STATS 32: Summary Statistics

Kenneth Tay

This document describes statistics that are widely used to summarize data.

Introduction

Imagine that you were recording measurements for something of interest and you ended up with these 100 measurements:

```
31  26  55   6  47  48  81  37  55  17  62  88  28  40  76  67  20
36  36  69  54  71  54  75  42  17  77  88  55  28  49  93  35  95
70  89  18  63  99  13  33  87  78  83  60  49  78  88  21 3001
33  20  24  27  59  25  12  23  60  21  46  65  96  68  45  36  46
45  25  69  41  33  57  97  66  62  86  77  83   9  46  60  92  98
4  58  73  25  30  73  91  21  36  45  91  39  52  13   3  77
```

If someone asked you to describe the measurements that you took, you're not going to say, "the first was 31, and the next was 26, …" First, it would take you a really long time to get through all 100 measurements. Second, your description does not convey any insight into the data that you have collected. What you should do instead is to present a concise description of your data using summary statistics.

Measures of central tendency

The first piece of information you might want to convey is the "center" of your data, i.e. what does the "average" measurement look like? "Average" is not a well-defined term: there are at least 3 different measures of "average":

- **Mean:** This is probably what most people mean (no pun intended) when they say average. We add up all the measurements, then divide the sum by the number of measurements. Because of this formula, the mean is mathematically and computationally simple, and is probably the most widely used measure.
- **Median:** We line up our measurements from smallest to largest, then take the measurement right in the middle (or the mean of the 2 measurements in the middle if we have an even number of measurements). While not as computationally tractable and mathematically "nice", the median is robust to outliers.
- **Mode:** This is the most commonly occuring value in the dataset.

These 3 measures of central tendency can be very different. In our dataset above, the mean is 81.71, the median is 53 while the mode is 36.

Measures of spread

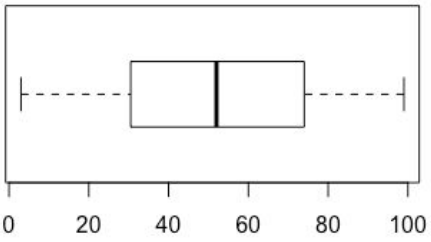
The second piece of information you might want to convey is how "spread out" your measurements are. Again, there is more than one way to describe this:

- **Variance:** This is essentially the sum of squares of the distance of each measurement from the mean, divided by a constant (roughly the number of measurements). The larger the variance, the more spread apart the data is.
- **Standard deviation (SD):** This is the square root of variance. It is more commonly reported than variance because it is on the same scale as the measurements themselves. For example, if the measurements are in inches, then variance is in square inches while the SD is in inches as well.
- **Interquartile range (IQR):** This the 3rd quartile minus the 1st quartile. Some explanation is in order:
    - The **x% percentile** is the value such that x% of the measurements lie below that value.
    - The 1st quartile is the 25% percentile, while the 3rd quartile is the 75% percentile.

The first 2 measures can be heavily influenced by outliers, while the third is robust to them. For our data, the variance is ~87,600, the standard deviation is ~296, and the interquartile range is 44.5.

Understanding boxplots

The boxplot is a simple plot which gives a lot of information on the distribution of your measurements. Here is a boxplot of the data above, excluding the outlier 3001:



The line in the middle of the box is the median, while the ends of the box represent the 1st and 3rd quartiles. The ends of the whiskers that extend out to the lowest measurement still within 1.5 IQR of the 1st quartile, and to the highest measurement still within 1.5 IQR of the 3rd quartile.

Any measurement more than 1.5 IQR from the 1st or 3rd quartiles is drawn as a circle. If we include the outlier 3001, it is drawn as a circle, but because it is so far out, the rest of the boxplot is squished to one side, becoming pretty useless as a visualization of the data.