vs $\text{H}_a : p_X - p_Y > \Delta_0$

(ii) $\text{H}_0 : p_X - p_Y \geq \Delta_0$ vs $\text{H}_a : p_X - p_Y < \Delta_0$

(iii) $\text{H}_0 : p_X - p_Y = \Delta_0$ vs $\text{H}_a : p_X - p_Y \neq \Delta_0$

where $\Delta_0 \in [-1, 1]$, we must assume that the number of successes and failures is greater than 10 for both samples. As the null hypotheses values for p_X and p_Y are not available we simply check that the sample successes and failures are greater than 10. By virtue of the C.L.T.

$$\hat{p}_X - \hat{p}_Y \overset{\text{H}_0}{\sim} N\left(\underbrace{p_X - p_Y}_{\Delta_0}, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right),$$

and test statistic (via Agresti-Coull method) would be constructed in the usual way.

$$T.S. = \frac{\tilde{p}_X - \tilde{p}_Y - \Delta_0}{\sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{\tilde{n}_Y}}} \overset{\text{H}_0}{\sim} N(0,1)$$

Remark 4.4. When $\Delta_0 = 0$ it is assumed that $p_X = p_Y$ which implies that *the two variances are equal* and therefore in lieu of Remark 4.1 we can replace $\hat{p}_X$ and $\hat{p}_Y$ in the variance by the pooled estimate

$$\tilde{p} = \frac{(x+1) + (y+1)}{\tilde{n}_X + \tilde{n}_Y}.$$

The test statistic is then

$$T.S. = \frac{\tilde{p}_X - \tilde{p}_Y - 0}{\sqrt{\tilde{p}(1-\tilde{p})(1/\tilde{n}_X + 1/\tilde{n}_Y)}} \overset{\text{H}_0}{\sim} N(0,1),$$

and the r.v. corresponding to the test statistic has a standard normal distribution under the null hypothesis.

Paired data

In the event that two samples are dependent, i.e. paired, such as when two different measurements are made on the same experimental unit, the inference methodology must be adapted to account for the dependence/covariance between the two samples.

Refer to Section 4.1.1, where we consider the data in the form of the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ and construct the one-dimensional, i.e. one-sample $D_1, D_2, \ldots, D_n$ where $D_i = X_i - Y_i$ for all $i = 1, \ldots, n$. As shown earlier, $\mu_D = \mu_X - \mu_Y$ and the variance term σ_D^2 incorporates the covariance between X and Y.

To test

(i) $\mathrm{H}_0 : \mu_X - \mu_Y = \mu_D \leq \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y = \mu_D > \Delta_0$

(ii) $\mathrm{H}_0 : \mu_X - \mu_Y = \mu_D \geq \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y = \mu_D < \Delta_0$

(iii) $\mathrm{H}_0 : \mu_X - \mu_Y = \mu_D = \Delta_0$ vs $\mathrm{H}_a : \mu_X - \mu_Y = \mu_D \neq \Delta_0$

perform a one-sample hypothesis test by using the test statistic

$$T.S. = \begin{cases} \dfrac{\bar{d} - \Delta_0}{\sigma_D/\sqrt{n}} \overset{\mathrm{H}_0}{\sim} N(0,1) \\ \dfrac{\bar{d} - \Delta_0}{s_D/\sqrt{n}} \overset{\mathrm{H}_0}{\sim} t_{n-1} \end{cases}$$

4.2 Inference for Population Variances

Now we extend the set up to two independent i.i.d. normal distribution samples $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ with variances σ_X^2 and σ_Y^2 respectively. It is known that

$$\frac{(n_X - 1)S_X^2}{\sigma_X^2} \sim \chi^2_{n_X - 1} \quad \text{and} \quad \frac{(n_Y - 1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{n_Y - 1}$$

but it is also known that a standardized (by dividing by the degrees of freedom) ratio of two χ^2's is an F-distribution. Therefore,

$$\frac{\frac{(n_X-1)S_X^2/\sigma_X^2}{n_X-1}}{\frac{(n_Y-1)S_Y^2/\sigma_Y^2}{n_Y-1}} = \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{n_X-1, n_Y-1}.$$

When comparing two variances, in order to use the F-distribution it is practical to make comparisons in terms or ratios rather than differences. For example,

- $\sigma_X^2/\sigma_Y^2 = 1 \Rightarrow \sigma_X^2 = \sigma_Y^2$
- $\sigma_X^2/\sigma_Y^2 > 1 \Rightarrow \sigma_X^2 > \sigma_Y^2$
- $\sigma_X^2/\sigma_Y^2 < 1 \Rightarrow \sigma_X^2 < \sigma_Y^2$

4.2.1 Confidence intervals

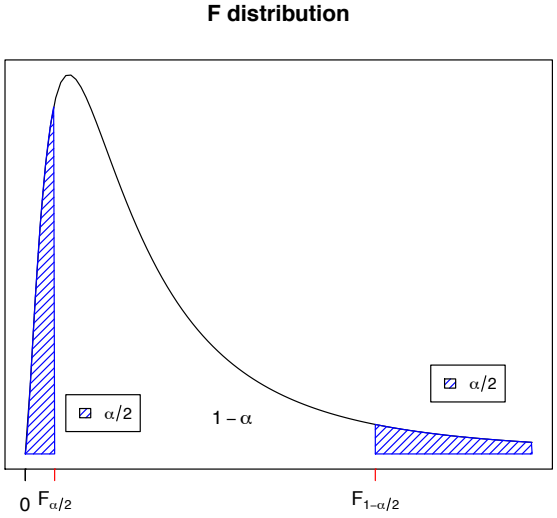


Figure 4.1: F_{ν_1,ν_2} distribution.

A $100(1-\alpha)\%$ C.I for σ_X^2/σ_Y^2 is constructed by

$$\begin{aligned} 1-\alpha &= P\left(F_{(\alpha/2;n_X-1,n_Y-1)} < \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} < F_{(1-\alpha/2;n_X-1,n_Y-1)}\right) \\ &= P\left(\frac{S_X^2}{S_Y^2}\frac{1}{F_{(1-\alpha/2;n_X-1,n_Y-1)}} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{S_X^2}{S_Y^2}\frac{1}{F_{(\alpha/2;n_X-1,n_Y-1)}}\right). \end{aligned}$$

Thus, the $100(1-\alpha)\%$ C.I. for σ_X^2/σ_Y^2 is

$$\left(\frac{s_X^2}{s_Y^2}\frac{1}{F_{(1-\alpha/2;n_X-1,n_Y-1)}}, \frac{s_X^2}{s_Y^2}\frac{1}{F_{(\alpha/2;n_X-1,n_Y-1)}}\right)$$

Example 4.6. The life length of an electrical component was studied under two operating voltages, 110 and 220. Ten different components were assigned to be tested under 110V and 16 under 220V. The times to failure (in 100's hrs) were then recorded. Assuming that the two samples are independent and normal we construct a 95% C.I. for $\sigma_{110}^2/\sigma_{220}^2$.

V	n	Mean	St.Dev.
110	10	20.1932	0.5688
220	16	9.9222	0.2408

Hence,

$$\left(\frac{0.5688^2}{0.2408^2}\underbrace{\frac{1}{F_{0.90;9,15}}}_{2.086209}, \frac{0.5688^2}{0.2408^2}\underbrace{\frac{1}{F_{0.10;9,15}}}_{0.4274191}\right) \to (1.786875, 21.032615)$$

We note that the value 1 is not in the interval. Therefore, we conclude that the variance are not equal and that in fact the variance of 110V is at least 78.69% larger than the variance of 220V and at most 2003.62%. In terms of the ratio of standard deviations the 95% C.I. for σ_X/σ_Y is

$$(\sqrt{1.786875}, \sqrt{21.032615}) \to (1.336740, 4.586133)$$

Critical values can be obtained using the quantile function `qf`, for example

`qf(0.95,9,15)`. In R once we create two vectors for the two datasets

```
> var.test(V110,V220,ratio=1,alternate="two.sided",conf.level=0.95)

F test to compare two variances

data:  V110 and V220
F = 5.5799, num df = 9, denom df = 15, p-value = 0.003624
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.786875 21.032615
sample estimates:
ratio of variances
          5.579894
```

http://www.stat.ufl.edu/~athienit/IntroStat/var_ratio.R

If creating one-sided C.I. simply replace $\alpha/2$ by α. If creating an upper, the lower limit will be 0, and if creating a lower, the upper limit is ∞.

4.2.2 Hypothesis tests

To test

(i) $\text{H}_0 : \sigma_X^2/\sigma_Y^2 \leq \Delta_0$ vs $\text{H}_a : \sigma_X^2/\sigma_Y^2 > \Delta_0$

(ii) $\text{H}_0 : \sigma_X^2/\sigma_Y^2 \geq \Delta_0$ vs $\text{H}_a : \sigma_X^2/\sigma_Y^2 < \Delta_0$

(iii) $\text{H}_0 : \sigma_X^2/\sigma_Y^2 = \Delta_0$ vs $\text{H}_a : \sigma_X^2/\sigma_Y^2 \neq \Delta_0$

$$T.S. = \frac{\frac{s_X^2}{s_Y^2}}{\Delta_0} \overset{\text{H}_0}{\sim} F_{n_X-1,n_Y-1}.$$

The p-value is calculated as

(i) $P(F_{n_X-1,n_Y-1} \geq T.S.)$

(ii) $P(F_{n_X-1,n_Y-1} \leq T.S.)$

(iii) $\begin{cases} 2P(F_{n_X-1,n_Y-1} \geq T.S.) & \text{if } P(F_{n_X-1,n_Y-1} > T.S.) \leq P(F_{n_X-1,n_Y-1} < T.S.) \\ 2P(F_{n_X-1,n_Y-1} \leq T.S.) & \text{if } P(F_{n_X-1,n_Y-1} > T.S.) > P(F_{n_X-1,n_Y-1} < T.S.) \end{cases}$

Rejection region criteria can be found in the relevant textbook section but are omitted here.

4.3 Distribution Free Inference

4.3.1 Wilcoxon rank-sum test

One of the most widely used two sample tests for location differences between two populations. Assume, that two independent samples $X_1, \ldots, X_{n_X}$ are i.i.d. with a c.d.f. $F_1(\cdot)$ and $Y_1, \ldots, Y_{n_Y}$ are i.i.d. with a c.d.f. $F_2(\cdot)$. The null hypothesis $\mathrm{H}_0 : F_1(x) = F_2(x)\ \forall x$ is tested against

(i) X's tend to be larger than the Y's by Δ_0 units, i.e. X's$> Y$'s$+\Delta_0$.

(ii) X's tend to be smaller than the Y's by Δ_0 units, i.e. X's$< Y$'s$+\Delta_0$.

(iii) One of the two populations is shifted from the other by Δ_0 units.

To conduct the test we

1. rank all the $N(= n_X + n_Y)$ data irrespective of sample
2. calculate the sum of the ranks associated with the first sample, (assuming X is first)
$$T.S. = S_X \overset{\mathrm{H}_0}{\sim} W_N$$

Under the null the test statistic has a wilcoxon sampling distribution. It really does not matter which sum rank we calculate since

$$S_X + S_Y = \sum_{i=1}^{N} i = \frac{N(N+1)}{2}$$

To find the p-value we will use software (rather than working with limited tables), where we can even obtain confidence intervals for the difference in the location of the center from the first population to the second..

Example 4.7. Two groups of 10 did not know whether they were receiving alcohol or the placebo and their reaction times (in seconds) was recorded.

(x) Placebo	0.90	0.37	1.63	0.83	0.95	0.78	0.86	0.61	0.38	1.97
(y) Alcohol	1.46	1.45	1.76	1.44	1.11	3.07	0.98	1.27	2.56	1.32

Test whether the distribution of reaction times for the placebo are shifted to the "left" of that for alcohol (case (ii)). The ranks are:

Placebo	7	1	16	5	8	4	6	3	2	18	70
Alcohol	15	14	17	13	10	20	9	11	19	12	140

The test statistic is $S_X = 70$.

```
> wilcox.test(placebo,alcohol,alternative="less",mu=0,conf.int=TRUE)
Wilcoxon rank sum test
data:  placebo and alcohol
W = 15, p-value = 0.003421
alternative hypothesis: true location shift is less than 0
95 percent confidence interval:
  -Inf -0.37
sample estimates:
difference in location
                -0.61
```

In R, the test statistic is calculated slightly differently but what we really care is the p-value. Since the p-value is tiny we reject the null and conclude that the center of placebo is generally smaller than the center of alcohol. We can also see this by the strictly negative 95% C.I.
http://www.stat.ufl.edu/~athienit/IntroStat/wilcox_1.R

4.3.2 Wilcoxon signed-rank test

To test for location differences between the X and Y components in the i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, we take the differences $D_i = X_i - Y_i$ (as in Section 4.1.1) and proceed just like the one sample Wilcoxon signed-rank test

H_0 : Distribution of D's is symmetric about the null value Δ_0, against the alternatives

(i) D's tend to be larger than Δ_0, i.e. X's$>$ Y's$+\Delta_0$.

(ii) D's tend to be smaller than Δ_0, i.e. X's$<$ Y's$+\Delta_0$.

(iii) D's tend to be consistently larger or smaller than Δ_0, i.e. X's tend to be consistently different than Y's by an amount of Δ_0 or greater.

Example 4.8. A city park department compared two fertilizers A and B on 20 softball fields. Each field was divided in half where each fertilizer was used. The effect of the fertilizer was measured in the pounds (lbs) of grass clippings produced.

Since not specified in the problem, we consider as an alternative hypothesis the (general) two-sided alternative (case (iii)) with $\Delta_0 = 0$.

Field	A	B	d	Rank($\lvert d\rvert$)	Field	A	B	d	Rank($\lvert d\rvert$)
1	211.4	186.3	25.1	15	11	208.9	183.6	25.3	17.5
2	204.4	205.7	-1.3	1	12	208.7	188.7	20.0	8
3	202.0	184.4	17.6	7	13	213.8	188.6	25.2	16
4	201.9	203.6	-1.7	2	14	201.6	204.2	-2.6	4
5	202.4	180.4	22.0	14	15	201.8	181.6	20.1	9
6	202.0	202.0	0	0	16	200.3	208.7	-8.4	6
7	202.4	181.5	20.9	13	17	201.8	181.5	20.3	10
8	207.1	186.7	20.4	11	18	201.5	208.7	-7.2	5
9	203.6	205.7	-2.1	3	19	212.1	186.8	25.3	17.5
10	216.0	189.1	26.9	19	20	203.4	182.9	20.5	12

$S_+ = 15 + 7 + 14 + 13 + 11 + 19 + 17.5 + 8 + 16 + 9 + 10 + 17.5 + 12 = 169$
$S_- = 1 + 2 + 3 + 4 + 6 + 5 = 21$

The test statistic is $S_+ = 169$ (although it would have been easier to calculate S_- and then deduce S_+). Since we are in the two sided case (case (iii)), to find the p-value we need to find $P(W_{19} \leq 21, W_{19} \geq 169)$ or use `R` (which will give a slightly different answer as it does not ignore the zero, run code and see warnings) as done in
http://www.stat.ufl.edu/~athienit/IntroStat/wilcox_2.R

```
> wilcox.test(A,B,paired=TRUE,conf.int=TRUE)
        Wilcoxon signed rank test with continuity correction
data:  A and B
V = 169, p-value = 0.003098
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  8.700026 22.650044
sample estimates:
(pseudo)median
      12.12435
```

where the p-value is tiny and we reject the null. Also, we note that the 95% C.I. is strictly positive implying that A tends to be 8.7 to 22.65 units larger than B.

Example 4.9. Continuing from the previous example, suppose that type B was the old fertilizer and that a sales agent approached the city council with a claim that their new fertilizer (type A) was better in that it would produce 5 or more pounds of grass clippings compared to B.

The alternative hypothesis is case (i) with $\Delta_0 = 5$. As a result we obtain the following table

Field	A−B	d	Rank($\lvert d\rvert$)	Field	A−B	d	Rank($\lvert d\rvert$)
1	25.1	20.1	16	11	25.3	20.3	18.5
2	-1.3	-6.3	2	12	20.0	15.0	9
3	17.6	12.6	7	13	25.2	20.2	17
4	-1.7	-6.7	3	14	-2.6	-7.6	5
5	22.0	17.0	15	15	20.1	15.2	10
6	0	-5.0	1	16	-8.4	-13.4	8
7	20.9	15.9	14	17	20.3	15.3	11
8	20.4	15.4	12	18	-7.2	-12.2	6
9	-2.1	-7.1	4	19	25.3	20.3	18.5
10	26.9	21.9	20	20	20.5	15.5	13

$S_+ = 16 + 7 + 15 + 14 + 12 + 20 + 18.5 + 9 + 17 + 10 + 11 + 18.5 + 13 = 181$
$S_- = 2 + 3 + 1 + 4 + 5 + 8 + 6 = 29$

The test statistic is $S_+ = 181$ and since the p-value is small and the C.I. is strictly greater than 5, we reject the null.

```
> wilcox.test(A,B,alternative="greater",mu=5,paired=TRUE,conf.int=TRUE)
        Wilcoxon signed rank test
data:  A and B
V = 181, p-value = 0.001576
alternative hypothesis: true location shift is greater than 5
95 percent confidence interval:
   9 Inf
sample estimates:
(pseudo)median
          11.8
```

4.3.3 Levene's test for variances

There are instances where we may wish to compare more than two population variations such as the variability in SAT examination scores for students using one of three types of preparatory material. The null assumption (for t populations) is

$$\mathrm{H}_0 : \sigma_1^2 = \cdots = \sigma_t^2$$

versus the alternative that not all σ_i^2's are equal.

Many different methods exist but we will focus on *Levene's test* that is the least restrictive in its assumptions as there are no assumptions regarding the sample sizes/distributions. We still require the assumptions of *independent populations/groups*. However, it is tedious calculation so we will rely once again on software.

For Levene's test, the sampling distribution of the test statistic is

$$T.S. \overset{H_0}{\sim} F_{t-1,N-t}$$

where $N = \sum_{i=1}^{n} n_i$, i.e. the grand total number of observations. Reject H_0 if the p-value $P(F_{t-1,N-t} \geq T.S.) < \alpha$ (area to the right of the test statistic is less than α.)

Example 4.10. Three different additives that are marketed for increasing fuel efficiency in miles per gallon (mpg) were evaluated by a testing agency. Studies have shown an average increase of 8% in mpg after using the products for 250 miles. The testing agency wants to evaluate the variability in the increase. (We will see in later sections how to compare the means).

Additive	% ↑ in mpg	Additive	% ↑ in mpg	Additive	% ↑ in mpg
1	4.2	2	0.2	3	7.2
1	2.9	2	11.3	3	6.4
1	0.2	2	0.3	3	9.9
1	25.7	2	17.1	3	3.5
1	6.3	2	51.0	3	10.6
1	7.2	2	10.1	3	10.8
1	2.3	2	0.3	3	10.6
1	9.9	2	0.6	3	8.4
1	5.3	2	7.9	3	6.0
1	6.5	2	7.2	3	11.9

By running `R` script we obtain a p-value of 0.1803 and fail to reject H_0.
http://www.stat.ufl.edu/~athienit/IntroStat/levene.R

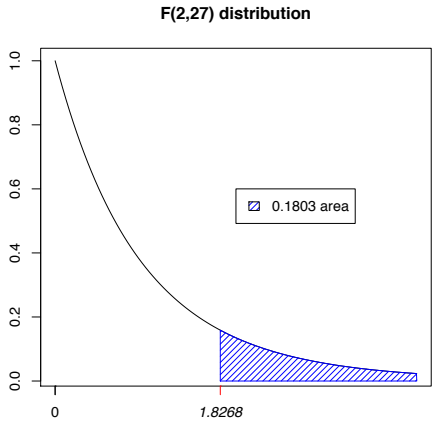


Figure 4.2: $F_{2,27}$ distribution and p-value.

4.4 Contingency Tables: Tests for Independence

Contingency tables are cross-tabulations of frequency counts where the rows (typically) represent the levels of the explanatory variable and the columns represent the levels of the response variable.

We motivate the methodology through an example. A personnel manager wants to assess the popularity of 3 alternative flexible time-scheduling plans among workers. A random sample of 216 workers yields the following frequencies.

	Office				
Favored Plan	1	2	3	4	Total
1	15	32	18	5	70
2	8	29	23	18	78
3	1	20	25	22	68
Total	24	81	66	45	216

Table 4.1: 3×4 contingency table of frequencies.

- Numbers within the table represent the numbers of individuals falling in the corresponding combination of levels of the two variables. Denote n_{ij} to be the frequency count for row i, column j. Let p_{ij} denote the proportion for that cell, e.g. $n_{24} = 18$.
- Row and column totals are called the marginal distributions for the two variables. Denote n_{i+} to the i^{th} row total and n_{+j} to be the j^{th} column total. Let p_{i+} denote the proportion for that row i and p_{+j} denote the proportion for that column j

In Section 2.4.1 we have seen that two events are independent if the joint probability can be written as a product of the marginal proportions (or estimated probabilities). Hence, under independence we expect

$$p_{ij} \overset{\text{ind}}{=} p_{i+}p_{+j} \quad i = 1, 2, 3 \; j = 1, 2, 3, 4.$$

We can model the situation using a multinomial distribution, an extension of the binomial, where there are multiple groups...not just successes or failures (two groups). Expected values, like the binomial, are simply np_{ij}, or

$$\begin{aligned} E_{ij} &= np_{ij} && \text{multinomial} \\ &= np_{i+}p_{+j} && \text{by ind.} \\ &= n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) \\ &= \frac{n_{i+}n_{+j}}{n} \end{aligned}$$

As a result $E_{11} = (70)(24)/216 = 7.7778$. Continuing in same way,

	Office				
Favored Plan	1	2	3	4	Total
1	15(7.7778)	32(26.2500)	18(21.3889)	5(14.5833)	70
2	8(8.6667)	29(29.2500)	23(23.8333)	18(16.2500)	78
3	1(7.5556)	20(25.5000)	25(20.7778)	22(14.1667)	68
Total	24	81	66	45	216

Table 4.2: Table of observed/(expected) frequencies.

To test

$$\mathrm{H}_0 : \text{Levels of one variable are independent of the other}$$

we use *Pearson's chi-square* χ^2 test statistic that is applicable if $E_{ij} = np_{ij} > 5\ \forall i, j$.. This is because this test is actually utilizing the C.L.T.

$$T.S. = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \overset{\mathrm{H}_0}{\sim} \chi^2_{(r-1)(c-1)}$$

where r is the number of rows and c is the number of columns. For a specified α, H_0 : is rejected if

$$T.S. > \chi^2_{1-\alpha,(r-1)(c-1)},$$

or if the p-value $P(\chi^2_{(r-1)(c-1)} \geq T.S.) < \alpha$ (the area to the right of the test statistic is less that α). For the example at hand, the T.S. is

$$T.S. = \frac{(15 - 7.7778)^2}{7.7778} + \cdots + \frac{(22 - 14.1667)^2}{14.1667} = 27.135,$$

the degrees of freedom are $2(3) = 6$ and the p-value is 0.0001366. Therefore, we reject H_0 and conclude that Favored Plan and Office are not independent.

Once dependence is established, of interest is to determine which cells in the contingency table have higher or lower frequencies than expected (under independence). This is usually determined by observing the *standardized residuals* (deviations) of the observed counts, n_{ij}, to the expected counts E_{ij}, i.e.

$$r^{\star}_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

	Office			
Favored Plan	1	2	3	4
1	3.3409	1.7267	-1.0695	-3.4306
2	-0.3005	-0.0732	-0.2563	0.6104
3	-3.0560	-1.6644	1.3428	2.8258

Table 4.3: Table of standardized residuals.

All this can be done in R by utilizing the `chisq.test` function. See http://www.stat.ufl.edu/~athienit/IntroStat/contingency.R

Hence, there appear to be more people than expected (under independence) in office 1 that choose plan 1, and less than expected in plan 3. The opposite applies for office 4.

Remark 4.5. Another test that can be used for any sample size, i.e. even if $E_{ij} \leq 5$ for some cells is the *Fisher's exact test.*

Example 4.11. A personnel director categorizes colleges as most desirable, good, adequate, and undesirable for purposes of hiring their graduates. The director collects data on 156 graduates on their performance and from which college they came from.

	Rating		
School	Outstanding	Average	Poor
Most desirable	21	25	2
Good	20	36	10
Adequate	4	14	7
Undesirable	3	8	6

Inference and conclusions are left as an exercise. An R script can be found at:
http://www.stat.ufl.edu/~athienit/IntroStat/rating.R

Part III

Modules 5-6

Module 5

Regression

We have seen and interpreted the population correlation coefficient ρ between two r.vs that measures the strength of the linear relationship between the two variables. In this chapter we hypothesize a linear relationship between the two variables, estimate and draw inference about the model parameters.

5.1 Simple Linear Regression

The simplest deterministic mathematical relationship between two mathematical variables x and y is a linear relationship

$$y = \beta_0 + \beta_1 x,$$

where the coefficient

- β_0 represents the y-axis intercept, the value of y when $x = 0$,
- β_1 represents the slope, interpreted as the amount of change in the value of y for a 1 unit increase in x.

To this model we add variability by introducing the random variable $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for each observation $i = 1, \ldots, n$. Hence, the statistical model by which we wish to model one random variable using known values of some predictor variable becomes

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{5.1}$$

where Y_i represents the r.v. corresponding to the response, i.e. the variable we wish to model and x_i stands for the observed value of the predictor. Therefore we have that

$$Y_i \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{5.2}$$

Notice that the Ys are no longer identical since their mean depends on the value of x_i.

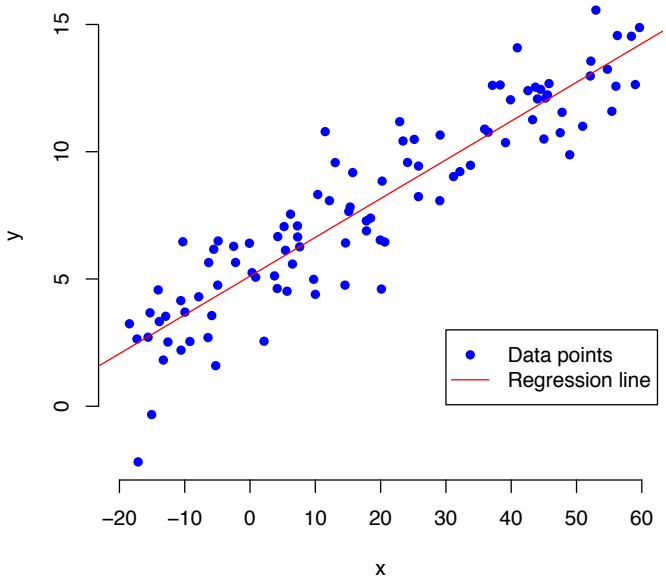


Figure 5.1: Regression model.

In order to fit a regression line one needs to find estimates for the coefficients β_0 and β_1 in order to find the prediction line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The goal is to have this line as "close" to the data points as possible. The concept, is to minimize the error from the actual data points to the predicted points (in the direction of Y, i.e. vertical)

$$\min \sum_{i=1}^{n} (Y_i - E(Y_i))^2 \quad \Rightarrow \quad \min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Hence, the goal is to find the values of β_0 and β_1 that minimizes the sum of the distances between the points and their expected value under the model. This is done by the following steps:

1. Taking the partial derivatives with respect to β_0 and β_1
2. Equate the two resulting equations to 0
3. Solve the simultaneous equations for β_0 and β_1
4. (Optional) Taking second partial derivatives to show that in fact they minimize, not maximize.

Therefore,

$$b_1 := \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} \left(= r\frac{s_Y}{s_X} \right) \tag{5.3}$$

and

$$b_0 := \hat{\beta}_0 = \bar{y} - b_1\bar{x}.$$

Remark 5.1. Do not extrapolate model for values of the predictor x that were not in the data, as it is not clear how the model behave for other values. Also, do not fit a linear regression for data that do not appear to be linear.

Next we introduce some notation that will be useful in conducting inference of the model. In order to determine whether a regression model is adequate we must compare it to the most naive model which uses the sample mean $\bar{Y}$ as its prediction, i.e. $\hat{Y} = \bar{Y}$. This model does not take into account any predictors as the prediction is the same for all values of x. Then, the total distance of a point y_i to the sample mean $\bar{y}$ can be broken down into two components, one measuring the error of the model for that point, and one measuring the "improvement" distance accounted by the regression model.

$$\underbrace{(y_i - \bar{y})}_{\text{Total}} = \underbrace{(y_i - \hat{y}_i)}_{\text{Error}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{Regression}}$$

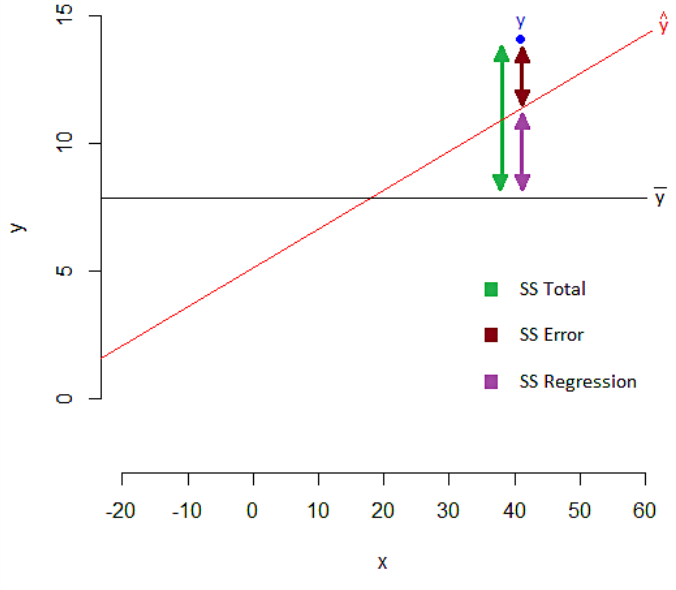


Figure 5.2: Sum of Squares breakdown.

Summing over all observations we have that

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSE}}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSR}}, \tag{5.4}$$

since it can easily be shown that the cross-product term $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$ is equal to 0.

Each sum of squares term has an associated *degrees of freedom* value.

$$\begin{array}{lccccc} \text{SS}: & \text{SST} & = & \text{SSE} & + & \text{SSR} \\ df: & (n-1) & = & (n-2) & + & (1) \end{array}$$

5.1.1 Goodness of fit

A goodness of fit statistic is a quantity that measures how well a model explains a given set of data. For regression, we will use the *coefficient of determination*

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}, \tag{5.5}$$

which is the proportion of variability in the response (to its naive mean $\bar{y}$) that is explained by the regression model, and $R^2\in[0,1]$.

Remark 5.2. For simple linear regression with (only) one predictor, the coefficient of determination is the square of the correlation coefficient, with the sign matching that of the slope, i.e.

$$r=\begin{cases}+\sqrt{R^2} & b_1>0\\ -\sqrt{R^2} & b_1<0\\ 0 & b_1=0\end{cases}$$

Example 5.1. Let x be the number of copiers serviced and Y be the time spent (in minutes) by the technician for a known manufacturer.

	1	2	$\cdots$	44	45
Time (y)	20	60	$\cdots$	61	77
Copiers (x)	2	4	$\cdots$	4	5

Table 5.1: Quantity of copiers and service time

The complete dataset can be found at
http://www.stat.ufl.edu/~athienit/STA4210/Examples/copiers.csv

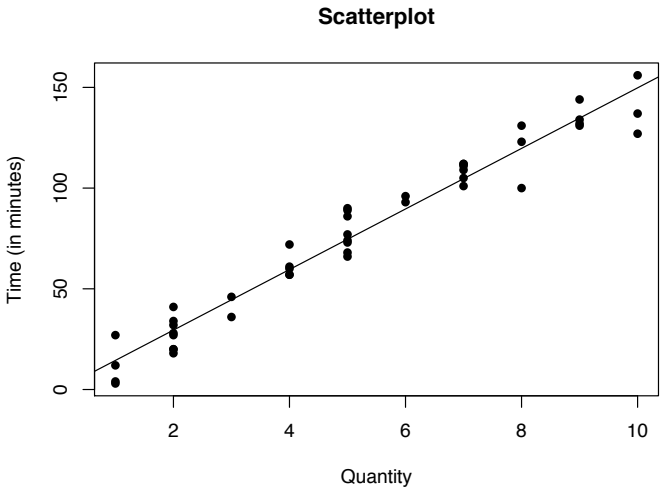


Figure 5.3: Scatterplot of Time vs Copiers.

The scatterplot shows that there is a strong positive relationship between the two variables. Below is the R output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5802     2.8039  -0.207    0.837
Copiers      15.0352     0.4831  31.123   <2e-16 ***
---

Residual standard error: 8.914 on 43 degrees of freedom
Multiple R-squared:  0.9575,Adjusted R-squared:  0.9565
F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

http://www.stat.ufl.edu/~athienit/STA4210/Examples/copier.R

The estimated equation is

$$\hat{y} = -0.5802 + 15.0352x$$

We note that the slope $b_1 = 15.0352$ implies that for each unit increase in copier quantity, the service time increases by 15.0352 minutes (for quantity values between 1 and 10). The coefficient of determination is $R^2 = 0.9575$ implying that 95.75% of the variability in time (to its mean as conveyed by SS Total) is explained by the model.

If we wish to estimate the time needed for a service call for 5 copiers that would be

$$-0.5802 + 15.0352(5) = 74.5958 \text{ minutes}$$

We can obtain SSR, SSE and SST from R, however everything we need was already provided in the output earlier from the `summary` function

- SSE = $\underbrace{(\text{Residual standard error})^2}_{s^2}(df_{\text{Error}})$ (from (5.6))
- SST = SSE$/(1 - R^2)$ (from (5.5))
- SSR = SST-SSE

5.1.2 Distribution of response and coefficients

For the regression model in equation (5.1), we assume that $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and hence $Y_i \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. The variance term σ^2, unlike the earlier chapters, represents the variance of the response Y to its mean as indicated by the model. Therefore,

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{df_{\text{Error}}} = \frac{\text{SSE}}{df_{\text{Error}}} \,(:= \text{MSE})\,. \tag{5.6}$$

The denominator is $n-2$ as we lose 2 degrees of freedom for estimating the two parameters β_0 and β_1. You will recall in earlier chapters the use of $n-1$ degrees of freedom which was due to losing 1 degree of freedom for estimating the center μ_Y by $\bar{y}$.

The coefficients b_0 and b_1 of equation (5.3) are linear combinations of the responses. Therefore, they have corresponding r.vs B_0 and B_1 and since the Ys are independent normal r.vs, by Proposition 2.8 are themselves normal r.vs. Re-expressing the r.v. B_1,

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} Y_i$$

it clear that B_1 is a linear combination of the responses with

$$E(B_1) = \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) \underbrace{E(Y_i)}_{\beta_0 + \beta_1 x_i} = \cdots = \beta_1$$

and

$$V(B_1) = \frac{1}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} \sum_{i=1}^n (x_i - \bar{x})^2 \underbrace{V(Y_i)}_{\sigma^2} = \cdots = \frac{\sigma^2}{\sum_{j=1}^n (x_j - \bar{x})^2}.$$

Thus,

$$B_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \tag{5.7}$$

Remark 5.3. The intercept term is not of much practical importance as it is the value of the response when the predictor value is 0 and inference is omitted. Also, whether statistically significant or not, it is always kept in the

model to create a parsimonious and better fitting model. It can be shown, in similar fashion to B_1 that

$$B_0 \sim N\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\sigma^2\right).$$

Remark 5.4. The larger the spread in the values of the predictor, the larger the $\sum_{i=1}^n (x_i - \bar{x})^2$ value will be and hence the smaller the variances for B_0 and B_1. Also, as $(x_i - \bar{x})^2$ are nonnegative terms when we have more data points, i.e. larger n, we are summing more non-negative terms and the larger the $\sum_{i=1}^n (x_i - \bar{x})^2$.

5.1.3 Inference on slope coefficient

The distribution of the r.v. corresponding to the slope coefficient, that is B_1 is given in equation (5.7). This is a scenario that we are all too familiar with, it is similar to equation (3.2) we have that

$$\frac{B_1 - \beta_1}{\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2},$$

where s stands for the conditional (upon the model) standard deviation of the response. The true variance σ is never known as there are infinite model variations and hence the Student's-t distribution is used, instead of the standard normal, irrespective of the sample size. Important to note is the fact that the degrees of freedom are $n - 2$, as 2 were lost due to the estimation of β_0 and β_1.

Therefore, a $100(1-\alpha)\%$ C.I. for β_1 is

$$\hat{\beta}_1 \mp t_{1-\alpha/2, n-2} s_{\beta_1}$$

where $s_{\beta_1} = s/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$. Similarly, for a null hypothesis value β_{1_0}, the test statistic is

$$T.S. = \frac{\hat{\beta}_1 - \beta_{1_0}}{s_{\beta_1}} \overset{\text{H}_0}{\sim} t_{n-2}$$

Example 5.2. Back to the copier example 5.1, a 95% C.I. for β_1 is

$$15.0352 \mp \underbrace{t_{1-0.025,43}}_{2.016692}(0.4831) \to (14.061010, 16.009486).$$

```
> confint(reg,level=0.95,type="Wald")
                 2.5 %    97.5 %
(Intercept) -6.234843  5.074529
Copiers     14.061010 16.009486
```

5.1.4 Confidence interval on the mean response

The mean is not longer a constant but it is a mean line.

$$\mu_{Y|X=x_{\text{obs}}} := E(Y|X = x_{\text{obs}}) = \beta_0 + \beta_1 x_{\text{obs}}$$

Hence, we can create an interval for the mean at a specific value of the predictor. We simply need to find a statistic to estimate the mean and find its distribution. The sample statistic is

$$\hat{y} = b_0 + b_1 x_{\text{obs}}$$

and the corresponding r.v. is

$$\hat{Y} = B_0 + B_1 x_{\text{obs}} = \sum_{i=1}^{n} \left[\frac{1}{n} + (x_{\text{obs}} - \bar{x}) \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i. \tag{5.8}$$

Note that $\hat{Y}$ can be expressed as a linear combination of the independent normal r.vs Y_i whose distribution is known to be normal (equation (5.2)). Therefore, $\hat{Y}$ is also a normal r.v. After some algebra, we have that

$$\hat{Y} \sim N\left(\beta_0 + \beta_1 x_{\text{obs}}, \left[\frac{1}{n} + \frac{(x_{\text{obs}} - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] \sigma^2 \right).$$

Thus, a $100(1-\alpha)\%$ C.I. for the mean response, $\mu_{Y|X=x_{\text{obs}}}$ is

$$\hat{y} \mp t_{1-\alpha/2,n-2} \underbrace{s \left(\sqrt{\frac{1}{n} + \frac{(x_{\text{obs}} - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}} \right)}_{s_{\hat{Y}}}.$$

Example 5.3. Refer back to Example 5.1. Assume we are interested in a 95% C.I. for the mean time value when the quantity of copiers is 5.

$$74.59608 \mp \underbrace{t_{1-0.025,43}}_{2.016692}(1.329831) \to (71.91422, 77.27794)$$

In R,

```
> predict.lm(reg,se.fit=TRUE,newdata=data.frame(Copiers=5),interval="confidence",l
$fit
       fit      lwr      upr
1 74.59608 71.91422 77.27794

$se.fit
[1] 1.329831

$df
[1] 43
```

5.1.5 Prediction interval

Once a regression model is fitted, after obtaining data $(x_1, y_1), \ldots, (x_n, y_n)$, it may be of interest to *predict a future value of the response*. From equation (5.1), we have some idea where this new prediction value will lie, somewhere around the mean response

$$\beta_0 + \beta_1 x_{\text{new}}$$

However, according to the model, equation (5.1), we do not expect new predictions to fall exactly on the mean response, but close to them. Hence, the r.v. corresponding to the statistic we plan to use is the same as equation (5.8) with the addition of the error term $\epsilon \sim N(0, \sigma^2)$

$$\hat{Y}_{\text{pred}} = B_0 + B_1 x_{\text{new}} + \epsilon$$

Therefore,

$$\hat{Y}_{\text{pred}} \sim N\left(\beta_0 + \beta_1 x_{\text{new}}, \left[1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right]\sigma^2\right),$$

and a $100(1-\alpha)\%$ prediction interval (P.I.) for , for a value of the predictor that is unobserved, i.e. not in the data, is

$$\hat{y}_{\text{pred}} \mp t_{1-\alpha/2,n-2} \underbrace{s\left(\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}\right)}_{s_{\text{pred}}}.$$

Example 5.4. Refer back to Example 5.1. Let us estimate the future service time value when copier quantity is 7 and create a interval around it. The predicted value is

$$-0.5802 + 15.0352(7) = 104.6666 \text{ minutes}$$

A 95% P.I. around the predicted value is

$$104.6666 \mp \underbrace{t_{1-0.025,43}}_{2.016692}(9.058051) \rightarrow (86.399, 122.9339)$$

```
> newdata=data.frame(Copiers=7)
> predict.lm(reg,se.fit=TRUE,newdata,interval="prediction",level=0.95)
$fit
       fit      lwr      upr
1 104.6666 86.39922 122.9339
$se.fit
[1] 1.6119
$df
[1] 43
```

Note that `se.fit` provided is the value for the CI not the PI. However, in the calculation of the PI the correct standard error term is used.
http://www.stat.ufl.edu/~athienit/STA4210/Examples/copier.R

5.2 Checking Assumptions and Transforming Data

Recall that for the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

we assume that $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$. However, once a model is fit, before any inference or conclusions are made based upon a fitted model, the assumptions of the model need to be checked.

These are:

1. **Normality**
2. **Independence**
3. **Homogeneity of variance**
4. **Model fit**

with components of model fit being checked simultaneously within the first three. The assumptions are checked using the residuals $e_i := y_i - \hat{y}_i$ for $i = 1 \ldots, n$, or the *standardized residuals*, which are the residuals divided by their standard deviation. Standardized residuals are usually the default residuals being used as their standard deviation should be around 1.

Although, exact statistical tests exist to test the assumptions (and on conjunction with graphical procedures can help provide a complete picture), linear regression is robust to slight deviations so only graphical procedures will be introduced here. Simply for reference here are a few tests:

1. Normality: Shapiro-Wilk
2. Independence: Runs test, Durbin-Watson
3. Homogeneity of variance: Cook-Weisberg, Levene
4. Model fit: Lack-of-Fit

More details concerning these please refer to Chapter 3 of
http://www.stat.ufl.edu/~athienit/STA4210/reg_notes.pdf

5.2.1 Normality

The simplest way to check for normality is with two graphical procedures:

- Histogram
- P-P or Q-Q plot

A histogram of the residuals is plotted and we try to determine if the histogram is symmetric and bell shaped like a normal distribution is. In addition, to check the model fit, we assume the observed response values y_i are centered around the regression line $\hat{y}$. Hence, the histogram of the residuals should be centered at 0. Referring to Example 5.1, we obtain the following histogram.

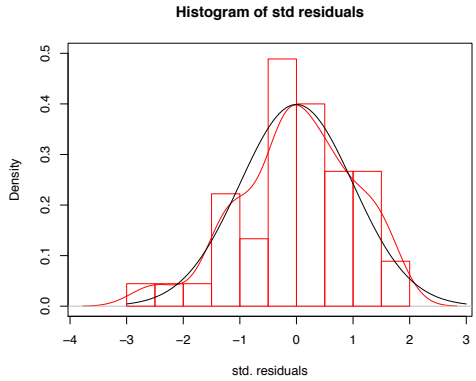


Figure 5.4: Histogram of standardized residuals.

We have referenced P-P and Q-Q plots in Section 2.7. Referring to Example 5.1, we obtain the following P-P plot of the residuals.

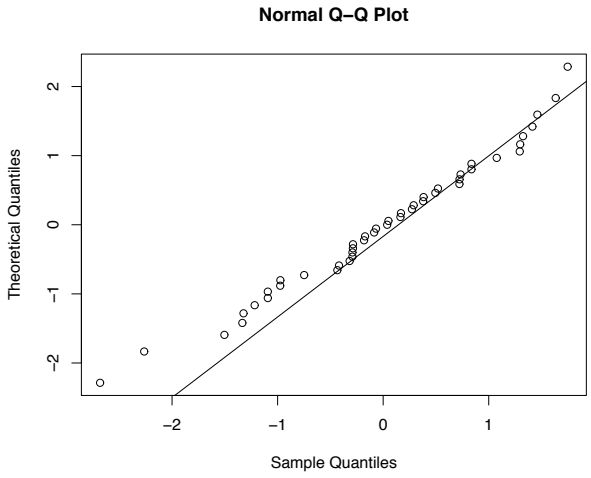


Figure 5.5: Q-Q Plot of standardized residuals.

5.2.2 Independence

To check for independence a time series plot of the residuals/standardized residuals is used, i.e. a plot of the value of the residual versus the value of its position in the data set (usually ordered by date and time). For example, the first data point (x_1, y_1) will yield the residual $e_1 = y_1 - \hat{y}_1$. Hence, the order of e_1 is 1, and so forth. Independence is graphically checked if there is no discernible pattern in the plot. That is, one cannot predict the next ordered residual by knowing the a few previous ordered residuals. Referring to Example 5.1, we obtain the following plot where there does not appear to be any discernible pattern.

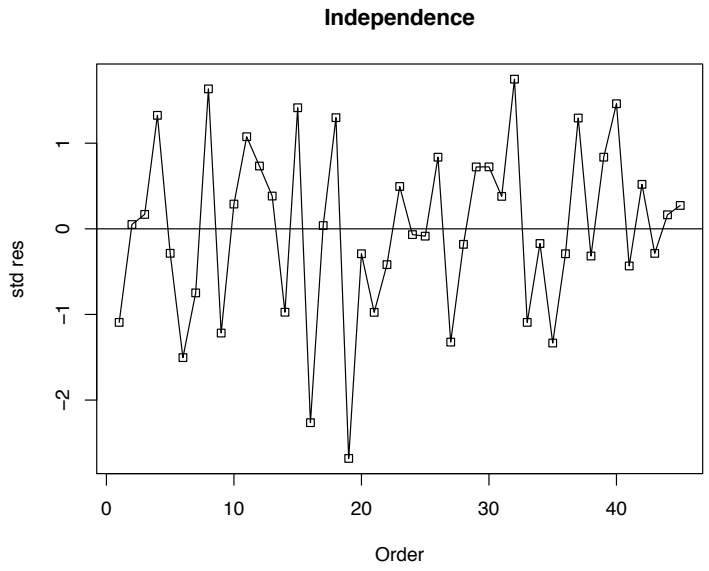


Figure 5.6: Time series plot of residuals.

When creating this plot the order in which the data was obtained must be the same as the way they are in the datasheet.

5.2.3 Homogeneity of variance/Fit of model

Recall that the regression model assumes that the errors ϵ_i have constant variance σ^2. In order to check this assumption a plot of the residuals (e_i) versus the fitted values $(\hat{y}_i)$ is used. If the variance is constant, one expects to see a constant spread/distance of the residuals to the 0 line across all the $\hat{y}_i$ values of the horizontal axis. Referring to Example 5.1, we see that this assumption does not appear to be violated.

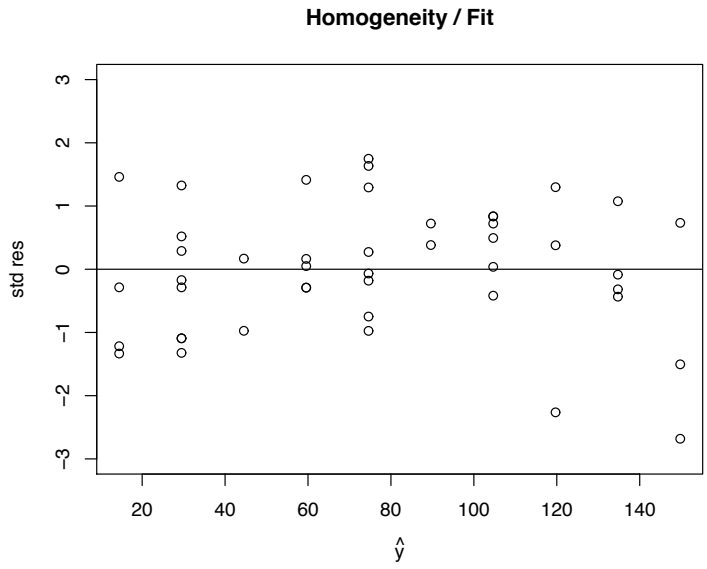


Figure 5.7: Residual versus fitted values plot.

In addition, the same plot can be used to check the fit of the model. If the model is a good fit, once expects to see the residuals evenly spread on either side of the 0 line. For example, if we observe residuals that are more heavily sided above the 0 line for some interval of $\hat{y}_i$, then this is an indication that the regression line is not "moving" through the center of the data points for that section. By construct, the regression line does "move" through the center of the data overall, i.e. for the whole big picture. So if it is underestimating (or overestimating) for some portion then it will overestimate (or underestimate) for some other. This is an indication that there is some curvature and that perhaps some polynomial terms should be added. (To be discussed in the next chapter).

http://www.stat.ufl.edu/~athienit/STA4210/Examples/copier.R

5.2.4 Box-Cox (Power) transformation

In the event that the model assumptions appear to be violated to a significant degree, then a linear regression model on the available data is not valid. However, have no fear, your friendly statistician is here. The data can be transformed, in an attempt to fit a valid regression model to the new transformed data set. Both the response and the predictor can be transformed but there is usually more emphasis on the response.

Remark 5.5. However, when we apply such a transformation, call it $g(\cdot)$, we are in fact fitting the mean line

$$E(g(Y)) = \beta_0 + \beta_1 x_1 + \dots$$

As a result we cannot back-transform, i.e. apply the inverse transformation $g^{-1}(\cdot)$ to make inference on $E(Y)$.

A common transformation mechanism is the Box-Cox transformation (also known as Power transformation). This transformation mechanism when applied to the response variable will attempt to remedy the "worst" of the assumptions violated, i.e. to reach a compromise. A word of caution, is that in an attempt to remedy the worst it may worsen the validity of one of the other assumptions. The mechanism works by trying to identify the (minimum or maximum depending on software) value of a parameter λ that will be used as the power to which the responses will be transformed. The transformation is

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda G_y^{\lambda-1}} & \text{if } \lambda \neq 0 \\ G_y \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

where $G_y = (\prod_{i=1}^n y_i)^{1/n}$ denotes the geometric mean of the responses. Note that a value of $\lambda = 1$ effectively implies no transformation is necessary. There are many software packages that can calculate an estimate for λ, and if the sample size is large enough even create a C.I. around the value.

Example 5.5. Referring to example 5.1, and figure 5.8 we see that $\hat{\lambda} \approx 0.75$.

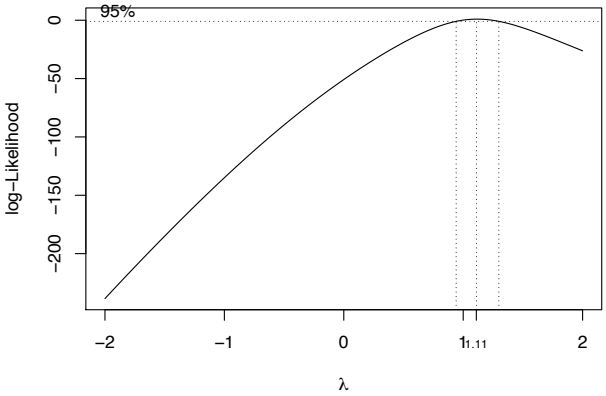


Figure 5.8: Box-Cox plot of example 5.1.

http://www.stat.ufl.edu/~athienit/STA4210/Examples/boxcox.R

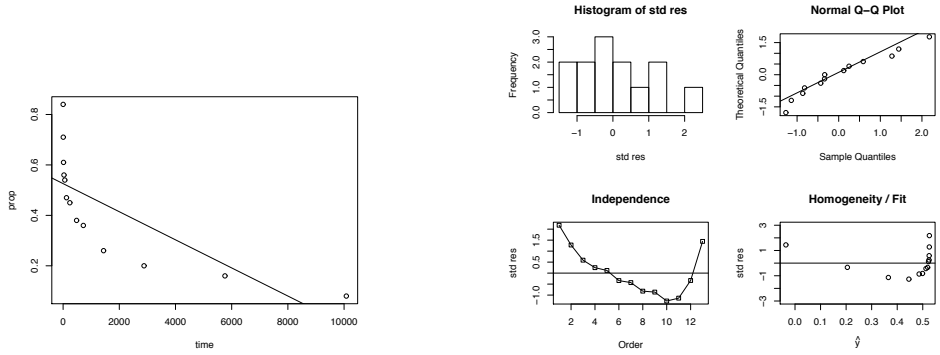
However, one could argue that the value is close to 1 and that a transformation may not necessarily improve the overall validity of the assumptions, so no transformation is necessary. In addition, we know that linear regression is somewhat robust to deviations from the assumptions, and it is more practical to work with the untransformed data that are in the original units of measurements. For example, if the data is in miles and a transformation is used on the response, inference will be on log(miles).

If the model fit assumption is the major culprit violated, a transformation of the predictor(s) will often resolve the issue without having to transform the response and consequently changing its scale.

Example 5.6. In an experiment 13 subjects asked to memorize a list of disconnected items. Asked to recall them at various times up to a week later.

- Response = proportion of items recalled correctly.
- Predictor = time, in minutes, since initially memorized the list.

Time	1	5	15	30	60	120	240
Prop	0.84	0.71	0.61	0.56	0.54	0.47	0.45
Time	480	720	1440	2880	5760	10080	
Prop	0.38	0.36	0.26	0.20	0.16	0.08	

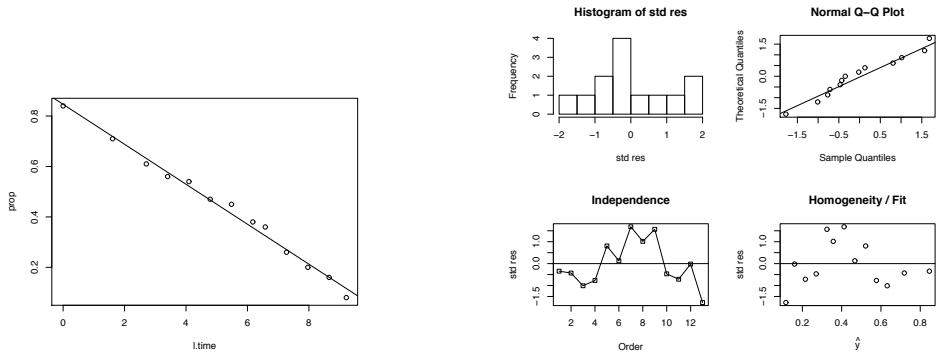



```
> powerTranform(dat$time)
bcPower Transformation to Normality

          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
dat$time     0.0617   0.1087          -0.1514           0.2748

Likelihood ratio tests about transformation parameters
                              LRT df         pval
LR test, lambda = (0)  0.327992  1 5.668439e-01
LR test, lambda = (1) 46.029370  1 1.164935e-11
```

It seems that a decent choice for λ is 0, i.e. a log transformation for time.



http://www.stat.ufl.edu/~athienit/IntroStat/reg_transpred.R

Remark 5.6. Software has a tendency to "zoom" in to where the data is and you may see patterns where there might not be if you were to "zoom" out. Is glass smooth? If you are viewing by eye then *yes*. If you are viewing it via an electron microscope then *no*. It is suggested that the axis where the standardized residuals are plotted are at least from -3 to 3. However, the `check` function was written that automatically adjusts for this.

5.3 Multiple Regression

5.3.1 Model

The multiple regression model is an extension of the simple regression model whereby instead of only one predictor, there are multiple predictors to better aid in the estimation and prediction of the response. The goal is to determine the effects (if any) of each predictor, controlling for the others.

Let p denote the number of predictors and $(y_i, x_{1,i}, x_{2,i}, \dots x_{p,i})$ denote the $p+1$ dimensional data points for $i = 1, \dots, n$. The statistical model is

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \epsilon_i \Leftrightarrow Y_i = \sum_{k=0}^{p} \beta_k x_{k,i} + \epsilon_i \quad x_{0,i} \equiv 1$$

for $\quad i = 1, \dots, n$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

Multiple regression models can also include polynomial terms (powers of predictors) such as

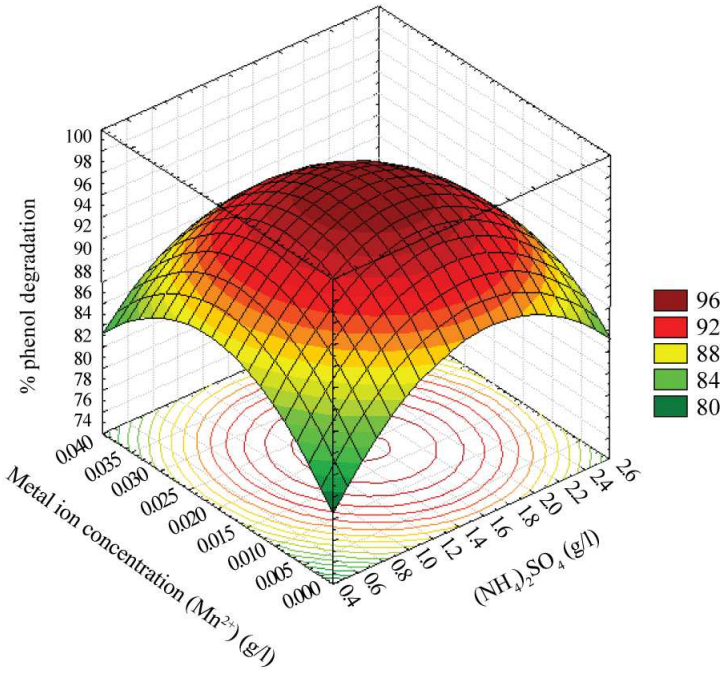
- $x_{2,i} := x_{1,i}^2$ a polynomial order 2 term
- $x_{4,i} := x_{1,i} x_{3,i}$ also a polynomial order 2 term, known as an *interaction* term of x_1 with x_3. Such terms are of particular usefulness when an interaction exists between two predictors, i.e. when the level/magnitude of one predictor has a relationship to the level/magnitude of the other.

The model is still linear as it is linear in the coefficients (β's). Polynomial terms are useful for accounting for potential curvature/nonlinearity in the relationship between predictors and the response.

An example of a regression with 5 predictors but with only 2 unique predictors could be:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + \beta_3 x_{2,i} + \beta_4 x_{1,i} x_{2,i} + \beta_5 x_{1,i}^2 x_{2,i} + \epsilon_i$$

In $p+1$ dimensions, we no longer use the term regression line, but a *response/regression surface*. Let $p = 2$, i.e. 2 predictors and a response. The resulting model may look like



The interpretation of the slope coefficients now requires an additional statement. A 1-unit increase in predictor x_k will cause the response, y, to change by amount β_k, *assuming all other predictors are held constant.* In a model with interaction terms special care needs to be taken as an increase in predictor also causes a change in a predictor that is an interaction term involving said predictor. Take for example

$$E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{x_3}$$

where a 1-unit increase in x_2, i.e. $x_2 + 1$, leads to

$$E(Y|x_1, x_2 + 1) = E(Y|x_1, x_2) + \beta_2 + \beta_3 x_1$$

The effect of increasing x_2 depends on the level of x_1.

5.3.2 Goodness of fit

Coefficient of determination

Goodness of fit is still measured by the coefficient of determination of R^2. Intuitively, we note that *SSR will never decrease, or that equivalently SSE will never increase, as the we include more predictors in the model.* This is because the fitted values $\hat{y}_i$ better fit the observed values of the response (y_i) as we add more predictors. It can never be worse off. Hence, any increase in SSR, no matter how minuscule, will cause R^2 to increase.

This implies that the addition of seemingly unimportant predictors to the model will lead to a tiny increase in R^2. This small "gain" in goodness of fit

is usually not of any practical value and the model may be overcomplicated with redundant predictors. This has lead to the introduction of the *adjusted* R^2, defined as

$$R^2_{adj} := R^2 - \underbrace{(1-R^2)\frac{p}{n-p-1}}_{\text{penalizing fcn.}} \quad \left(= 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}\right).$$

As p increases, R^2 increases, but the second term which is subtracted from R^2 also increases. Hence, the second term can be thought of as a *penalizing factor*.

Example 5.7. A linear regression model of 50 observation with 3 predictors may yield an $R^2_{(1)} = 0.677$, and an addition of 2 "unimportant" predictors yields a slight increase to $R^2_{(2)} = 0.679$. This increase does not seem to be worth the added model complexity. Notice,

$$R^2_{(1)adj} = 0.677 - (1-0.677)\frac{3}{46} = 0.6559$$
$$R^2_{(2)adj} = 0.679 - (1-0.679)\frac{5}{44} = 0.6425$$

that R^2_{adj} has decreased from model (1) to model (2).

Akaike Information Criterion

Similar to the coefficient of determination, the Akaike Information Criterion (AIC) utilizes a minimization of SSE with a penalizing function.

$$\text{AIC} = 2(p+1) + n\log\left(\frac{\text{SSE}}{n}\right) + c$$

where $p+1$ is the number of parameters, n is the sample size and c is a constant that is usually ignored. The constant c is ignored because when multiple models are compared, the c is common and it simply cancels out. A model with smaller AIC is preferable.

5.3.3 Inference

The sum of squares calculation remains as in equation (5.4). However, the degrees of freedom associated with SSE is now $n-(p+1)$. Therefore,

$$\begin{array}{lccccc} \text{SS}: & \text{SST} & = & \text{SSR} & + & \text{SSE} \\ df: & (n-1) & = & p & + & (n-p-1) \end{array}$$

and conditional variance

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1} = \frac{\text{SSE}}{n-p-1} = \text{MSE}.$$

The *Mean Squared Regression* (MSR) and the *Mean Squared Error* (MSE) are defined as

$$\text{MSR} = \frac{\text{SSR}}{p}, \quad \text{MSE} = \frac{\text{SSE}}{n-p-1}$$

Before we continue, it is important to note that there are (mathematical) limitations to how many predictors can be added to a model. As a guideline we usually have **one predictor per 10 observations**. For example, a dataset with sample size 60 should have at most 6 predictors.

Individual tests

Estimating the vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ now falls in the field of matrix algebra and will not be covered in this class. We will simply rely on statistical software.

Inference on the slope parameters β_j for $j = 1, \ldots, p$ is done as in Section 5.1.3 but under the assumption that

$$\frac{B_j - \beta_j}{s_{\beta_j}} \sim t_{n-p-1}.$$

An individual test on β_k, tests the significance of predictor k, **assuming all other predictors j for $j \neq k$ are included in the model**. This can lead to different conclusions depending on what other predictors are included in the model.

Consider the following theoretical toy example. Someone wishes to measure the area of a square (the response) using as predictors two potential variables, the length and the height of the square. Due to measurement error, replicate measurements are taken.

- A simple linear regression is fitted with length as the only predictor, $x =$ length. For the test $\text{H}_0 : \beta_1 = 0$, do you think that we would reject H_0, i.e. is length a significant predictor of area?
- Now assume that a multiple regression model is fitted with both predictors, $x_1 =$ length and $x_2 =$ height. Now, for the test $\text{H}_0 : \beta_1 = 0$, do you think that we would reject H_0, i.e. is length a significant predictor of area given that height is already included in the model?

This scenario is defined as *confounding*. In the toy example, "height" is a *confounding variable*, i.e. an extraneous variable in a statistical model that correlates with both the response variable and another predictor variable.

Example 5.8. In an experiment of 22 observations, a response y and two predictors x_1 and x_2 were observed. Two simple linear regression models were fitted:

```
(1)
y = 6.33 + 1.29 x1

Predictor    Coef  SE Coef     T      P
Constant    6.335    2.174  2.91  0.009
x1         1.2915   0.1392  9.28  0.000

S = 2.95954   R-Sq = 81.1%   R-Sq(adj) = 80.2%

(2)
y = 54.0 - 0.919 x2

Predictor     Coef  SE Coef      T      P
Constant    53.964    8.774   6.15  0.000
x2         -0.9192   0.2821  -3.26  0.004

S = 5.50892   R-Sq = 34.7%   R-Sq(adj) = 31.4%
```

Each predictor in their respective model is significant due to the small p-values for their corresponding coefficients. The simple linear regression model (1) is able to explain more of the variability in the response than model (2) with $R^2 = 81.1\%$. Logically one would then assume that a multiple regression model with both predictors would be the best model. The output of this model is given below:

```
(3)
y = 12.8 + 1.20 x1 - 0.168 x2

Predictor     Coef  SE Coef      T      P
Constant    12.844    7.514   1.71  0.104
x1          1.2029   0.1707   7.05  0.000
x2         -0.1682   0.1858  -0.91  0.377

S = 2.97297   R-Sq = 81.9%   R-Sq(adj) = 80.0%
```

We notice that the individual test for β_1 stills classifies x_1 as significant given x_2, but x_2 is no longer significant given x_1. Also, we notice that the coefficient of determination, R^2, has increased only by 0.8%, and in fact R^2_{adj} has decreased from 80.2% in (1) to 80.0% in (3). This is because x_1 is acting as a confounding variable on x_2. The relationship of x_2 with the response y is mainly accounted for by the relationship of x_1 on y. The correlation coefficient of

$$r_{x_1,x_2} = -0.573$$

which indicates a moderate negative relationship.

However, since x_1 is a better predictor, the multiple regression model is still able to determine that x_1 is significant given x_2, but not vice versa.

Remark 5.7. In the event that the correlation between x_1 and x_2 is strong, e.g. $|r_{x_1,x_2}| > 0.7$, both p-values for the individual tests in the multiple regression model would be large. The model would not be able to distinguish a better predictor from the two since they are nearly identical. Hence, x_1 given x_2, and x_2 given x_1 would not be significant.

Simultaneous tests

This far we have only seen hypotheses test about individual β's. In an experiment with multiple predictors, using only individual tests, the researcher can only test and potentially drop one predictor at a time and refitting the model at each step. However, a method exists for testing the statistical significance of multiple predictors simultaneously.

Let p denote the total number of predictors. Then, we can simultaneously test for the significance of $k(\leq p)$ predictors. For example, let $p = 5$ and the *full model* is

$$Y_i = \beta_0 + \beta_{x_1} x_{1,i} + \beta_{x_2} x_{2,i} + \beta_{x_3} x_{3,i} + \beta_{x_4} x_{4,i} + \beta_{x_5} x_{5,i} + \epsilon_i \tag{5.9}$$

Now, assume that after fitting this model and looking at some preliminary results, including the individual tests, we wish to test whether we can remove simultaneously the first, third and fourth predictor, i.e x_1, x_3 and x_4. Consequently, we wish to test the hypotheses

$$\mathrm{H}_0 : \beta_1 = \beta_3 = \beta_4 = 0 \quad \text{vs} \quad \mathrm{H}_a : \text{at least one of them} \neq 0$$

In effect we wish to test the full model in equation (5.9) to the *reduced model*

$$Y_i = \beta_0 + \beta_{x_2} x_{2,i} + \beta_{x_5} x_{5,i} + \epsilon_i \tag{5.10}$$

Remark 5.8. A full model does not necessarily imply a model with all the predictors. It simply means a model that has more predictors than the reduced model, i.e. a "fuller" model. For example, one may do a simultaneous test to determine if they can drop 2 predictors and hence compare a full versus reduced model. Assume that they do decide to go with the reduced model but then wish to perform an additional simultaneous test on the reduced model. In this second step, the reduced model becomes the new full model that will be compared to a further reduced model.

The SSE of the reduced model will be larger than the SSE of the full model, as it only has two of the predictors of the full model and can never fit the data better. The test statistic is based on comparing the difference in SSE of the reduced model to the full model.

$$T.S. = \frac{\dfrac{\mathrm{SSE}_{red} - \mathrm{SSE}_{full}}{df_{\mathrm{E}_{red}} - df_{\mathrm{E}_{full}}}}{\dfrac{\mathrm{SSE}_{full}}{df_{\mathrm{E}_{full}}}} \overset{\mathrm{H}_0}{\sim} F_{\nu_1,\nu_2}$$

where

- $\nu_1 = df_{\mathrm{E}_{red}} - df_{\mathrm{E}_{full}}$
- $\nu_2 = df_{\mathrm{E}_{full}}$

The p-value for this test is always the area to the right of the F-distribution, i.e. $P(F_{\nu_1,\nu_2} \geq T.S.)$.

Remark 5.9. Note that $\nu_1 = df_{\mathrm{E}_{red}} - df_{\mathrm{E}_{full}}$ always equals the number of coefficients being *restricted* under the null hypothesis in a simultaneous test.

If n denotes the sample size then for our example with $p = 5$ and testing 3 predictors,

$$\nu_1 = (n - 2 - 1) - (n - 5 - 1) = 3$$

Remark 5.10. In computer output a simultaneous test for testing the significance of all the predictors, $\mathrm{H}_0 : \beta_1 = \cdots \beta_p = 0$, is automatically given. This is called the *overall test* of the model. In this case, the reduced model has no predictors, hence

$$Y_i = \beta_0 + \epsilon_i \qquad \Leftrightarrow \qquad Y_i = \mu + \epsilon_i,$$

and thus SSE_{red}=SST and $df_{\mathrm{E}_{red}} = n - 1$. Therefore,

$$T.S. = \frac{\dfrac{\mathrm{SST} - \mathrm{SSE}}{(n-1)-(n-p-1)}}{\dfrac{\mathrm{SSE}}{n-p-1}} = \frac{\dfrac{\mathrm{SSR}}{p}}{\dfrac{SSE}{n-p-1}} = \frac{\mathrm{MSR}}{\mathrm{MSE}} \tag{5.11}$$

Example 5.9. In a biological experiment, researchers wanted to model the biomass of an organism with respect to a salinity (SAL), acidity (pH), potassium (K), sodium (Na) and zinc (Zn) with a sample size of 45. The full model yielded the following results:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 171.06949 1481.15956   0.115  0.90864
salinity     -9.11037   28.82709  -0.316  0.75366
pH          311.58775  105.41592   2.956  0.00527
K            -0.08950    0.41797  -0.214  0.83155
Na           -0.01336    0.01911  -0.699  0.48877
Zn           -4.47097   18.05892  -0.248  0.80576

Residual standard error: 477.8 on 39 degrees of freedom
Multiple R-squared:  0.4867,Adjusted R-squared:  0.4209
F-statistic: 7.395 on 5 and 39 DF,  p-value: 5.866e-05

AIC 690.4836
```

Assuming all the model assumptions are met, we first take a look at the overall fit of the model.

$$\mathrm{H}_0 : \beta_1 = \cdots = \beta_5 = 0 \quad \text{vs} \quad \mathrm{H}_a : \text{at least one of them} \neq 0$$

The test statistic value is $T.S. = 7.395$ with an associated p-value of approximately 0 (found using an $F_{5,39}$ distribution). Hence, at least one predictor appears to be significant. In addition, the coefficient of determination, R^2, is 48.67%, indicating that a large proportion of the variability in the response can be accounted for by the regression model.

Looking at the individual tests, pH is significant given all the other predictors with a p-value of 0.00527, but salinity, K, Na and Zn have large p-values (from the individual tests). Table 5.2 provides the pairwise correlations of the quantitative predictor variables.

	biomass	salinity	pH	K	Na	Zn
biomass	.	-0.084	0.669	-0.150	-0.219	-0.503
salinity	.	.	-0.051	-0.021	0.162	-0.421
pH	.	.	.	0.019	-0.038	**-0.722**
K	.	.	.	.	**0.792**	0.074
Na	.	.	.	.	.	0.117
Zn	.	.	.	.	.	.

Table 5.2: Pearson correlation and associated p-value

Notice that pH and Zn are highly negatively correlated, so it seems reasonable to attempt to remove Zn as its p-value is 0.80576 (and pH's p-value is small). Also, there is a strong positive correlation between K and Na and since both their p-values are large at 0.83155 and 0.48877 respectively, we should attempt to remove K (but not both). In addition, salinity has a large p-value. To test

$$\mathrm{H}_0 : \beta_{\text{salinity}} = \beta_{\text{K}} = \beta_{\text{Zn}} = 0 \quad \text{vs} \quad \mathrm{H}_a : \text{at least one of them} \neq 0,$$

the reduced model needs to be fitted.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -282.86356  319.38767  -0.886   0.3809
pH           333.10556   55.78001   5.972 4.36e-07
Na            -0.01770    0.01011  -1.752   0.0871

Residual standard error: 461.1 on 42 degrees of freedom
Multiple R-squared:  0.4851,Adjusted R-squared:  0.4606
F-statistic: 19.79 on 2 and 42 DF,  p-value: 8.82e-07

AIC 684.6179
```

The test statistic is

$$T.S. = \frac{(8928321 - 8901715)/3}{8901715/39} = 0.0389.$$

with p-value $P(F_{3,39} \geq 0.0389) = 0.9896$, and therefore fail to reject the null which implies that salinity, K and Zn are not statistically significant. Now we proceed with the reduced model as our current model.

At this point we see that Na is marginally significant with a p-value of 0.0871. Some may argue to remove it and some may not (due to its p-value being on the cusp). As the model is "simple" enough it is suggested to keep it. Arguments for keeping Na is that model without it yields

- a model with a higher conditional standard deviation, $s = \sqrt{\text{MSE}} = 472$ (compared to 461.1)
- smaller $R^2_{adj} = 0.4347$ (compared to 0.4606)
- larger AIC of 685.7907 (compared to 684.6179)

Similar to simple linear regression

- one can still create C.I. or P.I. for multiple regression but done via statistical software. For example, fit biomass for pH= 4.15 and Na= 10000 and create a 95% P.I.

```
> newdata=data.frame(pH=4.15,Na=10000)
> predict(linthurst.model.r, newdata, interval="prediction",level=0.95)
       fit       lwr      upr
1 922.4975 -29.45348 1874.448
```

- checking assumptions is done in the exact same way as in Section 5.2.

http://www.stat.ufl.edu/~athienit/IntroStat/linthurst.R

Remark 5.11. We could have reached the same final model choice by simply performing individual t-tests on the coefficients and refitting the model each time, i.e. find the coefficient with the highest p-value and if it is above some cutoff point such as 0.20 then remove it, refit and repeat. However, there are computer algorithms for this.

Example 5.10. Automated model selection example:
http://www.stat.ufl.edu/~athienit/IntroStat/cruise_model_selection.R

There is no one best/correct model, certain models meet certain criteria better than others.

5.4 Qualitative Predictors

Interpreting a regression model with qualitative predictors is slightly different. A qualitative predictor is a variable with groups or classification. The simple case with only two groups will be illustrated by the following example.

Example 5.11. A study is conducted to determine the effects of company size and the presence or absence of a safety program on the number of hours lost due to work-related accidents. A total of 40 companies are selected for the study. The variables are as follows:

$$\begin{aligned} y &= \text{lost work hours} \\ x_1 &= \text{number of employees} \\ x_2 &= \begin{cases} 1 & \text{safety program used} \\ 0 & \text{no safety program used} \end{cases} \end{aligned}$$

The proposed model,

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

implies that

$$Y_i = \begin{cases} (\beta_0 + \beta_2) + \beta_1 x_{1,i} + \epsilon_i & \text{if } x_2 = 1 \\ \beta_0 + \beta_1 x_{1,i} + \epsilon_i & \text{if } x_2 = 0 \end{cases}$$

When a safety program is used, i.e. $x_2 = 1$, the intercept is $\beta_0 + \beta_2$, but the slope (for x_1) remains the same in both cases. Fitting this model yields

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.39945    9.90247   3.171  0.00305 **
x1            0.01421    0.00140  10.148 3.07e-12 ***
x2          -54.21033    7.24299  -7.485 6.47e-09 ***

Residual standard error: 22.87 on 37 degrees of freedom
Multiple R-squared:  0.8154,Adjusted R-squared:  0.8054
F-statistic: 81.73 on 2 and 37 DF,  p-value: 2.658e-14
```

Therefore when $x_2 = 1$ the intercept is $31.399 - 54.210 = -22.811$. The regression line for when a safety program is used is parallel to the line when no safety program is used, and it lies below it.

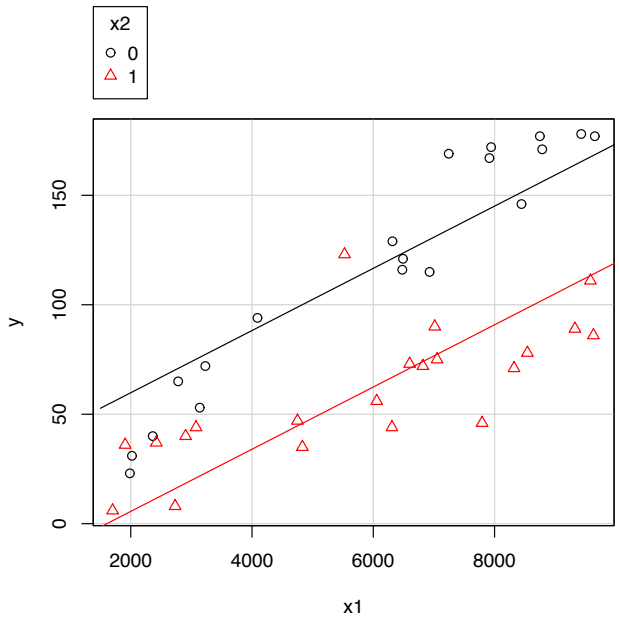


Figure 5.9: Scatterplot and fitted (parallel) regression lines.

Although the overall fit of the model seems adequate, from Figure 5.9 we see that the regression line for $x_2 = 1$ (red), does fit the data well - a fact that can also be seen by plotting the residuals in the assumption checking procedure. The model is too restrictive by forcing parallel lines. Adding an interaction term makes the model less restrictive.

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 (x_1 x_2)_i + \epsilon_i$$

which implies

$$Y_i = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1,i} + \epsilon_i & \text{if } x_2 = 1 \\ \beta_0 + \beta_1 x_{1,i} + \epsilon_i & \text{if } x_2 = 0 \end{cases}$$

Now, the slope for x_1 is allowed to differ for $x_2 = 1$ and $x_2 = 0$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.844082  10.127410  -0.182    0.857
x1           0.019749   0.001546  12.777 6.11e-15 ***
x2          10.725385  14.054508   0.763    0.450
x1:x2       -0.010957   0.002174  -5.041 1.32e-05 ***

Residual standard error: 17.75 on 36 degrees of freedom
Multiple R-squared:  0.8918,Adjusted R-squared:  0.8828
F-statistic:  98.9 on 3 and 36 DF,  p-value: < 2.2e-16
```

The overall fit of the new model is adequate with a $T.S. = 98.90$ but more importantly R^2_{adj} has increased and s has decreased. Figure 5.10 also shows the better fit.

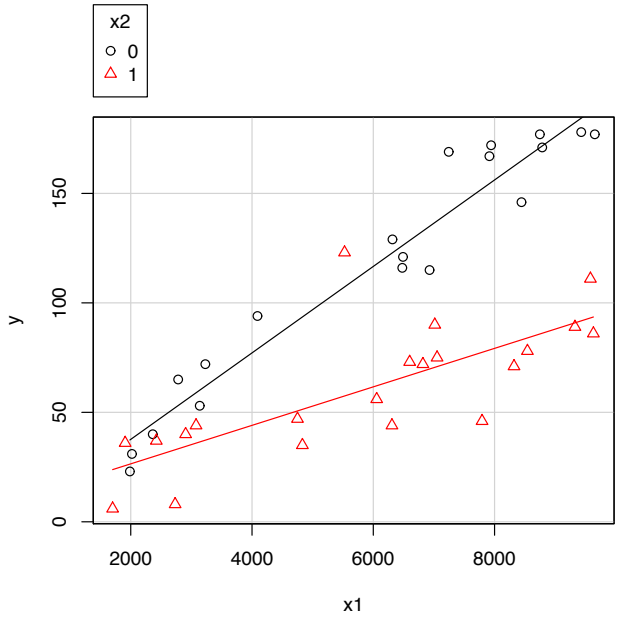


Figure 5.10: Scatterplot and fitted regression lines.

Remark 5.12. Since the interaction term x_1x_2 is deemed significant, then for model parsimony, all lower order terms of the interaction, i.e. x_1 and x_2 should be kept in the model, irrespective of their statistical significance. If x_1x_2 is significant then intuitively x_1 and x_2 are of importance (maybe not in the statistical sense).

Now lets try and to perform inference on the slope coefficient for x_1. From the previous equation we saw that the slope takes on two values depending on the value of x_2.

- For $x_2 = 0$, it is just β_1 and inference in straightforward...right?
- For $x_2 = 1$, it is $\beta_1 + \beta_3$. We can estimate this with $b_1 + b_3$ but the variance is not known to us. From equation (2.2) we have that

$$V(B_1 + B_3) = V(B_1) + V(B_3) + 2\text{Cov}(B_1, B_3)$$

The sample statistics for all the covariances among all the coefficients can easily be obtained in R using the `vcov` function (although we readily have available the variances, i.e. squared standard errors, for β_1 and β_3. Then create a $100(1-\alpha)\%$ C.I. for $\beta_1 + \beta_3$

$$b_1 + b_3 \mp t_{1-\alpha/2,n-p-1}\sqrt{s^2_{\beta_1} + s^2_{\beta_3} + 2s_{\beta_1\beta_3}}$$

Remark 5.13. The sample covariance is not part of the standard output and in R we use the `vcov()` function. Also, this concept can easily be extended to any linear combination of more that two coefficients.

```
> vc=vcov(reg2);vc
             (Intercept)            x1             x2          x1:x2
(Intercept)  102.564428 -1.433300e-02 -102.56442795  1.433300e-02
x1            -0.014333  2.389211e-06    0.01433300 -2.389211e-06
x2          -102.564428  1.433300e-02  197.52920433 -2.799714e-02
x1:x2          0.014333 -2.389211e-06   -0.02799714  4.724125e-06
> sum(reg2$coefficients[c(2,4)])+c(1,-1)*
+ qt(0.025,reg2$df.residual)*sqrt(vc[2,2]+vc[4,4]+2*vc[2,4])
[1] 0.005693056 0.011891084
```

http://www.stat.ufl.edu/~athienit/IntroStat/safe_reg.R

In the previous example the qualitative predictor only had two levels, the use or the the lack of use of a safety program. To fully state all levels only one dummy/indicator predictor was necessary. In general, if a qualitative predictor has k levels, then $k-1$ dummy/indicator predictor variables are necessary. For example, a qualitative predictor for a traffic light has three levels:

- red,
- yellow,
- green.

Therefore, only two binary predictors are necessary to fully model this scenario.

$$x_{\text{red}} = \begin{cases} 1 & \text{if red} \\ 0 & \text{otherwise} \end{cases} \qquad x_{\text{yellow}} = \begin{cases} 1 & \text{if yellow} \\ 0 & \text{otherwise} \end{cases}$$

Braking it down by case we have:

Color	x_{red}	x_{yellow}
Red	1	0
Yellow	0	1
Green	0	0

So the model would be

$$E(Y) = \beta_0 + \beta_1 x_{\text{red}} + \beta_2 x_{\text{yellow}}$$

and hence the mean line, piecewise is

$$E(Y) = \begin{cases} \beta_0 + \beta_1 & \text{if red} \\ \beta_0 + \beta_2 & \text{if yellow} \\ \beta_0 & \text{if green} \end{cases}$$

Testing

- $\beta_1 = 0$, is testing whether the mean for red is the same as for green
- $\beta_2 = 0$, is testing whether the mean for yellow is the same as for green
- $\beta_1 - \beta_2 = 0$, is testing whether the mean for red is the same as for yellow. Here we will need to create a C.I. as before.

The color variable has three categories, one may argue that color (in some context) is an ordinal qualitative predictor and therefore *scores* can be assigned, making it quantitative. For example, you can order a drink in 3 sizes: small, medium and large, and there is an inherent order of 1, 2 and 3.

Size	Score
Small	1
Medium	2
Large	3

Now assume we knew that the medium size is 50% larger than the small, and that the large drink was 350% larger than the small. More representative scores might be

Size	Score
Small	1
Medium	1.5
Large	3.5

So the model would be

$$E(Y) = \beta_0 + \beta_1 \text{score}$$

In terms of frequency (or wavelength) there is also an order of

Color	Frequency (THz)	Score
Red	400-484	442
Yellow	508-526	517
Green	526-606	566

Instead of creating 2 dummy/indicator variables we can create one quantitative variable using the midpoint of the frequency band.

Example 5.12. See relevant lecture video.
http://www.stat.ufl.edu/~athienit/IntroStat/example_dummy.R

Module 6

Analysis of Variance

For $t = 2$ samples/populations we have already seen in previous modules, various inference methods for comparing the central location of two populations. Next we introduce a statistical models and procedures that allow us to compare more than two (≥ 2) populations.

Design / Data	Parametric (Normal)	Nonparametric
Independent Samples (CRD)	1-Way ANOVA	Kruskal-Wallis Test
Paired Data (RBD)	2-Way ANOVA	Friedman's Test

6.1 Completely Randomized Design

The *Completely Randomized Design* (CRD) is a linear model (as is the regression model, and actually is special case of the regression model with a qualitative predictor) where for controlled experiments, subjects are assigned at random to one of t treatments and, and for observational studies, subjects are sampled from t existing groups with the purpose of comparing the different groups.

The CRD statistical model is

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \epsilon_{ij} \quad j = 1, \ldots, n_i, \; i = 1, \ldots, t \tag{6.1}$$

where $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and the restriction (to make model identifiable) that some $\alpha_i = 0$ or $\sum_{i=1}^{t} \alpha_i = 0$. The goal is test the statistical significance of the *treatment effects* α's. If all α's are 0 then it implies that the response can be modeled by a single mean μ rather than individual μ_i's for each treatment/sample.

To see the equivalence to the regression model, assume $t = 2$. The CRD model under the restriction $\alpha_1 = 0$ dictates that

$$E(Y_{ij}) = \begin{cases} \mu & \text{level 1} \\ \mu + \alpha_2 & \text{level 2} \end{cases}$$

However, if we treat this as regression, we have a qualitative predictor with two levels. Let,

$$x = \begin{cases} 0 & \text{level 1} \\ 1 & \text{level 2} \end{cases}$$

and the regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ dictates that

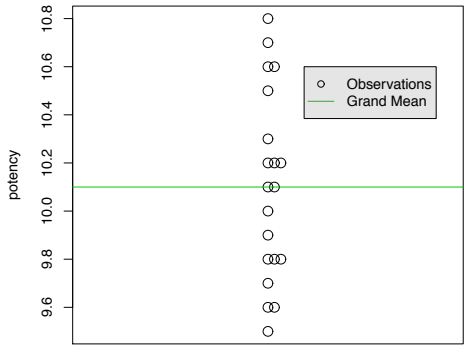
$$E(Y_i) = \begin{cases} \beta_0 \equiv \mu & \text{level 1} \\ \beta_0 + \beta_1 \equiv \mu + \alpha_2 & \text{level 2} \end{cases}$$

This concept can be extended easily to $t > 2$.

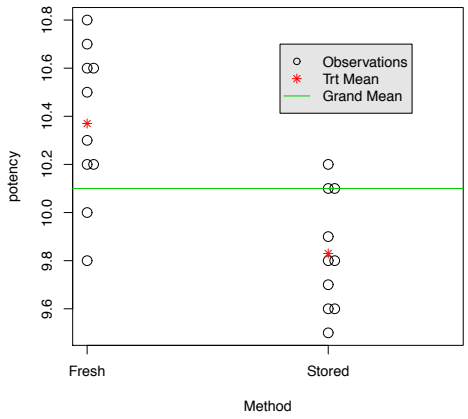
Example 6.1. Company officials were concerned about the length of time a particular drug retained its potency. A random sample of $n_1 = 10$ fresh bottles was retained and a second sample of $n_2 = 10$ samples were stored for a period of 1 year and the following potency readings were obtained.

Fresh	10.2	10.5	10.3	10.8	9.8	10.6	10.7	10.2	10.0	10.6
Stored	9.8	9.6	10.1	10.2	10.1	9.7	9.5	9.6	9.8	9.9

Under the assumption of no treatment effects $H_0 : \alpha_1 = \alpha_2 = 0$ it implies that there is no treatment effect but simply one grand mean (labeled as the "naive" model for regression).



Under the assumption of significant treatment effects, it implies that at least one treatment has a mean that is different from the grand mean (and the others).



The model in equation (6.1) is a linear model so we have the same identity for the sum of squares

$$\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{++})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i+})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i+}-\bar{y}_{++})^2}_{\text{SSTrt}}$$

where $\bar{y}_{i+}$ denotes the sample mean of group i, and $\bar{y}_{++} = \bar{y}$ denotes the sample grand mean. We can simplify each SS term

- $\text{SST} = (N-1)s_y^2$
- $\text{SSTrt} = \sum_{i=1}^{t} n_i(\bar{y}_{i+}-\bar{y}_{++})^2$
- $\text{SSE} = \sum_{i=1}^{t}(n_i-1)s_i^2$

This is once again the same as equation (5.4) in regression with $\hat{y}_i = \bar{y}_{i+}$. In addition, we have a similar identity for the degrees of freedom associated with each SS.

$$\underbrace{N-1}_{df_{\text{Total}}} = \underbrace{N-t}_{df_{\text{Error}}} + \underbrace{t-1}_{df_{\text{Trt}}}, \quad N = \sum_{i=1}^{t} n_i$$

Once again, ($s^2 = \text{MSE}$), and it can be shown (in more advanced courses) that the SS have χ^2 distribution and from that

$$E(\text{MSE}) = \sigma^2$$

$$E(\text{MSTrt}) = \sigma^2 + \frac{\sum_{i=1}^{t} n_i\alpha_i^2}{t-1}$$

As a consequence, under

$$\text{H}_0 : \alpha_1 = \cdots = \alpha_t = 0, \quad \Longrightarrow \quad E(\text{MSTrt})/E(\text{MSE}) = 1.$$

The test statistic for this hypothesis is, with sampling distribution of

$$T.S. = \frac{\text{MSTrt}}{\text{MSE}} \overset{\text{H}_0}{\sim} F_{t-1,N-t}$$

Reject p-value= $P(F_{t-1,N-t} \geq T.S.) < \alpha$. Equivalent to equation (5.11)

Remark 6.1. Checking the assumptions for the CRD model are exactly the same as for regression since both models belong to the same family of *linear models*. In addition, the Box-Cox transformation can also be used just as previously done for regression.

Example 6.2. A metal alloy that undergoes one of four possible strengthening procedures is tested for strength.

Factor	Alloy Strength					Mean	St. Dev.
A	250	264	256	260	239	253.8	9.757
B	263	254	267	265	267	263.2	5.4037
C	257	279	269	273	277	271.0	8.7178
D	253	258	262	264	273	262.0	7.4498

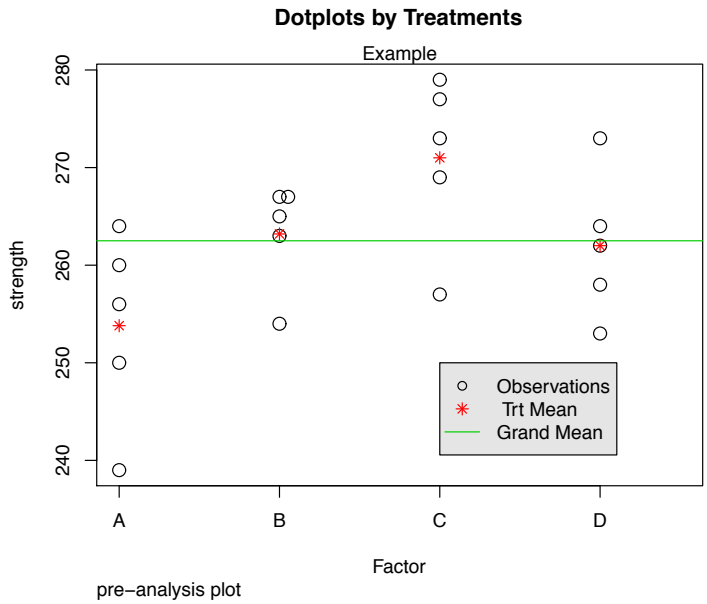


Figure 6.1: Dot plot of metal alloy.

Fitting the data into R we obtain the following ANOVA table.

```
           Df Sum Sq Mean Sq F value Pr(>F)
treat       3  743.4  247.80   3.873 0.0294 *
Residuals  16 1023.6   63.98
```

To test, $H_0 : \alpha_1 = \cdots \alpha_4 = 0$, the $T.S. = 3.873$ with a p-value of 0.0294. Hence, we conclude that not all factors have the same mean.

http://www.stat.ufl.edu/~athienit/IntroStat/anova1.R

6.1.1 Post-hoc comparisons

If differences in group means are determined from the F-test, researchers want to compare pairs of groups. Recall that each pairwise confidence interval, i.e. a C.I. for the difference of two means is equivalent to a hypothesis test. Hence, if each inference is done with $P(\text{Type I Error}) = \alpha$, then the question becomes, if we wish to perform joint (or simultaneous) inference on a certain number of confidence intervals and combine them into one conclusion, then surely the type I error cannot still be α.

Let α_I denote the individual comparison Type I error rate. Thus, $P(\text{Type I error}) = \alpha_I$ on each of the g tests. Now assume we wish to combine all the individual tests into an overall/combined/simultaneous test

$$\mathrm{H}_0 = \mathrm{H}_{0_1} \cap \mathrm{H}_{0_2} \cap \cdots \cap \mathrm{H}_{0_g}$$

H_0 is rejected if any, i.e. at least one, of the null hypotheses H_{0_i} is rejected.

The experimentwise error rate α_E, is the probability of falsely rejecting at least one of the g null hypotheses. If each of the g tests is done with α_I, then assuming each test is independent and denoting the probability of not falsely rejecting H_{0_i} by E_i

$$\begin{aligned} \alpha_E &= 1 - P(\cap_{i=1}^{g} E_i) \\ &= 1 - \prod_{i=1}^{g} P(E_i) && \text{independence} \\ &= 1 - (1-\alpha_I)^g \end{aligned}$$

For example, if $\alpha_I = 0.05$ and 10 inferences (usually pairwise comparisons) are made then $\alpha_E = 0.401$ which is very large.

However, if we do not know if the tests are independent, we use the *Bonferroni inequality*

$$P(\cap_{i=1}^{g} E_i) \geq \sum_{i=1}^{g} P(E_i) - g + 1$$

which implies

$$\begin{aligned} \alpha_E = 1 - P(\cap_{i=1}^{g} E_i) &\leq g - \sum_{i=1}^{g} P(E_i) \\ &= \sum_{i=1}^{g} [1 - P(E_i)] \\ &= \sum_{i=1}^{g} \alpha_I \\ &= g\alpha_I \end{aligned}$$

Hence, $\alpha_E \leq g\alpha_I$. So what we will do is choose an α to serve as an upper bound for α_E. That is we won't know the true value of α_E but we will now it is bounded above by α, i.e. $\alpha_E \leq \alpha$. For example, if we set $\alpha = 0.05$ then $\alpha_E \leq 0.05$, or that simultaneous C.I. from g individual C.I.'s, will have a confidence of *at least* 95% (if not more). Set

$$\alpha_I = \frac{\alpha}{g}$$

For example, if we have 5 multiple comparisons and wish that the overall error rate is 0.05, or simultaneous confidence of at least 95%, then each one (of the five) C.I's must be done at the

$$100\left(1 - \frac{0.05}{5}\right) = 99\%$$

confidence level.

For additional details the reader can read the multiple comparisons problem and the familywise error rate.

In order to control the experiment wise error rate, we will have to adjust the individual error rate of each test. Three popular methods include (from most conservative to least):

1. Bonferroni's Method: Adjusts individual comparison error rates so that all conclusions will be correct at desired confidence/significance level. Any number of comparisons can be made. Very general approach can be applied to any inferential problem.

2. Tukey's Method: Specifically compares all $t(t-1)/2$ pairs of groups. Utilizes special q distribution.

Bonferroni procedure

This is the most general procedure when we wish to test a priori g pairwise comparisons. When all pairs of treatments are to be compared $g = t(t-1)/2$. However, we shall see that the larger the g is the wider the intervals will be.

The steps are:

1. Choose an overall upper bound $\alpha(\geq \alpha_E)$ so that the overall confidence level is
   $100(1-\alpha) \leq 100(1-\alpha_E)\%$.

2. Decide how many and which pairwise comparisons are to be made, g.

3. Construct each pairwise C.I. (or test) with $\alpha_I = \alpha/g$, i.e. confidence level $100(1-\alpha/g)\%$. For comparing treatment means μ_i to μ_j will be

$$\bar{y}_{i+} - \bar{y}_{j+} \mp t_{1-\alpha/(2g),N-t}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Example 6.3. In our example we do not know before hand which comparisons we wish to make so let us perform all $4(3)/2 = 6$ pairwise comparisons with an overall confidence level of at least 95% (since $\alpha_E \leq 0.05$). This implies that each pairwise comparison must be made at the level of

$$100\left(1 - \frac{0.05}{6}\right) = 99.1667\%$$

so our intervals (for the means are)

$$\bar{y}_{i+} - \bar{y}_{j+} \mp \underbrace{t_{1-(0.05/6)/2,16}}_{3.008334} \sqrt{63.98\left(\frac{1}{5} + \frac{1}{5}\right)}$$

Since all our factors have the same sample size 5 we are lucky in that the margin of error is the same for all pairwise comparisons, 15.21813. The sample means were already calculated and hence,

	Difference	Lower	Upper	Differ?
A-B	-9.4	-24.6181	5.8181	0
A-C	-17.2	-32.4181	-1.9819	1
A-D	-8.2	-23.4181	7.0181	0
B-C	-7.8	-23.0181	7.4181	0
B-D	1.2	-14.0181	16.4181	0
C-D	9.0	-6.2181	24.2181	0

It is not unusual to present our conclusion using the following graphical summary, where all factors that are not significantly different are underlined with the same line.

$$\underline{\text{A D}}\ \underline{\text{D B}}\ \underline{\text{B C}}$$

Tukey's procedure

This procedure is derived so that the probability that at least one false difference is detected is α (experimentwise error rate). The C.I. is

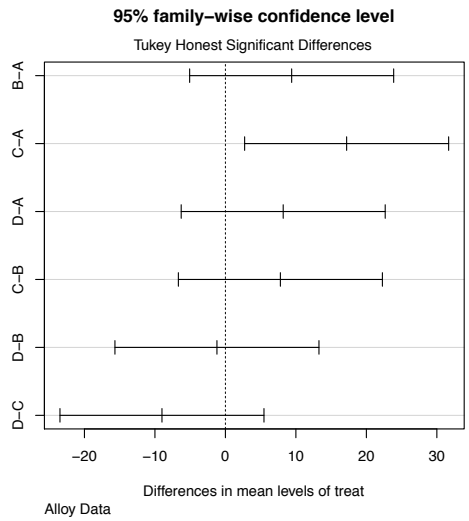
$$\bar{y}_{i+} - \bar{y}_{j+} \mp q_{1-\alpha;t,N-t}\sqrt{\frac{\text{MSE}}{n}}$$

where n is the common sample size for each treatment (which was 5 in the example). If the sample sizes are unequal use (harmonic mean)

$$n = \frac{t}{\frac{1}{n_1} + \cdots + \frac{1}{n_t}}$$

Example 6.4. Continuing with our example, R has a built in function that can create the C.I's and also create a plot for us.

```
    diff        lwr       upr     p adj
B-A  9.4  -5.072915 23.872915 0.2839920
C-A 17.2   2.727085 31.672915 0.0172933
D-A  8.2  -6.272915 22.672915 0.3953011
C-B  7.8  -6.672915 22.272915 0.4372295
D-B -1.2 -15.672915 13.272915 0.9951084
D-C -9.0 -23.472915  5.472915 0.3185074
```



http://www.stat.ufl.edu/~athienit/IntroStat/anova1.R

6.1.2 Distribution free procedure

The Kruskal-Wallis test is an extension of the Wilcoxon rank-sum test to ≥ 2 groups. It is a distribution free procedure (but still requires the assumptions of independence and constant variance). To test

H_0 : The t distributions (corresponding to the t treatments) are identical

the steps are:

1. Rank the observations across groups from smallest to largest, adjusting for ties.
2. Compute the sums of ranks for each group: $T_1, \ldots, T_t$

and then compute

$$T.S. = \frac{12}{N(N+1)} \left(\sum_{i=1}^{t} \frac{T_i^2}{n_i} \right) - 3(N+1) \overset{H_0}{\sim} \chi^2_{t-1}$$

where N is the grand sample size. Reject H_0 if p-value $= P(\chi^2_{t-1} \geq T.S.) < \alpha$

Remark 6.2. Alternate version of this test statistic exist when ties are present, but it is unnecessarily complicated. Usually software will use the "best" adjusted procedure. By hand we suggest to adjust for ties as we have always done.

Rank equivalent to Tukey's HSD

To compare groups i and j, the rank equivalent to Tukey's HSD is

$$\bar{T}_i - \bar{T}_j \mp q_{1-\alpha;t,\infty} \sqrt{\frac{N(N+1)}{24} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

where $\bar{T}_i$ is the average of the ranks corresponding to group i.

Example 6.5. In an experiment, patients were administered three levels of glucose and insulin release levels were then measured.

Glucose conc.	Insulin release			
Low	1.59	1.73	3.64	1.97
Medium	3.36	4.01	3.49	2.89
High	3.92	4.82	3.87	5.39

The corresponding ranks are

Glucose conc.	Insulin release				T	$\bar{T}$
Low	1	2	7	3	13	3.25
Medium	5	10	6	4	25	6.25
High	9	11	8	12	40	10

providing

$$T.S. = \frac{12}{12(13)}\left(13^2/4 + 25^2/4 + 40^2/4\right) - 3(13) = 7.0385$$

with a p-value of 0.02962 (using a χ^2_2 distribution).

With an $\alpha = 0.05$ the 95% pairwise rank analogue Tukey HSD comparisons (using $q_{0.95;3,\infty} = 3.314493$) are

	Difference	Lower	Upper	Differ?
Low-Medium	-3.00	-8.975	2.975	0
Low-High	-6.75	-12.725	-0.775	1
Medium-High	-3.75	-9.725	2.225	0

which summarized graphically is

L M H

http://www.stat.ufl.edu/~athienit/IntroStat/kruskal_wallis.R

Remark 6.3. Note that the Kruskal-Wallis test and the rank analogue to Tukey's may not always be in agreement as these are two different procedures and not equivalent. This is true for any two different methodologies as the "look" at the data in slightly different ways.

6.2 Randomized Block Design

Blocking (where applicable) is used to reduce variability so that treatment differences can be identified. Usually, experimental units constitute the blocks. In effect this is a 2-way ANOVA with treatment being a *fixed* factor, and the experimental unit being the *random* factor. Each subject, receives each treatment and the **order in which treatments are assigned to subjects must be random/arbitrary**.

- Fixed factor being a predictor with a fixed number of levels.
- Random factor being a predictor with potentially infinite levels but only a certain number observed in the experiment.

For example, consider a temperature predictor with levels: $20°\text{F}, 30°\text{F}, 40°\text{F}$. Is this fixed or random? It depends!

- If these are the only 3 settings on a machine, then **fixed**.
- If these were the 3 settings in a greenhouse experiment, where any temperature could be used, then **random**.

It is not always an easy task to determine if a factor should be treated fixed or random. Readers are encouraged to view
http://andrewgelman.com/2005/01/25/why_i_dont_use/.

The model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, t, \ j = 1, \ldots, b$$

with $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and independent $\beta_j \sim N(0, \sigma^2_\beta)$. Notice that the random factor has similar notation as the error. If we were performing a 1-way ANOVA it would have hidden inside it but now we try to account it and remove some "noise" from the model. We still have the same restrictions on the α's that for some $i, \alpha_i = 0$ (or that $\sum \alpha_i = 0$).

		Block			
		1	2	$\cdots$	b
Factor	1	y_{11}	y_{12}	$\cdots$	y_{1b}
	2	y_{21}	y_{22}	$\cdots$	y_{2b}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	t	y_{t1}	y_{t2}	$\cdots$	y_{tb}

This is an extension of the paired test to more than 2 populations. Notice we have more than 2 observations on the same experimental unit.

The sum of squares is

$$\text{SST} = \underbrace{\text{SSTrt} + \text{SSBlock}}_{\text{SSModel}} + \text{SSE}$$

The SSBlock is pulled out of the SSE term of a CRD model.

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{++})^2 \\
\text{SSTrt} &= \sum_{i=1}^{t} b(\bar{y}_{i+} - \bar{y}_{++})^2 \\
\text{SSBlock} &= \sum_{j=1}^{b} t(\bar{y}_{+j} - \bar{y}_{++})^2 \\
\text{SSE} &= \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2
\end{aligned}$$

where $\bar{y}_{i+}$ is the mean of treatment i, $\bar{y}_{+j}$ is the mean of block j, and $\bar{y}$ is the grand mean.

The ANOVA table is then

Source	SS	df	MS	E(MS)	F
Trt	SSTrt	$t-1$	$\frac{\text{SSTrt}}{t-1}$	$\sigma^2 + b\frac{\sum_{i=1}^{t}\alpha_i^2}{t-1}$	$\frac{\text{MSTrt}}{\text{MSE}}$
Block	SSBlock	$b-1$	$\frac{\text{SSBlock}}{b-1}$	$\sigma^2 + t\sigma_\beta^2$	
Error	SSE	$(b-1)(t-1)$	$\frac{\text{SSE}}{(b-1)(t-1)}$	σ^2	
Total	SST	$bt-1$			

In order to test whether there is a treatment effect, i.e. $\text{H}_0 : \alpha_1 = \cdots = \alpha_t = 0$, we notice that $E(\text{MSTrt}) \overset{\text{H}_0}{=} E(\text{MSE})$ and hence the test statistic is

$$T.S. = \frac{\text{MSTrt}}{\text{MSE}} \overset{\text{H}_0}{\sim} F_{t-1,(b-1)(t-1)}$$

with p-value$= P(F_{t-1,(b-1)(t-1)} \geq T.S.)$.

Remark 6.4. The block factor is used to reduce variability and we rarely care about its statistical significance, similar to the intercept term in regression. However, if we wished to test that $\text{H}_0 : \beta_1 = \cdots = \beta_b = 0 \Leftrightarrow \sigma_\beta^2 = 0$, the test statistics would be similar

$$T.S. = \frac{\text{MSBlock}}{\text{MSE}} \overset{\text{H}_0}{\sim} F_{b-1,(b-1)(t-1)}$$

Multiple comparison procedures are the same as in Section 6.1.1 with the exception that $n_i = b$ and degrees of freedom error are $(b-1)(t-1)$.

Definition 6.1. The *Relative Efficiency* of conducting a RBD as opposed to a CRD is the number of times of replicates that would be needed for each treatment/factor level in a CRD to obtain as precise of estimates of differences between two treatment means as we were able to obtain by using b experimental units per treatment in the RBD.

$$\text{RE(RBD,CRD)} = \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}}$$

Example 6.6. Data from a study quantifying the interaction between theophylline and two drugs (famotidine and cimetidine) in a three-period crossover study that included receiving theophylline with a placebo control (Bachman, et. al 1995). We would like to compare the mean theophylline clearances when it is taken with each of the three drugs.

In the RBD, we control for the variation within subjects when comparing the three treatments, i.e. we account for the sum of squares of subjects. In this example there are three treatments ($t = 3$) and fourteen subjects ($b = 14$).

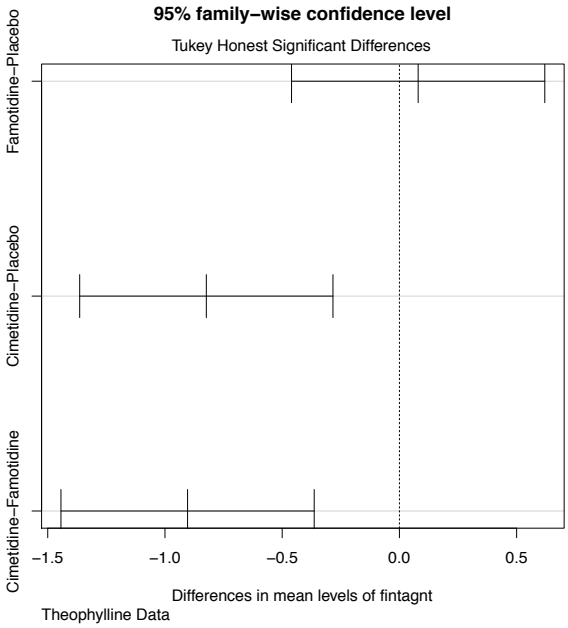
Subject	Placebo	Famotidine	Cimetidine
1	5.88	5.13	3.69
2	5.89	7.04	3.61
3	1.46	1.46	1.15
4	4.05	4.44	4.02
5	1.09	1.15	1.00
6	2.59	2.11	1.75
7	1.69	2.12	1.45
8	3.16	3.25	2.59
9	2.06	2.11	1.57
10	4.59	5.20	2.34
11	2.08	1.98	1.31
12	2.61	2.38	2.43
13	3.42	3.53	2.33
14	2.54	2.33	2.34
Mean	3.08	3.16	2.26

Fitting the model yields the following ANOVA table

```
Response: thcl
          Df  Sum Sq  Mean Sq  F value    Pr(>F)
fintagnt   2   7.005   3.5026   10.591 0.0004321
subj      13  71.811   5.5240   16.703 2.082e-09
Residuals 26   8.599   0.3307
```

We note that drug treatment, `fintagnt`, is highly significant with a p-value of 0.0004 and hence that not all means are equal. Next we perform Tukey's multiple comparison to determine which differ.

```
                          diff        lwr        upr     p adj
Famotidine-Placebo     0.0800000 -0.4601194  0.6201194 0.9282562
Cimetidine-Placebo    -0.8235714 -1.3636908 -0.2834521 0.0022563
Cimetidine-Famotidine -0.9035714 -1.4436908 -0.3634521 0.0008767
```



Therefore only Famotidine and Placebo do not differ.

Cimetidine Placebo Famotidine

Computing the relative efficiency

$$\text{RE(RBD,CRD)} = \frac{(14-1)(5.5240)+14(3-1)(0.3307)}{(14(3)-1)(0.3307)} = 5.9789 \approx 6$$

implying that we would have to have approximately $6(14) = 84$ subjects per treatment in a CRD to have as precise of comparisons between treatment means.
http://www.stat.ufl.edu/~athienit/IntroStat/RBD.R

6.2.1 Distribution free procedure

Friedman's test works by ranking the measurements corresponding to the t treatments within each block.

1. Each block is ranked 1 through t
2. The ranks are summed across subjects corresponding to each treatment yielding $R_{1+}, \ldots, R_{t+}$
3. Similar to the Kruskal-Wallis procedure we will only concern ourselves with SSTrt. In order to test the null hypothesis that all treatment populations are identical (and hence same location), use

$$\begin{aligned} T.S. &= \frac{12b}{t(t+1)} \sum_{i=1}^{t} \left(\bar{R}_{i+} - \frac{t+1}{2} \right)^2 \\ &= \left[\frac{12}{bt(t+1)} \sum_{i=1}^{t} R_{i+}^2 \right] - 3b(t+1) \end{aligned}$$

4. Under the null the sampling distribution of the test statistic is χ^2_{t-1}. As usual, we reject the null if the p-value$= P(\chi^2_{t-1} \geq T.S.) < \alpha$.

The follow-up multiple (pairwise) comparison at the α significance level for comparing treatment i to i', using a Bonferroni adjustment and working with rank sums not averages is

$$R_{i+} - R_{i'+} \mp z_{1-\frac{\alpha}{t(t-1)}} \sqrt{\frac{bt(t+1)}{6}}$$

Example 6.7. A crossover study was conducted to compare the absorption characteristics of a new formulation of valproate-depakote sprinkle in capsules (Carrigan, et al.,1990). There were $b = 11$ subjects, and each received the new formulation (capsule) in both fasting and non-fasting conditions. They also received an enteric-coated tablet. Each drug was given to each subject three times. Among the pharmacokinetic parameters measured was tmax, the time to maximum concentration. The mean tmax for each treatment (capsule-fasting, capsule-nonfasting, enteric-coated-fasting) is given for each subject, as well as the within subject ranks (in the parenthesis).

	Formulation		
Subject	cap f	cap nf	entct f
1	3.5(2)	4.5(3)	2.5(1)
2	4.0(2)	4.5(3)	3.0(1)
3	3.5(2)	4.5(3)	3.0(1)
4	3.0(1.5)	4.5(3)	3.0(1.5)
5	3.5(1.5)	5.0(3)	3.5(1.5)
6	3.0(1)	5.5(3)	3.5(2)
7	4.0(2.5)	4.0(2.5)	2.5(1)
8	3.5(2)	4.5(3)	3.0(1)
9	3.5(1.5)	5.0(3)	3.5(1.5)
10	3.0(1)	4.5(3)	3.5(2)
11	4.5(2)	6.0(3)	3.0(1)
R_{i+}	19.0	32.5	14.5

with

$$T.S. = \frac{12}{11(3)(4)}\left(19^2 + 32.5^2 + 14.5^2\right) - 3(11)(4) = 15.95455$$

and hence p-value= $P(\chi^2_2 \geq 15.955) = 0.00034$, so we reject the null and conclude that all treatments have the same center of location. In the `R` script provided for this example a function was created `friedman.test2` that provides the test statistic and all the pairwise comparisons. However, a built in function `friedman.test` does exist (that adjusts for ties differently) but does not compute the confidence intervals.

```
> friedman.test2(tmax~fformu|subj,data=cap,mc=TRUE)
[1] "95 %  Pairwise CIs on rank sums"
                 Difference    Lower   Upper Differ?
cap_f - cap_nf        -13.5 -18.1967 -8.8033       1
cap_f - entct_f         4.5  -0.1967  9.1967       0
cap_nf - entct_f       18.0  13.3033 22.6967       1

Friedman TS: 15.95455 on df= 2 with p-value= 0.0003431740767
```

Ef Cf Cnf

http://www.stat.ufl.edu/~athienit/IntroStat/friedman.R

Bibliography

[1] Navidi, W.C. *Principles of Statistics for Engineers and Scientists*. McGraw-Hill, 2010.

[2] Ott, R.L. and Longnecker, M.T. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, 2015.

[3] Kutner, M., Nachtsheim, C., Neter, J., Li, W. *Applied Linear Statistical Models*. McGraw-Hill, 2004