Week 3 - Numerical Summaries

Created	August 14, 2023 12:25 PM
Date	Empty
Status	Empty

+ Add a property

Add a comment…

0) Overview of numerical summaries

A numerical summary is a measure (or a “statistic”, a “single number”) that tries to describe as much as possible about data in the simplest way possible.

Major features we can summarize numerically include:

- Maximum
- Minimum
- Centre (mean, median)
- Spread (standard deviation, range, IQR)

0.1) Conventions to remember:

1. We always try to report measures of centre and spread together to minimize errors with numerical summaries. Specifically aim to either:
   a. report **mean with standard deviation** (such report is called coefficient of variation!), or
   b. report **median with interquartile range**

1) Centre

1.1) Some notation

- a dataset of size n can be represented by $x_1, x_2, \ldots, x_n$
- the ranked data set (ordered from smallest to largest) can be represented by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
- the sum of data is $\sum_{i=1}^{n} x_i$

1.2) Mean

Mean, denoted as $\bar{x}$, is the average of data; the unique point at which the data is balanced. Mean is defined as:

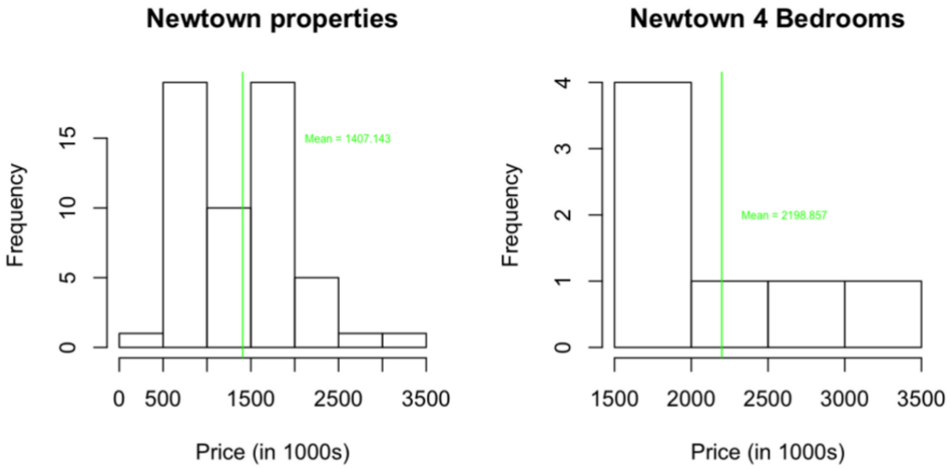
$$\text{Mean} = \frac{\text{sum of data}}{\text{size of data}}$$

or, formally, as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1.2.1) Mean on the histogram

In the histogram, “mean” is the balancing point where the expensive and cheap properties cancel each other out.



1.3) Median

The median is the middle data point, when the data is ordered from smallest to largest. An example of a median could be. Formally, the median is defined as:

- For an odd-sized dataset

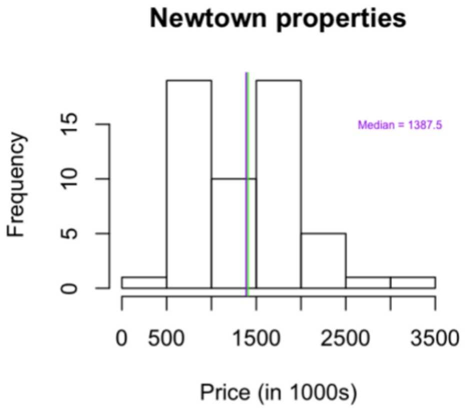
$$\text{Median} = x_{\left(\frac{n+1}{2}\right)}$$

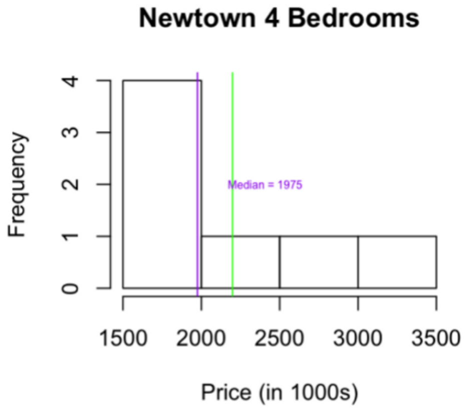
- Or, for an even-sized dataset (there are 2 midpoints)

$$Median = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

1.3.1) Median on the histogram

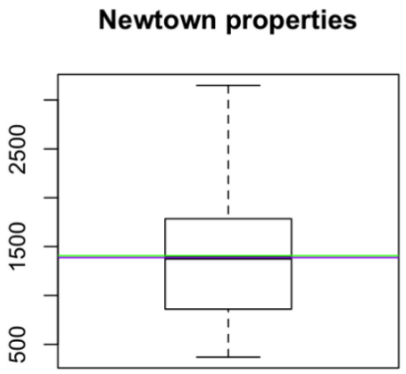
The median on the histogram is the halfway point in the histogram

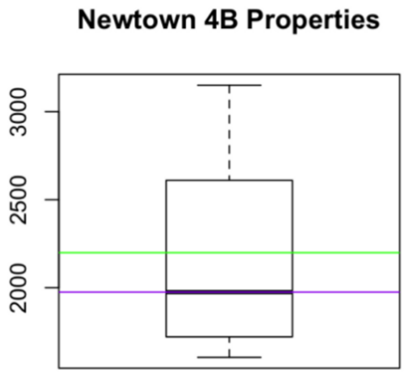




1.3.2) Median on the boxplot

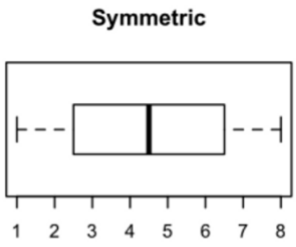
The median on the boxplot is the central line.

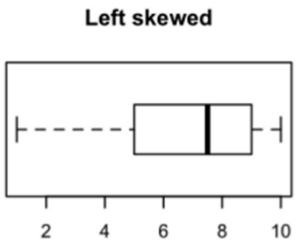


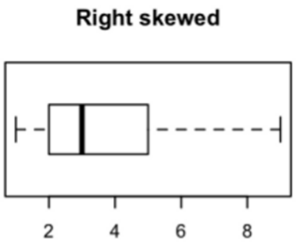


1.4) Comparing mean and median

- For symmetric data, we expect the mean and median to be the same.
- For left-skewed data, we expect the mean to be smaller than the median.
- For right-skewed data, we expect the mean to be larger than the median.

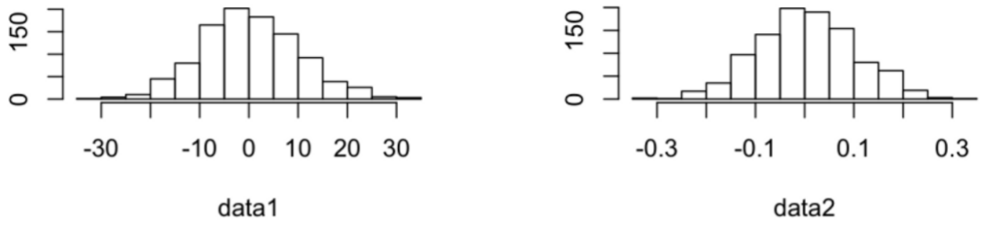






1.5) Mean and median must be paired with the spread!

Mean and median are generally good measures of the middle point of data. However, the mean and median of two data with vastly different spreads may be similar! See the figure below!



That's why you ought to take into account the measure of the spread of the data when you're putting forward a mean or median!

1.6) Robustness

Robustness refers to the quality of a numerical summary to resist the influence of outliers - and this is a very useful quality to have in a numerical summary when the data is heavily skewed. The following numerical summaries are said to be robust:

- median: a robust measure of the middle value
- interquartile range (IQR): a robust measure of the middle 50% of the data
- quantiles: a robust measure of central tendency and variability.

2) Spread

2.1) Gap

A "gap" refers to the difference between a data point and the mean of the data.

Example: Suppose we have a data {100, 200, 300, 400, 500}, which has a mean 200.

The gap from 500 and 200 is defined as: 500 - 200 = 300.

2.2) Population vs. sample

- Population is the group of subjects you want to draw conclusions from.
- Sample is the group of subjects you collect data from.

2.3) RMS and standard deviation

2.3.1) Root mean square (RMS)

RMS is a measure the mean of a set of numbers, regardless of the signs (positive or negative). To do RMS, you follow the reverse of its name:

1. **Square** the numbers
2. **Mean** the result
3. **Root** the result

Intuition — when you square a negative number, you get a positive, hence why we say that RMS ignore signs.

2.4) Standard deviation

Standard deviation is a measure of how much a data point deviates from the mean of the data.

2.4.1) Population standard deviation via RMS

You can use RMS to get the **population standard deviation** like so:

$$\sigma_{\text{pop}} : \sqrt{\text{Mean of }(\text{gaps})^2}$$

or formally:

$$\sigma_{\text{pop}} : \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

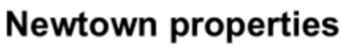
2.4.2) Sample standard deviation via RMS

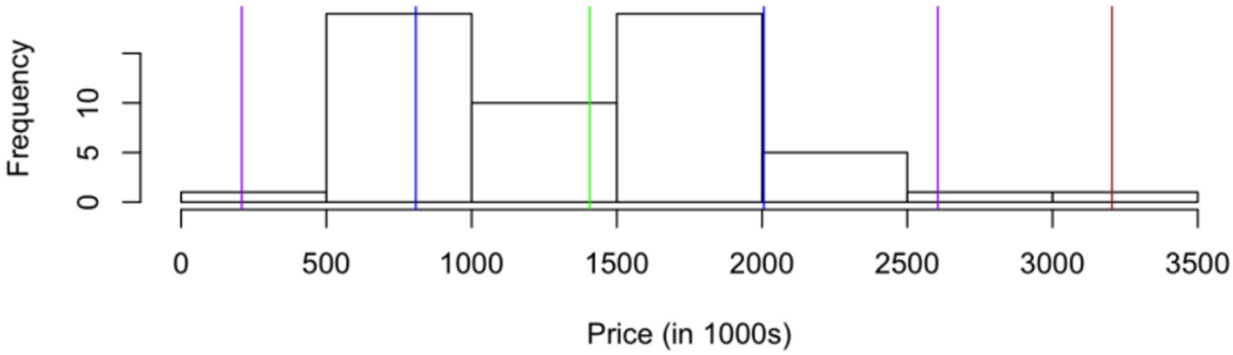
You can use RMS to get the **sample standard deviation** like so:

$$\sigma_{\text{sample}} : \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

2.4.3) Example of visualizing s.d. on the histogram

Context: Green is mean





2.4.4) Rule of thumb for s.d. regarding portions of data

Commonly, datasets tend to follow the pattern below:

Percentage of data	Distance from mean
68%	within 1 SD
95%	within 2 SDs
99.7%	within 3 SDs

2.4.5) Variance

Variance is another measure of spread, but it's basically just standard deviation squared. You may see variance being used instead of standard deviations. One example is the notation for normal distribution, i.e. $N(\mu, \sigma^2)$.

2.5) Z-score (standard units)

Z-score represents the **number of standard deviations a data point is away from the mean of the data.** It is defined as:

$$Z = \frac{\text{data point} - \text{mean}}{\text{SD}_{\text{sample}}}$$

or, using appropriate notations:

$$Z = \frac{x - \mu}{\sigma}$$

Realize that the higher z-score a data point has, the farther it is from the mean of the data, therefore the more unusual it is!!

2.6) Interquartile range (IQR)

IQR is a measure of the range of the middle 50% of the data. It is defined as:

$$\text{IQR} = Q_3 - Q_1$$

Where Q1 and Q3 and lower and upper quartiles of the data respectively.

2.6.1) Quartiles vs. Quantiles

Quartiles are 3 specific values that split the data into 4 segments (quarters).

- Q1: 25th percentile; the value under which 25% of data points lie.
- Q2: 50th percentile — **the median**; the value value under which 50% of data points lie.
- Q3: 75th percentile; the value under which 75% of data points lie.

Quantiles are 4 specific values that split the data into 5 segments (same logic as quartiles ^): 1) 20% quantile, 40% quantile, 60% quantile, 80% quantile.

2.6.2) Upper and lower thresholds

Upper and lower thresholds are values in a data useful for identifying outliers. Specifically, we follow that:

- any data greater than the upper threshold is an outlier
- any data smaller than the lower threshold is an outlier

Upper and lower thresholds are defined like so:

$$\text{UT} = Q_3 + 1.5(\text{IQR})$$

$$\text{LT} = Q_1 - 1.5(\text{IQR})$$

Realize that boxplots make it easy to work with quartiles, IQR, and thresholds!

2.7) Reporting centre together with spread

The convention is to either:

1. report **mean with standard deviation** (such report is called coefficient of variation!), or
2. report **median with interquartile range**

This means, when you're reporting the central tendency of data, it is preferred that for whichever of the two choices you pick, you report the former and latter numerical summaries.

2.7.1) Coefficient of variation (CV)

Coefficient of variation (CV) is a measure that combines mean and standard deviation, thereby allowing you to report one value to describe the central tendency of data instead of providing two values (1) mean and (2) sd. CV is defined as:

$$\text{CV} = \frac{\text{mean}}{\text{standard deviation}} = \frac{\bar{x}}{\sigma}$$