OpenIntro Statistics

CH 04B: Foundations for Inference

1

Confidence Intervals

2

Confidence intervals

A plausible range of values for the population parameter is called a *confidence interval*.

Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer and Chris Penny (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

3

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$\bar{x} = 3.2$ $s = 1.74$

The approximate 95% confidence interval is defined as point estimate ± 2 x SE

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$\bar{x} \pm 2 \times SE \rightarrow 3.2 \pm 2 \times 0.25$

$\rightarrow (3.2 - 0.5, 3.2 + 0.5)$

$\rightarrow (2.7, 3.7)$

4

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

5

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

6

A more accurate interval

A general formula for confidence intervals:

$$\textit{point estimate} \pm z^* \times SE$$

Conditions when the point estimate = $\bar{x}$

Independence: Observations in the sample must be independent

- random sample/assignment
- if sampling without replacement, $n < 10\%$ of population

Sample size / skew: $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ in the next chapter.

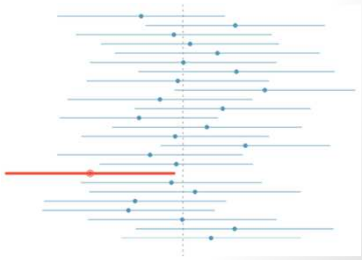
7

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* $\pm$ *2 x SE*.

Then about 95% of those intervals would contain the true population mean (μ).

The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



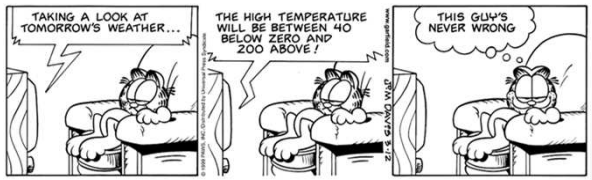
8

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

9

Changing the confidence level

point estimate ± z x SE*

In a confidence interval, *z* x SE* is called the *margin of error*, and for a given sample, the margin of error changes as the confidence level changes.

In order to change the confidence level we need to adjust *z** in the above formula. (See *t* table)

Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

For a 95% confidence interval, *z** = 1.96.

However, using the standard normal (*z*) distribution, it is possible to find the appropriate *z** for any confidence level.

10

Finding the Critical Value with a *t* table

For a particular confidence level *C (found at the top of the table)*, the appropriate *z** value is in the *z** row in the same column of the table.

- For 90% confidence, *z** = 1.645
- For 95% confidence, *z** = 1.960
- For 99% confidence, *z** = 2.576

Confidence Level C

df	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
*z**	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291
One-sided *P*	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
Two-sided *P*	0.50	0.40	0.30	0.20	0.10	0.05	0.04	0.02	0.01	0.01	0.002	0.001

11

Practice

Which of the below *Z* scores is the appropriate *z** when calculating a 98% confidence interval?

(a) *Z* = 2.05

(b) *Z* = 1.96

(c) *Z* = 2.33

(d) *Z* = -2.33

(e) *Z* = -1.65

12

Practice

Which of the below *Z* scores is the appropriate *z** when calculating a 98% confidence interval?

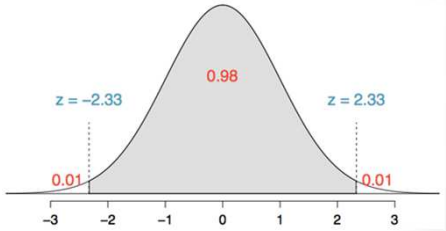
(a) *Z* = 2.05

(b) *Z* = 1.96

(c) *Z* = 2.33

(d) *Z* = -2.33

(e) *Z* = -1.65



13

Hypothesis Testing

14

A Trial as a Hypothesis Test

Hypothesis testing is very much like a court trial.

- H_0 : Defendant is innocent
  H_A : Defendant is guilty
- We then present the evidence - collect data.
- Then we judge the evidence - "Could these data plausibly have happened by chance if the null hypothesis were true?"
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately we must make a decision. How unlikely is unlikely?

Image from http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot

15

A Trial as a Hypothesis Test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of "not guilty".
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- Said statistically, we fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Therefore we never ``accept the null hypothesis".

16

A Trial as a Hypothesis Test (cont.)

- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs (the status quo), so it's the alternative hypothesis that we consider unusual and for which we must gather evidence.

17

Remember when...

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$\hat{p}_{males} = 21 / 24 = 0.88$ $\hat{p}_{females} = 14 / 24 = 0.58$

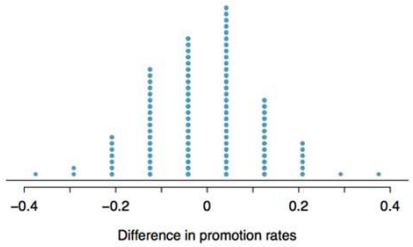
Possible explanations:

Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance. → null (nothing is going on)

Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance. → alternative (something is going on)

18

Simulation Results



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we decided to reject the null hypothesis in favor of the alternative.

19

Recap: hypothesis testing framework

We start with a *null hypothesis* (H_0) that represents the status quo.

We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.

We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).

If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

20

Number of college applications

A similar survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges.

Do these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

21

Setting the hypotheses

The *parameter of interest* is the average number of schools applied to by all Duke students.

There may be two explanations why our sample mean is higher than the recommended 8 schools.

- The true population mean is different.
- The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability

We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended): $H_0 : \mu = 8$

We test the claim that the average number of colleges Duke students apply to is greater than 8: $H_A : \mu > 8$

22

Number of college applications – conditions: sampling distribution of $\hat{p}$

Which of the following is *not* a condition that needs to be met to proceed with this hypothesis test?

a) Students in the sample should be independent of each other with respect to how many colleges they applied to.
b) Sampling should have been done randomly.
c) The sample size should be less than 10% of the population of all Duke students.
d) There should be at least 10 successes and 10 failures in the sample.
e) The distribution of the number of colleges students apply to should not be extremely skewed.

23

Number of college applications – conditions: sampling distribution of $\bar{x}$

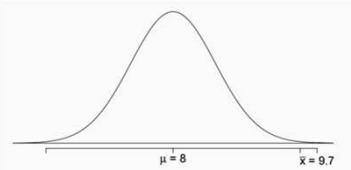
Which of the following is *not* a condition that needs to be met to proceed with this hypothesis test?

a) Students in the sample should be independent of each other with respect to how many colleges they applied to.
b) Sampling should have been done randomly.
c) The sample size should be less than 10% of the population of all Duke students.
d) There should be at least 10 successes and 10 failures in the sample.
e) The distribution of the number of colleges students apply to should not be extremely skewed.

24

Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

Yes, and we can quantify how unusual it is using a p-value.

25

p-values

We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

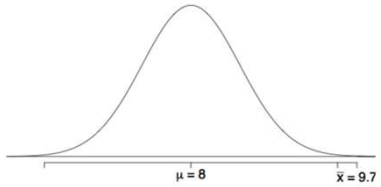
If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.

If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* H_0.

26

Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$

27

Number of college applications - Making a decision

p-value = 0.0003

If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.

This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.

Since p-value is *low* (lower than 5%) we *reject* H_0.

The data provide convincing evidence that Duke students apply to more than 8 schools on average.

The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

28

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

29

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

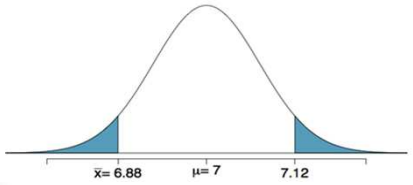
30

Two-sided hypothesis testing with p-values

If the research question was "Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?", the alternative hypothesis would be different

$H_0: \mu = 7$ $H_A: \mu \neq 7$

Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

31

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.

32

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.

33

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 circumstances?

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

34

Decision errors (cont.)

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

In general, which error do we want to make?

Neither! We don't WANT to make errors!

We would like to make the error rate for each type of error small.

In general, guarding against one type of error increases the chance of making the other type of error.

How can we make both error probabilities small?

35

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 if the null hypothesis is true.

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

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

36

Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.

We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

37

Choosing a significance level

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 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.

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. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

38

Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

39

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - H_0: μ = null value
   - H_A: μ < or > or ≠ null value
2. Calculate the point estimate
3. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or $n \geq 30$, no extreme skew -- or use the t distribution (Ch 5)

40

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
2. Calculate the point estimate
3. Check assumptions and conditions
4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

5. Make a decision, and interpret it in context
   - If p-value < α, reject H_0, data provide sufficient evidence for H_A
   - If p-value > α, do not reject H_0, data do not provide sufficient evidence for H_A

41

Inference for Other Estimates

42

Inference for other estimators

The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample proportions is also nearly normal when n is sufficiently large.

An important assumption about point estimates is that they are *unbiased*, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.

That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a "good" estimate. On average, it "hits" what it estimates.

43

Inference for other estimators

The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Some point estimates follow distributions other than the normal distribution, and some scenarios require statistical techniques that we haven't covered yet -- we will discuss these at the end of this section.

44

Confidence intervals for nearly normal point estimates

A confidence interval based on an unbiased and nearly normal point estimate is

$$\textit{point estimate} \pm z^* SE$$

where z^* is selected to correspond to the confidence level, and SE represents the standard error.

Remember that the value $z^* SE$ is called the *margin of error*.

45

Practice

One of the earliest examples of behavioral asymmetry is a preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first 6 months after birth. This is thought to influence subsequent development of perceptual and motor preferences. A study of 124 couples found that 64.5% turned their heads to the right when kissing. The standard error associated with this estimate is roughly 4%. Which of the below is *false*?

a) The 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 4%.

b) A higher sample size would yield a lower standard error.

c) The margin of error for a 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 8%.

d) The 99.7% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 12%.

Güntürkün, O (2003) Adult persistence of head-turning asymmetry. *Nature*. Vol 421.

46

Practice

One of the earliest examples of behavioral asymmetry is a preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first 6 months after birth. This is thought to influence subsequent development of perceptual and motor preferences. A study of 124 couples found that 64.5% turned their heads to the right when kissing. The standard error associated with this estimate is roughly 4%. Which of the below is *false*?

a) The 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 4%.

b) A higher sample size would yield a lower standard error.

c) The margin of error for a 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 8%.

d) The 99.7% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 12%.

Güntürkün, O (2003) Adult persistence of head-turning asymmetry. *Nature*. Vol 421.

47

Hypothesis testing for nearly normal point estimates

The third National Health and Nutrition Examination Survey collected body fat percentage (BF%) and gender data from 13,601 subjects ages 20 to 80. The average BF% for the 6,580 men in the sample was 23.9, and this value was 35.0 for the 7,021 women. The standard error for the difference between the average men and women BF%s was 0.114. Do these data provide convincing evidence that men and women have different average BF%s. You may assume that the distribution of the point estimate is nearly normal.

1. Set hypotheses
2. Calculate point estimate
3. Check conditions
4. Draw sampling distribution, shade p-value
5. Calculate test statistics and p-value, make a decision

48

Hypothesis testing for nearly normal point estimates (cont.)

1. The null hypothesis is that men and women have equal average BF%, and the alternative is that these values are different.

$H_0: \mu_{men} = \mu_{women}$ $H_A: \mu_{men} \neq \mu_{women}$

2. The parameter of interest is the average difference in the population means of BF%s for men and women, and the point estimate for this parameter is the difference between the two sample means:

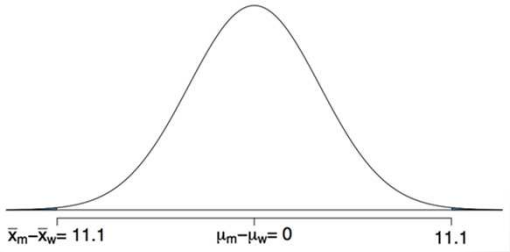
$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35.0 = -11.1$

3. We are assuming that the distribution of the point estimate is nearly normal (we will discuss details for checking this condition in the next chapter, however given the large sample sizes, the normality assumption doesn't seem unwarranted).

49

Hypothesis testing for nearly normal point estimates (cont.)

4. The sampling distribution will be centered at the null value ($\mu_{men} - \mu_{women} = 0$), and the p-value is the area beyond the observed difference in sample means in both tails (lower than -11.1 and higher than 11.1).



50

Hypothesis testing for nearly normal point estimates (cont.)

5. The test statistic is computed as the difference between the point estimate and the null value (-11.1 - 0 = -11.1), scaled by the standard error.

$$Z = \frac{11.1 - 0}{0.114} = 97.36$$

The Z score is huge! And hence the p-value will be tiny, allowing us to reject H_0 in favor of H_A.

These data provide convincing evidence that the average BF% of men and women are different.

51

Non-normal point estimates

- We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:
  - the sample size is too small for the normal approximation to be valid;
  - the standard error estimate may be poor; or
  - the point estimate tends towards some distribution that is not the normal distribution.
- For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

52

When to retreat

Statistical tools rely on the following two main conditions:

Independence. A random sample from less than 10\% of the population ensures independence of observations. In experiments, this is ensured by random assignment. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.

Sample size and skew. For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.

53

When to retreat

Whenever conditions are not satisfied for a statistical technique:

1. Learn new methods that are appropriate for the data.
2. Consult a statistician.
   - Like ME!
   - Pay him LOTS of Money!
   - Seriously, this should be done when planning your research (BEFORE you get stuck)
3. ~~Ignore the failure of conditions.~~ This last option effectively invalidates any analysis and may discredit novel and interesting findings.

54

Sample Size and Power

55

Finding a sample size for a certain margin of error

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

We know that the critical value associated with the 96% confidence level: $z^* = 2.05$.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

The minimum number of children required to attain the desired margin of error is 85.1. Since we can't sample 0.1 of a child, we must sample at least 86 children (round up, since rounding down to 85 would yield a slightly larger margin of error than desired).

56

Hypothesis testing possibilities

		Decision: fail to reject H_0	Decision: 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)

The power of a test is the probability of correctly rejecting H_0, and the probability of doing so is 1 – β

In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

57

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)

58

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?"

59

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 into two simpler problems

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

Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from N(mean = 132, SE = 25 / √100 = 2.5), i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

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

60

Problem 1

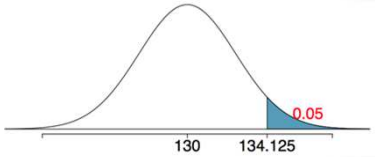
Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember H_0: $\mu = 130$, H_A: $\mu > 130$)

$$P(Z > z) < 0.05 \Rightarrow z > 1.65$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} > 1.65$$

$$\bar{x} > 130 + 1.65 \times 2.5$$

$$\bar{x} > 134.125$$



Any $\bar{x} > 134.125$ would be sufficient to reject H_0 at the 5% significance level.

61

Problem 2

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

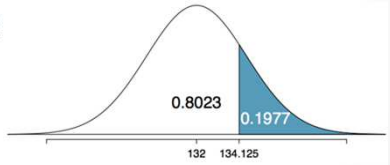
This is the same as finding the area above $\bar{x} = 134.125$ if $\bar{x}$ came from N(132, 2.5).

$$Z = \frac{134.125 - 132}{2.5}$$

$$= 0.85$$

$$P(Z > 0.85) = 1 - 0.8023$$

$$= 0.1977$$



The probability of rejecting H_0: $\mu = 130$, if the true average systolic blood pressure of employees at this company is 132 mmHg, is 0.1977 which is the power of this test.
Therefore, $\beta = 0.8023$ for this test.

62

Putting it all together

http://shiny.stat.tamu.edu:3838/eykolo/power/

63

Achieving desired power

There are several ways to increase power (and hence decrease the 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.

64

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 $\bar{x} \sim N(\mu_{H0} + \delta, SE)$.

65

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.

$\bar{x} > 130 + 1.65 \times 25 / n$

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*.

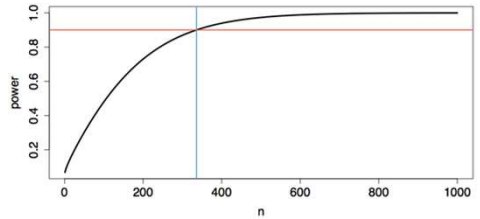
$$P\left(\bar{x} > 130 + 1.65\frac{25}{\sqrt{n}}\right) = 0.9$$

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

66

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.

67

Statistical vs Practical Significance

68

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

a) $n = 100$
b) $n = 10{,}000$

69

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

a) $n = 100$
b) $n = 10{,}000$

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

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad p\text{-value} = 0.0062$$

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

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

70

Test the hypothesis H_0: $\mu = 10$ vs. H_A: $\mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.

71

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

72

Are They Speeding?

College Terrace speed limit: 25 mph

Even after traffic-calming measures, a resident complains that cars still speed.

250 randomly selected cars were clocked with mean speed 25.55 mph, $s = 3.618$.

Is the mean speed of all cars greater than 25 mph?

H_0: $\mu = 25$ H_A: $\mu > 25$

73

Are They Speeding?

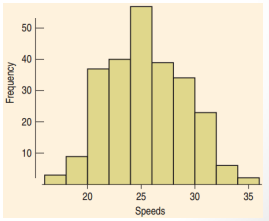
✓Independence Assumption:

The random sample makes the independence assumption reasonable.

The 10% condition is close enough to being met.

✓Nearly Normal Condition

The histogram is unimodal and symmetric. There are no outliers. The sample size is large.



74

Are They Speeding?

$n = 250$, df = 249, $\bar{y} = 25.55$, $s = 3.618$

p-value = 0.0085

A 95% confidence interval is (25.099, 26.001).

75

Are They Speeding?

P-value = 0.0085 is very small.

Reject the null hypothesis and conclude that the mean speed is greater than 25 mph.

This is statistically significant but is it practically significant?

Is 25.55 mph noticeably faster than 25 mph?

Even at the high end of the CI, 26 mph, should the City Council make an effort based on this finding?

76