



# HYPOTHESIS TESTING I

Sandeep Soni

---

09/27/2023

# CLASS LOGISTICS

---

- PS1 due today

# QUESTION FOR THE DAY

---

“How to statistically test a claim using text?”

Nabokov's



Favorite



Word Is



*Mauve*

Cover page of the book written by Ben Platt

# HYPOTHESES

---

# HYPOTHESES

---

- Many claims can be framed as hypotheses

# HYPOTHESES

---

- Many claims can be framed as hypotheses

## Examples

Gender bias is decreasing in books

Reframing a tweet can increase its retweet rate

Institutional mistrust is predictive of hate speech

Classifier A is better than classifier B

# NULL HYPOTHESIS

---

# NULL HYPOTHESIS

---

- Null hypothesis is a claim that is assumed to be true

# NULL HYPOTHESIS

---

- Null hypothesis is a claim that is assumed to be true

Hypothesis	$H_0$
Gender bias is decreasing in books	Gender bias remains the <b>same</b>
Reframing a tweet can increase its retweet rate	Reframing a tweet has <b>no</b> effect on retweet rate
Institutional mistrust is predictive of hate speech	Institutional mistrust <b>is not</b> predictive of hate speech
Classifier A is better than classifier B	Classifier A is the <b>same</b> as classifier B

# HYPOTHESIS TESTING

---

# HYPOTHESIS TESTING

---

- Null hypothesis gives us an expected result

# HYPOTHESIS TESTING

---

- Null hypothesis gives us an expected result
- We want to test that if the null hypothesis holds, how likely does the data match our expectation?

# HYPOTHESIS TESTING

---

# HYPOTHESIS TESTING

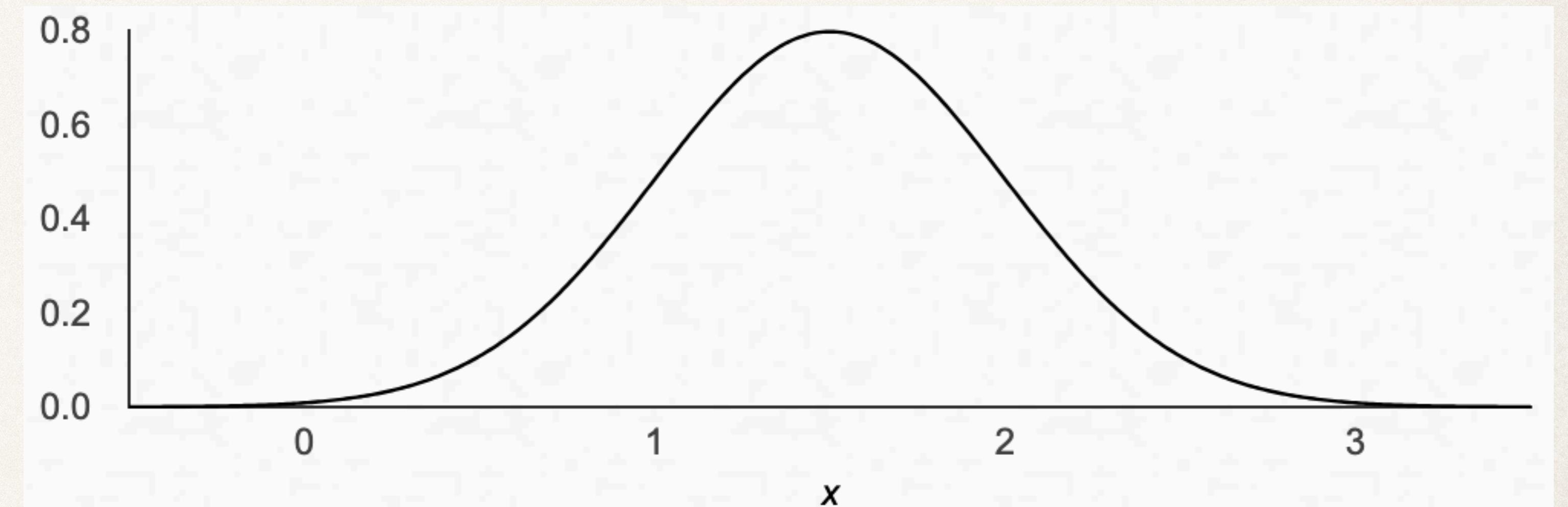
---

According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance

# HYPOTHESIS TESTING

---

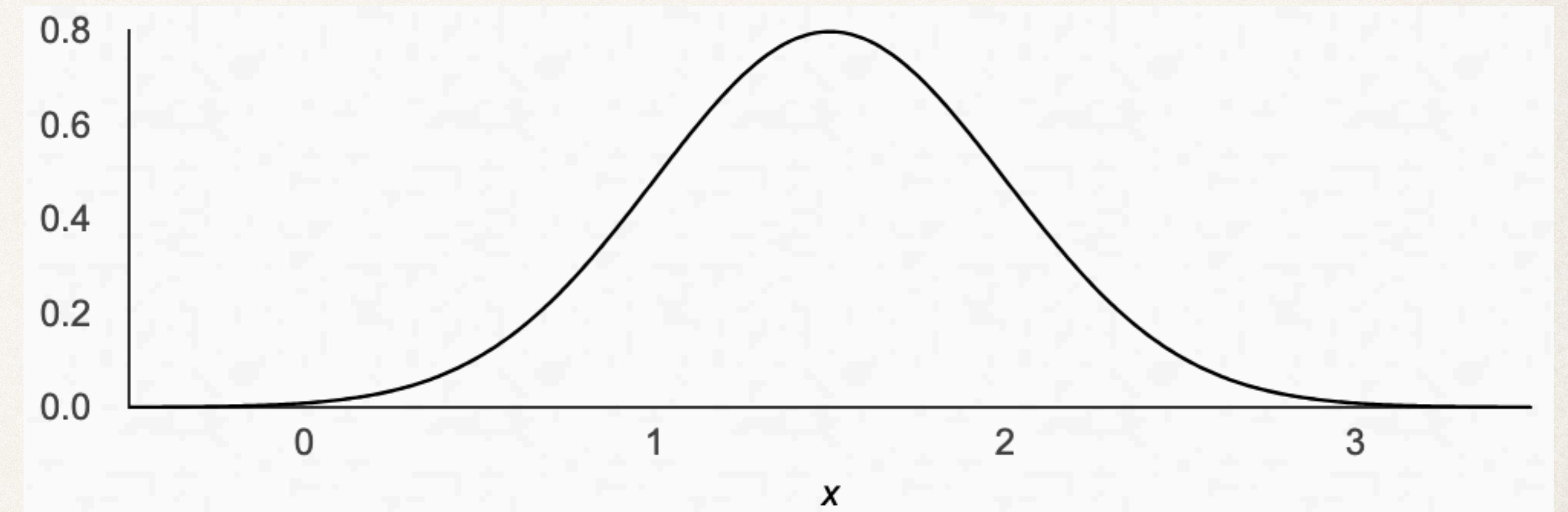
According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance



# HYPOTHESIS TESTING

---

According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance



This sets up expectation about the shape and position of the distribution

# HYPOTHESIS TESTING

---

# HYPOTHESIS TESTING

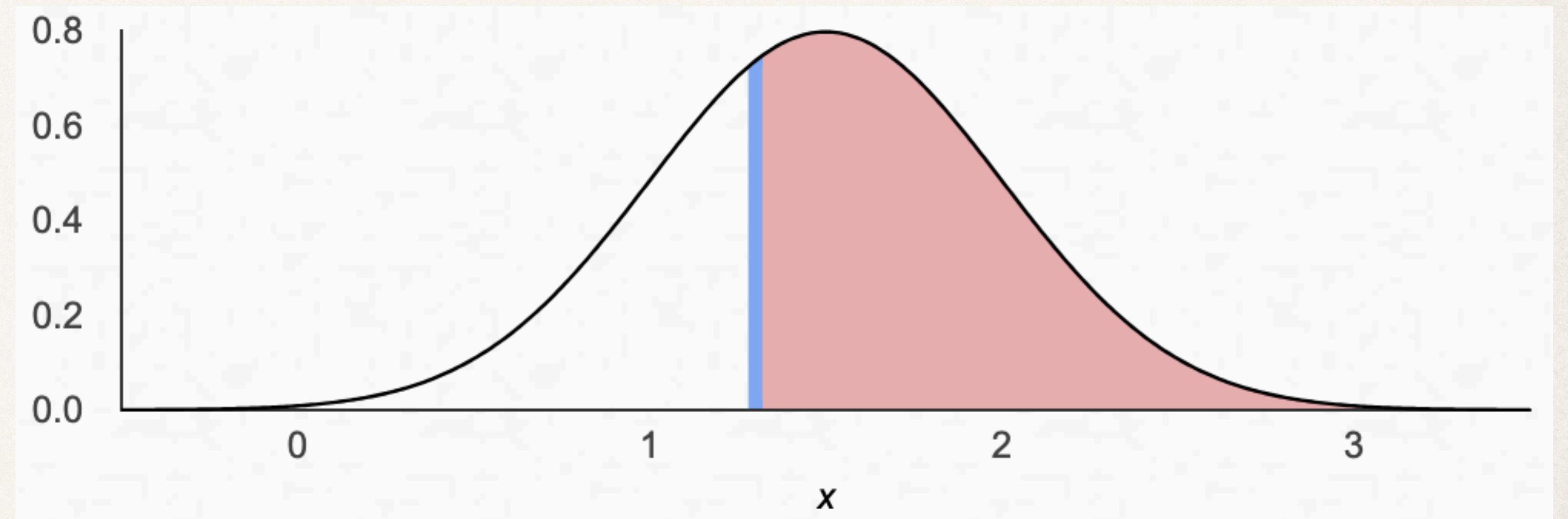
---

If we observe  
 $x$  to be 1.3,  
how likely are  
we to see that  
if the null  
hypothesis  
holds?

# HYPOTHESIS TESTING

---

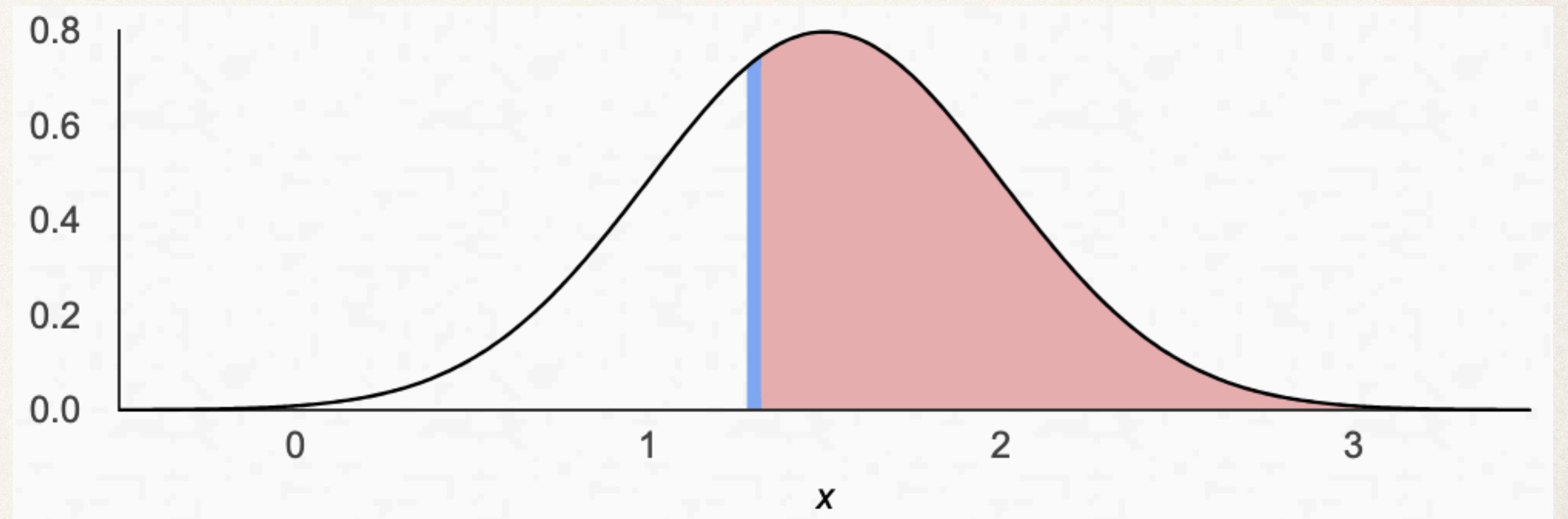
If we observe  $x$  to be 1.3, how likely are we to see that if the null hypothesis holds?



# HYPOTHESIS TESTING

---

If we observe  $x$  to be 1.3, how likely are we to see that if the null hypothesis holds?



We can quantify this by calculating the probability of the shaded area

# HYPOTHESIS TESTING

---

# HYPOTHESIS TESTING

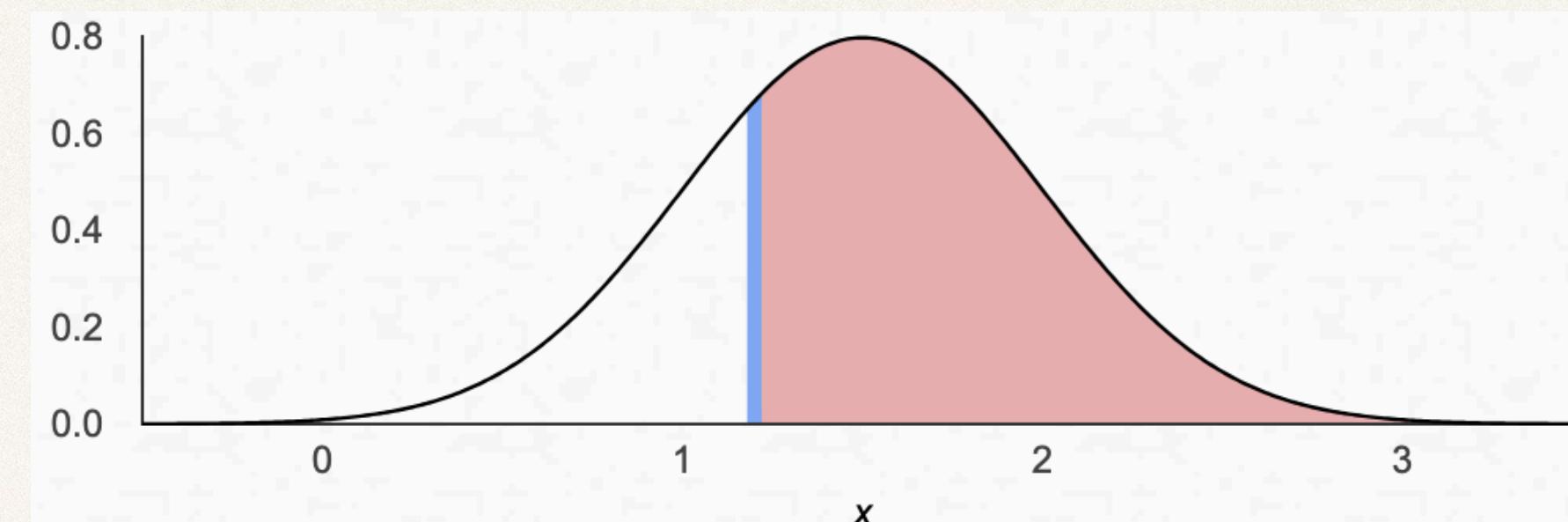
---

Which  
observation  
would you say  
is surprising  
if the null  
distribution  
holds?

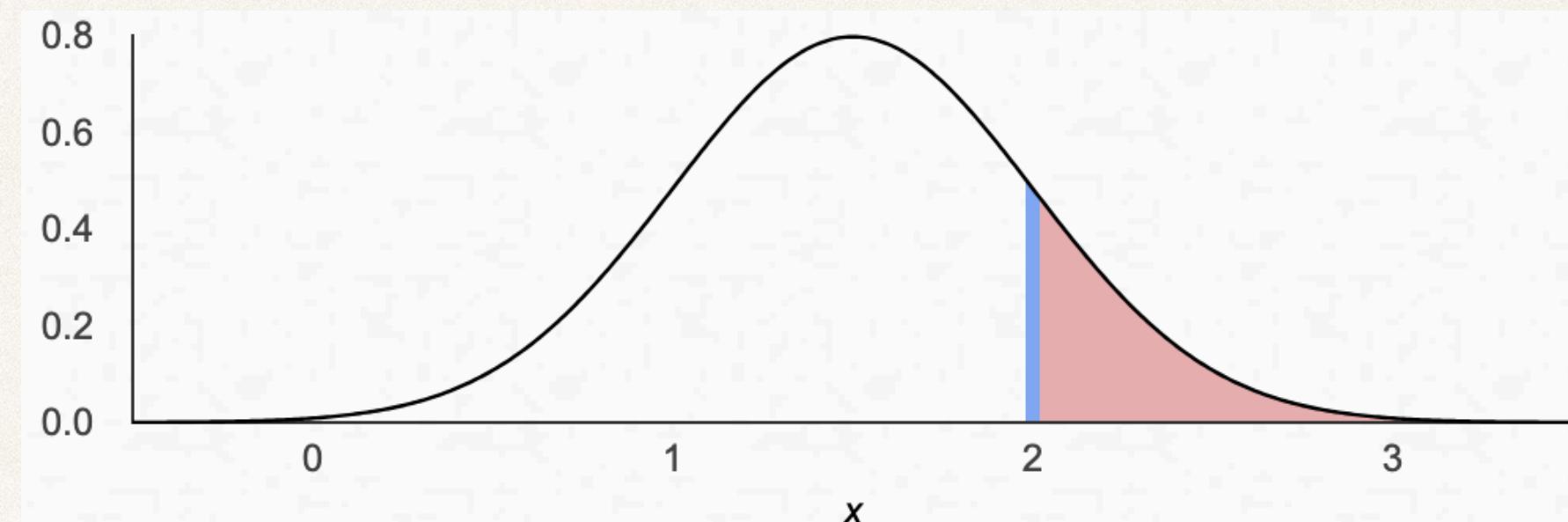
# HYPOTHESIS TESTING

Which observation would you say is surprising if the null distribution holds?

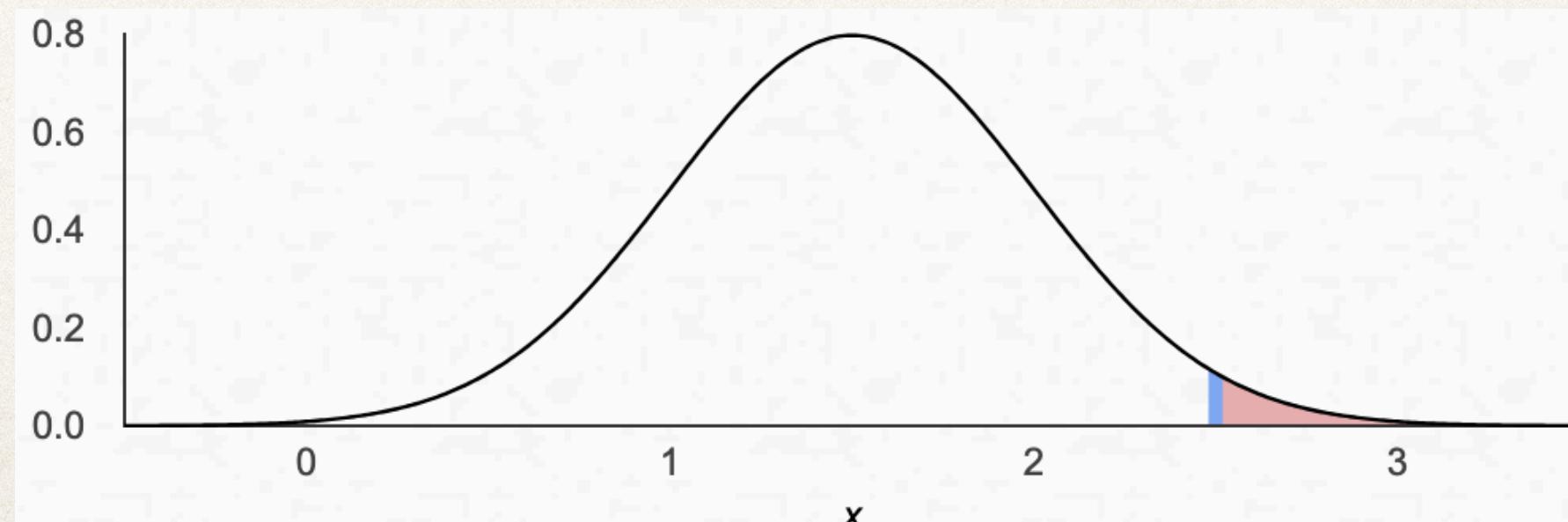
x=1.25



x=2



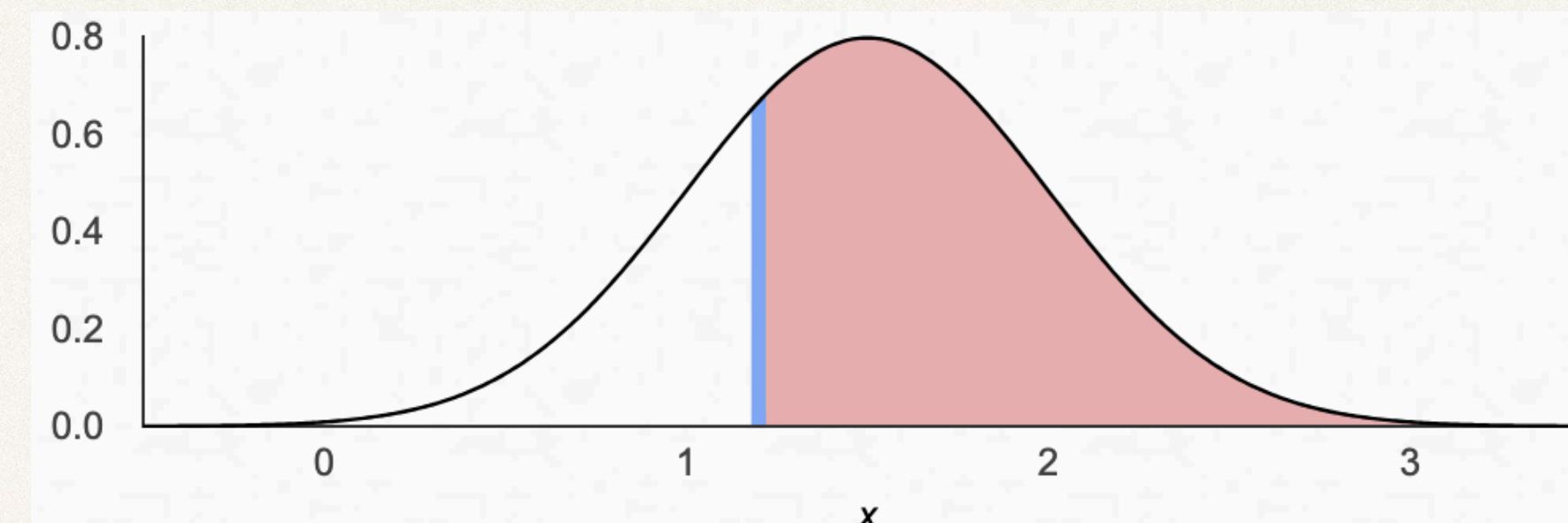
x=2.5



# HYPOTHESIS TESTING

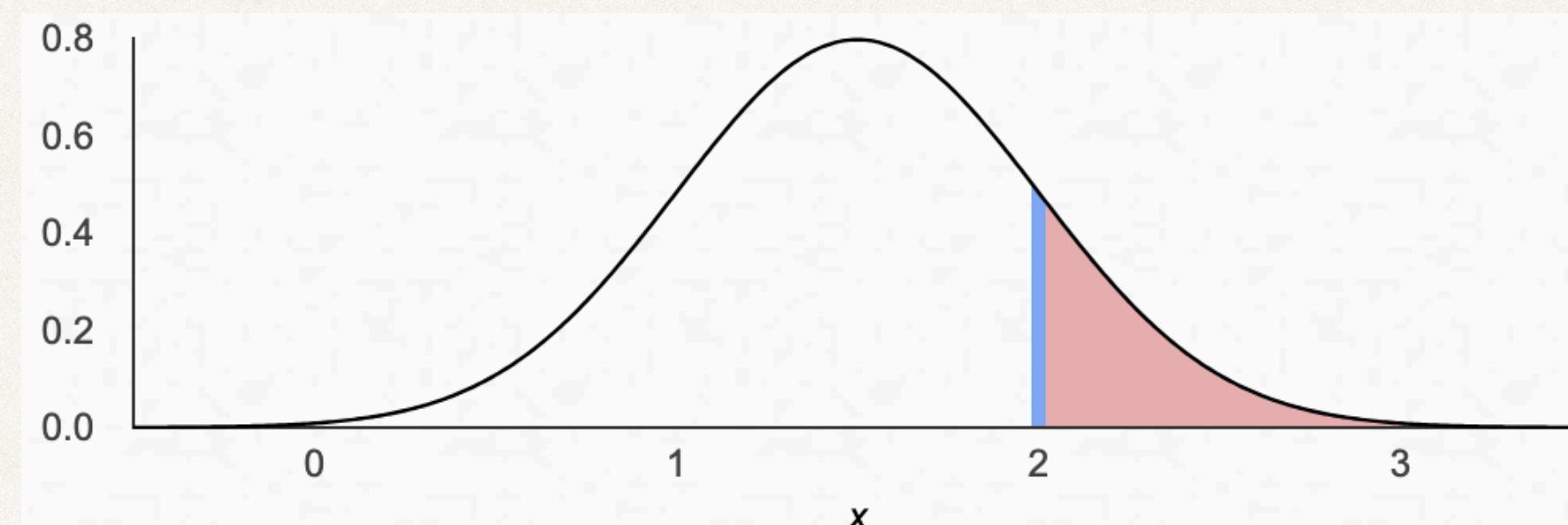
Which observation would you say is surprising if the null distribution holds?

x=1.25



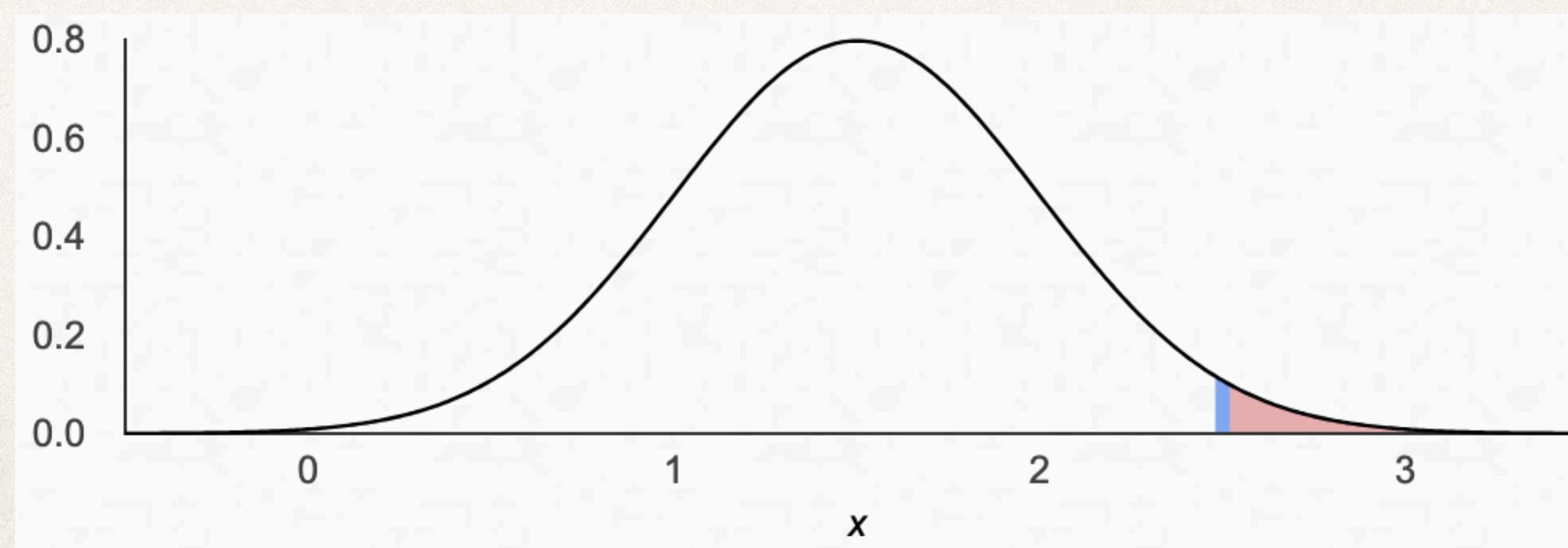
p=0.69

x=2



p=0.16

x=2.5



p=0.02

# SIGNIFICANCE LEVEL

---

# SIGNIFICANCE LEVEL

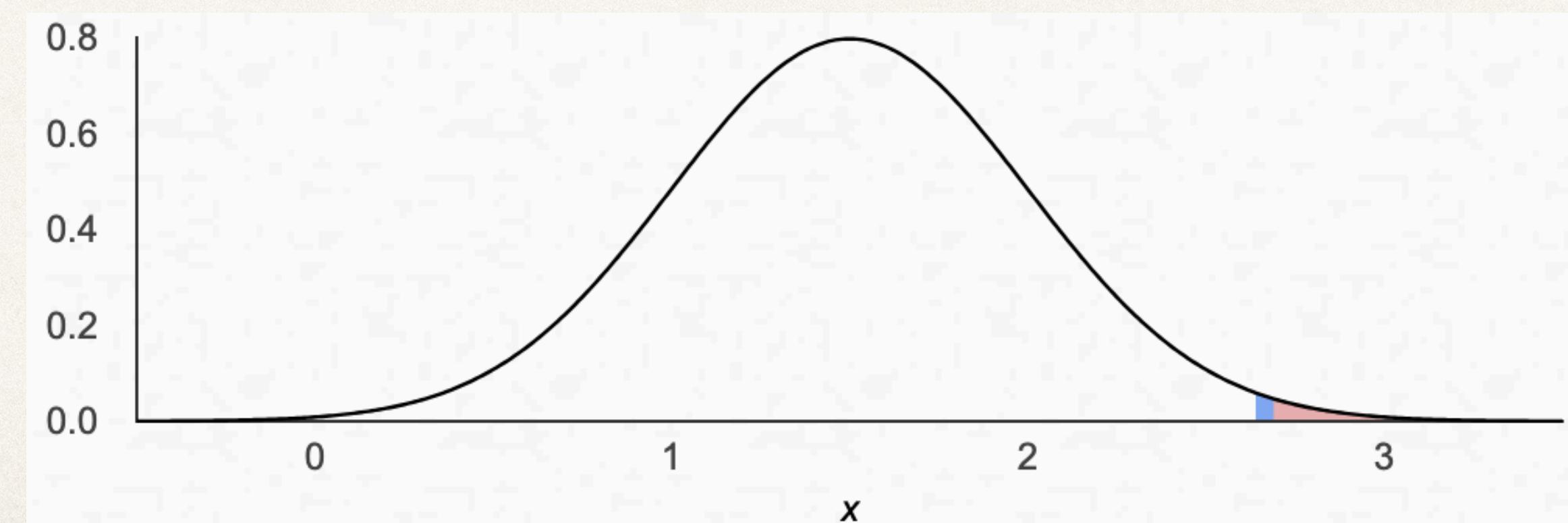
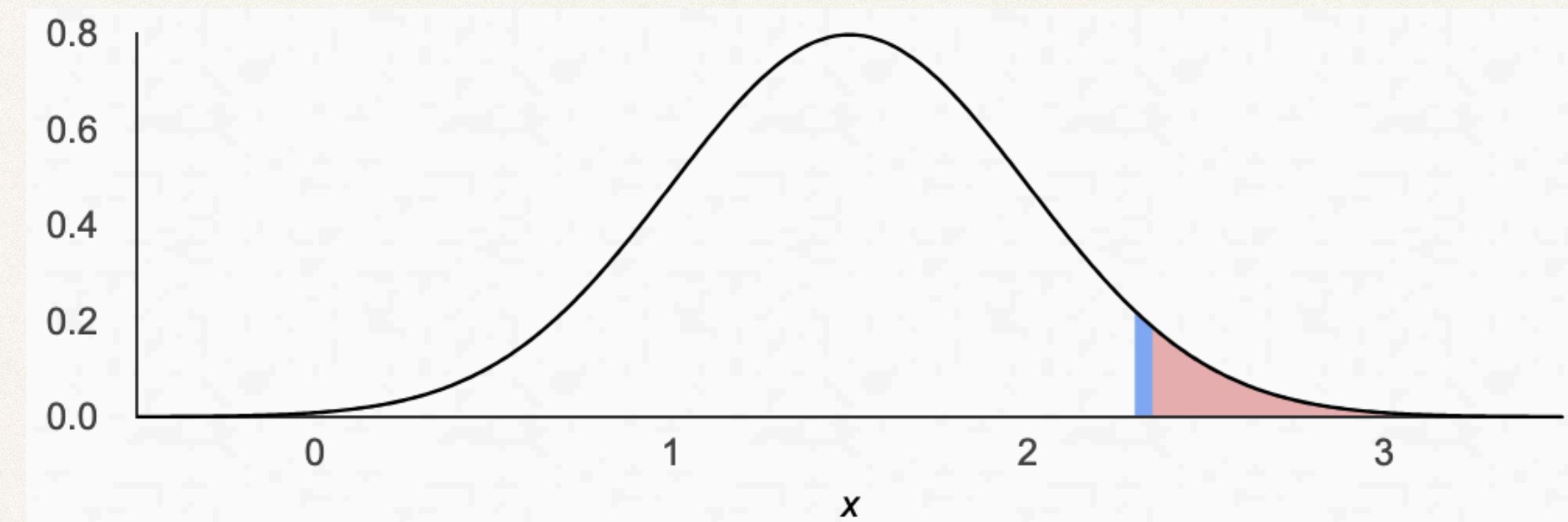
---

To make a decision, set up a rejection region by choosing the value of  $\alpha$

# SIGNIFICANCE LEVEL

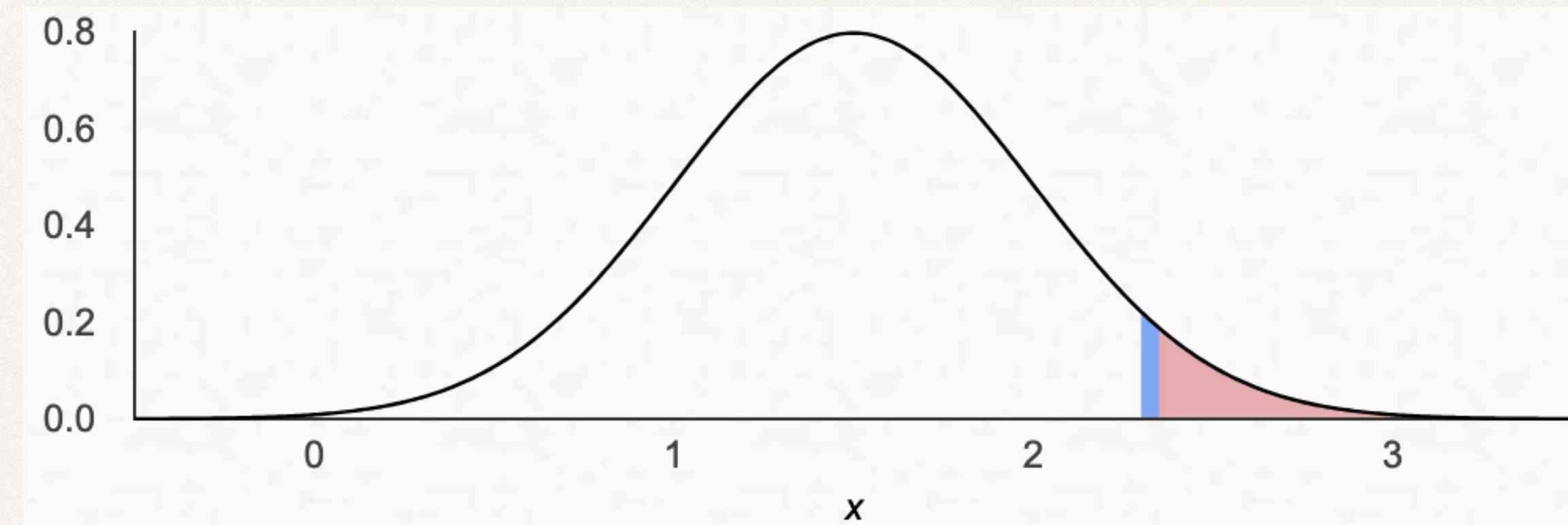
---

To make a decision, set up a rejection region by choosing the value of  $\alpha$

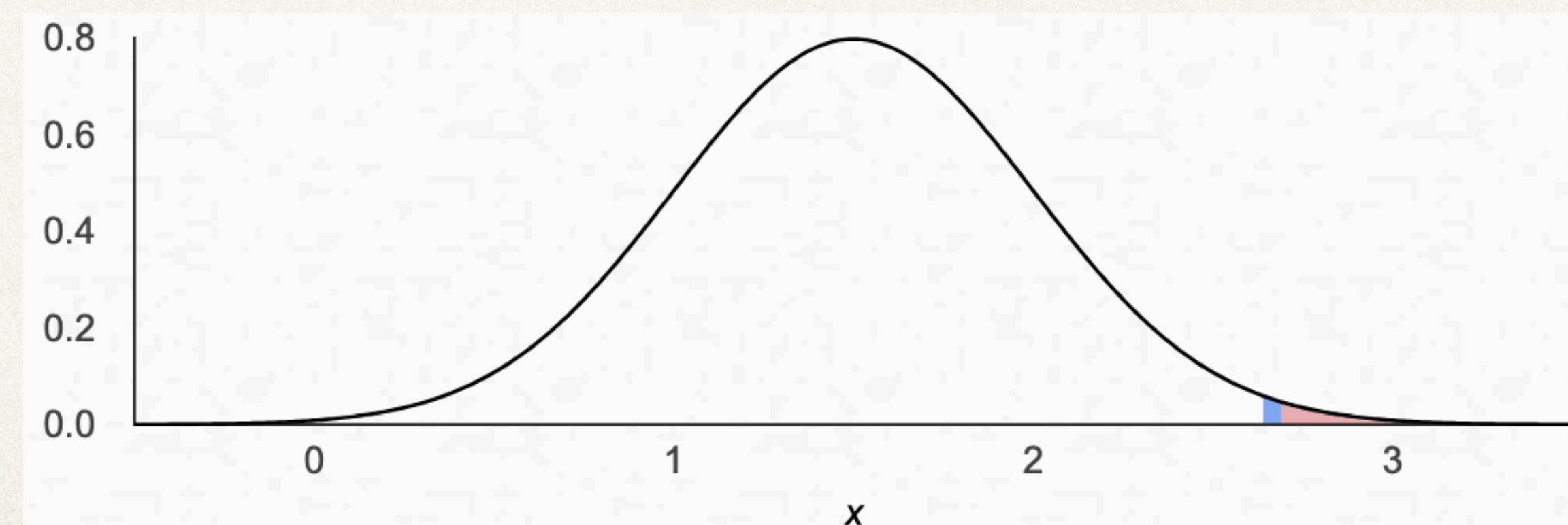


# SIGNIFICANCE LEVEL

To make a decision, set up a rejection region by choosing the value of  $\alpha$



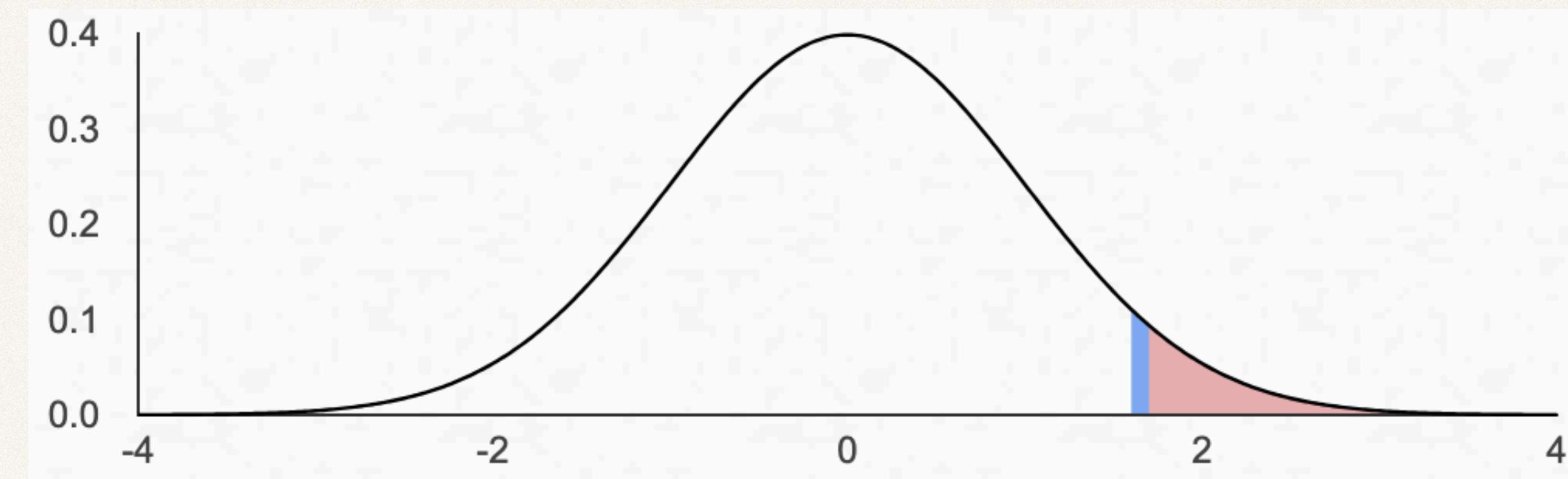
$$\alpha = 0.05$$



$$\alpha = 0.01$$

# Z-SCORE

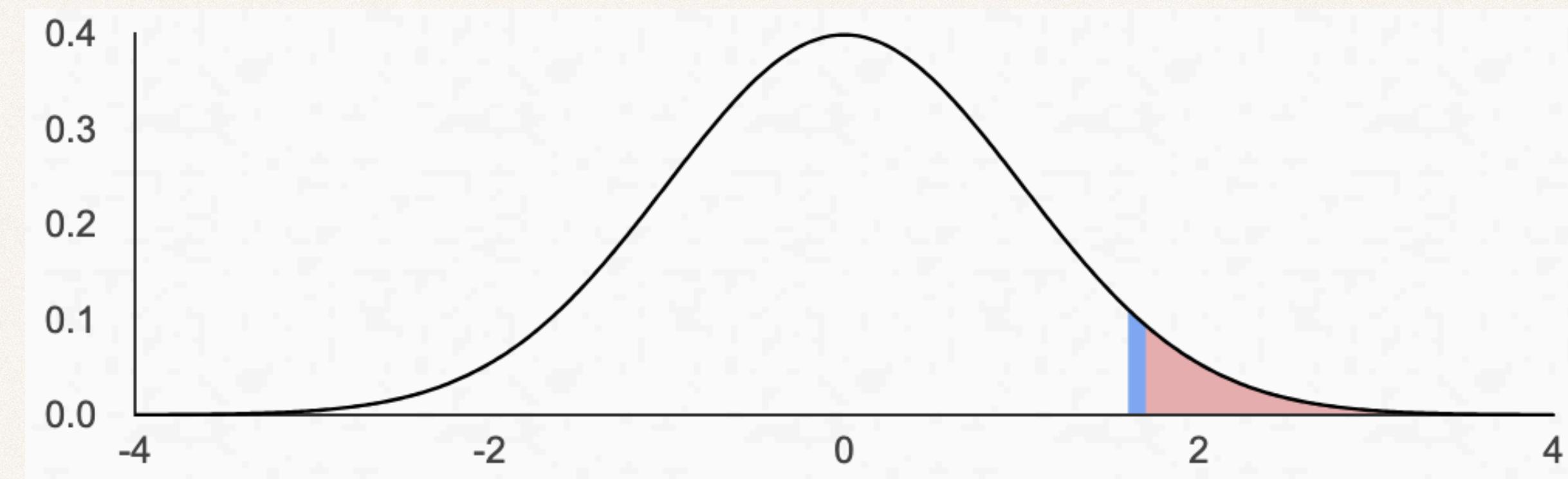
---



# Z-SCORE

---

Calculating probabilities and comparing with the significance level can be tedious

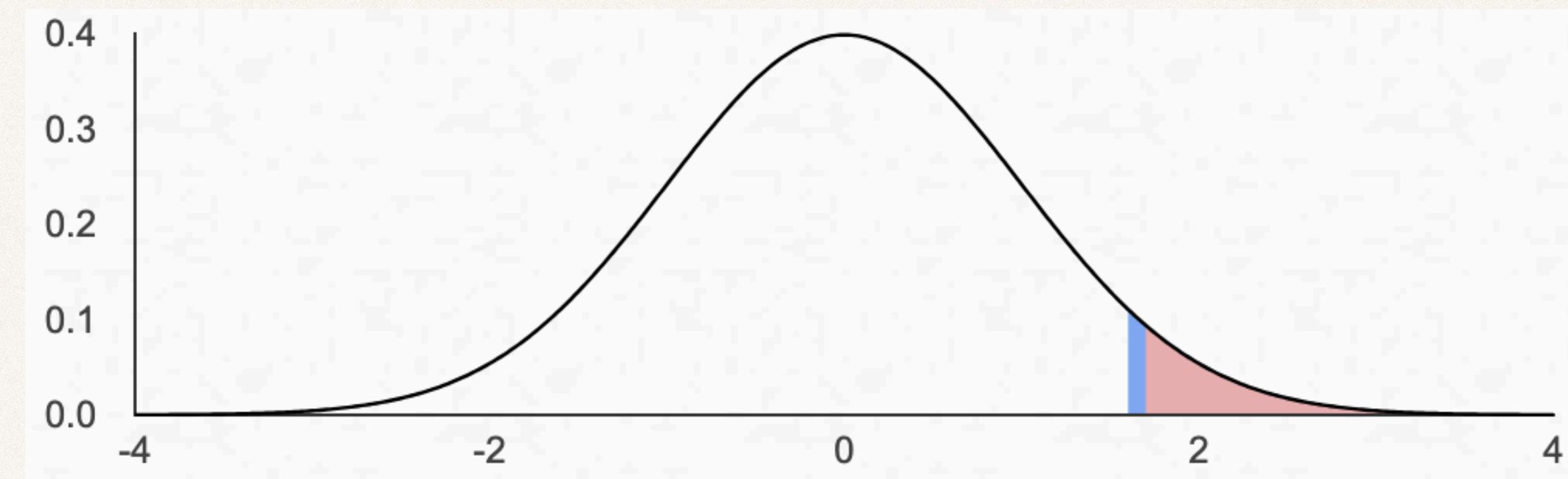


# Z-SCORE

---

Calculating probabilities and comparing with the significance level can be tedious

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

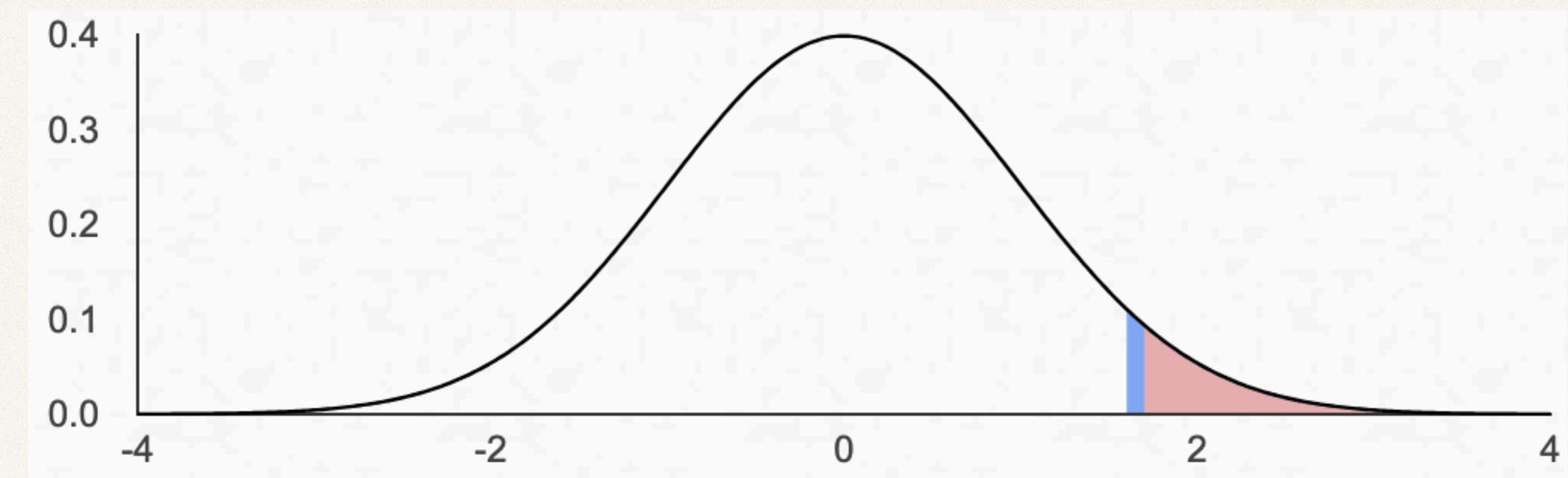


# Z-SCORE

---

Calculating probabilities and comparing with the significance level can be tedious

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

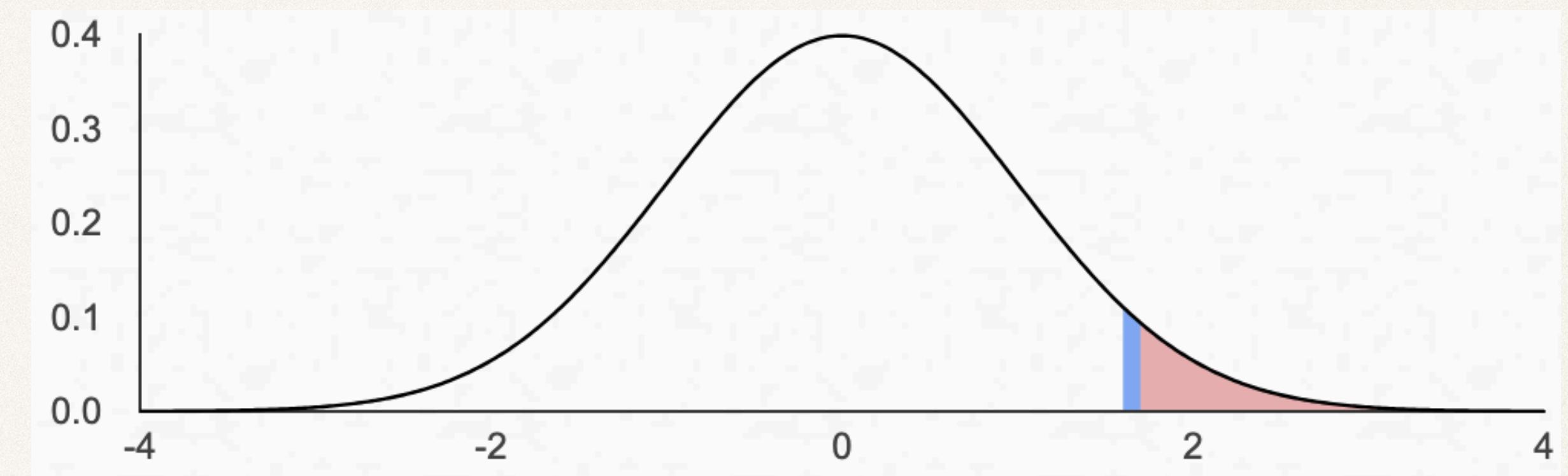


$z = 1.65$

# Z-SCORE

Calculating probabilities and comparing with the significance level can be tedious

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$



$z = 1.65$

We can simply calculate the z-score of the statistic and compare it to the z-score for the significance levels

# TAILS

---

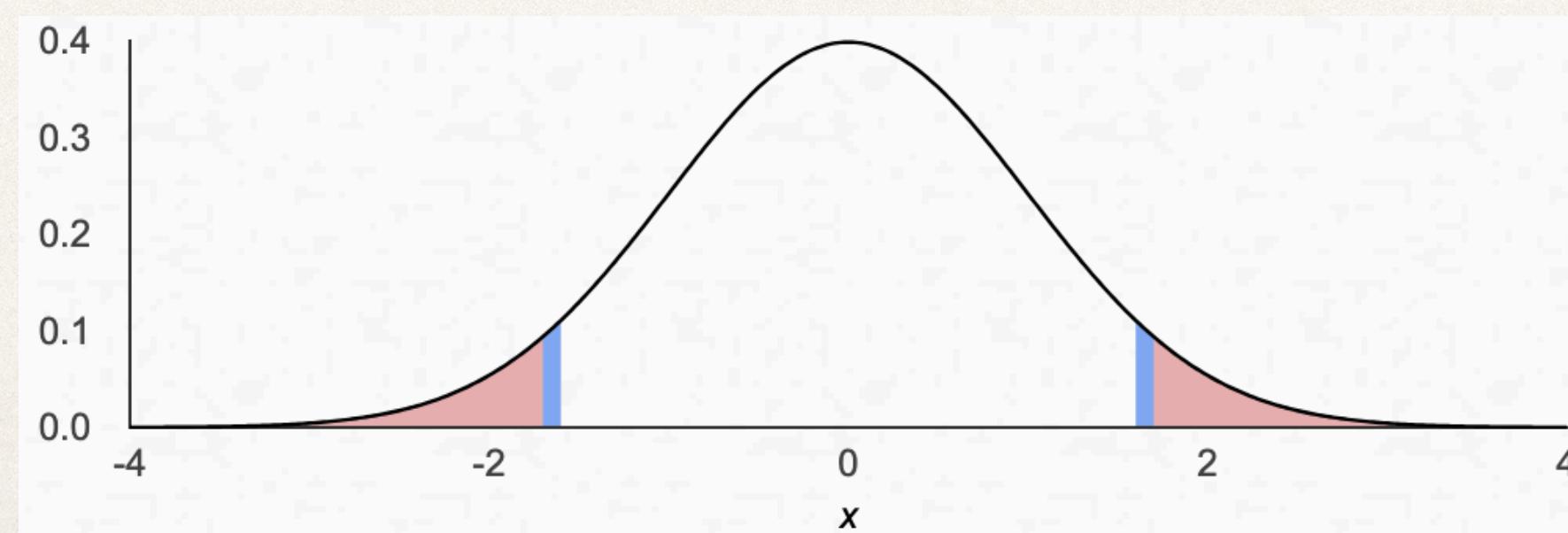
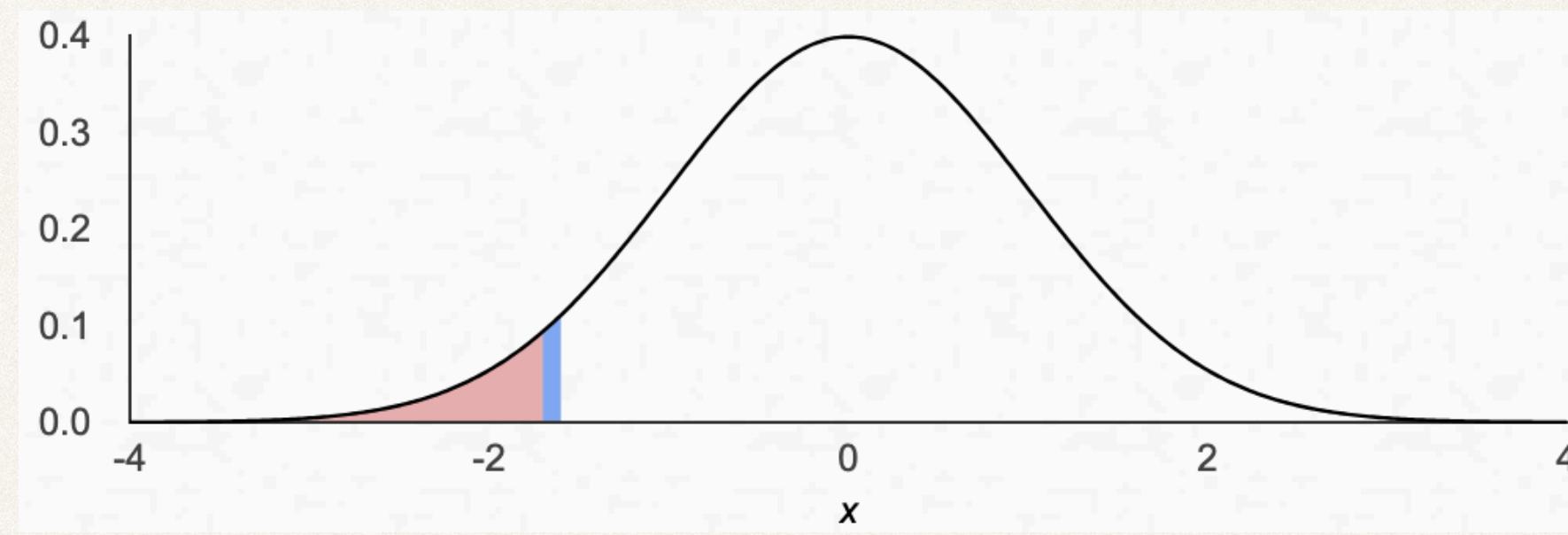
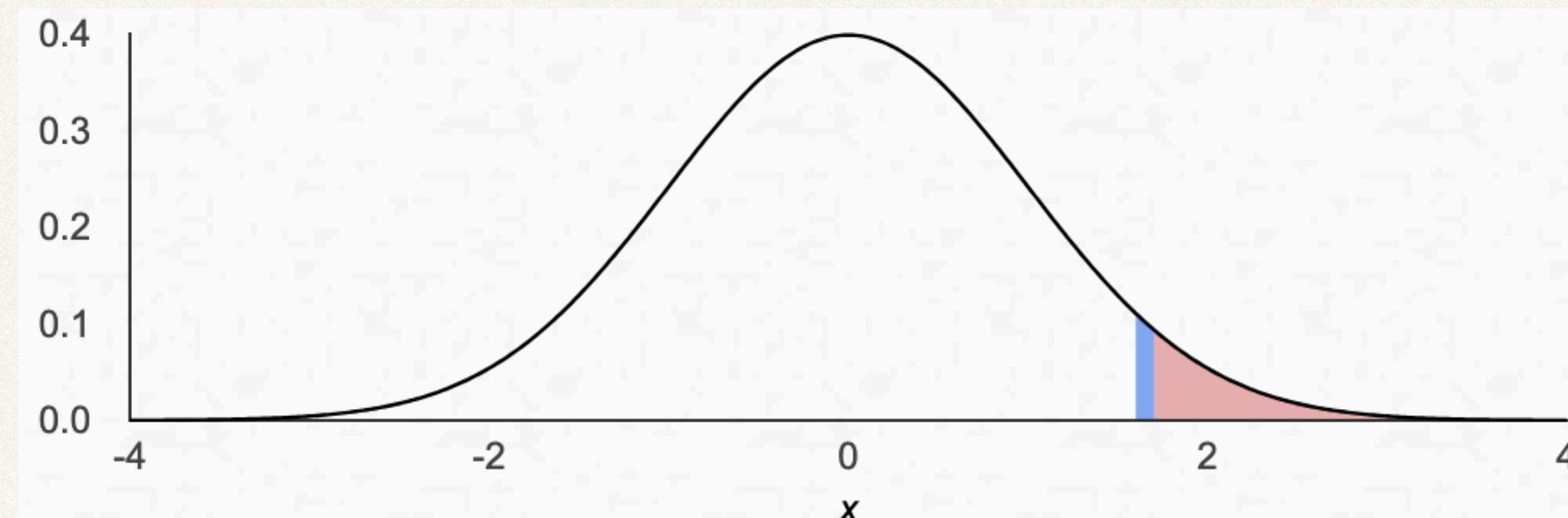
# TAILS

---

What is  
considered  
surprising is  
dependent on  
the problem at  
hand

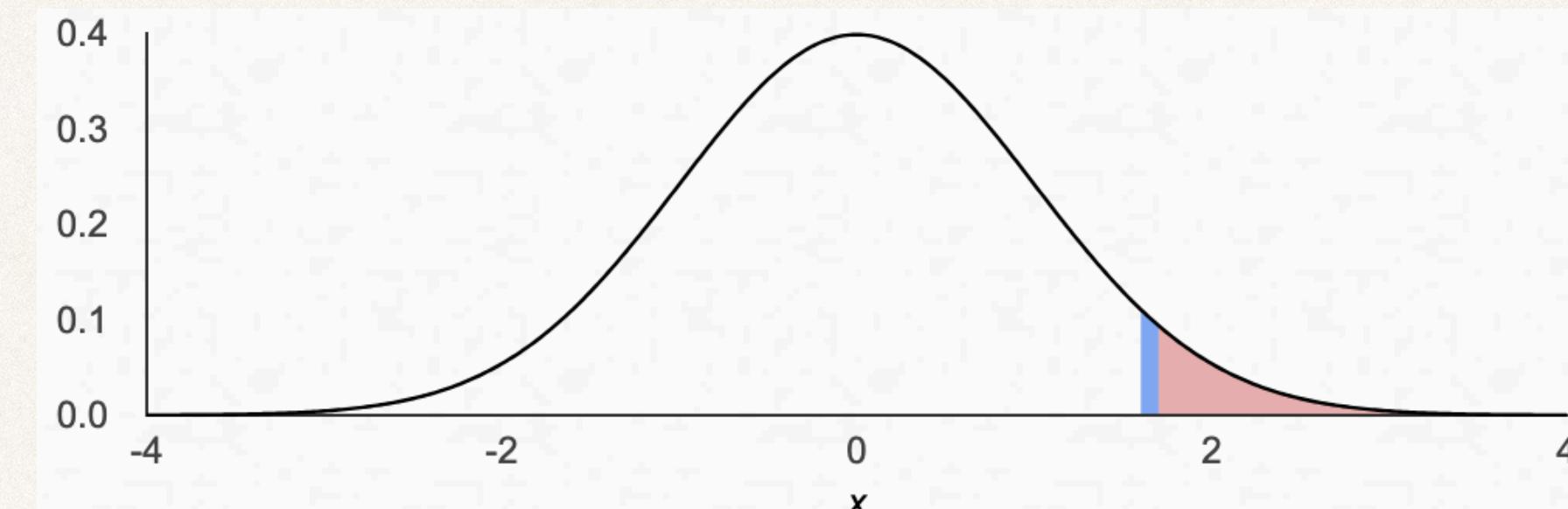
# TAILS

What is considered surprising is dependent on the problem at hand

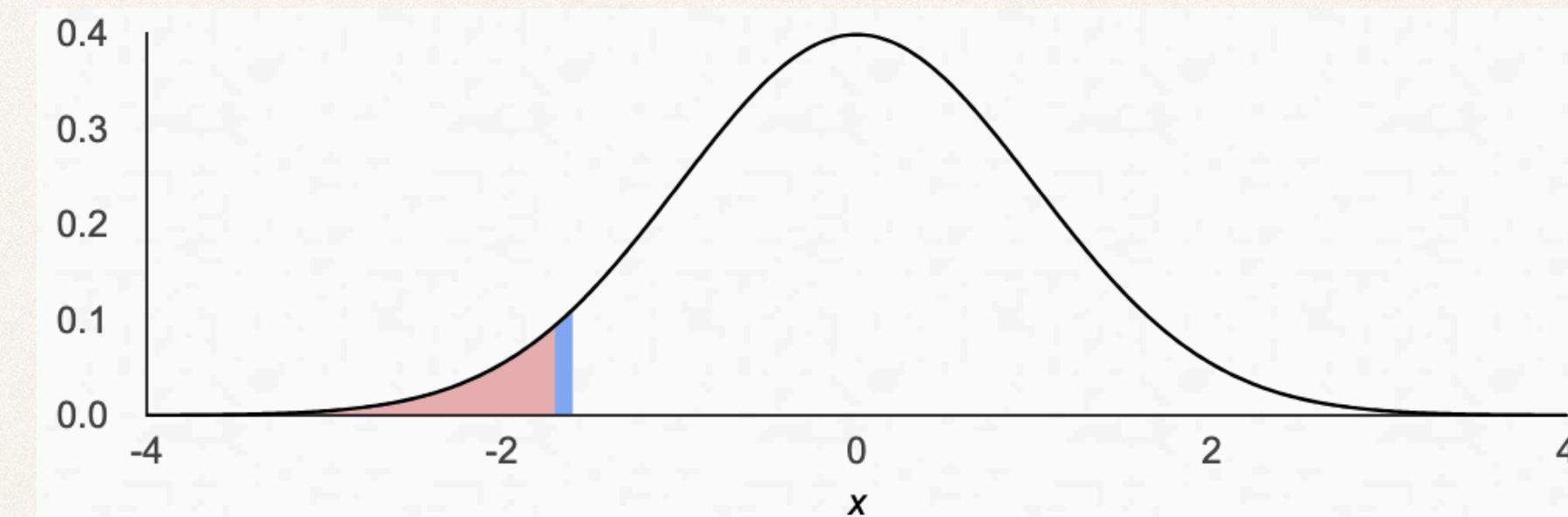


What is  
considered  
surprising is  
dependent on  
the problem at  
hand

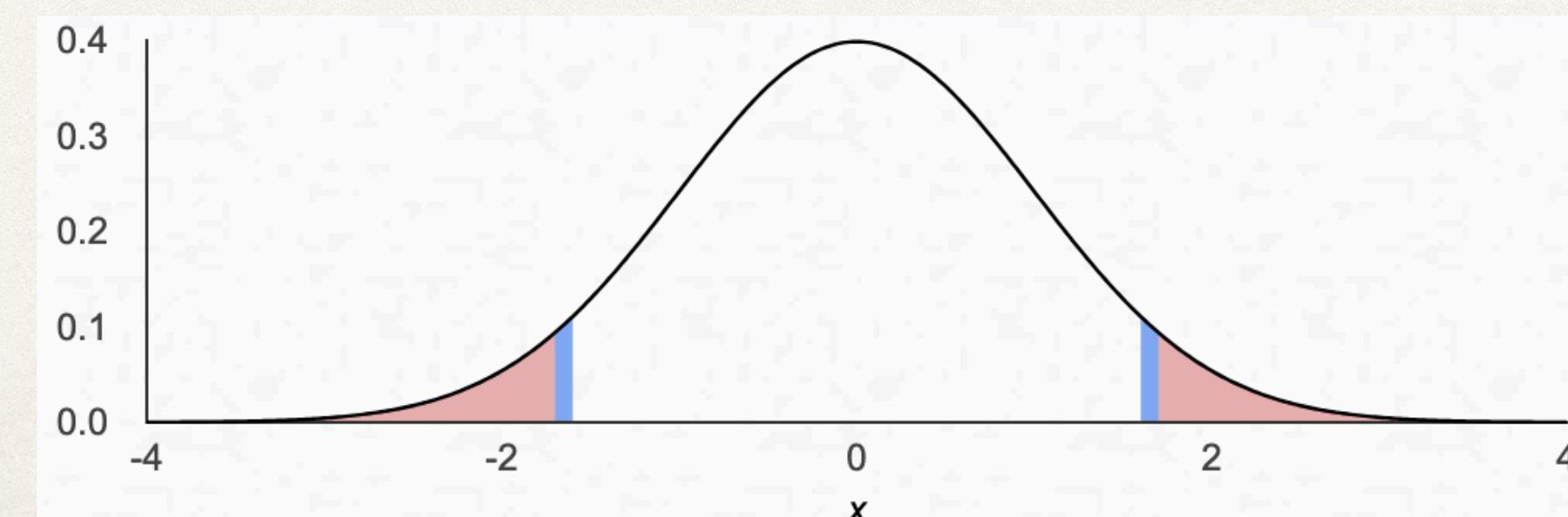
## TAILS



upper-tail: is  $x$   
significantly  
greater?



lower-tail: is  $x$   
significantly  
smaller?

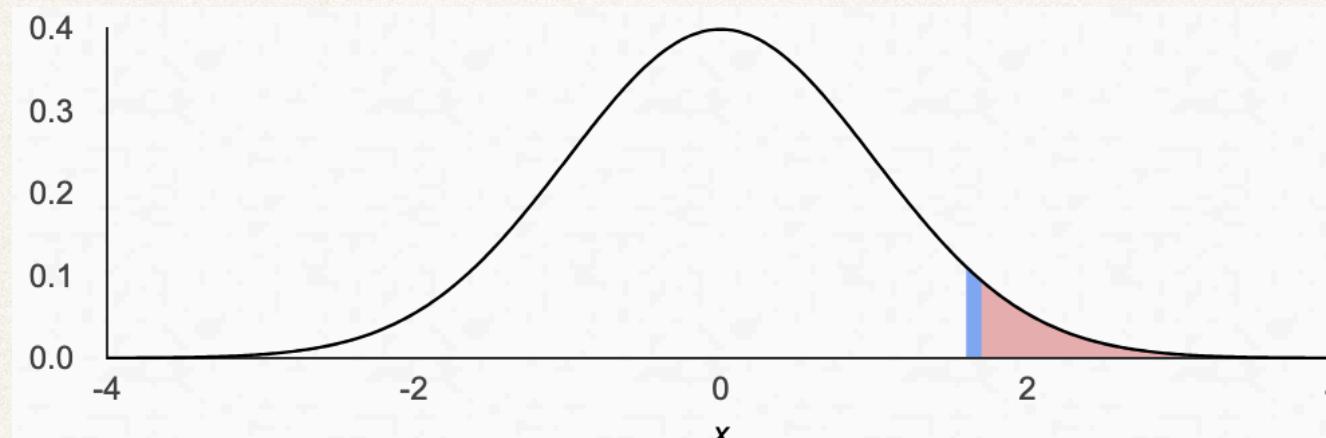


two-tail: is  $x$   
significantly  
different?

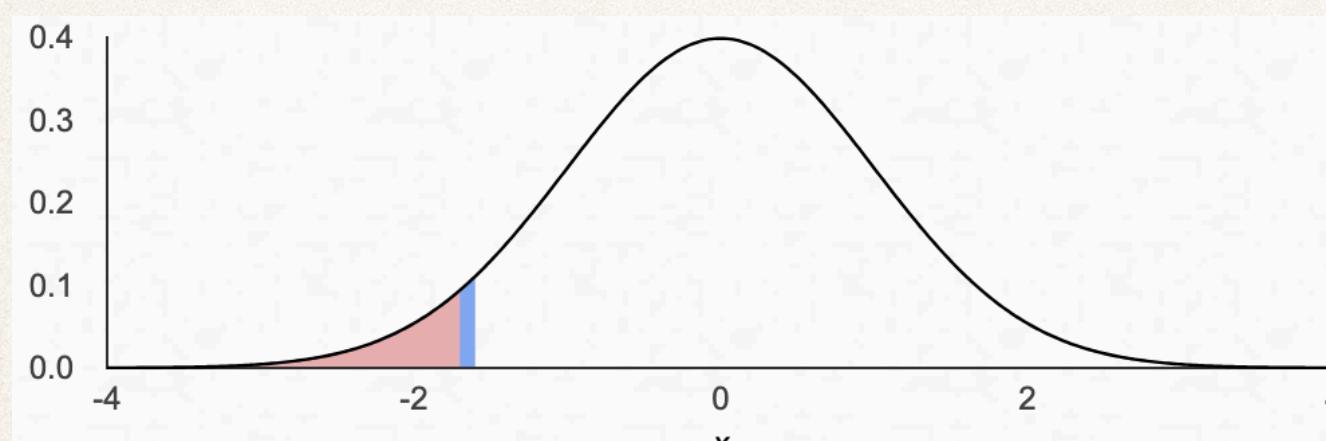
# P VALUES

---

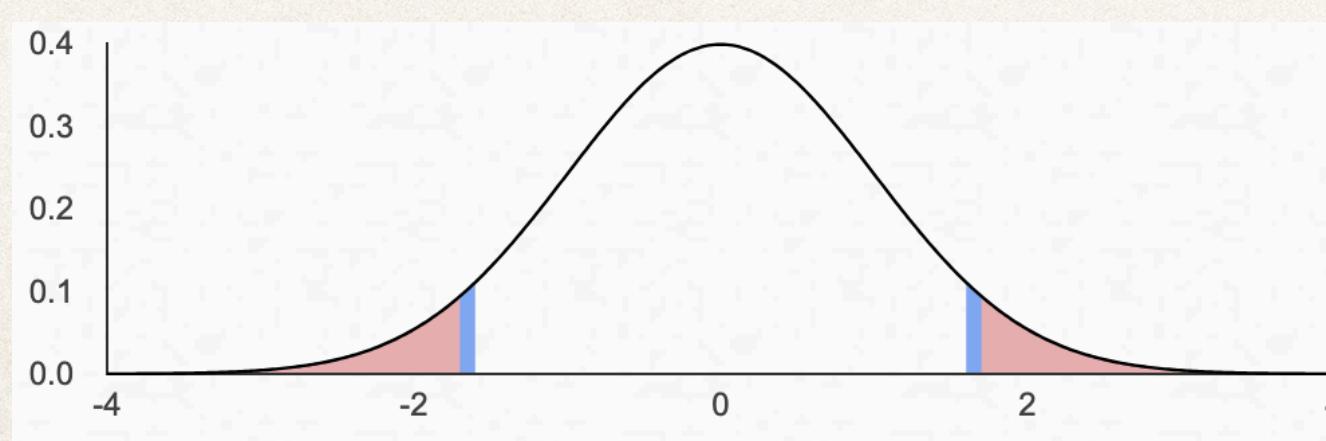
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.



$$\text{p-value } (x) = P(X \geq x | H_0) = 1 - P(X \leq x | H_0)$$



$$\text{p-value } (x) = 1 - P(X \leq x | H_0)$$



$$\text{p-value } (x) = 2 \times P(X \leq -|x| | H_0)$$

# RECIPE FOR HYPOTHESIS TESTING

---

# RECIPE FOR HYPOTHESIS TESTING

---

- Assume significance level  $\alpha$ ; state  $H_0$  and  $H_{\text{alternative}}$

# RECIPE FOR HYPOTHESIS TESTING

---

- Assume significance level  $\alpha$ ; state  $H_0$  and  $H_{\text{alternative}}$
- Calculate some statistic  $x$  whose conditional distribution is given by  $P(X|H_0)$

# RECIPE FOR HYPOTHESIS TESTING

---

- Assume significance level  $\alpha$ ; state  $H_0$  and  $H_{\text{alternative}}$
- Calculate some statistic  $x$  whose conditional distribution is given by  $P(X|H_0)$
- Calculate p-value

# RECIPE FOR HYPOTHESIS TESTING

---

- Assume significance level  $\alpha$ ; state  $H_0$  and  $H_{\text{alternative}}$
- Calculate some statistic  $x$  whose conditional distribution is given by  $P(X|H_0)$
- Calculate p-value
- If p-value falls in rejection region then null hypothesis can be rejected; else null hypothesis cannot be rejected

# ERRORS

---

# ERRORS

---

- Type I error: We incorrectly rejected the null hypothesis

# ERRORS

---

- Type I error: We incorrectly rejected the null hypothesis
- Type II error: We incorrectly failed to reject the null hypothesis

# ERRORS

- Type I error: We incorrectly rejected the null hypothesis
- Type II error: We incorrectly failed to reject the null hypothesis

		Test results	
		keep null	reject null
Truth	keep null	Type I error $\alpha$	Power
	reject null	Type II error $\beta$	



The Boy who Cried Wolf

- In the Aesop's fable –  
The Boy who Cried Wolf  
– the villagers make  
two types of errors
- They believe there  
was a wolf when there  
was none
- They believe there  
was no wolf when  
there was one



The Boy who Cried Wolf



The Boy who Cried Wolf

- In the Aesop's fable –  
The Boy who Cried Wolf –  
the villagers make two  
types of errors
  - They believe there was  
a wolf when there was  
none [TYPE I]
  - They believe there was  
no wolf when there was  
one [TYPE II]



The Boy who Cried Wolf

# ERRORS

---

# ERRORS

---

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power

# ERRORS

---

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power
- If  $\alpha = 0.05$  and we do 100 tests, we expect to make 5 mistakes

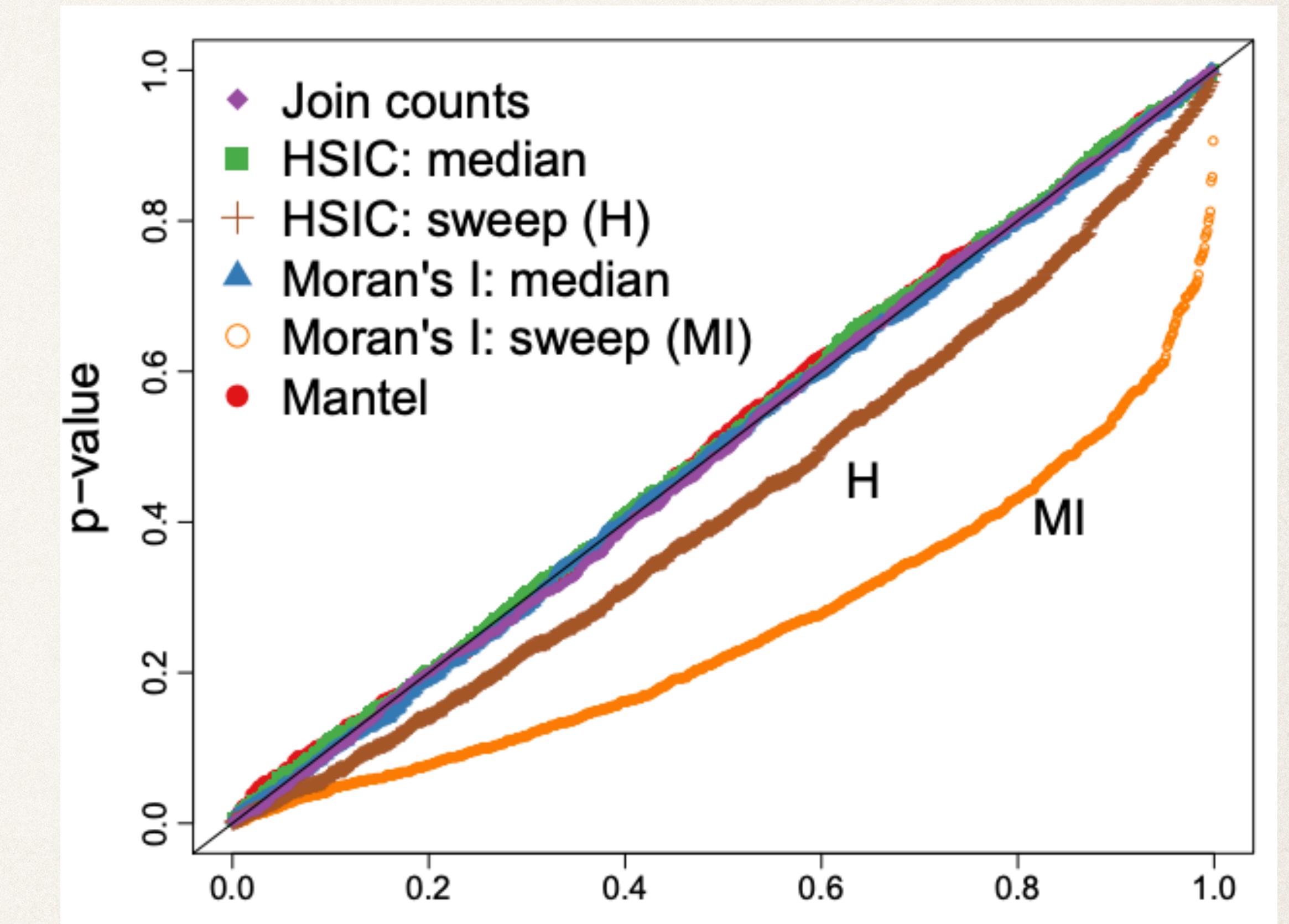
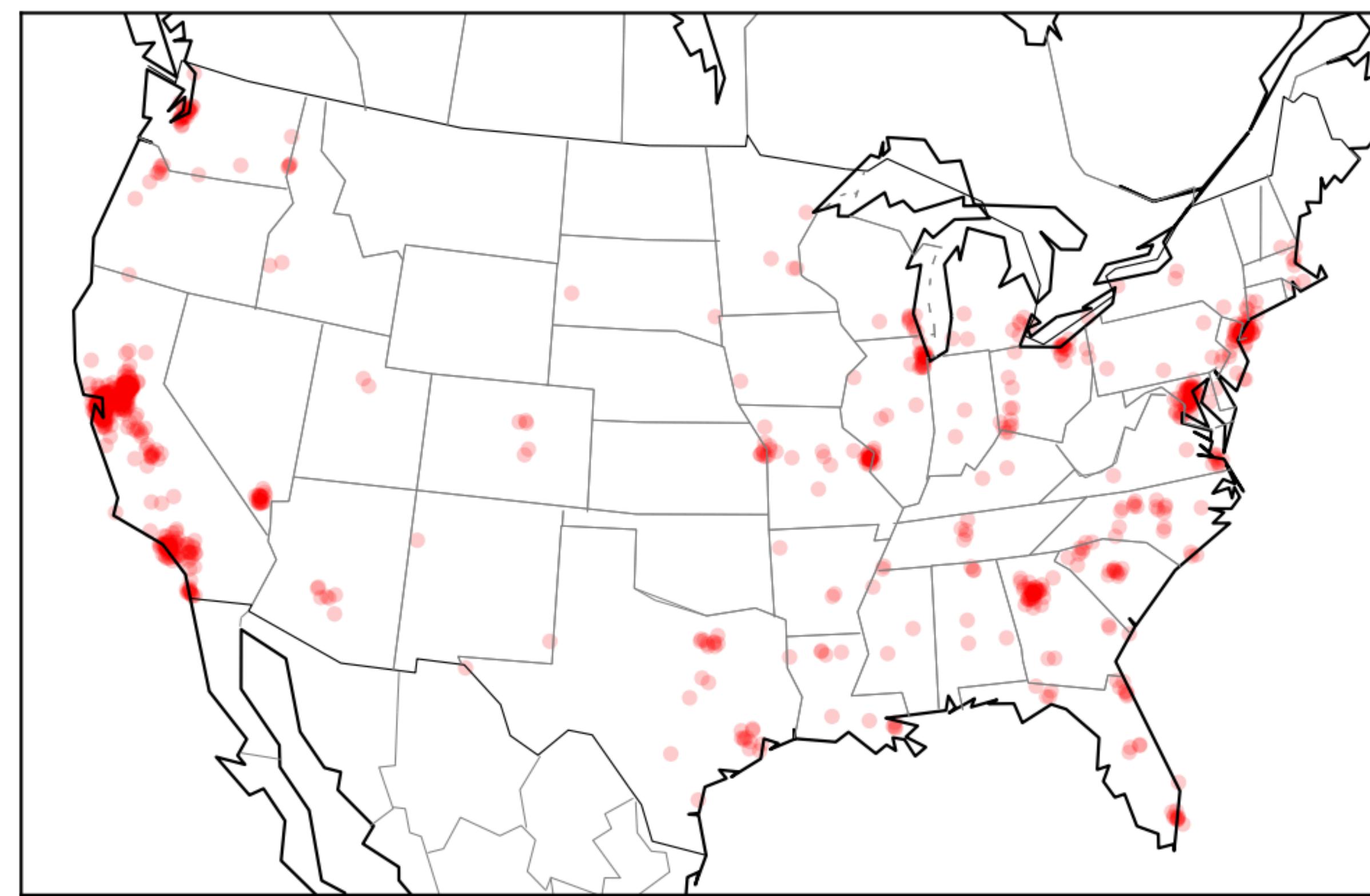
# ERRORS

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power
- If  $\alpha = 0.05$  and we do 100 tests, we expect to make 5 mistakes

		Test results	
		keep null	reject null
Truth	keep null	Type I error $\alpha$	Type II error $\beta$
	reject null		Power



# Test if lexical variation is independent of geography



# MULTIPLE HYPOTHESIS CORRECTIONS

---

# MULTIPLE HYPOTHESIS CORRECTIONS

---

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance

# MULTIPLE HYPOTHESIS CORRECTIONS

---

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance
- Apply bonferroni correction which conservatively sets a lower significance threshold based on number of tests

# MULTIPLE HYPOTHESIS CORRECTIONS

---

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance
- Apply bonferroni correction which conservatively sets a lower significance threshold based on number of tests

$$\alpha = \frac{\alpha_0}{n}$$

Here the significance level  $\alpha_0$  is adjusted

# CONFIDENCE INTERVALS

---

# CONFIDENCE INTERVALS

---

- In many instances, instead of doing a statistical test, we want to bound the error of the metric

# CONFIDENCE INTERVALS

---

- In many instances, instead of doing a statistical test, we want to bound the error of the metric
- Confidence intervals helps us quantify the range in which the observed metric will lie for the unobserved population

# CONFIDENCE INTERVALS

---

# CONFIDENCE INTERVALS

---

- The CI is statistically determined with some parametric assumptions

# CONFIDENCE INTERVALS

---

- The CI is statistically determined with some parametric assumptions
- The observed value is assumed to be a point estimate of the mean

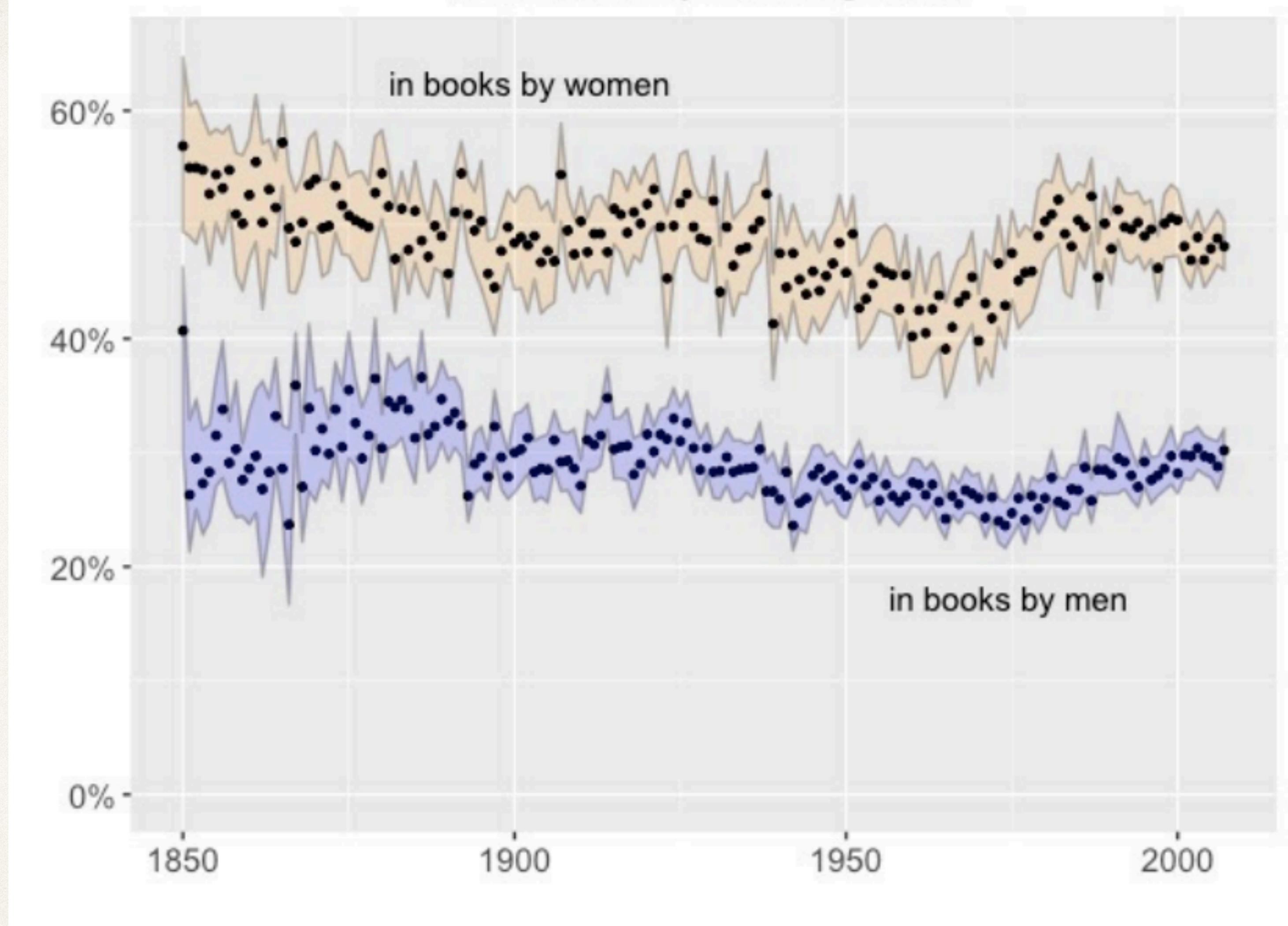
# CONFIDENCE INTERVALS

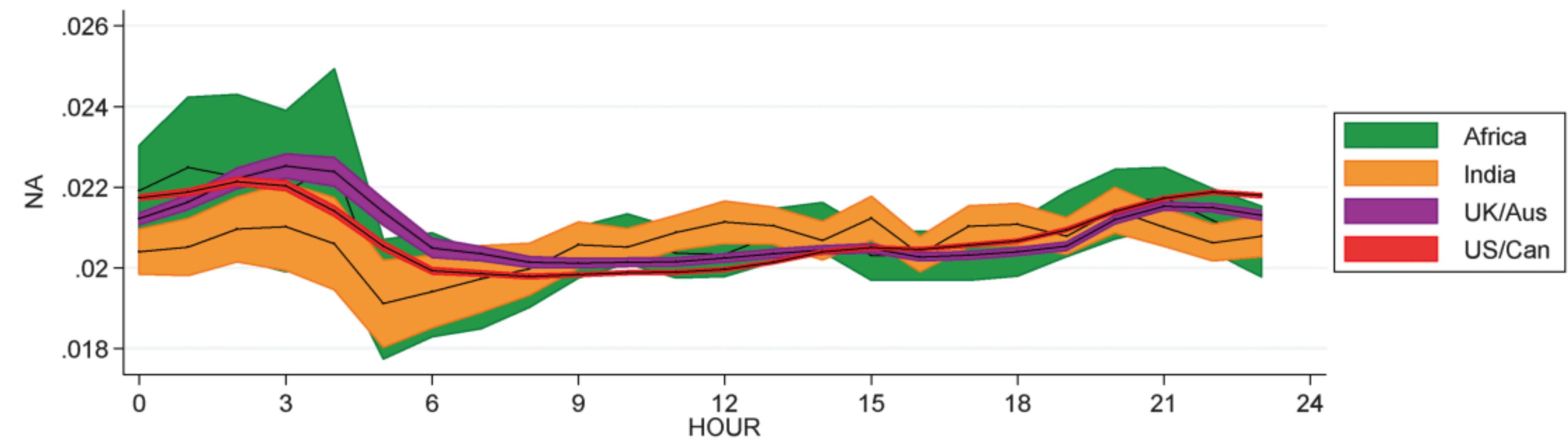
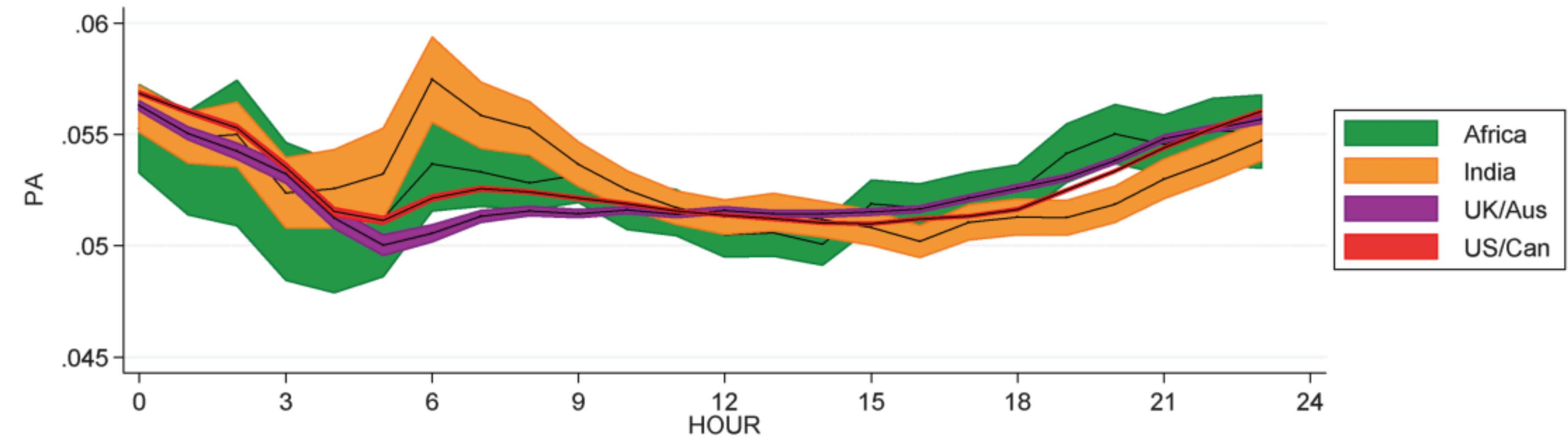
- The CI is statistically determined with some parametric assumptions
  - The observed value is assumed to be a point estimate of the mean

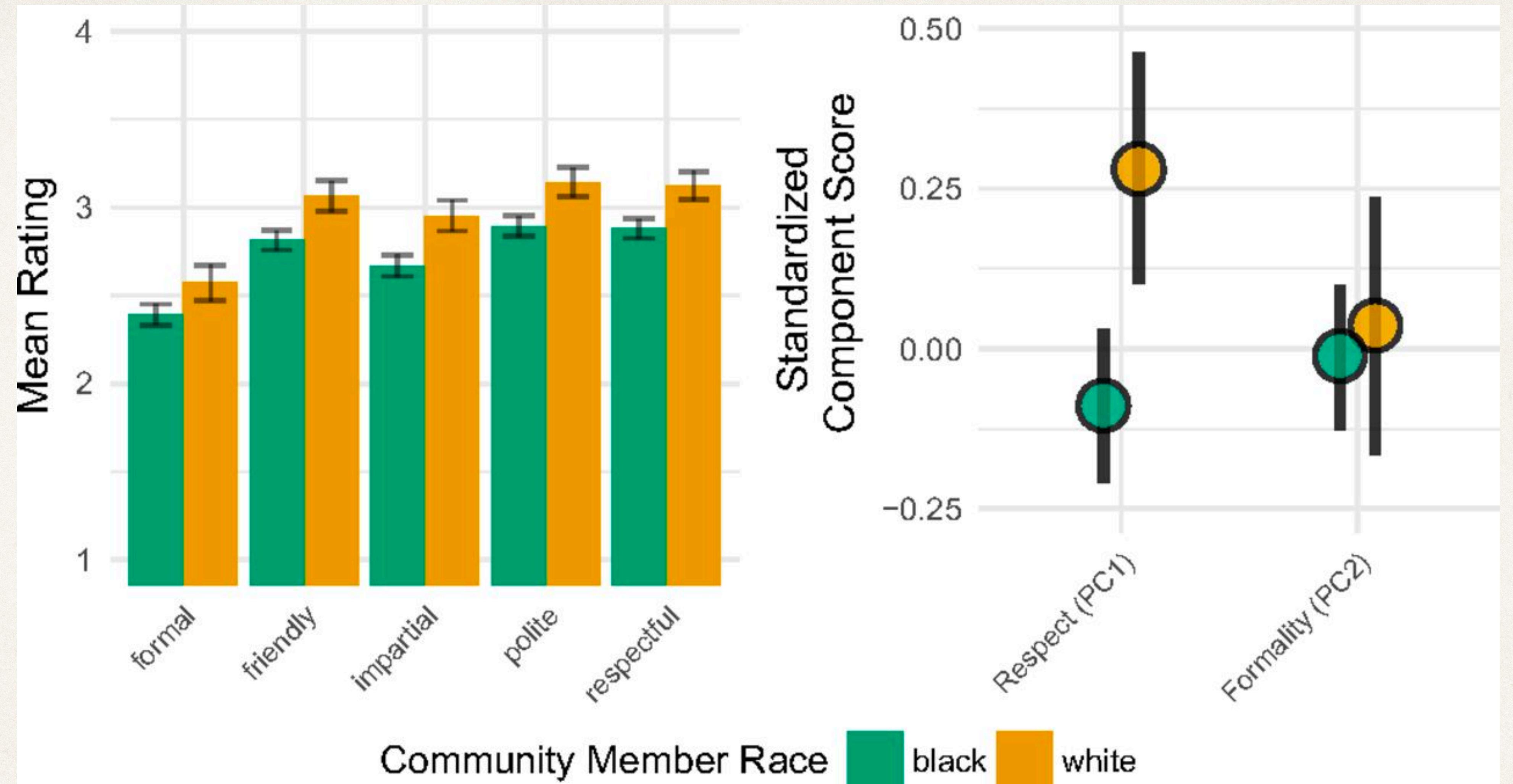
$$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

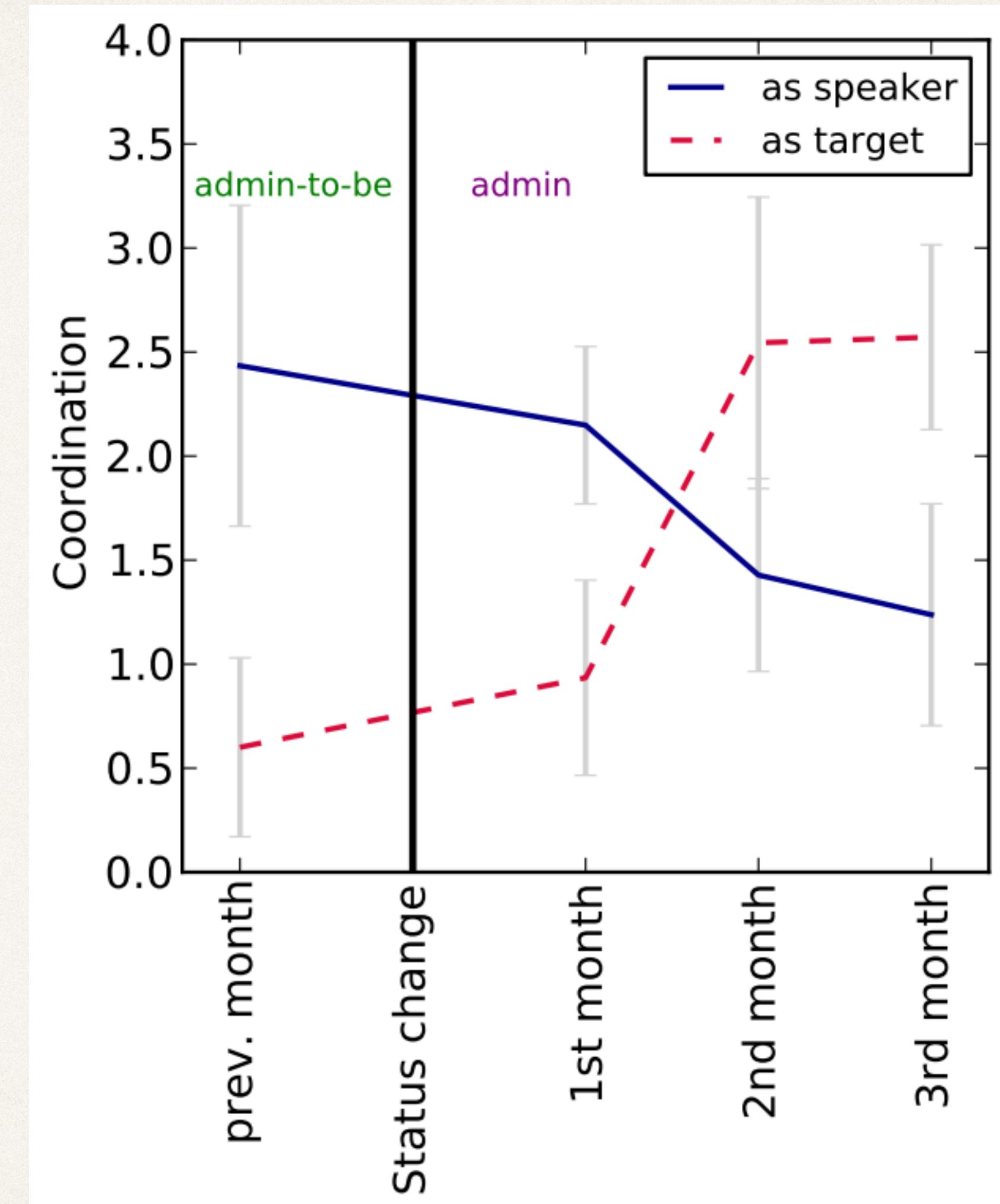
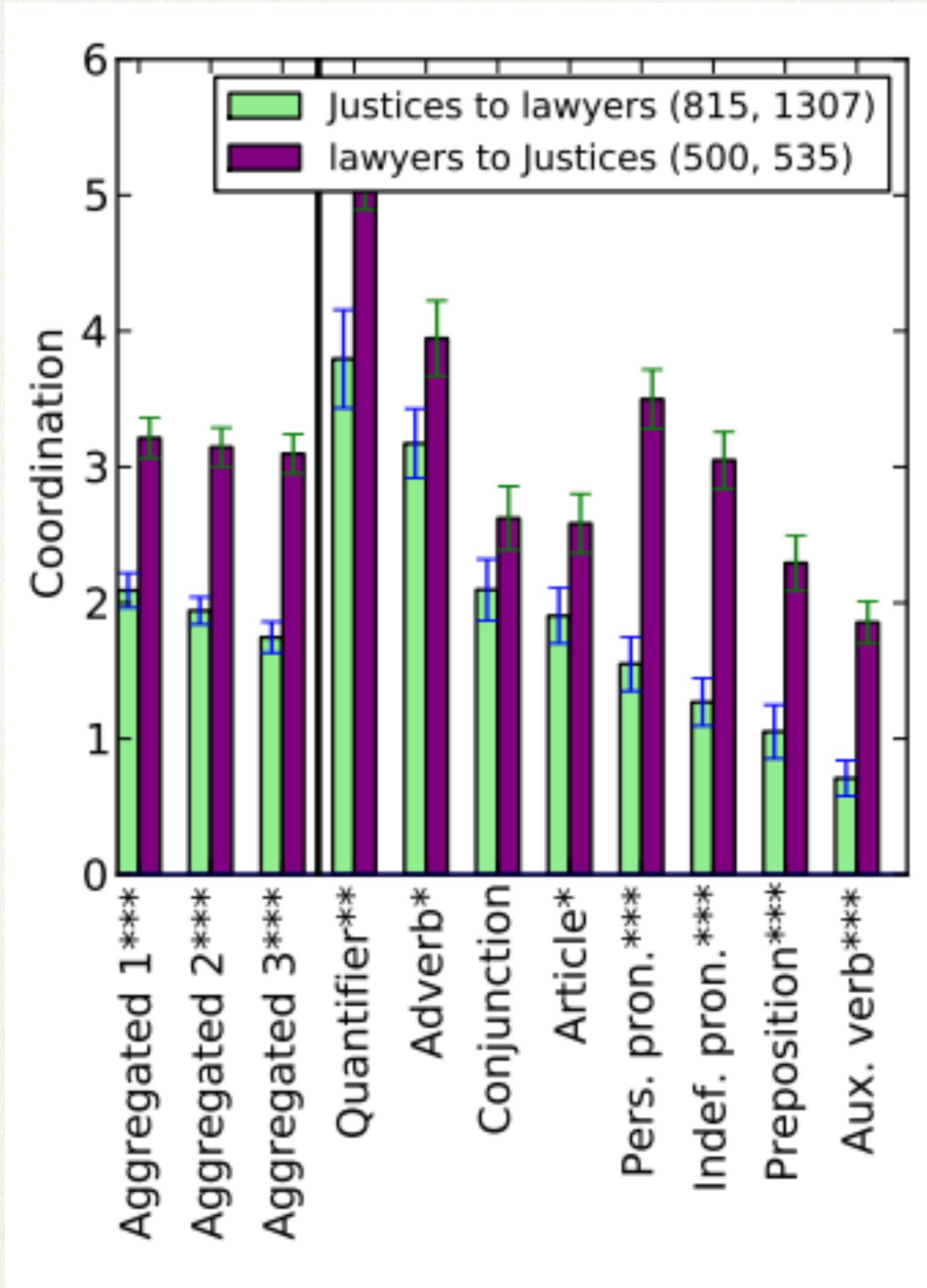
Mean value      Lower/Upper limit      z-value for the confidence level      Sample size

## Description of women, as a percentage of characterization, broken out by author gender









# ABLATION TESTING

---

# ABLATION TESTING

---

- Are a set of features really important for the model?

# ABLATION TESTING

---

- Are a set of features really important for the model?
- One can statistically test this:
  - Create model with all the features and evaluate
  - Create another model with test features removed and evaluate
  - Compare the difference in performance as a statistic and check for its statistical significance

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	<b>78.7</b>	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	<b>82.9</b>
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	<b>82.7</b>
(4)	bigrams	16165	pres.	77.3	<b>77.4</b>	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	<b>81.9</b>
(6)	adjectives	2633	pres.	77.0	<b>77.7</b>	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	<b>81.4</b>
(8)	unigrams+position	22430	pres.	81.0	80.1	<b>81.6</b>

# SUMMARY

---

# SUMMARY

---

- It's often a good idea to give a confidence interval for your estimated metric

# SUMMARY

---

- It's often a good idea to give a confidence interval for your estimated metric
- Statistical tests are useful to verify claims

# SUMMARY

---

- It's often a good idea to give a confidence interval for your estimated metric
- Statistical tests are useful to verify claims
- Use ablation testing to assess the importance of model components

# IN CLASS

---

- Parametric test