第10章 期末复习

感谢清华大学自动化系
瑞教授提供PPT

Basics of Statistics

统计学方法及其应用 | 统计学基础 | 随机抽样

Population

- 概念上，总体指研究对象的全体 **K**
- 统计上，总体指与全体相联系的某一数值特征的概率分布 $f(x)$
- 例如
  - 研究全中国儿童的身高
    - 总体为全体中国儿童
    - 因为关心的是身高这一数值特征，总体又指儿童身高的分布
  - 研究降压药物的降压作用
    - 总体为全体高血压病人
    - 因为关心的是血压这一数值特征，总体又指病人血压的分布

Random sample

The random variables $X_1,\ldots,X_n$ are called a **random sample of size $\boldsymbol{n}$ from the population $\boldsymbol{f(x)}$** if $X_1,\ldots,X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1,\ldots,X_n$ are called **independent and identically distributed (iid) random variables with pdf or pmf $\boldsymbol{f(x)}$**. A realization of these random variables, $x_1,\ldots,x_n$, is called **an observation** of the sample $X_1,\ldots,X_n$.

Statistic

Let $X_1,\ldots,X_n$ be a random sample of size n from a population and let $T(x_1,\ldots,x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1,\ldots,X_n)$. Then the random variable or random vector $T=(X_1,\ldots,X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution of $\boldsymbol{Y}$**.

统计量就是样本的函数，应不依赖于总体的参数。统计量的分布称为抽样分布。

Expectation of a random sample

Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}g(X_1)$ and $\mathrm{Var}\,g(X_1)$ exist. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{E}g(X_1),$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{Var}\,g(X_1).$$

Sample mean

The **sample mean** is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \bar{X}(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Sample variance and standard deviation

The **sample variance** is the statistic defined by

$$S^2 = S^2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Since

$$\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.$$

We have

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

The **sample standard derivation** is the statistic defined by

$$S = \sqrt{S^2}.$$

One normal sample

Let $X_1,\ldots,X_n$ be a random sample from a normal (μ,σ^2) population, then

(1) $\overline{X}$ and S^2 are independent random variables.

(2) $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

(3) $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ has a T_{n-1} distribution.

(4) $\displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ has a χ_n^2 distribution.

(5) $\displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2=\frac{(n-1)S^2}{\sigma^2}$ has a χ_{n-1}^2 distribution.

Two normal samples

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be two random samples from two normal populations $N(\mu_X,\sigma_X^2)$ and $N(\mu_Y,\sigma_Y^2)$, respectively.

Assume that $\sigma_X^2=\sigma_Y^2=\sigma^2$, then

(1) $$\frac{\left(\overline{X}-\overline{Y}\right)-\left(\mu_X-\mu_Y\right)}{\sigma\sqrt{1/m+1/n}}$$ has a standard normal distribution.

(2) $$\frac{\left(\overline{X}-\overline{Y}\right)-\left(\mu_X-\mu_Y\right)}{S_p\sqrt{1/m+1/n}}$$ has a T_{m+n-2} distribution,

where $$S_p^2=\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}.$$

(3) $$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2}=\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$ has a $F_{m-1,n-1}$ distribution.

Paired normal sample

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal population with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$, then

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{X-Y} / \sqrt{n}}$$

has a standard normal distribution, and

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_{X-Y} / \sqrt{n}}$$

has a student's t distribution with $n-1$ degrees of freedom,

where $\sigma_{X-Y}^2 = \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \sigma_Y^2$, and

$$S_{X-Y}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[(X_i - Y_i) - (\overline{X} - \overline{Y}) \right]^2.$$

The sufficiency principle

If $T(\mathbf{X})$ is a **sufficient statistic** for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X}$=$\mathbf{x}$ or $\mathbf{X}$=$\mathbf{y}$ is observed.

A sufficient statistic captures **ALL** the information about the parameter contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does **not** contain any more information about the parameter.

Sufficient statistics

A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

$$P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) = \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))}$$

$$= \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))}$$

$$P(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) = \frac{p(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)}$$

Sufficient statistics

Sufficient condition

If $p(\mathbf{x} \mid \theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t \mid \theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio

$$\frac{p(\mathbf{x} \mid \theta)}{q(T(\mathbf{x}) \mid \theta)}$$

is constant as a function of θ.

Factorization theorem

Sufficient and necessary condition

Let $f(\mathbf{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t \mid \theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)h(\mathbf{x}).$$

For exponential family

Sufficient statistic for the exponential family

Let $X_1, \ldots, X_n$ be iid random variables from a pdf or pmf that belongs to an exponential family given by

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right],$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d), d \leq k$. Then,

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

The likelihood function

Let $f(\mathbf{x} \mid \theta)$ denote the joint pmf or pdf of the sample $\mathbf{X} = (X_1, \ldots X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta)$$

is called the **likelihood function**.

The likelihood function measures the plausibility that the sample is observed under a certain parameter. Larger likelihood means the sample that we observed is more likely to have occurred due to the given parameter.

The likelihood principle

If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\theta \mid \mathbf{x})$ is proportional to $L(\theta \mid \mathbf{y})$, that is, there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\theta \mid \mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L(\theta \mid \mathbf{y}) \quad \text{for all } \theta,$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

$$\frac{L(\theta^{(1)} \mid \mathbf{x})}{L(\theta^{(2)} \mid \mathbf{x})} = \frac{C(\mathbf{x}, \mathbf{y}) L(\theta^{(1)} \mid \mathbf{y})}{C(\mathbf{x}, \mathbf{y}) L(\theta^{(2)} \mid \mathbf{y})} = \frac{L(\theta^{(1)} \mid \mathbf{y})}{L(\theta^{(2)} \mid \mathbf{y})}$$

Point Estimation

Introduction

- For a parametric model

$$f(x \mid \theta)$$

- The mathematical structure is already know
- The knowledge of the parameter yield the knowledge of the entire population
- We are interested in obtaining a good estimation of θ, Sometimes an estimation of a function of θ

Point estimator

A **point estimator** is any function $W(X_1,\ldots,X_n)$ of a sample; that is, any statistic is a point estimator.

Estimator: a function of the sample, a random variable.

$$W(X_1,\ldots,X_n)$$

Estimate: the realized value of an estimator, a number.

$$W(x_1,\ldots,x_n)$$

Method of moments

Let $X_1,\ldots,X_n$ be a sample from a population with k parameters $f(x \mid \theta_1,\ldots,\theta_k)$.
Define

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, & \mu'_1 &= \mathrm{E}X^1;\\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, & \mu'_2 &= \mathrm{E}X^2;\\
&\cdots\\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, & \mu'_k &= \mathrm{E}X^k.
\end{aligned}
$$

Solve the following system of equations for $(\theta_1,\ldots,\theta_k)$, in terms of $(m_1,\ldots,m_k)$

$$
\begin{aligned}
m_1 &= \mu'_1(\theta_1,\ldots,\theta_k);\\
m_2 &= \mu'_2(\theta_1,\ldots,\theta_k);\\
&\cdots\\
m_k &= \mu'_k(\theta_1,\ldots,\theta_k).
\end{aligned}
$$

Maximum likelihood estimate

Because a larger likelihood implies a bigger plausibility that a parameter is the true one. It is reasonable to choose the parameter $\theta^\star$ that can maximize the likelihood function $L(\theta \mid \mathbf{x})$ as our best guess of θ.
In other words,

$$\theta^\star = \arg\max_{\theta\in\Theta} L(\theta \mid \mathbf{x}).$$

Equivalently,

$$\theta^\star = \arg\max_{\theta\in\Theta} \log L(\theta \mid \mathbf{x}).$$

Obviously,

$$L(\theta^\star \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}), \text{ for any } \theta \in \Theta.$$

$\theta^\star$ is called the *maximum likelihood estimate* (MLE) of θ.

Maximum likelihood estimators

For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta \mid \mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator** (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

We need to find a **global** maximum!
Need to check boundary conditions!
Sometimes yielding optimization problems with constraints.

Refer to optimization books!

Missing data

- From the viewpoint of maximum likelihood estimation
  - Algorithm: EM algorithm
  - Application: Motif finding
  - Application: Gaussian mixture
- From the viewpoint of Bayesian estimation
  - Algorithm: Gibbs sampling
  - Application: Motif finding

The Basic Setting in EM

- $Y = (X, Z)$
  - Y: complete data ("augmented data")
  - X: observed data ("incomplete" data)
  - Z: hidden data ("missing" data)

- Given a fixed x, there could be many possible z's.
  - Ex: given a sentence x, there could be many state sequences in an HMM that generates x.

The Iterative Approach for MLE

- When missing data is available, it's hard to find the MLE directly

$$\theta_{ML} = \operatorname*{Argmax}_{\theta} \log \left(\sum_{Z} P(X, Z|\theta) \right)$$

- An alternative is to find a sequence

$$\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(t)}, \cdots,$$
$$\text{s.t.} \quad l(\theta^{(0)}) < l(\theta^{(1)}) < \cdots < l(\theta^{(t)}) < \cdots$$

Maximizing the Lower Bound

- The Jensen's inequality gives a lower bound to maximize,

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{Argmax}}\, E_{P(Z|X,\theta^{(t)})}\left[\log P(X, Z|\theta)\right]$$

- Q-function

$$Q(\theta|\theta^{(t)}) = E_{P(Z|X,\theta^{(t)})}\left[\log P(X, Z|\theta)\right]$$

Increasing the Likelihood

- Increasing the likelihood by maximizing the lower bound

$$l(\theta) - l(\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$$

$$Q(\theta^{(t+1)}|\theta^{(t)}) > Q(\theta^{(t)}|\theta^{(t)}) \Rightarrow l(\theta^{(t+1)}) > l(\theta^{(t)})$$

- Which means that a better estimation of the parameter.

Bayes estimators

$$p(\theta \mid \mathbf{x}) = \frac{p(\theta)p(\mathbf{x} \mid \theta)}{p(\mathbf{x})}$$

$$= \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}$$

$p(\theta):$ prior

$p(\mathbf{x} \mid \theta):$ likelihood

$p(\mathbf{x}):$ marginal likelihood (evidence)

$\mathrm{p}(\theta \mid \mathbf{x}):$ posterior

Conjugate prior

Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x \mid \theta)$ (indexed by θ). A class Π of prior distribution is a **conjugate family** of $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.

Binomial vs Beta

- Likelihood
- Prior
- Posterior

Multinomial vs Dirichelet

- Likelihood

$$p(\mathbf{n} \mid \theta) \propto \prod_{k=1}^{m} \theta_k^{n_k}$$

- Prior

$$p(\theta) = \frac{\Gamma(\Sigma_{k=1}^{m} \alpha_k)}{\Pi_{k=1}^{m} \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Posterior

$$p(\theta \mid \mathbf{x}) = \frac{\Gamma(\Sigma_{k=1}^{m} (\alpha_k + n_k))}{\Pi_{k=1}^{m} \Gamma(\alpha_k + n_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k + n_k}$$

Normal vs normal (mean)

- Likelihood

$$p(\mathbf{x} \mid \mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Normal prior

$$p(\mu \mid \xi, \tau^2) \propto \exp\left[-\frac{1}{2\tau^2}(\mu - \xi)^2\right]$$

- Posterior

$$\mu \mid \mathbf{x} \sim N\left(\frac{\sigma^2\xi + \tau^2 n\overline{x}}{\sigma^2 + \tau^2 n}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2 n}\right)$$

Normal vs Gamma (Precision)

- Likelihood

$$p(\mathbf{x} \mid \lambda) \propto \lambda^{\frac{n}{2}} \exp\left[-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Prior: $\mathrm{Gamma}(shape = \alpha, rate = \beta)$

$$p(\lambda \mid \alpha, rate = \beta) \propto \lambda^{\alpha-1}\exp(-\beta\lambda)$$

- Posterior

$$p(\lambda \mid \mathrm{x}; \alpha, \beta) \propto \lambda^{\tilde{\alpha}-1}\exp(-\lambda\tilde{\beta})$$

$$\tilde{\alpha} = \alpha + \frac{n}{2},\ \tilde{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Normal vs Inverse-Gamma (Variance)

- Likelihood

$$p(\mathbf{x} \mid \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Prior :

$$p(\sigma^2 \mid \alpha, \beta) \propto (1 / \sigma^2)^{\alpha+1} \exp(-\beta / \sigma^2)$$

- Posterior

$$p(\sigma^2 \mid \mathrm{x}; \alpha, \beta) \propto (1 / \sigma^2)^{\tilde{\alpha}+1} \exp(-\tilde{\beta} / \sigma^2)$$

$$\tilde{\alpha} = \alpha + \frac{n}{2},\ \tilde{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Normal vs Scaled-inverse-chi-square (Variance)

- Likelihood

$$p(\mathbf{x} \mid \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Prior

$$p(\sigma^2 \mid \eta, \tau^2) \propto (1 / \sigma^2)^{\eta/2+1} \exp[-\eta\tau^2 / (2\sigma^2)]$$

- Posterior

$$p(\sigma^2 \mid \mathrm{x}; \eta, \tau^2) \propto (1 / \sigma^2)^{\tilde{\eta}/2+1} \exp[-\tilde{\eta}\tilde{\tau}^2 / (2\sigma^2)]$$

$$\tilde{\eta} = \eta + n,\ \tilde{\tau}^2 = \frac{\eta\tau^2 + ns_n^2}{\eta + n}$$

Normal vs Normal-gamma (mean and precision)

- Likelihood

$$p(\mathbf{x} \mid \mu, \lambda) \propto \lambda^{\frac{n}{2}} \exp\left[-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Prior

$$p(\mu, \lambda \mid \xi, \kappa, \alpha, \beta) \ \propto \lambda^{\frac{1}{2}} \exp\left[-\frac{\kappa\lambda}{2}(\mu - \xi)^2\right]\lambda^{\alpha-1}\exp\left(-\beta\lambda\right)$$

- Posterior

$$p(\mu, \lambda \mid \mathrm{x}; \xi, \kappa, \alpha, \beta) \propto \lambda^{\frac{1}{2}} \exp\left[-\frac{\tilde{\kappa}\lambda}{2}(\mu - \tilde{\xi})^2\right]\lambda^{\tilde{\alpha}-1}\exp(-\lambda\tilde{\beta})$$

$$\tilde{\xi} = \frac{\kappa\xi + n\overline{x}}{\kappa + n}, \tilde{\kappa} = \kappa + n, \tilde{\alpha} = \alpha + \frac{n}{2}, \tilde{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{1}{2}\frac{n\kappa}{\kappa + n}(\overline{x} - \xi)^2$$

Normal vs normal-inverse-gamma (mean and variance)

- Likelihood

$$p(\mathbf{x} \mid \mu, \sigma^2) \propto (1 / \sigma^2)^{\frac{n}{2}} \exp\left[-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- Prior

$$p(\mu, \lambda \mid \xi, \kappa, \alpha, \beta) \propto (1 / \sigma^2)^{\frac{1}{2}} \exp\left[-\frac{\kappa}{\sigma^2}\frac{(\mu - \xi)^2}{2}\right](1 / \sigma^2)^{\alpha - 1} \exp(-\beta / \sigma^2)$$

- Posterior

$$p(\mu, \sigma^2 \mid \mathrm{x}; \xi, \kappa, \alpha, \beta) \propto (1 / \sigma^2)^{\frac{1}{2}} \exp\left[-\frac{\tilde{\kappa}}{2\sigma^2}(\mu - \tilde{\xi})^2\right](1 / \sigma^2)^{\tilde{\alpha}+1} \exp(-\tilde{\beta} / \sigma^2)$$

$$\tilde{\xi} = \frac{\kappa\xi + n\overline{x}}{\kappa + n}, \tilde{\kappa} = \kappa + n, \tilde{\alpha} = \alpha + \frac{n}{2}, \tilde{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{1}{2}\frac{n\kappa}{\kappa + n}(\overline{x} - \xi)^2$$

Normal vs Scaled-inverse-chi-square (mean and variance)

- Likelihood $$p(\mathbf{x} \mid \mu, \sigma^2) \propto (1/\sigma^2)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$
- Prior $$p(\mu, \sigma^2 \mid \xi, \kappa, \eta, \tau^2) \propto (1/\sigma^2)^{\frac{1}{2}} \exp\left[-\frac{\kappa}{\sigma^2}\frac{(\mu - \xi)^2}{2}\right](1/\sigma^2)^{\eta/2+1} \exp\left[-\frac{\eta\tau^2}{2\sigma^2}\right]$$
- Posterior

$$p(\mu, \sigma^2 \mid \mathbf{x}) \propto (1/\sigma^2)^{\frac{1}{2}} \exp\left[-\frac{\tilde{\kappa}}{\sigma^2}\frac{(\mu - \tilde{\xi})^2}{2}\right](1/\sigma^2)^{\tilde{\alpha}+1} \exp(-\tilde{\eta}\tilde{\tau}^2 / \sigma^2)$$

$$\mu, \sigma^2 \mid \mathbf{x} \sim \text{Normal-Scaled-inverse-chi-square}(\tilde{\xi}, \tilde{\kappa}, \tilde{\eta}, \tilde{\tau}^2)$$

$$\tilde{\xi} = \frac{\kappa\xi + n\overline{x}}{\kappa + n}, \qquad \tilde{\kappa} = \kappa + n,\ \tilde{\eta} = \eta + n,$$

$$\tilde{\eta}\tilde{\tau}^2 = \eta\tau^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{n\kappa}{\kappa + n}(\overline{x} - \xi)^2$$

Mean squared error

The **mean squared error** (MSE) of a point estimator W of a parameter θ is the function of θ defined by

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2.$$

The bias of W is defined by

$$\mathrm{Bias}_\theta W = \mathrm{E}_\theta W - \theta.$$

An estimator is called unbiased if

$$\mathrm{E}_\theta W = \theta$$

for all θ.

Best unbiased estimator

An estimator $W^\star$ is a **best unbiased estimator** (BUE) of $\tau(\theta)$ if

(1) It satisfies $\mathrm{E}_\theta W^\star = \tau(\theta)$ for all θ and,

(2) for any other estimator W with $\mathrm{E}_\theta W = \tau(\theta)$,

$$\mathrm{Var}_\theta W^\star \leq \mathrm{Var}_\theta W$$

for all θ.

$W^\star$ is also called a **uniform minimum variance unbiased estimator** (UMVUE) of $\tau(\theta)$.

If a best unbiased estimator exists, it is unique.

The Cramér-Rao Inequality

Let $X_1,\ldots,X_n$ be a sample with pdf $f(\mathbf{x}\mid\theta)$, and let $W(\mathbf{X}) = W(X_1,\ldots,X_n)$ be any estimator satisfying

$$\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) = \int_{\mathbf{x}\in\mathcal{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x}\mid\theta)]d\mathbf{x}$$

and

$$\mathrm{Var}_\theta W(\mathbf{X}) < \infty$$

Then

$$\mathrm{Var}_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\right)^2\right)}.$$

Information number

Attainment of the Cramér-Rao bound

Let $X_1, \ldots, X_n$ be iid random variables with pdf $f(x \mid \theta)$, where $f(x \mid \theta)$ satisfies the conditions of the Cramer-Rao theorem. Let $L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$ be the likelihood function of θ. If $W(\mathbf{X}) = (\mathrm{X}_1, \ldots X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao lower bound if and only if

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial \theta} \log L(\theta \mid \mathbf{x})$$

for some function $a(\theta)$.

Rao-Balckwell定理

- 设W是 $\tau(\theta)$ 的任一无偏统计量。T是 θ 的一个充分统计量，定义

$$\phi(T) = E(W|T)$$

则

$$E_\theta(\phi(T)) = \tau(\theta) \text{ and } Var\phi(T) \leq Var(W) \quad \forall\theta$$

即 $\phi(T)$ 是 $\tau(\theta)$ 的一致最优无偏估计

最优无偏估计量的唯一性

- 定理：如果W是$\tau(\theta)$的一个最优无偏估计量，则W唯一。
- 证明：令W'是另一个最优无偏估计量，考虑 $W^* = \frac{1}{2}(W + W'), E_\theta(W^*) = \tau(\theta).$

$$\begin{aligned} Var(W^*) &= \frac{1}{4}Var_\theta W + \frac{1}{4}Var_\theta W + \frac{1}{2}Cov_\theta(W, W') \\ &\leq \frac{1}{4}Var_\theta W + \frac{1}{4}Var_\theta W + \frac{1}{2}Var_\theta(W)][Var_\theta W')] \\ &= Var_\theta(W) \end{aligned}$$

最佳无偏估计的判断

- 定理：如果$E_\theta(W)=\tau(\theta)$, W是 $\tau(\theta)$的最佳无偏估计量的充分必要条件是W与0的所有无偏估计量不相关。

- 定理：设T是参数 θ 的完全充分统计量。$\phi(T)$ 是任意一个仅基于T的统计量. 则 $\phi(T)$ 是其期望的唯一最佳无偏统计量。

- 完全性表明：不存在0的非零无偏统计量

Hypothesis Testing

统计学方法及其应用 统计学基础 假设检验

Introduction

A hypothesis is a statement about a population parameter. The two complementary hypotheses in a hypothesis testing problem are called the **null hypothesis** and the **alternative hypothesis**. They are denoted by H_0 and H_1, respectively.

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_0^c$$
$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$
$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$
$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

Rejection Region and Acceptance Region

A **hypothesis testing procedure** or **hypothesis test** is a rule that specifies:

(1) For which smple values the decision is made to accept H_0 as true.

(2) For which smple values the decision is made to reject H_0 and accept H_1 as true.

The subset of the sample space for which H_0 will be rejected is called the **rejection region** (R) or **critical region**.

The complement of the rejection region is called the **acceptance region** $(A = R^c)$.

Test statistic

Certainly, the inference about the parameter θ is drawn by making use of the sample $\mathbf{X} = (X_1, \ldots, X_n)$, particularly, via a function of the sample, a **test statistic** $W = W(X_1, \ldots, X_n)$.

A hypothesis is a statement about the parameter, a subset of the parameter space.

A rejection region is a set of the sample observations, a subset of the sample space.

Neyman-Pearson tests (NPT)

Consider testing the simple hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta = \theta_1,$$

where the joint pdf or pmf of the sample, corresponding to θ_i is $f(\mathbf{x} \mid \theta_i)$ $(i = 0,1)$. A Neyman-Pearson Test is any test with rejection region

$$R = \{\mathbf{x} : f(\mathbf{x} \mid \theta_1) > kf(\mathbf{x} \mid \theta_0)\}$$

and acceptence region

$$R^c = \{\mathbf{x} : f(\mathbf{x} \mid \theta_1) < kf(\mathbf{x} \mid \theta_0)\},$$

where $k \geq 0$.

Likelihood ratio tests (LRT)

The **likelihood ratio test statistic** for a hypothesis testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_0^c$$

is

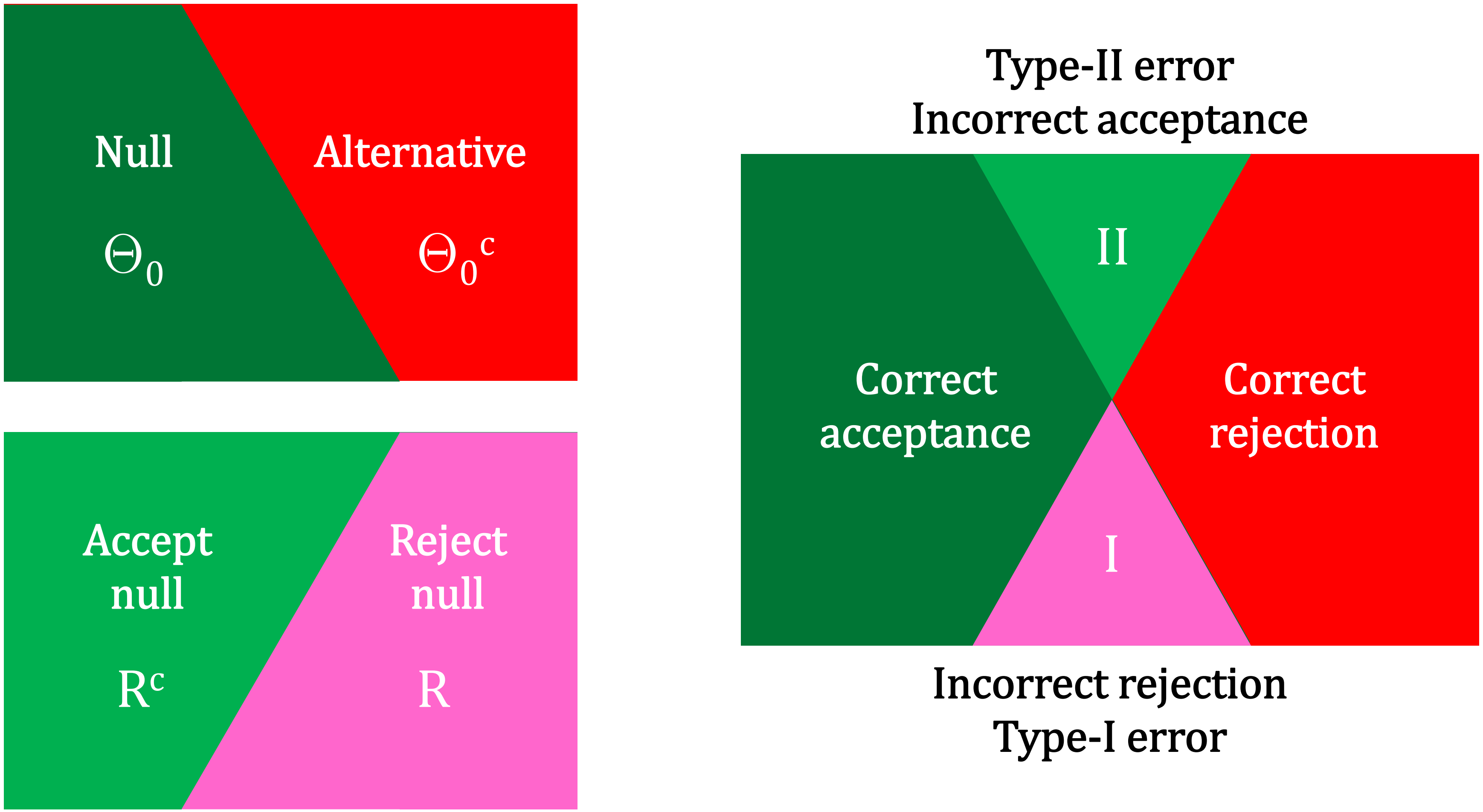
$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup\limits_{\Theta} L(\theta \mid \mathbf{x})}$$

A **likelihood ratio test (LRT)** is any test that has a rejection region of the form

$$R = \{\mathbf{x} : \lambda(\mathbf{x}) < c\},$$

where c is any number satisfying $0 \le c \le 1$.

Two hypotheses, Two actions



Error probabilities

- Confusion matrix

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($\boldsymbol{X} \in R$)	**Correct rejection**	**Type I error**
	Accept H_0 ($\boldsymbol{X} \in R^c$)	**Type II error**	**Correct acceptance**

- Type I error
  – When $\theta \in \Theta_0$ (H_0 is true), P(Type I error) = $P_\theta(\boldsymbol{X} \in R)$
- Type II error
  – When $\theta \in \Theta_0^c$ (H_1 is true), P(Type II error) = $P(\boldsymbol{X} \in R^c) = 1 - P_\theta(\boldsymbol{X} \in R)$

The power function

- Since
$$P_\theta(\mathbf{X} \in R) = \begin{cases} P(\text{Type I error}) & \text{if } \theta \in \Theta_0 \\ 1 - P(\text{Type II error}) & \text{if } \theta \in \Theta_0^c \end{cases},$$
we define

> The **power function** of a hypothesis test with rejection region R is the function of θ defined by
> $$\beta(\theta) = P_\theta(\mathbf{X} \in R).$$

and expect $\beta(\theta)$ to be near 0 for most $\theta \in \Theta_0$ and near 1 for most $\theta \in \Theta_0^c$.

Statistical significance

Size

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a **size α test** if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

Level

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a **level α test** if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

p-value

A $\boldsymbol{p}$ **- value** $p(\mathbf{X})$ is a test statistic that satisfies $0 \le p(\mathbf{x}) \le 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{x})$ give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \le \alpha \le 1$,

$$P_\theta(p(\mathbf{X}) \le \alpha) \le \alpha.$$

p-value

Let $W(\mathbf{X})$ be a test statistic such that **large** values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

Let $W(\mathbf{X})$ be a test statistic such that **small** values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \leq W(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

p-value

Let $W(\mathbf{X})$ be a test statistic such that **large** values of W give evidence that H_1 is true. Let $S(\mathbf{X})$ be a sufficient statistic for the parameter θ **under the null model**. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = P(W(\mathbf{X}) \geq W(\mathbf{x}) \mid S = S(\mathbf{x})).$$

Then, $p(\mathbf{X})$ is a valid p-value.

	H_0	H_1	σ^2 known	σ^2 unknown
One-sample	$\mu = \mu_0$	$\mu \neq \mu_0$	One-sample z test	One-sample t test
		$\mu > \mu_0$		
		$\mu < \mu_0$		
	$\mu \leq \mu_0$	$\mu > \mu_0$		
	$\mu \geq \mu_0$	$\mu < \mu_0$		
Two-sample	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y \neq \delta_0$	Two-sample z test	Two-sample t test
		$\mu_X - \mu_Y > \delta_0$		
		$\mu_X - \mu_Y < \delta_0$		
	$\mu_X - \mu_Y \leq \delta_0$	$\mu_X - \mu_Y > \delta_0$		
	$\mu_X - \mu_Y \geq \delta_0$	$\mu_X - \mu_Y < \delta_0$		
Paired-sample	$\mu_X - \mu_Y = \delta_0$	$\mu_X - \mu_Y \neq \delta_0$	Paired-sample z test	Paired-sample t test
		$\mu_X - \mu_Y > \delta_0$		
		$\mu_X - \mu_Y < \delta_0$		
	$\mu_X - \mu_Y \leq \delta_0$	$\mu_X - \mu_Y > \delta_0$		
	$\mu_X - \mu_Y \geq \delta_0$	$\mu_X - \mu_Y < \delta_0$		

	H_0	H_1	μ known	μ unknown
One-sample		$\sigma^2 \neq \sigma_0^2$	χ^2 test	χ^2 test
	$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$		
		$\sigma^2 < \sigma_0^2$		
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		
Two-sample		$\sigma_X^2 / \sigma_Y^2 \neq \lambda_0$	F test	F test
	$\sigma_X^2 / \sigma_Y^2 = \lambda_0$	$\sigma_X^2 / \sigma_Y^2 > \lambda_0$		
		$\sigma_X^2 / \sigma_Y^2 < \lambda_0$		
	$\sigma_X^2 / \sigma_Y^2 \leq \lambda_0$	$\sigma_X^2 / \sigma_Y^2 > \lambda_0$		
	$\sigma_X^2 / \sigma_Y^2 \geq \lambda_0$	$\sigma_X^2 / \sigma_Y^2 < \lambda_0$		

	H_0	H_1	Median	Symmetry
One-sample	$m = m_0$	$m \neq m_0$	Sign test	Wilcoxon Signed rank test
		$m > m_0$		
		$m < m_0$		
	$m \leq m_0$	$m > m_0$		
	$m \geq m_0$	$m < m_0$		
Two-sample	$m_X = m_Y$	$m_X \neq m_Y$	Sign test	Wilcoxon rank sum test (Mann-Whitney test)
		$m_X > m_Y$		
		$m_X < m_Y$		
	$m_X \leq m_Y$	$m_X > m_Y$		
	$m_X \geq m_Y$	$m_X < m_Y$		
Paired-sample	$m_X = m_Y$	$m_X \neq m_Y$	Sign test	Paired-sample Wilcoxon signed rank test
		$m_X > m_Y$		
		$m_X < m_Y$		
	$m_X \leq m_Y$	$m_X > m_Y$		
	$m_X \geq m_Y$	$m_X < m_Y$		

	H_0	H_1	Exact test	Approximation
One-sample	$p = p_0$	$p \neq p_0$	**Binomial exact test**	**Normal approximation** **χ^2 approximation**
		$p > p_0$		
		$p < p_0$		
	$p \leq p_0$	$p > p_0$		
	$p \geq p_0$	$p < p_0$		
Two-sample	$p_X = p_Y$	$p_X \neq p_Y$	**Fisher exact test**	**Normal approximation** **χ^2 approximation**
		$p_X > p_Y$		
		$p_X < p_Y$		
	$p_X \leq p_Y$	$p_X > p_Y$		
	$p_X \geq p_Y$	$p_X < p_Y$		
Multi-sample	Identical distributions			**χ^2 approximation**
	Independence			

Interval Estimation

Interval estimation

- Our knowledge about the parameter before observing the data

$$\theta \in (-\infty, \infty)$$

- After seeing the data, we made a decision

$$L(\mathbf{x}) \le \theta \le U(\mathbf{x})$$

  Shrank the parameter space from $(-\infty, \infty)$ to an interval

- The moderate precision

$$L(W(\mathbf{x}) \mid \mathbf{x}) \ge L(\theta \mid \mathbf{x}) \text{ for any } \theta \in [L(\mathbf{x}), U(\mathbf{x})]$$

- The gained confidence

$$P(L(\mathbf{X}) \le \theta \le U(\mathbf{X})) = P(L(\mathbf{X}) \ge \theta \text{ and } \theta \le U(\mathbf{X})) = 1 - \alpha$$

Interval estimator

An **interval estimate** of a real-valued parameter θ is any pair of functions, $L(x_1,\dots,x_n)$ and $U(x_1,\dots,x_n)$, of a sample that satisfy $L(\mathbf{x}) \le U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. If $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \le \theta \le U(\mathbf{x})$ is made. The **random interval** $[L(\mathbf{X}), U(\mathbf{X})]$ is called an **interval estimator** for θ.

An interval estimator is typically composed of TWO statistics.
An interval estimate is a pair of real numbers.
An interval estimator is a pair of statistics.

Compare them with point estimate and point estimator

Various forms of interval estimator

Two-sided:

$$[L(\mathbf{X}), U(\mathbf{X})]$$

One-sided:

$$[L(\mathbf{X}), \infty),\ (-\infty, U(\mathbf{X})]$$

Closed interval:

$$[L(\mathbf{X}), U(\mathbf{X})]$$

Open interval:

$$(L(\mathbf{X}), U(\mathbf{X})),\ (L(\mathbf{X}), \infty),\ (-\infty, U(\mathbf{X}))$$

Half open interval:

$$[L(\mathbf{X}), U(\mathbf{X})),\ (L(\mathbf{X}), U(\mathbf{X})],\ [L(\mathbf{X}), \infty),\ (-\infty, U(\mathbf{X})]$$

Coverage probability

For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ, the **coverage probability** of $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter.
The coverage probability is denoted by $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.
The **confidence coefficient** of $[L(\mathbf{X}), U(\mathbf{X})]$ is the infinum of the coverage probability, say,

$$\inf_{\theta \in \Theta} P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

An interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$, together with its confidence coefficient, is called a **confidence interval**. A confidence interval with the confidence coefficient $1-\alpha$ is called a $1-\alpha$ confidence interval.

Inverting a test statistic

For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\mathbf{x} \in \mathcal{X}$, define a set $C(\mathbf{x})$ in the parameter space by

$$C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}$$

Then the random set $C(\mathbf{X})$ is a $1-\alpha$ confidence set.

Conversely, let $C(\mathbf{x})$ be a $1-\alpha$ confidence set. For each $\theta_0 \in \Theta$, define a set in the sample space by

$$A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$$

Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Parameter free distributions

We have seen that for a random sample $X_1,\ldots,X_n$ from a normal population $N(\mu,\sigma^2)$.

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),\ \ \sigma^2\ \text{known}.$$

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim T_{n-1}.$$

$$K=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2,\ \ \mu\ \text{known}.$$

$$K=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2\sim\chi_{n-1}^2.$$

What's the common characteristics of these random variables?

Parameter free!

Pivotal quantities

A random variable $Q(\mathbf{X}, \theta) = Q(X_1, \dots, X_n, \theta)$ is a **pivotal quantity** (**pivot**) if the distribution of $Q(\mathbf{X}, \theta)$ is independent of all parameters. That is, if $\mathbf{X} \sim F(\mathbf{x} \mid \theta)$, then $Q(\mathbf{X}, \theta)$ has the same distribution for all values of θ.

One-way Analysis of Variance

ANOVA data

	Treatment					
Index	**1**	**2**	**3**	**…**	$\boldsymbol{k-1}$	$\boldsymbol{k}$
1	Y_{11}	Y_{21}	Y_{31}		$Y_{(k-1)1}$	Y_{k1}
2	Y_{12}	Y_{22}	Y_{32}		$Y_{(k-1)2}$	Y_{k2}
$\ldots$	Y_{13}	Y_{23}	Y_{33}		$Y_{(k-1)3}$	Y_{k3}
	$\ldots$	$\ldots$	$\ldots$		$\ldots$	$\ldots$
	Y_{1n_1}				$Y_{(k-1)n_{(k-1)}}$	
			Y_{3n_3}			
		Y_{2n_2}				
						Y_{kn_k}
$\boldsymbol{\theta}$	θ_1	θ_2	θ_3	$\ldots$	θ_{k-1}	θ_k
N	n_1	n_2	n_3	$\ldots$	n_{k-1}	n_k
$\overline{\overline{Y}}$	$\overline{Y}_{1\bullet}$	$\overline{Y}_{2\bullet}$	$\overline{Y}_{3\bullet}$		$\overline{Y}_{(k-1)\bullet}$	$\overline{Y}_{k\bullet}$

$$N = \sum_{i=1}^{k} n_i, Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}, \overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \overline{\overline{Y}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

ANOVA model

Random variables Y_{ij} are observed according to the model

$$Y_{ij} = \theta_i + \varepsilon_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i,$$

where

(i) $\mathrm{E}\varepsilon_{ij} = 0,\ \mathrm{Var}\varepsilon_{ij} = \sigma_i^2 < \infty,$ for all i and j.

$\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{st}) = 0$ for all $i, j, s,$ and t unless $i = s$ and $j = t$.

(ii) The ε_{ij} are independent and normally distributed (normal errors).

(iii) $\sigma_i^2 = \sigma^2$ for all i (equality of variance, homoscedasticity).

ANOVA normal families

(i) and (ii) and (iii) $\Rightarrow$

$$\varepsilon_{ij} \sim N(0, \sigma^2), \qquad \text{iid}, \; i = 1, \ldots, k, j = 1, \ldots, n_i.$$

$$Y_{ij} = \theta_i + \varepsilon_{ij} \text{ and } \varepsilon_{ij} \sim N(0, \sigma^2), \text{ iid} \Rightarrow$$

$$Y_{ij} \sim N(\theta_i, \sigma^2), \qquad i = 1, \ldots, k, j = 1, \ldots, n_i.$$

$$\bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \text{ and } Y_{ij} \sim N(\theta_i, \sigma^2) \;\Rightarrow$$

$$\bar{Y}_{i\bullet} \sim N(\theta_i, \sigma^2 / n_i), \qquad i = 1, \ldots, k$$

ANOVA hypothesis

Pair-wise two-sample t test over all possible combinations of:

$$H_0\colon \theta_i = \theta_j \qquad \text{versus} \qquad H_1\colon \theta_i \neq \theta_j$$

Is equal to

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{versus} \qquad H_1 : \theta_i \neq \theta_j \text{ for some } i, j$$

This is the classical ANOVA hypothesis

ANOVA t test

Two sample case.

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

The test is to reject H_0 if

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\sigma^2(1/m + 1/n)}} > t_{m+n-2,\alpha/2}.$$

ANOVA (k-sample) case.

$$H_0 : \sum_{i=1}^{k} a_i \theta_i = 0 \quad \text{versus} \quad H_1 : \sum_{i=1}^{k} a_i \theta_i \neq 0$$

The test is to reject H_0 if

$$\frac{\left|\sum_{i=1}^{k} a_i \bar{Y}_{i\bullet}\right|}{\sqrt{S_p^2 \sum_{i=1}^{k} (a_i^2 / n_i)}} > t_{N-k,\alpha/2}.$$

ANOVA *F* test

For the ANOVA hypothesis testing problem

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \quad \text{versus} \quad H_1 : \theta_i \neq \theta_j \text{ for some } i, j$$

We **reject H_0 if**

$$F = \frac{\sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \Big/ (k-1)}{S_p^2} > F_{k-1, N-k, \alpha}$$

The p - value is

$$p = P\left(F_{k-1, N-k} \geq \frac{\sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{\bar{y}})^2 \Big/ (k-1)}{s_p^2} \right)$$

Partitioning sums of squares

$$\mathbf{SST = SSB + SSW}$$

$$\sum_{j=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{\overline{Y}})^2 = \sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet}-\overline{\overline{Y}})^2 + \sum_{j=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet})^2$$

Dividing them by σ^2

$$\chi^2_{N-1} = \chi^2_{k-1} + \chi^2_{N-k}$$

$$\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{i=1}^{k} Y_{ij}, \overline{\overline{Y}} = \frac{1}{N}\sum_{i=1}^{k} n_i \overline{Y}_{i\bullet} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic	p value
Between treatment groups	$k-1$	SSB $\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2$	MSB $\frac{\text{SSB}}{k-1}$	$F = \frac{\text{MSB}}{\text{MSW}}$	$1 - F_{k-1,N-k}(F)$
Within treatment groups	$N-k$	SSW $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	MSW $\frac{\text{SSW}}{N-k}$		
Total	$N-1$	SST $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2$			

Linear Regression

统计学方法及其应用 统计学基础 回归分析

The data

	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{2}$	$\boldsymbol{\ldots}$	$\boldsymbol{k-1}$	$\boldsymbol{k}$
Y_1						
Y_2						
Y_3						
$\ldots$	$\mathbf{1}$		**Real numbers**			
Y_n						
$\boldsymbol{\beta}$	β_0	β_2	β_3	$\ldots$	β_{k-1}	β_k

Conditional normal model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, \ldots, n$$

Assume that

(i) $\mathrm{E}\varepsilon_i = 0,\ \mathrm{Var}\varepsilon_i = \sigma_i^2 < \infty,$ for all.

$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all i and j unless $i = j$.

(ii) The ε_i are independent and normally distributed (normal errors).

(iii) $\sigma_i^2 = \sigma^2$ for all i (equality of variance, homoscedasticity).

(iv) Y_i independent (but not identically distributed), $i = 1, \ldots, n$.

(iv) x_i known and fixed (not random variables), $i = 1, \ldots, n$.

It then follows that

$$\begin{aligned} Y_i \mid x_i &\sim N(\alpha + \beta x_i, \sigma^2) \\ \mathrm{E}\,Y_i &= \alpha + \beta x_i \\ \mathrm{Var}\,Y_i &= \sigma^2 \end{aligned}$$

Least squares estimates (LSE)

$$\text{RSS} = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

Solve $\displaystyle \min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$

$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \sum_{i=1}^{n} [(y_i - bx_i) - a]^2 \Rightarrow a = \frac{1}{n} \sum_{i=1}^{n} (y_i - bx_i) = \overline{y} - b\overline{x}$$

$$\sum_{i=1}^{n} [(y_i - bx_i) - (\overline{y} - b\overline{x})]^2 = \sum_{i=1}^{n} [(y_i - \overline{y}) - b(x_i - \overline{x})]^2 = b^2 S_{xx} - 2bS_{xy} + S_{yy}$$

$$= S_{xx} \left(b - \frac{S_{xy}}{S_{xx}} \right)^2 + \frac{S_{xx} S_{yy} - S_{xy} S_{xy}}{S_{xx}}$$

$$\Rightarrow b = \frac{S_{xy}}{S_{xx}}, a = \overline{y} - b\overline{x}$$

RSS: Residual Sum of Squares

Best linear unbiased estimators (BLUE)

$$
\begin{aligned}
Y_i &= \alpha + \beta x_i + \varepsilon_i, i = 1, \ldots, n \\
\mathrm{E}\, Y_i &= \alpha + \beta x_i \qquad (\mathrm{E}\varepsilon_i = 0) \\
\mathrm{Var}\, Y_i &= \sigma^2 \qquad\qquad (\mathrm{Var}\varepsilon_i = \sigma^2)
\end{aligned}
$$

We attempt to find estimators (functions of $\mathbf{Y}$) of α and β

Now, restrict our attention to the class of **linear estimators**, say,

$$\hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i \quad \text{and} \quad \hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i,$$

where ξ_i and δ_i are known, fixed constants.

We are interested in unbiased and minumum variance estimators.

Unbiased estimator of the slope

$$\hat{\boldsymbol{\beta}} = \frac{\boldsymbol{S}_{\boldsymbol{xY}}}{\boldsymbol{S}_{\boldsymbol{xx}}} \qquad \hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i$$

$$\mathrm{E}\hat{\beta} = \mathrm{E}\left[\sum_{i=1}^{n} \delta_i Y_i\right] = \sum_{i=1}^{n} \delta_i \mathrm{E}\, Y_i = \sum_{i=1}^{n} \delta_i (\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} \delta_i + \beta \sum_{i=1}^{n} \delta_i x_i$$

$\hat{\beta}$ is unbiased if and only if $\sum_{i=1}^{n} \delta_i = 0$ and $\sum_{i=1}^{n} \delta_i x_i = 1$.

Now, consider $\delta_i = \dfrac{x_i - \bar{x}}{S_{xx}}$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} \delta_i x_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} x_i = \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum_{i=1}^{n} (x_i - \bar{x}) \bar{x}\right] = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1$$

Therefore, $\hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i = \sum_{i=1}^{n} \dfrac{x_i - \bar{x}}{S_{xx}} Y_i = \dfrac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \dfrac{S_{xY}}{S_{xx}}$ is unbiased.

Unbiased estimator of the intercept

$$\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{Y}} - \hat{\boldsymbol{\beta}}\bar{\boldsymbol{x}} \qquad\qquad \hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i$$

$$\mathrm{E}\hat{\alpha} = \mathrm{E}\left(\sum_{i=1}^{n} \xi_i Y_i\right) = \sum_{i=1}^{n} \xi_i \mathrm{E}\, Y_i = \sum_{i=1}^{n} \xi_i (\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} \xi_i + \beta \sum_{i=1}^{n} \xi_i x_i$$

$\hat{\alpha}$ is unbiased if and only if $\sum_{i=1}^{n} \xi_i = 1$ and $\sum_{i=1}^{n} \xi_i x_i = 0$

Now, consider $\xi_i = \dfrac{1}{n} - \dfrac{(x_i - \bar{x})\bar{x}}{S_{xx}}$

$$\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right] = 1 - \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\bar{x} = 1$$

$$\sum_{i=1}^{n} \xi_i x_i = \sum_{i=1}^{n} \left[\frac{1}{n} x_i - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} x_i\right] = \bar{x} - \frac{\bar{x}}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) x_i = 0$$

Therefore, $\hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i = \bar{Y} - \hat{\beta}\bar{x}$ is unbiased.

Best linear unbiased estimator (BLUE)

$$\hat{\boldsymbol{\beta}} = \frac{\boldsymbol{S}_{xY}}{\boldsymbol{S}_{xx}}$$

$$\hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i, \ \ \delta_i = \frac{x_i - \overline{x}}{S_{xx}}$$

$$\mathrm{Var}\,\hat{\beta} = \mathrm{Var}\left(\sum_{i=1}^{n} \delta_i Y_i\right) = \sum_{i=1}^{n} \delta_i^2 \mathrm{Var}\, Y_i = \sum_{i=1}^{n} \delta_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} \delta_i^2 = \frac{\sigma^2}{S_{xx}}$$

$$\text{because } \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 = \frac{S_{xx}}{S_{xx} S_{xx}} = \frac{1}{S_{xx}}$$

It can be proved that $\mathrm{Var}\,\hat{\beta}$ is the minimum.

Therefore, $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β.

Best linear unbiased estimator (BLUE)

$$\boldsymbol{\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}}$$

$$\hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i,\ \ \xi_i = \frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}}$$

$$\mathrm{Var}\hat{\alpha} = \mathrm{Var}\left(\sum_{i=1}^{n} \xi_i Y_i\right) = \sum_{i=1}^{n} \xi_i^2 \mathrm{Var}\, Y_i = \sum_{i=1}^{n} \xi_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} \xi_i^2 = \sigma^2 \left(\frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right)$$

$$\sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}}\right)^2 = \sum_{i=1}^{n} \frac{1}{n^2} + \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})\overline{x}}{S_{xx}}\right)^2 = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} = \frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2$$

It can be proved that $\mathrm{Var}\hat{\alpha}$ is the minimum.

Therefore, $\hat{\alpha}$ is the best linear unbiased estimator (BLUE) of α.

Unbiased estimator of the variance

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (\alpha + \hat{\beta}x_i)]^2 = \frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2$$

Because

$$\mathrm{E}[\hat{\sigma}^2] = \frac{n}{n-2}\sigma^2$$

We have

$$\mathrm{E}[S^2] = \sigma^2$$

Recall biased and unbiased estimators for the normal variance

Sampling distribution of the slope

$$\hat{\boldsymbol{\beta}} = \frac{\boldsymbol{S}_{xY}}{\boldsymbol{S}_{xx}} \sim \mathbf{N}\left(\boldsymbol{\beta}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{S}_{xx}}\right)$$

$$\hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i, \delta_i = \frac{x_i - \overline{x}}{S_{xx}}, E(\hat{\beta}) = \beta, Var(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$$

Y_i is normally distributed, therefore the linear combination $\hat{\beta} = \sum_{i=1}^{n} \delta_i Y_i$ is also normally distributed. In other words, $\hat{\beta}$ has a normal distribution

$$\Rightarrow \qquad \hat{\beta} \sim \mathrm{N}(\beta, \sigma^2 / S_{xx}) \text{ or } \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}} \sim \mathrm{N}(0,1)$$

Sampling distribution of the intercept

$$\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{Y}} - \hat{\boldsymbol{\beta}}\overline{\boldsymbol{x}} \sim \mathbf{N}\left(\boldsymbol{\alpha}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}\boldsymbol{S}_{xx}} \sum_{i=1}^{n} \boldsymbol{x}_i^2\right)$$

$$\hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i, \xi_i = \frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}}, E(\hat{\alpha}) = \alpha, Var(\hat{\alpha}) = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2$$

Y_i is normally distributed, therefore the linear combination $\hat{\alpha} = \sum_{i=1}^{n} \xi_i Y_i$ is also normally distributed. In other words, $\hat{\alpha}$ has a normal distribution

$$\Rightarrow \qquad \hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right) \text{ or } \frac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2 (\sum_{i=1}^{n} x_i^2) / (nS_{xx})}} \sim \mathrm{N}(0,1)$$

Covariance of the intercept and slope

$$\hat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x}, \quad \widehat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = \operatorname{cov}\left(\sum_{i=1}^{n} \xi_i Y_i, \sum_{i=1}^{n} \delta_i Y_i\right) = \sum_{i=1}^{n} \xi_i \delta_i \operatorname{Var} Y_i = \sigma^2 \sum_{i=1}^{n} \xi_i \delta_i$$

$$\begin{aligned}
\sum_{i=1}^{n} \xi_i \delta_i &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}}\right)\left(\frac{x_i - \overline{x}}{S_{xx}}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) - \frac{\overline{x}}{S_{xx}} \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})(x_i - \overline{x})}{S_{xx}}\right) \\
&= -\frac{\overline{x}}{S_{xx}}
\end{aligned}$$

$$\Rightarrow \qquad \operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \overline{x}}{S_{xx}}$$

Sampling distribution of the variance

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - (\hat{\alpha} + \hat{\beta}x_i)\right]^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2$$

(1) $(\hat{\alpha}, \hat{\beta})$ **and** S^2 **are independent**

(2) $$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

$(n-2)S^2 = \sum_{i=1}^{n}\varepsilon_i^2$, **Residual sum of squares**

Hypothesis testing of the intercept

$$H_0 : \alpha = \alpha_0 \quad \text{versus} \quad H_1 : \alpha \neq \alpha_0$$

Since

$$\frac{\hat{\alpha} - \alpha}{\sqrt{S^2 (\sum_{i=1}^{n} x_i^2) / (nS_{xx})}} \sim T_{n-2}$$

We could reject H_0 at level ρ if and only if

$$\frac{|\hat{\alpha} - \alpha_0|}{\sqrt{S^2 (\sum_{i=1}^{n} x_i^2) / (nS_{xx})}} > t_{n-2,\rho/2}$$

p-value is

$$p = 2P\left(T_{n-2} \geq \frac{|\hat{\alpha} - \alpha_0|}{\sqrt{S^2 (\sum_{i=1}^{n} x_i^2) / (nS_{xx})}} \right)$$

Hypothesis testing of the intercept

$$H_0 : \alpha = \alpha_0 \quad \text{versus} \quad H_1 : \alpha \neq \alpha_0$$

Since

$$\frac{(\hat{\alpha} - \alpha)^2}{S^2 \left(\sum_{i=1}^{n} x_i^2\right) / (nS_{xx})} \sim F_{1,n-2}$$

We could reject H_0 at level ρ if and only if

$$\frac{(\hat{\alpha} - \alpha_0)^2}{S^2 \left(\sum_{i=1}^{n} x_i^2\right) / (nS_{xx})} > F_{1,n-2,\rho}$$

p-value is

$$p = P\left(F_{1,n-2} \geq \frac{(\hat{\alpha} - \alpha_0)^2}{S^2 \left(\sum_{i=1}^{n} x_i^2\right) / (nS_{xx})}\right)$$

Hypothesis testing of the slope

$$H_0 : \beta = \beta_0 \quad \text{versus} \quad H_1 : \beta \neq \beta_0$$

Since

$$\frac{\hat{\beta} - \beta}{\sqrt{S^2 / S_{xx}}} \sim T_{n-2}$$

We could reject H_0 at level ρ if and only if

$$\frac{|\hat{\beta} - \beta_0|}{\sqrt{S^2 / S_{xx}}} > t_{n-2,\rho/2}$$

p-value is

$$p = 2P\left(T_{n-2} \geq \frac{|\hat{\beta} - \beta_0|}{\sqrt{S^2 / S_{xx}}} \right)$$

Hypothesis testing of the slope

$$H_0 : \beta = \beta_0 \quad \text{versus} \quad H_1 : \beta \neq \beta_0$$

Since

$$\frac{(\hat{\beta} - \beta)^2}{S^2 / S_{xx}} \sim F_{1,n-2}$$

We could reject H_0 at level ρ if and only if

$$\frac{(\hat{\beta} - \beta_0)^2}{S^2 / S_{xx}} > F_{1,n-2,\rho}$$

p-value is

$$p = P\left(F_{1,n-2} \geq \frac{(\hat{\beta} - \beta_0)^2}{S^2 / S_{xx}}\right)$$

Estimation at a single point

Obviously, $\hat{\mu}_{Y_0} = \hat{\alpha} + \hat{\beta} x_0$ has a normal distribution

$$E\hat{\mu}_{Y_0} = E(\hat{\alpha} + \hat{\beta} x_0) = E(\hat{\alpha}) + x_0 E(\hat{\beta}) = a + \beta x_0$$

$$Var\hat{\mu}_{Y_0} = Var(\hat{\alpha} + \hat{\beta} x_0) = Var(\hat{\alpha}) + x_0^2 Var(\hat{\beta}) + 2x_0 Cov(\hat{\alpha}, \hat{\beta})$$

$$= \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2 + \frac{\sigma^2 x_0^2}{S_{xx}} - \frac{2\sigma^2 x_0 \overline{x}}{S_{xx}} = \frac{\sigma^2}{S_{xx}} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 + x_0^2 - 2x_0 \overline{x} \right)$$

$$= \frac{\sigma^2}{S_{xx}} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 + x_0^2 - 2x_0 \overline{x} + \overline{x}^2 \right)$$

$$= \frac{\sigma^2}{S_{xx}} \left\{ \frac{1}{n} \left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right] + (x_0 - \overline{x})^2 \right\}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]$$

Sampling distribution

$$\hat{\boldsymbol{\mu}}_{Y_0} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}\boldsymbol{x}_0 \sim \mathbf{N}\left(\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{x}_0, \boldsymbol{\sigma}^2\left[\frac{1}{\boldsymbol{n}} + \frac{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}})^2}{\boldsymbol{S}_{xx}}\right]\right)$$

$$\frac{(\widehat{\alpha} + \widehat{\beta}x_0) - (\alpha + \beta x_0)}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} \sim \mathrm{N}(0,1)$$

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2} \Rightarrow$$

$$\frac{(\widehat{\alpha} + \widehat{\beta}x_0) - (\alpha + \beta x_0)}{\sqrt{S^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} = \frac{(\widehat{\alpha} + \widehat{\beta}x_0) - (\alpha + \beta x_0)}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} \frac{1}{\sqrt{\frac{(n-2)S^2}{(n-2)\sigma^2}}} \sim T_{n-2}$$

Prediction at a single point

Assume that

$$Y_0 = \alpha + \beta x_0 + \varepsilon_0 \sim \mathrm{N}(\alpha + \beta x_0, \sigma^2)$$

Then,

$$Y_0 - \hat{\mu}_{Y_0} \text{ has a normal distribution}$$

$$E(Y_0 - \hat{\mu}_{Y_0}) = EY_0 - E\hat{\mu}_{Y_0} = (\alpha + \beta x_0) - (\alpha + \beta x_0) = 0$$

$$Var(Y_0 - \hat{\mu}_{Y_0}) = VarY_0 + Var\hat{\mu}_{Y_0} + 2Cov(Y_0, \hat{\mu}_{Y_0})$$

$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

Sampling distribution

$$\boldsymbol{Y}_0 - (\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}\boldsymbol{x}_0) \sim \mathbf{N}\left(0, \boldsymbol{\sigma}^2\left[1 + \frac{1}{\boldsymbol{n}} + \frac{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}})^2}{\boldsymbol{S}_{\boldsymbol{xx}}}\right]\right)$$

$$\frac{Y_0 - (\widehat{\alpha} + \widehat{\beta}x_0)}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} \sim \mathrm{N}(0,1)$$

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2} \Rightarrow$$

$$\frac{Y_0 - (\widehat{\alpha} + \widehat{\beta}x_0)}{\sqrt{S^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} = \frac{Y_0 - (\widehat{\alpha} + \widehat{\beta}x_0)}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} \frac{1}{\sqrt{\frac{(n-2)S^2}{(n-2)\sigma^2}}} \sim T_{n-2}$$

Prediction interval estimation

$$\frac{Y_0 - (\widehat{\alpha} + \widehat{\beta} x_0)}{\sqrt{S^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}} \sim T_{n-2}$$

It is a pivotal quantity.

Therefore, a $1-\rho$ prediction interval for Y_0 is

$$\widehat{\alpha} + \widehat{\beta} x_0 - t_{n-2,\rho/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \le Y_0 \le \widehat{\alpha} + \widehat{\beta} x_0 + t_{n-2,\rho/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

$$\hat{\alpha} + \hat{\beta} x_0 - t_{n-2,\rho/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \le \alpha + \beta x_0 \le \hat{\alpha} + \hat{\beta} x_0 + t_{n-2,\rho/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Illustration of confidence bands