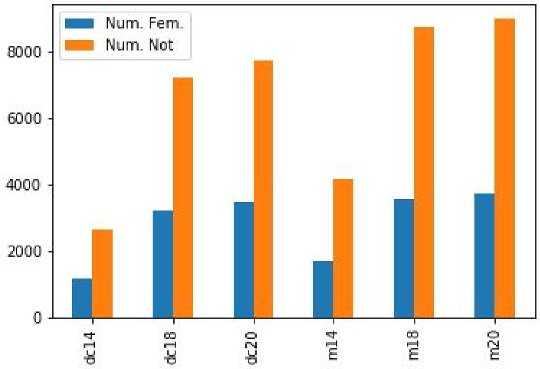
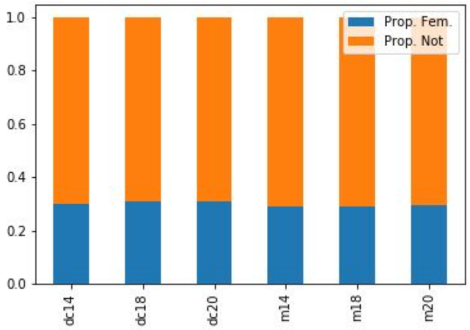
Ch. 10-12 Project: *Female Comic Characters*

Full project at: https://github.com/FrostForth/Comics

A longtime comics fan, Kaitlin decides to compare the proportion of female characters in DC and Marvel comics for her AP Statistics project. Using data from an article that scraped character data from the Marvel and DC wikis in 2014 as a model, she scrapes data for all characters on the DC and Marvel wikis in 2018 and in 2020, resulting in six separate groups. A sample of characters with more than five appearances was selected from each group and the number of female characters was recorded for each. The resulting data are shown in the table below.

	dc14	**dc18**	**dc20**	**m14**	**m18**	**m20**
Female	1153	3213	3484	1691	3546	3720
Not	2672	7201	7759	4189	8739	9009





a) Based on the graphs, does it appear that the proportion of female characters differs between the groups? Explain your answer.

b) Using significance level of 0.05, conduct an appropriate test to determine if the distribution of female characters differs across the scrapes. Assume all conditions are met.

c) Based on the result from part a and the table of contributions shown below, would it be more appropriate to perform chi-squared tests for each publisher or to perform z tests for differences in proportions for each year as follow-up analysis? Explain your reasoning.

Kaitlin decides to perform all five additional tests and finds that the p-value for the DC proportions is 0.615, the p-value for the Marvel proportions is 0.587, the 2014 p-value is 0.1429, the 2018 p-value is 0.0011, and the 2020 p-value is 0.003.

d) Based on these results, what conclusions about the data can be made?

DESIGN

a) Why were you interested in this question?

I was inspired by a FiveThirtyEight article I found when I was first starting out in data science a few years ago that analyzed diversity in characters in Marvel and DC Comics. I found the article and the data it used particularly interesting at the time but did not know how to conduct my own research on the topic so I was unable to do much with the additional data I collected. Now that I have a better understanding of statistics, however, I have decided to perform my own analysis on the topic for this project.

b) How did you collect the data? Be specific. You should have randomization of some sort. Explain.

Each scrape collected data on all characters on the respective wiki pages on the given date. The first scrape, on August 24, 2014, was linked to the article and I performed the additional scrapes December 27, 2018 and March 26, 2020. Once this data was collected, I took a sample of all characters with more than five appearances. This is to reduce bias resulting from characters with multiple entries due to their presence in multiple universes. The alternate versions of these characters are more likely to have appeared in special series or one-offs and would therefore be not as likely to be in the sample. Also, since the wiki pages are user-submitted less popular characters with fewer appearances may not have accurate data. From these samples, the number of female characters was recorded.

c) What is the explanatory variable? What is the response variable?

The explanatory variables are the date of the scrape and the comic publisher, while the response variable is the proportion of female characters in the sample.

d) If you did an experiment, explain the design of this study: completely randomized, randomized block, or matched pairs. If you did an observational study, was it prospective or retrospective? Explain.

Since the study period does not extend into the future and there are no experimental groups, this is a retrospective observational study.

e) What are possible limitations in your design and how could you address these? (“increase the sample size” is NOT what I’m looking for --- every group can say this)

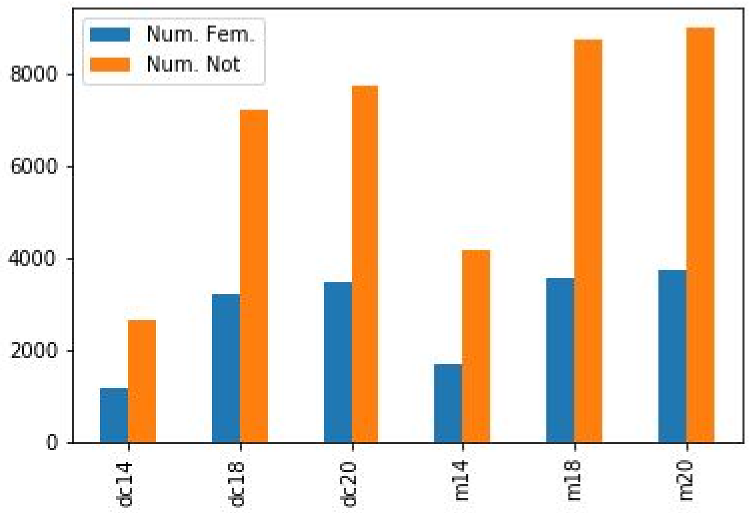
Since the samples were not randomly sampled, it may not be truly possible to infer the study's results to the population. However, random sampling cannot be used in this case, as many of the samples would have overlapping data. This also means that the individuals in the dataset from the same publisher are not independent. Therefore, we must merely assume that each sample of characters with more than 5 appearances is representative of all characters in their respective population.

DATA

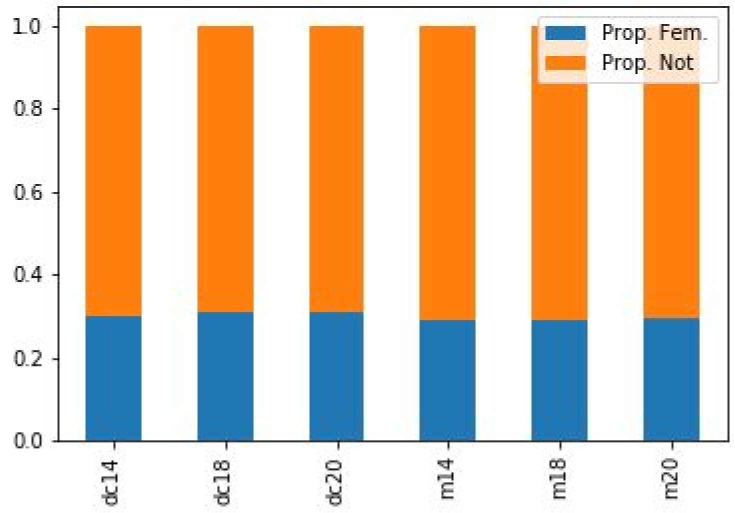
f) Include the **raw data** you collected (in an organized manner)

	Prop.	Num. Fem.	Num. Not	Total > 5	Total
dc14	0.301437908496732	1153	2672	3825	6896
dc18	0.30852698290762437	3213	7201	10414	24818
dc20	0.3098817041714845	3484	7759	11243	27071
m14	0.28758503401360547	1691	4189	5880	16376
m18	0.2886446886446886	3546	8739	12285	60276
m20	0.29224605232147066	3720	9009	12729	67022

g) Provide appropriate graphical displays of your data using technology (stapplet.com or artofstat.com). Describe what you see.

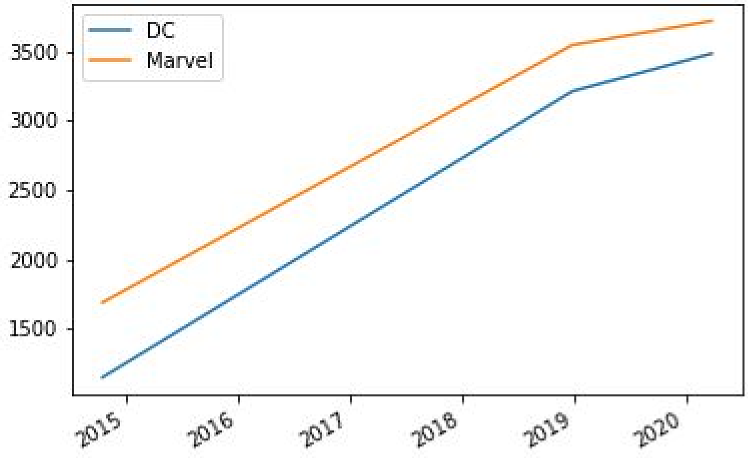


- All three Marvel groups have a higher total frequency of characters with more than five appearances
- Although the frequency of nonfemale characters in groups m18 and m20 are greater than the those in groups dc18 and dc20, the frequency of female characters is roughly similar across all four groups
- For both publishers, the frequency of nonfemales increases more between each scrape than the proportion of females



- All groups seem to have proportions close to .3
- For both publishers, the proportions seem to increase slightly over time

- The proportions in all three Marvel groups seem to be lower than the proportion in dc14



First scrape date: 10/13/2014

Second scrape date: 12/27/2018

Third scrape date: 3/26/2020

- The increase between the 2014 scrape and the 2018 scrapes are greater than the increase between the 2018 and 2020 scrapes for both publishers
- The amount of time between the first and second scrapes is about 3.5 times longer than between the second and third scrapes
- The increase in frequency between the first and third scrapes may not be linear but we do not have enough data to make a conclusion

h) Based on the graphs of the study results (the raw data) alone, do you think we will find statistically significant evidence for the alternative hypothesis? Explain.

There seems to be initial evidence that the proportion of female characters varies between each sample, as well as between each publisher and scrape.

SIGNIFICANCE TEST

i) State the hypotheses for this significance test. Define the parameter of interest.

Parameter: p_i = the true proportion of female characters from scrape i where i = [dc14, dc18, dc20, m14, m18, m20]

H_0: There is no difference in proportions between the six populations

H_A: At least one proportion is significantly different between the six populations

j) What is your initial evidence for your alternative hypothesis?

The raw data and graphs suggests differences between the groups.

k) What are two possible statistical explanations for the statistic (the initial evidence) you observed?

1. The observed differences in proportions over time and between publishers is a result of random chance
2. There really is a difference in at least one proportion

l) What type of significance test will you perform?

Initially a chi-squared test for homogeneity will be used, but follow up z interval tests for proportions will be used if necessary.

If evidence is found for the alternative hypothesis, follow-up analysis will be conducted. First, I will conduct a chi-squared test for homogeneity on the three samples from each publisher. If either of those tests succeed, I will perform follow-up z tests for differences in proportions. Additionally, I will conduct a z test for differences in proportions for each scrape date.

m) Would a confidence interval be appropriate for this study? If yes, include it, an interpretation of the interval, and an interpretation of the level. If an interval is not appropriate, explain why not.

No. Since the study involves more than two samples, it would not be appropriate to use an interval to approximate the difference in population proportions.

n) Perform the appropriate test of significance at the 5% level.

Data:

	dc14	**dc18**	**dc20**	**m14**	**m18**	**m20**
Female	1153	3213	3484	1691	3546	3720
Not	2672	7201	7759	4189	8739	9009

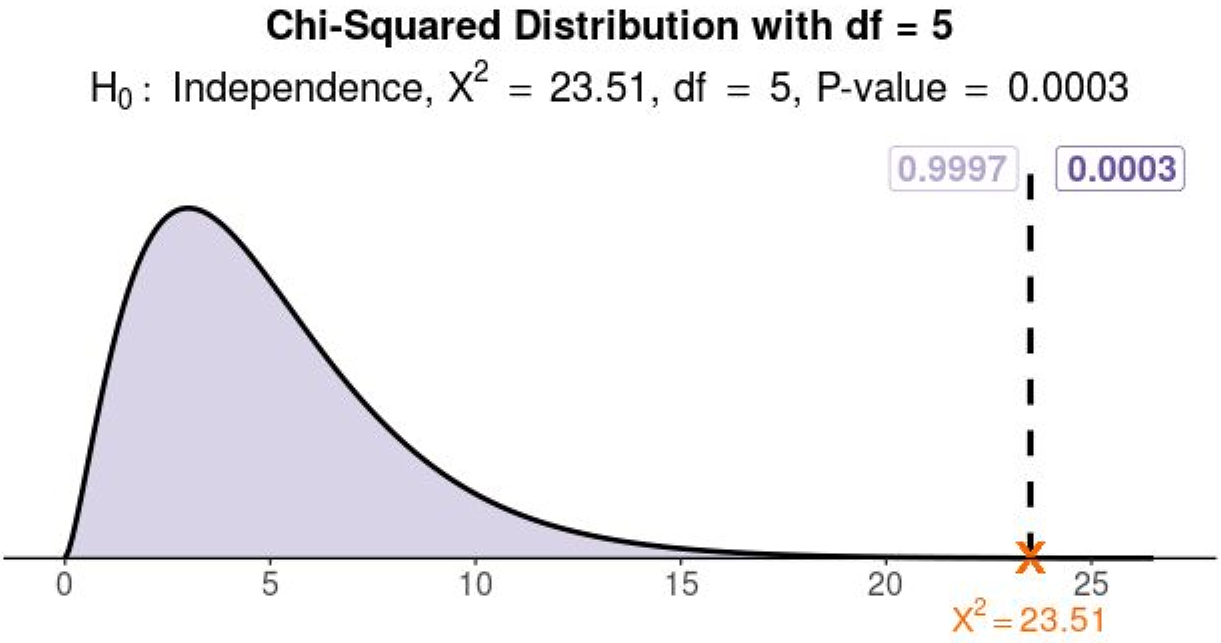
Exp:

	dc14	**dc18**	**dc20**	**m14**	**m18**	**m20**
Female	1140.3	3104.7	3351.8	1753.0	3662.4	3794.8
Not	2684.7	7309.3	7891.2	4127.0	8622.6	8934.2

Conditions:

alpha = 0.05

1. We assume each sample is representative of its respective population
2. All expected counts are > 5
3. We assume each observation is independent of other observations in each sample



chi-squared = [(1153 - 1140)^2]/ + [(2672 - 2685)^2]/ + ... = 23.513

p-value = 0.0002692

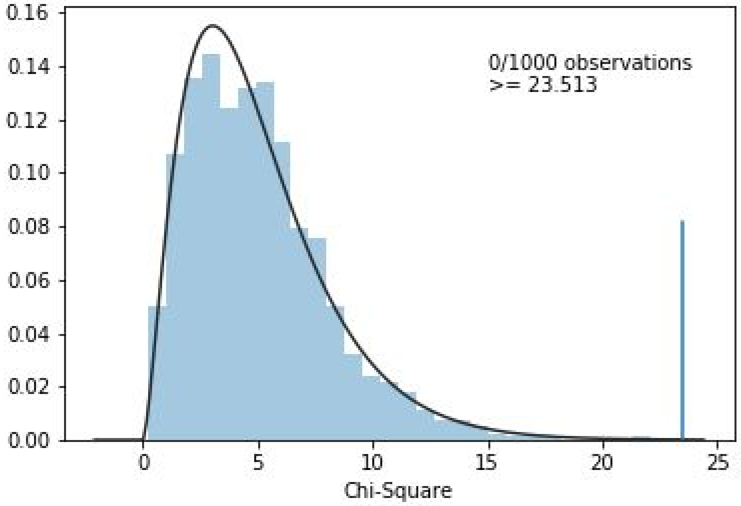
Assuming there is no difference in proportions between the six populations, there is a 0.00027 chance that the observed chi-squared value of 23.513 would occur by random chance.

Since 0 < .05, we reject the null hypothesis and conclude that we have evidence that there is a significant difference in the proportion of female characters in at least one population.

o) Interpret the p-value in context.

SIMULATION

p) Use one of the applets under “Statistical Inference” on http://www.rossmanchance.com/applets/index.html to run a simulation to estimate the p-value (use at least 500 trials under the “shuffle” option). State the p-value and include a screenshot of the graph of your 500 trials. See me if you need assistance.



p = 0

FOLLOW UP

Assuming all conditions are met for all tests.

Cont:

	dc14	**dc18**	**dc20**	**m14**	**m18**	**m20**
Fem	0.141	3.781	5.214	2.190	3.702	1.475
Not	0.060	1.606	2.215	0.930	1.572	0.626

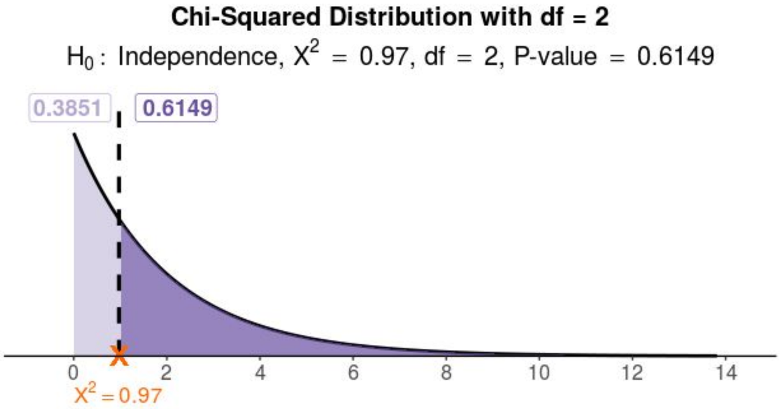
The three cells with the highest contributions are:

1. dc2020 female
2. dc18 female
3. m18 female

This suggests that the most successful tests will be the DC chi-squared test and the z tests for 2018 and 2020.

- **Additional chi-squared by publisher**

DC:

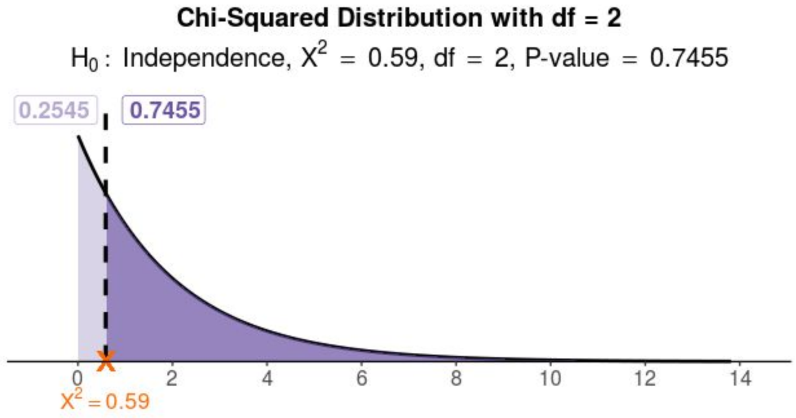


chi-squared = 0.9726

p = 0.615

Since 0.615 > .05, we fail to reject the null hypothesis and conclude that we do not have evidence that the proportion of female characters in DC Comics has changed significantly between the scrapes.

Marvel:



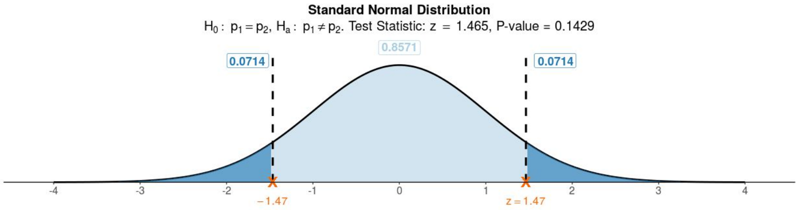
chi-squared = 0.587

p = 0.7455

Since 0.7455 > .05, we fail to reject the null hypothesis and conclude that we do not have evidence that the proportion of female characters in Marvel Comics has changed significantly between the scrapes.

- **Additional Z tests by scrape**

2014:

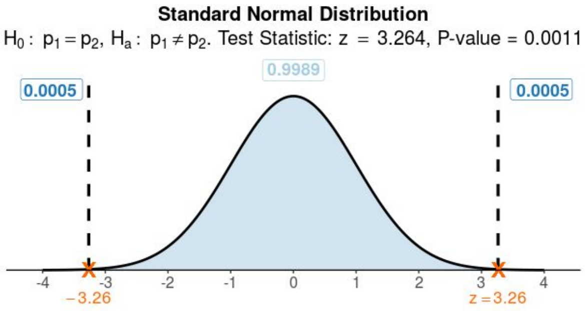


z = 1.465

p = 0.1429

Since .143 > .05, we fail to reject the null hypothesis and conclude that we do not have evidence that the true proportion of female characters was significantly different between the two publishers in the 2014 scrape.

2018:

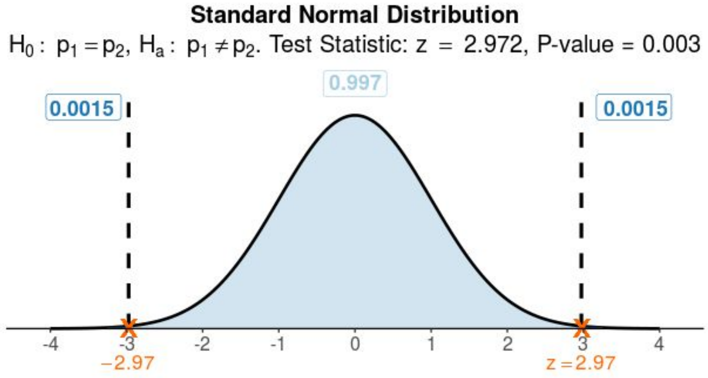


z = 3.264

p = 0.0011

Since .0011 < .05, we reject the null hypothesis and conclude that we have evidence that there is a significant difference in the true proportion of female characters between the publishers in the 2018 scrape.

2020



z = 2.972

p = 0.003

Since .003 < .05, we reject the null hypothesis and conclude that we have evidence that there is a significant difference in the true proportion of female characters between the publishers in the 2020 scrape.

CONCLUSION

q) Conclude the significance test in context based on the P-value from the test (not the simulation).

The initial test found evidence of at least one sample's proportion of female characters being statistically significantly different from the six populations. Upon further analysis, we discovered that there was not statistically significant evidence that the proportion had changed over time for either publisher. However, there is statistically significant evidence that the proportion of female characters is different between the publishers in both 2018 and 2020, but not in 2014.

r) What type of error—Type I or Type II—could you have made? Describe it in context and a possible consequence.

A type I error may have occurred in this test. This would mean that we found evidence that at least one proportion is significantly different when in reality, the proportions were not significantly different. This would lead to unnecessary follow-up analysis.

s) To what population can we generalize the results?

We should be able to generalize the results to all characters recorded in the Marvel and DC wikis.

t) Can we infer cause and effect?

Since this is an observational study and not an experiment, we cannot infer a cause-and-effect relationship.

u) If a group next year wanted to do a similar project as this one, what three pieces of advice would you give them?

1. Set up a schedule and pace the steps of the project appropriately
2. Do research on what needs to be built beforehand, such as programs to collect the data and perform the simulation
3. Choose data that is independent

v) What could be improved in the study?

Since most of the data is cumulative over time, using the proportions between scrapes may have impacted the data and contributed to the lack of evidence in both additional chi-squared tests. A better test of the proportion of female characters over time for each publisher may be using the first appearance dates for each character to create a regression line. By plotting the number or proportion of female characters in each month or release cycle, a clearer association between the proportion of female characters and time may emerge. However, this data is the most commonly incomplete section on the wiki pages and the most difficult to scrape, I did not use this method in this study.