Statistics and Sampling Distributions

- Statistics are the numerical descriptive measures that you calulate from the random sample of a population (eg. sample mean, sample variance)

- these staticstics can vary or change for each different random sample ⇒ they are random variables

= relative frequency histogram size → 0

- the probability distribution of these statistics, of this random variable, is called sampling distributions

> The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population.

- 3 ways of finding sampling distribution

① mathematically from laws of probability

② Simulation
calc a lot of stats from a lot of samples and plot the relative frequency histogram

③ Use statistical theorem

bellcurve's importance

The central limit theorem ♡ v. imp
+ Sample mean

- it is a statistical theorem that describes the sampling distribution of statistics that are sums or averages

- theorem :
sums and means of random samples of measurements drawn from a population tend to have an approximately normal distribution

- as n increases, the distribution of $\bar{x}$ becomes more and more normal + skew decreases

- The CLT can be restated to apply to the sum of sample measurements Σx_i → as n becomes large has an approximately normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$

- to apply to the mean of sample measurements $\bar{x}$, as $n \to$ large, has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$

WHEN TO USE CLT

- if sampled population is normal, sampling distribution of $\bar{x}$ will be normal $\forall n$

- if sampled population is approximately symmetric, CLT holds for relatively small n

- if sampled polution is skewed, for $n \geq 30$, $\bar{x}$ will be approximately normal

Standard Error of Sample Mean

- standard deviation of a statistic = standard error of the estimator (SE)

- $\because$ standard deviation of $\bar{x} = \frac{\sigma}{\sqrt{n}}$
  = standard error of the mean (SE($\bar{x}$) SEM)

- finding probabilities for the sample mean $\bar{x}$
  assuming $\bar{x} \approx$ normal

① $\mu, \bar{x}, \sigma$ ← calc
$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ ← calc

② calc z value . $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

③ use normal table to compute the probability of $\bar{x}$

- note: population mean should be the peak of the sample mean's normal distribution

mean and std of the normal dist of sample dist give population mean and std.

- greater the n (sample size), less the deviation for mean $\because \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- $np \geq 10$, $n(1-p) \geq 10$ then normal

Properties of sampling distribution of $\underbrace{\text{sample proportion}}$

binomial population : success or failure

$\hat{p} = \frac{\text{no of success}}{\text{no of trials}}$.

's parameter 'p'

prob success in sample

p = probability of success for whole population

$sd = \sigma = \sqrt{pq}$
$\mu = p$
} of random variable $Y = \begin{cases} a & \text{success} \\ b & \text{failure} \end{cases}$

when plotting sampling distribution of $\hat{p}$

μ mean = p
$\sigma/\sqrt{n}$ SEM = $\sqrt{\frac{pq}{n}}$
} because its of $\hat{p}$

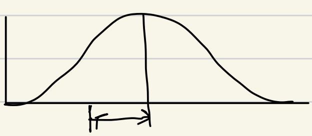
no of samples / observations

mean = $\frac{\text{mean of sample}}{n} = np/n$

sd = $\frac{\text{sd sample}}{n} = \frac{\sqrt{npq}}{n}$

can be approximated for $np > 5$ AND $nq > 5$

calculate probabilities for $\hat{p}$ the same way.



p

$SE = \sqrt{\frac{pq}{n}}$

calcing skew : less than 5 success → right
less than 5 failure → left
at least 5 s + f → normal
neither → uniform

Point Estimation

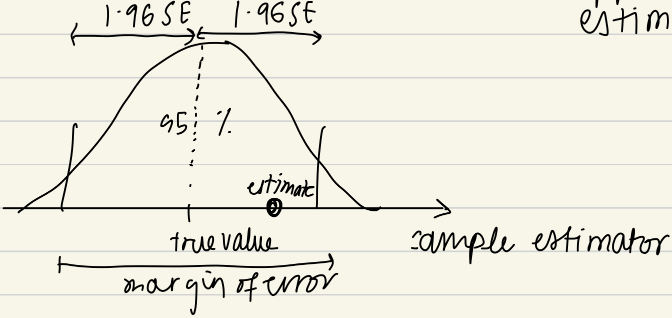
- point estimator (pe) → a statistic used to estimate the value of an unknown parameter of a population
- practically, we can have many pe's ← how to pick best?
- characteristics of pe.

① sampling dist of pe should be centered over true value of parameter ($\mu_{pe} - \mu_p$)
↑
should be unbiased

② variance of pe's sampling dist should be as small as possible

③ error of estimation : difference b/w an estimate & true value of the parameter measured in SE of estimator's sampling dist

margin of error = 95% margin of error = practical upper bound for error of estimation.



sample sd

- est. pop mean, $n \geq 30$, margin of error $= \pm 1.96\left(\frac{s}{\sqrt{n}}\right)$
- est population proportion, $n\hat{p} > 5$ & $n\hat{q} > 5$, margin of error $= \pm 1.96\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

eg.

sample size $n = 50$
sample mean $= 980$ $\bar{x}$
sample dev $= 105$ s
pop mean ?

$$1.96SE = 1.96\left(\frac{s}{\sqrt{n}}\right) = 1.96\left(\frac{105}{\sqrt{50}}\right) = 29.10 \approx 29 \text{ pounds}$$

sample estimate 980 is within ± 29 of population mean

eg.

$n = 100$
$\hat{p} = 0.73$
$\sqrt{\hat{p}\hat{q}} = 0.44$
estimate p

$$1.96\left(\frac{\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}\right) = 0.09.$$

0.73 is with $p \pm 0.09$
$\Rightarrow p = 0.73 \pm 0.9$

interval estimation

- interval estimator : rule for calculating 2 numbers (a, b) which contains the parameter of interest might

- probability that it will contain the parameter = confidence coefficient = $1-\alpha$

- constructing a confidence interval for a sampling distribution of a point estimator

- 95% ⇝ PE ± 1.96 SE

variable center

confidence limits for confidence coeff $(1-\alpha)$

pe ± $Z_{\alpha/2}$ (SE) of estimator

upper and lower confidence limits

random quantities

use s ← sample sd ← if we don't know σ → $\frac{\sigma}{\sqrt{n}}$

good interval is as narrow as possible, w/ large $(1-\alpha)$

but increasing confidence w/o increasing width can only happen by changing sample size n

for population proportion p, $\hat{p} \leftarrow$ sample proportion is the best estimator

$$\hat{p} \pm z_{\alpha/2} \left(\sqrt{\frac{pq}{n}} \right)$$

assuming $np > 5$ & $nq > 5$ & independent, p constant

population mean : $\mu = \sum x_i p_i$
population sd : σ
sample size : n theres $^{N}C_n$ possible samples
population size : N

CLT applies to random variable x_i
$n\mu, \sigma\sqrt{n}$ ← sum of x_i (which is also a random var)
$\mu, \sigma/\sqrt{n}$ ← mean : $\sum x_i p_i$ (—— " ——)

pop normal ⇒ sampling dist normal $\forall n$
pop approx symmetric ⇒ for small n
pop is skewed ⇒ $n \geq 30$.

standard error = standard deviation of a satistic = SEM
SE(mean) = $\frac{\sigma}{\sqrt{n}}$ mean : μ

SE(sum) = $\sigma \cdot \sqrt{n}$ mean : $n\mu$

calculating probabilities : very simple → you know what to do.

standard error for sampling proportion

$\mu = p$, $\sigma = \sqrt{\frac{pq}{n}}$ p = population proportion

$np > 5$, $nq > 5$ ← can apply CLT

less than 5 success → right
less than 5 failure → left
at least 5 s + f → normal
neither → uniform

population parameter estimation :

- mean : $\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$ can be s
- proportion : $\hat{p} \pm z_{\alpha/2} \left(\sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$

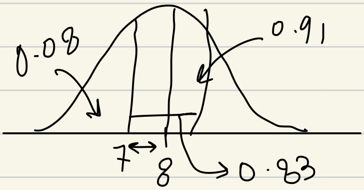
---

other key points

- $E(X^2) = \mu_X^2 + \sigma_X^2$
- $var(\Sigma X_i) = \Sigma(var(X_i))$ $\rightsquigarrow$ $var(\text{sum of } X_i) = n \cdot \sigma^2$
  $\therefore sd = \sqrt{n} \cdot \sigma$

some sums

pg30 ① $\mu = 8$ pop mean
$\sigma = 4$ pop sd
$n = 30$ sample size
enough to approx
∵ it is skewed.
records avg duration.
mid = 8
SEM = $4/\sqrt{30} < 1$.



KA lesson 2 1. $\mu = 210$
2. $\mu = 30$
$\sigma = 1.5$
$n = 3$
mean : $\mu = 30$
$\sigma = 1.5/\sqrt{3}$

3. $\mu = 60$
$\sigma = 0.5$
$n = 36$
$0.5/6 =$

pg42 7.10 $N = 500$
$p = 0.6$ $\sigma = \sqrt{0.6 \times 0.4} = \sqrt{0.24}$
sd : $\mu_{\hat{p}} = 0.6$
$\sigma_{\hat{p}} = \sqrt{\frac{0.24}{500}} = 0.022$