Lessons 3 and 4: Statistics

- **Working with Stata**

1. **Probability Example: CLT in action**
2. **Populations have parameters, Samples have estimators**
3. **Estimators & Estimates**
4. **Parameters have distributions: From Probability to Statistics**
5. **Confidence Intervals for Parameters**
   E.g. Pick a digit
6. **Hypothesis Testing: terms**
7. **Hypothesis Testing: steps**
8. **P-value**
9. **Properties of Estimators**
10. **Efficiency of Sample Mean**
11. **Monte Carlo Demonstration of CLT**

1. Probability Example: Confidence Interval for sample mean

- A_i - the outcome of some event for individual i
  E.g. p(lefthanded)=.085
- How often should that happen in a class of this size?
- frequency of "~~4~~"s (lefties) = N x sample mean of "~~4~~"s (lefties)
- Standard error (std. deviation of mean)
  - $V(A_i) = p(A_i=1)[1-p(A_i=1)]$
- Normal approximation
- Confidence interval for sample mean
- Application: Fair bet?

Copyright © 2008 by Pearson Education, Inc.

$$A_i = \begin{cases} 1 & i \text{ is lefty} \\ 0 & \text{otherwise} \end{cases}$$

$\bar{A}_n$ = proportion lefties in sample of size n $= \dfrac{\sum_{i=1}^{n} A_i}{n}$; # of lefties $= n \times \bar{A}_n$

For C.L.T we need:

- some kind of average ✓
- large sample ✓
- randomly sampled data ✓ (I owe Victor)
- mean $= 8.5\% = \mu = p$

$E(A_i) = 0 \times (1-p) + 1 \times p = p$

$E(\bar{A}_n) = E\left(\dfrac{\sum A_i}{n}\right) = \dfrac{1}{n} E(\sum A_i) = \dfrac{1}{n} n E(A_i) = p$

- variance

$V(A_i) = E[A_i - E(A_i)]^2$

$= [0-p]^2(1-p) + [1-p]^2 p$

$= p^2(1-p) + (1-p)^2 p = p(1-p)[p + (1-p)] = p(1-p)$

$V(\bar{A}_n) = V\left(\dfrac{\sum A_i}{n}\right) = \dfrac{1}{n^2} V(\sum A_i) = \dfrac{1}{n^2} \sum V(A_i) = \dfrac{n\,p(1-p)}{n^2} = \dfrac{p(1-p)}{n}$

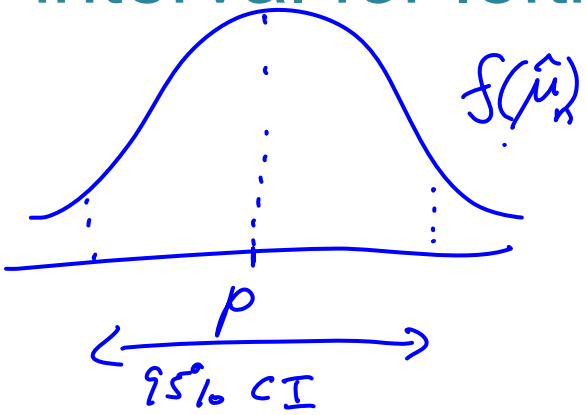
Copyright © 2003 by Pearson Education, Inc.

Calculating Confidence interval for lefties

$\bar{A}_n \overset{A}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$

$\vdots$

$\hat{\mu}$



$p = 0.085$

$s.d.(\hat{\mu}) \equiv s.d.(\bar{A}_n)$

$= \sqrt{V(\bar{A}_n)}$

$= \sqrt{\frac{p(1-p)}{n}}$

$= \sqrt{\frac{0.085(0.915)}{119}}$

$= 0.0256$

$P(p - 1.96\, s.d.(\hat{\mu}) \le \hat{\mu} \le p + 1.96\, s.d.(\hat{\mu})) =$

$P(.085 - 1.96(.0256) \le \hat{\mu} \le .085 + 1.96(.0256)) = .95$

$P(.085 - .050 \le \hat{\mu} \le .085 + .050) = 0.95$

$P(.035 \le \hat{\mu} \le .135)$

$P(4.165 \le \text{lefties} \le 16.065) = 0.95$

# of lefties = 12 in Winter 08 120B

$\bar{A}_n = 0.101 = \hat{\mu} = \hat{p}$

Copyright © 2003 by Pearson Education, Inc.

2. Populations have parameters, Samples have estimators

Population	Sample
μ, σ, β "Greek to me"	$\hat{\mu}, \hat{\sigma}, \hat{\beta}$ or m, s, b
e.g. U.S. Residents	*C.P.S.*
All coin flips	*N coin flips*
Parameters (typically Greek)	Estimators
e.g. Pop. mean	*Sample mean*
Pop. Variance	*Sample variance*

Copyright © 2003 by Pearson Education, Inc.

3. Estimators and Estimates

Estimators and Estimates

An **estimator** is a function of a sample of data to be drawn randomly from a population. An **estimate** is the numerical value of the estimator when it is actually computed using data from a specific sample. An estimator is a random variable because of randomness in selecting the sample, while an estimate is a nonrandom number.

e.g. $\bar{A}_n$ was an estimator of p

10.1% " " estimate of p

Copyright © 2003 by Pearson Education, Inc.

4. Parameters have distributions: From Probability to Statistics

- Probability: use information from populations to learn about samples
  (Lefthanded example)
- Statistics: use information from samples to learn about populations
- How to make the transition

Copyright © 2003 by Pearson Education, Inc.

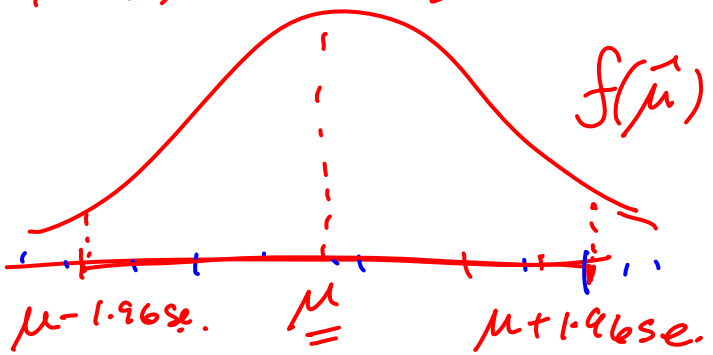
From Probability to Statistics…

CI for $\hat{\mu}$, the sample mean, $\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}_n$

$E(\bar{x}_n) = \mu$, $V(\bar{x}_n)$

$$t = \frac{\bar{x}_n - \mu}{\sqrt{V(\bar{x}_n)}} \overset{A}{\sim} N(0,1) \text{ when } n \to \infty$$

s.e.



$\Rightarrow$

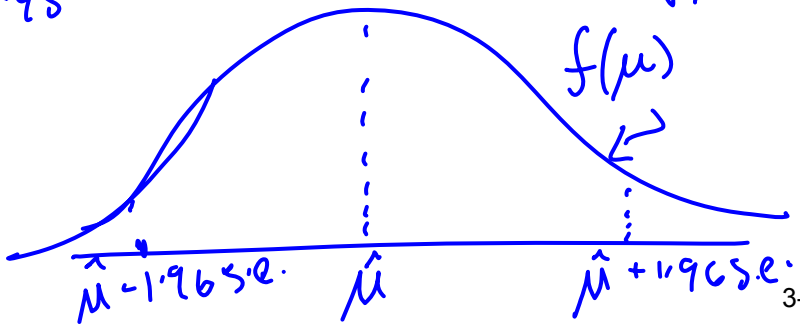
$P(\mu - 1.96\,s.e. \le \hat{\mu} \le \mu + 1.96\,s.e.) = 0.95$

$P(-1.96\,s.e. \le \hat{\mu} - \mu \le +1.96\,s.e.) = 0.95$

$P\left(-1.96 \le \frac{\hat{\mu} - \mu}{s.e.} \le 1.96\right) = 0.95$

$P(\hat{\mu} + 1.96\,s.e. \ge \mu \ge \hat{\mu} - 1.96\,s.e.) = 0.95$

we don't know μ, we do know $\hat{\mu}$ so it's useful to think of μ as a R.V.



Copyright © 2003 by Pearson Education, Inc.

5. Confidence Intervals for parameters: Pick a number example

- Find the distribution of estimator
- Interpret as distribution of parameter

$P(\hat{\mu} - 1.96 s.e. \le \mu \le \hat{\mu} + 1.96 s.e.) = 0.95$ $\quad \hat{\mu} = \frac{11}{120} = .0917$

$P(.0917 - 1.96(.0263) \le \mu \le .0917 + 1.96(.0263)) = 0.95$ $\quad s.e. = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{.0917(1-.0917)}{120}}$

$P(\ .040 \le \mu \le .143\ )$

Assume $N = 120$ $\quad = .0263$

Copyright © 2003 by Pearson Education, Inc.

Confidence Intervals

Confidence Intervals for the Population Mean

A 95% two-sided confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of its applications. When the sample size n is large, 95%, 90%, and 99% confidence intervals for μ_Y are:

$$95\% \text{ confidence interval for } \mu_Y = \{\overline{Y} \pm 1.96SE(\overline{Y})\}.$$
$$90\% \text{ confidence interval for } \mu_Y = \{\overline{Y} \pm 1.64SE(\overline{Y})\}.$$
$$99\% \text{ confidence interval for } \mu_Y = \{\overline{Y} \pm 2.57SE(\overline{Y})\}.$$

Copyright © 2003 by Pearson Education, Inc.

6. Hypothesis testing: terms

The Terminology of Hypothesis Testing

The prespecified rejection probability of a statistical hypothesis test under the null hypothesis is the **significance level** of the test. The **critical value** of the test statistic is the value of the statistic for which the test just rejects the null hypothesis at the given significance level. The set of values of the test statistic for which the test rejects the null is the **rejection region**, and the values of the test statistic for which it does not reject the null is the **acceptance region**. The probability that the test actually incorrectly rejects the null hypothesis when the null is true is the **size** of the test, and the probability that the test correctly rejects the null when the alternative is true is the **power** of the test.

The *p*-value is the probability of obtaining a test statistic, by random sampling variation, at least as adverse to the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct. Equivalently, the *p*-value is the smallest significance level at which you can reject the null hypothesis.

Copyright © 2003 by Pearson Education, Inc.

7. Hypothesis Testing: Steps

Testing the Hypothesis $E(Y) = \mu_{Y,0}$ Against the Alternative $E(Y) \neq \mu_{Y,0}$

1. Compute the standard error of $\overline{Y}$, $SE(\overline{Y})$ (Equation (3.14)).
2. Compute the t-statistic (Equation (3.10)).
3. Compute the p-value (Equation (3.13)). Reject the hypothesis at the 5% significance level if the p-value is less than .05 (equivalently, if $|t^{act}| > 1.96$).

DIFFERENCES FROM CI CALCULATION:

A) $f(\hat{\mu})$ rather than $f(\mu)$ [Exercise in Probability]

B) Calculate std. error (in this binary case) using μ from H_0.

e.g. $se. = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.1)(1-0.1)}{n}}$.

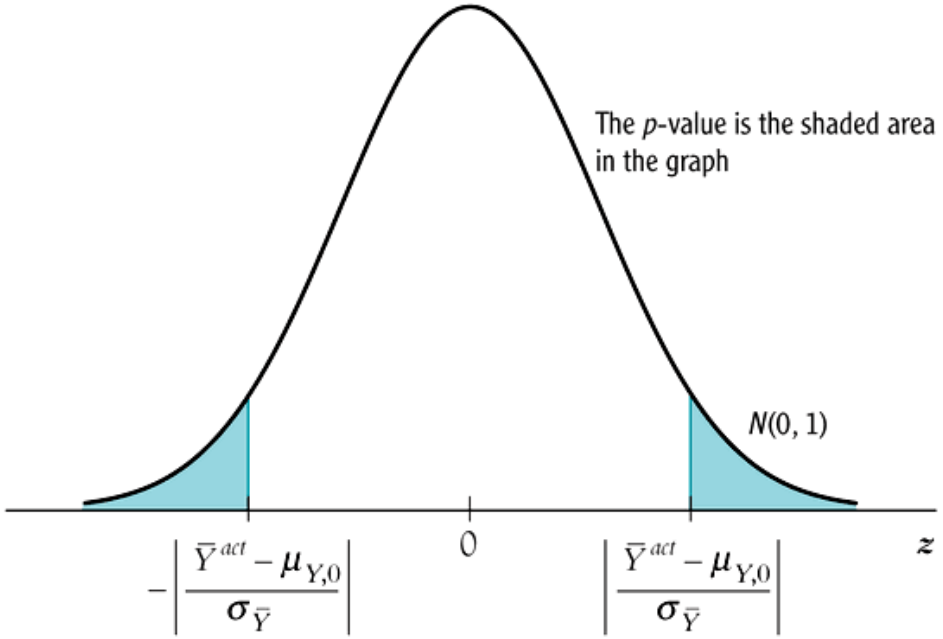
Copyright © 2003 by Pearson Education, Inc.

8. P-value:

How likely were we to miss by at least that much, if Ho is true?

FIGURE 3.1 Calculating a *p*-value

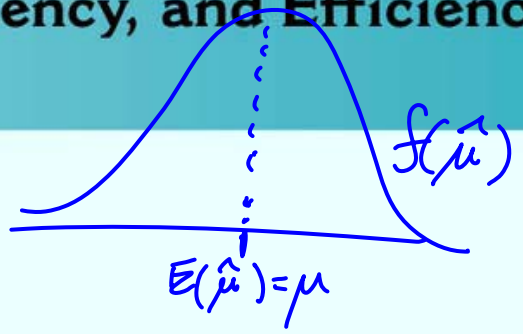
The *p*-value is the probability of drawing a value of $\overline{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\overline{Y}^{act}$. In large samples, $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$ under the null hypothesis, so $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ is distributed $N(0, 1)$. Thus the *p*-value is the shaded standard normal tail probability outside $\pm|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.



Copyright © 2003 by Pearson Education, Inc.

9. Properties of Estimators

Bias, Consistency, and Efficiency



Let $\hat{\mu}_Y$ be an estimator of μ_Y. Then:

- The **bias** of $\hat{\mu}_Y$ is $E(\hat{\mu}_Y) - \mu_Y$.
- $\hat{\mu}_Y$ is an **unbiased estimator** of μ_Y if $E(\hat{\mu}_Y) = \mu_Y$.
- $\hat{\mu}_Y$ is a **consistent estimator** of μ_Y if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$.
- Let $\tilde{\mu}_Y$ be another estimator of μ_Y, and suppose that both $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ are unbiased. Then $\hat{\mu}_Y$ is said to be more **efficient** than $\tilde{\mu}_Y$ if $\text{var}(\hat{\mu}_Y) < \text{var}(\tilde{\mu}_Y)$.

If we choose $\hat{\mu} = \bar{Y} = \frac{\sum Y}{N}$ then $E(\hat{\mu}) = E(\bar{Y}) = \mu$ UNBIASED

Copyright © 2003 by Pearson Education, Inc.

10 Efficiency of Sample Mean

Efficiency of $\overline{Y}$

Let $\hat{\mu}_Y$ be an estimator of μ_Y that is a weighted average of $Y_1, \ldots, Y_n$, that is, $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are nonrandom constants. If $\hat{\mu}_Y$ is unbiased, then $\text{var}(\overline{Y}) < \text{var}(\hat{\mu}_Y)$ unless $\hat{\mu}_Y = \overline{Y}$. That is, $\overline{Y}$ is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$.

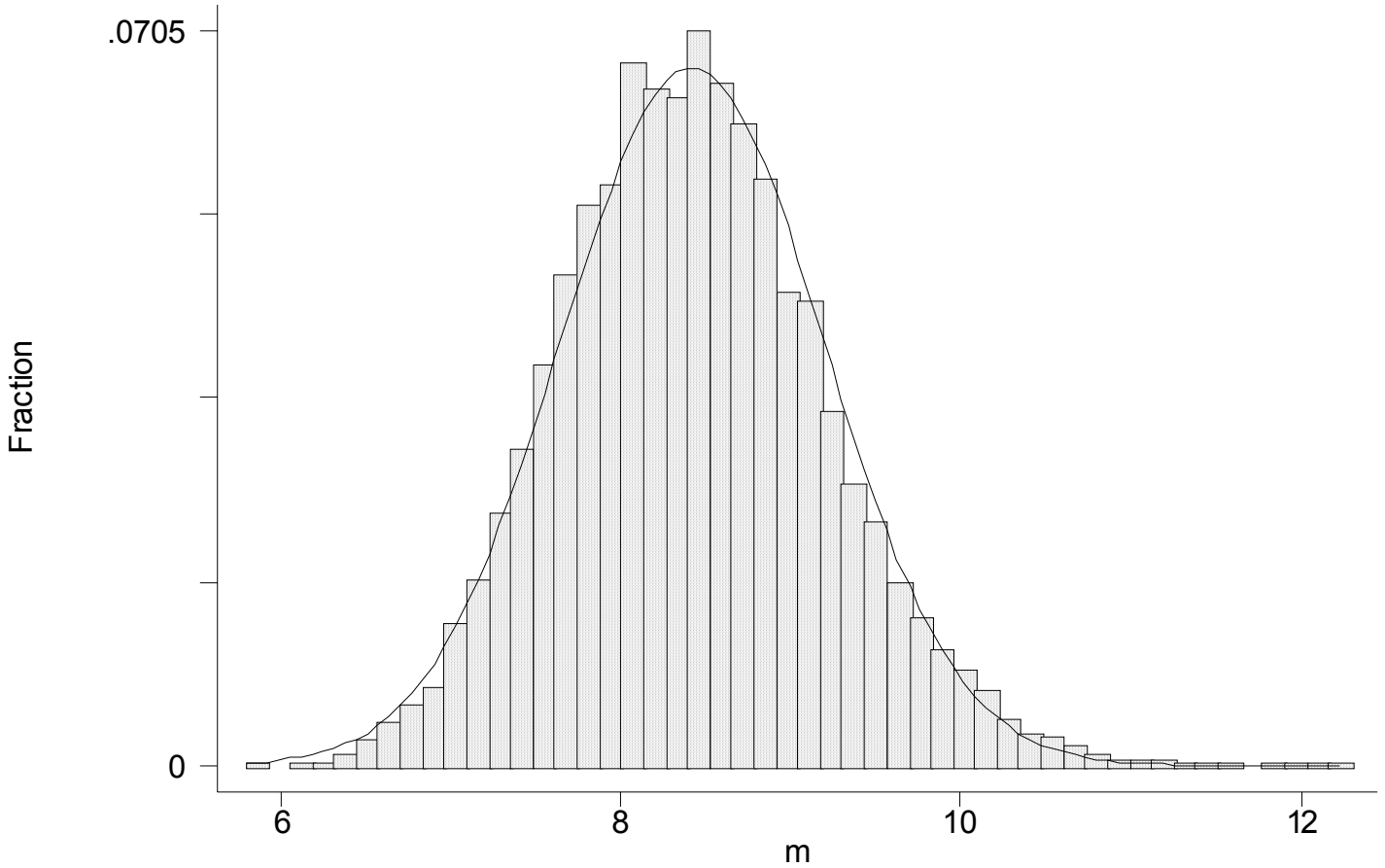
Copyright © 2003 by Pearson Education, Inc.

11. Monte Carlo Demonstration of CLT

- Imagine estimating the mean hourly wage μ, by drawing samples of size N from the distribution of hourly wages, say in 1984.
- Stata will do this for us, maybe 10,000 times.

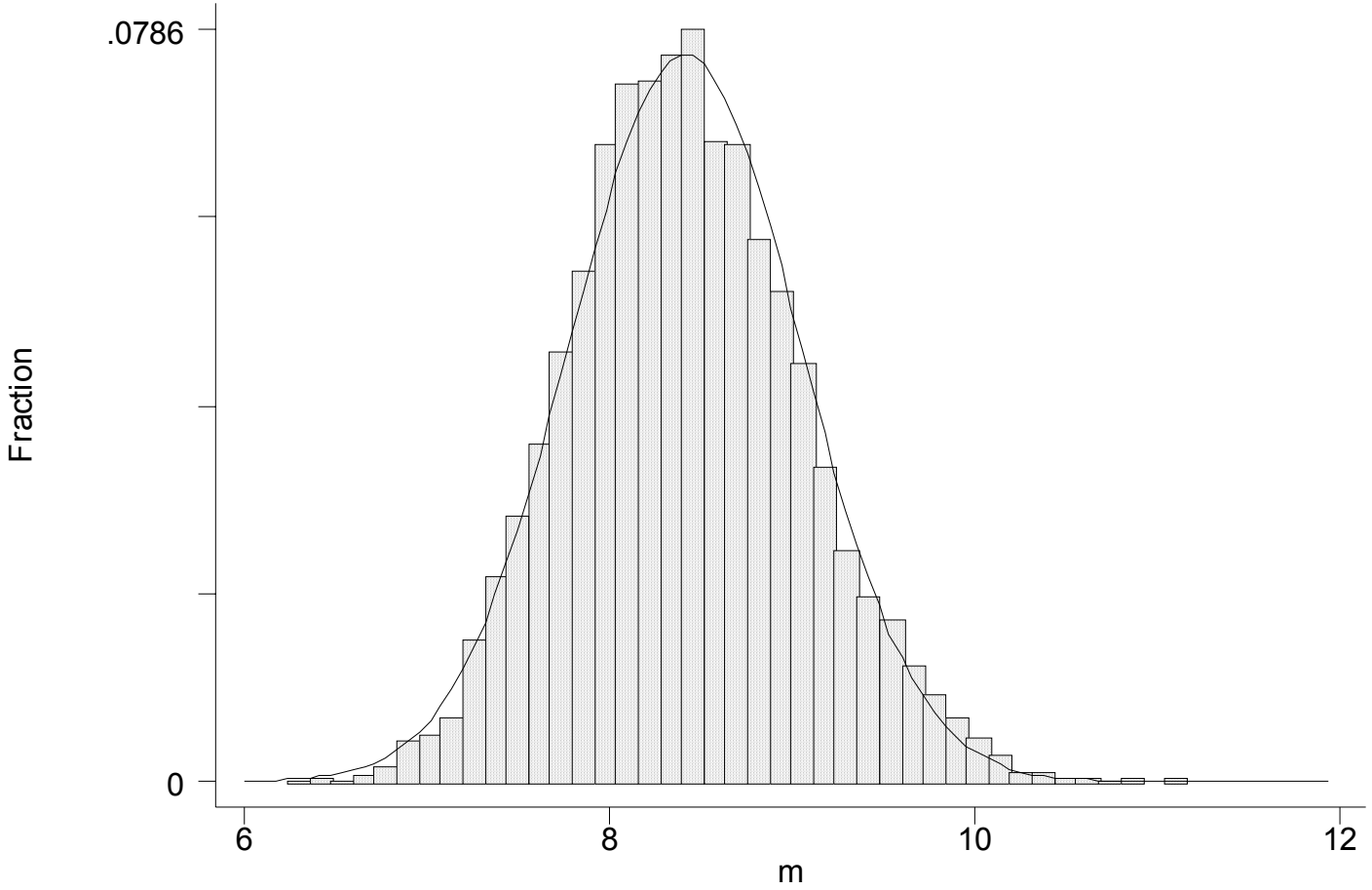
Copyright © 2003 by Pearson Education, Inc.

N=40, 10,000 draws



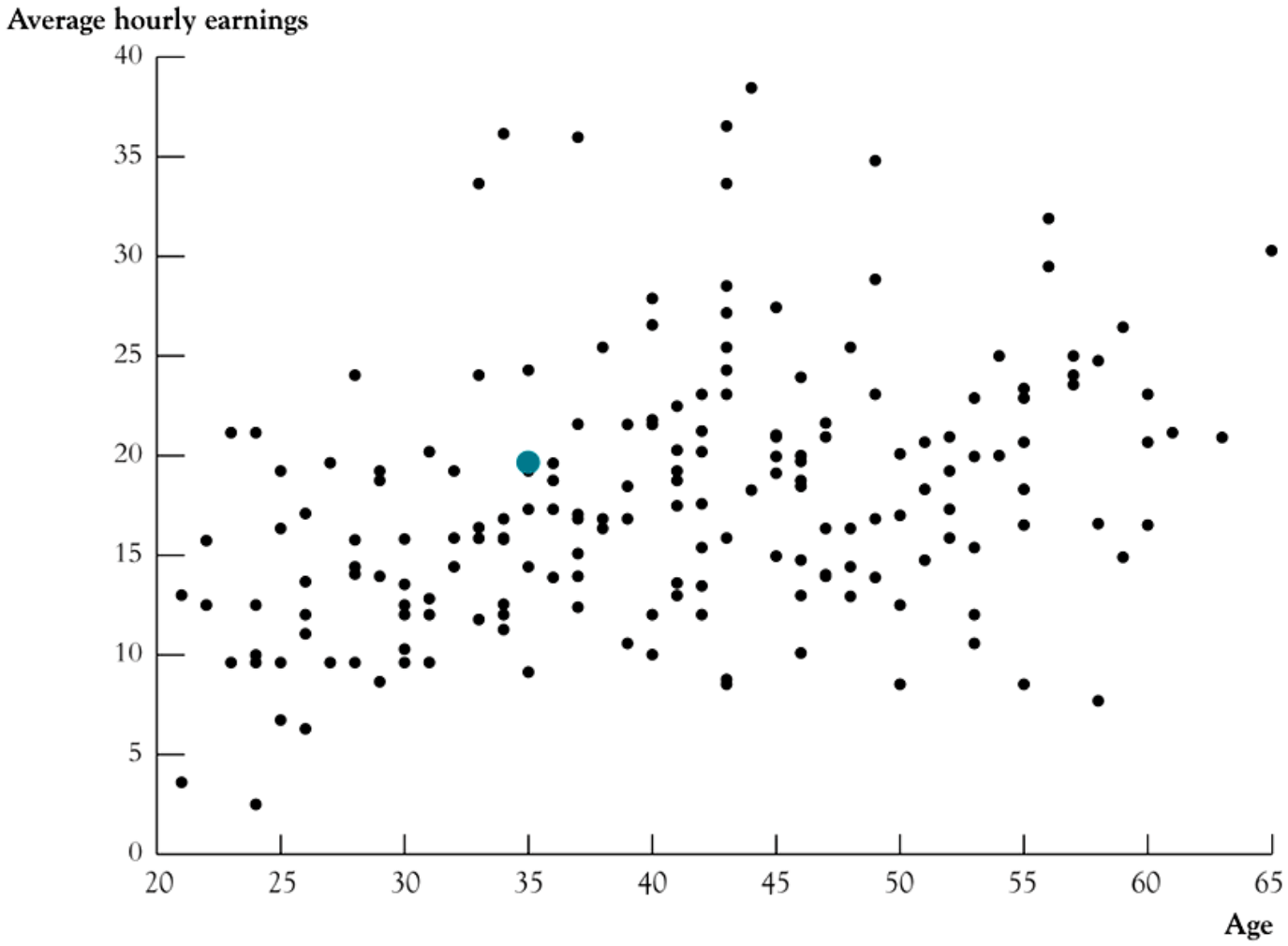
Copyright © 2003 by Pearson Education, Inc.

N=60, 10,000 draws



Copyright © 2003 by Pearson Education, Inc.

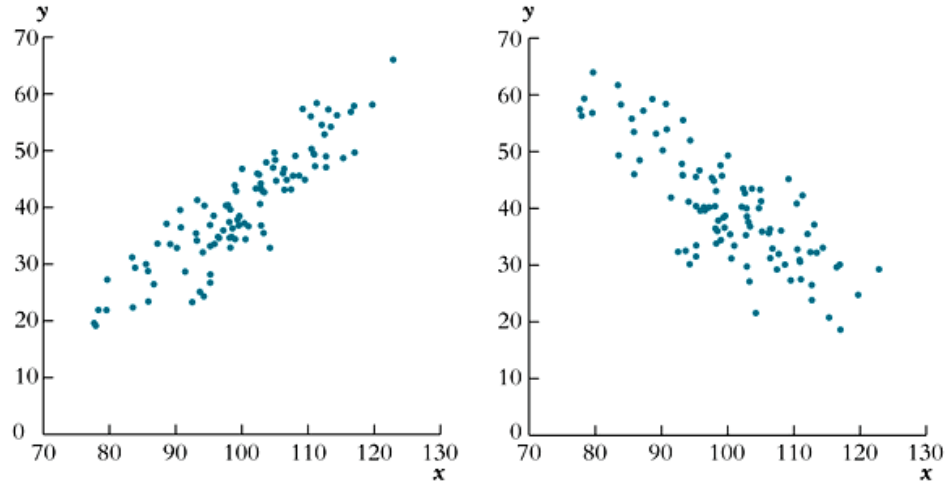
FIGURE 3.2 Scatterplot of Average Hourly Earnings vs. Age



Each point in the plot represents the age and average earnings of one of the 184 workers in the sample. The colored dot corresponds to a 35-year-old worker who earns $19.61 per hour. The data are for technicians in the communications industry without college degrees from the March 1999 CPS.

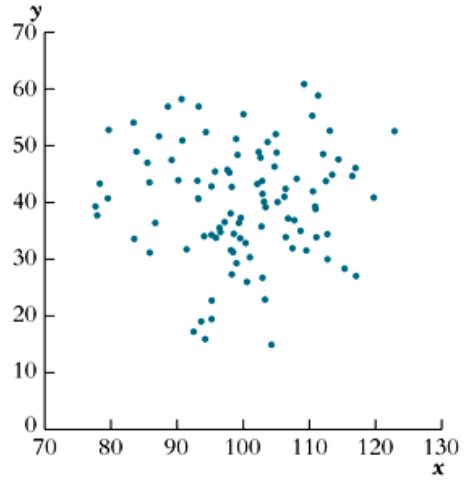
Copyright © 2003 by Pearson Education, Inc.

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets

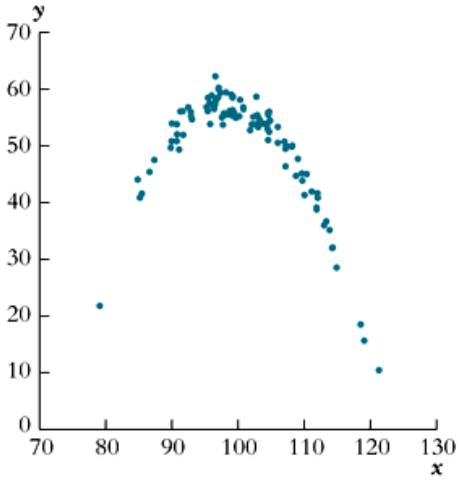


(a) Correlation = +0.9

(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between *X* and *Y*. In Figure 3.3c, *X* is independent of *Y* and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.

The next two slides each present one half of Figure 3.3.

Copyright © 2003 by Pearson Education, Inc.

TABLE 3.1 Hourly Earnings in the United States of Working College Graduates, Aged 25–34: Selected Statistics from the Current Population Survey, in 1998 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\bar{Y}_m$	s_m	n_m	$\bar{Y}_w$	s_w	n_w	$\bar{Y}_m - \bar{Y}_w$	$SE(\bar{Y}_m - \bar{Y}_w)$	95% Confidence Interval for d
1992	17.57	7.50	1591	15.22	5.97	1371	2.35**	0.25	1.87–2.84
1994	16.93	7.39	1598	15.01	6.41	1358	1.92**	0.25	1.42–2.42
1996	16.88	7.29	1374	14.42	6.07	1235	2.46**	0.26	1.94–2.97
1998	17.94	7.86	1393	15.49	6.80	1210	2.45**	0.29	1.89–3.02

These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

Copyright © 2003 by Pearson Education, Inc.

Summary - Lessons 3 and 4: Statistics

1. **Probability Example: CLT in action**
2. **Populations have parameters, Samples have estimators**
3. **Estimators & Estimates**
4. **Parameters have distributions: From Probability to Statistics**
5. **Confidence Intervals for Parameters**
   E.g. Pick a digit
6. **Hypothesis Testing: terms**
7. **Hypothesis Testing: steps**
8. **P-value**
9. **Properties of Estimators**
10. **Efficiency of Sample Mean**
11. **Monte Carlo Demonstration of CLT**

Appendix 3.1

The U.S. Current Population Survey

Each month the Bureau of Labor Statistics in the U.S. Department of Labor conducts the "Current Population Survey" (CPS), which provides data on labor force characteristics of the population, including the level of employment, unemployment, and earnings. Approximately 65,000 U.S. households are surveyed each month. The sample is chosen by randomly selecting addresses from a database comprised of addresses from the most recent decennial census augmented with data on new housing units constructed after the last census. The exact random sampling scheme is rather complicated (first small geographical areas are randomly selected, then housing units within these areas are randomly selected); details can be found in the *Handbook of Labor Statistics* and on the Bureau of Labor Statistics website (www.bls.gov).

The survey conducted each March is more detailed than in other months and asks questions about earnings during the previous year. The statistics in Table 3.1 were computed using the March surveys. The CPS earnings data are for full-time workers, defined to be somebody employed more than 35 hours per week for at least 48 weeks in the previous year.

Copyright © 2003 by Pearson Education, Inc.