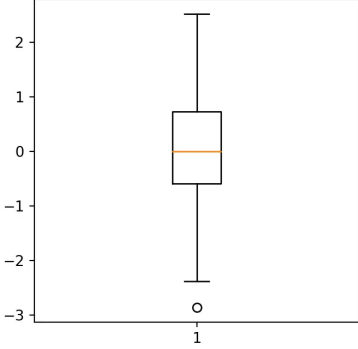


# DATA1001 Topic 3 Flashcards

## Numerical Summaries

Abyan Majid

Question	My answer
What is a numerical summary? Identify one major advantage and one major disadvantage that comes with it!	A numerical summary groups multiple information about data into one number called a statistic.  Advantage: It eases communication of insights about some features of the data  Disadvantage: The grouping of many values into a statistic often loses much information.
What are major features of data that can and is commonly summarized numerically?	Centre, spread, maximum, minimum
What is a mean? Recite the formula for mean!	Mean represents the average of a set of numbers/data points. $\text{Mean} = \frac{\text{Sum of data points}}{\text{Number of data points}}$ Or formally, $\text{Mean} = \frac{\sum_{i=1}^N x_i}{N}$
Is mean the balancing point of data? Why or why not?	Yes, mean is the balancing point of the data because it is the point from which all lower and upper readings cancel each other out.
What is a median? Recite the formulas for a median given an odd-sized data and even-sized data respectively!	Median represents the middle number if you were to sort the set of data points in ascending order.  $\text{Median} = X[\frac{n+1}{2}] \text{ for even-sized data}$ $\text{Median} = \frac{X[\frac{n}{2}] + X[\frac{n+1}{2}]}{2} \text{ for odd-sized data}$
What do we mean when we say that a numerical summary is robust? That said, is mean robust and is median robust?	A numerical summary is robust when it is not significantly affected by outliers. Median is robust, mean is not.
Fill in the three ellipses!  For symmetric data, we expect the mean and median to be ...	(1) Relatively the same/similar (2) smaller than the (3) greater than the

<p>For left-skewed data, we expect the mean to be ... median</p> <p>For right-skewed data, we expect the mean to be ... median</p>	
 <ol style="list-style-type: none"> <li>1. What does the center line on the boxplot represent?</li> <li>2. What do the sides of the box that are connected to the whiskers represent?</li> <li>3. What does the length of the box represent?</li> <li>4. What do the lines at the end of the whiskers represent?</li> <li>5. What does any point beyond the whiskers represent?</li> </ol>	<ol style="list-style-type: none"> <li>1. Median</li> <li>2. Lower and upper quartiles</li> <li>3. Interquartile range</li> <li>4. Lower and upper thresholds/bounds</li> <li>5. Outliers</li> </ol>
<p>Which of mean or median, or both is/are optimal for describing center, given:</p> <ol style="list-style-type: none"> <li>1. A symmetric data</li> <li>2. Skewed data</li> </ol>	<ol style="list-style-type: none"> <li>1. Both mean and median are arguably equally optimal</li> <li>2. Median</li> </ol>
<p>What do we do with each point in data in order to compute spread?</p>	<p>We measure the gap between the data point and the mean</p>
<p>What does Root Mean Square (RMS) do? What are the steps?</p>	<p>Root Mean Square computes the average of a set of data points regardless of sign. To do an RMS, we follow its name in reverse:</p> <ol style="list-style-type: none"> <li>1. Square the numbers</li> <li>2. Compute the mean</li> <li>3. Root the mean</li> </ol>
<p>Recite the formula for population and sample standard deviations in terms of RMS? Then, identify how we can use R to find each!</p>	$SD_{pop} = RMS\ of\ gaps = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$

	$SD_{sample} = \text{Adjusted RMS of gaps} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
How to find variance?	Square the standard deviation (assuming it was given)
What do we mean by the Z-score (standard unit) of a data point? Recite the formula for Z-score!	<p>Z-score is a measure of how many standard deviations a data point is from the mean.</p> $Z = \frac{\text{Data point} - \text{Mean}}{SD}$
What is interquartile range (IQR)? Recite the formula for IQR!	<p>Interquartile range represents the middle 50% of the data.</p> $IQR = Q_3 - Q_1, \text{ where}$ <ul style="list-style-type: none"> <li>• <math>Q_3</math> is the upper quartile (75th percentile)</li> <li>• <math>Q_1</math> is the lower quartile (25th percentile)</li> </ul>
How do we identify an outlier? Recite any formulas where necessary!	<p>We consider any data point beyond the lower and upper thresholds to be an outlier.</p> $LT = Q_1 - (1.5 \times IQR)$ $UT = Q_3 + (1.5 \times IQR)$ <p>Any data point smaller than LT or greater than UT is considered an outlier</p>
We can report the center and spread as a pair of values. Identify two commonly reported center-spread pairs!	<p>(Mean, SD) (Median, IQR)</p>
What is the coefficient of variation? Recite its formula!	<p>Coefficient of variation is a statistic that combines mean and standard deviation.</p> $CV = \frac{SD}{Mean}$
Suppose you have data of size $N = 15$ , with sample standard deviation $SD = 2.7$ . What is the population standard deviation?	<p>By the formula</p> $SD_{pop} = SD_{sample} \times \sqrt{\frac{n-1}{n}}$ <p>We have</p> $SD_{pop} = 2.7 \times \sqrt{\frac{14}{15}} \approx 2.61$