APPLIED BIOSTATISTICS FOR THE PULMONOLOGIST

DR. VISHWANATH GELLA

- Statistics is a way of thinking about the world and decision making-*By Sir RA Fisher*

Why do we need statistics?

- **A man with one watch** always knows what time it is
- **A man with two watches** always searches to identify the correct one
- **A man with ten watches** is always reminded of the difficulty in measuring time

Objectives

- Overview of Biostatistical Terms and Concepts
- Application of Statistical Tests

Types of statistics

- Descriptive Statistics
  - identify patterns
  - leads to *hypothesis generation*
- Inferential Statistics
  - distinguish true differences from random variation
  - allows *hypothesis testing*

Study design

- Analytical studies

Case control study(Effect to cause)

Cohort study(Cause to effect)

- Experimental studies

Randomized controlled trials

Non-randomized trials

Sample size estimation

- ‘Too small’ or ‘Too large’
- Sample size depends upon four critical quantities: Type I & II error states (alpha & beta errors), the variability of the data $(S.D)^2$ and the effect size(d)
- For two group parallel RCT with a continuous outcome-sample size(n) per group = $16(S.D)^2/d^2$ for fixed alpha and beta values
- Anti hypertensive trial- effect size= 5 mmHg, S.D of the data-10 mm Hg. n= 16 X 100/25= 64 patients per group in the study
- Statistical packages - PASS in NCSS, n *query* or sample power

TYPES OF DATA

- Quantitative(“how much?”) or categorical variable(“what type?”)
- Quantitative variables

✓ continuous- Blood pressure, height, weight or age

✓ Discrete- No. of people in a family, no of attacks of asthma/wk

- Categorical variables

✓ Ordinal (ordered categories)- grade of breast cancer, stage of carcinoma

✓ Nominal(unordered categories)- sex(male, female), alive or dead, blood group(O, A, B, AB)

Descriptive statistics

- Identifies patterns in the data
- Identifies outliers
- Guides choice of statistical test

Describing the data with numbers

Measures of Central Tendency

- MEAN – average
- MEDIAN -- middle value
- MODE -- most frequently observed value(s)

Describing the data with numbers

Measures of Dispersion

- RANGE(difference between highest and lowest)
- STANDARD DEVIATION
- SKEWNESS

Standard deviation

- It is a summary measure of differences of each observation from the mean
- Squares of the differences are added, divided by (n-1), square root applied $SD = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

- Measure of scatter of observations about mean
- Significance can be better understood in relationship to a normal distribution curve

Distribution

- Normal(Mean=median=mode)
- Binomial (yes or no, positive or negative)
- Poisson
- Very important - decides the further application of statistical tests
  - Kolmogorov-Smirnov test

- Multiple random sample means will conform to a normal distribution- Central limit theorum
- Standard error of mean- the standard deviation of the sample means (s/√n)
- E.g- 25 males aged 20-24 years, mean temp-98.14^{0} F, S.D- 0.6 , S.E=0.12; 95% CI- 97.9-98.38
- Standard error of proportion, standard error of difference between two means and standard error of difference between proportions

Inferential Statistics

Used to determine the likelihood that a conclusion based on data from a sample is true

Terms

- **p value**: the probability that an observed difference could have occurred by chance
- $P<0.05$(1 in 20)-significant, $P<0.01$(1 in 100)
- Z score(relative deviate)- The no. of SDs that a specific score is above or below the mean in a distribution
- E.g- The pulse of a group of healthy males was 72 with a standard deviation of 2. The probability that a male chosen at random would be found to have a pulse of 80 or more- $Z=X-\text{Mean}/\text{S.D}=80-72/2=4$

Null hypothesis & Errors

- Sampling uncertainty (chance) is the reason for the observed result (null hypothesis)
- P value < 0.05 – rejects the null hypothesis
- Type 1 error(false positive): null hypothesis is incorrectly rejected
- Type 2 error(false negative): null hypothesis is incorrectly accepted

Which statistical test to use?

- Is there a study hypothesis?
- Decide whether hypothesis are confirmatory or exploratory?
- Are the data independent or dependent?
- What type of data is being analysed?(categorical or quantitative, Normally distributed or not?)
- No. of groups being analysed

Quantitative data-parametric tests

- Test the null hypothesis that there is no difference between two means
  - to test if sample mean differs significantly from a given population mean (**one-sample t-test**)
  - to test if the population means estimated by two independent samples differ significantly (**unpaired t-test**)
  - to test if the population means estimated by two dependent samples differ significantly (**paired t-test**)

Quantitative data-parametric tests

- T-test should be used only for two groups
- Although it is possible to divide three groups into three different pairs and use the t-test for each pair, this will increase the chance of making a type I error
  - Bonferroni's correction(p/√n)
  - Analysis of variance (ANOVA)
  - Repeated measurements

Quantitative data- nonparametric tests

- Mann-Whitney U test (identical to the Wilcoxon rank sum) equivalent to the unpaired Student's t-test
- Wilcoxon signed ranks test - equivalent to the paired Student's t-test
- Kruskal-Wallis test - equivalent of one-way ANOVA
- Friedman's test - repeated measures ANOVA

Categorical data

- Chi-square test – compare independent groups of categorical data (Yates' correction factor - sample size is small)
- Fisher's exact test - alternative for analyzing data from 2 x 2 tables
- McNemar's test – compare paired groups of categorical data

Terms

- Incidence rate: dividing the number of adverse events by number of patient-years
- Relative risk: rate therapy/ rate control group
- Relative risk reduction(RRR)- 1- RR
- Absolute risk reduction(ARR)- difference between 2 incidence rates(Rate therapy- rate control)
- Number needed to treat(NNT)- (1/ARR)

- E.g- The mortality rate for patients given Drug A is 25% and for patients who did not receive any treatment-50%

a) Relative risk-25/50=1/2

b) RRR= 1-1/2= 0.5

c) Absolute risk reduction=50-25=25%

d) NNT= 1/0.25= 4

Survival analysis

- Survival analysis is used when analysing time between entry into a study and a occurrence of subsequent event
- Originally used for time from treatment until death
- Two survival curves can be compared using the logrank test

Kaplan-meir survival curve

APPLIED BIOSTATISTICS FOR THE PULMONOLOGIST- PART 2

DR. VISHWANATH GELLA

Accuracy of diagnostic tests

- Accuracy of diagnsotic test – sensitivity and specificity
- Sensitivity- Probability of positive result in patients who have the condition or true-positive rate or Tp/Tp+Fn
- Specificity- ability to identify those patients who do not have the condition or the true negative rate or Tn/Tn+Fp
- Determined by administering the test to two groups i.e; diseased and non-diseased persons

PROBABILITY

- Prior probability-Quantitative expression of the confidence in the diagnosis before test (e.g- prevalence of the disease, calculated by prediction calculators
- Posterior probability- Revised statement of confidence in the diagnosis after test

- PPV- proprotion of patients with positive test result who actually have disease (Tp/Tp+Fp)
- NPV- proportion of patients with negative test results who actually do not have the disease (Tn/Tn+Fn)
- Tests having high sensitivity are useful for ruling out a disease in when the test is negative, screening tests are highly sensitive

Likelihood ratio

- Defined as the ratio of the probability of a given test result (e.g., “positive”or “negative”) in a patient with disease to the probability of that result in a patient without disease
- Likelihood Ratio(LR)$^{+}$ - sensitivity/False positive, value of 1 no change in disease likelihood
- Likelihood Ratio(LR)$^{-}$ –False negative/True negative
- Eg- Pleural fluid ADA- sensitivity and specificity of 90% (LR^{+}-9 OR 90/10) & (LR^{-} -0.11 OR 10/90)
- Higher values for LR^{+} (>5)more accurate diagnostic test & smaller values for LR- (<0.2)more accurate

- Probability(proportion)- No of times an outcome occurs/all occurences e.g- Blood sample
- Odds(ratio)- No of times a given outcome occurs/No of times that specific outcome does not occur e.g-
- Pretest odds = pretest prob/1-pretest prob
- Pretest odds x LR= Post test odds
- Posttest prob = posttest odds/1+posttest odds = PPV

Baye’s theorem

- Simple mathematical model to calculate post test probability based on sensitivity(a), specificity (b)and pre test probability(c)
- Post test probability= ac/ac+ (1-c) X Test FP rate
- Nomogram version – Accuracy of diagnostic test in question is summarised by LR
- Post test probability depends on prevalence of disease

Terms

- Incidence rate: dividing the number of adverse events by number of patient-years
- Relative risk: rate therapy/ rate control group
- Relative risk reduction(RRR)- 1- RR
- Absolute risk reduction(ARR)- difference between 2 incidence rates(Rate therapy- rate control)
- Number needed to treat(NNT)- (1/ARR)

Odds ratio vs. Risk ratio

	Outcome				
Exposure	Yes	No			
Yes	a	b	a/(a+b) Risk Rate of Outcome in Exposed	**Relative Risk** (Risk Ratio) = (a/(a+b)) ----------- (c/(c+d))	**Relative Risk Reduction** = (a/(a+b))-(c/(c+d)) ---------------------- (c/(c+d))
No	c	d	c/(c+d) Risk Rate of Outcome in non-exposed		
	a/c Odds of exposure in Cases	b/d Odds of exposure in Controls	ARR= a-b NNT = is the reciprocal of the absolute risk reduction		
	Odds Ratio = a/c ---- b/d				

ROC(RECEIVER OPERATING CHARACTERSTIC) CURVE

- Previous methods for revising pretest probability of a disease or condition on the basis of a diagnostic test apply if the outcome is simply positive or negative
- For test values measured on a continuum sensitivity and specificity depend on where the cutoff is set between positive and negative -E.g
- A more efficient way to display the relationship between sensitivity and specificity for tests having continuous outcomes is ROC curve.

Characterstics of ROC curve

- Initially developed in communication fields to display signal-noise relationship(signal-TP & noise-FP)
- Plot of the sensitivity (True-positive rate) to the false-positive rate (1- specificity)
- Closer to upper left-hand corner of the graph: TP-1 & FP-0

Characterstics of ROC curve

- Are useful graphic methods for comapring the two or more diagnostic tests or for selecting cutoff levels for a test
- A Statistical test can be performed to evaluate an ROC curve or to determine whether two ROC curves are significantly different
- Commonly used procedure-

a)Area uder each ROC curve

b) Comparison of areas using a modification of Wilcoxon rank sum test

Correlation & regression

- Correlation and regression are used to describe the relationship between two numerical variables
- Correlation- to denote an association between two quantitative variables
- Correlation coefficient(r)- Pearson’s correlation coefficient (measure of linear association), varies from +1 through 0 to -1
- Correlation coefficient calculated by a formula after plotting a scatter diagram using a formula

- Spearman rank correlation(non-parametric procedure)- for data which is not normally distributed
- Linear correlation-strengh of association, coefficient if large(positive or negative) next step is to develop a model for the relationship
- Regression analysis- Line of best fit
- Based upon formula y=mx+c

- Logistic regression analysis-Outcome of interest is a dichotomous categorical variable
- Linear regression-Outcome of interest is a continuous variable

Cox proportional hazards model

- It is a multivariate technique to analyse ‘Time to event’ curves
- Time to event curves-Uses all available information, including patients who fail to complete the trial
- Depicted by Kaplan-Meier curve
- Hazard ratio-hazard rate in treatment vs control group
- Hazard rate –It is the probability that the event in question if it has not occurred, the prob that it will ocuur in next time interval divided by the length of that interval
- HR-2 eg- a treated patient who has not yet healed by a certain time has twice the chance of being healed at the next point in time compared to some one in the control group

FAGAN’S NOMOGRAM FOR BAYE’S THEOREM

THANK YOU