Review of Statistics

Ming-Ching Luoh

2005.9.15

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

$$s_{XY} \xrightarrow{p} \sigma_{XY}$$

- One natural way to estimate the population mean, μ_Y, is simply to compute the sample average $\bar{Y}$ from a sample of n i.i.d. observations. This can also be motivated by law of large numbers.
- But, $\bar{Y}$ is not the only way to estimate μ_Y. For example, Y_1 can be another **estimator** of μ_Y.
- In general, we want an estimator that gets as close as possible to the unknown true value, at least in some average sense. In other words, we want the sampling distribution of an estimator to be as **tightly** centered around the unknown value as possible.
- This leads to three specific desirable characteristics of an estimator.

Three desirable characteristics of an estimator. Let $\hat{\mu}_Y$ denote some estimator of μ_Y,

- Unbiasedness: $\mathrm{E}(\hat{\mu}_Y) = \mu_Y$.
- Consistency: $\hat{\mu}_Y \xrightarrow{p} \mu_Y$.
- Efficiency. Let $\tilde{\mu}_Y$ be another estimator of μ_Y, and suppose 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 $\mathrm{Var}(\hat{\mu}_Y) < \mathrm{Var}(\tilde{\mu}_Y)$.

Properties of $\bar{Y}$

It can be shown that $\mathrm{E}(\bar{Y}) = \mu_Y$ and $\bar{Y} \xrightarrow{p} \mu_Y$ (from law of large numbers), $\bar{Y}$ is both unbiased and consistent.

But, is $\bar{Y}$ efficient?

Examples of alternative estimators.

Example 1: The first observation Y_1?

Since $\mathrm{E}(Y_1) = \mu_Y$, Y_1 is an unbiased estimator of μ_Y. But,

$$\mathrm{Var}(Y_1) = \sigma_Y^2 \geq \mathrm{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n},$$

if $n \geq 2$, $\bar{Y}$ is more efficient than Y_1.

Example 2:

$$\tilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right),$$

where n is assumed to be an even number. The mean of $\tilde{Y}$ is μ_Y and its variance is

$$\text{Var}(\tilde{Y}) = \frac{1.25\sigma_Y^2}{n} > \text{Var}(\bar{Y})$$

Thus $\tilde{Y}$ is unbiased and, because $\text{Var}(\tilde{Y}) \to 0$ as $n \to \infty$, $\tilde{Y}$ is consistent.
However, $\bar{Y}$ is more efficient than $\tilde{Y}$.

In fact, $\bar{Y}$ is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, \cdots, Y_n$. (Weighted average implies that the estimators are all unbiased.)

Hypothesis Testing

The **hypothesis testing** problem (for the mean): make a provisional decision, based on the evidence at hand, whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

$$
\begin{aligned}
H_0 : \mathrm{E}(Y) &= \mu_{Y,0} \; vs. H_1 : \mathrm{E}(Y) > \mu_{Y,0} \;\; (1 - \text{sided}, >) \\
H_0 : \mathrm{E}(Y) &= \mu_{Y,0} \; vs. H_1 : \mathrm{E}(Y) < \mu_{Y,0} \;\; (1 - \text{sided}, <) \\
H_0 : \mathrm{E}(Y) &= \mu_{Y,0} \; vs. H_1 : \mathrm{E}(Y) \neq \mu_{Y,0} \;\; (2 - \text{sided})
\end{aligned}
$$

- 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.

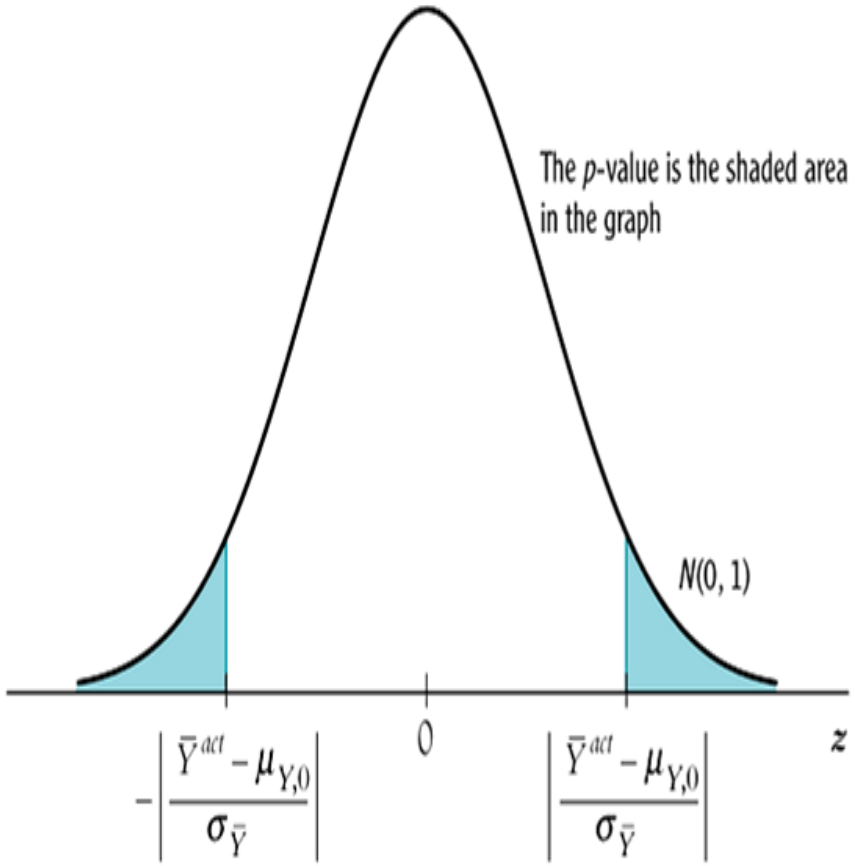
- To compute the p-value, you need to know the distribution of $\bar{Y}$.
- If n is large, we can use the large-n normal approximation.

$$
\begin{aligned}
p - \text{value} &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|] \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}|] \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|] \\
&\cong \text{probability under left + right } N(0, 1) \text{ tails}
\end{aligned}
$$

where $\sigma_{\bar{Y}}$ denotes the std. dev. of the distribution of $\bar{Y}$.

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 $\pm|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.



In practice, $\sigma_{\bar{Y}}$ is unknown - it too must be estimated.
Estimator of the variance of Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Fact: If $(Y_1, \cdots, Y_n)$ are i.i.d. and $\mathrm{E}(Y^4) < \infty$, 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.
- Technical note: we assume $\mathrm{E}(Y^4) < \infty$ because here the average is not of Y_i, but of its square.

Computing the p-value with σ_Y^2 estimated:

$$
\begin{aligned}
p-\text{value} &= \Pr_{H_0}[|\bar{Y}-\mu_{Y,0}| > |\bar{Y}^{act}-\mu_{Y,0}|] \\
&= \Pr_{H_0}[|\frac{\bar{Y}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}| > |\frac{\bar{Y}^{act}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}|] \\
&\cong \Pr_{H_0}[|\frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}| > |\frac{\bar{Y}^{act}-\mu_{Y,0}}{s_Y/\sqrt{n}}|](\text{ large } n) \\
&= \Pr_{H_0}[|t| > |t^{act}|] \\
&\cong \text{probability under normal tails (large } n)
\end{aligned}
$$

where $t = \frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}$.

The p-value and the significance level

With a prespecified significance level (e.g. 5%):

- reject if $|t| > 1.96$
- equivalently: reject if $p \leq 0.05$.
- The p-value is sometimes called the **marginal significance level**.

Digression: The Student t-distribution

If Y is distributed $N(\mu_Y, \sigma_Y^2)$, then the t-statistic has the Student t-distribution (tabulated in back of all stats books)

Some comments:

- For $n > 30$, the t-distribution and $N(0, 1)$ are very close.
- The assumption that Y is distributed $N(\mu_Y, \sigma_Y^2)$ is rarely plausible in practice (income? number of children?)
- The t-distribution is an historical artifact from days when sample sizes were very small.
- In this class, we won't use the t distribution - we rely solely on the large-n approximation given by the CLT.

Confidence Intervals

A 95% **confidence interval** for μ_Y is an interval that contains the true value of Y in 95% of repeated samples.
Digression: What is random here? the confidence interval - it will differ from one sample to the next; the population parameter, μ_Y, is not random, we just don't know it.

A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.

$$\begin{aligned}
& \{\mu_Y :| \frac{\bar{Y} - \mu_Y}{s_y/\sqrt{n}} |\leq 1.96\} \\
= & \{\mu_Y : -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_y/\sqrt{n}} \leq 1.96\} \\
= & \{\mu_Y : -1.96\frac{s_Y}{\sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1.96\frac{s_Y}{\sqrt{n}}\} \\
= & \{\mu_Y \in (\bar{Y} - 1.96\frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96\frac{s_Y}{\sqrt{n}})\}
\end{aligned}$$

This confidence interval relies on the large-n results that $\bar{Y}$ is approximately normally distributed and $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

Summary:

From the assumptions of:

(1) simple random sampling of a population, that is, $\{Y_i, i = 1, \cdots, n\}$ are i.i.d.

(2) $0 < \mathrm{E}(Y^4) < \infty$.

we developed, for large samples (large n):

- Theory of estimation (sampling distribution of $\bar{Y}$)
- Theory of hypothesis testing (large-n distribution of t-statistic and computation of the p-value).
- Theory of confidence intervals (constructed by inverting test statistic).

Are assumptions (1) & (2) plausible in practice? Yes

Tests for Difference between Two Means

Let μ_w be the mean hourly earning in the population of women recently graduated from college and let μ_m be population mean for recently graduated men. Consider the null hypothesis that earnings for these two populations differ by certain amount d, then

$$H_0 : \mu_m - \mu_w = d \; vs \; H_1 : \mu_m - \mu_w \neq d.$$

Since $\bar{Y}_m \sim N(\mu_m, \frac{\sigma_m^2}{n_m})$ and $\bar{Y}_w \sim N(\mu_w, \frac{\sigma_w^2}{n_w})$, then

$$\bar{Y}_m - \bar{Y}_w \sim N(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w})$$

Replace population variances by sample variances, we have the standard error

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

and the t-statistic is

$$t = \frac{\bar{Y}_m - \bar{Y}_w - d}{SE(\bar{Y}_m - \bar{Y}_w)}$$

If both n_m and n_w are large, the t-statistic has a standard normal distribution.

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

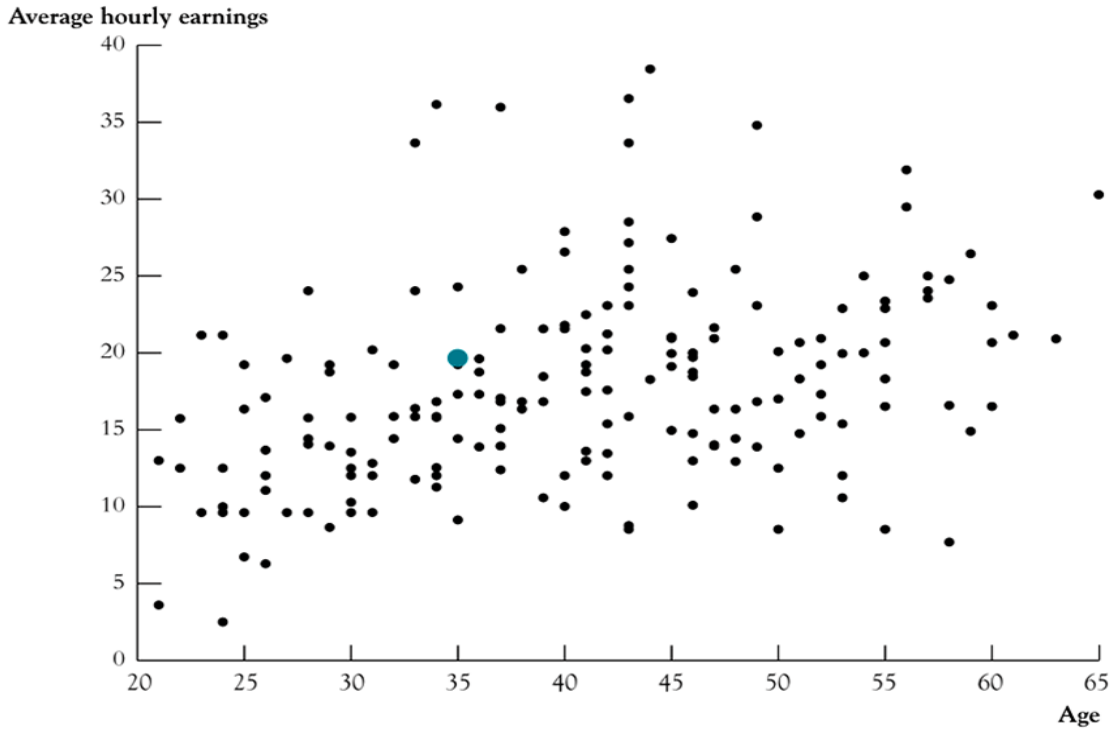
These estimates are computed using data on all full-time workers aged 25–34 from the CPS for the indicated years. The difference is significantly different from zero at the *5% or **1% significance level.

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

Summarize the relationship between variables

Scatterplots:

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.

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

The population covariance and correlation can be estimated by the **sample covariance** and **sample correlation**.

The **sample covariance** is

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

The **sample correlation** is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}, |r_{XY}| \leq 1$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

It can be shown that under the assumptions that (X_i, Y_i) are i.i.d. and that X_i and Y_i have finite fourth moments,

$$\begin{aligned} s_Y^2 &\xrightarrow{p} \sigma_Y^2 \\ s_{XY} &\xrightarrow{p} \sigma_{XY} \\ r_{XY} &\xrightarrow{p} \text{Corr}(X_i, Y_i) \end{aligned}$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

Prove that $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

$$
\begin{aligned}
s_Y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
(Y_i - \bar{Y})^2 &= [(Y_i - \mu_Y) - (\bar{Y} - \mu_Y)]^2 \\
&= (Y_i - \mu_Y)^2 - 2(Y_i - \mu_Y)(\bar{Y} - \mu_Y) \\
&\quad +(\bar{Y} - \mu_Y)^2
\end{aligned}
$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

Substituting $(Y_i - \bar{Y})^2$, collect terms and the fact that $\sum_{i=1}^n (Y_i - \mu_Y) = n(\bar{Y} - \mu_Y)$, we have

$$\begin{aligned} s_Y^2 &= \frac{1}{n-1}\sum_{i=1}^n (Y_i - \mu_Y)^2 - \frac{2}{n-1}\sum_{i=1}^n (Y_i - \mu_Y)(\bar{Y} - \mu_Y) \\ &\quad + \frac{1}{n-1}\sum_{i=1}^n (\bar{Y} - \mu_Y)^2 \\ &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_Y)^2 - \frac{n}{n-1}(\bar{Y} - \mu_Y)^2 \end{aligned}$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

From law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 \xrightarrow{p} \mathrm{E}(Y_i - \mu_Y)^2 = \sigma_Y^2,$$

$\bar{Y} \xrightarrow{p} \mu_Y$ and thus $(\bar{Y} - \mu_Y)^2 \xrightarrow{p} 0$,
and finally $\frac{n}{n-1} \to 1$ as $n \to \infty$, therefore

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

Prove that $s_{XY} \xrightarrow{p} \sigma_{XY}$.

$$
\begin{aligned}
& s_{XY} \\
= & \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\
= & \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \mu_X) - (\bar{X} - \mu_X)][(Y_i - \mu_Y) - (\bar{Y} - \mu_Y)] \\
= & \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) - \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X} - \mu_X)(Y_i - \mu_Y) \\
& - \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(\bar{Y} - \mu_Y) + \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)
\end{aligned}
$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

Use the fact that $\sum_{i=1}^n (Y_i - \mu_Y) = n(\bar{Y} - \mu_Y)$, $\sum_{i=1}^n (X_i - \mu_X) = n(\bar{X} - \mu_X)$ and collect terms, we have

$$\begin{aligned} s_{XY} &= \left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y) \\ &\quad - \left(\frac{n}{n-1}\right)(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \end{aligned}$$

It is easy to see that the second term converges in probability to zero because $\bar{X} \xrightarrow{p} \mu_X$ and $\bar{Y} \xrightarrow{p} \mu_Y$ so $(\bar{X} - \mu_X)(\bar{Y} - \mu_Y) \xrightarrow{p} 0$ by Slutsky's theorem.

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

By the definition of covariance, we have $\mathrm{E}[(X_i - \mu_X)(Y_i - \mu_Y)] = \sigma_{XY}$. To apply the law of large numbers on the first term, we need to have

$$\mathrm{Var}[(X_i - \mu_X)(Y_i - \mu_Y)] < \infty$$

which is satisfied since

$$\begin{aligned} \mathrm{Var}[(X_i - \mu_X)(Y_i - \mu_Y)] &= \mathrm{E}[(X_i - \mu_X)^2(Y_i - \mu_Y)^2] \\ &\leq \sqrt{\mathrm{E}(X_i - \mu_X)^4 \mathrm{E}(Y_i - \mu_Y)^4} \\ &< \infty \end{aligned}$$

The second inequality follows by applying the Cauchy-Schwartz inequality, and the last inequality follows because of the finite fourth moments for (X_i, Y_i).

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

The Cauchy-Schwartz inequality is

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\mathrm{E}(Y^2)}$$

Applying the las of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \xrightarrow{p} \mathrm{E}[(X_i - \mu_X)(Y_i - \mu_Y)] = \sigma_{XY}$$

Also, $\frac{n}{n-1} \to 1$, therefore

$$s_{XY} \xrightarrow{p} \sigma_{XY}$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

The Cauchy-Schwartz inequality is

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\mathrm{E}(Y^2)}$$

Applying the las of large numbers, we have

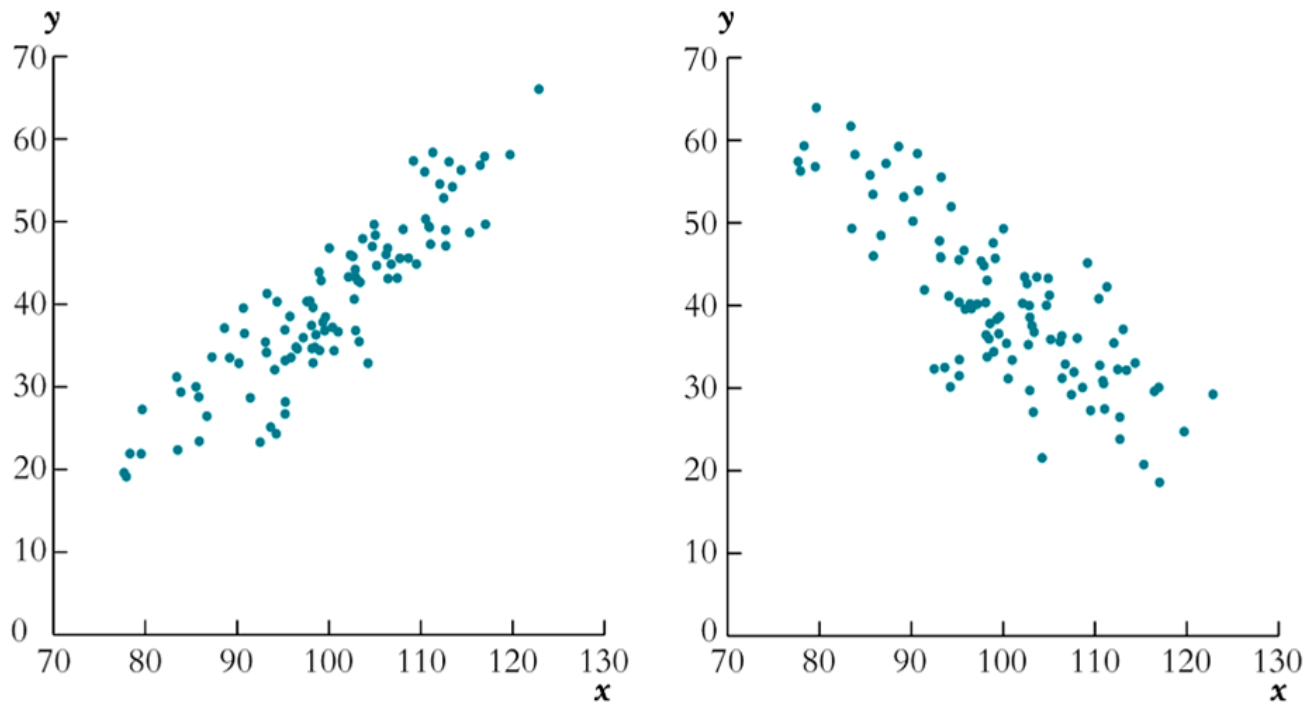
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \xrightarrow{p} \mathrm{E}[(X_i - \mu_X)(Y_i - \mu_Y)] = \sigma_{XY}$$

Also, $\frac{n}{n-1} \to 1$, therefore

$$s_{XY} \xrightarrow{p} \sigma_{XY}$$

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets

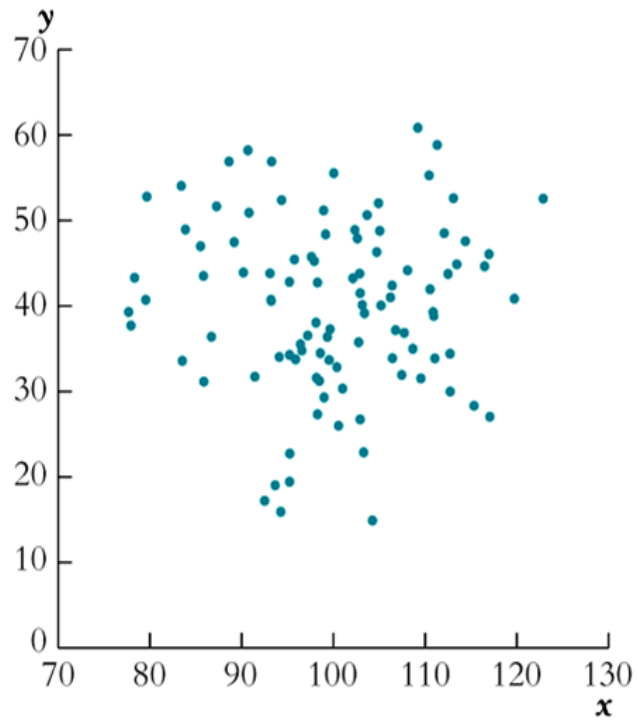


(a) Correlation = +0.9 (b) Correlation = –0.8

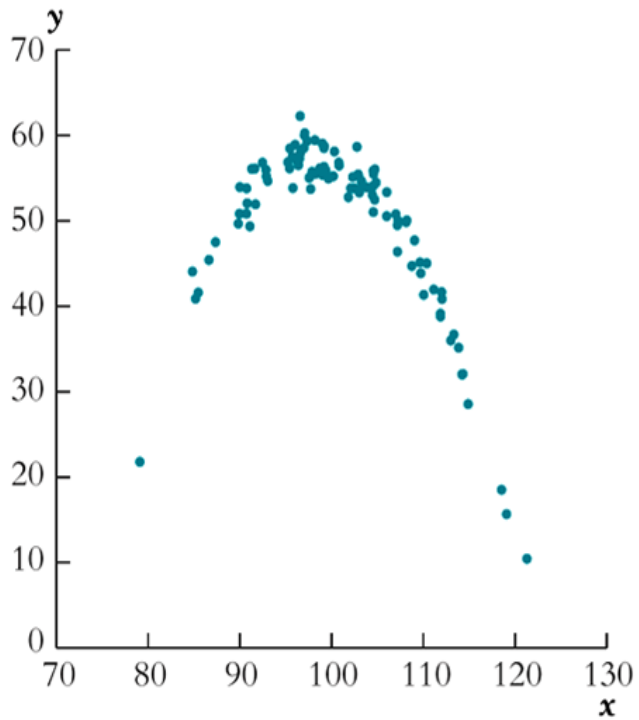
The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y. In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.

$s_Y^2 \xrightarrow{p} \sigma_Y^2$
$s_{XY} \xrightarrow{p} \sigma_{XY}$

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



(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.