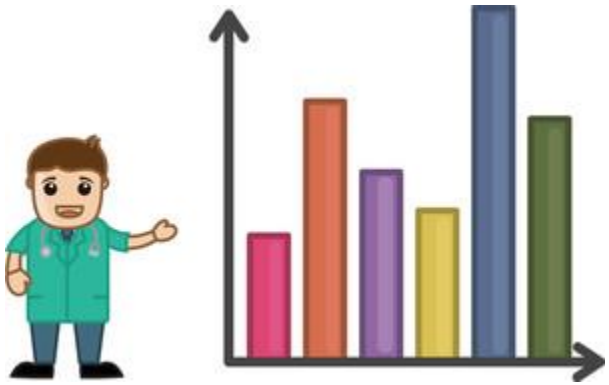


Medical Statistics

Professor Mark E. O'Donnell

DipSEM(GB&I) MB BCH BAO(Dist) MFSEM(UK) MFSEM(RCSI&RCPI) MFSTEd MMedSc(Dist) MD ECFMG RPVI(ARDMS) FCPhleb FRCSEd(Gen&VascSurg) FEBVS(Hon)

Consultant Vascular and Endovascular Surgeon, Belfast Health and Social Care Trust and Visiting Professor, Sport and Exercise Sciences Research Institute, Faculty of Life and Health Sciences, Ulster University, Northern Ireland.

Northern Ireland Vascular Trainee Teaching Program – 6th December 2017

Royal Victoria Hospital


53


53


LONDON
itv


LONDON
DNSDALE
itv

Learning Outcomes

- Basic Statistical Terminology.
- Biostatistics.
- Statistical Analyses.

Basic Statistical Terminology

- Mean, mode and median

Basic Statistical Terminology

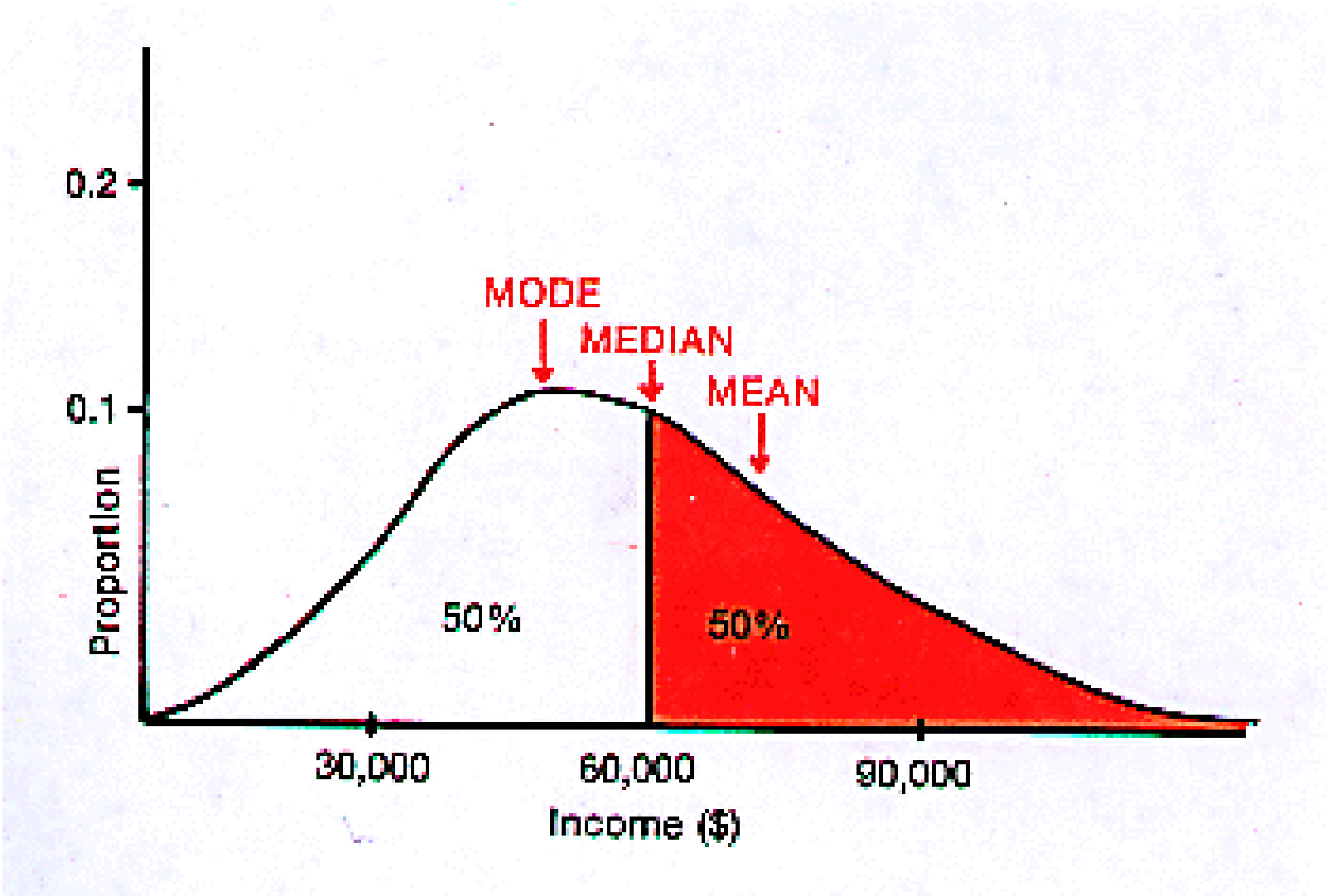
- Mean, mode and median are three kinds of "averages".

Basic Statistical Terminology

- Mean:
  - The "mean" is the "average" you're used to. You add up all the numbers (scores) and then divide by the number of variables (participants).

- Mode:
  - The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

- Median:
  - The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order, so you may have to rewrite your list first.


0.2
0.1
Proportion
MODE
MEDIAN
MEAN
50%
50%
30,000
60,000
90,000
Income ($)

Basic Statistical Terminology

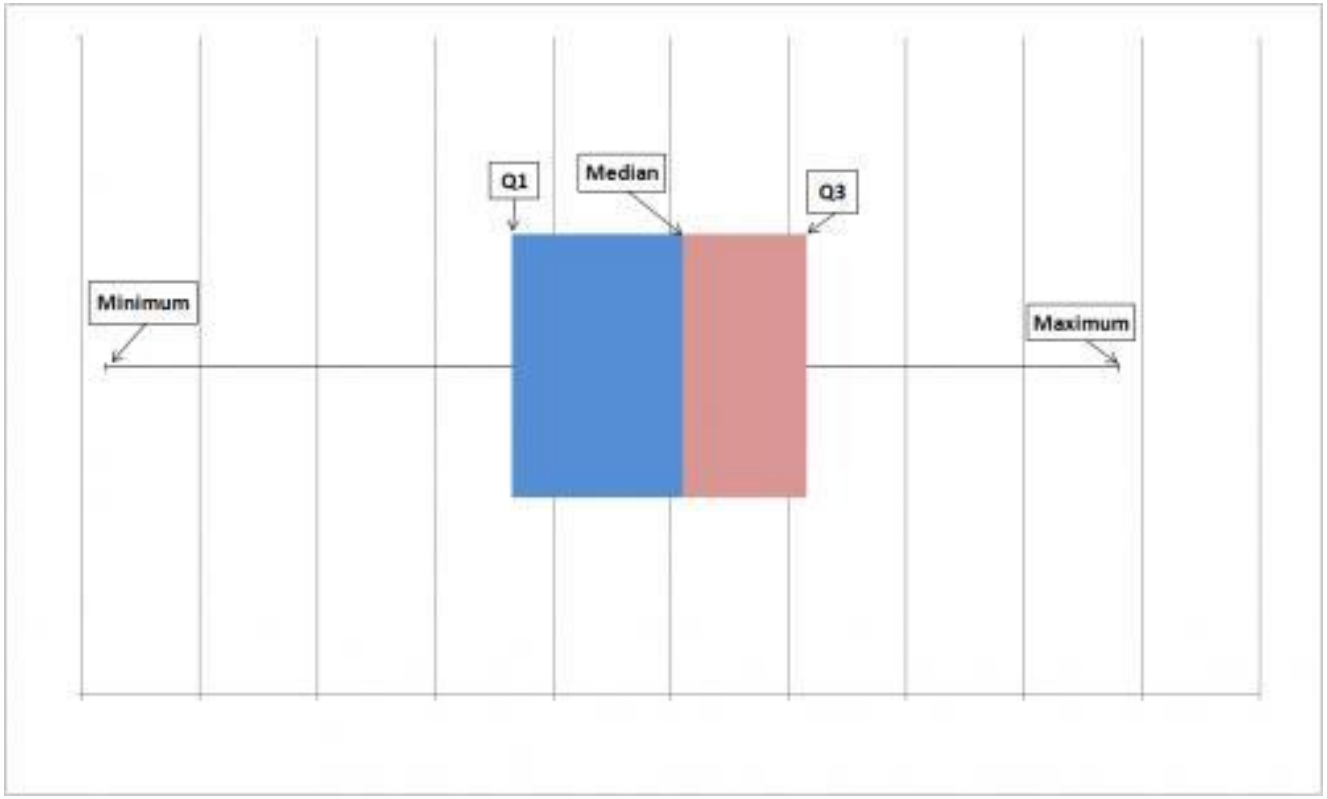
- Range:

- Interquartile Range:

Basic Statistical Terminology

- Range:
  - The "range" is just the difference between the largest and smallest values.

- Interquartile Range:
  - Middle fifty equal to the difference between the upper (75^{th}) and lower (25%) quartile.
  - Is a measure of statistical dispersion and is the most robust measure of scale.


Q1
Median
Q3
Minimum
Maximum

Biostatistics

- Incidence / Prevalence.
- Hypotheses.
- Statistical Significance / Confidence Intervals.
- Study Power / Errors.
- Sensitivity / Specificity.
- Predictive Values / Relative Risk / Odds Ratio.

Incidence

Incidence

- Defined as the number of **new** cases of a given disease per year:
  - Four cases of HIV per year.

Prevalence

Prevalence

- Defined as the total number of **existing** cases of a given disease in the entire population:
  - 20 people have HIV (right now).

Hypotheses

- Null Hypothesis (H_0):

Hypotheses

- Null Hypothesis (H_0):
  - A type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations.
  - The null hypothesis attempts to show that no variation exists between variables or that a single variable is no different than zero.

Hypotheses

- Null Hypothesis (H_0):
  - There is no significant difference between specified populations where any observed difference being due to sampling or experimental error.
  - It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.

Statistical Significance

Statistical Significance

- The p-value expresses the likelihood that an observed outcome was due to random chance.

- A p-value <0.05 is generally accepted as indicating that an outcome is statistically significant and not due to chance.

Statistical Significance

- The p-value expresses the likelihood that an observed outcome was due to random chance.
- A p-value <0.05 is generally accepted as indicating that an outcome is statistically significant and not due to chance.
- This means that you expect less than 5% of the observed outcomes to be due to chance. Therefore a difference in outcome is most likely related to a variance or influence on the study population such as a treatment.

Statistical Significance

- The p-value expresses the likelihood that an observed outcome was due to random chance.
- A p-value <0.05 is generally accepted as indicating that an outcome is statistically significant and not due to chance.
- If the p-value is greater than 0.05 then it can be assumed that any difference in outcome between the study populations is a chance event and not related to the treatment.

Statistical Significance

- Before you run any statistical test, you must first determine your α level, which is also called the "significance level."
- By definition, the α level is the probability of rejecting the null hypothesis when the null hypothesis is true.
- It's the probability of making a wrong decision.

How a p-value affects the Null hypothesis

- Null Hypothesis (H_0):
  - There is no significant difference between specified populations where any observed difference being due to sampling or experimental error.

How a p-value affects the Null hypotheses

- Null Hypothesis (H_0):
  - There is no significant difference between specified populations where any observed difference being due to sampling or experimental error.
  - A p-value <0.05 will therefore reject the null hypothesis as it suggests that a treatment effect rather than a chance event has effected a change or difference in the population.
  - The alternative hypothesis (H_1) will therefore be true meaning that the treatment has made the difference.

Confidence Intervals

Confidence Levels

- Confidence intervals are constructed at a confidence level selected by the used such as 95 %.
- This means that if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases.
- A confidence interval stated at a $1-\alpha$ level can be thought of as the inverse of a significance level α.

Confidence Intervals

- The purpose of taking a random sample from a lot of the population and computing a statistic, such as the mean from the data, is to approximate the mean of the population.
- How well the sample statistic estimates the underlying population value is always an issue.
- A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

Confidence Intervals

- Similar in concept to p-value, the confidence interval expresses the certainty that the observation is real or a product of random chance.

- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.

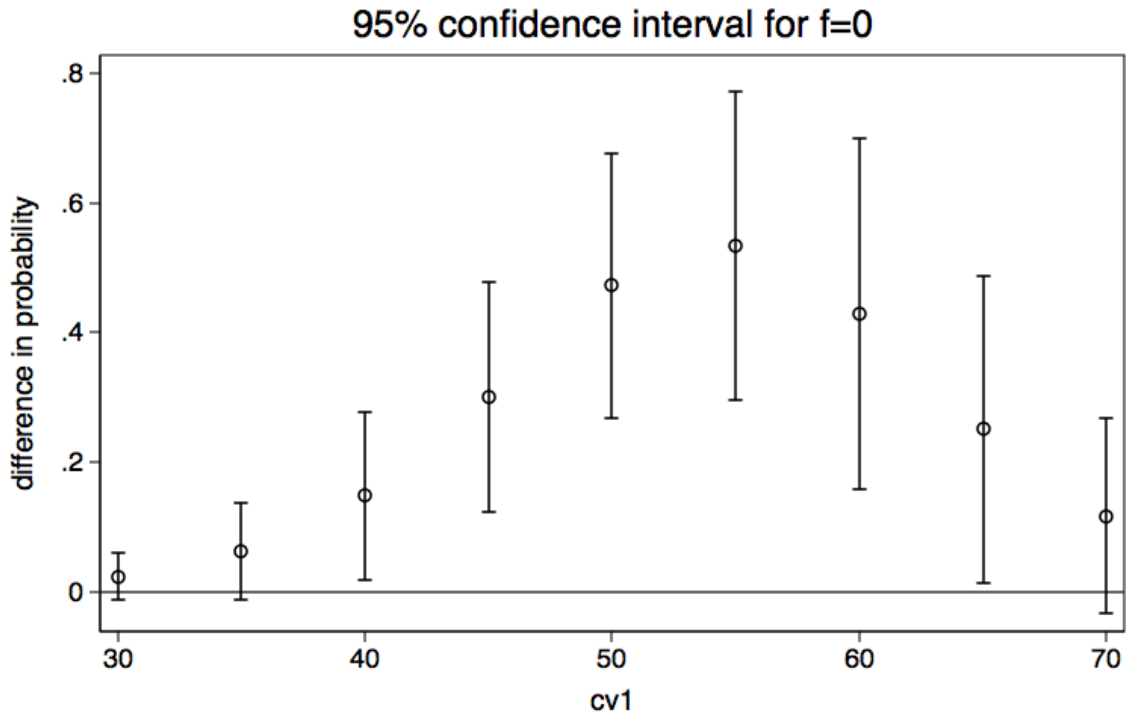
Confidence Intervals

- Confidence intervals are usually calculated so that this percentage is 95%, but we can produce 90%, 99%, 99.9% (or whatever) confidence intervals for the unknown parameter.

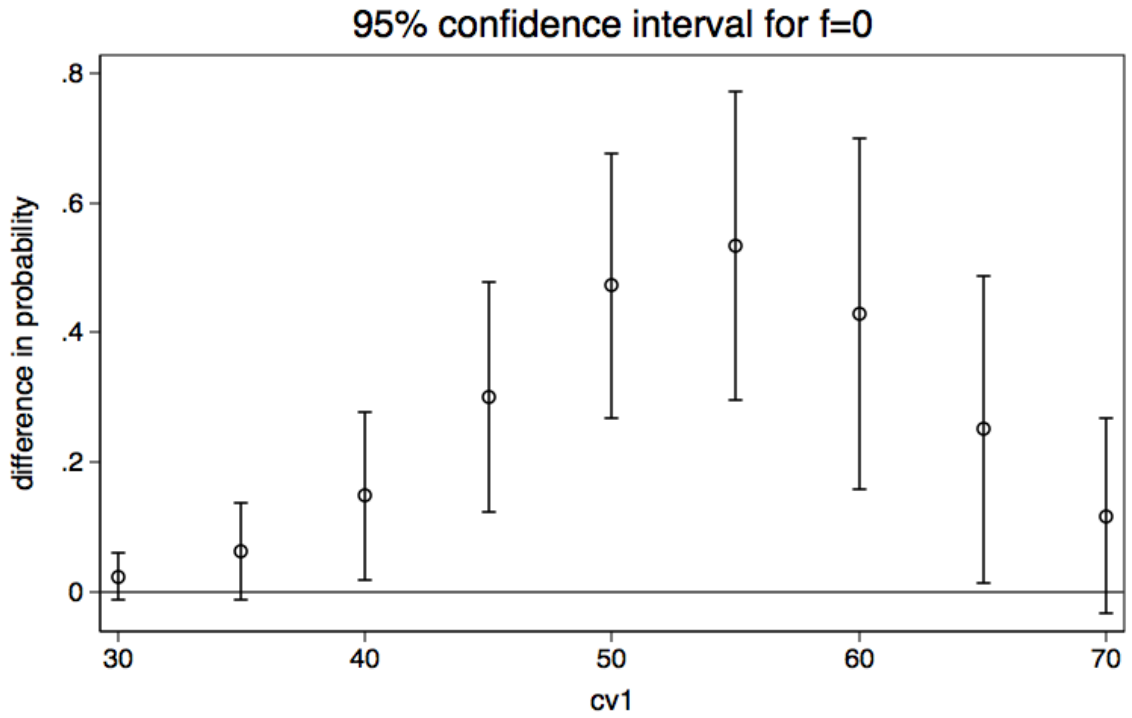
- The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter. A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.

- Confidence intervals are more informative than the simple results of hypothesis tests (where we decide "reject H0" or "don't reject H0") since they provide a range of plausible values for the unknown parameter.

Confidence Intervals



Confidence Intervals



If a 95% confidence interval crosses 0 - it indicates that the results are not significant.

Study Power

Study Power

- The power or sensitivity of a statistical test is the probability that it correctly rejects the null hypothesis (H_0) when it is false.

- It can be equivalently thought of as the probability of correctly accepting the alternative hypothesis (H_1) when it is true - that is, the ability of a test to detect an effect, if the effect actually exists.

Study Power

- Statistical power may depend on a number of factors;
  - the statistical significance criterion used in the test.
  - the magnitude of the effect of interest in the population.
  - the sample size used to detect the effect.

Statistical Errors

Statistical Errors

- Type I:
  - Incorrect rejection of a true null hypothesis (a "false positive").
  - Is detecting an effect that is not present.

- Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't.
  - A test that shows a patient to have a disease when in fact the patient does not have the disease.
  - An experiment indicating that a medical treatment should cure a disease when in fact it does not.

Statistical Errors

- Type II:
  - Failure to reject a false null hypothesis (a "false negative").
  - Is failing to detect an effect that is present.
  - A blood test failing to detect the disease it was designed to detect, in a patient who really has the disease.
  - A clinical trial of a medical treatment failing to show that the treatment works when really it does.

Study Power

- As the power increases, the chances of a Type II error (false negative), termed false negative rate (β) decrease.

- Power = 1−β.

Study Power

- As the power increases, the chances of a Type II error (false negative), termed false negative rate (β) decrease.
- Power = 1−β.
- Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size.
- Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size.

Sensitivity

Sensitivity

- The probability that a person with a disease will have a positive result on a given test.

- Measure of the “**TRUE POSITIVE**”.

Sensitivity

- The probability that a person with a disease will have a positive result on a given test.
- Measure of the "**TRUE POSITIVE**".
- The higher the sensitivity;
  - Higher the true positive rate.
  - Lower the rate of false negative.
- A highly sensitive test is useful as a screening test as the goal is to identify everyone with a given disease.

Specificity

Specificity

- The probability that a person without the disease will have a negative result.

- Measure of the “**TRUE NEGATIVE**”.

Specificity

- The probability that a person without the disease will have a negative result.
- Measure of the "**TRUE NEGATIVE**".
- The higher the specificity;
  - Higher the true negative rate.
  - Lower the rate of false positive.
- A highly specific test is desirable as a confirmatory test.
  - E.g.. HIV ELISA test is confirmed with a highly specific Western Blot test.

Summary – Statistical Errors

- Type I:
  - False Positive.
  - The higher the specificity, the lower the false positive rate (and higher the true negative rate).

- Type II:
  - False Negative.
  - The higher the sensitivity, the lower the false negative rate (and higher the true positive rate).

Beware

- It is rare to have both a highly sensitive and highly specific test.
- A test that is highly sensitive but not specific will yield many false positive.
- A test that is highly specific but not sensitive will yield many false negative.

SPIN SNOUT

- The higher the specificity, the lower the false positive rate – this means that each positive result is more likely to be true. Therefore, a highly specific test is good to "Rule-In" a disease process.

SPIN SNOUT

- The higher the specificity, the lower the false positive rate – this means that each positive result is more likely to be true. Therefore, a highly specific test is good to "Rule-In" a disease process.

- The higher the sensitivity, the lower the false negative rate – this means that each negative test is more likely to be negative. Therefore, a highly sensitive test is good to "Rule-Out" a disease process.

Predictive Values

Predictive Values

- Positive Predictive Value:
  - Probability that a person with a positive test result has the disease.
  - True Positives / All Positives.
- If a disease has a greater prevalence then the PPV will be higher.

Predictive Values

- Negative Predictive Value:
  - Probability that a person with a negative result is disease free.
  - True Negatives / All Negatives.
- A test has a higher NPV when a disease has a lower prevalence.

TABLE 6-1. Determination of PPV and NPV

	Disease Present	No Disease	
Positive test	a	b	PPV = a/(a + b)
Negative test	c	d	NPV = d/(c + d)
	Sensitivity = a/(a + c)	Specificity = d/(b + d)	

Relative Risk

Relative Risk

- Used to evaluate the results of prospective studies.
- Compares the incidence of a disease in a group exposed to a particular risk factor with the incidence in those not exposed to the risk factor.
  - RR < 1 means that the event is less likely in the exposed group.
  - RR > 1 signifies that the event is more likely in that group.

Odds Ratio

Odds Ratio

- Used in retrospective studies.
- Compares the rate of exposure among those with and without a disease.
  - Considered less accurate than relative risk.
  - However, in rare diseases the OR approximates the RR.

TABLE 6-2. Determination of RR and OR

	Disease Develops	No Disease	
Exposure	a	b	RR = [a/(a + b)]/[c/(c + d)]
No exposure	c	d	OR = ad/bc

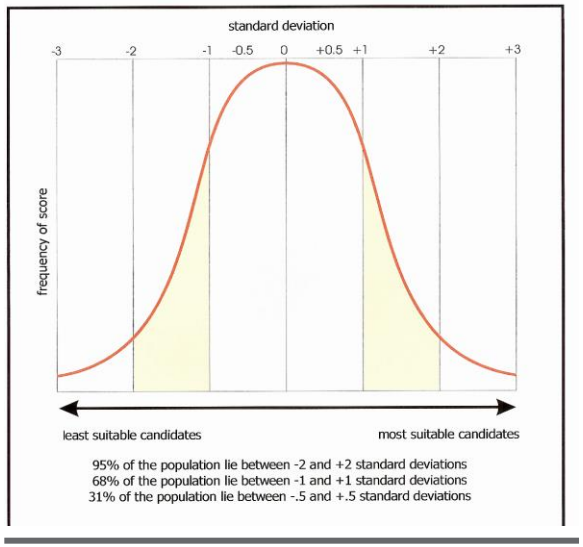
Basic Statistical Analyses

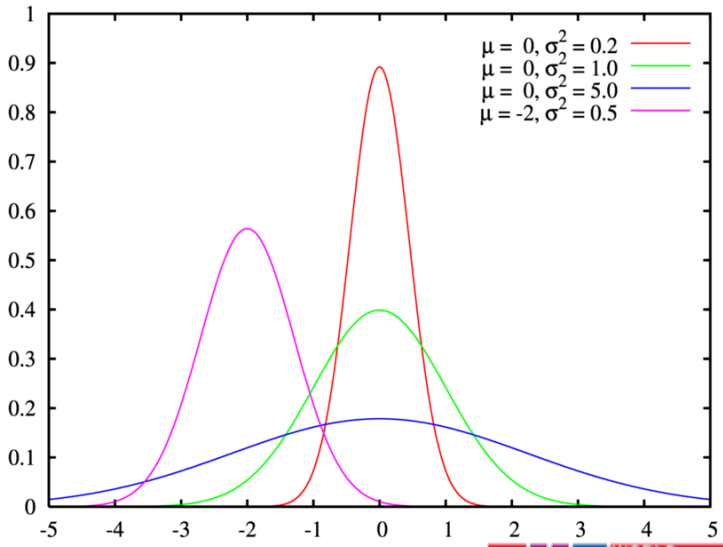
- Parametric vs. Non-parametric.
- Assessment of different variables.
- Assessment of variables with multiple outcomes.
- Correlation.
- Survival analysis.

Parametric vs. Non-Parametric

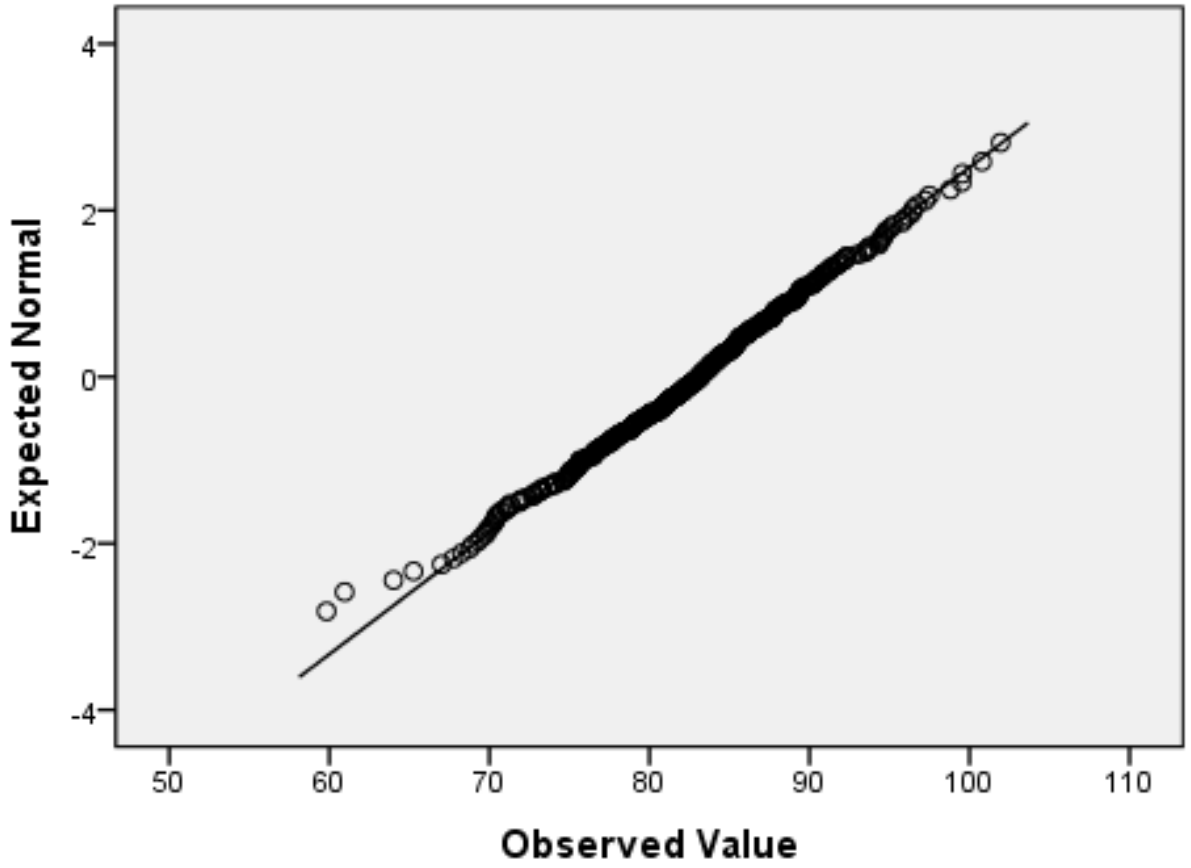
Parametric vs. Non-Parametric

- **Parametric:**
  - Normally distributed data.
  - Larger patient studies.
  - Higher statistical strength.
  - Independent T-Test / Paired T-Test.



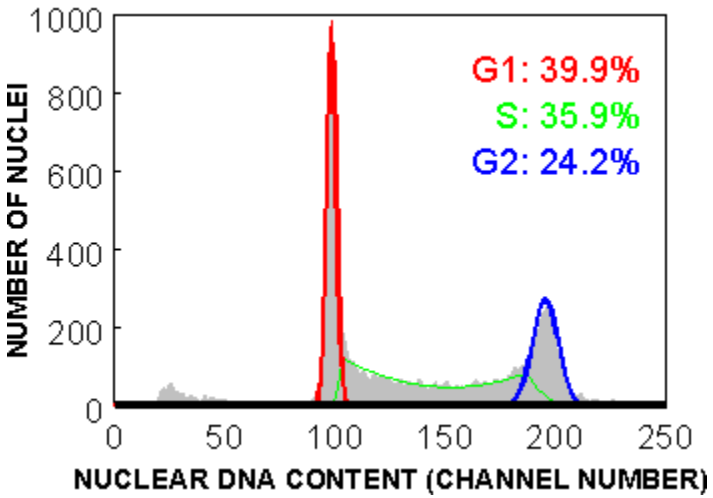


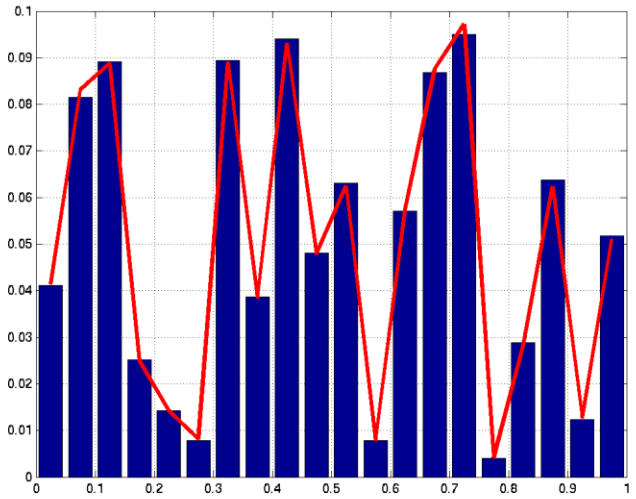
Q-Q Test of Normality



Parametric vs. Non-Parametric

- **Non-Parametric:**
  - Not normally distributed.
  - Smaller numbers.
  - Lower statistical strength.
  - Mann-Whitney U / Wilcoxon Signed Rank.





Belfast Health and Social Care Trust
caring supporting improving together

Ulster University

Assessment of Different Variables

Assessment of Different Variables

- **Two variables at same point:**

Assessment of Different Variables

- **Two variables at same point:**
  - Parametric: Independent T-Test.
  - Non-parametric: Mann-Whitney U Test.
    - E.g.. Medical co-morbidities between two populations.

Assessment of Different Variables

- **Two variables at same point:**
  - Parametric: Independent T-Test.
  - Non-parametric: Mann-Whitney U Test.
    - E.g.. Medical co-morbidities between two populations.

- **Assessment of same variable over two time-points:**

Assessment of Different Variables

- **Two variables at same point:**
  - Parametric: Independent T-Test.
  - Non-parametric: Mann-Whitney U Test.
    - E.g.. Medical co-morbidities between two populations.

- **Assessment of same variable over two time-points:**
  - Parametric: Paired T-Test.
  - Non-parametric: Wilcoxon-Signed Rank.
    - E.g.. Study population weight before and after eating.

Assessment of Different Variables

- **Two variables at same point:**
  - Parametric: Independent T-Test.
  - Non-parametric: Mann-Whitney U Test.
    - E.g.. Medical co-morbidities between two populations.

- **Assessment of same variable over two time-points:**
  - Parametric: Paired T-Test.
  - Non-parametric: Wilcoxon-Signed Rank.
    - E.g.. Study population weight before and after eating.

- **Assessment of same variable over multiple time-points:**

Assessment of Different Variables

- **Two variables at same point:**
  - Parametric: Independent T-Test.
  - Non-parametric: Mann-Whitney U Test.
    - E.g.. Medical co-morbidities between two populations.

- **Assessment of same variable over two time-points:**
  - Parametric: Paired T-Test.
  - Non-parametric: Wilcoxon-Signed Rank.
    - E.g.. Study population weight before and after eating.

- **Assessment of same variable over multiple time-points:**
  - Parametric: Analysis of Variance (ANOVA).
  - Non-parametric: Friedman Test (ordinal) / Kruskal-Wallis (rank).
    - E.g.. Effect of walking exercise pre-, 6-weeks and 24 weeks post-exercise programme.

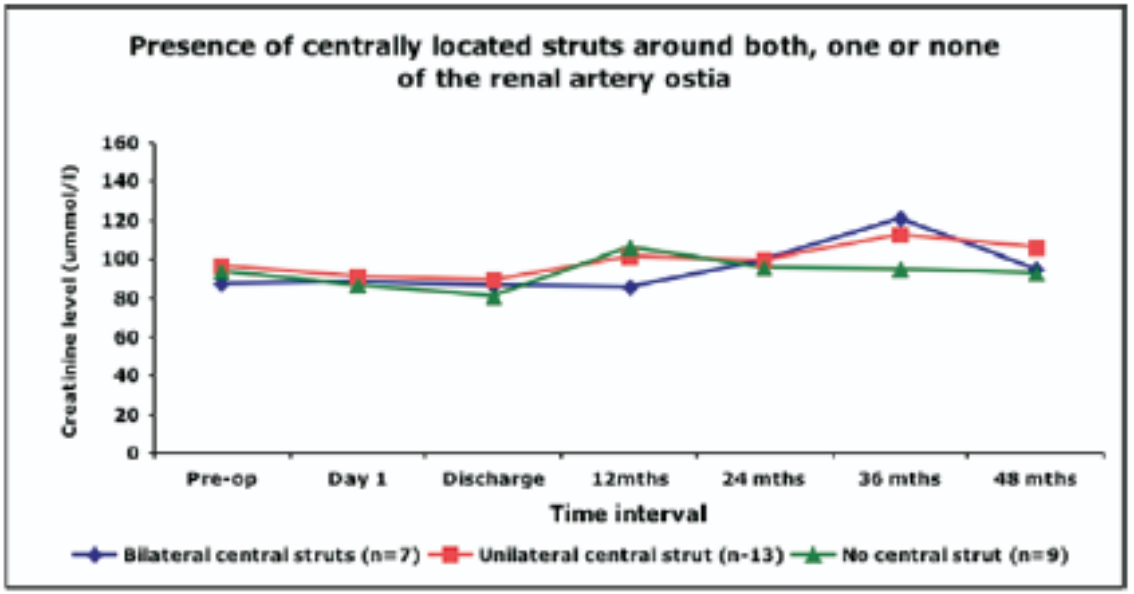


Fig 4. Analysis variance test for linear trend in serum creatinine concentration (mmol/L) for the anatomic configuration subgroup (presence of centrally located struts around one, both, or none of the renal artery ostia).

Variables with Multiple Outcomes

Variables with Multiple Outcomes

- Applies to a particular combination of characteristics relating to two or more classifications rather than a variable defined by the absence or presence of a characteristic.

- Frequency tables documented in contingency variables.

	History		
Cotinine Test	Nonsmokers	Ex-smokers	Active Smokers
+ve	13	29	66
−ve	127	157	12

Variables with Multiple Outcomes

- Method of Analysis varies according to;
  - Number of categories.
  - Whether the categories are ordered or not.
  - Number of independent groups or subjects.
  - Nature of objective.

- X^2 Test.

- Fishers Exact Test if n is small.

	History		
Cotinine Test	Nonsmokers	Ex-smokers	Active Smokers
+ve	13	29	66
−ve	127	157	12

Correlation

Correlation

- Measure of the degree to which two variables are related:
  - Identifying related characteristics (height and weight).
  - Evaluating validity (whether a variable, such as skinfold thickness, can be used as a measure of another variable, such as body fat)
  - Predicting the value of one variable using the values of a second variable (such as using mock exam scores as a predictor of examination grades).

Correlation

- If two variables are related, the value of one variable tells us something about the value of the other variable:
  - How much are the values of one variable associated with the values of another variable?
  - How much does one variable change as another variable changes?

Correlation

- Correlation is measured by evaluating the extent to which the deviations from the mean in one variable correspond to the deviations from the mean in another variable.

- The ***correlation coefficient*** ranges from -1.00 to +1.00.

Correlation

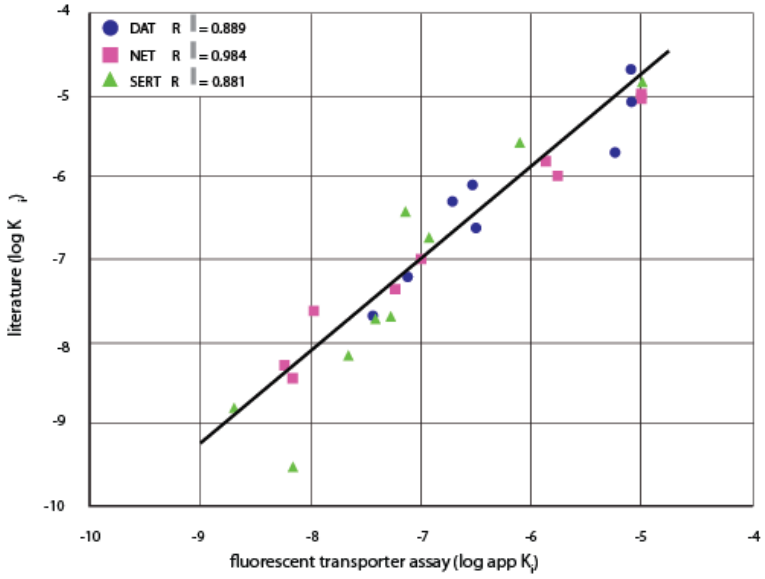
- Correlation coefficients are interpreted by their ***magnitude*** and ***sign****;*
  - A **correlation coefficient of +1.00** means that every subject's scores are *exactly* the same standardized distance and the same direction from the means for both variables.
  - A **correlation coefficient of -1.00** means that every subject's scores are the exactly same standardized distance but in opposite directions from the means of both variables.
  - A **correlation coefficient of 0** means that the two variables, age and height, are unrelated to one another. In this graph, older people are not systematically taller or shorter than younger people.

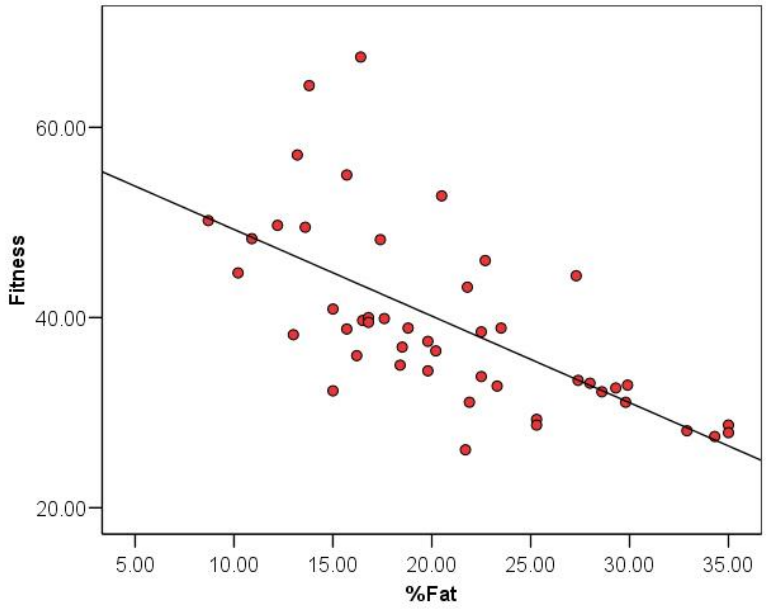
Correlation Analyses

- Parametric – Pearson's.

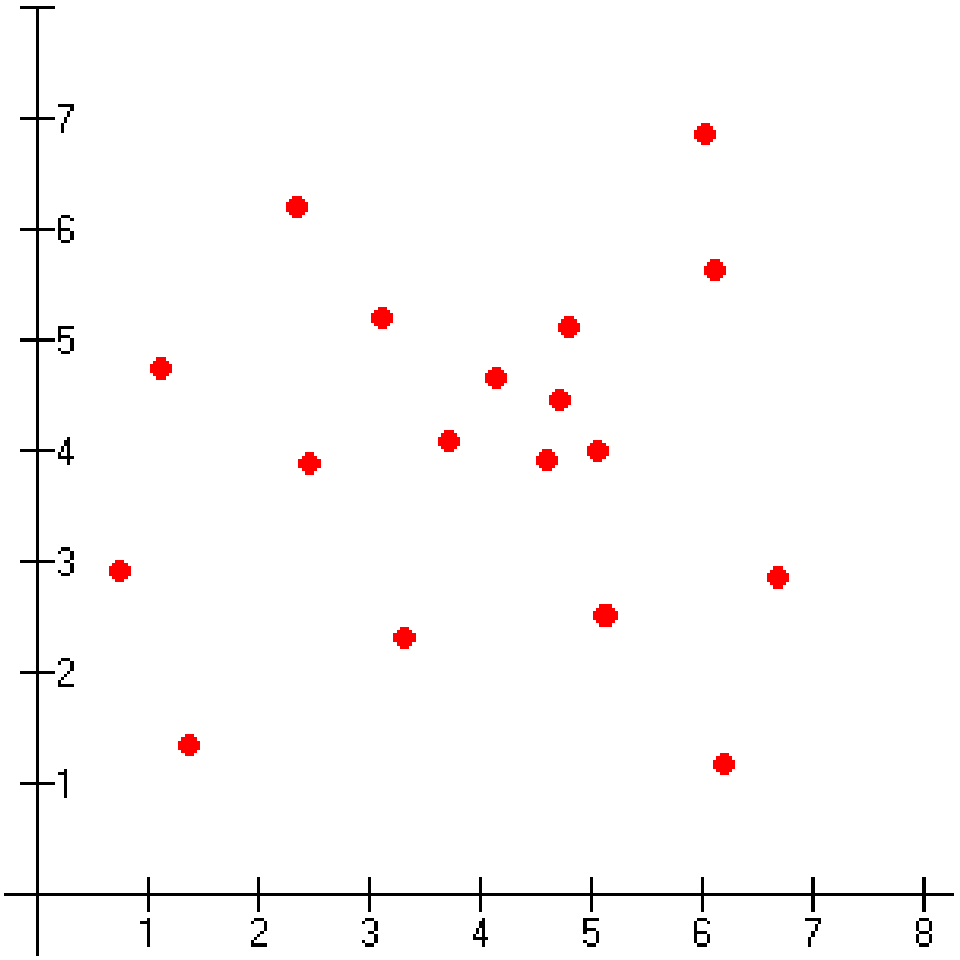
- Non-Parametric – Spearman's.

Correlation





Correlation



Correlation

A, *Correlation between intensive care unit stay and intra-abdominal pressure*

IAP time points	*Correlation coefficient*	P
2 hours	0.54	.002
6 hours	0.61	<.001
Day 1	0.49	.006
Day 2	0.74	<.001
Day 3	0.72	<.001
Day 4	0.73	<.001
Day 5	0.74	<.001

Survival Analyses

Survival Analyses

- Kaplan-Meier Plot.

- Analysed using Log-Rank Test.

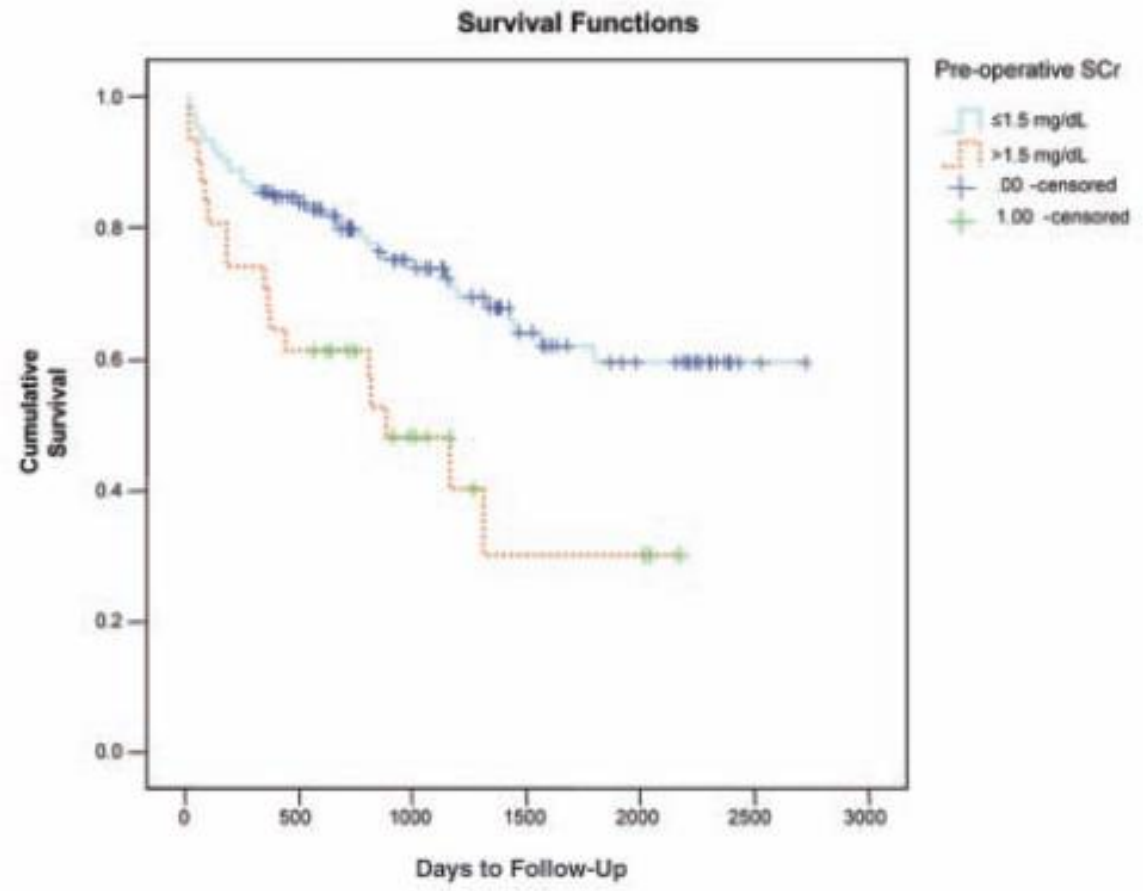


Figure 1. Comparison of cumulative survival in patients with normal and raised serum creatinine, $P < .02$. SCr = serum creatinine.

Logistic Regression

Logistic Regression

- Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

- Logistic regression can be binomial, ordinal or multinomial.
  - Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (for example, "dead" vs. "alive" or "win" vs. "loss").
  - Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "disease A" vs. "disease B" vs. "disease C") that are not ordered.

- Stepwise.

Receiver Operator Characteristic

Receiver Operator Characteristic

- A receiver operating characteristic (ROC) or ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

- Essentially the curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

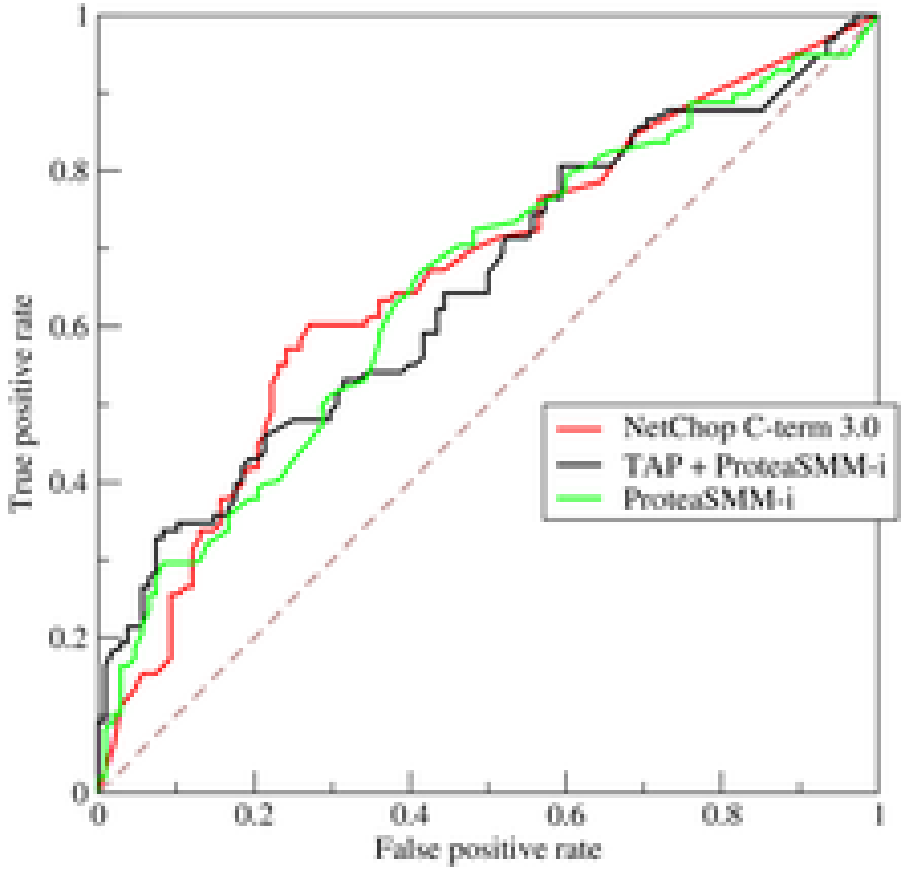
- Remember TPR (Sensitivity) and FPR (1-Specificity).

Receiver Operator Characteristic

- ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.

- ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

ROC Curve


True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
NetChop C-term 3.0
TAP + ProteaSMM-i
ProteaSMM-i

Problems with Multiple Comparisons

Problems with Multiple Comparisons

- In statistics, the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values.

- It is also known as the look-elsewhere effect.

- Errors in inference, including confidence intervals that fail to include their corresponding population parameters or hypothesis tests that incorrectly reject the null hypothesis, are more likely to occur when one considers the set as a whole.

Problems with Multiple Comparisons

- Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared.

- These techniques generally require a higher significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

Problems with Multiple Comparisons

- Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low.
- If multiple comparisons are done or multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases.
- The Bonferroni correction is based on the idea that if an experimenter is testing m hypotheses, then one way of maintaining the familywise error rate (FWER) is to test each individual hypothesis at a statistical significance level of 1/m times the desired maximum overall level.

Problems with Multiple Comparisons

- If the desired significance level for the whole family of tests is alpha , then the Bonferroni correction would test each individual hypothesis at a significance level of alpha / m.

- For example, if a trial is testing m = 8 hypotheses with a desired alpha = 0.05 then the Bonferroni correction would test each individual hypothesis at alpha / m = 0.05 / 8 = 0.00625.

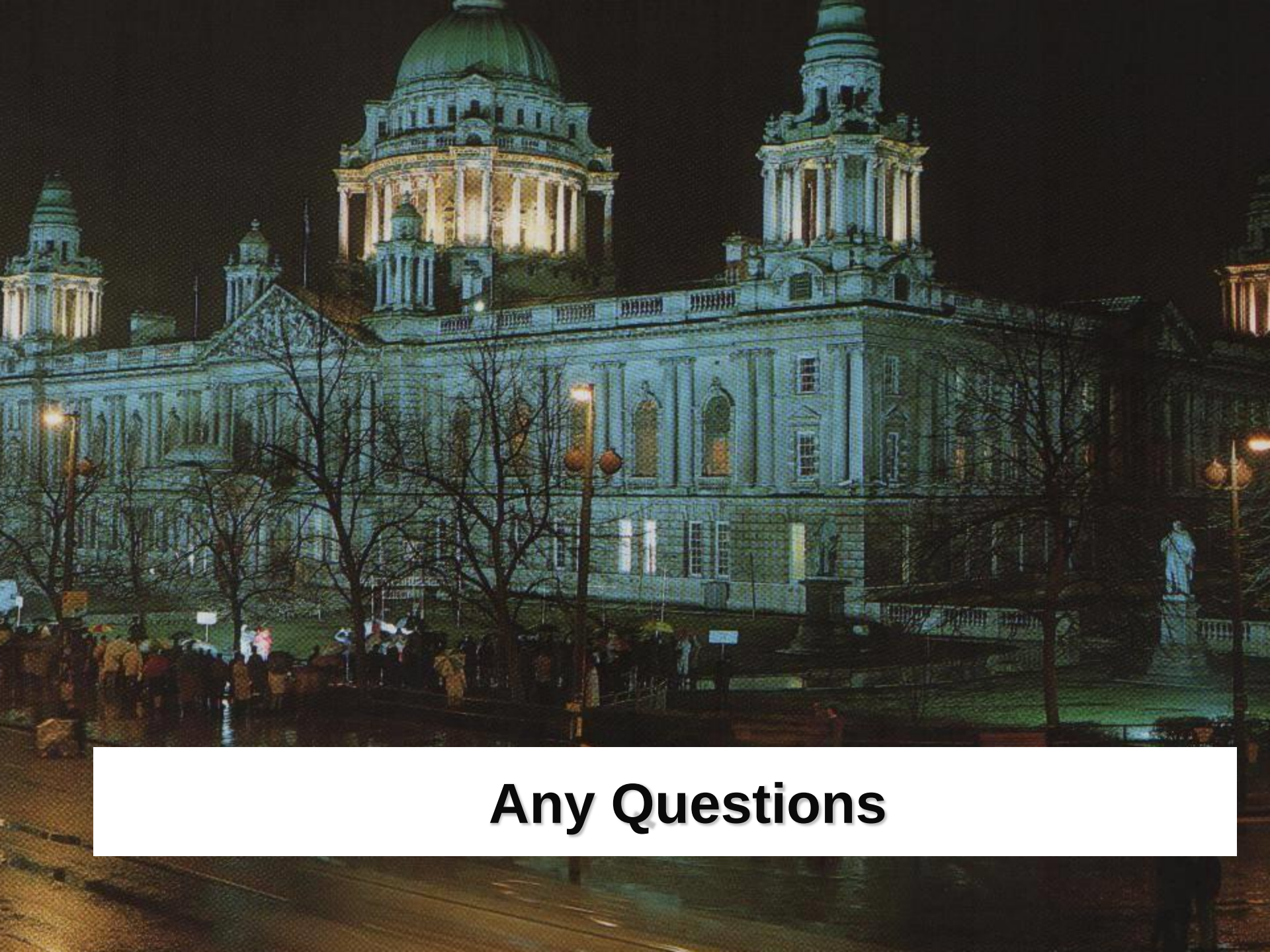
- Therefore level of significance becomes $p<0.00625$ not $p<0.05$.

Pitfalls in Analyses

- Outliers.
- Skew Data.
- Non-Independence.
- Alternative Aims / Results / Conclusions.
- Statistical Manipulation.

Common Mistakes in Research

- No or biased goals.
- Unsystematic approach.
- Incorrect performance metrics.
- Unrepresentative workload.
- Wrong evaluation technique, measurements, simulations, analytical modelling.
- Overlooking important parameters or ignoring significant factors.



Any Questions