Discussion Question

Which expression best predicts federal spending in 2017?

```
A. fed_spending_2016*(g)
B. fed_spending_2016*(1+g)
C. fed_spending_2000*(17+g)
D. fed_spending_2016**(g)
```

Discussion Question

Which expression best predicts federal spending in 2017?

A. `fed_spending_2016*(g)`
B. `fed_spending_2016*(1+g)`
C. `fed_spending_2000*(17+g)`
D. `fed_spending_2016**(g)`

Actual prediction is 4,062,223,000,000 dollars.

Discussion Question

Assume you have run the following statements.

```
x = 3
y = '4'
z = '5.6'
```

Choose the expression that will be evaluated without an error.

```
A. x + y
B. x + int(y + z)
C. str(x) + int(y)
D. str(x) + z
```

Discussion Question

Assume you have run the following statements.

```
x = 3
y = '4'
z = '5.6'
```

What is the source of the error in the other expressions?

```
A. x + y
B. x + int(y + z)
C. str(x) + int(y)
D. str(x) + z
```

(Demo)

Discussion Question

Assume you have run the following statements.

```
x = make_array(2, 3, 4)

y = np.arange(2, 3, 4)

z = np.range(3)
```

Which lines will cause an error?

A. `x + y`

B. `x + z`

C. `x.item(0)+y.item(0)`

D. `x.item(1)+y.item(1)`

Leibniz Formula for Pi

$$\pi = 4 \cdot \left(1 - \tfrac{1}{3} + \tfrac{1}{5} - \tfrac{1}{7} + \tfrac{1}{9} - \tfrac{1}{11} + \ldots\right)$$

```
a = np.arange(1, 12, 4)
```

```
a
```

```
a + 2
```

```
4 * sum(1/a - 1/(a+2))
```

Which of the following is false?

A. `sum` is a function being applied to an array
B. `a` is an array
C. `a` is a range
D. The last line is equivalent to
         `4 * (1/a + -1/(a+2))`

(Demo)

Leibniz formula for π

Discussion Question

Assume you have run the following statements.

```
x = 3
y = '4'
z = '5.6'
```

Choose the expression that will be evaluated without an error.

```
A. x + y
B. x + int(y + z)
C. str(x) + int(y)
D. str(x) + z
```

Discussion Question

Assume you have run the following statements.

```
x = 3
y = '4'
z = '5.6'
```

What is the source of the error in the other expressions?

A. `x + y`

B. `x + int(y + z)`

C. `str(x) + int(y)`

D. `str(x) + z`

(Demo)

Discussion Question

Assume you have run the following statements.

```
x = make_array(2, 3, 4)

y = np.arange(2, 3, 4)

z = np.range(3)
```

Which lines will cause an error?

A. `x + y`

B. `x + z`

C. `x.item(0)+y.item(0)`

D. `x.item(1)+y.item(1)`

Discussion Question

Longitude	Latitude	City	Direction	Survivors
32	54.8	Smolensk	Advance	145000
33.2	54.9	Dorogobouge	Advance	140000
34.4	55.5	Chjat	Advance	127100
37.6	55.8	Moscou	Advance	100000
34.3	55.2	Wixma	Retreat	55000
32	54.6	Smolensk	Retreat	24000
30.4	54.4	Orscha	Retreat	20000
26.8	54.3	Moiodexno	Retreat	12000

How would you calculate the average of the numbers in the last column?

A. `sum(minard.select('Survivors'))/minard.num_rows`
B. `sum(minard.column('Survivors'))/minard.num_rows`
C. Both **A** and **B** work.
D. Neither **A** nor **B** work.

(Demo)

Discussion Question

To create a table of the highest-paid players in each position:

```
nba.sort(3, descending=True).sort(1, distinct=True)
```

Which code creates a table of the lowest-paid players in each position?

A. `nba.sort(3, descending=True).sort(1, distinct=False)`

B. `nba.sort(3, descending=False).sort(1, distinct=True)`

C. `nba.sort(3, descending=False).sort(1, distinct=False)`

D. `nba.sort(3, descending=True).sort(1, distinct=True)`

Discussion Question

The table **nba** has columns **PLAYER**, **POSITION**, **TEAM**, **SALARY**.

Order the snippets of code to calculate the total salary of all small forwards (**SF**)

A. `nba` `.column(3)` `.sum()` `.where(1,'SF')`

B. `nba` `.where(1,'SF')` `.sum()` `.column(3)`

C. `nba` `.column(3)` `.where(1,'SF')` `.sum()`

D. `nba` `.where(1,'SF')` `.column(3)` `.sum()`

Discussion Question

The table **nba** has columns **PLAYER**, **POSITION**, **TEAM**, **SALARY**.

```
nba.where(1,'SF').column(3).sum()/nba.where(1,'SF').num_rows
```

What does this code compute?

(Demo)

Discussion Question

The table **nba** has columns **PLAYER**, **POSITION**, **TEAM**, **SALARY**.

Create an array containing the names of all point guards (**PG**) who make more than $15M/year

Discussion Question

The table **nba** has columns **PLAYER**, **POSITION**, **TEAM**, **SALARY**.

Create an array containing the names of all point guards (**PG**) who make more than $15M/year

```
nba.where(1, 'PG').where(3, are.above(15)).column(0)
```

(Demo)

Discussion Question

The table **nba** has columns **PLAYER**, **POSITION**, **TEAM**, **SALARY**.

What is the output when we execute a cell containing these two lines of code?

```
nba.with_row(['Jazz Bear','Mascot','Utah Jazz',100])
nba.where('PLAYER', are.containing('Bear'))
```

A. A table with one row for Jazz Bear
B. An empty table with no rows
C. An error message

(Demo)

Practice

The table `menu` has a row for each item on a restaurant's menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that produces the same table without a row for `Cheeseburger.`

Practice

The table **menu** has a row for each item on a restaurant's menu. The columns are **Item** and **Price**, in that order. One of the menu items is **Cheeseburger.**

Write one line of code that produces the same table without a row for **Cheeseburger.**

```
menu.where('Item', are.not_equal_to('Cheeseburger'))
```

Practice

The table `menu` has a row for each item on a restaurant's menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Which line of code finds the number of items on the menu at this restaurant?

A. `menu.num_rows`
B. `menu.column(0).num_rows`
C. `menu.column(0).length`
D. `menu.column(1).size`
E. `More than one of the above`

Practice

The table `menu` has a row for each item on a restaurant's menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that evaluates to
a)   the name of a menu item that has the lowest possible price.

b)   **Challenge**: a table containing the name of **all** menu items that have the lowest possible price.

Practice

The table `menu` has a row for each item on a restaurant's menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that evaluates to
a)   the name of a menu item that has the lowest possible price.

```
menu.sort('Price').column(0).item(0)
```

b)   **Challenge**: a table containing the name of **all** menu items that have the lowest possible price.

Practice

The table **menu** has a row for each item on a restaurant's menu. The columns are **Item** and **Price**, in that order. One of the menu items is **Cheeseburger.**

Write one line of code that evaluates to
a) the name of a menu item that has the lowest possible price.

```
menu.sort('Price').column(0).item(0)
```

b) **Challenge**: a table containing the name of **all** menu items that have the lowest possible price.

```
menu.sort('Price').where('Price',
menu.sort('Price').column('Price).item(0)).select('Item')
```

Analyzing Census Data

Leads to the discovery of interesting features and trends in the population

(Demo)

Question

Suppose we execute this code:

```
aged = top.with_column("Age", 2017-top.column('Year'))
aged.group('Age').barh('Age')
```

What type of bar graph will be produced?

A. A bar for each movie.
    The length of the bar is the age of the movie.
B. A bar for each age.
    The length of the bar is the number of movies of that age.
C. A bar for each year.
    The length of the bar is the age of movies made that year.

(Demo)

Combining Bins

What should happen to our histogram if we combine the two bins [20, 40) and [40, 60) into one large bin [20, 60)?

A. The new histogram should have four bars of equal width.

B. The height of the bar for bin [20, 60) should be the sum of the heights of the bars for bins [20, 40) and [40, 60).

C. The area of the bar for bin [20, 60) should be the sum of the areas of the bars for bins [20, 40) and [40, 60).

D. More than one of the above.
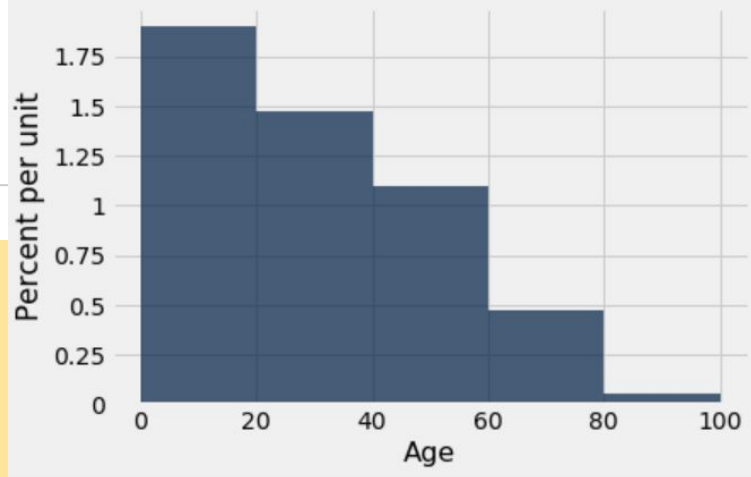
bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

(Demo)

Combining Bins

Suppose we combine the two bins [20, 40) and [40, 60) into one large bin [20, 60). What is the density of the new bin?

A. The new bin has about twice as many movies as each original bin, so it is about twice as dense as each original bin.

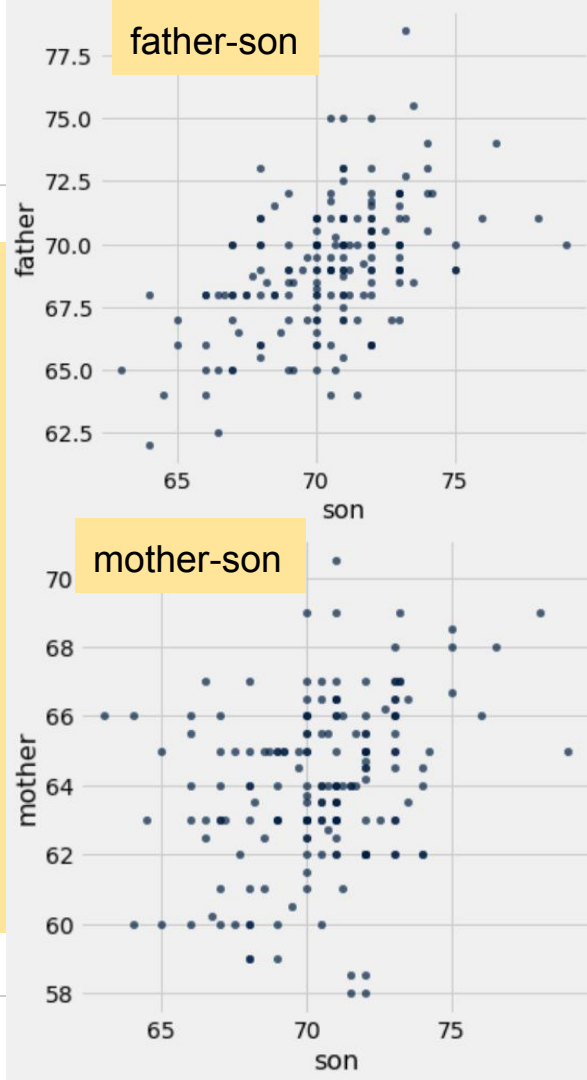B. The new bin is about twice as big as each original bin, so it is about half as dense as each original bin.

C. The new bin has about twice as many movies and is twice as big as each original bin, so it is about the same density as each original bin.
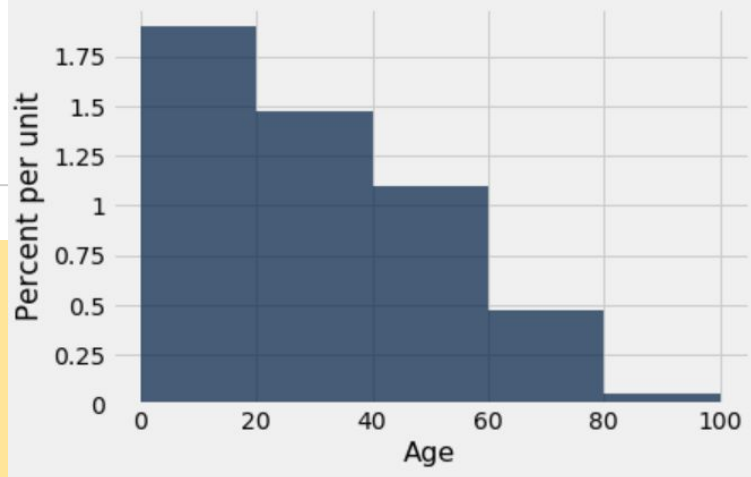
bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

Father or Mother?

Is a son's height more influenced by his father's height or his mother's height?

A. Father, because difference between father and son height is smaller than difference between mother and son height.
B. Mother, because there is more variability in mother's heights than father's heights.
C. Father, because the points on the father-son plot more strongly resemble a line than those on the mother-son plot.
D. Father, because the points on the father-son plot form a steeper curve than the those on the mother-son plot.

father-son

mother-son

Father or Mother?

Is a son's height more influenced by his father's height or his mother's height?

A. Father, because difference between father and son height is smaller than difference between mother and son height.
B. Mother, because there is more variability in mother's heights than father's heights.
C. Father, because the points on the father-son plot more strongly resemble a line than those on the mother-son plot.
D. Father, because the points on the father-son plot form a steeper curve than the those on the mother-son plot.
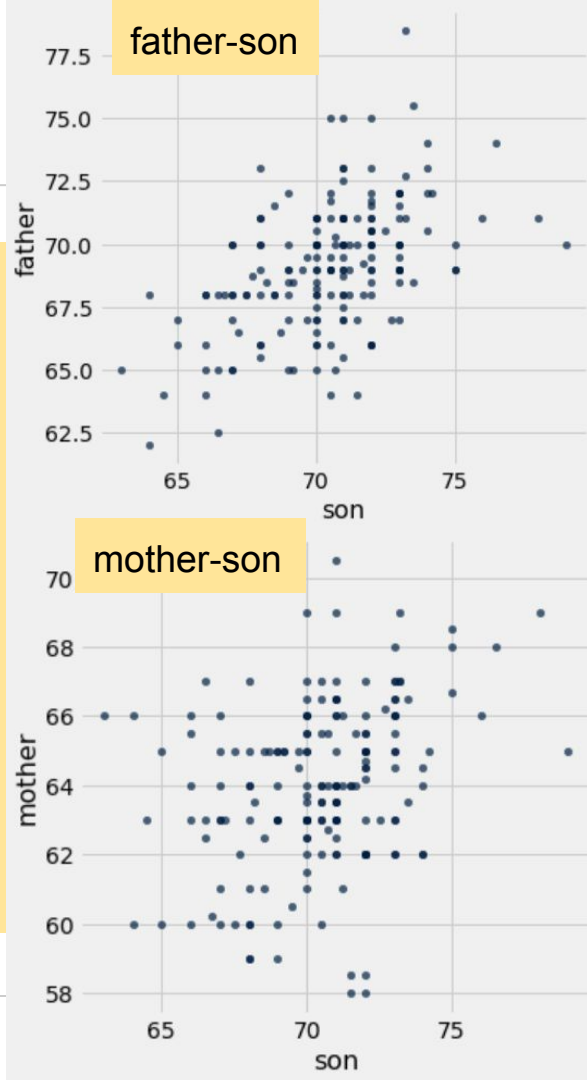
Discussion Question

```
def f(s):
    return np.round(s / sum(s) * 100, 2)
```

What does this function do?   What kind of input does it take?   What output will it give?

A. Computes average   array of numbers   one number
B. Computes average   one number   one number
C. Computes percents   array of numbers   array of numbers
D. Computes percents   array of numbers   one number
E. None of the above

What's a reasonable name?   (Demo)

Discussion Question

Input:

Title
Avatar
Star Wars: Episode I - The Phantom Menace
Star Wars
Star Wars: Episode III - Revenge of the Sith
Star Wars: Episode II - Attack of the Clones
Return of the Jedi

Output:

```
array(['Avatar: length is 6',
       'Star Wars: Episode I - The Phantom Menace: length is 41',
       'Star Wars: length is 9',
       'Star Wars: Episode III - Revenge of the Sith: length is 44',
       'Star Wars: Episode II – Attack of the Clones: length is 44',
       'Return of the Jedi: length is 18',
```

If the name of the table is **top** and the name of our function is **str_len**, how do we find the length of each movie title?

A.   `top.apply(str_len(string), 'Title')`
B.   `top.apply(str_len( ), 'Title')`
C.   `top.apply(str_len, 'Title')`
D.   `Title.apply(str_len, 'top')`
E.   `Title.apply(str_len(), 'top')`

(Demo)

Extra Practice

```python
def my_func():
    x = 10
    print("Value inside function:",x)

x = 20
my_func()
print("Value outside function:",x)
```

What is the value of x after this code is executed?

A. 10
B. 20
C. This code will not run because of an error.
D. None, x has no value since it is defined inside a function.

Extra Practice

Input:

Title	Studio	Gross	Gross (Adjusted)	Year
Avatar	Fox	760507625	846120800	2009
Star Wars: Episode I - The Phantom Menace	Fox	474544677	785715000	1999
Star Wars	Fox	460998007	1549640500	1977

Output:

Title	Studio	Gross	Gross (Adjusted)	Year	Difference
Avatar	Fox	760507625	846120800	2009	85613175
Star Wars: Episode I - The Phantom Menace	Fox	474544677	785715000	1999	311170323
Star Wars	Fox	460998007	1549640500	1977	1088642493
Star Wars: Episode III - Revenge of the Sith	Fox	380270577	516123900	2005	135853323

Discuss how you would create the output table.

Way 1) Defining a function and using `apply`.

Way 2) Without defining a function or using `apply`.

(Demo)

Discussion Question

- A *starter* for a team is the player with the highest salary on that team in that position.
- The name of the table shown is *starters*.

TEAM	POSITION	SALARY max
Atlanta Hawks	C	12
Atlanta Hawks	PF	18.6717
Atlanta Hawks	PG	8
Atlanta Hawks	SF	4
Atlanta Hawks	SG	5.74648
Boston Celtics	C	2.61698
Boston Celtics	PF	5
Boston Celtics	PG	7.73034
Boston Celtics	SF	6.79612
Boston Celtics	SG	3.42551

Which will rank the teams in order of their highest-paid starter?

A. `starters.group('TEAM', max).sort(1, descending = True)`

B. `starters.drop('POSITION').group('TEAM', max).sort(1, descending = True)`

C. `starters.select('TEAM', 'SALARY').group('TEAM', max).sort(1, descending=True)`

D. `starters.select('TEAM', 'SALARY max').group('TEAM', max).sort(1, descending = True)`

E. More than one of the above

(Demo)

Discussion Question

```
d = np.arange(6) + 1
```

What happens when we evaluate the following 2 expressions?

- `np.random.choice(d, 1000) + np.random.choice(d, 1000)`
- `2 * np.random.choice(d, 1000)`

A. Gives the same result; Describing the same process
B. Gives the same result; Describing different processes
C. Gives different results; Describing the same process
D. Gives different results; Describing different processes
E. None of the above

Discussion Question

```
a = True
b = True
not(((not a) and b) or ((not b) or a) )
```

What does the expression evaluate to?

    A. True
    B. False
    C. I'm lost

(Demo)

Discussion Question

```
def func(a, b):
    if (a + b > 4 and b > 0):
        return 'print 1'
    elif (a*b >= 4 or b < 0):
        return 'print 2'
    else:
        return 'print 3'
```

What is returned when *func(2, 2)* is called?

A. print 1
B. print 2
C. print 3
D. More than one print statement
E. Error

(Demo)

Discussion Question

```python
def func(a, b):
    if (a + b > 4 and b > 0):
        return 'print 1'
    elif (a*b >= 4 or b < 0):
        return 'print 2'
    else:
        return 'print 3'
```

What is returned when *func(2, 2)* is called?

A. print 1
B. print 2
C. print 3
D. More than one print statement
E. Error

(Demo)

Discussion Question

I have three cards: red, blue and green.

What is the chance that I chose a card at random and it is green, then I chose another random card and it is red?

A: 1/6
B: 1/3
C: 2/3
D: None of the above

Discussion Question

I have three cards: red, blue and green. I pick two cards at the same time. What is the probability that I pick one red and one green?

A. ⅙
B. ⅓
C. ⅚
D. None of the above

Discussion: At Least One Head

I have a fair coin.

Find the probability of at least one head in 3 tosses.

Discussion: At Least One Head

I have a fair coin.

Find the probability of at least one head in 3 tosses.

o Any outcome *except* TTT

o P(TTT) = (½) x (½) x (½) = ⅛

o P(at least one head) = 1 - P(TTT) = ⅞ = 87.5%

Discussion Question

Every time I call my Grandma, the probability that she answers her phone is ⅓. If I call my Grandma **two** times today, what is the chance that I will talk to her?

A.   1/3
B.   2/3
C.   1/2
D.   1
E.   None of the above

Discussion Question

Every time I call my Grandma, the probability that she answers her phone is ⅓. If I call my Grandma **three** times today, what is the chance that I will talk to her?

A. 1/3
B. 2/3
C. 1/2
D. 1
E. None of the above

Discussion question

If you saw these serial numbers, what would be your estimate of N?

170 271   285 290   48

235 24 90   291   19

A: 291

B: 350

C: 470

D: Not enough information

Question

	purple	red	orange	green	yellow	blue
Distribution 1	0	1/5	1/5	1/5	1/5	1/5
Distribution 2	1	0	0	0	0	0
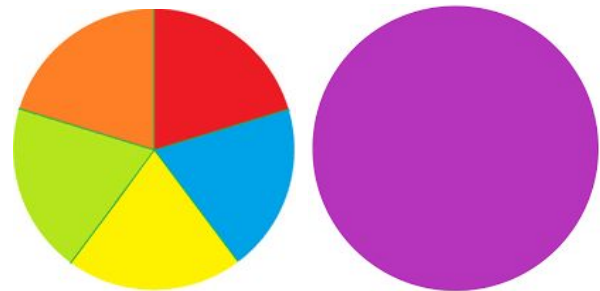
What is the total variation distance between the two distributions above?

A: 1/5   D: 7/5

B: 3/5   E: 2

C: 1

The Bootstrap

- Need another random sample that looks like the population

- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population

- So we sample at random from the original sample!

- A technique for simulating repeated random sampling

Question

What should be the size of your new sample?

A: 25% of the original sample

B: 50% of the original sample

C: 75% of the original sample

D: 100% of the original sample

E: Depends on the problem

Question

How should we obtain this new sample?

A: WITH replacement

B: WITHOUT replacement

C: Depends on the problem

Can You Use a C.I. Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:
- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

A: True
B: False
C: I'm lost

(Demo)

Discussion Question

Which are `True`, when `s = [1, 7, 3, 9, 5]`?

1. `percentile(10, s) == 0`
2. `percentile(39, s) == percentile(40, s)`
3. `percentile(40, s) == percentile(41, s)`
4. `percentile(50, s) == 5`

A: 1, 2
B: 2, 3
C: 2, 4
D: 1, 4
E: None of the above

(Demo)

Question
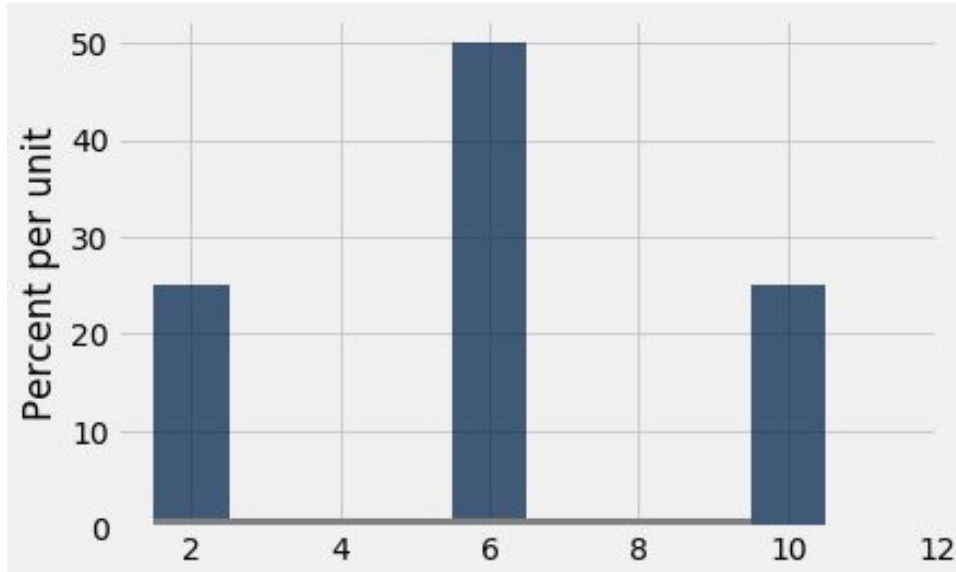
How can you calculate the mean?
- A. (2 + 6 + 10)/3
- B. (2 + 6 + 10)/4
- C. (2 + 6 + 6 + 10)/3
- D. (2 + 6 + 6 + 10)/4
- E. None of the above

(Demo)

Discussion Question

Three gorilla siblings are 2, 3, and 4 years old.

What is the standard deviation of gorilla ages?

- A. 1
- B. 2
- C. sqrt(2)
- D. sqrt(⅔)
- E. None of the above.

SD = root mean square of deviations from average

Which Has Larger SD?
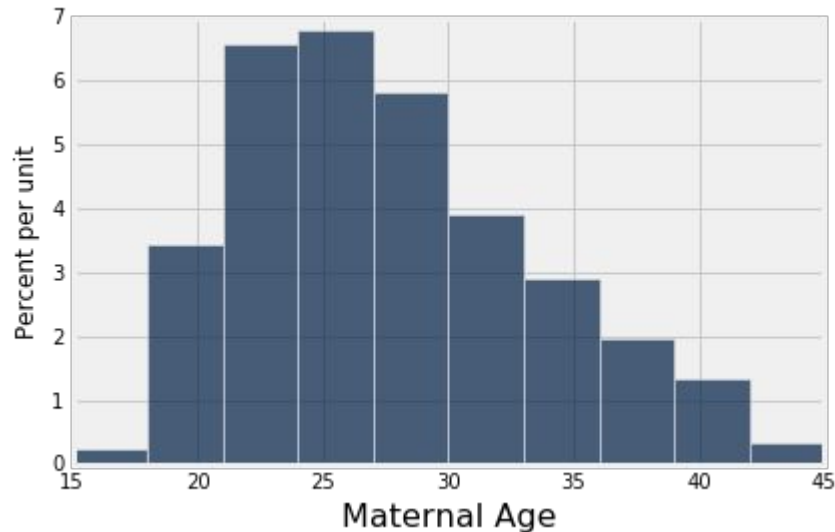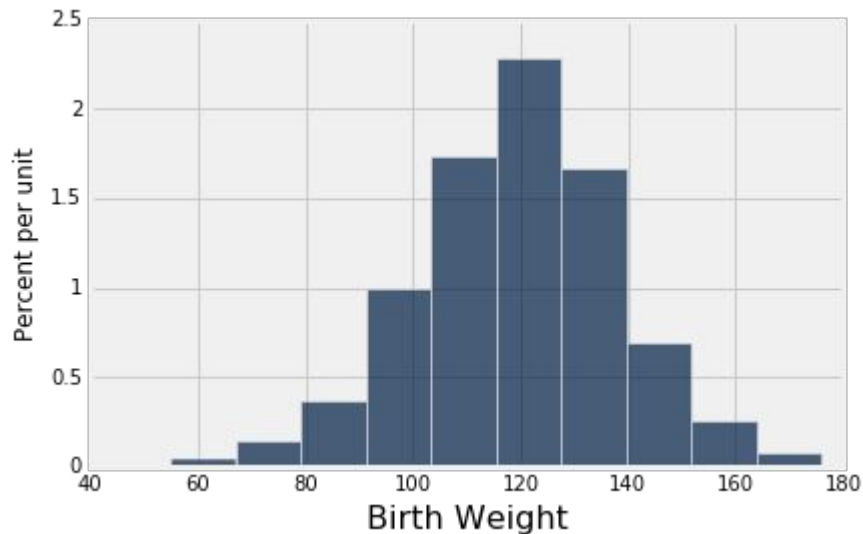
A. Birth Weight (Left)
B. Maternal Age (Right)
C. Cannot tell from the histograms

(Demo)

Question
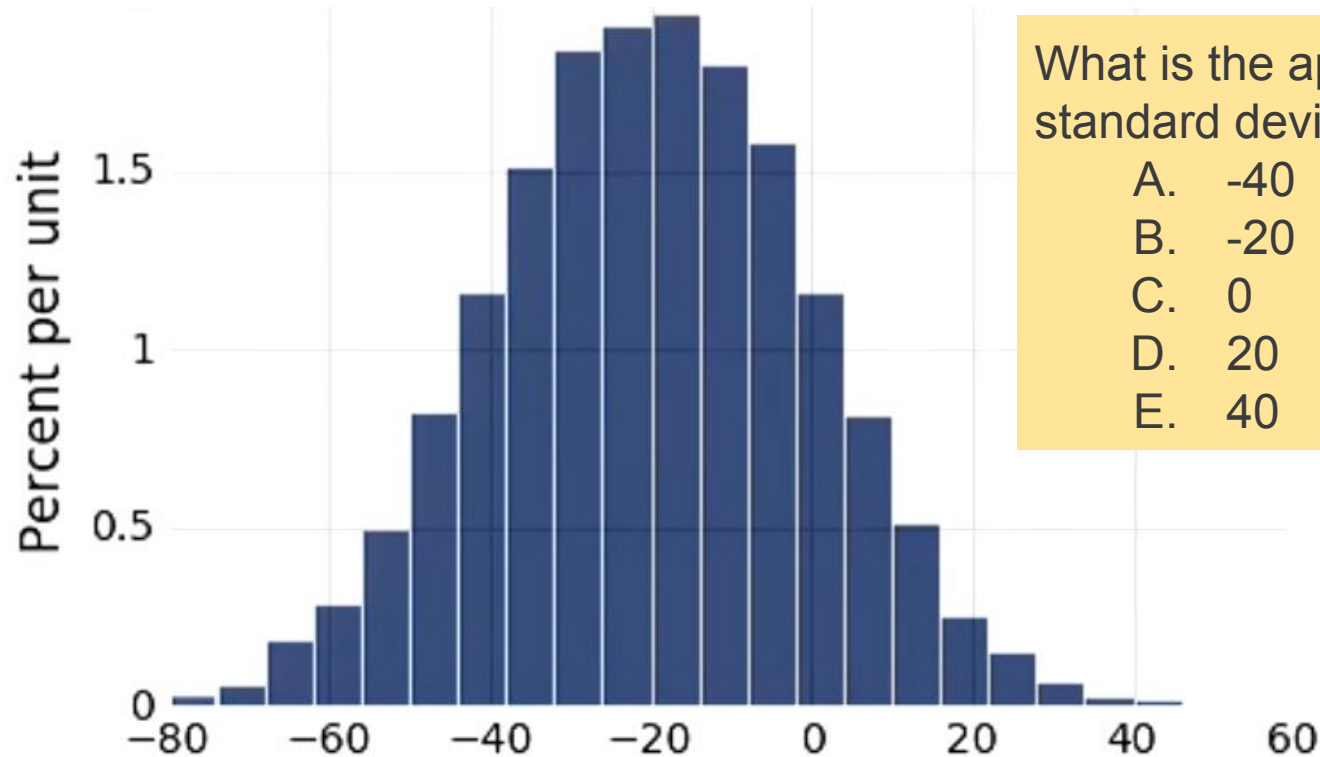
What is the approximate standard deviation?
- A. -40
- B. -20
- C. 0
- D. 20
- E. 40

Discussion Question 1

Sampling from the flight delay distribution.
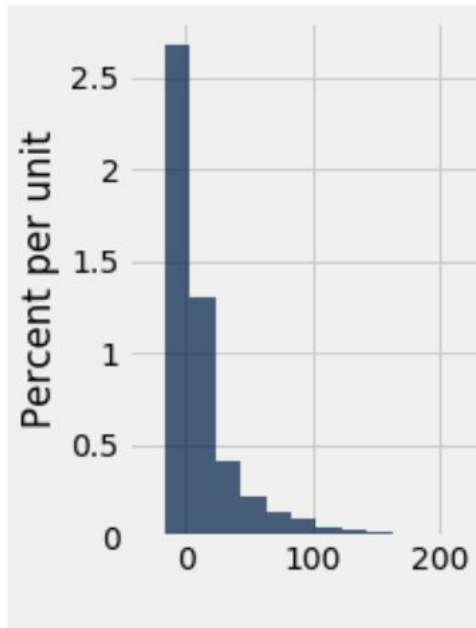
If you repeatedly compute the mean from a sample size of **1,** what will be the shape of the probability histogram?

A.   Impossible to predict

B.   Bell - shaped

C.   Resembles the original histogram

(Demo)

Discussion Question 2

Sampling from the flight delay distribution.
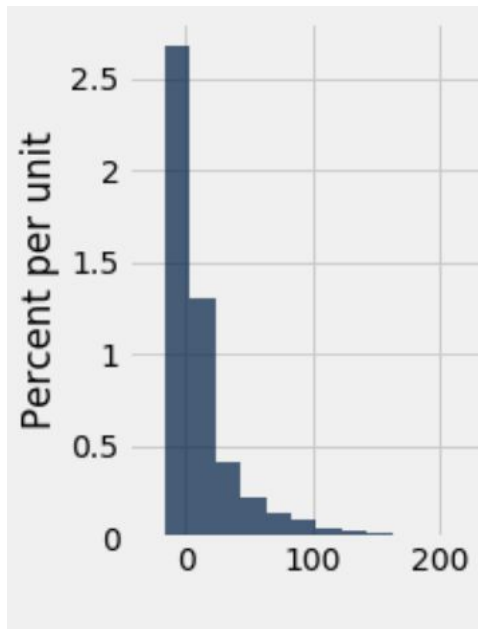
If you repeatedly compute the mean from a sample size of **1,** what will be the SD of the probability histogram?

A. 1
B. 0
C. Impossible to predict
D. Same as original SD

Discussion Question 3

Population: Incomes with mean $10,000 and SD $20,000

Sample: 100 chosen uniformly at random with replacement

What's the chance that the sample average is above $14,000?

A.   2.5%

B.   37%

C.   75%

D.   I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

Discussion Question 3

Population: Incomes with mean $10,000 and SD $20,000

Sample: 100 chosen uniformly at random with replacement

What's the chance that the sample average is above $14,000?

A. 2.5%

B. 37%

C. 75%

D. I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Percent in Range	All Distributions	Normal Distribution
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

Discussion Question 5

You want to estimate the height of the tallest person on campus. You sample 100 people at random and compute a 99.9999% confidence interval using the bootstrap. Its upper bound is 6'4".

A 6'5" person walks by! What might have gone wrong?

A. Standard deviation of the population is too large to estimate
B. Sample size is too small for 99.9999% confidence interval
C. Height of tallest person is difficult to estimate with bootstrap
D. Empirical distribution of height of tallest person is not bell-shaped
E. More than one of the above

Discussion Question 6

You want to estimate the average compensation for SF workers by randomly sampling workers.

How many workers should you sample at random in order to get a 95% confidence interval with a width of $10,000 or less?

(Demo)

Discussion Question

A course has a midterm (average 70; standard deviation 10)
and a really hard final (average 50; standard deviation 12)

If the scatter diagram comparing midterm & final scores for students has a typical oval shape with correlation 0.75, then...

What is the average final score for students who scored 90 on the midterm?
A. 76
B. 90
C. 68
D. 82
E. 67.5

Review discussion question

Given a table with 3 columns:

Week

Beer: number of bottles of beer consumed in San Diego that week

Weddings: the number of weddings in San Diego that week

Let r be the *correlation* between beer and weddings.

Which statement is True?
A. A possible value for r is 1.5.
B. If r is between -0.05 and 0.05, there is little association between beer consumption and weddings.
C. If r = 1, then an increase in weddings causes an increase in beer consumption.
D. More than one of the above.
E. None of the above.

Discussion Question
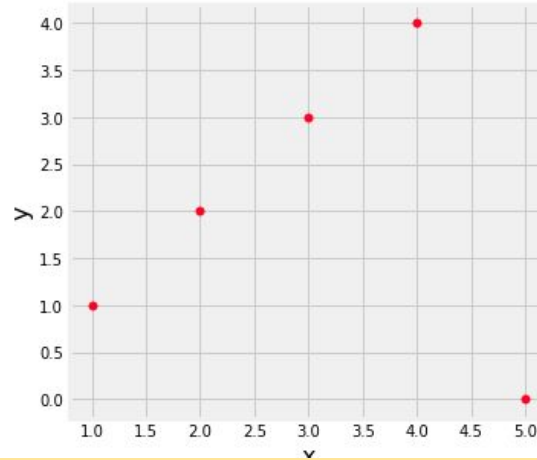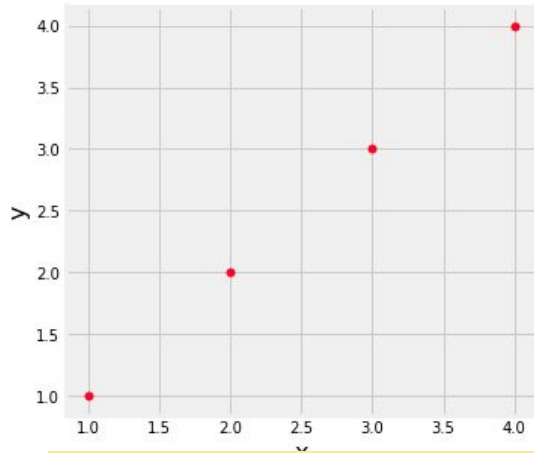
What are the correlations for these scatter plots? Note one outlier on the right plot.
- A. r = 1,   r ~ 0.9
- B. r = 1,   r ~ 0.5
- C. r = 1,   r = 0
- D. r = 0,   r ~ 0.5
- E. None of the above

Discussion Question

A course has a midterm (average 70; standard deviation 10)
and a really hard final (average 50; standard deviation 12)

If the scatter diagram comparing midterm & final scores for students has a typical oval shape with correlation 0.75, then...

What is the average final score for students who scored 90 on the midterm?
A. 76
B. 90
C. 68
D. 82
E. 67.5

Discussion question

```python
def my_func(c):
    if c < -2:
        return 4
    elif c > 2:
        return 4
    else:
        return abs(c)+2
```

Pick the option that best completes the sentence:

"The expression **minimize(my_func)** evaluates to…"

A: -3
B: 0
C: 1
D: 2
E: 4

Discussion Question

How does the SD of the fitted values relate to r?

    A. (SD of fitted) / (SD of y) = r

    B. (SD of fitted) / (SD of y) = |r|

    C. (SD of fitted) / (SD of residuals) = r

    D. (SD of fitted) / (SD of residuals) = |r|

Discussion Question 1

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$
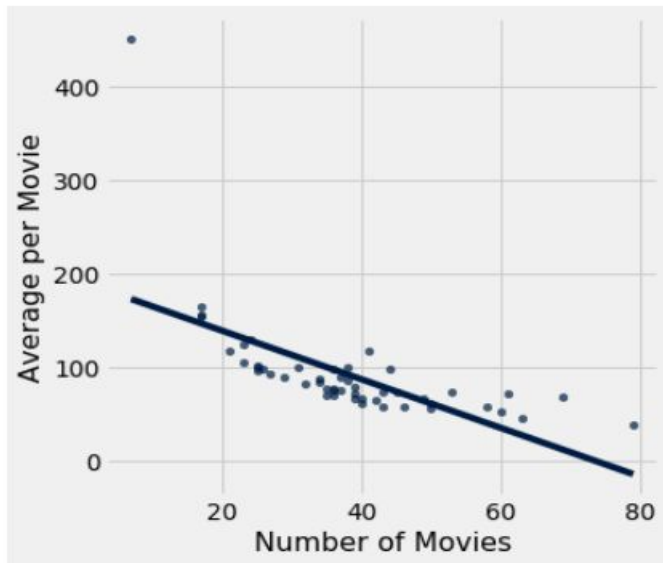

What are the units of the slope of the regression line?

A. no units
B. movie per dollar
C. dollar per movie
D. million dollars per movie
E. movie per million dollars

Discussion Question 2

slope of the regression line $= r \cdot \dfrac{\text{SD of } y}{\text{SD of } x}$ **= - 2.5**

intercept of the regression line $=$ average of y $-$ slope $\cdot$ average of x
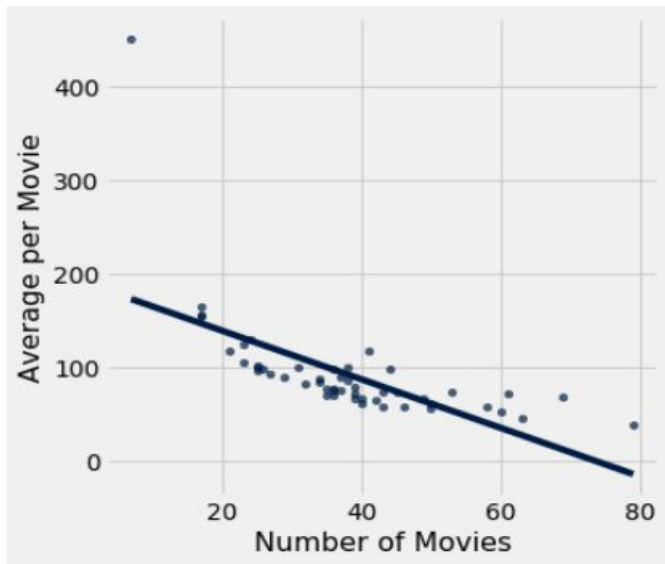
Actor A appeared in *m* movies, which made an average of 87 million each. If Actor B appeared in *m+2* movies, estimate his average earnings per movie.

A. 82 million dollars
B. 84 million dollars
C. 84.5 million dollars
D. 89 million dollars
E. 89.5 million dollars

Discussion Question 3

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$
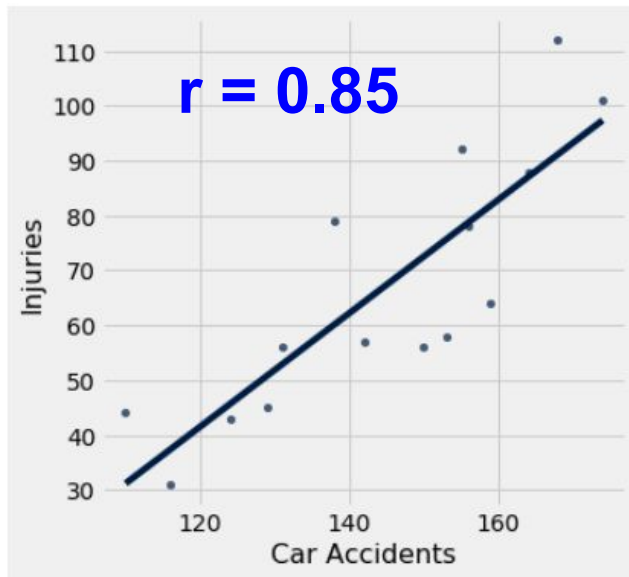
r = 0.85

If r = 0.85, and we know that 150 car accidents occurred this month, estimate the number of car accident-related injuries this month.

A. 150 * 0.85
B. 150 / 0.85
C. 150 * sqrt(1-0.85)
D. 150 ** 0.85
E. None of the above

Discussion Question 3

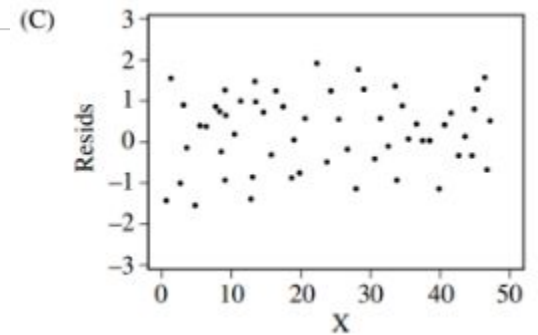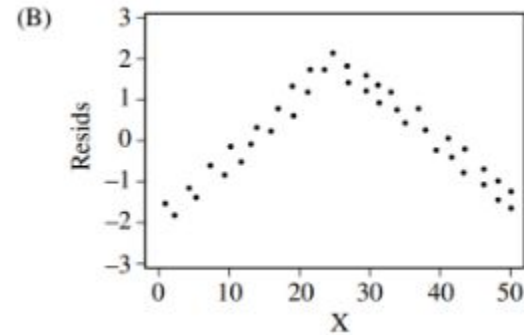$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$
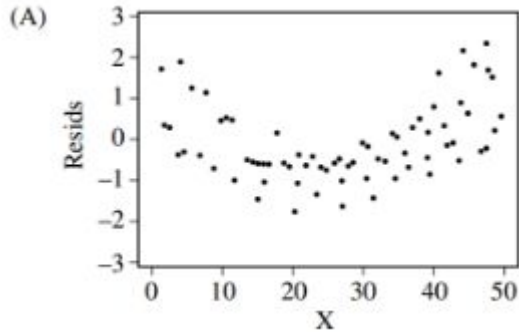
$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$
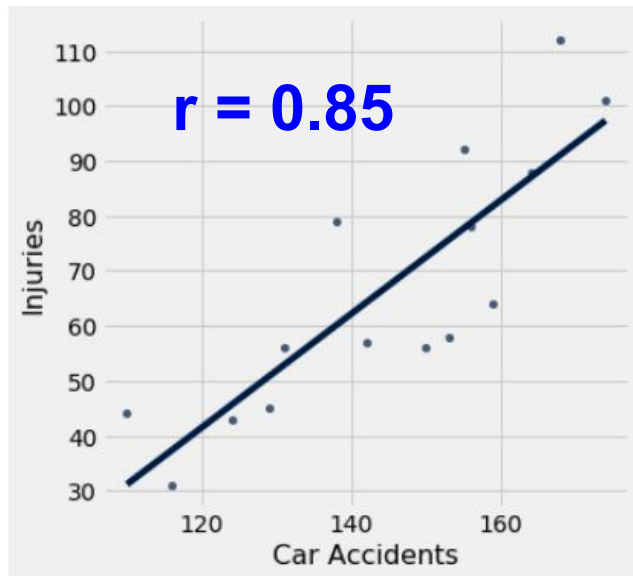
r = 0.85

If r = 0.85, and we know that 150 car accidents occurred this month, estimate the number of car accident-related injuries this month.

	mean	sd
Car Accidents	144	19
Injuries	70	23

Discussion Question
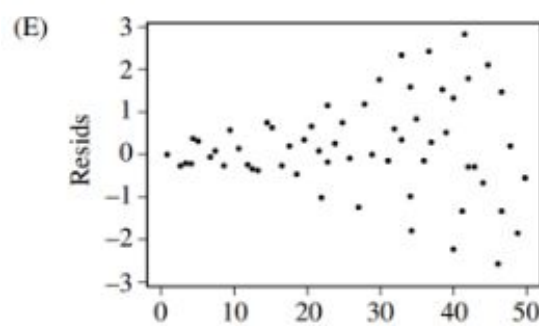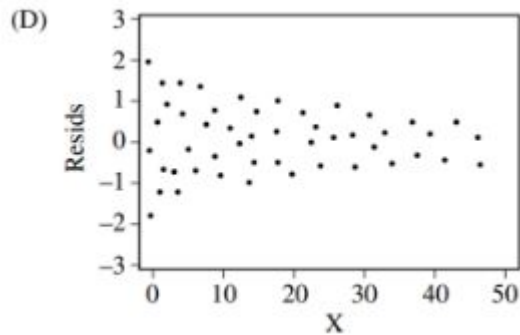
Which of the plots provides the strongest evidence that the regression line is an appropriate model?

Discussion Question

What kind of test results will lead you to conclude that a patient does not have Chronic Kidney Disease ("0")?

A: Hemoglobin below average   Glucose below average
B: Hemoglobin above average   Glucose below average
C: Hemoglobin below average   Glucose above average
D: Hemoglobin above average   Glucose above average
E: More than one possible answer