The Central Limit Theorem

Summer 2003

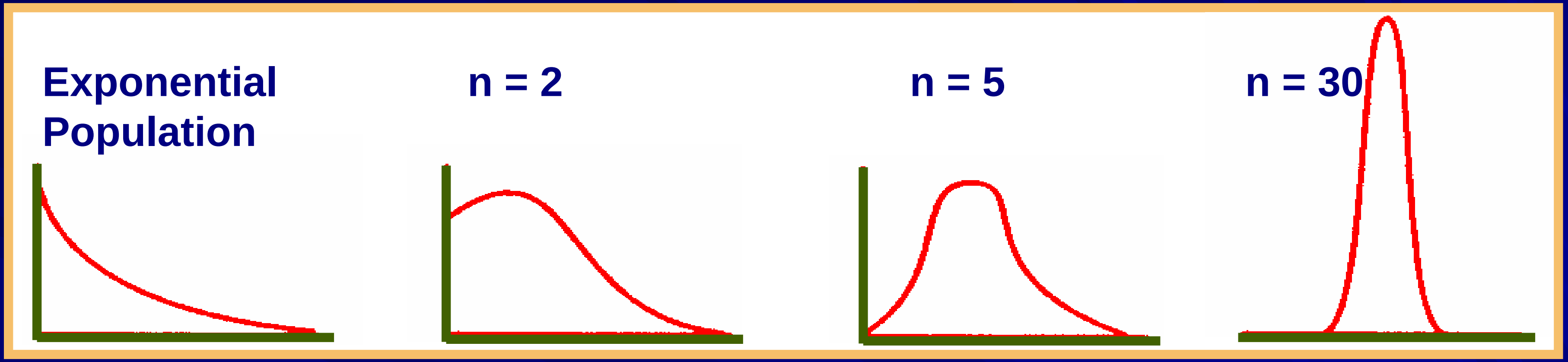
The Central Limit Theorem (CLT)

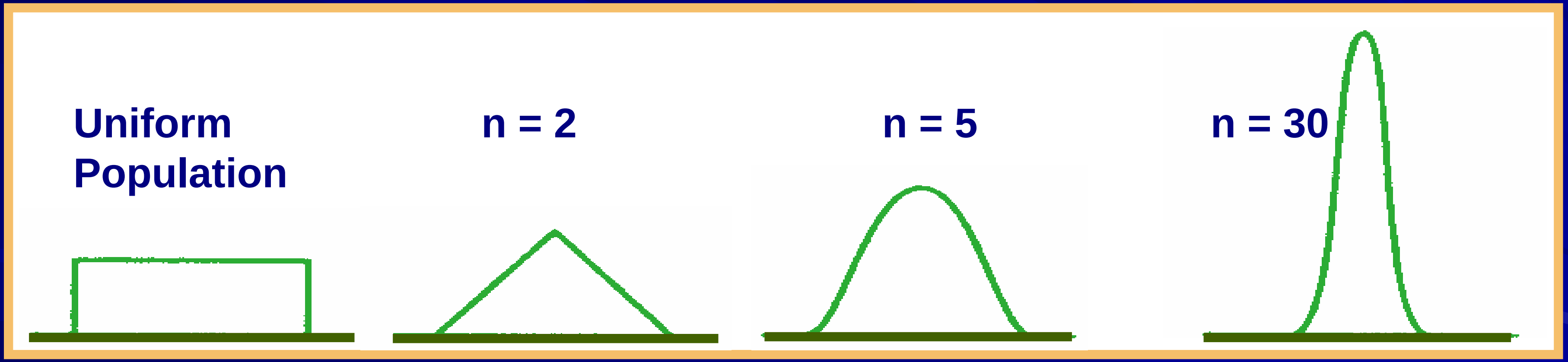
If random variable S_n is defined as the sum of n independent and identically distributed (i.i.d.) random variables, $X_1, X_2, \ldots, X_n$; with mean, μ, and std. deviation, σ.

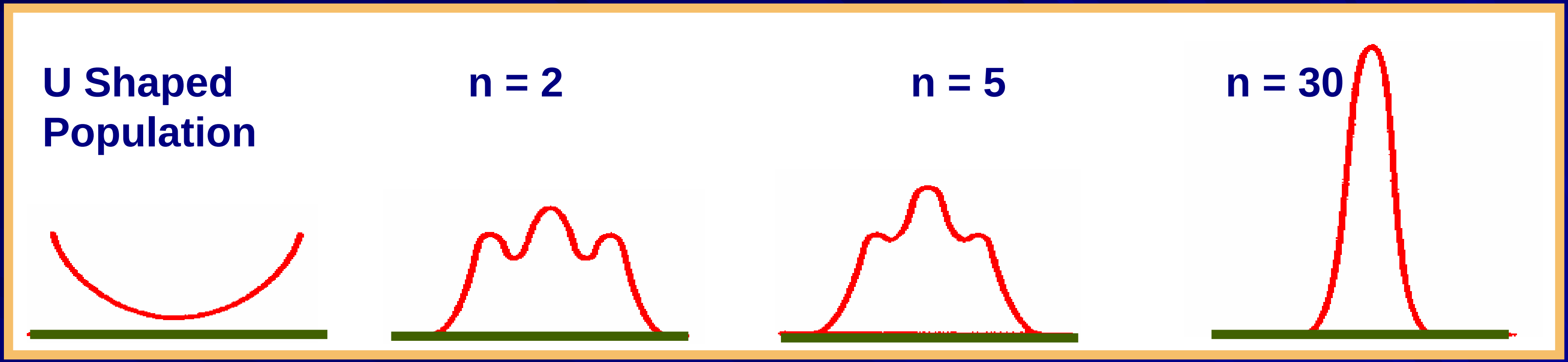
Then, for large enough n (typically $n \geq 30$), S_n is approximately Normally distributed with parameters: $\mu_{Sn} = n\mu$ and $\sigma_{Sn} = \sqrt{n}\,\sigma$

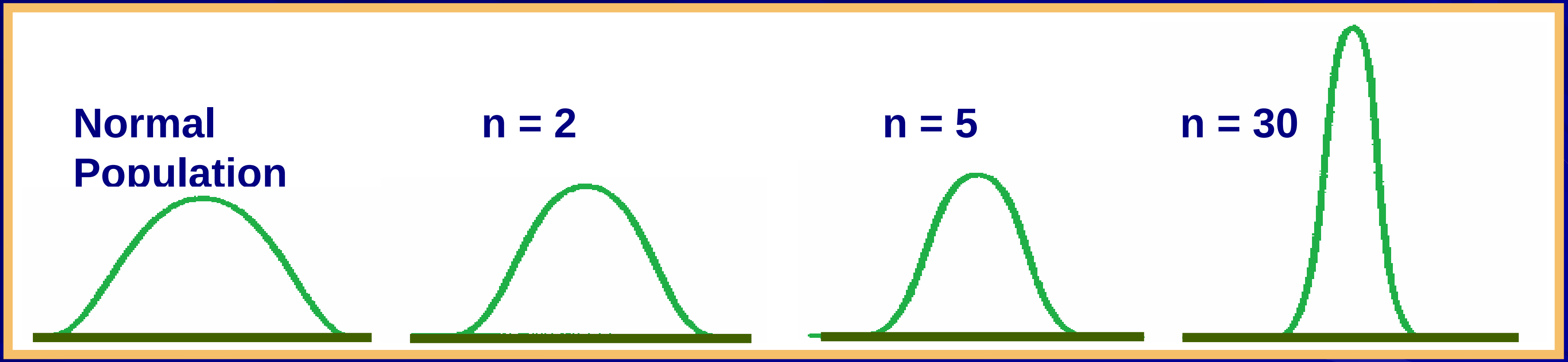
This result holds regardless of the shape of the X distribution (i.e. the Xs don't have to be normally distributed!)

Examples


Exponential
Population
n = 2
n = 5
n = 30

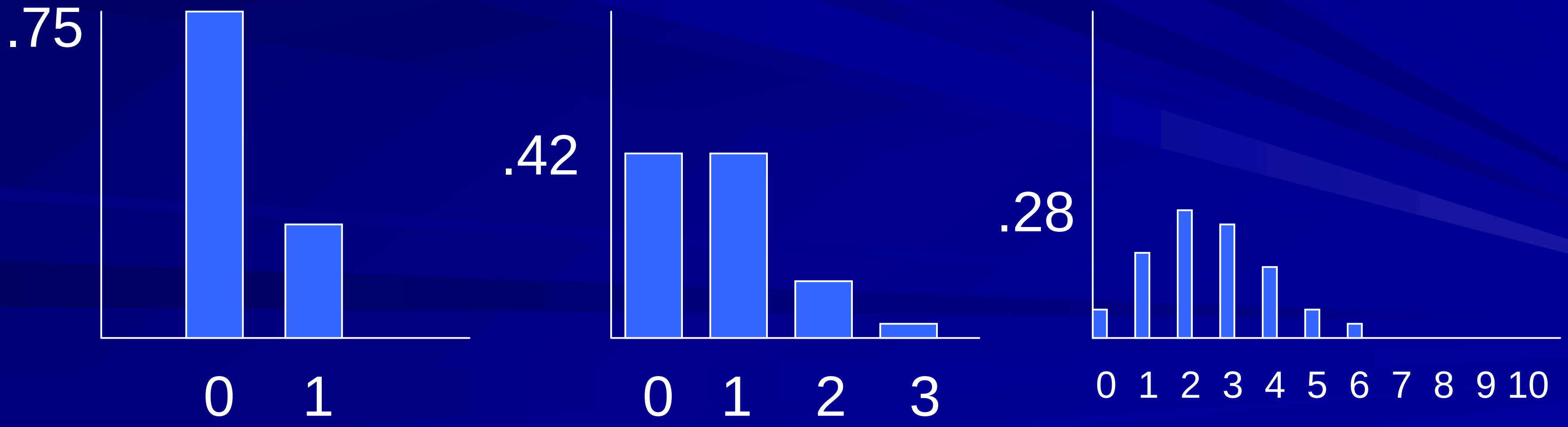

Uniform
Population
n = 2
n = 5
n = 30


U Shaped Population
n = 2
n = 5
n = 30


Normal Population
n = 2
n = 5
n = 30

An Example

- Each person take a coin and flip it twice (a pair)
- The distribution of two heads vs. other is binomial
- Now flip your coin 3 new pairs, report #(two heads)
- New variable H3 = sum of n=3 binomial (p=.25)
- Now flip each coin 10 new pairs, report #(two heads)

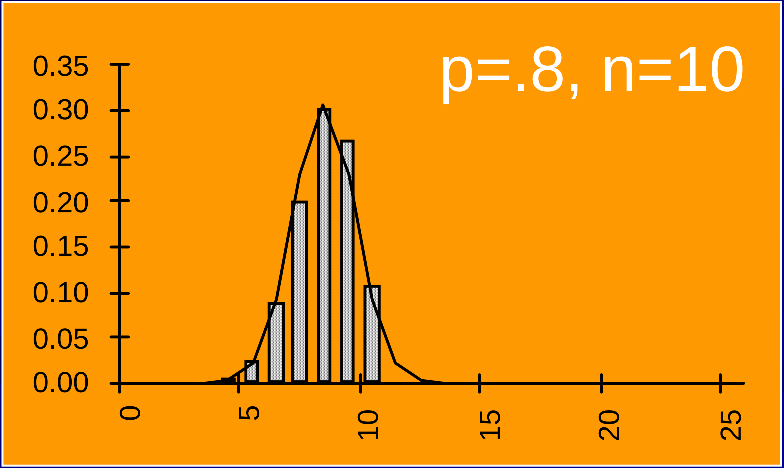


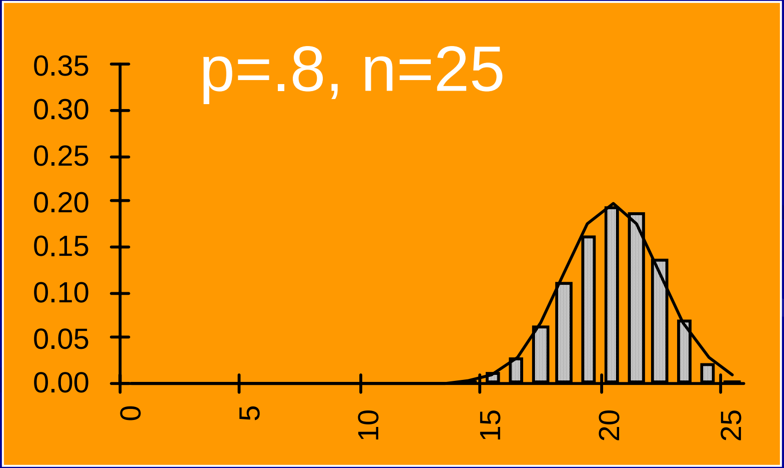
For any binomial r. v. X (n,p)

- X can be seen as the sum of n i.i.d. (independent, identically distributed) 0-1 random variables Y, each with probability of success p (i.e., P(Y=1)=p).

  $$X = Y_1 + Y_2 + \dots Y_n$$

- In general we can approximate r.v. X binomial (n,p) using r.v. Y Normal: $\mu = np$; $\sigma = \sqrt{np(1-p)}$





Using the Normal Approximation to The Binomial…

- If r.v. X is Binomial (n, p) with parameters:

  $E(X) = np; \quad VAR(X) = np(1-p);$

- We can use Normal r.v. Y with mean np and variance np(1-p) to calculate probabilities for r.v. X (i.e., the binomial)

- The approximation is good if n is large enough for the given p, i.e, must pass the following test:

$$\text{Must have: } np \geq 5 \text{ and } n(1-p) \geq 5$$

Small Numbers Adjustment

To calculate binomial probabilities using the normal approximation we need to consider the “0.5 adjustment”:

1. Write the binomial probability statement using “≥” and “≤”: e.g. $P(3<X<9)= P(4 \leq X \leq 8)$
2. Draw a picture of the normal probability Y you want to calculate and enlarge the area making a 0.5 adjustment(s) to the edge(s). This is because each discrete probability is represented by a range in the normal probability, e.g., $P(X=4)$ is ~ $P(3.5<Y<4.5)$
3. Calculate the size of the area (Normal probability)

   (The book ignores this adjustment. The example on page 139 should have ~ $P(Y \geq 9.5)$)

This adjustment is less important as n becomes larger.

Example: An electrical component is guaranteed by its suppliers to have 2% defective components. To check a shipment, you test a random sample of 500 components. If the supplier's claim is true, what is the probability of finding 15 or more defective components?

X = number of defective components found during the test.

X is Binomial(500, 0.02).

We want $P(X \geq 15) = P(X=15) + P(X=16) \ldots + P(X=500)$

Can we use r.v. Y Normal with:

mean=500(0.02) = 10 and sd = sqrt{500*.02(1-.02) = 3.13 ?

Yes! np = 500 * 0.02 = 10 and n (1-p) = 500 * 0.98=490 (> 5)

Using the ".5 adjustment" we see that $P(X \geq 15) \sim P(Y \geq 14.5)$

Easiest way is to calculate $P(Y \geq 14.5) = 1-P(Y<14.5)$

$= 1-F(z) =(14.5-10)/3.13=1.44)= 1-F(1.44) = 1- 0.9251$

$= 0.0749$

The Central Limit Theorem (for the mean)

If random variable $\overline{X}$ is defined as the average of n independent and identically distributed random variables, X_1, X_2, …, X_n; with mean, μ, and Sd, σ. Then, for large enough n (typically $n \geq 30$), $\overline{X}_n$ is approximately Normally distributed with parameters: $\mu_x = \mu$ and $\sigma_x = \sigma/\sqrt{n}$

Again, this result holds regardless of the shape of the X distribution (i.e. the Xs don't have to be normally distributed!).

The CLT for the mean and statistical sampling: (chapter 4)

Point estimate:

$$\overline{X} = \frac{\sum X}{n}$$

Interval Estimate:

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

or

$$\overline{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z \frac{\sigma}{\sqrt{n}}$$

- Idea: If we take a large enough random sample (i.e. n>=30) for r.v. X (i.e., the population of interest), then we can estimate the mean, μ , for r.v. X even if we do not know the distribution of X. Note: use the sample SD, s, if the population sd, σ, is not known:

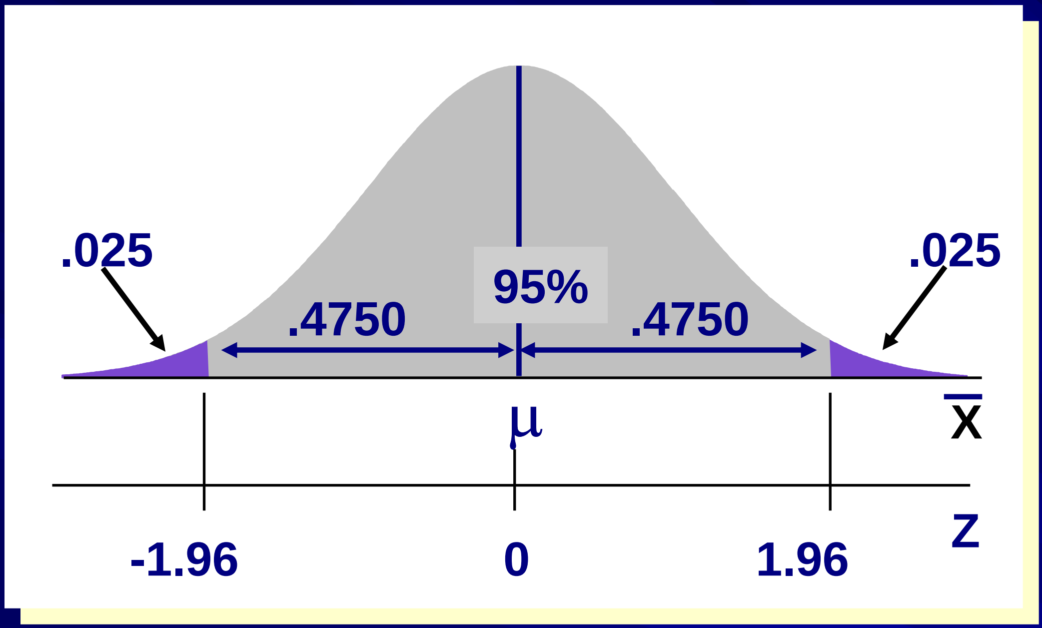
$$S^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1}$$

$$S = \sqrt{S^2}$$

More on s vs. σ later

- The value of z is determined by the confidence level assigned to the interval (see next slide)

Values of Z for selected confidence levels:



Confidence Level	Z Value
90% (α=0.1)	1.645
95% (α=0.05)	1.96
98% (α=0.02)	2.33
99% (α=0.01)	2.575

We would for example say that we are 95% confident the true mean for x falls in the interval:

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Confidence Limits Are a Way of Knowing What We Know

- Estimates and forecasts are difficult to evaluate for quality or degree of confidence
- What will the Dow Jones be six months from now?

Estimation and Confidence Limits

- How many employees (in total) did IBM have worldwide on Dec. 31, 2002?
- After making your best estimate, give a low estimate and high estimate so you are 95% sure that the true answer falls within these limits

Low____________ High_____________

315,889

Overconfidence

Respondant	***Topic***	***Target***	***Result***
Harvard MBAs	Trivia facts	2%	46%
Kellogg MBAs	Starting salary	49%	85%
Chemical employees	Industry & co. facts	10%	50%
Computer managers	Business: co. facts:	5% 5%	80% 58%

More Overconfidence

- “A severe depression like that of 1920-1921 is outside the range of probability”

  Harvard Econ. Soc’ty W’kly Letter, Nov. 16, 1929
- “With over 50 foreign cars already on sale here, the Japanese auto industry isn’t likely to carve out a big slice of the U.S. market for itself”

  Business Week, August 2, 1968
- “There is no reason anyone would want a computer in their home”

  Ken Olson, DEC founder, 1977

Overcoming Overconfidence

- Commercial Loan Officer, Midwest Bank: “Are we overconfident about the competition?”
- First, convince the boss.

  Tactic: overconfidence test
- Second, avoid the mistakes

  Tactic: competitor alert file
- Result: within 3 weeks, saved $160K account

Summary and Look Ahead

- The Central Limit Theorem allows us to use the Normal distribution, which we know a lot about, to approximate almost anything, as long as some requirements are met (e.g., $n>=30$)
- Confidence limits are a way of estimating our degree of knowledge
- People typically think they know more than they do (we don't like uncertainty)
- Next class we use the same tools to look at statistical sampling
- Homework #2 is due!!