

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Lecture 20 - Power Analysis

Sta230 / Mth230

Colin Rundel

June 12, 2014

Decision errors (cont.)

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

- 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 0.05, $\alpha = 0.05$.
- 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.

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

- This is why we prefer to small values of α – increasing α increases the Type 1 error rate.

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 is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- Power* of a test is the probability of correctly rejecting H_0 , and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

Type 2 error rate

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.
- 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)

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?

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 preceding 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 $N(\text{mean} = 132, SE = \frac{25}{\sqrt{100}} = 2.5)$, i.e. what is the probability that we can obtain such an \bar{x} from this distribution?

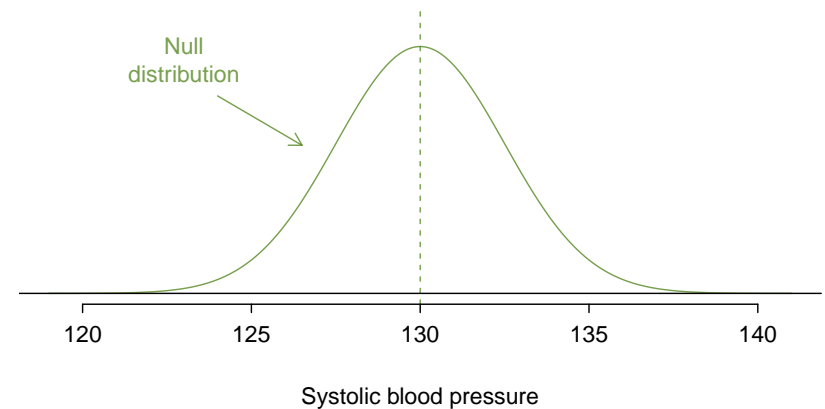
Problem 1

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

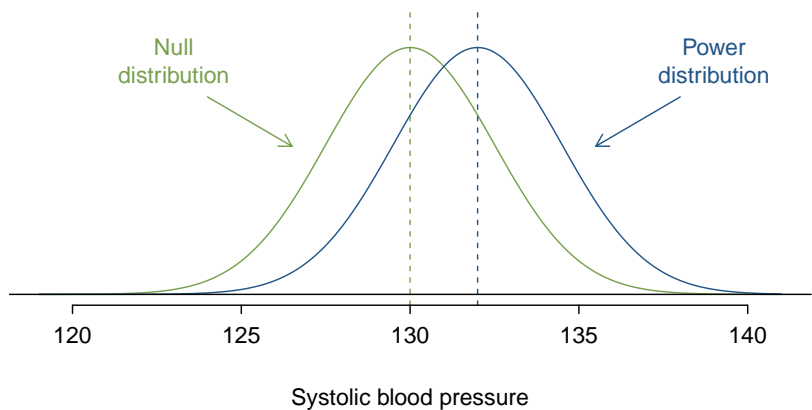
Problem 2

What is the probability that we would reject H_0 if \bar{x} did come from $N(\text{mean} = 132, SE = 2.5)$.

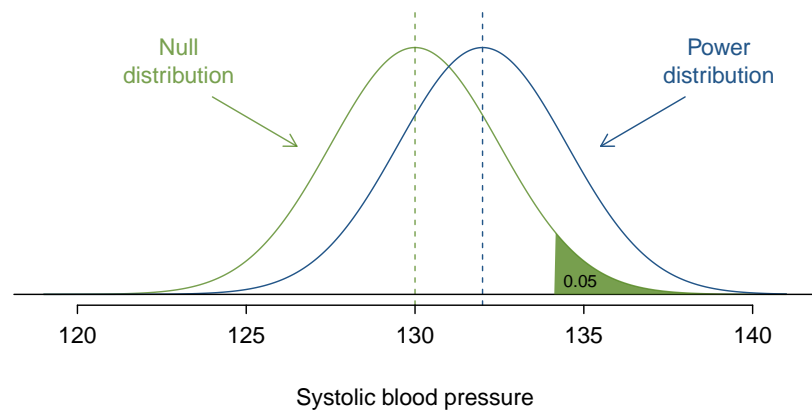
Putting it all together



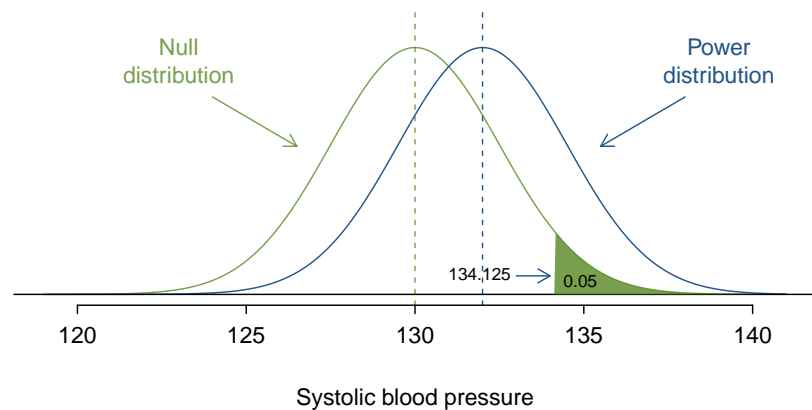
Putting it all together



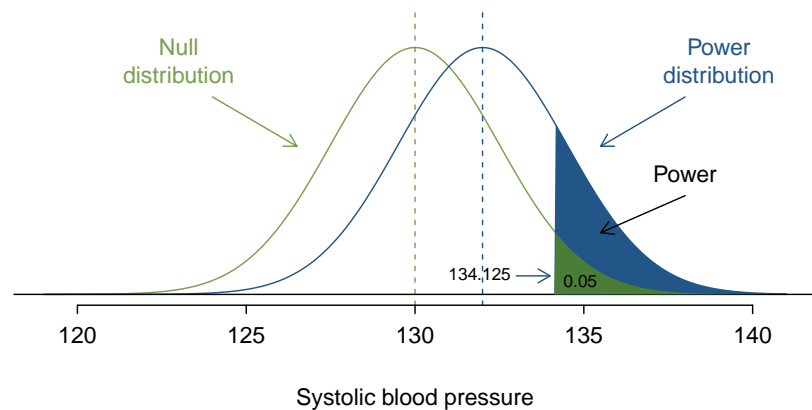
Putting it all together



Putting it all together



Putting it all together



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 of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is 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.

Recap - Calculating Power

- Begin by picking a meaningful effect size δ and a significance level α
- Calculate the range of values for the point estimate beyond which you would reject H_0 at the chosen α level.
- Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\bar{x} \sim N(\mu_{H_0} + \delta, SE)$

Example - Calculating 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:

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

Step 1:

$$P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96$$

$$\bar{x} > 130 + 1.96 \frac{25}{\sqrt{625}} \text{ or } \bar{x} < 130 - 1.96 \frac{25}{\sqrt{625}}$$

$$\bar{x} > 131.96 \text{ or } \bar{x} < 128.04$$

Step 2:

$$\bar{x} \sim N(\mu + \delta, SE) = N(134, 1)$$

$$P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) = P(Z > [131.96 - 134]/1) + P(Z < [128.04 - 134]/1)$$

$$= P(Z > -2.04) + P(Z < -5.96) = 0.979 + 0 = 0.979$$

Example - Effect Size vs. Power - One Sided

Lets go back to the blood pressure example, as see if we can find an equation for power in terms of an effect size δ change in average blood pressure for the hypothesis that the population average is greater than 130 mmHg at a 95% significance level for a sample of 625 patients?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu > 130, \quad \alpha = 0.05, \quad n = 625, \quad \sigma = 25, \quad \delta = \delta$$

Step 1:

$$P(Z > z) < 0.05 \quad \Rightarrow \quad z > 1.645$$

$$\bar{x} > 130 + 1.645 \frac{25}{\sqrt{625}}$$

$$\bar{x} > 131.645$$

Step 2:

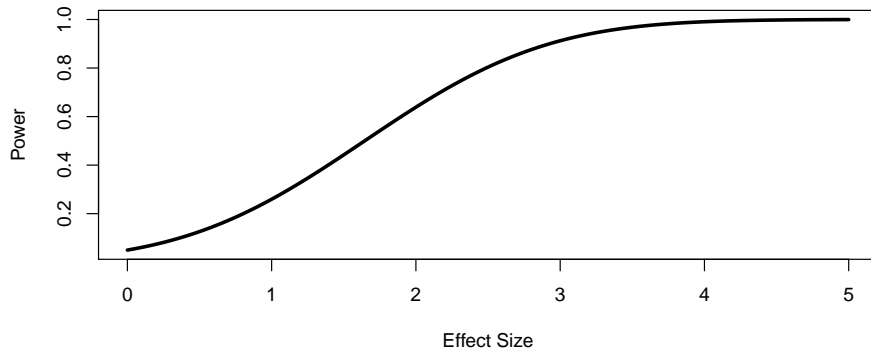
$$\bar{x} \sim N(\mu + \delta, SE) = N(130 + \delta, 1)$$

$$P(\bar{x} > 131.645) = P(Z > [131.645 - 130 - \delta]/1)$$

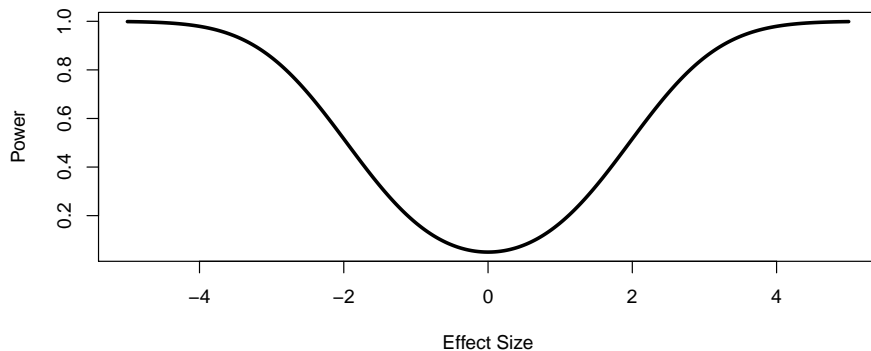
$$= P(Z > 1.645 - \delta)$$

$$= 1 - \Phi(1.645 - \delta)$$

Example - Effect Size vs. Power - One Sided, cont.



Example - Effect Size vs. Power - Two Sided, cont.



Example - Effect Size vs. Power - Two Sided

We can do the same for a two sided hypothesis as well.

Step 0:

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

Step 1:

$$P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96$$

$$\bar{x} > 130 + 1.96 \frac{25}{\sqrt{625}} \text{ or } \bar{x} < 130 - 1.96 \frac{25}{\sqrt{625}}$$

$$\bar{x} > 131.96 \text{ or } \bar{x} < 128.04$$

Step 2:

$$\bar{x} \sim N(\mu + \delta, SE) = N(130 + \delta, 1)$$

$$P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) = P(Z > [131.96 - 130 - \delta]/1) + P(Z < [128.04 - 130 - \delta]/1)$$

$$= P(Z > 1.96 - \delta) + P(Z < -1.96 - \delta)$$

$$= 1 - \Phi(1.96 - \delta) + \Phi(-1.96 - \delta)$$

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:

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

Step 1:

$$P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96$$

$$\bar{x} > 130 + 1.96 \frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96 \frac{25}{\sqrt{n}}$$

Step 2:

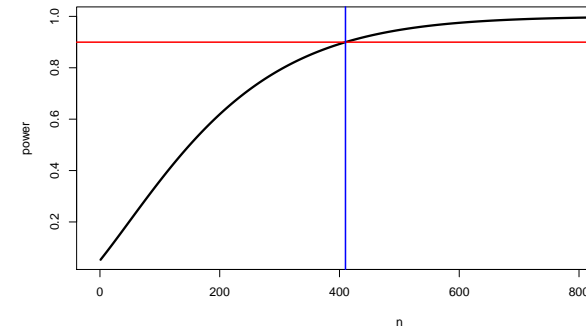
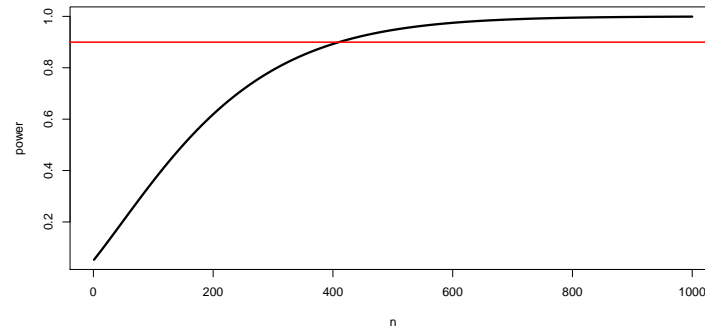
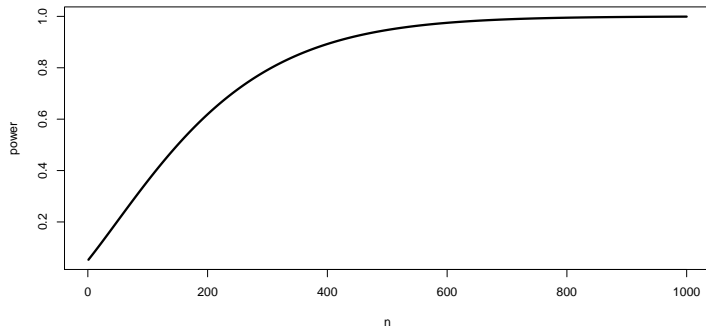
$$\bar{x} \sim N(\mu + \delta, SE) = N(134, 25/\sqrt{n})$$

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

$$P\left(Z > 1.96 - 4 \frac{\sqrt{n}}{25} \text{ or } Z < -1.96 - 4 \frac{\sqrt{n}}{25}\right) = 0.9$$

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

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



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.