Sampling Distributions—Statistics as *RV*s

Previously Defined Terms

Parameter---a numerical descriptive measure of a population

Sample statistic---a numerical descriptive measure of a sample; calculated from the observations in a sample

----

Sample values are measurements or observations of *RV*s ➔ the value for a sample statistic will vary in a random manner from sample to sample

Sample statistics are *RV*s because different samples can lead to different values for the sample statistics

Sampling Distributions

New Definition

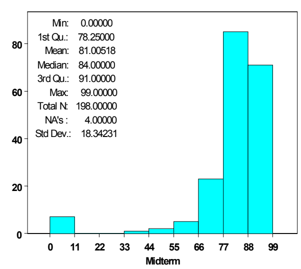
Sampling distribution **(*of a sample statistic*)**---the distribution of possible values of a statistic for **repeated samples of the same size** from a population

Statistical Inference

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

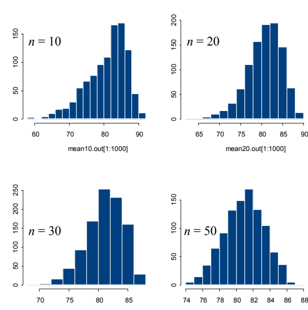
- **Confidence Intervals**—interval of values that the researcher is fairly certain will cover the true, unknown value of the population parameter

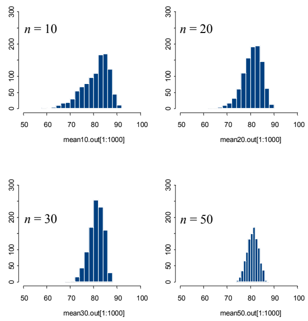
- **Hypothesis Tests (significance testing)**—uses sample data to attempt to reject a hypothesis about the population

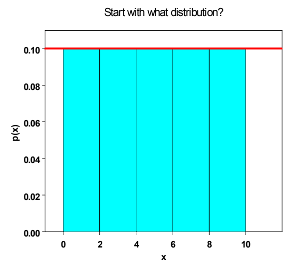


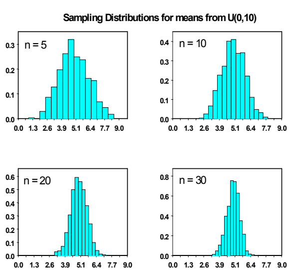
Mean midterm score = 81.01;
SD midterm score = 18.34

	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



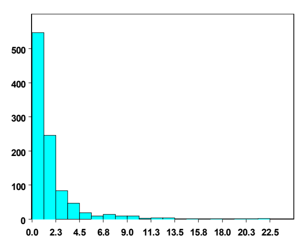


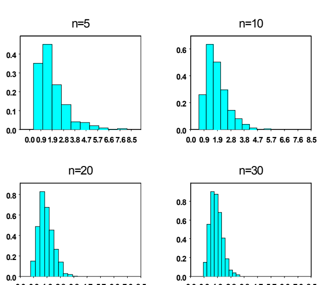




1000 observations from Lognormal(0,1)

$$X \sim \text{Lognormal}(\mu,\sigma) \Rightarrow Y = \log(X) \sim N(\mu,\sigma)$$





Sampling Distributions

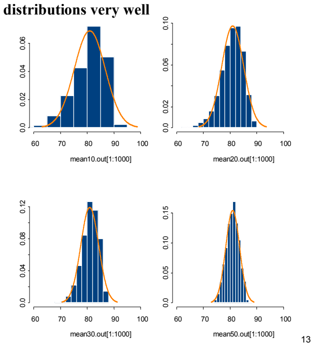
Many common, practical problems involve estimating mean values from samples; interest is really in making an inference about the mean, μ , of some population

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

<u>**Example**</u>: Midterm exam scores—we considered the class data to represent the population

Look again at histograms for $\bar{x}$ based on n = 10, 20, 30 and 50 samples

The normal probability distribution approximates the computer-generated sampling distributions 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

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

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

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

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 $\bar{x}$

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.

Central Limit Theorem (CLT)

If $X \sim N(\mu_X, \sigma_X)$, then $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$, regardless of the value of *n*.

If $X \sim ?$ with parameters mean = μ_X and SD = σ_X ,

then $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$, for *n* sufficiently large.

What does this mean?

- Many of the procedures we will be learning have an assumption of normality
- May have to access normality

Accessing Normality

- Look at a histogram—look for unimodal, approximately symmetric, no profound outliers
- Identify potential outliers; reject assumption of normality of there is more than one outlier present
- Normal quantile (probability) plot—use software
- Statistical tests for normality
    - Chi-square goodness-of-fit
    - Shapiro Wilks
    - Others

If data deemed not normal, what can you do?

- Consider data transformations to achieve approximate normality
- Use Nonparametric methods