1

6/21/11

Lecture #15

Admin

- HW #4 almost done and LONG
most important HW set of ~~the~~ semester.

Plan

- a little bit more about CI's
- χ^2_{df} distr. (not on test)
- tests for variance
- ~~sum~~ ind of r.v.'s (ch. 10)
- cont. tables (ch. 5.1)(ch 5.3)
(Cramér V not covered)

Sec 15.4 manipulating CI's

Let's say we have a CI for the profit made at one store and all stores are essentially alike

$$CI_{\mu(\text{profit one store}),95\%} = [38{,}000,\ 40{,}000]$$

If we have 10 stores, we can just multiply by 10 to get a CI for all stores

$$CI_{\mu(\text{all stores}),95\%} = 10\, CI_{\mu\,\text{profit (one store)},95\%} = [380{,}000,\ 400{,}000]$$

book talks about monotonic transforms ... you'll see one on HW

What is the Chi-sq distribution?

$$\chi^2_{df} \stackrel{d}{=} Z_1^2 + Z_2^2 + \dots + Z_{df}^2$$ where $Z_1, Z_2, \dots Z_{df}$ independent

$$= \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2 + \dots + \left(\frac{X_{df} - \mu_{df}}{\sigma_{df}}\right)^2$$

Greek letter chi and a squared

Turns out many things have the above configuration, so it is very useful.

Example: beer manufacturer making 40's. Need nozzle to be $\sigma = 0.2 oz$ and not higher otherwise there's overflow and underflow. They have separate QA for the mean such as we did yesterday. How to do this? Take an SRS $n = 15$ beers at setup test:

H_0: $\sigma = 0.2 oz \iff \sigma^2 = 0.04 oz^2$

H_a: $\sigma > 0.2 oz \iff \sigma^2 > 0.04 oz^2$

Take sample: $s = 0.27 oz \iff s^2 = .0729 oz^2$

Is this due to chance? Or is it too unlikely under the null hyp? Need test statistic with known distribution

$$Q \stackrel{\Delta}{=} (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

proving this you need a PhD

draw from χ^2_{24}

$$q = (25-1)\frac{.0729 oz^2}{.04 oz^2} = 43.74$$

If we look on sheet, at $\alpha = 0.05$, $\chi^2_{0.05,24} = 36.4$

Since $q > 36.4 \Rightarrow$ reject H_0.

$pval = P\left(\chi^2_{24} \geq 43.74\right) = .0082 < \alpha = 5\% \Rightarrow$ reject H_0

something is probably wrong with the machine

We're done with univariate statistics! that is, looking at data for one variable e.g. all heights, all volumes, all wages for one company, etc

We now move on to bivariate analysis where we look at data between variables and how they relate. For instance:

how is a person's height associated with weight?

how does gender affect SAT score → interval vs. interval

→ nominal categorical vs. interval

how does Location Region affect purchase (ch 5)

→ nominal categorical vs. nominal categorical

The rest of the class will be looking & analyzing bivariate data.

We start with categorical vs. categorical. ↗ shows data categories upon other variables

The way we show this data is through a "contingency table"

Host

Purchase?	MSN	Recipe Source	Yahoo	
Y	6977	8282	5848	17,103
N	275	1	230	516
	7252	8283	6078	17,619 = n

P. 78

"cells" are mutually exclusive & collectively exhaustive of the data

Entries in the "margins" are totals and represent the "marginal distr."

Why? Regardless of host, the # of people purchasing is:

Y	17103
N	516

The marginal distr. is the univariate distr.

We can also see conditional distributions:

	MSN	Recipe Source	Yahoo	
Y	96%	99.98%	96.22%	97.07%
N	3.93%	0.08%	3.28%	2.93%

	MSN	RS	Yahoo
Y	40.6%	28.9%	34.2%
N	39.7%	0%	60.3%
	41.2%	28.8%	34.5%

Purchase | Yahoo

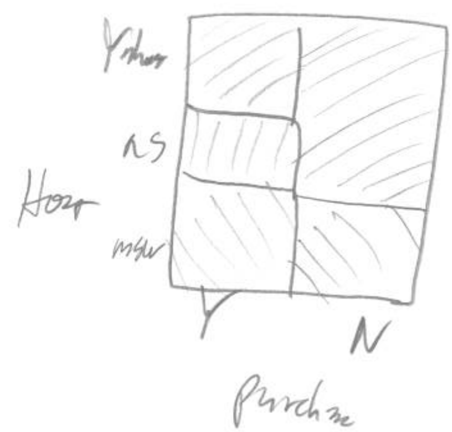
each column and row is a conditional distr.

You must calculate the %age differently in each case

It seems that the proportion of purchases varies based on the host. That means the host and the decision to purchase are associated.

The different vocabulary words: correlated, dependent, causation
we'll get to these at the end of the week.

Another way to visualize data is via the segmented bar chart (p 81) and mosaic plots (p82, 83). These charts make it easy to see association.



You can visualize association much more easily

How strong is association?

We'd like to test the association (in categ. var's) if possible because there's always some due to chance...

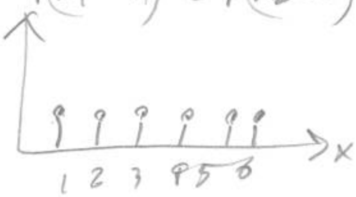
Let's review r.v.'s again... bit of ch. 10.2

Two dice $X_1, X_2 \overset{iid}{\sim}$ Dice

identical distr.

PMF

$P(X_1 = x) = P(X_2 = x)$



What does independence really mean?

$P(X_1 = x_1, X_2 = x_2)$ ← "Joint" PMF

AKA Joint prob. distribution

$X_1 \backslash x_2$	1	2	3	4	5	6	
1	1/36	1/36					1/6
2							1/6
3							1/6
4							1/6
5							1/6
6							1/6
	1/6	1/6	1/6	1/6	1/6	1/6	1

marginal distr of X_1

1 ← always 1

marginal distr of X_2

Independence is when $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)\, P(X_2 = x_2)$

for all possible states of x_1, x_2 ... (p.222)

Are our two dice independent above?

Ho: independence Ha: not independence

We essentially want to test whether or not variables are independent... If they're independent, the conditionals must be the same as the marginals. For instance: example in book p89-91

Poll 200 people attitude towards sharing music online

Ho: Group, Attitude $\perp$
Ha: ~
$\alpha = 5\%$

Observed

Attitude

Group	OK	Not OK	
Staff	30	70	100
Students	50	50	100
	80	120	200 = n

$\Longrightarrow$

Expected

	OK	Not OK		
Staff	40 (20%)	60 (30%)	100	(50%)
Students	40 (20%)	60 (30%)	100	(50%)
	80	120	200 = n	
	(40%)	(60%)		

Under the null...

num rows

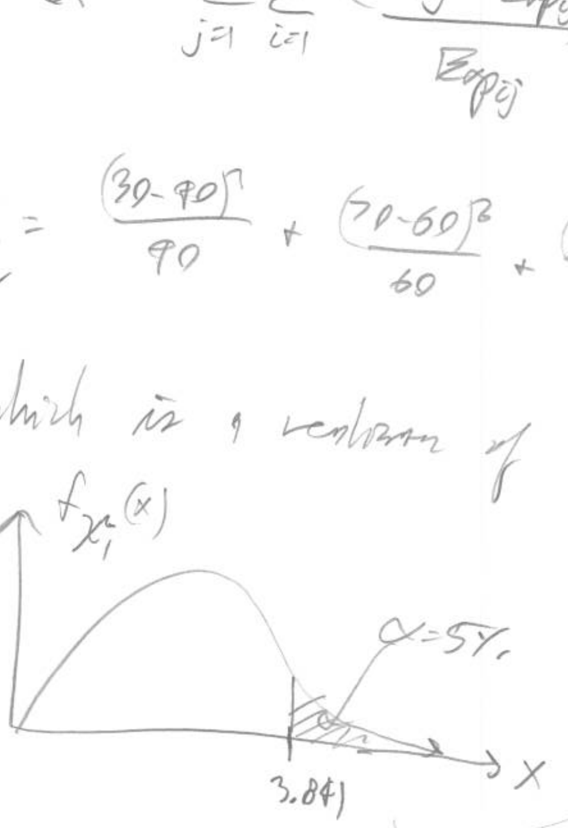
num cols

$$Q = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(Obs_{ij} - Exp_{ij}\right)^2}{Exp_{ij}} \sim \chi^2_{df} \quad \text{st. } df = (r-1)(c-1) = (2-1)(2-1) = 1$$

$$q = \frac{(30-40)^2}{40} + \frac{(70-60)^2}{60} + \frac{(50-40)^2}{40} + \frac{(50-60)^2}{60} = 8.33 > 3.841 = \chi^2_{5\%,1}$$

$\Rightarrow$ reject Ho

which is a realization of χ^2_1



$$p\text{val} = P\left(\chi^2_1 \geq 8.33\right) = .0039 < 5\% = \alpha$$

$\Rightarrow$ reject Ho

In-Class

Experiment: are two coin flips independent?

H_0: they are ind.

H_a: they aren't ind.

$\alpha = 1\%$

Observed

Coin 2

Coin 1	H	T	
H			
T			

Expected

Coin 2

Coin 1	H	T	
H			
T			

$$Q = \sum_{j=1}^{c} \sum_{i=1}^{r} \qquad \text{etc....}$$

~~Forget Cramer's V... p92/93 (not covered)~~

Just because two variables show association, does that mean we can say there's causation? No

$$\text{association} \not\Longrightarrow \text{causation}$$

Most subtle point in all of statistics... probably most important message of semester.. will do this tomorrow..