Lecture 11

NONPARAMETRIC STATISTICAL INFERENCE

1. Motivation for using Nonparametric Inference.

1.1. General purpose of Nonparametric Procedures

There are several reasons that can be pointed out as a motivation to use Nonparametric

Procedures. Let us mention but a few:

- For many variables of interest we do not know for sure if they follow a normal distribution. For example, is income distributed normally in the population? Most likely not! The incidence rates of rare diseases are **not normally distributed**, the number of car accidents at a given cross road within given time interval (which number is a discrete variable to start with) is, of course, also not normally distributed. Since normality of observations is a basic assumption in applying `t`-test, for example, one can not justify its application for the above mentioned non-normal data.

- Another factor that limits the applicability of tests based on the assumption of *asymptotic* normality is the **sample size.** Consider a test that uses the sample mean as a test statistic. We can assume that the sampling distribution of this statistic is asymptotically normal by relying on the **Central Limit Theorem**. However, if the sample is *very small* and the single observations are not normally distributed, the normal approximation to the distribution of the sample mean would be very inaccurate and inference that uses this approximation may be flawed.

- Applications of tests that are based on normality assumption are further limited by a **lack of precise measurement**. On a more extreme scale, we may only have **rank** ordering of the observations and not their precise values.
  Hence, a need is there for statistical procedures that allow us to process data of **"low quality", from small samples, on variables about which nothing or very little is known concerning their distribution.**

In more technical terms, nonparametric methods **do not rely much on estimation of parameters describing the distribution of the variable of interest in the population.** Very often, the distribution of the test-statistic used in a nonparametric procedure does **not depend** on the individual distribution of each of the n i.i.d. observations used in its construction. Therefore, these methods are sometimes called *distribution-free* methods. Obviously, the distribution-free property extends the applicability of these procedures and is desirable.
Note: While often it is impossible to get away with **no assumptions at all** on the population that generates the data, it **is** possible to derive procedures whose assumptions are minimal, for example assuming only continuity of the density. Such procedures are also called distribution-free.

Also, historically, in non-parametric inference, much more research has been done on hypothesis testing than on point- or interval-estimation and this will be reflected in our discussions on the topic of non-parametric inference.

1.2. General remarks about scales of measurement. It is preferable to use **continuous scale** of measurement when sampling data. Within this scale, ideally no information loss occurs and this scale should be preferred when available. Not all data used in nonparametric statistics are of such type, however. For example, in contingency tables analysis, we typically use data that are dichotomous, i.e. success-failure, male-female, smoker-no-smoker or , more generally, data that may be classified according to different criteria and into two or more separate classes by each criterion. **No ordering** of any sort of the classes is assumed or is available and we call this type of data **nominal.** Testing independence for such type of data is usually performed by chi-square test. The chi-square test may also be used where more than just nominal information concerning the data is available but for different reasons (speed, ease of calculation, abundance of data etc.) some information contained in the data is disregarded and the data are reduced to nominal type. Clearly, this results in some **information loss** which manifests itself in lower power of the tests used. The loss could be reduced if more of the information in the data could be utilized by ordering it at least on an **ordinal scale.** We utilize the **ranks** in the further analysis. For example, instead of considering three non-ordered categories, we may attach ranks 1,2,3 to the nonnumeric data "good, better, best". If the data is numeric and $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$) then rank-order statistic can be any function r such that $r(X_{(i)}) \leq r(X_{(j)})$ holds whenever $i \leq j$. For any set of n sample observations, the simplest set of numbers to use to indicate relative positions is the set of first n positive integers. Rank statistics are statistics that are functions of the ranks. Advantages of rank-statistics include:

- They are particularly useful in non-parametric statistics since they are *distribution-free*. (Indeed, no matter what the original (continuous) distribution F_X of the data, the mere fact that we are dealing with n i.i.d. observations tells us that a particular ordering $R_1, R_2, \ldots, R_n$ of the ranks of the realizations is just one of the possible permutations of the numbers $1, 2, \ldots, n$ and has the same probability $1/n!$ of occurring.)
- testing using rank statistics is usually simple and quick to apply
- the loss of efficiency when using them, is usually small

1.3. Topics in nonparametrics to be discussed in our course.

The following areas are of great interest in nonparametric statistics and will be discussed here:

- For a sample from one population (that is **possibly non-normal**). We will **consider the counterparts of the usual t-test for the population mean** (that is, of the test that we would have used if the sample **was** normal). We will discuss the *Sign Test* and the *Wilcoxon signed rank test.* The latter tests only use the signs or the ranks of the observations (**not** their actual numerical values) and have broader applicability.
- Testing differences between **two dependent** groups. Here, we will be looking for **nonparametric alternatives of the t-test for paired observations**. These are again the *Sign test* and *Wilcoxon's test* but now applied on the **differences** of the

$N(\mu, \sigma^2)$

$i)\ T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim t_{n-1}$ $H_0 : \mu = \mu_0$

$(T) \geq t_{\frac{\alpha}{2}-1}$

(x_i, y_i) $N(\mu_x, \sigma_x^2)$ $N(\mu_y, \sigma_y^2)$ $H_0 : \mu_x = \mu_y \Rightarrow \bar{H}_0 : \mu_x - \mu_y = 0$

$i = 1, 2, \ldots n$ $D_i = X_i - Y_i$ $H_0 = \mu_D = 0$

$H_1 = \mu_0 \neq 0$

$T = \frac{\sqrt{n}\,\bar{D}}{S} \sim t_{n-1}$

$|T| > t_{\frac{\alpha}{2}, n-1}$

paired observations. If the variables of interest are 0-1 variables (dichotomous) then *McNemar's Chi-Square* test is appropriate.

- Instead of comparing **two** dependent samples, we may need to compare more than 2,say, **k** related samples. Under normality assumption, the method applied for this is the ANOVA (analysis of variance). We will be looking at a nonparametric alternative to ANOVA (analysis of variance). This will lead us to discussing the *Friedman's two-way analysis of variance* and *Cohran's Q* test.

- Testing differences between **two independent** samples. Here, we are looking for nonparametric alternatives of the `t`-test for non-paired observations. Such alternatives are the *Wald-Wolfowitz runs test, the Mann-Whitney U test*, and the *Kolmogorov-Smirnov two-sample test.*

- If we have multiple groups (k groups (samples) where $\mathbf{k} > 2$) we would be looking for nonparametric alternatives of the classical (normality-based) analysis of variance technique. We will discuss analysis of variance in its nonparametric variant- the *Kruskal-Wallis analysis of ranks.*

- **Relationships** between two variables. The standard measure for strength of the linear relationship (association) between two quantitative variables is the **correlation coefficient**. We can speak about the correlation between height and weight in a given age group of males. But, also for a non-quantitative data, we might be interested in a degree of association, for example, between sex (M/F) and voting intention (Labor/Liberal). For a ranked data, we might be interested, say, if there is a tendency for high GPA exam scores to be assiciated with high GMAT scores. Here, **nonparametric counterparts** of the standard correlation coefficient for variables that are measured on at least an ordinal (rank order) scale, are the *Spearman's R* and *Kendal's Tau*. If the two variables of interest are categorical in nature (like sex and voting intention), then appropriate measure is the *Phi* coefficient. The strength of the relationship can tested by *Chi-square Test*. As it turns out, the *Phi* coefficient is a suitable function of the value of the chi-square test-statistic.

1.4. Advantages and disadvantages of Nonparametric Procedures

We could point out the following **advantages** of the Nonparametric procedures:

- The extreme generality of nonparametric methods of inference and their wide scope of usefulness outside the normal family settings are definite advantages in application.

- They are usually simpler and easier to apply than their parametric counterparts.

- They are quite easy to understand.

- Often, they are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods).

- Usually, they are relatively insensitive to outliers; they use "simpler data"- usually the nonparametric procedure would require just the ranks of the observations rather than their actual magnitude.

(V)

For multiple groups $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, $i = 1, 2, \ldots, n$

$$\hat{\rho} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0, (1-\rho^2)^2)$$

$$\hat{\rho} \approx N\left(\rho, \frac{(1-\rho^2)^2}{n-3}\right)$$

$$\frac{1}{2} \ln \frac{1+\hat{\rho}}{1-\hat{\rho}} \approx N\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n}\right)$$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$\rho > 0$

$\rho < 0$

WHY IS IT NORMAL!

$$z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

↓

Fishers z Transform

$$\Rightarrow \hat{z} = \frac{1}{2} \ln \frac{1+\hat{\rho}}{1-\hat{\rho}}$$

$$\Rightarrow \sqrt{n}(\hat{z} - z) \xrightarrow{d} N(0,1)$$

The following could be pointed out as a **disadvantage** of nonparametric procedures:

They may have less efficiency (in estimation) or power (in testing) for a specific families of distributions.

Comparisons with parametric, robust or other nonparametric procedures to assess performance are interesting and informative but they generally do not provide the sought-for comprehensive analysis under more nonparametric conditions. We can even say that specific comparisons are contrary to the spirit of nonparametric methods. This is why it is very difficult to give definite rules of choice. Because of this and because of time limitations, we will not discuss extensively such comparisons in our course.

2. Order statistics

Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ denote a random sample from a population with a continuous distribution function F_X. Since F_X is assumed to be *continuous*, the probability of any two of these random variables assuming the same value is zero. After reordering the n values we get $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)})$ in which, as mentioned, the $\leq$ sign could also be replaced by $<$. These values are collectively termed the *order statistic* of the random sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$. The subject of order statistics generally deals with properties of $X_{(r)}$ $(r = 1, 2, \ldots, n)$ which is called the r-th order statistic.

Order statistics are particularly useful in nonparametric statistics because of the following:

Theorem 2.1 (Probability-integral transformation). If the random variable X has a continuous cdf F_X then the random variable $Y = F_X(X)$ has the uniform probability distribution over the interval (0,1). Further, given a sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ of n i.i.d. random variables with cdf F_X, the transformation $U_{(r)} = F_X(X_{(r)})$ produces a random variable $U_{(r)}$ which is the r-th order statistic from the uniform population in (0,1), *regardless* of what F_X is, i.e. $U_{(r)}$ is distribution-free.

Proof: For $y \in (0, 1)$, define $\xi = F_X^{-1}(y)$ to be the largest number satisfying the relationship $F_X(\xi) = y$. Then:

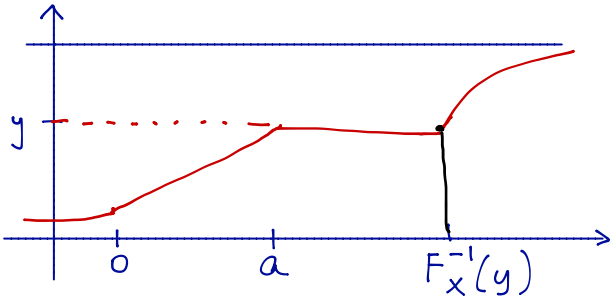
$$P(Y < y) = P(F_X(X) < y) = P(X < F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y \qquad (1)$$

Equality (1) means precisely that Y is uniformly distributed in the interval $(0, 1)$. Since F (being a cumulative distribution function) is monotone, ordering the X_i values as $X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)}$ will result in ordering the values $F_X(X_{(1)}) \leq F_X(X_{(2)}) \leq \ldots \leq F_X(X_{(n)})$, that is, of the values $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$.

$(U_{(1)} < U_{(2)} < \ldots < U_{(n)})$ $Y = F_X(X)$ $F_X(F_X^{-1})(\ldots)$

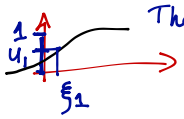
Note 1: For given $\alpha \in [0, 1]$, the above defined quantity $\xi = F_X^{-1}(\alpha)$ is called the α-quantile of the distribution F_X. If the random variable X represents a random payoff of a portfolio then this ξ represents a threshold in dollar terms below which the portfolio value falls with probability α. If this value is large in magnitude and with a negative sign it implies that the portfolio bears a high risk. The value ξ above is also called value at risk $(V@R_\alpha(X))$ at level α. Its disadvantage is, however, that it is not informative about the magnitude of the very large losses above the $V@R$ level and it is replaced by the *average*

$X_{(m)} =$ Uniform: $\frac{1}{m}(X_{(1)} + X_{(2)} + \cdots + X_{(m)})$



Take $Y \sim U[0,1] \rightarrow \xi = F_X^{-1}(Y) \Rightarrow \xi \sim F_X(\cdot)$

Generate Uniform Number!



Then ξ_1 will be an uniform number.

$P(F_X^{-1}(Y) < x) = P(F_X^{-1}(F_X(X)) < x)$
$= P(X < x) = F_X(x)$

$V@R$ (called $AV@R_\alpha(X)$):

$$AV@R_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha F_X^{-1}(p)dp = \frac{1}{\alpha}\int_0^\alpha V@R_p(X)dp.$$

Note 2: The above Probability-integral transformation theorem is very important in non-parametric statistics and will be used repeatedly in our course. The theorem has also an extremely important practical application in the generation **(computer simulation)** of observations from any specific continuous distribution function. There are several well-developed **uniform random number generators** that implement methods to generate sequences of uniform in (0,1) pseudo-random numbers. These numbers are **pseudo** since in fact they are generated by a deterministic algorithm (therefore are not random) **but** look as **random** (hence the word pseudo-random) in the sense that they pass usual statistical tests about randomness of the generated sequence. Every program system (Fortran, SPLUS, C, SAS, etc.) has such uniform random number generators and we will not discuss their specific implementation here. What we would like to discuss is how we could use these **uniform** random number generators to generate random numbers with **arbitrary continuous** cumulative distribution function F_X. The answer is:

1) Generate Y as uniformly distributed in (0,1) using the uniform random number generator

2) Calculate $\xi = F_X^{-1}(Y)$.

Then ξ is distributed according to $F_X(.)$ since its cumulative distribution function is:

$$P(F_X^{-1}(Y) < x) = P(F_X^{-1}(F_X(X)) < x) = P(X < x) = F_X(x).$$

Some obvious applications of order statistics are listed below:

- $X_{(n)}$ is of interest in studying floods, earthquakes and other extreme phenomena, sports records, financial markets etc.
- $X_{(1)}$ is useful, for example, in estimating strength of a chain that would depend on the weakest link.
- the sample median defined as $X_{[(n+1)/2]}$ for n odd and any number between $X_{(n/2)}$ and $X_{(n/2+1)}$ for n even, is a measure of location and an estimate of the population central tendency.
- the sample midrange $(X_{(n)} + X_{(1)})/2$ is also a measure of central tendency, whereas the sample range $X_{(n)} - X_{(1)}$ is a measure of dispersion.

2.2. Multinomial distribution. At this point, we will give the definition of the multinomial distribution. It is used to prove in an easy way many of the results related to distributions of order statistics and for that reason, it will be given first:
Suppose a single trial can result in k ($k \geq 2$) possible outcomes numbered $1, 2, \ldots, k$ and let $w_i = P(\text{a single trial results in outcome } i)$ ($\sum_{i=1}^k w_i = 1$). For n independent trials, let X_i denote the number of trials resulting in outcome i (then $\sum_{i=1}^k X_i = n$). Then we say that the distribution of $(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n; w_1, w_2, \ldots, w_k)$ and it holds

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} w_1^{x_1} w_2^{x_2} \ldots w_k^{x_k}, 0 < w_i < 1, \sum_{i=1}^k w_i = 1$$

Note that when $k = 2$ this is just the familiar Binomial distribution, i.e. the Multinomial can be considered as its generalization. It is also easy to see that

$$E(X_i) = nw_i; Var(X_i) = nw_i(1 - w_i), i = 1, 2, \ldots, k$$

holds by noting that the **marginal** distribution of the i-th component is in fact binomial ($\text{Bin}(n; w_i)$). In a bit more complicated way (for example, using moment generating functions), one can also show that $Cov(X_i, X_j) = -nw_i w_j$ holds for $j \neq i$.

$* \; w_i \geq 0 \;,\; \sum_{i=1}^{k} w_i = 1 \;,\; X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \;,\; \sum_{i=1}^{k} X_i = n$

Problem!

$\Rightarrow \varphi_X(t) = E[e^{t'X}]$

$X \in \mathbb{R}^k, t \in \mathbb{R}^k$

$\frac{\partial}{\partial t_i} \varphi_X(t)\Big|_{t=0} = E[X_i]$

$\frac{\partial^2}{\partial t_i \partial t_j} \varphi_X(t)\Big|_{t=0} = E[X_i X_j]$