In the procedure we have learned, the last step involves comparing the p-value with $\alpha$. That practice is (slowly) becoming "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, $\alpha$ 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 $H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$.

We assume $H_0$ true (i.e. $\mu = \mu_0$), then compute a p-value.

If p-value $\leq \alpha$, then reject $H_0$ in favor of $H_1$.

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

How often will that happen?

For $H_0, H_1$ given here

$p$-value = $\text{prob}(X > x_{\text{obs}})$

Assuming $H_0$ is true ($\mu = \mu_0$)

How frequently is $X$ in the red? $\alpha$

$H_1: \mu > \mu_0$

$\alpha = \text{prob}(\text{p-value} \leq \alpha \mid H_0 = \text{true})$

So, $\alpha = \text{prob(Reject } H_0 \text{ in favor of } H_1 \mid H_0 = \text{true} )$

"Bad" error

"False Alarm Rate" (convicting an innocent person)

This is how you decide the value of $\alpha$. You ask:

"How much bad error can I tolerate in the long run?"
The other error is called Type II, and it’s not as bad!

\[ \beta = \text{prob (Data cannot reject } H_0 \mid H_0 = \text{False)} \]
\[ \text{in favor of } H_1 \]
\[ H_1 = \text{True} \]
(Releasing a guilty person.)

And \( \text{power} = 1 - \beta = \text{prob (Data reject } H_0 \mid H_0 = \text{False)} \]
\[ \text{in favor of } H_1 \]

\[ \alpha, \beta \text{ (The probs of the 2 types of errors) have a complex but mostly inverse relationship, depending on } n \]
(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 \( \alpha \), suggests another way to set-up \( H_0 / H_1 \):

**Example: Guilt or Innocence?**

\[ \text{Bad error} = \text{Data say guilty} \mid \text{innocent} \]
\[ \text{Type I} = \text{Data say } H_1 \mid H_0 = \text{T} \]
\[ \text{Power} = \text{prob (Data say guilty} \mid \text{guilty}) \]

You can see why we usually set the value of \( \alpha \) to very small.
Another example: NASA

A company manufactures computer screens for use by astronauts on space missions. If more than 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?

\[ \pi = \text{true/pop. prop. of defective pixels} \]

\[ H_0: \pi = 0.1 \quad H_a: \pi > 0.1 \]

\[ H_1: \pi > 0.1 \quad H_0: \pi < 0.1 \]

**OK error**

(Data say \( \pi > 0.1 \) | \( \pi < 0.1 \))

**Bad error**

(Data say \( \pi < 0.1 \) | \( \pi > 0.1 \))

\[ \alpha = \Pr(\text{Data say } H_1 \mid H_0 = \text{true}) \]

\[ p-value = \Pr(\frac{Z}{\pi_{obs}} > 0.1 \mid H_0 = \text{true}) = \Pr\left( \frac{Z}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < \frac{\pi_{obs} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \mid \pi = 0.1 \right) \]

\[ \pi_{obs} = \frac{1}{16} \]

\[ = \Pr\left( \frac{Z}{0.0625 - 0.1}{\sqrt{1.0016}} \right) = \Pr\left( Z < \frac{-0.0375}{0.34} \right) \]

\[ = \Pr( Z < -0.5 ) = 0.3085 \]

Since the p-value is less than \( \alpha \), we cannot reject \( H_0 \) in favor of \( H_1 \).

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

Now, you need to decide! Give screen to NASA or not?
Suppose you are testing whether a drug has \( \mu > 0 \).

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

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 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 \( \alpha \). After all, the above are all problems with \( \alpha \), not the p-value!

In this class, we will continue to compare it with \( \alpha \), but be aware of this "defect."
Here is another example that shows up frequently in a set of problems called “classification problem” (as opposed to “regression problems”). Machine-learning methods like dealing with classification problems.

Suppose you have a AI machine that classifies/predicts all incoming emails as “Safe” or “Unsafe”. The following Table (called a contingency Table, or confusion matrix) is often used to summarize things.

For example, suppose we take 200 safe emails, and 100 unsafe emails, and throw them into our classifier, and get the following counts:

<table>
<thead>
<tr>
<th></th>
<th>Classified as</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unsafe</td>
<td>safe</td>
</tr>
<tr>
<td>Truly</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

These are the BAD (Type I) errors.

These errors are not as bad (Type II)
Something New!

When we compare 2 props \((\pi_1, \pi_2)\), e.g. \(H_0: \pi_1 - \pi_2 < 0.1\)
the implication is that we have two populations, each with
2 groups/cats. (e.g. Boys, Girls).
In that case, ONE proportion (e.g. prop of boys, \(\pi_{\text{Boy}}\)) is enough
to describe each pop., because the other prop. (e.g. \(\pi_{\text{Girl}}\)) is
fixed by \(1 - \pi_{\text{Boy}}\). The 2-sample \(z\)-test we have developed
involves TWO proportions, one from each of TWO populations.
So an example would be \(\pi_1 = \pi_{\text{Boys}}\) in Northern hemisphere.
\(\pi_2 = \pi_{\text{Boys}}\) in Southern hemisphere.

Note that both \(\pi_1\) and \(\pi_2\) 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
\(H_0: \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \ldots, \pi_k = \pi_{0k}\) \(\text{prop. of } k^{th}\text{ cate. in pop.}\)

\(H_1: \text{At least one of } \pi_i \text{ is wrong} \quad \sum_{i=1}^{k} \pi_{0i} = 1\)
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
described by one random variable with 2 levels. Similarly, a pop.
with \(k\) groups can be described with one r.v. with \(k\) levels.
Does data provide sufficient evidence to support an association between climate and tornado activity?

<table>
<thead>
<tr>
<th>El Nino</th>
<th>La Nina</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 = 14 )</td>
<td>( n_2 = 28 )</td>
<td>( n_3 = 44 )</td>
</tr>
</tbody>
</table>

\[ \text{proportion: } \frac{14}{86} = 0.16 \quad 0.33 \quad 0.5 \]  
\[ \text{Data.} \]

\[ \text{proportion: } \frac{12}{54} = 0.22 \quad 0.32 \quad 0.46 \]  
\[ \text{(1)} \]

\( H_0: \) There is no association, i.e. the true prop. of tornado days in El Nino years, etc.

\( H_0: \) \( P_1 = 0.22 \quad P_2 = 0.32 \quad P_3 = 0.46 \)

\( H_1: \) At least one of these assignments is wrong.

If \( H_0 = \) True, how many tornadoes do you expect in each of the \( k \) categories?

**Expected counts:**
\[ 0.22(86) \quad 0.32(86) \quad 0.46(86) \]

**Observed counts:**
\[ 14 \quad 28 \quad 44 \]

\[ \frac{(\text{Exp.-obs})^2}{\text{Exp.}}: \]
\[ (4.9)^2 \quad (-0.5)^2 \quad (-4.4)^2 \]

\[ \chi^2 = \sum_{i=1}^{3} \frac{(\text{exp.-obs})^2}{\text{exp.}} = 1.77 \]
If there were really no difference at all in the # of tornadoes between the 3 categories, then this would be near zero.

So, is this $X_{\text{obs}}^2$ far away from 0 to reject $H_0$ (in favor of $H_1$)? Note: $X^2$ is non-negative, unlike $z$, $t$

We need to know the sample 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 distr.? It's just another Table (VII). But FYI,

\[ p\text{-value} = \text{prob}(X^2 > X_{\text{obs}}^2) = \text{prob}(X^2 > 1.77) > 0.1 \]
\[ \uparrow \]
\[ \text{df} = 3 - 1 = 2 \]

For the chi-squared test, p-value is always right area. (See below).

Conclusion (at $\alpha = 0.01$): $p\text{-value} > \alpha$

In words: Cannot reject $H_0$ in favor of $H_1$.

In English: There is no evidence from data to suggest that the 3 props are not 0.22, 0.32, 0.46, i.e.

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

The chi-squared density function is (FYI)

\[ f(x) = \frac{d^f - 1}{2} e^{-\frac{x}{2}} \frac{d^f}{\Gamma(d_f)} \frac{d^f}{\Gamma(d_f)} \]

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

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

\[
\begin{align*}
\pi_1 &= \text{proportion of cat. 1} \quad \text{Null params} \quad \overline{\text{Example}} \\
\pi_2 &= \quad \text{2's} \quad \pi_{01} \quad 0.22 \\
\pi_3 &= \quad \text{3's} \quad \pi_{02} \quad 0.32 \\
\pi_4 &= \quad \text{4's} \quad \pi_{03} \quad 0.46 \\
\end{align*}
\]

**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:

\[ n \pi_{01} = 18.9 \]

\[ n \pi_{02} = 27.5 \]

\[ n \pi_{03} = 39.6 \]

... let \( \sum n_i = n \)

But according to data, we observe this many:

\[ \begin{align*}
\sum n_1 &= 14 \\
\sum n_2 &= 28 \\
n_3 &= 44 \\
\end{align*} \]

**Punch line:**

Thus the theorem tells us that

\[ X^2_{\text{obs}} = \sum_i \frac{(\text{exp. - obs})^2}{\text{exp.}} = \sum_i \frac{(n \pi_{0i} - n_i)^2}{n \pi_{0i}} \]

... has a chi-sq. distr with df = \( k-1 \).

\[ p\text{-value} = \text{prob}(X^2 > X^2_{\text{obs}}) \]
Interpretation/Diagnosis

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

**Example**: In the tornado example

Suppose we had found $p_{value} < 0.05$, i.e., there is evidence that climate does affect tornadoic activity. Then $x_{obs}^2$ would be big. But what makes it big? The 3 terms contributing to $x_{obs}^2$ are:

<table>
<thead>
<tr>
<th>El Nino</th>
<th>La Nina</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.27 (large)</td>
<td>0.009 (small)</td>
<td>0.49</td>
</tr>
</tbody>
</table>

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

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

How about the "direction" of the association? E.g., 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:

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

**Note**: order!

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

So in El Nino years: $\text{exp} > \text{obs} \Rightarrow$ less tornadoic than expected.

in La Nina years: $\text{exp} < \text{obs} \Rightarrow$ more tornadoic.
Note that the above $H_0, H_1$ is just a generalization of

$H_0 : \pi = \pi_0$  \hspace{1cm} (z-test).

$H_1 : \pi \neq \pi_0$

to more than 2 categories in the population.

However, there are no i-sided / 2-sided varieties of chi-sq'd.

When $X^2_{obs}$ is small (say $\approx 0$), then the observed counts
are consistent with the expected counts if $H_0$ is true
(i.e. $\pi_1 = \pi_0, \pi_2 = \pi_0, \ldots, \pi_k = \pi_0$). So, if $X^2_{obs}$ is large,
then at least one of these specifications must be wrong.

In other words, the appropriate hypotheses are

$H_0 : \pi_1 = \pi_0, \pi_2 = \pi_0, \ldots, \pi_k = \pi_0.$

$H_1 : \text{At least one of these specifications is wrong.}$

And it is the “At least” which gives us

$p$-value = $\text{prob}(X^2 \geq X^2_{obs})$ \hspace{1cm} (Table VII).

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

Said differently for the chi-sq'd 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

In Ch. 7, we learned how to build CIs for either 1 prop, \( \hat{p} \), or the difference between 2 props, \( \hat{p}_1 - \hat{p}_2 \), where \( \hat{p}_1 \) = prop of something (e.g., boys) in population 1, and \( \hat{p}_2 \) = "same thing" in 2.

We also learned how to do hyp. tests on \( \hat{p}_1 \), or \( \hat{p}_1 - \hat{p}_2 \). [Note \( \hat{p}_1 + \hat{p}_2 \neq 1 \), because \( \hat{p}_1, \hat{p}_2 \) are from 2 different populations]

But in all of these situations, the 2 pops have 2 categories (boy/girl) and \( \hat{p}_i \) is the prop. of 1 of them.

The tornado/climate e.g. deals with the situation where one population has 3 categories. For \( k \) categories:

We learned that the relevant dist. is chi-squared with \( df = k-1 \). And the quantity that follows that dist. is like \( Z \),

\[
\chi^2 = \sum_{i=1}^{k} \left( \frac{\text{obs}_i - \text{exp}_i}{\text{exp}_i} \right)^2
\]

where \( \text{obs}_i \) and \( \text{exp}_i \) are observed and expected counts in the \( i \)th category (still of 1 population). The latter are computed assuming \( H_0 \) is true, where

\[
H_0: \hat{p}_1 = \hat{p}_{o1}, \hat{p}_2 = \hat{p}_{o2}, \ldots, \hat{p}_k = \hat{p}_{ok}
\]

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

This time \( \hat{p}_1 + \hat{p}_2 + \cdots = 1 \), \( \hat{p}_{o1} + \hat{p}_{o2} + \cdots = 1 \)
How to use Table VII:

Table VII gives the area to the right of some value of $x^2_{\text{obs}}$, i.e., 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.

- $x^2_{\text{obs}} = 8.49$, df = 4 $\Rightarrow$ p-value = 0.075
- $x^2_{\text{obs}} = 8.66$, 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 $x^2_{\text{obs}} = 8.55$ with df = 4. Then we can say $0.070 < \text{p-value} < 0.075$. That is good enough if $\alpha = 0.05$, because p-value > $\alpha$, and so we cannot reject $H_0$ in favor of $H_1$. 

This is 8.49, but fill-in the boxes 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

(a, G) (b, B) (s, S) (c, S) (c, G) (s, B) (b, S) (g, B) (g, G)

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

Category 1: (g, G), (b, B), (s, S)
Category 2: (c, S) (c, G) (s, B) (b, S)
Category 3: (g, B) (g, G)

The expected counts in each of these 3 categories are:

Category 1

Category 2

Category 3

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

Category 1

Category 2

Category 3

d) Compute $X^2_{obs}$ (show work here, put answer here)

$\square$

e) Compute the p-value.

$\square$
f) What's the conclusion "in English"?
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 " 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 \( \alpha = 0.05 \).