Statistics for Oncology

A Course for Scottish Trainees

by… The Edinburgh Cancer Informatics Research Group

https://edin.ac/oncology-statistics

Preparation for Stats Exam

7/1/26

Dr Anna Cowell
Consultant Clinical Oncologist

Plan for today

- Exam format
- Resources
- “Exam definitions”
- Equations

- Choosing a hypothesis test – 6 example MCQs
- Common themes – 11 example MCQs

- Different learning styles – Mnemonics, odd associations, aide memoires…

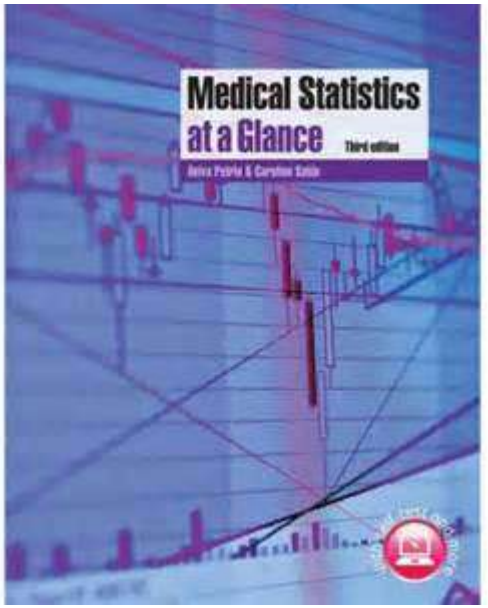
Exam format

- 40 Questions, 2 hours
- Generally enough time to recheck answers

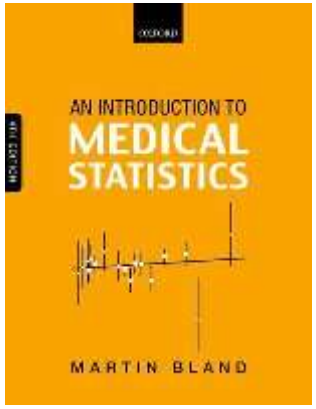
	All Candidates		UK-trained Candidates		UK 1st attempt Candidates	
Cancer Biology & Radiobiology	62%	117/188	77%	62/81	79%	48/61
Clinical Pharmacology	58%	107/183	71%	50/70	72%	44/61
Medical Statistics	54%	105/195	76%	54/71	78%	47/60
Physics	64%	136/212	65%	56/86	65%	39/60

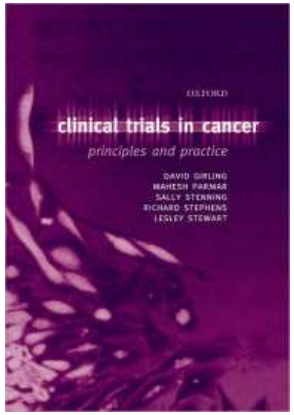
Resources

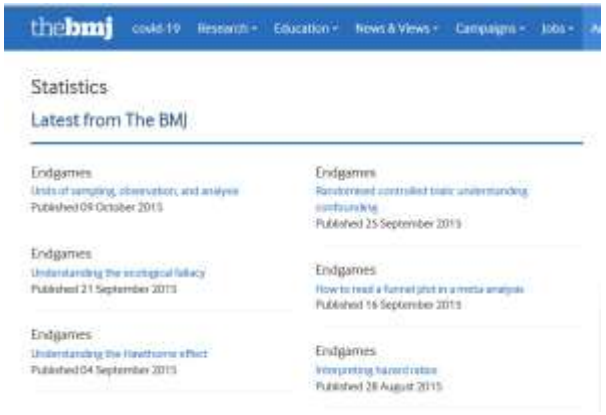
- This course and booklet
- Medical Statistics at a glance
  - Workbook
  - Online MCQs corresponding to the Chapters
- Mock exam questions



http://higheredbcs.wiley.com/legacy/college/petrie/1119167817/quiz/c01.html







“Exam definitions”

- **P value:** Assuming the null hypothesis is true, this is the probability we would obtain our results or more extreme results.
- **95% Confidence interval**: There is a 95% chance **the true population value (mean/OR/RR/HR)** lies between these two values, and a **2.5% chance** above and below.
- **Reference range:** 95% of sample data lies between these two values (if normally distributed).
- **Power:** Probability of finding a difference when one exists OR the probability of rejecting the null hypothesis when it is false.
- **Standard Error of the Mean:** An estimate of the precision of the sample mean.
- **Intention to treat analysis:** patients get analysed in the groups they were randomised to. The main aim is to **reduce allocation/selection bias**. It is not to provide real life clinical scenarios (although that is a benefit).
- **Progression free survival:** Time from randomisation until progression **or death**.
- **Phase I trial**: Main aim is to define a Phase 2 dose.
- **Phase II trial**: Main aim is to test efficacy of the drug.
- **Censoring**: The event did NOT occur during the time we observed the individual, and we only know the total number of days in which it didn’t occur. Advantage is we can include this data in the analysis.
- **Randomisation**: So each patient gets an equal chance of each drug.

Equations

Standard Deviation (SD) =

Population	Sample
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

95% Confidence Interval (CI) = Mean ± 1.96 x SEM

Variance = SD^2

Standard error of the mean (SEM) = SD/ $\sqrt{n}$

Standard error (proportion) = $\sqrt{\frac{p\,(1-p)}{n}}$

Sensitivity = True positives/All diseased

Specificity = True negatives/All healthy

PPV = True positives/All positive tests

NPV = True negatives/All negative tests

Likelihood ratio of a positive test = sens/(1-spec)

Likelihood of a negative test = (1-sens)/spec

Relative Risk Reduction (RRR) = ARR/CER

Absolute risk reduction (ARR) = Control Event Rate (CER) - Experimental Event Rate (EER)

Number Needed to Treat (NNT) = 1/ARR

Choosing a hypothesis test for 2 groups

- Comparing means
  - Independent – (Student's) two sample t-test
  - Paired – Paired t-test
- Comparing medians
  - Independent – Mann-Whitney U (Also called Wilcoxon Rank Sum test)
  - Paired – Wilcoxon Signed Rank
- Comparing proportions
  - CHI squared
  - Small numbers – (<5 expected per cell) – Fisher's exact test
  - Paired – McNemar's
  - *If categories linked – CHI squared test with trend*
- Comparing time to an event (Usually hazard ratios)
  - Log Rank (just tests if there is a difference, Cox proportional hazards model quantifies difference)
  - *Stratified Log Rank (Corrected for age/sex)*
  - *Log Rank test for trend (Ordered groups* e.g. deprivation level and stage of presentation)

Choosing a hypothesis test for >2 groups

- Comparing means
  - ANOVA
- Comparing medians
  - Kruskal-Wallis

Hypothesis Tests ★

Common Questions

- Which graph to use: **Stats at a glance**
- Which type of data is it: **Stats at a glance**
- Skewedness of data (positive or negative)
- Definition of P value
- Interpreting Confidence intervals
- Increasing/decreasing prevalence and effect on PPV/NPV
- Main aim of phase I or II trials
- Log rank on Kaplan Meier Curve.
- SAE/SAR/SUSAR
- Who reviews emerging safety and efficacy data in interim analysis in a RCT – data monitoring committee
- Random nonsense
  - Funnel plot is for publication bias
  - Actuarial survival curve is the probability of surviving each year and does NOT have the steps of a Kaplan Meier curve
  - Degrees of freedom in a Chi Squared Test (number of rows -1) x (number of columns -1) excluding headings and totals

Examiners report Spring 2025

*We would encourage future candidates to be clear with respect to their knowledge on sampling theory - particularly the relationships and differences between **samples and populations**. We would also encourage candidates to revise decisions around **choosing statistical tests with both categorical and continuous outcomes**, and survival analysis, particularly the principles and interpretation of **Kaplan-Meier** analyses.*

Regression

- Regression analysis is a statistical method used for the estimating the strength of the relationship between a dependent variable and one or more independent variables.

- Small print - Poisson regression – relative rate

$$\textbf{Rate} = \frac{\text{Number of events occurring}}{\text{Total number of years of follow-up for all individuals}}$$
$$= \frac{\text{Number of events occurring}}{\text{Person-years of follow-up}}$$

Regression Analyses

Continuous/explanatory variable as outcome (e.g. Creatinine Clearance, gentamicin dose)

Patient factors (height, weight, sex, star sign)

Multiple linear regression (Multivariate analysis)

Formula to calculate explanatory variable and 'rejects variables' not contributing

Known BINARY outcome (positive lymph nodes/no positive lymph nodes, metastasis/no metastasis, alive/dead)

Patient factors (tumour size, age) Retrospective – defined time point/period

Logistic regression

Odds ratios on how each factor does or does not influence the binary outcome.

Event occurrence, usually survival or PFS, over TIME.

Patient factors (drug A/Drug B, surgery/no surgery)

Cox Regression (Cox proportional hazard model)

Hazard ratio produced to quantify difference in event occurrence over time

Conclusion

- Don’t need to know the methodology/intricacies of the statistical tests/regression models etc – need to know how to select them.
- Select chapters of MSAAG – nicely laid out
- Some common questions – if you practice MCQs, you can pick up marks using pattern recognition.
- I will try to help you, but I am not a statistician ☹