Chapter 7: Inference for numerical data

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

One-sample means with the *t* distribution

Notations

Recall: Sample mean, $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_1, x_2, \cdots, x_n$ represent the n observed values.

Recall: Sample variance, s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

where $x_1, x_2, \cdots, x_n$ represent the n observed values.

$s = \sqrt{s^2}$ is the sample standard deviation.

Central Limit Theorem(CLT) for the Sample Mean

Central Limit Theorem for the Sample Mean

When we collect a sufficiently large sample of n independent observations from a population with mean μ and variance σ^2, sample mean $\bar{X}$ will be nearly normally distributed with mean μ and variance σ^2/n.

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note: *(Exercise 3.47) If $X_1, \cdots, X_n$'s are independent observations from a distribution with mean μ and variance σ^2,*

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu,$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

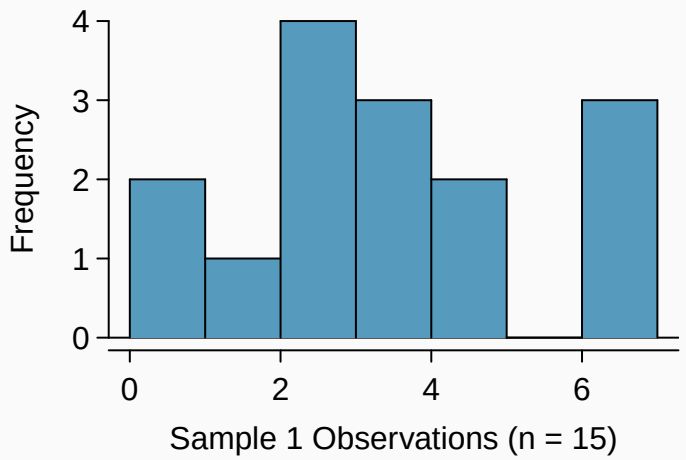
CLT - conditions

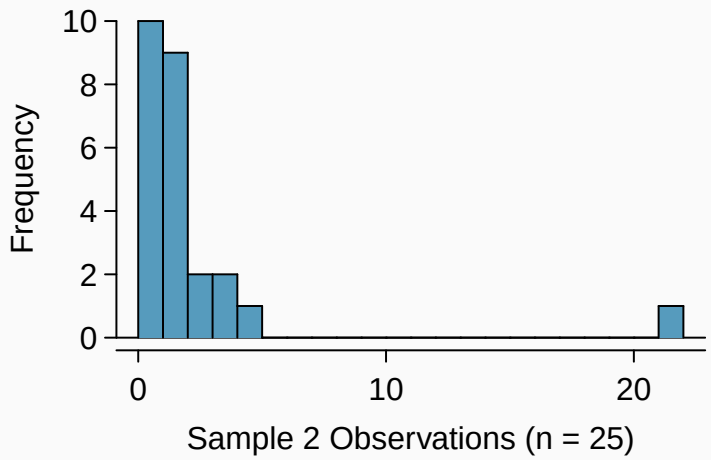
Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent.
    - This is usually satisfied if random sampling is used.
2. *Normality:* When the sample size is small, the sample observations should come from a normally distributed population.
    - **If** $n < 30$: The data should have no clear outliers to assume normality.
    - **If** $n \geq 30$: The sample size is large enough to apply the CLT.

Practice

Consider the following two plots that come from random samples from different populations. Are the independence and normality conditions met in each case?





Introducing the t-Distribution

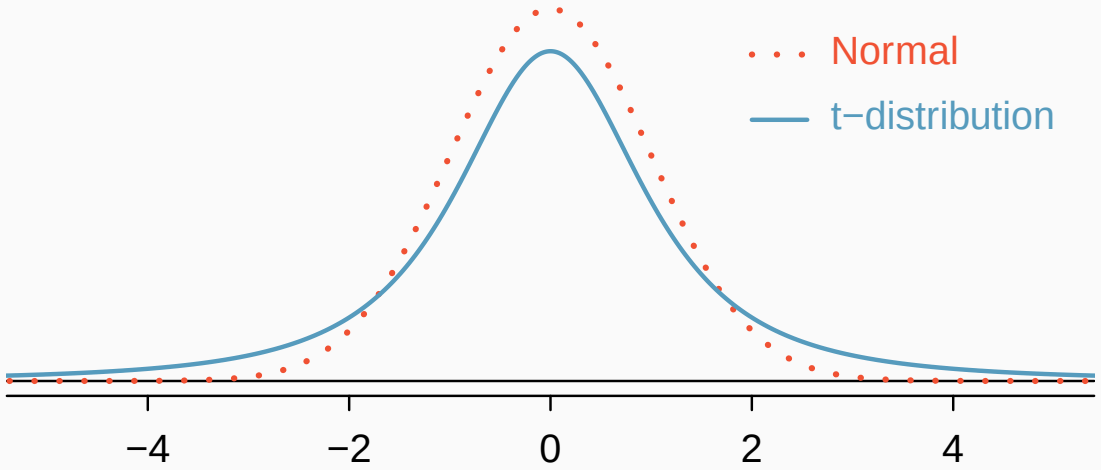
- In practice, we cannot directly compute the standard error for $\bar{x}$ since the population standard deviation σ is unknown.
- Instead, we estimate σ using the sample standard deviation s:

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

- This estimate works well for large samples but is less precise for small samples.
- To address this issue, we use the *t-distribution* instead of the normal distribution.

Comparison of the *t*-Distribution and Normal Distribution

- The *t*-distribution is similar to the standard normal distribution but has thicker tails.

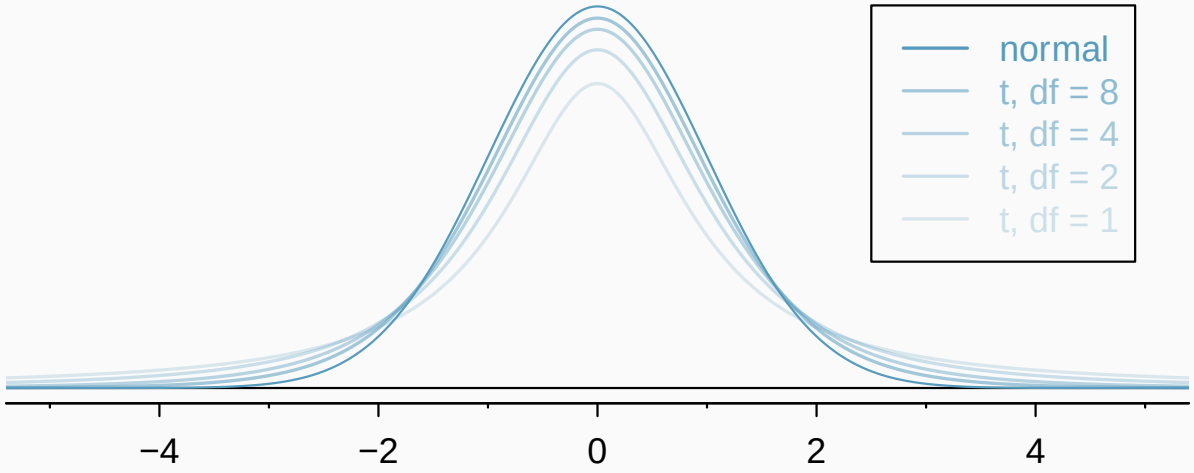


Degrees of Freedom (*df*)

- The *t*-distribution is always centered at zero and has a single parameter: **degrees of freedom (df)**.
- If a RV T has a t-distribution with degrees of freedom, df, write

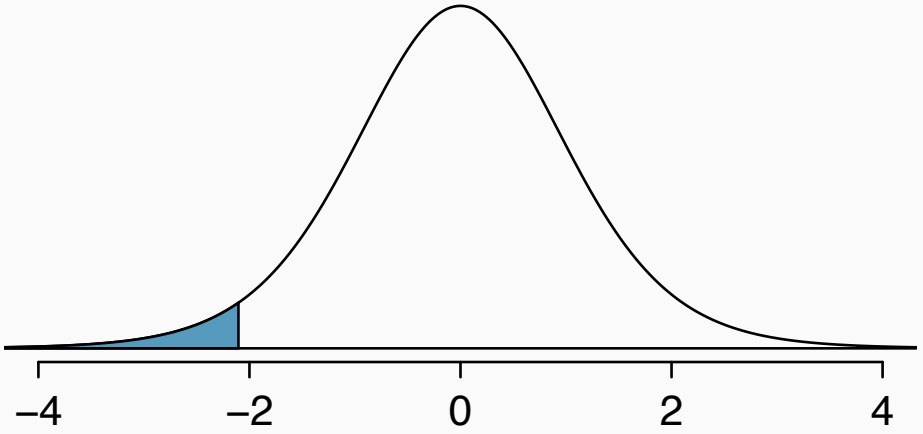
$$T \sim t(df).$$

- As *df* increases, the *t*-distribution becomes more similar to the normal distribution.



What proportion of the *t*-distribution with 18 degrees of freedom falls below -2.10?

We shade the area below -2.10 in the *t*-distribution with $df = 18$.




```
> pt(q = -2.10, df = 18, lower.tail = TRUE)
[1] 0.0250452
```

Confidence interval of mercury content in Risso's dolphins

Photo by Mike Baird (*www.bairdphotos.com*)

- Dolphins accumulate mercury from the ocean, which can be harmful to them and to humans who consume them.
- We will identify a confidence interval for the average mercury content in dolphin muscle.

Checking conditions for inference

n	$\bar{x}$	s	min	max
19	4.4	2.3	1.7	9.2

Table 1: Mercury content in 19 Risso's dolphins (in μg/wet g).

- **Independence:**
- **Normality:**

Recall: A z-confidence interval for the *proportion*

If a sample proportion $\hat{p}$ closely follows a normal model, then a $100(1-\alpha)\%$ confidence interval for the population proportion p is

$$\text{point estimate} \pm z_{\alpha/2} \times SE = \hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $z_{\alpha/2}$ is the $\alpha/2$-th upper quantile of the z-distribtuion.

A t-confidence interval for the *mean*

Based on a sample of n independent and normal observations, a $100(1-\alpha)\%$ confidence interval for the population mean μ is

$$\text{point estimate} \pm t_{\alpha/2}(df) \times SE = \bar{x} \pm t_{\alpha/2}(df) \times \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}(df)$ is the $\alpha/2$-th upper quantile of the t-distribtuion with degree of freedom $df = n-1$.

- Using $s = 2.3$ and $n = 19$:

$$SE =$$

- Using $\alpha = 0.05$ and $df =$, $t_{\alpha/2}(df) =$

```
> qt(p = 0.025, df = 18, lower.tail = FALSE)
[1] 2.100922
```

- Using $\bar{x} = 4.4$, 95% confidence interval for the mean is

$$\text{point estimate} \pm t_{\alpha/2}(df) \times SE \quad \rightarrow$$

- We are 95% confident that the true mean mercury content in Risso's dolphins is between 3.29 and 5.51 μg/wet gram.

One Sample t-Tests

To conduct a hypothesis test on the population mean μ:

$$H_0 : \mu = \mu_0, \quad H_A : \mu \neq \mu_0,$$

we use *T-score(or T-statistics)*.

Test statistic for inference on mean

The test statistic and its null distribution are

$$T = \frac{\bar{X} - \mu_0}{SE} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t(n-1),$$

where μ_0 is the null value, $\bar{X}$ is the sample mean, and S is the sample standard deviation.

Example

Research Question: has the average run time in the Cherry Blossom Race changed in 2017 compared to 2006?

- 2006 mean time: $\mu_0 = 93.29$ minutes.
- Sample of 100 runners from 2017.
- Hypotheses:

$$H_0 : \mu = 93.29 \quad \text{(No change in mean run time)}$$

$$H_A : \mu \neq 93.29 \quad \text{(Mean run time has changed)}$$

Checking conditions for inference:

- **Independence:** The sample is randomly collected, so independence is reasonable.
- **Normality:** The number of observation n is greater than 30.

Computing the Test Statistic

- Sample statistics:

$$n = 100, \quad \bar{x} = 97.32, \quad s = 16.98$$

- Compute standard error:

$$SE =$$

- Compute the observed T-score:

$$t =$$

Finding the P-value

P-value of the two sided test $= 2 \times P(T > 2.37)$, where $T \sim$

```
> pt(2.37, df = 99, lower.tail = FALSE)
[1] 0.009863209
```

Therefore, the P-value is

- Since P-value is (greater / less) than $\alpha = 0.05$, we (reject / don't reject) H_0.
- The data (provide / do not provide) strong evidence that the mean run time in 2017 was different from 2006.

```
> t.test(time, alternative = "two.sided", mu = 93.29)

 One Sample t-test

data:  time
t = 2.3701, df = 99, p-value = 0.019721
alternative hypothesis: true mean is not equal to 93.29
95 percent confidence interval:
  93.93546 100.70783
sample estimates:
mean of x
 97.32167
```

Remarks on t-distribution and T-statistics

Note: Definition of t-distribution: Suppose $Z \sim N(0,1)$ *and* $V \sim \chi^2(k)$ *are independent. Then*

$$\frac{Z}{\sqrt{V/k}} \sim t(k).$$

Note: Derivation of T-statistics: If $X_1, X_2, \cdots, X_n$ *are independent observations from a normal distribution with mean* μ *and variance* σ^2*, we have*

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } V := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ *and* $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ *are independent. From the definition of t-distribution, we have*

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} = \underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{Z} \times \left(\underbrace{\frac{(n-1)S^2}{\sigma^2}}_{V} \frac{1}{n-1} \right)^{-1/2} = \frac{Z}{\sqrt{V/(n-1)}} \sim t(n-1).$$

Remarks on variance estimation

Note: S^2 *is unbiased for* σ^2 *:* *Let* $X_1, \ldots, X_n$ *be a random sample from a population with mean* μ *and variance* σ^2*. The sample mean and variance is:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

To check S^2 *is unbiased for* σ^2*, we show that* $E[S^2] = \sigma^2$*:*

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2]\right) \qquad (1) \\ &= \frac{1}{n-1}\left(n\sigma^2 - n \cdot \frac{\sigma^2}{n}\right) = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2, \end{aligned}$$

and (1) *holds because*

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{0}.$$

Paired data

Amazon vs UCLA Bookstore

- We investigate whether Amazon prices were different from UCLA Bookstore prices.

	Subject	Bookstore Price	Amazon Price	Price Difference
1	American Indian Studies	47.97	47.45	0.52
2	Anthropology	14.26	13.55	0.71
3	Arts and Architecture	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	35.96	32.40	3.56

Paired Observations

Paired data

Two sets of observations are **paired** if each observation in one set corresponds with exactly one observation in the other set.

- Each textbook has two prices: UCLA Bookstore and Amazon.
- Differences computed as:

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

- Example differences:
  - $47.97 - 47.45 = 0.52$
  - $14.26 - 13.55 = 0.71$
  - $13.50 - 12.53 = 0.97$

Histogram of Price Differences

Does UCLA Bookstore prices seem different from Amazon?

Can we use *t*-test to compare the two prices?

Paired t-test

Consider a hypothesis test on the difference of two means μ_{diff}:

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

Test statistic for paired data

The test statistic and its null distribution are

$$T = \frac{\bar{X}_{diff} - 0}{SE(\bar{X}_{diff})} = \frac{\bar{X}_{diff}}{S_{diff}/\sqrt{n_{diff}}} \overset{H_0}{\sim} t(n_{diff} - 1),$$

where $\bar{X}_{diff}$, S_{diff}, and n_{diff} are the sample mean, the standard deviation, and the number of differences respectively.

Hypothesis test for Price Differences

- Hypotheses:

$$H_0 : \mu_{diff} = 0 \text{ (No difference in average textbook prices)}$$

$$H_A : \mu_{diff} \neq 0 \text{ (Difference in average prices)}$$

- To analyze a paired data set, we analyze the differences.
- Summary statistics for price differences

n_{diff}	$\bar{x}_{\text{diff}}$	s_{diff}
68	3.58	13.42

Computing the Test Statistic

- Sample statistics:

$$n_{diff} = 68, \quad \bar{x}_{diff} = 3.58, \quad s_{diff} = 13.42$$

- Compute standard error:

$$SE =$$

- Compute the observed T-score:

$$t =$$

Finding the P-value

P-value of the two sided test $= 2 \times P(T > 2.20)$, where $T \sim$

```
> pt(2.20, df = 67, lower.tail = FALSE)
[1] 0.01562998
```

Therefore, the P-value is

- Since P-value is (greater / less) than $\alpha = 0.05$, we (reject / don't reject) H_0.
- The data (provide / do not provide) strong evidence that Amazon prices are different from UCLA bookstore prices.

```
> t.test(UCLA, Amazon, alternative = "two.sided",
          mu = 0, paired = TRUE)

 Paired t-test

data:  UCLA and Amazon
t = 2.2012, df = 67, p-value = 0.03117
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.3340641 6.8324064
sample estimates:
mean difference
       3.583235
```

Practice

(Exercise 7.19) We analyzed temperature data from 197 NOAA stations where records were available for both 1948 and 2018, comparing the number of days exceeding 90°F. The average difference (2018 - 1948) was 2.9 days with a standard deviation of 17.2 days. We seek evidence that there were more days in 2018 that exceeded 90°F from NOAA's weather stations using **a paired t-test**.

1. State the null and alternative hypotheses.(Use a one-sided test.)

2. Find the null distribution of the test statistic.

Practice

(Exercise 7.19) We analyzed temperature data from 197 NOAA stations where records were available for both 1948 and 2018, comparing the number of days exceeding 90°F. The average difference (2018 - 1948) was 2.9 days with a standard deviation of 17.2 days. We seek evidence that there were more days in 2018 that exceeded 90°F from NOAA's weather stations using **a paired t-test**.

3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Difference of two means (1)

Stem cells and heart function

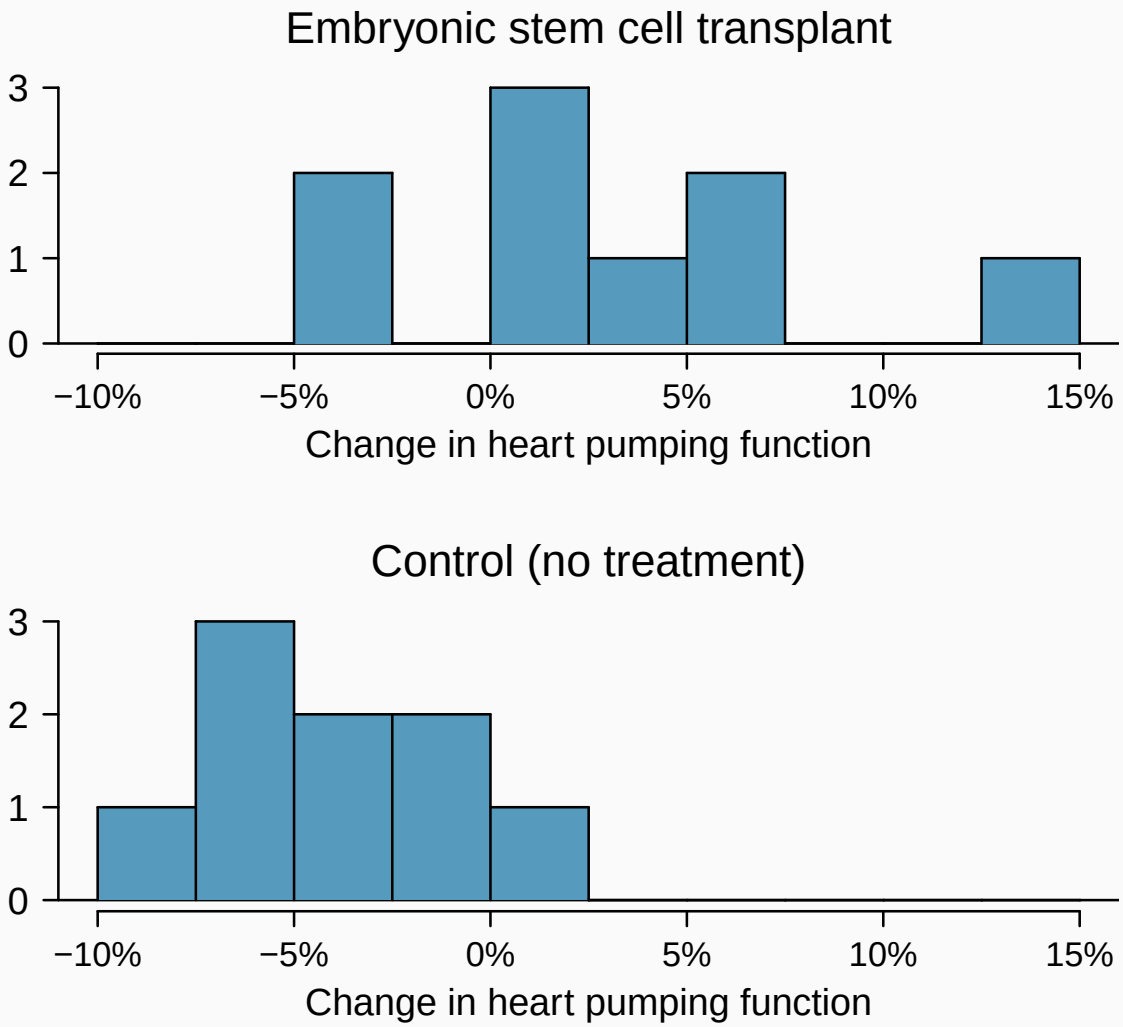
Study on embryonic stem cells (ESCs) and heart function in sheep.

- Sheep were *randomly assigned* to the ESC or control group.
- Change in heart pumping capacity was measured.
- Positive value corresponds to increased pumping capacity.

Group	n	$\bar{x}$	s
ESCs	9	3.50	5.17
Control	9	-4.33	2.76

Table 2: Summary statistics of the ESC study.

Are the ESC group and control group paired?



Can we use *t*-test to compare the two groups?

Suppose $\bar{X}_1$ and $\bar{X}_2$ are independent sample means of size n_1 and n_2 from the two distributions with mean μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. From Exercise 3.47,

$$Var(\bar{X}_1) = \frac{\sigma_1^2}{n_1}, \quad Var(\bar{X}_2) = \frac{\sigma_2^2}{n_2}.$$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent,

$$Var(\bar{X}_1 - \bar{X}_2) =$$

which leads to

$$SE(\bar{X}_1 - \bar{X}_2) =$$

Two-sample t-test

Consider a hypothesis test on a difference in two means $\mu_1 - \mu_2$.

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_A : \mu_1 - \mu_2 \neq 0$$

Test statistic for a difference in means

The test statistic and its null distribution are

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{H_0}{\sim} t(df),$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of groups 1 and 2, S_1^2 and S_2^2 are the sample variances, and n_1 and n_2 are the sample sizes. The exact df is complex, so we may use $df = \min(n_1, n_2) - 1$.

Note: *Exact df is the Satterthwaite's df:* $\psi = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$

Back to the ESC example

We want to determine whether ESC help improve heart function.

- Using one-sided test, hypotheses are:

$$H_0 : \mu_{esc} - \mu_{control} = 0$$
$$H_A :$$

Computing the Test Statistic

- Sample statistics:

$$n_{esc} = 9, \quad \bar{x}_{esc} = 3.50, \quad s_{esc} = 5.17$$

$$n_{control} = 9, \quad \bar{x}_{control} = -4.33, \quad s_{control} = 2.76$$

- Compute standard error:

$$SE =$$

- Compute the observed T-score:

$$t =$$

- Compute the Satterthwaite's degrees of freedom:

$$\psi = \frac{(5.17^2/9 + 2.76^2/9)^2}{(5.17^2/9)^2(9-1) + (2.76^2/9)^2/(9-1)} = 12.22$$

Finding the P-value

P-value of the one-sided test is

```
> pt(4.02, df = 12.22, lower.tail = FALSE)
[1] 0.0008202582
```

- P-value is less than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- The data (provide / do not provide) convincing evidence that ESCs help improve heart function following a heart attack.

t-test using Satterthwaite's degrees of freedom

```
> t.test(esc, ctrl, alternative = "greater", mu = 0)

 Welch Two Sample t-test

data:  esc and ctrl
t = 4.0073, df = 12.225, p-value = 0.0008388
alternative hypothesis: true difference
                        in means is greater than 0
95 percent confidence interval:
 4.35469     Inf
sample estimates:
mean of x mean of y
 3.500000 -4.333333
```

Practice

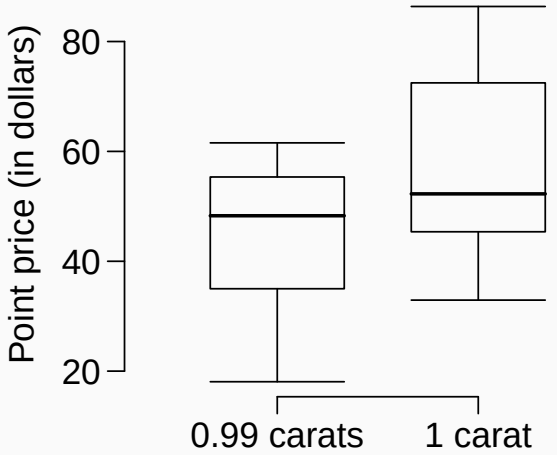
(Exercise 7.24) Diamond prices are known to increase with carat weight—1 carat diamonds are often priced much higher than 0.99 carat diamonds. We compare two random samples of 23 diamonds each (0.99 and 1 carat) using standardized prices, calculated by dividing the price by 100 times the carat weight. Conduct a hypothesis test to determine if 1 caret diamonds are more expensive than 0.99 carat diamonds.

1. State the null and alternative hypotheses.(Use one-sided test.)

2. Find the null distribution of the test statistic.

Practice

	0.99 carats	1 carat
Mean	\$44.51	\$56.81
SD	\$13.32	\$16.13
n	23	23



3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.
   (Use $\psi \approx 42.48$.)

Difference of two means (2)

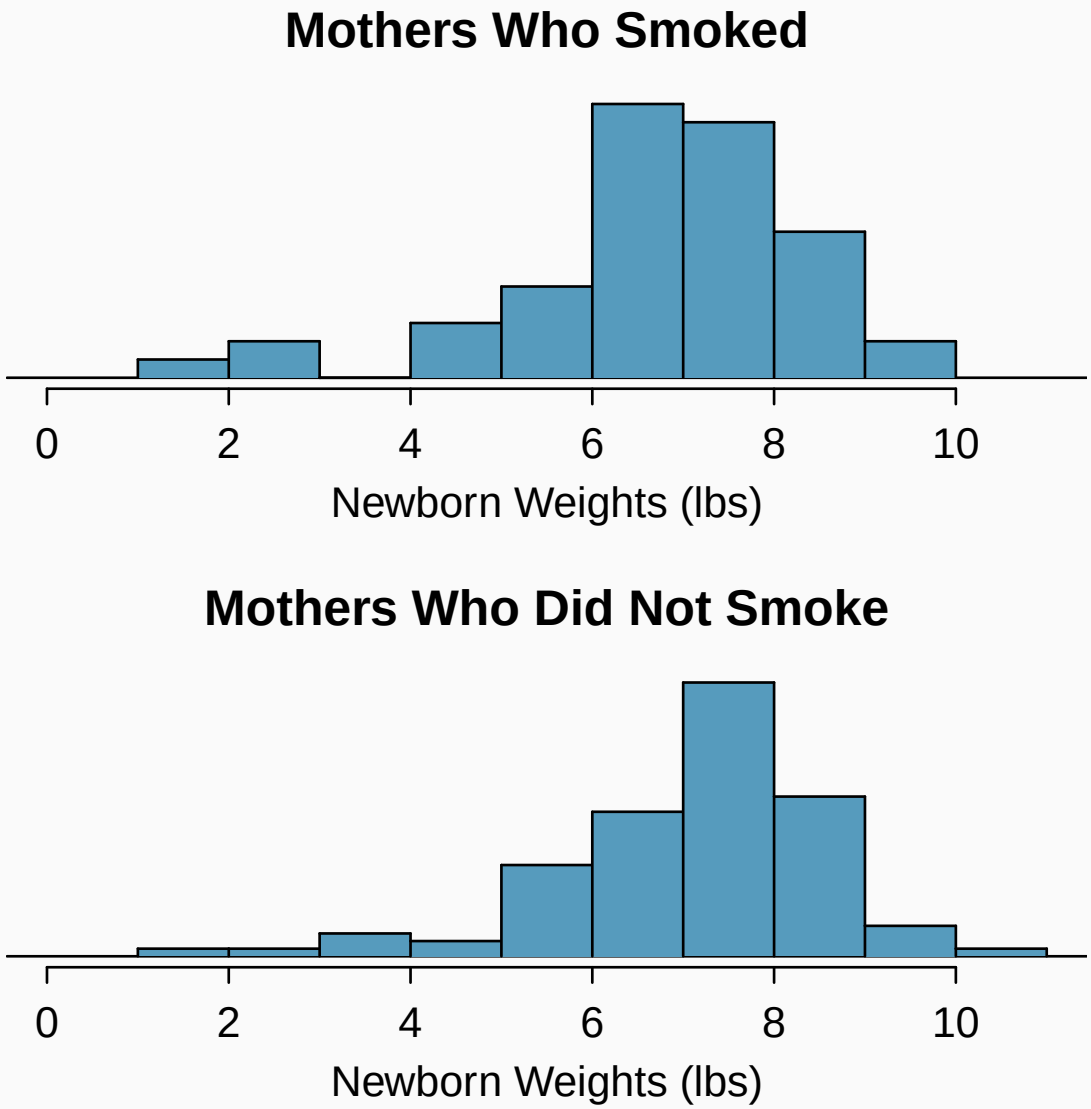
Smoking and birth weights of newborns

- We analyze the difference in birth weights of newborns based on whether the mother smoked during pregnancy.
- Data set: **ncbirths** - a random sample of 150 cases from North Carolina.
- Variables of interest:
  - **weight**: newborn's weight (in pounds)
  - **smoke**: whether the mother smoked (yes/no)

Sample Data

	fage	mage	weeks	weight	sex	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Table 3: Four cases from the **ncbirths** data set.



Careful statistical analysis suggested that the population variances of the two groups(smokers and non-smokers) are the same.

Suppose $\bar{X}_1$ and $\bar{X}_2$ are independent sample means of size n_1 and n_2 from the two distributions with mean μ_1 and μ_2 respectively and with *equal* variance σ^2. From Exercise 3.47,

$$Var(\bar{X}_1) = \frac{\sigma^2}{n_1}, \quad Var(\bar{X}_2) = \frac{\sigma^2}{n_2}.$$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent,

$$Var(\bar{X}_1 - \bar{X}_2) =$$

To estimate σ^2, we use pooled sample variance defined as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

where S_1^2 and S_2^2 are sample variances of two groups. This leads to

$$SE(\bar{X}_1 - \bar{X}_2) =$$

Two-sample t-test with equal variances

Suppose $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and consider a hypothesis test on $\mu_1 - \mu_2$:

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_A : \mu_1 - \mu_2 \neq 0$$

Test statistic for a difference in means with equal variances

The test statistic and its null distribution are

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \overset{H_0}{\sim} t(df),$$

where the pooled sample variance is defined as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

$\bar{X}_1$ and $\bar{X}_2$ are the sample means of groups 1 and 2, S_1^2 and S_2^2 are the sample variances, n_1 and n_2 are the sample sizes, and $df = n_1 + n_2 - 2$.

Remarks on the pooled sample variance

Note: Derivation of pooled sampled variance: *If $X_{11}, X_{12}, \cdots, X_{1n_1}$ are independent observations from a normal distribution with mean μ_1 and variance σ^2 and $X_{21}, X_{22}, \cdots, X_{2n_2}$ are from a normal distribution with mean μ_2 and the same variance σ^2, we have*

$$Z := \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1),$$

$$V_1 := \frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1), \text{ and } V_2 := \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1),$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means , S_1^2 and S_2^2 are the sample variances of groups 1 and 2. V_1 and V_2 are independent, so the additivity of χ^2 distribution gives

$$V := V_1 + V_2 = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} = \frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2).$$

It follows that

$$T := \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \times \left(\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \frac{1}{n_1+n_2-2} \right)^{-1/2}$$

$$= \frac{Z}{\sqrt{V/(n_1+n_2-2)}} \sim t(n_1+n_2-2)$$

Back to the example

Using one-sided test, hypotheses are:

$H_0 : \mu_N - \mu_S = 0$ (No difference in average birth weight)

H_A :

Checking conditions:

- Data is from a random sample, ensuring independence.
- Sample sizes are large: 50 (smokers), 100 (non-smokers).

Computing the Test Statistic

- Sample statistics:

$$n_N = 100, \quad \bar{x}_N = 7.18, \quad s_N = 1.60$$
$$n_S = 50, \quad \bar{x}_S = 6.78, \quad s_S = 1.43$$

- Compute pooled variance:

$$s_p^2 = \frac{(n_N - 1)s_N^2 + (n_S - 1)s_S^2}{n_N + n_S - 2} =$$

- Compute standard error:

$$SE =$$

- Compute the observed T-score:

$$t =$$

Finding the P-value

P-value of the one-sided test is

```
> pt(1.48, df = _______________, lower.tail = FALSE)
[1] 0.07049939
```

- P-value is larger than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- There is (sufficient / insufficient) evidence to conclude that the babies born from women who smoke are smaller on average.

However, a larger data set with an increased sample size shows that smoking is associated with lower birth weight.

```
> t.test(nonsmoker, smoker, alternative = "greater",
+        mu = 0, var.equal = T)

 Two Sample t-test

data:  nonsmoker and smoker
t = 1.4824, df = 148, p-value = 0.07050
alternative hypothesis: true difference
                        in means is greater than 0
95 percent confidence interval:
 -0.046407        Inf
sample estimates:
mean of x mean of y
   7.1795    6.7790
```

Practice

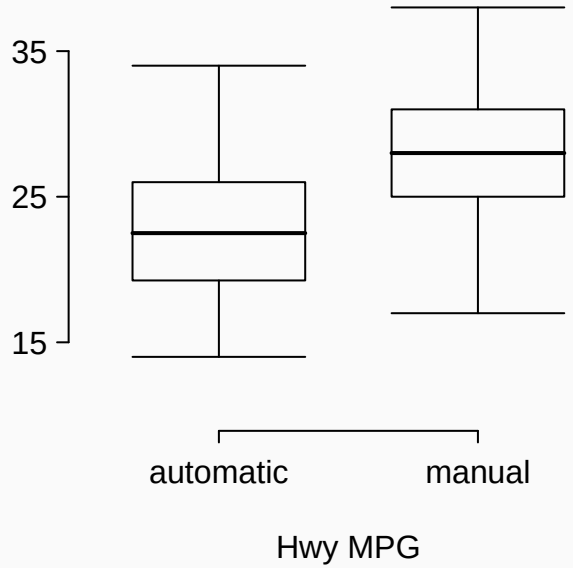
(Exercise 7.30*) The US Environmental Protection Agency (EPA) collects fuel economy data annually. Below are summary statistics on highway fuel efficiency (MPG) from random samples of cars with automatic and manual transmissions. Conduct a hypothesis test to determine if there is a difference in average highway mileage between the two transmission types. **Assume that the variances of the two type's MPG are equal.**

1. State the null and alternative hypotheses.

2. Find the null distribution of the test statistic.

Practice

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26



3. Compute the observed test statistic.

4. Compute the p-value and complete the hypothesis test.

Power calculations for a difference of means

Blood pressure example

A pharmaceutical company is testing a new drug for lowering blood pressure. Patients taking a standard blood pressure medication are randomly assigned to:

- Control group: Continue with the current medication.
- Treatment group: Receive the new drug.

Hypotheses:

$$H_0 : \mu_{trmt} - \mu_{ctrl} = 0, \quad H_A : \mu_{trmt} - \mu_{ctrl} \neq 0$$

Previously published studies suggest that the *standard deviation* of the blood pressure is about 12 mmHg.

Suppose $\bar{X}_1$ and $\bar{X}_2$ are independent sample means of size n_1 and n_2 from the two normal distributions with mean μ_1 and μ_2 and *known* variances σ_1^2 and σ_2^2 respectively.

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent,

$$\bar{X}_1 - \bar{X}_2 \sim$$

Two-sample Z-test with known variances

Suppose σ_1^2, σ_2^2 are known and consider a hypothesis test on $\mu_1 - \mu_2$:

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_A : \mu_1 - \mu_2 \neq 0$$

Test statistic for a difference in means with known variances

The test statistic and its null distribution are

$$\bar{X}_1 - \bar{X}_2 \overset{H_0}{\sim} N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of groups 1 and 2, σ_1^2 and σ_2^2 are the known variances.

Back to the example

- Sample sizes and standard deviations of the two groups are

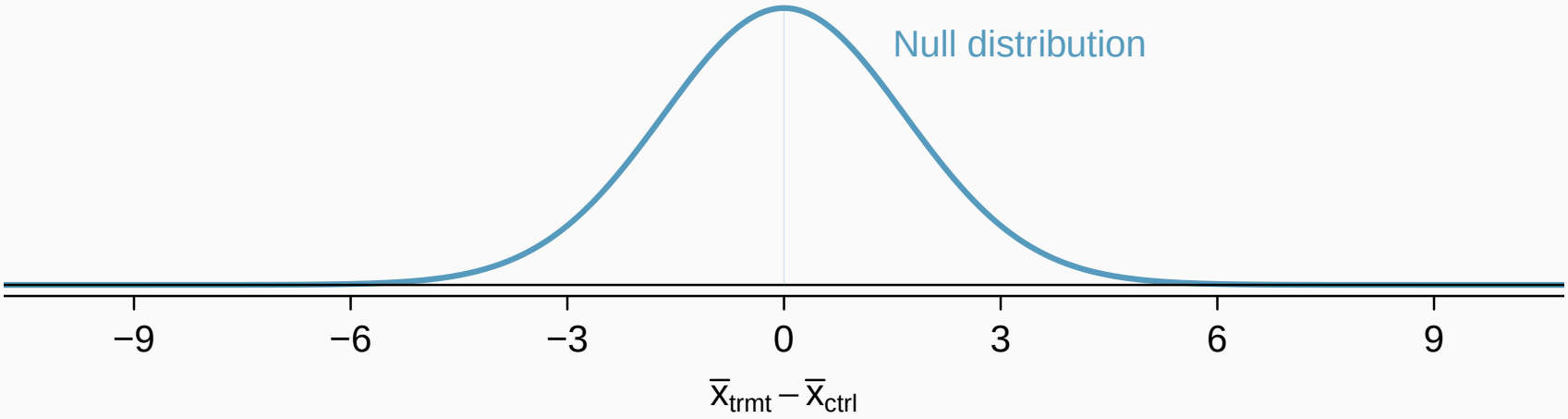
$$n_{trmt} = n_{ctrl} = 100, \quad \sigma_{trmt} = \sigma_{ctrl} = 12.$$

- Standard error of $\bar{X}_{trmt} - \bar{X}_{ctrl}$:

$$SE(\bar{X}_{trmt} - \bar{X}_{ctrl}) =$$

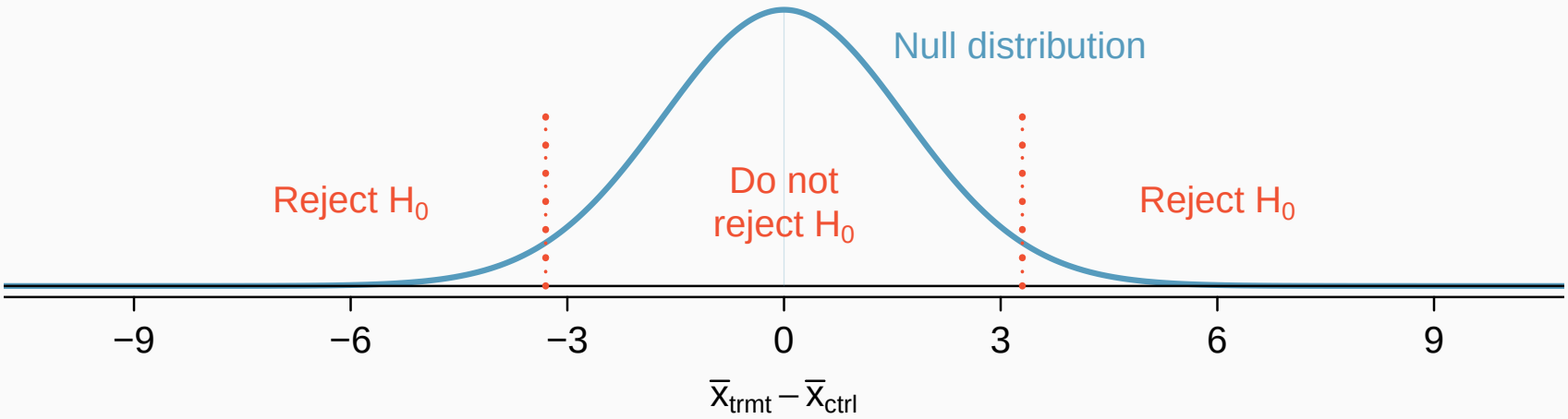
Null Distribution

- $\bar{X}_{trmt} - \bar{X}_{ctrl}$ is normally distributed.
- Its standard error (standard deviation) is 1.70.
- Under H_0, the mean difference in the population is 0.



Rejection Regions for $\alpha = 0.05$

- Lower 2.5% tail: $-z_{0.025} \times SE = -1.96 \times 1.70 = -3.33$ mmHg.
- Upper 2.5% tail: $z_{0.025} \times SE = 1.96 \times 1.70 = 3.33$ mmHg.
- Reject H_0 if $\bar{x}_{trmt} - \bar{x}_{ctrl}$ falls outside $(-3.33, 3.33)$.



We say *rejection region* is $|\bar{x}_{trmt} - \bar{x}_{ctrl}| > 3.33$.

Rejection region for a two-sided test

The rejection region for a two-sided test

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_A : \mu_1 - \mu_2 \neq 0$$

with significance level α is

$$|\bar{x}_1 - \bar{x}_2| > z_{\alpha/2} \times SE.$$

Rejection region for a one-sided test

The rejection region for a one-sided test

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_A : \mu_1 - \mu_2 > 0$$

with significance level α is

Power of a test

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

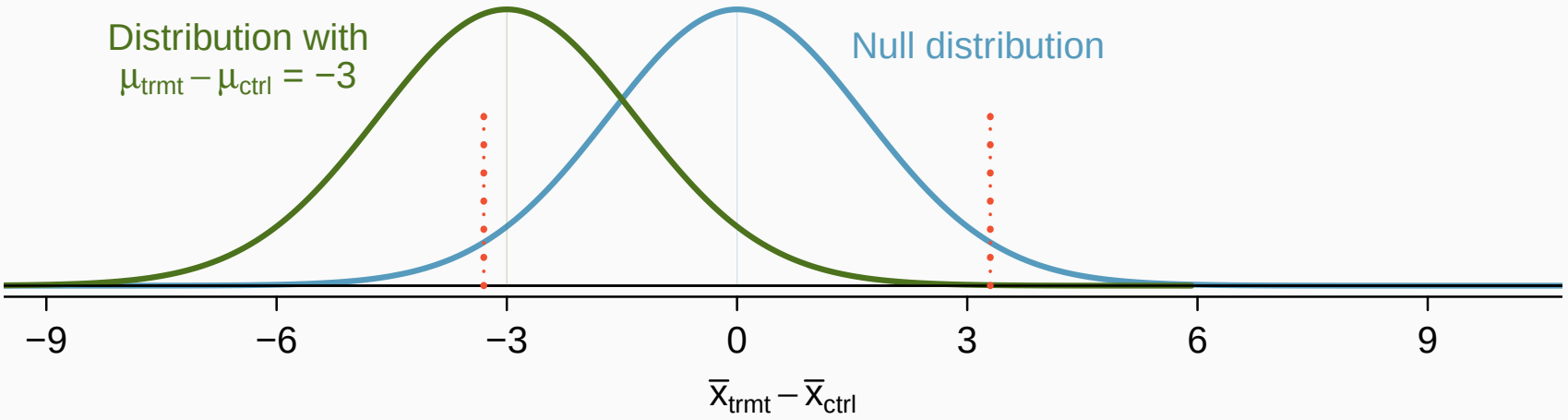
- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true})$.
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta = P(\text{Reject } H_0 \mid H_A \text{ is true})$

Effect Size

- Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger versus the standard medication.
- In our case, a 3 mmHg drop in blood pressure is the minimum *effect size* of interest.
- We want to determine how likely we are to detect this effect.

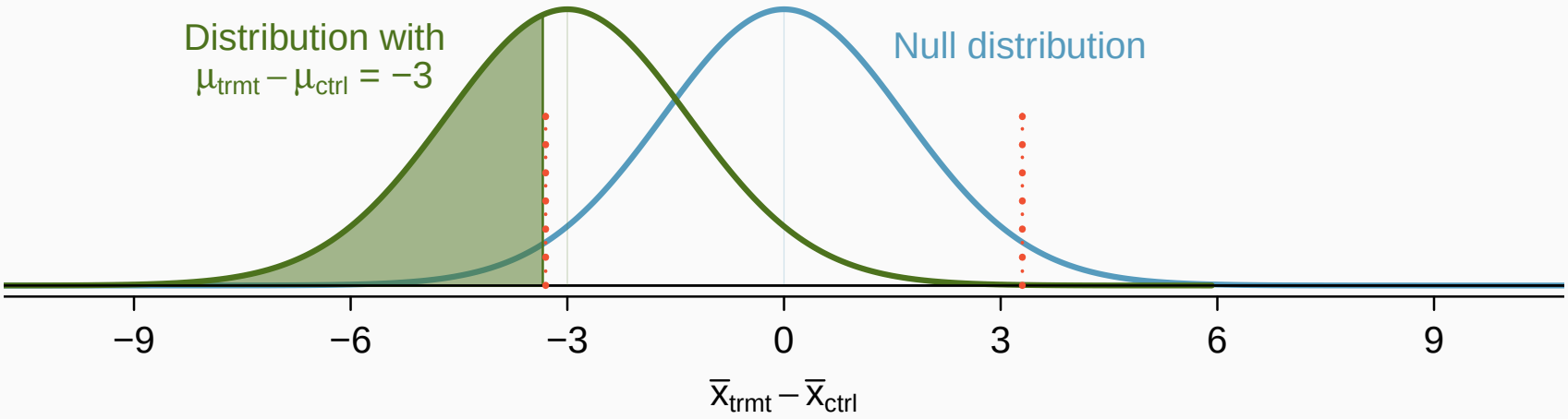
Shifted Sampling Distribution

- Under H_0, the distribution of $\bar{X}_{trmt} - \bar{X}_{ctrl}$ is centered at 0.
- If the true mean difference $\mu_{trmt} - \mu_{ctrl}$ is -3 mmHg under H_A, the distribution shifts left by 3.
- The standard deviation remains the same as in the null distribution.



Identifying the Rejection Regions

- The rejection regions remain unchanged.
- We reject H_0 if $|\bar{x}_{trmt} - \bar{x}_{ctrl}| > 3.33$.
- We calculate the fraction of the new distribution that falls in the rejection region.



Computing the power

The power of the test is

$P(\text{Reject } H_0 \mid H_A \text{ is true}) =$

First, $P(\bar{X}_{trmt} - \bar{X}_{ctrl} < -3.33 \mid \mu_{trmt} - \mu_{ctrl} = -3)$

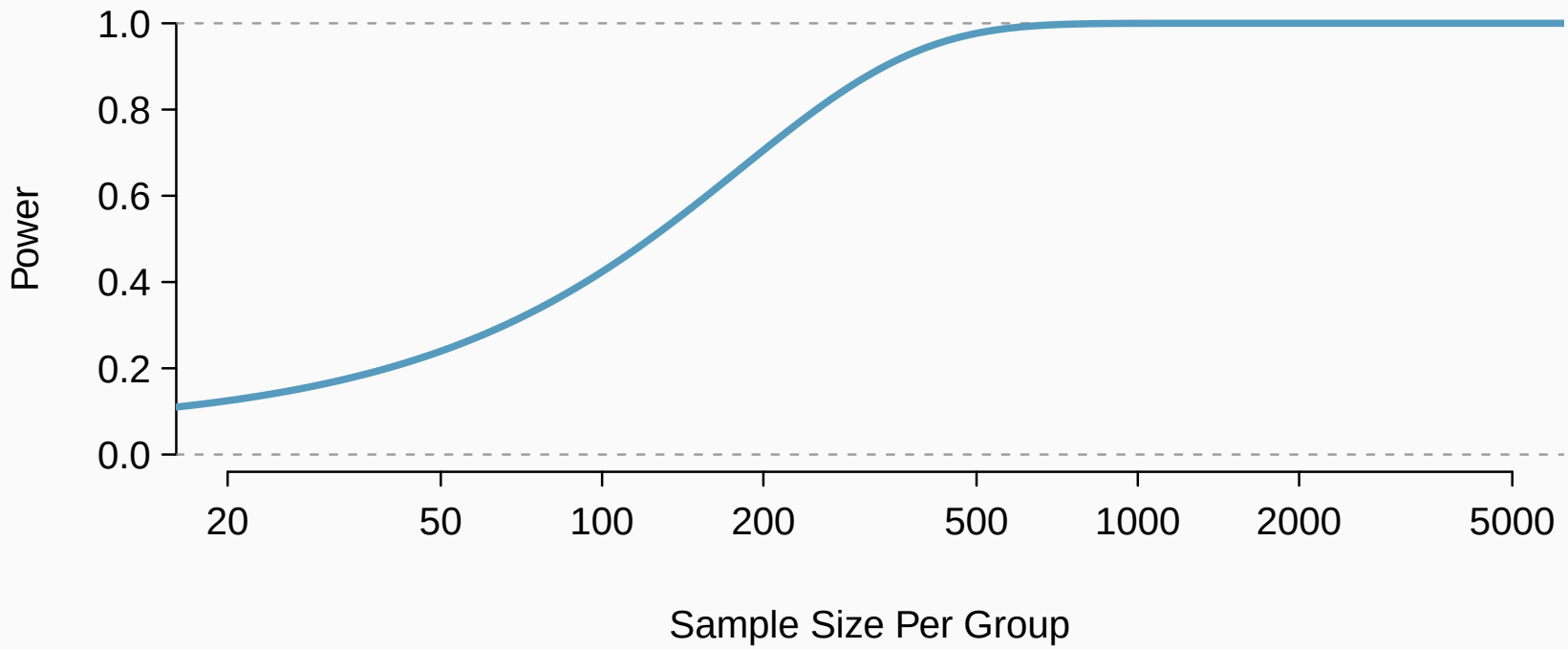
$$= P\left(\frac{\bar{X}_{trmt} - \bar{X}_{ctrl} - (-3)}{1.70} < \frac{-3.33 - (-3)}{1.70}\right) = P(Z < -0.2) = 0.42$$

Secondly, $P(\bar{X}_{trmt} - \bar{X}_{ctrl} > 3.33 \mid \mu_{trmt} - \mu_{ctrl} = -3)$

$$= P\left(\frac{\bar{X}_{trmt} - \bar{X}_{ctrl} - (-3)}{1.70} > \frac{3.33 - (-3)}{1.70}\right) = P(Z > 3.72) \approx 10^{-4}$$

So the power of the test is $0.42 + 10^{-4} = 0.42$.

As power increases with a larger sample size, we may calculate the required sample size for a desired level of power.



Practice

D Electronics' mobile phones use batteries from Company A. Company B claims their batteries have a longer lifespan. To test this at $\alpha = 0.05$, D Electronics samples $n_A = n_B = 16$ batteries from each company. Let $X_{A1}, \cdots, X_{A16} \sim N(\mu_A, 64)$ and $X_{B1}, \cdots, X_{B16} \sim N(\mu_B, 80)$ denote the lifespans of batteries from Companies A and B, respectively.

1. State the null and alternative hypotheses(Use **one-sided** test).

2. Find the null distribution of the test statistic.

Practice

D Electronics' mobile phones use batteries from Company A. Company B claims their batteries have a longer lifespan. To test this at $\alpha = 0.05$, D Electronics samples $n_A = n_B = 16$ batteries from each company. Let $X_{A1}, \cdots, X_{A16} \sim N(\mu_A, 64)$ and $X_{B1}, \cdots, X_{B16} \sim N(\mu_B, 80)$ denote the lifespans of batteries from Companies A and B, respectively.

3. Find the rejection region.

4. Given that $\bar{x}_A = 60$, $\bar{x}_B = 66$, complete the hypothesis test.

Practice

D Electronics' mobile phones use batteries from Company A. Company B claims their batteries have a longer lifespan. To test this at $\alpha = 0.05$, D Electronics samples $n_A = n_B = 16$ batteries from each company. Let $X_{A1}, \cdots, X_{A16} \sim N(\mu_A, 64)$ and $X_{B1}, \cdots, X_{B16} \sim N(\mu_B, 80)$ denote the lifespans of batteries from Companies A and B, respectively.

5. Find the power of this test assuming $\mu_A - \mu_B = -6$.

Summary of Chapter 7

	Null hypothesis	Test statistic and its null distribution $\left(\frac{\text{point estimate} - \text{null value}}{\text{SE}}\right)$	df
One-sample t-test	$H_0 : \mu = \mu_0$	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(df)$	$n - 1$
Paired t-test	$H_0 : \mu_{diff} = 0$	$T = \frac{\bar{X}_{diff}}{S_{diff}/\sqrt{n_{diff}}} \sim t(df)$	$n_{diff} - 1$
Two-sample t-test	$H_0 : \mu_1 - \mu_2 = 0$	$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(df)$	ψ
Two-sample t-test (with equal variances)	$H_0 : \mu_1 - \mu_2 = 0$	$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(df)$	$n_1 + n_2 - 2$
Two-sample Z-test (with known variances)	$H_0 : \mu_1 - \mu_2 = 0$	$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$	

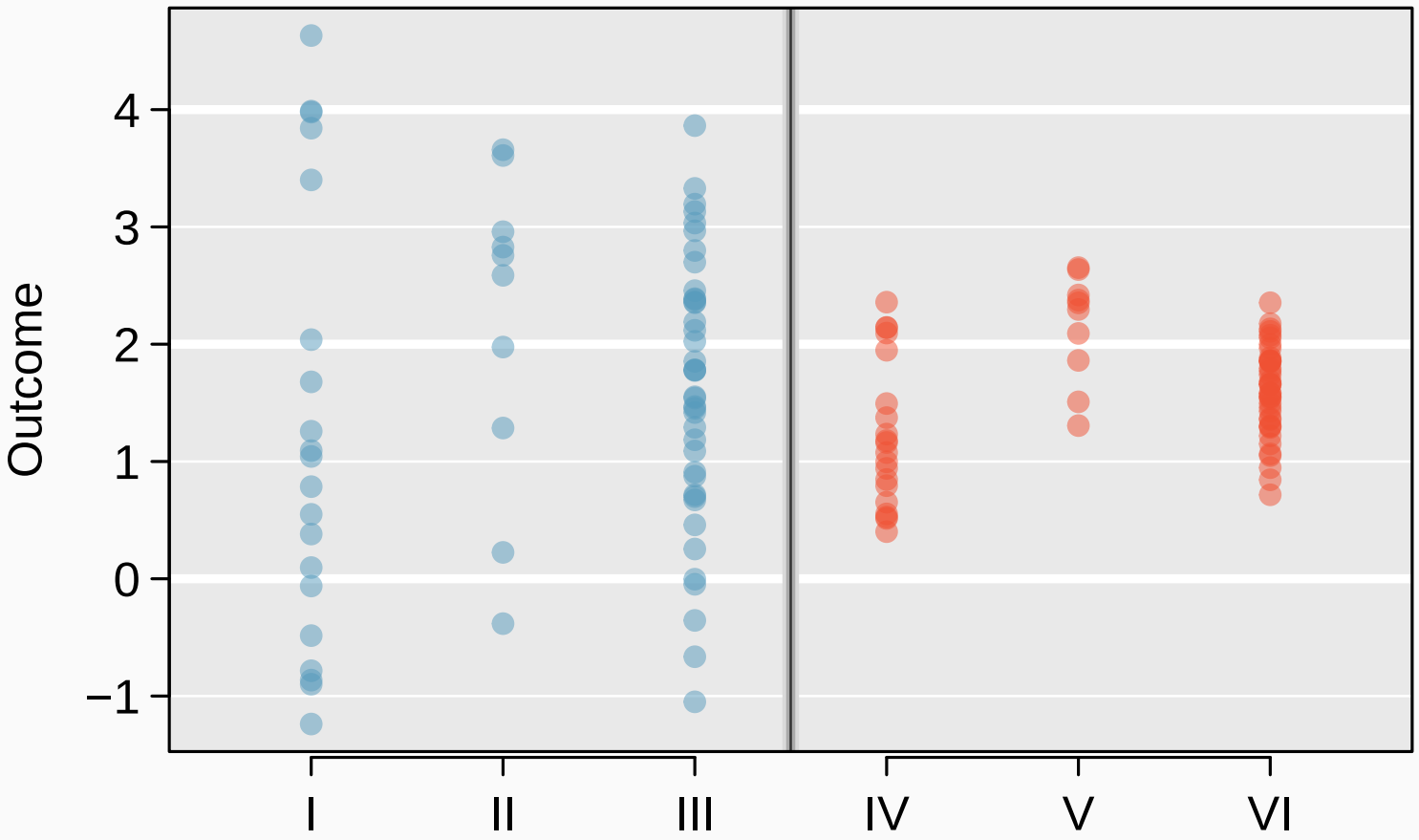
Comparing many means with ANOVA

Motivating Example: Exam Scores in Three Classes

- A statistics department runs three lectures: A, B, and C.
- We want to determine if there are significant differences in first exam scores.
- Hypotheses:
  - H_0: The average score is the same in all lectures ($\mu_A = \mu_B = \mu_C$).
  - H_A: The average score differs in at least one class.

Comparing Variability

Compare groups I, II, and III. Is there a difference between the mean scores across the three groups? What about for groups IV, V, and VI?



ANOVA: ANalysis Of VAriance

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

H_0 : The mean outcome is the same across all categories,

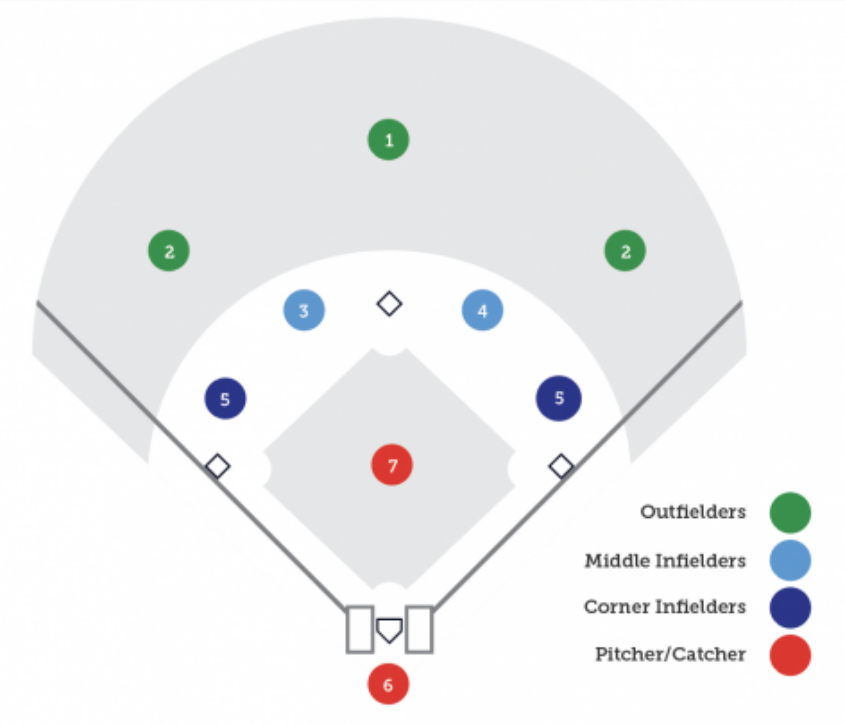
$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i.

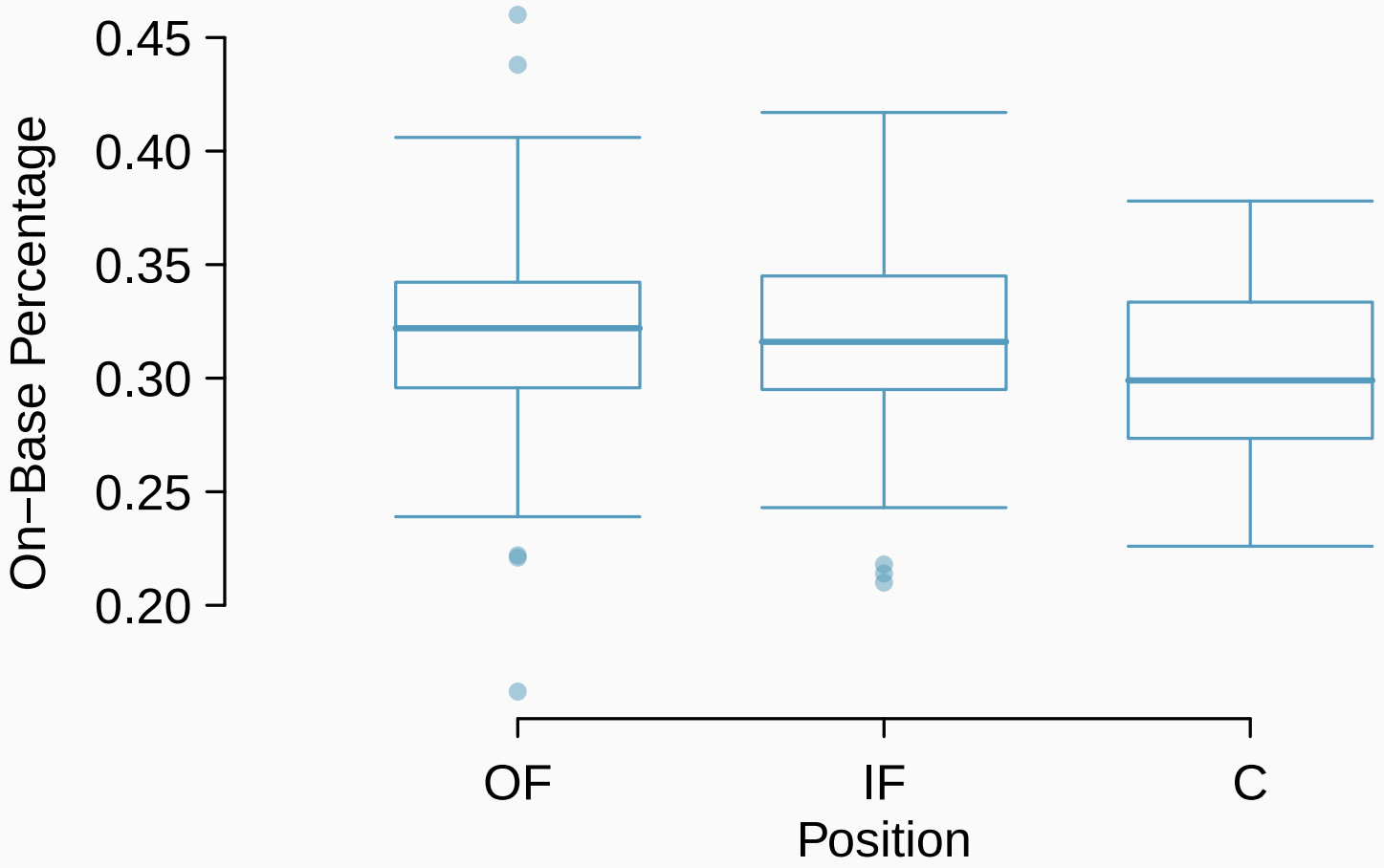
H_A : At least one mean is different than others.

Example: MLB Batting Performance

- Investigating batting performance by player position:
  - Outfield, Infield, Catcher
- Using OBP as the measure of batting performance.



Box Plot of On-Base Percentage



Overview of ANOVA

Uses the F-statistic: $F = \dfrac{MSG}{MSE}$.

- MSG (Mean Square between Groups) measures variability among group means.
- MSE (Mean Square Error) measures within-group variability.
- The F-statistic compares between-group and within-group variability.

Mean Square between Groups (*MSG*)

- Measures variability among group means.
- Computed as:

$$MSG = \frac{SSG}{df_G} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{df_G},$$

  where *SSG* is the Sum of Squares between Groups.
- Degrees of freedom: $df_G = k - 1$.

group number	1	2	$\cdots$	k	All
Sample Size	n_1	n_2	$\cdots$	n_k	$n = n_1 + \cdots + n_k$
Sample Mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$	$\bar{x}$
Sample SD	s_1	s_2	$\cdots$	s_k	s

Mean Square Error (*MSE*)

- Measures within-group variability.
- Computed as:

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{df_E},$$

  where SSE is the Sum of Squares Error.
- Degrees of freedom: $df_E = n - k$.

group number	1	2	$\cdots$	k	All
Sample Size	n_1	n_2	$\cdots$	n_k	$n = n_1 + \cdots + n_k$
Sample Mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$	$\bar{x}$
Sample SD	s_1	s_2	$\cdots$	s_k	s

Consider a hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \qquad H_A : \text{ not } H_0$$

Test statistic for ANOVA

The test statistic and its null distributions are

$$F = \frac{MSG}{MSE} = \frac{SSG/df_G}{SSE/df_E} \overset{H_0}{\sim} F(df_G, df_E),$$

where $df_G = k - 1$, $df_E = n - k$,

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2, \qquad SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

for group sample size n_i, the group sample means, $\bar{x}_i$, the group sample variances s_i^2 and the whole sample mean, $\bar{x}$.

Note: Decomposition of Sum of Squares: Let x_{ij} be the j-th observation from i-th group. Then

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2,$$

$$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2.$$

Define Sum of Squares Total(SST):

$$SST = (n-1)s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2.$$

Then $SST = SSG + SSE$ because

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \\ &= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}_{=SSE} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2}_{=SSG} + 2\sum_{i=1}^{k}(\bar{x}_i - \bar{x})\underbrace{\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)}_{=0} \end{aligned}$$

Note: Definition of F-distribution: *Suppose $V_1 \sim \chi^2(k_1)$ and $V_2 \sim \chi^2(k_2)$ are independent. Then the random variable $F := \frac{V_1/k_1}{V_2/k_2}$ follows F distribution with degrees of freedom k_1 and k_2:*

$$F := \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2).$$

Note: Derivation of F-statistic: *Now, suppose*

$$x_{ij} = \mu_i + e_{ij}, \quad i = 1, \cdots, k, \quad j = 1, \cdots, n_i,$$

and $e_{ij} \sim N(0, \sigma^2)$ are independent. It is known that

$$V_2 := SSE/\sigma^2 \sim \chi^2(n-k), \; SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2.$$

If H_0 is true in the sense that $\mu_1 = \mu_2 = \cdots = \mu_k$, it is also known that

$$V_1 := SSG/\sigma^2 \sim \chi^2(k-1), \; SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2.$$

Independence of V_1 and V_2 gives

$$F = \frac{V_1/(k-1)}{V_2/(n-k)} = \frac{SSG/(k-1)}{SSE/(n-k)} = \frac{MSG}{MSE} \overset{H_0}{\sim} F(k-1, n-k).$$

Note: Relationship between t-distribution and F-distribution: Suppose $Z \sim N(0,1)$ and $V \sim \chi^2(k)$ are independent. Then by the definition of t-distribution, the random variable $T := \frac{Z}{\sqrt{V/k}}$ follows t distribution with degrees of freedom k:

$$T := \frac{Z}{\sqrt{V/k}} \sim t(k).$$

Furthermore, by the definition of χ^2-distribution, the random variable Z^2 follows χ^2-distribution with degrees of freedom 1. Since Z^2 and V are independent, the definition of F-distribution gives

$$F := \frac{Z^2/1}{V/k} = T^2 \sim F(1,k).$$

In other words, if $T \sim t(k)$ and $F \sim F(1,k)$ then $F = T^2$.

Example: MLB Player Positions

Summary statistics:

	Outfield	Infield	Catcher	All
Sample Size (n)	160	205	64	429
Sample Mean ($\bar{x}$)	0.320	0.318	0.302	0.317
Sample SD (s)	0.043	0.038	0.038	0.040

Compute the Sum of Squared between Groups:

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 =$$

Compute the Sum of Squared Errors:

$$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2 =$$

Example: MLB Player Positions

Summary statistics: $SSG = 0.0161$, $SSE = 0.6740$,

	Outfield	Infield	Catcher	All
Sample Size (n)	160	205	64	429
Sample Mean ($\bar{x}$)	0.320	0.318	0.302	0.317
Sample SD (s)	0.043	0.038	0.038	0.040

Compute the Mean Squared between Groups:

$$MSG =$$

Compute the Mean Squared Errors:

$$MSE =$$

Compute the F statistic:

$$F =$$

Interpreting the *F*-Distribution

Larger *F* statistic indicates more significant group differences.

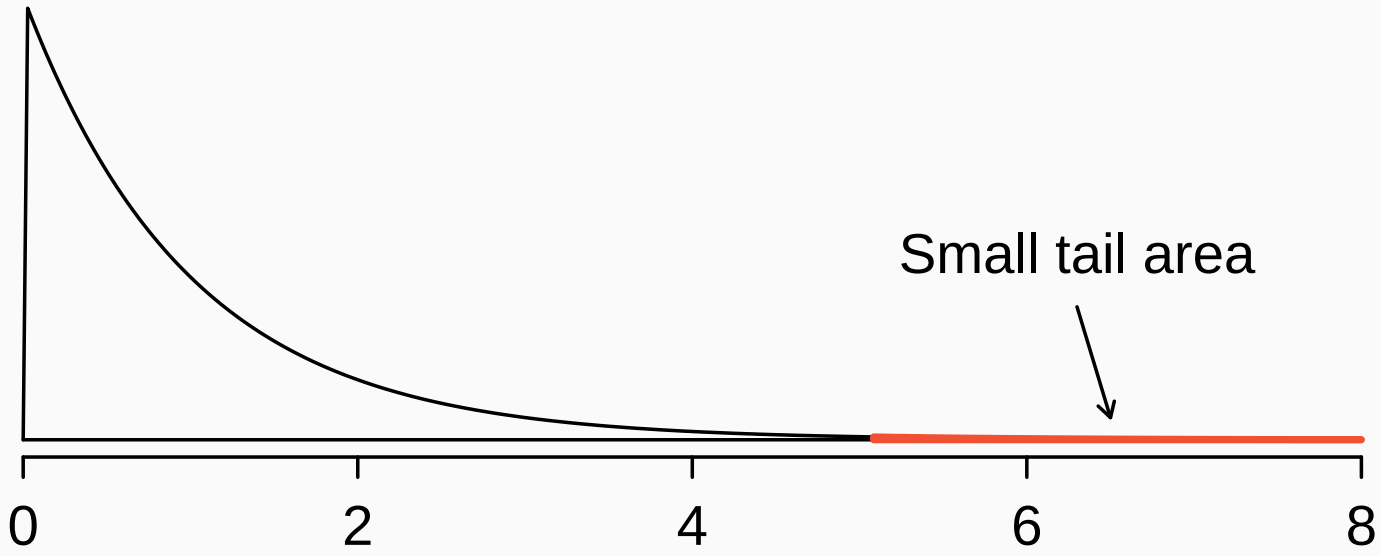


Figure 1: An *F* distribution with $df_1 = 2$ and $df_2 = 426$.

Finding the P-value

P-value is

```
> pf(5.08, df1 = ____, df2 = ____, lower.tail = FALSE)
[1] 0.006602098
```

- P-value is less than $\alpha = 0.05$, so we (reject / don't reject) H_0.
- There (is / isn't) significant evidence to conclude that OBP differs across player positions.

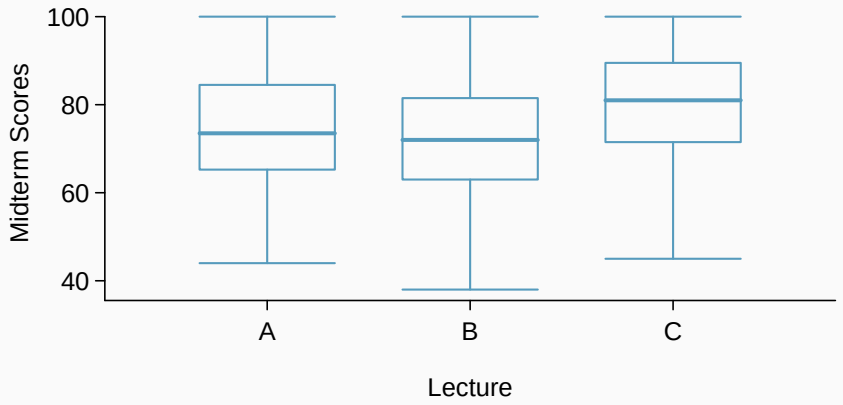
```
> mod <- lm(OBP ~ position, data = mlb_players_18)
> anova(mod)
Analysis of Variance Table

Response: OBP
            Df  Sum Sq   Mean Sq F value   Pr(>F)
position     2 0.01606 0.0080314  5.0766 0.006624 **
Residuals  426 0.67395 0.0015820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
```

Practice

ANOVA was conducted for the exam score data, and partial summary results are shown below. Fill in the blank and complete the hypothesis test.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

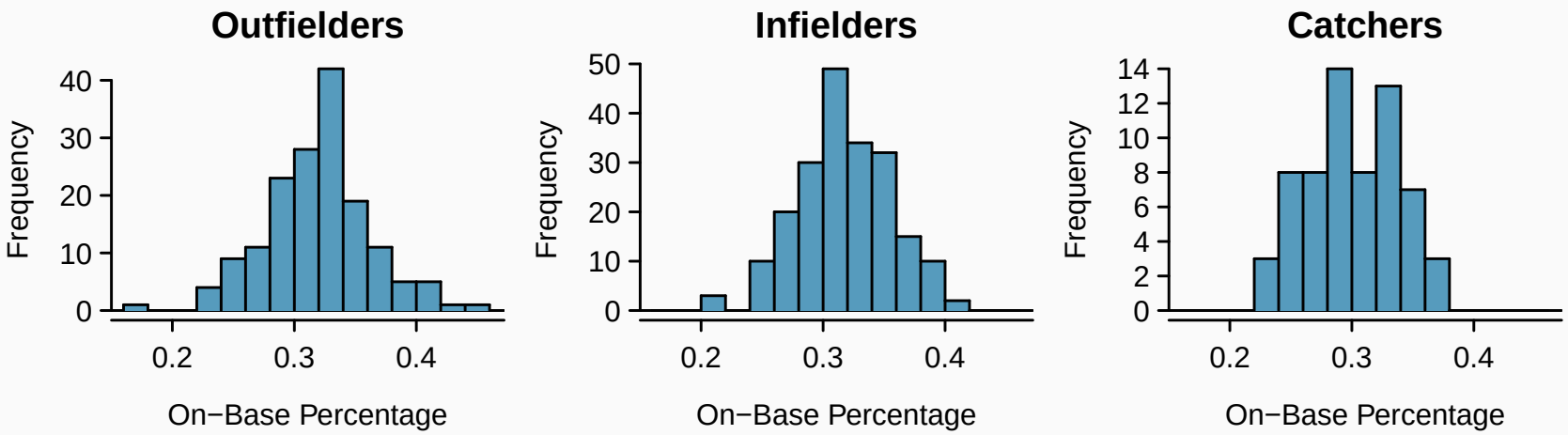


	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	___	1290.11	_____	_____	_____
Residuals	___	29810.13	_____		

Checking conditions

Three conditions to check for an ANOVA analyis:

1. **Independence**
   - All observations must be independent.
2. **Approximately normal**
   - The data in each group must be nearly normal.
3. **Constant variance**
   - The variance within each group must be approximately equal.



Exercises

Exercises in OpenIntro Statistics 4th ed.

- One-sample means with t-distribution: Exercise 7.7, 7.11(b)
- Paired data: Exercise 7.20(c, d, e)
- Difference of two means: Exercise 7.29, 7.47
- Power calculations for a difference of means: Example 7.35
- Comparing many means with ANOVA: Exercise 7.39 (a, c, d)