BIOSTATISTICS

- REVIEW : - SAMPLING DISTRIBUTIONS
  - ESTIMATION - CONFIDENCE INTERVAL
  - HYPOTHESIS TESTING

SAMPLING DISTRIBUTIONS

DISTRIBUTION of the SAMPLE MEAN

Def let $X_1, \dots, X_N$ be random samples from a distribution with mean μ and variance σ^2. The sample mean is the r.v.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Fact $E[\bar{X}] = \mu$, $\quad var[\bar{X}] = \frac{\sigma^2}{n}$ $\quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ is the STANDARD ERROR

If $X_1, \dots, X_N$ are sampled from NORMAL POPULATION, then $\bar{X}$ is normal. $\bar{X} \sim N(\mu, \frac{\sigma^2}{N})$

If $X_1, \dots, X_n$ are sampled from non-normal population, then $\bar{X}$ is approx. normal in accord with CENTRAL LIMIT THEOREM (CLT)

Related result : SAMPLING DISTRIBUTION of $\bar{X}_1 - \bar{X}_2$
SAMPLING DISTRIBUTION of PROPORTION $\hat{P}$

SUMMARY TABLE p. 140 in textbook

ESTIMATION

Def An estimator T of a statistical parameter θ is an UNBIASED ESTIMATOR of θ if $E(T) = \theta$.

Example : $\bar{X}$ is an unbiased estimator of μ since $E(\bar{X}) = \mu$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$, sample variance, is an unbiased estimator of the variance $E[S^2] = \sigma^2$.

CONFIDENCE INTERVAL of a population mean

→ it relies on properties of sampling distribution

Example 6.2.1 Want to estimate avg level of some enzyme in a human population. Sample size $N=10$. Find $\bar{x} = 22$.

Suppose you know variable is normal with known $\sigma^2 = 45$

95% C.I. is $\bar{X} \pm z_{\frac{\alpha}{2}} \sigma_{\bar{X}} = 22 \pm 1.960 \sqrt{\frac{45}{10}}$

Interpretation : The probability that the true mean is contained in the C.I. is 95%

In general, when sampling from a normal population with known variance σ^2, we are $100(1-\alpha)$ percent confident that the interval $\bar{x} \pm Z_{(1-\frac{\alpha}{2})} \sigma_{\bar{x}}$ contains the population mean μ.

$90\% \to Z_{0.05} = 1.645$; $95\% \to Z_{0.025} = 1.960$; $99\% \to Z_{0.005} = 2.58$

- If population is non-normal, $N > 30$, σ^2 known, we can still use same formula for C.I.
- If population variance is unknown, we can estimate S^2. In ~~this~~ case the population is normal and σ^2 is unknown we need to use the t-distribution

Example 6.3.1: We study effectiveness of weight-bearing and ankle mobilization therapy. Isometric gastrocsoleus muscle strength is measured to verify recovery. We sample $N = 19$ patients and find $\bar{x} = 250.8$ N with $S = 130.9$. Assume population is NORMAL

Find 95% C.I. of population mean

$$\bar{x} \pm t_{(1-\frac{\alpha}{2};n-1)} \frac{S}{\sqrt{n}} = 250.8 \pm (2.109)(30.031)$$
$$= (187.7, 313.9)$$

Related results: Confidence interval for the difference between two population means.
Confidence interval for a population proportion
Confidence interval for the difference between two population prop.
Determining the sample size for estimating means and proportions

SUMMARY TABLE p. 182-183 in textbook

Example 6.4.1 Compare serum uric acid levels in patients with and without Down syndrome.

$N_1 = 12$ $\bar{x}_1 = 4.5$ mg/100 ml $\sigma_1^2 = 1$ Assume normal pop.

$N_2 = 15$ $\bar{x}_2 = 3.4$ $\sigma_2^2 = 1.5$

C.I. $$\bar{x}_1 - \bar{x}_2 \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

95% C.I. $1.1 \pm (1.96)(0.4282) = (0.26, 1.94)$

We are 95% confident that the true difference between μ_1 and μ_2 is in this interval. Since both values of interval are > 0, then we are 95% confident $\mu_1 > \mu_2$.

Example 6.7.1 Want to estimate avg daily protein intake in a population of adolescents.
The nutritionist would like the C.I. to be 10 g wide, that is the estimate should be within $h = 5$ g of the pop. mean.
Set $\alpha = 0.05$, that is 95% C.I. I assume that $\sigma = 20$.

Recall C.I. $\bar{x} \pm z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$

Here want $h = z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} \iff 5 = 1.960 \cdot \frac{20}{\sqrt{n}}$

$$\iff n = \frac{(1.960)^2 \cdot 20^2}{5^2} = 61.47$$

Need $n = 62$ or more samples.

Example 6.8.1 Want to estimate what proportion of families in a certain area are medically indigent. Want a 95% C.I. and an estimate within 0.05 from exact proportion. Find n

Recall C.I. $\hat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}}$

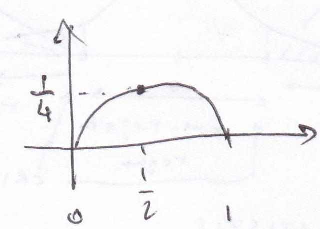
$$0.05 = 1.960\sqrt{\frac{p(1-p)}{n}} \le 1.960\sqrt{\frac{1}{4n}}$$ *

($p = 0.5$ is the worst case scenario)

$$n \ge \frac{(1.960)^2}{(0.05)^2}\frac{1}{4} = 384.16$$

* NOTE

$f(x) = x(1-x), \quad x \in [0,1]$



$|f(x)| \le \frac{1}{4}$ on $[0,1]$

By choosing $\frac{1}{4}$, we are sure we can handle any possible value of proportion p.

HYPOTHESIS TESTING

This method does not prove an hypothesis but simply tests if an hypothesis is supported by the data.

When we fail to reject a null hypothesis, we do not conclude that the hypothesis is true but that IT MAY BE TRUE.

Example 7.2.1 (Single pop. mean)

Want to investigate the mean age of a certain population.

Question: is mean age (of population) $\neq$ 30 y? {

DATA: $N = 10$ samples $\bar{x} = 27$

ASSUMPTION: Assume population is normal and $\sigma^2 = 20$

HYPOTHESIS Null hyp. $H_0: \mu_0 = 30$

Alt. hyp. $H_1: \mu_0 \neq 30$

TEST STATISTIC Based on our assumption $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$

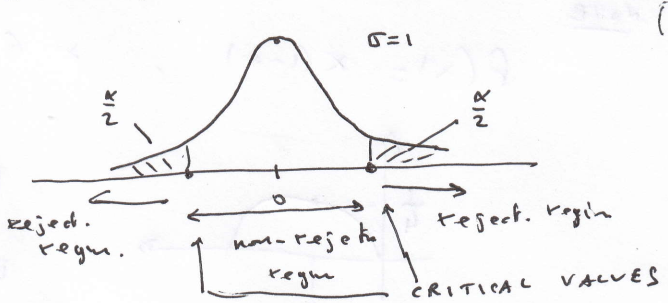
DISTRIBUTION OF TEST STATISTIC Based on sampling distribution and the normal pdf assumption, $z \sim N(\mu = 0, \sigma^2 = 1)$, if H_0 is true

DECISION RULE This rule prescribes to reject H_0 if the test statistic value falls in REJECTION REGION and to fail to reject H_0 if the test statistic value does NOT fall in rejection region.

The size of rejection region depends on the SIGNIFICANCE LEVEL α

= the size of the probability of committing a type I error

type I error = to reject a true null hypothesis



$\alpha = 0.05$

$\alpha/2 = 0.025$

$1 - \alpha = 0.95$

$z_{1-\frac{\alpha}{2}} = 1.960$

CALCULATION OF TEST STATISTIC $z = \dfrac{27 - 30}{\sqrt{20/10}} = -2.12$

STATISTICAL DECISION Since $z < -1.960$, then z is in the REJ. REGION and we can REJECT H_0

The computed value of test statistic is significant at the 0.05 level

CONCLUSION We conclude that $\mu \neq 30$

p-value $P(z \geq 2.12 \text{ or } z \leq -2.12) = 2\Phi(-2.12) = 0.0340$

This is the smallest value of α for which the NULL HYP. can be rejected.

Reporting this value is more informative than simply reporting decision.