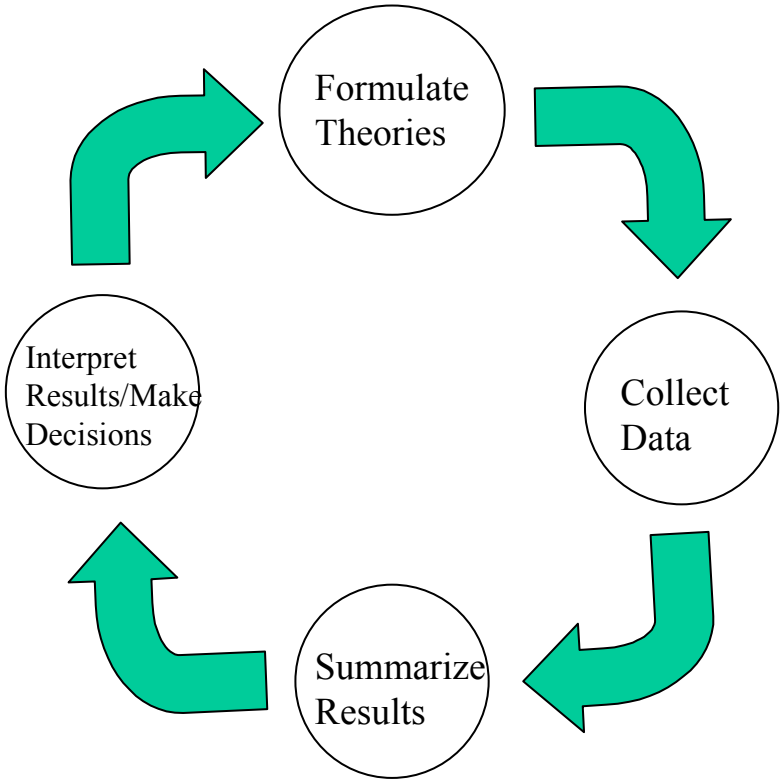
Chapter Goals

To use information from the sample to make inference about the population.



Statistical Sampling

Definition *Parameters are numerical descriptive measures of populations.*

Definition *Statistics are numerical descriptive measures of samples (more generally, a quantity computed from the observations in a sample).*

Definition *Estimators are sample statistics that are used to estimate the unknown population/process parameter. E.g., the sample mean $\overline{X}$ is a point estimator of the population mean μ. Note that $\overline{X}$ is also a random variable.*

Question? How close is our sample statistic to the true, but unknown population parameter?

Notations

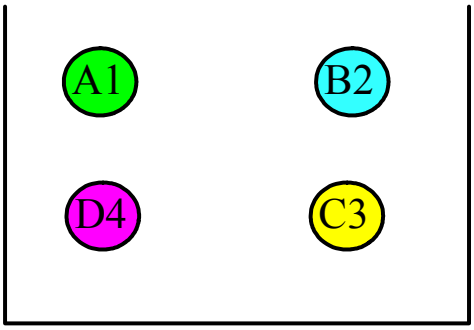
Parameters	Statistics
μ	$\overline{X}, Md,$ mode, $\hat{\mu}$
σ^2	$s^2, \hat{\sigma}$
p	$\hat{p}$

Sampling Distributions

Objective: To find out how the sample mean $\bar{X}$ varies from sample to sample. In other words, we want to find out the sampling distribution of the sample mean.

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

Sampling Distribution: An Example



Population Characteristics

Population consists of $N = 4$ items A, B, C, and D, with values 1, 2, 3, and 4 respectively.

Summary measures:

$$\mu = \frac{\sum_{i=1}^{4} X_i}{4} = \frac{1+2+3+4}{4} = 2.5$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} = \frac{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2}{4} = 1.25$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.25} = 1.118$$

Sampling Distribution of the Sample Mean

Now let us draw all possible samples of size $n = 2$.

16 possible samples				
1st Observation	2nd Observation			
	A	B	C	D
A	AA	AB	AC	AD
B	BA	BB	BC	BD
C	CA	CB	CC	CD
D	DA	DB	DC	DD

16 corresponding sample means				
1st Observation	2nd Observation			
	A	B	C	D
A	1.0	1.5	2.0	2.5
B	1.5	2.0	2.5	3.0
C	2.0	2.5	3.0	3.5
D	2.5	3.0	3.5	4.0

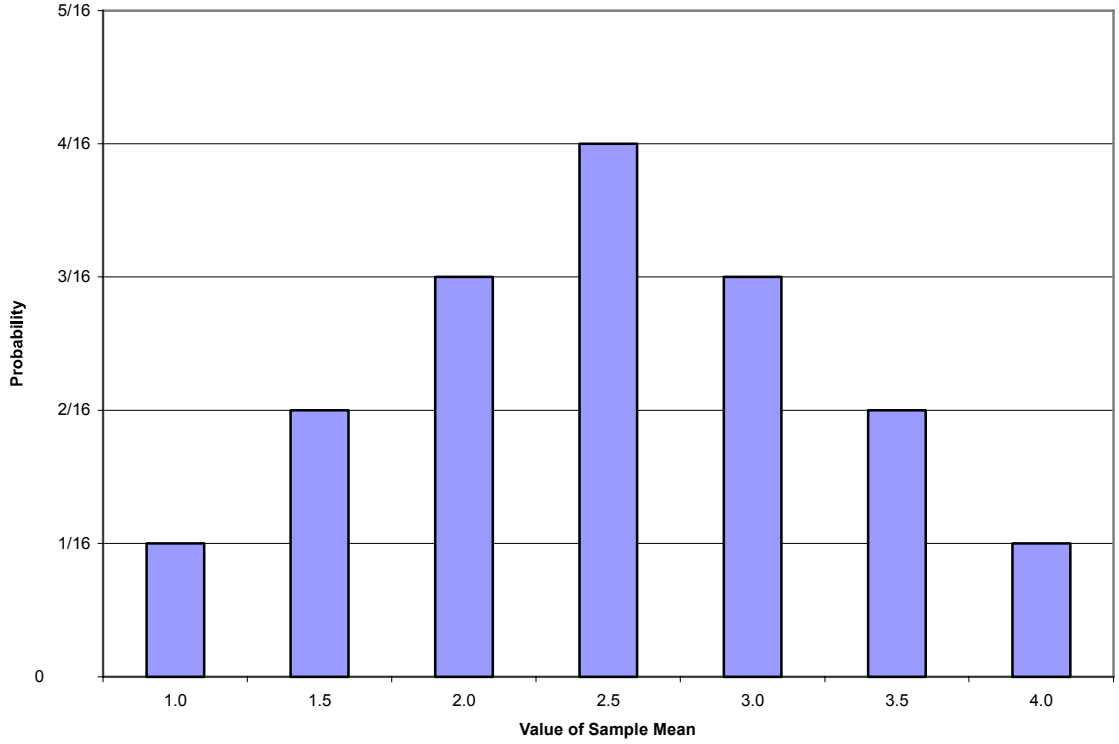
Summary measures:

$$\mu_{\overline{X}} = \frac{\sum_{i=1}^{16} \overline{X}_i}{N} = \frac{1.0+1.5+...+4.0}{16} = 2.5$$

$$\sigma^2_{\overline{X}} = \frac{\sum_{i=1}^{16} (\overline{X}_i - \mu_{\overline{X}})^2}{N} = \frac{(1.0-2.5)^2+(1.5-2.5)^2+...+(4.0-2.5)^2}{16} = .625$$

$$\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}} = \sqrt{.6241} = .79$$

Sampling Distribution of $\overline{X}$



$E(\bar{X}) =$

Comparison of Population and Sampling Distribution

	Population	Sampling
μ (Mean)	2.5	2.5
σ^2 (Variance)	1.25	0.625
σ (Standard Deviation)	1.118	0.79

Question: What is the relationship between the variance in the population and sampling distributions?

Empirical Derivation of Sampling Distribution of $\overline{X}$

- Select a random sample of n observations from a given population or process.
- Compute $\overline{X}$.
- Repeat steps (1) and (2) a "large" number of times.
- Construct a relative frequency histogram of the resulting set of $\overline{X}$'s.

Points to Remember about the Sampling Distribution of $\overline{X}$

1. The mean of the sampling distribution of $\overline{X}$ is the same as the mean of the population/process being sampled from. That is, $E(\overline{X}) = \mu_{\overline{X}} = \mu_X = \mu$.
2. The variance of the sampling distribution of $\overline{X}$ is equal to the variance of the population/process being sampled from divided by the sample size. That is, $E\left[(\bar{X} - \mu_{\overline{X}})^2 \right] = \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$
3. If the original population is normally distributed, then for **any** sample size n the distribution of the sample mean is also normal. That is, if $X \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
4. If the distribution of the original population is not known, but n is sufficiently "large", the distribution of the sample mean is **approximately** normal with mean and variance given as $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. This result is known as the **central limit theorem (CLT)**.

Example *Let* X *the length of pregnancy be* $X \sim N(266, 256)$.

- What is the probability that a randomly selected pregnancy lasts more than 274 days. I.e., what is $P(X > 274)$?

- Suppose we have a random sample of $n = 25$ pregnant women. Is it less likely or more likely (as compared to the above question), that we might observe a sample mean pregnancy length of more than 274 days. I.e., what is $P(\overline{X} > 274)$?

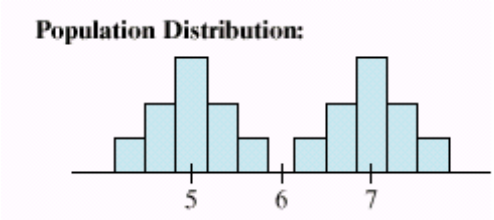
Let's do it! 9.8

Exercise *The model for breaking strength of steel bars is normal with a mean of* 260 *pounds per square inch and a variance of* 400. *What is the probability that a randomly selected steel bar will have a breaking strength greater than* 250 *pounds per square inch?*

Exercise *A shipment of steel bars will be accepted if the mean breaking strength of a random sample of* 10 *steel bars is greater than* 250 *pounds per square inch. What is the probability that a shipment will be accepted?*

Let's do it! 9.9

The following histogram shows the population distribution of a variable X. How would the sampling distribution of $\overline{X}$ look, where the mean is calculated from random samples of size 150 from the above population?



Desirable Characteristics of Estimators

Definition *An estimator $\hat{\theta}$ is **unbiased** if the mean of its sampling distribution is equal to the population parameter θ to be estimated. That is, $\hat{\theta}$ is an unbiased estimator of θ if $E\{\hat{\theta}\} = \theta$. For example, $E(\overline{X}) = \mu$, therefore $\overline{X}$ is an unbiased estimator of μ.*

Definition *An estimator is a **consistent** estimator of a population parameter θ if the larger the sample size, the more likely it is that the estimate will be closer to θ. Is $\overline{X}$ a consistent estimator?*

Definition *The **efficiency** of an unbiased estimator is measured by the variance of its sampling distribution. If two estimators based on the same sample size are both unbiased, the one with the smaller variance is said to have greater relative efficiency than the other.*