Lecture 25 (Ch-8)

In the procedure we have learned, the last step involves comparing the p-value with α. That practice is (slowly) become "old style". More recently, one reports the p-value itself, because by itself it's useful - it reflects the evidence from data against H_0.

But, α does have an important interpretation nevertheless. We know that it is the largest prob at which we are confident to reject H_0 in favor of H_1. But there is more to it!

Suppose we are testing $H_0: \mu \le \mu_0$ vs. $H_1: \mu > \mu_0$.

We assume H_0 = True (ie. $\mu = \mu_0$), then compute a p-value.

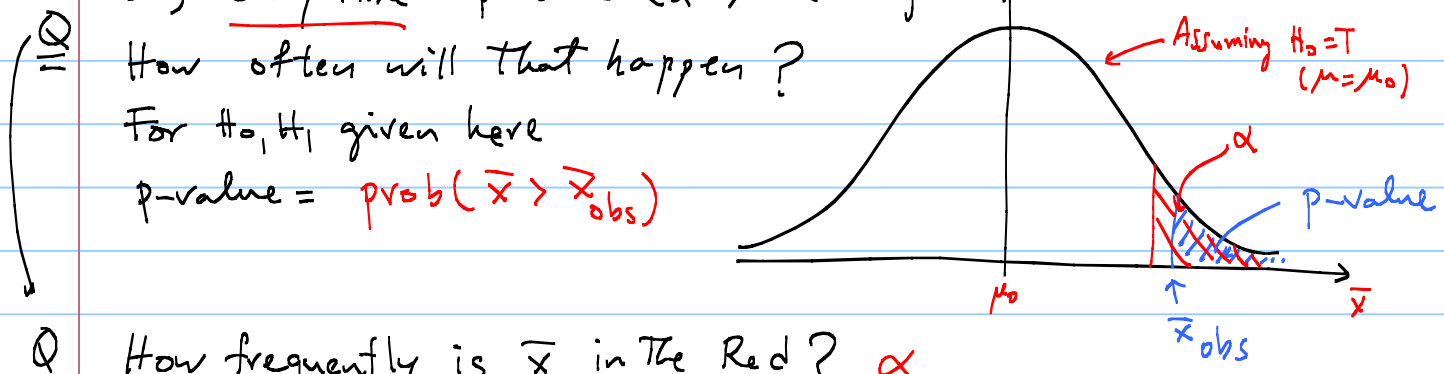
If p value $< \alpha$, Then Reject H_0 in favor of H_1.

So, every time p-value $< \alpha$, we reject.

How often will that happen?

For H_0, H_1 given here

p-value = prob($\bar{x} > \bar{x}_{obs}$)



How frequently is $\bar{x}$ in the Red? α

Note: α = prob(p-value $< \alpha \mid H_0 = T$)

So, α = prob(Data Reject H_0 in favor of $H_1 \mid H_0 = T$)

Type I error

"Bad" error

"False Alarm Rate"

(Convicting an innocent person.)

This is how you decide on α.

How much bad error can you tolerate in the long run?

Why not set $\alpha = 0$, so that we will not have any bad (Type I) errors?

Because there is another kind of error:

β = prob (Data cannot reject H_0 in favor of H_1 | H_0 = False)

Type II (Releasing a guilty person.)

Setting $\alpha = 0 \Rightarrow \beta = 1$.

α, β (The probs of the 2 types of errors) have a complex but mostly inverse relationship, depending on n (p. 389-391)

So, given that α is the prob of the bad error, we generally set α at a fixed, but low, value.

Obviously, this will lead to some nonzero β, and it is important to compute it for your own specific problem. Sometimes, people look at $1-\beta$ (instead of β).

$1-\beta$ = power = prob(Rejecting H_0 ---- | H_0 = F)

(Convicting a guilty person).

If there is time, we'll return to power.

The understanding that $\alpha = pr(\text{Type I error}) = pr(\text{Bad error})$ offers another way of setting H_0/H_1 correctly.

Example: NASA

A company manufactures computer screens for used by astronauts on space missions. If more that 10% of the pixels on a given screen are defective, then the company does not release the screen to NASA, because otherwise disasters will occur.

which is the appropriate H_0/H_1 ?

π = prop. of defective pixels. μ = mean # of defective pixels.

(A) $H_0: \pi < 0.1$ $H_1: \pi > 0.1$

(B) $H_0: \pi = 0.1$ $H_1: \pi \neq 0.1$

(C) $H_0: \pi > 0.1$ $H_1: \pi < 0.1$

OK error

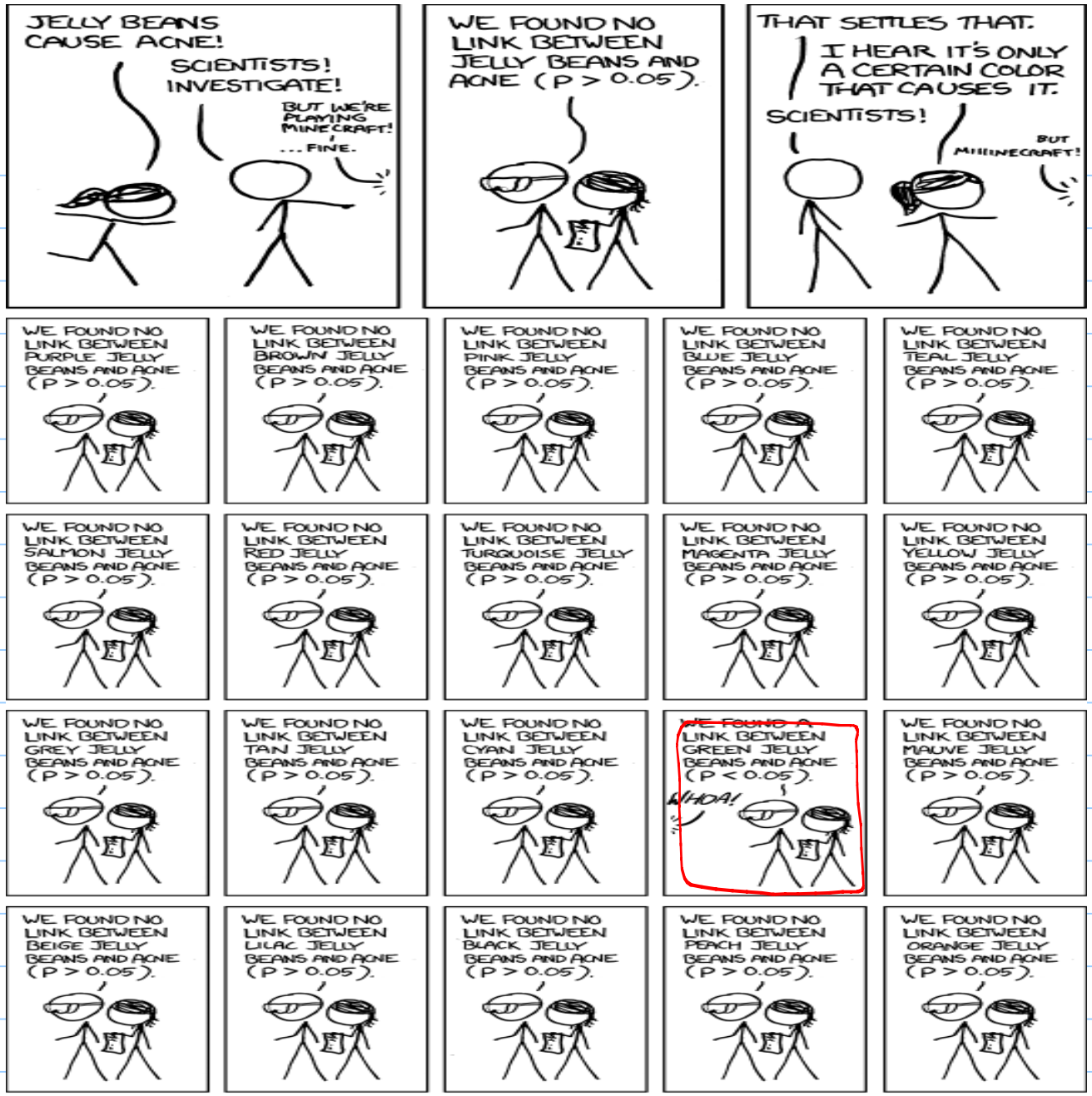
(Data say $\pi > 0.1 \mid \pi < 0.1$)

Bad error

(Data say $\pi < 0.1 \mid \pi > 0.1$)

$\alpha = pr(\text{Data say } H_1 \mid H_0 = T)$

α is Dangerous!



Suppose you are testing whether a drug has $\mu > 0$.

So:

$H_0: \mu < 0$, $H_1: \mu > 0$

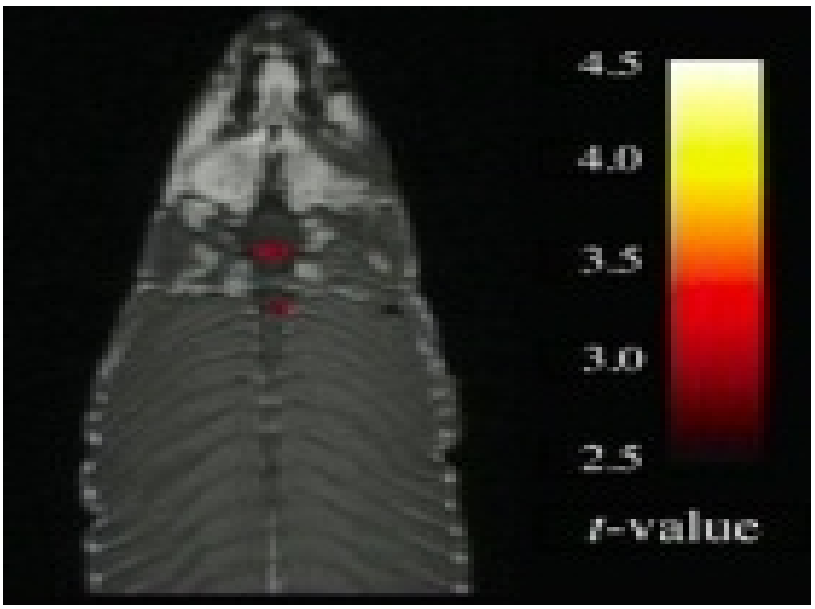
Suppose you compute the p-value and find p-value $> \alpha$, i.e. there is no evidence that $\mu > 0$. If you repeat the experiment many times, eventually you will find p-value $< \alpha$, ie. there is evidence that $\mu > 0$. This will happen (at most) α% of the time even if, in fact, $\mu < 0$. I.e. α% of the time, you will make a type I error.

The conf. level (e.g. 0.95), and the significance level (e.g. 0.05) (α) play an important role in every thing we've done. The last revealed what α really is: $\alpha = \text{prob}(\text{Type I}) = \text{prob}(\text{Bad error})$

$= \text{pr}(\text{Data reject } H_0 \ldots \mid H_0 = T)$

So, we now have another way of setting H_0, H_1 ↗.

Also, we now see that we will be wrong (Bad wrong) $\alpha\%$ of the time. One example was the Jelly Bean example, here is another:



Dead thinking Salmon!

FYI

There exist other decision-making frameworks which avoid such problems (e.g. check out
- multiple hypothesis testing
- False Discovery Rate)

Alternatively, in some situations, one can simply report the p-value, without comparing it to α.

In this class, we will continue to compare it with α, but be aware of this "defect"

Summary

We are done with 1-sample and 2-sample, z and t-tests, for paired and unpaired data, but all of that has dealt with the pop. means. What about pop. proportions?

Easy! Follow the pattern:

1 sample

C.I. for μ_x:

$\bar{x} \pm z^* \frac{\sigma_x}{\sqrt{n}}$ or $\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$, $df = n-1$

C.I. for π_x:

$p \pm z^* \sqrt{\frac{p(1-p)}{n}}$ ~~No $p \pm t^* \sqrt{\frac{p(1-p)}{n}}$~~

test for μ:

$H_0: \mu \square \mu_0$ $H_1: \mu \square \mu_0$

$z_{obs} = \frac{\bar{x}_{obs} - \mu_0}{\sigma_x/\sqrt{n}}$ or $t_{obs} = \frac{\bar{x}_{obs} - \mu_0}{s_x/\sqrt{n}}$, $df = n-1$

p-value = ...

Test for π:

$H_0: \pi \square \pi_0$ $H_1: \pi \square \pi_0$

$z_{obs} = \frac{p_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ ← No p! Because we assumed $H_0 = T$, ie. $\pi = \pi_0$.

p-value = ...

2 sample

C.I. for $\mu_2 - \mu_1$:

$\bar{x}_2 - \bar{x}_1 \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $\pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, df = Welch

Test for $\pi_2 - \pi_1$:

$p_2 - p_1 \pm z^* \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ ~~No t~~

Test for $\mu_2 - \mu_1$:

$H_0: \mu_2 - \mu_1 \square \Delta$ H_1: ...

$z_{obs} = \frac{(\bar{x}_2 - \bar{x}_1)_{obs} - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ or $t_{obs} = \frac{(\bar{x}_2 - \bar{x}_1)_{obs} - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, df = Welch

Test for $\pi_2 - \pi_1$:

$H_0: \pi_2 - \pi_1 \square \Delta$ ...

$z_{obs} = \frac{(p_2 - p_1) - \Delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$

When we compare 2 props (π_1, π_2), e.g. $H_0: \pi_1 - \pi_2 < 0.1$ the implication is that we have two populations, each with 2 groups/categ. (eg. Boys, Girls).
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 $1 - \pi_{Boys}$. The 2-sample z-test we have developed involves two proportions, one from each of two populations.
So, an example would be $\pi_1 = \pi_{Boys}$ in Northern hemisphere.

$\pi_2 = \pi_{Boys}$ " Southern " .

Note that both π_1 and π_2 refer to Boys, but in 2 different populations (e.g. Northern and Southern hemispheres).

But there are situations where we have ONE population, with more than 2 categories, and we want to test some claim about the proportions of each category.

If we have 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.

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

Of course, given that there is only ONE pop., we have $\sum_{i=1}^{k} \pi_{0i} = 1$
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.

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

H_0: true prop. of tornadic days in El Nino years. Etc.
There is no association, ie.

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

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 count:	0.22 (86) ≈ 18.9	0.32 (86) ≈ 27.5	0.46 (86) ≈ 39.6	(86)
Observed count:	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 χ^2_{obs} far away from 0 to reject H_0 (in favor of H_1)?
Note: χ^2 is non-negative, unlike z, t

A: We need to know the sampl. distr. of χ^2, when $H_0 = T$.

Theorem: Under the null hypothesis, χ^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.

p-value = $prob(\chi^2 > \chi^2_{obs}) = prob(\chi^2 > 1.77) > 0.1$ (see a few pages down)
$\uparrow$ $df = 3-1 = 2$

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

In words: Cannot reject H_0 in favor of H_1. (at least 1 is wrong.)
$(\pi_1 = .22, \pi_2 = .32, \pi_3 = .46)$

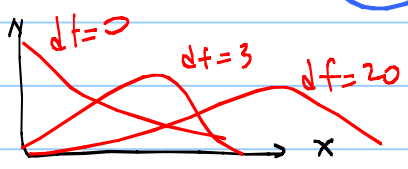
In English: There is no evidence from data to suggest that the 3 props are NOT .22, .32, .46, ie.

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

For the chi-squared test, this sign is always $>$! See below for expl.

The chi-squared density function is (FYI)

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



Summary / generalization

Now, let's generalize the above example to k categories (3, above):

Let π_i = proportion of cases in category i:

		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

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

counts, not proportions!

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

Note that the above H_0, H_1 is just a generalization of

$H_0: \pi = \pi_0$
$H_1: \pi \neq \pi_0$ (z-test).

to more than 2 categories in the population. (See hw)

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

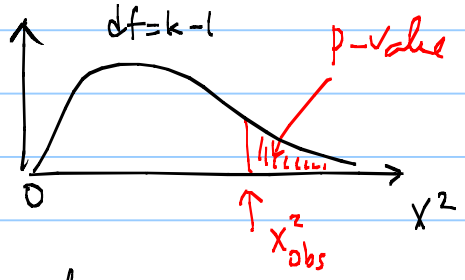
p-value = prob($X^2 \geq X^2_{obs}$) (Table VII)

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.

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 get $x^2_{obs} = 8.55$ with $df = 4$. Then we can say $0.070 < \text{p-value} < 0.075$. That is good enough if $\alpha = .05$, because p-value $> \alpha$, and so we cannot reject H_0 in favor of H_1.

FYI The same chi-sqd test that we have developed above can be used to see if the row-variable and the column-variable in a matrix of counts are independent.

Eg.

X \ Y	0	1
0	10	20
1	15	35

where 10 is the number cases with $x=0$, $y=0$, etc.

This chi-squared test is called the test of homogeniety

Ask me, if you need to know more.

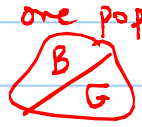
Summary

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.

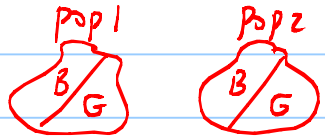
We also learned how to do hyp. tests on π_x, or $\pi_1 - \pi_2$.

[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.

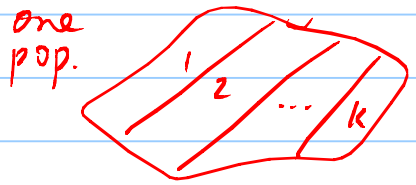


$\pi_B = ?$



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

The tornado/climate eg. in this 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

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{obs_i - exp_i}{\sqrt{exp_i}} \right)^2$$

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: $\pi_1 = \pi_{10}$, $\pi_2 = \pi_{20}$, ..., $\pi_k = \pi_{k0}$
H_1: At least one of these is wrong.

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

FYI

Further Diagnosis:

The magnitude of the k terms in X^2_{obs} is useful for diagnosing which of the k proportions differ most from the Null values.

Example: In the tornado example

Suppose we had found p-value $< \alpha$, ie. there is evidence that climate does affect tornadic activity. In that case X^2_{obs} would have been very large (so that p-value would be small).

The 3 terms contributing to x^2_{obs} are

1.27	0.009	0.49
El Nino	La Nina	Normal
large	small	

Then we could conclude that it is the EL Nino years which are most different (in terms of Tornadic activity) from what would be expected by chance (ie. if climate had no effect on tornadic activity).
In other words, we could conclude that the effect of climate on tornados is most in the El Nino years.
Tornadic activity in La Nina years, in fact, seems to be pretty close to what one would expect by chance.

Note: none of this tells us anything about the "direction" of the association. Are there more tornadoes in El Nino years than in Normal years? That's a different question that can be addressed by looking at the data directly. For example, one can look at # of tornadic days per EL Nino years, per LaNina years, ...
$\left(\frac{14}{12}, \frac{28}{17}, \frac{44}{25}\right) = (1.17, 1.64, 1.76)$. EL Nino years have "fewer" tornados.

hw-lect25-1

~~hw-lect24-1~~

hw-lect21-1 asked does it appear that π_x (the true proportion of defective screws) is at most 2.5%?
There, the appropriate interval is the upper conf. Bound for π_x.

a) which of the following is/are appropriate pair of hypotheses.

A)	B)	C)
$H_0: \pi_x \leq 2.5\%$	$H_0: \pi_x \geq 2.5\%$	$H_0: \pi_x = 2.5\%$
$H_1: \pi_x > 2.5\%$	$H_1: \pi_x < 2.5\%$	$H_1: \pi_x < 2.5\%$

b) Compute the p-value (using the data in hw-lect21-1)

c) At $\alpha = .05$, is the conclusion consistent with the conclusion from the CI approach in hw-lect21-1?

---

hw-lect25-2

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?

hw-lect25-3

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? Let $\alpha = .05$.