Week 4: Statistics 1

Dr Giuseppe Brandi

Northeastern University London

Outline

What is Statistics?

Statistics is the science of collecting, analysing, presenting, and interpreting data. Big organizations or governments need for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics.

Currently, the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics. (Britannica.com, 2024)

Types of Statistics

Two types of statistics:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

In Descriptive Statistics, the data is summarized through the given observations. This summarization is one from a sample of the population using parameters such as the mean or standard deviation.

It is a way to organize, represent, and describe a collection of data using tables, graphs, and summary measures, e.g., the collection of people in a city using the internet or television.
Four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

Inferential Statistics

Inferential Statistics are used to interpret the meaning of Descriptive Statistics. Once the data has been collected, analyzed, and summarized, we use these stats to describe the meaning of the collected data.

Inferential Statistics allow us to use information collected from a sample to make decisions, predictions, or inferences about a population.

They enable us to make statements and interpretations that go beyond the available data or information, e.g., deriving estimates from hypothetical research.

Single Set of Data

One obvious description of any data set is simply the data itself:

```
num_friends = [100, 49, 41, 40, 25, ... ]
```

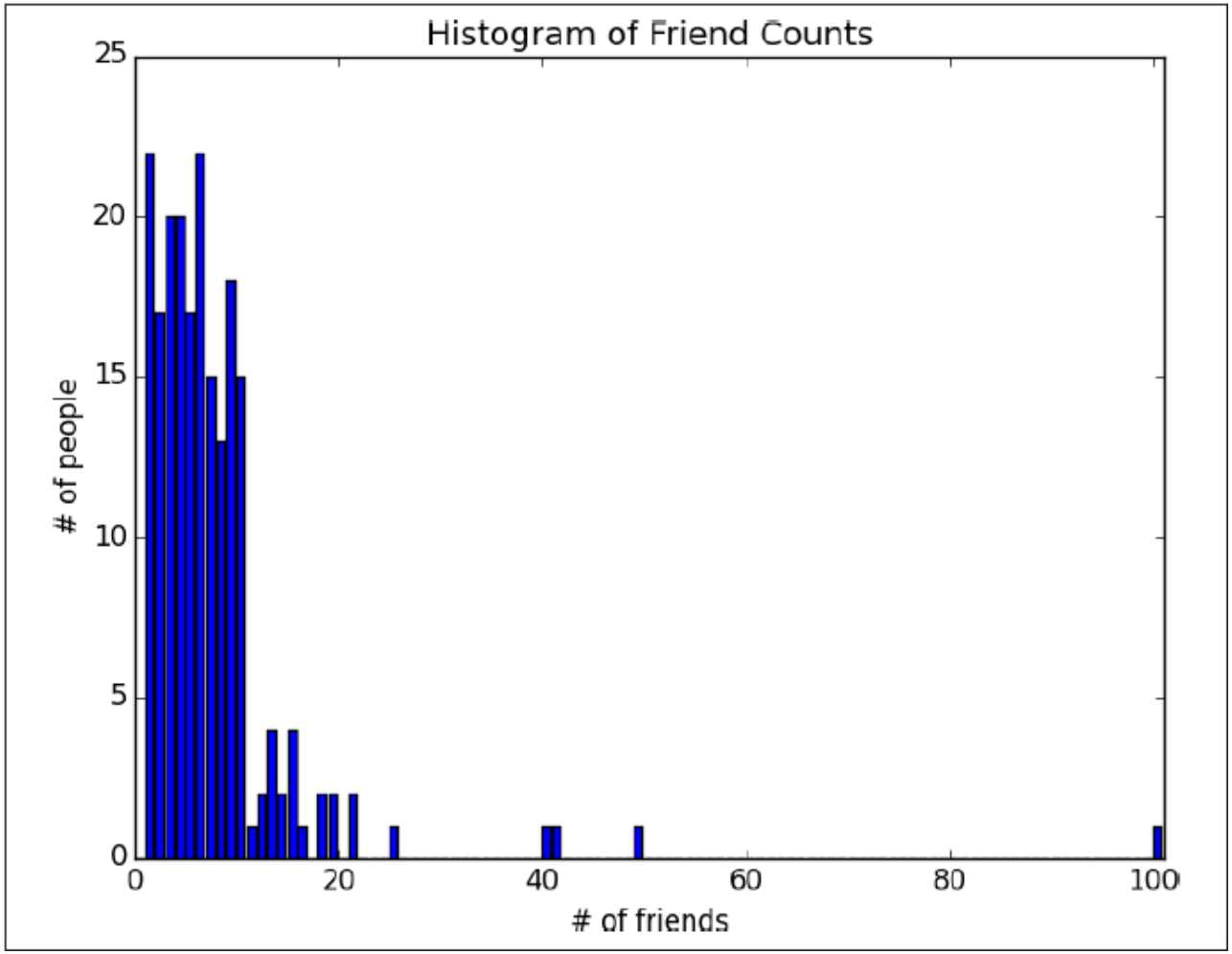
Single Set of Data

For a small enough data set, this might even be the best description.

But for a larger data set, this is unwieldy and probably opaque, e.g., imagine staring at a list of 1 million numbers.

Use statistics to represent and communicate relevant features of the data.

Single Set of Data



Single Set of Data

This chart is still too difficult to interpret correctly. Instead, use statistics, i.e., the simplest statistic is the number of data points:

```
1 num_points = len(num_friends) # 204
```

Single Set of Data

Consider:

```
largest_value = max(num_friends)     # 100
smallest_value = min(num_friends)    # 1
```

Single Set of Data

Special cases of computing the values in specific positions:

```
sorted_values = sorted(num_friends)
smallest_value = sorted_values[0]          # 1
second_smallest_value = sorted_values[1]   # 1
second_largest_value = sorted_values[-2]   # 49
```

Central Tendencies

Notion of where our data is centered:
Use the **mean** (or **average**), which is just the sum of the data divided by its count.

```
1 def mean(x):
2     return sum(x) / len(x)
3
4 mean(num_friends)   # 7.333333
```

Central Tendencies

For two data points, the mean is the point halfway between them.

As you add more points, the mean shifts around, but it always depends on the value of every point.

The **median** is the middle-most value (if the number of data points is odd) or the average of the two middle-most values (if the number of data points is even).

Central Tendencies

Example:

If we have five data points in a sorted vector x, the median is `x[5 // 2]` or `x[2]`.

If we have six data points, we want the average of `x[2]` (the third point) and `x[3]` (the fourth point).

Unlike the mean, the median does not depend on every value in the data, e.g., if you make the largest point larger (or the smallest point smaller), the middle points remain unchanged; so does the median.

Central Tendencies: Mean vs Median

```python
def median(v):
    """finds the 'middle-most' value of v"""
    n = len(v)
    sorted_v = sorted(v)
    midpoint = n // 2

    if n % 2 == 1:
        # if odd, return the middle value
        return sorted_v[midpoint]
    else:
        # if even, return the average of the middle values
        lo = midpoint - 1
        hi = midpoint
        return (sorted_v[lo] + sorted_v[hi]) / 2


median(num_friends) # 6.0
```

Central Tendencies: Mean and Median

Mean is simpler to compute, and it varies smoothly as our data changes.

In order to find the median, we have to sort our data. Changing one of our data points by a small amount might increase the median by that amount, less than that amount, or not at all (depending on the rest of the data).

Central Tendencies: Quantile

A generalization of the median is the quantile, which represents the value less than which a certain percentile of the data lies.

In that sense, the median represents the value less than which 50% of the data lies.

Central Tendencies: Quantile

```python
def quantile(x, p):
    """returns the pth-percentile value in x"""
    p_index = int(p * len(x))
    return sorted(x)[p_index]

quantile(num_friends, 0.10) # 1
quantile(num_friends, 0.25) # 3
quantile(num_friends, 0.75) # 9
quantile(num_friends, 0.90) # 13
```

Dispersion

Dispersion refers to measures of how spread out our data is.

We use statistics for which values near zero signify not spread out at all and for which large values signify very spread out.

A very simple measure is the range, which is just the difference between the largest and smallest elements.

Dispersion: Range

```python
# "range" already means something in Python, so we'll use a different name
def data_range(x):
    return max(x) - min(x)

data_range(num_friends) # 99
```

Dispersion: Range

The range is zero precisely when the max and min are equal, which can only happen if the elements of x are all the same, meaning the data is as undispersed as possible.

If the range is large, then the max is much larger than the min and the data is more spread out.

But like the median, the range does not depend on the whole data set.

Dispersion: Variance σ^2

A more complex measure of dispersion is the variance σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Where μ is the mean of the data. It has units that are the square of the original units!

Dispersion: Standard Deviation σ

Standard deviation σ is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

It is a measure of how dispersed the data is in relation to the mean.

- Low (small) standard deviation indicates data is clustered tightly around the mean.

- High (large) standard deviation indicates data is more spread out.

Dispersion: Standard Deviation σ

Both mean and standard deviation are sensitive to the data.

```
def standard_deviation(x):
    return math.sqrt(variance(x))

standard_deviation(num_friends) # 9.03
```

Covariance

Whereas variance measures how a single variable deviates from its mean, covariance measures how two variables vary in conjunction from their means.

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

Covariance

- When corresponding elements of X and Y are both above their means or both below their means, a positive number enters the sum.

- When one is above its mean and the other below, a negative number enters the sum.

- A large positive covariance means that X tends to be large when Y is large and small when Y is small.

- A large negative covariance means the opposite.

- A covariance close to zero means no such relationship exists.

Correlation

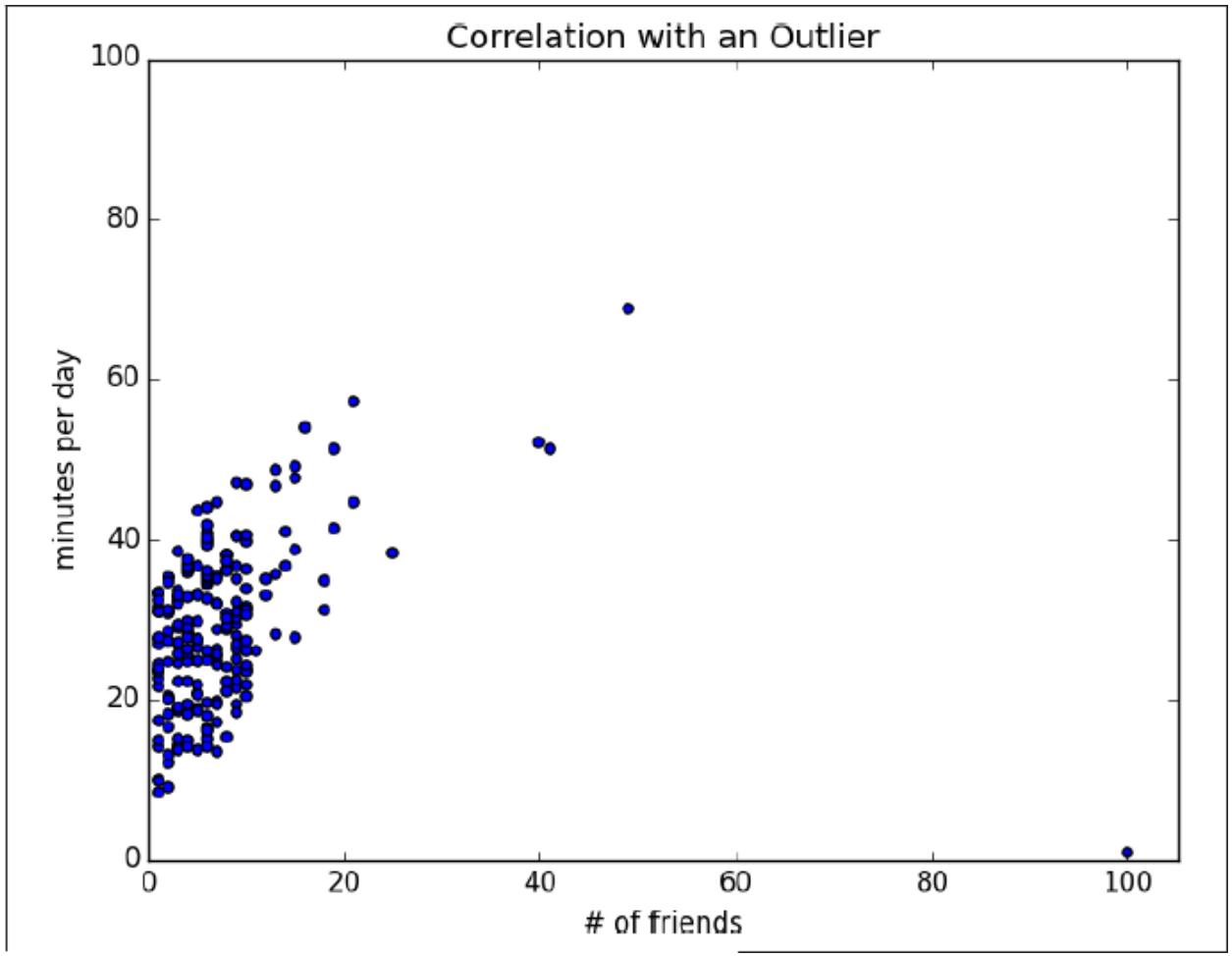
Correlation divides out the standard deviations of both variables.

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

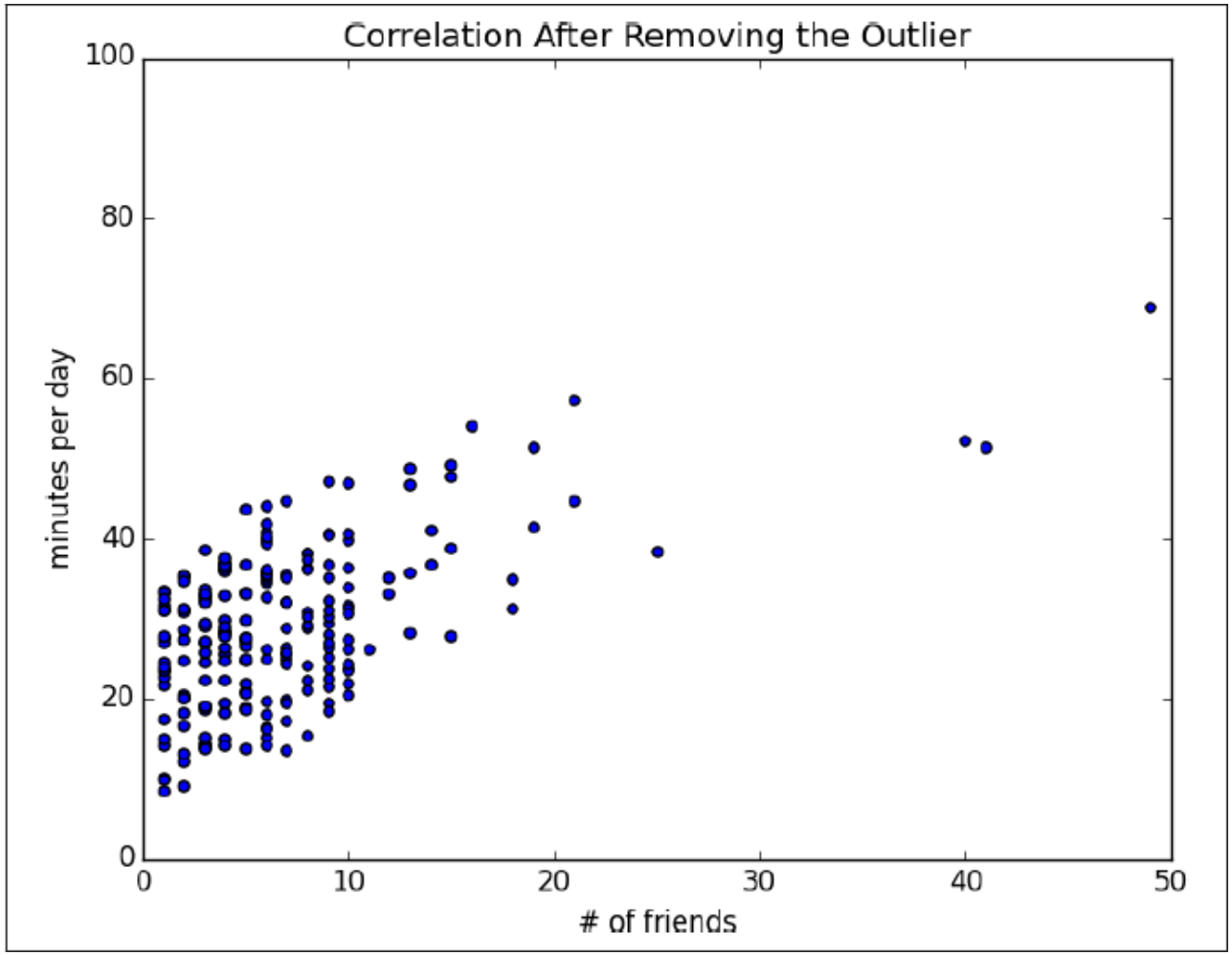
- The correlation is unitless and always lies between -1 (perfect anti-correlation) and 1 (perfect correlation).

- A number like 0.25 represents a relatively weak positive correlation.

Correlation with an Outlier



Correlation after Removing the Outlier



Correlation

Simpson's Paradox

One not uncommon surprise when analyzing data is Simpson's Paradox, in which correlations can be misleading when confounding variables are ignored.

Correlation and Causation

Correlation is not causation.

This is an important point: if X and Y are strongly correlated, that might mean:

- That X causes Y
- That Y causes X
- That each causes the other
- That some third factor causes both
- Or it might mean nothing

Probability

What is a probability?

Probability is a way of quantifying the uncertainty associated with events chosen from some universe of events.

Consider rolling a die: it consists of all possible outcomes.

Any subset of these outcomes is an event; for example, "the die rolls a one" or "the die rolls an even number."

Probabilities: Dependence vs Independence

We say that two events E and F are dependent if knowing something about whether E happens gives us information about whether F happens (and vice versa).

Otherwise, they are independent.

Probabilities: Dependence vs Independence

Mathematically, we say that two events E and F are independent if the probability that they both happen is the product of the probabilities that each one happens:

$$P(E \text{ and } F) = P(E) \times P(F)$$

Probabilities: Conditional Probability

The probability of event E given that event F has occurred is called the conditional probability of E given F:

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

Probabilities: Bayes's Theorem

Bayes's Theorem is a way of "reversing" conditional probabilities.

Assume we need to know the probability of some event E conditional on some other event F occurring, but we only have information about the probability of F conditional on E occurring. Bayes's Theorem states:

$$P(E|F) = \frac{P(F|E) \times P(E)}{P(F)}$$

Probabilities: Bayes's Theorem

This theorem allows us to update our beliefs based on new evidence.

It is foundational in fields like statistics, machine learning, and data science.