Lecture 18

Reminder: Inference for other estimators

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 (we'll talk about this in detail soon).

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 (we'll talk about this in detail in two weeks).

- 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.

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 (we'll talk about this in detail in two weeks).

- 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.

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

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 (we'll talk about this in detail in two weeks).

- 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.

    - 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 later in the course, and in the next course

Confidence intervals for nearly normal point estimates

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

$$\text{point estimate} \pm z^{\star} SE$$

where $z^{\star}$ is selected to correspond to the confidence level, and SE represents the standard error. (We've seen this before!)

Remember that the value $z^{\star} SE$ is called the **margin of error**.

Hypothesis testing for nearly normal point estimates

The Procedure

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

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.

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.

The *t* Distribution

Review: Purpose of Large Sample

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that:

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error (SE) , as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size as long as the population distribution is nearly normal.

While this is a helpful special case, it's inherently difficult to verify normality in small data sets.

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.

For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The t Distribution

When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.

This distribution also has a bell shape, but its tails are thicker than the normal model's.

Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)

Confidence intervals for a small-sample mean

Confidence intervals are always of the form

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

ME is always calculated as the product of a critical value (e.g., z^*) and SE.

Since small-sample means follow a t distribution (and not a z distribution), the critical value is a t^* (as opposed to a z^*).

$$\text{point estimate} \pm t^* \times \text{SE}$$

Recap: Inference using a small sample mean

If $n < 30$, sample means follow a t distribution with $\text{SE} = \frac{s}{\sqrt{n}}$.

1. **Conditions**:
   - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
   - $n < 30$ and no extreme skew (often phrased as ``follows a normal distribution'')

2. **Hypothesis Testing**:
$$t_{\text{df}} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1.$$
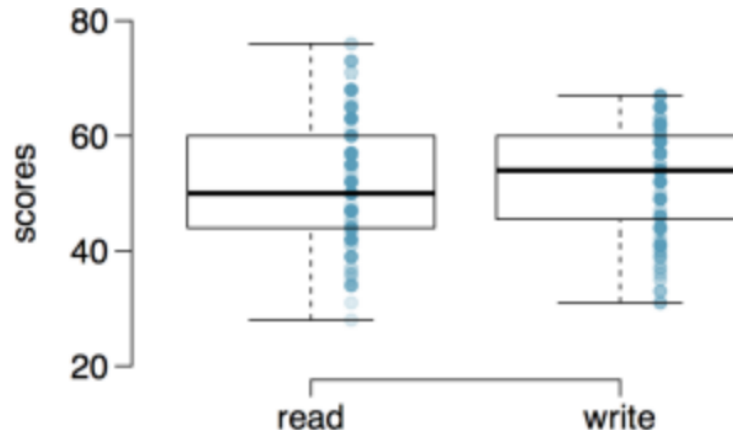
3. **Confidence Interval**:
$$\text{point estimate} \pm t^*_{df} \times SE.$$

Paired Data

Paired Observations

200 observations were randomly sampled from the *High School and Beyond* survey. The same students took a reading and writing test and their scores are shown below. At a first glance, does there appear to be a difference between the average reading and writing test score?

Paired Observations

The same students took a reading and writing test and their scores are shown below. Are the reading and writing scores of each student independent of each other?

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

1. Yes (b) No

Paired Observations

The same students took a reading and writing test and their scores are shown below. Are the reading and writing scores of each student independent of each other?

|	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

1. Yes *(b) No*

Analyzing Paired Data

When two sets of observations have this special correspondence (not independent), they are said to be **paired**.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

$$\text{diff} = \text{read - write}$$

It is important that we always subtract using a consistent order, from student to student or patient to patient, or whatever our operational unit is.

id	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
⋮	⋮	⋮	⋮	⋮
200	137	63	65	-2

Differences in scores (read – write)
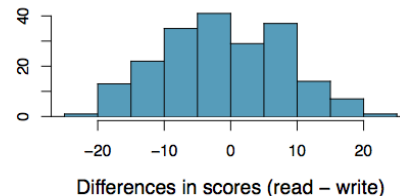
Parameter and Point Estimate

Parameter of interest: Average difference between the reading and writing scores of **all** high school students.

$$\mu_{\text{diff}}$$

Point estimate: Average difference between the reading and writing scores of **sampled** high school students.

$$\bar{x}_{\text{diff}}$$

Setting the Hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

$$0$$

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

$H_0 :$ There is no difference between the average reading and writing score, $\mu_{\text{diff}} = 0$

versus

$H_A :$ There is a difference between the average reading and writing score, $\mu_{\text{diff}} \neq 0$

Nothing new here

The analysis is no different than what we have done before.

We have data from one sample: differences.

We are testing to see if the average difference is different than 0.

Checking assumptions and conditions

Which of the following is true?

- Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another.

- The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

- In order for differences to be random we should have sampled with replacement.

- Since students are sampled randomly and are less than 10% of all students, we can assume that the sampling distribution of the average difference will be nearly normal.

Checking assumptions and conditions

Which of the following is true?

- *Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another.*

- The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

- In order for differences to be random we should have sampled with replacement.

- Since students are sampled randomly and are less than 10% of all students, we can assume that the sampling distribution of the average difference will be nearly normal.

Calculating the test statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.
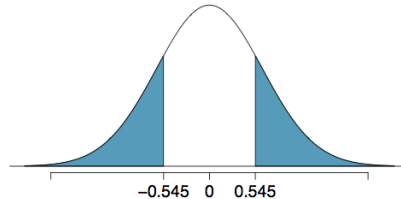
$$Z = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.626}$$

$$= -0.87$$

with p-value of $0.1949 \times 2 = 0.3898$.

Conclusion

How did we get this p-value?

```
2 * pt(-0.87, df = 199, lower.tail = TRUE)
```

```
## [1] 0.3853486
```

Since p-value > 0.05, we fail to reject, and the data do not provide convincing evidence of a difference between the average reading and writing scores.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

- Probability that the average scores on the reading and writing exams are equal.

- Probability that the average scores on the reading and writing exams are different.

- Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.

- Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

· Probability that the average scores on the reading and writing exams are equal.

· Probability that the average scores on the reading and writing exams are different.

· *Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.*

· Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true.

Hypothesis Test ⇔ Confidence Interval

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

· yes

· no

· cannot tell from the information given

Hypothesis Test $\Leftrightarrow$ Confidence Interval

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

- *yes*
- no
- cannot tell from the information given

$$-0.545 \pm 1.96 \frac{8.887}{\sqrt{200}} = -0.545 \pm 1.96 \times 0.626 = (-1.775, 0.685)$$

General Hypothesis Test Conclusion

There are actually three separate ways we can **finish** a NHST scenario.

- compute the p-value, compare it to α
- compute a critical value of a statistic, compare it to the test statistic
- compute the confidence interval, check if it contains the null hypothesis

Any one of these is sufficient: they will **always** agree.

Note: we did not cover the second of these two methods, only the first and third.

Practice

WeBWorK

We're now going to finish the three additional problems from WeBWorK **2018-PracticeProblems8b**, and then if we have time, we'll also solve some problems from **2018-PracticeProblems9b**.