Inference for one mean

Stat 120

May 10 2023

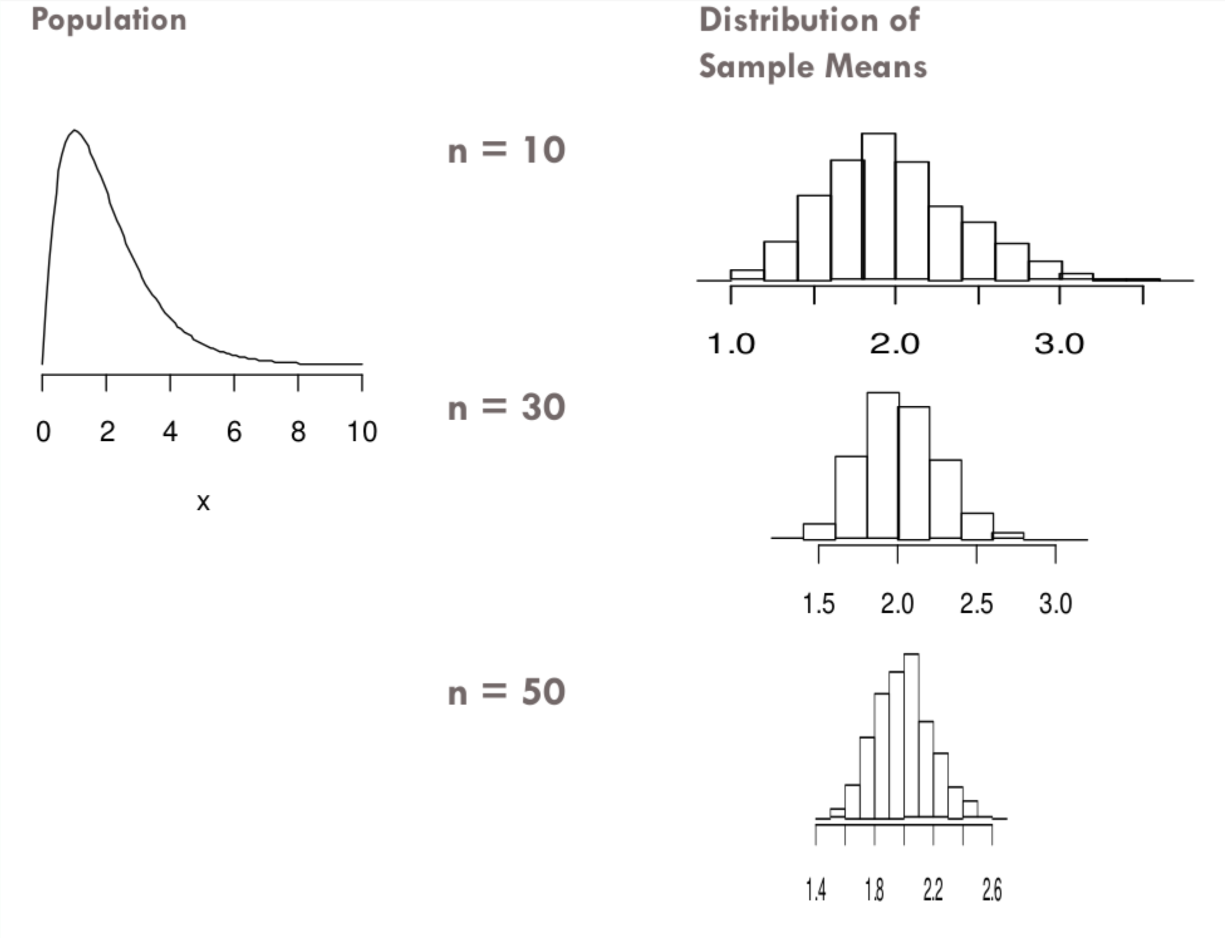
CLT for a Mean

If $n \geq 30^*$, then

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

*Smaller sample sizes may be sufficient for symmetric distributions, and 30 may not be sufficient for very skewed distributions or distributions with high outliers

CLT for mean



Population
Distribution of Sample Means
n = 10
n = 30
n = 50
0 2 4 6 8 10
x
1.0 2.0 3.0
1.5 2.0 2.5 3.0
1.4 1.8 2.2 2.6

Standard Deviation

The standard deviation of the population is

a) σ

b) s

c) $\frac{\sigma}{\sqrt{n}}$

▶ Click for answer

Standard Deviation

The standard deviation of the sample is

a) σ

b) s

c) $\frac{\sigma}{\sqrt{n}}$

▶ Click for answer

Standard Deviation

The standard deviation of the sample mean is

a) σ

b) s

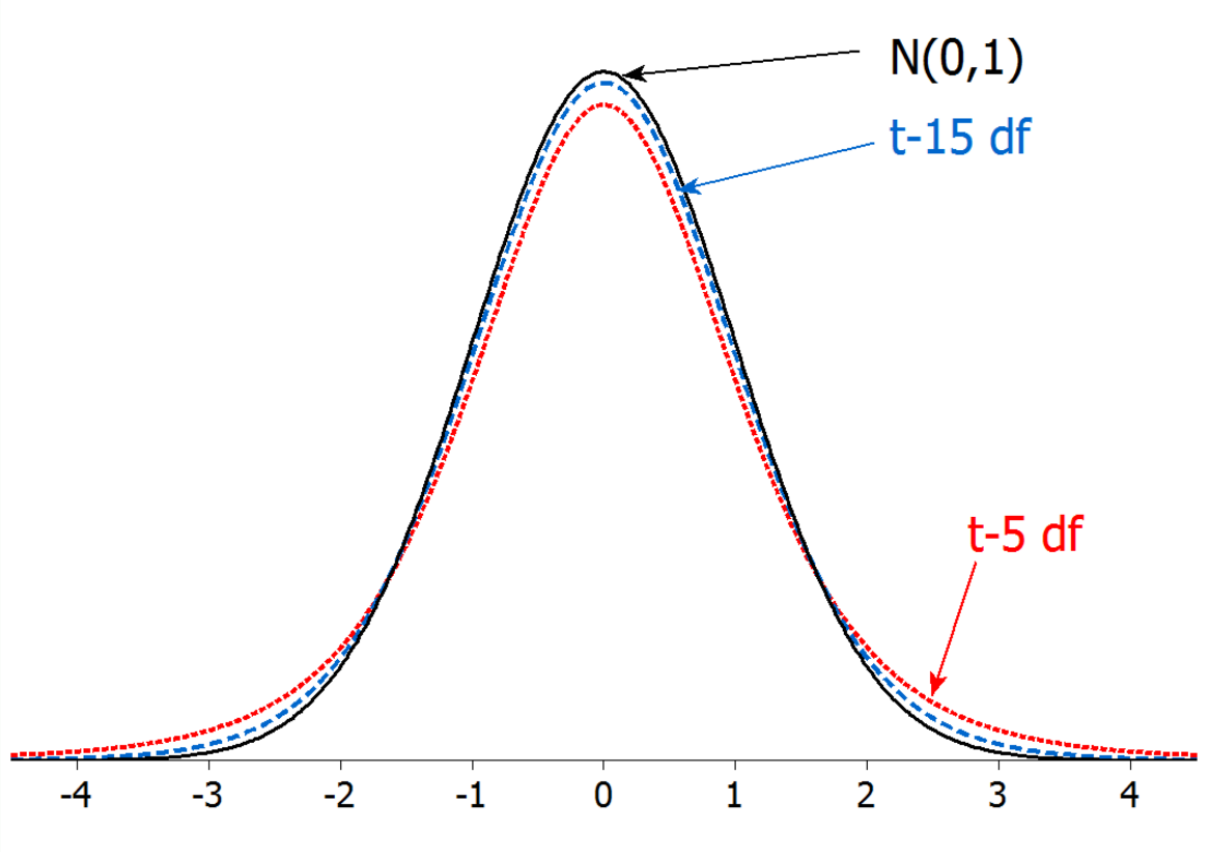
c) $\frac{\sigma}{\sqrt{n}}$

▶ Click for answer

T-distribution

- *Replacing σ with s changes the distribution of the z-statistic from a normal distribution to a t-distribution*
- *The t distribution is very similar to the standard normal, but with slightly fatter tails to reflect this added uncertainty*

T-distribution



Inference for means

Check: Check the one sample size conditions for the CLT

Tests: Use t-ratios of the form

$$t = \frac{\text{stat} - \text{null value}}{SE}$$

P-values computed from a t-distribution with appropriate df

- `pt(t, df= )` gives the area to the left of t

Confidence intervals: CI of the form

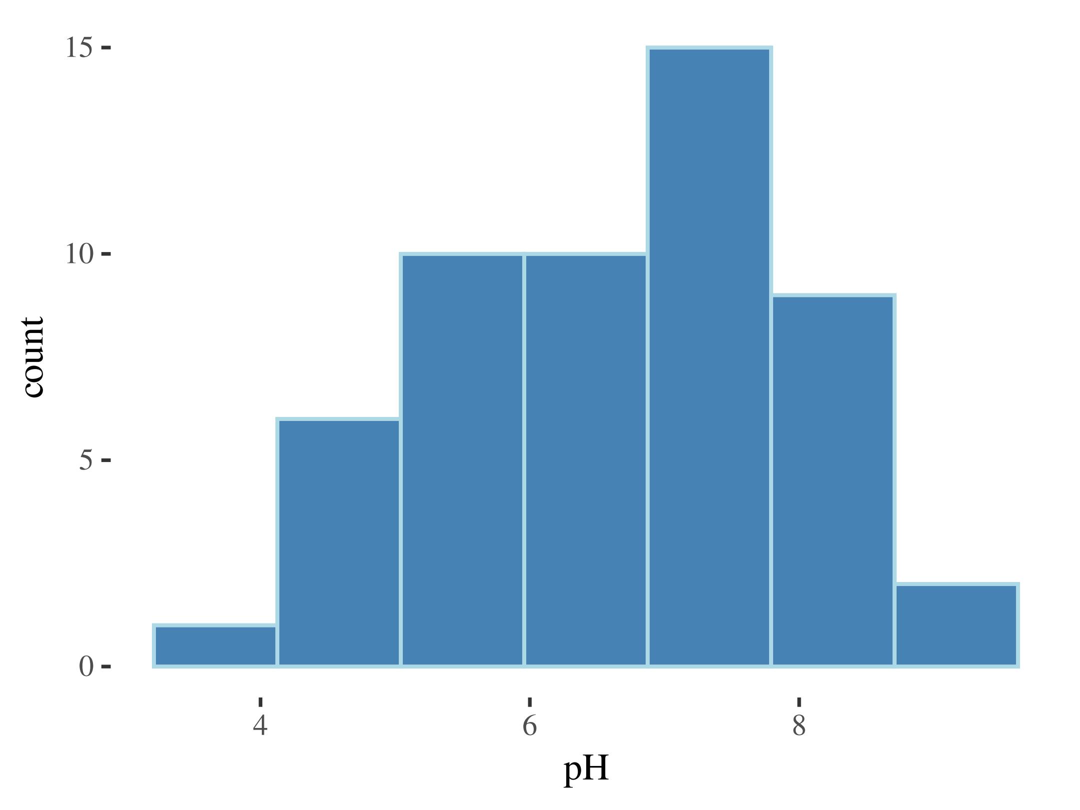
$$\text{stat} \pm t^* SE$$

The t^* multiplier comes from a t-distribution with appropriate df

- `qt(0.975, df= )` gives t^* for 95% confidence

Florida lakes

```
library(ggplot2)
lakes <- read.csv("http://www.lock5stat.com/datas
ggplot(lakes, aes(x = pH)) +
  geom_histogram(fill = "steelblue",
                 bins = 7,
                 col = "lightblue")
```



Florida lakes

$$H_0 : \mu = 7 \qquad H_A : \mu \neq 7$$

Data: The average pH was $\bar{x} = 6.591$ *with a standard deviation of* $s = 1.288$.

```
mean(lakes$pH)
[1] 6.590566
```

```
sd(lakes$pH)
[1] 1.288449
```

- *The t-test stat is*

$$t = \frac{6.591 - 7}{1.288/\sqrt{53}} \approx -2.31$$

- *Interpret t: The observed mean of 6.591 is 2.31 SEs below 7.*

Florida lakes

$$H_0 : \mu = 7 \qquad H_A : \mu \neq 7$$

- *p-value* $2 \times P(t < -2.31)$, *or double left tail area below -2.31*
  - *use t-distribution with* $df = 53 - 1 = 52$

```
2*pt(-2.31, df=53-1)  # df = n-1
[1] 0.02489032
```

- *Interpret:* *The p-value is 0.025. If the mean pH of all lakes is 7, then we would see a sample mean that is at least 2.31 SEs away from 7 about 2.5% of the time in samples of 53 lakes.*
- *Conclusion:* *There is a statistically significant difference between the observed mean pH of 6.591 and the hypothesized mean of 7 (t=-2.31, df=52, p=0.025).*

Florida lakes: `t.test` in R

- We can also use `t.test` in R !

```
t.test(lakes$pH, mu  = 7)

    One Sample t-test

data:  lakes$pH
t = -2.3134, df = 52, p-value = 0.02469
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.235425 6.945707
sample estimates:
mean of x
 6.590566
```

Florida lakes

How different is the population mean from 7?

- ***95% CI for μ:***

$$6.591 \pm 2.0066 \frac{1.288}{\sqrt{53}} = 6.591 \pm 0.355 = (6.236, 6.946)$$

where t^* corresponds to 95% confidence (97.5th percentile):

```
qt(.975, df=53-1)
[1] 2.006647
```

.

We are 95% confident that the mean pH of all lakes is between 6.236 and 6.946 (slightly acidic)

Confidence Interval for a Mean

Gribbles

Gribbles are small marine worms that bore through wood, and the enzyme they secrete may allow us to turn inedible wood and plant waste into biofuel

- ***A sample of 50 gribbles finds an average length of*** $3.1\,\mathrm{mm}$ ***with a standard deviation of*** $0.72\,\mathrm{mm}$***.***
- ***Give a*** 90% ***confidence interval for the average length of gribbles.***

Gribbles

A sample of 50 gribbles finds an average length of $3.1\,\mathrm{mm}$ *with a standard deviation of* $0.72\,\mathrm{mm}$*. For a* 90% *confidence interval for the average length of gribbles, what is* t^* *?*

a). 1.645

b). 1.677

c). 1.960

d). 1.690

▶ Click for answer

Gribbles

A sample of 50 gribbles finds an average length of $3.1\,\mathrm{mm}$ with a standard deviation of $0.72\,\mathrm{mm}$. For a 90% confidence interval for the average length of gribbles, what is the standard error?

a). 0.171

b). 0.720

c). 1.960

d). 0.102

▶ Click for answer

Gribbles

A sample of 50 gribbles finds an average length of $3.1\,\mathrm{mm}$ *with a standard deviation of* $0.72\,\mathrm{mm}$*. For a* 90% *confidence interval for the average length of gribbles, what is the margin of error?*

a). 0.171

b). 0.720

c). 1.960

d). 0.102

▶ Click for answer

Gribbles

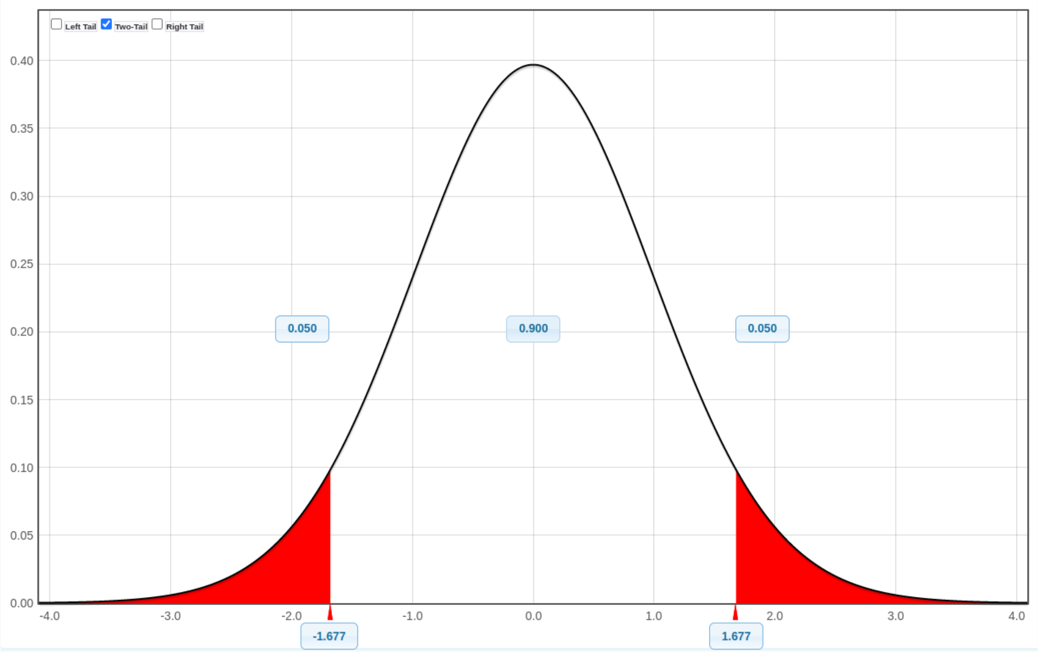
$$\text{statistic } \pm t^* \cdot SE$$

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

$$3.1 \pm 1.677 \cdot \frac{0.72}{\sqrt{50}}$$

$$3.1 \pm 0.17$$

$$(2.93, 3.27)$$



We are 90% confident that the average length of gribbles is between 2.93 and $3.27\,\mathrm{mm}$.

Margin of error

$$\mathrm{CI}: \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

For a single mean, what is the margin of error?

a) $\frac{s}{\sqrt{n}}$

b) $t^* \cdot \frac{s}{\sqrt{n}}$

c) $2 \cdot \mathrm{t}^* \cdot \frac{s}{\sqrt{n}}$

▶ Click for answer

GPA

Suppose we want to estimate average GPA at a college (where GPA's go from 0 to 4.0), with a margin of error of 0.1 with 95% *confidence. How large a sample size do we need?*

(a) About 100

(b) About 400

(c) About 800

(d) About 1000

▶ Click for answer

Hypothesis Test for a Mean

Chips Ahoy!

A group of Air Force cadets bought bags of Chips Ahoy! cookies from all over the country to verify this claim. They hand counted the number of chips in 42 bags.

$$\bar{x} = 1261.6 \text{ chips}$$
$$s = 117.6 \text{ chips}$$
$$n = 42 \text{ bags}$$

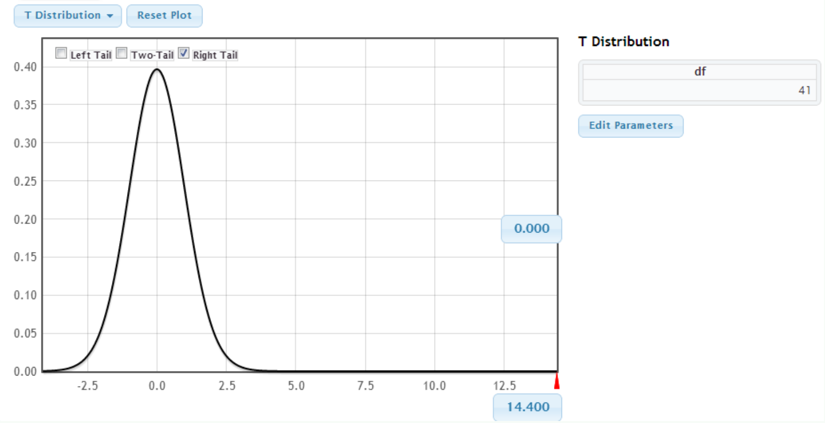
Source: Warner, B. & Rutledge, J. (1999). "Checking the Chips Ahoy! Guarantee." Chance, 12(1).

Chips Ahoy! Hypothesis Test

1. ***State hypotheses:***
   $H_0 : \mu = 1000 \quad \& \quad H_a : \mu > 1000$
2. ***Check conditions:*** $n = 42 \geq 30$ ✅
3. ***Calculate test statistic:***

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{1261.6 - 1000}{117.6/\sqrt{42}} = 14.4$$

4. ***Compute p-value:*** $p-$ ***value*** ≈ 0
5. ***Interpret in context:***



This provides extremely strong evidence that the average number of chips per bag of Chips Ahoy! cookies is significantly greater than 1000.

Chips Ahoy! Give a 99% confidence interval for the average number of chips in each bag.

1. ***State hypotheses:***

   $H_0 : \mu = 1000 \quad \& \quad H_a : \mu > 1000$

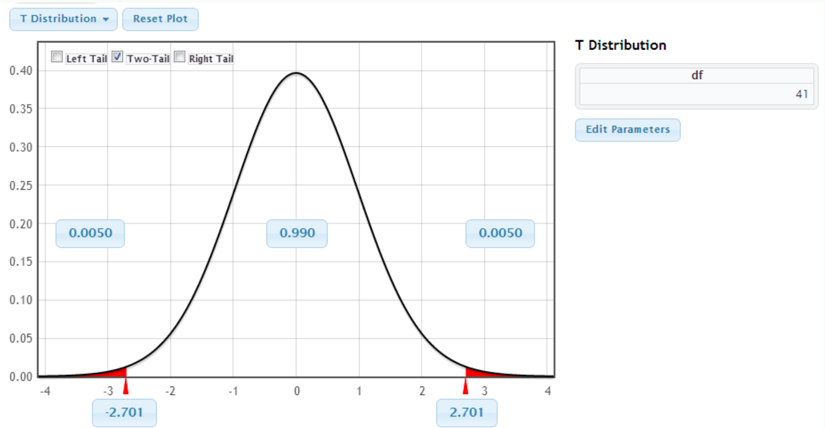
2. ***Check conditions:*** $n = 42 \geq 30$ ✅

3. ***Find t^*:*** $t^* = 2.7$

4. ***Compute confidence interval:***

   $$\bar{X} \pm t^* \times \frac{s}{\sqrt{n}} = 1261.6 \pm 2.7 \times \frac{117.6}{\sqrt{42}}$$
   $$= (1212.6, 1310.6)$$

5. ***Interpret in context:***



We are 99% confident that the average number of chips per bag of Chips Ahoy! cookies is between 1212.6 and 1310.6 chips.

Sanity check!

Which of the following properties is/are necessary for $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ to have a t-distribution?

a) the data is normal

b) the sample size is large

c) the null hypothesis is true

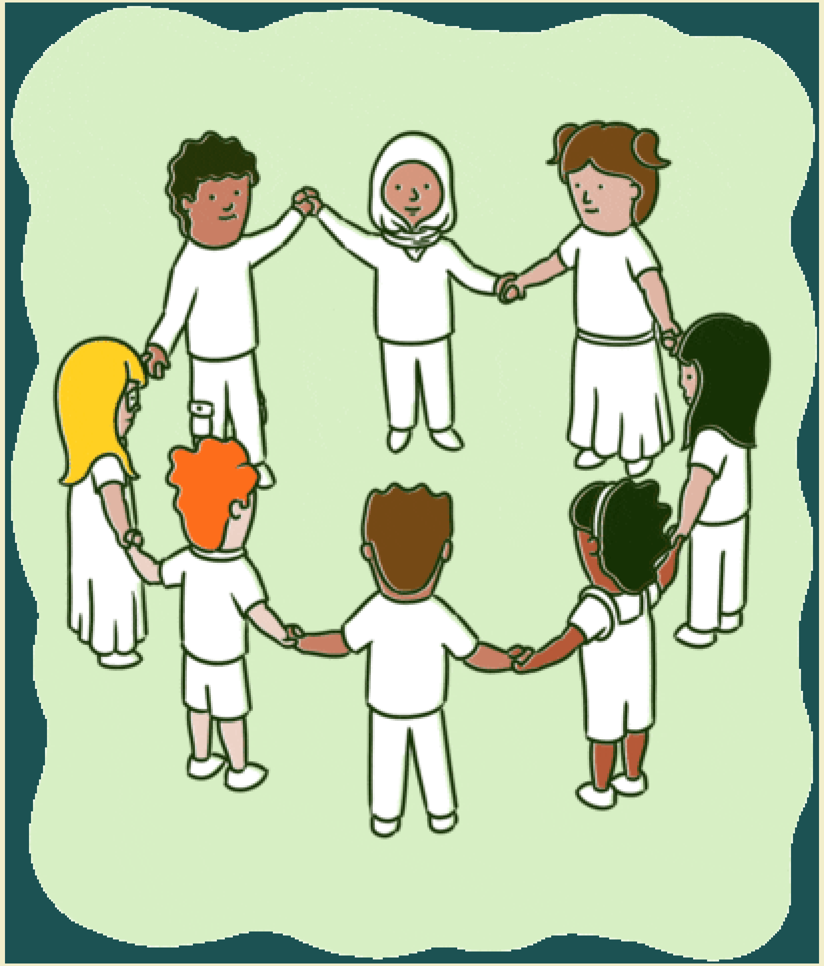
d) a ***or*** b

e) d ***and*** c

- Click for answer

YOUR TURN 1

10:00



Let's go over to the class activity .Rmd file and complete the tasks for today.