MBC 638

LIVE SESSION WEEK 4

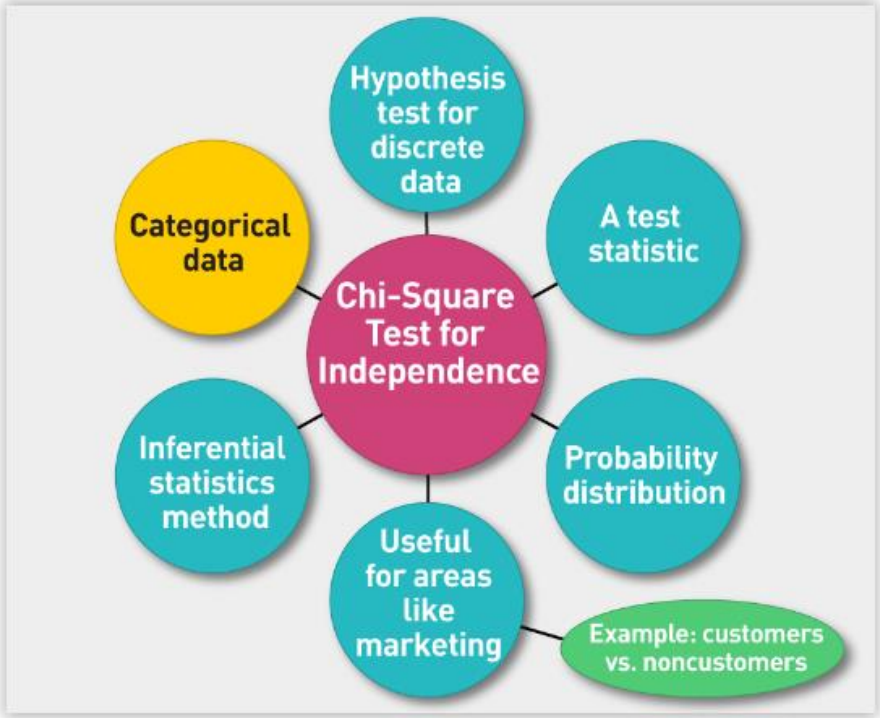
Agenda

Topic	Time	Thursday Section	Sunday Section
Introduction	5 min	9:00-9:05PM	6:30-6:35PM
Highlights from Week 4 Video	40 min	9:05-9:45PM	7:35-7:15PM
Breakout on Additional Example	20 min	9:45–10:05PM	7:15–7:35PM
Review of Upcoming Assignments and Open Question	25 min	10:05-10:30PM	7:35-8:00PM

Highlights: Video Segment 4.3:Chi-Square Test of Independence

Hypothesis test for discrete data, categorical data – i.e. customer or non-customer, do we need to market to them differently

Verifying the validity of a claim we are making about a sample, this test should point you in a direction



Highlights: Video Segment 4.3:Chi-Square Test of Independence

χ^2 test for independence: a procedure used to determine if two variables are related or are statistically independent

Convention:

H_0: Categorical Variable 1 and Categorical Variable 2 are independent (i.e., there is *no* relationship).

H_a: Categorical Variable 1 and Categorical Variable 2 are not independent (i.e., there *is* a relationship).

How it works:

Compares "observed" counts and "expected" counts or frequencies

Does not give kind (positive/negative) or intensity of relationship

All it tells us is, if there is a relationship – not directional

Highlights: Video Segment 4.3:Chi-Square Test of Independence

Setup for Analysis of Two-Way Tables

H0: Categorical Variable 1 and Categorical Variable 2 are independent.
Ha: Categorical Variable 1 and Categorical Variable 2 are not independent.

Example: Does day of week affect car sales? Variable 2

Example: Does day of week affect car sales?

		Variable 2							
		MON	TUES	WED	THUR	FRI	SAT	SUN	**Totals**
Variable 1 Customer lead resulting in:	Sale of car	Count data	Count data	Count data	Count data	Count data	Count data	Count data	Row total
	Not selling a car	Count data	Count data	Count data	Count data	Count data	Count data	Count data	Row total
	Totals	Col. total	Col. total	Col. total	Col. total	Col. total	Col. total	Col. total	Grand total

Put your data in a two-way table.
Depending on the variables, use two or more columns or rows.
In each cell, count the frequency of simultaneous occurrence.
Discrete data: need a lot of data (> 5 counts per cell)
No exact sample size specified

Highlights: Video Segment 4.4:Chi-Square Example

1. Write your hypothesis statements
H0: Age and feature preference are independent (no relationship).
Ha: Age and feature preference are not independent (relationship).

2. Fill your table

	Feature A	Feature B	Feature C	Totals
Young	200	38	20	258
Old	157	45	5	207
Totals	357	83	25	465

3. Reorganize the table
1-turn the data into columns

Category	f(Observed)
Young A	200
Young B	38
Young C	20
Old A	157
Old B	45
Old C	5

4- calculate F(Expected)
i.e. Young A = (258X357)/465=198.1
Old A = (207X357)/465=158.9

Category	f(Observed)	F(Expected)
Young A	200	198.1
Young B	38	46.1
Young C	20	13.9
Old A	157	158.9
Old B	45	36.9
Old C	5	11.1
Totals	465	

5– calculate Chi and then add them all up

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

i.e. Young A = ((200-198.1)^2)/198.1=.02

Category	f(Observed)	F(Expected)	$\frac{(f-F)^2}{F}$
Young A	200	198.1	0.02
Young B	38	46.1	1.41
Young C	20	13.9	2.71
Old A	157	158.9	0.02
Old B	45	36.9	1.75
Old C	5	11.1	3.38
Totals	465		**9.288**

Highlights: Video Segment 4.4:Chi-Square Example

6– Calculate degrees of freedom df

Notes: Degrees of Freedom: # of independent pieces of data that you have, (For example, imagine you have four numbers (a, b, c and d) that must add up to a total of m; you are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to m - thus your degree of freedom is three.) http://www.statsdirect.com/help/default.htm#basics/degrees_freedom.htm

$$df = (r-1)(c-1)$$
$$= (2-1)(3-1) = 2$$

- Gives which row to reference in χ^2 distribution table

7. Use the table to find the p value, probability is across the top, df on the left column, look for your value inside the table, look above to find your probability =.01

8. Is p lower than our alpha of .05? Yes, p is lower than .05 so we must reject Ho and accept the alternative. Yes, the data is related – age and feature preference are not independent – they have a relationship.

Chi Square Values

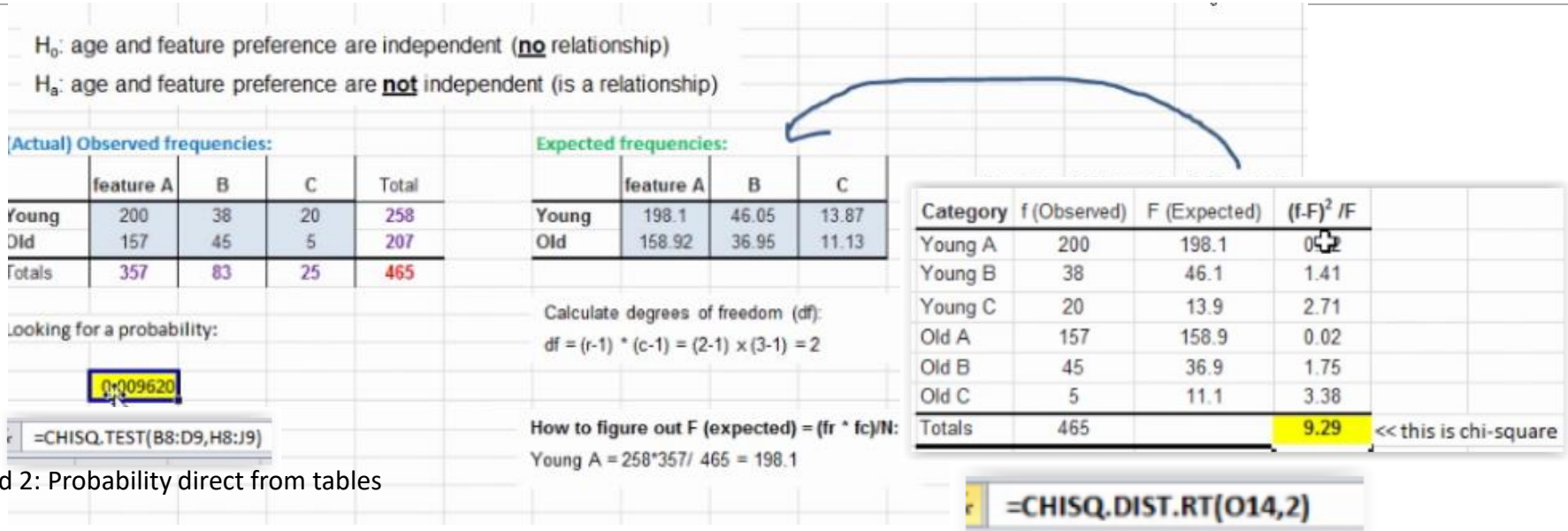
Probabilities

Chi-Square Distribution Table

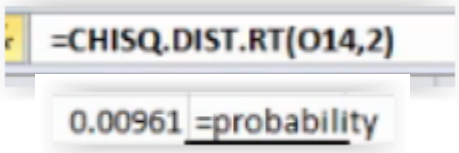
Table E: Chi-Square (χ^2) Distribution

Degrees of freedom	Area to the right of critical value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Highlights: Video Segment 4.5:Chi-Square in Excel



Method 2: Probability direct from tables



Method 1: This gives you the probability if you have calculated your Chi Square and df.

Why is Chi Square always a right tail test?

The chi-squared test is essentially *always a one-sided test*. Here is a loose way to think about it: the chi-squared test is basically a 'goodness of fit' test. Sometimes it is explicitly referred to as such, but even when it's not, it is still often in essence a goodness of fit. For example, the chi-squared test of independence on a 2 x 2 frequency table is (sort of) a test of goodness of fit of the first row (column) to the distribution specified by the second row (column), and vice versa, simultaneously. Thus, when the realized chi-squared value is way out on the right tail of it's distribution, it indicates a poor fit, and if it is far enough, relative to some pre-specified threshold, we might conclude that it is so poor that we don't believe the data are from that reference distribution.

If we were to use the chi-squared test as a two-sided test, we would also be worried if the statistic were too far into the *left* side of the chi-squared distribution. This would mean that we are worried the fit might be *too good*. This is simply not something we are typically worried about. (As a historical side-note, this is related to the controversy of whether Mendel fudged his data. The idea was that his data were too good to be true. See here for more info if you're curious.)

http://stats.stackexchange.com/questions/22347/is-chi-squared-always-a-one-sided-test

Highlights: Chi-Square for Projects Examples

Does the day of the week and how much I exercise have a relationship?

Ho: Weekend/Weekday and Amount of Exercise are independent
Ha: Weekend/Weekday and Amount of Exercise are NOT independent

	Weekday	**Weekend**	Totals
High Exercise Min(>60)			
Low Exercise Min(<60)			
Totals			

Does the amount of time I work have a relationship with how much sugar I consume?

Ho: Amount of time worked and amount of sugar consumed are independent
Ha: Amount of time worked and amount of sugar consumed are NOT independent

	<20 grams of sugar	**20-40 grams of sugar**	**>40 grams of sugar**	Totals
>480 min work day				
<480 min work day				
Totals				

Highlights: Video Segment 4.6:Chi-Square in Excel - Coffee

1 – Look at coffee example and alternative method for calculation

Agenda

Topic	Time	Thursday Section	Sunday Section
Introduction	5 min	9:00-9:05PM	6:30-6:35PM
Highlights from Week 4 Video	40 min	9:05-9:45PM	7:35-7:15PM
Breakout on Additional Example	20 min	9:45–10:05PM	7:15–7:35PM
Review of Upcoming Assignments and Open Question	25 min	10:05-10:30PM	7:35-8:00PM

Chi-Square Example and Breakout

1. What level of alpha would you reject Ho?
2. Is this a reasonable alpha to use?
3. If not what alpha would you use instead?
4. What would you conclude about your data set?

Review of Upcoming Assignments

Topic	Thursday Section Due Dates	Sunday Section Due Dates
Homework #2 in LaunchPad • Chapter 9 Online Practice Quiz • Ch 11 StatTutor Expected counts in 2-way tables	Sunday 10/28 Midnight EST	Wednesday 10/31 Midnight EST
Project Activities	Measure should be wrapping up in the next week to two weeks, and Analyze should be getting started. Some ideas on getting started with Analyze • Look at the list of tools • Start making graphs to look for patterns in the data • Calculate some basic descriptive statistics: mean, median, mode, etc. • Chi Square test for independence – is there a relationship between X and Y	
Quiz #2 Prep	I also posted a Quiz 2 prep file to help you study for Quiz 2, I'll post the answer document next week. We will be reviewing the questions and answers over the next 2 classes prior to the quiz.	
Start working on HMWK #3 in LaunchPad, Chapter 8 Online Practice Quiz	Sunday, 11/4 Midnight EST	Wednesday,11/7 Midnight EST