# Notes on Statistics

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

The following statements are always true:

|  |  |  |
| --- | --- | --- |
| **is a subset of** | and are **mutually exclusive** | and are **independent** |
|  |  |  |

# Combinations and Permutations

Tools of trade:

Choosing **distinct** objects from objects:

Arrange **distinct** objects   
–- in **one row**:   
–- in **one** non-flip-able **circle**:   
–- in one **flip-able** circle:

If there are **objects that are the same** (as well as objects, objects …):   
**divide** by (and …)

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.  Use a random number generator to generate n distinct integers between 1 and.  *Units* numbered with the integers generated are chosen. | Number the *units* from 1 to m.  Use a random number generator to generate an integer between 1 and.  *Units* numbered are chosen. .  ( 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 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 |

# Probability distributions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | no. of trials  probability of success | average no. of occurrences | |  |
| Additivity | nil. need to approximate | or else need to approximate | |  |
| Recom-mended SOP | Express until GC-friendly | | | As calculator takes input, define as |
| Approximate | Binomial to Poisson | | Poisson to Normal | Binomial to Normal |
| large | |  | large |
| 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 |  |  |  |  |
|  | |  | |
| Probability |  |  |  |  |
| Output |  |  |  |  |

# Hypothesis testing

unbiased estimate of population mean   
unbiased estimate of population variance

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

Let be is r.v. denoting the \_\_\_\_\_\_ and be the mean of the \_\_\_\_\_\_.

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

|  |  |  |
| --- | --- | --- |
|  | unknown need to calculate: | known |
| large  (if not normal, CLT applies) |  |  |
| small  (need to assume normal, if not) |  |

Using a **two/one-tailed test** at  **significance level**, **reject if Using GC**, \_\_\_\_\_\_, \_\_\_\_\_, or \_\_\_\_\_, \_\_\_\_\_:   
, so we **(do not) reject** . There is **(in)sufficient evidence at 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 .

# 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.