Estimates of Location & Variability

Terminology

1. **Robust** - insensitive to outliers
2. **Outlier** - an extreme value that is different from most other data
3. **Deviations** (AKA residuals) - difference between the observed value and estimate of location

Estimates of Location

- **Mean**: sum of all values divided by the number of values
  - $\bar{x} = \frac{\sum x_i}{n}$
- **Weighted mean**: sum of all values times a weight divided by a sum of weights
  - $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$
  - Why use weighted mean?
    1. Some values are more valuable than others. If one sensor is known to be less accurate, we can down-weight it
    2. Data does not equally represent all the groups that we interested in. If a group is underrepresented, we can add higher weight to it
- **Trimmed mean**: the average of values dropping some extremes
  - $\bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$
    - $x_{(i)}$ - sorted x values
  - Trimmed mean is a middle ground between median and mean. It's fairly robust, but still uses reflect more data than median
  - Usual drop is bottom and top 10% of the data
- **Median**: the value such that half of the data lies below it
  - Median is *robust* to outliers
  - However, unlike mean, median uses less data to calculate it
- **Weighted Median**: the value such that one-half of the sum of the weights lies below the sorted data
  - Used for similar reasons as weighted mean, i.e. to balance the groups discrepancy and downplay certain values
- **Percentile**: the value such that P% of the data lies below it

Estimates of Variability

- **Variance** (AKA mean-squared-error): sum of *squared* deviations from the mean divided by $n-1$
  - $Var = \frac{\sum (x_i - \bar{x})^2}{n-1}$
  - Need squares so that positive and negative differences don't cancel each other out
  - Preferred over MAE since it's easier to work with squares
  - $n-1$ is used to account for **degrees of freedom**. Degrees of freedom are the number of parameters that can be moved freely. For example, in the triangle, if I set two angles to some values, the third one has to be 180-two angles. so, triangle has 2 degrees of freedom. In this case for variance, since mean depends on the data, we are not free to change it, hence $n-1$ degrees of freedom
    - $n-1$ provided **unbiased** estimation of variance ($E(\hat{\sigma^2}) = \sigma^2$)

- **STD**: square root of Variance
- **Mean Absolute Error** (AKA l1-norm, Manhattan Norm): mean of the *absolute* values of the deviations from the mean
  - $MAE = \frac{\sum |x_i - \bar{x}|}{n}$
- **Median absolution deviation from the median** (MAD) - the median of the absolute values deviations from the median
  - $MAD = Median(|x_1 - m|, |x_2 - m|, \dots)$
  - MAD is robust, unlike Variance, STD, or MAE
- **Range** - difference between largest and smallest value in the set
- **IQR** - difference between 75th and 25th percentiles

Notes

- Statisticians use **estimates**, when data scientists tend to call them **metrics**. Statistics account for uncertainty (we don't know true mean of the population, hence the *estimate of the mean*), when data science deals with concrete objectives