Some terminology for testing statistical hypotheses:

p-value = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

The ***significance level*** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on $\bar{Y}$:

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom)

Calculating the p-value, ctd.

- To compute the *p*-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if n is small.
- If n is large, you can use the normal approximation (CLT):

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

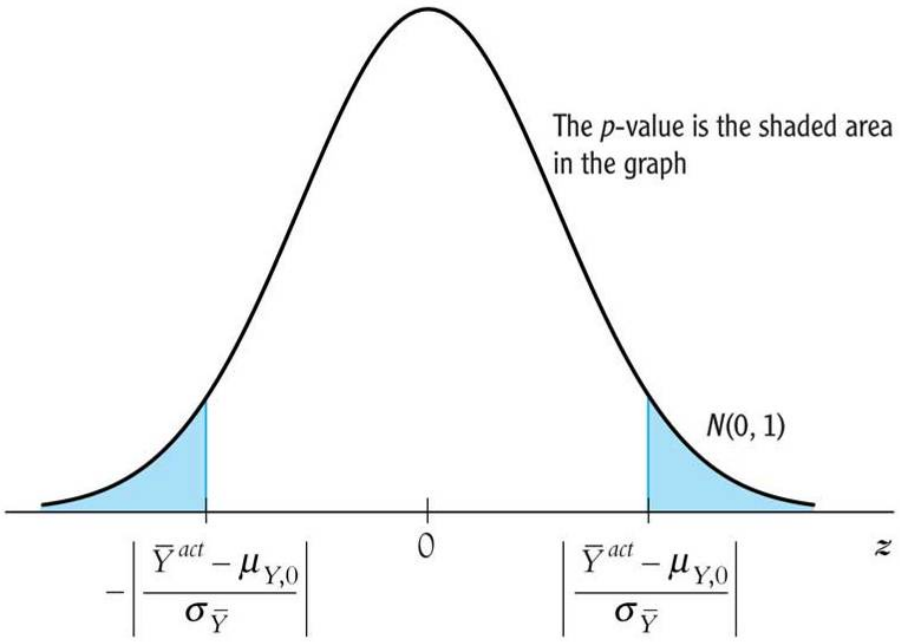
$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$\cong \text{probability under left+right } N(0,1) \text{ tails}$$

where $\sigma_{\bar{Y}}$ = std. dev. of the distribution of $\bar{Y} = \sigma_Y/\sqrt{n}$.

Calculating the p-value with σ_Y known:



- For large n, p-value = the probability that a $N(0,1)$ random variable falls outside $|(\bar{Y}^{act}-\mu_{Y,0})/\sigma_{\bar{Y}}|$
- In practice, $\sigma_{\bar{Y}}$ is unknown – it must be estimated

Estimator of the variance of Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“sample variance of } Y\text{”}$$

Fact:

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $E(Y^4) < \infty$, then $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Why does the law of large numbers apply?

- Because s_Y^2 is a sample average; see Appendix 3.3
- Technical note: we assume $E(Y^4) < \infty$ because here the average is not of Y_i, but of its square; see App. 3.3

The Gender Gap of Earnings of College Graduates in the United States

The box in Chapter 2, "The Distribution of Earnings in the United States in 2008," shows that, on average, male college graduates earn more than female college graduates. What are the recent trends in this "gender gap" in earnings? Social norms and laws governing gender discrimination in the workplace have changed substantially in the United States. Is the gender gap in earnings of college graduates stable, or has it diminished over time?

Table 3.1 gives estimates of hourly earnings for college-educated full-time workers aged 25–34 in the United States in 1992, 1996, 2000, 2004, and 2008, using data collected by the Current Population Survey. Earnings for 1992, 1996, 2000, and 2004 were adjusted for inflation by putting them in 2008 dollars using the Consumer Price Index (CPI).[1] In 2008, the average hourly earnings of the 1838 men surveyed was \$24.98, and the standard deviation of earnings for men was \$11.78. The average hourly earnings in 2008 of the 1871 women surveyed was \$20.87, and the standard deviation of earnings was \$9.66. Thus the estimate of the gender gap in earnings for 2008 is \$4.11 (= \$24.98 − \$20.87), with a standard error of \$0.35 ($= \sqrt{11.78^2/1838 + 9.66^2/1871}$). The 95% confidence interval for the gender gap in earnings in 2008 is $4.11 \pm 1.96 \times 0.35 = (\$3.41, \$4.80)$.

The results in Table 3.1 suggest four conclusions. First, the gender gap is large. An hourly gap of \$4.11 might not sound like much, but over a year it adds up to \$8220, assuming a 40-hour work week and 50 paid weeks per year. Second, from 1992 to 2008, the estimated gender gap increased by \$0.87 per hour in real terms, from \$3.22 per hour to \$4.11 per hour; however, this increase is not statistically significant

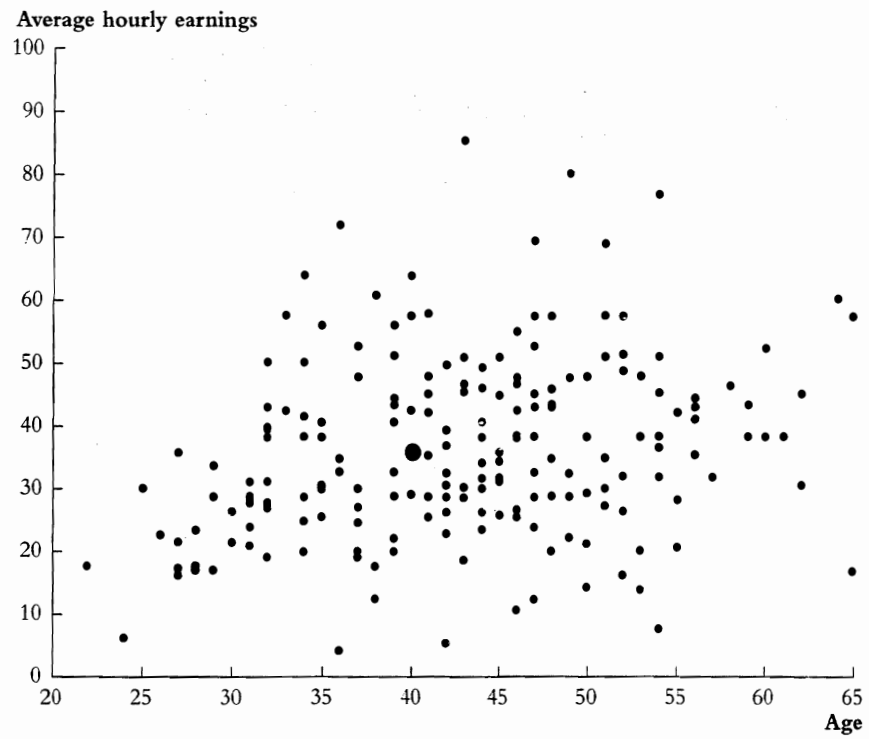
TABLE 3.1 Trends in Hourly Earnings in the United States of Working College Graduates, Ages 25–34, 1992 to 2008, in 2008 Dollars

	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	23.27	10.17	1594	20.05	7.87	1368	3.22**	0.33	2.58–3.88
1996	22.48	10.10	1379	18.98	7.95	1230	3.50**	0.35	2.80–4.19
2000	24.88	11.60	1303	20.74	9.36	1181	4.14**	0.42	3.32–4.97
2004	25.12	12.01	1894	21.02	9.36	1735	4.10**	0.36	3.40–4.80
2008	24.98	11.78	1838	20.87	9.66	1871	4.11**	0.35	3.41–4.80

These estimates are computed using data on all full-time workers aged 25–34 surveyed in the Current Population Survey conducted in March of the next year (for example, the data for 2008 were collected in March 2009). The difference is significantly different from zero at the **1% significance level.

continued

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 200 workers in the sample. The highlighted dot corresponds to a 40-year-old worker who earns $35.78 per hour. The data are for computer and information systems managers from the March 2009 CPS.

the respective population means. When n is large, it makes little difference whether division is by n or $n - 1$.

The **sample correlation coefficient**, or **sample correlation**, is denoted r_{XY} and is the ratio of the sample covariance to the sample standard deviations:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \tag{3.25}$$

The sample correlation measures the strength of the linear association between X and Y in a sample of n observations. Like the population correlation, the sample correlation is unitless and lies between -1 and 1: $|r_{XY}| \leq 1$.

The sample correlation equals 1 if $X_i = Y_i$ for all i and equals -1 if $X_i = -Y_i$ for all i. More generally, the correlation is ± 1 if the scatterplot is a straight line. If the