Lecture 3: Review of Statistics

Zheng Tian

Outline

The goal of estimation

- Suppose we draw n random samples, $Y_1, \ldots, Y_n$, and $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$.
- The goal is to estimate μ_Y given these n samples. A natural way is to compute the sample average, $\overline{Y}$.

Estimators

- 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 realization of the estimator.

Estimators of μ_Y

- $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$ is an estimator of μ_Y.
- Y_1, the first observation, can also be used as an estimator because it is indeed a function of sample data.
- As such, we can have many different estimators of μ_Y. How can we judge which estimator is better than another?

Definition of unbiased estimators

- Let $\hat{\mu}_Y$ be an estimator of μ_Y. The estimator $\hat{\mu}_Y$ is said to be unbiased if

$$\mathrm{E}(\hat{\mu}_Y) = \mu_Y$$

where $\mathrm{E}(\hat{\mu}_Y)$ is the expectation of the sampling distribution of $\hat{\mu}_Y$.

Are $\overline{Y}$ and Y_1 unbiased?

- $\overline{Y}$ is an unbiased estimator of μ_Y.

  In Lecture 2, we have already shown that $\mathrm{E}(\overline{Y}) = \mu_Y$ when $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$.

- Y_1 is also an unbiased estimator.
  $\mathrm{E}(Y_1) = \mu_Y$ when Y_1 is drawn from $IID(\mu_Y, \sigma_Y^2)$.

Definition of consistent estimators

- $\hat{\mu}_Y$ is a consistent estimator of μ_Y if $\hat{\mu}_Y$ is convergent in probability to μ_Y. That is, $\hat{\mu}_Y$ is consistent if

$$\hat{\mu}_Y \xrightarrow{\text{p}} \mu_Y \text{ as } n \to \infty$$

Are $\overline{Y}$ and Y_1 consistent?

- $\overline{Y}$ is a consistent estimator of μ_Y.

  The law of large number ensures that $\overline{Y} \xrightarrow{\text{p}} \mu_Y$ is true when $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$, and $\sigma_Y^2 < \infty$.

- However, we cannot assess the consistency for Y_1 because it cannot be written as the form of an average.

Definition of efficient estimators

- When both $\tilde{\mu}_Y$ and $\hat{\mu}_Y$ are two unbiased estimators of μ_Y, we choose the estimator with the tightest sampling distribution, which means the smallest variance.
- $\hat{\mu}_Y$ is said to be more efficient than $\tilde{\mu}_Y$ if

$$\text{Var}(\hat{\mu}_Y) < \text{Var}(\tilde{\mu}_Y)$$

- In words, $\hat{\mu}_Y$ is more efficient than $\tilde{\mu}_Y$ because $\hat{\mu}_Y$ uses the information in the data more efficiently than does $\tilde{\mu}_Y$.

$\overline{Y}$ is more efficient than Y_1?

- In Lecture 2, we compute the variance of $\overline{Y}$ to be σ_Y^2/n when $Y_i \sim IID(\mu_Y, \sigma_Y^2)$.
- The variance of Y_1 is σ_Y^2.
- When $n > 1$, $\overline{Y}$ is more efficient than Y_1.

BLUE

- $\overline{Y}$ happens to be the Best Linear Unbiased Estimator (BLUE).
- It means that among all linear unbiased estimator, $\overline{Y}$ has the smallest variance.

Linear estimators and $\overline{Y}$ is BLUE

- A linear estimator of μ_Y is a weighted average of $Y_1, \ldots, Y_n$, written as

$$\tilde{\mu}_Y = \frac{1}{n} \sum_{i=1}^{n} \alpha_i Y_i$$

  where $\alpha_1, \ldots, \alpha_n$ are nonrandom constants.
- If $\tilde{\mu}_Y$ is another unbiased estimator of μ_Y, then we always have $\mathrm{Var}(\overline{Y}) \leq \mathrm{Var}(\tilde{\mu}_Y)$, and the equality holds only if $\tilde{\mu}_Y = \overline{Y}$. It means that $\overline{Y}$ is BLUE.

A linear model for the population mean

- Consider the following model

$$Y_i = \alpha + u_i \text{ for } i = 1, 2, \ldots, n$$

where α is a nonrandom intercept to be estimated.

- u_i is the error term, which is a random variable with $\mathrm{E}(u_i) = 0$. Thus, we have $E(Y_i) = \alpha = \mu_Y$.
- u_i can be seen as the error of predicting Y_i with α for each i, and we use

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

to measure the total prediction errors.

- A natural choice of an estimator of α is the one that minimizes this sum of squared errors.

The least squares estimator

- The least squares estimator of μ_Y (or α) is obtained by solving the following problem

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

- The solution of this minimization problem is just $a = \overline{Y}$.

The proof for $\overline{Y}$ is the least square estimator

- The first order condition for the minimization problem is

$$\frac{d}{da}\sum_{i=1}^{n}(Y_i - a)^2 = -2\sum_{i=1}^{n}(Y_i - a) = -2\sum_{i=1}^{n} Y_i + 2na = 0$$

- Solving the equation for a, we get $a = 1/n\sum_{i=1}^{n} Y_i = \overline{Y}$.

The null hypothesis

- Hypothesis testing is thus to make a provisional decision based on the evidence at hand on.
- The hypothesis of the population mean, $\mathrm{E}(Y)$, taking on a specific value, $\mu_{Y,0}$. So the null hypothesis, denoted as H_0, is

$$H_0 : E(Y) = \mu_{Y,0}$$

The alternative hypothesis

- The alternative hypothesis, denoted as H_1
  - The two-sided alternative: $H_1 : E(Y) \neq \mu_{Y,0}$
  - The one-sided alternative: $H_1 : E(Y) > \mu_{Y,0}$
- The language
  One thing should be kept in mind is that we usually do not say "accept the null hypothesis" when the hypothesis test is in favor of the null, but say "fail to reject the null".

The z-statistic when σ_Y is known

- We know that when $Y_i \sim IID(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$, $E(\overline{Y}) = \mu_Y$ and $\text{Var}(\overline{Y}) = \sigma_{\overline{Y}}^2 = \sigma_Y^2/n$.
- In the null hypothesis, we specify $\mu_Y = \mu_{Y,0}$.
- So given that σ_Y is known, the z-statistic is computed as

$$z = \frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$$

- As $n \to \infty$, by the central limit theorem, we know $z \xrightarrow{\text{d}} N(0, 1)$.

The t-statistic when σ_Y is unknown

- Of course, σ_Y is the standard deviation of the population variance that is usually unknown. So we need to replace σ_Y with its estimator.

The sample variance and standard deviation

- The sample variance s_Y^2 is is an estimator of the population variance σ_Y^2, which is computed as

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

- The sample standard deviation, s_Y, is the square root of s_Y^2
- The sample variance, s_Y^2, is a consistent estimator of the population variance, that is, as

$$n \to \infty, s_Y^2 \xrightarrow{\text{p}} \sigma_Y^2$$

The standard error of $\overline{Y}$

- The standard error of $\overline{Y}$, denoted as $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$, is an estimator of the standard deviation of $\overline{Y}$, $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, with s_Y replacing σ_Y.

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = \frac{s_Y}{\sqrt{n}}$$

The t-statistic

- When σ_Y is unknown, by replacing σ_Y with s_Y, we have the t statistic

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} = \frac{\overline{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

- The asymptotic distribution of the t statistic is $N(0,1)$ because s_Y is a consistent estimator of σ_Y.
- When Y_i for $i = 1, \ldots, n$ are i.i.d. from $N(\mu_Y, \sigma_Y^2)$, we can show that the exact distribution for the t statistic is the student t distribution with $(n-1)$ degrees of freedom. That is

$$t \sim t(n-1)$$

Hypothesis testing with a pre-specified significance level

With the null and alternative hypotheses being the goal of the test and test statistics being the tools, we need a rule to make a judgment: When can we reject (or fail to reject) the null hypothesis if the test statistic takes on what values? To do so, we need to first define some concepts.

Type I and type II errors

- A statistical hypothesis test can make two types of mistakes:
  - Type I error. The null hypothesis is rejected when in fact it is true.
  - Type II error. The null hypothesis is not rejected when in fact it is false.

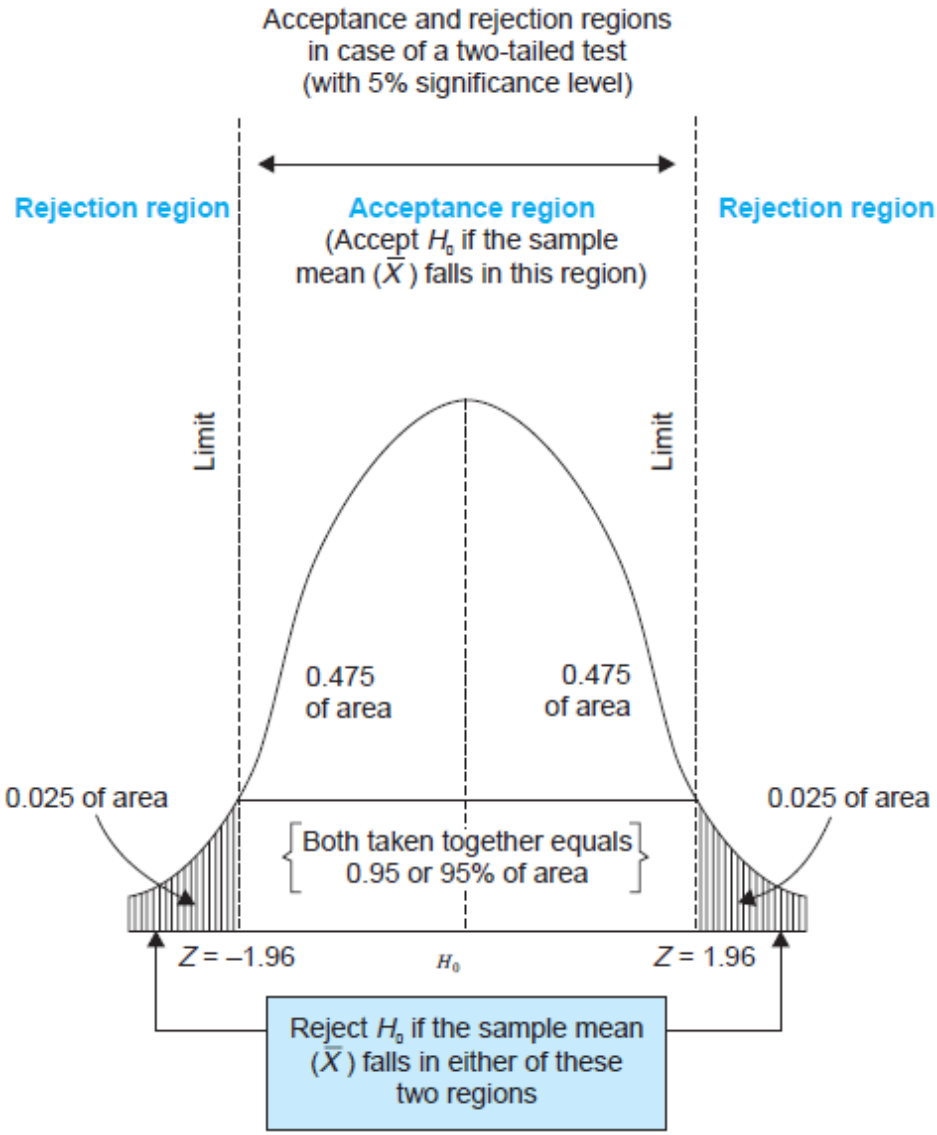
The significance level and the critical value

- The significance level is the pre-specified probability of type I error. Usually, we set the significance level to be $\alpha = 0.05, 0.10,$ or 0.01.
- The critical value, denoted as c_α, is the value of the test statistic for which the test rejects the null hypothesis at the given significance level. The $N(0,1)$ critical value for a two-sided test with a 5% significance level is 1.96.

The rejection rule and rejection region

- The rejection rule. For a two-sided test, we reject the null hypothesis when $|z^{act}| > c_\alpha$.
- The rejection region is the set of values of the test statistic for which the test rejects the null, and the acceptance region is the vice.

The rejection region illustrated



The power and the size of the test

- The size of the test is the probability that the test actually incorrectly rejects the null hypothesis when it is true. That is, the size of the test is just the significance level.
- The power of the test is the probability that the test correctly rejects the null when the alternative is true. That is,

$$\text{power} = 1 - \Pr(\text{type II error})$$

The p-value

- The p-value, also called the significance probability, is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct.
- The p-value provides more information than the significance level. In fact, the p-value is also named the marginal significance level, which the smallest significance level at which you can reject the null hypothesis.

Rejection rule with the p-value

- The rejection rule of rejecting the null is then the p-value $< \alpha$.
- Mathematically, the p-value is computed as

$$p\text{-value} = \begin{cases} \Pr_{H_0}(|z| > |z^{act}|) = 2\Phi(-|z^{act}|) \text{ when } \sigma_Y \text{ is known} \\ \Pr_{H_0}(|t| > |t^{act}|) = 2\Phi(-|t^{act}|) \text{ when } \sigma_Y \text{ is unknown} \end{cases}$$

One-sided alternatives

- For a one-sided alternative hypothesis, $H_1 : \mathrm{E}(Y) > \mu_{Y,0}$, we can compute the p-value as

$$p\text{-value} = \Pr_{H_0}(t > t^{act}) = 1 - \Phi(t^{act})$$

- The $N(0,1)$ critical value for a one-sided test with a 5% significance level is 1.64. The rejection region for this test is all values of the t-statistic exceeding 1.64.

Definitions

- A confidence set is the set of values that contains the true population mean μ_Y with a certain prespecified probability.
- A confidence level is the prespecified probability that μ_Y is contained in the confidence set. confidence level $= 1 -$ significance level.
- A confidence interval is the confidence set when it is an interval.
- In the case of a two-sided test for μ_Y, we say that a 95% confidence interval is an interval constructed so that it contains the true value of μ_Y in 95% of all possible random samples.

Constructing a confidence interval based on the t statistic

- Step 1: we compute the t statistic for the two-sided test

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} \xrightarrow{\text{d}} N(0,1)$$

- Step 2: we know that we fail to reject the null at the 5% level if $|t| < 1.96$.
- Step 3: we plug in the definition of t and solving for $|t| \leq 1.96$, we get

$$-1.96 \leq \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} \leq 1.96$$

$$\overline{Y} - 1.96SE(\overline{Y}) \leq \mu_{Y,0} \leq \overline{Y} + 1.96SE(\overline{Y})$$

The 95%, 90%, and 99% confidence interval

- The 95% confidence interval two-sided confidence interval for μ_Y is

$$\{\overline{Y} \pm 1.96SE(\overline{Y})\}$$

- 90% confidence interval for $\mu_Y = \{\overline{Y} \pm 1.64SE(\overline{Y})\}$
- 99% confidence interval for $\mu_Y = \{\overline{Y} \pm 2.58SE(\overline{Y})\}$

Hypothesis tests for the difference between two means

- The question is whether there is a difference in earnings between male college graduates and female college graduates.
- Let $Y_{m,i}$ for $i = 1, \ldots, n_m$ be n_m i.i.d. samples from the population of earnings of male college graduate, i.e.,

$$Y_{m,i} \sim IID(\mu_m, \sigma_m^2) \text{ for } i = 1, \ldots, n_m$$

- Let $Y_{w,j}$ for $j = 1, \ldots, n_w$ be n_w i.i.d. samples from the population of earnings of female college graduate, i.e.,

$$Y_{w,j} \sim IID(\mu_w, \sigma_w^2) \text{ for } j = 1, \ldots, n_w$$

- Also, we assume that $Y_{m,i}$ and $Y_{w,j}$ are independent.

The null and alternative hypotheses

- The hypothesis to be tested is whether the mean earnings for the male and female graduates differ by a certain amount, that is,

$$H_0 : \mu_m - \mu_w = d_0, \quad \text{vs.} \quad H_1 : \mu_m - \mu_w \neq d_0$$

The test procedures: step 1

- Calculate the sample average earnings:
  - $\overline{Y}_m$ for the male and $\overline{Y}_w$ for the female.
  - As n_m and n_w get large, we know $\overline{Y}_m \xrightarrow{\text{d}} N(\mu_Y, \sigma_m^2/n_m)$, and $\overline{Y}_w \xrightarrow{d} N(\mu_w, \sigma_w^2/n_w)$.
  - Given that $\overline{Y}_m - \overline{Y}_w$ is a linear function of $\overline{Y}_m$ and $\overline{Y}_w$, and $Y_{m,i}$ and $Y_{w,j}$ are independent, we know that

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

Step 2

- When σ_m^2 and σ_w^2 are known, we use the z statistic

$$z = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{\left(\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right)^{1/2}} \xrightarrow{\text{d}} N(0,1)$$

- When σ_m^2 and σ_w^2 are unknown, we the t statistic

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \xrightarrow{d} N(0,1)$$

where

$$SE(\overline{Y}_m - \overline{Y}_w) = \left(\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}\right)^{1/2}$$

$$s_m^2 = \frac{1}{n_m - 1}\sum_{i=1}^{n_m}(Y_{m,i} - \overline{Y}_m)^2$$

$$s_w^2 = \frac{1}{n_w - 1}\sum_{i=1}^{n_w}(Y_{w,i} - \overline{Y}_w)^2$$

Step 3

- Calculate the p value: The p value for the two-sided test is calculated as

$$p\text{-value} = 2\Phi(-|t|)$$

- For a two-sided test at the 5% significant level, we can reject the null hypothesis when the p value is less than 5%, or, equivalently, when $|t| > 1.96$.

Confidence intervals for the difference between two means

- The 95% confidence interval can be constructed as usual based on the t statistic we have computed above.
- The 95% confidence interval for $d = \mu_m - \mu_w$ is

$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96 SE(\overline{Y}_m - \overline{Y}_w)$$

Differences-of-Means Estimation of Causal Effects Using Experimental Data

- We define the outcome of a randomized controlled experiment to be Y and the binary treatment variable to be X, $X = 1$ for the treatment group and $X = 0$ for the control group.
- Then the causal effect of the treatment can be conveniently expressed as the difference in the conditional expectation

$$E(Y \mid X = 1) - E(Y \mid X = 0)$$

Scatterplots

- Exploratory data analysis. Drawing graphs is an important aspect of exploratory data analysis to visualize the patterns of the variables of interests.
- A scatterplot is a plot of n observations on X_i and Y_i, in which each observation is represented by the point (X_i, Y_i)

An example of scatterplot

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.

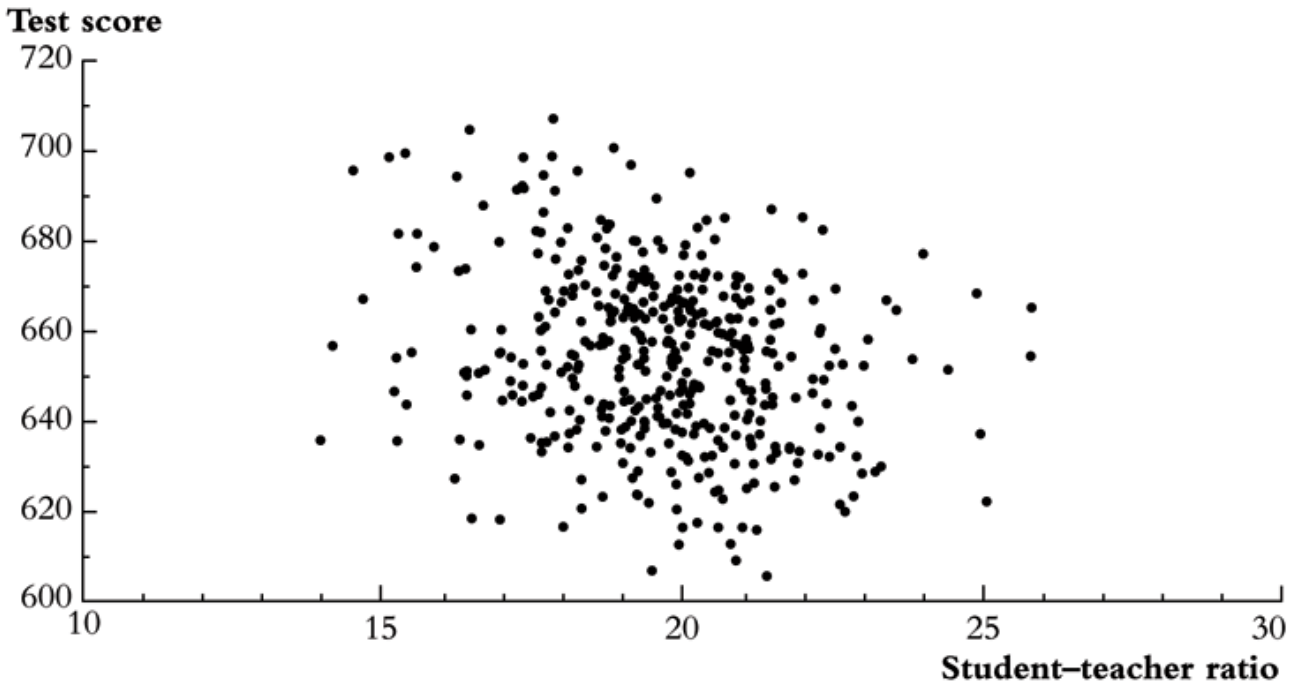


Figure: The scatterplot between test scores and student-teacher ratios

Sample covariance and sample correlation coefficient

- The sample covariance, denoted as s_{XY}, is

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

- The sample correlation coefficient, denoted as r_{XY}, is

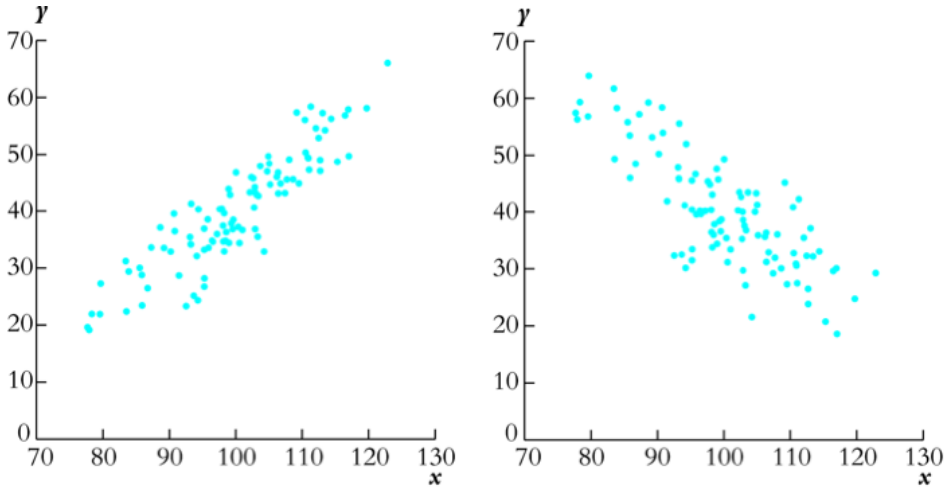
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

  and we have $|r_{XY}| \leq 1$.
- If $(X_i,\ Y_i)$ are i.i.d. and X_i and Y_i have finite fourth moments, then

$$s_{XY} \xrightarrow{\text{p}} \sigma_{XY} \text{ and } r_{XY} \xrightarrow{\text{p}} \rho_{XY}$$

The correlation coefficient measures the linear association

We should emphasize that the correlation coefficient is a measure of linear association between X and Y.



(a) Correlation = +0.9

(b) Correlation = −0.8