Lecture 12 - Decisions and Power

Sta102 / BME 102

March 7th, 2016

Colin Rundel

Statistical vs. Practical Significance

Example - Sample Size

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?

Example - Sample Size

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?

$$T_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.007$$

$$T_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

Example - Sample Size

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?

$$T_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.007$$

$$T_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

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

Example - Sample Size 2

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.9$, and $H_A : \mu > 49.9$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?

Example - Sample Size 2

Suppose $\bar{X} = 50$, $s = 2$, $H_0 : \mu = 49.9$, and $H_A : \mu > 49.9$.

Will the p-value be lower if $n = 100$ or $n = 10,000$?

$$T_{n=100} = \frac{50 - 49.9}{\frac{2}{10}} = \frac{0.1}{0.2} = 0.5, \quad \text{p-value} = 0.309$$

$$T_{n=10000} = \frac{50 - 49.9}{\frac{2}{100}} = \frac{0.1}{0.02} = 5, \quad \text{p-value} = 2.87 \times 10^{-7}$$

Statistical vs. Practical Significance

- Differences between the point estimate and null value are easier to detect with larger samples
- Large samples can result in statistical significance even for tiny *effect sizes*, even when the difference is not practically significant
- This is particularly 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).

"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

Decisions and Decision Errors

Decision errors

- Hypothesis Tests and Confidence Intervals both make mistakes.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision using statistical inference methods as well.
- The difference is that we have the ability to quantify / adjust how often we make errors using statistical inference.

Decision errors for HTs

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 errors for HTs

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		
	H_A true		

Decision errors for HTs

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	✓	
	H_A true		

Decision errors for HTs

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	✓	
	H_A true		✓

Decision errors for HTs

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		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

Decision errors for HTs

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.

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 cirumstances?

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

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 cirumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

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 cirumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Hypothesis Test as a trial (cont.)

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

Hypothesis Test as a trial (cont.)

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

"better that ten guilty persons escape than that one innocent suffer"
– William Blackstone

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

"better that ten guilty persons escape than that one innocent suffer"
– William Blackstone

Implications for statistical inference:

- Both types of errors are bad and we want to avoid them - but there is a trade off.
- Generally, type I errors are considered to be worse - so we tune our inference procedures to minimize them.

Type 1 error rate

As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* (α) of 0.05.

Type 1 error rate

As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* (α) of 0.05.

This means that, for those cases where H_0 is actually true, we will incorrectly reject it at most 5% of the time.

Type 1 error rate

As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* (α) of 0.05.

This means that, for those cases where H_0 is actually true, we will incorrectly reject it at most 5% of the time.

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

Type 1 error rate

As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* (α) of 0.05.

This means that, for those cases where H_0 is actually true, we will incorrectly reject it at most 5% of the time.

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

This is why we prefer small values of α – decreasing α decreases our Type 1 error rate.

Filling in the table...

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

Filling in the table...

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true		

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

Filling in the table...

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

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

Type 2 error rate - $\beta = P(\text{Failing to reject } H_0 \mid H_A \text{ is true})$

Filling in the table...

		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	

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

Type 2 error rate - $\beta = P(\text{Failing to reject } H_0 \mid H_A \text{ is true})$

Power - $1 - \beta = P(\text{Rejecting } H_0 \mid H_A \text{ is true})$

Filling in the table...

		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,* $1-\beta$

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

Type 2 error rate - $\beta = P(\text{Failing to reject } H_0 \mid H_A \text{ is true})$

Power - $1-\beta = P(\text{Rejecting } H_0 \mid H_A \text{ is true})$

Type 2 error rate

The type 2 error is defined as

$$\beta = P(\text{Failing to reject } H_0 \mid H_A \text{ is true}) = \,?$$

Type 2 error rate

The type 2 error is defined as

$$\beta = P(\text{Failing to reject } H_0 \mid H_A \text{ is true}) = ?$$

How do we calculate this probability (or its complement)? It is not immediately obvious but we can come up with some basic rules:

- If the true population average is very close to the null hypothesis value (δ likely to be small), it will be difficult to detect the difference (and reject H_0).
- If the true population average is very different from the null hypothesis value (δ likely to be large), it will be easy to detect the difference.

Type 2 error rate - intuition

Intuitively, β depends on

- δ (effect size)
- α (significance level)
- n (sample size)

Type 2 error rate - intuition

Intuitively, β depends on

- δ (effect size)
- α (significance level)
- n (sample size)

to increase power / decrease β:

- increase n,
- increase δ, and/or
- increase α

Power

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is *greater* than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is *greater* than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

$$H_0 : \mu = 130$$
$$H_A : \mu > 130$$

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is *greater* than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

$$H_0 : \mu = 130$$
$$H_A : \mu > 130$$

We'll start with a very specific question – "What is the power of this hypothesis test to correctly detect an *increase* of 2 mmHg in average blood pressure?"

Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?

Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let's break this down into two simpler problems:

Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let's break this down into two simpler problems:

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?

Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let's break this down into two simpler problems:

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
2. Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from a distribution with $\mu = 132$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

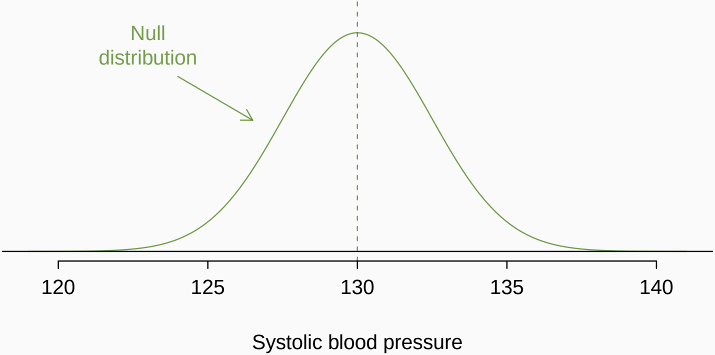
Problem 1

Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130$, $H_A : \mu > 130$)

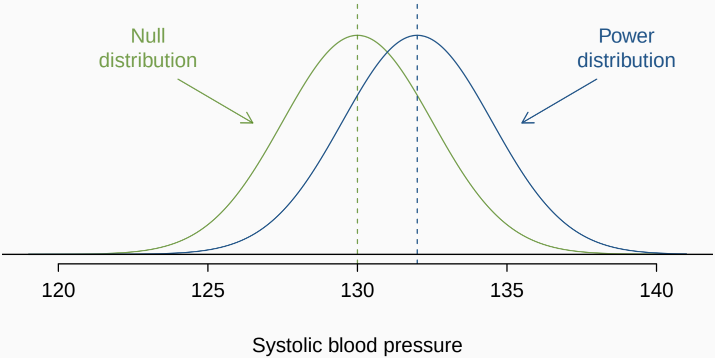
Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ came from a distribution where $\mu = 132$.

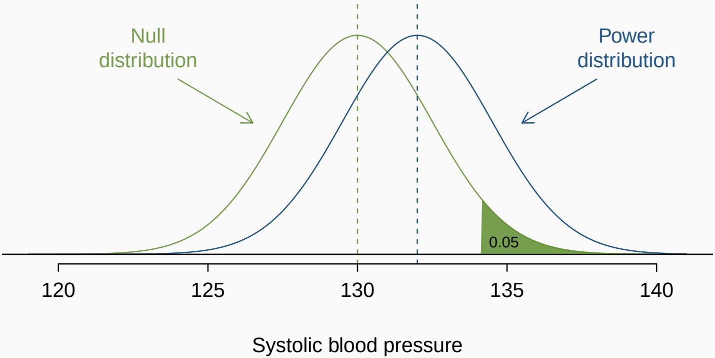
Putting it all together



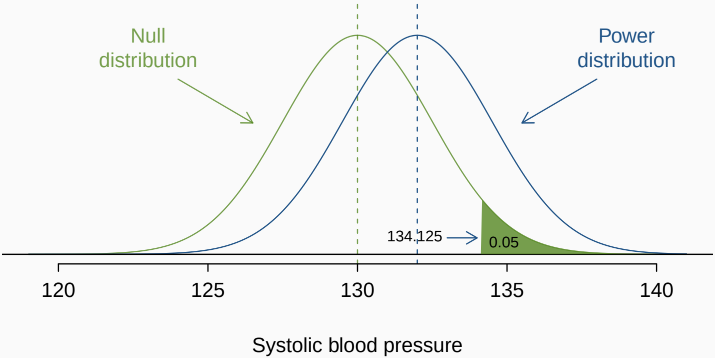
Putting it all together



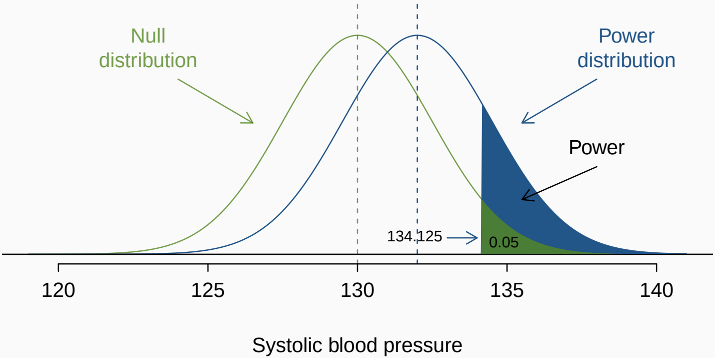
Putting it all together



Putting it all together



Putting it all together



Recap - Calculating Power

- *Step 0:* Pick a meaningful effect size δ and a significance level α
- *Step 1:* Calculate the range of values for the point estimate beyond which 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 derived from a population where $\mu = \mu_{H_0} + \delta$

Example - Power for a two sided hypothesis test

Going back to the blood pressure example, what would the power be to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level for a sample of 625 patients?

Example - Power for a two sided hypothesis test

Going back to the blood pressure example, what would the power be to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level for a sample of 625 patients?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad n = 625, \quad \sigma = 25, \quad \delta = 4, \quad 1 - \beta = ?$$

Step 1:

$$P(T > t \text{ or } T < -t) < 0.05 \quad \Rightarrow \quad t > 1.96$$
$$\bar{x} > 130 + 1.96\frac{25}{\sqrt{625}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{625}}$$
$$\bar{x} > 131.96 \text{ or } \bar{x} < 128.04$$

Step 2: Assume $\mu = \mu_{H_0} + \delta = 134$

$$\begin{aligned} P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) &= P(T > [131.96 - 134]/1) + P(T < [128.04 - 134]/1) \\ &= P(T > -2.04) + P(T < -5.96) \\ &= 0.979 + 0 = 0.979 \end{aligned}$$

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level?

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4, \quad n = ?$$

Step 1:

$$P(T > t \text{ or } T < -t) < 0.05 \quad \Rightarrow \quad t > 1.96$$
$$\bar{x} > 130 + 1.96\frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{n}}$$

Step 2: Assume $\mu = \mu_{H_0} + \delta = 134$

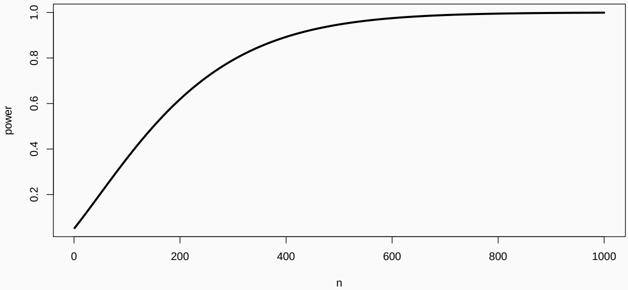
$$P\left(\bar{x} > 130 + 1.96\frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{n}}\right) = 0.9$$
$$P\left(T > 1.96 - 4\frac{\sqrt{n}}{25} \text{ or } T < -1.96 - 4\frac{\sqrt{n}}{25}\right) = 0.9$$

Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?

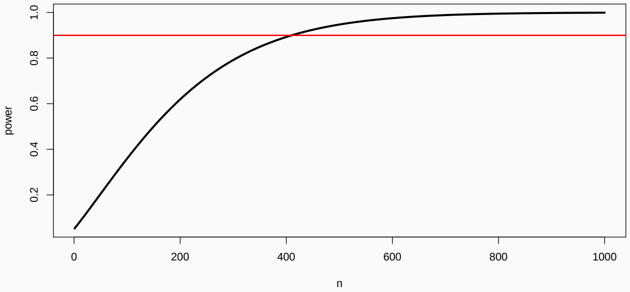
Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?



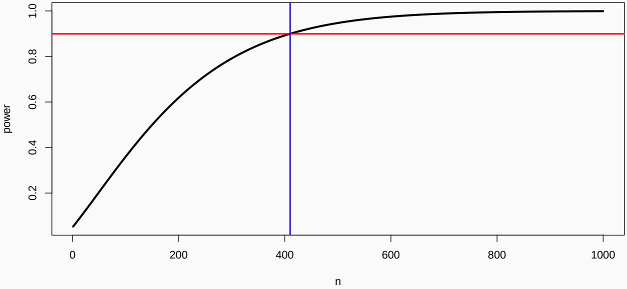
Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?



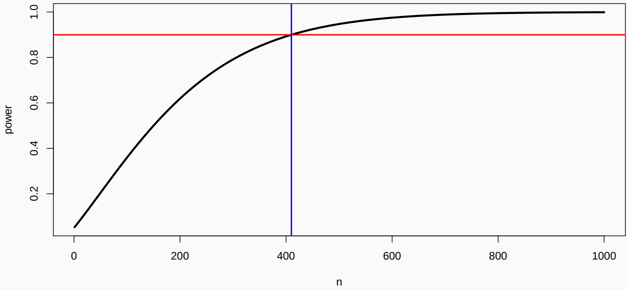
Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?



Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?



For $n = 410$ the power $= 0.8996$, therefore we need 411 subjects in our sample to achieve the desired level of power for the given circumstance.