Lecture 21 (Ch. 8)

In the p-value approach to hypothesis testing, at the end of the procedure, we compare it with a pre-specified prob. α.
Last time we realized that

$$\alpha = pr(\text{Type I error})$$

where

$$\text{Type I error} = \left(\begin{array}{l} \text{Data reject } H_0 \text{ in} \\ \text{favor of } H_1 \end{array} \middle| \; \underbrace{H_0 = T}_{H_1 = F} \right)$$

The other error

$$\text{Type II} = \left(\begin{array}{l} \text{Data do not reject } H_0 \\ \text{in favor of } H_1 \end{array} \middle| \; \underbrace{H_0 = F}_{H_1 = T} \right)$$

The prob. of Type II error is denoted β.

And power $= 1 - \beta = pr\left(\begin{array}{l} \text{Data reject } H_0 \\ \text{in favor of } H_1 \end{array} \middle| \; H_0 = F \right)$

α, β (The probs of the 2 types of errors) have a complex but mostly inverse relationship, depending on n (Fig. 8.14, p.401)

By convention, we assign the "Bad error" to Type I.
Who decides what's a bad error? You do!
And this understanding of α, suggests another way to set-up H_0/H_1:

E.g. Guilt or Innocence?

Bad error = (Data say guilty | innocent)
Type I = (Data say H_1 | $H_0 = T$)

$\Rightarrow \begin{cases} H_0: \text{ innocent} \\ H_1: \text{ guilty.} \end{cases}$

power = pr(Data say guilty | guilty)

→ You can see why we usually set the value of α to very small.

Another example:

A company manufactures computer screens for use by NASA astronauts on space missions. If more that 10% of the pixels on a given screen are defective, then the company does not give the screen to NASA, because otherwise disaster will occur. For one screen, 16 pixels are examined, and it is found that 1 is bad. Should the company give the screen to NASA?

π = true/pop. prop. of defective pixels

$H_0: \pi \le 0.1$ $\quad$ $H_0: \pi = 0.1$ $\quad$ $\boxed{\begin{array}{l} H_0: \pi \geq 0.1 \\ H_1: \pi < 0.1 \end{array}}$

$H_1: \pi > 0.1$ $\quad$ $H_1: \pi \neq 0.1$

OK error $\quad$ Bad error

(Data say $\pi > 0.1 \mid \pi < 0.1$) $\quad$ (Data say $\pi < 0.1 \mid \underline{\pi > 0.1}$)

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

$$\text{p-value} = pr(p < p_{obs} \mid \overbrace{H_0 = T}^{\pi \geq 0.1}) = pr\left(\overbrace{\frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}}^{z} < \frac{p_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \;\middle|\; \pi = 0.1\right)$$

see summary in last lect.

sufficient to test $H_0: \pi = \pi_0$

$p_{obs} = \frac{1}{16}$

$$= pr\left(z < \frac{0.0625 - 0.1}{\sqrt{\frac{.1(.9)}{16}}}\right) = pr\left(z < \frac{-0.0375}{(.3/4)}\right)$$

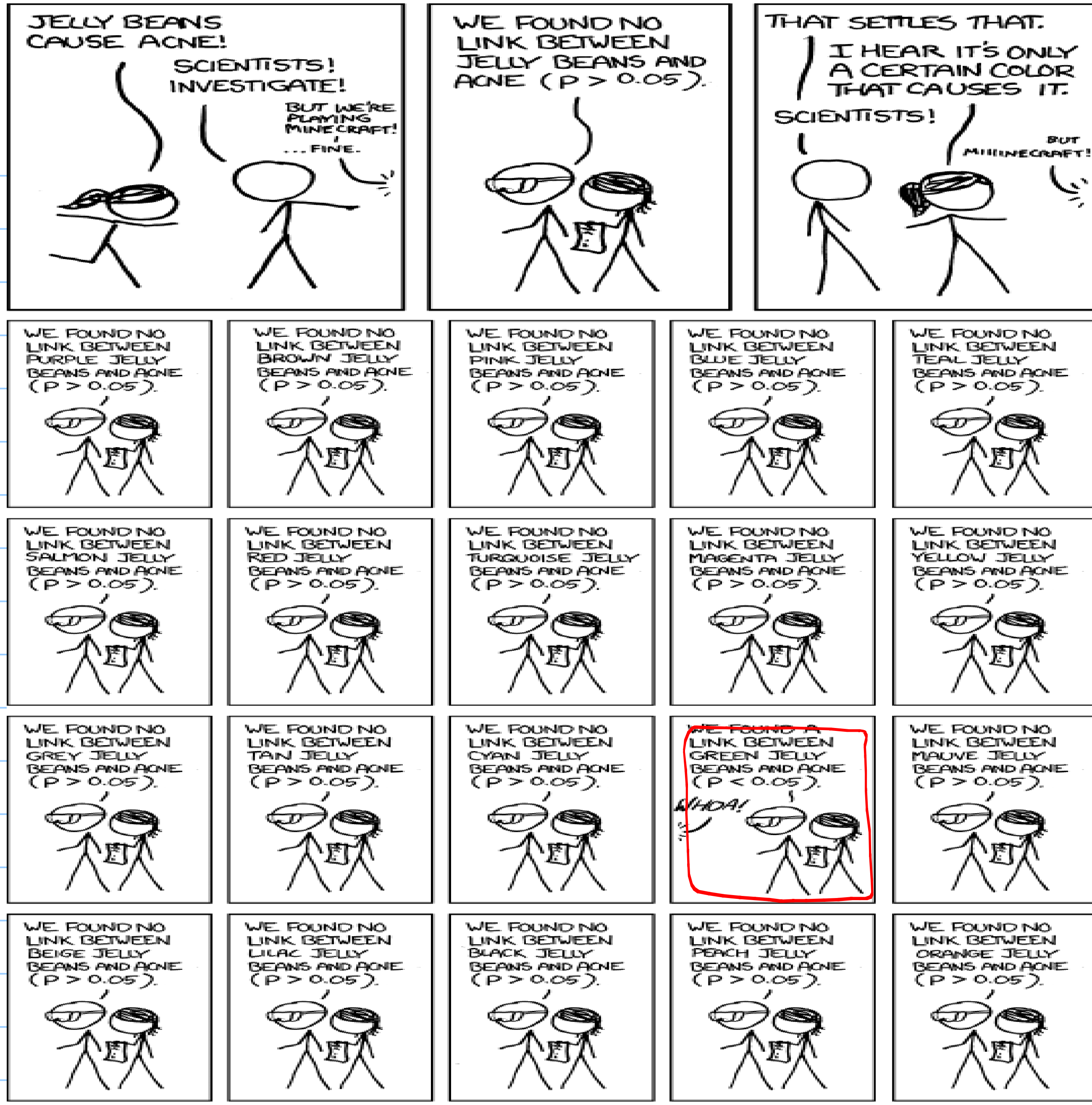
$$= pr(z < -0.5) = 0.3085$$

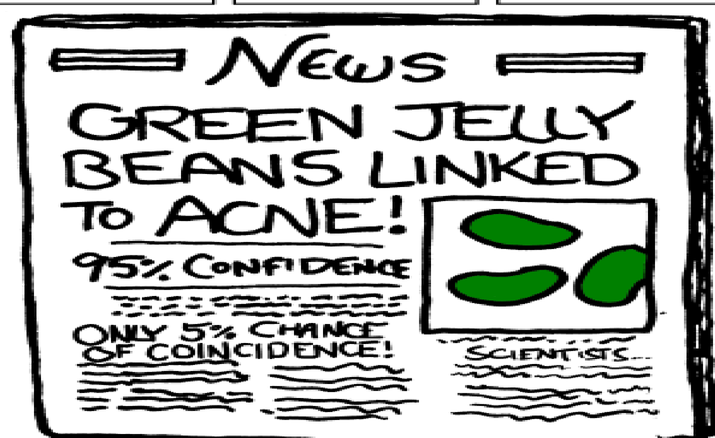
Since p-value < α, we cannot reject $\overset{\pi > 0.1}{H_0}$ in favor of $\overset{\pi < 0.1}{H_1}$

"In English": There is no evidence that the screens are OK.

Now, you need to decide! Give screen to NASA or not?

α is Dangerous!





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

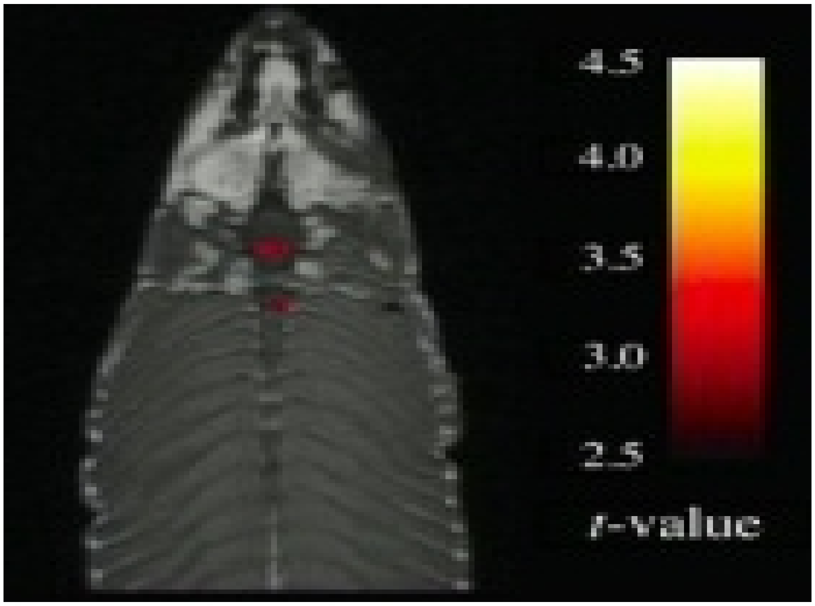
So:

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

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, $\mu \leq 0$. Ie. α% of the time, you will make a type I error.

Another example:

Dead thinking Salmon!



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

FYI

Alternatively, in some situations, one can simply report the p-value, without comparing it to α. After all, the above are all problems with α, not the p-value!

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

Something New!

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

I'll explain This later.

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 can be 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)

$\underbrace{\qquad\qquad\qquad}_{\text{Data.}}$

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

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

H_0: There is no association, i.e.

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 $H_0 = T$.

$\boxed{\text{Theorem: Under the null hypothesis, } X^2 \text{ has a chi-squared distr. with } df = k-1 \; (= 3-1 = 2)}$

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

see a few pages down

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

$\uparrow df = 3-1 = 2$

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

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

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

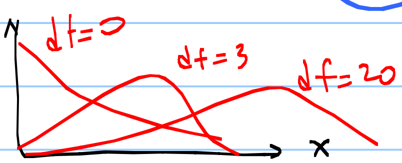
In English: There is no evidence from data to suggest that the 3 props are NOT .22, .32, .46, ie.

Ie. 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 next lecture!

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 $\underline{\underline{k}}$ 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

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

Then in a sample of size n, how many would

we $\boxed{\text{expect}}$ 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 $\boxed{\text{observe}}$ this many :	n_1	14
	n_2	28
	n_3	44

$\boxed{\text{Punch line}}$:

then the theorem tells us that

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

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

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

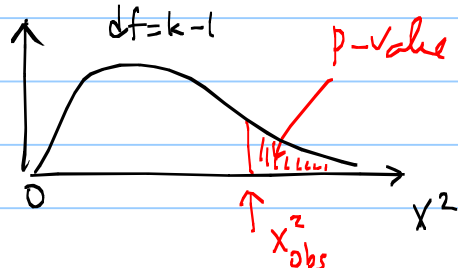
to more than 2 categories in the population. See how

How to use Table VII:

Table VII gives the area to the right of some value of χ^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.

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

$\chi^2_{obs} = 8.66$, 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 $\chi^2_{obs} = 8.55$ 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.

print out This page, point out The fallacy, and explain it.

hw_lect21_1

A student (Thuan) asked a good question, which I have structured into a hw problem. I hope it will help in better understanding p-values, and The logic of hypothesis testing.

Suppose we are testing $\begin{cases} H_0: \mu \leq 1 \\ H_1: \mu > 1 \end{cases}$

Here are 2 arguments, which have opposite conclusions, and so, one of Them must be wrong.

a) Which one is wrong?

b) point out where The fallacy is, and explain why you Think it's a fallacy.

1) If $H_0 = T$

Then any large $\bar{x}$ is evidence against H_0.

Then pvalue $= pr(\bar{x} > \bar{x}_{obs} \mid \mu \leq 1)$ measures evidence against H_0.

Then, if That prob. is small (e.g. $< \alpha$),

our assumption must have been wrong, and so we must reject H_0 in favor of H_1.

2) If $H_0 = T$

Then any large $\bar{x}$ is evidence against H_0.

Then pvalue $= pr(\bar{x} > \bar{x}_{obs} \mid \mu \leq 1)$ measures evidence against H_0, ie. evidence in favor of H_1.

Then, if That prob. is small (e.g. $< \alpha$),

There is small evidence in favor of H_1, and so we must not reject H_0 in favor of H_1.

Print out This page, and fill in the boxes

hw-lect 21-2 This is 8.49, but fill-in The boxes $\square$, below.

In This problem, there are 9 categories (in 1 pop.). The null hypothesis is $H_0: \pi_1 = \frac{1}{9}, \pi_2 = \frac{1}{9}, \ldots \pi_9 = \frac{1}{9}$.

a) If $H_0 = T$,

The expected counts in each of The 9 categories is

(G, G) (B, B) (S, S) (G, S) (S, G) (S, B) (B, S) (G, B) (B, G)

b) The 9 categories are combined into 3 new categories:

categ. 1 : (G, G), (B, B), (S, S)

categ. 2 : (G, S) (S, G) (S, B) (B, S)

categ. 3 : (G, B) (B, G)

The expected counts in each of These 3 categories are:

categ. 1 categ. 2 categ. 3

c) The observed counts in each of The 3 categories:

categ. 1 categ. 2 categ. 3

d) Compute X^2_{obs}.

e) Compute The p-value.

f) What's The conclusion "in English"?

hw-lect21-3

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-lect21-4

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 $\alpha = .05$.