

# MCQS - Statistics course 1

**Written by: Robin Beaumont e-mail: robin@organplayers.co.uk**

**Date last updated Wednesday, 03 November 2010**

**Version: 1**

This document lists all the MCQs in Statistics course 1 that may be in the timed online MCQ exam.

## Contents

1. Defining data.....	2
2. Defining the centre.....	3
3. Graphics.....	5
4. Spread.....	8
5. Sample / Populations.....	10
6. Assessing a single mean.....	12
7. Assessing two means.....	15
8. Assessing Ranks.....	17
9. Correlation.....	19
10. Simple regression.....	21
11. Proportions and Chi square.....	23
12. Risk, rates and odds.....	23
13. Survival analysis.....	24
14. Hypotheses, Power and sample size.....	26
15. Simple logistic regression.....	28

# 1. Defining data

1. I suggest two reasons why I feel people fall foul at the first hurdle of learning statistics. Which of the following are they? (two correct choices)

- a. 'user friendly' introductions under emphasising basic concepts
- b. 'user friendly' introductions incorrectly explaining basic concepts
- c. statistics presented as a poorly defined subjective discipline
- d. over emphasis on the use of computers
- e. statistics presented as a clear cut subject with clearly defined rules

2. Which of the following is an example of nominal data? (one correct choice)

- a. Number of people on a course
- b. Cancer staging scale
- c. List of different species of bird visiting a garden over the past week
- d. Popularity rating of UK top ten television programmes
- e. Heart rate

3. Which of the following are examples of Interval/Ratio data? (two correct choices)

- a. Number of people on a course
- b. Cancer staging scale
- c. List of different species of bird visiting a garden over the past week
- d. Popularity rating of UK top ten television programmes
- e. Heart rate

4. Which of the following are examples of Ordinal data? (two correct choices)

- a. Number of people on a course
- b. Cancer staging scale
- c. List of different species of bird visiting a garden over the past week
- d. Popularity rating of UK top ten television programmes
- e. Heart rate

5. Which of the following is the correct listing of data from the simplest to the most complex? (one correct choice)

- a. Nominal -> Ordinal -> Interval -> Transcendental
- b. Nominal -> Ordinal -> Interval -> Ratio
- c. Qualitative -> Ordinal -> Interval -> Discrete
- d. Qualitative -> Ordinal -> Interval -> Ratio
- e. Nominal -> Ordinal -> Interval -> Quantitative

6. Which of the following is an **incorrect** statement about Ranking a dataset? (one correct choice)

- a. You can rank any dataset as long it is not Nominal
- b. Each value in a dataset should only occur once
- c. The process of ranking a dataset involves ordering it and then assigning a 'rank' value to each score from 1 to the number of scores in the dataset.
- d. When ranking a dataset tied scores receive the average of the rank value given to the ties.
- e. The result of ranking a dataset means that you loose the effect of magnitude if the data were Interval/Ratio

## 2. Defining the centre

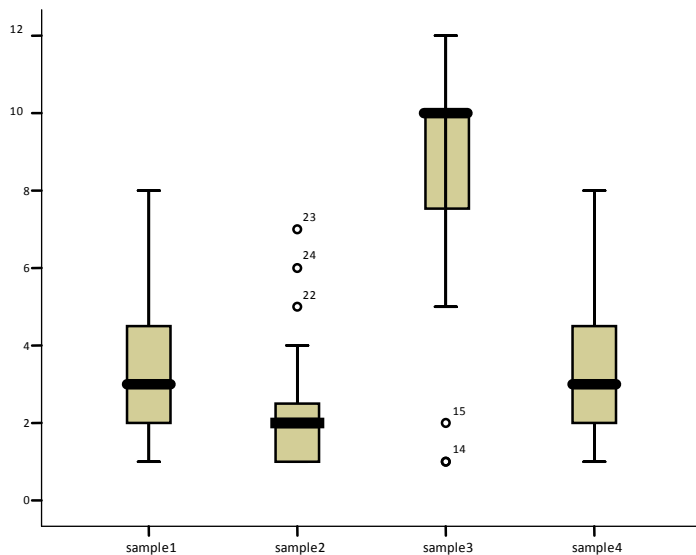
1. What is the main aim of descriptive statistics (one correct answer)?
  - a. Reduce the number of scores to a smaller more typical dataset
  - b. Create a comparable dataset
  - c. Reduce the number of scores to one value that provides a description of some aspect of the dataset
  - d. Increase the dataset to represent a population
  - e. Provide a narrative interpretation of a dataset
  
2. Which of the following are measures that attempt to describe the typical score of a dataset (three correct answers)?
  - a. Deviation
  - b. Mode
  - c. Median
  - d. Mean
  - e. Residual
  - f. Range
  - g. Frequency
  
3. Which of the following provides the best definition of a **frequency** when the term is applied to a dataset?
  - a. The number of occurrences for a range of values that a variable takes in a data set
  - b. The number of occurrences for zero values that a variable takes in a data set
  - c. The number of occurrences for one, or a range of values that a variable takes in a data set
  - d. The number of occurrences for the mean value that a variable takes in a data set
  - e. The number of occurrences of inappropriate values that a variable takes in a data set
  
4. Which of the following provides the best definition of the term **relative frequency** when the term is applied to a dataset?
  - a. The same as the percentage for a particular value a variable may take in a dataset
  - b. The number of occurrences of the mean value that a variable takes in a data set divided by the total number of scores in the dataset
  - c. The number of occurrences for one, or a range of values that a variable takes in a data set divided by the total number of scores in the dataset
  - d. The number of occurrences of negative deviations divided by the total number of scores in the dataset
  - e. The number of occurrences of positive deviations divided by the total number of scores in the dataset
  
5. For which type of data is the mode the most appropriate descriptive statistic?
  - a. Ordinal
  - b. Interval/ Ratio
  - c. Nominal
  - d. Textual
  - e. Quantitative

6. To work out the median by inspecting the scores what must you first do to the dataset?
- Remove any negative values
  - Rank
  - Work out the mean
  - Count the total number of scores
  - Know the formula to use
7. What is the main difference between the median and mean?
- The median uses the ranked values whereas the mean uses the frequencies
  - The median uses the ranked values whereas the mean uses the actual values
  - The mean uses the ranked values whereas the median uses the actual values
  - There is no difference
  - The median uses deviations whereas the mean uses the actual values
8. When calculating the median for a dataset consisting of an even number of scores (i.e. 2,4,6 etc.) which of the following is correct?
- Calculate the average value of the three middle ranked scores
  - Calculate the mean for the whole dataset which would provide the same answer in this instance
  - Calculate the average value of the two middle ranked scores
  - Calculate the mode and use instead
  - Choose either the upper or lower value of the two
9. Which of the following statements concerning the mean is **incorrect** (choose one)?
- The mean is not suitable for nominal data
  - The mean is sensitive to a single extreme value
  - The mean should always be used as the preferred measure to indicate a typical score
  - The mean is a more complex descriptive statistic than either the mode or median
  - The mean provides the most sensible result when the interval/ratio dataset has a symmetrical set of scores
10. The mean can be interpreted as (choose one)?:
- The centre of gravity of a dataset
  - The average of the mode and median values of a dataset
  - The weight of all the scores
  - The weight of all the positive deviations
  - The relative frequency with the highest value
11. In a positively skewed dataset the various measures suggesting a typical value lie in the following order (choose one)
- median -> mode -> mean
  - mode -> median -> mean
  - mean -> mode -> median
  - mean -> median -> mode
  - mode -> mean -> median

### 3. Graphics

#### Exercise 2

The following four boxplots provide summary information from four datasets. Please answer the following multiple choice questions (MCQs).



1. Which sample has the highest median (one correct answer)?

- a. Sample1
- b. Sample2
- c. Sample3
- d. Sample4
- e. None of them

2. Which sample has a median value of 2 (one correct answer)?

- a. Sample1
- b. Sample2
- c. Sample3
- d. Sample4
- e. None of them

3. Which sample has outliers with low scores (one correct answer)?

- a. Sample1
- b. Sample2
- c. Sample3
- d. Sample4
- e. None of them

4. Which sample has outliers with high scores (one correct answer)?

- a. Sample1
- b. Sample2
- c. Sample3
- d. Sample4
- e. None of them

5. Which sample has 50% of its scores which do not overlap the other sets of scores to a great extent (one correct answer)?

- a. Sample1
- b. Sample2
- c. Sample3
- d. Sample4

The following histograms represent four different datasets. Study them carefully before answering the questions on the following page.

Chart 1

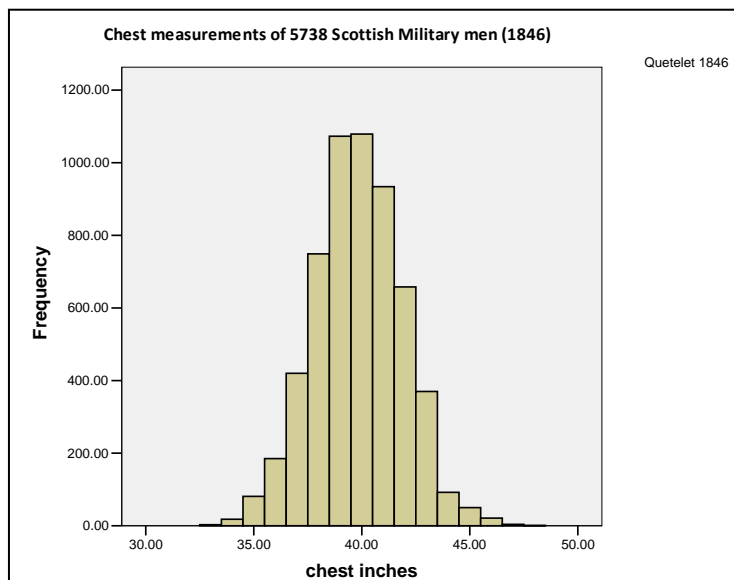
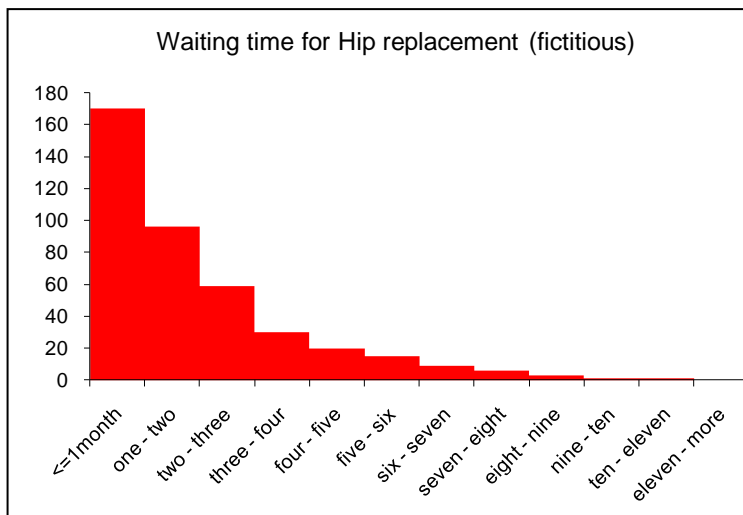


Chart 2

Chart 3

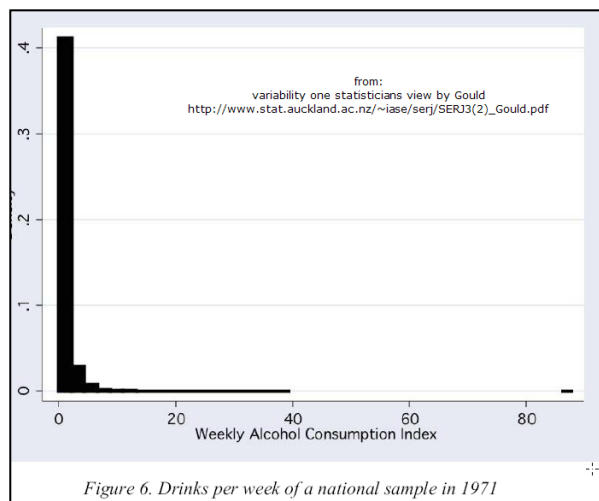


Figure 6. Drinks per week of a national sample in 1971

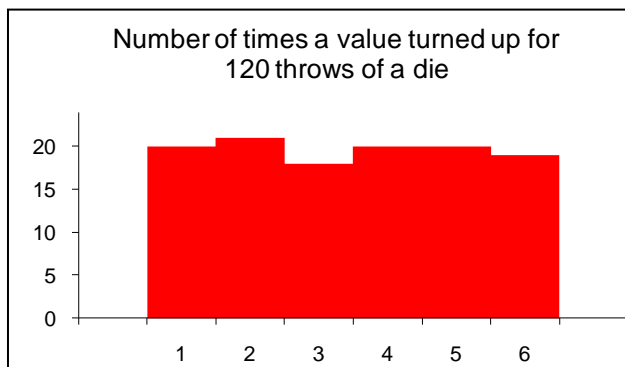


Chart 4

1. Which one of the charts suggests that the data form a uniform distribution (one correct answer)?

- a. Chart 1
- b. Chart 2
- c. Chart 3
- d. Chart 4

2. Which one of the charts suggests that the data form a normal distribution (one correct answer)?

- a. Chart 1
- b. Chart 2
- c. Chart 3
- d. Chart 4

3. Which two of the charts suggests that the data form a negative exponential distribution (one correct answer)?

- a. Chart 1 and 2
  - b. Chart 2 and 3
  - c. Chart 3 and 4
  - d. Chart 3 and 1
-

## 4. Spread

1. The interquartile range includes the following scores? (one correct choice)
  - a. 50% of the un ranked scores
  - b. 25% of the ranked scores
  - c. 70% of the rank scores
  - d. 50% of the ranked scores
  - e. 70% of the un ranked scores
  
2. Summing (adding together) all the deviations from the mean produces the following value? (one correct choice)
  - a. half the standard deviation
  - b. the standard deviation
  - c. 0
  - d. the mean value for the set of scores
  - e. the median value for the set of scores
  
3. What is an alternative name for the deviation from the mean? (**two** correct choices)
  - a. residual from the mean
  - b. derivation from the mean
  - c. residual from the median
  - d. error from the mode
  - e. error from the mean
  
4. Why does the standard deviation formula have a square root as part of it? (**one** correct choice)
  - a. to make it add up to the mean
  - b. to reverse the effect of squaring the deviations
  - c. to provide a standard (i.e. mean=0; sd=1) unit of measure
  - d. to provide a smaller value
  - e. none of these
  
5. Which of the following Greek letters represents the mean of a population? (**one** correct choice)
  - a.  $\beta$
  - b.  $\alpha$
  - c.  $\mu$
  - d.  $\epsilon$
  - e.  $\lambda$
  
6. Sigma squared represents? (**one** correct choice)
  - a. Population variance
  - b. Sample standard deviation
  - c. Population standard deviation
  - d. Population range
  - e. Sample variance



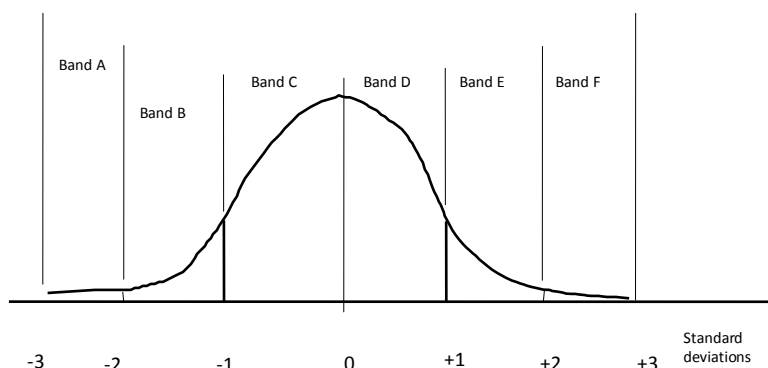
7. What specific strategy do I recommend when you come across Greek letters in statistical equations?

- a. Ignore them
- b. Replace them with familiar names
- c. Write the Greek name above them and practice saying the equation as a sentence.
- d. Write a familiar English name above them and practice saying the equation as a sentence.
- e. Write the Greek name above them.

8. For a set of data that follow a normal distribution how many scores can one expect to find within one standard deviation on each side of the mean, that is two standard deviations in total? **(one correct choice)**

- a. 54%
- b. 99%
- c. 50%
- d. 88%
- e. 68%.

9. A mother has a child and tells all her friends that he has an IQ of 113 on the Wechsler scale and is truly exceptionally intelligent. Given that the mean is 100 and the standard deviation 15 in which band on the graph opposite does he fit in? **(one correct choice)**



- a. Band A
- b. Band B
- c. Band C
- d. Band D
- e. Band E
- f. Band F

10. How 'truly exceptional' is the above child, which of the following most accurately reflects this situation? **(one correct choice)**

- a. Slightly under 14% of children would have a score less than his
- b. Slightly under 54% of children would have a score less than his
- c. Slightly under 84% of children would have a score less than his
- d. Slightly under 94% of children would have a score less than his
- e. He is in the top 1% of children.

11. A friend of the above lady also had her child tested and discovered that her daughter had an IQ of 2 standard deviations above the average IQ, she assumes that this must be far less than the very gifted boy. What is her child's IQ? **(one correct choice)**

- a. 115
- b. 120
- c. 125
- d. 130
- e. 145

12. A sample of data is highly negatively skewed? **(one correct choice)**

- a. Standard deviations should never be used to report the spread of such scores.
- b. Standard deviations are always the most appropriate measure to report the spread of such scores.
- c. Dependent upon the Standard deviation values it may be an inappropriate measure to report the spread of such scores.
- d. The degree of skewedness is irrelevant in deciding to use the standard deviation.
- e. In this instance the standard deviation should be divided by the number of scores to obtain a more valid measure.

## 5. Sample / Populations

1. The total area represented by a probability histogram is equal to: (one correct choice)

- a. The p value
- b. undefined
- c. infinity
- d. 1
- e. n

2. Within statistics the term pdf stands for: (one correct choice)

- a. Probability disease function
- b. Probability deviance function
- c. Portable Document format
- a. Probability density function
- b. Portable density function

3. The pdf is a function that considers all the values for a particular random variable and allocates the following: (one correct choice)

- a. A residual
- b. A Probability
- c. A odds
- d. A odds ratio
- e. A survival function

4. The normal pdf takes two parameters to fully define it, they are: (two correct choices)

- a. Mean
- b. Median
- c. Mode
- d. Variance
- e. Range
- f. Skewness
- g. Kurtosis
- h. M estimator
- i. t value

5. For the normal pdf a value of 1.96 standard deviations each side of the mean is where approximately X percent of the scores lie. X is equal to: (one correct choice)

- j. 25%
- k. 50%
- l. 75%
- m. 85%
- n. 95%
- o. 100%

6. The degrees of freedom concept can be summed up as: (one correct choice)

- a. The number of data items that are not free to vary, that is the parameter estimates
- b. The number of data items that are free to vary plus those used for parameter estimation
- c. The number of data items that are not free to vary plus those used for parameter estimation
- d. The number of data items that are free to vary
- e. None of the above

7. Which of the following best describes what is meant by the sampling distribution of the mean? : (one correct choice)
- The theoretical process of non-randomly sampling from a population and recording the mean value of each sample to produce a distribution of sample means
  - The theoretical process of randomly sampling from a population and recording the range of values of each sample to produce a distribution.
  - The theoretical process of randomly sampling from a population and recording the mean value of each sample to produce a standard deviation
  - The theoretical process of non-randomly sampling from a population and recording the median value of each sample to produce a distribution of sample medians
  - The theoretical process of randomly sampling from a population and recording the mean value of each sample to produce a distribution of sample means
8. When we theoretically consider an infinite number of sample means the estimate of the standard deviation of them is called the: (one correct choice)
- Standard error of the sample
  - Standard error of the median
  - Standard deviation of the mean
  - Standard error of the population
  - Standard error of the mean
9. The standard error of the mean (SEM) has the following formula: (one correct choice)
- Sample variance divided by the square root of number in sample
  - Standard deviation of sample divided by the square root of number in sample
  - Standard deviation of sample divided by the number in sample
  - Estimated standard deviation of population divided by the number in sample
  - Standard deviation of sample multiplied by the square root of number in sample
10. The standard error of the mean (SEM) is related to sample size, specifically: (one correct choice)
- As sample size increases, SEM increases
  - As sample size increases, SEM stays constant
  - As sample size increases, SEM becomes less stable
  - As sample size increases, SEM decreases
  - None of the above
11. Standardized scores (also called Z scores) allow values to be compared with the standard normal pdf. They are calculated in the following manner: (one correct choice)
- $(\text{Score mean} - \text{population mean})/\text{standard deviation}$
  - $(\text{Score} - \text{mean})/\text{standard deviation}$
  - $((\text{Score} - \text{mean})/\text{standard error})$
  - $((\text{Score} - \text{mean})/\text{SEM})$
  - $((\text{Score} - \text{mean})/n)$
12. The process of estimation is an essential aspect of inferential statistics, it can be defined as: (one correct choice)
- The process of calculating unbiased, efficient, consistent values from populations to sample parameters
  - The process of calculating uniquely varying sensitive values from samples of population parameters
  - The process of calculating unbiased, efficient, consistent values from both samples or populations
  - The process of calculating unbiased, efficient, consistent values from samples of population means
  - The process of calculating unbiased, efficient, consistent values from samples of population parameters

13. A confidence interval of X% is best described in the following statement: (one correct choice)

- a. We are confident that the estimated parameter from our single study will equal the population value X% of the time in the long run
- b. We are confident that the interval obtained from our single study will NOT contain the estimated parameter X% of the time in the long run
- c. We are confident that the interval obtained from our single study will contain the estimated parameter but in the long run only X% of the time
- d. We are confident that the interval obtained from our single study will contain the estimated parameter X% of the time in the long run
- e. We are confident that the interval obtained from our single study will equal the estimated parameter (100-X)% of the time in the long run

14. A confidence interval of the mean of 90% is best described in the following statement: (one correct choice)

- a. We are confident that the mean obtained from our single study will equal the population mean 90% of the time in the long run
- b. We are confident that the interval obtained from our single study will contain the population mean 90% of the time in the long run
- c. We are confident that the interval obtained from our single study will NOT contain the population mean 90% of the time in the long run
- d. We are confident that the specific interval obtained from our single study contains the population mean but in the long run only 90% of intervals will contain the population mean
- e. We are confident that the interval obtained from our single study will contain the population mean with a probability of 0.9 in the long run

15. The width of a confidence interval varies over samples because of the standard error, but what happens when sample size increases: (one correct choice)

- a. Confidence interval increases in size
- b. Confidence interval decreases in size
- c. Confidence interval stays constant
- d. Confidence interval becomes less stable
- e. None of the options

## 6. Assessing a single mean

1. The  $t$  pdf has a mean value of: (one correct choice)

- a. 0
- b. 1
- c. 2
- d. 3
- e. 4

2. The one sample  $t$  statistic, according to Norman and Streiner (2009) can be interpreted as: (one correct choice)

- a.  $(\text{Observed difference in means})/(\text{pooled standard deviation}) = \text{signal/noise}$
- b.  $(\text{Observed difference in means})/(\text{expected variability in means due to random sampling}) = \text{noise/signal}$
- c.  $(\text{Observed difference in means})/(\text{expected variability in means due to random sampling}) = \text{signal/noise}$
- d.  $(\text{Observed mean})/(\text{expected variability in means due to random sampling}) = \text{noise/signal}$
- e.  $(\text{Observed difference in medians})/(\text{expected variability in medians due to random sampling}) = \text{signal/noise}$

3. The one sample  $t$  statistic, is suitable in the following situation: (one correct choice)
- Comparison of a sample mean to that of a population mean
  - Comparison of a sample proportion to that of a population proportion
  - Comparison of a sample mean to that of a population one, where the sampling distribution is exponential
  - Comparison of a sample distribution to that of a population
  - Comparison of a sample mean to that of a population one over a time period
4. The one sample  $t$  statistic, has a degrees of freedom equal to: (one correct choice)
- Number of observations in sample plus one
  - Number of observations in sample
  - Number of observations in sample minus one
  - Number of observations in sample minus two
  - Number of observations in sample minus three
5. The  $p$  value associated with the one sample  $t$  statistic, assumes the following: (one correct choice)
- Mean of sample is not equal to the comparator
  - Mean of sample less than that of the comparator
  - Mean of sample greater than that of the comparator
  - Mean of sample and comparator are identical
  - None of the above
6. The effect size measure (i.e. clinical importance measure) associated with the one sample  $t$  statistic, is calculated as: (one correct choice)
- $(\text{sample mean} - \text{population mean})/\text{standard error}$
  - $(\text{sample mean} - \text{population mean})/\text{standard deviation}$
  - $(\text{sample mean} - \text{population mean})/\text{number in sample}$
  - $(\text{sample mean} - \text{population mean})/\text{sample mean}$
  - $(\text{sample mean} - \text{population mean})/1$
7. The effect size measure (i.e. clinical importance measure) associated with the one sample  $t$  statistic, provides: (one correct choice)
- The difference between the hypothesised and observed mean
  - The probability of obtaining the observed difference in means
  - The probability of obtaining the effect size observed
  - The probability of the null hypothesis being true
  - A standardised measure of the difference between the hypothesised and observed mean

8. The **paired** sample  $t$  statistic, is suitable in the following situation: (one correct choice)
- Comparison of a sample proportion to that of a population proportion of 0.5
  - Comparison of a sample mean to that of a population one, where the sampling distribution is exponential
  - Comparison of a sample distribution to that of a population
  - Comparison of a sample mean of zero to that of a population one over a time period
  - Comparison of a sample mean to that of a population mean of zero
9. If we obtained a  $p$ -value of 0.034 ( $n=13$ , two tailed) from a paired sample  $t$  statistic, how would we initially interpret this outside of the decision rule approach (i.e. hypothesis testing): (one correct choice)
- We will obtain the same  $t$  value from a random sample of 13 observations 34 times in every thousand on average, given that the population mean is zero.
  - We will obtain the same  $t$  value from a random sample of 13 observations 34 times, or more in every thousand on average, given that the population mean is zero.
  - We will obtain the same of a more extreme  $t$  value from a random sample of 13 observations 34 times in every thousand on average.
  - We will obtain the same or a more extreme  $t$  value from a random sample of 13 observations 34 times in every thousand on average, given that the population mean is zero.
  - We are 0.966 (i.e.  $1-0.034$ ) sure that the null hypothesis is true.
10. If interval/ratio data are paired in a research design such as pre and post test a paired sample  $t$  statistic . . . : (one correct choice)
- Is the most appropriate test, regardless of the differences being normally distributed
  - Is the most appropriate test, if the differences are normally distributed
  - Is the most appropriate test, if the differences are NOT normally distributed
  - Is sometimes the appropriate test, if the differences are normally distributed and centred around zero
  - Is the least appropriate test, regardless of the differences being normally distributed
11. A  $p$  value is a special type of probability with two fundamental characteristics what are they . . . : (one correct choice)
- Conditional probability, range of values representing area(s) under PDF curve
  - Conditional probability, of a specific single value representing a  $x$  value along the PDF curve
  - Non-conditional probability, range of values representing area(s) under PDF curve
  - Conditional probability, always representing a single area under PDF curve
  - Non-conditional probability, representing a  $x$  value along the PDF curve
12. The conditional probability for a  $p$  value, is **usually** re-interpreted as . . . : (one correct choice)
- Parameter value = zero = specific alternative hypothesis
  - Parameter value = zero = alternative hypothesis
  - Parameter value = zero = null hypothesis
  - Parameter value = zero = not related to any hypothesis
  - Parameter value not equal to zero = probability of the null hypothesis being true
13. Before calculating a single sample or paired sample  $t$  statistic it is essential to . . . : (one correct choice)
- Perform graphical statistics. Review study design.
  - Perform descriptive/graphical statistics to assess assumptions. Review study design.
  - Not perform descriptive/graphical statistics to assess assumptions. Review study design.
  - Assess the difference between the mean and median. Review study design.
  - Not perform description statistics to assess assumptions nor review study design.

## 7. Assessing two means

1. The two independent samples  $t$  statistic, according to Norman and Streiner (2009) can be interpreted as: (one correct choice)

- a.  $(\text{Observed difference in means})/(\text{pooled standard deviation}) = \text{signal/noise}$
- b.  $(\text{Observed difference in means})/(\text{expected variability in means due to random sampling}) = \text{noise/signal}$
- c.  $(\text{Observed difference in means})/(\text{expected variability in means due to random sampling}) = \text{signal/noise}$
- d.  $(\text{Observed treatment mean})/(\text{expected variability in means due to random sampling}) = \text{noise/signal}$
- e.  $(\text{Observed difference in medians})/(\text{expected variability in medians due to random sampling}) = \text{signal/noise}$

2. The two independent samples  $t$  statistic, makes an additional assumption, compared to that of the one sample/paired  $t$  statistic, that is assessed by Levenes statistic what is this: (one correct choice)

- a. Variances of both samples is due to random sampling
- b. Variances of both samples is due to sampling bias
- c. Variances of both samples is due to sample size
- d. Variances of both samples is significantly different
- e. Means of both samples is due to random sampling

3. The sampling distribution of Levenes statistic follows a particular theoretical distribution which of the following is it? (one correct choice)

- a. Standard normal
- b.  $t$
- c. F
- d. Chi square
- e. Exponential

4. Traditionally, when evaluating a null hypothesis one makes use of a critical value. A critical value is . . .? (one correct choice)

- a. a value set by the computer to create a decision rule regarding acceptance/rejection of the null hypothesis
- b. a value you set to create a decision rule regarding effect size
- c. a value set by the computer to create a decision rule regarding acceptance/rejection of the null hypothesis
- d. a value you set to create a confidence interval regarding acceptance/rejection of the null hypothesis
- e. a value you set to create a decision rule regarding acceptance/rejection of the null hypothesis

5. Traditionally a critical value is set at one of the following. . .? (one correct choice)

- a. 0.05, 0.01, 0.00001
- b. 0.05, 0.01, 0.001
- a. 0.5, 0.1, 0.001
- b. 0.5, 0.01, 0.001
- c. 0.005, 0.001, 0.005

6. The two independent sample  $t$  statistic, is suitable in the following situation: (one correct choice)

- a. Comparison of two independent sample means where the samples are  $<30$
- b. Comparison of two independent sample means where the samples are  $>30$  or normally distributed
- c. Comparison of two independent sample means where the samples are exponentially distributed
- d. Comparison of a sample distribution to that of a independent population
- e. Comparison of a specified mean to that of a population one over a time period

7. The two independent samples  $t$  statistic, has a degrees of freedom equal to: (one correct choice)
- Number of observations in both samples plus one
  - Number of observations in both samples
  - Number of observations in both samples minus one
  - Number of observations in s both samples minus two
  - Number of observations in both samples minus three
8. The  $p$  value (two sided) associated with the two independent samples  $t$  statistic, assumes the following: (one correct choice)
- Mean of samples identical
  - Mean of sample one is not equal to that of sample two
  - Mean of sample one is less than that of sample two
  - Mean of sample one is greater than that of sample two
  - None of the above
9. Given that  $s_1$  = sample one and  $s_2$  = sample 2. The effect size measure (i.e. clinical importance measure) associated with the two independent samples  $t$  statistic, is calculated as: (one correct choice)
- $(s_1 \text{ mean} - s_2 \text{ mean})/\text{standard error}$
  - $(s_1 \text{ mean} - s_2 \text{ mean})/\text{standard deviation}$
  - $(s_1 \text{ mean} - s_2 \text{ mean})/\text{number in sample}$
  - $(s_1 \text{ mean} - s_2 \text{ mean})/\text{sample mean}$
  - $(s_1 \text{ mean} - s_2 \text{ mean})/1$
10. Given that  $s_1$  = sample one and  $s_2$  = sample 2. The effect size measure (i.e. clinical importance measure) associated with two independent samples  $t$  statistic, provides: (one correct choice)
- The difference between  $s_1$  mean and  $s_2$  mean
  - The probability of obtaining the observed difference in means
  - The probability of obtaining the effect size observed
  - The probability of the null hypothesis being true
  - A standardised measure of the difference between  $s_1$  mean and  $s_2$  mean
11. The two independent samples  $t$  statistic, is suitable in the following situation: (one correct choice)
- Comparison of a sample mean to that of a population mean of zero
  - Comparison of more than two sample means
  - Comparison of a sample mean to that of another sample mean
  - Comparison of a sample distribution to that of a population
  - Comparison of two sample means to that of zero
12. If we obtained a  $p$ -value of 0.034 ( $n=7,8$ , two tailed) from an independent samples  $t$  statistic, how would we initially interpret this outside of the decision rule (i.e. hypothesis testing) approach: (one correct choice)
- We will obtain the same  $t$  value from two independent random samples of the specified size 34 times in every thousand on average, given that both samples come from a population with the same mean.
  - We will obtain the same, or a more extreme,  $t$  value from two independent random samples of the specified size 34 times in every thousand on average.
  - We will obtain the same or a more extreme  $t$  value from a single random sample of the specified size 34 times, or more in every thousand on average, given that both samples come from a population with the same mean.
  - We are 0.966 (i.e.  $1-0.034$ ) sure that the null hypothesis is true.
  - We will obtain the same, or a more extreme  $t$  value from two independent random samples of the specified size 34 times in every thousand on average, given that both samples come from a population with the same mean.



13. If two independent samples (both less than 30 observations) of interval/ratio data are produced in a research design an independent samples  $t$  statistic . . . : (one correct choice)

- a. Is the most appropriate test, regardless of the scores being normally distributed or not
- b. Is the most appropriate test, if the scores are normally distributed
- c. Is the most appropriate test, if the scores are NOT normally distributed
- d. Is sometimes the appropriate test, if the scores are normally distributed and centred around zero
- e. Is the least appropriate test, regardless of the scores being normally distributed

## 8. Assessing Ranks

1. Rank order statistics assume the scale of measurement is . . . : (one correct choice)

- a. Nominal
- b. Ordinal
- c. Interval
- d. Ratio
- e. Binary

2. Which of the following gives the reasons for using rank order statistics . . . : (one correct choice)

- a. Normal distributions, or **not** ordinal data or sample size less than 20
- b. Non normal distributions, or ordinal data or sample size greater than 20
- c. Non normal distributions, or **not** ordinal data or sample size less than 20
- d. Normal distributions, or ordinal data, or sample size greater than 20
- e. Non normal distributions, or ordinal data, sample size irrelevant

3. Which of the following statistics is often called the non parametric equivalent to the two independent samples  $t$  statistic? (one correct choice)

- a. Wilcoxon
- b. Chi square
- c. Mann Whitney U
- d. Sign
- e. Kolmogorov – Smirnov (one sample)

4. Non parametric statistics use the ranks of the data, in so doing which of the following characteristics of the original dataset may be lost? (one correct choice)

- a. Range/ magnitude
- b. Median
- c. Rank order
- d. Group membership
- e. Number in each group

5. When investigating an ordinal data set which of the following is the most appropriate method of assessing values graphically? (one correct choice)

- a. Barchart with SEM bars
- b. Barchart with CI bars
- c. Boxplots
- d. Histograms
- e. Funnel plots

6. When carrying out a Wilcoxon matched- pairs statistic on a small dataset (i.e.  $n < 50$ ), what method of p-value computation is the most appropriate? (one correct choice)

- a. Asymptotic
- b. Bootstrapped
- c. Simulated
- d. Z score approximation
- e. Exact method

7. When carrying out a Wilcoxon matched- pairs statistic which of the following is NOT a sample data assumption? (one correct choice)

- a. Must be ordinal/interval or ratio scale
- b. Paired observations independent
- c. Number of tied ranks must not be excessive
- a. Distributions should be symmetrical
- b. Not normally distributed

8. When carrying out a Wilcoxon matched- pairs statistic what method do I suggest you use to obtain the confidence intervals? (one correct choice)

- a. Carry out calculations by hand
- b. Use R
- c. Use SPSS
- d. Don't bother
- e. None of the above

9. The unstandardized effect size measure (i.e. clinical importance measure) associated with the Wilcoxon matched-pairs statistic is the . . (one correct choice)

- a. Signed-rank statistic (S) with an expected value of zero
- b. Signed -rank statistic (S) with an expected value of 1
- c. Signed -rank statistic (S) with an expected value equal to its maximum value
- d. Signed -rank statistic (S) with an expected value of n
- e. Signed -rank statistic (S) with an expected value of  $2n$

10. The Mann Whitney U statistic measures the degree of . . . (one correct choice)

- a. Enfoldment/ separation between the two groups
- b. Difference in the medians between the two groups
- c. Difference in the means between the two groups
- d. Difference in spread (interquartile range) between the two groups
- e. None of the above

11. The effect size measure (i.e. clinical importance measure) associated with the Mann Whitney U statistic is the . . (one correct choice)

- a. Difference between the means in the two groups
- b. Difference between the medians in the two groups divided by the pooled standard deviation
- c. Difference between the modes in the two groups
- d. Difference between the medians in the two groups
- e. Difference between the means divided by the pooled standard deviation

## 9. Correlation

1. Correlation is a measure that makes use of a particular distribution, what is it? (one correct choice)
  - a. Normal
  - b. Exponential
  - c. Chi square ( $df=1$ )
  - d. Bivariate normal
  - e. Uniform
2. Correlation is often assessed by eye, which type of plot is usually used for this purpose? (one correct choice)
  - a. Histogram
  - b. Bar chart
  - c. Boxplot
  - d. Scatter plot
  - e. Funnel plot
3. Which of the following statements is true concerning correlation? (one correct choice)
  - a. A correlation is always between -2 and 2, a zero value indicates no clustering towards line
  - b. A correlation is always between -1 and 1, a zero value indicates all points on line
  - c. A correlation is always between -2 and 2, a zero value indicates all points on line
  - d. A correlation is always between -1 and 1, a zero value indicates no clustering towards line
  - e. A correlation is always between -1 and 1, a zero value indicates all points on a horizontal line
4. The correlation coefficient is based upon another measure, what is it? (one correct choice)
  - a. Variance
  - b. Co-relation
  - c. Contingency coefficient
  - d. Covariance
  - e. Cooks distance
5. The calculation of the confidence interval for the correlation coefficient is . . . ? (one correct choice)
  - a. No different from other statistics
  - b. More complex than usual because of the restricted range
  - c. Needs to be interpreted with extreme caution
  - d. Un-defined
  - e. Equivalent to the coefficient of determination
6. There are a number of effect size measures for the correlation coefficient. Which of the following is **not** considered to be one? (one correct choice)
  - a. Coefficient of determination ( $r^2$ )
  - b. Cohens d
  - c. Correlation coefficient
  - d. Cooks distance
  - e. Correlation coefficient squared
7. The coefficient of determination can be interpreted a number of ways. Which of the following is one of them? (one correct choice)
  - a. Proportion of explained variation
  - b. Proportion of unexplained variation (i.e. residual)
  - c. Proportion of mean variation
  - d. Proportion of variance variation
  - e. Proportion of points on the line

8. There is a special variety of the correlation coefficient used in the situation where the x and y values are interchangeable such as when comparing two measures, this intraclass correlation can be calculated easily by? (one correct choice)

- a. Appending the y scores to the x scores and then performing a standard correlation.
- b. Appending the y scores to the x scores and then performing a rank correlation
- a. Appending the y scores to the x scores and appending the x scores to the y ones then performing a standard correlation.
- b. Appending the y scores to the x scores and appending the x scores to the y ones then performing a rank correlation.
- c. Appending the y scores to the x scores and appending the x scores to the y ones then performing a paired *t* statistic .

9. Which is the most important assumption that is relaxed when considering Rank correlation compared to those for the Pearson correlation coefficient? (one correct choice)

- a. Linear relationship
- b. Normal distribution
- c. Observation pairs are independent
- d. Sample is randomly selected
- e. Data cannot be nominal

10. Which of the following statements concerning the correlation coefficient is **not** correct? (one correct choice)

- a. Correlation does not imply causation
- b. Usual correlation techniques only consider monotonic/linear associations
- c. Non-homogenous groups can affect the correlation
- d. A significant p value provides evidence that the population correlation is equal to that observed
- e. Correlation was originally developed by Sir Francis Galton

11. Which of the following provides the most accurate interpretation of a Pearson correlation coefficient of .733 ( $p=.0001$ )? (one correct choice)

- a. We are likely to observe a correlation of .733 given that the population correlation is equal to .773 around once in ten thousand times on average in the long run.
- b. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is **not** equal to zero around once in ten thousand times on average in the long run.
- c. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is equal to zero around once in a hundred times on average in the long run.
- d. We are likely to observe a correlation of .0001 or one more extreme given that the population correlation is equal to .733 in the long run.
- e. We are likely to observe a correlation of .733 or one more extreme given that the population correlation is equal to zero around once in ten thousand times on average in the long run.

# 10. Simple regression

1. The aim of simple linear regression is to? (one correct choice)

- a. Minimise the sum of vertical (y) errors (residuals), using the least squares method that creates model parameters ( $\alpha$ ,  $\beta$ ) that maximises the likelihood of the observed data.
- b. Minimise the sum of squared horizontal (x) errors (residuals), using the least squares method that creates model parameters ( $\alpha$ ,  $\beta$ ) that maximises the likelihood of the observed data.
- c. **Maximise** the sum of squared vertical (y) errors (residuals), using the least squares method that creates model parameters ( $\alpha$ ,  $\beta$ ) that maximises the likelihood of the observed data.
- d. Minimise the sum of squared vertical (y) errors (residuals), using the least squares method that creates model parameters ( $\alpha$ ,  $\beta$ ) that maximises the likelihood of the observed data.
- e. Minimise the sum of squared vertical (y) errors (residuals), using the least squares method that creates model parameters ( $\alpha$ ,  $\beta$ ) that minimises the likelihood of the observed data.

2. The dependent variable in simple linear regression is also called the? (one correct choice)

- a. Criterion or response or item
- b. Criterion or response or causal
- c. Criterion or response or explanatory
- d. Explanatory or predictor or independent
- e. Criterion or response or outcome

3. In the simple linear regression equation  $y = a + bx + e$  which of the following correctly describes the equation? (one correct choice)

- a.  $a$ =intercept,  $b$ =slope,  $e$ = random error with mean zero, unknown distribution
- b.  $a$ = slope,  $b$ =intercept,  $e$ = normally distributed random error with mean zero
- c.  $a$ =intercept,  $b$ =slope,  $e$ = normally distributed random error with mean zero
- d.  $a$ =intercept,  $b$ =slope,  $e$ = normally distributed random error with mean equal to mean of x variable
- e.  $a$ = slope,  $b$ =intercept,  $e$ = random error with mean zero, unknown distribution

4. The term 'simple' in simple linear regression is because? (one correct choice)

- a. There are no independent variables
- b. There is one independent variable
- c. There is more than one independent variable
- d. There are multiple dependent and independent variables
- e. The dependent variable is dichotomous

5. The **one parameter model** in simple linear regression attempts to? (one correct choice)

- a. Fit the data to the mean value of the dependent variable
- b. Fit the data to the mean value of the independent variable
- c. Fit the data within the 95% CI limit, by transforming the x values
- d. Fit the data, by transforming the x values to z scores
- e. Fit the data by using both intercept and slope parameters

6. In simple linear regression the total sum of squares is divided into two components, what are they? (one correct choice)

- a. Error and group
- b. Error and mean
- c. Error and correlational
- d. Error and regression
- e. Error and interaction

7. In simple linear regression the model is assessed by two methods, which happen to be equivalent in this case what are they? (one correct choice)

- a. Anova table (F ratio) and means for each parameter estimate
- b. Chi square and t statistics for each parameter estimate
- c. Chi square and t statistics for each parameter estimate
- d. Anova table (F ratio) and t statistic for first parameter estimate
- e. Anova table (F ratio) and t statistics for each parameter estimate

8. In simple linear regression it is possible to calculate two intervals along the line, one is the confidence interval (also called the mean prediction interval) and the other is the (individual) prediction interval. For a given tolerance level one is closer to the regression line and the other more distant, in which order are they? (one correct choice)

- a. Prediction interval closer; Confidence interval further away
- a. Confidence interval closer; prediction interval further away
- b. Confidence interval and prediction interval together (because same %)
- c. Confidence interval closer initially then crosses over at mean x value
- d. Prediction interval closer initially then crosses over at mean x value

9. In simple linear regression the model is assessed by various influence statistics. Which of the following is NOT a reason for using them? (one correct choice)

- a. Identify unduly influential points that affect the regression line
- b. Identify invalid points due to data entry error
- c. Identify points that you may wish to omit from a subsequent analysis
- d. Identify points that are the furthest away from the regression line
- e. Identify points that may belong to a subgroup

10. Simple linear regression has a number of sample data assumptions, what are they? (one correct choice)

- a. Linearity, Independence, Normality, Unequal variance
- b. Linearity, Independence, non-normality, Equal variance
- c. Linearity, Independence, Normality, Equal variance
- d. Linearity, Independence, Normality, Unequal range between x and y variables
- e. Linearity, Independence, Normality, Unequal variance

11. In Simple linear regression a process of regression diagnostics is carried out, for what two purposes is this undertaken, for the assessment of . . .? (**two** correct choices)

- a. Normality of residuals
- a. Normal distribution of independent variable
- b. Normal distribution of dependent variable
- c. Equal variance over y axis range
- d. Equal variance over x axis range
- e. Equal variance between independent and dependent variables
- f. Purpose not given above

12. While Simple linear regression can demonstrate a mathematical relationship between two variables, to demonstrate causality one needs to consider an additional set of criteria, by what name do these criteria go under? (one correct choice)

- a. Bradford-Hill criteria (1965)
- b. Bevan-Hill criteria (1965)
- c. Brewis-Hill criteria (1965)
- d. Banford-Hill criteria (1965)
- e. Barkley-Hill criteria (1965)

## 11. Proportions and Chi square

No MCQs in current run.

## 12. Risk, rates and odds

No MCQs in current run.

# 13. Survival analysis

1. Burton and Walls 1987 investigated the survival of patients on one of three types of renal replacement therapy, peritoneal dialysis, haemodialysis and transplantation details given opposite. What is the usual name for the exponential coefficient column? (one correct choice)

- a. Hazard Rate (HR)
- b. Hazard Ratio (HR)
- c. Hazard probability
- d. Hazard proportion
- e. Hazard logarithm

2. Considering the results from Burton and Walls 1987 given opposite. Which is the most appropriate way of interpreting the values in the exponential coefficient column (one correct choice)

- a. Odds
- b. Probability
- c. Time to event
- d. Proportion failing
- e. Odds ratio

Burton P R, Walls J 1987 Selection-adjusted comparison of life-expectancy of patients on continuous ambulatory peritoneal dialysis, haemodialysis, and renal transplantation		
Variables that significantly influenced probability of survival		
Variable	Exponential coefficient (risk multiplying factor)	Statistical significance
<b>Adverse</b>		
Age (each additional decade)	1.68	p<0.0001
Amyloidosis	8.26	p<0.0001
Acute or acute-on-chronic presentation	2.73	p<0.005
Ischaemic heart disease	1.65	p<0.025
Convulsions	3.17	p<0.03
<b>Beneficial:</b>		
Male sex	0.48	p<0.001
Parenthood	0.45	p<0.001
Pyelonephritis	0.48	p<0.02
Residence in Leicestershire	0.64	p<0.05

3. Considering the results from Burton and Walls 1987 given above. Which variable represents the greatest hazard (one correct choice)

- a. Age (in decades)
- b. Amyloidosis
- c. Convulsions
- d. Ischaemic heart disease
- e. Acute or acute on chronic presentation

4. Considering the results from Burton and Walls 1987 given above. Which variable represents the greatest benefit (one correct choice)

- a. Male sex
- b. Parenthood
- c. Pyelonephritis
- d. Residence in Leicestershire
- e. Absence of Ischaemic heart disease

5. Considering the results from Burton and Walls 1987 given above. If anyone were considering dropping a variable from the model which one would it most likely be? (one correct choice)

- a. Male sex
- b. Parenthood
- c. Pyelonephritis
- d. Residence in Leicestershire
- e. Absence of Ischaemic heart disease

6. Considering the results from Burton and Walls 1987 given above. What is the Exponential coefficient value likely going to be for the female sex? (one correct choice)

- a. 0
- b. .5
- c. 1
- d. 1- 0.48
- e. 1+ 0.48



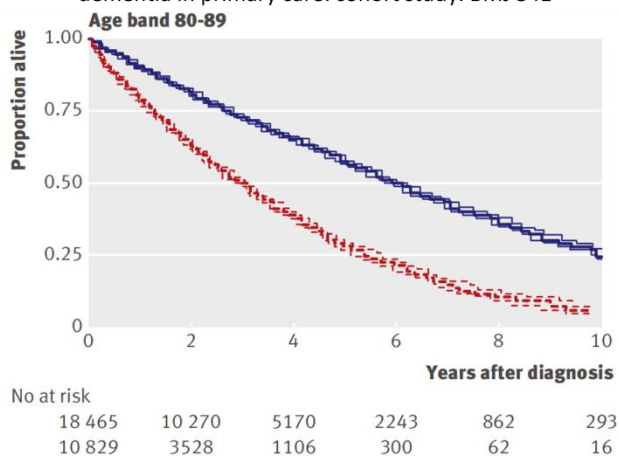
7. Considering the results from Rait et al 2010 given opposite. What is the more usual term for the x axis? (one correct choice)

- a. Survival function  $S(t)$
- b. Logit
- c. Inverse hazard
- d. Actuarial survival
- e. Proportion censored

8. Considering the results from Rait et al 2010 given opposite. The cohort detail below the x axis are? (one correct choice)

- a. Irrelevant and should not be shown
- b. Confuse the issues
- c. More important than the graph
- d. Provide useful additional information
- e. Can be calculated from the graph

Rait et al. 2010 Survival of people with clinical diagnosis of dementia in primary care: cohort study. BMJ 341



9. When gathering the failure times to calculate the Kaplan Meier plot which of the following statements is correct? (one correct choice)

- a. Its accurate measurement is of minimal importance
- b. Can be grouped into equal intervals
- c. Can be calculated from other measures
- d. Its accurate measurement is of major importance
- e. It is best to collect them at the end of the study period only

10. Censored observations do **not** include . . .? (one correct choice)

- a. Those who experience the event during the followup period of the study
- b. Those that are lost to followup
- c. Those that fail to provide event data
- d. Those subjects whose survival time is less than the followup period of the study
- e. Those who experience the event after the followup period of the study

11. Censored observations are . . .? (one correct choice)

- a. More important than non-censored ones in survival analysis
- b. Are assumed to be normally distributed over time
- c. Are assumed to have the same survival chances as uncensored observations
- d. Are essential to allow calculation of the Kaplan Meier plot
- e. Are allocated to the baseline survival curve

12. A Cox regression analysis . . .(one correct choice)

- a. Is used to analyse survival data when individuals in the study are followed for varying lengths of time.
- b. Can only be used when there are censored data
- c. Always assumes that the relative hazard for a particular variable is constant at all times
- d. Uses the logrank statistic to compare two survival curves
- e. Relies on the assumption that the explanatory variables (covariates) in the model are Normally distributed.

## 14. Hypotheses, Power and sample size

1. Within the R A Fisher approach which of the following is not true: (one correct choice)
  - a. P value =Probability of the observed data (statistic) or more extreme given that the null hypothesis is true
  - b. P value is interpreted on a individual experiment basis
  - c. The critical value is specific to a experiment
  - d. P value is interpreted as evidence against the null hypothesis, lower values greater strength of evidence
  - e. Decision rules form a major component in Fishers approach
  
2. Within the Neyman Pearson approach Alpha ( $\alpha$ ) is interpreted as: (one correct choice)
  - a. Probability of rejecting the null hypothesis assuming it is true
  - b. Probability of accepting the null hypothesis assuming it is true
  - c. Probability of rejecting the specific alternative hypothesis assuming it is true
  - d. Probability of accepting the specific alternative hypothesis assuming it is true
  - e. None of the above
  
3. Within the Neyman Pearson approach power is : (one correct choice)
  - a. Probability of rejecting the null hypothesis assuming it is true
  - a. Probability of accepting the null hypothesis assuming it is true
  - b. Probability of rejecting the specific alternative hypothesis assuming it is true
  - c. Probability of accepting the specific alternative hypothesis assuming it is true
  - d. None of the above
  
4. Statistical Power is affected by several factors which of the following is **false**: (one correct choice)
  - a. Effect size, increasing it, increases power
  - b. Sample size, increasing it, increases power
  - c. Type one error ( $\alpha$ ), increasing it, increases power
  - d. Type two error ( $\beta$ ), increasing it, increases power
  - e. Variance decreasing it, increases power
  
5. Statistical Power when considering the Null and specific alternative pdfs is graphically: (one correct choice)
  - a. Area of the  $h_1$  pdf covering the values not in the critical region in the  $h_0$  pdf
  - b. Area of the  $h_0$  pdf covering the same x values as the critical region in the  $h_0$  pdf
  - c. Area of the  $h_0$  pdf covering the values not in the critical region in the  $h_0$  pdf
  - d. Area of the  $h_1$  pdf covering the same x values as the critical region in the  $h_0$  pdf
  - e. Value of the  $h_1$  pdf at the critical value of the  $h_0$  pdf

6. Analysis of Statistical Power can be undertaken after an investigation in certain circumstances. Which of the following is one of them? (one correct choice)

- a. Failure to achieve a significant (i.e. P value within critical region) result
- b. Obtaining a significant (i.e. P value outside critical region) result
- c. Failure to achieve expected sample size
- d. Effect size measure greater than that expected
- e. Greater loss to followup than expected

7. An original investigation reports a Power of 0.5, the investigator then recruits approximately 30% more subjects and carries out an analysis on the larger dataset. What are the chances of her obtaining a significant result (i.e. p value in critical region) if the specific alternative hypothesis is assumed to be true now? (one correct choice)

- a. 0%
- b. Below 50%
- c. 50% approximately
- d. Above 50%
- e. (50 + 30)%

8. Which of the following applications is most frequently used to carry out a power analysis? (one correct choice)

- a. SPSS
- b. Epi Info
- c. Gpower
- d. Excel
- e. Word

9. When carrying out a Statistical Power analysis a graph of the following variables is most frequently produced: (one correct choice)

- a. Effect size, power
- b. Sample size, power
- c. P value, power
- d. Effect size,  $\alpha$
- e.  $\beta$ ,  $\alpha$

10. When carrying out a Statistical Power analysis which one of the following statements is NOT correct: (one correct choice)

- a. A power analysis can indicate the minimum required sample for a given effect size
- b. A power analysis can indicate the expected p value for a given effect size
- c. A power analysis can be undertaken for both parametric and non parametric tests
- d. A power analysis carried out after the investigation is only appropriate in specific circumstances
- e. A power analysis before a investigation can provide important information

11. Which of the following statements is the most accurate interpretation of a p value of 0.036: (one correct choice)

- a. We are 3.6% sure that the null hypothesis is true
- b. We are 96.4% (i.e. 1-.036) sure that the alternative hypothesis is true
- c. We are 3.6% sure that the null hypothesis is true given our observed data
- d. We are 3.6% sure that the our observed data are the result of random sampling
- e. We will observe data or that more extreme 3.6% of the time in the long run when the null hypothesis is true.

12. Which of the following statements represents the transposed conditional incorrect interpretation of the p value, where D = observed data or those more extreme;  $h_0$  = null hypothesis;  $h_1$  = alternative hypothesis: (one correct choice)

- a.  $p(D | h_0)$
- b.  $p(h_0 | D)$
- c.  $p(h_1 | D)$
- d.  $p(D | h_1)$
- e.  $p(h_0)$

13. Which of the following statements most accurately describes the size fallacy concerning the p value: (one correct choice)

- a. Interpreting a p value as a probability of support for the alternative hypothesis
- b. Interpreting a p value as the size of alpha
- c. Interpreting a p value as the size of beta
- d. Interpreting a p value as a probability of support for the null hypothesis
- a. Interpreting a p value as a measure of effect size

## 15. Simple logistic regression

Chapter contains MCQs but these will not form part of the time online MCQ for this run.

End of document