**Main Title**

* Hello test
  + Point 1
  + Point 2

**Reading for Lectures 1A and 1B: Kirkwood and Sterne Chp1-8**

Chapter 1: Using this book

* Meta-analysis: statistical methods for combining the results of a number of studies

Chapter 2: Defining the data

* Sampling variation: difference across samples due to chance
* Sample size: number of individuals
* Variable: any aspect of an individual that is measured or recorded
  + Numerical/quantitative: either continuous (weight) or discrete (limited number of values)
  + Categorical/qualitative: non-numeric and maybe binary/dichotomous (only 2 values) or ordered categorical (have natural ordering like social class) or just categorical
  + Rates: over time
* Derived variables: derived from those originally recorded, can be categorical (eg. based on threshold values of a measured variable) or numerically derived (BMI) or based off a reference curve that is based on standard population values
* Transformed variables: change of scale perhaps to satisfy assumptions needed for statistical method (eg. Logarithmic transformation)
* Outcome vs Exposure variables: Outcome as the event that we want to understand aka dependent/response/y/case-control while exposure as factors that influence the size/occurrence of outcome variable aka explanatory/independent/x/risk factor/treatment group 🡨 trtment group used when clinical trial vs risk factor when case-control study
  + Statistical analysis used to quantify the magnitude of association between ≥1 exposure variable and the outcome variable
  + Type of outcome variable is impt in determining the most appropriate statistical method

Chapter 3: Displaying the data

* Frequencies: usually used for categorical variables where count no. of observations per category. Can also be presented as relative frequencies that are proportionate or percentages of total population size
  + Visualised via bar chart/diagram or pie chart where length/size references the frequency size, and frequency distribution for numerical variables or for population, histograms, frequency polygon (used when comparing ≥2 frequency distributions on the same figure)
    - General rule for histogram: if intervals are not the same width, make the heights of the rectangles proportional to the frequencies divided by the widths so that the areas are proportional to the frequencies (eg x:14-15, y:5 and x:15-17, y:1; when making x:14-16, instead of summing y to 6, make it 3 as it’s the sum of frequencies for that range divided by the width of two intervals)
    - Frequency polygon method: drawn by indicating the y value for each x (like a histogram) and connecting the midpoints of the tops of the rectangles with free lines. The endpoints of the resulting line are then joined to the horizontal x axis at the midpoints of the groups that are immediately above and below the highest and lowest non-zero frequencies
  + Frequency distribution: If >20 observations, numerical variables are summarized into a frequency distribution (a table showing no. of observations at different values or within certain ranges). If discrete variable, frequencies may be tabulated for either each value or groups of values. If continuous variable, groups have to be formed.
    - Method: Count the number of observations and identify highest and lowest values. Then decide if data should be grouped (and what interval to use if yes) 🡨 should aim for 5-20 groups as a rough guide depending on number of observations
    - Confidence for frequency distribution depends on number of individuals where larger sample = finer grouping interval that is chosen = smoother histogram/frequency polygon = more closely resembles the population distribution
    - Has upper and lower tails and three main shapes of symmetrical aka bell-shaped, or asymmetrical aka skewed either to the right (positively) or the left (negatively). A single peak implies it’s a unimodal distribution whereas bimodal has two peaks (usually suggests data is a mix of two separate distributions) like for hormone levels of males and females. Other distributions include reverse j-shaped and the uniform distribution.



* Cumulative frequency distributions: summation over groups until reaches full sample size/percentage and is usually drawn as a step function (vertical jumps correspond to increases in cumulative % observed like in Kaplan-Meier plots of cumulative survival probabilities over time). An even frequency distribution = increasing at a constant rate cumulative distribution.
  + Advantage is that they display the shape of distributions without need for grouping (needed in histograms) but shape is less clearly seen here
  + Visualised via median and quartiles (divide distribution into evenly sized groups aka 0, 25, 50, 75, 100), as well as quantiles and percentiles (equal sized divisions where tertiles: divide into 3 groups and quintiles: divide into 5 groups and deciles: into 10 groups)
    - Median: midway value aka half the distribution lies above and below it = (n+1)/2 where n is number of observations that are ordered
    - Chart, box and whisker chart

      Description automatically generatedLower quartile: (n+1)/4 aka 25th percentile of ordered observations where n is no. of observation
    - Upper quartile: 3(n+1)/4 aka 75th percentile of ordered observations where n is no. of observation
    - Range: highest value – lowest value



* + - IQR (Interquartile range): upper quartile – lower quartile (used in box and whiskers plot)
      * Box and whiskers plot: box is drawn from lower quartile to upper quartile with its length giving the IQR. Horizontal line represents median. Whiskers are drawn on either end of box to indicate max and min values



* + - * + Can also be used to display relationship between numerical and discrete categorical variables
    - Kth percentile: point below which k% of distribution values lie = k(n+1)/100
* Between two variables, association seen via cross tabulations in a contingency table (rows correspond to exposure vals aka x and columns correspond to outcomes aka y) and scatter plots (positive/negative relationship between two numeric variables)
  + Contingency plots can be further improved via marginal totals and percentages inclusion, where percentages should correspond to exposure variable.
  + Scatter plots may have overlapping points, reducing interpretability. Solution is jittering where points are scattered randomly across the horizontal axis (when x axis has categorical values).
* Time trends: visualized using graphs that show absolute changes over time unless a logarithmic scale is used (useful when comparing rates of progress between absolute regions)
  + Breaks and discontinuities should be clearly marked on graphs and avoided

Chapter 4: Means, standard deviations and standard errors

* Mean, median and mode: used to summarise a numerical variable and present the average value
  + Mean: Σx/n = sum of values divided by number of values aka arithmetic mean. Is the preferred measure as it accounts for each individual observation and is most useful for statistical analysis. If distribution is positively skewed (to the right), a geometric mean is more appropriate.
  + Median: (n+1)/2 = divides distribution by half. If there is an even no. of observations, there is no middle one and the average of the two middle ones is denoted as the median. Is useful as a descriptive measure if there are some extreme outliers which make the mean unrepresentative of the data.
  + Mode: value which occurs most often. If no value repeats, there is no estimate of the mode. Seldom used as may be misleading especially when sample size is small.
* Measures of variation: gives spread of values
  + Range and IQR: Range based on only two values, giving no insight into distribution between these two values and tends to be larger given a larger sample size aka highly sensitive to sample size. IQR as less sensitive to sample size (assuming sample size is not too small), with lower and upper quartiles as more stable than the extreme range end values
  + Variance s2: Σ(x-mean)2 / (n-1) and uses all observations, larger if scattered over considerable distance and small if bunched closely about mean. Cannot average deviation as will always =0, with positive (above mean) and negative deviations (below mean) balancing out 🡨 can average the sizes of deviation while ignoring their sign but not mathematically malleable/controllable so we average the squares of the deviations.
    - Degrees of freedom: (n-1) rather than n as gives better estimate of variance, with only n-1 of the deviations (x – mean) are independent of each other, with the last one being calculable from the others as all n of them add up to 0.
    - Standard Deviation (sd) s: sqrt(Σ(x-mean)2 / (n-1)) = sqrt(Σx2 – [(Σx)2/n] /(n-1)) Counters disadvantage of variance being measured in square of units used by expressing the square root of variance, causing variation to be in original units again. Usually 70% of observations lie within 1sd while 95% lie within 2 sd, based on the normal distribution.
      * Adding/subtracting a constant from the observations changes the mean by the same amount that is changed but standard deviation is constant,
      * Multiplying/dividing a constant from the observations changes both the mean and the standard deviation in the same amount through multiplication/division.
    - Coefficient of variation cv: s/mean\*100 expresses the standard deviation as a percentage o the sample mean, and is useful when size of the variation relative to size of the observation is of interest. Advantageous that it is independent of units of observation.
* Sampling Error (stopped at page 38): the sample mean is unlikely to be exactly = to population mean since a different sample would give a different estimate due to sampling variation. This sampling error is the standard error of the sample mean = population sd / sqrt(n) and measures how precisely the population mean is estimated by the sample mean, with the standard error size depending on both size of sample (inversely) and variation within sample (directly)
  + Point 1
  + Point 2

Chapter 5

* Normal distribution
  + Point 1
  + Point 2
* Sampling distribution
  + Point 1
  + Point 2

Chapter 6

* Confidence intervals for single variable:
  + Point 1
  + Point 2

Chapter 7

* Confidence intervals when comparing two groups:
* P-values
  + Point 1
  + Point 2

Chapter 8

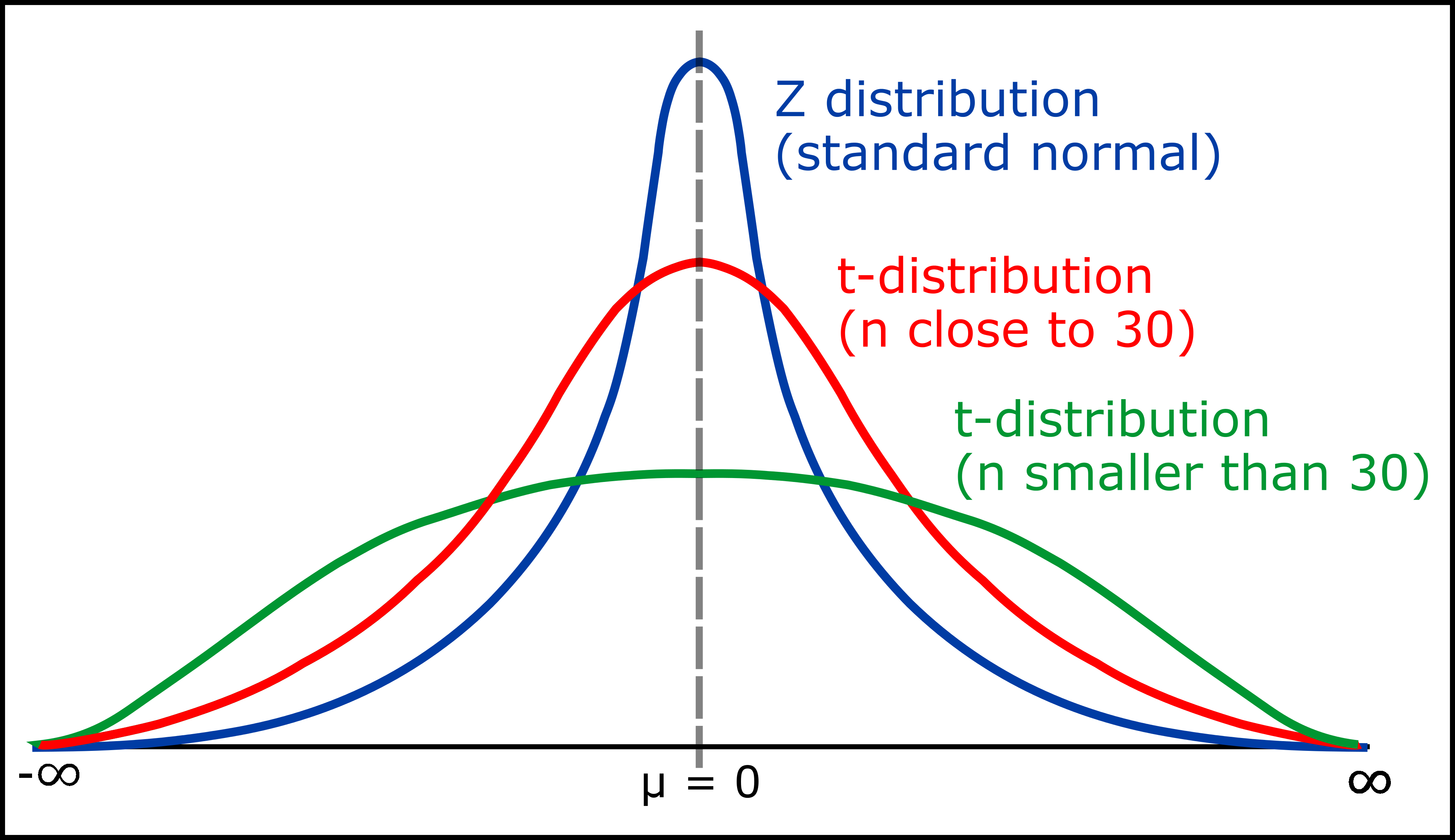
* P-values
  + Point 1
  + Point 2

**Lecture 1A: Principles of Inference, Sampling, Normal Distribution**

* Principles of Inference:
  + Point 1
  + Point 2
* Sampling:
  + Point 1
  + Point 2
* Normal distribution:
  + Point 1
  + Point 2
* Standard deviation or variance of sample is measure of amount of variation in our population
* Standard error is the measure of variability of sample mean; precision estimate

**Lecture 1B: Confidence intervals, Hypothesis testing, and p-values**

* Confidence interval (Chp 6): range of values within which size of association in population is likely to lie, accounting for sampling variation and standard error (inherent variation between individuals)
  + A 95% CI suggests that 95% of the time, we expect the confidence interval to contain the population mean; CANNOT say that there is a 95% probability that the CI contains the true mean since the true mean is not a random but fixed and what is random is the bounds of the interval calculated off of the true mean and probability is used when there is randomness present.
* Hypothesis testing (Chp 7) aka significance tests: used to assess evidence strength against the null hypothesis that there is no true association in population
  + Methods include t-tests or chi-squared tests (when comparing two exposure groups), regression models (examining effect of several exposure variables)
* P-values (Chp 7):



**Reading for Applied Lab 1: Kirkwood and Sterne Chp38**

**Reading for Lectures 2A and 2B: Kirkwood and Sterne Chp9-10,12**

Chapter 9: Comparison of means from several groups; analysis of variance

Used when the exposure variable has >2 categories

* One way analysis of variance: used when subgroups to be compared are defined by one exposure (eg ethnic group or socioeconomic group).Thus, the mean of a numerical outcome variable is compared where the exposure group is classified by just one variable, examining how much of the overall variation in the outcome variable can be attributed to difference between the exposure group means.
  + Sum of squares (SS = sum(x-mean)^2 = sum(x^2) – (sum x)^2/n) as part of the variation can be due to either difference between group means or due to differences between observations within each group aka residual sum of squares. Assuming there are k groups, total degrees of freedom (n-1) where n is no. of observations is divided into the between groups SS (k-1 df) = sum n\*(x-mean)^2 = sum n\*(mean xi^2) – (sum x)^2/n and residual SS or within groups SS (n-k df) = sum si^2\*(ni – 1).
  + Mean square: MS = sum of squares/df that gives the amount of variation per degree of freedom. 🡨 MS is important as comparison of between group and within group MS is used to determine if the mean outcome differs between exposure groups; if variation is due to chance, variation between group means is about the same as those between individuals aka within group. If variation is due to real difference, the between groups variation > within group variation. MS is compared using the variance-ratio or F test where F = between group MS/within group MS and df = between group df, within group df = k-1, n-k. 🡨 F distribution (unlike most distributions) is specified by a PAIR of df with k-1 in numerator and n-k in denominator. F test is usually reported with p value! Small p val = evidence to reject null hypothesis. F test also assumes that (1) outcome is normally distributed and that (2) population value of standard deviation between individuals is the same in each exposure group, estimated by sqrt(within group MS).
    - F about 1: no real differences between groups
    - F>1: there are differences
  + When there are only 2 groups, one way analysis of variance gives exactly same result as t test AKA F stat with (1, n-2 df) == **t stat^2** (n-2 df).
* Two way analysis of variance: used when subdivision is based on **two factors** (such as age AND sex)

If equal no. of observations within each group == balanced design. Else, unbalanced. Balanced designs can be with replication (>1 observation in each group) and without replication.

* + Balanced with replication aka where every group has the same number of observations/individuals: Total sum of squares is divided into FOUR components of (1) main effect of **factor1** where df = no. of factor – 1, (2) main effect of the factor2 where df = no. of factor2 -1, (3) SS due to **interaction between** the two factors where df = (df of factor1)\*(df of factor2) and (4) residual SS due to differences between individuals in each factor-factor group where df = no. of factor 1\*no. of factor 2\*(n in each group – 1). F test is used to compare mean squares of factors with residual mean square where F = MS effect (aka from each individual factor and from factor-factor interaction)/MS residual. 🡨 Thus, F only exists for both factors and inter factor interactions.
  + Balanced without replication: (page 85)
  + Unbalanced:
* Fixed and random effects
  + Point 1

Chapter 10: Linear regression and correlation

Simple linear regression: only one exposure variable is considered; y = mx + c where only one x

* Linear regression: equation used is y = beta0 + beta1\*x aka c + mx where m is gradient, found by change y/change x aka rise/run and c is the y intercept where x=0.
  + Parameters are estimated via least squares: m=beta1=sum[(x-xbar)(y-ybar)]/sum[(x-xbar)^2] and c=beta0=ybar – beta1\*xbar.
  + Sampling variation: Calculated values are estimates of the population values of the intercept and slope and are thus subject to sampling variation. Thus, precision is measured via standard errors. Standard error wrt gradient beta1 = s/sqrt( sum(x-xbar)^2 ) and standard error wrt intercept beta0 = s\*sqrt( 1/n + xbar^2/sum(x-xbar)^2 ). Standard deviation of the points about the line = s = sqrt( [sum(y-ybar)^2 – beta1^2 \* sum(x-xbar)^2] / (n-2) ). Thus, standard deviation has n-2 df aka sample size – no. of parameters for regression equation (2)
* Correlation
* Using variance approach for simple linear regression
* Correlation coefficient and variance analysis

Chapter 12: Goodness of fit and regression diagnostics

Used when the