SAMPLING DISTRIBUTION OF MEANS & PROPORTIONS

PPSS

The situation in a statistical problem is that there is a population of interest, and a quantity or aspect of that population that is of interest. This quantity is called a parameter. The value of this parameter is unknown.

To learn about this parameter we take a sample from the population and compute an estimate of the parameter called a statistic.

Populations & Samples

- **Population**
  - All Saudis
  - All inpatients in KKUH
  - All depressed people
- **Sample**
  - A subset of Saudis
  - A subset of inpatients
  - The depressed people in Riyadh.

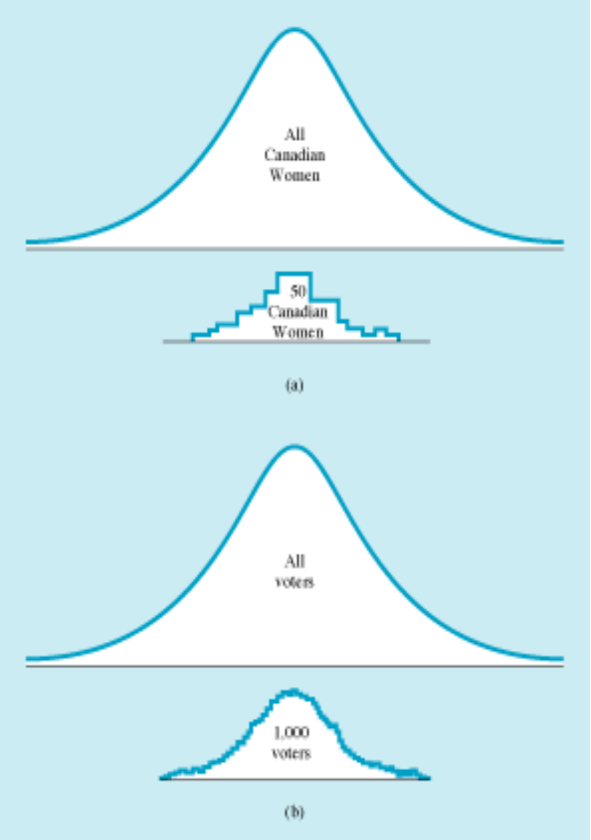
Samples and Populations-

- Sample

Relatively small number of instances that are studied in order to make inferences about a larger group from which they were drawn

- Population

The larger group from which a sample is drawn



Samples And Population

- It is usually not practical to study an entire population
  - in a random sample each member of the population has an equal chance of being chosen
  - a representative sample might have the same proportion of men and women as does the population.

Statistics and Parameters-

- **a parameter is a characteristic of a population**
  - e.g., the *average* heart rate of all Saudis.
- **a statistic is a characteristic of a sample**
  - e.g., the *average* heart rate of a sample of Saudis.
- **We use statistics of samples to estimate parameters of populations.**

$$Statistic \rightarrow estimates \rightarrow Parameter$$

$\overline{X} \rightarrow estimates \rightarrow \mu$	μ "mew"
$s \rightarrow estimates \rightarrow \sigma$	σ "sigma"
$s^2 \rightarrow estimates \rightarrow \sigma^2$	
$r \rightarrow estimates \rightarrow \rho$	ρ "rho"

- Inference – extension of results obtained from an experiment (sample) to the general population.

- use of sample data to draw conclusions about entire population

- *P*arameter – number that describes a *p*opulation
  - Value is not usually known
  - We are unable to examine population
- *S*tatistic – number computed from *s*ample data
  - Estimate unknown parameters
  - Computed to estimate unknown parameters
    - Mean, standard deviation, variability, etc..

SAMPLING DISTRIBUTION:

The sample distribution is the distribution of all possible sample means that could be drawn from the population.

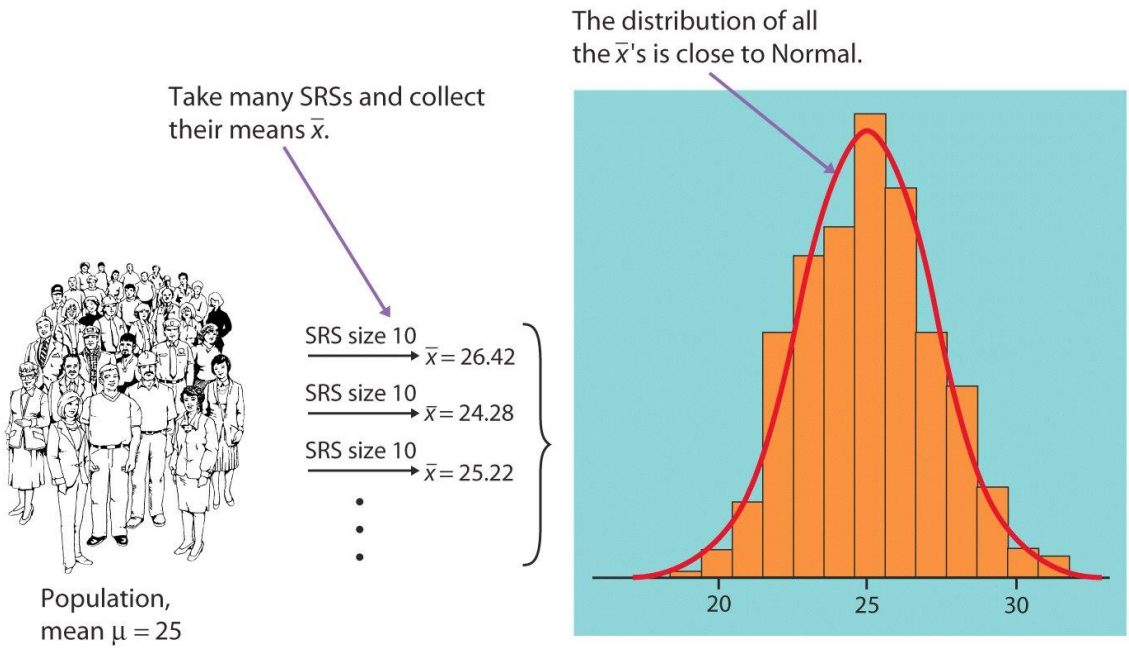
What would happen if we took many samples of 10 subjects from the population?

Steps:

1. Take a large number of samples of size 10 from the population
2. Calculate the sample mean for each sample
3. Make a histogram of the mean values
4. Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers and other deviations

SAMPLING DISTRIBUTION

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.



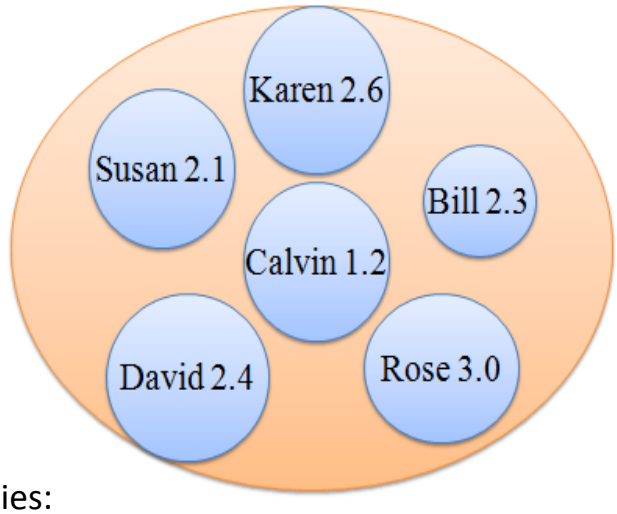
- How can experimental results be trusted? If $\bar{x}$ is rarely exactly right and varies from sample to sample, how it will be a reasonable estimate of the population mean μ?
- How can we describe the behavior of the statistics from different samples?
    - E.g. the mean value

- Very rarely do sample values coincide with the population value (parameter).
- The discrepancy between the sample value and the parameter is known as sampling error, when this discrepancy is the result of random sampling.
- Fortunately, these errors behave systematically and have a characteristic distribution.

A sample of 3 students from a class
of a population of 6 students and measure students GPA

Student	GPA
Susan	**2.1**
Karen	**2.6**
Bill	**2.3**
Calvin	**1.2**
Rose	**3.0**
David	**2.4**

Draw each possible sample from this 'population'



With samples of n = 3 from this population
of N = 6 there are 20 different sample possibilities:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{6\times5\times4\times3\times2\times1}{3\times2\times1(3\times2\times1)} = \frac{720}{36} = 20$$

- Note that every different sample would produce a different mean and S.D.

ONE SAMPLE = Susan + Karen +Bill / 3

= 2.1+2.6+2.3 / 3

$\overline{X}$ = 7.0 / 3 = 2.3

Standard Deviation:

$(2.6-2.3)^2 = 0.3^2 = .09$

$(2.3-2.3)^2 = 0^2 = 0$

s^2 = 0.13 / 3 and $s = \sqrt{.043} = 0.21$

So this one sample of 3 has a mean of 2.3 and a S.D of 0.21

What about the other samples ? ?

A SECOND SAMPLE

= Susan + Karen + Calvin

= 2.1 + 2.6 + 1.2

$\overline{X}$ = 1.97

S.D = .58

20th SAMPLE

= Karen + Rose + David

= 2.6 + 3.0 + 2.4

$\overline{X}$ = 2.67

S.D = .25

- Assume the true mean of the population is known, in this simple case of 6 people and can be calculated as 13.6/6 = μ =2.27
- The mean of the sampling distribution (i.e., the mean of all 20 samples) is 2.30.
- Sample mean is a random variable.
- If the sample was randomly drawn, then any differences between the obtained sample mean and the true population mean is due to sampling error.
- Any difference between $\overline{X}$ and μ is due to the fact that different people show up in different samples

- If $\overline{X}$ is not equal to μ , the difference is due to sampling error.

- "Sampling error" is normal, it is to-be-expected variability of samples

What is a Sampling Distribution?

- A distribution made up of every conceivable sample drawn from a population.

- A sampling distribution is almost always a hypothetical distribution because typically you do not have and cannot calculate every conceivable sample mean.

- The mean of the sampling distribution is an unbiased estimator of the population mean with a computable standard deviation.

LAW OF LARGE NUMBERs

If we keep taking larger and larger samples, the statistic is guaranteed to get closer and closer to the parameter value.

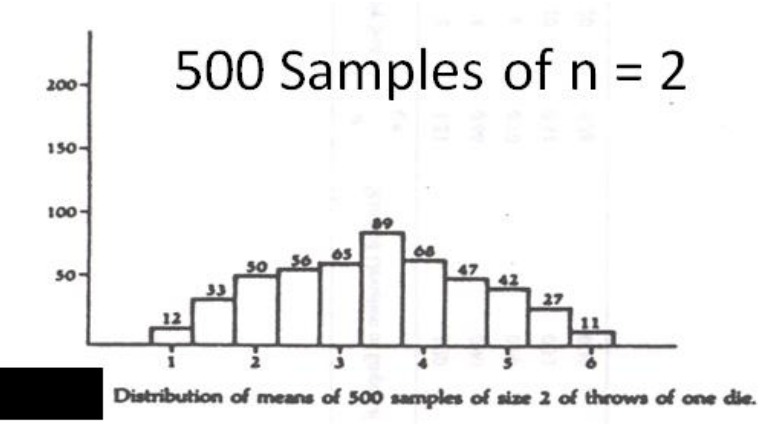
LAW OF LARGE NUMBERS

Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\overline{x}$ of the observed values gets closer and closer to the mean μ of the population.

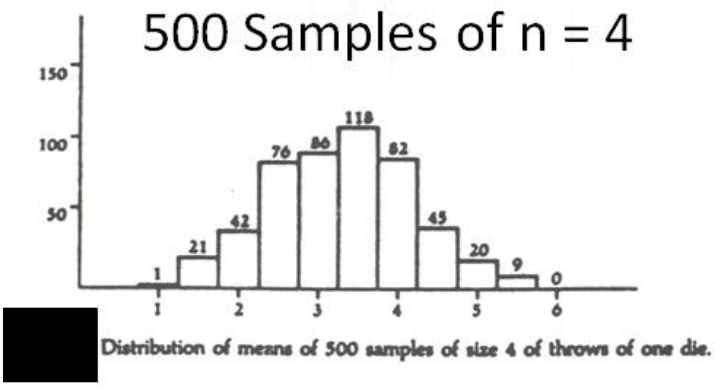
ILLUSTRATION OF SAMPLING DISTRIBUTIONS

Draw 500 different SRSs.

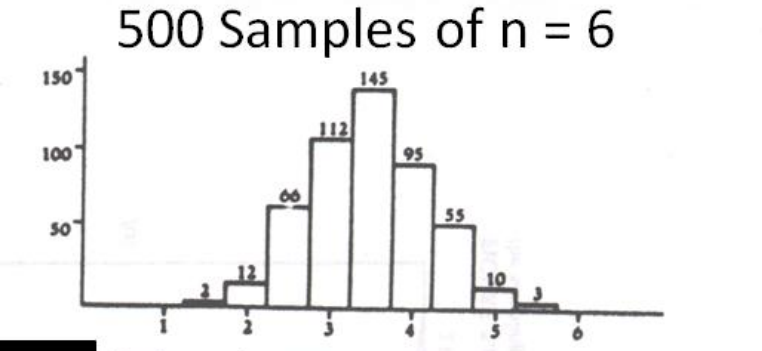
What happens to the shape of the sampling distribution **as the size of the sample increases**?



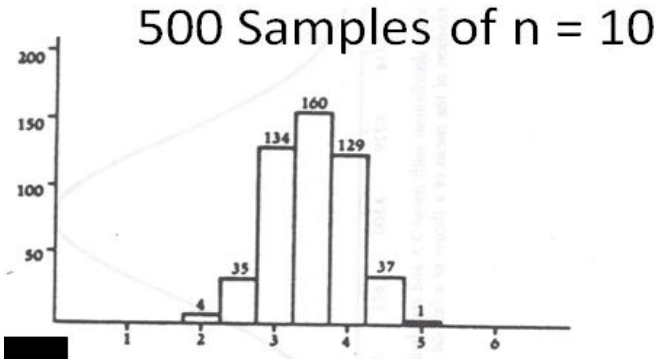
Distribution of means of 500 samples of size 2 of throws of one die.



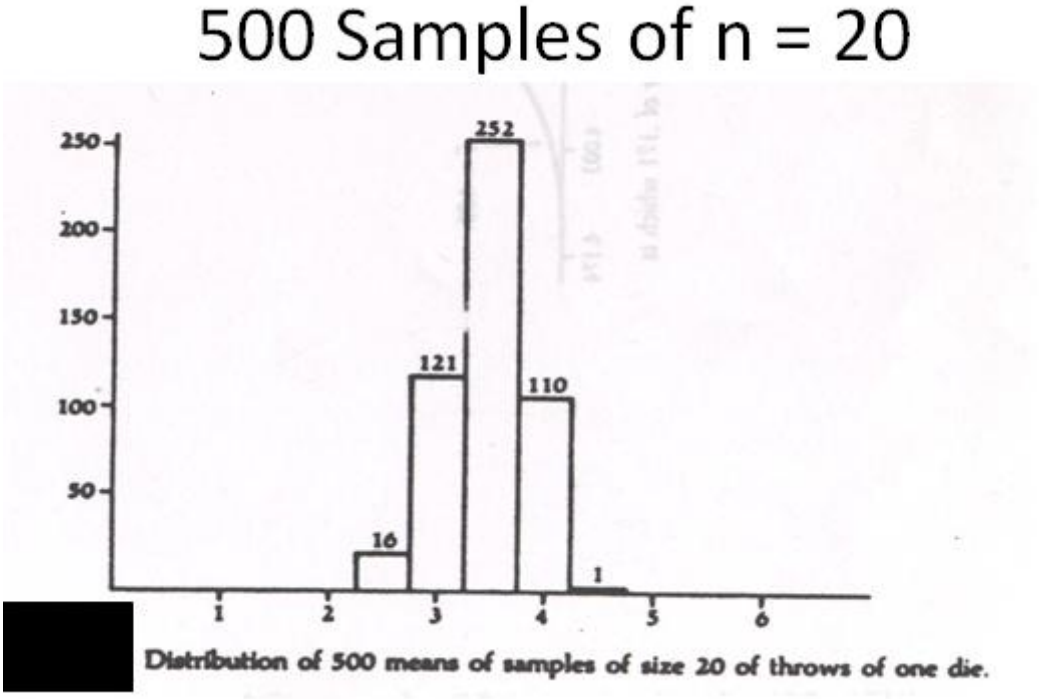
Distribution of means of 500 samples of size 4 of throws of one die.



Distribution of 500 means of 500 samples of size 6 of throws of one die



Distribution of 500 means of samples of size 10 of throws of one die.



Distribution of 500 means of samples of size 20 of throws of one die.

Key Observations

As the sample size increases the mean of the sampling distribution comes to more closely approximate the true population mean, here known to be m = 3.5

AND -this critical- the standard error - that is the standard deviation of the sampling distribution gets systematically narrower.

Three main points about sampling distributions

1- Probabilistically, as the sample size gets bigger the sampling distribution better approximates a normal distribution.

2- The mean of the sampling distribution will more closely estimate the population parameter as the sample size increases.

3- The **standard error (SE)** gets narrower and narrower as the sample size increases. Thus, we will be able to make more precise estimates of the whereabouts of the unknown population mean.

Don't get confuse with the terms of

STANDARD DEVEIATION and STANDARD ERROR

Standard deviation

measures the variation of a variable in the sample.

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

Standard error

of mean is calculated by

$$s_{\bar{x}} = sem = \frac{s}{\sqrt{n}}$$

Standard Deviation versus Standard Error

- The *standard deviation* (s) describes variability between individuals in a sample.
- The *standard error* describes variation of a sample statistic.
  - The standard deviation describes how individuals differ.
  - The standard error of the mean describes the precision with which we can make inference about the true mean.

Standard Error of the mean:

- Standard error of the mean (sem):

$$s_{\bar{x}} = sem = \frac{s}{\sqrt{n}}$$

- Comments:
  - n = sample size
  - even for large s, if n is large, we can get good precision for sem
  - always smaller than standard deviation (s)

ESTIMATING THE POPULATION MEAN:

We are unlikely to ever see a sampling distribution because it is often impossible to draw every conceivable sample from a population and we never know the actual mean of the sampling distribution or the actual standard deviation of the sampling distribution. But, here is the good news:

- We can estimate the where abouts of the population mean from the sample mean and use the sample's standard deviation to calculate the standard error. The formula for computing the standard error changes, depending on the statistic you are using, but essentially you divide the sample's standard deviation by the square root of the sample size.

P-Hat

The situation in this section is that we are interested in the proportion of the population that has a certain characteristic.

This proportion is the population parameter of interest, denoted by symbol p.

We estimate this parameter with the statistic p-hat – the number in the sample with the characteristic divided by the sample size n.

P-Hat Definition

$$\hat{p} = X / n$$

Sample proportions

The proportion of an "event of interest" can be more informative. In statistical sampling the sample proportion of an event of interest is $\hat{p}$ used to estimate the proportion *p* of an event of interest in a population.

For any SRS of size *n,* the sample proportion of an event is:

$$\hat{p} = \frac{\text{count of event in the sample}}{n} = \frac{X}{n}$$

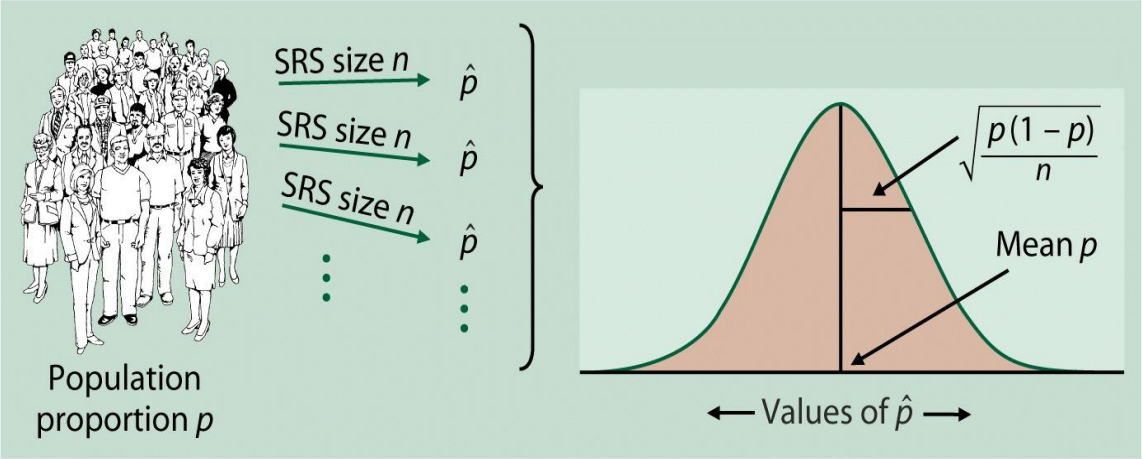
- In an SRS of 50 students in an undergrad class, 10 are O +ve blood group:
  $\hat{p}$ = (10)/(50) = 0.2 *(proportion of O +ve blood group in sample)*
- The 30 subjects in an SRS are asked to taste an unmarked brand of coffee and rate it "would buy" or "would not buy." Eighteen subjects rated the coffee "would buy."
  $\hat{p}$ = (18)/(30) = 0.6 *(proportion of "would buy")*

Sampling Distribution of p-hat

How does p-hat behave?

To study the behavior, imagine taking many random samples of size n, and computing a p-hat for each of the samples.

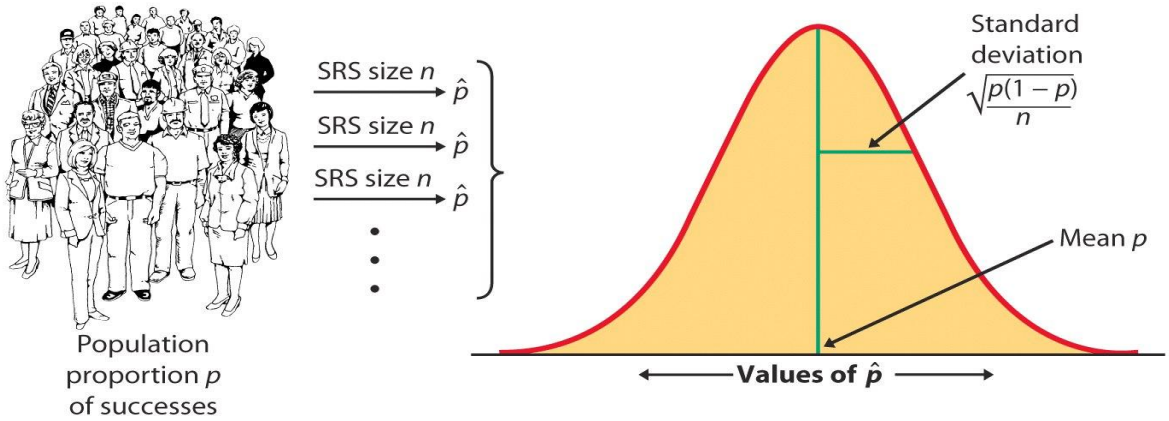
Then we plot this set of p-hats with a histogram.



Sampling distribution of the sample proportion:

The sampling distribution of $\hat{p}$ is never exactly normal. But as the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.

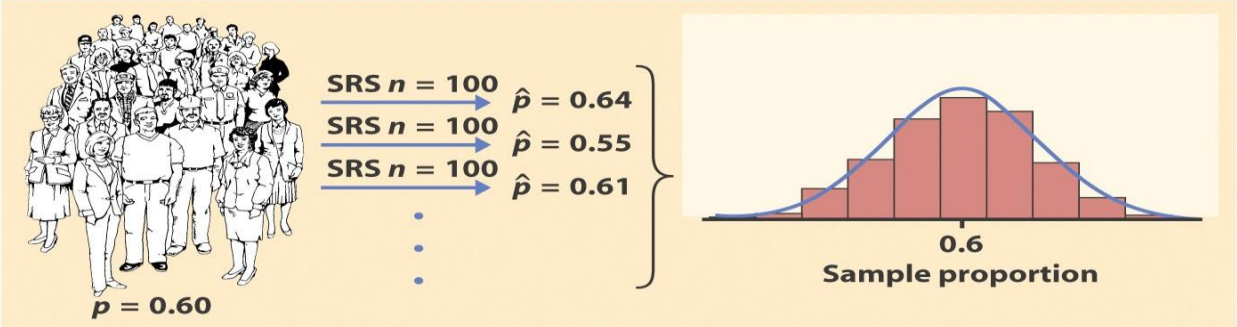
The normal approximation is most accurate for any fixed *n* when *p* is close to 0.5, and least accurate when *p* is near 0 or near 1.



Reminder: Sampling variability

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called sampling variability.

If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern.



Properties of p-hat:

- When sample sizes are fairly large, the shape of the p-hat distribution will be normal.
- The mean of the distribution is the value of the population parameter p.
- The standard deviation of this distribution is the square root of p(1-p)/n.

$$sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Sampling Distribution for Proportion:

Example: Proportion

- Suppose a large department store chain is considering opening a new store in a town of 15,000 people.
- Further, suppose that 11,541 of the people in the town are willing to utilize the store, but this is unknown to the department store chain managers.
- Before making the decision to open the new store, a market survey is conducted.
- 200 people are randomly selected and interviewed. Of the 200 interviewed, 162 say they would utilize the new store

Example: Proportion

What is the population proportion p?

11,541/15,000 = 0.77

What is the sample proportion $\hat{p}$?

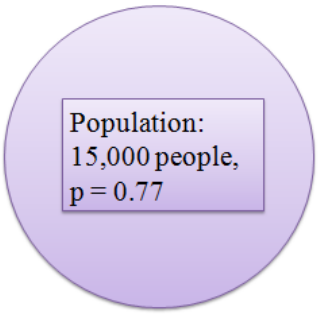
162/200 = 0.81

What is the approximate sampling distribution (of the sample proportion)?

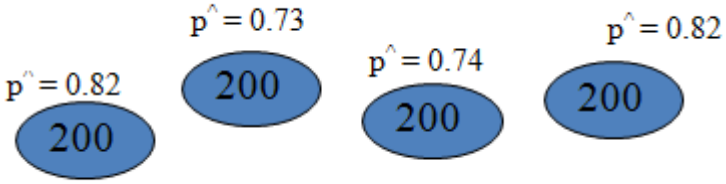
$$\hat{p} \sim Normal\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right) = Normal\left(0.77, 0.0297^2\right)$$

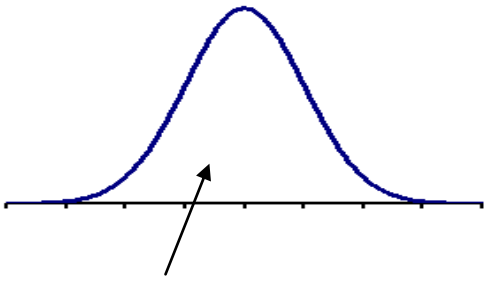
What does this mean??

Suppose we take many, many samples (of size 200):



Then we find the sample proportion for each sample





The sample we took fell here.

Example: Proportion

- The managers didn't know the true proportion so they took a sample.
- As we have seen, the samples vary.
- However, because we know how the sampling distribution behaves, we can get a good idea of how close we are to the true proportion.
- This is why we have looked so much at the normal distribution.
- Mathematically, the normal distribution is the sampling distribution of the sample proportion, and, as we have seen, the sampling distribution of the sample mean as well.

Two Steps in Statistical Inference Process-

1- Calculation of **"confidence intervals"** from the sample mean and sample standard deviation within which we can place the unknown population mean with some degree of probabilistic confidence

2- Compute **"test of statistical significance"** (Risk Statements) which is designed to assess the probabilistic chance that the true but unknown population mean lies within the confidence interval that you just computed from the sample mean.