Goodness of fit and the chi-squared

Announcements

Goals for today

- Goodness of fit: looking at one variable with multiple categories
- Introduce the chi-squared

One variable with multiple categories

One categorical variable with more than 2 categories

- With one continuous variable we tested whether the mean was equal to a hypothesized null (Z or one sample T)
- With one categorical variable with two categories (binary, yes/no) we tested that the proportion was equal to a hypothesized null (one sample test of proportions)
- What do we do with a categorical variable when there are more than 2 categories?

One categorical variable with more than 2 categories

The general pattern we will follow for these types of variables is:

- estimate how many observations we would expect in each category under our null hypothesis
- compare the number of observations in each category to the expected value
- summarize these differences and compare them to a theoretical distribution

Jury Selection example

Suppose that the following number of people were selected for jury duty in the previous year, in a county where jury selection was supposed to be random.

Ethnicity	White	Black	Latinx	Asian	Other	Total
Number selected	1920	347	19	84	130	2500

You read online about concerns that jury was not selected randomly. How can you test this evidence?

- Example derived from this video.

Jury Selection example

Consider the distribution of race/ethnicity in the county overall:

Ethnicity	White	Black	Latinx	Asian	Other	Total
% in the population	42.2%	10.3%	25.1%	17.1%	5.3%	100%

How do we determine the counts that are expected (E) under the assumption that selection was random?:

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected count						2500

Jury Selection example

Ethnicity	White	Black	Latinx	Asian	Other	Total
% in the population	42.2%	10.3%	25.1%	17.1%	5.3%	100%

- To fill in the table, multiple the total size of the jury by the % of the population of each race/ethnicity:

Expected counts under the assumption that selection is random from the county:

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected	2500 $\times$	2500 $\times$	2500 $\times$	2500 $\times$	2500 $\times$	2500
count	0.422	0.103	0.251	0.171	0.053	
$=$	1055	257.5	627.5	427.5	132.5	2500

Jury Selection example

How far off does the observed counts of race/ethnicities in the sample differ from what we would expect if the jury had been selected randomly?

Jury Selection example

Here are the counts we observed (O):

Ethnicity	White	Black	Latinx	Asian	Other	Total
Observed count	1920	347	19	84	130	2500

Which we can compare to our expected (E):

Ethnicity	White	Black	Latinx	Asian	Other	Total
Expected count	1055	257.5	627.5	427.5	132.5	2500

Jury Selection example

This plot shows the deviations between the observed and expected number of jurors. What is the chance of observed deviations of these magnitudes (or larger) under the null hypothesis?

Jury Selection example

- Recall the usual form of the test statistic:

$$\frac{estimate - null}{SE}$$

- We want an estimate that somehow quantifies how different the observed counts (O) are from the expected counts (E) across the 5 race/ethnicities.

The Chi-square test statistic

The chi^2 test statistic quantifies the magnitude of the difference between observed and expected counts under the null hypothesis. It looks like this:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- k is the number of cells in the table. Here, k is the number of race/ethnicity groups. That is, $k = 5$
- O_i is the observed count for the i^{th} group (here race/ethnicity)
- E_i is the expected count for the i^{th} group
- χ^2 is a distribution, like t or Normal.

The Chi-square test statistic

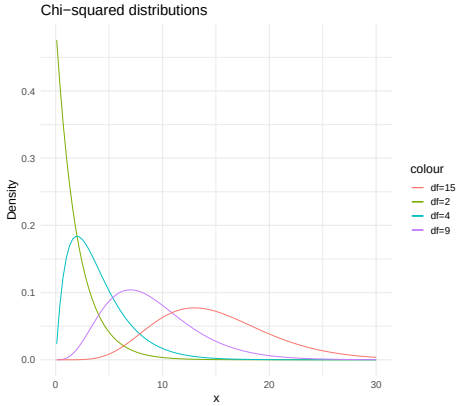
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- The numerator measures the squared deviations between the observed (O) and expected (E) values. Bigger deviations will make the test statistic larger (which means that its corresponding p-value will be smaller)
- The denominator makes this magnitude *relative* to what we expect. This adjusts for the different magnitude of expected counts. For example, with our example, we would *expect* the number of white jurors to be close to 1055, but we would expect the number of Latinx jurors to be close to 628. Therefore, we divide by these expectations such that a difference of 100 fewer Latinx jurors than expected counts for more than a difference of 100 few white jurors.

The Chi-Square distribution

The Chi-square distribution

The chi-square distribution is a new distribution to us. Like the t-distribution, the chi-square distribution only has one parameter: a degrees of freedom. The degrees of freedom is equal to the number of groups (here, race/ethnicities) - 1. Or, $df = k - 1$.



The shape of the Chi-square

- As the `df` is increased, the distribution's central tendency moves to the right.
- This means that there will be more probability out in the right tail when the degrees of freedom is higher.
- The chi-square distribution is also positive. We only ever compute upper tail probabilities for the chi-square test because there is only one form to the H_a.

Back to the jury example

State the null and alternative hyptoheses.

- The null hypothesis is that the proportions of each race/ethnicity in the jury pool is the same as the proportion of each group in the county. That is:

$H_0 : p_{white} = 42.2\%,\ p_{black} = 10.3\%,\ p_{latinx} = 25.1\%,\ p_{asian} = 17.1\%,\ p_{other} = 5.3\%$

H_a : At least one of p_k is different than specified in H_0, for k being one of white, black, latinx, asian, or other.

Back to the jury example

Calculate the chi-square statistic using the jury data.

Ethnicity	White	Black	Latinx	Asian	Other	Total
O	1920	347	19	84	130	2500
E	1055	257.5	627.5	427.5	132.5	2500

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(1920-1055)^2}{1055} + \frac{(347-257.5)^2}{257.5} + \frac{(19-627.5)^2}{627.5} + \frac{(84-427.5)^2}{427.5} + \frac{(130-132.5)^2}{132.5}$$

$$\chi^2 = 709.218 + 31.10777 + 590.0753 + 276.0053 + 0.04716981$$

$$\chi^2 = 1606.454$$

Back to the jury example

Calculate the p-value (what is the approprate degrees of freedom?).

```
pchisq(q = 1606.454, df = 4,lower.tail = F)
```

```
## [1] 0
```

The probability of seeing this pool of people chosen for jury duty under the null hypothesis of random sampling from the county is so small that R rounded the p-value to 0!

Chi-square test in R

Run the chi-square test using the `chisq.test` command in R.

```
chisq.test(x = c(1920, 347, 19, 84, 130), # x is vector of observed counts
           p = c(.422, .103, .251, .171, .053)) # p is probability under the
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(1920, 347, 19, 84, 130)
## X-squared = 1606.5, df = 4, p-value < 2.2e-16
```

Interpretation

- Which race/ethnicities appear to deviate the most from what was expected under the null hypothesis?
  - Compare the proportion observed vs. proportion expected
  - Compare the count observed vs. the count expected
  - Compare the 5 contributions to the chi-square test from each race/ethnicity. We see that whites, Latinx, and Asians contribute the most to the χ^2 statistic. This agrees with what we saw in the data visualization in terms of the size of the gaps between observed and expected counts.

Example 2: Births by day of the week (Ex. 21.7)

A random sample of 700 births from local records shows the distribution across the days of the weeks:

Day	M	T	W	Th	F	Sa	Su
Births	110	124	104	94	112	72	84

Is there evidence that the proportion of births occurring on any given day of the week is not random?

Example 2: Births by day of the week (Ex. 21.7)

State the null and alternative hypotheses

$H_0 : p_1 = 1/7,\ p_2 = 1/7,\ p_3 = 1/7,\ p_4 = 1/7,\ p_5 = 1/7,\ p_6 = 1/7,\ p_7 = 1/7,\ .$
Written another way:

$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$

H_a : At least one of these p_k differ from $1/7$. Or: not all p_k equal $1/7$.

Example 2: Births by day of the week (Ex. 21.7)

Calculate the expected counts under H_0

Day	M	T	W	Th	F	Sa	Su
Expected births	?	?	?	?	?	?	?

Example 2: Births by day of the week (Ex. 21.7)

Calculate the expected counts under H_0

- Use the fact that the total number of births equaled 700. Then 700*(1/7) = 100. We would expect to see around 100 births on each day if the births occurring randomly over the course of the week.

Day	M	T	W	Th	F	Sa	Su
Expected births	100	100	100	100	100	100	100

Example 2: Births by day of the week (Ex. 21.7)

Calculate the chi^2 test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(110-100)^2}{100} + \frac{(124-100)^2}{100} + \frac{(104-100)^2}{100} + \frac{(94-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(72-100)^2}{100} + \frac{(84-100)^2}{100}$$

$$\chi^2 = 1 + 5.76 + 0.16 + 0.36 + 1.44 + 7.84 + 2.56$$

$$\chi^2 = 19.12$$

- Based on the individual contributions of each day to the chi-square statistic, which days were most different from the expected value under H_0?

Example 2: Births by day of the week (Ex. 21.7)

Calculate the p-value

```
pchisq(q = 19.12, df = 6, lower.tail = F)
```

```
## [1] 0.003965699
```

Interpret the p-value

Based on a p-value of 0.39%, there is very strong evidence against the null hypothesis in favor of an alternative hypothesis where the proportion of births across the seven days of the week are not evenly distributed.

Example 3: cheating at dice?

Suppose there is a game in which the objective is to roll sixes as possible using 3 die. Over 100 rolls, one of the players seems to be winning quite often, we see the following

Number of 6s	0	1	2	3
Observed rolls	47	35	15	3

We suspect they are using a loaded die or cheating in some way.

Are they cheating? Or just lucky (within the bounds of chance)?

Example derived from this site

Example 3: cheating at dice?

What would we expect?
The rolls of dice should follow a binomial distribution (# of successes in # trials)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

What is P here? What is K?

Example 3: cheating at dice?

Remember dbinom?

'dbinom(#successes,size,probability of success)'

This function calculates the probability of observing x successes when $X \sim Binom(n, p)$

```
Expect_0<-dbinom(0,size=3,prob=0.166666667)
Expect_1<-dbinom(1,size=3,prob=0.166666667)
Expect_2<-dbinom(2,size=3,prob=0.166666667)
Expect_3<-dbinom(3,size=3,prob=0.166666667)
Expected<-c(Expect_0,Expect_1,Expect_2,Expect_3)
Expected
```

```
## [1] 0.57870370 0.34722222 0.06944444 0.00462963
```

Example 3: cheating at dice?

Number of 6s	0	1	2	3
Observed rolls	47	35	15	3
Expected rolls	57.9	34.7	6.9	0.46

Exampe 3: cheating at dice?


```
chisq.test(x = c(47,35,15,3), # x is vector of observed counts
           p = Expected) # p is probability under the null
```

```
## Warning in chisq.test(x = c(47, 35, 15, 3), p = Expected): Chi-squared
## approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(47, 35, 15, 3)
## X-squared = 25.292, df = 3, p-value = 1.342e-05
```

Conditions to perform a chi-square test

- Fixed n of observations
- All observations are independent of one another. What does this mean in the first example? In the second example?
- Each observation falls into just one of the k mutually exclusive categories
- The probability of a given outcome is the same for each observation.

Counts requirement

- At least 80% of the cells have 5 or more observations ($O_i \geq 5$ for $\geq 80\%$ of the cells)
- All k cells have expected counts > 1 ($E_i > 1$)

Recap

- We learned about the chi-square χ^2 test
- We used the test to look at the distribution of one categorical variable to test the null hypothesis

$$H_0 : p_1 = \#_1, p_2 = \#_2, ..., p_k = \#_k$$

where $\#_1$, $\#_2$, ..., $\#_k$ were provided in the question.

- This test is called the chi-square goodness of fit test
  - How good do the expected counts "fit" the observed counts?

Recap of the chi-square goodness of fit test (for one categorical variable)

- The chi-square test statistic (Or, the "Old McDonald" test statistic: "E-i, E-i, O!"):

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

- We can also use the chi-square test to investigate the relationship between two categorical variables
- The form of the test statistic is the same!

Think back to Chapter 5...

- In Chapter 5, we learned about two-way tables and talked about how to calculate the conditional probability of one variable given another.
- For example, what is the conditional probability of lung cancer among smokers vs. among non-smokers?
- Recall also the definition of explanatory and response variables. In the case of smoking and lung cancer, which was explanatory and which was response?

Hypotheses for the chi-square test for two categorical variables

- H_0 : Response and explanatory variables are independent.

Stated another way:

- H_0 : The probability distribution for lung cancer among smokers is equal to the probability distribution among non-smokers
- If you remember our probability independence rules P(A|B)=P(A)... how does this apply here?

Alternative hypothesis:

- H_a : Response and explanatory variables are dependent.
- H_a : The probability distribution for lung cancer among smokers is different from the probability among non-smokers.
  - The alternative hypothesis is not one-sided or two-sided. It is non-specific and allows for any kind of difference from the null. Does this mean we look at 2 sides of the distribution?

Chi-square test of independence

- As with the goodness of fit test, we compare observed cell counts (O_i) to expected cell counts (E_i), but this time we have a two-way table showing the distribution of data across two variables.

Steps of the chi-squared test based on these data.

1. Make the two-way table.
2. Calculate the expected values.
3. Calculate the test statistic.
4. Calculate the degrees of freedom and p-value.
5. Interpret the p-value and assess the evidence.

Also: assess whether the conditions are met to conduct the test.

Sample size conditions for the chi-square test of independence

- $E_i \geq 5$ for at least 80% of the cells
- All $E_i > 1$
- If table is 2X2, all four cells need $E_i \geq 5$

Statistical assumptions for the chi-square test of independence

Must have either data arising from:

- Independent SRSs from ≥ 2 populations, with each individual classified according to one category (i.e., each individual can only belong to one cell in the table so the categories need to be mutually exclusive)
- A single SRS, with each individual classified according to each of two categorical variables.

Example : gastroenteritis outbreak

From Gross et al. Public health reports vol 104, March-April 1989, 164-169

Group	Sandwich	No Sandwich	Row total
Ill	109	4	113
Not Ill	116	34	150
Column total	225	38	263

- The inner four cells are the observed cell counts
- The outer row and column are the table margins
- The margins are important for the computations, so be sure to calculate the marginal counts if they aren't computed for you.

Example : gastroenteritis outbreak

Group	Sandwich	No Sandwich	Row total
Ill	109	4	113
Not Ill	116	34	150
Column total	225	38	263

- What would these data look like under the null hypothesis of no association between sandwiches and getting sick?
- That is, what are the expected counts under the null hypothesis?

Example : gastroenteritis outbreak

To help us get the expected counts, calculate the marginal percentages and remove the data from the inner cells

Group	Sandwich	No Sandwich	Row total
Ill	?	?	113 (43%)
Not Ill	?	?	150 (57%)
Column total	225 (85.6%)	38 (14.4%)	263

- Recall that if A and B are independent then $P(A\&B) = P(A)P(B)$. That is, if sandwiches and illness are independent, then $P(Sandwich\&Illness) = P(S)P(I) = .855 * .43 = .368 = 36.8\%$
- What is the expected count for the S&I cell under the null hypothesis?
  - 0.368*263 = 96.7

Example : gastroenteritis outbreak

- What is the expected count for the S&I cell under the null hypothesis?
  - 0.368*263 = 96.7

Group	Sandwich	No Sandwich	Row total
Ill	96.7	16.3	113 (43%)
Not Ill	128.3	21.7	150 (57%)
Column total	225 (85.6%)	38 (14.4%)	263

- What are the expected counts for the other cells under H_0?
  - S' & I: $0.144 \times 0.43 \times 263$
  - S & I': $0.856 \times 0.57 \times 263$
  - S' & I': $0.144 \times 0.57 \times 263$
- Note that once you compute two of the cells you can use subtraction from the marginal counts to get the other two values. Thus, only do as much calculation as you need and then get the rest by subtracting from the margins.

A trick for calculating the expected counts

- On the previous slides, we first calculated the marginal probabilities and multiplied them together and with the sample size to calculate the expected counts.
- We started with this calculation so you could see the intuition for why it worked.
- But there is a quicker way!:

$$E_i = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

A trick for calculating the expected counts

Worked calculations:

- S&I = 225*113/263 = 96.7
- S&I' = 225*150/263 = 128.3
- S'&I = 38*113/263 = 16.3
- S'&I' = 38*150/263 = 21.7
- Use this trick for faster calculation

Compare E_i and O_i

Group	Sandwich	No Sandwich
Ill	E=96.7 vs. O=109	E=16.3 vs. O=4
Not Ill	E=128.3 vs. O=116	E=21.7 vs. O=34

- Think about the direction of the deviations. When is the observed higher than the expected? When is it the other way around? Does this jibe with the association you're expecting?

Calculate the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

$$\chi^2 = \frac{(96.7-109)^2}{96.7} + \frac{(16.3-4)^2}{16.3} + \frac{(128.3-116)^2}{128.3} + \frac{(21.7-34)^2}{21.7}$$

$$\chi^2 = 1.5645 + 9.2816 + 1.1792 + 6.972 = 18.9973$$

Calculate the degrees of freedom

- Like last class, we need a degrees of freedom for the test statistic.
- When we only had one variable the degrees of freedom equaled $k - 1$
- Here we have two variables. The degrees of freedom equals $(r - 1)(c - 1)$, where r is the number of inner row cells and c is the number of inner column cells (here $r = 2$ and $c = 2$)
- For these data, df = (2-1)(2-1) = 1

Calculate the p-value for the chi-square test

```
pchisq(q = 18.9972, df = 1, lower.tail = F) #df = (2-1)(2-1) = 1
```

```
## [1] 1.309104e-05
```

- Remember for the chi-squared test we always do an upper tail test!

Interpret the p-value: Assuming no association between sandwiches and illness, there is less than a 0.01% chance of the chi-square value we calculated or a larger one. This probability is small enough that there is evidence in favor of the alternative hypothesis that there is a relationship between sandwiches and illness.

Chi-squared test of independence in R

Chi-square test of independence in R

To compute the chi-square test in R, we need to first put this two-way table into a data frame:

```
library(tibble)
two_way <- tribble(~ sandwich, ~ nosandwich,
                   109,         4, #row for Illness
                   116,         34) #row for no Illness
```

Chi-square test of independence in R

Then, we use `chisq.test()`. We set `correct=F` to get a value closer to what we calculated by hand - there will be some differences here because of rounding:

```
chisq.test(two_way, correct = F) #not using Yates' correction for continuity
```

```
##
##  Pearson's Chi-squared test
##
## data:  two_way
## X-squared = 19.074, df = 1, p-value = 1.257e-05
```

Yates' continuity correction

Continuity correction

- The χ^2 is a continuous distribution and we are using discrete observations to estimate a χ^2 value.
- When there are many degrees of freedom and/or a large number of observations, this is a reasonable approximation
- In a 2x2 table (df=1) with a small sample size this may be less reasonable.
- The correction looks like this

$$\chi^2 = \sum_{i=1}^{k} \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

What do you thing this will do to the χ^2 value?

Chi-square test of independence in R

Compare to the result where `correct = T` (the default with correction):

```
chisq.test(two_way, correct = T) #using Yates' correction for continuity
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  two_way
## X-squared = 17.558, df = 1, p-value = 2.786e-05
```

- A common practice is to incorporate the Yate's continuity correction when n < 100 or any O_i < 10. Reference

Relationship between the chi-square test and the two-sample z test

- The topic of one of the labs.

Extending the 2 X 2 to a more generic R X C

Extending the 2 X 2 to a more generic R X C

- we have looked at 2 x 2 as an example of how you would compare two categorical variables
- 2 X 2 tables are common as many variables that we look at are classified as binary
- however the chi-squared test works the same way for variables with more than 2 categories

Another example: HPV Status and age group

Suppose you had these data of HPV status vs. age group.

Age Group	HPV +	HPV -	Row total
14-19	160	492	652 (33.9%)
20-24	85	104	189 (9.8%)
25-29	48	126	174 (9.1%)
30-39	90	238	328 (17.1%)
40-49	82	242	324 (16.9%)
50-59	50	204	254 (13.2%)
Col total	515 (26.8%)	1406 (73.2%)	1921

- Which variable is explanatory and which is response?
- Can you formulate a null and alternative hypothesis using these data?

Welcome back to the dodged histogram

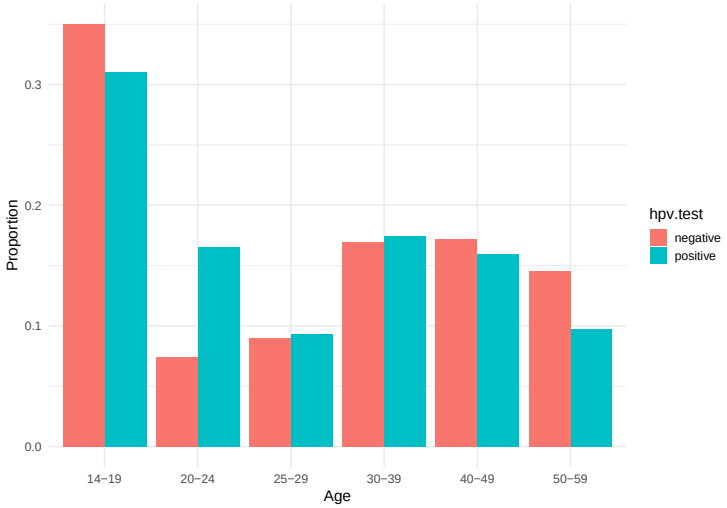
- Recall that we used dodged histograms to compare the conditional distribution of one variable across the levels of another variable.
- These plots are useful to make before we conduct the hypothesis test.

Remember the syntax: geom_bar(aes(fill = outcome), stat = "identity", position = "dodge")

The "identity" option tells R that the values are already calculated

Welcome back to the dodged histogram

Is there visual evidence of a difference between the conditional distribution of HPV status by age group?



Example: HPV Status and age group

- Conduct all stages of the chi-square hypothesis test for independence (state the null and alternative hypotheses, calculate the test statistic, calculate the degrees of freedom and the p-value, interpret the p-value and assess whether there is evidence against the null in favor of the alternative.)

Example: HPV Status and age group

Expected values

Age Group	HPV +	HPV -	Row total
14-19	174.7	477.3	652 (33.9%)
20-24	50.7	138.3	189 (9.8%)
25-29	46.6	127.4	174 (9.1%)
30-39	87.9	240.1	328 (17.1%)
40-49	86.8	237.2	324 (16.9%)
50-59	68.1	185.9	254 (13.2%)
Col total	515 (26.8%)	1406 (73.2%)	1921

Calculate the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

$$\chi^2 = \frac{(174.7-160)^2}{174.7} + \frac{(477.3-492)^2}{477.3} + \frac{(50.7-85)^2}{50.7} + ...$$

$$\chi^2 = 40.55$$

Example: HPV Status and age group

```
pchisq(40.55, df=5, lower.tail=F)
```

```
## [1] 1.15665e-07
```

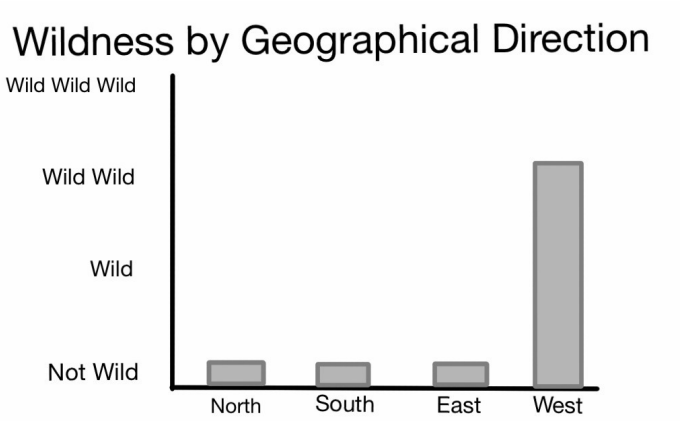
Example: HPV Status and age group

```
library(tibble)
n_way <- tribble(~ HPV, ~ noHPV,
                    160,492,
                    85,104,
                    48,126,
                    90,238,
                    82,242,
                    50,204)
chisq.test(n_way, correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  n_way
## X-squared = 40.554, df = 5, p-value = 1.155e-07
```

Congratulations!

Parting Graphic



Source: The Escape Club, Will Smith