**Summarizing Data – Numerical Methods**

**Topics**

**Measures of Centre**

**Measures of Position**

**Measures of Variation**

**Measures of Central Tendency:**

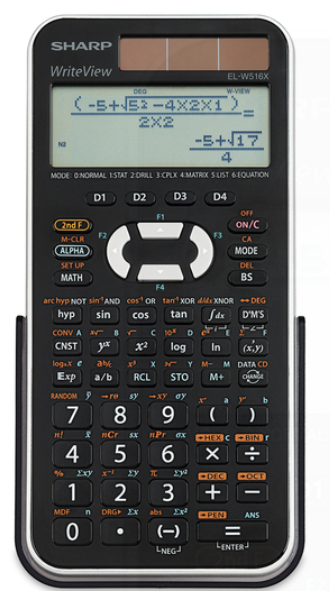
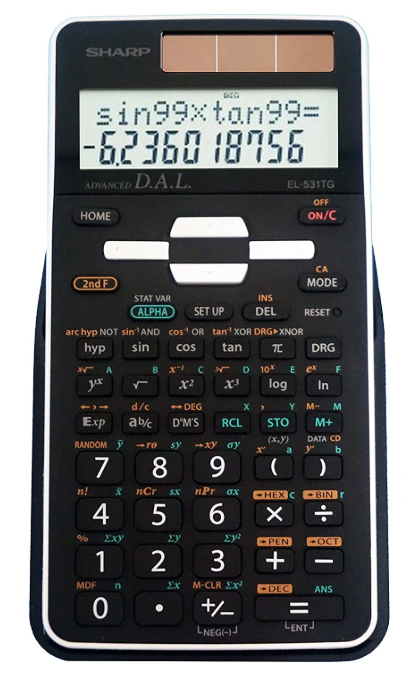
**Measures of central tendency are single numerical values (in most instances) which are intended to indicate the center or middle region of the distribution of values. The most common measure of central tendency is the *mean* , which is more precisely known as the *arithmetic mean* and more generally known as the *average.***

**The arithmetic mean of a set of scores is the value obtained by adding the scores and dividing the total by the number of scores.**

**Sample mean  (statistic) n is the sample size**

**Population mean  (parameter) N is the size of the population**

**You can use your calculator’s statistical functions to compute the mean:**

****

**We will do this with the first 20 values for eruption times for Old Faithful:**

3.6 1.8 3.3 2.3 4.5 2.9 4.7 3.6 1.9 4.3 1.8 3.9 4.2 1.8 4.7 2.2

1.8 4.8 1.6 4.2

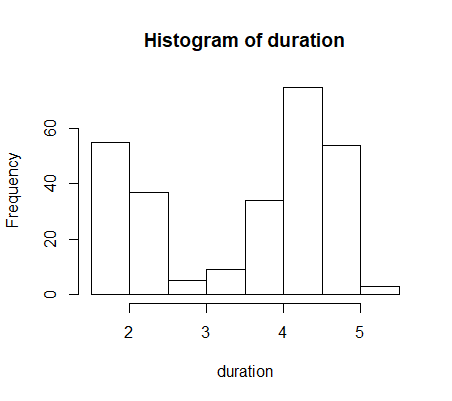
**Clearly, you would not want to use your calculator to find means of very large datasets. In R, the command mean(faithful$eruptions) will find the mean eruption time for the entire *faithful* dataset.**

**Note: the mean duration for all 272 eruptions of Old Faithful is 3.488 minutes.**

**Mean from grouped data**

**Technical journals often only present summaries. When data are summarized in a frequency table or frequency histogram, we can approximate the mean by assuming each value in a class equals the class midpoint (commonly called *class mark*).**

**Suppose we only have the following frequency histogram:**



**Approximate mean for grouped data:**



x = class mark

= mid-point of the classes

= mean of the upper and

lower class limit of a class

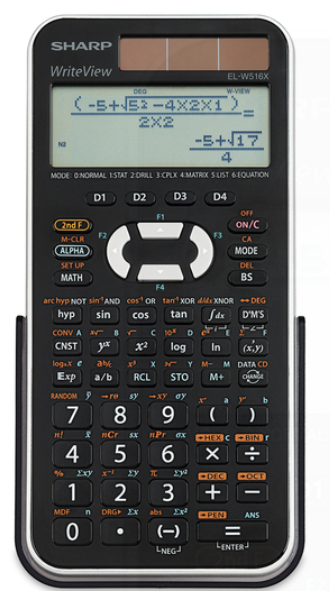
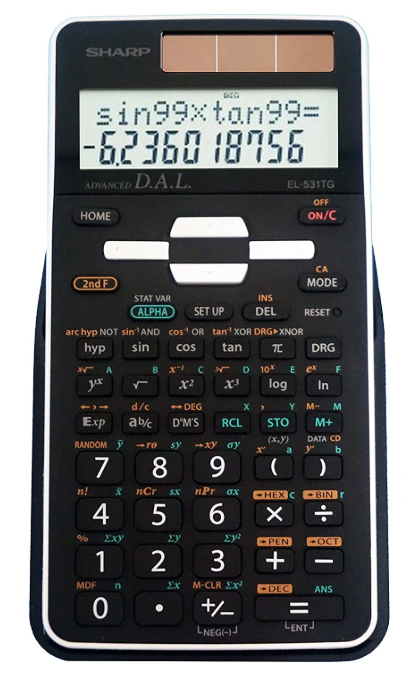
f = frequency

**Our best estimate of the mean approximates each duration with the centre of its class, or *class mark*.**

**The class marks are: 1.75, 2.25, 2.75, 3.25, 3.75, 4.25, 4.75, 5.25**

**The frequencies are : 51 41 5 7 30 73 61 4**

**You can use your calculator’s statistical functions to compute group mean.**

****

**The estimated mean is 3.498 minutes.**

**The actual value was 3.488 minutes.**

**Median**

**The median of a set of scores is the middle value when the scores are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by  (pronounced “x-tilde”)**

**If there is an odd number of data points, the median is located in the exact middle of the data set.**

**If there is an even number of data points, the median is found by computing the mean of the two middle numbers.**

**The median duration for the Old Faithful eruptions is 4 minutes.**

**Note that this is noticeably different from the mean value. If we have a set of data that is symmetric, then the mean and median will be exactly equal. The difference between the mean and median gives a measure of how *skewed* the data is. We will quantify the degree of skew once we have some more tools.**

**Quartiles, Percentiles**

**The median gave us a number that was larger than half of the data. We can generalize this concept.**

**The kth –Percentile, Pk, is a value, which divides the data or population into two parts: the lower k% of the values and the upper (100-k)% of the values.**

Pk

**Note: Data must be sorted in order to locate percentiles. In R, the command quantile(faithful$eruptions, 0.42) will find the duration that is larger than 42% of eruptions**

quantile(faithful$eruptions, 0.42)

42%

3.73012

**Here, we can see that 42% of eruptions from Old Faithful are 3.73012 minutes or less in duration.**

**Note that there are no eruptions that are exactly 3.73012 minutes in duration. There are several choices for P42. Statistical software uses interpolation methods; different types of statistical software may give slightly different results. However, all methods will return values between 3.717 (which is larger than 41.9% of the data) and 3.733 (which is larger than 42.3% of the data). The differences in methods will be more noticeable in small datasets, or when the data is very spread apart.**

**Quartiles are just percentiles in multiples of 25.**

* **the first quartile, Q1 = P25 , value separating the elements of a population or set of data into the lower 25% and the upper 75%.**

**25%**

**25%**

**25%**

**25%**

Q3

Q2

Q1

* **Q2 = P50 , Q3 = P75.**

**By default, in R, the quantile() command provides quartile data.**

> quantile(faithful$eruptions)

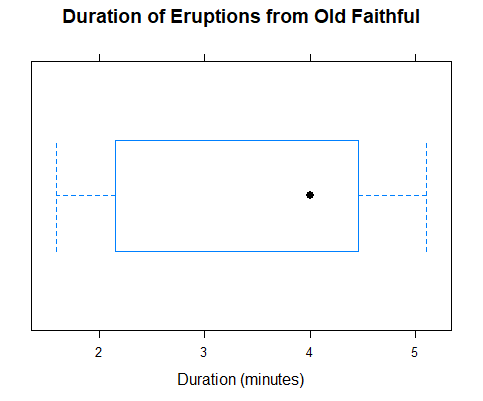
0% 25% 50% 75% 100%

1.600 2.163 4.000 4.454 5.100

**This very simple output gives us very detailed and useful information about the spread of the data:**

**It may be more useful to see this information visually. We can do this with a *boxplot* (or *box-and-whisker plot*). We do this by first obtaining the *five-number summary,* which consists of the minimum, Q1, median, Q3, and maximum of a set of data. The quantile function gives all of this information by default.**

**Boxplots are basically a way of graphing a five-number summary. Here is one we created from the *faithful* data in R:**

**  
Note that the boxplot divides the data into four groups, each containing a quarter of the data. A glance at a boxplot then tells you visually the span of values between the smallest and largest observation, the interval containing the middle 50% of the observations (the box), and the location of the center of the data, as represented by the median.**

**What does this boxplot tell us about the frequency distribution of this set of data?**

**Although single boxplots by themselves are informative, probably the most common use of the boxplot is in comparing two or more sets of data. The relative horizontal positions of the side-by-side boxplots allow you to compare the general distributions of the various data sets.**

**Example: Two types of drugs were tested on 10 students who had trouble sleeping. The dataframe *sleep* provides results. *Group* gives the two kinds of drugs (listed here as just “1” and “2”). *Extra* gives the number of extra hours of sleep that each student had on each kind of drug. We can get a quick visual representation of the effectiveness of the two types of drugs by creating side-by-side boxplots for each kind of drug:**

|  |  |
| --- | --- |
|  |  |

**What can we say about the two types of drugs based on this data?**

**Outliers**

**A difficult issue in statistical work is the question of what to do about outliers (suspicious or very unusual observations).**

**Are unusual values the results of errors or are the naturally occurring phenomena?**

**Should we delete outliers?**

**There is no absolute solution, because that would require people to know when they made a mistake that they don’t know about!**

**We need an objective rule to identify possible outliers so that the decision does not depend on the individual researcher’s mood.**

**Outliers:**

**One of the most commonly used rules for identifying outliers is based on the five-number summary and the IQR (*Inter quartile Range* = Q3 – Q1 ).**

**lower inner fence = Q1 - 1.5 IQR**

**upper inner fence = Q3 + 1.5 IQR**

**lower outer fence = Q1 - 3 IQR**

**upper outer fence = Q3 + 3 IQR**

**To represent outliers in a boxplot, we proceed as follows:**

* **Draw a box around Q1, , Q3**
* **Draw the fences**
* **From Q1, extend the lines to the minimum value that does not lie beyond the lower inner fence. If any values lie to the left of the lower inner fence, they are outliers; designate them with asterisks.**
* **Similarly, from Q3, extend the lines to the maximum value that does not lie beyond the upper inner fence. If any values lie to the right of the upper inner fence, they are outliers; designate**

**The dataframe *precip* gives annual rainfalls in 70 American cities.**

|  |  |
| --- | --- |
|  |  |

**Note that the highest value, 67 inches, is much higher than the second highest value of 59.8 inches. By representing the larger value as an outlier, we can clearly see how far it is from the rest of the data.**

**Measures of Variation**

**Often we are interested in how spread apart the data is. Quantities that measure this property are called measures of variation**. **Just like measures of centre, there are several ways to measure variation.**

* **Range**

**The range of a dataset is simply the difference between its largest and smallest values.**

* **Interquartile Range**

**As we saw in the last section, the interquartile range is the range of the middle 50% of the data.**

**What advantage does the IQR have over the range?**

**What disadvantage does the IQR have over the range?**

**We want a measure of variation that takes all the data into consideration. This will give us an idea of the overall spread of the data. There are several such measures. The most common one is the *standard deviation*.**

* Standard Deviation

**The standard deviation of a data set is a measure of variation of scores about the mean. It can be thought of as a measure of the “average” distance that the data is from the mean of that data set. We use the standard deviation because it is easy to work with mathematically, has the same units as the data, and takes every score into account.**

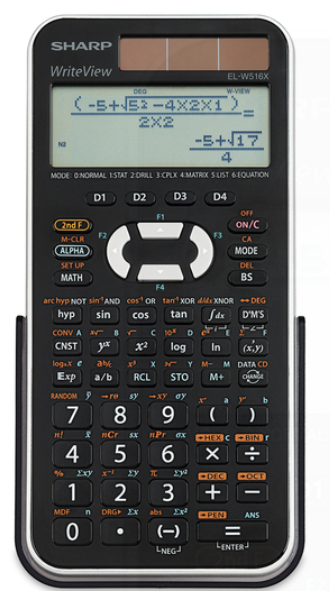
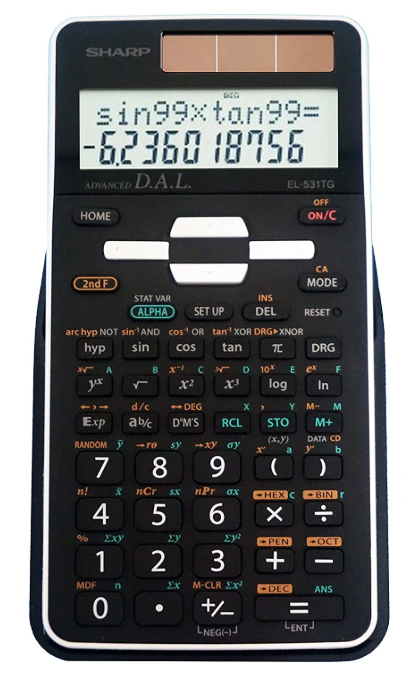
**Sample standard deviation  (statistic)**

**Population standard deviation  (parameter)**

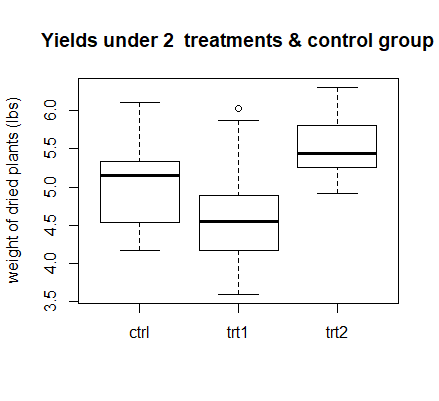
**Why do we take the squares of the differences ?**

**Why do we take the square root of the result?**

**You can compute the standard deviation on your calculator.**

****

**Example (Plant Growth): The dataframe *PlantGrowth* gives results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions.We can compare the treatments with boxplots:**

****

**The standard deviation represents the spread of the data. Which of the three datasets should have the largest standard deviation?**

**Which of the datasets should have the smallest standard deviation?**

**We can compute the three standard deviations using the *sd* command in R.**

> sd(data=PlantGrowth, weight~group)

ctrl trt1 trt2

0.5830914 0.7936757 0.4425733

**The units are \_\_\_\_\_\_\_\_\_\_\_\_.**

**Which of the three groups gives the most consistent results?**

**A standard deviation does not necessarily mean much by itself. For example, is a standard deviation of 0.1 cm small or large? It depends on the dataset. For that reason, it is often more useful to compute the standard deviation as a percentage of the mean.**

**Coefficient of Variation**

**The coefficient of variation, is defined as the ratio**

** **

**expressed either as a fraction or in percentage form. You can view it as a measure of relative dispersion or relative variability.**

**The CV is dimensionless, indicating how large the variability of the data is *in relation to the value of the mean*.**

**Example:**

**If we collected a sample of 20 car tires, and found that their mean diameter was 50 cm. Their standard deviation of .1 cm would indicate a very uniform sample of tires.**

**On the other hand, if we collected a sample of 15 screws and found their mean diameter was 1 cm. A standard deviation of .1 cm would indicate a quite a non-uniform set of screws.**

**The uniform set of tires would have **

**The diverse set of screws would have **

**This indicates that the variation of the screws is more than 20 times are large as the variation of the tires.**

**The CV also allows us to compare variability of data sets with different measurement units (such as ppm vs. %).**

**We can compare the coefficients of variation for the three plant groups:**

> sd(treatment1$weight)/mean(treatment1$weight)

[1] 0.1702801

> sd(treatment2$weight)/mean(treatment2$weight)

[1] 0.08008927

> sd(control$weight)/mean(control$weight)

[1] 0.1158767

**We can use the mean, median, and standard deviation to determine how *skewed* a set of data is.**



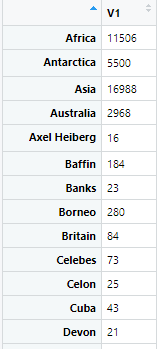
**If a dataset is symmetric, its mean and median are equal. If the dataset is skewed to the left, its median is larger than its mean. If a dataset is skewed to the right, its mean is larger than its median. The size of the difference gives a measure of the extent of the skew. We can also estimate skewness from a boxplot: if the median line is to the left of centre in the box, the distribution is skewed to the right and vice versa.**

**A common formula for skewness is the *Pearsonian coefficient of Skewness*:**

****

**If Sk=0, the data is perfectly symmetric. Data considered skewed if Sk<-1 or Sk >+1**

**The dataset *islands*, which lists the 48 largest land masses, is an example of a skewed dataset. There are a very small number of huge landmasses, such as North America, and many much smaller ones, such as Vancouver Island.**

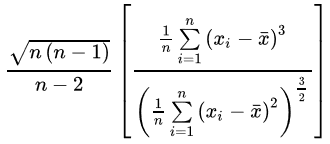


> 3\*(mean(islands)-median(islands))/sd(islands)

[1] 1.078324

**The Pearsonian coefficient of skewness for this dataset is greater than 1; therefore the set of landmasses is skewed to the right.**

**Note: there are other formulas to measure skewness. The *skewness* function in R, which is part of the *moments* package, uses the formula**

****

**sk =**

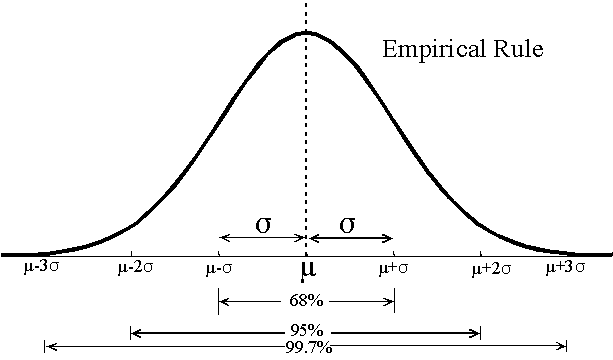
**to compute sample skewness. This formula does a better job than the previous one of capturing how skewed a dataset “looks”, but it’s annoying to compute and most calculators don’t provide a quick way to calculate it. If you’re using R, you should use the *skewness* function, but if you only have a calculator on hand, the simpler function will do.**

**Understanding Standard Deviation**

**Throughout this course, we will frequently encounter datasets whose distributions follow a bell-shaped, or *normal* distribution. When this is the case – and it often is – then the standard deviation gives us some very specific information about where most of our data lies.**

**Empirical Rule (68-95-99 Rule)**

**The empirical rule describes how values in sets of data or populations that are normally distributed are clustered around the mean value. It is exact for situations that are exactly normally distributed, but it can be a reasonable approximation in situations where the normal distribution is followed approximately.**



**These percentages are accurate when the distribution is exactly a normal distribution.**

**68% of the values are within 1 standard deviation.**

**95% of the values are within 2 standard deviations.**

**99.7% of the values are within 3 standard deviations.**

**However, not all data is normally distributed. In such cases, the standard deviation still gives us some information about where to find our data – however, this information is not quite as specific as what the empirical rule tell us about normally distributed data.**

**Chebyshev's Theorem**

**Chebyshev's Theorem is a more general version of the Empirical Rule. This theorem makes no assumptions about the shape of the data distribution.**

**The fraction of a population occurring within k standard deviations of the mean is always at least .**

**This rule is true for any positive value of k > 1 (not just whole numbers).**

**Examples:**

* **when k = 2, , indicating that at least 75% of the data values will be found within two standard deviations of the mean. (When we are confident enough that the data is approximately normally distributed so that we can use the empirical rule, we're able to make the statement that at least 95% of the data falls within this interval -- the lack of information about the data distribution results in this rule hedging by 20 percentage points.) Note that since we haven't assumed that the distribution is symmetric about the mean here, we can't say anything about where the residual 25% of the data may be -- just that up to 25% of the data may be as much as two standard deviations different from the mean.**
* **when k = 3, , allowing us to say that at least 89% of the data or the population will be found within three standard deviations of the mean.**

**Chebyshev's theorem has the advantage that what it says is guaranteed to be true regardless of the shape of the data. Its disadvantage is that what it says is often so imprecise that it is of little practical use. On the other hand, it you assume the empirical rule when it really isn't justified, you may get some very specific results, but they are untrue.**

**Standard Scores (z-scores)**

**Often it is useful to know how many standard deviations a measurement is from the mean. The *z-score* or *standard score* is the number of standard deviations that a given value x is above or below the mean. If x is below the mean, the z-score (represented by the letter z) is negative. If x is above the mean, z is positive.**

**It is quite straightforward to find the z-score.**

**For example, suppose a dataset has mean 42 and standard deviation 6. What is the z-score corresponding to the value 36? (Equivalently: what is the number of standard deviations the value 36 is from the mean?)**

**In general: If a data set has mean µ and standard deviation σ, then the z-score of the value x is:**

**z =**

**Thus, if z = 1, it means that the corresponding value of x is one standard deviation greater than the mean; that is x = μ + σ. If z=-2, the corresponding value of x is two standard deviations less than the mean; that is x = μ - 2σ. In fact, in general, we could rearrange the formulae to give**

** or **

**Examples:**

**Suppose a population has a mean μ = 275, and a standard deviation σ = 22.3. Compute the standard scores corresponding to x = 250, 275, and 280.**

**Solution:**

**To do this, we simply substitute each of these values of x into the formula:**

**x = 250 ⇒**

**x = 275 ⇒**

**x = 280 ⇒**

**Notice that the mean value always gives a standard score of \_\_\_\_\_.**

**z-scores are used to:**

1. **Differentiate between ordinary and unusual values (outliers).**
2. **Compare scores from different populations with different means and different standard deviations.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **µ** | **σ** | **Student’s grade** |
| **Test 1** | **70%** | **5%** | **80%** |
| **Test 2** | **80%** | **10%** | **90%** |

**Eg, a student takes two tests in a class. The class average, standard deviation, and the student’s mark are given in the table above. On which test did the student get a more impressive score?**

1. **Calculate probabilities from normal distributions. (This topic will be addressed at a later date.)**

**Note that we could restate both the empirical rule and Chebyshev's theorem in terms of standard scores. The empirical rule becomes:**

* **approximately 68% of all data will have a standard score between -1 and +1**
* **approximately 95% of all data will have a standard score between -2 and +2**
* **approximately 99.7% of all data will have a standard score between -3 and +3.**

**As a general rule, we say that ordinary values fall within two standard deviations of the mean and unusual values are considered more than 2 standard deviations from the mean.**

**Usual values: -2 ≤ z ≤ 2**

**Unusual values: z < -2 or z >2**

**Example:**

**The heights of all adult males have a mean μ = 1.75m, a standard deviation of σ = 0.07m, and a distribution that is bell shaped. Basketball player LeBron James is 2.03m. Is he unusually tall compared to the general population of adult men?**