BIOSTATISTICS 701: Introduction to Biostatistics Theory and Methodology

Contents

Module 1: Random Variables

1. Video

Click here to watch the video.

2. Background Information

Data generating mechanism

An "experiment" is an action or event with an observable result. We deal with experiments whose results aren't fully predictable. Those results are described through random variables. We sometimes call the process by which random variables are created the data generating mechanism.

Sample Space

The set of all possible outcomes of an experiment is called the sample space. Outcomes can be:

- Discrete and finite (e.g., "democrat", "independent", "other", "republican")
- Discrete and countably infinite (e.g., 0, 1, 2, . . .)
- Continuous (e.g., systolic blood pressure, also known as SBP)

There are nuances:

- Continuity is an abstraction, since we only observe things to a specified level of precision
- Discrete and finite and discrete and infinite are usually lumped together into discrete – so the main classification is discrete versus continuous
- Discrete can also include ordered (e.g., small, medium, large), where ordered variables are often treated as a special case
- Discrete, ordered outcomes with large numbers of categories (e.g., a scale with integer values from 0-100) are often treated as if they are continuous

Studies usually contain multiple individuals, and so a study generates observed values of random variables for each individual. When we are considering the possibility of multiple similar studies, a single study (with observed values for random variables for multiple individuals) is termed a "realization" of that study.

As notation, if the random variable is named Y, when we are speaking about it in general we denote it by Y, whereas what we observe is $Y = y$. For example, if we observe SBP for a single individual, $Y = 140$.

For our purposes, it is usually sufficient to discuss the values of a random variable for a single individual.

However, if it is necessary to distinguish a SBP of 140 from participant 1 from a SBP of 150 from participant 2, we use the notation $Y_1 = 140$ and $Y_2 = 150$.

Events

As noted, the set of possible outcomes for a RV is termed the sample space. For example, the sample space for one roll of a die is $\{1, 2, 3, 4, 5, 6\}$. For each individual, we observe one member of the sample space (i.e., no more than one, no less). Thus, the elements of the sample space are mutually exclusive and exhaustive.

Events are subsets of the sample space. For example, the subset $\{2, 4, 6\}$ is a proper subset, and this event can be labelled {rolling an even number}. It's OK for an event to include the entire sample space, in which its certain to occur, or none of the elements of the sample space, in which case it's certain not to occur. Of course, the interesting cases fall between these two extremes.

3. Problem Set

Problem I

The probability that a fair die will land on any number is 1/6 (so, for example, $Pr\{Y = 3\} = 1/6$. Calculate the probability for these events:

- Y = any of $\{1, 2, 3, 4, 5, 6\}$
- Y = not any of $\{1, 2, 3, 4, 5, 6\}$
- Y = even number $\{2, 4, 6\}$
- Y = less than 3 $\{1, 2\}$
- Y = prime number $\{1, 2, 3, 5\}$
- Y = not a prime number $\{4, 6\}$
- Y = prime and even number $\{2\}$

Problem II

Health status, denoted by Y, can be classified as terrible, poor, fair, good or excellent.

i. What type of random variable is Y: discrete with a finite number of possible values, discrete with a countably infinite number of possible values, continuous.

ii. Is Y ordinal?

iii. Suppose the probability that Y falls into various categories is given by the table below:

Category	Probability
Terrible	.05
Poor	.10
Fair	.15
Good	.25
Excellent	.45

Calculate the probability for these events:

- Y is excellent
- Y is good
- Y is good or better
- Y is good or worse
- Y is poor or worse

Module 2: Bernoulli Random Variables

1. Video

Click here to watch the video.

2. Example SAS Program

On Sakai, you should find "module 2 SAS program.SAS"

3. Background

At this point, a typical inference course might digress into a deep discussion of set theory, but let's instead take a detour into a simple and important distribution: namely, the Bernoulli.

The Bernoulli distribution has only 2 possible outcomes; for example, $\{live, die\}$, $\{good, bad\}$, $\{win, lose\}$, etc. It's useful to denote these possible outcomes by $\{0, 1\}$. $Y = 1$ is often called a success with $Y = 0$ termed a failure. It doesn't really matter which outcome you assign to 1 and which you assign to 0 — you'll get the same result (once you translate from the mirror image). Often, there's a natural way to assign the 0 and 1 which will assist in communication with the investigators.

One part of the definition of a Bernoulli random variable (RV), and, indeed, of any RV, is its set of possible outcomes — here, this set is $\{0, 1\}$. As noted above, for every individual one and only one of these outcomes will occur. A Bernoulli RV is discrete with a finite number of possible values (i.e., 2).

The other part of the definition is a formula/rule which assigns a probability to each possible value (for now, probability will be loosely defined). Denote the probability of a success, namely, the probability that $Y = 1$, by θ. Thus, in symbols: $Pr\{Y = 1\} = \theta$. Since the event in question either will or won't occur, it follows that $Pr\{Y = 0\} = 1 - \theta$ (in other words, the probabilities associated with all the elements of the sample space sum to 1, and $\theta + (1 - \theta) = 1$.

The table below represents a complete description of a Bernoulli RV:

Possible Value	Probability
0	$1 - \theta$
1	θ

The second column of this table is called the probability mass function (PMF). In somewhat more formal

terms, the PMF is $Pr\{Y = y\}$ for all possible values of Y, and 0 otherwise.

The PMF provides equivalent information as the cumulative distribution function (CDF). Formally, the CDF is the $Pr\{Y \leq y\}$ for all values of $Y = y$. For discrete RVs it is a step function, where the steps occur at the possible values of Y (for contiguous RVs, the CDF has an equivalent definition, but isn't a step function).

The following table summarizes the CDF of a Bernoulli RV:

Y	CDF
$Y < 0$	0
$0 <= Y < 1$	$1 - \theta$
$Y >= 1$	1

We'll go into more detail about this later, but for now we will go ahead and use the PMF to calculate the mean (which measures central tendency) and variance (which measures spread) of Y. The table below summarize the calculation for the mean.

Possible Value of Y	$Pr\{Y = y\}$	$Y * Pr\{Y = y\}$
0	$1 - \theta$	$0 * (1 - \theta) = 0$
1	θ	$1 * \theta = \theta$
sum		$0 + \theta = \theta = mean$

So, the mean of Y is θ. More generally, to obtain the mean, take the sum of $[Y * Pr\{Y = y\}]$ for all possible values of Y. This can be understood to be a weighted sum of the possible values of Y, where the weights are $Pr\{Y = y\}$.

The variance of Y is another weighted sum: as above, the weighting factor is $Pr\{Y = y\}$ and the quantity to be summed is $(Y - Y_m)^2$, where Y_m denotes the mean of Y: here, $Y_m = \theta$. The table summarizes the calculation for the variance.

Possible Value of Y	$Y - Y_m$	$(Y - Y_m)^2$	$Pr\{Y = y\}$	$(Y - Y_m)^2 * Pr\{Y = y\}$
0	$0 - \theta = -\theta$	$(-\theta)^2$	$1 - \theta$	$\theta^2 * (1 - \theta)$
1	$1 - \theta$	$(1 - \theta)^2$	θ	$(1 - \theta)^2 * \theta$
sum				$\theta * (1 - \theta) = variance$

So, after applying a little bit of algebra to simplify the sum, the variance of Y is $\theta * (1 - \theta)$. The standard deviation is the square root of the variance.

Individual Bernoulli random variables are useful in their own right. For example, when writing a program in R, a logical condition taking the values TRUE or FALSE is a Bernoulli random variable, with the probability of TRUE being θ. We are also interested in sequences (sets) of Bernoulli random variables. In the standard case, the probability of success is assumed to be the same across all individuals, and the results are assumed to be independent across individuals. In other words, the Bernoulli random variables within the sequence are assumed to be independent and identically distributed. The statistical literature does not always make an explicit distinction between the individual random variable and the sequence of such variables. Instead, it relies upon the context. For the standard version of a Bernoulli random variable, the probability of success is assumed to be the same across all individuals.

Here, θ is the single parameter of a Bernoulli distribution. Once we know θ, we know everything we possibly can about that distribution. Moreover, even though the value of θ might be 0.80 in one application and 0.47 in another, all Bernoulli distributions have the same essential structure. Once you've decided that you're working with a Bernoulli distribution, the next step in an actual analysis is to use the data to make an educated guess about the value of θ and also the degree of precision you can attach to that guess.

In summary, the 3 conditions for a standard sequence of Bernoulli RVs are:

- 2 possible outcomes {success,failure}
- Results are independent across individuals
- Probability of success is the same for every individual

4. Problem Set

Note: You are welcome to make your calculations in R, SAS or another language. The video describes various programming techniques using SAS.

Problem I

i. Simulate 1 study, with $n = 10$ individuals, having Y as a Bernoulli variable with $\theta = 0.55$. Use your software to calculate the mean and variance of Y. How closely does this match the formulas for the actual mean and variance?

ii. Repeat using $n = 100$, then $n = 1000$. Do the mean and variance get closer to the correct values?

Problem II

i. Simulate another study, with $n = 1,000$ individuals, having Y as a Bernoulli variable with $\theta = 0.55$. Calculate the mean and variance. How much did the answer differ from the previous simulation?

ii. Repeat the simulation, but for each individual replacing $\theta = 0.55$ with a random draw from a uniform distribution on the interval [0.45,0.65]. Calculate the mean and variance. (Note: The uniform distribution will be described in detail later in the course. For now, apply the appropriate R function, $runif()$)

iii. Repeat the simulation, but for each individual replacing $\theta = 0.55$ with a random draw from a uniform distribution on the interval [0.25,0.85]. Thus, θ will be even more variable than before. Calculate the mean and variance. When comparing the results of parts $i., ii.$, and $iii.$, what is happening to the variance of Y (i.e., is it increasing, decreasing, or constant)? Why?

Problem III

Simulate 1,000 studies, each with $n = 100$ individuals, having Y as a Bernoulli variable with $\theta = 0.55$. For each study, calculate the sum of Y. The result is a binomial distribution. Create a histogram of this binomial random variable (the histogram should be based on 1,000 observations).

Module 3: Joint and Conditional Probabilities

1. Video

Click here to watch the video.

2. Some Axioms

We now consider a limited group of axioms from set theory and probability. Among others, we're ignoring cardinality, De Morgan's laws, and various techniques used to deal with events and their probabilities for countably infinite sets.

- When considering two sets, a union (denoted by $A \cup B$) is defined as the objects in either set.
- When considering two sets, an intersection (denoted by $A \cap B$) is defined as the objects in both sets.
- The complement of a set (denoted by A^c) is all objects not in that set.

A probability, P, is a function with particular restrictions:

- Probabilities fall within [0,1]
- $Pr\{\text{null set}\} = 0$
- $Pr\{\text{entire sample space}\} = 1$
- $Pr\{A^c\} = 1 - Pr\{A\}$
- $Pr\{A \cup B\} = Pr\{A\} + Pr\{B\} - Pr\{A \cap B\}$
    - This can also be written as $Pr\{A \text{ or } B\} = Pr\{A\} + Pr\{B\} - Pr\{A \text{ and } B\}$
- If events are mutually exclusive (termed "disjoint") then $Pr\{A \text{ and } B\} = 0$, and so $Pr\{A \cup B\} = Pr\{A\} + Pr\{B\}$
- Since the components of the sample space are mutually exclusive, the probability of an event is the sum of the component probabilities
    - For example, $Pr\{\text{die is even}\} = Pr\{Y = 2\} + Pr\{Y = 4\} + Pr\{Y = 6\}$

3. Joint Probability

The joint probability of two events A and B is the probability that A and B both occur.

4. Conditional Probability

The conditional probability of A given B, denoted by $Pr\{A \mid B\}$, is

$$\frac{Pr\{A \cap B\}}{Pr\{B\}}$$

assuming that $Pr\{B\} > 0$. This is can also be written as

$$Pr\{A \mid B\} = \frac{Pr\{A \text{ and } B\}}{Pr\{B\}}$$

As covered in the PPV exercise, these two approaches are equivalent:

1. Consider the original sample space, and use the formula

$$Pr\{A \mid B\} = \frac{Pr\{A \text{ and } B\}}{Pr\{B\}}$$

2. Or, create a new sample space limited to B, and recalculate the probability of A within that new sample space

5. Total Probability

If the possible values of X are $\{0, 1\}$, then $Pr\{Y = 1\} = Pr\{Y = 1 \text{ and } X = 0\} + Pr\{Y = 1 \text{ and } X = 1\}$. This idea extends to more than 2 values of X, so long as they are disjoint.

6. Bayes' Theorem

Bayes' Theorem is directly derived from the above. One version of the theorem is:

$$Pr\{Y = 1 \mid X = 1\} = \frac{Pr\{Y = 1\} * Pr\{X = 1 \mid Y = 1\}}{Pr\{Y = 0\} * Pr\{X = 1 \mid Y = 0\} + Pr\{Y = 1\} * Pr\{X = 1 \mid Y = 1\}}$$

7. Example: Joint and Conditional Probability, Bayes' Theorem, Discrete Case

Consider the following table, which cross-classifies 300 patients according to the results of a diagnostic test (X) and the presence of a disease (Y). The observed data are italicized, other quantities are derived by summation.

	$Y = 0$: disease absent	$Y = 1$: disease present	Marginal distribution of test results
$X = 0$: test is negative	*90*	*40*	130
$X = 1$: test is positive	*10*	*160*	170
Marginal distribution of disease	100	200	Total=300

For example, if a patient is randomly selected from this population, the probability that the disease is present is 200/300=0.67.

"Joint" events consider both X and Y, whereas "marginal" events consider one but not the other. The probabilities for the 4 joint events are:

- $Pr\{X = 0 \text{ and } Y = 0\} = 90/300$
- $Pr\{X = 0 \text{ and } Y = 1\} = 40/300$
- $Pr\{X = 1 \text{ and } Y = 0\} = 10/300$
- $Pr\{X = 1 \text{ and } Y = 1\} = 160/300$

The 4 joint events are mutually exclusive and exhaustive, and thus constitute a sample space (i.e., the set of all possible outcomes). The probabilities for the 4 marginal events are:

- $Pr\{X = 0\} = 130/300$
- $Pr\{X = 1\} = 170/300$
- $Pr\{Y = 0\} = 100/300$
- $Pr\{Y = 1\} = 200/300$

The probabilities for the marginal events are simply obtained by using the information in the margins of the table (i.e., the numbers that are not italicized). However, they can also be constructed using the joint probabilities. For example:

$$Pr\{X = 1\} = Pr\{X = 1 \text{ and } Y = 0\} + Pr\{X = 1 \text{ and } Y = 1\} = \frac{10}{300} + \frac{160}{300} = \frac{170}{300}$$

This illustrates the law of total probability, which holds when the events $Y = 0$ and $Y = 1$ are disjoint.

Conditional probabilities, like $Pr\{Y = 1 \mid X = 1\}$, can be calculated in two ways. One approach is to treat those individuals with $X = 1$ as a new population of 170 individuals and notice that 160 of them have $Y = 1$, implying that the desired probability is 160/170. Another approach uses the joint probabilities and the formula

$$Pr\{Y = 1 \mid X = 1\} = \frac{Pr\{X = 1 \text{ and } Y = 1\}}{Pr\{X = 1\}}$$

which in turn equals

$$\frac{160/300}{170/300}$$

which is identical to 160/170 after cancellation. The second approach illustrates the law of conditional probability.

We could apply the law of total probability to the denominator and obtain:

$$Pr\{Y = 1 \mid X = 1\} = \frac{Pr\{X = 1 \text{ and } Y = 1\}}{Pr\{X = 1\}} = \frac{Pr\{X = 1 \text{ and } Y = 1\}}{Pr\{Y = 0 \text{ and } X = 1\} + Pr\{Y = 1 \text{ and } X = 1\}}$$

which equals

$$\frac{160/300}{10/300 + 160/300}$$

the same answer as before.

Finally, we can rearrange

$$Pr\{Y = 1 \mid X = 1\} = \frac{Pr\{X = 1 \text{ and } Y = 1\}}{Pr\{X = 1\}}$$

and plug this into the formula:

$$Pr\{Y = 1 \mid X = 1\} = \frac{Pr\{Y = 1\} * Pr\{X = 1 \mid Y = 1\}}{Pr\{Y = 0\} * Pr\{X = 1 \mid Y = 0\} + Pr\{Y = 1\} * Pr\{X = 1 \mid Y = 1\}}$$

which equals:

$$\frac{200/300 * 160/200}{100/300 * 10/100 + 200/300 * 160/200}$$

the same answer after cancellation. This final version of the formula is one version of Bayes' Theorem.

8. How This is Used in Statistics

Having multiple versions of the same formula might seem to be more trouble than it is worth, but the advantage is that you might have one set of probabilities but want the other. For example, a patient has a positive test and wants to know the probability that they have the disease. The information we actually have pertains to the disease's prevalence (i.e., $Pr\{Y = 1\}$) and the operating characteristics of the test: namely, $Pr\{X = 1 \mid Y = 1\}$ and $Pr\{X = 1 \mid Y = 0\}$. These operating characteristics might have been derived from a separate experiment whereby, for example, a sample of patients with the disease are given the test and a sample of patients without the disease are given the test.

Bayes' Theorem allows us to derive what we want from what we have, and we could tell the patient with a positive test that their chance of having disease is 160/170. Indeed, one way of quantifying the impact of a positive test is that it changed the physician's estimate of the probability of disease from the baseline value of $200/300 = 0.67$ to $160/170 = 0.94$. Similarly, a negative test would change the estimate of probability of disease from 0.67 to $40/130 = 0.31$.

Bayes' Theorem also allows us to make the same calculation for a hypothetical population with a different prevalence of disease, in similar fashion.

9. Practice Problems

Problem I

Consider a diagnostic test with 95% sensitivity and 80% specificity, and a disease with 50% prevalence. If the test for a particular patient is positive, what is the probability that they have the disease (this quantity is called the positive predictive value)? (Hint: To answer this question, the sample size doesn't matter. For concreteness: set the number of patients with disease to 100, the number of patients without the disease to 100, and use the table below.)

	$Y = 0$: disease absent	$Y = 1$: disease present	Marginal distribution of test results
$X = 0$: test is negative	*90*	*5*	95
$X = 1$: test is positive	*10*	*95*	105
Marginal distribution of disease	100	100	Total=200

Problem II

Continuing to assume that the diagnostic test has 95% sensitivity and 80% specificity, suppose that the disease prevalence is 1%.

i. What is the positive predictive value?

ii. Based on that result, what is the problem with screening for rare diseases?

Problem III

Now, consider a table with an essentially identical structure, but with the variables renamed:

	$Y = 0$: no actual effect	$Y = 1$: actual effect	Marginal distribution
$X = 0$: observed $p < .05$	A	C	
$X = 1$: observed $p \geq .05$	B	D	
Marginal distribution			

i. The "type 1 error rate", α, is the probability of a statistically significant result ($p < .05$) when there is no actual effect. For example, for a t-test, the type 1 error rate is the probability of a statistically

significant result when the true, but unable to observe, group means are identical. In terms of A, B, C, and D, what is the type 1 error rate?

ii. The type 2 error rate, β, is the probability of a non-significant result ($p \geq .05$) when there is an actual effect ("statistical power" is $1 - \beta$). Suppose that the investigator is on a fishing expedition, and the probability that, for any particular statistical test, there is an actual effect of 1%. Also assume that $\alpha = .05$ and $\beta = .20$. The investigator observes a statistically significant result. What is the probability that this result is real?

Module 4: Independence

1. Video

Click here to watch the video.

2. Explanation

Independence has a substantive meaning and a mathematical meaning, with the latter being a representation of the former. The substantive meaning is that the events in question are physically and/or causally unrelated. If the science says "unrelated," the statistics can usually say "independent." Indeed, when independence is assumed, rather than demonstrated, it is because of the science.

The mathematical definition of independence is $Pr\{A \mid B\} = Pr\{A\}$, which should be intuitive, because if B and A are unrelated, knowing the value of B has no impact on the probability of A.

In general, $Pr\{A \text{ and } B\} = Pr\{B\} * Pr\{A \mid B\}$. If independence holds, $Pr\{A \text{ and } B\} = Pr\{A\} * Pr\{B\}$. This is a very useful result, because it allows you to calculate a joint probability only using marginal probabilities (i.e., probabilities associated with one event at a time).

Independence is sometimes assumed, and at other times it is assessed. As an example of assuming independence, consider 3 tosses of a fair coin, which is a form of sampling with replacement. The tosses are unrelated to one another, and so we can assume independence. The probability of 3 consecutive heads can be denoted as $Pr\{\text{coin 1=H and coin 2=H and coin 3=H}\}$, which, by independence, is simply the product of the probabilities of the 3 events: $Pr\{\text{coin 1=H}\} * Pr\{\text{coin 2=H}\} * Pr\{\text{coin 3=H}\} = .5 * .5 * .5 = .125$.

Card problems typically assume sampling without replacement, because you are dealt one card after another, and the relevant construct is sometimes called conditional independence. For example, we can calculate the probability of drawing 2 consecutive diamonds as follows:

> A deck of cards contains 52 cards, 13 of which are diamonds. The probability of drawing the first diamond is 13/52. Conditional on this happening, the new deck has 51 cards, 12 of which are diamonds. The probability of drawing the second diamond is 12/51. The probability of both cards being diamonds is $13/52 * 12/51$. It doesn't matter which diamond happens to have been selected first.

As a tiny digression around card problems and the like, they fall within the branch of statistical inference where the outcome is discrete and countable (e.g., the first draw has exactly 52 possibilities), where every

element of the sample space has equal probability. In that case, the probability of an event equals

$$\frac{\text{number of elements of sample space with the event}}{\text{number of elements of sample space}}$$

For example, to solve the above card problem using counting, we would generate the denominator by counting the number of possible 2–card hands; we would generate the numerator by counting the number of possible 2–card hands with 2 diamonds, etc.

Traditional inference courses go into detail about all this counting, and the resulting combinatorial results are used in the derivation of the binomial and hypergeometric distributions, among others. However, this information is not crucial to the daily practice of biostatistics, and thus in this course we treat most of the material pertaining to counting as supplemental. For our purposes, we simply assign probabilities to elements of the sample space, whether or not those probabilities are identical, and work with those probabilities without treating equally likely outcomes as separate cases.

For completeness, the number of ways to produce a sample size k from a population of size n is:

- For sampling with replacement: n^k
- For sampling without replacement: $n(n-1)(n-2)...(n-k+1)$, often denoted ${}_nC_k$

As an example of assessing independence, consider the 2x3 table below:

Population	Poor	Good	Excellent	Total
Drug A				100
Drug B				100
Overall	40	60	100	200

There are 200 patients overall, 100 of whom received Drug A and 100 of whom received Drug B. Overall, 40 patients had poor outcomes, 60 had good outcomes, and 100 had excellent outcomes. The logic underpinning a chi-square test is:

- The assigned treatment (Drug A or Drug B) can be labelled as event X, and the patient outcome {poor, good, excellent} can be labelled as event Y.
- If the null hypothesis that the drugs work identically holds, then the outcome Y is independent of the drug (i.e., it is independent of the treatment X).
- In other words, if the null hypothesis holds, then $Pr\{X \text{ and } Y\} = Pr\{X\} * Pr\{Y\}$
- $Pr\{X\}$ is obtained from the margin of the table; for example, the $Pr\{\text{X=Drug A}\} = 100/200 = 0.50$
- $Pr\{Y\}$ is obtained from the margin of the table; for example, $Pr\{\text{Y=poor outcome}\} = 40/200 = 0.20$
- To translate probabilities into expected numbers within the table, multiply by the overall sample size
- Thus, if the null hypothesis holds, we expect the raw data will, approximately, show the following pattern in the table (the expected numbers are italicized):

Population	Poor	Good	Excellent	Total
Drug A	.5*.2*200=*20*	.5*.3*200=*30*	.5*.5*200=*50*	100
Drug B	.5*.2*200=*20*	.5*.3*200=*30*	.5*.5*200=*50*	100
Overall	40	60	100	200

- The chi-square test compares the observed and expected numbers — if the observed numbers are similar to the expected numbers, then we conclude that the drugs work identically (a more precise statement of this conclusion will be covered under hypothesis testing)

In other words, suppose the actual data are as follows:

Population	Poor	Good	Excellent	Total
Drug A	25	30	45	100
Drug B	15	30	55	100
Overall	40	60	100	200

Statistical inference essentially asks whether those data are "close enough" to what would be expected if the drugs work identically in order to draw that conclusion.

3. Problem Set

Problem I

Using a fair coin, what is the probability of throwing 6 consecutive heads?

Problem II

You're analyzing a very complicated laboratory experiment — one which compares the active ingredients in two drugs. It's hard to precisely quantify how much better the drugs perform, but for each replicate of the experiment it's easy to determine which of the two drugs performed better than the other. Drug B performed better than Drug A in 6 experiments out of 6. How strong is this information in favor of Drug B? (Hint: Is this similar to the previous question? This turns out to be a derivation of a non-parametric test called the sign test.)

Module 5: Multinomial and Hypergeometric Distributions

1. Video

Click here to watch the video.

The discrete distributions with a finite number of possible values that you will typically encounter include the Bernoulli, binomial, multinomial, and hypergeometric. The binomial is the most common. This module covers the multinomial and hypergeometric distributions. They are closely related to the last exercise from the previous module, where we compared 2 drugs according to their pattern of 3 possible outcomes (i.e., "Poor," "Good," and "Excellent").

2. Multinomial Distribution

The multinomial distribution is often used in the analysis of R^*C "contingency tables," where R denotes the number of rows, C denotes the columns, and the interior of the table contains counts of the number of individuals falling within each cross-classification between the rows and the columns.

The data table from the previous module is copied below. It is a 2x3 contingency table:

Population	Poor	Good	Excellent	Total
Drug A	25	30	45	100
Drug B	15	30	55	100
Overall	40	60	100	200

Suppose that the investigator reclassified patients into either "Poor" or "Good/Excellent." The 2x2 contingency table becomes:

Population	Poor	Good/Excellent	Total
Drug A	25	75	100
Drug B	15	85	100
Overall	40	160	200

This table is simply a count of Bernoulli random variables. Indeed, a statistical report might begin with:

> "We observed that 75% of patients receiving Drug A had good to excellent outcomes, whereas 85% of patients receiving Drug B had good to excellent outcomes."

The original data array for this study has 200 rows — these rows are depicted in the table below, in compressed form:

Drug (Predictor)	Outcome (Response)	Number of Copies
A	Poor	25
A	Good	30
A	Excellent	45
B	Poor	15
B	Good	30
B	Excellent	55

The modified data array also has 200 rows; in compressed form:

Drug	Outcome	Number of Copies
A	Poor	25
A	Good/Excellent	75
B	Poor	15
B	Good/Excellent	85

The original version of the outcome variable, with 3 categories, is a generalization of the Bernoulli distribution, called the multinomial distribution. For the multinomial distribution, the outcome will fall into exactly one of k categories. For the Bernoulli distribution, $k = 2$.

The probability that the outcome will fall into category j is θ_j, where $\theta_1 + \theta_2 + ... + \theta_k = 1$, and the PDF is based on the these probabilities.

One application of the multinomial distribution involves random sampling with replacement. If the data set to be sampled from has k individuals, and we want to randomly select one of these individuals, then this is equivalent to using a multinomial distribution with equal selection probability (i.e., all the values of θ_j) of 1/k for each individual.

3. Hypergeometric Distribution

Consider another randomized trial comparing Drugs A and B, and using a binary outcome for simplicity, suppose that we are also interested in a rare but serious complication. The data might appear as in the table below:

Population	Complication	No Complication	Total
Drug A	$0 = a$	$210 = b$	$210 = a + b$
Drug B	$6 = c$	$184 = d$	$190 = c + d$
Overall	$6 = a + c$	$394 = b + d$	$400 = a + b + c + d = n$

One possible analysis compares the 0/210 with the 6/190. The complication is rare for both drugs, and the drugs probably aren't statistically distinguishable when the two proportions are directly compared.

Another approach to comparing the two drugs uses conditional logic:

- Conditional on 400 total observations:
    - 6 patients experienced complications
    - 394 patients experienced no complications
    - 210 patients were administered Drug A
    - 190 patients were administered Drug B
- Equivalently stated, conditional on $n = 400$:
    - 6=a+c
    - 210=a+b

- Therefore, conditional on the margins of the table, there are 7 possible data tables. These tables are listed below:

Population	Complication	No Complication	Total
Drug A	**0**	210	*210*
Drug B	6	184	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**1**	209	*210*
Drug B	5	185	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**2**	208	*210*
Drug B	4	186	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**3**	207	*210*
Drug B	3	187	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**4**	206	*210*
Drug B	2	188	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**5**	205	*210*
Drug B	1	189	*190*
Overall	*6*	*394*	*400*

Population	Complication	No Complication	Total
Drug A	**6**	204	*210*
Drug B	0	190	*190*
Overall	*6*	*394*	*400*

The margins of the table are fixed, with fixed quantities being italicized. Fixing/conditioning on $n = 400$ and $a + c = 210$ is unremarkable, as this follows from the design of the study. Additionally, fixing/conditioning on $a + b = 6$ is what distinguishes the hypergeometric distribution. In essence, what we will ask is:

> "Given that exactly 6 patients had a complication, is the observed value of the hypergeometric random variable Y consistent with the assumption that the complication rates for Drugs A and B are the same."

The random variable Y is in bold and underlined in each table above. Since the margins are fixed, once we know the random variable, the other 3 quantities within the table can be derived. Accordingly, this table "contains only one moving part."

Since the number of patients receiving Drug A is approximately the same as the number of patients receiving Drug B, if the true complication rates for the 2 drugs are equal, you should expect to see values of Y near 3. The Y values of 0 and 6 would provide the strongest evidence against the hypothesis that the complication rates are equal. This is the core of the logic underpinning the Fisher Exact Test (FET) for 2x2 tables. The FET can be extended to larger tables.

We won't work all the way through this logic, but, in brief, this can be conceptualized as an "urn problem:"

> An urn contains 6 red balls and 394 green balls. You make 190 selections. The selections aren't returned to the urn. Thus, the last selection is made out of an urn containing 201 balls. If the complication rates for the 2 drugs are identical, then each selection is made with equal probability,

and we're left with applying a counting argument. Eventually, this PDF is derived:

$$Pr\{Y = a\} = \frac{\text{(a+b)!} \quad \text{(c+d)!} \quad \text{(a+c)!} \quad \text{(b+d)!}}{\text{a!} \quad \text{b!} \quad \text{c!} \quad \text{d!} \quad \text{n!}}$$

The derivation is based on sampling without replacement. In contrast, the multinomial distribution assumes sampling with replacement, where each individual's outcome is independent of the other. This is conceptually equivalent to repeatedly sampling from the same set of $\{\theta_1, \theta_2, ..., \theta_k\}$.

The binomial distribution (covered in the next module) is also based on sampling without replacement. This is because it is the count of the number of successes in n independent Bernoulli trials.

4. Practice Problems

Problem I

Create a table, as per the Bernoulli example, of the probability mass function of the outcomes of a 6–sided die (this is a multinomial distribution with $k = 6$). The probabilities for the outcomes of a multinomial distribution don't have to be identical, although they are in this case. Calculate the mean and variance of this multinomial random variable.

Problem II

R has a function that allows you to directly simulate multinomial random variables. However, as an exercise, suppose that the only tool available to you is a function for simulating Bernoulli random variables. How could you repeatedly apply this function to simulate a multinomial random variable? (Note: Describing the algorithm is sufficient, but if you can program this algorithm in R, then this is much better.)

Problem III

Using the complication data from the above drug trial, assume that the true, but unknown, probability of a complication is the same for both drugs. Derive the probability that $Y = 0$.

To derive this probability, imagine the urn, as above, with 6 red balls and 394 green balls. Since 210 patients received Drug A, you will draw 210 balls without replacement from that urn. All these balls will be green (i.e., since, $Y = 0$).

- What is the probability that the first ball will be green?

- Conditional on the first ball being green, what is the probability that the second ball will be green, as well? (Hint: the urn now contains 399 balls — what are their colors?)
- Conditional on balls 1–209 all being green, what is the probability that ball 210 will be green, as well? Use an R function to check your answer.

Problem IV

This question is an empirical demonstration of one reason why the FET is only used for small samples: namely, in larger samples, sampling without replacement (i.e., the basis for the hypergeometric distribution and the FET) is similar to sampling with replacement (i.e., the basis for the binomial distribution and the chi–square test, among others). In smaller samples, this distinction is important enough to matter.

i. Consider an urn with 10,000 balls, 5,000 of which are red and 5,000 of which are green. Draw a ball from the urn, and let C_1 denote its color. What is the probability that C_1 is red? Return the ball to the urn (the urn continues to contain 5,000 red balls and 5,000 green balls).

ii. Draw another ball from the urn, and let C_2 denote its color. What is the probability that C_2 is red?

iii. Repeat this experiment, with C_1 being red, but do not return the ball to the urn (the urn now contains 4,999 red balls and 5,000 green balls). Draw another ball from the urn, and let C_2 denote its color.

- What is the probability that C_2 is red?
- Are $Pr\{C_1 = red\}$ and $Pr\{C_2 = red \mid C_1 = red\}$ similar?
- What about $Pr\{C_3 = red \mid C_1 = red \text{ and } C_2 = red\}$?
- What is $Pr\{C_6 = red \mid C_1\text{–}C_5 \text{ are } red\}$?

iv. Now repeat this experiment with an urn containing 10 balls, 5 of which are red and 5 of which are green.

- Are $Pr\{C_1 = red\}$ and $Pr\{C_2 = red \mid C_1 = red\}$ similar?
- What about $Pr\{C_3 = red \mid C_1 = red \text{ and } C_2 = red\}$?
- What is $Pr\{C_6 = red \mid C_1\text{–}C_5 \text{ are } red\}$?

Module 6: Binomial Distribution and Central Limit Theorem

1. Video

Click here to watch the video.

2. Binomial Distribution

Consider two Bernoulli trials. The probability of observing the sequence $\{1, 1\}$ (i.e., two successes in a row) is $\theta * \theta = \theta^2$.

This follows from the conditional independence:

$$Pr\{Y_1 = 1 \text{ and } Y_2 = 1\} = Pr\{Y_1 = 1\} * Pr\{Y_2 = 1 \mid Y_1 = 1\}$$

However, since Y_1 and Y_2 are independent, we can replace $Pr\{Y_2 = 1 \mid Y_1 = 1\}$ with $Pr\{Y_2 = 1\}$. Similarly, the probability of observing the sequence $\{1, 1, 1, 1\}$ is θ^4. We can label $\{1, 1, 1, 1\}$ as $\{4$ successes in a row$\}$, and also as $\{$exactly 4 successes$\}$, since there is only one sequence that yields 4 successes.

Now consider the sequence $\{1, 1, 0, 1\}$. The probability that this sequence will occur is $\theta * \theta * (1 - \theta) * \theta$, which can be rewritten as $\theta^3 * (1 - \theta)^1$. Indeed, we could have written the probability of observing $\{1, 1, 1, 1\}$ as $\theta^4 * (1 - \theta)^0$, since anything to the power of 0 equals 1.

In summary, we have derived the probability that the sequence $\{1, 1, 0, 1\}$ will occur. However, this event isn't equivalent to $Pr\{$exactly 3 successes$\}$, since $\{0, 1, 1, 1\}$, $\{1, 0, 1, 1\}$, and $\{1, 1, 1, 0\}$ also generate exactly three successes. In other words:

$$Pr\{\text{exactly 3 successes}\} = Pr\{\{0, 1, 1, 1\} \text{ or } \{1, 0, 1, 1\} \text{ or } \{1, 1, 0, 1\} \text{ or } \{1, 1, 1, 0\}\}$$

These events are disjoint, so the probabilities add. Moreover, the component probabilities are identically $\theta^3 * (1-\theta)^1$, and so $Pr\{\text{exactly 3 successes}\} = 4 * \theta^3 * (1-\theta)^1$.

The binomial RV represents the probability of exactly Y successes out of n trials. Its PDF is the following:

$$Pr\{Y = y\} =_n C_y * \theta^y * (1-\theta)^{n-y} \text{ for } x = 0, 1, 2, ..., n; \quad 0 \quad otherwise$$

The parameters of the binomial distribution are n and θ. R has a function to calculate the value of the PDF for small moderate values of y and n, and similarly for the CDF. When n is very large, it is usually better to use an approximation to the binomial distribution (this will be discussed later in the course).

In summary, use the binomial distribution when the goal is to count the total number of successes (equivalently stated, the proportion of successes) out of a series of n Bernoulli trials with the same probability of success for each trial. The order of the successes doesn't matter. For example, when the event in question is the presence or absence of a surgical complication, we only care about the number of people with complications, not their order within the data set.

The above discussion illustrates deriving the binomial distribution from first principles. The binomial distribution can also be understood to be the sum of the $\{0, 1\}$ outcomes from n independent Bernoulli RVs. It turns out that conceptualizing the binomial distribution as the sum of independent Bernoulli RVs makes deriving its mean and variance simple: the mean is simply $\theta + \theta + \theta + ... + \theta = n * \theta$, and the variance is $n * \theta * (1-\theta)$. The following is a discussion of how we arrived at this conclusion. The discussion starts with some basics about the mean and variance of sums of independent random variables, and then relates this information to the case of independent Bernoulli trials. Finally, it links this information to the binomial distribution.

To understand the following discussion, we need to know some basic facts about expectation. We'll cover expectation in more detail later in the course. For now, it is sufficient to accept the following rules:

1. So long as $Y_1, ..., Y_n$ are independent, the mean of $S_n = Y_1 + Y_2 + ... + Y_n$ is equal to $\mu_1 + \mu_2 + ... + \mu_n$. In other words, the mean of the sum of independent random variables is equal to the sum of their means.
2. Letting c denote a constant, the mean of $c * Y$ is $c * \mu_Y$. In other words, the mean of a constant times a random variable is equal to the constant times the mean of that random variable.

With these rules in hand, let's consider the sample mean. The sample mean can be written as $M_n = (1/\text{n}) * S_n$. It is a random variable because its value can't be predicted ahead of time. By application of rules 1 and 2 above, we can say that M_n has a mean of $1/\text{n} * (\mu_1 + \mu_2 + ... + \mu_n)$ as follows:

$Mean(M_n) = Mean((1/n) * S_n) = (1/n) * Mean(S_n)$ (by application of Rule 2 [$c = 1/n$])

$Mean(M_n) = (1/n) * (\mu_1 + \mu_2 + ... + \mu_n)$ (by application of Rule 1)

So far, we've shown the mean of the sample mean, M_n is $(1/n) * (\mu_1 + \mu_2 + ... + \mu_n)$ for independent random variables $Y_1, ..., Y_n$. When $Y_1, ..., Y_n$ are identically distributed (in addition to being independent), they all have the same mean, μ. Thus, we can simplify our expression for $Mean(M_n)$ by replacing μ_1 with μ, μ_2 with μ, and etc. In other words, the term $\mu_1 + \mu_2 + ... + \mu_n$ simplifies to $n * \mu$. Therefore, M_n has a mean of $(1/n) * n * \mu$, which reduces to μ.

> *Thus, in general, when $Y_1, ..., Y_n$ are independent and identically distributed, M_n has a mean of μ; that is, the mean of the original random variable.*

To find the standard deviation, we apply a similar argument that is also based on two more basic rules. These rules are about the variance (we'll cover this topic in more detail later in the course):

1. When the random variables in question are independent, the variance of their sum is equal to the sum of their variances: $Var(S_n) = Var(Y_1) + Var(Y_2) + ... + Var(Y_n)$
2. The variance of $c * Y = c^2 * Var(Y)$

Using Rule 1, and denoting the variance of Y by σ^2, we arrive at: $Var(S_n) = n * \sigma^2$. Now, returning to the sample mean:

$Var(M_n) = Var\{(1/n) * S_n\} = (1/n)^2 * Var(S_n)$ (by application of Rule 2)

$Var(M_n) = (1/n)^2 * n * \sigma^2$ (substituting $n * \sigma^2$ from above)

$Var(M_n) = \sigma^2/n$

The standard deviation is, of course, the square root of the variance, which is $\sigma/\sqrt{n}$. The standard deviation of a sample statistic, like the mean, is called the *standard error*. So, that's how we derived the mean and variance for the sample mean for independent and identically distributed random variables $Y_1 + Y_2 + ... + Y_n$.

Applying all this information to the binomial random variable Y, we can write Y as $S_n = Y_1 + Y_2 + ... + Y_n$, where each of the independent Y_i are Bernoulli trials with success probability θ. Remember, the binomial random variable, $Y = S_n$, is the number of successes. Since each Bernoulli random variable Y_i has mean θ, a binomial random variable Y (the number of successes) has mean $n * \theta$. Furthermore, since a Bernoulli random

variable Y_i has variance $\theta * (1 - \theta)$, the binomial random variable Y has variance $n * \theta * (1 - \theta)$. Often, it is useful to transform binomial random variables into proportions: $Y^* = Y/n$. Note, Y^* is equivalent to our definition above about a sample mean:

$$Y^* = Y/n = (1/n) * S_n = M_n$$

Thus, Y^* will have mean θ (i.e., the mean of the underlying Bernoulli random variables that are summed in S_n) and variance $[\theta * (1 - \theta)]/n$ (we showed above that the variance of M_n is σ^2/n and $\sigma^2 = \theta * (1 - \theta)$, in this case).

When sample sizes are large, we can approximate a binomial proportion with the continuous normal distribution. The reasons to consider doing so include:

1. To avoid a problem
2. To simplify the analysis

The problem to avoid is, for example, calculating the probability of 600,000 successes out of 1,000,000 independent Bernoulli trials with success probability θ. We would need to calculate:

$$\frac{1,000,000!}{(600,000)!(400,000)!} * \theta^{600,000} * (1 - \theta)^{400,000}$$

This formula generates numbers that are too extreme to work with directly (e.g., 1000000! is a very, very large number). The simplification follows from:

1. In many cases, inference using a continuous normal distribution is more straightforward to execute than inference using a discrete binomial distribution
2. With large sample sizes, many distributions approach the normal distribution, and thus, separate approaches for each distribution can be replaced by a single generalized approach using normality.

This is where the Central Limit Theorem (CLT) enters the picture. We won't state it precisely. For our purposes it suffices to say:

> *As the sample size grows sufficiently large, and the variables are independent and identically distributed, a sample mean not only has mean μ and variance σ^2/n, it is also normally distributed.*

The next module focuses on the normal distribution. As a curriculum note:

> We place little emphasis on deriving the moments of standard distributions, as this information is readily available elsewhere. For example, moment generating functions are relegated to

supplemental material. The exception is when those and similar manipulations are used elsewhere in the curriculum. For example, the survival analysis class covers parametric modeling using the exponential distribution, and therefore, this class practices using that distribution.

3. Problem Set

Problem I

Suppose that you observe exactly 1 success for a binomial variable representing n Bernoulli trials.

i. Demonstrate that, conditional on observing 1 success, the distribution of the specific trial where the success was observed is multinomial.

ii. What are the parameters of this multinomial distribution? (Hint: this problem can be solved by logic — brute force algebra isn't needed.)

Problem II

Two physicians, equally well trained, are asked to determine whether a patient has a particular diagnosis. The 2x2 table below summarizes the results from 100 patients:

	Doctor 2 Diagnosis = Yes	Doctor 2 Diagnosis = No
Doctor 1 Diagnosis = Yes	60	6
Doctor 1 Diagnosis = No	0	34

One investigator argues that the physicians are similarly calibrated: Doctor 1 diagnoses 66% of patients, whereas Doctor 2 diagnoses 60% of patients (these percentages turn out to be statistically similar). Another investigator argues that, although the physicians agree most the time, when they disagree it is always the case that Doctor 1 says the disease is present and Doctor 2 says the disease is absent. How could this investigator derive a statistical test to support their logic? (Hint: Use a conditional binomial distribution. What are the values of n and θ for this distribution?)

Module 7: PDF and CDF

1. Video

Click here to watch the video.

2. Discussion of PDFs and CDFs

We're now going to shift our focus from discrete to continuous distributions. This module focuses on PDFs and CDFs; namely, the similarities and differences across types of distributions.

For continuous RVs, the probability density function (PDF) has a similar purpose as the probability mass function (PMF) for discrete variables; namely, to differentiate between values of the RV that are relatively likely, relatively unlikely, and impossible. For example, for a discrete binomial variable with $n = 10$ and $\theta = 0.5$, the PMF would demonstrate that a value of 5 successes is relatively likely, a value of 9 successes is relatively unlikely, and a value of 11 successes is impossible.

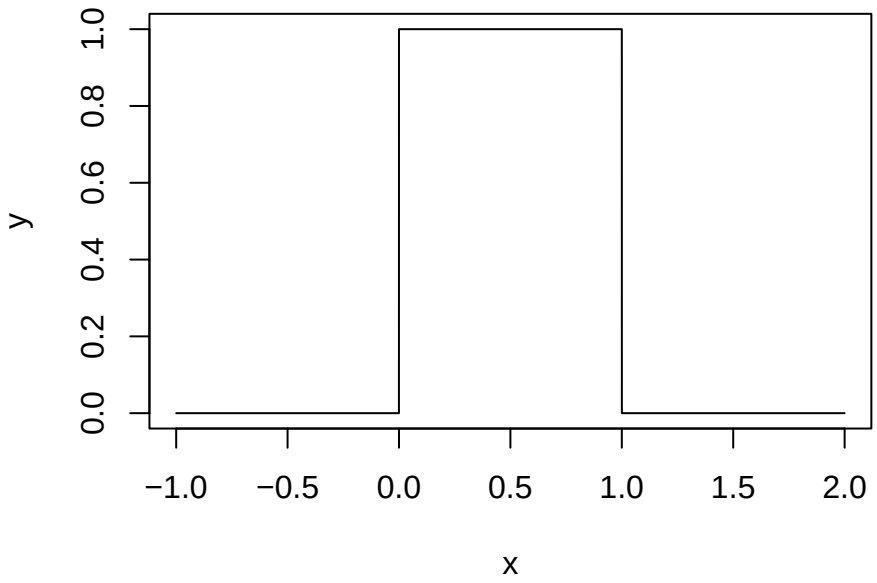
For continuous RVs, the PDF makes the same distinction. The main difference is, as per the logic of calculus, the probability that Y will take on any particular value $Y = y$ is 0. Accordingly, probabilities are defined within intervals rather than at discrete time points.

To obtain the probability that a continuous RV will fall within the interval $[a, b]$ (it doesn't matter whether the endpoints of the interval are open or closed), integrate the PDF and evaluate it at a and b. We will use the uniform distribution as an example. The PDF for the uniform distribution is $1/(b - a)$, where a and b define the minimum and maximum values. Stated formally:

$$PDF = f(y) = \begin{cases} 1/(b-a) & \text{for } y \in [a, b] \\ 0 & otherwise \end{cases}$$

Actually, uniform refers to a "family" of distributions defined by different values of a and b. For example, a commonly used member of the uniform family of distributions is called the "standard uniform distribution." It is defined by $a = 0$ and $b = 1$, and therefore, it's PDF is $1/(b - a) = 1/(1 - 0) = 1$. Uniform distributions are usually written as $U(a, b)$, and therefore, standard uniform is written as $U(0, 1)$. If a random variable, Y, is distributed uniform on the interval 0 to 1, then we write: $Y \sim U(0, 1)$. If we plot the PDF of a standard uniform distribution in R, it looks like this:

Standard Uniform Distribution



The shape of the PDF is rectangular. This is the case with any of the uniform distributions. Try plugging in different values of a and b using the "min" and "max" arguments to the $dunif()$ function to see what other uniform distributions look like.

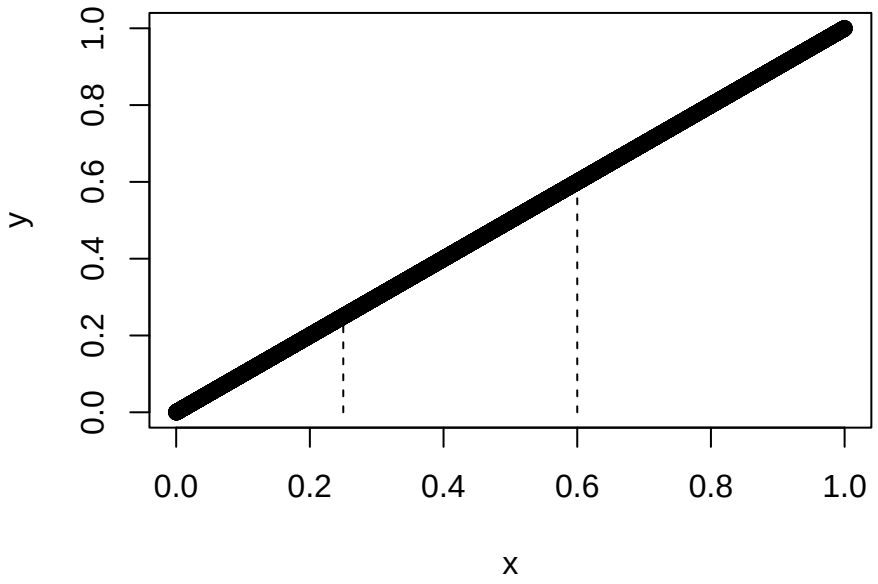
Let's say $Y \sim U(0, 1)$, and we want to know the probability that Y takes on a value between 0.25 and 0.60. To answer this question, we can integrate the PDF of the standard uniform distribution, which is the constant 1, with respect to y and arrive at y (i.e., the integral of a constant with respect to y is the constant times y). If we evaluate the integral at the values 0.25 and 0.60, then we get the following: 0.60 - 0.25 = 0.35. What we've done is derive the CDF of the standard uniform distribution, and we've used it to calculate a probability! The result can be visualized in two ways:

1. We can plot the CDF of the standard uniform distribution, which is the function $f(y) = y$ for y in $[0, 1]$
2. We can annotate the plot of the PDF at the bounds 0.25 and 0.60

Recall from calculus that "evaluating" an integral means to:

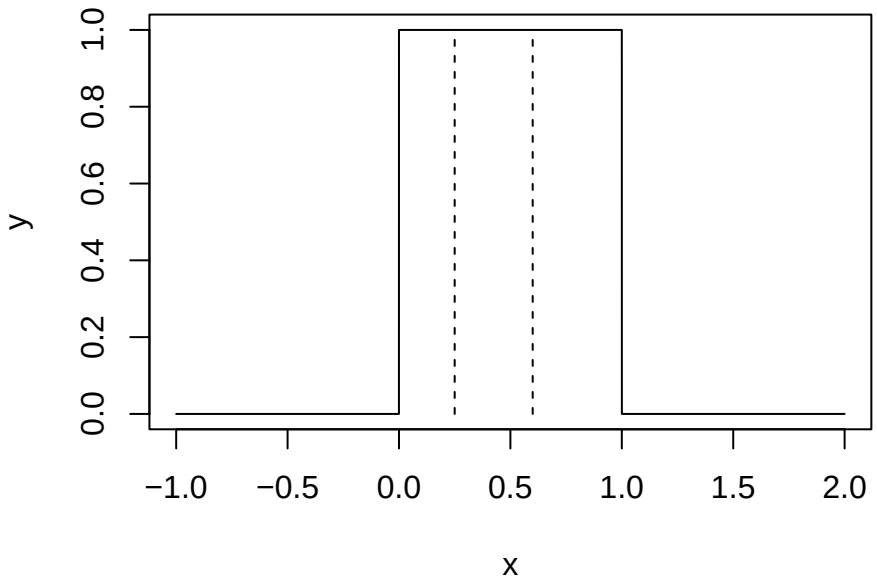
1. Plug in the upper value: $Y = 0.60$
2. Plug in the lower value: $Y = 0.25$
3. Subtract #2 from #1

CDF



The above is a plot of the CDF for the standard uniform distribution with lines drawn at the values 0.25 and 0.60. Clearly, 60% of the area under the line is contained from the origin to the point 0.60 on the x-axis. Likewise, 25% of the area under the line is contained from the origin to the point 0.25. If we subtract these two areas, we get the area in the middle; that is, 35% or 0.35.

PDF



The above is a plot of the PDF of the standard uniform distribution with the same values marked off, as seen in the plot of the CDF, along the x-axis. We can almost eyeball the area between the dashed lines as being slightly more than 1/3 of the area under the density function.

In the above example, we integrated the PDF of $U(0,1)$ from 0.25 to 0.60 to find the area under the curve between these points. Essentially, this is the same process we follow to derive the CDF of the uniform family of distributions, as shown below:

$$PDF = f(Y = y) = \begin{cases} 1/(b-a) & \text{for } y \in [a,b] \\ 0 & otherwise \end{cases}$$

$$\int_a^x f(y)dy \; = \int_a^x \frac{1}{b-a}dy \; = \frac{1}{b-a}y\Big|_a^x \; = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a}$$

The above math shows the PDF of the general family of uniform distributions and its integral. The integral is evaluated starting at the beginning of the interval and ending at some arbitrary point, x, which is less than or equal to the end of the interval, b (i.e., recall, $f(y)$ is defined for y in $[a, b]$). The quantity $1/(b-a)$ is a constant by definition, and the integral of a constant with respect to x is that constant times x. In this case, we need to evaluate the integral at a and x. Therefore, we use some subtraction and algebra to find that the CDF is $(x-a)/(b-a)$.

We can check that our work complies with the standard uniform distribution, as follows:

- $Y \sim U(0,1)$ implies $a = 0$ and $b = 1$
- $Pr[0.25 \leq x \leq 0.60] = Pr[x \leq 0.60] - Pr[x \leq 0.25] = (0.60-0)/(1-0) - (0.25-0)/(1-0) = 0.60 - 0.25 = 0.35$

In this example, we've used the CDF to compute the area between $a = 0.25$ and $b = 0.60$, by first computing the area between $a = 0$ and $b = 0.60$, and then by subtracting the area from $a = 0$ to $b = 0.25$ (this logic is similar to how we used the above plot of the CDF). This is an example of a common use of the CDF. That is, once we know the CDF, we can use it in clever ways to compute the probability that a random variable takes on any value within a given range, without having to integrate the PDF repeatedly.

Now that we've shown the connection between PDFs and CDFs; that is, the CDF is the integral of the PDF — and by extension, the PDF if the derivative of the CDF — we need to emphasize an important property of the PDF that is implicit in the demonstration we made. Namely, PDFs for continuous random variables integrate to 1. For example, using what we learned in this module, we can see that any uniform distribution integrates to 1. The reason why is as follows:

- Given $Y \sim U(a,b)$ then $CDF[f(Y=y)] = (y-a)/(b-a))$
- Therefore, $CDF[f(Y=b)] = (b-a)/(b-a) = 1$

A distribution that integrates to 1 is called a "proper distribution," provided it also generates non-negative values. Stated alternatively, this is the law of total probability; that is, the probability of all possible values $Y = y$ sum to 1.

For discrete RVs, the mean is a weighted sum of Y. That is, the mean is weighted by the value of the PMF at $Y = y$. For a continuous RV, by analogy, the mean is defined as follows: here, $f(y)$ is the PDF, and the integration is taken over the range of the possible values of the RV (called the "support"):

$$E[Y] = \int yf(y)dy$$

For $Y \sim U(0,1)$, this integral becomes:

$$PDF = f(Y = y) = \begin{cases} 1 & \text{for } y \in [a,b] \\ 0 & otherwise \end{cases}$$

$$E[Y] = \int_0^1 yf(y)dy \; = \int_0^1 (y)(1)dy \; = \frac{y^2}{2} \; |_0^1 = \frac{1}{2} - \frac{0}{2} = \frac{1}{2}$$

The integral in this case is $y^2/2$, which is evaluated at $y = 1$ and $y = 0$, and thus, is equivalent to $1/2 - 0/2 = 0.50$. Again, this approach generalizes the entire uniform family of distributions. In other words, for $Y \sim U(a,b)$, we get the following:

$$PDF = f(Y = y) = \begin{cases} 1/(b-a) & \text{for } y \in [a,b] \\ 0 & otherwise \end{cases}$$

$$\begin{aligned} E[Y] &= \int_a^b yf(y)dy = \int_a^b (y)(\frac{1}{b-a})dy = (\frac{1}{b-a}) \int_a^b ydy \\ &= (\frac{1}{b-a})(\frac{y^2}{2}\Big|_a^b) \\ &= (\frac{1}{b-a})(\frac{b^2}{2} - \frac{a^2}{2}) = \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2} = \frac{1}{2}(a+b) \end{aligned}$$

The variance of Y is obtained in a similar fashion by integrating the PDF weighted by the square of the difference between the range of observed values of y and the mean of Y. This can be done as follows (the mathematics have been truncated for brevity):

$$\int_a^b (y - \frac{a+b}{2})^2 (\frac{1}{b-a})dy = \frac{1}{12}(b-a)^2$$

The mean and the variance are the first two "central moments" of a random variable. The method shown above (i.e., finding the second central moment based on the first) is one way of deriving the moments. We can also find the moments by using a moment generating function, which will be mentioned in the supplemental materials. Alternatively, we can also calculate $E(X^2) - [E(X)]^2$.

3. Practice Problems

Problem I

In survival analysis, a key quantity is the "survival function" — namely, the proportion of the original population that survive until time t. By definition, $S(0) = 1$ — that is, 100% of the population is alive at $t = 0$, which might be birth or the time of an event (e.g., disease diagnosis and etc.). The survival function $S(t)$ is related to the CDF: how are they related?

Problem II

In survival analysis, another key quantity is the "hazard function." It is the instantaneous rate of death at time t, conditional on having survived until time t. If $f(y)$ is the PDF of the survival time Y, how can the hazard function be written in terms of the PDF and CDF?

Problem III

Write an R function to plot the PDF of an arbitrary uniform distribution with any parameters, a and b. (Hint: Refer to the code in the .rmd file and the R documentation for the $dunif()$ function.)

Problem IV

Write another R function to compute the mean and variance of an arbitrary random uniform variable with parameters a and b, again using equations rather than generating data. Use the new function to perform the following:

i. Plot the PDF for a random variable $Y \sim U(1, 2)$

ii. Compute the probability that Y takes on values between 0.5 and 1

iii. Compute the probability hat Y takes on values between 3 and 5 (Hint: this will test software engineering skills!)

iv. Compute the mean of Y

v. Compute the variance of Y

Problem V

Now, generate data from the uniform distribution $U(1, 2)$, and use this data to verify the computations in parts 2 through 5 above. Do this step in a series, first generating a data set with only 10 observations, then

with 100, then with 1,000, and finishing with 5,000. Do the computations from the data match what was obtained from the theoretical values?

Bonus Problem

Solve the integral shown above to demonstrate that the 2nd central moment (i.e., the variance) of the uniform family of distributions is $(1/2) * (b - a)^2$.

Module 8: Agreement

1. Video

Click here to watch the video.

2. Discussion

Following up on an example from Module 6, suppose we ask two physicians to assess whether or not a patient has a certain disease. One physician is a generalist, and the other is a specialist. The specialist is an expert in the condition under study, and their diagnosis is always correct. On the other hand, there are too few specialists, and it is natural to ask whether the generalist's diagnostic abilities are "good enough."

The results for 100 patients are summarized in Table 1.

Table 1

	Specialist: No Disease	Specialist: Disease	
Generalist: No Disease	65	20	85
Generalist: Disease	10	5	15
	75	25	100

The two physicians agree in 70 patients out of 100 (65 agreements on no disease, 5 agreements on disease), which is part, but not all, of the story. Indeed, the "directional" results are also quite relevant:

- If the generalist says "No Disease" they will be correct in 65/85 cases.
- If the generalist says "Disease" they will be correct in 5/15 cases.
- If the specialist says "No Disease" the generalist will agree in 65/75 cases.
- If the specialist says "Disease" the generalist will agree in 5/25 cases.
- The specialist finds a greater prevalence of disease: 25/100 versus 15/100.

These results, along with the consequences of the two possible wrong decisions, would be used to determine when the generalist should refer the patient to a specialist for a second opinion. Here, the main problem is that the generalist is missing most of the actual cases, and also wrongly diagnosing some non–cases as cases.

Although some "observer agreement studies" compare experts with non–experts, most compare 2 or more observers in the absence of a gold standard. For simplicity, we will limit consideration to 2 observers, which are considered to be representative of a larger population. The results are reported in a single summary table with the same structure as Table 1, but now with the entries denoted by "Observer A" and "Observer B." We are no longer interested in directional relationships, and so the agreement measures are defined differently:

Table 2

	Observer B: No Disease	Observer B: Disease	
Observer A: No Disease	65	20	85
Observer A: Disease	10	5	15
	75	25	100

- The overall agreement is 70/100.
- If one observer says "No Disease" the other will agree in (65+65)/(65+65+20+10) of the cases.
- If one observer says "Disease" the other will agree in (5+5)/(5+5+20+10) of the cases.
- The "chance–adjusted agreement" (i.e., kappa) is (OBS-EXP)/EXP, where OBS=.70 and EXP=(.85∗.75) + (.15∗.25). Here, kappa=(.70-.675)/.675=.04.

For the second bullet point, please note that there are 65+65+20+10 ratings of "No Disease," which appropriately double–counts the cases where the two observers agree.What's being estimated is the probability that if one physician diagnoses a patient as being without disease a second opinion will show the same. A similar analysis applies to bullet point 3.

The fundamental idea behind the kappa statistics is that some level of agreement can be achieved simply by guessing.Here, once it is stipulated that "Observer A" has a 15% prevalence of disease and "Observer B" has a 25% prevalence, the expected proportions within the table are calculated as per a chi–square test for independence.

Table 3: Expected Proportions

	Observer B: No Disease	Observer B: Disease	
Observer A: No Disease	.85∗.75	.85∗.25	.85
Observer A: Disease	.15∗.75	.15∗.25	.15
	.75	.25	1

The expected agreement is the sum of the entries on the main diagonal.

The kappa statistic is conservative, in the sense that it doesn't give the observers credit for independently coming to similar conclusions about disease prevalence. Due to this, it is at risk for misinterpretation. Nevertheless, it is commonly used, in part because of its relationship to a natural quantity to estimate for observer agreement studies with continuously–scaled outcomes. (We won't expand on this analogy).

Expanding observer agreement, and kappa, to discrete outcomes with more than 2 categories involves assigning scores to each type of disagreement. For example, in the 2x2 case, these scores are:

Table 4: Agreement Scores for 2x2 Table

	Observer B: No Disease	Observer B: Disease
Observer A: No Disease	1	0
Observer A: Disease	0	1

The usual estimate of agreement is also the weighted sum of the category–specific values, weighted by these scores: Agreement= $(65/100) * 1 + (20/100) * 0 + (10/100) * 0 + (5/100) * 1$.

For outcomes with more than 2 categories, agreement is assigned a score of 1, the worst possible disagreement is (usually) assigned a score of 0, and less severe disagreements are assigned scores between 0 and 1, depending on the practical significance of the disagreement. Here, disagreements between "Fair" and "Poor" were considered to be subtle misclassifications, and disagreements between "Good" and "Poor" were considered to be more serious. A weighted sum is calculated as before, and the result is a "weighted kappa statistic."

Table 5: Agreement Scores for 3x3 Table

	Observer B: Poor	Observer B: Fair	Observer B: Good
Observer A: Poor	1	0.8	0
Observer A: Fair	0.8	1	0.2
Observer A: Good	0	0.2	1

In practice, creating a weighted kappa statistic requires discussion with the investigator about the practical implications of various types of disagreement. Fortunately, research has shown that the value of weighted kappa is relatively robust to reasonable selection of the weights.

The presentation to date has been limited to observer agreement studies for discrete outcomes. Without attempting a systematic treatment of continuous outcomes, the main idea can be illustrated by restructuring the calculation of the above weighted kappa statistic, after assigning values of 0 to "Poor," 0.2 to "Fair," and 1 to "Good." (This assignment retains the agreement weights). With two observers and 4 patients to be rated, one possible data array is:

Patient	Observer 1	Observer 2
1	Poor=0	Poor=0
2	Poor=0	Fair=0.2
3	Good=1	Fair=0.2
4	Good=1	Good=1

This is the raw data array, and the calculation of kappa would include derived agreement scores of $\{1, 0.8, 0.2, 1\}$.

The raw data could also be organized in a long–thin format:

Patient	Observer	Result
1	1	Poor=0
1	2	Poor=0
2	1	Poor=0
2	2	Fair=0.2
3	1	Good=1
3	2	Fair=0.2
4	1	Good=1
4	2	Good=1

Using the numeric version of the scores as the outcome variable, this is a type of regression model that will be encountered in BIOSTAT 702 and BIOSTAT 705: the outcome is continuously scaled, and the categorical predictors are "Patient" and "Observer." This data structure generalizes to multiple observers, and so observer agreement for continuously–scaled predictors turns out to be a special case of a more general statistical model.

This is also an example of treating an ordinal variable as being continuously scaled, the appropriateness of which should be assessed on a case–by–case basis. For example, the usual modeling assumptions pertaining to normality are unlikely to apply.

Module 9: Normal Distribution

1. Video

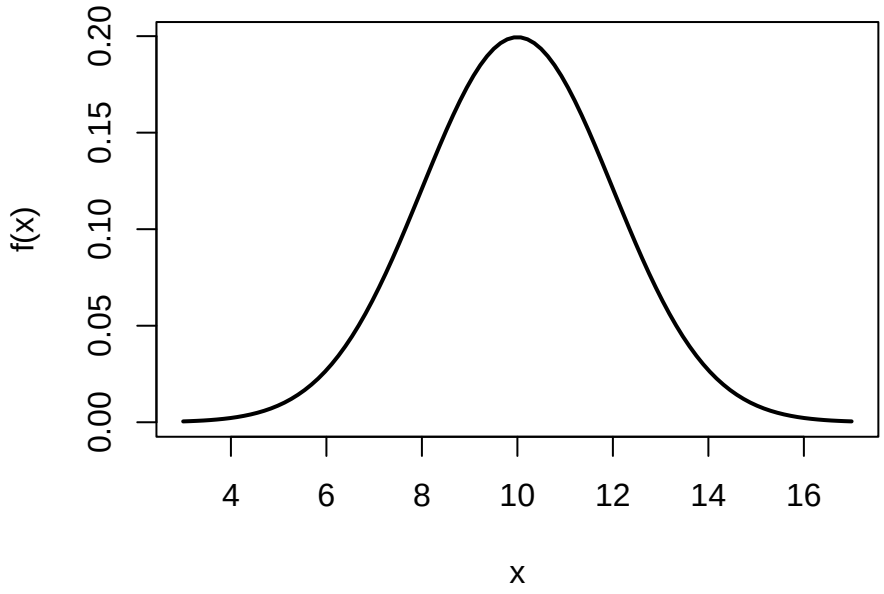
Click here to watch the video.

Note: There is a minor notation mistake, where we used a capital F to name the PDF for the normal distribution. Typically, we use a lowercase f for the PDF and an uppercase F for the CDF.

2. Introduction

The normal distribution is described by two parameters that indicate the location and scale of the distribution. The location parameter is the mean, which we write as μ, and the scale parameter is the variance, which we write as σ^2. In statistics, "scale parameters" reveal the dispersion, or spread, of a distribution. The "location parameter" reveals the center of the distribution. Here's a plot of a normal distribution with mean=10 and variance=4.

Normal Distribution



The distribution is located, or centered, at the mean value of 10 on the x-axis. The variance of 4 describes the spread of the curve. Sometimes, it is more helpful to understand dispersion in terms of standard deviation

(i.e., the square root of the variance), because it is on the same scale as the x values. In this example, the standard deviation is 2, and it's easy to see based on how the x-axis is labeled, that the majority of area under the curve falls between +/- 2 standard deviations from the mean (i.e., x=6 and x=14).

3. Probability Density Function

The plot above is a picture of a specific normal distribution. We can see that by changing the location and scale parameters, we can obtain an infinite number of normal distributions. We can represent this in mathematical notation by writing the PDF for the normal distribution with mean, μ, and variance, σ^2, as follows:

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

Note, the value of $f(x)$ is what we plotted above on the y-axis. It is for the specific case where $\mu = 10$ and $\sigma^2 = 4$.

Sometimes, we will see the PDF written slightly different, like this:

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

These equations are obviously algebraically equivalent, but can lead to confusion when writing the notation for normally distributed random variables. It is conventional to write $X \sim N(\mu, \sigma^2)$ to indicate that X is normally distributed with mean μ and variance σ^2. For example, $X \sim N(10, 4)$ says X is normally distributed with mean 10 and variance 4. However, the standard deviation in this case is 2. This can be confusing, since most of us prefer to work with the standard deviation rather than the variance (see the note below on Chebyshev's inequality). For this reason, some textbooks use the format $X \sim N(\mu, \sigma)$: therefore, the previous example could be read as $X \sim N(10, 2)$. In summary, make certain to understand the format being used!

The PDF for a "standard normal" distribution (i.e., mean=0 and variance=1) is a simplified version of the above PDF, where we plug $\mu = 0$ and $\sigma^2 = \sigma = 1$ as follows:

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

4. Using the Standard Normal PDF

Facility with calculating and working with standard normal probabilities is helpful for the practicing biostatistician. Various applications of the normal distribution use the process of "standardizing" a normal random variable. For example, $X \sim N(\mu, \sigma)$ then $(X - \mu)/\sigma \sim N(0, 1)$. In other words, subtracting the mean and dividing by the standard deviation transforms the variable into a standard normal. Standardizing allows us to treat all problems dealing with normal probabilities within a general framework.

The exercises emphasize this process, but for now, the key insight is:

> Since the standard normal distribution has a standard deviation of 1, probabilities for the standard normal correspond to probabilities for standard deviation multiples for other normal distributions. For example, we know the probability that a standard normal random variable will fall within the interval (-1.96, +1.96) is 95%. Equivalently, recognizing that the mean of a standard normal random variable is 0 and the standard deviation is 1, the interval can written as (0-1.96*1, 0+1.96*1). More generally, this interval can described as (μ-1.96σ, μ+1.96σ). For example, for a normal random variable with mean 50 and standard deviation 10, the probability that the standard normal variable will fall within (50-1.96*10, 50+1.96*10)=(30.4, 69.6) is 95%.

Note: Chebyshev's inequality gives the probability that an observation will be further than a given number of standard deviations of the mean for any distribution. Formally stated:

$$Pr[\Big|X - \mu\Big|] \geq \kappa\sigma] \leq \frac{1}{\kappa^2}$$

For $k = 2$, the inequality states that the probability an observation is more than 2 standard deviations from the mean is 0.25 (more precisely, it is no more tan 0.25, because of the $\leq$). This implies the probability that an observation is within 2 standard deviations of the mean is 0.75. In other words, 75% of observations are expected to be within 2 standard deviations of the mean. Chebyshev's inequality holds for any random variable with non-zero variance, regardless of the specific distribution. However, if we know the distribution of the random variable in question, we can be more specific than Chevyshev's theorem allows. For example, if X is a normally distributed random variable, we know (as discussed above) that 95% of the observations will fall within 1.96 standard deviations of the mean.

5. Cumulative Distribution Function

The CDF is the integral of the PDF from $-\infty$ to y. There isn't a closed form, which isn't a problem as R has a function that makes this calculation. Given its widespread application to statistics, it would be a surprise if the normal PDF didn't integrate to 1; in fact, it does. The derivation is tedious, and is treated as supplemental information. Here is a video that explains it.

6. How Normal Distributions are used in Statistics

There are 3 reasons why normal distributions are common in statistics. In fact, when we read about the "assumptions of normality," it is usually one of these reasons that is being referenced.

1. Loosely speaking, the Central Limit Theorem (CLT) states that the sum of N independent identically distributed random variables divided by N (i.e., the sample mean) converges toward a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, as the sample size N increases. The variables in question don't have to be normally distributed, so long as the sample is sufficiently large. If the variables are normal, then the sample mean is normally distributed, regardless of sample size.
2. If the underlying system producing realizations of the random variable has a large number of small perturbations (i.e., errors), and these perturbations are independent, the random variable will have a normal distribution. In other words, many, but not all, physical systems produce distributions that are approximately normal.
3. As a member of the exponential family, the normal distribution has a number of convenient mathematical properties. We will discuss this topic later in the course.

7. Practice Problems

Problem I

For the standard normal distribution Z:

i. What is $Pr\{Z > 0\}$?

ii. What is $Pr\{Z < 0\}$?

iii. What is $Pr\{Z = 0\}$?

iv. What is $Pr\{Z > 2\}$?

v. What is $Pr\{1 < Z < 2\}$?

vi. What is $Pr\{-1 < Z < 2\}$?

Problem II

For a normal distribution Y with mean 50 and standard deviation 10:

i. What is $Pr\{Y > 50\}$?

ii. What is $Pr\{Y > 55\}$?

iii. What is $Pr\{40 < Z < 65\}$?

iv. What is the 95th percentile of Y?

Problem III

This problem illustrates the CLT:

i. Plot the PDF of a binomial distribution with $n = 10$ and $\theta = 0.9$.

ii. Plot the PDF of a binomial distribution with $n = 20$ and $\theta = 0.9$.

iii. Plot the PDF of a binomial distribution with $n = 30$ and $\theta = 0.9$.

iv. Plot the PDF of a binomial distribution with $n = 50$ and $\theta = 0.9$.

v. Plot the PDF of a binomial distribution with $n = 100$ and $\theta = 0.9$.

vi. Plot the PDF of a binomial distribution with $n = 500$ and $\theta = 0.9$.

vii. Plot the PDF of a binomial distribution with $n = 1000$ and $\theta = 0.9$.

viii. At what sample size does the PDF appear to be approximately normal?

viiii. Do the same exercises for $\theta = 0.7$. At what sample size does the PDF appear to be approximately normal?

Module 10: Distributions Related to the Normal Distribution

1. Video

Click here to watch the video.

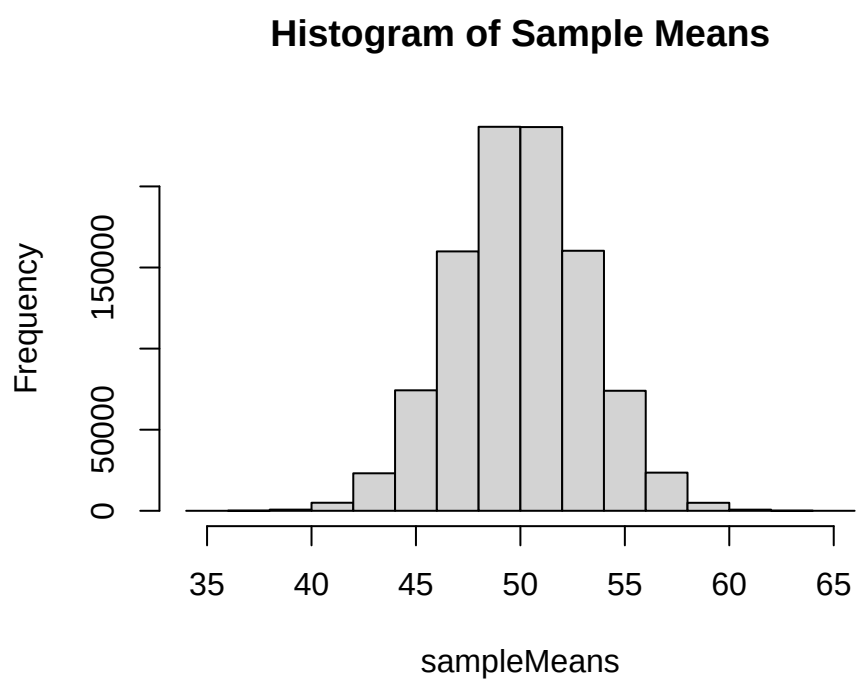
2. The Student's t-Distribution

Let $Y \sim N(\mu, \sigma^2)$, and let $\overline{y} = \frac{1}{N}\sum y_i$ be the sample mean. Imagine taking a series of samples from the population with replacement and computing the mean, $\overline{y}$, of each sample. As a concrete illustration, consider the following R program that generates 1 million samples of size $N = 10$ from a population that has a mean 50 and a variance 100 (standard deviation 10). Note, the size of the population is unimportant, but in most applications, it is assumed to be infinite.

```
sigma <- 10   #Parameters for the normally distributed population we will sample from
mu <- 50
N <- 10       #Sample Size
K <- 1000000  #Number of samples to draw from the population

set.seed(123) #Draw K samples of size N from the population. We use a random number
              #seed to make our results reproducible. The samples are stored in a
              #NxK matrix. This means each column of the matrix is a single sample.

generateData <- function(){rnorm(n = N, mean = mu, sd = sigma)}
samples <- replicate(K, generateData())
sampleMeans <- apply(X = samples, MARGIN = 2, FUN = mean)
hist(sampleMeans, main = "Histogram of Sample Means")
```



This is an empirical example of the "sampling distribution" for the sample mean. This is a very important distribution in statistics, because it is the basis for statistical inference (i.e., p-values and confidence intervals) for the population mean. For example, we can imagine using the sampling distribution to determine the probability of observing a sample mean of 60 or higher, if the true population mean is actually 50 (this would be the area under the curve to the right of 60 in the above histogram, which is essentially 0). We'll be discussing this topic, as well as various other features of the sampling distribution for the sample mean, throughout the rest of this course. Now, we will focus on the standardized difference of each sample mean from the population mean. This relationship can be described by the random variable Z:

$$Z = \frac{\overline{y} - \mu}{(\sigma/\sqrt{N})} \sim N(0, 1)$$

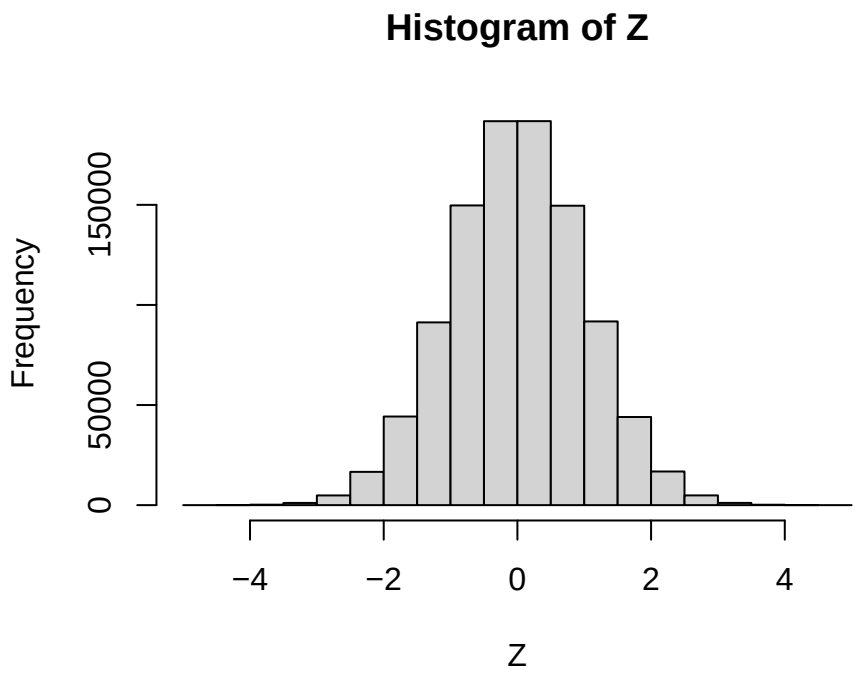
Recall from the last module, any observation can be standardized by subtracting the mean from that observation and dividing by the standard deviation. By analogy, we have done the following:

1. Treat each sample mean as an "observation"
2. Subtract the population mean (the mean of all sample means)
3. Divide by the standard deviation of the observations (the sample means)

- The standard deviation of the observations, in this instance, is a function of the standard deviation of the population, σ, and the size of the samples we drew from the population, N (Note: The reason why the standard deviation is $(\sigma/\sqrt{N})$ will be discussed later in the course.)

We can verify this logic by adding a few lines of code to the above R program.

```
Z <- (sampleMeans - mu) / ( sigma / sqrt(N))
hist(Z)
mean(Z)
sd(Z)
```



The output from the *mean*() and *sd*() functions in R generate a mean of 0.0003644679 and standard deviation of 1.00053. These values are effectively 0 and 1, which provides empirical support for our mathematical assertions about the distribution of Z.

Why is are these results of interest? Recall from the previous module on normal distributions, standardizing allows us to treat all problems dealing with normal probabilities within a general framework. In this case, we are standardizing the sampling distributions for the sample mean. This implies we will be interested in working with probabilities for obtaining a range of values for the sample mean. Thus, the sampling distribution for the sample mean is the fundamental basis for inference about the sample mean.

There is a crucial assumption that allows us to use Z as a basis for inference about the sample mean. The assumptions is we know the population variance, σ^2. Unfortunately, in most cases, both the mean and variance of the population are unknown and must be estimated from the data. In this case, we substitute the sample standard deviation, s, for the population standard deviation, σ, in the above equation. This yields a random variable T that has a different distribution from Z:

$$t = \frac{\overline{y} - \mu}{(s/\sqrt{N})} \sim T(N-1)$$

The notation $\sim T(N-1)$ is read as "T-distributed with N-1 degrees of freedom." The t-distribution is the standardized sampling distribution for the sample mean when:

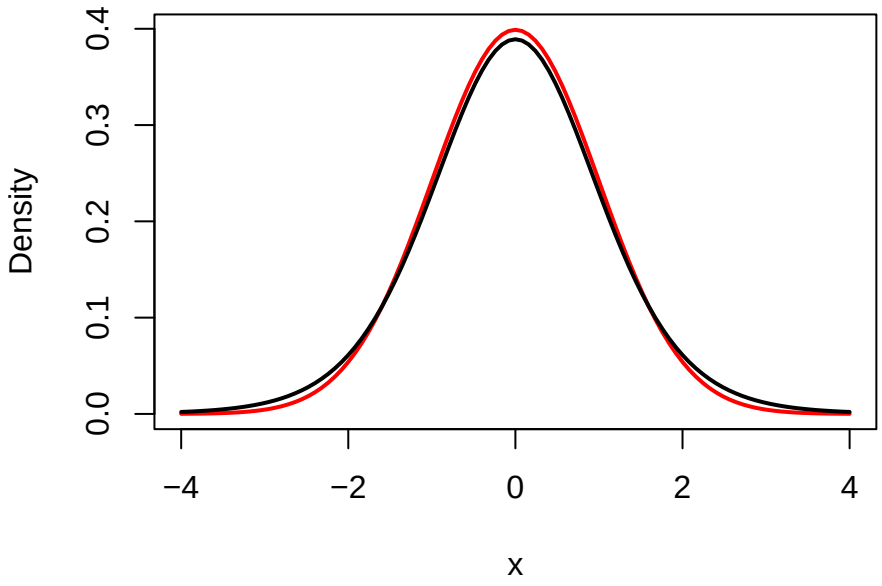
1. The mean is unknown and must be estimated from the data
2. The variance is unknown and must be estimated from the data
3. The sample is drawn from a normal population

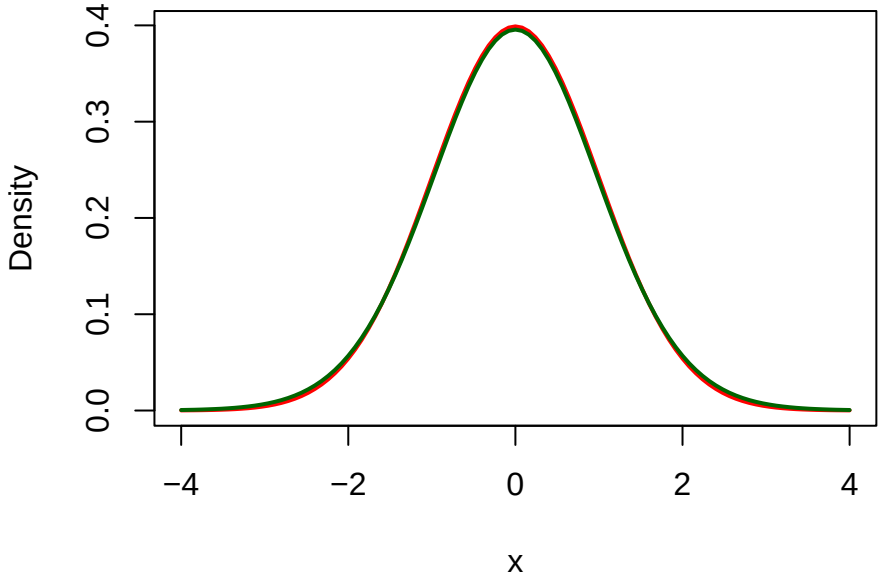
The difference between T and Z is most pronounced for small samples, and the difference gradually disappears as the sample size increases. The T and Z distributions are barely distinguishable for sample sizes $N \geq 30$. So, the normal distribution (Z) is typically used for statistical inference about the population mean for large sample sizes, while it is crucial to use T instead of Z for small samples.

The following plots illustrate the differences between T and Z. The top plot shows a normal distribution in red, and a T distribution, with 10 degrees of freedom and a sample of size of 11, in black. We can see the tails of the T distribution are a bit larger than the normal. This difference may be small, but it is important for obtaining correct p-values and confidence intervals for tests and interval estimates of the population mean. The bottom plot shows a normal distribution in red, and a T distribution with 30 degrees of freedom and a sample size of 31, in green. As mentioned above, the T distribution begins to approximate the normal distribution when the number of degrees of freedom is 30 or higher.

```
curve(dnorm(x), -4, 4, col = "red", ylab = "Density", lwd = 2) #Top plot
curve(dt(x, df = 10), add = TRUE, col = "black", lwd = 2)

curve(dnorm(x), -4, 4, col = "red", ylab = "Density", lwd = 2) #Bottom plot
curve(dt(x, df = 30), add = TRUE, col = "darkgreen", lwd = 2)
```


Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

3. Degrees of Freedom

The term *degrees of freedom* is used to distinguish the actual sample size from the effective sample size for a number of statistical testing procedures. For example, consider taking a random sample of n independent observations $y_1, y_2, ..., y_n$ from a normally distributed population. Imagine we want to estimate the population mean using these data. The best estimate of the population mean is the sample mean, $\overline{y} = \frac{1}{N}\sum y_i$. The effective sample size for estimating the population mean is equal to the total sample size N. In other words, any set of N randomly sampled observations from the population will suffice to estimate the population mean. What if we wanted to estimate the population variance, in addition to the mean? Recall, the best estimate for the population variance is:

$$s = \frac{1}{N-1}\sum(y_i - \overline{y})^2$$

The variance is almost an arithmetic mean of the squared deviation of each observation from the sample mean. We use the term *almost*, because the divisor is $N-1$ instead of N. This is because the effective sample size, or degrees of freedom, for estimating the population variance is 1 less than the total sample size. Why is this the case?

Carefully look at the definition of the sample variance. It requires computation of the sample mean, $\overline{y}$. Let's take a look at a concrete example with $N = 5$.

$$Y = \{2, 18, 9, 11, 10\}$$

In this example, $\overline{y} = \frac{1}{N}(2 + 18 + 9 + 11 + 10) = 50/5 = 10$. So, we can now use the value of 10 to compute the sample variance. However, we have placed a constraint on the sample observations by computing the sample mean. Specifically, the sum of all 5 observations in the sample must equal 50 to obtain a sample mean of 10. Thus, it is no longer the case that we can replace all 5 observations with new random draws from the population, as we could when we were interested in only estimating the mean, not the variance. However, we could replace 4 of the observations and still obtain a sum of 50. How? Imagine we sample 4 observations at random:

$$Y = \{1, 19, 10, 12\}$$

The sum of these is 42. Therefore, 5th observation in the sample must be 8 for all 5 observations to sum to 50. Thus, the number of independent observations (the effective sample size) for subsequent estimates

from the same sample has been reduced by 1, once we have computed the sample mean. Therefore, when we compute the sample variance, we are taking an average over $N-1$ independent observations, rather than N. Relating this information back to the Student's t-distribution, notice how the sample standard deviation is apart of the t statistic:

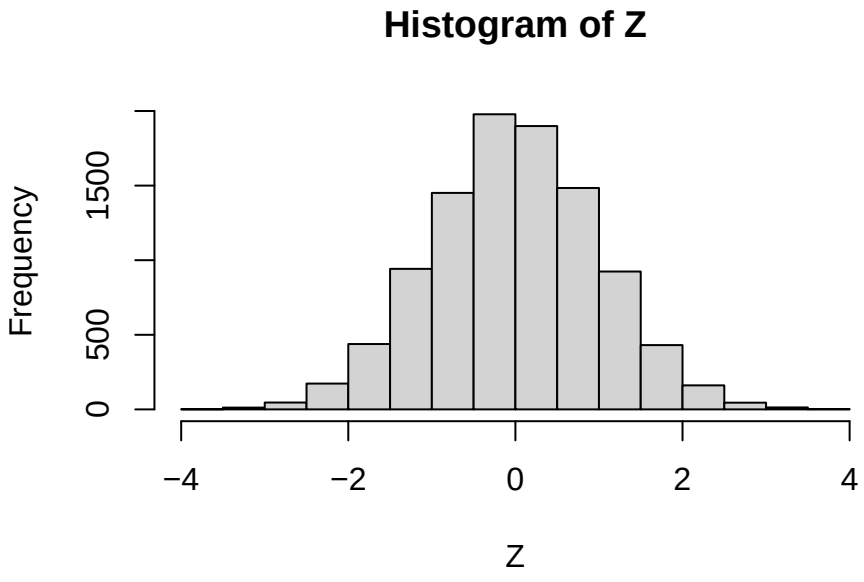
$$t = \frac{\overline{y} - \mu}{s/\sqrt{N}} \sim T(N-1)$$

Thus, the effective sample size is important to specify when determining the probability density function for the t statistic.

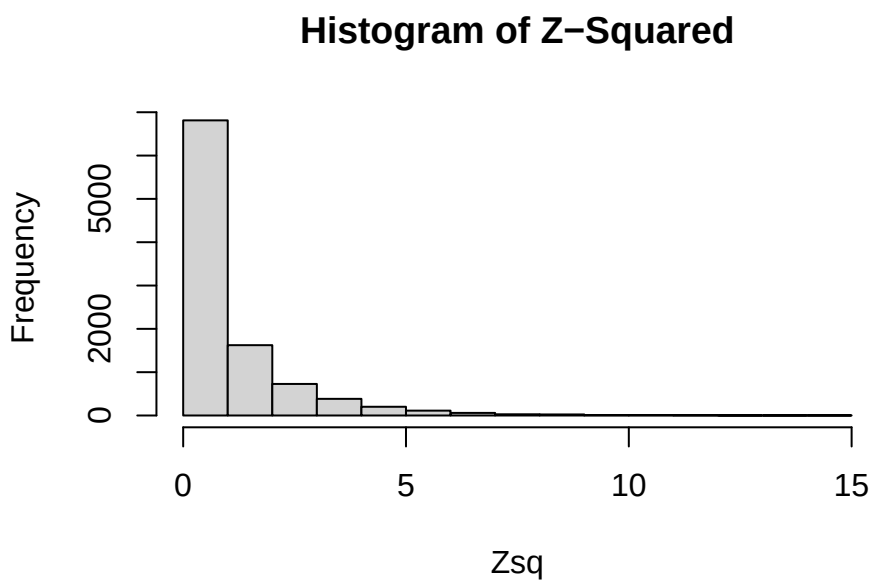
4. The Chi-square Distribution

If $Z_1, Z_2, ..., Z_k$ are independent and identically distributed standard normal RVs, then $C_k = Z_1^2 + Z_2^2 + ... + Z_k^2$ is distributed chi-square with k degrees of freedom. In other words, the Chi-square distribution is generated from the sums of squares of standard normal RVs. For example, consider the case of a single standard normal random variable Z. The $Z^2 \sim \chi^2_{df=1}$. A visual depiction is shown below:

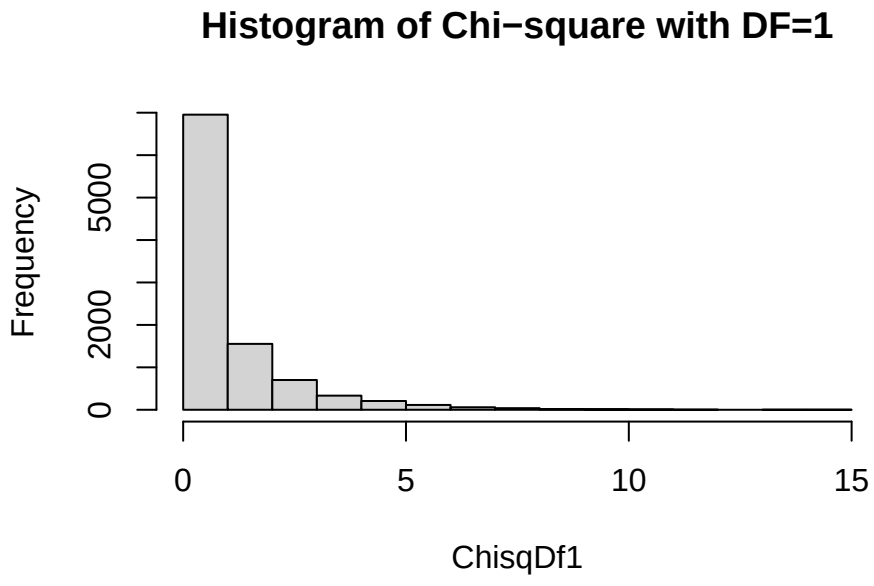
```
set.seed(123)
Z <- rnorm(n = 10000, mean = 0, sd = 1)
hist(Z, main = "Histogram of Z")
```



```r
Zsq <- Z^2
hist(Zsq, main = "Histogram of Z-Squared")
```

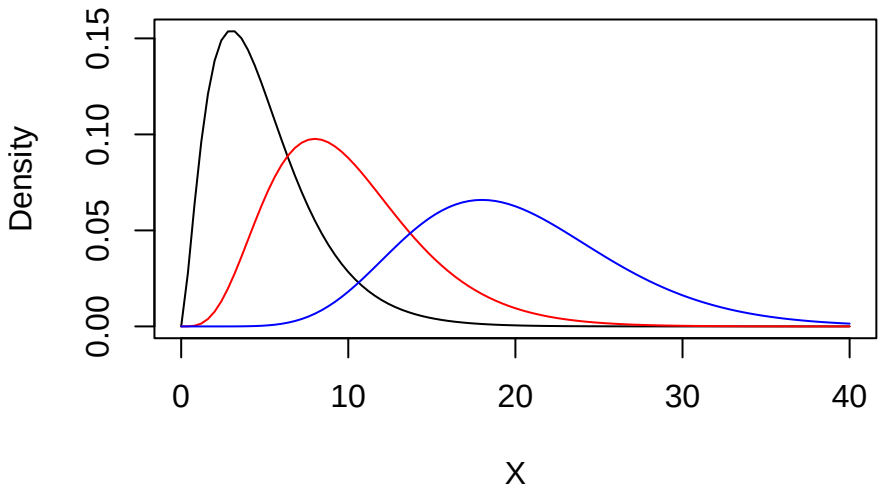



```r
ChisqDf1 <- rchisq(n = 10000, df = 1)
hist(ChisqDf1, main = "Histogram of Chi-square with DF=1")
```



It is easy to see that Chi-square distributions will have different shapes depending on the number of degrees of freedom (i.e., the number of squared normal variates). Here's an example of a few Chi-square distributions overlaid on the same plot:

```
curve(dchisq(x, df = 5), from = 0, to = 40, ylab = "Density", xlab = "X")
curve(dchisq(x, df = 10), from = 0, to = 40, add = TRUE, col = "red")
curve(dchisq(x, df = 20), from = 0, to = 40, add = TRUE, col = "blue")
```



The black line shows a distribution with DF=5, the red line has DF = 10, and the blue line has DF = 20. There are a few things to note about these distributions. First, they are all bounded on the left side by zero. Second, they are not symmetrical; however, they become more symmetrical as the DF increases. Try plotting a Chi-square distribution with DF=50 or DF=100 (Hint: We recommend expanding the x-axis range in these plots to see the entire curve). We will discuss the PDF and CDF for the Chi-distribution later in the course.

Many statistics are distributed chi-square, and this distribution is used a lot in statistics for hypothesis testing. The following isn't intended to be either comprehensive or self–contained. Instead, it provides a heads–up for when we encounter this information later in the course. The Chi-square distribution is used to make inference about the population variance. The key insight is that, for a sample size N, the quantity $\sum \frac{(y_i - \overline{y})^2}{\sigma^2}$ is distributed chi-square with $n - 1$ degrees of freedom (more on this later). The Chi-square distribution also contributes to the standard methods for modeling continuous outcome variables. Let $\hat{y}, \overline{y},$ and y denote the predicted value of the outcome, the mean of the outcome, and the observed value of the outcome, respectively.

The "ANOVA table," which summarizes the results of the modeling, contains a sum of $(\hat{y} - \overline{y})^2$ (the "model sum of squares"), and sum of $(\hat{y} - y)^2$ (the "error sum of squares"). When the additional assumption of normal errors is made, these become sum of squares of normal RVs, and thus, have Chi-square distributions.

Another application of the Chi-square distribution in modeling is through the Wald statistic:

$$(\frac{\hat{\beta}}{SE(\hat{\beta})})^2$$

Here, $SE(\hat{\beta})$ denotes the standard error of the regression coefficient β. Under the null hypothesis that the predictor(s) in question have no relationship with the outcome, the Wald statistic has a Chi-square distribution, with the degrees of freedom equaling the degrees of freedom associated with β (which could be a vector). For a single scalar predictor, this is a Chi-square distribution with 1 degree of freedom.

5. The F Distribution

Let χ_a χ_b Chi-square distributions with a and b degrees of freedom, respectively. Then, $(\chi_a/a)/(\chi_b/b)$ is an F distribution with degrees of freedom (a, b). As with the Chi-square and t-distributions, the F distributions all have different shapes according to the number of degrees of freedom (try plotting some F distributions in R). Here, mean squares are sums of squares divided by their degrees of freedom.

In the ANOVA table associated with modeling continuously scaled outcome variables, the numerator of the F–test is the model (explained) mean squared error, and the denominator of the F-test is the error (unexplained) mean squared error. Under the null hypothesis that the predictors in the model have no relationship with the outcome, this F-statistic has an expected value of 1. Larger values of the F-statistic support the conclusion that the model has predictive capacity. It is also interesting to note that the square of a random variable with a t-distribution, namely t^2, is a F-distribution with 1 and v degrees of freedom. This result is applied to partial regression coefficients in the linear regression model.

A theme that we are beginning to highlight is there are relationships among many probability distributions that are used in statistics!

6. Practice Problems

Note: These instructions assume you know how to plot PDFs and CDFs in R, and find quantiles and tail probabilities. There are R functions for doing all these tasks for the t and Z distributions. Some of these functions have been demonstrated in the above examples, and some can be determined through Google searches or reading R documentation. Ask questions if you get stuck!

Problem I

i. Use R to plot the PDF of a t-distribution with 5 degrees of freedom.

ii. Plot the CDF, too.

iii. Use R to obtain the 95th quantile of this distribution (i.e., the value of C where $Pr\{T \leq c\} = 0.95$).

Problem II

i. Calculate the values of the 95th quantile of the t-distribution with 5, 10, 15, 20, 25, 30, 50, 100, and 500 degrees of freedom.

ii. Compare the values obtained in i. with the 95th quantile of a standard normal distribution.

iii. At what point do these distributions become nearly identical?

Problem III

i. Consider the following random sample from a normally distributed population:

$$Y = \{44.40, 47.70, 65.59, 50.71, 51.29, 67.15, 54.61, 37.35, 43.13, 45.54\}$$

Compute the sample mean and variance.

ii. Treat the variance obtained from i. as known (i.e., assume it is the population variance) and the find the probability of observing the sample mean, or something larger, if the true population mean is 45.

iii. Treat the variance as unknown (i.e., estimate the population variance from the data) and the find the same probability as in ii.. Are these values different? Why or why not?

Problem IV

i. Simulate a data set with 100 individuals, each drawn from a standard normal distribution. Square these values, then add them.

ii. Is this sum near 100?

iii. What distribution has been simulated?

iv. Can you obtain similar results using an R function that samples from this distribution directly?

Problem V

i. Simulate 2 data sets with 100 individuals, with each sample being drawn from a standard normal distribution. Note, these samples are drawn from the same population, and should be statistically indistinguishable. Calculate an F-statistic.

ii. Is the value of this F-statistic near 1? Why, or why not?

Module 11: Other Distributions

1. Video

Click here to watch the video.

We will introduce some discrete distributions with countably infinite numbers of possible values (geometric, negative binomial, and Poisson) and a continuous distribution (exponential). Apart from those previously illustrated, these are the distributions which are most likely to be encountered by a practicing biostatistician. There are, of course, many other named distributions.

2. The Geometric Distribution

Given a sequence of Bernoulli trials, the geometric distribution gives the number of trials required to obtain the first success. Unless the probability of success for each trial (i.e., the Bernoulli parameter θ) equals 0, a success will eventually occur. This suggests that this is a proper distribution whose PMF sums to 1. On the other hand, there isn't a maximum number of trials (e.g., we could start with 10 failures, 100 failures, 1 million failures, etc.), meaning the number of possible trials is infinite. Moreover, it is countably infinite because these possible values can be placed in a one–to–one correspondence with the positive integers. The above statements are loose, and a more advanced inference course would make them more precise.

The PMF for the geometric distribution is the following:

$$Pr[Y = y] = \theta(1-\theta)^{y-1} \text{ for } Y = 1, 2, 3, ...$$

Note: The fact that the PMF is defined for $Y \geq 1$ indicates that the trial that resulted in the first success is included in the total number of trials required to obtain a success — more on this later.

The PMF for the geometric distribution can be derived from first principles. For example, imagine a doctor is trying to find a nausea medicine that works for a cancer patient who is getting sick from their chemotherapy treatments. Suppose there are many drugs the doctor could try, and the probability that any single drug works for the patient is 0.25. The doctor will keep trying drugs, until she finds one that works for the patient. Based on this scenario, we can define the following:

$\theta = Pr[\text{drug works}] = 0.25$ (a success)

$1 - \theta = Pr[\text{drug does not work}] = 1 - 0.25 = 0.75$ (a failure)

Y is equal to a random variable representing the number of drugs that must be tried to find one that works

(the number of trials required to obtain the first success; the trial where the success occurs is included in the number of trials that are required)

$Y \sim Geo(\theta)$ (Y has a geometric distribution with success probability θ)

Given this information, we can ask what is the probability that the first drug will work for the patient? If the first drug works, then $Y = 1$ (i.e., we had to try only 1 drug to find one that worked). Based on the PMF:

$$Pr[Y = y] = 0.25 * (0.75)^0 = 0.25$$

This is the probability of success on a single Bernoulli trial that we have defined. This should be intuitive (i.e., the chance that the first drug we try works is 0.25).

What is the probability that we'll have to try 2 drugs to find one that works? In this case, we are asking what is the probability that $Y = 2$? By the PMF, this probability is:

$$Pr[Y = 2] = 0.25 * (0.75)^1 = 0.1875$$

This can also be derived from first principles, based on statistical independence:

$$\begin{aligned} Pr[Y = 2] &= Pr[\text{trial 1 fails and trial 2 succeeds}] \\ &= Pr[\text{trial 1 fails}] * Pr[\text{trial 2 succeeds}] && \text{(Statistically independent probabilities)} \\ &= 0.75 * 0.25 && \text{(Plugging in 1-}\theta\text{ and }\theta\text{)} \\ &= 0.1875 \end{aligned}$$

The process proceeds similarly for larger values of Y. It should be intuitive that the probability will be small that we will need to try a large number of drugs (i.e., a large Y), before we find a drug that works.

*Note: There is an alternative definition of the geometric distribution that may be encountered in future coursework. This alternative definition gives the distribution for the number of trials required before a success is found (i.e., the number of failures that must occur prior to a success happening). The following is the PMF for this version of the geometric distribution (**pay attention to the difference in the exponent in the PMF, and the fact that the support set includes 0**)*:

$$Pr[Y = y] = \theta(1 - \theta)^y \text{ for } Y = 0, 1, 2, 3, ...$$

3. The Negative Binomial Distribution

The negative binomial distribution is an extension of the geometric distribution, and gives the number of Bernoulli trials required to obtain r successes (when $r = 1$, we have the geometric distribution).

We can derive the PMF from first principles:

Y = number of trials required to find r successes (Y is a random variable, but r is fixed in advance)
θ = probability of success on a single trial

For example, if $Y = 10$ and $r = 2$, then this is equivalent to saying that we needed to conduct 10 independent Bernoulli trials before we obtained 2 successes. Moreover, the probability statement would be about Y, not about r. The probability statement is: *What is the probability that 10 trials are required to obtain 2 successes?*

Even though the probability statement is about Y (the number of trials), we can use the PMF for the number of successes (r) to find the probability of interest. First, consider that one way to obtain 2 successes in 10 trials is:

1. Observe 1 success in 9 trials
2. Have the final trial result in a success

The probability for (1) can be obtained from the binomial distribution:

$$\begin{aligned} Pr(\text{r successes in n trials}) &= ({}_nC_r)\theta^r(1-\theta)^{n-r} \\ Pr(\text{1 success in 9 trials}) &= ({}_9C_1)\theta^1(1-\theta)^{9-1} \\ &= 9\theta^1(1-\theta)^8 \end{aligned}$$

The probability for (2) is θ, which is the probability of success on the 10th trial. Then, the probability of 10 trials resulting in 2 successes becomes the product of the probability of 1 success in 9 trials.

The probability of success on the 10th trial is (it's a product because these are independent events):

$$\begin{aligned} Pr(\text{2 successes in 10 trials}) &= Pr(\text{1 success in 9 trials}) * Pr(\text{success on 10th trial}) \\ &= ({}_nC_1)\theta^1(1-\theta)^{9-1}\theta \\ &= 9\theta^2(1-\theta)^8 \end{aligned}$$

Generally, the probability of k trials being required to obtain r successes is:

$$\begin{aligned} Pr(Y=k\Big|r,\theta) &= ({}_{n-1}C_{r-1})\theta^{r-1}(1-\theta)^{(n-1)-(r-1)}\theta \\ &= ({}_{n-1}C_{r-1})\theta^r(1-\theta)^{n-r} \end{aligned}$$

We have used the binomial distribution in an unconventional way. Here, we've fixed the number of successes (r) in advance. We normally use the binomial distribution to treat the number of successes as random.

4. The Poisson Distribution

The Poisson distribution is often applied to count data. That is, if we have a count, like the number of events occurring within a certain time interval or a certain spatial area, then the Poisson distribution should be considered.

The PMF for the Poisson distribution is:

$$Pr\{Y=y\} = \frac{\lambda^y e^{-\lambda}}{y!} \text{ for } y = 0, 1, 2, ...$$

The Poisson distribution is a limiting form of the binomial distribution. The proof, generally, begins with the binomial distribution. We make the number of trials large, the probability of success in any trial small, and keep $\lambda = n * \theta$ (the expected number of events per experiment) constant.

The Poisson approximation is ideal when:

- $N > 20$ and $\theta < 0.5$
- $N > 100$ and $n\theta < 10$

5. The Exponential Distribution

The exponential distribution is the continuous analog of the geometric distribution. It is often used to model the time to the first event.

The PDF:

$$P(Y = y) = \lambda e^{-\lambda y} \text{ for } Y \geq 0$$

The single parameter of the exponential distribution is λ, which is also known as the rate parameter. The exponential distribution "lacks memory," and is a natural choice when the past behavior of the system provides no information about future behavior. One application of the exponential distribution is in survival analysis when the hazard of death is constant.

6. Practice Problems

Problem I

Simulate data using Bernoulli trials with $\theta = 0.1$

i. How long did it take to obtain success number 1?

Replicate this simulation 1000 times (i.e., produce 1000 estimates of this waiting time).

ii. What is the maximum number of trials to obtain 1 success?

iii. What is the mean number of trials to obtain 1 success?

iv. How can the mean number of trials be used to estimate θ?

Problem II

Use the same simulation structure as in **Problem I**.

i. What is the maximum number of trials required to obtain 5 successes?

ii. What is the mean number of trials required to obtain 5 successes?

Problem III

Simulate the distribution of the number of successes out of 100 Bernoulli trials, each with success probability $\theta = 0.01$ How similar are the results to a Poisson distribution with a random variable with $\lambda = 1$?

Problem IV

For the exponential distribution, plot the CDF for $\lambda = 0.5, 1, 2,$ and 5.

Problem V

To demonstrate that the exponential distribution "lacks memory," consider $Pr\{Y > s + t \Big| Y > t\}$. By the definition of conditional probability, this equals $Pr\{Y > s + t \text{ and } Y > t\}/Pr\{Y > t\}$. But, $Y > s + t$ implies $Y > t$, and so $Pr\{Y > s + t \text{ and } Y > t\} = Pr\{Y > s + t\}$. Thus, $Pr\{Y > s + t \Big| Y > t\} = Pr\{Y > s + t\}/Pr\{Y > t\}$. The CDF for the exponential distribution is $Pr\{Y \leq k\} = exp\{-\lambda k\}$. Complete the algebra to demonstrate that $Pr\{Y > s + t \Big| Y > t\} = exp\{-\lambda s\}$, which does not depend on t, and this is equivalent to demonstrating that the exponential distribution lacks memory.

Module 12: Extra Material About Random Variables

1. Video

Click here to watch the video.

2. Properties of Simple Functions of One Random Variable

Denote the mean of the random variable Y as $E[Y]$ and the variance of Y as $Var[Y]$.

To shift the values of Y, take $Y^* = Y + c$, where c is a constant.

- The mean of $Y^* = E[Y] + c$
- The variance of $Y^* = Var[Y]$

To rescale the values of Y, take $Y^* = aY$, where a is a constant.

- The mean of $Y^* = a * E[Y]$
- The variance of $Y^* = a^2 * Var[Y]$

In general, the mean of y is:

$$E[Y] = \int y_i f(y_i) dy$$

Where the y_i are the possible values of y and $f(y)$ is the PDF, i.e., $f(y_i) = Pr[Y = y]$

The expected value of a function Y is similar:

$$E[g(Y)] = \int g(y_i) f(y_i) dy$$

3. Properties of Pairs of Random Variables

Many, but not all, of the properties of pairs of RVs will extend to no more than 2 RVs. When modeling large numbers of predictors, a key input is the "variance–covariance matrix." This contains all possible variances and covariances among pairs of variables. Thus, even though the general modeling theory applies to more than 2 predictors, a key input is limited to pairs:

When X and Y are independent:

- The mean of $X + Y = E[X] + [Y]$
- The variance of $X + Y = Var[X] + Var[Y]$

When X and Y are not independent:

- The mean of $X + Y = E[X] + [Y]$
- The variance of $X + Y = Var[X] + Var[Y] + 2 * Cov(X, Y)$

The covariance of X and Y is:

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

When X and Y have mean 0, $Cov(X, Y) = E(XY)$. Thus, when Y rises as the value of X increases, the covariance is positive. When Y falls as the value X increases, the covariance is negative. When X and Y are unrelated (i.e., independent), the covariance is 0.

The correlation is a scaled version of the covariance: $Corr(X, Y) = Cov(X, Y)/(\sigma_x \sigma y)$. When X and Y are standardized to have variances of 1, the covariance and correlation are identical. Correlations fall between -1 and 1.

The regression coefficient, β, resulting from fitting the line $Y = \alpha + \beta X$, is a rescaled version of the correlation, and also the covariance :

$$\beta = (\sigma_y/\sigma_x)Corr(X, Y)$$

Thus, the sign of β indicates whether Y rises or falls as the value of X increases.

Some properties of covariance and related quantities include:

- $Cov(X, Y) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, c) = 0$
- $Cov(aX, Y) = aCov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(X, Y)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Var(X + c) = Var(X)$
- $Cov(cX) = c^2 Var(X)$
- If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$
- If X is constant, then $Var(X) = 0$

There are two equivalent ways to write the variance:

- $E[(Y - E[Y])^2]$
- $E[Y^2] - E[Y]^2$

4. Laws of Conditional Expectation

The derivation of Baye's Theorem makes use of the law of total probability. There is an analogous law of total expectation. This law is useful in the causal inference course, among others. To derive the law of total expectation in the discrete case, we start with the law of conditional expectation:

$$E_x(X|Y = y) = \sum_x (xPr\{X = x|Y = y\})$$

Here, the value of Y is fixed, and the element that varies is X, and so the notation E_x is used to emphasize that the expectation is taken over X. Now, take the expectation of the above quantity over Y:

$$E_y\{E_x(X|Y = y)\} = \sum_y \{\sum_x (xPr\{X = x|Y = y\})\}Pr\{Y = y\}$$

This can be rewritten in terms of a joint density:

$$E_y\{E_x(X|Y = y)\} = \sum_y \{\sum_x (xPr\{X = x \text{ and } Y = y\})\}$$

The order of summations can be reversed (Note: for continuous RVs, this step requires that a regularity condition hold.):

$$E_y\{E_x(X|Y = y)\} = \sum_x x[\sum_y (Pr\{X = x \text{ and } Y = y\})]$$

Applying the law of total probability:

$$E_y\{E_x(X|Y = y)\} = \sum_x (xPr\{X = x\})$$

Then simplifying:

$$E_y\{E_x(X|Y = y)\} = E[X]$$

This is often written in the rather cryptic notation: $E(E[X|Y]) = E[X]$, which is suppressing the variables

the expectations in question are being taken with respect to.

Here are some properties of conditional expectation that will prove useful in 709:

- $E[E[Y|X]] = E[Y]$
- If X and Y are independent, then $E[Y|X = x] = E[Y]$ and $E[X|Y = y] = E[X]$
- $E[g(x)Y|x] = g(x)E[Y|x]$
- $E[aY_1 + bY_2 \Big| X] = aE[Y_1|X] + bE[Y_2|X]$

5. Practice Problems:

Problem I

Show that $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.

Problem II

Show that the regression coefficient in a simple linear regression model, $Y = \alpha + \beta X$, can be expressed as $\beta = Cov(X, Y)/Var(X)$.

Problem III

Let $X \sim N(10, 100)$ and $Y \sim N(20, 100)$ be independent normal random variables. What are the mean and the variance of $X + Y$?

Problem IV

Now assume X and Y from **Problem III** are repeated measures of the same outcome in one arm of a randomized controlled trial. Meaning, instead of being statistically independent, X and Y are now correlated with $r = 0.7$. The outcome of the trial is the change score, i.e., $X - Y$. What are the mean and the variance of $X - Y$? (Hint: The answer to this question will require methodology similar to *i.*, but consider that $X + Y$ is equivalent to $X + (-Y)$.)

Problem V

Show that $E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2$

Module 13: Additional Information on Functions

1. Video

Click here to watch the video.

2. Discussion

To date, we have been rather informal in referring to functions. Without devolving into a fully axiomatic treatment, a brief discussion of functions from the perspective of mathematics will be helpful, as this information will be used in this and other courses.

A function is a rule that takes an input, x, and produces an output, $f(x)$ over a range of possible values of the input, x. For example, if the function is $f(x) = x + 1$, and if the input is $x = 1$, then the output is $f(x) = 2$. Often, $f(x)$ is denoted by y.

When there is a single input to the function, the input can be plotted against the output — in the above example, the result is a straight line with a slope of 1 and an intercept of 1. This behavior can also be described as $\{(x, f(x) = x + 1)\}$, where the complete definition of the function also requires specifying the range of possible values of x.

Let $Y = g(X)$. Then for a discrete RV:

$$\begin{aligned} Pr\{Y = y\} &= Pr\{g(X) = y\} \\ &= Pr\{X = x_1\} + Pr\{X = x_2\} + ... + Pr\{X = x_n\} \end{aligned}$$

where $x_1, x_2, ..., x_n$ are the values of X for which $g(X) = y$. A similar principle applies to continuous RVs, although expressing it precisely would require somewhat different notation.

A function is 1–1 if, and only if, $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. That is, each x in the domain is mapped to a unique y in the range.

A function is invertible if every value of the output, y, maps to a unique input, x. In other words, for an invertible function, we can go back and forth between the function and its inverse. For example, the function $f(x) = x^2$ is not invertible, because $f(x) = 4$ maps to both $x = -2$ and $x = 2$. However, this function is invertible when the range of possible values of x is limited to positive numbers.

The above discussion in mathematical notation: If $f^{*-1}(f(x)) = x$ then f is invertible and $f^{*-1}(x)$ is the inverse function of $f(x)$. That is, the inverse of a function is a function.

If $g(x)$ is invertible, then $Pr\{Y = y\} = Pr\{X = x\}$ where $x = g^{*-1}(y)$.

Change of variables: If $Y = g(X)$ is invertible, then $f_y(y) = f_x(x)^* \left| dx/dy \right|$.

We don't cover the extension to the change of variables formula to the multivariable case.

3. The Indicator Function and Boolean Algebra

A common function in statistics is the indicator function, denoted as $I\{^*\}$, where:

$$I\{\text{logical expression}\} = 1 \text{ if true, } 0 \text{ if false}$$

One application of this idea is it changes an element of logic, which isn't in a directly computable form, into a RV (i.e., it can be added, multiplied, it has a mean, etc.). Indeed, the RV in question has a Bernoulli distribution.

For example, this table maps the value of Y into the indicator variable $I\{Y > 5\}$.

Y	$Y > 5$	$I\{Y > 5\}$
$Y \leq 5$	No	0
$Y > 5$	Yes	1

Boolean algebra starts with Bernoulli RVs, typically derived from logical conditions, and then uses calculations to represent operations, such as "and" and "or." For example, "and" is the same as $I\{A\} * I\{B\}$ (i.e., the result will only be 1 if $A = 1$ and $B = 1$). "Or" is $(1 - I\{A\}) * (1 - I\{B\})$. (Note: Technically, this is an "exclusive or" condition, also denoted XOR, where the expression evaluates to 1 only when either A or B is 1, and evaluates to 0 when both A and B are equal to 1).

4. Taylor's Theorem

Taylor's theorem allows us to approximate the value of a function, as long as it has at least k derivatives at the point $x = a$. An advanced inference course will use a more precise statement. This discussion is for exposure, because it may be encountered later in the curriculum.

Any $f(x)$ can be approximated at $x = a$ by $f(a) + f^{'}(x)(x - a) + 2!f^{"}(x)(x - a)^2 + ...$

For example, to approximate $f(x) = e^x$ at 0, we utilize the facts that $f^{'}(e^x) = e^x$ and $e^0 = 1$. The Taylor series expansion then becomes:

$$f(x) = e^a + e^x(x - a) + 2e^x(x - a)^2 + ...$$

Now plug in $x = 0$ and simplify:

$$1 + x + 2!x^2 + ... + k!x^k$$

5. Practice Problems

Mathematical functions have an analog in computer programming languages. For example, R functions take input and return output like a mathematical function. Write the following functions in R. Make sure to test these functions for valid and invalid inputs (e.g., the binomial distribution cannot have probability less than 0 or greater than 1).

Problem I

Write a function that returns the CDF for the following distributions. Show that each function produces the correct answer by comparing the function output with the analogous built-in function in R for returning the CDF.

i. Standard Uniform Distribution

ii. An arbitrary Uniform Distribution

iii. Binomial Distribution

Module 14: Intro to Likelihood

1. Video

Click here to watch the video.

2. Introduction to Inference

Until now, we have considered random variables, RV, their distributions, and some properties of those distributions. In actual practice, distributions are selected based on the characteristics of the population under study. For example, if the distribution of LDL cholesterol values appears to be approximately bell–shaped, then the normal distribution is a natural choice to model these data. Distributions can be assigned based on empirical considerations (e.g., the data appear normal), theoretical considerations (e.g., LDL values are determined by a large number of small, independent perturbations), or both. From now on, we assume that the distributions have been well chosen.

The link between actual data and statistical inference is through the parameters of a well chosen distribution that is used as a model for the data. For example, for LDL values assumed to be normally distributed, we can ask which values of μ are "sufficiently consistent" with the data. Asking such questions about μ; generally, asking such questions about parameters of any distribution, is the essence of statistical inference.

3. The Likelihood Function

Consider 7 independent Bernoulli trials, each with success probability $\theta = 0.6$. We are interested in the total number of successes, Y. Therefore, the distribution in question is binomial with $N = 7$ and $\theta = 0.6$. The PMF is given below:

PMF of Bin(n=7,$\theta = 0.6$)

X	$Pr\{X = x \mid \theta = 0.6\}$
0	0.00164
1	0.01720
2	0.07741
3	0.19354
4	**0.29030**
5	0.26127
6	0.13064
7	0.02799

Notice that the experiment has not yet been performed, so we do not know the value of Y, which will be observed. Y is random and θ is fixed, and these facts are illustrated by the following notation:

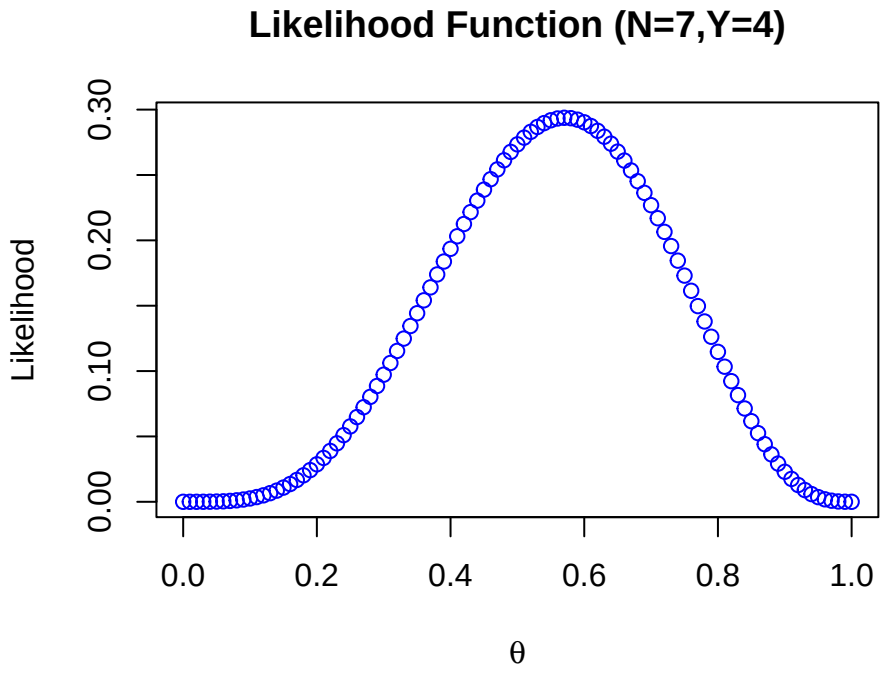
$$Pr\{Y|\theta\} = {}_nC_y\theta^y(1-\theta)^{n-y}$$

Now suppose that we observe $Y = 4$, and want to ask whether the data are consistent with the value $\theta = 0.6$. The of value of $Y = 4$ will be observed almost 30% of the time, which provides informal support to the notion that θ might be 0.6. Here, θ is unknown to the investigator, so we can only derive an educated guess.

Plugging $\theta = 0.5$ into the above formula should yield $Pr\{Y = 4 \mid \theta = 0.5\} = 0.27344$. The value of $Y = 4$ will observed almost 30% of the time, which provides support to the notion that θ might be 0.5, although there is marginally less support for $\theta = 0.5$ than for $\theta = 0.6$.

Now, let's consider the same information from a different perspective. In particular, we will treat Y as fixed (i.e., the experiment has already occurred) and vary the value of θ, the unknown parameter of interest.

An exercise asks that the values of θ vary from 0 to 1 by 0.01. The resulting plot is given below. We find that the maximum value of θ occurs at 0.57, which also happens to be 4/7. We are using the same formula as before, but with 2 key differences. First, the value of Y is now fixed to be what was observed in the data. Second, the formula is interpreted to be a function of θ, and describes how $Pr\{Y = 4\}$ varies as a function of θ.



The function can be used to make inferences about the true, but unknown, value of θ. Indeed, we've already done this informally: the data seem consistent with both $\theta = 0.60$ and $\theta = 0.50$, and the value of θ most consistent with the data is 0.57. Another way to phrase "most consistent with the data" is "the most likely value of θ given the data." Therefore, the function in question is termed the "likelihood function." We denote it by $L\{\theta \big| Y\}$ to emphasize that this is a function of θ, conditional on a fixed value of Y. In this example, we may use the notation $L\{\theta \big| Y = 4\}$.

The value of θ that maximizes the likelihood function is called the maximum likelihood estimator (MLE) of θ, denoted by $\hat{\theta}$. For that matter, the value of θ that maximizes the likelihood function also maximizes the value of the log of the likelihood function. It is often simpler to work with the log-likelihood.

4. Practice Problems

Problem I

Write some R code to plot the likelihood function, from $\theta = 0$ to 1 by 0.01, for the above experiment.

Problem II

Now perform 100 Bernoulli trials and observe $Y = 30$ successes.

i. Guess what MLE for θ should be

Plot the likelihood function from $\theta = 0$ to 1 by 0.01.

ii. Is the answer in i. correct?

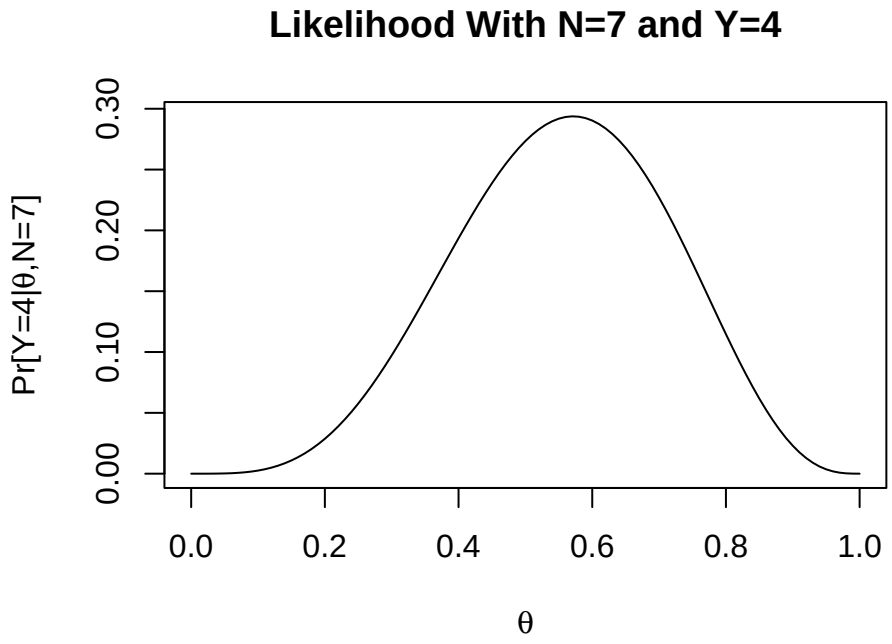
Module 15: Properties of the Likelihood Function

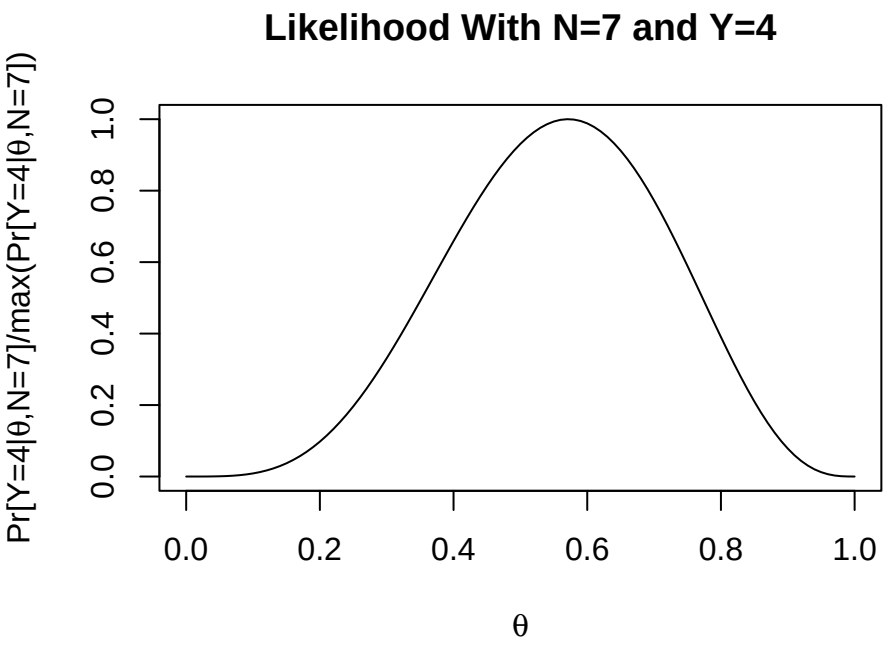
1. Video

Click here to watch the video.

2. Background on Plotting the Likelihood Function

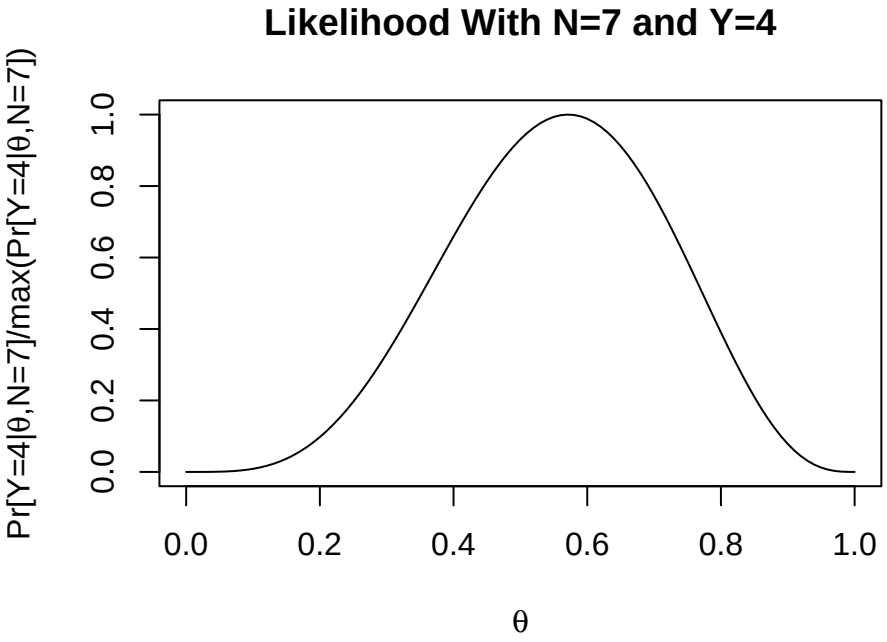
We begin this discussion by returning to our earlier example of the binomial likelihood function for 4 successes out of 7 trials. It is common for likelihood functions to be "scaled" before they are plotted. In other words, the likelihood at each value of θ is divided by the largest value of the likelihood, such that the y-axis of the plot ranges from 0 to 1. An example is given below of the scaled and unscaled binomial likelihoods for $Y = 4$ and $N = 7$. In the rest of the course, we can assume all plotted likelihood functions have been scaled. We have provided supplemental code for these plots.

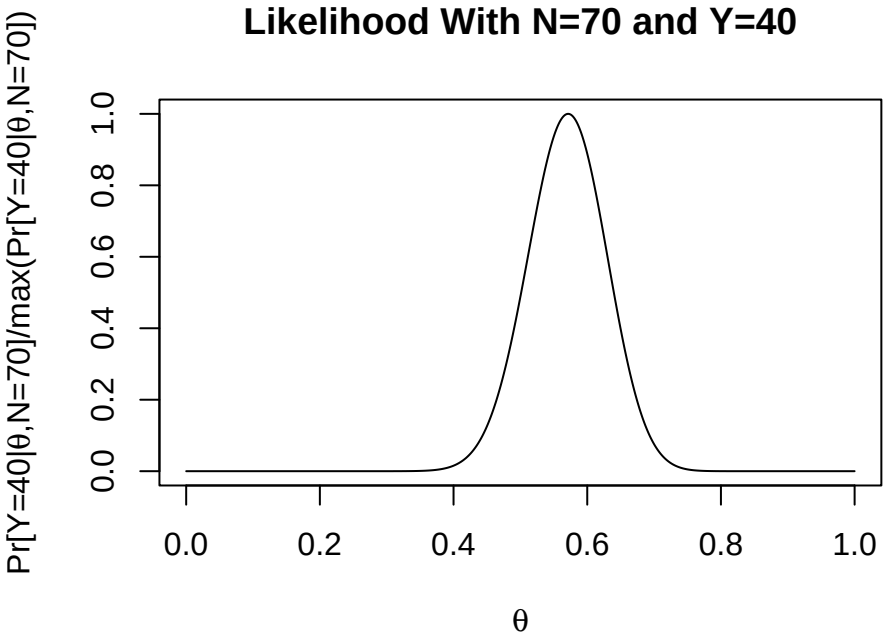



Likelihood With N=7 and Y=4
Pr[Y=4|θ,N=7]/max(Pr[Y=4|θ,N=7])
0.0
0.2
0.4
0.6
0.8
1.0
θ

3. Using the Likelihood Function for Statistical Inference

Much of statistical inference is based on the notion of generating parameter estimates using maximum likelihood. As an illustration, we've plotted the likelihood functions for 2 binomial experiments: the first has 4 successes out of 7 trials (the same plot as above) and the second has 40 successes out of 70 trials.





The signal is the same. That is, they yield the same value of the maximum likelihood estimate (MLE) for θ (the peak of the curve is at $\theta = 0.57$). The second experiment has less "noise." That is, it is consistent with a smaller range of likely values of θ, because it was generated from a larger sample size. For example, in the first experiment a value of θ of 0.40 is relatively consistent with the data, whereas in the second experiment a value of 0.40 is not.

The noise in the estimate of θ (the imprecision in the signal) is related to the likelihood function's curvature. Relatively speaking, the first curve is flat and the second curve is steep.

Flat curves are:

1. Consistent with a greater range of plausible values of θ
2. Have smaller second derivatives (in absolute value) than steep curves

Recalling from calculus that the second derivative of a function evaluated at a specific point quantifies its curvature at that point. Another way to state the above is that the larger the second derivative (in absolute value), the steeper the likelihood function and the more "information" the data provides about the actual value of θ.

The likelihood function (LF) can be used to generate point estimates, confidence intervals, and hypothesis tests, and these are the central components of statistical inference.

4. Using the LF to Generate Point Estimates

In the binomial example (a sequence of N Bernoulli trials) the MLE for θ is Y/N (the number of successes divided by the number of trials). More generally, this "point estimate" is associated with the maximum value of the likelihood function. Recall that calculus can be used to find the global maximum for a function. To apply that concept here, we would find the first derivative of the likelihood function, set it equal to zero, and solve for θ.

In reality, we don't work directly with the likelihood function. Instead, we work with the natural log of the likelihood function, because this makes our work with calculus easier. The properties of the likelihood function that we discuss in this course all relate to the log–likelihood.

For a vector of observed data Y and a vector of possible values of θ, the likelihood function is:

$$L(\theta) = Pr[Y|\theta]$$

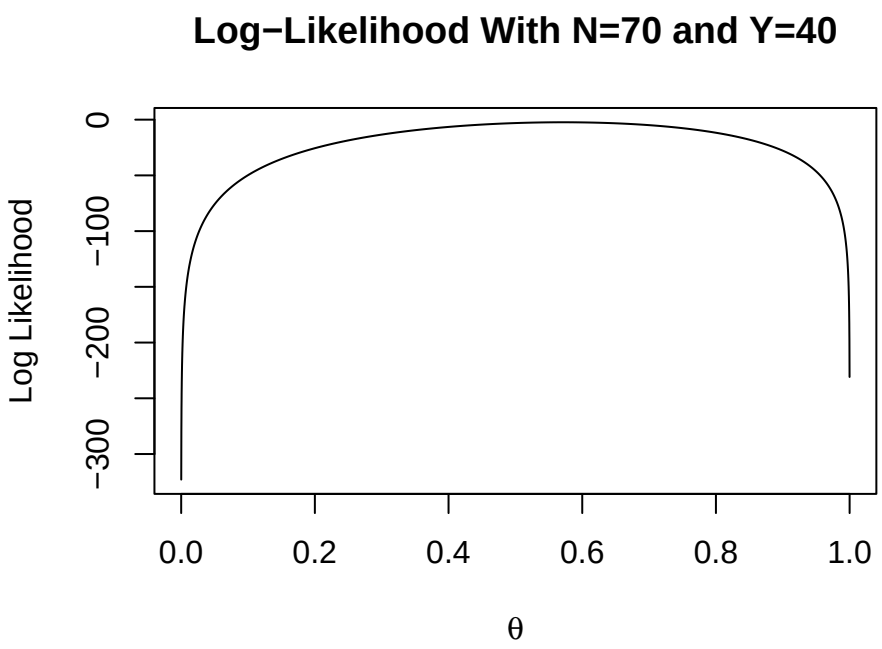
We already have a function for $Pr[Y\Big|\theta]$, which is the probability mass function for a binomial random variable (y is the observed number of successes and θ is the probability of success on each trial):

$$L(\theta) = Pr[Y|\theta] = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

The log-likelihood is:

$$logL(\theta) = ln\Big[\binom{n}{y}\theta^y(1-\theta)^{n-y}\Big] = ln\Big[\binom{n}{y}\Big] + yln(\theta) + (n-y)ln(1-\theta)$$

As an example, here's what the plot of the log-likelihood looks like for $Y = 40$ and $N = 70$:



To find the maximum of the log-likelihood function, we take the first derivative, set it equal to zero, and solve for θ. The first derivative of the log–likelihood is called the *score function*, and it is usually denoted as $S(\theta)$:

$$S(\theta) = \frac{d}{d\theta}logL(\theta)$$

For the binomial likelihood, we have the following (some of the math has been skipped for brevity):

$$\begin{aligned}
S(\theta) &= \frac{d}{d\theta}\Big[ln\Big[\binom{n}{y}\Big] + yln(\theta) + (n-y)ln(1-\theta)\Big] \\
&= \frac{d}{d\theta}\Big[ln\Big[\binom{n}{y}\Big] + \frac{d}{d\theta}yln(\theta) + \frac{d}{d\theta}(n-y)ln(1-\theta)\Big] \\
&= 0 + \frac{d}{d\theta}yln(\theta) + \frac{d}{d\theta}(n-y)ln(1-\theta)\Big] \\
&= \frac{y}{\theta} - \frac{n-y}{1-\theta}
\end{aligned}$$

Notice in the third line that the natural log of the combination, $ln\Big[\binom{n}{y}\Big]$, is a constant whose derivative is zero. For this reason, we will see that the log–likelihood of the binomial, and other PMF/PDF that have constants in them, is written without the constant term. We skipped most of the math to expand the remaining parts of the derivative. However, the work can be accomplished by applying basic rules for derivatives and simplifying using algebra.

Now, if we set the score function equal to zero and solve for θ, we should get an unsurprising result:

$$\begin{aligned}
S(\theta) &= 0 \\
\frac{y}{\theta} - \frac{n-y}{1-\theta} &= 0 \\
\frac{y}{\theta} &= \frac{n-y}{1-\theta} \\
(\frac{y}{\theta})\theta(1-\theta) &= (\frac{n-y}{1-\theta})\theta(1-\theta) \\
y(1-\theta) &= (n-y)\theta \\
y - y\theta &= n\theta - y\theta \\
y &= n\theta \\
\theta &= \frac{y}{n}
\end{aligned}$$

Therefore, the MLE for θ is Y/N, the sample proportion.

Keen observers will notice that we skipped the work of identifying y/n as a global maximum. However, we can see by plotting the function that there are no local maxima in this case. This actually leads to an important statement in likelihood inference regarding *regularity*. In simple terms, a log–likelihood is *regular* if it can be approximated by a quadratic function. In the case of regular likelihoods, we can use the simple rules of calculus to find characteristics, like the maximum and the curvature.

5. Extra Words on Point Estimates

An area of statistical inference covered elsewhere discusses optimal properties of estimators, and in some instances, produces rules under which optimal estimators can be derived. The Wikipedia entry under "sufficient statistics" contains an accessible summary of this material (including a discussion, with examples, of how it all comes together). The presentation of this material typically begins by limiting the estimators under consideration to "unbiased estimators" (i.e., estimators whose expected value is the parameter under consideration). For example, for a binomial random variable, $E(Y/N) = \theta$, so Y/N is an unbiased estimator of θ.

In general, we'd also like estimators to have as small a variance as possible. If we have two unbiased estimators, we'd prefer the one with smaller variance. In practice, the criteria around bias and variance can conflict. This is called the "bias–variance trade–off." An implication is we might occasionally prefer an estimator with a modest amount of bias, if its variance is much less than the variance of an unbiased estimator.

We will discuss later in the course properties of estimators and how estimators are derived.

6. Using the LF to Generate Confidence Intervals

In general, the "confidence interval" for θ is the set of all values of θ whose LF values are sufficiently large (i.e., plausible values of θ, given the observed data). The confidence interval will always contain the point estimate.

The simplest way to generate a confidence interval is to use information contained in the peak of the likelihood function. Recall from the above discussion that a steep curve represents a scenario where we are almost certain about the value of θ. This curvature is described by the second derivative of the function. We are specifically interested in quantifying the curvature of the LF at the MLE. This curvature is called the amount of *Fisher information*, and it is defined as below:

$$I(\theta) = -\frac{d^2}{d\theta^2} logL(\theta)$$

Notice that we define $I(\theta)$ as the negative second derivative. This is because the second derivative itself is negative. By changing the sign, we can interpret larger numbers as corresponding to tighter peaks, and thus, more certainty about the value of θ. In the case of the binomial distribution, this looks like the following (much of the mathematics has been skipped for brevity):

$$
\begin{aligned}
I(\theta) &= -\frac{d^2}{d\theta^2} logL(\theta) \\
&= -\frac{d}{d\theta} S(\theta) \\
&= -\frac{d}{d\theta}\Big[\frac{y}{\theta} - \frac{n-y}{1-\theta}\Big] \\
&= \frac{y}{\theta^2 + \frac{n-y}{(1-\theta)^2}}
\end{aligned}
$$

Remember, we are interested in the information at the MLE. What does it mean for us to find this number? The MLE is our estimate for θ:

$$\hat{\theta} = \frac{y}{n}$$

It is also true that the expected number of successes is the product of $\hat{\theta}$ and n:

$$y = \hat{\theta} n$$

We can plug this value for y into the equation for $I(\theta)$ and obtain the observed information at the MLE:

$$I(\hat{\theta}) = \frac{\hat{\theta}}{\hat{\theta^2}} + \frac{n - n\hat{\theta}}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

Note, this quantity is the inverse of the variance of a binomial random variable. We can assume that for regular likelihood functions that it is acceptable to conclude:

$$Var(\hat{\theta}) = I^{-1}(\hat{\theta})$$

And likewise, the standard error of $\hat{\theta}$ is:

$$SE(\hat{\theta}) = I^{-1/2}(\hat{\theta})$$

Now, we can use the inverse of the observed Fisher information to construct a Wald 95% confidence interval. The Wald interval assumes that $\hat{\theta}$ is normally distributed. The Wald 95% confidence interval is:

$$\hat{\theta} \pm 1.96 * SE(\hat{\theta})$$

For our experiment with $Y = 40$ and $N = 70$, we have:

$$\frac{40}{70} \pm 1.96 * \sqrt{\frac{\frac{40}{70}(1 - \frac{40}{70})}{70}} = (0.46, 0.69)$$

The interval for our experiment with $Y = 4$ and $N = 7$ is:

$$\frac{4}{7} \pm 1.96 * \sqrt{\frac{\frac{4}{7}(1 - \frac{4}{7})}{7}} = (0.20, 0.94)$$

Notice that the MLE in both cases is the same, but in the second experiment, which has a smaller sample size (number of trials), the interval is wider. This represents the fact that there is more uncertainty about θ in the smaller experiment. This uncertainty is due directly to the larger curvature of the likelihood function which is represented by the standard error. The standard error is the square root of the inverse of the observed Fisher information, which measures the curvature of the likelihood function.

Note, the Wald interval is not the only way to construct a confidence interval. Other approaches to building confidence intervals will be discussed later. The Wald interval is sufficient when regularity conditions hold, which happens for the binomial random variables when the number of trials is large and the success probability is not close to 0 or 1.

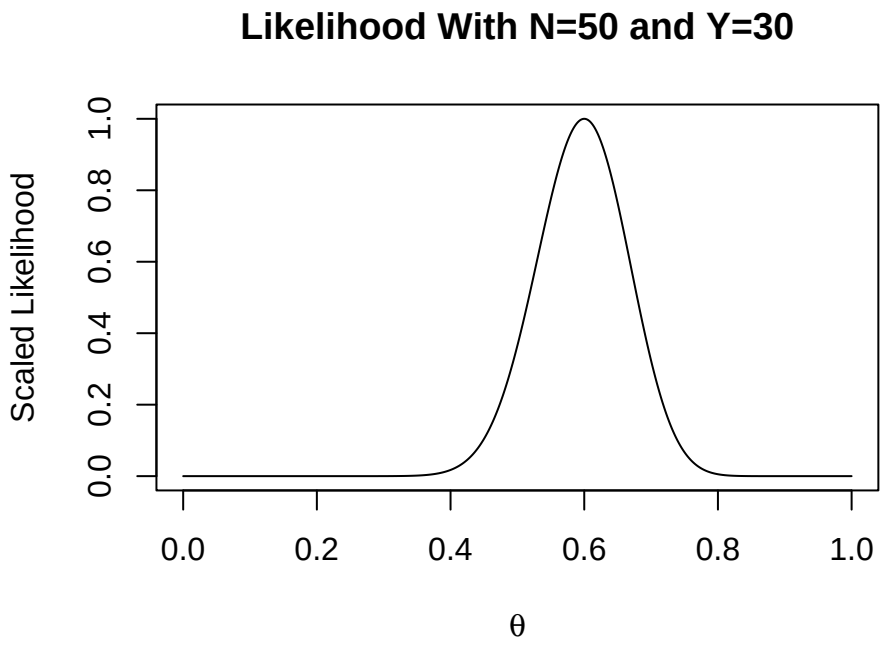
7. Using the LF to Generate Hypothesis Tests

Consider two hypothesized values of θ; namely, θ_0 and θ_1. The "strength of the evidence in favor of θ_0" is quantified by $L(\theta_0)$. Similarly, the strength of the evidence in favor of θ_1 is quantified by $L(\theta_1)$.

The strength of the evidence in favor of the two hypothesized values could potentially be compared in various ways — it turns out that $L(\theta_0)/L(\theta_1)$ has good statistical properties. This is a "likelihood ratio (LR) hypothesis test" — the closer the LR is to 1, the less distinguishable are the two hypotheses. In practice, we usually take logarithms, and then calculate $-2(logL(\theta_0) - logL(\theta_1))$. This value has a chi–square distribution under the null hypothesis (i.e., that $\theta = \theta_0$). The degrees of freedom are 1 for simple hypothesis tests, like this.

8. Practice Problems

The below plot shows the likelihood function for a binomial experiment with 50 trials and 30 observed successes.



Problem I

Based on the plot, what is the MLE for θ?

Problem II

If we are defining a confidence interval for θ as all values of θ for which $LF > 0$, what is the confidence interval for θ, based on the plot?

Problem III

What is the 95% Wald confidence interval for θ?

Problem IV

Do a likelihood ratio test comparing the possible values $\theta = 0.6$ and $\theta = 0.4$ for the probability of success. Show your work and write your interpretation of the results of this test.

Module 16: Sampling Distributions

1. Video

Click here to watch the video.

2. Background Discussion

When thinking about estimating parameters, a natural place to start is a point estimator, often the MLE. For example, with 60 successes out of 100 Bernoulli trials, the point estimate for θ is 60/100=0.60. However, we still need to estimate its precision. For example, 0.60 $\pm$ 0.01 yields a very different, and more precise, conclusion than 0.60 $\pm$ 0.25.

To derive this precision, the sampling distribution is key. If we could replicate a study many times, we could directly observe the variance of the estimator Y/N (Y is the number of successes and N is the number of trials for the binomial random variable). For example, if we observe that, over replications of the study, the point estimates varied around the MLE within a range of 0.10 units, then our estimate is MLE $\pm$ 0.05. Of course, if we actually had replications of the study, we'd want to incorporate information from those studies into the estimate of θ. Even without having actually replicated the study, we nevertheless still want to know how much the point estimates of θ vary from study to study, and this is what the sampling distribution will tell us.

The general approach to inference is to use the observed data to derive the MLE, and the sampling distribution to derive the precision of that estimate. Unfortunately, it is often the case we have only a single study. Amazingly, the sampling distribution can be estimated from that study, rather than directly observed.

Within statistics, there are multiple ways to apply this paradigm to generate confidence intervals. Two common approaches are:

1. Likelihood-based methods
2. Bootstrapping

3. Likelihood-Based Methods

We have already seen an introduction to the likelihood-based approach in the last module (i.e., we use the curvature of the likelihood function an an indicator of the information the sample has about the parameter). This gets translated into a SE that can be used to create the confidence interval using the familiar form: MLE $\pm$ 1.96SE. Although it was tacit in the last module, this approach to creating a confidence interval embeds an implicit assumption: the MLE (denoted in module 15 as $\hat{\theta}$) is normally distributed. Meaning, in repeated samples from the same population, the sample mean follows a normal distribution. This fact was echoed in earlier modules as a statement of the Central Limit Theorem (CLT): sample means are approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

For example, suppose that Y is continuously scaled, $n = 100$, we observed a mean of 50 in a single sample, and that the standard deviation is 10. Applying the CLT, we can say that the sampling distribution for the sample mean for studies of size $n = 100$ is approximately normal with mean μ (the population mean) and standard deviation $\sigma/10$. Recall, we also know that for any normal random variable 95% of observations will fall within 1.96 standard deviations of its mean. Moreover, we know that if $\mu = 50$ and $\sigma = 10$, then in 95% of studies the sample mean will be between 50-(1.96)(10/10)=48.04 and 50+(1.96)(10/10)=51.96.

Of course, we don't know μ. However, we do know that $\bar{y}$, observed from the raw data, can be considered to be a single draw from a sampling distribution. That sampling distribution has the property that, 95% of the time, a random draw from it will fall within $(\mu - 1.96SE, \mu + 1.96SE)$, where SE is the standard deviation of the sampling distribution. If we apply this information to the example we just showed for the mean of a continuously scaled outcome, then $SE = \sigma/\sqrt{n}$. The standard error might look different for other types of random variables (e.g., binomial random variable); the form is similar, meaning that it is still the standard deviation of the population divided by the square root of the sample size.

If we consider the observed data, a 95% confidence interval for μ is $(\bar{y} - 1.96SE, \bar{y} + 1.96SE)$. If σ is unknown, as is usually the case, we replace it with s, the sample standard deviation. If the sample size is large enough, then estimating the population standard deviation through the sample standard deviation is not much of a concern.

4. Bootstrapping

Bootstrapping is a synthetic way to estimate the sampling distribution. Once the sampling distribution is estimated, the confidence interval is generated in essentially the same manner as above. To bootstrap a study with sample size $n = 100$, we create R replicates of the study by sampling 100 observations with replacement from the original data set. As a result, the samples will tend to be similar to the original data set, but not identical. This is a desired trait in a sampling distribution. The 95% confidence interval can then be based on this distribution, either by computing the standard error and using the above formula or by using 2.5 and 97.5 percentiles of the distribution. If the distribution is reasonably symmetric, which should be the case if R is large enough, then the two approaches yield similar results.

The frequentist interpretation of the confidence interval isn't that "we're 95% confident" about anything; instead, it is that we have a process which, if repeated numerous times, will "get it right" 95% of the time. Here, "get it right" means "true, but unknown value of μ falls within our confidence interval."

5. Practice Problems

consider a study generating a binomial RV with 100 individuals and 30 successes.

Problem I

What is the MLE for θ?

Problem II

What is the Wald 95% confidence interval for θ?

Problem III

What is a 95% confidence interval using 100, 1000, and 10000 bootstrapped samples?

Module 17: Hypothesis Testing

1. Video

Click here to watch the video.

2. Background Discussion

Hypothesis tests are closely related to confidence intervals. For example, a hypothesis test wil be statistically significant at $\alpha = 0.05$ if, and only if, the 95% confidence interval fails to contain the null value. For example, suppose that an investigator hypothesizes that for a binomial random variable $\theta = 0.40$ and the 95% confidence interval is (0.20-0.30). This interval does not contain the null value of 0.40, so the null hypothesis is rejected. We conclude that the value of $\theta \neq 0.40$.

Hypothesis tests and confidence intervals are closely related because they share building blocks in "signal" and "noise" (in the above example, the maximum likelihood estimate, or MLE, is the signal and the width of the confidence interval is the noise). The noise depends on the sample size. The role of the sampling distribution is similar in both hypothesis testing and confidence intervals. Likewise, the role of the sampling distribution is similar in both of the common approaches to inference discussed in module 16: likelihood methods and bootstrap. The interpretation of a hypothesis test is the same regardless of how the sampling distribution is estimated.

Hypothesis testing is used throughout statistics to answer two types of questions:

- Is a hypothesized value of the parameter plausible?
- Is an observed relationship in the data "real," or is it consistent with what we'd expect due to sampling variability (this is often called "chance")?

The above example illustrates the first question by asking whether $\theta = 0.40$. An example of the second question, in the context of linear regression, is asking if $\beta = 0$ (this would mean there is no relationship between the predictor and the outcome). The second question is a special case of the first: namely, where the parameter in question is specified to imply that there's no relationship.

Hypothesis testing terminology and process remain unchanged, regardless of the which of the two above questions we are asking. The parameter value that is directly specified, which in the latter question is the "null value" if no relationship exists (e.g., $\beta = 0$), is called the "null hypothesis." The null hypothesis has a special status, because, counterintuitively, it is the more important of the two hypotheses that are formed. Indeed, careful language states hypothesis testing outcomes as "fail to reject the null hypothesis" or "reject

the null hypothesis." To be clear, rejecting a null hypothesis is no affirmation that we have accepted the alternative hypothesis. However, it is simpler to interpret the mechanics of hypothesis testing in terms of deciding on one explanation or the other. This is how the material is presented here.

Hypotheses can be classified as 1–sided or 2–sided, depending on the specification of the parameter values under the alternative hypothesis. Continuing the above binomial example, $\theta > 0.40$ is an example of a 1–sided alternative hypothesis, whereas $\theta > 0.40$ or $\theta < 0.40$ is an example of a 2–sided alternative hypothesis.

When forming 95% confidence intervals for a 2–sided test, we use the 2.5th and the 97.5th percentiles of the sampling distribution. For a 1–sided test, we use the parameter values above the 5th percentile or below the 95th percentile, depending on the direction of the alternative hypothesis. The distinction between the 1–sided and 2–sided hypothesis tests is not always clear. For example, when dealing with multivariable modeling, a simple conceptualization of "direction" doesn't apply. We can also have a situation with a composite null hypothesis and, typically, a 1–sided alternative hypothesis. For example, H_0 could be $\theta \leq 0.40$, meaning H_a would be $\theta > 0.40$. We will not go into detail, but this situation is resolved by analyzing the test as if the null hypothesis is $\theta = 0.40$.

In the above linear regression example, where the null hypothesis is "no (linear) relationship exists between the predictor and the outcome" (i.e., $\beta = 0$), the investigator must choose one of two conclusions: the relationship is real or it is not. This merely means the investigator must make a judgment call from the observed data. There are 4 possibilities for the investigator's conclusion. Among the 4 outcomes, there are two ways the investigator can be correct and there are two ways the investigator can be incorrect. The investigator's decision about H_0 is based on the observed data. This information is summarized in the table below:

Investigator's Decision	**Truth (Unknown to Investigator)**	
	Relationship Real	Relationship Not Real
Relationship Present	Correct	Type I Error (False Positive)
Relationship Not Present	Type II Error (False Negative)	Correct

Recall that confidence intervals describe the results of applying a process many times. The Type I error rate (α) and the the Type II error rate (β) do not refer to a singular hypothesis test, rather they both refer to a large number of tests. Power, the probability of observing a relationship when it is real, is denoted by $1 - \beta$. Power will be discussed in another module.

It is convention that $\alpha = 0.05$ and that $\beta = 0.10$ or $\beta = 0.20$. These are set by the statistician when the

study is designed, prior to data collection. Interestingly, the sample size required for the study is determined partly by the specification of these error rates.

To use the likelihood approach to perform a hypothesis test:

1. Form a confidence interval for θ, and then decide whether θ_0 is within that interval. If θ_0 is not in that interval, then reject H_0, or
2. If H_a contains a single value (e.g., $H_a : \theta = 0.30$), compare the value of the LF at $H_0 : \theta = 0.40$ with the value of the LF at H_a (i.e., compare $L(\theta = 0.40)$ with $L(\theta = 0.30)$. If the H_a contains a range (e.g., all values of θ not equal to 0.40), take the maximum value of the LF within H_a and compare this to the value of the LF at H_0.

To use bootstrap to perform a hypothesis test:

1. Form a confidence interval for θ based on the bootstrap sampling distribution (e.g., use the values of θ at the 2.5 and 97.5 percentiles), and then determine whether θ_0 is within that interval. If θ_0 does not appear in that interval, then reject H_0. This is essentially the same process as the likelihood–based approach.
2. Alternatively, use the bootstrap sampling distribution to determine the proportion of estimated values of θ that are larger or smaller than θ_0. Meaning, we look at the proportion of estimated values of θ that are "at least as extreme" than θ_0.

The second bootstrap approach is effectively equivalent to estimating a p–value. The p–value is a continuous value that tells us how much information the data have against the H_0. We will illustrate this idea using the PDF approach. Suppose that for a study with a binomial RV, $n = 100$ and $\theta_0 = 0.40$, observed values of $Y = 40$ are consistent with the H_0, values like $Y = 38$ and $Y = 43$ are quite consistent, and values like $Y = 100$ are the most inconsistent. We can do the following to generate a p–value:

1. Calculate the PDF under H_0
2. Using this PDF and the observed value of $Y = y$, we can calculate the probability that either $Y = y$ or $Y = y^*$, which is a more "extreme" value than the observed $Y = y$

Loosely speaking, a p–value assumes that the H_0 is true, and then it asks how extreme are the observed data. If the observed data are sufficiently extreme, the possible explanations are either "the H_0 is true and we have bad luck" or "the H_0 is false." If the p–value is sufficiently low, we decide on the latter explanation. The number where the p–value is considered "sufficiently low" is a hotly debated topic, because this is the instance at which interpretation of a p–value becomes a hypothesis test. For example, we only reject H_0 when the p–value is less than 0.05. Since the α level for a test is arbitrary, many have argued for abandoning such a black–and–white interpretation. Instead, these people suggest that p–values should be interpreted on a continuum. Nonetheless, hypothesis testing remains common, especially in highly regulated disciplines where clear boundaries for decision making are required (e.g., drug development). This is not to say that H_0 significance testing is the only avenue for decision making (e.g., there are Bayesian methods based on posterior probabilities for θ exceeding a given value), but it remains a standard. This statement from the American Statistical Association is a good place to begin to learn more.

3. Practice Problems

Consider an experiment with a binomial random variable, where the number of observed successes is 35 out of 100 trials.

Problem I

i. Test $H_0 : \theta = 0.40$ (i.e., $H_a : \theta \neq 0.40$) using likelihood-based and bootstrap methods

ii. Report the p–value from both methods and write an interpretation of both values

iii. Write an interpretation of the p–values as continuous numbers against the common threshold of $\alpha = 0.05$

Problem II

i. Estimate the 95

ii. Write an interpretation of both intervals

iii. Use each of the intervals to test $H_0 : \theta = 0.40$ against $H_a : \theta \neq 0.40$

Module 18: More About Hypothesis Testing

1. Video

Click here to watch the video.

2. Discussion

For our purposes, a statistical hypothesis is a statement about a parameter θ. This parameter could be a scalar, as for the binomial probability of success, or a vector, as for the mean and standard deviation of a normal distribution. As mentioned in module 17, if the hypothesis completely specifies the PDF $f(x, \theta)$, then it is a simple hypothesis; otherwise, it is a composite hypothesis. For example, $\theta = 0.40$ is a simple hypothesis. Whereas, $\theta \leq 0.40$ is a composite hypothesis. A composite hypothesis does not precisely specify the value of its parameter, and instead specifies a range.

The null and alternative hypotheses correspond to different subsets of the "parameter space" (i.e., the set of possible values of θ). For a binomial random variable, the parameter space for θ ranges from 0 to 1. The null and alternative hypotheses do not have to fill the entire parameter space. For example, $H_0 : \theta = 0.40$ and $H_a : \theta = 0.30$.

As discussed in module 17, suppose Y is a binomial random variable and $H_0 : \theta = 0.40$ (the null hypothesis often specifies a single parameter value). An example of a 2–sided alternative hypothesis is $H_a : \theta > 0.40$ or $\theta < 0.40$. An example of a 1–sided alternative hypothesis is: $H_a : \theta > 0.40$.

A "statistic" is a value calculated from the data; for example, a sample mean. Sample statistics are used to make inference about population parameters; for example, to derive point estimates or confidence intervals, and to perform hypothesis tests. Sample statistics are random variables.

The "rejection region" (or "critical region") for a hypothesis test consists of all the values of the test statistic that lead to rejecting the null hypothesis. Similarly, the "acceptance region" consists of all values of the test statistic that we lead to failing to reject the null hypothesis. In the binomial example, if $H_0 : \theta = 0.40$ and $H_a : \theta = 0.30$, the rejection region will consist of small values of Y. One role of statistical theory is to set appropriate boundaries between the acceptance and rejection regions. The placement of the boundaries depends on, among others, the choice of the Type I error rate. When p–values are used instead of a previously specified value of α, the shape of the acceptance and the rejection regions is the same as if α were a fixed value. The next step is to take each possible value of the test statistic and classify it as being more extreme, less extreme, or as extreme as the value that was observed.

One–sided hypotheses are controversial. Continuing with the binomial example, where $H_0 : \theta = 0.40$:

- $\theta > 0.40$ might be impossible
- $\theta > 0.40$ might be uninteresting

The latter case is the more problematic of the two.

Efficiency is an advantage gained from using a 1–sided hypothesis. Rather than stacking 2.5% of values somewhere below 0.40 and the other 2.5% of values somewhere above 0.40, the rejection region can contain the values of Y corresponding to the top 5% of the distribution. Thus, some large values of Y will fall into the rejection region of a 1–sided test, but not of a 2–sided test. This is why investigators often choose to use 1–sided hypotheses. One way investigators justify their selection of a 1–sided test is by setting $\alpha = 0.025$. Meaning, if we are to conduct a 1–sided hypothesis test at $\alpha = 0.025$ and a 2–sided test at $\alpha = 0.05$, then the rejection region in the 1–sided test covers the same values of the test statistic as the analogous side of the rejection region in the 2–sided test. Thus, the 1–sided test at $\alpha = 0.025$ and the 2–sided test at $\alpha = 0.05$ maintain the same Type I error rate. They also require the same sample size for a given power and detectable effect size, but we will discuss this topic in more detail later in the next module. However, there are reasons, aside from statistical error rates, why one prefers to use a 2–sided test over a 1–sided test. For example, it might be important to know why a drug's performance is unexpectedly worse than the placebo. Regardless, users of 1–sided tests and 2–sided tests both agree that the hypotheses should be specified in full ahead of time.

For a simple null hypothesis, the probability of rejecting a true H_0 (denoted as α, the probability of making a Type I error) is called the significance level of the test. For a composite null hypothesis, the size of the test, or size of the critical region, is the maximum probability of rejecting H_0 when it is true (maximized over the values of the parameter under H_0). Meaning, the size of the test represents the worse-case scenario for the false positive rate.

In general, a Type I error is considered to be more serious than a Type II error. For example, incorrectly declaring a useless drug to be efficacious leads to bad consequences, both legally and in terms of the public health. Science is intentionally conservative, in the sense that strong evidence is required to overturn the status quo (this is represented by the H_0). However, a study with low power places subjects at risk for no good reason (i.e., failing to reject the H_0 is baked into the cake), and should be avoided.

Therefore, it is best that both Type I and Type II error rates be small. However, there is a trade–off in that decreasing one error rate simultaneously increases the other. The only way to improve both error rates is to

increase the sample size. If this cannot be done, then it is best to improve the study design. The Type I and Type II error rates are often set by tradition. However, it is best to be cautious of cases where the traditional settings do not apply. For example, in drug development for fatal diseases, one might be willing to have a larger value of α in the early stages of drug development. This will increase the power of the study and will avoid missing a promising candidate for the drug. (Recall that β and the power of the study are providing essentially the same information.) The power function, $\pi(\theta)$, provides the probability of rejecting H_0 when the true value of the parameter is θ. This topic will be discussed in more detail in the next module.

The following information is to supply background knowledge for future discussion. The "Neyman–Pearson lemma" essentially describes how to build a test with maximum power. Simply stated, given a specific value of α for a simple H_0:

- Place values of the test statistic into the critical region until α is reached
- These values should be more likely under H_a than H_0, with "more likely" defined by $L(\theta_a)/L(\theta_0)$
- The first value of the test statistic should be "most likely" under H_a, the second value of the test statistic should be the second most likely under H_a, and etc.

To illustrate the use of this lemma, consider a binomial random variable with $N = 100$ and the hypotheses $H_0 : \theta = 0.40$ and $H_a : \theta > 0.40$. To make matters simpler, we will specify $H_a : \theta = 0.80$. Furthermore, we will stipulate that θ cannot be equal to 1 under H_a (since probabilities can never be exactly 1). The first value to be placed into the critical region is $Y = 100$, the second value to be placed into the critical region is $Y = 99$, and the third value is $Y = 98$, and etc. To illustrate the likelihood for $Y = 100$ (out of 100 trials) under $H_0 : \theta = 0.40$:

$$L(\theta = 0.40|Y = 100, n = 100) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{100}{100}0.4^{100}(1-0.4)^{100-100} = 1.6(10^{-40})$$

This expression gives the probability of observing 100 successes in 100 trials if the true probability of success on each trial is 0.40. The probability is small, and in fact, it is much smaller than 0.05. Therefore, $Y = 100$ belongs in the rejection region for the test of $H_0 : \theta = 0.40$.

Notice that the likelihood of observing 100 successes in 100 trials under $H_a : \theta = 0.40$ is:

$$L(\theta = 0.80|Y = 100, n = 100) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{100}{100}0.8^{100}(1-0.8)^{100-100} = 2(10^{-10})$$

The ratio of these likelihoods is (2(10^{-10}))/(1.6(10^{-40}))), which is a large number, 1.25(10^{30}). Meaning, 100 successes in 100 trials is much more likely under the assumption that $\theta = 0.8$ than $\theta = 0.4$. In other words, larger values of Y are more consistent with H_a, whereas smaller values of Y are more consistent with H_0.

If we repeat this process for successively smaller values of Y (i.e., $Y = 99$, $Y = 98$, etc.), then we will find a likelihood under $H_0 : \theta = 0.40$ that is greater than 0.05. At that point, we will have found the boundary of the rejection region. In this example, we have only considered a boundary in one direction (i.e., values of $\theta > 0.40$), because of the phrasing of the hypotheses.

Note, for discrete random variables, like the binomial example, we might not be able to precisely find a value of Y that lands exactly on the desired $\alpha = 0.05$. Although there exist more nuanced approaches, a conservative solution is to let the Type I error rate fall below 0.05, while being as close to 0.05 as possible.

Practice Problems

Problem I

Use $\alpha = 0.05$, a binomial RV with $n = 100$, and $H_0 : \theta = 0.60$ to answer the following:

i. Find the boundary of the rejection region for $H_a : \theta = 0.8$ (it is fine to complete this problem with programming statements, instead of using pen and paper)

ii. Write an interpretation of the boundary

iii. What would the boundary be if $\alpha = 0.025$?

Module 19: Power

1. Video

Click here to watch the video.

2. Background Discussion

The power function, $\pi(\theta)$, produces the probability of rejecting H_0 when the true value of the parameter is θ. We will use a computer simulation to develop this idea.

Consider a simple randomized experiment with control and treatment groups. In such a study, participants are randomly assigned to treatment or control, they receive their designation, and their outcome is measured some time later (for the moment, we will not consider adjustment for baseline values). Assume the outcome in the control group is sampled from a population with mean 50 and standard deviation 10. If the treatment is to be helpful, it would have to increase the mean of the outcome among treated participants by at least 5 points. We will assume that prior research suggests that the treatment is capable of accomplishing this goal. It is our responsibility as biostatisticians to design this experiment such that it has the greatest possibility of identifying whether the treatment is helpful at the desired level of 5 points. Meaning, we want out experiment to have "high power."

Based on the description above, we have two random variables, both with hypothesized distributions:

- $C \sim N(50, 10^2)$
- $T \sim N(55, 10^2)$

The variable C represents the distribution of the outcome in the population of people who would receive the control therapy. The parameters of this distribution are usually set based on previous research (e.g., if the control is a standard of care that has been studied thoroughly, then we probably have good estimates of the mean and standard deviation for planning our new trial). The variable T represents the distribution that the outcome *would need to have* among treated patients for the treatment to be considered helpful. We do not know if T follows this distribution, but if it does, then we want to be sure our experiment has a high probability of distinguishing patients sampled from either T or C.

> Brief digression: In this example, we have no reason to believe that the standard deviation of the outcome would be any different in treated patients than it is in patients who receive the control. Therefore, we have assumed that the standard deviation for T and C are equivalent. This "common variance" assumption may or may not be applicable to our experiment. It is based on the specific therapeutic area we are working in. Making this assumption in this example simplifies the problem.

To continue, the parameter we want to estimate is the difference in the mean outcome comparing patients who receive the treatment and patients who receive the control; namely:

$$\theta = \mu_T - \mu_C$$

This parameter, and others like it that give the size of the difference between randomized groups (e.g., hazard ration, odds ration, etc.), is occasionally called the "treatment effect" in clinical trials. The null and alternative hypotheses we want to test about this parameter are as follows:

- $H_0 : \mu_T - \mu_C = 0$
- $H_a : \mu_T - \mu_C > 0$ or $\mu_T - \mu_C < 0$

The H_a is 2–sided. Thus, we could have stated H_a as $\mu_T - \mu_C \neq 0$. The H_a is a composite hypothesis, because it specifies a range of values for the parameter θ. The "clinically meaningful" value for θ is 5, which is the value of θ under the H_a that we will use to calculate power.

Since we are comparing means of normally distributed random variables, we will apply the two–sample t–test. Suppose that we would like to know whether 50 patients in both the treatment group and the control group is satisfactory for detecting a shift in means from 50 to 55 points (i.e., $\theta = 5$) with a common standard deviation of 10 points in both groups. We could do the following to estimate the power for this t–test:

1. Generate multiple replications of the experiment (e.g., 5000)
    a. For each replication, simulate data according to the specifications of the power analysis (i.e., simulate 50 random draws from $N(50, 10^2)$) and 50 random draws from $N(55, 10^2)$
2. For each replication, perform the t–test and obtain a p–value
3. For each replication, create an indicator variable denoting the presence of statistical significance (e.g., whether the p–value is less than 0.05)
4. Calculate the mean of the indicator variable across the multiple simulation replications. The result is the estimated power (i.e., the proportion of simulated studies that yielded the correct rejection of the null hypothesis). This proportion is an estimate of the following conditional probability:

$$\pi(\theta = 5, n = 50, \sigma = 10, \alpha = 0.05) = Pr[\text{Reject } H_0 | \theta = 5]$$

Here, the conditioning event implies that the H_0 is false and the H_a is true; specifically, the true treatment effect is 5 points. Meaning, power is estimated for the detection of a 5–unit difference in means, assuming a common standard deviation of 10 with 50 patients per group and a Type I error rate of 5%.

Here is syntax for the algorithm:

```
n_studies=1000

for i=1 to n_studies{
    control=50          #random draws from N(50,100)
    treatment=50        #random draws from N(55,100)
    p_value=t.test(treatment,control).pValue
    if p_value<0.05 then sig=1
    else if p_value>=0.05 then sig=0
}

power=mean(sig)
print(power)
```

However, this algorithm only estimates power for 1 value of θ, which is the clinically relevant value. To create a "power function," we need to keep the assumptions about the control group (i.e., $C \sim N(50, 10^2)$), vary θ across a range (e.g., 0 to 10 increment by 1), and calculate the power for each value of θ using the above method. Note, varying θ across a range results in sampling from a different treatment group distribution (i.e., $T \sim N(50 + \theta, 10^2)$). When $\theta = 0$, we are actually simulating a case where the null hypothesis is true.

We expect, in this scenario, the proportion of studies where the null hypothesis is rejected is equal to 5%, because all these rejections would be Type I errors. In fact, for a fixed sample size per experiment, we find that the power increases as the absolute value of θ increases. Recall that the treatment effect can be in two directions, and our study will still have greater power for increasingly negative or positive values of θ. Negative values of θ imply the treatment is worse than the control. If statistical significance in the negative direction is observed, then we probably do not want to use the treatment.

More generally, a power calculation has 3 elements (i.e., sample size, effect size, and Type I error). In the t–test example, the effect size is the difference between the group means, in standard deviation units. Large effect sizes correspond to variables whose impact is large. During study design, a power calculation helps the investigator choose a sample size with good statistical qualities:

- Good power at an important value of θ under the H_a
- Modest loss of power if the estimate of θ is optimistic (i.e., the estimate of θ lies a relatively flat part of the curve)
- Not wasteful (i.e., we cannot obtain similar power with a much smaller sample size)

The calculations can also work in the opposite direction. For example, given a previously specified sample size and statistical power, the investigator might ask what magnitude of effects can be observed. This version of the calculation is used when the investigator is constrained in the number of individuals they can study (e.g., the size of the patient population, the budget, etc.).

In simple cases, power calculations have closed-form solutions; thus, a formula can be applied. Nevertheless, it is good practice to check the results using simulation to verfiy that we have used the formula correctly. When 2 tests are being considered, efficiency quantifies the ratio of their powers. For example, non–parametric tests (illustrated in another course) can be surprisingly efficient in comparison with their parametric analogs.

3. Practice Problems

Problem I

i. Use simulation to plot the power function, as described above, for $\theta = \{0 \text{ to } 10 \text{ by } 1\}$, assuming 50 patients per group, a control group distribution of $C(50, 10^2)$, and a 2–sided alpha of 5

ii. Use the t–test in the simulation, assuming equal variances

iii. Generate 5000 replications of the study

Problem II

Repeat the simulation, but now use the t–test under the assumption that the population variances are not the same.

Problem III

Repeat the simulation, and instead of using a t–test, apply the Wilcoxon rank sum test. Use the normal approximation for the Wilcoxon rank sum test, if there is such an option in the statistical software that you are using. Note, this option may be the default setting.

Problem IV

i. Describe the differences observed in power and Type I error across the 3 tests, based on the simulations created in Problems I, II, and III

ii. Which test do you recommend using, and why?

Module 20: Multiple Testing

1. Video

Click here to watch the video.

2. Discussion

The issue of multiple testing is relevant to randomized trials, genomics, and variable selection, among others.

The core insight is that individual hypothesis tests build in the possibility of false positive conclusions, through the selection of the Type 1 error rate, α. In other words, since every test has the possibility of generating a false positive, it follows that every positive (i.e., statistically significant) result might be a false positive. Moreover, when more tests are performed, there is expected to be more statistically significant results. Accordingly, it is important to be able to estimate the likely number of false positives. For example, if the observed number of false positives is no greater than the expected number, the statistically significant results might not be considered seriously.

These ideas can be most simply introduced *via* independent tests. It should be recognized that, in actual practice, tests are not necessarily independent. There are various specific procedures which account for multiple testing, and these might be optimal. For a practicing biostatistician, mastery of general principles around multiple testing is usually satisfactory.

Two tasks around multiple testing include:

1. Estimating the number of false positive tests
2. Estimating the probability that at least one test is a false positive

The probability in (2) is called the experiment-wise error rate, to distinguish it from the test-wise error rate for any particular test. The experiment-wise error rate might or might not be a relevant criterion to apply to any given study. This is because there is more to interpreting the results of an experiment than statistical significance. Often, one can manage the occasional false positive result by considering the number, magnitude, direction, and clinical plausibility of the statistically significant (and non-significant) results.

The expected number of false positive results is $k\alpha$, where k is the number of tests and α is the Type 1 error rate per test. More generally, it is the sum of the $\alpha's$, which allows the α to differ from test to test. An application in genomics appears when the test-wise error rate for k tests is set to $.05/k$ (referred to as the Bonferroni correction). However, in genetics it is often unrealistic to strictly control the experiment-wise

Type 1 error rate in this fashion. This is because k is typically very large. In these cases, a statistician might turn to a different approach, which is to control the expected number of false positive tests instead.

If all the tests are independent and are performed with a test–wise $\alpha = .05$, then considering the experiment-wise false positive rate, the number of false positive tests is binomial with k trials (i.e., one trial per test) and there is a success probability of .05 for each test. Thus, the experiment-wise Type 1 error rate is $1 - .95^k$. More generally, it is $1 - [(1 - \alpha_1) * (1 - \alpha_2) * ... * (1 - \alpha_k)]$.

One application of this idea appears when the investigator in a randomized trial wants to take multiple looks at the data (e.g., in order to stop the study early in the presence of strong evidence about efficacy or futility), maintain a frequentist perspective, and keep an overall experiment-wise error rate of α. For example, the first of two tests might use $\alpha = .01$ with the second test using $\alpha = .04$. Often, the first looks at the data use very small values of α, and thus the investigator will only stop the trial early if there is very strong evidence in favor of the intervention. This application is discussed under the topic of sequential and group sequential hypothesis testing in the clinical trials literature.

Another version of the multiple testing problem is encountered in randomized trails with more than 2 intervention groups. For example, if A, B, and C denote two versions of the intervention and the control, respectively, some of the comparisons of interest might include:

- A vs B
- A vs C
- B vs C
- Average of A and B versus C

The BIOSTAT 705 course will cover some of the relevant adjustment methods (e.g., Tukey, Bonferonni, Scheffe) that can be used. The tests in question might not be independent, which is a complicating factor.

A simple tool that the practicing biostatistician can use to assess the multiple comparison problem is to leverage the knowledge that under the null hypothesis, p–values have a U(0,1) distribution. Accordingly, a simple way to assess whether an ensemble of p–values is consistent with an experiment-wide null hypothesis is to create a Q–Q plot, which plots the expected quantiles of the U(0,1) distribution against the observed quantiles of p–values. Departures from the 45–degree line suggest the presence of actual relationships, as well as the number of such relationships.

3. Practice Problems

Problem I

Suppose we are testing 100000 genes.

i. If we use $\alpha = .05$ per test, how many false positive results are expected?

ii. What should the test-wise Type 1 error rate be set to in order to have $\alpha = .05$ for the entire experiment?

Problem II

i. Simulate p–values from 100 independent tests for which the null hypothesis of no relationship holds

ii. Verify that the Q–Q plot performs as expected

Problem III

i. Take the smallest p–values from Problem II and divide them by 5

ii. Describe the change in the Q–Q plot

iii. Perform the same procedure for the smallest 10 p–values (i.e., divide the smallest 10 p–values by 5)

iv. Describe the change in the Q–Q plot

Module 21: Non-Parametric Statistics

1. Background Discussion

Entire courses can be devoted to the topic of non–parametric tests. Our goal here is not to be comprehensive, but instead to illustrate the thinking behind such tests. A key idea is that the distributional assumptions that we have been making to date can be weakened. This implies that the resulting tests are more generalizable, but at the trade–off of being less powerful when the distributional assumptions in question actually apply. We will assume that we are analyzing a paired data set where the analysis variable is a difference score. For example, the first 3 observations might be as follows:

Observational Unit	Outcome for Subunit A	Outcome for Subunit B	Y=difference score
1	14	17	-3
2	32	30	2
3	20	21	-1

The observational unit might be an individual person, with data taken at 2 points in time. Alernatively, the observational unit might be a pair of individuals with similar characteristics, one of which receives intervention A and the other receives intervention B. The ultimate outcome variable is the difference score Y, and a key inferential question is whether values of Y consistently differ from 0.

If the distribution of Y is approximately normal, then the natural statistical test is a 1–sample paired t–test. Here, "paired" emphasizes that the analysis variable is derived from a difference score and "1–sample" emphasizes that we are interested in hypotheses, like $H_0 : \mu_Y = 0$ — in other words, that the hypotheses pertain to the population as a whole and do not involve any predictor variables (other than the one used to create the difference score). For simplicity of exposition, we assume that Y was derived from the difference in outcomes from receiving drugs A and B.

The following illustrates a data set with 6 experimental units, as well as some transformations of Y. The normality assumption underlying the t–test does not apply, $Y = 99$ is an outlier that would have a disproportionate influence on a t–test, but not necessarily on a non–parametric test.

Observational Unit	Y	$\|Y\|$	Rank($\|Y\|$)	Signed Rank	$I\{Y > 0\}$
1	-3.2	3.2	2	-2	0
2	-4.5	4.5	3	-3	0

Observational Unit	Y	$\|Y\|$	Rank($\|Y\|$)	Signed Rank	$I\{Y > 0\}$
3	-9.3	9.3	5	-5	0
4	99.0	99.0	6	6	1
5	-4.7	4.7	4	-4	0
6	-0.2	0.2	1	-1	0

To form a non–parametric null hypothesis, we start by considering some of the implications of drugs A and B having identical efficacy. Assuming for simplicity that Y is continuous and so $Pr[Y = 0] = 0$, one implication is that positive and negative values of Y are equally likely. In other words, under the null hypothesis, $I\{Y > 0\}$ has a binomial distribution with $N = 6$ and $\theta = 0.50$. We observed exactly 1 success for this binomial RV; for example, a 1–sided p–value would be derived from $Pr[Y = 0|\theta = .50] + Pr[Y = 1|\theta = .50]$. More generally, this procedure is called the "sign test."

The sign test considers direction rather than magnitude – in other words, it asks whether $Y > 0$. It is not concerned with how much $Y > 0$. This feature reduces its power, but increases its generalizability. Sometimes, the sign test can be helpful in proving the obvious in situations where modeling is difficult. For example, suppose that study design is a clustered randomized trial, where each observational unit represents a pair of medical practices, one is the intervention and the other is the usual care. Suppose that the intervention outperformed usual care in all 10 pairs, but that the pairs were so different from one another that it is tedious to include them in a general model. The p–value under the null hypothesis is $.5^{10}$, or $2 * .5^{10}$, depending on whether the alternative hypothesis is 1– or 2–sided. In either case, the p–value is sufficiently small as to demonstrate that the intervention is superior to usual care. In this illustration, the sign test is termed "exact," rather than approximate, because it is based on an enumeration of all possible outcomes, and a precise calculation of the probabilities of each possible outcome under the null hypothesis.

A more powerful non–parametric test recognizes that the ranks of Y sum to 21, and also that, under the null hypothesis, the ranks associated with positive values of Y ought to be similar to the ranks associated with the negative values of Y. The ranks in question pertain to the absolute values of Y. Here, the signed ranks associated with positive values of Y sum to 6, whereas the signed ranks associated with negative values of Y sum to 15. The impact of the outlier is modest, because of the transformation into ranks. The likelihood function, and thus the p–value, is derived from recognizing that, under the null hypothesis, each of the 6 ranks contributing to the score is effectively a single Bernoulli trial with success probability 0.5.

In practice, it is best to consider non–parametric tests as follows:

1. A simple approach to exploring the robustness of the conclusions from the usual parametric analysis
2. As one of the ways to reduce the impact of outliers

2. Practice Problems

The examples given above imply an important nuance about non–parametric tests. When using non–parametric tests, the process of translating scientific hypotheses into statistical language results in different null and alternative hypotheses for the non–parametric (e.g., sign test and signed rank test) and corresponding parametric tests (e.g., paired t–test). The following exercise is intended to help explore this nuance.

Problem I

Refer to the data set in the example above with 6 observations of Y. Imagine the observed values of Y represent the difference in a continuous outcome for each patient under a treatment and control condition. The treatment increases the outcome relative to the control, such that positive differences for an individual participant represent a beneficial effect from the treatment. Our task is to analyze the data three different ways:

1. Using a paired t–test
2. Using the sign test
3. Using the signed rank test

i. Before performing any of the analyses, write down the scientific hypothesis that is being tested

ii. For each of the analyses, translate the scientific hypothesis into statistical hypotheses (i.e., null and alternative hypotheses)

iii. Conduct the analyses and report the results

iv. The description of the results for each analysis should include an interpretation of the statistical and scientific hypotheses

Module 22: Non-Parametric Two-Sample T-Tests

1. Video

Click here to watch the video.

In the previous module we considered some non–parametric alternatives to a 1–sample t–test. Let us do the same for the 2–sample t–test. We will illustrate the following:

1. Applying a t–test to a rank–transformed version of the outcome variable
2. A permutation test

2. Using Parametric Tests on Ranked Data

The notion of transforming the outcome variable often appears within the context of generating something which satisfies the t–test's assumption of normality. Transformation is useful in other cases as well. For example, when the impact of an intervention is multiplicative rather than additive, performing a log–transformation places the outcome on a more natural scale relative to the underlying science. Rank–transforming the data has a different purpose entirely: namely, as a simple approach to applying a non–parametric test. More precisely, in the two–sample case, the procedure is to transform the combined (from both samples) data into ranks, separate the ranked data back into their respective groups, and then apply the usual parametric two–sample t–test.

Why do this? The short answer is that this approach is a simple way to construct a non–parametric test. In fact, the result is nearly identical to that of the Wilcoxon rank sum test. However, there is no expectation that the ranks of the data will have a normal distribution. There is also no expectation that the mean of the ranks will have a scientific interpretation.

Let us look at a small example. Consider the following realizations of two continuous random variables.

X	Y
-0.1	6.4
0.5	3.9
4.1	0.5
1.1	1.6
1.3	2.1

The following R code illustrates the process of ranking the data:

```
x <- c(-0.1,0.5,4.1,1.1,1.3)

y <- c(6.4,3.9,0.5,1.6,2.1)

rank_all <- rank(c(x,y))

rank_x <- rank_all[1:length(x)]

rank_y <- rank_all[-(1:length(x))]
```

This code produces the following ranking:

X	Rank of X	Y	Rank of Y
-0.1	1	6.4	10
0.5	2.5	3.9	8
4.1	9	0.5	2.5
1.1	4	1.6	6
1.3	5	2.1	7

The rank of 2.5 is assigned to the tied values of $X = 0.5$ and $Y = 0.5$. If these data points had been different from each other (e.g., $X = 0.5$ and $Y = 0.6$), then they would occupy ranks 2 and 3. However, they are equal, so we average the ranks 2 and 3 to arrive at a common rank of 2.5 for each of the tied values of 0.5.

Performing a two–sample t–test on the ranks (assuming equal variances) gives a p–value of 0.23. The Wilcoxon rank sum test gives a p–value of 0.25. These are very close (we will explore the correspondence between these tests in more depth in the practice problems).

```
t.test(rank_x, rank_y)

wilcox.test(x, y)
```

3. Permutation Tests

An alternative non–parametric test is the permutation test. The following are key insights to performing a permutation test:

1. The input data to the statistical test contain the outcome variable Y and the group label X (i.e., X might be 0 or 1, denoting whether a patient received drug A or B)
2. If the null hypothesis holds, the group labels do not matter (i.e., they do not predict the outcome)
3. Accordingly, we can estimate the distribution of the test statistic under H_0 by keeping values of the outcome variable as is and randomly assigning the labels

The table below illustrates the idea for the observed data ("iteration 0") and 2 replications of a study with 4 patients per group. The randomization is accomplished by randomly assigning each individual the value of a uniform RV and then assigning the top 4 values of that RV to group 0 and assigning the rest to group 1.

Actual data ("iteration 0"):

ID	Group	Y
1	0	50
2	0	64
3	0	58
4	0	71
5	1	39
6	1	63
7	1	48
8	1	59

The observed value of the test statistic is based upon the difference between the observed group means:

- For group 0: the mean of 50, 64, 58, 71
- For group 1: the mean of 39, 63, 48, 59

Iteration 1

ID	Y	Random Number	New Group
1	50	0.41	1
2	64	0.20	0
3	58	0.99	1
4	71	0.81	1
5	39	0.78	1

ID	Y	Random Number	New Group
6	63	0.03	0
7	48	0.19	0
8	59	0.36	0

To estimate the sampling distribution under the null hypothesis, the contribution from iteration 1 uses the difference between the new group means:

- For group 0: the mean of 64, 63, 48, 59
- For group 1: the mean of 50, 58, 71, 39

Iteration 2

ID	Y	Random Number	New Group
1	50	0.03	0
2	64	0.07	0
3	58	0.98	1
4	71	0.78	1
5	39	0.48	1
6	63	0.35	0
7	48	0.91	1
8	59	0.23	0

The contribution to the estimated sampling distribution from iteration 2 uses the difference between the new group means:

- For group 0: the mean of 50, 64, 63, 59
- For group 1: the mean of 58, 71, 39, 48

It is easy to imagine that repeating this process (e.g., 5000 times) would produce a nice distribution of differences in means that is consistent with a true null hypothesis (i.e., the assigned group does not influence the outcome). The observed mean difference (from iteration 0) can be compared to the estimated sampling distribution to compute a p–value. This is a basic permutation test.

4. Practice Problems

Problem I

Write a simulation that examines the agreement between the p–value from the Wilcoxon rank sum test and the p–value from the two–sample t–test performed on ranked data. A general algorithm for the simulation is as follows:

Repeat K times :

Generate N observations from $X \sim N(1, 2^2)$

Generate N observations from $Y \sim N(1, 2^2)$

Rank the data

Run the t–test on the ranks and capture the p–value

Run the Wilcoxon test and capture the p–value

The result will be a K x 2 matrix where the rows represent runs 1, ..., K and the two columns represent the p–values for the t–test on the ranked data and the Wilcoxon test, respectively.

i. Scatter plot the p–values from the K Wilcoxon tests and the K t–tests on the ranks to visualize the results of the simulation.

ii. Use $K = 5000$ and $N = 5$. Then, increase to $N = 10$ through 100 in increments of 10 (i.e., $N = 10, N = 20, N = 30$, etc.).

iii. Describe the results of these simulations. Do the results seem to agree or disagree with one another? What impact does the sample size have on the results, if any?

Problem II

Repeat the simulation above, but make random draws from $X \sim U(0,1)$ and $Y \sim U(0.5, 1.5)$. Does anything change in comparison to the results from the simulation that generated data from normal distributions?

Problem III

We received a data set from a randomized trial of Drug A versus Placebo. The outcome measure is continuous. Write an R program to do a permutation test of the null hypothesis that the mean outcome on Drug A is equivalent to the mean outcome on Placebo.

$$A = 9.02, 5.38, 12.01, 8.58, 8.62, 12.05, 9.43, 7.56, 10.36, 9.72$$

$$B = 30.01, 30.77, 29.26, 31.29, 29.56, 30.66, 32.19, 30.87, 29.35, 32.30$$

Module 23: What is Not Covered

Discussion

This is a brief description of some of the topics that are covered in a typical second inference course. We are not intending to provide actionable information here, but instead a big picture view of how everything fits together.

Among others, a second inference course considers criteria for comparing estimators and, ideally, for identifying and creating optimal estimators. Much, but not all, of this content is limited to unbiased estimators, although the principle of the bias-variance trade–off reminds us that a modestly biased estimator with a small variance might be preferable to an unbiased estimator with a much larger variance. Indeed, some of this content pertains to different approaches for making the trade–off between bias and precision.

Within the class of unbiased estimators, a second inference course considers how to develop lower bounds on the variance (an estimator that achieves this lower bound can be considered to be optimal), and also describes rules for creating such estimators which apply to some, but not all, situations. It turns out that one circumstance where optimal estimators can be defined occurs when the distribution in question is part of the exponential family, a characteristic shared by all the distributions we have covered in this course. So, a second course in inference will show how to create optimal estimators in a large number of practically relevant situations.

The other main theme in a second inference course pertains to the distinction between small sample problems and large sample ones (i.e., the difference between exact and asymptotic inference). Exact inference is based on sampling distributions from the underlying variable, which often is not normal. Asymptotic inference is based on the central limit theorem and its supporting architecture. In essence, asymptotic inference makes more precise the notion of what it means for the behavior of an estimator to converge to something as the sample size increases. The ultimate destination is the recognition that, under certain regularity conditions, the maximum likelihood estimate (MLE) is unique, consistent, asymptotically normal with mean θ and asymptotic variance $1/n * I(\theta)$ and asymptotically efficient. In other words, and with various qualifications, the MLE is the best we can do.