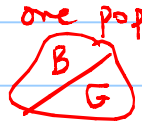
Lecture 22 (Ch. 8)

In Ch. 7, we learned how to build CIs for either 1 prop, π_x, or the difference between 2 props, $\pi_1 - \pi_2$, where
π_1 = prop of something (e.g. boys) in population 1, and
π_2 = " " same thing " " 2.

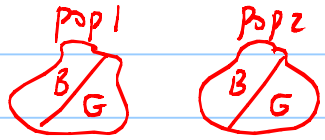
We also learned how to do hyp. tests on π_x, or $\pi_1 - \pi_2$.

[Note $\pi_1 + \pi_2 \neq 1$, because π_1, π_2 are 2 different populations]

But in all of these situations, the 2 pops have 2 categories (boy/girl) and π_i is the prop. of 1 of them.

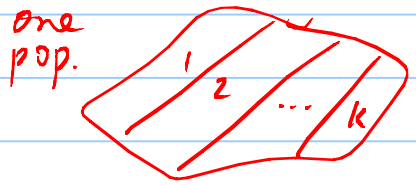


$\pi_B = ?$



$\pi_{1B} - \pi_{2B} = ?$

The tornado/climate eg. in prev. lecture deals with the situation where ONE population has 3 categories. For k categories:



$\pi_1 = ?$, $\pi_2 = ?$, ..., $\pi_k = ?$

We learned that the relevant dist. is chi-squared with df = k-1. And the quantity that follows that dist. is

like z, t $$\chi^2 = \sum_{i=1}^{k} \left(\frac{obs_i - exp_i}{\sqrt{exp_i}} \right)^2$$ → Order matters in interpretation (below).

where obs_i and exp_i are observed and expected **counts** in the i^{th} category (still of 1 population). The latter are computed assuming H_0 is true, where

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

H_1: At least one of these is wrong.

[This time $\pi_1 + \pi_2 + \cdots = 1$
$\pi_{01} + \pi_{02} + \cdots = 1$]

Note that the above H_0, H_1 is 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 [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 (ie. $\pi_1 = \pi_{01}$, $\pi_2 = \pi_{02}$, ... $\pi_k = \pi_{0k}$). 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: $\pi_1 = \pi_{01}$, $\pi_2 = \pi_{02}$, ... $\pi_k = \pi_{0k}$

H_1: At least one of these↑ specifications is wrong.

And it is the "At least" which gives us

p-value = prob($X^2 \geq X^2_{obs}$) (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.

Interpretation / Diagnosis

The magnitudes of the k terms in X^2_{obs} are important in deciding which of the k proportions are most different from the expected props. (under H_0).

Example: In the tornado example

Suppose we had found p-value < .a, ie. there is evidence that climate does affect tornadic activity. Then X^2_{obs} would be big. But what makes it big? The 3 terms contributing to X^2_{obs} are:

El Nino	La Nina	Normal
1.27 (Large)	0.009 (small)	0.49

Then we could conclude that it is the El Nino years which differ most (in terms of tornadic activity) from what we would expect under H_0 (ie. if climate had no effect on tornadic activity).

And we could say that tornadic activity in La Nina years is pretty close to what one would expect by chance.

How about the "direction" of the association? Eg.
Are there more tornadoes in El Nino years than in Normal years?
Look at the signs of the k term is X^2, BEFORE squaring:

El Nino	La Nina	Normal

$$X^2_{obs} = \frac{(+4.9)^2}{18.9} + \frac{(-0.5)^2}{27.5} + \frac{(-4.4)^2}{39.6}$$

Note order!

In this formula, we looked at $(\text{expected} - \text{obs})^2$

So in El Nino years: exp. > obs $\Rightarrow$ Less tornadic than expected
in La Nina years: exp < obs $\Rightarrow$ More tornadic " "

Q1: An information retrieval system has 3 storage locations, and it is designed with the expectation that the long-run proportion of requests for the 3 locations is $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, respectively. According to observations, however, of the 12 requests made over some period of time, the number of requests from the 3 locations is 2, 6, 4, respectively. We want to know if the design expectations are inconsistent with observations. So, we do a chi-squared test. The value of X^2_{obs} is

A) $\frac{(\frac{1}{4} - \frac{2}{12})^2}{\frac{2}{12}} + \frac{(\frac{1}{2} - \frac{6}{12})^2}{6/12} + \frac{(\frac{1}{4} - \frac{4}{12})^2}{4/12}$

B) $\frac{(\frac{12}{4} - 2)^2}{12/4} + \frac{(\frac{12}{2} - 6)^2}{12/2} + \frac{(\frac{12}{4} - 4)^2}{12/4}$ ← $\frac{(\text{exp.} - \text{obs})^2}{\text{exp.}}$

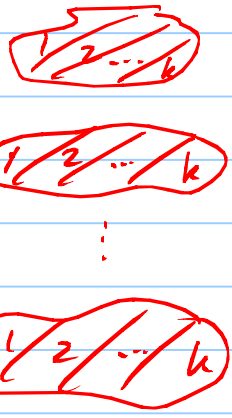
counts not props.

C) $\frac{(\frac{1}{4} - 2)^2}{1/4} + \frac{(1/2 - 6)^2}{1/2} + \frac{(1/4 - 4)^2}{1/4}$

D) $\frac{(\frac{12}{4} - \frac{2}{12})^2}{12/4} + \frac{(\frac{12}{2} - \frac{6}{12})^2}{12/2} + \frac{(12/4 - 4/12)^2}{12/4}$

The chi-sqd distr. shows up in 2 other situations.

2) $\boxed{k}$ props across $\boxed{r}$ populations:



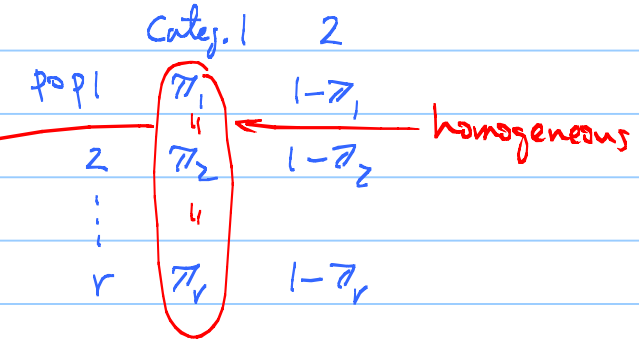
H_0: r pops are homogeneous w.r.t. k categories
H_1: not ↑ big concept!

Homogeneous means

pop 1	π_1	π_2	π_3	$\cdots$ π_k	$\rightarrow \sum_{i=1}^{k} \pi_i = 1$
pop 2	π_1	π_2	π_3	$\cdots$ π_k	$\rightarrow \sum_{i=1}^{k} \pi_i = 1$
⋮					
pop r	π_1	π_2	π_3	$\cdots$ π_k	$\rightarrow \sum_{i=1}^{k} \pi_i = 1$

Not homogeneous means that at least 1 of these is wrong.

E.g. for $k=2$, the above H_0/H_1 translate to:



H_0: $\pi_1 = \pi_2 = \cdots = \pi_r$
H_1: At least 2 πs are different

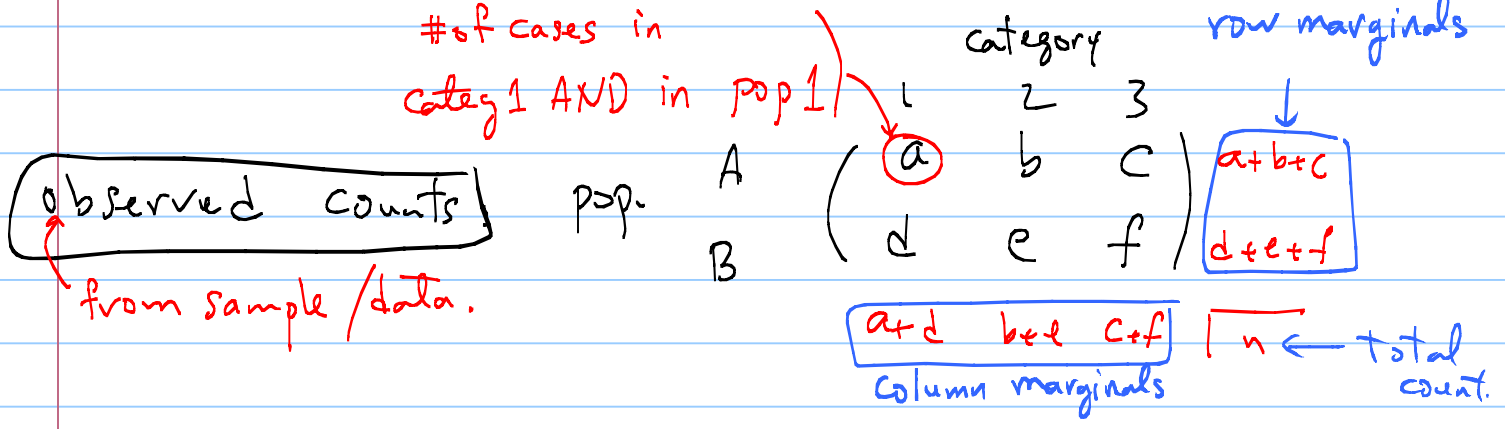
[Note: This is diff. from $\pi_1 = \pi_{01}, \pi_2 = \pi_{02} \cdots$]

3) Independence of 2 categorical variables, one with $\boxed{k}$ levels, the other with $\boxed{r}$ levels.

Same test, diff. conclusion

H_0: 2 categ. vars. are indep.
H_1: " " " " not indep.

In such problems, the data are shown as a Contingency Table:



The good news is that one can test homogeneity with a chi-sqd dist., but with $df = (k-1)(r-1)$.

k: # of cols / # of categories; r: # of rows / # of populations

I.e. compute $X^2_{obs} = \sum_{\text{all cells}} \left(\frac{\text{obs} - \text{exp}}{\sqrt{\text{exp}}}\right)^2$, and p-value $= pr(X^2 > X^2_{obs})$

(counts, not props)

Table VII.

The only question is what are the expected counts?

Expected counts: (Assuming $H_0 = T$.)

$$\begin{pmatrix} \frac{(a+b+c)(a+d)}{n} & \frac{(a+b+c)(b+e)}{n} & \cdots \\ \frac{(d+e+f)(a+d)}{n} & \cdots & \cdots \end{pmatrix}$$

(row-marginal, column marginal)

Remember this result a "$\frac{\text{row} \times \text{col. marginals}}{n}$".

(It's not easy to see this.)

Eg. Are boys & girls homogeneous w.r.t. belief in afterlife?

Here are the data in the form of a Contingency Table:

	Yes	Un.	No	
Boys	435	58	89	582
Girls	375	50	84	509
	810	108	173	1091

Expected Counts

$$\begin{pmatrix} \frac{(582)(810)}{1091} & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \begin{matrix} 582 \\ 509 \end{matrix}$$

810 108 177

$$= \begin{pmatrix} 432.1 & 57.6 & 92.3 \\ 377.9 & 50.4 & 80.7 \end{pmatrix}$$

$$X^2_{obs.} = \frac{(435-432.1)^2}{432.1} + \frac{(58-57.6)^2}{57.6} + \cdots$$

See interpretation/diagnosis below.

$$= .019 + .0028 + \boxed{0.118} +$$

$$0.022 + .0032 + \boxed{0.135}$$

$$= 0.3$$

"big" numbers, both in the "No" category.

Signs of (obs − exp)

$$\begin{pmatrix} + & + & - \\ - & - & + \end{pmatrix}$$

$df = (2-1)(3-1) = 2 \Rightarrow$ p-value > 0.1 (huge)

← Table VII

Cannot reject H_0 in favor of H_1, at $\alpha = .01$ (or .05)

I.e. there is no evidence to think that Boys and Girls are not homogeneous ("different") w.r.t. their belief in afterlife.

Also There is no evidence that Gender and Belief in afterlife are not independent. (ie. They "are" independent.)

Interpretation / Diagnosis

So, based on this data, we cannot say that there is a difference between boys & girls w.r.t. their belief in afterlife. Mathematically, the reason is that X^2_{obs} was too small.

But suppose, X^2_{obs} had turned out to be huge. Then we could conclude that there is a difference between boys & girls in terms of their belief in afterlife. Then just as before, we can look at the relative size of the various terms in X^2_{obs} to see which ones make the X^2_{obs} big.

In this example, the big terms are 0.118, 0.135, which correspond to the "No" category.

In short, if the result had turned out to be statistically significant (ie. X^2_{obs} = huge, p-value $< \alpha$), then we could go further and say that the biggest difference between boys and girls (in terms of their belief in afterlife) is in the non-believer category.

The signs can be interpreted, too; but we'll skip it for now.

Summary: Chi-sqd shows up in 3 situations:

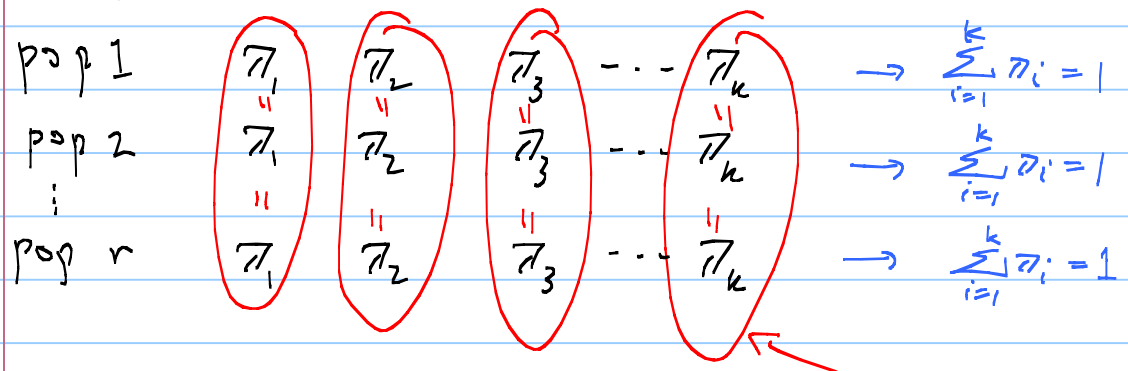
I) 1 pop. (1 variable) with k categories.

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

H_1: At least one of π is wrong.

Because there is only 1 pop, must have $\sum_{i=1}^{k} \pi_{0i} = 1$.

II) r pops. each with k categories:



H_0: The above π's are equal as specified.
(i.e. The r pops are homogeneous w.r.t. the k categories)

H_1: For at least 1 of the k categories, the proportions are not equal for all pops.
(i.e. the r pops are not homogeneous ...)

III) Are 2 categorical variables independent?

H_0: they are independent.

H_1: they are not.

hw_lect22-1 By hand

Have you ever wondered whether soccer players suffer adverse effects from hitting "headers"? The authors of the article "No Evidence of Impaired Neurocognitive Performance in Collegiate Soccer Players" (The Amer. J. of Sports Medicine, 2002: 157Â162) investigated this issue. The paper reported that 45 of the 91 soccer players in their sample had suffered concussion, 28 of 96 nonsoccer athletes had suffered concussion, and only 8 of 53 student controls had suffered concussion. Denote
pi1 = pop. proportion of concussions among soccer players,
pi2 = pop. proportion of concussions among non-soccer players,
pi3 = pop. proportion of concussions among control group.
Set up this problem as a test of homogeniety of three populations with respect to 2 categories. Specifically,

a) State the hypotheses in terms of pi1, pi2, pi3.
b) Write the data in the form of a contingency table.
c) Compute the expected counts.
d) Compute the p-value (or specify a range for it).
e) State the conclusion "in English."
f) Diagnose the various terms appearing in X_obs^2.

hw_lect22-2 By R

The accompanying data resulted from an experiment in which seeds of five different types were planted and the number that germinated
within 5 weeks of planting was observed for each seed type ("Nondestructive Optical Methods of Food Quality Evaluation," Food
Science and Nutr., 1984: 232-279). Carry out a chi-squared test at level .01 to see whether the proportion of seeds that germinate in the
specified period varies according to type of seed.
#
Seed type: 1 2 3 4 5
Germinated: 31 57 87 52 10
Failed to germinate: 7 33 60 4 4 19
#
Specifically,
#
a) Does the statement of the problem require a test of homogeniety of 2 populations with respect to 5 categories, or vice versa?
b) Compute the X^2, the df, and the p-value corresponding to your answer in part a.
c) State your conclusion "in English," at significance level 0.05.
d) Diagnose the magnitude of the various terms in X_obs^2

hw_lect22-3

Consider 1 pop. with 2 categories (say A, B), and let π_A, π_B denote the proportion of A's and B's in the pop. Note $\pi_A + \pi_B = 1$.

In a hw problem (about Bell computers) you see that the 1-sample 2-sided z-test of $\begin{cases} H_0: \pi_A = \pi_0 \\ H_1: \pi_A \neq \pi_0 \end{cases}$ gives the same p-value as the

chi-squared test of proportions in 1 pop.

$$\begin{cases} H_0: \pi_A = \pi_0 \text{ , } \pi_B = (1-\pi_0) \\ H_1: \text{At least one of these is wrong} \end{cases}$$

This equivalence can be seen at the level of H_0 and H_1, too; note that "At least one of these is wrong" translates to $\pi_A \neq \pi_0$, because $\pi_A + \pi_B = 1$; and if $\pi_A = \pi_0$ is wrong, then so is $\pi_B = 1 - \pi_A$.

Now, consider 2 populations, each with 2 categories. Denote the 2 pops with "1", "2", and the 2 categories as "A", "B". So, π_{1A} = prop. of A's in 1st pop., π_{2A} = ..., etc.

a) Write down the H_0/H_1 that test homogeneity of pops w.r.t. categories

b) Show that the H_0/H_1 in part a) are equivalent to H_0, H_1 of the 2-sample 2-sided z-test.