Sta 111 - Summer II 2017
Probability and Statistical Inference

11. Decision errors & power of a test

Lu Wang

Duke University, Department of Statistical Science

July 19, 2017

Outline

1. Decision errors

1. Type 1 and 2 error rates are traded off against each other

2. Power of a test

1. A two-step process for calculating the power
2. A priori power calculations determine desired sample size
3. Power goes up with effect size and sample size, and is inversely proportional with significance level and standard error

3. Results that are statistically significant are not necessarily practically significant

Decision errors

- Hypothesis tests are not flawless. We could make a wrong decision in hypothesis tests.
- But we have the tools necessary to quantify how often we make errors in statistics.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer."

– William Blackstone

Type 1 error rate - significance level α

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. $\alpha = 0.01$). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

Type 2 error rate

- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.
- If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?
- The answer is not obvious. We denote the Type 2 error rate as β.
  - If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
  - If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
  - Clearly, β depends on the *effect size* δ (difference between the true population parameter and the null value).
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. $\alpha = 0.10$). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Example - Blood Pressure (BP), minimum effect size required to reject H_0

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication, and half of the subjects are given the new drug (treatment) and the other half continue to take their current medication through generic-looking pills to ensure blinding (control). What are the hypotheses for a two-sided hypothesis test in this context?

BP: minimum effect size required to reject H_0

Suppose researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric. If we had 100 patients per group, what would be the approximate standard error for difference in sample means of the treatment and control groups?

$$SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

BP: minimum effect size required to reject H_0

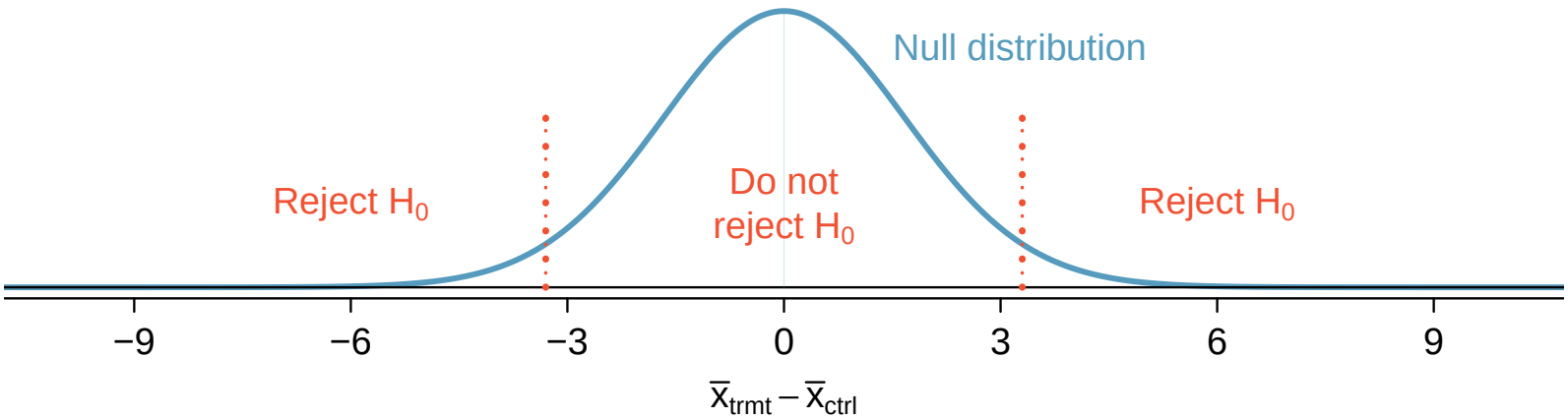
For what values of the difference between the observed averages of blood pressure in treatment and control groups (effect size) would we reject the null hypothesis at the 5% significance level?

The difference should be at least

$$1.96 * 1.70 = 3.332$$

or at most

$$-1.96 * 1.70 = -3.332.$$



Type 1 and 2 error rates are traded off against each other.

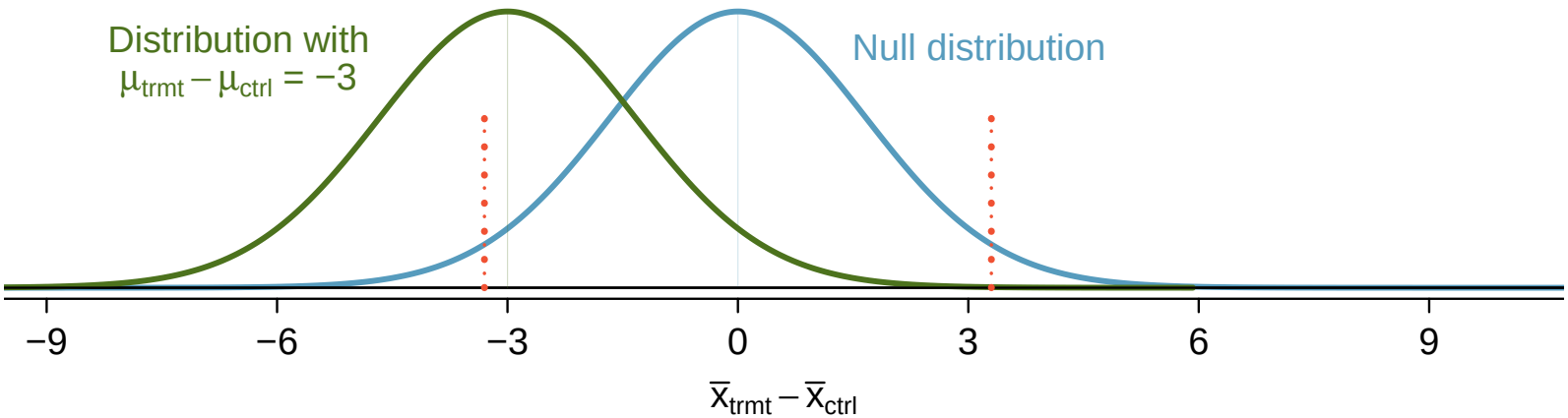
- For any given sample set, the effort to reduce one type of error generally results in increasing the other type of error.
- For a given test, sometimes the only way to reduce both error rates is to increase the sample size.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1-\beta$

The *power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$.

Blood Pressure Example, power

Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger. What is the power of the test that can detect this effect? i.e. What is the probability that we would reject this H_0 if $\bar{x}_{trmt} - \bar{x}_{crtl}$ had come from a distribution with $\mu_{trmt} - \mu_{crtl} = -3$?



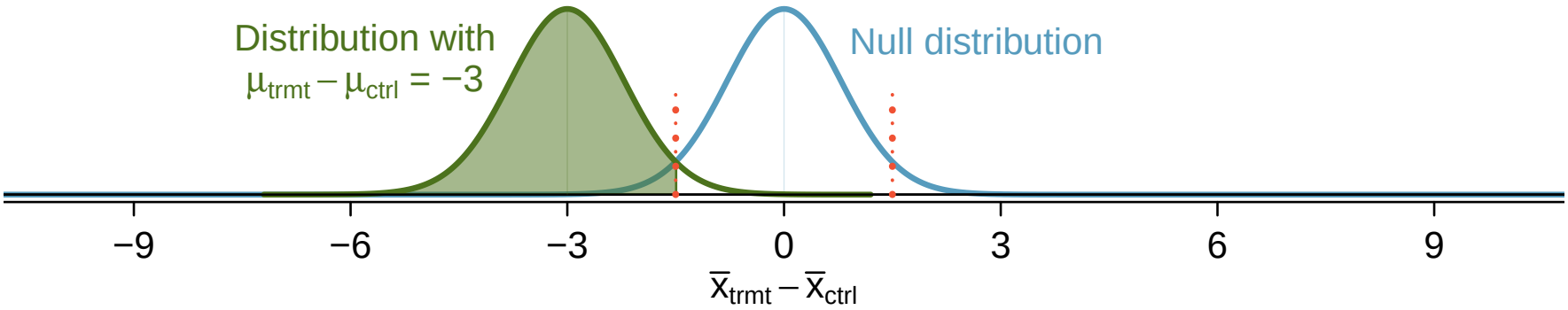
$$Z = \frac{-3.332 - (-3)}{1.70} = -0.20$$

$$P(Z < -0.20) = 0.4207$$

Recap - Calculating Power

- *Step 0:* Pick a meaningful effect size δ and a significance level α
- *Step 1:* Find the rejection region for the point estimate where you would reject H_0 at the chosen α level.
- *Step 2:* Calculate the probability of observing a value from preceding step if the sample was drawn from a population where $\mu = \mu_{H_0} + \delta$

What sample size will lead to a power of 80% for this test?




```
> qnorm(p=0.8)
[1] 0.8416212
```

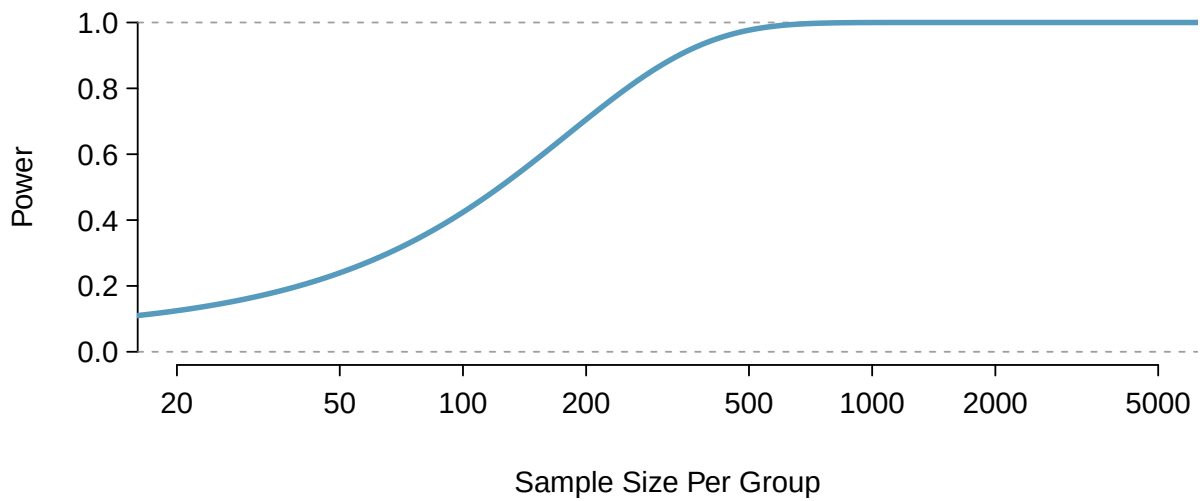
$$SE = \frac{3}{0.84 + 1.96} = \frac{3}{2.8} = 1.07142$$

$$1.07142 = \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = 250.88 \rightarrow n \geq 251$$

Power vs sample size

- Calculate required sample size for a desired level of power (usually 80% or 90%).
- Calculate power for a range of sample sizes, then choose the sample size that yields the target power,



Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation *s* of the sample. Smaller *s* leads to smaller *SE* and hence we have a better chance of distinguishing the null value from the observed point estimate.
    - This is often achieved by increasing the sample size, but sometimes cautious measurement process and limiting the population to be more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

Suppose $\bar{x} = 5$, $s = 2$, $H_0 : \mu = 4.5$, and $H_A : \mu > 4.5$. Will p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

$$Z_{n=100} = \frac{5 - 4.5}{\frac{2}{\sqrt{100}}} = \frac{5 - 4.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.0062$$

$$Z_{n=10000} = \frac{5 - 4.5}{\frac{2}{\sqrt{10000}}} = \frac{5 - 4.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

As n increases - $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

Statistical vs Practical Significance

- When n is large, even small deviations from the null, which may be considered practically insignificant, can yield statistically significant results.
- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value, even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. " - R.A. Fisher