

Statistics Lecture

Claudio Ortelli ¹

¹Finance Institute
Università della Svizzera italiana

Advanced Learning and Research Institute, 2011

Conditional Distribution and Expectation

Let A and B be two events and $P(B) \neq 0$. The conditional probability of the event A , given that event B is realized, is by definition

$$P(A | B) = \frac{P(A \cap B)}{P(B)}.$$

Let X be a random variable and define B to be the event that $X = x$. The conditional probability $P(A | X = x)$ of the event A is then

$$P(A | X = x) = \frac{P(A \text{ and } X = x)}{P(X = x)} = \frac{P(A \cap [X = x])}{P(X = x)} =$$

provided of course that $P(X = x) \neq 0$.

Conditional Distribution and Expectation

Definition

1) Conditional pmf. Let X and Y be discrete random variables with joint pmf $p(x, y)$. The conditional pmf of Y given X is

$$\begin{aligned} p_{Y|X}(y | x) &= P(Y = y | X = x) \\ &= \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p_X(x)} \end{aligned}$$

if $p_X(x) \neq 0$ and 0 otherwise.

2) The conditional distribution function of a random Variable Y (not necessarily discrete) given a discrete random variable X is

$$F_{Y|X}(y | x) = P(Y \leq y | X = x) = \frac{P(Y \leq y \text{ and } X = x)}{P(X = x)}$$

for all y and all x such that $P(X = x) \neq 0$.

Example

- Server cluster with two servers labeled A and B .
- Incoming jobs are independently routed to A and B with probability p and $q = 1 - p$, respectively.
- The number X of arriving jobs per unit of time is Poisson distributed with intensity λ .
- Determine the number of jobs, Y , received by server A , per unit of time.

$$P_{Y|X}(k, n) = \begin{cases} P_{Y|X}(Y = k, X = n) = \binom{n}{k} p^k q^{n-k}, & 0 \leq k \leq n. \\ 0 & \text{otherwise.} \end{cases}$$

Example

(Continued) Recall that $P(X = n) = e^{-\lambda} \lambda^n / n!$ so that

$$\begin{aligned} p_Y(k) &= \sum_{n=k}^{\infty} p_{Y|X}(k | n) p_X(n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \frac{e^{-\lambda} \lambda^n}{n!} \\ &= \lambda^k p^k e^{-\lambda} \sum_{n=k}^{\infty} \binom{n}{k} \frac{1}{n!} q^{n-k} \lambda^{n-k} \\ &= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{(q\lambda)^{n-k}}{(n-k)!} \end{aligned}$$

so that finally $p_Y(k) = \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{q\lambda} = \frac{(\lambda p)^k}{k!} e^{-\lambda p}$, i.e. Y is Poisson distributed with intensity λp .

Conditional Distribution and Expectation

If X is a continuous random variable then $P(X = x) = 0$ for all $x \in \mathbb{R}$ so that the previous definition $\frac{P(Y=y, X=x)}{P(X=x)}$ of conditional probability is not satisfactory.

However when X and Y are jointly continuous we can define the conditional pdf of Y given X :

Definition

Let X and Y be continuous r.v. with joint pdf $f(x, y)$. The conditional density $f_{Y|X}$ is

$$f_{Y|X}(y | x) = \begin{cases} \frac{f(x, y)}{f_X(x)}, & \text{if } 0 < f_X(x) < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Distribution and Expectation

From the definition of conditional density it follows that

$$f(x, y) = f_{Y|X}(y | x)f_X(x) = f_{X|Y}(x | y)f_Y(y),$$

and if X and Y are independent, then

$$f(x, y) = f_X(x)f_Y(y).$$

Furthermore,

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

which is the continuous analog of the theorem of total probability.

- The conditional pdf can be used to obtain the conditional probability:

$$P(a \leq Y \leq b \mid X = x) = \int_a^b f_{Y|X}(y \mid x) dy, \quad a \leq b.$$

- The conditional distribution function is defined analogously

$$\begin{aligned} F_{Y|X}(y \mid x) &= P(Y \leq y \mid X = x) \\ &= \frac{\int_{-\infty}^y f(x, t) dt}{f_X(x)} \\ &= \int_{-\infty}^y f_{Y|X}(t \mid x) dt. \end{aligned}$$

Example

Consider a series system of two *independent* components with respective lifetime distributions $X \sim \text{EXP}(\lambda_1)$ and $Y \sim \text{EXP}(\lambda_2)$. We are interested in the probability of event A that component 2 causes the system failure, i.e.

$$P(A) = P(X \geq Y).$$

The conditional pdf is $F_{X|Y}(t, t) = P(X \leq t \mid Y = t) = F_X(t)$ by the independence of X and Y . By the total prob. theorem (continuous version)

$$\begin{aligned} P(A) &= \int_0^\infty P(X \geq t, Y = t) f_Y(t) dt \\ &= \int_0^\infty [1 - F_X(t)] f_Y(t) dt = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \end{aligned}$$

Conditional Distribution and Expectation

Exercise

Consider the three-dimensional vector $X = (X_1, X_2, X_3)$ having the following joint density function

$$f_X(x_1, x_2, x_3) = \begin{cases} 6x_1x_2^2x_3, & \text{if } 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, 0 \leq x_3 \leq \sqrt{2}. \\ 0, & \text{otherwise.} \end{cases}$$

- 1 Compute the conditional density functions $f_{X_1, X_2 | X_3}(x_1, x_2 | x_3)$ and $f_{X_3 | X_1}(x_3 | x_1)$.
- 2 Verify if the three random variables X_1, X_2, X_3 are independent.

Exercise

X_1 and X_2 are independent r. v. with Poisson distribution, having respective parameters α_1 and α_2 . Show that the conditional pmf of X_1 given $X_1 + X_2$, $p_{X_1 | X_1 + X_2}(X_1 = x_1 | X_1 + X_2 = y)$, is binomial. Determine its parameters.

Exercise

Let the execution times X and Y of two independent parallel processes be uniformly distributed over $(0, t_X)$ and $(0, t_Y)$, respectively, with $t_X \leq t_Y$. Find the probability that the former process finishes execution before the later.

Conditional Distribution and Expectation

Mixture distributions

- Consider a file server whose workload may be divided into r distinct classes.
- For a job of class i ($1 \leq i \leq r$) the CPU time is exponentially distributed with parameter λ_i .
- Let Y denote the service time of a job and let X be the job class. Then

$$f_{Y|X}(y | i) = \lambda_i e^{-\lambda_i y}, \quad y > 0.$$

- Assume that the probability $p_X(i)$ that a randomly chosen job belongs to class i is equal to $\alpha_i > 0$. It follows $\sum_{i=1}^r \alpha_i = 1$.

The joint density of X and Y is then

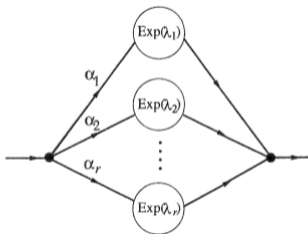
$$f(i, y) = f_{Y|X}(y | i)p_X(i) = \alpha_i \lambda_i e^{-\lambda_i y}, \quad y > 0.$$

Conditional Distribution and Expectation

Mixture distributions

The marginal density of Y is then

$$f_Y(y) = \sum_{i=1}^r f(i, y) = \sum_{i=1}^r \alpha_i \lambda_i e^{-\lambda_i y}, \quad y > 0, \text{ i.e.}$$



Y has an r -stage hyperexponential distribution!

Conditional Distribution and Expectation

Mixture distributions

In general the conditional distribution of Y does not have to be exponential!
Denoting $f_{Y|X}(y | i) = f_{Y_i}(y)$ and $F_{Y|X}(y | i) = F_i(y)$ then the unconditional pdf of Y is

$$f_Y(y) = \sum_{i=1}^r \alpha_i f_i(y)$$

and the unconditional CDF of Y is

$$F_Y(y) = \sum_{i=1}^r \alpha_i F_i(y).$$

Applying the definition of the mean and higher moments we obtain

$$\begin{aligned} E[Y] &= \sum_{i=1}^r \alpha_i E[Y_i], \\ E[Y^k] &= \sum_{i=1}^r \alpha_i E[Y_i^k]. \end{aligned}$$

Conditional Distribution and Expectation

- If X and Y are continuous random variables, we can for instance compute the conditional density $f_{Y|X}$.
- Since $f_{Y|X}$ has all properties of a density function of a continuous random variable, we can talk about its moments.
- Its mean (if exists) is called the conditional expectation of Y given $X = x$ and is denoted $E[Y | X = x]$ or $E[Y | x]$:

$$E[Y | x] = \begin{cases} \int_{-\infty}^{\infty} yf(y | x)dy, & \text{if } 0 < f(x) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

- In case the random variables X and Y are discrete, $E[Y | x]$ is defined as

$$E[Y | X = x] = \sum_y yP(Y = y | X = x) = \sum_y yp_{Y|X}(y | x).$$

Similar arguments hold when X and Y are discrete. The conditional expectation is then defined as

$$E[Y | X = x] = \sum_y y P(Y = y | X = x) = \sum_y y p_{Y|X}(y | x).$$

Definition

The quantity

$$m(x) = E[Y | x]$$

considered as a function of x is known as the *regression function* of Y on X .

Definition

The conditional expectation of a function $\phi(Y)$ is defined as

$$E[\phi(Y) | X = x] = \begin{cases} \int_{-\infty}^{\infty} \phi(y) f_{Y|X}(y | x) dy, & \text{if } Y \text{ is continuous,} \\ \sum_i \phi(y_i) p_{Y|X}(y_i | x), & \text{if } Y \text{ is discrete.} \end{cases}$$

We may take expectation of the regression function to obtain the unconditional expectation of $\phi(Y)$

$$E[\phi(Y)] = \begin{cases} \sum_x E[\phi(Y) | X = x] p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} E[\phi(Y) | X = x] f_X(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

This last formula is known as the **theorem of total expectation**.

Theorem

(Chebyshev) Let X be a random variable with expected value μ and finite variance $\sigma^2 < \infty$. Then, for all $t > 0$, the following inequality holds

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Definition

(Convergence in probability) Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables. We say that the sequence converges in probability to $c \in \mathbb{R}$, write $X_n \xrightarrow{p} c$ or $p\lim X_n = c$, if, for all $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - c| \geq \varepsilon) = 0.$$

Limit Theorems

Theorem

Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables with common expectation μ and finite variance $\sigma_n^2 < \infty$. If $\lim_{n \rightarrow \infty} \sigma_n^2 = 0$, then

$$X_n \xrightarrow{p} \mu.$$

Proof.

Apply Chebyshev inequality. □

Example

Let $\{X_n\}_{n \in \mathbb{N}}$ be an *i.i.d.* $\sim (\mu, \sigma^2)$ (independent and identically distributed) sequence of random variables. Define the sequence $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. From the linearity of the expectation and the properties of the variance we know that $\bar{X}_n \sim (\mu, \frac{\sigma^2}{n})$. The sequence $\{\bar{X}_n\}_{n \in \mathbb{N}}$ converges in probability to μ : $\text{plim } \bar{X}_n = \mu$.

Limit Theorems

The result in the previous Example is also known as the **weak law of large numbers** (WLLN). In order for the WLLN to apply the existence of the second moment (the variance) is not required. The WLLN holds just under the assumption that the $\{X_n\}_{n \in \mathbb{N}}$ i.i.d. sequence have finite expected value μ .

Theorem

Consider two sequences $\{X_n\}_{n \in \mathbb{N}}$ and $\{Y_n\}_{n \in \mathbb{N}}$ of random variables converging in probability to $a < \infty$ and $b < \infty$, respectively. Then

① $p\lim (X_n + Y_n) = p\lim X_n + p\lim Y_n = a + b.$

② $p\lim (X_n \cdot Y_n) = p\lim X_n \cdot p\lim Y_n = a \cdot b.$

③ $b \neq 0,$

$$p\lim \left(\frac{X_n}{Y_n} \right) = \frac{p\lim X_n}{p\lim Y_n} = \frac{a}{b}.$$

④ Function g continuous in a : $p\lim g(X_n) = g(p\lim X_n) = g(a).$

Definition

(Standardization) Let X be a random variable with expected value μ and finite variance σ^2 . The location and scale transform

$$Z = \frac{X - \mu}{\sigma}$$

defines the standardization of X . From the properties of expectation it is straightforward to prove that $Z \sim (0, 1)$.

Example

Let $\{X_n\}_{n \in \mathbb{N}}$ be an independent sequence of random variables with $X_n \sim (\mu, \sigma^2)$ for all $n \in \mathbb{N}$. Define the sequence $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim (0, 1) \text{ for all } n \in \mathbb{N}.$$

Theorem

The Central Limit Theorem (CLT). Let $\{X_n\}_{n \in \mathbb{N}}$ be independent random variables with a finite mean $E[X_n] = \mu_n$ and a finite variance $\text{Var}(X_n) = \sigma_n^2$. Define the normalized random variable

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

so that $E[Z_n] = 0$ and $\text{Var}(Z_n) = 1$ for all n . Then under regularity conditions the limiting distribution of Z_n is standard normal, denoted $Z_n \rightarrow N(0, 1)$, i.e.

$$\lim_{n \rightarrow \infty} F_{Z_n}(t) = \lim_{n \rightarrow \infty} P(Z_n \leq t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Remark: the special condition X_n independent with $\text{Var}(X_n) = \sigma^2$ for all n is sufficient for the CTL to apply.

exercises ...

The object under study is

- 1 the probability distribution function F of a random experiment or random variable X , or
- 2 the statistical distribution function F of a given attribute of a population of individuals, users, devices,

We assume that F is known up to a vector of unknown parameters θ .

Definition

The family of distributions $\mathcal{P} = \{F_\theta\}_{\theta \in \mathbb{R}^n}$, $n \in \mathbb{N}$ finite, is called parametric model. The parametric model is usually specified in terms of probability mass or density functions.

Example

We assume that the number of e-mails per minute arriving to an e-mail server follows a Poisson distribution. The Poisson family of distributions is parametrized by a single parameter $\lambda > 0$

$$\mathcal{P} = \left\{ p_{\lambda}(j) = \frac{\lambda^j}{j!} \exp^{-\lambda}, j = 0, 1, \dots \mid \lambda > 0 \right\}.$$

Example

The delivery time of the Google search engine is normal distributed. The Normal family is parametrized by two parameters $\theta = (\mu, \sigma)$

$$\mathcal{P} = \left\{ f_{\theta}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} \mid \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

Example

In the previous Example, instead of assuming the Normal distribution, we use the Logistic distribution, that is defined by the following distribution function (in this case $\theta = (\mu, \beta)$) :

$$\mathcal{P} = \left\{ F_{\theta}(x) = \frac{1}{1 + \exp^{-(x-\mu)/\beta}} \mid \mu \in \mathbb{R}, \beta > 0 \right\}.$$

Example

We are interested in the percentage of female students at USI. The distribution of the gender attribute of the USI population is Bernoulli distributed with parameter $\theta \in [0, 1]$.

In all previous examples the “true” vector θ of parameters identifying the correct distribution within the corresponding family is *unknown* and must be estimated by means of a sample (a random drawn subset of the population).

Parametric estimation theory deals with the problem of approximating the unknown parameters by means of the information collected from the sample.

Example

In order to estimate the percentage of female students we randomly select n students (sampling with replacement). In this way we obtain n independent realizations of Bernoulli distributed random variables X_i where

$$P(X_i = \text{"female"}) = \theta.$$

Identifying with 1 the female gender, the sample space is composed by 2^n points $x = (x_1, \dots, x_n)$, where $x_i \in \{0, 1\}$ and $p_\theta(x) = \theta^{m(x)}(1 - \theta)^{n - m(x)}$ with $m(x) = \sum_{i=1}^n x_i$. Given a sample x it seems reasonable to estimate θ by the proportion of successes i.e. $m(x)/n$.

Definition

The set of random variables X_1, \dots, X_n is said to constitute a *random sample* of size n from the population with the distribution function $F(x)$, provided that they are *i.i.d.*, i.e.

- mutually independent
- identically distributed

with distribution function $F_{X_i}(x) = F(x)$ for all i and all x .

Point estimation problem

The basic situation in point estimation is as follows.

- We observe a realization of random variables X_1, \dots, X_n .
- The joint distribution function of X_1, \dots, X_n depends on an unknown parameter θ that is known to be in some given set Θ .
- The problem to find the value of θ is a problem of point estimation.
- We estimate θ by some function of the observations x_1, \dots, x_n .

Definition

Any function of the random variables that are being observed, say $\hat{\Theta}(X_1, \dots, X_n)$, is called a *statistic*. It is also called an *estimator of θ* . Since X_1, \dots, X_n are random variables, $\hat{\Theta}$ is a random variable too. A particular realization of the estimator, say $\hat{\Theta}(x_1, \dots, x_n)$, is called an *estimate of θ* and denoted by $\hat{\theta}$.

Point estimation problem

Remark

- The function $\hat{\Theta}$ must be known, that is, it does not have to depend on unknown parameters: $\hat{\Theta} = \bar{X}_n$ is a statistics, $\hat{\Theta} = \bar{X}_n - \mu$ is not a statistics if μ is unknown.
- The distribution of the estimator $\hat{\Theta}$ is called the sampling distribution.
- Since the joint distribution of X_1, \dots, X_n depends on θ , the sampling distribution will depends on θ too.

Example

Let X_1, \dots, X_n be an *i.i.d.* sequence of $N(\mu, 1)$ distributed random variables. From the properties of the Normal distribution we know that the sample distribution of $\hat{\Theta} = \bar{X}_n$ is $N(\mu, \frac{1}{n})$.

In many cases the sampling distribution is an unknown function of θ .

Point estimation problem

We would like our estimator of θ to be exactly θ on average.

Definition

We say that $\hat{\Theta}(X_1, \dots, X_n)$ is an unbiased estimator of θ if

$$E \left[\hat{\Theta}(X_1, \dots, X_n) \right] = \theta.$$

Example

The function

$$\hat{\Theta}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

is an estimator of the population variance. It is possible to show (see Trivedi, page 642) that this estimator is biased, i.e.

$$E \left[\hat{\Theta}(X_1, \dots, X_n) \right] = \sigma^2 - \frac{1}{n} \sigma^2 \neq \sigma^2.$$

Example

The formula

$$\hat{\Theta}(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i$$

is an unbiased estimator of the expected value $\mu = E[X]$ provided that the real weights a_i satisfy the condition $\sum_{i=1}^n a_i = 1$.

- The previous example shows that many unbiased estimators of the same parameter may exist.
- We need a criterion to choose the “best” unbiased estimator.
- Recall that if $\hat{\Theta}$ is unbiased, then from chebyshev's inequality

$$P\left(|\hat{\Theta} - \theta| \geq \varepsilon\right) \leq \frac{\text{Var}[\hat{\Theta}]}{\varepsilon^2} \text{ for } \varepsilon > 0.$$

Definition

Efficiency. An estimator $\hat{\theta}_1$ is said to be a more efficient estimator of the parameter θ than the estimator $\hat{\theta}_2$, provided that

- 1 $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ ;
- 2 $\text{Var}[\hat{\theta}_1] \leq \text{Var}[\hat{\theta}_2]$, for all $\theta \in \Theta$;
- 3 $\text{Var}[\hat{\theta}_1] < \text{Var}[\hat{\theta}_2]$, for some $\theta \in \Theta$.

Example

The sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

is the most efficient (minimum-variance) linear estimator of the population mean μ (provided μ exists). We also say that \bar{X} is BLUE (*Best Linear Unbiased Estimator*) for μ (see Trivedi, page 644).

Point estimation problem

As in the case of the sample mean, the variance of the sampling distribution of an estimator generally decreases with increasing n . This fact leads us to the following property of an estimator.

Definition

Consistency. An estimator $\hat{\theta}$ of the parameter θ is said to be consistent if $\text{plim } \hat{\theta} = \theta$, i.e. if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0.$$

From the Chebyshev inequality we conclude that any unbiased estimator $\hat{\theta}$ of the parameter θ such that

$$\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$$

is consistent.

Example

The empirical distribution function,

$$\hat{F}(x) = \frac{\# \text{ observations } \leq x}{n}$$

is for all $x \in \mathbb{R}$ a consistent estimator of the true distribution function $F(x)$.

Proof: see Trivedi, page 645.

Method of least squares

Starting point:

- We observe a random sample y_1, \dots, y_n from a population with distribution function F .
- We are interested in estimating the mean θ of the distribution.

We rewrite our model as

$$y_i = \theta + \varepsilon_i$$

where $\varepsilon_i := y_i - \theta$ is the zero mean deviation error of the i -th observation from the mean of the underlying distribution. Note that $\text{Var}[\varepsilon_i] = \text{Var}[y_i] = \sigma^2$ for all observations. One possible criterion to estimate θ is to choose an estimate such that the sum of squared errors is as small as possible, i.e.

$$\hat{\theta}(y_1, \dots, y_n) = \min_{\theta} \sum_{i=1}^n \varepsilon_i^2 = \min_{\theta} \sum_{i=1}^n (y_i - \theta)^2.$$

The solution is $\hat{\theta}(y_1, \dots, y_n) = \bar{y}$ the mean of the sample!

Method of least squares

We can extend the argument to the case where the expected value of the random variable Y_i depends on a concomitant **deterministic** and **observable** variable x_i . If the dependence is linear, we have that

$$E[Y_i] = \theta_1 + \theta_2 x_i,$$

where $\theta' = (\theta_1, \theta_2)$ is a vector of unknown parameters. As before, under this further assumptions we can write our model as

$$y_i = \theta_1 + \theta_2 x_i + \varepsilon_i$$

and estimate θ_1 and θ_2 by those number minimizing

$$\min_{\theta} \sum_{i=1}^n \varepsilon_i^2 = \min_{\theta} \sum_{i=1}^n (y_i - \theta_1 - \theta_2 x_i)^2.$$

Method of least squares

Finally, we can generalize the argument to the case of K explanatory variables $x'_i = (x_{i,1}, \dots, x_{i,K})$. This means that, for the i -th observation y_i , the model is

$$y_i = \theta_1 x_{i,1} + \theta_2 x_{i,2} + \dots + \theta_K x_{i,K} + \varepsilon_i$$

$$y_i = x'_i \theta + \varepsilon_i$$

We can rewrite the model in a more compact way using matrix notation

$$\underset{(n \times 1)}{y} = \underset{(n \times K)}{X} \underset{(K \times 1)}{\theta} + \underset{(n \times 1)}{\varepsilon}$$

where $y' = (y_1, \dots, y_n)$, $\varepsilon' = (\varepsilon_1, \dots, \varepsilon_n)$ and

$$X = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,K} \end{bmatrix}.$$

Method of least squares

Again, the estimates of the K parameters $\theta_j, j = 1, \dots, K$ are obtained by minimizing the sum of squared

$$\min_{\theta} \sum_{i=1}^n \varepsilon_i^2 = \min_{\theta} \varepsilon' \varepsilon = \min_{\theta} (y - X\theta)'(y - X\theta).$$

- This method of estimation is called the *method of least squares*.
- Any minimizing value $\hat{\theta}(y)$ is called the *least-square estimate* of θ .
- The function $\hat{\theta} : \mathbb{R}^n \rightarrow \mathbb{R}^K$ is a *least-squares estimator*.

The solution of the minimization problem is obtained by differentiating with respect to θ the quadratic form $(y - X\theta)'(y - X\theta)$ and solving the first order conditions

$$X'X\theta = X'y.$$

Method of least squares

Any solution of the system of equations $X'X\theta = X'y$ is a least-square estimate. If $\text{rank } X = K$, then $X'X$ is non-singular and the unique least-square estimate is

$$\hat{\theta}(y) = (X'X)^{-1}X'y.$$

If $\text{rank } X < K$, then $X'X$ is singular and there is a family of least-square estimates.

Exercise

Assume the following regression model

$$y_i = \theta + \varepsilon_i \quad i = 1, \dots, n.$$

- 1 *Express the model in matrix form and give the dimension and the content of the matrix X .*
- 2 *Show that $\hat{\theta}(y) = (X'X)^{-1}X'y = \bar{y}$.*

Exercise

Assume the following regression model

$$y_i = \theta_1 + \theta_2 x_i + \varepsilon_i \quad i = 1, \dots, n.$$

- ① *Express the model in matrix form and give the dimension and the content of the matrix X in terms of the constant 1 and the explanatory variable x_i , $i = 1, \dots, n$.*
- ② *Show that $\hat{\theta}_1 = \bar{y} - \hat{\theta}_2 \bar{x}$ and $\hat{\theta}_2 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$.*

Maximum likelihood estimation

Suppose that each night a certain lion has three possible states of activity:

- ① very active, denoted by θ_1 ;
- ② moderately active, denoted by θ_2 ;
- ③ lethargic, denoted by θ_3 .

Each night this lion eats i people with probability $p(i|\theta)$, $\theta \in \Theta$ where $\Theta = \{\theta_j, j = 1, 2, 3\}$. The numerical values are given in the following table:

i	0	1	2	3
$p(i \theta_1)$	0.00	0.05	0.05	0.90
$p(i \theta_2)$	0.05	0.05	0.80	0.10
$p(i \theta_3)$	0.90	0.08	0.02	0.00

Question: If we know that $X = i_0$ people were eaten last night, how should we estimate the lion's activity state?

Maximum likelihood estimation

Answer: take as the estimate of θ that $\theta_j \in \Theta$ for which the probability $p(i_0|\theta)$ is largest! In this example we have $\hat{\theta}(i_0 = 0) = \theta_3$, $\hat{\theta}(i_0 = 1) = \theta_3$, $\hat{\theta}(i_0 = 2) = \theta_2$, $\hat{\theta}(i_0 = 3) = \theta_1$.

Suppose that we know for sure that $\theta \in \Theta' = \{\theta_1, \theta_2\}$. The estimate is no longer unique since it can be either θ_1 or θ_2 .

Definition

Let X_1, \dots, X_n be n random variables with distribution function $F(x_1, \dots, x_n|\theta)$ where $\theta \in \Theta$ is unknown. The *likelihood function* is

$$L(\theta) = \begin{cases} f_{X_1, \dots, X_n}(x_1, \dots, x_n|\theta), & \text{if } F \text{ has a density function } f, \\ p_{X_1, \dots, X_n}(x_1, \dots, x_n|\theta), & \text{if } F \text{ has a prob. function } p. \end{cases}$$

Any $\hat{\theta} = \hat{\Theta}(X_1, \dots, X_n) \in \Theta$ such that

$$L(\hat{\theta}) = \sup\{L(\theta) : \theta \in \Theta\}$$

is called *maximum-likelihood estimator* (MLE) of θ .

Maximum likelihood estimation

Remark: Those $\hat{\theta}$'s that maximize the likelihood function $L(\theta)$ are precisely those that maximize the *log-likelihood function* $\log L(\theta)$. It turns out that maximization of the log-likelihood function is often easier.

Example. Suppose we observe X_1, \dots, X_n independent random variables each $N(\theta, \sigma^2)$, with σ^2 known. What is the MLE of $\theta \in \Theta = \mathbb{R}$?

Here

$$L(\theta) = f_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f_{X_i}(x_i | \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2}$$

and it is convenient to take

$$\log L(\theta) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \theta)^2.$$

Maximum likelihood estimation

Now we wish to maximize $\log L(\theta)$ with respect to θ . This is equivalent to minimize $g(\theta) = \sum_{i=1}^n (X_i - \theta)^2$. Since

$$g'(\theta) = \frac{dg(\theta)}{d\theta} = -2 \sum_{i=1}^n (X_i - \theta),$$

$$g''(\theta) = 2 \sum_{i=1}^n 1 = 2n.$$

From the basic course in analysis, we get the MLE solving $g'(\theta) = 0$. We find that the MLE, denoted by $\hat{\theta}$, in this case is $\hat{\theta} = \bar{X}_n$ the sample mean.

Exercise. Suppose we observe X_1, \dots, X_n independent random variables each $N(\mu, \sigma^2)$ with μ known and σ^2 unknown. What is the MLE of $\theta = \sigma^2$?

Maximum likelihood estimation

Remark. The justification of the method of least squares requires no knowledge of the form of the distribution of the error vector apart its mean and variance. In contrast, the method of maximum likelihood is applicable mainly in situations where the true distribution on the sample space is known apart from the values of a finite number of unknown real parameters.

In many applied problems, it is desired to estimate *not the parameter θ itself, but rather a function of the parameter*. The following theorem aids us in this.

Theorem. Invariance principle of maximum-likelihood estimation. *Suppose we have likelihood function $L(\theta) = f(X_1, \dots, X_n | \theta)$, $\theta \in \Theta$. Suppose that $\hat{\theta}$ is a MLE of θ and $g(\cdot)$ a function. Then $g(\hat{\theta})$ is a MLE of $g(\theta)$. This results holds even if θ is a vector.*

Maximum likelihood estimation

Example. Let X_1, \dots, X_n be independent random variables, each Bernoulli with probability of success p . Then $E[X_1] = p, \text{Var}(X_1) = p(1 - p)$. We compute MLE of the mean and the variance. Here

$$L(p) = p^k (1 - p)^{n-k}, \text{ where } k = \sum_{i=1}^n X_i, \text{ or}$$

$$\log L(p) = \sum_{i=1}^n X_i \log(p) + (n - \sum_{i=1}^n X_i) \log(1 - p).$$

Thus

$$\frac{d}{dp} \log L(p) = \frac{\sum_{i=1}^n X_i}{p} - \frac{(n - \sum_{i=1}^n X_i)}{1 - p}, \quad \frac{d^2}{dp^2} \log L(p) = -\frac{\sum_{i=1}^n X_i}{p^2} - \frac{(n - \sum_{i=1}^n X_i)}{(1 - p)^2}.$$

Setting the first derivative equal to zero and solving, we find

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n.$$

Since the second derivative is negative at \hat{p} , this is the MLE of the mean.

Using the previous theorem, the variance's MLE is $g(\hat{p}) = \bar{X}_n(1 - \bar{X}_n)$.

Exercise. In medical applications one often encounters random variables X_i such that the logarithm of X_i has a normal distribution. Let X_1, \dots, X_n be independent random variables each with the same lognormal distribution. Thus $Y_i = \log(X_i) \sim N(\mu, \sigma^2)$. Find the MLE of $E[X_1]$ and $Var(X_1)$.

Exercise. Let X_1, \dots, X_n be a random sample from a $N(\mu, \sigma^2)$ distribution with *both* mean μ and variance σ^2 unknown. Find a MLE of μ and σ^2 , or, equivalently, the MLE of $\theta = (\mu, \sigma^2)$. Find a MLE of σ .

The method of moments

We introduce a third method of estimation known as the **method of moments**. This method, like maximum likelihood, generates possible good estimators.

Definition. Let X_1, \dots, X_n be independent and identically distributed random variables, each with distribution function $F(x \mid \theta)$ for some fixed $\theta \in \Theta \subset \mathbb{R}^r$. If the moments indicated exist, then the **method of moments estimators (MME)** of $\theta_1, \dots, \theta_r$ are the solutions of the equations

$$\mu'_k = m'_k, \text{ for } k = 1, 2, \dots, r.$$

Thus, the method of moments attempts to equate the first r population moments to the first r sample moments and take the resulting solution $\hat{\theta}_1, \dots, \hat{\theta}_r$ as an estimator of $\theta_1, \dots, \theta_r$.

The method of moments

Example. Suppose that X_1, \dots, X_n are independent, each being a $N(\mu, \sigma^2)$ random variable with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ both unknown. We compute the method of moments estimators of μ and σ^2 .

In this case $r = 2$ and the equations to be solved are

$$\begin{cases} \mu'_1 = m'_1, \\ \mu'_2 = m'_2 \end{cases}$$

or

$$\begin{cases} \mu = E[X_1] = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n, \\ \sigma^2 + \mu^2 = E[X_1^2] = \frac{\sum_{i=1}^n X_i^2}{n}. \end{cases}$$

Solving this system we get that the MME of (μ, σ^2) are

$$\begin{cases} \hat{\mu} = \bar{X}_n, \\ \hat{\sigma}^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \hat{\mu}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}. \end{cases}$$

The MME are also the MLE. This is often the case.

The method of moments

Example. Although the basic MME technique requires solving as many equations as there are unknown parameters, sometimes this technique fails. Consider the Laplace distribution centered at the origin and with shape parameter $\beta > 0$:

$$f(x | \beta) = \frac{1}{2\beta} \exp(-|x|/\beta), \quad x \in \mathbb{R}.$$

If X has this distribution function, then $E[X] = 0$. If X_1, \dots, X_n is a random sample of size n from this distribution, then the equation $E[X] = \bar{X}_n$ does not yield an estimate of the single unknown parameter β .

However, we can find that $E[X^2] = 2\beta^2$, which is a function of β . Hence, setting $E[X^2] = \sum_{i=1}^n X_i^2/n$ yields

$$\hat{\beta} = \left(\frac{1}{2} \sum_{i=1}^n \frac{X_i^2}{n} \right)^{1/2}.$$

Exercise. Check that in this case the MLE is not the same as the MME, but rather

$$\beta^* = \sum_{i=1}^n |X_i|/n.$$

Test of hypothesis

A hypothesis-testing problem usually results from questions like

- Does smoking cause cancer?
- Do seat belts reduce car accident injuries?
- Do atomic power plants increase radiation levels?

Thus, we have an underlying parameter θ (in the examples above, this could be a cancer rate, an injury rate, or a radiation level) and we wish to determine whether it changes in specified ways when an element of a system is changed (i.e. smoking, seat belts, or atomic power plants are added).

Formally, let us consider X a random variable with distribution function $F(x | \theta)$ where $\theta \in \Theta$ is unknown (X and θ may be multivariate).

Test of hypothesis

Example. Let us now talk in terms of the question “Do atomic power plants increase radiation levels?”. Then, a long history of precise study has shown that to date the average level of radiation per person is $\theta_0 = 1.2$ units on some standard scale in the area within a 25-km radius of a proposed model atomic power plant site. Based on the study, we can assume that a person chosen at random has a level $X_1 \sim N(\theta_0, \sigma^2)$ of radiation per year, where $\theta_0 = 1.2$ and $\sigma^2 = 0.09$.

If n persons' radiation exposure is measured in this area next year, we will observe independent random variables X_1, \dots, X_n that are each $N(\theta, \sigma^2)$, where: $\theta = \theta_0$ if an atomic power plant *does not* operate in the area; $\theta \in \Theta = \mathbb{R}$ unknown, if an atomic power plant *does* operate in the area.

The hypothesis that the atomic power plant has no effect (or a null effect) is called **null hypothesis H_0** , and is the assertion that $\theta = \theta_0$. The **alternative hypothesis H_1** that it has an effect (either of increase or of decrease) is the assertion that $\theta \neq \theta_0$ (two-sided alternative hypothesis: $\theta > \theta_0$ or $\theta < \theta_0$).

Test of hypothesis

From the above example, we note that **each hypothesis is a subset of Θ** . It also useful to distinguish whether, for any specified hypothesis $H \subseteq \Theta$, the distribution of X is completely determined or not.

Definition. Suppose that X is random variable with distribution function $F(x | \theta)$ where $\theta \in \Theta$ is unknown. Any $H \subseteq \Theta$ is called a **hypothesis**. Often one hypothesis H_0 is called **null hypothesis**, and another hypothesis H_1 is called **alternative hypothesis**.

Moreover, a hypothesis $H \subseteq \Theta$ is called **simple** if knowing $\theta \in H$ completely specifies $F(x | \theta)$. Otherwise, H is called a **composite hypothesis**.

Example. In the above example, H_0 is a simple null hypothesis, whereas H_1 is a composite alternative hypothesis. Note that it is not the case that H_0 and H_1 are unique for a particular problem. Thus, the alternative hypothesis in the atomic power plant example could just as well be taken to be $H_1^* = \{\theta \in \Theta : \theta > \theta_0\}$. This would be an alternative of increase instead of an alternative of change (+ or -).

Exercises

1. In the following hypothesis testing problems, $X \sim N(\mu, \sigma^2)$ and $\sigma^2 > 0$ is known. Determine whether each hypothesis is simple or composite.
 - (i) $H : \mu = \mu_0$, μ_0 a known constant.
 - (ii) $H : \mu > \mu_0$, μ_0 a known constant.
 - (iii) $H : \mu \neq \mu_0$, μ_0 a known constant.
 - (iv) $H : \mu \in I$, where I is a specified interval.

2. Are the following hypotheses simple or composite?
 - (i) $H : X$ has an exponential distribution with some parameter θ .
 - (ii) $H : F_X(x | \theta) = F_X(x | \theta_0)$, θ_0 a fixed known number.
 - (iii) When $X \sim N(\mu, \sigma^2)$ with μ and σ^2 both unknown, $H : \mu = \mu_0$, where μ_0 a fixed known number.
 - (iv) When $X \sim N(\mu, \sigma^2)$ with μ and σ^2 both unknown, $H : \{\mu = \mu_0, \sigma^2 = \sigma_0^2\}$, where μ_0 and σ_0^2 fixed known numbers.

Hypothesis testing problem

Let X be a random variable with distribution function $F(x \mid \theta)$ for an unknown parameter $\theta \in \Theta$. Suppose we have the null hypothesis H_0 and an alternative hypothesis H_1 . Our inference as to whether H_0 or H_1 contains the true parameter θ will be based on observing X . So, for some outcomes $X = x$ we will decide $\theta \in H_0$ and for some others $\theta \in H_1$.

Definition. The problem of deciding (after observing X) either that $\theta \in H_0$ (called *accepting the null hypothesis*) or that $\theta \in H_1$ (called *rejecting the null hypothesis*) is called a *hypothesis-testing problem*. H_0 is called *true* if $\theta \in H_0$, and *false* if $\theta \in H_1$.

If the set of possible outcomes of X is \mathcal{H} , then a division of \mathcal{H} into two disjoint and exhaustive subsets A called the *acceptance region* and R called the *critical region* such that if we find $X \in A$ we accept H_0 and if we find $X \in R$ we reject H_0 , is called a *test of H_0 versus H_1* .

Hypothesis testing problem

When we conduct a test in a hypothesis-testing problem, **two errors** are possible:

- We can reject the null hypothesis H_0 when it is true;
- we can accept the null hypothesis H_0 when it is false.

These errors are of central importance and have special names.

Definition. *In a hypothesis testing problem, rejecting H_0 when it is true is called an **error of type I**. The probability of rejecting a true H_0 depends on the test employed and the true $\theta \in H_0$, and is called the **level (of significance) of the test** or the **size of the critical region**.*

*Accepting H_0 when it is false is called an **error of type II**. The probability of rejecting a false H_0 depends on the test employed and the true $\theta \in \Theta \cap \overline{H_0}$, and is called the **power of the test**. Note that the power is defined for all $\theta \in \Theta$ that are not in H_0 ; this is often a larger set than H_1 .*

Hypothesis testing problem

Example. Consider again the atomic power plant example. There we observe $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{H} = \mathbb{R}^n$, where the X_i 's are independent random variables each $N(\theta, \sigma^2)$ with $\theta \in \mathbb{R}$ unknown and $\sigma^2 = 0.09$. Our null hypothesis is $H_0 = \{\theta_0\}$ where $\theta_0 = 1.2$, and our alternative hypothesis is $H_1 = \{\theta \in \mathbb{R}, \theta \neq \theta_0\}$. In our decision to accept or reject H_0 we will need to decide, for each possible value $\mathbf{x} = (x_1, \dots, x_n)$ of \mathbf{X} whether $\mathbf{x} \in A$ or $\mathbf{x} \in R$.

A reasonable way to do this would be to base our decision on a sufficient statistic for the problem. For example, we know that \bar{X}_n is sufficient in this case (and also a BUE of θ). So it may make sense to compare \bar{X}_n with θ_0 in order to make our acceptance or rejection decision.

As a test, we take the one that rejects H_0 if and only if $|\bar{X}_n - \theta_0| > a$. We set a such that the level of the test equals $\alpha = 0.05$. That is,

$$\begin{aligned} P_{\theta_0} \left[|\bar{X}_n - \theta_0| > a \right] &= 1 - P_{\theta_0} \left[|\bar{X}_n - \theta_0| \leq a \right] = 1 - P_{\theta_0} \left[-a \leq \bar{X}_n - \theta_0 \leq a \right] \\ &= 1 - P_{\theta_0} \left[-\frac{a}{\sigma/\sqrt{n}} \leq Y \leq \frac{a}{\sigma/\sqrt{n}} \right] \\ &= 1 - \left[\Phi(a\sqrt{n}/\sigma) - \Phi(-a\sqrt{n}/\sigma) \right] = 2 \left[1 - \Phi(a\sqrt{n}/\sigma) \right], \end{aligned}$$

where Y is a $N(0, 1)$ random variable and $\Phi(z) = P[Y \leq z]$.

Hypothesis testing problem

To obtain level $\alpha = 0.05$ for the test, we set

$$0.05 = 2\left[1 - \phi(a\sqrt{n}/\sigma)\right] \iff \phi\left(a\frac{\sqrt{n}}{\sigma}\right) = 0.975 \iff a\frac{\sqrt{n}}{\sigma} = 1.96$$

and $a = 0.588/\sqrt{n}$. So if, for example, $n = 100$, then $a = 0.0588$; that is, we reject H_0 if and only if $|\bar{X}_n - \theta_0| > 0.0588$.

Similarly, we compute the power of the test for $\theta \in \Theta \cap \bar{H}_0$, i.e. for $\theta \neq \theta_0$,

$$\begin{aligned} P_\theta[|\bar{X}_n - \theta_0| > a] &= 1 - P_\theta[-a \leq \bar{X}_n - \theta_0 \leq a] \\ &= 1 - P_\theta[-a + \theta_0 - \theta \leq \bar{X}_n - \theta \leq a + \theta_0 - \theta] \\ &= 1 - P_\theta\left[\frac{-a + \theta_0 - \theta}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \theta}{\sigma/\sqrt{n}} \leq \frac{a + \theta_0 - \theta}{\sigma/\sqrt{n}}\right] \\ &= 1 - \left[\phi\left(\frac{a + \theta_0 - \theta}{\sigma/\sqrt{n}}\right) - \phi\left(\frac{-a + \theta_0 - \theta}{\sigma/\sqrt{n}}\right)\right] \\ &= 1 - \left[\phi\left(1.96 + \frac{\theta_0 - \theta}{\sigma}\sqrt{n}\right) - \phi\left(-1.96 + \frac{\theta_0 - \theta}{\sigma}\sqrt{n}\right)\right]. \end{aligned}$$

As a function of θ , say $f(\theta)$, the rejection probability has the properties: $f(\theta_0) = \alpha = 0.05$; $f(\theta_2) > f(\theta_1)$ for all $\theta_2 > \theta_1 \geq \theta_0$; and $\lim_{\theta \rightarrow +\infty} f(\theta) = 1$.

Hypothesis testing problem

Exercises

1. In the example of the atomic power plant, consider as a second possible test the test that rejects H_0 if and only if $\bar{X}_n > \theta_0 + b$. Find b such that the level of the test equals $\alpha = 0.05$. Compute the power of the test for a given $\theta \neq \theta_0$.
2. The manufacturer of an imported automobile claims that the average miles per gallon (mpg) of this type of car is at least 30. Suppose that the mpg of a randomly selected car is a normally distributed random variable with mean μ and variance $\sigma^2 = 25$. To check the claim, we consider the hypothesis-testing problem $H_0 : \mu = 30$ versus $H_1 : \mu > 30$. We take a random sample of 36 cars from this manufacturer's output, find mpg for each of the cars, and then compute \bar{X}_{36} of the mpg's of the 36 cars. Suppose it is decided to reject H_0 if and only if $\bar{X}_n \geq 32$. Compute the size of the critical region (i.e. α), the power of the test, and the probability $\beta \equiv (1 - \text{power})$ of type II error at $\mu > 30$.
3. Consider testing the simple null hypothesis $H_0 : \theta = \theta_0$ versus the simple alternative hypothesis $H_1 : \theta = \theta_1$, where θ is the mean of a normal population with known variance σ^2 . Suppose we take a random sample of size n (to be determined) and use the test that rejects H_0 if and only if $\bar{X}_n > c$. How large an n is needed to obtain probability of type I error of α and power of $1 - \beta$?
Suppose now that $\theta_0 = 5$ and $\theta_1 = 8$. If $\sigma^2 = 4$, find the values of c and n such that $\alpha = 0.05$ and $\beta = 0.05$.

Hypothesis testing problem

Exercise

4. Consider a Poisson distribution with mean $\lambda \in \{2, 3\}$. Consider testing $H_0 : \lambda = 3$ versus $H_1 : \lambda = 2$. A random sample of size 5 is taken and the test calls for rejection of H_0 if and only if $\bar{X}_n < c$. If the probability of type I error α has to be 0.05 (or the closest number you can find in the Poisson table), find the critical region to use.