Lecture 12 - Decisions and Power

Sta102 / BME 102

Colin Rundel

October 14th, 2015

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 $n = 10,000$?

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

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

Statistical vs. Practical Significance

- 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 and Confidence Intervals 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 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.
- 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 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

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 will incorrectly reject it at most 5% of the time.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error.

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

- This is why we prefer 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, but
  - 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.
- Therefore, β must depend on the *effect size* (δ) in some way

To increase power / decrease β: increase n, increase δ, or increase α

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?

Let's 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 a distribution with $\mu = 132$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

Problem 1

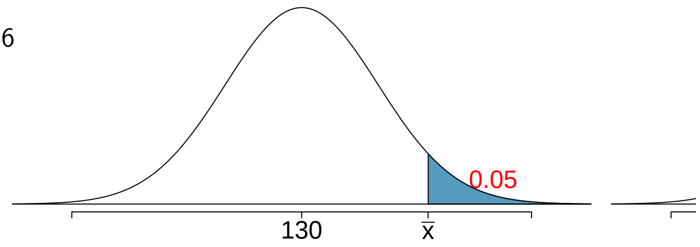
Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130,\ H_A : \mu > 130$)

$$P(T > t) < 0.05 \ \Rightarrow \ t > 1.66$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} > 1.66$$

$$\bar{x} > 130 + 1.66 \times 2.5$$

$$\bar{x} > 134.15$$



Any $\bar{x} > 134.15$ would be sufficient to reject H_0 at the 5% significance level.

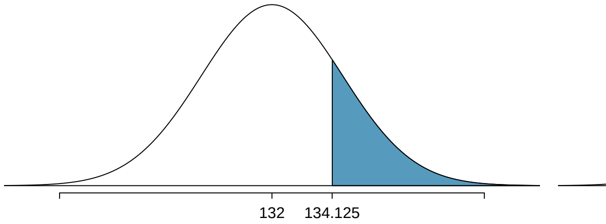
Problem 2

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

This is the same as finding the area above $\bar{x} = 134.125$ if the sampling distribution were centered at 132.

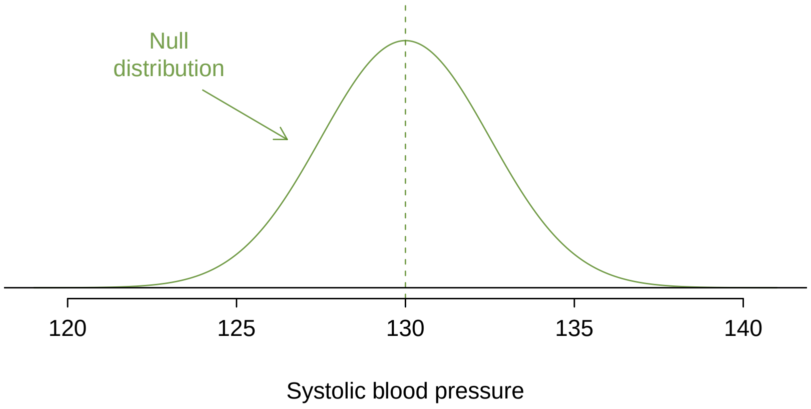
$$T = \frac{134.125 - 132}{2.5}$$
$$= 0.85$$

$$P(T > 0.85) = 1 - 0.801$$
$$= 0.199$$

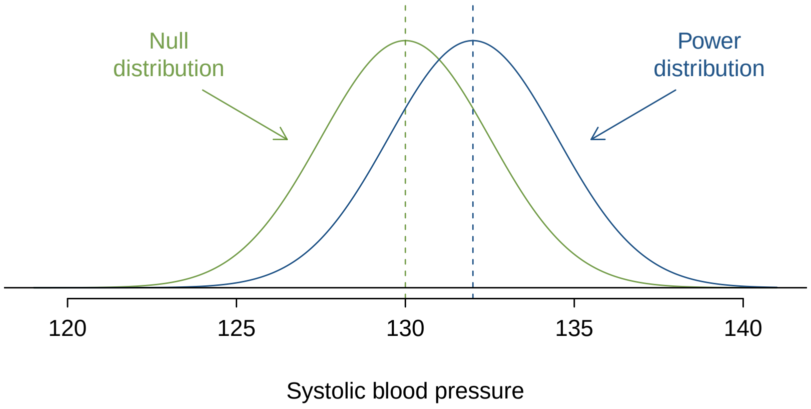


The probability of rejecting $H_0 : \mu = 130$, if the true average systolic blood pressure of employees at this company is 132 mmHg, is 0.199 which is the power of this test.

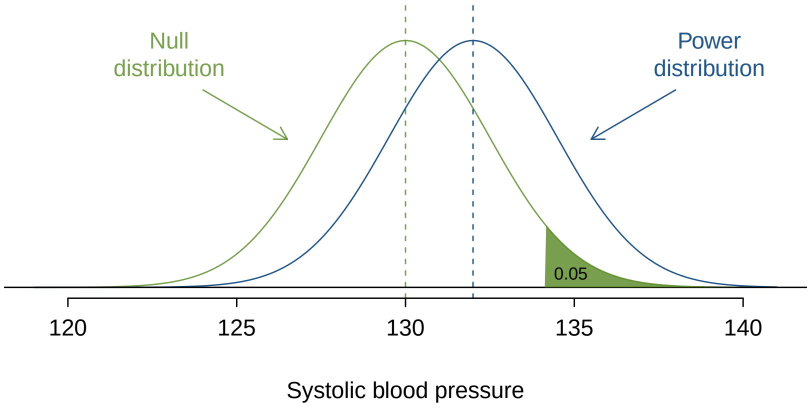
Putting it all together



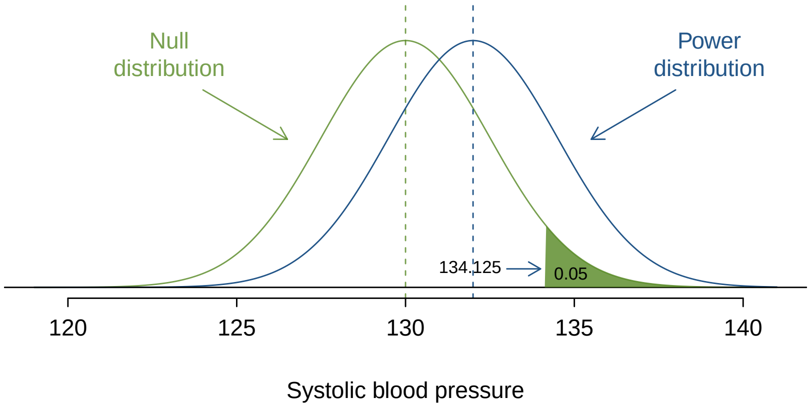
Putting it all together



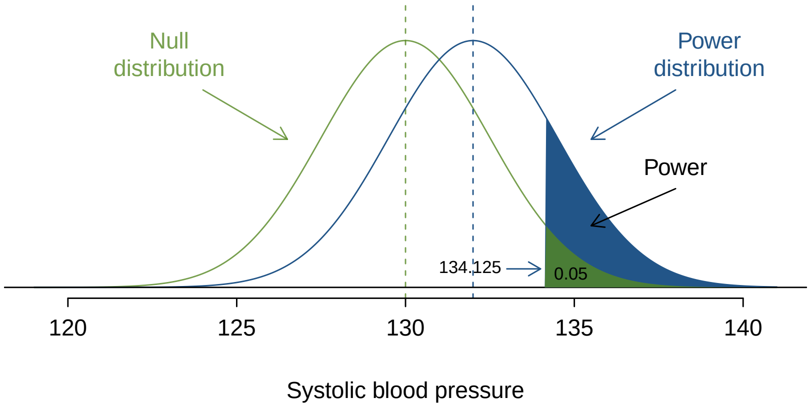
Putting it all together



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 is equivalent to 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

- *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?

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?

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:

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

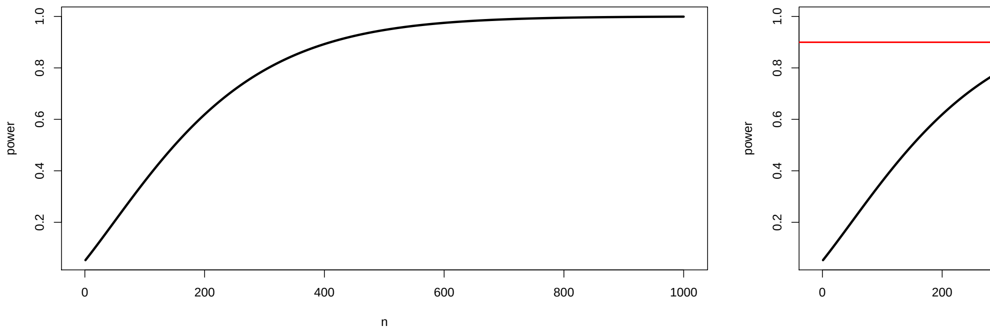
Step 2: Assume $\mu = \mu_{H_0} + \delta = 134$

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

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

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

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

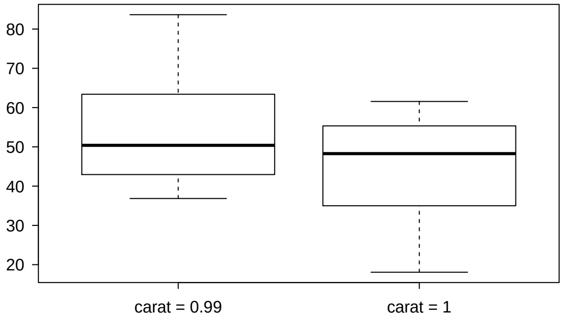


For $n = 410$ the power $= 0.8996$, therefore we need 411 subjects in our sample to achieve the desired level of power for the given circumstance.

Example - Diamonds

- Weights of diamonds are measured in carats.
- 1 carat = 100 points, 0.99 carats = 99 points, etc.
- The difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond.
- We are going to test to see if there is a difference between the average prices of 0.99 and 1 carat diamonds.
- In order to be able to compare equivalent units, we divide the prices of 0.99 carat diamonds by 99 and 1 carat diamonds by 100, and compare the average point prices.

Data



	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

These data are a random sample from the `diamonds` data set in the `ggplot2` R package.

Parameter and point estimate

- *Parameter of interest:* Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds.

$$\mu_{pt99} - \mu_{pt100}$$

- *Point estimate:* Average difference between the point prices of *sampled* 0.99 carat and 1 carat diamonds.

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

- *Hypotheses:* testing if the average per point price of 1 carat diamonds ($_{pt100}$) is higher than the average per point price of 0.99 carat diamonds ($_{pt99}$)

$$H_0 : \mu_{pt99} = \mu_{pt100}$$
$$H_A : \mu_{pt99} < \mu_{pt100}$$

Hypothesis test

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

$$\begin{aligned} T &= \frac{\text{point estimate} - \text{null value}}{SE} \\ &= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\ &= \frac{-8.93}{3.56} \\ &= -2.508 \end{aligned}$$

What is the correct df for this hypothesis test?

$$\begin{aligned} df &= min(n_{pt99} - 1, n_{pt100} - 1) \\ &= min(23 - 1, 30 - 1) \\ &= min(22, 29) = 22 \end{aligned}$$

p-value

What is the correct p-value for the hypothesis test?

$$T = -2.508$$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79

Synthesis

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

- p-value is small so we rejected H_0. The data provide convincing evidence to suggest that the per point price of 0.99 carat diamonds is lower than the per point price of 1 carat diamonds.
- Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is significantly cheaper.

Critical value

What is the appropriate $t^\star$ for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds that would be equivalent to our hypothesis test?

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

$$\begin{aligned}(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^\star_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\ &= -8.93 \pm 6.12 \\ &= (-15.05, -2.81)\end{aligned}$$

We are 90% confident that the average point price of a 0.99 carat diamond is $15.05 to $2.81 lower than the average point price of a 1 carat diamond.

Power

What is the power of our hypotheses and data to detect a difference of $9 per point?

Step 0:

$$H_0 : \mu_{99} = \mu_{100}, \quad H_A : \mu_{99} < \mu_{100}$$

$$\alpha = 0.05, \quad n_{99} = 23, \quad n_{100} = 30, \quad SE = 3.56, \quad df = 22, \quad \delta = 9, \quad 1 - \beta = ?$$

Step 1:

$$P(T > t) < 0.05 \quad \Rightarrow \quad t > 1.72$$
$$P\left(\frac{\bar{x}_{100} - \bar{x}_{99} - 0}{3.56} > 1.72\right) = 0.05$$
$$\bar{x}_{100} - \bar{x}_{99} > 0 + 1.72 \times 3.56$$
$$\bar{x}_{100} - \bar{x}_{99} > 6.12$$

Step 2: Assume $\mu_{100} - \mu_{99} = \delta = 9$

$$\begin{aligned} &P(\bar{x}_{100} - \bar{x}_{99} > 6.12 | \mu_{100} - \mu_{99} = 9) \\ &= P\left(T > \frac{6.12 - 9}{3.56}\right) \\ &= P(T > -0.8089) \\ &= 0.786 \end{aligned}$$