

Advanced Statistics

DS2003 (BDS-4A)

Lecture 13

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

05 April, 2022

Previous Lecture

- Chi-Square test of Independence
 - Example: popular dataset, student goals and grades (years), are they independent?
 - Example: Asthma and smoking, across different continents

Example: Asthma and Smoking

$$\chi^2 = \frac{(339-258)^2}{258} + \frac{(33-59.46)^2}{59.46} + \frac{(61-88.28)^2}{88.28} + \frac{(34-61.26)^2}{61.26} + \frac{(377-458)^2}{458} + \frac{(132-105.54)^2}{105.54} + \frac{(184-156.72)^2}{156.72} + \frac{(136-108.74)^2}{108.74} = 90.2987$$

Exact p-Value (using R on <https://rdr.io/snippets/>)

```
t2stat = pchisq(q = 90.2987, df = 3, lower.tail = FALSE)
print(t2stat)
```

OUTPUT: **1.889728e-19** **→** we can reject the H_0 **→** p-value much smaller than 0.05

One-sample mean with the t-distribution

Friday the 13th

- Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and hospital admissions on Friday 13th and the previous Friday, Friday 6th. Below is an excerpt from this data set on traffic flow. We can assume that traffic flow on given day at locations 1 and 2 are independent.

	type	date	6 th	13 th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Friday the 13th

- We want to investigate if people's behavior is different on Friday the 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6th and 13th are equal.
 H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

No!

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

A. $H_0: \mu_{6th} = \mu_{13th}$

$H_A: \mu_{6th} \neq \mu_{13th}$

B. $H_0: p_{6th} = p_{13th}$

$H_A: p_{6th} \neq p_{13th}$

C. $H_0: \mu_{diff} = 0$

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

D. $H_0: \bar{x}_{diff} = 0$

$H_A: \bar{x}_{diff} \neq 0$

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

A. $H_0: \mu_{6th} = \mu_{13th}$

$H_A: \mu_{6th} \neq \mu_{13th}$

B. $H_0: p_{6th} = p_{13th}$

$H_A: p_{6th} \neq p_{13th}$

C. $H_0: \mu_{diff} = 0$

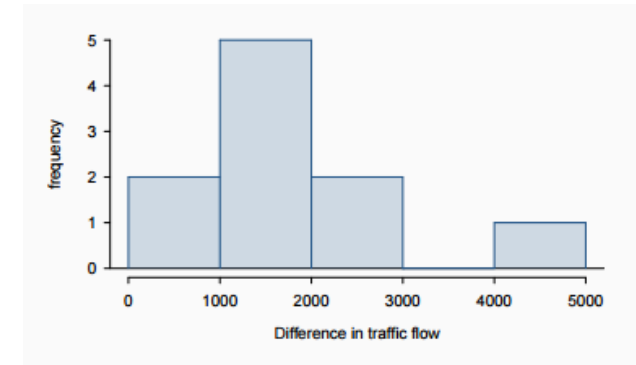
$H_A: \mu_{diff} \neq 0$

D. $H_0: \bar{x}_{diff} = 0$

$H_A: \bar{x}_{diff} \neq 0$

Conditions

- *Independence*: We are told to assume that cases (rows) are independent
 - *Sample size / skew*:
 - The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.
 - We do not know σ and n is too small to assume s is reliable estimate for σ
- So what do we do when the sample size is small?



Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that..

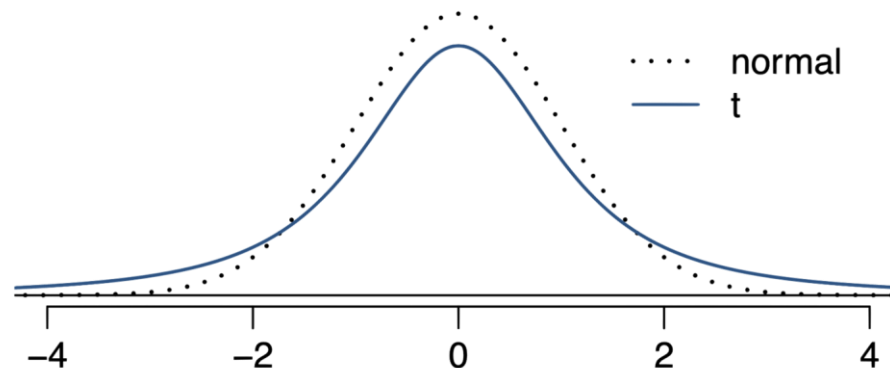
- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

- The CLT (central limit theorem), which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also to think about where the data come from
 - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

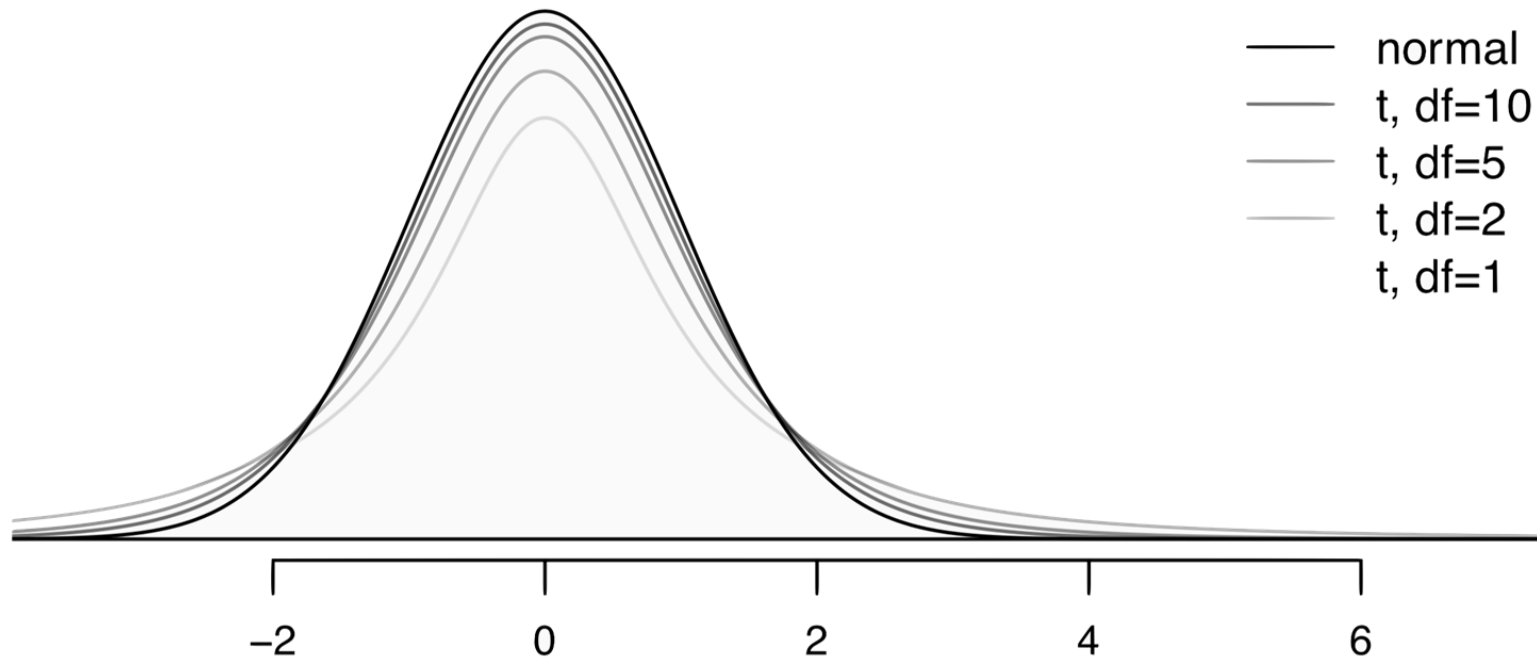
The t distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.
- This distribution also has a bell shape, but its tails are **thicker** than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution
- These extra thick tails are helpful for resolving our problem with a less reliable estimate of the standard error (since n is small)



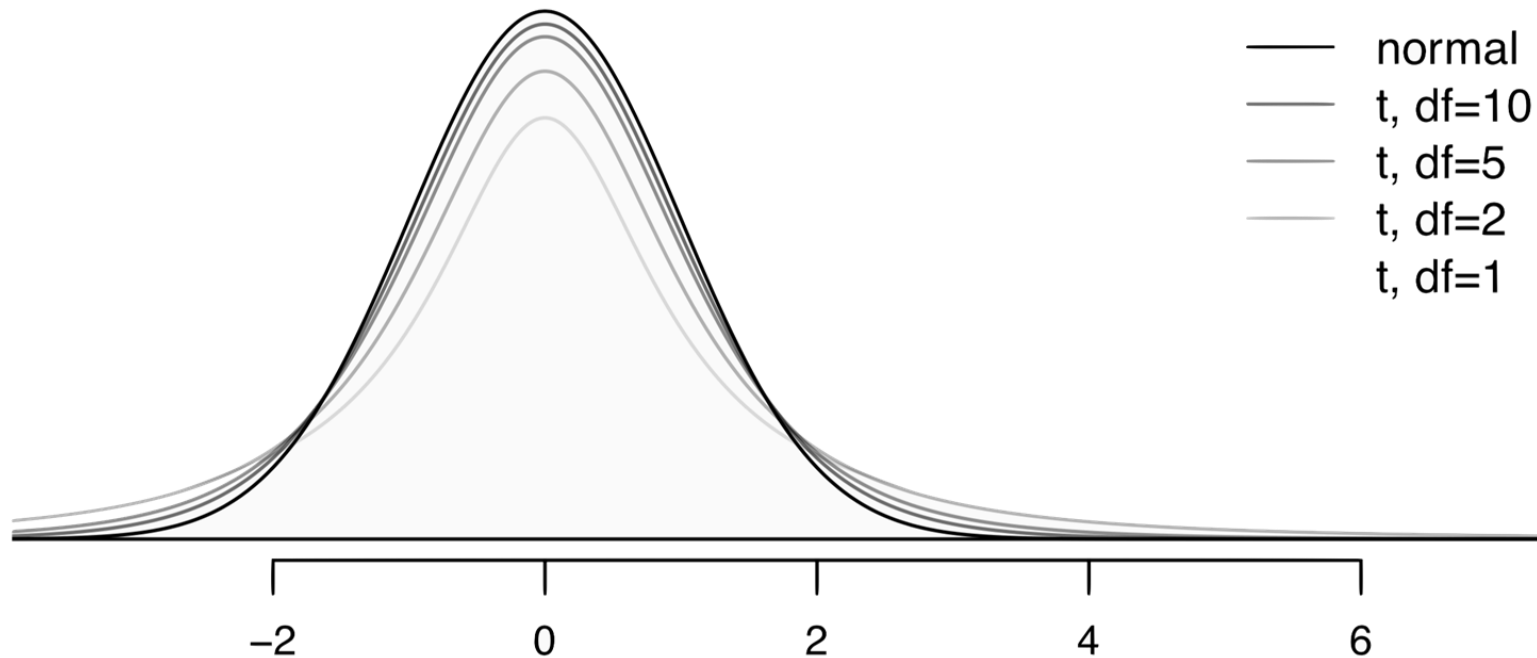
The t distribution (continued)

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (df).



The t distribution (continued)

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (df).



What happens to the shape of the t distribution as df increases?

Approaches normal

Back to Friday the 13th

	type	date	6 th	13 th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Back to Friday the 13th

	type	date	6 th	13 th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2


$$\bar{x}_{diff} = 1836$$

$$s_{diff} = 1176$$

Note: Null value is 0 because in the null hypothesis we set $\mu_{\text{diff}} = 0$

Find the test statistic

Test statistic for inference on a small sample mean

- The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

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

in context...

$$\text{point estimate} = \bar{x}_{\text{diff}} = 1836$$

$$SE = \frac{s_{\text{diff}}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372$$

$$T = \frac{1836 - 0}{372} = 4.94$$

$$df = 10 - 1 = 9$$

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution

- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)  
[1] 0.0008022394
```

- Using a web app:

http://gallery.shinyapps.io/dist_calc/

Complete R code

```
num_diff <- c(698, 1104, 1037, 1889, 1911, 2416, 2761, 4382, 1839, 321)
print(paste("Sample mean = ", mean(num_diff)))
print(paste("Standard deviation of sample = ", sd(num_diff)))
std_error = sd(num_diff)/sqrt(length(num_diff))
print(paste("Standard error of sample = ", std_error))
t_stat_diff = (mean(num_diff) - 0)/std_error
print(paste("T statistic = ", t_stat_diff))
print(paste("p-value = ", 2 * pt(t_stat_diff, df = length(num_diff)-1, lower.tail = FALSE)))
```

Complete R code (<https://rdrr.io/snippets>)

```
num_diff <- c(698, 1104, 1037, 1889, 1911, 2416, 2761, 4382, 1839, 321)
print(paste("Sample mean = ", mean(num_diff)))
print(paste("Standard deviation of sample = ", sd(num_diff)))
std_error = sd(num_diff)/sqrt(length(num_diff))
print(paste("Standard error of sample = ", std_error))
t_stat_diff = (mean(num_diff) - 0)/std_error
print(paste("T statistic = ", t_stat_diff))
print(paste("p-value = ", 2 * pt(t_stat_diff, df = length(num_diff)-1, lower.tail = FALSE)))
```

Output:

```
[1] "Sample mean = 1835.8"
[1] "Standard deviation of sample = 1176.01386991065"
[1] "Standard error of sample = 371.888238886661"
[1] "T statistic = 4.93642930331951"
[1] "p-value = 0.000806184359845231"
```

http://gallery.shinyapps.io/dist_calc/

Distribution Calculator

Distribution:

Degrees of freedom:

Model: $P(X < a \text{ or } X > b)$

Find Area:

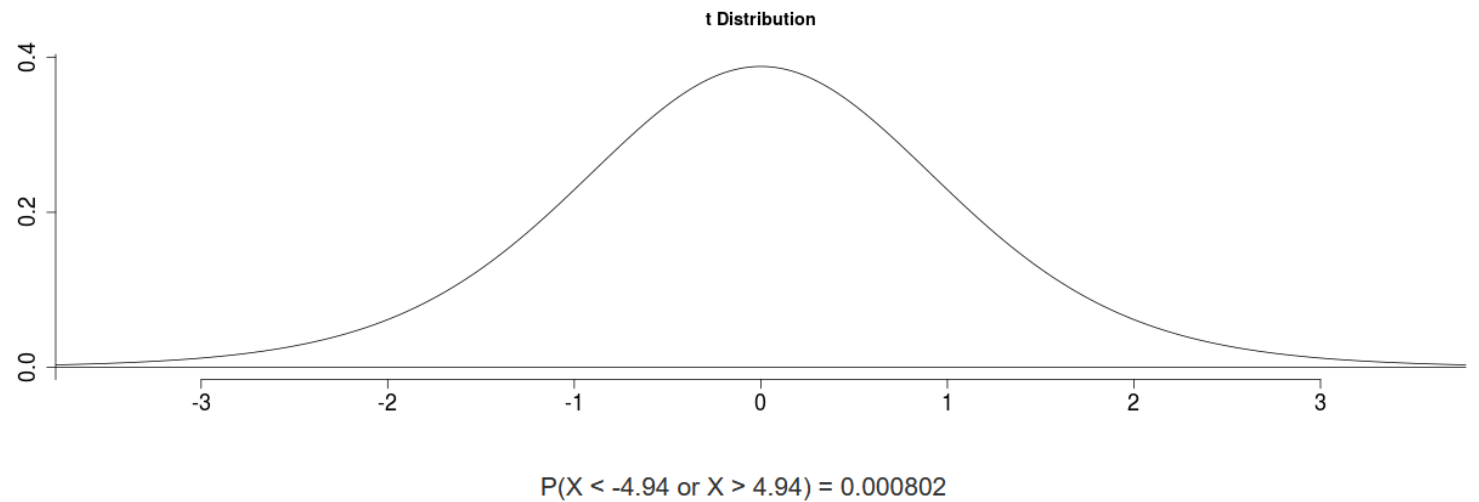
a:

b:

[View code](#)

[Check out other apps](#)

[Learn more for free!](#)



Conclusion of the test

What is the conclusion of this hypothesis test?

Since the p-value is quite low, we conclude that the data provide strong evidence of a difference between traffic flow on Friday 6th and 13th.

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6th and 13th
- But it would be more interesting to find out what exactly this difference is
- We can use a confidence interval to estimate this difference

Confidence interval for a small sample mean

- Confidence intervals are always of the form

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

- ME is always calculated as the product of a critical value and SE
- Since small sample means follow a ***t*** distribution (and not a *z* distribution), the critical value is a ***t**** (as opposed to a *z**?).

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

Finding the critical value (t^\star)

Using R (getting to t-statistic for a specific p-value):

```
> qt(p = 0.975, df = 9)  
[1] 2.262157
```


Here, we want to t-statistic for a two-tailed 95% confidence interval, hence we find the t-statistic that gives prob 0.975 for 1-tail

$$P(T_{df=9} < 2.26) = 0.975, \text{ thus } \rightarrow P(-2.26 < T_{df=9} < 2.26) = 0.95$$

Constructing a CI for a small sample mean

- Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \quad s_{diff} = 1176 \quad n = 10 \quad SE = 372$$

- A. $1836 \pm 1.96 \times 372$
- B. $1836 \pm 2.26 \times 372$  (995, 2677)
- C. $1836 \pm -2.26 \times 372$
- D. $1836 \pm 2.26 \times 1176$

Interpreting the CI

- Which of the following is the *best* interpretation for the confidence interval we just calculated?

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that...

- A. the difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2,677
- B. on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average
- C. on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average
- D. on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average

Interpreting the CI

- Which of the following is the *best* interpretation for the confidence interval we just calculated?

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that...

- A. the difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2,677
- B. on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average
- C. on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average
- D. on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average**

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

No, this is an observational study. We have just observed a significant difference between the number of cars on the road on these two days. We have not tested for people's beliefs

Imagine we have collected a random sample of 31 energy bars from a number of different stores to represent the population of energy bars available to the general consumer. The labels on the bars claim that each bar contains 20 grams of protein.

Energy Bar - Grams of Protein						
20.70	27.46	22.15	19.85	21.29	24.75	
20.75	22.91	25.34	20.33	21.54	21.08	
22.14	19.56	21.10	18.04	24.12	19.95	
19.72	18.28	16.26	17.46	20.53	22.12	
25.06	22.44	19.08	19.88	21.39	22.33	25.79

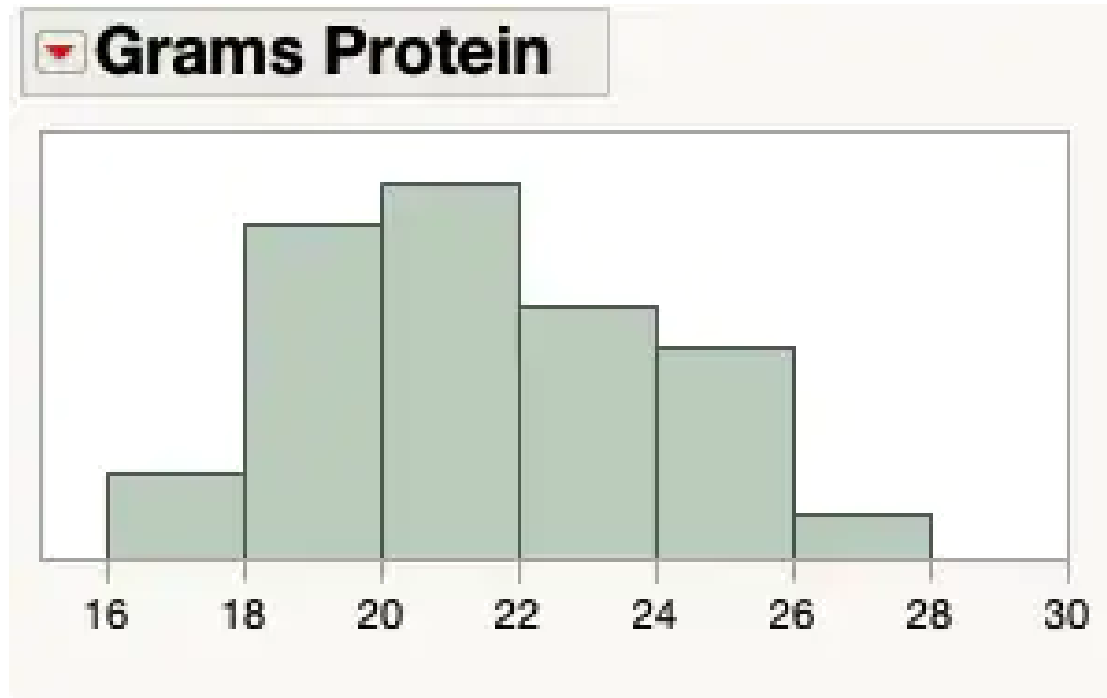


Figure 1: Histogram and summary statistics

Mean = 21.399

St_dev = 2.54186

St_error = $2.54186 / \sqrt{31} = 0.4565$

DF = $31 - 1 = 30$

R code:

```
t_stat = (21.399 - 20) / 0.4565  
print(2 * pt(t_stat, df = 30, lower.tail = FALSE))
```

Output:

0.004578209 (p-value indicates we can reject the null hypothesis, it is less than 0.05)

Reject the claim that the energy bar contains 20 grams of protein

Sources

- openintro.org/os (Chapter 7, Section 7.1)

Helpful Links (jbstatistics on YouTube):