# Binomial Distribution

Muhammad Samir Assawalhy

March 12, 2023

## 1 Binomial distribution

Binomial coeficients is a very useful way to count. It is an essential part of combinatorics (the math of counting). It is basically a way to answer the following question: How many way I can choose k item from n set of items?

$$^nC_r = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{^nP_r}{k!}$$

- Permutations cares about the order of element, if we swaps two elements we get another different permutation.
- Combinations doesn't care out order, so we need to know how many ways to get a subset of size $k$ of a set of items of size $n$.
- We can think about combinations as permutations but also dividing it by $k!$ which is the number of ways to order these $k$ items.
- So at the end we get the above formula simply proven.

**Binomial distribution:**

Binomial distribution can be use with any events with two outputs, i.e heads or tails, buy or not but, fraud or not fraud.

Let think about the coin flip example. This is an event with two outputs, head or tail. If we need to calculate the probability to get $k$ heads after fliping a coin $n$ times with probability of head equal $p$.

This is the formula to calculate this probability:

$$\frac{n!}{(n-k)!k!} \cdot p^k(1-p)^{(n-k)}$$

- Notice that the sequence of heads and tails doesn't matter
- What matter is how many heads we got $k$
- Hence the probability of any sequence of $n$ flips which contains $k$ heads is $p^k(1-p)^{(n-k)}$.

- To count how many events with $k$ flips we simply use combinations (binomial coeficients).

# 2 Statistics Summary

## 2.1 Descriptive Statistics

- We have different types of quantities, quantitative and categorical data.
- Quantitative:
  - Math operations can be used
  - Can be either discrete or continuous
- Categorical
  - Non-numerical values such as types of order
- When describing **quantitative** data we generally discuss 4 main aspects:
  1. Center
     1. Mean: $\sum_{i=1}^{n} x_i/n$
     2. Median: middle point
     3. Mode: most frequent value
        1. If all with the same frequency there is no median
        2. If multiple values have the same frequency all are the mode
  2. Spread
     1. 5 number summary gives values for calculative range and interquartile range.
     2. First quartile value $Q_1$ is the median of the first half
     3. Third quartile value $Q_3$ is the median of the second half
     4. Second quartile value $Q_2$ is the median of all values
     5. When the dataset has odd number of values the middle number ($Q_3$) is not considered in the first nor the second half
     6. Range $=$ max $-$ min
     7. Interquartile range IQR $= Q_3 - Q_1$
  3. Shape
     1. Left Skewed
     2. Right Skewed
     3. Symmetric
     4. Normal Distribution and Bell Curve
  4. Outliers
     1. Are obvious visually
     2. Point far away from the mean

## 2.2 Simpson's Paradox

- Simpson's paradox is a phenomenon in probability and statistics

- Properties appears in several groups of data but disappears or reverses when the groups are combined.

- Results of statistics can be hugely impacted by how we choose to look at our data.

## 2.3 Probability

- Statistics and probability are different but strongly related fields in mathematics.
- There is also almost and opposite relation in these two.
- Probability of opposite event $P(\neg A)$ equals $1 - P(A)$.
- Probability of composite events $P(A, B) = P(A) \cdot P(B)$.