Statistics Overview

1. Binomial Distribution-fixed sample size, independent trials, fixed probability of success p

$$X \sim B(n, p) \rightarrow P(X = r) = {}^{n}C_{r}(q)^{n-r}(p)^{r}$$

2. Poisson Distribution-rare occurrences, parameter of occurrence is proportional to frame of measurement; eg number of defects in certain length of cloth, number of cars rented in a week etc

$$X \sim P_o(\lambda) \rightarrow P(X = r) = \frac{e^{-\lambda}(\lambda)^{r}}{r!}$$

3. Normal Distribution-bell shaped curve with defined mean and variance

$$X \sim N(\mu, \sigma^2)$$

Standardisation to standard normal Z variable with mean 0 and variance 1:

$$P(X < \alpha) = P\left(Z < \frac{\alpha - \mu}{\sigma}\right)$$

Standard normal variable operations:

$$X \sim N(\mu_1, {\sigma_1}^2) \text{ and } Y \sim N(\mu_2, {\sigma_2}^2)$$

$$aX \pm bY \sim N(a\mu_1 \pm b\mu_2,\ a^2{\mu_1}^2 + b^2{\mu_2}^2)$$

(Care must be exercised when interpreting the question:

twice the weight of a durian has distribution $2X \sim N(2\mu, 4\sigma^2)$

and **two** durians selected has distribution $X_1 + X_2 \sim N(2\mu, 2\sigma^2)$)

4. Approximations:

Binomial to Poisson: $np < 5 \rightarrow X \sim P_o(np)$ approximately

Binomial to Normal: $np > 5,\ nq > 5,\ n > 30 \rightarrow X \sim N(np, npq)$

(Note: Continuity correction is required)

Poisson to Normal: $\lambda > 10 \rightarrow X \sim N(\lambda, \lambda)$

(Note: Continuity correction is required)

5. Central Limit Theorem (CLT):

Any **non normal** distribution with a large sample size can be approximated using CLT under the following 2 axioms:

(1) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ approximately

[Keywords in question: **sample mean, average**]

eg $X \sim B(20,\ 0.3)$ originally; if 60 samples of X were taken and the mean of which is to be investigated, then since 60>50, then by CLT, the distribution of the sample mean $\overline{X}$ can be modelled as being a normal

distribution approximately where $\overline{X} \sim N(20\,(0.3),\ \frac{20(0.3)(0.7)}{60})$

(2) $X_1 + X_2 + X_3 + ... + X_n \sim N(n\mu, n\sigma^2)$ approximately
[Keywords in question: **sum, total**]

(Note that CLT **MUST NOT** be used if original distribution is normal since the standard operations on normal distributions can be used directly-see (3))

6. Sampling:
True population mean and variance is unknown, hence a sample is extracted to estimate these population parameters.

Unbiased estimate of population mean = $\frac{\sum x}{n} = \frac{\sum (x-a)}{n} + a$

Unbiased estimate of population variance = $\frac{1}{n-1}\left(\sum (x-\overline{x})\right)^2$

$$= \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$$

$$= \frac{1}{n-1}\left(\sum (x-a)^2 - \frac{(\sum (x-a))^2}{n}\right)$$

$$= \frac{n}{n-1} \text{ (sample variance)}$$

(Note that sampling will be largely utilised in hypothesis testing when certain population parameters are not given in the question.)

7. Hypothesis testing:

Formulation of hypothesis: $H_0 :\ \mu = \mu_0$ (null hypothesis)

$H_1 : \mu \neq \mu_0$ OR $\mu < \mu_0$ OR $\mu > \mu_0$

(Note: the alternate hypothesis must be interpreted accurately-keywords such as **underestimating**, **overestimating**, **difference** in mean values aid in laying out H_1 correctly)

Test to be conducted: Z test (original distribution is normal, variance known)
Z test (original distribution is normal, variance unknown-must be estimated through sampling, sample size large)
t test (original distribution is **normal**, **variance unknown**-must be estimated through sampling, sample size **small**)

p value generated must be compared against level of significance α

$p > \alpha \Rightarrow H_0$ is NOT rejected

$p < \alpha \Rightarrow H_0$ is rejected and H_1 accepted

Based on rejection/acceptance of the null hypothesis, provide conclusion accordingly employing the following sentence template-**insufficient/sufficient evidence at the α % level of significance that……**

8. Regression:

-ability to provide a qualitative explanation of what is meant be least squares regression line
-ability to formulate equations of required regression lines accordingly
-appreciate the utilisation of regression lines for prediction purposes and being able to employ the relevant line equation (either y on x OR x on y) for such a purpose.
- understand the significance of the product moment correlation coefficient in assessing strength of linear associations amongst 2 variables, and its limitations since it is not capable of deducing other possible forms of associations between the variables (eg weak r value when there is clear indication based on graphical evidence that a quadratic association between 2 variables exist)

Regression line of y on x :

$$y = a + bx$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} = \frac{\sum (x-a)(y-a) - \frac{\sum (x-a)\sum (y-a)}{n}}{\sum (x-a)^2 - \frac{\left(\sum (x-a)\right)^2}{n}}$$

Since the point $(\bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n})$ lies on the regression line, once b is found from the above formula, a can be realised very easily through substitution.

Regression line of x on y :

$$x = c + dy$$

$$d = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}} = \frac{\sum (x-a)(y-a) - \frac{\sum (x-a)\sum (y-a)}{n}}{\sum (y-a)^2 - \frac{\left(\sum (y-a)\right)^2}{n}}$$

Since the point $(\bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n})$ lies on the regression line, once d is found from the above formula, c can be realised very easily through substitution.

Product moment correlation coefficient:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\left(\sum x^2 - \frac{\left(\sum (x)\right)^2}{n} \right)^{\frac{1}{2}} \left(\sum y^2 - \frac{\left(\sum (y)\right)^2}{n} \right)^{\frac{1}{2}}}$$

$$= \frac{\sum (x-a)(y-a) - \frac{\sum (x-a) \sum (y-a)}{n}}{\left(\sum (x-a)^2 - \frac{\left(\sum (x-a)\right)^2}{n} \right)^{\frac{1}{2}} \left(\sum (y-a)^2 - \frac{\left(\sum (y-a)\right)^2}{n} \right)^{\frac{1}{2}}}$$

Also note that $r^2 = bd$, and r will follow the sign (ie positive or negative) depending on the signs of b and d (either **both** positive or both negative).