Statistical Thinking for Informed Decision Making

Leslie Myint

Contents

Preface

This is the class textbook for Statistical Thinking for Informed Decision Making. This book will serve as the source for pre-class reading on statistical concepts. Boxed sections entitled **Reflection Question** are for you to think about on your own and will be used as a source of discussion during class.

Chapter 1

Hypothesis Testing

In this chapter, we will learn about statistical hypothesis testing, one of the most ubiquitous frameworks for making statistical inferences - for learning about the world. We will start by learning the general conceptual framework underlying hypothesis tests. This will allow us to understand a large variety of statistical tests that are commonly used in scientific practice.

1.1 Aside: what is statistical inference?

I have said that hypothesis testing is a framework for making statistical inferences. What exactly is statistical inference? Statistical inference refers to the process of using data to make conclusions about the world. It deals with estimating true underlying quantities and expressing our uncertainty about those estimates. For example, we may assume that there is some true prevalence of malaria in a particular population which is unknown to us because we cannot collect malaria outcome data on everyone in the population. Instead, we may collect data on malaria outcomes within a certain research center. Statistical inference deals with using that collected data to estimate the true prevalence of malaria and to quantify our uncertainty about that estimate, typically with a range. This range is most often a confidence interval, a concept we will review in this chapter.

Hypothesis testing, we will see, is useful for comparing groups and is most often used to identify where there are differences between groups. This is a form of statistical inference because we are still using data to learn a truth about the world - the truth being whether or not these groups of interest are different. Differences are of fundamental interest in public health and in science because we care about comparisons: comparisons of different demographic groups, of policies, of old and new treatments.

1.2 Conceptual framework

A statistical test aims to answer the question: is there a difference? This question can be made more specific in different situations:

- Is there a difference in the levels of a trait between groups? e.g. CD4 levels in different HIV risk groups.
- Is there a difference in the proportion of a trait between groups? e.g. Birth defects in the children of Zika-exposed and unexposed mothers.
- Is there a difference between the average level of a trait in population X and a meaningful cutoff value? e.g. Rate of operating room mortalities in a certain hospital as compared to the national average.

All of the above questions are common in public health practice and can be examined in the hypothesis testing framework. Let's develop these ideas using the example of adverse pregnancy outcomes in the children of Zika-exposed and unexposed mothers. In this case, the true underlying quantities of interest are (1) the probability of adverse pregnancy outcomes in Zika-exposed mothers p_E and (2) the probability of adverse pregnancy outcomes in Zika-unexposed mothers p_U. p_E and p_U are also referred to as **parameters**.

The first component of a hypothesis test is the **null hypothesis**, denoted H_0. This is a statement about the true parameters that describes the situation where *nothing interesting is happening.* It is interesting if there is a difference in the rate of birth defects between Zika-exposed and Zika-unexposed mothers. What would be uninteresting? If there were no difference in the rate of birth defects between the two groups. In other words, it would be uninteresting if the true probabilities of birth defects are the same. We can write this mathematically as:

$$H_0 : p_E - p_U = 0$$

Hypothesis testing is akin to proof by contradiction. In a proof by contradiction, we assume that the opposite of what we wish to prove is true. Under this assumption, we determine what must logically follow. If we end up with a statement that is false, then our original assumption must have been wrong. For example, let's say I want to prove the statement "Not all umbrellas are green." In a proof by contradiction, I would assume the opposite: "All umbrellas are green." What logically follows is that every umbrella I see must be green. I'm likely to stumble upon a counterexample very quickly, which proves my original assertion.

Hypothesis testing works similarly. We set up an assumption, the null hypothesis. Under this assumption, we use the tools of probability to determine how likely our data is. If our data is unlikely under this assumption, then perhaps our assumption was wrong to begin with. The key idea is that we set up our assumption to be a description of the world with nothing interesting going on. If this assumption might be wrong, then perhaps there *is* something interesting going on. This uncertainty is what differentiates hypothesis testing from the proof by contradiction. The moral of statistics, and perhaps science in general, is that few things can ever be proven without a doubt but that we can get close with accumulation of evidence.

The table below summarizes the comparison between proof by contradiction and hypothesis testing.

Step	Proof by contradiction	Hypothesis testing
State what you want to show	Not all umbrellas are green.	There is a difference in the rate of birth defects between children of Zika-exposed and unexposed mothers.
Make an assumption that is the opposite of what you want to show	All umbrellas are green.	There is no difference in the rate of birth defects between children of Zika-exposed and unexposed mothers.
Collect data	Obtain umbrellas and record their colors.	Collect information on birth outcomes for Zika-exposed and unexposed mothers.
Evaluate discrepancies	A single counterexample of a non-green umbrella is enough to prove our original assertion.	Use a statistical test that provides a discrepancy measure (often a p-value).

1.3 Statistical details

The previous section set up the big picture ideas of hypothesis testing. In this section, we will delve into more of the statistical details. Let's continue with our example of determining if there is a difference in the rate of adverse pregnancy outcomes in Zika-exposed and unexposed mothers. Recall that the probability of adverse pregnancy outcomes in Zika-exposed mothers is p_E and p_U for unexposed mothers. The null hypothesis (which we aim to gather data to refute) describes the situation where nothing interesting is happening:

$$H_0 : p_E - p_U = 0$$

This states that the probabilities of birth defects are equal in both groups. Recall from the previous section that the next step is to use collected data to evaluate if it is discrepant with this null hypothesis. Let's look at the following data from a cohort study published in late 2016:

	Zika-positive	Zika-negative	Total
Adverse pregnancy outcomes	58	7	65
No adverse pregnancy outcomes	67	54	121
Total	125	61	186

Note that $p_E - p_U$ in the formulation of the null hypothesis above denotes the true difference in probabilities. We can obtain an **estimate** of this true value by subtracting the sample proportions:

$$\frac{58}{125} - \frac{7}{61} = 0.35$$

This represents a resonable guess from our data of the difference in adverse outcome rates. We want to compare this estimate to 0, also called the **null value**. The null value gives the value of true difference in probabilities if nothing interesting were going on. Can't we just use the difference between our estimate and the null value? This is a step in the right direction but how do we know if this difference is big? What if differences of 0.35 happen quite often just by chance? We need to take into account the uncertaintly/variability of the estimate (called the **standard error**) to see how much this difference exceeds what we might reasonably see by chance. The quantity that is computed in statistical tests that accounts for all of this information (the estimate, the standard error, and the null value) is called a **test statistic**. Often, but not always, a test statistic has the form

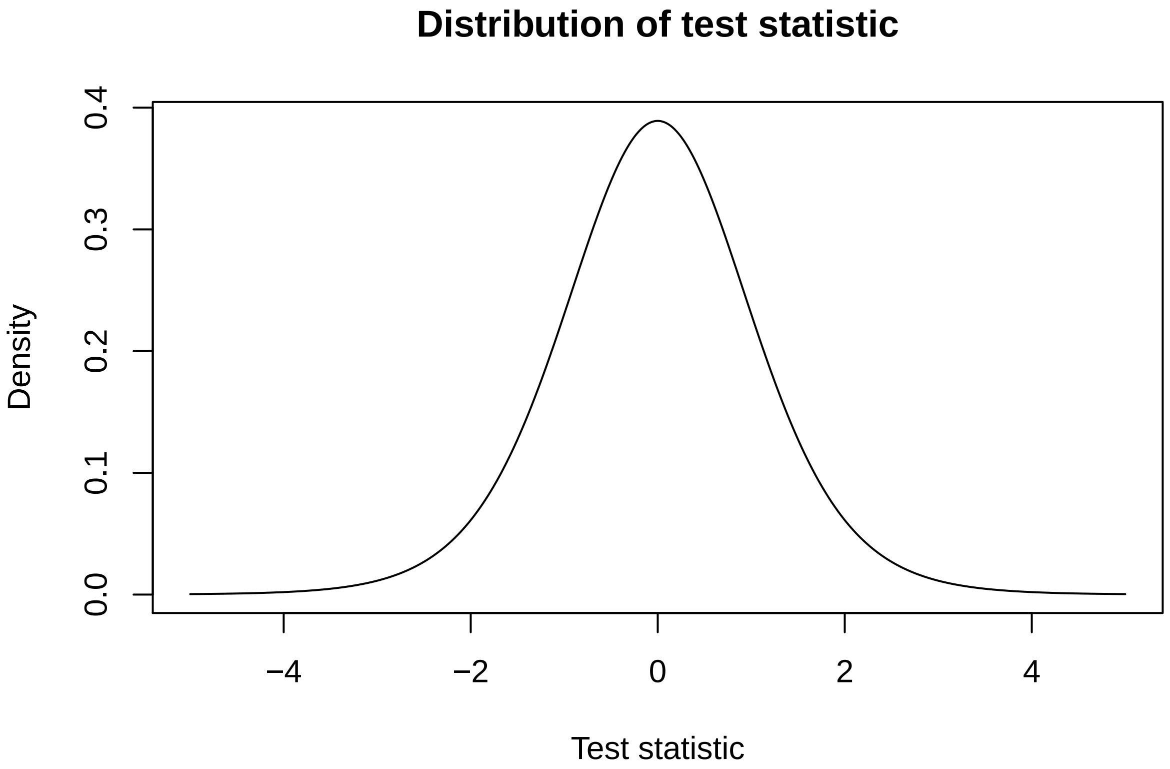
$$\text{test statistic} = \frac{\text{estimate} - \text{null value}}{\text{standard error of estimate}}$$

What is common to all test statistics is that they are used to give a measure of discrepancy with the null hypothesis. When test statistics are large, we reject the null hypothesis. In this case, we would say that there is indeed a difference in rates of adverse pregnancy outcomes between the two groups of mothers. Let's look at the formula above to see why it makes sense that larger test statistics suggest a higher level of discrepancy with the null hypothesis. We see that the test statistic is large when two things happen: (1) the estimate is far from the null value and (2) the standard error of the estimate is low. When (1) happens, our data is telling us that the quantity that we're interested in is quite different from the null value. When (2) happens, our estimate is more reliable. So if our estimate of the truth is far from the null value *and* is reliable, there is some suggestion that we should reject the null hypothesis. Although not all test statistics have this form, they do all have the general use that large values correspond to decisions to reject the null hypothesis.

How large does a test statistic need to be to reject the null hypothesis? Is a threshold of 4 suitable? Statistical theory is able to help us here. It turns out that different thresholds lead to different error rates - we may incorrectly reject the null hypothesis when we should not or we may incorrectly fail to reject the null hypothesis when we should reject it. In other words, errors occur if we claim differences when there are none or fail to notice differences when they truly exist. With this setup, the following outcomes are possible:

	H_0 true	H_A true
Reject	False positive Type I error α	True positive Power β
Fail to reject	True negative $1 - \alpha$	False negative Type II error $1 - \beta$

To understand why different thresholds on the test statistic affect error rates, we need to understand how test statistics vary from dataset to dataset. The estimate of the true quantity of interest $(p_E - p_U)$ varies from dataset to dataset, and because it is used to compute the test statistic, the test statistic also varies from dataset to dataset. In other words, the test statistic comes from a distribution which might look something like this.

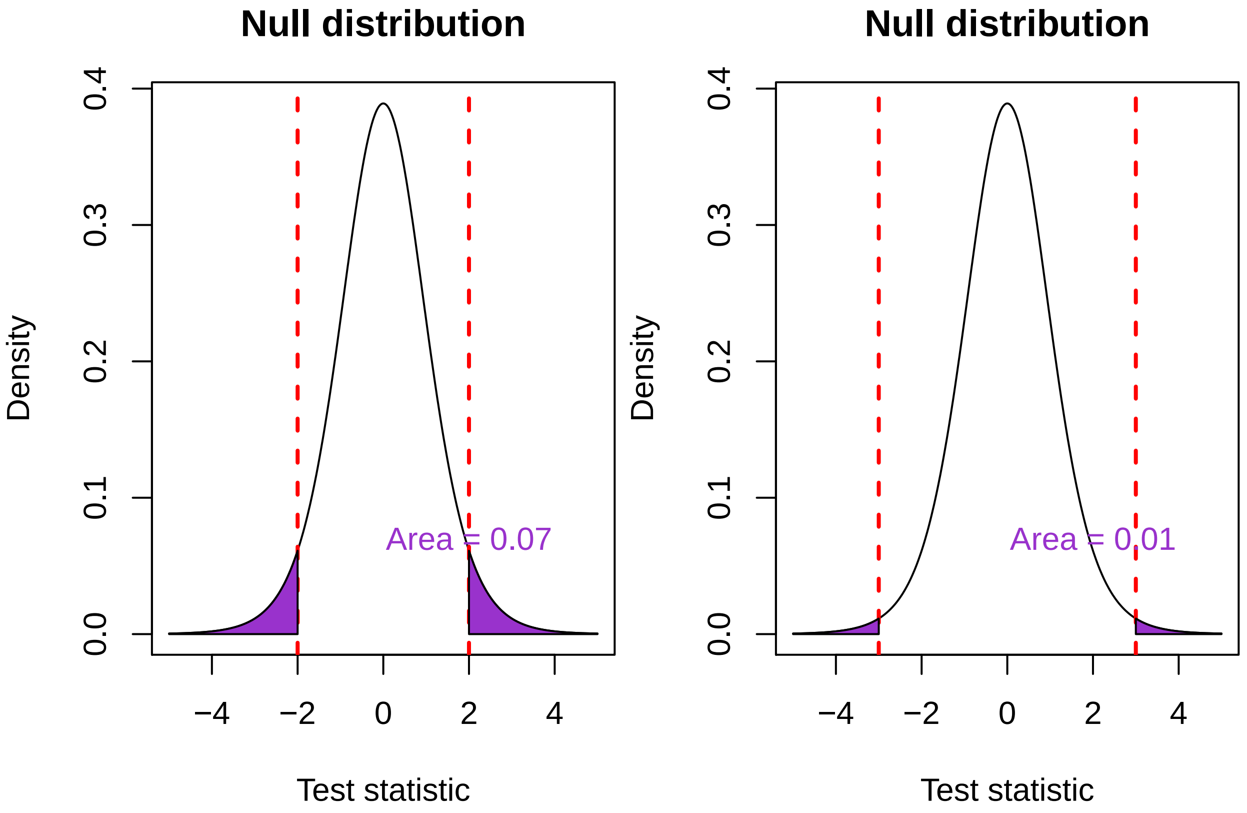


Note that this distribution is centered around zero. This turns out to be a good description of what the test statistics might look like if the null hypothesis were true. Why is that? If the null hypothesis is true, the true difference in probabilities is zero. Thus we expect estimates of the true difference in probabilities to be around zero. This means that the numerator of the test statistic will tend to be around zero, and the entire test statistic will tend to be around zero. All of this is under the assumption that the null hypothesis is true. This distribution is called the **null distribution** and is also described as the distribution of the test statistic *under the null hypothesis*.

If the null hypothesis is true, any rejection of it is an error. Such an error is called a **type I error**, and the probability of making this type of error is used to pick thresholds for decision making in hypothesis testing. The probability of making a type I error is called the **type I error rate**, is denoted by α and can be written in probability notation as:

$$\alpha = P(\text{reject} \mid H_0)$$

For historical reasons, this rate is typically set to be 0.05: if the null hypothesis is true, then we expect to incorrectly reject the null hypothesis 5% of the time. In other words, if there really is no difference in the probability of adverse outcomes between the two groups, then we can expect to wrongly conclude that there is a difference 5% of the time. This 5% is very commonly adopted but can vary from discipline to discipline depending on desired levels of stringency. Lower values of α are more stringent because we are saying that the probability of a type I error should be lower. Statistical theory tells us how test statistic thresholds correspond to type I error probabilities. For example, with the null distribution we saw above, test statistic thresholds of 2 and 3 correspond to type I error rates of 0.073388 and 0.0133437, illustrated in the picture below.



A threshold of 2 would not work if we wanted to maintain a type I error rate of 0.05 because it is associated with a higher type I error rate of 0.073388. It is too low of a threshold because using it would give a higher rate of type I errors than we want. A threshold of 3 is too high of a threshold. While the associated error rates are lower than our 0.05 threshold, using too high of a threshold on the test statistic means that we are also less likely to reject the null when it truly is false. In this example, the threshold that corresponds to $\alpha = 0.05$ is 2.23.

A very closely related concept is the **p-value**. A p-value is calculated from a particular test statistic and represents the probability of seeing a test statistic as or more extreme than the one seen if the null hypothesis were true. Let's say that our statistical test gave us a test statistic of t, then we can express the p-value as:

$$\text{p-value} = P(|\text{test statistic}| \geq t \mid H_0)$$

Similarly to how high values of the test statistic indicate higher discrepancy with the null hypothesis, low p-values indicate higher discrepancy with the null hypothesis. Why? Remember that test statistics provide a measure of discrepancy with the null hypothesis. A low p-value tells us that, if the null were true, it would be unlikely to see a test statistic as or more extreme than the one we saw.

We have seen how test statistics and p-values give discrepancy measures with the null hypothesis. We have not focused our attention on looking directly at our estimate of the true difference in adverse event rates. Recall that our estimate of $p_E - p_U$ was 0.35. Surely, the difference in probabilities is not *exactly* 0.35. It would be nice to express this estimate with some sort of "wiggle" room, with some degree of uncertainty. This can be achieved with a **confidence interval**.

A confidence interval can generally be written as

$$\text{estimate} \pm k \times \text{standard error of estimate}$$

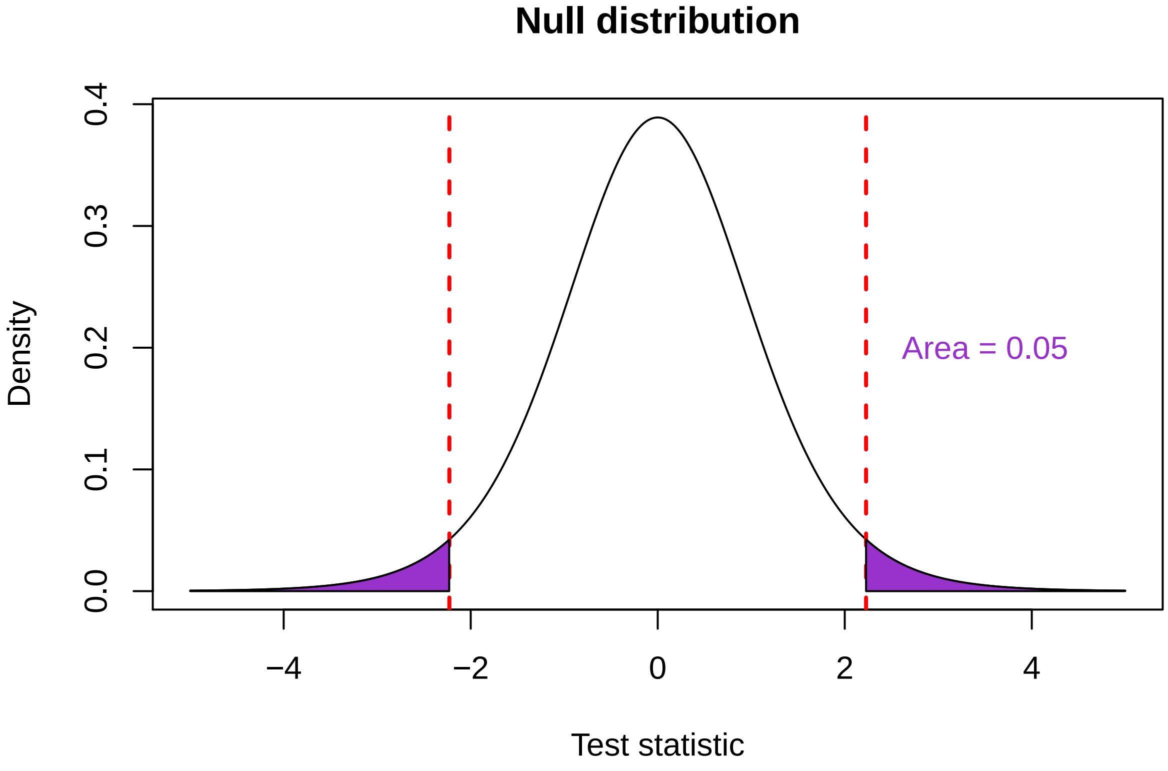
The value of k determines the **coverage probability** of the confidence interval. The coverage probability gives the proportion of times over many data collections that such a confidence interval contains the true parameter of interest - contains the true value for the difference in adverse event rates. A coverage probability of 0.95 would indicate: “if I were to collect many different datasets and use this procedure each time to construct a confidence interval, I would expect 95% of those intervals to contain the true value of the parameter of interest.” Because confidence interval calculations often rely on approximations, the nominal coverage probabilities are sometimes not equal to the actual coverage probabilities. For example, a nominal 95% confidence interval may actually only cover the true value, say, 90% of the time.

1.3.1 One and two-tailed tests

You will often see in scientific literature expressions such as one- and two-tailed tests. What does this mean? This refers to how p-values are calculated. Recall that a p-value is calculated from a particular test statistic and represents the probability of seeing a test statistic as or more extreme than the one seen if the null hypothesis were true. Let’s say we had a test statistic of 3. For a two-tailed test, we consider “more extreme” to be values greater than 3 or less than -3. In other words, we consider “more extreme” as being in both the positive and negative directions. This corresponds to writing the null (H_0) and alternative (H_A) hypotheses as:

$$H_0 : p_E - p_U = 0$$

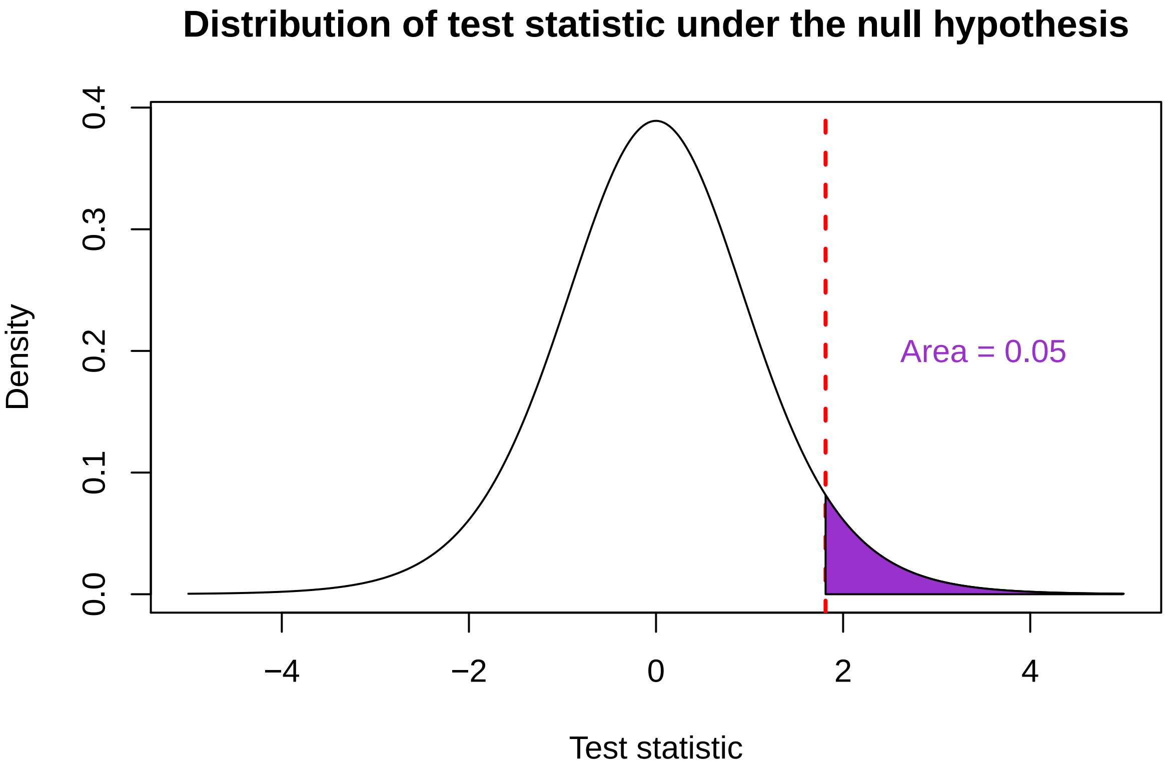
$$H_A : p_E - p_U \neq 0$$



For a one-tailed test, we consider “more extreme” to be in one direction only. If in the positive direction, the situation looks like this:

$$H_0 : p_E - p_U = 0$$

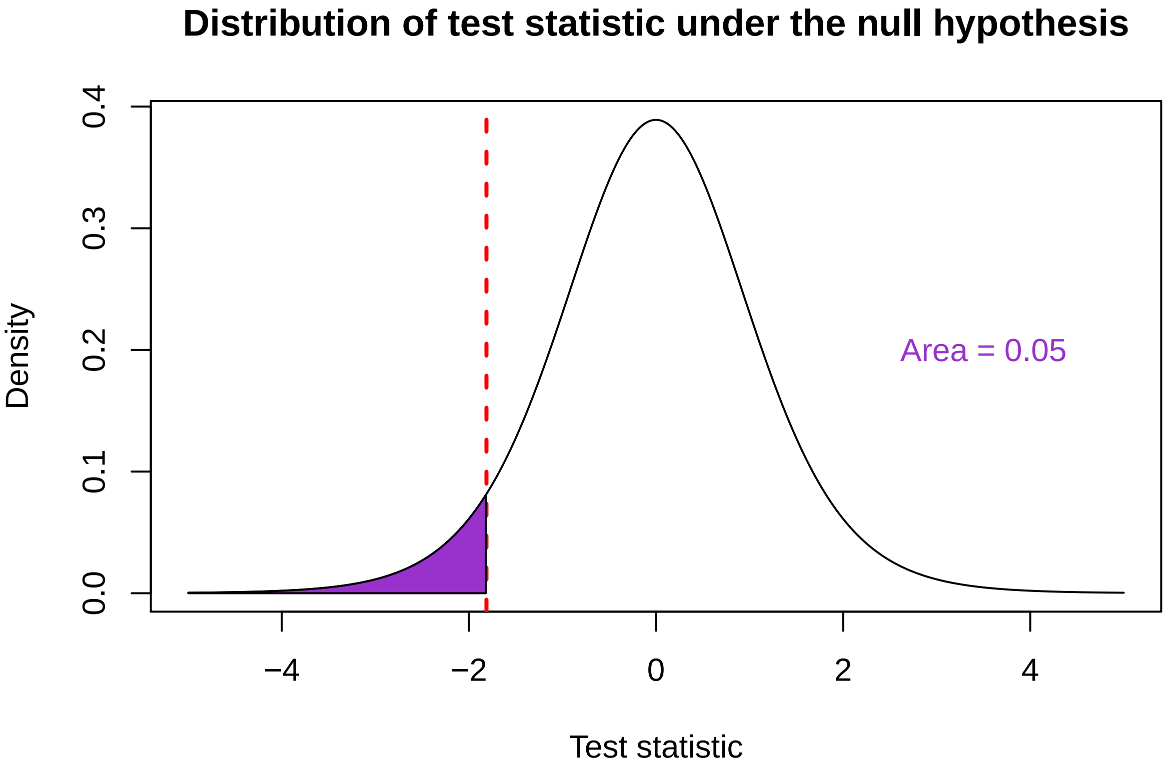
$$H_A : p_E - p_U > 0$$



If in the negative direction, the situation looks like this:

$$H_0 : p_E - p_U = 0$$

$$H_A : p_E - p_U < 0$$

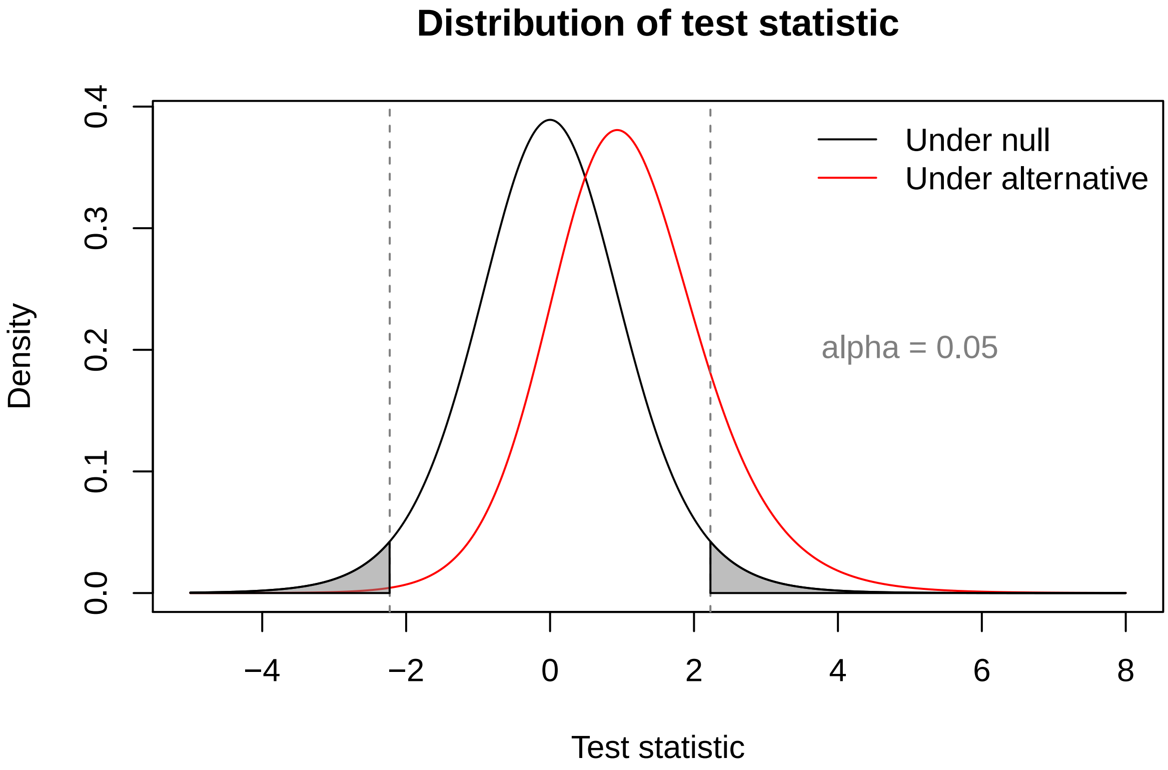


So when would we use a two-tailed vs one-tailed test? Two-tailed tests are used when we don't know a priori if the difference should be positive or negative. In this case, we might *feel* that the difference $p_E - p_U$ should be positive because it seems that the rate of birth defects for Zika-exposed mothers would be higher. However, we don't know for sure. One-tailed tests can be used if we know for sure that the difference is of a certain sign. People will often raise an eyebrow when one-tailed test results are reported because it is "easier" to obtain a statistically significant results. We can see from the plots above that the test statistic thresholds are lower when using one-tailed tests.

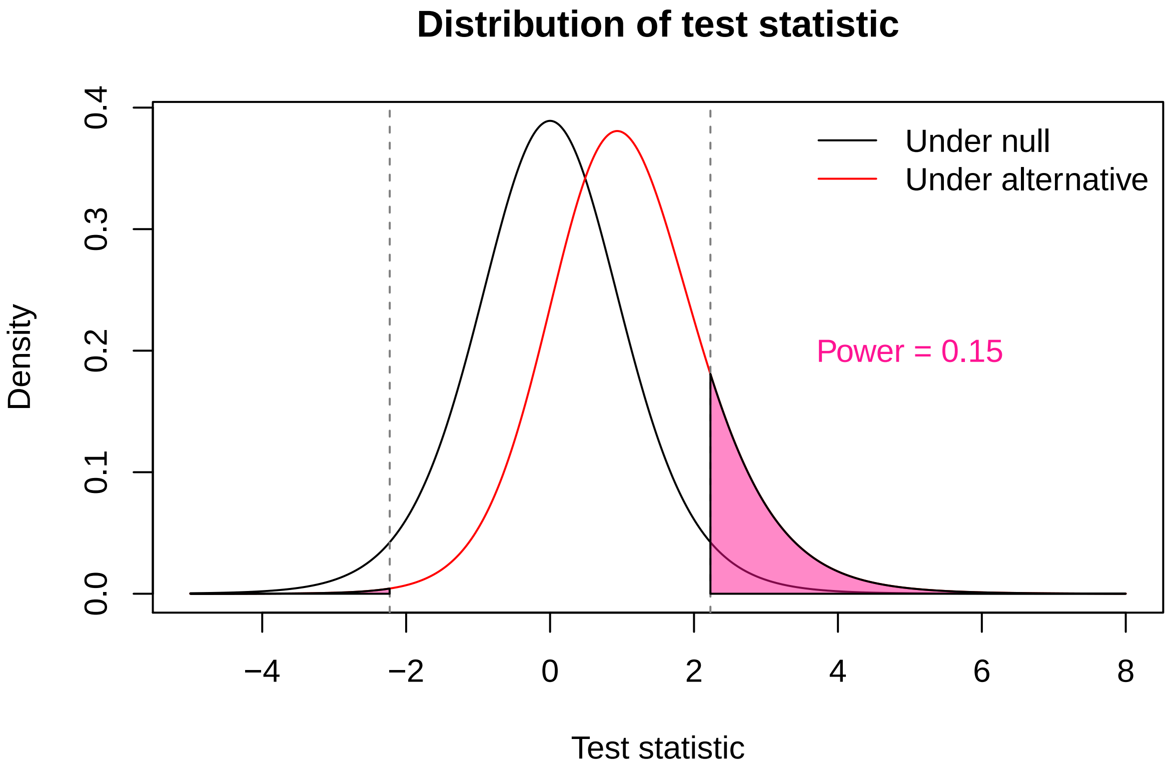
1.3.2 Statistical power

Statistical power is the probability of detecting an association **given that there truly is an association**. While p-values require us to know the distribution of the test statistic "under the null", calculating power requires us to know the distribution "under the alternative".

The gray lines indicate the test statistic thresholds (two-tailed) for rejection for a type I error rate $\alpha = 0.05$. The gray area is equal to to α.



Because the thresholds (gray lines) indicate when we reject the null, we can look at the pink area underneath the alternative distribution (red curve) to calculate power.



There are three main determinants of statistical power:

- Type I error rate (α): Being more stringent (lower) with this rate is nice for controlling false positives, but it also decreases statistical power (imagine the gray lines moving outward).
- Effect size: This measures the distance between the null and alternative distributions (how far apart the black and red curves are). The further the alternative distribution is from the null distribution, the higher the power.
- Sample size: Sample size determines the precision of our estimates. (Recall that sample size always comes up in formulas for standard errors.) Estimate precision is reflected in the width of the distributions. The narrower the distributions (the more precise our estimates and the higher our sample size), the higher the power.

1.4 Common statistical tests

1.4.1 Tests for comparing continuous data

The most common statistical tests used to compare continuous values in two groups are t-tests and Wilcoxon rank sum tests. With these tests, we want to know if one group has higher values than another group.

In t-tests, we are essentially comparing the means of two groups and accounting for the variability of the observations. The test statistic in a t-test can be writen as

$$\frac{\text{difference in means}}{\text{SE(difference in means)}}$$

The above is the situation for what is called an **unpaired t-test**. It is so called because the observations in the two groups aren't linked in any way. In contrast, in a **paired t-test** the two "groups" are linked. This is most often seen in pre-post designs when each individual has both a pre-intervention and post-intervention measurement. In this case, we are interested in if the pre/post differences (also called change scores) are equal to 0. Here the test statistic can be written as:

$$\frac{\text{mean of the change scores}}{\text{SE(change scores)}}$$

T-tests rely on the assumption that the observations come from a normal distribution or that the sample size is large enough for an approximation (this has to do with something called the **central limit theorem**) to kick in. An often used rule of thumb is 20 observations per group. When these assumptions are not met, the prescribed type I error rate α, typically set to 0.05, may not hold. That is, the actual type I error rate if you use the t-test with unmet assumptions might not actually be 0.05. When this is the case researchers use **nonparametric tests** called rank tests. Nonparametric tests do not make assumptions about the distribution of the data. The nonparametric equivalent of an unpaired t-test is called the **Wilcoxon rank sum test**. It is conceptually similar to a t-test where we replace the actual values with their ranks. Say for example that we have the following data:

Group 1: 10 20 30

Group 2: 15 35 40

A t-test would operate on the actual values above. The rank sum test uses the ranks of the observations in the full data (combined over the two groups):

Group 1: 1 3 4

Group 2: 2 5 6

If one group has lower values than another group, then its ranks will be lower than the other groups. The nonparametric equivalent of a paired t-test is called the **signed rank test** and works similarly - it compares the ranks of change scores that are negative and the ranks of change scores that are positive.

1.4.2 Tests for comparing categorical data

Studies frequently wish to compare variables that are categorical. This is commonly seen in case-control studies where we wish to see if there is an association between case status (a binary variable) and exposure status (a binary variable). With categorical data tests, we are interested in assessing if two categorical variables (each variable having two or more categories) are independent. Using our Zika example, if birth defect status is independent of Zika exposure status, then there is no link between Zika and birth defects. For categorical data we can set up our data using a **contingency table** that contains the counts that appear in the category combinations. For example, the Zika example results in a 2-by-2 contingency table:

	Zika-positive	Zika-negative
Adverse pregnancy outcomes	58	7
No adverse pregnancy outcomes	67	54

As with continous data, there are parametric tests and nonparametric tests. The **chi-squared test** is a parametric test for categorical data. It assumes a certain distribution of the test statistic (the chi-squared distribution) that holds up as long as there are high enough counts in each cell of the contingency table. An often used rule of thumb is a count of at least 5 per cell. How does the chi-squared test work? It is helpful to look at two quite different contingency tables:

	Disease	No disease
Exposed	20	20
Unexposed	20	20

	Disease	No disease
Exposed	33	7
Unexposed	9	31

In the first table, the counts are evenly distributed between the cells, so there is no sign of an association between exposure and disease. In particular, half of the study population is exposed and half are diseased. So if there is no association, we would expect $0.5 \times 0.5 = 0.25$ of the population to be both exposed and with disease. This is what the first table shows. If there were some association, we would expect either more or less than 25% of the population to be both exposed and with disease. We expect more than 25% if there is some positive association and les sthan 25% if there is some negative association. This is what we see in the second table. It is still the case that half of the study population is exposed and half are diseased, but now more than 25% are both exposed and with disease. This suggests some interaction, some association between exposure and disease. This comparison of observed counts to the counts we would expect with no association is how the chi-squared test works.

A nonparametric version of the chi-squared test is called **Fisher's exact test**. It is similar in spirit to the chi-squared test but does not rely on assuming a distribution for the test statistic. It essentially counts how many tables are more extreme than the actual contingency table. (For example, the second table is "more extreme" than the first.)

1.5 Multiple testing

It is often the case that studies perform many hypothesis tests. This may be because they are looking at many outcomes or covariates. Let's say for example that a team wishes to perform 100 comparisons for their study, all at a type I error rate of 0.05. Say also that they happen to be studing something fruitless - none of their comparisons have a real difference (i.e. the null hypothesis is true for all of them). Just by chance, we expect them to see 5 (100×0.05) comparisons that show a statistically significant result! This is the problem of multiple testing and portrayed well by this xkcd comic.

When scientists perform several hypothesis tests, they can make adjustments to their p-value thresholds for rejecting the null hypothesis. For example, the **Boneferroni correction** divides the typical type I error rate of 0.05 by the number of tests. In this way, the type I error rate of 0.05 is now expected over *all of the tests* instead of just a single test.

Chapter 2

Regression Modeling

Models, generally speaking, are simplified descriptions of the world that distill the workings of a system into its essential parts. We construct models mainly for two purposes: to explain our observations and to predict what will happen in the future using data we have now. Much of public health research has the goal of understanding the relationship between outcomes and exposures or other characteristics. To this end, we will expand our toolbox to include a class of models, called **generalized linear models**, that can handle different types of data we see in public health research. Because common study designs in public health research violate the assumptions for this class of models, we will also look at other modeling strategies that can be used in such cases.

2.1 Regression overview

Regression is a class of techniques that is used to describe outcomes as a function of predictor variables, also called **covariates**. Because there are different types of outcomes, there are also different types of regression methods that are suited for each type. You likely have learned about linear and logistic regression in previous classes. We will review both in the following sections.

Throughout I will use Y to denote the outcome variable and $x_1, \dots, x_p$ to denote p covariates. $E[Y]$ denotes the **expected value** of Y and is another way of writing the *mean* of Y. It describes the average value of the outcome we would see if we had data on a very large number of outcomes. $E[Y]$ is an example of a **parameter** as we talked about in Chapter @ref(chap__hypo__test). It is a true underlying value that describes a characteristic of the population.

It will also be helpful to know the term **linear combination**. If $x_1, \dots, x_p$ are covariate values, then a linear combination of these covariates is written as:

$$LC = a_1x_1 + \cdots + a_px_p$$

where $a_1, \dots, a_p$ are numbers and the result LC is a number as well.

2.2 Linear regression

In linear regression, we want to describe continuous measures as a function of covariates. To be concrete, let's say that our outcome measure Y is the concentration of HIV particles in the blood and that x_1, x_2, x_3 indicate clinical and demographic covariates that might reasonably affect viral particle concentration, say age in years, sex (1 for females, 0 for males), and antiretroviral (ART) drug dosage in milligrams (mg).

In linear regression, we model the expected value of Y as a linear combination of covariates:

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$\beta_0, \beta_1, \beta_2, \beta_3$ are called **regression coefficients** and are simply numbers. **Fitting** a linear regression model is the computational process of estimating the numeric values of the β's (column 1 in the table below).

```
##          term   estimate  std.error  statistic      p.value
## 1 (Intercept)  3.4904393 0.66590221  5.2416695 2.358704e-07
## 2         age  0.3906799 0.01659026 23.5487538 7.255804e-83
## 3         sex  0.1537585 0.36077712  0.4261869 6.701566e-01
## 4         ART -0.1161197 0.07827203 -1.4835399 1.385662e-01
```

2.2.1 Interpreting coefficients

The regression coefficients in front of the covariates ($\beta_1, \beta_2, \beta_3$) have the nice interpretation of being the *expected change in outcome per 1 unit increase in the predictor, holding all other variables constant.* To see this, let's look at age (x_1). We will compare two people who are identical in terms of sex and ART dosage but who differ in age by 1.

Person 1: Age = a, sex = s, ART dosage = d

Person 2: Age = $a + 1$, sex = s, ART dosage = d

We can write the expected HIV particle concentration for each of them:

Person 1: $E[Y_1] = \beta_0 + \beta_1 a + \beta_2 s + \beta_3 d$

Person 2: $E[Y_2] = \beta_0 + \beta_1 (a + 1) + \beta_2 s + \beta_3 d$

The expected change in outcome comparing person 2 to person 1 is $E[Y_2] - E[Y_1]$:

$$E[Y_2] - E[Y_1] = \beta_1$$

Thus, we see that β_1 is the expected change in HIV concentration per year increase in age, holding sex and ART dosage constant.

What about β_0? β_0 is called the **intercept** and represents the expected outcome (mean HIV particle concentration) for a person who has age 0, is male, and has an ART dosage of 0. It is somewhat odd to imagine someone with age 0, so for this reason, predictor variables like age are often **mean-centered**.

$$E[Y] = \beta_0 + \beta_1 (x_1 - \bar{x_1}) + \beta_2 x_2 + \beta_3 (x_3 - \bar{x_3})$$

The numeric values and the interpretations of $\beta_1, \beta_2, \beta_3$ don't change, but the numeric value and interpretation of β_0 will change. It is now interpreted as something more sensible: the expected outcome for someone of average age $\bar{x_1}$, male, and of average ART dosage $\bar{x_3}$.

2.2.2 Interaction

The coefficients in the model we looked at above are all called **main effects**. They describe the effects of covariates holding constant the other covariates. Important to notice in this interpretation is that the effect of some factor is the *same across all individuals.*

Often times, we wish to understand if effects are *different across different groups.* This can be achieved by including interaction terms in a regression model. Most often researchers will include an interaction between a continuous variable and a categorical one or between two categorical variables. For example:

- How does the age effect differ across different socioeconomic categories?
- How does the race effect differ across different countries?

2.3 Logistic regression

While linear regression is used to describe continuous measures as a function of covariates, logistic regression is used to describe the probability of a binary event as a function of covariates. For logistic regression, it is helpful to remember the definition of **odds**. The odds of an event is the ratio of the probability of the event happening to the the probability of the event not happening. If p is the probability of the event happening, then the odds of the event can be written as:

$$\text{odds} = \frac{p}{1-p}$$

In logistic regression, we model the expected log odds of a binary event as a linear combination of covariates:

$$E[\text{log odds}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

We can go through the same process as with linear regression to obtain interpretations of coefficients in logistic regression. In linear regression, coefficients were interpreted as the expected change in outcome per unit change in covariate holding all other factors constant. In logistic regression, we replace "outcome" with "log odds of the binary event." So in logistic regression, the coefficients give the difference in log odds of the event, more commonly expressed as the log odds ratio. For example, if the binary event were disease, the interpretation of an age coefficient would be

$$\log\left(\frac{\text{odds of having disease at age a+1}}{\text{odds of having disease at age a}}\right)$$

In publications, these are almost always presented in exponentiated form:

$$\frac{\text{odds of having disease at age a+1}}{\text{odds of having disease at age a}}$$

2.4 Generalized linear models

With linear and logistic regression we have dealt with continous and binary outcomes respectively. What about categorical outcomes with more than two categories, which are common in surveys and medical situations where severity is an outcome? What about count data, which arise commonly in public health via incidence rates?

There is a class of models in statistics called **generalized linear models** that allows a variety of outcome variables to be used as the dependent variable - not just continuous and binary as we have seen with linear and logistic regression, but also count and categorical data with more than two outcomes.

Generalized linear models have the general form:

$$f(E[Y]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In words, the left side $f(E[Y])$ just describes some function of the expected outcome. The right side is the familiar linear combination of covariates.

For linear regression, the function f is the identity function $f(x) = x$ and $E[Y]$ indicates the expected/mean outcome which depends on the covariates.

For logistic regression, the function f is the logit function:

$$f(E[Y]) = \log\left(\frac{E[Y]}{1 - E[Y]}\right)$$

In logistic regression, the outcome Y is the binary 0/1 indicator of group and $E[Y]$ indicates the probability of being in group 1 (e.g. the probability of being a case in a case-control study). So in more familiar form, the above expression is the log odds of being in group 1:

$$f(p) = \log\left(\frac{p}{1 - p}\right)$$

Similar ideas apply for when Y is a different type of outcome variable (e.g. general categorical, count). The main idea with generalized linear models is that it is possible to perform regression for a wide variety of outcome variables.

2.5 Model selection

What variables should be included in a regression model? It is tempting to just include them all and let the computer do the work, but we need to be careful. Say we have 100 subjects and 50 predictors. There is no way we have enough information with only 100 subjects to get reliable information on all of those predictors. This is an idea called the curse of dimentionality: the more variables we have for a fixed size dataset the less information we have to learn about each variable. Model selection techniques are used to identify the most relevant variables in explaining an outcome. If we are able to identify the most important variables, we can use this subset for regression. We will discuss techniques for this in class.

Chapter 3

Causal Inference: Background

Up until this point we've been looking at various statistical tools that can be used to study the relationship between outcomes and covariates. These relationships, also called associations, are interesting but not what we fundamentally care about. Ultimately science cares about identifying causes of an outcome, but unfortunately, this is a very difficult endeavor in health research. Why is it so difficult to determine if some exposure truly **causes** an outcome? To be confident in making such a strong claim, we would ideally need to compare many people who have had the exposure to many people who have not and who are otherwise identical. This is difficult to impossible in most settings! In this chapter, we are going to learn about tools and study designs that researchers use to attempt to get closer to causal effects as opposed to just associations. These tools and study designs fall under a general biostatistical area known as **causal inference.**

3.1 Bradford Hill criteria

In this chapter, we are going to be learning about study designs and techniques that have statistical tools associated with them. Before we jump into that content, it is worth being aware of a set of causality criteria known as the Bradford Hill criteria[1]. These criteria were set forth by a statistician named Sir Austin Bradford Hill in 1965 to provide some useful guidelines for assessing if a causal claim is supported by scientific literature. This is typically used in reviewing a large body of observational studies and lends itself to more qualitative discussions. The criteria he propsed are as follows:

1. **Strength**: causal effects should be strong or noticeable in some sense (large effect size)
2. **Consistency**: causal effects should be able to be shown across multiple studies (reproducibility)
3. **Specificity**: causal effects should be specific to a certain system, population, environment
4. **Temporality**: a cause should occur before any effects
5. **Biological gradient**: higher levels of the causal agent should result in higher intensities of the outcome (dose-response relationship)
6. **Plausibility**: causal effects should be supported by mechanistic explanations (answers: how?)
7. **Coherence**: epidemiological results and laboratory results should support each other
8. **Experiment**: is there experimental evidence where the supposed causal agent was intervened upon?
9. **Analogy**: is the supposed causal agent similar to another agent which we strongly believe to be causally related to an outcome?

As an example, there have been two recent papers[2][3] using the Bradford Hill criteria to review the body of literature studying the association beteen Zika virus infection in pregnant mothers and birth defects. Feel free to look at how these authors interpreted and applied these criteria for this situation.

[1] Wikipedia

[2] Zika Virus and Birth Defects - Reviewing the Evidence for Causality

[3] Does Zika Virus Cause Microcephaly - Applying the Bradford Hill Viewpoints

3.2 Exercise: comparing two statisticians

Before we dive into new material, let's consider the following scenario. University officials are interested in determining if the food served at the main cafeteria is causing weight gain in the freshman class. Specifically they want to know whether this weight gain is different in men and women. They have data on the September and May weights of all members of last year's freshman class. Two statisticians are called in to conduct an analysis.

Statistician A computes the weight change from September to May for each freshman and separates these differences by sex. He compares these weight changes in males and females using a t-test. He does not find a statistically significant difference between men and women and concludes that there is no evidence that the cafeteria is causing differential weight gain in male and female freshmen.

Statistician B models weight change as a function of September weight and gender using linear regression. He finds a statistically significant difference in weight changes between male and female freshmen.

3.3 Rubin causal model

The situation above is known as Lord's paradox. What made it tricky for the two statisticians? We know that the university did not provide any information about student weight outcomes in the absence of the cafeteria, but the two statisticians seem to have forgotten about this. The university is interested in the *causal effect of the cafeteria diet* and wants to compare this causal effect between men and women.

When thinking about our scientific questions that attempt to get at causal effects, it is useful to have some sort of structure, some sort of guidelines in a sense, that helps us formulate the information we need. The **Rubin causal model** is one such framework that is widely used in the field of causal inference. It contains three components:

- Treatment
- Units
- Potential outcomes

A **treatment** is some program, consumable item, or the like, that can be administered to participants in the study. These participants are called **units**. A key idea is that the treatment must be able to be administered as an intervention. This is of practical importance because if the treatment does actually *cause* some change in outcomes of interest, it is useful for us to be able to intervene on the treatment (by increasing it or removing it, say) to affect outcomes in the real world. In the table below, the first column lists some general concepts that we might be interested in, the second lists typical treatments that we might think about but that are not suitable treatments in this framework, and the third column lists suitable formulations of the treatment.

Idea	Non-examples	Examples
BMI	"Having a high BMI"	Weight loss program
Sex	"Being female"	Gender partial policies
Preferences	"Liking Pepsi"	Advertising campaigns for Pepsi

We often think about causal questions such as, "Does having a high BMI cause an increased risk for heart disease?" It is a relevant and important scientific question, but it does not provide a good example of a treatment in this framework because it does not provide us a means of *studying* the effect of BMI. In a little bit, we will get to the idea of **assignment mechanisms** for treaments. An assignment mechanism tells us how different units are assigned to treatment or control and are used in the mathematics of estimating treatment effects. Being able to be assigned to treatment or control necessitates that the treatment be one that can be administered as an intervention, which is not true for "having a high BMI." Rather, we can assign people to diet and exercise programs which lead them to change their BMI.

Units, as we discussed above, are the entities (person, place, or thing) to which we administer treatment or withold it (control). A person, place, or thing at two different times is considered as two different units. So, yourself at this moment and yourself 10 days from now are considered different units. This is sensible given that you're going to change, albeit slightly, in the next 10 days.

Potential outcomes are outcomes that could be observed for each unit under the different levels of the treatment (typically two levels). In other words,

- $Y_i(0)$ is the potential outcome for study unit i under the control condition.
- $Y_i(1)$ is the potential outcome for study unit i under the treatment condition.

If a study unit receives a particular level of the treatment in real life, then the outcome under the other level of the treatment is called the **counterfactual** outcome. We don't get to observe this counterfactual outcome! So if a unit received the treatment in real life, their counterfactual outcome would be the outcome under the control condition. Think of potential outcomes as splitting the world into two parallel universes

If we knew a unit's potential outcome under control $Y(0)$ and under treatment $Y(1)$, then we could simply subtract the two quantities to obtain the causal effect of the treatment for that unit. Easy, right? We can just observe the unit under the control condition first and then the treatment condition (or vice-versa) to obtain these two potential outcomes. But wait! Recall that a person, place, or thing at two different times is considered as two different units. This is the case here - after we observe the unit under the control condition, it becomes a different unit which we then observe under the treament condition. The two outcomes we observe are no longer directly comparable. This leads us to the **fundamental problem of causal inference**: we can only ever observe one potential outcome for each unit. This problem is illustrated in the table below.

Units	$Y_i(1)$	$Y_i(0)$	T_i
1	30	?	1
2	?	15	0
.	.	.	.
10	22	?	1

We see that unit 1 was exposed to the advertising campaign and has the potential outcome under the treatment condition filled in. Conversely, unit 2 was not exposed to the ad campaign and has the potential outcome under the control condition filled in. Ideally, we would subtract column three from column two and average these resulting differences to estimate the causal effect of a Pepsi advertising campaign on dollars spent on Pepsi per week.

3.3.1 Revisiting Lord's paradox

We can frame the situation using the Rubin causal model as indicated in the table below. To understand the causal effect of university diet on end-of-year weights, our potential outcomes $(Y(0), Y(1))$ must be May weights under treatment (cafeteria diet) and under control (no cafeteria diet). It is clear when we write it this way that there was no control group. Then how are we to estimate causal effects? The two statisticians implicitly made assumptions about the counterfractual outcome without realizing it.

Units	Sex, Sept. weight	$Y_i(0)$	$Y_i(1)$	T_i
1	M,140	?	141	1
2	F,120	?	119	1
.	.	.	.	.
N	M,150	?	144	1

Statistician 1 assumed that the causal effect of the cafeteria diet for all freshmen was just the September to

May weight change. In other words, he assumed that the potential outcome under control, $Y(0)$, was simply that freshman's September weight.

Units	Sex, Sept. weight	$Y_i(0)$	$Y_i(1)$	T_i
1	M,140	140	141	1
2	F,120	120	119	1
.	.	.	.	.
N	M,150	150	144	1

Statistician 2 used the following regression for his analysis of the data under treatment (university diet) and assumed that it also held under control (no university diet).

$$E[\text{weight change}] = \beta_0 + \beta_1 \text{sex} + \beta_3 \text{Weight}_{\text{Sept}}$$

$$E[\text{June weight under control}] = \beta_0 + \beta_1 \text{sex} + \beta_3 \text{Weight}_{\text{Sept}} + \text{Weight}_{\text{Sept}}$$

Units	Sex, Sept. weight	$Y_i(0)$	$Y_i(1)$	T_i
1	M,140	$\beta_0 + \beta_1 + 140(\beta_3 + 1)$	141	1
2	F,120	$\beta_0 + 120(\beta_3 + 1)$	119	1
.	.	.	.	.
N	M,150	$\beta_0 + \beta_1 + 150(\beta_3 + 1)$	144	1

3.4 Types of causal effects

There are two types of causal effects that are usually estimated in causal inference studies: the **average treatment effect (ATE)** and the **average treatment effect for the treated (ATE)**.

Average treatment effect (ATE): average effect for everyone in the population

$$ATE = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

Average treatment effect for the treated (ATT): average effect for only those treated

$$ATT = \frac{1}{N} \sum_{i \in T} (Y_i(1) - Y_i(0))$$

where T is the set of people who received the treatment ($T_i = 1$)

Let's look at two examples to see how the ATE and ATT differ.

We see from these examples that ATE is a useful estimate for neutral or beneficial interventions because there is no problem with theoretically administering the treatment to the entire population. The ATT generally applies to harmful interventions or interventions that we would not want our broader population of inference to experience.

3.5 How do we learn about causal effects?

So far we have set up the Rubin causal model as a framework for specifying the study of causal effects. It is useful for thinking carefully about what our treatment/intervention should be and who will receive the treatment and control conditions. There are few more ideas that we need in order to use this thinking in practice for real studies.

- Replication
- Stable Unit Treatment Value Assumption (SUTVA)
- Assignment mechanism

3.5.1 Replication

We need data on multiple units where there is a mix of units receiving the treatment and control conditions. On the surface, this is perhaps obvious, but it was not so clear in our example on Lord's paradox. In that example, there were multiple units under the treatment (cafeteria food) condition but none under the control condition. It is not uncommon in scientific investigations for all subjects to only receive the treatment, yet investigators wish to use this data to estimate the (causal) effect of the treatment.

3.5.2 Stable Unit Treatment Value Assumption (SUTVA)

This assumption has two parts:

1. No interference between units: treatment assignment of one unit does not affect potential outcomes of another unit. e.g. One person's drug use (treatment) doesn't affect someone else's outcomes.
2. Treatment only has one version and control only has one version. e.g. "Heavy" drug use doesn't have distinct subcategories such as "very heavy" or "moderately heavy." "Low" drug use doesn't have distinct subcategories such as "very low" or "moderately low."

What is useful about having these assumptions? The second part is easier to tackle - if the treatment and/or control has more than one version, then comparisons between the treatment and control groups become murky. For example, comparing heavy drug users to low drug users is unclear if both categories have gradations. Am I obtaining the effect of high-ish drug use versus low-ish drug use?

The first part of the assumption is for tractability of the analysis. Imagine that I have just two people in my study: Alice and Bob. The outcome of interest is liver function as a score from 1 to 100. If Alice and Bob never interact, then the potential outcomes might look like this:

Units	Heavy drug use	Low drug use
Alice	10	50
Bob	14	48

Both Alice and Bob have potential liver function scores under the heavy and low drug use conditions. If, however, Alice and Bob had the chance to interact, it is reasonable that they might influence each other's outcomes (say by encouraging poor behavior). Then the potential outcomes become more complicated:

Units	Alice: heavy, Bob: heavy	Alice: heavy, Bob: low	Alice: low, Bob: heavy	Alice: low, Bob: low
Alice	10	10	50	50
Bob	8	48	48	14

When there was no interference, Alice only had two potential outcomes. Now with the potential for interference, Alice's two potential outcomes now also depend on Bob's drug use which increases the number of potential outcomes to 4. The same happens for Bob. With more subjects in the study, the number of potential outcomes increases exponentially. We then must ask ourselves: Which columns do we compare to estimate causal effects? Do we compare all possible combinations? Answering such questions is not easy, so we need to be careful in designing, conducting, and evaluting studies to be sure that both parts of SUTVA holds. In summary:

- If part 1 of SUTVA could be violated, we must question the purity of the comparison being made. What exactly is meant by treatment group versus control group if there are gradations of treatment and control?
- If part 2 of SUTVA could be violated, we must also question the purity of the comparison being made. With interference, the treatment group now has gradations such as "treatment with strong influence from control group" and "treatment with moderate influence from control group." In a way, part 1 of SUTVA is being violated here as well.

3.5.3 Assignment mechanism

An assignment mechanism is a process that determines which treatment each unit receives. The idea of an assignment mechanism generalizes the idea of selection bias. Why is this? Say for example that the treatment of interest is the flu shot. Selection bias could occur if unhealthier people tend to not have access to clinics where they can obtain the flu shot. Thus, any comparison of flu outcomes in individuals with and without the flu shot is going to give biased results. In particular, we are likely to overestimate the benefit of the shot because we would be comparing healthy people who get the shot to unhealthy people who didn't get the shot. This selection bias is an example of a **confounded assignment mechanism**. Assignment into the treatment or control group is confounded by underlying health status. Useful comparisons of the treatment and control group cannot be made because of this confounding. An example of an **unconfounded assignment mechanism** occurs with randomized controlled trials in which people are randomly assigned to treatment and control groups. This assignment is unconfounded because we don't expect any covariate differences between the groups due to the randomization. More on randomized trials will be discussed in the next chapter.

Study designs and techniques in causal inference try to capitalize on some feature of the world that allows for knowledge about the assignment mechanisms. If we have knowledge of the assignment mechanism, we can try to counteract it to still obtain useful estimate of causal effects.