

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 **the 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 parameter θ

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

Standard error of an estimator is its standard deviation $\text{Std}(\hat{\theta})$

It is estimated by $\widehat{\text{Std}}(\hat{\theta})$. It shows the accuracy, reliability of estimator $\hat{\theta}$.

Estimation of a mean

Sample (X_1, \dots, X_n) is collected from a population with $\mathbf{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

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

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

Properties:

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

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

So, \bar{X} is **unbiased**, and its **standard error** is

$$SE(\bar{X}) = \text{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Estimation of a variance

Estimate *the population variance*

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

by *a sample variance*

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

It is also unbiased: $\mathbf{E}(S^2) = \sigma^2$.

Then, the standard error of \bar{X} is estimated by

$$\widehat{\text{Std}}(\bar{X}) = \frac{S}{\sqrt{n}} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n(n-1)}}$$

Estimation of a proportion

Sample (X_1, \dots, X_n) is collected from **Bernoulli** population with parameter p .

Estimate *the population proportion* $p = \mathbf{E}(X_i)$ by a *sample proportion*

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{\text{number of } X_i = 1}{n}$$

Special case of a sample mean \bar{X}

$$\mathbf{E}(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$

So, \hat{p} is **unbiased**;

its **standard error** is $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

typically estimated by $\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

General Methods of Estimation

1. Method of Moments

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

To estimate d parameters, solve the system of d equations

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

M_1, \dots, M_d are known from the sample; μ_1, \dots, μ_d are functions of unknown parameters

Example: X_1, \dots, X_n are Exponential(λ)

Estimate λ .

The number of parameters is $d = 1$. So, we need 1 equation.

$$M_1 = \bar{X}; \mu_1 = \frac{1}{\lambda}$$

Solve

$$M_1 = \mu_1 \Rightarrow \bar{X} = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{mom} = \frac{1}{\bar{X}}$$

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 .

Example: X_1, \dots, X_n are Exponential(λ)

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

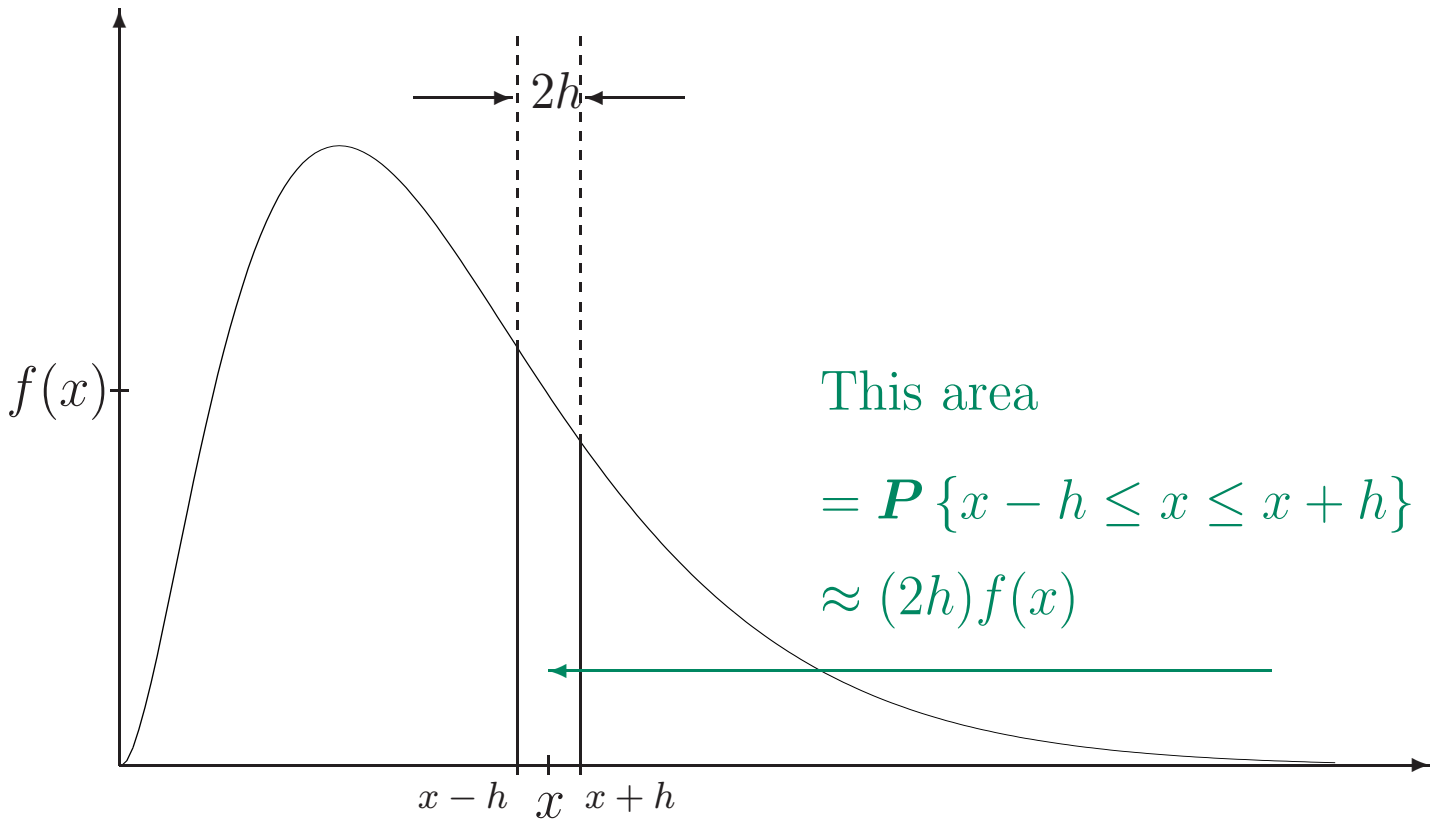
$$\begin{aligned} \ln f(X_1, \dots, X_n | \theta) &= \sum_{i=1}^n \ln(\lambda e^{-\lambda X_i}) \\ &= n \ln \lambda - \lambda \sum_{i=1}^n X_i \end{aligned}$$

$$\frac{\partial}{\partial \lambda} \ln f(X_1, \dots, X_n | \theta) = \frac{n}{\lambda} - \sum_{i=1}^n X_i =: 0$$

Solve for λ ,

$$\hat{\lambda}_{mle} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}$$

Maximum likelihood



This area
 $= P \{x - h \leq x \leq x + h\}$
 $\approx (2h)f(x)$

Probability of observing “almost” $X = x$

Confidence Intervals

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

That is,

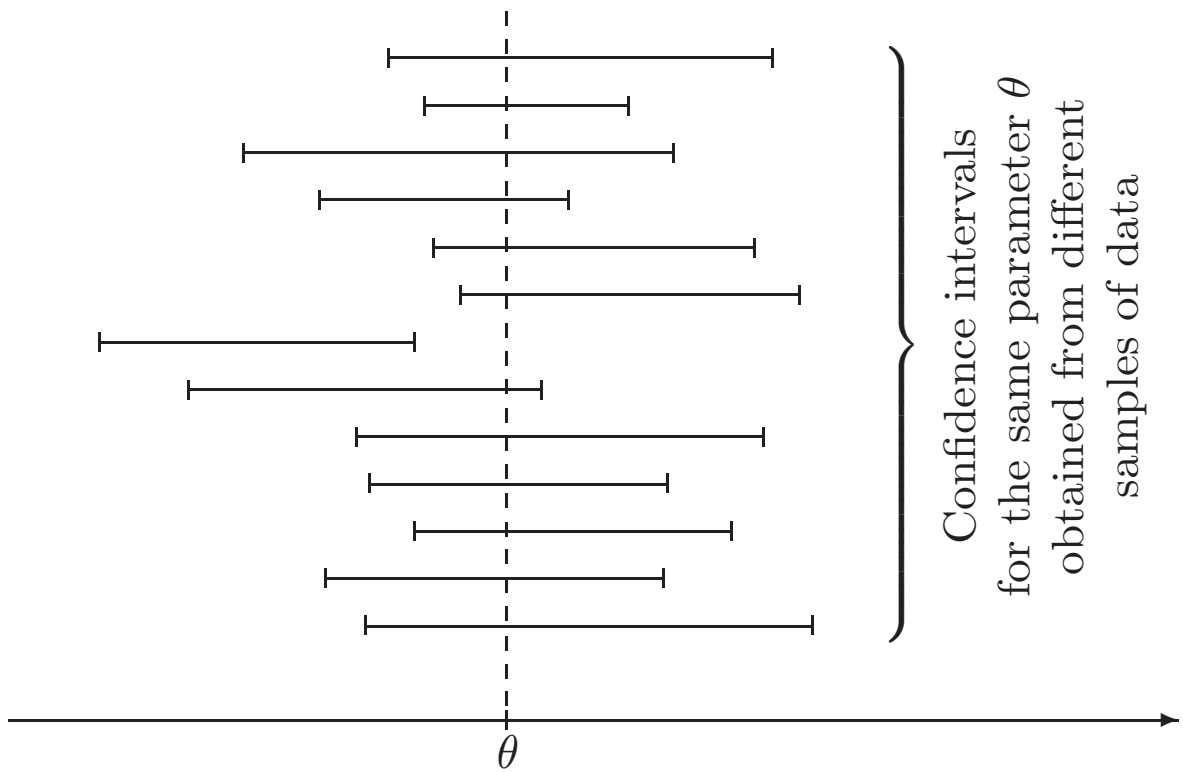
$$P\{a \leq \theta \leq b\} = 1 - \alpha$$

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.

Confidence intervals



Confidence intervals and coverage of parameter θ .

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

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

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

Therefore,

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

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

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

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 .

\bar{X} is approximately Normal for large n and *any* distribution of X_1, \dots, X_n .

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.

Hypotheses testing, Z-tests

Null hypothesis H_0	Parameter, estimator $\theta, \hat{\theta}$	If H_0 is true:		Test statistic $Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}}$
		$E(\hat{\theta})$	$\text{Var}(\hat{\theta})$	
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, \bar{X}	μ_0	$\frac{\sigma^2}{n}$	$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, \hat{p}	p_0	$\frac{p_0(1-p_0)}{n}$	$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y, \bar{X} - \bar{Y}$	D	$\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$	$\frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2 = D$	$p_1 - p_2, \hat{p}_1 - \hat{p}_2$	D	$\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$	$\frac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$

Hypothesis testing, t-tests

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n ; unknown σ	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m ; unknown but equal $\sigma_X = \sigma_Y$	$t = \frac{\bar{X} - \bar{Y} - D}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m ; unknown, unequal $\sigma_X \neq \sigma_Y$	$t = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Special formula