Statistics

1. Basic Statistics for Reference

1.1. Rules for variances and standard deviations

1) Independence and addition

- Recalls that the mean of a sum of random variables is *always* equal to the sum of their means.
- This is *only* true for variance when the variances are independent. That is to say, that knowing *x* provides no knwoledge of *Y*. Most probability models assume independence of variance for two random variables, but you should ensure that this assumption is reasonable.

2) Correlation

- When random variables are not independent, the variance of their sums depends on the correlation between them as well as on their individual variances.
- The correlation will fall between -1 and 1. The correlation between two independent variables is 0.

Rules for variances and standard deviations

Rule 1. Because the variance is the mean of the squared deviations from the mean, multiplying each value of *x* by a constant b multiplies σ_x^2 by b^2. This is given as:

$$\sigma_{b*x}^2 = b^2 * \sigma_x^2$$

Rule 2. Adding a constant *a* to every value of *x* does not change the deviation of each point from the mean and therefore does not change the variance.

$$\sigma_{x+a}^2 = \sigma_x^2$$

Rule 3. When *x* and *Y* are *independent*, the variance of $x \pm Y$ is equal to: var(*x*) + var(*Y*).

$$\sigma_{x\pm Y}^2 = \sigma_x^2 + \sigma_Y^2$$

Rule 4. When *x* and *Y* are *correlated* at the population level by the parameter ρ:

$$\sigma_{x\pm Y}^2 = \sigma_x^2 + \sigma_Y^2 + 2\rho\sigma_x\sigma_{\ Y}$$

1.2. IQR, mean median, mode. Robust statistics

1.3. *p-values*

p-values are used in frequentist statistics. Frequentist (i.e., significance) tests are formal procedures for comparing observed data to a hypothesis. The result of the test is often the *p-value*, which is a measure of how well the data and hypothesis agree.

Imagine an example (Section 6.2 of *Introduction to the Practice of Statistics 6e*) where we have a data set comparing the change in student debt from the year 1997 to 2002. The mean difference in debt is $7500. We calculate the probability of observing this difference *under the assumption that there is no difference*. The probability of observing this is 0.00004, or 4 in 100,000. In this case, we have two possible explanations:

A. We have observed a very unlikely event, or
B. The assumption that we used to perform our calculation, that is, that there is no difference in the means, is incorrect.

If the probability is very low, we may decide that (A) is less likely than (B).

1.3.1. Base-rate fallacy

A *p-value* describes the probability of observing the given outcome *under the assumption* that the null hypothesis is true. However, it does not take into account the *base error-rate* - the probability that the null hypothesis is true (and your hypothesis is false) to begin with.

For example:

- Let's say that in our field of research 100 out of every 1000 hypotheses is truly correct. Most of them are incorrect, because that's the nature of science.
- If we power our studies to 80%, we will detect 80 of these correct hypotheses.
- If we accept an alpha level of 0.05, we will erroneously conclude that 45 out of 900 incorrect hypotheses are true.

25 studies were found to be positive at $p < 0.05$. However, the probability that one of these given hypotheses are correct is only ${}^{80}/_{125} = 64\%$. With a base error-rate of 90%, and a power of 80% (both optimistic estimates), the false discovery rate is 36%, *not* 5%.

1.3.2. Interpreting 'non-significance'

From a single study

Under frequentist statistics, a *p-value* greater than 0.05 does not provide evidence that the null hypothesis is true, or that the alternative hypothesis is false. A *p-value* greater than 0.05 should be interpreted as *a lack of evidence against* the null hypothesis and not evidence *for* it.

The distribution of *p-values* from a sample drawn from the 'null hypothesis population' is uniform; that is, obtaining a test result with a *p-value* between 0.95 and 1.0 is exactly the same as the probability of obtaining a test result with a *p-value* between 0.05 and 0. This can be visualized using a simple simulation: > Scripts\Statistics Demos\Power and p-value distribution simulation.R

From replication studies

If you perform 3 identical studies with $\alpha = 0.05$ and $\beta = 0.80$, you can still expect to get 'non-significant results' quite often.

Imagine 2 out of 3 studies are significant, and one is not. The probability of this occuring if we assume H_0 is true is 0.2%:

$$0.05 \times 0.05 \times 0.95 = 0.0024$$

The probability of this occuring if we assume that H_1 is true is 13%:

$$0.8 \times 0.8 \times 0.2 = 0.128$$

Which gives a likelihood ratio of:

$${}^{0.128}/_{0.0024} = 53.9$$

Therefore, the probability that this is a true effect is *still* much higher than the probability that it is a false effect (although in practice studies are rarely identical in design, so you can't just multiply *p-values* and *power* for each study).

1.3.3. Lindley's Paradox: Nominal significance in the presence of high power

With very high theoretical power, the distribution of *p-values*, assuming a real effect, will be very steep, with the majority of *p-values* falling far below the α cut-off. If we observe a *p-value* that is nominally significant, it is actually more likely that the null hypothesis is true than the alternative (ex., if we take the likelihood ratio).

This can also be visualized using the simulation above.

1.3.4. Bias

All null hypothesis statistical testing procedures are some sort of ratio of signal to noise. Therefore, reducing variance (i.e., noise) will improve the power of the test. This is most commonly done by increasing sample size. However, bias has a *net* magnitude and direction, and increasing sample size doesn't reduce bias, and may in fact magnify it.

p-values cannot take into account bias, and so interpretation of results can only really be done in the absence of bias (most other metrics can't take into account bias either, but *p-values* are often used as a single-number summary of the strength of a result).

1.4. Visualizing data: boxplots, bar graphs, histograms, contingency tables

1.4.1.Histograms

A histogram is a graphical representation/estimate of the probability distribution of a continuous variable. Frequency is plotted on the y-axis and the variable of interest is plotted on the x-axis into bins (usually of equal size).

Histograms can be used to assess the probability distribution of a variable, allowing one to:

- Evaluate normality / visualize skew
- Estimate the mean, median, and mode
- Estimate the probability of occurence for a specific value or range of values (i.e., dnorm() / pnorm())

**When examining histograms, it is suggested to view the data using several different bin sizes.

*For histograms to evaluate normality, see "> **Modeling a normal distribution using histograms**"

> Histograms

```
hist()
        x                    - numeric. Object that will be plotted.
        prob = FALSE         - True will convert the histogram into a density plot (versus a frequency plot)
        breaks               - integer. sets the width of the bins
```

- After calling `hist(x, prob = TRUE)`, you can get a density curve for your histogram using: `lines( density(x) )`

1.4.2.Boxplots

1.4.3.Bar graphs

1.4.4.Contingency tables

2. Distributions of Random Variables

2.1. Random sampling and simulations (R)

> 'd' 'p' 'q' 'r' prefixes for distributions in R

# The 'd' 'p' 'q' and 'r' prefixes are available for many functions in order to model distributions (ex., normal, t, binomial distributions)

1. **'p'** – returns the cumulative probability density function for the value that you entered.
   - Ex., `pnorm(1.96)` will return a probability of 0.975, which is the probability of obtaining a value below 0.975 in a normal distribution.
   - Using `pnorm( -abs(x) )` will return the one-sided p-value.

2. **'q'** – returns the inverse cumulative probability density function. the 'q' prefix is the inverse of the 'p' prefix. If you enter a probability, it will return the associated value (for discrete data, it will round to the closest value).
   - Ex., `qnorm(0.975)` will return 1.959964.

3. **'d'** – returns the height of the probability density function at the specified point. That is to say, it will answer the question, "what is the probability associated with *this* observation?"

   - This is useful for discrete data. For continuous distributions, the total heights of probabilities will be greater than 1. I have yet to figure out what it does for continuous data.

4. **'r'** – returns a randomly generated set of numbers (as a vector) for the specified function / distribution. This function is very useful when performing simulations.

> Sampling and simulating null distributions in R

The `sample()` command is very useful for simulations. When population data are available for an experiment, you can use the command in order to generate a large number of random samples to construct a null distribution.

`sample()`

`x` - The object being sampled. I.e., what are the possible values that can be sampled?

- Ex., for a 6-sided die, this would be 1:6

`size` - The number of cases to take per sample.

`replace = FALSE` - Should observations be able to be taken more than once per sample? If TRUE, will treat each case as independent.

`replicate()` - A large number of samples can be generated using `replicate()`

`n` - Integer. Number of times to replicate the function.

`function` - The function to apply (and all of the arguments within them).

# You can construct a null distribution (if you have the population data) using a `for()` loop:

```
N <- 10000 # Number of samples to take
n <- 25 # sample size
Prob <- 0 # value doesn't matter but it must be defined before-hand
for(i in 1:N){
        Null <- sample(x = N, size = n)    # Take a random sample from the population
        Prob[i] <- ( mean (Null) - mean (Population) )
}
```

# `Null` is a vector so-named because it is a simulated "treatment" condition (i.e., drawn from the population) under the assumption that H_o is true.

```
hist(Prob)
effect_size <- mean(Treatment) - mean(Controls)
```

```
p-value <- mean( abs(Prob) ) > effect_size
p-value
```

#This will generate the probability that a simulated “treatment” condition will have a greater effect size than the actual treatment (i.e., more extreme).

2.2. The normal distribution

Many variables are nearly normal, but almost none are truly normal. The normal distribution is rarely a perfect approximation of your data set, but it is often a useful approximation nonetheless.

The normal distribution is a model which dsecribes a curve that is:

1. Symmetric
2. Unimodal
3. Bell-shaped (i.e., has two points of inflexion)

The reason that the normal distiburion is special, and the reason why we really want our data to be normally distributed, is because the entire distribution can be described by only two numbers: **μ**, and **σ**.

2.2.1.Standardizing normal distributions with z-scores

While different normal distributions can take on different values of **μ**, and **σ**, all normal distributions can be standardized into “normal” or “**z**” units. This is done by subtracting the mean from each observation and dividing by the sd.

$$z = \frac{x - \mu}{\sigma}$$

When applied to all observations, the distribution can be described as: $N(\mu = 0, \sigma = 1)$

This is useful because we no longer have to look at raw values. We can instead describe observations by their distance from the mean. For example, We know that a score of *2* is large, and a score of -*3* is very small, without knowing the mean or the units of measurement.

2.2.2.Calculating probability with z-scores

Because z-scores are standardized, we can easily determine the probability of observing a particular values once we know the z-score. Probabilities can be determined from a probability table or computed using software.

In general

- 68% of data fall within +/- 1 sd of the mean
- 95% of data fall within +/- 2 sd of the mean
- 99.7% of data fall within +/- 3sd of the mean

`> Computing probability distribution functions in R`

#The `pnorm()` function returns the cumulative density function for a normal distribution; as a default, mu = 0 and sd = 1.

#Ex., `pnorm(0)` will return 0.5; it is equivalent to asking for for probability of getting a z-score of 0 or below.

#However, you can also calculate probability distribution functions for non-standardized normal distributions by inputing values. Ex.,

```
pnorm(0, mean = 12, sd = 2)
```

#The `qnorm()` function is the inverse of pnorm. That is, if you input a probability, it will return the value associated with that probability. I.e., "What value will have a cumulative probability of 0.5? Ex,

```
qnorm(0.5)
```

2.2.3.Evaluating normal approximations

Histograms

> Modeling a normal distribution using histograms

# This is a technique that allows you to visually inspect to see if a data set appears to follow a normal distribution.

# The fast way:

```
hist(x, prob = TRUE)
curve(dnorm( mean = mean(x), sd = sd(x)), add = TRUE)
```

# Another way:

# Define the x-parameter (i.e., range of the x-axis) and the y-parameter (i.e., the values of y that a normal distribution with the given mean and sd would take on) for the density probability function.

# The x-parameter follows the general outline: xfit <- c(n_1:n_2, length = n_3).

# The y= parameter uses the dnorm() function

```
hist(x, prob = TRUE)
xfit <- seq( min(x), max(x), length = length(x) )
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
lines(xfit, yfit, col = "blue")
```

Q-Q Plots

Normality can also be inspected visually using a normal probability plot (a special case of the Q-Q plot). The normal probability plot will plot the 'observed' value (i.e., z-score) of a data point against the 'expected' value of that data point if it were to follow a normal distribution.

The closer the data are to following the 'expected' values, the more normal they are. Conversely, plots with a consistent curvature are likely not normal; in this case, looking at the y-axis may help to determine which direction the data are skewed.

> Modeling a normal distribution using Q-Q plots

```
qqnorm(x)
qqline(x)
```

3. Foundations for frequentist inference

The frequentist approach is one of many ways of performing inference. This approach first *assumes that the null hypothesis is true, and then tries to disprove it*. This is great for *disproving* things, but it is not very good at providing evidence *for* things. For this reason, this branch of statistics is somewhat controversial (particularly because of the interpretation of *p-values*).

Likelihood and Bayesian statistics provide alternative approaches that are slightly better at communicating evidence *for* a given hypothesis. Either way, frequentist statistics are the most commonly used and widely-understood - so suck it up and learn.

3.1. Inference in a nutshell

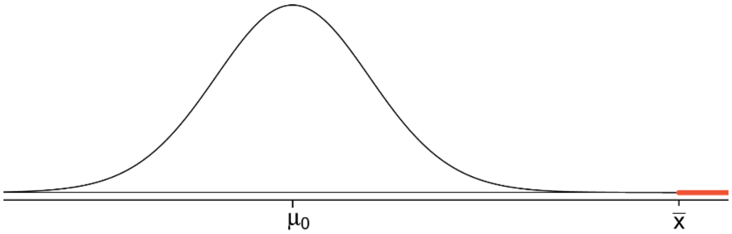
A typical inferential question will ask:

- “Is this group that we’ve sampled different from the general population?” (one-sample)
    - We answer this by comparing our sample distribution to the real or estimated population distribution(i.e., null distribution).
- “Are these two groups that we’ve sampled different from each other?” (two-sample)
    - We answer this by comparing the distribution of the difference between the samples to the estimated population distribution under the assumption that the difference in means is = 0 (i.e., null distribution).

The same approach is used to answer both these questions.

Inference in a nutshell

1. Identify the population distribution. That is to say: what kind of *values you should expect in your sample if your sample is taken from this population*.
    - You can do this by looking at population data (uncommon) OR by *estimating* the population distribution based on the characteristics of your sample (common).



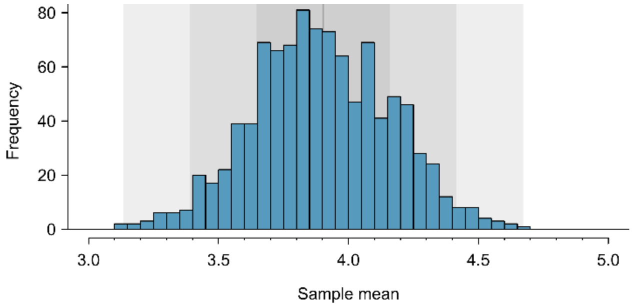
2. Compare your sample mean $\bar{x}$ to this distribution; does it look like it belongs to the population distribution?
    - We can calculate the probability (either exact or approximate) that *our sample mean or a value even more extreme* was drawn from the population by chance. This is a *p*-value.

3. In this case, we would say "*no, the probability that our sample was drawn from this population by chance is too low*." In this case, the assumption is that the sample belongs to a different population.

3.2. Point estimates and standard error of the mean

The mean, $\bar{x}$, is a point estimate of the population mean, μ. If you had a to choose a single value to represent the population distribution, it would be the mean. However, $\bar{x}$ will vary between samples because of sampling variance. Therefore, point estimates are not exact.

Ex., imagine that we take 1000 samples where n = 100. If we plot the mean for each of these samples, we would get a normal distribution. This is a sampling distribution for the mean.



From this figure, we see:

- There is variability in the sampling; not all values of $\bar{x}$ are the same.
- $\bar{x}$ measurements tend to "fall around" the true mean, μ, and appear normally distributed.

In summary: The statistic $\bar{x}$ is a normally distributed random variable with a sampling distribution with a mean of μ and a standard deviation of $\sigma/\sqrt{n}$. The standard deviation of $\bar{x}$ is called the Standard Error of the Mean (sem, or $SE_{\bar{x}}$)

$$sem = \frac{\sigma}{\sqrt{n}}$$

Where σ denotes the population sd
n is the size of each sample.

Note: The true sem requires knowledge of the population sd. This is not avilable most of the time; however, we can estimate it using our sample sd for samples larger than 30 and when the population distribution is not skewed (cf. Central Limit Theorem).

3.3. Confidence intervals

Confidence intervals (CI) are a way to describe the population mean using a range of plausible values.
*While we discuss CIs in the context of the population mean, it should be noted that CIs can be constructed for any *parameter* (ex., μ, σ). It should also be noted that CIs never refer to *statistics* (ex., $\bar{x}$, sd).

The general equation for calculating CIs is as follows:

$$Confidence\ interval = point\ estimate\ \pm margin\ of\ error$$

For example, for a normal distribution we can construct a CI around $\bar{x}$ based on critical z-values. For 95%, the critical value is 1.96 (two-tailed).

$$95\%\ Confidence\ interval = \bar{x} \pm 1.96 * sem$$

**A 95% CI means that: 95% of the confidence intervals built around $\bar{x}$ *from different samples* will contain the true mean, μ.

- We *can not* say: that there is a 95% probability that the interval contains μ; once calculated, the CI either contains μ or it does not.
- We *can* say: that there is a 95% probability that the calculated CI in a replicate study will contain μ.
- We *can* say: that we arrived at this interval by a method that gives correct results 95% of the time.

> Computing confidence intervals using z-scores

You should have already decided if you need to use a t-distribution or z-distribution. You can find how to make this decision in this document or online. For computing CIs using a t-distribution, see "**4.2.2. Confidence intervals using the t-distribution**."

```
# Compute the sample sd. Recall that this will be used as our estimate when computing sem.
sd_variable1 <- sd(variable1)
# Calculate the sem.
n <- length(variable1)
sem_variable1 <- sd_variable1 / sqrt( n )
# For a z-statistic, simply input the critical z-score.
mean(variable1) + 1.96 * c(-sem_variable1, sem variable1)
```

3.4. Null distributions

Imagine we are asking the question, *"Do mice that are fed a high fat diet weigh more than mice that are fed mouse chow?"* We purchase 24 young mice from a company with a colony of 3,000 mice and randomly assign 12 to a high fat diet and 12 to a chow diet. By 8 weeks old, the mice on the high fat diet weigh on average 30g and the chow mice weigh 26g. Do we conclude that the high fat mice do indeed weigh more? How do we know this is not due to chance (i.e., variability due to sampling from the population)?

We can answer this question using a **null distribution**: the distribution of sample means that would be observed if we just sampled randomly from the whole population.

Where population data are available:

- In some cases, we may have access to the whole population. For example, if it is small enough that we can measure all of the individuals. In this case, we have these data because we purchased the mice from a company that has records of the weights for all 3,000 of their mice.
- We can run a simulation, and sample 12 mice many times (ex., 10,000). We can count the proportion of times that the observed sample mean was <u>more extreme</u> than what we observed in our experiment.
- For example, perhaps only 1 in every 100 random samples has a mean weight greater than 30g. This means that *in our experiment, the probability of observing a result at least as extreme as what we observed is 1%* (i.e., p = 0.01).

<u>Where population data are not available</u>:

- More commonly, we do not have access to data for the whole population. In this case, how can we construct a null distribution in order to determine a *p-value*?
- We make use of the **<u>Central Limit Thereom</u>** <u>(described below)</u> to approximate our sample to the normal distribution, allowing us to calculate z-scores and t-statistics.

3.5. The Central Limit Theorem (CTL)

The CTL states that, <u>when certain conditions are met</u>, the mean (or other mathematical function) of a <u>sufficiently a sufficiently large sample</u> of a random variable will be approximately normally distributed (if infinite samples are taken), regardless of the underlying distribution.

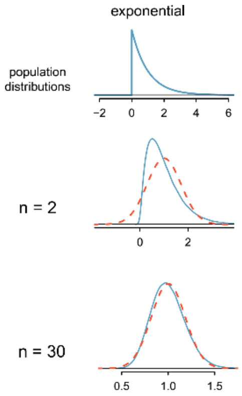
**That is to say: <u>The sampling distribution of $\bar{x}$ is normally distributed if n is large</u>.

<u>These conditions are</u>

- The observations must be independent
- The interations must have the same distribution. This means that they must have the same variance and expected value (i.e., expected mean).

<u>What determines what a "sufficiently large sample" is:</u>

- The sampling distribution will tend to appear more normal than the initial population distribution, but as you increase the sample size, the sampling distribution better approximates a normal distribution.
- The degree of skew in the <u>population distribution</u> and the number of outliers will affect this.
    - With greater skew and more/extreme outliers, you will need more iterates for the normal approximation to be reasonable.
- For most distributions, a sample size $n \geq 30$ is a reasonable number to apply a normal distribution. However, as this number increases, our normal approximation becomes more accurate.



<u>Why is this important?</u>

**As long as we have a large enough sample size (and meet the above conditions), we can assume that the sampling distribution is normally distributed. This means that we can construct a null distribution without any information about the population distribution.
((Recall: the null distribution is the distribution of sample means that would be observed if we just sampled randomly from the whole population, which can be used to calculate *p-values*)

Note: The CLT refers to the tendency for the distribution of sampling distributions (i.e., the distribution of $\overline{x}$) to approximate a normal distribution. It does provide any information about the type of the distribution that you will get from a sample from a population.

- i.e., Do not expect that the values *within* a sample will be normally distributed.

3.6. Pitfalls in Inference

3.6.1. Statistical tests underestimate error

Statistical tests only account for random sampling error (in the standard error term). Other sources of error are not taken into account, including:

- Measurement error (which may be biased in the case of non-response, or simply lead you to underestimate noise)
- Bias from non-random selection, non-response, survivorship, etc.

4. Inference for numerical data (distributions)

4.1. z-statistic

When data are nearly, normal, we can test hypotheses using z-statistics. If we assume that the population distribution is normal, we can construct a null distribution that is normal. The z-statistic itself is discussed in more detail in earlier sections; this section will just discuss hypothesis testing.

4.1.1. One-sided z-test

1. State your hypotheses in plain language
2. Then in mathematical notation.
3. Take an appropriate sample and identify the mean.
4. Verify conditions to ensure that the sem is reasonable and that the sampling distribution for $\overline{x}$ is nearly normal (cf. **2.2. The normal distribution**)
    - Visualize distribution of data, perform tests for normality
    - Conceptually, do you expect data to be normally distributed?

- Are outliers rare?

5. Compute the sem and construct the null distribution.
6. Compute the *p-value* using the z-score against the null distribution.

4.2. t-statistic

While x̅ will follow a normal distribution when sample sizes are large, the sem is more difficult to calculate accurately without population data.

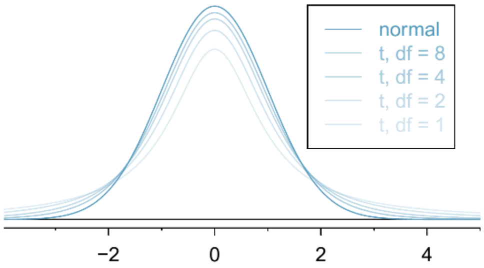
$$\text{Recall: } Z = \frac{\bar{x}- \mu}{\sigma/\sqrt{n}}$$

Often, we will estimate σ with the sample sd. However, σ is a constant value, and sd is a statistic (i.e., it has a sampling distribution). Therefore, the statistic calculated from the sample sd will not be normally distributed. Instead, this is called the t-distribution:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

with *n – 1* degrees of freedom

Because the t-statistic is calculated with *s*, which varies, the t-distribution looks like the normal distribution, except with greater variance.



Properties of a t-distribution

- The t-distribution is always centred around 0.
- The t-distribution can be described by a single parameter: degrees of freedom (df), which describes the shape of the curve.
  - As df increases, the t-distribution more closely resembles a normal distribution.
  - df is related to the sample size (*df* = *n -1*)

*For any given confidence interval (ex., 95%) the critical t-statistic(for df = 30, t = 2.04) will *always* be greater than the critical z-statistic (z =1.96) for the same population. This is because, with greater variance, a larger proportion of the values will fall in the tails. The exact value of the t-statistic depends on the number of degrees of freedom in the distribution (at df = ∞, t = 1.96)

Conditions for using a t-distribution for inference

- Observations are independent
- The population distribution is nearly normal.
  - However, *t* procedures are quite robust against non-normality. When sample sizes are equal and the distributions are similar, the normality condition is not as important. Therefore, *t*-tests are most robust when sample sizes are similar (and studies should be designed accordingly).

  - With large sample sizes, the sample sd will still accurately estimate the population sd and will therefore be appropriate even if the population distribution is non-normal.
- The assumption that the population distribution is nearly normal rules out outliers. However, *t* procedures are not robust against outliers because the mean and sd are not resistant to outliers.
- For samples sizes less than 15, *t* procedures should only be used if you know that the data are close to normal.

4.2.1.t-distributions versus z-distribution

z-distribution:

- When you can accurately calculate the sem:
  - When the population σ is known, a z-distribution should be used (i.e., almost never).
  - When you have large samples sizes (cf. CLT). However, if the sample size is not *extremely large*, you will underestimate the margin of error (cf. t-statistics).

t-distribution:

- When you cannot accurately calculate the sem:
  - When the population σ is not known, a t-distribution should be used. (i.e., almost always)
  - With large sample sizes, the t-distribution becomes equivalent to the z-distribution. However, at larger sample sizes, there is no real reason to choose a z-distribution over a t-distribution.
- For sample sizes less than 15, use *t* procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, to not use *t* (or Z).

4.2.2.Confidence intervals using the t-distribution

Recall, for a normal distribution: $95\%\ Confidence\ interval = \bar{x} \pm z * sem$

CIs are calculated the same way when using a t-distribution as the null distribution, with the t-statistic instead of the z-score:

$$95\%\ Confidence\ interval = \bar{x} \pm t_{df} * sem$$

> Computing 1-sample confidence intervals under the t-distribution in R

# 1. The fast way

```
t.test(variable1, conf.level = 0.95)
```

# 2. The conceptual way

# For a t-statistic, use the `qt()` function to compute the margin of error. This provides the quantile function (i.e., the inverse of the cumulative probability distribution function); that is to say, if you input a probability and df, it will return the associated t-statistic.

```
Margin_of_Error <- qt( (1 - 0.05/2) , df = n - 1 )  * sem_variable
mean(variable1) + c(-Margin_of_Error, Margin_of_Error)
```

Note that for a 95% CI, we computed 97.5% one-sided margins of error, which add up to 95%.

4.2.3. One-sample t-test

A one-sample t-test is performed the same way as a z-test, only using the t-statistic. Construct a null distribution for your population using t_{df}, and find the t-statistic that corresponds to the mean of your experimental group. Finally, find the p-value (using the constructed null distribution + software, or by consulting a t-table).

* Because the two-sample t-test is more commonly used, assumptions and theory of the *t* test are discussed in **4.3.2. Two-sample t-test.**

* See **> Computing t-tests in R** for instructions on how to use `t.test()`.

Calculating a two-sample t-test (manually, assuming equal variance)

Our hypothesis would look like:

$H_o\ \mu = \bar{x}$; $|t_{df}| < t_{critical}$

$H_A\ \mu \neq \bar{x}$; $|t_{df}| > t_{critical}$

1. Calculate the degrees of freedom. To do this manually, simply use the *df* from the smaller *n*.
2. Find the critical *t*-statistic, $t_{critical}$
3. Calculate the t-statistic. Recall: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
4. Make a decision on whether or not to reject the null hypothesis based on the observed t-statistic.

4.3. Inference for 2-samples (focus on *t* procedures)

4.3.1. Inference for two samples

In general, inference for two-samples differs from one-samples in that:

- $\bar{x}_{diff}$ is used instead of $\bar{x}$
- The sem uses information from both samples using the equation:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

- The sampling distribution for $\bar{x}_{diff}$ is approximately normally distributed with a mean of μ_1 - μ_2 and a $SE_{\bar{x}_1 - \bar{x}_2}$ of $\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$. Note however, that if σ is replaced by the estimator sd, the resulting distribution will not have a *t*-distribution (this only works when a single σ is replaced by sd).

Rationale for $\bar{x}_{diff}$

- While $\bar{x}_1$ is variable, the "expected" value of $\bar{x}_1$ is μ_1. The same is true for $\bar{x}_2$.
- Therefore, on average: $\bar{x}_{diff} = \bar{x}_1 - \bar{x}_2 = \mu_1 - \mu_2$
- The difference between two sample means is the best unbiased representation of the difference between two population means.

Rationale for SE$_{\bar{x}_1-\bar{x}_2}$

- If two random variables are independent, then the *variance of their difference* is equal to the *sum of the two variances* (cf. Variance Sum Law):

$$Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$$

Distribution

- The sampling distributions for $\bar{x}_1$ and $\bar{x}_2$ are: $\bar{x}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$; $\bar{x}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$
- Following the rationale form above, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is: $\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2})$
- While this distribution does not strictly follow the *t*-distribution, it is a good approximation. We use the *t*-distribution as the null distribution.

4.3.2. Two-sample t-test

Conditions of a two-sample t-test:

- Observations are independent
- The population distribution is nearly normal.
  - However, *t* procedures are quite robust against non-normality. When sample sizes are equal and the distributions are similar, the normality condition is not as important. Therefore, *t*-tests are most robust when sample sizes are similar (and studies should be designed accordingly).
  - With large sample sizes, the sample sd will still accurately estimate the population sd and will therefore be appropriate even if the population distribution is non-normal.
- The assumption that the population distribution is nearly normal rules out outliers. However, *t* procedures are not robust against outliers because the mean and sd are not resistant to outliers.

Inference for small samples

Small samples require care; there are not enough observations to examine the distributions or moderate outliers. The power of these tests tends to be low, and the confidence intervals are large. However, if effect sizes are large, valid conclusions can still be drawn from these studies.

When sample are equal in size, ≥ 5, and have similar distributions, the *t* procedure is still robust against non-normality.

Degrees of Freedom

- The two-sample *t* statistic, $\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2$, does not follow a *t*-distribution, we can still approximate the null distribution using a *t*-distribution by approximating the degrees of freedom using the Satterhwaite equation. This is the approximation made by most statistical software. It is accurate when both sample sizes are 5 or larger.

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

- An alternate and more conservative (i.e., less sensitive) approach is to select the smaller *n* – 1. The *df* from this approach will always be less than that of the Satterhwaite approximation.
- Note, $n_1 + n_2 - 2$ apparently can also be used to calculate *df* but the other approximations are more conservative. This equation is used most commonly to calculate pooled variance, not *df*.

Pooled Variance

- When the variance of two population are exactly the same, they *will* follow exactly a *t* distribution. When variances are similar, they can be pooled. The equation for this weighs the variance of each sample depending on the sample size (i.e., *df*).

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

- Under circumstances of pooled variance, the t-statistic can be simplified to:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  *Note that in both of the above equations, we can calculate values for the z-dististribution by using the true population sd.

- However, because we do not actually know the population sd, we will not actually know if population variances are equal. There is no definitive way to know if the population variances are the same.
  - Differences may arise in our estimate (sample sd) just by chance.
  - The pooled *t* procedure is robust against unequal variance when sample sizes are similar, but can be sensitive to unequal variance when samples are different in size (unless sample sizes are large).
  - Pooled variance may increase your power if appropriate, but may increase the false positive rate if inappropriate.
- We can check for equal variance using the F-test (with many limitations; see “ F-test for equality of variance”). Far safer is to simply plot the distribution of your samples, with special atention to skewness and outliers.

4.3.3. F-test for equality of variance

The two most basic characteristics of a distribution are its centre (mean) and spread (variance). We have discussed in this book how to make inference on population means; we can also do the same with spread. However, procedures that test this are much more sensitive to deviations from assumptions than *t* procedures. The *Introduction to the Practice of Statistics 6e* states, "**Don't do it without expert advice**." (Section 7.3)

The F-test compares the spread of two Normal populations. However:

- The F-test is *extremely* sensitive to non-Normal distributions, which does not improve with larger samples. Therefore, the F-test is considered extremely non-robust.
  - Consequently, it is difficult in practice to tell if a significant *F*-value is evidence of unequal variance or simply evidence that the populations are not Normal. It is difficult to exclude this possibility, since we rarely know if the population is normal.
- While it was once common practice to test the equality of variance to decide on whether to pool variance, it is no longer widely accepted. It is better to check the distributions graphically, with special attention to skewness and outliers, and to use software-based two-sample *t* procedures that do not require equal variance.

The *F*-statistic

- The *F* statistic is given by: $F = \frac{s_1^2}{s_2^2}$ where s_1^2 and s_2^2 are the sample variances from independent samples of sizes n_1 and n_2 drawn from Normal populations.
- The *F*-statistic has the *F* distribution with n_1-1 and n_2-1 degrees of freedom when H_0: $\sigma_1 = \sigma_2$ is true
- The *F* distributions are a family of distributions with two parameters: The degrees of freedom for each sample variance in the numerator and denominator of the *F* statistic.
  - Interchanging the degrees of freedom changes the distribution, so the order matters.
  - The *F* distribution can be described using the notation *F*(*j*,*k*), where *j* is the *df* in the numerator, and *k* is the *df* in the denominator.
  - Because it is a ratio, it cannot take on negative values; therefore, the *F* distribution is right-skewed with a centre around 1. Values far from 1 in either direction are evidence against equal variance.
  - Because of skew and the fact that both *df* in the numerator and denominator are needed, *F* tables are awkward. The best way to evaluate an *F* test is using software.

4.3.4. Computing a two-sample t-test

Calculating a two-sample t-test (manually)

Our hypothesis would look like:

$\overline{x}_1 - \overline{x}_2 = \overline{x}_{diff}$

H_o $\overline{x}_{diff} = 0$; $| t_{df} | < t_{critical}$

H_A $\overline{x}_{diff} \neq 0$; $| t_{df} | > t_{critical}$

1. Ensure all conditions for a t-test are met.

2. Calculate the sem either assuming equal or unequal variance. Note that for a t-distribution you will use the sample sd:

   Equal variance: $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$; $SE_{\bar{x}_1-\bar{x}_2} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

   Unequal variance: $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}$
3. Calculate the degrees of freedom. To do this manually, simply use the *df* from the smaller *n*.
4. Find the critical *t*-statistic, $t_{critical}$
5. Calculate the t-statistic. Recall: $t_{df} = \frac{\bar{x}_{diff}-0}{SE_{\bar{x}_1-\bar{x}_2}}$
6. Compare the t-statistic to the $t_{critical}$. Use software to determine precise *p*-values. Make a decision on whether or not to reject the null hypothesis.

> Computing t-tests in R

Before beginning, ensure all conditions for a t-test are met. Transform the ‘x’ and ‘y’ variables into vectors.

`t.test()`

`x` - x_1: group to be tested (numeric vector)

`y` - x_2: only for 2-sample t-tests; second group to be tested (numeric vector)

`mu` - numeric. Specify the population mean for a one-sided t-test

`paired = FALSE` - Logical indicating if test is paired or not. If TRUE, ‘x’ and ‘y’ must be the same length

`alternative` - Must be one of:

- “t” for two-sided (default)
- “g” for greater (i.e., x_1 is greater than x_2), or
- “l” for lesser.

`var.equal = FALSE` - Logical argument indicating whether the variances of the two groups are equal. If TRUE, then pooled variance is used. If FALSE (default), the Welch approximation is used.

`conf.level = 0.95` - Confidence level

- Returns: t-statistic, df, p-value, confidence interval, and mean of x_1 and x_2. These values can be extracted using the ‘$’ operator.
- Note: This function calculates variance using n – 1

If not all data are available, you can stil compute a t-test given the mean, sd, and n of each group. Follow the steps above in **4.3.4. Computing a two-sample t-test** to calculate the t-statistic ‘t’. Then use the following code to calculate the critical t-statistic and p-value.

In order to compute the critical t-statistic (given at the 95% level):

```
tcrit <- qt( (1 - 0.05/2), df = n - 1)
```

```
tcrit
# To compute the p-value:
p <- 2 * pt(-abs(t), df = n - 1)
```

This is for a two-sided test (hence the 2*); t <- the previously calculated t-statistic.

4.3.5. Paired t-test

A paired t-test will ask if there is a difference between two groups, but the data are some how paired (ex., the same subject before- and after, or the same product in two different stores). A paired t-test is similar to a one-sample t-test; but instead of using the sample mean ($\overline{x}$) we use the mean difference between the pairs in the two data sets ($\overline{x}_{diff}$), and instead of using the population mean (μ), we assume the difference is = 0 (though other values can be used). The variance that we use is the variance between *differences between each observation*.

**See "> Computing t-tests in R" for instructions on how to use `t.test()`.

Calculating a paired t-test (manually)

Our hypothesis would look like:

$\overline{x}_1 - \overline{x}_2 = \overline{x}_{diff}$

H_o $\overline{x}_{diff} = 0$; $| t_{df} | < t_{critical}$

H_A $\overline{x}_{diff} \neq 0$; $| t_{df} | > t_{critical}$

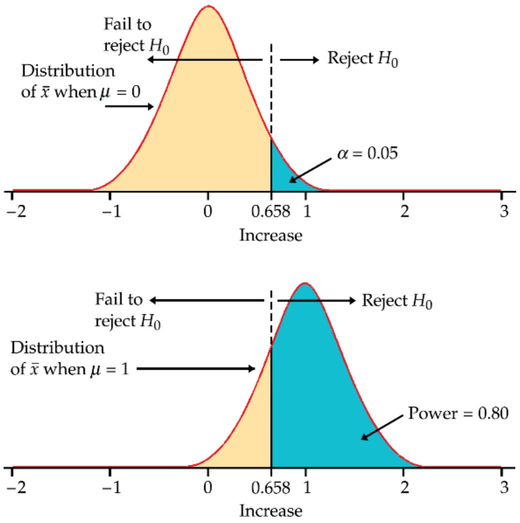
1. Ensure all conditions for a t-test are met.
2. Calculate the degrees of freedom. To do this manually, simply use the *df* from the smaller *n*.
3. Find the critical *t*-statistic, $t_{critical}$
4. Calculate the t-statistic. Recall: $t = \frac{\bar{x}_{diff} - 0}{s/\sqrt{n}}$
5. Make a decision on whether or not to reject the null hypothesis based on the observed t-statistic.

4.3.6. Power of a t-test

Power should be calculated prior to designing any study involving statistics.

Recall that power is the probability to detect a difference between a sample mean ($\overline{x}$) and the null hypothesis (or mean of a second sample) **assuming that there is a true difference**.

This is in contrast to α, which is the probability to detect a difference **assuming that there is no difference**.



There are three basic steps to calculate power:

1. State H_o and H_A, and the significance level. Power should be computed using the test that you will use to test the data.
2. Find the least extreme value of $\bar{x}$ required to reject H_o.
3. Calculate the probability of observing a value of $\bar{x}$ at least this extreme assuming H_A is true.

The methods of calculating power are similar for most procedures. Because the t-test is so important, we will go into depth about these methods.

Calculating the power of a t-test

1. Specify the following:
   a. An alternative value for μ_1 - μ_2 (i.e., your effect size).
      i. Your effect size may be based off of previous research / pilot studies OR based on an alternative value that you consider to be important / "clinically significant."
   b. The sample sizes, n_1 and n_2.
   c. The significance level, α
   d. An guesstimate of the population standard deviation, σ, (best estimated based on previous research).
      i. Since it's a complete shot in the dark, we will often assume the same sd for both groups (and thus use the pooled variance).
2. Find the degrees of freedom and the critical t_{crit} value required for rejection.
3. Calculate the noncentrality parameter (the distribution that our sample came from, assuming H_o is FALSE; i.e., the opposite of the null distribution):

$$\delta = \frac{|\bar{x}_{diff}|}{SE_{\bar{x}_1 - \bar{x}_2}} \quad ; \quad SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}$$

Assuming equal variance: $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$; $SE_{\bar{x}_1 - \bar{x}_2} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

4. Determine the probability of observing the noncentrality parameter:

$$Power = P(\delta \geq t_{crit})$$

*Often we will want to calculate the required n to achieve a desired level of power. This can be done by re-arranging the equation (see "> Computing the power of a two-sample t-test (quick)").

```
> Computing the power of a two-sample t-test (quick)
```

#Ensure that a t-test is appropriate. Use the function:

```
power.t.test(n, delta, sd, sig.level, power)
```

`n` - Integer. Number of observations per group (assumes equal groups). Default = NULL `delta` - Numeric. The effect size (hypothesized "true" difference in means). Default = NULL

`sd = 1` - Numeric. Estimated population standard deviation. Default = 1

`power` - Numeric. Power (1 - beta). Default = NULL

- This function uses the input of 3 out of the above 4 values to calculate the fourth

`sig.level` - Numeric. Alpha level. Default = 0.05)

`type` - String. Takes on one of:

  - `t` - two.sample (default)
  - `o` - one.sample
  - `p` - paired

`alternative` - String. Takes on one of:

  - `t` - two.sided (default)
  - `l` - less
  - `g` - greater

`strict` - Default = FALSE. If TRUE, the power will include the probability of rejection in the opposite direction of the true effect in the two-sided case. Without this the true power will be half the true significance level if the true difference is 0.

> Worked example: computing the power of a two-sample t-test in R (slow)

```
# This will use example numbers to demonstrate how to calculate power with a given sample size.
# 0. First ensure that the conditions of a t-test will be met.
# 1. Imagine an example with the following parameters under the null hypothesis: mu1 - mu2 = 0 ;  and an alternate hypothesis where we assume: mu1 - mu2 = 2
mu1 <- 1 # This is estimated
mu2 <- 3 # This is a value that we chose because it may be clinically significant
sigma1 <- 2 # This is a guess
sigma2 <- 3 # This is also a guess (for this example, we assume unequal variance)
n1 <- 40 # This is chosen based on what is feasible to our lab (for example)
n2 <- 40
alpha <- 0.05

#2. Determine the df and tcrit:
degreesf <- n1 - 1 # The software approximation of df requires knowledge of the sample variance, while the df = n1 + n2 - 2 is not as conservative.
tcrit <- qt(1 - alpha/2, df = degreesf) # This is for a two-sided test
tcrit

#3. Determine the noncentrality parameter:
```

```
delta <- abs(mu1 - mu2) / sqrt(
  sigma1^2/n1 +
    sigma2^2/n2
)
delta
```

#4. Determine the probability observing the noncentrality parameter (i.e., power):

```
power <- 1 - pt(d, df = degreesf, ncp = delta)
power
```

4.3.7. Confidence interval for a difference of means

Suppose you want to compare the efficacy of some treatment versus placebo. One way to test this is to construct a confidence interval. CIs are easily calculated in R using the `t.test()` function. Calculating CIs for two-samples is very similar to calculating CIs for one sample, except we use the difference in means and the sem calculated from both samples.

*For details on interpretation of CIs, see **3.3. Confidence intervals**.

Constructing a 2-sample CI (manually)

1. Find the sem. Recall: $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}$
2. Identify the *df*. This can be done using software. Manually, simply calculate the *df* of the smaller *n*.
3. Once *df* has been calculated, find the critical *t*-statistic for a 95% CI (using a t-table).
4. Calculate the CI:

$$95\%\ CI = (\bar{x}_1 - \bar{x}_2) \pm t_{df} * SE_{\bar{x}_1-\bar{x}_2}$$

5. Report the CI: “We are 95% confident that the mean improvement in the outcome due to the treatment is between *(lower margin)* to *(upper margin)*.”

5. Inference for Proportions

5.1. Inference for a single proportion

Single proportions follow the binomial distribution, if the population is much larger (i.e., 20-times larger) than the sample size *n*. Recall that the binomial distribution has two parameters, the probability of success *p* and the number of trials *n* ('trials' is synonymous with sample size). In this context, describes the spread of the observed proportion $\hat{p}$.

- For very small sample sizes, the binomial distribution doesn't work well because of its 'discreteness.'

- With larger sample sizes, the binomial distribution begins to resemble the Normal distribution (and by corollary the count of successes X and the proportion $\hat{p}$ are approximately normal). In general, we limit inferential procedures to larger sample sizes.
    - By extension, inference on proportions follows the same procedures as normally distributed data.

If the sample size is sufficiently large, the sampling distribution for $\hat{p}$ is approximately normal with mean and variance:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

Where p is the true unknown population proportion p.

5.1.2. Confidence intervals for a single proportion

When estimating the sd from the data, we simply replace p with $\hat{p}$. Thus, the SE and confidence intervals are given by:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$Confidence\ Interval = \hat{p}\ \pm z\ \times SE_{\hat{p}}$$

- In general, we want at least 15 successes and 15 failures to construct a confidence interval.
- For smaller sample sizes, it is generally better to use exact methods that use the binomial distribution.
- For intermediate cases, there is the "plus four" method.
    - This method simply involves adding 4 additional observations to your sample, 2 of which are successes.
    - You can then apply the "large sample" approach to construct confidence intervals. This adjustment is strange but apparently works well in practice (*Section 8.1 of Introduction to the Practice of Statistics 6e)*.

5.1.3. z-test for a single proportion

This is the same as the z-test for numerical data. To test the null hypothesis that $p = p_0$:

$$z = \frac{\hat{p} - p_0}{SE_{p_0}}$$

- The expected number of successes and failures (based off p_0) should be at least 10.
- The population should be ≥ 20-times the size of the sample.

Note that we use the standard error for p_0 rather than for $\hat{p}$. This is mentioned elsewhere, but we do this because we are testing the hypothesis under the assumption that $p = p_0$ (as opposed to $p = \hat{p}$). Under most circumstances (like with numerical data), we can't know SE_{p_0} so we estimate it from our observed data. When we are testing proportions we can know SE_{p_0}, since it is a calculated value.

* Note: we still use $SE_{\hat{p}}$ for confidence intervals. This is because confidence intervals provide an estimate of the parameter and do not directly test a hypothesis.

5.2. Inference for two proportions

Inference for two proportions follows the same procedures as inference for two means that are Normally distributed (these are discussed in section **4.3.2. Two-sample t-test**, except replace the *t* statistic with *z*). Whereas with single proportion procedures we are trying to estimate the true proportion, with two proportion procedures we are testing to see if they are different. By the addition rule for means and the sum rule for variances we get:

$$D = \hat{p}_1 - \hat{p}_2$$

$$\sigma_D^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\mu_D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2}$$

$$\sigma_D^2 = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2$$

5.2.1. Confidence intervals for two proportions

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$Confidence\ Interval = D \pm z \times SE_D$$

* Note: the "Plus four" estimate method can also be applied to confidence intervals for two proportions. Simply add 2 to *n* and 1 to *p* for both groups. Both sample sizes should be at least 5.

5.2.2. z-test for two proportions

The z-test follows the same procedures here as with numerical data, with one notable difference. Much like with single proportions, we use the standard error of the mean for our null hypothesis, which is that $p_1 = p_2$. If we assume that this is true, then both our samples came from the same population; instead of using individual proportions and the sum of their variances, we pool them together:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{p_1 + p_2}{2}$$

$$SE_{D_{pooled}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{D}{SE_{D_{pooled}}}$$

Where *D* is the difference between the two proportions

5.3. Relative risk

Relative risk (RR) is a useful way of comparing two proportions, especially if they are small. It is simply the ratio of the two proportions. A relative risk of 1 indicates that the two proportions are equal. Confidence intervals can also be calculated for relative risk, using similar procedures as above; however, the details are a bit more complicated. Software can handle these computations. Notably, confidence intervals for RR are not always symmetrical.

6. Inference for categorical data

6.1. Inference for contingency tables

Situation that might be tested using contingency tables include:

- Comparing ≥ 2 populations across a response variable with ≥ 2 categories
- Testing if 2 categorical variables are independent

Contingency tables provide summaries of count data for two different variables (in the below example: gender and yes/no response to some question). These tables are useful because we can use them to easily summarize marginal, joint, and conditional probabilities. In general, we assign the "independent variable" as the column variable, though we frequently do not know which variable is the independent variable.

	Male	Female	Total (margin)
Yes	**7** r 19.40% c 18.40% j 6.00%	**29** r 80.60% c 36.70% j 24.80%	**36** r 100.00% c - j 30.80%
No	**31** r 38.30% c 81.60% j 26.50%	**50** r 61.70% c 63.30% j 42.70%	**81** r 100.00% c - j 69.20%
Total (margin)	**38** r - c 100.00% j 32.50%	**79** - c 100.00% j 67.50%	r **117** r 100.00% c 100.00% j 100.00%

r - Conditional probability $P(c|r)$

c - Conditional probability $P(r|c)$

j - Joint probability $P(r \cap c)$

The 'totals' at the right and the bottom provide the marginal probabilities.

6.1.1. The chi-square (χ^2) test

The chi-square statistic tests the null hypothesis for a contingency table: there is no association between the row variable and the column variable. The alternate hypothesis is that there *is* an association. χ^2 may be large if there is a difference from expected counts in any direction. Therefore, we cannot describe H_a as being either one-sided or two-sided.

Calculating the X^2 statistic

To calculate the X^2, we first calculate the expected cell counts, which are the counts we would expect to see in each cell if the null hypothesis is true. This is simply expressed by:

$$expected\ cell\ count = \frac{row\ total\ \times column\ total}{n}$$

The expected values are simply the marginal proability of the response variable and apply this to all cells. In the example above, the marginal probability of '*yes*' is 30.8% across all genders; if there is no association, we expect that 30.8% of all men and 30.8% of all women to respond '*yes*.' Therefore, we expect that 30.8% of 38 = 11.7 men will respond '*yes*'.

The next step in calculating the χ^2 statistic is to square the differences between the expected and observed cell counts. Divide the differences by the expected cell count (this normalizes large vs. small cell counts). Finally, sum all of thes values.

$$X^2 = \sum \frac{(observed\ count\ - expected\ count)^2}{expected\ count}$$

X^2 provides a measure of how much teh observed cell counts diverge from the expected cell counts. To test if a difference is statistically significant, we compare the computed X^2 to the critical values on the χ^2 distribution.

Note: the test does not tell us anything about what kind of relationship is present, only that the observed data diverge from what is expected. You will have to report percentages to tell what is going on. Additionally, the test does not imply anything about causality.

6.1.2. The chi-square (χ^2) distribution

Recall that when a binomial distribution has a sufficient number of failures and successes (typically **≥ 15 each**), the distribution approximates the Normal distribution - this is the χ^2 distribution.

- Like the *t* distributions, the χ^2 distributions form a family of distributions described by a single parameter, *df*.

$$df = (r-1) \times (c-1)$$

- χ^2 distributions are right-skewed and only take on positive values (since they are squared).

6.1.3. The chi-square goodness of fit test

The χ^2 goodness of fit test is used to compare the sample distribution of a categorical variable from a population with a hypothesized distribution.

Procedurally, this test is similar to the χ^2 test, but there are a few differences:

- The data are summarized by a table with a single row and k number of cells.
- The null hypothesis that specifies the probability of each outcome *k*. We observe the data for *n* observations with *k* outcomes and compare this to the probability specified by the null hypothesis
    - We can use this to test "no association" by specifying a null hypothesis with a uniform distribution, but we can also test other distributions.
- Under the null hypothesis, X^2 has approximately the χ^2 distribution with $df = k - 1$.

7. Probability

We use probability to understand processes that are *apparently* random. A large number of processes that are seemingly random are often not; however, we can still model these processes as random for many purposes.

The probability of an outcome is the proportion of times this outcome would occur if the process occured an *infinite* number of times. The actual proportion of outcomes will almost always deviate from the "probability" when the process occurs a finite number of times.

This concept is important, and is known as the *Law of Large Numbers*: As the number of trials approaches infinity, the observed proportion, $\hat{p}$, will approach the true probability.

7.1. Notation / nomenclature for probability

Marginal probability - The probability of an outcome based on only a single variable.

Joint probability - The probability of an outcome based on the outcome of another variable (also the conditional probability).

Notes on notation:

$P(A)$	- "The probability of A"
$P(A^c)$	- "The probability of not A (i.e., the complement of A)"
$P(A \cup B)$	- "The probability of A *or* B"
$P(A \cap B)$	- "The probability of A *and* B occuring together"
$P(A\|B)$	- "The probability of A given that B has occured"

7.2. Rules for probability

7.2.1. The Addition Rule for disjoint outcomes

The sum of all possible *mutually exclusive* outcomes of a random process is 1.

$$P(A \cup B) = P(A) + P(B) = 1$$

Ex., Flipping a coin:

$$\begin{aligned} P(heads\ or\ tails) &= P(heads) + P(tails) \\ &= 0.5 +\ 0.5 \end{aligned}$$

7.2.2. The General Addition Rule for non-disjoint outcomes

When outcomes are *non-disjoint* (i.e., not mutually exclusive), the General Addition Rules applies. The General Addition Rule states that the probability that *either* outcome will occur is given by the sum of their marginal probabilities minus the probability that they occur together:

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

Ex., The probability of drawing a diamond or a face card from a deck of cards:

$$P(diamond\ or\ face) = P(diamond) + P(face) - P(diamond\ and\ face)$$
$$= \frac{13}{52} + \frac{12}{52} - \frac{3}{52}$$

7.2.3. The Multiplication Rule for independent processes

Two random processes are said to be independent if the outcome of one does not affect / provide knowledge of the outcome of the second. The probability that two outcomes from two independent random processes will occur together is given by their product:

$$P(A \cap B) = P(A) \times P(B)$$

Note how the addition rules describe the probability that *either* outcome will occur, while the Multiplication Rule describes the probability that *both* outcomes will occur.

7.2.4. Complements

For disjoint events: if the probability of *A* is equal to *x,* then the probability of *Not A* is equal to *1 - x*. This is called the complement. This is simply a re-statement of the Addition Rule.

$$1 = P(A) + P(A^c)$$

The complement does not need to be a single outcome, but we represent it as a single outcome (i.e., every outcome that wasn't *A*). In many cases, finding the complement may be easier and faster than finding the outcome directly.

Calculating marginal probabilities

This rule can be used to calculate marginal probabilities from joint probabilities. Most often, you will know the marginal probabilities for each outcome for a variable, since you collected that data. However, it is also possible to compute the marginal probabilities by adding up the joint probabilities (which is the same as the complement of *A*):

$$P(B) = P(A|B) + P(A^c|B)$$

Ex., The probability of being obese:

$$P(obese) = P(male\ and\ obese) + P(female\ and\ obese)$$

7.2.5. The General Multiplication Rule

Each conditional probability has two parts: the outcome of interest, and the condition. The condition is a previous outcome that may provide information about the outcome of interest.
The conditional probability of *A given B* is the same as the probability of *A and B* divided by the probability of *B* alone:

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

When re-arranged, this gives the General Multiplication Rule:

$$P(A \cap B) = P(A|B) \times P(B)$$

7.3. Bayes' Theorem

Bayes' Thereom is simply a generalization of conditional probabilities, but becomes useful when there are many possibly outcomes per process.

The conditional probability of outcome A_1 *given B* is equal to the conditional probability of *B given* A_1 multiplied by the marginal probability of A_1 divided by the sum of the conditional probabilities of *B given every possible outcome of* $A_{1\ldots k}$ multiplied by the marginal probability of everypossible outcome of $A_{1\ldots k}$.

$$P(A_1|B) = \frac{P(B|A_1) \times P(A_1)}{P(B|A_1) \times P(A_1) + \cdots P(B|A_k) \times P(A_k)}$$

The numerator describes the probability of having both outcomes. Recall from the General Multiplication Rule:

$$P(A \cap B) = P(B|A_1) \times P(A_1)$$

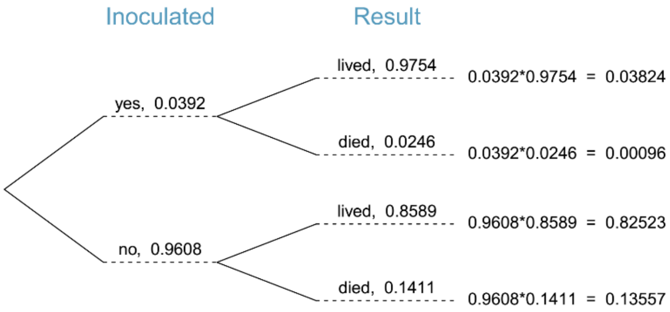
The demonimator is simply the marginal probability of *B*. We apply complements (a corollary of the Addition Rule)

$$P(A \cap B) = P(B|A_1) \times P(A_1)$$
$$P(A^c \cap B) = P(B|A_2) \times P(A_2) + \cdots P(B|A_k) \times P(A_k)$$

Recall that the sum of all joint probabilities A_x *given B* is equal to the marginal probability of *B*.

7.3.1. Tree diagrams

For less complex conditional probabilities, tree diagrams can be used in place of Bayes' Therorem.



8. Distributions (other than the Normal Distribution)

8.1. Continuous probability distributions

8.1.1.Beta distribution

The *Beta* distribution is a continuous distribution for the variable θ with two parameters (α, β). Note that these letters are entirely different from those used to represent type-I and type-II errors. Also note that β is a parameter of the *Beta* distribution. Its pdf is given by:

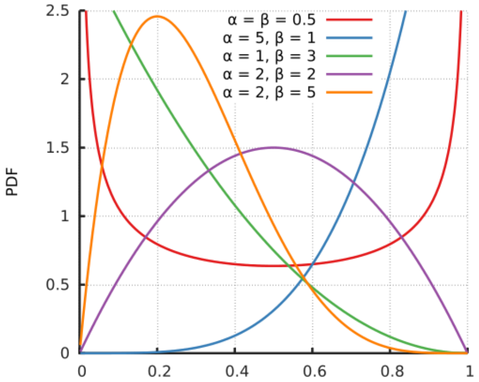
$$\int(\theta, \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$B(, \alpha, \beta)$ is the *Beta* function

This function is only defined for the range of: $\theta \in [0,1]$

The *Beta* distribution is commonly used to express prior knowledge about the probability of a binomial event in Bayesian statistics. By changing the parameters α and β, we can significantly vary how we express this prior knowledge. Since the first term of the pdf is just a constant, we can simplify it:

$$\int(\theta, \alpha, \beta) \ \propto \ \theta^{\alpha-1}(1-\theta)^{\beta-1}$$



Below are some examples:

$$P(\theta|\alpha = 0.5, \beta = 0.5) \ \propto \ \frac{1}{\theta^{0.5}(1-\theta^{0.5})}$$

- When either α or $\beta > 1$, the term exponent is negative; we get asymptotes that approach ∞ as θ approaches 0 or 1 (respectively).

$$P(\theta|\alpha = 2, \beta = 2) \ \propto \ \theta^{1}(1-\theta)^{1}$$

- Increasing α or β will skew the pdf right or left, respectively.
- The magnitude of α or β will determine the maximum height of the pdf.

You can get a better sense of how parameters affect the *Beta* distribution using the following simulator > Scripts\Statistics Demos\Bayes binomial posterior calculator.R

9. Likelihood Statistics

Likelihood statistics are an alternative way of quantifying evidence against a null hypothesis. Rather than the frequentist approach, which *assumes the null hypothesis and then seeks to disprove it*, likelihood approaches *compare how likely one hypothesis is, relative to another, given the observed data*.

9.1. Likelihood functions

A likelihood function gives the function of a parameter given the observed data. The simplest example is the binomial distribution, since it has a well defined variance when $\hat{p}$ is known. However, likelihood functions exist for other distributions

$$L(\theta) = \frac{n!}{x!\,(n-x)!} \times \theta^{x} \times (1-\theta)^{n-x}$$

Where θ is the true proportion of 'successes',

$L(\theta)$ is the likelihood function (i.e., estimated distribution) for the parameter θ,

and x is the number of successes.

We can calculate the likelihood of two opposing hypotheses (we can still call these *H_o* and *H₁*) given the data quite simply:

$$Likelihood\ ratio_{H_1} = \frac{Maximum\ likelihood_{H_1}}{Maximum\ likelihood_{H_0}}$$

- In general, likelihood ratios of **8** and **32** indicate moderate evidence and strong evidence, respectively.
- Likelihoods are relative evidence, so we can actually make statements about how strongly we believe in an alternative hypothesis.
- Note: Because likelihood ratios are relative and only use two points, they can be misleading if both hypotheses fail to capture the data.

10. Bayesian Statistics

10.1. Basic concepts in Bayesian Statistics

Prior odds

Posterior probabilities are calculated using a *prior*. This is both a strength and a weakness, as it allows/requires you to input subjective beliefs.

Posterior odds

A *p-value* gives $P(data|H_0)$.

The *posterior* gives $P(H_0|data)$. In this way, you can actually estimate the probability that the hypothesis is true.

$$Posterior\ odds = Likelihood\ ratio\ \times\ Prior\ odds$$

$$\frac{P(H_1|data)}{P(H_0|data)} = \qquad \frac{P(data|H_1)}{P(data|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

Bayes factor

With a strong prior, the posterior is less influenced by data. Likewise, the posterior will be more influenced by strong data. The ratio of posterior odds to the prior odds is the *Bayes Factor*, which is a representation of 'how much we have changed our beliefs.' A Bayes Factor > 1 indicates that the hypothesis is more likely than before, given the data.

$$Bayes\ factor = \frac{Posterior\ odds}{Prior\ odds}$$

10.1.1. Binomial Bayesian Inference

In frequentist statistics, we assume that there is a *true* population parameter that we are trying our best to estimate. In Bayesian statistics, we assume that parameters are *random variables* that have some degree of uncertainty. These can be quantified by their probability distribution.

For example, for a binomial distribution, we calculate the prior using the *Beta* distribution. The *Beta* distribution has two parameters, α and β. Note that these letters are entirely different from those used to represent type-I and type-II errors. Also note that β is a parameter of the *Beta* distribution. The pdf for the *Beta* distribution is given by:

$$\int(x, \alpha, \beta) = \frac{1}{B(, \alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Where $B(, \alpha, \beta)$ is the *Beta* function,

α is the number of successes,

and β is the number of failures

See **8.1.1. Beta distribution** for more on this distribution (I still don't understand the math, though).

Prior odds under the *Beta* distribution

The ratio of α and β determine the skew of the distribution, while the magnitude of the numbers determines the spread. Higher values of α and β convey greater "certainty" of the prior odds. Values of α = 1 and β = 1 are "uninformative", and (supposedly) do not contribute to the posterior odds.

Posterior odds under the *Beta* distribution

The posterior odds are calculated the general same way ($Posterior\ odds = Likelihood\ ratio\ \times\ Prior\ odds$). Under the beta distribution, the posterior odds are given by:

$$B(\alpha_{post}, \beta_{post})$$

Where:

$$\alpha_{post} = \alpha_{prior} + \alpha_{data}$$

$$\beta_{post} = \beta_{prior} + \beta_{data}$$

This distribution has a mean:

$$mean\ of\ B(\alpha, \beta) = \frac{\alpha}{\alpha + \beta}$$

Calculator: > Scripts\Statistics Demos\Bayes binomial posterior calculator.R

Bayesian estimation

The posterior distribution summarizes our belief about the expected outcome. We can also calculate a *credible interval*, which is like a Bayesian version of a confidence interval (but has a different interpretation). A confidence interval assumes that the parameter is fixed and that the data will vary; a credible interval assumes that the data are fixed and the parameter will vary. A 95% credible interval represents our belief that there is a 95% probability that the parameter falls within an interval. The 95% credible interval is simply the area under the posterior distribution that falls below 0.025 and above 0.975.

In R this is calculated easily with the `qbeta()` function. See > Scripts\Statistics Demos\Bayes binomial credible interval calculator.R

When an 'uninformative' prior is used, the credible interval and confidence interval are numerically the same, but their interpretations still differ.

11. Other concepts

11.1. Multiple comparisons

Performing multiple statistical tests without the proper precautions will result in an increased false positive rate. However, applying multiple comparisons corrections will also increase the rate of false negatives. Which corrections should be applied, if any, in a given context, will depend on your tolerance for false positives versus false negatives. For this reason, in exploratory data analysis, we often don't apply any corrections, even if we're measuring 50 different things. Whatever the case, the corrections that will be applied should always be determined *a priori*.

Instances when we are concerned with multiple comparisons

- When we are comparing >2 groups using a single statistical test.
    - Certain statistical tests that perform multiple-comparisons have correction procedures built into them.
- When we are comparing >2 variables between 2 groups, often using multiple statistical tests.
- When performing interim analyses or testing with "optional stopping."

Instances where we may apply looser (or no) multiple comparison corrections

- As mentioned above, exploratory data analysis often avoids multiple comparisons because it is not hypothesis driven and is more concerned with discovery.
- A statistical test may perform many comparisons and provide statistics for those comparisons, but we may only be interested in a few of them.
    - For example, a 2x2x2 ANOVA compares 3 main effects, 3 two-way interactions, and 1 three-way interaction. However, if we are only interested in the 4 interactions, we may decide that this is the "family" of tests that we are interested in and only correct for those.

11.1.1. Family-wise error procedures

11.1.1.1. Bonferroni procedure

The simplest method of controlling for multiple comparisons.

$$\alpha_{bonferroni} = \frac{\alpha}{number\ of\ tests} \quad or \quad p_{bonferroni} = p \times number\ of\ tests$$

This method is very conservative and leads to false negatives, but also removes pretty much all false positives.

11.1.1.2. Holm correction

This correction is supposedly more efficient than the Bonferroni, but generally yields the same results. The procedure involves taking all *p-values* that you've obtained, and ranking them from highest to lowest. Then multiple each *p-value* by its rank.

For example:

Unadjusted *p*	Rank	Adjusted *p*
0.015	3	0.045
0.02	2	0.04
0.3	1	0.3

In general, the Holm correction produces similar results to the Bonferroni correction (from simulations), but is slightly less conservative. The procedure is less punishing than Bonferroni in instances where you have many *p-values* that are nominally significant.

11.1.2. False discovery rate procedures

Some recent approaches are concerned with controling the false discovery rate (FDR), rather than the family-wise error rate. The most famous of these is the Benjamini-Hochberg FDR procedure. These procedures are most appropriate when dealing with large amounts of data where family-wise error-rates may be too conservative.

11.1.2.1. Benjamini-Hochberg FDR

unfinished.

11.1.3. Hierarchical step-down testing

This method is most appropriate for clinical trials or surveys where there is a well-defined primary end-point and supporting secondary end-points, though I'm not sure how common this is. I don't see this being used for pretty much anything else, *but maybe I'm wrong*.

Genovese et. al (2016) and Taylor et. al (2017) use this particular method to test the efficacy of baricitinib in clinical trials for rheumatoid arhritis. Their primary end-point was ACR20 followed by a hierarchy of supporting secondary end-points. These were determine *a priori* and when an outcome was not found to be significant, all of the outcomes below it on the hierarchy were not tested. In this way, each additional outcome was used as evidence to support the primary outcome rather than evidence that could be used in place of the primary outcome. This technique is clever and seems reasonable overall, and requires essentially no math.

13.3.1. Optional Stopping Corrections

A common problem in research is that: Researchers will collect data and analyze them, and if the results are not significant, they will recruit more subjects and re-analyze the data. While this may seem reasonable, repeatedly looking at the same data will inflate the false positive rate (since researchers will continue until $p < 0.05$).

Mathematically, optional stopping will *always* result in a significant result *eventually*. Sequential analysis techniques exist to try to control the type-I error rate when re-analyzing data. These generally work by lowering the α level each time that you look at the data (similar to Bonferroni correction) so that the *cumulative alpha level* remains the same (i.e., 0.05).

11.1.3.1. Pocock boundary

The Pocock boundary similar to a simple Bonferroni correction where we divide the α level by the number of analyses, but it is apparently more efficient (and has slightly more lax α levels compared to Bonferroni). It is necessary that the number of interim analyses are fixed and determind before the first analysis.

You can look up a table of *p-values* to use if you ever go down this route.

11.2. Missing Data

11.2.1. Types of missing data

Missing completely at random (MCAR)

- As it implies, values are missing completely at random. I.e., missingnes is not related to any property of the observation, any property of the variable, the value of other variables, or the missingness of other variables. That is to say, missingness is due only to noise.
  - This almost never happens.
- If you drop these cases you will not introduce bias, but the efficiency of your estimate will be lower (likely by increasing variance from lost data). You can model these variables using a random model to improve efficiency.

Missing at random (MAR)

- The occurence of missing values is random but associated with other variables that have been measured. That is to say, missingness can be modeled.

- In a way, MAR is a conditional form of MCAR. Once MAR has been modeled, it is essentially MCAR in that only noise remains.
- If you drop these observations you will introduce bias (selection bias) and reduce efficiency. However, if you model this you can eliminate bias and improve efficiency.

Not missing at random (NMAR or MNAR)

- The occurence of missing values is associated with variables that are unknown or unmeasured.
- You can't do anything about this. Unfortunately, it is not even possible to distinguish NMAR from MAR *post hoc* because the data required to test this are missing - by definition.
  - You can attempt to re-gather data on missing entries and test if these entries differ significantly from non-missing entries (ex., following up with survey non-respondents).

11.2.2. Ways of dealing with missing data

This section deals primarily with how to handle missingness for data MAR. MCAR is dealt with essentially the same way (but simpler), and NMAR data cannot be modeled. Long-story short: multiple imputation is generally the best way to estimate missing values to improve efficiency and reduce bias.

1. Listwise deletion: drop all observations with any missing data
   - Default in most statistical packages, including R
   - Results in maximum loss of information

2. Mean/median/multivariate imputation: replace missing data with the observed mean/median (or draws from the multivatiate distribution of the variable)
   - Understates the variability in the imputed variable, leading to an underestimation of the variance and standard errors
   - Makes no (or a limited) attempt to recover associations between the variables
   - Does not introduce bias and improves efficiency (but as stated above, it may underestimate standard errors)

3. Regression-based imputation: predicts missing values based on linear/logistic regression for all other available variables.
   - Not a bad method
   - Note: it is okay to use the dependent variable here to predict the missing value.
   - Understates the variability of the imputed variable, as the quality of the estimate for the missing value is not included when using it to estimate the dependent variable.
     - I.e., the imputed value is an uncertain estimate, but this uncertainty is not used when making further estimates in future models.
   - This will improve efficiency but will likely lead to an underestimation of standard errors in future models.

4. Interpolation of panel data: for time-series/panel data, replaces missing observations using either the last observation or the mean of the other panels.
   - Not terrible - frequently used in clinical trials for drop-outs/missed follow-ups
   - There is some evidence that this technique introduces bias and overconfidence of estimates
     - There is no way to know *a priori* if this will improve or reduce the quality of your estimate
   - Only works for panel data

5. Multiple imputation: See below.

11.2.3. Multiple Imputation

Multiple imputation uses multiple models (ex., regression) to predict missing observations from non-missing observations. Each model has slight variations, so the variation in the predictions generated by the different models provides a measure of uncertainty in our estimate.

A general work flow for a regression model is as follows:

1. The variable *x* can be given as a function of noise and other variables: $x \sim \alpha_0 + \alpha_1 z + \alpha_2 y$. Each α follows a distribution (recall: for a linear regression model to apply this should be a normal distribution)
2. We draw *m* number of each of α_0, α_1, and α_2 for each missing entry and create *m* number of models with these draws to predict *m* number of values of *x*.
   - This produces *m* number of data sets that differ in their predicted values of *x*.
3. We plug each of the *m* predictions to create *m* estimates of β_0, β_1, and β_2 for *y*.
4. Finally, we pool each estimate of β_0, β_1, and β_2, which is done simply by taking the mean and variance across all the estimates.

$$Var_\beta = W + \left(1 + \frac{1}{m}\right) B$$

W is an estimate of the within-imputation variation (i.e., the mean of the within-imputation variances).

B is an estimate of the between-imputation variation (i.e., the mean of the variance between different estimates of β.

$$W = \sum\nolimits_{m=1}^{M} \frac{s_m^2}{m} \qquad ; \qquad B = \sum_{m=1}^{M} \frac{(\hat{\beta}_m - \hat{\beta})^2}{m-1}$$

The $\boldsymbol{\left(1 + \frac{1}{m}\right) B}$ term is the inflation in the standard errors of $\hat{\beta}$ that we use to correct for imputation (i.e., the additional noise that we introduce by imputing estimates for *x*). From this term we see that increasing the number of models (*m*) increases the confidence of our estimate by decreasing the amount of noise that we introduce.

In practice, you'll never have to do this by hand. There are several statistical packages that will do the computations for you.

11.2.4. MICE: Multiple imputation through chained equations

MICE uses an iterative boot-strapping approach to create multiple imputations.

The process for MICE is as follows:

1. Discard all observations for which all variables are missing.
2. For all missing entries, fill in the missing data with random draws from the observed values.
3. Move through the columns of variables and perform single-variable imputation using the method of choice, replacing the original (quasi-random) replacements with the fitted replacements.
    - The first variable will be imputed using the quasi-random entries drawn in step 2.
    - As you progress through the variables, missing entries will be imputed using the imputed values of the preceding variables.
4. Repeat step 3 for *o* number of iterations.
5. Repeat steps 1-4 *m* number of times to create *m* number of imputed data sets.
6. Make estimates using each of the *m* data sets. Pool these estimates together (as described above).

The critical steps for MICE are 3 and 4.

- There are many ways to impute the variables:
    - Regression (linear, logistic, or multinomial)
        - Pick the maximum likelihood $\hat{y}$ OR sample from a distribution of *y*'s.
    - Predictive mean matching (PMM) - the default for MICE for continuous variables in R.
        - Creates a predicted value for missing entries from a regression model. It then picks *q* number of *real entries* that have the closest predicted value (i.e., closest Euclidean distance) and randomly chooses one of those impute.
        - I.e., the missing entry takes on the value of a real entry that is close to value predicted by regression, but there is an element of randomness involved.
- You must choose how many data sets (*m*) to use (5 a good minimum) and how many iterations (*o*) of imputation should be performed for each *m*. For PMM, you must choose how many neighbours (*q*) the prediction should be drawn from (default = 3).

11.2.4.1. Multiple imputation on data that are NMAR

Contrary to popular belief, `mice` is not restricted to data that are MAR. While it is true that imputation techniques commonly assume MAR, the theory of multiple imputation is generalizeable to NMAR. It is true that, when data are NMAR, the model fitted to complete cases will be incorrect for the incomplete cases; however, multiple imputation methods are still better than other alternatives[1].