Notes on Statistics

(Please answer according to the context of the question.)

The following statements are always true:

$$P(A) + P(A') = 1$$

$$P(A \cup B) + P(A \cap B) = P(A) + P(B)$$

$$P(A \cap B) = P(A)\, P(B \mid A)$$

A **is a subset of** B	A and B are **mutually exclusive**	A and B are **independent**
$P(A \cup B) = P(B)$	$P(A \cap B) = 0$	$P(B) = P(A \mid B)$

Combinations and Permutations

Tools of trade:

Choosing r **distinct** objects from n objects: ${}^{n}C_{r}$

Arrange n **distinct** objects
–- in **one row**: $n!$
–- in **one** non-flip-able **circle**: $(n-1)!$
–- in one **flip-able** circle: $(n-1)!/2$

If there are p **objects that are the same** (as well as q objects, r objects …):
divide by $p!$ (and $q!\,r!$ …)

Methods to consider:

Addition of **mutually exclusive** cases
Exclusion of **complementary** cases

Sampling Methods

A good sample should be
unbiased (random – every sampling unit has an **equal chance** of selection)
representative of the population (taking into account of population structure)
and **sufficiently large**.

Method	**Simple Random Sampling**	**Systematic**	**Stratified**	**Quota**
Choosing a sample of size n	Number the *units* from 1 to m. Use a random number generator to generate n distinct integers between 1 and m. *Units* numbered with the integers generated are chosen.	Number the *units* from 1 to m. Use a random number generator to generate an integer k between 1 and $a = \lceil \mathrm{m}/\mathrm{n} \rceil$. *Units* numbered $\mathrm{k} + z\mathrm{a}$ are chosen. ($z \in \mathbb{Z}_0^+, 0 \leq z \leq n-1$). ($\mathrm{k}$ should be the only variable in your answer – calculate all the others)	Using the population data, classify the *units* into strata. (Draw table) Number of *units* in each stratum to be sampled is proportional to the number of *units* in the population <Use SRS within each strata> Choose the sample for the other strata accordingly.	Split the population into strata (Draw table) Decide on the number of *units* in each stratum to be sampled (may not be proportional) Carry out the sampling (describe how the interviewer might sample these *units* – by standing on a street corner to ask people)
Info needed	Need sampling frame	Does not always need sampling frame Neither the population size is needed to sample a% of the population	Need sampling frame with population data to classify	Does not need sampling frame Neither the population size is needed to sample first n units from each strata
Efficiency	Time consuming Unavailable *units* cannot be replaced			A quicker method Unavailable *units* can be replaced
Bias	**Random** (every sampling *unit* has an **equal chance** of selection)			Biased as they might prefer – who are easier to interview – who are more approachable – from a certain strata (sample not proportional to population)
Rep.	Clustering may occur	Avoids clustering, but not if periodic	More representative as population strata is taken into account	Not representative because selection is already biased
	A random sample may not be representative of the population. It is possible for sample (esp. small ones) to have lopsided characteristics.			
Suitable for	Small, up to date population	Larger population	Population with significant strata, and its information known	For data to be collected quickly

<u>Probability distributions</u>

Distribution	Binomial	Poisson	Normal
Assumption	Each trial has **same probability** of success. The outcome of each trial is **independent** of the outcomes of other trials.	The events occur at **constant average rate**. The events occur **independently** of one another.	
Declaration	$X \sim B(n, p)$ $n =$ no. of trials $p =$ probability of success $E(X) = np$ $Var(X) = np(1-p)$	$X \sim P_0(\lambda)$ $\lambda =$ average no. of occurrences $E(X) = \lambda$ $Var(X) = \lambda$	$X \sim N(\mu, \sigma^2)$ $E(X) = \mu$ $Var(X) = \sigma^2$
Additivity	nil. need to approximate	$aX + bY \sim$ $P_0(a\lambda_1 + b\lambda_2)$ $a > 0, b > 0$ or else need to approximate	$X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2)$ $\overline{X} \sim N(\mu, \sigma^2/n)$ $aX + bY \sim$ $N(a\mu_1 + b\mu_2, a^2{\sigma_1}^2 + b^2{\sigma_1}^2)$
Recom-mended SOP	Express until GC-friendly $P(X > 4) = 1 - P(X \leq 4)$		As calculator takes input σ, define as $\sqrt{\text{variance}}^2$
Approximate	Binomial to Poisson	Poisson to Normal	Binomial to Normal
	$n > 50$ large $np < 5$ $X \sim P_0(np)$	$\lambda > 10$ $X \sim N\left(\lambda, \sqrt{\lambda}^2\right)$	$n > 50$ large $np > 5$ $n(1-p) > 5$ $X \sim N\left(np, \sqrt{np(1-p)}^2\right)$
Notes	Define success as failure if needed	Need to use c.c. (continuity correction)	

Standardising normal distribution, and using t-distribution (which is already standardised):

Given	$A \sim N(4, \sigma^2)$	$A \sim N(\mu, 1.23^2)$	$B \sim t(n-1)$	$B \sim t(n-1)$
	$P(A \leq 4) \leq 0.2$		$P(B \leq 4) \leq 0.2$	
Probability	$P\left(Z \leq \frac{4 - 6.2}{\sigma}\right) \leq 0.2$	$P\left(Z \leq \frac{\mu - 6.2}{1.23}\right) \leq 0.2$	$P\left(t(n-1) \leq \frac{4 - 6.2}{s}\right) \leq 0.2$	$P\left(t(n-1) \leq \frac{\mu - 6.2}{1.23}\right) \leq 0.2$
Output	$\frac{4 - 6.2}{\sigma} \leq invN(0.2)$	$\frac{\mu - 6.2}{1.23} \leq invN(0.2)$	$\frac{4 - 6.2}{s} \leq invT(0.2, n-1)$	$\frac{\mu - 6.2}{1.23} \leq invT(0.2, n-1)$

Hypothesis testing

unbiased estimate of population mean $\mu = \bar{x} = \frac{\sum x}{n}$

unbiased estimate of population variance $\sigma^2 = s^2 = \frac{n}{n-1}(\text{sample variance}) = \frac{n}{n-1}\left(\frac{\sum x^2}{n} - (\bar{x})^2\right)$

Since the **sample size n is large**, by **Central Limit Theorem**, the **distribution of the sample mean $\bar{X}$ is already normal**.
Therefore there is **no need for an assumption that X is normal** (but there is **still a need to assume that X is unbiased**).

Let X be is r.v. denoting the ______ and μ be the mean of the ______.
H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$

Since **population variance** σ^2 is known/unknown and
sample size n is large/small, (**assuming**, if X is not normal, and CLT does not apply) population is **normal**, under H_0, test statistic:

	σ^2 unknown need to calculate: $\mathbf{s} = \sqrt{\frac{n}{n-1}(\text{sample variance})}$	σ^2 known
$n \geq 50$ large (if not normal, CLT applies)	$Z = \frac{\bar{X} - \mu_0}{\mathbf{s}/\sqrt{n}} \sim N(0,1)$	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
$n \leq 50$ small (need to assume normal, if not)	$T = \frac{\bar{X} - \mu_0}{\mathbf{s}/\sqrt{n}} \sim t(n-1)$	

Using a **two/one-tailed test** at **α% significance level**, **reject H_0 if $p < \alpha$%**
Using GC, μ_0 =______, $\bar{x}$ =_____, σ or s =_____, n =_____:
$\mathbf{p - value}$ = _____ < α%, so we **(do not) reject H_0**. There is **(in)sufficient evidence at α% significance level** to (reject the claim made).

What do you understand by **1% level of significance**?
There is 1% chance of **concluding** that the *mass* is (**more or less/less/more**) **than** 8.5g **when it is currentl**y 8.5g.

A **p-value** of 0.121 means there is a probability of 0.121 that the **sample mean** is **as extreme as or more extreme than the observed value of sample mean**, **assuming that the population mean** is μ.

Correlation and Regression

If y (the dependent factor) is based on x (independent factor or controlled)
– Plot y on x

If both variables are random (cannot determine which one is controlled), if x is being estimated,
– Plot x on y

What consider in determining **whether model is appropriate**:

Shape of the scatter plot
Possibility of values taken
Likely long term behaviour

Sample answers, please use according to context:

A linear model is not likely to be appropriate as:	A quadratic model is not likely to be appropriate as:
The scatter diagram clearly did not indicate a straight line.	It would eventually have a maximum and then decrease increasingly steeply.
It would predict a continuous increase, eventually above 100%, which is impossible.	

Even if r takes a value close to 1, the predicted value of t is unreliable as **extrapolation** is involved.