Section 10.5

Confidence Intervals

置信区间

Confidence Intervals

For an arbitrary random variable Y and constant $c > 0$,

$$\mathsf{P}\left[|Y - \mu_Y| \geq c\right] \leq \frac{\mathsf{Var}[Y]}{c^2}.$$

- $$\mathsf{P}\left[|M_n(X) - \mu_X| < c\right] \geq 1 - \frac{\mathsf{Var}[X]}{nc^2} = 1 - \alpha \qquad (1)$$

  Equation (10.35) contains two inequalities.

- One inequality,

  $$|M_n(X) - \mu_X| < c, \qquad (2)$$

  defines an event.

- This event states that the sample mean is within $\pm c$ units of the expected value.

- The length of the interval that defines this event, $2c$ units, is referred to as a *confidence interval*.

- The other inequality states that the probability that the sample mean is in the confidence interval is at least $1 - \alpha$.

- We refer to the quantity $1 - \alpha$ as the *confidence coefficient*.

- If α is small, we are highly confident that $M_n(X)$ is in the interval $(\mu_X - c, \mu_X + c)$.

Example 10.7 Problem

Suppose we perform n independent trials of an experiment and we use the relative frequency $\hat{P}_n(A)$ to estimate $\mathrm{P}[A]$. Find the smallest n such that $\hat{P}_n(A)$ is in a confidence interval of length 0.02 with confidence 0.999.

$$\mathrm{P}\left[|M_n(X) - \mu_X| < c\right] \geq 1 - \frac{\mathrm{Var}[X]}{nc^2} = 1 - \alpha$$

$$\Downarrow$$

$$\mathrm{P}\left[\left|\hat{P}_n(A) - \mathrm{P}[A]\right| < c\right] \geq 1 - \frac{\mathrm{P}[A](1 - \mathrm{P}[A])}{nc^2} \qquad p(1-p) \leq 0.25$$

Example 10.7 Solution

Recall that $\hat{P}_n(A)$ is the sample mean of the indicator random variable X_A. Since X_A is Bernoulli with success probability $\mathsf{P}[A]$, $\mathsf{E}[X_A] = \mathsf{P}[A]$ and $\mathsf{Var}[X_A] = \mathsf{P}[A](1-\mathsf{P}[A])$. Since $\mathsf{E}[\hat{P}_n(A)] = \mathsf{P}[A]$, Theorem 10.5(b) says

$$\mathsf{P}\left[\left|\hat{P}_n(A) - \mathsf{P}[A]\right| < c\right] \geq 1 - \frac{\mathsf{P}[A](1-\mathsf{P}[A])}{nc^2}. \tag{1}$$

In Example 10.6, we observed that $p(1-p) \leq 0.25$ for $0 \leq p \leq 1$. Thus $\mathsf{P}[A](1-\mathsf{P}[A]) \leq 1/4$ for any value of $\mathsf{P}[A]$ and

$$\mathsf{P}\left[\left|\hat{P}_n(A) - \mathsf{P}[A]\right| < c\right] \geq 1 - \frac{1}{4nc^2}. \tag{2}$$

For a confidence interval of length 0.02, we choose $c = 0.01$. We are guaranteed to meet our constraint if

$$1 - \frac{1}{4n(0.01)^2} \geq 0.999. \tag{3}$$

Thus we need $n \geq 2.5 \times 10^6$ trials.

Example 10.8 Problem

Suppose we perform n independent trials of an experiment. For an event A of the experiment, calculate the number of trials needed to guarantee that the probability the relative frequency of A differs from $\mathsf{P}[A]$ by more than 10% is less than 0.001.

$$\mathsf{P}\left[|M_n(X) - \mu_X| < c\right] \geq 1 - \frac{\mathsf{Var}[X]}{nc^2} = 1 - \alpha$$

Example 10.8 Solution

In Example 10.7, we were asked to guarantee that the relative frequency $\hat{P}_n(A)$ was within $c = 0.01$ of $\mathrm{P}[A]$. This problem is different only in that we require $\hat{P}_n(A)$ to be within 10% of $\mathrm{P}[A]$. As in Example 10.7, we can apply Theorem 10.5(a) and write

$$\mathrm{P}\left[\left|\hat{P}_n(A) - \mathrm{P}[A]\right| \geq c\right] \leq \frac{\mathrm{P}[A](1 - \mathrm{P}[A])}{nc^2}. \tag{1}$$

We can ensure that $\hat{P}_n(A)$ is within 10% of $\mathrm{P}[A]$ by choosing $c = 0.1\,\mathrm{P}[A]$. This yields

$$\mathrm{P}\left[\left|\hat{P}_n(A) - \mathrm{P}[A]\right| \geq 0.1\,\mathrm{P}[A]\right] \leq \frac{(1 - \mathrm{P}[A])}{n(0.1)^2\,\mathrm{P}[A]} \leq \frac{100}{n\,\mathrm{P}[A]}, \tag{2}$$

since $\mathrm{P}[A] \leq 1$. Thus the number of trials required for the relative frequency to be within a certain percentage of the true probability is inversely proportional to that probability.

Example 10.9 Problem

Theorem 10.5(b) gives rise to statements we hear in the news, such as,

> Based on a sample of 1103 potential voters, the percentage of people supporting Candidate Jones is 58% with an accuracy of plus or minus 3 percentage points.

The experiment is to observe a voter at random and determine whether the voter supports Candidate Jones. We assign the value $X = 1$ if the voter supports Candidate Jones and $X = 0$ otherwise. The probability that a random voter supports Jones is $\mathrm{E}[X] = p$. In this case, the data provides an estimate $M_n(X) = 0.58$ as an estimate of p. What is the confidence coefficient $1 - \alpha$ corresponding to this statement?

Example 10.9 Solution

Since X is a Bernoulli (p) random variable, $\mathrm{E}[X] = p$ and $\mathrm{Var}[X] = p(1-p)$. For $c = 0.03$, Theorem 10.5(b) says

$$\mathrm{P}\left[|M_n(X) - p| < 0.03\right] \geq 1 - \frac{p(1-p)}{n(0.03)^2} = 1 - \alpha. \tag{1}$$

We see that

$$\alpha = \frac{p(1-p)}{n(0.03)^2}. \tag{2}$$

Keep in mind that we have great confidence in our result when α is small. However, since we don't know the actual value of p, we would like to have confidence in our results regardless of the actual value of p. Because $\mathrm{Var}[X] = p(1-p) \leq 0.25$. We conclude that

$$\alpha \leq \frac{0.25}{n(0.03)^2} = \frac{277.778}{n}. \tag{3}$$

Thus for $n = 1103$ samples, $\alpha \leq 0.25$, or in terms of the confidence coefficient, $1 - \alpha \geq 0.75$. This says that our estimate of p is within 3 percentage points of p with a probability of at least $1 - \alpha = 0.75$.

Interval Estimates

- A confidence interval estimate of a parameter consists of a range of values and a probability that the parameter is in the stated range.

- If the parameter of interest is r, the estimate consists of random variables A and B, and a number α, with the property

$$\mathsf{P}\left[A \leq r \leq B\right] \geq 1 - \alpha. \tag{1}$$

- In this context, $B - A$ is called the *confidence interval* and $1 - \alpha$ is the *confidence coefficient*.

- Since A and B are random variables, *the confidence interval is random*.

- The confidence coefficient is now the probability that the deterministic model parameter r is in the random confidence interval.

- An accurate estimate is reflected in a low value of $B - A$ and a high value of $1 - \alpha$.

More on Interval Estimates

- In most practical applications of confidence-interval estimation, the unknown parameter r is the expected value $\mathsf{E}[X]$ of a random variable X and the confidence interval is derived from the sample mean, $M_n(X)$, of data collected in n independent trials.

- In this context, Equation (10.35) can be rearranged to say that for any constant $c > 0$,

$$\mathsf{P}\left[M_n(X) - c < \mathsf{E}\left[X\right] < M_n(X) + c\right] \geq 1 - \frac{\mathsf{Var}[X]}{nc^2}. \tag{1}$$

- In comparing Equations (10.45) and (10.46), we see that

$$A = M_n(X) - c, \qquad B = M_n(X) + c, \tag{2}$$

  and the confidence interval is the random interval $[M_n(X) - c, M_n(X) + c]$.

- Just as in Theorem 10.5, the confidence coefficient is still $1 - \alpha$, where $\alpha = \mathsf{Var}[X]/(nc^2)$.

Example 10.10 Problem

Suppose X_i is the ith independent measurement of the length (in cm) of a board whose actual length is b cm. Each measurement X_i has the form

$$X_i = b + Z_i, \tag{1}$$

where the measurement error Z_i is a random variable with expected value zero and standard deviation $\sigma_Z = 1$ cm. Since each measurement is fairly inaccurate, we would like to use $M_n(X)$ to get an accurate confidence interval estimate of the exact board length. How many measurements are needed for a confidence interval estimate of b of length $2c = 0.2$ cm to have confidence coefficient $1 - \alpha = 0.99$?

Example 10.10 Problem

Suppose X_i is the ith independent measurement of the length (in cm) of a board whose actual length is b cm. Each measurement X_i has the form

$$X_i = b + Z_i, \tag{1}$$

where the measurement error Z_i is a random variable with expected value zero and standard deviation $\sigma_Z = 1$ cm. Since each measurement is fairly inaccurate, we would like to use $M_n(X)$ to get an accurate confidence interval estimate of the exact board length. How many measurements are needed for a confidence interval estimate of b of length $2c = 0.2$ cm to have confidence coefficient $1 - \alpha = 0.99$?

$$\mathrm{P}\left[M_n(X) - c < \mathrm{E}\left[X\right] < M_n(X) + c\right] \geq 1 - \frac{\mathrm{Var}[X]}{nc^2}$$

$$\mathrm{P}\left[M_n(X) - 0.1 < b < M_n(X) + 0.1\right] \geq 1 - \frac{1}{n(0.1)^2}$$

Example 10.10 Solution

Since $E[X_i] = b$ and $\text{Var}[X_i] = \text{Var}[Z] = 1$, Equation (10.46) states

$$P\left[M_n(X) - 0.1 < b < M_n(X) + 0.1\right] \geq 1 - \frac{1}{n(0.1)^2} = 1 - \frac{100}{n}. \qquad (1)$$

Therefore, $P[M_n(X) - 0.1 < b < M_n(X) + 0.1] \geq 0.99$ if $100/n \leq 0.01$. This implies we need to make $n \geq 10{,}000$ measurements. We note that it is quite possible that $P[M_n(X) - 0.1 < b < M_n(X) + 0.1]$ is much less than 0.01. However, without knowing more about the probability model of the random errors Z_i, we need 10,000 measurements to achieve the desired confidence.

Theorem 10.14

Let X be a Gaussian (μ, σ) random variable. A confidence interval estimate of μ of the form

$$M_n(X) - c \leq \mu \leq M_n(X) + c$$

has confidence coefficient $1 - \alpha$, where

$$\alpha/2 = Q(c\sqrt{n}/\sigma) = 1 - \Phi(c\sqrt{n}/\sigma).$$

Proof: Theorem 10.14

We observe that

$$\begin{aligned}\mathsf{P}\left[M_n(X) - c \le \mu_X \le M_n(X) + c\right] &= \mathsf{P}\left[\mu_X - c \le M_n(X) \le \mu_X + c\right] \\ &= \mathsf{P}\left[-c \le M_n(X) - \mu_X \le c\right]. \end{aligned} \tag{1}$$

Since $M_n(X) - \mu$ is the Gaussian(0,$\sigma/\sqrt{n}$) random variable,

$$\begin{aligned}\mathsf{P}\left[M_n(X) - c \le \mu \le M_n(X) + c\right] &= \mathsf{P}\left[\frac{-c}{\sigma/\sqrt{n}} \le \frac{M_n(X) - \mu}{\sigma/\sqrt{n}} \le \frac{c}{\sigma/\sqrt{n}}\right] \\ &= 1 - 2Q\left(\frac{c\sqrt{n}}{\sigma}\right). \end{aligned} \tag{2}$$

Thus $1 - \alpha = 1 - 2Q(c\sqrt{n}/\sigma)$.

Example 10.11 Problem

Z_i is a random variable with expected value and standard deviation $\sigma_Z = 1$ cm.

In Example 10.10, suppose we know that the measurement errors Z_i are iid Gaussian random variables. How many measurements are needed to guarantee that our confidence interval estimate of length $2c = 0.2$ has confidence coefficient $1 - \alpha \geq 0.99$?

$$M_n(X) - c \leq \mu \leq M_n(X) + c$$

confidence coefficient $1 - \alpha$, where

$$\alpha/2 = Q(c\sqrt{n}/\sigma) = 1 - \Phi(c\sqrt{n}/\sigma)$$

Example 10.11 Solution

As in Example 10.10, we form the interval estimate

$$M_n(X) - 0.1 < b < M_n(X) + 0.1. \tag{1}$$

The problem statement requires this interval estimate to have confidence coefficient $1 - \alpha \geq 0.99$, implying $\alpha \leq 0.01$. Since each measurement X_i is a Gaussian $(b, 1)$ random variable, Theorem 10.14 says that $\alpha = 2Q(0.1\sqrt{n}) \leq 0.01$, or equivalently,

$$Q(\sqrt{n}/10) = 1 - \Phi(\sqrt{n}/10) \leq 0.005. \tag{2}$$

In Table 4.2, we observe that $\Phi(x) \geq 0.995$ when $x \geq 2.58$. Therefore, our confidence coefficient condition is satisfied when $\sqrt{n}/10 \geq 2.58$, or $n \geq 666$.

Example 10.12 Problem

Y is a Gaussian random variable with unknown expected value μ but known variance σ_Y^2. Use $M_n(Y)$ to find a confidence interval estimate of μ_Y with confidence 0.99. If $\sigma_Y^2 = 10$ and $M_{100}(Y) = 33.2$, what is our interval estimate of μ formed from 100 independent samples?

$$\mathrm{P}\left[M_n(Y) - c \leq \mu \leq M_n(Y) + c\right] = 1 - \alpha$$

$$\alpha/2 = 1 - \Phi\left(\frac{c\sqrt{n}}{\sigma_Y}\right)$$

Example 10.12 Solution

With $1-\alpha = 0.99$, Theorem 10.14 states that

$$\mathrm{P}\left[M_n(Y) - c \leq \mu \leq M_n(Y) + c\right] = 1 - \alpha = 0.99, \tag{1}$$

where

$$\alpha/2 = 0.005 = 1 - \Phi\left(\frac{c\sqrt{n}}{\sigma_Y}\right). \tag{2}$$

This implies $\Phi(c\sqrt{n}/\sigma_Y) = 0.995$. From Table 4.2, $c = 2.58\sigma_Y/\sqrt{n}$. Thus we have the confidence interval estimate

$$M_n(Y) - \frac{2.58\sigma_Y}{\sqrt{n}} \leq \mu \leq M_n(Y) + \frac{2.58\sigma_Y}{\sqrt{n}}. \tag{3}$$

If $\sigma_Y^2 = 10$ and $M_{100}(Y) = 33.2$, our interval estimate for the expected value μ is $32.384 \leq \mu \leq 34.016$.

Quiz 10.5

X is a Bernoulli random variable with unknown success probability p. Using n independent samples of X and a central limit theorem approximation, find confidence interval estimates of p with confidence levels 0.9 and 0.99. If $M_{100}(X) = 0.4$, what is our interval estimate?

$$M_n(X) - c \le p \le M_n(X) + c$$

confidence coefficient $1 - \alpha$, where

$$\alpha/2 = Q(c\sqrt{n}/\sigma) = 1 - \Phi(c\sqrt{n}/\sigma)$$

Joint Random Variable

Joint Cumulative Distribution Function

- Joint cumulative distribution function of random variables X and Y is

$$F_{X,Y}(x, y) = P[X \leq x, Y \leq y]$$

- Properties:

For any pair of random variables, X, Y,

(a) $0 \leq F_{X,Y}(x, y) \leq 1$,
(b) $F_{X,Y}(\infty, \infty) = 1$,
(c) $F_X(x) = F_{X,Y}(x, \infty)$,
(d) $F_Y(y) = F_{X,Y}(\infty, y)$,
(e) $F_{X,Y}(x, -\infty) = 0$,
(f) $F_{X,Y}(-\infty, y) = 0$,
(g) If $x \leq x_1$ and $y \leq y_1$, then

$$F_{X,Y}(x, y) \leq F_{X,Y}(x_1, y_1)$$

Joint Probability Mass Function

- Joint probability mass function of discrete random variables X and Y is
$$P_{X,Y}(x,y) = P[X = x, Y = y]$$
- Probability of the event $\{(X,Y) \in B\}$ is
$$P[B] = \sum_{(x,y)\in B} P_{X,Y}(x,y)$$
- Marginal probability mass function:
$$P_X(x) = \sum_{y\in S_Y} P_{X,Y}(x,y) \qquad P_Y(y) = \sum_{x\in S_X} P_{X,Y}(x,y)$$

$P_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$P_X(x)$
$x=0$	0.01	0	0	0.01
$x=1$	0.09	0.09	0	0.18
$x=2$	0	0	0.81	0.81
$P_Y(y)$	0.10	0.09	0.81	

Joint Probability Density Function

- Joint probability density function of continuous random variables X and Y is

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}, \qquad F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$$

- Properties

(a) $f_{X,Y}(x,y) \geq 0$ for all (x,y), (b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$

- Probability of the event $\{(X,Y) \in B\}$ is

$$f_{X,Y}(x,y) = \begin{cases} c & x \geq 0, y \geq 0,\ x+y \leq 1 \\ 0 & otherwise \end{cases}$$

$$P[B] = \iint_B f_{X,Y}(x,y)dxdy$$

- Marginal probability density function:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Independence, Covariance and Correlation

- Random variable X and Y are independent if and only if

$$\text{Discrete:} \qquad [E]P_{X,Y}(x,y) = P_X(x)P_Y(y)$$

$$\text{Continuous:} \qquad f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

- Covariance of two random variables X and Y is

$$\text{Cov}[X,Y] = \text{E}[(X - \mu_X)(Y - \mu_Y)].$$

$$Cov[X,Y] = E[X \cdot Y] - \mu_x\mu_Y$$

Cov >0, =0, <0.
Independent = uncorrelated ?

- Correlation coefficient of two random variables X and Y is

$$\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\,\text{Var}[Y]}} = \frac{\text{Cov}[X,Y]}{\sigma_X\sigma_Y}.$$

- Correlation of X and Y is

$$r_{X,Y} = E[XY]$$

Expectation

- For random variables X and Y, the expected value of W=g(X,Y) is

$$\text{Discrete:} \qquad [E]\,\mathrm{E}[W] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) P_{X,Y}(x,y)$$

$$\text{Continuous:} \qquad \mathrm{E}[W] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy.$$

- Properties

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{E}[(X-\mu_X)(Y-\mu_Y)].$$

Exercise Problem

Random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} cxy & x = 1,2,3,4;\, y = 1,3 \\ 0 & otherwise \end{cases}$$

1. What is the value of c? Hint: $\sum_{X,Y} P_{X,Y}(x,y) = 1$

2. What is $P[Y<X]$? Hint:

3. What is $P[Y>X]$? Hint:

4. What is P[Y=X]? Hint: Really?

5. Find the marginal PMF $P_X(x)$ and $P_Y(y)$. Hint: $P_X(x) = \sum_{y \in S_Y} P_{X,Y}(x,y)$ $\quad P_Y(y) = \sum_{x \in S_X} P_{X,Y}(x,y)$

6. Determine if X and Y independent. Justify your answer. Hint: $P_{X,Y}(x,y) = P_X(x)P_Y(y)$?

Exercise Problem

Random variables X and Y have the joint PMF

$$P_{X,Y}(x,y) = \begin{cases} cxy & x = 1,2,3,4; y = 1,3 \\ 0 & otherwise \end{cases}$$

1. Find the expected value of W=Y/X?

$$\mathrm{E}[W] = \sum_{x\in S_X} \sum_{y\in S_Y} g(x,y) P_{X,Y}(x,y)$$

2. Find the correlation $r_{X,Y}$

$$r_{X,Y} = E[XY]$$

3. Find covariance Cov[X,Y].

$$Cov[X,Y] = E[X \cdot Y] - \mu_x \mu_Y$$

4. Find the correlation coefficient, $\rho_{X,Y}$.

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}.$$

5. Find the variance Var[X+Y].

$$Var[X+Y] = Var[X] + Var[Y] + 2Cov(X,Y)$$

Bivariate Gaussian Random Variables

- Random variables X and Y have a bivariate Gaussian probability density function if

$$f_{X,Y}(x,y) = \frac{\exp\left[-\frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho_{X,Y}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}{2(1-\rho_{X,Y}^2)}\right]}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}},$$

- Probability density function of random variable X and Y

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}}e^{-(x-\mu_X)^2/2\sigma_X^2}, \qquad f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}e^{-(y-\mu_Y)^2/2\sigma_Y^2}.$$

- Linear combination of Gaussian distribution is still a Gaussian distribution