INTRODUCTION

The term **DISTRIBUTION or PROBABILITY DISTRIBUTION** or **FREQUENCY DISTRIBUTION** is a "list" of *all* possible values a r.v. can take, and the corresponding frequency, relative frequency or probability. Distributions can be expressed

- □ Using descriptions: Normal / Uniform / Binomial etc.
- □ Graphically: Stemplot, Boxplot, Histogram, Scatterplot, Segmented Bar Graph
- □ In table format: horizontal [as in a probability distribution table] or vertical [as in a frequency distribution]

Distribution ~

- For a univariate data set ⇨ Center (Mean / Median), Shape (Skewed / Symmetric), Spread (Range, s.d., IQR)
- For a Bivariate data set ⇨ Form, Association, Strength
- For r.v. X ⇨ Probability Distribution Table

X	X_1	X_2	.	.	.	X_n
P(X)	P_1	P_2	.	.	.	P_n

From *this*, determine **C**enter (mean, median), **S**pread (σ) and **S**hape (*compare* position of Mean and Median).

For combinations of r.v. such as: [Sampling] Distribution of

$X_1 \pm X_2$ }
$X \pm Y$ }
$X_1 + X_2 + \ldots X_n$ }
$(X_1 + \ldots + X_n) \pm (Y_1 + \ldots + Y_n)$ }
XB }
$XB_1 - XB_2$ }
$XB_{1\text{-}2}$ }
p' }
$p'_1 - p'_2$ }

Assuming SRS and Independence of outcomes, find E(r.v.) ~ Centre, σ(r.v.)~ Spread, Shape ~ Normal?

SECTION 1: Describing Distributions

1. To calculate, Mean, Median, Q1, Q3, IQR, σ for a data set, use **1-Varstat** command:

1-Varstat L1	L1: x-values; freq = 1
1-Varstat L1, L2	L1: x-values; L2: freq / probabilities

Average ~ Mean ~ Expected Value

2. For Histograms / Scatterplots, use **scale** ~ (Max–Min)/10

3. To describe univariate (1-variable) distributions: *in context,* describe **CSS**:
a) Centre **(C)**: mean, median
b) Shape **(C):** Symmetrical, Right/Left-skewed; Outliers / Gaps / Clusters, if any; Unimodal / Bimodal
c) Spread (**S**): Range, IQR, s.d.

To *compare* distributions,
i) Observe consistent / systematic *patterns* [< or >] between the **Centre** & **Spread** of *both* distributions; and differences *re* **Shape**
ii) Mention *both* populations.

4. **Ogive**: S-shaped curve of X-Values [X-axis] vs. Relative Cumulative Frequencies or Percentiles [Y-axis].

◦ **A.** For right-skewed distributions, the beginning of the Ogive is steep while the "right-tail" is long and flat.

◦ **B.** For left-skewed distribution, the "left-tail" is long and flat while the end is steep.

◦ **C.** For (symmetric) distributions, the S-shape is smooth, the ends are flat while the middle is steep.

5. If the ***p*-th Percentile** is X, then *p%* of ALL observations are ≤ X ⇨ Left "Area" in a Histogram or Continuous Curve where N ~ 100%. Alternately, if X corresponds to the p-th percentile, then p% of ALL values [incomes, weights, heights, length of pregnancies, etc.] are ≤ X.

Given *any* data set or distribution, we can find
i) the percentile corresponding to a given X
ii) the X corresponding to a given percentile
iii) the Z-score for a specified X: $Z = (X - \mu) / \sigma$

6. **IQR**: measure of spread / dispersion / variability / Range of the middle 50% of the data-set, $Q_3 - Q_1$ = 75th – 25th Percentiles

7. **Boxplots**: reveal *degree* of symmetricity ONLY; *not* precise shape or normality of distribution (⇨ use Histogram / Stemplot / Normal Probability plot)!

Symmetric Boxplots / distributions need *not* be Normal.

8. **Histograms**: reveal rough center, shape and spread; *not* Numerical Summaries ⇨ use 1-Varstat L1, L2 for **CSS**

9. For **skewed distributions** do *not* use XB or σ as measures of Centre and Spread, since X is influenced by skewness ⇨ use Median and IQR since they are unaffected by extreme values.
For ~ **symmetric distributions**, use XB and σ.

10. For **left-skewed** distributions, the smaller values of X occur with low frequencies [~ rarely!] while larger values, more frequently: Mode > Median > Mean
For **right-skewed** distributions, the larger values of X occur rarely, while the smaller values occur more often: Mode < Median < Mean
For **symmetric distributions**: Median ≈ Mean

11. **Outlier Rule of Thumb**:
For ~ symmetric distributions ⇨ observations *beyond* $\pm 3\sigma$.
For skewed distributions ⇨ observations *beyond* $Q3 + 1.5\,IQR$ and $Q1 - 1.5\,IQR$

12. **Transformations**: Measures of Centre and Position [Mean, Median, Percentiles] and Measures of Spread [Range, IQR and s.d.] are affected by transformations in a similar fashion.

I *Adding* a constant, ± *k*, to each data value, X: changes measures of Center [Percentiles, Mean] by ± k units; measures of Spread [Range, s.d., IQR] is *not* affected.

II *Multiplying* each data value, X, by a constant *k*: changes measures of Center [Percentiles, Mean] *and* Spread [Range, s.d., IQR] by *k* units.

13. **Standard Deviation**, $\sigma = \sqrt{[\sum(X_i - \mu)^2/n]}$, is a measure of the average variability of each X-value from the mean, μ, obtained by computing the sums of squares of deviations of each X-value from μ, and "adjusting" for the sample size, *n*, and units [by taking √].

14. **Properties of S.D.**:
a) S.d. is strongly influenced by extreme values [since the mean is!].
b) Sum of deviations from mean is zero: $\Sigma(X_i - \mu) = 0$.
c) Farther the observations from the mean, greater the spread and variability, and larger the standard deviation.
d) When $\sigma = 0$, *all* observations are *equal* to μ (since there is no deviation from the mean!) so that Mean = Median!

15. **Population Variance,** $V = \sigma^2$; **Sample Variance,** S^2
Sample s.d., $S = \sqrt{[\sum(X_i - \mu)^2/(n - 1)]}$

SECTION 2: Normal Distributions

1. For any Q involving Normal Distributions:
a. Define variable(s) and its distribution $X \sim N(\mu, \sigma)$.
b. Use Probability Notation to describe the Q. [Get *all* variables to the L.H.S., if required!]
c. Write the *distribution* of the L.H.S. variables.
d. Draw a graph, label and shade the appropriate region.
e. Use a calculator command to solve the problem!
For unknown area: **NormalCdf** (Left Limit, Right Limit, μ, σ)
For unknown X-value: **InvNorm**(Left Area, μ, σ)

The terms Probability, Proportion, Percentage, Relative Frequency are used *interchangeably* and refer to the same idea.

2. **Notation:** to find probability, P (Area / Proportion) when given *a*, use $P(X \geq a)$ or $P(X \leq a)$ or $P(a \leq X \leq b)$.
To find X-Value (*a*) when probability, ***P*** ("left"Area / proportion) is given, use $P(X \leq a) = \boldsymbol{P}$.

3. **Z-score** corresponding to X gives the number of s.d. (σ) X, is from (above/below) the mean, μ: $Z = (X - \mu) / \sigma$.
Note: We can compute Z-scores for *any* value for any distribution or a dataset; the distribution *doesn't* have to be Normal.

Since sample mean, $XB \sim N(\mu, \sigma/\sqrt{n})$ if $N(\mu, \sigma)$ or if $n \rightarrow 30$, then, $(XB - \mu) / (\sigma/\sqrt{n})$ follows the $Z \sim N(0, 1)$ distribution.

Since sample mean, $XB \approx N(\mu, \sigma/\sqrt{n})$ if $n \rightarrow 30$ by CLT, then, $(XB - \mu) / (\sigma/\sqrt{n}) \approx Z \sim N(0, 1)$ distribution.

Likewise, if nP > 5 and *n*Q > 5, the sample proportion, p' ≈ N(P, √PQ/n), so that (p' – P) / (√PQ/n) ≈ Z ~ N(0, 1) distribution.

4. **Properties of Z-Scores**: If Z-scores are calculated for *each* observation in a data set, then, *irrespective* of the "population" distribution [Centre / Shape / Spread]:
a) The sum of *all* Z-scores is always ZERO.
b) The mean of all Z-scores is always ZERO.
c) The s.d. of *all* Z-scores is always ONE.
d) The distribution of the Z-scores has the same shape as the *original* data set. For a Normal distribution, the distribution of Z-scores is ~ N(0, 1).

5. When **comparing** 2 Normal Distributions (income, test scores, etc.) use *Z-scores* [if distribution is symmetric and mean and s.d. are known] OR *Percentiles* [if entire data-set or the precise distribution e.g. Binomial, Normal, etc. is known].

6. For missing parameter problems [missing **μ** or **σ]**:
I find the Z-score corresponding to the given percentile [≤ Area] by using InvNorm(Percentile, 0, 1)
II Use **X = μ ± *Z*σ** to back-solve for **unknown μ** or **σ**.
NOTE: If **μ** and **σ** are missing, then use $\text{distance}_{\text{Z-scores}} \cdot \boldsymbol{\sigma} = \text{distance}_{\text{X-values}}$

7. **Empirical Rule:** in *any* Normal Distribution,
a) ≈68% of all values lie within 1 s.d. of the mean: i.e. μ – 1σ and μ + 1σ
b) ≈95% of all values lie within 2 s.d. of the mean: i.e. μ – 2σ and μ + 2σ
c) ≈99.7% of all values lie within 3 s.d. of the mean: i.e. μ – 3σ and μ + 3σ.
Corollary: ≈68% of Z-scores lie between -1 and +1; ≈95% of Z-scores lie between -2 and +2; ≈99.7% of Z-scores lie between -3 and +3.

8. An event is said to **Rare or Unusual** if the probability of its occurrence is ≤ 5%. Alternately, Rare Events occur at beyond the 5th or 95th percentiles. Rare events are interpreted as Inequalities *even though* theyre framed as equalities, by finding the *smaller* "extreme" area.

For Normal distributions, to determine if an event is Rare:
a) **Check if** P(Outcome) < 5% or it lies *beyond* the 5th or 95th percentiles
b) **Check if** Z-score lies *beyond* 1.645s.d. from the Mean
c) **Check if** the X-value lies *beyond* the Rare-event cut-off for X, by using **X = μ ±** 1.645**σ**

9. To compare 2 distributions,
a) compare Percentiles: comparable scores have same percentiles **OR**
b) compare Z-scores: comparable scores have same Z-scores

10. For Normal distributions,
→ given Z-scores, we can calculate the corresponding percentile using NormalCdf
→ given percentiles, we can calculate the corresponding Z-score using InvNorm

11. **Percentiles** If X corresponds to the p-th percentile, then p% of ALL values [incomes, weights, heights, length of pregnancies, etc.] are ≤ X.

A problem that *asks for* the Percentile seeks a Proportion / Probability / Percentage / *left* Area for a given X-value → use NormalCdf.

A problem that *gives* a Percentile involves finding an X-value for a given Proportion / Probability / Percentage / *left* Area → use InvNorm.

12. Normal distributions are bell-shaped: unimodal and symmetric, and approximations of a continuous function. They're only mathematical approximations of Histograms. Since theyre continuous, P(X = *a*) = 0 [area under the *line* x = *a* is ZERO!] so that, *theoretically*, P(X ≥ *a*) ~ P(X > *a*).

13. To check for Normality, apply the Empirical Rule or perform a Chi-Square Goodness of Fit Test.

14. There is no reason for **μ > σ** *always* or that **μ >>> σ;** this implies that [from the Empirical Rule], that certain X values are *negative* [which is OK!]. For instance. stock market returns are quite "volatile" / risky: s.d. >> Mean [so that one may *lose* money ~ negative return]; differences [in Matched Pair situations e.g. score improvements, weight-loss, etc.] could have a Mean of, say, 10 with a s.d. of 15, or a Mean of -5 and a s.d. of 15.

15. For any Normal distribution, with *only* the Min, *a*, and Max, *b*, of values provided, Mean ≈ (*a* + *b*)/2 and s.d. ≈ Range / 6 i.e. (*b* – *a*)/6 because of the Empirical

Rule [virtually *all* observations lie within 3 s.d. of the Mean].

Use the Empirical Rule to determine *if* a distribution is ~ Normal or skewed, given it's Min and Max values, or given its Mean and s.d.

SECTION 3: Regression, Transformations and Inference

1. Describing appropriateness of Linear Model:
a) Describe Residual Plot: Plot of X-values vs. RESIDUALS; Haphazard Scatter ⇨ Strong Linear Model
b) Interpret R-Sq: __% of the variation in (Y-variable) can be explained by the linear relationship of (Y-variable on X-variable)
c) Describe the distribution / Scatterplot: **F**orm, **A**ssociation (interpret this) and **S**trength IN CONTEXT
d) Interpret Correlation-coefficient, *r*: Value and Sign ⇨ both, the strength of linear relationship i.e. the closeness of scatter about the LSRL, and the Association
e) Perform a Test of Significance to assess linearity.

---

2. The **LSRL** is $y = a + bx$ (Use Option 8: **LinReg**) suggests a *linear relationship* between the Response variable (Y) and Explanatory variable (X), to predict Y for a given X.

Y-intercept, $a = YB - b\cdot XB$; slope, $b = \frac{\Delta Y}{\Delta X} = r\cdot Sy/Sx$

Properties of the LSRL:
a) Since the Sum [and Average] of Residuals is *always* zero, the LSRL minimizes the Sum of the Squares of the Residuals in the *vertical* direction: $\sum R^2 = \sum(y_i - y^\wedge_i)^2$
b) The LSRL *always* passes through the Means (XB, YB).
c) There are 2 LSRLs: the *other* minimizes the Sum of the Squares of the Residuals in the *horizontal* direction

3. **Slope**: when (X-variable) increases by 1 unit, the (Y-variable) rises/falls by an *estimated* (slope value). $b = \Delta Y/\Delta X$, on average.

4. **Y-Intercept**: when the (X-variable) is close to ZERO units, the (Y-variable) is *estimated* to be, on average, the (Y-Intercept value)...which may not make sense if X = 0 lies outside the given domain of X-values for which the LSRL is valid, due to *extrapolation*.

5. **Residual**: is the error in predicting Y for a given X when using the LSRL. It is the underestimate or overestimate of the predicted Y value from the actual Y value.
Positive Residual ⇨ Underestimate;
Negative Residual ⇨ Overestimate

RAP: Residual $_{\text{Y-value}}$ = Actual $_{\text{Y-Value}}$ – Predicted $_{\text{Y-Value}}$

6. A **residual plot** is a *scatterplot* of Residuals ~ Y-variable) vs. (X-values or *predicted* Y-values, y_i). Using a residual plot and RAP, the original X-Y data-set can be constructed "backwards".

If the residual plot suggests no clear pattern, then the LSRL model is appropriate. A fan pattern in the residual plot indicates the LSRL to be flawed for small / large X-values. Examine *size* of the residuals along with *patterns*.

7. **Exponential Model**: Plot of *log* y vs. x is linear; used when y/x have ≈ common ratio for equidistant x-values
$\log y^\wedge = a + bx$
⇨ $y^\wedge = 10^{a+bx} = 10^a\cdot(10^b)^x$
⇨ $y^\wedge = A\cdot B^x$ → *x* is the Exponent.

Power Model: Plot of *log* y vs. *log* x is linear; used when x and/or y are physical quantities
$\log y^\wedge = a + b \log x$
⇨ $y^\wedge = 10^{a + b\log x} = 10^a\cdot(10^{b\log x})$
⇨ $y^\wedge = A\cdot X^b$ → *x* is the Base.

8. If a *transformed* linear model is appropriate [haphazard residual plot for the *transformed* model], the *original* lineal model *would* have been **in**appropriate [curvilinear residual plot for the *original* model]. And **vice versa**.

9. **Statistical Inference for the LSRL**
The *true* LSRL is: $y = \alpha + \beta x$, where α = true y-intercept and β = true slope; *a* and *b* derived from the sample LSRL, $y = a + bx$, are used to estimate α and β.

10. The **assumptions** for performing Tests of Hypothesis and constructing C.I. for the true slope, β, form **l.i.n.e** as a mnemonic!

a) The *underlying* model is **l**inear [if it *isn't* linear, why bother!]. **How to Check:** the Linearity of Scatterplot **or** the haphazardness of residual-plot.

b) For any x-value,
i) **assume** that the y-values are **i**ndependent of each other, in repeated samples. **How to Check:** assume OR data from a random sample or experiment.
ii) the y-values are distributed **n**ormally [with s.d. σ]. **How to Check:** A Boxplot of the residuals is ≈ symmetric **or** a Histogram of the residuals is reasonably bell-shaped **or** a Normal Probability plot [plot #6] of the residuals is reasonably linear.

c) the y-values have **e**qual variability. **How to Check:** Scatter in scatterplot or residual-plot is ≈ of uniform width.

11. **Standard Error of the Residuals** For any x-value, the y-values are distributed normally [with s.d., σ]. Se [or ***s*** in the computer printout] is simply an estimate of σ with (n – 2) *d.f.*
1: The SE(residual), Se, is a measure of the average *variability* of the actual Y-values from the predicted Y-values as given by the LSRL between Y and X.
2: The SE(residual), Se, is a measure of the average variability of the Y-values for a *given* / fixed X-value, in repeated samples, when describing the relationship between Y and X.
3: The SE(residual), Se, is the "typical" prediction error when estimating Y for a given X, when using the LSRL between Y and X.
4: The SE(residual), Se, indicates that, on average, the actual Y-values differ from the predicted Y-values as by the LSRL between Y and X...by Se.

12. SE(b), the **Standard Error of the Slope** (b), is the standard deviation of the sampling distribution of sample slopes, and
1. is a measure of the average variability of the sample slopes from the true slope of the LSRL of Y on X, in repeated samples.
2. suggests that, on average, the sample slopes differ from the true slope of the LSRL between Y and X by...SE(b).

13. **Tests of Inference for the true slope of the true LSRL**: Ho: The true relationship is NOT linear ie. β = 0
H1 : The true relationship is linear ie. β ≠ 0, β > 0, β < 0

For a **C.I. for the true slope**, β: $b \pm t^*_{n-2} \cdot \mathrm{SE}(b)$ [Read off the statistics from ***printout***!] There is NO command for this. Do the simple arithmetic of $b - t_{n-2}^* \cdot \mathrm{SE}(b)$ and $b + t_{n-2}^* \cdot \mathrm{SE}(b)$ by "hand".

For a **Hypothesis Test for β**: Test statistic, $t_{n-2} = b$ / SE(b)
P-value, P = $2 \cdot P(t > t_{n-2})$ or $2 \cdot P(t < t_{n-2})$ using the **tcdf** command [in case of a 2-tailed Test...which is the common situation]

14. Correlation does *not* imply Causation [*except* when a randomized-controlled *experiment* is the source of the data]. Association is accountable due to confounding factors.

15. **Properties of correlation coefficient,** r:

- It has no units
- $r = \sqrt{R^2}$ and has the same sign of the slope of the LSRL (from printout!)
- + r ⇨ + association and -r ⇨ - association
- r measures degree of *linear* association ONLY in the *range* of given values: the underlying relationship may be non-linear (Check Residual Plot!)
- $-1 \le r \le 1$
- It is *not* affected by linear transformations upon X or Y ie. Adding and / or Multiplying *each* X / Y value leaves r unchanged
- $r \approx 0$ [and $b = 0$ ⇨ flat LSRL] for a haphazard scatter-plot **or** U-shaped scatter- plot ⇨ $r \approx 0$ does not imply that there is no relationship
- It is affected by outliers so that r can change from positive to negative association [and vice versa] **or** from no association to positive / negative association [and vice versa]

16. **Influential Point**: value in the extreme x-direction (v. small or v. large) whose removal / inclusion affects at least 2 amongst R^2, slope and y-intercept. An influential point may not necessarily have a large residual or be an outlier!

17. **Extrapolation**: is to *unwisely* make predictions for X beyond the given domain of the explanatory variable using the LSRL (which applies for the range of given X-values only!).

18. **Extreme Values:** influence XB, Sx (or μ, σ), Range, r and R^2; outliers do not affect Percentiles (including Median) and IQR.

19. When X and Y are standardized, $\mu Z_x = \mu Z_y = 0$ and $\sigma Z_x = \sigma Z_y = 1$ so that $b = r$ and $a = 0$. If the original LSRL is $y = a + bx$, then the transformed LSRL for the Z-scores of Y and X is: $Y' = rX$.

20. **Compare** bivariate distributions by comparing *r*, R^2, Residual Plots [scale], Se and P-value for slope for both data-sets.

SECTION 4: Rules of Probability

1. The **probability** of any outcome is the proportion of times it would occur in a large number of repetitions.

2. Sum of the Probability of ALL Outcomes of an event = 1; $0 \leq P(A) \leq 1$; $P(A') = 1 - P(A)$

3. The **OR Rule**: P(A **or** B) = P(A) + P(B) – P(A **and** B) The OR Rule holds for *any* 2 events.

4. **Disjoint or Mutually Exclusive Events**: A and B CANNOT occur simultaneously. For Disjoint Events, P(A or B) = P(A) + P(B) since P(A **and** B) = 0.

5. **Conditional Probability:** If A and B are 2 events, the conditional probability of A, given B is:
$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$. This holds for *any* 2 events.
Thus, $P(A \text{ and } B) = P(B) \cdot P(A \mid B) = \cdot P(A) \cdot P(B \mid A)$

6. **Independent Events**: are those such that their outcomes do not depend on / influence each other.

- $P(A \mid B) = P(A)$ ← Definition
- $P(B \mid A) = P(B)$ ← Definition
- P(A **and** B) = P(A)·P(B)
- P(A **or** B) = P(A) + P(B) – P(A)·P(B)

7. **Disjoint events** *cannot* be independent.

8. P(*at least* 1 occurrence of Event) = 1 – P(Event does *not* occur)

9. Probability Qs [from *this* section] involve:
a) Sampling w Replacement ≈ Binomial Situation (Independent Events)
b) Sampling w/o Replacement
c) Conditional Probability w Venn Diagram / Table: Use *definition* of Conditional Probability to interpret Q

10. Sampling w/o Replacement ≈ Sampling w Replacement when $N \geq 10n$ since the outcomes are more or less Independent then.

SECTION 5: Mean and Variances of r.v.

1. **Probability Distribution**: refers to the Distribution of Probabilities corresponding to each Outcome or Value taken by r.v. X

X	X1	X2	X3	...	X_n
P(X)	P(X1)	P(X2)	P(X3)	...	$P(X_n)$

- List ALL possible Outcomes of the event
- Calculate corresponding Probabilities
- List corresponding values taken by r.v. X
- For Mean and s.d. use 1-Varstat L1, L2 (X: L1, and P: L2)

2. **Mean**, μ or Expected Value, $E(X) = \sum X \cdot P(X)$
S.d., $\sigma = \sqrt{[\sum (X - \mu)^2 \cdot P(X)]}$
Variance, $V(X) = \sigma^2$

3. **Linear Combination of r.v.**:
a. $E(k) = k$
b. $E(kX) = kE(X)$
c. $E(kX \pm Y + a) = kE(X) \pm E(Y) + a$
d. $V(k) = 0$
e. $V(kX) = k^2 V(X)$
f. $V(kX \pm Y + a) = k^2 V(X) + V(Y)$
g. $E(k_1X_1 \pm k_2X_2 \pm \ldots k_nX_n + a) = k_1E(X_1) \pm k_2E(X_2) \pm \ldots k_nE(X_n) + a$
h. For $\sigma(k_1X_1 \pm k_2X_2 \pm \ldots k_nX_n + a)$, *assuming independence*, first find $V(k_1X_1 \pm k_2X_2 \pm \ldots k_nX_n + a) = k_1^2V(X_1) + k_2^2V(X_2) + \ldots + k_n^2V(X_n)$
i. If $X_1, X_2 \ldots X_n$ are independently and Normally Distributed, then $k_1X_1 \pm k_2X_2 \pm \ldots k_nX_n \sim$ Normally Distributed

SECTION 6: Binomial & Geometric Distributions

1. Conditions for a **Binomial Distribution**:

- There are *n* trials
- There are 2 outcomes for each trial, S or F
- P(S) for each trial = P, constant
- The trials are identical and independent.
- We are interested in *number* of successes, *X*, in *n* trials [Note: there are n – *x* failures in *n* trials]

2. For a Binomial Distribution,
a. $P(X = x)$ = **BinomPdf**(n, P, *x*) = ${}^nC_x \, P^x \, Q^{n-x}$

b. $P(X \leq k)$ = **BinomCdf**(n, P, x) = $\sum[x = 0 \textit{ to } x = k]\ {}^{n}C_{x}\ P^{x}\ Q^{n-x}$

Caution: For $P(X \mathbf{<} x)$, $P(X \mathbf{\geq} x)$ and $P(X \mathbf{>} x)$ Qs, since the calculator can perform *only* ≤ operations for BinomCdf...you must "transform" the Q: if uncertain, draw a number line and use common-sense / logic!

c. Expected or Average # of successes, E(X) = nP
d. $\sigma(X) = \sqrt{nPQ}$ and V(X) = nPQ

CAUTION! Only when the Q deals with expressions as $P(X > a)$ or $P(X \geq b)$ do we rewrite it as: $1 - P(X \leq c)$, and for which we use the ∑ Symbol!

For expressions as $P(X < a)$ or $P(X \leq b)$, there is ***no need*** to subtract from 1...because the calculator processes ≤ probabilities!

For Rare-events: interpret the Q as an *inequality* and choose the inequality such that it doesn't cross the Mean: E(X) = nP; E(p') = P

3. **Normal Approximation**: If X ~ Bin(n, P) and
- Samples are drawn randomly
- Outcomes are independent: N > 10*n*
- Samples are Large: nP ≥ 5 and nQ ≥ 5,

then $X \approx N(nP, \sqrt{nPQ})$

4. A Binomial Distribution is
- Right-skewed if the average # of successes, nP, is small relative to the number of trials, *n* **OR** P → 0
- Left-skewed if the expected # of successes, nP is large relative to *n* **OR** P → 1
- Reasonably symmetric if *n* is large and nP > 5 and nQ > 5 **OR** P ≈ 50% so that nP ≈ ½*n.*

5. Conditions for a **Geometric Distribution**:
- There are 2 outcomes for each trial, S or F
- P(S) for each trial = P, constant
- The trials are identical and independent.
- We are interested in the *number of trials for the 1st success* ⇨ Number of trials, n, is *not* fixed!

6. For Geometric Distribution,
- $P(X = x) = (1 - P)^{x-1} P$
- E(X) = 1/p

7. Geometric Distributions are *always* right-skewed.

SECTION 7: SAMPLING DISTRIBUTIONS

1. The term **Sampling Distribution of a Statistic** refers to how frequently the different values a *sample statistic* can assume occurs ie. it is the *frequency distribution* of a statistic. For this, *repeated samples* of size *n* are taken, and the sample statistic [sample mean, median, mode, range, s.d., proportion, variance, 75th percentile, slope, etc.] is computed from each sample; then, a "list" is made of *all* possible values the sample statistic can take, and with what frequency or probability.

Statistic	Value1	Value2	Value3	...	Value_n
P(Statistic)	P(Value1)	P(Value2)	P(Value3)	...	$P(\text{Value}_n)$

2. **Parameter**: refers to values pertaining to the **P**opulation like True Mean, True Proportion, True Standard Deviation, True Median, True Range, etc. that *need* to be estimated. **Statistic / Estimate**: refers to values related to the **S**ample like Sample Mean, Sample Proportion, Sample Standard Deviation, Sample Median, Sample Range, etc. *used* to estimate Parameters.

3. **Sampling Distribution of Means**: Suppose the population [/ "parent"] distribution of r.v. X is such that it has mean μ and s.d. σ. Then, if repeated samples of size *n* are drawn and the sample means computed for *each* sample, then

i) Centre: For *any* sample size, the expected value [/ "average"] of all sample means, XB, is *always* the population mean: E(XB) = μ ⇨ the sample mean is *always* an ***unbiased*** **estimator** of the population mean!

ii) Spread: For *any* sample size, assuming the Xs are independent [or N > 10n], the s.d. of all sample means, XB, is *always* $\sigma/\sqrt{n}$: $\sigma(XB) = \sigma/\sqrt{n}$
- For n > 1, the variability in the sample means, σ(XB) < variability of the population, σ.
- Larger the sample size, smaller the "error" in estimating μ.
- **Specifically,** the s.d. is *inversely* related to the $\sqrt{n}$...so that as sample size rises by a factor of *k,* the s.d. *falls* by a factor of $\sqrt{k}$. Alternately, to reduce the s.d. by a factor of *k,* the sample size should be increased by a factor of k^2.

iii) Shape [CLT]: If *n* → 30, then $XB \approx N(\mu, \sigma/\sqrt{n})$, if the distribution of X is unknown.

4. For $n << 30$, the distribution of sample means is less skewed than the original distribution. But for large n, whatever be the distribution of X and however *skewed* it might be, the distribution of XB approaches normality. In other words, if you take *repeated* samples and calculate XB for each sample, the distribution (or shape) of XB shall be approximately Normal, for $n \rightarrow 30$. To summarize:

- If $X\sim N(\mu, \sigma)$, then for any sample size, n, $XB \sim N(\mu, \sigma/\sqrt{n})$.
- If X is *not* Normal but $n \rightarrow 30$, then by CLT, $XB \approx N(\mu, \sigma/\sqrt{n})$.

5. **Sampling Distribution of Proportions**: If X is a Binomial r.v., $X \sim Bin(n, P)$ and p' is the sample proportion of successes in *n* trials [i.e. $p' = X/n$], then

i) Centre: For *any* sample size, the expected value [/ "average"] of all sample proportions, p', is *always* the population proportion, P: $E(p') = P \Rightarrow$ the sample proportion is *always* an *unbiased* estimator of the population proportion!

ii) Spread: For *any* sample size, assuming the trials are independent [$N > 10n$], the s.d. of all sample proportions, p', is *always* $\sqrt{PQ/n}$: $\sigma (p') = \sqrt{PQ/n}$

- Larger the sample size, smaller the "error" in estimating P.
- **Specifically,** the s.d. is *inversely* related to the $\sqrt{n}$*...*so that as sample size rises by a factor of *k,* the s.d. *falls* by a factor of $\sqrt{k}$. Alternately, to reduce the s.d. by a factor of *k,* the sample size should be increased by a factor of k^2.

iii) Shape: If $nP \geq 5$ and $nQ \geq 5$ then $p' \approx N(P, \sqrt{PQ})/n)$

6. **For Rare-events:** interpret the Q as an *inequality* and choose the inequality such that it doesn't cross the Mean: $E(XB) = \mu$; $E(p') = P$

SECTION 8: Hypotheses Tests and Confidence Intervals for Means & Proportions

1. **Check the validity of assumptions in context** If you're *assuming* something to be true for using the procedure, state it clearly. If some fact is *given* explicitly, state it clearly! CHECK FOR ASSUMPTIONS whether or not the Q explicitly requires it!

If you have any doubts about the validity of a certain assumption, STATE IT: say something like 'we must proceed cautiously with the inference since it may NOT be valid!'

2. The 3 conditions ALWAYS are
i) Randomization / SRS: For surveys, the sample is to be representative of the population, so that one can make inferences or generalizations about the entire population. For experiments, the subjects must be assigned the treatments randomly [on occasion, the subjects are representative of the population also]. Frame the check as: It is *given* that...It is *reasonable* to *assume* that…

ii) Independence of Outcomes: For surveys, the σ(Estimate) calculations require that the outcomes or responses or characteristics are independent of each other: for this, the population needs to be at least 10 times the sample size [$N > 10n$]. For 2 samples, the responses of the individuals or subjects must be independent, *between* and *within* the 2 samples. For carefully-designed randomized controlled comparative experiments, we can assume independence of outcomes.

iii) Normality of sampling distribution OR population: Inference procedures require that the sample was derived from a parent population that is Normal or that the distribution of the Statistic (p' or XB) is ≈ Normal.

3. **For constructing C.I.:**
a) Define the Parameter of Interest in Symbols and Words
b) Check Conditions *in context* pertaining to Random Sample / Randomization, Independence and Normality of the parent population or Statistic
c) Perform Calculations:
(i) State the formula for the C.I. of the PoI *in context*: **Estimate ± M.E.** or **Estimate ± C.V.·S.E.(Estimate)**
(ii) Substitute.
(iii) By calculator, compute the C.I.
d) Interpret the C.I. in context. If necessary, interpret the Confidence Level.
e) If necessary, write a pair of Hypotheses and test the claim about the Parameter using the C.I. Answer the question based on the Confidence Level. Write a decision.

SUMMARY: General Format of a **Confidence Interval**: **Estimate ± M.E.** or **C.V.·S.E.(Estimate)**

Steps: State the Parameter in Symbols and Words ⇨ State the Conditions (Random Sample / Random Assignment; Independence; Normality) ⇨ Perform Calculations ⇨ Write Conclusions (Interpretation of Confidence Interval; Interpretation of Confidence Level; Decision / Final Conclusion in terms of the Parameter)

Confidence Interval formulas:
a) Single Mean: $XB \pm t^*_{n-1}(Sx/\sqrt{n})$
a) Difference of Means: $(XB1 - XB2) \pm t^* \sqrt{[S1^2 / n1 + S2^2 /n2]}$
c) Single Proportion: $p' \pm Z^* \sqrt{(p' \cdot q'/n)}$
d) Mean of Differences: $XB_d \pm t^*_{n-1}(S_d/\sqrt{n})$
e) Difference of Proportions: $(p'1 - p'2) \pm Z^* \sqrt{(p'1 \cdot q'1/n1 + p'2 \cdot q'2/n2)}$

4. **Interpreting the C.I.**: *We are* **x** *% confident that the true (mean / proportion) / lies between* [*and*].

Interpreting the Confidence Level: What an **x** % C.I. means: *if we were to take 100 SRS / perform the study or experiment a 100 times, and computed the* **x***% C.I. for the parameter (true mean/ proportion/difference of means / proportions, etc) from each study, then, in the long run,* **x%** *of the intervals shall contain the parameter.*

5. **Margin of Error**,
M.E. = C.V.·S.E.(Estimate)
= ½ (U.C.L. – L.C.L.) = ½(Width of the C.I)

Interpretation of ME: What M.E. of (say) **3%** for a **x**% Confidence Level means: *if we were to perform the study / experiment 100 times, and computed the* **x***% C.I. for the parameter (true mean/ proportion/difference of means/proportions, etc), then, in the long run, in* **x***% of the intervals,*
a) *the parameter (true mean/ proportion/difference of means/proportions, etc) lies within 3% of the estimate (sample mean/ proportion/difference of means/ proportions, etc).*

OR

b) *the difference between the parameter (true mean/ proportion/difference of means/ proportions, etc) and the estimate (sample mean/ proportion/difference of means/proportions, etc) is at most 3%.*

6. For proportions, M.E. = $Z^* \cdot \sqrt{p' \cdot q'/n}$

The Margin of Error and Width of the Interval depends on
a) the Confidence Level or Z*: Higher the C.L. ⇨ higher the Z* ⇨ larger the M.E., and wider the interval.
b) the sample proportion, p' ⇨ Higher the sample proportion (till 0.5) higher the M.E., and wider the interval.
c) the sample size, *n*: Larger samples ⇨ accurate estimates ⇨ smaller S.E.(estimate) ⇨ smaller M.E. and narrower intervals.

Specifically, the M.E. and S.E. are *inversely* related to the $\sqrt{n}$...so that as sample size rises by a factor of *k,* the M.E. *falls* by a factor of $\sqrt{k}$. Alternately, to reduce the M.E. by a factor of *k,* the sample size should be increased by a factor of k^2.

7. For Means, M.E. = $t^*_{n-1} \cdot Sx/\sqrt{n}$
The Margin of Error depends on
a) Confidence Level or t^*_{n-1} : Higher the C.L. ⇨ higher the t* ⇨ larger the M.E. and wider the interval.
b) Sample s.d., Sx : More variable the data-set ⇨ larger the M.E. and wider the interval.
c) Sample size, *n*: Larger samples ⇨ more accurate estimates ⇨ smaller S.E.(estimate) ⇨ smaller M.E. and narrower intervals. **Specifically,** the M.E. and S.E. are *inversely* related to the $\sqrt{n}$...so that as sample size rises by a factor of *k,* the M.E. *falls* by a factor of $\sqrt{k}$. Alternately, to reduce the M.E. by a factor of *k,* the sample size should be increased by a factor of k^2.

8. **Conditions for surveys**
Random Sample
ONE SAMPLE We are given / need to assume that the individuals were randomly chosen from the population of interest so that the characteristic / responses shall be representative of the population.
FOR MATCHED PAIRS We are given / need to assume that the individuals were randomly chosen from the population of interest so that the ***differences*** in the characteristic / responses shall be representative of the population.
TWO SAMPLES We are given / need to assume that the individuals from BOTH GROUPS were randomly chosen from THEIR RESPECTIVE POPULATIONS so that the characteristic / responses shall be representative of their population.

Independence
ONE SAMPLE We need to assume that the

characteristic / responses / outcomes are independent of each other [$N \geq 10n$].
FOR MATCHED PAIRS We are given / need to assume that the ***differences*** in the characteristic / responses shall be independent of each other [$N \geq 10n$].
TWO SAMPLES We need to assume that the characteristic / responses / outcomes are independent of each other, BETWEEN AND WITHIN THE 2 SAMPLES [$N1 \geq 10n1$, $N2 \geq 10n2$].

Conditions for Experiments
Randomization
FOR MATCHED PAIRS We are given / need to assume that the order of the treatments was randomized [**or** the order in which the subjects received the treatments was random].
TWO SAMPLES We are given / need to assume that the subjects were randomly assigned to the 2 treatments groups.

Independence
FOR MATCHED PAIRS In a well-designed controlled experiment, we can safely assume that the ***differences*** in the characteristic / responses shall be independent of each other.
TWO SAMPLES In a well-designed controlled experiment, we can safely assume that the outcomes / responses are independent of each other BETWEEN AND WITHIN THE 2 SAMPLES.

Normality Condition for Means and Proportions
For **Means**:
I IF data is given, we either make the Boxplot **OR** comment using the boxplot or Stemplot [if given] and depending on the symmetricity [lack of outliers or strong skewness] we say: *We cannot rule out Normality.*
II IF data is not given but *n* > 30, we say *since n > 30, by CLT, XB ≈ Normal.*
III IF data is not given and *n* < 30, we say *we need to assume that the parent population is Normal.*
IV In the case of 2 samples, we make 2 Boxplots **OR** *if* n1 + n2 → 30, since the t-distribution is **robust**, we still proceed.
For **Proportions**:
I for 1 sample: nP, nQ > 5 [Hypotheses Test] and n·p', n·q' > 5 [Confidence Interval]
II for 2 samples: n1·P', n1·Q', n2·P' and n2·Q' are > 5 where Pooled Estimate, P' = (1 + x2)/(n1 + n2) [Hypotheses Test] and n1·p'1, n1·q'1, n2·p'2 and n2·q'2 > 5 [Confidence Interval]

There is <u>no</u> *n* → 30 rule for Proportions!

Note: In case the Normality Condition fails for Means, say: *We shall proceed with Caution.* In case of Outliers, check for Condition upon removal.

In case the Normality Condition fails for Proportions, proceed using a Binomial Distribution.

9. The different Parameters for which C.I. can be calculated are:
Single Mean: μ
Single Proportion: P
Difference of Means: $\mu_1 - \mu_2$
Mean Difference (Matched Pairs): μ_{1-2}
Difference of Proportions (NO Pooling!): $P_1 - P_2$

10. **Missing Sample Size** The *smallest* sample size that would yield a certain Margin of Error for confidence intervals is calculated at a specified confidence level by using:
M.E.(Means) = t_{n-1}*•Sx/√n (use Z* as an approximation of t*)
M.E.(Proportions) = Z*√(p'·q'/n)
When the sample proportion is unknown i.e. there is no preliminary estimate for p, use the “worst-case” or conservative estimate of p' = 0.5.

11. **Missing Confidence Level** ~ middle “area” of a C.I. can be obtained by
a) Normalcdf(L.C.L., U.C.L., Estimate, S.E.(Estimate))
b) Using M.E. = C.V.·S.E.(Estimate) to find the C.V., Z*, and back-solving to find the middle area using Normalcdf(-Z*, +Z*, 0, 1).

12**. For Testing Hypothesis:**
a) State Hypothesis in Symbols; Define the Parameter of Interest in Symbols and Words
b) Check Conditions *in context* pertaining to Random Sample / Randomization, Independence and Normality of the parent population or Statistic
c) Perform Calculations:
Under Ho,
i) Calculate the Test-Statistic, t or Z
ii) Calculate the P-Value while noting if the Hypothesis Test 1 / 2-sided!
iii) State Critical Value, Z or t* (n-1, α) while noting if the Hypothesis Test 1 / 2-sided!
iv) State Decision I based on P-Value.
v) State Decision II based on Critical Value, Z*or t*.
d) Write a Conclusion with the following aspects:

- Interpret the P-Value: If indeed [Ho was true], our P-value of __ % indicates that we'd get a result as extreme as that observed, only __ % of the time.
- Compare the P-Value to the Significance Level, α, and state if the results are *statistically significant* or not.
- Write a Decision in terms of Ho: Reject / not reject Ho...**Tip!** you can *always* cleverly frame this in terms of accepting / rejecting H1!] and answer the Q asked!

SUMMARY: General Format of **Test Statistic = (Estimate – E(Estimate))/ S.D.(Estimate)**

Steps for Tests of Hypotheses: State the Hypotheses; Define the Parameter of Interest ⇨ Check for the Conditions (Random Sample / Random Assignment; Independence; Normality) ⇨ Perform Calculations (Under Ho, find Test-Statistic and P-value) ⇨ Write Conclusions (Interpretation of P-value; Statement of Statistical Significance ⇨ Decision / Final Conclusion in terms of the Parameter]

Test Statistic formulas**:**
a) Single Mean: $t_{n-1} = (XB - \mu)/ (Sx/\sqrt{n})$
b) Difference of Means: $t = (XB1 - XB2) / \sqrt{[S1^2 / n1 + S2^2 /n2]}$
c) Single Proportion: $Z = (p' - P) / \sqrt{(PQ/n)}$
d) Mean of Differences: $t_{n-1} = (XB_d - \mu_d)/ (S_d/\sqrt{n})$
e) Difference of Proportions: $Z = (p'1 - p'2) / \sqrt{P' Q' (1/n1 + 1/n2)}$ where $P' = (x1 + x2)/(n1 + n2)$

13. **Factors affecting Statistical Significance**
Since Test-statistic = [Estimate – E(Estimate)]/ σ(Estimate), an outcome is *likelier* to be significant if

- the observed statistic [~ Estimate] is farther away [i.e. more "extreme"] from the expected value
- the S.E.(Statistic) is smaller [smaller spread and / or larger sample size]

In both cases, this results in a larger Test-statistic value ⇨ smaller P-value ⇨ outcomes are significant.

14. **Making Decisions using P-value**
If P-value < significance level, α, then, under Ho [i.e. if Ho were True], the observed result *is* rare i.e. the observed result *is* statistically significant i.e. we couldn't have got it due to sampling variations ⇨ the hypothesis in Ho *cannot* be true ⇨ **Decision:** Reject Ho in favour of H1

If P-value > significance level, α %, then, under Ho [i.e. if Ho were True] the observed result is not rare i.e. the observed result is not statistically significant i.e. we could have got it due to sampling variations the hypothesis in Ho *could* be true ⇨ **Decision:** Do not Reject Ho: we did not find sufficient evidence in favour of H1

15. For **Matched Pairs**:
a) The sample sizes for BOTH populations is the same; and the complete paired-data set is given OR X_d and S_d is given.
b) For some factor / criteria, the data *was collected* for the 2 groups *together*: it *made sense* to pair the data! BOTH samples are related / alike in some way or dependent on each other.
d) The parameter is μ_d [μ_{1-2} or μ_{2-1}]. Matching implies the underlying context to be Change / Difference / Increase /Decrease / Gain / Loss / Improvement.
e) When data is provided, enter the *differences* into L1 and use 1-Varstat (L1) to find X_d and S_d
f) Check the assumptions for the *differences*: For Observation Studies, the *differences* are random and representative of the population of *differences*, and independent of each other. For Experiments, the order in which the subjects were assigned to the Treatments was random, so that the *differences* may be regarded as random and independent of each other. The boxplot of the *differences* must be reasonably symmetric so that normality cannot be ruled out.

For **Difference of Means**, conservative estimate for *d.f.* ~ min. [n_1 *and* n_2] – 1. **Tip!** For C.I. and Tests of Hypothesis for 2 Means, use the Calculator estimate of the *d.f.*

16. Factors affecting Statistical Significance:
a. (Numerator) The difference between the Parameter and the Statistic i.e. the Observed and Expected values ⇨ Greater the difference, greater the Test-Statistic ⇨ Smaller the P-value ⇨ Reject Ho.
b. (Denominator) Variability of the Sampling Distribution of the Statistic ⇨ Greater the variability, smaller the Test-Statistic ⇨ Smaller the P-value ⇨ Do Not Reject Ho.
c. (Denominator) Sample size(s) ⇨ Larger the samples, larger the Test-Statistic (since the S.E.(Estimate) is smaller) ⇨ Smaller the P-value ⇨ Reject Ho.
d. Significance-level, α ⇨ Larger the level of significance, easier to Reject Ho.

17. **When to Pool** The *only* instance when we Pool estimates is a Test of Hypothesis for Difference in Proportions. We *never* pool for Confidence Intervals [since there is *no* initial premise about "there is no significant difference between..."] and we do not pool for Test of Hypothesis for Difference in Means [since *that* would require the Variances of the populations to be equal, too, something rarely true / unknowable in real life!]. *Re* Difference of Proportions, for a Test of Hypotheses, use the Pooled Estimate, $P' = (x_1 + x_2)/(n_1 + n_2)$ [under Ho: P1 = P2] to

a) check for Normality: n1·P', n1·Q', n2·P' and n2·P' > 5

b) calculate $S.E.(p'_1 - p'_2) = \sqrt{[P'Q'(1/ n_1 + 1/n_2]}$ so that Test Statistic, $Z = (p_1 - p_2) / [\sqrt{P' Q' (1/n1 + 1/n2)}]$

For C.I., do *not* Pool the estimate.

18**. Features of the t-distribution**:

a) *t*-curve is bell-shaped centred at Zero.

b) *t*(n) is shorter and more variable than $Z \sim N(0, 1)$ curve: it approaches the *x*-axis at a higher level than the Z-curve. It has a Mean of 0 but a s.d. > 1.

c) *t*-curve approaches $Z \sim N(0, 1)$ curve as $n \to 30$

d) As *n* increases, the spread decreases: Spread of t(*n*) curve is less than t(*n* – 1) curve

e) The key **Condition** to use the t-distribution is that the *parent* population [of the Xs] be Normal.

f) **Use t-distribution when population s.d., σ, is unknown, and sample s.d., Sx, is used instead [**it is not relevant if *n* is small!]

g) Use t-distributions cautiously in case of Outliers or EXTEREME skew

19. **Missing Parameters in Case of Sampling Distributions**

Given a probability for how extreme an outcome is i.e. P(statistic ≥ Value) = *p* or P(statistic ≤ Value) = *p*, we can determine missing parameters such as sample size, *n*; Sample s.d., Sx; and Estimate, by using the Z- or t-score: Estimate – E(Estimate) / S.E.(Estimate) and back-solving!

20. **Calculator Commands & Formulas for Hypothesis Tests and Confidence Intervals**

i) For single sample Mean: $S.E.(X) = Sx/\sqrt{n}$
Command: **T-Test** or **TInterval**

ii) For two (dependent) sample Means [Matched Pairs]: $S.E.(XB_d) = S_d/\sqrt{n}$
Command: **T-Test** or **TInterval**

iii) For Differences of two (Independent) sample Means: $S.E.(XB1 - XB2) = \sqrt{(S_1^2/n_1 + S_2^2/n_2)}$
Command: **2-SampTTest** or **2-SampTInterval**

iv) For single sample Proportion: $S.E.(p') = \sqrt{(PQ/n)}$ or $\sqrt{(p' \cdot q'/n)}$
Command: **1-PropZTest** or **1-PropZInterval**

v) For Differences of two sample Proportions:
For Hypothesis Tests: $S.E.(p1 - p2) = \sqrt{[P'Q'(1/ n_1 + 1/n_2]}$ where $P' = (x_1 + x_2)/(n_1 + n_2)$;
For C.I.: $S.E.(p'1 - p'2) = \sqrt{[p_1 \cdot q_1/ n_1 + p_2 \cdot q_2/n_2]}$
Command: **2-PropZTest** or **2-PropZInterval**

Warning: For Mean situations, if you *have* the data [list of values], choose Data; if you have the X and Sx values *given,* choose Stats. For proportions, if p is given instead of *x*, calculate the Number of Successes, *x* (using p = *x*/n) and **round it** suitably!

21. **Testing Hypothesis using a C.I. involving Differences** Using a C.I., to test the appropriate Hypothesis for Difference in Proportions, Difference in Means and Mean Difference:

I if both limits of the C.I. are > 0, then (obviously!) *parameter1 > parameter2* ⇨ Reject Ho!

II if both limits of the C.I. are < 0, then *parameter 1 < parameter2* ⇨ Reject Ho!

III if both limits of the C.I. have opposite signs, then *we don't have sufficient evidence to conclusively favour H1* ⇨ *Cannot* Reject Ho!

Only if *both* signs of the confidence limits are positive [or negative], do we have conclusive evidence that, at the *stated* confidence level, $\mu_1 > \mu_2$ [always!] or $\mu_1 < \mu_2$ [always!]. Likewise, $P_1 > P_2$ or $P_1 < P_2$.

22. **Testing a Hypotheses *via* a computer printout** Students must be able to interpret a computer printout to

I) identify the Null and Alternate Hypotheses,

II) recognize the Hypothesis Test or C.I. procedure performed

III) ignore irrelevant data and identify the relevant statistics

IV)observe the Test-Statistic and identify the P-value or C.I.

V) make a decision

23. **CAUTIONS re Statistical Inference for Means & Proportions!**

a. Identify the Q as dealing with Proportions or Means. The notation for both situations is *different*. **Do not mix-up** μ and XB; P and p'; μ and P; X and p'.

b. Identify if the Q deals with one sample or two. Do *not* confuse the population mean / proportion for the sample mean / proportion, and imagine the former to constitute *sample statistics*!

c. In the absence of *clear* indication whether a 1-sided or 2-sided Hypothesis Test is applicable [since, at times, the AP question simply states: *Test a suitable Hypothesis*...] use your judgment. You can *always* go for a 2-tailed test [!] OR go for a 1-tail test in the *direction* of the estimate, while *not* crossing the mean of the sampling distribution. [**Caution:** Misinterpreting the problem and choosing the *wrong* inequality may yield a P-value → 100%...when we're *supposed* to dealing with rare-events with P-values ≈ 5-10%!]

d. For 2-tailed tests, remember to multiply the probability, $P(Z > \#)$ or $P(Z < \#)$ or $P(t < \#)$ or $P(t > \#)$ by 2, to find the P-value!

e. On the Z / t-distribution, label *both*, the Test statistic (Z or t) *and* the Critical Value (Z^* or t^*) so making the decision [reject / not reject Ho] is apparent; ***if*** a mistake were made in calculating the P-value, it'd be readily revealed!

f. If the Normality Condition for a Test of Hypotheses for a Single Proportion fails, then work on X, the Number of Successes, since the underlying distribution is Binomial!

g. It is critical to mention the Confidence Level / Significance Level (α) while writing the Conclusions.

SECTION 9: Chi-Square Tests, Type I & II Errors

1a. Chi-Square Distributions are used to compare 2 or more proportions.

b. The $X^2(n)$ curve is strongly right-skewed and right-tailed but approaches normality for $n \geq 30$.

X^{2-}Tests are *always* 2-sided.

c. The Conditions for a Chi-Square Test are that the samples are SRS / representative of their respective populations **or** the subjects of an experiment are randomly assigned to Treatments, the outcomes are independent of each other (between and within categories), the data are counts [even if given as %, the frequencies can be *calculated*], and that Expected Frequencies are > 5.

d. X^2-value = $\Sigma(O - E)^2/E$ with Observed Values, O: L1 and Expected Values, E: L2.
Degrees of freedom: $(r - 1)\cdot(c - 1)$

e. By examining $(O - E)^2/E$ values, we can determine *which* "cell" or combination of factors contributes *most* to the X^2 value, making the outcomes *more* statistically significant.

f. For **P-value**, use P(Test statistic, $X^2_{df} > X^2$-value) = **X^2cdf**(X^2- value, 9999, *d.f.*).

g. Statistical Significance If $X^2 = \Sigma(O - E)^2/E$ is small, it indicates that the deviations between O and E are small ⇨ large P-value ⇨ the observed differences are *not* statistically significant.

2a. Chi-Square Goodness of Fit Test: there is one population from which **one sample** is drawn [cars, students, objects, accidents, births] then classified on the basis of 1 variable into *k* categories [brand, ethnicity, colour, days of the week, month #]. For the Goodness of Fit Test, a "model" is stated in terms of multiple proportions, with sample data to "verify" it.

b. We are interested in:
Ho: the sample data "fit" [is consistent with] the claims: $P_1 = a\%$, $P_2 = b\%$, ...$P_k = n\%$ for *k* Categories
H1: the sample data do not "fit" [is *not* consistent with] the claim **or** At least one proportion in Ho is different.

c. The multiple proportions *may* be the same [in the case of a "uniform" distribution"] so that
Ho: $P_1 = P_2 = P_3 = ...P_n = p\%$ **or** the distribution is uniform
H1: At least 1 proportion in Ho is different **or** the distribution is *not* uniform.

d. CALCULATIONS: Compute the Expected frequencies by first finding the Totals [= Sum of the Observed Frequencies ~ Sample Size, *n*]. Then, use the hypothesized % [in Ho] to find the individual Expected Frequencies by multiplying the % by sample size, *n*.

e. Goodness of Fit Test Qs only have 2 columns / rows: Observed and Expected Frequencies. The Degrees of Freedom is always: $k - 1$, where k ~ # of categories.

f. Goodness of Fit situations are *almost* never presented as 2 X 2, 2 X 3, 3 X 2, 3 X 3, etc. table format. The data *can* be presented as:

	Category 1	2	3	…	*k*
Population %	P_1	P_2	P_3	…	P_k
Sample % (*n* given)	p_1	p_2	p_3		p_k

so that the Observed *and* Expected Frequencies can be determined from the sample size, *n*.

3a. Chi-Square Test Independence: <u>one sample</u> is drawn from the population and then classified on the basis of 2 "factors" [e.g. Age vs. Height; Hair Color vs. Gender; Fatalities vs. Race, etc.].

b. We are interested in examining the claim:
Ho: Factor A and B are independent OR there is no association / relationship between Factor A and B.
H1: Factor A and B are not independent OR there is an association / relationship between Factor A and B.

c. CALCULATIONS: Compute the Expected Frequencies by making a Row X Column table and then using the formula: For cell with "coordinates" (*i*, *j*), **Row *i* Total · Column *j* Total / Grand Total**

It is customary to write the Expected values in ( ) to the right of / below the Observed Frequencies.

4a. Chi-Square Test of Homogeneity of Proportions: <u>several independently chosen samples</u> (say, *k*) are drawn from (*k*) *different* populations and the proportions of the occurrence of a certain characteristic from each sample is calculated.

b. We are interested in examining the claim:
Ho: There is no significant difference in the *k* proportions.
H1: There ***is*** a significant difference in the proportions. i.e. at least one proportion in Ho is different.

c. CALCULATIONS: Compute the Expected Frequencies by making a Row X Column table and then using the formula: For cell with "coordinates" (*i*, *j*), Row *i* Total X Column *j* Total / Grand Total

5. Independence Vs. Homogeneity
The Chi-Square Tests of Independence and Homogeneity differ in 2 fundamental ways:
a) <u>How the data was collected</u>: For Independence, we take *one* sample and classify it on the basis of 2 variables / factors / categories. Whereas, for Homogeneity, *2 or more* samples are taken from their respective populations, and the outcomes for 1 variable is noted.
b) <u>The Premise for the **Row Total · Column Total / Grand Total** formula</u>: while the Independence test uses the concept of Independence [Under Ho: Factor 1 and 2 are independent, so P(A and B) = P(A) • P(B)] to arrive at the formula, the Test of Homogeneity uses Pooling [Under Ho: there is no significant difference between the population proportions, so we can pool the estimates] to reach the same formula.

6. Graphical Display of r X c Tables for Test of Homogeneity and Independence
a. For Test of Homogeneity, use a **segmented bar-graph**, one bar for each sample, and each bar composed of the Relative Frequencies (%) of the different outcomes.
b) For Test of Independence, use a **segmented bar-graph**, one bar for each level of the either of the 2 factors, and each bar composed of the Relative Frequencies (%) of the different outcomes for *that* factor.

7. The different **Hypothesis Tests** are:
i) Single Mean: $\mu = \mu o$
ii) Single Proportion: P = Po
iii) Difference of Means: $\mu_1 - \mu_2 = 0$
iv) Mean Difference (Matched Pairs): $\mu_{1-2} = 0$
v) Difference of Proportions (Pooled): $P_1 - P_2 = 0$
vi) X^2-Test of Goodness of Fit: *hypothesized model is valid*
vii) X^2-Test of Independence: *the 2 variables are independent* **OR** *the 2 variables have NO association*
viii) X^2-Test of Homogeneity of Proportions: $P_1 = P_2 = \ldots P_n$ for *n* samples
ix) Linearity of the LSRL: $\beta = 0$

8. Distributions, Degrees of Freedom
ZAP-TAX ⇨ use Z-score for Proportions, t-score for Means.
For Single Means, the *d.f.* is t_{n-1}.
In case of Difference of Means, use conservative value of $t_{\min[n1,\, n2]-1}$.
For X^2 distributions, the *d.f.* is $X^2_{(r-1)(c-1)}$
For T-test for slope, the *d.f.* is t_{n-2}.

9. Interpretation of P-value: if the Null Hypothesis, Ho, was True, the P-value is the probability that we

would get sample values as extreme or more extreme than those observed. **OR**
If Ho were true, there is ONLY an *p*% chance of obtaining observations as extreme as the sample data in repeated sampling.

10. Statistical Significance An outcome [or statistic] is considered to be statistically significant [or simply, *significant*] if it is so "extreme" that its occurrence cannot be attributed to random chance / natural sampling variations, *were the hypothesized value of the parameter true*.

A treatment is said to be **statistically significant** when the observed effect – that is, the effect on a particular treatment on the response variable relative to other treatments – is so large that it would only rarely occur by chance. In other words, the effect of the treatment cannot be attributed to chance variation.

In the case of 2 or more samples, [under Ho] <u>assuming that there are indeed no significant differences between the populations</u>, the differences are **statistically significant** [or simply, *significant*], if the observed differences are so "extreme" or large that the outcomes cannot be attributed to random chance / sampling variations.

11. Type I Error: is the error of rejecting Ho *when it is indeed True*. That is, we find evidence in favour of H1 when Ho is really true! The probability of a Type I error is α, the significance level.

A Type I Error is also called a **False Positive** since P(False Positive) = P(test + | condition absent ~ Ho is true)

Since Type I Errors occur when we detect a statistically significant difference when, in truth, there is none, to reduce the probability of a Type I Error, decrease the significance level, α...so that we're less likely to reject Ho!

12. Type II Error: is the error of failing to reject the null hypothesis, Ho, when it is actually false. In other words, we *don't* find evidence in favour of H1 when H1 is *really* true! The probability of a Type II error is denoted by β.

A Type II Error is also called a **False Negative** since P(False Negative) = P(test - | condition present ~ H1 is True).

13. The **Power** of a Test is the probability of rejecting Ho when it is false i.e. it is the probability of correctly rejecting Ho: 1 – β.

A **Type II Error** is the error of failing to observe a difference when, in truth, there *is* one. Therefore, to reduce a Type II Error – correctly reject Ho / increase the Power of the Test – we need to be able to detect differences between the Treatment groups!

To reduce Type II error (β) or increase the Power of a Test, we *want* to detect differences between the 2 treatments *or* between the statistic and the hypothesized parameter...or, what is the same thing, find the observed differences to be statistically significant, so we

1. increase the significance level, α: this way, we're more likely to reject Ho!
2. increase the sample size, *n*: with a more precise estimate [lower S.E.(Estimate)], we're less likely to find outcome in favour of Ho [due to large test-statistic ~ low P-value]!
3. make the differences between the Treatments **or** differences between the observed and hypothesized values *as large as possible* by controlling for the effects of potentially confounding factors thereby reducing variability and the S.E.(Estimate) ⇨ larger Test-statistic ⇨ lower P-value. This can be achieved by controlling for extraneous Sources of Variation by

a) Direct Control
b) Blocking for a key Extraneous Factor.