5

INTRODUCTION TO STATISTICS

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. Wells

Statistics has been likened to a telescope. The latter enables one to see further and to make clear objects which were diminished or obscured by distance. The former enables one to discern structure and relationships which were distorted by other factors or obscured by random variation.

D. J. Hand, *Psychological Medicine* (1985)

This chapter covers some elementary concepts of statistics, which are necessary for probabilistic analysis when an appropriate probabilistic model is not given.

Main Topics

- Sampling, Sample Mean, and Sample Variance
- Empirical Distributions
- Parameter Estimation
- Hypothesis Testing
- Linear Regression and Curve Fitting

5.1 Introduction

Previous chapters are devoted to the *analysis* of a random phenomenon, where it is assumed that a probabilistic model of the phenomenon (e.g., the distribution of the random variable) is given. In reality, however, it is quite often the case that the probabilistic model is not known. In this case, statistics is essential in establishing the correct probabilistic model. In this chapter we study some elementary concepts of statistics.

Statistics is a science dealing with data subject to uncertainties. It has an extremely wide spectrum of application. Examples of the application areas are

- Quality control
- Instrumentation
- Insurance
- Poll taking
- Weather forecasting.

Statistics is a magic weapon: Given a collection of data/facts, one can arrive at almost any conclusion he/she wants by *abusing* statistics.

Terminology

- The entire collection of data being studied is called a ***population***. It is a RV whose possible values are the values of the data.
- A ***sample*** $(X_1, \ldots, X_n)$ is a subset of the population selected at random.
- The *population size* is the number of data pieces that make up the population.
- The ***sample size*** is the number of pieces of data that make up the sample.

For example, if we conduct a poll by asking 1,000 people to predict the outcome of a U.S. presidential election, then the population consists of all people in the U.S. who are eligible to vote. The sample for this poll is those people surveyed and the sample size is 1,000. The population size is the number of eligible people in the U.S.

It is almost always assumed that the data (i.e., RVs $X_1, X_2, \ldots, X_n$) making up a sample are ***independent and identically distributed*** (***i.i.d.***).

5.2 Sample Mean and Sample Variance

5.2.1 Sample Mean

The ***sample mean*** (or ***sample average***) of a sample $(X_1, \ldots, X_n)$ is defined as the average value of the sample:

$$\boxed{\hat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \tag{5.1}$$

The sample mean is used in practice to estimate (approximate) the unknown true mean of the population:

$$\bar{x} = E[X] \approx \hat{X} \tag{5.2}$$

The sample mean is a RV since it is an (evenly weighted) sum of RVs $X_1, \ldots, X_n$. Given a realization $(x_1, \ldots, x_n)$ of the sample (i.e., given the values $x_1, \ldots, x_n$ of the RVs $X_1, \ldots, X_n$), the value (or realization)

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

of the sample mean is a nonrandom number.

The expected value (mean) of the sample mean is equal to the true mean:

$$E[\hat{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \bar{x} = \bar{x} \tag{5.3}$$

which has the following important interpretation: *The sample mean equals the true mean on the average.*

The variance of the sample mean as a RV is, since $X_1, \ldots, X_n$ are i.i.d.,

$$\sigma_{\hat{x}}^2 = \text{var}(\hat{X}) = \text{var}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] \overset{(4.32)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{1}{n}\sigma_x^2 \tag{5.4}$$

where σ_x^2 is the variance of the population (i.e., of the RV X). $\sigma_{\hat{x}}^2$ provides a measure of the variation of the sample mean. Thus, the variation of the sample mean is reduced by increasing the sample size n.

Since the sample mean is the sum of independent RVs, if the size is large, the sample mean is approximately Gaussian distributed in view of the central limit theorem.

Example 5.2: Probabilistic Analysis of the Average Score of a Class

Suppose that the score of a student in a school is approximately a $\mathcal{N}(70, 10^2)$ RV. Consider the average score of a class of 20 students.

Here we are given a sample of size 20 and $\bar{x} = 70, \sigma_x = 10$.

(a) Find the expected value of the average score of this class: The average score is the sample mean and thus its expected value is

$$E[\hat{X}] = \bar{x} \text{ (true mean)} = 70$$

(b) Find the variance of the average score of the class: The variance of the average score is the variance of the sample mean

$$\sigma_{\hat{x}}^2 \overset{(5.4)}{=} \sigma_x^2/n = 10^2/20 = 5$$

(c) Is the average score of this class a Gaussian RV? The average score is the sample mean of the score of a student in this school. Since it is a weighted sum of independent Gaussian RVs for this example, from Example 3.19 it is a Gaussian RV.

(d) Find the probability that a student's score will be in $(60, 85)$: The score X of an arbitrary student is a $\mathcal{N}(70, 10^2)$ RV. Thus, $\tilde{X} = \frac{X-\bar{x}}{\sigma_x} = \frac{X-70}{10}$ is a standard Gaussian RV. Hence

$$\begin{aligned} P\{60 < X < 85\} &= P\Big\{\frac{60-70}{10} < \frac{X-70}{10} < \frac{85-70}{10}\Big\} = P\{-1 < \tilde{X} < 1.5\} \\ &= \Phi(1.5) - \Phi(-1) \overset{\Phi \text{ table}}{=} 0.9332 - (1 - 0.8413) = 77.45\% \end{aligned}$$

(e) Find the probability that the average score of this class will be in $(65, 75)$: Since $E[\hat{X}] = 70, \sigma_{\hat{x}}^2 = 5$ and the average score $\hat{X}$ is a Gaussian RV, it is a $\mathcal{N}(70, 5)$ RV. Thus, $\tilde{\hat{X}} = \dfrac{\hat{X} - \bar{x}}{\sigma_{\hat{x}}} = \dfrac{\hat{X} - 70}{\sqrt{5}}$ is a standard Gaussian RV. Hence

$$\begin{aligned} P\{65 < \hat{X} < 75\} &= P\Big\{\frac{65-70}{\sqrt{5}} < \frac{\hat{X}-70}{\sqrt{5}} < \frac{75-70}{\sqrt{5}}\Big\} \\ &= P\{-2.236 < \tilde{\hat{X}} < 2.236\} = \Phi(2.236) - \Phi(-2.236) \\ &= 0.9873 - (1 - 0.9873) = 97.46\% \end{aligned}$$

From (d) and (e), the distribution of the average score is much more concentrated around its mean than the score of a student. Does this make sense?

5.2.2 Sample Variance

The ***sample variance*** of a sample $(X_1, X_2, \ldots, X_n)$ is defined as

$$\boxed{\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{X})^2} \tag{5.5}$$

Sample variance is used in practice to estimate the variance of the population:

$$\sigma_x^2 \approx \hat{V} \tag{5.6}$$

Sample variance is also a RV. Its value, given the value of a sample, will be denoted by $\hat{v}$.

Note that

$$\begin{aligned} E\left[(X_i - \hat{X})^2\right] &= E\left[[(X_i - \bar{x}) - (\hat{X} - \bar{x})]^2\right] \\ &= E\left[(X_i - \bar{x})^2 + (\hat{X} - \bar{x})^2 - 2(X_i - \bar{x})(\hat{X} - \bar{x})\right] \\ &= \sigma_x^2 + \sigma_{\hat{x}}^2 - 2E\Big[(X_i - \bar{x})\Big(\frac{1}{n}\sum_{i=1}^{n}(X_j - \bar{x})\Big)\Big] \\ &= \sigma_x^2 + \frac{1}{n}\sigma_x^2 - 2E\Big[(X_i - \bar{x}) \cdot \frac{1}{n}[(X_1 - \bar{x}) + \cdots + (X_n - \bar{x})]\Big] \\ &\overset{?}{=} \sigma_x^2 + \frac{\sigma_x^2}{n} - \frac{2\sigma_x^2}{n} = \sigma_x^2 - \frac{\sigma_x^2}{n} = \frac{n-1}{n} \cdot \sigma_x^2 \end{aligned}$$

where $\overset{?}{=}$ follows from the fact that $X_1, X_2, \ldots, X_n$ are independent and thus uncorrelated. Thus, the expected value of $\hat{V}$ is

$$\begin{aligned} E[\hat{V}] = E\Big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{X})^2\Big] &= \frac{1}{n-1}\sum_{i=1}^{n} E\left[(X_i - \hat{X})^2\right] \\ = \frac{1}{n-1}\Big[n \cdot \frac{n-1}{n} \cdot \sigma_x^2\Big] &= \sigma_x^2 \end{aligned}$$

That is*, the sample variance is equal to the true variance of the population (i.e., the RV) on the average*.

Note that were the sample variance defined as $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X})^2$, it would not be equal to the true variance on the average, although it is sometimes also called the *sample variance*. If the true mean $\bar{x}$ of the population is known, however, the sample variance should be defined as $\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$.

The square root of the sample variance is called the ***standard error***.

Example 5.3: Determination of the Gaussian Curve of Test Scores

Suppose that the score of a student in a school is approximately a $\mathcal{N}(\bar{x}, \sigma^2)$ RV and that 30 students were surveyed whose scores are: 79, 43, 66, 99, 91, 88, 89, 74, 78, 83, 77, 68, 89, 85, 69, 58, 90, 84, 75, 77, 83, 94, 57, 63, 65, 79, 66, 68, 74, 75.

(a) The average score of the class is the sample mean, given by

$$\hat{x} = \frac{1}{30}[79 + 43 + 66 + \cdots + 68 + 74 + 75] = 76.2$$

(b) The sample variance is a measure of the deviation of an individual score from the average score of the class. It has the value

$$\hat{v} = \frac{1}{30-1}\Big[(79-76.2)^2 + (43-76.2)^2 + \cdots + (75-76.2)^2\Big] = 154.23$$

Hence, the *standard error* is $\sqrt{\hat{v}} = \sqrt{154.23} = 12.42$.

(c) The standard deviation of the sample mean is approximately given by

$$\sigma_{\hat{x}} = \sqrt{\sigma^2/n} \approx \sqrt{\hat{v}/30} = \sqrt{154.23/30} = 2.27$$

It is a measure of the variation (randomness) of the average score.

We may thus conclude that the test score of a student in this school is approximately a $\mathcal{N}(76.2, 154.23)$ RV and the average score is approximately a $\mathcal{N}(76.2, 2.27^2)$ RV. The Gaussian curve of the test scores provided by your professor may have been generated in this way.

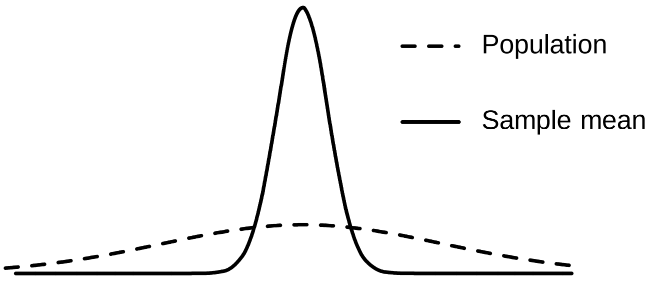


Figure 5.1: PDFs of the population and sample mean.

5.3 Empirical Distributions

In the previous chapter, it was assumed that the distribution of a RV X is given. In reality, it is more often the case that the distribution is not known but the values $x_1, \ldots, x_n$ of a sample of X are given. Note that the elements $X_1, \ldots, X_n$ of a sample are assumed to be independent. How do we obtain an approximate distribution of the RV X based on the data $x_1, \ldots, x_n$?

The ***empirical CDF*** of a RV X given the values $x_1, \ldots, x_n$ of a sample of X is defined as

$$\bar{F}(x|x_1, \ldots, x_n) = \frac{\text{number of sample values } x_1, \ldots, x_n \text{ not greater than } x}{n}$$

The empirical CDF can be used to approximate the true CDF. This is justified by the so-called Glivenko theorem which states that the empirical CDF uniformly converges to the true CDF with probability one as $n \to \infty$.

$\bar{F}(x|x_1, \ldots, x_n)$ is actually a histogram of a stairway type. Its value at x is the percentage of the points $x_1, \ldots, x_n$ that are not larger than x. It is often more convenient to use a histogram of the PDF type. This can be done based on the relation $f_X(x) \approx P\{x < X \leq x + \Delta x\}/\Delta x$ for small Δx. Thus the ***empirical PDF*** can be defined as

$$\bar{f}(x|x_1, \ldots, x_n) = \frac{\text{number of sample values } x_1, \ldots, x_n \text{ in } [x, x + \Delta x)}{n\Delta x}$$

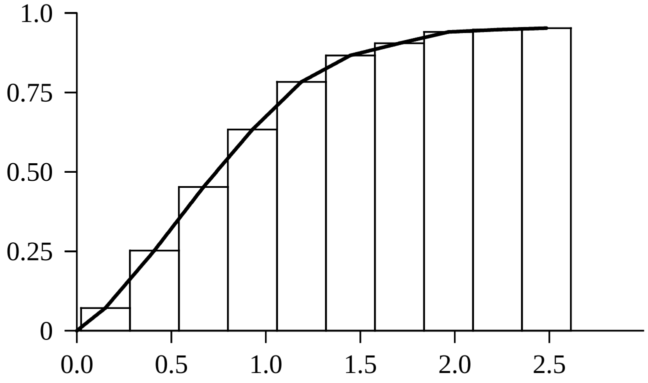


Figure 5.2: The empirical CDF of Example 5.4.

Example 5.4: Identification of Distribution From Data

The file `e5_4.dat` in the companion software P&R contains a record of 400 pieces of data. The distribution of the population from which the data was drawn can be identified as follows.

S1. Use P&R to plot the empirical PDF (i.e., histogram of the PDF type), as shown in Fig. 5.3. It can be observed that the empirical PDF resembles the PDF of a Weibull or Rayleigh RV. Since Weibull includes Rayleigh as a special case, let us assume it is a Weibull RV.

S2. Use P&R to compute the sample mean and sample variance of the data as $\hat{x} = 0.8891$ and $\hat{v} = 0.2299$. From Table 3.1, it is found that the parameters $a \approx 1.946$ and $b \approx 0.9930$.

S3. Use P&R to overlay the empirical PDF and the Weibull PDF with $a = 1.946$ and $b = 0.993$, as shown in Fig. 5.4. They match quite well and thus it can be concluded that the population is probably Weibull distributed with parameters $a = 1.946$ and $b = 0.993$.

In fact, the data was drawn from a population that is Weibull distributed with parameters $a = 2$ and $b = 1$ (i.e., Rayleigh distributed).

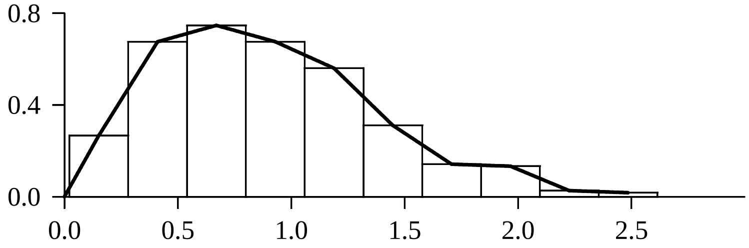


Figure 5.3: Empirical PDF of data `e5_4.dat`.

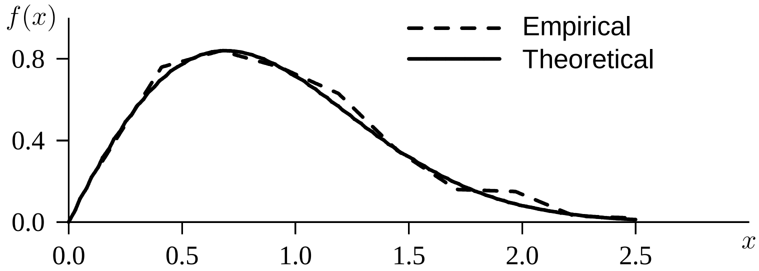


Figure 5.4: Comparison of the empirical and theoretical PDFs.

5.4 Statistical Inference

Statistical inference consists of two parts: Estimation and decision making concerning some unknown parameters of the population.

- ***Estimation*** provides an approximate value of the parameter that is close to the true value.
- ***Decision*** or ***hypothesis testing*** decides whether a given or hypothesized value of the parameter should be rejected as the true value or not.

Example 5.5: Estimation of Mean and Variance of a Gaussian Population

The mean and variance of a population $X \sim \mathcal{N}(\bar{x}, \sigma^2)$ are not known, where X is the age of a member of a professional society. A random sample $(X_1, \ldots, X_n)$ of X is available. Many estimates of $\bar{x}$ and σ^2 are possible. For example, the sample mean and sample variance can be their estimates, respectively, that is, $\bar{x} \approx \hat{X}, \sigma^2 \approx \hat{V}$. $\bar{x}$ and σ can also be estimated by $\bar{x} \approx X_m$ and $\sigma \approx R/d_n$, where X_m is the ***sample median*** and R is the ***sample range***, defined by

$$X_m = \text{middle value of the sample} = \left\{\begin{matrix} X_{i+1} & n \text{ odd} \\ \frac{1}{2}(X_i + X_{i+1}) & n \text{ even} \end{matrix}\right\}, \quad X_i \leq X_{i+1}$$

$$R = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$$

and $d_n \approx n(n - \frac{1}{2})^{-1/2}$ is a constant. If the sample has the values: 39, 41, 55, 34, 52, 45, 36, then we have

$$\hat{x} = 43.14, \quad \hat{v} = 63.12, \quad x_m = 41, \quad r = 22 \Longrightarrow (r/d_n)^2 = 64.20$$

Note that $\hat{x} \approx x_m$ and $\hat{v} \approx (r/d_n)^2$.

Example 5.6: Hypothesis Testing on Mean of Measurement Error

The measurement error of a device has variance $\sigma^2 = 0.0004$ and an unknown mean $\bar{x}$. Suppose we hypothesize that $\bar{x} = 0$ (i.e., no bias). Should we reject this hypothesis given a sample $(0.01, -0.06, -0.09, 0.04, -0.05, 0.08, -0.03, 0.07)$ of the measurement error? This is the problem of hypothesis testing. Clearly, it would not work by simply comparing the hypothesized value with an estimate of the mean (which cannot be equal in general).

5.5 Parameter Estimation

Many methods are available for parameter estimation. We shall focus on two of them: the maximum likelihood method and the method of moments.

The basic idea of the ***maximum likelihood method*** is the following. If an event occurs in a single observation, then we can reasonably assume that it has a large probability — its likelihood is large. As such, the value of the unknown parameter that is most likely to have produced that event (i.e., that particular sample of data) may be used as the estimate of the parameter.

The ***likelihood function*** $L(x_1, \ldots, x_n; \theta)$ of a parameter θ given a sample $(X_1, \ldots, X_n)$ is the joint PDF of $(X_1, \ldots, X_n)$ pretending that the parameter is known. It is in general a function of the parameter. The maximum likelihood estimate (MLE) of θ is the maximum point (i.e., the peak location — θ value of the peak) of the likelihood function $L(x_1, \ldots, x_n; \theta)$.

Example 5.7: Maximum Likelihood Estimation of Failure Rate

The time X to failure of a system is an exponentially distributed RV with PDF $f(x) = \lambda e^{-\lambda x} u(x)$. However, the failure rate λ is unknown. Given a sample $(X_1, \ldots, X_n)$, we use the maximum likelihood method to estimate λ.

The likelihood function is, since $X_1, \ldots, X_n$ are independent,

$$\begin{aligned} L(x_1, \ldots, x_n; \lambda) &= f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\text{given } \lambda) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i) = \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n} u(x_1) \cdots u(x_n) \\ &= \lambda^n e^{-\lambda(x_1 + \cdots + x_n)} u(x_1) \cdots u(x_n) = \lambda^n e^{-\lambda n \hat{x}} \prod_{i=1}^{n} u(x_i) \\ \ln L &= n(\ln \lambda - \hat{x}\lambda) + \sum_{i=1}^{n} \ln u(x_i) \end{aligned}$$

where $\hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Clearly, L and $\ln L$ achieve their maximum at the same value of λ, which is obtained easily by setting $\frac{d}{d\lambda} \ln L = 0$ as

$$\hat{\lambda} = 1/\hat{X}$$

That is, the maximum likelihood estimate (MLE) of the failure rate is the reciprocal of the sample mean.

The underlying idea of the ***method of moments*** is that the sample moment should be close to the true moment. Thus, the value of the unknown parameter that makes the moments equal can be used as an estimate of the parameter. Usually, a lower (e.g., the first or second) moment is preferred.

Example 5.8: Method of Moments Estimation of Failure Rate

We now use the method of moments to estimate the failure rate λ of the above example. For this example we will use the first moment. Since the true mean (expected value) of the exponential RV X is equal to $1/\lambda$ (see Example 3.21), by letting

$$1/\lambda = \text{true mean} = \text{sample mean} = \hat{X}$$

we have $\hat{\lambda} = 1/\hat{X}$, which turns out to be equal to the MLE.

Example 5.9: Maximum Likelihood and Method of Moments Estimation

Consider a sample $(X_1, \ldots, X_n)$ of a population X with the following PDF:

$$f(x) = (\theta+1)x^\theta, \qquad 0 < x < 1, \theta > -1$$

where θ is an unknown parameter to be estimated. The likelihood function is

$$L(x_1, \ldots, x_n; \theta) = (\theta+1)x_1^\theta \cdots (\theta+1)x_n^\theta = (\theta+1)^n (x_1 \cdots x_n)^\theta$$
$$\ln L = n \ln(\theta+1) + \theta(\ln x_1 + \cdots + \ln x_n)$$

Setting $\frac{d}{d\theta} \ln L = 0$ leads to

$$\frac{n}{\theta+1} + \sum_{i=1}^n \ln x_i = 0$$

Solving this equation yields the maximum likelihood estimate

$$\hat{\theta} = -\Big[\frac{n}{\sum_{i=1}^n \ln X_i} + 1\Big]$$

Note that

$$\bar{x} = \int_0^1 x(\theta+1)x^\theta dx = \frac{\theta+1}{\theta+2} x^{\theta+2}\Big|_0^1 = \frac{\theta+1}{\theta+2}$$

Let $\bar{x} = \hat{x}$. Then the method of moments estimate of θ is $\hat{\theta} = \frac{2\hat{X}-1}{1-\hat{X}}$, which is different from the maximum likelihood estimate.

5.6 Hypothesis Testing

In ***hypothesis testing***, we have one or more hypotheses about the values of some unknown parameters of the population, and we want to decide whether the information contained in the sample supports or rejects the hypotheses.

Consider testing the following single hypothesis on an unknown parameter θ of the population

$$H_0 : \theta = \theta_0$$

where θ_0 is a given constant. Clearly, there are two possible decision errors (mistakes):

- ***Type I error***: We decide that H_0 is false but in fact it is true.
- ***Type II error***: We decide that H_0 is true but in fact it is false.

The type I error probability α is known as the ***significance*** and $(1 - \alpha)$ the ***confidence*** of the test.

Example 5.11: Instrument Calibration — Test on Population Mean

An instrument makes a random measurement error X that is Gaussian distributed with zero mean: $X \sim \mathcal{N}(0, 0.01)$, if it is well calibrated according to the product specification. It has been calibrated some time ago and we want to decide if further calibration is needed by checking if the mean of its measurement error can be accepted as zero. That is, our hypothesis is H_0: $\bar{x} = 0$. We have obtained the following sample of the measurement error:

$$0.1294, -0.0336, 0.1714, 0.2624, 0.0308, 0.1858, 0.2254, -0.0594, -0.044, 0.157$$

Since the sample mean $\hat{X}$ has $\mathcal{N}(\bar{x}, \sigma^2/n) = \mathcal{N}(0, 0.01/10)$ distribution, from Example 3.38, it should be within the interval $(-1.96\sigma/\sqrt{10}, 1.96\sigma/\sqrt{10}) = (-0.062, 0.062)$ with 95% probability if H_0 is true. Based on the above sample, the sample mean $\hat{X} = 0.1025$ is outside the interval. Thus, we should reject H_0 with 95% confidence. After further calibration, we have a sample:

$$-0.040, 0.069, 0.0816, 0.0712, 0.129, 0.0669, 0.1191, -0.1202, -0.002, -0.0157$$

The sample mean $\hat{X} = 0.0359$ is in the interval and thus we should accept H_0 with 95% confidence and no more calibration is needed.

5.7 Linear Regression and Curve Fitting

Regression techniques are statistical tools that handle the *statistical relation* between two or more variables.

Consider two RVs X and Y. Assume they are related by

$$Y = a + bX + V$$

where $V \sim \mathcal{N}(0, \sigma^2)$ and the coefficients a and b do not depend on X. The problem of ***linear regression*** is to find the estimates $\hat{a}$ and $\hat{b}$ of a and b given a sample (data) $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of the pair of RVs (X, Y).

Geometrically, the sample can be plotted in a scatter diagram, as illustrated in Fig. 5.5. The problem of regression is then that of ***curve fitting*** — find a curve (or a straight line for *linear* regression) that best fits the data points. Assume the fitted straight line is given by

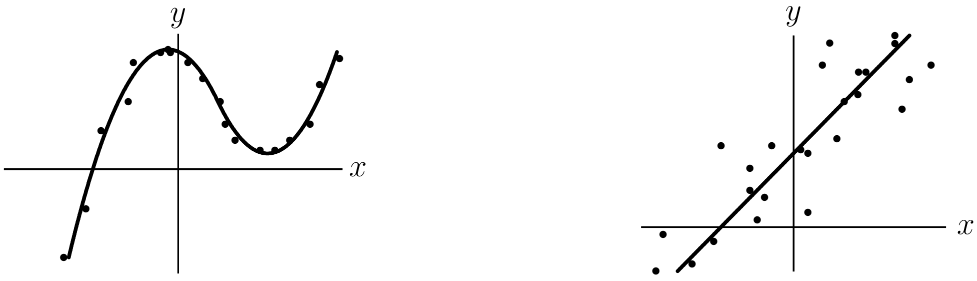
$$\check{Y} = \hat{a} + \hat{b}X$$

It can be shown that $\hat{a} = \hat{y} - \hat{b}\hat{x}$, where $\hat{x}$ and $\hat{y}$ are the values of the sample means $\hat{X}$ and $\hat{Y}$. Thus

$$\check{Y} = \hat{y} + \hat{b}(X - \hat{x}) \tag{5.7}$$

It can be seen that the regression line passes through the centroid $(\hat{x}, \hat{y})$ of the data points in the scatter diagram. It can be shown that least squares and maximum likelihood estimation both lead to

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^{n}(x_i - \hat{x})^2} \tag{5.8}$$



(a) Nonlinear regression

(b) Linear regression

Figure 5.5: Curve fitting by linear or nonlinear regression.

Example 5.14: Linear Regression by P&R

A sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size 50

x_i	y_i
1.9140	11.168
1.1175	7.6677
$\vdots$	$\vdots$
1.8757	11.455

of the pair of RVs (X, Y) is given in data file `e5_14.dat` in the companion software P&R. We have

$$\begin{aligned}\hat{x} &= [1.9140 + 1.1175 + \cdots + 1.8757]/50 = 1.6228\\ \hat{y} &= [11.168 + 7.6677 + \cdots + 11.455]/50 = 8.6861\\ \hat{b} &= \frac{[(1.9140)(11.168) + \cdots + (1.8757)(11.455)] - 50(1.6228)(8.6861)}{[1.9140^2 + \cdots + (1.8757)^2] - 50(1.6228)^2}\\ &= 2.9536\\ \hat{a} &= 8.6861 - (2.9536)(1.6228) = 3.8931\end{aligned}$$

Consequently, the linear regression of Y on X is given by

$$\check{Y} = 3.8931 + 2.9536X$$

In fact, the above regression equation can be obtained by P&R as follows. Following the procedure described in Example 4.8, the correlation of the data in the file `e5_14.dat` can be obtained. The results are

$$\hat{x} = 1.6228, \quad \hat{y} = 8.6861, \quad \hat{v}_x = 1.616, \quad \hat{v}_y = 15.7545, \quad \rho = 0.94596$$

From the analysis on page 164, we have $\frac{(\check{Y}-\bar{y})}{\sqrt{\hat{v}_y}} = \rho\frac{(X-\bar{x})}{\sqrt{\hat{v}_x}}$. Comparing it with (5.7) yields

$$\hat{b} = \rho\sqrt{\hat{v}_y/\hat{v}_x} = (0.94596)(\sqrt{15.7545})/(\sqrt{1.616}) = 2.9536$$

Consequently, the linear regression of Y on X is given by

$$\check{Y} = \hat{y} + \hat{b}(X - \hat{x}) = 8.6861 + 2.9536(X - 1.6228) = 3.8931 + 2.9536X$$

If $X = 4.294$, the best guess of Y is $3.8931 + (2.9536)(4.294) = 16.5759$.