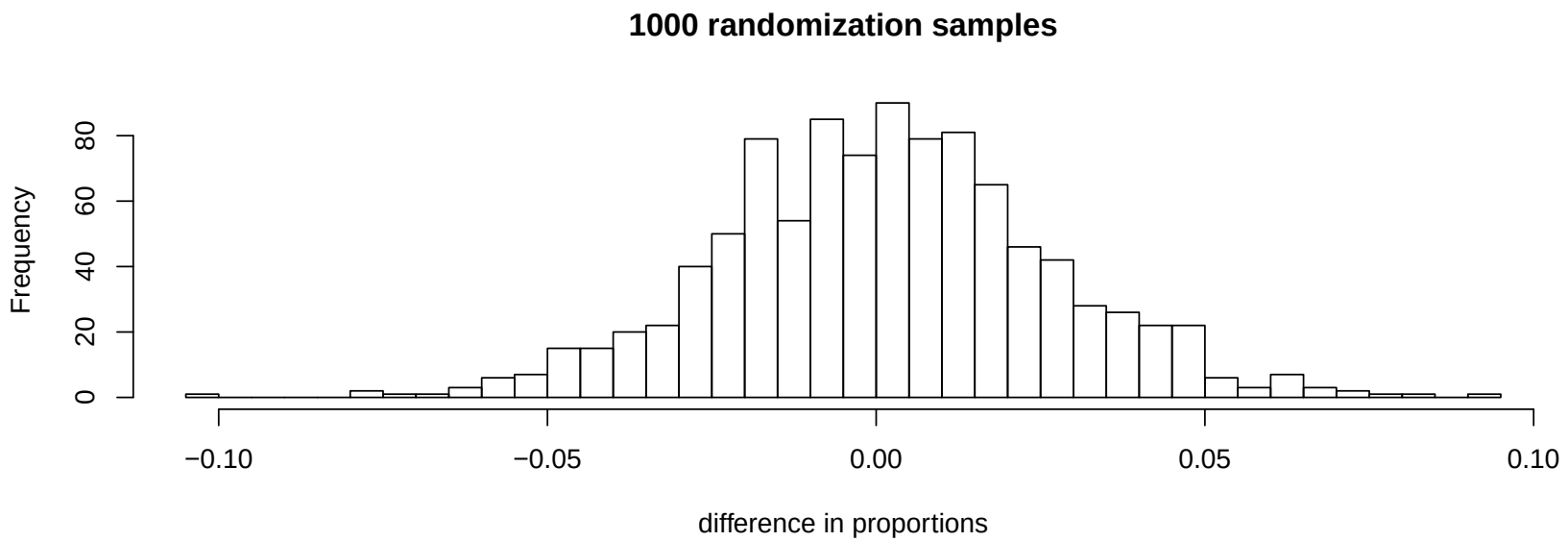
Classical Inference for Proportions

Ryan Miller

Normal Distributions

In our work with distributions, you may have noticed the prevalence of a certain shape:



- Nearly all of the bootstrap or randomization distributions we've seen were symmetric and bell-shaped
- This is not a coincidence, it's actually a foundation of classical statistics

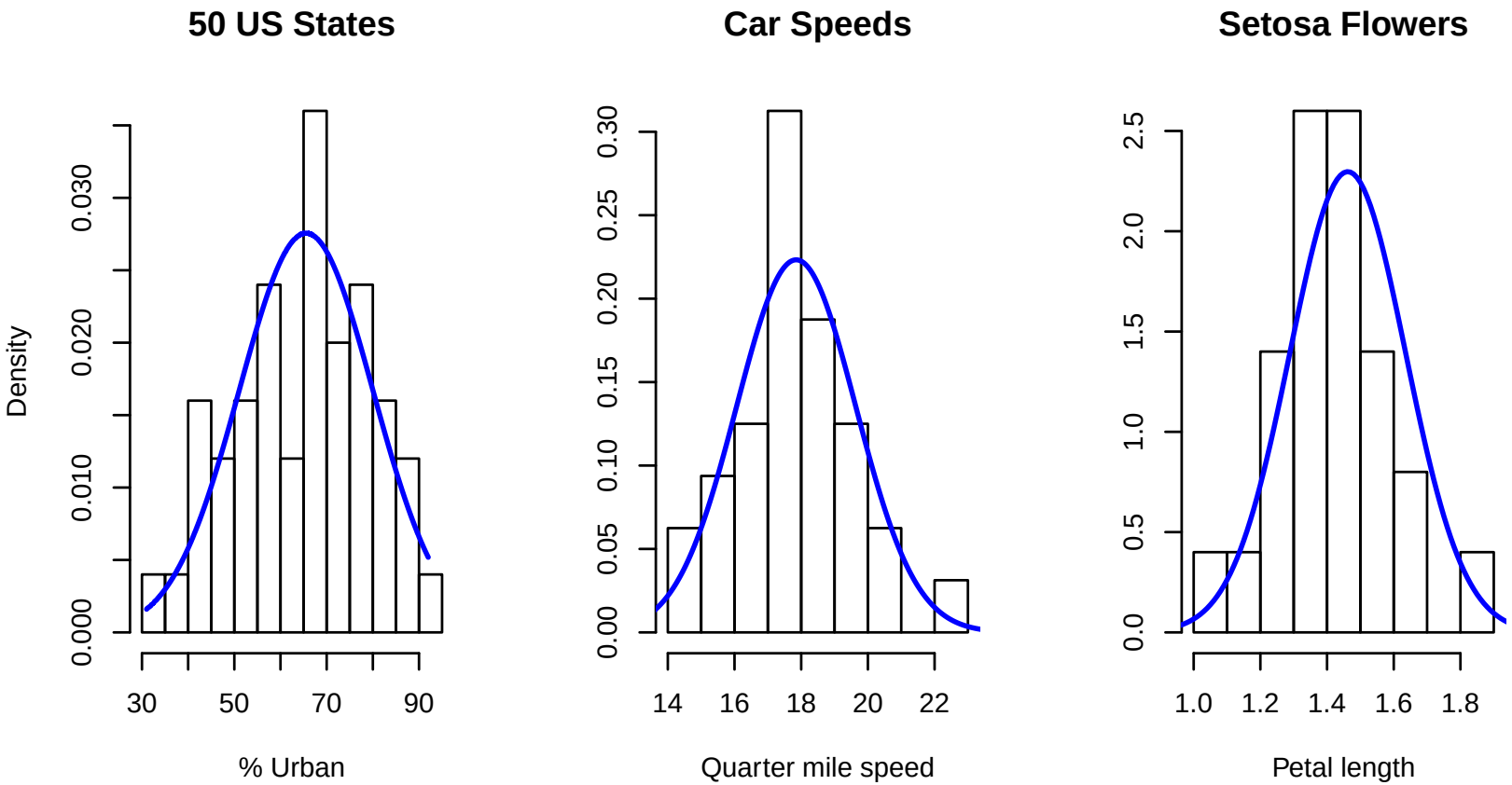
Normal Distributions

- These symmetric, bell-shaped distributions can be characterized by the curve:

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

- This curve defines the **Normal Distribution**
  - μ is the center (mean) of the distribution
  - σ is the standard deviation of the distribution
  - Denote normal distributions using $N(\mu, \sigma)$, for example: N(3,1)
- You aren't expected to know the formula of the normal curve, but you should know that it depends upon μ and σ

Normal Approximation (Examples)



Normal Approximation

Normal approximation provides another way to estimate:

1. The *sampling distribution* (we had been using the bootstrap distribution)
2. The *null distribution* (we had been using the randomization distribution)

To do so we need to mathematically figure out the mean and standard deviation of these distributions for the scenario we are analyzing (since the normal curve is entirely defined by μ and σ)

Normal Approximation

We can pretty easily determine the mean of sampling and null distributions:

- The mean of the *sampling distribution* is the observed sample statistic ($\bar{x}$ or $\hat{p}$ or $\hat{p}_1 - \hat{p}_2$ or $\bar{x}_1 - \bar{x}_2$)
- The mean of the *null distribution* is the value specified by H_0 (most often 0)

But what about the standard deviation of these distributions? (which we call the *standard error* to avoid confusion)

The Central Limit Theorem

We won't get into the details, but the **Central Limit Theorem** (CLT), one of the most well-known results in the history of statistics, provides the necessary information for situations (notably those involving means and proportions)

For a single proportion, provided the sample size, n, is sufficiently large:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

In words, the sample proportion $\hat{p}$ follows a normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$

The Central Limit Theorem (Single Proportion)

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

How do we make use of this result when we don't know p?

- We should use the value that *best reflects* our goal
- If we are constructing a confidence interval, we should use $\hat{p}$, because it is our best estimate of p
- If we are conducting a hypothesis test, we should use p_0, the value *specified in the null hypothesis*

The Standard Normal Distribution

Under the null hypothesis, Central Limit Theorem suggests:

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

Because there are so many different normal curves, it is convenient to *standardize*:

$$z_{\text{test}} = \frac{\hat{p} - p_0}{\sqrt{(p_0(1-p_0))/n}} \sim N(0,1)$$

- z_{test} is the z-score of the observed proportion amongst the possible proportions that are possible when the null hypothesis is true
- z_{test} is called a **test statistic**, and has been used historically to calculate p-values via the **standard normal distribution**, $N(0,1)$

Example #1

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months

1. Find the 95% confidence interval estimate for the proportion of all babies born in the US at 25 weeks gestation that will survive at least 6 months (assuming babies born at Johns Hopkins Hospital are representative of this population)
2. An article on Wikipedia suggests 70% of babies born at 25 weeks gestation survive. Use the Johns Hopkins study to evaluate this claim.

Example #1 (Solution)

1. $\hat{p} = 31/39 = 0.795$, using the normal approximation provided by CLT, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.795(1-0.795)}{39}} = 0.065$; this suggests the 95% CI: $0.795 \pm 2 * 0.065 = (0.668, 0.922)$
2. For this we want to test $H_0 : p = 0.7$ versus $H_A : p \neq 0.7$, this suggests the null distribution is $N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$, which is $N(0.7, 0.073)$. Using StatKey to locate our sample proportion ($\hat{p} = 0.795$) on this distribution, we find a two-sided p-value of 0.194. We conclude the Johns Hopkins study is consistent with this claim

Note: We could have used the test statistic: $z_{\text{test}} = \frac{0.795-0.7}{0.073} = 1.3$ and found the same p-value using the standard normal distribution.

Example #2

The John Hopkins researchers also investigated gestations earlier than 25 weeks. At 22 weeks gestation, they found 0/29 babies in their hospital's records survived at least 6 months

1. Use the normal approximation approach to construct a 95% confidence interval for the proportion of babies born at 22 weeks gestation who will survive. Would you report this interval?
2. Use Minitab to construct a 95% confidence interval for this proportion by selecting "Stat -> Basic Statistics -> 1 Proportion" and then selecting "Summarized data" from the drop-down menu. Minitab uses an exact approach rather a normal approximation. Is this interval more reasonable that the one you calculated in question 1? Why do you think that is?

How Large is "Sufficiently Large"?

The normal approximation suggested by Central Limit Theorem only works well when n is sufficiently large

- For a single proportion, sufficiently large depends upon the value of p
- A common rule of thumb for whether or not the approximation will be reasonable is to check:

1. $n * p \geq 10$
2. $n * (1 - p) \geq 10$

If either of these conditions are not met, we should consider another approach to estimating the sampling/null distribution (ie: bootstrapping or randomization)

The Exact Binomial Test

- Prior to modern computing, the normal approximation was one of the only ways to perform hypothesis tests or construct confidence intervals
- The **exact binomial test** is the modern standard in tests (or confidence intervals) on a single proportion since it does not rely on an approximation of the null (or sampling) distribution
- By default, Minitab uses the exact binomial test for 1 proportion tests, which may lead to slightly better results than by-hand calculations that rely on the normal approximation
- We won't cover the details of this test, but you should know that it exists (and why it's used)

The Central Limit Theorem (Difference in Proportions)

For a difference in proportions, provided n_1 and n_2 are sufficiently large, CLT suggests:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

In this scenario, checking the conditions of this normal approximation is a little more tedious:

1. $n_1 * p_1 \geq 10$
2. $n_1 * (1 - p_1) \geq 10$
3. $n_2 * p_2 \geq 10$
4. $n_2 * (1 - p_2) \geq 10$

Example #3

A common stereotype is that medical doctors have terrible handwriting. If true, this might contribute to errors in medical prescriptions that rely on hand-written forms. A 2010 study took two groups of doctors that had similar error rates before the study and randomly assigned half of them to electronic prescription form, while the other half continued using written prescriptions. After 1 year, the error rate of each group was recorded:

	Error	Non-errors	Total
Electronic	254	3594	3848
Hand-written	1478	2370	3848

- Find the 95% confidence interval for difference proportions (e-prescriptions minus hand-written prescriptions)

Example #3 (Solution)

1. $\hat{p}_e - \hat{p}_{hw} = 254/2848 - 1478/3848 = 0.066 - 0.384 = -0.318$; while $SE = \sqrt{\frac{0.066(1-0.066)}{3848} + \frac{0.384(1-0.384)}{3848}} = 0.017$; thus the 95% Confidence Interval is given by:

$$-0.318 \pm 0.017 = (-0.335, -0.301)$$

We can conclude that the reduction in errors is very large.

Pooled Proportions

- The standard error when hypothesis testing is a little tricky when dealing with a difference in proportions
- Hypothesis testing puts us in a hypothetical world where the null hypothesis is true, and the null distribution reflects that
- Because the null hypothesis stipulates that $p_1 = p_2$, the standard error of the null distribution requires use of a **pooled proportion**:

$$\hat{p}_{1+2} = \frac{\text{Frequency in } n_1 + \text{Frequency in } n_2}{n_1 + n_2}$$

In difference of proportions tests we use $\hat{p}_{1+2}$ in place of *both* p_1 and p_2 in the standard error calculation

Example #4

For the hand-written prescriptions example, test the hypothesis that electronic and hand-written prescriptions result in equal amounts of errors:

	Error	Non-errors	Total
Electronic	254	3594	3848
Hand-written	1478	2370	3848

Example #4 (solution)

$$\hat{p}_{1+2} = \frac{\text{Frequency in } n_1 + \text{Frequency in } n_2}{n_1 + n_2} = \frac{254 + 1479}{3848 + 3848} = 0.225$$

$$z_{\text{test}} = \frac{-0.318 - 0}{\sqrt{\frac{0.225(1-0.225)}{3848} + \frac{0.225(1-0.225)}{3848}}} = -33.4$$

The p-value is near zero

Conclusion

Right now you should...

1. Understand how a normal approximation can be used to describe the sampling and null distributions
2. Know how to perform hypothesis tests or construct confidence intervals for scenarios involving one proportion, or a difference in proportions, using the normal approximation
3. Be aware of the assumptions required for the normal approximation be reasonable
4. Be aware of the exact binomial test

These notes cover Ch 5 and parts of Ch 6 of our textbook, I encourage you to read through those sections and their examples