Hypothesis Testing Continued

Collin Nolte

February 15, 2022

Review and Preview

What we did last time:

- Null hypothesis and p-values
- Standard normal distribution
- Reject and fail to reject

What we do today:

- Type I and Type II Errors
- Sensitivity and Specificity
- Power and effect size

Review

Null hypothesis associated with a null distribution, $H_0 : \mu = \mu_0$ and $X \sim N(\mu_0, \sigma^2/n)$

Specify critical values z_α to set width of confidence intervals, with coverage probability $1 - \alpha$

p-value describes relationship between observed data and hypothesis:

$$p = P(\text{observed data} \mid H_0)$$

We reject H_0 if observed data is unlikely to come from null distribution, i.e., $p < \alpha$

Drawing Conclusions

In actuality, a null hypothesis is either true or false, and based on the data, we may reject or fail to reject this null. As a consequence, there are two ways in which we might make a mistake.

	True State of Nature	
Test Result	H_0 True	H_0 False
Fail to reject H_0	Correct $(1 - \alpha)$	Incorrect Type II Error (β)
Reject H_0	Incorrect Type I Error (α)	Correct $(1 - \beta)$

- Type I error $=$ P(Reject $H_0 | H_0$ true) $=$ false alarm
- Type II error $=$ P(Fail to reject $H_0 | H_A$ true) $=$ missed opportunity

Controlling Errors

While all mistakes aren't great, some are worse than others, and the design of our study can influence which errors are more likely to occur.

The *Type I* error can be controlled by setting the level of significance, α. The smaller the value of α, the more evidence required to reject H_0. In other words, we can require the p-value to be such that $p < \alpha$

The *Type II* error is controlled by β. The quantity $1 - \beta$ is called the *power* of a study. More powerful studies have lower probabilities of Type II errors

Unfortunately, these values are often in conflict: if we always reject the null, we will never commit a Type II error. Similarly, if we never reject the null, the probability of a Type I error is zero. Obviously, neither is ideal

A working example

Let's suppose we are interested in determining the effect of drug A on some particular metric, say blood pressure

Drug A either has an effect or it does not, and to determine that, we initiate a study and plan on recruiting *n* participants

If drug A doesn't have an effect, I would like to know so that I don't invest any more resources into investigating it. Alternatively, if drug A does have an effect, I would like to know so that I can commit more resources to it

In actuality, I will not know the truth, but I can conduct a study, a random process that will give me some estimate of the effect of drug A. My null hypothesis is having no effect, with $H_0 : \mu = 0$

Type I error

If drug A doesn't work, I want to know

- "If H_0 is true, I want to be reasonably sure"
- "I can be more confident by collecting more evidence"
- "Evidence, in this case, would mean that my observed $\overline{X}$ is extreme, given the null distribution based on H_0"
- "I can set my threshold for how much evidence I would need by my choice of α, the Type I error rate"
- "Smaller values of α indicate that I need stronger evidence. This requires that I have a smaller p-value, with $p < \alpha$"

Since this is basically what we have done the last few weeks, here is a good time to have a list of questions

Type II error

If drug A does work, I want to know

- "If H_0 is false, I want to be sure to reject it"
- "This means I want to be more confident about my estimate of μ"
- "This is difficult to do if there is a lot of variability. I can reduce the amount of variability by increasing my sample size"
- "This can be expensive, though, so I should know how many I need in order to be reasonably sure I have enough. This is called estimating my *power*, $(1 - \beta)$"
- "This will also depend on my *effect size*. A larger effect size requires less evidence, while a smaller effect size requires more"

Effect Size

In short, an effect size, often denoted δ, is a measure of the strength of the relationship between two variables

Many different methods of determining effect size, with standardized differences being the most common. This consists of considering the absolute difference (or ratio) between two metrics, adjusted for the amount of variability

The null hypothesis is *never* correct, as nothing has a truly null effect. What is important to consider, then, is how much of an effect is considered meaningful

Power

Recall that our probability of committing a Type II error is given as

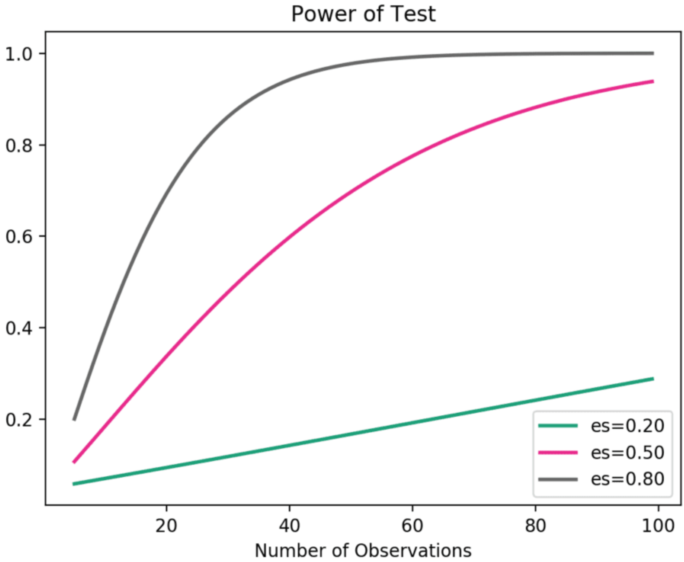
$$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ is false})$$

Our power, then, is the probability of correctly rejecting our null, written $1 - \beta$.

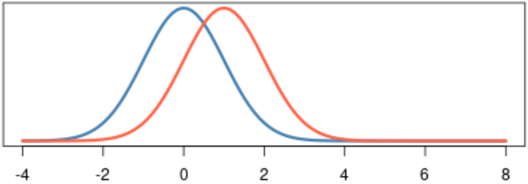
Given a statistical test, there are three things that impact our power

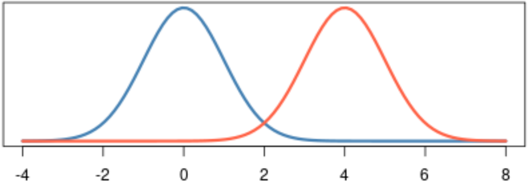
- Magnitude of departure of the observed data from the null or the *effect size*, δ
- The variability of the population or response being studied
- Sample size

Power and Effect Size

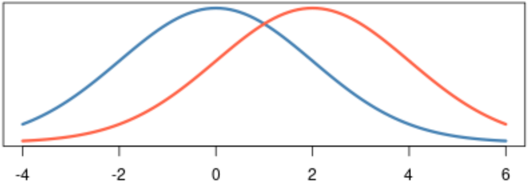


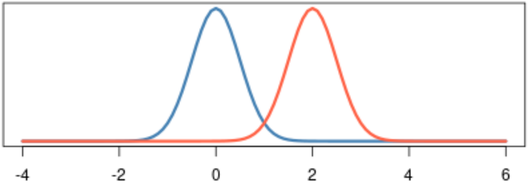
Magnitude of departure



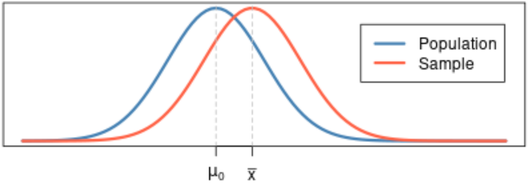


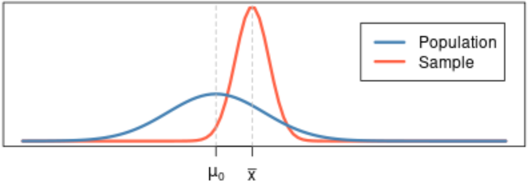
Variability of Population





Sample Size





Things to Consider

Prior to starting a study, it's important to determine a meaningful effect size, δ. A treatment wouldn't be worth pursuing, for example, if it was certain to have an effect on blood pressue that only lowered pressue 2 points

We should also try to anticipate the amount of variability we might see, usually an informed guess based on prior studies

Having determined a meaningful effect size and an estimate of variability, a *sensitivity analysis* can be performed to estimate power based on a number of sample sizes and different estimates of variability

Remarks on sample sizes

Some statistical test are more powerful than others, which can reduce the number of subjects needed. However, this is at the cost of greater assumptions

Attrition needs to be accounted for in all studies. Some subjects drop out, others lost for follow up, etc.,.

Depending on the cost of the study, one may have to concede the need for a larger effect size to account for a smaller number of participants

Sample sizes that are too large can be unnecessarily expensive, and samples that are too small may not be powerful enough, wasting resources. It's therefore often of critical importance to determine a reasonable sample size at the onset

Sensitivity and Specificity

We have a specific instance of of Type I and Type II errors associated with 2×2 tables, often used in diagnostic testing or classification (with H_0 being no disease, a positive test being a rejection of H_0)

Test Result	True State of Nature	
	No Condition	Have Condition
Negative Test	True Positive	False Positive
Positive Test	False Negative	True Negative

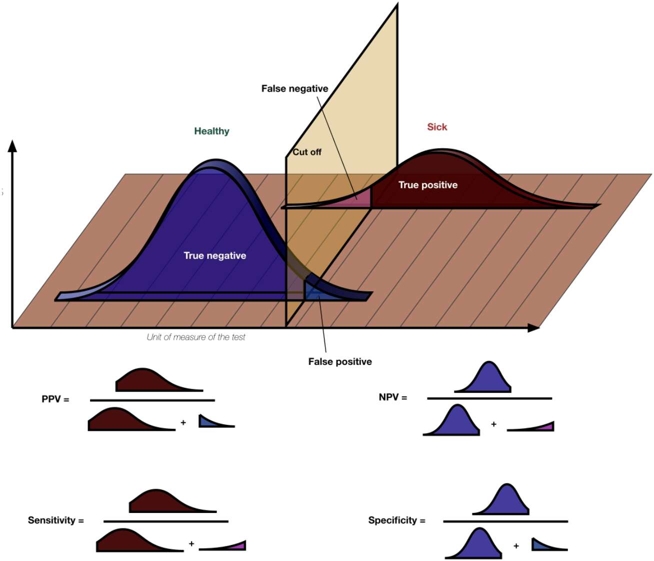
$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$= \text{Probability of positive test, given patient has condition}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

$$= \text{Probability of negative test, given that patient is well}$$

Sensitivity and Specificity, cont.



Review

- Type I and Type II errors
- Effect Size
- Power
- Properties of samples and testing
- Sensitivity and Specificity