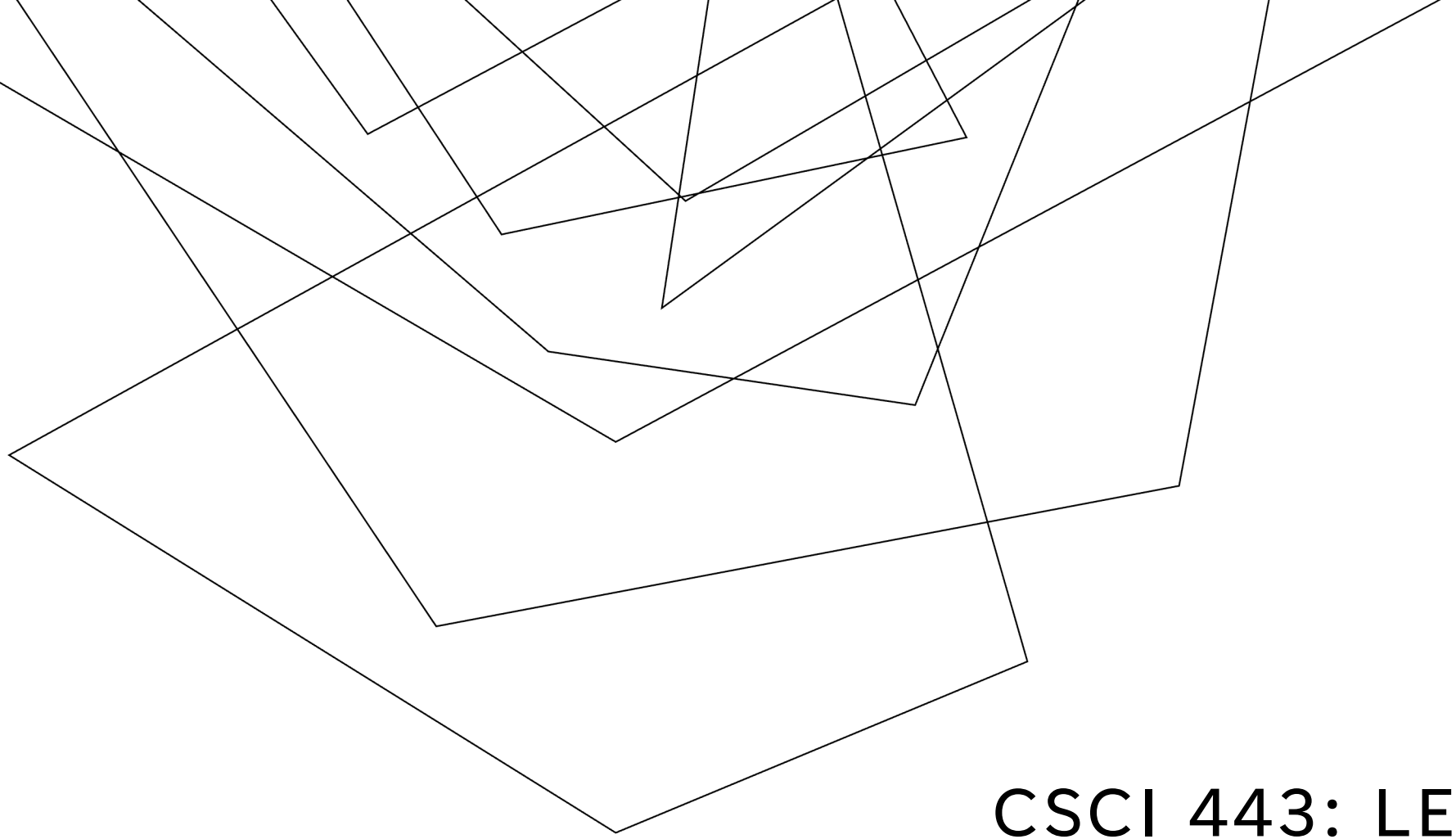


CSCI 443: LECTURE 12 MIDTERM, SKEW

Professor David Harrison

OFFICE HOURS

Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

HOMEWORK 4

Will be handed out after break.

DATES OF INTEREST

March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break
March 19	Homework 4 handed out
March 28	Homework 4 due

BLACKBOARD & GITHUB

Slides up through lecture 11 on blackboard.

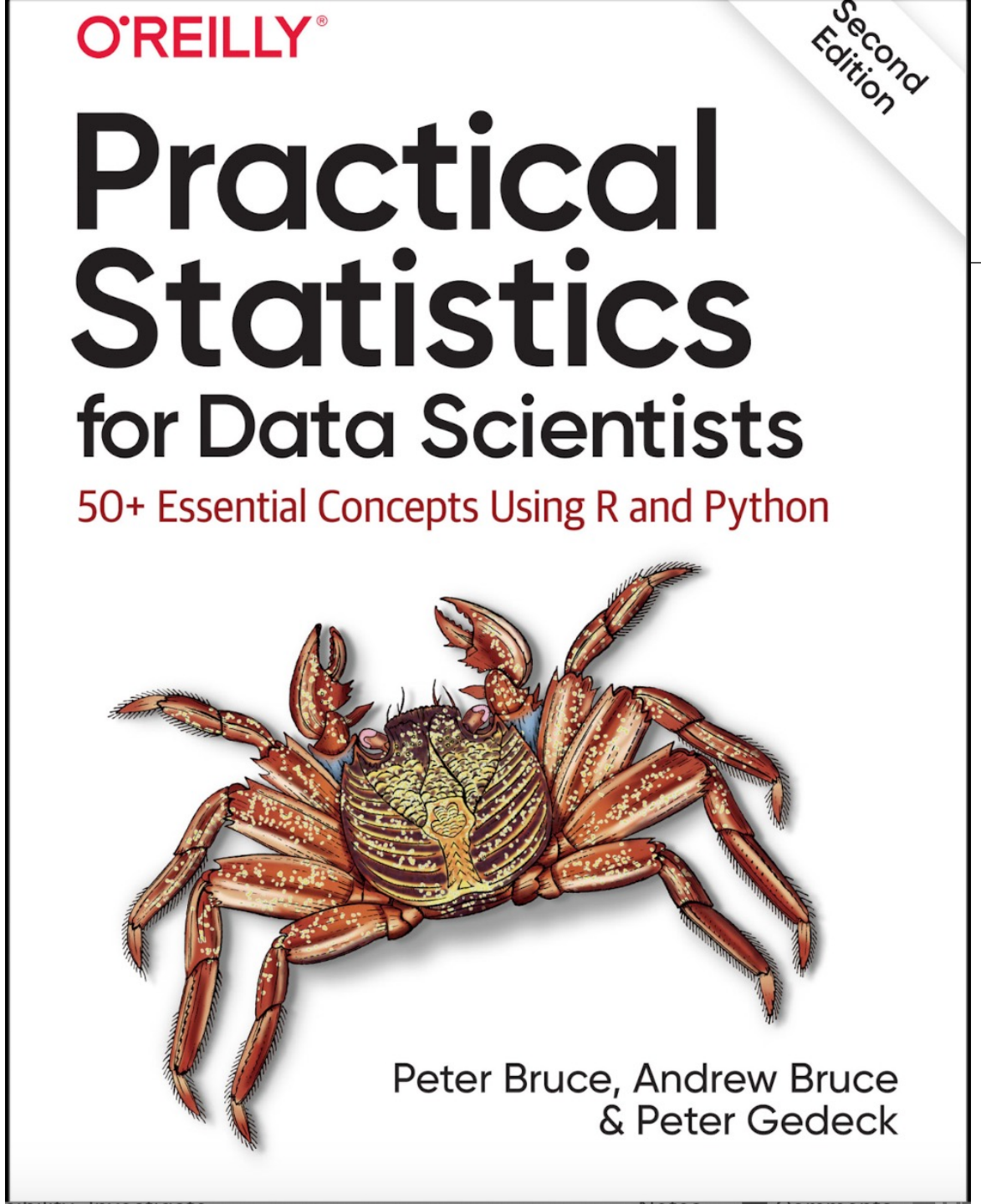
Lecture slides and examples committed to GitHub also up through lecture 11.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

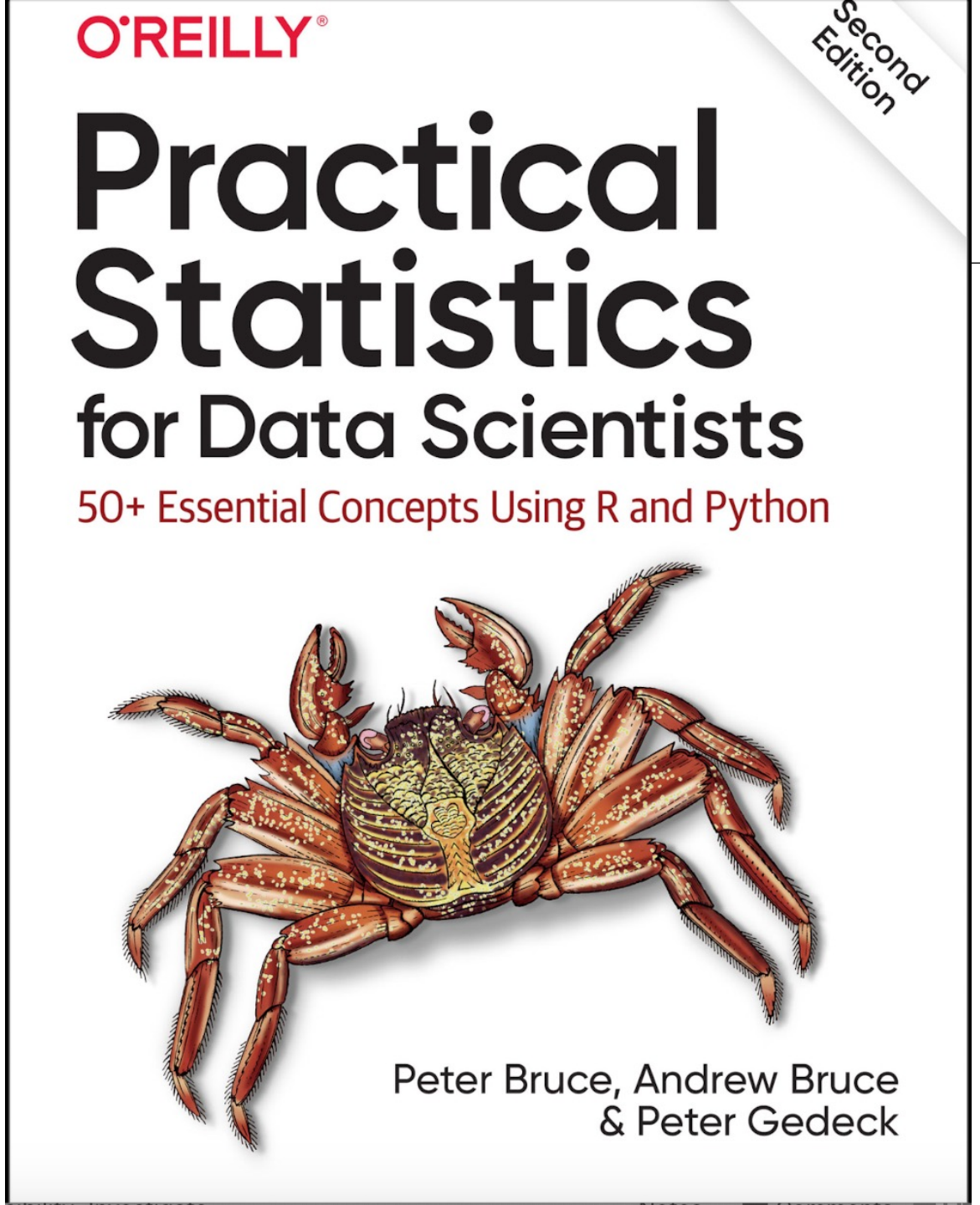
READ ABOUT

- QQ Plots
- Long tailed distributions
- Student t-distribution
- Binomial distribution
- Chi-Squared distribution



THINGS I WANT TO COVER TODAY

- Review Exam
- Skewness
- Long tailed distributions
- Binomial distribution
- Chi-Squared distribution



REVIEW PROBLEM #1

- “No math, not numeric”
- ”Ordinal, think order.
  Nominal think not.”

REVIEW PROBLEM #1

b) Rankings of a movie (e.g., Excellent, Good, Fair, Poor).

•

REVIEW PROBLEM #1

b) Rankings of a movie (e.g., Excellent, Good, Fair, Poor).

Categorial (ordinal).

REVIEW PROBLEM #1

a) Temperature in degrees Celsius.

REVIEW PROBLEM #1

a) Temperature in degrees Celsius.

Numeric

REVIEW PROBLEM #1

a) Temperature in degrees Celsius.

Numeric

REVIEW PROBLEM #1

f) The zip codes of addresses in a city.

REVIEW PROBLEM #1

f) The zip codes of addresses in a city.

Numeric

REVIEW PROBLEM #1

f) The zip codes of addresses in a city.

~~Numeric~~

REVIEW PROBLEM #1

f) The zip codes of addresses in a city.

Categorical (nominal)

REVIEW PROBLEM #2

d) True/False: The number of pages in the first edition of "To Kill a Mocking Bird" by Harper Lee is a random variable.

?

REVIEW PROBLEM #2

d) True/False: The number of pages in the first edition of "To Kill a Mocking Bird" by Harper Lee is a random variable.

FALSE. Not random.

REVIEW PROBLEM #2

i) True/False: The bus arrives early on Monday, Tuesday, and Wednesday is an event.

?

REVIEW PROBLEM #2

i) True/False: The bus arrives early on Monday, Tuesday, and Wednesday is an event.

TRUE.

If you can assign a probability to it, it is an event.

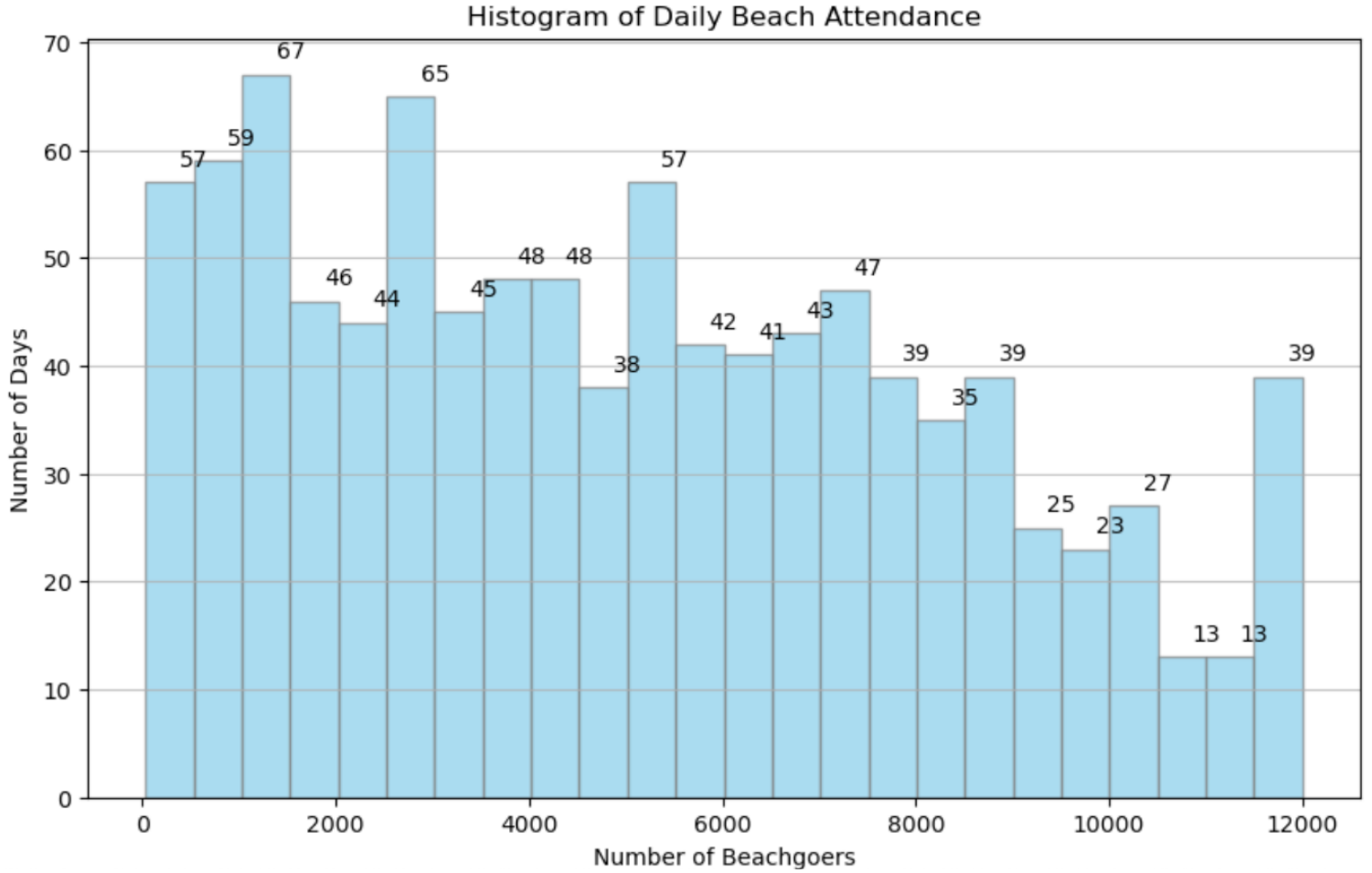
REVIEW PROBLEM #3

The Dreemes Job

Dreemes is a seaside town known for its beach. The Dreemes Chamber of Commerce has hired you to do some data science with some of the data it has collected about beach attendance and weather. Every year, Dreemes staffs a beach for 100 days covering the summer months. For the last ten years they have gathered data for three random variables (N=1000):

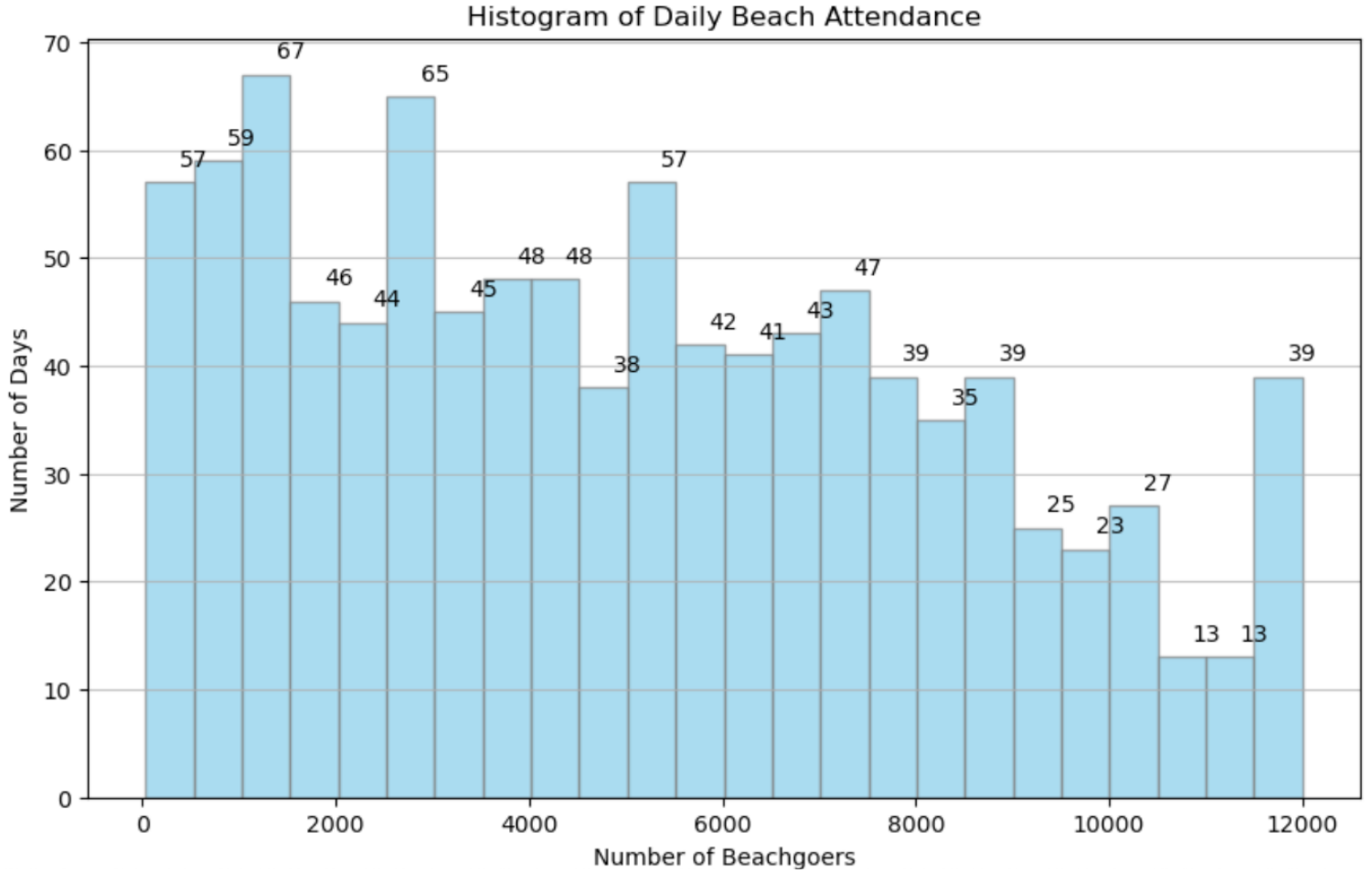
- number of daily beachgoers
- daily high temperature
- inches of daily precipitation

REVIEW PROBLEM #3



d) How many days have more than 15000 beachgoers?

REVIEW PROBLEM #3

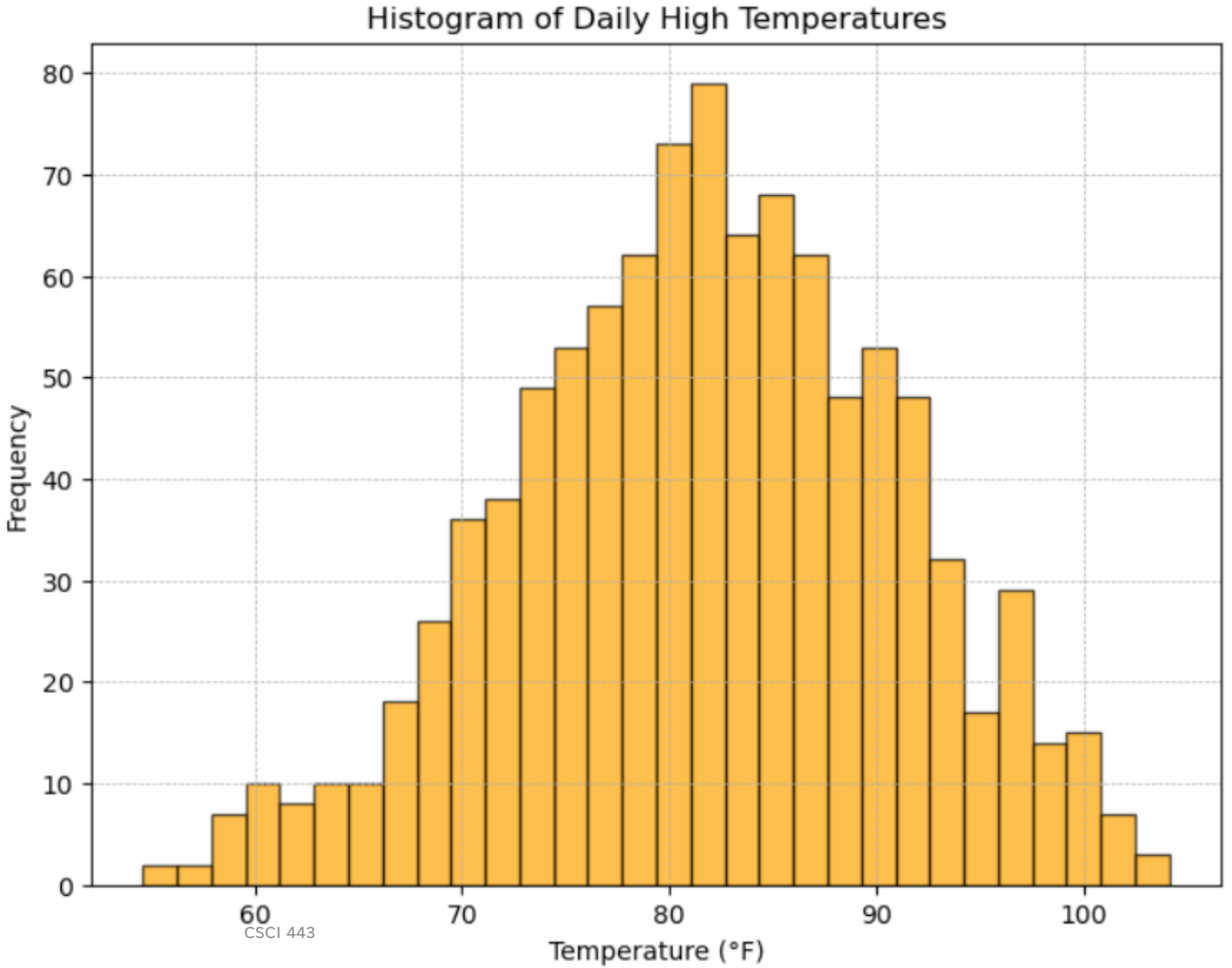


d) How many days have more than 15000 beachgoers?

Zero. There are 1000 days. All are accounted for. Max is 12000.

REVIEW PROBLEM 4

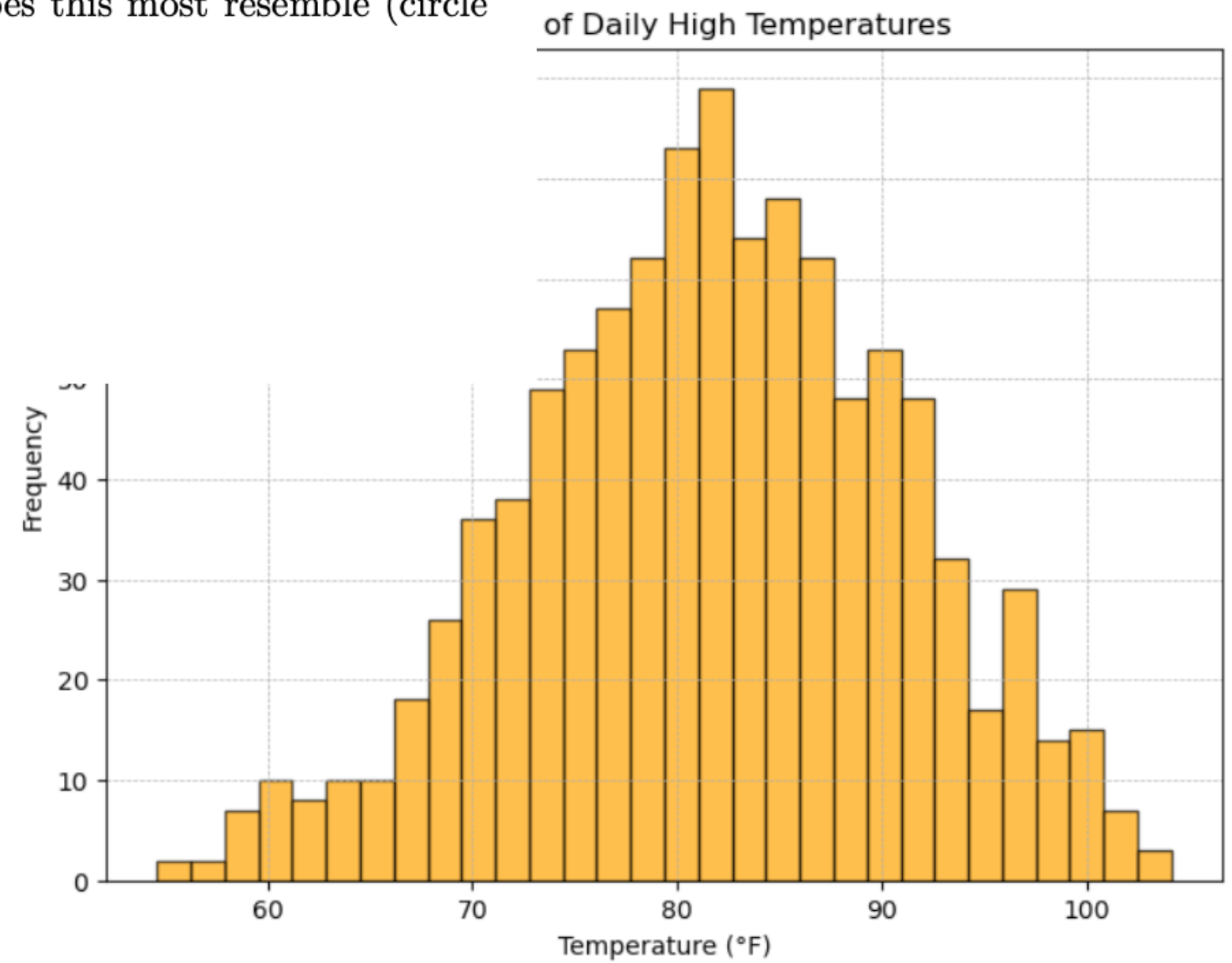
Histogram of daily high temperatures.



REVIEW PROBLEM 4

a) Which of the following distributions does this most resemble (circle one)?

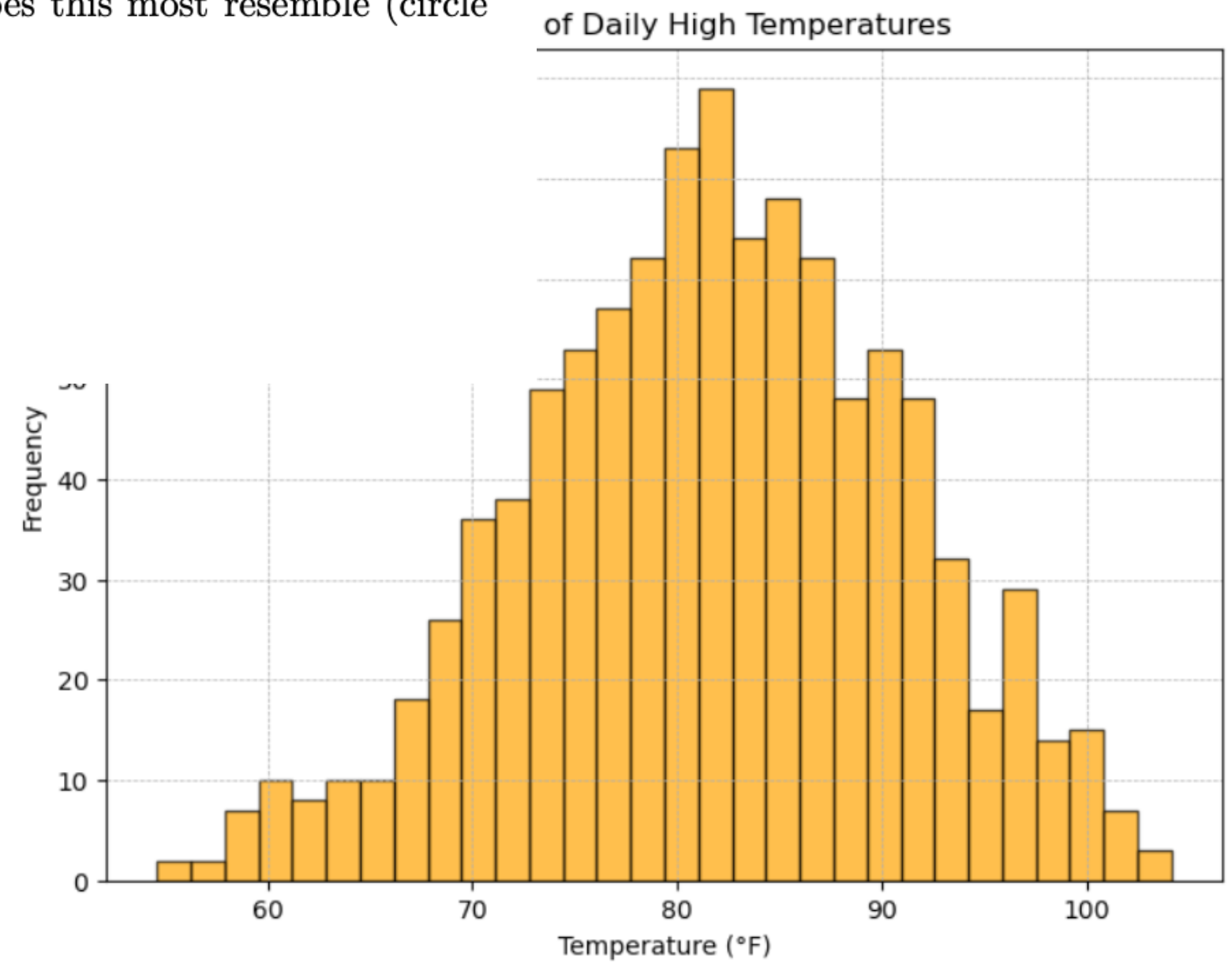
- uniform
- exponential
- Gaussian
- bimodal



REVIEW PROBLEM 4

a) Which of the following distributions does this most resemble (circle one)?

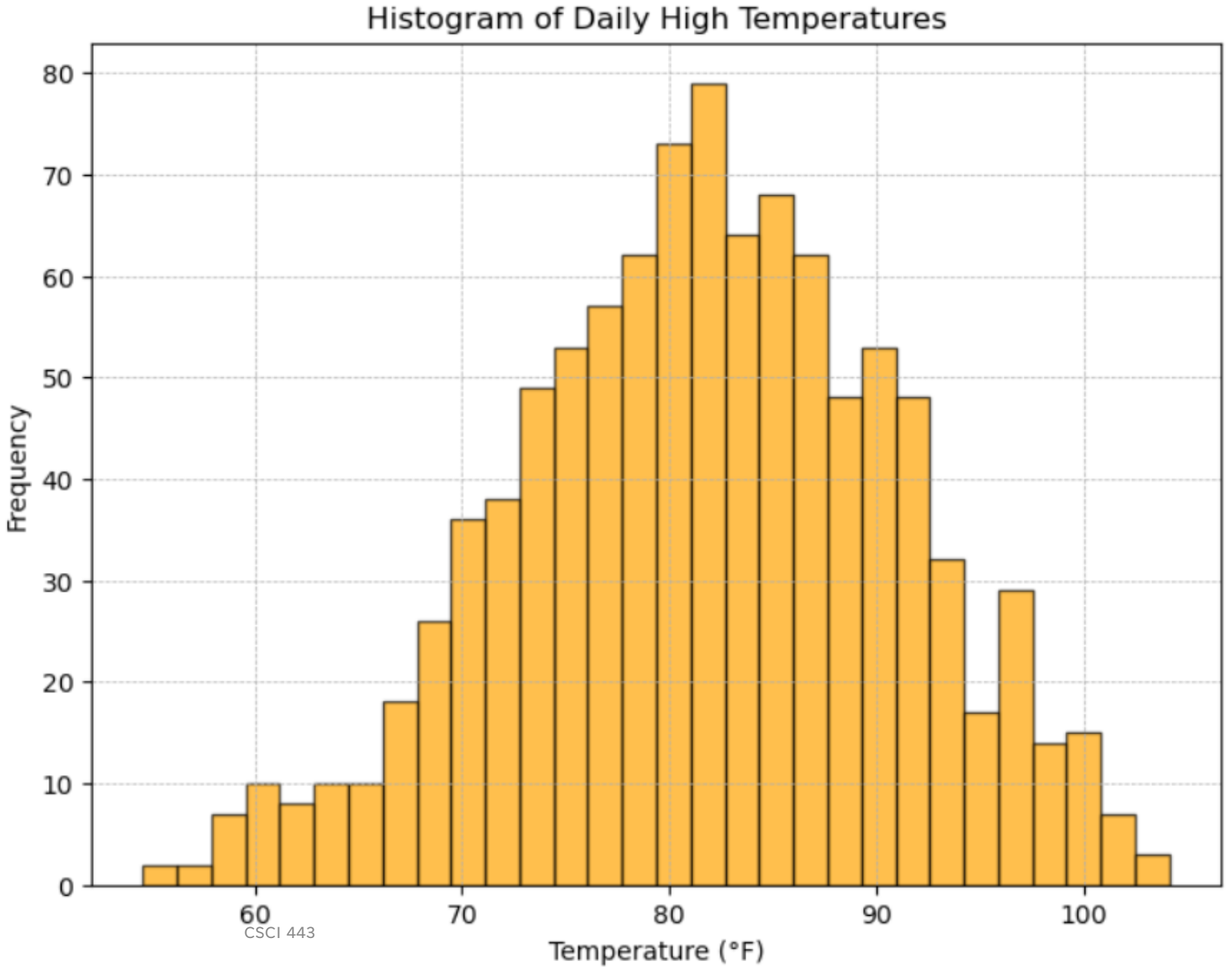
- uniform
- exponential
- Gaussian
- bimodal



REVIEW PROBLEM 4

Histogram of daily high temperatures.

Is this skewed?



SKEWNESS

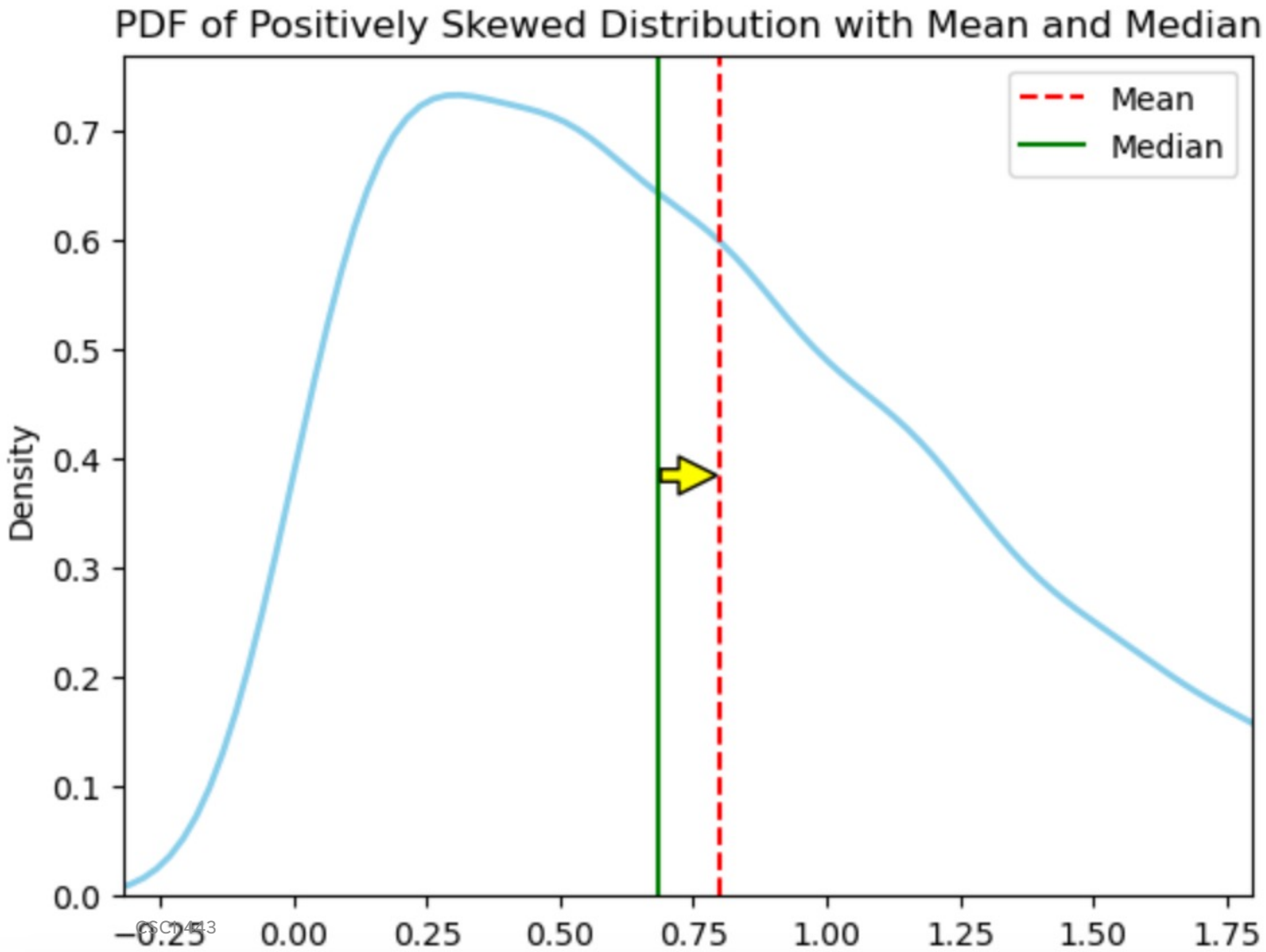
When is a distribution skewed?

SKEWNESS

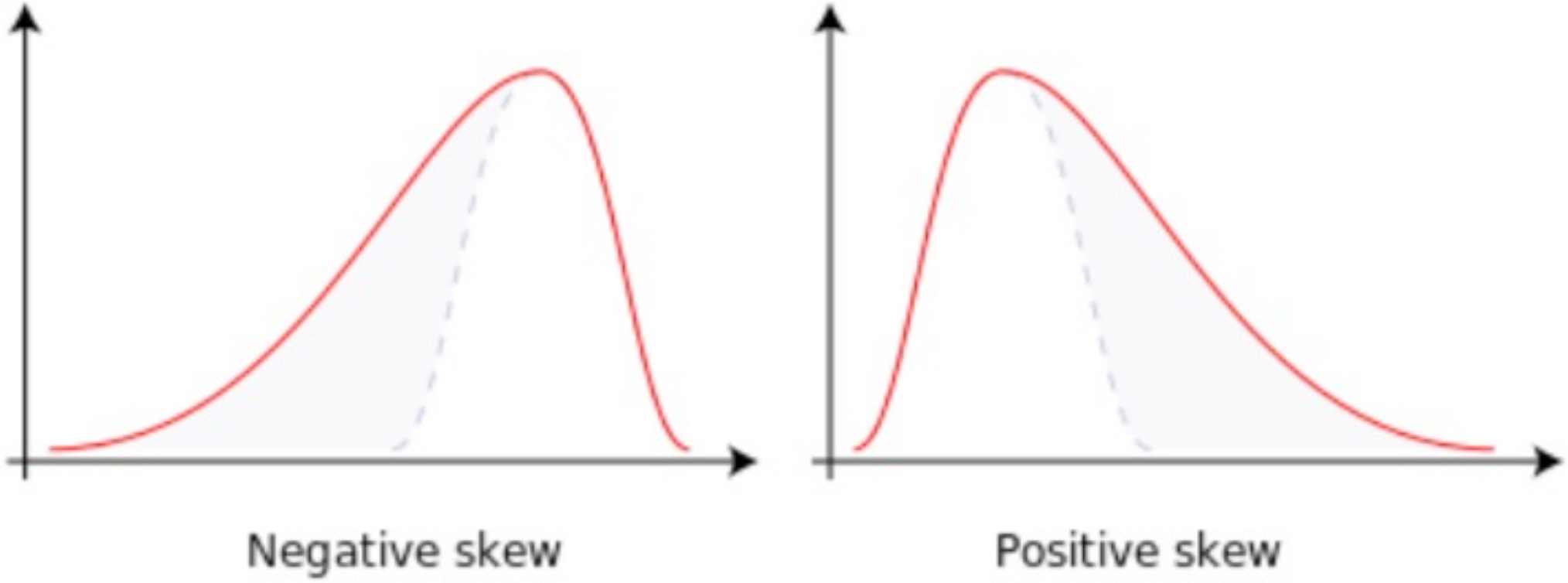
When is a distribution skewed?

Rule of thumb: “When the mean deviates from the median.”

$$\gamma_1 = \frac{E[(X - \mu)^3]}{\sigma^3}$$



SKEWNESS



Follow the tail...

SKEWNESS EXAMPLES: SKEWNORM

Play with scipy.stats.skewnorm

$$f(x;\alpha) = 2\phi(x)\Phi(\alpha x)$$

When $\alpha = 0$ this becomes

$$f(x;\alpha) = 2\phi(x)\Phi(0) = 2\phi(x)\cdot\frac{1}{2} = \phi(x)$$

$\alpha > 0$ causes right skew

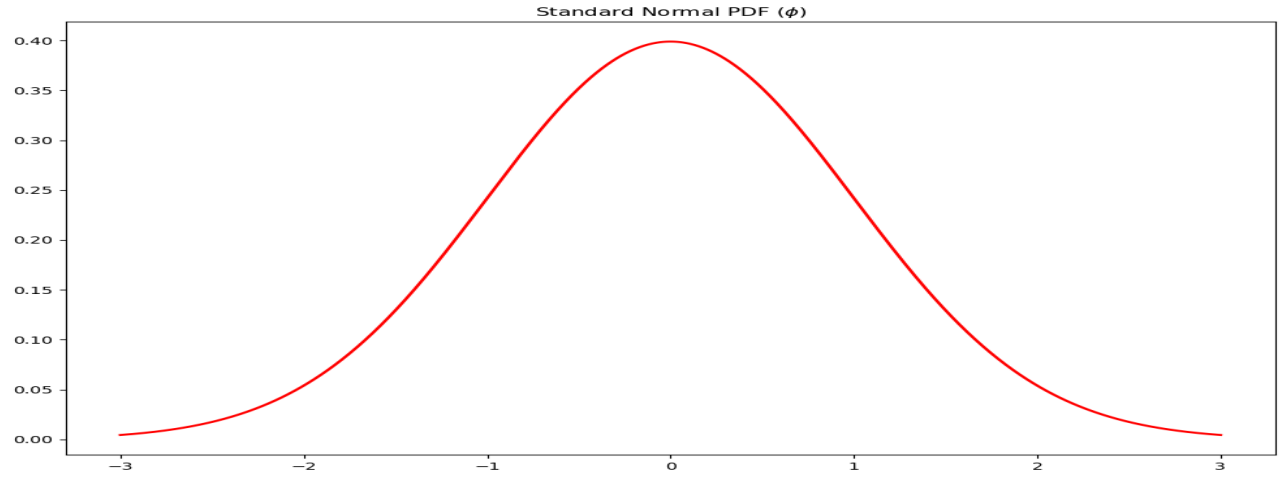
$\alpha < 0$ causes left skew

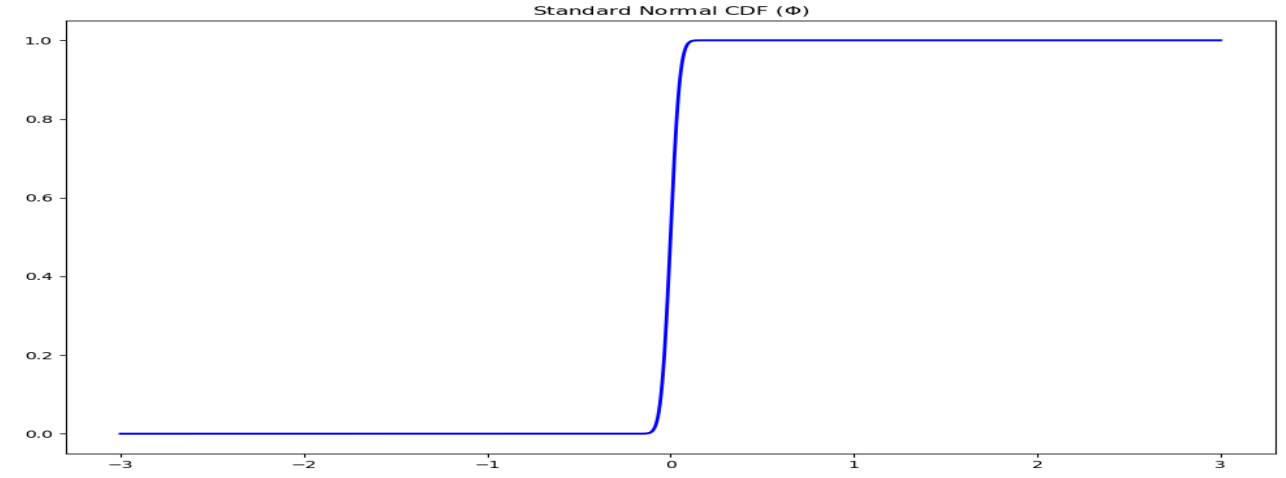
SKEWNORM

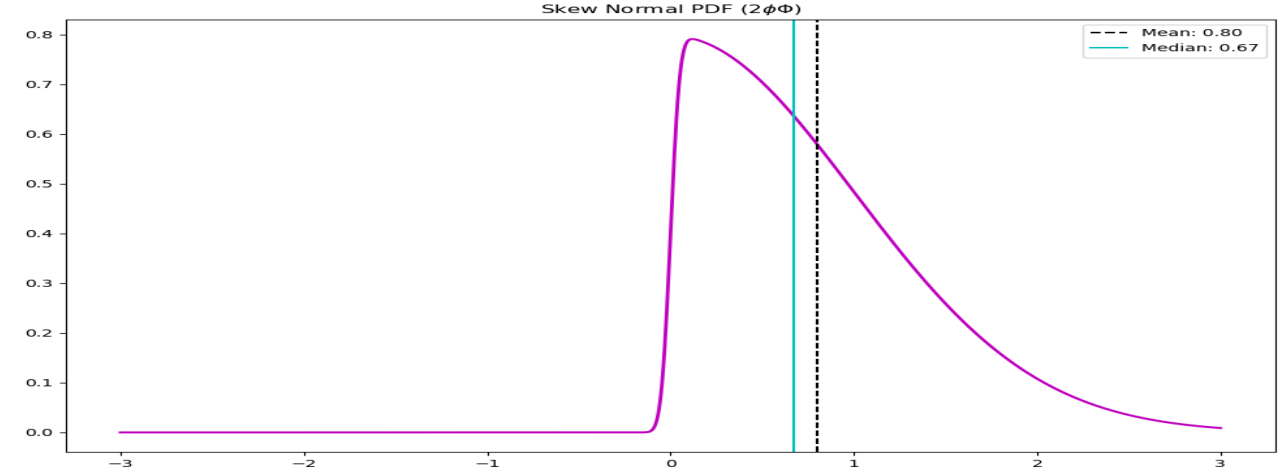
Play with
scipy.stats.skewnorm

$$f(x;\alpha) = 2\phi(x)\Phi(\alpha x)$$

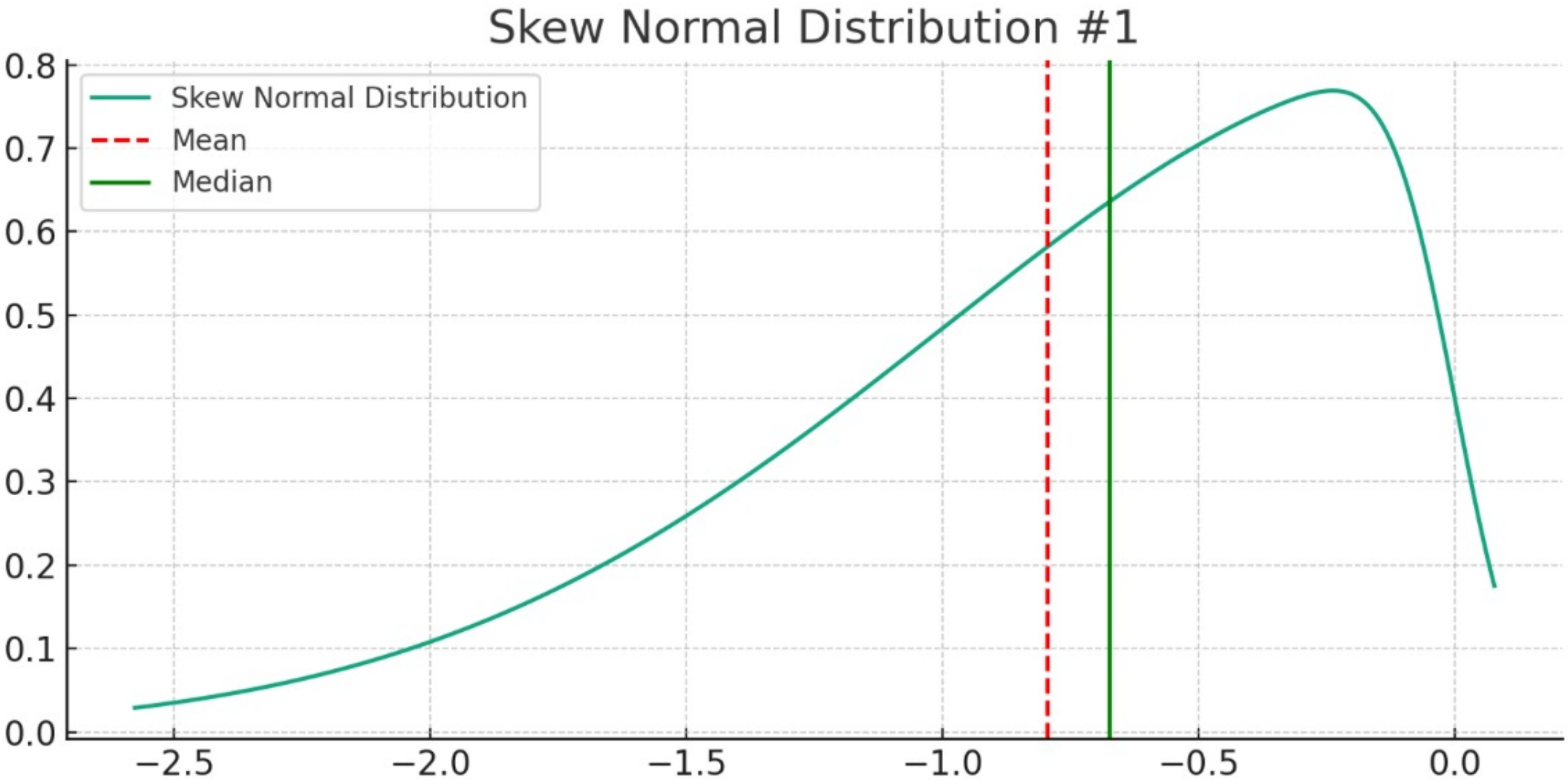
α = 25



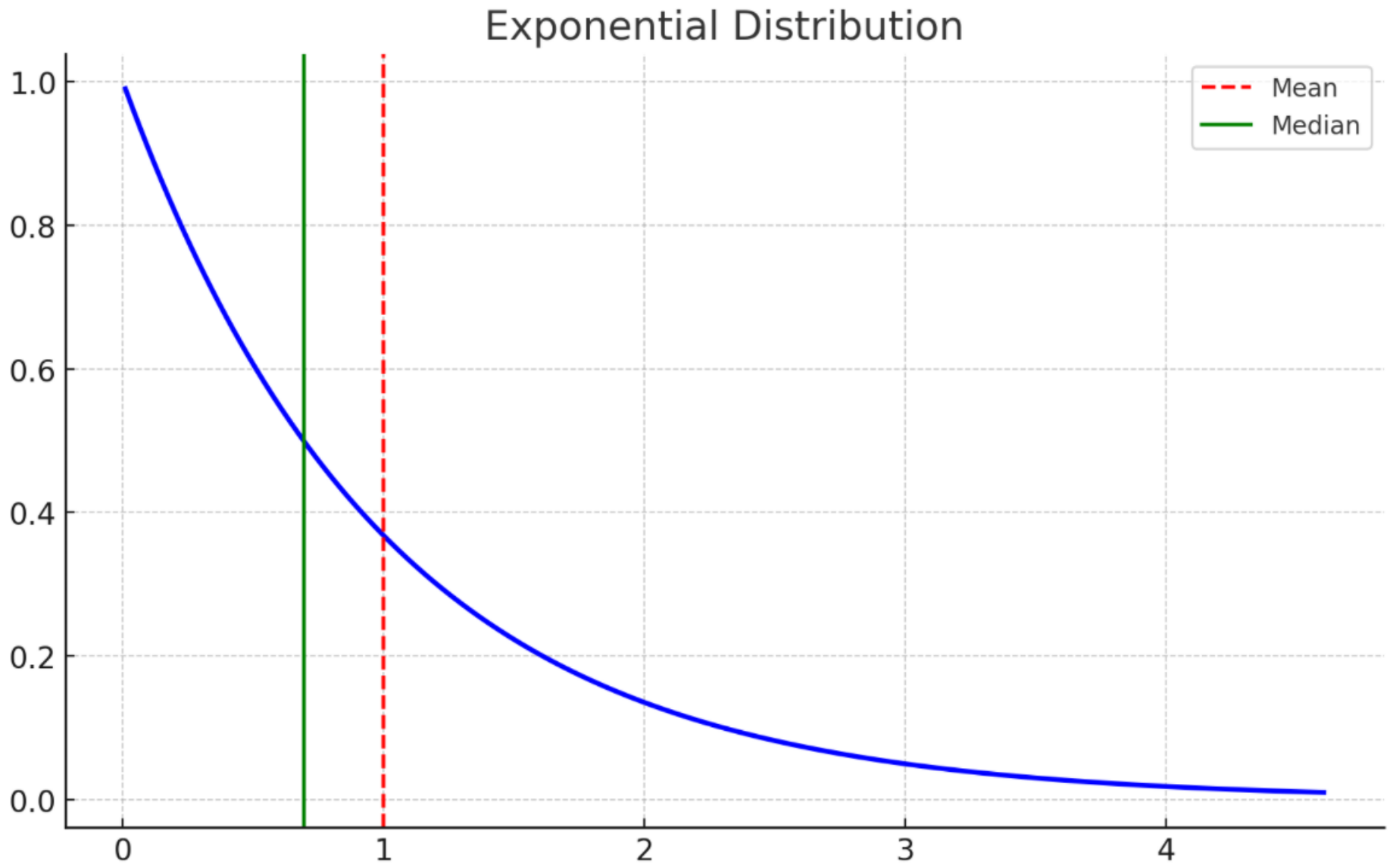




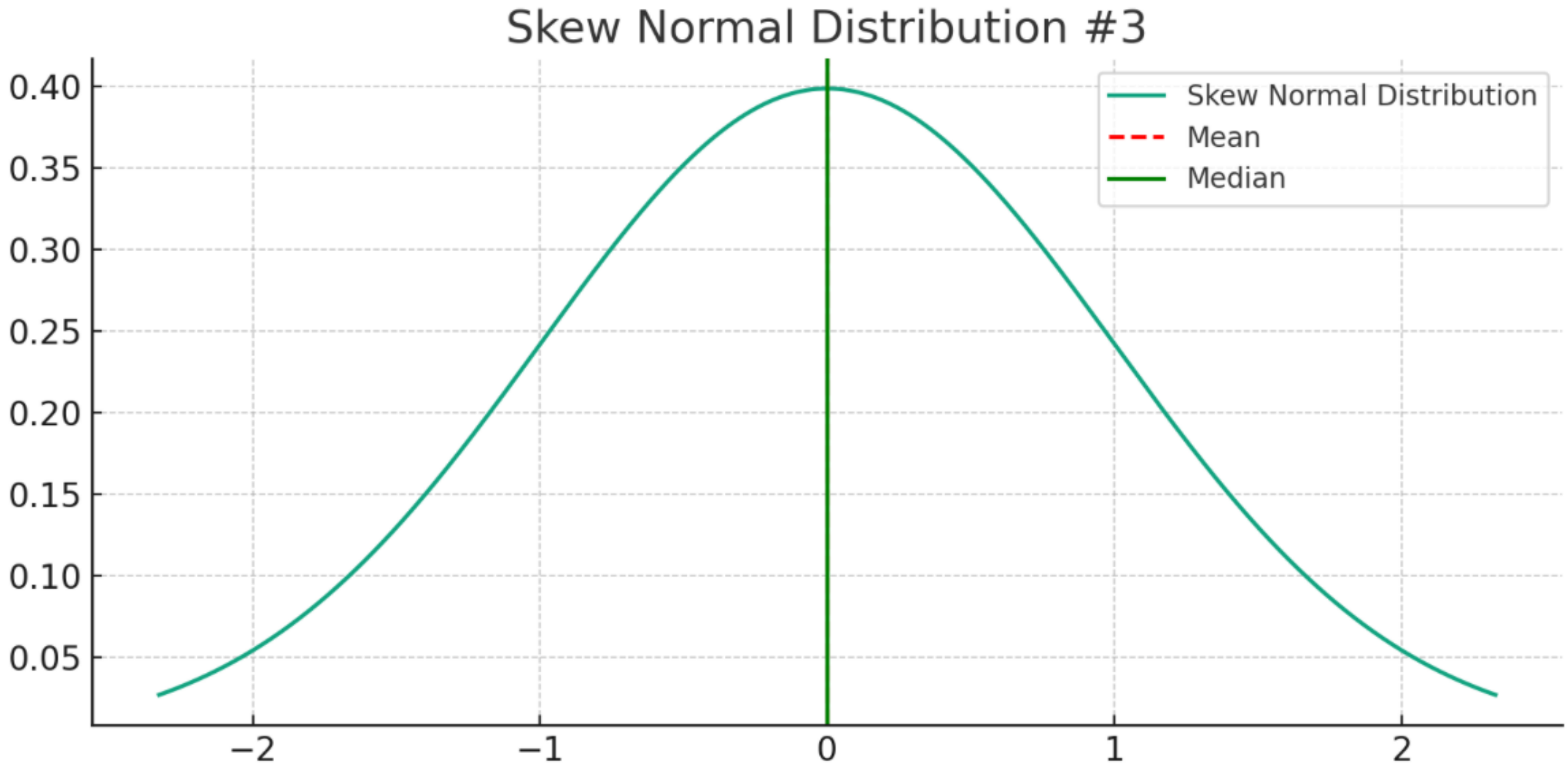
WHICH DIRECTION IS THE SKEW?



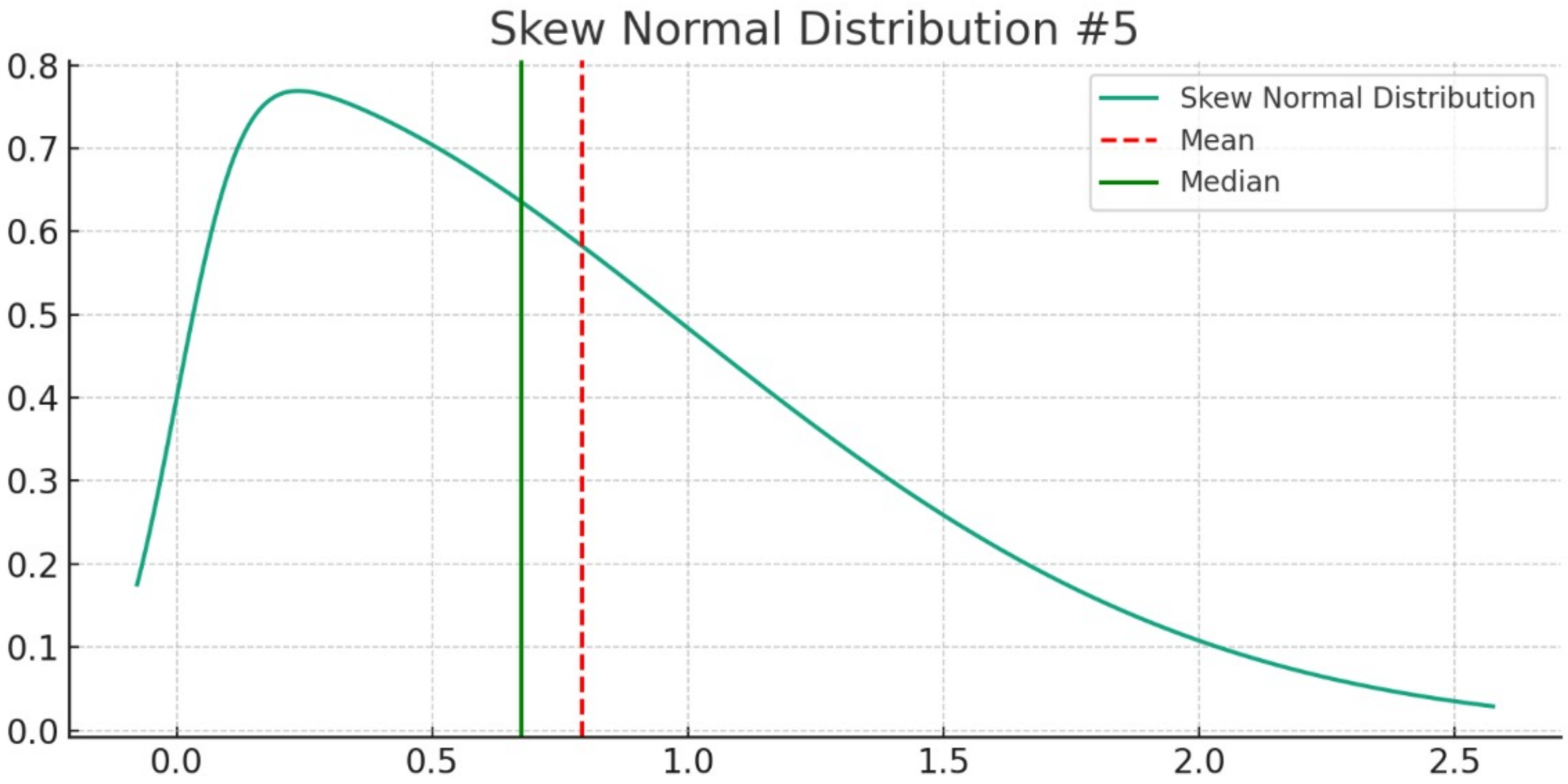
WHICH DIRECTION IS THE SKEW?



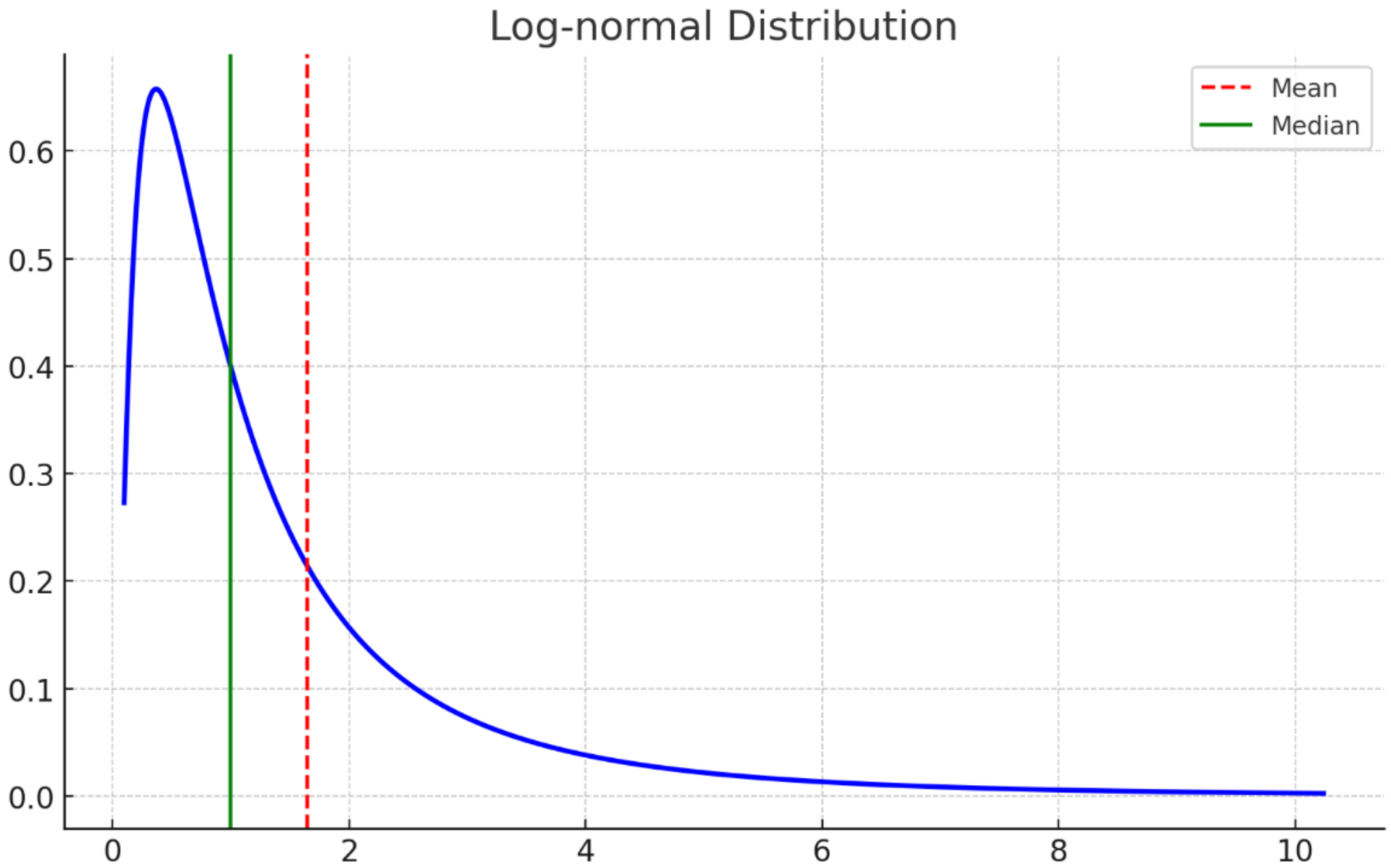
WHICH DIRECTION IS THE SKEW?



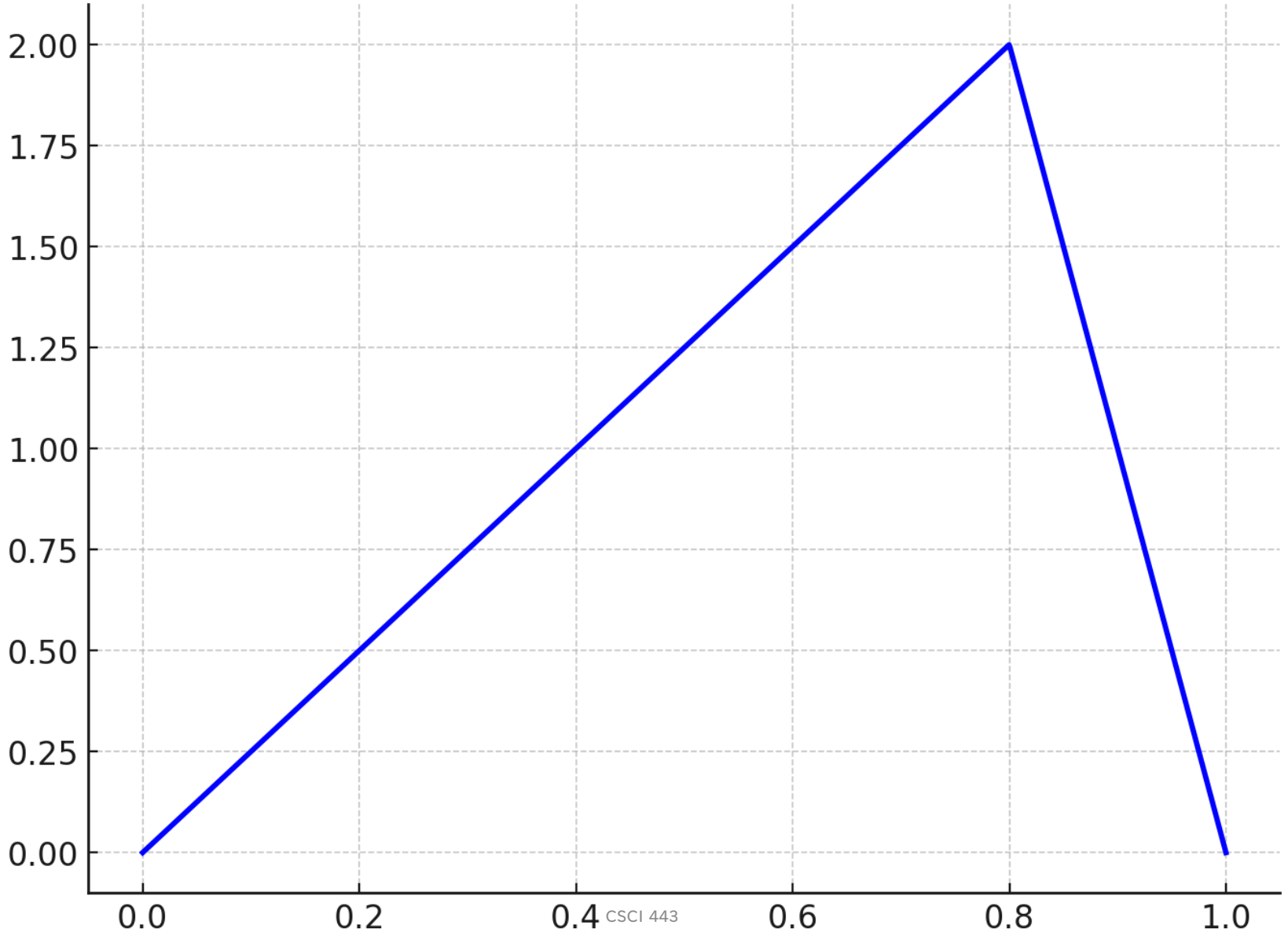
WHICH DIRECTION IS THE SKEW?



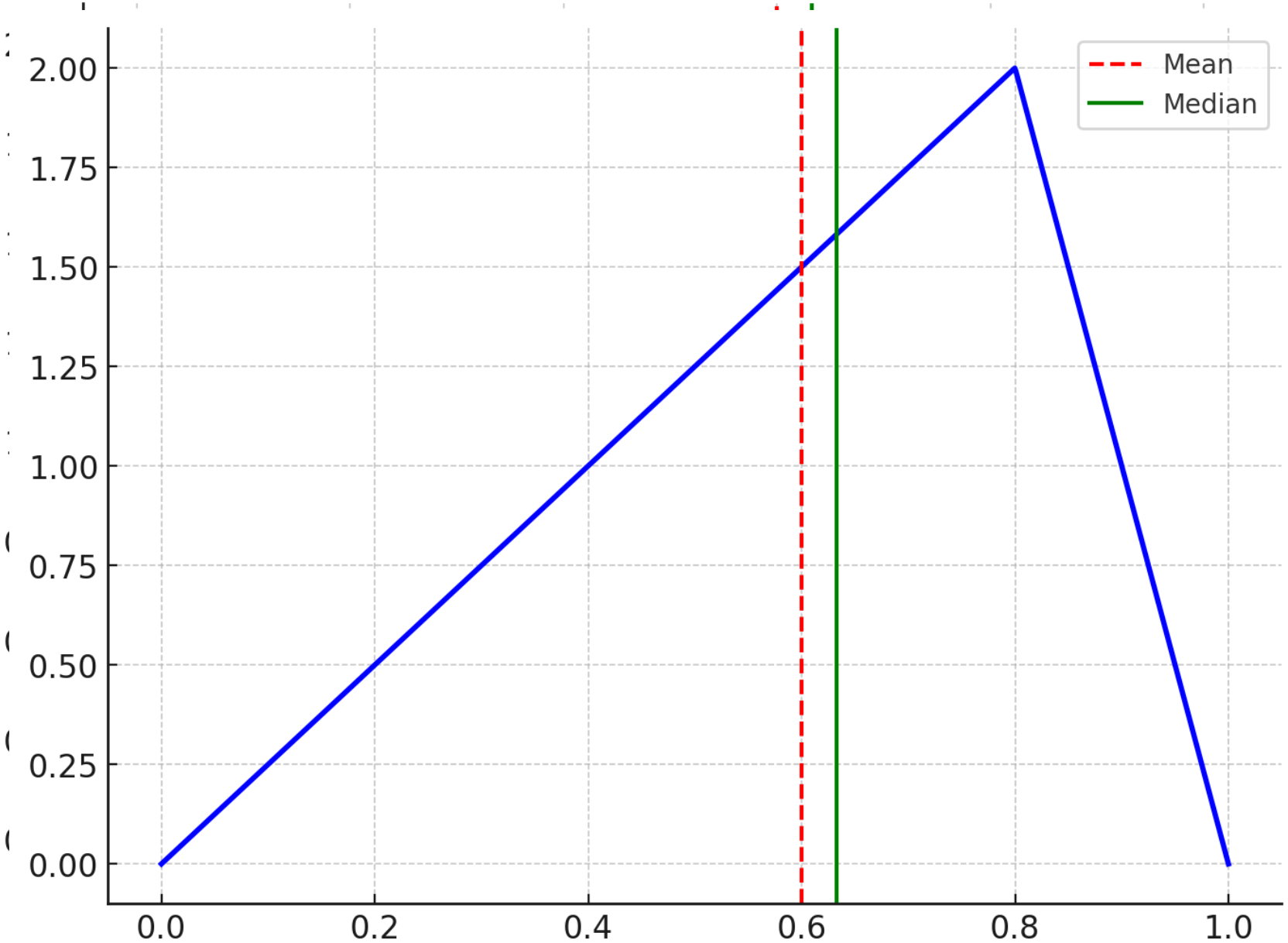
WHICH DIRECTION IS THE SKEW?



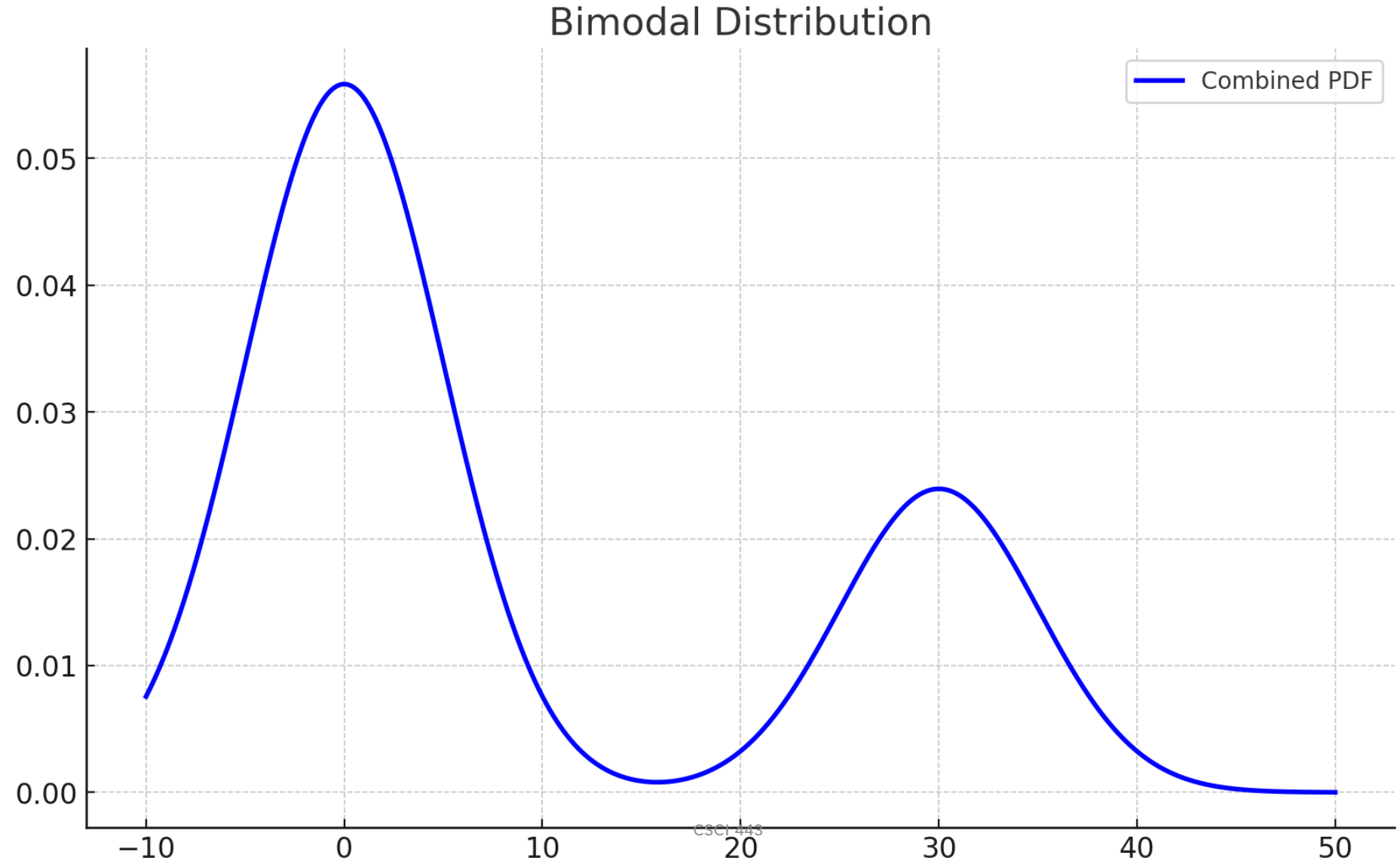
WHICH DIRECTION IS THE SKEW?



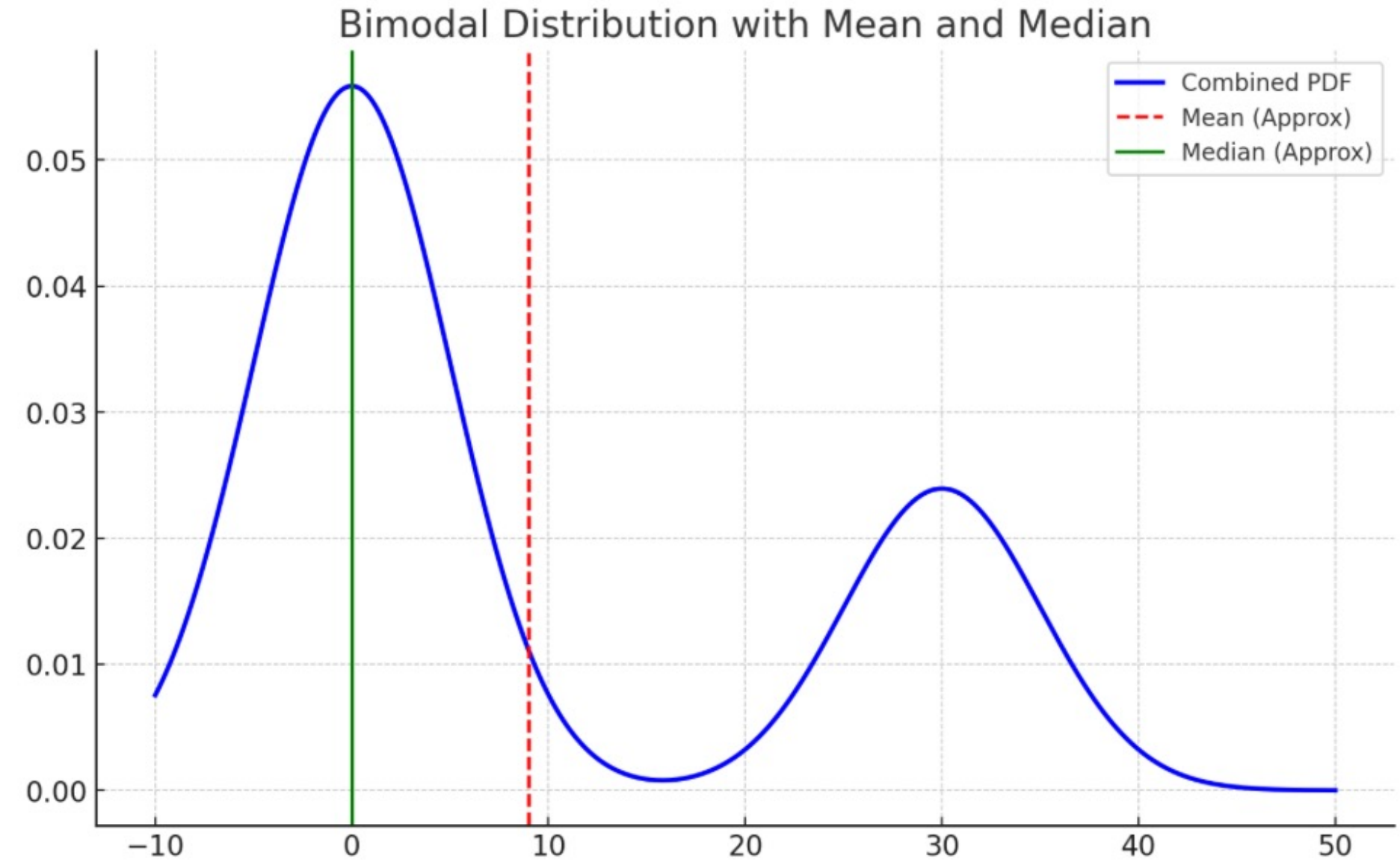
WHICH DIRECTION IS THE SKEW?



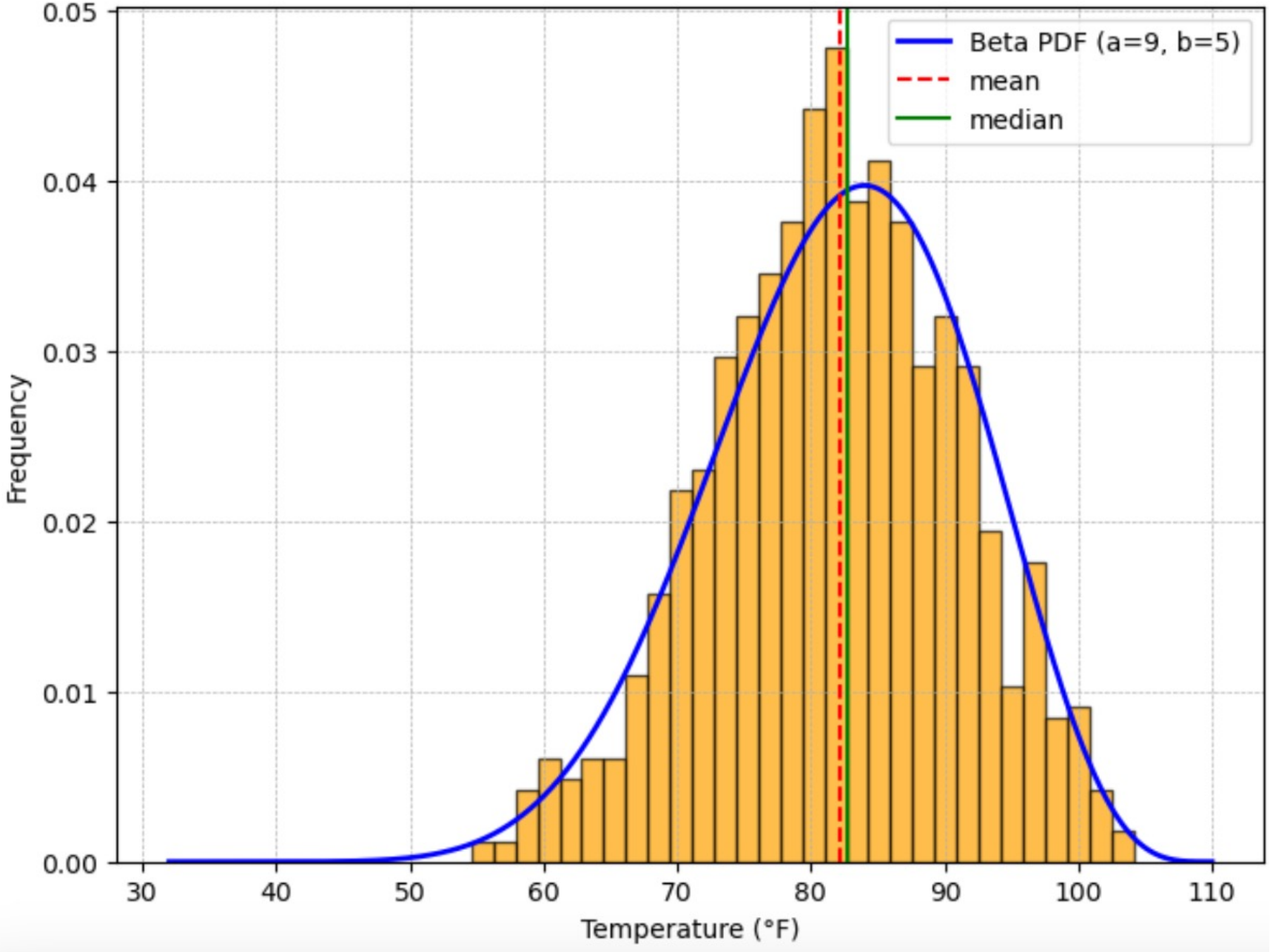
WHICH DIRECTION IS THE SKEW?



WHICH DIRECTION IS THE SKEW?



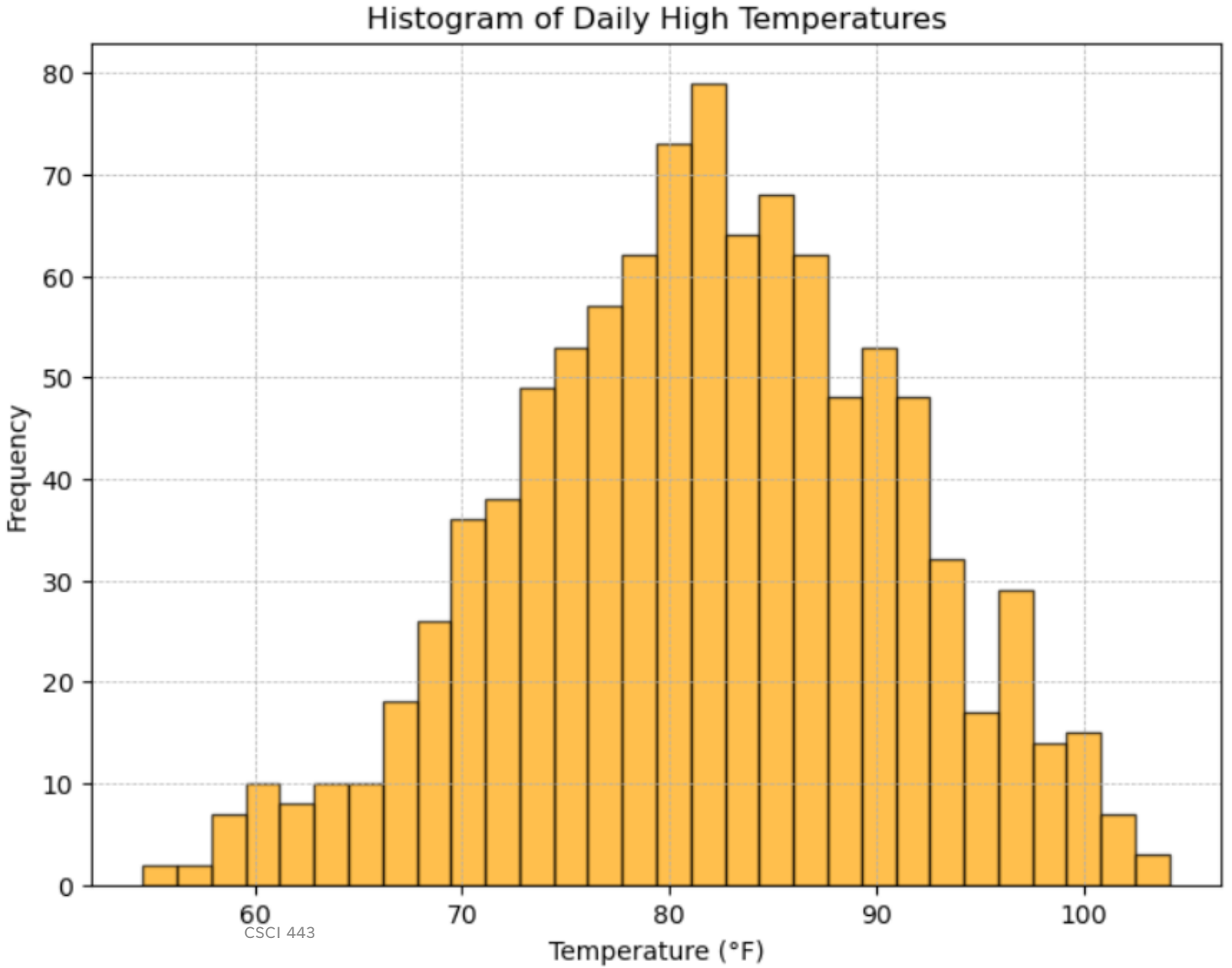
WHICH DIRECTION IS THIS SKEWED?



REVIEW PROBLEM 4

Histogram of daily high temperatures.

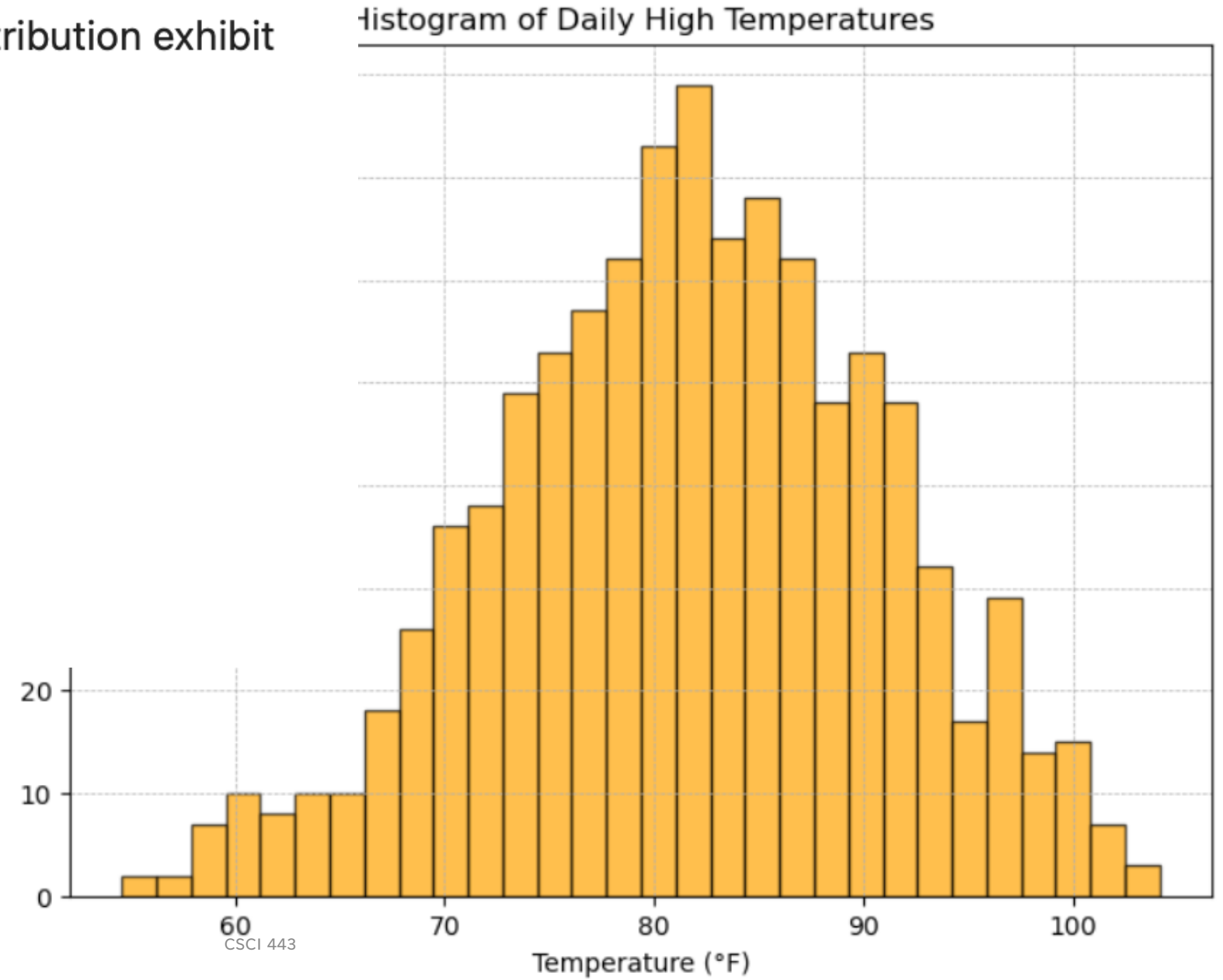
Is this skewed?



REVIEW PROBLEM 4

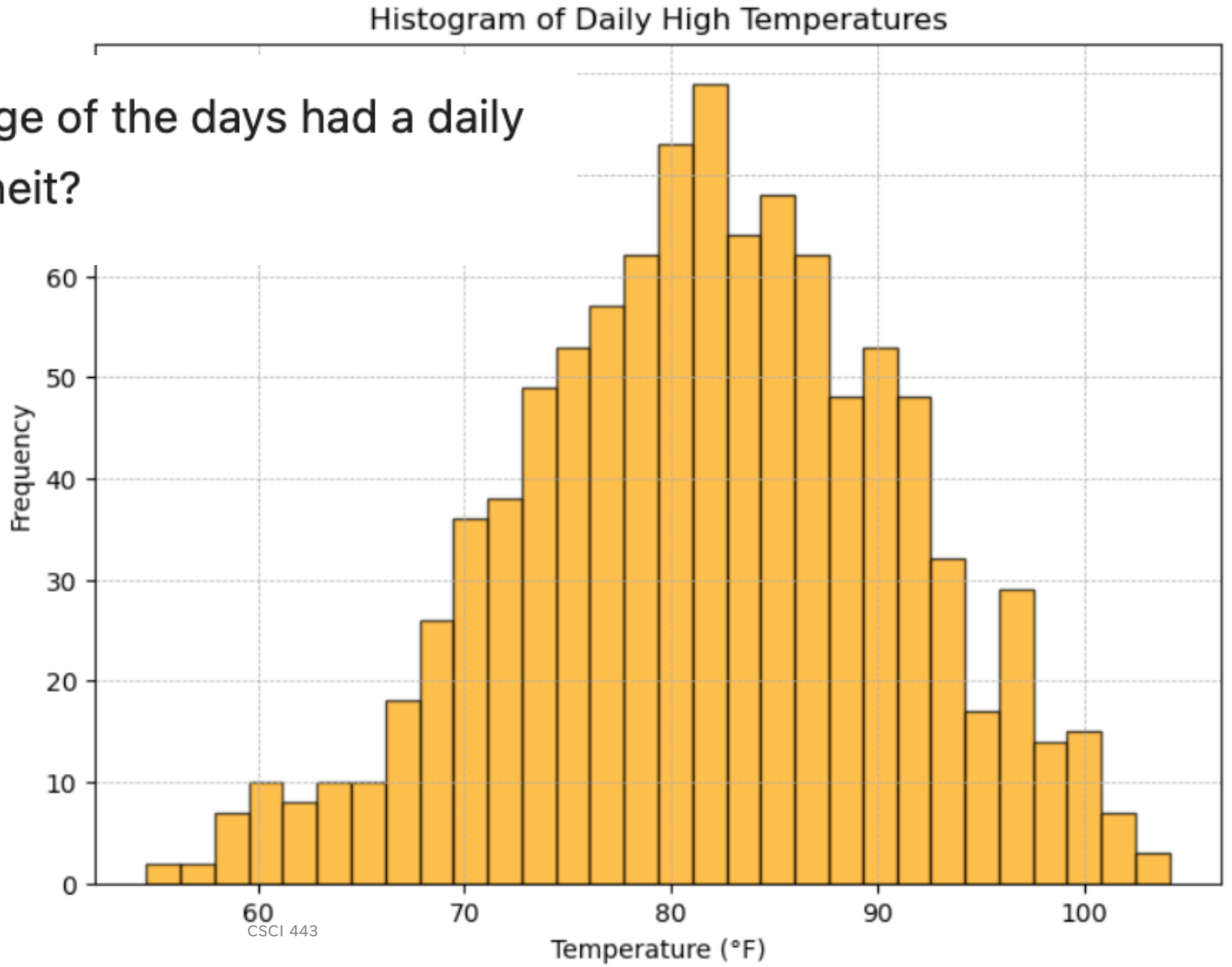
b) Which of the following does this distribution exhibit (circle one)?

- strong left skew
- slight left skew
- no skew
- slight right skew
- strong right skew



REVIEW PROBLEM 4

e) Approximately what percentage of the days had a daily high above 100 degrees Fahrenheit?



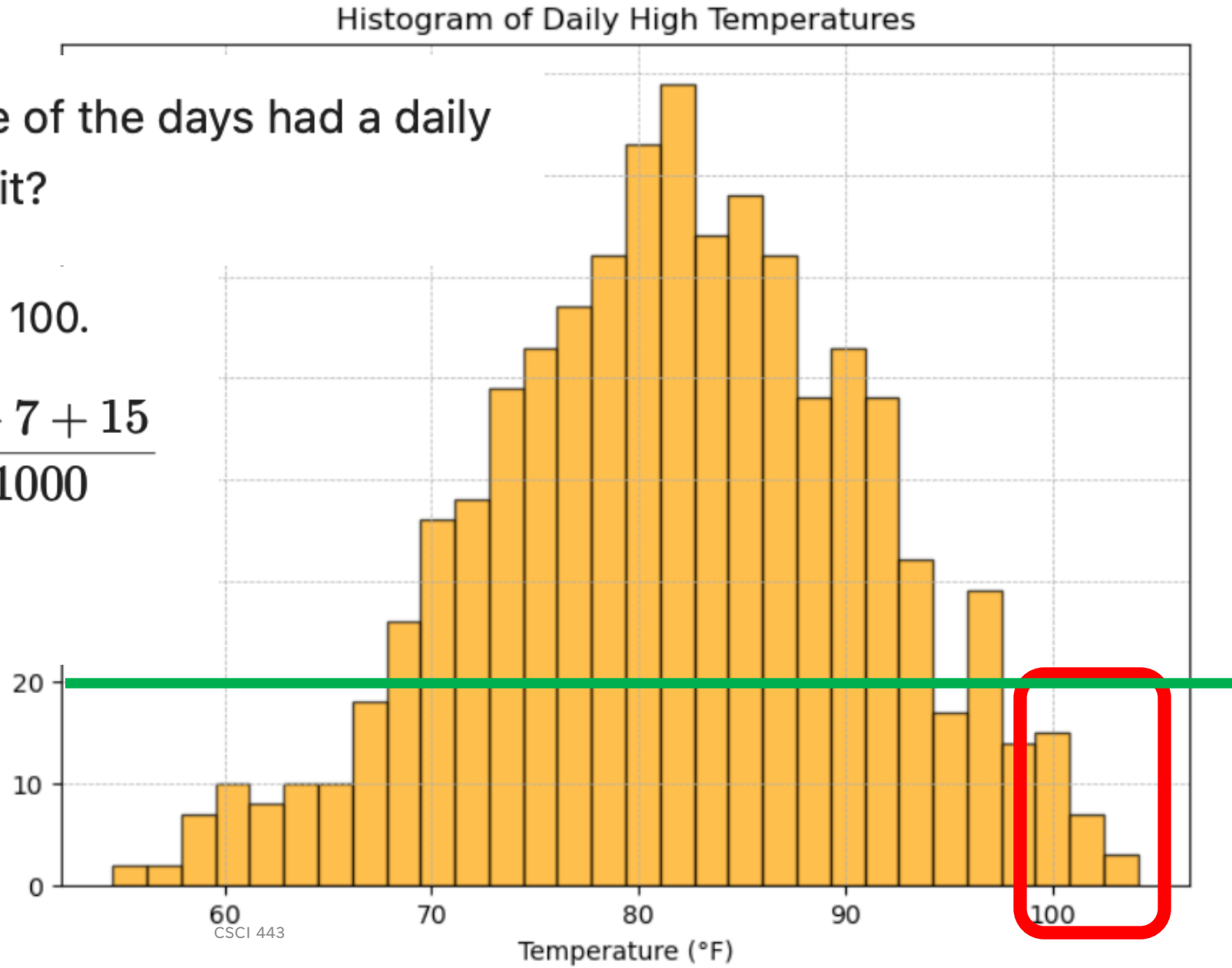
REVIEW PROBLEM 4

e) Approximately what percentage of the days had a daily high above 100 degrees Fahrenheit?

Let $p\%$ denote the percentage above 100.

$$100 \cdot \frac{2+6}{1000} \leq p\% \leq 100 \cdot \frac{3+7+15}{1000}$$

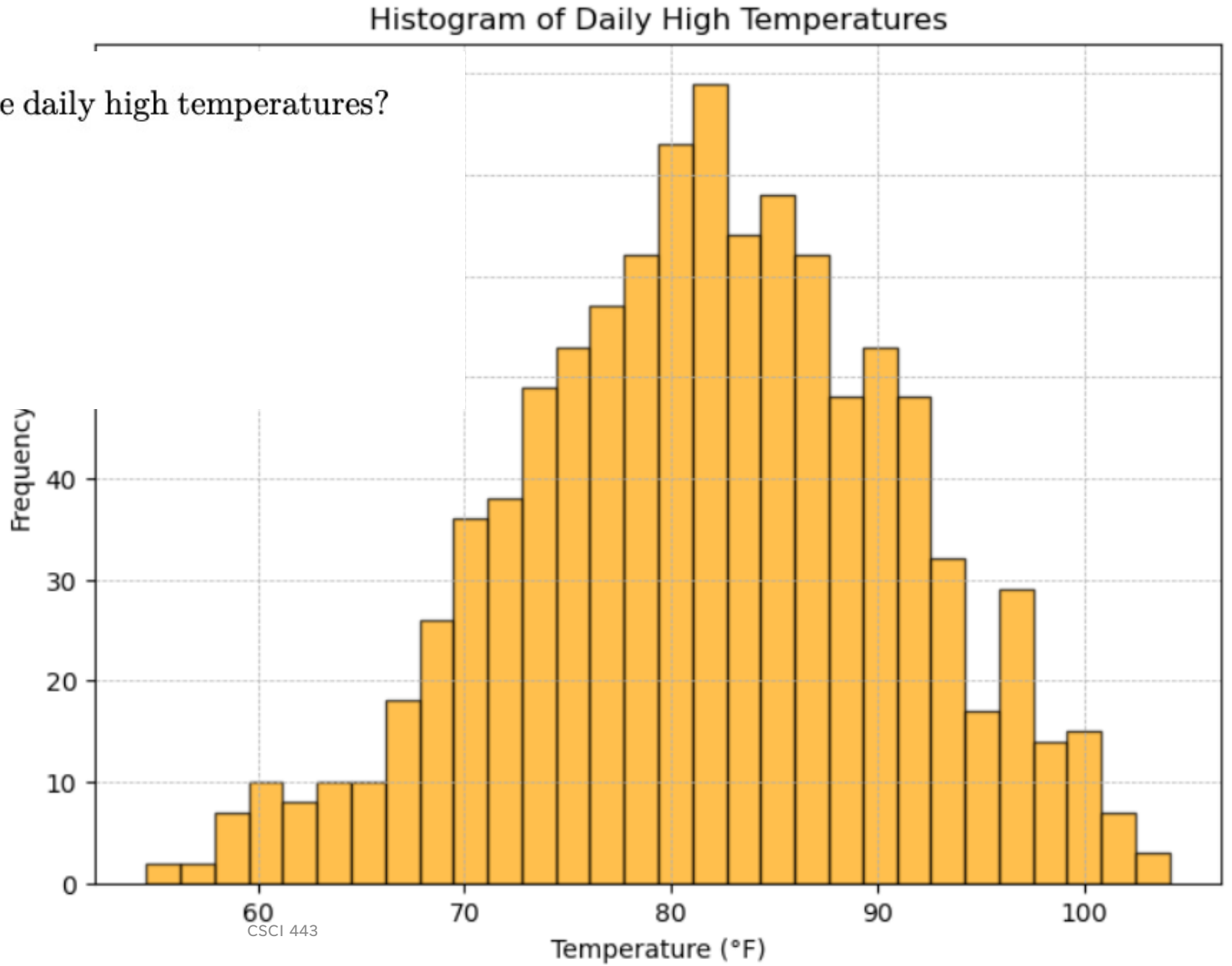
$$\boxed{0.8\% \leq p\% \leq 2.5\%}$$



REVIEW PROBLEM 4

f) What best describes the sample mean of the daily high temperatures?

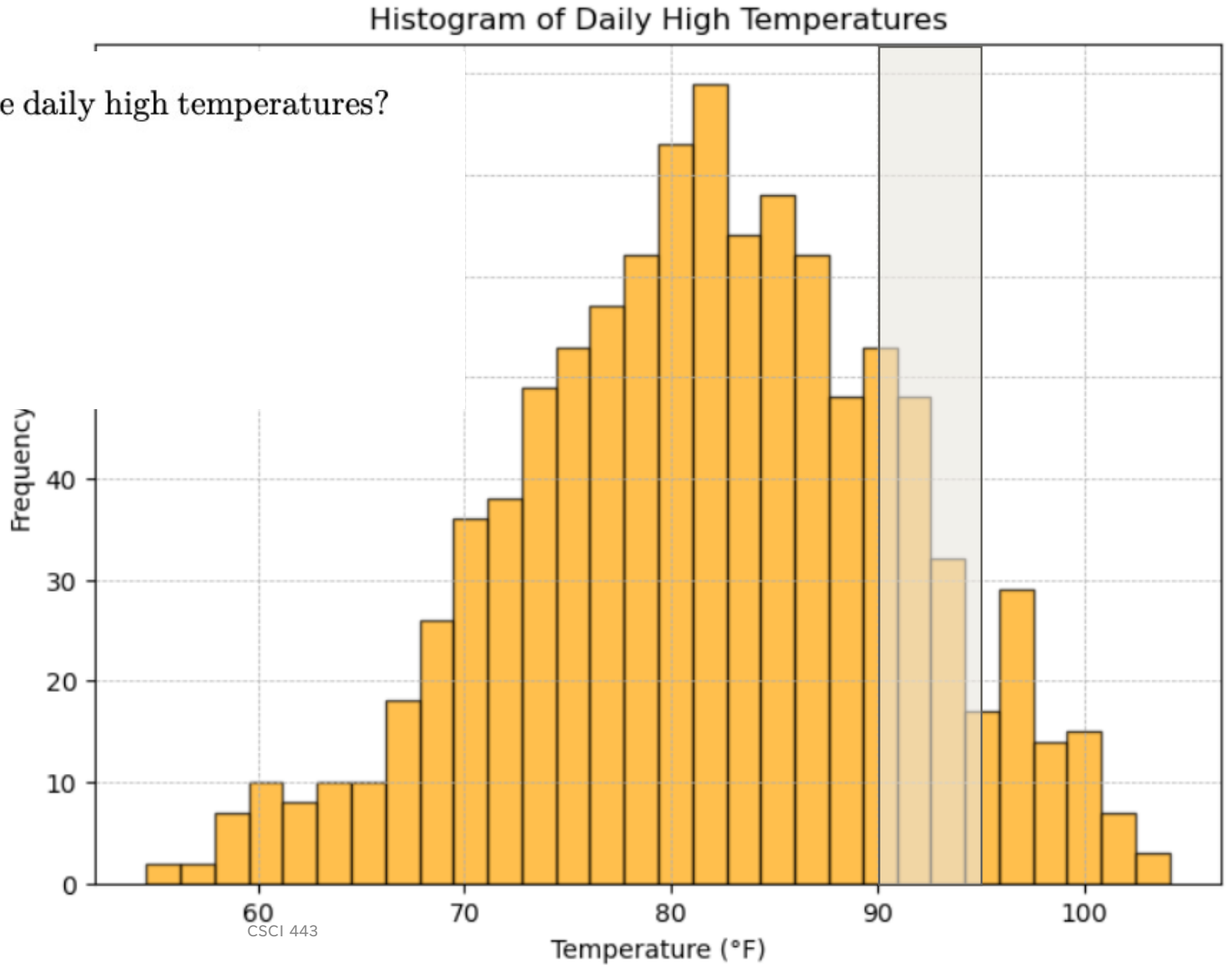
- the mean is between 80 and 85 degrees
- the mean is between 70 and 75 degrees
- the mean is between 85 and 90 degrees
- the mean is between 90 and 95 degrees.



REVIEW PROBLEM 4

f) What best describes the sample mean of the daily high temperatures?

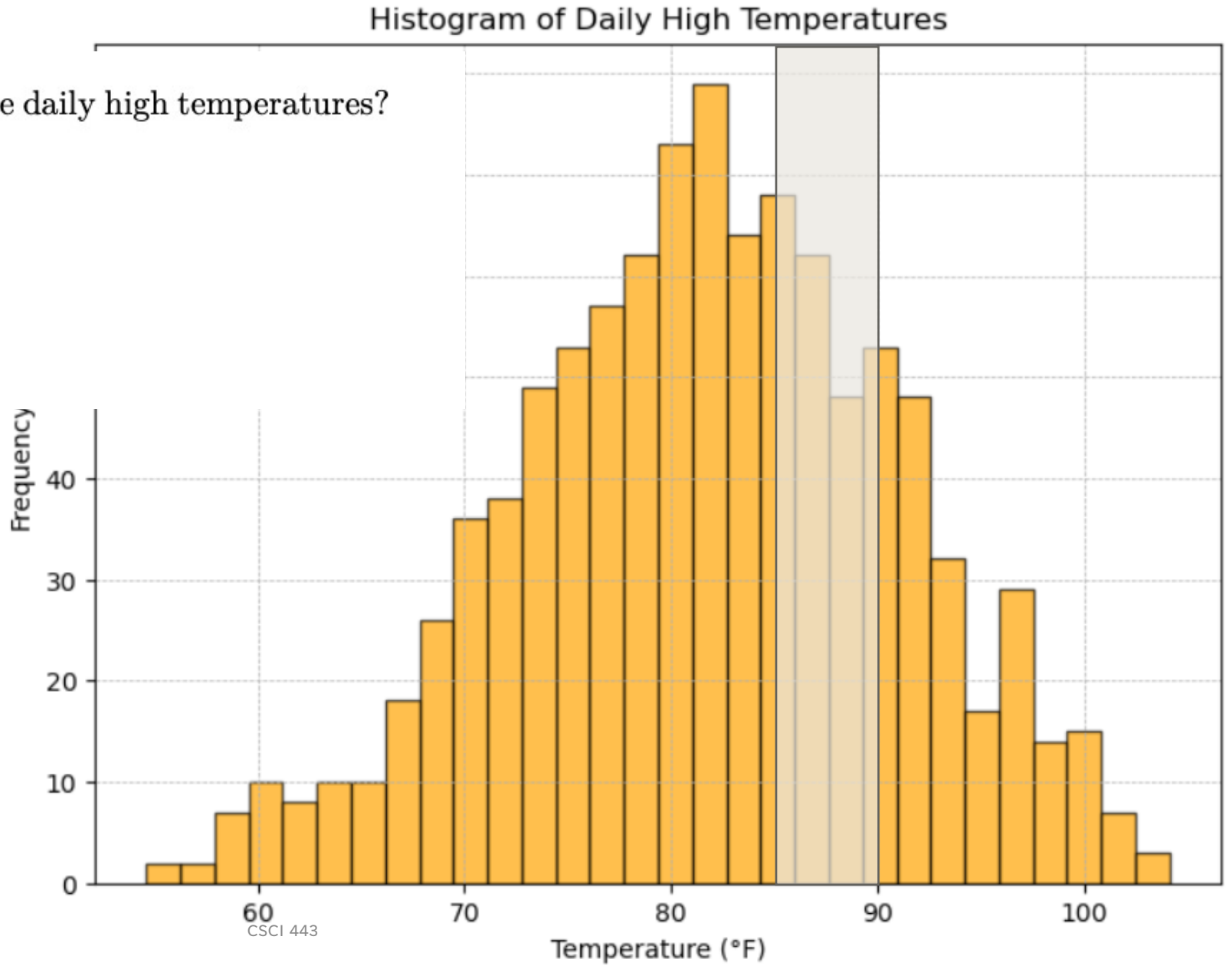
- the mean is between 80 and 85 degrees
- the mean is between 70 and 75 degrees
- the mean is between 85 and 90 degrees
- the mean is between 90 and 95 degrees.



REVIEW PROBLEM 4

f) What best describes the sample mean of the daily high temperatures?

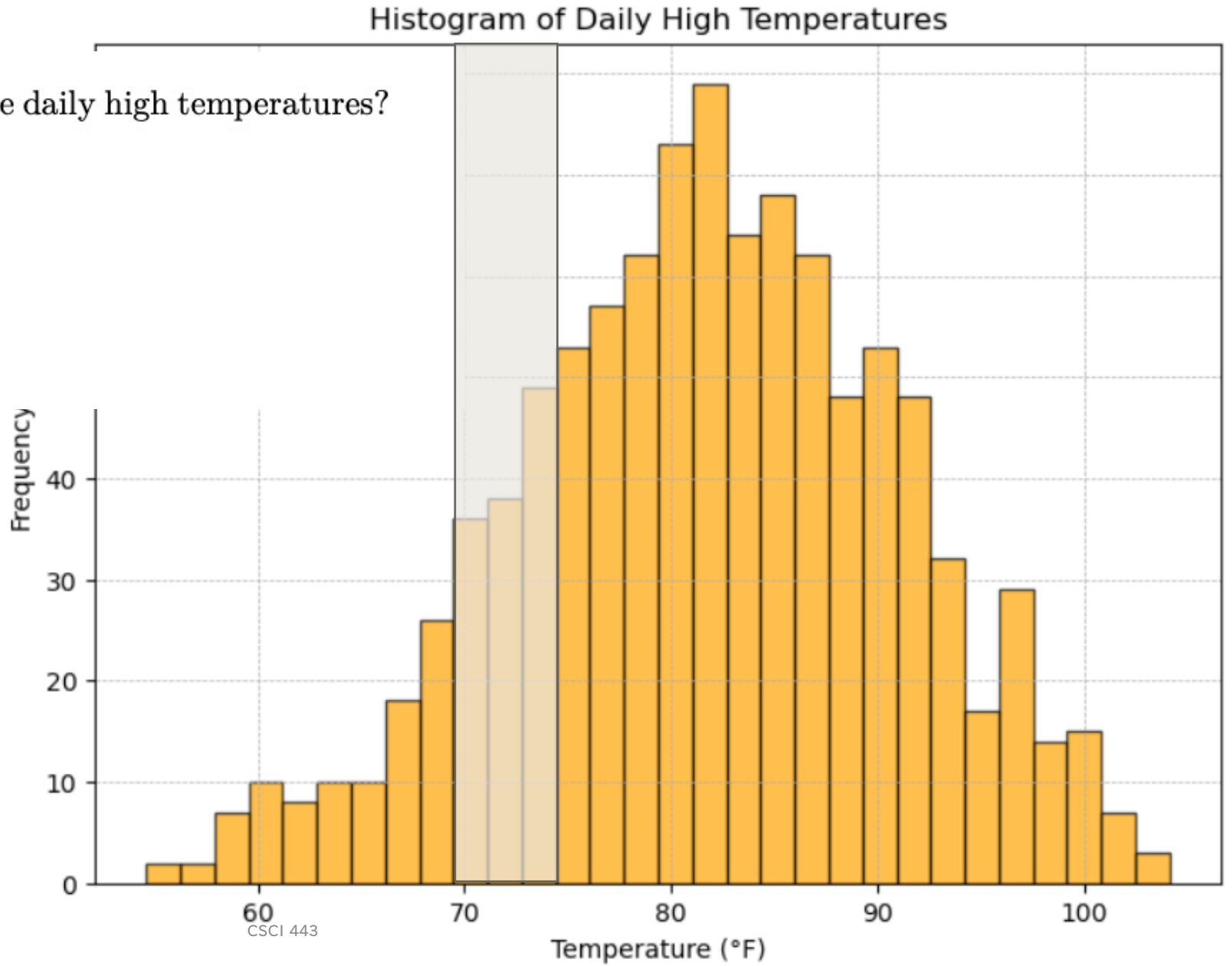
- the mean is between 80 and 85 degrees
- the mean is between 70 and 75 degrees
- the mean is between 85 and 90 degrees
- the mean is between 90 and 95 degrees.



REVIEW PROBLEM 4

f) What best describes the sample mean of the daily high temperatures?

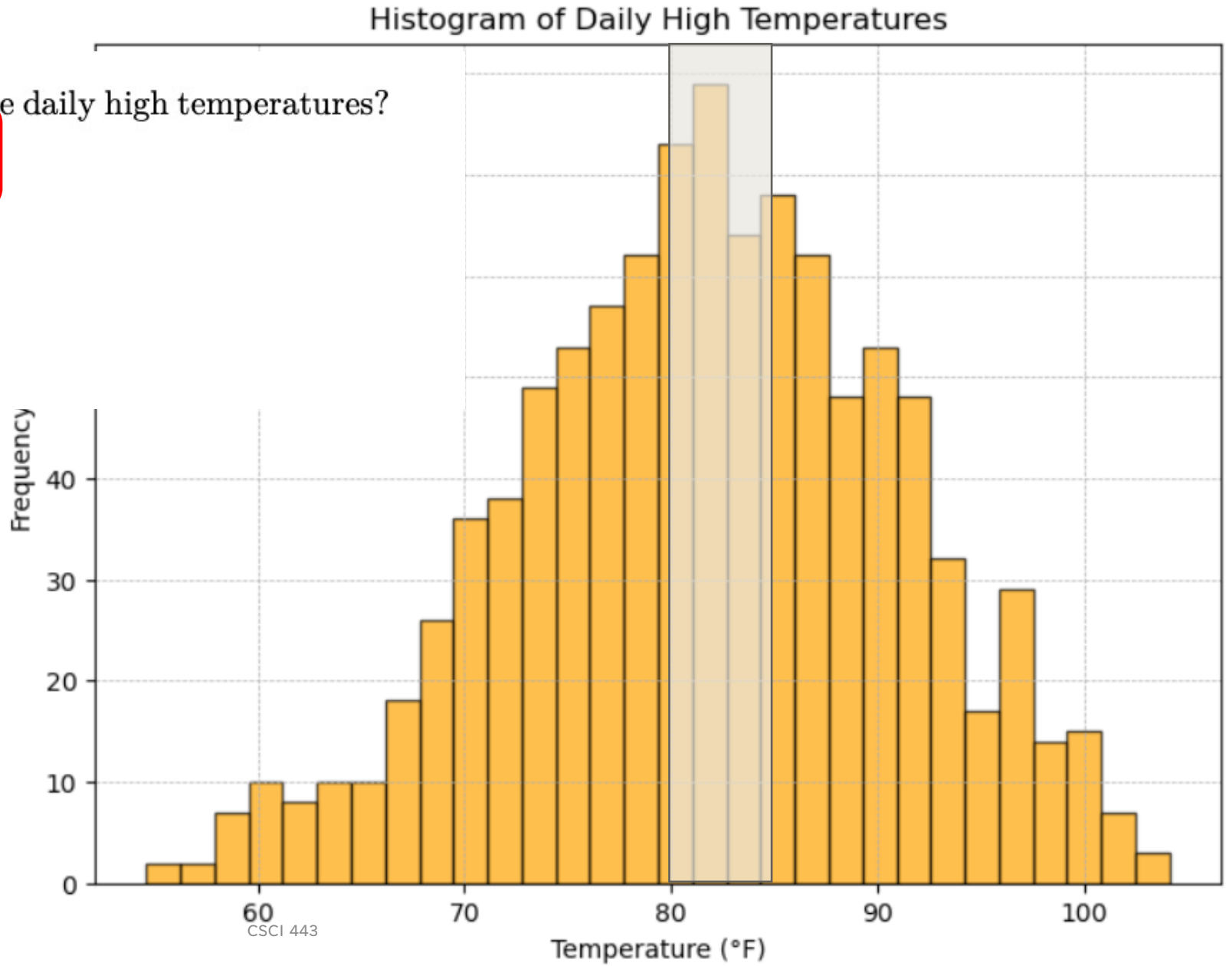
- the mean is between 80 and 85 degrees
- the mean is between 70 and 75 degrees
- the mean is between 85 and 90 degrees
- the mean is between 90 and 95 degrees.



REVIEW PROBLEM 4

f) What best describes the sample mean of the daily high temperatures?

- the mean is between 80 and 85 degrees
- the mean is between 70 and 75 degrees
- the mean is between 85 and 90 degrees
- the mean is between 90 and 95 degrees.



REVIEW PROBLEM 4

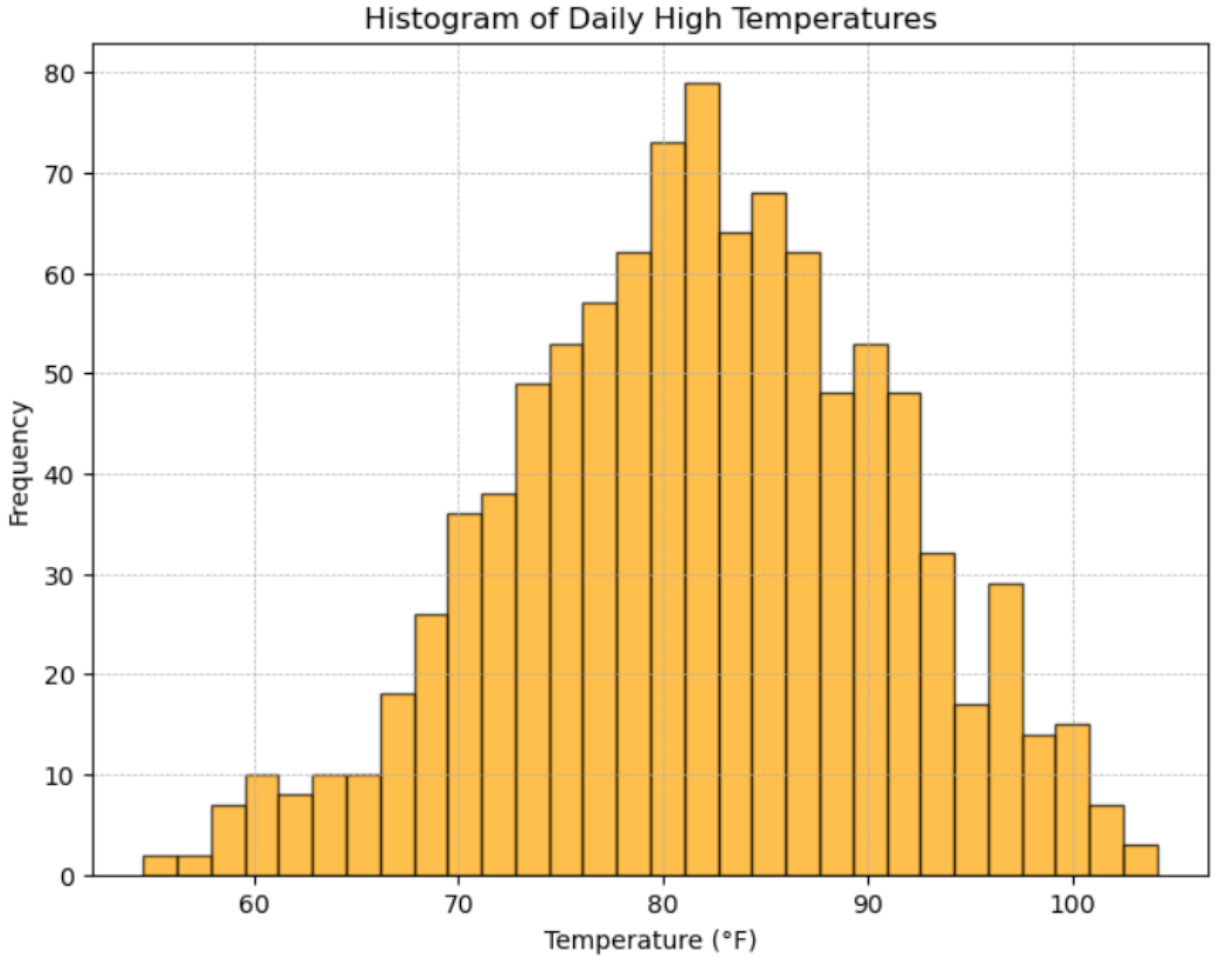
Verification (if you have time in the exam):

Sum = 0

For each bucket *j*:

1) Let x_j be the midpoint of the bucket
2) Count how many fell the jth bucket histogram bucket.
3) For each sample that fell in the bucket add x_j to the sum.

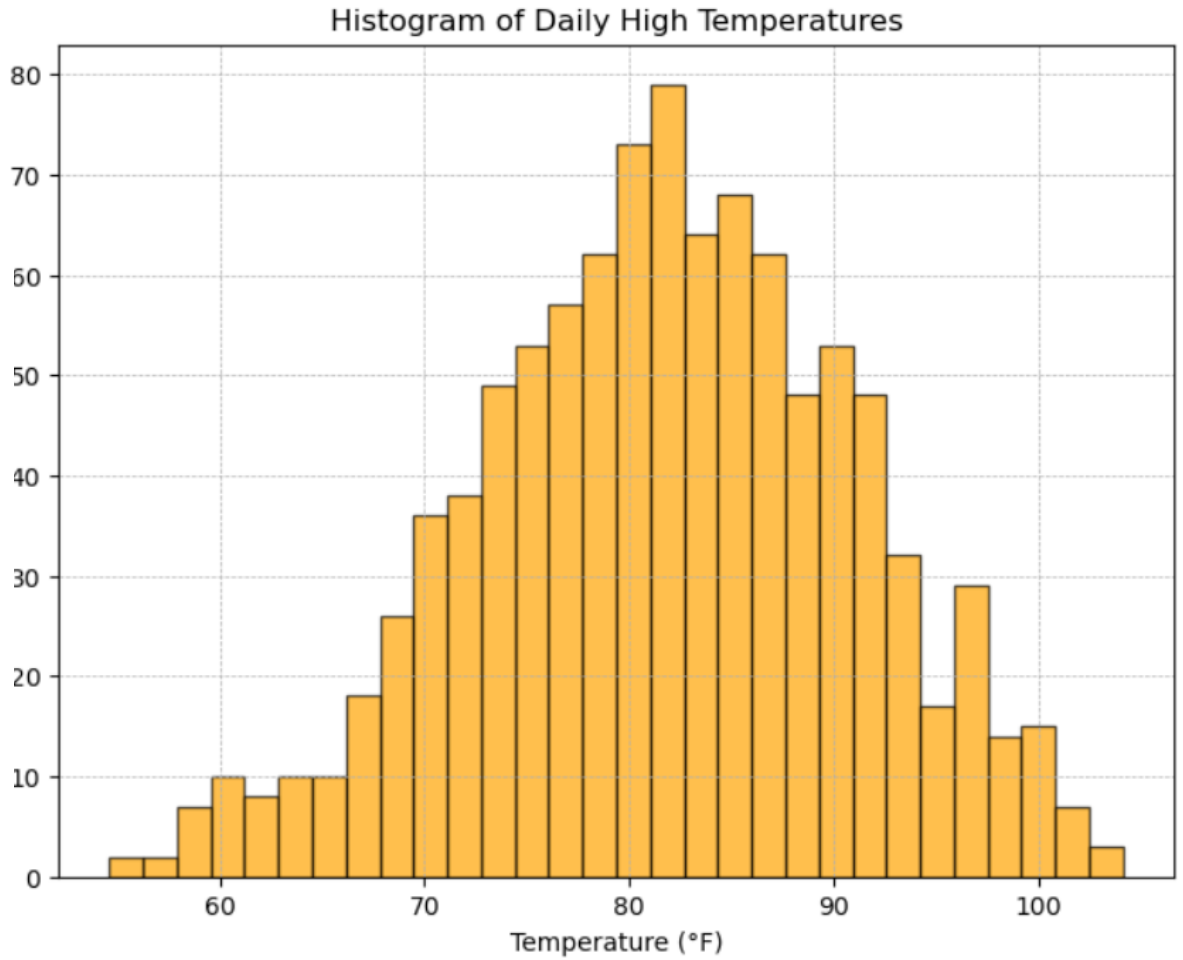
Divide by n.



$$\overline{x} \approx \frac{1}{n}\left((55+55)+(57+57)+(58+\cdots+58)+\cdots\right)$$

REVIEW PROBLEM 4

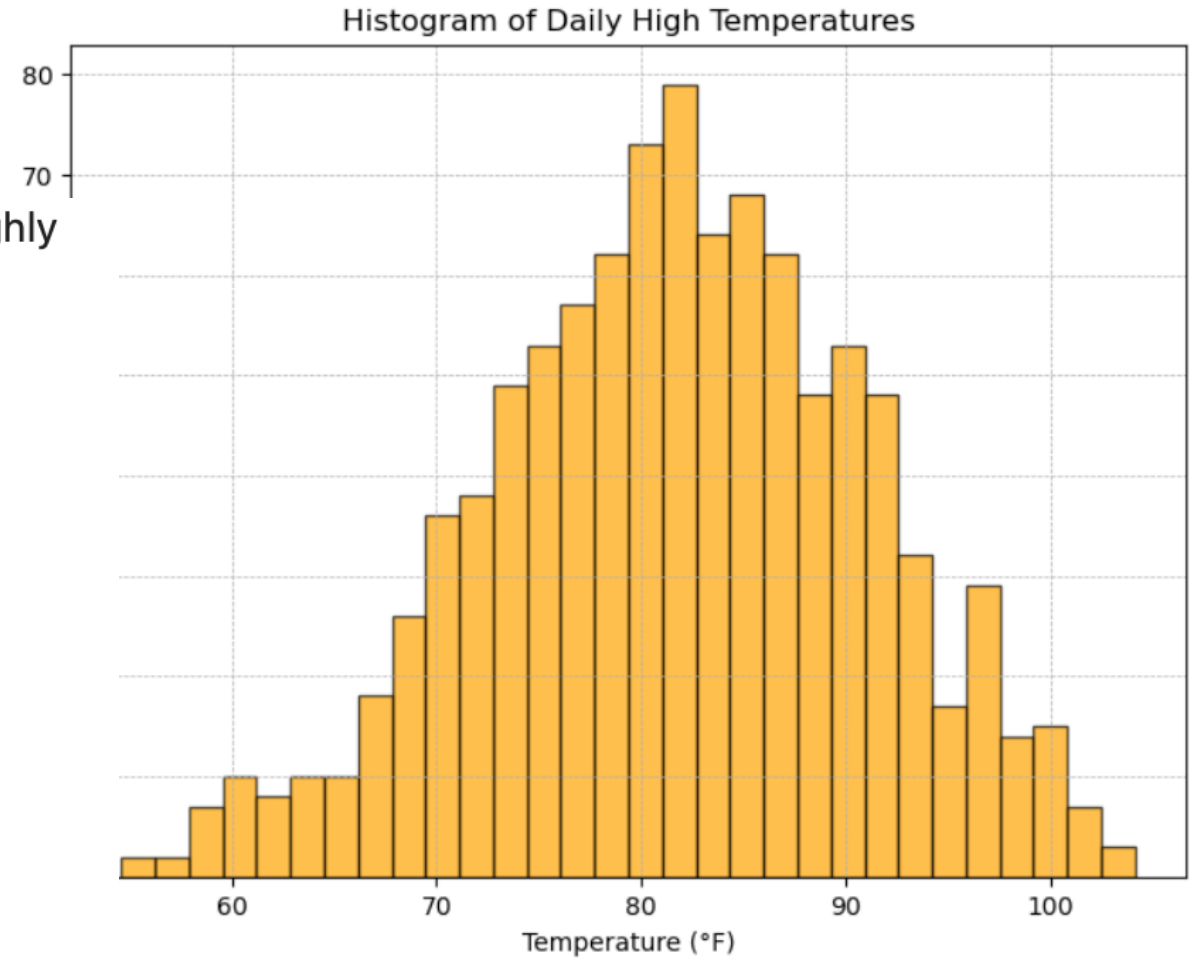
```python
plt.figure(figsize=(8, 6))
n, bins, patches = plt.hist(temperatures, bins=30,
                            alpha=0.7, color='orange',
                            edgecolor='black')
# estimating mean from bins.
# . n contains the heights of the bins (frequencies)
# . bins contains the edges of the bins
sum = 0
sumn = 0
for nj, bj in zip(n, bins):
    sum += nj * bj
    sumn += nj
```



$$\frac{1}{1000}(2 \cdot 55 + 2 \cdot 57 + 7 \cdot 58 + 10 \cdot 61 + \cdots + +7 \cdot 102 + 3 \cdot 103)$$

REVIEW PROBLEM 4: FAST ESTIMATE

Group buckets then compute weighted mean of the groups.



In the above plot we could leftmost group would span roughly from 55 to 63.

$$x_{g1} = \frac{55 + 63}{2} = 59$$

The number of samples falling in these 5 buckets is approximately

$$2 + 2 + 7 + 10 + 8 = 29$$

Perform weighted sum

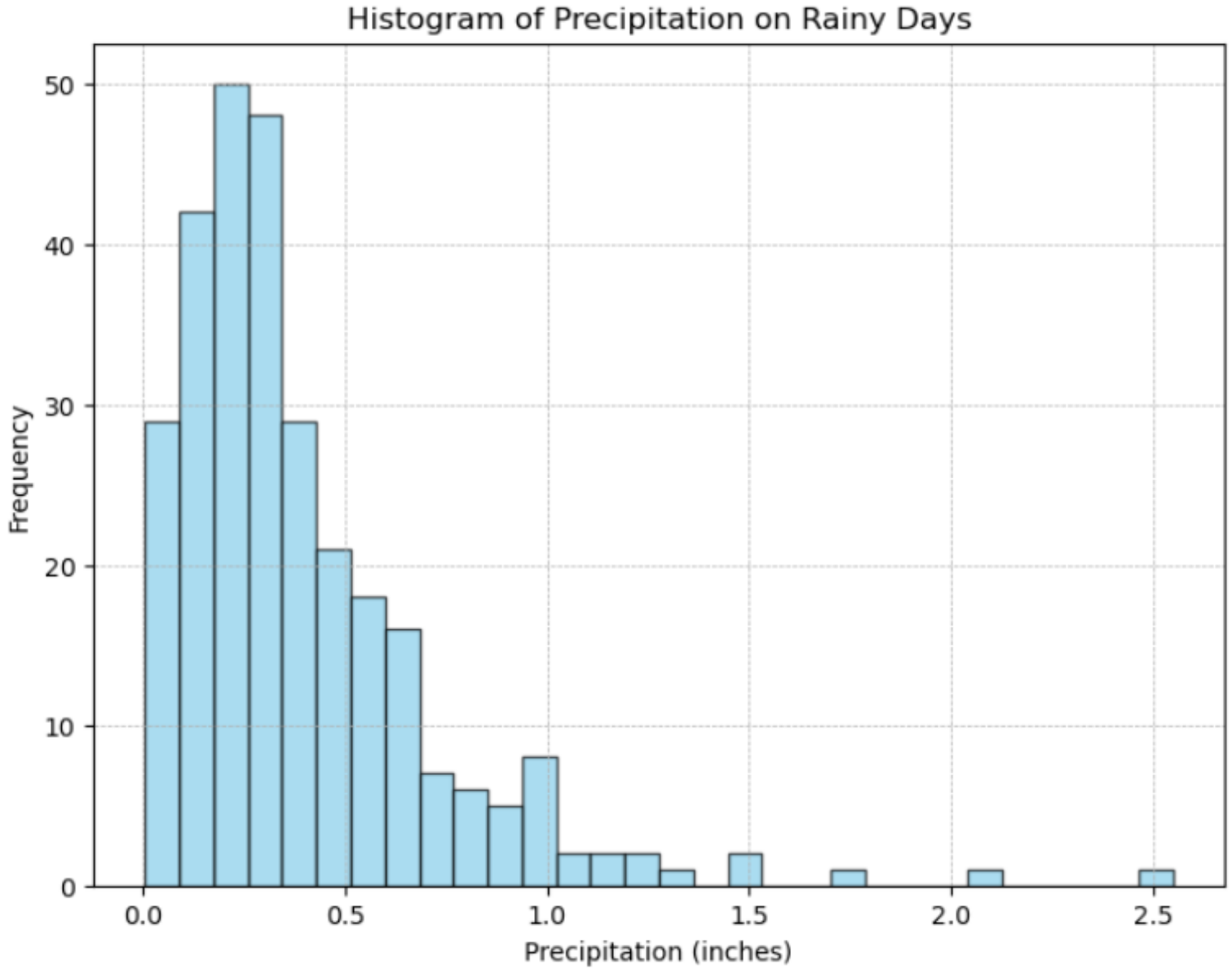
$$\frac{1}{n} \sum_{j=1}^{n_g} n_{gj} x_{gj} = (29 \cdot 59 + \cdots) \approx 81.7$$

Actual sample mean?

$$\overline{x} = 81.662$$

REVIEW PROBLEM 6

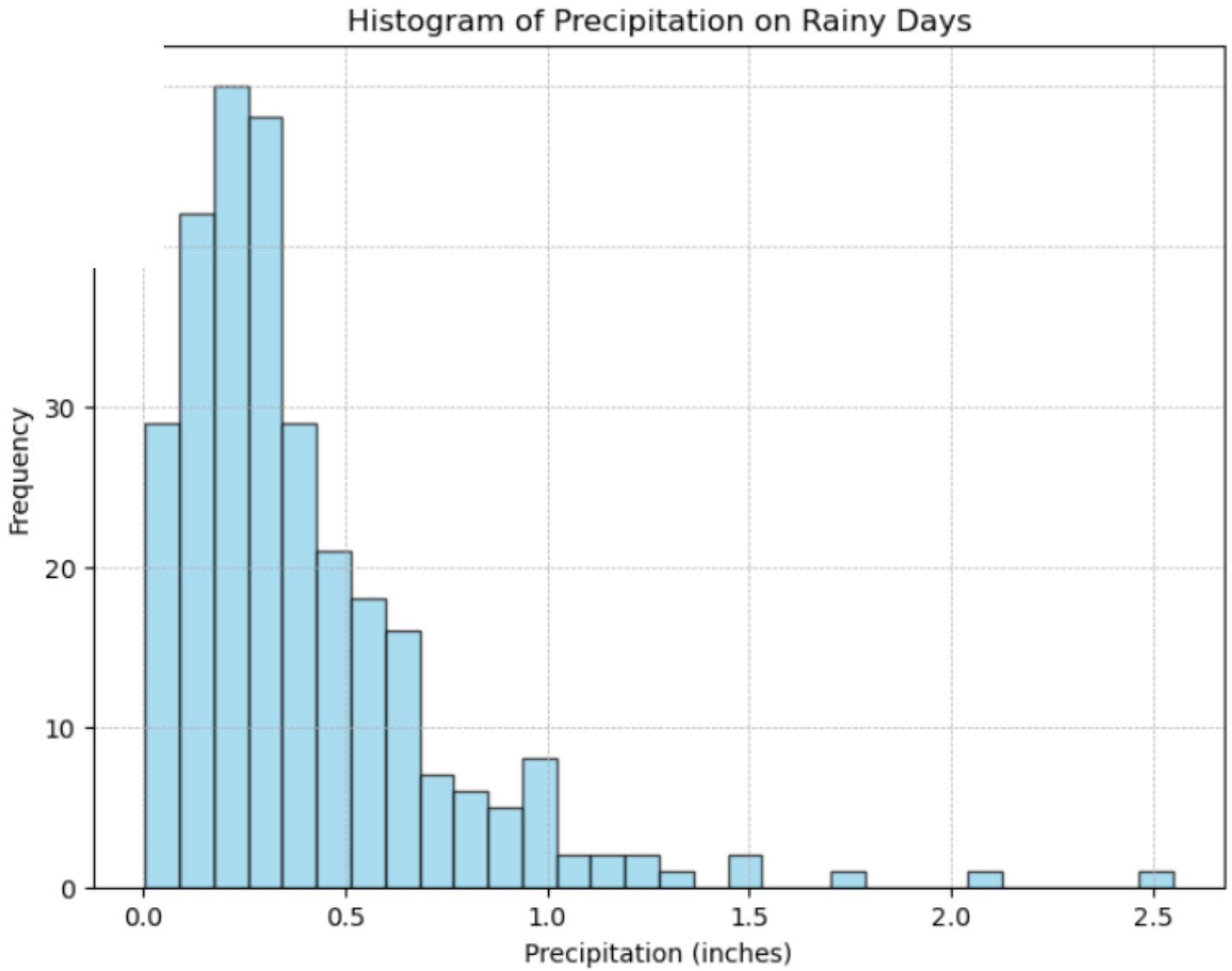
Skewed? Which direction?



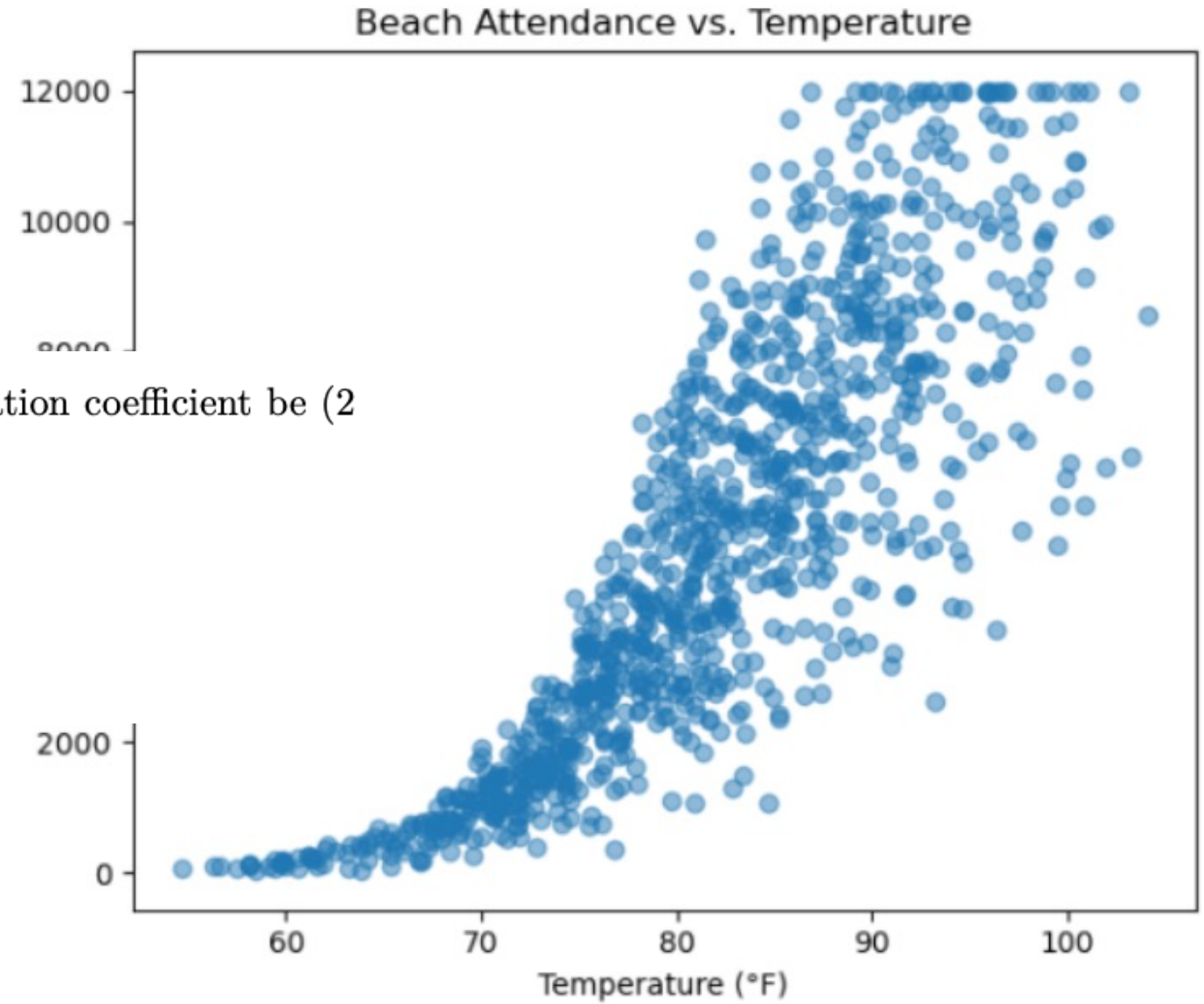
REVIEW PROBLEM 6

c) How many days in the observed period had more than 2 inches of rain? (1 point)

- Zero
- More than 20
- More than 5
- Between 0 and 5



REVIEW PROBLEM 7



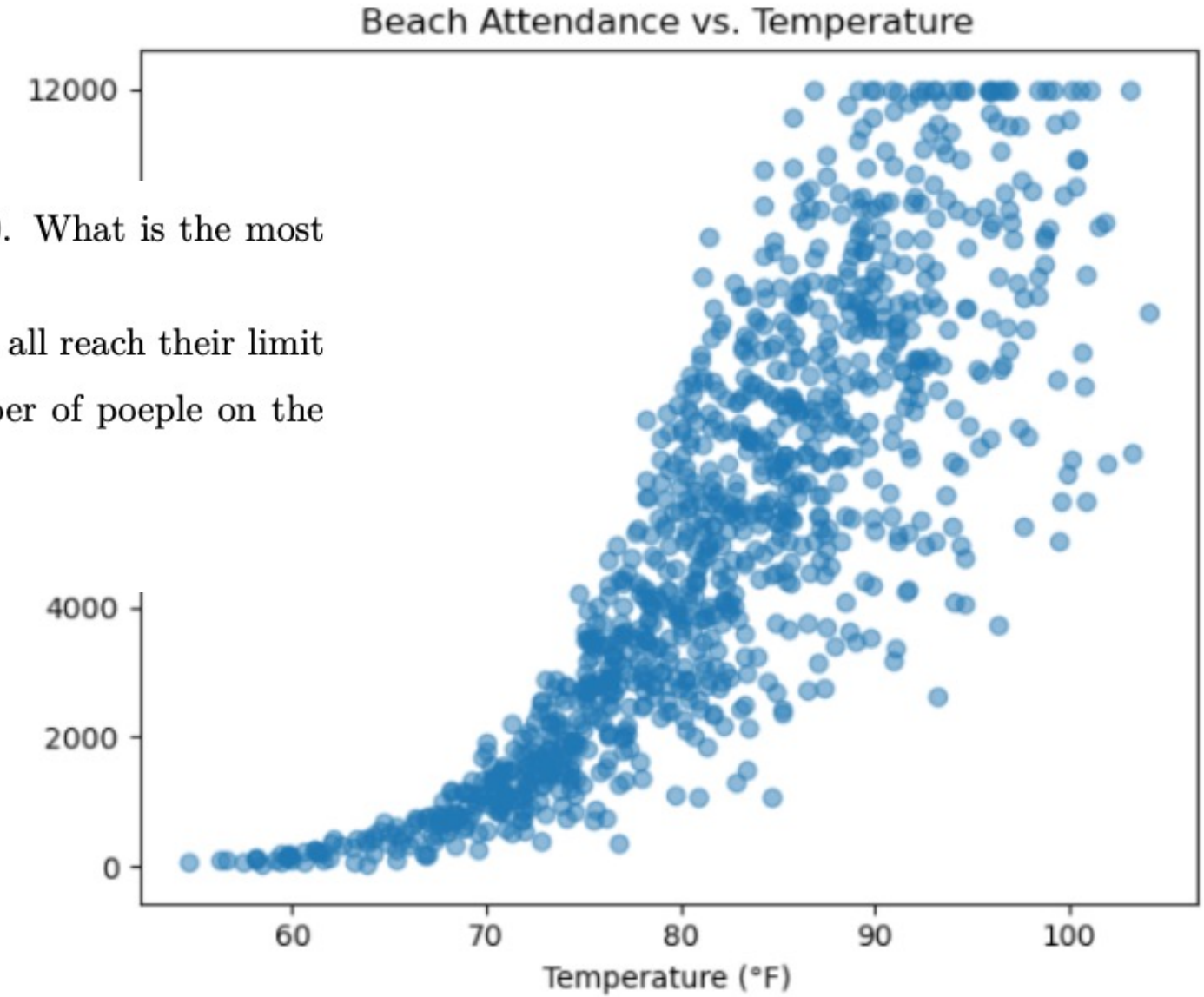
c) For this distribution would the Pearson correlation coefficient be (2 points)

– positive

– negative

– near zero

REVIEW PROBLEM 7



b) The number seems to peak or saturate at 12000. What is the most reasonable explanation? (1 point)

– people don't like large crowds and at 12000 they all reach their limit

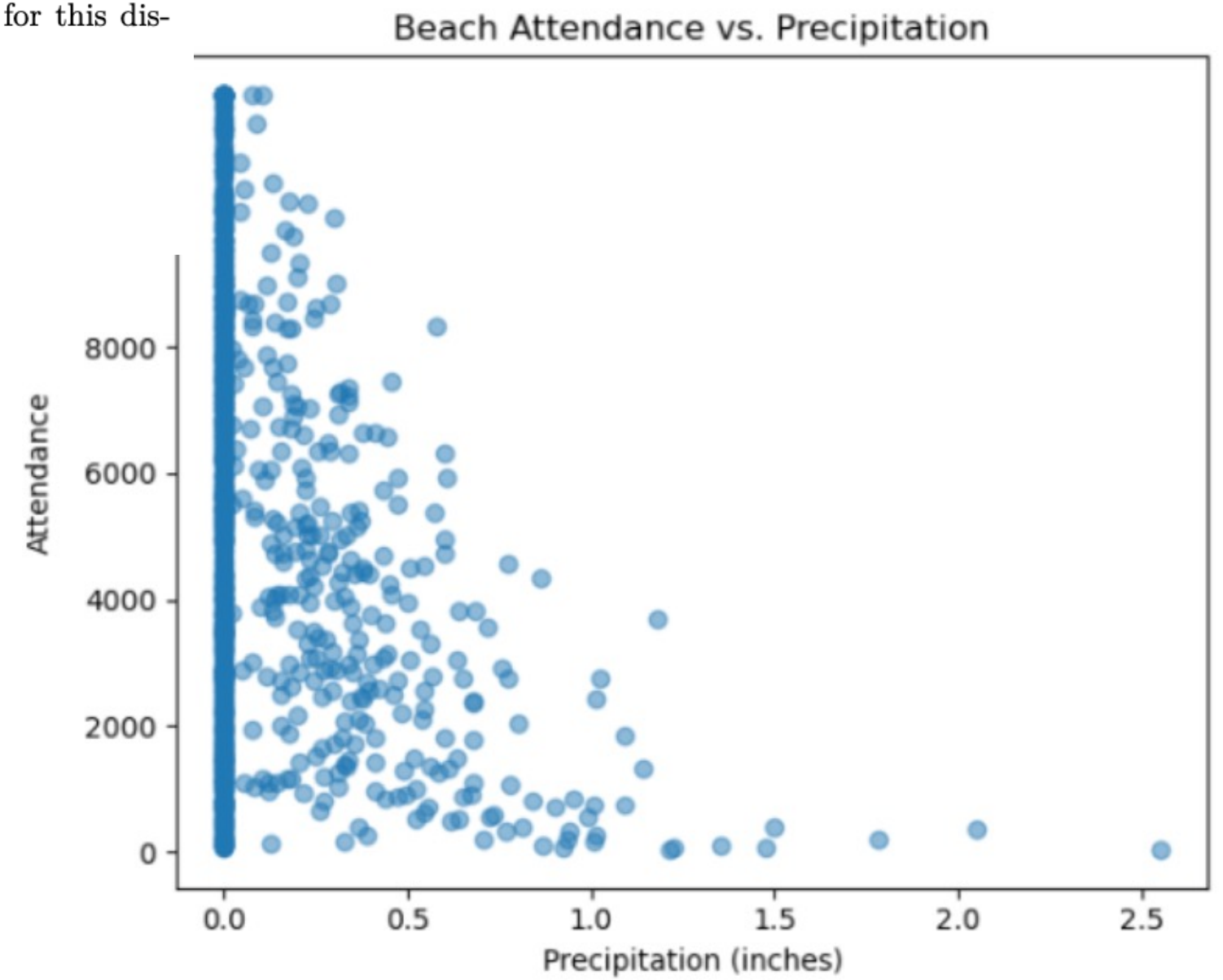
– some limitation is imposed restricting the number of poeple on the beach.

– this is just an artifact of randomness in nature

REVIEW PROBLEM 7

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

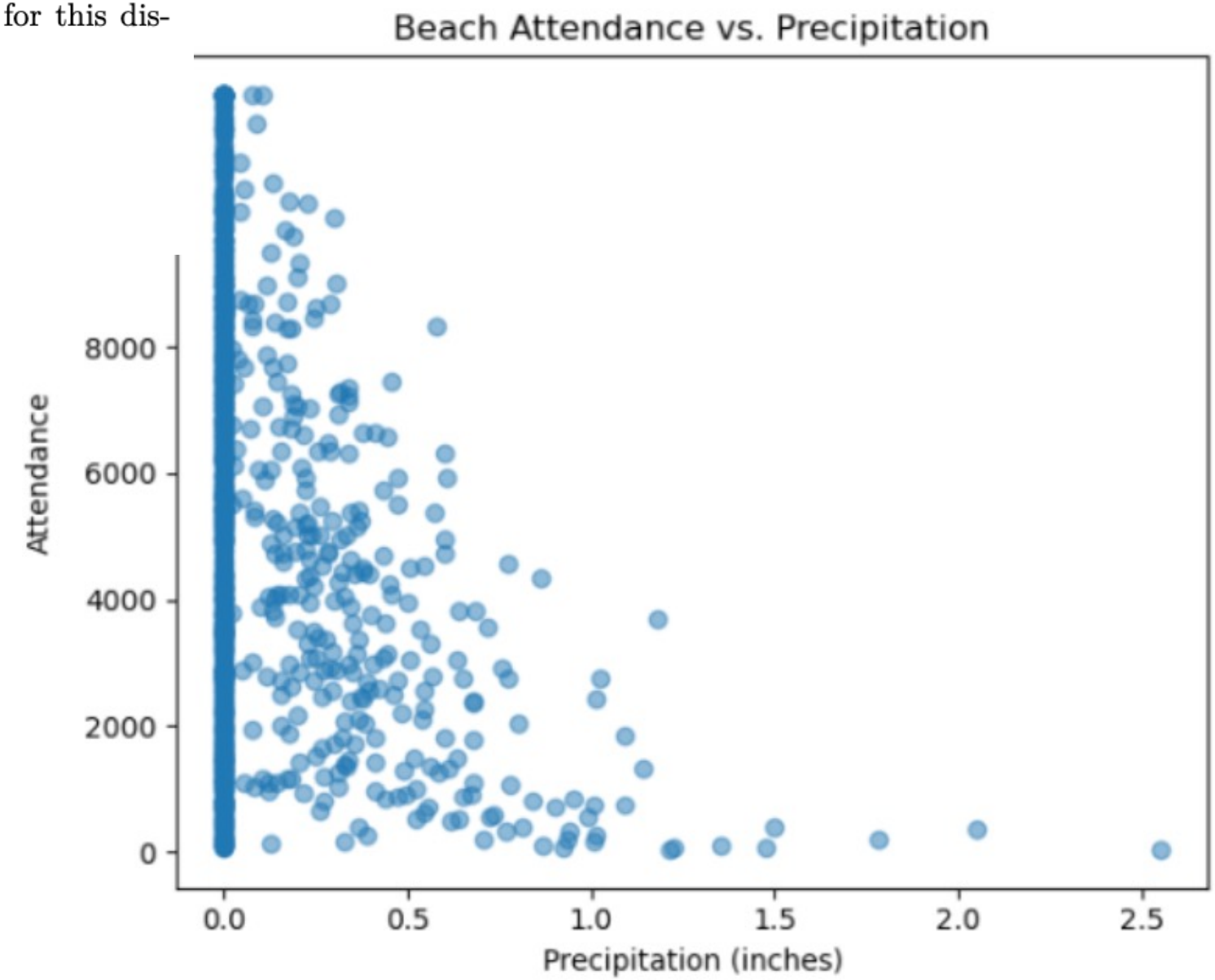
- Negative
- Near zero
- Positive



REVIEW PROBLEM 7

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

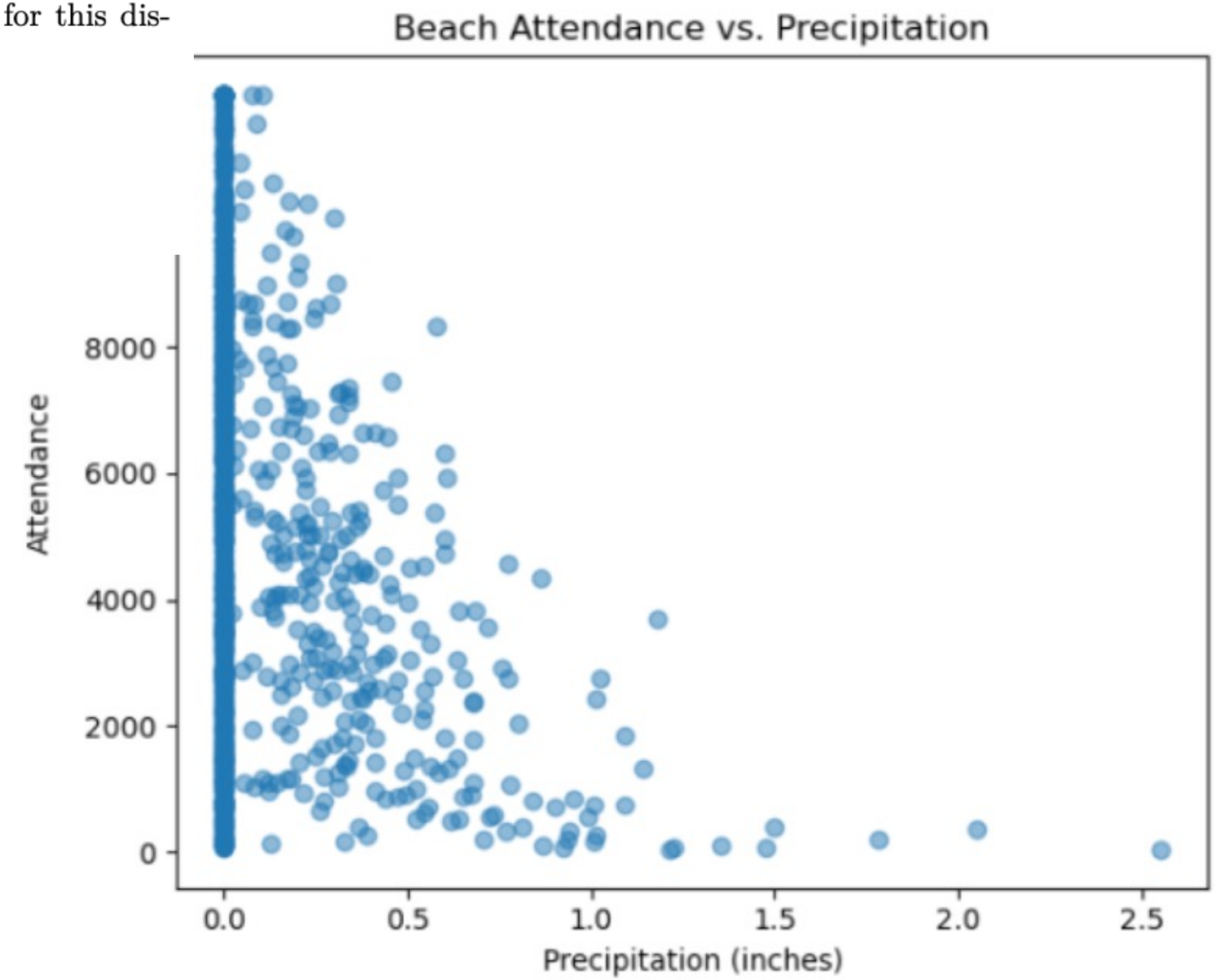
- Negative
- Near zero
- Positive



REVIEW PROBLEM 7

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

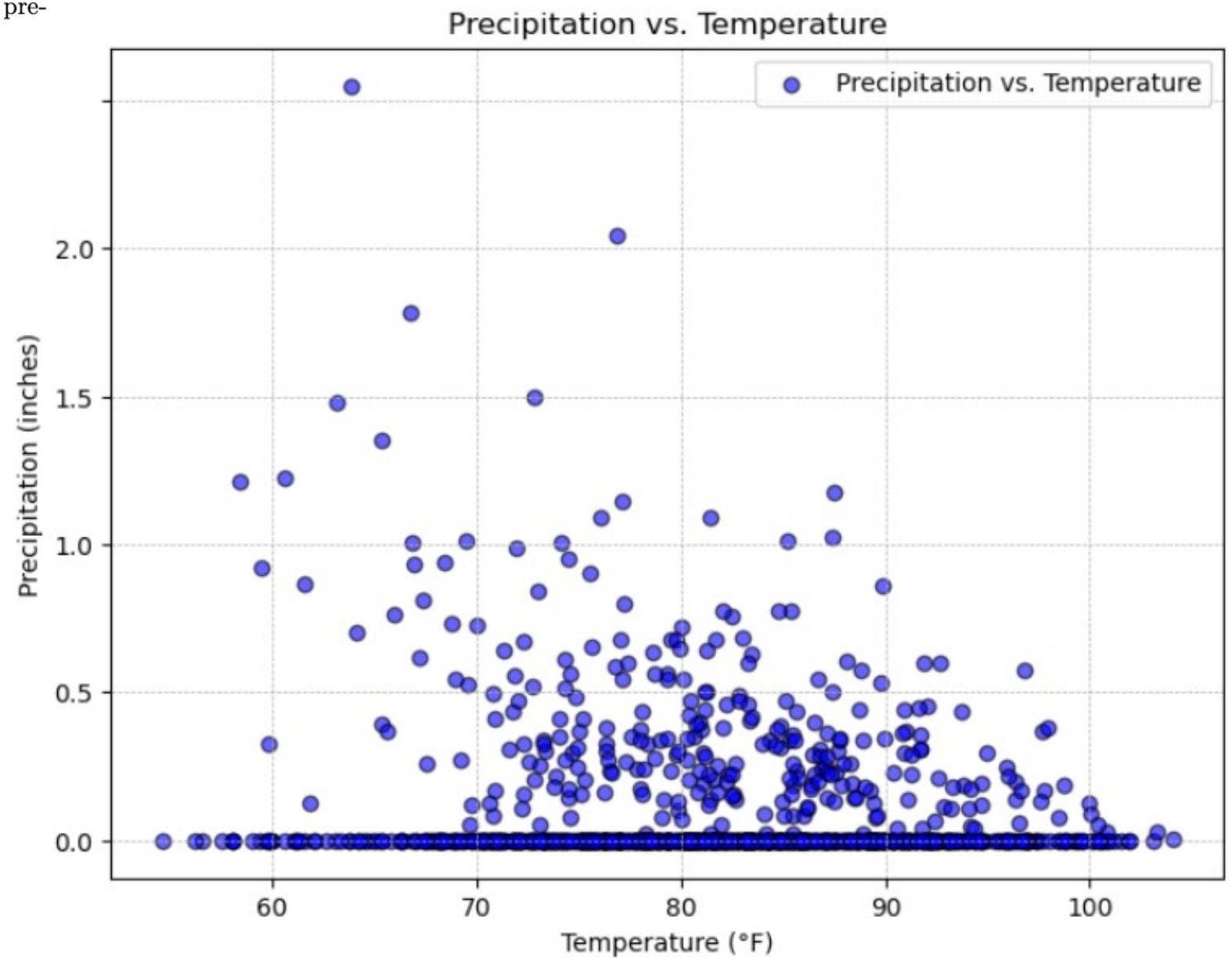
- Negative
- Near zero
- Positive



REVIEW PROBLEM 8

a) What best describves the relationship between temperature and precipitation? (2 points)

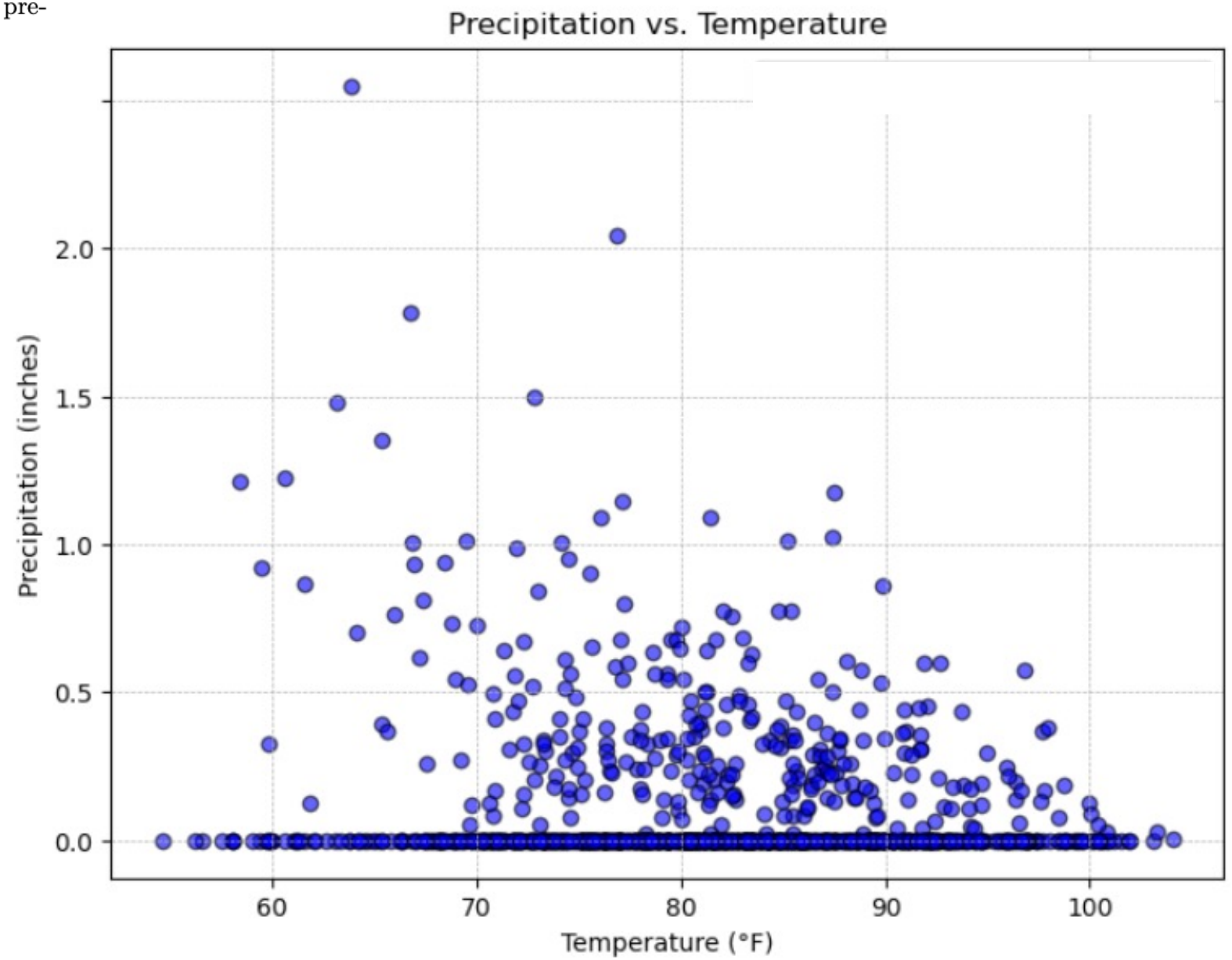
- The amount of precipitation increases sharply with temperature
- The amount of precipitation diminsihes with temperature
- Precipitation and temperature are unrelated.



REVIEW PROBLEM 8

a) What best describves the relationship between temperature and precipitation? (2 points)

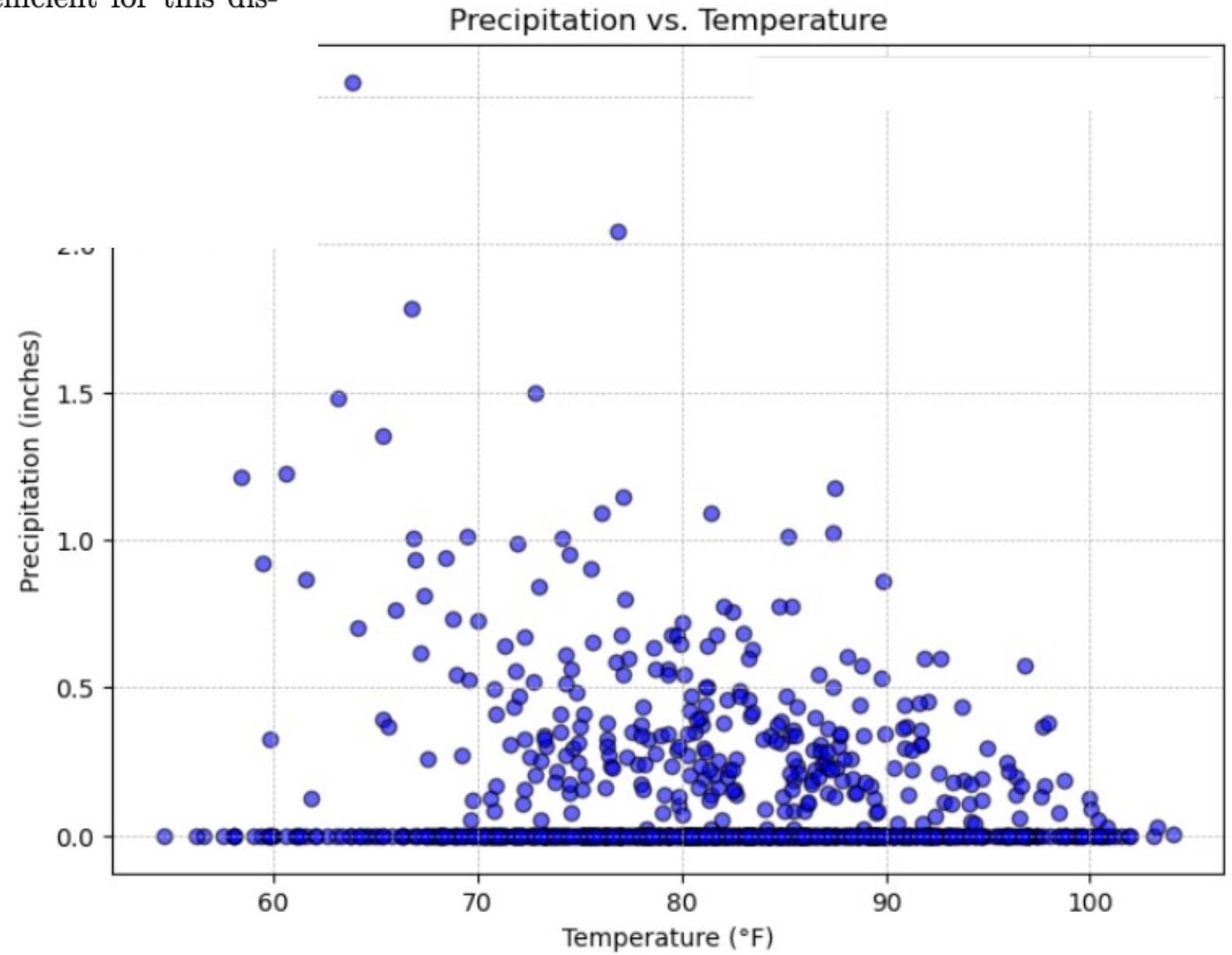
- The amount of precipitation increases sharply with temperature
- The amount of precipitation diminsihes with temperature
- Precipitation and temperature are unrelated.



REVIEW PROBLEM 8

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

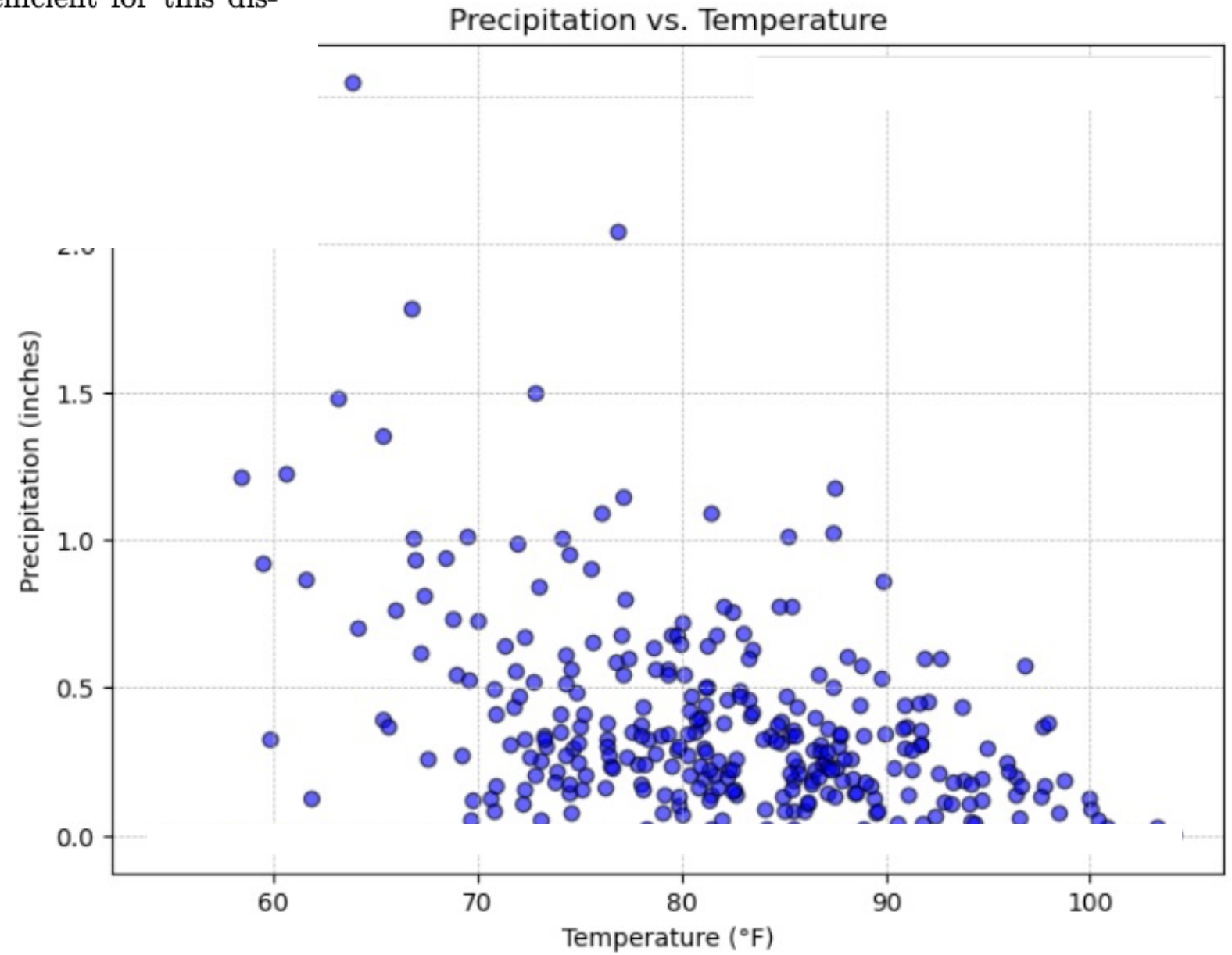
- Negative
- Near zero
- Positive



REVIEW PROBLEM 8

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

- Negative
- Near zero
- Positive



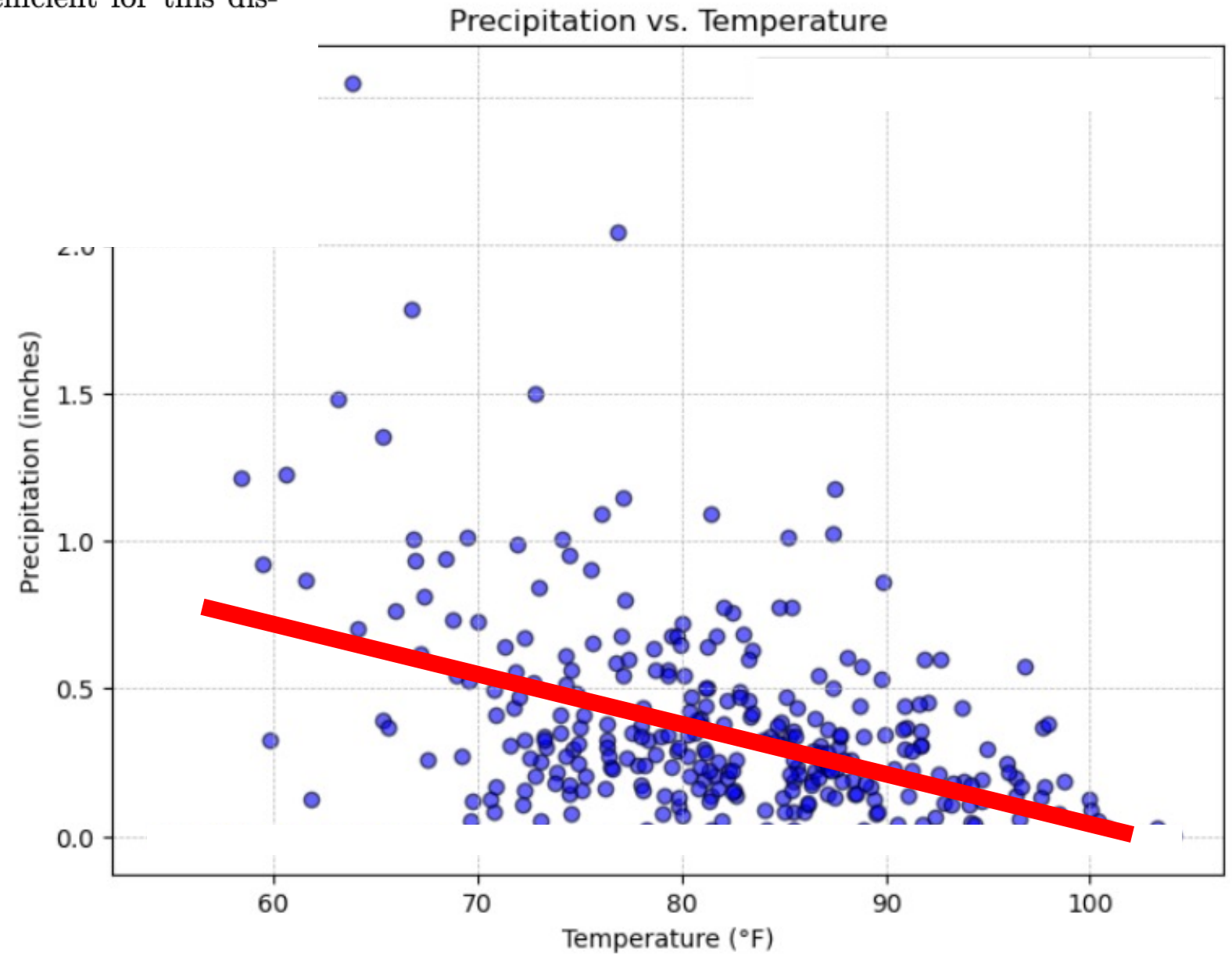
REVIEW PROBLEM 8

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

- Negative
- Near zero
- Positive

Pearson correlation coefficient.

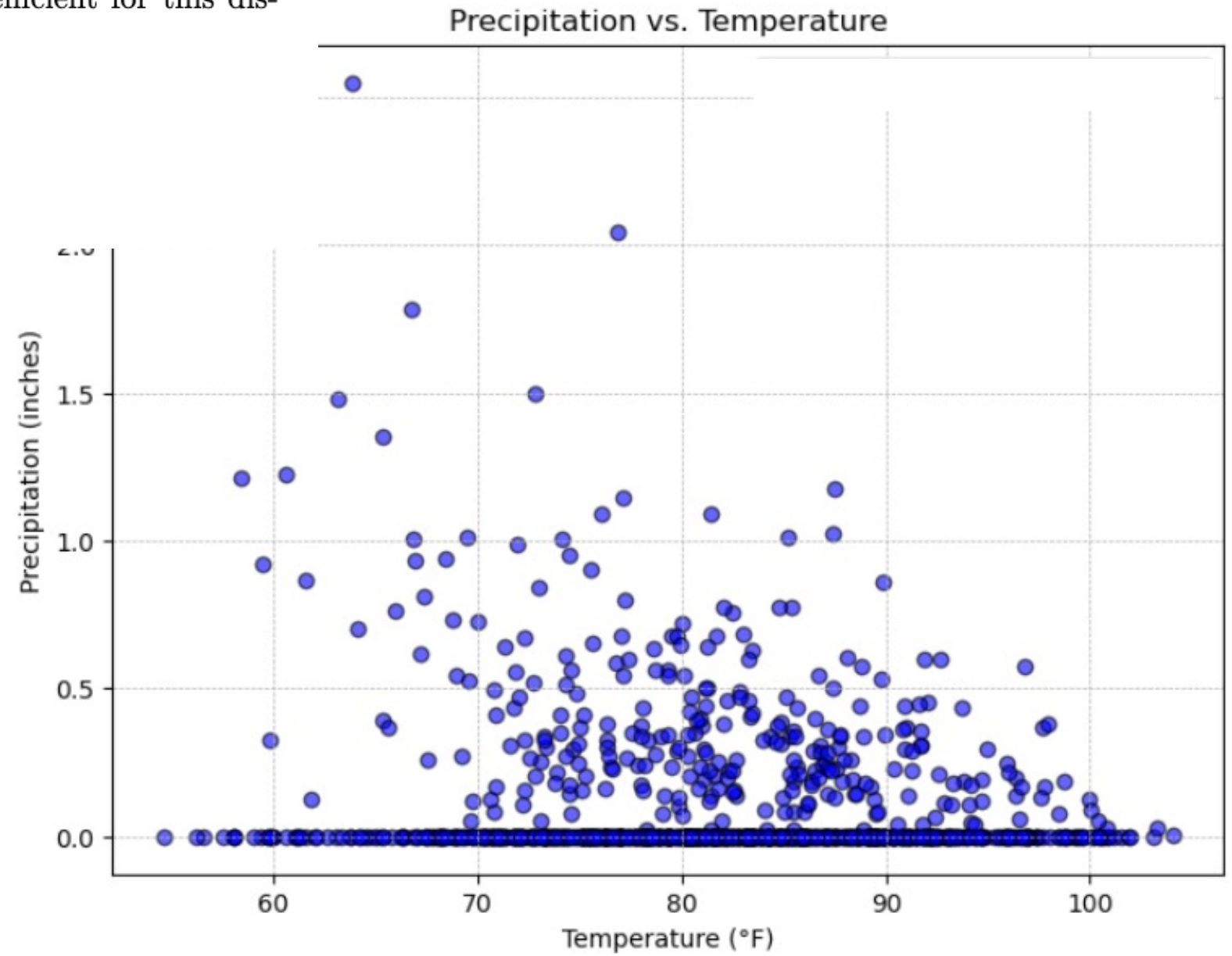
```
-0.4356766085370711
```



REVIEW PROBLEM 8

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

- Negative
- Near zero
- Positive

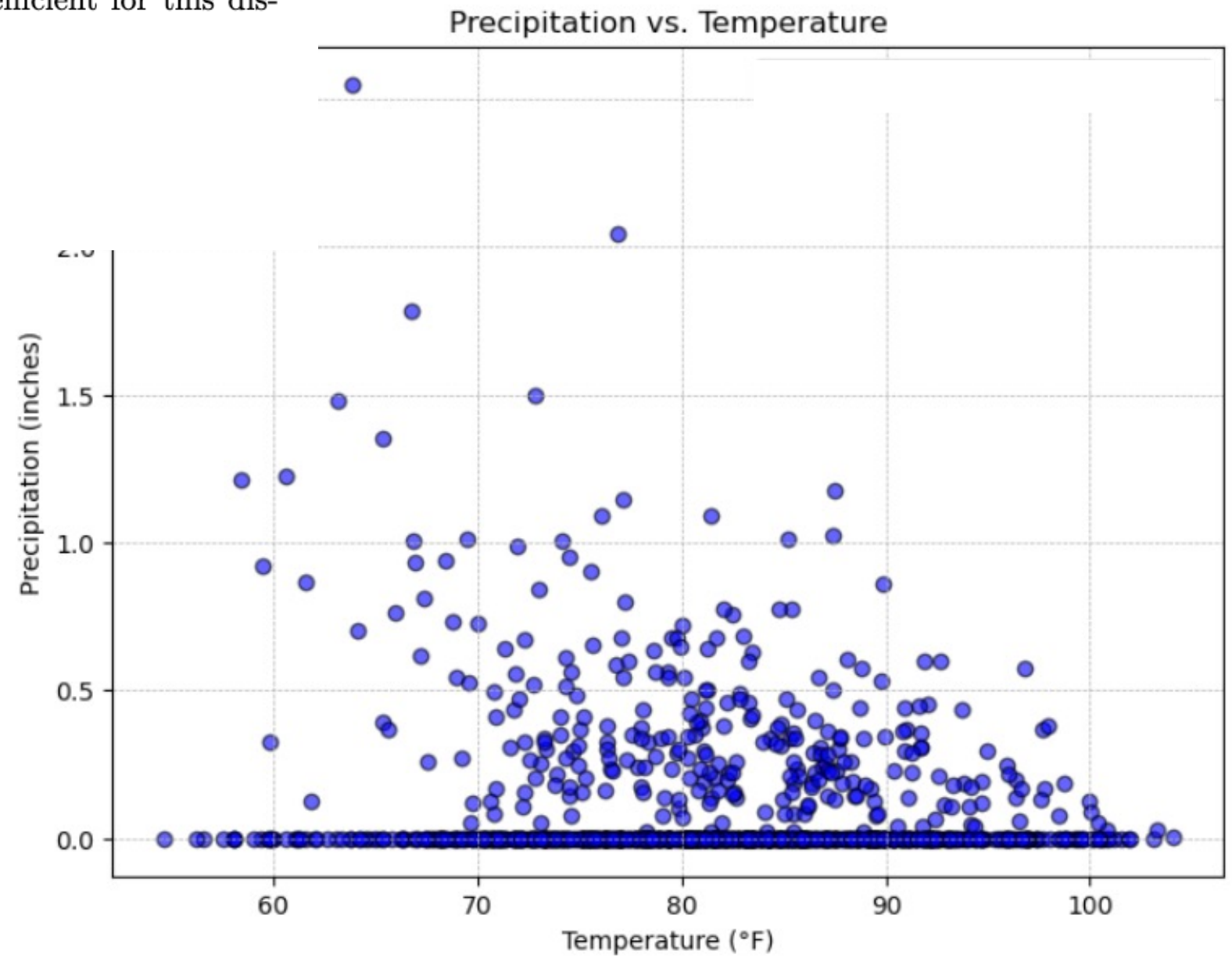


REVIEW PROBLEM 8

b) What best describes the Pearson correlation coefficient for this distribution? (2 points)

- Negative
- Near zero
- Positive

Approximately 70% of the samples fall on the horizontal



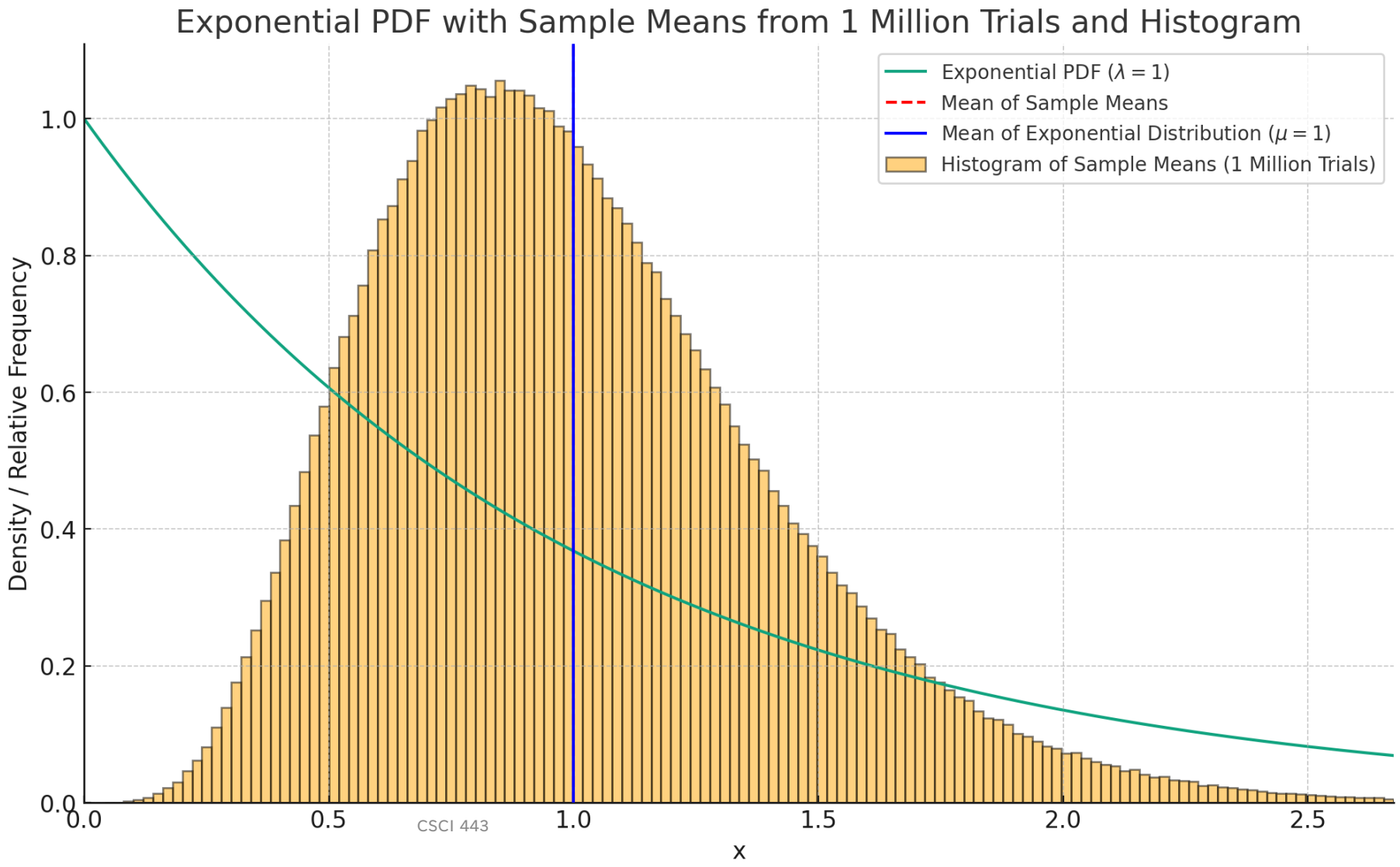
SAMPLE MEAN IS ALSO RANDOM

n=6

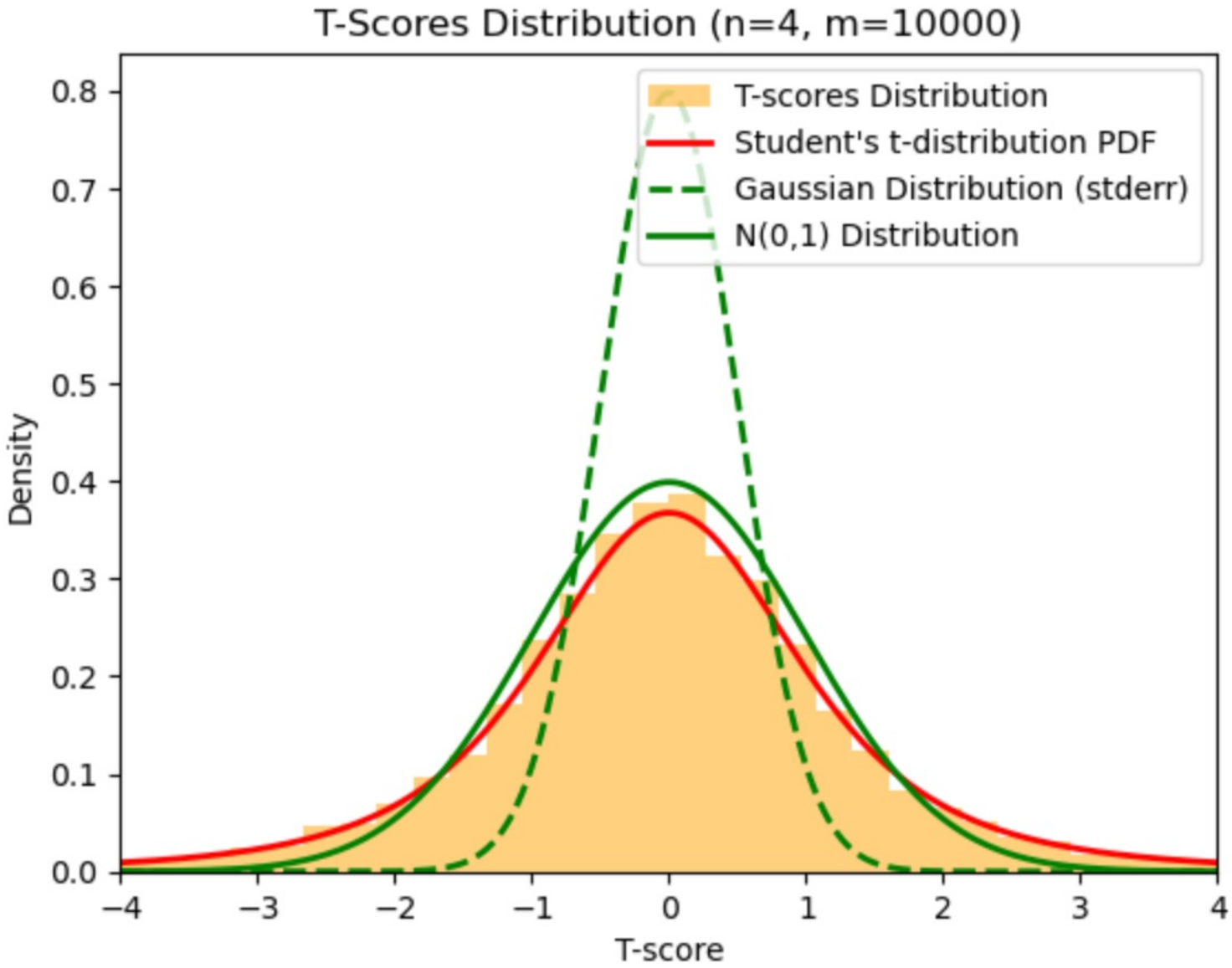
1 million trials (sample means) Looks kind of like a slightly skewed Gaussian.

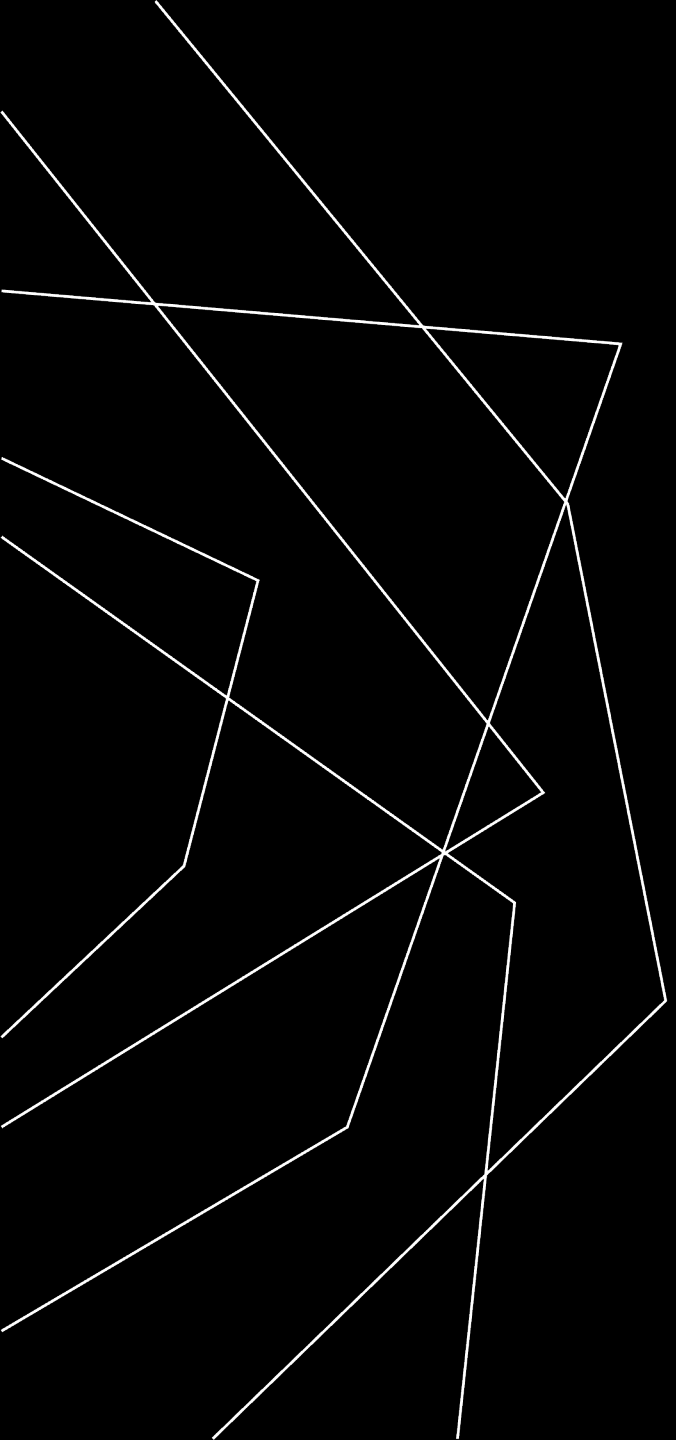
With small n in each sample mean, the distribution of sample means may remain skewed.

CLT's effectiveness depends on increasing *n*.



STUDENT'S T-DISTRIBUTION





THANK YOU

David Harrison

Harrison@cs.olemiss.edu