STA220H1: The Practice of Statistics I

Elena Tuzhilina

March 7, 2023

Please turn on your videos :)

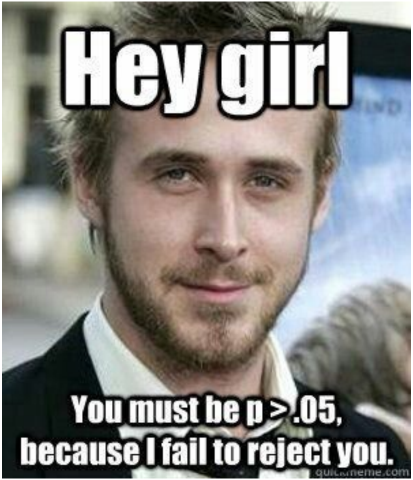


Figure 1: [picture source]

Announcements

1. Midterm 2 is next week at 7:20-9:40 PM in EX 100.
2. Same rules. Bring your ID.
3. Online review session at 6:00-7:00PM, you can stay in EX 100.

Agenda for today

- Recap: CLT, confidence intervals
- Statistical testing: H_0 and H_a, process, p-value

Recap: confidence intervals

We want to study the average life expectancy in Canada μ.

We take a sample of n people, record their ages of death

$$x_1, \ldots, x_n$$

and compute the sample mean $\bar{x}$.

We claim that it is an **estimate** of the average life expectancy in Canada.

$$\bar{x} \approx \mu$$

How confident are we in our estimate? Can we use our sample to find a range for μ?

Recap: central limit theorem

Central limit theorem: for n large enough

$$\bar{X} \text{ approximately} \sim \textit{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

When CLT is true?

- $X_1, \ldots, X_n$ should be independent and identically distributed
- If $X_1, \ldots, X_n$ are normal then $\bar{X}$ is exactly normal
- If $n > 30$ then $\bar{X}$ is approximately normal

Recap: confidence intervals

CLT:

$$\bar{X} \text{ approximately} \sim \textit{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardization:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ approximately} \sim \textit{Normal}\,(0, 1)$$

Distribution table:

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

With probability 0.95, the population parameter μ belongs to

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Recap: confidence intervals

How to use the sample to construct the confidence interval?

If σ is known

$$x_1, \ldots, x_n \Rightarrow \bar{x}$$

and the 95% confidence interval is

$$\left[\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Recap: confidence intervals

How to use the sample to construct the confidence interval?

If σ is unknown we estimate it $s \approx \sigma$

$$x_1, \ldots, x_n \Rightarrow \bar{x}, s$$

and could probably use the 95% confidence interval

$$\left[\bar{x} - 1.96 \cdot \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{s}{\sqrt{n}}\right]$$

However it is not accurate...

Recap: confidence intervals

Both $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are RVs!

Standardization:

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

Distribution table:

$$P\left(-... \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq ...\right) = 0.95$$

With probability 0.95, the population parameter μ belongs to

$$\left[\bar{X} - ... \cdot \frac{S}{\sqrt{n}}, \bar{X} + ... \cdot \frac{S}{\sqrt{n}}\right]$$

Recap: confidence intervals

How to use the sample to construct the confidence interval?

If σ is unknown we estimate it $s \approx \sigma$

$$x_1, \ldots, x_n \Rightarrow \bar{x}, s$$

and the 95% confidence interval is

$$\left[\bar{x} - ... \cdot \frac{s}{\sqrt{n}}, \bar{x} + ... \cdot \frac{s}{\sqrt{n}}\right]$$

where ... is found from the distribution table.

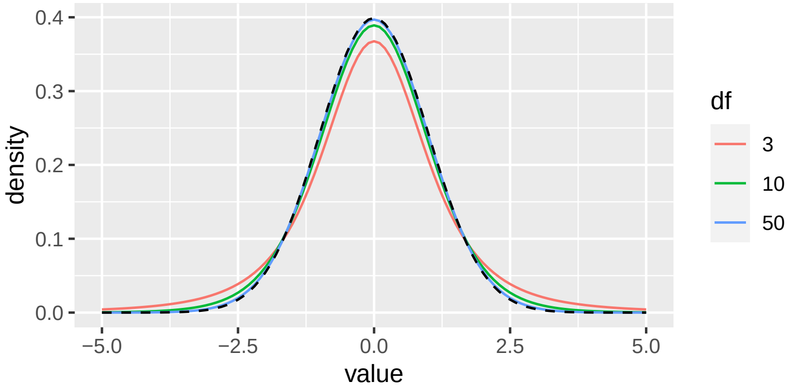
Recap: t-distribution

Normal: $\ldots = 1.96$

t with df = 3: $\ldots = 3.18$

t with df = 10: $\ldots = 2.23$

t with df = 50: $\ldots = 2.01$



Confidence intervals

Known σ, 95% CI is $\left[\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$

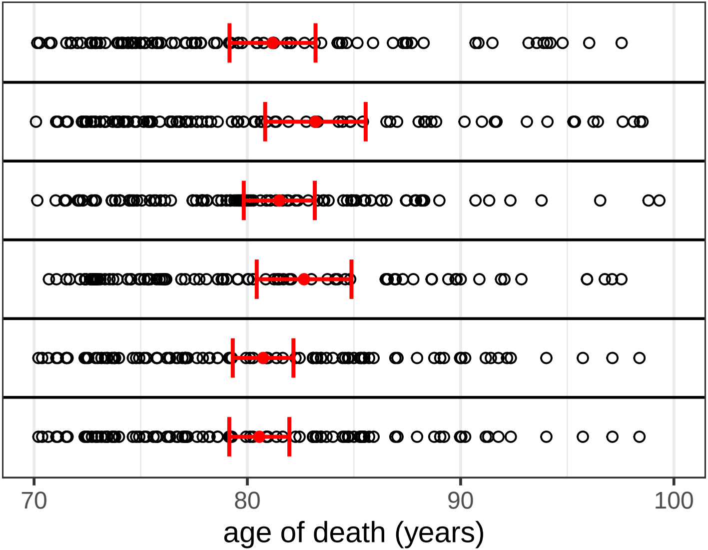
Unknown σ, 95% CI is $\left[\bar{x} - ... \cdot \frac{s}{\sqrt{n}}, \bar{x} + ... \cdot \frac{s}{\sqrt{n}}\right]$

Properties:

- CI is centered at $\bar{x}$
- CI covers μ with high probability
- CI size decreases with the growth of n
- CI size decreases with the decrease in confidence
- CI bounds depend on the sample $x_1, ..., x_n$
- If σ is know the CI does not depend on $x_1, ..., x_n$

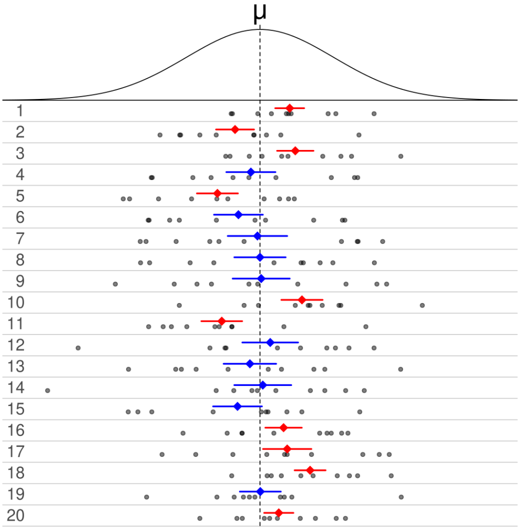
Exercise

Is σ known for these confidence intervals?



Confidence intervals: interpretation

95% confidence means that for 95% samples CI will cover μ.



Confidence intervals: interpretation

Alternative view:

- We have n random variables $X_1, \ldots, X_n$
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are **random variables**
- CI bounds $LB = \bar{X} - ... \cdot \frac{S}{\sqrt{n}}$ and $UB =, \bar{X} + ... \cdot \frac{S}{\sqrt{n}}$ are **random variables**
- Each CI is a realization of $[LB, UB]$

95% confidence means that $[LB, UB]$ will cover μ with probability 0.95.

Exercise

We want to study the average height of people in Canada. We took 500 samples of size 50 and used each sample to compute 90% confidence interval (500 CIs in total). How many of them do you think will not contain the true average height?

Statistical testing

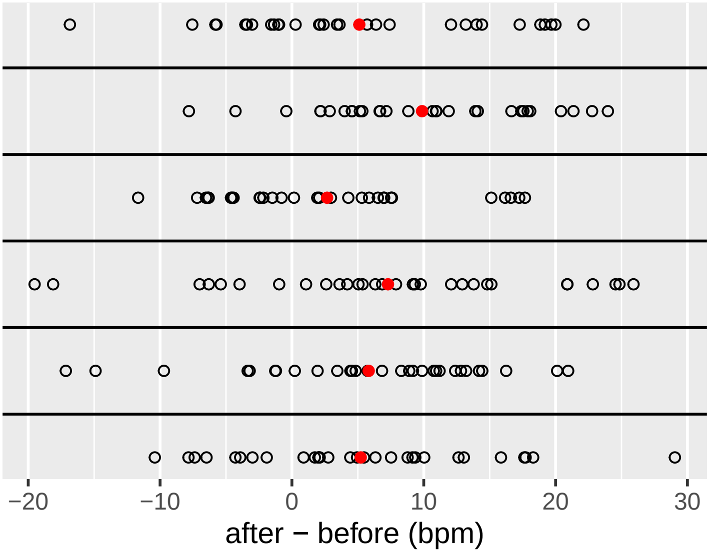
Statistical tests use data to answer questions about the population.

- *Caffeine causes a dramatic increase in the heart rate!*

For 30 participants of the experiment, the heart rates before and after coffee was measured the difference was computed. The average value for the difference is 5 bpm.

Statistical testing

Why can't we just compare the average difference to zero?



Statistical testing

Goal: determine whether the collected data provides enough evidence for us to believe in a claim about the theoretical world.

Statistical testing

Step 1: state your **null** hypothesis and the **alternative** hypothesis.

- **Null** $H_0 : \mu = \mu_0$
- **One sided alternative** $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$
- **Two sided alternative** $H_a : \mu \neq \mu_0$

Do our data provide enough evidence against the null?

Statistical testing

Step 1: state your **null** hypothesis and the **alternative** hypothesis.

H_0 : the before and after coffee heart rates are the same

H_a : the after coffee heart rates is higher than the before one

What are μ and μ_0 here?

Statistical testing

Step 2: summarize the data into a **test statistic**.

- Test statistic is constructed assuming that H_0 is true

How extreme is our test statistic assuming that H_0 *is true?*

Statistical testing

Step 2: summarize the data into a **test statistic**.

For 30 participants of the experiment, the heart rates before and after coffee was measured the difference was computed. The sample mean is 5 bpm, the sample standard deviation is 0.5.

$$t_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

```
tobs = (5-0)/(0.5*sqrt(30))
tobs
```

```
## [1] 1.825742
```

Statistical testing

Step 3: compute **p-value**.

- It is a number between 0 and 1
- It quantifies how strong is the evidence against H_0
- The smaller the value the stronger the evidence that our data contradict H_0

P-value measures how likely the observed data would be IF the null hypothesis is true.

Statistical testing

Step 3: compute **p-value**.

Recall that $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

Thus we can find $p\text{-}value = P(T > t_{obs})$

```
pt(tobs, df = 29, lower.tail = FALSE)
```

```
## [1] 0.03910166
```

Statistical testing

Step 4: draw the conclusion.

- If p-value is small we have enough evidence against H_0
- We reject H_0 in favor of H_a
- We say that the result is **statistically significant**

How small should be p-value?

- The smaller the more significant evidence we have to reject H_0

General rule: pre-select some **significance level** α and check if $p\text{-value} < \alpha$

- Statisticians prefer $p\text{-value} < 0.05$

Statistical testing

Step 4: draw the conclusion.

p-value < 0.05, thus we can reject H_0 in favor of H_a.

The after coffee heart rates is higher than the before one!

Statistical testing and tails

Statistical testing: more examples

- *The student was randomly guessing on the exam!*

A student took a test with 100 Yes/No questions. They received the tests results and they got 65 questions correctly.

How to compute p-value?

Statistical testing: more examples

Step 1: state your **null** hypothesis and the **alternative** hypothesis.

$H_0 : p = 0.5$ and $H_a : p \neq 0.5$

Statistical testing: more examples

Step 2: summarize the data into a **test statistic**.

Since $Z = \frac{\bar{X}-p}{\sqrt{p(1-p)/n}} \sim Normal(0,1)$ we use

$$z_{obs} = \frac{0.65 - 0.5}{\sqrt{0.5(1-0.5)/100}}$$

```
zobs = (0.65-0.5)/sqrt(0.5 * (1-0.5)/100)
zobs
```

```
## [1] 3
```

Statistical testing: more examples

Step 3: compute **p-value**.

The probability to observe such an extreme statistic is *p-value* $= P(|Z| > |z_{obs}|)$

```
2*pnorm(zobs, lower.tail = FALSE)
```

```
## [1] 0.002699796
```

Statistical testing: more examples

Step 4: draw the conclusion.

p-value < 0.05, thus the student did not guess on the exam!

Statistical testing and tails

Statistical testing: more examples

- *Lottery is scamming people!*

The lottery company claims that 10% of their tickets win. A customer bought 500 tickets and won only 30 times.

How the procedure will change?

Statistical testing: more examples

Step 1: state your **null** hypothesis and the **alternative** hypothesis.

$H_0 : p = 0.1$ and $H_a : p < 0.1$

Step 2: summarize the data into a **test statistic**.

$$z_{obs} = \frac{0.06 - 0.1}{\sqrt{0.1(1 - 0.1)/500}}$$

```
zobs = (0.06-0.1)/sqrt(0.1 * (1-0.1)/500)
zobs
```

```
## [1] -2.981424
```

Statistical testing: more examples

Step 3: compute **p-value**.

The probability to observe such an extreme statistic is *p-value* $= P(Z < z_{obs})$

```
pnorm(zobs, lower.tail = TRUE)
```

```
## [1] 0.001434556
```

Step 4: draw the conclusion.

p-value < 0.05, thus the probability to win is less than announced!

Statistical testing and tails

Statistical testing

What if p-value is more than 0.05?

- We do not have enough evidence to reject H_0
- This is not the same as to accept H_0!

TO DO

1. Module 8. The Process of Statistical Tests
2. Quiz 8 due Monday (March 13) @ 11:59 PM (EST)
3. Practice Problem Set 8