Lecture 2

Descriptive statistics

Most of the statistics we will cover are for numerical data. Categorical data uses mostly proportions only – what fraction of the population falls into that category. Frequency tables can be used to give an overview of the counts or proportions in a data set.

Frequency Tables – a table that lists the categories/levels of a data set and the counts from the data for each category.

Method of Travelling	Number of children
Walking	8
Car	9
Bus	4
Cycle	5
Train	1
Taxi	3

A Relative Frequency table converts the counts to percentages.

Method of Traveling	**Number of Children**	**Relative Frequency of Children**
Walking	8	$\frac{8}{30} \approx 26.7\%$
Car	9	$\frac{9}{30} = 30\%$
Bus	4	$\frac{4}{30} \approx 13.3\%$
Cycle	5	$\frac{5}{30} \approx 16.7\%$
Train	1	$\frac{1}{30} \approx 3.3\%$
Taxi	3	$\frac{3}{30} = 10\%$
Total	30	$\frac{30}{30} = 100\%$

We can also make cumulative Frequency and Cumulative Relative Frequency to the table.

Method of Traveling	Number of Children	Cumulative Number of Children	Relative Frequency of Children	Cumulative Relative Frequency of Children
Walking	8	8	$\frac{8}{30} \approx 26.7\%$	$\frac{8}{30} \approx 26.7\%$
Car	9	17	$\frac{9}{30} = 30\%$	$\frac{17}{30} \approx 56.7\%$
Bus	4	21	$\frac{4}{30} \approx 13.3\%$	$\frac{21}{30} = 70\%$
Cycle	5	26	$\frac{5}{30} \approx 16.7\%$	$\frac{26}{30} \approx 86.7\%$
Train	1	27	$\frac{1}{30} \approx 3.3\%$	$\frac{27}{30} = 90\%$
Taxi	3	30	$\frac{3}{30} = 10\%$	$\frac{30}{30} = 100\%$
Total	30		$\frac{30}{30} = 100\%$	

The cumulative frequency is the sum of the count in the current category plus all the counts in the preceding categories. The cumulative relative frequency is similar, but converted to a percentage. In the cumulative frequency, the last category should always display the total.

Tables like this can also be used to get an initial glimpse of numerical data as well. This is easily done for discrete variables, especially when the values are few in number. In most cases, numerical data is binned (grouped into ranges, similarly as we will see to what is needed to build a histogram by hand).

Category	Frequency
10-19	0
20-29	1
30-39	3
40-49	7
50-59	9
60-69	12
70-79	7
80-89	3
90-99	1
	43

Descriptive statistics for numerical data come in several types: measures of center, measures of variation and measures of location.

Measures of center

Think of measures of center as an answer to the question "what is a typical value for the observation?"

Mean = average. Add up all the observations and divide by the total number of observations. Means are affected by extreme values. Round your answer to one decimal place more than the original data.

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

μ is the symbol used for population means, while $\bar{x}$ for samples.

Median – is the value that divides the ordered data set into two equal halves. 50th percentile. If the data set has an odd number of values, then the location of the median in the ordered list of observations is $\frac{n+1}{2}$. If there is an even number of observations in the data set, then this formula will give you a fraction. The median will be the average of the two values at the floor of this number (round down), and the one at the ceiling of this number.

Suppose there are 17 observations. Then $\frac{n+1}{2} = \frac{18}{2} = 9$. Arrange the observations in order from smallest to largest. The ninth observation is the median.

Suppose instead that there are 18 observations. Then $\frac{n+1}{2} = \frac{19}{2} = 9.5$. After arranging the values in order as before, you want the to average the observations at the 9th position and the 10th position to obtain the median.

The median is not impacted by extreme values and so it is often preferred when the data is not symmetrically distributed (will say more about this in future lectures).

There is no common notation for the median, but some books use $\tilde{\mu}$ for population median, but others just abbreviate Med, or other m (with M for mode) or other methods. You should adopt a consistent notation. I will use $\tilde{x}$ for the median and M for the mode.

Mode – the most is the value in the data set that appears most often. Sometimes, more than one value will appear most frequently (the biggest categories have the same counts). In general, we report modes if there is one mode, or two, but many texts recommend resorting to "no mode" if there are three or more modes. Continuous data frequently has no repeated values and thus no mode.

We can extend the definition of mode to something more useful: the idea of a modal class. When the data is in categories—either because the data is categorical or because it has been binned into groups, then the category with the highest count is the modal class. This is useful when looking at histograms in the next lecture. If we look back at our table of the methods children get to school, "car" is the model class. For the table of what looks like test scores, the modal class is 60-69.

Weighted average – we can use this method for calculating a mean from a frequency table (or estimating it for non-discrete data) – or for calculating course grades.

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

To calculate, multiply each observation by its weight (or frequency) and then divide not by the number of observations, but by the sum of the weights. We'll see in a future lecture that is the same method used to calculate the mean of a discrete probability distribution.

Trimmed mean – the trimmed mean is something of a compromise between the mean and the median. Trimmed means are described by the percentage of values trimmed off of each end of the data. For example, a 10%-trimmed mean takes 10% of the data off the high end, and another 10% off the low end, leaving only 80% of the data left to take the mean of. By removing extreme values on both ends, this tends to make the trimmed mean more similar to the value of the median, but can still be affected by data with a strong skew. The notation used is to indicate the percentage trimmed: $\bar{x}_{tr(10)}$ is a 10% trimmed mean, or sometimes you may see $\bar{x}_{10\%}$.

Measures of variation

Variance – variance is a method of measuring the absolute distance from the mean of observations in a data set. To make the distances positive, the formula squares values rather than using the absolute value (partly to make the calculus easier). The variance of a population uses σ^2, while the sample statistic is s^2.

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

The $n - 1$ is something that we'll discuss more in a future lecture on unbiased estimators (after the first exam). If you have the whole population, then the formula becomes

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

In general, we will not be working with the whole population, so don't use this formula unless the problem specifically mentions it's population data. We will derive this version of the formula at some point later in the course.

Some texts will abbreviate $(x_i - \bar{x})^2$ as S_{xx} to make creating complex formulas a bit easier to read.

Standard deviation – this standard deviation is the square root of the variance. Since we work with standard deviation (when building confidence intervals and doing hypothesis tests) more often than directly with variance, the variance takes the squared notation rather than using the square root to define the standard deviation. We use σ for the population value, and s for the sample standard deviation.

Take the previous formula and take the square root.

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

Conceptually, we can think of this as approximately the average distance to the mean. It's not quite the same (which would be mean absolute deviation), but that's the idea. Absolute values are problematic for calculus, though.

Typically, round your standard deviation values to two decimal places more than was used in the original data.

There are lots of alternative methods of calculating something like a standard deviation. Median absolute deviation, or even the IQR described below. These alternatives are uncommon but do sometimes find useful purchase in specific fields.

Range – the range is not generally considered a reliable descriptive statistic because it tends to increase as the sample size increases. It is calculated from the difference between the maximum value and the minimum value.

IQR – the interquartile range is the range of the middle 50% of the data. We will more carefully define the quartiles below, but the IQR is the difference between the third quartile and the first quartile. You can think of it also similarly to the trimmed mean: drop the top and bottom 25% of the data (from each end), and then calculate the range of what remains. We will see that this is an important value when constructing box plots, and is often used as an alternative to standard deviation if the data set is strongly skewed.

Measures of Position/Location

Rank – rank is the position of an observation in an ordered list. We used rank to calculate the median. Rank is not used often in parametric statistics, but we will see later in the course that rank is sometimes used in calculating nonparametric statistics.

Quartiles – Quartiles are generally calculated as an extension of our procedures for finding the median. The process goes something like this: Find the median and then divide the dataset into two halves. Then the first quartile is the median of the lower half, and the third quartile is the median of the upper half. One potential complication here is whether to include the median (if it's an observation) in neither half, or both halves. There is not a definitive answer to this question. Indeed, Excel includes formulas for both methods, but the default approach seems to be to include it in both halves if the median is an observation.

The first quartile thus roughly represents the point at which 25% of the data is less than or equal to that value. The third quartile roughly represents the point at which 75% of the data is smaller than the value. Sometimes the median is referred to as the second quartile. The 0^{th} quartile is the minimum. The 4^{th} quartile is the maximum.

The first quartile is often abbreviated Q1, and the third quartile Q3.

Percentiles – percentiles divide the data set into 100 groups instead of just 4. The 70th percentile means that 70% of the data is at or below the specified value. Percentiles are always defined by the percentage of the data less than the given value.

It's typical to round percentiles to the nearest whole percentage in the middle of the data set, but when we get to percentiles below 1% or above 99%, we don't round as much. It's common to refer to 99.9th percentile. 0th percentile and 100th percentile are almost never used unless they are the absolute minimum or maximum value possible.

Calculating a percentile by hand involves ordering the list of observations and finding it's rank (its location in the ordered list). Divide that location position by the total number of observations.

$$percentile = \frac{rank}{n}$$

To find the value that corresponds to a particular percentile, multiply the percentile by the number of observations. This will give you the approximate rank (location in the ordered list). Choose the closest observation. Some people will also estimate the value in the middle of a gap between observations.

$$rank = percentile \times n$$

Extreme Values

Unusual Values – unusual values are typically defined as values which occur less than 5% of the time. In a symmetric data set, this is equivalent to being more than two standard deviations from the mean, so this is an equivalent method of calculating them instead of percentiles. The 5% occurrence rate can also be used with categorical data, however.

Fences – the term fences comes from its use in constructing boxplots which we'll look at next week. Finding the fences uses the IQR.

Lower fence:

$$lower\ fence = Q1 - 1.5 \times IQR$$

Upper fence:

$$upper\ fence = Q3 + 1.5 \times IQR$$

This produces a number in each case. Values that are lower than the lower fence are considered extreme values or outliers. Values that are above the upper fence are also considered extreme values or outliers.

Extreme outliers can be found outside the range of the two outer fences:

Lower outer fence:

$$LOF = Q1 - 3 \times IQR$$

Upper outer fence:

$$UOF = Q3 + 3 \times IQR$$

Extreme outliers can also be found using 3 standard deviations from the mean (similar to the procedure for the unusual values but further away).

Summary tables

Summary tables of statistics can be generated in many statistical software packages, including R, that quickly calculate a range of common statistics. A typical example looks like this:

Table 1:

Statistic	Mean	St. Dev.	Median	Min	Max
rating	64.63	12.17	65.5	40	85
complaints	66.60	13.31	65	37	90
privileges	53.13	12.24	51.5	30	83
learning	56.37	11.74	56.5	34	75
raises	64.63	10.40	63.5	43	88
critical	74.77	9.89	77.5	49	92
advance	42.93	10.29	41	25	72

Different packages may calculate different statistics.

R may produce a table like this (here, in subgroups within the data)

```
 Descriptive statistics by group
group: java
           vars n mean   sd min  max range  se
id            1 2  1.5 0.71   1    2     1 0.5
subjects      2 2  NaN   NA Inf -Inf  -Inf  NA
marks         3 2 89.5 0.71  89   90     1 0.5
percentage    4 2 83.5 7.78  78   89    11 5.5
------------------------------------------------------------------------
group: python
           vars n mean   sd min  max range  se
id            1 2  3.5 0.71   3    4     1 0.5
subjects      2 2  NaN   NA Inf -Inf  -Inf  NA
marks         3 2 83.0 8.49  77   89    12 6.0
percentage    4 2 72.0 8.49  66   78    12 6.0
------------------------------------------------------------------------
group: R
           vars n mean sd min  max range se
id            1 1    5 NA   5    5     0 NA
subjects      2 1  NaN NA Inf -Inf  -Inf NA
marks         3 1   89 NA  89   89     0 NA
percentage    4 1   90 NA  90   90     0 NA
```

Excel's summary statistics looks like this:

	A	B	C	D	E
1	Scores		*Column1*		
2	82				
3	93		Mean	81.21428571	
4	91		Standard Error	4.045318243	
5	69		Median	85	
6	96		Mode	93	
7	61		Standard Deviation	15.13619489	
8	88		Sample Variance	229.1043956	
9	58		Kurtosis	-1.426053506	
10	59		Skewness	-0.402108004	
11	100		Range	42	
12	93		Minimum	58	
13	71		Maximum	100	
14	78		Sum	1137	
15	98		Count	14	
16					

How to do these calculations in R will be covered in the lab.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. https://www.excel-easy.com/examples/descriptive-statistics.html
4. https://www.geeksforgeeks.org/how-to-create-summary-tables-in-r/
5. https://i.stack.imgur.com/YZz9g.png
6. https://en.wikipedia.org/wiki/Median_absolute_deviation
7. https://www.softschools.com/math/probability_and_statistics/frequency_table_numerical_data_categories_are_a_range_of_values/
8. https://www.investopedia.com/ask/answers/021215/what-difference-between-standard-deviation-and-average-deviation.asp
9. https://www.pinterest.com/pin/456974693425279527/