TMA4268 SL M1
08.01.2018

Notation & useful results

n: number of units, individuals, elements

Y_i: response, output: univariate $\underbrace{\text{random variable}}$ (RV)

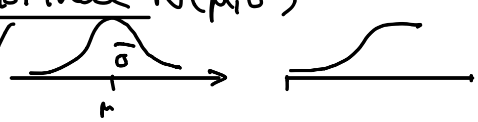
[multivariate L 17.01] (cov.mat + mvN + F)

- population under study (BP in Framingham)
- draw unit from population
  - observe Y_i

simple random sample

continuous (BP) → normal $N(\mu, \sigma^2)$

binary, categorical
(spam or not) → binomial$(1, p)$
(iris species) → multinomial



f: pdf, $F = P(Y_i \le y_i)$: cdf, $\mu = E(Y_i)$, $\sigma^2 = Var(Y_i)$ $E((Y_i-\mu)^2)$

⇒ see Rintermediate for "repetition with R"

Q: how to explain what f is for Y continuous?

$i = 1, \ldots, n$ indedendent Y_i's $\sim N(\mu, \sigma^2)$ USEFUL RESULTS

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad \frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

new to many of you

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1} \quad + \text{ also } \quad F = \frac{\chi^2_a / a}{\chi^2_b / b} \sim F_{a,b}$$

p : number of inputs, predictors, features, covariates

(p+1 with intercept)

$$\begin{array}{c} \\ \underset{n \times p}{\mathbf{X}} \end{array} = \begin{array}{ccc} \text{input } 1 & \dots & \text{input } p \\ \left[\begin{array}{ccc} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{array}\right] \end{array} \begin{array}{l} \leftarrow \text{obs } 1 \\ \vdots \\ \text{obs } n \end{array} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ \mathbf{x}_p]$$

$i = 1, \dots, n$
$j = 1, \dots, p$

x_i^T row vector with $x_{i1}, \dots, x_{ip}$

(not bf in book)

$\mathbf{x}_j$ column vector with $x_{1j} \cdots x_{nj}$

(bf in book)

Book : RV's : capital normal (not bold) font : Y

Matrices : bold capitals : $\mathbf{X}$

Vectors : lower case normal font, but $x_i \in \mathbb{R}^p$

lower case bold font if column vector of length n : $\mathbf{x}_j$

lecturer: add dimensions if unclear

Matrix algebra: A^T transpose , A^{-1} inverse $(AA^{-1} = I)$

AB matrix multiplication

we will also work with eigenvalues and eigenvectors (PCA)

I identity matrix $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
diag(1)

[LINK Härdle & Simar (2015) Ch 2

$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ column vector of 1's

Observe : (Y_i, x_i) $i = 1, \dots, n$ pairs

$\mathbb{R}$; $\{0,1\}$ $x_i \in \mathbb{R}^p$

$\{1,2,3\}$

often independent pairs.

What is the connection between x_i and Y_i?

Often assume $Y_i = f(x_i) + \varepsilon_i$ where ε_i are errors

but also other connection possible ($E(Y_i)$ vs x_i's).

3

— ADDED AFTER CLASS FROM HERE —

FINALLY:

- parameter estimator (bias, variance)
- confidence interval CI
- hypothesis test
- (prediction)

the key topics of statistical inference

$Y_1, \dots, Y_n$ random sample (independent)

and let $E(Y_i) = \mu$, $Var(Y_i) = \sigma^2$

Statistic T (e.g $\bar{Y} = \frac{1}{n}\sum Y_i$ or $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

or $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$)

with some sampling distribution

Parameter estimator T for τ. If $E(T) = \tau$ = unbiased

else $(E(T) - \tau)^2$ squared bias

$Var(T)$ variance of T. Before, mainly focus on $\overset{\text{T's where}}{E(T) = \tau}$, but now also $\overset{\text{biased estimator end on}}{\text{on } Var(T)}$ (and thus also on biased estimators)

$$\text{Mean squared error} = \overset{\substack{\text{distr. of } T \\ \downarrow}}{E}\left((T - \tau)^2 \right)$$

T = statistic to estimate τ

$$= \int_{-\infty}^{\infty} (t - \tau)^2 f(t)\, dt$$

$E(T)$ into this

→ bias & variance term next week.

Confidence interval: (via "most popular case")

Let $Y_1, \dots, Y_n$ i.i.d $N(\mu, \sigma^2)$, then $T = \dfrac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

and a $(1-\alpha)100\%$ CI for μ is

need statistic with known distribution!

quantiles from this!

$$P\left(-t_{\frac{\alpha}{2}, n-1} \le T \le t_{\frac{\alpha}{2}, n-1}\right) = 1-\alpha \iff$$

$$\left[\bar{Y} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}\right]$$ has probability $(1-\alpha) \cdot 100\%$

for covering the true μ. And when data for $\bar{Y}$ and S^2 inserted, then we have a $(1-\alpha)100\%$ CI.

Not assuming N-data, then asymptotically $(1-\alpha)100\%$ CI

n large

$$\left[\bar{y} \pm z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

$\alpha = 0.05 \to z_{0.025} = 1.96$

We have $(1-\alpha)100\%$ confidence that the true μ is in the interval.

Hypothesis test and p-value

Ex: $Y_1, .., Y_n$ i.i.d $N(\mu, \sigma^2)$, for example BP

$H_0: \mu \le 120$ vs $H_1: \mu > 120$

↑ ideal

[120-140 prehigh
140- high]

Does members of this population have an average systolic BP above 120? (or 140?)

Reject H_0 when $\bar{Y}$ is high.

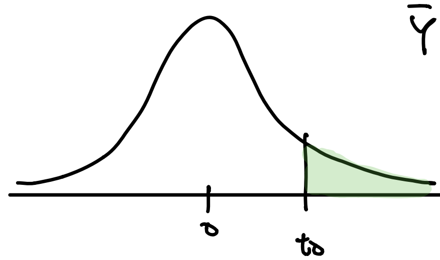
observed ↓

p-value : $P(\text{observe ave. systolic bp} \ge \bar{y} \mid H_0 \text{ true})$

↑ $\bar{Y} \sim N(120, \sigma^2/n)$

$= P(T_0 \ge t_{obs})$

$T_0 = \dfrac{\bar{Y} - 120}{S/\sqrt{n}}$, $\quad t_{obs} = \dfrac{\bar{y} - 120}{s/\sqrt{n}}$



Reject H_0 when p-value is low (below chosen α).

Type I error = crime of injustice = reject H_0 when H_0 is true

type II error = not detect H_0 wrong when it is wrong

power = $1 - P(\text{Type II error})$

↑ $> 80\%$ ideally!