APPENDIX 1

Are You Ready to Test Yourself? Chapter 2

Q: Which of the following is true?

1. An accurately measured variable remains constant for all participants in a sample.
2. As long as studies are accurately conducted and measured, the sample mean will remain the same for different samples.
3. If one chooses the sample wisely and calculates variables accurately, the resulting sample mean will be the same as the population mean.
4. Every variable has a unique characteristic; it cannot be converted.
5. Sample distribution is unique to the type of variable.

A: 5 is correct.

A variable is a measured characteristic of the population and would naturally vary from person to person. The sample mean will vary between different samples. Unless we observe the whole population, our sample mean will never be the same as the population mean. It is correct that every variable has a unique characteristic. However, measured variables can be expressed in more than one fashion for the sake of statistical analyses without affecting the integrity of the measurement. For example, one can measure alcohol consumption in terms of both a numerical variable (how many units one consumed in a week) or a categorical variable (alcohol consumed or not, or the amount consumed: mild, moderate, heavy etc.) Different types of variables have different types of distributions that dictate the particular types of statistical tests.

There are no accompanying questionnaires for Chapters 1 and 19.

Chapter 3

Q: Which of the following is true?

1. We can calculate the median or range from a box and whisker plot.
2. The interquartile range contains the data between two quartiles.
3. Mean is the sum of all values divided by the number of observations, it is not an actual value in the dataset.
4. The mode can be a suitable measure of central tendency for numerical data.
5. Median and range are the respective suitable measures of central tendency and spread for skewed data.

A: 1 is correct.

We can calculate median and range from the box and whisker plot, although the box indicates the interquartile range (IQR), the range is indicated by the whiskers and the outliers if any. The IQR is the range of data specifically between the 3rd and the 1st quartile, not just of any two quartiles. IQR contains the middle 50% of the observations (when data are ordered), 25% observations (1st quartile) are below this range and 25% above (4th quartile). A mean does not necessarily have to be an actual value in the dataset, but a real value can be the mean value if the calculation dictates so. The mode is not suitable for numerical data. It is uncommon for multiple numerical variables to have the same value. Even if there are numerous identical values, they may be outliers and may not represent data well. The mean is usually the suitable measure of central tendency for numerical data. However, when data are skewed or have outliers, these can unduly influence mean. In these circumstances, the median is more suited as a measure of central tendency. However, in this case, the most suitable measure of spread is the IQR.

Chapter 4

Q: Which of the following is true?

1. Variance and standard deviation (SD) have the same units of measurement.
2. Standard deviation is more precise and a better point estimate compared to the mean.
3. The sum of squared differences from the mean value is divided by the number of observations to calculate the variance.
4. We could use either the variance or the SD to describe the measure of the spread of data.
5. Mean and SD are always a more suitable measure of data than the median and interquartile range (IQR) when the numerical data are symmetric.

A: 5 is correct.

Variance and standard deviation (SD) do not have the same units of measurement. We square the differences to calculate the variance. Therefore, the unit is squared for the variance. It is not identical to the SD, which is the square root of the variance. SD is a measure of the spread of the data, whereas the mean is a measure of the central tendency. Both are essential and complementary. When one presents symmetrical numerical data, one should mention both mean and SD. We calculate variance by dividing the sum of squared differences from the mean by 'n-1' and not by 'n' (the number of the sample size). Variance is an artificially inflated number; it is not mentioned; instead, SD is presented to describe the measure of spread. When numerical data are symmetrical, i.e., there are no outliers or skew, the mean and the SD represent data better compared to the median and the IQR. The median and the IQR should only be used in numerical data when outliers skew the data distribution.

Chapter 5

Q: Which of the following is true?

1. We can calculate the *z*-score in a dataset where the mean and median have the same value.

2. The *z*-score has the same unit of measurement as the mean value.

3. The mean value is situated to the left of the mode in a dataset where outliers are larger than the bulk of the dataset.

4. The mean and the SD of standard normal distribution (SND) would vary according to the sample chosen.

5. Continuous random variables are more suitably displayed using a box and whisker plot.

A: 1 is correct.

To qualify to be converted to the *z*-score, the raw score has to come from a normal distribution, where the mean and median have the same or nearly identical value. To calculate the *z*-score, we first convert the distribution to a standard normal distribution with a mean of '0' and SD of '1'. The *z*-score does not have any unit of measurement. It is essentially the number of SDs away from the mean (*z*-score of 1 is 1 SD away from the mean.) The mean value is pulled by the outliers in a dataset. Therefore, the mean is larger than the mode in right-skewed data. The bulk of the data are smaller than the outliers in right-skewed data, and the mean value will be on the right side of the mode value. SND is a particular type of theoretical probability distribution with a mean value of '0' and SD of '1'. SND does not vary. Continuous random variables, in general, assume a normal distribution. A box and whisker plot is more suitable for data that are not normally distributed although it can be used for normally distributed data.

Chapter 6

Q: Which of the following is true?

1. A 95% confidence interval (CI) indicates a 95% probability that the mean of a different sample from the same population can also be found in this range.

2. 95% CI includes 95% of the dataset.

3. CI is a useful measure of descriptive statistics.

4. 95% CI gives us a precise measurement of the population mean.

5. Predicting the results of an election by polling a selected sample would require making inferences.

A: 5 is correct.

The CI is an estimate of the uncertainty of the population mean for a sample mean. It does not allow any inference about the mean from a different sample, even if it is from the same population. 95% CI gives us an estimate about the range of the values within which the actual population mean may be found. To find 95% of the dataset we need a mean ± 2 SD. CI is an inferential statistic. Despite its superiority as a statistical measure, there is no need to use it for descriptive statistics where we are just describing different data parameters but not making any inferences. CI is used to make inferences about the unknown population. For example, when describing the age range of a sample, it is enough to describe the mean value and the range if age data are normally distributed rather than the mean and 95% CI. We would need mean and 95% CI if we wanted to guess the age range of the population from our sample. 95% CI is an estimate of the unknown population mean and is not a precise measurement. We never know if the actual but unknown population mean will be found within this interval. Since we choose a sample and wish to predict the probable results of the polling from this sample, we have to use inferential statistics; this is usually done in the form of the CI (typically described as the margin of error of the poll).

Chapter 7

Q: Which of the following is true?

1. P<0.01 has a more potent treatment effect than p<0.05.
2. P=0.06 means the two treatment arms are not different in outcome.
3. A large p-value confirms that the null hypothesis is true.
4. P-value may be affected by the size of the sample.
5. A p-value <0.05 represents a real treatment effect.

A: 4 is correct.

P-value is the strength of evidence against the null hypothesis; the p-value does not give us any idea about the strength of the treatment effect. P<0.01 means there is less than a 1 in 100 chance of incorrectly rejecting the null hypothesis. P=0.06 means that no statistically significant difference was noted between the groups, but this does not necessarily prove there was no difference between them. The null hypothesis is never proven, it is only disproved. A large p-value represents strong evidence in favour of the null hypothesis but does not confirm it is true. All things being equal, the p-value is affected by the sample size. For the same size of treatment effect, the larger the sample, the easier it is for hypothesis tests to find a statistically significant difference between two different treatments if it exists. We have learnt that p<0.05 represents a 1 in 20 chance that we might have rejected the null hypothesis by mistake. Hence, not all significant results represent the true treatment effect.

Chapter 8

Q: Which of the following is true?

1. The power calculation indicates the maximum effect size detectable for a given sample size.

2. The null hypothesis is true when the p-value is large in an adequately powered study.

3. A post-hoc power calculation can be useful especially when the results are non-significant.

4. Since power is 1-β, increasing the α does not affect the power of a study.

5. The larger the α the more likely it is that we could reject the null hypothesis that is not false.

A: 5 is correct.

Power calculation allows us to calculate the minimum effect size likely to be detected for a given sample size, not the maximum effect size. If a study is adequately powered, even if the p-value is large, there is no guarantee that the null hypothesis is necessarily true. We never know whether the null hypothesis is true, only if it is not false. Power calculation should always be performed a priori. Post-hoc power calculations are not encouraged as the sample size and effect estimate are fixed and the results are unreliable. When studies report a non-significant result, we should investigate whether a priori and appropriate sample size calculation had been performed to reassure us that the study was adequately powered, not perform a post-hoc calculation ourselves. Type I and 2 errors have an inverse relation. Increasing the probability of type I error will, all things being equal, decrease the probability of type II error and affect the study power. α is the significance level where we reject the null hypothesis. If α is increased from 0.05 to 0.10, the corresponding probability of rejecting the null hypothesis, if it is not false, increases from 5% to 10%.

Chapter 9

Q: Which of the following is true?

1. A well-designed randomised controlled trial (RCT) does not suffer from bias.
2. In a controlled trial, the therapeutic benefit noticed in the intervention arm is only due to the intervention.
3. Detection bias cannot be controlled in an open-label trial.
4. When confounding factors are measured in a trial, their effect can be mitigated.
5. Intention-to-treat (ITT) analysis has the potential to commit a type I error when performed in a non-inferiority trial.

A: 5 is correct.

A well-designed RCT can certainly reduce bias to a large extent, but all trials suffer from bias. The therapeutic benefit noticed in a trial in the intervention arm is due to a combination of factors; undoubtedly, the therapeutic effect is likely to be predominant. Still, these patients also benefit from the natural course of the disease, the placebo effect, regression to the mean effect etc. Whether detection bias is present or not depends on the outcome of the trial. If outcomes are objective, for example, birth, death etc. then the risk of detection bias is immaterial even if the trial is not blinded. It does not matter whether confounding factors are measured or not, they can affect the study outcomes unless the sample is randomised to equally distribute these factors. ITT is the preferred analysis option in RCTs. However, in the case of non-inferiority trials, where ITT has the potential to favour no difference between treatments, a per-protocol analysis may be favoured. In a non-inferiority trial, ITT may favour the alternative hypothesis when it is not valid and thus lean towards a non-inferiority result. The null hypothesis in a non-inferiority trial is that the intervention arm is inferior to the control arm by a predetermined margin; the alternative hypothesis is that the intervention arm is not inferior to the control arm. Therefore, ITT may erroneously support the alternative hypothesis and commit a type I error.

Chapter 10

Q: Which of the following is true?

1. A paired sample *t*-test cannot be performed if the variables are not normally distributed.

2. The independent samples *t*-test assumes that the sample sizes in the two groups are similar.

3. A parametric test should not be performed if sample data are not normally distributed.

4. A large *t*-statistic is a confirmation of strong evidence against the null hypothesis.

5. The *t*-distribution is unique for each degrees of freedom.

A: 5 is correct.

It is not the variables themselves but the paired difference that needs to be normally distributed for the paired samples *t*-test to be valid. Even if the variables are not normal, as long as the distribution of the paired differences is normal the test is still valid. There is no need for sample sizes to be equal for an independent samples *t*-test to be valid, although the difference must not be substantial. Although it is ideal to perform a non-parametric test when the assumption of normality is not met, this is not necessarily always the case. Parametric tests can accommodate a degree of departure from normality. For minor departures, a parametric test is entirely valid provided other assumptions are also met. Another option is to perform both parametric and non-parametric tests and accept the more conservative of the two estimates. Besides, one can also transform the data to allow a parametric test to be performed. A large *t*-statistic is likely to indicate strong evidence against the null hypothesis but is not enough on its own. The *t*-statistic needs to be compared against the *t*-distribution (with the degrees of freedom of the sample) to derive the actual p-value. Degrees of freedom affect the shape of the *t*-distribution. With larger degrees of freedom, the tails get flatter, and the hump gets wider.

Chapter 11

Q: Which of the following is true?

1. The alternative hypothesis of the ANOVA test accepts that some of the population means may be the same.

2. If the null hypothesis is rejected in ANOVA all the population means are different from each other.

3. If the null hypothesis were true, the probability of obtaining an *F*-statistic larger than the critical value would be less than 0.05.

4. Post-hoc tests can be performed regardless of the value of the *F*-statistic to determine if there are individual pairwise differences.

5. A non-parametric equivalent test should be performed if the data distribution is not normal.

A: 1 is correct.

The alternative hypothesis in the ANOVA test accepts that at least one of the means is different. It does not indicate the superiority of one specific mean, nor does it suggest that all means are different; it may well be that if five means are tested only one is different and the others are not significantly different. We conclude that the null hypothesis is false ($p<0.05$) when the *F*-statistic is larger, not smaller, than the critical value. The purpose of the post-hoc test is to determine which means are different when the *F*-statistic is larger than the critical value and we reject the null hypothesis. If we cannot reject the null hypothesis this signifies that the means are not significantly different from each other; in this scenario, there is no benefit in performing post-hoc tests to investigate which pairs of means are different from each other. When the distribution curve is skewed, a non-parametric test is best. However, when the departure is minor, as seen on visual inspection of the histogram, the ANOVA test can still be performed.

Chapter 12

Q: Which of the following is true?

1. In a Chi-Squared test with a 5x3 contingency table, at least 12 cells must have frequencies >5.

2. A small p-value suggests a strong association between the row and the column variables.

3. The Chi-Squared test can be performed to investigate if the risk of bronchial asthma rises with an increasing degree of air pollution.

4. The calculation of expected frequencies in the Chi-Squared test is based on the alternative hypothesis.

5. This test could be used to compare the prevalence of cigarette smokers between lung cancer cases and a matched-control population.

A: 3 is correct.

The requirements for frequencies are for expected ones, not observed frequencies. 5x3 = 15 cells, 15x80% = 12; at least 12 cells must have expected frequencies >5 and all >1. The Chi-Squared test is a test of the null hypothesis. It merely informs us whether there is a statistically significant association between the variables, but not the strength of the relationship. The test can be performed for ordinal data. Air pollution can be divided into increasing categories of severity (normal, mild, moderate, severe, extreme) and risk of bronchial asthma (present or absent). The Chi-Squared test for trend can be used in this case. The test compares the expected frequencies against the observed frequencies. Expected frequencies are calculated based on the scenario that the null hypothesis was not false. This test cannot be performed for paired observations. The independence of data is a fundamental assumption of this test. Although the observations are not repeated from the same group, the pairs are matched. The Chi-Squared test will not be valid. McNemar's test would be more appropriate.

Chapter 13

Q: Which of the following is true?

1. A negative correlation coefficient signifies no correlation between the *x*- and the *y*-axis.

2. A linear relation is best measured with linear regression.

3. A regression model can be used to predict any value of *y* from *x*.

4. The equation of the simple linear regression will be affected if the dependent and independent variables are changed over.

5. A strong correlation coefficient means a strong causative relationship between the *x*- and the *y*-axis.

A: 4 is correct.

A negative correlation does not indicate there is no correlation between the *x*- and the *y*-axis. The correlation could well be strong, in a negative direction. A negative correlation means that with an increase in one variable, the other moves in the opposite direction. Linear regression is not able to measure correlation, we need a correlation coefficient for this measure. The regression model cannot be extrapolated beyond the range of values of *x* in the model. If the variables are changed the relationship between the *x* and the *y* variables will change resulting in a change in the equation of the simple linear regression. The correlation coefficient is a measure of the linear relationship and does not imply a causative relationship.

Chapter 14

Q: Which of the following is true?

1. A model that is statistically significant in the Hosmer and Lemeshow test suggests satisfactory goodness of fit of the model.

2. The event rates in the sample dictate the number of independent variables that can be reliably included in a logistic regression model.

3. A model with a good fit has excellent predictive ability.

4. An odds ratio of 0.95 for an event means that the event is 95% more likely to take place.

5. When a significant predictor from univariable analysis becomes non-significant after multivariable analysis, it suggests considerable confounding.

A: 2 is correct.

The Hosmer and Lemeshow test is conducted to check for the lack of goodness of fit of the model. When the test is statistically significant, it means that there is a large discrepancy between the observed and expected events and the model does not have a good fit. There is a rule of thumb of a minimum of 10 events for each independent variable in a logistic regression model. A model that has a good fit does not necessarily indicate a good predictive ability. Prediction of the model is judged by the pseudo R2 value, not by the goodness of fit. An odds ratio of 0.95 means that the event is less likely. Since the probability is 1, the event is 1-95% = 5% less likely to occur. It is true that when multivariable analyses are performed to correct for confounding, some factors deemed significant in the univariable analysis may be found to be non-significant due to the effect of confounding. However, although this confirms that there was confounding, the analysis does not inform us regarding the size of confounding. It may well be that the size of confounding was minimal; the difference between the crude and the adjusted odds ratios will give us a better idea concerning the likely size of confounding.

Chapter 15

Q: Which of the following is true?

1. Relative risk (RR) and odds ratios (ORs) can be used interchangeably.

2. A patient with higher baseline risk would have greater absolute benefit from a beneficial intervention than those at a lower risk.

3. The smaller the magnitude of the Numbers needed to harm (NNH), the more beneficial the treatment.

4. Odds ratios and relative risk estimate can be controlled for confounders and the effect of covariates.

5. A drug with a 50% relative risk reduction against another agent is likely to have a clinically significant treatment benefit.

A: 2 is correct.

When event rates are low, RR and OR approximate each other, but they should not be used interchangeably, the calculations are different. Absolute risk reduction is higher when the event rate or baseline risk is higher, and vice versa. A patient with a higher baseline risk would have a larger gain from a beneficial treatment. In contrast to the Number needed to treat, the smaller the NNH the more harmful the treatment. ORs can be adjusted for confounding by performing a logistic regression analysis. RR cannot be adjusted for confounders and covariates. Although 50% RR reduction is impressive, on its own it is not enough; we also need information on the baseline risk and absolute risk reduction to decide if the improvement is clinically significant.

Chapter 16

Q: Which of the following is true?

1. The calculation of likely survival probability in a Kaplan-Meier analysis is accurate as long as the sample size is representative.

2. The hazard rate does not change irrespective of time.

3. A patient with complete follow-up is not censored.

4. The median survival time is not always calculated.

5. The magnitude of difference in survival between two different curves can be investigated with the Log-rank test.

A: 4 is correct.

Since it is not possible to be sure about the fate of the patients who were censored, the Kaplan-Meier analysis is at best an estimate. As soon as the first patient gets censored, the curve becomes an estimate, not accurate. The hazard rate is the instantaneous event rate in the at-risk group and will change throughout the study period. To construct Cox proportional hazards modelling, an underlying assumption has to be fulfilled that the Hazard ratio remains constant throughout time. Even if the follow-up is complete if a patient does not experience the event of interest that patient is censored. Although median survival time is the preferred indicator of average survival, it is not always possible to calculate it. This is because it is not possible to extrapolate data. If the cumulative probability of survival does not reach 0.5, it is not possible to calculate the median survival probability. The Log-rank test is a significance test; it only informs us if two curves are significantly different, it gives no information about the magnitude of difference.

Chapter 17

Q: Which of the following is true?

1. The authors investigated the diagnostic accuracy of asymmetric skin creases for the diagnosis of pathological developmental dysplasia of the hip (DDH). The reliability of asymmetric skin creases in the diagnosis of pathological DDH will not change if the study is repeated in a different setting.

2. Concerning the previous question, it is important to ascertain the study prevalence before deciding that the clinical sign is unreliable.

3. The Receiver operating characteristic (ROC) curve would allow us to graphically plot the sensitivity against the specificity of the diagnostic test to assess for optimal test threshold.

4. Researchers found that synovial biopsy had a positive predictive value of 95% and a negative predictive value of 100%. The test was more useful to rule in the disease.

5. Researchers investigating the accuracy of a test for diagnosis of periprosthetic joint infection found the negative likelihood ratio of the test to be 0.248. Patients incorrectly labelled as negative outnumbered those who were correctly labelled negative.

A: 2 is correct.

Predictive values depend on the prevalence of the disease, and Positive predictive value (PPV) improves with the increasing prevalence of the disease. Therefore, if the test is repeated in a different clinical setting with improved prevalence, the PPV is likely to improve. It is, therefore, essential to establish what was the prevalence of the condition in the study population. Ideally, the prevalence of the study sample should match that of the population. Otherwise, the values are unreliable. ROC curves plot sensitivity versus 1-specificity (false positive). The test had a negative predictive value (NPV) of 100%, which means that if the test is negative in the target population, we can confidently rule out the disease. The test was more useful to rule out the disease. The negative likelihood ratio (NLR) is a ratio of False negative rate (FNR) ÷ True negative rate (TNR). Therefore, cases that were correctly labelled as negative outnumbered those that were incorrectly labelled as negative.

Chapter 18

Q: Which of the following is true?

1. Meta-analysis cannot be performed when the treatment effect varies between studies.

2. Larger studies, due to the trial size, get more weight in a meta-analysis.

3. In a fixed-effects model, even if we were to collect infinite sample sizes of each study, the study results would still be the same.

4. In a random-effects model, even if we were to collect infinite sample sizes of each study, the study results would still be the same.

5. An asymmetric funnel plot is confirmative of publication bias.

A: 3 is correct.

Variation in treatment effect is suggestive of statistical heterogeneity. The decision to proceed to a meta-analysis in the presence of statistical heterogeneity is a nuanced one. One should assess the quality of the evidence and investigate the causes of heterogeneity before deciding whether or not to proceed with meta-analysis. The sample size is a crucial consideration in the weightage of studies, but more importance is placed on the precision of the estimate in the form of variance rather than just sample size on its own. Because the fixed-effects model assumes that the actual effect is fixed across studies, even if we take infinite samples, the result should remain identical. The results would vary in a random-effects model since the underlying assumption is that the actual effect has its own distribution and randomly varies across studies. An asymmetric funnel plot is suggestive of publication bias but is not confirmative. Methodological bias can also contribute to asymmetry.