2/16/21

8. Statistical Inference

Estimation:

$X_1, X_2, \ldots X_n$ → sample observation

statistic → Functions of sample observation

$\bar{x}, s^2$

Estimation is choose a good estimator to estimate unknown population parameter.

$\bar{x}$ is estimator for μ:

$\bar{X}$ → numerical value is estimate

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

↓ estimator ↳ estimate

Point Estimation: single value as estimate

$\mu = 25$

$\sigma^2 = 25$

Interval Estimation: Range of values as estimate

$\mu \in (10, 20)$ → Confidence interval

Methods for finding estimators:

Method of moments, Maximum likelyhood estimation.

Properties of Estimators:

Unbiased Estimator: $T(X_1, X_2, \ldots X_n) = T(\bar{X})$ is Estimation

unbiased estimator if $E(T(x)) = g(\theta)$

↓ parameter of population

$T(x)$ is unbiased estimator of $g(\theta)$

$E(T(x)) = g(\theta) + b(\theta) \Rightarrow T(x)$ is biased estimator of $g(\theta)$

↓ bias

Ex: $X \sim Bin(n, P)$

$$E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{nP}{n} = P = g(\theta)$$

$\frac{X}{n}$ is unbiased estimator of P.

$$E\left(\frac{X(X-1)}{n(n-1)}\right) = \frac{1}{n(n-1)} E(X(X-1))$$

$$= \frac{1}{n(n-1)} n(n-1)p^2$$

$$= p^2 = g(\theta)$$

$\frac{X(X-1)}{n(n-1)}$ is unbiased estimator of p^2.

Ex: $X_1, X_2, \ldots X_n \sim Poison(\lambda)$

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$E(\bar{X}) = \frac{1}{n} \cdot n\lambda = \lambda$$

$\bar{X}$ is unbiased estimator of λ.

$\rightarrow$ $\bar{X}$ is unbiased estimator of μ $(E(\bar{X}) = \mu)$

s^2 is unbiased estimator of σ^2 $(E(s^2) = \sigma^2)$

Consistency: An estimator is said to be consistent estimator if

$$\lim_{n \to \infty} T(X_1, X_2 \ldots X_n) \longrightarrow g(\theta)$$

$$\lim_{n \to \infty} \left(P\left(|T(x) - g(\theta)|\right) > \epsilon\right) \longrightarrow 0$$

$\lim_{n\to\infty} P\left(|T(x) - g(\theta)| < \epsilon\right) \to 1$

Ex $X_1, X_2, \ldots X_n \quad E(X_i) = \theta, \quad Var(X_i) = \sigma^2$

$\bar{X}$ is consistent for θ

$$P(|\bar{X} - \theta| > \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2} \quad \text{(Chebyshew's Inequality)}$$

$$\leq \frac{\sigma^2}{n\epsilon^2}$$

$$\to 0$$

As $n \to \infty$ RHS $\to 0$

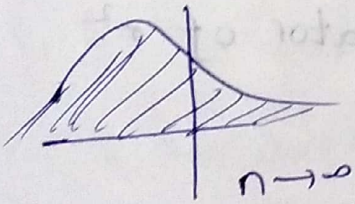
$\bar{X}$ is consistent estimator of θ

Ex $T(X) = \frac{X_1 + X_2 + \cdots + X_n}{n} = \bar{X}$ Normal distribution

$$P\{|T(x) - \mu| > \epsilon\} = 2\, P\{(T(x) - \mu) > \epsilon\}$$

$$= 2\left[1 - P\left((T(x) - \mu) < \epsilon\right)\right]$$

$$= 2\left(1 - P\left\{\frac{\sqrt{n}(T(x) - \mu)}{\sigma} < \frac{\epsilon\sqrt{n}}{\sigma}\right\}\right)$$



$n \to \infty$

$P \to 1$

$$= 2(1 - 1)$$

$$= 0$$

Mean squared error:

$$MSE = \frac{1}{n} \sum \left(T(X_i) - E(T(X_i))\right)^2$$

($T(X_i)$: actual value; $E(T(X_i))$: Expected)

$$= E\left(T(X_i) - E(T(X_i))\right)^2$$

$MSE = Var(T(X))$

Smaller the value of variance of estimator better the estimate

Ex: $X_1, X_2 \ldots X_n$ · random sample

$E(X_i) = \theta \qquad Var(X_i) = \sigma^2$

$T_1(X) = X, \quad T_2(X) = \bar{X}$

$E(X_1) = \theta, \quad E(X_i) = \theta$ $X_1, \bar{X}$ are unbiased

$Var(X_1) = \sigma^2, \quad Var(\bar{X}) = \frac{\sigma^2}{n}$

$\sigma^2 \geq \frac{\sigma^2}{n}$ → less variance

$\bar{X}$ is better estimator.

⟶ $E(s^2) = \sigma^2$ s^2 is unbiased estimator of σ^2

$E\left(\frac{n-1}{n} s^2\right) = \frac{n-1}{n} \sigma^2$

$= \sigma^2 - \frac{\sigma^2}{n}$ → bias.

$\frac{n-1}{n} s^2$ is biased estimator of σ^2

Method of Moments:

$X_1, X_2 \ldots X_n$ from a population

$\theta = (\theta_1, \theta_2, \ldots \theta_k)$

↓
parameter vector

Consider non central moments μ_j'

$E(X_i^j) = \mu_j' \qquad \mu_j' = \frac{1}{n} \Sigma X_i^j$

↓
Fixed probability.
Each sample has equal chance of choosing.

$\mu_1' = g_1(\theta_1, \theta_2 \ldots \theta_k)$

$\mu_2' = g_2(\theta_1, \theta_2 \ldots \theta_k)$

$\vdots$

$\mu_k' = g_k(\theta_1, \theta_2 \ldots \theta_k)$

Number of unknown parameters

Solve for $\theta_1, \theta_2 \ldots \theta_k$

$\theta_1 = h_1(\mu_1', \mu_2' \ldots \mu_k') \rightarrow$ estimators

$\theta_2 = h_2(\mu_1', \mu_2' \ldots \mu_k')$

$\vdots$

$\theta_k = h_k(\mu_1', \mu_2' \ldots \mu_k')$

h_i is estimator for θ_i

22/6/21

Ex:

① $X_1, X_2, \ldots X_n \sim \text{Poison}(\lambda)$ $\lambda > 0$

λ ↓ unknown parameter

$\mu_1' = \frac{1}{n} \sum X_i$

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

$\mu_1' = \lambda$

Estimator of λ is $\mu_1' = \frac{1}{n} \sum X_i = \bar{X}$

② $X_1, X_2 \ldots X_n \sim N(\mu, \sigma^2)$

μ, σ^2 ↓ 2 unknown parameter

$\mu_1' = \frac{1}{n} \sum X_i = \frac{1}{n} \cdot n\mu = \mu$

$\mu_2' = \sigma^2 + \mu^2$

$\text{Var}(X) = E(X^2) - (E(X))^2$

$\sigma^2 = E(X^2) - \mu^2$

$\mu_2' = \sigma^2 + \mu^2$

$$\mu_1' = \mu$$

$$\mu_2' = \sigma^2 + \mu^2 \Rightarrow \mu_2' = \sigma^2 + (\mu_1')^2$$

$$\Rightarrow \sigma^2 = \mu_2' - (\mu_1')^2$$

So μ_1' is estimator for μ.

$$\mu_2' - (\mu_1')^2 = \frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2$$

$$= \frac{1}{n}\sum X_i^2 - (\bar{X})^2 \text{ is estimator for } \sigma^2$$

→ Method of moments guarantees consistency but not unbiasedness.

Method of Likelyhood Estimation: (discovered by R.A Fisher ↓ F-dist)

→ Form of distribution.

$X_1, X_2 \ldots X_n$ be the observed sample.

Joint probability pmf/pdf

$f(x_1, x_2 \ldots x_n, \theta)$ is called likelyhood function

$L(\theta, x)$ ↓ unknown parameter $\quad x = (x_1, x_2 \ldots x_n)$

Independent

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

$$L(\theta, x) = \prod_{i=1}^{n} f(x_i, \theta) \text{ is likelyhood function.}$$

Most pdf's and pmf's are exponential in nature So we take log

$$l(\theta, x) = \text{Log } L(\theta, x) \rightarrow \text{linear function}$$

↓

log likelyhood function

→ A statistic $\theta(x)$ or $T(x)$ is said to be more/max likelyhood estimator of θ if

$$L(\theta'(x), x) \geq L(\theta, x) \quad \text{for } \forall \theta$$

Ex: Let $X_1, X_2, \ldots X_n \sim Ber(1, P)$

$$L(P, x) = \prod_{i=1}^{n} P^{x_i} (1-P)^{n-x_i}$$

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

$$l(P, x) = \text{Log } L = \sum x_i \log P + (n - \sum x_i) \log (1-P)$$

$$\frac{\partial l}{\partial P} = \frac{\sum x_i}{P} + (n - \sum x_i) \frac{(-1)}{1-P}$$

$$= \frac{\sum x_i}{P} - \frac{(n - \sum x_i)}{1-P}$$

$$(1-P) \sum x_i - P(n - \sum x_i) = 0$$

$$(1-P) \sum x_i = P(n - \sum x_i)$$

$$\sum x_i - P \sum x_i = Pn - P \sum x_i$$

$$\sum x_i = nP$$

$$P = \frac{\sum x_i}{n} = \bar{x}$$

$P = \bar{x}$ is maximum likelyhood estimator.

Ex

$$L(\lambda, x) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod_{i=1}^{n} (x_i!)}$$

$$\log l = -n\lambda + \sum x_i \log \lambda - \log\left(\prod_{i=1}^{n} (x_i!)\right)$$

$$\frac{\partial l}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda}$$

$$-n + \frac{\sum x_i}{\lambda} = 0$$

$$\lambda = \frac{\sum x_i}{n} = \bar{x}$$

$\lambda = \bar{x}$ is maximum likelyhood estimator.

Confidence Interval: An interval is provided within the parameter is supposed to lie.

Ex $X \sim N(\mu, \sigma^2)$, μ, ~~are~~ is unknown
σ is known.

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

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$P\left(-2.576 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 2.576\right) = 0.99$$

$$P\left(-2.576 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 2.576 \frac{\sigma}{\sqrt{n}}\right) = 0.99$$

$$P\left(-2.576 \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq 2.576 \frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0.99$$

$$P\left(\bar{x} - 2.576\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2.56\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

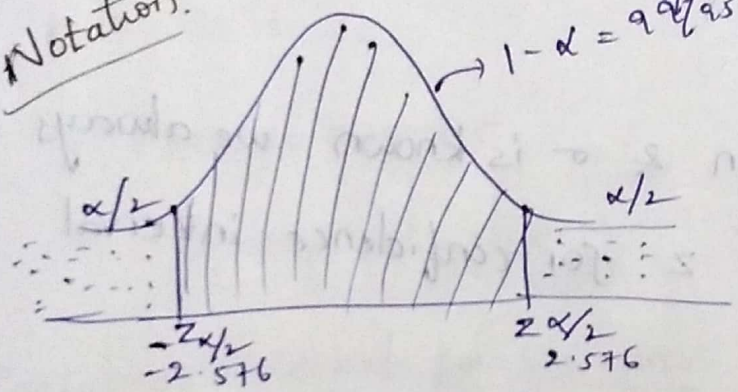
Population mean lies between $\bar{x} - 2.576\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + 2.576\frac{\sigma}{\sqrt{n}}$ with confidence level 99%.

$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$ is confidence interval in which μ is supposed to lie with 99% probability.

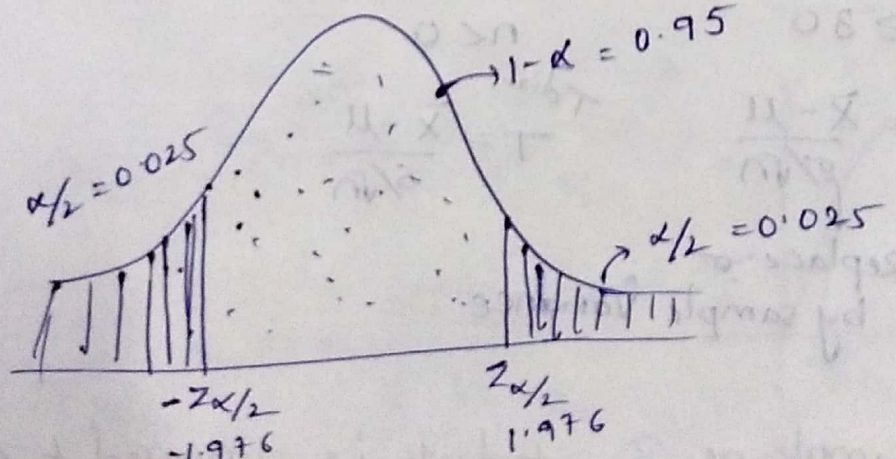
As you increase

23/6/21

Notation:



Confidence interval $\left(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$



→ Bigger the confidence level wider the confidence interval.

→ Confidence level is $1-\alpha$.

Ex: A random sample of 100 employees is taken & mean height 180cm & Variance of population is $49\,cm^2$ find 95% confidence interval for population.

Sol:

$$\mu \in \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x} = 180$

$\sigma^2 = 49$

$$\mu \in \left(180 - 1.976\left(\frac{7}{10}\right), 180 + 1.976\left(\frac{7}{10}\right)\right)$$

→ μ is unknown & σ is known we always use statistic z for confidence interval for any n.

→ μ is unknown & σ is unknown

$n \geq 30$

Normal $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

↓ Replace σ by sample variance

$n < 30$

T dist $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Ex: A sample of 80 students is surveyed & Avg amount spent by them is 593.84 & S.D is 369.34. For 95% confidence level find confidence interval for mean.

$$\mu \in \left(593.84 \pm 1.976 \cdot \left(\frac{369.34}{\sqrt{80}}\right)\right)$$

Ex: A sample size 15 is taken from a population where $\bar{X} = 12$ & $s^2 = 25$ what is 95% confindence interval for μ?

Sol: $$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$$

Interval $\left(\bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$

$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$

$n - 1 = 15 - 1 = 14$

$t_{0.025,14} = 2.145$

$$\mu \in \left(12 \pm 2.145 \frac{5}{\sqrt{15}}\right)$$

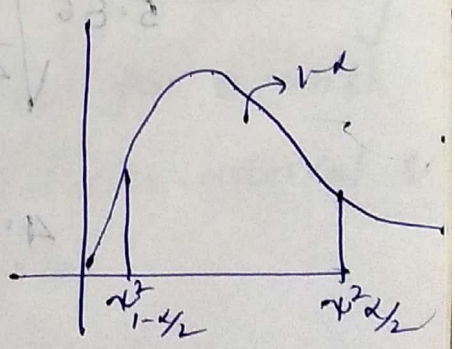
Confidence interval for Variance:

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

$$\frac{1}{\chi^2_{\alpha/2}} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{\chi^2_{1-\alpha/2}}$$

$$\sqrt{\frac{s^2(n-1)}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}}$$



Ex: A statistician chooses 27 dates when examining occupancy records of particular hotel finds s.D on 5.86 rooms rented. Find 95% confidence interval for population s.D

sol

$$n = 27$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

$$n - 1 = 26$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$\sqrt{\frac{s^2(n-1)}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}}$$

$$\sqrt{\frac{(5.86)^2(26)}{\chi^2_{0.025,26}}} < \sigma < \sqrt{\frac{(5.86)^2(26)}{\chi^2_{0.975,26}}}$$

$$\chi^2_{0.025,26} = 41.923$$

$$\chi^2_{0.975,26} = 13.844$$

$$5.86\sqrt{\frac{26}{41.923}} < \sigma < \sqrt{\frac{(5.86)^2 26}{13.844}}$$

$$4.614 < \sigma < 8.0307$$

Test of Hypothesis: There will be two claims about values of parameter & determine which is correct.

→ Decision making problem

$\mu = 50$ ✓
$\mu \neq 50$ ✗ mutually exclusive

Statistical Hypothesis:

... Statement about the population

Null Hypothesis: (H_0)

prediction is not true

Alternative Hypothesis: (H_A or H_1)

Prediction to te true

25/6/21

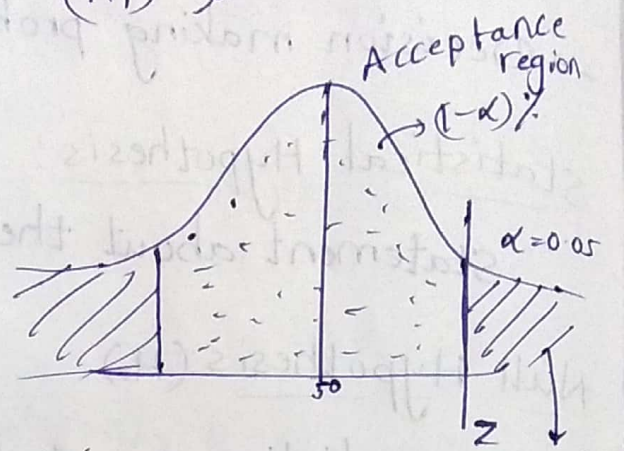
Confidence Interval: Let $X_1, X_2, \ldots X_n$ be the sample observation θ be the unknown parameter of population if there exist statistic L_n & U_n such that $P(L_n < \theta < U_n) = 1-\alpha$ then we say (L_n, U_n) as confidence interval for θ with ~~$\alpha(100-\alpha)\%$~~ $100(1-\alpha)\%$ confidence interval & $(1-\alpha)$ is called confidence level.

Testing of Hypothesis

Ex:

Vaccine 1 for covid 19 (H_0) $\mu = 50$

Vaccine 2 (H_1) $\mu > 50$



Condition of Null Hypothesis (H_0)

		True	False
Possible Action	Accepting	Correct action	Type II error $P(II) = \beta$
	Reject H_0	Type I error $P(I) = \beta$	Correct action

Type 1 error → rejecting Null hypothesis when it is true $\alpha = p$

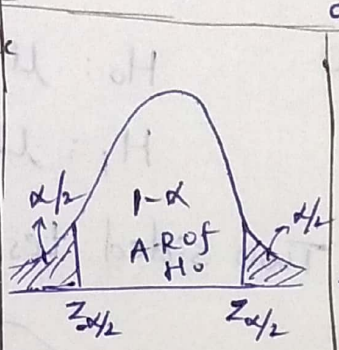
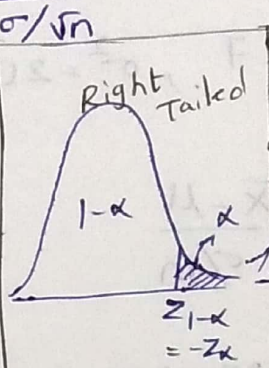
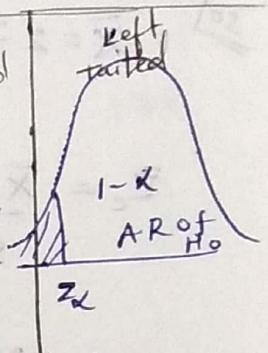
Type 2 error → Failed to reject Null Hypothesis.

Testing hypothesis for the mean μ:

When the value of sample size(n)

- Population is normal or not normal $(n \geq 30)$
  - σ is known: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - σ is not known: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- population is normal $(n < 30)$
  - σ is known: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - σ is not known: $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Test Procedures:

Hypotheses	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$
Test statistic	Calculate the value of: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$		
Rejection region & Acceptance Region R.R & A.R of H_0			
Critical value(s)	$Z_{\alpha/2}$ & $Z_{-\alpha/2}$	$Z_{1-\alpha} = -Z_\alpha$	Z_α
Decision	We jer reject H_0 (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two sided test	$Z > Z_{1-\alpha} = -Z_\alpha$ One sided test	$Z < Z_\alpha$ One sided test

α = P(type I error)

β = P(type II error) = P(failing to reject H_0 when it is false)

Size of the test = α

Power of test = $1 - \beta$ = P(reject H_0 when it is false)

Ex: Average mean age of population of random sample of 10 individuals from the population has mean age 27 & Variance of population is 20. Can we conclude mean is different from 30

Sol

$\bar{x} = 27$, $\sigma^2 = 20$

$H_0: \mu = 30$

$H_1: \mu \neq 30$

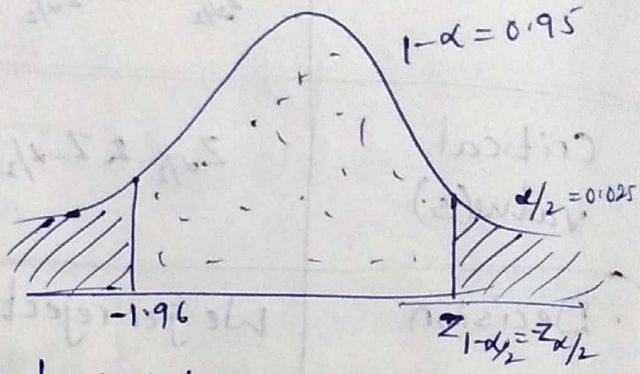
$$Z_c = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Two sided test

$$= \frac{27 - 30}{\sqrt{20}/\sqrt{10}}$$

Calculated statistic

$Z_c = -2.121$



$Z_{0.025} = 1.96$

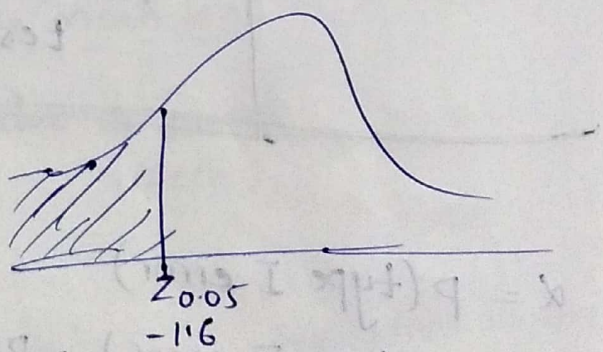
Two Test: $Z_c \geq |z|$ Accept H_0

Since $Z_c < Z$ (−2.12 < 1.96) → Reject H_0

Ex: $H_0: \mu = 30$

$H_1: \mu < 30$

$Z_c = -2.12$



Reject H_0

if $Z_c < Z_{0.05}$ reject H_0

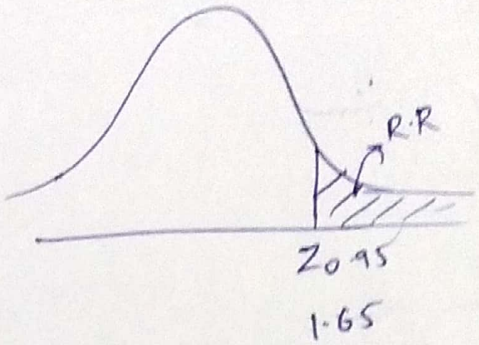
Ex $n = 157$ Avg B.P $\bar{x} = 146$

$S = 27$

$\alpha = 0.01$

$H_0 : \mu = 140$

$H_1 : \mu > 140$



$$Z_c = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{146 - 140}{27/\sqrt{157}}$$

$$= \frac{2}{\cancel{27}\,9}\sqrt{157}$$

$Z_c = 2.78$

$Z_c > Z_{1-\alpha}$ Reject H_0.

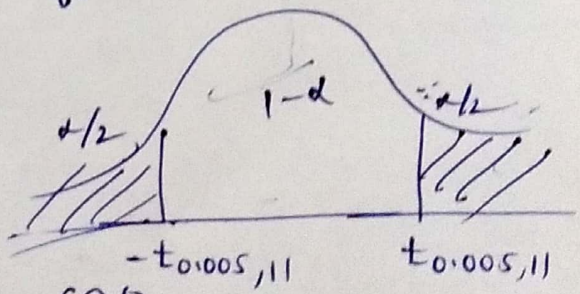
Ex T test $n = 12$, $\bar{x} = 672.6$

$\alpha = 0.01$ $S = 43.72$

New process yield is different from 690.

$H_0 : \mu = 690$

$H_1 : \mu \neq 690$



$$T_c = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{672.6 - 690}{43.72/\sqrt{11}} = -1.319$$

$T_{0.005,11} = 3.106$

Accept H_0.