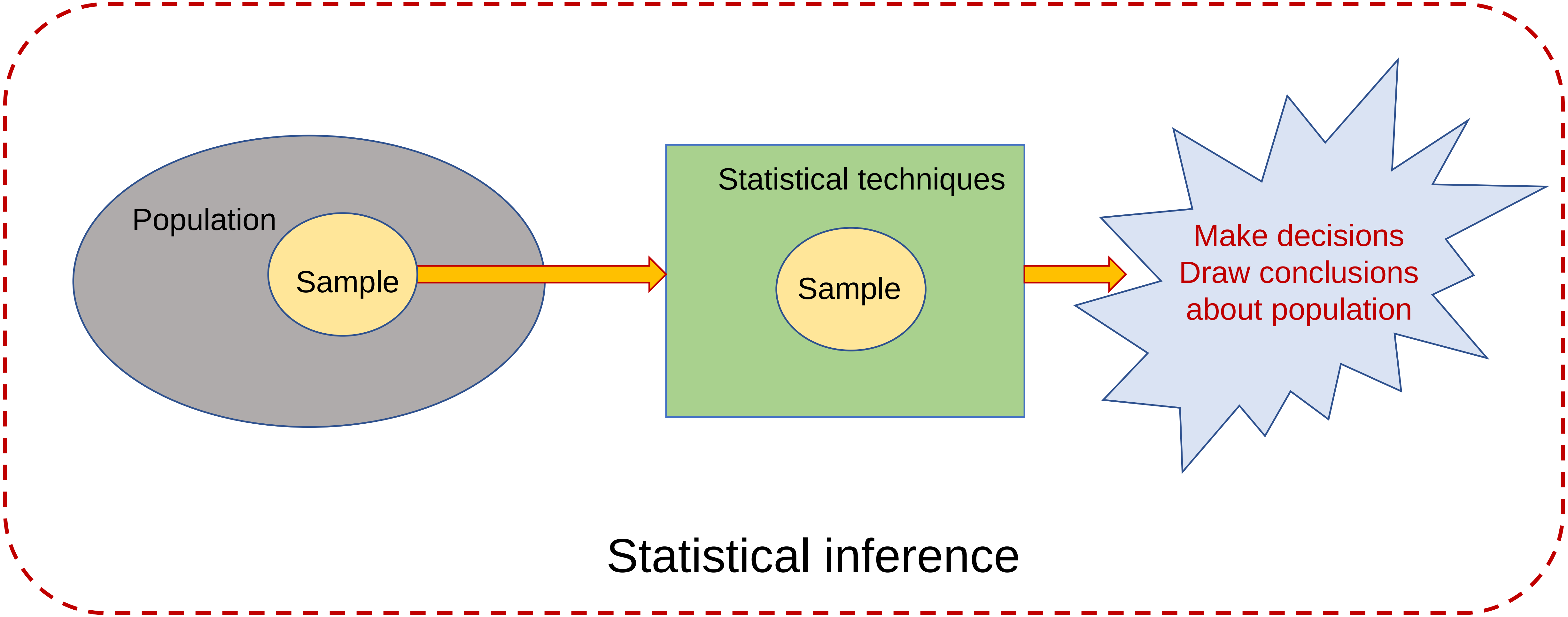
7

Sampling Distributions & Point Estimation of Parameters

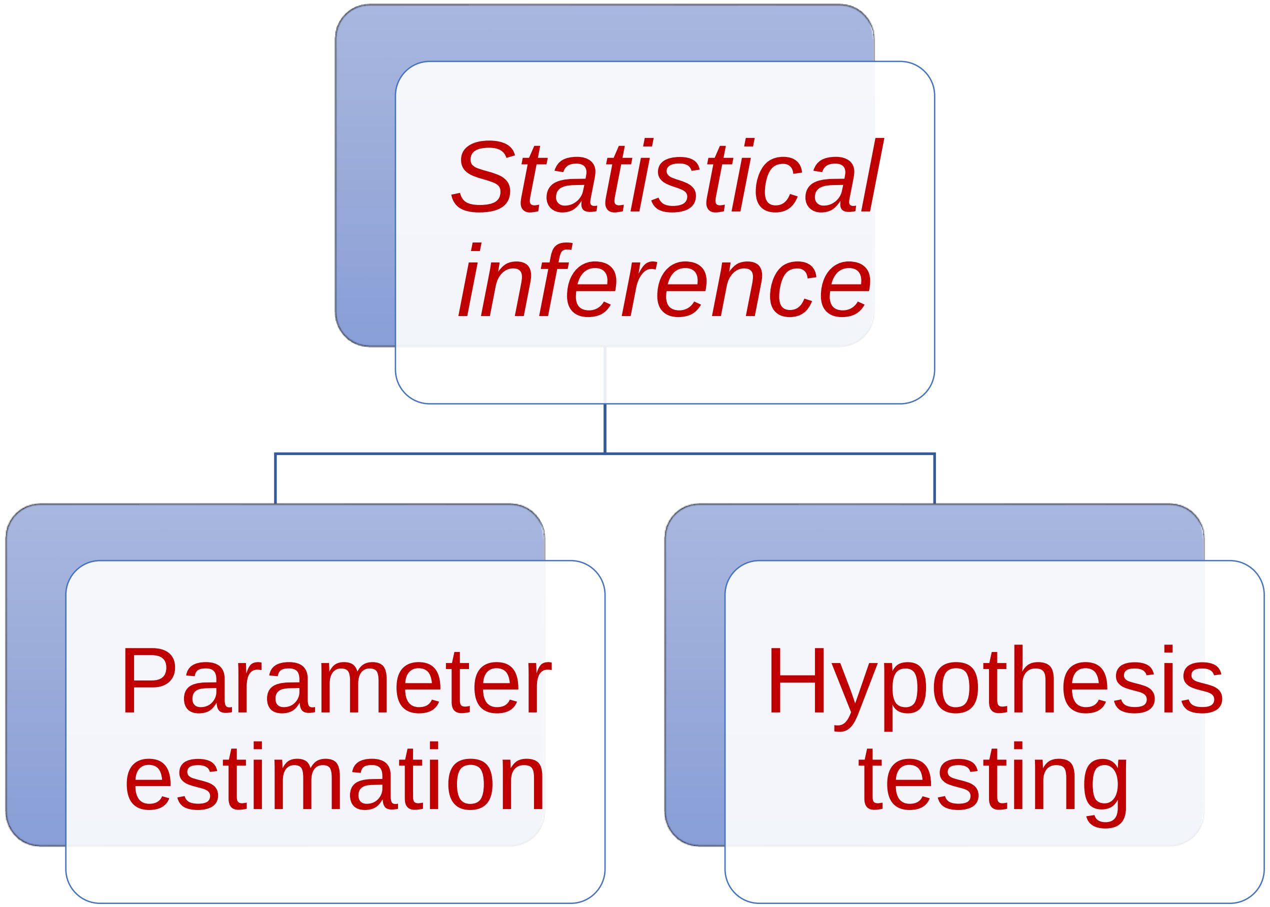
LO

- Discuss sampling distributions and use the *Central Limit Theorem* to analyze sampling distribution of the means
- Discuss sampling distributions; find the point estimation of a parameter.
- Apply Central Limit Theorem to calculate probabilities regarding the sampling distribution of the means.

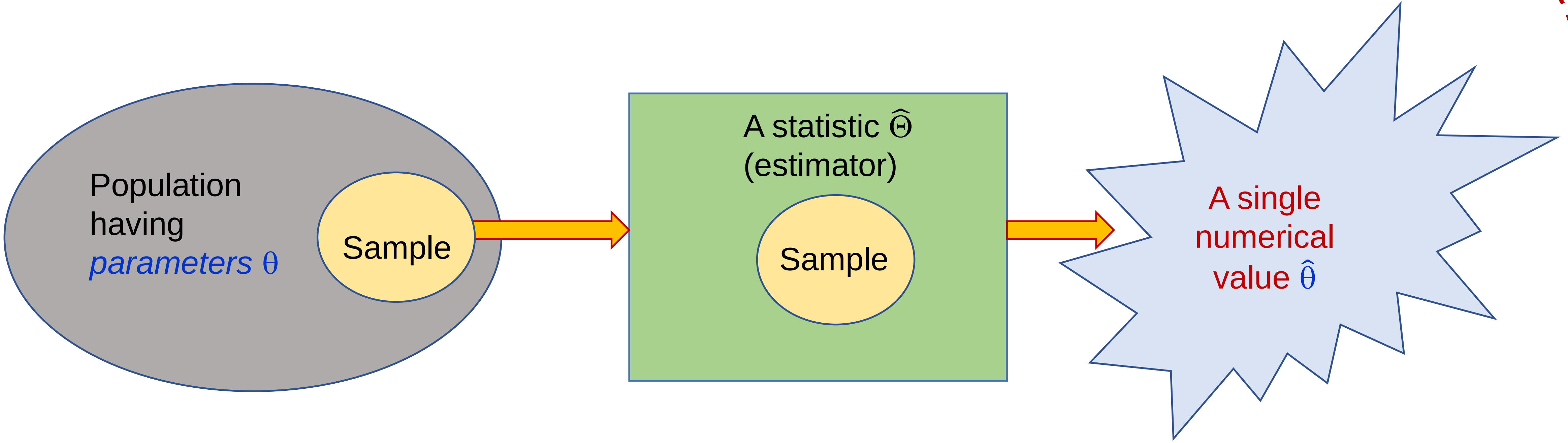
Introduction



Statistical inference

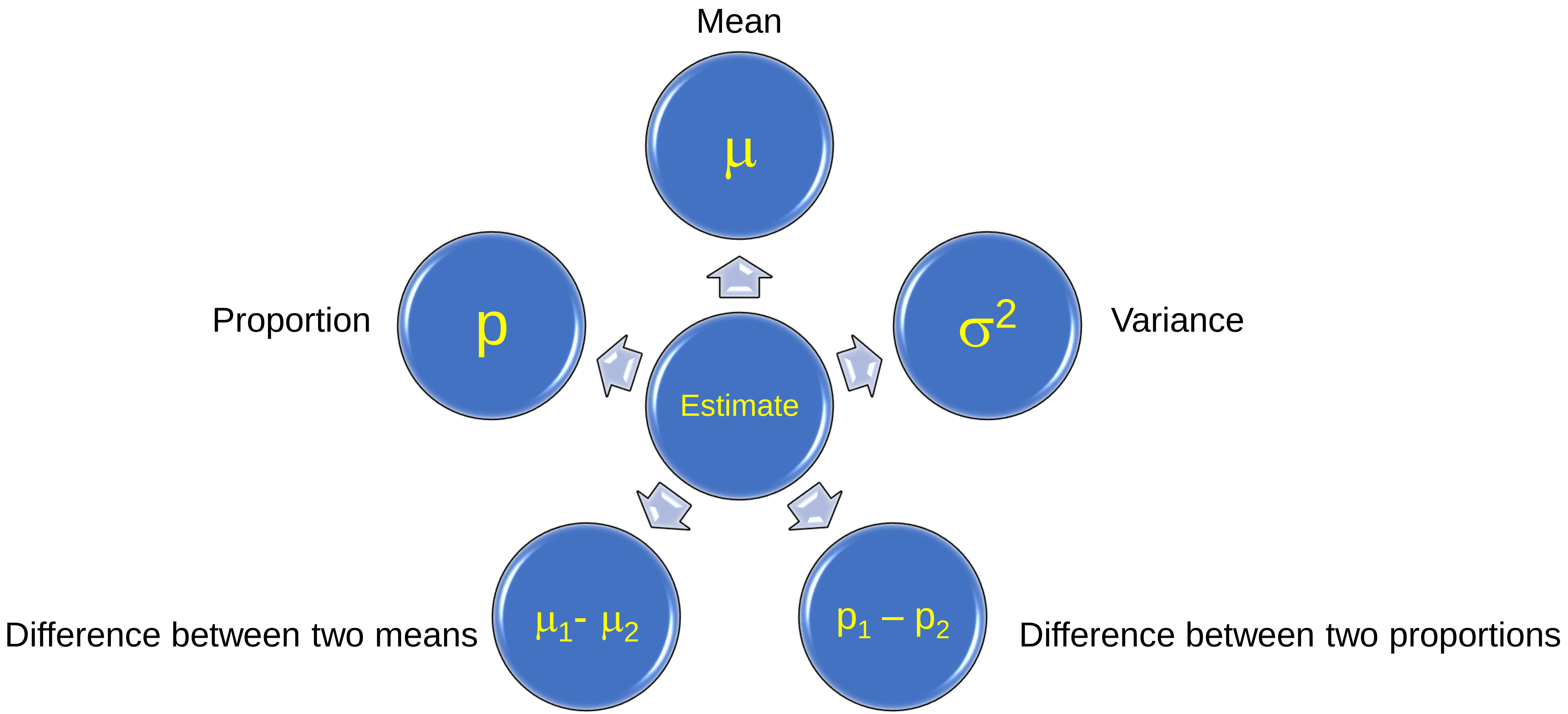


Point estimate



A *statistic* $\widehat{\Theta}$ is a *function* of the observations in a random sample $X_1, X_2, \ldots, X_n$.
Distribution of a statistic $\widehat{\Theta}$ is called *sampling distribution.*
For example, the probability distribution of $\overline{X}$ is the *sampling distribution of the mean*.

(Point) Estimation problems

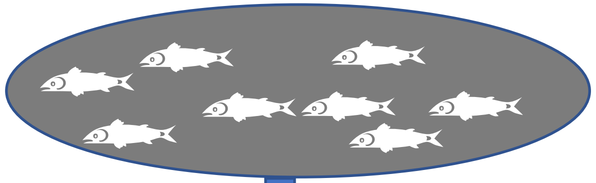


Reasonable estimates

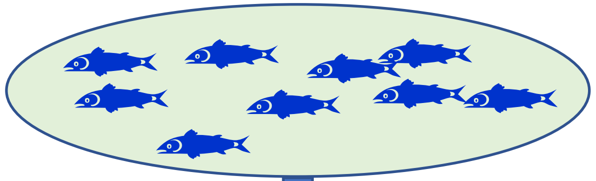
Population parameter	Estimate
Mean μ	$\hat{\mu} = \bar{x}$, the sample mean
Variance σ^2	$\hat{\sigma}^2 = s^2$, the sample variance
Proportion p	$\hat{p} = x/n$, the sample proportion
$\mu_1 - \mu_2$	$\widehat{\mu_1} - \widehat{\mu_2} = \bar{x}_1 - \bar{x}_2$, the different sample means of two ind. rand. samples
$p_1 – p_2$	$\widehat{p}_1 - \widehat{p}_2$, the different sample proportions computed from two ind. rand. samples

Example

Pond 1 = population 1



Pond 2 = population 2



Sample S1

Weight: 1.2, 3.0, 2.3, 1.0; 1.9; 2.1; 1.4; 2.2; 0.7; 1.3; 0.5; 0.8; 2.3; 3.3; 4.1; 3.5; 2.7; 1.3; 3.0; 1.4

Sample S2

Weight: 1.0, 2.3, 1.3; 1.5; 0.3; 1.6; 2.3; 2.6; 3.3; 4.2; 0.8; 2.8; 3.7; 0.5; 4.1; 3.3; 2.1; 3.6; 1.8; 2.1

1) Estimated average weight, variance, standard deviation in each pond.
2) Compare the weight average, variance, standard deviation of the number of fish in two ponds.

Sampling Distributions
The Central Limit Theorem

Statistics: $\bar{X}$

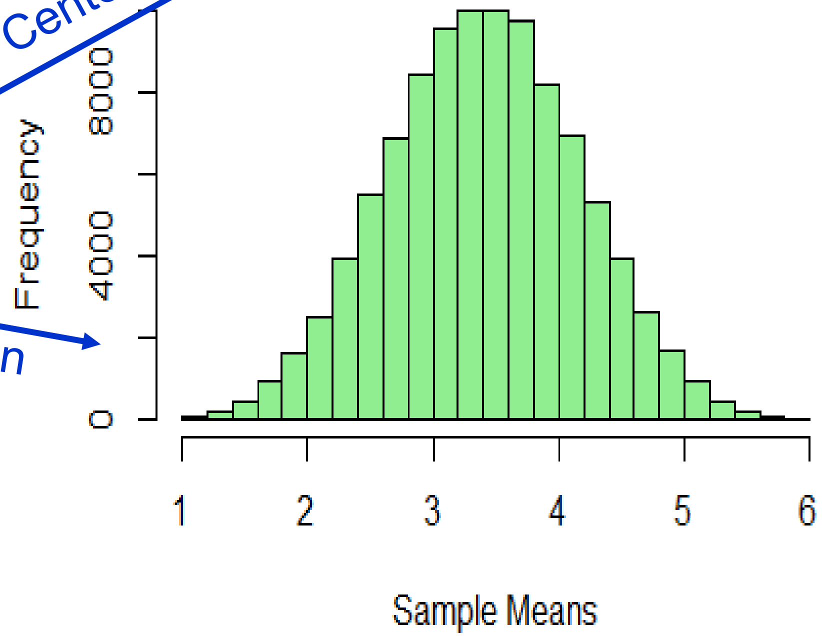
Sampling procedure:
Roll a die 5 times (n=5)
and compute $\bar{x}$

Population parameter:
$\mu = 3.5$

Sample 1 → 4.2
Sample 2 → 3.6
Sample 3 → 4.8
⋮
Sample 10^5 → 2.2

Measure of Center → Mean: 3.49922

Distribution →



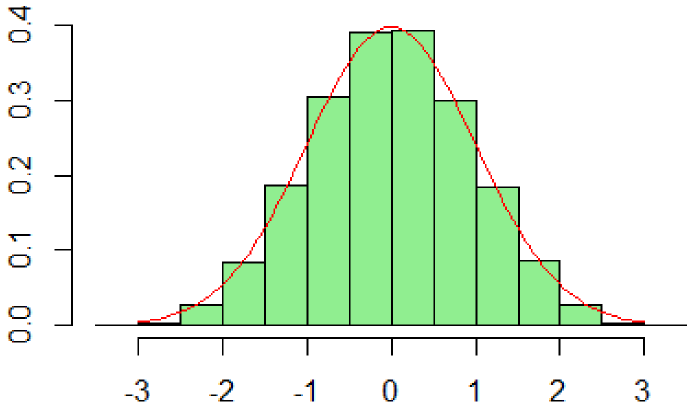
Specific results from 100,000 trials:
The population mean is 3.5; the mean of the 100,000 trials is 3.49922.
Notice the distribution is "normal."

Sampling Distributions
The Central Limit Theorem

If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean μ and variance σ^2/n, if the sample size n is large.

Theorem. (CLT). If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

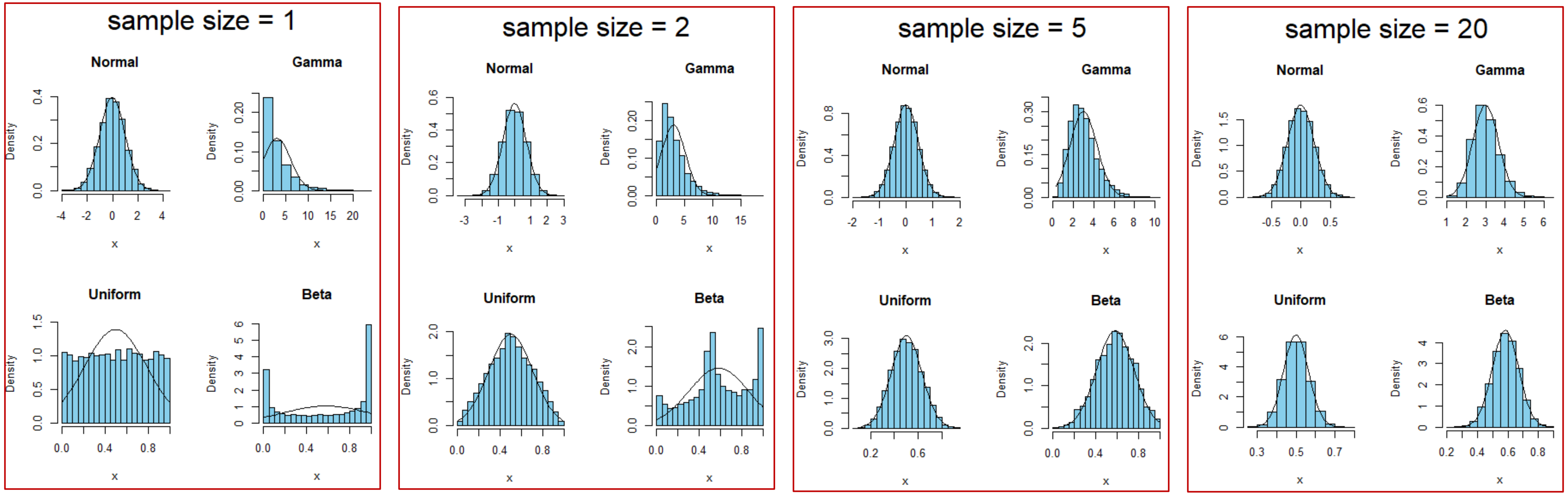
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$



as $n \to \infty$, is the *standard normal distribution*.

$\sigma/\sqrt{n}$: standard error of sample mean

CLT – different sample size

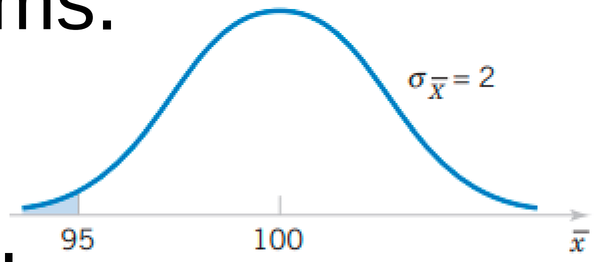


Sampling distribution tends to *standard normal distribution*
as sample size gets larger

CLT - Example

An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal.

Find the probability that a random sample of n 25 resistors will have an average resistance less than 95 ohms.



Note that the sampling distribution of $\overline{\mathrm{X}}$ is normal

with mean 100 ohms and std $\sigma/\sqrt{n}$ = 10/$\sqrt{25}$ = 2.

P($\overline{\mathrm{X}}$ < 95) = P($\frac{\bar{X}-100}{2}$ < $\frac{95-100}{2}$) = P(Z < -2.5) = 0.0062

CLT – Ex2

Customers at a popular restaurant are waiting to be served. Waiting times are independent and exponentially distributed with mean $1/\lambda = 30$ minutes. If 16 customers are waiting what is the probability that their average wait is less than 25 minutes?

Let $\bar{X}$ be the average waiting time of 16 customers.

Since waiting time is exponentially distributed, the mean and standard deviation of individual waiting time is $\mu = \sigma = 30$.

By the central limit theorem, $Z = (\bar{X} - 30)/(30/\sqrt{16}) \sim N(0, 1)$

$P(\bar{X} < 25) = P(Z < -0.667) = 0.252$

Approximate Sampling Distribution of a Difference in Sample Means

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \longrightarrow N(0, 1)$$

Ex. The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of n_1 = 16 components is selected from the "old" process and a random sample of n_2 = 25 components is selected from the "improved" process.

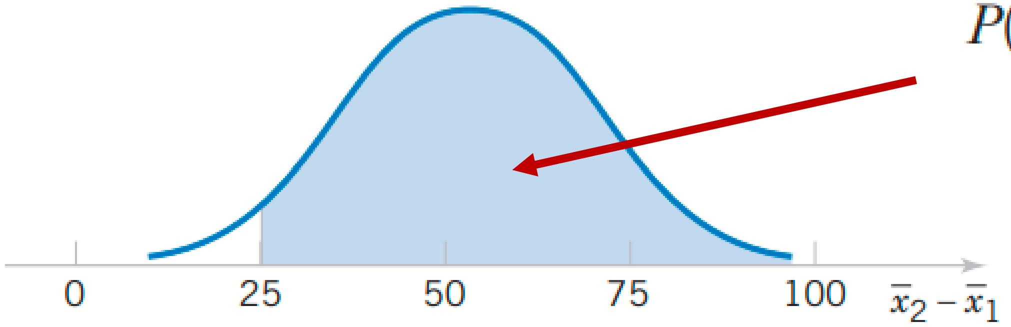
What is the probability that the difference in the two sample means is at least 25 hours?

Assume that the old and improved processes can be regarded as independent populations.

Example (cont.)

- We have
  - $\bar{X}_1 \sim N(\mu_1, \sigma_1{}^2/n_1\ ) = N(5000, 100)$
  - $\bar{X}_2 \sim N(\mu_2, \sigma_2{}^2/n_2\ ) = N(5050, 36)$

➔ $\bar{X}_2 - \bar{X}_1 \sim N(\mu_1 - \mu_2, \sigma_1{}^2/n_1 + \sigma_2{}^2/n_2) = N(50, 136)$



$$P(\bar{X}_2 - \bar{X}_1 \geq 25) = P(Z \geq -2.14)$$
$$= 0.9838$$

Like hurricanes and earthquakes, geomagnetic storms are natural hazards with possible severe impact on the Earth. Severe storms can cause communication and utility breakdowns, leading to possible blackouts. The National Oceanic and Atmospheric Administration beams electron and proton flux data in various energy ranges to various stations on the Earth to help forecast possible disturbances. The following are 25 readings of proton flux in the 47-68 kEV range (units are in p / (cm2-sec-sterMeV)) on the evening of December 28, 2011: 2310 2320 2010 10800 2190 3360 5640 2540 3360 11800 2010 3430 10600 7370 2160 3200 2020 2850 3500 10200 8550 9500 2260 7730 2250

a/ Find a point estimate of the mean proton flux in this time period.

b/ Find a point estimate of the standard deviation of the proton flux in this time period.

c/ Find an estimate of the standard error of the estimate in part a/.

Exercises

- $(3x + 13) \bmod 21$
- $(5x + 15) \bmod 21$
- $(7x + 17) \bmod 21$