Sampling Distributions—Statistics as RV's

- Sample values are measurements or observations of RVs ➔ the value for a sample statistic will vary in a random manner from sample to sample
- Sample statistics are RVs because different samples can lead to different values for the sample statistics

Sampling Distributions

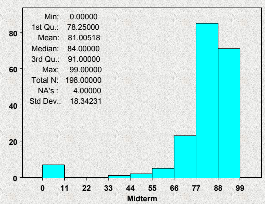
New Definition

Sampling distribution **(*of a sample statistic*)**---the distribution of possible values of a statistic for **repeated samples of the same size** from a population

Statistical Inference

Want to make conclusions about population parameters on the basis of sample statistics.

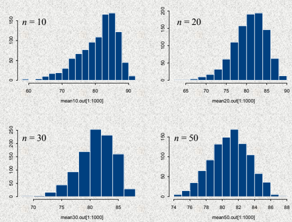
- **Confidence Intervals**—interval of values that the researcher is fairly certain will cover the true, unknown value of the population parameter
- **Hypothesis Tests (significance testing)**—uses sample data to attempt to reject a hypothesis about the population

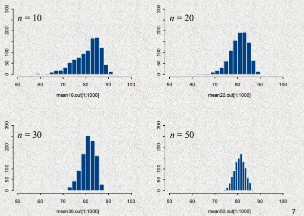


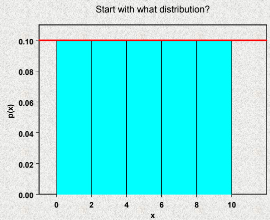
Mean Midterm Score = 81.01

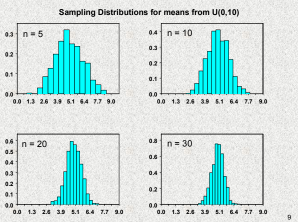
SD Midterm Score = 18.34

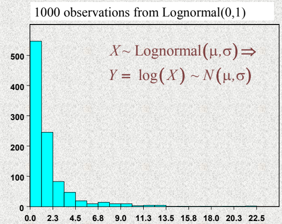
	mean10	mean20	mean30	mean50
Sample 1	86.89	83.05	84.45	82.22
Sample 2	77.40	73.16	82.76	82.90
Sample 3	85.90	78.68	84.30	82.06
Sample 4	79.90	73.15	82.14	81.06
Sample 5	86.40	81.10	80.31	78.51
Sample 6	80.40	82.61	81.47	82.52

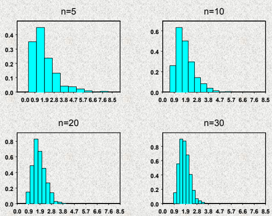












Sampling Distributions

Many common, practical problems involve estimating mean values from samples; interest is really in making an inference about the mean, μ, of some population

What do we know: the sample mean, $\overline{x}$, is often a good estimator of μ

Example: Midterm exam scores—we will consider the class data to represent the population

Consider histograms for $\overline{x}$ based on n = 10, 20, 30 and 50 samples

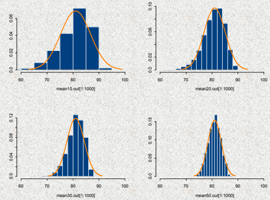
μ = 81.01

σ = 18.34

} We are using parameter designators since we are taking the full class data to be the population

	mean10	mean20	mean30	mean50
Sample 1	86.89	83.05	84.45	82.22
Sample 2	77.40	73.16	82.76	82.90
Mean of Sampling Distribution based on 1000 draws of size *n*	80.86	80.71	81.10	80.95

the normal probability distribution approximates the computer-generated sampling distribution very well



μ = 81.01

σ = 18.34

	mean10	mean20	mean30	mean50
Sample 1	86.89	83.05	84.45	82.22
Sample 2	77.40	73.16	82.76	82.90
Mean of Sampling Distribution based on 1000 draws of size *n*	80.86	80.71	81.10	80.95
SD of Sampling Distribution based on 1000 draws of size *n*	5.78	4.13	3.09	2.36
Sigma/sqrt(*n*)	5.80	4.10	3.35	2.59

$18.34/\sqrt{10}$

Sampling Distributions

Properties of the sampling distribution of $\bar{x}$

1. Mean of sampling distribution equals mean of sampled population:

$$\mu_{\bar{x}} = E(\bar{x}) = \mu$$

2. Standard deviation of sampling distribution equals

$$\frac{\text{Standard deviation of sampled population}}{\text{Square root of sample size}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distributions

For the midterm scores example:

$$\sigma_{\bar{x}} = \frac{18.34}{\sqrt{10}} = 5.80 \qquad \sigma_{\bar{x}} = \frac{18.34}{\sqrt{30}} = 3.35$$

$$\sigma_{\bar{x}} = \frac{18.34}{\sqrt{20}} = 4.10 \qquad \sigma_{\bar{x}} = \frac{18.34}{\sqrt{50}} = 2.59$$

Central Limit Theorem (CLT)

Consider a random sample of *n* observations selected from a population (*any population*) with mean μ and standard deviation σ

Then, when *n* is sufficiently large, the sampling distribution of $\bar{x}$ will be **approximately a normal distribution** with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

The larger the sample size, the better will be the normal approximation to the sampling distribution of $\bar{x}$

If *X* is normally distributed then $\bar{x}$ will be normally distributed

Central Limit Theorem (CLT)

The sum of a random sample of n observations, $\sum x$, will also possess a sampling distribution that is approximately normal for large samples;

$$\mu_{\sum x} = n\mu; \quad \sigma_{\sum x} = \sqrt{n}\sigma$$

How large is large?

The greater the skewness of the sampled population distribution, the larger the sample size must be before the normal distribution is an adequate approximation for the sampling distribution of $\bar{x}$; for many sampled population, sample sizes of $n \geq 30$ will suffice.

Z-scores

$$Z \approx \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \text{ where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$Z \xrightarrow{\text{approaches}} N(0,1) \text{ as } n \to \infty$$

$$\text{If } X \sim N(\mu, \sigma^2) \text{ then } Z \sim N(0,1)$$

We do not need to depend on the CLT to "induce" normality; the normality is exact, not approximate