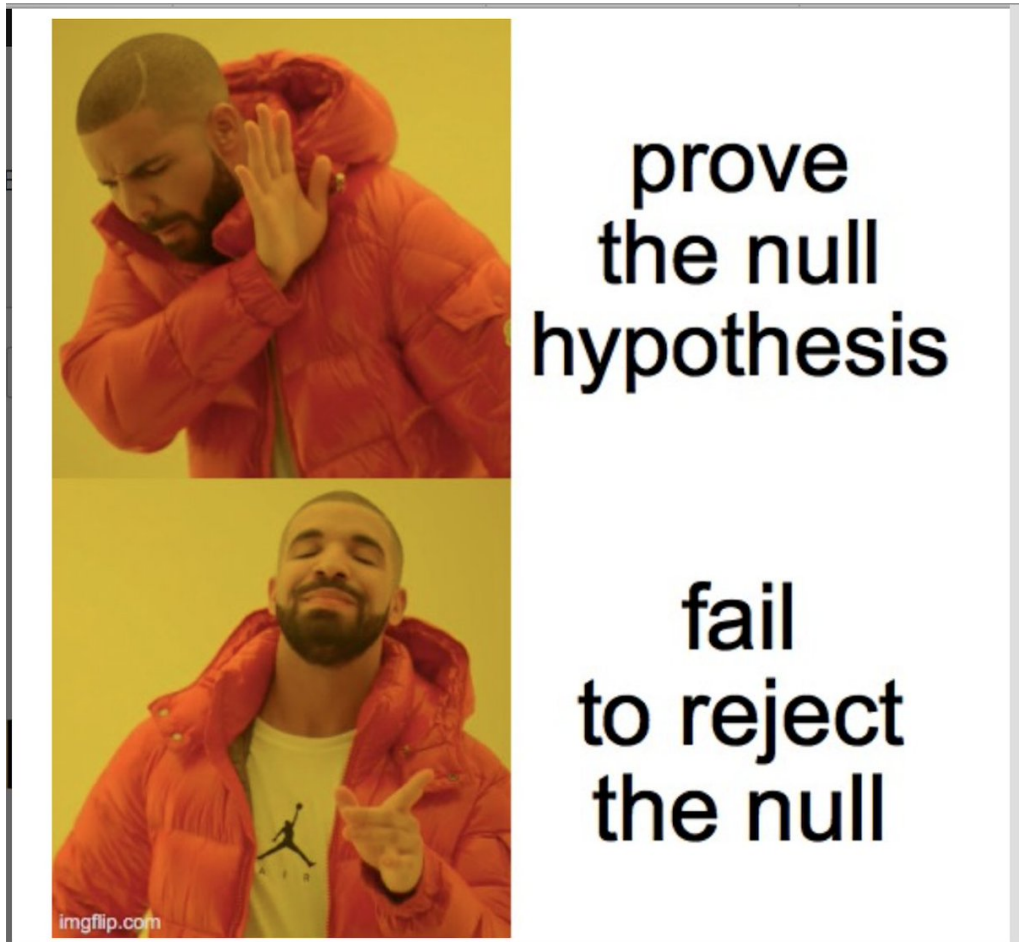
STA220H1: The Practice of Statistics I

Elena Tuzhilina

March 28, 2023

Please turn on your videos :)



Announcements

1. We have two Quizzes left before the final exam.
2. The Midterm 2 grades will be released in a couple of days, you will have one week to submit your regrade request (we will use email again).
3. Final exam is on April 20 at 3-5 pm in MY.
4. We will hold additional office hours before the final exam.

Agenda for today

- Recap: testing and connection to confidence intervals, power, type I and II error
- Statistical testing for two samples: matching and non-matching pairs

Statistical testing and confidence intervals

There is a connection between statistical testing and CI.

$$CI = \left[\bar{x} - t_{n-1}^{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{n-1}^{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$$

- If CI does not cover μ_0, then we can reject $H_0 : \mu = \mu_0$ in favor of $H_a : \mu \neq \mu_0$ at significance level α
- If CI covers μ_0, we do not have enough evidence to reject H_0 at significance level α

Example

If 90% confidence interval is [2, 10] what can we say about the hypotheses $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$?

If 90% confidence interval is [2, 10] what can we say about the hypotheses $H_0 : \mu = 5$ vs. $H_a : \mu \neq 5$?

Two-sided confidence interval

Two-sided confidence interval is $[a, b]$ that covers μ.

Standardization:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ approximately} \sim t_{n-1}$$

Distribution table:

$$P\left(-t_{n-1}^{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1}^{\alpha/2}\right) = 1 - \alpha$$

With probability $1 - \alpha$, the population parameter μ belongs to

$$\left[\bar{X} - t_{n-1}^{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1}^{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

One-sided confidence interval

Upper one-sided confidence interval is $[a, +\infty)$ that covers μ.

Standardization:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ approximately} \sim t_{n-1}$$

Distribution table:

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1}^{\alpha}\right) = 1 - \alpha$$

With probability $1 - \alpha$, the population parameter μ belongs to

$$\left[\bar{X} - t_{n-1}^{\alpha}\frac{S}{\sqrt{n}}, \ \ +\infty\right)$$

One-sided confidence interval

Lower one-sided confidence interval is $(-\infty, a]$ that covers μ.

Standardization:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ approximately} \sim t_{n-1}$$

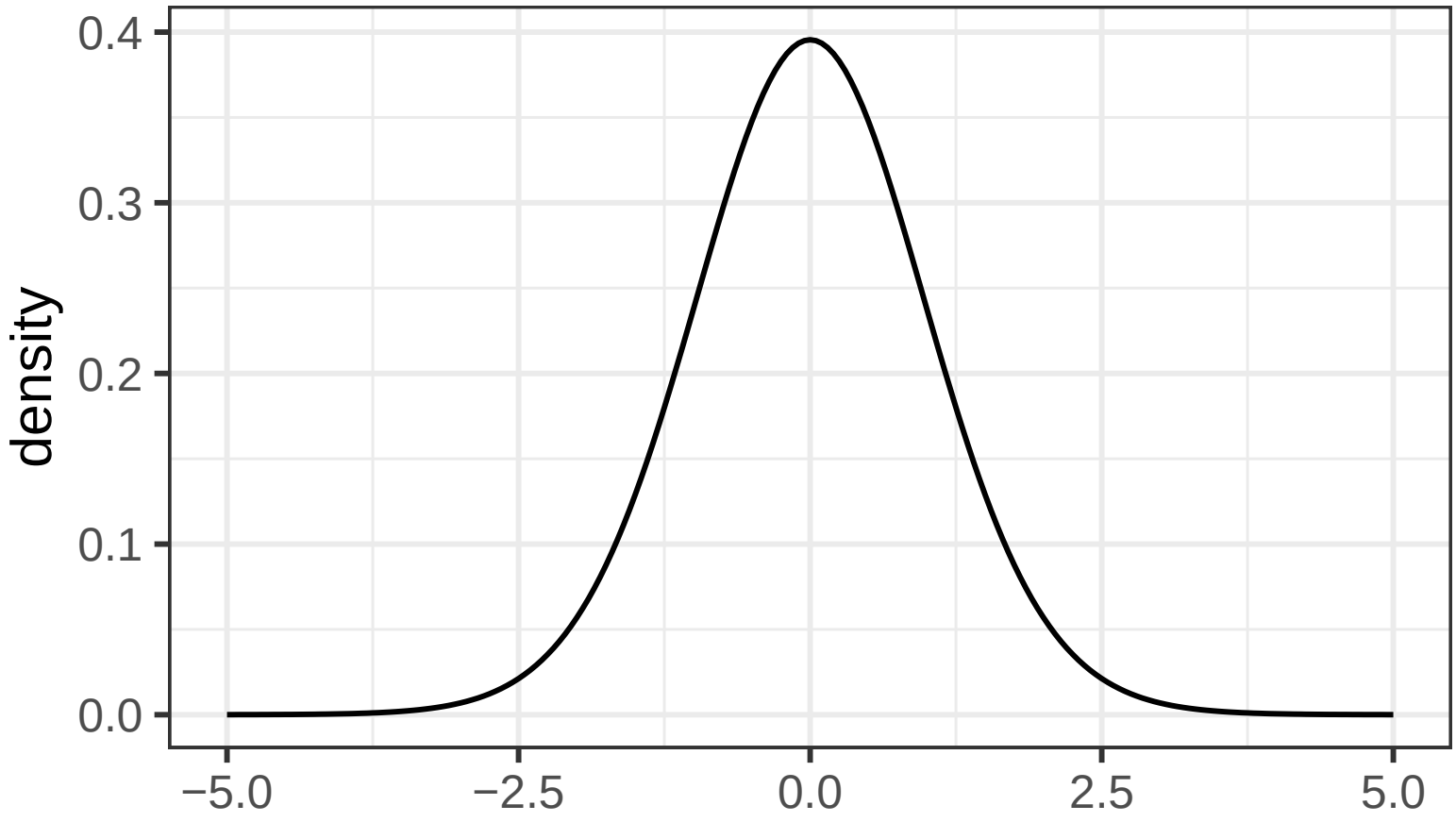
Distribution table:

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \geq -t_{n-1}^{\alpha}\right) = 1 - \alpha$$

With probability $1 - \alpha$, the population parameter μ belongs to

$$\left(-\infty,\ \bar{X} + t_{n-1}^{\alpha}\frac{S}{\sqrt{n}}\right]$$

One-sided vs two-sided quantiles



One-sided confidence intervals

To find **one-sided confidence interval**

- Use sample $x_1, \ldots, x_n$ to find $\bar{x}, s$
- Find quantile t^{α}_{n-1} that corresponds to the upper α-tail
- Compute $\left[\bar{x} - t^{\alpha}_{n-1}\frac{s}{\sqrt{n}},\ +\infty\right)$ or $\left(-\infty,\ \bar{x} + t^{\alpha}_{n-1}\frac{s}{\sqrt{n}}\right]$

What if σ^2 was known?

Statistical testing and confidence intervals

There is a connection between one-sided statistical testing and one-sided CI.

$$CI = \left[\bar{x} - t_{n-1}^{\alpha}\frac{s}{\sqrt{n}}, \ \ +\infty\right)$$

- If CI does not cover μ_0, then we can reject $H_0 : \mu = \mu_0$ in favor of $H_a : \mu > \mu_0$ at significance level α

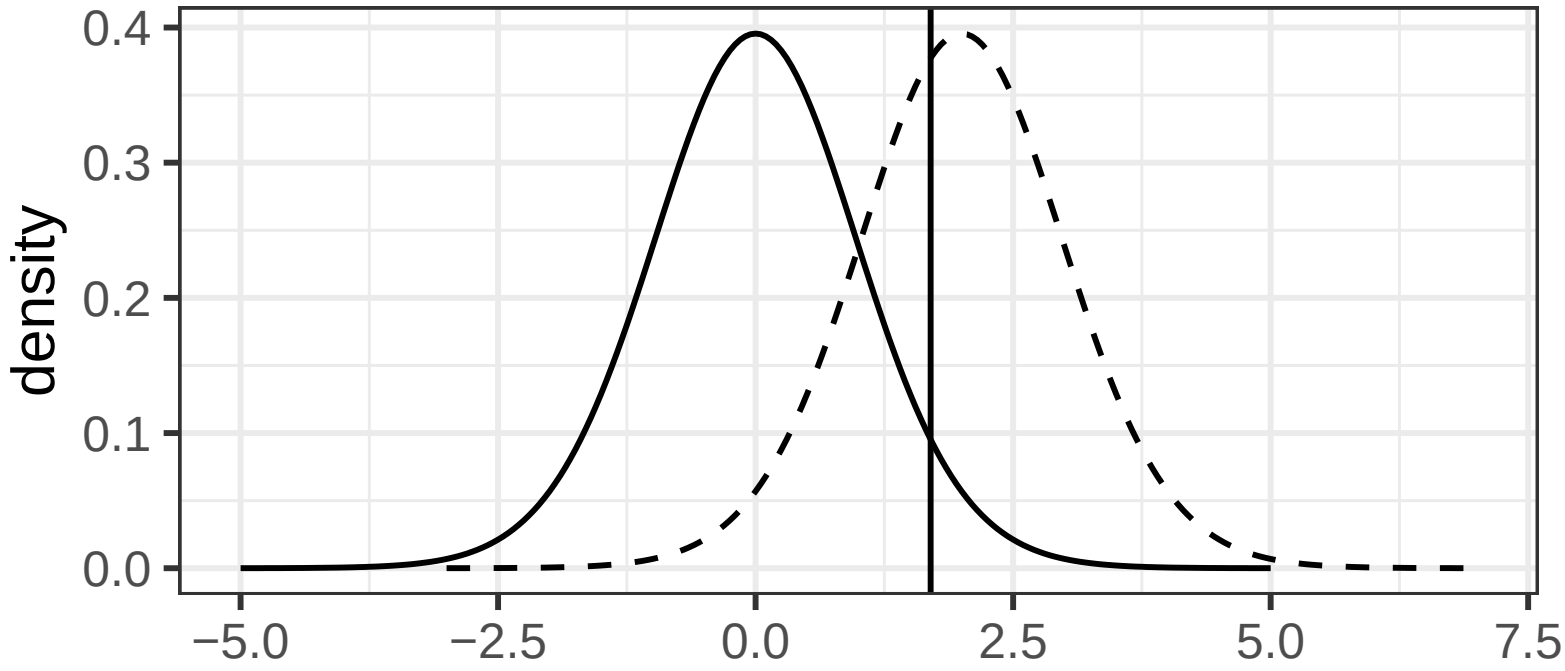
$$CI = \left(-\infty, \ \bar{x} + t_{n-1}^{\alpha}\frac{s}{\sqrt{n}}\right]$$

- If CI does not cover μ_0, then we can reject $H_0 : \mu = \mu_0$ in favor of $H_a : \mu < \mu_0$ at significance level α

Statistical testing: Type I type II errors and power

Type I error: we rejected H_0 when H_0 was true.

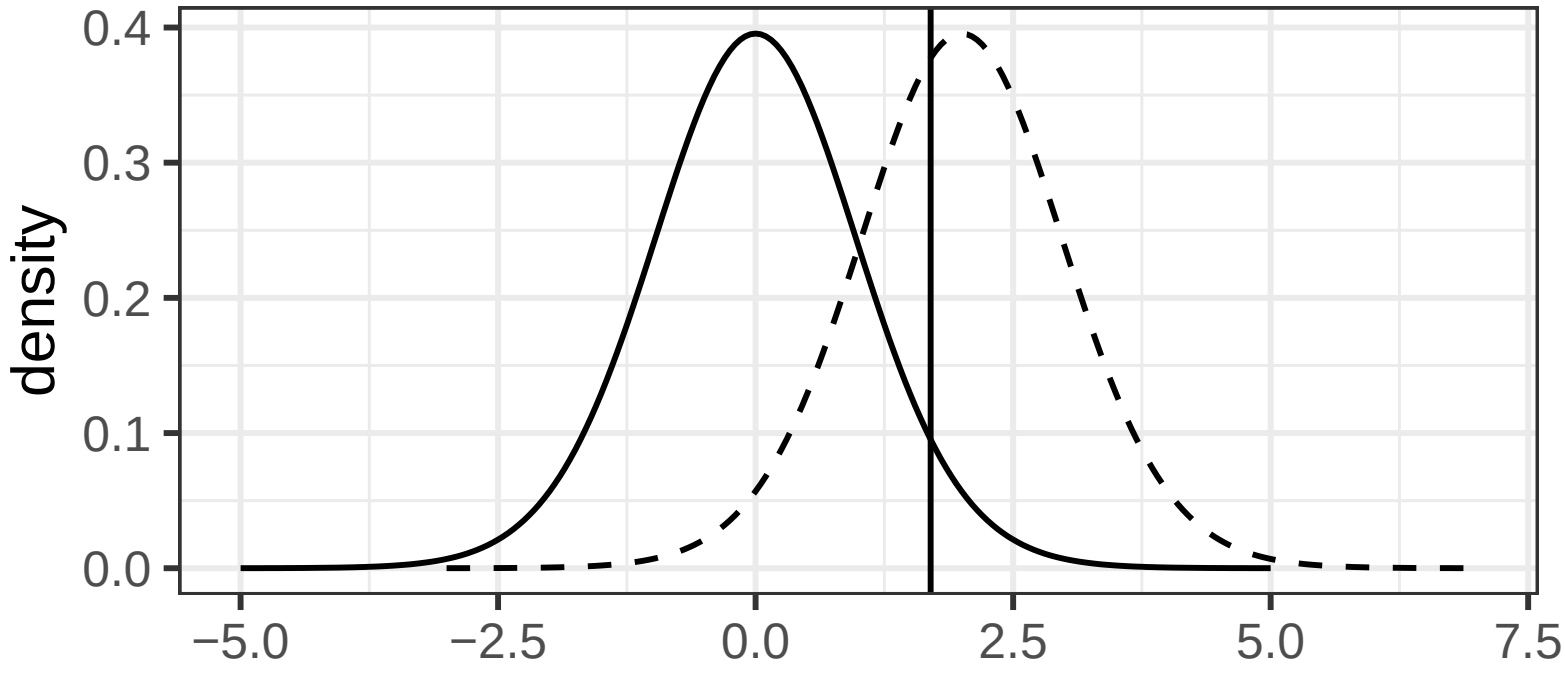
$$\alpha = P(\textit{reject } H_0 | H_0 \textit{ is true})$$



Statistical testing: Type I type II errors and power

Type II error: we failed to reject H_0 when H_a was true.

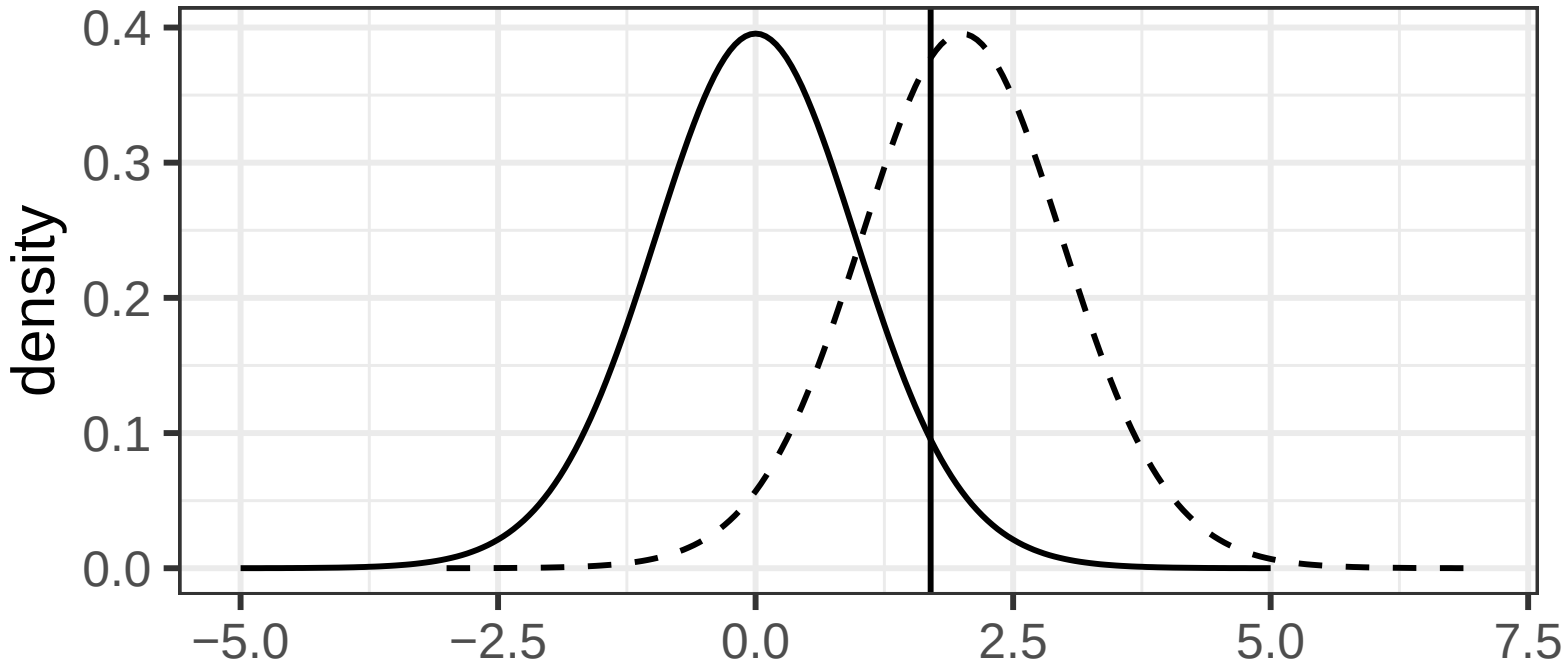
$$\beta = P(\textit{fail to reject } H_0 | H_a \textit{ is true})$$



Statistical testing: Type I type II errors and power

Power: chances to correctly reject H_0 when H_a is true.

$$1 - \beta = P(\textit{reject } H_0 | H_a \textit{ is true})$$



Statistical testing for two groups: mathching samples

Paired t-test: compare two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ with **matching observations**.

- Create a sample that shows the difference in measurements:

$$d_1, \ldots, d_n \text{ where } d_i = x_i - y_i$$

- Perform statistical test on differences testing $H_0 : \mu_d = 0$ vs $H_a : \mu_d \neq 0$.

Example

30 chickens were fed with sunflower seeds for 1 month. Their weight gain (in gramms) was recorded.

The same chickens were fed with corn for 1 month. The new weight gain was recorded.

Is any diet better for the weight gain?

chicken	diet1	diet2	difference
1	248.4738	229.5219	18.95187
2	213.6821	272.2564	-58.57428
3	285.9187	227.6527	58.26604
4	160.2935	310.4249	-150.13137
5	140.8080	219.8894	-79.08139
6	249.5029	201.6018	47.90115

Statistical testing for two groups: mathching samples

To test $H_0 : \mu_d = 0$ vs $H_a : \mu_d \neq 0$ you need to

- Compute sample mean $\bar{d}$ and sample standard deviations s_d for the differences
- Compute $t_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}}$
- Get p-value from the table and interpret the results

This requires normal approximation!

Statistical testing for two groups: mathching samples

- $X_1, \ldots, X_n$ is the outcome for diet 1
- $Y_1, \ldots, Y_n$ is the outcome for diet 2

Then for differences $D_i = X_i - Y_i$, t-test requires that

$$\bar{D} \sim Normal(\mu_d, \sigma_d^2)$$

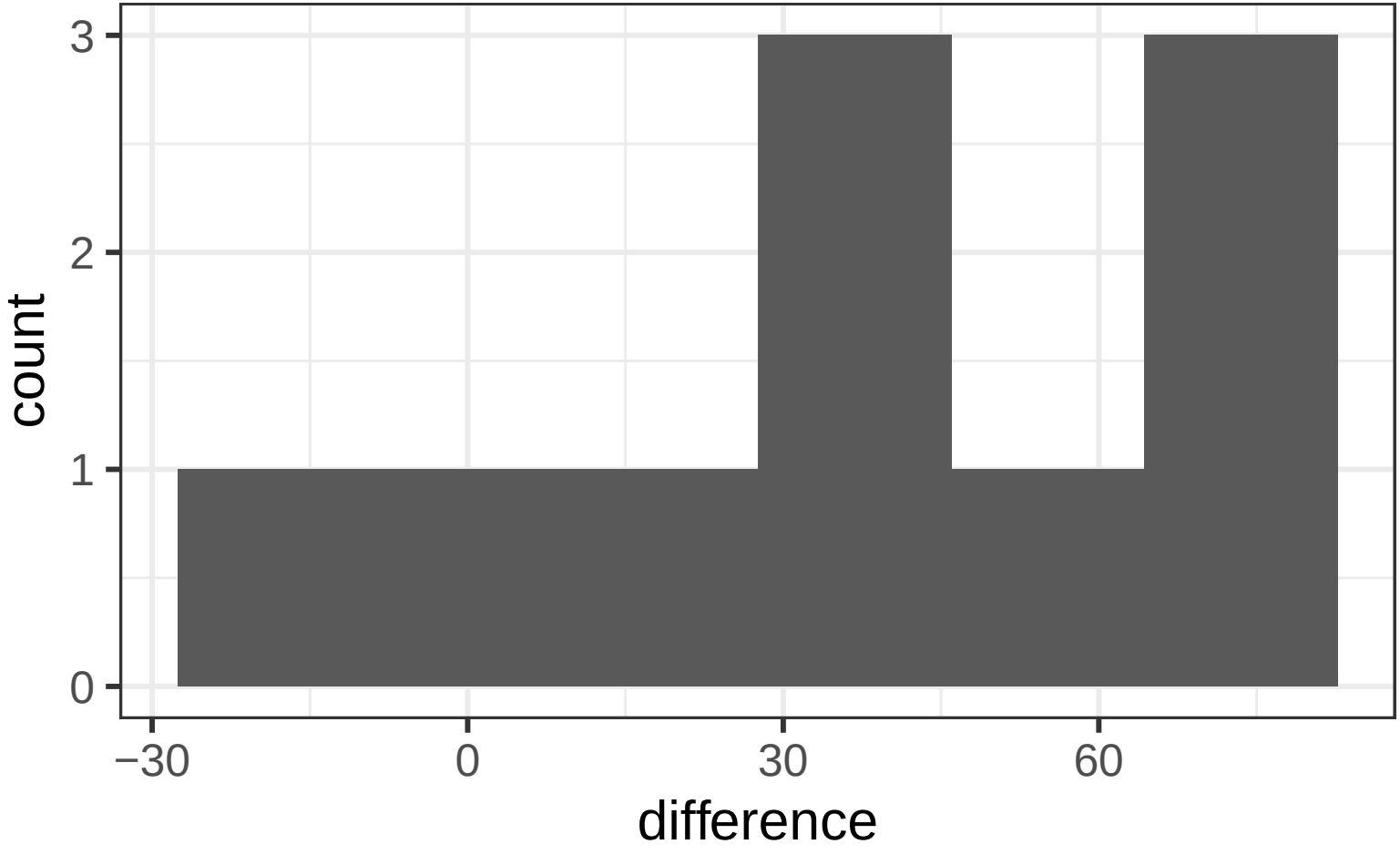
which is true when

- n is large (CLT)
- D_i are normal

Can we compare matching samples if n is small and differences are not normal?

Statistical testing for two groups: mathching samples

Assume we collected the data only for 10 chickens.



Statistical testing for two groups: mathching samples

Let's look at the **sign of the differences**.

chicken	diet1	diet2	difference	sign
1	183.8958	158.2390	25.656797	1
2	193.4912	157.0623	36.428881	1
3	211.5568	177.4809	34.075889	1
4	241.7387	165.3641	76.374552	1
5	178.1514	180.7937	-2.642283	-1
6	240.8551	169.9080	70.947102	1
7	245.0208	178.7047	66.316034	1
8	219.4718	189.6762	29.795558	1
9	216.6203	165.2014	51.418857	1
10	165.5608	181.0978	-15.537045	-1

Statistical testing for two groups: mathching samples

Signed test compares how often the difference is positive (or negative).

- New random variable $S_i = 1$ if $D_i \geq 0$ and $S_i = 0$ if $D_i < 0$
- Then $S_1, \ldots, S_n \sim Bernoulli(p)$
- If $\mu_d = 0$ then $p = 0.5$

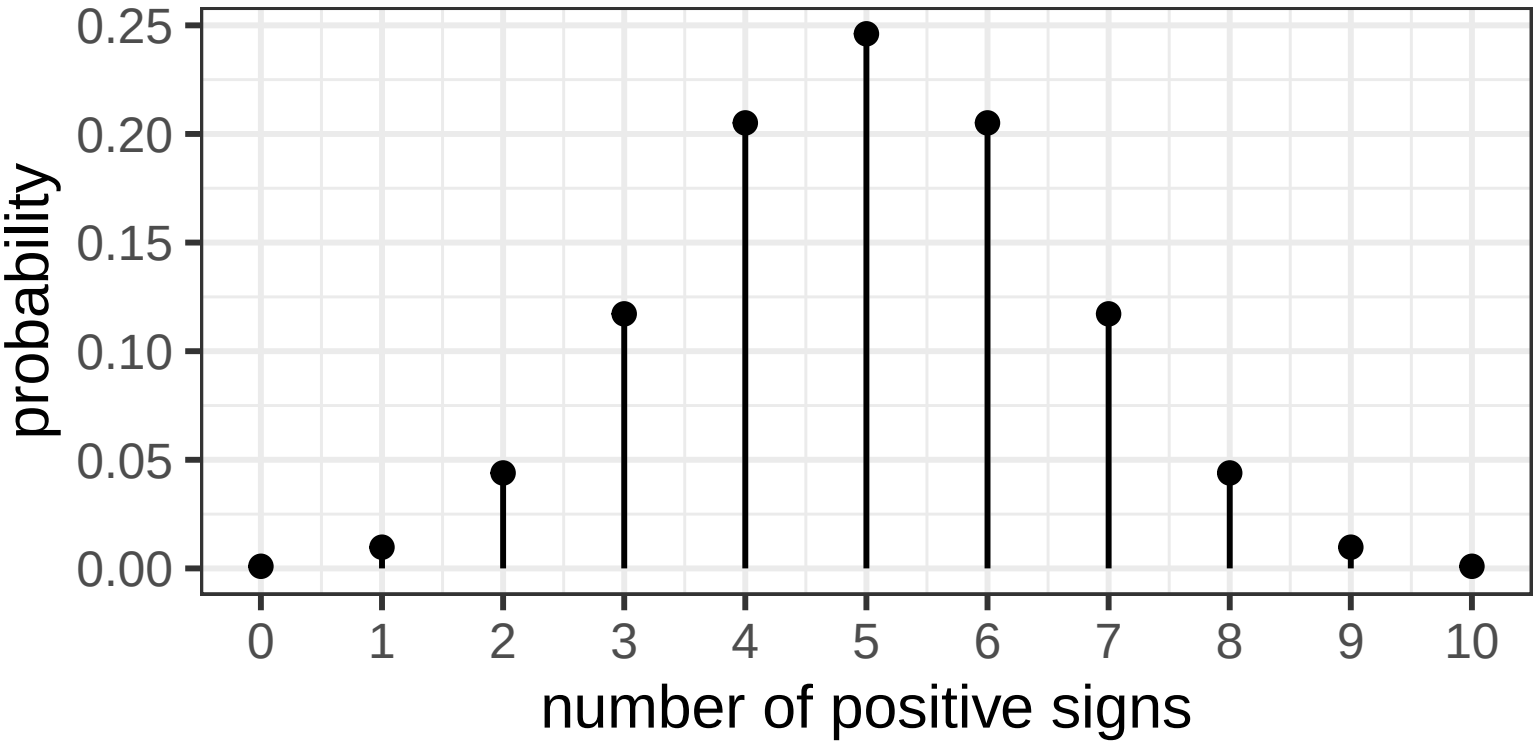
Signed test tests the hypotheses $H_0 : p = 0.5$ vs $H_a : p \neq 0.5$

- The observed statistic is $N = \sum_{i=1}^{n} S_i$, i.e. the **number of positive signs**
- The null distribution is $N \sim Bernoulli(n, 0.5)$

Statistical testing for two groups: mathching samples

We observe 8/10 positive differences.

$p\text{-}value = P(N \geq 8) + P(N \leq 2)$



Non-matching samples: proportions

50 patients received Moderna vaccine. The hospital recorded if they got COVID during the following 6 months.

```
##  [1] 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0
```

Other 30 patients received Pfizer vaccine and were monitored during the following 6 months

```
##  [1] 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0
```

Can we say if any vaccine is more efficient against COVID?

Non-matching samples: proportions

Let's compare the proportions of patients that got COVID.

```
table(moderna)
```

```
## moderna
##  0  1
## 42  8
```

```
table(pfizer)
```

```
## pfizer
##  0  1
## 26  4
```

Non-matching samples: proportions

We are given two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ with **non-matching observations**.

- $X_1, \ldots, X_n \sim Bernoulli(p)$ is the outcome for each Moderna recipient
- $Y_1, \ldots, Y_m \sim Bernoulli(q)$ is the outcome for each Pfizer recipient

From CLT and properties of normal distribution

$$\bar{X} - \bar{Y} \text{ approximately} \sim N\left(p - q, \frac{p(1-p)}{n} + \frac{q(1-q)}{m}\right)$$

Non-matching samples: confidence intervals for proportions

Standardization:

$$Z = \frac{(\bar{X} - \bar{Y}) - (p - q)}{\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}} \text{ approximately} \sim N(0, 1)$$

With probability $1 - \alpha$ the difference is $p - q$ in

$$(\bar{X} - \bar{Y}) \pm z^{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}$$

Confidence interval for $p - q$:

$$(\bar{x} - \bar{y}) \pm z^{\alpha/2} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n} + \frac{\bar{y}(1-\bar{y})}{m}}$$

Exercise

Find 95% confidence interval for $p - q$ in the vaccine example.

Statistical testing for two groups: non-matching samples

Do we observe significant difference in $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_m$*?*

Ultimate goal: test hypotheses $H_0 : p = q$ vs. $H_a : p \neq q$

Statistical testing for two groups: non-matching samples

Test hypotheses $H_0 : p = q$ vs. $H_a : p \neq q$

- Test statistic is constructed under the null hypothesis
- Need to use both x and y samples to find a common estimate for p and q

Idea: "pool" two samples into $x_1, \ldots, x_n, y_1, \ldots, y_m$ and approximate both by $p, q \approx \frac{n\bar{x}+m\bar{y}}{n+m}$

Statistical testing for two groups: non-matching samples

To test hypotheses $H_0 : p = q$ vs. $H_a : p \neq q$ use test statistic

$$z_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{n\bar{x}+m\bar{y}}{n+m}\left(1 - \frac{n\bar{x}+m\bar{y}}{n+m}\right)\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Then, use normal table to find p-value and draw the conclusion...

Exercise

Use statistical testing to check if there is a difference between Prizer and Moderna vaccines. What will change if we want to check that Pfizer has higher success rate than Moderna?

Non-matching samples: means

A normal pregnancy can range from 38 to 42 weeks.

50 pregnant patients got COVID in the first trimester. The hospital recorded the gestational age of each born baby (in weeks).

```
## [1] 36.3 36.9 37.9 39.5 36.0 39.5
```

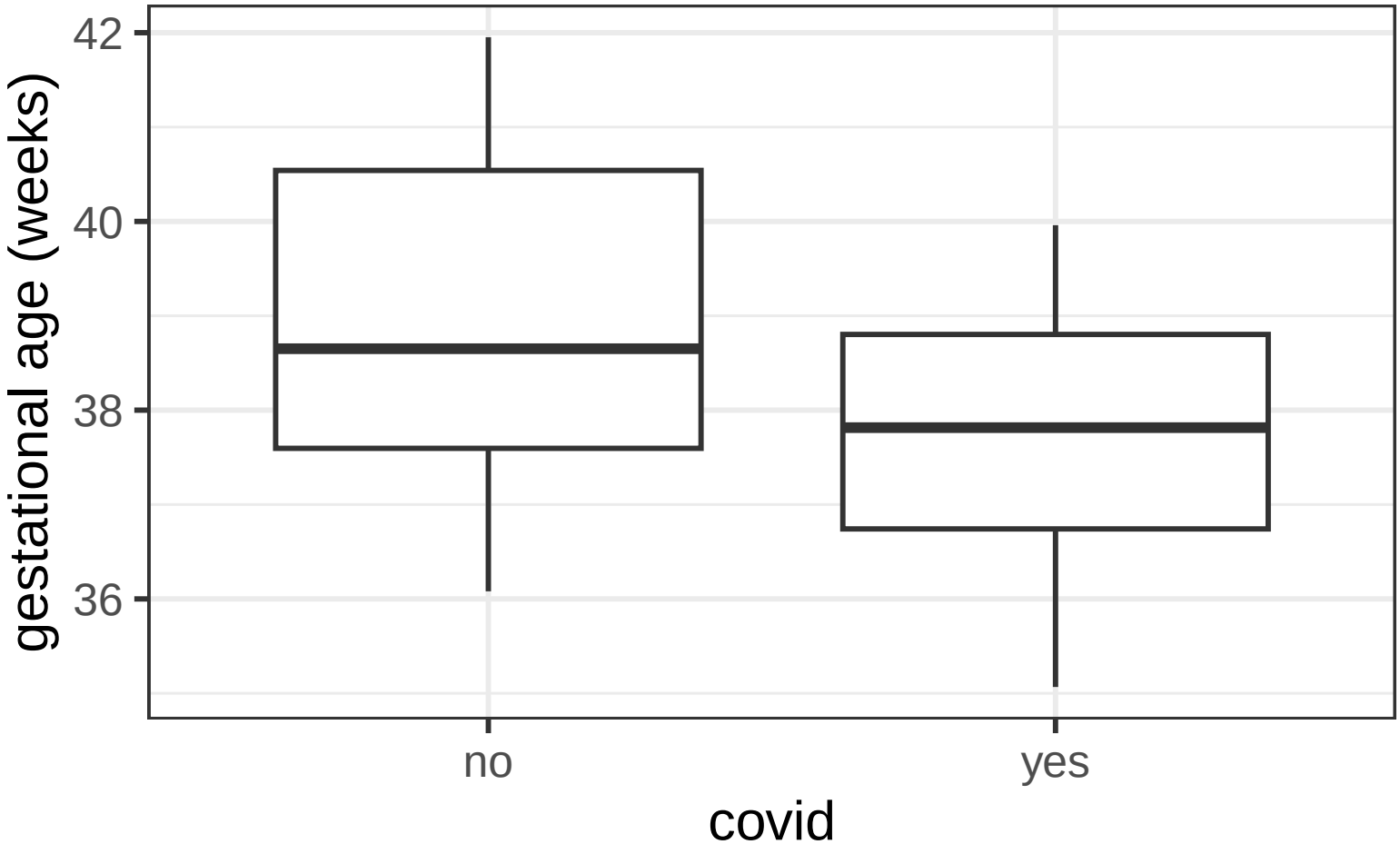
Other 30 pregnant patients did not get COVID in the first trimester.

```
## [1] 37.6 38.2 39.4 41.4 37.2 41.4
```

Does COVID lead to a preterm birth?

Non-matching samples: means

Let's also compare the boxplots.



Does COVID lead to a preterm birth?

Non-matching samples: means

- $X_1, \ldots, X_n \sim (\mu_x, \sigma_x^2)$ is the outcome patients with COVID
- $Y_1, \ldots, Y_m \sim (\mu_y, \sigma_y^2)$ is the outcome patients without COVID

From CLT and properties of normal distribution

$$\bar{X} - \bar{Y} \text{ approximately} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

Non-matching samples: means

Standardization:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \text{ approximately} \sim N(0, 1)$$

But, we do not know σ_x and σ_y

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \text{ approximately} \sim t_{df}$$

With probability $1 - \alpha$ the difference is $\mu_x - \mu_y$ in

$$(\bar{X} - \bar{Y}) \pm t_{df}^{\alpha/2} \cdot \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

Non-matching samples: confidence intervals for means

Confidence interval for $\mu_x - \mu_y$:

$$(\bar{x} - \bar{y}) \pm t_{df}^{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

Well, this is not the end of the story! You need to estimate df . . . :(

Non-matching samples: confidence intervals for means

Confidence interval for $\mu_x - \mu_y$:

$$(\bar{x} - \bar{y}) \pm t_{df}^{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

"Pooled" degrees of freedom are estimated as

$$df = \frac{\left(s_x^2/n + s_y^2/m\right)^2}{\frac{\left(s_x^2/n\right)^2}{n-1} + \frac{\left(s_y^2/m\right)^2}{m-1}}$$

Exercise

Find *df* for the pregnancy data example.

```
sd(nocovid)
```

```
## [1] 1.77168
```

```
sd(covid)
```

```
## [1] 1.361195
```

Find 95 confidence interval for $\mu_x - \mu_y$.

```
mean(nocovid)
```

```
## [1] 39.05731
```

```
mean(covid)
```

```
## [1] 37.66296
```

Statistical testing for two groups: non-matching samples

Do we observe significant difference in $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_m$*?*

To test hypotheses $H_0 : \mu_x = \mu_y$ vs. $H_a : \mu_x \neq \mu_y$

- Use test statistic

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

- Compute *df* from the ugly formula
- Compute p-value from the t-distribution table

Exercise

Use statistical testing to check if there is a difference in gestational ages of babies for COVID and non-COVID patients. What will change if we want to check that COVID leads to a premature birth?

Exercise

```
t.test(covid, nocovid, alternative = "two.sided", mu = 0,
       paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  covid and nocovid
## t = -3.7043, df = 49.505, p-value = 0.0005341
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -2.1505816 -0.6381185
## sample estimates:
## mean of x mean of y
##  37.66296  39.05731
```

Statistical testing for two groups: non-matching samples

Can we simplify computations for statistical testing by "pooling" the samples?

Yes, but only if $\sigma_x = \sigma_y$.

Idea: "pool" two samples into $x_1, \ldots, x_n, y_1, \ldots, y_m$ and approximate s_x^2, s_y^2 by

$$s^2 \approx \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Then

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

with $df = n + m - 2$.

Exercise

Assuming that variances for two samples are equal, use statistical testing to check if there is a difference in gestational ages of babies for COVID and non-COVID patients.

Exercise

```
t.test(covid, nocovid, alternative = "two.sided", mu = 0,
       paired = FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  covid and nocovid
## t = -3.9546, df = 78, p-value = 0.0001676
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -2.0963005 -0.6923995
## sample estimates:
## mean of x mean of y
##  37.66296  39.05731
```

TO DO

1. Module 10. Comparing Two Groups
2. Quiz 11 due Monday (April 3) @ 11:59 PM (EST)
3. Practice Problem Set 11