Sampling Distributions

Outline

- Sampling plans and experimental designs
- Statistics and sampling distributions
- The Central Limit Theorem
- The sampling distribution of the sample mean
- The sampling distribution of the sample proportion

Introduction

- Type of distribution and corresponding parameters (e.g. p(Success) in binomial, μ in Poisson, and μ and σ in normal distribution) are required to determine the probability of a sample outcome.

- For most practical situations, the distribution type can be decided fairly easily, but the values of the parameters are unknown.

- *Samples* are normally required to estimate these parameters – but for the resulting shape of the distribution to truly reflect the population, the samples need to be selected in a certain way.

Sampling plans and experimental designs

- **Sampling plans** or **experimental designs** – the way a sample is selected and affect the quantity of information in the sample.
- **Simple random sampling -** every sample of size n equally likely to be selected. The resulting sample is called a **random sample**.
- **Observational study** – data already existed before you decided to observe/analyse its characteristics, e.g. sample surveys. Potential problems include nonresponse, undercoverage, and wording bias
- **Stratified random sampling** – selecting a simple random sample from each of the given number of subpopulations or **strata**.
- **Cluster sampling** – random sampling of existing clusters in the population, e.g. randomly picking whole households (clusters) in the population.

Sampling plans and experimental designs

- **1-in-k systematic random sample** – random selection of 1 in the first k elements in an ordered population, and then every k^{th} element thereafter.
- Non-random sampling techniques exist but MUST NOT be used to infer the population parameters.

Statistics and Sampling Distributions

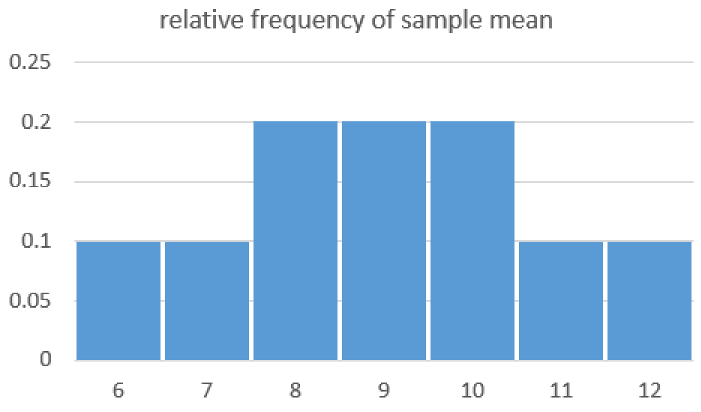
- **Statistics** – numerical descriptive measures calculated from the *sample*, and are *random variables*
- **Sampling distribution of a statistic** – probability distribution of possible values of a statistic of random samples drawn from a population – can be determined
  - *Mathematically* using the law of probability (when the population is *small*)
  - Using *simulation* – repeatedly drawing large number of samples and calculate the corresponding statistic
  - Using statistical theorems
- Sampling distributions can be discrete or continuous.

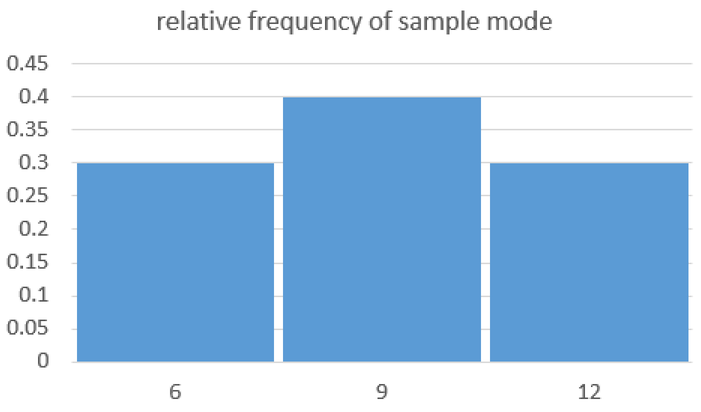
Statistics and Sampling Distributions

Example: Randomly picks 3 values from the list of 5 numbers 3, 6, 9, 12, 15, and determine the resulting sampling distribution for mean and mode.

Which one would we choose as the estimate for population mean?

Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12





The Central Limit Theorem

- Draw random samples of n observations from a non-normal population that has mean μ and standard deviation σ.
- The sampling distribution of sample mean $\bar{x}$ is approximately normally distributed.
- The mean of this distribution is μ, standard deviation is $\sigma/\sqrt{n}$.
- This approximation becomes more accurate as n gets larger.
- Many estimators used in inferencing about *population parameters* are sums or averages of *sample measurements*.
- CLT allows us to use normal distribution to describe the behaviours of these estimators in repeated sampling and evaluate the probability of observing certain sample outcomes.

The Central Limit Theorem

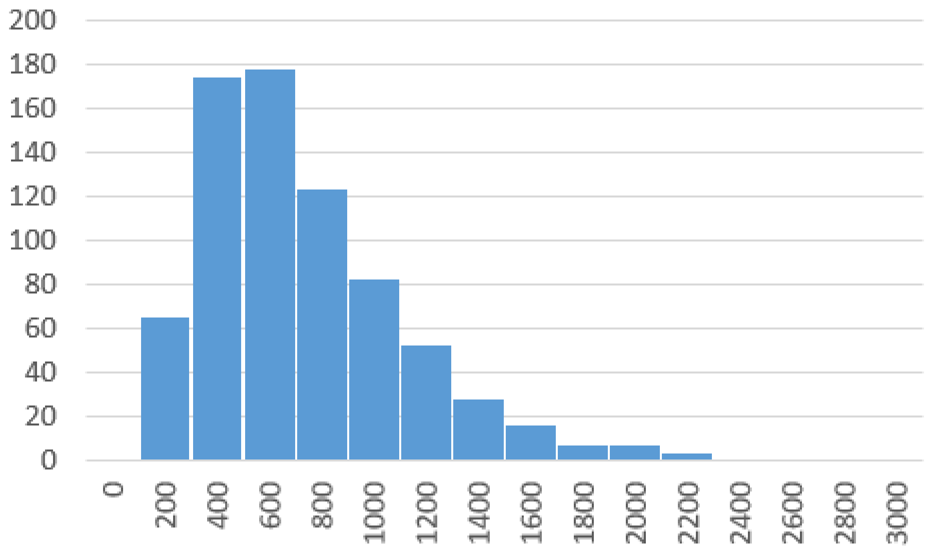
Example: The Calcium data – distribution not normal

1.Randomly pick n=5 values from the population of measurements and calculate the sample average, which is an estimate of the true mean.

2. Repeat the random sampling for 200 times.

3. Calculate the standard deviation of the sample average and plot the histogram of sample average

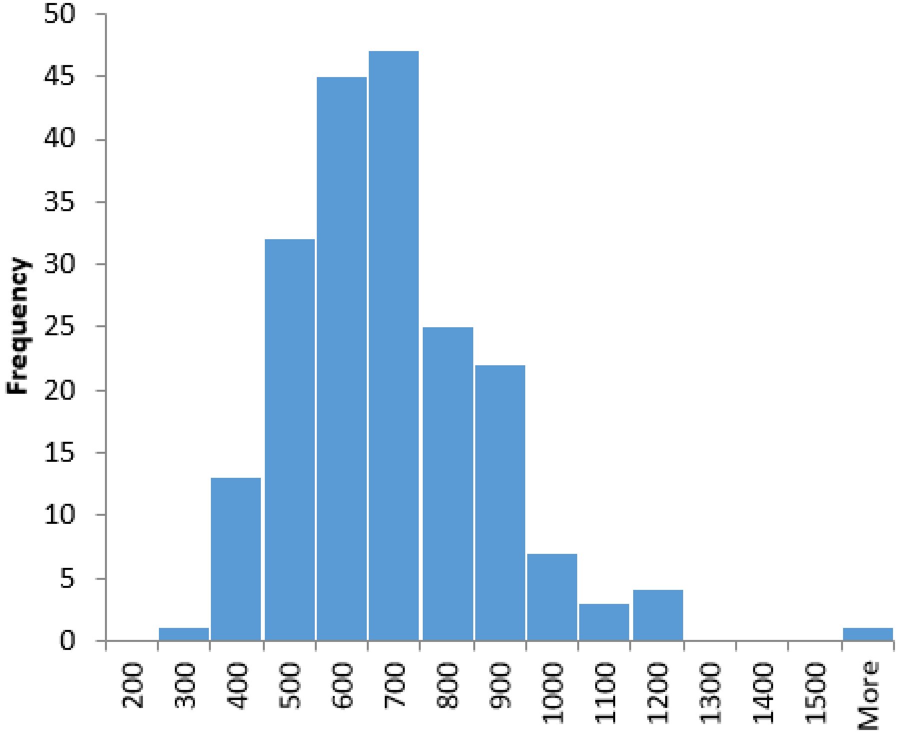
4. Repeat steps 1-3 for n=10, n=30, n=50



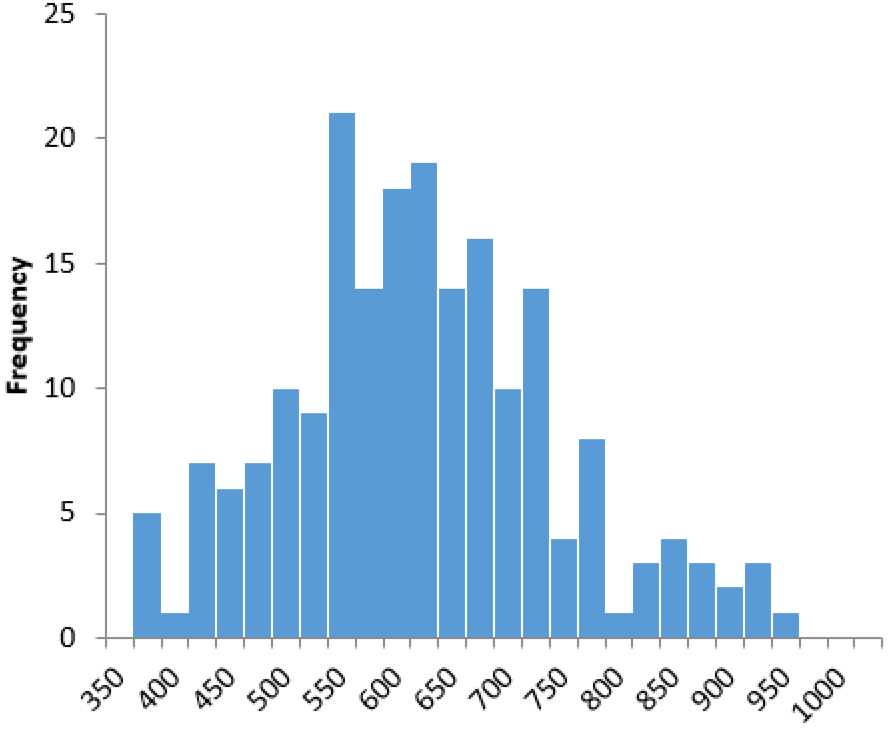
The Central Limit Theorem

$$true\ mean\ \mu = 624.05, \sigma = 397$$

$$n = 5, \bar{x} = 641, s = 191.38, \frac{\sigma}{\sqrt{n}} = 177.5$$



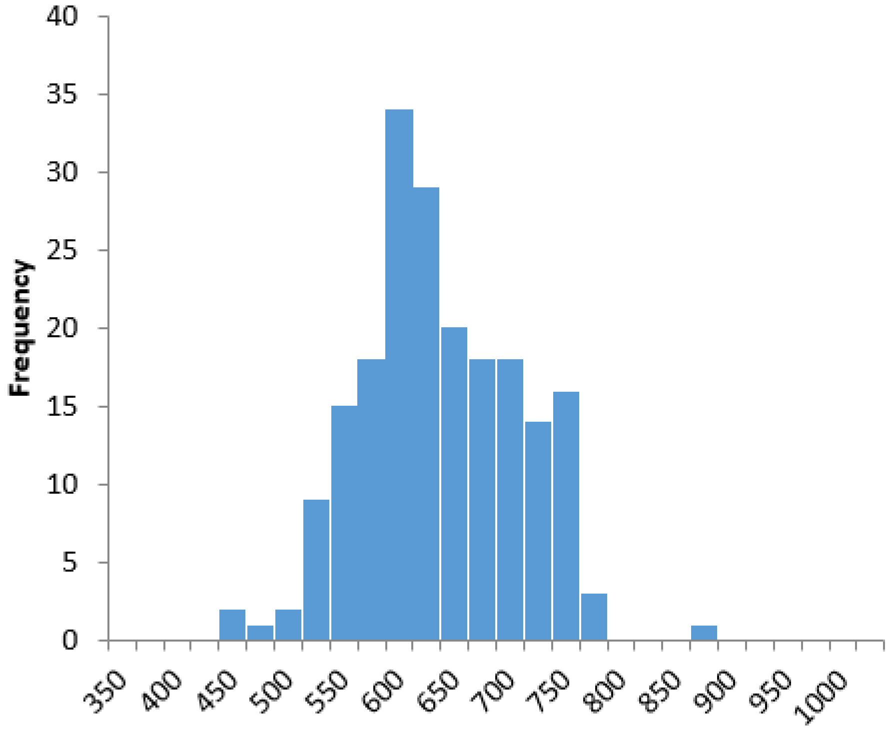
$$n = 10, \bar{x} = 609, s = 122.54, , \frac{\sigma}{\sqrt{n}} = 125.54$$



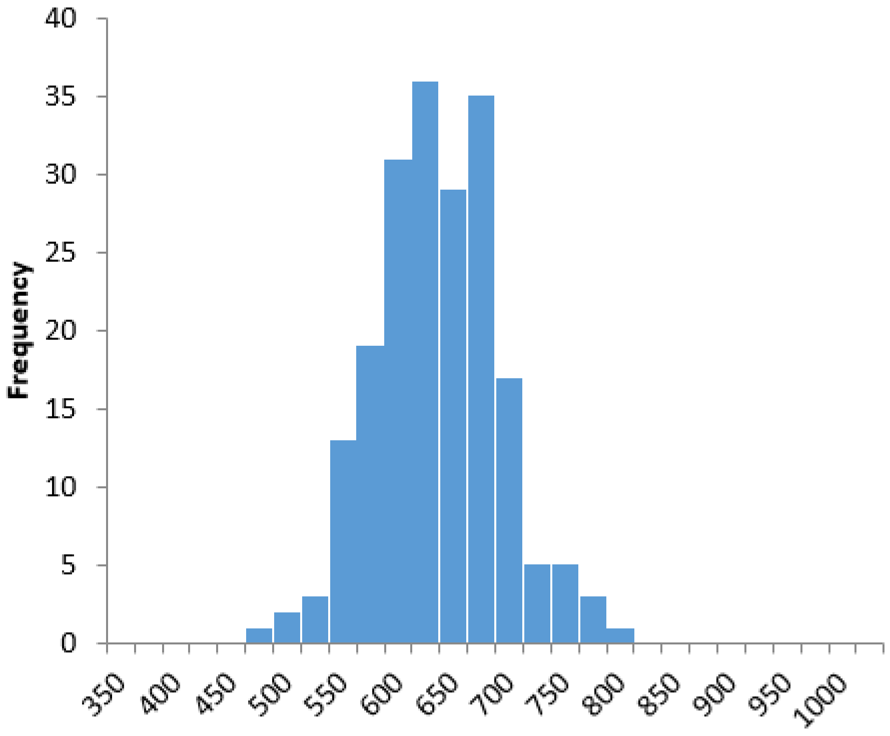
The Central Limit Theorem

$$true\ mean\ \mu = 624.05, \sigma = 397$$

$$n = 30, \bar{x} = 624, s = 71.01, \frac{\sigma}{\sqrt{n}} = 72.5$$



$$n = 50, \bar{x} = 623, s = 56.03, \frac{\sigma}{\sqrt{n}} = 56.14$$



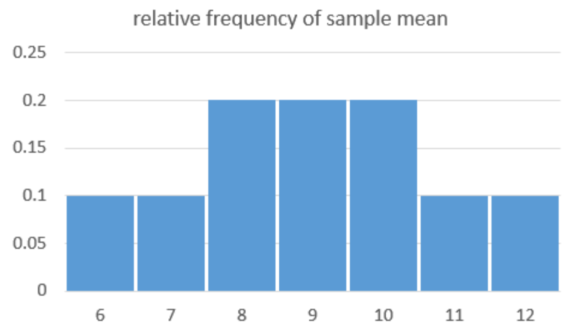
The Central Limit Theorem

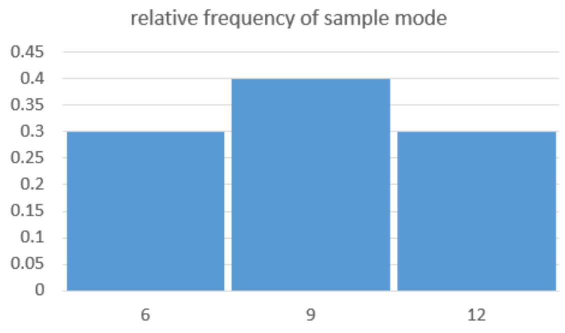
- When n is large enough, thanks to CLT, many sample statistics will follow normal distribution and allows us to make statistic inference, for example evaluating the probability of observing certain sample results

- How large n should be? There is no clear answer but some guidelines
  - If the sampled population is *normally distributed*, the sampling distribution of $\bar{x}$ is also normal, regardless of the sample size
  - If the sampled population is *symmetric*, the sampling distribution of $\bar{x}$ becomes normal with fairly small n.
  - If the sampled population is *skewed*, the sampling distribution of $\bar{x}$ will only become approximately normal at a fairly large n, e.g. $n>30$.
  - If the sampled population is a binomial population with p or (1-p) very small, n needs to be fairly large

The Sampling Distribution of the Sample Mean

- If the population mean μ is unknown, sample statistics such as sample mean $\bar{x}$ and sample median m can be used. But which one?





- Sample mean $\bar{x}$ has properties not shared by other sample statistics, i.e. its sampling distribution follows the normal distribution, regardless of the population distribution (with large sample size)

The Sampling Distribution of the Sample Mean

- If a random sample of n measurements is selected from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\sigma/\sqrt{n}$

- **Standard error** – standard deviation of a statistic used in estimating a population parameter.
- The standard deviation of $\bar{x}$, calculated by $\sigma/\sqrt{n}$, is referred to as standard error of the mean, denoted as $\mathrm{SE}(\bar{x})$ or simply SE.

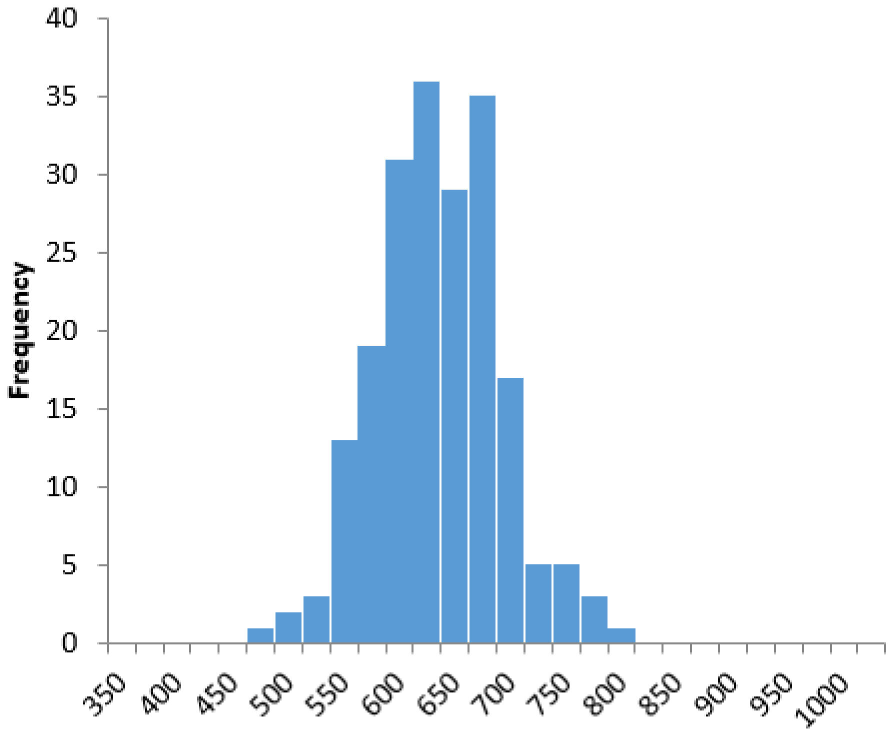
The Sampling Distribution of the Sample Mean

Example. Refer to the Calcium example above, the true mean of Calcium reading is μ = 624.05 mg and standard deviation is σ = 397 mg. If we pick 50 samples, what is the probability having the average Calcium reading $\bar{x}$ less than 700mg?

Solutions.

We can reasonably assume that the sampling distribution with n=50 is normal. The standardized z-score of 400mg is $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{700-624.05}{397/\sqrt{50}} = 1.35$

Probability of $\bar{x}$ less than 700mg equals $\mathrm{P}(z < 1.35) =.9115$



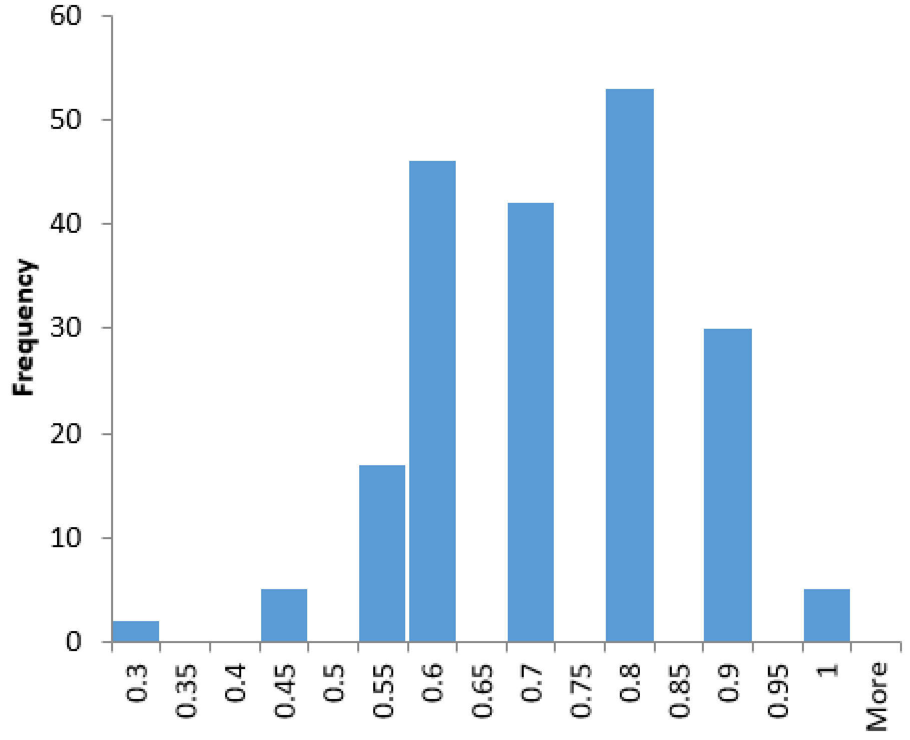
The Sampling Distribution of the Sample Proportion

Assuming we have a binomial population of 1200 observations and 71% of them are Successes (p=0.71). Let's pretend that we do not know the true value of p and we try to estimate p by repeated random sampling from the population.

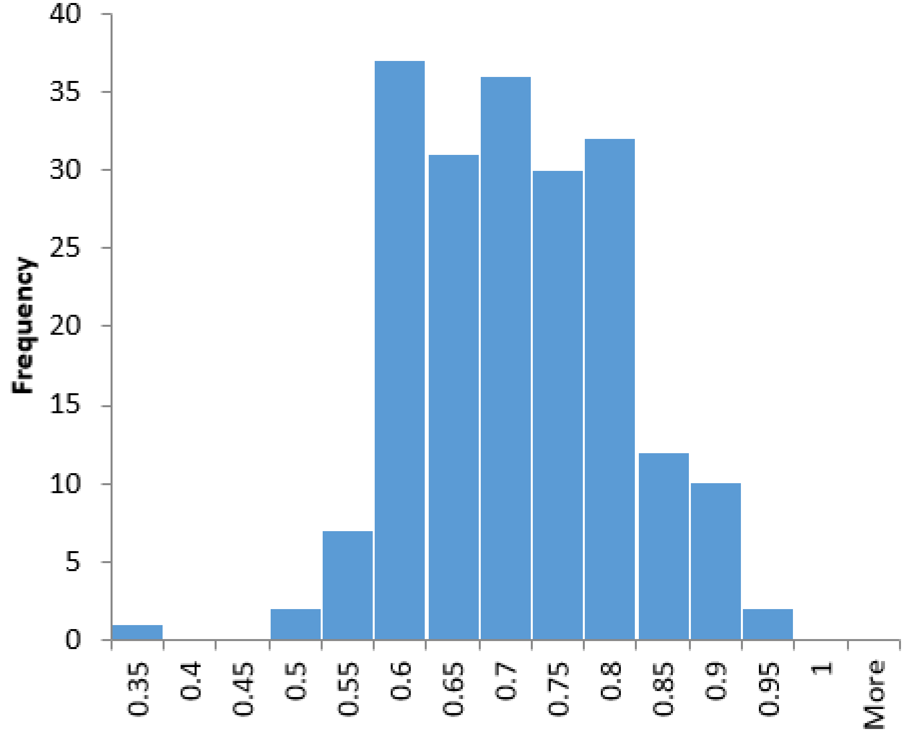
1. Randomly pick n=10 elements from this population and calculate the average. This average is the estimate of p, denoted $\hat{p}$, based on the 10-element sample.
2. Repeat the random sampling process for 200 times.
3. Calculate the standard deviation of the estimated $\hat{p}$ and plot the histogram of $\hat{p}$ from the 200 random samples.
4. Repeat step 1-3 for n=20, n=30, and n=100

The Sampling Distribution of the Sample Proportion

$n = 10, \bar{x} \equiv \hat{p} = .712, s = .143$

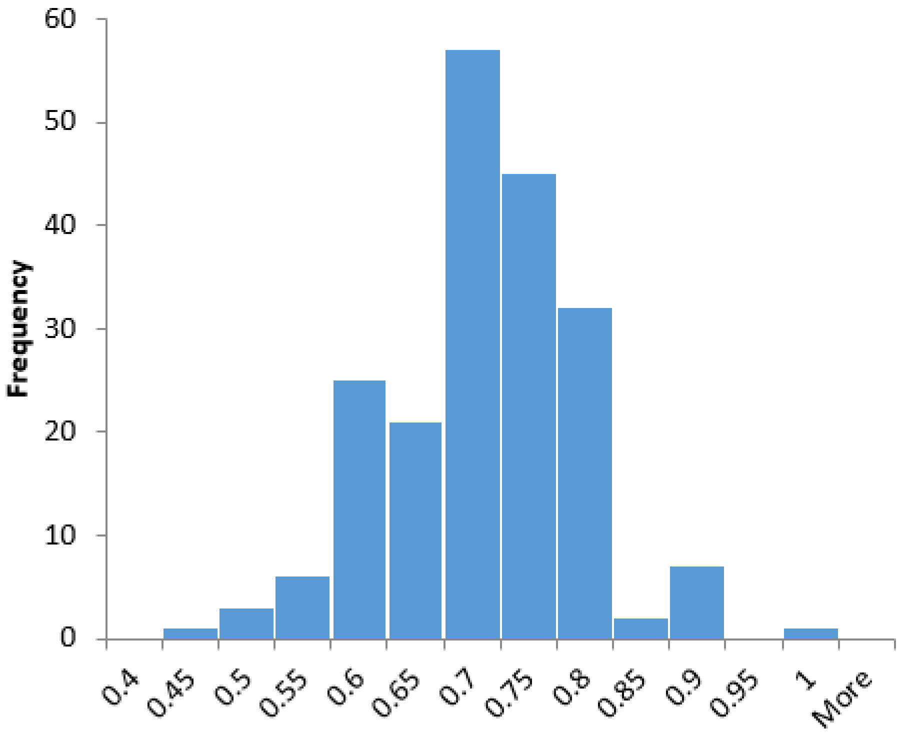


$n = 20, \bar{x} \equiv \hat{p} = .706, s = .106$

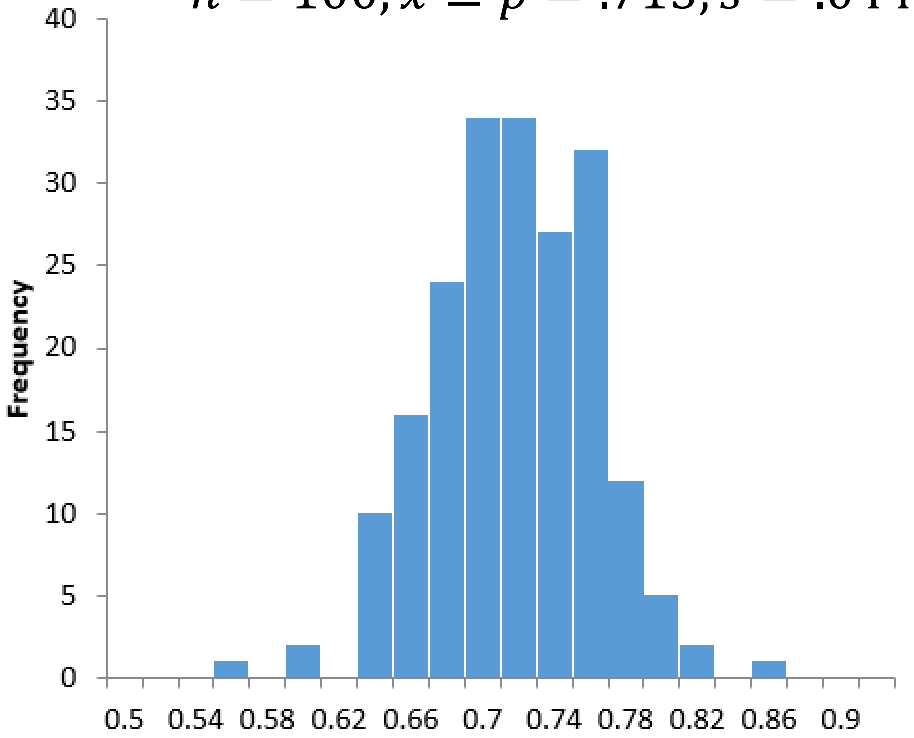


The Sampling Distribution of the Sample Proportion

$n = 30, \bar{x} \equiv \hat{p} = .696, s = .084$



$n = 100, \bar{x} \equiv \hat{p} = .713, s = .044$



The Sampling Distribution of the Sample Proportion

If a random sample of n observations is selected from a binomial population with parameter p

- the sampling distribution of $\hat{p}$ will have the mean p
- the standard deviation of this distribution is $\mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$