Unit 3: Foundations for inference
Lecture 3: Decision errors, significance levels, sample size, and power

Statistics 101

Mine Çetinkaya-Rundel

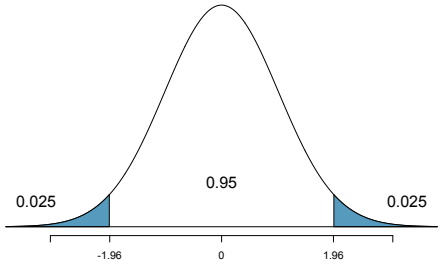
October 1, 2013

Announcements

- Project proposal due 5pm on Friday, on Sakai + hard copy at my office
- Remaining office hours for the week:
  - Tuesday - Mine: 4:30-8pm
  - Wednesday - Mine: 1:30-3pm & George: 5-7pm
  - Thursday - Mine: 1-2pm & Daiana: 4:30-5:30pm
- MT review materials posted on the course website (past midterm + practice problems)
- Optional: MT Review session: Sunday (10/6), 3-4pm (Location TBA)

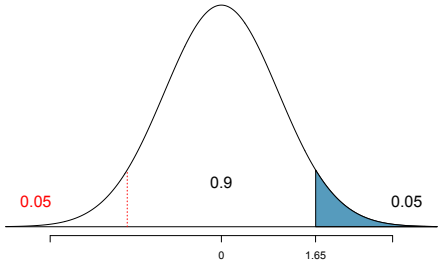
Significance level vs. confidence level

Two sided



Two sided HT with $\alpha = 0.05$ is equivalent to 95% confidence interval.

One sided



One sided HT with $\alpha = 0.05$ is equivalent to 90% confidence interval.

Agreement of CI and HT

- Confidence intervals and hypothesis tests (almost) always agree, as long as the two methods use equivalent levels of significance / confidence.
  - A two sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - \alpha$.
  - A one sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - (2 \times \alpha)$.
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

Clicker question

A 95% confidence interval for the average waiting time at an emergency room is (128 minutes, 147 minutes). Which of the following is false?

(a) A hypothesis test of $H_A : \mu \neq 120$ min at $\alpha = 0.05$ is equivalent to this CI.

(b) A hypothesis test of $H_A : \mu > 120$ min at $\alpha = 0.025$ is equivalent to this CI.

(c) This interval does not support the claim that the average wait time is 120 minutes.

(d) The claim that the average wait time is 120 minutes would not be rejected using a 90% confidence interval.

Sample Size

Clicker question

All else held equal, will p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Statistical vs. Practical Significance

- 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 (*effect size*), 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

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.

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 : Defendant is innocent

H_A : 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

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?"

Application exercise: 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?

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

Time permitting: Determine how power changes as sample size, standard deviation of the sample, α, and effect size increases.

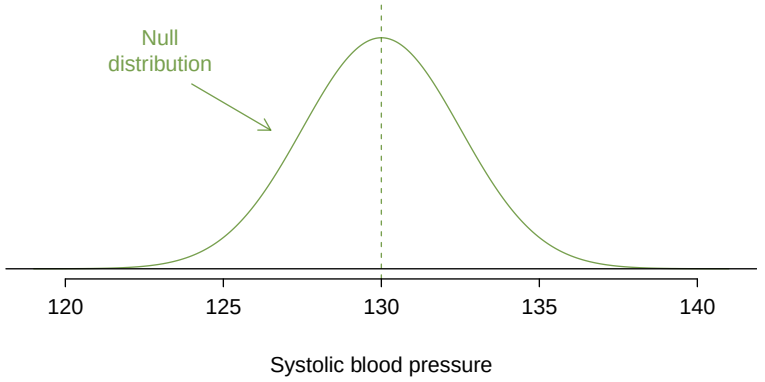
Problem 1

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

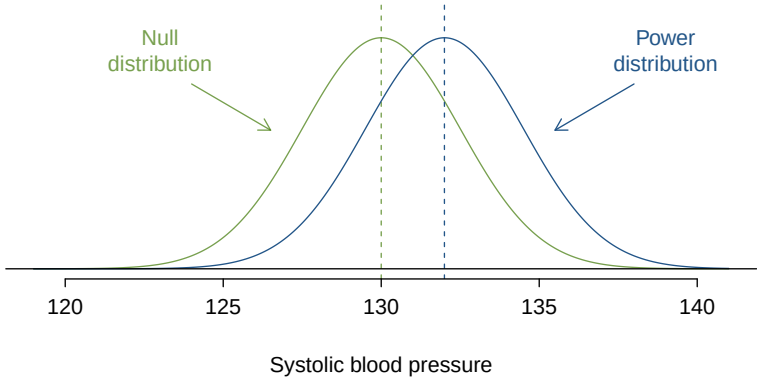
Problem 2

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

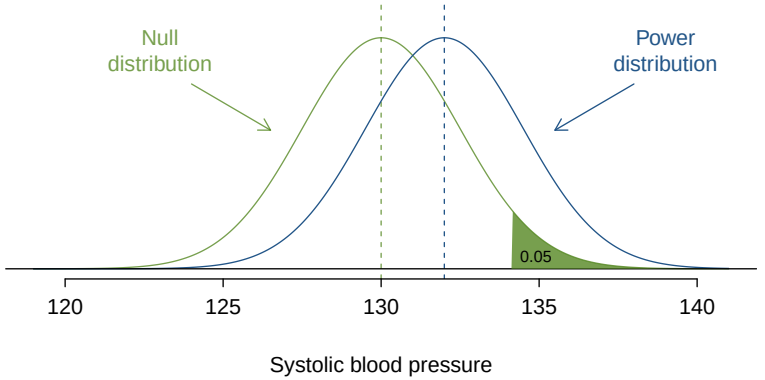
Putting it all together



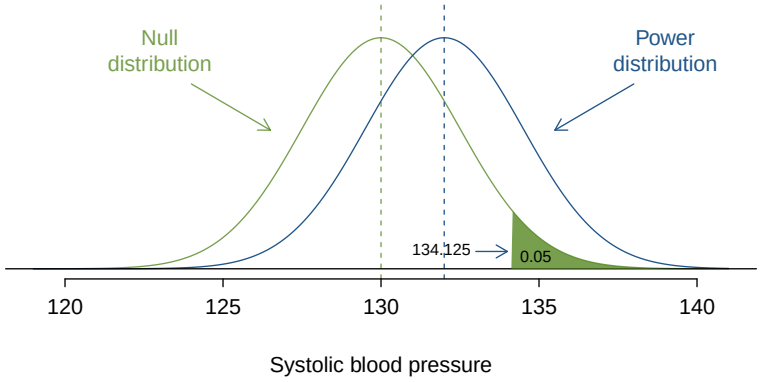
Putting it all together



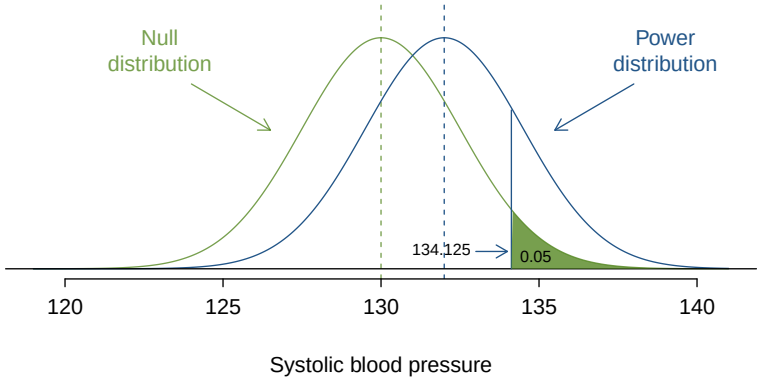
Putting it all together



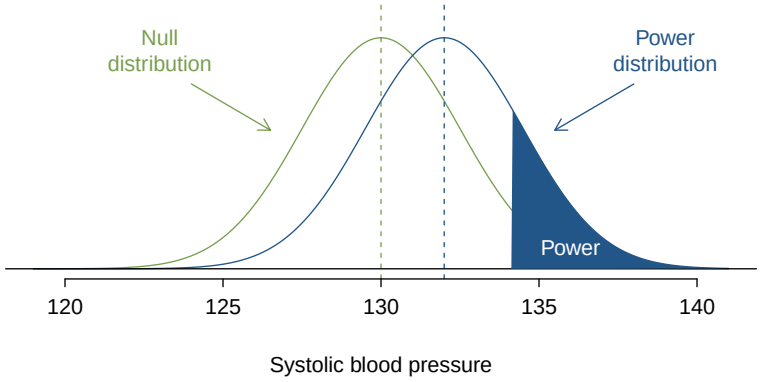
Putting it all together



Putting it all together



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 - 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 greater than 130 mmHg at $\alpha = 0.05$?

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

Step 1: Determine the cutoff – in order to reject H_0 at $\alpha = 0.05$, we need a sample mean that will yield a Z score of at least 1.65.

$$\bar{x} > 130 + 1.65\frac{25}{\sqrt{n}}$$

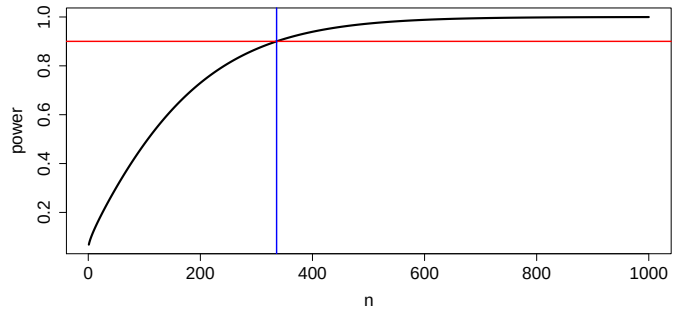
Step 2: Set the probability of obtaining the above $\bar{x}$ if the true population is centered at 130 + 4 = 134 to the desired power, and solve for n.

$$P\left(\bar{x} > 130 + 1.65\frac{25}{\sqrt{n}}\right) = 0.9$$

$$P\left(Z > \frac{\left(130 + 1.65\frac{25}{\sqrt{n}}\right) - 134}{\frac{25}{\sqrt{n}}}\right) = P\left(Z > 1.65 - 4\frac{\sqrt{n}}{25}\right) = 0.9$$

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

You can either directly solve for n, or use computation to calculate power for various n and determine the sample size that yields the desired power:



For $n = 336$, power $= 0.9002$, therefore we need 336 subjects in our sample to achieve the desired level of power for the given circumstance.