Throughout this semester we learned some key concepts and tools to help us interpret what data tell us about scientific hypothesis.

We have learned how to use these tools, when to use them and what their underlying purpose is.

We have learned that statistics is more that simply getting a *p-value* - there are often many different ways to do that - e.g. exact or approximate, assuming equal variances or not etc.

Statistics is an ‘art’ because we must use our own judgment in determining which method to use. This is the key to a good statistical analysis (anybody can shove data into a computer package and get a p-value).

We have learned that a good criteria by which to choose a particular method is by how often that methods does what it is supposed to do (e.g., how often does the test reject when it is supposed to). This is called a frequency criterion.

We have learned that some criteria are linked - e.g., we can not increase power and decrease the type I error at the same time if we hold everything else constant.

More fundamentally we have learned:

How to display, summarize, and tabulate data.

How to calculate probabilities of observing certain events and/or data.

That confounding is nothing more than comparing two weighted averages when their weights are different (making said comparison unfair).

That any statistic (a rate, mean, proportion, odds ratio, or a relative risk) can be confounded.

Statistical methods like hypothesis testing and confidence intervals.

We have learned the basic probabilities models on which these tests are based.

We have learned how to perform exact versions of statistical methods using those probability models.

We have learned how to perform approximate versions of those methods using standardization techniques that work in large samples.

We have learned when it is appropriate to use those approximate versions.

We have learned how to project the sample size for a study that will use these (approximate) methods.

We have learned that the p-value is noting more than the probability of observing your data or data more extreme assuming the Null hypothesis is true.

(Remember, we often represent the data with a summary statistic, such as a sample mean or proportion, which estimates the unknown population parameter specified in the null hypothesis.)

$$\begin{aligned}
\text{p - value } &= \text{P(data or data more extreme } | \, \text{H}_0) \\
&= P\left(\overline{X} > \overline{X}_{obs} \mid \text{H}_0\right) \\
&= P\left(\hat{\theta} > \hat{\theta}_{obs} \mid \text{H}_0\right) \\
&= P\left(\overline{X} - \overline{Y} > \overline{X}_{obs} - \overline{Y}_{obs} \mid \text{H}_0\right) \\
& etc.
\end{aligned}$$

The observed values of these sample estimates come from the data. So we are asking what is the probability of observing a sample proportion greater than what I saw in this data set given the null hypothesis (which specifies the population proportion).

We learned how to calculate this in two ways: by computing it exactly or by standardizing to get an approximation from the z-table.

We have learned about significance testing and hypothesis testing, but we did not talk much about how to distinguish between them. To make matters complicated, these two testing procedures use the exactly the same mathematical framework.

The basic idea of hypothesis testing: Determine if you should reject or accept the null hypothesis based on the data at hand. Report your decision to reject or accept and your Type I error, but do not report the p-value. This is a decision oriented methodology.

The basic idea of significance testing: Calculate the probability of observing your data or data more extreme under the null hypothesis (i.e., the p-value). The smaller the p-value, the more evidence you have that the null hypothesis is not supported by the data at hand (this is the ideal, unfortunately it is not quite true…but that is another story).

Note that it is common to confuse these two approaches by first calculating the p-value, checking to see if is less than the pre-specified type I error and then rejecting the hypothesis and quoting the p-value.

Hypothesis testing

Neyman and Pearson (1933)

“The problem of testing statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B…”

“Thus to accept a hypothesis means only to decide to take action A rather than action B. This does not mean that we necessarily believe that the hypothesis H is true.” (Neyman 1950)

Concept of controlling type I and II errors is attractive; leads to ‘inductive behavior’

No concept of statistical evidence

Significance testing

‘P-values measure the strength of evidence against a hypothesis’ (Fisher 1958)

But the interpretation of p-values depends on sample size:

1. Berkson (1942), Cornfield (1966): If two p-values are equal the one from the larger sample size confer more evidence (because those data are more precise).

2. Lindley & Scott (1984), Peto et al. (1976): If two p-values are equal the one from the smaller study confers more evidence (because you have observed a larger departure from the null).

3. Royall 1986, Morrison & Henkel (1970): Both the above are correct. It depends on if you report the p-value or if you just report your decision to reject or accept.

Confusion between p-value and type I errors causes multiple comparisons and multiple looks problems.

Over time p-values have been given a post-hoc type I error interpretation, leading to irresolvable controversies over multiple comparisons and multiple looks

Why? Because here p-values represent two distinct concepts:

(1) the measure of the strength of evidence

(2) the measure of the potential for that evidence to be misleading

And these are clearly different: You can have very strong evidence that is not likely to be misleading or you have very strong evidence that is not likely to be misleading. So these concepts can not be represented by the same mathematical quantity (tail area probability).

Finally, every statistical analysis reflects the following three elements:

1. Data
2. Probability model (assumptions)
3. Context (generalizability)

Statistics do not lie; they merely reflect the three elements listed above.

Statisticians (and students who have taken statistics) are concerned not only with the data they observed, but with what assumptions they used to make those data "speak" to the hypothesis of interest.