

STATISTICAL INFERENCE

POPULATION AND SAMPLE

Population = all elements of interest
Characterized by a distribution F
with *some* parameter θ



Sample = the data X_1, \dots, X_n ,
selected subset of the population

n = sample size

Examples of F : Bernoulli(p), Normal(μ, σ),
Gamma(n, λ), Poisson(λ), etc.

Statistical Inference

= inference about a population based on a sample

- Parameter estimation
- Confidence intervals
- Hypothesis testing
- Model fitting

Parameter Estimation

Statistic = any function of data $W(X_1, \dots, X_n)$

Estimator of θ = any statistic used to estimate θ

Estimator $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta$

Example 1. Estimation of a mean:

Estimate *the population mean* $\theta = \mu = E(X_i)$ by a *sample mean*

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

Unbiased: $E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E X_i = \frac{1}{n} \sum_{i=1}^n \theta = \theta.$

Example 2. Estimation of a variance:

Estimate *the population variance*

$$\theta = \sigma^2 = \text{Var}(X_i)$$

by *a sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Also unbiased: $E(S^2) = \sigma^2$.

Methods of Estimation

1. Method of Moments

$$\begin{array}{ll} k^{th} \text{ population moment} & \mu_k = EX^k \\ k^{th} \text{ sample moment} & M_k = \frac{1}{n} \sum_{i=1}^n X_i^k \end{array}$$

To estimate d parameters, solve the system

$$\begin{cases} M_1 = \mu_1 \\ \dots \\ M_d = \mu_d \end{cases}$$

2. Method of Maximum Likelihood

Maximize the probability (pmf, pdf) of seeing the really observed data

Implementation

Observe X_1, \dots, X_n from pdf or pmf $f(x | \theta)$.

Maximize

$$f(X_1, \dots, X_n | \theta) = \prod_{i=1}^n f(X_i | \theta)$$

in θ .

Simplification: maximize

$$\ln f(X_1, \dots, X_n | \theta) = \sum_{i=1}^n \ln f(X_i | \theta)$$

Typically, compute

$$\frac{\partial}{\partial \theta} \ln f(X_1, \dots, X_n | \theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i | \theta)$$

equate to 0 and solve in θ .

Confidence Intervals

$100(1 - \alpha)$ %-confidence interval is an interval that contains parameter θ with probability γ .

That is,

$$P\{a \leq \theta \leq b\} = 1 - \alpha$$

where

$$a = a(X_1, \dots, X_n) \text{ and } b = b(X_1, \dots, X_n)$$

are statistics. So, a and b are random, θ is not.

Example: X_1, \dots, X_n from $\text{Normal}(\mu, \sigma)$ with unknown μ , known σ

1. Estimate $\theta = \mu$ by its estimator $\bar{X} = \frac{1}{n} \sum X_i$.
2. Find its distribution: *Normal* with

$$E(\bar{X}) = \mu$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_1^n \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Therefore,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is Normal}(0,1)$$

3. Find *critical values* $\pm z_{\alpha/2}$ such that

$$P \left\{ -z_{\alpha/2} < Z < z_{\alpha/2} \right\}$$

for $Z \sim \text{Normal}(0,1)$.

4. Then we have

$$P \left\{ -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right\} = 1 - \alpha$$

Solve for μ :

$$P \left\{ \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

5. Hence,

$$\bar{X} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \left[\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right]$$

is a $(1 - \alpha)100\%$ confidence interval for μ .

When σ is unknown

Data X_1, \dots, X_n from Normal(μ, σ) with unknown μ , **unknown** σ

1. Estimate σ by $S = \sqrt{\frac{1}{n-1} \sum_1^n (X_i - \bar{X})^2}$
2. Use t -distribution with $(n-1)$ degrees of freedom instead of Normal.

For large n , use Normal approximation.

Result:

$$\bar{X} \pm \frac{t_{\alpha/2, n-1} S}{\sqrt{n}}$$

TESTING HYPOTHESES

Hypothesis H_0 and alternative $H_A =$ mutually exclusive statements about the unknown parameter θ .

Collect data



Conduct a test



State if there is sufficient evidence to reject H_0 in favour of H_A .

Conclusion	Reject H_0	Accept H_0
H_0 is true	Type I error	correct
H_0 is false	correct	Type II error

Control the **significance level**

$$\alpha = P \{ \text{Type I error} \}$$

Data: X_1, \dots, X_n from Normal(μ, σ) with unknown μ , known σ

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.

1. Find $\pm z_{\alpha/2}$. Acceptance region: $[-z_{\alpha/2}, z_{\alpha/2}]$.
2. Compute the *test statistic*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

3. If Z belongs to the acceptance region, do not reject H_0 .
Otherwise, reject H_0 .
-

If H_0 is true, Z has Normal(0,1) distribution, and

$$\mathbf{P} \{ \text{Type I error} \} = \mathbf{P} \{ |Z| > z_{\alpha/2} \} = \alpha$$

One-sided, right-tail tests

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$.

1. Find z_α . The *acceptance region* is $(-\infty, z_\alpha]$.
2. Compute the *test statistic*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

3. If Z belongs to the acceptance region, do not reject H_0 .
Otherwise, reject H_0 .

One-sided, left-tail tests

Test $H_0 : \mu = \mu_0$ vs $H_A : \mu < \mu_0$.

1. Find z_α . The *acceptance region* is $[-z_\alpha, +\infty)$.
2. Compute the *test statistic*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

3. If Z belongs to the acceptance region, do not reject H_0 .
Otherwise, reject H_0 .

Case of unknown variance

1. Estimate σ by $S = \sqrt{\frac{1}{n-1} \sum_1^n (X_i - \bar{X})^2}$
2. Use t -distribution with $(n - 1)$ degrees of freedom.

For large n , use Normal approximation.

LINEAR REGRESSION

Observe pairs (X_i, Y_i) for $i = 1, \dots, n$.

X_i = predictor

Y_i = response

Fitting a *linear model*

$$Y_i = a + X_i b + \varepsilon_i$$

ε_i = error term

Goals:

- Estimate parameters a and b
- Predict Y for a new value of X

Method of Least Squares

For each X_i , the model predicts

$$\hat{Y}_i = a + X_i b$$

Minimize $G(a, b) = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$:

$$\begin{cases} \frac{\partial G}{\partial a} = 2 \sum_{i=1}^n (a + X_i b - Y_i) = 0 \\ \frac{\partial G}{\partial b} = 2 \sum_{i=1}^n (a + X_i b - Y_i) X_i = 0 \end{cases}$$

$$\begin{cases} a n + b \sum X_i = \sum Y_i \\ a \sum X_i + b \sum X_i^2 = \sum X_i Y_i \end{cases}$$

Least squares estimates:

$$\begin{cases} a = \bar{Y} - \bar{X}b \\ b = \frac{S_{XY}}{S_{XX}} = r \frac{S_Y}{S_X} \end{cases}$$

where

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = (n - 1)Cov(X, Y),$$

$$S_{XX} = \sum (X_i - \bar{X})^2 = (n - 1)S_X^2,$$

$S_X^2, S_Y^2 =$ sample variances,

$r = \frac{Cov(X, Y)}{S_X S_Y} =$ correlation
coefficient

Sample regression line:

$$y = a + xb$$

Prediction for new X_0 :

$$\hat{Y}_0 = a + X_0b$$