Chi-Squared Tests

Prof. Wells

STA 209, 5/3/23

Outline

In this lecture, we will. . .

- Determine whether data follows a certain distribution
- Investigate the chi-squared distribution.
- Use the chi-squared statistic to determine whether two variables are independent

Section 1

The Chi-Squared Test for Goodness of Fit

Inference for Categorical Variables

Suppose we want to investigate either 1 categorical variable or the relationship between 2 categorical variables.

Inference for Categorical Variables

Suppose we want to investigate either 1 categorical variable or the relationship between 2 categorical variables.

- If the single variable has just 2 levels, we can consider the proportion p for one level

Inference for Categorical Variables

Suppose we want to investigate either 1 categorical variable or the relationship between 2 categorical variables.

- If the single variable has just 2 levels, we can consider the proportion p for one level
- If both response and explanatory variables have 2 levels, we can consider the difference in proportions $p_1 - p_2$.

Inference for Categorical Variables

Suppose we want to investigate either 1 categorical variable or the relationship between 2 categorical variables.

- If the single variable has just 2 levels, we can consider the proportion p for one level
- If both response and explanatory variables have 2 levels, we can consider the difference in proportions $p_1 - p_2$.

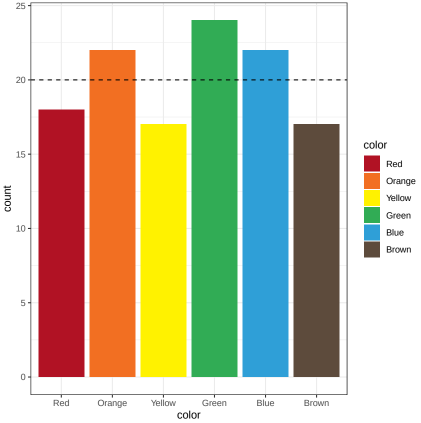
What can we do if one or both the variables are categorical with more than 2 levels?

M&Ms

Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:

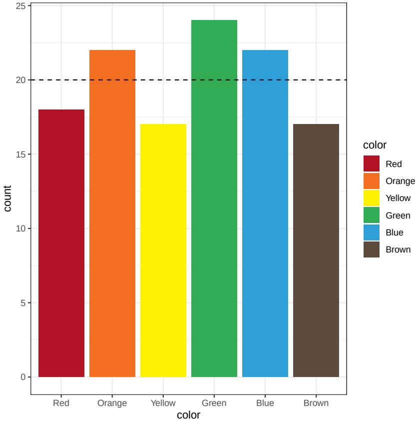
M&Ms

Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:



M&Ms

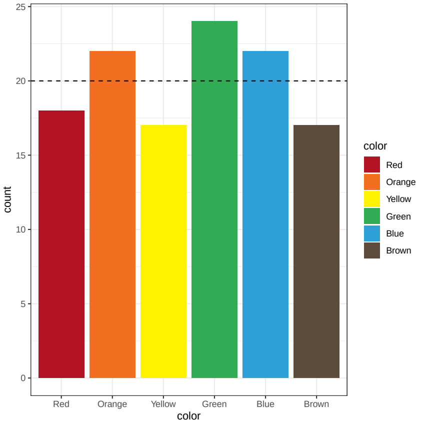
Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:



- Note that Green M&Ms exceed by the expected count by 20%.

M&Ms

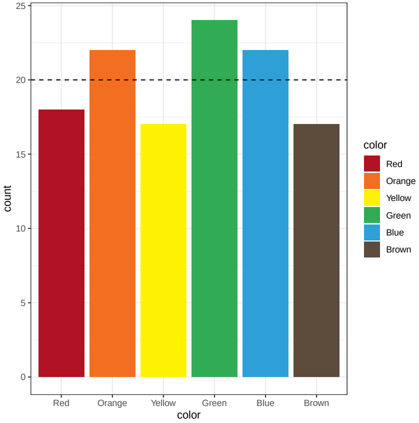
Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:



- Note that Green M&Ms exceed by the expected count by 20%.
- Does this give good evidence that M&M colors appear at different rates?

M&Ms

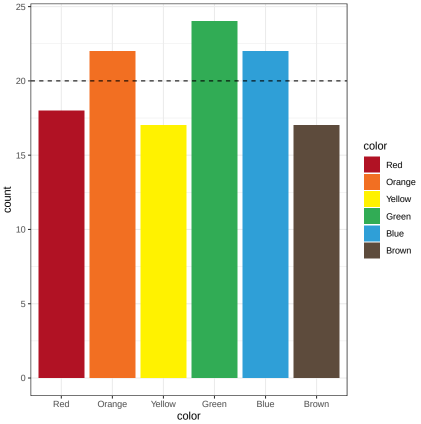
Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:



- Note that Green M&Ms exceed by the expected count by 20%.
- Does this give good evidence that M&M colors appear at different rates?
    - Suppose we had 20 colors instead of 6. . .

M&Ms

Suppose we are interested in whether the 6 colors of M&Ms appear with equal frequency. Data from 1 jumbo bag of 120 M&Ms is summarized in the graphic below:



- Note that Green M&Ms exceed by the expected count by 20%.
- Does this give good evidence that M&M colors appear at different rates?
  - Suppose we had 20 colors instead of 6. . .
  - Would it really be unusual for 1 color to be over- or under-represented?

Data

- Let's consider some numeric data:

Color	Red	Orange	Yellow	Green	Blue	Brown
Frequency	.15	.183	.142	.2	.183	.142
Counts	18	22	17	24	22	17
Expected Counts	20	20	20	20	20	20
Difference (Obs - Exp)	-2	2	-3	4	2	-3

Data

- Let's consider some numeric data:

Color	Red	Orange	Yellow	Green	Blue	Brown
Frequency	.15	.183	.142	.2	.183	.142
Counts	18	22	17	24	22	17
Expected Counts	20	20	20	20	20	20
Difference (Obs - Exp)	-2	2	-3	4	2	-3

- We are interested in assessing whether the colors are all equally represented.

Data

- Let's consider some numeric data:

Color	Red	Orange	Yellow	Green	Blue	Brown
Frequency	.15	.183	.142	.2	.183	.142
Counts	18	22	17	24	22	17
Expected Counts	20	20	20	20	20	20
Difference (Obs - Exp)	-2	2	-3	4	2	-3

- We are interested in assessing whether the colors are all equally represented.
  - We can represent this as a statement about the theoretical proportion of each each color, $p_r, p_o, p_y, p_g, p_b, p_{br}$

Data

- Let's consider some numeric data:

Color	Red	Orange	Yellow	Green	Blue	Brown
Frequency	.15	.183	.142	.2	.183	.142
Counts	18	22	17	24	22	17
Expected Counts	20	20	20	20	20	20
Difference (Obs - Exp)	-2	2	-3	4	2	-3

- We are interested in assessing whether the colors are all equally represented.
  - We can represent this as a statement about the theoretical proportion of each each color, $p_r, p_o, p_y, p_g, p_b, p_{br}$
- We want to test the following hypotheses:

$$H_0: p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a: \text{at least one of the } p\text{'s is not as specified above}$$

Randomization

- Since we have theoretical values for each proportion, we can simulate samples under the null hypothesis

Randomization

- Since we have theoretical values for each proportion, we can simulate samples under the null hypothesis

```
## # A tibble: 6 x 8
##   color  Sample_1 Sample_2 Sample_3 Sample_4 Sample_5 expected observed
##   <chr>  <chr>    <chr>    <chr>    <chr>    <chr>    <chr>    <chr>
## 1 Blue   22       10       22       13       18       20       22
## 2 Brown  15       25       17       17       24       20       17
## 3 Green  28       17       24       23       18       20       24
## 4 Orange 19       21       23       29       26       20       22
## 5 Red    19       20       23       19       13       20       18
## 6 Yellow 17       27       11       19       21       20       17
```

Randomization

- Since we have theoretical values for each proportion, we can simulate samples under the null hypothesis

```
## # A tibble: 6 x 8
##   color  Sample_1 Sample_2 Sample_3 Sample_4 Sample_5 expected observed
##   <chr>  <chr>    <chr>    <chr>    <chr>    <chr>    <chr>    <chr>
## 1 Blue   22       10       22       13       18       20       22
## 2 Brown  15       25       17       17       24       20       17
## 3 Green  28       17       24       23       18       20       24
## 4 Orange 19       21       23       29       26       20       22
## 5 Red    19       20       23       19       13       20       18
## 6 Yellow 17       27       11       19       21       20       17
```

- How does the observed data compare?

The χ^2 Statistic

- We need a statistic that measures how much our sample counts differ from the expected counts. This statistic should...

The χ^2 Statistic

- We need a statistic that measures how much our sample counts differ from the expected counts. This statistic should...
  - Take large values when the sample and expected counts differ
  - Take small values when the sample and expected values are similar
  - Be standardized, so that large expected counts don't on their own lead to large values of the statistic.

The χ^2 Statistic

- We need a statistic that measures how much our sample counts differ from the expected counts. This statistic should. . .
  - Take large values when the sample and expected counts differ
  - Take small values when the sample and expected values are similar
  - Be standardized, so that large expected counts don't on their own lead to large values of the statistic.
- The most commonly used statistic for this purpose is the chi-squared statistic χ^2:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The χ^2 Statistic

- We need a statistic that measures how much our sample counts differ from the expected counts. This statistic should...
  - Take large values when the sample and expected counts differ
  - Take small values when the sample and expected values are similar
  - Be standardized, so that large expected counts don't on their own lead to large values of the statistic.
- The most commonly used statistic for this purpose is the chi-squared statistic χ^2:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

  - The sum is taken over all values of the categorical variable.

The χ^2 Statistic

- We need a statistic that measures how much our sample counts differ from the expected counts. This statistic should...
  - Take large values when the sample and expected counts differ
  - Take small values when the sample and expected values are similar
  - Be standardized, so that large expected counts don't on their own lead to large values of the statistic.
- The most commonly used statistic for this purpose is the chi-squared statistic χ^2:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

  - The sum is taken over all values of the categorical variable.
  - For example, if the categorical variable has 6 levels, this sum has 6 terms.

Observed Statistic

- What is the χ^2 statistic for our observed sample?

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

- But what counts as *large*?

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

- But what counts as *large*?
- Let's compute the χ^2 statistic for each of the previous 5 simulated samples

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

- But what counts as *large*?
- Let's compute the χ^2 statistic for each of the previous 5 simulated samples

```
## # A tibble: 5 x 2
##   r          chi2
##   <chr>     <dbl>
## 1 Sample_1    5.2
## 2 Sample_2    9.2
## 3 Sample_3    6.4
## 4 Sample_4    7.5
## 5 Sample_5    5.5
```

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

- But what counts as *large*?
- Let's compute the χ^2 statistic for each of the previous 5 simulated samples

```
## # A tibble: 5 x 2
##   r          chi2
##   <chr>     <dbl>
## 1 Sample_1    5.2
## 2 Sample_2    9.2
## 3 Sample_3    6.4
## 4 Sample_4    7.5
## 5 Sample_5    5.5
```

- So our statistic is much smaller than the statistics for these 5 samples.

Observed Statistic

- What is the χ^2 statistic for our observed sample?

$$\chi^2 = \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(17-20)^2}{20} = 2.3$$

- But what counts as *large*?
- Let's compute the χ^2 statistic for each of the previous 5 simulated samples

```
## # A tibble: 5 x 2
##   r          chi2
##   <chr>     <dbl>
## 1 Sample_1    5.2
## 2 Sample_2    9.2
## 3 Sample_3    6.4
## 4 Sample_4    7.5
## 5 Sample_5    5.5
```

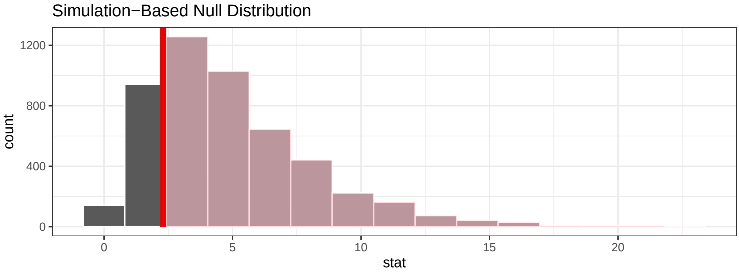
- So our statistic is much smaller than the statistics for these 5 samples.
  - But is this a fluke?

Distribution of χ^2 statistics

- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution

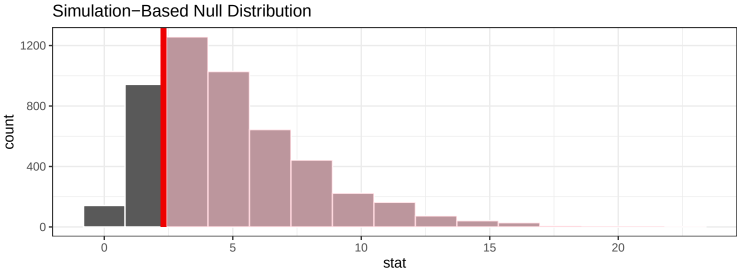
Distribution of χ^2 statistics

- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution



Distribution of χ^2 statistics

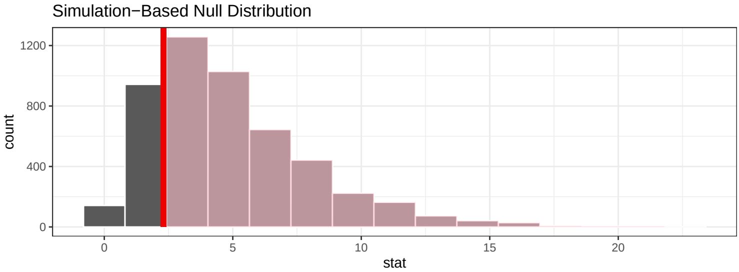
- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution



- For this data, it seems that most statistics are between 0 and 10.

Distribution of χ^2 statistics

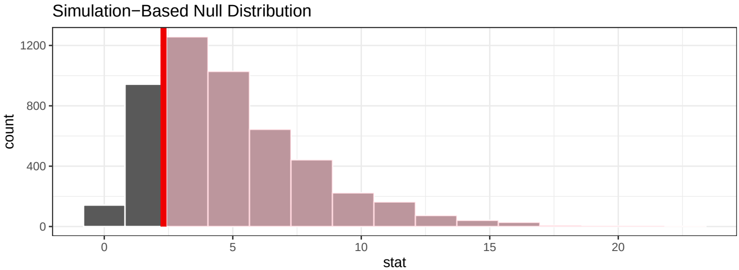
- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution



- For this data, it seems that most statistics are between 0 and 10.
  - Almost no statistic is greater than 15. And NONE are greater than 20.

Distribution of χ^2 statistics

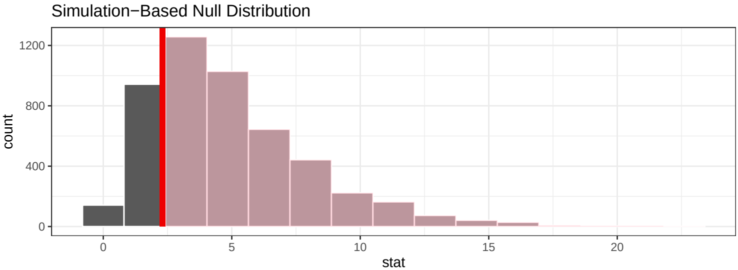
- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution



- For this data, it seems that most statistics are between 0 and 10.
  - Almost no statistic is greater than 15. And NONE are greater than 20.
- Our observed statistic of $\chi^2 = 2.3$ is very moderate

Distribution of χ^2 statistics

- Let's calculate the χ^2 statistic for several thousand other samples and plot the distribution



- For this data, it seems that most statistics are between 0 and 10.
  - Almost no statistic is greater than 15. And NONE are greater than 20.
- Our observed statistic of $\chi^2 = 2.3$ is very moderate
  - A statistic more extreme would occur about 80% of the time!

Using `infer`

- How do we find the probability that a particular χ^2 value would occur?

Using `infer`

- How do we find the probability that a particular χ^2 value would occur?
  - Use `infer`!

Using infer

- How do we find the probability that a particular χ^2 value would occur?
  - Use infer!

```
set.seed(1)
MMs %>% specify(response = color) %>%
  hypothesize(null = "point",
              p = c("Red" = 1/6, "Orange" = 1/6, "Yellow" = 1/6,
                    "Green" = 1/6, "Blue" = 1/6, "Brown" = 1/6))  %>%
  generate(reps = 5000, type = "simulate") %>%
  calculate(stat = "Chisq") %>%
  get_p_value(obs_stat = 2.3, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.813
```

Conclusions

- We tested the following hypotheses:

$$H_0 : p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a : \text{at least one of the } p\text{'s is not as specified above}$$

Conclusions

- We tested the following hypotheses:

$$H_0: p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a: \text{at least one of the } p\text{'s is not as specified above}$$

- Our observed statistic $\chi^2 = 2.3$ had a simulated p-value of approximately 0.8

Conclusions

- We tested the following hypotheses:

$$H_0: p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a: \text{at least one of the } p\text{'s is not as specified above}$$

- Our observed statistic $\chi^2 = 2.3$ had a simulated p-value of approximately 0.8
- We do not reject H_0 at the $\alpha = 0.05$ significance level (or at any reasonable level)
  - It is likely that such a difference in counts would arise due to chance, if the null hypothesis were true.

Conclusions

- We tested the following hypotheses:

$$H_0: p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a: \text{at least one of the } p\text{'s is not as specified above}$$

- Our observed statistic $\chi^2 = 2.3$ had a simulated p-value of approximately 0.8
- We do not reject H_0 at the $\alpha = 0.05$ significance level (or at any reasonable level)
  - It is likely that such a difference in counts would arise due to chance, if the null hypothesis were true.
- The test provides inconclusive evidence that frequency differs among colors.

Conclusions

- We tested the following hypotheses:

$$H_0 : p_r = \frac{1}{6},\ p_o = \frac{1}{6},\ p_y = \frac{1}{6},\ p_g = \frac{1}{6},\ p_b = \frac{1}{6},\ p_{br} = \frac{1}{6}$$
$$H_a : \text{at least one of the } p\text{'s is not as specified above}$$

- Our observed statistic $\chi^2 = 2.3$ had a simulated p-value of approximately 0.8
- We do not reject H_0 at the $\alpha = 0.05$ significance level (or at any reasonable level)
  - It is likely that such a difference in counts would arise due to chance, if the null hypothesis were true.
- The test provides inconclusive evidence that frequency differs among colors.
  - Importantly, it does not verify that colors ARE equally distributed.

Theory-based p-values for χ^2 statistics

If we have independent observations on a categorical variable with k levels, and each observed count is at least 5,

Theory-based p-values for χ^2 statistics

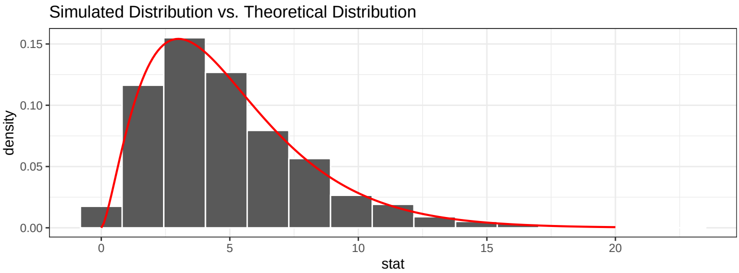
If we have independent observations on a categorical variable with k levels, and each observed count is at least 5,

- Then χ^2 is approximately the Chi-Squared distribution with $k-1$ degrees of freedom.

Theory-based p-values for χ^2 statistics

If we have independent observations on a categorical variable with k levels, and each observed count is at least 5,

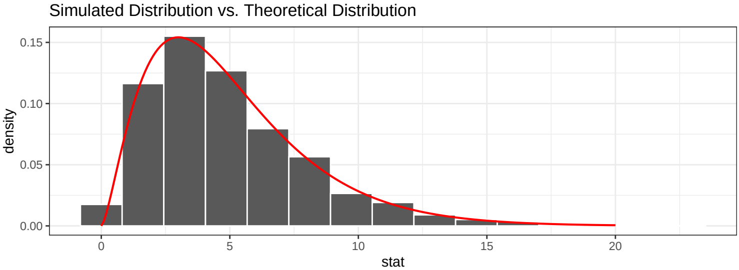
- Then χ^2 is approximately the Chi-Squared distribution with $k-1$ degrees of freedom.



Theory-based p-values for χ^2 statistics

If we have independent observations on a categorical variable with k levels, and each observed count is at least 5,

- Then χ^2 is approximately the Chi-Squared distribution with $k-1$ degrees of freedom.

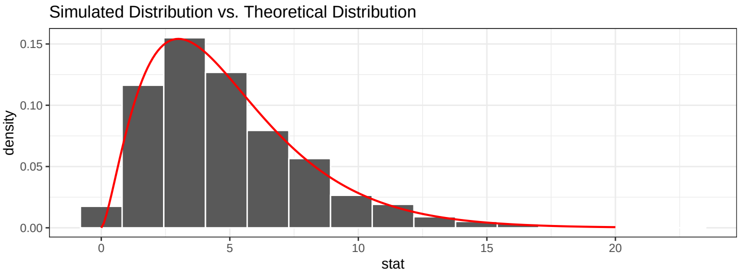


- Use `pchisq(q = ..., df = ..., lower.tail = F)` to find the area to the right of the observed statistic q.

Theory-based p-values for χ^2 statistics

If we have independent observations on a categorical variable with k levels, and each observed count is at least 5,

- Then χ^2 is approximately the Chi-Squared distribution with $k-1$ degrees of freedom.



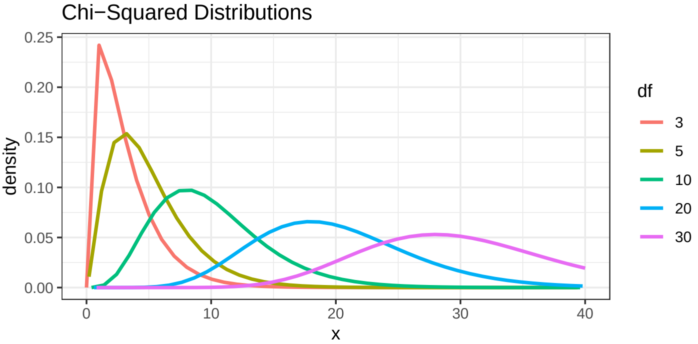
- Use `pchisq(q = ..., df = ..., lower.tail = F)` to find the area to the right of the observed statistic q.

```
pchisq(q = 2.3, df = 5, lower.tail = F)
```

```
## [1] 0.8062669
```

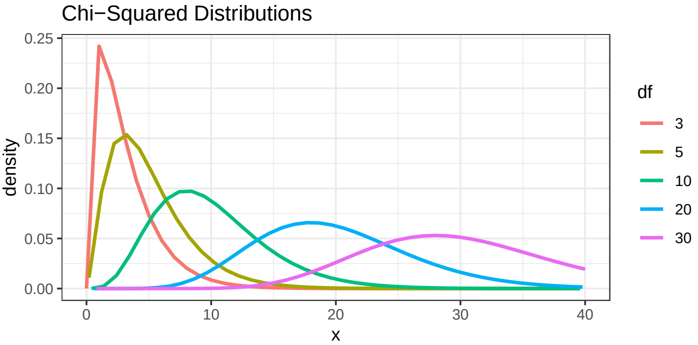
The Chi-Squared Distribution

Just Normal distributions are described by their mean μ and standard deviation σ, the Chi-Square distribution is described by its degrees of freedom *df*.



The Chi-Squared Distribution

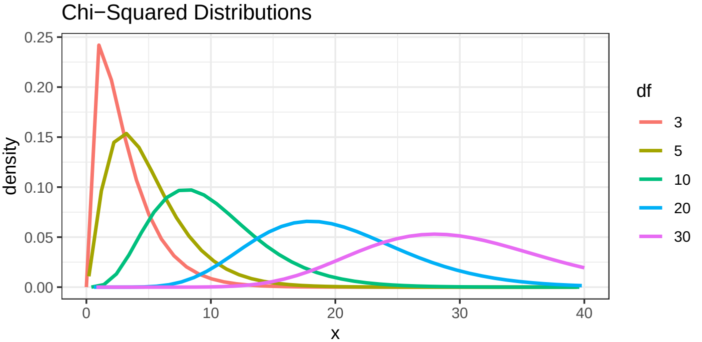
Just Normal distributions are described by their mean μ and standard deviation σ, the Chi-Square distribution is described by its degrees of freedom *df*.



- The mean of a chi-square distribution is *df*, while the standard deviation is $\sqrt{2 \cdot df}$

The Chi-Squared Distribution

Just Normal distributions are described by their mean μ and standard deviation σ, the Chi-Square distribution is described by its degrees of freedom df.



- The mean of a chi-square distribution is df, while the standard deviation is $\sqrt{2 \cdot df}$
- For Chi-Squared tests, larger degrees of freedom require larger χ^2 statistics to reject H_0.

Section 2

Chi-Square Test for Independence

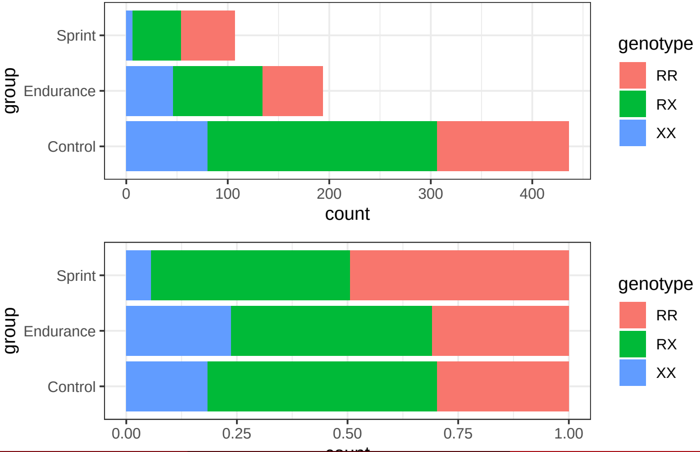
Genetic Basis for Fast Twitch Muscles

A study on genetics and fast-twitch muscles includes a sample of sprinters, endurance athletes, and a control group of non-athletes.

Genetic Basis for Fast Twitch Muscles

A study on genetics and fast-twitch muscles includes a sample of sprinters, endurance athletes, and a control group of non-athletes.

- Is there an association between a genotype classification (RR, RX, or XX) and group?



Contingency Table

Consider the contingency table for group and genotype

```
table(twitch$group, twitch$genotype) %>%
  addmargins()
```

```
##
##              RR  RX  XX Sum
##   Control   130 226  80 436
##   Endurance  60  88  46 194
##   Sprint     53  48   6 107
##   Sum       243 362 132 737
```

```
table(twitch$group, twitch$genotype) %>%
  prop.table( 1)
```

```
##
##                RR    RX    XX
##   Control   0.298 0.518 0.183
##   Endurance 0.309 0.454 0.237
##   Sprint    0.495 0.449 0.056
```

Contingency Table

Consider the contingency table for group and genotype

```
table(twitch$group, twitch$genotype) %>%
  addmargins()
```

```
##
##              RR  RX  XX Sum
##   Control   130 226  80 436
##   Endurance  60  88  46 194
##   Sprint     53  48   6 107
##   Sum       243 362 132 737
```

```
table(twitch$group, twitch$genotype) %>%
  prop.table( 1)
```

```
##
##                 RR    RX    XX
##   Control   0.298 0.518 0.183
##   Endurance 0.309 0.454 0.237
##   Sprint    0.495 0.449 0.056
```

- If group and genotype were independent, we would expect proportions to all be equal to the marginal proportions for genotype:

```
table(twitch$genotype) %>% prop.table()
```

```
##
##   RR   RX   XX
## 0.33 0.49 0.18
```

Expected Counts

If the null hypothesis is true, we can multiply the marginal proportions of `genotype` by the observed counts for `group` to get expected counts for each `genotype`-`group` pair:

	RR	RX	XX
Control	(0.33)(436)	(0.49)(436)	(0.18)(436)
Endurance	(0.33)(194)	(0.49)(194)	(0.18)(194)
Sprint	(0.33)(107)	(0.49)(107)	(0.18)(107)

Expected Counts

If the null hypothesis is true, we can multiply the marginal proportions of `genotype` by the observed counts for `group` to get expected counts for each `genotype-group` pair:

	RR	RX	XX
Control	(0.33)(436)	(0.49)(436)	(0.18)(436)
Endurance	(0.33)(194)	(0.49)(194)	(0.18)(194)
Sprint	(0.33)(107)	(0.49)(107)	(0.18)(107)

	RR	RX	XX
Control	144	214	78
Endurance	64	95	35
Sprint	35	52	19

Expected Counts

If the null hypothesis is true, we can multiply the marginal proportions of `genotype` by the observed counts for `group` to get expected counts for each `genotype`-`group` pair:

	RR	RX	XX
Control	(0.33)(436)	(0.49)(436)	(0.18)(436)
Endurance	(0.33)(194)	(0.49)(194)	(0.18)(194)
Sprint	(0.33)(107)	(0.49)(107)	(0.18)(107)

	RR	RX	XX
Control	144	214	78
Endurance	64	95	35
Sprint	35	52	19

- We can compare to the observed data:

	RR	RX	XX	Sum
Control	130	226	80	436
Endurance	60	88	46	194
Sprint	53	48	6	107
Sum	243	362	132	737

Expected Counts

If the null hypothesis is true, we can multiply the marginal proportions of `genotype` by the observed counts for `group` to get expected counts for each `genotype-group` pair:

	RR	RX	XX
Control	(0.33)(436)	(0.49)(436)	(0.18)(436)
Endurance	(0.33)(194)	(0.49)(194)	(0.18)(194)
Sprint	(0.33)(107)	(0.49)(107)	(0.18)(107)

	RR	RX	XX
Control	144	214	78
Endurance	64	95	35
Sprint	35	52	19

- We can compare to the observed data:

	RR	RX	XX	Sum
Control	130	226	80	436
Endurance	60	88	46	194
Sprint	53	48	6	107
Sum	243	362	132	737

- As before, we compute the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = 25$$

The Null Distribution

Under the null hypothesis, `group` and `genotype` are independent.

The Null Distribution

Under the null hypothesis, `group` and `genotype` are independent.

- We can simulate data under H_0 by permuting the `group` labels among individuals. (Just like we did for hypothesis tests for 2 proportions)
    - After each permutation, we compute a new χ^2 statistic.
    - The distribution of these statistics gives the null distribution.

The Null Distribution

Under the null hypothesis, group and genotype are independent.

- We can simulate data under H_0 by permuting the group labels among individuals. (Just like we did for hypothesis tests for 2 proportions)
  - After each permutation, we compute a new χ^2 statistic.
  - The distribution of these statistics gives the null distribution.

```
##   ID     group genotype
## 1  1 Endurance       RX
## 2  2    Sprint       XX
## 3  3   Control       XX
## 4  4    Sprint       RX
## 5  5   Control       RX
## 6  6    Sprint       RR
```

```
##   ID     group genotype
## 1  1 Endurance       RX
## 2  2    Sprint       RX
## 3  3   Control       XX
## 4  4    Sprint       RR
## 5  5   Control       XX
## 6  6    Sprint       RX
```

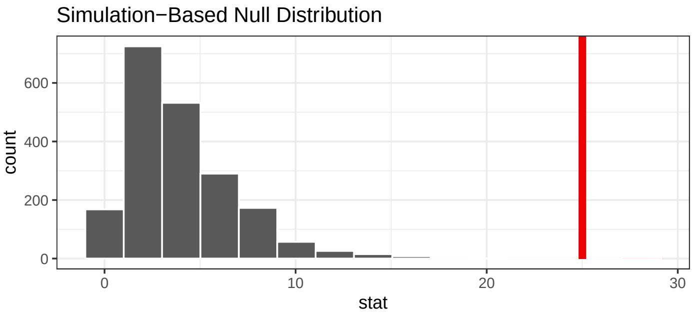
Chi-Square Statistic in `infer`

Using `infer`...

Chi-Square Statistic in infer

Using infer...

```
set.seed(49)
twitch_null <- twitch %>%
  specify(genotype ~ group) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 2000, type = "permute") %>%
  calculate(stat="Chisq")
twitch_null %>% visualize()+shade_p_value(obs_stat = 25, direction = "right")
```



P-value and conclusions

Using `infer`, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

P-value and conclusions

Using `infer`, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

- Accordinging to theory-based methods, the χ^2 statistic follows the χ^2 distribution with $(k-1) \cdot (p-1)$ degrees of freedom, where k and p are the number of levels of the response and explanatory variables.

P-value and conclusions

Using `infer`, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

- Accordinging to theory-based methods, the χ^2 statistic follows the χ^2 distribution with $(k-1)\cdot(p-1)$ degrees of freedom, where k and p are the number of levels of the response and explanatory variables.

```
pchisq(q = 25, df = 4, lower.tail = F)
```

```
## [1] 5e-05
```

P-value and conclusions

Using `infer`, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

- Accordinging to theory-based methods, the χ^2 statistic follows the χ^2 distribution with $(k-1)\cdot(p-1)$ degrees of freedom, where k and p are the number of levels of the response and explanatory variables.

```
pchisq(q = 25, df = 4, lower.tail = F)
```

```
## [1] 5e-05
```

- At significance $\alpha = 0.01$, we reject H_0 in favor the alternative:
    - This sample gives good evidence that `group` and `genotype` are associated.

P-value and conclusions

Using `infer`, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

- Accordinging to theory-based methods, the χ^2 statistic follows the χ^2 distribution with $(k-1)\cdot(p-1)$ degrees of freedom, where k and p are the number of levels of the response and explanatory variables.

```
pchisq(q = 25, df = 4, lower.tail = F)
```

```
## [1] 5e-05
```

- At significance $\alpha = 0.01$, we reject H_0 in favor the alternative:
  - This sample gives good evidence that `group` and `genotype` are associated.
- What association is there?

P-value and conclusions

Using infer, the approximate p-value is

```
twitch_null %>% get_p_value(obs_stat = 25, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0005
```

- Accordinging to theory-based methods, the χ^2 statistic follows the χ^2 distribution with $(k-1)\cdot(p-1)$ degrees of freedom, where k and p are the number of levels of the response and explanatory variables.

```
pchisq(q = 25, df = 4, lower.tail = F)
```

```
## [1] 5e-05
```

- At significance $\alpha = 0.01$, we reject H_0 in favor the alternative:
  - This sample gives good evidence that group and genotype are associated.
- What association is there?
  - We'll need to further study and experiment to find out.