Hypothesis Testing - Categorical Data

Ryan Miller

Introduction

Previously, we introduced **hypothesis testing**, a general statistical approach used to measure the compatibility of sample data with a null model. Hypothesis testing is a multi-step process:

1) Specify an appropriate null model
2) Find the corresponding null distribution
3) Use the null distribution to calculate a p-value
4) Use the p-value to make a decision regarding the plausibility of the null model

Our introduction glossed over Step #2 (and to some degree Step #1), which are arguably the most challenging aspects of hypothesis testing. This presentation will focus on those steps for *categorical data*.

Transmission Disequilibrium

- In genetics, a common question is whether a certain gene is linked to a certain trait
  - This is a challenging question, as humans have over 30,000 genes and countless traits that likely depend upon numerous factors (gene-environment interactions, etc.)

Transmission Disequilibrium

- In genetics, a common question is whether a certain gene is linked to a certain trait
  - This is a challenging question, as humans have over 30,000 genes and countless traits that likely depend upon numerous factors (gene-environment interactions, etc.)
- One approach is to find child-parent pairs where the child has the trait of interest and the parent is *heterozygous* for the gene of interest (ie: they have one copy of each version of the gene)

Transmission Disequilibrium

- In genetics, a common question is whether a certain gene is linked to a certain trait
  - This is a challenging question, as humans have over 30,000 genes and countless traits that likely depend upon numerous factors (gene-environment interactions, etc.)
- One approach is to find child-parent pairs where the child has the trait of interest and the parent is *heterozygous* for the gene of interest (ie: they have one copy of each version of the gene)
  - Under normal circumstances, the parent is equally likely to pass on either version of the gene
  - Thus, if a gene is unrelated to trait, we'd expect 50% of children with the trait to have either version of the gene

Type I Diabetes - Introduction

- A study published in *Genetic Epidemiology* collected data on 124 children with Type 1 diabetes whose parent was heterozygous for the gene FP'1
  - Among these 124 children, 78 had the "class 1" version of FP'1, while 46 did not
- Is this sufficient evidence to link FP'1 with Type 1 diabetes?

Type I Diabetes - Hypothesis Testing

- **Step #1**: Specify an appropriate null model

Type I Diabetes - Hypothesis Testing

- **Step #1**: Specify an appropriate null model
  - The proportion of *all children with Type 1 diabetes and a heterozygous parent* that have the "class 1" version of FP'1 is 50%, expressed statistically as $H_0 : p = 0.5$

Type I Diabetes - Hypothesis Testing

- **Step #1**: Specify an appropriate null model
  - The proportion of *all children with Type 1 diabetes and a heterozygous parent* that have the "class 1" version of FP'1 is 50%, expressed statistically as $H_0 : p = 0.5$
- **Step #2**: Find the corresponding null distribution

Type I Diabetes - Hypothesis Testing

- **Step #1**: Specify an appropriate null model
  - The proportion of *all children with Type 1 diabetes and a heterozygous parent* that have the "class 1" version of FP'1 is 50%, expressed statistically as $H_0 : p = 0.5$
- **Step #2**: Find the corresponding null distribution
  - Option #1: Simulation
  - Option #2: Binomial distribution
  - Option #3: Normal distribution

The Simulation Approach

- Simulation is a general approach that can be used to generate a null distribution in a wide variety of scenarios
  - Proper use of this approach preserves as many aspects of the data as possible while satisfying the null model

The Simulation Approach

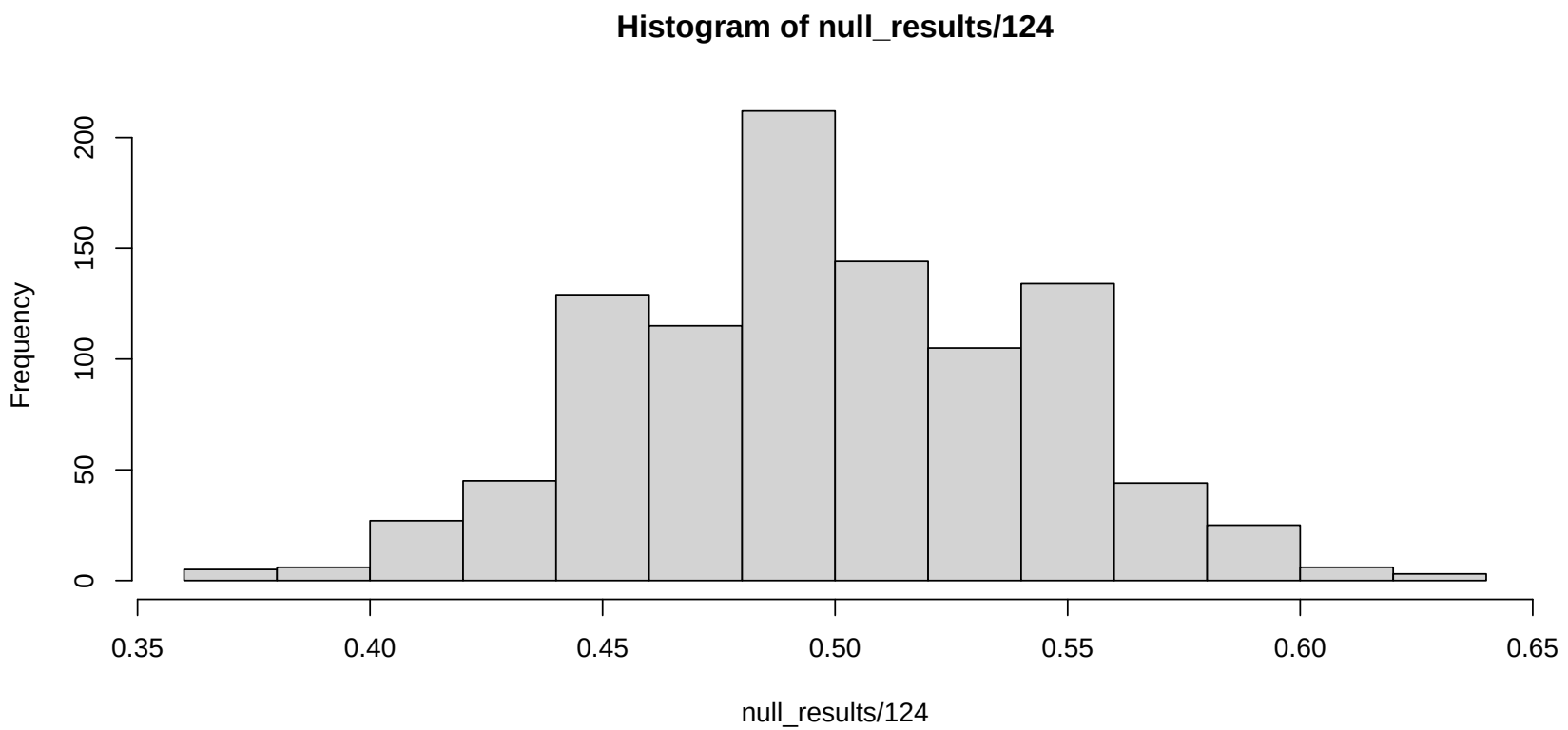
- Simulation is a general approach that can be used to generate a null distribution in a wide variety of scenarios
  - Proper use of this approach preserves as many aspects of the data as possible while satisfying the null model
- In our example, the observed data are 78 of 124 children ($\hat{p} = 0.63$) with the "class 1" gene and the null model is $H_0 : p = 0.5$, how might you use simulation to generate a null distribution for this scenario?

The Simulation Approach

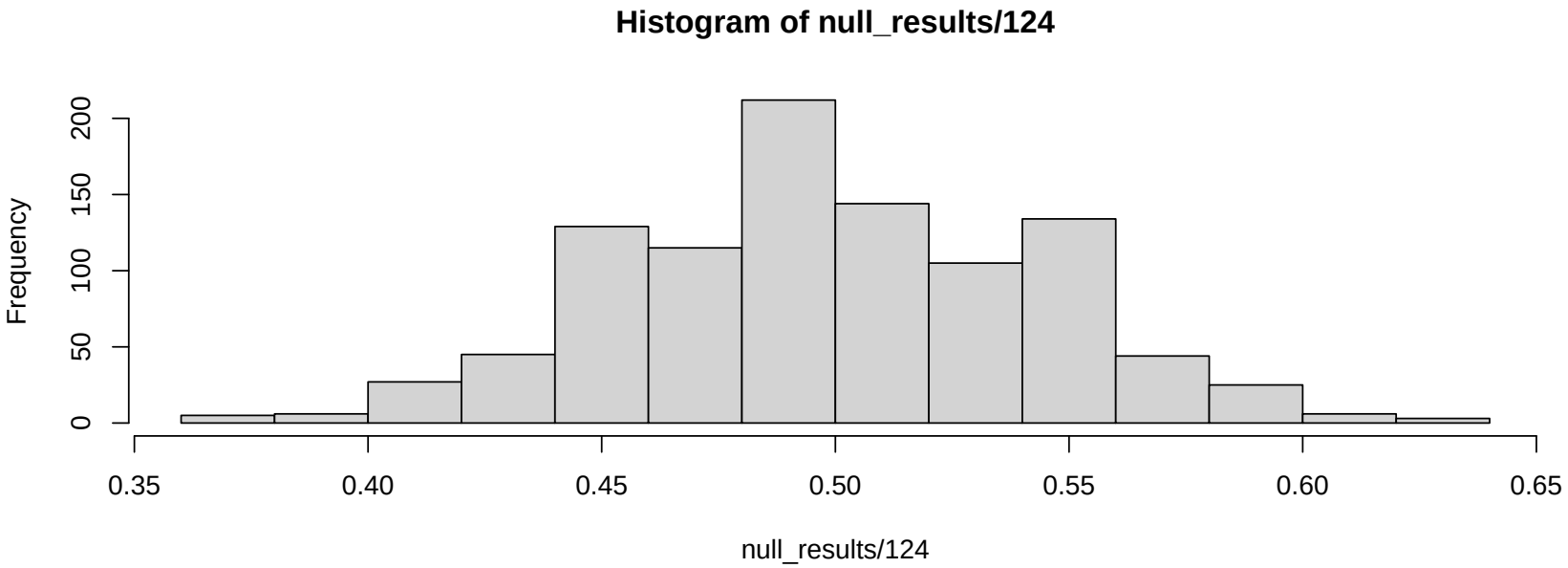
- Coin flips provide a suitable null model for random process of each child having (or not having) the "class 1" gene
- Thus, the proportion of "heads" across a large number of sets of 124 coin flips can be used as the null distribution
  - The `rbinom()` function allows us to simulate sets of coin flips, can you use it to generate a null distribution and find the p-value of this test?

The Simulation Apporach (solution)

```r
### Generate 1000 sets of n = 124 coin-flips
set.seed(123)
null_results <- rbinom(1000, size = 124, prob = 0.5)
hist(null_results/124, breaks = 15)
```



The Simulation Approach (solution)




```
## Calculate the two-sided p-value
upper_tail <- sum(null_results/124 >= 78/124)/1000
2*upper_tail
```

```
## [1] 0.004
```

Simulation - Advantages/Disadvantages

Advantages:

- Extremely general
  - Can be used for lots of different statistical summaries, not just proportions/averages

Simulation - Advantages/Disadvantages

Advantages:

- Extremely general
  - Can be used for lots of different statistical summaries, not just proportions/averages

Disadvantages:

- Approximate
  - Different simulation seeds will result in slightly different null distributions
- Computationally involved
  - Lots of simulations needed to get a stable null distribution
- Requires creativity (sometimes)
  - Its not always easy to determine how to generate data under the null model while preserving all of the key aspects of the original study

The Binomial Distribution Approach

- You may have recognized that *simulating* sets of 124 coin flips is unnecessary
  - The **binomial distribution** can be used to exact probabilities of each possible result that could arise from this null model
  - How might you find the *two-sided* p-value (for this scenario) using the binomial distribution (via the `pbinom()` function)?

The Binomial Distribution Approach

- You may have recognized that *simulating* sets of 124 coin flips is unnecessary
  - The **binomial distribution** can be used to exact probabilities of each possible result that could arise from this null model
  - How might you find the *two-sided* p-value (for this scenario) using the binomial distribution (via the `pbinom()` function)?

```
upper_tail <- pbinom(78 - 1, size = 124, prob = .5,
                     lower.tail = FALSE)
2*upper_tail
```

```
## [1] 0.005161225
```

```
binom.test(78, n = 124)$p.value
```

```
## [1] 0.005161225
```

Comments on the use of `pbinom()`

- It might seem odd to subtract 1 from the number of "successes" in `pbinom`
  - But remember, the p-value calculation involves every possible outcome *at least as extreme* as the observed outcome
- The argument `lower.tail = FALSE` tells `pbinom()` to calculate $P(X > x)$, which doesn't include $P(X = x)$
  - Thus, subtracting 1 is a quick fix that will start the summation at the observed number of successes

Binomial - Advantages/Disadvanteges

Advantages:

- Exact
  - Eliminates the uncertainty/randomness involved in simulation

Binomial - Advantages/Disadvanteges

Advantages:

- Exact
  - Eliminates the uncertainty/randomness involved in simulation

Disadvantages:

- Computationally involved (somewhat)
  - Behind the scenes `R` is summing a lot of different binomial probabilities
- Not generalizable
  - Only useful when analyzing a single proportion

The Normal Distribution Approach (Z-Test)

- A final approach uses *Central Limit Theorem* to come up with a Normal model for the null distribution

The Normal Distribution Approach (Z-Test)

- A final approach uses *Central Limit Theorem* to come up with a Normal model for the null distribution
- Recall that if $H_0 : p = 0.5$ and $n = 124$, then CLT suggests:

$$\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{124}}\right)$$

The Normal Distribution Approach (Z-Test)

- A final approach uses *Central Limit Theorem* to come up with a Normal model for the null distribution
- Recall that if $H_0 : p = 0.5$ and $n = 124$, then CLT suggests:

$$\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{124}}\right)$$

Based upon this distribution, might consider the standardized Z-value of our observed sample proportion:

$$Z = \frac{\hat{p}-p_0}{SE} = \frac{0.63-0.50}{\sqrt{0.5(1-0.5)/124}} = 2.9$$

So, our sample proportion is 2.9 standard deviations higher than what we'd expect under the null model, let's formalize how unusual this is with a p-value

The Normal Distribution Approach (Z-Test)

The Z-test in R:

```
Z <- 2.9
upper_tail <- pnorm(Z, mean = 0, sd = 1,
                    lower.tail = FALSE)
2*upper_tail
```

```
## [1] 0.003731627
```

The Normal Distribution Approach (Z-Test)

The Z-test in R:

```
Z <- 2.9
upper_tail <- pnorm(Z, mean = 0, sd = 1,
                    lower.tail = FALSE)
2*upper_tail
```

```
## [1] 0.003731627
```

Recognize we could do the same test without any standardization:

```
phat <- 0.63
upper_tail <- pnorm(phat, mean = 0.5,
                    sd = sqrt(.5*.5/124),
                    lower.tail = FALSE)
2*upper_tail
```

```
## [1] 0.003788718
```

Z-Test - Advantages/Disadvanteges

Advantages:

- Generalizable
  - Can be used for proportions/averages and linear combinations of them
- Computationally easy
  - Statisticians could easily use the Z-test prior to modern computing

Z-Test - Advantages/Disadvanteges

Advantages:

- Generalizable
  - Can be used for proportions/averages and linear combinations of them
- Computationally easy
  - Statisticians could easily use the Z-test prior to modern computing

Disadvantages:

- Approximate
  - Uses a large-sample theoretical result that might not be accurate for finite samples

Comments on these Three Approaches

- In a real statistical analysis you'd chose only one of these three approaches to use/report
- Notice the p-values were very similar:
  - Simulation yielded $p = 0.0040$
  - The exact binomial test yielded $p = 0.0051$
  - The Z-test yielded $p = 0.0038$
- Regardless of the statistical test, the conclusion is that "class 1" gene seems related to Type I diabetes

Comments on these Three Approaches

- In a real statistical analysis you'd chose only one of these three approaches to use/report
- Notice the p-values were very similar:
  - Simulation yielded $p = 0.0040$
  - The exact binomial test yielded $p = 0.0051$
  - The Z-test yielded $p = 0.0038$
- Regardless of the statistical test, the conclusion is that "class 1" gene seems related to Type I diabetes
- These similarities shouldn't be surprising
  - Simulation is an approximation of the exact binomial
  - The Central Limit Theorem normal result will be reasonably accurate when $n * p_0 \geq 10$ and $n * (1 - p_0) \geq 10$

Non-Binary Categorical Variables

- In the previous example, we were able to reduce the scenario to a test on a *single proportion*, the proportion of child-adult pairs where the child had the "class 1"
  - However, not every application involving categorical data can be summarized using a single proportion

Non-Binary Categorical Variables

- In the previous example, we were able to reduce the scenario to a test on a *single proportion*, the proportion of child-adult pairs where the child had the "class 1"
  - However, not every application involving categorical data can be summarized using a single proportion
- Below is the distribution of correct answers for 400 randomly selected AP Stats Exam questions:

A	B	C	D	E
85	90	79	78	68

Non-Binary Categorical Variables

- In the previous example, we were able to reduce the scenario to a test on a *single proportion*, the proportion of child-adult pairs where the child had the "class 1"
  - However, not every application involving categorical data can be summarized using a single proportion
- Below is the distribution of correct answers for 400 randomly selected AP Stats Exam questions:

A	B	C	D	E
85	90	79	78	68

1. If correct AP Exam answers are truly random, what proportion of these answers would you expect to be "A's"?
2. Would a z-test on the proportion of "A" answers in this sample provide enough evidence to decide if AP Exam's answers are randomly distributed?

AP Exam Answers

- First, you'd expect (on average) 1/5 of answers in a random sample to be "A"
- Second, a z-test based upon $\hat{p}_A$ and the null proportion of 0.2 would *not* be sufficient - even if the proportion of "A" answers aligns with what we'd expect, the proportions of other answers might not
  - Thus, you need *four* proportions to describe these five outcomes (why not a fifth?)

AP Exam Answers

- Four different hypothesis tests seems like overkill for such a simple frequency table
- Instead, it's more sensible to do a single test of the hypotheses:

$$H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$$

$$H_A : p_i \neq 0.2 \text{ for at least one } i \in \{A, B, C, D, E\}$$

- To see if we can come up with a test of this hypothesis, we'll begin by assuming the null hypothesis is true
  - So, had we randomly sampled 400 AP questions under this null hypothesis, what frequencies would you expect for each answer choice?

Expected Counts

- The most likely frequencies under the null hypothesis are called the **expected counts**
- For the AP Exam data, they are:

A	B	C	D	E
80	80	80	80	80

Expected Counts

- The most likely frequencies under the null hypothesis are called the **expected counts**
- For the AP Exam data, they are:

A	B	C	D	E
80	80	80	80	80

- In general, we calculate the expected counts for each of i possible categories as:

$$\text{Expected}_i = n * p_i$$

- This is easy with the AP Exam data because the proportions, p_i, are the same for every category (under the null hypothesis)
  - Note that this won't always be the case

Chi-Square Testing

- To evaluate $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$ we can compare the **observed counts** with those we'd expect if the null hypothesis was true:

Answer	A	B	C	D	E
Expected Count	80	80	80	80	80
Observed Count	85	90	79	78	68

- In this framework, we seek to answer the question: "If the null hypothesis is true, do the observed counts deviate from the expected counts by more than we'd reasonably expect due to random chance"

Chi-Square Testing

- To evaluate $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$ we can compare the **observed counts** with those we'd expect if the null hypothesis was true:

Answer	A	B	C	D	E
Expected Count	80	80	80	80	80
Observed Count	85	90	79	78	68

- In this framework, we seek to answer the question: "If the null hypothesis is true, do the observed counts deviate from the expected counts by more than we'd reasonably expect due to random chance"
- Think about how you'd summarize the distance between the observed and expected counts?
  - Is the distance between 79 and 80 the same as the distance between 80 and 79?
  - Is it the same as the distance between 4 and 5?

The Chi-Square Statistic

- We evaluate H_0 (as previously defined) using the **Chi-Square Test**, the test statistic is given below:

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

The Chi-Square Statistic

- We evaluate H_0 (as previously defined) using the **Chi-Square Test**, the test statistic is given below:

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

- Like other test statistics, it compares the observed data to what we'd expect under the null hypothesis, while standardizing the differences
  - Different is that we must sum over the variable's i categories
  - Also different is that the numerator is squared so that positive and negative deviations won't cancel each other out

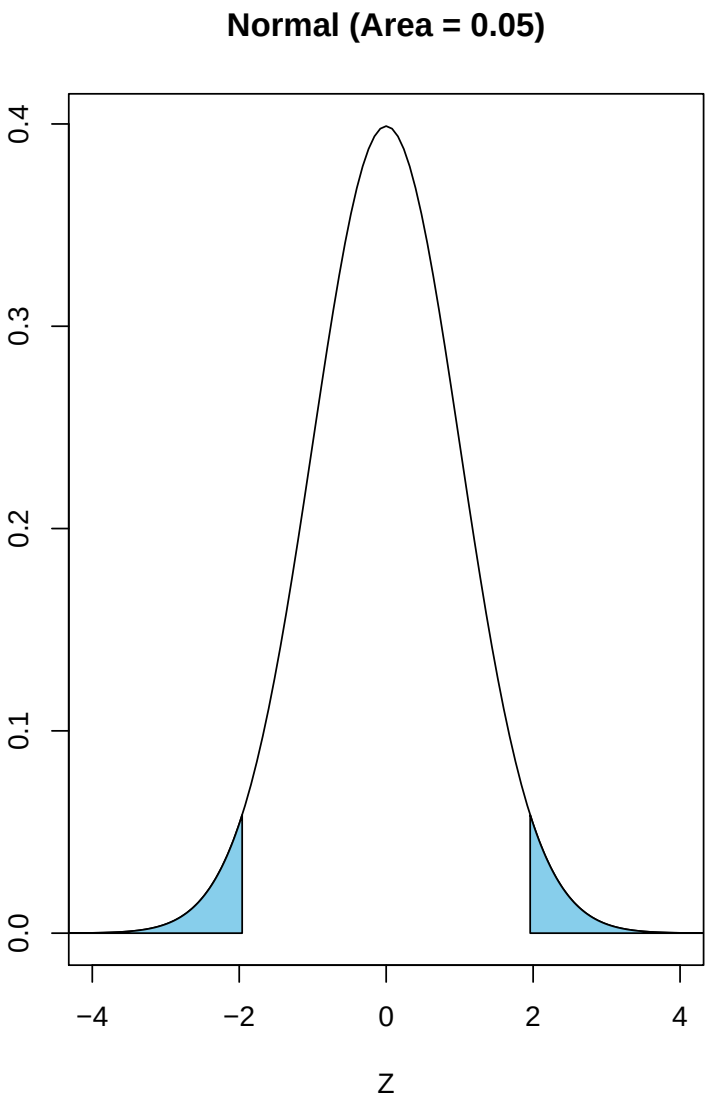
The Chi-Square Distribution

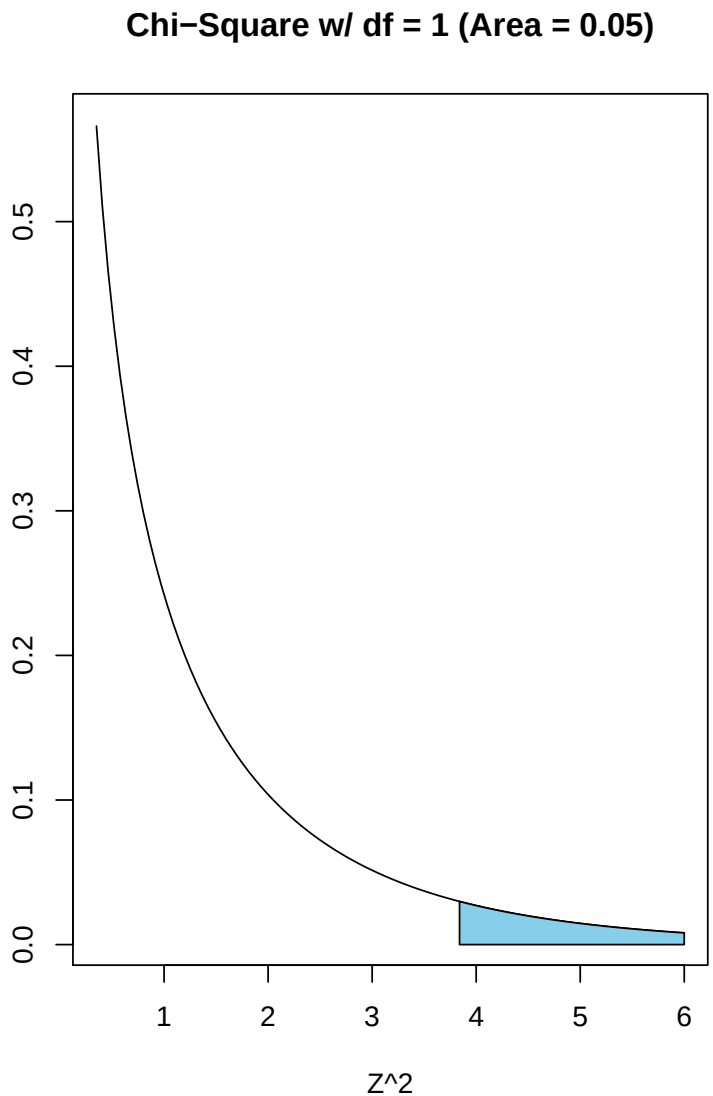
- The Chi-Square test requires us to learn a new distribution, the χ^2 curve
- Fortunately, the χ^2 distribution is related to the standard normal distribution

The Chi-Square Distribution

- The Chi-Square test requires us to learn a new distribution, the χ^2 curve
- Fortunately, the χ^2 distribution is related to the standard normal distribution
  - Suppose we generated lots of data from the standard normal distribution, the histogram of these data would look like the normal curve
  - Now suppose we took these observations and squared them, this histogram looks like the χ^2 curve (with $df = 1$)

The Chi-Square Distribution





The Chi-Square Distribution

- The relationship between the χ^2 distribution and the normal distribution is clearly illustrated by looking at the test statistic for the Z-test:

$$z_{\text{test}} = \frac{\text{observed} - \text{null value}}{SE}$$

$$z^2_{\text{test}} = \frac{(\text{observed} - \text{null value})^2}{SE^2}$$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected count})^2}{\text{expected count}}$$

The Chi-Square Distribution

- The relationship between the χ^2 distribution and the normal distribution is clearly illustrated by looking at the test statistic for the Z-test:

$$z_{\text{test}} = \frac{\text{observed} - \text{null value}}{SE}$$

$$z^2_{\text{test}} = \frac{(\text{observed} - \text{null value})^2}{SE^2}$$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected count})^2}{\text{expected count}}$$

- Essentially, the χ^2 test is just a squared version of the z-test
  - This makes the test naturally two-sided, even though we only calculate p-values using the right tail of the χ^2 curve
  - Under H_0, the SE of each category count is approximately the square root of the expected value of that count

Degrees of Freedom

- There are many different χ^2 distributions depending upon how many unique categories we must sum over
- Letting k denote the number of categories of a categorical variable, the χ^2 test statistic for testing a single categorical variable has $k - 1$ degrees of freedom
  - This is because the category proportions are constrained to sum to 1

Degrees of Freedom

- There are many different χ^2 distributions depending upon how many unique categories we must sum over
- Letting k denote the number of categories of a categorical variable, the χ^2 test statistic for testing a single categorical variable has $k - 1$ degrees of freedom
  - This is because the category proportions are constrained to sum to 1
  - The mean and standard deviation of the χ^2 curve both depend upon its degrees of freedom
  - We can use `pchsq()` to calculate areas under the various different χ^2 curves in `R`

Performing the Chi-Square Test (Quick Example)

1. State the Null Hypothesis:
   $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$

Performing the Chi-Square Test (Quick Example)

1. State the Null Hypothesis:
   $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$
2. Calculate the expected counts under the null:

$$E_A = 0.2 * 400 = 80, \; E_B = 0.2 * 400 = 80, \; \ldots$$

Performing the Chi-Square Test (Quick Example)

1. State the Null Hypothesis:
   $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$
2. Calculate the expected counts under the null:

$$\mathrm{E}_A = 0.2 * 400 = 80, \ \mathrm{E}_B = 0.2 * 400 = 80, \ \ldots$$

3. Calculate the χ^2 test statistic:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\ &= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\ &= 3.425\end{aligned}$$

Performing the Chi-Square Test (Quick Example)

1. State the Null Hypothesis:
   $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$
2. Calculate the expected counts under the null:

$$\mathrm{E}_A = 0.2 * 400 = 80, \ \mathrm{E}_B = 0.2 * 400 = 80, \ \ldots$$

3. Calculate the χ^2 test statistic:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\ &= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\ &= 3.425\end{aligned}$$

4. Locate the χ^2 test statistic on the χ^2 distribution with $k-1$ degrees of freedom to find the p-value: $p = 0.49$

Chi-Squared Test in R

Using the X^2 test statistic:

```
pchisq(3.425, df = 4, lower.tail = FALSE)
```

```
## [1] 0.4893735
```

Using the sample data directly:

```
observed <- c(85, 90, 79, 78, 68)
chisq.test(observed, p = c(.2, .2, .2, .2, .2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 3.425, df = 4, p-value = 0.4894
```

Another Example

- Pools of prospective jurors are supposed to be drawn at random from the eligible adults in that community
  - The American Civil Liberties Union (ACLU) studied the racial composition of the jury pools for a sample of 10 trials in Alameda County, California
  - The 1453 individuals included in these jury pools are summarized below. For comparison, census data describing the eligible jurors in the county is included

Another Example

- Pools of prospective jurors are supposed to be drawn at random from the eligible adults in that community
  - The American Civil Liberties Union (ACLU) studied the racial composition of the jury pools for a sample of 10 trials in Alameda County, California
  - The 1453 individuals included in these jury pools are summarized below. For comparison, census data describing the eligible jurors in the county is included

Race/Ethnicity	White	Black	Hispanic	Asian	Other
Number in jury pools	780	117	114	384	58
Census percentage	54%	18%	12%	15%	1%

Directions: Use a Chi-Squared test to determine whether the racial composition of jury pools in Alameda County differs from what is expected based upon the census

Example - Solution

$$H_0 : p_w = 0.54, p_b = 0.18, p_h = 0.12, p_a = 0.15, p_o = 0.01$$

$$H_A : \text{At least one } p_i \text{ differs from those specified in } H_0$$

Race/Ethnicity	White	Black	Hispanic	Asian	Other
Observed Count	780	117	114	384	58
Expected Count	1453*.54 = 784.6	1453*.18 = 261.5	1453*.12 = 174.4	1453*.15 = 218	1453*.01 = 14.5

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

$$= \frac{(780 - 784.6)^2}{784.6} + \frac{(117 - 261.5)^2}{261.5} + \frac{(114 - 174.4)^2}{174.4} + \frac{(384 - 218)^2}{218} + \frac{(58 - 14.5)^2}{14.5}$$

$$= 357$$

- The p-value of this test is near zero and provides strong evidence that the jury pools don't match the racial proportions of the census
- Comparing the observed vs. expected counts, it appears that Blacks and Hispanics are underrepresented while Asians and Others are overrepresented in the jury pools

Summary

- This presentation covered methods for hypothesis tests involving a *single* categorical variable
- If we could summarize the variable using a single proportion, we could use:
  - Simulation
  - Exact Binomal
  - Z-test
- If summarizing the variable requires multiple proportions, we might use:
  - Simulation (not shown)
  - Exact Multinomial (not shown)
  - Chi-Squared Goodness of Fit test
- The next presentation will cover methods for testing *relationships* between *two* categorical variables