Week 9 - Introduction to Inferential Statistics

Reading: 351-357

Introducting Null Hypothesis Significance Testing (NHST)

- A hypothesis is a statement or prediction about a relationship or effect that can be tested through scientific research. It is an educated guess or a proposed explanation for a phenomenon based on existing knowledge and observation. In scientific research, a hypothesis is usually tested by conducting an experiment or collecting data.
- The goal of a psychological experiment is often to test the **null hypothesis** and see if there is enough evidence to reject it and support the alternative hypothesis. A **null hypothesis** is a statement that there is no significant difference or relationship between two variables or groups. The null hypothesis is the default (albeit 'fictional') assumption that is made when there is no evidence to suggest otherwise. In other words, it is the hypothesis that any observed effect is due to chance.
- Null Hypothesis Significance Testing (NHST) is a statistical method used to determine whether there is enough evidence to reject the null hypothesis in favor of an alternative hypothesis. Some issues with NHT are:
- Because the null hypothesis is 'imaginary', it may not be correct
- NHT assumes a normally distributed dataset, which is not always the case - Just because a null hypothesis is 'rejected' does not necessarily mean the alternate hypothesis is correct
- These issues will be expanded on in later sections.

A null hypothesis is an example of a *statistical* hypothesis that describes the pattern to be investigated. In NHT, this includes two components:

- The *null* hypothesis (H_O), which states there is no difference between our comparisons
- The *alternative* hypothesis (H_A), which states there *is* a difference

The goal of NHST is to evaluate the evidence for or against the research hypothesis by comparing it to the null hypothesis. This is done by collecting data and applying statistical tests to determine the probability that the results observed are due to chance (if the null hypothesis is true). If the probability of obtaining the results by chance is low enough (usually below a certain threshold, such as p < .05), the null hypothesis is rejected and the research hypothesis is supported.

A *research* hypothesis describes the phenomenon of interest (e.g., does reduced physical activity increase depression?) This is a statement or prediction about a relationship that can be tested through research. It is an educated guess for a phenomenon based on existing knowledge and observation. Research hypotheses are usually formulated before data is collected and analyzed.

> Research hypotheses are tested by statistical hypotheses. The latter are designed to evaluate the evidence of the research hypothesis by comparing it to the null hypothesis. NHST is a common statistical approach within Psychology that helps evaluate the tenability of a research hypothesis based on a sample of data.

Testing hypotheses through statistical analyses

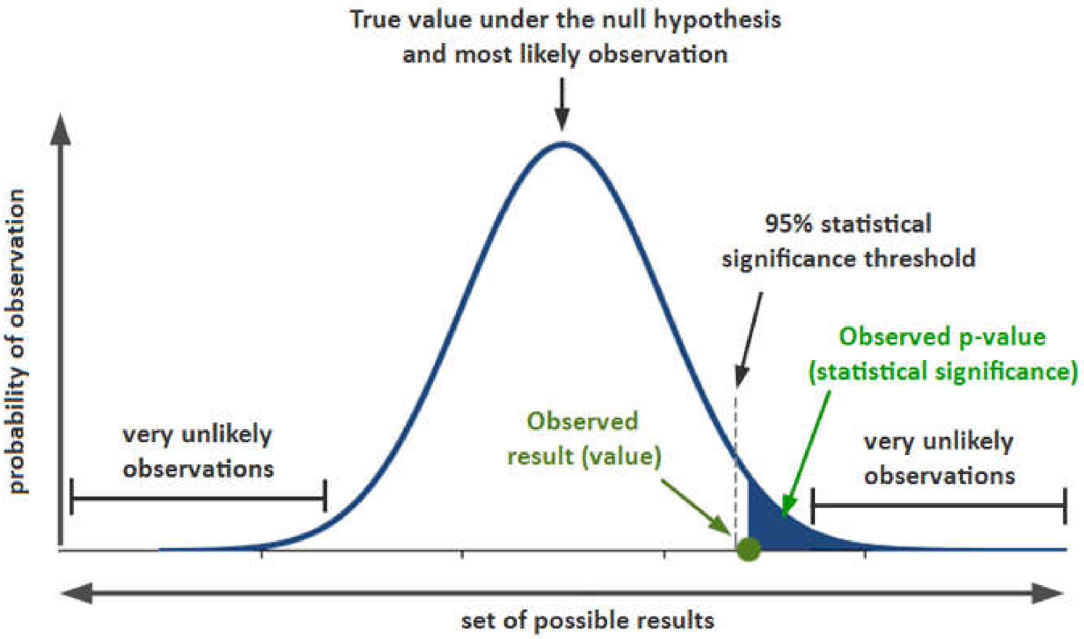
Null hypotheses are tested through statistical analyses by comparing the observed data to what would be expected if the null hypothesis were true. Statistical analyses involves estimating and comparing across variable *parameters* (values of interest). This can include *Means* and *Medians* as *point* estimates, as well as standard deviations (SD) and variance as *range* estimates.

We typically compare across *samples*, which are assumed to represent the overall *population*. In practice, the sample may not be a ‘true’ representation of the population - there will always be a degree of *sampling error* (random variance across individuals that may not be present in the actual population)

Statistics involves the quantitative description and interpretation of sample parameters

---

- Recall that the goal of NHST is to determine whether we can *retain/accept* or *reject* H_0 (the null hypothesis that there is no difference). Note that we *cannot* make claims about H_A (the alternative hypotheseis) using NHST!
- When performing NHSTs, we aim to determine the probability of observing the results obtained from the sample data if H_0 were true. This probability is known as the *p*-value. If the *p*-value is less than a pre-determined level of significance, usually set at 0.05, we conclude that the data is statistically significant and that the results are unlikely to have occurred by chance.
- If you see $p < .05$, you can (usually) reject the null hypothesis as this means that the observed data falls outside 95% of what would be expected assuming the sample is representative of the true population.


True value under the null hypothesis and most likely observation
probability of observation
95% statistical significance threshold
Observed p-value (statistical significance)
very unlikely observations
Observed result (value)
very unlikely observations
set of possible results

The probability of our result being observed *if* the null hypothesis was true is less than 5%, in which case we would conclude that the data statistically significant ($p < .05$).

A note on statistical significance

> Just because you see statistically different effects do not *necessarily* mean that the latter are practically important!

- There are several reasons why statistically significant effects may not be practically important:

1. *Small effect size*: Even though an effect may be statistically significant, it may be very small and imply little practical impact. It is important to consider the magnitude of the effect, not just its statistical significance.
2. *Large sample size*: With large enough samples, even small differences between groups can be statistically significant, but they may not be practically important.
3. *Lack of practical relevance*: Even if an effect is statistically significant, it may not be relevant to the real-world problem or question being studied.
4. *Multiple testing*: When performing multiple tests on the same data set, the chances of finding a statistically significant result by chance increases, which can lead to false positives (Type-1 error).

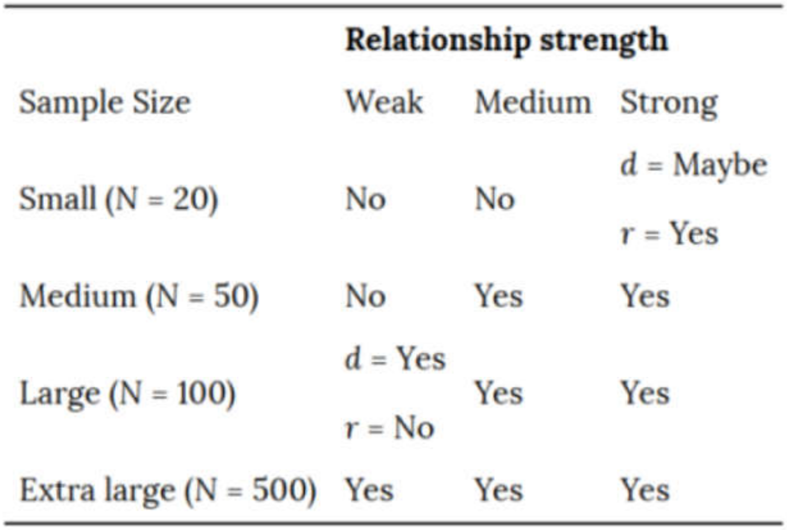
It's important to consider both statistical significance and practical importance when evaluating research findings. This can be done by looking at effect size, confidence intervals, and the relevance of the results to the real-world problem or question being studied.

Sample Sizes

- An important consideration when undertaking NHST is **sample size**, as this affects the ability to detect a difference or relationship between variables or groups. The larger the sample size, the greater the power of the test, which means that it is more likely to detect a real difference or relationship if one exists. However, with large sample sizes, even small differences between groups can be statistically significant, but they may not be practically important as noted above.
- Sample size must be chosen carefully, based on the research question, the expected *effect size* (a measure of the magnitude of the difference or relationship between variables or groups), and the desired level of *power* (the probability of correctly rejecting the null hypothesis when the latter is false).

Some rules of thumb for effective sample sizes…
$n \leq 20$: Only large/strong effects can be detected.
$20 < n < 50$: Medium-to-strong effects can be detected.
$n \geq 50$: Weak-to-medium effects (but not *correlations*) may be detected.

Table 13.1 How Relationship Strength and Sample Size Combine to Determine Whether a Result Is Statistically Significant

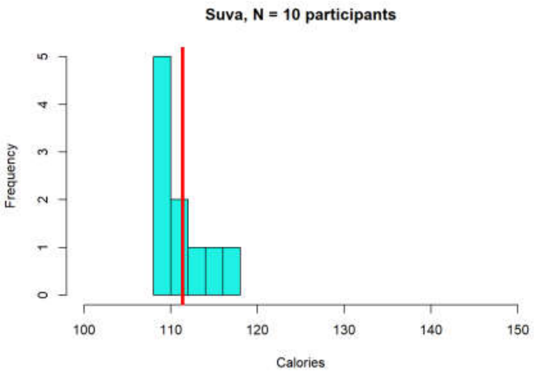


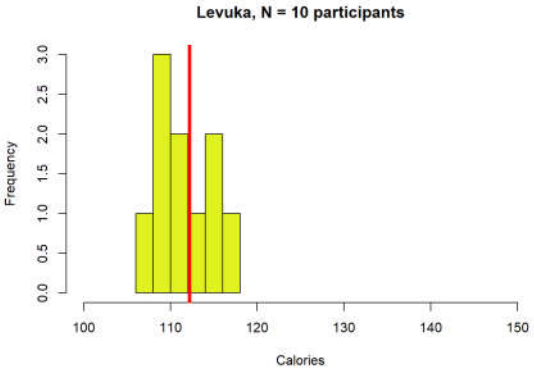
	Relationship strength		
Sample Size	Weak	Medium	Strong
Small (N = 20)	No	No	*d* = Maybe *r* = Yes
Medium (N = 50)	No	Yes	Yes
Large (N = 100)	*d* = Yes *r* = No	Yes	Yes
Extra large (N = 500)	Yes	Yes	Yes

Example of a hypothetical study using NHST

Suppose you are a nutritionist interested in obesity in Fiji. You want to test whether typical snacks in Suva are calorically different from typical snacks in Levuka.

You have the raw calorie data of 10 Suva residents (108, 117, 113, 111, 109, 116, 110, 110, 109, and 111 calories) and the raw calorie data of 10 Levuka residents (110, 112, 114, 112, 110, 109, 115, 118, 115, and 107 calories). We can explore this data by constructing two histograms.



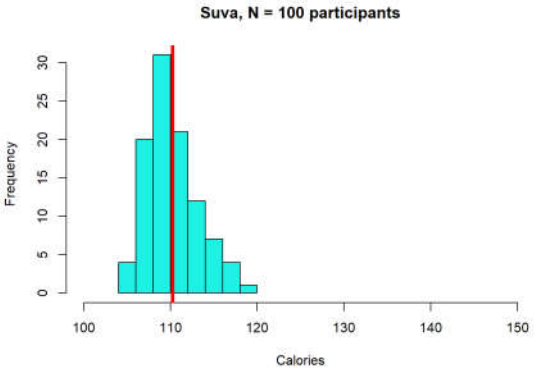


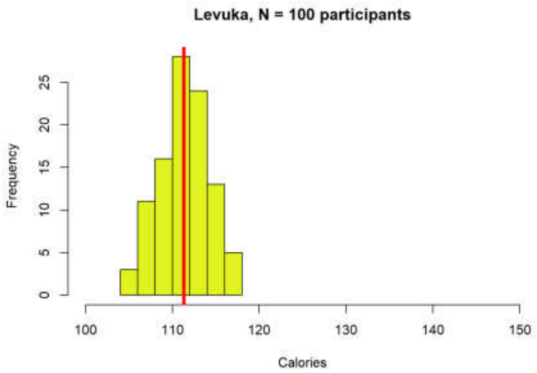
From the raw data, we can estimate the mean and standard deviations from Suva ($M = 111.4, SD = 3$) and Levuka ($M = 112.2, SD = 3.3$). There seems to be a marginal difference of 1~2 calories, which is likely to be neither significant nor important…

A two-sample *t*-test can confirm that the null hypothesis (of there being no difference between Suva and Levuka calories) is statistically supported (*p* = 0.581, which is larger then the typical threshold of *p* = .05)

We decide to increase our sample size to 100 Suva and 100 Levuka residents (raw data not shown to conserve space).

Now our histograms might look like the following:





- The difference between calories in Suva ($M = 110.3, SD = 3$) and Levuka ($M = 111.3, SD = 2.8$) remains virtually unchanged when looking at means and SDs.
- Yet running a t-test now produces $p = 0.009$, which is clearly smaller then $p < .05$, meaning we can claim that the difference is statistically significant.

> A difference of 1~2 calories was too weak to be detected when sample sizes were $n = 10$. Yet when the same parameters were noted for samples containing $n = 100$ each, our test reached significance.

It is common (and incorrect) to keep collecting data the significance threshold is passed. Known as *p-hacking*, this is an ill-informed practice within *frequentist* statistics. Doing so in isolation can generate false positives (Type-1 error) and highlight effects that may not be practically meaningful.

Some common NHST tests

An overview of some common statistical tests used in Psychology are provided below:

- **One-sample t-test**
  - You have **one group/sample** ($k = 1$) and you know the population parameter
  - The average grade for a Research Methods class of 100 students is 70 ($M = 70$). The average grade for all RM classes is 69 ($\mu = 69$).
  - H_0: The grade average for the current RM class is statistically equivalent to the population grade average ($M = \mu$).
  - During analysis, we estimate test statistics from the differences observed. The probability of observing the test statistics given our null hypothesis informs us whether (any) difference is significant.
- **Two-sample t-test**
  - You have two groups ($k = 2$) with two parameter estimates (e.g., Group 1 = μ_1; Group 2 = μ_2).
  - H_0: The two groups are statistically equivalent ($\mu_1 = \mu_2$). ++ When we measure the same group twice, we run *pairwise/repeated* tests. For example, if we measure depression scores

before and after lockdowns and note whether there is a difference between the two times (where $H_0 : Time1_{Depression} = Time2_{Depression}$).
++ When we measure two seperate groups of participants, we run *independent* tests. For example, if we measure average calories consumed by Levuka and Suva residents (e.g., $H_0 : Levuka_{Calories} = Suva_{Calories}$)

- **Analysis of Variance (ANOVA)**
    - You have **more than two groups** ($k \geq 3$).
    - You want to find out whether the groups being tested are different from one another ($\mu 1 = \mu 2 = \mu 3$).
    - If an ANOVA is significant (meaning H_0 is rejected), we can run two-sample (post-hoc) tests to find out which group differences are significant.
    - For example, if we want to test that the mean (μ) calories between Levuka, Suva and Labasa are statistically equivalent, we test the null hypothesis of $\mu 1 = \mu 2 = \mu 3$. If a significant effect is found, then we can test whether $\mu 1 = \mu 2$, $\mu 1 = \mu 3$ and $\mu 2 = \mu 3$

Similar to t-tests, ANOVAs produce a test statistic (called the F-ratio) which can be used to estimate a p-value. The latter tells us whether observed variances are statistically different between groups.

There are varieties of ANOVAs for independent, repeated and mixed contrasts. These include:

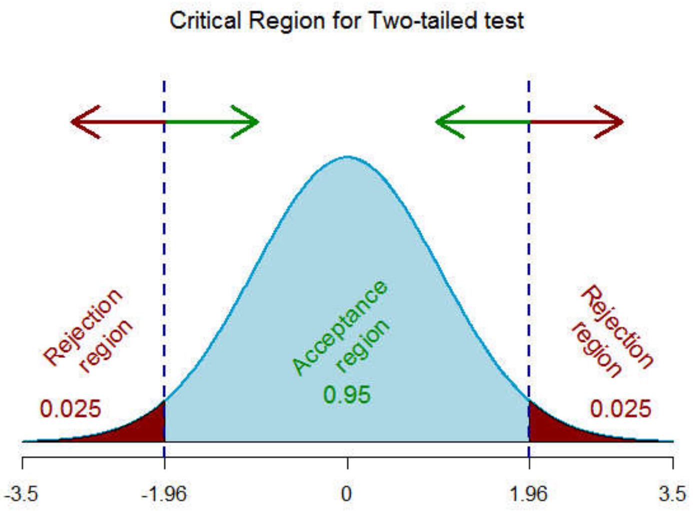
- **One-way ANOVA**
    - You have $k \geq 3$ levels along a *single* independent variable and you want to test whether there is an overall difference
    - Are the calorie densities of kokoda between Levuka, Suva and Tavua are statistically equivalent ($p \geq .05$) or different ($p < .05$)?
- **Repeated ANOVA**
    - You have $k \geq 3$ measurements of the same group. For example, do academic performances varies across summer, fall and winter semesters?
- **Independent/one-way ANOVA**
    - You have $k \geq 3$ groups that are independent of each other
    - Are the academic performances across three schools in Suva statistically different?
- **Mixed ANOVA**
    - You have $k \geq 3$ independent groups that you sample across $k \geq 2$ times. For example, are the academic performances across three schools in Suva statistically different between summer and winter sessions?
- **Factorial ANOVA**
    - You have (at least) 2 independent variables. For example, is school location and the amount of time spent on social media predictive of academic performance?

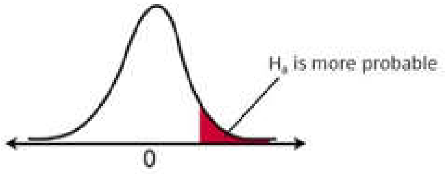
Across tests, we want to know whether the difference between means is statistically significant ($p < .05$). Statistical significance refers to the likelihood that the results of a study are due to chance rather than a real difference or relationship between variables or groups. It is typically determined by a *p*-value, which is the probability of observing the results obtained from the sample data if the null hypothesis were true.

Two-sided vs One-sided tests

Significance thresholds are further informed by whether you decide to run *two-sided* or *one-sided* tests.

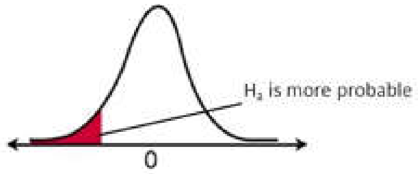
- A two-sided test checks for a difference or relationship **in either direction**. For example, a two-sided test for the mean of a population would check if the sample mean is significantly different from the population mean **in either direction, either higher or lower**.
- A one-sided test checks for a difference or relationship **in one specific direction**. For example, a one-sided test for the mean of a population would check if the sample mean is significantly different from the population mean in one direction, **either higher or lower, but not both**.





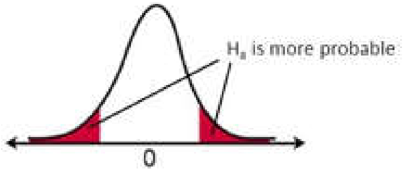
Right-tail test

H_a: $\mu >$ value



Left-tail test

H_a: $\mu <$ value



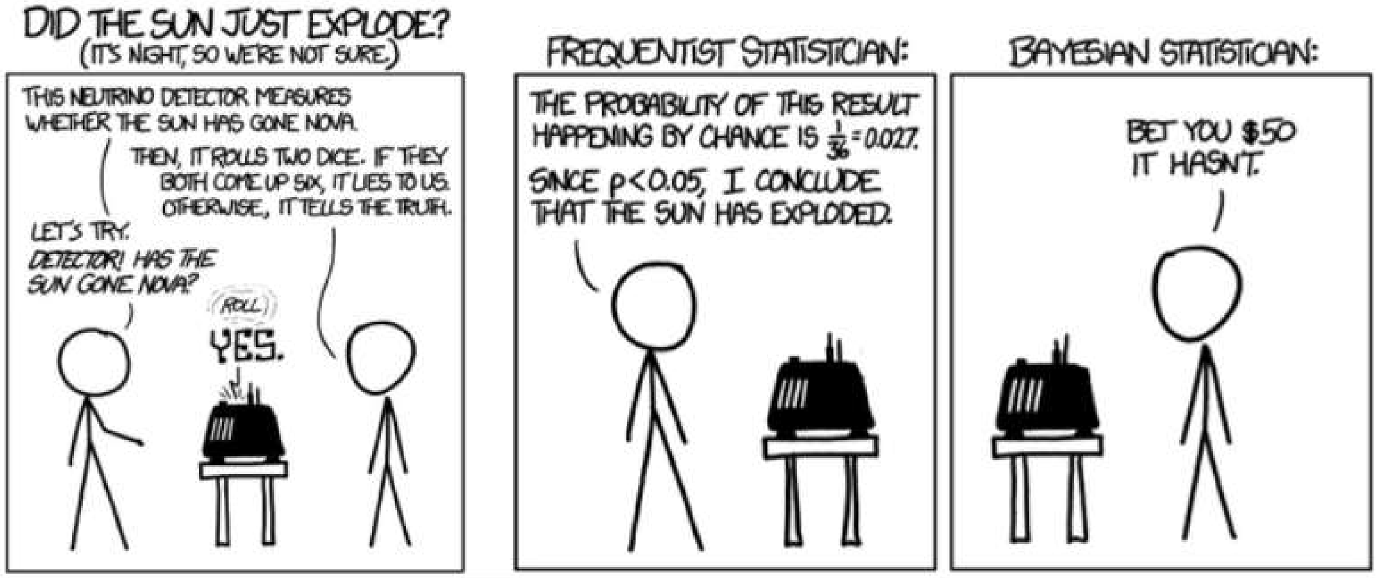
Two-tail test

H_a: $\mu \neq$ value

Conclusion

- NHSTs are useful for rejecting or retaining null hypotheses, but they cannot tell us whether the *alternative* claims are true.

- p-values tell us whether an observed difference is statistically different ('significant') or not, but is silent as to whether the difference is practically meaningful.
- Effect size estimates help determine whether an observed difference is practically important. For example, Cohen's difference score (d_{Cohen}) is a standardized estimate which tells us hoe many SD's two groups differ by (most commonly used for t-tests, assuming some assumptions are met) (https://toptipbio.com/cohens-d/)
- Restricted to binary claims (either an effect is significant or it is not). Not informative for estimating continuous interpretations.
- Time to become a Bayesian? (https://medium.datadriveninvestor.com/bayesian-vs-frequentist-for-dummies-58ce230c3796) (not all statistics are frequentist!)


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

http://xkcd.com/1132/