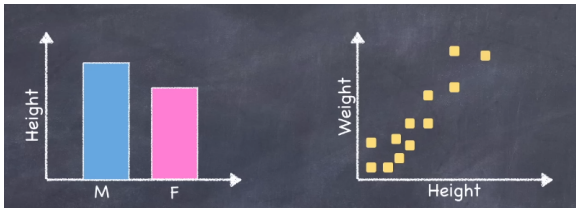
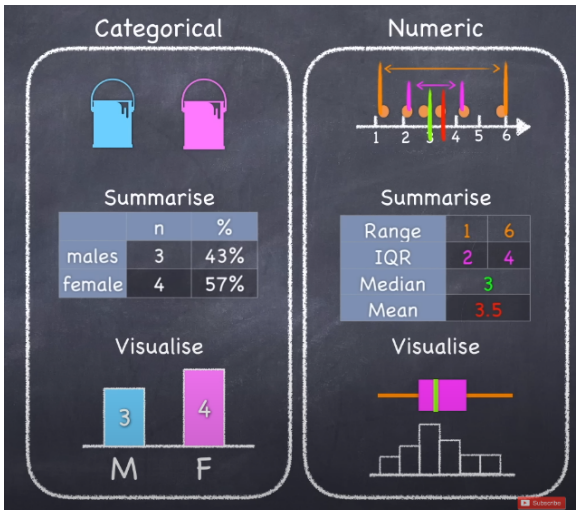
Hypothesis testing HT, null hypothesis NH, p-value PV https://youtu.be/S2eKynREGM4?t=437 , confidence interval CI



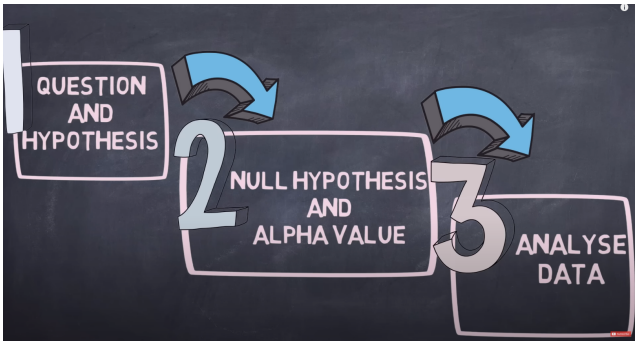
Explore by setting up multiple contingency tables!:
try plotline p9
-stats.describe(0,10).mean /.variance /.
Restrict to categorical nominal variables.

- When looking at sample data, one can generally see two things, differences between groups, like men weigh more than women, or relationships between variables, like is weight associated with height,

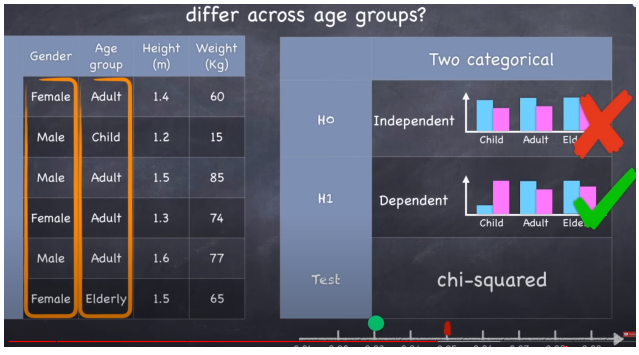


  - We can then run statistical tests to determine if those relationships are significant, real.
  - Depending on the variables and relationships one has several common statistical tests to choose from to test those relationships for signifigance.
  - because we can't test the whole population, we take a random sample of the population.
  - For our categorical variables we start by summarizing and visualizing. after checking the data for cleanliness.
  - For our numeric variables, we use things like IQR, mean,median, range, STDdev, and the distribution of the data.Visualize with IQR, use HIST for shape of data.
  - To see if what we are seeing in the sample data has implications about the larger population we perform a stat test to see if it is statistically significant, can we infer anything?

- https://youtu.be/I10q6fjPxJ0?t=520 chi square: hyp: the number of men women observed in each group is DEPENDENT on age group



  - not good science to stab around blindly hoping to find something statistically significant SS. We must first:
    - define research questions
    - define null hypothesis and alpha
      - H0/indie - there is/is no significant relationship.
      - Hypothesis1/dependent- the proportion of men women observed in each group is DEPENDENT on age (DEPENDS ON WHAT YOU SEE IN THE DATA FIRST!)

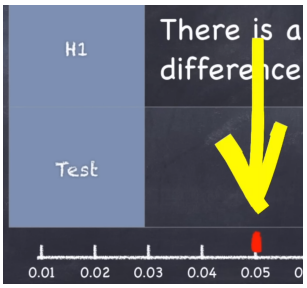


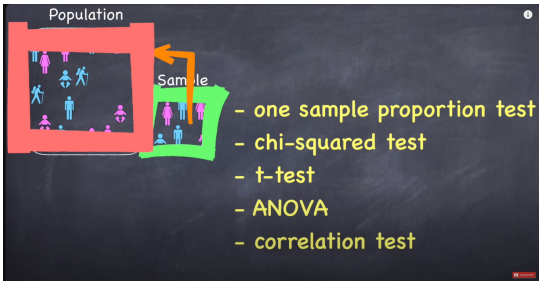
      - Well Let's "Test the idea that they are indie of age (H0)" readmin and other categorical variables.

  - "then analyze. Chi gives us p-value , if pV is less than alpha, we can reject the null. AND our observation is statistically significant.

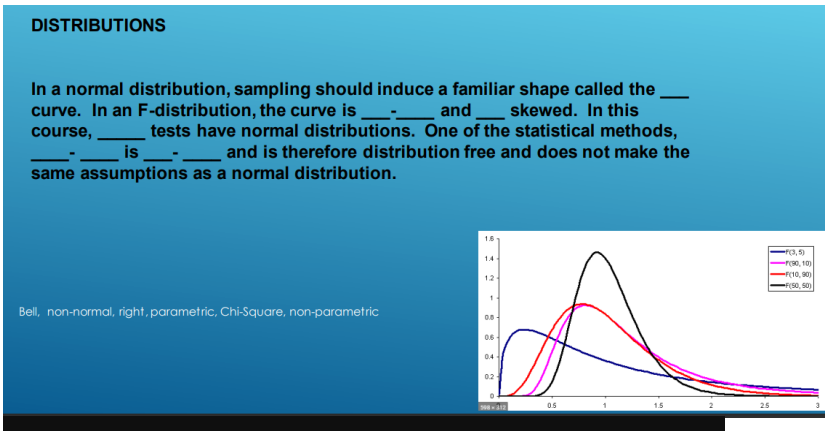
- CHECK:
    - IF the NULL HYPOTHESIS is TRUE, what is the probability that we would see what we saw in our random sample (H1 = what we saw)
    - If we can show that that probability is LOW, then we can have a degree of confidence that the null hypothesis is wrong and we can reject it.

- Before we calculate the probability, we must be clear about how small is small enough
- For this assessment, I will look for potential relationships amongst many cat variables and ReADmin, and select one to perform the chi_square2 test on for PA. purposes.
- *Q: Below What value of p (p-value) would we reject the null H0?* We must decide that cutoff before we calculate anything. We call that cutoff the *Alpha-Value.*





- For discussion, talk about how the test was Not exhaustive, then talk about doing other tests with other combinations of variables and other stat tests. drag it out.
- also talk about things like sample size, not knowing more info and other possibilities for error or confounding
- When we increase the sample size, we increase our statistical power, decrease the usable alpha value, and make it possible to detect smaller effects. In effect, we are asking what sample size is needed to pick up effect size x at statistical power y and significance level z.



anova

finding Chi-Square critical values

Using the following Contingency table, we want to calculate Chi-Square

Observed Table

	Regular Exercise	No Regular Exercise	TOTAL
Male	112	104	216
Female	96	88	184
TOTAL	208	192	400

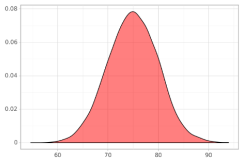
- **Start with Null: nothing to see here, nothing interesting, just random chance…..**
- **with enough evidence we can *reject the null* -**
- **otherwise we *fail to reject the null***

Two hypotheses

Null hypothesis

A = B

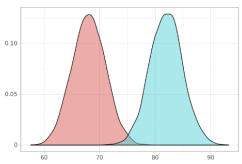
- Observed patterns are the product of random chance



Alternative hypothesis

A != B

- Difference between samples represents a real difference between the populations

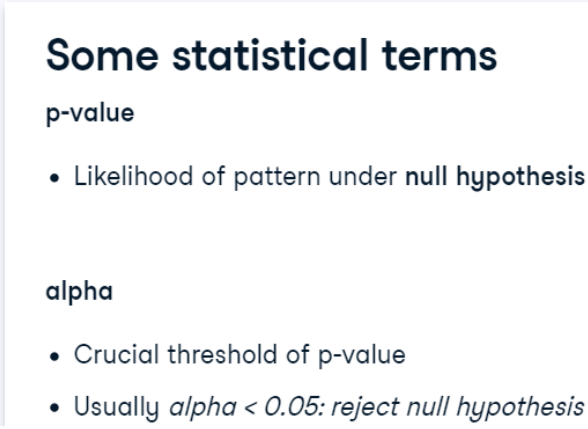


- https://campus.datacamp.com/courses/experimental-design-in-python/the-basics-of-statistical-hypothesis-testing?ex=5

To draw a conclusion, we will need to distinguish between two cases or hypotheses. In statistics, our starting point is the "null hypothesis": that there isn't anything interesting happening and the observed patterns are just the product of random chance. With enough evidence, we can reject the null hypothesis and turn to the more interesting "alternative hypothesis": that the difference between these samples represents a real difference between the populations.

But when do we know to reject the null hypothesis? Here we turn to two statistics. The p-value represents the likelihood that the distribution of values observed would occur if the null hypothesis were correct. We can't be 100 percent sure that our pattern couldn't have emerged due to random chance but we can quantify the probability that random chance would produce a given pattern; this is the p-value. The smaller the p-value is, the less likely it is that the null hypothesis can account for our observations. When p falls below a critical value, which we call alpha, we reject the null hypothesis. A standard value for alpha is 0 point 05. So, below a 5 percent probability that random chance would produce the pattern observed, it's usually considered safe to reject the null hypothesis.

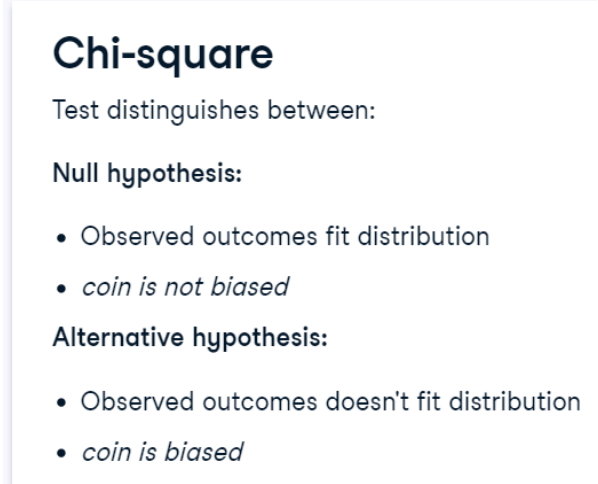
Student's t-test



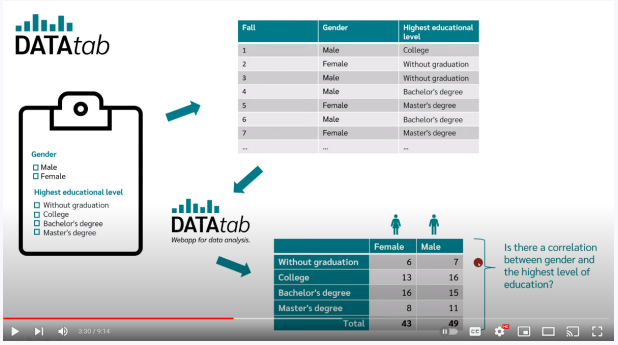
t-tests: compare means of continuous variables

Testing proportion and correlation

Chi_square: examine proportions of discrete categories



Explore by setting up multiple contingency tables!:



expected value mean / variance

Binomial distribution expected value and variance (Cont.)

What are the expected value and variance for one fair coin flip?

```
binom.stats(n=1, p=0.5)
```

```
(array(0.5), array(0.25))
```

What are the expected value and variance for one biased coin flip, with 30% probability of success?

```
binom.stats(n=1, p=0.3)
```

```
(array(0.3), array(0.21))
```

In this course …

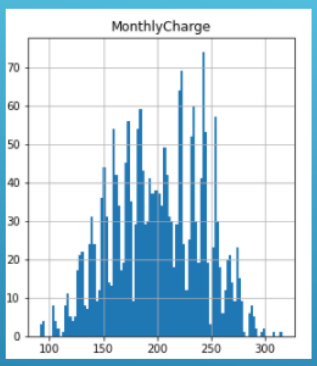
You have studied three statistical methods:
two are parametric: t-tests and ANOVA.
one is non-parametric: Chi-Square.

You must choose to examine a dataset and decide which method to use. Two require continuous data. The latter, Chi-Square, requires categorical.

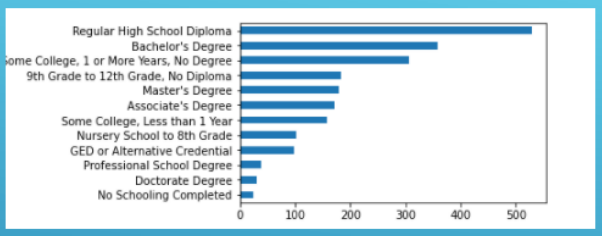
Success in this course is defined by your ability to analyze the data and results.

COURSE NOTES

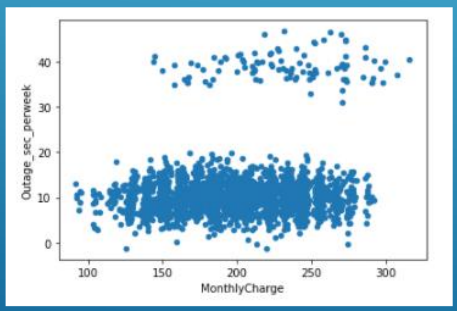
Monthly Charge in a histogram would appear like this:



The rubric requires that you display the distribution of the variables like this: For example, Education in a horizontal bar would look like this:



In a bivariate relationship, you could show a scatter plot like this:



Be sure to show the statistics for the variables as well. In R, that would be a summary() or in Python it would be a describe().

19. The Chi-Square Test (χ^2) has the following features:

A. It is a nonparametric test (distribution free). The data that it tests violate the assumptions of equal variance or homoscedasticity.

B. The difference between Chi-square and t-test is that a t-test tests a null hypothesis about two means to see if they are equal. The difference between them is zero. A chi-square test tests a null hypothesis about the relationship between two variables.

C. Based on frequencies and not on parameters like mean and standard deviation.

D. Used for testing the hypothesis and not for estimation.

E. The advantage over a Z test is that it can be applied to smaller samples as well as large samples.

F. All of these.

A Chi-Square Test Example in Python

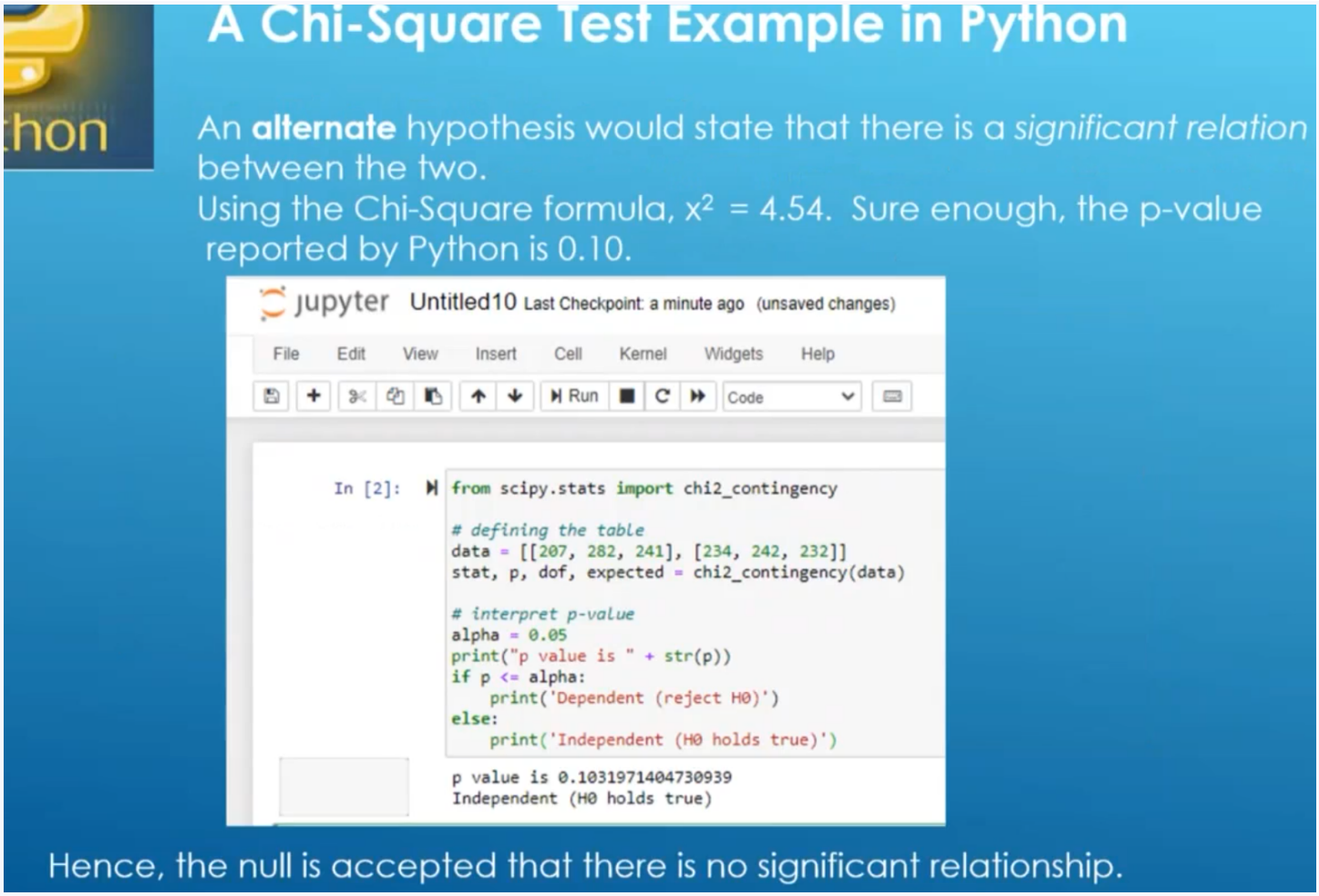
Buying habit of pets by gender
null hypothesis (**H0**) which states that there is *no relation* between the variables.
An **alternate** hypothesis would state that there is a *significant relation* between the two.
Using the Chi-Square formula, $x^2 = 4.54$

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

Critical values of the Chi-square distribution with *d* degrees of freedom

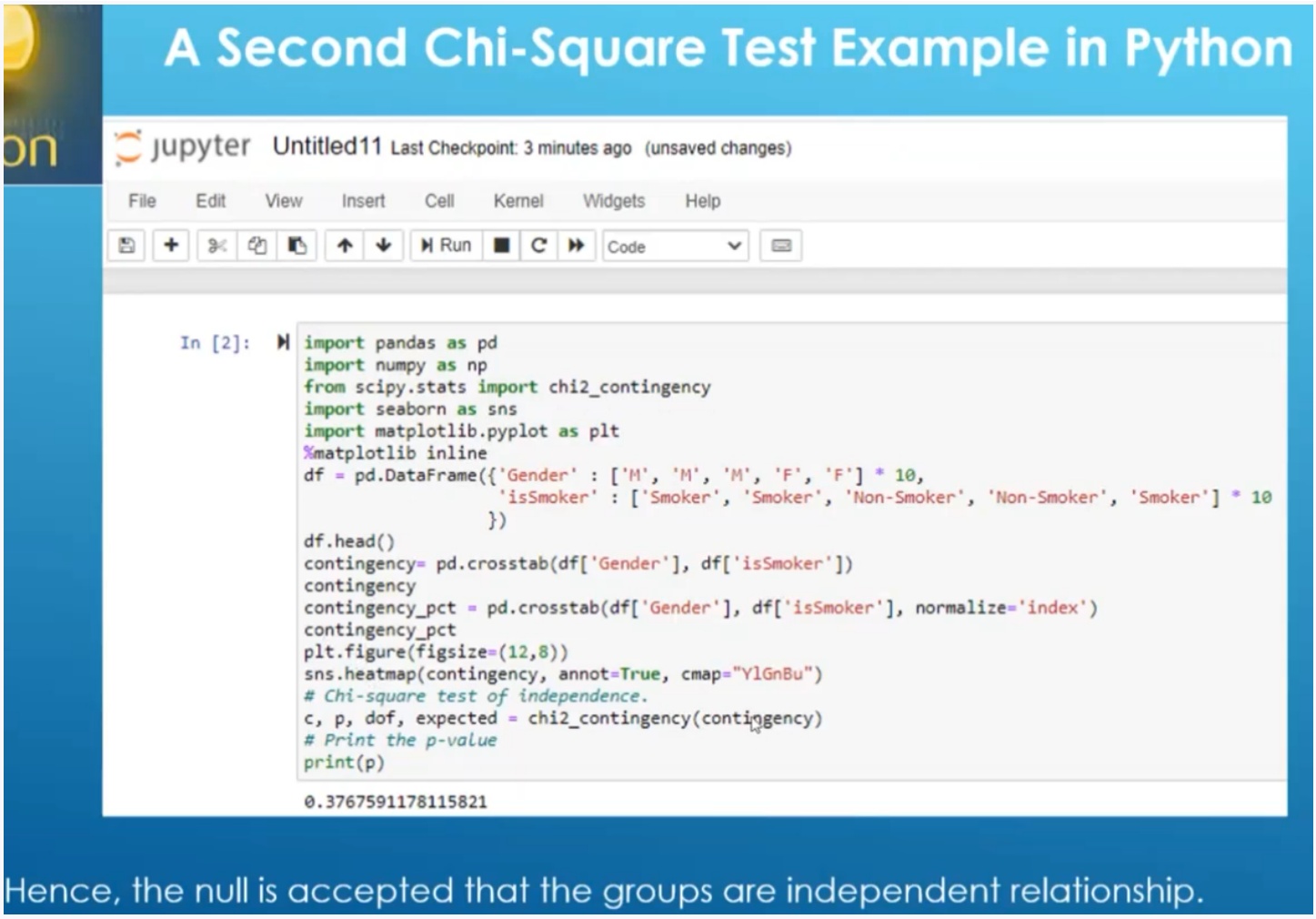
	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Hence, the null is accepted that there is no significant relationship.



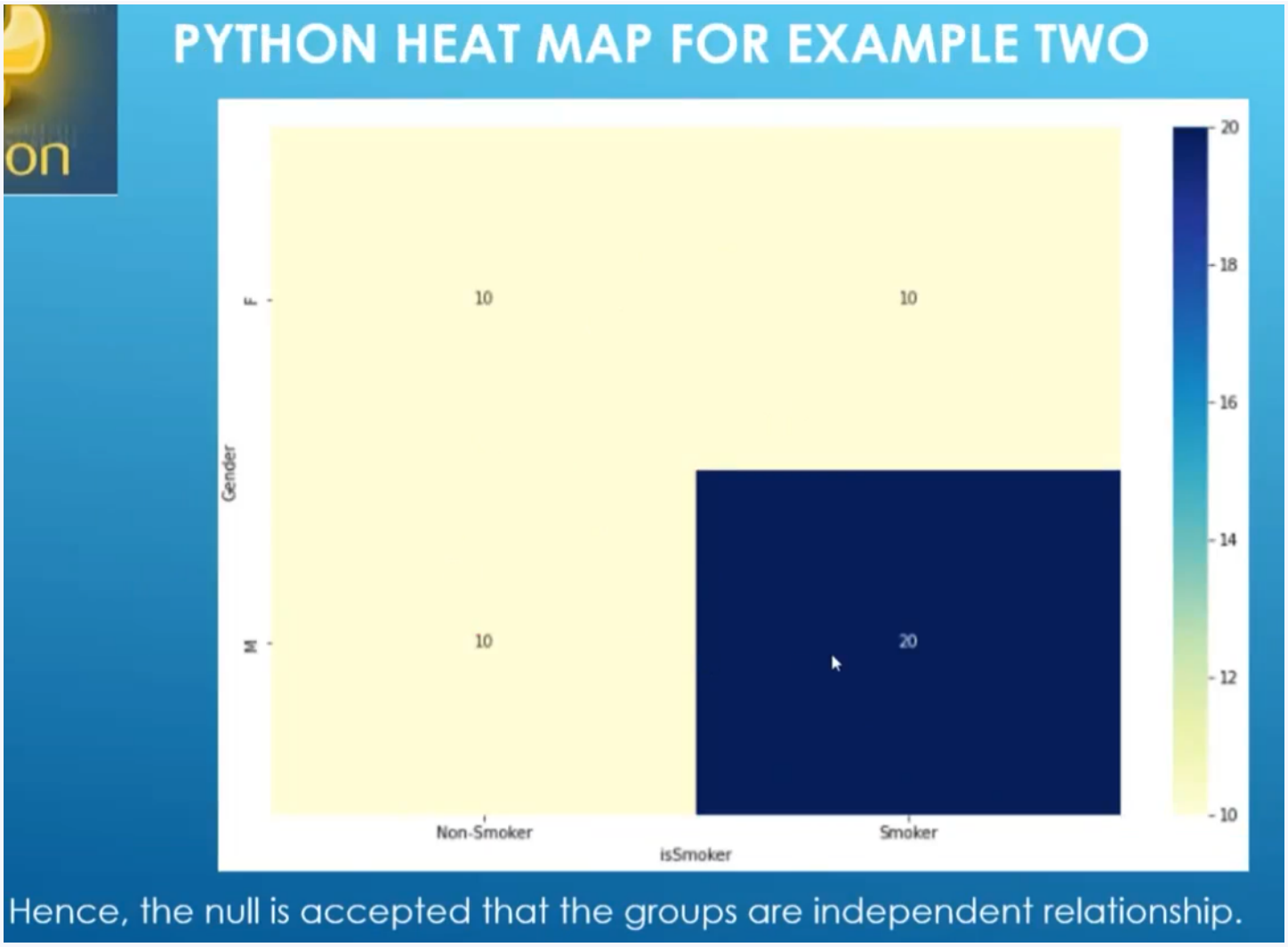
using contingency because 2 groups:

A Second Chi-Square Test Example in Python



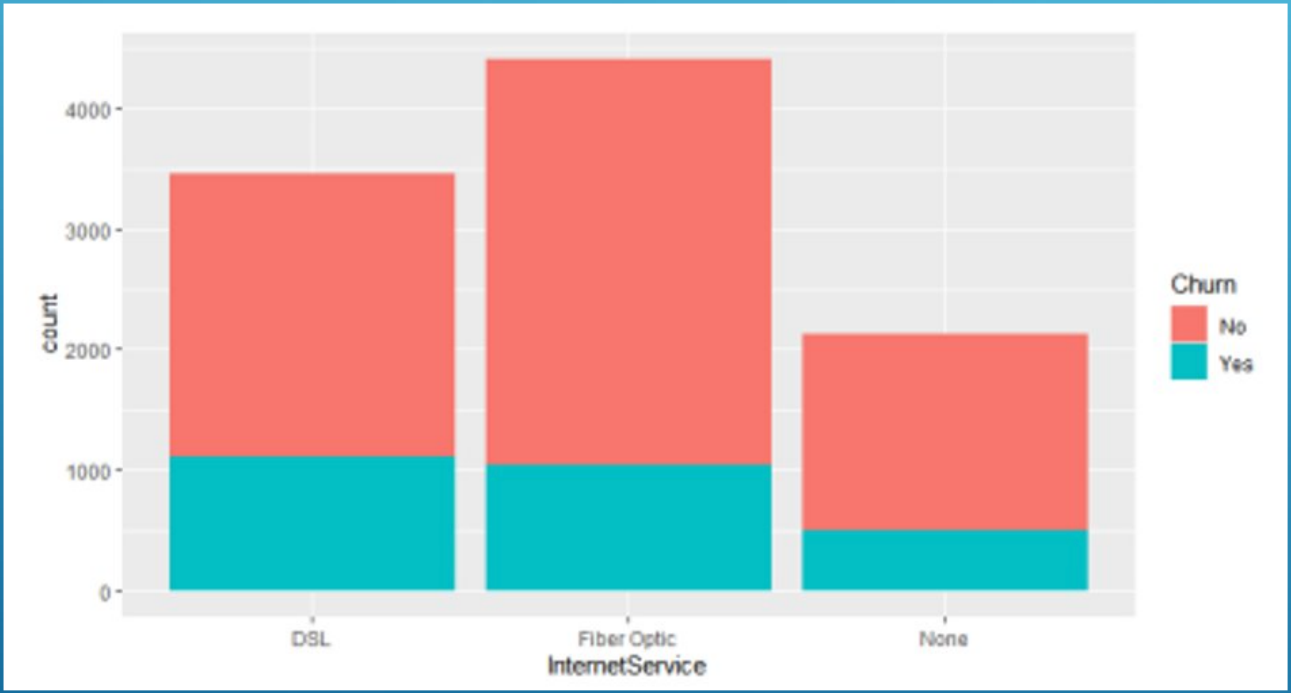
Hence, the null is accepted that the groups are independent relationship.

heatmap of above

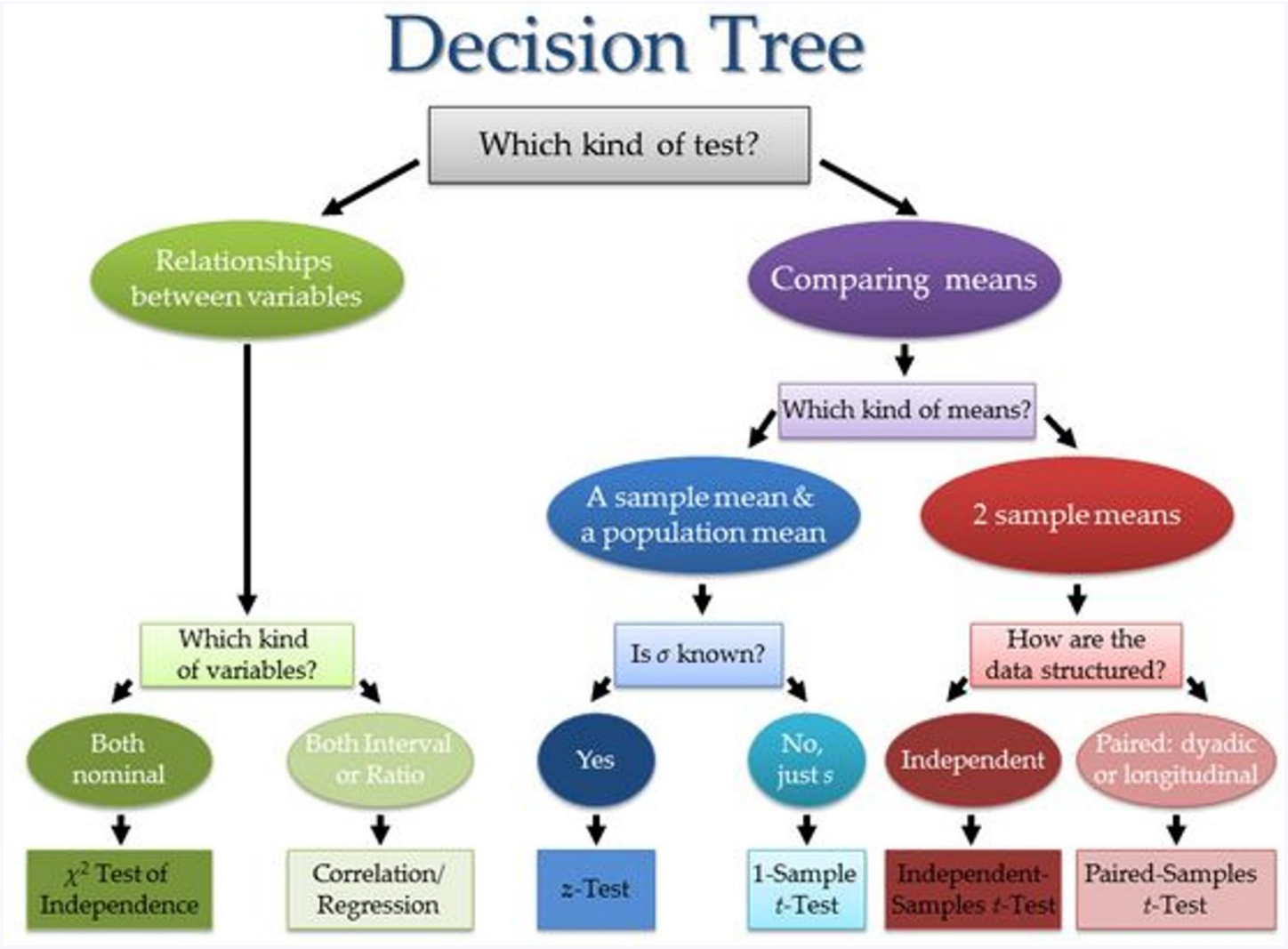


Bivariate Stacked Bar in R – categorical v. categorical

```
ggplot(df, aes(x = InternetService,
fill = Churn))
+ geom_bar(position = "stack")
```



Decision Tree



In this course …

You have studied three statistical methods:
two are parametric: t-tests and ANOVA.
one is non-parametric: Chi-Square.

You must choose to examine a dataset and decide which method to use. Two require continuous data. The latter, Chi-Square, requires categorical.

Success in this course is defined by your ability to analyze the data and results.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/#:~:text=The%20assumptions%20of%20the%20Chi,the%20variables%20are%20mutually%20exclusive