Nonparametric Statistics

Sangita Kulathinal

1 April - 13 May 2009

University of Helsinki

Introduction

- Aim of any scientific investigation is to obtain information about some population on the basis of a sample drawn from it.
- Suppose the distribution (unknown) of the characteristic measured is continuous with distribution function F, some member of the family $\mathcal{F}$.
- We wish to guess about the true distribution using the information from the sample.
- This is a statistical inference problem (in a very geenral sense).

Parametric and nonparametric

- Parametric: For example, $F \sim N(\mu, \sigma^2)$. Each member of the normal family is determined by the values of two characteristics (parameters) μ, and σ^2. A family $\mathcal{F}$ of distribution functions is a parametric family if each memeber of the family $\mathcal{F}$ can be uniquely identified by the values of a finite number of real parameters.
- Nonparametric: For example, F is a continuous distribution. A family $\mathcal{F}$ of distribution functions which is not a parametric family is called a nonparametric family.
- Conceptual paradox: often nonparametric means more parameters.

Statistical inference

- An inference problem, where $\mathcal{F}$ is a parametric family is called a *parametric inference* problem.

- An inference problem where the family $\mathcal{F}$ is a nonparametric family is a *nonparametric inference* problem.

Distribution-free methods

- Inference procedures whose validity do not rest on a specific model for the population distributions are termed as distribution-free inference procedures.
- The term *nonparametric* relates to the property of the inference problem itself.
- The term *distribution-free* pertains to the property of the methodology used in solving inference problem.

Statistical hypothesis testing problems

- A statistical hypothesis is a statement about the population distirbution - form of the distribution or the numerical values of one or more parameters of the distribution.
- Two statements
  - Null hypothesis (H_0): the hypothesis which we want to test. For example, $\mathcal{F}$ be the family of all possible distribution functions. If F the population distribution belongs to a proper subset $\mathcal{F}_0$ then $H_0 : F \in \mathcal{F}_0$.
  - Alternative hypothesis (H_1): states the forms of the distribution when H_0 is nto true. For example, $H_1 : F \in \mathcal{F} - \mathcal{F}_0$.

Statistical hypothesis testing problems

- The family $\mathcal{F}$ decides whether the hypothesis testing problem is parameteric or nonparametric.
- If $\mathcal{F}$ is parametric then the testing problem is parametric otherwise it is nonparametric.
- Suppose H_0: population mean is 0.5, against H_1: population mean is not 0.5.
  If $\mathcal{F}$ is a parametric family like (i) all normal distirbutions, (ii) all normal distributions with variance one, then the problem is parametric testing problem.
  If $\mathcal{F}$ is a nonparametric family like (i) all continuous distributions, (ii) all continuous distribution on [0,1], then the testing problem is a nonparametric problem.

General method for solving a problem

- Consider a statistic T which is a function of observations $(X_1, \ldots, X_n)$
  - distribution of T is completely known under H_0
  - some values of T are more likely under H_0 and hence, favour H_0, whereas some other are more likely under H_1 and hence, favours H_1
- Question is what should be the cut-off points?
- Possible consequences of decision:
  - Correct decisions - Accept H_0 when it is true or reject H_0 when it is not true.
  - Type I error - Probability of (Reject H_0 when H_0 is in fact true).
  - Type II error - Probability of (Accept H_0 when it is false).

Power of a test

Power of a test is the probability that the test statistic will lead to the rejection of H_0. This is the probability of a correct decision and Power = 1 - Type II error.
Power depends on the following four variables:

- Degree of falseness of H_0
- Size of the test
- Number of observable random variables involved in the test statistic, generally sample size
- Rejection region R

Choosing between two or more tests

- *Most powerful test*: A test is the most powerful for a specified alternative hypothesis if no other test of the same size has greater power against the same alternative.
- *Uniformly most powerful test*: A test is uniformly most powerful against a class of alternatives if it is the most powerful with respect to each specific alternative within that class.
- *Consistent*: A test is consistent for a specified alternative if the power of the test when that alternative is true approaches 1 as the sample size approaches ∞.

Choosing between two or more tests

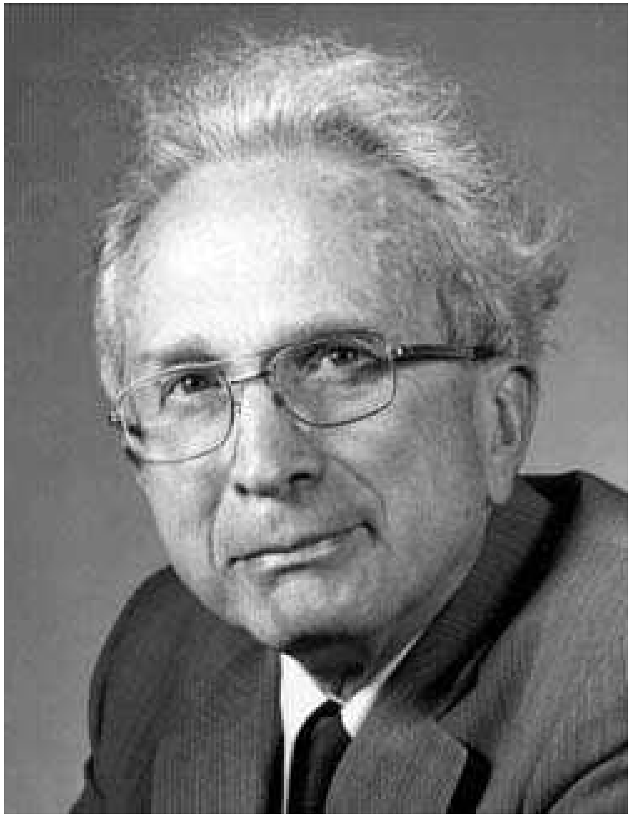
- *Power efficiency*: Power efficiency of a test *A* relative to a test *B*, where both tests are for the same simple null and alternative hypotheses, the same type of rejection region, and the same significance level, is the ratio (n_b/n_a), where n_a is the number of observations required by test *A* for the power of the test *A* to be equal to test *B* when (n_b) observations are used.
- *Asymptotic Relative Efficiency (ARE)*: The ARE of test *A* relative to test *B* is the limiting value of the ratio (n_b/n_a), where n_a and n_b are as defined above and when $n_b \to \infty$ and $H_1 \to H_0$.

Nonparametric methods

- Applications: widely used for ranked order data (such as relative scores in terms of 1-5 levels) but no clear numerical interpretation, for data on an ordinal scale.
- Methods are based on fewer assumptions, and hence, their applicability is much wider than the corresponding parametric methods.
- Are easier to use.
- The term nonparametric was first used by Wolfowitz, 1942.

Jacob Wolfowitz

Born: 19 March 1910 in Warsaw, Russian Empire (now Poland)
Died: 16 July 1981 in Tampa, Florida, USA



http://www-history.mcs.st-and.ac.uk/history/Biographies/Wolfowitz.html

Commonly used tests

- Test based on runs: used for testing randomness
- Sign test: used for testing whether median of the distribution is a specified value
- Wilcoxon signed-rank test
- Mann-Whitney U or Wilcoxon rank sum test
- Wald-Wolfowitz runs test
- Kolmogorov-Smirnov test
- Median test
- Measures of association: Spearman's rank correlation coefficient and Kendall's tau

Order statistics: Definition

- Data $X_1, \ldots, X_n$ from a population with continuous distribution F_x
- Suppose $X_{(1)}$ is the smallest of $X_1, \ldots, X_n$; $X_{(2)}$ is the second smallest, etc.; and $X_{(n)}$ is the largest.
- $X_{(1)} < \ldots < X_{(n)}$ denote the original sample which is arranged in the increasing order of their magnitudes.
- $X_{(1)} < \ldots < X_{(n)}$ are order statistics of the random sample $X_1, \ldots, X_n$.
- $X_{(r)}$, for $1 \leq r \leq n$ is the rth order statistic.

Ranks

- The ith rank-order statistic $r(X_i)$ is called the rank of the ith observation in the unordered sample. The value it assumes is $r(x_i)$ which is the number of observations x_j, $j = 1, \ldots, n$ such that $x_j \leq x_i$. That is $r(x_i) = \sum_{j=1}^{n} I(x_j \leq x_i)$.
- Note that $r(x_{(i)}) = i$.
- Data are in terms of relative importance for example, assessing preferences.

Order statistics (1)

- *Probability-integral transformation*: Let X have the cdf F_X. If F_X is continuous, the random variable $Y = F_X(X)$ has the uniform probability distribution over the interval $(0, 1)$.
- If $X_{(1)} < \ldots < X_{(n)}$ are order statistics of the original sample $X_1, \ldots, X_n$ then $F(X_{(1)}) < \ldots < F(X_{(n)})$ are order statistics from the uniform distribution on $(0, 1)$.
- These order statistics may be termed distribution-free, in the sense that their probability distribution is known to be uniform regardless of the original distribution F_X as long as it is continuous.

Order statistics (2)

- *Sample median*: $X_{([n+1]/2)}$ for n odd, and any number between $X_{(n/2)}$ and $X_{(n/2+1)}$ for n even. It is a measure of location and an estimate of the population central tendency
- *Sample midrange*: $(X_{(1)} + X_{(n)})/2$, measure of central tendency
- *Sample range*: $(X_{(n)} - X_{(1)})$, measure of dispersion
- *Sample interquartile range*: $(Q_3 - Q_1)/2$, measure of dispersion
- Sampling process which ceases after observing r failures out of n results into data $X_{(1)}, \ldots, X_{(r)}$ where $r \leq n$
- Useful in studying outliers or extreme observations

Order statistics: Distributions

- Joint distribution of $X_{(1)} < \ldots < X_{(n)}$
- Marginal distribution of $X_{(i)}$
- Joint distribution of $(X_{(i)}, X_{(j)})$

Empirical distribution: Definition

- True cdf of a r.v. is unknown in practice.
- We make educated guess about it and one way is to observe several observations from the unknown distribution and constructing a graph which may be used as an estimate of cdf.
- *E*mpirical distribution function: Let $X_1, \ldots, X_n$ be a random sample from cdf F. The *E*mpirical distribution function $F_n(x)$ is a function of x, which equals the fraction of $X_i's$ that are less than or equal to x for each x, $-\infty < x < \infty$.

$$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

Empirical distribution: Properties

- step function and is nondecreasing taking values between 0 and 1
- jumps at the observed value
- jump size is $1/n$ (when all observations are distinct)

Quantile

- pth *quantile*: The pth quantile, $(0 < p < 1)$, Q_p of the r.v. X with cdf F is the number such that $P(X < Q_p) \leq p$ and $P(X > Q_p) \leq 1 - p$.
- pth *sample quantile*: Let $X_1, \ldots, X_n$ be a random sample from cdf F. The pth sample quantile q_p is the number such that $\sum I(X_i < q_p)/n \leq p$ and $\sum I(X_i > q_p)/n \leq (1 - p)$.

Relative measure of association (1)

- For any two independent pairs (X_i, Y_i) and (X_j, Y_j) of random variables which follow this bivariate distribution, the measure will equal $+1$ if the relationship is direct and perfect in the sense that $X_i < X_j$ whenever $Y_i < Y_j$ or $X_i > X_j$ whenever $Y_i > Y_j$. This relationship will be referred to as perfect concordance (agreement).
- For any two independent pairs (X_i, Y_i) and (X_j, Y_j) of random variables which follow this bivariate distribution, the measure will equal -1 if the relationship is indirect and perfect in the sense that $X_i < X_j$ whenever $Y_i > Y_j$ or $X_i > X_j$ whenever $Y_i < Y_j$. This relationship will be referred to as perfect discordance (disagreement).

Relative measure of association (2)

- If neither criterion 1 nor 2 is true for all pairs, the measure will lie in between the two extremes -1 and $+1$.
- The measure will equal zero if X and Y are independent.
- The measure for X and Y will be the same as for Y and X, or $-X$ and $-Y$, or $-Y$ and $-X$.
- The measure for $-X$ and Y, or X and $-Y$ will be the negative of the measure for X and Y.
- The measure should be invariant under all transformations of X and Y for which order of magnitude is preserved.

Relative measure of association (3)

- Last criterion seems especially desirable in nonparametric statistics, as inferences must usually be determined by relative magnitudes as opposed to absolute magnitudes of the variables.
- Probabilities of events involving inequalities relations between the variables are invariant under all order-preserving transformations.
- A measure of association which is a function of such probabilities will satisfy the last criterion as well as all the other criteria.