Indian Institute of Technology Madras

Module II

Nirav Bhatt
Email: niravbhatt@iitm.ac.in

Statistical Data Analysis

Module I

- Descriptive statistics: Data analysis through
  - Numerical computation of sample statistics: Mean, variance, mode, range, ...
  - Graphical representation: Organize, summarize, and visualize in terms of different types of graphs, Box plots, scattered plots...

Statistical Data Analysis

Module II

- Module II- Inference or inductive statistics: Data analysis for decision making
  - Parameter estimation: Determine unknown parameters from sample data
  - Hypothesis testing: Verify or validate a postulate (or hypothesis) regarding population(s) or parameters using the data

Module II

Statistical Hypothesis testing and confidence intervals

- Topics:
  - Point estimation of parameters → distribution / model, MOM / MLE
  - Confidence interval computation → $|x - \mu| = t, n, \sigma, \alpha$
  - Statistical hypothesis testing → p-value, z, t, χ^2-test, single popn.
- Learning Outcomes: Students should be able to
  - estimate parameters from observations
  - compute confidence intervals
  - formulate statistical hypothesis and run tests using data

Data types, form and variables

- Format: Images, texts, numbers, videos...
- Types:
  - Numerical (or quantitative)
    - Interval: Ordering of scale and difference between two values in data is meaningful
      GATE Scores, IQ Scores, credit score
    - Ratio: Interval with clear definition of absolute zero
      Height in meters, Weight in Kg, Concentrations...
  - Categorical (or qualitative)
    - Nominal: Categories with no order
      Patient's name or ID, color of t-shirt...
    - Ordinal: Categories with order but no different between values
      Grades, Weight in Healthy, overweight, obese

Parameter Estimation

- Example: Lethal dose of a medicine
- Important to know for assessing the overall efficacy of the medicine
- Variability due to Gender, BMI, Age, Geography...
- FDA needs a representative value or a range for lethal dose of a medicine
- Use sample data to compute a reasonable value of lethal dose
  Point estimate of lethal dose

Hypothesis testing

- Example: Two medicines A and B for a disease
- Scientist conjectures that A is better medicine for the disease
- How can you prove or reject the conjecture?
- If the scientists can perform experiments on different sets of patients having the same disease with both medicines and shows that A is better
- Need to collect data and a procedure to show that A is better medicine than B
  Statistical hypothesis testing
- Emphasis on the better medicine

Parameter Estimation

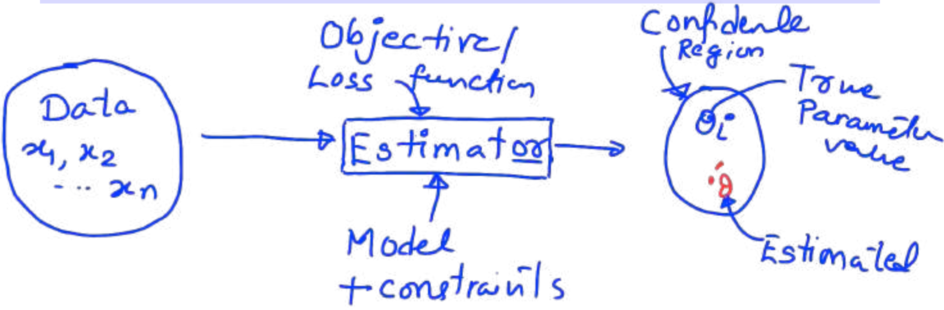
- Two problems of parameter estimation
  - Estimate parameters of a distribution from data — μ, σ^2
  - Estimate parameters of models from data $y = \alpha_1 x_1 + \alpha_2 x_2$, Estimate α_1, α_2
- Objectives of estimation: (i) Estimating parameters, and (ii) provide a goodness of estimated parameters
- Parameter estimation involves two steps:
  - Estimating parameters using methods for estimation
  - Assessing the "goodness" of the estimated parameters and provide bounds on variables

Parameter Estimation

Elements

Definition : Estimator

It is the process of inferring unknown parameters in a model or distribution from a given set of data and other information using a *mathematical map* between the unknowns parameters and the known information and a decision criterion.



Parameter Estimation

Types

- Point estimators: Produce single-valued estimates (more common)
  Examples: kinetic parameter estimates from data, mean height of person in the classroom, expected life of a mobile device....
- Interval estimators: Produce an interval
  Examples: catalyst particle size, age of the students in BT5450
- Other types: Non-parametric, Parametric, and semi-parametric
  Depends on the information available such as function and/or density distribution forms

Parameter Estimation

Random Sample

Random Sample

Consider RVs $X_1, X_2, \ldots, X_n$. These RVs are random sample of size n if
(i) the X_i's are independent RVs
(ii) item Each X_i is drawn from the same probability distribution

Random sample should be
- Representation of population
- Bias free

Statistics

A statistic is any function of the observation in a random sample, $\hat{\Theta} = g(X_1, X_2, \ldots, X_n)$

- $\hat{\Theta}$ is a random variable

- Example: means

m Random samples

$$\left\{\begin{array}{l} \{x_1^1, x_2^1, \ldots, x_n^1\} \longrightarrow \hat{\mu}_1 \\ \{x_1^2, x_2^2, \ldots, x_n^2\} \longrightarrow \hat{\mu}_2 \\ \vdots \\ \{x_1^m, x_2^m, \ldots, x_n^m\} \longrightarrow \hat{\mu}_m \end{array}\right\} \text{Different values}$$

Parameter Estimation

Sampling distribution

Sampling distribution

The probability density function of a statistic is called a sampling distribution

→ sample mean, $\bar{x}$ computed from random sample: $x_1 \ldots x_n$ for a population with $N(\mu, \sigma^2)$

→ Sampling distribution of $\bar{x}$

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Parameter Estimation

Point Estimator

- Random sample: $X_1\, X_2, \ldots, X_n$ with $f(x, \theta)$: Density function
- θ: Unknown parameters in column- vector form

Point Estimator

A point estimate of some population parameters θ is a single numerical vector-value $\hat{\theta}$ of a statistic. The statistic is called the point estimator.

Normal distribution:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{with } \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

objective fn $\rightarrow$ Estimator

Data $\rightarrow$ $\boxed{\text{Estimator}}$ $\rightarrow$ $\hat{\mu} = 2$

$f(x, \theta) \rightarrow$ Estimator; $\hat{\sigma}^2 = 3.1$

Parameter Estimation

Estimator

- Statistical properties of the estimate
  - Accuracy: How accurate is the estimate on the average?
  - Precision: Variability of the estimates obtained from different random samples?
- The given estimator gives an estimate with the least variability?
- What about true value of θ (θ_t) and $\hat{\theta}$ obtained from the estimator?
- How does the sample size n affect the value of estimate?

Role of n:

$n \to \infty$ (large sample size)

$\hat{\theta} \to \theta_t$?

Parameter Estimation

Unbiased Estimators

- How accurate is the estimate on the average?
  Closeness of estimate to the true values
- How close values can be computed using an Estimator $\hat{\Theta}$?

$$E(\hat{\Theta}) = \theta_t$$

Unbiased estimator

A point estimator $\hat{\Theta}$ is an unbiased estimator for the parameter θ if

$$E(\hat{\Theta}) = \theta_t$$

Parameter Estimation

Unbiased Estimators

Bias of an estimator

If the estimator is not unbiased estimator, the bias (b) can be computed as

$$b = E(\hat{\Theta}) - \theta_t$$

$E(\hat{\theta}) = \theta_b$, Then

Bias, $b = \theta_b - \theta_t$

If $\boxed{\theta_b \cong \theta_t}$, $b = 0$ & $\boxed{\hat{\theta} \rightarrow \text{unbiased Estimator}}$

Parameter Estimation

Example: Show that sample mean and variance are unbiased

Random samples : $X_1, X_2, \ldots, X_n$

$x_i \sim P(\mu, \sigma^2)$ P: Distribution

$i = 1, \ldots, n$

sample mean, $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

$$E[\bar{x}] = E\left[\sum_{i=1}^{n} \frac{x_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n} E[x_1 + x_2 + \ldots + x_n]$$

$$= \frac{1}{n} \{E[x_1] + E[x_2] + \ldots + E[x_n]\}$$

$$= \frac{1}{n} [\mu + \mu + \ldots + \mu] = \mu$$

$$\text{Bias} = E[\bar{x}] - \mu = \mu - \mu = 0$$

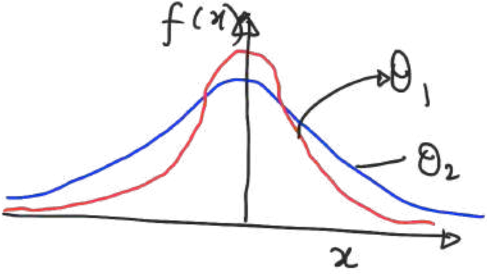
Parameter Estimation

Variance of a Point Estimator

- Two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$

$$E(\hat{\Theta})_1 = \boldsymbol{\theta}_t, \quad E(\hat{\Theta})_2 = \boldsymbol{\theta}_t$$

Question: Which one to choose?



$Var(\Theta_1) < Var(\Theta_2)$

choose the estimator with minimum variance

Parameter Estimation

Variance of a Point Estimator

Minimum variance unbiased estimator (MVUE)

Consider all the unbiased estimators (say total m) of $\boldsymbol{\theta}$ $(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m,)$, the one with the smallest variance is called MVUE.

$$\mathrm{Var}(\hat{\Theta}_2) < \mathrm{Var}(\hat{\Theta}_1) < \ldots \mathrm{Var}(\hat{\Theta}_m)$$

Θ_2 is MVUE

Parameter Estimation

Standard Error

- Precision and variability of an estimate?
  Standard error of the estimate

Standard Error of an Estimator

The standard error of of an estimator $\hat{\Theta}$ is its standard deviation given by

$$\hat{\sigma}_{\hat{\Theta}} = \sqrt{\text{Var}(\hat{\Theta})}$$

$E[\hat{\Theta}] = \mu$, $var(\hat{\Theta}) = \sigma^2$

$\hat{\sigma}_{\hat{\Theta}} = \sigma$

$\rightarrow$ Point estimate $\pm$ standard error

$\mu \pm \sigma$

Parameter Estimation

Mean Squared error of an Estimator

- Only biased estimators are available
  How to select an estimator?
- Means squared error of an estimator $\hat{\Theta}$ of the parameter θ

Means squared error

$$\text{MSE}(\hat{\Theta}) = E[(\hat{\Theta} - \theta_t)^2]$$

or

$$\begin{aligned}\text{MSE}(\hat{\Theta}) &= (E[\hat{\Theta} - E(\hat{\Theta})])^2 + (\theta_t - E(\hat{\Theta}))^2 \\ &= Var(\hat{\Theta}) + (\text{Bias})^2\end{aligned}$$

- Bias-variance trade-off : MSE $(\hat{\theta})$ is total of variance & Bias

Parameter Estimation

Mean Squared error of an Estimator

- Two estimators of the parameter θ: $\text{MSE}(\hat{\Theta}_1)$ and $\text{MSE}(\hat{\Theta}_2)$
- Relative efficiency of estimators

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)}$$

- Relative efficiency < 1: $\hat{\Theta}_1$ is a more efficient $\hat{\Theta}_2$

Parameter Estimation

Methods of Point Estimation

- Method of moments:
  Equate population moments to sample moments

Random Sample: $X_1, X_2, \ldots, X_n$ from a PMF or PDF with unknown p parameters $\boldsymbol{\theta}$. The moment estimators $\hat{\Theta}_1, \ldots, \hat{\Theta}_p$ can be found by equating the first p population moments to the first p sample moments and solving the set of nonlinear equations

Parameter Estimation

Methods of Point Estimation

- Method of moments:
  Equate population moments to sample moments

Random Sample: $X_1, X_2, \ldots, X_n$ from a PMF or PDF with unknown p parameters θ. The moment estimators $\hat{\Theta}_1, \ldots, \hat{\Theta}_p$ can be found by equating the first p population moments to the first p sample moments and solving the set of nonlinear equations

k^{th} moment: $E[X^k]$; $k=1$; $E[X] \rightarrow$ mean

$$\underbrace{E[X^k]}_{k^{th} \text{ moment for population}} = \underbrace{\hat{\Theta}^k(X_1, X_2, \ldots X_n)}_{k^{th} \text{ moment computed from data}}$$

Parameter Estimation

Method of moments : Example

Data: $x_1, x_2 \ldots x_n$; Drawn from $\mathrm{Exp}(\lambda)$

λ: Unknown parameter. Estimate λ using MoM

$$E[X^1] = \frac{1}{n} \sum_{i=1} x_i$$

$$\frac{1}{\lambda} = \frac{1}{n} \sum x_i$$

$$\Rightarrow \lambda = \frac{n}{\sum x_i}$$

Parameter Estimation

Maximum Likelihood Estimation

- RV $X \sim f(x, \theta)$, θ: Unknown parameters
- Observations $x_1, x_2, \ldots, x_n$
- The likelihood function of the sample is

$$L(\theta) = L(\theta / x_1, x_2, \ldots, x_n) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot f(x_3, \theta) \cdot \ldots \cdot f(x_n, \theta)$$

$\max_{\theta} L(\theta)$: Maximum likelihood estimator

$$\max_{\theta} L(\theta) = \max_{\theta} \prod_{i=1}^{n} f(x_i, \theta)$$

Parameter Estimation

Maximum Likelihood Estimation

Data: $x_1, \ldots x_n \sim$ Bernoulli R.V.

$$PMF: f(x, \theta) = p^{x}(1-p)^{1-x}, \quad x = 0,1$$
$$= 0 \quad , \text{otherwise}$$

Parameter to be estimated: p

$$\theta = [p] = p$$

$\rightarrow$ Construct $L(\theta)$

$$L(\theta) = f(x_1, p) \cdot f(x_2, p) \cdot \ldots f(x_n, p)$$
$$= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \ldots p^{x_n}(1-p)^{1-x_n}$$
$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Parameter Estimation

Maximum Likelihood Estimation

$$L(p) = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

$$\Leftrightarrow \ln L(p) = \sum x_i \ln p + (n - \sum x_i) \ln(1-p)$$

$$\frac{\partial L}{\partial p} = \frac{\partial \ln L(p)}{\partial p} = 0$$

$$\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0 \Rightarrow p = \frac{\sum x_i}{n}$$

Verify $\frac{\partial^2 \ln L(p)}{\partial p} < 0$

Parameter Estimation

Maximum Likelihood Estimator: Properties

- Unbiased estimator : For large n
- Variance of $\hat{\Theta}$ is nearly as small as the one that could be obtained with any other estimator
- $\hat{\Theta}$: An approximate normal distribution

Invariance Property

$$\left.\begin{matrix} \theta_1, \theta_2 \ldots \theta_p \\ \downarrow \text{MLE} \\ \hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_p \end{matrix}\right\} \rightarrow \begin{matrix} h(\theta_1, \ldots \theta_p) \\ \text{MLE} \\ h(\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p) \end{matrix}$$

Parameter Estimation

Bootstrapping Estimation

- Non-parametric approach of estimation *Unlike MLE and MoM:*

  No need for assumption about underlying distribution

- Often used for computing standard error and confidence intervals for relatively small sample size

- Uses sampling with replacement strategies

R. S. = {1, 2, 8, 9, 3, 6, 7}

sample 1: 1, 8, 9, 3, 9

Sample 2: 8, 8, 3, 6, 6

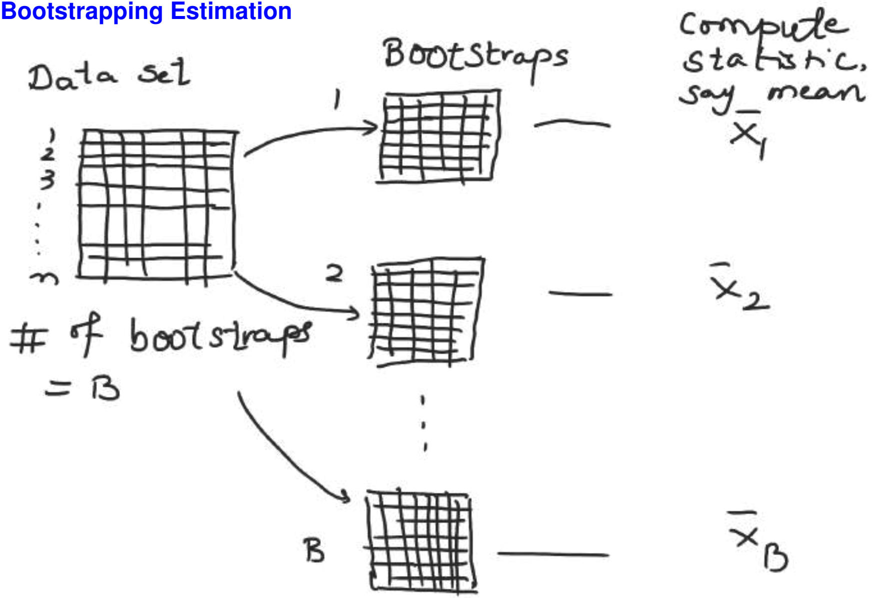
Parameter Estimation

Bootstrapping Estimation

- Samples: $X_1, X_2, \ldots, X_n$ drawn from independent and identical but unknown distribution
- Let $\hat{\Theta} = \hat{\Theta}(X_1, X_2, \ldots, X_n)$ be statistic

Parameter Estimation

Bootstrapping Estimation



Parameter Estimation

Bootstrapping Estimation

- Bootstrap means

$$
\begin{aligned}
\bar{X}_1 = &\quad \text{mean}(X_1^{*,1}, \ldots, X_n^{*,1}) \\
\bar{X}_2 = &\quad \text{mean}(X_1^{*,2}, \ldots, X_n^{*,2}) \\
\vdots & \\
\bar{X}_B = &\quad \text{mean}(X_1^{*,B}, \ldots, X_n^{*,B})
\end{aligned} \tag{1}
$$

- Bootstrap estimate of the variance

$$
\text{var}(\bar{X}) = \frac{1}{B-1} \sum_{i=1}^{B} (\bar{X}_i - \bar{X}_B)^2, \text{ with } \bar{X}_B = \frac{1}{B} \sum_{i=1}^{B} \bar{X}_i
$$

Confidence interval

Introduction

- Point estimate: How close to true value?
- Interested in knowing the variability of the population parameters
- Range of plausible values: Confidence interval
- An interval estimate for a population parameter is called a confidence interval
- *Confidence*: Specifies level of confidence 90%,95%, 99%
- Constructed so that it contains true unknown population parameter(s)

Confidence Interval

Introduction

- Random sample: $X_1, X_2, \ldots, X_n$
- Unknown Distribution, Unknown mean μ and Known variance σ^2
- Sample mean $\bar{X} \sim F(\mu, \sigma^2/n)$
- Standardize $\bar{X}$, New R.V.

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

Three types of intervals

1. Confidence Interval
2. Tolerence Interval
3. Prediction Interval

Confidence Interval

Introduction

- Confidence interval: lower and upper bounds,

$$l \leq \mu \leq u \qquad l, u : \text{Unknown}$$

- l and u: End points computed from the data
- l and u: Values of random variable L and U
- Question: How do we determine values of l and u

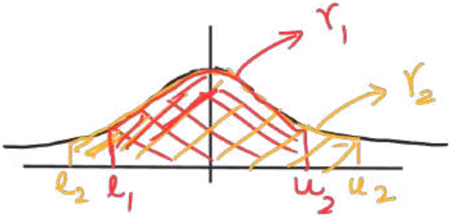
Random sample: $x_1, x_2, \ldots, x_n$

$$l = L(x_1, x_2, \ldots x_n)$$

$$u = U(x_1, x_2, \ldots x_n)$$

Confidence Interval

Introduction



- L and U: RVs
- Determine values of these RVs such that

$$P\{L \leq \mu \leq U\} = \gamma = 1 - \alpha, \quad 0 \leq \gamma, \alpha \leq 1$$

- From samples $x_1, x_2, \ldots, x_n$, l and u can be computed to determine CI with (1-α) probability

$$l \leq \mu \leq u$$

l and u: Lower and upper-confidence bounds

Confidence Interval

Introduction

Normal distribution $X \sim N(\mu, \sigma^2)$

Data: $X_1, X_2, \ldots, X_n$

μ is unknown and σ^2: known

Objective: Find interval for unknown μ.

Standard Normal R.V., $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

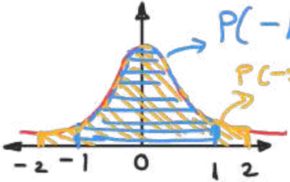
compute $\bar{x}$ from data

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Confidence Interval

Introduction

z-distribution:



$P(-1 \le Z \le 1) = 0.68 = 1 - 0.32$

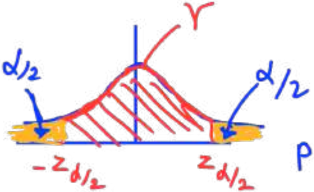
$P(-2 \le Z \le 2) = 0.96 = 1 - 0.04$

For $z \in [-1, 1]$, $\alpha = 0.32$

$z \in [-2, 2]$, $\alpha \cong 0.04$

Given α, $z_{\alpha/2}$ can be computed from z-tables

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = \gamma = 1 - \alpha$$



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

$$P\left(\underbrace{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{l} \le \mu \le \underbrace{\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{u}\right) = 1 - \alpha$$

Confidence Interval

CI and Precision

- 90% or 95% or 99%?
- $z_{\alpha/2}$ for $\alpha = 0.05$ and $\alpha = 0.01$

$$z_{0.025} = 1.96$$

$$z_{0.005} = 2.58$$

- Length of confidence intervals

$$95\%, \text{ Length} = 2(1.96\sigma/\sqrt{n}) = 3.92\sigma/\sqrt{n}$$

$$99\%, \text{ Length} = 2(2.58\sigma/\sqrt{n}) = 5.16\sigma/\sqrt{n}$$

Confidence Interval

Introduction

One-sided CI on the mean, Known σ^2

$100(1-\alpha)\%$ upper-confidence bound

$$\mu \le \bar{X} + z_\alpha \sigma/\sqrt{n}$$

Lower-confidence bound

$$\mu \ge \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Confidence Interval

Introduction

Large Sample, n, CI for μ

- σ^2 unknown but $n \sim$ large
- compute sample variance, s^2

$$Z \cong \bar{X} - \mu / s\sqrt{n} \quad \text{(Approximately z-distribution)}$$

- CI

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- Typically, in practice, $n \geq 40$

Confidence Interval

Introduction

- Error=$\|\bar{x} - \mu\|$
- For given σ, specify E, and α, then *n*: number of samples required

$$n = \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2$$

- For example, E=0.5, $\sigma = 2$, $\alpha = 0.05$, Then *n*

$$n = \left(\frac{(1.96)(2)}{0.5} \right)^2 = 61.5$$

- For $E = 0.25$, $n =$? $\sigma = 2$, $\alpha = 0.05$, $n \cong 246$
- $\sigma = 1$, $n =$? $E = 0.5$, $\alpha = 0.05$, $n = 16$
- For $\alpha = 0.01$, $n =$?, $z_{0.005} = 2.58$, $n = 107$

Confidence Interval

One-sided Confidence Bounds

- One-sided Confidence Bounds for a given α: Provides
  - Lower bound $l \leq \mu$
  - Upper bound $u \geq \mu$
- For a given *alpha* Computed by
  - Lower bound $\bar{x} - z_\alpha \sigma/\sqrt{n} \leq \mu \leq \infty$
  - Upper bound $\bar{x} + z_\alpha \sigma/\sqrt{n} \geq \mu \geq -\infty$

Confidence Interval

Unknown Population variance

- So far
  - n random samples, unknown μ, and known σ^2
- n random samples, unknown μ, and σ^2?
- Confidence interval for μ
- Sample variance, S^2 can be computed from n observations
- A statistic can be computed (on same line as z-statistic, $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$)

$$T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$$

- T is RV from t-distribution with $n-1$ degrees of freedom

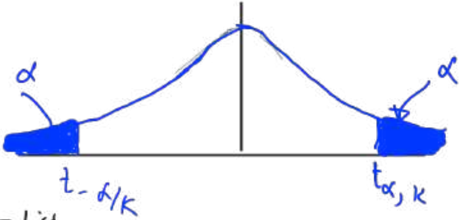
$$f(u) = \frac{\Gamma[n/2]}{\sqrt{\pi(n-1)}\,\Gamma(n-1/2)} \frac{1}{\left[\left(\frac{u^2}{n-1}\right)+1\right]^{n/2}} \quad -\infty < u < \infty$$

Confidence Interval

Unknown Population variance

- $T \sim$ t-dist. with $n-1$ degrees of freedom
- symmetric like z-dist.
- CI for μ with unknown σ^2
- $100(1-\alpha)\%$ CI

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} s/\sqrt{n} \le \mu \le \bar{x} + t_{-\alpha/2, n-1} s/\sqrt{n}$$



- Large n, z-dist. $\sim$ t-dist.

Confidence Interval

CI for σ^2 of a Normal Distribution

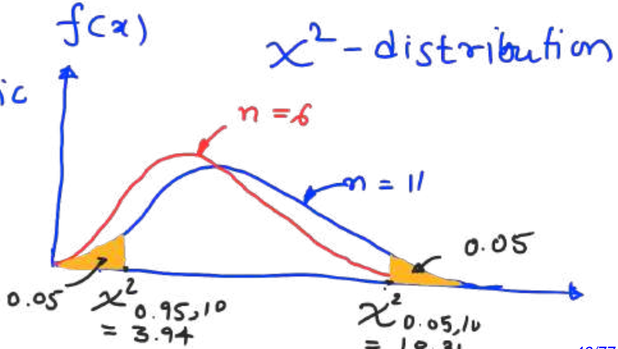
- χ^2-distribution for n samples from $\mathcal{N}(\mu, \sigma^2)$
- X^2-statistic (or RV)

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

- $X^2 \sim \chi^2_{n-1}$

- Not symmetric

$$P(X^2 > \chi^2_{\alpha,k}) = \int_{\chi^2_{\alpha,k}}^{\infty} f(u)\,du = \alpha$$



Confidence Interval

confidence Interval on the variance

Two-sided $100(1-\alpha)\%$ with n observations

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}, \quad s^2: \text{computed sample variance}$$

Interpolating of CI:

- L & V : Random Variables
- CI : Random interval
- Interpretation: If large number of random samples are collected then $100(1-\alpha)\%$ of these CI will contain the true value of statistic (mean, variance)

Confidence Interval

Parameter of interest	symbol	Other parameter	Confidence Interval $100(1-\alpha)\%$	
Mean: Normal distribution	μ	σ^2: known	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	Z
Mean: Arbi. distribution large size	μ	σ^2: Not known compute s^2 from data	$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$	Z
Mean: Normal distribution	μ	σ^2: Not known compute s^2 from data	$\bar{x} - t_{\alpha/2, n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$	T
Variance: Normal distribution	σ^2	Mean μ unknown and estimate μ & s^2	$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$	χ^2

Confidence Interval

Summary : $100(1-\alpha)\%$

Parameter of interest	Lower-bound	Upper-bound
μ, both σ^2 known & σ^2 unknown but large n	$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu$	$\bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \ge \mu$
μ & σ^2 unknown	$\bar{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}} \le \mu$	$\bar{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}} \ge \mu$
σ^2 & μ unknown	$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \le \sigma^2$	$\sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}$

Hypothesis Testing
Example

- Treatments for a disease: T-A and T-B
- Claim: T-A is better than T-B
- Practitioners question: Does T-A better than T-B?
- Approach to answer practitioners question: **Hypothesis testing: Decision making process**

Hypothesis Testing

Introduction

- Claim: T-A is better than T-B
- Claim(s) or statement(s): Statistical Hypothesis (es)
  Statement about the parameters of one or more populations
- Claim: T-A is better than T-B: Claim to population's parameter
- Practitioners' interest: mean number of days to recuperate from the appearance of clinical symptoms

Hypothesis Testing

Introduction

- Practitioners' interest: mean number of days to recuperate from the appearance of clinical symptoms

Population mean

- μ_{T-A} and μ_{T-B}: mean days to recuperate for treatment T-A and T-B
- $\mu_{T-A} > \mu_{T-B}$
- Formal re-casting of statement as two hypotheses:

$$H_0 : \mu_{T-A} = \mu_{T-B}$$

$$H_1 : \mu_{T-A} > \mu_{T-B}$$

- H_0: Null hypothesis: Both treatments are same
- H_1: Alternative hypothesis: T-A is better than T-B

Hypothesis Testing

Introduction

- One-sided alternative hypothesis

$$H_0 : \mu_{T-A} = \mu_{T-B} \quad H_1 : \mu_{T-A} > \mu_{T-B}$$

  or

$$H_0 : \mu_{T-A} = \mu_{T-B} \quad H_1 : \mu_{T-A} < \mu_{T-B}$$

- Claim: Mean number of days to recuperate for T-A is 8 days or $\mu_{T-A} = 8$
- Two-sided alternative hypothesis

$$H_0 : \mu_{T-A} = 8 \ \ days \quad H_1 : \mu_{T-A} \neq 8 \ \ days$$

- By convention: Null hypothesis is an equality claim

Hypothesis Testing
Elements

- *Hypothesis: A statement about the population or model or distribution not about the sample*

  Use sample to verify hypothesis

- Truth or falsity of a claim (or hypothesis) is never known in practical situation

- Hypothesis testing: Probabilistic approach to reach a conclusion based on population parameter(s)

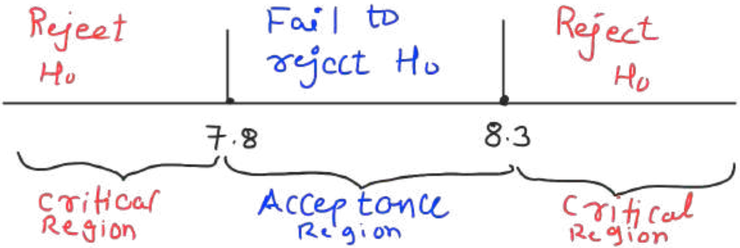
Hypothesis Testing

Elements

- Two-sided alternative hypothesis

$$H_0: \mu_{T-A} = 8 \;\; days \quad H_1: \mu_{T-A} \neq 8 \;\; days$$

- Samples are available for T-A different patients
- Sample mean: Take on many different values
- Let us define range (recall CI): $7.8 \leq \bar{x} \leq 8.3$
- Acceptance region: any value in the range
- Critical regions: outside the acceptance region



Hypothesis Testing
Elements

- Pitfall I:
  - $\mu_{T-A} = 8$: Truth
  - Random sample selected: $\bar{x} = 8.7$
  - Outside acceptance region ($7.8 \leq \bar{x} \leq 8.3$)
  - Reject H_0 in favor of H_1
    Wrong conclusion $\rightarrow$ Type I error

Type I Error

Rejecting H_0 when it is true is defined as a type I error

Type II Error

Failing to reject H_0 when it is false is defined as a type II error

Hypothesis Testing
Elements

Decision	H_0 is true	H_0 is false
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

- Quantifying Type I and II errors
- Type I error: Probability of rejecting H_0 when H_0 is true

Hypothesis Testing

Elements

Decision	H_0 is true	H_0 is false
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

- Quantifying Type I and II errors
- Type I error: Probability of rejecting H_0 when H_0 is true

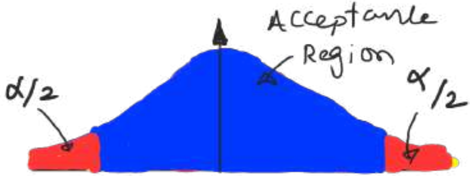
Hypothesis Testing
Elements

Type I error

Rejecting the H_0 when it is true is defined as a type I error.

Probability of Type I error

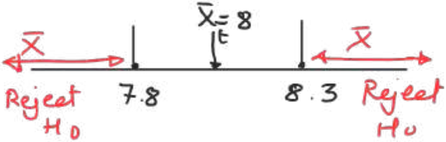
α= P(type I error)=P(reject H_0 when H_0 is true)



Hypothesis Testing

Elements

$$P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$



$\sigma = 1, \; n = 100$

$$P(\text{Type I error}) = P(\bar{x} \leq 7.8 \text{ when } \bar{x}_t = 8) + P(\bar{x} \geq 8.3 \text{ when } \bar{x}_t = 8)$$

$$= P\left(z_1 \leq \frac{7.8 - 8}{1/10}\right) + P\left(z_2 \geq \frac{8.3 - 8}{1/10}\right)$$

$$= \alpha$$

Hypothesis Testing

Elements

Type II error

β=P(Probability of Type II error)=P(fail to reject H_0 when H_0 is false)

claim: Average weight of students: 50 kg

$H_0: \mu = 50, \quad H_1: \mu \neq 50$

True mean, $\mu = 52$

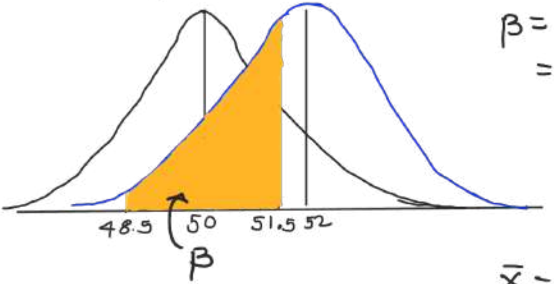
Acceptanle region: $48.5 \leq \bar{X} \leq 51.5$

$\beta = P(48.5 \leq \bar{x} \leq 51.5 \text{ when } \mu = 52)$

Hypothesis Testing

Elements

$H_0: \mu = 50 \quad H_1: \mu = 52$



$$\beta = P(48.5 \le X \le 50.5, \mu = 52)$$
$$= P(X \le 51.5) - P(X \le 48.5)$$

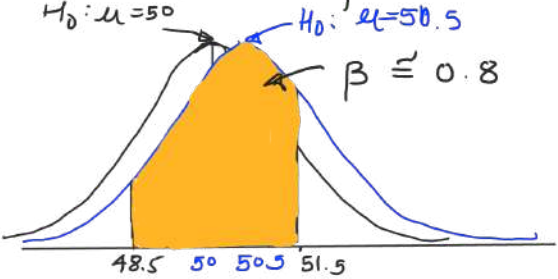
For $\sigma/\sqrt{n} = 0.8$, $\quad z_1 = \dfrac{\bar{x} - 52}{\sigma/\sqrt{n}}$

$$\beta = P(z_1 \le -0.62) - P(z_1 \le -2.5)$$
$$= 0.26$$

Hypothesis Testing

Elements

Instead of $\mu = 52$, $\mu = 50.5$



$$\beta = P(48.5 \le X \le 51.5,\ \mu = 50.5)$$
$$\cong 0.8$$

Type II error is higher when $\mu = 50.5$.

Hypothesis Testing

Elements

Important points

1. The size of the critical region can be reduced by type I error, α.

2. For given sample size, n, decrease in the probability of one type error results in an increase in the other type.

3. For given α, increase n reduces β

4. Value of β decreases as the diff. between the true mean and the hypothesized value increases

Hypothesis Testing

Elements

- β: Not constant, depends on true value of parameter and sample size
- Extent of falsity of null hypothesis
- Accept H_0: Weak conclusion
- Failing to reject H_0: Strong conclusion

↳ Does not mean "H_0 is true", but, it means "more data are required to make strong conclusion".

Hypothesis Testing

Power

- Power:
  Power of statistical test: Probability of rejecting null Hypothesis H_0 when H_1 is true
- Power $=1 - \beta$
- Power: Probability of correctly rejecting a false H_0

→ Power is used to compare two statistical test

→ Useful measure of sensitivity of a statistical test.

Hypothesis Testing

Elements

one-sided hypothesis:

claim involving phrases "greater than", less than or at least ...

"one-sided hypothesis testing"

- Appropriate alternative hypothesis test has to be chosen.

- one-sided H_1; Rejecting H_0 is a strong conclusion.

Hypothesis Testing

Elements

- α: Fixed significance level
- α: Doesn't provide any idea location of parameters in critical region

P-value

The P-value is the smallest level of significance that would lead to rejection of H_0 with the given data.

P-value: Observed significance level

— Provides how significant the data are

Hypothesis Testing

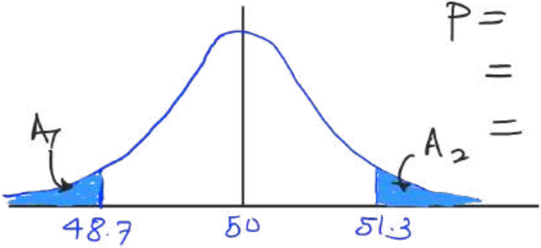
Elements

Two-sided hypothesis test

$H_0: \mu = 50, \quad H_1: \mu \neq 50, \; n=16,$

$\sigma = 2.5$

- Observed sample mean, $\bar{x} = 51.3$



$$P = A_1 + A_2$$

$$= 1 - P(48.7 < \bar{x} < 51.3)$$

$$= 1 - 0.962 = 0.038$$

It indicates $\bar{x} = 51.3$ is a rare event when $p = 0.038$

Hypothesis Testing

General Procedure for Hypothesis tests

Context specific

1. Parameter of interest: Identify the parameter of interest for a context
2. Null hypothesis, H_0: State the null hypothesis
3. Alternative hypothesis, H_1: Specify an appropriate alternative hypothesis
4. Test statistic: Determine an appropriate test statistic
5. Reject H_0 if: State the rejection criteria for the null hypothesis
6. Computations: Compute any necessary sample quantities and value of test statistic
7. Draw conclusions: Decide whether or not H_0 should be rejected and report that in the problem context

P1: Population, σ^2 known, mean S1

S2 $H_0: \mu = \mu_0$, S3 $H_1: \mu \neq \mu_0$

Data: $X_1, X_2, \ldots, X_n$

Hypothesis test on the mean

S4 $$\boxed{Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}, \quad z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

(S5) choose α value

$\rightarrow$ $z_{\alpha/2}$ & $-z_{\alpha/2}$

If Z_0 computed from data

$-z_{\alpha/2} \le z_0 \le z_{\alpha/2}$, Fail to reject H_0

We Reject H_0 if

$z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$

(5b) compute sample mean $\bar{x}$ from data & compute z_0

(S7) z_0^c location if

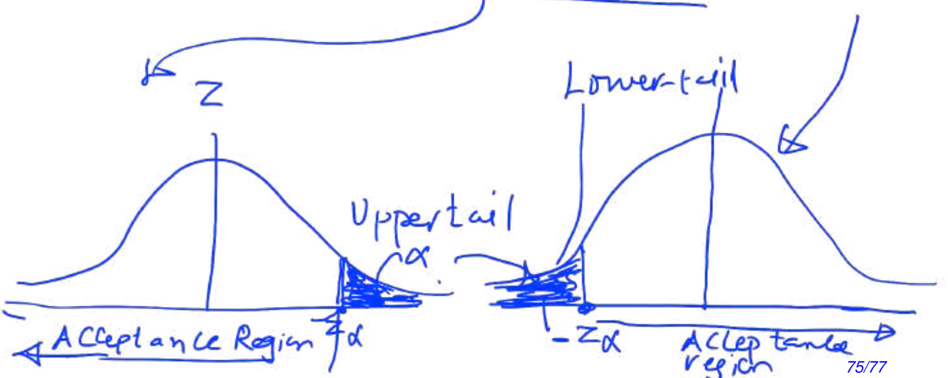
$z_0^c < -z_{\alpha/2}$ or $z_0^c > z_{\alpha/2}$

Reject H_0

(S1) mean

(S2) One-sided; $H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$



S4 $Z \to$ test statistic

S5 $\mu > \mu_0$ or $\mu < \mu_0$

$Z_0 > Z_\alpha$ $\quad$ $Z_0 < -Z_\alpha$

S6 $$z_0^c = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

S7 $z_0^c > Z_\alpha$ or $z_0^c < -Z_\alpha$

for Reject H_0

P2: Population mean is parameter of interest. n : Large size, σ^2 unknown

Apply P1: $Z_0 = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ sample S.D.

P3: Population mean : parameter of interest. $n \neq$ large, σ^2 unknown

$$T_0 = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1} \text{ - distn.}$$

S5 Two-sided test;

$t_0 > t_{\alpha/2, n-1}$; $t_0 < t_{-\alpha/2, n-1}$

Reject H_0

One-sided test:

Reject H_0 $\begin{cases} t_0 > t_{\alpha, n-1} & \text{for } H_1: \mu > \mu_0 \\ t_0 < -t_{\alpha, n-1} & \text{for } H_1: \mu < \mu_0 \end{cases}$

P4: Variance is parameter of interest, Population μ & σ^2 unknown

Null Hypothesis

$H_0: \sigma^2 = \sigma_0^2$

Reject H_0

Two-sided $H_1: \sigma^2 \neq \sigma_0^2$ $\quad \chi_0^{c\,2} > \chi^2_{\alpha/2, n-1}$ or $\chi_0^{c\,2} < \chi^2_{1-\alpha/2, n-1}$

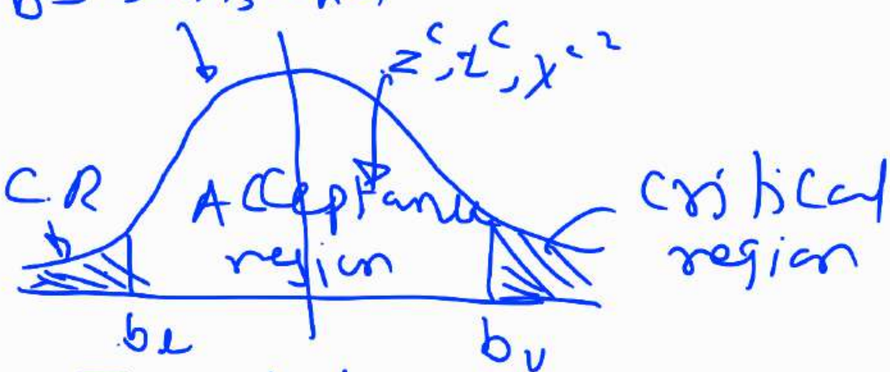
One-sided $H_1: \sigma^2 > \sigma_0^2 \longrightarrow \chi_0^{c\,2} > \chi^2_{\alpha, n-1}$

$H_1: \sigma^2 < \sigma_0^2 \longrightarrow \chi_0^{c\,2} < \chi^2_{1-\alpha, n-1}$

Test statistics

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

$b = z, t_{n-1}, \chi^2_{n-1}$



Two-sided

$$\left.\begin{array}{ll} z & b_l = -z_{\alpha/2}, \quad b_u = z_{\alpha/2} \\ t & b_l = -t_{\alpha/2, n-1}, \quad b_u = t_{\alpha/2, n-1} \\ \chi^2 & b_l = \chi^2_{\alpha/2, n-1}, \quad b_u = \chi^2_{1-\alpha/2, n-1} \end{array}\right\} \alpha$$

---

one-sample hypothesis testing

$P_1 : \mu = \mu_1$, σ_1^2 $\qquad$ $P_2 : \mu = \mu_2$, σ_2^2

$H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$

Z-

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$