STAT 231

November 4, 2016.

Roadmap.

- PREDICTION INTERVALS USING THE GAUSSIAN DISTRIBUTION.

- Introduction to Hypothesis Testing

  Null hyp

  Alternate "

  p-value

  Type I / II error.

Time Series Analysis

$Y_1, \dots Y_n$ r.v.s

Gaussian: $Y_1, \dots Y_n \sim G(\mu, \sigma)$

indep.

μ and σ unknown.

Sample $= \{y_1, y_2, \dots y_n\}$

HISTORICAL DATA.

Objective: To construct a 95% Prediction Interval for Y_{n+1}, the next value in your sequence.

Applications

(i) y_i = # of children born in Canada in month i

Predict

$(y_1, \ldots y_n)$ $\quad$ y_{n+1} ?

(ii) <u>Secretary problem</u> :

Hiring a person for a job.

You have to predict the future values of the quality of applicants.

(iii) $Y_1, \ldots Y_n$ = Stock price of Blackberry for the last n months

Predict $\underline{Y_{n+1}}$.

$\{Y_1, \ldots, Y_n\}$

Model:

$$Y_i \sim G(\mu, \sigma)$$

$i = 1, \ldots, n.$

Predict $Y_{n+1} = ?$

$$Y_{n+1} \sim G(\mu, \sigma) \quad \text{— (1)}$$

$$\bar{Y} = \frac{1}{n} \sum Y_i \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{— (2)}$$

Y_{n+1} and $\bar{Y}$ are independent

$$Y_{n+1} - \bar{Y} \sim G\left(0, \sqrt{\sigma^2 + \frac{\sigma^2}{n}}\right).$$

$$X \sim G(\mu_1, \sigma_1^2)$$

$$Y \sim G(\mu_2, \sigma_2^2)$$

$$X - Y \sim G(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$\boxed{Y_{n+1} - \bar{Y} \sim G\left(0, \sigma\sqrt{1 + \frac{1}{n}}\right)}$$

(3)

$$V(aX + bY) = a^2 V(X) + b^2 V(Y)$$

$$\left.\begin{aligned} a &= 1 \\ b &= -1 \end{aligned}\right\}$$

$$\frac{Y_{n+1} - \bar{Y}}{\sigma\sqrt{1+\frac{1}{n}}} = Z$$

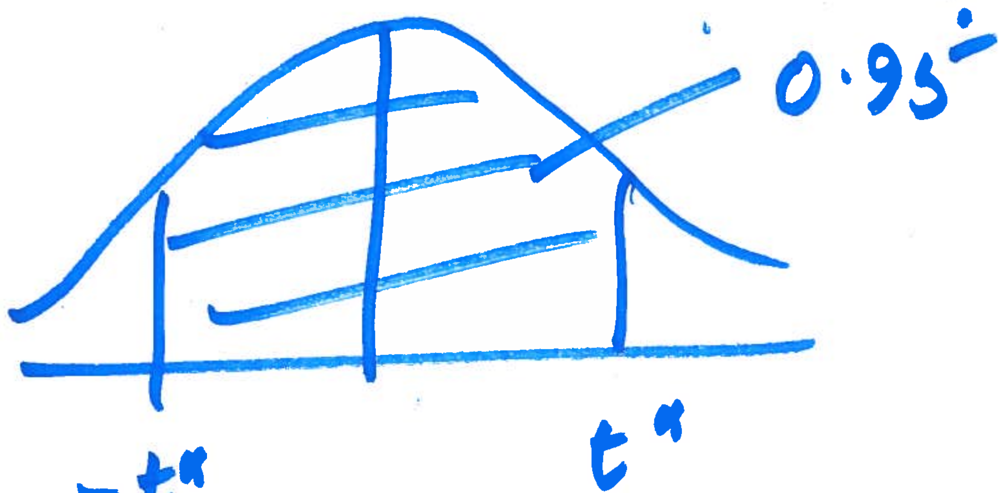
PIVOTAL DISTRIBUTION

PIVOTAL QTY

$$\boxed{\frac{Y_{n+1} - \bar{Y}}{S\sqrt{1+\frac{1}{n}}} \sim T_{n-1}}$$

Suppose $n = 10$ 95% P.I

Go to the t-table and find t^*



$df = n - 1.$ Column = 0.975

$$P\left(-t^\alpha < T < t^\alpha\right) = 0.95$$

$$P\left(-t^\alpha < \frac{Y_{n+1} - \bar{Y}}{S\sqrt{1 + \frac{1}{n}}} < t^\alpha\right) = 0.95$$

Coverage Interval

$$\boxed{\bar{Y} \pm t^{\alpha} S\sqrt{1+\frac{1}{n}}}$$

Confidence Interval

$$\boxed{\bar{y} \pm t^{\alpha} s\sqrt{1+\frac{1}{n}}}$$

Unrealistic, since, in most cases Y_i's are correlated, not independent.

Model needs to be modified for correlated variables.

(worry about trends and seasonality).

HYPOTHESIS TESTING.

<u>Definition</u>: A hypothesis is a statement about the population (or some attribute of the population)

Two competing hypotheses

H_0 : H_{nought} = Null hypothesis

(conventional wisdom / current belief)

H_1 : H_A : Alternate hypothesis

(Challenger)

Our job is to collect a sample and check whether there is sufficient evidence for/against H_0

We use the p-value approach.

Analogy: Legal System.

H_0: Banerjee is innocent

H_1: Banerjee is guilty

Jury has two choices

Convict: (Rejecting H_0, if there is strong evidence against it)

Acquit (do not have enough evidence to reject H_0)

p-value = P(observing your evidence (or worse) given that H_0 is true)

Reject the null hypothesis for low p-values.

$p < 0.05$ (cut-off)

H_0 and H_1 are not treated symmetrically

The burden of proof is on the challenger, not on H_0.

Unless there is "overwhelming evidence" against H_0, we do not reject it

	Reject H_0	Do not Reject H_0	
H_0 is true	✗	✓	Type I error
H_1 is true.	✓	✗	Type II error

The two errors are not treated symmetrically.

We typically want the Type I error $< 5\%$
