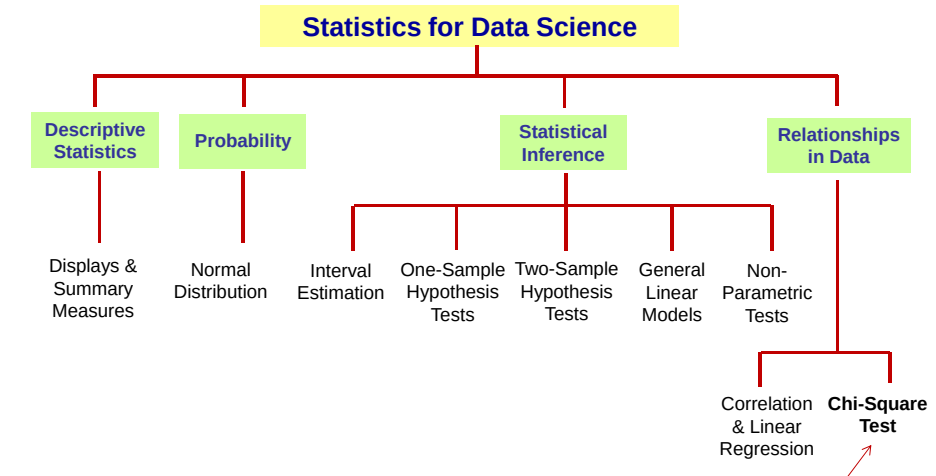
MATH 4044
Statistics for Data Science

Relationships in categorical data. Chi-square test

Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 18, 18.1-18.5.

Week 12

Topics to be covered

- Describing relationships in categorical data.
- Chi-square test for two-way contingency tables:
  - Association;
  - Independence;
  - Agreement.

Example: Identifying spam email

- Statistics can be used to filter spam from incoming email messages.
- By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy.
- One of those characteristics is whether the email contains no numbers, small numbers, or big numbers.

Example: Identifying spam email

LEVELS

Table of number by spam

FACTOR →		spam		
↓		No	Yes	Total
number				
big	Frequency	407	138	545
	Percent	10.38	3.52	13.90
	Row Pct	74.68	25.32	
	Col Pct	12.33	22.26	
none	Frequency	470	79	549
	Percent	11.99	2.01	14.00
	Row Pct	85.61	14.39	
	Col Pct	14.24	12.74	
small	Frequency	2424	403	2827
	Percent	61.82	10.28	72.10
	Row Pct	85.74	14.26	
	Col Pct	73.43	65.00	
Total	Frequency	3301	620	3921
	Percent	84.19	15.81	100.00

LEVELS

We can use the data from the two-way table for different types of calculations, and to test hypotheses.

Example: Identifying spam email

Table of number by spam

		spam		
		No	Yes	Total
number				
big	Frequency	407	138	545
	Percent	10.38	3.52	13.90
	Row Pct	74.68	25.32	
	Col Pct	12.33	22.26	
none	Frequency	470	79	549
	Percent	11.99	2.01	14.00
	Row Pct	85.61	14.39	
	Col Pct	14.24	12.74	
small	Frequency	2424	403	2827
	Percent	61.82	10.28	72.10
	Row Pct	85.74	14.26	
	Col Pct	73.43	65.00	
Total	Frequency	3301	620	3921
	Percent	84.19	15.81	100.00

Proportions

$$\text{proportion} = \frac{\textit{cell value}}{\text{table/row/column total}}$$

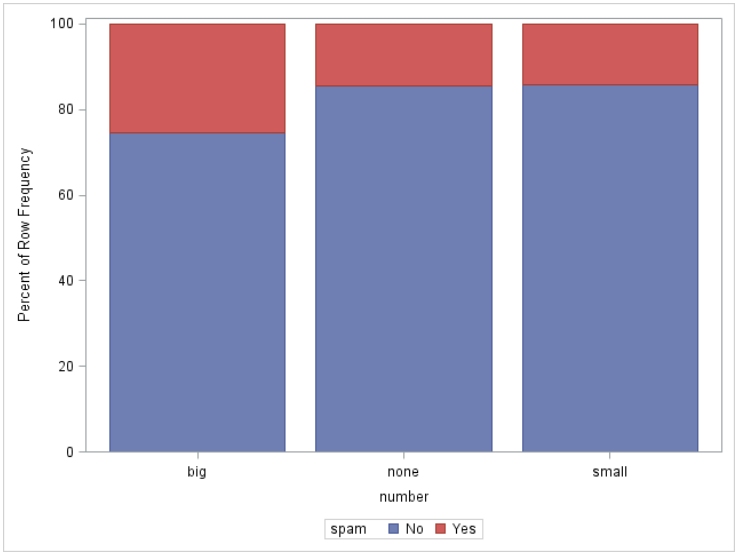
Proportion of 'big' and 'yes' $= \frac{138}{3921} = 0.0352$

Expected Counts

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

Expected count of 'big' and 'yes' $= \frac{545 \times 620}{3921} = 86.177$

Example: Identifying spam email



Three ways to look at the data

- Three ways to look at the data:
  - Compare appropriate proportions
    - Which outcomes occur with quite different probabilities?
    - *Answer using a hypothesis test of homogeneity.*
  - Compare observed and expected cell counts
    - Which cells have more or fewer observations expected if H_0 were true?
    - *Answer using a hypothesis test of independence.*
  - Better understand the chi-square statistic:
    - Which cells contribute the most to the value of the test statistic?

Hypothesis testing: Chi-square tests

- The study design indicates whether the hypothesis test is for *homogeneity* or *independence*.
- Chi-Square test of independence:
  - Tests independence of the row and column variables.
  - One simple random sample is collected.
- Example: A set of 3921 emails from 2012 were examined.
  - The emails were classified as spam or not spam.
  - It was also recorded whether they contained no numbers, small numbers or big numbers.
  - Is an email being spam independent of it containing numbers?

Hypothesis testing: Chi-square tests

- Chi-Square test of homogeneity:
  - Tests whether different populations have the same proportions, based on a variable of interest.
  - Independent random samples are collected from each population.
- Example:
  - A set of 620 spam emails have been selected and examined whether they contained any numbers.
  - A second set of 3301 non-spam emails were also selected and examined whether they contained numbers.
  - Are there differences in frequency of big, small or no numbers in emails that are and are not spam?

Chi-square test set-up

Independence	Homogeneity
FORMULATE	FORMULATE
H_0: There is no association between the two variables.	H_0: All population proportions are equal.
H_1: There is an association between the two variables.	H_1: Not all population proportions are equal.
SOLVE	SOLVE
Compute χ^2 and degrees of freedom.	Compute χ^2 and degrees of freedom.
If *P-value* ≤ α reject H_0.	If *P-value* ≤ α reject H_0.
If *P-value* > α fail to reject H_0.	If *P-value* > α fail to reject H_0.
CONCLUDE	CONCLUDE
Return to the practical question to describe your results.	Return to the practical question to describe your results.

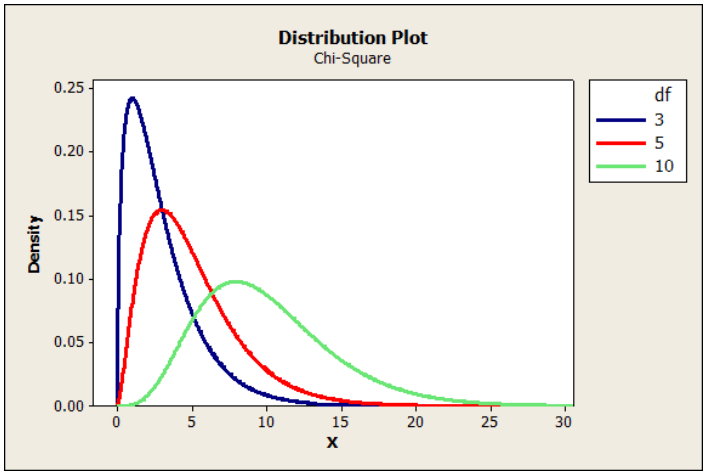
Chi-Square test statistic

- Using the observed and expected counts we can compute the chi-square test statistic χ^2

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \sum \frac{(O-E)^2}{E}$$

- We will use this statistic in the two hypothesis tests and reject H_0 if the value of the chi-square statistic is too large.
- The χ^2 statistic has chi-square distribution with degrees of freedom (# rows – 1) x (# columns – 1).
- The expected counts are calculated under the implicit assumption that H_0 is true.

Chi-square distribution



The P-value is the area under the curve to the right of the χ^2 test statistic value.

Conditions for the chi-square test

- Observations are independent, and each observation falls into just one of a finite number *k* of complementary and mutually exclusive outcomes.
- The expected frequency for each cell is 5 or greater.
- You can safely use the chi-square test with critical values from the chi-square distribution when:
  - No more than 20% of the *expected counts* are less than 5 and
  - All individual *expected counts* are greater than or equal to 1.
- If more than 20% of the cells have expected counts that are less than 5:
  - In 2x2 tables use Fisher's exact test (available by default in SAS).
  - In larger tables, you can collapse rows or columns so that the cell frequencies are larger.

Example: Identifying spam email

Table of spam by number					
		number			
		1:None	2:Small	3:Big	Total
spam					
1:Yes	Frequency	79	403	138	620
	Expected	86.809	447.01	86.177	
	Cell Chi-Square	0.7026	4.3336	31.164	
2:No	Frequency	470	2424	407	3301
	Expected	462.19	2380	458.82	
	Cell Chi-Square	0.132	0.8139	5.8533	
Total	Frequency	549	2827	545	3921

Statistic	DF	Value	Prob
Chi-Square	2	42.9994	<.0001
Likelihood Ratio Chi-Square	2	38.5438	<.0001
Mantel-Haenszel Chi-Square	1	24.4116	<.0001
Phi Coefficient		0.1047	
Contingency Coefficient		0.1042	
Cramer's V		0.1047	

Fisher's Exact Test	
Table Probability (P)	<.0001
Pr <= P	<.0001

H_0: There is no association between number and spam.

H_1: There is an association between number and spam.

$\alpha = 0.05$

The test statistic is $\chi^2 = 42.9994$ with 2 degrees of freedom.

Since the P-value is <0.0001, we reject H_0

Thus, at 5% level of significance, there is an association between number and spam.

A closer look at the two-way table

Table of spam by number					
		number			
		1:None	2:Small	3:Big	Total
spam					
1:Yes	Frequency	79	403	138	620
	Expected	86.809	447.01	86.177	
	Cell Chi-Square	0.7026	4.3336	31.164	
2:No	Frequency	470	2424	407	3301
	Expected	462.19	2380	458.82	
	Cell Chi-Square	0.132	0.8139	5.8533	
Total	Frequency	549	2827	545	3921

SAS can produce the expected counts and cell chi-square values.

These counts indicate whether the sample over- or under-represents parts of the population.

The largest contribution (31.164) to the chi-square statistic (42.9994) comes from spam emails with big numbers.

The actual count (138) for spam emails with big numbers was much higher than the expected count (86.177) assuming independence.

Example: SAS code

```
proc format;
       value Spam 1='1:Yes' 0='2:No';
       value $Num 'none'='1:None' 'small'='2:Small'
                  'big'='3:Big';
run;
```

Formatting to produce informative labels and rearrange rows and columns in the contingency table.

```
proc freq data=work.email order=formatted;
       tables spam * number / chisq exact expected cellchisq
                              nocol norow nopercent;
       format spam Spam. number $Num.;
run;
```


Example: Google search algorithms

- Google regularly runs experiments help improve their search engine.
- In an experiment, 10,000 google.com queries are split into three algorithm groups:
  - The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 each for two test algorithms.
- The ultimate goal is to see whether there is a difference in the performance of the algorithms:
  - H_0: The algorithms each perform equally well.
  - H_1: The algorithms do not perform equally well.

Example: Google search algorithms

- Performance is measured by whether the search results align with the user's interests:
  - If the user clicked one of the links provided and did not try a new search, the initial search is considered successful;
  - If the user performed a related search, the initial search was not successful.

New Search	Algorithm			Total
	Current	Test 1	Test 2	
No	3511	1749	1818	7078
Yes	1489	751	682	2922
Total	5000	2500	2500	10000

Example: Google search algorithms

Test of homogeneity

Table of NewSearch by Algorithm

		Algorithm			
		Current	Test1	Test2	Total
NewSearch					
No	Frequency	3511	1749	1818	7078
	Expected	3539	1769.5	1769.5	
	Cell Chi-Square	0.2215	0.2375	1.3293	
Yes	Frequency	1489	751	682	2922
	Expected	1461	730.5	730.5	
	Cell Chi-Square	0.5366	0.5756	3.2201	
Total	Frequency	5000	2500	2500	10000

Statistic	DF	Value	Prob
Chi-Square	2	6.1203	0.0469
Likelihood Ratio Chi-Square	2	6.1749	0.0456
Mantel-Haenszel Chi-Square	1	4.1154	0.0425
Phi Coefficient		0.0247	
Contingency Coefficient		0.0247	
Cramer's V		0.0247	

H_0: The proportions of new searches are equal across all three algorithms.

H_1: Not all population proportions are equal.

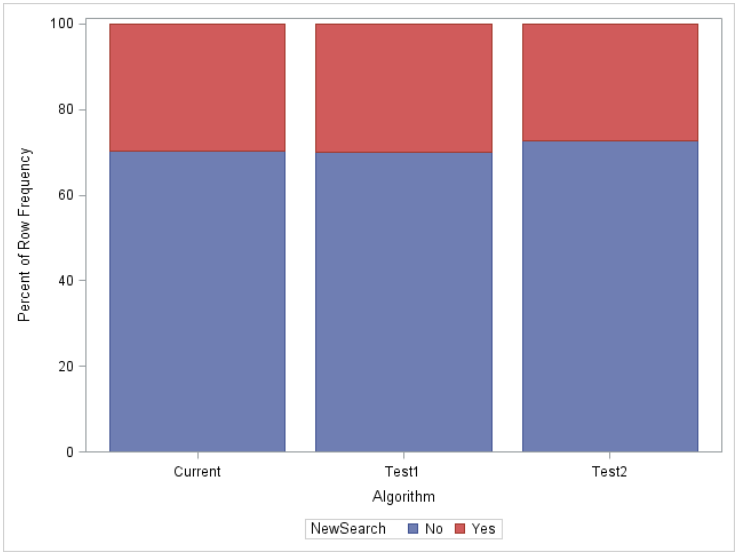
$\alpha = 0.05$

The test statistic is $\chi^2 = 6.1203$ with 2 degrees of freedom.

Since the P-value is $0.0469 < 0.05$, we reject H_0.

At 5% level of significance, there is a statistically significant difference among proportions of new searches for the three algorithms.

Example: Google search algorithms



Example: SAS code

```
data google;
input NewSearch $ Algorithm $ Count;
datalines;
No Current 3511
No Test1 1749
No Test2 1818
Yes Current 1489
Yes Test1 751
Yes Test2 682
;
proc freq data=work.google;
tables NewSearch * Algorithm / chisq expected cellchisq nocol
                               norow nopercent;
weight Count;
run;
```

Creating a data file containing counts from our contingency table.

Statement to specify frequencies for each combination of factor levels in the contingency table.

Example: Google quality raters

- Quality raters are Google's fact checkers – the people who work to make sure the algorithm is doing what it is supposed to do.

- Data from quality raters not only serves as quality control on existing search engine results pages, but it helps validate potential algorithm changes.
- A sample of 33 searches were evaluated for relevance by two raters.
  - Was any agreement between raters due simply to chance?

McNemar's test

- You would perform McNemar's test if you were interested in the marginal frequencies of two binary outcomes.
- These binary outcomes may be the same outcome variable on matched pairs (like a case-control study) or two outcome variables from a single group.
  - These counts can be considered in a two-way contingency table.
- For our rater example, the null hypothesis is that both raters assess searches as relevant or irrelevant at the same rate (or that the contingency table is symmetric).

Example: Google quality raters

Table of Rater1 by Rater2				
		Rater2		
		No	Yes	Total
Rater1				
No	Frequency	12	3	15
	Percent	36.36	9.09	45.45
	Row Pct	80.00	20.00	
	Col Pct	85.71	15.79	
Yes	Frequency	2	16	18
	Percent	6.06	48.48	54.55
	Row Pct	11.11	88.89	
	Col Pct	14.29	84.21	
Total	Frequency	14	19	33
	Percent	42.42	57.58	100.00

McNemar's Test	
Statistic (S)	0.2000
DF	1
Pr > S	0.6547

$$\chi^2 = \frac{(3-2)^2}{3+2}$$

Simple Kappa Coefficient	
Kappa	0.6927
ASE	0.1262
95% Lower Conf Limit	0.4453
95% Upper Conf Limit	0.9402

Test of H0: Kappa = 0	
ASE under H0	0.1737
Z	3.9870
One-sided Pr > Z	< .0001
Two-sided Pr > |Z|	< .0001

Kappa is the coefficient of agreement:
1 = perfect agreement;
0 = agreement purely by chance

Example: Google quality raters

- The test statistic for McNemar's test is $\chi^2 = 0.2$, with 1 degree of freedom and a *P*-value = 0.6547 > 0.05.
  - We conclude that the two raters assess searches as relevant or irrelevant at the same rate.
- The estimate of the kappa coefficient is $\kappa = 0.6927$, with a *P*-value < 0.0001.
  - The agreement between the two raters was therefore significant and not due to chance.

Example: SAS code

```
data work.raters;
input Rater1 : $3. Rater2 : $3. @@;
```

Creating two character variables of length 3

Several observations are to be created from a single line of data

```
datalines;
No No Yes Yes No Yes Yes No No No No No No
No Yes Yes Yes Yes Yes Yes No No No Yes No No Yes
Yes Yes No Yes Yes Yes Yes Yes Yes No No No No Yes
Yes Yes Yes No Yes No No No No Yes Yes Yes Yes Yes
Yes Yes Yes No No Yes Yes No No Yes Yes
;
proc freq data=work.raters;
      tables Rater1 * Rater2 / agree;
      test kappa;
      run;
```

To obtain results of McNemar's test