Hypothesis Testing I

Sandeep Soni

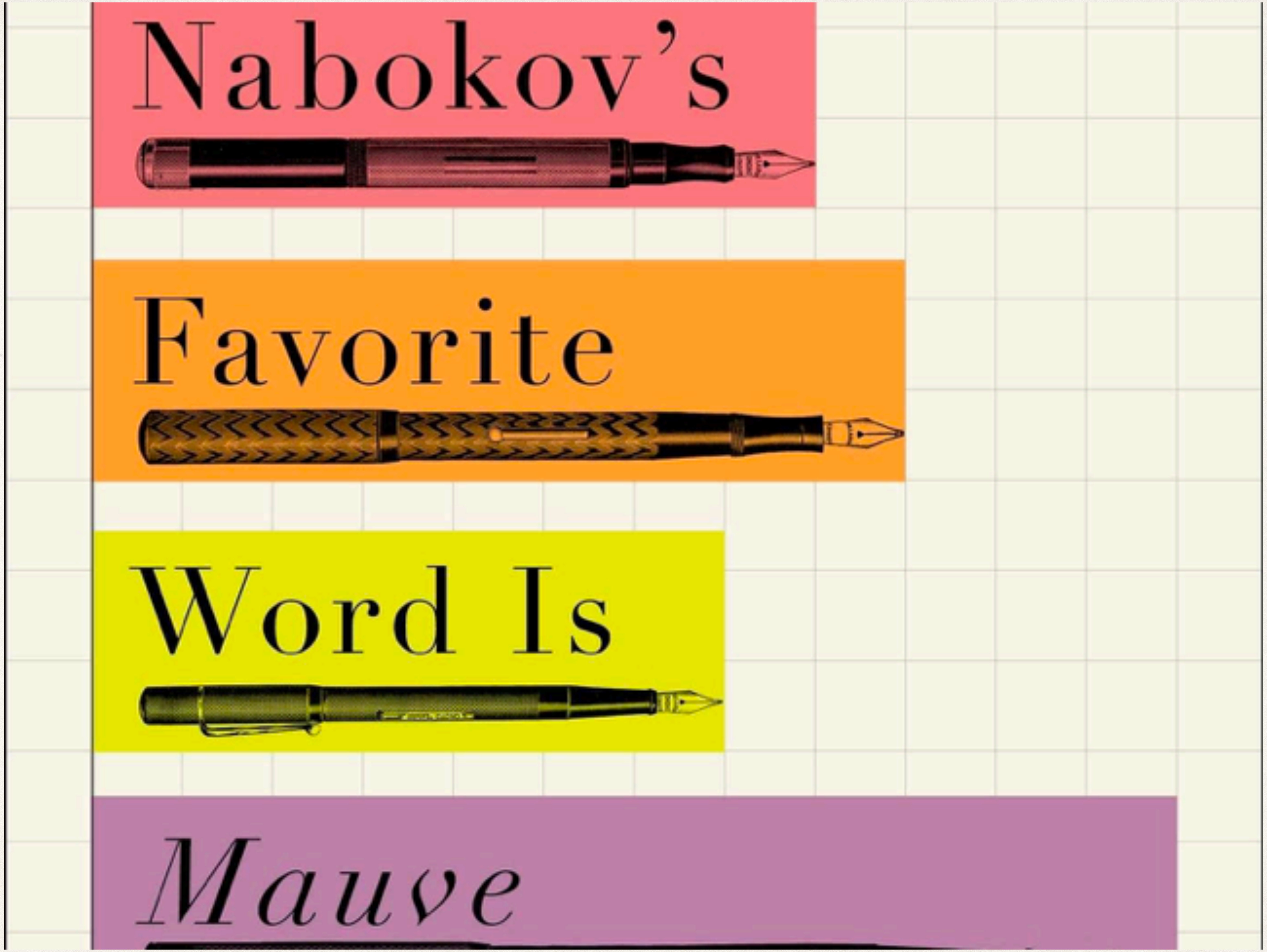
10/02/2023

Class logistics

- PS1 due today

Question for the day

“How to statistically test a claim using text?”



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 same
Reframing a tweet can increase its retweet rate	Reframing a tweet has no effect on retweet rate
Institutional mistrust is predictive of hate speech	Institutional mistrust is not predictive of hate speech
Classifier A is better than classifier B	Classifier A is the same 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?

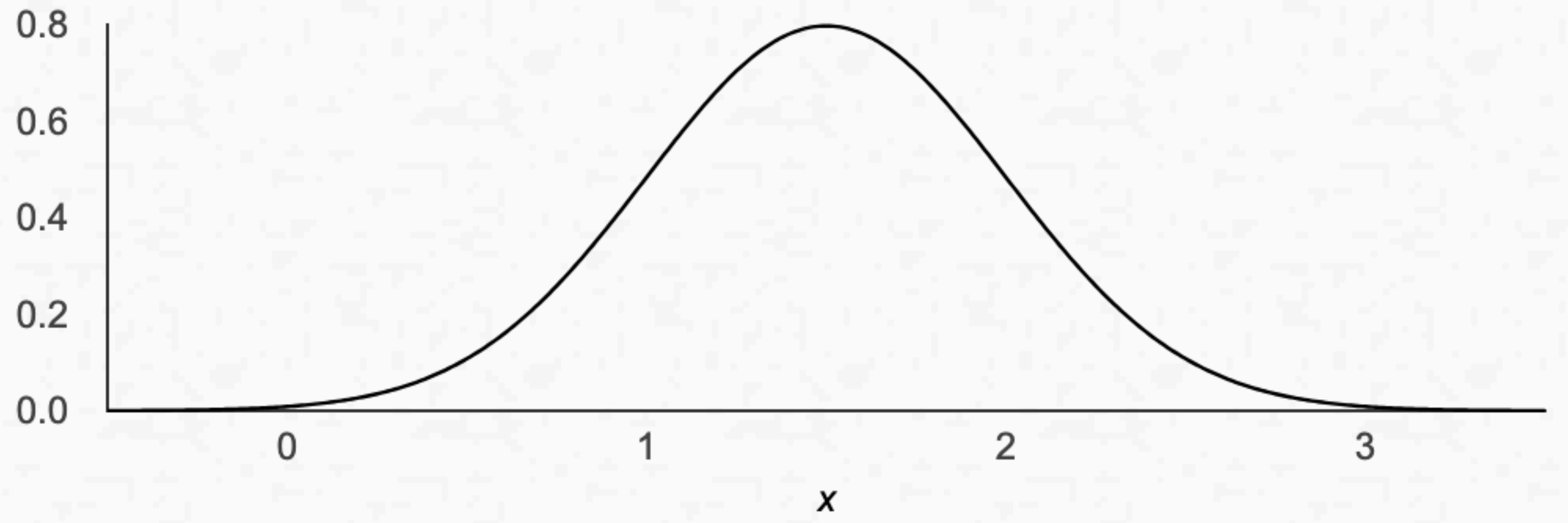
Hypothesis testing

Hypothesis testing

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

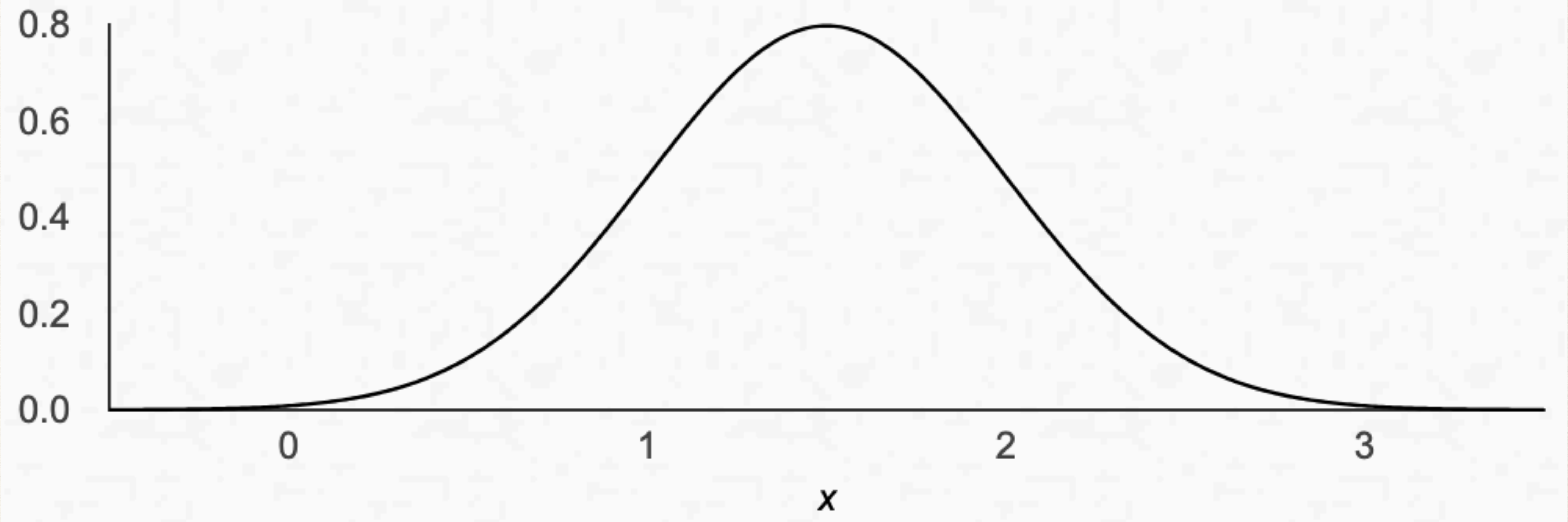
Hypothesis testing

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

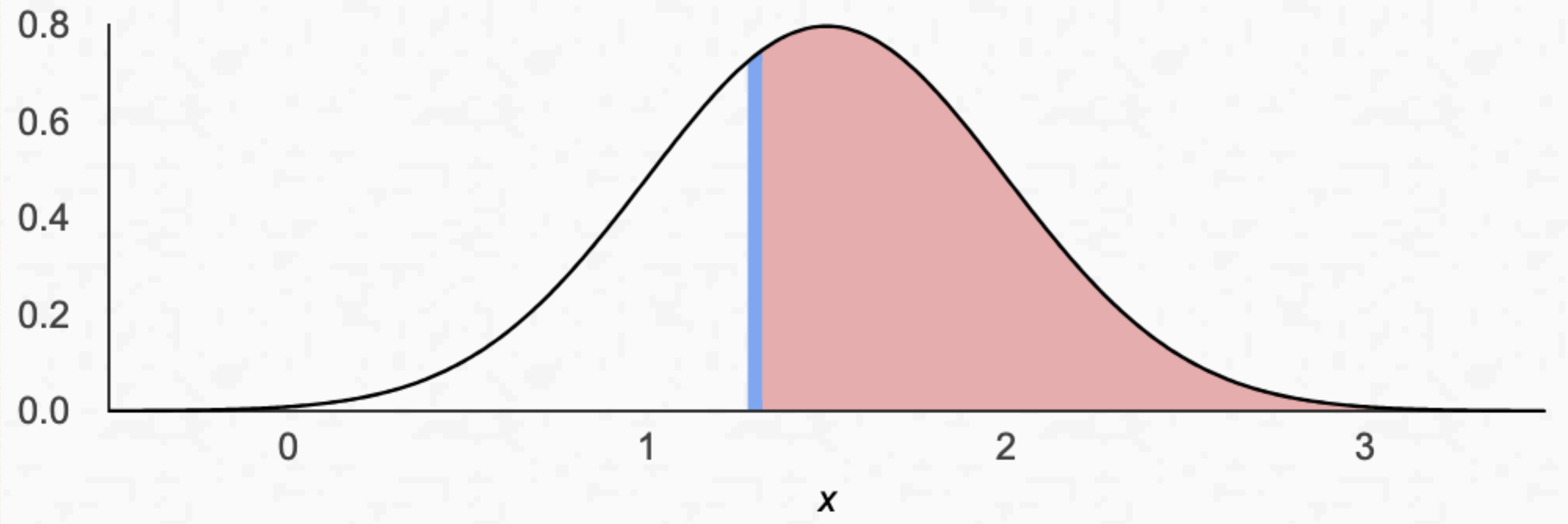
Hypothesis testing

Hypothesis testing

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

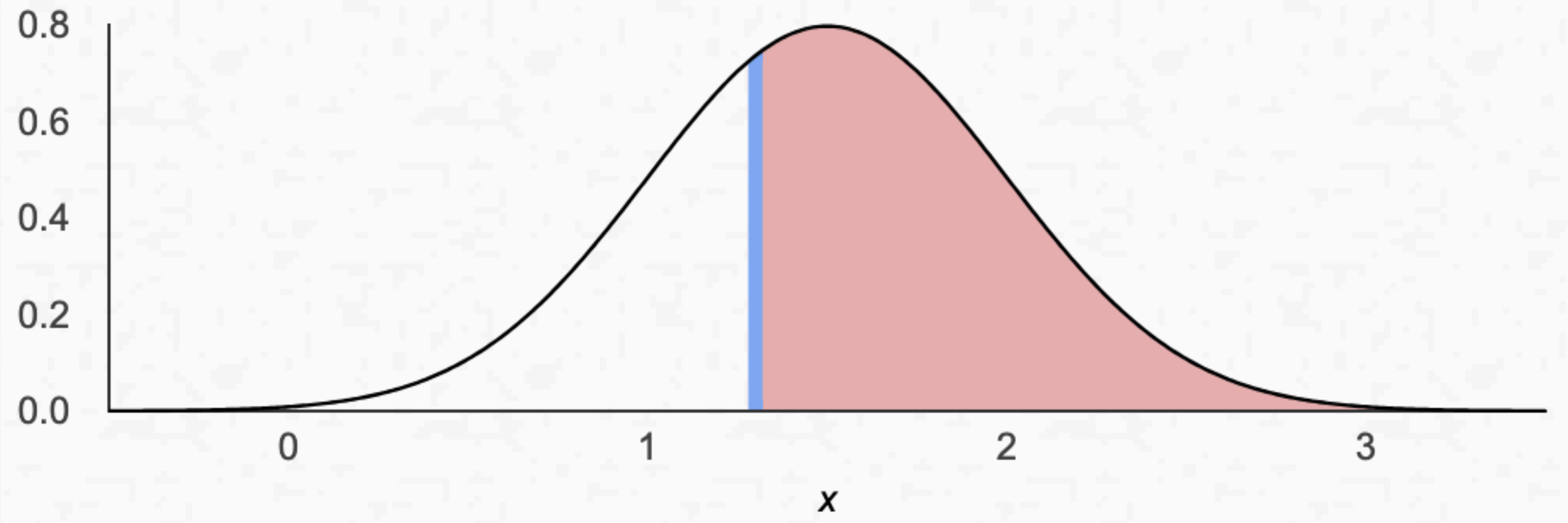
Hypothesis testing

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

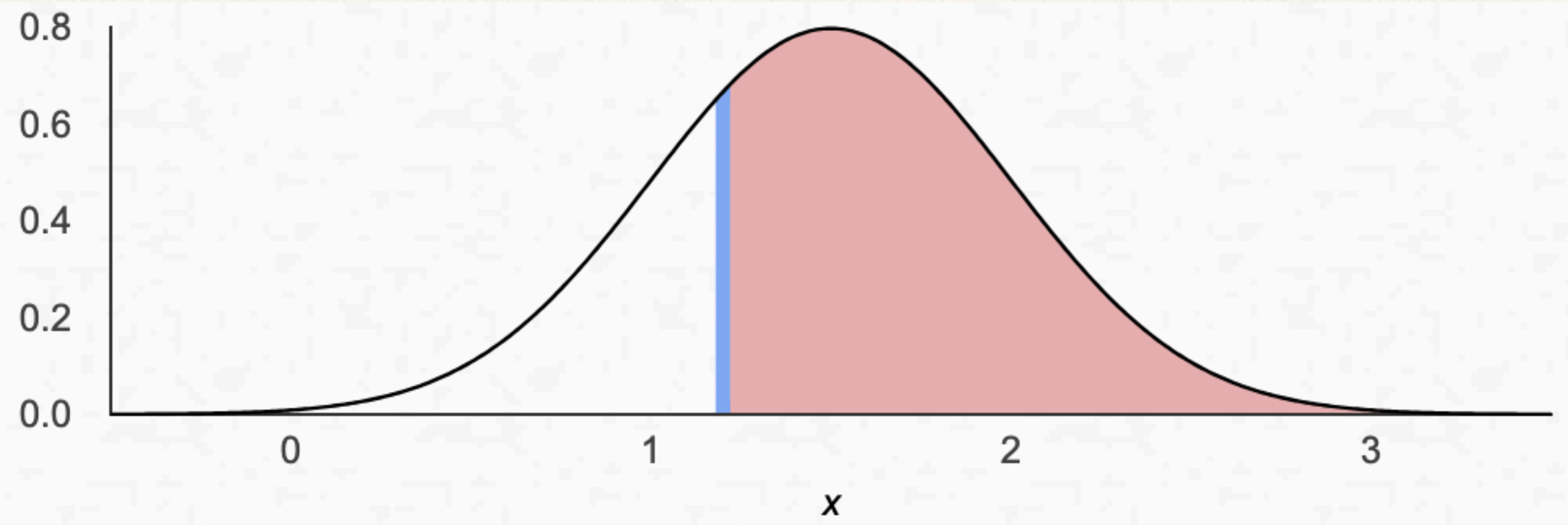
Hypothesis testing

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

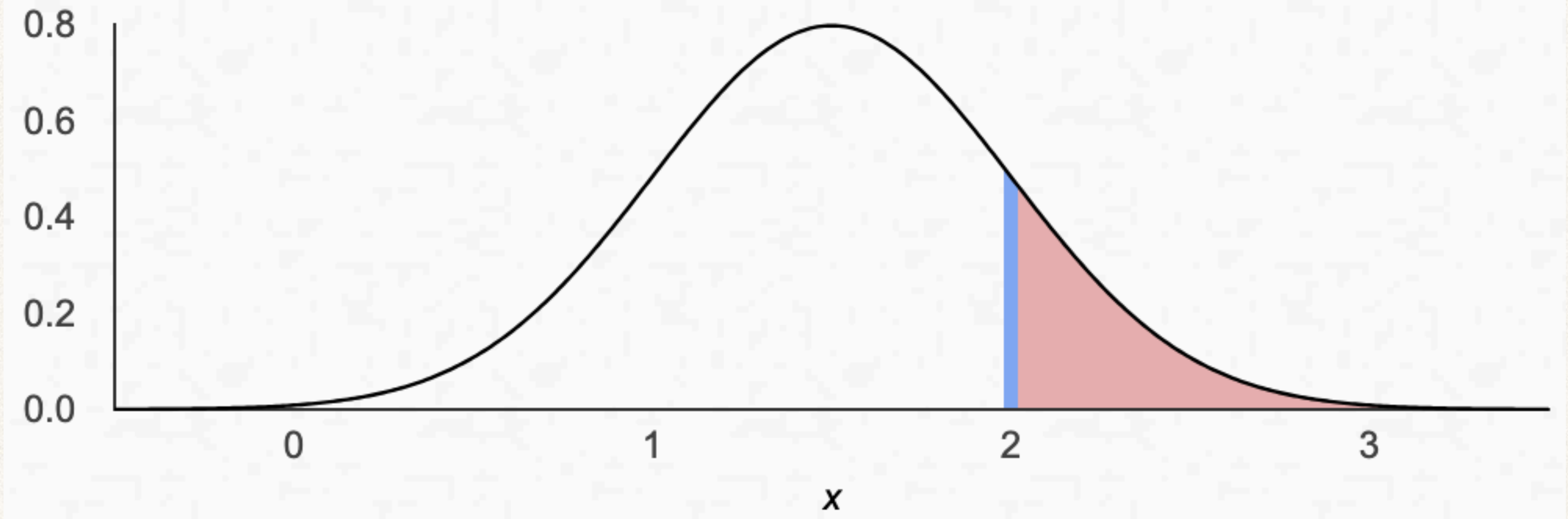
Hypothesis testing

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

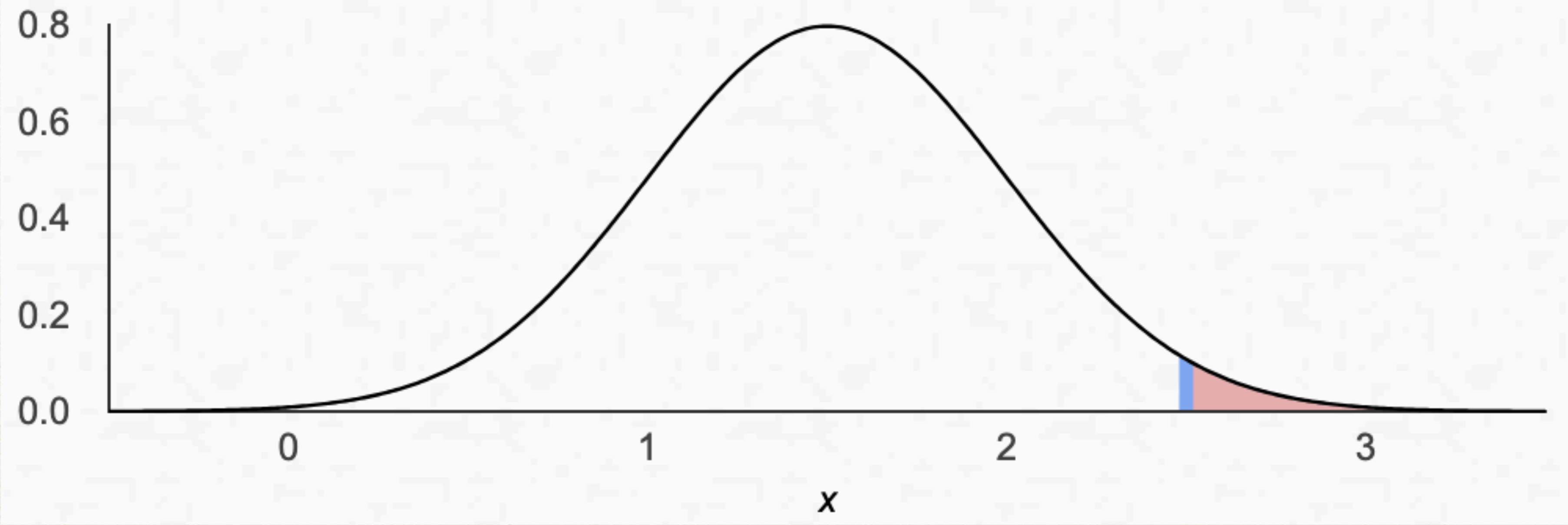
x=1.25



x=2



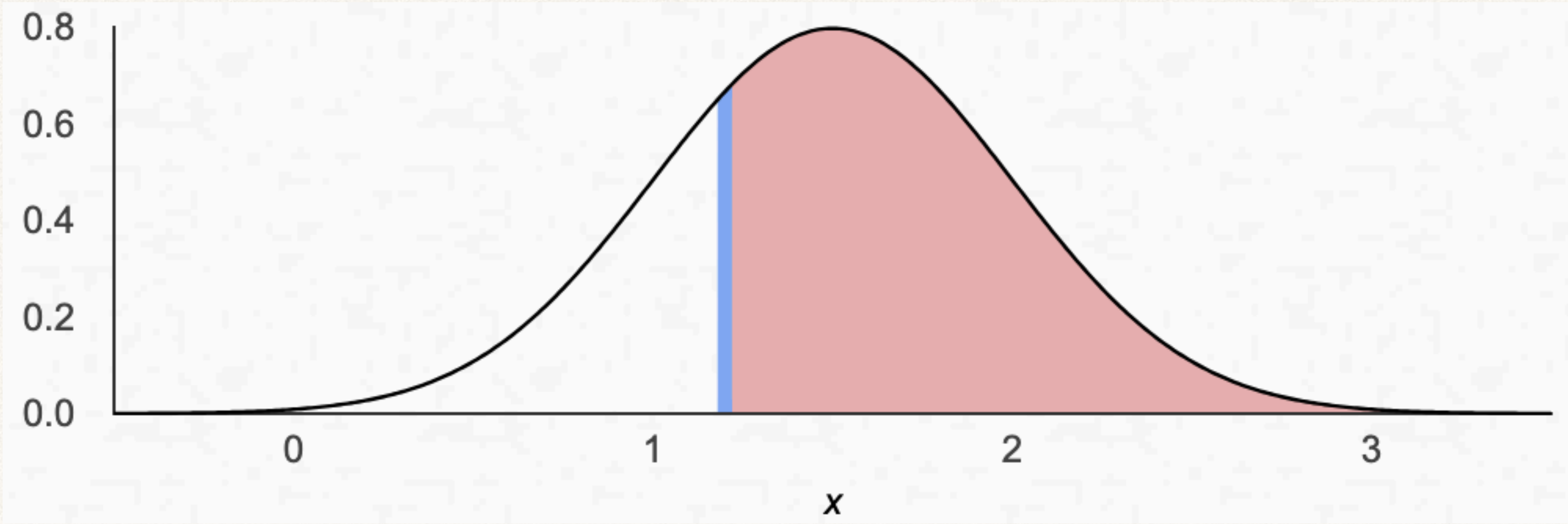
x=2.5



Hypothesis testing

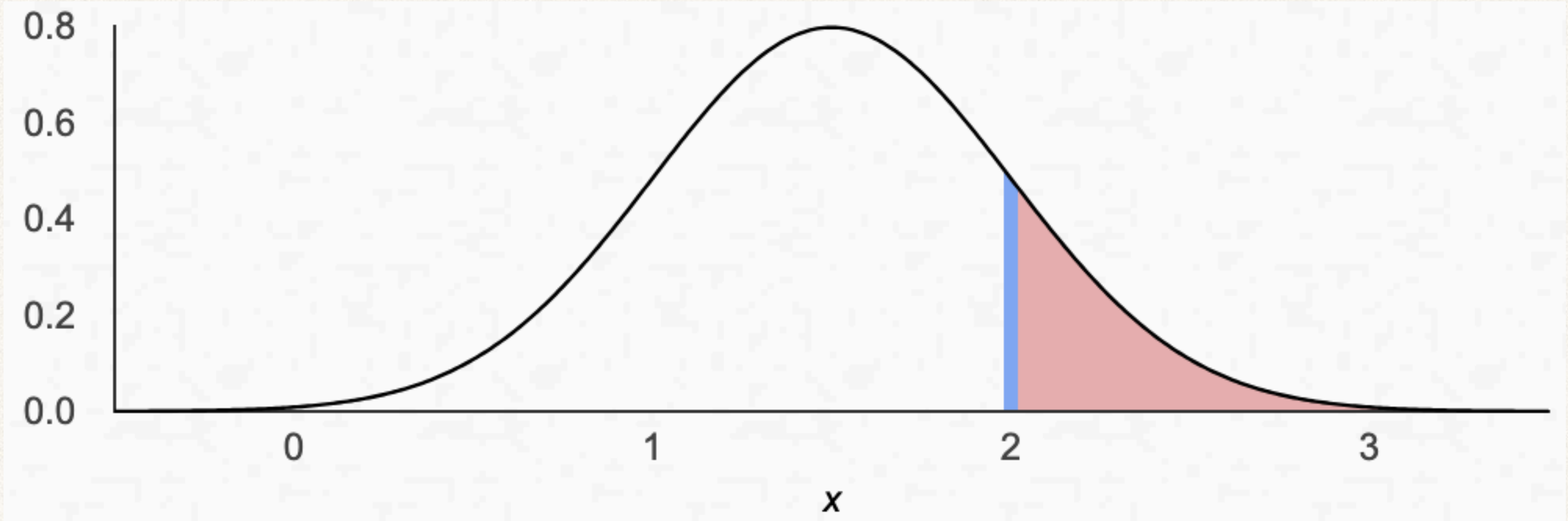
Which observation would you say is surprising if the null distribution holds?

x=1.25



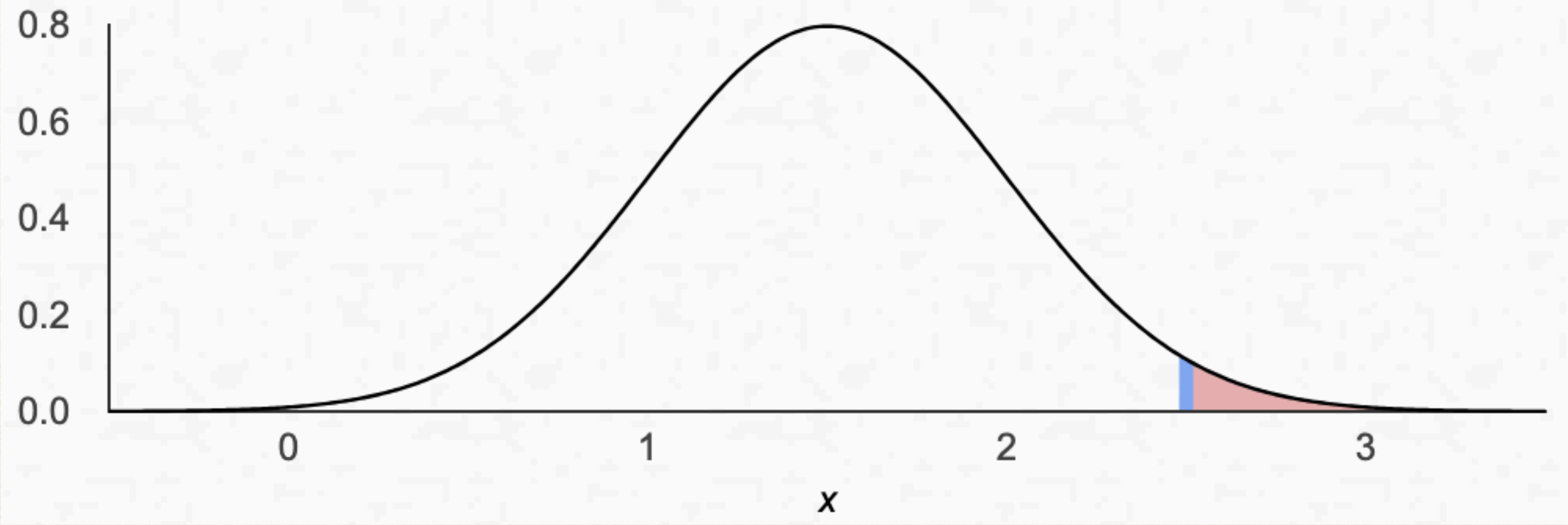
p=0.69

x=2



p=0.16

x=2.5



p=0.02

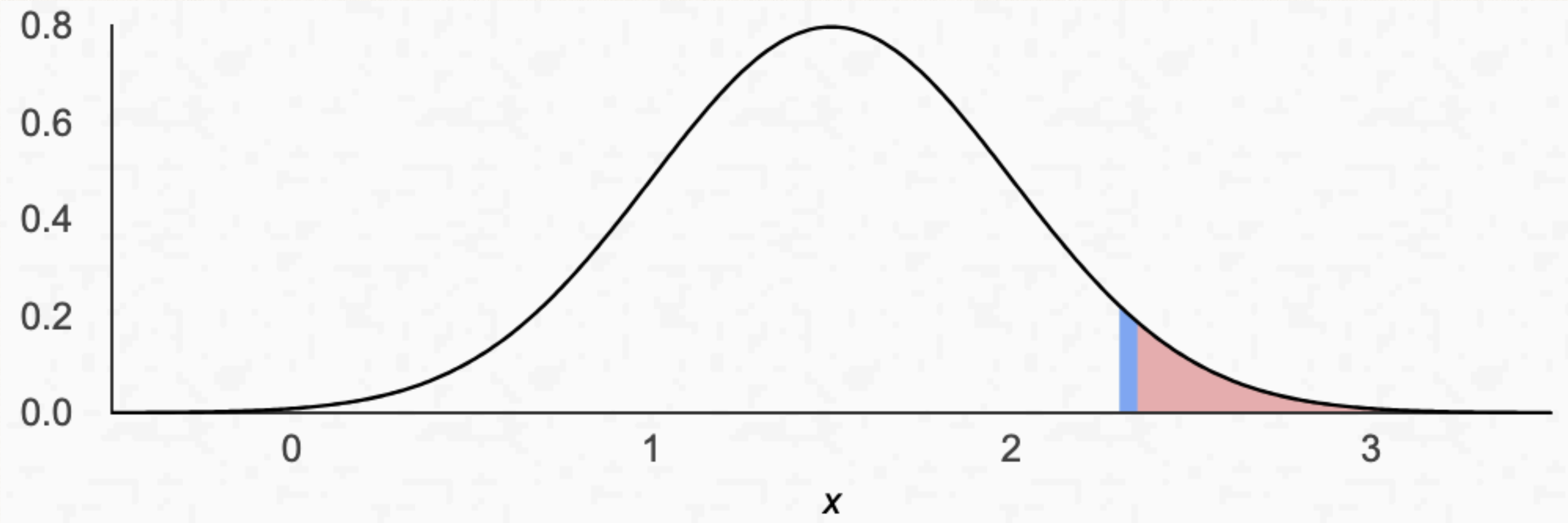
Significance level

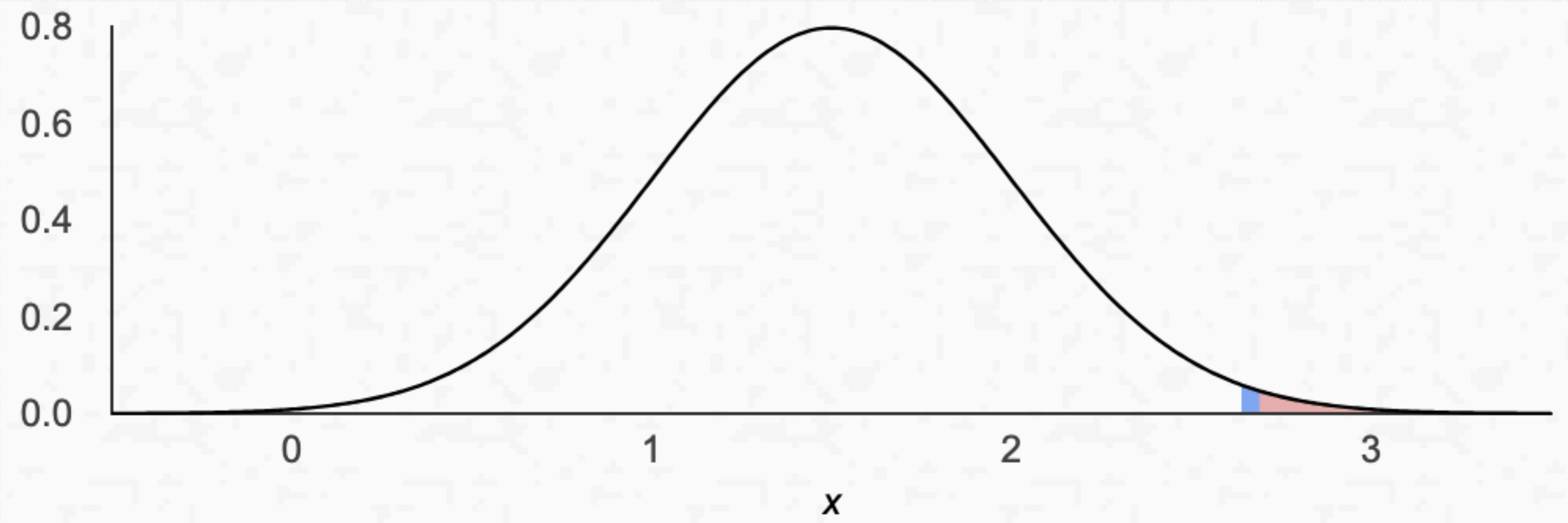
Significance level

To make a decision, set up a rejection region by choosing the value of α

Significance level

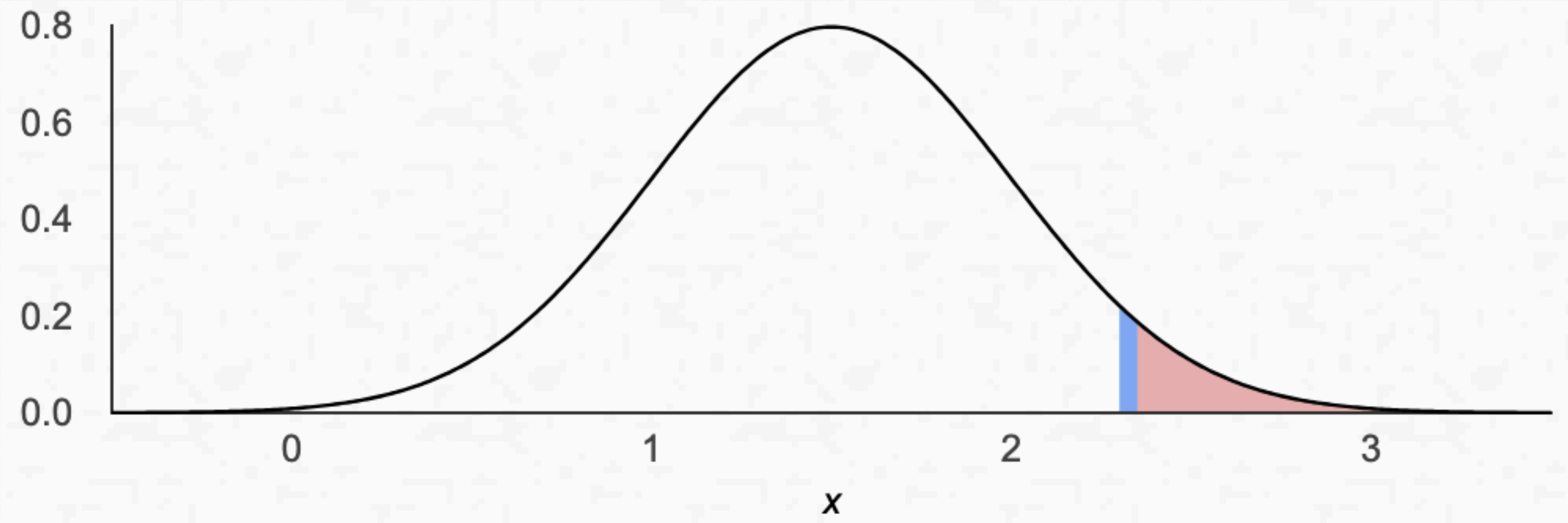
To make a decision, set up a rejection region by choosing the value of α



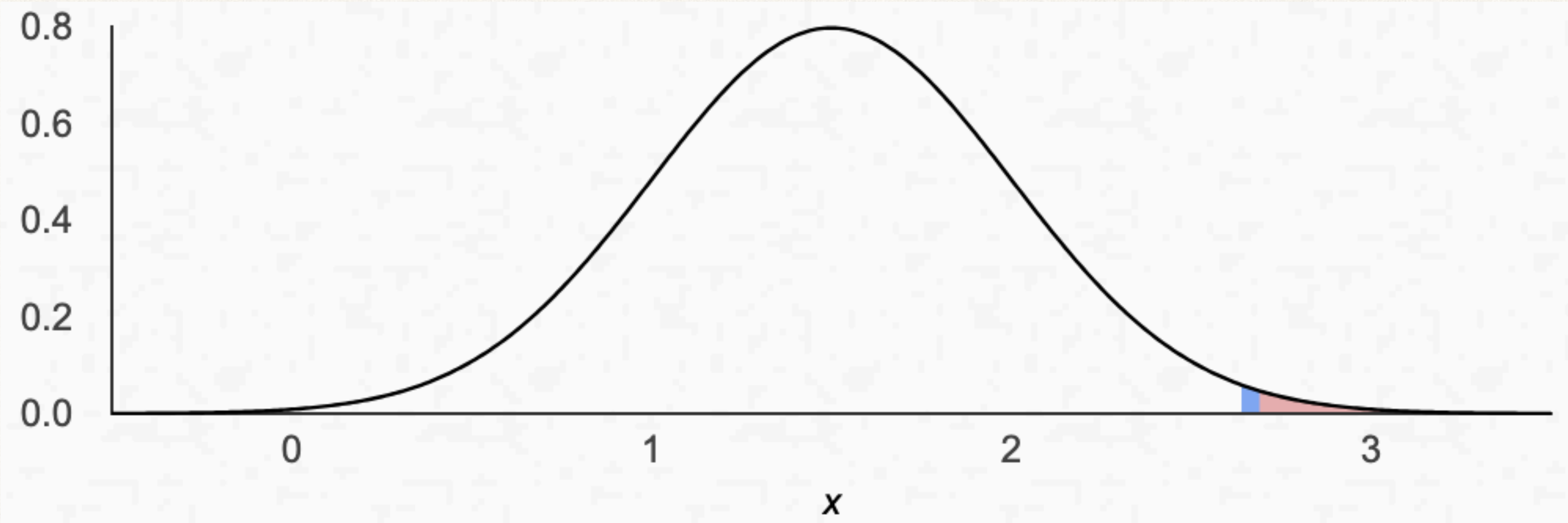


Significance level

To make a decision, set up a rejection region by choosing the value of α

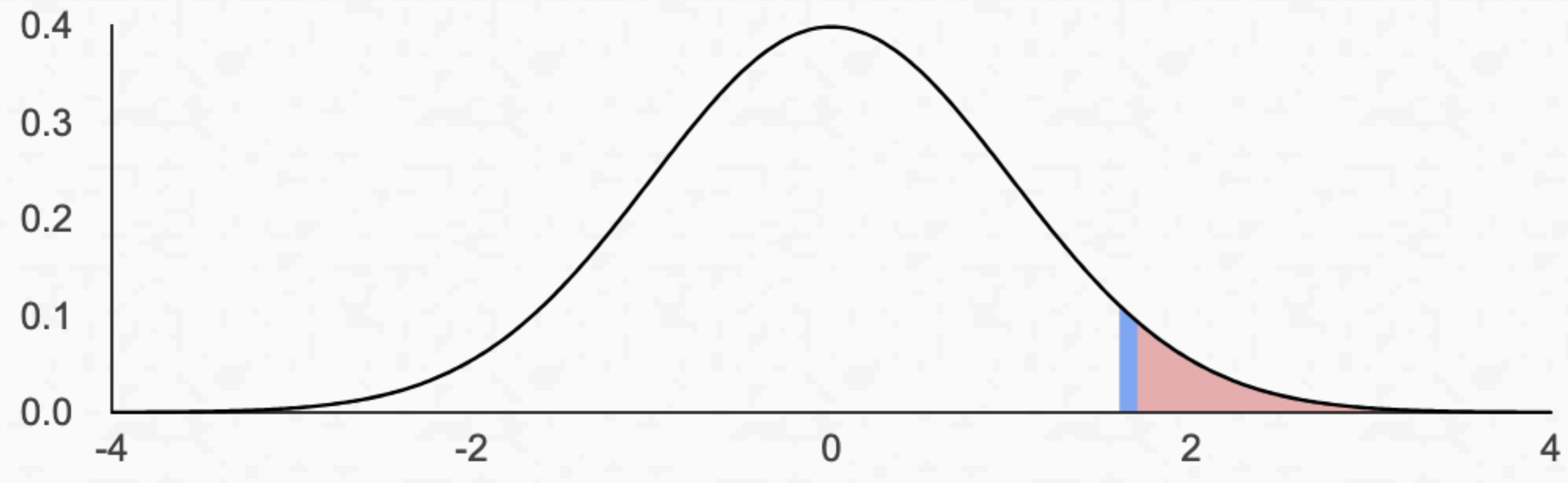


α=0.05



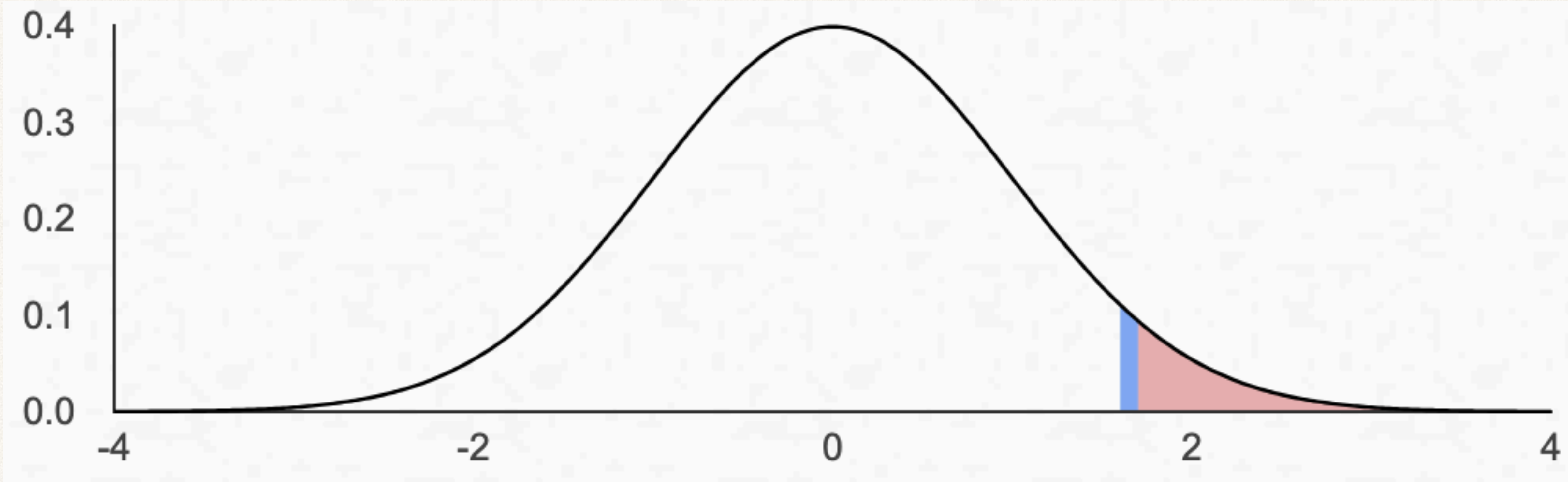
α=0.01

Z-score



Z-score

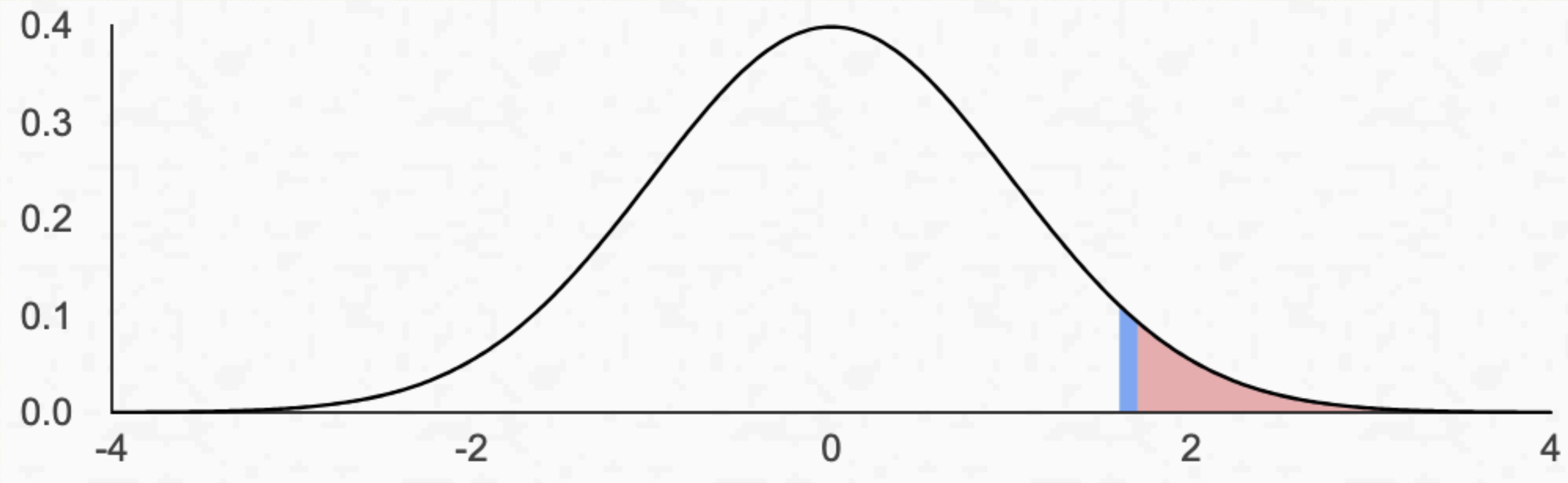
Calculating probabilities and comparing with the significance level can be tedious



Z-score

Calculating probabilities and comparing with the significance level can be tedious

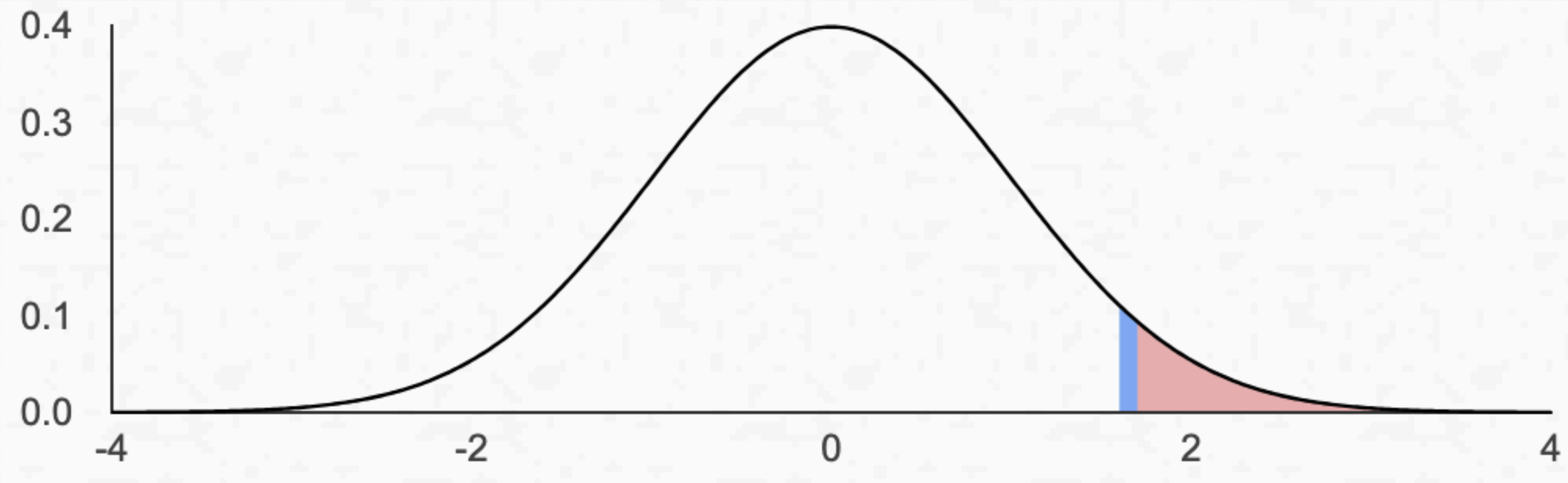
$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$



Z-score

Calculating probabilities and comparing with the significance level can be tedious

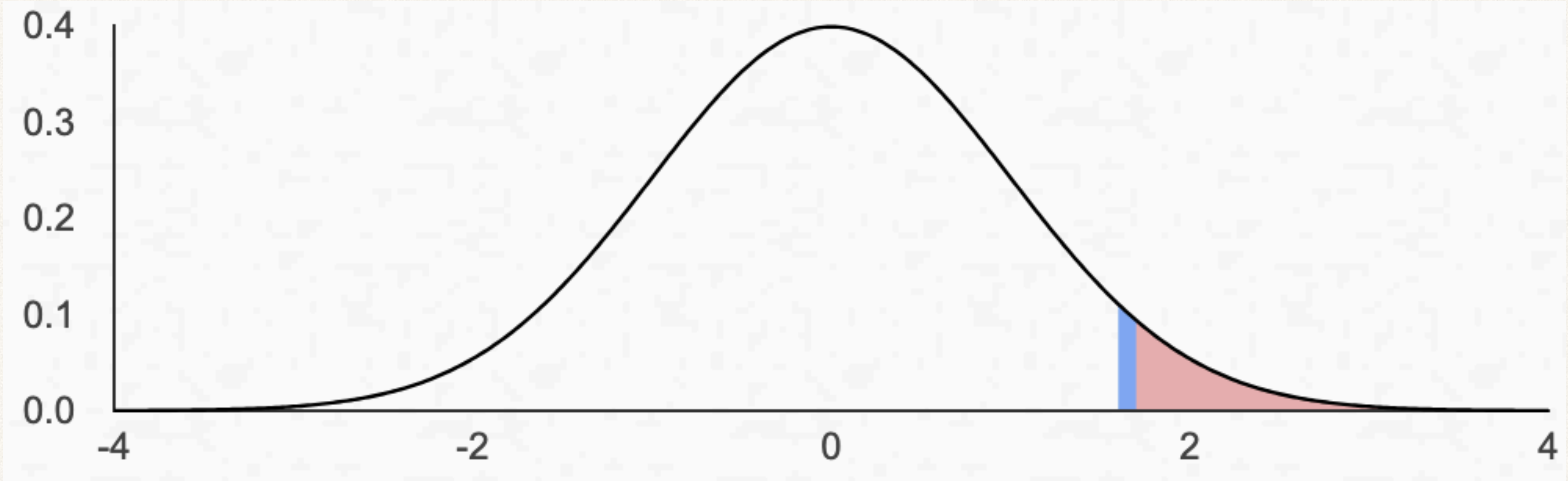
$$z = \frac{x - \mu}{\sigma / \sqrt{n}}$$



$\boxed{z = 1.65}$

Z-score

Calculating probabilities and comparing with the significance level can be tedious



$\boxed{z = 1.65}$

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

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

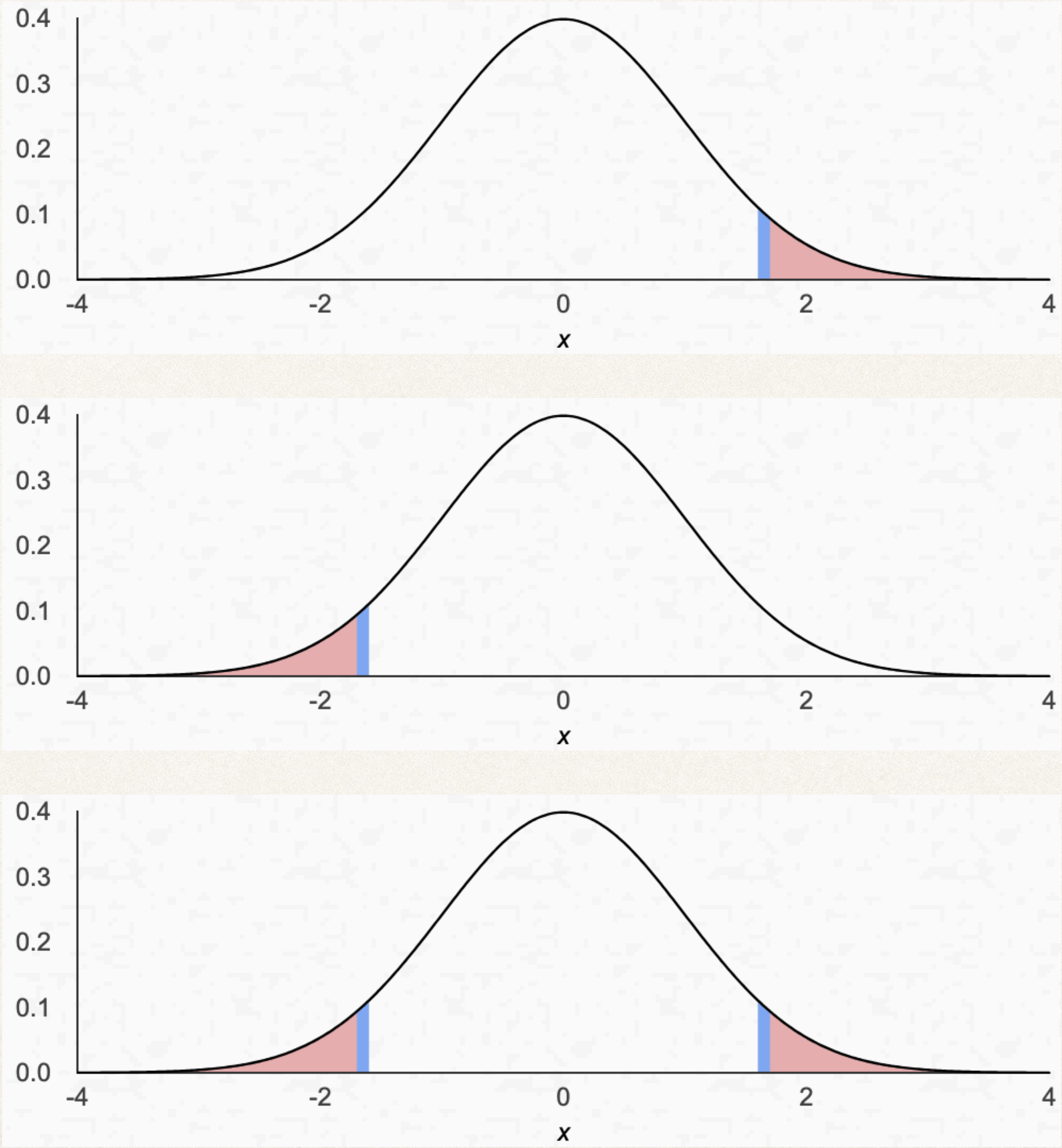
Tails

Tails

What is considered surprising is dependent on the problem at hand

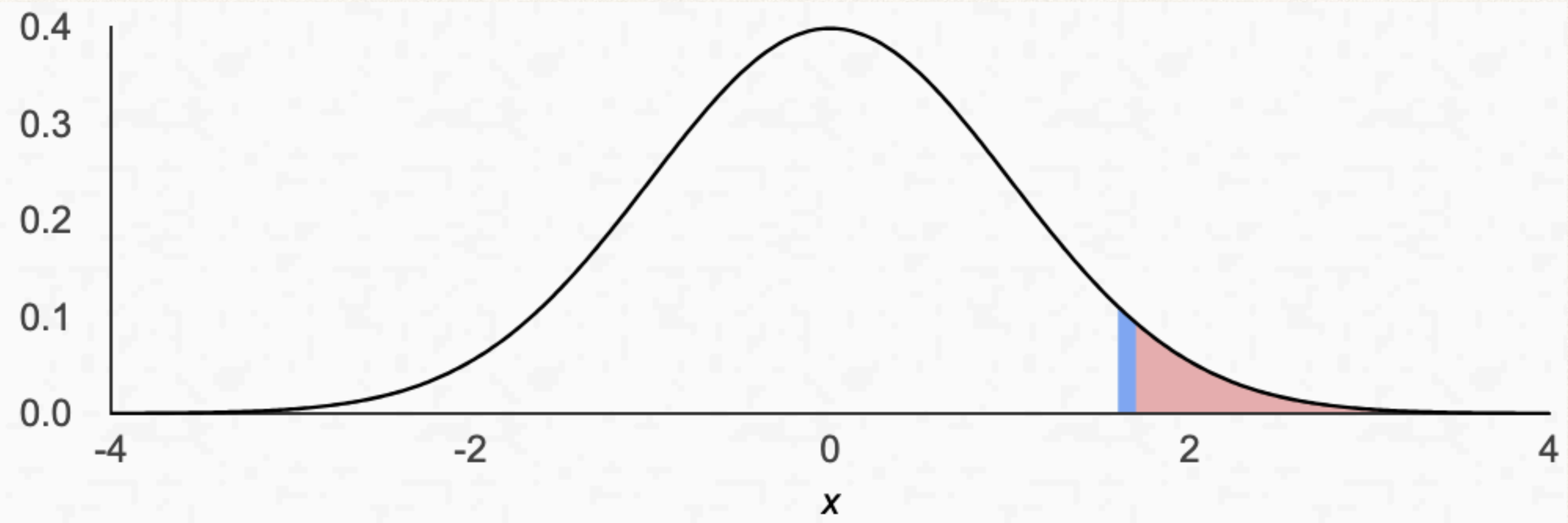
Tails

What is considered surprising is dependent on the problem at hand

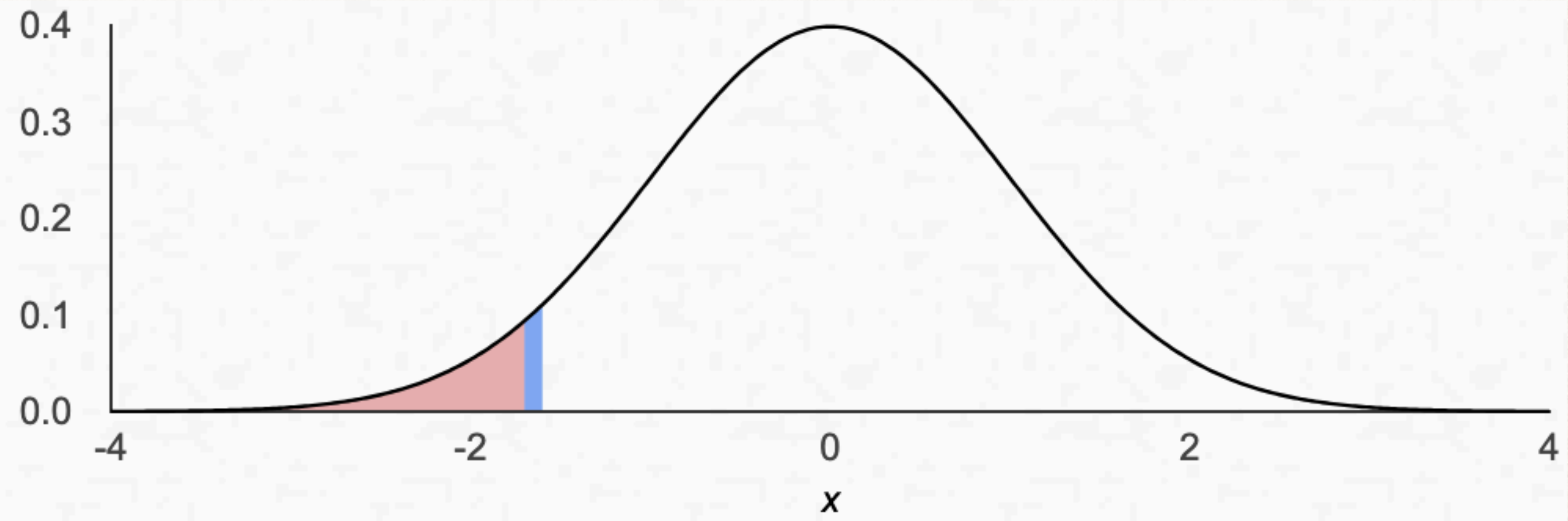


Tails

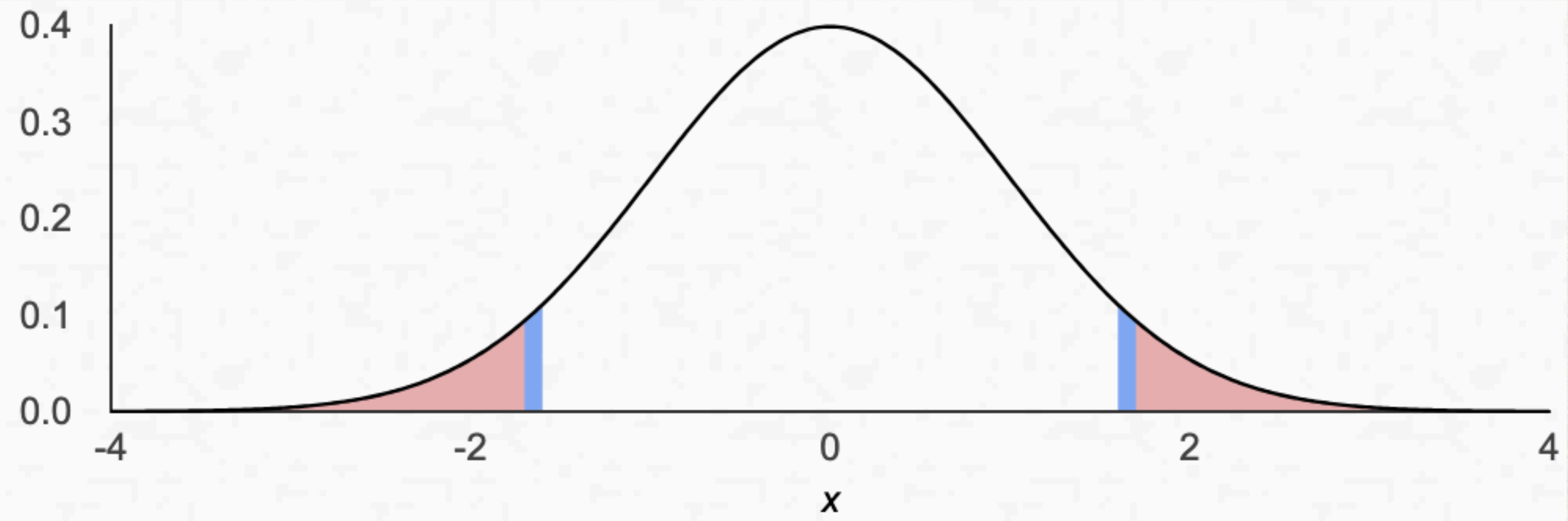
What is considered surprising is dependent on the problem at hand



upper-tail: is x significantly greater?



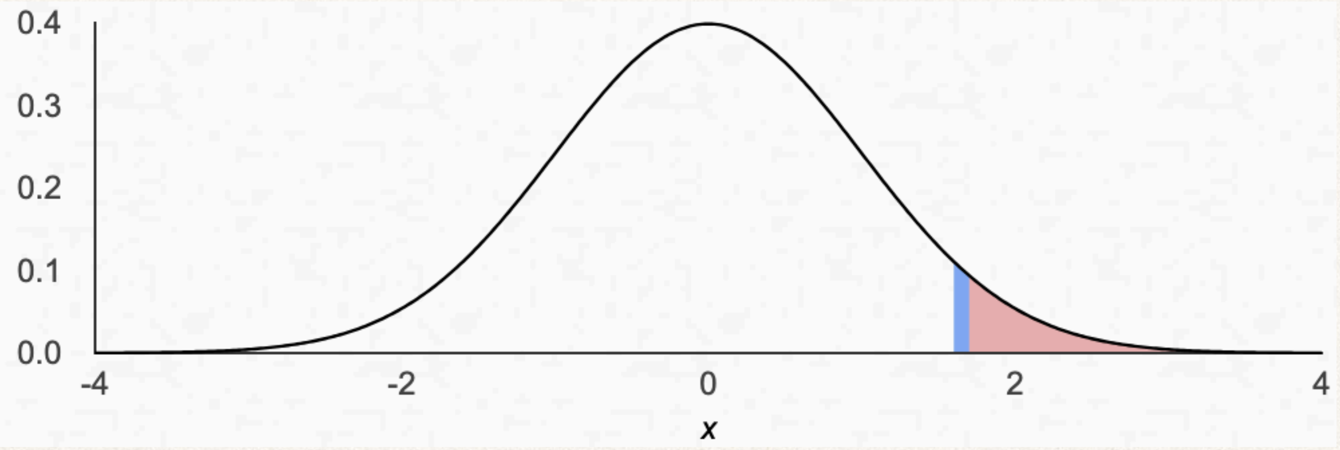
lower-tail: is x significantly smaller?



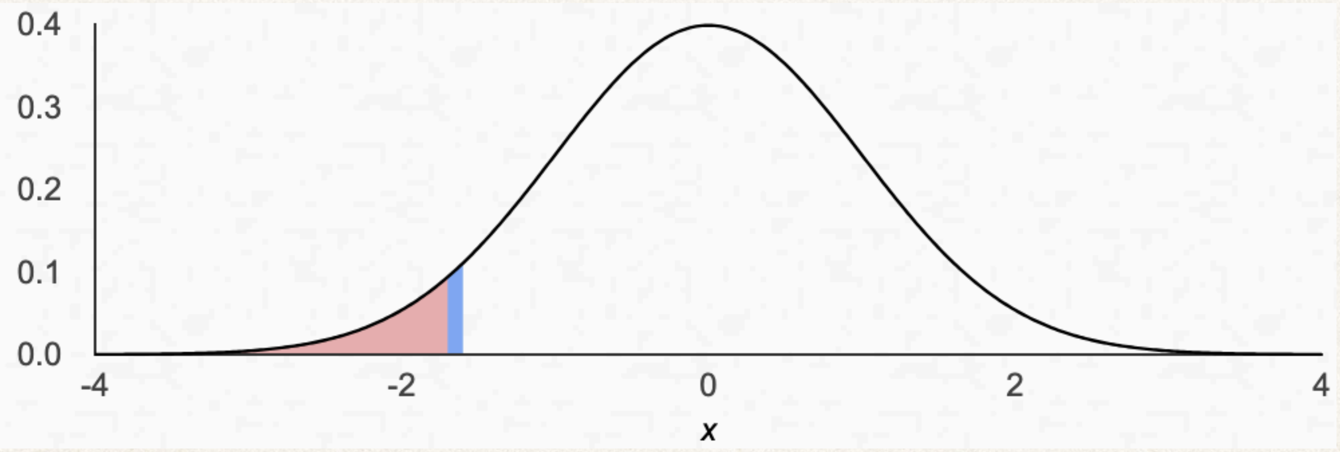
two-tail: is x significantly different?

P VALUES

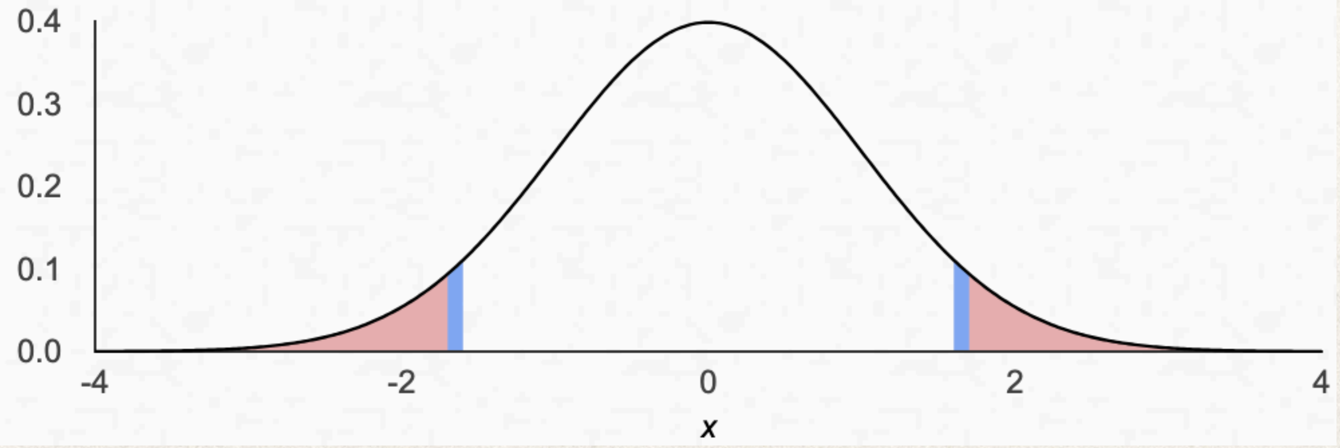
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.



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



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



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

Recipe for hypothesis testing

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$

Recipe for hypothesis testing

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

Recipe for hypothesis testing

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

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- 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 α
	reject null	Type II error β	Power

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

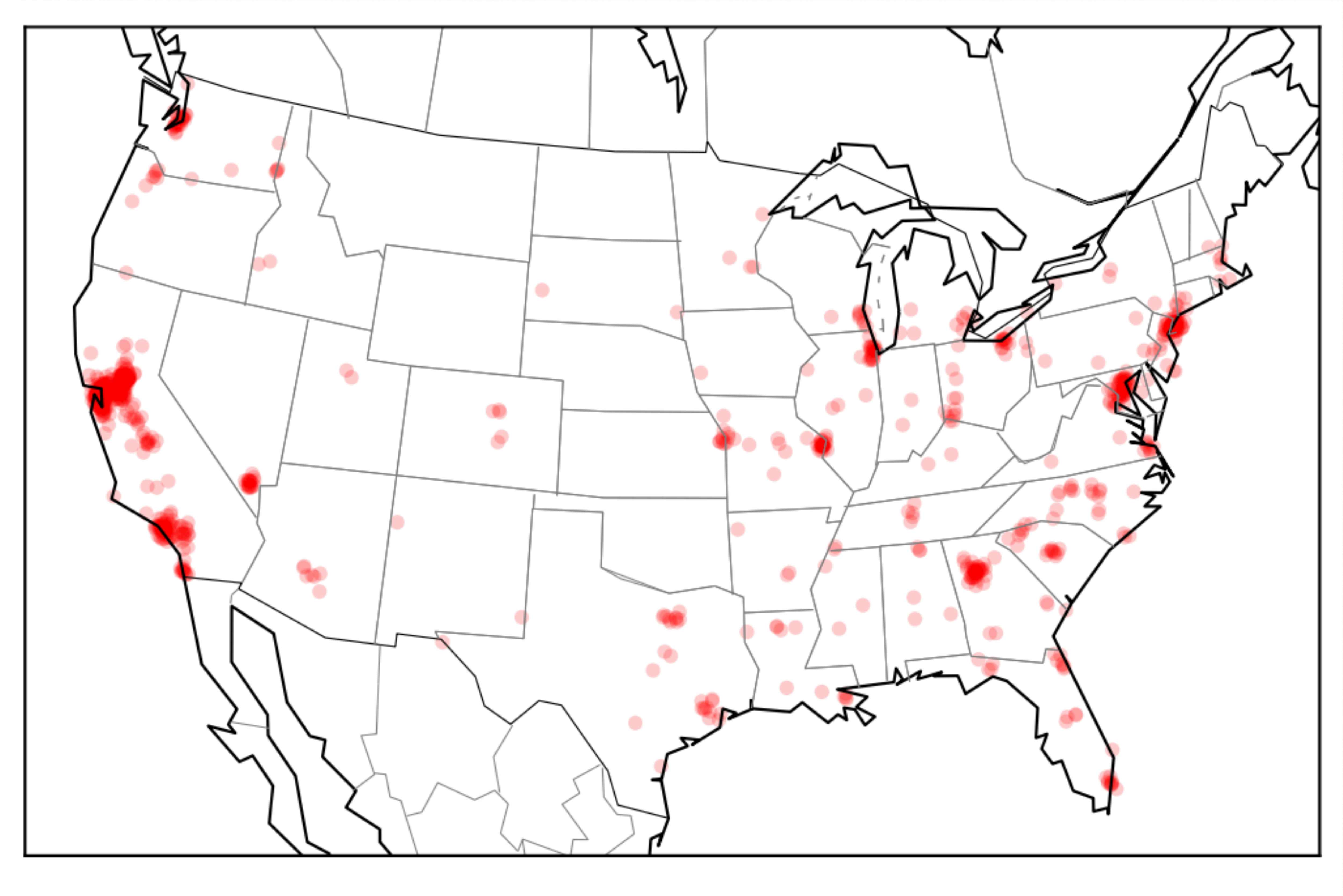
- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power

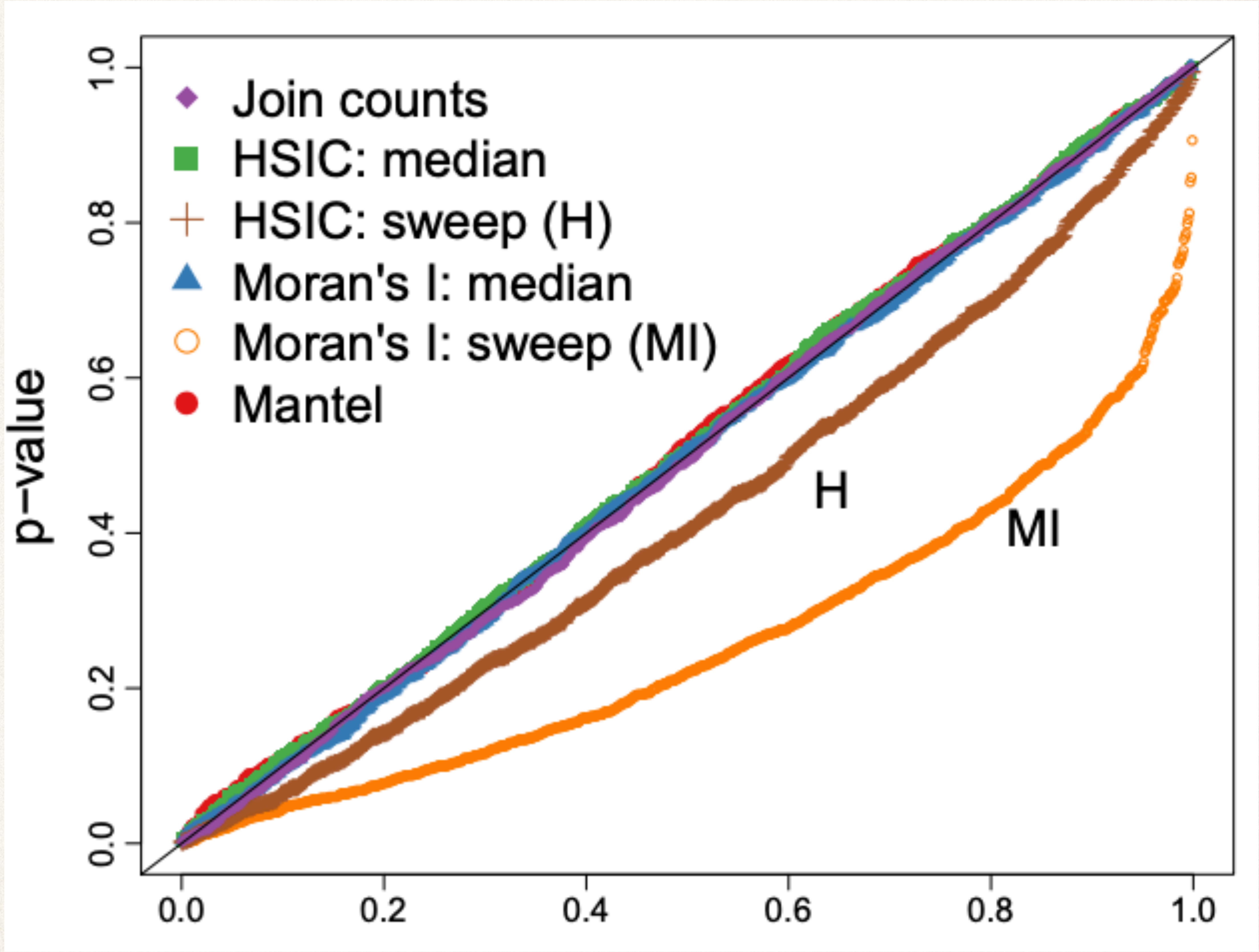
- If $\alpha = 0.05$ and we do 100 tests, we expect to make 5 mistakes

		Test results	
		keep null	reject null
Truth	keep null		Type I error α
	reject null	Type II error β	Power

Dong Nguyen and Jacob Eisenstein. 2017. A Kernel Independence Test for Geographical Language Variation. Computational Linguistics, 43(3):567–592.

Test if lexical variation is independent of geography





Dong Nguyen and Jacob Eisenstein. 2017. A Kernel Independence Test for Geographical Language Variation. Computational Linguistics, 43(3):567–592.

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 α_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

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

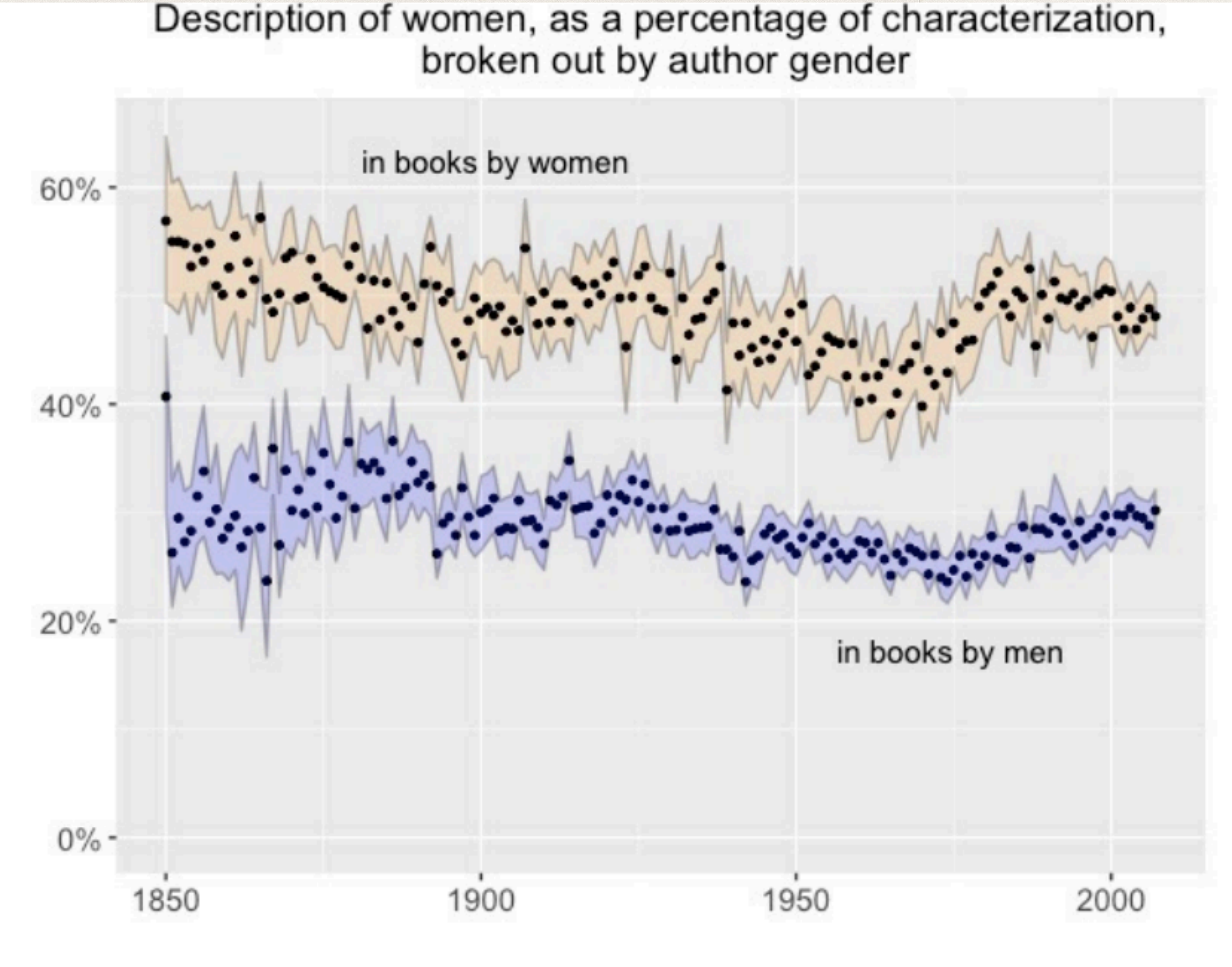
Standard deviation

Mean value

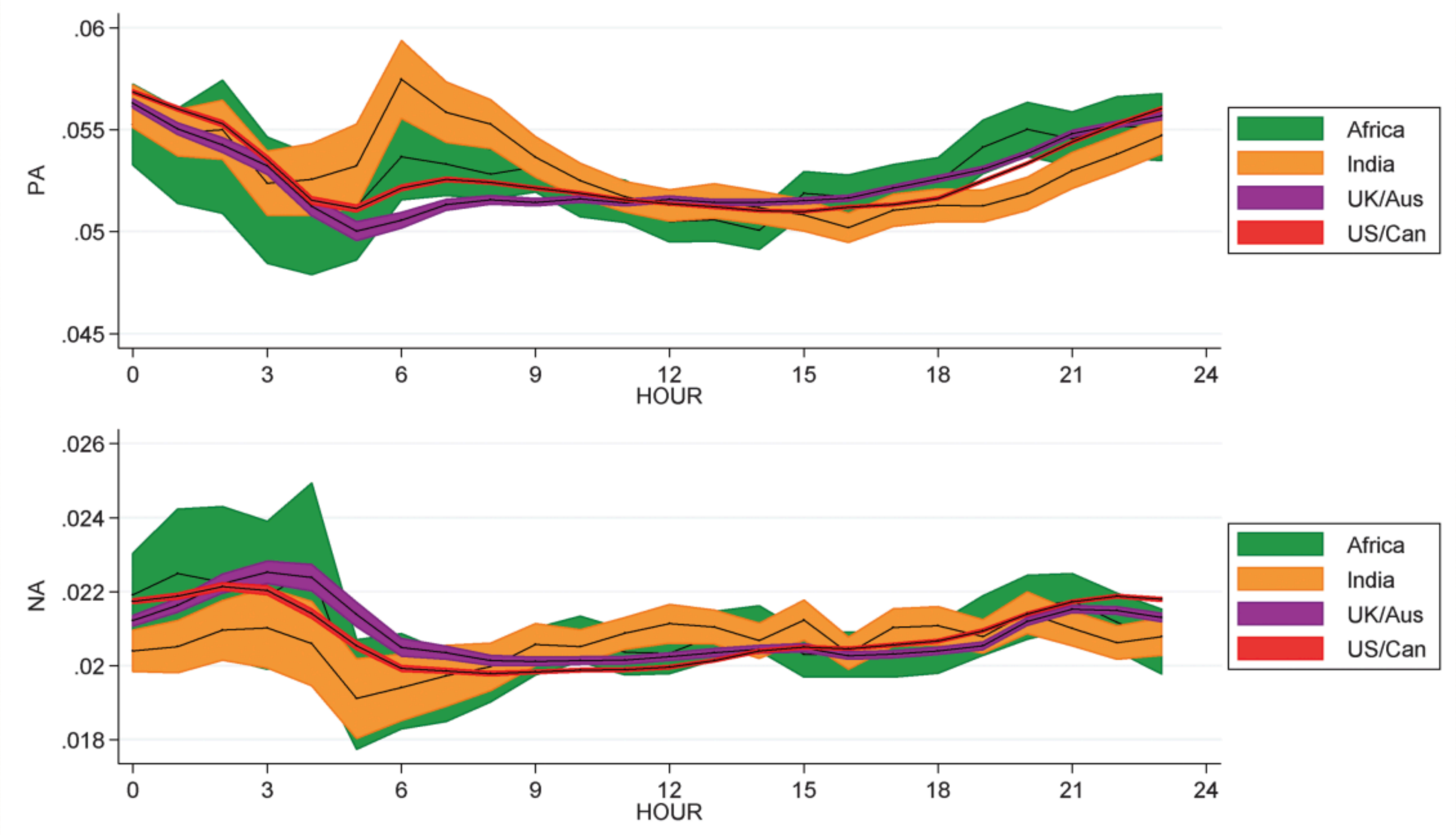
Sample size

Lower/Upper limit

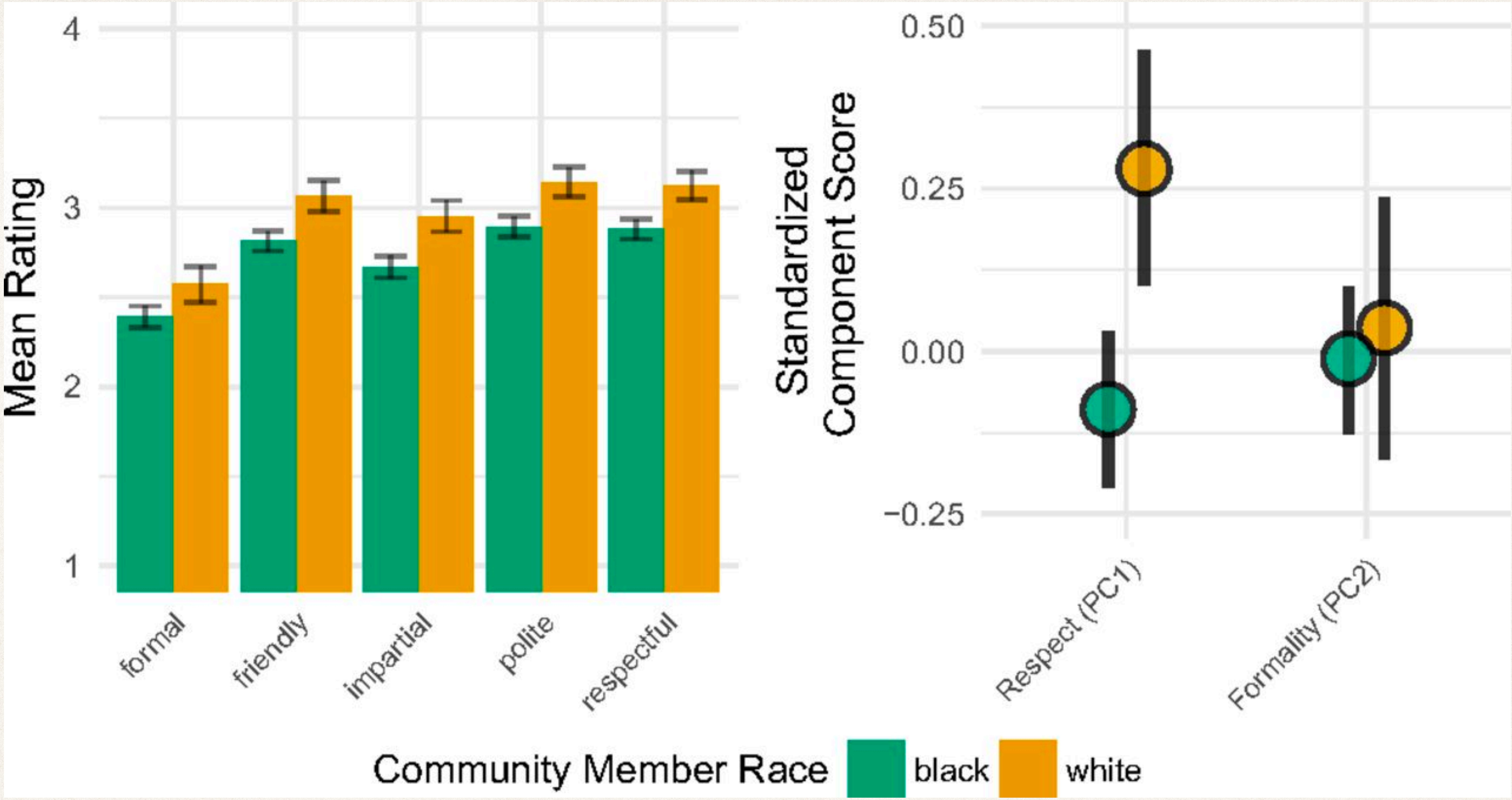
z-value for the confidence level



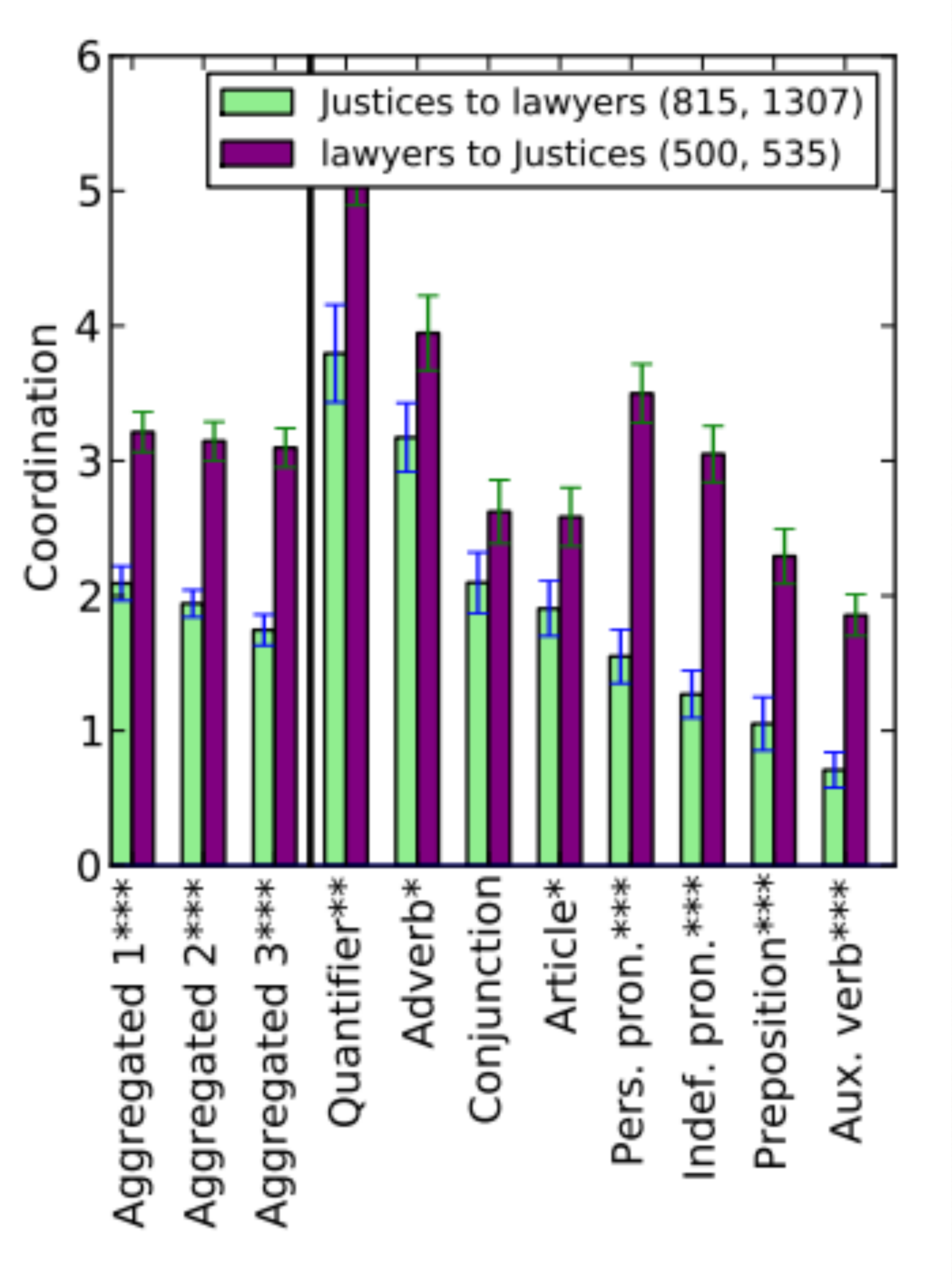
Underwood, Ted, David Bamman, and Sabrina Lee. "The transformation of gender in English-language fiction." Journal of Cultural Analytics 3.2 (2018).

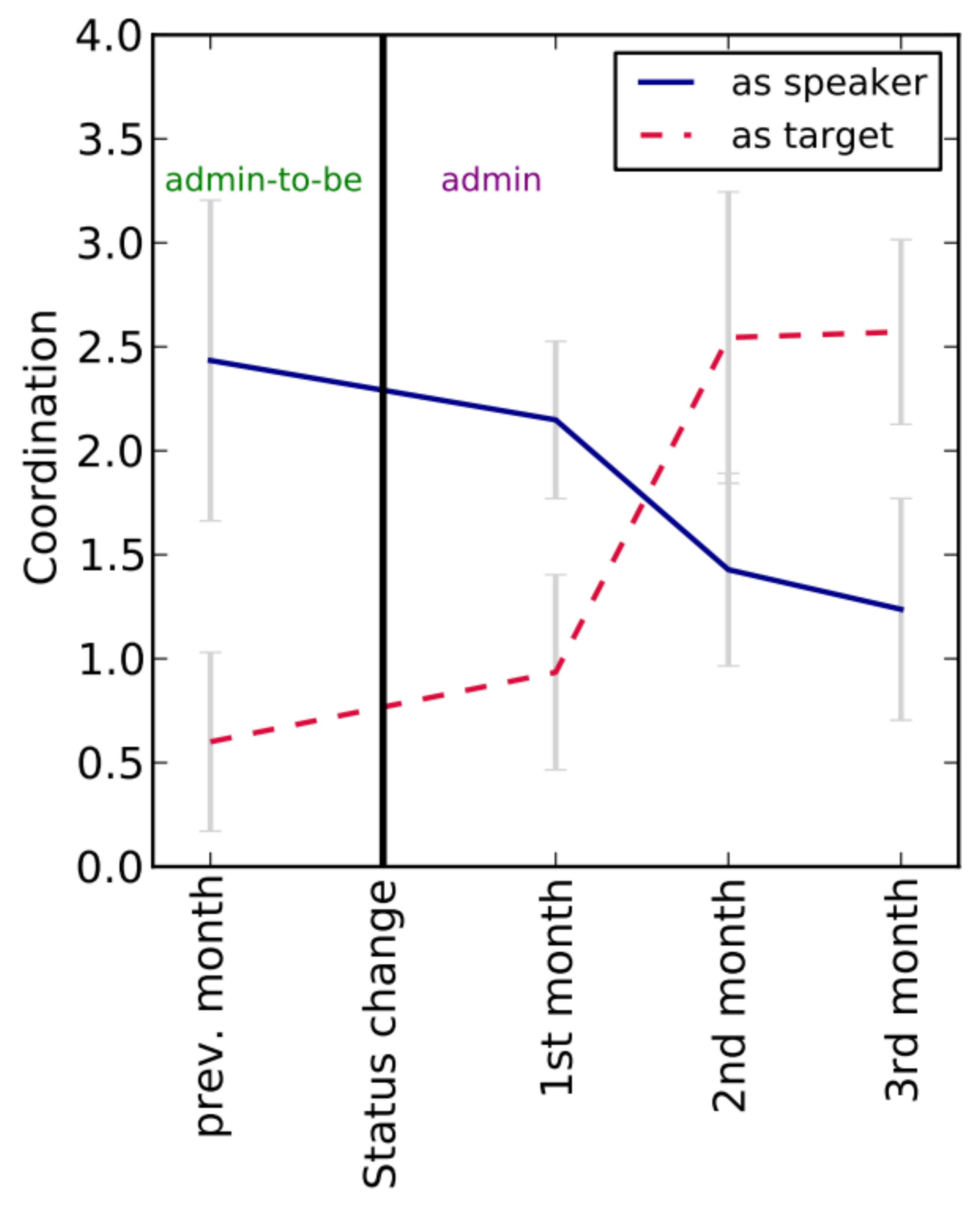


Golder, Scott A., and Michael W. Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." Science 333.6051 (2011): 1878-1881.



Voigt, Rob, et al. "Language from police body camera footage shows racial disparities in officer respect." Proceedings of the National Academy of Sciences 114.25 (2017): 6521-6526.





Danescu-Niculescu-Mizil, Cristian, et al. "Echoes of power: Language effects and power differences in social interaction." Proceedings of the 21st international conference on World Wide Web. 2012.

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.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

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