Sampling distributions

BIG IDEA: Given a population of cases and a variable on these cases, we consider the random variable "take a sample of a fixed size n and calculate its summary statistic."

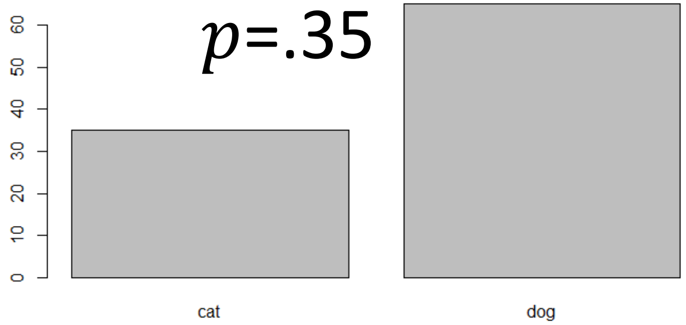
The Sample Proportion $\hat{p}$ (p-hat):
Starting with a population and a categorical variable, with a particular category singled out as "success", with probability p of success and q of failure, take a sample of n cases, and record the fraction of successes in the sample.

The Sample Mean $\bar{y}$ (y-bar):
Starting with a population and a quantitative variable, take a sample of n cases, and record the mean of the sample.

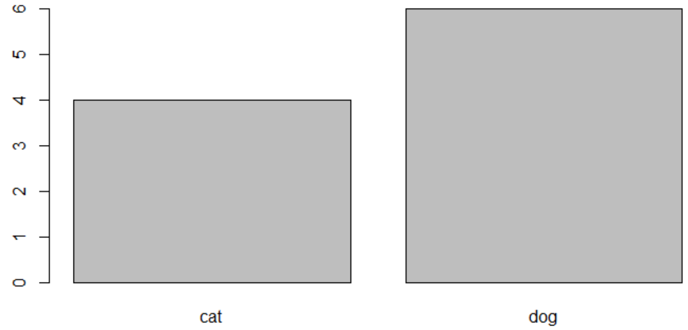
The Sample Proportion $\hat{p}$ (p-hat):

- $\hat{p}$ is the sample proportion, p is the population proportion.
- $\hat{p}$ is variable and p is constant.
- Each sample only gives a single value of $\hat{p}$. To get the distribution of $\hat{p}$, we have to take many samples.
- **Warning:** Even though the original variable is categorical, the proportion of successes in each sample, $\hat{p}$, is quantitative.
- **Amazing:** The sampling distribution of $\hat{p}$ can be approximated by a normal model, with mean p and standard deviation $\sqrt{\frac{pq}{n}}$.

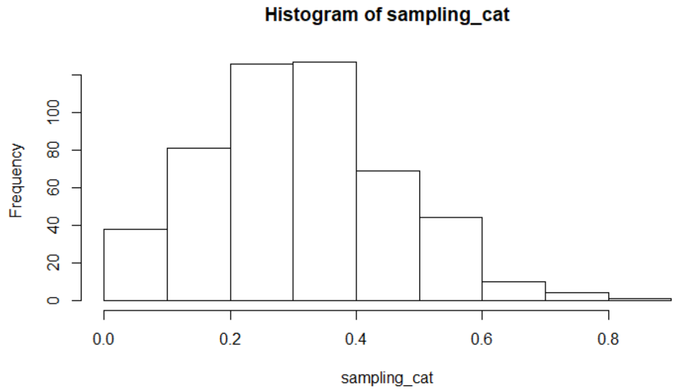
p=.35



One value of $\hat{p}$ is .4



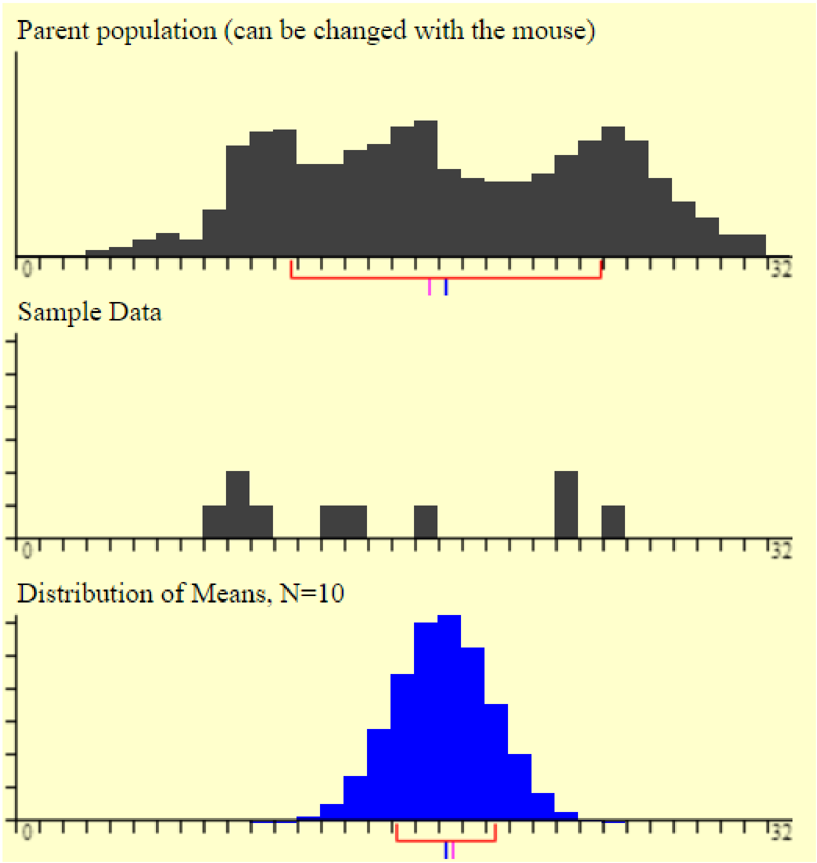
All values of $\hat{p}$



The Sample Mean $\bar{y}$ (y-bar):

- $\bar{y}$ is the sample mean, μ is the population mean.
- $\bar{y}$ is variable and mu is constant.
- Each sample only gives a single value of $\bar{y}$. To get the distribution of $\bar{y}$, we have to take many samples.
- **Warning**: There are three different sets of data, with three different histograms!

  1. the population, 2. the sample, 3. means of multiple samples ($\bar{y}$ is #3.)



- **Central Limit Theorem.** The sampling distribution of $\bar{y}$ can be approximated by a normal model, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. As n gets larger, this approximation gets better.