chapter 6 - estimation

Inferential statistics
- Estimation.
- Hypothesis Testing.

Estimation
- Point estimation (specific values)
- Interval estimation (range within which the parameter might fall)

population & sample

Reference / Target / Study population

This is where we obtain our samples from.

↳ The Group we want to study.

populations
- finite (can be enumerated) — (Random sampling)
- infinite (can't be enumerated) — (assumed to be random sampling)

not every "large" population is infinite

Sample → Some members of a population.

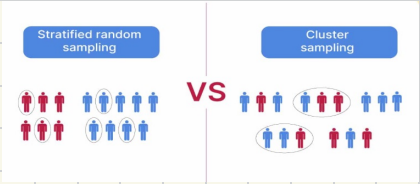
a Random sample :- the sample is random if each member is chosen independently and has a known (non-zero) probability of being chosen

A Simple Random Sample → is a Random sample where each member has an equal chance of being selected.

▶ Types of probability Random samples
- Simple random sample
- Cluster sample (explained in RCT)
- Stratified sample.
- sysmtematic sample.

▷ cluster sampling → when geographic, spatial charcteristics naturally divide the target population into random groups (clusters) like schools, neighborhoods

▷ Stratification :- method of dividing a target population into subgroups (strata) based on characteristics thought to be important (age, gender, clinical condition)



Notation:-

The probability distribution of a Sample Statistic is called the sampling distribution

↳ like : $\bar{x}$, $\hat{p}$, s^2

① estimation of $\bar{x}$

$\hat{\theta}$ → sample statistic

θ → population parameter

If: $E(\hat{\theta}) = \theta$

then $\hat{\theta}$ is a good estimator of θ

$\bar{x}$ is a good estimator of μ because →

▷ $E(\bar{x}) = \mu$ → population mean

▷ $\bar{x}$ (sample mean) is called Minimum Variance unbiased estimator (more accurate)

Notation →

even when using a good estimator like $\bar{x}$; the precision of the estimation can still be affected by →
- Sample size (n)
- σ^2 (underlying variance of individual observations)

▷ Standard Error of Sample Mean →

The variance (σ^2) of the sample mean

$Var(\bar{x}) = \frac{\sigma^2}{n}$ ⇒ the standard deviation of $\bar{x}$ is called the standard error $\left[\frac{\sigma}{\sqrt{n}}\right]$

the standard error can be approximated by:- (when σ is unknown.)

$\frac{s}{\sqrt{n}}$ ← s: sample standard deviation

▷ The Central limit theorem

if

- x is from other distribution
  - $n < 30$ → The CLT can't be applied
  - $n \geq 30$ → $\bar{x} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$
- $x \sim N(\mu, \sigma^2)$ → Then $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ Regardless of n ($n \geq 30$ or $n < 30$)

• Standardization of $\bar{x}$

whether
- $n \geq 30$
- $x \sim N(\mu, \sigma^2)$

→ $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

Z score of $\bar{x}$ can be found using :-

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

Standard normal distribution.

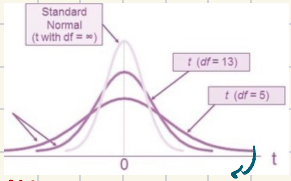
---

t-distribution

- bell shaped (symmetric)
- has mean = 0
- variance = $\frac{df}{df-2}$, df → degrees of freedom, $df = n - 1$ (sample size)
  - greater than 1

- The t distribution has "thicker tails" [higher variance] than the standard normal distribution

But as the sample size (n) increases the variance decreases such that $df \to \infty$ ⇒ The t distribution is the same as the standard normal.



n < then variance ↗

degrees of freedom

* example on Degrees of freedom

Consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degree of freedom is four.

Degrees of freedom = n - 1

In the case of the t-distribution, the degrees of freedom are N-1 as one degree of freedom is reserved for estimating the mean, and N-1 degrees remain for estimating the variability.

1 Degree of freedom for estimating the mean

n-1 for estimating the variability

when to use Z or t ?

if σ

- unknown σ
  - $n \geq 200$ → use Z: $Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
  - $n < 200$ → use t: $t_{(n-1)} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ (degrees of freedom)
- known: always use Z score, regardless of n: $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

notation

when $n \geq 30$

the t distribution variance is nearly (1)

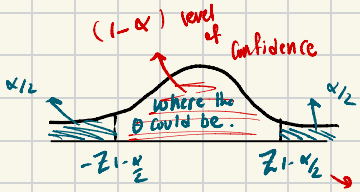
so there won't be a huge difference between the Z of t distribution!

So Don't worries about that please ☺

interval estimation of mean

Confidence Interval

- Level of Confidence $(1-\alpha)$



CI = Point estimator ± [critical value] [standard error]

to construct a $(1-\alpha)100\%$ CI for (μ)

- σ is not known
  - $n \geq 200$: $\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
  - $n < 200$: $\bar{x} \pm t_{(n-1)(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$ (critical value)
- σ is known: $\bar{x} \pm Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ (Critical Value)

gives an Idea of the precision of the Point estimate

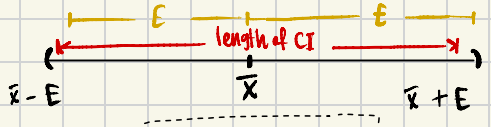
meaning of $(1-\alpha)100\%$ CI for μ

Over the collection of all 95% CIs that could be constructed from repeated random samples of size n, 95% will contain the parameter μ.

$(1-\alpha)$ 100%.

length of CI

▷ length of Confidence interval = upper bound - lower bound



CI = Point estimator ± (critical value)(standard error)

margin of error

length of CI = 2E

example → $L = 2 \times Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ (L: length)

factors that affect the length of CI

1. level of Confidence $(1-\alpha)$ directly
2. α ⇒ Inversly
3. sample size (n) Inversly
4. standard error directly

$\sqrt{\frac{pq}{n}}$ / $\frac{s}{\sqrt{n}}$ → in that case $s \propto$ length of CI directly

Length $\propto (1-\alpha) \propto \frac{1}{\alpha} \propto s \propto \frac{1}{n}$

▶ estimation for the binomial distribution

$X \sim Bin(n, p)$

↳ population proportion

▶ $\hat{p}$ (sample proportion)

$\hat{p} = \frac{X}{n}$ → number of successes

→ sample size

↑ The sample proportion $\hat{p}$ is a good point estimator of p

- $E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n}E(X) = \frac{np}{n} = p$
- $Var(\hat{p}) = Var(\frac{X}{n}) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}(npq) = \boxed{\frac{pq}{n}} \approx \frac{\hat{p}\hat{q}}{n}$ if p is unknown ($q = 1-p$)
- Standard error of $\hat{p} = \frac{pq}{n} \approx \frac{\hat{p}\hat{q}}{n}$ if p is unknown.

from the central limit theorem

$\hat{p} = \bar{x}$ is normally distributed

$\hat{p} \sim N(p, \frac{pq}{n})$, as long as $n\hat{p}\hat{q} \geq 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

confidence interval of binomial proportion

$$\hat{p} \pm \sqrt{\frac{\hat{p}\hat{q}}{n}} z_{1-\alpha/2} \qquad n\hat{p}\hat{q} \geq 5$$

↳ maximum standard error ↑, ↑ length is when $\hat{p} = \hat{q} = \frac{1}{2}$