Lecture 11 - Hypothesis Tests for a Mean

Sta102/BME102

June 1, 2016

Colin Rundel & Mine Çetinkaya-Rundel

Recap

Recap - Null Value Hypothesis Testing

Regardless of the sample statistic of interest, all null value hypothesis testing takes exactly the same form:

1. Define the null and alternative hypotheses
2. Check assumptions and conditions
3. Calculate the appropriate test statistic and use that to find the p-value
4. Make a decision, and interpret it in context of the research question

Recap - Null Value Hypothesis Testing - Single Mean

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

Recap - Null Value Hypothesis Testing - Single Mean

1. Set the hypotheses
   - $H_0 : \mu = \textit{null value}$
   - $H_A : \mu <$ or $>$ or $\neq \textit{null value}$
2. Check assumptions and conditions
3. Calculate the appropriate test statistic and use that to find the p-value
4. Make a decision, and interpret it in context of the research question

Recap - Null Value Hypothesis Testing - Single Mean

1. Set the hypotheses
   - $H_0 : \mu = \textit{null value}$
   - $H_A : \mu <$ or $>$ or $\neq$ *null value*
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality/Sample size: nearly normal population or n large enough, w/ no extreme skew or tail weirdness
3. Calculate the appropriate test statistic and use that to find the p-value
4. Make a decision, and interpret it in context of the research question

Recap - Null Value Hypothesis Testing - Single Mean

1. Set the hypotheses
   - $H_0 : \mu = \textit{null value}$
   - $H_A : \mu <$ or $>$ or $\neq$ *null value*
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality/Sample size: nearly normal population or n large enough, w/ no extreme skew or tail weirdness
3. Calculate the appropriate test statistic and use that to find the p-value

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

4. Make a decision, and interpret it in context of the research question

Recap - Null Value Hypothesis Testing - Single Mean

1. Set the hypotheses
    - $H_0 : \mu = \textit{null value}$
    - $H_A : \mu <$ or $>$ or $\neq \textit{null value}$
2. Check assumptions and conditions
    - Independence: random sample/assignment, 10% condition when sampling without replacement
    - Normality/Sample size: nearly normal population or n large enough, w/ no extreme skew or tail weirdness
3. Calculate the appropriate test statistic and use that to find the p-value

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

4. Make a decision, and interpret it in context of the research question
    - If p-value $< \alpha$, reject H_0
    - If p-value $> \alpha$, do not reject H_0

Recap - Confidence Interval - Single Mean

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n - 1$ when the CLT holds.

Recap - Confidence Interval - Single Mean

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n - 1$ when the CLT holds.

Conditions (same as NVHT/CLT):

- Independence: random sample/assignment, 10% condition when sampling without replacement
- Normality/Sample size: nearly normal population or n large enough, w/ no extreme skew or tail weirdness

Recap - Confidence Interval - Single Mean

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n - 1$ when the CLT holds.

Conditions (same as NVHT/CLT):

- Independence: random sample/assignment, 10% condition when sampling without replacement
- Normality/Sample size: nearly normal population or n large enough, w/ no extreme skew or tail weirdness

Confidence interval:

$$\bar{X} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}, \text{ where } df = n - 1$$

Statistical vs. Practical Significance

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 $n = 10,000$?

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 $n = 10,000$?

$$T_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.007$$

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

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 - $SE \downarrow$, $Z \uparrow$, 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 $n = 10,000$?

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.

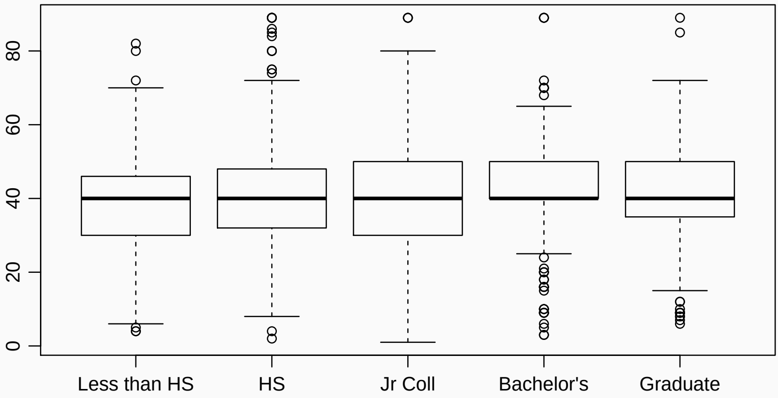
Hypothesis Tests for the difference of two means

Example - GSS

The General Social Survey (GSS) is an annual Census Bureau survey covering demographic, behavioral, and attitudinal questions. To facilitate time-trend studies many of the questions have not changed since 1972. Below is an excerpt from the 2010 survey. The variables are number of hours worked per week and highest educational attainment.

	degree	hrs1
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
⋮		
1172	HIGH SCHOOL	40

Exploratory analysis



What can we say about the relationship between educational attainment and hours worked per week?

Collapsing levels

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.

Collapsing levels

- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- We can combine the levels of education into:
    - `hs or lower` $\leftarrow$ less than high school or high school
    - `coll or higher` $\leftarrow$ junior college, bachelor's, and graduate

Collapsing levels

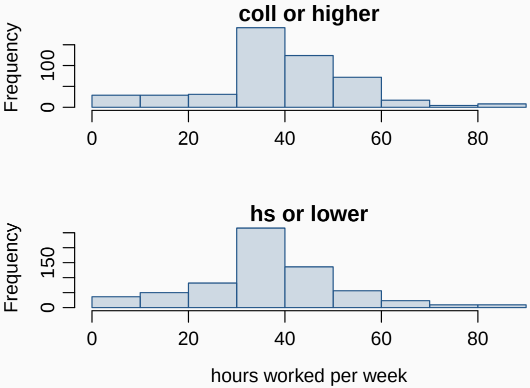
- Say we are only interested the difference between the number of hours worked per week by college and non-college graduates.
- We can combine the levels of education into:
    - `hs or lower` ← less than high school or high school
    - `coll or higher` ← junior college, bachelor's, and graduate
- Here is how you can do this in R:

```
# create a new empty variable
gss$edu = NA

# if statements to determine levels of new variable
gss$edu[gss$degree == "LESS THAN HIGH SCHOOL" ♦
        gss$degree == "HIGH SCHOOL"] = "hs or lower"
gss$edu[gss$degree == "JUNIOR COLLEGE" ♦
        gss$degree == "BACHELOR" ♦
        gss$degree == "GRADUATE"] = "coll or higher"
```

Exploratory analysis - another look

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667



Parameter and point estimate

We want to be able to make useful statments the difference between average hours worked per week by Americans with and without a college degree. What is the parameter of interest and its point estimate?

Parameter and point estimate

We want to be able to make useful statments the difference between average hours worked per week by Americans with and without a college degree. What is the parameter of interest and its point estimate?

- *Parameter of interest:* Average difference between the number of hours worked per week by *all* Americans with a college degree and those with a high school degree or lower.

$$\mu_C - \mu_{hs}$$

Parameter and point estimate

We want to be able to make useful statments the difference between average hours worked per week by Americans with and without a college degree. What is the parameter of interest and its point estimate?

- *Parameter of interest:* Average difference between the number of hours worked per week by *all* Americans with a college degree and those with a high school degree or lower.

$$\mu_c - \mu_{hs}$$

- *Point estimate:* Average difference between the number of hours worked per week by *sampled* Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c - \bar{x}_{hs}$$

Difference of Means and the CLT

We can think about our observations as being samples from two distributions A and B,

$$X_1,\ X_2,\ \ldots,\ X_m \sim A$$
$$Y_1,\ Y_2,\ \ldots,\ Y_n \sim B.$$

Difference of Means and the CLT

We can think about our observations as being samples from two distributions A and B,

$$X_1,\ X_2,\ \ldots,\ X_m \sim A$$
$$Y_1,\ Y_2,\ \ldots,\ Y_n \sim B.$$

We now want to know what the distribution of $\bar{x} - \bar{y}$ will be so that we can perform inference.

Difference of Means and the CLT

We can think about our observations as being samples from two distributions A and B,

$$X_1,\ X_2,\ \ldots,\ X_m \sim A$$
$$Y_1,\ Y_2,\ \ldots,\ Y_n \sim B.$$

We now want to know what the distribution of $\bar{x} - \bar{y}$ will be so that we can perform inference.

From our work with a single sample means, we know that (from the CLT)

$$\bar{x} \sim N(\mu = E(A),\ \sigma^2 = Var(A)/m),$$
$$\bar{y} \sim N(\mu = E(B),\ \sigma^2 = Var(B)/n)$$

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

This proposition then tells us that

$$\bar{x} - \bar{y} \sim N\left(E(\bar{x} - \bar{y}),\ Var(\bar{x} - \bar{y})\right),$$

where

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

This proposition then tells us that

$$\bar{x} - \bar{y} \sim N\left(E(\bar{x} - \bar{y}),\ Var(\bar{x} - \bar{y})\right),$$

where

$$E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y$$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

This proposition then tells us that

$$\bar{x} - \bar{y} \sim N\left(E(\bar{x} - \bar{y}),\ Var(\bar{x} - \bar{y})\right),$$

where

$$E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y$$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

Did I make any assumptions here?

Difference of Means and the CLT (cont.)

Proposition - the sum or difference of normal RVs is also normally distributed. (Not terribly hard to prove, but requires more probability theory than we've covered).

This proposition then tells us that

$$\bar{x} - \bar{y} \sim N\left(E(\bar{x} - \bar{y}),\ Var(\bar{x} - \bar{y})\right),$$

where

$$E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y$$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

Did I make any assumptions here?

Yes - variance result requires that $\bar{x}$ and $\bar{y}$ are independent. We call this independence between groups.

Checking assumptions & conditions

1. *Independence:*
    1.1 *Independence within groups:*
        - Both the college graduates and those with HS degree or lower are sampled randomly.

Checking assumptions & conditions

1. *Independence:*
    1.1 *Independence within groups:*
        - Both the college graduates and those with HS degree or lower are sampled randomly.
        - $505 < 10\%$ of all college graduates and $667 < 10\%$ of all students with a high school degree or lower.

Checking assumptions & conditions

1. *Independence:*
   1.1 *Independence within groups:*
      - Both the college graduates and those with HS degree or lower are sampled randomly.
      - $505 < 10\%$ of all college graduates and $667 < 10\%$ of all students with a high school degree or lower.

      We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.

Checking assumptions & conditions

1. *Independence:*
   1.1 *Independence within groups:*
      - Both the college graduates and those with HS degree or lower are sampled randomly.
      - $505 < 10\%$ of all college graduates and $667 < 10\%$ of all students with a high school degree or lower.

   We can assume that the number of hours worked per week by one college graduate in the sample is independent of another, and the number of hours worked per week by someone with a HS degree or lower in the sample is independent of another as well.

   1.2 *Independence between groups:*
   Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower.

Checking assumptions & conditions

2. *Sample size / Nearly Normal:*
   Both distributions look reasonably symmetric, and the sample sizes are large.

Therefore we can reasonably conclude that the sampling distribution of average number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Additionally, we then also can conclude that the sampling distribution of the difference of the averages will also be nearly normal.

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

- Always, $\textit{ME} = \textit{critical value} \times \textit{SE of point estimate}$

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

- Always, $ME = \textit{critical value} \times \textit{SE of point estimate}$
- In this case the point estimate is $\bar{x} - \bar{y}$

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

- Always, $\textit{ME} = \textit{critical value} \times \textit{SE of point estimate}$
- In this case the point estimate is $\bar{x} - \bar{y}$
- Since the population σ for the difference is unknown, the critical value is $t^{\star}$. We will define $df = \min(n_x - 1,\ n_y - 1)$ which is wrong (but in the conservative direction).

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

- Always, $\textit{ME} = \textit{critical value} \times \textit{SE of point estimate}$
- In this case the point estimate is $\bar{x} - \bar{y}$
- Since the population σ for the difference is unknown, the critical value is $t^\star$. We will define $df = \min(n_x - 1,\ n_y - 1)$ which is wrong (but in the conservative direction).
- So the only new concept is the standard error of the difference between two means...

Confidence interval for difference between two means

- All confidence intervals have the same form:

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

- Always, $ME = \textit{critical value} \times \textit{SE of point estimate}$
- In this case the point estimate is $\bar{x} - \bar{y}$
- Since the population σ for the difference is unknown, the critical value is $t^\star$. We will define $df = \min(n_x - 1,\ n_y - 1)$ which is wrong (but in the conservative direction).
- So the only new concept is the standard error of the difference between two means...

$$SE = \sqrt{Var(\bar{x} - \bar{y})} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \approx \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
college or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
college or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE = \sqrt{\frac{s_c^2}{n_c} + \frac{s_{hs}^2}{n_{hs}}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
college or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE = \sqrt{\frac{s_c^2}{n_c} + \frac{s_{hs}^2}{n_{hs}}} = \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}}$$

Let's put things in context

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
college or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$SE = \sqrt{\frac{s_c^2}{n_c} + \frac{s_{hs}^2}{n_{hs}}} = \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}} = 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE = 0.89$$

$$df = \min(505 - 1,\ 667 - 1) = 504 \qquad t^{\star}_{df=504} = 1.96$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE = 0.89$$

$$df = \min(505 - 1,\ 667 - 1) = 504 \qquad t^\star_{df=504} = 1.96$$

$$\begin{aligned}(\bar{x}_c - \bar{x}_{hs}) \pm t^\star \times SE_{(\bar{x}_c - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 = (0.66, 4.14)\end{aligned}$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_c = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE = 0.89$$

$$df = \min(505 - 1,\ 667 - 1) = 504 \qquad t^\star_{df=504} = 1.96$$

$$\begin{aligned}(\bar{x}_c - \bar{x}_{hs}) \pm t^\star \times SE_{(\bar{x}_c - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\ &= 2.4 \pm 1.74 = (0.66, 4.14)\end{aligned}$$

We are 95% confident that college grads work on average between 0.66 and 4.14 more hours per week than those with a HS degree or lower.

Hypothesis Test

If instead we wanted to conduct a hypothesis, what would the hypotheses be for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

Hypothesis Test

If instead we wanted to conduct a hypothesis, what would the hypotheses be for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

H_0: $\mu_c = \mu_{hs}$

There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

Hypothesis Test

If instead we wanted to conduct a hypothesis, what would the hypotheses be for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

H_0: $\mu_c = \mu_{hs}$

There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

H_A: $\mu_c \neq \mu_{hs}$

There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Hypothesis Test

If instead we wanted to conduct a hypothesis, what would the hypotheses be for testing if there is a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower?

H_0: $\mu_c = \mu_{hs} \rightarrow \mu_c - \mu_{hs} = 0$
There is no difference in the average number of hours worked per week by college graduates and those with a HS degree or lower. Any observed difference between the sample means is due to natural sampling variation (chance).

H_A: $\mu_c \neq \mu_{hs} \rightarrow \mu_c - \mu_{hs} \neq 0$
There is a difference in the average number of hours worked per week by college graduates and those with a HS degree or lower.

Calculating the test-statistic and the p-value

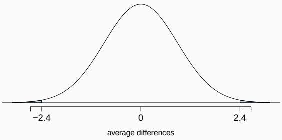
H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4, \; SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

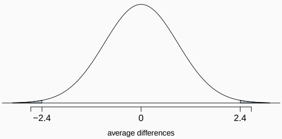
$$\bar{x}_c - \bar{x}_{hs} = 2.4, SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$



Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4,\ SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$

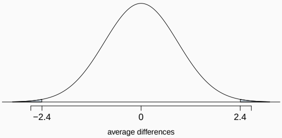


$$T = \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE}$$

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4, \; SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$

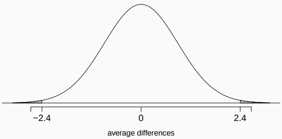


$$\begin{aligned} T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4 - 0}{0.89} = 2.70 \end{aligned}$$

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4,\ SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$

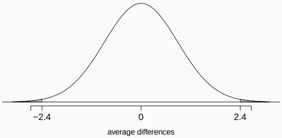


$$\begin{aligned} T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4 - 0}{0.89} = 2.70 \\ P(T > 2.70) &= 1 - 0.9965 = 0.0035 \end{aligned}$$

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4,\ SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$

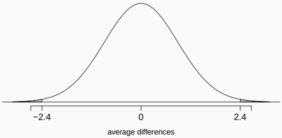


$$\begin{aligned} T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4 - 0}{0.89} = 2.70 \\ P(T > 2.70) &= 1 - 0.9965 = 0.0035 \\ p - value &= 2 \times P(T > 2.70) = 0.007 \end{aligned}$$

Calculating the test-statistic and the p-value

H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

$$\bar{x}_c - \bar{x}_{hs} = 2.4, \; SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$



$$
\begin{aligned}
T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\
&= \frac{2.4 - 0}{0.89} = 2.70 \\
P(T > 2.70) &= 1 - 0.9965 = 0.0035 \\
p - value &= 2 \times P(T > 2.70) = 0.007
\end{aligned}
$$

Reject H_0 - the data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

Inference using difference of two means

- Conditions:
    - independence within groups (random sample / $n < 10\%$ of population if sampling w/o replacement)
    - independence between groups
    - Sample sizes (n_1 and n_2) large enough relative to skew and or thick/thin tails in each sample.

Inference using difference of two means

- Conditions:
  - independence within groups (random sample / $n < 10\%$ of population if sampling w/o replacement)
  - independence between groups
  - Sample sizes (n_1 and n_2) large enough relative to skew and or thick/thin tails in each sample.
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(\mu_1 - \mu_2) - (\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Inference using difference of two means

- Conditions:
  - independence within groups (random sample / $n < 10\%$ of population if sampling w/o replacement)
  - independence between groups
  - Sample sizes (n_1 and n_2) large enough relative to skew and or thick/thin tails in each sample.
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(\mu_1 - \mu_2) - (\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

- Confidence interval:

$$CI = \text{point estimate} \pm CV \times SE = (\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

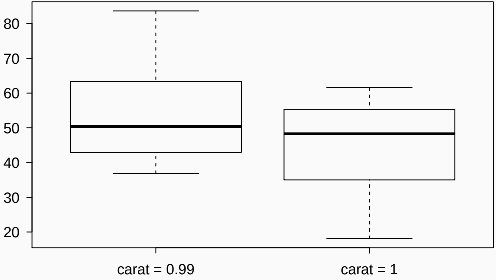
$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \approx \min(n_1 - 1, n_2 - 1)$$

Diamond Example

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.

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

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.

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

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

Hypothesis test

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

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

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?

Hypothesis test

	0.99 carat pt99	*1 carat* 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?

$$T = -2.508 \qquad df = 22$$

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

p-value

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

$$T = -2.508 \qquad df = 22$$

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?

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

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.

Confidence interval

Calculate the interval, and interpret it in context.

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

Confidence interval

Calculate the interval, and interpret it in context.

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

$$(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^\star_{df} \times SE \quad = \quad (44.50 - 53.43) \pm 1.72 \times 3.56$$

Confidence interval

Calculate the interval, and interpret it in context.

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

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

Calculate the interval, and interpret it in context.

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

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

Confidence interval

Calculate the interval, and interpret it in context.

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

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

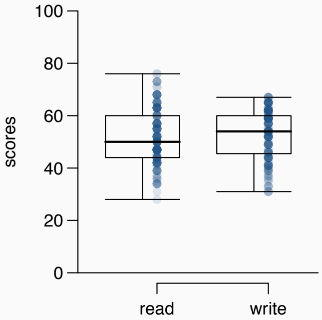
We are 90% confident that the average point price of a 0.99 carat diamond is \$15.05 to \$2.81 lower than the average point price of a 1 carat diamond.

Paired Tests of Two Means

Example - Reading and Writing

200 randomly selected high school 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?

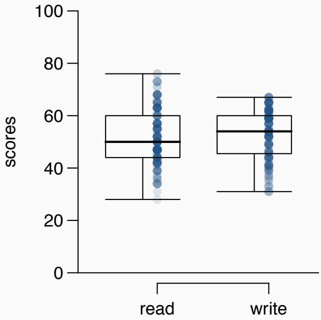
	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65



Example - Reading and Writing

200 randomly selected high school 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?

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



Do you think reading and writing scores are independent?

Analyzing paired data

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

Analyzing paired data

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

To analyze paired data, we will only examine the difference in outcomes of each pair of observations.

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

Analyzing paired data

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

To analyze paired data, we will only examine the difference in outcomes of each pair of observations.

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

	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

Parameter and point estimate

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

$$\mu_{diff}$$

Parameter and point estimate

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

$$\mu_{diff}$$

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

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

Setting the hypotheses

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

Setting the hypotheses

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.

$$\mu_{diff} = 0$$

H_A: There is a difference between the average reading and writing score.

$$\mu_{diff} \neq 0$$

Nothing new here

We have already done this kind of analysis previously.

- We have data from *one* numeric variable - the difference.
- We are testing to see if this variable is or is not equal to 0.

Nothing new here

We have already done this kind of analysis previously.

- We have data from *one* numeric variable - the difference.
- We are testing to see if this variable is or is not equal to 0.

	diff
$\bar{x}$	-0.545
s	8.89
n	200

$$H_0 : \mu_{diff} = 0$$
$$H_A : \mu_{diff} \neq 0$$

Nothing new here

We have already done this kind of analysis previously.

- We have data from *one* numeric variable - the difference.
- We are testing to see if this variable is or is not equal to 0.

	diff
$\bar{x}$	-0.545
s	8.89
n	200

$$H_0: \ \mu_{diff} = 0$$
$$H_A: \ \mu_{diff} \neq 0$$

$$T = \frac{\bar{X} - \mu}{SE} = \frac{-0.545 - 0}{8.89/\sqrt{200}} = -0.877$$

Nothing new here

We have already done this kind of analysis previously.

- We have data from *one* numeric variable - the difference.
- We are testing to see if this variable is or is not equal to 0.

	diff
$\bar{x}$	-0.545
s	8.89
n	200

$$H_0: \ \mu_{diff} = 0$$
$$H_A: \ \mu_{diff} \neq 0$$

$$T = \frac{\bar{X} - \mu}{SE} = \frac{-0.545 - 0}{8.89/\sqrt{200}} = -0.877$$

$$\begin{aligned}\text{p-value} &= P(T < -0.877 \text{ or } T > 0.877) \\ &= 2 \times P(T < -0.877) = 2 \times 0.19 = 0.38\end{aligned}$$

Example - Zinc

Trace metals in drinking water affect the flavor and unusually high concentrations can pose a health hazard. Data were collected by measuring zinc concentration at the bottom and at the surface of 10 randomly sampled wells in Wake country.

We would like to evaluate whether the true average concentration of zinc at the bottom of the well water exceeds that of the surface water. Data are given below.

well	zinc	location
1	0.43	bottom
2	0.266	bottom
3	0.567	bottom
4	0.531	bottom
5	0.707	bottom
6	0.716	bottom
7	0.651	bottom

well	zinc	location
8	0.589	bottom
9	0.469	bottom
10	0.723	bottom
1	0.415	surface
2	0.238	surface
3	0.39	surface
4	0.41	surface

well	zinc	location
5	0.605	surface
6	0.609	surface
7	0.632	surface
8	0.523	surface
9	0.411	surface
10	0.612	surface

Tidying the data

We prefer data where each row represents a *unit of observation* - in this case a well. What does that look like?

Tidying the data

We prefer data where each row represents a *unit of observation* - in this case a well. What does that look like?

well	zinc bottom	zinc top
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

Tidying the data

We prefer data where each row represents a *unit of observation* - in this case a well. What does that look like?

well	zinc bottom	zinc top	diff
1	0.43	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.39	0.177
4	0.531	0.41	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
7	0.651	0.632	0.019
8	0.589	0.523	0.066
9	0.469	0.411	0.058
10	0.723	0.612	0.111

Inference

Lets use a confidence interval to evaluate the difference in zinc concentration between the bottom and top of a well.

$$\bar{x}_{diff} = 0.08, \quad s = 0.052, \quad n = 10$$

Inference

Lets use a confidence interval to evaluate the difference in zinc concentration between the bottom and top of a well.

$$\bar{x}_{diff} = 0.08, \quad s = 0.052, \quad n = 10$$

95% Confidence Interval:

$$PE \pm CV \times SE$$

$$\bar{x}_{diff} \pm t^{\star}_{df=9} \times \frac{s}{\sqrt{n}}$$

$$0.08 \pm 2.26 \times \frac{0.052}{\sqrt{10}}$$

$$(0.043, 0.118)$$