# Lecture 12 - Power, Tests of Two Means

Sta102 / BME 102

Colin Rundel

February 25th, 2015



## Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

$$H_0 : \mu = 130$$
$$H_A : \mu > 130$$

We'll start with a very specific question – "What is the power of this hypothesis test to correctly detect an increase of 2 mmHg in average blood pressure?"





## Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject $H_0$ when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let's break this down intro two simpler problems:

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject $H_0$?
2. Problem 2: What is the probability that we would reject $H_0$ if $\bar{x}$ had come from a distribution with $\mu = 132$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?





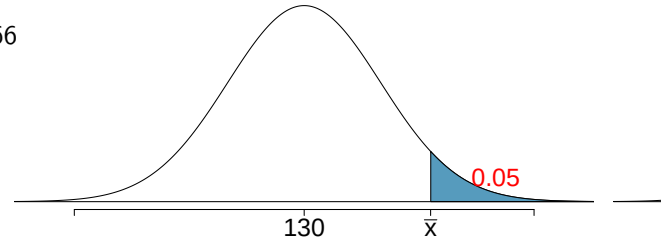
## Problem 1

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

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

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

$$\bar{x} > 130 + 1.66 \times 2.5$$
$$\bar{x} > 134.15$$




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

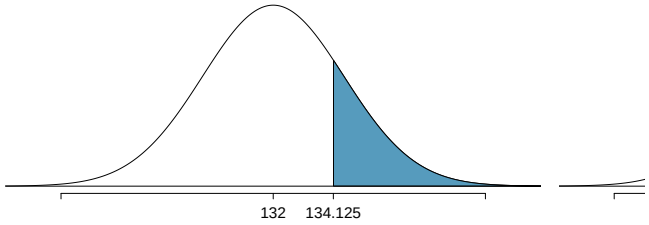
## Problem 2

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

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

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

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

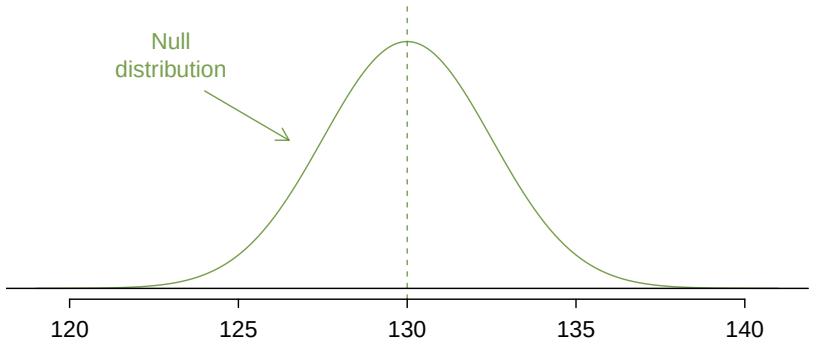



The probability of rejecting $H_0 : \mu = 130$, if the true average systolic blood pressure of employees at this company is 132 mmHg, is 0.199 which is the power of this test.





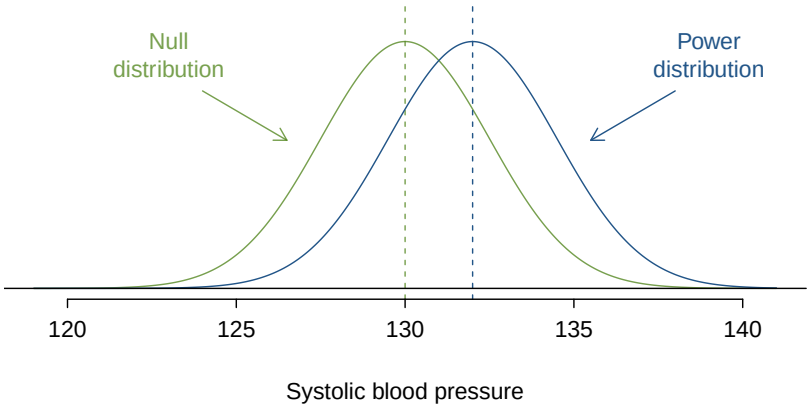
## Putting it all together







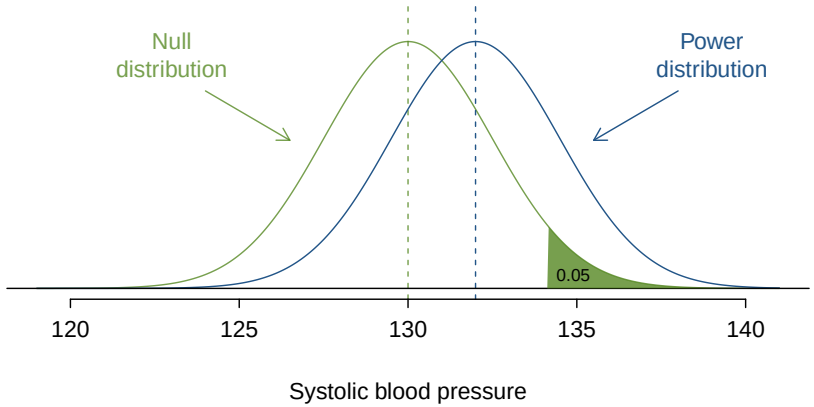

## Putting it all together



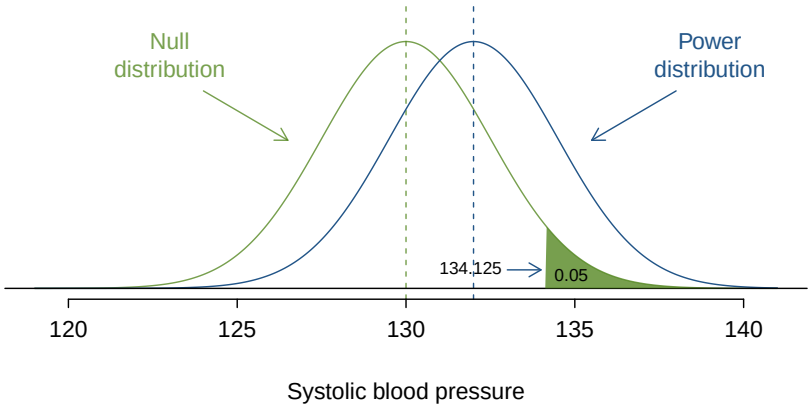





## Putting it all together



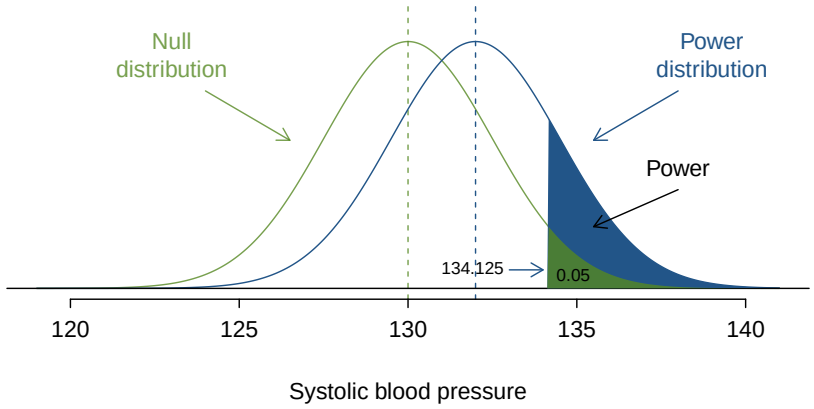
## Putting it all together








## Putting it all together








## Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which is equivalent to increasing the sample size (it will decrease the standard error). With a smaller $s$ we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase $\alpha$, which will make it more likely to reject $H_0$ (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.





## Recap - Calculating Power

- *Step 0:* Pick a meaningful effect size $\delta$ and a significance level $\alpha$
- *Step 1:* Calculate the range of values for the point estimate beyond which you would reject $H_0$ at the chosen $\alpha$ level.
- *Step 2:* Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\mu = \mu_{H_0} + \delta$

## Example - Power for a two sided hypothesis test

Going back to the blood pressure example, what would the power be to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level for a sample of 625 patients?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad n = 625, \quad \sigma = 25, \quad \delta = 4, \quad 1 - \beta = ?$$

Step 1:

$$P(T > t \text{ or } T < -t) < 0.05 \quad \Rightarrow \quad t > 1.96$$
$$\bar{x} > 130 + 1.96\frac{25}{\sqrt{625}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{625}}$$
$$\bar{x} > 131.96 \text{ or } \bar{x} < 128.04$$

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

$$\begin{aligned} P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) &= P(T > [131.96 - 134]/1) + P(T < [128.04 - 134]/1) \\ &= P(T > -2.04) + P(T < -5.96) \\ &= 0.979 + 0 = 0.979 \end{aligned}$$





## Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4, \quad n = ?$$

Step 1:

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

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

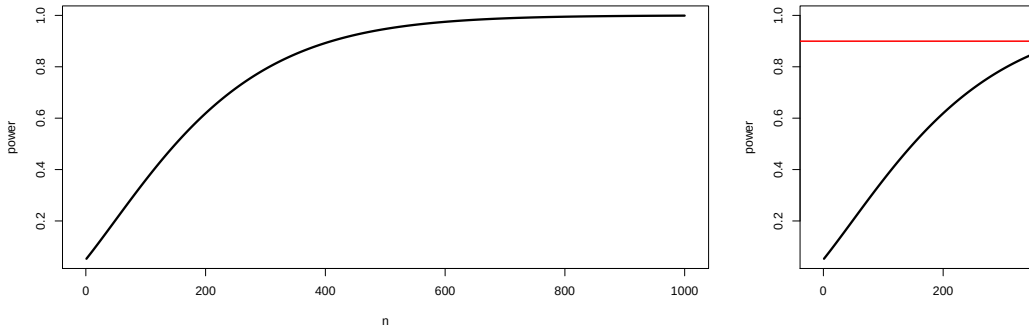
$$P\left(\bar{x} > 130 + 1.96\frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{n}}\right) = 0.9$$
$$P\left(T > 1.96 - 4\frac{\sqrt{n}}{25} \text{ or } T < -1.96 - 4\frac{\sqrt{n}}{25}\right) = 0.9$$





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

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




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



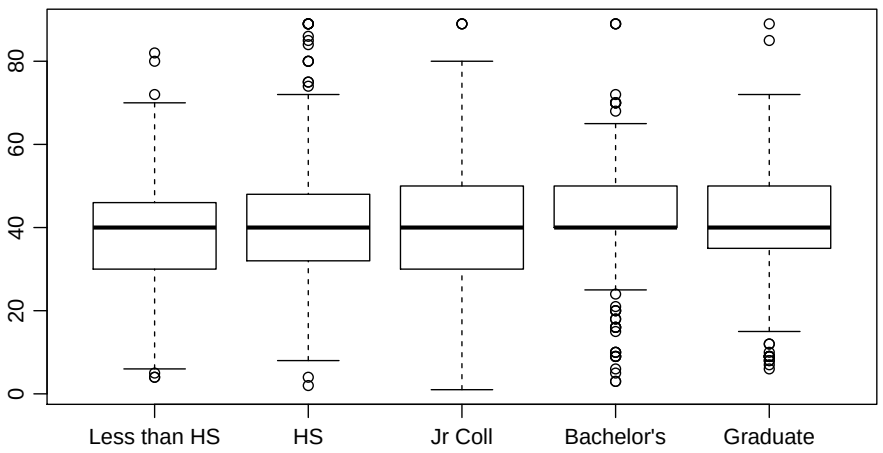


## Example - GSS

The General Social Survey (GSS) conducted by the Census Bureau contains a standard 'core' of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. Below is an excerpt from the 2010 data set. 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.
- 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:

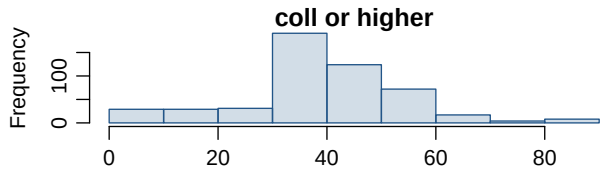
```
# create a new empty variable
gss$edu = NA

# 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"

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

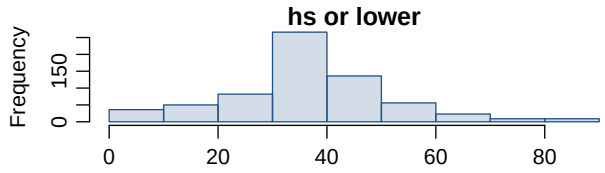




## Exploratory analysis - another look

| | $\bar{x}$ | $s$ | $n$ |
|---|---|---|---|
| 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 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 $D_y$,

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

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 both

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





## 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 then 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

$$\bar{x} - \bar{y} = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i - \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$$





## Expected Value of the Difference

$$\begin{aligned}
E(\bar{x} - \bar{y}) &= E\left(\frac{1}{n_x}\sum_{i=1}^{n_x} x_i - \frac{1}{n_y}\sum_{i=1}^{n_y} y_i\right) \\
&= E\left(\frac{1}{n_x}\sum_{i=1}^{n_x} x_i\right) - E\left(\frac{1}{n_y}\sum_{i=1}^{n_y} y_i\right) \\
&= \frac{1}{n_x}\sum_{i=1}^{n_x} E(x_i) - \frac{1}{n_y}\sum_{i=1}^{n_y} E(y_i) \\
&= \frac{1}{n_x}\sum_{i=1}^{n_x} \mu_x - \frac{1}{n_y}\sum_{i=1}^{n_y} \mu_y \\
&= \frac{n_x\mu_x}{n_x} - \frac{n_y\mu_y}{n_y} = \mu_x - \mu_y
\end{aligned}$$





## Variance of the Difference

$$\begin{aligned}
Var(\bar{x} - \bar{y}) &= Var\left(\frac{1}{n_x}\sum_{i=1}^{n_x} x_i - \frac{1}{n_y}\sum_{i=1}^{n_y} y_i\right) \\
&= Var\left(\frac{1}{n_x}\sum_{i=1}^{n_x} x_i\right) + Var\left(\frac{1}{n_y}\sum_{i=1}^{n_y} y_i\right) \\
&= \frac{1}{n_x}\sum_{i=1}^{n_x^2} Var(x_i) + \frac{1}{n_y^2}\sum_{i=1}^{n_y} Var(y_i) \\
&= \frac{1}{n_x^2}\sum_{i=1}^{n_x} \sigma_x^2 + \frac{1}{n_y^2}\sum_{i=1}^{n_y} \sigma_y^2 \\
&= \frac{n_x\sigma_x^2}{n_x^2} + \frac{n_y\sigma_y^2}{n_y^2} = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}
\end{aligned}$$

Did I make any assumptions here?

## Checking assumptions & conditions

1. *Independence:*
   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.
   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.
2. *Sample size / skew:*
   Both distributions look reasonably symmetric, and the sample sizes are at least 30, 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 have the same form:

$$point\ estimate \pm ME$$

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

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





## 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 |

$$SE_{(\bar{x}_c-\bar{x}_{hs})} = \sqrt{\frac{s_c^2}{n_c} + \frac{s_{hs}^2}{n_{hs}}} = \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}} = 0.89$$





## 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.

$$\bar{x}_c = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_c-\bar{x}_{hs})} = 0.89$$

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

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

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?

$H_0$: $\mu_c = \mu_{hs} \rightarrow \mu_c - \mu_{hs} = 0$
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.

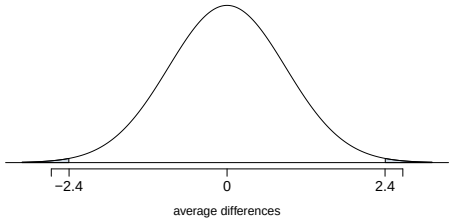




## Calculating the test-statistic and the p-value

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

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




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

Since the p-value is small, we 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.





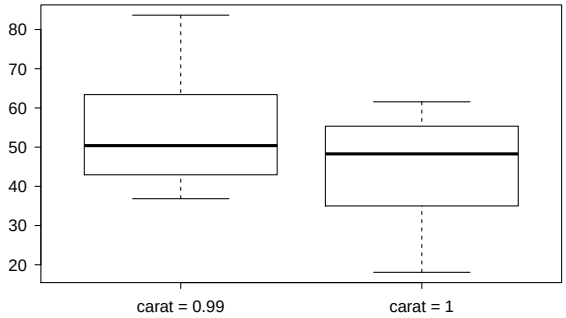
## Example - Diamonds

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





## Data




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

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

## Parameter and point estimate

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

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

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

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

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

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





## Hypothesis test

| | *0.99 carat* 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?

$$T = -2.508$$

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





## Synthesis

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

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

## Critical value

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

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





## Confidence interval

Calculate the interval, and interpret it in context.

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

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

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





## 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 (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - 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:

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