Module 1.2: Statistics Overview

Statistics: objectives

This section will cover basic statistical methods commonly used in cleaning, analyzing and visualization of data. The relationship between statistics and data science will be discussed. The python skills learned thus far will be used to practice translating a mathematical equation into code.

1. Explain the difference between and calculate mean, median, mode, range, standard deviation, variance of a dataset
2. Differentiate when and why it is appropriate to use mean vs median
3. Use a histogram and box plot to understand the distribution of a dataset
4. Identify outliers and decide what actions to take
5. Normalize and scale a feature and describe when and why it is necessary

Statistics: define

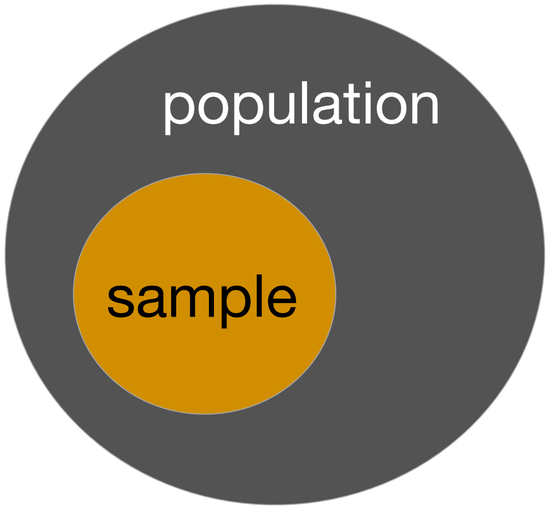
Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring

Statistics: define

Population - a collection of persons, things, or objects under study.

To study the population, we select a **sample** - a portion (or subset) of the larger population. Study this portion (sample) to gain information about the population.

Data are the result of sampling from a population.



Statistics: variables

Variables - a characteristic or measurement for each member of a population

Numerical - numbers (weight, bytes, time)

Categorical - a category that members can be put into (user/bot, location)

Statistics: types of data

QuaLitative - aka **categorical** data, non-numeric data (user status, url, gender)

QuanTitative - numeric data

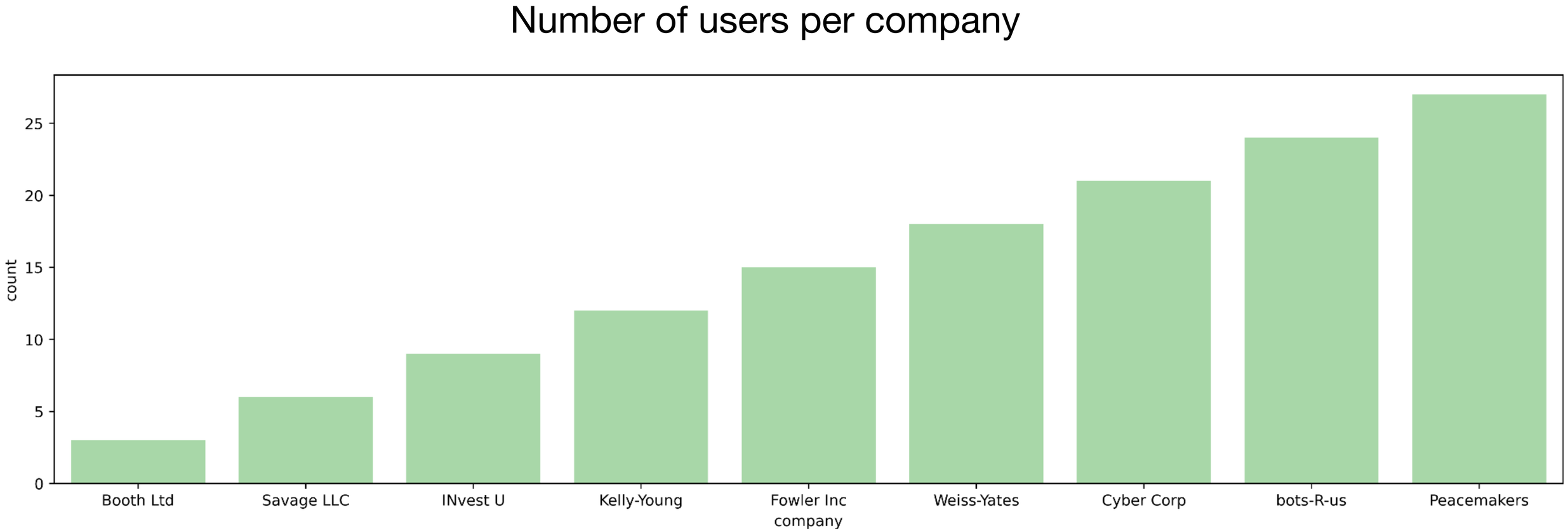
Discrete - integers, like the result of counting (# of site visits, log ins)

Continuous - decimal numbers, fractions (weight, amount of time)

Statistics: data example

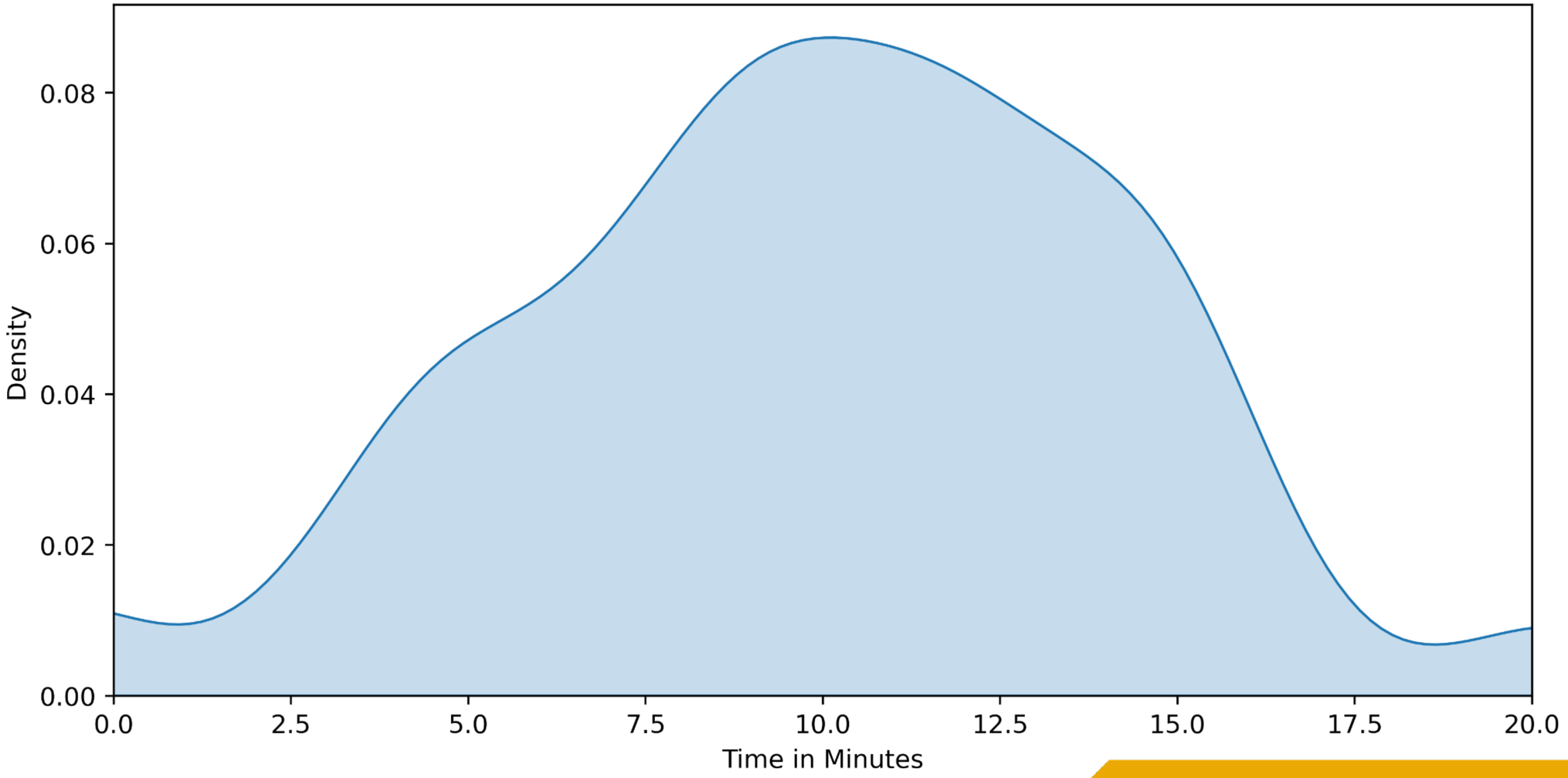
company	username	name	mail	connected
Peacemakers	susan63	Allison Hanson	denisealexander@yahoo.com	22.498529
bots-R-us	gerald76	David Martinez	hernandezmichele@yahoo.com	13.470739
Kelly-Young	christophergomez	Heather Bolton	robert00@gmail.com	13.967764
bots-R-us	mariacraig	Melissa Fisher	katierussell@hotmail.com	7.517620
Cyber Corp	lesliebarnett	Jamie Adams	morganfields@gmail.com	2.734185

Statistics: quantitative discrete



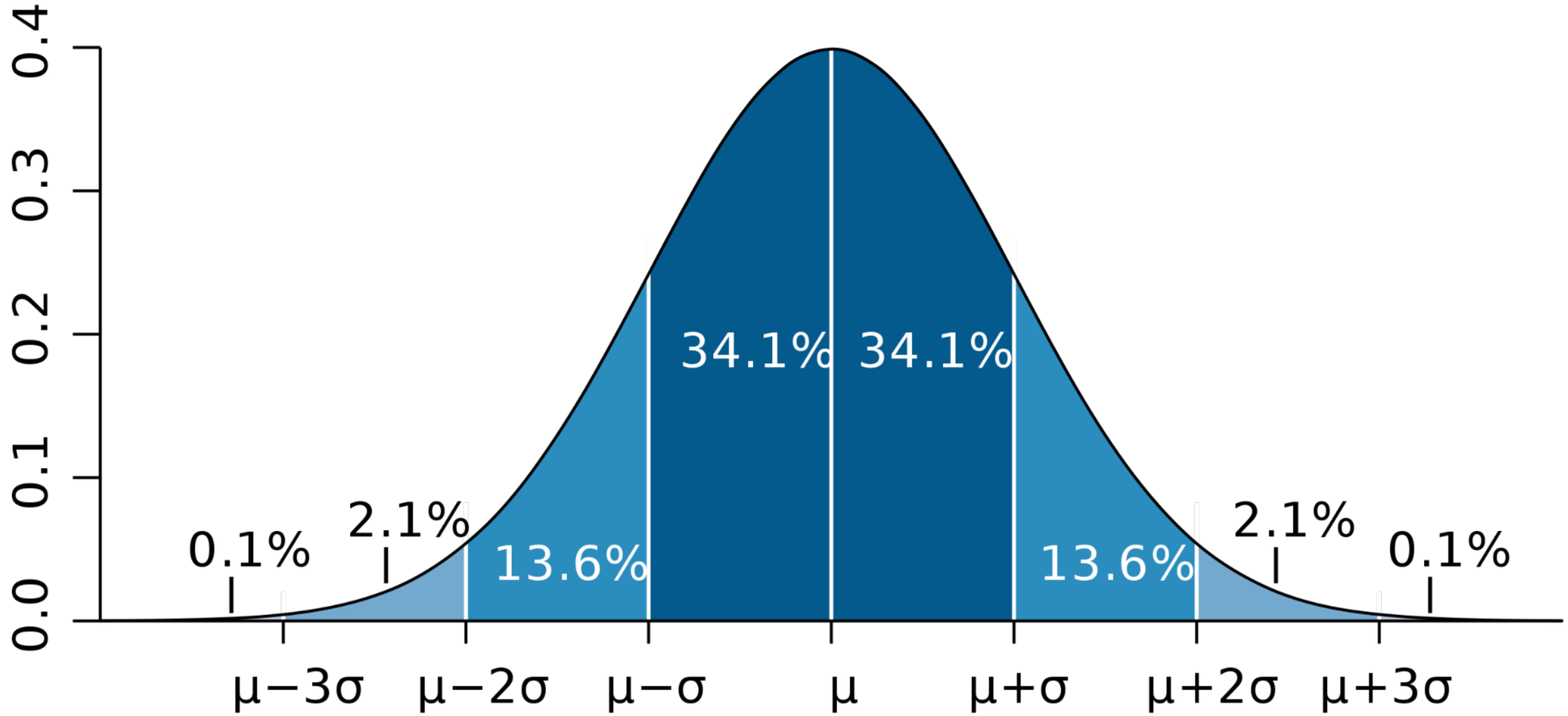
Statistics: quantitative continuous

Amount of time users were logged into server



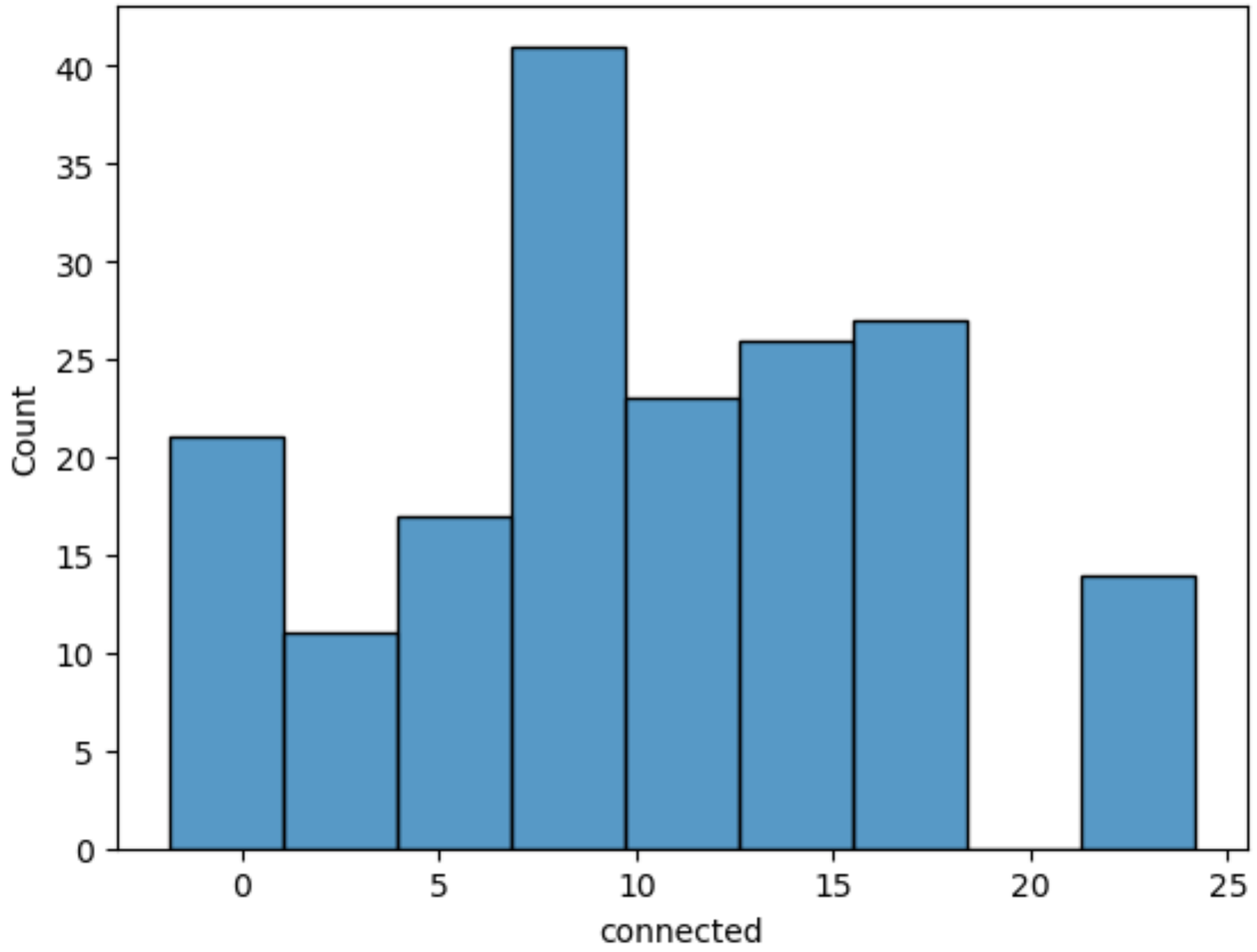
Statistics: normal distribution

Normal distribution (Gaussian)



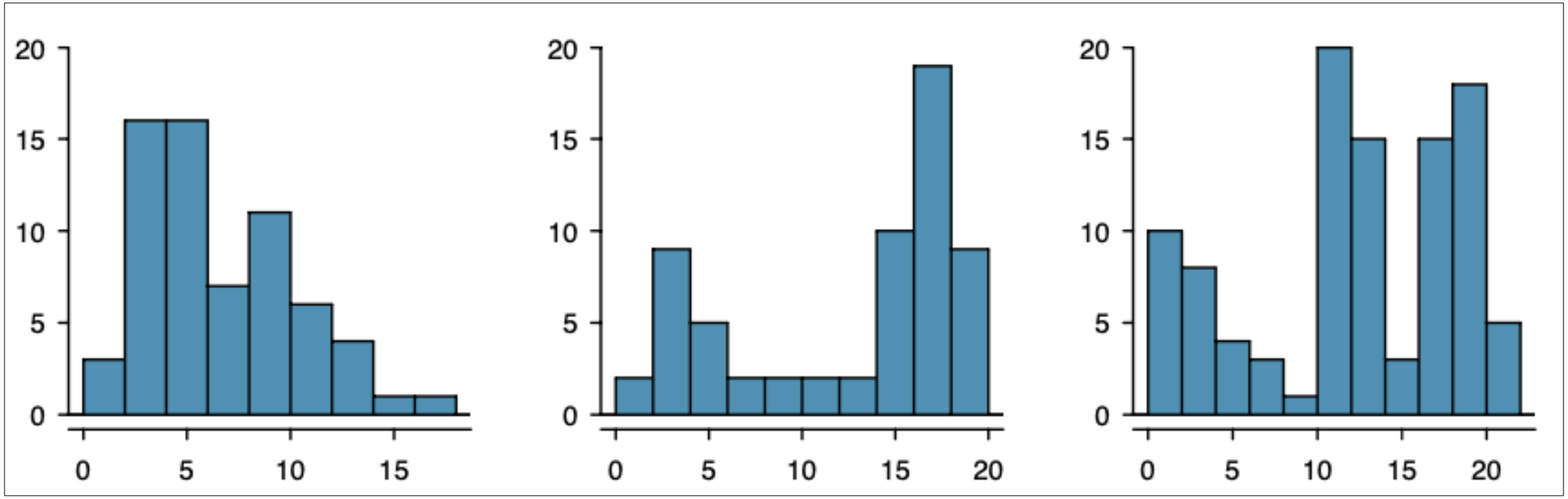
Statistics: histogram

Amount of time users were logged into server



Statistics: mode

Mode - value with the most occurrences



Unimodal Bimodal Multimodal

Statistics: mean

Mean - Measure of central tendency.

μ = population mean

$\bar{x}$ = sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

We say that x bar is the sum of the values of x from 1 to n, multiplied by 1 over the number of values n.

n is the total number of data points

$x_i, i = [1, 2, 3...]$ are the values

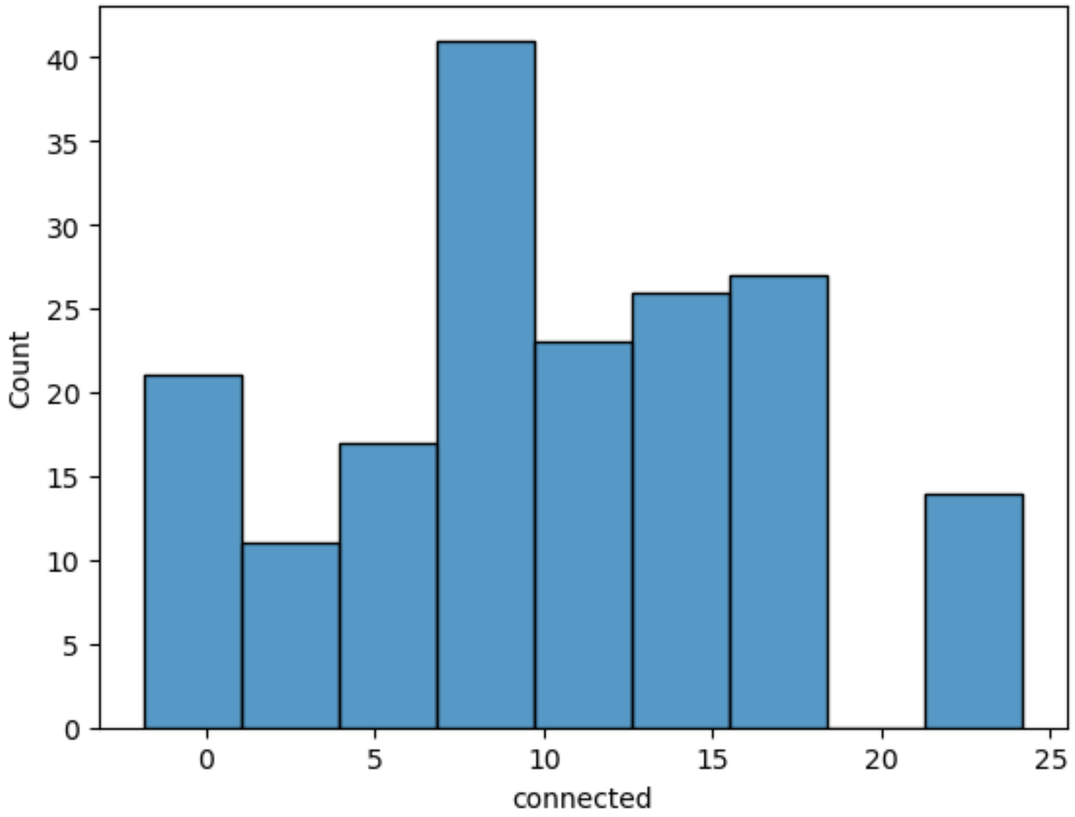
```
(3 + 2 + 9 + 4 + 1) / 5
```

Note: Sensitive to outliers

Statistics: mean

Amount of time users were logged into server

Mean = 10.11



Statistics: variation

Variation - all data has variation (spread of data around the mean)

Variance - the mean squared deviation

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Deviation - square root of the variance

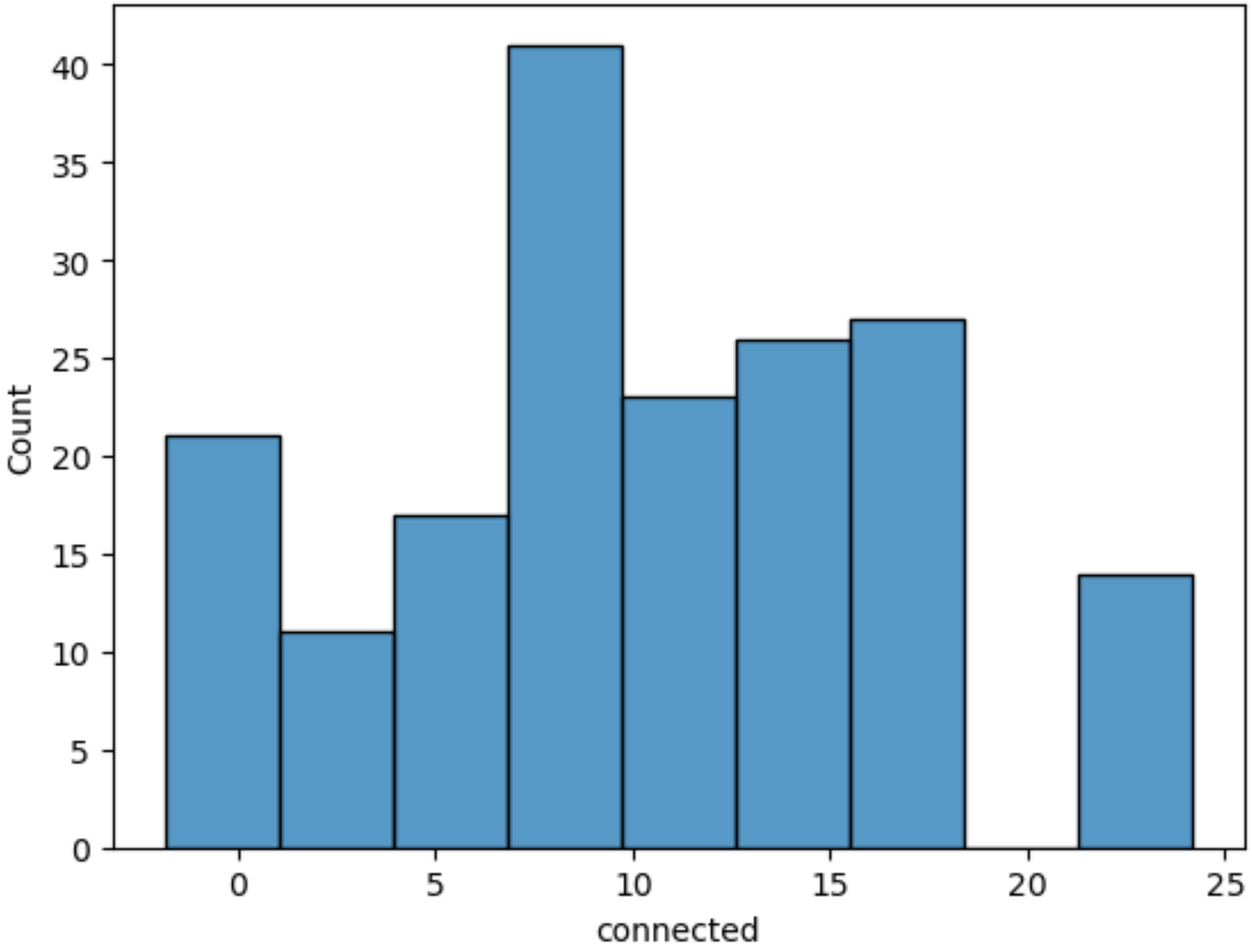
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note: Sensitive to extreme values

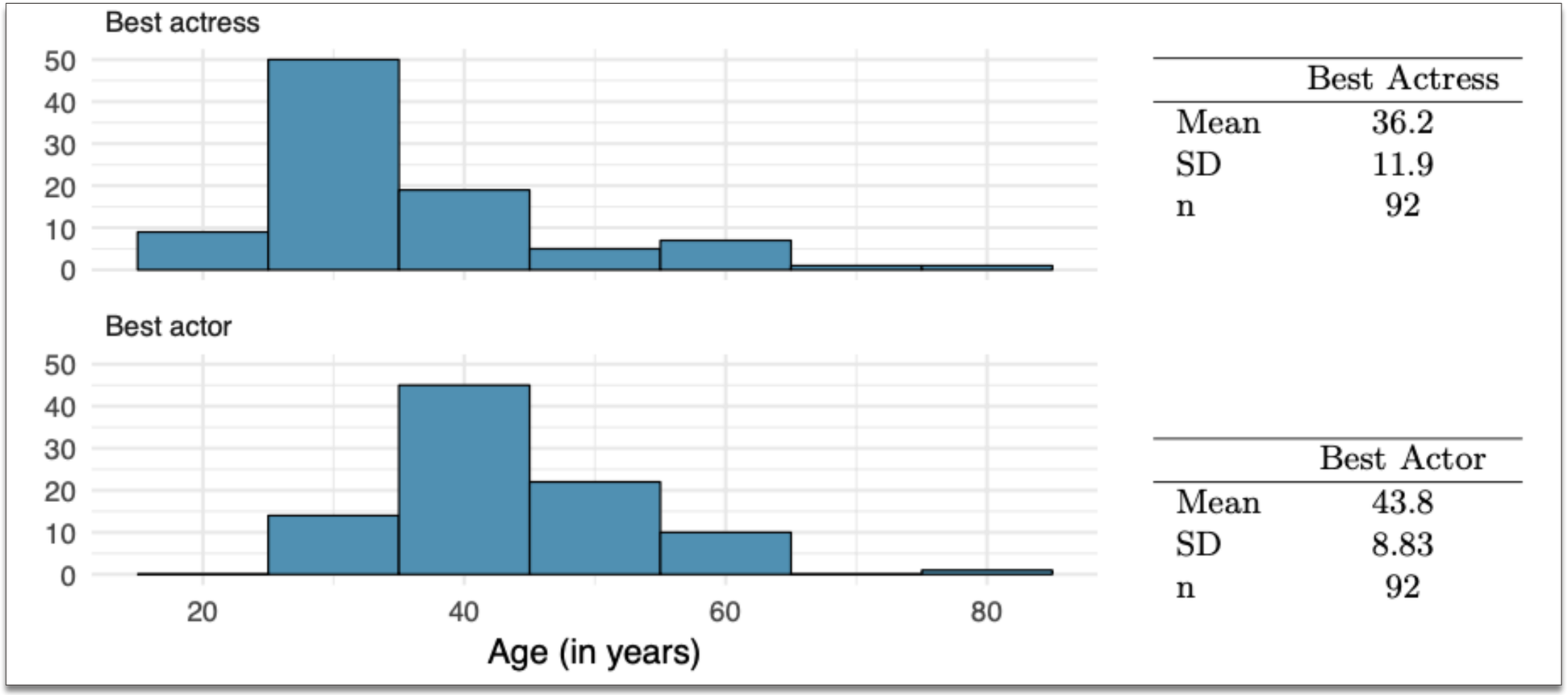
Statistics: variation

Variance = 44.73

Standard Deviation = 6.68



Statistics: mean & variation

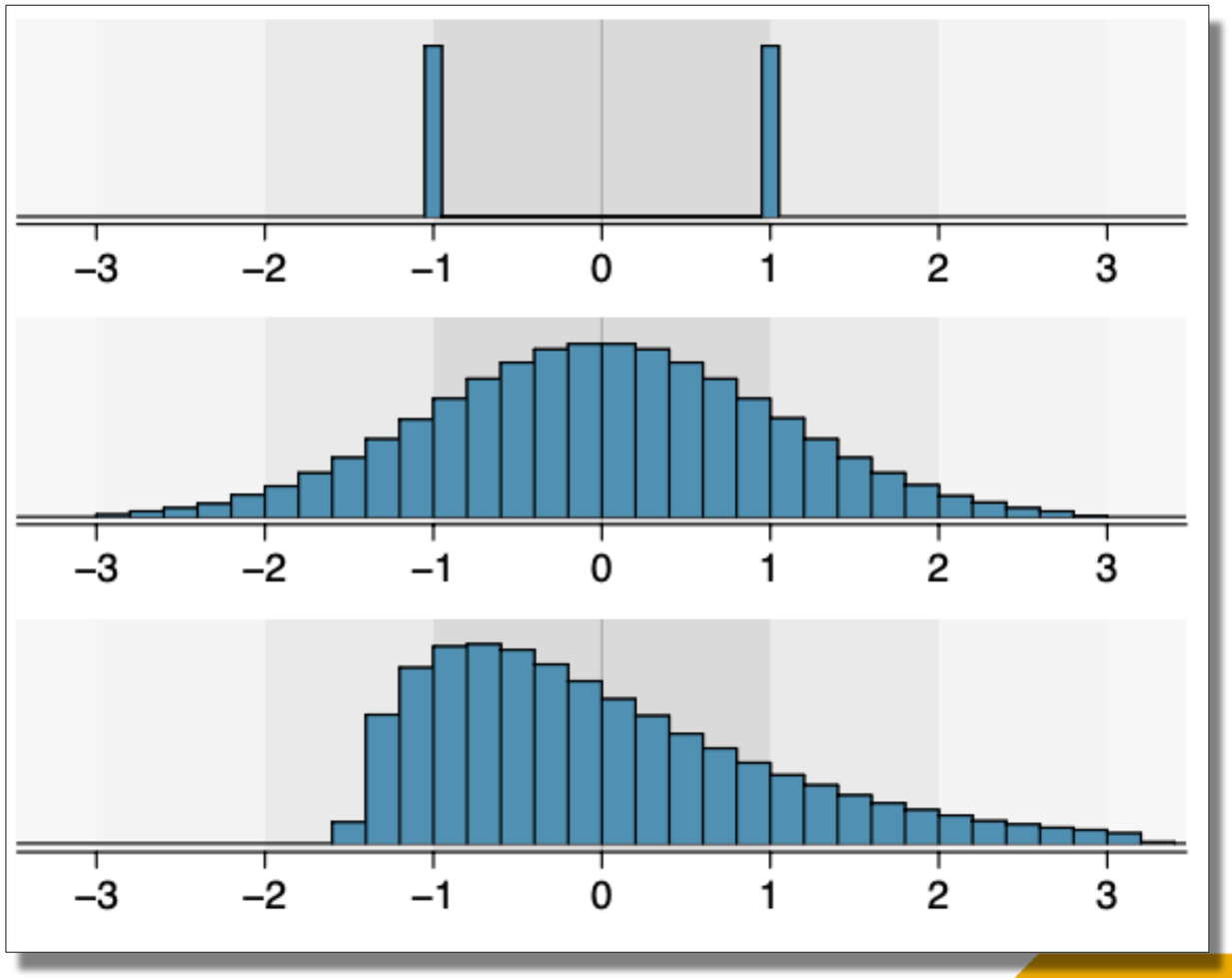


	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

Statistics: median

Median - Middle value separating the greater and lesser halves of a data set



Statistics: median

Median - Middle value separating the greater and lesser halves of a data set

```
s_list = [8,3,1,8,3,9,2,0]
```

Sort the values

```
s_list = [0,1,2,3,3,8,8,9]
```

middle values

Note: If number of values is even, take mean of middle 2 numbers

median = `(3+3)/2 = 3`

Intro Stats: Covariance

Covariance, like the variance, is a measure of spread, however it also measures how closely two datasets track each other.

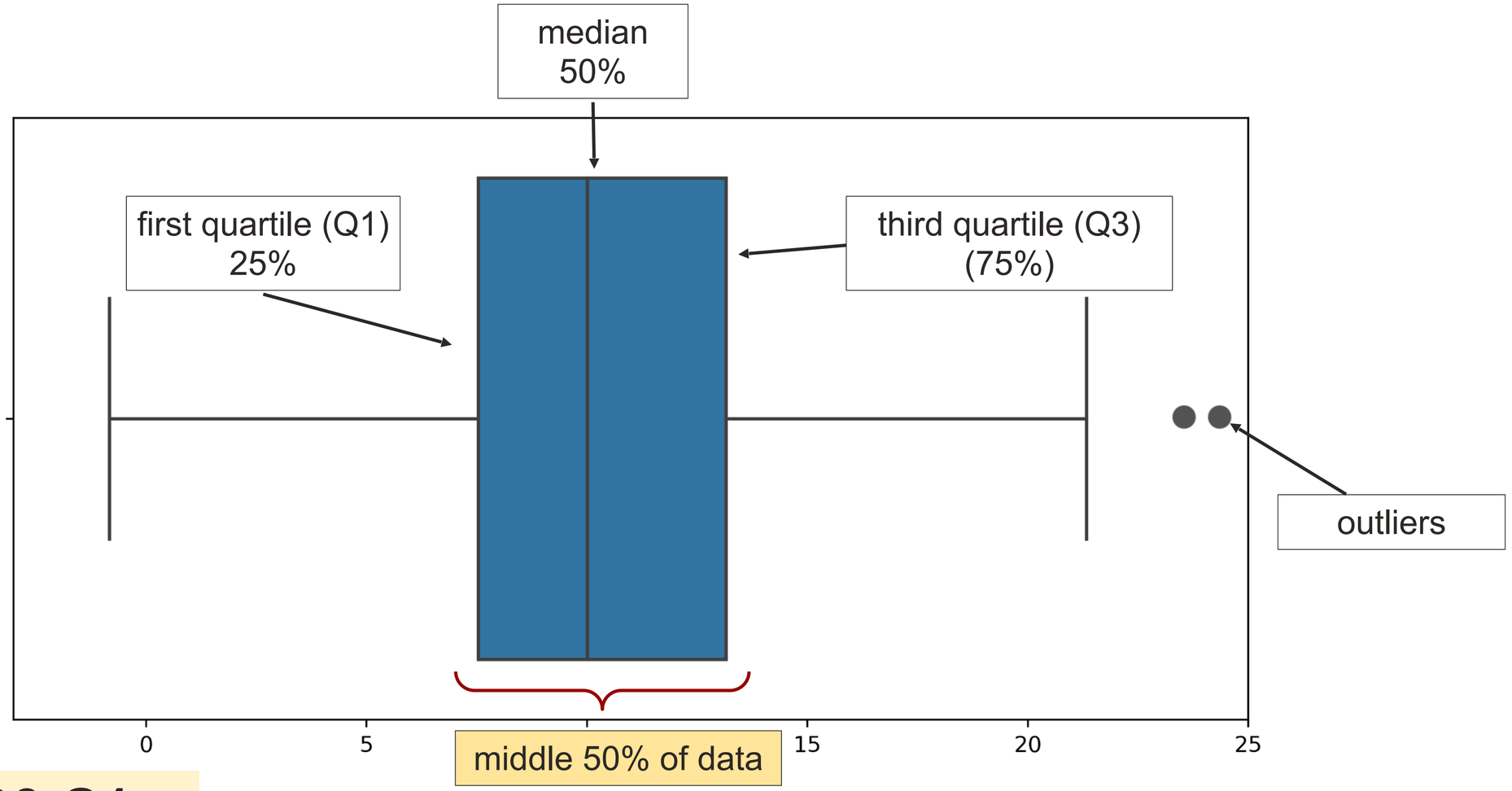
- **Covariance** is a squared quantity, so it is not on the same scale as the mean
- **Covariance** of different pairs of variables can have completely different scales

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

Tukey 5 Number Summary

- **Minimum:** The smallest value in the dataset
- **Lower Quartile:** Smallest 25% of the dataset
- **The Median:** The middle value of the dataset
- **Upper Quartile:** The largest 25% of the dataset
- **Maximum:** The largest value in the dataset

Statistics: IQR (interquartile range)

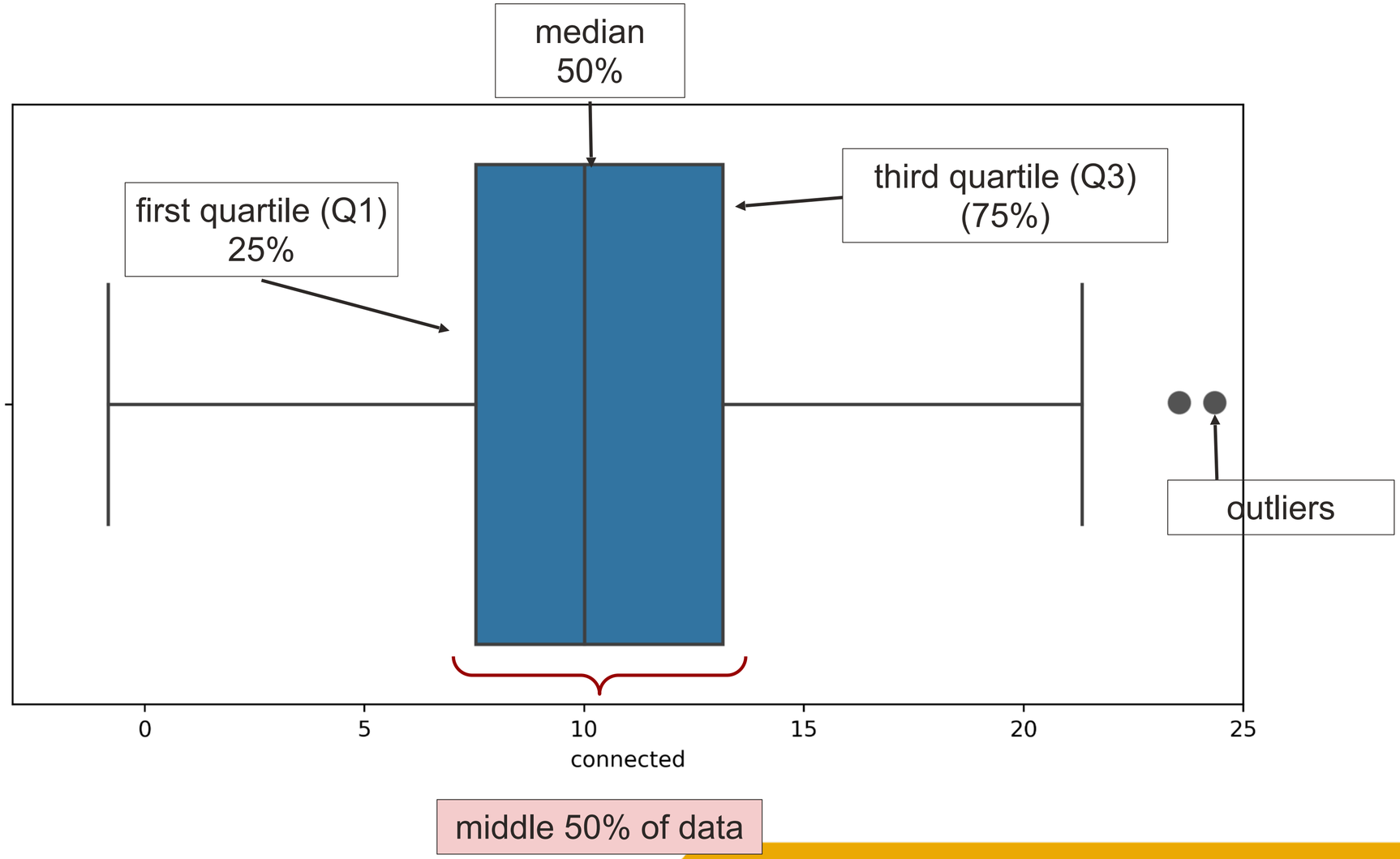


IQR = Q3-Q1

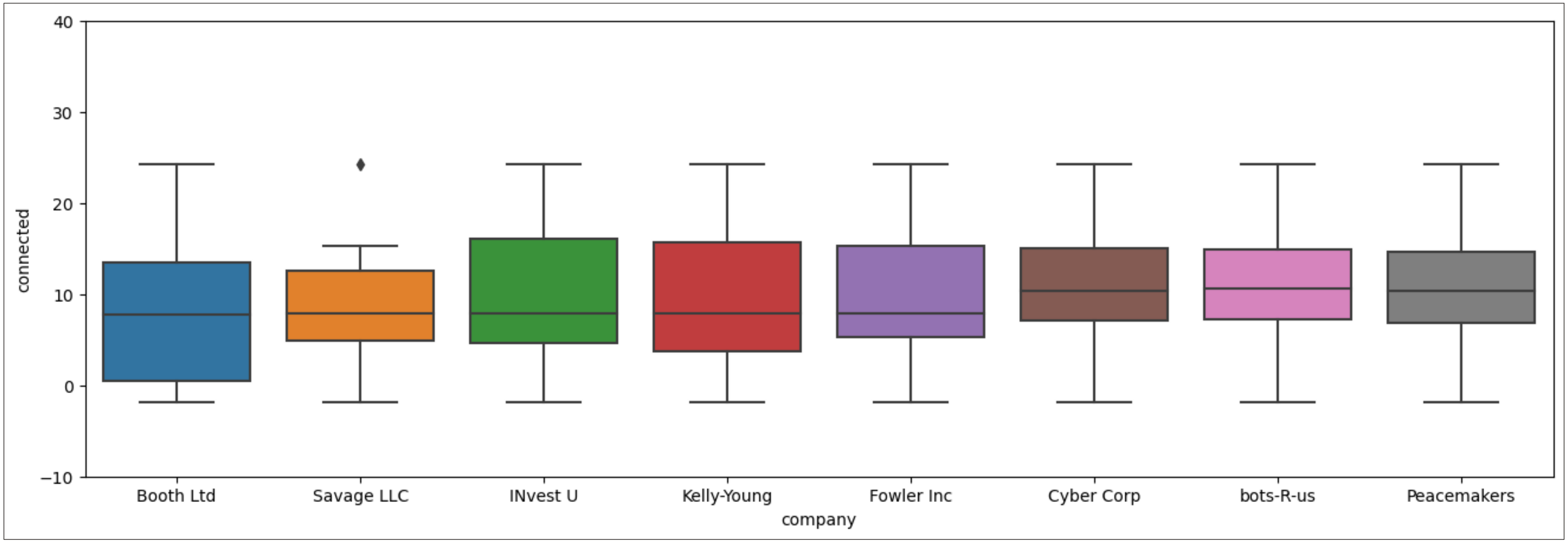
Statistics: IQR (interquartile range)

```
users.connected.describe()
```

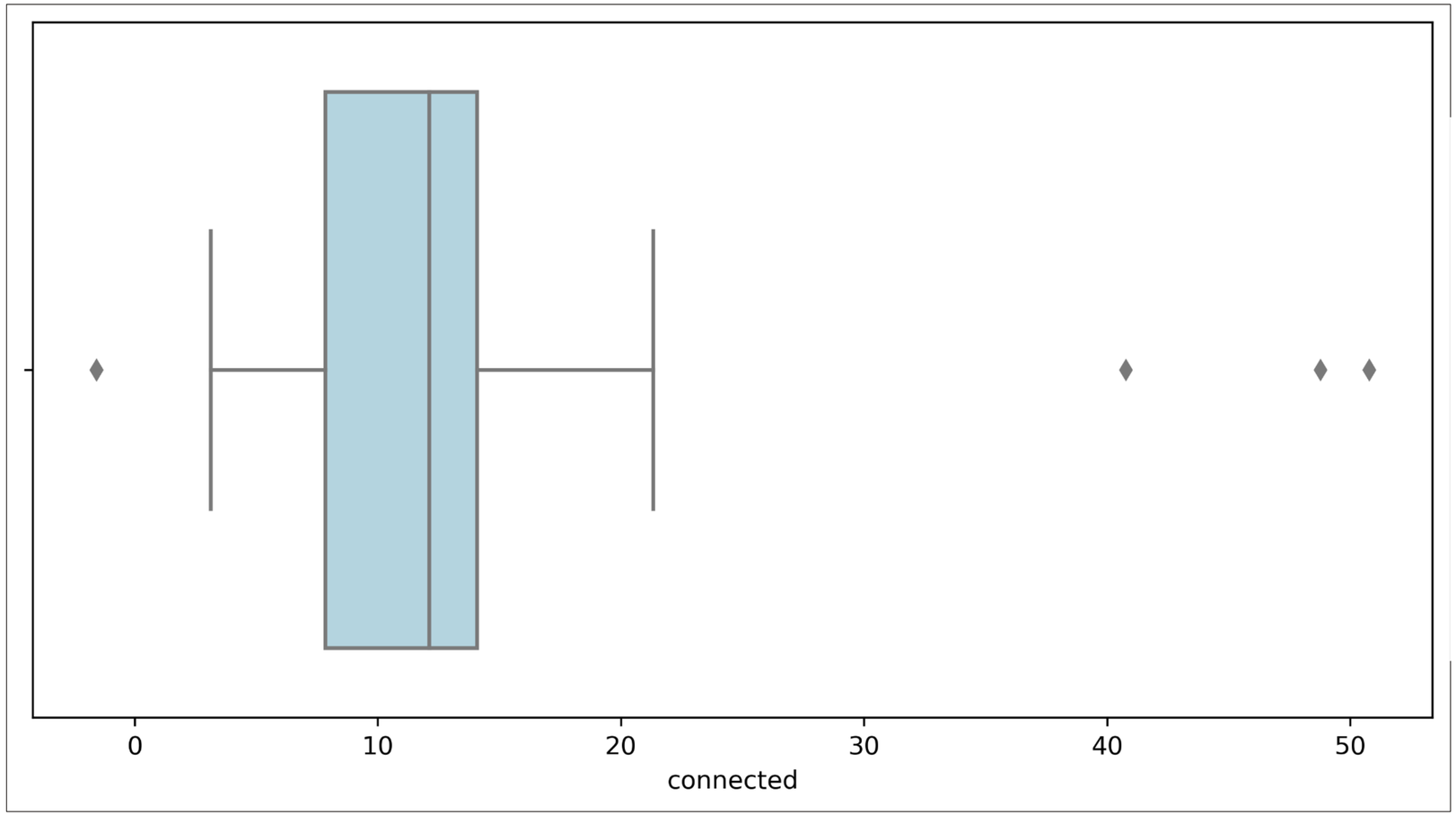
count	180.000000
mean	10.111818
std	6.688669
min	-1.869387
25%	5.307170
50%	9.326331
75%	15.352434
max	24.194889

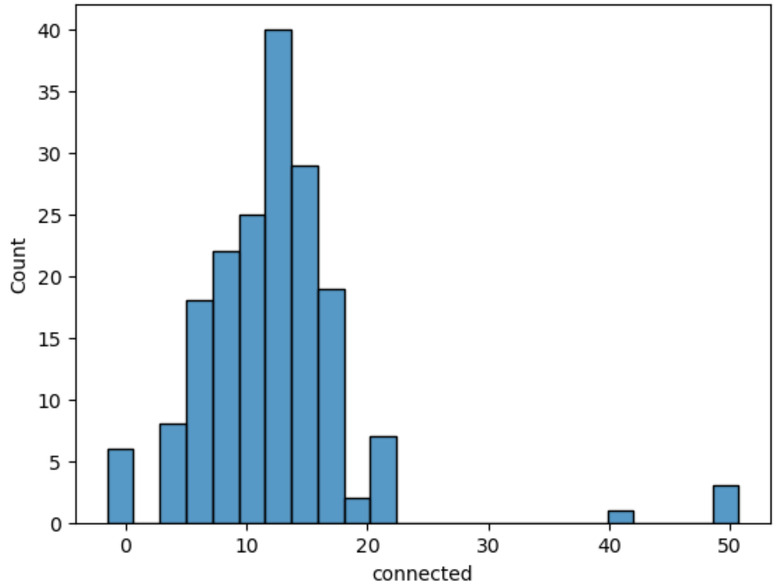


Statistics: IQR



Statistics: outliers





Statistics: Normalize

connected	number_logins
21.333161	2
3.120090	8
16.598087	3
10.711065	12
9.877369	11

Normalization (min-max scaler) - get data on a common scale

$$x_{scaled} = \frac{x - min}{max - min}$$

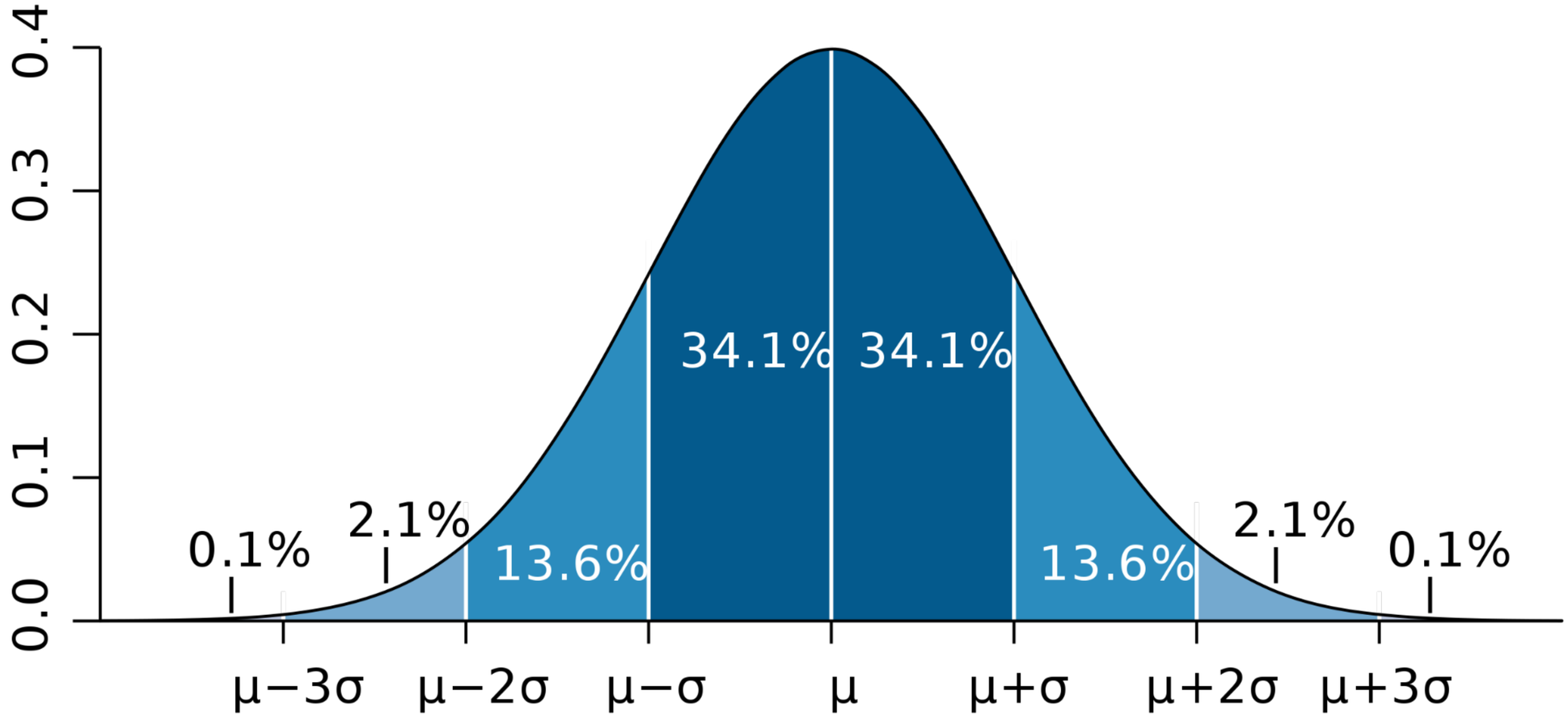
Standardization - rescale so that mean = 0, std = 1

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

	connected	number_logins
count	180.000000	180.0000
mean	11.980229	7.9500
std	7.269764	3.3095
min	-1.565569	1.0000
25%	7.834576	5.0000
50%	12.123636	8.0000
75%	14.088452	11.0000
max	50.770000	13.0000

Statistics: hypothesis testing

Normal distribution (Gaussian)



`Series.abs()`	Absolute Value of the Series
`Series.count()`	Returns number of non-empty values in the series
`Series.max()`	Returns maximum value in the Series
`Series.mean()`	Returns the mean of a Series
`Series.median()`	Returns the median of a Series
`Series.min()`	Returns the minimum value in a Series
`Series.mode()`	Take a guess..
`Series.quantile([q])`	Returns the quantiles of a Series
`Series.sum`	Returns the sum of a series
`Series.std`	Returns the standard deviation of a Series

In Class Exercise
Back at xx:xx

Please take 10 minutes and complete

Worksheet 1.2: Exploring One Dimensional Data