Introduction to Statistical Ideas and Methods

Sampling Distributions

The Mean and Variance of a Proportion

In this document we investigate the behaviour of a random variable that is a proportion.

Lets start with a familiar example of flipping a coin and measuring the proportion of times that it comes up heads. If we flip the coin twice, we could get 0, 1, or 2 heads, corresponding to proportions of 0, 0.5 or 1. Anytime we flip a coin twice we do not know what proportion of H (heads) we will get since it is random. But we know the probability distribution that describes how likely we are to get one of these possible proportions:

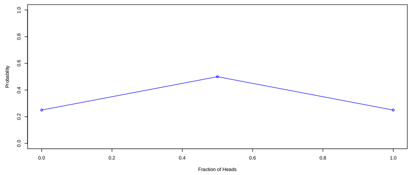

Figure 1: Probabilities when Flipping 2 Coins

The proportion of half is most likely since we are flipping a coin that is equally likely to come up H or T.
We also saw the probability distributions for the proportions of H if we flip our coin 3 or 10 or 100 times:

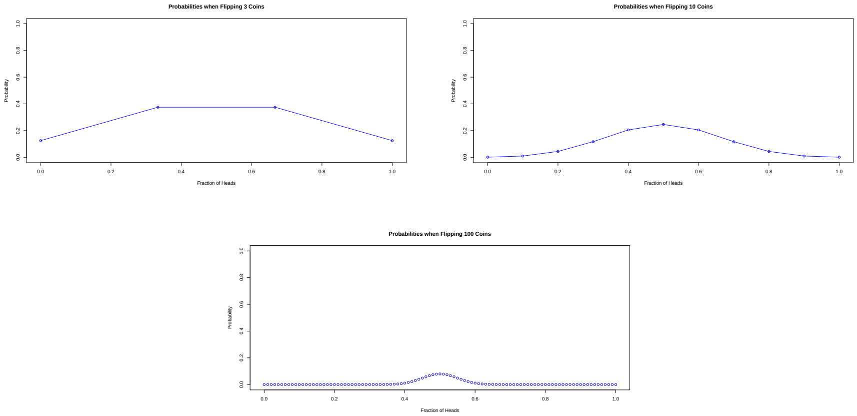

Figure 2: Proportion probabilities when Flipping 3, 10 and 100 Coins

Note that for each of these, the most likely value is 0.5. Also as the number of flips of the coin increases, the most likely proportions are increasingly concentrated around 0.5. Therefore 0.5 is not just the most likely value, for large number of flips we are unlikely to get a proportion that is far from 0.5.

We also saw that as the number of flips of the coin increases, the probability distribution for the proportion of H approaches a Normal distribution. This was due to the **Central Limit Theorem** (CLT).

Central Limit Theorem: If an experiment is repeated over and over, then the probabilities for the average results, or the proportion of successes, will converge to a Normal distribution.

Now the main question is: *What are the mean and standard deviation of this Normal distribution?*

Lets introduce some more general notation, that can apply to measuring the proportion of H for flipping a coin, but also for studying other proportions that might not be 0.5.

Suppose we carry out n trials.
Each trial comes out success or failure.
$p =$ probability of success on a trial.

In our example, a success is flipping a coin and getting a H with $p = 0.5$.

$\hat{p} =$ proportion of successes observed in the n trials.
In statistics, the hat notation is often used for an estimate, here we can think about the proportion of successes we observe as an estimate for the probability of getting a success on a trial.

p is a property of the experiment we are carrying out, and so it is fixed (not random). On the other hand, $\hat{p}$ varies randomly. When carrying out n trials, the proportion of successes we get will not always be exactly the same. Therefore $\hat{p}$ is a random variable.

We already know that $\hat{p}$ has a Normal distribution (as long as n is large enough). To fully specify its distribution we need to find mean and standard deviation.

For each of our n trials, we record a random variable X:

$$X = \begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases}$$

We know that X is a Bernoulli random variable with the probability of a success p:

$$X \sim \text{Bernoulli}(p)$$

The expected value of X can be calculated with the formula for the expectation of a discrete random variable:

$$E(X) = \sum_x xP(x) = 0 \times (1 - p) + 1 \times p = p$$

Similarly the variance of X can be calculated with the formula for the variance of a discrete random variable:

$$Var(X) = \sum_x (x - E(X))^2 P(x) = (0-p)^2 \times (1-p) + (1-p)^2 \times p = p(1-p)$$

We can think of our n trials as n independent Bernoulli random variables $X_1, , X_n$. We are estimating the proportion of successes, which is the number of successes divided by the number of trials. Each X is 0 or 1, so we get:

$$\text{Number of successes} = X_1 + \cdots + X_n$$

and

$$\hat{p} = \text{Proportion of successes} = \frac{\text{Number of successes}}{n} = \frac{X_1 + \cdots + X_n}{n}$$

The key thing to note is that $\hat{p}$ is an average of Bernoulli random variables. That is why the Central Limit Theorem applies to it. Also now we can easily find the mean and standard deviation of the resulting normal distribution.

We already know, that in general,

$$E(\bar{X}) = E(X), \quad Var(\bar{X}) = \frac{Var(X)}{n}$$

Therefore we get:

$$E(\hat{p}) = p, \quad Var(\hat{p}) = \frac{p(1-p)}{n}, \quad SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

So we have that $\hat{p}$, the proportion of times we get a success, has expectation equal to p. It is a good property if we are using the proportion to estimate the probability. And an estimator that has this property ($E(\hat{p}) = p$) is called **unbiased**.

Also note that, standard deviation of sample proportion decreases with n. So the larger our sample size n, that is the more times we flip the coin, the less variability in the proportion of times the coin comes up H.

Summary:
For large n, $\hat{p}$ (proportion of successes in n trials) where each trial has probability p of a success, has approximately normal distribution with mean p and variance $p(1-p)/n$:

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

For this to be true, we need n large enough for the CLT to apply.