Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

EEU33E03: Probability and Statistics

Lecture 4: Numerically Summarizing Data

Arman Farhang (arman.farhang@tcd.ie)

2023

Outline

Organizing and Summarizing Qualitative Data

The most common measures (statistics) used to describe numerical variables can be classified into two categories.

1. Measures of **centrality**, i.e. measure of the centre of the distribution.
2. Measures of **variability**, i.e. the spread, dispersion of the data.

Notation and Measures of Centrality

Let x denote a particular numerical variable, e.g. height.

Considering n observations of this variable (in the same sequence as they are observed),

$$x_1, x_2, \ldots, x_n, \tag{1}$$

then n ordered observations are denoted by,

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}, \tag{2}$$

where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. This is simply a list of the observations from the smallest to the largest.

If a value appears k times in the sample, it must appear k times in this ordered list. In this case, the index is the rank of the observation.

The Sample and Population Means

The sum of the data is denoted by,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n. \tag{3}$$

The **sample mean** (which we observe) is given by $\overline{x}$, where,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{4}$$

The **population mean** (which we do not observe) is given by μ, where,

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{5}$$

where N is the population size.

The Sample Median

— When n is **odd**, the sample median, denoted by '*Med*', is the observation that appears in the middle of the list of ordered observations.

— When n is **even**, the sample median is the average of the two observations in the middle of this list.

That is to say,

— For **odd** values of n, the median has rank $\frac{n+1}{2}$.

— For **even** values of n, the median is the average of the two observations with ranks $\frac{n}{2}$ and $\frac{n}{2}+1$.

Note: Half the data are less than or equal to the median and the other half the data are strictly greater than the median.

The Sample Mode

The mode is useful when we are dealing with discrete or categorised data (**not continuous data**).

The mode is the observation (or category, as appropriate) that occurs most frequently in a sample.

Sample Mean, Median and Mode: Example

Example 1: The number of children in 14 families is given below:

$$4, 2, 0, 0, 4, 3, 0, 1, 3, 2, 1, 2, 0, 6.$$

Calculate,

a) the mean;

b) the median;

c) and the mode,

of number of children.

Sample Mean, Median and Mode: Example

Solution:

a)

$$\overline{x} = \frac{4+2+0+0+4+3+0+1+3+2+1+2+0+6}{14} = 2$$

Sample Mean, Median and Mode: Example

Solution:

a)

$$\overline{x} = \frac{4+2+0+0+4+3+0+1+3+2+1+2+0+6}{14} = 2$$

b) Since $n = 14$, the median will be the average of the observations with rank 7 and 8. Ordering these observations, we obtain

$$0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 6$$

The 7th and 8th observations in this list are both 2. Thus, the Median is 2.

Sample Mean, Median and Mode: Example

Solution:

a)

$$\overline{x} = \frac{4+2+0+0+4+3+0+1+3+2+1+2+0+6}{14} = 2$$

b) Since $n = 14$, the median will be the average of the observations with rank 7 and 8. Ordering these observations, we obtain

$$0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 6$$

The 7th and 8th observations in this list are both 2. Thus, the Median is 2.

c) 0 appears most frequently (4 times). Thus, the mode is 0.

Outliers

The mean is much more sensitive to extreme values or errors in data, i.e., **outliers**, than the median.

Outliers

The mean is much more sensitive to extreme values or errors in data, i.e., **outliers**, than the median.

For example:

Data Set 1: $3.7, 4.6, 3.9, 3.7, 4.1$

median: 3.9 mean: 4

Outliers

The mean is much more sensitive to extreme values or errors in data, i.e., **outliers**, than the median.

For example:

Data Set 1: $3.7, 4.6, 3.9, 3.7, 4.1$

median: 3.9 mean: 4

For example:

Data Set 1: $37, 4.6, 3.9, 3.7, 4.1$

median: 4.1 mean: 10.66

Outliers

The mean is much more sensitive to extreme values or errors in data, i.e., **outliers**, than the median.

For example:

Data Set 1: $3.7, 4.6, 3.9, 3.7, 4.1$

median: 3.9 mean: 4

For example:

Data Set 1: $37, 4.6, 3.9, 3.7, 4.1$

median: 4.1 mean: 10.66

Note: Outliers are a real problem for data analysts. If a population truly contains outliers, but they are deleted from the sample, the sample will not characterize the population correctly.

Use of Median for Asymmetric distributions

If the distribution (histogram) is symmetric or close to being **symmetric**, then,

$$\overline{x} \approx \text{Median} \tag{6}$$

In this case, the sample might contain a few outliers. Either measure may be used (the mean is normally used as it has nicer statistical properties).

If the distribution is **clearly asymmetric** these measures will be significantly different.

In this case the median should be used as a measure of centrality, since there will be outliers which have a large influence on the mean.

The Trimmed Mean

Like the median, the **trimmed mean** is a measure of center that is designed to be unaffected by outliers.

The trimmed mean is computed by arranging the sample values in order, "trimming" an equal number of them from each end, e.g., $p\%$, which is called $p\%$ trimmed mean.

There are no hard-and-fast rules on how many values to trim. The most commonly used trimmed means are the 5%, 10%, and 20% trimmed means.

Since the number of data points trimmed must be a whole number, it is impossible in many cases to trim the exact percentage of data.

In this case, the simplest thing to do is to round $np/100$ to the nearest whole number and trim that amount where n is the sample size.

The Trimmed Mean: Example

Example 2: In the article "Evaluation of Low-Temperature Properties of HMA Mixtures" (P. Sebaaly, A. Lake, and J. Epps, Journal of Transportation Engineering, 2002: 578–583), the following values of fracture stress (in mega pascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

Compute the mean, median, and the 5%, 10%, and 20% trimmed means.

The Trimmed Mean: Example

Solution:

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

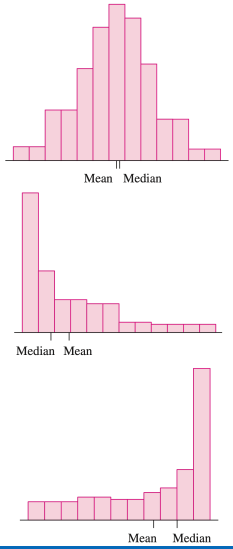
Mean is found by averaging all 24 numbers as 195.42.

Since $n = 24$, the median is the average of the observations with ranks 12 and 13, i.e., $(191 + 223)/2 = 207$.

To calculate the 5% trimmed mean, we must drop 5% of the data from each end, i.e., $(5 \times 24)/100 = 1.2$ which is rounded off to 1. Similarly, to compute 10% and 20% trimmed mean values, we round off $(10 \times 24)/100 = 2.4$ and $(20 \times 24)/100 = 4.8$ to 2 and 5. Thus, the 5%, 10%, and 20% trimmed means can be found as 190.45, 186.55, and 194.07, respectively.

Symmetry and Skewness

- A histogram is perfectly **symmetric** if its right half is a mirror image of its left half.
- Histograms that are not symmetric are referred to as skewed. In practice, no sample has a perfectly symmetric histogram.
- A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**.
- For a histogram that is skewed to the right, the mean is greater than the median as the mean is near the center of mass of the histogram.
- A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**.



Measures of Variability (Dispersion)
Range

These measures show the amount of dispersion in the variable. Dispersion is the degree to which the data are spread out.

1. **Range:** The range, R, is the difference between the largest and the smallest values in a sample, i.e.,

$$R = x_{(n)} - x_{(1)}. \tag{7}$$

This is simple to calculate, but is very sensitive to extreme values.

Since it depends only on the two extreme values, it is rarely used, as it provides no information about the rest of the sample.

Measures of Variability (Dispersion)

The Sample Variance

2. **Sample Variance:** The sample variance measures how closely the data are placed around the mean.

If all the data are equal, then the variance will be zero. The more dispersed data around the mean, the greater the variance.

The sample variance is given by s^2, where,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big). \tag{8}$$

It is a measure of the average squared deviations from the mean.

Note: A serious drawback of the sample variance as a measure of spread is that its units are different than the sample values, i.e., they are the squared units!

Measures of Variability (Dispersion)

Standard Deviation

3. **Standard deviation:** To obtain a measure of spread whose units are the same as those of the sample values, we simply take the square root of the variance. This quantity is known as the sample standard deviation, s, i.e.,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big)}. \tag{9}$$

It is preferred to the variance as a descriptive measure, since it is a measure of the average distance of an observation from the mean.

It is thus measured in the same units as the observations themselves.

Note: On scientific calculators the standard deviation is denoted as s_{n-1} or σ_{n-1}.

Measures of Variability (Dispersion)

Coefficient of Variation

4. **Coefficient of variation (CV):** The coefficient of variation is a measure that allows for comparison in spread between two or more data sets by describing the amount of spread per unit mean, i.e.,

$$\mathrm{CV} = \frac{s}{\bar{x}}. \tag{10}$$

CV is calculated by dividing the standard deviation by the sample mean. The sample with the largest CV has the largest relative spread.

A major advantage of the coefficient of variation is that it is unitless and thus, it is insensitive to units.

Measures of Variability: Example

Example 3: For the following data sets, calculate the coefficient of variation.

$\mathcal{D}_1 = \{100, 100, 100\}$ $\mathcal{D}_2 = \{90, 100, 110\}$ $\mathcal{D}_3 = \{8, 10, 12, 14, 16, 18, 20\}$

Measures of Variability: Example

Example 3: For the following data sets, calculate the coefficient of variation.

$\mathcal{D}_1 = \{100, 100, 100\}$ $\mathcal{D}_2 = \{90, 100, 110\}$ $\mathcal{D}_3 = \{8, 10, 12, 14, 16, 18, 20\}$

Solution:

Since all the data in $\mathcal{D}_1$ is constant, its standard deviation is 0 and its average is 100. Thus, $\mathrm{CV}_1\% = 100 \times (\frac{0}{100}) = 0\%$.

Measures of Variability: Example

Example 3: For the following data sets, calculate the coefficient of variation.

$\mathcal{D}_1 = \{100, 100, 100\}$ $\mathcal{D}_2 = \{90, 100, 110\}$ $\mathcal{D}_3 = \{8, 10, 12, 14, 16, 18, 20\}$

Solution:

Since all the data in $\mathcal{D}_1$ is constant, its standard deviation is 0 and its average is 100. Thus, $\mathrm{CV}_1\% = 100 \times (\frac{0}{100}) = 0\%$.

The data in $\mathcal{D}_2$ has more variability. Its standard deviation is 10 and its average is 100. Thus, $\mathrm{CV}_2\% = 100 \times (\frac{10}{100}) = 10\%$.

Measures of Variability: Example

Example 3: For the following data sets, calculate the coefficient of variation.

$\mathcal{D}_1 = \{100, 100, 100\}$ $\mathcal{D}_2 = \{90, 100, 110\}$ $\mathcal{D}_3 = \{8, 10, 12, 14, 16, 18, 20\}$

Solution:

Since all the data in $\mathcal{D}_1$ is constant, its standard deviation is 0 and its average is 100. Thus, $\mathrm{CV}_1\% = 100 \times (\frac{0}{100}) = 0\%$.

The data in $\mathcal{D}_2$ has more variability. Its standard deviation is 10 and its average is 100. Thus, $\mathrm{CV}_2\% = 100 \times (\frac{10}{100}) = 10\%$.

The data in $\mathcal{D}_3$ has more variability again. Its standard deviation is approximately 4.32 and its average is 14. Thus, $\mathrm{CV}_3\% = 100 \times (\frac{4.32}{14}) = 30.85\%$.

Measures of Variability: Example

Example 4: Consider the data set in Example 1, i.e.,

$$0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 6.$$

Calculate the sample variance, standard deviation, range and coefficient of variance.

Measures of Variability: Example

Example 4: Consider the data set in Example 1, i.e.,

$$0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 6.$$

Calculate the sample variance, standard deviation, range and coefficient of variance.

Solution:

In Example 1, we calculated $\overline{x} = 2$, thus, the sample variance can be obtained as

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{4(0-2)^2 + 2(1-2)^2 + 2(3-2)^2 + 2(4-2)^2 + (6-2)^2}{13} = \frac{44}{13} \approx 3.3846 \end{aligned}$$

Measures of Variability: Example

Solution (continued):

The standard deviation is given by,

$$s = \sqrt{s^2} = \sqrt{3.3846} \approx 1.8397$$

Measures of Variability: Example

Solution (continued):

The standard deviation is given by,

$$s = \sqrt{s^2} = \sqrt{3.3846} \approx 1.8397$$

The range is given by $R = 6 - 0 = 6$ as the largest and smallest observations are 6 and 0, respectively.

Measures of Variability: Example

Solution (continued):

The standard deviation is given by,

$$s = \sqrt{s^2} = \sqrt{3.3846} \approx 1.8397$$

The range is given by $R = 6 - 0 = 6$ as the largest and smallest observations are 6 and 0, respectively.

The coefficient of variance is given by,

$$\mathrm{CV} = \frac{1.8397}{2} \approx 0.9199$$

Grouped Data

Data from a discrete distribution may well be displayed in tabular form, e.g. the data from the previous example may be presented as:

Number of children	0	1	2	3	4	5	6
Number of families	4	2	3	2	2	0	1

The first row gives the value of the observation x_i (here no. of children).

The second row gives the frequency of each observation n_i (the number of observations of x_i children). The sum of these numbers is the total number of observations.

Since 3 families have 2 children, these families account for $3 \times 2 = 6$ children (in mathematical notation $n_3 \times 2$, where n_3 is the number of times 2 children were observed).

The Sample Mean and Variance of Grouped Data

It can be seen that summing $x_i n_i$ over all the possible observations of the variable (number of children), we obtain the sum of the observations.

It follows that the sample mean is given by,

$$\overline{x} = \frac{\sum x_i n_i}{\sum n_i} \tag{11}$$

Similarly, the sample variance and standard deviation are given by,

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2 n_i}{(\sum_i n_i) - 1}, \qquad s = \sqrt{\frac{\sum_i (x_i - \overline{x})^2 n_i}{(\sum_i n_i) - 1}},$$

where $n = \sum_i n_i$ is the total number of observations.

Grouped Data

Calculation of the Sample Mean and Variance

The above formulae can be used to calculate the mean and the sample variance in the following way.

If we are not given the total number of observations, n, we first calculate it,

$$n = \sum_{i=1}^{7} n_i = 4 + 2 + 3 + 2 + 2 + 0 + 1 = 14.$$

The sample mean is then given by,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{7} x_i n_i = \frac{(0 \times 4) + (1 \times 2) + (2 \times 3) + (3 \times 2) + (4 \times 2) + (0 \times 5) + (6 \times 1)}{14}$$

$$= 2.$$

Grouped Data

Calculation of the Sample Mean and Variance

The sample variance is given by,

$$
\begin{aligned}
s^2 &= \frac{1}{(\sum_i n_i) - 1} \sum_i (x_i - \overline{x})^2 n_i \\
&= \frac{4(0-2)^2 + 2(1-2)^2 + 2(3-2)^2 + 2(4-2)^2 + (6-2)^2}{13} = \frac{44}{13} \approx 3.3846
\end{aligned}
$$

Grouped Data

Calculation of the Sample Mean and Variance

The sample variance is given by,

$$s^2 = \frac{1}{(\sum_i n_i) - 1} \sum_i (x_i - \overline{x})^2 n_i$$
$$= \frac{4(0-2)^2 + 2(1-2)^2 + 2(3-2)^2 + 2(4-2)^2 + (6-2)^2}{13} = \frac{44}{13} \approx 3.3846$$

Note: The median for the grouped data is calculated in the same way as for data given individually, by ranking (ordering) the data.

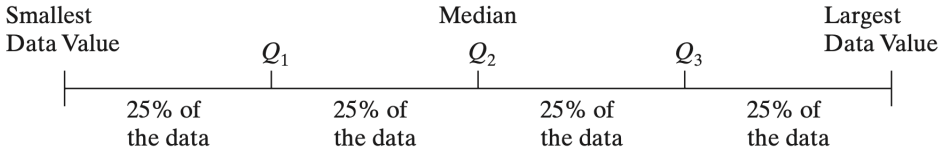
Percentiles and Box Plots

The k^{th} percentile of a sample for k between 0 and 100, divides the sample so that as nearly as possible k% of the sample values are less than the k^{th} percentile, and $(100 - k)$% are greater.

The k^{th} percentile is denoted by P_k.

The percentiles that split the ordered data into 4 subsamples of equal size are called **quartiles**.

The median divides the sample in half. **Quartiles** divide it as nearly as possible into quarters. A sample has three quartiles.



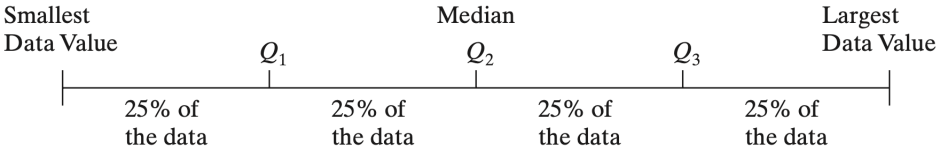
Percentiles and Box Plots
Finding Quartiles

Quartiles are found by taking the following steps:

Step 1 Arrange the data in ascending order.

Step 2 Determine the median, M, or second quartile, Q_2.

Step 3 Divide the data set into halves: the observations below (to the left of) M and the observations above M. The first quartile, Q_1, is the median of the bottom half of the data and the third quartile, $Q3$, is the median of the top half of the data.



Percentiles and Box Plots
The Interquartile Range (IQR)

The interquartile range (IQR) is a measure of the spread of the data.

IQR is the range of the middle 50% of the observations in a data set. It is the difference between the third and first quartiles and is found using the formula

$$\text{IQR} = Q_3 - Q_1. \tag{12}$$

The interpretation of the interquartile range is similar to that of the range and standard deviation.

The more spread a set of data has, the higher the interquartile range will be.

Percentiles and Box Plots

Box Plots

A **box plot** is a graphic that presents the median, the first and third quartiles, and any outliers that are present in a sample.

For the purpose of drawing box plots, by convention, any point that is more than $1.5 \times \mathrm{IQR}$ above the third quartile, or more than $1.5 \times \mathrm{IQR}$ below the first quartile, is considered an *outlier*.

Some texts define a point that is more than $3 \times \mathrm{IQR}$ from the first or third quartile as an **extreme outlier**.

The bottom and top side of the box plot are the first and third quartiles, respectively.

An horizontal line is drawn at the median, the "outliers" are plotted individually and are indicated by crosses in the plot.

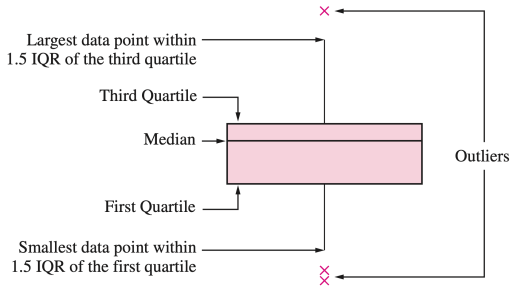
Extending from the top and bottom of the box are vertical lines called "whiskers".

Construction of a Box plot

Step 1 Compute the median and the first and third quartiles of the sample.

Step 2 Find the largest sample value that is no more than $1.5 \times \mathrm{IQR}$ above the third quartile, and the smallest sample value that is no more than $1.5 \times \mathrm{IQR}$ below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points.

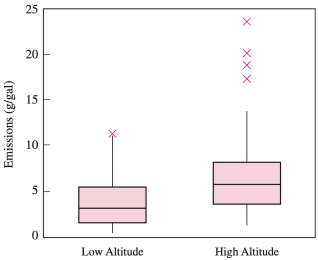
Step 3 Points more than $1.5 \times \mathrm{IQR}$ above the third quartile, or more than $1.5 \times \mathrm{IQR}$ below the first quartile, are designated as outliers. Plot each outlier individually.



Description and Comparison of Box Plots

A major advantage of boxplots is that several of them may be placed side by side, allowing for easy visual comparison of the features of several samples.

As an example, comparative box plots for PM emissions data for vehicles driven at high versus low altitudes are shown in the figure on the right.



The box plots show that vehicles driven at low altitude tend to have lower emissions.

There are several outliers among the data for high-altitude vehicles whose values are much larger than any of the values for the low- altitude vehicles.

We conclude that at high altitudes, vehicles have somewhat higher emissions in general, the spread in values is slightly larger than low-altitude vehicles and it is much larger when the outliers are considered.

Symmetry and Skewness

The figures below show three histograms and their corresponding box plots for different data sets that are skewed to the right, symmetric and skewed to the left.