



STT 231

STATISTICAL METHODS

Chapter 2: Introduction to
Inference

Lecture 2-1: Statistical models & parameters

Consider each of the following uses of the term *model*:

1. The law became a **model** for dozens of laws banning nondegradable plastic products.
2. The research method will be **modeled** on previous work.

In statistics, we often create and apply

models to represent the generative process that creates our sample data.

Many statistical models involve **parameters**, values that govern a statistical model.

Statistical inference can be defined as the use of sample data to evaluate and/or estimate these parameters and the models they govern.

Definitions

- A statistical model is a set of assumptions (often mathematical) concerning the process that generates data and the relationship between one or more random variables.
- A parameter is a number that describes some aspect of a **statistical model**.
- A statistic is a number that describes some aspect of a **sample**.
- In the context of statistical inference, a **statistic** is used as an estimate/approximation of a parameter.

Notation

Notation for parameters and statistics			
Variable type	Quantitative Summary	Parameter (summarizing a generative process)	Statistic (summarizing observed sample)
Quantitative	A quantitative mean	μ "mu"	\bar{x} "x-bar"
Quantitative	A difference in quantitative means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Quantitative	A standard deviation	σ "sigma"	s
Categorical	A proportion / rate of a categorical variable	p	\hat{p}
Categorical	A difference in proportions / rates	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$

Distinguishing between parameters & statistics

In each of the following cases, identify the parameter(s) of interest and their corresponding sample statistics.

1. Minori Mori, a 12th grader at Meikei High School in Tsukuba, Japan, set out to investigate whether auxin and phosphates increased the chance of a growing a four-leaf clover.

In a controlled study, 4 of 372 clovers in the control group had more than three leaves, whereas 31 of 444 clovers in the treatment group receiving auxin and phosphate fertilizers had more than three leaves.

Parameter:

$$P_1 - P_2$$

Sample statistic:

$$\hat{p}_1 - \hat{p}_2 = \frac{4}{372} - \frac{31}{444} = -0.0591$$



Distinguishing between parameters & statistics

2. Researchers believe that the rate at which children experience severe reactions to diphtheria, tetanus, and pertussis vaccines (booster shots) is the same regardless of whether the vaccines are administered at the thigh or the upper arm. A randomized experiment showed that children receiving vaccines at the upper arm had a 6% higher reaction rate.

Parameter :

$$p_1 - p_2$$

Statistic

$$\hat{p}_1 - \hat{p}_2 = 0.06$$



Distinguishing between parameters & statistics

3. Last year, data emerged showing that fireworks bring choking pollution to many people during the Indian festival of Diwali, an annual four-day Hindu religious celebration.

Last December, a meteorologist presented data on particulate matter (or PM) from Diwali fireworks in his city. Scientists measure such pollution in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$) of air.

In a 12-hour period on the holiday, PM values soared to $500.5 \mu\text{g}/\text{m}^3$. That rise was about 21 to 27 percent higher than before the fireworks went off.

Parameter:

μ

Statistic:

$\bar{x} = 500.5 \mu\text{g}/\text{m}^3$



Distinguishing between parameters & statistics

4. After computing sample statistics, a researcher claims that the standard deviation of shell measurement (the length of the anterior adductor muscle scar, standardized by dividing by length) for a particular species of mussel is roughly 0.08 mm.

Parameter:

σ

Statistic:

$S = 0.08 \text{ mm}$



Independence Models

- An important situation in statistics occurs when the two variables turn out to be independent.

Definition:

Two variables A, B are said to be independent if ...

the likelihood of variable A taking on a particular value is not influenced by (i.e., independent of) the value taken on by variable B.

Example 2.1: Dolphin Therapy

- Is swimming with dolphins therapeutic for patients suffering from clinical depression?
- Researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment.
- These 30 subjects went to Honduras, where they were randomly assigned to one of two treatment groups.
 - Both groups engaged in the same amount of swimming and snorkeling each day (the outdoor nature program), but one group did so in the presence of bottlenose dolphins and the control group did not.
 - Each subjects' level of depression was evaluated at the beginning of the study and then again at the end. Across both groups, **13** of the **30** patients showed substantial improvement in depressive symptoms.

Creating an independence model

- a. An independence model would assume

treatment had absolutely no relationship with outcome. Complete the table below to show the study results you would expect to see under such a model.

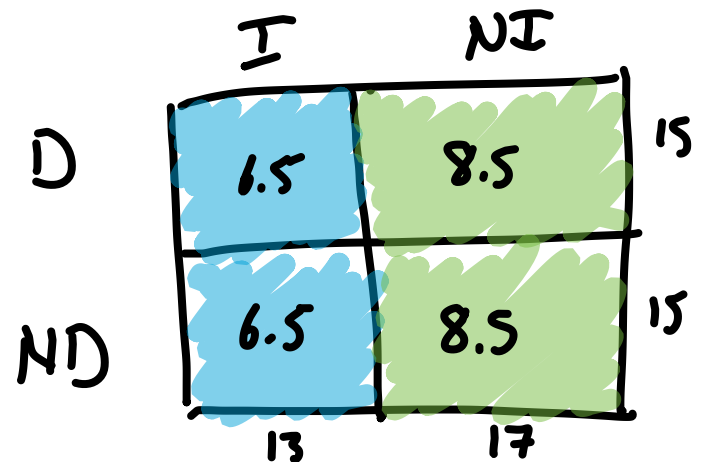
Treatment	Substantial improvement	No substantial improvement	Total
Dolphins	6.5	8.5	15
No Dolphins	6.5	8.5	15
Total	13	17	30

Creating an independence model

Treatment	Substantial improvement	No substantial improvement	Total
Dolphins	6.5	8.5	15
No dolphins	6.5	8.5	15
Total	13	17	30

b. Sketch a mosaic plot to visualize the results expected by our independence model. What is the expected difference in improvement rates across groups? In other words, what is the *parameter* that summarizes this model?

$$P_D - P_{ND} = \frac{6.5}{15} - \frac{6.5}{15} = 0$$



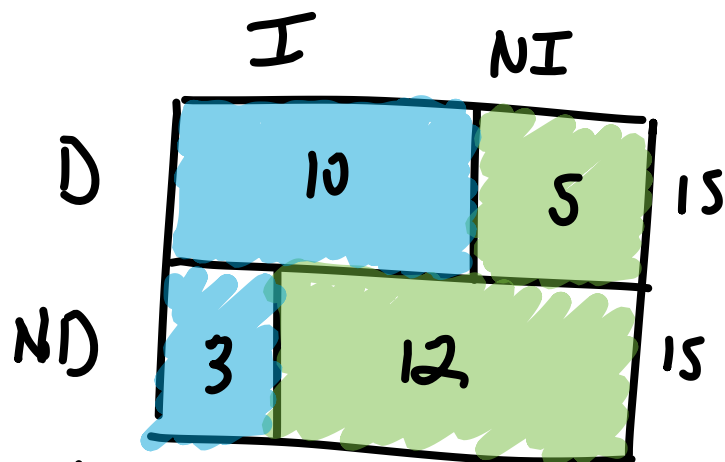
Comparing observed results to model expectations

Treatment	Substantial improvement	No substantial improvement	Total
Dolphins	10	5	15
No dolphins	3	12	15
Total	13	17	30

d. Sketch a mosaic plot to visualize the actual results observed by (Antonioli & Reveley, 2005). To what extent do you think the model captures the observed relationship between treatment type and patient outcome?

$$\hat{p}_1 - \hat{p}_2 = \frac{10}{15} - \frac{3}{15} = \underline{\underline{0.4667}}$$

The model's prediction is not at all compatible with our observed value of $\hat{p}_1 - \hat{p}_2$.



It doesn't do a good job describing the relationship between Treatment & Outcomes.



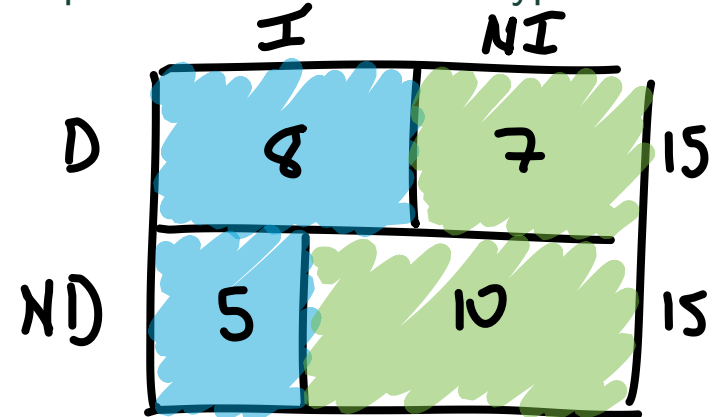
Comparing observed results to model expectations

Treatment	Substantial improvement	No substantial improvement	Total
Dolphins	8	7	15
No dolphins	5	10	15
Total	13	17	30

e. Suppose, instead, that the results of the study were as seen below. Sketch a mosaic plot to visualize these hypothetical results. To what extent would you think the model captures the observed relationship between treatment type and patient outcome?

$$\text{Here, } \hat{p}_1 - \hat{p}_2 = \frac{8}{15} - \frac{5}{15} = 0.2.$$

Not quite what the model of independence predicted, but certainly much closer.



Example 2.2: Medical testing

Treatment	Substantial improvement	No substantial improvement	Total
Dolphins	8	7	15
No dolphins	5	10	15
Total	13	17	30

e. Suppose, instead, that the results of the study were as seen below. Sketch a mosaic plot to visualize these hypothetical results. To what extent would you think the model captures the observed relationship between treatment type and patient outcome?



HIV testing

- It is estimated that 0.5% of adults in the United States are HIV positive. HIV screening is commonly performed as part of routine prenatal care, as the virus is transmissible from mother to child during birth.
- A rapid screening test will accurately diagnose HIV positive subjects 75% of the time.
- This same test will also misidentify 4% of healthy people as HIV positive.

What do you think the chances are?

- If a pregnant woman receives a diagnosis of HIV from this testing procedure, what is the probability she is actually HIV positive? Without doing any calculations, give an estimate for the probability (as a percentage.)
 - A. Less than 10%
 - B. Between 10% and 25%
 - C. Between 25 and 50%
 - D. Between 50% and 75%
 - E. More than 75%



Analyzing the situation (United States)

- To help us analyze this situation, let's make a table of possibilities using a theoretical group of 10,000 patients:

HIV Status	Positive Test Result	Negative Test Result	Total
Patient is actually HIV positive	0.75(50) 37.5	0.25(50) 12.5	50
Patient is not HIV positive	0.04(9950) 398	0.96(9950) 9552	9950
Total	435.5	9564.5	10000

Results (in the United States)

- a. Of the patients who receive a positive test result, what proportion are actually HIV positive?

$$\frac{37.5}{435.5} = 0.086$$

- b. Of the patients who receive a positive test result, what proportion are actually **not** HIV positive?

$$\frac{398}{435.5} = 0.914$$

- c. Of the patients who receive a negative test result, what proportion are actually HIV positive?

$$\frac{12.5}{9564.5} = 0.0013$$

- d. Do you think this is a useful diagnostic test?

Definitely not. Although it very rarely falsely diagnoses patients with HIV as healthy, nearly everyone it diagnoses as having the disease (91.4%!!) is also healthy.

Base rate

- In addition to the accuracy of the diagnostic test, the prevalence of the disease is also an important element to consider when interpreting results. As a contrast to the example above, consider if the diagnostic test is used in Swaziland, where an estimated 26% of adults are HIV positive.
- Let's repeat our calculations under this scenario.
Remember:
 - A particular rapid screening test will accurately diagnose HIV positive subjects 75% of the time.
 - This same test will also misidentify 4% of healthy people as HIV positive.

Analyzing the situation (Swaziland)

- Again, let's make a table of possibilities using a theoretical group of 10,000 patients in Swaziland:

HIV Status	Positive Test Result	Negative Test Result	Total
Patient is actually HIV positive	1950	650	2600
Patient is not HIV positive	296	7104	7400
Total	2246	7754	10000

Results (in Swaziland)

e. Of the patients who receive a positive test result in Swaziland, what proportion are actually HIV positive?

$$\frac{1950}{2246} = 0.868$$

Do you think this is a useful diagnostic test?

Obviously, a positive test result should be verified by a second diagnosis (because it only has an 86.8% rate of being correct), but this seems much more useful:



Example 2.3: The models of an oil mogul

- Petroleum engineering is a field concerned with activities related to the production of hydrocarbons, specifically crude oil or natural gas, with a focus on maximizing economic recovery of hydrocarbons from subsurface reservoirs.
- When exploring an area for oil, oil company surveyors look for leads, formations on the survey that suggest the possibility of oil deeper below. Suppose that 6% of leads actually have oil present.
- The tests are not fool-proof, however; suppose you are tasked with deciding which test is better for a particular oil company.

Example 2.3: The models of an oil mogul

■ The facts:

1. The average cost to drill and operate a well is \$20 million USD and the average yield over a well's lifespan is 105 million USD.
2. **Test A** is a symmetric test, and is 95% accurate for both the absence and presence of oil. (That is, if oil is present, the probability the test will conclude oil is present is 0.95. If oil is absent, the probability the test will conclude oil is absent is 0.95.)
3. **Test B** is an a-symmetric test, and is 99% accurate for the presence of oil but only 90% accurate for the absence of oil. (That is, if oil is present, the probability the test will conclude oil is present is 0.99. If oil is absent, the probability the test will conclude oil is absent is 0.90.)

Example 2.3: The models of an oil mogul

a. Model the results you might expect if the company implemented Test A.

Oil Present	Positive Test	Negative Test	Total
Yes	570	30	600
No	470	8930	9400
Total	1040	8960	10,000

Example 2.3: The models of an oil mogul

b. Model the results you might expect if the company implemented Test B.

Oil Present	Positive Test	Negative Test	Total
Yes	594	6	600
No	940	8460	9400
Total	1534	8466	10,000

Example 2.3: The models of an oil mogul

c. Recall the average cost to drill and operate a well is \$20 million USD and the average yield over a well's lifespan is 105 million USD.

Which test would you recommend the company implement in the field? Provide a 1-2 sentence justification of your choice.

TLDR: Given costs & revenues associated with false/true positives/negatives, Test A is preferable.

Models suggest it would lead to an additional 15.73 billion USD in profit.



Lecture 2-2: Sampling distributions & the CLT

Recall in Chapter 1-1, we sampled responses to two questions regarding MSU students' *handedness*:

1. Do you identify as left- or right-handed?
2. What is your handedness score (as represented by an average of responses to various questions)?
 - a. How many cases are there in the population of responses from that day?
 - b. What proportion of this population identified as right-handed?
 - c. What was the average handedness score of this population?

Example 2.4: A sampling distribution for \hat{p}

- a. Quickly sketch a bar chart of the population of responses for question (1) above.

Suppose we wanted to estimate the proportion of cases that answered “Right-handed” but only sampled $n = 30$ students, instead of conducting a census of the entire population.

We usually think of a parameter as a Static Value while the sample statistic Varies from sample to sample, depending on which cases were selected at random to be in the sample.

Example 2.4: A sampling distribution for \hat{p}

b. We can use technology to simulate this process, drawing with-replacement samples of size 30 and recording the corresponding sample proportion \hat{p} . Sketch the histogram of 5,000 of these repetitions below.

Example 2.5: A sampling distribution for \bar{x}

c. Quickly sketch a bar chart of the population of responses for question (2) above. Suppose we wanted to estimate the mean “Handedness” score but only sampled $n = 30$ students, instead of conducting a census of the entire population.

Example 2.5: A sampling distribution for \bar{x}

d. We can use technology to simulate this process, drawing with-replacement samples of size 30 and recording the corresponding sample mean \bar{x} . Sketch the histogram of 5,000 of these repetitions below.

NOTE: the histogram here describes a *hypothetical* sampling distribution, and not data we actually observed. It is a *model* of the types of statistics we could expect to see if we repeated a sampling process many times over and over that describes the typical values of such statistics and how much they vary.

Sampling distributions

Definitions

A sampling distribution is a statistical model that describes the behavior of sample statistics computed for different samples of the same size from the population / generative process.

The **standard error** of a statistic, denoted SE, is the approximate standard deviation of the sample statistic.

We interpret the standard error as the approx. avg. distance between a statistics and the parameter they estimate.

Example 2.6: BRFSS

- The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States collected by the Centers for Disease Control and Prevention (CDC).
- It is designed to identify risk factors in the adult population and report emerging health trends.
- Respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage.



Example 2.6: BRFSS

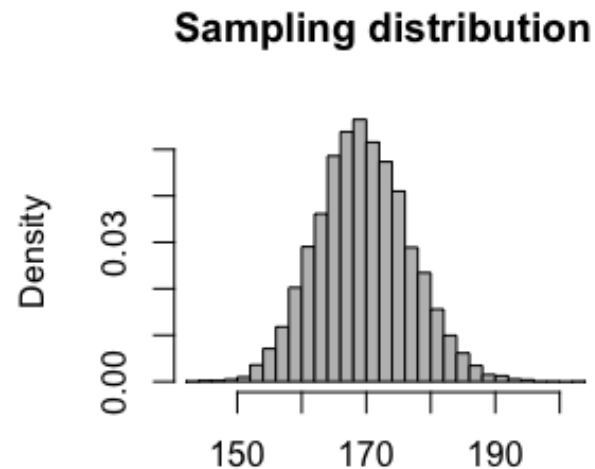
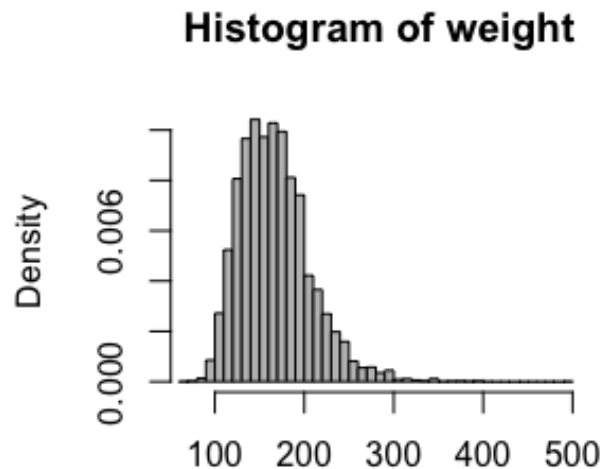
a. Classify each variable by its type and determine what parameter would be appropriate to use as a summary of its distribution.

Name	Variable type	Parameter
Genhlth	<i>nominal</i>	p
Hlthplan	<i>nominal</i>	p
Height	<i>continuous</i>	μ
Weight	<i>continuous</i>	μ
Wtdesire	<i>continuous</i>	μ
Age	<i>continuous</i>	μ

Relating populations to sampling distributions

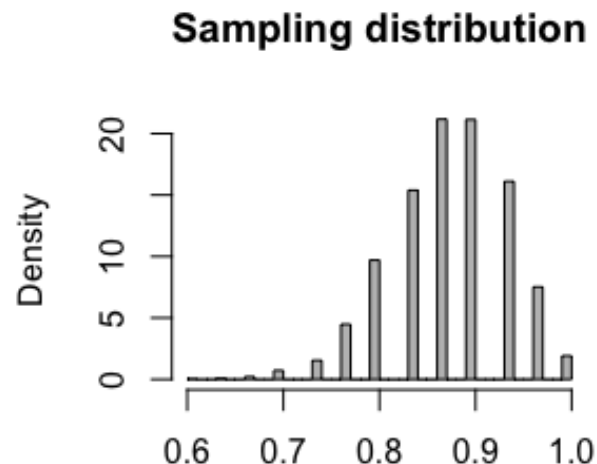
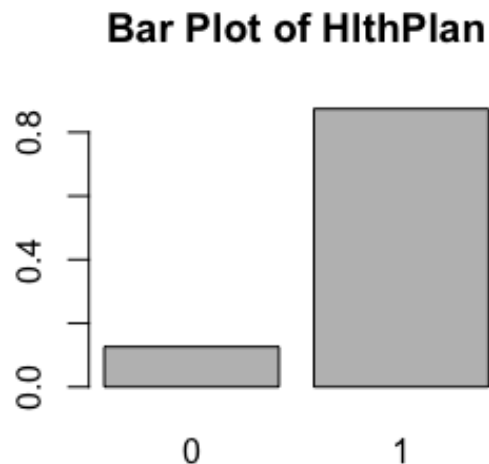
b. Consider the variable **Weight**. A histogram of the population is presented below. What is the approximate mean weight μ of this population?

To the right is the sampling distribution of \bar{x} for repeated samples of size $n = 30$. What is the approximate average of this distribution?



Relating populations to sampling distributions

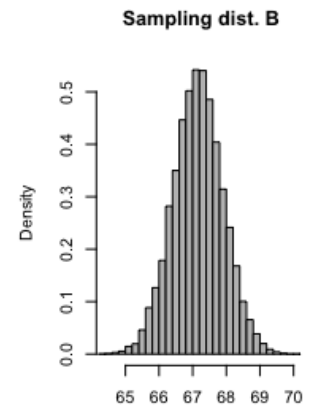
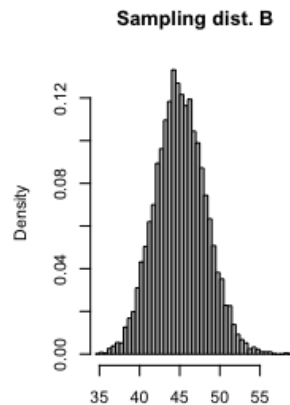
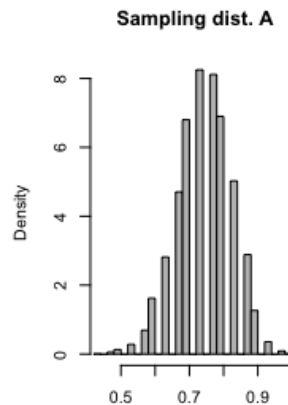
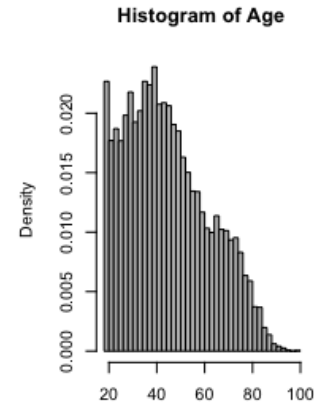
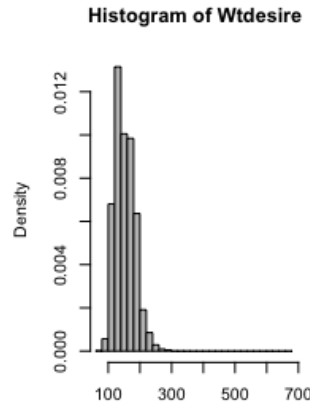
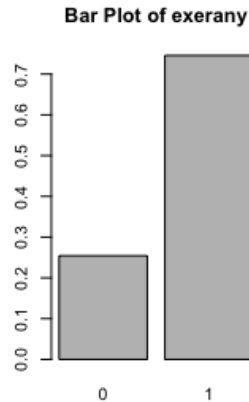
c. Consider the variable **Hlthplan**. A histogram of the population is presented below. What is the approximate proportion of individuals with a health plan p of this population? (**Note 1 = has healthplan.**) To the right is the sampling distribution of \hat{p} for repeated samples of size $n = 30$. What is the approximate average of this distribution?



Relating populations to sampling distributions

d. Below are the population-level data of **Exerany**, **Weight**, and **Age**. Match each sampling distribution to its corresponding population data.

A = Exerany
B = Age
C = Wt desire



Additional notes!

What are the key take-aways of Lecture 2-2?

Lecture 2-3: Introducing hypothesis tests

In the following lectures, we explore how to use *hypothesis tests* to answer questions such as:

1. Is the flu vaccine still effective when delivered at a 50% dosage?
2. Are antimicrobial ingredients having the opposite effect?
3. Did Kristen Gilbert kill her patients?
4. Does swimming with dolphins reduce symptoms associated with clinical depression?



HT 1: Flu-vaccine

In 2004, the USA experienced a shortage of flu vaccine when the supplies from a major pharmaceutical manufacturer was found to be contaminated.

- A study was done to see whether smaller dosages could be used successfully; if so, vaccine material could be divided into more flu shots. The usual amount of vaccine was injected into the muscle of half of the patients at random.
- The other half had only a small amount of vaccine injected under the skin (not into the muscle).
- Response was measured by looking at patients' production of antibodies.

HT 1: Flu-vaccine

Dose Type	Vigorous antibody response	Lack of vigorous antibody response	Total
Full dose	19	2	21
Small dose	17	4	21
Total	35	7	42

What statistic could be used to answer whether the flu vaccine is still effective when delivered at a 50% dosage?

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



Steps of a hypothesis test

A hypothesis test is used to determine whether results from a sample are convincing enough to allow use to conclude something about a population. Although hypothesis tests take many different forms, they always follow the same four steps:

1. Express research question in terms of parameter. *hypotheses about a*
2. Generate a sampling distribution under null model. *to see what is typical*
3. Compute p-value to quantify results under the null. *likelihood of observed*
4. Draw conclusion re: reasonableness of null model.

Step 1: Setting up hypotheses

Many research questions can be expressed as two competing claims that might be correct for a population. These two statements are called the **null** and the **alternative hypotheses**.

1. **Definition:** The null hypothesis is often denoted by H_0 , and is a statement that there is no effect, no difference, no change, nothing of interest to observe.

The null hypothesis is usually referred to as the

status quo. It is the claim that any differences we see in sample results compared to the status quo is due to chance, that is, to uninteresting variation or randomness in the sampling distribution that is expected by the model.

Step 1: Setting up hypotheses

Many research questions can be expressed as two competing claims that might be correct for a population. These two statements are called the **null** and the **alternative hypotheses**.

2. **Definition:** The **alternative hypothesis** is denoted by H_a ,

and is a statement that there is

an effect, a difference, something of scientific interest.

It is the claim that the difference in sample results compared to the status quo is difficult to explain as randomness expected by the model

and is NOT due to chance .

Step 1: Setting up hypotheses

Many research questions can be expressed as two competing claims that might be correct for a population. These two statements are called the **null** and the **alternative hypotheses**.

3. Key idea: In a hypothesis test, we examine whether sample data provide evidence against the null hypothesis and support the alternative hypothesis.

4. Key idea: In general, the null hypothesis H_0 is a statement of equality, while the alternative hypothesis uses notation indicating inequality, depending on the question of interest.

Example 2.7 – setting up hypotheses

In each case, state the null and alternative hypothesis for the statistical test described.

Testing to see if there is evidence that a proportion is greater than 0.3.

$$H_0: \underline{p = 0.3} \quad \text{vs.} \quad H_a: \underline{p > 0.3}$$

Example 2.7 – setting up hypotheses

In each case, state the null and alternative hypothesis for the statistical test described.

Testing to see if there is evidence that the mean of group A is different than the mean of group B.

$$H_0: \underline{\mu_A = \mu_B} \quad \text{vs.} \quad H_a: \underline{\mu_A \neq \mu_B}$$



Example 2.7 – setting up hypotheses

In each case, state the null and alternative hypothesis for the statistical test described.

Testing to see if there is evidence that the rate at which individuals given the flu vaccine at a 50% dosage display a vigorous antibody response is different than that of individuals given the vaccine at full dose.

$$H_0: \frac{p_F}{p_S} = \frac{p_F}{p_S} \quad \text{vs.} \quad H_a: \frac{p_F}{p_F} \neq \frac{p_S}{p_S}$$

NOTE: The direction of the alternative hypothesis is dictated by the motivation of the research question.



Step 2: Seeing what's typical – null distributions based on H_0

The appropriate hypotheses associated with the research question, *Is the flu vaccine effective when delivered at 50% dosage?* are...

$$H_0: \underline{p_F = p_S} \quad \text{vs.} \quad H_a: \underline{p_F \neq p_S}$$

...what was the observed difference in proportions $\hat{p}_F - \hat{p}_S$?

$$\underline{19/21 - 17/21 = 0.0952}$$

Is this a typical result or an atypical result **if, in fact**, the small dose and full dose were equally effective?

To help answer this question, we'll create a special type of sampling distribution called a

null distribution.

Definition: Null distribution

To evaluate the quality of a null model, we need to generate a *sampling distribution* for our sample statistic **using a process that generates data based on H_0** .

This distribution is what we will call a

 null distribution . It is the sampling distribution we would expect to see if H_0 adequately described the phenomenon we are studying. It will be centered at the value the null hypothesis thinks we are most likely to see and will show what values of the sample statistic are likely to occur by random chance if, in fact, the null hypothesis is correct.

Step 2: Seeing what's typical – null distributions based on H_0

a. Suppose that the doses are equally effective, that is, *Dose type* and *Response* are independent. What type of results would we expect to get?

Dose Type	Vigorous	Lack of Vigorous	Total
Full	18	3	21
Small	18	3	21
Total	35 36	6 6	42

Step 2: Seeing what's typical – null distributions based on H_0

Dose Type	Vigorous	Lack of Vigorous	Total
Full	18	3	21
Small	17	3	21
Total	36	6	42

Under the assumption that the null is true, the expected difference $\hat{p}_F - \hat{p}_S = \underline{0}$.

Thus, given that the null hypothesis is correct, any sample data we collect should give us an observed difference that varies slightly from 0.

But by how much should it vary?

Creating a null distribution

Dose Type	Vigorous	Lack of Vigorous	Total
Full	18	3	21
Small	18	3	21
Total	36	6	42

We can simulate what would happen if *DOSE* did not influence *RESPONSE*.

In the simulation, we would write full on 21 cards and small on 21 cards.

Then we would shuffle the cards and deal 36 cards into one pile to represent patients displaying a vigorous antibody response and put the remaining 6 cards in a pile to represent those who did not.

Then we tabulate the results and determine the difference in vigorous response rates, $\hat{p}_F - \hat{p}_S$.

Creating a null distribution

Dose Type	Vigorous	Lack of Vigorous	Total
Full	18	3	21
Small	18	3	21
Total	25 36	6 6	42

Simulation 1: $\hat{p}_F - \hat{p}_S = \frac{15}{21} - \frac{21}{21} = -0.2857$

Dose Type	Vigorous	Lack of Vigorous	Total
Full	15	6	21
Small	21	0	21
Total	36	6	42

Creating a null distribution

Dose Type	Vigorous	Lack of Vigorous	Total
Full	18	3	21
Small	18	3	21
Total	36	6	42

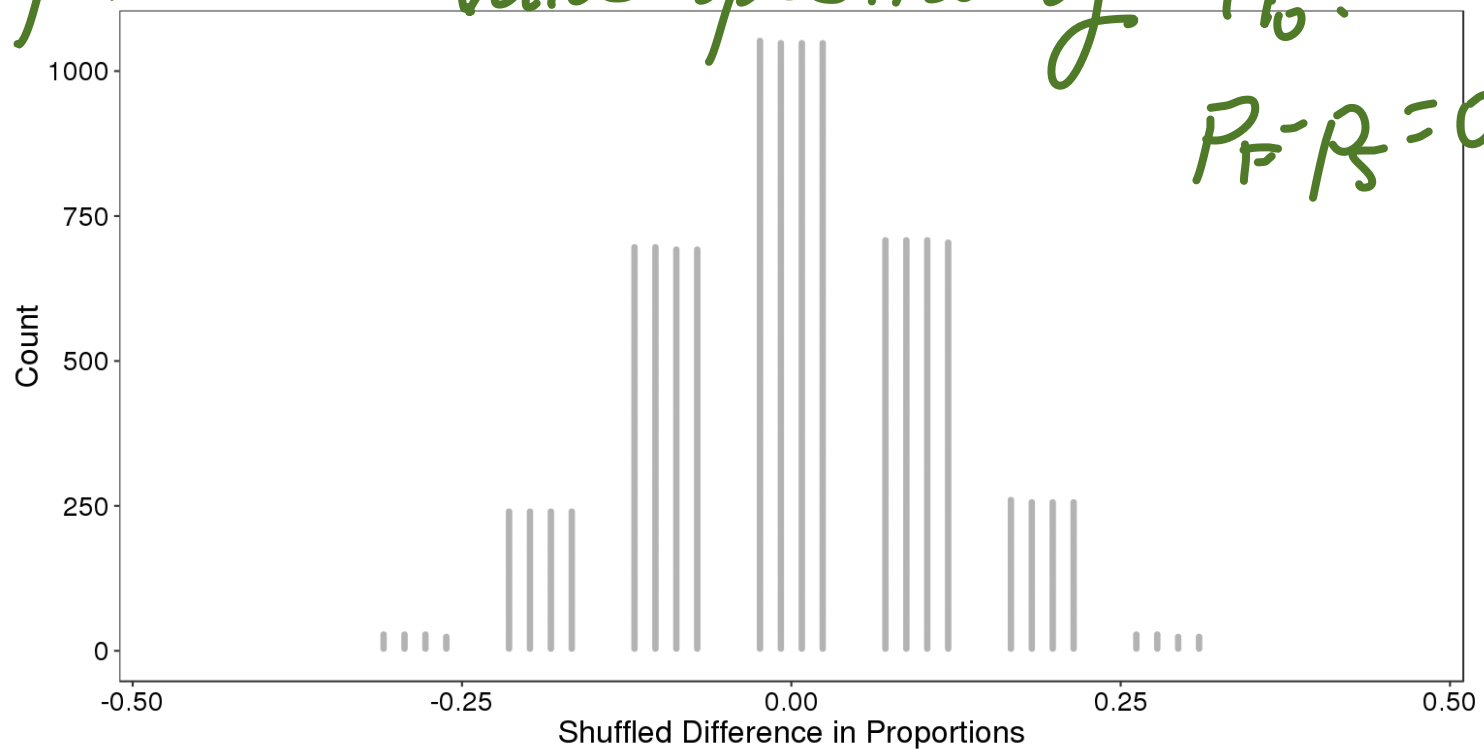
Simulation 2: $\hat{p}_F - \hat{p}_S = \underline{\underline{\frac{20}{21} - \frac{16}{21} = 0.1905}}$

Dose Type	Vigorous	Lack of Vigorous	Total
Full	20	1	21
Small	16	5	21
Total	36	6	42

Creating a null distribution

b. What is the approximate center of this distribution? Could we have expected this in advance?

Distribution is centered at 0. This is the parameter value specified by H_0 :



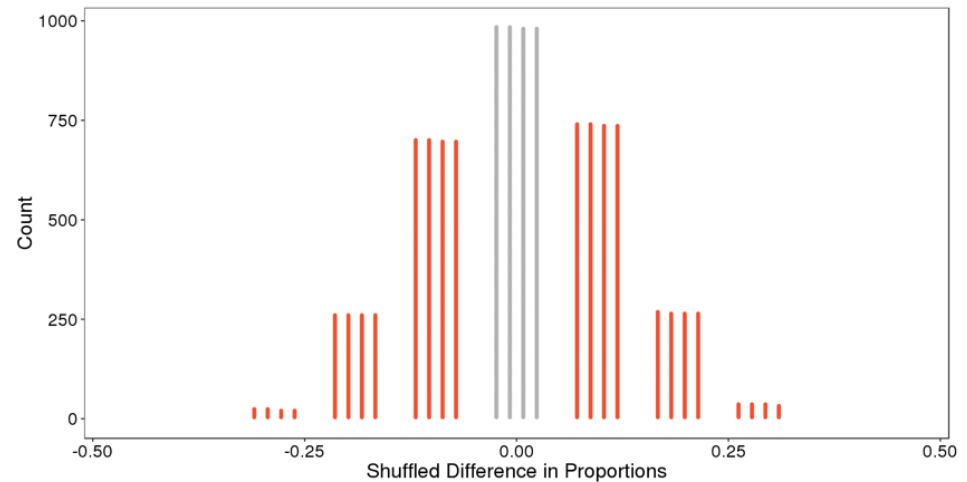
Step 3 – the p-value: evaluating evidence against null model

c. What proportion of our simulated studies resulted in a difference of vigorous response rates $|\hat{p}_F - \hat{p}_S| \geq 0.0952$?

Approx.
68% of
simulations

showed

$$|\hat{p}_F - \hat{p}_S| \geq 0.0952.$$



This value is referred to as a p-value !

Definition: p-values

The **p-value** of a test is a probability calculated using a model based on the null hypothesis being tested.

It is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, **using the null model**.

We can use p-values to help us evaluate how well the null hypothesis model explains or “fits” our observed sample results.

Definition: p-values

If the p-value is large, this indicates that our observed results look like they could be a result of the natural variation that we expect to see when we take random samples.

The smaller a p-value is, the less inclined we will be to think that our sample result is simply due to natural variation.

In other words, *small p-values give us reason to doubt that the null model is a good explanation of the observed results we have.*

If the p-value is:	Greater than 0.10 ($p > 0.10$)	Between 0.05 and 0.10 ($0.05 < p \leq 0.10$)	Between 0.01 and 0.05 ($0.01 < p \leq 0.05$)	Between 0.001 and 0.01 ($0.001 < p \leq 0.01$)	Less than 0.001 ($p \leq 0.001$)
we will say we have:	little evidence	some evidence	strong evidence	very strong evidence	extremely strong evidence

Step 4 – making a conclusion in context

Recall our two hypotheses regarding the rate of vigorous antibody responses across patients given full and small doses.

- $H_0: p_F = p_S$ The variables **Dose** and **Response** are **independent**. Whether you receive a full dose or a small dose **does not** affect whether you display a vigorous antibody response. The observed difference in sample rates is attributable to chance variation expected by our null model.
- $H_A: p_F \neq p_S$ The variables **Dose** and **Response** are **NOT independent**. Whether you receive a full dose or a small dose **does** affect whether you display a vigorous antibody response. The observed difference in sample rates isn't well-explained as chance variation expected by our null model, and warrants further study.

Step 4 – making a conclusion in context

- $H_0: p_F = p_S$
- $H_A: p_F \neq p_S$

a. Draw a conclusion regarding these hypotheses that includes mention of our p-value.

Because our p-value was 0.68, we have very little evidence against the hypothesized model that assumes antibody response rate is the same across both dosages.

Example 2.8: Are antimicrobial ingredients having the opposite effect?

Triclosan is a compound often added to products in soaps, lotions, and toothpastes. It is antimicrobial, so we expect it to lower one's chance of having a staph infection. However, the opposite was found in a recent study.

Microbiologists swabbed the noses of 100 people and recorded which had detectable levels of triclosan and which had evidence of carrying staph bacteria, which greatly increases one's chance of having a serious staph infection.

Are antimicrobial ingredients having the opposite effect?

Group	Staph	No Staph	Total
Triclosan	24	23	47
No Triclosan	15	38	53
Total	39	61	100

a. What is the observed difference in staph rates across the 'Triclosan' and 'No Triclosan' groups? Use the appropriate notation.

$$\hat{p}_{Tr} - \hat{p}_{NoTr} = \frac{24}{47} - \frac{15}{53} = 0.2276$$



Are antimicrobial ingredients having the opposite effect?

Group	Staph	No Staph	Total
Triclosan	24	23	47
No Triclosan	15	38	53
Total	39	61	100

a. What is the observed difference in staph rates across the 'Triclosan' and 'No Triclosan' groups? Use the appropriate notation.



Step 1 – Establishing hypotheses:

b. We wish to see if the effect found in the study (that Triclosan actually *increases* the risk of staph infections) generalizes to the broader population. What are the appropriate hypotheses to be tested?

$$H_0: \frac{P_{\text{Tr}}}{n} = \frac{P_{\text{NoTr}}}{n} \quad \text{vs.} \quad H_a: \frac{P_{\text{Tr}}}{n} > \frac{P_{\text{NoTr}}}{n}$$



Step 2 – Null distribution:

Assume there truly is no difference in Staph infection rates for Triclosan and No Triclosan groups. Write

Triclosan on 47 cards and
No Triclosan on 53

cards. Then we would thoroughly shuffle the cards and deal 39 cards into one pile to represent those

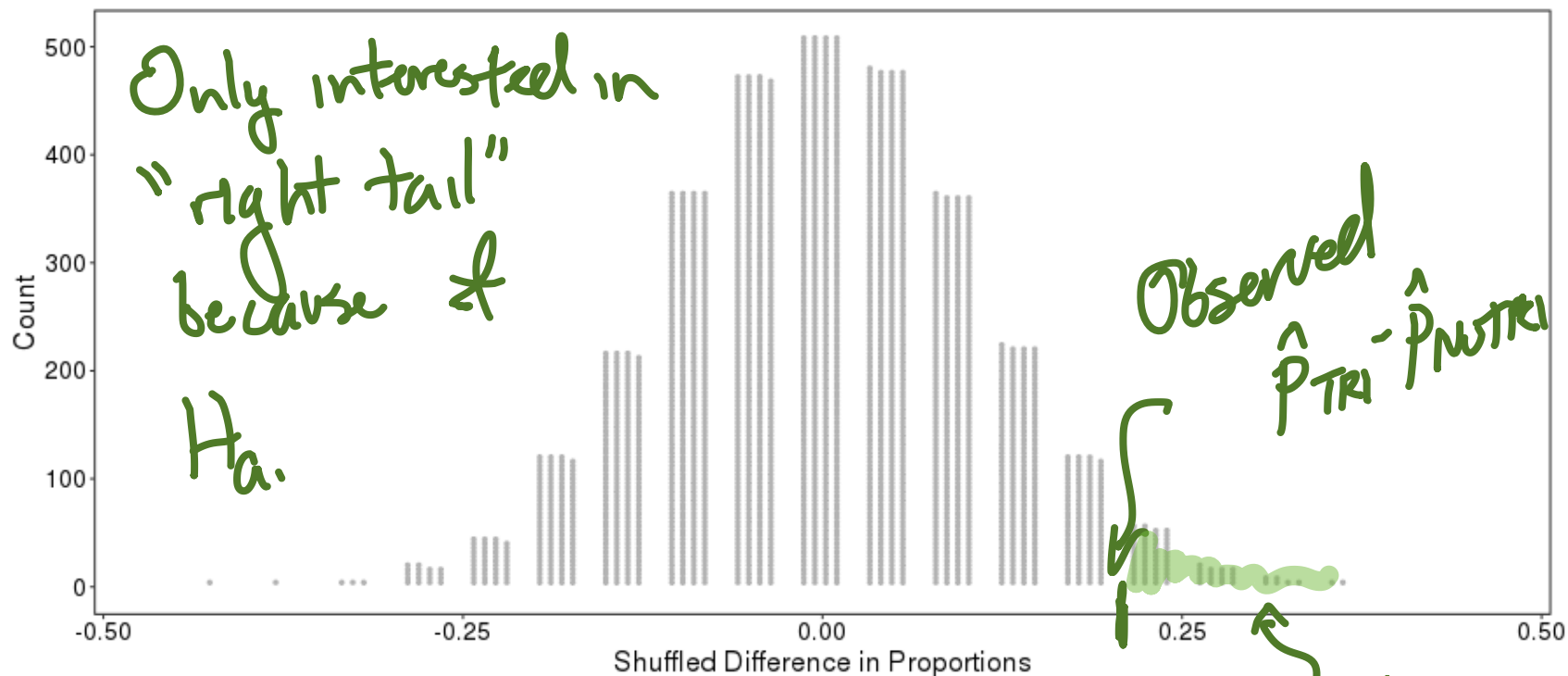
cases where staph was present and put the remaining 61 cards in a pile to represent where staph was

absent. Then we tabulate the results and determine the difference in precipitation rates, $\hat{p}_{Tri} - \hat{p}_{NoTri}$.

Plotting these differences in precipitation rates would create a *null distribution*, a sampling distribution of $\hat{p}_{Tri} - \hat{p}_{NoTri}$ that assumes H_0 is true.

Step 3 – the p-value

d. What proportion of the simulated studies in the null distribution were as extreme or more extreme than the observed difference of Staph infection rates?



About 1.86%

Step 4 – draw a conclusion

e. What amount of evidence is there against the null model?

If H_0 was exactly correct, we'd witness

$\hat{p}_{TRI} - \hat{p}_{NOTRI} \geq 0.2276$ in only about 1.6%

of repetitions. This is strong evidence against

the model hypothesized by H_0 and

suggests further study is warranted.

If the p-value is:	Greater than 0.10 ($p > 0.10$)	Between 0.05 and 0.10 ($0.05 < p \leq 0.10$)	Between 0.01 and 0.05 ($0.01 < p \leq 0.05$)	Between 0.001 and 0.01 ($0.001 < p \leq 0.01$)	Less than 0.001 ($p \leq 0.001$)
we will say we have:	little evidence	some evidence	strong evidence	very strong evidence	extremely strong evidence

Example 2.9: Did you kill them, Kristin?!

For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts.

Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. Part of the evidence used against Gilbert in her criminal trial was an analysis of more than one thousand 8-hour shifts during the time she worked in the ICU. Data are below:

Example 2.9: Did you kill them, Kristin?!

a. What is the observed difference in death rates across the shifts where Gilbert was present and when she was absent? Use the appropriate notation.

Shift	At least one person died	No one died	Total
Gilbert Absent	34	1350	1384
Gilbert Present	40	217	257
Total	74	1567	1641

$$\hat{p}_A - \hat{p}_P = -0.1312$$



Example 2.9: Did you kill them, Kristin?!

b. **Step 1 – Set up hypotheses:** We wish to see if the difference in mortality rates found in the courtroom (that Gilbert's presence actually *increases* the risk of a patient dying) should be considered evidence of a **true** difference in mortality rates or an aberration of collected data. What are the appropriate hypotheses to be tested?

$$H_0: \underline{p_A = p_P} \quad \text{vs.} \quad H_a: \underline{p_A < p_P}$$



Example 2.9: Did you kill them, Kristin?!

c. **Step 2 – Seeing what the null thinks is typical:** Assume there truly is no difference in mortality infection rates for across the shift types.

Write absent on 1384 cards and present on 257 cards.

Then we would thoroughly shuffle the cards and deal 74 cards into one pile to represent shifts where at least one patient died and put the remaining 1567 cards in a pile to represent where no one did. Then we tabulate the results and determine the difference in mortality rates, $\hat{p}_{NG} - \hat{p}_G$.

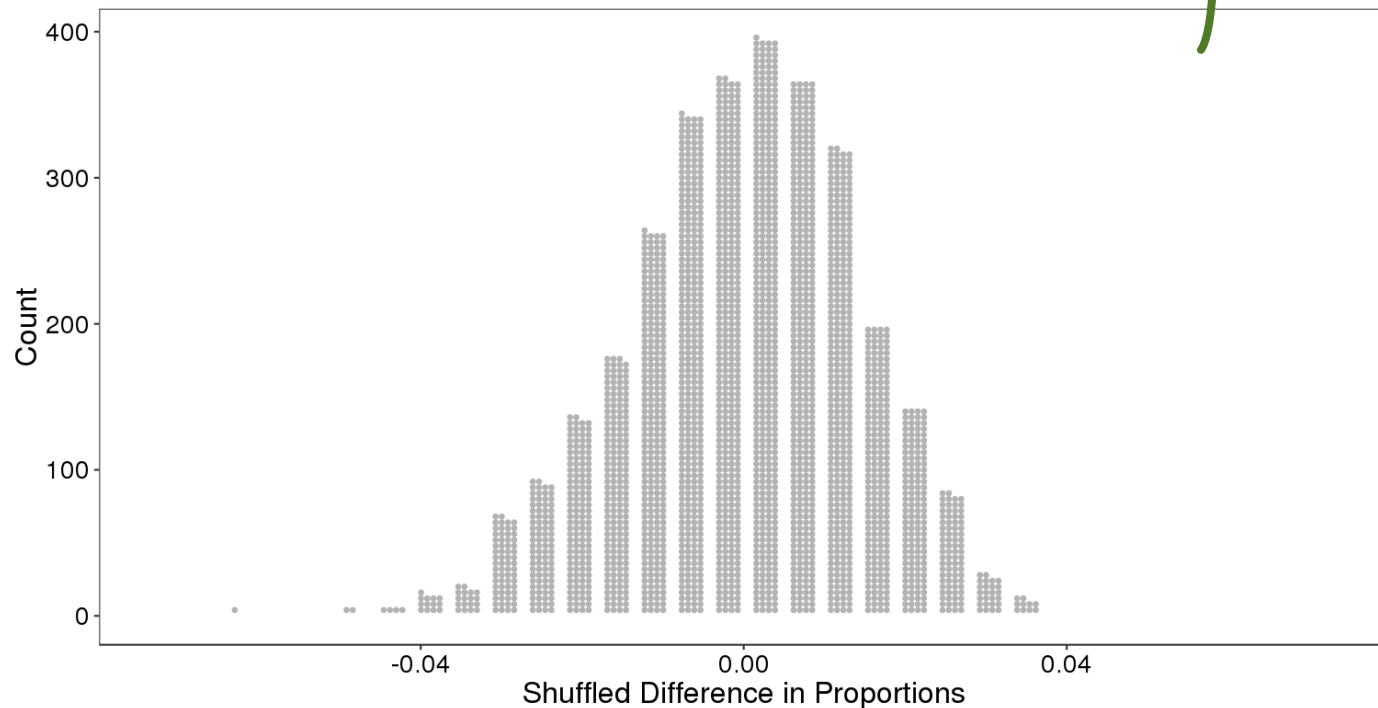
Plotting these differences in mortality rates would create a *null distribution*, a sampling distribution of the values of $\hat{p}_{NG} - \hat{p}_G$ expected by H_0 .



Example 2.9: Did you kill them, Kristin?!

d. **Step 3 – evaluating evidence against the null: the p-value.** What proportion of the simulated studies in the null distribution were as extreme or more extreme than the observed difference of Staph infection rates?


p-value = 0.



Example 2.9: Did you kill them, Kristin?!

e. What amount of evidence is there against the null model?

If the p-value is:	Greater than 0.10 ($p > 0.10$)	Between 0.05 and 0.10 ($0.05 < p \leq 0.10$)	Between 0.01 and 0.05 ($0.01 < p \leq 0.05$)	Between 0.001 and 0.01 ($0.001 < p \leq 0.01$)	Less than 0.001 ($p \leq 0.001$)
we will say we have:	little evidence	some evidence	strong evidence	very strong evidence	extremely strong evidence



Lecture 2-4: More hypothesis tests

This lecture provides a few brief additional examples of randomization-based hypothesis tests.

In contrast to Lecture 2-3, which dealt with tests concerning a difference of proportions between two groups, we now look at randomization-based tests for a single proportion.

Example 2.10: Fewer side effects

A pharmaceutical product is known to cause adverse events (side effects such as headaches, dizziness, stomach aches, etc.) in about 15% of all patients. The manufacturer hopes incorporating a new enteric-coating formulation (a polymer barrier often applied to oral medication to prevent disintegration in gastric environments) should lessen the rate of adverse effects.

They administer the product to a sample of $n = 200$ patients, 21 of whom report adverse events.

Do these sample results suggest that the enteric coating lessens the rate of adverse effects?

a. What is the observed rate of adverse events? Use the appropriate notation.

$$\hat{p} = \frac{21}{200} = 0.105$$

Fewer side effects

Step 1: Setting up hypotheses – What are the appropriate hypotheses to be tested?

$$H_0: \underline{p = 0.15}$$

vs.

$$H_a: \underline{p < 0.15}$$

Fewer side effects

Step 2: Seeing what the null thinks is typical – We can use randomization to simulate what we would expect to happen under the null model that specifies enteric-coating does not lessen the rate of adverse events.

Take 15 green cards and 85 white cards to create a deck that represents the null value.

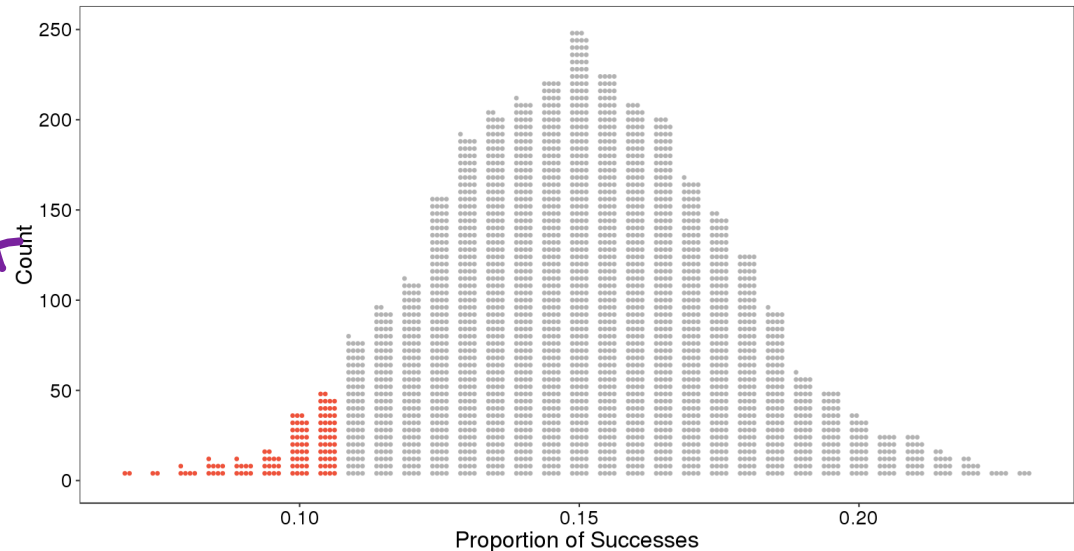
Shuffle and draw a card. Record the color and place the card back in the pile. Do this 200 times to create a single randomized sample.

Calculate the sample proportion of green cards in each set of draws to form a distribution of sample proportions we would expect to see under the null model.

Fewer side effects

Step 3: The p-value

$$p\text{-value} = \frac{122}{3000} = 0.0407$$



Overlay Normal Curve

Show summary statistics

Mean = 0.15 SD = 0.0252

Count Samples

less than

.105

122/3000 (0.0407)

d. What proportion of the simulated studies in the null distribution had adverse event rates as low or lower than the rate in the observed sample?

Fewer side effects

Step 4: Draw a conclusion

Example 2.11: Rubik's Cube Competitions

At competitions for solving the Rubik's Cube, competitors can participate in both speed events and blindfolded events. In blindfolded events, the competitor must memorize all information about the puzzle before making any turns, and then puts on a blindfold and solves the cubes without looking.

One such event is the "Multiple Blindfolded" event, where a competitor must memorize and successfully solve as many Rubik's cubes as possible without failing. The world record holder, Marcin Kowalczyk, has solved a perfect 41 cubes out of 41 attempted in under an hour.

Marcin unofficially attempted to solve 50 Rubik's Cubes blindfolded and successfully solved 49 cubes.

Suppose that it is believed that a typical blindfolded competitor successfully solves a cube blindfolded in 80% of their attempts. However, Marcin claims that he is better than the typical competitor and wants to use his 100 cube attempt results to put that claim to the test.

Example 2.11: Rubik's Cube Competitions

a. What is Marcin's observed successful solve rate? Use the appropriate notation.

$$\hat{p} = \frac{49}{50} = 0.98$$

b. **Step 1: Setting up hypotheses** – What are the appropriate hypotheses to be tested?

$$H_0: \underline{p = 0.8}$$

vs.

$$H_a: \underline{p > 0.8}$$



Example 2.11: Rubik's Cube Competitions

c. **Step 2: Seeing what the null thinks is typical** – We can use randomization to simulate what we would expect to happen under the null model that specifies Marcin's solving rate is no better than others'.

Take 8 green cards and 2 white cards to create a deck that represents the null value.

Shuffle and draw a card. Record the color and place the card back in the pile. Do this 50 times to create a single randomized sample.

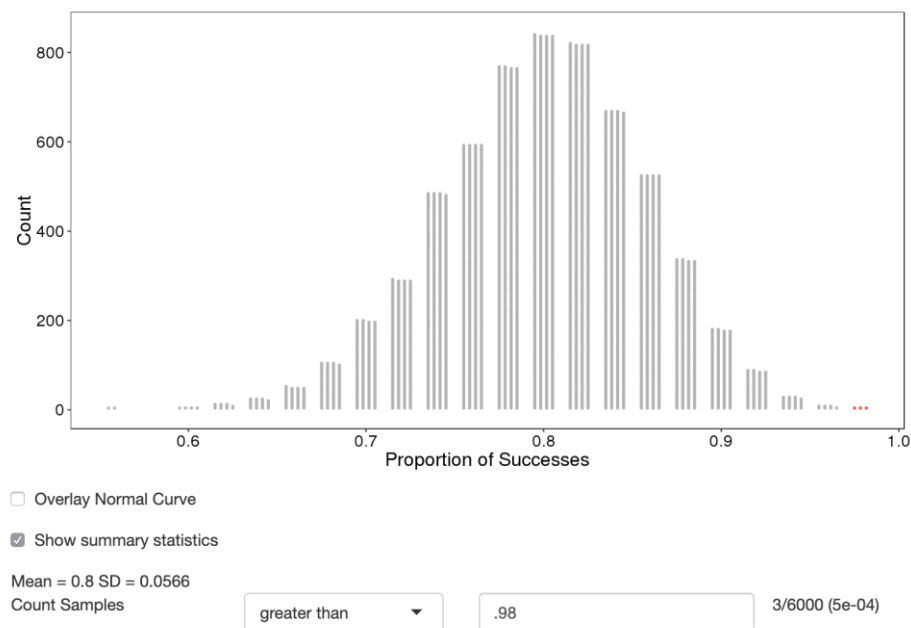
Calculate the sample proportion of green cards in each set of draws to form a distribution of sample proportions we would expect to see under the null model.



Example 2.11: Rubik's Cube Competitions

Step 3: the p-value.

$$\frac{3}{6000} = 0.0005$$



d. What proportion of the simulated studies in the null distribution that assumes Marcin only has an 80% success rate shows him have an observed success rate of at least 95%?



Example 2.11: Rubik's Cube Competitions

Step 4: draw a conclusion

If the p-value is:	Greater than 0.10 ($p > 0.10$)	Between 0.05 and 0.10 ($0.05 < p \leq 0.10$)	Between 0.01 and 0.05 ($0.01 < p \leq 0.05$)	Between 0.001 and 0.01 ($0.001 < p \leq 0.01$)	Less than 0.001 ($p \leq 0.001$)
we will say we have:	little evidence	some evidence	strong evidence	very strong evidence	extremely strong evidence



Lecture 2-5: The normal approximation

Step 4: draw a conclusion

Lecture 2-5: The normal approximation

In lecture 2-2, we created many sampling distributions to describe how a statistic behaves over repeated sampling. In lectures 2-3 & 2-4, we created many *null* distributions to describe how a statistic behaves under some null model.

What characteristics did they share?

1. Unimodal
2. Bell-shaped
3. Centered at population (or hypothesized) parameter.

The Central Limit Theorem

The Central Limit Theorem requires two conditions:

1. The observations are independent .
Independence is often guaranteed in an observational study by taking a random sample from a population. It can also be guaranteed in the context of a controlled experiment if we randomly assign individuals to treatment groups.
2. Sample is sufficiently large .
We must gather a *sufficiently large* sample of data, regardless of whether it is an observational study or controlled experiment, for the Central Limit Theorem to take effect. Just how large is large enough? That differs from one context to the next, and we'll provide guidelines as we encounter them through the rest of the semester.

The Central Limit Theorem

If these conditions are met...

...then the sampling distribution of many sample statistics can be well-approximated by a mathematical function called

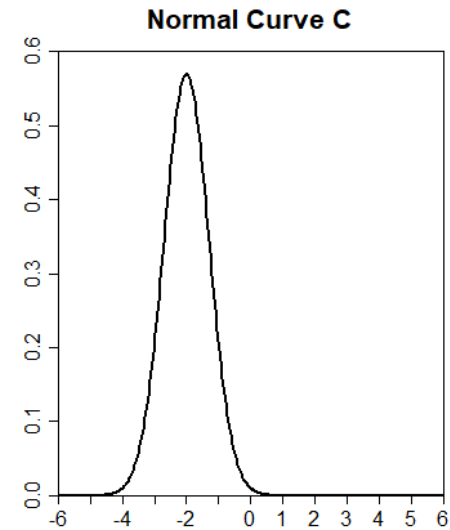
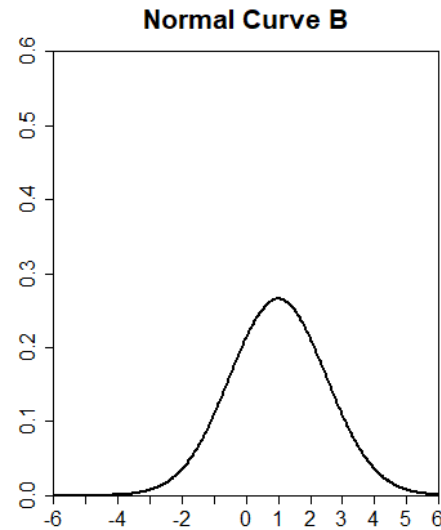
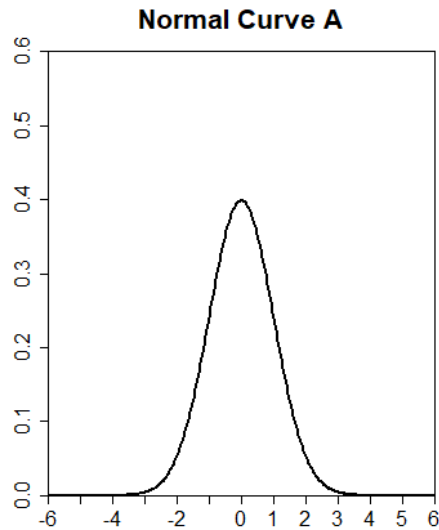
a normal density function.

The equation of a normal curve is given by:

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

Basic facts of the Normal Curve

Consider each of the three normal distributions below, Curves A, B, and C.



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

Basic facts of the Normal Curve

Consider each of the three normal distributions below, Curves A, B, and C.

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

The normal distribution can be adjusted using two parameters, the **mean μ** and the **standard deviation σ** .

Changing the mean of a normal curve

shifts curve left ; right .

Changing the standard deviation of a normal curve

stretches / compresses curve .

Basic facts of the Normal Curve

If a normal curve has mean μ and standard deviation σ , statisticians will abbreviate the equation of the curve

as $N(\mu, \sigma)$.

When a normal curve has mean $\mu = 0$ and standard deviation $\sigma = 1$, we label the curve the

standard normal curve.

Basic facts of the Normal Curve

Two crucial facts:

Although the normal curve has an

infinite domain, most of our

interest is focused on the interval

$(\mu - 3\sigma, \mu + 3\sigma)$.

The integral of the normal curve over its domain

(i.e., from $-\infty$ to ∞) is equal to exactly 1.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

Using your Calculator to Find Probabilities, Areas, and Percentiles

To find a probability if a data value is known:

2nd Vars – “normalcdf” – enter “lower limit, upper limit, mean, sd”

Example: $P(900 \leq X \leq 1200)$

Enter 2nd Vars – normalcdf (900, 1200, 1060, 195) enter.

Answer 0.557644

To find data values when given an area (or percentage):

2nd Vars – “invnorm” – enter (enter area to the left as decimal, mean, sd)

Example: Find the score or data value corresponding to the 80th percentile.

2 nd Vars – “invnorm” – (0.80, 1060, 195) enter

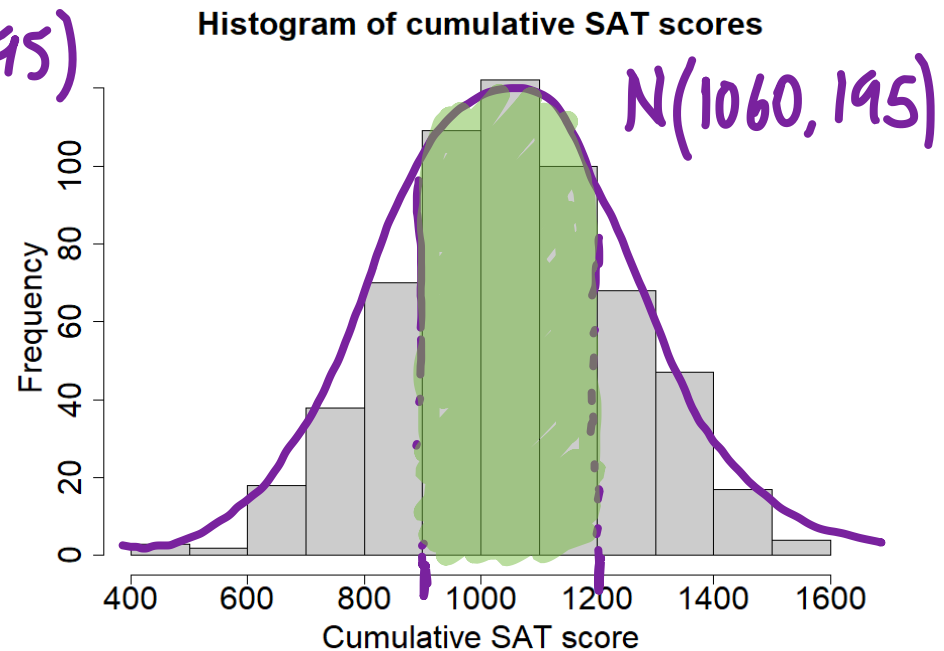
Answer: 1224

Example 2.8: Integrating the normal curve

Cumulative SAT scores are approximated well by a normal model, $N(1060, 195)$. A histogram of a sample of these scores is shown, and we want to be able to answer questions like the following:

- Approximately what proportion of test takers score between 900 and 1200 on the SAT?

$$\text{normalcdf}(900, 1200, 1060, 195) \\ = 0.5576$$

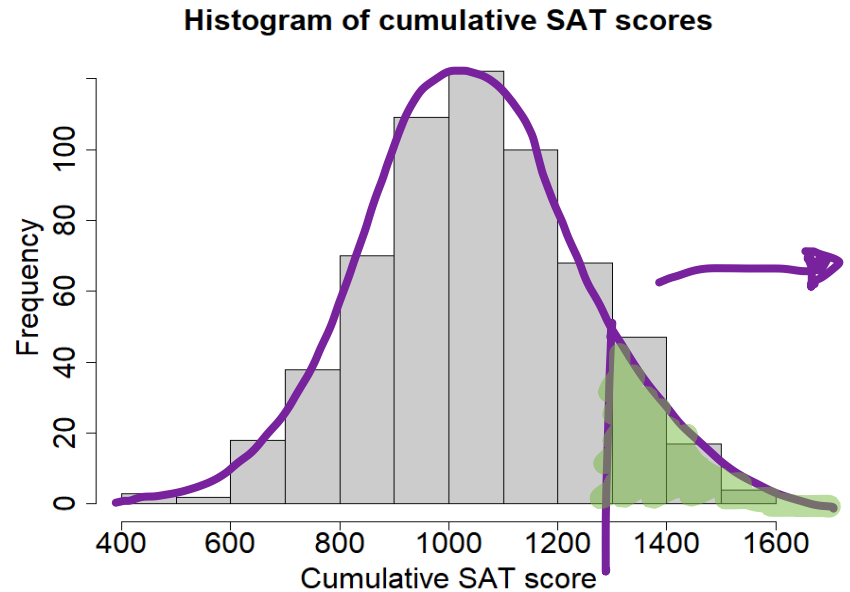


Example 2.8: Integrating the normal curve

Cumulative SAT scores are approximated well by a normal model, $N(1060, 195)$. A histogram of a sample of these scores is shown, and we want to be able to answer questions like the following:

b. A randomly-selected SAT test-taker is about to sit for the test. Nothing is known about her aptitude. What is the probability that she scores at least 1300 on her SATs?

$$\text{normalcdf}(1300, 10^{10}, 1060, 195) \\ = 0.1092$$

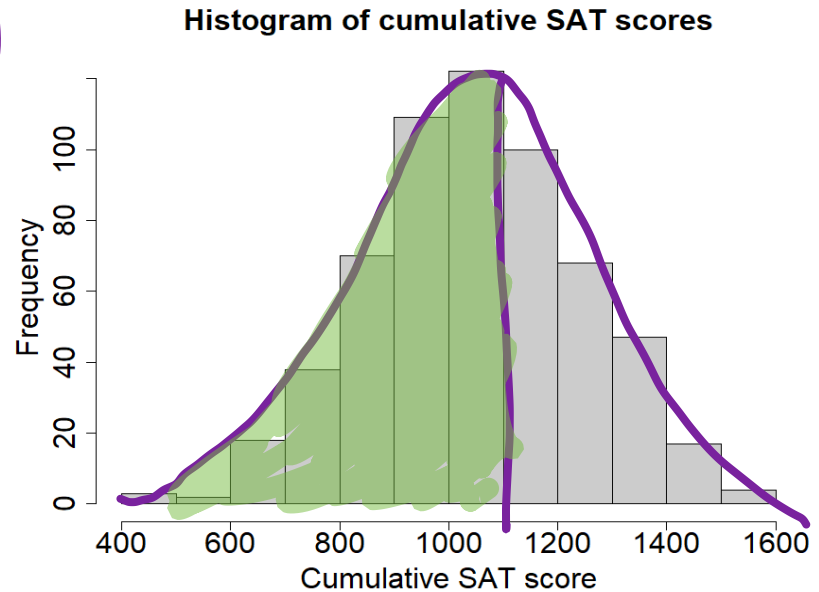


Example 2.8: Integrating the normal curve

Cumulative SAT scores are approximated well by a normal model, $N(1060, 195)$. A histogram of a sample of these scores is shown, and we want to be able to answer questions like the following:

c. Another SAT test-taker is taking the SAT for a second time after earning a 1100 on his first attempt. What was the percentile of his first score?

$$\text{normalcdf}(-10^{10}, 1100, 1060, 195) \\ = 0.5813$$

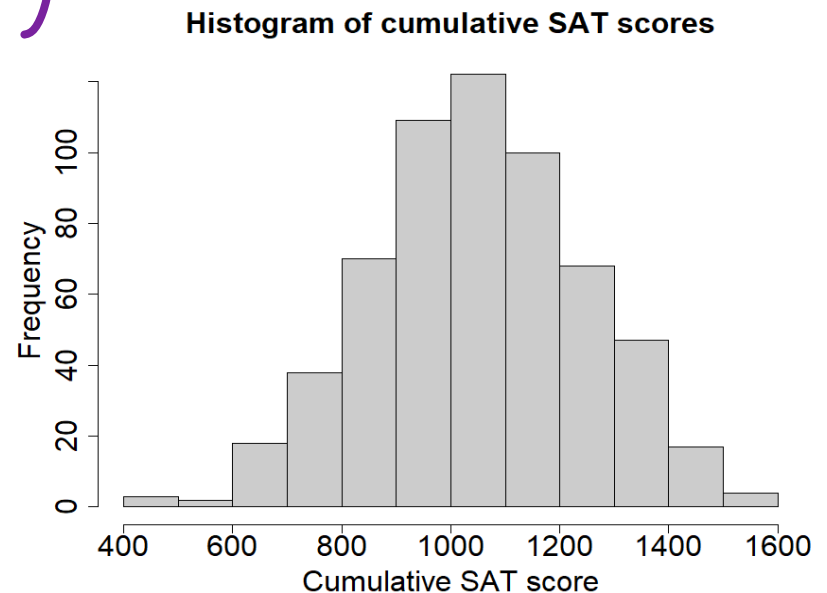


Example 2.8: Integrating the normal curve

Cumulative SAT scores are approximated well by a normal model, $N(1060, 195)$. A histogram of a sample of these scores is shown, and we want to be able to answer questions like the following:

d. What is the SAT score of someone who scores at the 80th percentile?

$$\text{inv Norm } (0.8, 1060, 195) \\ = 1224$$



Example 2.9: Using Z-scores to make comparisons

There are two major tests of readiness for college, the ACT and the SAT. Both are well-modeled by a normal curve.

- ACT scores are reported on a scale from 1 to 36 with mean = 20.8 and sd = 4.8.
- SAT scores are reported on a scale from 400 to 1600 with
 - mean = 1060 and sd = 195.

Suppose Tonya took the SAT and scored 1320. Jessica took the ACT and scored 28.

If we assume that both tests measure the same thing, who has the higher score?

Example 2.9: Using Z-scores to make comparisons

e. We can answer this by computing the z-score for each student.

Tonya: $z = 1.41$

Jessica: $z = 1.5$

f. Who did better on their college prep test based on the z-scores?

Jessica



Example 2.10: More z-score practice

1. Find $P(Z \leq 1.22)$. Without any calculations, find $P(Z < 1.22)$.

$$0.8889$$

2. What z-scores provide the bounds for the middle 50% of the standard normal distribution?

$$\pm 0.67$$

3. What z-scores provide the bounds for the middle 95% of the standard normal distribution?

$$\pm 1.96$$

4. Without any calculations, find $P(Z \leq -7.2)$. ≈ 0

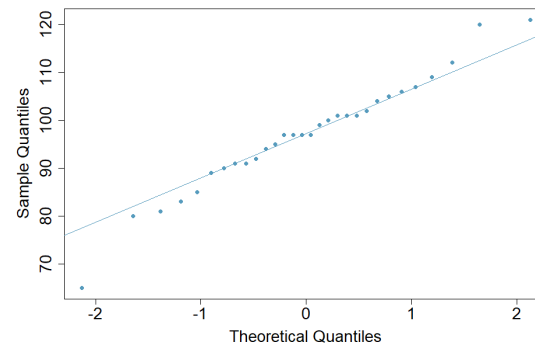
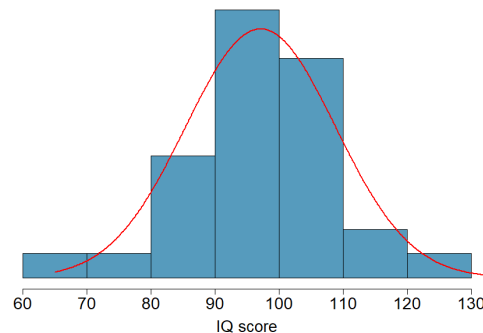


Evaluating the normal approximation

While normal models are helpful and convenient, remember that they are *only an approximation*.

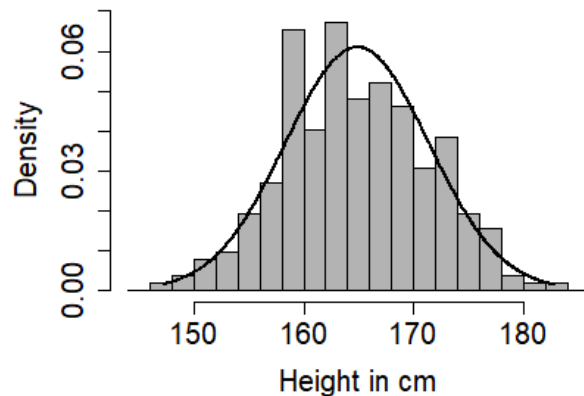
There are two simple visual ways to assess whether a normal approximation is appropriate:

1. Plot histogram & see if data is bell-shaped
2. Construct a QQ-plot & assess points for linearity

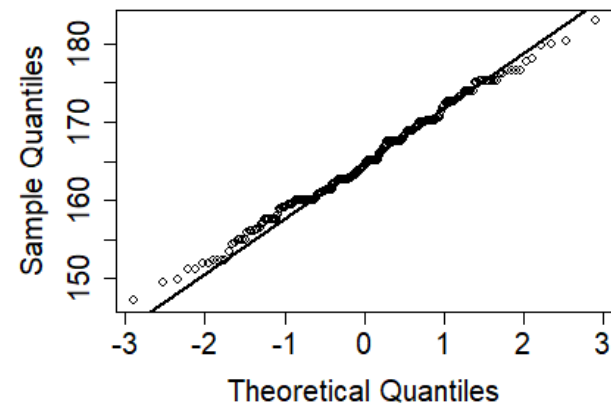


Evaluating the normal approximation

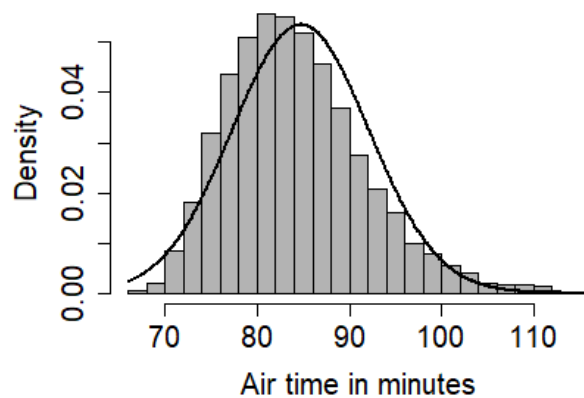
Histogram of Women Heights



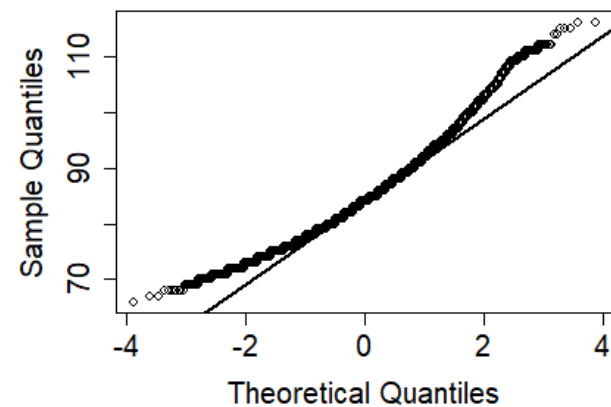
QQ-Plot of Women Heights



Flight Times

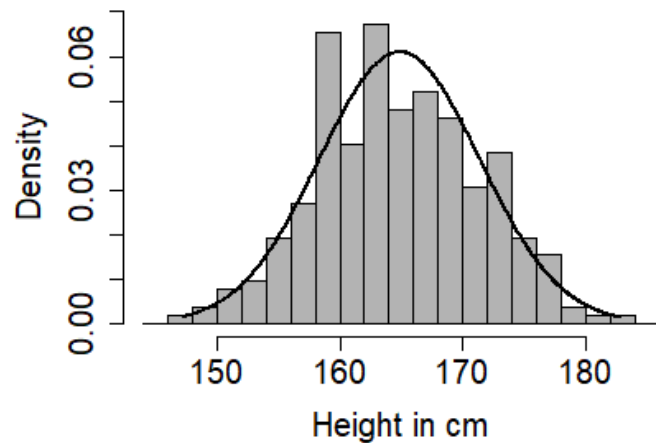


QQ-Plot of Flight Times

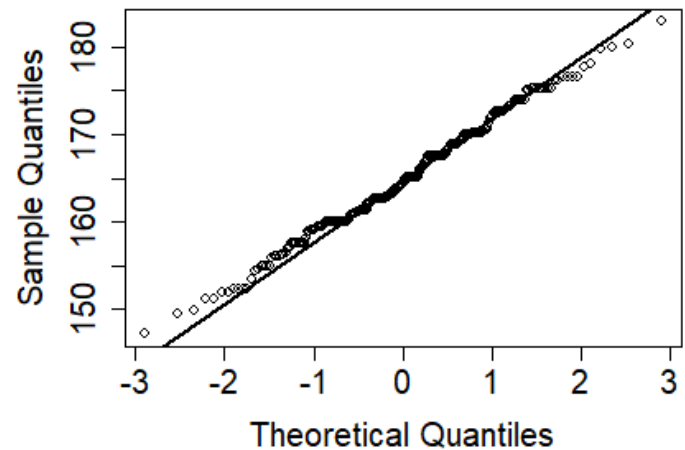


Evaluating the normal approximation

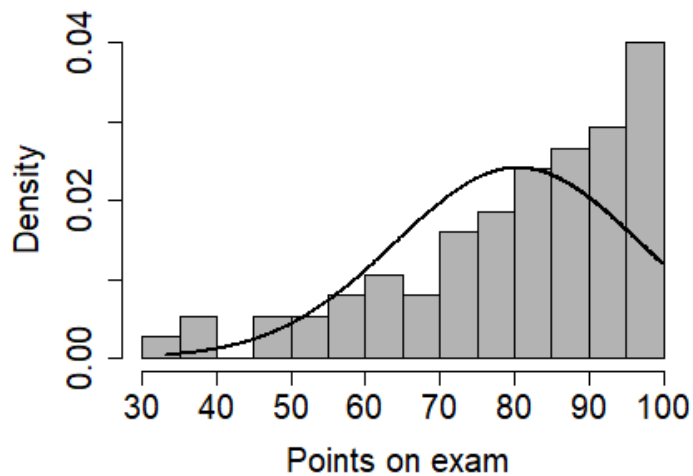
Histogram of Women Heights



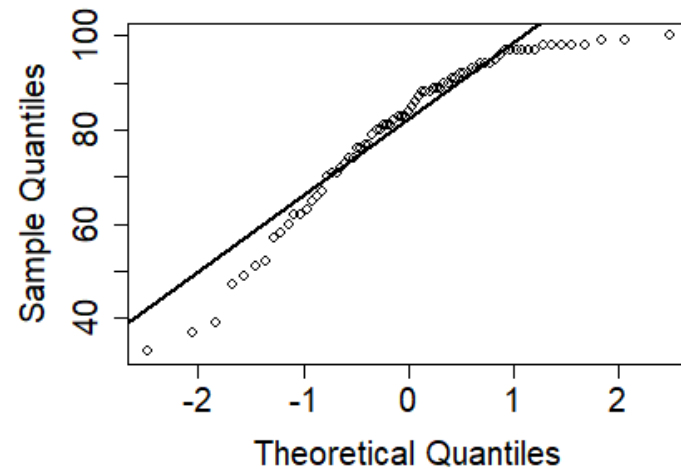
QQ-Plot of Women Heights



Exam Scores



QQ-Plot of Exam Scores



Normal approximations for hypothesis tests

Recall an earlier research scenario investigating whether enteric-coating formulation (a polymer barrier often applied to oral medication to prevent disintegration in gastric environments) should lessen the 15% rate of adverse effects.

The researchers administered the product to a sample of $n = 200$ patients, 21 of whom report adverse events.

The hypotheses that were tested were:

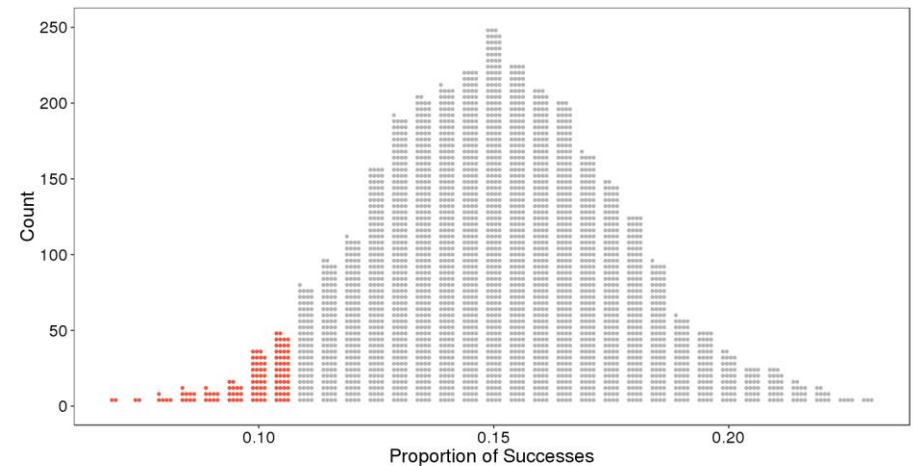
$$H_0: p = 0.15 \quad \text{vs.} \quad H_a: p < 0.15$$

Normal approximations for hypothesis tests

The hypotheses that were tested were:

$$H_0: p = 0.15 \quad \text{vs.} \quad H_a: p < 0.15$$

a. The randomization-based distribution that displayed the results expected under the null is below. What was the p-value and conclusion to this test?



Overlay Normal Curve

Show summary statistics

Mean = 0.15 SD = 0.0252

Count Samples

less than

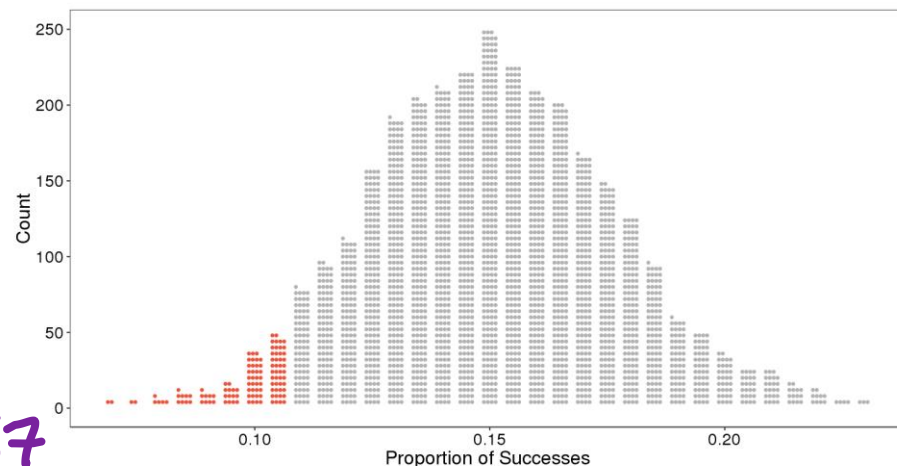
.105

122/3000 (0.0407)

Normal approximations for hypothesis tests

b. Try to replicate these results using a normal approximation of the null distribution in (a). Calculate a Z score using the observed 'adverse event' rate of $\frac{21}{200} = 0.105$, along with the mean and standard deviation of the null distribution. [Notice how we use the standard error of the statistic as the standard deviation for the z score.]

$$\begin{aligned}
 Z &= \frac{\text{observed} - \text{expected}}{\text{standard deviation}} \\
 &= \frac{\text{sample statistic} - \text{null value}}{\text{standard error}} \\
 &= \frac{0.105 - 0.15}{0.0252} = -1.7857
 \end{aligned}$$



Overlay Normal Curve

Show summary statistics

Mean = 0.15 SD = 0.0252

Count Samples

less than

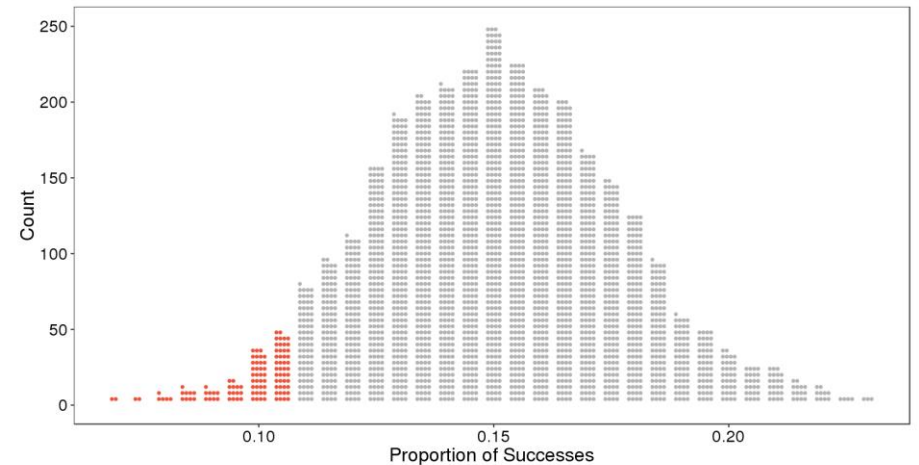
.105

122/3000 (0.0407)

Normal approximations for hypothesis tests

c. Identify the p-value corresponding to this z score. How does it compare to the p-value from the randomization simulation? Would we make the same evaluation regarding the null hypothesis?

$$\text{normalcdf}(-10^{99}, -1.7837, 0, 1) = 0.037$$



Overlay Normal Curve

Show summary statistics

Mean = 0.15 SD = 0.0252

Count Samples

less than

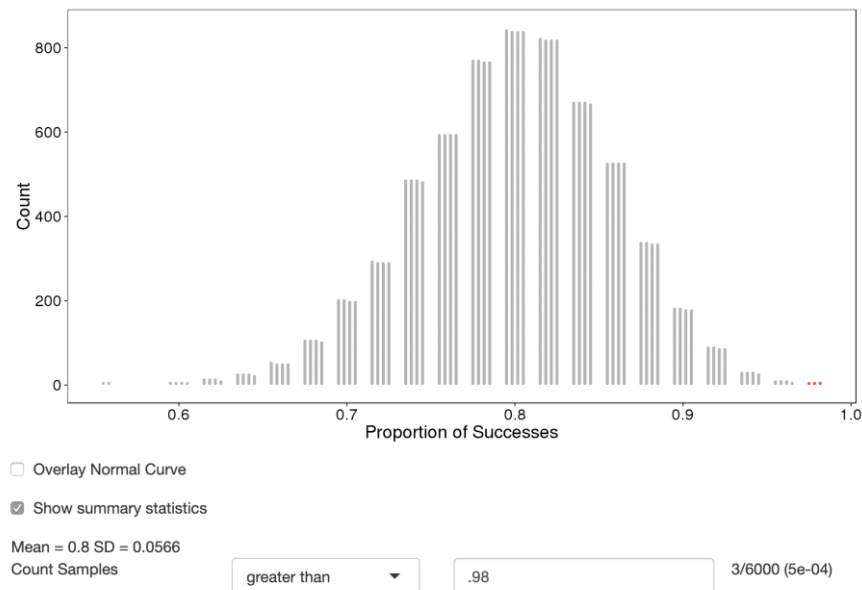
.105

122/3000 (0.0407)

Normal approximations for hypothesis tests

Now consider a second research scenario where Marcin Kowalczyk, a Rubik's Cubing champion, unofficially attempted to solve 50 Rubik's Cubes blindfolded and successfully solved 49 cubes.

d. The randomization-based distribution that displayed the results expected under the null is below. What was the p-value and conclusion to this test?



Normal approximations for hypothesis tests

e. Try to replicate these results using a normal approximation of the null distribution in (d). Calculate a Z score using Marcin's observed solve rate of $\frac{49}{50} = 0.98$, along with the mean and standard deviation of the null distribution. [Notice how we use the standard error of the statistic as the standard deviation for the z score.]

$$Z = \frac{0.98 - 0.8}{0.0566} = 3.1802$$

f. Identify the p-value corresponding to this z score. How does it compare to the p-value from the randomization simulation? Would we make the same evaluation regarding the null hypothesis?

$$\text{normal cdf}(3.1802, 10^{-10}, 0, 1) \\ = 0.0007$$



Normal approximations for hypothesis tests

Think about it!

Consider the results from (c) and (f). Why did the normal approximation closely resemble the randomization-based p-value in one study but not the other?