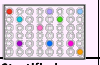
Sampling Methods

Simple Random Sample (SRS)



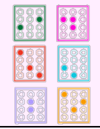
Every sample of certain size has the same chance of being selected (as every other sample of the same size)

e.g. 150 cars in a parking lot. Want to pick 3 of them

Label cars 001-150 since want blocks of 3

- If using a number line table: 456124321150240984312102364219814
  Split the number line up into blocks of 3 and select 3 cars
  456 124 321 150 240 984 312 102 364 219 814
- If using calculator: Randint (1, 150, 3) will select 3 cars

Stratified



Split population into smaller groups called strata (subgroups) based on common characteristics/shared attributes.

e.g. A company has 320 drivers, 80 office workers and 40 mechanics. We want to select a committee of 11 to represent all the employees.

number for each category $= \frac{\text{category total}}{\text{complete total}} \times$ sample size required

Drivers$=\frac{320}{440}\times 11 = 8$, Office $=\frac{80}{440}\times 11 = 2$, Mech$=\frac{40}{440}\times 11$=1

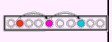
Label: Drivers 001-320, Office 321-400, Mechanics 401-440

Drivers: Randint (1,320,8) will select 8 drivers

Office workers: Randint (321,400,2) will select 2 office workers

Mechanics: Randint (1,320,1) will select 1 mechanic

Systematic



Choose subjects in a systematic, orderly/logical way by sampling every k^{th} element/value

e.g. Select a sample of 15 people from a business survey of 1200 people

Label the people 0001 to 1200

$\frac{1200}{15} = 80$ i.e. we'll take every 80th value

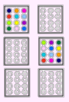
- If using number line: 12436289461234560124
  Split this up into blocks of 4: 1243 6289 4612 3456 0124
  124 is our starting point
  124+80=204, 204+80=284, 284+80=364 etc
  124,204,284,364,444, 524, 604,684, 764,844,924,1004,1084, 1164, 044

If using a calculator: Randint (1,1200,1) gives 24 so start at 24

24+80=104, 104+ 80 = 184, 184+80=264 etc

24,104,184,264,344,424,504,584,664,744,824,904,984,1064,1144

Cluster



Split population into smaller groups called clusters and sample EVERYONE from randomly chosen subgroups. Normally we split up based on area/geographical location. This is similar to stratified sampling, but with stratified sampling the randomness is applied when selecting within each subgroup whereas with cluster sampling the randomness is applied when selecting between each subgroup. When we select the subgroup, we choose EVERY number inside. Therefore not all subgroups will be chosen and therefore won't be represented in the sample unlike with stratified.

Confidence Intervals (CI)

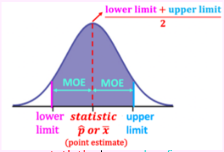
Terminology and Form

$\hat{p} = \frac{x}{n} = \frac{subset}{sample}$

$\bar{x}$ =sample mean

p, μ =population proportion/mean (parameters)

$\hat{p}, \bar{x}$ =sample proportion/mean (statistics/point estimates)



$$\text{CI} = \begin{cases} \text{statistic} \pm \text{margin of error} \\ \text{statistic} \pm (\text{critical value})(\text{standard error}) \end{cases}$$

Note: The standard error is just the S.D. of the sampling distribution

$$= (\text{statistic} - \text{margin of error}, \text{statistic} + \text{margin of error})$$
$$= (\text{lower limit}, \text{upper limit}) = (a, b)$$

Statistic/ Point Estimate

$$= \begin{cases} \frac{x}{n} \text{ if given subset } x \text{ of the sample size } n \\ \frac{a+b}{2} \text{ if given confidence interval } (a, b) \end{cases}$$

Margin Of Error (MOE)

This is the part of the formula after the $\pm$ in the confidence interval

$$= \begin{cases} \frac{b-a}{2} & \text{if given interval (a, b)} \\ z_c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} & \text{(if proportion)} \\ Z_c\frac{s}{\sqrt{n}} & \text{(if mean and using Z)} \\ T_c\frac{s}{\sqrt{n}} & \text{(if mean and using T)} \end{cases}$$

To solve for n: Use either $z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$ or $T_c\frac{s}{\sqrt{n}} / Z_c\frac{s}{\sqrt{n}}$

If don't know $\hat{p}$ then use $\hat{p}(1-\hat{p}) = 0.25$ unless told otherwise

Interpretation: If we repeated this procedure multiple times then statistic would be within ...(MOE) of the true value ... (%) of the time

Standard Error

Proportion: $\sqrt{\frac{\hat{p}\hat{q}}{n}}$, Mean: $\frac{s}{\sqrt{n}}$

Formulae

All these formulae have STATISTIC/ SAMPLE values as their values, not population values

Ask yourself first: Is proportion or mean and then is it 1 or 2 samples

Use Z if σ known OR $n \geq 30$

Use T if σ unknown OR $n < 30$

AP Stats: σ unknown always, so can always use T for mean tests.

We pool if T Test and $\sigma_1 \approx \sigma_2$, but you will also notice that the mark scheme normally doesn't pool regardless of size of σ

- 1 prop: $\hat{p} \pm z_c\sqrt{\frac{\hat{p}(\hat{q})}{n}}$ where $\hat{p} = \frac{x}{n}, \hat{q} = 1 - \hat{p}$
  Tests, 1 Prop Z interval
- 2 prop: $\widehat{p_1} - \widehat{p_2} \pm z_c\sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$ where $\widehat{p_1} = \frac{x_1}{n_1}, \widehat{p_2} = \frac{x_2}{n_2}$
  Tests, 2 Prop Z interval
- 1 mean: $\bar{X} \pm Z_c\frac{\sigma}{\sqrt{n}}$ OR $\bar{X} \pm T_c\frac{s}{\sqrt{n}}$ (σ unknown/ small sample)
  Tests, Z OR T Interval
  z_c : Invnorm with area $\frac{1-\%}{2}, \mu = 0, \sigma = 1$
  T_c: InvT with area $\frac{1-\%}{2}$, df$= n - 1$
- 2 means: $\bar{X_1} - \bar{X_2} \pm Z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ / $\bar{X_1} - \bar{X_2} \pm T_c s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ (pooled formula)
  We pool when $\sigma_1 \approx \sigma_2$: $S_1 < 2S_2$ or $S_2 < 2S_1$
  If don't pool use formula: $\bar{X_1} - \bar{X_2} \pm T_c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  z_c : Invnorm with area $\frac{1-\%}{2}, \mu = 0, \sigma = 1$
  T_c: InvT with area $\frac{1-\%}{2}$, df$=n_1 + n_2 - 2$) (no pool)
  df= min(n_1-1, n_2-1) (no pool)
  Tests, 2 Samp Z or T interval
- 2 means paired: $\bar{d} \pm T_c\frac{s_d}{\sqrt{n}}$
  Tests, Z OR T Interval. We usually use T
  T_c: InvT with area $\frac{1-\%}{2}$, df$= n - 1$
- Slope of regression line (2 possible formulae)
  - slope$\pm T_c(s_b)$ where s_b is standard error (SE)
  - slope$\pm T_c\sqrt{\frac{(1-r^2)s_y^2}{(n-2)s_x^2}}$ where s_x, s_y are the S.D.'s

  T_c: InvT $\left(\frac{1-\%}{2}, df\right)$, $df = n - 2$
  Tests, LinRegTint (can only use if have raw data)

Assumptions

- Random: SRS
- Independent: $N \geq 10n$ or $n \leq .10n$ (pop size is 10 times sample size or sample size is 10 percent of pop size)
- Normality:

Proportion:	Mean (Z)	Mean (T)
$n\hat{p} \geq 1$ $n(1-\hat{p}) \geq 10$ $n \geq 30$ Note: Do for both n's & $\hat{p}$'s if 2 samples	$n \geq 30$ Central Limit Theorem applies	$n < 30$ box plot(should look normal) X list: L_1 (data) Y List: $L_2=\frac{L_1-\bar{x}}{s}$

Interpretation of 2 sample CI:

(+,-) cannot say whether a diff

(−, −): 2nd > 1st

(+, +): 1st > 2nd

- ONE INTERVAL in context/ lying BETWEEN 2 values: We are...% confident that the true population prop/mean of ... lies between ___ and ___
- LEVEL/Capture Rate/ALL intervals: If we repeated this procedure many times, about ...% of the confidence intervals that we compute would contain the true pop value

Increase/ Decrease CI

- Smaller % ⟹smaller CI (since smaller critical value)
- Smaller n ⟹larger CI (larger SE)
- Larger SD ⟹larger CI (larger MOE)

Hints for conceptual understanding

- If see probability not true unless says 0% or 100%
- CI does not talk about sample, it talks about population
- We use the CI for sample to talk about CI for **population,** but we use sample values in the actual confidence interval formulae

Linear Regression

Variables

x =explanatory/independent

y =response/dependent

y is the variable that you're testing to see whether it gets affected by x (y on x or y versus x)

Line of best fit aka least squares regression line (LSRL)

Slope interpretation: for every 1 increase in unit of x, y increases/decreases by b (increases is + slope and decreases if – slope)

y intercept interpretation: when x=0 then y=a i.e. the starting value

To plot line of best fit on TI 84:
Enter data and go Linreg
Store RegEq : y vars, function Y_1 ,calculate
Stat plot, select plot 1, x list: L_1, y list L_2
Press zoom, (9:zoomstat if can't see graph

Way 1: calculator (raw data)
Enter the data
Stat,Calc, LineReg
Freq list: leave blank

Way 2: easier formula not given raw data but given
$y = a + bx$, where $a = \bar{y} - b\bar{x}$ and $b = r\frac{s_y}{s_x}$
Note:
$\bar{x}$ and $\bar{y}$ = means
s_x and s_y = standard deviations

Way 3: harder formula
$$y - \bar{y} = \frac{s_{xy}}{s_{xx}}(x - \bar{x}), \text{ where,}$$
$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = \sum x^2 - n\bar{x}^2$$
$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = \sum y^2 - n\bar{x}\bar{y}$$

Making Predictions

we can plug x values into the line of best fit and solve for y and vice versa

Interpolation

plugging values into line of best fit that ARE within the data range (safe/accurate)

Extrapolation

plugging values into line of best fit that aren't within the data range (not safe/not accurate)

Correlation coefficient (r)

Note:
s_x and s_y = standard deviations

0=no correlation
1=perfect positive correlation
−1=perfect negative correlation

0.00-0.19 very weak
0.20-0.39 weak
0.40-0.59 moderate
0.60-0.79 strong
0.8-1 very strong

Way 1: calculator (if have raw data)
Stat,Calc LineReg(ax+b)
if r is not appearing go to:Catalog, Diagnostic on

Way 2: easier formula
$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$ or $r = \frac{s_x}{s_y}b$ where $y = a + bx$

Way 3: harder formula
$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$
Where
$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = \sum x^2 - n\bar{x}^2$$
$$s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = \sum y^2 - n\bar{y}^2$$
$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = \sum xy - n\bar{x}\bar{y}$$

Determination of variation/proportion of variation (r^2)

This tells us the percent of variation in y that is due to x

Way 1:
Square r value found above

Way 2: use a formula (never use this)
$$r^2 = 1 - \frac{SSE}{SST}$$
SSE=$\sum(y_i - \hat{y})^2 = \sum(\text{residuals})^2$
SST=sum of squares for total (sum of squared deviations from the mean)=$\sum(y_i - \bar{y})^2$

Residual

To draw residual plot on TI 84:
Stat plot, select plot 1, x list: L_1,
Y List: Resid (2nd, stat, Resid)
Make sure you have already stored the regression equation in line of best fit section

actual data value – predicted value from line

This is the vertical distance from the points to the line. The less the gap between points and line, the lower the residual value

Points lie above line means positive residual
Points lie below line means negative residual

The mean of the residuals is always zero

Computer Printout Analysis (you are asked to locate and interpret the values in red on the table to the right)

Response variable is ....

Variable	Coefficient	Std Dev	T	P
Constant	INTERCEPT	...	...	...
Amount.	SLOPE	SE Slope	...	...

S = SD residuals R-Sq = r^2 value

Standard Error of Slope Parameter (SE)

Standard deviation of the estimated slope for predicting y by using an amount if x

Standard Deviation of Residuals

Measures the typical amount of variability in the vertical from the actual data value (observed y variable) to the regression line i.e measure of variation in y variable for a given amount of x variable

To Check Whether A Good Model

Scatter plot looks like a straight line
High r value
High r^2 value
Residual plot has no pattern with uniform variation across x (starry night)

Linear Regression - Transformations

Power

$y = ax^b$ where $x \to log x, y \to \log y$
On calculator: PwrReg(L_1, L_2) or LinReg (log L_1,log L_2)

Exponential

$y = ab^x$ where $x \to x, y \to \log y$
On calculator: ExpReg(L_1, L_2) or ExpReg (L_1,log L_2)

Interpretations

Transformed scatter plot must look linear
Residual plot of TRANSFORMED no pattern starry night

Probability

Probability of event A: $P(A) = \frac{n(A)}{n(U)} = \frac{number\ of\ favourable\ outcomes}{number\ of\ possible\ outcomes}$

Complementary Events: P(A')=1−P(A) i.e. probabilities add to 1

Combined Events (Addition Rule): P(A∪B)=P(A)+P(B)−P(A ∩ B)

Mutually Exclusive Events: $P(A \cap B) = 0$
addition rule becomes: P(A∪B)=P(A)+P(B)

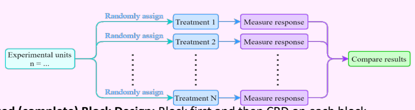
Independent Events: $P(A \cap B) = P(A)P(B)$
addition rule becomes:P(A∪B)=P(A)+P(B)−P(A)P(B)

Conditional "A given B": $P(A|B) = \frac{P(A \cap B)}{P(B)}$
If independent: $P(A|B) = P(A)$

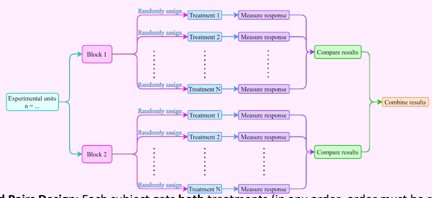
Bayes Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$

Experiment Templates

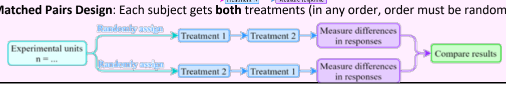
Completely Randomise Design (CRD): Each subject receives **only one** treatment



Randomized (complete) Block Design: Block first and then CRD on each block



Matched Pairs Design: Each subject gets **both** treatments (in any order, order must be random)



Distributions

Binomial Distribution
Binompdf (=), Binomcdf (≤)

$x \sim B(n, p)$
E(X)=Mean=np, Var(X)=$np(1-p)$, $P(X = x)=\binom{n}{x}p^x(1-p)^{n-x}$

Normal Distribution
Normcdf (given x, want prob)
Invnorm (given prob, want x, μ, σ)

$x \sim N(\mu, \sigma^2)$
Standardised variable $z = \frac{x - \mu}{\sigma}$

Geometric Distribution (how long until 1st success)
Geopdf (=)
Geocdf (≤)

$X \sim Geo(p)$
Mean= $\frac{1}{p}$, variance=$\frac{1-p}{p^2}$
P(X= x)= $p(1-p)^{x-1}$ P(X> r)= $(1-p)^x$
If not using calculator: need to turn all into = or >

Expectation Algebra

$E(aX \pm b) = aE(X) \pm b$, $VAR(aX \pm b) = a^2Var(X)$
If X and Y independent:
$E(XY) = E(X)E(Y)$, $Var(aX \pm bY) = a^2Var(X) + b^2Var(Y)$

Expected Value Discrete

E(X)= $\sum xP(X = x)$. Multiply $x's$ by their prob and add

Variance Discrete

Var(X)= $\sum x^2P(X = x) - E(X)^2$
Multiply $x^{2'}s$ by their prob and add and then take away $E(X)^2$

Normal Approximation To Binomial (don't forget continuity correction before finding the prob)

n large, p close to $\frac{1}{2}$ OR $n\hat{p} \geq 10$, n$(1-\hat{p}) \geq 10$
$X \sim B(n, p) \Longrightarrow X \sim N(np, np(1-p))$
Mean is np and s.d. is $\sqrt{np(1-p)}$

Sampling Distributions

SAMPLE Mean Distribution (Central Limit Theorem)
Conditions: $n \geq 30$, $s \approx \sigma$

$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Mean is μ and s.d. is $\sqrt{\frac{\sigma}{n}}$

SAMPLE Proportion Distribution
Condition: $n\hat{p} \geq 10$, n$(1-\hat{p}) \geq 10$, $\hat{p} = p$

$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$. Mean is p and s.d. is $\sqrt{\frac{p(1-p)}{n}}$
use POPULATION value p, not sample value $\hat{p}$ which is $\frac{x}{n}$
(we only use $\hat{p}$ for the conditions, intervals and test statistic)

Hypothesis Testing – Test Statistics

Assumptions: See confidence interval section

Hypothesis Templates

Use population values and ask yourself
- whether 1 or 2 sample
- whether Z or T

1 sample:
H_0: $p = \cdots$ or $\mu > \cdots$
H_1 : $p > \cdots$ or $\mu > \cdots$
$p < \cdots$ or $\mu < \cdots$
$p \neq \cdots$ or $\mu \neq \cdots$

2 sample:
H_0: $p_1 - p_2$= 0 i.e. p_1=p_2 or $\mu_1 - \mu_2$= 0 i.e. μ_1=μ_2
H_1 : $p_1 - p_2$< 0 i.e. p_1<p_2 or $\mu_1 - \mu_2$< 0 i.e μ_1<μ_2
$p_1 - p_2$> 0 i.e. p_1>p_2 or $\mu_1 - \mu_2$> 0 i.e μ_1>μ_2
$p_1 - p_2 \neq$ 0 i.e. $p_1 \neq p_2$ or $\mu_1 - \mu_2$= 0 i.e $\mu_1 \neq \mu_2$

2 sample paired:
H_0: $\mu_d = 0$
H_1/H_a : $\mu_d > 0$
$\mu_d < 0$
$\mu_d \neq 0$

Slope of regression line:
H_0: $\beta = 0$
H_1: $\beta \neq 0$
$\beta < 0$
$\beta > 0$

Test Statistic (TS)

AP Stats: σ unknown always, so can always use T for mean tests.

We pool if T Test and $\sigma_1 \approx \sigma_2$, but you will also notice that the mark scheme normally doesn't pool regardless of size of σ

Either use TS or P value to decide whether to reject:
TS Method:
<: TS< CV reject
>: TS> CV reject
≠: TS< CV or TS>CV reject

P value Method:
If P value < α reject H_0

1 prop TS: Z= $\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$

2 prop TS: Z =$\frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $p_1 - p_2 = 0$

$\widehat{p_1} = \frac{x_1}{n_1}, \widehat{p_2} = \frac{x_2}{n_2}, \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$

1 mean TS: Z = $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$, T=$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ df =n-1

2 mean TS: Z = $\frac{\bar{X_1} - \bar{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, T=$\frac{\bar{X_1} - \bar{X_2} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $\mu_1 - \mu_2 = 0$ and $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$

If don't pool: T= $\frac{\bar{X_1} - \bar{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

df=$n_1 + n_2 - 2$ (if pool)/ min (n_1- 1, n_2-1) (if no pool)

2 means paired TS: Z/T = $\frac{\bar{x_d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$ df =n-1

Slope of regression line TS: T= $\frac{slope}{s_b} = \frac{b - \beta}{s_b} = \frac{b}{s_b}$
where $\beta = 0$, df =n-2

Conclusion

There is sufficient/insufficient evidence at the...% level to reject H_0 & we can conclude...

P value

To calculate: Find TS First
<: P(Z<TS)
>: P(Z>TS)
≠: 2P(Z>TS) or 2P(Z<TS)
Find prob's using normcdf with μ and $\frac{\sigma}{\sqrt{n}}$ or Tcdf with df (see test statistics section above for df for Tcdf)

If p values is than α reject H_0 i.e $P < \alpha$

Interpretation:
Assume H_0 is true, then prob of being in the correct tail is the p value (in lower tail if < test, in upper tail if > test and in double either tail if ≠ test) the p value

Larger sample size ⟹ smaller p value

Chi- Squared

If $x^2_{calc} = \sum\frac{(O-E)^2}{E}$. Reject: $x^2_{calc} > x^2_{critical}$
Hypothesis:H_0: are independent/in the ratio/... distributed
H_1: are not independent/in the ratio/... distributed
df=(row-1)(columns-1) for independence and n−1 for others
df= $n - 2$ if approximating p or μ/σ but this never comes up

Errors

Type 2 Error Steps:
Step 1: Find CV (using invnorm)
<: area = α (left), u=0, 0, σ=1
>: area = α (right), u=0, 0, σ=1
≠ : area $= \frac{\alpha}{2}$ (left), u=0, σ=1
$= \frac{\alpha}{2}$ (right), u=0, σ=1

Step 2: Find error (normcdf)
<: Lower=−100, upper=CV
u=new μ, σ=$\frac{\sigma}{\sqrt{n}}$
>: Lower=CV, upper=100
u=new μ, σ=$\frac{\sigma}{\sqrt{n}}$
≠: Lower=CV1, upper=CV2
u=new μ, σ=$\frac{\sigma}{\sqrt{n}}$

Definitions:
Type 1: H_0 true, but we say it is false i.e. reject it
Type 2: H_0 false, but we say true i.e accept it

Calculations:
Type 1: This is α (prob of rejecting the test i.e. being in the critical region)
Type 2 = β: Prob of NOT being in the critical region-see blue on left for the method of ow to find this

Note: There is an alternative method which is harder. Only use this method if you're forced to.
Re-arrange $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ to get $\bar{x} = \mu + z_c\frac{\sigma}{\sqrt{n}}$

<: P$\left(\bar{x} < \mu + z_c\frac{\sigma}{\sqrt{n}}\right)$ where μ=original mean, z_c=CV
Lower= = −100, upper= $\mu + z_c\frac{\sigma}{\sqrt{n}}$,$u$=new μ, σ=$\frac{\sigma}{\sqrt{n}}$
>: P$\left(\bar{x} > \mu + z_c\frac{\sigma}{\sqrt{n}}\right)$ where μ=original mean, z_c=CV
Lower= $\mu + z_c\frac{\sigma}{\sqrt{n}}$, upper=100, u=new μ, σ=$\frac{\sigma}{\sqrt{n}}$
≠:P$\left(\mu - z_c\frac{\sigma}{\sqrt{n}} < \bar{x} > \mu + z_c\frac{\sigma}{\sqrt{n}}\right)$ where μ=original mean
Lower= $\mu - z_c\frac{\sigma}{\sqrt{n}}$, upper= $\mu + z_c\frac{\sigma}{\sqrt{n}}$, u=new μ, σ=$\frac{\sigma}{\sqrt{n}}$

To Increase/Decrease The Errors:
Power: 1 minus prob of type $2 = 1 - \beta$
Increase Type 1: increase sig level. Changing sample size does nothing.
Increase Type 2: dec sig level i.e. dec type 1 error, take smaller samples, increase α, increase SE

Data

Mean: If no frequency:$\bar{x} = \frac{\sum x}{n}$, If frequency: $\bar{x} = \frac{\sum fx}{\sum f}$

Variance
Note:can also use formula $\frac{s_{xx}}{n}$

If no frequency: $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{\sum(x-\mu)^2}{n}$
If frequency: $\sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{\sum f(x-\bar{x})^2}{\sum f}$

Standard Dev: $\sigma = \sqrt{variance}$

s_{xx}

$$s_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$
$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = \sum y^2 - n\bar{x}\bar{y}$$
$$s_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = \sum xy - n\bar{x}\bar{y}$$

Unbiased Estimator ($\bar{x}, S_x$)

$$\bar{x} = \frac{\sum x}{n}, S_x = \sqrt{\frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{n}{n-1}}\sigma$$

Quartiles
Being in the pth percentile means you did better than p % of people.
e.g. lower quartile (25th percentile)
$P(X < a) = 0.25$
25% were below the value a

Median $= Q_2 = \frac{n^{th}}{2}$ or $\frac{50}{100}n^{th}$ value
Lower Quartile $= Q_1 = \frac{n^{th}}{4}$ or $\frac{25}{100}^{th}$
Upper Quartile $= Q_3 = \frac{3n^{th}}{4}$ or $\frac{75}{100}^{th}$
pth percentile $= \frac{p}{100}^{th}$

Shape
Left skew: Mean<Median<Mode or $Q_3 - Q_2 < Q_2 - Q_1$
Right Skew: Mode<Mean<Median or $Q_3 - Q_2 > Q_2 - Q_1$
Symmetrical: Mode=Median=Mean or $Q_3 - Q_2 = Q_2 - Q_1$

Centre: Mean vs Median. Use median if skew

Spread: Range vs IQR (use if skew) vs Standard Deviation

Outliers: Any values > UQ + 1.5(IQR) or < LQ − 1.5(IQR)