We have developed a hypothesis testing procedure that gives a p-value (i.e., prob. of observing more extreme data than that observed). The p-value is useful by itself, and more people these days report it as a measure of evidence from data. But traditionally (i.e., old style!) one compares it with \( \alpha \) to make a reject/no-reject decision. So, the value of \( \alpha \) is important.

We know that it's the largest prob. at which we feel "confident" to reject \( H_0 \) in favor of \( H_1 \). But there is more:

Suppose we are testing \( H_0: \mu = \mu_0 \) vs. \( H_1: \mu > \mu_0 \).

We assume \( H_0 \) True (i.e., \( \mu = \mu_0 \)), 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

\[
\text{p-value} = \text{prob}(\bar{x} > \bar{x}_{\text{obs}})
\]

\( \alpha \)

How frequently is \( \bar{x} \) in the red? \( \alpha \)

\[ \text{Note: } \alpha = \text{prob}(\text{p-value} < \alpha) \]

So, \( \alpha = \text{prob}(\text{Data Reject } H_0 \text{ in favor of } H_1 \mid H_0 = \text{T}) \) "Bad" error

Type I error

"False Alarm Rate"

(Convicting an innocent person)

This is how you decide on \( \alpha \).

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:

$$\beta = \text{prob} \left( \text{Data cannot reject } H_0 \mid H_0 = \text{False} \right)$$

in favor of $H_1$

Type II

(Releasing a guilty person.)

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

$\alpha, \beta$ (the probs of the 2 types of errors) have a complex but mostly inverse relationship, depending on $n$ (p. 389-391).

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

Obviously, this will lead to some nonzero $\beta$, and it is important to compute it for your own specific problem.

Sometimes, people look at $1 - \beta$ (instead of $\beta$).

$$1 - \beta = \text{power} = \text{prob}(\text{Rejecting } H_0 \mid H_0 = \text{False})$$

(Convincing a guilty person).

If there is time, we'll return to power.
Suppose you are testing whether a drug has \( \mu > 0 \). 

So: 
\[ H_0: \mu \leq 0, \quad H_1: \mu > 0 \]

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 < \( \alpha \), i.e., there is evidence that \( \mu > 0 \). This will happen (at most) \( \alpha \% \) of the time even if, in fact, \( \mu < 0 \). I.e. \( \alpha \% \) of the time, you will make a type 1 error.
Another example of how fixing $\alpha$ is dangerous.

Dead Thinking Salmon!

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 $\alpha$.

In this class, we will continue to compare it with $\alpha$, but be aware of this "defect"
This lecture was a bit different in that you did not learn any new formulas! But it's a very important lecture because it reveals what α is, and the dangers of using the hyp. test methodology carelessly.

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?

Recall 1-sample test for μ: \( H_0: \mu = \mu_0 \), \( H_1: \mu \neq \mu_0 \), distr. = \( z, t \).

Statistic: \( t_{\text{obs}} = \frac{X_{\text{obs}} - \mu_0}{\sigma/\sqrt{n}} \), p-value = ...

Compare with the C.I. for μ: \( \bar{x} \pm z^* \frac{SE}{\sqrt{n}} \), \( \bar{x} \pm t^* \frac{S}{\sqrt{n}} \)

Recall 1-sample C.I. for \( p \): \( p \pm z^* \sqrt{\frac{p(1-p)}{n}} \)

⇒ 1-sample z-test for \( p \): \( H_0: \pi = \pi_0 \), \( H_1: \pi \neq \pi_0 \)

Statistic: \( Z_{\text{obs}} = \frac{p_{\text{obs}} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \), p-value = ...

\( \not\Rightarrow \) p! Because we assume \( H_0 = \text{True} \), i.e. \( \pi = \pi_0 \)!
By now, we know how to test the difference between 2 means:

\[ H_0: \mu_1 - \mu_2 = \Delta \quad , \quad H_1: \mu_1 - \mu_2 \neq \Delta \quad \text{dist}: Z, t. \]

2 props:

\[ H_0: \pi_1 = \pi_2 = \cdots = \pi_\Delta \quad , \quad H_1: \pi_1 - \pi_2 \neq 0 \quad \text{dist}: Z \]

When we compare 2 props from 2 populations \((\pi_1, \pi_2)\), one implication is that each population can be described with one proportion. That means that the population consists of 2 groups (e.g., boy/girl), because when we estimate the proportion of one group in one of the pops (say \(\pi_1\), boys in Northern hemisphere), then the proportion of the other group (say \(\pi_2\), girls in Northern hemisphere) is estimated automatically, because \(\pi_2 = 1 - \pi_1\). What the tests do for us is to compare \(\pi_1\) boys in Northern with \(\pi_1\) boys in Southern hemisphere.

What if we have 1 population, but multiple (k) groups?

Note: The 2-group case can be thought of as the population consisting of 1 binary variable. Similarly, the k-group case can be thought of as a categorical variable with k levels.

The \(H_0 / H_1\) we can test is this:

\[ H_0: \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \cdots, \pi_k = \pi_{0k} \quad \text{dist: Chi-sq} \]

\[ H_1: \text{At least one of } \pi_i \text{ is wrong} \]

I'll explain this later.
Does data provide sufficient evidence to support an association between climate and tornado activity?

<table>
<thead>
<tr>
<th></th>
<th>El Nino</th>
<th>La Nina</th>
<th>Normal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of years</td>
<td>12</td>
<td>17</td>
<td>25</td>
<td>54</td>
</tr>
<tr>
<td>Proportion</td>
<td>( \frac{12}{54} = 0.22 )</td>
<td>( \frac{17}{54} = 0.32 )</td>
<td>( \frac{25}{54} = 0.46 )</td>
<td>1</td>
</tr>
</tbody>
</table>

# of days with violent tornadoes: 
- \( n_1 = 14 \) in El Nino
- \( n_2 = 28 \) in La Nina
- \( n_3 = 44 \) in Normal

Proportion: \( \frac{14}{86} = 0.16 \) in El Nino, \( \frac{28}{86} = 0.33 \) in La Nina, \( \frac{44}{86} = 0.51 \) in Normal

**Null Hypothesis (H₀):** There is no association, i.e., \( \pi_1 = \pi_2 = \pi_3 = 0.46 \)

**Alternative Hypothesis (H₁):** At least one of these assignments is wrong.

If H₀ is True, how many tornadoes do you expect in each of the \( k^3 \) categories?

**Expected Count:**
- \( 0.22 \times 86 = 18.9 \)
- \( 0.32 \times 86 = 27.5 \)
- \( 0.46 \times 86 = 39.6 \)

**Observed Count:**
- 14
- 28
- 44
\[
\frac{(\text{Exp.} - \text{obs})^2}{\text{Exp}} = \begin{array}{c}
(4.9)^2 & (-0.5)^2 & (-4.4)^2 \\
1.27 & 0.0009 & 0.49 \\
\end{array}
\]

\[
\sum_{i=1}^{3} \frac{(\text{exp.} - \text{obs})^2}{\text{exp.}} = 1.77
\]

If there were really no difference at all in the # of tomatoes between the 3 categories, then this would be near 0.

Q: So, is it far away from 0 to call it not 0?

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

\[X^2\text{ has a chi-squared dist. with } df = k - 1 \quad (= 3 - 1 = 2)\]

The area under the chi-sq. dist. is in Table VII.

p-value = prob\(X^2 >= x^2_{\text{obs}}\) = prob\(X^2 >= 1.77\) > 0.1 \(\Rightarrow\) df = 3-1 = 2

Conclusion (at \(\alpha = 0.01\)): p-value > \(\alpha\)

In words: cannot reject \(H_0\) in favor of \(H_1\).

\((\hat{p}_1 = 0.22, \hat{p}_2 = 0.32, \hat{p}_3 = 0.46)\)

In English: The evidence from this data does not suggest that there is an association between climate and tomato activity.

For the chi-sq test, this sign is always > 1. I'll explain that, later!
Summary / Generalization

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

Let \( \pi_i = \text{proportion of cases in category } i \):

- \( \pi_1 = \text{proportion of category 1} \)
- \( \pi_2 = \text{proportion of category 2} \)
- \( \pi_3 = \text{proportion of category 3} \)

Null params:

- \( \pi_{01} = 0.22 \)
- \( \pi_{02} = 0.32 \)
- \( \pi_{03} = 0.46 \)

If \( H_0 = \text{True} \), \( H_0: \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \ldots \)

Then in a sample of size \( n \), how many would we expect in category 1:

\[ n \pi_{01} \]

- Category 1:
- Category 2:
- Category 3:

\[ \sum n_i = n \]

But according to data, we observe this many:

\[ \sum \begin{cases} n_1 \rightarrow 14 \\ n_2 \rightarrow 28 \\ n_3 \rightarrow 44 \end{cases} \]

Punch line:

Thus the theorem tells us that:

\[ X_{\text{obs}}^2 = \sum \frac{(\text{exp} - \text{obs})^2}{\text{exp}} = \sum \frac{(n \pi_{0i} - n_i)^2}{n \pi_{0i}} \]

counts, not proportions!

has a chi-squared distribution with \( df = k-1 \).
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)? Specifically,

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?

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 students do not like Lab
- 15 students 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 = 0.05.