INTRODUCTION TO STATISTICS

Session 12

July 3rd 2017

Madoka Takeuchi

Review

- T-distribution- similar properties to the normal distribution
  - What is different? How do we look up the critical value/p-value for a t-distribution?
- 2 sample test- extension of a one sample test
  - Use a t-test
  - Pooled vs non-pooled variance
    - When do we pool the variances??

SPSS

- Until now we have done hypothesis testing by hand.
  - In SPSS, hypothesis testing for determining whether a sample comes from a population with a specific mean is done through a one- sample t-test.
  - In class we used the z-test. The z-test is often more powerful if all the assumptions are met, but the t-test is more conservative.

Q. want to see if the average score in the sample is the same as the population`s score of 80- collected data on 5 people- scores were 34, 50,67,19,90

SPSS- one sample t-test

Input data in SPSS

1. Click **Analyze > Compare Means > One-Sample T Test…** on the main menu

2. The **One-Sample T Test** box will pop up and transfer the variable of interest
   into the test variable box

Input the comparison value of interest into the test-value box. This is the population mean that you are comparing the sample to.

- In the score data, input 80 into the test value box

Output

One-Sample Statistics table- gives the samples- summary stat

One-sample t-test table- reports the results of the one-sample t-test

- The top row provides the value of the known or hypothesized population mean you are comparing your sample data
- The "t" column is the calculated test statistic
- The "df" column is the degrees of freedom (N-1)
- The "Sig. (2-tailed)" column is the statistical significance (p-value) of the one sample t-test- This test compares the sample mean to the population mean.
  - If $p > .05$, the difference between the sample-estimated mean and the comparison population mean would not be statistically significantly different.

SPSS- two sample t-test

independent-samples t-test (or independent t-test) compares the means between two unrelated groups

- For example 10 students are weighed and want to see if the weight is different among the regions
  - Japan- 45,56,53,78,65
  - US-120,138,159,112,162
- What is the null and alternative hypothesis?

1. Input data into SPSS (should have at least two variables)
   - In the example above- should have a variable for gender and a test score variable

2. Click **Analyze > Compare Means > Independent-Samples T Test**… on the top menu
3. A **Independent-samples T-test** box will appear
4. Drag the dependent variable into the test variables box- which is the dependent variable? weight or region?
5. Drag the independent variable into the grouping variable box
6. Click OK

Output

- **Group Statistics table**-provides useful descriptive statistics for the two groups that you compared, including the mean and standard deviation
- **Independent samples Test table**-provides the actual results from the independent t-test.
  - the results are given for when the variances are assumed to be equal (pooled) and notequal.
  - can report when group means are significantly different when the value in the "Sig. (2-tailed)" row is less than 0.05

Inference of categorical variables

-We are often interested in the relationship of two or more variables.

-As discussed in previous sessions, we can use tools such as correlation, regression, t-tests, Analysis of variance and cross tabulation.

-This session will focus on cross tabulation/ contingency tables

-Cross tabulation is based on two or more categorical variables

Unlike continuous variables, the range of possible values of categorical variables are limited- normal assumption cannot be approximated.

-for example, the possible values for sex, ethnicity, disease status are limited. When we plot the histogram of the data, will not be bell shaped curved.

-Statistical tests (correlation, regression, t-test) require the assumption of normality- CANNOT use this methods to analyze categorical data.

Contingency table

A contingency table is visual method to show how two categorical variables are associated.

- i.e In a population of 47, 26 were male. Of the 26 males, 14 wore glasses. In the population there were a total of 28 people who wore glasses. Is the proportion of people who wear glasses the same for males and female?

-With the given information, we see that there are four possible groups (Male/ no glasses) (Male/ glasses) (Female/ no glasses) (Female/ glasses)

-to summarize, create a table to show the number of people in the each of the four groups

In a population of 47, 26 were male. Of the 26 males, 14 wore glasses. In the population there were a total of 28 people who wore glasses. Is the proportion of people who wear glasses the same for males and female?

Information given

	Male	Female	Total
Glasses	14	?	28
No Glasses	?	?	?
Total	26	?	47

Inferred Information

	Male	Female	Total
Glasses	14	14	28
No Glasses	12	7	19
Total	26	21	47

In class exercise

- Ms A wanted to know the proportion of students (by year) who lived on campus. She believed the proportion of students living on campus differed by school year.
- She surveyed 388 students. Of the 388 there were 90 sophomores, 98 juniors. Of the remaining, there are 74 more freshman than seniors
- Of those who lived on campus, 100 were freshman, 48 were sophomores and 1 was a senior.
- There were a total of 231 people who lived off campus
- Create a contingency table

Going back to Glasses by gender

Q-What do we want to know?

A- Is the proportion of people who glasses the same in males and females?

	Male	Female	Total
Glasses	14	14	28
No Glasses	12	7	19
Total	26	21	47

Q- What is the Null and Alternative hypothesis?

A- H_o: The proportion of male glasses = female glasses ($p_m=p_f$)

H_A: The proportion of male glasses female glasses (p_m pf)

Under the null hypothesis the expected number of females and males who wear glasses should the same as the overall sample of people who wear glasses.

The overall proportion or the pooled estimate of people who wear glasses in the sample is 28/47= 0.5957

Under the Null hypothesis…

- The pooled estimate of the people wearing glasses was 0.5957
- Under the null hypothesis, the proportion of females wearing glasses is expected to be 0.5957.
  - This is interpreted as $x/21 = 0.5957$
  - It is expected that 12.5 females are expected to wear glasses
- -Under the null hypothesis, the proportion of males wearing glasses is expected to be 0.5957.
  - This is interpreted as $x/26 = 0.5957$
  - It is expected that 15.5 males are expected to wear glasses
- -With this information we can fill out the table for the expected numbers

	Male	**Female**	**Total**
Glasses	15.5	12.5	28
No Glasses	10.5	8.5	19
Total	26	21	47

More formally…

- The expected values for each of the groups (cells) can be computed as:

- **expected cell count= (row total * column total)/ n**

- A formal test of the null hypothesis of no association between the row and column variables is done by comparing the observed values and the expected values under the Null hypothesis.

The formal test is called the chi-squared () test

In class exercise continued…

Before we do formal testing, the chi-squared test…

What is the expected number of students living on and off campus by school year?

observed	Off campus	On Campus	Total
Freshman	37	100	137
Sophomore	42	48	90
Junior	90	8	98
Senior	63	1	63
Total	231	157	388

expected	Off campus	On Campus	Total
Freshman	?	?	137
Sophomore	?	?	90
Junior	?	?	98
Senior	?	?	63
Total	231	157	388

Solution

	Off campus	On Campus	Total
Freshman	(137*231)/388		137
Sophomore	(90*231)/388		90
Junior			98
Senior			63
Total	231	157	388

	Off campus	On Campus	Total
Freshman	81.56	55.44	137
Sophomore	53.58	36.42	90
Junior	58.35	39.65	98
Senior	37.51	25.49	63
Total	231	157	388

Before actual testing- Assumptions for the chi-square

1. Random sampling is not required
2. The observations are independent
3. The frequencies, cell counts should not be less than 1 and no more than 20% of the expected values should be less than 5.
4. The row and column variables must be mutually exclusive meaning that the categories should not overlap or any observation can be classified into 2 different categories

Properties of the Chi-square

1. Chi-square is a statistic that compares the counts of the categories
2. Chi-square is never a negative number
3. Chi-square is non-symmetric
4. There are many chi-square distributions- one for each degree of freedom
5. The degrees of freedom when working with a single population variance is n-1.

Chi-square test for independence of 2 categorical variables

The hypothesis of interest is:

- H_o: The rows and columns are independent- there is no association
- H_A: The rows and columns are dependent- there is an association

The chi-square test statistic is

- O= observed
- E= expected

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The degrees of freedom = (#rows-1) * (#columns-1)

The p-value is found by looking up the chi-square table

- The alternative hypothesis is two sided but use one tail distribution since the numerator of the test statistic is squared- takes into account the two tails/sides

Table 5-2

Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Chi-square testing for glasses and gender

Going back to the example of females and males wearing glasses…

H_o: gender and wearing glasses are not associated/ independent

H_A: gender and wearing glasses are associated/ dependent

Chi-square testing for glasses and gender cont`d

- Use the information given and calculate the expected

observed	Male	Female	Total
Glasses	14	14	28
No Glasses	12	7	19
Total	26	21	47

expected	Male	Female	Total
Glasses	15.5	12.5	28
No Glasses	10.5	8.5	19
Total	26	21	47

- The Chi-square test statistic =

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$(14-15.5)^2/15.5 + (14-12.5)^2/12.5 + (12-10.5)^2/10.5 + (7-8.5)^2/8.5$

$=0.793$

Degrees of freedom and p-value

- The degrees of freedom = (#rows-1) * (#columns-1)

  = (2-1)*(2-1)

  = 1

- Look up 0.793 with df=1 in the chi square table→ somewhere between 0.1 and 0.5
- Since the p-value is >0.05 we do not have enough evidence to reject the null hypothesis, thus we fail to reject the null hypothesis
- Can conclude the wearing glasses is not associated with gender- wearing glasses is not dependent on gender

Short cut method for calculating Chi-squared

It is very tedious to calculate the observed and expected values from the contingency table to then calculate the test statistic

There is a simple method to calculate the chi-square test statistic

The general notation for a 2x2 contingency table is

	Data 1	**Data 2**	**Total**
Category 1	a	b	a+b
Category 2	c	d	c+d
Total	a+c	b+d	N

The chi-square test statistic can be calculated by the following equation

Going back to the on/off campus living Q

- Is there an association with school year and living on/off campus?

observed	Off campus	On Campus	Total
Freshman	37	100	137
Sophomore	42	48	90
Junior	90	8	98
Senior	63	1	63
Total	231	157	388

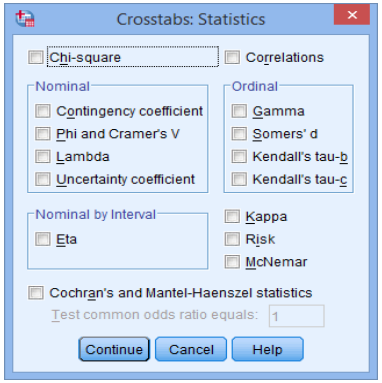
Expected	Off campus	On Campus	Total
Freshman	81.56	55.44	137
Sophomore	53.58	36.42	90
Junior	58.35	39.65	98
Senior	37.51	25.49	63
Total	231	157	388

SPSS-

Steps for chi-square test for independence analysis

1. Input data- for example in the gender and glasses example, have two variables gender and glasses.
2. Click Analyze > Descriptive Statistics > Crosstabs…
3. The crosstabs box will pop-up
4. Transfer one variable into the Row(s): box and the other variable into the Column(s): box

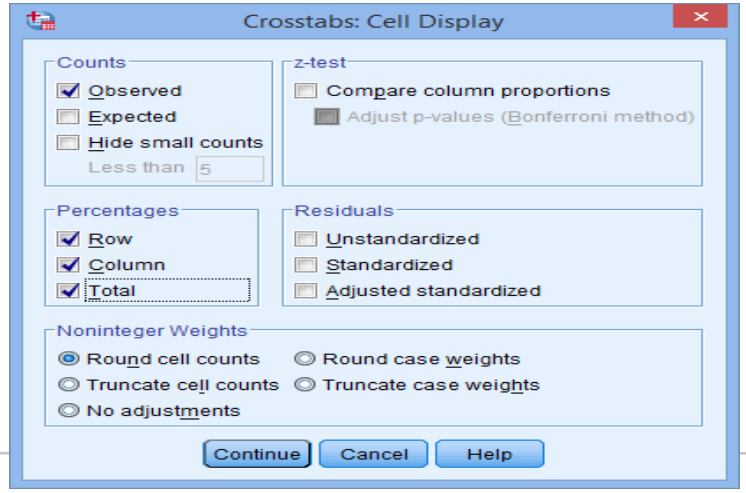
5. Click on the statistics button-



5. Select chi-square and phi and Cramer and click continue

6. Click the cells button

7. Select observed from the counts section, and rows, column and total from the percentages→ continue



8. Click OK to get output

SPSS- Output

Should have many tables in the output viewer

- Cross tabulation table (gender and eye glasses)
- Chi-square tests table
  - when reading this table we are interested in the results of the "Pearson Chi-Square" row
    - The value under the value column is the chi-square test stat
    - df= is the degrees of freedom
    - Asymp sig is the p-value for the chi-square test statistic