Elementary Statistics

Dom Hutchinson

October 3, 2020

Statistics is the analysis of data from *past* events. While *Probability* is the study of predicting the likelihood of *future* events.

Remark 0.1 - *Frequentist & Bayesian Approach to Statistics*
There are two main approachs to probability: *Frequentist*; and, *Bayesian*.

- The *Frequentist* approach defines the probability of an event to be limit of the *relative frequency* of that event, over many trials.
- The *Bayesian* approach is based on *Bayes Theorem* and treats probability as a *degree of belief* in an event. This belief is made up of prior beliefs and from oberserved data.

1 Definitions

Definition 1.1 - *Statistic*
A *Statistic* is a quanity computed from a dataset (or sample). Typically *Statistics* are used to quantify features of the sample.

Definition 1.2 - *Order Statistic*
An *Order Statistic* is a dataset which is ordered in order of increasing value.
$x_{(i)}$ denotes the value with the i^{th} smallest value (the datapoint with *rank* i).

Definition 1.3 - *Sample Mean,* $\bar{x}$
The *Sample Mean* of a dataset is the arithmetic average of the dataset and represents the mid point of a dataset, weighted by the value of datapoints. The *Mean* is considered the *Expected Value* when sampling from a dataset.

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

Definition 1.4 - *Trimmed Sample Mean*
The *Trimmed Sample Mean* is the mean from a dataset, with a defined proportion of the most extreme data ignored. $\Delta\%$ denotes the *Trimmed Sample Mean* with $n\Delta$ smallest and largest values removed.

$$\begin{aligned} \text{TSM}(\{x_1, \dots, x_n\}, \Delta) &= \text{Mean}(\{x_{(k+1)}, \dots, x_{(n-k-1)}\}) \text{ where } k = \left\lfloor \tfrac{n\Delta}{100} \right\rfloor \\ &= \tfrac{1}{n-2k} \sum_{i=k+1}^{n-k-1} x_{(i)} \end{aligned}$$

Definition 1.5 - *Median, H_2*
The *Median* is the true midpoint of a dataset.

$$\text{median}(\{x_1, \dots, x_n\}) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases}$$

Definition 1.6 - *Mode*
The *Mode* of a dataset is the most common value.

Definition 1.7 - *Sample Variance, s*
The *Sample Variance* of a dataset is a measure of spread of data around the mean. A lower *Sample Variance* indicates data is more concentrated around the mean.

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

Standard Deviation is the square root of *Sample Variance*.

Definition 1.8 - *Hinges, H_k*
Hinges describe the spread of data, while trying to ignore extreme values.

- *Lower Hinge*, H_1 is the *median* of the set of values with rank less than, or equal to, the rank of the *sample median*.
- *Upper Hinge*, H_3, is the *median* of the set of values with rank greater than, or equal to, the rank of the *sample median*.

N.B. - H_2 is the median.

Definition 1.9 - *Quartiles, Q_k & Percentiles, P_k*
Quartiles & Percentiles describe the distribution of values in a data set.

- *Quartiles* partition the dataset into four equally sized groups

$$\tfrac{nk}{4} \text{ values are less than } Q_k \text{ for } k \in [1, 3].$$

- *Percentiles* partition the dataset into one-hundred equally sized groups

$$\tfrac{nk}{100} \text{ values are less than } P_k \text{ for } k \in [1, 99].$$

The *Inter-Quartile Range* of a data set is the difference between Q_1 and Q_3.

$$\text{IQR} := Q_3 - Q_1$$

Definition 1.10 - *Outliers*
A data point x is considered an *Outlier* if it is more than $\frac{3}{2}$IQR from its nearest hinge.

$$\max\left(|x - H_3|, |x - H_1|\right) > \frac{3}{2}\text{IQR}$$

Definition 1.11 - *Skew*
Skew is a measure of the *asymetry* of a dataset. A dataset is:

- *Left Skewed* if $|H_3 - H_2| > |H_2 - H_1|$.
- *Right Skewed* if $|H_2 - H_1| > |H_3 - H_2|$.

2 Theorems

Theorem 2.1 - *Central Limit Theorem*
Let $X_1, \dots, X_m$ be iid random variables with $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \sigma^2$ and $\bar{X}_n$ be the sample mean of the first n. Then for large n

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$