Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Sample size.
2. Statistic and parameter.
3. Categorical and quantitative variables.
4. Statistical significance and testing.
5. Null and alternative hypotheses.
6. Z statistic.
7. Simulating null distributions.
8. p-values.
9. Heart transplant example.
10. Standardized statistic.
11. Note on 1-sided and 2-sided tests.

Read through chapter 1.
The course website is http://www.stat.ucla.edu/~frederic/13/W24
No class Mon Jan15.

1. Sample size.

Each record typically corresponds to an *observational unit*, and the number of observed units in the analysis is called the sample size, *n*.

In some situations, the population size might be known and you might have a Simple Random Sample (SRS) from the population. The sample size then is the number of people in your sample.
For instance, there are 4 million births every year in the United States.
Suppose we sample 1,000 of them at random from this population, and record for each pregnancy, the number of weeks of pregnancy, and the height, weight and gender of the baby at birth.
Here n = 1,000. Each baby is an observational unit.

2. Statistic and parameter.

A statistic is a numerical description of your sample. Another word for statistic is *random variable*. The sample is typically considered random, and if a different sample were obtained, then the statistic might be different.

A parameter, however, is a property of the whole population. If a different sample were obtained, the parameter would not change.

Parameters are properties of the population. Typically unknown. Represented by Greek letters (like μ or σ).
Statistics are properties of the sample.
Represented by Roman letters (like $\bar{x}$ or s).
Typically, you're interested in a value of a parameter. But you can't know it. So you *estimate* it with a statistic, based on the sample.

There are two means and two standard deviations.
The sample mean $\overline{x}$ and sample std deviation *s* are statistics.
Define the population average μ as the sum of all values in the population ÷ the number of subjects in the population. (parameter).
It turns out $\bar{x}$ is an unbiased estimate of μ.
That is, $\bar{x}$ is neither higher nor lower, on average, than μ, if we sampled repeatedly.

3. Categorical and quantitative variables.
For a quantitative variable, the responses are all numbers and the difference between two observations has a natural interpretable meaning. For categorical variables there is no such meaning to the difference between two observations.
The line between the two terms can sometimes be a bit blurry.
e.g. gender of baby would be categorical.
height, weight, and number of weeks would be quantitative.
eye color, birth type, or pain medication used might be examples of categorical variables here with multiple possibilities.

4. Statistical significance and Testing.

According to the CDC, 4 million babies were born in the U.S. in 2014 and 10% were born preterm (< 37 weeks). Suppose you take a simple random sample (SRS) of women with HG and you want to test whether the proportion preterm among women with HG might really be different from 10%.

Suppose in the sample of n=254 mothers with HG, $\hat{p}$ = 39/254 (15.35%) are preterm.

You want to test whether something like this could reasonably have happened just by chance alone, if the populations were actually identical with respect to delivery time.

Otherwise we conclude that the two population proportions are probably not equal, i.e. the difference observed is *statistically significant*.

There are different tests, but we'll just talk about the Z-test (or normal test) for now.

Assumptions:

SRS (or obs are known to be independent)

AND n is large (or pop is known to be normally distributed).

For testing proportions, there should be ≥ 10 of each type of response in the sample.

Here we have 39 preterm and 215 not preterm.

We will talk later in the course about these assumptions and also about the t test. If n is small, pop. is normal, and σ is unknown, then use t instead of Z.

After checking assumptions, the remaining steps in testing are

* stating the hypotheses,
* computing the test statistic (Z in this case),
* computing the p-value, and
* concluding.

5. Null and alternative hypotheses.

Let π be the proportion preterm in the population from which the sample was drawn.

Null hypothesis (Ho): $\pi = 10\%$.

This means that any observed difference between the sample proportion, $\hat{p}$, and 10%, was due to chance alone. Usually we specify these hypotheses numerically.

Alternative hypothesis (Ha): $\pi \neq 10\%$. Difference is not due to chance alone. (2-sided test.)

Or Ha: $\pi > 10\%$. Or Ha: $\pi < 10\%$. (1-sided tests). We will talk about this next lecture.

When in doubt, do a two-sided test, unless there is a specific reason to do a 1-sided test.

6. Z-statistic.

A test statistic is a summary of the strength of the evidence in your data.

Z-statistic here = $(\hat{p} - 10\%) \div SE$.

SE means Standard Error. We will talk about ways to get the SE either analytically or via simulations in a bit. For proportion problems like this, $SE = \sqrt{[\pi(1-\pi)/n]}$.

Here, analytically, the SE would be $\sqrt{[.10 \times .90 / 254]} \sim 1.88\%$.

$Z = (15.34\% - 10\%)/1.88\% = 2.84$.

The book calls Z a *standardized* test statistic.

It indicates how many SDs the observed statistic is above its hypothesized value under Ho.

The book also calls the SE the "standard deviation of the null distribution" but it is usually called the standard error or SE.

A value of Z far from 0 (more than 2 or less than -2) indicates strong evidence against the null hypothesis. A value of Z between -2 and 2 indicates weak evidence against the null.

$|Z| > 3$ indicates very strong evidence against the null.

7. Simulating null distributions and Standard Errors.

We observe $\hat{p}$ = 15.34% in our sample, and under Ho, the population percentage π = 10%. So we see a difference of 5.34%. This is our quantity of interest, and it is usually a difference like this. We want to see if that quantity of interest, 5.34%, is bigger than what we'd expect by chance under the null hypothesis.

The Standard Error (SE) is the standard deviation of the quantity of interest under the null hypothesis.

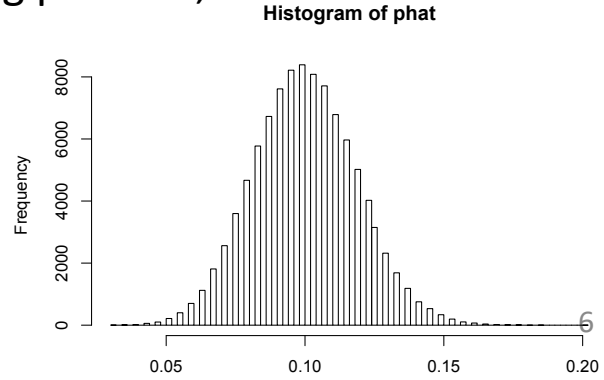
Many stat books just tell you the formulas to get the SE. Your book is different. They want to emphasize that in many cases you can estimate the SE by simulations.

In this example, under Ho, women with HG are just like the rest in terms of probability of delivering preterm. We have a SRS of size 254 from a population with π = 10% having preterm delivery. We can simulate 254 draws on the computer, where each draw is independent of the others and has a 10% chance of being preterm, and then see what results we get. In R, I did

```
x = runif(254)
y = (x<0.1)
phat = mean(y)
```

The first time, I got phat = 0.1259843. 12.60%.

I tried it many times, and here is what I got.



Simulating null distributions and Standard Errors.

We observe p = 15.34% in our sample, and under Ho, the population percentage π = 10%. So we see a difference of 5.34%. This is our quantity of interest, and it is usually a difference like this. We want to see if that quantity of interest, 5.34%, is bigger than what we'd expect by chance under the null hypothesis.

The Standard Error (SE) is the standard deviation of the quantity of interest under the null hypothesis.

Many stat books just tell you the formulas to get the SE. Your book is different. They want to emphasize that in many cases you can estimate the SE by simulations.

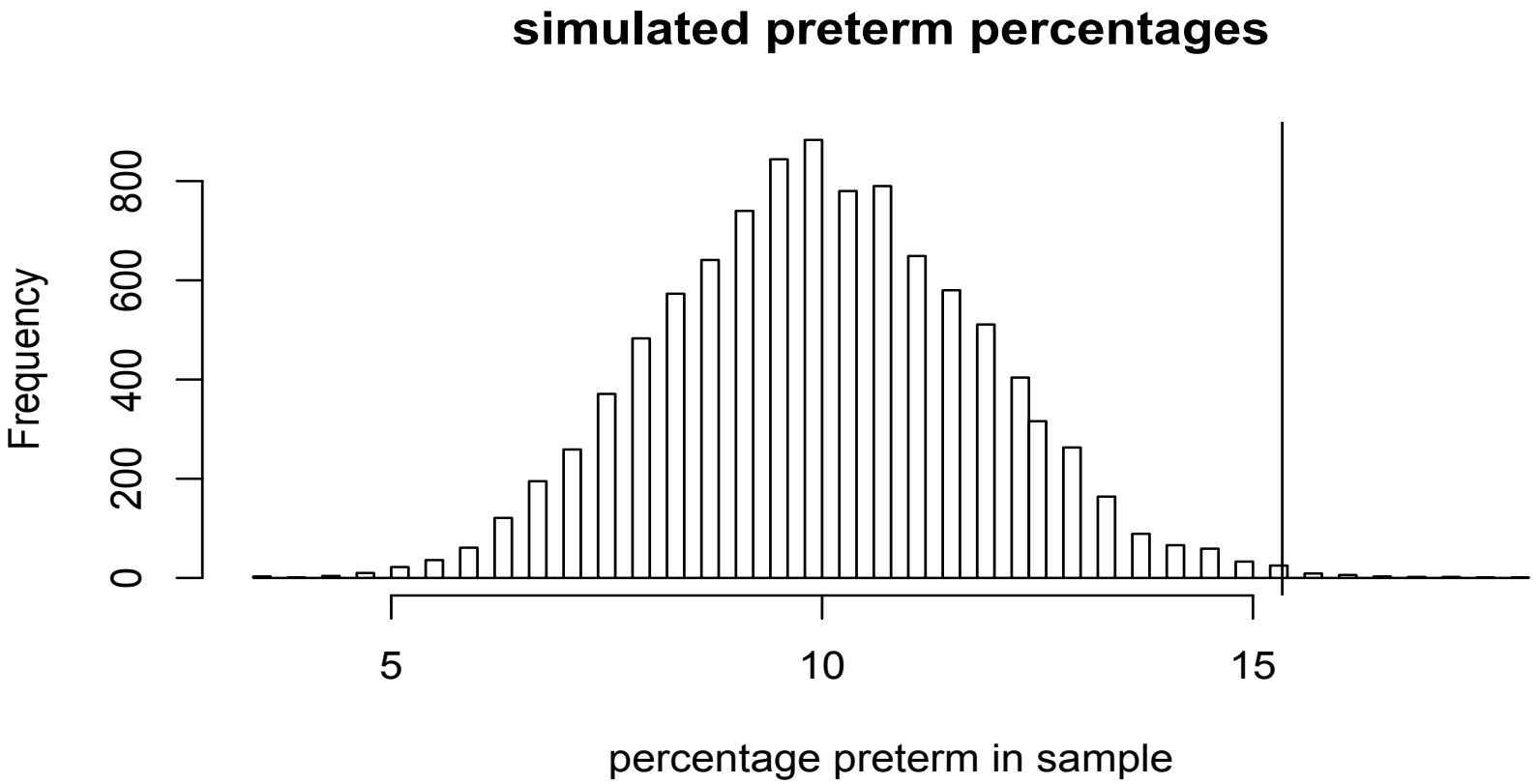
In this example, under Ho, women with HG are just like the rest in terms of probability of delivering preterm. We have a SRS of size 254 from a population with π = 10% having preterm delivery. We can simulate 254 draws on the computer, where each draw is independent of the others and has a 10% chance of being preterm, and then see what results we get. In R, I did

```
x = runif(254)
y = (x<0.1)
phat = mean(y)
```

The first time, I got phat = 0.1259843. 12.60%.

I tried it many times, and here is what I got.

```
a = rep(0,10000)
for(i in 1:10000){ x = runif(254); a[i] = mean(x<.1)}
hist(a*100,main="simulated preterm percentages", nclass=100,
      xlab="percentage preterm in sample")
abline(v=15.34)l
sd(a)                    ## 0.01885409
sqrt(.10 * .90 / 254) ## 0.01882367
```



8. p-values.

The p-value is the probability, assuming Ho is true, that the test statistic will be at least as extreme as that observed.

"What are the chances of that?"

The key idea is that the convention is to compute the probability of getting something as extreme as you observed or more extreme.

e.g. n = 5, π_o = 50%, $\hat{p}$= 4/5. The probability that $\hat{p}$ = 4/5 is 15.625%.

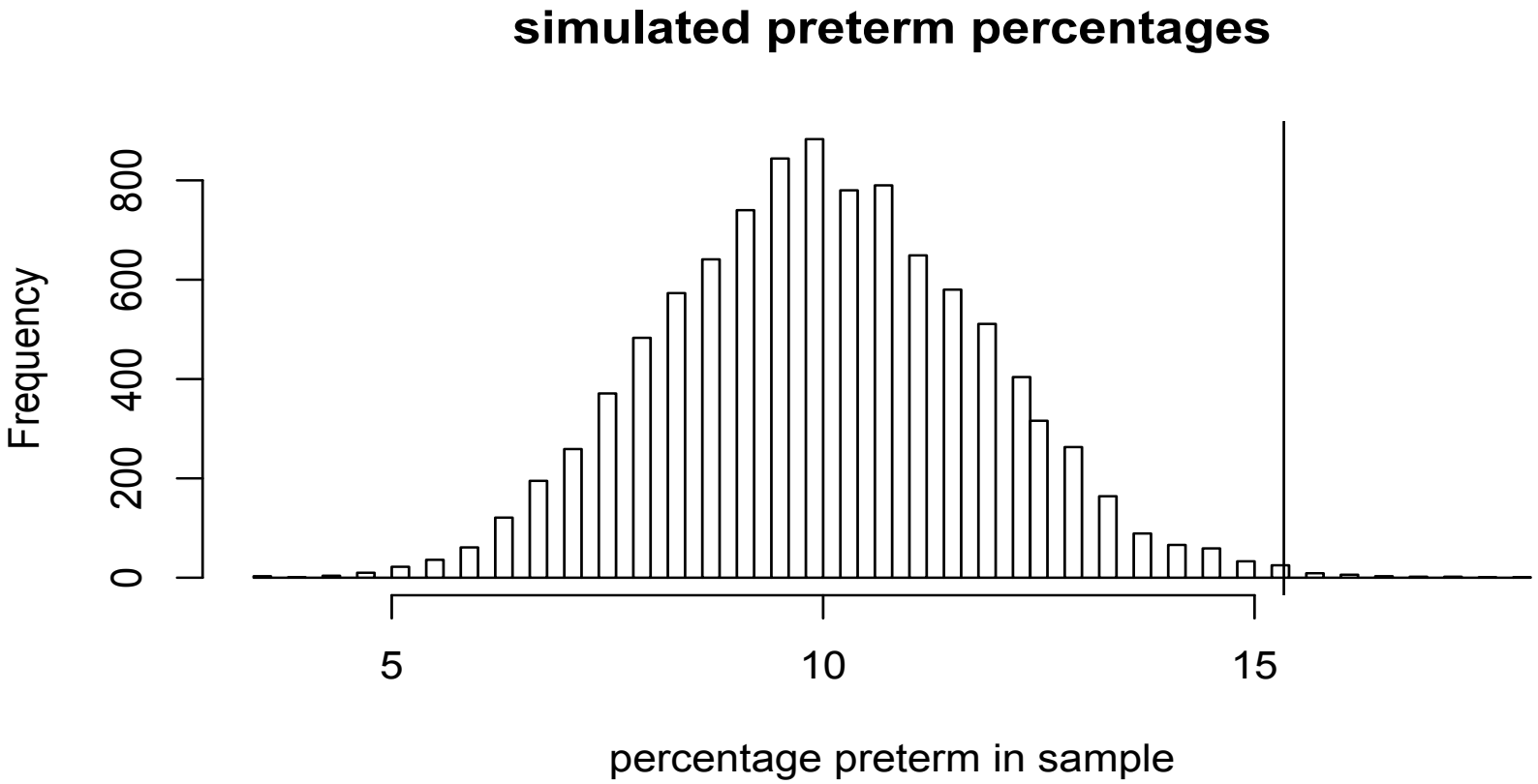
However, what if n = 400, π_o = 50%, and $\hat{p}$ = 201/400? Now the probability of getting 201/400 is 3.97%, but obviously the data are consistent with the null hypothesis that π = 50%.

Typically, one does a two-sided test, which means that by "extreme", we mean extreme in either direction. We want to see how in line our observed value of $\hat{p}$ =15.34% is with our null hypothesis of a population percentage of 10%. Could our sample of 15.34% preterm have come from a population of 10% preterm? A simulation with $\hat{p}$ > 15.34% would be more extreme than what we observed, and also a simulation with $\hat{p}$ < 4.66% would be more extreme than what we observed.

Guidelines for evaluating strength of evidence from p-values

- p-value >0.10, not much evidence against null hypothesis
- 0.05 < p-value $\leq$ 0.10, moderate evidence against the null hypothesis
- 0.01 < p-value $\leq$ 0.05, strong evidence against the null hypothesis
- p-value $\leq$ 0.01, very strong evidence against the null hypothesis

```
phat = rep(0,10000)
for(i in 1:10000){ x = runif(254); phat[i] = mean(x<.1)}
hist(phat*100,main="simulated preterm percentages", nclass=100,
      xlab="percentage preterm in sample")
abline(v=15.34)l
mean(abs(phat-.10)>.0534)        ## 0.0051
```



Continuing the HG example, using simulations of Ho we obtained samples of 254 values, and in 0.51% of these samples, at least 15.34% or more were preterm or less than 4.66% were preterm.
So we'd say the p-value is 0.51% for this two-sided test.
The observed difference is highly significant, and we have strong evidence against the null hypothesis of HG pregnancies having a 10% chance of being preterm like other pregnancies.

9. Heart Transplant Example.

Example 1.3

Heart Transplants

- The *British Medical Journal* (2004) reported that heart transplants at St. George's Hospital in London had been suspended after a spike in the mortality rate
- Of the last 10 heart transplants, 80% had resulted in deaths within 30 days
- This mortality rate was over five times the national average.
- The researchers used 15% as a reasonable value for comparison.

Heart Transplants

- Does a heart transplant patient at St. George's have a higher probability of dying than the national rate of 0.15?
- Observational units
  - The last 10 heart transplantations
- Variable
  - If the patient died or not
- Parameter
  - The actual probability of a death after a

Heart Transplants

- **Null hypothesis:** Death rate at St. George's is the same as the national rate (0.15).
- **Alternative hypothesis:** Death rate at St. George's is higher than the national rate.

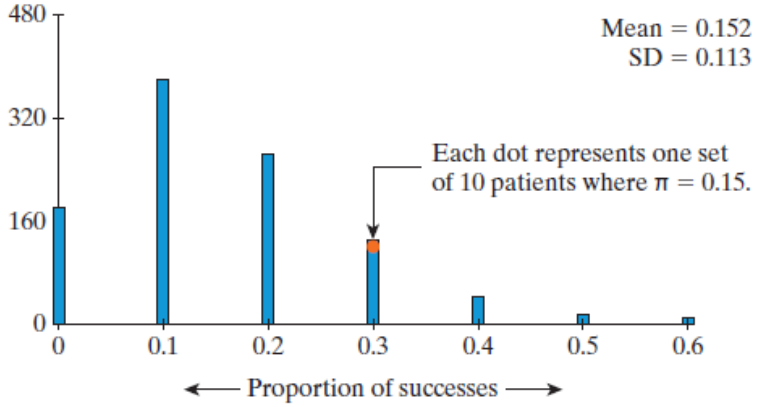
- $\mathbf{H_0}$**:** $\pi = 0.15$ $\mathbf{H_a}$**:** $\pi > 0.15$

- Our **statistic** is 8 out of 10 ($\hat{p} = 0.8$)

Heart Transplants

Simulation

- Null distribution of 1000 repetitions of drawing samples of 10 “patients” where the probability of death is equal to 0.15.



What is the p-value?

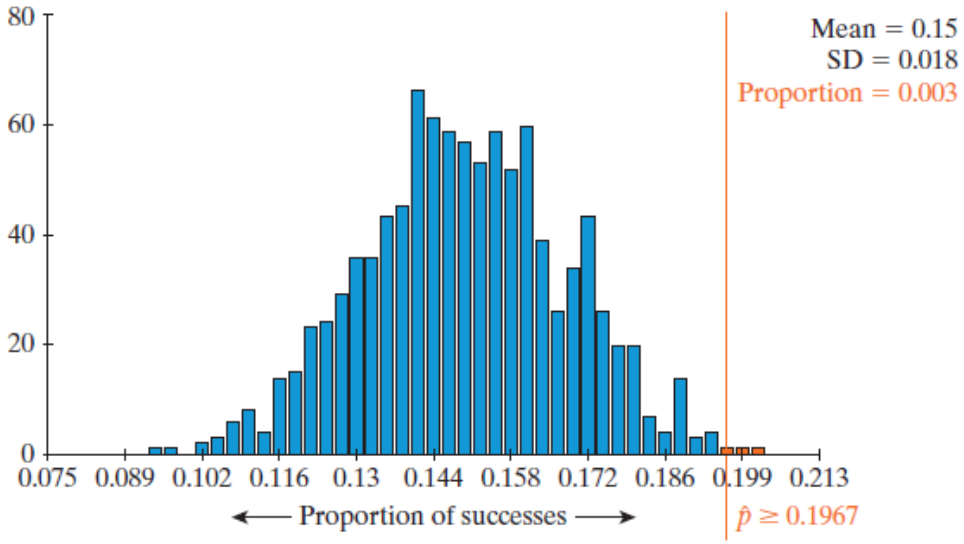
Heart Transplants

Strength of Evidence

- Our p-value is 0, so we have very strong evidence against the null hypothesis.
- Even with this strong evidence, it would be nice to have more data.
- Researchers examined the previous 361 heart transplantations at St. George's and found that 71 died within 30 days.
- Our new statistic, $\hat{p}$, is 71/361 ≈ 0.1967

Heart Transplants

- Here is a null distribution and p-value based on the new statistic.



Heart Transplants

- The p-value was about 0.003
- We still have very strong evidence against the null hypothesis, but not quite as strong as the first case

- Another way to measure strength of evidence is to ***standardize*** the observed statistic

Heart Transplants

- The p-value was about 0.003
- We still have very strong evidence against the null hypothesis, but not quite as strong as the first case

- Another way to measure strength of evidence is to ***standardize*** the observed statistic

10. The Standardized Statistic

- The ***standardized statistic*** is the number of standard deviations our sample statistic is above the mean of the null distribution (or below the mean if it is negative).

- $$z = \frac{\textit{statistic} - \textit{mean of null distribution}}{\textit{standard deviation of null distribution}}$$

- The sd of the null distribution is the *standard error.*
- For a single proportion, we will use the symbol *z* for standardized statistic.
- In the formula above, for the mean, we should use the long-term proportion (probability) given in the null hypothesis. If you do simulations, the mean of the simulated statistics should be close to this.

The Standardized Statistic

- Here are the standardized statistics for our two studies.

$$z = \frac{0.80 - 0.15}{0.113} = 5.75 \qquad z = \frac{0.197 - 0.15}{0.018} = 2.61$$

- In the first, our observed statistic was 5.75 standard deviations above the mean.
- In the second, our observed statistic was 2.61 standard deviations above the mean.
- Both of these are very strong, but we have stronger evidence against the null in the first.

Guidelines for strength of evidence

- If a standardized statistic is below -2 or above 2, we have strong evidence against the null.

Standardized Statistic	Evidence Against Null
between -1.5 and 1.5	not much
below -1.5 or above 1.5	moderate
below -2 or above 2	strong
below -3 or above 3	very strong

11. A quick note on 1-sided versus 2-sided tests.

- On my exams, I will tell you explicitly whether to do a 1 or 2 sided test.
- On hw problems, you might have to decide whether to do a 1-sided or 2-sided test.
- With the hw, if in the problem you are given that you are only looking for evidence in one direction as evidence against the null hypothesis, then you do a 1-sided test. If you are looking for *any* difference in proportions as evidence against the null hypothesis, then do a 2-sided test.

Two-Sided Tests

- The change to the alternative hypothesis affects how we compute the p-value.
- Remember that the p-value is the probability (assuming the null hypothesis is true) of obtaining a proportion that is equal to or **more extreme** than the observed statistic
- In a *two-sided test*, **more extreme** goes in both directions.