Neyman-Pearson Tradeoffs

BIOS 6611

CU Anschutz

Week 4

1. **Hypothesis Testing and Fisher's Review**

2. **The Neyman-Pearson Approach to Data Testing**

Hypothesis Testing and Fisher's Review

Hypothesis Tests

Hypothesis testing is a method of making inferences about a population quantity from a data sample. We begin with a statement or "hypothesis" about the population and use data to determine if the hypothesis is supportable or not.

A **hypothesis** is a claim or statement about a population parameter (or parameters). A **hypothesis test** is a statistical method of quantifying evidence (using sample information) to reach a decision about a hypothesis.

e.g. Recommended daily allowance of zinc for males over 50 is 15 mg/day. A study found a sample of 115 men aged 65-74 had an average intake of 11.3 mg/day and the s.d. of intake was 6.4 mg/day. Does the study indicate too little zinc for these men?

Fisher's Approach Review

Fisher's approach to testing data focused on:

- permutation test
- calculation of a "p-value"
- only defining a null hypothesis
- comparisons done *a posteriori*

The Neyman-Pearson Approach to Data Testing

Jerzy Neyman and Egon Pearson

- Neyman was a Polish statistician who eventually taught at UC-Berkeley
- Pearson was a British statistician who taught at University College London
- Neyman and Pearson felt Fisher's approach was lacking (and they were low-key rivals)
- They developed an approach that was more mathematical and focused on *a priori* considerations

Null Hypotheses

We first define a **null hypothesis** (H_0).

The null hypothesis *is a claim that is initially assumed to be true*, and usually has a form similar to:

- "There is *no change* between..."
- "...no difference...", "...no effect of...", "...no association..."

H_0 is where we place the burden of proof for data–what we could actually like to *disprove*.

Alternative Hypotheses

In the Neyman-Pearson approach, we then state an opposing or **alternative hypothesis** (H_1 or H_A).

H_1 contradicts H_0 so that both cannot be true, and is the statement we would like to prove to be true.

The form of H_1 is similar to H_0, but we would generally indicate $>$, $<$, or $\neq$ in place of $=$.

Type I and Type II Errors

We collect data assuming H_0 is true, we then test that assumption and make a decision about the truth of H_0.

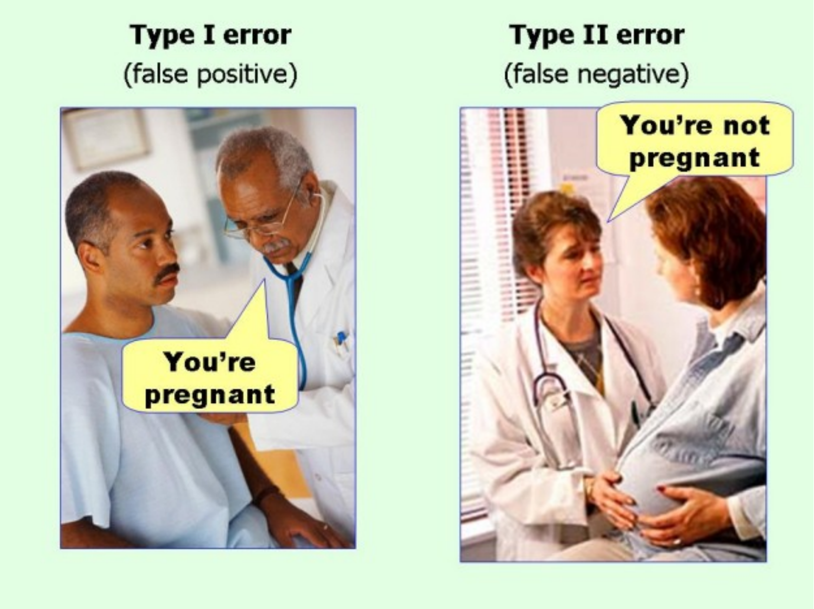
Based on our data, we have 4 possible outcomes:

1. H_0 is true and we fail to reject H_0 (i.e., we say it is "true")
2. H_0 is true and we reject H_0 (i.e., we say it is "false")
3. H_0 is false and we fail to reject H_0
4. H_0 is false and we reject H_0

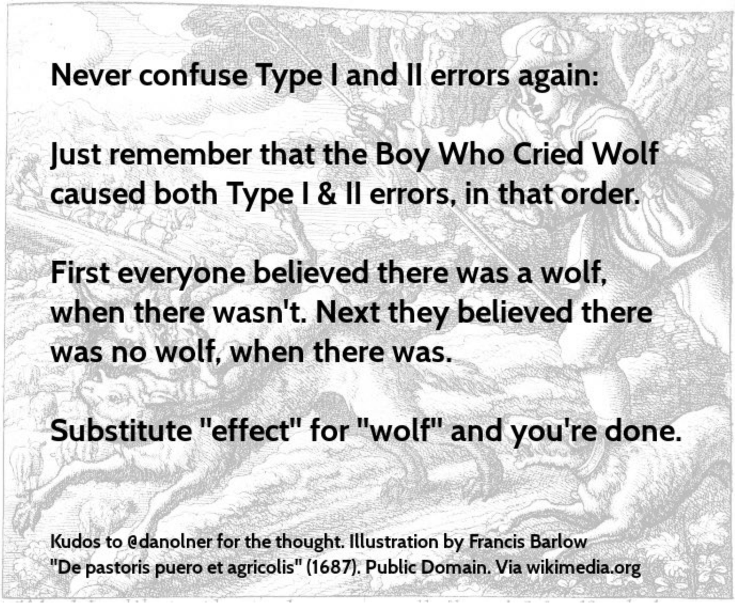
Two of these scenarios are incorrect conclusions and represent

- **Type I Error:** probability of rejecting H_0 when it is true (outcome $\#2$) *and is usually considered the more serious error*
- **Type II Error:** probability of failing to reject H_0 when it is false (outcome $\#3$)

Type I and Type II Errors



Type I and Type II Errors



Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow
"De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org

α, β, and Power

Based on the data we collect to address H_0, we make a decision to reject or not reject H_0. *Note that we don't "accept H_0" or say "H_0 is true"*, all we can say is that we have evidence to reject it or we don't: we "reject H_0" or we "fail to reject H_0".

	Reality	
What we decide	**$\mathbf{H_0}$ True**	**$\mathbf{H_0}$ False/$\mathbf{H_1}$ True**
Fail to reject $\mathbf{H_0}$	*Correct* Probability of correct decision $=$ $1-\alpha =$ level of confidence	Type II Error P(Type II Error) $= \beta$
Reject $\mathbf{H_0}$	Type I Error P(Type I Error) $= \alpha$ (**Level of significance**)	*Correct* Probability of correct decision $=$ $1-\beta =$ **Power**

An Important Assumption

Neyman and Pearson assumed that α and β were in terms of *the long run* (i.e., over infinite repeated samples).

Unfortunately, this is an unrealistic assumption since we cannot conduct infinite repeated samples, and in practice we often do not try to even reproduce a study once.

Fortunately, we can leverage these properties in our simulation studies, since we can simulate as many samples as we desire and summarize the number of correct and incorrect decisions.

Revisiting Our 4 Possible Outcomes

Based on our data, we have 4 possible outcomes:

1. H_0 is true and we fail to reject $H_0 \implies 1 - \alpha =$ level of confidence
2. H_0 is true and we reject $H_0 \implies \alpha =$ P(type I error)
3. H_0 is false and we fail to reject $H_0 \implies \beta =$ P(type II error)
4. H_0 is false and we reject $H_0 \implies 1 - \beta =$ power

Note, both #1 and #4 are "correct" decisions, but they represent different probabilities of making a correct decision.

How We Make Our Decision with Neyman-Pearson

Neyman-Pearson introduced the concept of **rejection regions** (also called **critical regions**) of a test. It represents the range of potentially observable values for which we would reject H_0.

This region is defined based on the desired α-level. For the mean it would be defined as

$$P(c_1 \leq \bar{X} \leq c_2 | H_0 \text{ is true}) = 1 - \alpha$$

Deriving the Rejection Region for the Mean

Assume $X \sim N(\mu, \sigma^2)$, so that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. If $H_0 = \mu_0$, we can solve for the rejection region by leveraging the standard normal distribution:

$$\begin{aligned}
1 - \alpha =& P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{1-\frac{\alpha}{2}} | H_0 \text{ is true}\right) \\
=& P\left(-Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu_0 < Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} | H_0 \text{ is true}\right) \\
=& P\left(\mu_0 - Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu_0 + Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} | H_0 \text{ is true}\right) \\
=& P\left(c_1 < \bar{X} < c_2 | H_0 \text{ is true}\right)
\end{aligned}$$

What we have essentially calculated is a **confidence interval** around μ_0.

Rejection Region Example

Assume we conduct a study measuring cholesterol with $n = 12$, $\bar{X} = 217$ mg/dL, $\sigma^2 = 46^2$ (mg/dL)^2, $\mu_0 = 211$ mg/dL, and $\alpha = 0.05$. We can note that $Z_{1-\frac{0.05}{2}} = Z_{0.975} = 1.96$ (you can check this with `qnorm(0.975)` in R).

$$
\begin{aligned}
1 - 0.05 =& P\left(\mu_0 - Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu_0 + Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} | H_0 \text{ is true}\right) \\
0.95 =& P\left(211 - 1.96 \times \frac{46}{\sqrt{12}} < \bar{X} < 211 + 1.96 \times \frac{46}{\sqrt{12}} | H_0 : \mu_0 = 211\right) \\
0.95 =& P(185 \text{ mg/dL } < \bar{X} < 237 \text{ mg/dL})
\end{aligned}
$$

In other words, we would *fail to reject* H_0 if our sample mean is between 185 and 237 mg/dL.

Therefore, in our sample, we fail to reject the null hypothesis that the sample came from a population with a mean cholesterol level of 211 mg/dL.

Neyman-Pearson Approach Summary

We can see that this approach has some similarities to Fisher's but is different in many ways:

- We explicitly define H_1
- There are no p-values
- Repeated sampling is assumed for properties like type I and II errors
- α and β should be defined *a priori* and a study designed based on these assumptions