Stats for Psych

September 4th, 2023

PREFACE

PSY1205 notes.

Contents

PART 1

Introduction to statistics

Here's how this tutoring thing is gonna go. First we will cover basic probability and stat theory, then we will cover each statistical test that your course pdf listed. For each test, we will do a short maths intro then move on to maths practice problems then practice in Jamovi. Every time we learn a new concept, I will have a new document on that topic written up. The list of tests and the order in which we will cover them is

(a) tbd...

Things I can change. Amount of time we meet each week (length of each session, number of sessions per week), topics covered (we might be spending too much time on a topic that is important but that you already understand among other reasons), the amount of intro maths explanation for each, amount of maths questions, amount of Jamovi practice. Regarding the readings, you can ask me to change the length of the reading, the amount of maths, the tone used (too conversational? too serious?), or the type of content (maybe more examples and less abstract with the concepts?). Obviously feel free to ask me to change stuff outside of this list since I may not have thought of everything and I will do my best.

Based on what I've seen from the textbook and class notes, the goal of the stats courses you are going through seem to be to give you the bare minimum amount of theoretical knowledge to answer questions about data that you might encounter as a psychologist using the statistical tests they want you to be proficient with. I will try and follow this idea but clearly what I consider the acceptable amount of theoretical knowledge to be using statistical concepts in real life is different from what the people who designed your course think. I also think that more than just statistical tests should be used to answer questions about data but I don't know what a psychologist might need or if you will be taught that in the future so I will let that go for now. If you think I am ever focusing too much on the maths and not the problems solving, tell me and I will fix that.

So before we get into all the fancy mathy bits about understanding stats and all that, I hope that this intro will show you the importance of being very precise about what each concept in maths or stats means and how being precise can help you in both your exams and thinking about stats in general.

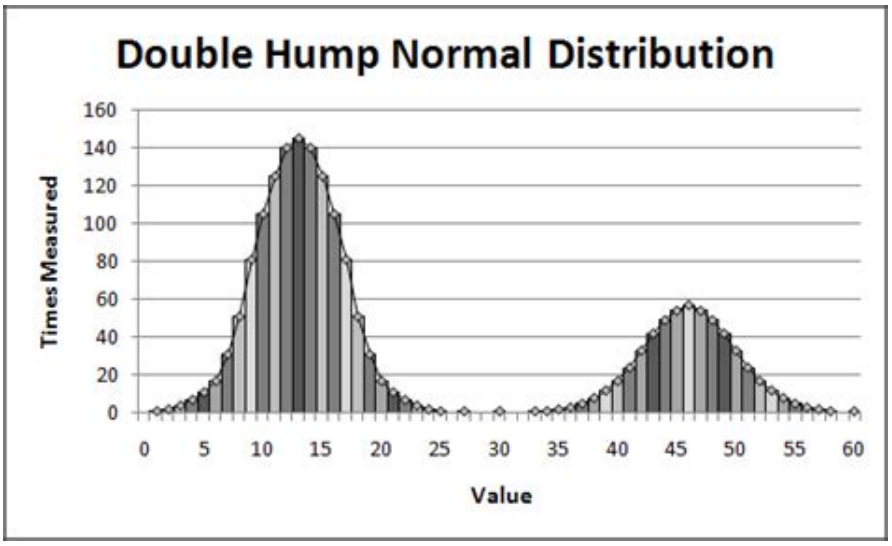
1.1

THE LOGIC BEHIND MATHEMATICS

So your textbook argues that teaching statistics without maths is beneficial to students because the authors believe it should be the case that "understanding a very concrete concept such as the arithmetical mean would be a good deal easier than understanding a rather vague psychological concept such as 'an attitude" (p.1). This is interesting for two reasons. The first is that it suggests that the arithmetic mean is not only more concrete than 'an attitude' but is also simpler than all the forms that 'an attitude' can take. The second is that the authors seem to be under the misunderstanding that a more "concrete" concept should be easier to understand than a "vague" one.

Let's look at the first point and ask if the arithmetic mean is a less complex idea than whatever 'an attitude' is in psychology. The arithmetic mean is a popular idea that everyone with any level of education in probability or statistics will be somewhat familiar with.

So now we turn to our second point. An acceptable verbal definition of the arithmetic mean in a conversational context when applied to a psychological experiment may be "the attitude that a normal person takes". In an psychological experiment, it is almost certainly true that if we take the arithmetic mean of some data set that we want to try and find some number that represents the behaviour of a "normal" or "middling" person. However, if your data has two peaks like the figure below, would it be correct to say that the arithmetic mean is a good representation of "normal", "average", "middling", or some other synonym of a typical subject?



If we just told everyone that the arithmetic mean meant that it is the value that occurs the most, it would be confused as the mode, but if we told everyone it is a value that is probably somewhere in the middle, it would be confused as the median.

However, I will agree that in almost all stats courses, the interpretation of the maths is under-emphasised and that students would have a much easier time learning what the maths means in words than just working with fancy symbols. My approach will be different from the book in that I will show you all these complicated looking mathematics, treat these fancy symbols as something like another language and translate them back into English, and hopefully when you think of stats in the future you can think about how the logic connects instead of just "well I've passed the t- or z-test, whatever that means, so my findings are correct". I hope that all my explanations will lay out clearly, both what the mathematical statements that we will learn mean, and also just as importantly (as we have seen the previous paragraphs), what they do not mean.

imp topic (anova)

PART 2

Probability Theory

2.1

SET THEORY

Definition 2.1.1. The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

Definition 2.1.2. An *event* is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).

Example 2.1.3 (Event operations)**.** Selecting cards...

Theorem 2.1.4. *For any three events, A, B, C, defined on a sample space S,*

a. *Commutativity*

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A;$$

b. *Associativity*

$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C;$$

c. *Distributive Laws*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

d. *DeMorgan's Laws*

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

Proof can be done as an exercise

Definition 2.1.5. Two events A and B are disjoint (or *mutually exclusive*) if $A \cap B = \emptyset$. The events $A_1, A_2, ...$ are *pairwise disojoint* (or *mutually exclusive*) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 2.1.6. If $A_1, A_2, ...$ are pairwise disjoint and $\bigcup_{i=0}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ forms a *partition* of S.

Basics of probability Theory

Theorem 2.1.7. *If P is a probability function and A is any set in S, then*

a. $P\left(\emptyset\right) = 0$, *where $\emptyset$ is the empty set;*
b. $P(A) \leq 1$;
c. $P(A^c) = 1 - P(A)$.

Theorem 2.1.8. *If P is a probability function and A and B are any sets in $\mathcal{B}$, then*

a. $P(B \cap A^c) = P(B) - P(A \cap B)$;
b. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
c. *If $A \subset B$, then $P(A) \leq P(B)$.*

Theorem 2.1.9. *If a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, ..., k$, then the entire job can be done in $n_1 \times n_2 \times ... \times n_k$ ways.*

This is sometimes known as the Fundamental Theorem of Counting. The proof for this can be done as an exercise.

Example 2.1.10 (Lottery - II)**.** Although the Fundamental Theorem of Counting is a reasonable place to start...

Definition 2.1.11. For a positive integer n, $n!$ (read n factorial) is the product of all of the positive integers less than or equal to n. That is,

$$n! = n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1.$$

Furthermore, we define $0! = 1$.

Definition 2.1.12. For nonnegative integers n and r, where $n \geq r$, we define the symbol $\binom{n}{r}$, read n *choose* r, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Example 2.1.13 (Poker)**.** Consider choosing a five-card poker hand...

Example 2.1.14 (Sampling with replacement)**.** Consider sampling $r = 2$ items from $n = 3$ items...

Some authors argue that it is appropriate to assign equal probabilities to the unordered outcomes when "randomly distributing r indistinguishable balls into n distinguishable urns." That is, an urn is chosen at random and a ball placed in it, and this is repeated r times. The order in which the balls are placed is not recorded so, in the end, an outcome such as $\{1, 3\}$ means one ball is in urn

1 and one ball is in urn 3.
But here is the problem with this interpretation. Suppose two people observe this process, and Observer 1 will assign probability $\frac{2}{9}$ to the event $\{1,3\}$. Observer 2, who is observing exactly the same process, should also assign probability $\frac{2}{9}$ to this event. But if the six unordered outcomes are written on identical pieces of paper and one is randomly chosen to determine the placement of the balls, then the unordered outcomes each have probability $\frac{1}{6}$. So Observer 2 will assign probability $\frac{1}{6}$ to the event $\{1,3\}$.
The confusion arises because the phrase "with replacement" will typically be interpreted with the sequential kinda of sampling we described above, leading to assigning a probability $\frac{2}{9}$ to the event $\{1,3\}$. This is the correct way to proceed, as probabilities should be determined by the sampling mechanism, not whether the balls are distinguishable or indistinguishable.

Example 2.1.15 (Caldulating an average)**.** An illustration of the distinguishable/indistinguishable approach...
If there are k places and we have m different numbers repeated $k_1, k_2, ..., k_m$ times, then the number of ordered samples is $\frac{k!}{k_1!k_2!...k_m!}$. This type of counting is related to the *multinomial distribution*, which we will see in Section 4.6.
Figure 1.2.2 (p. 19 of book) is an elementary version of a very important statistical technique known as the *bootstrap* (Efron and Tibshirani 1993). We will return to the bootstrap in Section 10.1.4

Conditional Probability and Independence

Definition 2.1.16. If A and B are events in S and $P(B) > 0$, then the *conditional probability* of A *given* B, written $P\left(A|B\right)$, is

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

Note that what happens in the conditional probability calculation is that B becomes the sample space: $P\left(B|B\right) = 1$. The intuition is that our original sample space, S, has been updated to B. All further occurrences are then calibrated with respect to their relation to B. In particular, note what happens to conditional probabilities of disjoint sets. Suppose A and B are disjoint, so $P(A \cap B) = 0$. It then follows that $P\left(A|B\right) = P\left(B|A\right) = 0$.

Theorem 2.1.17 (Bayes' Rule)**.** *Let $A_1, A_2, ...$ be a partition of the sample space, and let B be any set. Then for each $i = 1, 2, ...,$*

$$P\left(A_i|B\right) = \frac{P\left(B|A_i\right)P(A_i)}{\sum_{j=1}^{\infty} P\left(B|A_j\right)P\left(A_j\right)}.$$

Definition 2.1.18. Two events, A and B, are *statistically independent* if

$$P(A \cap B) = P(A)\,P(B).$$

Note that independence could have been equivalently defined by either $P\left(A|B\right) = P\left(A\right)$ or $P\left(B|A\right) = \frac{P\left(A|B\right)P(B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P\left(B\right)$. The advantage of defining it this way is that it treats the events symmetrically and will be easier to generalise to more than two events.

Theorem 2.1.19. *If A and B are independent events, then the following pairs are also independent:*

a. *A and B^c*
b. *A^c and B*
c. *A^c and B^c*

All three statements can be proved as an exercise.

Definition 2.1.20. A collection of events $A_1, ..., A_n$ are *mutually independent* if for any subcollection $A_{i_1}, ..., A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P\left(A_{i_j}\right).$$

Random Variables

Definition 2.1.21. A *random variable* is a function from a sample space S into the real numbers.

Distribution Functions

Definition 2.1.22. The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X\left(x\right)$, is defined by

$$F_X\left(x\right) = P_X\left(X \leq x\right), \ \forall x.$$

Theorem 2.1.23. *The function $F\left(x\right)$ is a cdf if and only if the following three conditions hold:*

a. *$\lim_{x\to-\infty} F\left(x\right) = 0$ and $\lim_{x\to\infty} F\left(x\right) = 1$.*
b. *$F\left(x\right)$ is a nondecreasing function of x.*
c. *$F\left(x\right)$ is right-continuous; that is, for every number x_0, $\lim_{x\downarrow x_0} F\left(x\right) = F\left(x_0\right)$.*

Whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the association is such that it is convenient to define continuous random variables in this way.

Definition 2.1.24. A random variable X is *continuous* if $F_X\left(x\right)$ is a continuous function of x. A random variable is *discrete* if $F_X\left(x\right)$ is a step function of x.

Definition 2.1.25. The random variables X and Y are *identically distributed* if, for every set $A \in \mathcal{B}$, $P\left(X \in A\right) = P\left(Y \in A\right)$.

Theorem 2.1.26. *The following two statements are equivalent:*

a. *The random variables X and Y are identically distributed.*
b. *$F_X(x) = F_Y(x)$ for every x.*

Sufficiency can be proven as an exercise, necessity is very hard. See page 34.

Density and Mass Functions

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases. Both pdfs and pmfs are concerned with "point probabilities" of random variables.

Definition 2.1.27. The *probability mass function (pmf)* of a discrete random variable X is given by

$$f_X(x) = P(X = x) \ \forall x.$$

Definition 2.1.28. The *probability density function* or *pdf*, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \ dt \ \forall x.$$

Note (notation). The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$", where we read the symbol "$\sim$" as "is distributed as". We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$.

Since $P(X = x) = 0$ if X is a continuous random variable,

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

It should be clear that the pdf (or pmf) contains the same information as the cdf and we should choose whichever one makes the problem simpler.

Theorem 2.1.29. *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

a. *$f_X(x) \geq 0$ for all x.*
b. *$\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (pdf).*

2.2

EXPECTATION

Expectation of a Random Variable

Definition 2.2.1. The expected value, or mean, or first moment, of X is defined to be

$$\mathbb{E}[X] = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{If } X \text{ is discrete} \\ \int x f(x)\, dx & \text{If } X \text{ is continuous} \end{cases} \tag{1}$$

assuming that the sum (or integral) is well defined. Use the following notation to denote the expected value of X

$$\mathbb{E}[X] = \mathbb{E}X = \int x dF(x) = \mu = \mu_X. \tag{2}$$

$\int x dF(X)$ is used as a convenient unifying notation so we do not have to write $\sum_x x f(x)$ for discrete RVs and $\int x f(x)\, dx$ for continuous ones but be aware that $\int x dF(X)$ has a precise meaning that is discussed in real analysis course.

To ensure that $\mathbb{E}[X]$ is well defined, we say that $\mathbb{E}[X]$ exists if $\int_x |x|\, dF_X(x) < \infty$. Otherwise, we say that the expectation does not exist.

To compute $\mathbb{E}[Y]$ when $Y = r(X)$, one way is to find $f_Y(y)$ then compute $\mathbb{E}[Y] = \int y f_Y(y)\, dy$.

Theorem 2.2.2 (The Rule of the Lazy Statistician)**.** *Let $Y = r(X)$. Then*

$$\mathbb{E}[Y] = \mathbb{E}\left[r(X)\right] = \int r(x)\, dF_X(x). \tag{3}$$

$$Z = r(X,Y) \Rightarrow \mathbb{E}[Z] = \mathbb{E}\left[r(X,Y)\right] = \int\int r(x,y)\, dF(x,y). \tag{4}$$

The k^{th} moment of X is defined to be $\mathbb{E}\left[X^k\right]$ assuming that $\mathbb{E}\left[|X|^k\right] < \infty$.

Theorem 2.2.3. *If the k^{th} moment exists and if $j < k$ then the j^{th} moment exists.*

Proof.

$$\begin{aligned}\mathbb{E}\left[|X|^j\right] &= \int_{-\infty}^{\infty} |x|^j f_X(x)\ dx \\ &= \int_{|x|\leq 1} |x|^j f_X(x)\ dx + \int_{|x|>1} |x|^j f_X(x)\ dx \\ &\leq \int_{|x|\leq 1} f_X(x)\ dx + \int_{|x|>1} |x|^k f_X(x)\ dx \\ &\leq 1 - \mathbb{E}\left[|X|^k\right] < \infty.\end{aligned}$$

■

The k^{th} central moment is defined to be $\mathbb{E}\left[(X-\mu)^k\right]$.

Properties of Expectations

Theorem 2.2.4. *If $X_1, ..., X_n$ are RVs and $a_1, ..., a_n$ are constants, then*

$$\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}\left[X_i\right]. \tag{5}$$

Theorem 2.2.5. *Let $X_1, ..., X_n$ be independent RVs. Then*

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_i \mathbb{E}\left[X_i\right]. \tag{6}$$

Note the summation rule does not require independence but the product rule does.

PART 3

Statistical inference

Statistical inference, or "learning" as it is called in comp sci, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is

Given a sample $X_1, ..., X_n \sim F$, how do we infer F?

3.1

PARAMETRIC AND NONPARAMETRIC MODELS

A **statistical model** $\mathscr{F}$ is a set of distributions (or densities or regression functions). A **parametric model** is a set $\mathscr{F}$ that can be parameterized by a finite number of parameters. For example, assuming the data comes from a Normal distribution, the model is

$$\mathscr{F} = \{f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2\sigma^2}(x-\mu)^2,\ \mu \in \mathbb{R}, \sigma > 0\}.$$

This is a two-parameter model. Writing the density as $f(x;\mu,\sigma)$ shows that x is a value of the RV while μ and σ are parameters. In general, a parametric model takes the form

$$\mathscr{F} = \{f(x;\theta) : \theta \in \Theta\}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ. A **nonparametric model** is a set $\mathscr{F}$ that cannot be parameterized by a finite number of parameters. For example, $\mathscr{F}_{ALL} = \{\text{all CDF's}\}$ is nonparametric[1].

Example 3.1.1 (Regression, prediction, and classification)**.** Suppose we observe pairs of data $(X_1, Y_1), ..., (X_n, Y_n)$. X is called a predictor or regressor or feature or independent variable. Y is called the outcome or the response variable or the dependent variable. We call $r(x) = \mathbb{E}\left[Y|X = x\right]$ the **regression function**. If we assume that $r \in \mathscr{F}$ where $\mathscr{F}$ is finite dimensional, then we have a **parametric regression model**. If $\mathscr{F}$ is not finite dimensional then we have a **nonparametric regression model**. The goal of predicting Y for a new patient based on their X value is called **prediction**. If Y is discrete, then prediction is called **classification**. If we wish to estimate the function r, then we call this **regression** or **curve estimation**. Regression models are sometimes written as

$$Y = r(X) + \epsilon \tag{7}$$

where $\mathbb{E}[\epsilon] = 0$. We can always rewrite a regression model this way. To see this, define $\epsilon = Y - r(X)$ and hence $Y = Y + r(X) - r(X) = r(X) + \epsilon$. Moreover, $\mathbb{E}[\epsilon] = \mathbb{E}\mathbb{E}\left[\epsilon|X\right] = \mathbb{E}\left[\mathbb{E}\left[Y - r(X)\right]|X\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right] - r(X)\right] = \mathbb{E}\left[r(X) - r(X)\right] = 0$.

Notation: If $\mathscr{F} = \{f(x;\theta) : \theta \in \Theta\}$ is a parametric model, we write $\mathbb{P}_\theta(X \in A) = \int_A f(x;\theta)\,dx$ and $\mathbb{E}_\theta\left[r(X)\right] = \int r(x) f(x;\theta)\,dx$. The subscript θ indicates that the probability or expectation is w.r.t $f(x;\theta)$, not that we are averaging over θ. Similarly, write $\mathbb{V}_\theta$ for the variance[2].

[1] The distinction between parametric and nonparametric is more subtle than this but we don't need a rigorous definition for our purposes

[2] There's no need to understand the logic behind the notation here, this is more for me so I don't lose track of what's happening

Most inferential problems can be divided into either estimation, confidence sets, or hypothesis testing. We are interested mainly in hypothesis testing so I will only provide a basic description of the other two.

Point Estimation
Point estimation refers to providing a single "best guess" of some quantity of interest. We denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$.

Confidence Sets
A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, ..., X_n)$ and $b = b(X_1, ..., X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \ \forall \theta \in \Theta. \tag{8}$$

In words, (a, b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

Warning! C_n is random and θ is fixed. If θ is a vector then we use a **confidence set** instead of an interval.

Interpretation: Don't think of it as, if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time. This is correct but useless since experiments are rarely repeated that many times. A better interpretation is that if you construct 95 percent confidence intervals with many parameters (e.g. $\theta_1, \theta_2, \theta_3, ...$) then 95 percent of the intervals you construct will trap the true parameter value.

Note: Remember that a confidfence interval is not a probability statement about θ.

Hypothesis Testing
In hypothesis testing, we start with some default theory called a **null hypothesis** and ask if the data provides sufficient evidence to reject the theory. If not, we retain the null hypothesis.

Example 3.1.2 (Test if a coin is fair)**.** Let

$$X_1, ..., X_n \sim \text{Bernoulli}(p)$$

be n independent coin flips. Suppose we want to test if the coin is fair. Let H_0 denote the hypothesis that the coin is fair and let H_1 denote the hypothesis. H_0 is called the **null hypothesis** and H_1 the **alternative hypothesis**. They can be written as

$$H_0 : p = \frac{1}{2} \text{ versus } H_1 : p \neq \frac{1}{2}.$$

It seems reasonable to reject H_0 if $T = \left|\hat{p}_n - \frac{1}{2}\right|$ is large. When we discuss hypothesis testing in detail, we will be more precise about how large T should be to reject H_0.

Another question we could ask is if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rates in the two groups. Consider the following two hypotheses:

The Null Hypothesis: The disease rate is the same in the two groups.
The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis. This is an example of hypothesis testing. More formally, suppose that we partititon the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{9}$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis.

Let X be a random variable and let $\mathscr{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathscr{X}$ called the rejection region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:
$X \in R \Rightarrow$ reject H_0
$X \notin R \Rightarrow$ retain (do not reject) H_0.

Usually the rejection region R is of the form

$$R = \{x : T(x) > c\} \tag{10}$$

where T is a test statistic and c is a critical value. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that they are guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. Rejecting H_0 when H_0 is true is called a type I error and retaining H_0 when H_1 is true is called a type II error.

3.2

HYPOTHESIS TESTING

Summing up hypothesis testing. Every nonparametric procedure will have these steps.

First, state the hypotheses. There are two types of hypotheses, null and alternate. The null hypothesis says that no difference exists between the conditions, groups, or variables. The alternate hypothesis, also called a research hypothesis, predicts a difference or relationship between conditions, groups, or variables.

The alternate hypothesis may be directional or nondirectional. A directional (one-tailed) hypothesis predicts a statistically significant change in a particular direction. A nondirectional (two-tailed) hypothesis predicts a statistically significant change, but in no particular direction.

We then set the risk (or level of significance) associated with the null hypothesis. Whenever we perform a test, there is always some chance that the results we get are due to chance instead of any real difference. Therefore, when we perform such tests, we state the level of risk that we are willing to accept. The two types of errors we can make are type I, where we claim that there is a difference (alternate is true and null is false) when in reality there is no difference and the null hypothesis is true, and type II, where we claim there is no difference (null is true and alternate is false) when in reality there is a real difference and the null hypothesis is false. The commonly accepted way of stating your risk levels is in terms of the probability you allow yourself to make type I errors α. Most of the time, $\alpha = 0.05$ is used, which means that when we claim our alternate hypothesis is true, we are correct 95% of the time.

We also choose a suitable test statistic based on the characteristics of the data. For example, some tests are appropriate for two sample tests, while others are more appropriate for three or more samples. Different tests may also be suitable for different measurement scales.

We then compute the test statistic. This is usually done with a computer program and the interpretation is different for each statistic so there is not much to say here.

Determine what values the test statistic can take in order to reject the null hypothesis using the appropriate table of critical values for the particular statistic. Finding this critical value may require you to use data characteristics such as the degrees of freedom, number of samples, and/or number of groups.

Compare test statistic value with critical value in the table. Interpret the results then report them.

PART 4

Wilcoxon

Imagine you give an attitude test to a small group of people. After you deliver some treatment, say a daily vitamin supplement for a few weeks, you give the same group of people another attitude test. You then compare the two measures to see if there is any meaningful difference between the two sets of scores.

The characteristic to note is that every test score is paired. Since everyone is tested twice, for each initial test score, there is another test score for that same person after the treatment. This forms a pair for each test score[1]. The parametric equivalent to these tests is called the Student's t-test, t-test for matched pairs, t-test for paired samples, or t-test for dependent samples.

Note that the procedures for the Wilcoxon signed-rank and sign tests change depending on whether the samples are small or large.

[1] Assuming no attrition, of course

4.1

RANKING DATA

Many nonparametric procedures involve ranking data values. To rank all values from the first

Students who ate breakfast	Students who skipped breakfast
87	93
96	83
92	79
84	73

Table 1 Healthy breakfast or not

table, we simply place them all in order in a new table from smallest to largest. The table below shows how to do this with our breakfast example, keeping the values for the students who ate breakfast bolded. On the surface, it seems that the students who ate breakfast scored higher.

Students who ate breakfast	Students who skipped breakfast
73	1
79	2
83	3
84	4
87	5
92	6
93	7
96	8

Table 2 Ranking data

However, if you wish to claim statistical significance, some type of procedure or test is required.

What if we have ties? Imagine we replaced the score for the student with a rank of 4 (scoring 84 points) with 83 points. This student is now tied with the student with a rank of 3. We simply give the students the average of their rank values. See the table below for how this would work. The rows in bold are now the tied values instead of values for those who ate breakfast. Most nonparametric statistical tests require a different formula when a sample of data contains ties.

Students who ate breakfast	Students who skipped breakfast
73	1
79	2
83	**3.5**
83	**3.5**
87	5
92	6
93	7
96	8

Table 3 Ties in data

4.2

WILCOXON SIGNED-RANK/WILCOXON MATCHED-PAIRS

The formula for computing the Wilcoxon T for small samples is:

$$T = \text{smaller of } \sum \mathbb{R}_+ \text{ and } \sum \mathbb{R}_- \tag{11}$$

where $\sum \mathbb{R}_+$ is the sum of the ranks with positive differences and $\sum \mathbb{R}_-$ is the sum of the ranks with negative differences.

We then examine the T statistic using the relevant table of critical values. However, if the number of pairs n exceeds those available from the table, then a large sample approximation may be performed. Simply put, we compute a z-score and use a table with the normal distribution to obtain a critical region of z-scores. First calculate:

$$\bar{x}_T = \frac{n\,(n+1)}{4} \tag{12}$$

where $\bar{x}_T$ is the mean and n is the number of matched pairs included in the analysis,

$$s_T = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}} \tag{13}$$

where s_T is the standard deviation. The z-score approximation is therefore:

$$z^* = \frac{T - \bar{x}_T}{s_T}. \tag{14}$$

Let's work through a sample matched-pairs test with small data samples. The counseling staff of Clear Creek County School District has implemented a new program this year to attempt reducing bullying at primary schools. To evaluate the effectiveness of our anti-bullying program, the school district has decided to compare the percentage of successful interventions last year before the program began with the percentage of successful interventions this year with the program in

place[1]. The next table shows 12 elementary school counselors and their reported percentage of successful interventions last year and this. Since the samples are small, we need a nonparametric

Participant number	Last year	This year
1	31	31
2	14	14
3	53	50
4	18	30
5	21	28
6	44	48
7	12	35
8	36	32
9	29	34
10	29	34
11	17	27
12	40	42

Table 4 Percentage of successful interventions

procedure. Since we are comparing two related, or paired, samples, we will use the Wilcoxon matched-pairs test. We will use the steps laid out in the previous section.

State the null and alternative hypotheses. The null hypothesis states that the counselors reported no difference in the percentages last year and this year. The alternative states that the counselors observed some differences between this year and last year. Note that our alternative hypothesis does not claim that these differences are supposed to be positive ones, so our alternative is a two-tailed, nondirectional hypothesis. They can be written as

$$H_0 : \mu_D = 0$$
$$H_A : \mu_D \neq 0.$$

Set the level of risk. It is standard to choose $\alpha = 0.05$, which states that we will make type I errors 5% of the time.

Choose the appropriate test statistic. We have already shown that the Wilcoxon matched-pairs test is appropriate.

Compute the test statistic. Exclude data points where the difference between two matched pairs is zero, then rank the data again using the **absolute values**. With the new ranks, check that the ranks with positive differences are $9, 7, 4.5, 10, 1, 6, 8$, and 2. Adding the ranks with positive difference we get $\sum \mathbb{R}_+ = 47.5$. The ranks with negative differences are 3 and 4.5, summing to $\sum \mathbb{R}_- = 7.5$. We choose the smaller of the two rank sums and therefore the Wilcoxon is $T = 7.5$.

[1]You may realise that since we have not stated what kind of program this is nor defined clearly what "interventions" mean or what a "successful intervention" would look like, this is not a very good experiment. Hopefully in any experiments you run in real life, you will have defined criteria for success better than I have here but for now, we will assume that we have some well-defined concept of a "successful intervention".

Determine the critical value needed to reject the null hypothesis. To find the critical value of the Wilcoxon matched-pairs test, we need the sample size, level of significance, and type of hypothesis. Note that since we discarded the two participants with no reported differences between this year and the last, we now have a sample size of 10 instead of 12. The table in the course handbook says that for a significance level of 0.05 with a two-tailed hypothesis and $n = 10$, the critical value is $T = 8$. An obtained value less than or equal to 8 will lead us to reject our null hypothesis.

Compare obtained value with critical value. Since our obtained value $T = 7.5$ is less than the critical value, 8, we reject the null hypothesis.

Interpret and report results. By rejecting the null hypothesis, we suggest that a real difference exists between last year's percentages and this year's percentages. Reporting the results would go something like this.
For this experiment, the Wilcoxon signed rank (matched-pairs) test ($T = 7.5, n = 12, p < 0.05$) indicated that the difference in percentage of successful interventions was statistically significant. In addition, the sum of the positive difference ranks ($\sum \mathbb{R}_+ = 47.5$ was larger than the sum of the negative difference ranks ($\sum \mathbb{R}_- = 7.5$), showing a positive impact from the program. Therefore, our analysis supports that the new anti-bullying program provides positive benefits towards the improvement of the aspect of student behaviour that school counselors perceive.

Confidence Intervals

Now we aim to construct a confidence interval based on the Wilcoxon signed rank test for matched pairs. Such a confidence interval provides a range of values that fall within the population with a $100\,(1-\alpha)$ percent chance. So now instead of just answering yes or no, we can say that the difference in successful interventions due to the new anti-bullying program is some interval $[x, y]$.

First, let us introduce ourselves with some notation. Index each participant as some $i = 1, \ldots, n$. Each counselor i has matched pair $\left(x_1^i, x_2^i\right)$, where x_1^i is the i's reported percentage of successful interventions before the anti-bullying program was implemented and x_2^i the reported percentage after program implementation. Compute value $D_i = x_1^i - x_2^i$ for each participant i. Then compute u_{ij} as defined below:

$$u_{ij} = \frac{D_i + D_j}{2} \text{ for all} 1 \leq i \leq j \leq n.$$

So let's pause here to give some interpretation. All we have done here is said that the difference between the reported percentages of counselor i is represented by D_i so for each of our matched pair data points, we have another number D_i. Then we calculate u_{ij} for all unique combinations of individuals. For example:

$$u_{11} = \frac{D_1 + D_1}{2} = \frac{-3 + -3}{2} = -3$$
$$u_{12} = \frac{D_1 + D_2}{2} = \frac{-3 + 12}{2} = \frac{9}{2}.$$

You may notice that the number of u_{ij}s we have is a counting problem. If you wish to practice, note that $u_{12} = u_{21}$ so we have an unordered problem with replacement. Then order these averages from smallest to largest. The median of the ordered averages gives a point estimate of the population median difference.

Use the table in the handbook to find the endpoints of the confidence interval. First, determine T from the table that corresponds with the sample size and desired confidence such that $p = \alpha/2$. We seek to find a 95% confidence interval and we have $n = 10$ and $p = 0.05/2$. For single-tailed hypotheses, the table provides $T = 8$. Note that we are using a single-tailed hypothesis because we wish to provide a confidence interval around the point estimate of the population median. Therefore, we construct two single-tailed hypotheses around the median with a level of significance that adds up to $\alpha = 0.05$.

The endpoints of the confidence interval are the Kth smallest and Kth largest values of u_{ij}, where $K = T + 1$. So for our case, $K = 9$. The ninth smallest value is 0.5 and the ninth largest is 12. Therefore, we can say that we are 95% confident that the difference of successful interventions due to the new bullying programs lies between 0.5 and 12.

I can go through the same test with a large data sample and normal approximation if required.

4·3

WILCOXON SIGN TEST/RANK-SUM TEST

Related samples can be analyzed more efficiently if the data is reduced to, or already comes in, dichotomous (binary) results. This method is based on Gibbons and Chakraborti (2010).

We start by identifying the sign of the difference demonstrated by each set from the related data samples, then sum all positive and negative differences as n_p and n_n, respectively.