3251 Foundations of Data Science

Data Science Fundamentals Certificate

Module 5

INFERENCE OF CATEGORICAL VARIABLES

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference Part 1
5	Inference Part 2
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 5: Learning Objectives

- Inference for Categorical Data

Key Topic Overview – Inference

- Inference for Categorical Data

Required/Recommended Readings

Required:

1) OpenIntro, 3rd edition, by David Diez, Christopher Barr, Mine Centinkaya-Rundel, Copyright © 2015 OpenIntro.org

Recommended:

1) Think Stats: Probability & Statistics for Programmes Version 1.6.0, by Allen Downey, Copyright © 2011 Green Tea Press

2) Using R for introductory statistics, John Versani, Chapman and Hall/CRC Press

Section sub-header

INFERENCE OF CATEGORICAL DATA

Inference of Categorical Data

- Categorical variables represent types of data which may be divided into groups
- Examples of categorical variables are race, sex, age group, and educational level
- While the age group and education level may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups
- Analysis of categorical data generally involves the use of data tables
- A two-way table presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns

Two-way frequency table

	Pool	Video Games(Inside)	Total
Boys	4	8	12
Girls	10	3	13
Total	14	11	25

Two-way frequency table with respect to the total

	Pool	Video Games(Inside)	Total
Boys	0.16	0.32	0.48
Girls	0.40	0.12	0.52
Total	0.56	0.44	1.00

Inference of Categorical Data

- The analysis of categorical data generally involves the proportion of “observations" in a given population.
- Given a simple random sample of size n from a population, the number of “observations" x divided by the sample size n provides the sample proportion, which is an estimate of the population proportion
- The distribution of the sample proportion follows a binomial distribution
- By the CLT, the binomial distribution is approximately normal for large sample sizes (checked using the condition: $np \geq 10$ and $n(1-p) \geq 10$) with mean p and variance $(p(1-p))/n$

Inference of Categorical Data

- To calculate the test of significance and confidence interval for a single proportion, we use a *z*-statistic
- The standard error calculation for the confidence interval and the hypothesis test are different when dealing with proportions
  - For confidence intervals

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - For hypothesis tests

$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

  - where $\hat{p}$ is the observed sample proportion and p_0 is the null probability
- Such a discrepancy does not exist when conducting inference for means, since the mean does not factor into the calculation of the standard error

Inference of Categorical Data

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

The 1% level of significance means that $\alpha = 0.01$

H_0: $p = 0.50$; H_a: $p \neq 0.50$. This is a 2-tailed test.

p' = the percent of first-time brides who are younger than their grooms.
The problem contains no mention of an average, only percentages.
Use the distribution for p', the estimated proportion.
$p' \sim N(p, \sqrt{p \cdot (1-p)/ n})$; therefore, $P' \sim N(0.5, \sqrt{0.5 \cdot 0.5 /100})$ where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$.

$\mu = p = 0.50$ comes from H_0, the null hypothesis
p'= 0.53. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

- Using p-value calculator with $p = 0.50$, $\sqrt{p \cdot (1-p)/ n}=0.05$, $p'=0.53$, we find p-value = 0.27425; since 2-sided, therefore p-value=0.5485
- Since $\alpha = 0.01$ and p-value = 0.5485. Therefore, $\alpha <$ p-value.
- Since $\alpha <$ p-value, we cannot reject H_0

Inference of Categorical Data

- The current shampoo kills 25% fleas. In order to test a new shampoo we test how many fleas are killed from an initial known population: kills 17 fleas out of 42 p'= 17/42 = 0.4048
- Level of significance 0.01. Use the new shampoo or not?
- H_0: p = 0.25; H_a: p > 0.25
- p' = The proportion of fleas that are killed by the new shampoo
- Distribution to use for the test: N(0.25, $\sqrt{(0.25)(1-0.25)/42}$)
- z-statistic = (0.4048-0.25)/ 0.0668 = 2.3168
- Calculate the p-value using the normal distribution for proportions: p-value = 0.0103
- If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 or more.
- At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% Confidence Interval for the experimental mean or proportion:

- Using $SE_{p'} = \sqrt{p'(1-p')/n} = \sqrt{0.4048\ (1-0.4048)\ /\ 42} = 0.0757$
- 2 SE away from 0.4048 = [0.234,0.5562] for the 95% confidence interval

Choosing a sample size when estimating a proportion

- The margin of error for a sample proportion is

$$z^\star \sqrt{\frac{p(1-p)}{n}}$$

- How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} < 0.04$$

- There are two unknowns, p and n. If we don't have an ideal of p the practice is to use 0,5 (worst case)
- We will need a sample of at least 600

Inference of Categorical Data

- Comparing two proportions, like comparing two means, is common
- If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance
- A hypothesis test can help determine if a difference in the estimated proportions ($P'_A - P'_B$) reflects a difference in the populations
- The difference of two proportions follows an approximate normal distribution
- Generally, the null hypothesis states that the two proportions are the same, $H_0 : p_A = p_B$.
- To conduct the test, we use a pooled proportion

$$p_{pool} = \frac{x_A + x_B}{n_A + n_B}$$

Inference of Categorical Data

- The z-statistic is

$$z = \frac{\left(p'_A - p'_B\right) - \left(p'_A - p'_B\right)}{\sqrt{p_{pool}\left(1 - p_{pool}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Inference of Categorical Data

- Two types of medication for hives are being tested to determine if there is a difference in the percentage of adult patient reactions.
- 20 out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication.
- 12 out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

- $H_0 : p_A = p_B$ or $p_A - p_B = 0$
- $H_a : p_A \neq p_B$ or $p_A - p_B \neq 0$
- The test is two-tailed

- $p_C = X_A+X_B \,/\, n_A+n_B = 20+12 \,/\, 200+200 = 0.08$; $1 - pc = 0.92$
- Therefore, $p'_A - p'_B \sim N(0, \sqrt{(0.08) \cdot (0.92) \cdot (1/200 + 1/200)})$

- $p'_A - p'_B$ follows an approximate normal distribution
- Calculate the p-value using the normal distribution: p-value = 0.1404

- Estimated proportion for group A: $p'_A = X_A/n_A = 20/200 = 0.1$
- Estimated proportion for group B: $p'_B = X_B/n_B = 12/200 = 0.06$
- $p'_A - p'_B = 0.1 - 0.06 = 0.04$

Inference of Categorical Data

Calculate the p-value from the normal distribution calculator using mean (of $p_A - p_B$)=0;
$SE_{P'A - P'B}=\sqrt{(0.08) \cdot (0.92) \cdot (1/200 + 1/200)} =0.02713$;
$p'_A - p'_B = 0.1 - 0.06 = 0.04$,

We obtain $P(X \leq 0.04) = 0.92981$.
Therefore, p-value = 2 * (1-0.92981) because it is a 2-tailed test.
Half the p-value is above 0.04 and half is below -0.04.
p-value= 0.14038

$\alpha = 0.01$ and the p-value = 0.1404.
$\alpha <$ p-value.
Since $\alpha <$ p-value, we cannot reject H_0.

Inference of Categorical Data

- The reason for the difference in calculations of standard error is the same as in the case of the single proportion: when the null hypothesis claims that the two population proportions are equal, we need to take that into consideration when calculating the standard error for the hypothesis test, and use a common proportion for both samples

Inference of Categorical Data

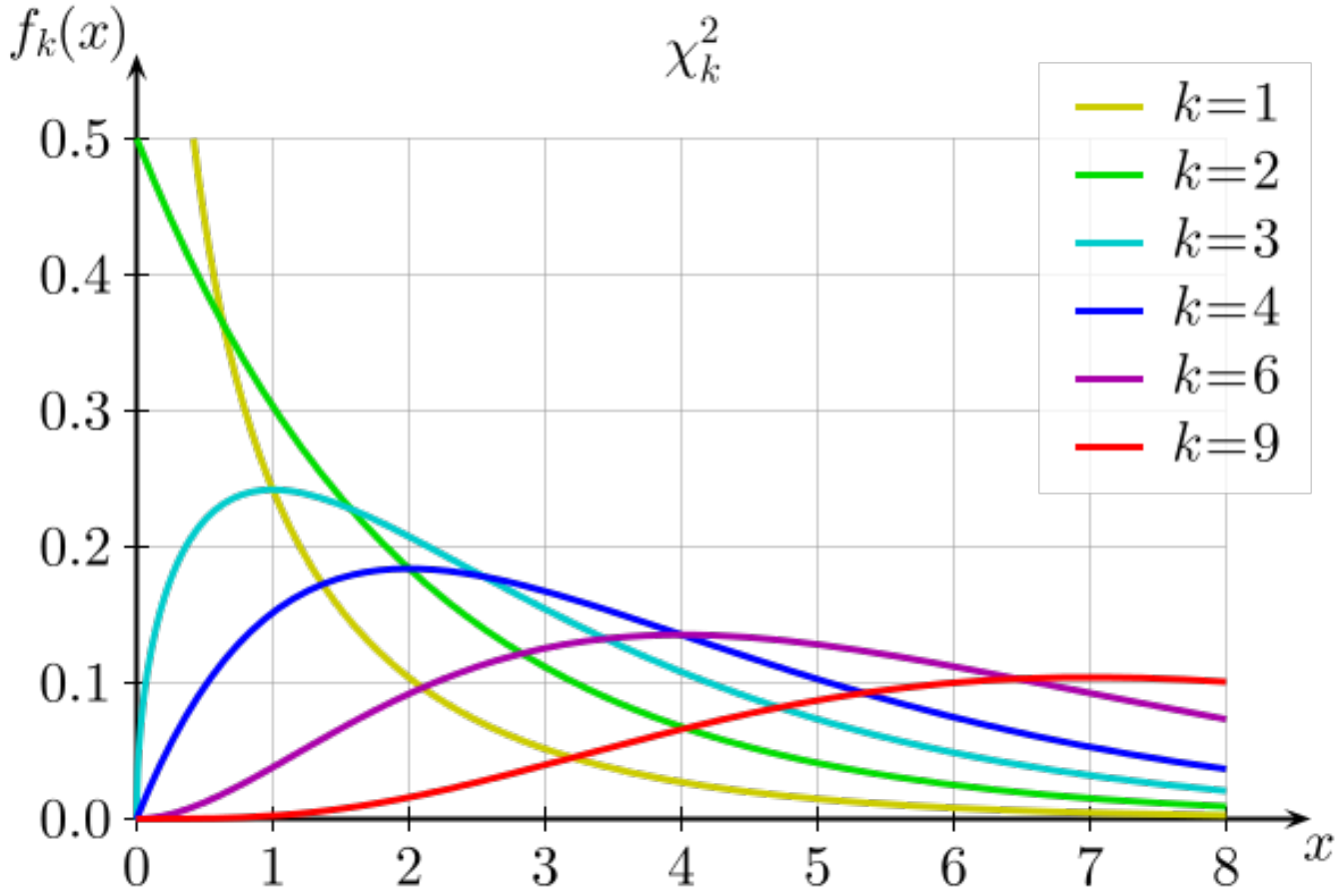
- Let's turn our attention to 2 categorical variables from a single population and we want to assess if there is significant association between the 2 variables
  - H_0 : The two variables are independent.
  - H_a : The two variables are dependent.
- For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference

X^2 (Chi Square) Distribution

- If Z has a standard normal distribution then by definition Z^2 has a X_1^2 distribution, i.e. a chi-square distribution with one degree of freedom.
- If Z_1, …, Z_k are independent random variables, each with a standard normal distribution, then by definition $Z_1^2 + Z_2^2 + \cdots + Z_k^2$ has a distribution X_k^2.
- When testing for proportions it is common to construct the chi-square as:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

X^2 (Chi Square) Distribution



By Geek3 - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9884213

Inference of Categorical Data

- When to use the chi-square test for independence testing?
  - The sampling method is simple random sampling
  - The variables under study are each categorical
  - If sample data are displayed in a contingency table (2-way table), the expected frequency count for each cell of the table is at least 5

Testing for Goodness of Fit using Chi-Square

- A goodness of fit test is a stipulation concerning the expected pattern of frequencies in a set of categories
- The expected pattern may conform to the assumption of equal likelihood and may therefore be uniform, or the expected pattern may conform to such patterns as the binomial, Poisson, or normal.
- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population
- In general the chi-square value for testing the difference between obtained and expected patterns of frequencies is:

$$\chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e}$$

Inference of Categorical Data

- We can also use a chi-square test of goodness of fit to evaluate if the distribution of levels of a single categorical variable follows a hypothesized distribution.
  - H_0 : The distribution of observed counts follow the hypothesized distribution, and any observed differences are due to chance
  - H_A : The distribution of observed counts do not follow the hypothesized distribution
- Calculate the expected counts for a given level (cell) in a one-way table as the sample size times the hypothesized proportion for that level
- Calculate the chi-square test statistic

$$\chi = \sum_{i=1}^{k} \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

- with k = number of cells and degrees of freedom (df) = k-1

Example

Determine if the sample is representative of the general population:

- Consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, and we would like to determine if these jurors are racially representative of the population.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Example Cont.

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_a: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

Calculate the statistics: (The standard error for the point estimate of the count in binned data is the square root of the count under the null)

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54 \qquad Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39 \qquad Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

Example Cont.

- Calculate the statistic:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

- Find the critical value: there are k = 4 categories: white, black, Hispanic, and other. According to the rule above, the test statistic $\chi 2$ should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true
- From tables we obtain that $p \sim 0.15$, we fail to reject the null, the data do not provide convincing evidence of racial bias in the juror selection

Testing for independence in two-way tables

- Google might test three algorithms using a sample of 10,000 google.com search queries
- H_0: The algorithms each perform equally well.
- H_a: The algorithms do not perform equally well.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Example Cont.

- Expected counts in two-way tables
- We estimate the proportion of users who were satisfied with their initial search (no new search) as 7078/10000 = 0.7078.
- If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, about 70.78% of the 5000 would be satisfied with the initial search: 0.7078 × 5000 = 3539 users

Search algorithm	current		test 1		test 2		Total
No new search	3511	(3539)	1749	(1769.5)	1818	(1769.5)	7078
New search	1489	(1461)	751	(730.5)	682	(730.5)	2922
Total	5000		2500		2500		10000

Example Cont.

- We did:

$$0.7078 \times (\text{column 1 total}) = 3539$$
$$0.7078 \times (\text{column 2 total}) = 1769.5$$
$$0.7078 \times (\text{column 3 total}) = 1769.5$$

- In general:

$$\text{Expected Count}_{\text{row } i,\ \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

Example Cont.

- Calculate the statistics:

General formula	$\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\frac{(3511-3539)^2}{3539} = 0.222$
Row 1, Col 2	$\frac{(1749-1769.5)^2}{1769.5} = 0.237$
$\vdots$	$\vdots$
Row 2, Col 3	$\frac{(682-730.5)^2}{730.5} = 3.220$

$$\chi^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120$$

- For two way tables, the degrees of freedom:

$$df = (R-1) \times (C-1)$$

Example Cont.

- The test statistic, $\chi 2 = 6.120$ yields a p-value is between 0.02 and 0.05. Then the null hypothesis is rejected at a significance of 0.05. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Assignment #4

- Section 6.5 of the textbook covers how simulation can assist when one of the assumptions / requirements of the Chi-Square distribution are not met.
- The text goes through a step by step explanation.
- The assignment is to follow the steps using Python. Results are given so you will now if you are in the right direction.

Questions?