Lecture 23 (ch. 8 - end)

Back to proportions

When we compare 2 props (π_1, π_2) (e.g. $H_1: \pi_1 - \pi_2 < 0.1$)
we have TWO populations, each with 2 groups/categ. (e.g. Boy/Girl)
In That case, ONE proportion (eg. prop of boys, π_{Boy}) is enough
to describe each pop., because The other prop. (eg. π_{Girls}) is
fixed by $1 - \pi_{Boys}$. In other words, our 2-sample CIs/tests
involve TWO proportions, one from each of Two populations.

So, an example would be $\begin{cases} \pi_1 = \boxed{\pi_{Boys}} & \text{in Northern hemisphere.} \\ \pi_2 = \boxed{\pi_{Boys}} & \text{" Southern " .} \end{cases}$

Note That both π_1 and π_2 refer to Boys, but in 2 different
populations (e.g. Northern and Southern hemispheres).
Then, we can test, e.g. $H_0: \pi_1 - \pi_2 = 0$ vs. $H_1: \pi_1 - \pi_2 < 0.1$

But There are situations where each population has
more Than 2 categories, and we want to test
some claim about The proportions of each category.
In fact, we can even have ONE population with many categs.

If we have $\boxed{\text{ONE}}$ pop, with k categories, we can test

$H_0: \pi_1 = \pi_{01}, \ \pi_2 = \pi_{02}, \ \ldots, \ \pi_k = \pi_{0k}$, prop. of k^{th} categ. in pop.

H_1: Atleast one of ↑ is wrong

I'll explain This later.

$$\boxed{\sum_{i=1}^{k} \pi_{0i} = 1}$$

Of course, given That There is only ONE pop., we must have
Below, we will see how to do This test.
There will be a new distribution: **Chi-squared.**
Also note That a pop. with 2 groups can be Thought of as being
discribed by one random variable with 2 levels. Similarly, a pop.
with k groups canbe described with one r.v. with k levels.

E.g. Monthly Weather Review, 2008: Vol. 136, p. 3121. Cook & Schaefer.

Does data provide sufficient evidence to support an association between climate and tornadic activity?

	El Nino	La Nina	Normal	
# of Days with violent tornadoes in each climate category :	$n_1 = 14$	$n_2 = 28$	$n_3 = 44$	(86)
proportion :	$\frac{14}{86} = 0.16$	0.33	0.51	(1)

Data.

generalization of Binomial to more than 2 categories. FYI: Multinomial dist.

# of years classified as	12	17	25	(54)
proportion :	$\frac{12}{54} = 0.22$	0.32	0.46	(1)

H_0: There is no association, i.e.

true prop. of tornadic days in El Nino years. Etc.

H_0: $\pi_1 = 0.22$ $\quad \pi_2 = 0.32$ $\quad \pi_3 = 0.46$

H_1: At least one of these assignments is wrong.

Q If H_0 = True, how many tornadoes do you expect in each of the $k=3$ categories?

A

Expected counts :	$0.22(86) \simeq 18.9$	$0.32(86) \simeq 27.5$	$0.46(86) \simeq 39.6$	(86)
observed counts:	14	28	44	
$(\text{Exp.} - \text{obs})^2$:	$(4.9)^2$	$(-0.5)^2$	$(-4.4)^2$	
$\frac{(\text{Exp} - \text{obs})^2}{\text{Exp}}$:	1.27	0.009	0.49	

Like z_{obs}, t_{obs},

$$\chi^2_{obs.} \equiv \sum_{i=1}^{3} \frac{(\text{exp.} - \text{obs})^2}{\text{exp.}} = 1.77$$

If There were really no difference at all in The # of tornadoes between The 3 categories, then This would be near zero.

Q So, is This X^2_{obs} far away from 0 to reject H_0 (in favor of H_1)?

Note: X^2 is non-negative, unlike z, t

A We need to know The sampl. distr. of X^2, when $H_0 = T$.

Theorem: Under The null hypothesis, X^2 has a chi-squared distr. with $df = k-1$ $(= 3-1 = 2)$

What's a chi-squared dist? It's just another Table (VII). But FYI.

see two pages down

p-value = $\text{prob}(X^2 > X^2_{obs}) = \text{prob}(X^2 > 1.77) > 0.1$

$df = 3-1 = 2$

R: pchisq(1.77, df=3-1, lower.tail=F) = 0.41

For The chi-squared test, p-value is alway right area.

Conclusion (at $\alpha = .01$): p-value > α

In words: Cannot reject H_0 in favor of H_1. atleast 1 is wrong.

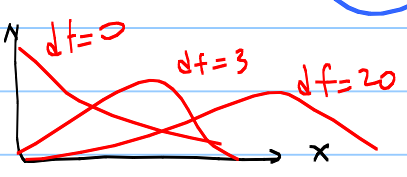
$(\pi_1 = .22, \pi_2 = .32, \pi_3 = .46)$

In English: There is no evidence from data to suggest That The 3 props are NOT .22, .32, .46, ie.

I.e. There is no evidence from data That There is an association between tornadic activity and climate.

The chi-squared density function is FYI

$$f_{df}(x) = \frac{x^{\frac{df}{2}-1} e^{-x/2}}{\Gamma(\frac{df}{2})\, 2^{df/2}}$$



(*)

Here is a generalization of the above to any k (above k=3).
I propose that you do NOT use the formula at the bottom, but instead do things like I did above. After you are completely comfortable with the steps, then you can use this page.
But Do read the IMPORTANT Note at the bottom.

Let π_i = proportion of cases in category i (where $i = 1, 2, \cdots, k$).

	Null params	Tornado Example
π_1 = proportion of categ. 1's	π_{01}	0.22
π_2 = - - - - 2's	π_{02}	0.32
π_3 = - - - - 3's	π_{03}	0.46

If H_0 = True, H_0: $\pi_1 = \pi_{01}$, $\pi_2 = \pi_{02}$, ...

Then in a sample of size n, how many would

we expect in category 1 :	$n\pi_{01}$	18.9
" " " " 2 :	$n\pi_{02}$	27.5
" " " " 3 :	$n\pi_{03}$	39.6
⋮		

$\sum_{i=1}^{k} n_i = n$

But according to data, we observe this many :	n_1	14
	n_2	28
	n_3	44

Punch line:
Then the theorem tells us that

$$X^2_{obs} = \sum_i^k \frac{(\text{exp.} - \text{obs})^2}{\text{exp.}} = \sum_{i=1}^{k} \frac{(n\pi_{0i} - n_i)^2}{n\pi_{0i}}$$

has a chi-sqd. distr with df = k-1

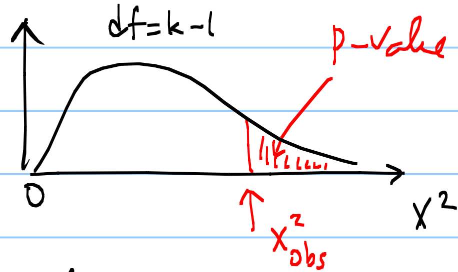
p-value = prob($X^2 > X^2_{obs}$)

IMPORTANT

First, note that we are dealing with props, not means (i.e. our r.v. is discrete)
But, even though we wrote a lot of things in terms of props, the data are all in counts. This chi-sqd test deals with count data

How to use Table VII:

Table VII gives the area to the right of some value of x^2_{obs}, ie. it gives a p-value. However, it does not give all p-values; the only ones it provides are listed in the left-most column. E.g.



$x^2_{obs} = 8.49$, df=4 $\Rightarrow$ p-value = 0.075

$x^2_{obs} = 8.66$, df=4 $\Rightarrow$ p-value = 0.070

One might think that putting bounds on p-value is not enough for hypothesis testing, but it often is. For example, suppose we got $x^2_{obs} = 8.55$ with df=4. Then we can say $0.070 < \text{p-value} < 0.075$. That is good enough if $\alpha = .05$, because p-value > α, and so we cannot reject H_0 in favor of H_1.

Additional comments:

Note that the first H_0, H_1 above are just a generalization of

$H_0: \pi = \pi_0$ (z-test).

$H_1: \pi \neq \pi_0$

to more than 2 categories in the population.

However, there are $\boxed{\text{no}}$ 1-sided / 2-sided varieties of chi-sqd.

When X^2_{obs} is small (say ~ 0), then the observed counts are consistent with the expected counts if $H_0 = T$ (i.e. $\pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \ldots \pi_k = \pi_{0k}$). So, if X^2_{obs} is large, then at least one of these must be wrong.

In other words the appropriate hypotheses are

$H_0: \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \ldots \pi_k = \pi_{0k}$

H_1: At least one of these specifications is wrong.

And it is the "At least" which gives us

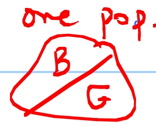
p-value = prob($X^2 > X^2_{obs}$) (Table VII)

I.e. We are always interested in the upper tail area only

Said differently for the chi-sqd test of the above H_0/H_1, the p-value is only the right area, because violation of each part of H_0, increases X^2.

Summary

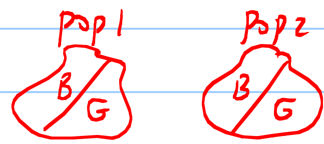
⇒ 1 pop., 1 prop.



$\pi_B = ?$ $\quad (\pi_B + \pi_G = 1)$

Data: $p_{obs} = 13/15$ $\quad z_{obs} = (p_{obs} - \pi_0) \Big/ \sqrt{\frac{\pi_0(1-\pi_0)}{n}} \Rightarrow$ p-value.

⇒ 2 pop.'s, 2 props.

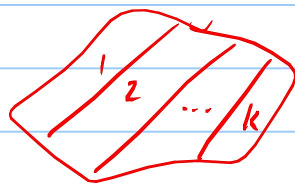


$\pi_{1B} - \pi_{2B} = ?$ $\quad (\pi_{1B} + \pi_{2B} \neq 1)$

Data: $p_1 = 13/15$ $\quad p_2 = 23/40$ $\quad z_{obs} = \frac{(p_1 - p_2) - (\pi_{1B} - \pi_{2B})}{\sqrt{\cdots}} \Rightarrow$ p-value

Note that when we write $H_0: \pi_1 - \pi_2 = \cdots$, the π_i are the proportion of 1 of the 2 categories in either pop. e.g.
π_1 = prop of something (e.g. boys) in population 1, and
π_2 = " " same thing " " 2.

⇒ 1 pop., k categories
(The tornado example)



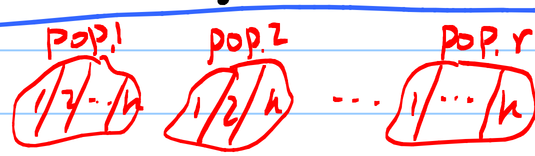
$\pi_1 = ?, \ \pi_2 = ?, \ \ldots, \ \pi_k = ?$
$(\pi_1 + \pi_2 + \cdots + \pi_k = 1.)$

Data: 14 / 28 / 44
observed counts

$X^2_{obs} = \sum \left(\frac{obs - exp}{\sqrt{exp}} \right)^2 \sim$ chisqd
df = k − 1

Note that when we talk about $\pi_1, \pi_2, \ldots, \pi_k$, the indices refer to k levels of some categorical r.v.

⇒ r pops, k categories



$i = 1, \ldots, k$
$j = 1, \ldots, r$

Data: Contingency Table
(Confusion Matrix)

Category

pop.	1	2	...	k
1				
2				
⋮				
r				

counts

Skipped

Skipped

$X^2 = \sum_{\text{all cells}} \left(\frac{obs - exp}{\sqrt{exp}} \right)^2 \sim$ chisqd $\Rightarrow$ p-value.
df = (r−1)(k−1)

hw_lect23_1 (By hand)
Consider the data from an example in a past lecture where a survey of students in 390 yielded the following data:
17 students like lab
48 do not like lab
15 have no opinion
Suppose I believed that the proportion of students in each of the 3 categories (like, no-like, no-opinion) was equal. Does this data contradict that belief? Use alpha = 0.05

hw_lect23_2 (By hand)
A sample of 210 Bell computers has 56 defectives. Theory suggests that a third of all Bell computers should be defective. Does this data contradict the theory (at alpha=0.05)? Specifcially,
a) Do a z-test ,
b) Do a chi-squared test with k=2 categories. Hint: The pi's (and pi_0's) of the k categories must sum to 1.
c) Are the conclusions in a and b consistent?