6/22/11

Lecture #16

Plan
- review
- Simpson's Paradox
- Experimentation
- Bivariate Analysis

$$\text{pval} \triangleq P(\text{our data or more extreme} \mid H_0)$$

If pval $\downarrow$ $\Rightarrow$ H_0 not plausible

If pval $\uparrow$ $\Rightarrow$ H_0 highly plausible

What determines if it's low or high? α

Coin A

		H	T	
Coin C	H	6	5	11
	T	5	4	9
		11	9	20 = n

Check if T_A indep H_C

$$p(H_C, T_A) \stackrel{?}{=} P(H_C)\,P(T_A)$$

$$\frac{5}{20} \stackrel{?}{=} \frac{6}{11} \cdot \frac{5}{9}$$

$$\Rightarrow \frac{5}{20} \neq \frac{30}{99}$$

$\Rightarrow H_C, T_A$ dependent!

Why is this false?

<u>errors</u>

a) Prob is <u>long run freq</u> .. this is not long run freq

b) Even in any finite n, we expect some random variation

$\Rightarrow$ this is the reason for the χ^2 test of independence

Another ex: UC Berkeley was sued in 1973 on allegations of gender bias in grad school. Why?

Admitted?

Cond probs given gender

		Y	N	
Gender	M	3714 (44%)	4728	8442
	F	1512 (35%)	2809	4321
		5226	7537	12763 = n

Admitted | Male

Admitted | Female

HW: run χ^2 test of independence. Answer: admitted & gender are associated.

Can we make the jump from association $\Rightarrow$ causation?

Can we say being a woman <u>causes</u> you to not be admitted to Berkeley grad school?

Are there other variables that affect admission rates?

What about which dept you apply to? It's prob. harder to get into pure math than history...

Maybe there's a hidden bias in men and women to choose harder or easier depts?

Let's analyze admissions by gender & dept:

Dept	Men Admitted (%)	Total	Women Admitted %	Total
A	62%	825	82%	108
B	63%	560	68%	85
C	37%	325	34%	593
D	33%	417	35%	375
E	28%	191	24%	393
F	6%	272	7%	341

If anything there's now a bias in favor of women.
Analysis was published in 1973 and is used in all stats courses to exemplify the following three concepts

lurking variable: a concealed variable that can affect the relationship between the two variables of interest

confounding: when a variable is affected by a lurking variable so strongly that the original analysis is changed

Simpson's Paradox: an abrupt change in the association of the two original variables when "controlling" for a lurking variable ...

Ch5.2 Wikipedia

What's going on in Simpson's Paradox? We will see this again in the other types of binomial analysis. Take this simple example:

Lisa and Bart both edit and improve Wikipedia articles:

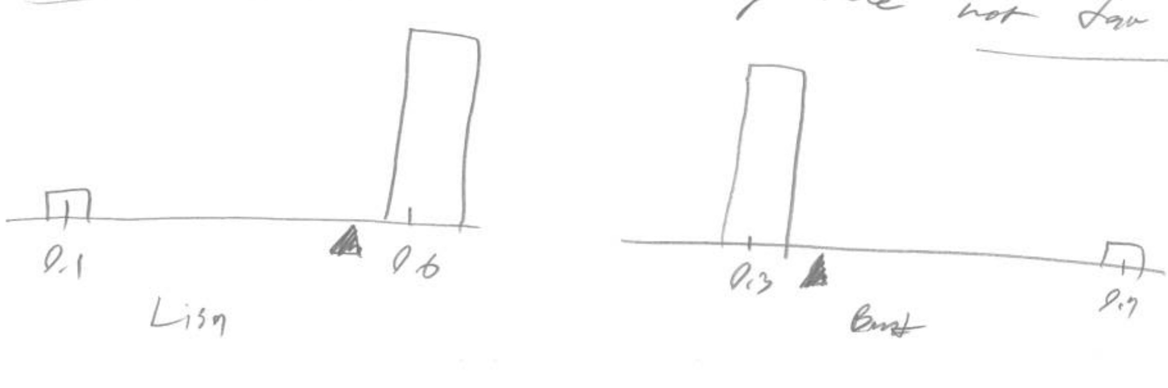
Lisa 61/110

Bart 39/110

Overall Lisa improves a greater %age of articles. But, week by week:

	Week 1	Week 2	Total
Lisa	60/100	1/10	61/110
Bart	9/10	30/100	39/110

Bart wins week to week because he has a higher ratio each week, but the sample sizes were not the same so they were not fair comparisons.



The paradox is seen if we only look at %ages.

	Week 1	Week 2	
Lisa	60%	10%	55.5%
Bart	90%	30%	35.5%

"Week" is a lurking variable

You can look at this and ponder...

How about another example:

	1995	1996	1997	
Derek Jeter	.250	.314	.291	.300
David Justice	.253	.321	.329	.298

"Year" is a lurking variable

But... sample sizes not the same

	1995	1996	1997	
Derek Jeter	12/48	183/582	190/654	385/1284
David Justice	104/411	45/140	163/495	312/1046

Back to Berkeley...

Women apply to more competitive programs in larger #'s than men. Since dept by dept they are not discriminated against, changes were flipped...

Cond ch 5.2 ——— association $\not\Rightarrow$ causation

Other examples of confounding / Simpson's Paradox

Students who finish early on exams have better grades

finishing early cause better grades?

No... confounded by IQ and study habits...

People who own BMW's have more money than people who own Fords... does owning a BMW cause you to have more money? No... confounded by job level, etc...

If there's always the possibility of variables to be confounded, how do you conclude anything causal?

Observational Study: tries to draw inference about certain effects by merely observing them in the real world... suffers from confounding... you cannot conclude anything causal (interested... Rosenbaum Stat 501)

Why? There's no control. People who opt in for certain activities have selection bias to other activities... Example: people who buy BMW's tend to have a better jobs. Example: hospitals improve health? but people who go to hospitals tend to be sick so people who don't go are healthier than those that do...

How do we eliminate selection bias?

Randomized Experiments Ch 18.1

This is how the scientific method works! Experiments can get you causal data.

factor: the variable manipulated e.g. drugs in a study

treatment: the factor given to a randomized group in the study...

100 rats.. randomized into two groups: drug and placebo. There's no option for the rats to ~~opt~~ opt into taking the drug.. They're randomized.

Obs study: 100 people. 50 people take a drug 50 people don't. Problem: the 50 people who take the drug may be disproportionately alcoholics, skydivers, live in areas with pollution, etc tons of possible lurking vars biases

Experiment: 100 people 50 of them are randomly selected to take the drug. No possibility of selection bias.

Obs study vs. Experiments: Why do people do obs studies?

- Unethical to do experiments...

Imagine 100 people... force 50 of them to smoke randomly.. and force them to keep smoking!

- don't have authority

-

⇒ there is no experimental evidence (in humans) that smoking causes lung cancer only obs. data.

Cigarette companies paid for the large stats ... "Simpson's Paradox" on monotonous usage ... all sorts of things..

Famous statistician Fisher (inventor of hyp. testing) smoked because he never believed the data was causal. He died of colon cancer...

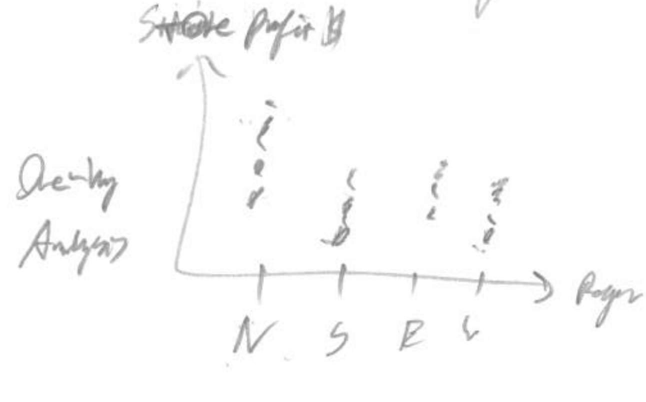
Bivariate Analysis

~~Categorical data vs. categorical data (ch 5)~~

Categorical data vs. numerical data (today tomorrow) (ch 18 + stuff not in book)

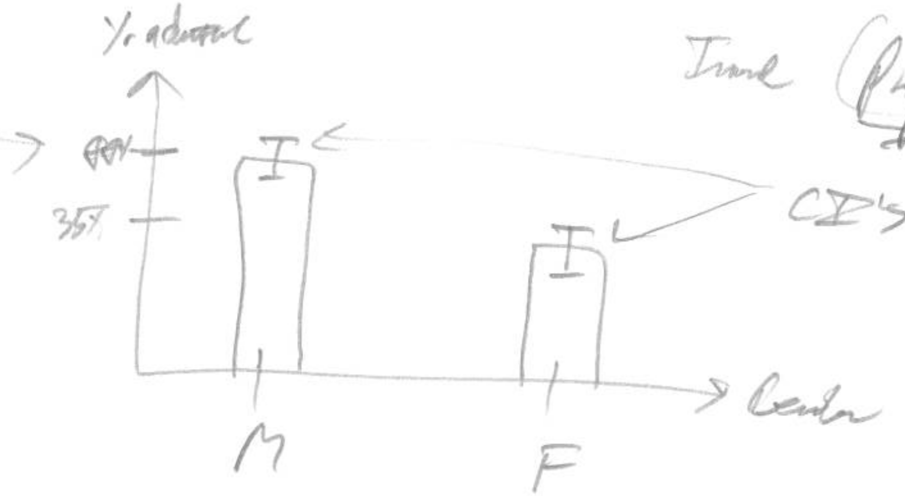
Numerical data vs. numerical data (ch 6, 19)

Two ways to present:

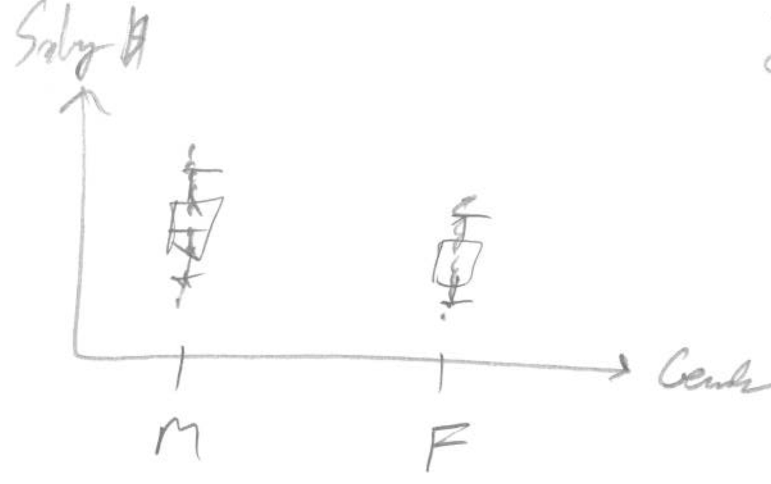


Comparison... for Stat 101, we only look at two categories vs numerical... In Stat 102 you'll look at many categories vs. numerical (ANOVA).

Inference (Proportions)



Inference



We'll see more examples tomorrow...

Ch 18: two-proportion and two-sample hypothesis testing.