Advanced Topics in Software Engineering: Statistical Inference

Prof. Michele Loreti

Advanced Topics in Software Engineering
Corso di Laurea in Informatica (L31)
Scuola di Scienze e Tecnologie

Statistical Inference. . .

Statistical Inference (or learning) is the process of using data to infer the distribution that generated the data.

Statistical Inference. . .

Statistical Inference (or learning) is the process of using data to infer the distribution that generated the data.

Problem: We observe $X_1, \ldots, X_n$ having CDF F, we want to infer/estimate/learn F or some feature of F (such as its mean).

Statistical Inference...

Statistical Inference (or learning) is the process of using data to infer the distribution that generated the data.

Problem: We observe $X_1, \ldots, X_n$ having CDF F, we want to infer/estimate/learn F or some feature of F (such as its mean).

In our context X_i will be the outcome of a simulation!

Statistical models...

A statistical model is a set of distributions $\mathfrak{F}$.

Statistical models. . .

A statistical model is a set of distributions $\mathfrak{F}$.

A parametric model is a set $\mathfrak{F}$ that can be parametrised by a finite number of parameters.

Statistical models. . .

A statistical model is a set of distributions $\mathfrak{F}$.

A parametric model is a set $\mathfrak{F}$ that can be parametrised by a finite number of parameters.

A nonparametric model is a set $\mathfrak{F}$ that cannot be parametrised by a finite number of parameters

$$\mathfrak{F} = \left\{ F \middle| F \text{ is a CDF} \right\}$$

Parametric models. . .

Example. Parametric model for data coming from a *Normal distribution*:

$$\mathfrak{F} = \left\{ f(x : \mu, \sigma) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \middle| \mu \in \mathbb{R}, \sigma > 0 \right\}$$

Parametric models. . .

Example. Parametric model for data coming from a *Normal distribution*:

$$\mathfrak{F} = \left\{ f(x : \mu, \sigma) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \middle| \mu \in \mathbb{R}, \sigma > 0 \right\}$$

In general a parametric model takes the form:

$$\mathfrak{F} = \left\{ f(x; \theta) \middle| \theta \in \Theta \right\}$$

where:

- θ is an unknown parameter (or vector of parameters);
- Θ is the parameter space.

Regression, prediction and classification. . .

Suppose that we observe pairs of data. . .

$$(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$$

Regression, prediction and classification. . .

Suppose that we observe pairs of data. . .

$$(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$$

Example: X_i is the *blood pressure* of a subject i and Y_i is how long they live.

Regression, prediction and classification. . .

Suppose that we observe pairs of data. . .

$$(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$$

Example: X_i is the *blood pressure* of a subject i and Y_i is how long they live.

X is called predictor/regressor/feature/independent variable.

Regression, prediction and classification. . .

Suppose that we observe pairs of data. . .

$$(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$$

Example: X_i is the *blood pressure* of a subject i and Y_i is how long they live.

X is called predictor/regressor/feature/independent variable.

Y is called outcome/response variable/dependent variable.

Regression, prediction and classification. . .

Suppose that we observe pairs of data. . .

$$(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$$

Example: X_i is the *blood pressure* of a subject i and Y_i is how long they live.

X is called predictor/regressor/feature/independent variable.

Y is called outcome/response variable/dependent variable.

The regression function is the function

$$r(x) = E(Y|X = x)$$

Regression, prediction and classification. . .

Let us assume that $r \in \mathfrak{F}$. . .

Regression, prediction and classification...

Let us assume that $r \in \mathfrak{F}$...

...if we know something about the structure of $\mathfrak{F}$ (e.g. is a set of linear function) we have a parametric regression model

Regression, prediction and classification. . .

Let us assume that $r \in \mathfrak{F}$. . .

. . . if we know something about the structure of $\mathfrak{F}$ (e.g. is a set of linear function) we have a parametric regression model

The goal of predicting Y for a new patients based on their X values is called prediction.

Regression, prediction and classification...

Let us assume that $r \in \mathfrak{F}$...

...if we know something about the structure of $\mathfrak{F}$ (e.g. is a set of linear function) we have a parametric regression model

The goal of predicting Y for a new patients based on their X values is called prediction.

If Y is *discrete* the prediction is called classification.

Corner Stones of Inference...

Point Estimation **Confidence Sets** **Hypothesis Testing**

Point Estimation. . .

Point Estimation refers to providing a single best guess of some quantity of interest.

Point Estimation…

Point Estimation refers to providing a single best guess of some quantity of interest.

The quantity of interest can be…

- a parameter in a parametric model;
- a CDF F;
- a regression function r;
- …

Point Estimation...

Point Estimation refers to providing a single best guess of some quantity of interest.

The quantity of interest can be...

- a parameter in a parametric model;
- a CDF F;
- a regression function r;
- ...

By convention we denote a point to estimate of θ by $\hat{\theta}$...

Point Estimation...

Point Estimation refers to providing a single best guess of some quantity of interest.

The quantity of interest can be...

- a parameter in a parametric model;
- a CDF F;
- a regression function r;
- ...

By convention we denote a point to estimate of θ by $\hat{\theta}$...

- θ is a fixed, unknown quantity;
- $\hat{\theta}$ is a random variable.

Point Estimation...

Let X_1,..., X_n be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of X_1,...,X_n:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

Point Estimation...

Let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

We let the bias of $\widehat{\theta}_n$

$$\text{bias}(\widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta$$

Point Estimation...

Let $X_1,\ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1,\ldots,X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

We let the bias of $\widehat{\theta}_n$

$$\mathrm{bias}(\widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta$$

A model is unbiased if $E(\widehat{\theta}_n) = \theta$.

Point Estimation...

Let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

We let the bias of $\widehat{\theta}_n$

$$\text{bias}(\widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta$$

A model is unbiased if $E(\widehat{\theta}_n) = \theta$.

A point estimator $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \xrightarrow{P} \theta$.

Point Estimation. . .

The distribution $\widehat{\theta}_n$ is called the sampling distribution.

Point Estimation. . .

The distribution $\widehat{\theta}_n$ is called the sampling distribution.

The standard deviation of $\widehat{\theta}_n$ is called the standard error, denoted by se:

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{V[\widehat{\theta}_n]}$$

Point Estimation...

The distribution $\widehat{\theta}_n$ is called the sampling distribution.

The standard deviation of $\widehat{\theta}_n$ is called the standard error, denoted by se:

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{V[\widehat{\theta}_n]}$$

Often it is not possible to compute the standard error but usually we can estimate it. The estimated standard error is denoted by $\widehat{\mathsf{se}}$.

Point Estimation. . .

The quality of a point estimate is sometimes assessed by the mean squared error (MSE):

$$\mathsf{MSE} = E(\widehat{\theta}_n - \theta)$$

Point Estimation...

The quality of a point estimate is sometimes assessed by the mean squared error (MSE):

$$\text{MSE} = E(\widehat{\theta}_n - \theta)$$

MSE can be written as

$$\text{MSE} = \text{bias}(\widehat{\theta}_n)^2 + V(\widehat{\theta}_n)$$

Point Estimation...

The quality of a point estimate is sometimes assessed by the mean squared error (MSE):

$$\text{MSE} = E(\widehat{\theta}_n - \theta)$$

MSE can be written as

$$\text{MSE} = \text{bias}(\widehat{\theta}_n)^2 + V(\widehat{\theta}_n)$$

If bias $\rightarrow 0$ and se $\rightarrow 0$ as $n \rightarrow \infty$ then $\widehat{\theta}_n$ is consistent, that is $\widehat{\theta}_n \xrightarrow{P} \theta$

Point Estimation. . .

An estimator is asymptotically Normal if

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1)$$

Point Estimation. . .

An estimator is asymptotically Normal if

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1)$$

Mean is asymptotically Normal!

Confidence Sets

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where

$$a = g_a(X_1, \ldots, X_n) \qquad b = g_b(X_1, \ldots, X_n)$$

are functions such that:

$$Pr(\theta \in C_n) \geq 1-\alpha$$

for all $\theta \in \Theta$.

Confidence Sets
Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence,
$p = Pr(X_i = 1) = E[X_i]$.

Confidence Sets

Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = Pr(X_i = 1) = E[X_i]$.

The $1 - \alpha$ confidence interval for p can be computed as:

$$C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n) \qquad \text{where} \qquad \epsilon_n = \frac{\log(\frac{2}{\alpha})}{2n}$$

Confidence Sets
Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = Pr(X_i = 1) = E[X_i]$.

The $1 - \alpha$ confidence interval for p can be computed as:

$$C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n) \qquad \text{where} \qquad \epsilon_n = \frac{\log(\frac{2}{\alpha})}{2n}$$

This because $X_1, \ldots, X_n$ have a Bernulli distribution with parameter p and that for any $\epsilon > 0$:

$$Pr(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Normal-based Confidence Interval

Suppose that $\widehat{\theta}_n \approx N(\theta, \mathsf{se}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, that is (where $Z \sim N(0,1)$):

$$Pr(Z > z_{\alpha/2}) \geq \alpha/2 \qquad \text{and} \qquad Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

Normal-based Confidence Interval

Suppose that $\widehat{\theta}_n \approx N(\theta, \mathsf{se}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$, that is (where $Z \sim N(0,1)$):

$$Pr(Z > z_{\alpha/2}) \geq \alpha/2 \qquad \text{and} \qquad Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Let

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{se}, \widehat{\theta}_n + z_{\alpha/2}\widehat{se})$$

Normal-based Confidence Interval

Suppose that $\widehat{\theta}_n \approx N(\theta, \mathrm{se}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, that is (where $Z \sim N(0,1)$):

$$Pr(Z > z_{\alpha/2}) \geq \alpha/2 \qquad \text{and} \qquad Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$$

Let

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{se}, \widehat{\theta}_n + z_{\alpha/2}\widehat{se})$$

Then

$$Pr(\theta \in C_n) \to 1-\alpha$$

Normal-based Confidence Interval

Suppose that $\widehat{\theta}_n \approx N(\theta, \mathsf{se}^2)$. Let Φ be the CDF of a standard Normal and let $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$, that is (where $Z \sim N(0, 1)$):

$$Pr(Z > z_{\alpha/2}) \geq \alpha/2 \qquad \text{and} \qquad Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Let

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{se}, \widehat{\theta}_n + z_{\alpha/2}\widehat{se})$$

Then

$$Pr(\theta \in C_n) \to 1 - \alpha$$

For $1 - \alpha = 0.95$ (95% confidence interval) $\alpha = 0.05$, $z_{\alpha/2} = 1.96 \approx 2$.

Normal-based Confidence Interval

Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = Pr(X_i = 1) = E[X_i]$.

Normal-based Confidence Interval
Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = Pr(X_i = 1) = E[X_i]$.

We know by the *Central Limit Theorem* that $\widehat{p}_n \approx N(p, \widehat{\text{se}}^2)$ where:

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \widehat{\text{se}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

Normal-based Confidence Interval

Example

Consider flipping a coin for which the probability of *heads* is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = Pr(X_i = 1) = E[X_i]$.

We know by the *Central Limit Theorem* that $\widehat{p}_n \approx N(p, \widehat{\text{se}}^2)$ where:

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \widehat{\text{se}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

An approximate $1-\alpha$ confidence interval is:

$$\widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

From theory to practice. . .

Empirical Distribution Function

Let $X_1, \ldots, X_n \sim F$ be IID where F is a CDF on the real line.

Empirical Distribution Function

Let $X_1, \ldots, X_n \sim F$ be IID where F is a CDF on the real line.

We can estimate F via the empirical distribution function $\widehat{F}_n$ defined as follows:

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}$$

Plug-in estimator

A statistical functional $T(F)$ is any function of F, such at *mean*, *variance*, *median*. . .

Plug-in estimator

A statistical functional $T(F)$ is any function of F, such at *mean*, *variance*, *median*. . .

The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

Plug-in estimator

A statistical functional $T(F)$ is any function of F, such at *mean*, *variance*, *median*. . .

The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

This is obtained by using $\widehat{F}_n$ for the unknown F.

Plug-in estimator

A statistical functional $T(F)$ is any function of F, such at *mean*, *variance*, *median*...

The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

This is obtained by using $\widehat{F}_n$ for the unknown F.

For $T(\widehat{F}_n)$ we can compute a Normal-based interval by assuming:

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\text{se}})$$

Plug-in estimator

A statistical functional $T(F)$ is any function of F, such at *mean*, *variance*, *median*...

The plug-in estimator of $\theta = T(F)$ is defined by

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

This is obtained by using $\widehat{F}_n$ for the unknown F.

For $T(\widehat{F}_n)$ we can compute a Normal-based interval by assuming:

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\text{se}})$$

Warning: calculating $\widehat{\text{se}}$ is not easy in the general case!

Plug-in estimator

Examples

Let $X_1, \ldots, X_n \sim F$ be IID where F is an unknown CDF:

Plug-in estimator
Examples

Let $X_1, \ldots, X_n \sim F$ be IID where F is an unknown CDF:

Mean:

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

Plug-in estimator

Examples

Let $X_1, \dots, X_n \sim F$ be IID where F is an unknown CDF:

Mean:

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

Variance:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

The two are equivalent for large values of n.

From theory to practice. . .

Bootstrap method

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals.

Bootstrap method

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals.

Let $T_n = g(X_1, \ldots, X_n)$ be a statistic, that is, any function of the data.

Bootstrap method

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals.

Let $T_n = g(X_1, \ldots, X_n)$ be a statistic, that is, any function of the data.

Suppose we want to know $V_F(T_n)$, the variance of T_n. **This value depends on the unknown distribution function F.**

Bootstrap method

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals.

Let $T_n = g(X_1, \ldots, X_n)$ be a statistic, that is, any function of the data.

Suppose we want to know $V_F(T_n)$, the variance of T_n. **This value depends on the unknown distribution function F.**

The idea of bootstrap method is to approximate F with $\widehat{F}_n$.

Simulation

Suppose we draw an IID sample $Y_1, \ldots, Y$ď$_B$ from a distribution G. By the law of large numbers we have that when $B \to \infty$:

$$\overline{Y}_n = \frac{1}{B} \sum_{j=1}^{B} Y_j \xrightarrow{P} E[Y]$$

Simulation

Suppose we draw an IID sample $Y_1, \ldots, Y$ď$_B$ from a distribution G. By the law of large numbers we have that when $B \to \infty$:

$$\overline{Y}_n = \frac{1}{B} \sum_{j=1}^{B} Y_j \xrightarrow{P} E[Y]$$

We can use sample mean $\overline{Y}_n$ to approximate $E[Y]$. In a simulation we can make B as large as we like.

Simulation

More generally, if h us a function with finite mean then when $B \to \infty$ then:

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{P} E[Y]$$

Simulation

More generally, if h us a function with finite mean then when $B \to \infty$ then:

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{P} E[Y]$$

In particular:

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y})^2 = \frac{1}{B}\sum_{j=1}^{B} Y_j^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y_j\right)^2 \xrightarrow{P} V[Y]$$

Bootstrap Variance Estimation

We can approximate $V_{\widehat{F}_n}[T_n]$ by simulation.

Bootstrap Variance Estimation

We can approximate $V_{\widehat{F}_n}[T_n]$ by simulation.

Boostrap Variance Estimation

1. Draw $X_1, \dots, X_n$ from F
2. For $i = 0$ to m do:
    - Sample $X^*_{i_1}, \dots, X^*_{i_n}$ from $\widehat{F}_n$
    - Let $T^*_i = g(X^*_{i_1}, \dots, X^*_{i_n})$
3. Let

$$v_{boot} = \frac{1}{m} \sum_{i=1}^{m} \left(T^*_i - \frac{1}{m} \sum_{j=1}^{m} T^*_j \right)$$

Bootstrap Variance Estimation

Warning: We are using **two** approximations:

$$V_F[T_n] \approx V_{\widehat{F}_n}[T_n] \approx v_{boot}$$

From Theory to Practice

Bootstrap Confidence Interval

Normal Interval: let $\widehat{se}_{boot}$ be the bootstrap estimate of the standard error

$$T_n \pm z_{\alpha/2}\widehat{se}_{boot}$$

Bootstrap Confidence Interval

Normal Interval: let $\widehat{se}_{boot}$ be the bootstrap estimate of the standard error

$$T_n \pm z_{\alpha/2}\widehat{se}_{boot}$$

The interval is not accurate unless T_n is close to Normal.

Bootstrap Confidence Interval

Let $\theta = T(F)$ and $\widehat{\theta_n} = T(\widehat{F}_n)$. The pivot $R_n = \widehat{\theta}_n - \theta$.

Bootstrap Confidence Interval

Let $\theta = T(F)$ and $\widehat{\theta_n} = T(\widehat{F}_n)$. The pivot $R_n = \widehat{\theta}_n - \theta$.

Let $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,m}$ denote the bootstrap replications of $\widehat{\theta}_n$.

Bootstrap Confidence Interval

Let $\theta = T(F)$ and $\widehat{\theta_n} = T(\widehat{F}_n)$. The pivot $R_n = \widehat{\theta}_n - \theta$.

Let $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,m}$ denote the bootstrap replications of $\widehat{\theta}_n$.

Let $H(r)$ denote the CDF of the pivot:

$$H(r) = Pr_F(R_n \leq r)$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$Pr(a \leq \theta \leq b)$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$Pr(a \leq \theta \leq b) \quad = \quad Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n)$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$\begin{aligned} Pr(a \leq \theta \leq b) &= Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ &= Pr(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \end{aligned}$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$\begin{array}{rcl} Pr(a \leq \theta \leq b) & = & Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ & = & Pr(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\ & = & H(\widehat{\theta} - a) - H(\widehat{\theta} - b) \\ & = & H\left(H^{-1}(1 - \frac{\alpha}{2})\right) - H\left(H^{-1}(\frac{\alpha}{2})\right) \end{array}$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$\begin{aligned} Pr(a \leq \theta \leq b) &= Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ &= Pr(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\ &= H(\widehat{\theta} - a) - H(\widehat{\theta} - b) \\ &= H\left(H^{-1}(1 - \tfrac{\alpha}{2})\right) - H\left(H^{-1}(\tfrac{\alpha}{2})\right) \\ &= 1 - \tfrac{\alpha}{2} - \tfrac{\alpha}{2} = 1 - \alpha \end{aligned}$$

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$\begin{array}{rcl} Pr(a \leq \theta \leq b) & = & Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ & = & Pr(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\ & = & H(\widehat{\theta} - a) - H(\widehat{\theta} - b) \\ & = & H\left(H^{-1}(1 - \frac{\alpha}{2})\right) - H\left(H^{-1}(\frac{\alpha}{2})\right) \\ & = & 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{array}$$

Good news: C_n^* is an exact $1 - \alpha$ interval for θ!

Bootstrap Confidence Interval

We can consider $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{and} \qquad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

We have that (a and b are random variables):

$$\begin{array}{rcl} Pr(a \leq \theta \leq b) & = & Pr(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ & = & Pr(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\ & = & H(\widehat{\theta} - a) - H(\widehat{\theta} - b) \\ & = & H\left(H^{-1}(1 - \frac{\alpha}{2})\right) - H\left(H^{-1}(\frac{\alpha}{2})\right) \\ & = & 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{array}$$

Good news: C_n^* is an exact $1 - \alpha$ interval for θ!
Bad news: H is unknown!

Bootstrap Confidence Interval

We can form a bootstrap estimate of H:

$$\widehat{H}(r) = \frac{1}{m}\sum_{j=1}^{m} I(R^*_{n,j} \leq r)$$

where $R^*_{n,j} = \widehat{\theta}^*_{n,j} - \widehat{\theta}_n$. We let r^*_β and θ^*_β denote the β sample quantiles of $(R^*_{n,1}, \dots, R^*_{n,m})$ and $(\theta^*_{n,1}, \dots, \theta^*_{n,m})$.

Bootstrap Confidence Interval

We can form a bootstrap estimate of H:

$$\widehat{H}(r) = \frac{1}{m}\sum_{j=1}^{m} I(R^*_{n,j} \leq r)$$

where $R^*_{n,j} = \widehat{\theta}^*_{n,j} - \widehat{\theta}_n$. We let r^*_β and θ^*_β denote the β sample quantiles of $(R^*_{n,1}, \ldots, R^*_{n,m})$ and $(\theta^*_{n,1}, \ldots, \theta^*_{n,m})$.

An approximate $1 - \alpha$ confidence interval $C_n = (\widehat{a}, \widehat{b})$ is:

$$\begin{array}{lllllll} \widehat{a} & = & \widehat{\theta}_n - \widehat{H}^{-1}(1 - \frac{\alpha}{2}) & = & \widehat{\theta}_n - r^*_{1-\frac{\alpha}{2}} & = & 2\widehat{\theta}_n - \theta^*_{1-\frac{\alpha}{2}} \\ \widehat{b} & = & \widehat{\theta}_n - \widehat{H}^{-1}(\frac{\alpha}{2}) & = & \widehat{\theta}_n - r^*_{\frac{\alpha}{2}} & = & 2\widehat{\theta}_n - \theta^*_{\frac{\alpha}{2}} \end{array}$$

Bootstrap Confidence Interval

We can form a bootstrap estimate of H:

$$\widehat{H}(r) = \frac{1}{m}\sum_{j=1}^{m} I(R^*_{n,j} \leq r)$$

where $R^*_{n,j} = \widehat{\theta}^*_{n,j} - \widehat{\theta}_n$. We let r^*_β and θ^*_β denote the β sample quantiles of $(R^*_{n,1}, \dots, R^*_{n,m})$ and $(\theta^*_{n,1}, \dots, \theta^*_{n,m})$.

An approximate $1 - \alpha$ confidence interval $C_n = (\widehat{a}, \widehat{b})$ is:

$$\begin{array}{lclclcl} \widehat{a} & = & \widehat{\theta}_n - \widehat{H}^{-1}(1 - \frac{\alpha}{2}) & = & \widehat{\theta}_n - r^*_{1-\frac{\alpha}{2}} & = & 2\widehat{\theta}_n - \theta^*_{1-\frac{\alpha}{2}} \\ \widehat{b} & = & \widehat{\theta}_n - \widehat{H}^{-1}(\frac{\alpha}{2}) & = & \widehat{\theta}_n - r^*_{\frac{\alpha}{2}} & = & 2\widehat{\theta}_n - \theta^*_{\frac{\alpha}{2}} \end{array}$$

Bootstrap pivotal confidence interval C_n is typically pointwise, asymptotic confidence interval.

Bootstrap Confidence Interval

Bootstrap percentile interval is defined as

$$C_n = \left(\theta^*_{\frac{\alpha}{2}}, \theta^*_{1-\frac{\alpha}{2}}\right)$$

From Theory to Practice

Hypothesis testing. . .

In our bike sharing system we have to choose among two different allocation policy $P1$ and $P2$ in terms of balanced use of resources.

Hypothesis testing. . .

In our bike sharing system we have to choose among two different allocation policy $P1$ and $P2$ in terms of balanced use of resources.

We built two models (M_{P1} and M_{P_2}) that (based on some assumptions about the utilisation environment) can be used to predict system behaviour.

Hypothesis testing. . .

In our bike sharing system we have to choose among two different allocation policy $P1$ and $P2$ in terms of balanced use of resources.

We built two models (M_{P1} and M_{P_2}) that (based on some assumptions about the utilisation environment) can be used to predict system behaviour.

We can consider two hypothesis:

- **The Null Hypothesis**, $P1$ is worst than $P2$;
- **The Alternative Hypothesis**, $P1$ is not worst than $P2$.

Hypothesis testing. . .

In our bike sharing system we have to choose among two different allocation policy $P1$ and $P2$ in terms of balanced use of resources.

We built two models (M_{P1} and M_{P_2}) that (based on some assumptions about the utilisation environment) can be used to predict system behaviour.

We can consider two hypothesis:

- **The Null Hypothesis**, $P1$ is worst than $P2$;
- **The Alternative Hypothesis**, $P1$ is not worst than $P2$.

If we observe that performance in M_{P1} is much better than that observed in M_{P2} we reject the null hypothesis in favour of alternative hypothesis.

Hypothesis testing. . .

Suppose that we partition the parameter space Θ in two disjoint sets Θ_0 and Θ_1 and that we wish to test:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

Hypothesis testing...

Suppose that we partition the parameter space Θ in two disjoint sets Θ_0 and Θ_1 and that we wish to test:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

We call:

- H_0 the null hypothesis;
- H_1 the alternative hypothesis.

Hypothesis testing. . .

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subseteq \mathcal{X}$ called the rejection region.

Hypothesis testing. . .

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subseteq \mathcal{X}$ called the rejection region.

If $X \in R$ we **reject** the null hypothesis, otherwise, we **do not reject** the null hypothesis:

- $X \in R \Longrightarrow$ reject H_0;
- $X \notin R \Longrightarrow$ retain (do not reject H_0.

Hypothesis testing...

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subseteq \mathcal{X}$ called the rejection region.

If $X \in R$ we **reject** the null hypothesis, otherwise, we **do not reject** the null hypothesis:

- $X \in R \implies$ reject H_0;
- $X \notin R \implies$ retain (do not reject H_0.

Usually the rejection region R is of the form $R = \{x : T(x) > c\}$ where T is a test statistic and c is a critical value.

Hypothesis testing...

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subseteq \mathcal{X}$ called the rejection region.

If $X \in R$ we **reject** the null hypothesis, otherwise, we **do not reject** the null hypothesis:

- $X \in R \Longrightarrow$ reject H_0;
- $X \notin R \Longrightarrow$ retain (do not reject H_0.

Usually the rejection region R is of the form $R = \{x : T(x) > c\}$ where T is a test statistic and c is a critical value.

The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate cutoff value c.

Hypothesis testing. . .

When hypothesis testing is applied there are two types of errors we can make.

Hypothesis testing. . .

When hypothesis testing is applied there are two types of errors we can make.

Type I error: we reject H_0 when H_0 is true.

Hypothesis testing. . .

When hypothesis testing is applied there are two types of errors we can make.

Type I error: we reject H_0 when H_0 is true.

Type II error: we reject H_1 when H_1 is true.

Hypothesis testing. . .

When hypothesis testing is applied there are two types of errors we can make.

Type I error: we reject H_0 when H_0 is true.

Type II error: we reject H_1 when H_1 is true.

Possible outcomes of hypothesis testing are:

	Retain Null	**Reject Null**
H_0 is true	OK	Type I error
H_1 is true	Type II error	OK

Hypothesis testing...

Power function, size and level of a test

The power function of a test with rejection region R is defined by

$$\beta(\theta) = Pr_{\theta}(X \in R)$$

Hypothesis testing. . .

Power function, size and level of a test

The power function of a test with rejection region R is defined by

$$\beta(\theta) = Pr_{\theta}(X \in R)$$

The size of a test is defined to be:

$$sup_{\theta \in \Theta_0} \beta(\theta)$$

Hypothesis testing. . .

Power function, size and level of a test

The power function of a test with rejection region R is defined by

$$\beta(\theta) = Pr_\theta(X \in R)$$

The size of a test is defined to be:

$$sup_{\theta \in \Theta_0} \beta(\theta)$$

A test is said to have a level α if its size is less than or equal to α.

Hypothesis testing. . .

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.

Hypothesis testing. . .

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.

A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called composite hypothesis.

Hypothesis testing...

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.

A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called composite hypothesis.

A test of the form

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_0 : \theta \neq \theta_0$$

is called two-sided test.

Hypothesis testing...

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.

A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called composite hypothesis.

A test of the form

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_0 : \theta \neq \theta_0$$

is called two-sided test.

Hypothesis testing. . .

A hypothesis of the form $\theta = \theta_0$ is called a simple hypothesis.

A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called composite hypothesis.

A test of the form

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_0 : \theta \neq \theta_0$$

is called two-sided test.

A test of the form

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_0 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \qquad \text{versus} \qquad H_0 : \theta < \theta_0$$

is called one-sided test.

The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\text{se}}$ be the estimated standard error of $\widehat{\theta}$.

The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

Consider testing:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

Consider testing:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

Assume that $\widehat{\theta}$ is asymptotically Normal:

$$\frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0, 1)$$

The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

Consider testing:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

Assume that $\widehat{\theta}$ is asymptotically Normal:

$$\frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1)$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$ where:

$$W = \frac{\widehat{\theta} - \theta}{\widehat{\mathsf{se}}}$$

The Wald Test

The Wald test has asymptotically size α:

$$Pr_{\theta_0}(|Z| > z_{\alpha/2}) \to \alpha$$

as $n \to \infty$.

The Wald Test

The Wald test has asymptotically size α:

$$Pr_{\theta_0}(|Z| > z_{\alpha/2}) \to \alpha$$

as $n \to \infty$.

Suppose that θ is $\theta_\star \neq \theta_0$. The power $\beta(\theta_\star)$ (that is the probability of correctly rejecting the null hypothesis) is (approximatively):

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathrm{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathrm{se}}} - z_{\alpha/2}\right)$$

Compute $z_{\alpha/2}$

1. Divide α by two;
2. Subtract what you obtain from .5;
3. Find the value in the $z - table$.

Compute $z_{\alpha/2}$

1. Divide α by two;
2. Subtract what you obtain from .5;
3. Find the value in the $z - table$.

Easy approach, use table for recurrent values of α:

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.1	0.05	1.645
95%	0.05	0.025	1.96
98%	0.02	0.01	2.326
99%	0.01	0.005	2.576

Wald test...

Comparing two means...

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y$ď$_n$ be two independent samples from populations with means μ_1 and μ_2, respectively.

Wald test. . .

Comparing two means. . .

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y$ď$_n$ be two independent samples from populations with means μ_1 and μ_2, respectively.

Let's test for null hypothesis that $\mu_1 = \mu_2$, that we can write as:

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_1 : \delta \neq 0$$

where $\delta = \mu_1 - \mu_2$.

Wald test. . .

Comparing two means. . .

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y$ď$_n$ be two independent samples from populations with means μ_1 and μ_2, respectively.

Let's test for null hypothesis that $\mu_1 = \mu_2$, that we can write as:

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_1 : \delta \neq 0$$

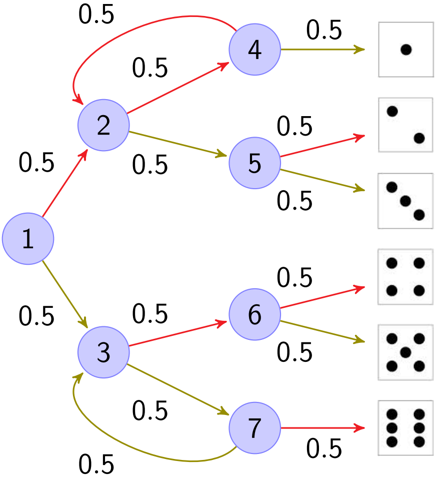
where $\delta = \mu_1 - \mu_2$.

The size α Wald test reject H_0 when $|W| > z_{\alpha/2}$

$$W = \frac{\widehat{\delta} - 0}{\widehat{\text{se}}} = \frac{\overline{X} - \overline{Y}}{\widehat{\text{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Wald Test in Action. . .

We can use Wald test to check correctness of Knut-Yao Algorithm:



To be continued. . .