Lecture 13 - Difference of Means

Sta102/BME102

March 9, 2016

Colin Rundel

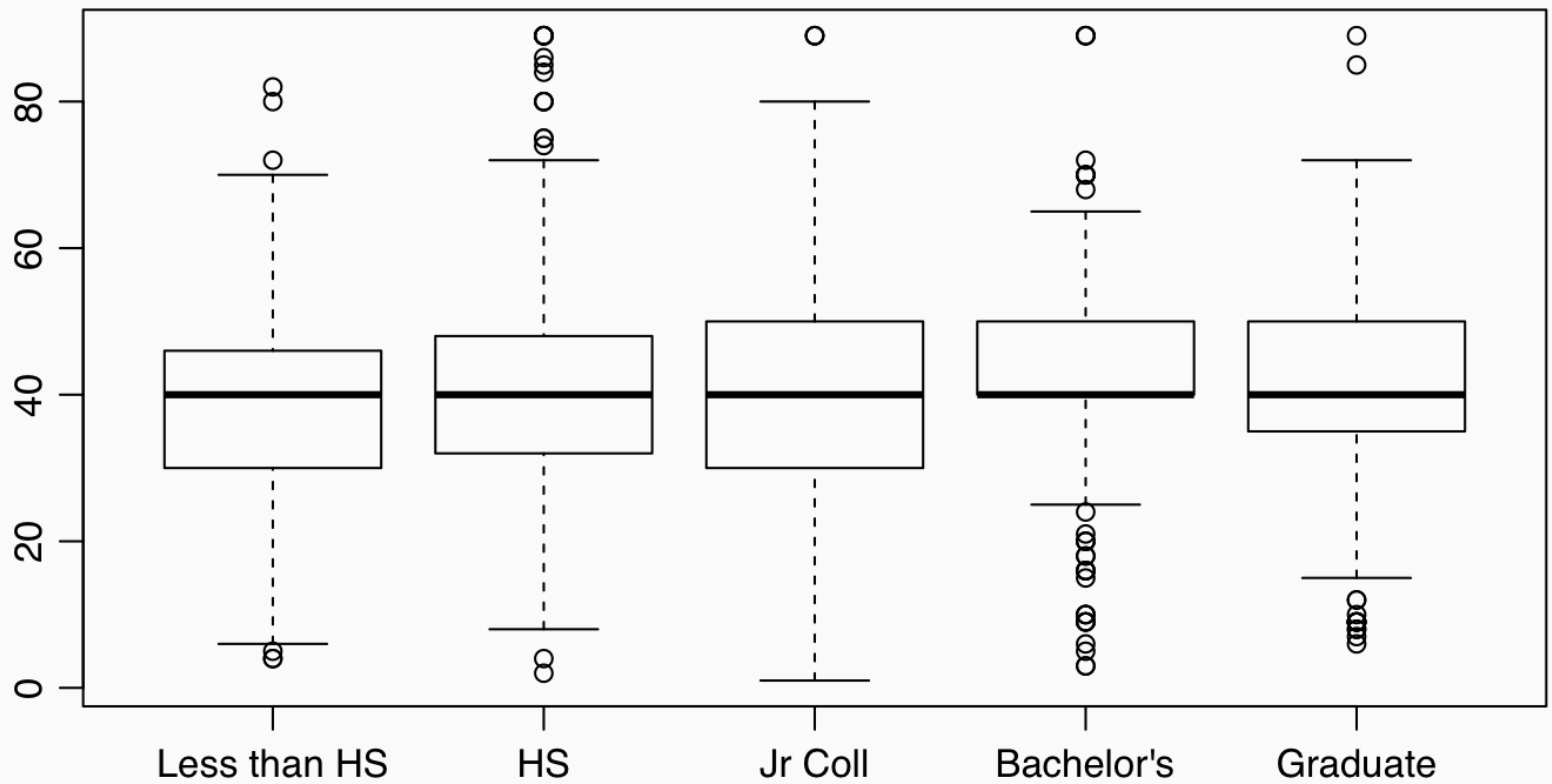
Testing the difference of two means

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` ← less than high school or high school
- `coll or higher` ← junior college, bachelor's, and graduate

Collapsing levels (in R)

Here is how we can collapse levels in R:

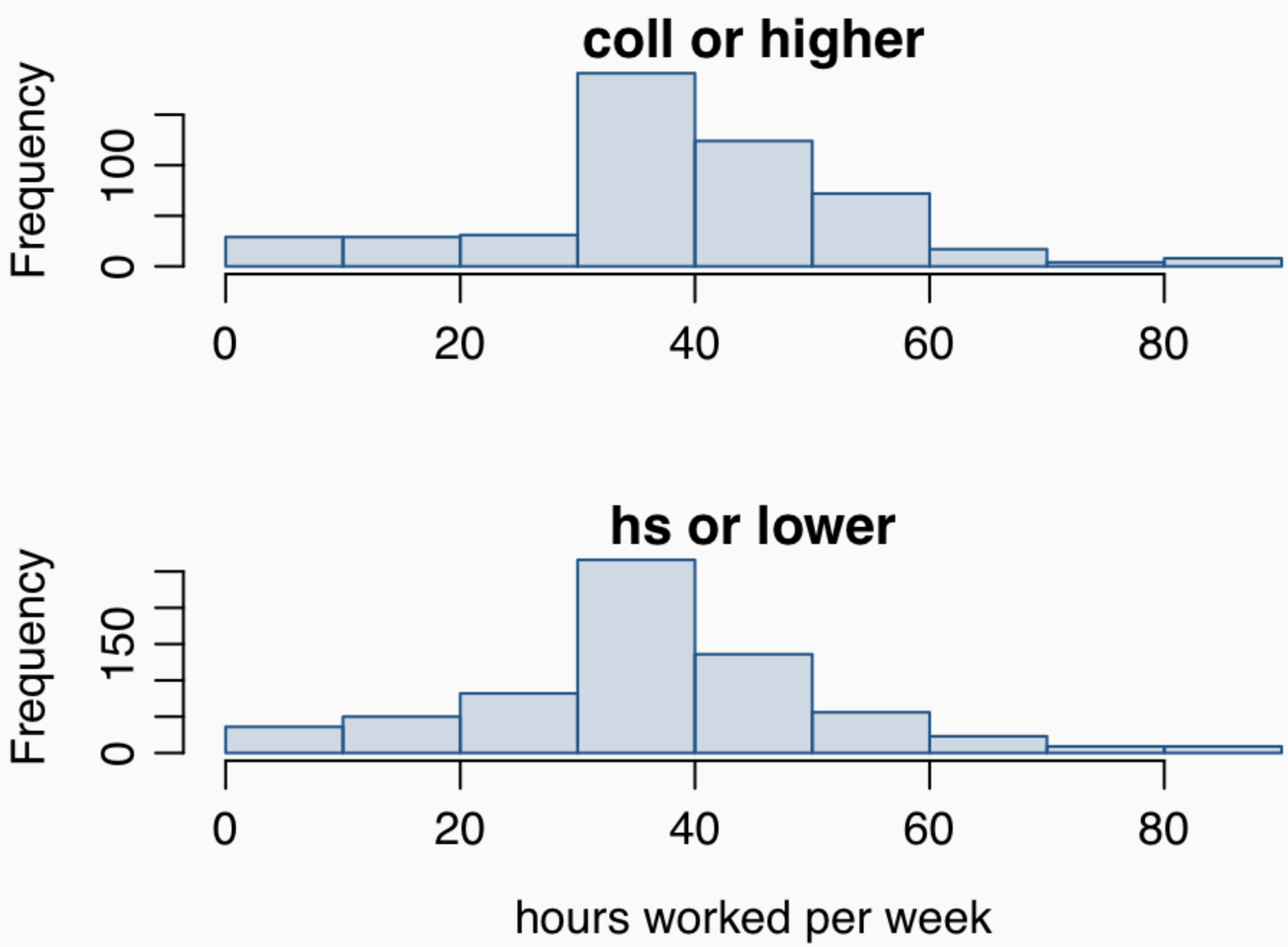
```
# create a new empty variable
gss$edu = NA

# conditional statements to determine levels of new vari-
able
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"

# make sure new variable is categorical
gss$edu = as.factor(gss$edu)
```

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 construct a 95% confidence interval for 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. What are the parameter of interest and the point estimate?

Parameter and point estimate

We want to construct a 95% confidence interval for 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. What are the parameter of interest and the 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 construct a 95% confidence interval for 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. What are the parameter of interest and the 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 D_x and D_y,

$$X_1,\ X_2,\ \ldots,\ X_{n_x} \sim D_x$$
$$Y_1,\ Y_2,\ \ldots,\ Y_{n_y} \sim D_y.$$

Difference of Means and the CLT

We can think about our observations as being samples from two distributions D_x and D_y,

$$X_1,\ X_2,\ \ldots,\ X_{n_x} \sim D_x$$
$$Y_1,\ Y_2,\ \ldots,\ Y_{n_y} \sim D_y.$$

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 D_x and D_y,

$$X_1,\ X_2,\ \ldots,\ X_{n_x} \sim D_x$$
$$Y_1,\ Y_2,\ \ldots,\ Y_{n_y} \sim D_y.$$

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 the CLT tells us that

$$\bar{x} \sim N(E(D_x),\ Var(D_x)/n_x),$$
$$\bar{y} \sim N(E(D_y),\ Var(D_y)/n_y),$$

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).

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 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 tells us that

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

where

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

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

Assumes $\bar{x}$ indep. of $\bar{y}$

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 tells us that

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

where

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

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

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 tells us that

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

where

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

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

Did I make any assumptions here?

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

Checking assumptions & conditions

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

Checking assumptions & conditions

1. *Independence:*
   a. *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:*
   a. *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:*
   a. *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.

   b. *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 / skew:*
   Both distributions look reasonably symmetric, and the sample sizes are large, therefore we can assume that the sampling distribution of number of hours worked per week by college graduates and those with HS degree or lower are nearly normal. Hence the sampling distribution of the average difference will be nearly normal as well.

Confidence interval for difference between two means

All confidence intervals will have the same form:

$$point\ estimate \pm ME$$

$$point\ estimate \pm CV \times SE$$

Confidence interval for difference between two means

All confidence intervals will have the same form:

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

$$\textit{point estimate} \pm \textit{CV} \times \textit{SE}$$

- In this case, the point estimate is $\bar{x} - \bar{y}$

Confidence interval for difference between two means

All confidence intervals will have the same form:

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

$$\textit{point estimate} \pm \textit{CV} \times \textit{SE}$$

- In this case, the point estimate is $\bar{x} - \bar{y}$
- Since the population σs are not known, the critical value will be $T^{\star}$ with

$$df = \min(n_x - 1,\ n_y - 1)^{\star}$$

Confidence interval for difference between two means

All confidence intervals will have the same form:

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

$$\textit{point estimate} \pm CV \times SE$$

- In this case, the point estimate is $\bar{x} - \bar{y}$
- Since the population σs are not known, the critical value will be $T^\star$ with

$$df = \min(n_x - 1,\ n_y - 1)^\star$$

- The standard error is the standard deviation of sampling distribution

Confidence interval for difference between two means

All confidence intervals will have the same form:

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

$$\textit{point estimate} \pm CV \times SE$$

- In this case, the point estimate is $\bar{x} - \bar{y}$
- Since the population σs are not known, the critical value will be $T^\star$ with

$$df = \min(n_x - 1,\ n_y - 1)^\star$$

- The standard error is the standard deviation of sampling distribution

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

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.

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

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

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

PE CV SE

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.

Setting the hypotheses

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?

Setting the hypotheses

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).

Setting the hypotheses

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.

Setting the hypotheses

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

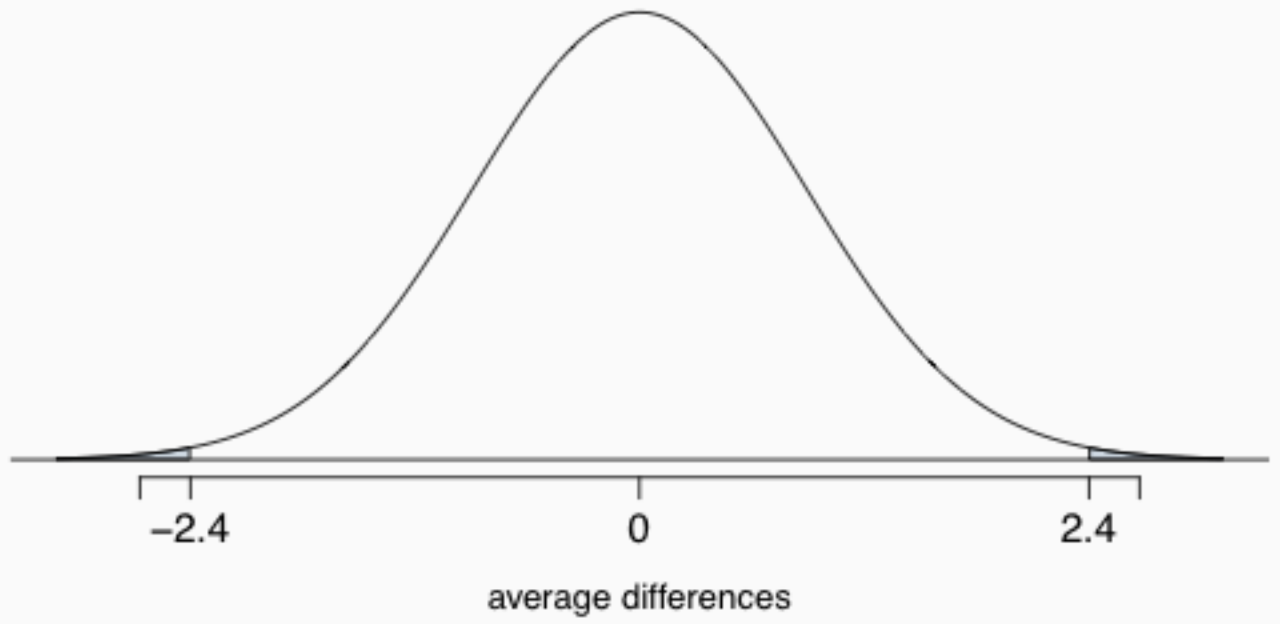
H_0: $\mu_c - \mu_{hs} = 0$
H_A: $\mu_c - \mu_{hs} \neq 0$

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

Calculating the test-statistic and the p-value

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

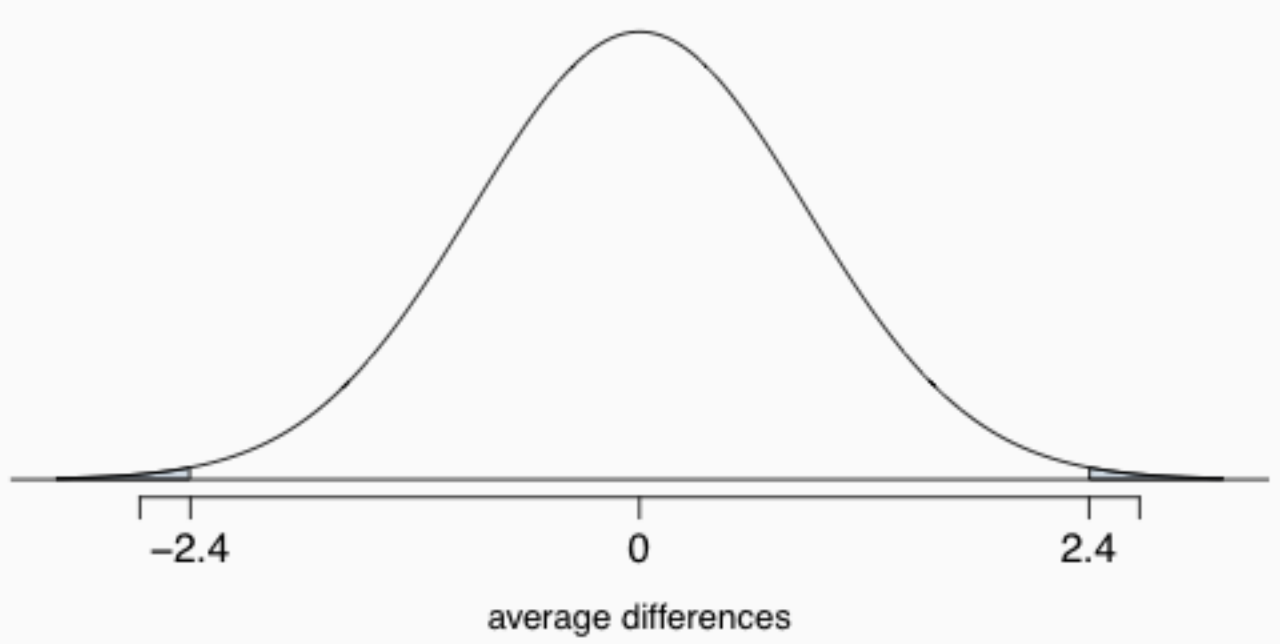
$$\bar{x}_c - \bar{x}_{hs} = 2.4, SE_{\bar{x}_c - \bar{x}_{hs}} = 0.89$$



Calculating the test-statistic and the p-value

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

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

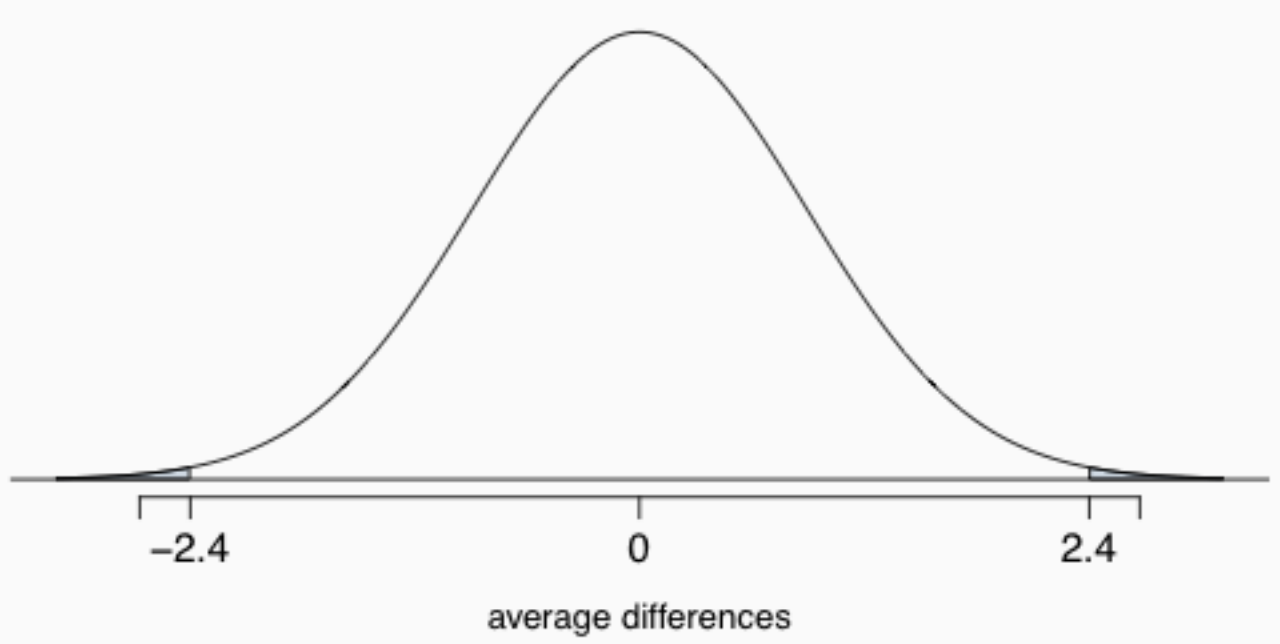


$$T = \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE}$$

Calculating the test-statistic and the p-value

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

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



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

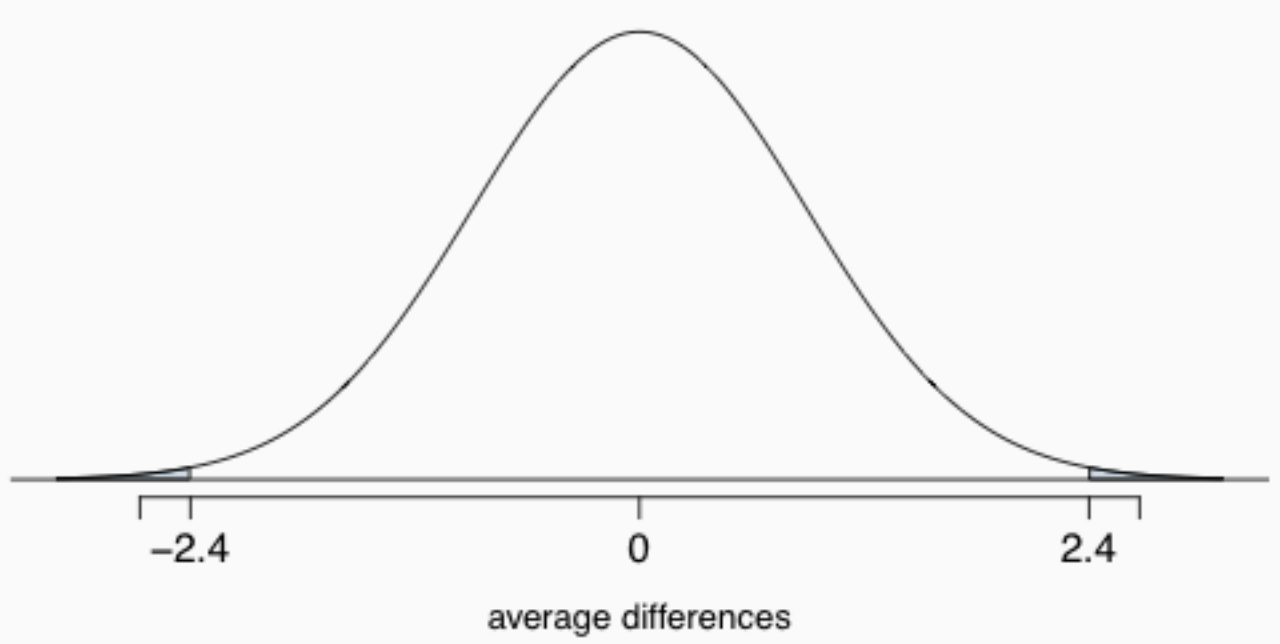
$$\text{Pvalue} = P(T > 2.70 \text{ or } T < -2.70)$$

$$= 2 \times P(T > 2.70)$$

Calculating the test-statistic and the p-value

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

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

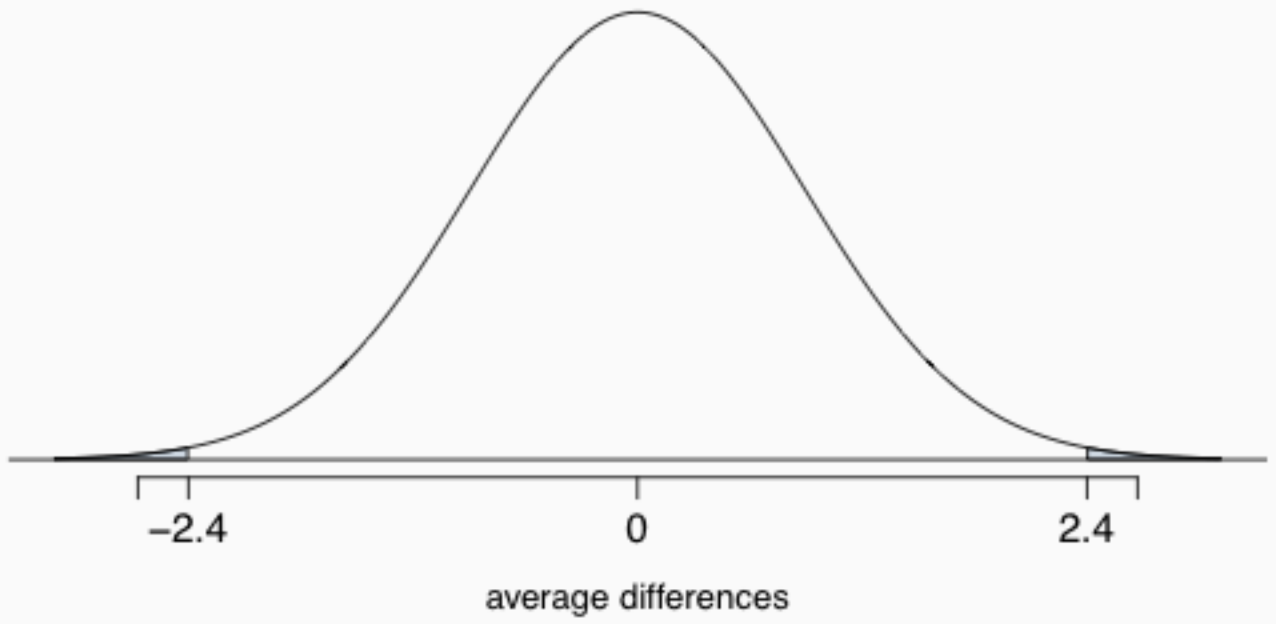


$$\begin{aligned} T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4}{0.89} = 2.70 \\ P(T > 2.70) &= 1 - 0.9965 = 0.0035 \end{aligned}$$

Calculating the test-statistic and the p-value

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

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

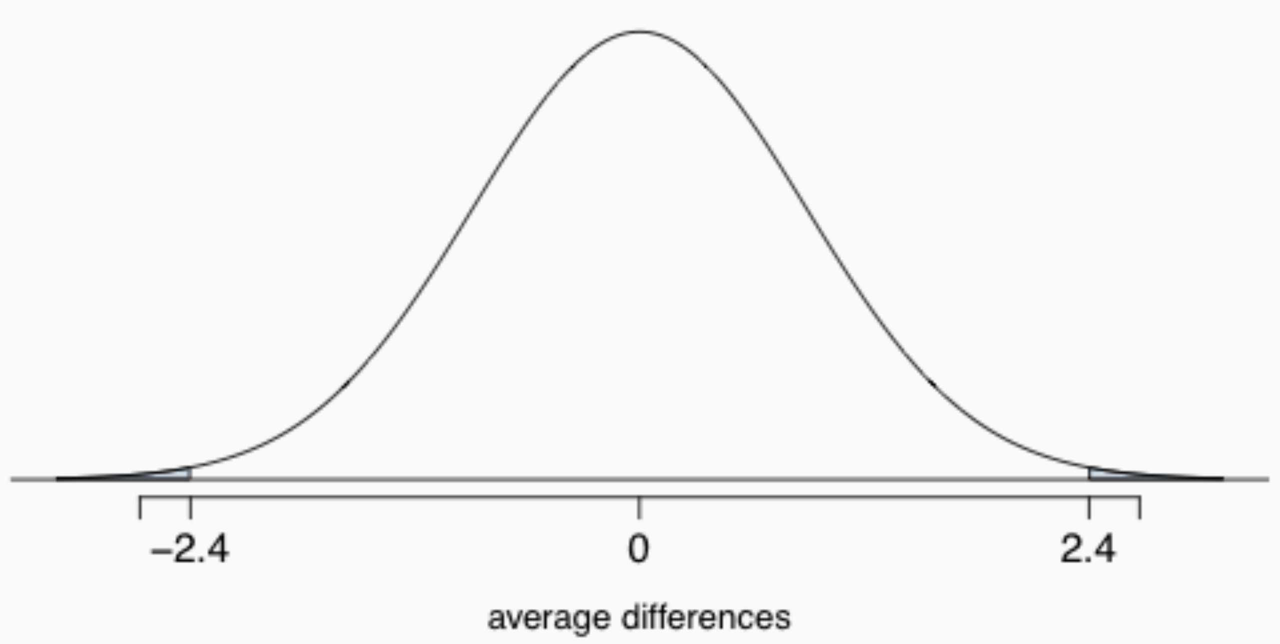


$$\begin{aligned} T &= \frac{(\bar{x}_c - \bar{x}_{hs}) - (\mu_c - \mu_{hs})}{SE} \\ &= \frac{2.4}{0.89} = 2.70 \\ P(T > 2.70) &= 1 - 0.9965 = 0.0035 \\ p - value &= 2 \times P(T > 2.70) = 0.007 \end{aligned}$$

Calculating the test-statistic and the p-value

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

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



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

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

For sufficiently large sample size (of both groups), the distribution of the difference between the sample means has a $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ and follows a T distribution with $df = \min(n_1 - 1,\ n_2 - 1)^*$.

Inference using difference of two means

For sufficiently large sample size (of both groups), the distribution of the difference between the sample means has a $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ and follows a T distribution with $df = \min(n_1 - 1,\ n_2 - 1)^*$.

Conditions:

- independence within groups
- independence between groups
- Sample sizes (n_1 and n_2) large enough relative to skew and or think/thin tails in either sample.

Inference using difference of two means

For sufficiently large sample size (of both groups), the distribution of the difference between the sample means has a $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ and follows a T distribution with $df = \min(n_1 - 1,\ n_2 - 1)^*$.

Conditions:

- independence within groups
- independence between groups
- Sample sizes (n_1 and n_2) large enough relative to skew and or think/thin tails in either sample.

Hypothesis testing:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

Inference using difference of two means

For sufficiently large sample size (of both groups), the distribution of the difference between the sample means has a $SE \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ and follows a T distribution with $df = \min(n_1 - 1,\ n_2 - 1)^*$.

Conditions:

- independence within groups
- independence between groups
- Sample sizes (n_1 and n_2) large enough relative to skew and or think/thin tails in either sample.

Hypothesis testing:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

Confidence interval:

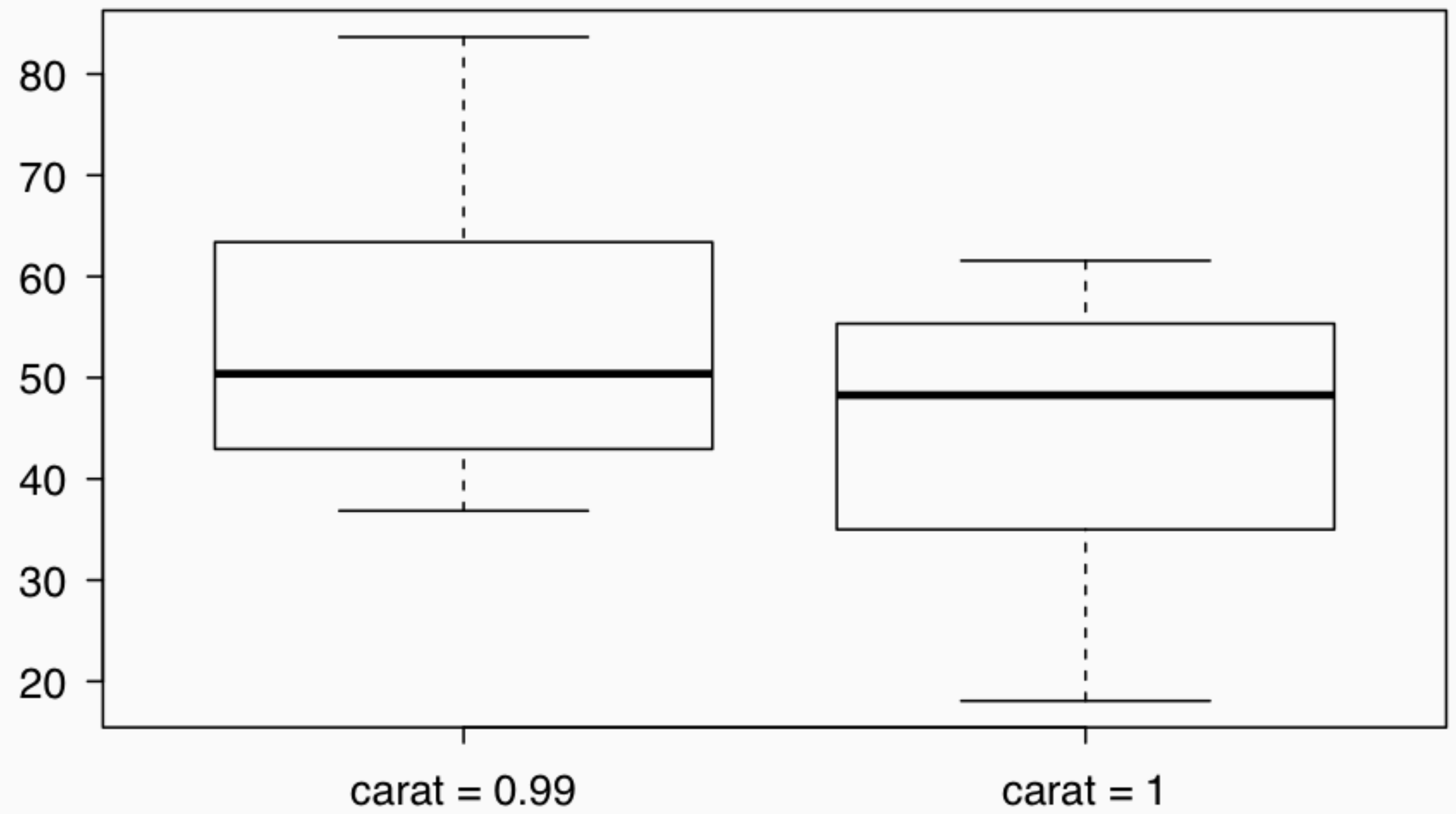
$$\text{point estimate} \pm T^\star \times SE$$

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 pt99	*1 carat* 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 pt99	*1 carat* 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 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?

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?

Synthesis

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

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

Critical value

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

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

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.

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

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

Confidence interval

Calculate the interval, and interpret it in context.

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

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

Confidence interval

Calculate the interval, and interpret it in context.

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

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

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

Power

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

Power

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

Step 0:

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

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

Power

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

Step 0:

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

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

Step 1:

$$P(T > t) < 0.05 \quad \Rightarrow \quad t > 1.72$$

$$P\left(\frac{\bar{x}_{100} - \bar{x}_{99} - 0}{3.56} > 1.72\right) = 0.05$$

$$\bar{x}_{100} - \bar{x}_{99} > 0 + 1.72 \times 3.56$$

$$\bar{x}_{100} - \bar{x}_{99} > 6.12$$

Power

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

Step 0:

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

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

Step 1:

$$P(T > t) < 0.05 \quad \Rightarrow \quad t > 1.72$$

$$P\left(\frac{\bar{x}_{100} - \bar{x}_{99} - 0}{3.56} > 1.72\right) = 0.05$$

$$\bar{x}_{100} - \bar{x}_{99} > 0 + 1.72 \times 3.56$$

$$\bar{x}_{100} - \bar{x}_{99} > 6.12$$

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

$$P(\bar{x}_{100} - \bar{x}_{99} > 6.12 | \mu_{100} - \mu_{99} = 9)$$

$$= P\left(T > \frac{6.12 - 9}{3.56}\right) = P(T > -0.8089)$$

$$= 0.786$$