STA220H1: The Practice of Statistics I

Elena Tuzhilina

March 7, 2023

Please turn on your videos :)

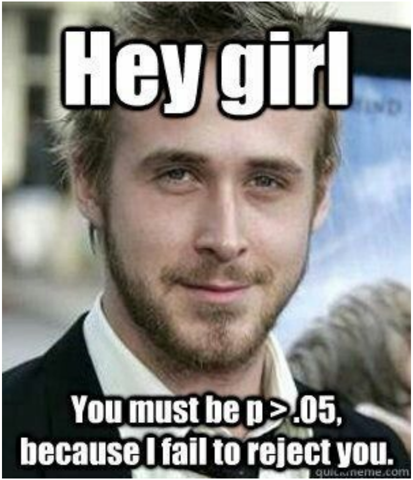


Figure 1: [picture source]

Announcements

1. Midterm 2 is next week at 7:20-8:40 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

$$\frac{X_1 + \dots + X_n}{n} = \bar{X} \text{ approximately} \sim \textit{Normal}\left(\underset{E(X_i)}{\mu}, \frac{\overset{Var(X_i)}{\sigma^2}}{n}\right)$$

When CLT is true?

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

Recap: confidence intervals

CLT:

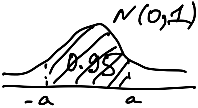
$$\bar{X} \text{ approximately} \sim \textit{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardization:

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

Distribution table:

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



With probability 0.95, the population parameter μ belongs to

RV

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

Recap: confidence intervals

How to use the sample to construct the confidence interval?

If σ is known

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

number

and the 95% confidence interval is

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

Recap: confidence intervals

How to use the sample to construct the confidence interval?

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

$$x_1, \ldots, x_n \Rightarrow \bar{x}, s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

RV

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 $s \approx \sigma$

number

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

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

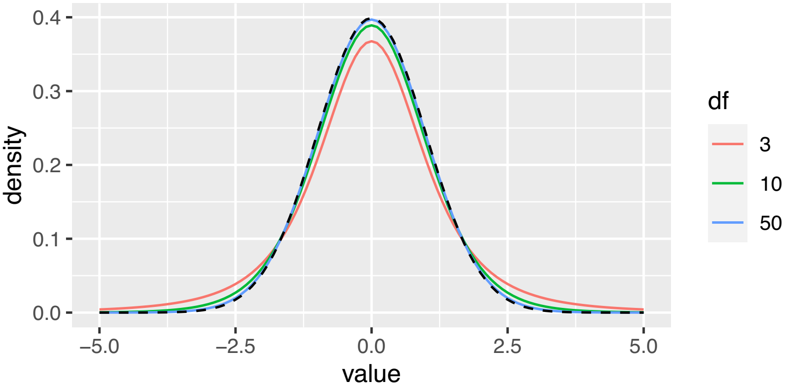
Normal: $\ldots = 1.96$

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

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

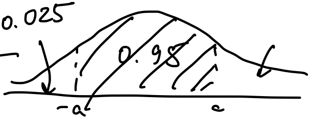
t with df = 50: $\ldots = 2.01$

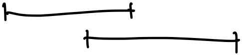
$t_{n-1} \to N(0,1)$
$n \to \infty$



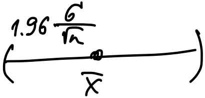
Confidence intervals

$P(X \le -a) = 0.025$

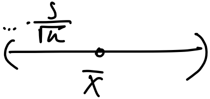




Known σ, 95% CI is $\left[\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$



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

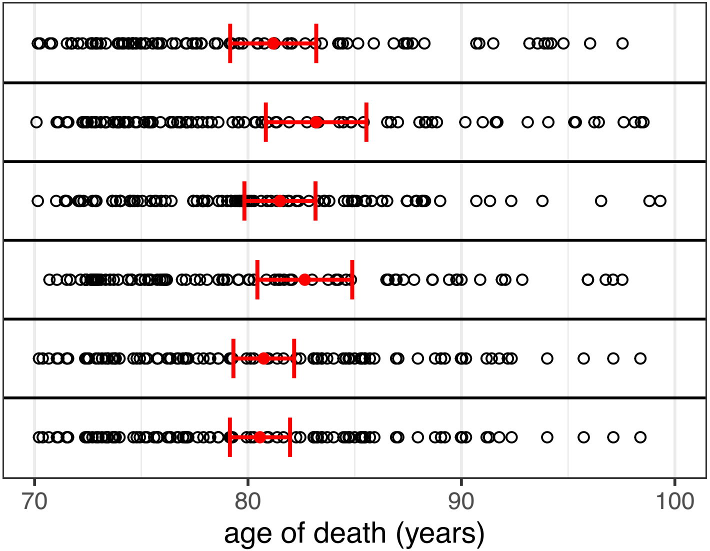


Properties:

- 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 90% 95%
- CI bounds depend on the sample $x_1, \ldots, x_n$
- If σ is known then $\underbrace{\text{CI}}_{\text{width}}$ does not depend on $x_1, \ldots, x_n$

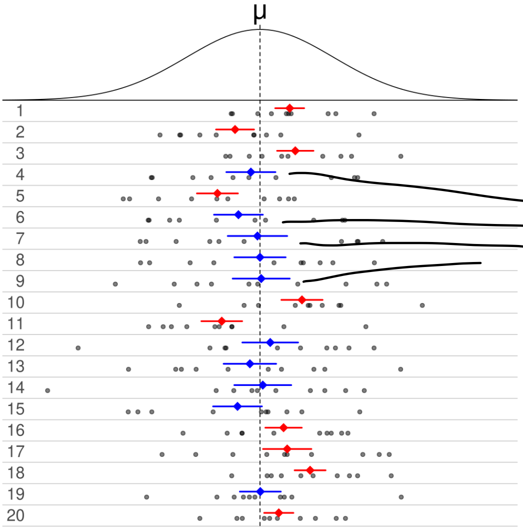
Exercise

Is σ known for these confidence intervals?



Confidence intervals: interpretation

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



1/20 does not cover μ

95% intervals

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

$$\boxed{n=50} \longrightarrow \left.\begin{array}{l} x_1 \dots x_{50} \to CI_1 \\ x_1 \dots x_{50} \to CI_2 \\ \\ x_1 \dots x_{50} \to CI_{500} \end{array}\right\} 500$$

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? μ

90% ⟹ 10% samples will produce CI that does not cover μ

$500 \Rightarrow \boxed{50 \text{ CI}}$

Statistical testing

Statistical tests use data to answer questions about the population.

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

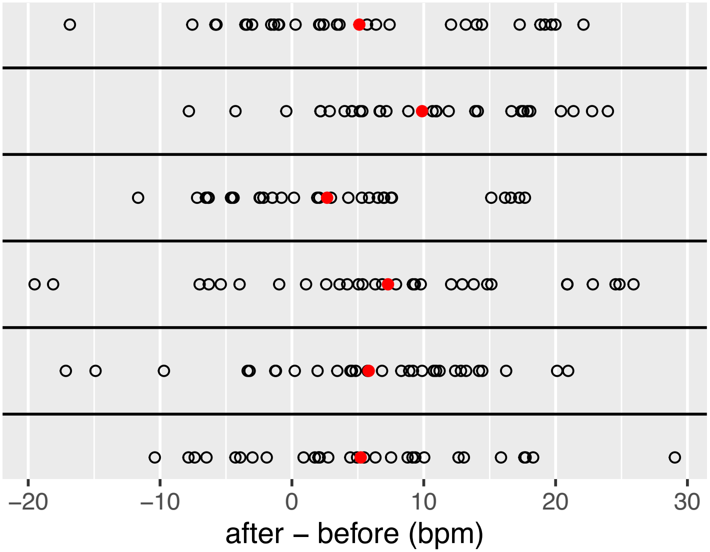
For 30 participants of the experiment, the heart rates before and after coffee was measured the difference was computed. The average value for the difference is 5 bpm.

$$\text{difference} = \text{after} - \text{before}$$

$$x_1, \ldots x_{30} \longrightarrow \bar{x} = 5\,\text{bpm}$$

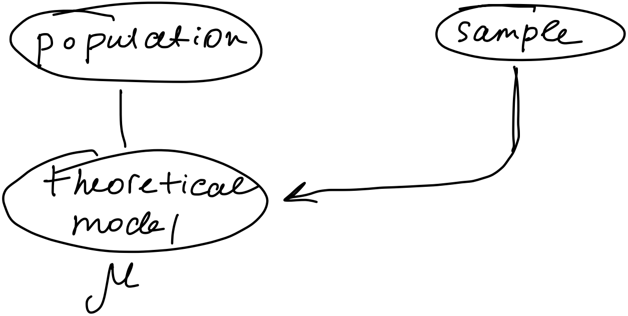
Statistical testing

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.

pop. par

some value

- **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?

$H_0 \; \mu = \mu_0$ vs $H_a \; \mu > \mu_0$

$H_0 \; \mu = \mu_0$ vs $H_a \; \mu < \mu_0$

$H_0 \; \mu = \mu_0$ vs $H_a \; \mu \neq \mu_0$

$H_0 : \sigma^2 = 1 \quad H_a :$

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

after − before > 0

What are μ and μ_0 here?

X = difference

$$\boxed{\begin{aligned} H_0 &: \mu = 0 \\ H_a &: \mu > 0 \end{aligned}}$$

$$\boxed{E(X) = \mu}$$

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.

$\bar{X}$ S

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

μ: 0 (H_0)

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

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

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

Statistical testing

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

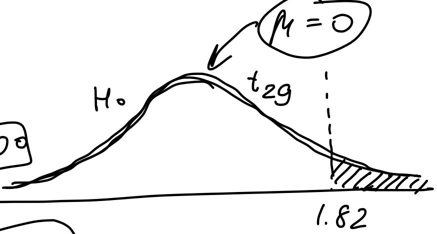
- It is a number between 0 and 1
- It quantifies how strong is the evidence against H_0
- The smaller the value the stronger the evidence that our data contradict H_0

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

Statistical testing

$(n=30)$

$n>500$



Step 3: compute **p-value**. $N(0,1)$

$n>30$

Recall that $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{\cancel{n-1}}$ $\quad t_{29}$

Thus we can find *p-value* $= P(T > t_{obs}) = P(T > 1.82)$, $T \sim t_{29}$

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

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

$=$ p-value

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

$0.05 \to 0.01 \to 0.1$

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

- Statisticians prefer *p-value* $<$ *0.05*

Statistical testing

Step 4: draw the conclusion.

0.035

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!

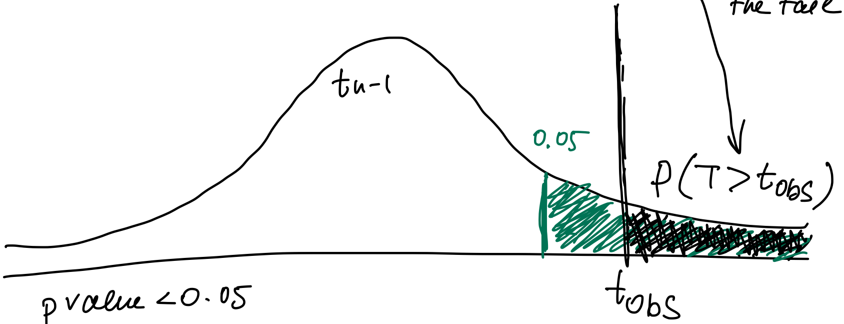
Statistical testing and tails

$H_0: T \sim t_{n-1}$

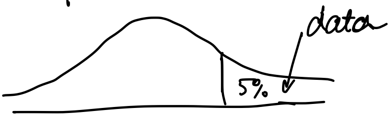
Data: $x_1 .. x_n \to t_{obs} \to$ p-value

p-value < 0.05

"size of the tail"



p value < 0.05



$\boxed{\text{Our data is rare. It is in 5\%}}$

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?

$$\bar{X} = \frac{65}{100} = 0.65$$

Statistical testing: more examples

x_1 - - - x_n $\quad n = 100$

$\shortparallel$

0 or 1

incorrect correct

X = proportion of correctly guessed questions

p = prob. to get answer correctly

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

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

H_0: student is guessing

H_a: not guessing

Statistical testing: more examples

$X_1 \ldots X_n \sim Bern(p) \Rightarrow \bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$ σ^2

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

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

$\sigma / \sqrt{n}$

$\bar{X}$ = proportion of correct questions

$\sqrt{p(1-p)} = \sigma$

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

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

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

$X_1 \ldots X_n \sim Bern(p)$

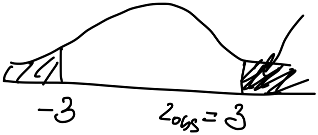
$E(X_i) = p \quad Var(X_i) = p \cdot (1-p)$

$E(\bar{X}) = p \quad Var(\bar{X}) = \frac{p(1-p)}{n}$

Statistical testing: more examples

$H_0: p = 0.5 \qquad H_a: p > 0.5$

$P(T > t_{obs})$



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

The probability to observe such an extreme statistic is *p-value*
$= P(|Z| > |z_{obs}|) = P(Z > 3) + P(Z < -3)$ $\vee$ $\boxed{0.05}$

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

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

$=$ p-value

$Z \sim N(0,1)$ under null

$H_a: p \neq 0.5$

or $p > 0.5$
$p < 0.5$

Statistical testing: more examples

$H_0 : \cancel{p = 0.5}$ $\boxed{H_a : p \neq 0.5}$

Step 4: draw the conclusion.

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

$\geqslant 0.05$

$X_1 \ldots X_n$

$\sim Bern(p)$

$\boxed{E(X_i) = \mu}$

$\boxed{E(X_i) = p}$

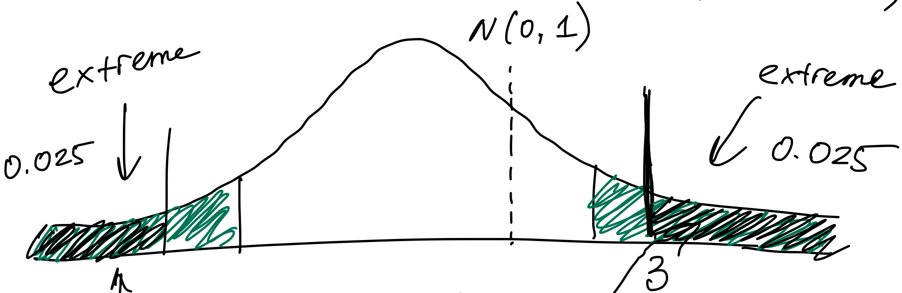
Statistical testing and tails

$H_0: Z \sim N(0,1)$

$x_1, \dots, x_n \longrightarrow z_{obs}$

Reject $p = 0.5$ in favor of $p \neq 0.5$

$P(|Z| > |z_{obs}|)$



reject H_0 if data is extreme

$\boxed{P(Z > z_{obs}) \quad \text{compare to } 0.025}$

$P(Z < -z_{obs}) + P(Z > z_{obs})$ compare to 0.05

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?

0.06

$H_a: p > 0.1$

p-value

Statistical testing: more examples

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

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

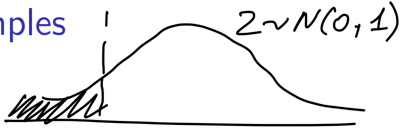
Step 2: summarize the data into a **test statistic**.

$$z_{obs} = \frac{0.06 - 0.1}{\sqrt{0.1(1 - 0.1)/500}} \qquad \frac{\overset{\frac{30}{500}}{\overset{\shortparallel}{\bar{X}}} - p}{\sqrt{p(1-p)/n}}$$

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

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

Statistical testing: more examples



$z_{obs} = -2.98$ H_0 $p < 0.1$

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
```

= p-value

Step 4: draw the conclusion.

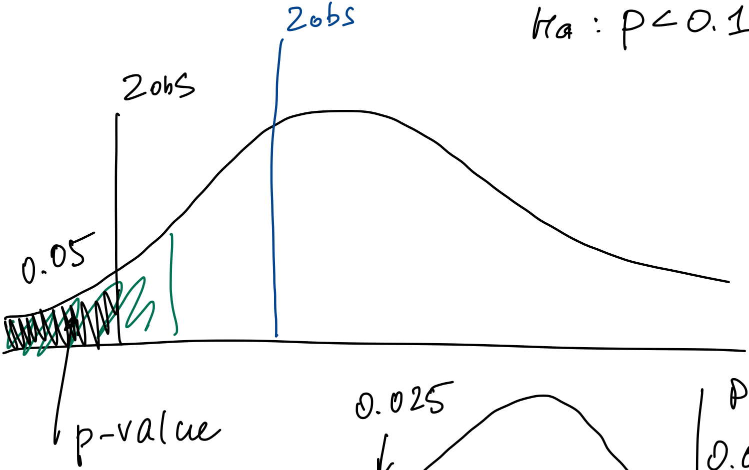
p-value < 0.05, thus the probability to win is less than announced!

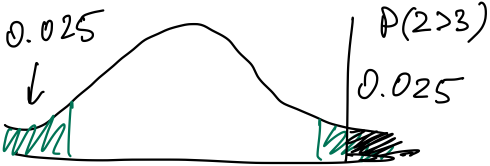
$H_0 : p = 0.1$ $H_a : p < 0.1$

Statistical testing and tails

$H_0 : Z \sim N(0.1)$

$H_a : p < 0.1$





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