Review of Statistics

Ming-Ching Luoh

2022.2.

Estimation of the Population Mean

- 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.

Estimators (估計式) and Their Properties

- The sample average $\bar{Y}$ is a natural way to estimate μ_Y, but, $\bar{Y}$ is not the only way. For example, the first observation Y_1 can be another **estimator** of μ_Y.
- What makes one estimator “better" than another? What are desirable characteristics of the sampling distribution of an estimator?

- 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: unbiasedness, consistency, and efficiency.

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

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

Thus $\tilde{Y}$ is unbiased and, because $\mathrm{Var}(\tilde{Y}) \to 0$ 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.)
- Said differently, $\bar{Y}$ is the Best Linear Unbiased Estimator (BLUE).
- It is the most efficient (best) estimator among all estimators that are unbiased and are linear function of $Y_1, \cdots, Y_n$.

$\bar{Y}$ is the least squares estimator of μ_Y.

- The sample average $\bar{Y}$ provides the best fit to the data in the sense that the average squared differences between the observation and $\bar{Y}$ are the smallest of all possible estimators.
- The solution to the problem of minimizing

$$\sum_{i=1}^{n}(Y_i - m)^2$$

is $\hat{m} = \bar{Y}$, which is called the least squares estimator.

Hypothesis Tests Concerning the Population Mean

Null and Alternative Hypotheses

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

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

- If the null hypothesis is "accepted," this does not mean that it is true. It is accepted tentatively with the recognition that it might be rejected later based on additional data.
- The p-value is the 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.
- For the case of population mean, the p-value is the probability of drawing $\bar{Y}$ at least as far in the tails of its distribution under the null hypothesis as the sample average you actually computed.

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

$$p-\text{value} = \Pr_{H_0}\left(\left|\bar{Y}-\mu_{Y,0}\right| > \left|\bar{Y}^{act}-\mu_{Y,0}\right|\right),$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed.

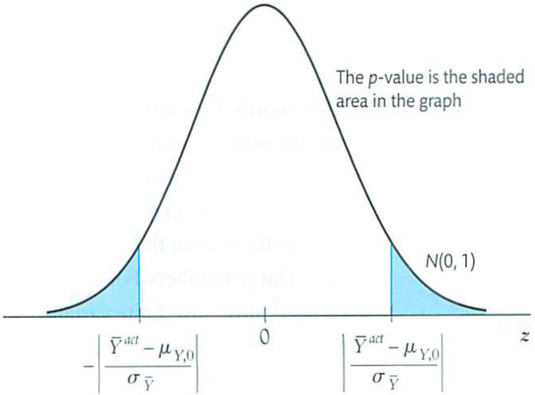
- To compute the p-value, we need to know the sampling distribution of $\bar{Y}$ under the null hypothesis.
- If n is large, $\bar{Y}$ is well approximated by a normal distribution.

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

where $\sigma_{\bar{Y}}$ denotes the standard deviation 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 $\pm|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.



The Sample Variance, Sample Standard Deviation, and Standard Error

- In practice, $\sigma_{\bar{Y}}$ is unknown and needs to 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 $E(Y^4) < \infty$, 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 $E(Y^4) < \infty$ because here the average is not of Y_i, but of its square.

Prove that $\underline{s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \xrightarrow{p} \sigma_Y^2}$.

First,
$$\begin{aligned}
& (Y_i - \bar{Y})^2 \\
= & \left[(Y_i - \mu_Y) - (\bar{Y} - \mu_Y)\right]^2 \\
= & (Y_i - \mu_Y)^2 - 2(Y_i - \mu_Y)(\bar{Y} - \mu_Y) + (\bar{Y} - \mu_Y)^2
\end{aligned}$$

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

For the first term,

- Define $W_i = (Y_i - \mu_Y)^2$, then $\mathrm{E}(W_i) = \sigma_Y^2$, and $W_i, \cdots, W_n$ are i.i.d.
- $\mathrm{E}(W_i^2) = \mathrm{E}\left[(Y_i - \mu_Y)^4\right] < \infty$ because $\mathrm{E}(Y_i^4) < \infty$.
- Thus $W_i, \cdots, W_n$ are i.i.d. and $\mathrm{Var}(W_i) < \infty$, so $\bar{W} \xrightarrow{p} \mathrm{E}(W_i) = \sigma_Y^2$, and $\frac{n}{n-1} \to 1$.
- Therefore, $\frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)^2\right] = \frac{n}{n-1}\bar{W} \xrightarrow{p} \sigma_Y^2$.

For the second term, because $\bar{Y} \xrightarrow{p} \mu_Y$, $(\bar{Y} - \mu_Y)^2 \xrightarrow{p} 0$.
Therefore, $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

Computing the p-value with estimated σ_Y^2:

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

where $t = \frac{\bar{Y}-\mu_{Y,0}}{s_Y/\sqrt{n}}$ is the t-statistic or t-ratio.

The p-value and the significance level

- Type I error: the null hypothesis (無罪) is rejected when in fact it is true. (誤判)
- Type II error: the null hypothesis (無罪) is not rejected when in fact it is false. (縱放)
- The prespecified probability of type I error is the significance level of the test.
- With a prespecified significance level (e.g. 5%):
  - reject if $|t| > 1.96$.
  - equivalently: reject if $p \leq 0.05$.

- The probability that the test actually incorrectly rejects the null hypothesis when it is true is the **size** of the test.
- The probability that the test correctly rejects the null hypothesis when the alternative is true is the **power** of the test.

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 Central Limit Theorem.

Confidence Intervals for the Population Mean

- Because of random sampling error, it is impossible to learn the exact value of the population mean of Y using only the information in a sample.
- It is possible to use data from a random sample to construct a set of values that contains the true population mean μ_Y with a certain prespecified probability.

- 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.

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.

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

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

Comparing Means from Different Populations

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

$$H_0: \mu_m - \mu_w = d \ v.s. \ H_1: \mu_m - \mu_w \neq d.$$

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)

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

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 Trends in Hourly Earnings in the United States of Working College Graduates, Ages 25–34, 1992 to 2012, in 2012 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	24.83	10.85	1594	21.39	8.39	1368	3.44**	0.35	2.75–4.14
1996	23.97	10.79	1380	20.26	8.48	1230	3.71**	0.38	2.97–4.46
2000	26.55	12.38	1303	22.13	9.98	1181	4.42**	0.45	3.54–5.30
2004	26.80	12.81	1894	22.43	9.99	1735	4.37**	0.38	3.63–5.12
2008	26.63	12.57	1839	22.26	10.30	1871	4.36**	0.38	3.62–5.10
2012	25.30	12.09	2004	21.50	9.99	1951	3.80**	0.35	3.11–4.49

These estimates are computed using data on all full-time workers ages 25–34 surveyed in the Current Population Survey conducted in March of the next year (for example, the data for 2012 were collected in March 2013). The difference is significantly different from zero at the **1% significance level.

Another Example.

TABLE 3.1 Differences in Household Income According to Childhood Socioeconomic Circumstances, Grouped by Level of Highest Qualification

	Father's NS-SEC = Higher			Father's NS-SEC = Routine			Difference, Higher vs. Routine		95% Confidence Interval for d	
Qualification	$\bar{Y}_h$	s_h	n_h	$\bar{Y}_r$	s_r	n_r	$\bar{Y}_h - \bar{Y}_r$	$SE(\bar{Y}_h - \bar{Y}_r)$		
None	£2,223.13	£2,115.12	1129	£1,842.98	£1,487.29	6383	£380.15	£65.64	£251.38	£508.93
GCSE/O-Level	£2,837.18	£1,819.73	1962	£2,596.93	£1,738.47	4042	£240.25	£49.35	£143.49	£337.00
A-Level	£3,045.99	£2,451.81	1216	£2,745.70	£1,912.50	1169	£300.30	£89.85	£124.11	£476.49
Undergraduate degree or more	£3,690.51	£2,743.55	4359	£3,370.96	£2,443.58	2505	£319.55	£64.11	£193.86	£445.23
All categories	£3,215.71	£2,497.73	8666	£2405.45	£1,886.86	14099	£810.25	£31.18	£749.13	£871.38

Source: Understanding Society.

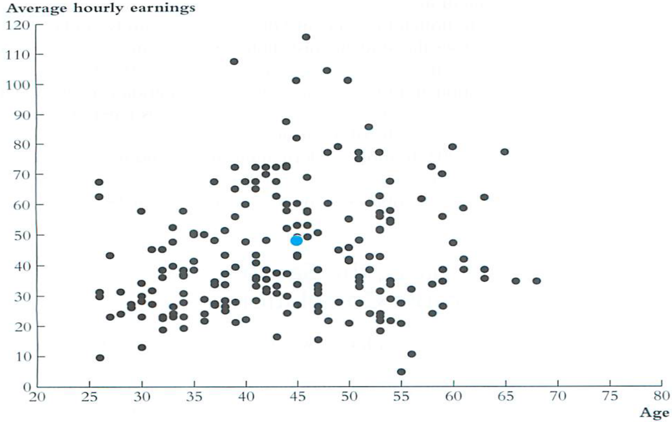
Scatterplots, the Sample Covariance, and the Sample Correlation

Three ways to summarize the relationship between two variables

- scatterplot,
- sample covariance,
- sample correlation coefficient.

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 200 workers in the sample. The highlighted dot corresponds to a 45-year-old worker who earns $49.15 per hour. The data are for computer and information systems managers from the March 2016 CPS.

Sample Covariance and Correlation

- 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

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

- 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,

$$
\begin{aligned}
s_Y^2 &\xrightarrow{p} \sigma_Y^2 \\
s_{XY} &\xrightarrow{p} \sigma_{XY} \\
r_{XY} &\xrightarrow{p} \mathrm{Corr}(X, Y)
\end{aligned}
$$

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

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

- 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.

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

$$\mathrm{Var}\left((X_i - \mu_X)(Y_i - \mu_Y)\right) < \infty$$

which is satisfied since

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

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).

- *The Cauchy-Schwartz inequality* is

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

- Applying the law of large numbers, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \\ &\xrightarrow{p} \mathrm{E}\left((X - \mu_X)(Y - \mu_Y)\right) = \sigma_{XY} \end{aligned}$$

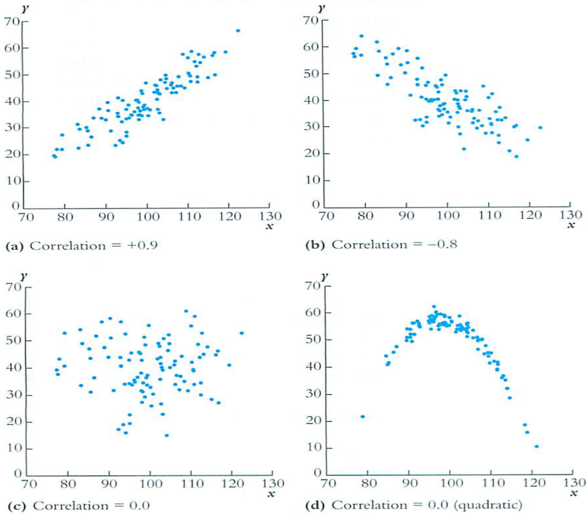
- Also, $\frac{n}{n-1} \to 1$, therefore

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

Scatterplots and Sample Correlation

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets

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.



(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)