More Hypothesis Testing, Linear Regression

March 5, 2019
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

Announcements

- Projects: check in with your mentor TA!
- Extensions for sickness, etc.
- Google Cloud Credits!
- Its that time of year…when we can officially say we are behind schedule :-D
- Anything from you guys?

Today

- Hypothesis Testing and p-values continued

- Linear Regression (Part 1)

Hypothesis Testing in General

Hypothesis Testing in General

- Null hypothesis (H_0) — the "nothing to see here" assumption

Hypothesis Testing in General

- Null hypothesis (H_0) — the "nothing to see here" assumption
- Alternative hypothesis (H_a)—the thing you know will lead to an explosive headline and are really hoping is true but you are a good scientist, so you will look to the data to confirm

Hypothesis Testing in General

Thing you can model

- Null hypothesis (H_0) — the "nothing to see here" assumption
- Alternative hypothesis (H_a)—the thing you know will lead to an explosive headline and are really hoping is true but you are a good scientist, so you will look to the data to confirm

Hypothesis Testing in General

- Assume the null hypothesis is true—i.e., don't deviate from status quo without good reason :)

Hypothesis Testing in General

- Assume the null hypothesis is true—i.e., don't deviate from status quo without good reason :)
- If there is enough evidence to suggest that H_0 is highly unlikely, then we can say we "reject the null hypothesis"

Hypothesis Testing in General

- Assume the null hypothesis is true—i.e., don't deviate from status quo without good reason :)

- If there is enough evidence to suggest that H_0 is highly unlikely, then we can say we "reject the null hypothesis"

- If there is not enough evidence, we "fail to reject it"

Hypothesis Testing in General

- Assume the null hypothesis is true—i.e., don't deviate from status quo without good reason :)

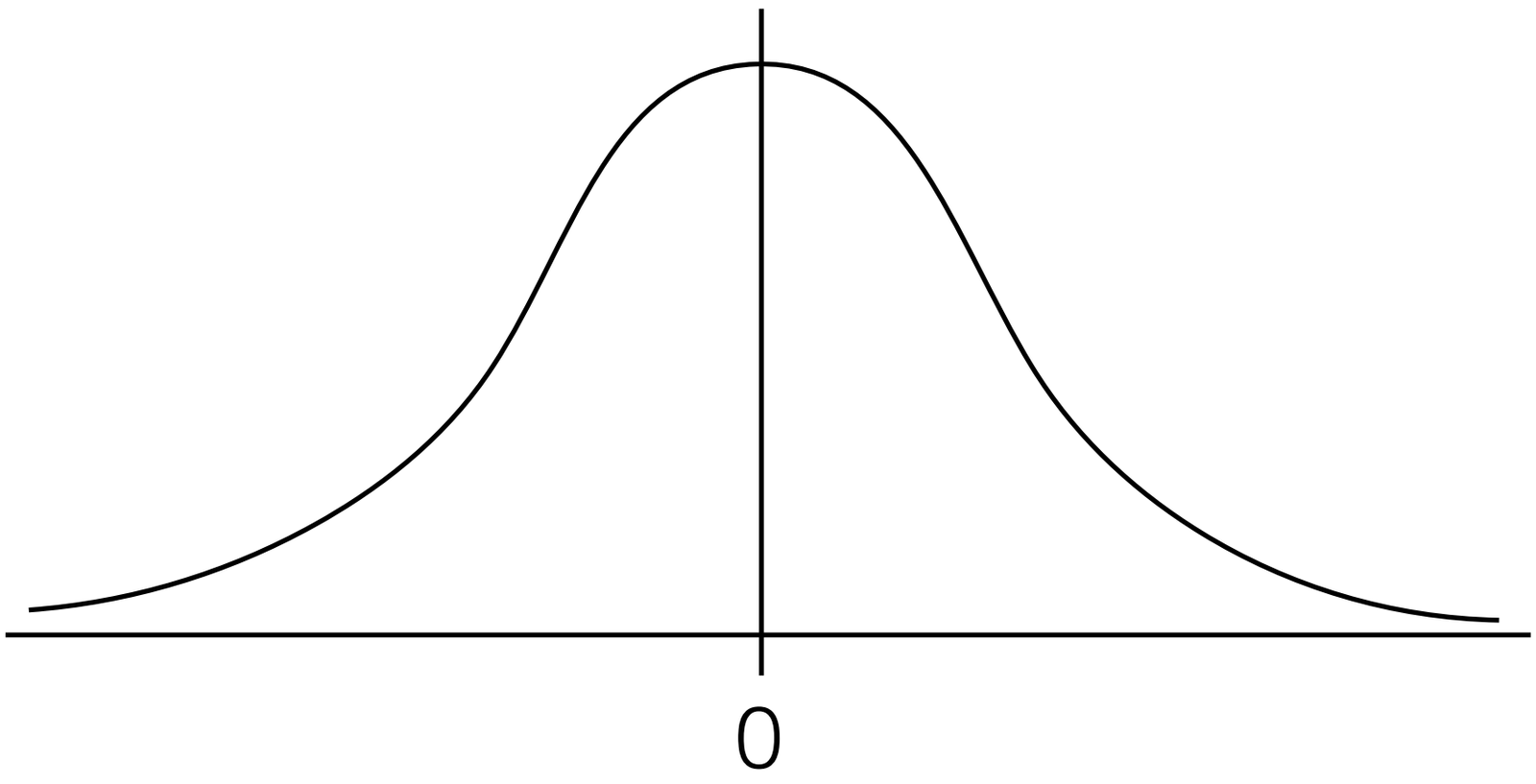
- If there is enough evidence to suggest that H_0 is highly unlikely, then we can say we "reject the null hypothesis"

- If there is not enough evidence, we "fail to reject it"

- We don't "accept" or "prove" H_0 or H_a

Standard Normal Distr.

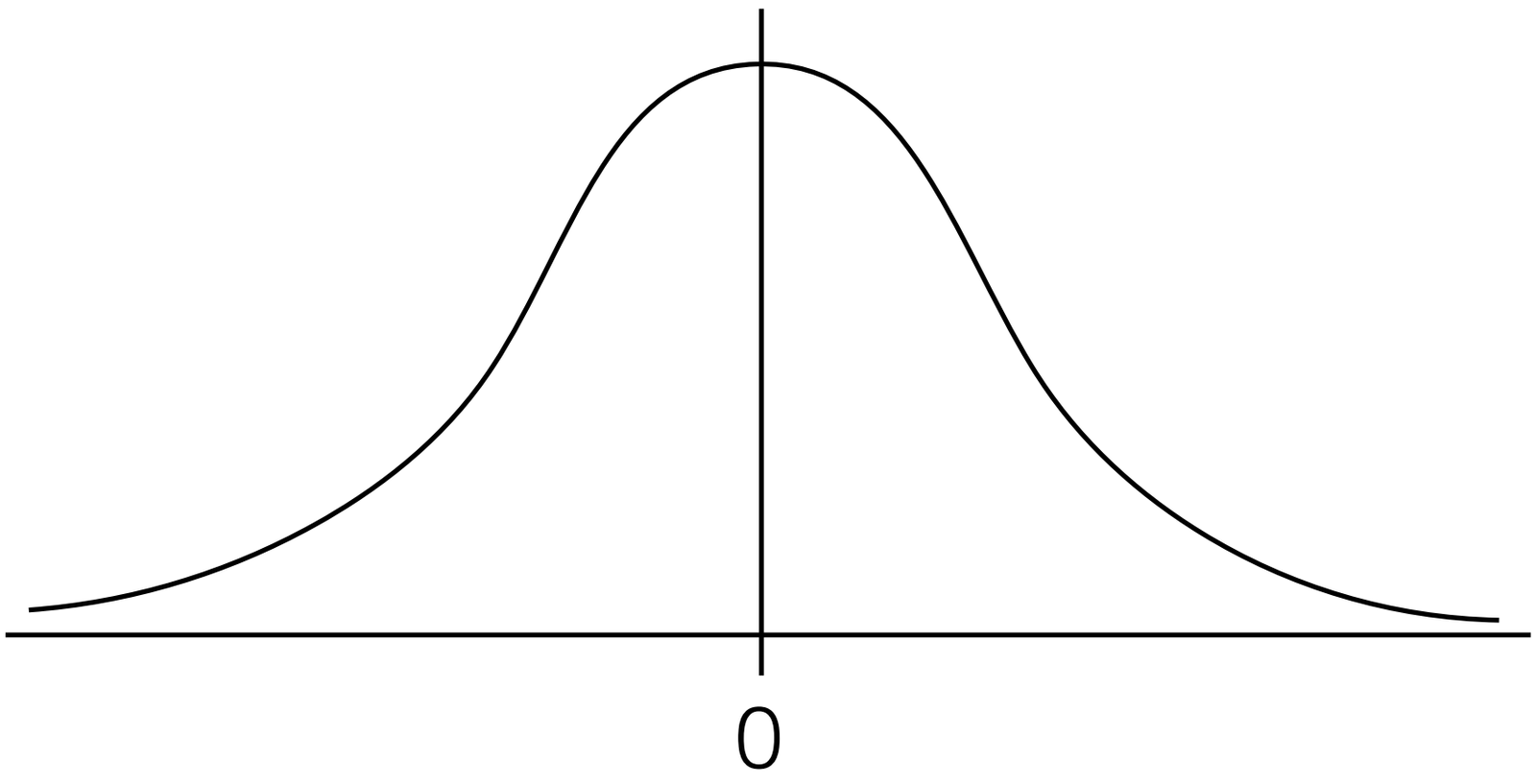
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



z = distance from mean in std units

Standard Normal Distr.

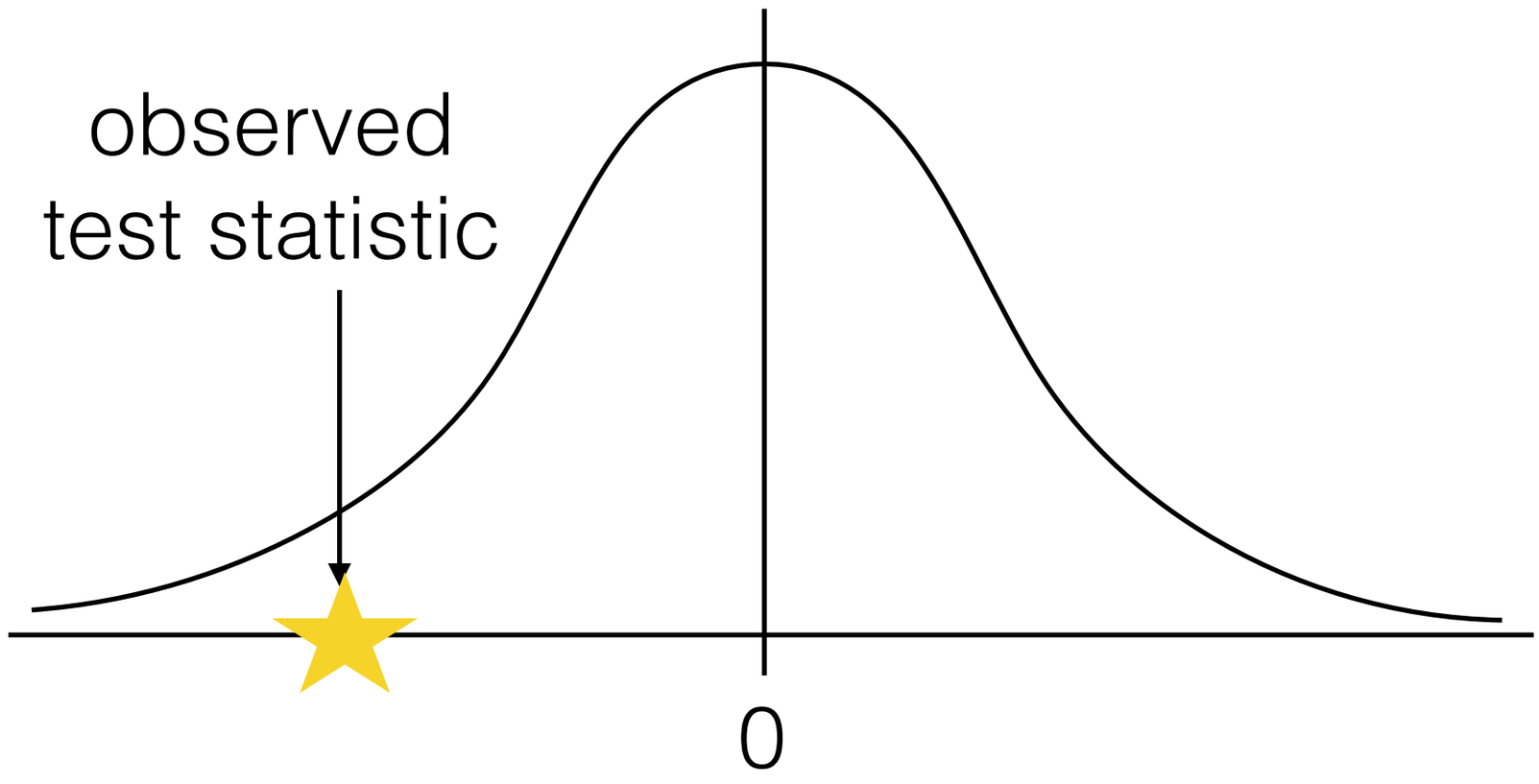
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



z = (observed - expected) / standard deviation

Standard Normal Distr.

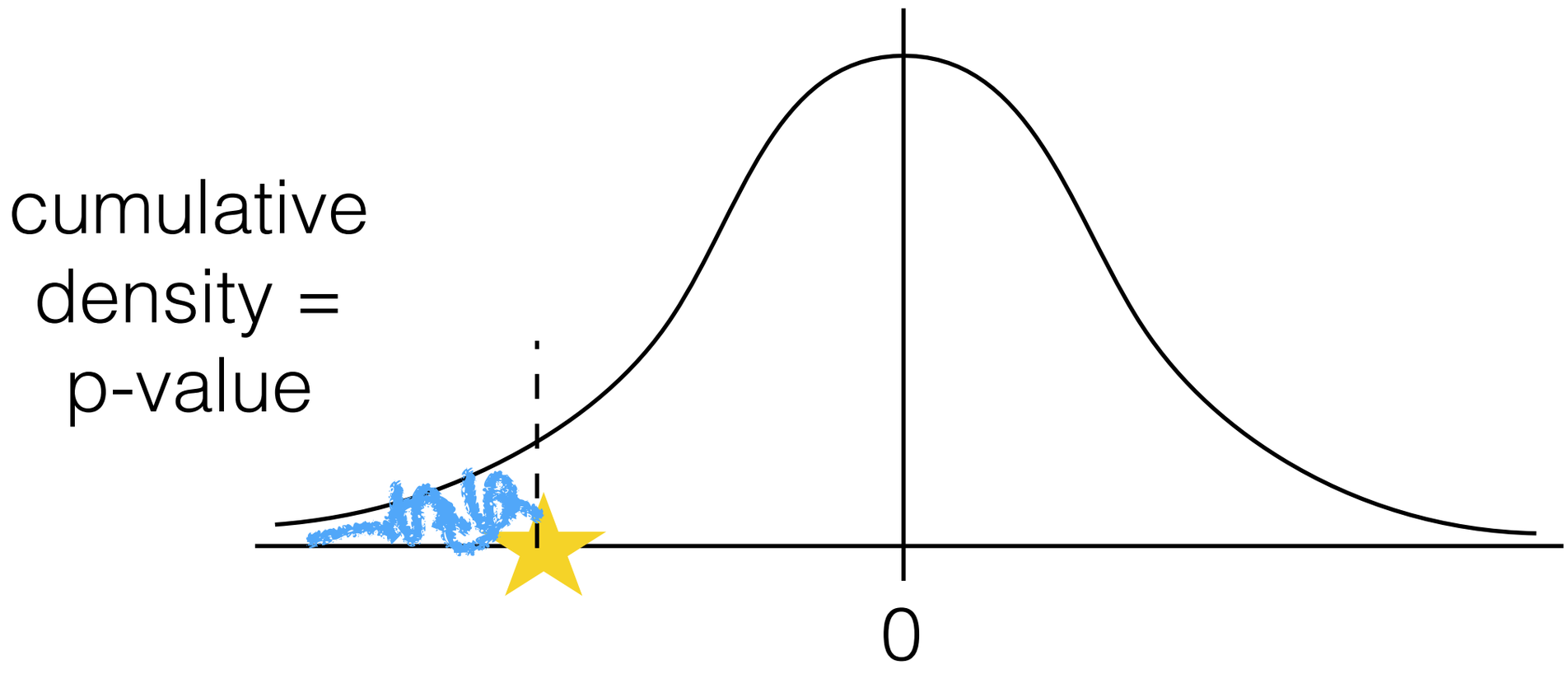
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



z = (observed - expected) / standard deviation

Standard Normal Distr.

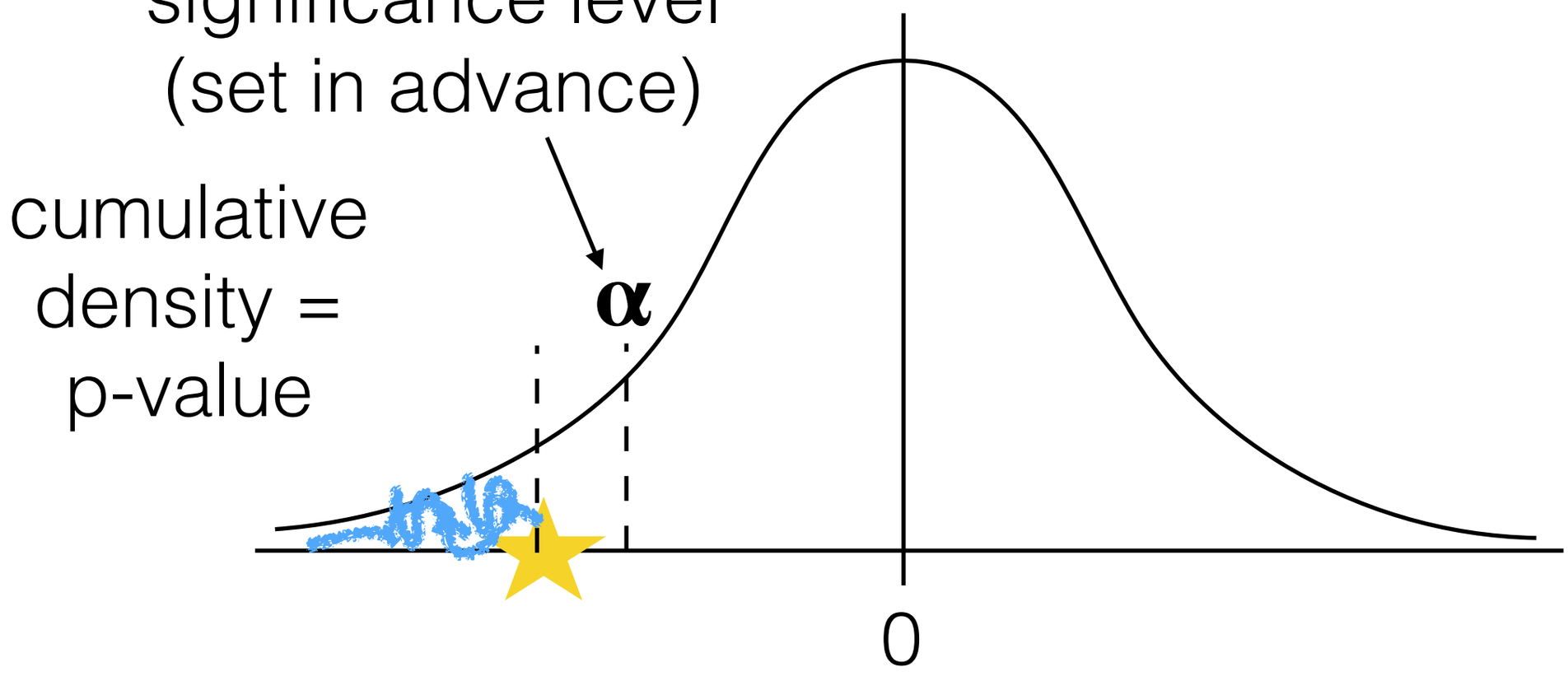
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



z = (observed - expected) / standard deviation

Standard Normal Distr.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



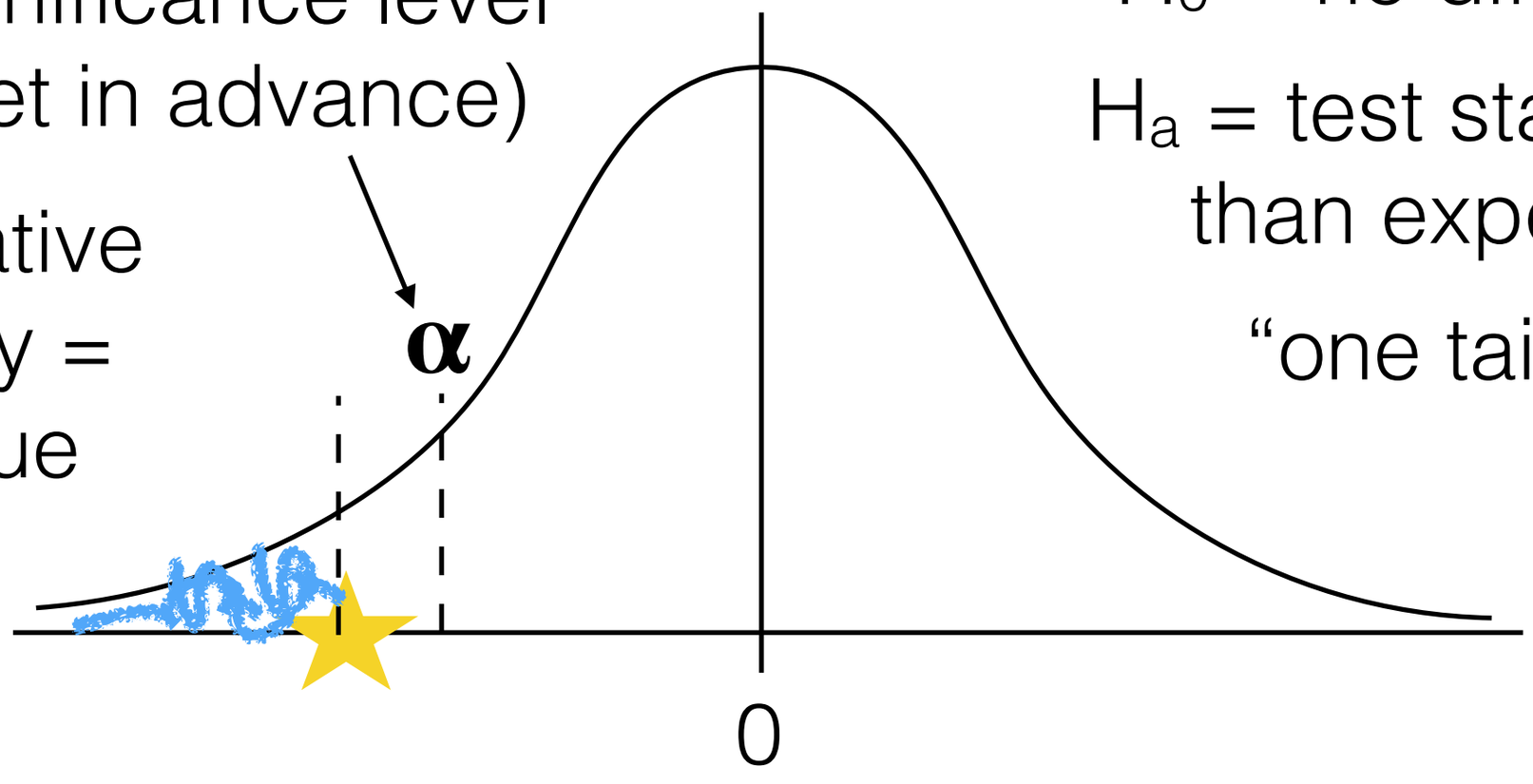
z = (observed - expected) / standard deviation

Standard Normal Distr.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

significance level
(set in advance)

cumulative
density =
p-value



H_0 = no difference

H_a = test stat is less than expected

"one tailed"

z = (observed - expected) / standard deviation

Standard Normal Distr.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

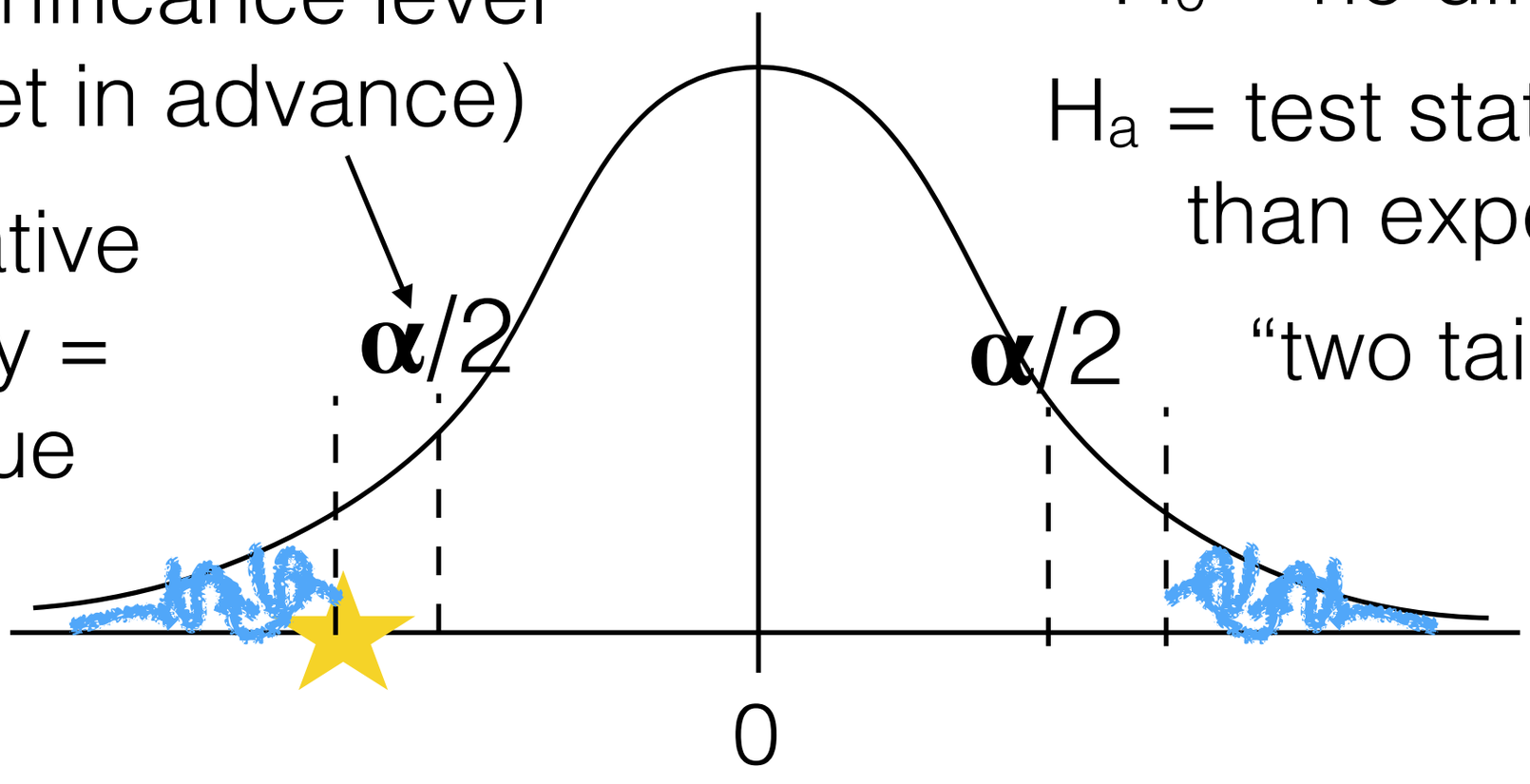
significance level
(set in advance)

cumulative
density =
p-value

H_0 = no difference

H_a = test stat different
than expected

“two tailed”



z = (observed - expected) / standard deviation

Interpreting P-Values

Credit: CS1951A Spring 2017

Interpreting P-Values

☑ Probability of obtaining an effect equal to or more extreme than the one observed, presuming the null hypothesis is true

Credit: CS1951A Spring 2017

Interpreting P-Values

- ☑ Probability of obtaining an effect equal to or more extreme than the one observed, presuming the null hypothesis is true

- ❑ ***NOT*** the probability that the null or the alternative hypothesis are correct or incorrect

Credit: CS1951A Spring 2017

Interpreting P-Values

- ☑ Probability of obtaining an effect equal to or more extreme than the one observed, presuming the null hypothesis is true

- ❑ ~~***NOT*** the probability that the null or the alternative hypothesis are correct or incorrect~~

Credit: CS1951A Spring 2017

Interpreting P-Values

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with P = 0.05, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect. This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true."

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

Credit: CS1951A Spring 2017

Interpreting P-Values

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with P = 0.05, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect. This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true."

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

Credit: CS1951A Spring 2017

Clicker Question!

Clicker Question!

If P=0.05, the null hypothesis has only a 5% chance of being true

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

If P=0.05, the null hypothesis has only a 5% chance of being true

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

If P=0.05, the null hypothesis has only a 5% chance of being true

a) Agree
b) Disagree
c) Don't know don't care

If we flip a coin four times and observe four heads, two-sided P =.125. This does not mean that the probability of the coin being fair is only 12.5%.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

If we observe a non-significant difference between two groups, (e.g., P=0.1), this means there is no difference between the groups.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

If we observe a non-significant difference between two groups, (e.g., P=0.1), this means there is no difference between the groups.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

If we observe a non-significant difference between two groups, (e.g., P=0.1), this means there is no difference between the groups.

a) Agree
b) Disagree
c) Don't know don't care

A non-significant difference only means the null effect is statistically consistent with the observation, not necessarily most likely.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

I test a new cancer treatment and find a significant decrease in tumor size for patients receiving the treatment compared to a control group. I should prescribe this treatment to all of my patients now.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

I test a new cancer treatment and find a significant decrease in tumor size for patients receiving the treatment compared to a control group. I should prescribe this treatment to all of my patients now.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

I test a new cancer treatment and find a significant decrease in tumor size for patients receiving the treatment compared to a control group. I should prescribe this treatment to all of my patients now.

a) Agree
b) Disagree
c) Don't know don't care

The P value carries no information about the magnitude of an effect. Significance doesn't correspond to clinical/practical relevance.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

P=0.05 means that the probability of data we have observed, plus anything more extreme, would only occur 5% of the time assuming the null hypothesis is true.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

P=0.05 means that the probability of data we have observed, plus anything more extreme, would only occur 5% of the time assuming the null hypothesis is true.

a) Agree
b) Disagree
c) Don't know don't care

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Clicker Question!

P=0.05 means that the probability of data we have observed, plus anything more extreme, would only occur 5% of the time assuming the null hypothesis is true.

a) Agree
b) Disagree
c) Don't know don't care

Yes, this is the definition. Internalize it. Live it. Breathe it. Tattoo it on your arm.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

Credit: CS1951A Spring 2017

Test for population proportion

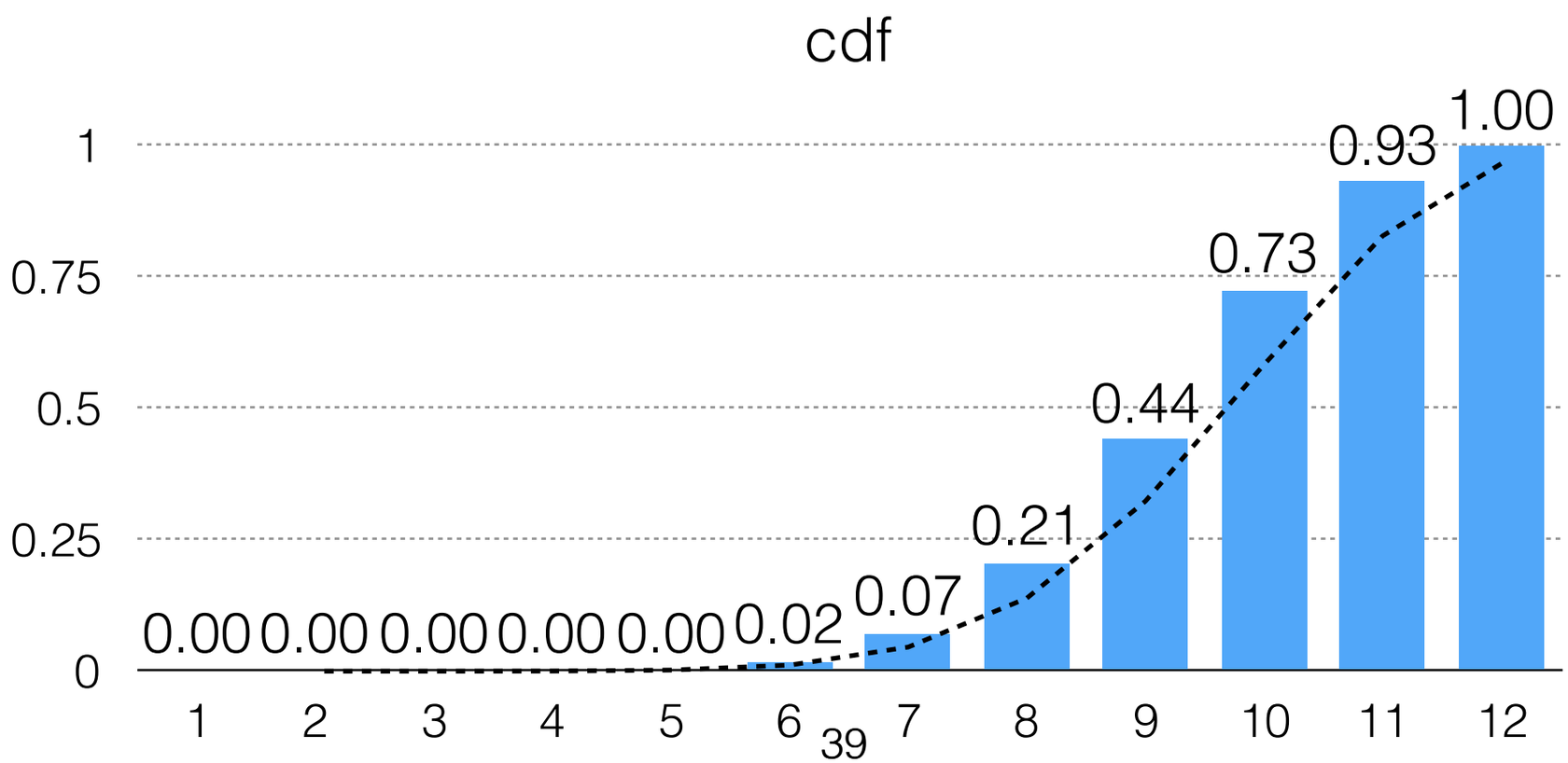
Test for population proportion

$\langle \Omega, F, P \rangle$

{b, not b} p(B) = 0.8

H_0 = proportion of (b) is 80%

H_a = proportion of (b) is not 80%

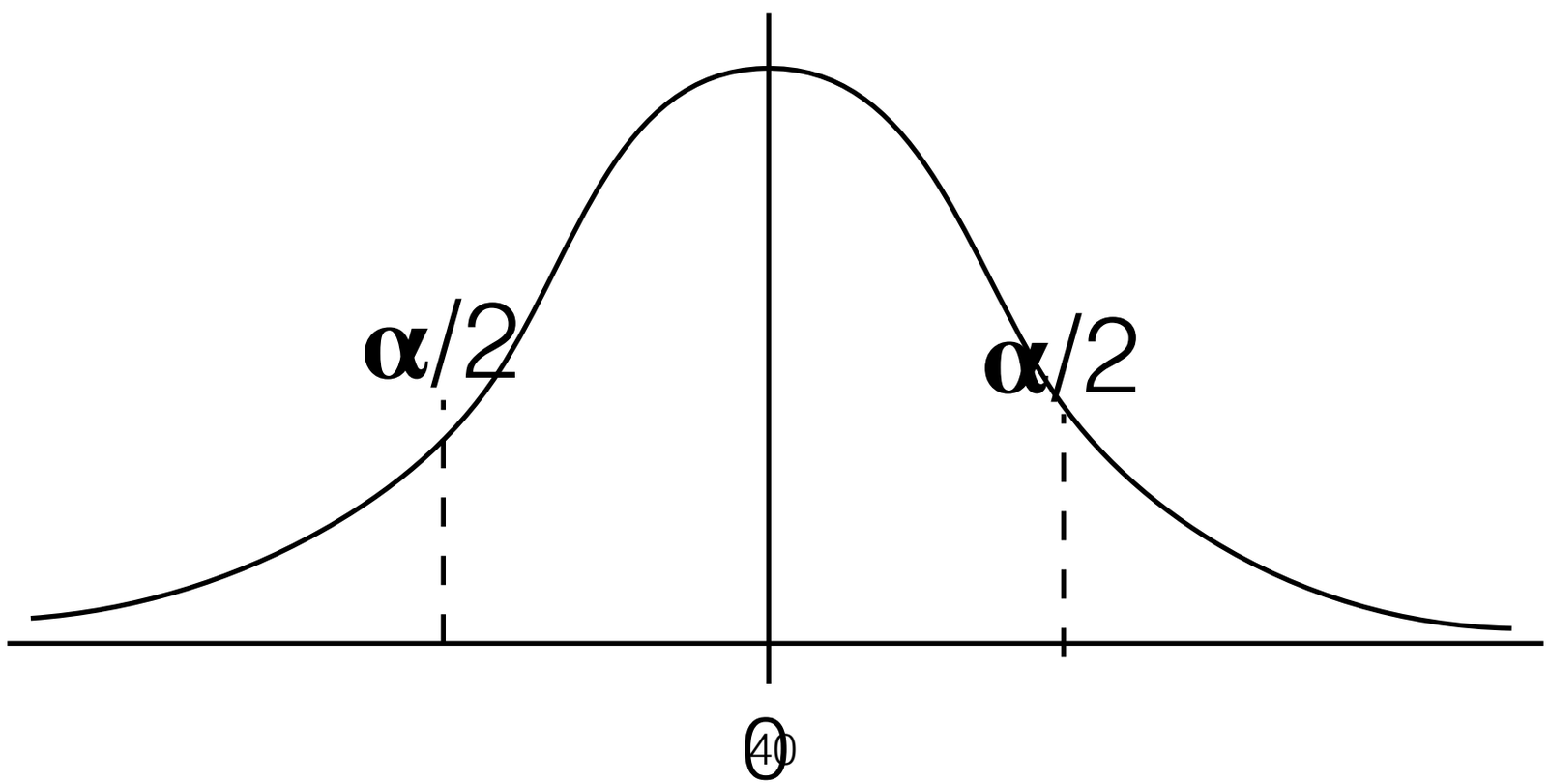


Test for population proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

H_0 = proportion of (b) is 80%

H_a = proportion of (b) is not 80%



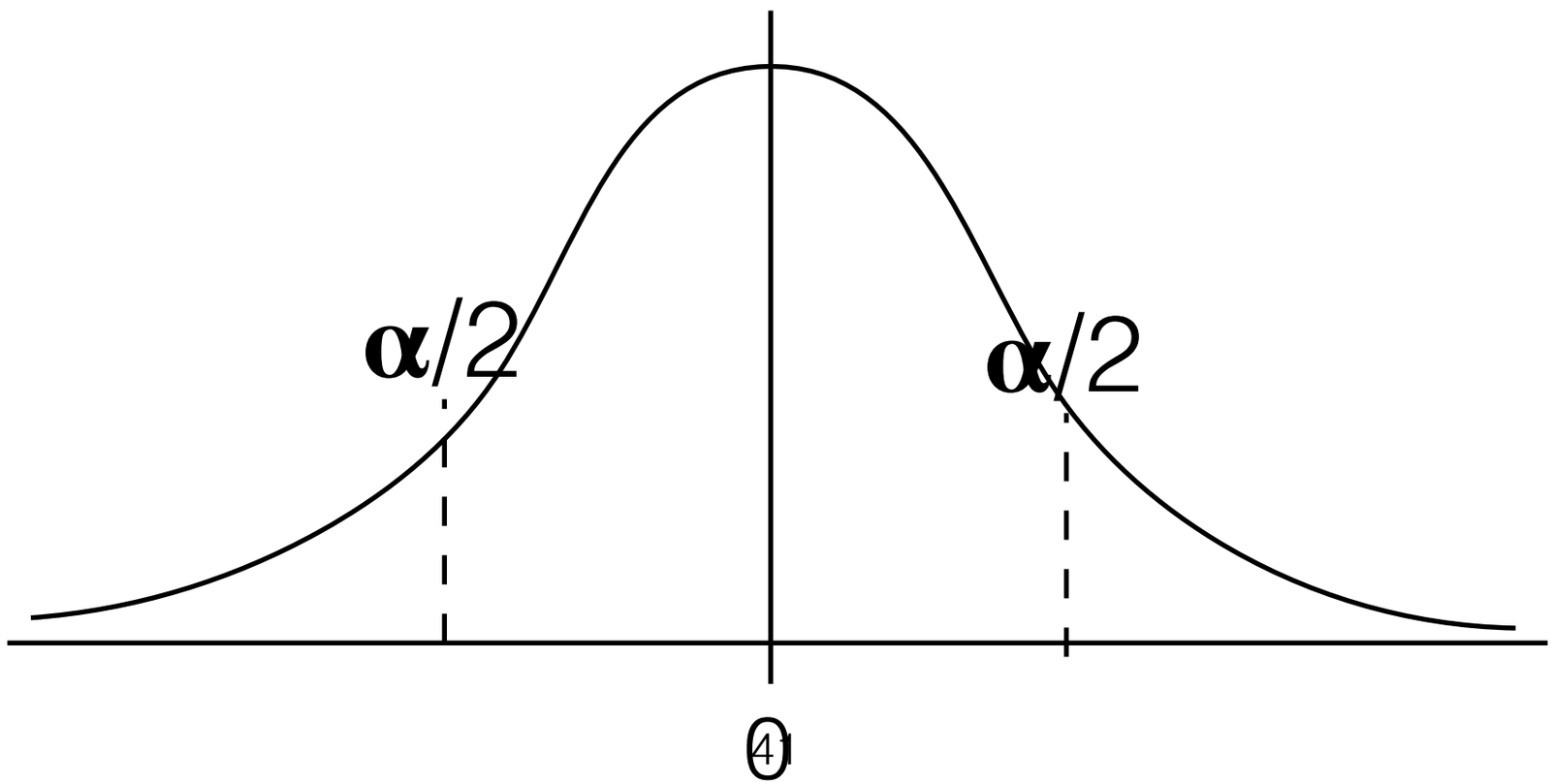
Test for population proportion

observed proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

H_0 = proportion of (b) is 80%

H_a = proportion of (b) is not 80%



Test for population proportion

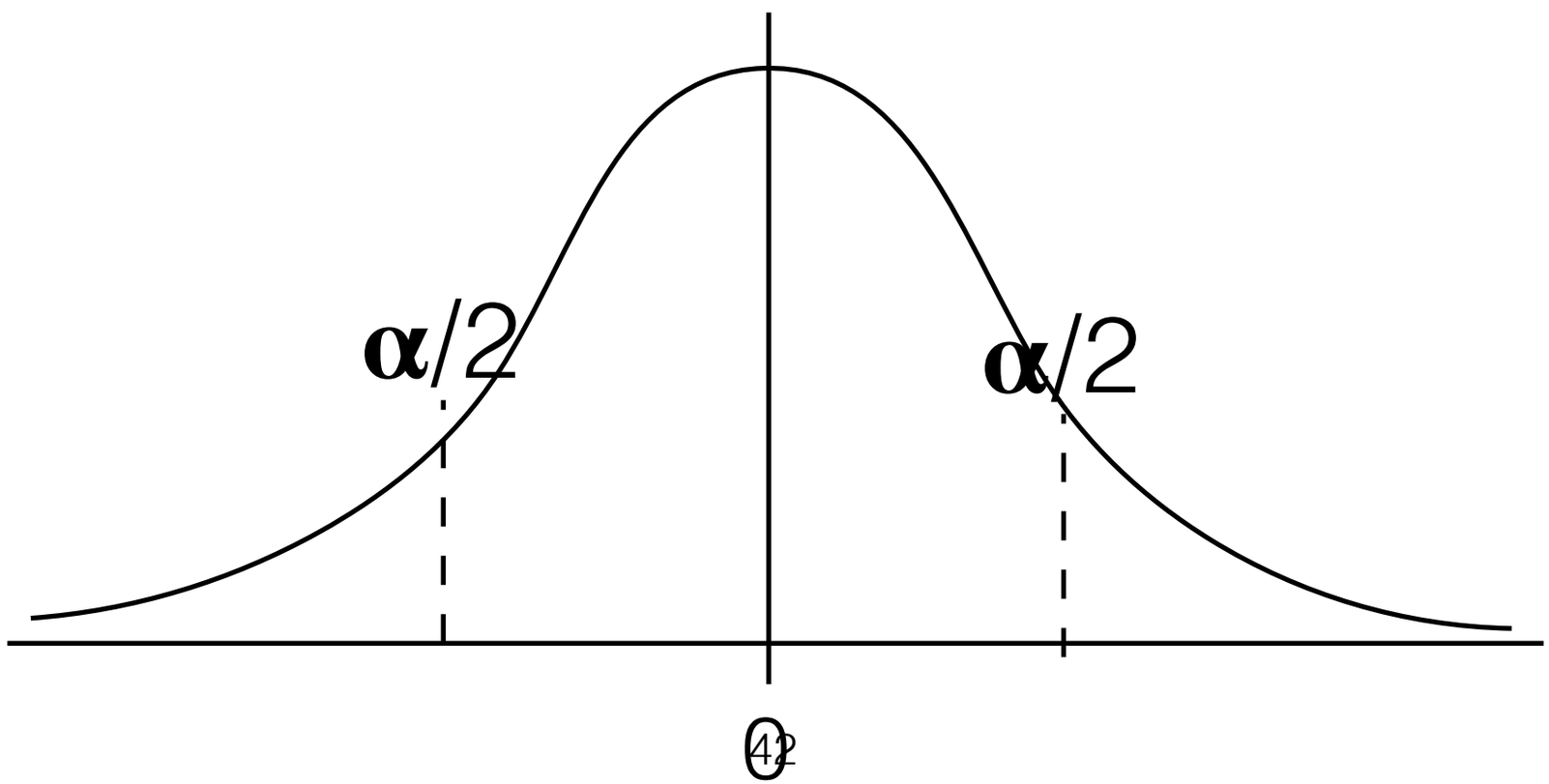
observed proportion

assumed proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

H_0 = proportion of (b) is 80%

H_a = proportion of (b) is not 80%



Test for population proportion

observed proportion

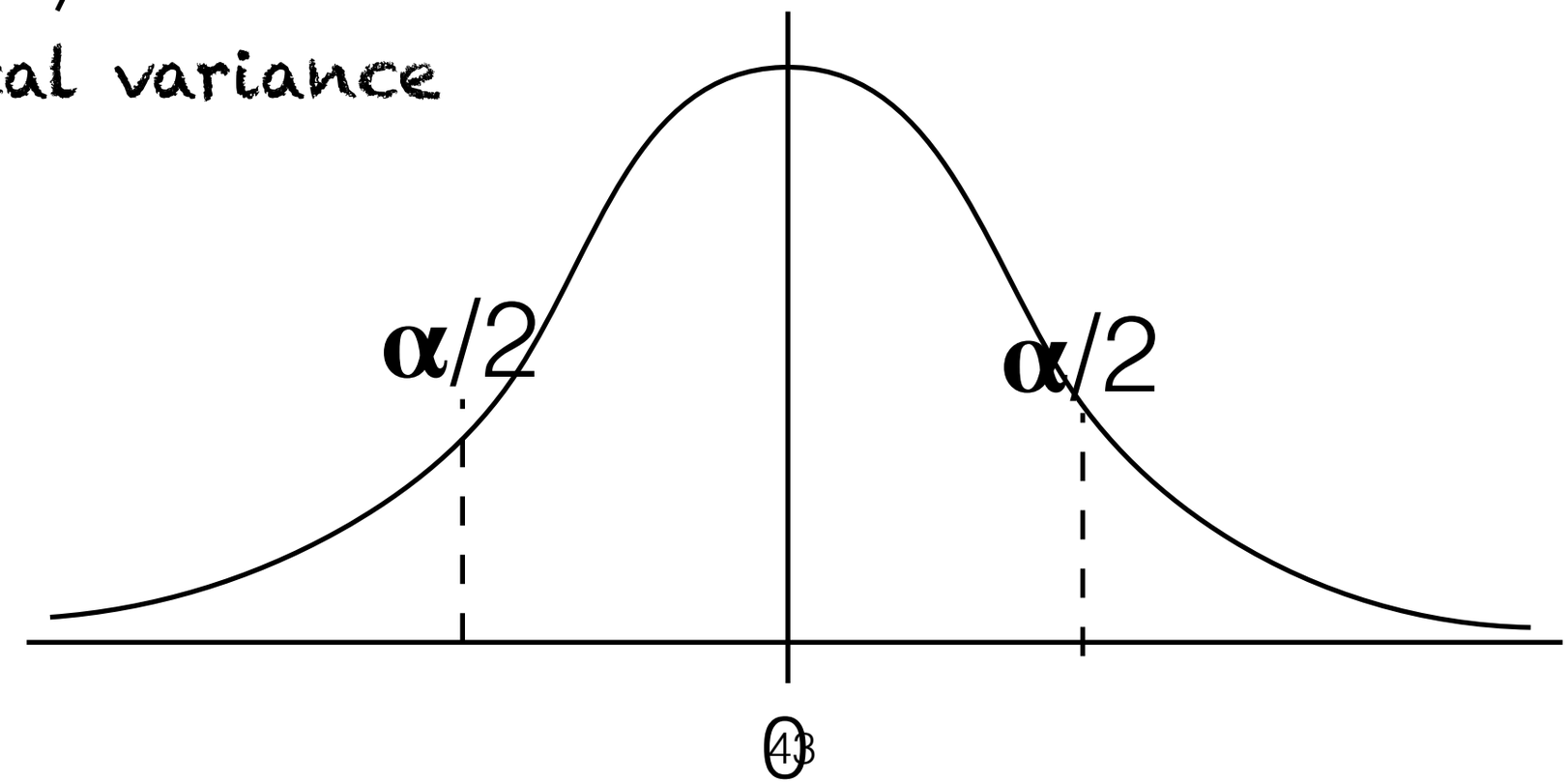
assumed proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

theoretical variance

H_0 = proportion of (b) is 80%

H_a = proportion of (b) is not 80%



Clicker Question!

Clicker Question!

Why can we use a normally-distributed test statistic to evaluate a binomial distribution like this?

a) Because its a random variable
b) Because of regression to the mean
c) Because of the law of large numbers
d) Because of the central limit theorem
e) The limit does not exist!

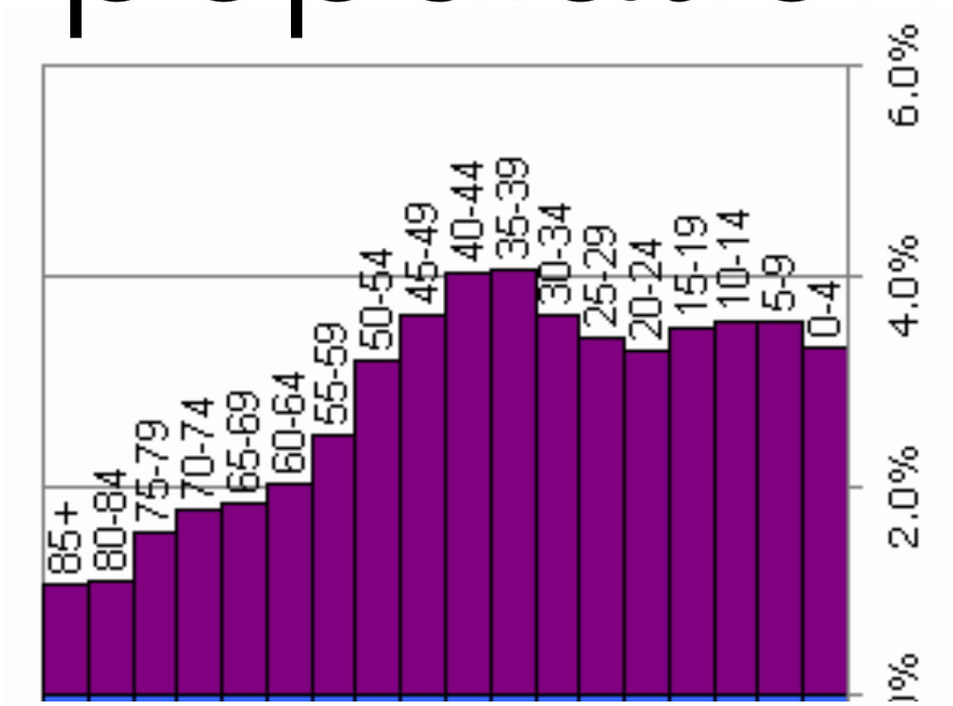
Clicker Question!

Why can we use a normally-distributed test statistic to evaluate a binomial distribution like this?

a) Because its a random variable
b) Because of regression to the mean
c) Because of the law of large numbers
d) Because of the central limit theorem
e) The limit does not exist!

Test for population means

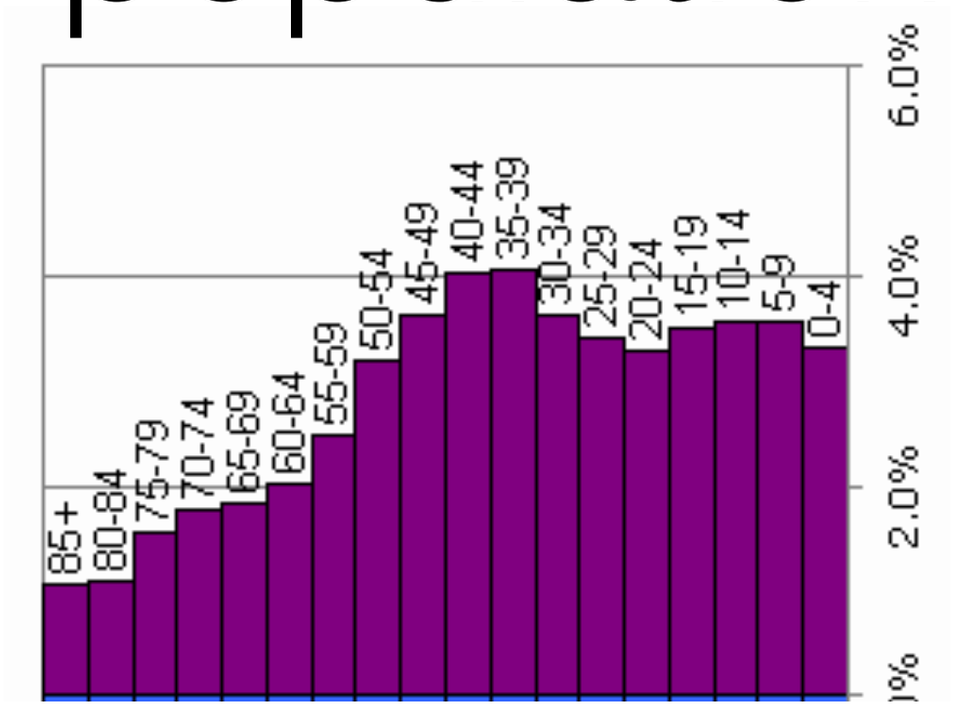
Test for population means



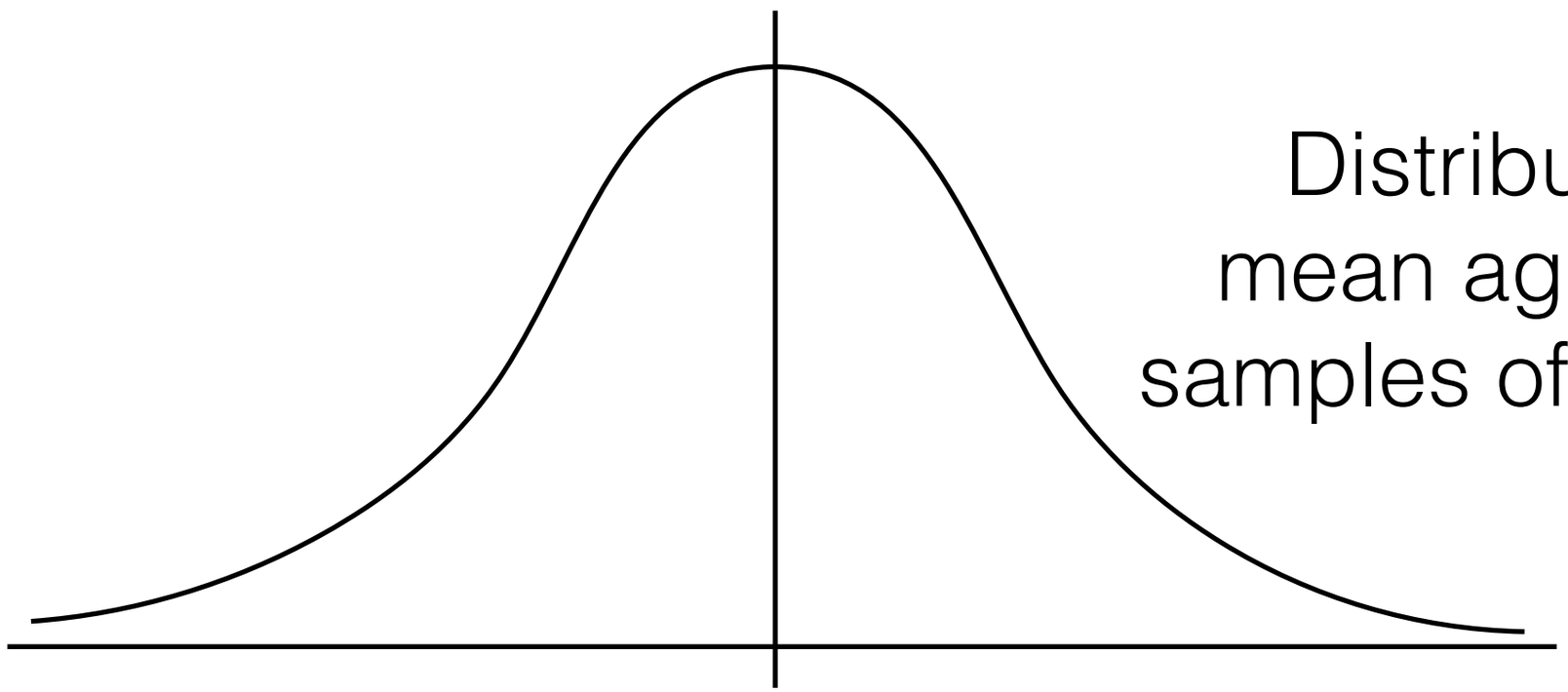
Distribution of ages in the US

http://www.censusscope.org/us/chart_age.html

Test for population means



Distribution of ages in the US

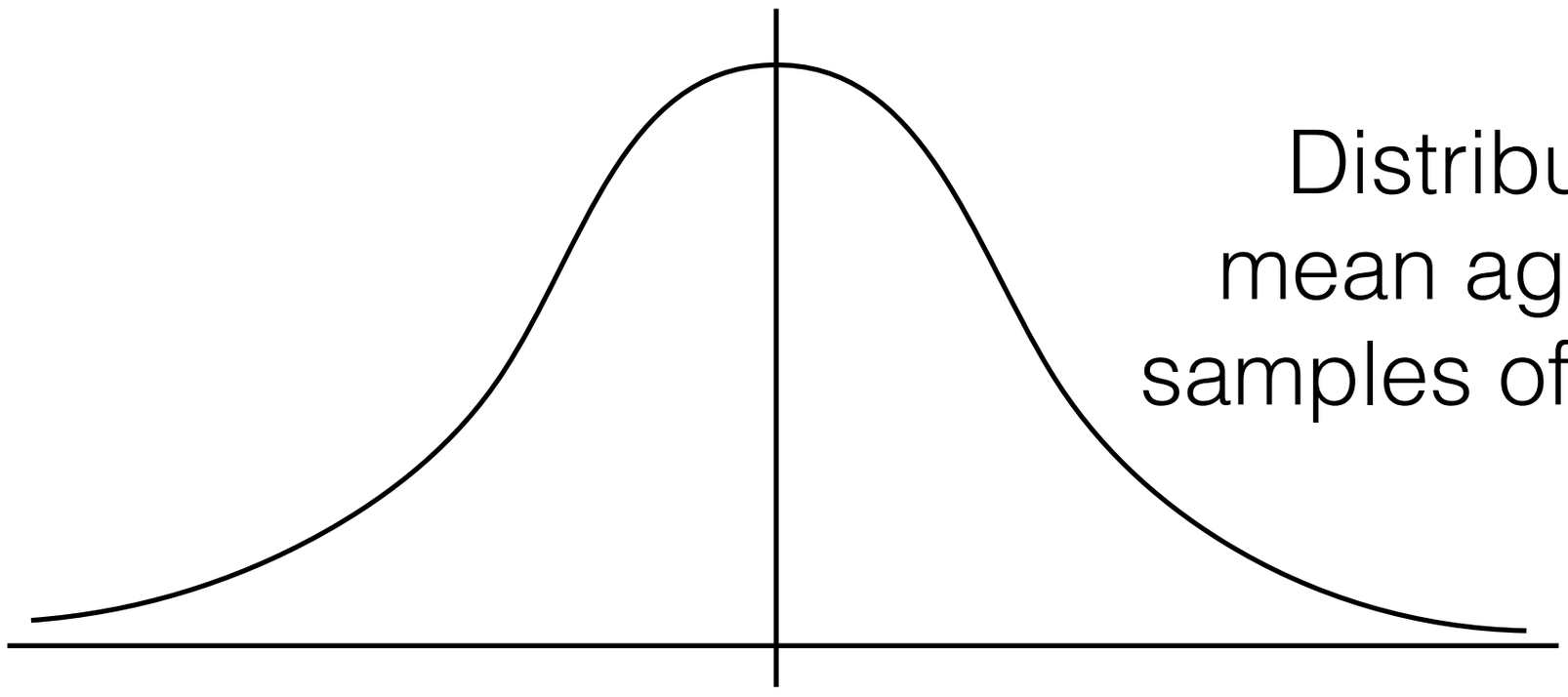


Distribution of mean ages given samples of US adults

http://www.censusscope.org/us/chart_age.html

Test for population means

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$



Distribution of mean ages given samples of US adults

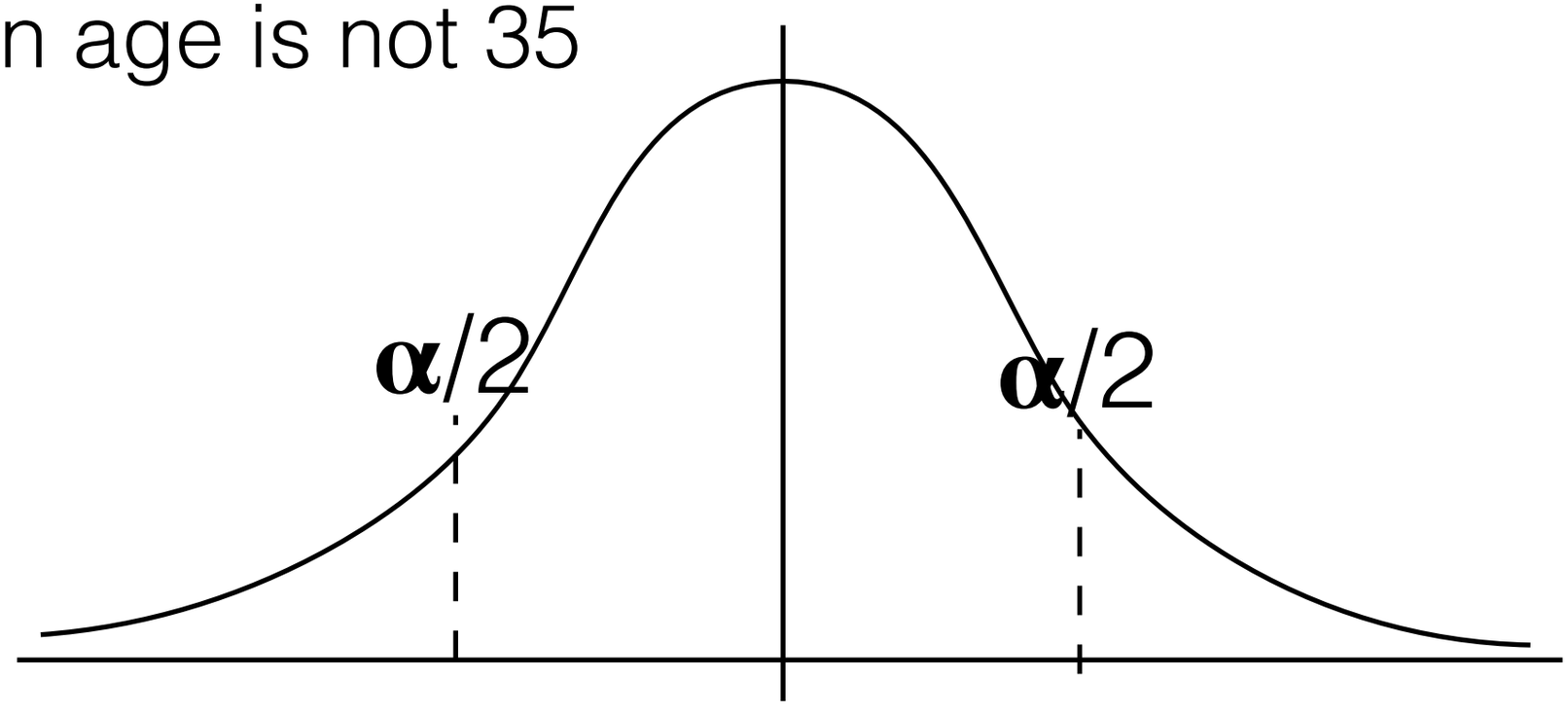
http://www.censusscope.org/us/chart_age.html

Test for population means

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

H_0 = mean age is 35

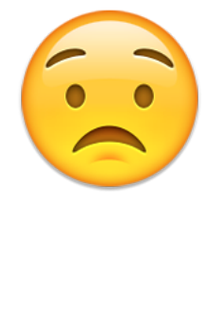
H_a = mean age is not 35



http://www.censusscope.org/us/chart_age.html

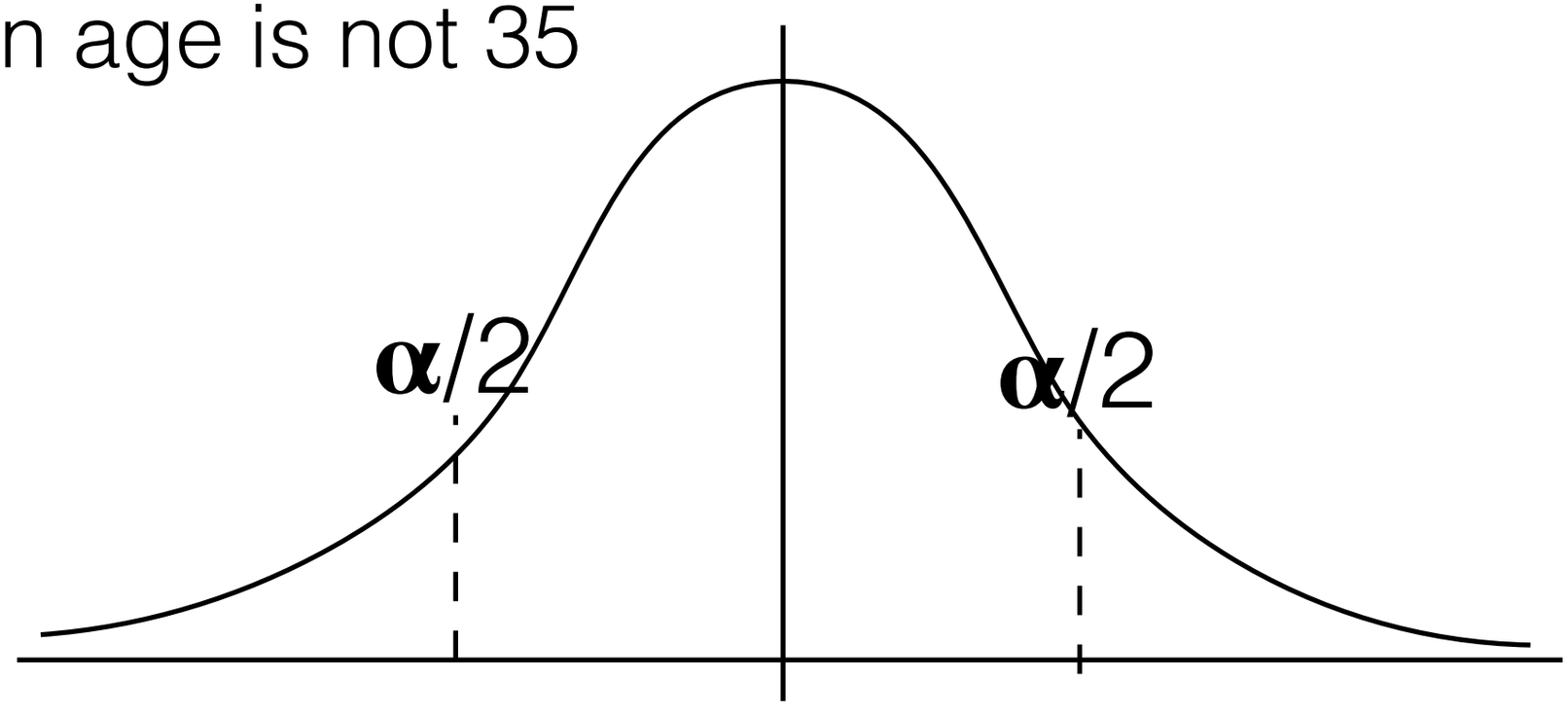
Test for population means

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$



H_0 = mean age is 35

H_a = mean age is not 35



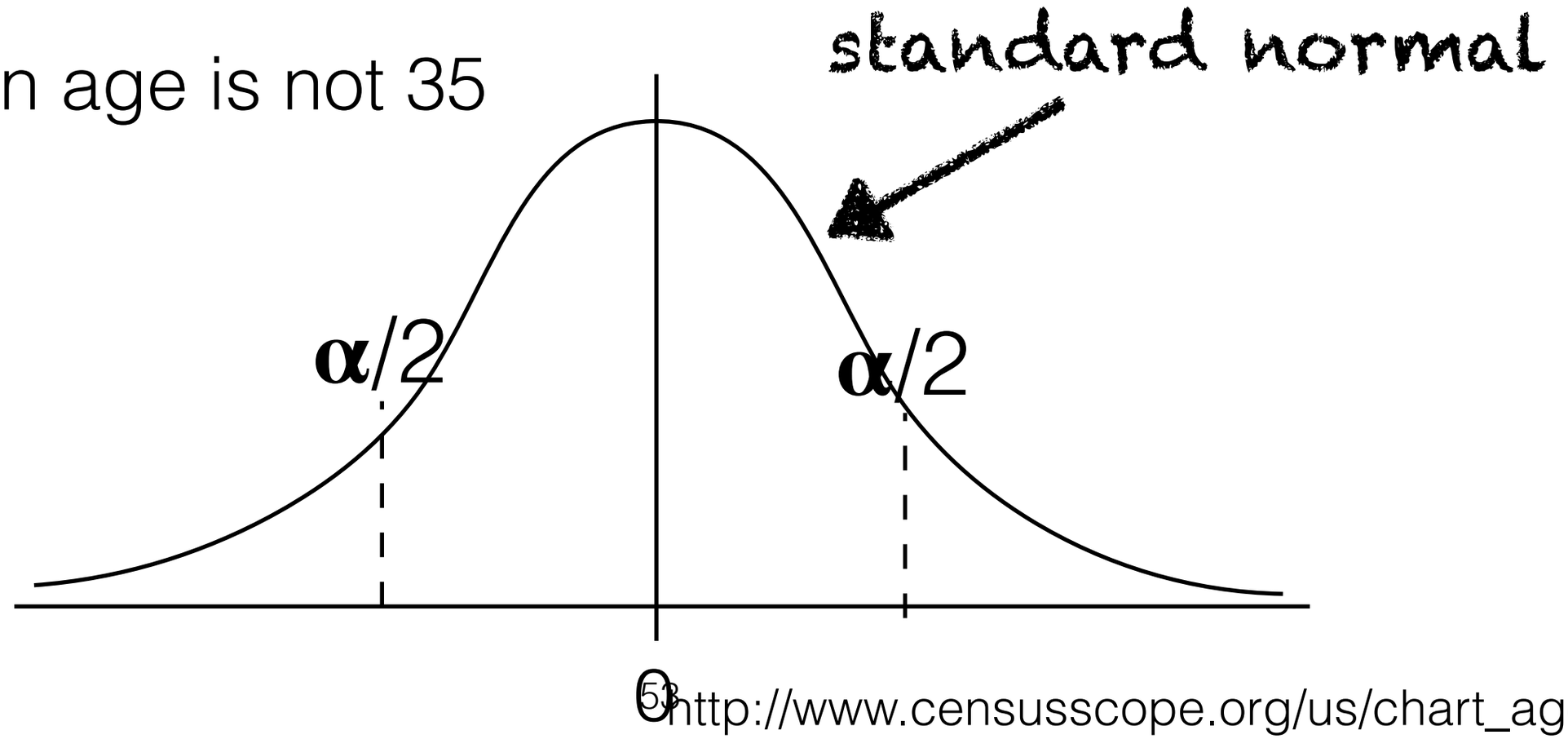
http://www.censusscope.org/us/chart_age.html

Test for population means

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

H_0 = mean age is 35

H_a = mean age is not 35



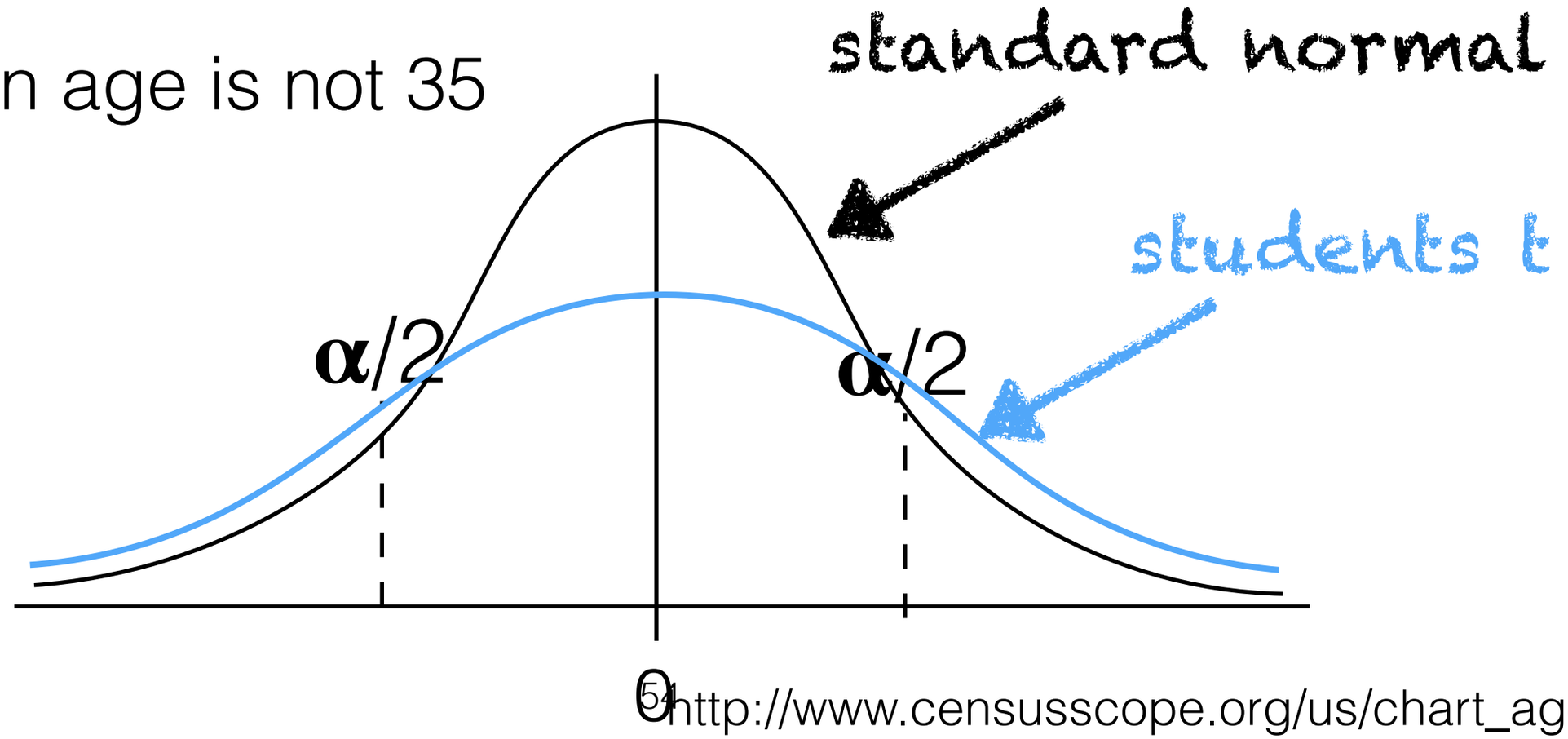
http://www.censusscope.org/us/chart_age.html

Test for population means

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

H_0 = mean age is 35

H_a = mean age is not 35



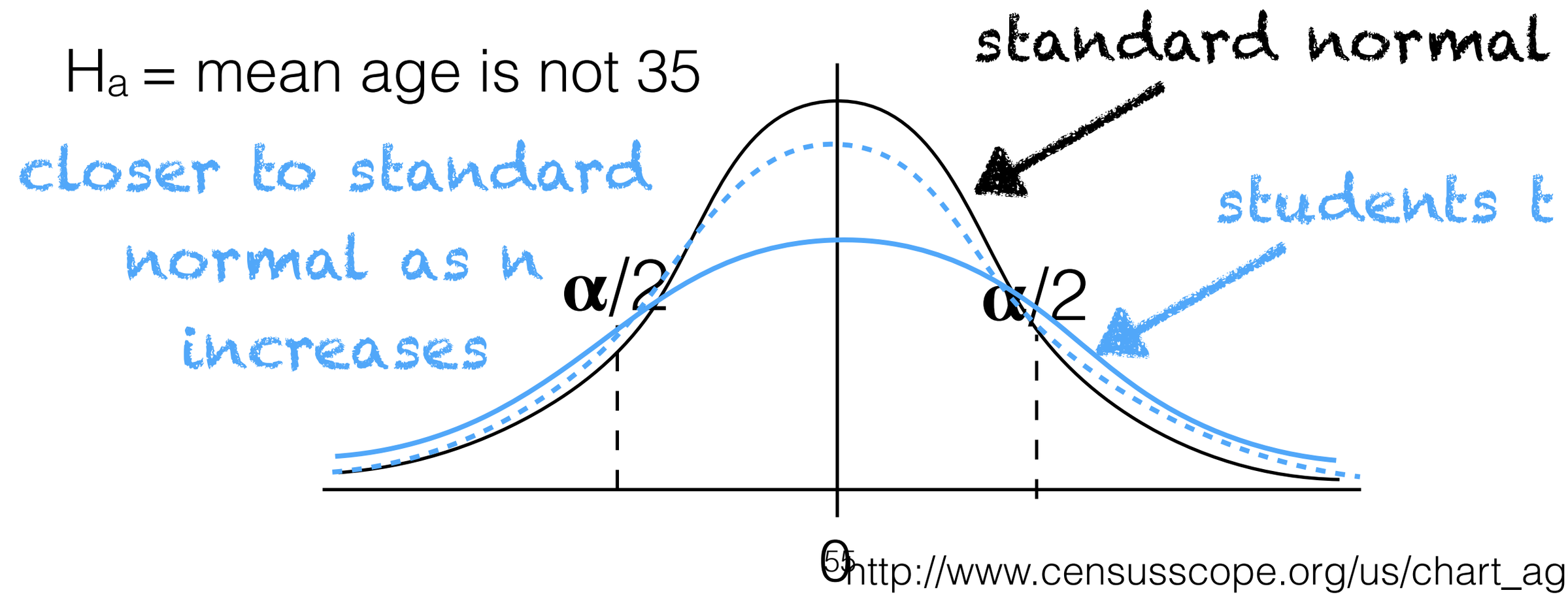
http://www.censusscope.org/us/chart_age.html

Test for population means

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

H_0 = mean age is 35

H_a = mean age is not 35



http://www.censusscope.org/us/chart_age.html

Permutation Test

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

- Still trying to determine the probability of the test statistic under the null hypothesis…same old same old emiright?

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

- Still trying to determine the probability of the test statistic under the null hypothesis…same old same old emiright?

- But don’t have an analytic solution, maybe because

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

- Still trying to determine the probability of the test statistic under the null hypothesis…same old same old emiright?

- But don’t have an analytic solution, maybe because

    - Form of underlying distribution is complex or hard to write down

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

- Still trying to determine the probability of the test statistic under the null hypothesis…same old same old emiright?

- But don’t have an analytic solution, maybe because

  - Form of underlying distribution is complex or hard to write down

  - Assumptions about analytic solution are suspect (e.g. sample size not large enough)

Permutation Test

- “Non-parametric” — doesn’t assume a particular form for the distribution

- Still trying to determine the probability of the test statistic under the null hypothesis…same old same old emiright?

- But don’t have an analytic solution, maybe because

    - Form of underlying distribution is complex or hard to write down

    - Assumptions about analytic solution are suspect (e.g. sample size not large enough)

- Related: bootstrapping (won’t discuss this today)

Permutation Test

H_a: CS students sleep less than the rest of Brown students

Permutation Test 

H_a: CS students sleep less than the rest of Brown students

Permutation Test

H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students

Permutation Test

H_0: CS students sleep the same amount as everyone else

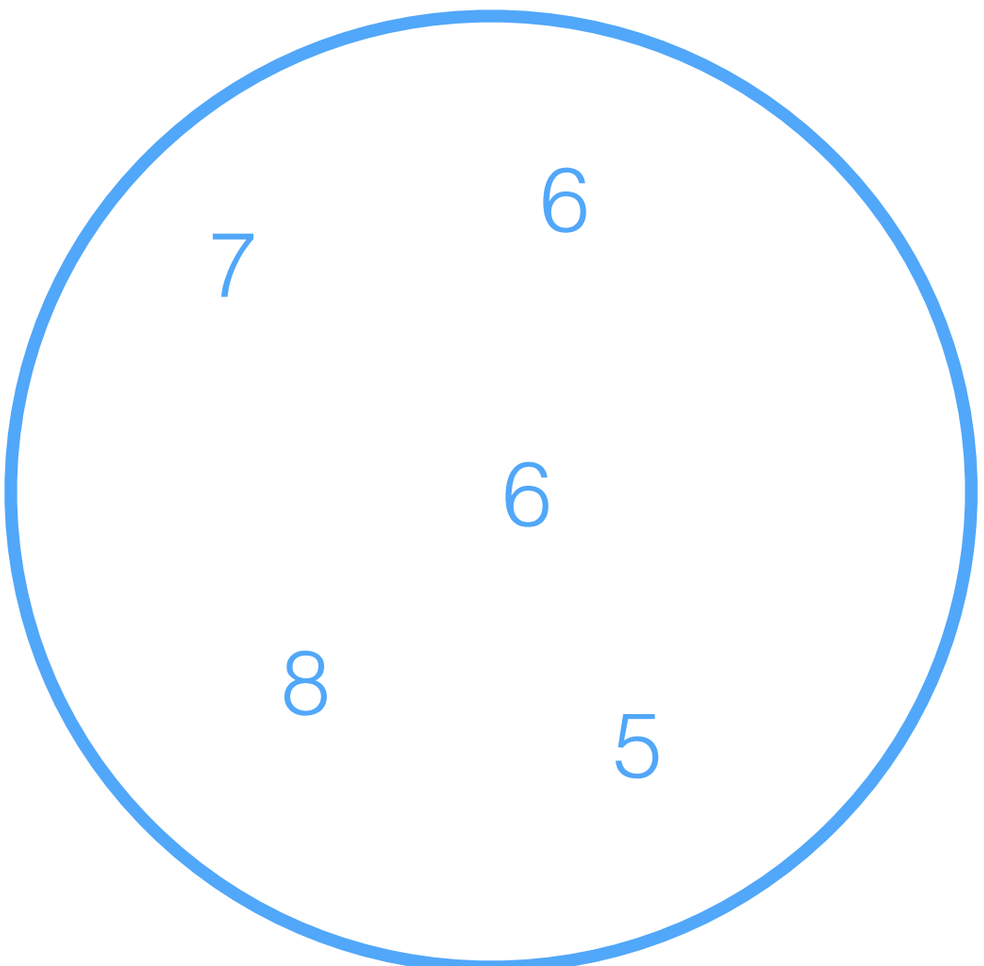
H_a: CS students sleep less than the rest of Brown students

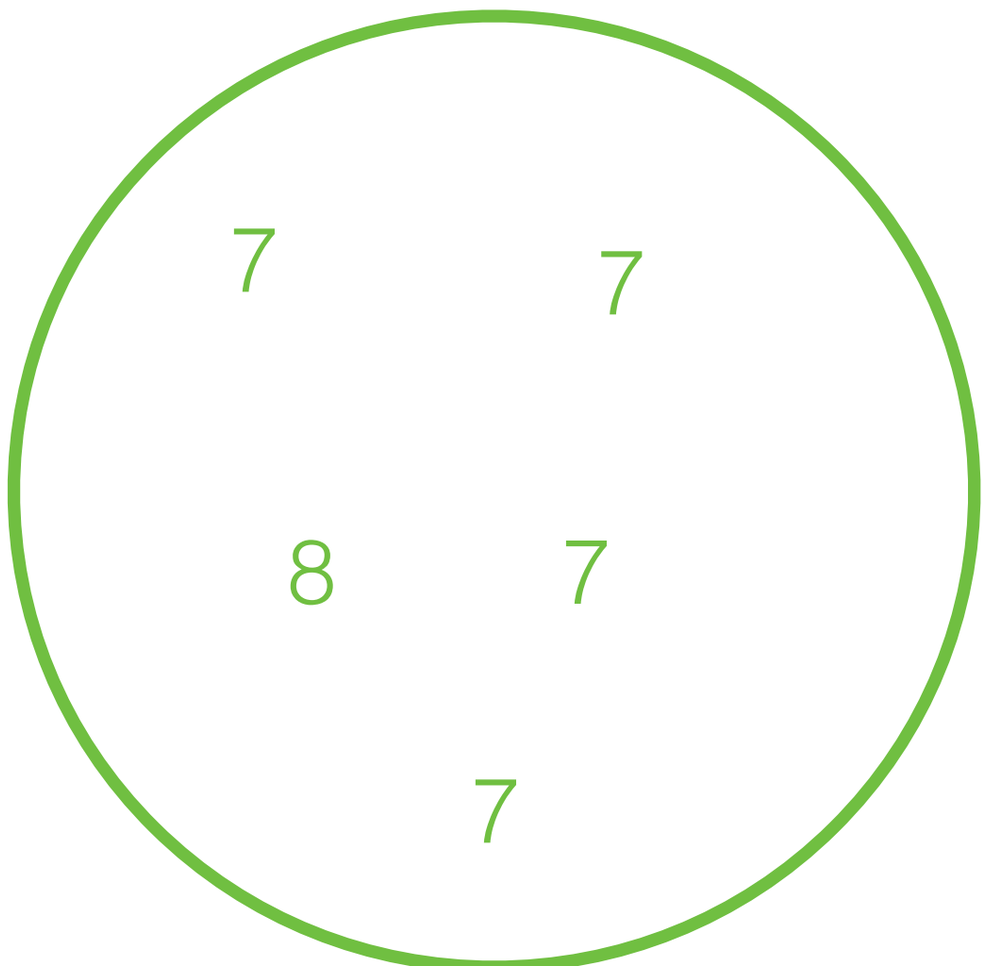
CS Students

6.4

Brown Overall

7.2





Permutation Test

H_0: CS students sleep **the same** amount as everyone else

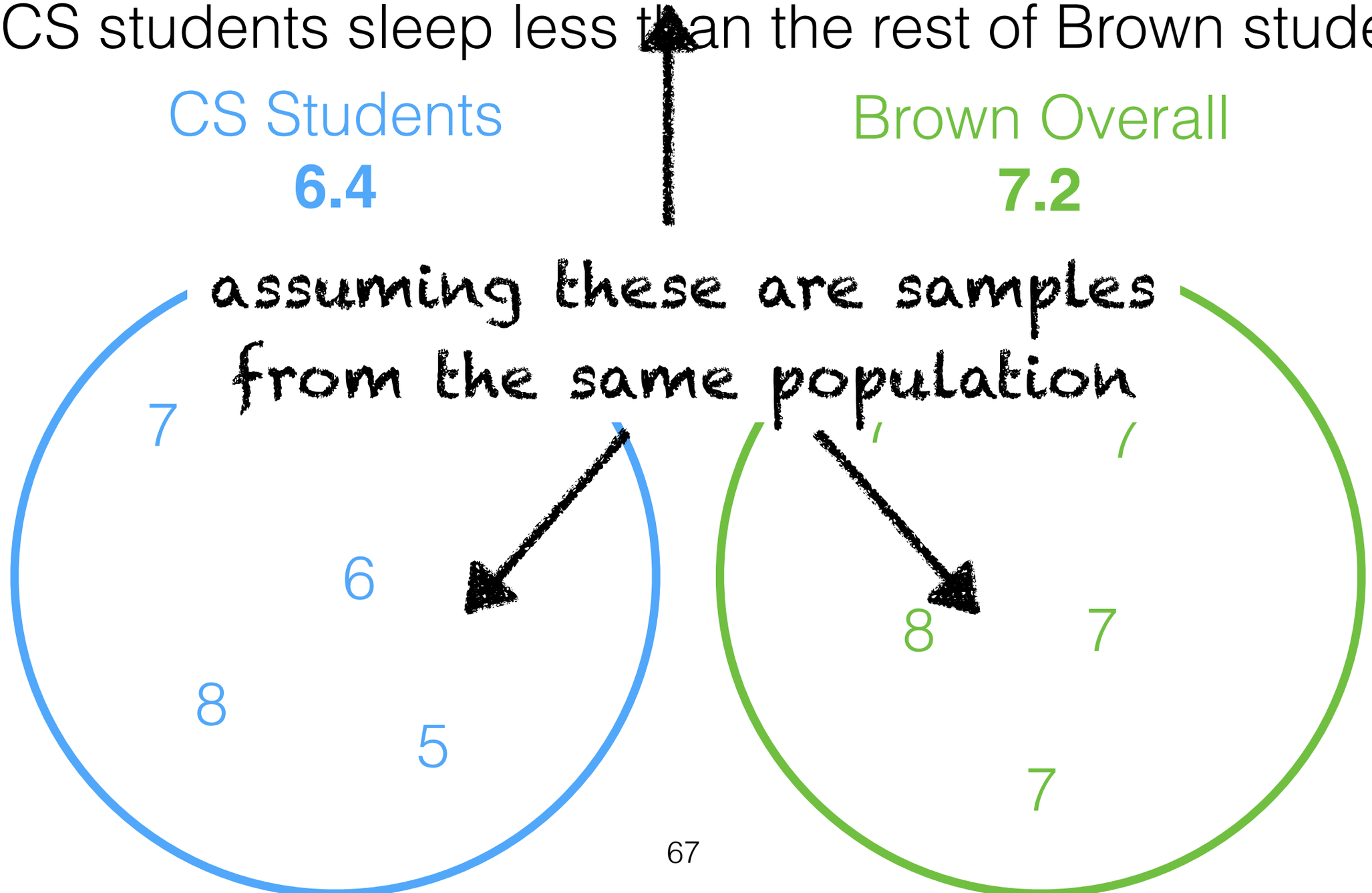
H_a: CS students sleep less than the rest of Brown students

CS Students

6.4

Brown Overall

7.2



Permutation Test

H_0: CS students sleep the same amount as everyone else

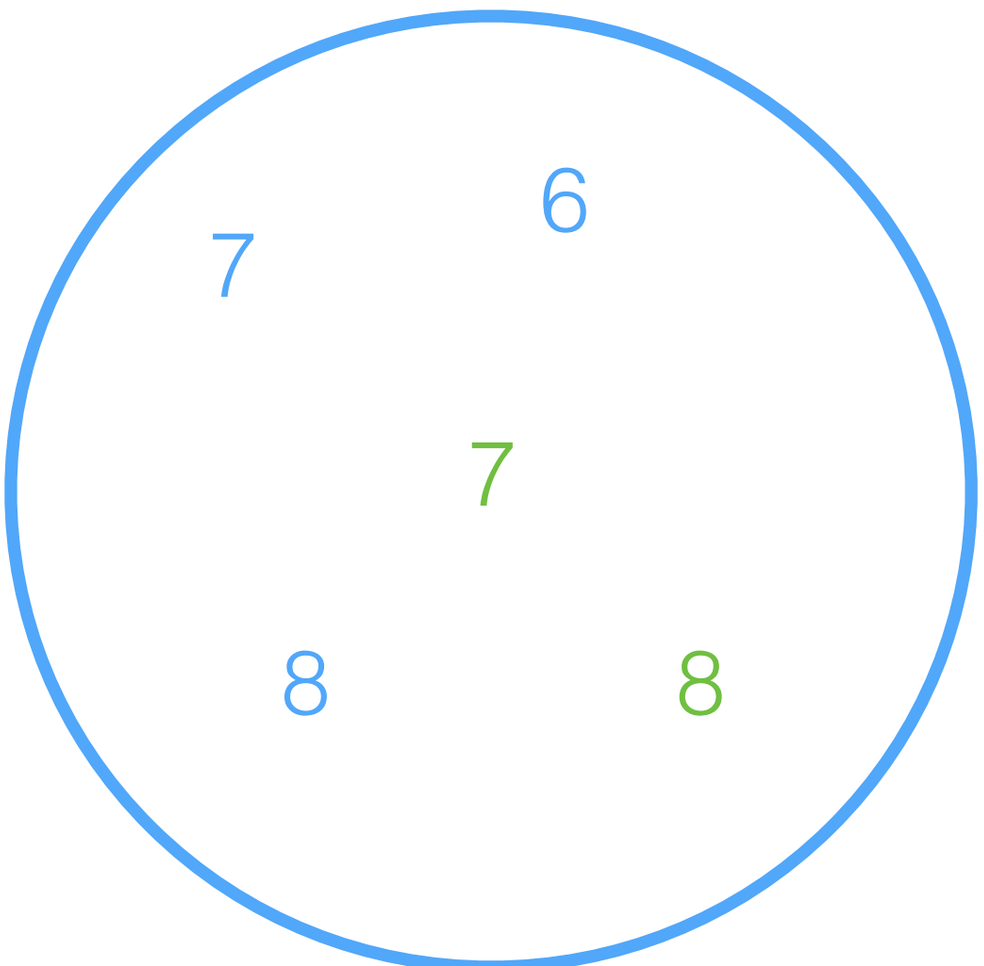
H_a: CS students sleep less than the rest of Brown students

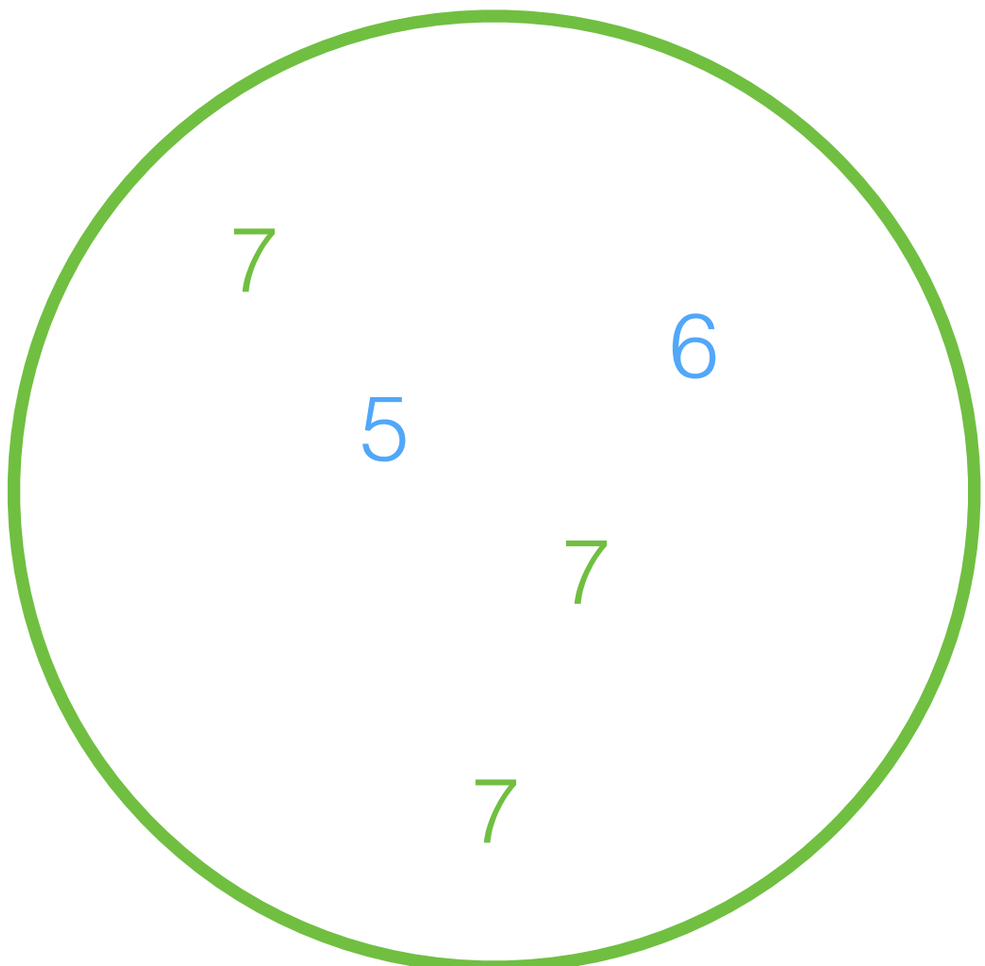
CS Students

7.2

Brown Overall

6.4





Permutation Test

H_0: CS students sleep the same amount as everyone else

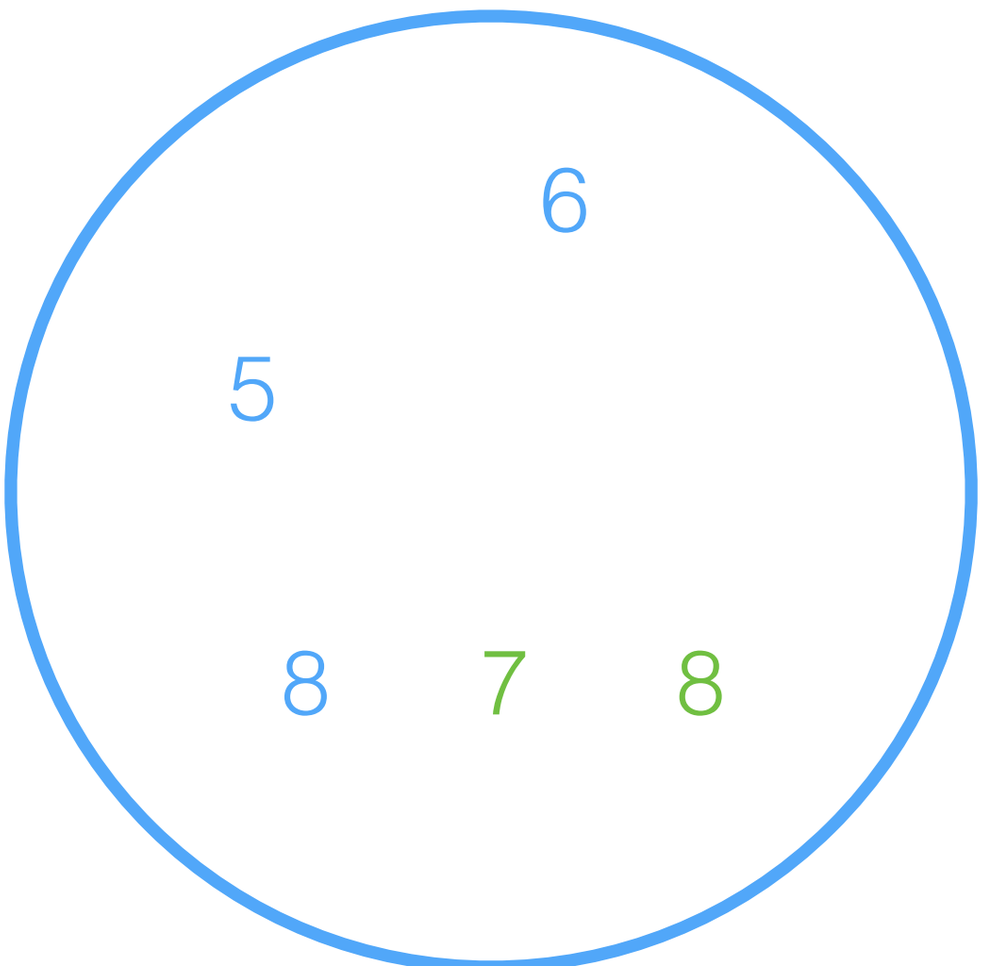
H_a: CS students sleep less than the rest of Brown students

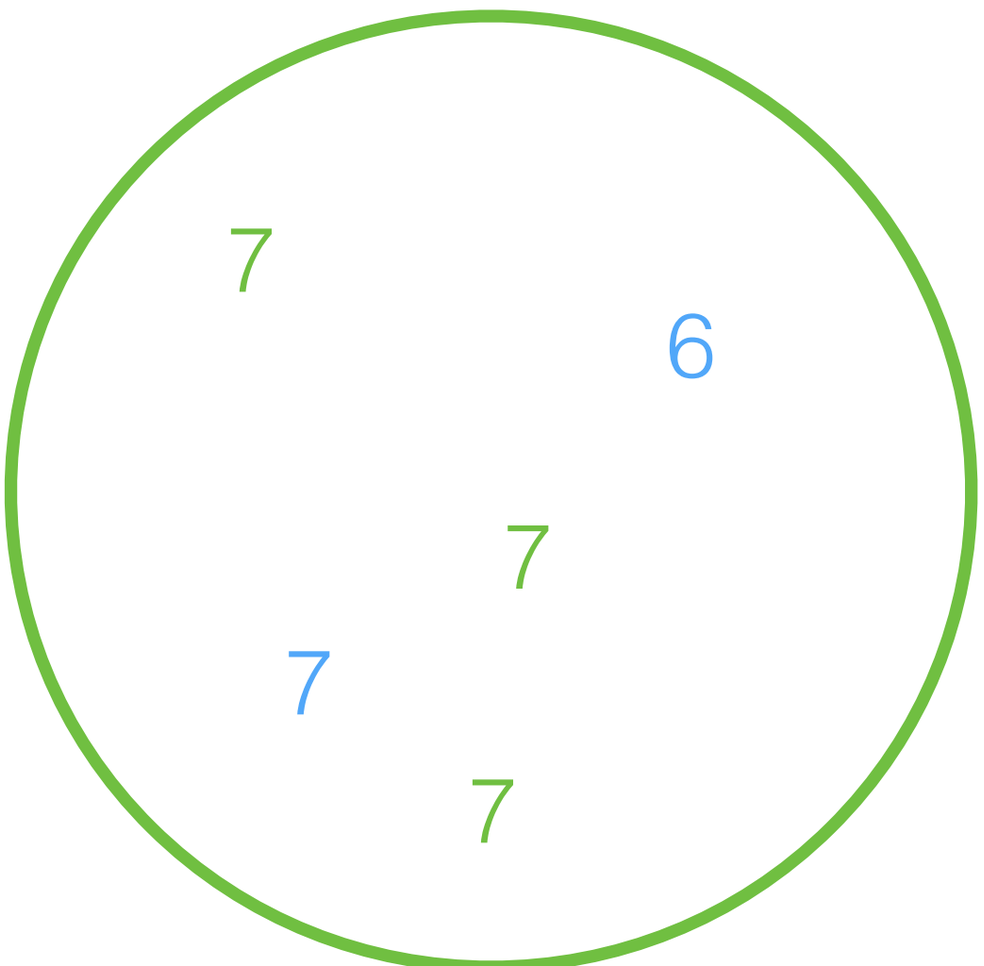
CS Students

6.8

Brown Overall

6.8





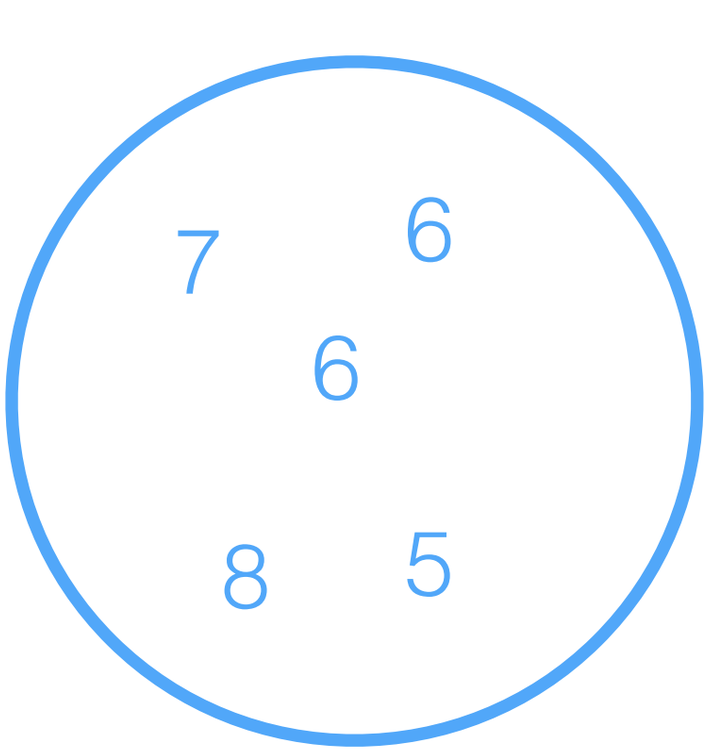
Permutation Test

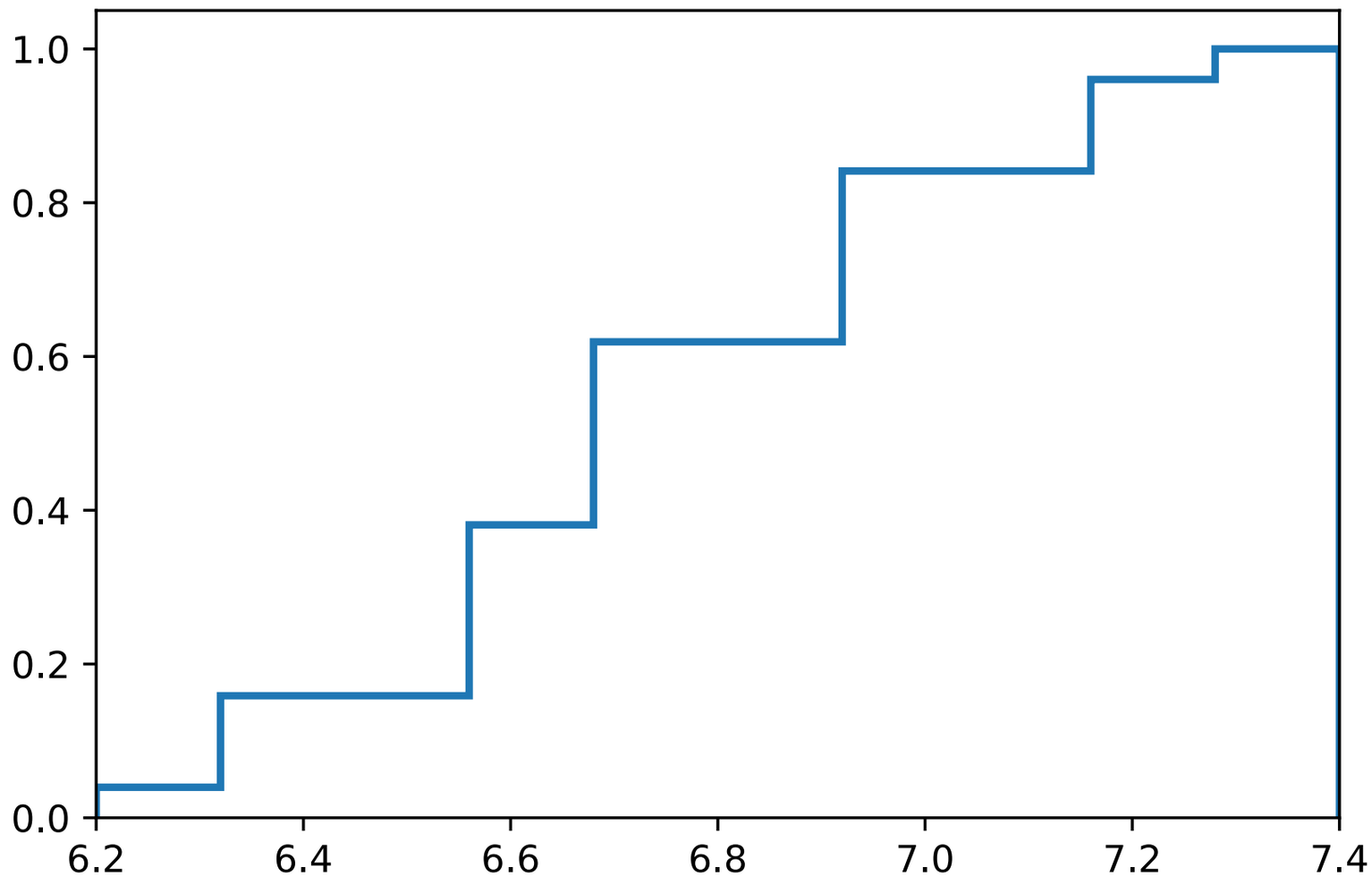
H_0: CS students sleep the same amount as everyone else

H_a: CS students sleep less than the rest of Brown students

CS Students

6.4





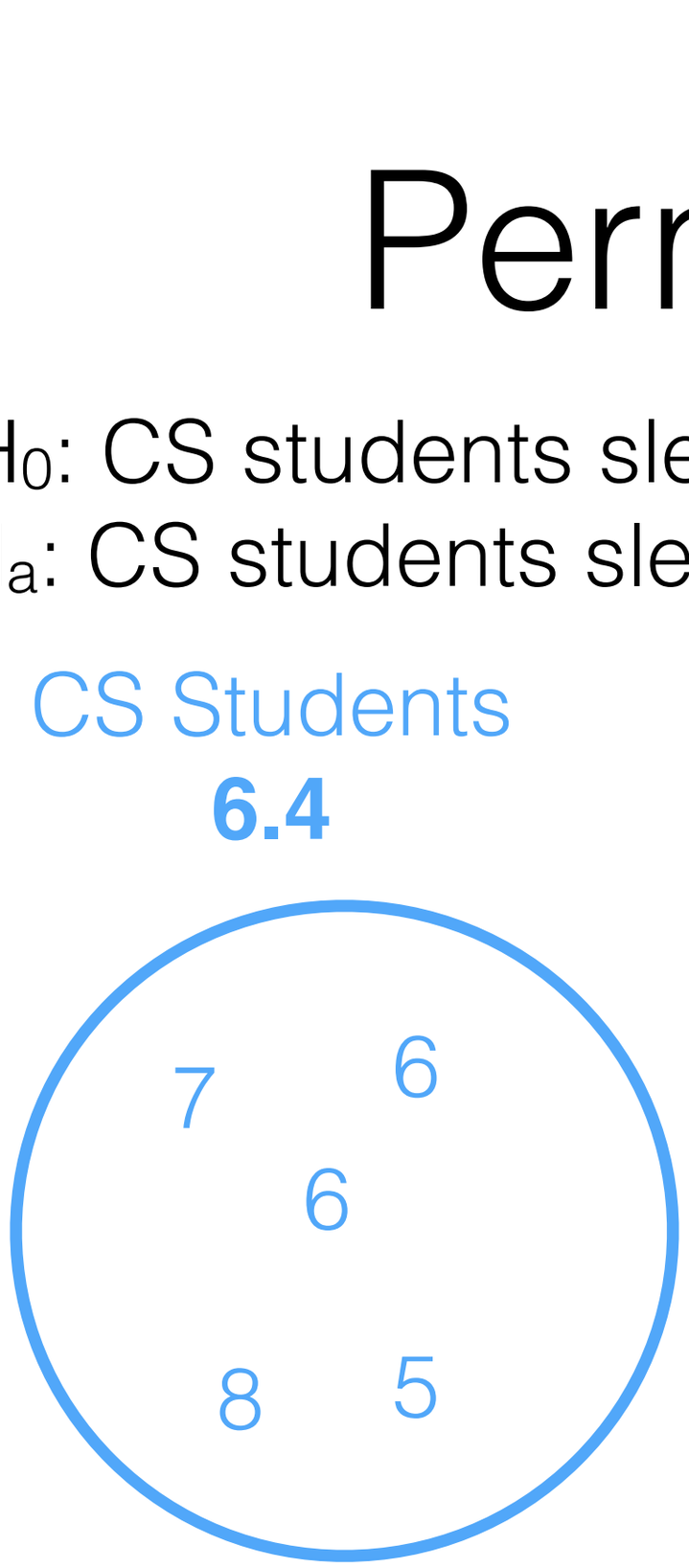
Permutation Test

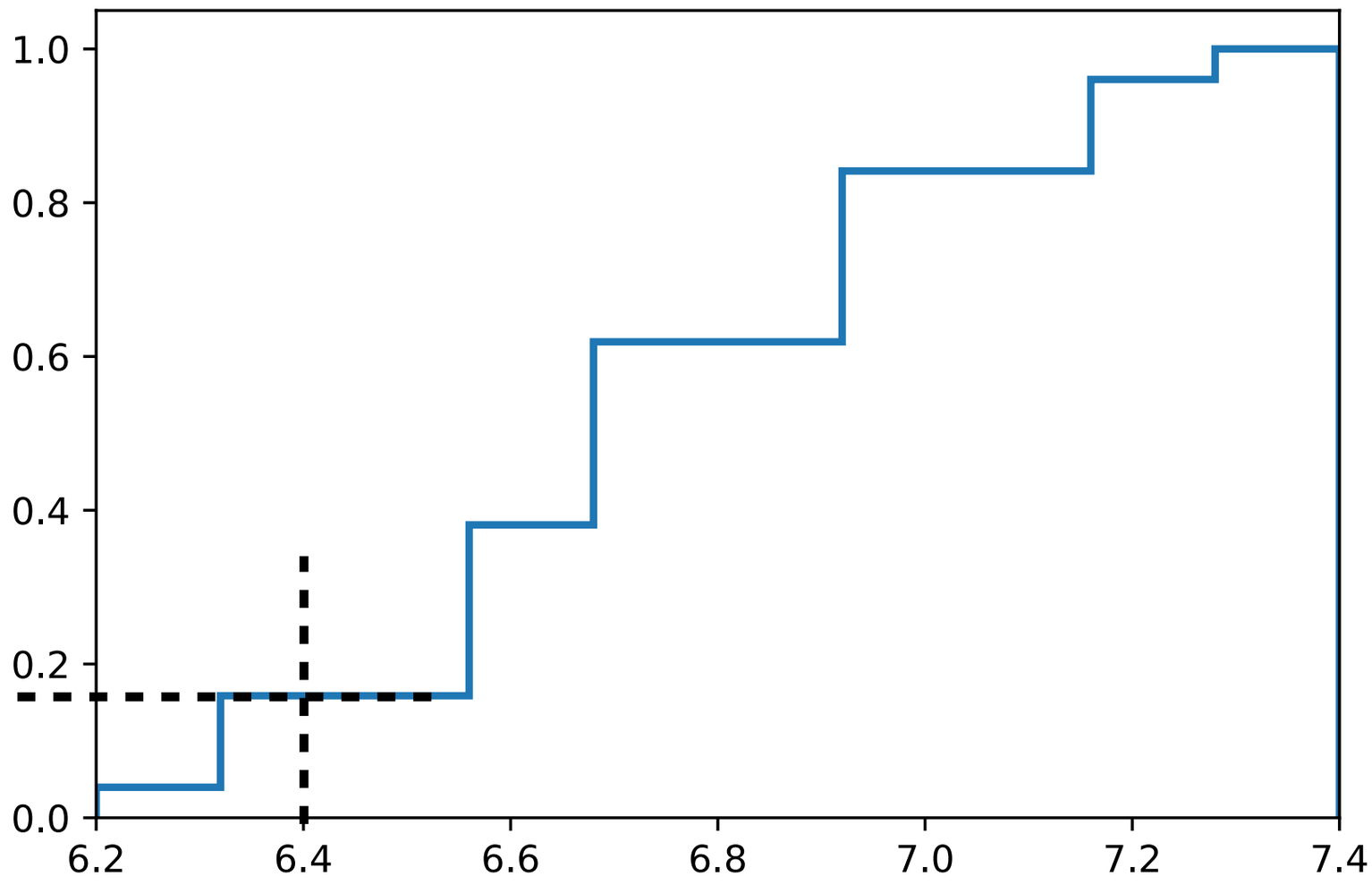
H_0: CS students sleep the same amount as everyone else

H_a: CS students sleep less than the rest of Brown students

CS Students

6.4





Regression

Regression

$$y = f(x)$$

Regression

```
cholesterol = f(mg eucalyptus oil)
```

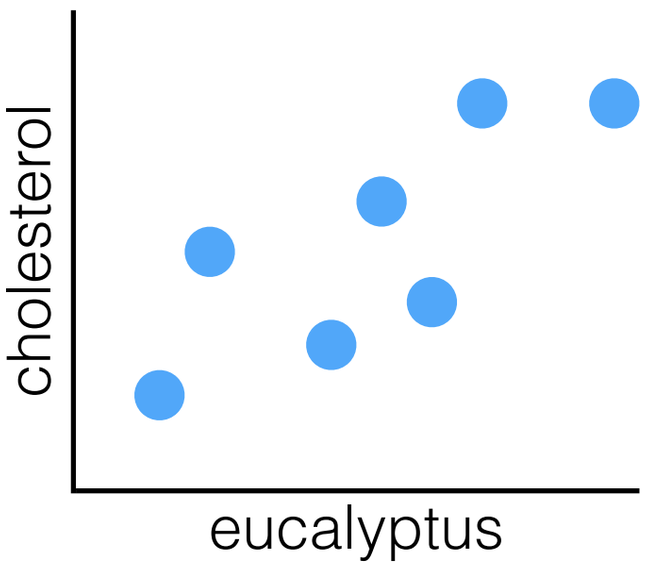
Regression

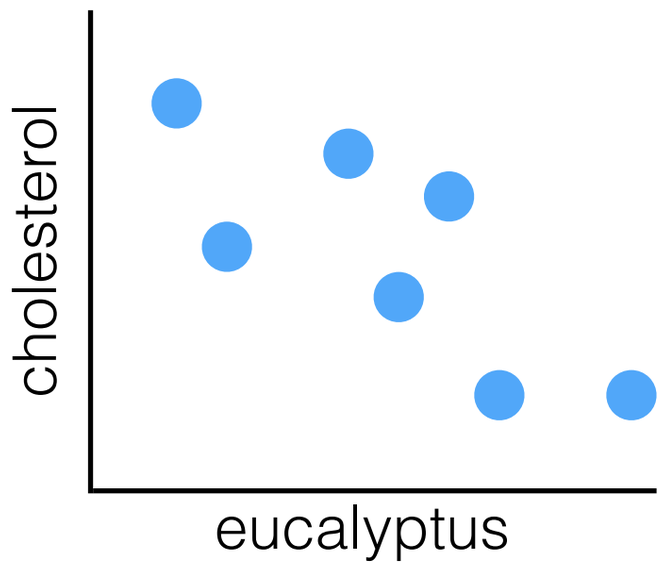
```
cholesterol = f(mg eucalyptus oil)
```

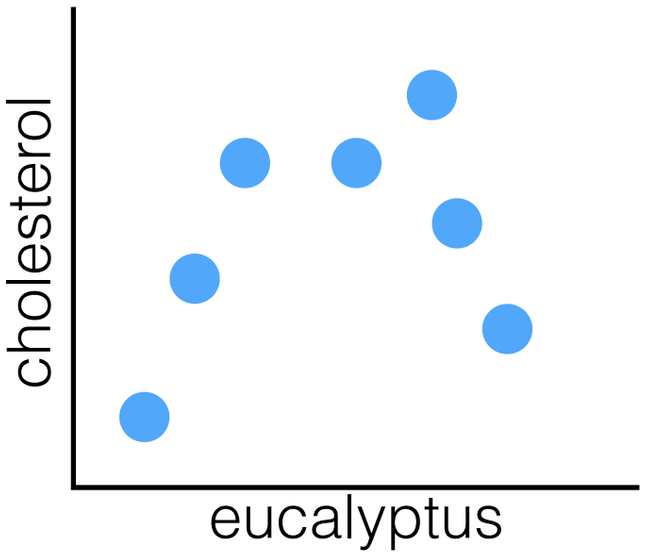
look at your data!
plot early. plot often.

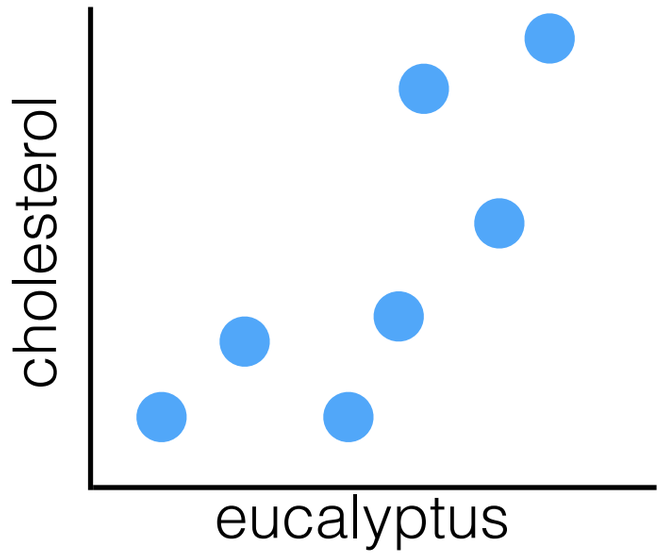
Regression

`cholesterol = f(mg eucalyptus oil)`





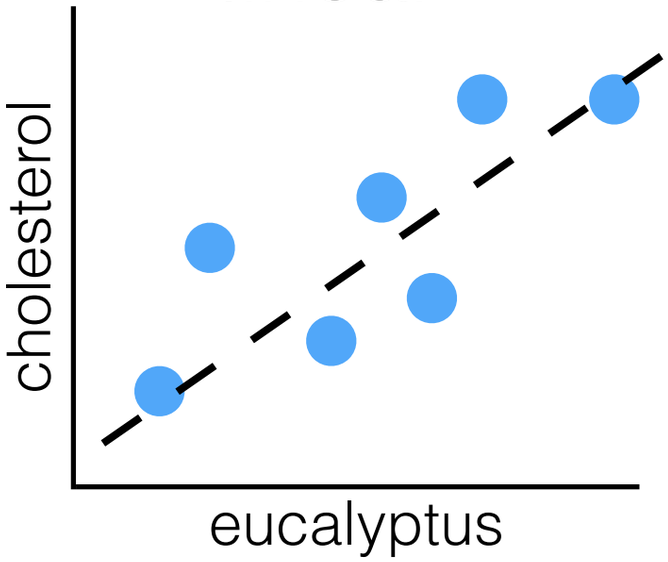


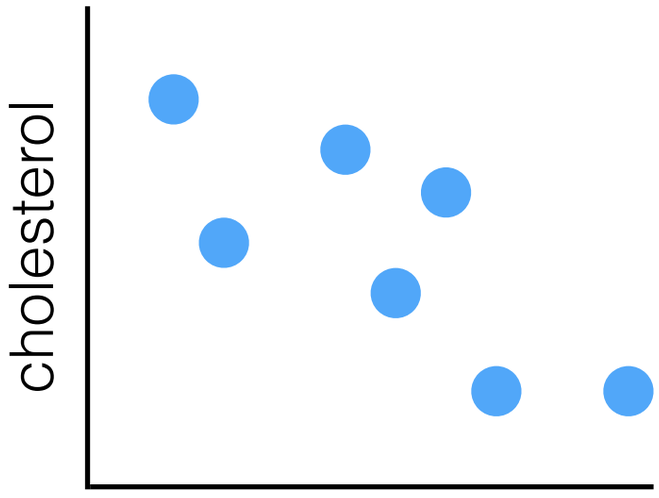


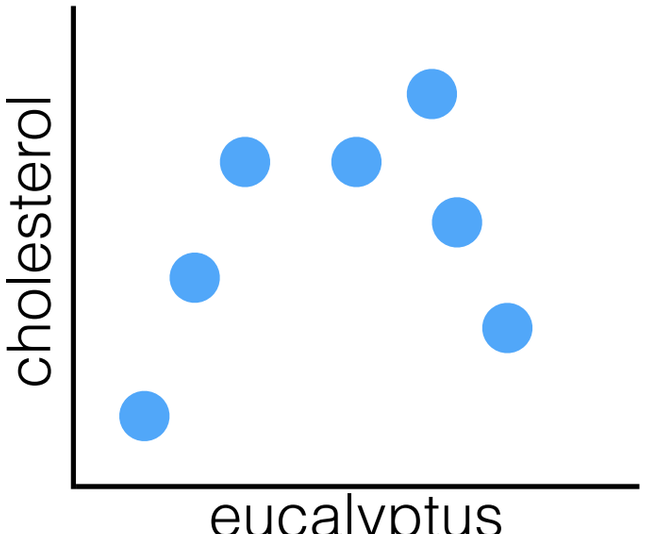
Regression

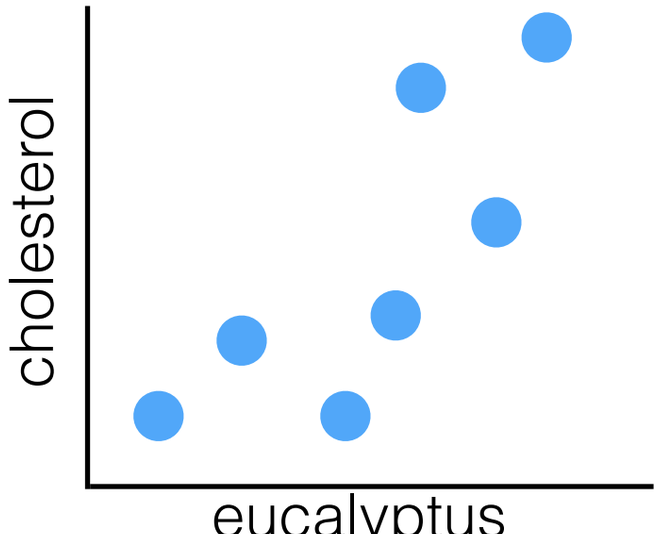
`cholesterol = f(mg eucalyptus oil)`

linear





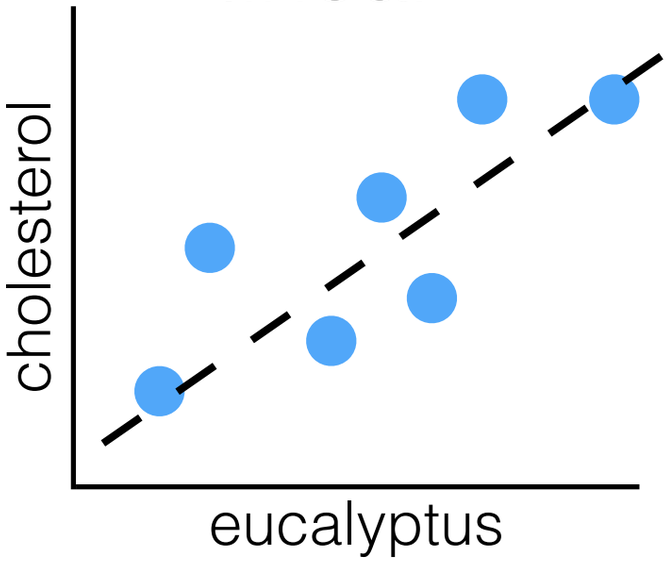




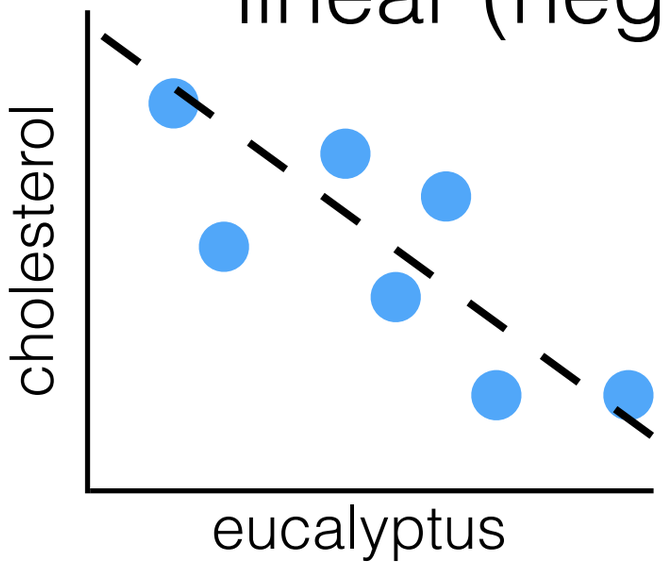
Regression

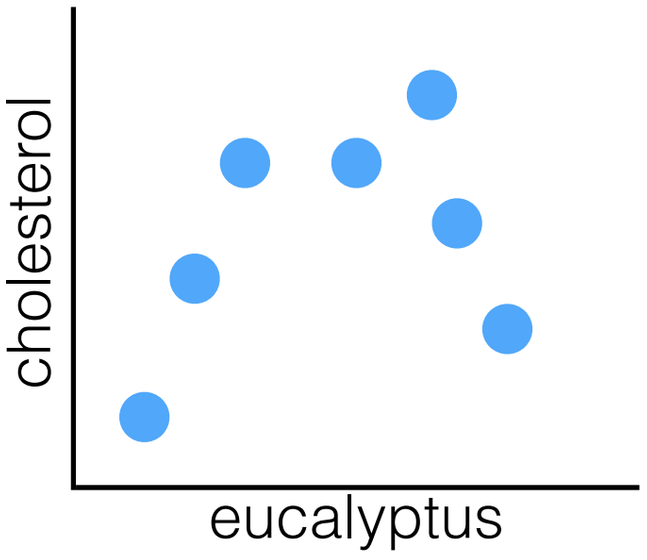
```
cholesterol = f(mg eucalyptus oil)
```

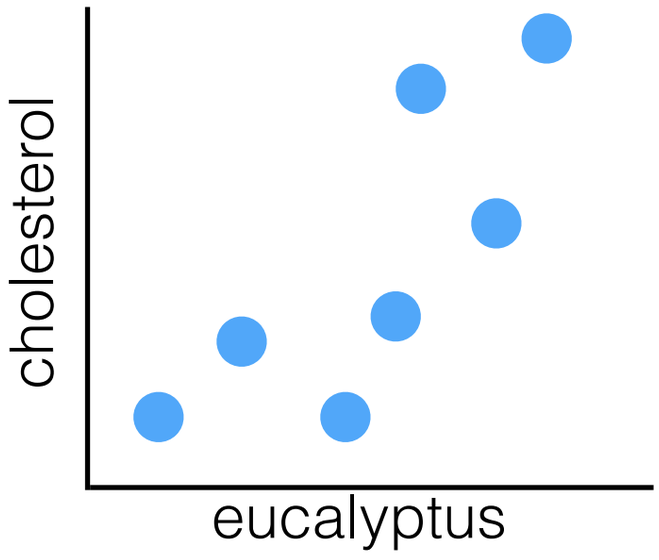
linear



linear (negative)



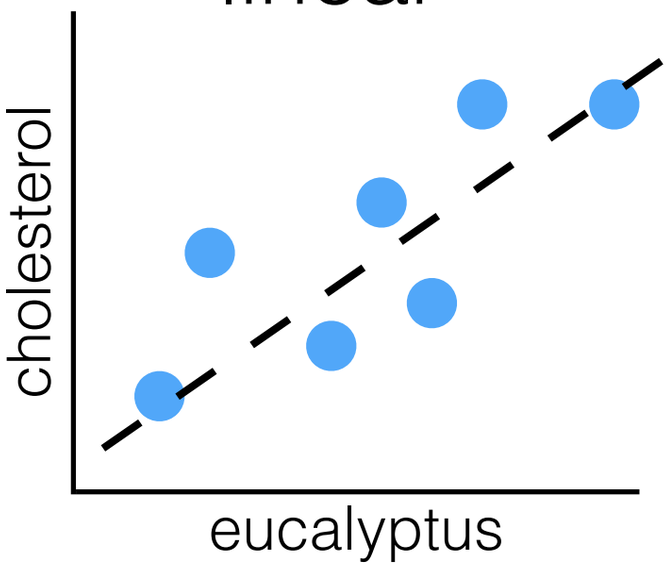




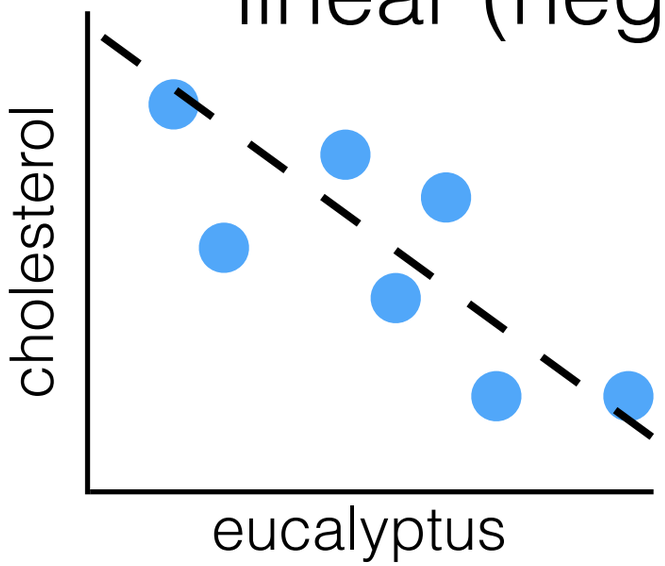
Regression

`cholesterol = f(mg eucalyptus oil)`

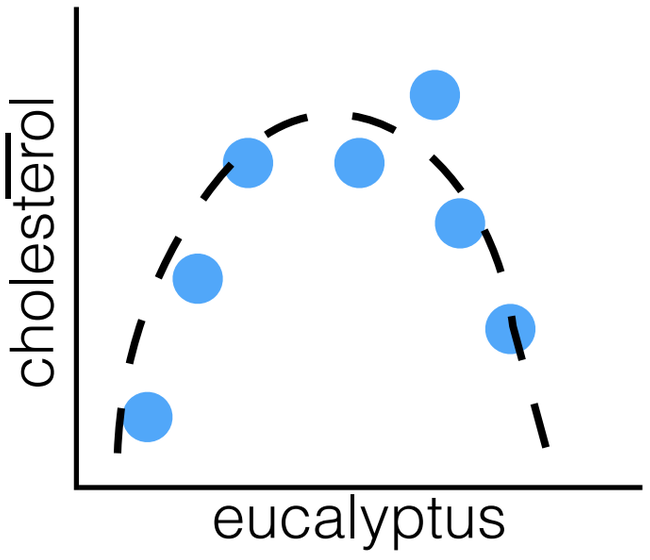
linear

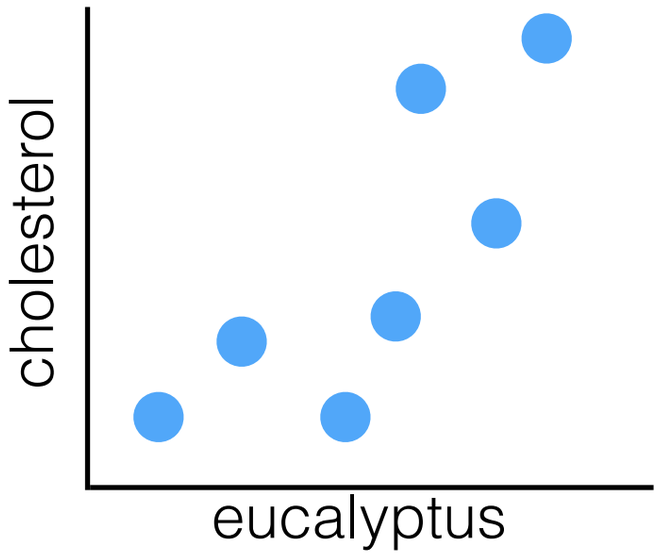


linear (negative)



polynomial

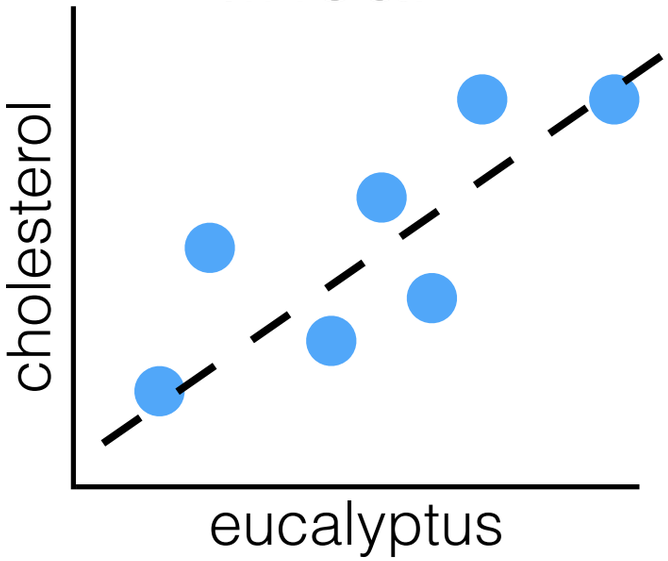




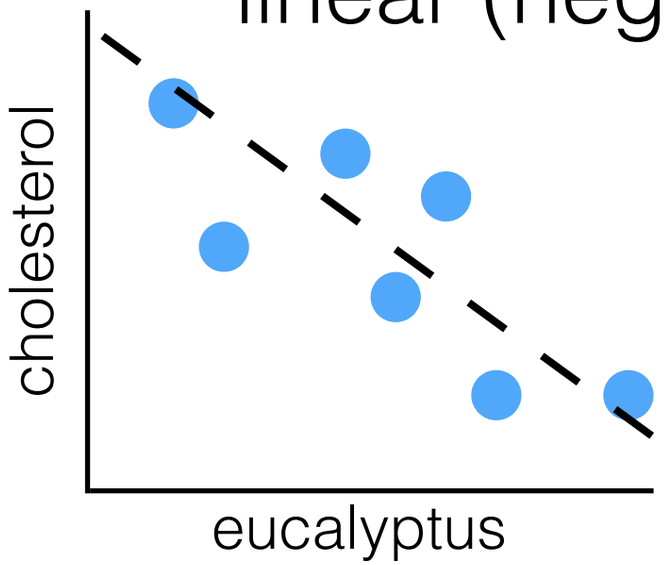
Regression

`cholesterol = f(mg eucalyptus oil)`

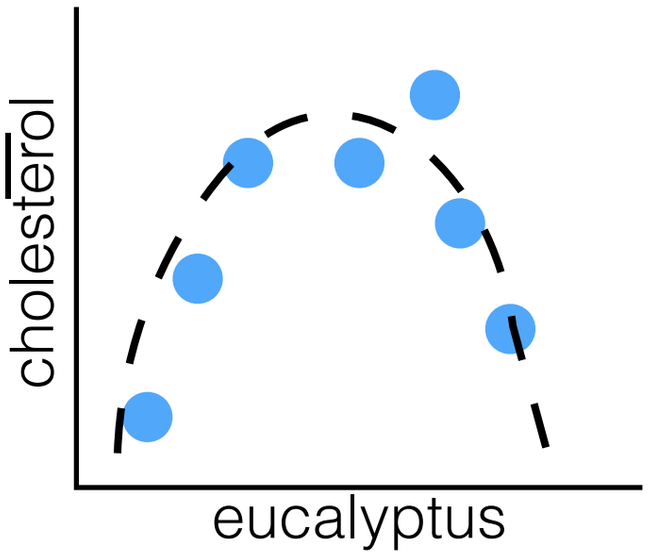
linear

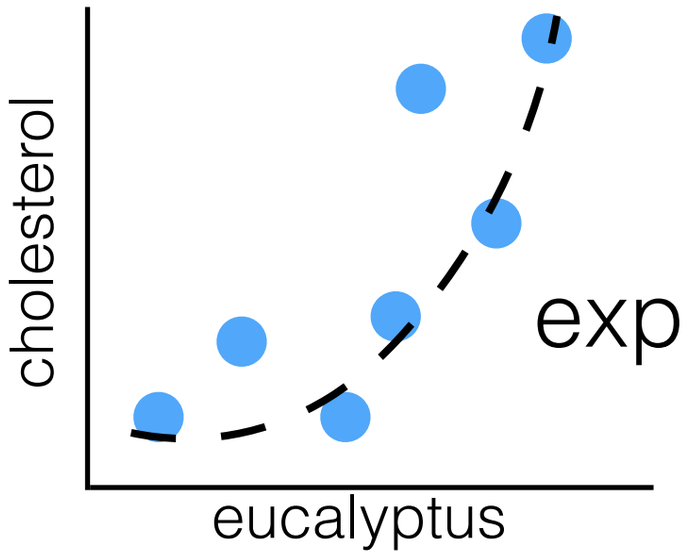


linear (negative)



polynomial





exponential

Linear Regression

$$y = mx + b + e$$

Linear Regression

$$y = mx + b + e$$

dependent variable (cholesterol)

Linear Regression

independent variable (mg eucalyptus oil)

$$y = mx + b + e$$

dependent variable (cholesterol)

Linear Regression

independent variable (mg eucalyptus oil)

$$y = mx + b + e$$

dependent variable (cholesterol)

slope (co-efficient) expected delta cholesterol for 1mg increase in eucalyptus oil

Linear Regression

independent variable (mg eucalyptus oil)

intercept expected cholesterol when eucalyptus = 0

$$y = mx + b + e$$

dependent variable (cholesterol)

slope (co-efficient) expected delta cholesterol for 1mg increase in eucalyptus oil

Linear Regression

independent variable (mg eucalyptus oil)

intercept expected cholesterol when eucalyptus = 0

$$y = mx + b + e$$

random (🙏) error

dependent variable (cholesterol)

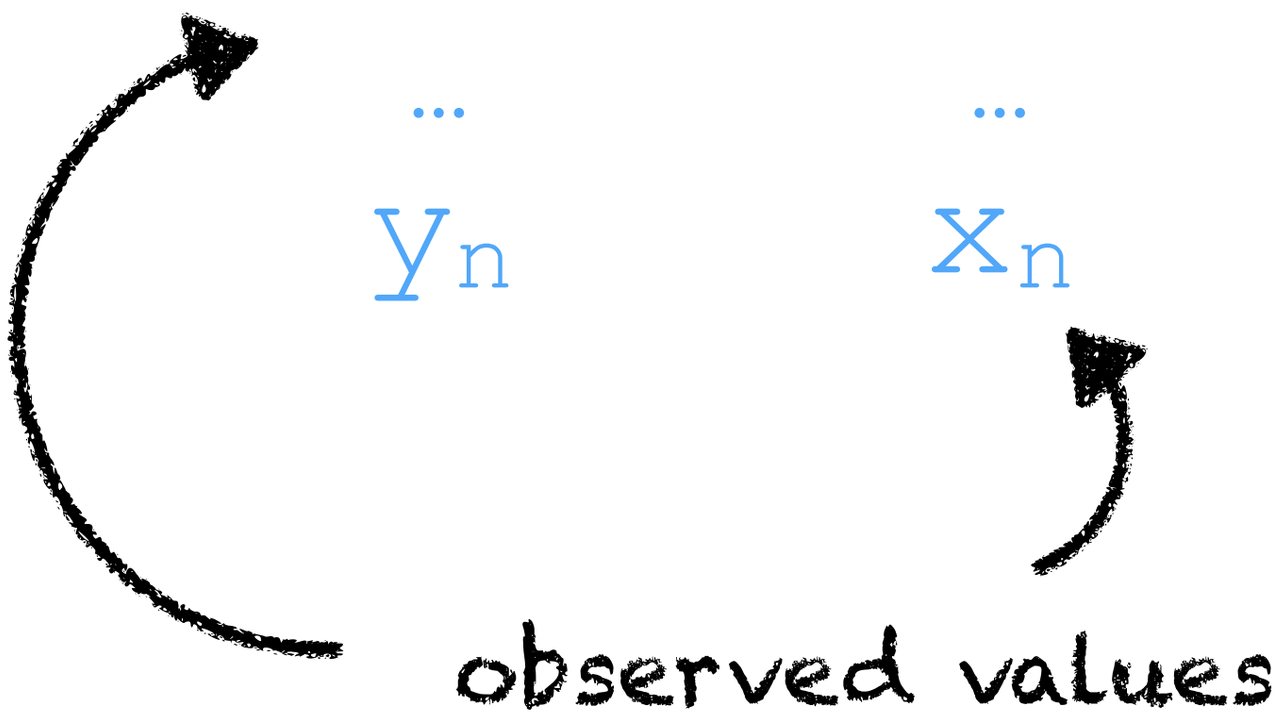
slope (co-efficient) expected delta cholesterol for 1mg increase in eucalyptus oil

Linear Regression

$$
\begin{array}{lllll}
y_1 & & x_1 & & e_1 \\
y_2 & & x_2 & & e_2 \\
y_3 & = m & x_3 & + b + & e_3 \\
\dots & & \dots & & \dots \\
y_n & & x_n & & e_n
\end{array}
$$

Linear Regression

$$
\begin{array}{lll}
y_1 & x_1 & e_1 \\
y_2 & x_2 & e_2 \\
y_3 = m\ x_3 + b + e_3 & & \\
\dots & \dots & \dots \\
y_n & x_n & e_n
\end{array}
$$



Linear Regression

y_1 x_1 e_1

y_2 x_2 e_2

$y_3 = m\ x_3 + b + e_3$

… … …

y_n x_n e_n

Linear Regression

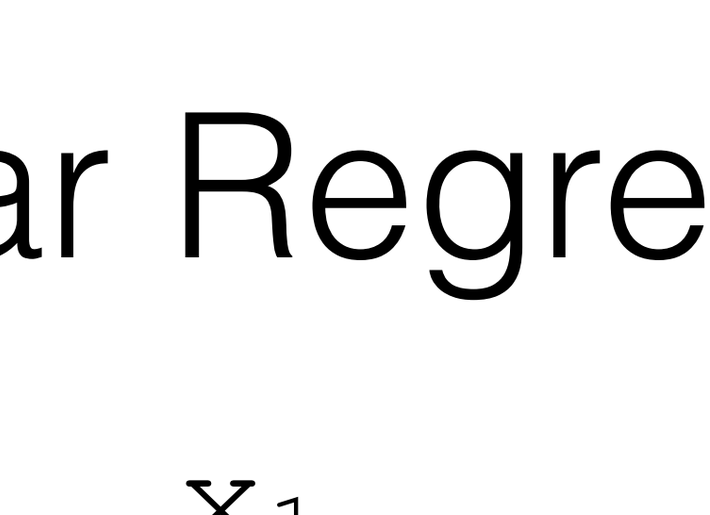
y_1 x_1 e_1

y_2 x_2 e_2

$y_3 = m\ x_3 + b + e_3$

… … …

y_n x_n e_n



assumed to be shared across the population

Linear Regression

y_1 x_1 e_1

y_2 x_2 e_2

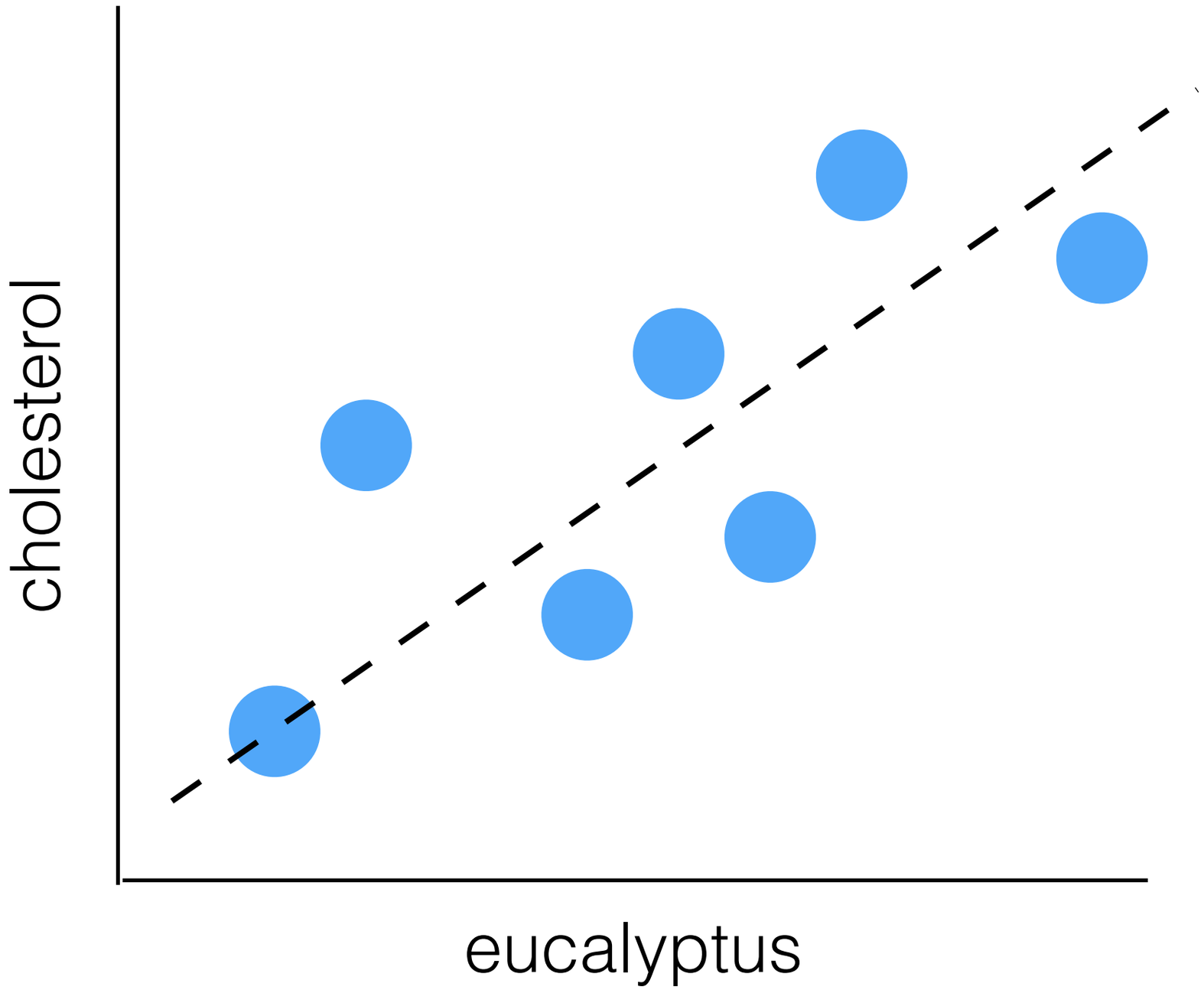
$y_3 = m\ x_3 + b + e_3$

… … …

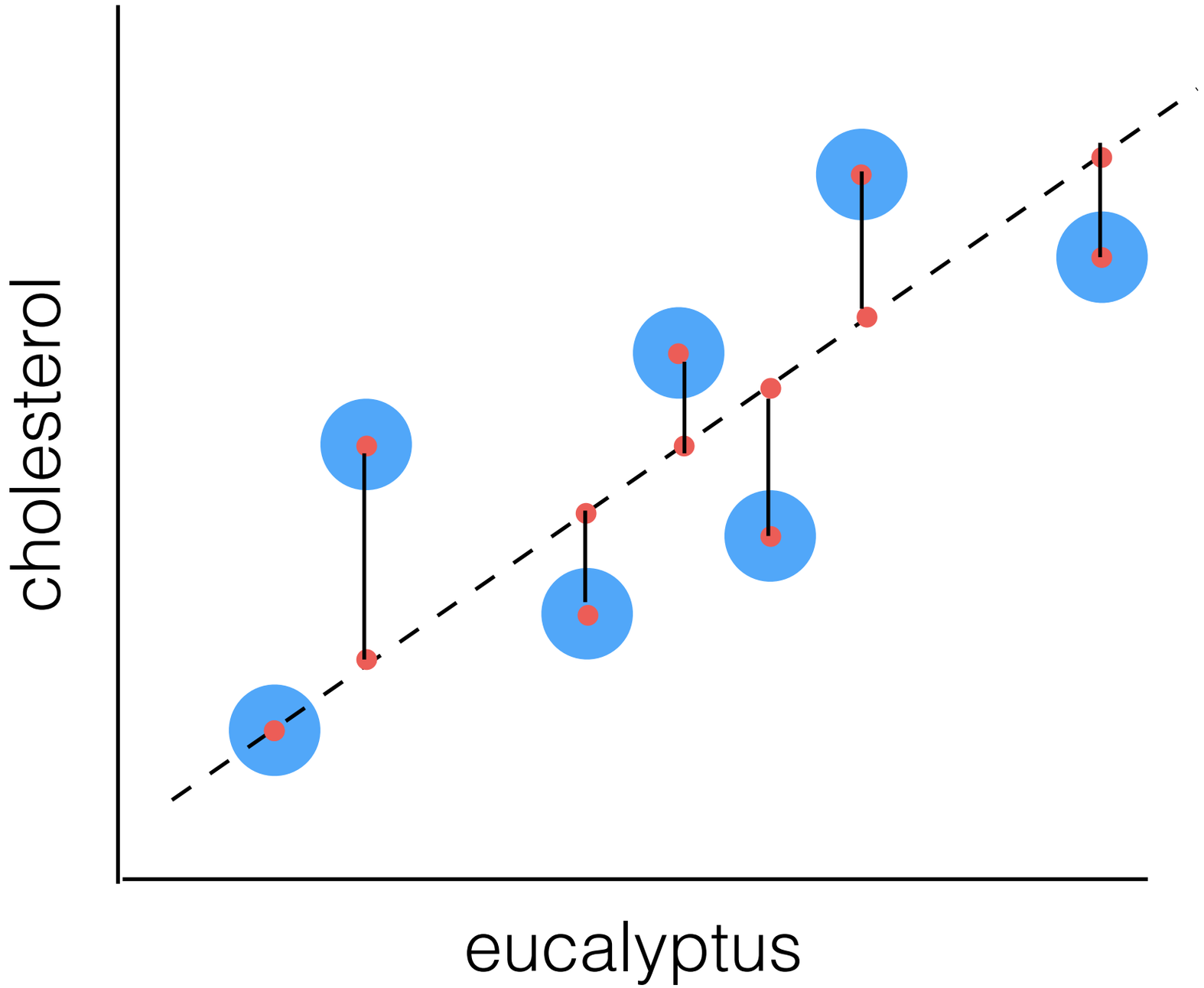
y_n x_n e_n

what we want to minimize

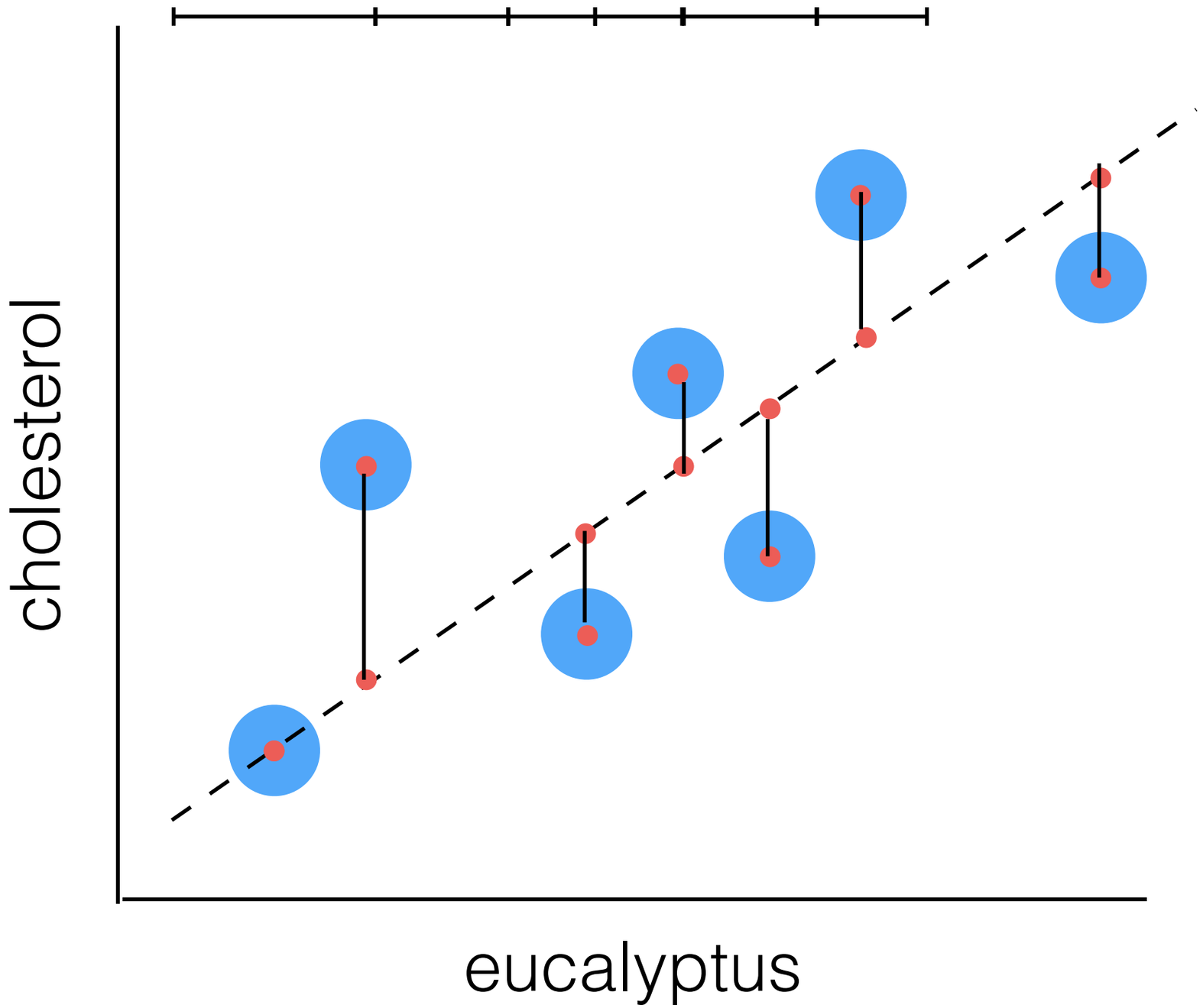
Linear Regression



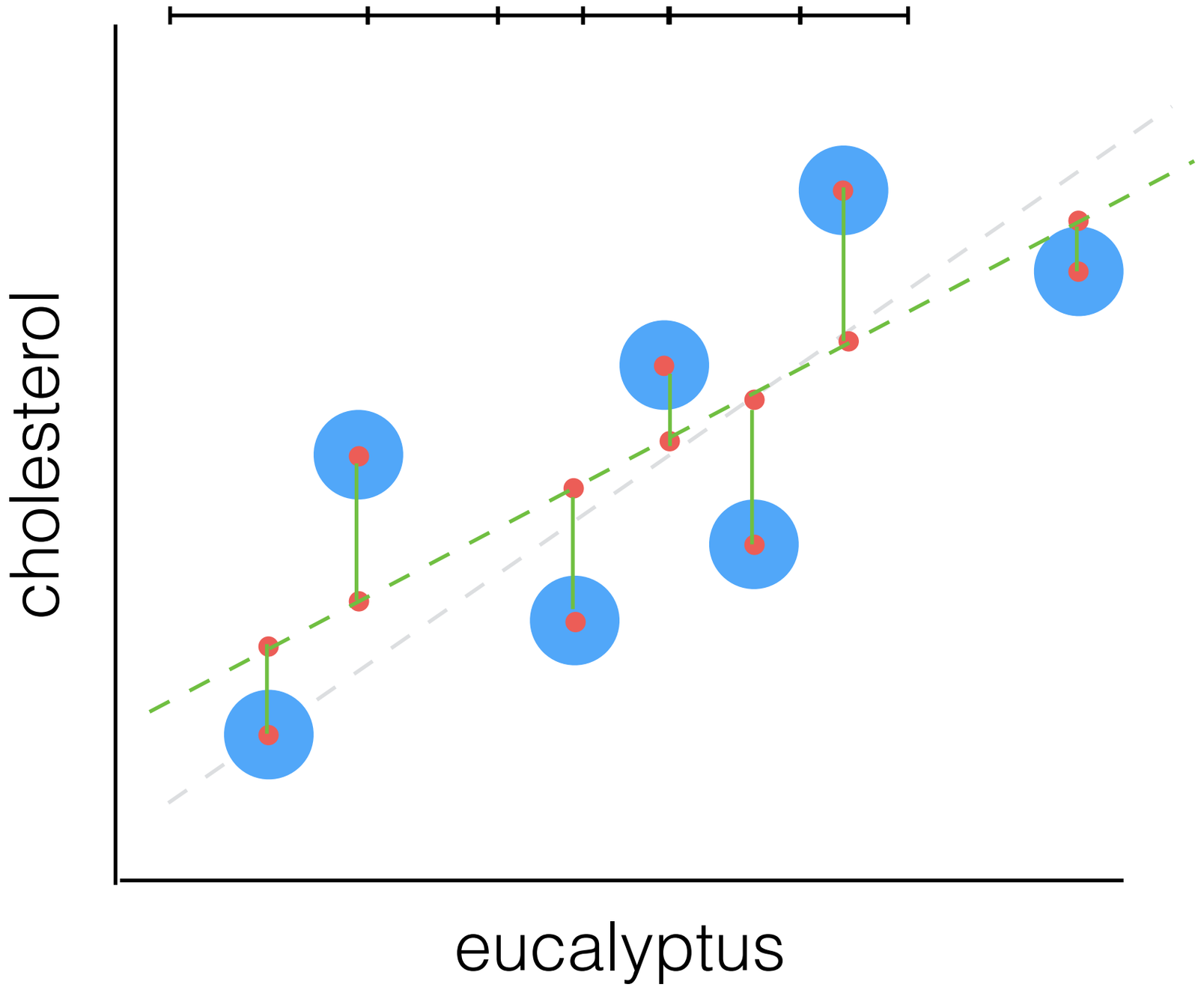
Linear Regression



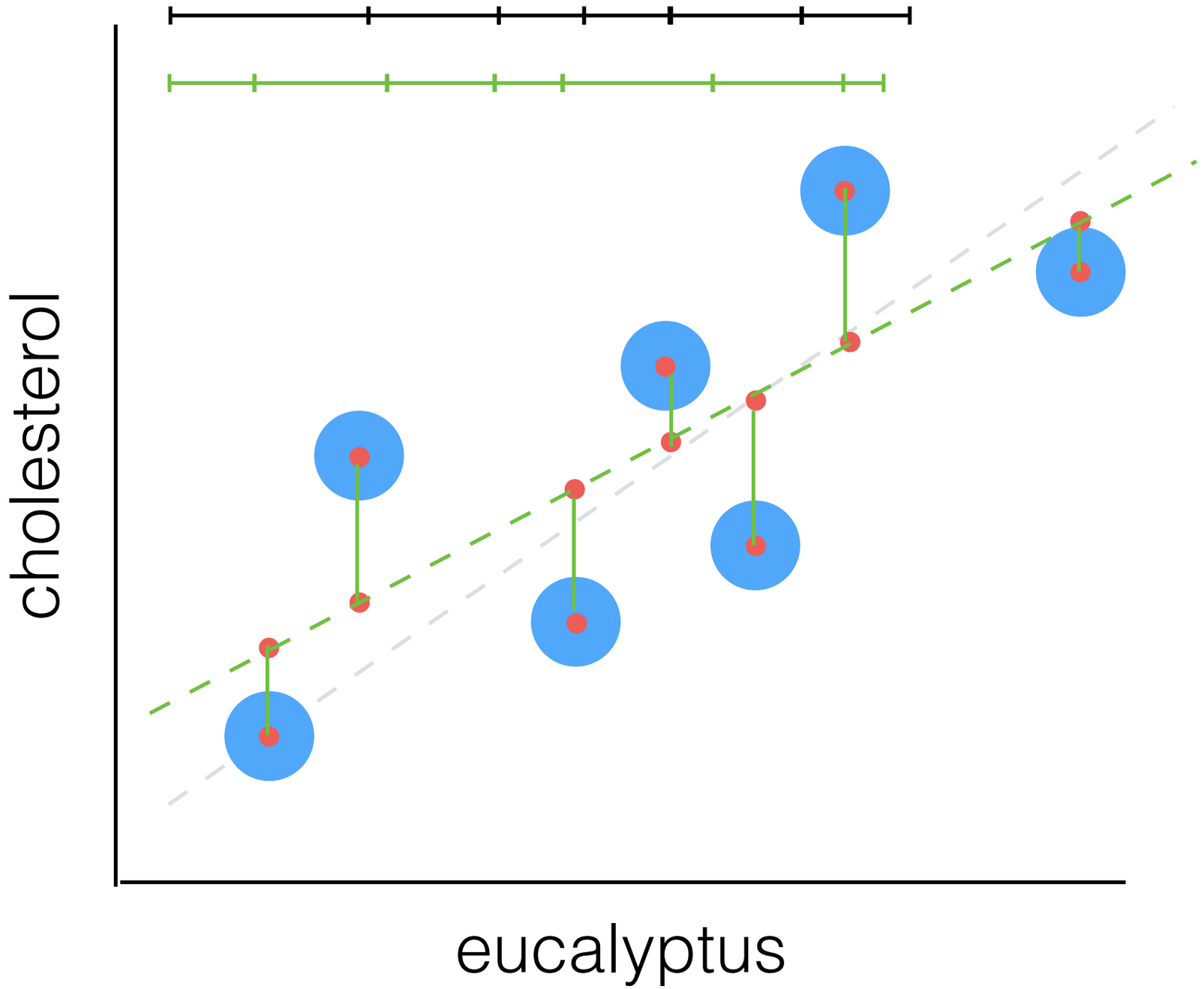
Linear Regression



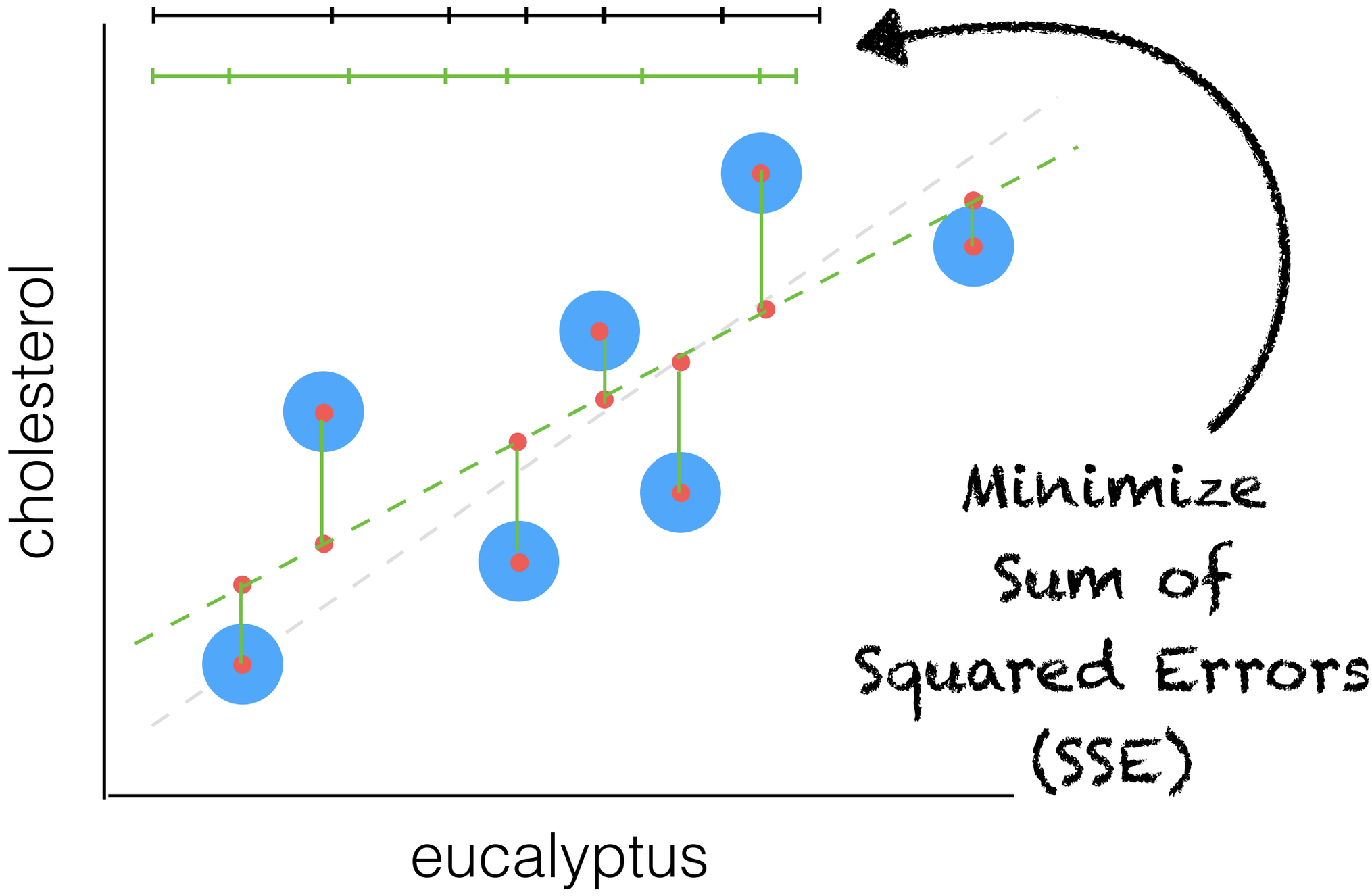
Linear Regression



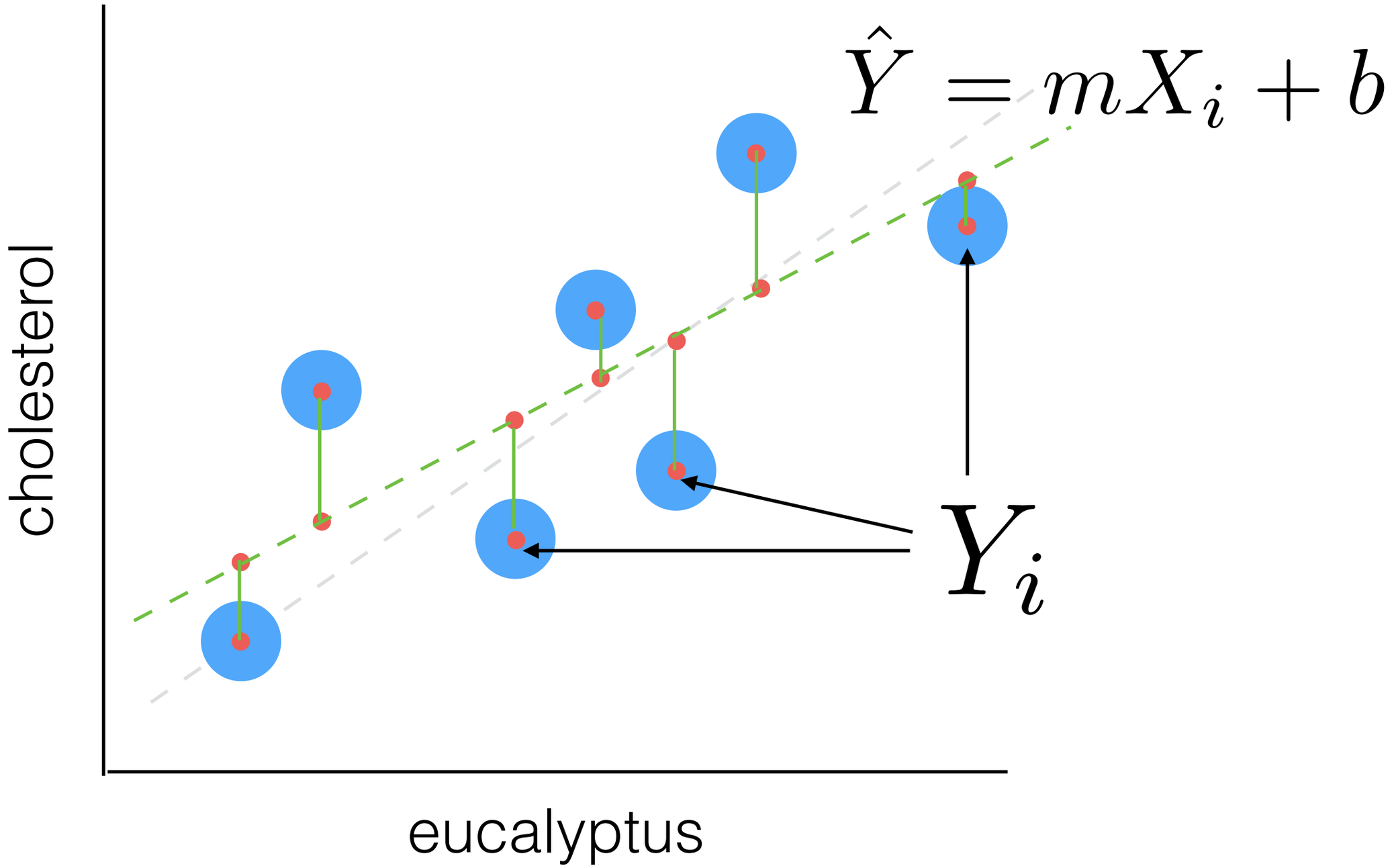
Linear Regression



Linear Regression



Linear Regression



Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

intercept at minimum

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

intercept at minimum

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

slope at minimum

Linear Regression

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

Linear Regression

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$2\left(nb + m\sum X_i - \sum Y_i\right) = 0$$

$$i=1 \qquad\qquad i=1$$

Linear Regression

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$2\left(nb + m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

Linear Regression

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$2\left(nb + m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Linear Regression

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$2\left(nb + m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

$$b = \bar{Y} - m\bar{X}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

intercept at minimum

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

slope at minimum

Linear Regression

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

$$\sum_{i=1}^{n} -2(Y_i X_i - bX_i - mX_i^2) = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

$$\sum_{i=1}^{n} -2(Y_i X_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

Linear Regression

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

$$\sum_{i=1}^{n} -2(Y_iX_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

$$\sum_{i=1}^{n} -2(Y_iX_i - \bar{Y}X_i + m\bar{X}X_i - mX_i^2) = 0$$

Linear Regression

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b - mX_i) = 0$$

$$\sum_{i=1}^{n} -2(Y_iX_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

$$\sum_{i=1}^{n} -2(Y_iX_i - \bar{Y}X_i + m\bar{X}X_i - mX_i^2) = 0$$

$$\sum_{i=1}^{n}(Y_iX_i - \bar{Y}X_i) - m\sum_{i=1}^{n} X_i^2 - \bar{X}X_i = 0$$

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

$$m = \frac{\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i)}{\sum_{i=1}^{n} X_i^2 - \bar{X} X_i}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

$$m = \frac{\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i)}{\sum_{i=1}^{n} X_i^2 - \bar{X} X_i}$$

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i \bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i \bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

ok ok, go go go