Basic Biostatistics

Statistics for Public Health Practice

B. Burt Gerstman

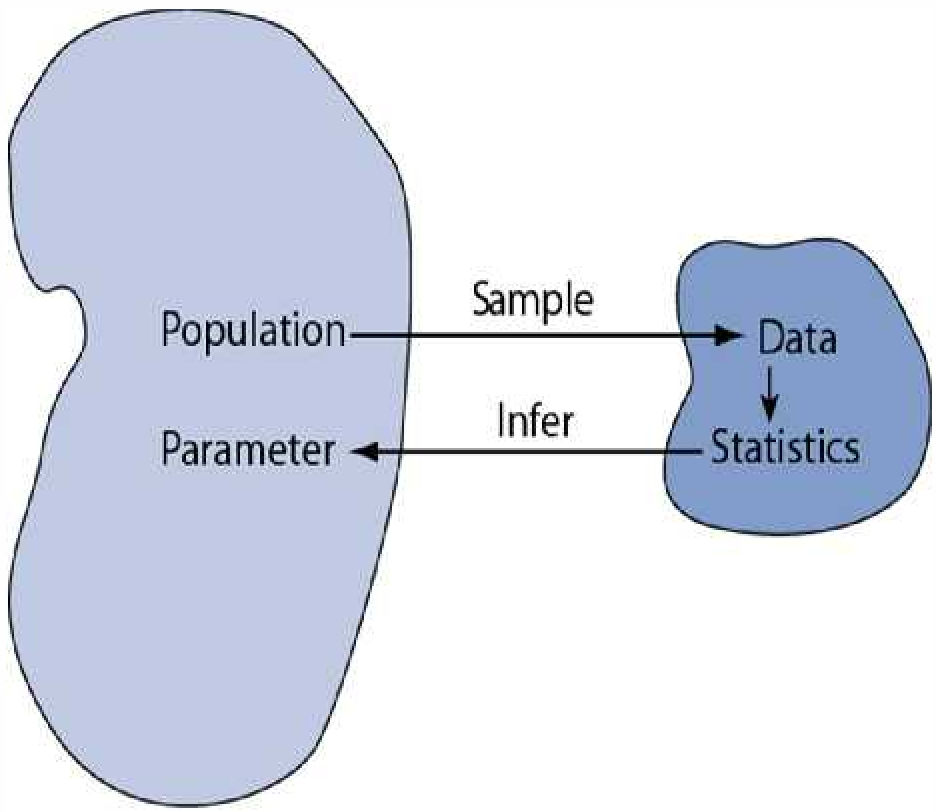
Chapter 8: Introduction to Statistical Inference

In Chapter 8:

§8.1: Concepts

Statistical inference is the act of generalizing from a **sample** to a **population** with calculated degree of certainty.

We want to learn about population *parameters* …



…but we can only calculate *sample statistics*

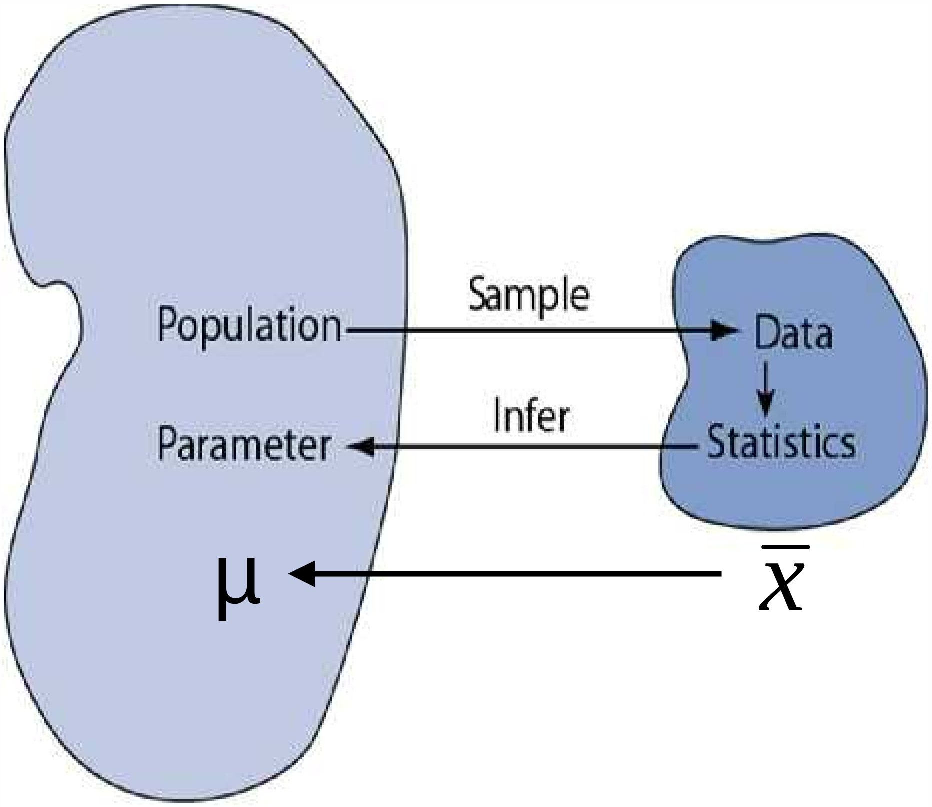
Parameters and Statistics

It is essential that we draw distinctions between parameters and statistics

	Parameters	**Statistics**
Source	Population	Sample
Calculated?	No	Yes
Constants?	Yes	No
Examples	μ, σ, p	$\bar{x}, s, \hat{p}$

Parameters and Statistics

We are going to illustrate inferential concept by considering how well a given sample mean “x-bar” reflects an underling population mean µ



Precision and reliability

- How **precisely** does a given sample mean (x-bar) reflect underlying population mean (μ)? How reliable are our inferences?
- To answer these questions, we consider a **simulation experiment** in which we take all possible samples of size *n* taken from the population

Simulation Experiment

- Population (Figure A, next slide)

  N = 10,000

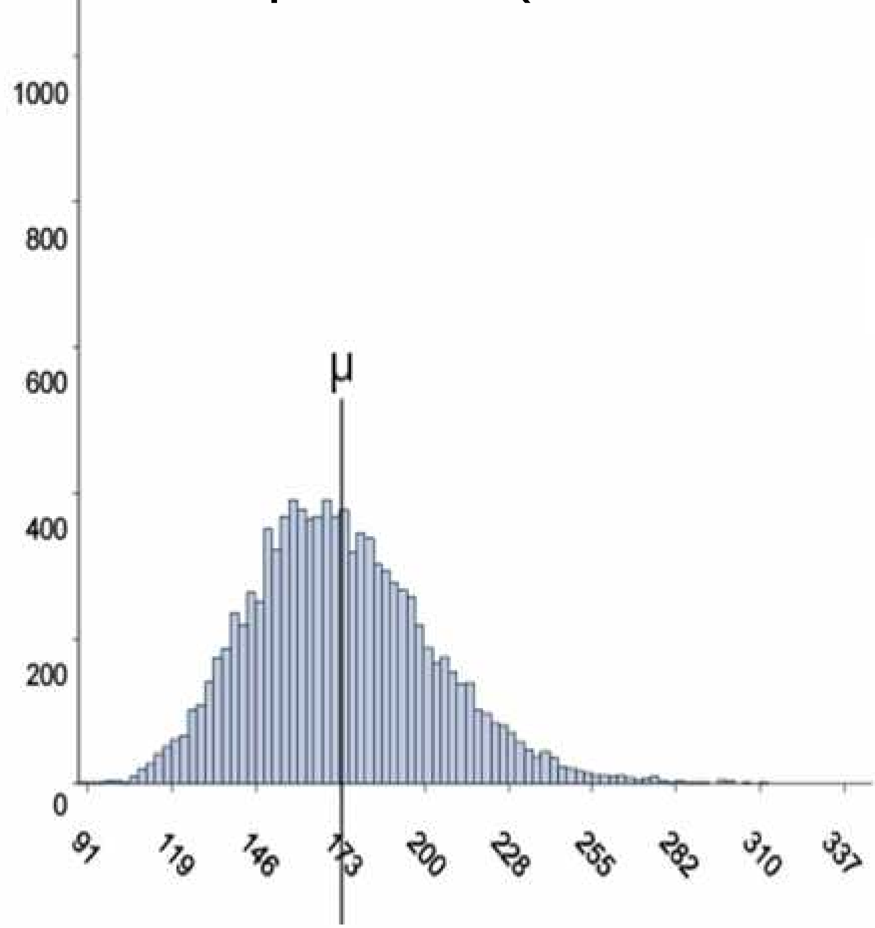
  Lognormal shape (positive skew)

  $\mu = 173$

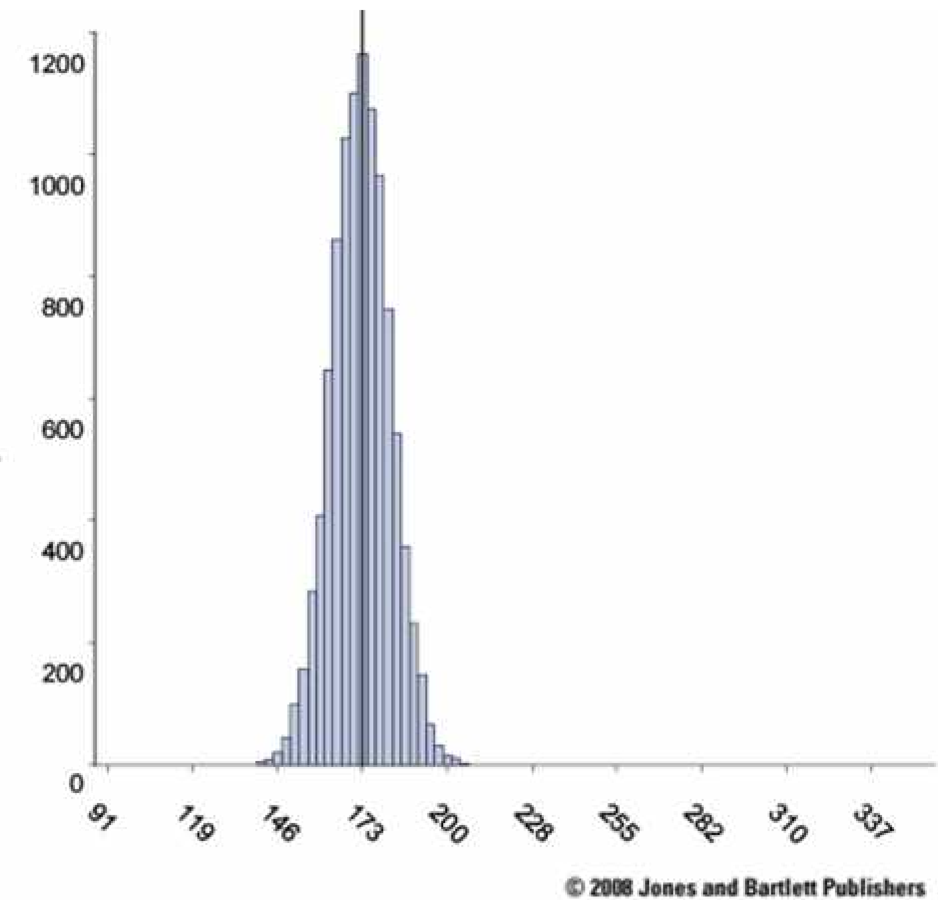
  $\sigma = 30$

- Take repeated SRSs, each of $n = 10$
- Calculate x-bar in each sample
- Plot x-bars (Figure B , next slide)

A. Population (individual values)

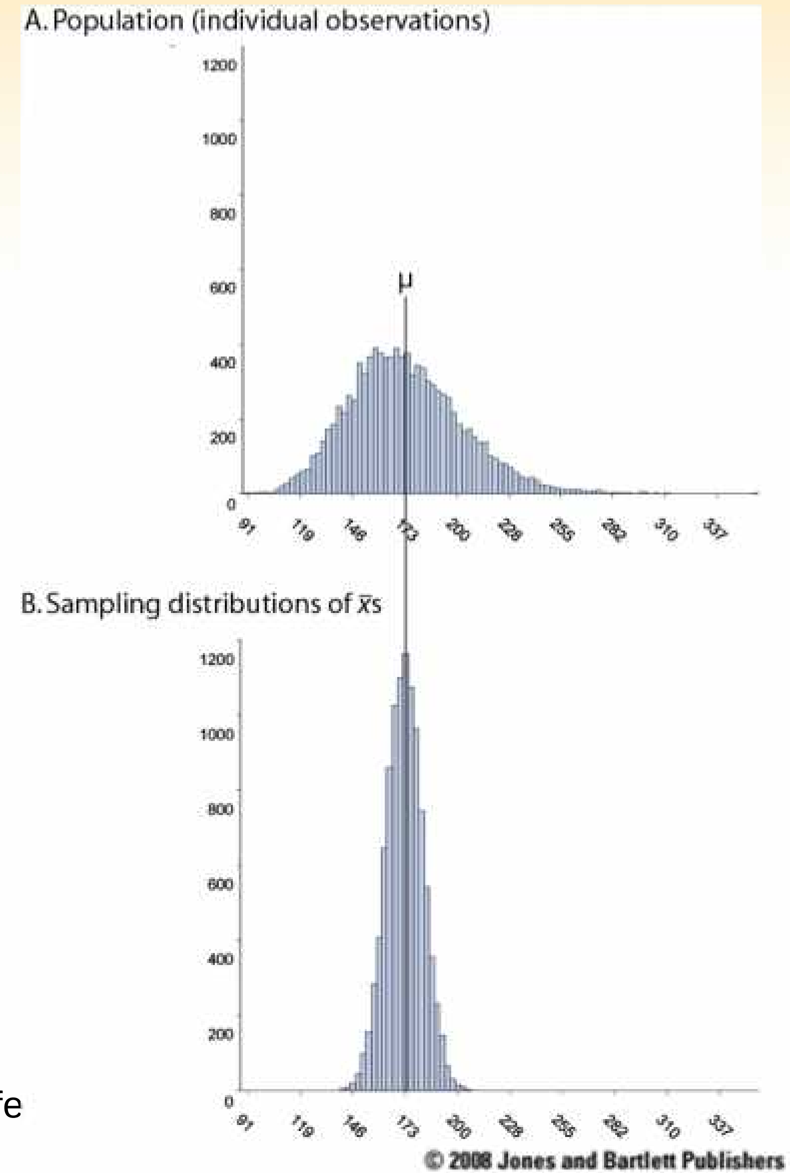


B. Sampling distribution of x-bars



Simulation Experiment Results

1. Distribution B is more Normal than distribution A ę **Central Limit Theorem**
2. Both distributions centered on µ ę **x-bar is unbiased estimator of µ**
3. Distribution B is skinnier than distribution A ę related to **“square root law”**



Reiteration of Key Findings

- **Finding 1 (central limit theorem):** the sampling distribution of x-bar tends toward Normality even when the population is not Normal (esp. strong in large samples).
- **Finding 2 (unbiasedness):** the expected value of x-bar is μ
- **Finding 3** is related to the **square root law,** which says:

$$\pm_{\bar{x}} = \frac{\pm}{\sqrt{n}}$$

Standard Deviation of the Mean

- The standard deviation of the sampling distribution of the mean has a *special name*: **standard error of the mean** (denoted σ_{xbar} or SE_{xbar})
- The square root law says:

$$\pm_{\bar{x}} \equiv SE_{\bar{x}} = \frac{\pm}{\sqrt{n}}$$

Square Root Law
Example: σ = 15

For n = 1 $\Rightarrow SE_{\bar{x}} = \frac{\pm}{\sqrt{n}} = \frac{15}{\sqrt{1}} = 15$

For n = 4 $\Rightarrow SE_{\bar{x}} = \frac{\pm}{\sqrt{n}} = \frac{15}{\sqrt{4}} = 7.5$

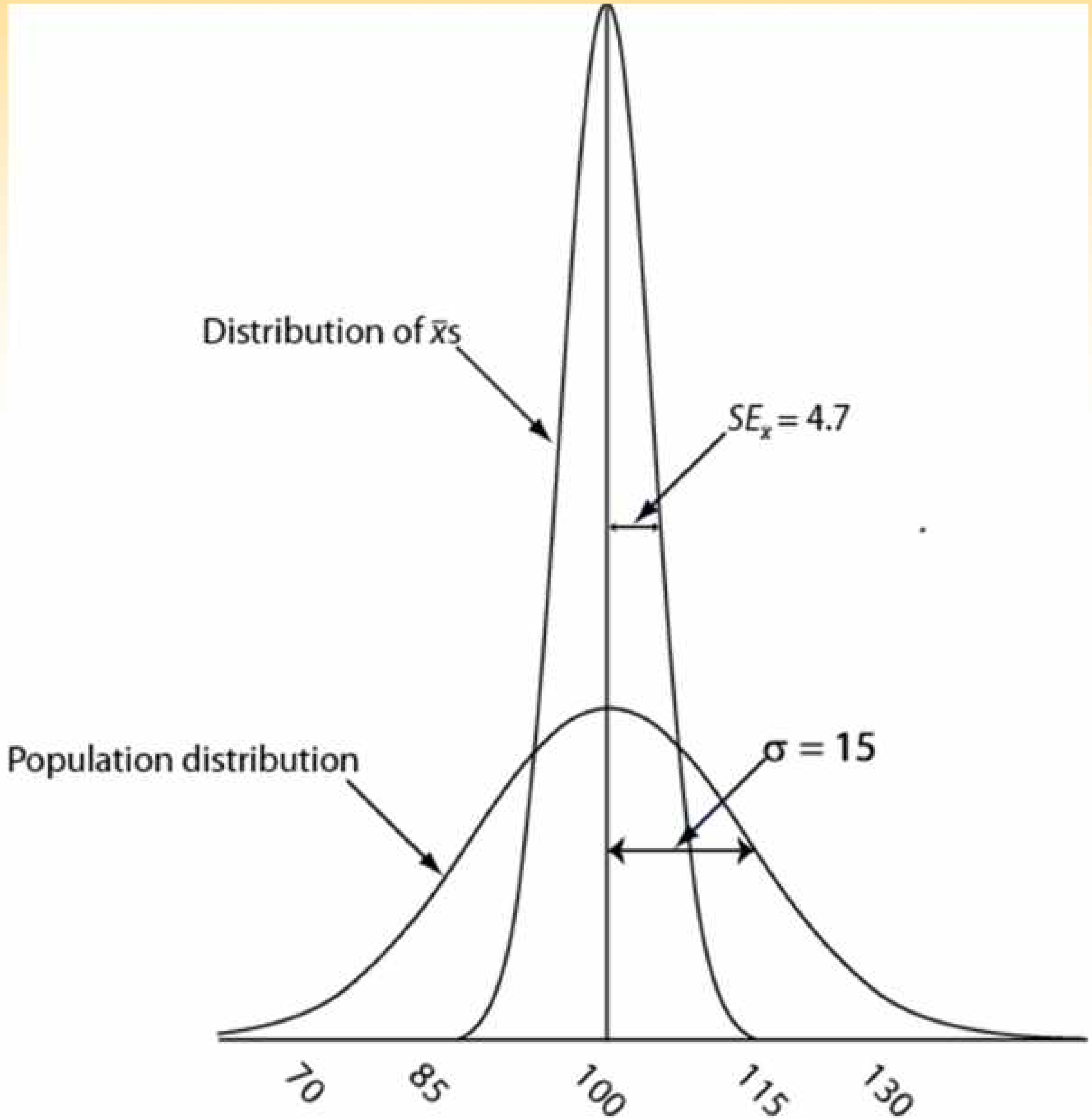
For n = 16 $\Rightarrow SE_{\bar{x}} = \frac{\pm}{\sqrt{n}} = \frac{15}{\sqrt{16}} = 3.75$

Quadrupling the sample size cuts *the standard error of* the mean in half

Putting it together: $\overline{x}$ ~ N(μ, SE)

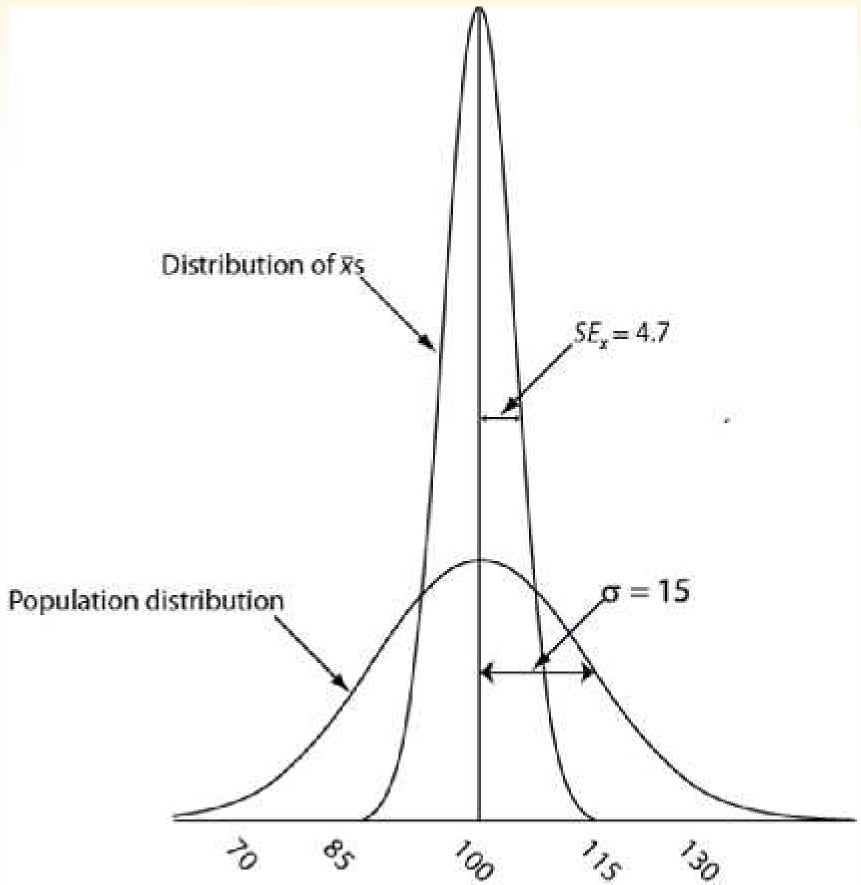
- The sampling distribution of x-bar tends to be Normal with mean μ and $\sigma_{xbar} = \sigma / \sqrt{n}$
- Example: Let X represent Weschler Adult Intelligence Scores; X ~ N(100, 15).
  - Take an SRS of $n = 10$
  - $\sigma_{xbar} = \sigma / \sqrt{n} = 15/\sqrt{10} = 4.7$
  - Thus, xbar ~ N(100, 4.7)

Individual WAIS (population) and mean WAIS when n = 10



68-95-99.7 rule applied to the SDM

- We've established xbar ~ N(100, 4.7). Therefore,
- 68% of x-bars within $\mu \pm \sigma_{xbar}$ $= 100 \pm 4.7$ $= 95.3$ to 104.7
- 95% of x-bars within $\mu \pm 2 \cdot \sigma_{xbar}$ $= 100 \pm (2 \cdot 4.7)$ $= 90.6$ to 109.4



Law of Large Numbers

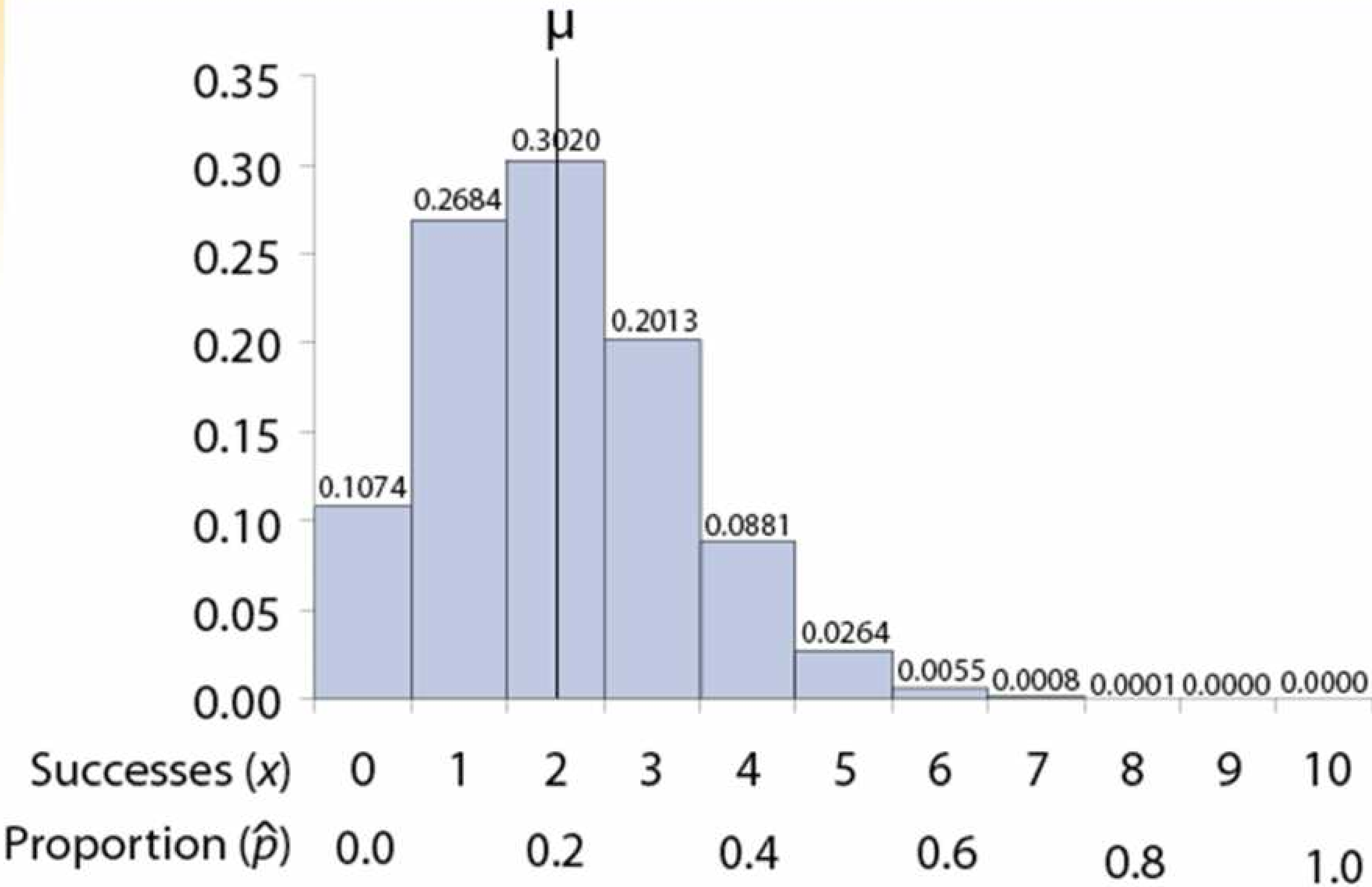
As a sample gets larger and larger, the x-bar approaches μ. Figure demonstrates results from an experiment done in a population with μ = 173.3

ones and Bartlett Publishers

8.3 Sampling Behavior of Counts and Proportions

- Recall Chapter: binomial random variable represents the random number of successes in *n* independent Bernoulli trials each with probability of success *p;* otation X~b(*n*,*p*)
- X~b(10,0.2) is shown on the next slide. Note that

  $\mu = 2$
- Reexpress the counts of success as proportion p-hat = *x* / *n.* For this re-expression, $\mu = 0.2$


μ
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.1074
0.2684
0.3020
0.2013
0.0881
0.0264
0.0055
0.0008
0.0001
0.0000
0.0000
Successes (x) 0 1 2 3 4 5 6 7 8 9 10
Proportion ($\hat{p}$) 0.0 0.2 0.4 0.6 0.8 1.0

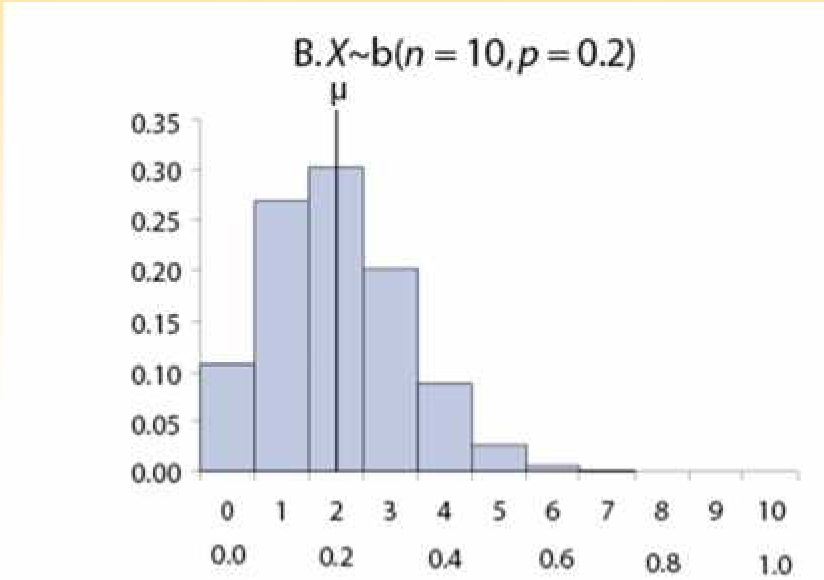
Normal Approximation to the Binomial ("*npq* rule")

- When *n* is large, the binomial distribution approximates a Normal distribution ("the Normal Approximation")
- How large does the sample have to be to apply the Normal approximation? $\Rightarrow$ One rule says that the Normal approximation applies when $npq \geq 5$

Top figure:

X~b(10,0.2)

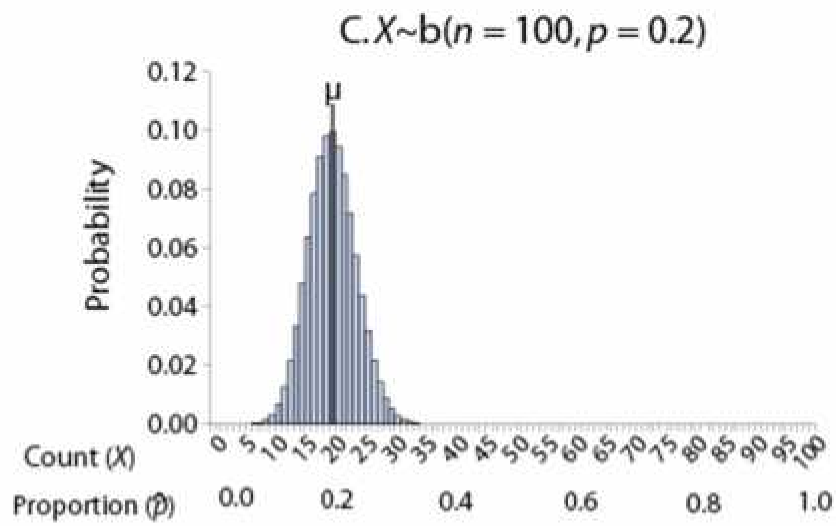
$npq = 10 \cdot 0.2 \cdot (1–0.2)$
$= 1.6 \Rightarrow$ Normal approximation does ***not*** apply



Bottom figure:

X~b(100,0.2)

$npq = 100 \cdot 0.2 \cdot (1-0.2)$
$= 16 \Rightarrow$ Normal approximation applies



Normal Approximation for a Binomial Count

$$\text{«} = np \text{ and } \pm = \sqrt{npq}$$

When Normal approximation applies:

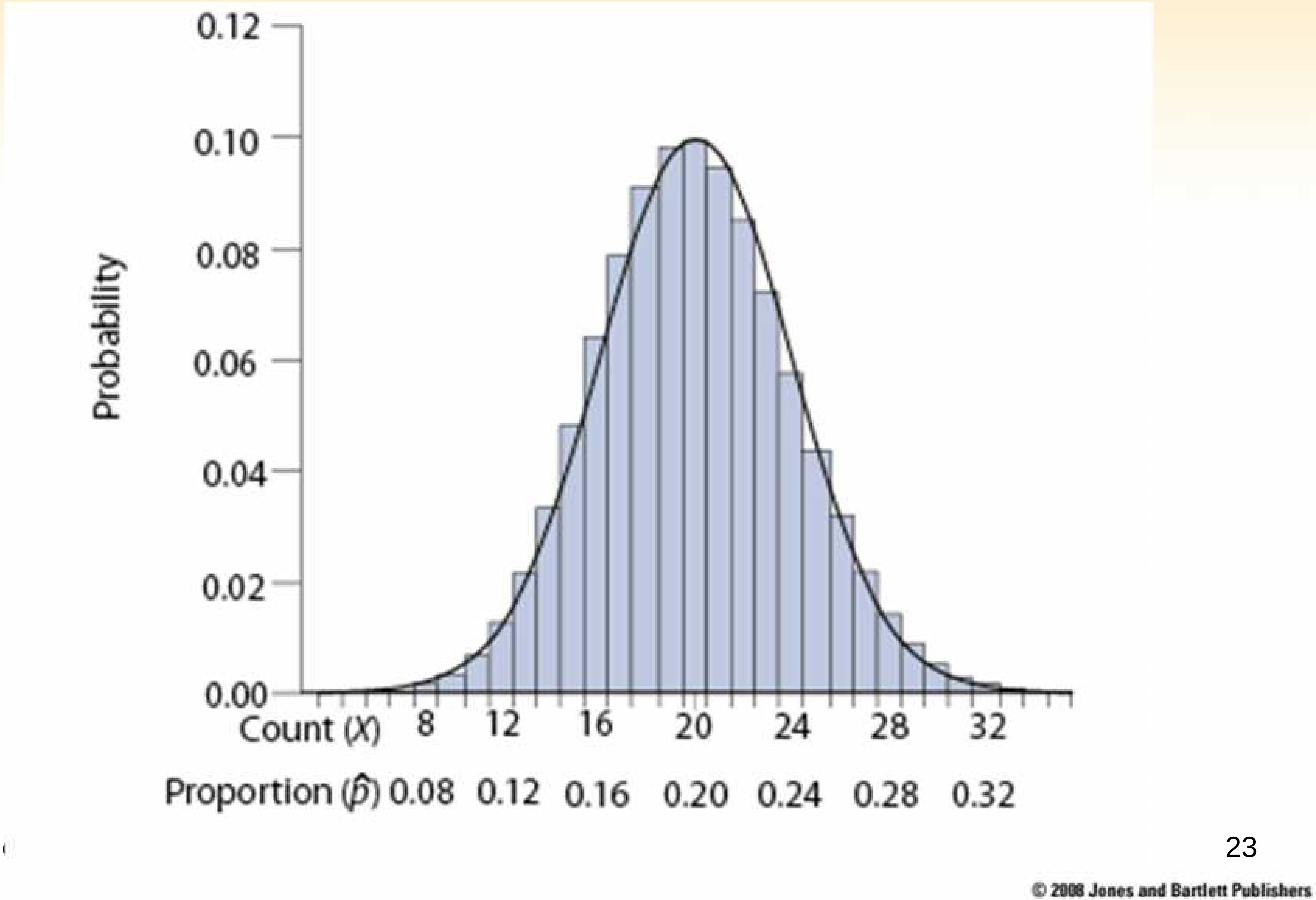
$$X \sim N\left(np, \sqrt{npq}\right)$$

Normal Approximation for a Binomial Proportion

$$\text{«} = p \text{ and} \pm = \sqrt{\frac{pq}{n}}$$

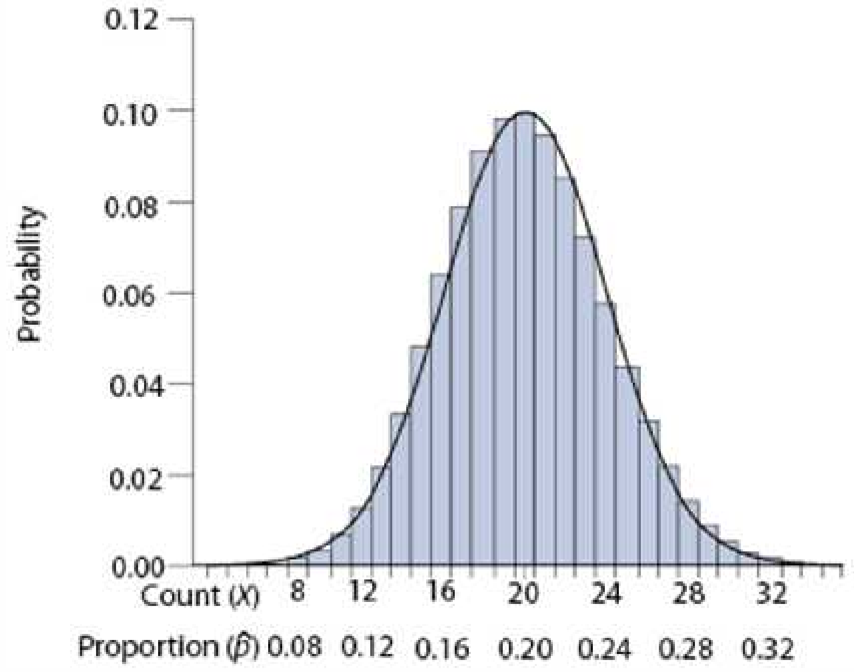
$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

“p-hat” represents the sample proportion



Illustrative Example: Normal Approximation to the Binomial

- Suppose the prevalence of a risk factor in a population is 20%
- Take an SRS of n = 100 from population
- A variable number of cases in a sample will follow a binomial distribution with n = 20 and p = .2



Illustrative Example, cont.

The Normal approximation for the **count** is:

$$\text{«} = np = 100 \cdot .2 = 20$$

$$\text{and} \pm = \sqrt{npq} = \sqrt{100 \cdot .2 \cdot .8} = 4$$

$$X \sim N(20{,}4)$$

The Normal approximation for the **proportion** is:

$$\text{«} = p = .2$$

$$\pm = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.2 \cdot .8}{100}} = 0.04$$

$$\hat{p} \sim N(0.2{,}0.04)$$

Illustrative Example, cont.

1. Statement of problem: Recall $X \sim N(20, 4)$ Suppose we observe 30 cases in a sample. What is the probability of observing at least 30 cases under these circumstance, i.e., $\Pr(X \geq 30) = ?$

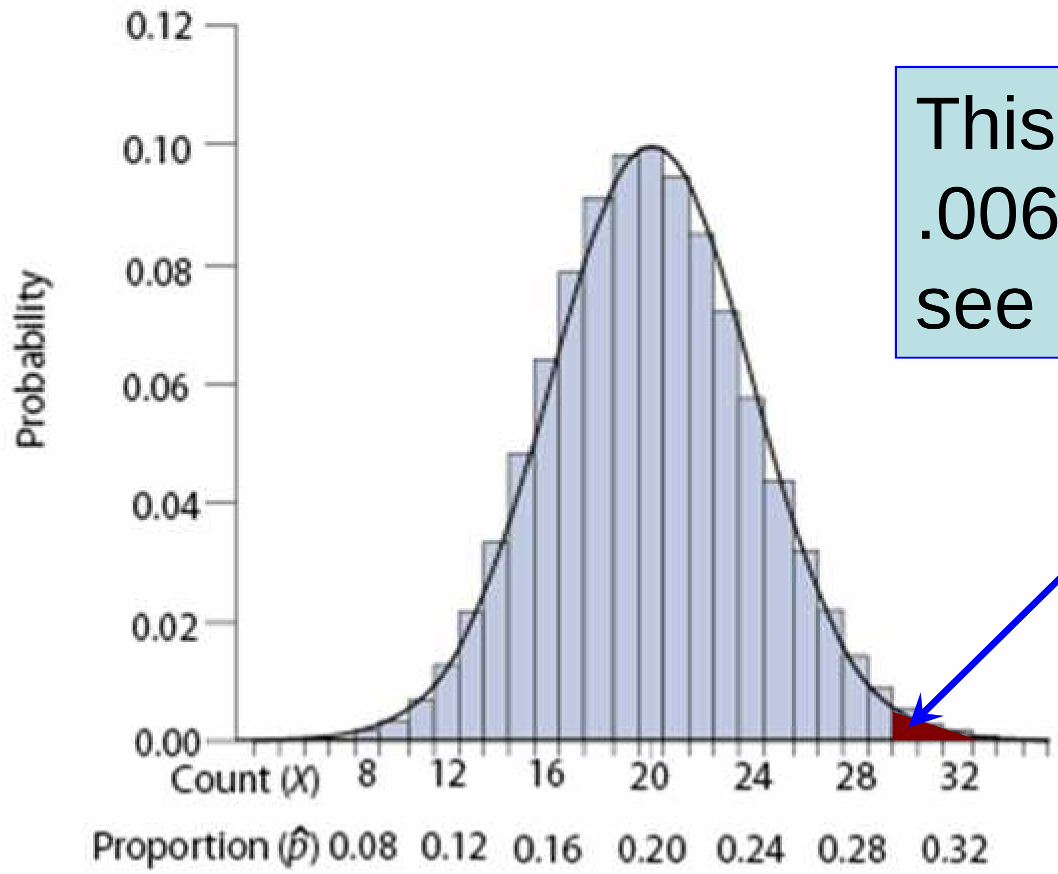
2. Standardize: $z = (30 - 20) / 4 = 2.5$

3. Sketch: next slide

4. Table B: $\Pr(Z \geq 2.5) = 0.0062$

Illustrative Example, cont.

Binomial and superimposed Normal distributions



This model suggests .0062 of samples will see 30 or more cases.