

## Lecture 16

### 5.1. K-sample problems ( $K > 2$ ).

The procedures in this lecture are considered generalizations of the corresponding procedures for comparing  $K = 2$  groups of samples. Difference should be made between the cases of **related** and **unrelated (independent)** samples. In the case  $K = 2$  and for normal populations, such differences have been discussed in your basic Statistics course. They result in the 2 different forms of **t tests**, namely the two-sample **t test** (for unrelated samples) and the application of the simple one-sample **t test** on the **differences** of the observations in the case of **paired observations**. In non-parametric setting, for the case of  $K = 2$ , we also made a different treatment of the two situations: we applied the Wald-Wolfowitz runs, Mann-Whitney (or equivalently Wilcoxon rank sum) test for the case of 2 independent samples. On the contrary, for related (paired samples) we were dealing with the paired-sample variant of the Signed Test or of the Signed Rank Test for one sample. We shall continue this parallel also for the case of  $K > 2$  samples.

### 5.2. The case of $K$ independent (unrelated) samples (Kruskal-Wallis test).

First, we want to generalize the testing procedures for the two-sample location problem to situations in which the data consists of  $K \geq 3$  independent random samples, one sample from each of  $K$  populations. The null hypothesis is always that of no difference in locations (medians). Under the null hypothesis, the  $K$  samples can be treated as a single (combined) sample from one population. The alternatives correspond to a variety of restricted nonnull relationships between the locations (medians).

The above procedure is treated by the ANOVA technique in case of normal populations but we explicitly assume here that the normality assumption is not justifiable and are looking for alternative, nonparametric variant of the ANOVA technique. As is custom in nonparametric statistics, we would like to get a distribution free procedure by performing the analysis on the **ranks** of the observations.

#### 5.2.1. General notation and formulation of the Kruskal-Wallis test

The Kruskal-Wallis test is based on the **ranks** of the observations and can be considered as a generalization of the Mann-Whitney rank test for two samples to the case of  $K$  samples.

In another sense, the Kruskal-Wallis test can be considered a true alternative of the ANOVA technique for the case when normality assumption for the samples is not justifiable.

We assume that  $n = \sum_{i=1}^K n_i$  observations are available, with  $n_i$  observations from the  $i$ th treatment.

Treatment 1	Treatment 2	...	Treatment k
$X_{11}$	$X_{21}$	...	$X_{K1}$
$X_{12}$	$X_{22}$	...	$X_{K2}$
...	...	...	...
...	...	...	...
$X_{1n_1}$	$X_{2n_2}$	...	$X_{Kn_K}$

We make the following assumptions:

- The  $n$  observations are independent
- For each fixed  $i$ , the  $n_i$  variables  $X_{i1}, \dots, X_{in_i}$  are a random sample from a continuous

distribution with cumulative distribution function  $F_i$

- The distribution functions  $F_1, \dots, F_K$  are related through the relationship  $F_i(t) = F(t - \tau_i)$  (i.e.  $\tau_i$  can be considered unknown treatment effect for the  $i$ th population).

Under the null hypothesis about no difference among the treatment effects ( $H_0 : \tau_1 = \tau_2 = \dots = \tau_K$ ), we have  $n$  observations from the same population. Let us recall the main idea of the ANOVA technique if for a moment we assume that the observations in each of the  $K$  samples were normal. Since **always** the following decomposition holds:

$$\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 = \underbrace{\sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X}_{..})^2}_{\text{variability around the total mean}} + \underbrace{\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}_{\text{variability within distribution } i.} = SSTr + SSE$$

Due to treatment      Due to error

it is clear that under the assumption of equal means for the  $K$  samples, the first summand on the RHS would tend to be smaller than the second and the normed ratio  $F = \frac{SSTr/(K-1)}{SSE/(n-K)}$  will tend to be small. The test is based on this  $F$  statistic. **Under normality assumption**,  $F$  is f-distributed with degrees of freedom  $(K-1), (n-K)$  and therefore, the threshold constant to compare with, is  $f_{\alpha; (K-1), (n-K)}$ . For a given level  $\alpha$ , the test (called **F test**) rejects the hypothesis of equality of the  $K$  means when  $F > f_{\alpha; (K-1), (n-K)}$  holds.

The above distribution theory **does not apply** though when the normality hypothesis is violated and we shall be looking for alternative rank-based tests.

We shall still exploit the idea that under the null hypothesis of all shifts equal, we are dealing with  $n$  observations from the same population and hence all possible assignments of **ranks**  $1, 2, \dots, n$  to the  $K$  samples are equally likely and we would expect the ranks to be intermingled in the samples. On the contrary, if the null hypothesis is violated, then some samples will consists mostly of observations with small ranks whereas some other samples will consist of observations with predominantly large ranks. Denote by  $R_{ij}$  and by  $\bar{R}_i$  respectively the sum and the average of the ranks in the  $i$ th sample. Under  $H_0$  we have  $E(R_{ij}) = E(\bar{R}_i) = (n+1)/2$  (WHY !). The **Kruskal-Wallis** statistic is defined as a **weighted sum** of the squared differences of  $\bar{R}_i$  and their means under  $H_0$ , the weights being chosen properly so that a limiting distribution of the resulting statistic under  $H_0$  could be derived (see Section 5.2.2 below).

$$H = \frac{12}{n(n+1)} \sum_{i=1}^K n_i \left( \bar{R}_i - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^K \frac{R_i^2}{n_i} - 3(n+1)$$

Total sum of all rank average       $\bar{R}_i = R_i/n_i$       ONLY WORK FOR RANKS!

Statistic       $SSTr - SSE = \text{const.}$

The second line gives the same formula in an easier to compute manner (it involves the totals, not averages and in addition, requires only one subtraction). Apparently, the rejection region of a test of  $H_0$  is in the form  $H \geq c$  where  $c$  is chosen to satisfy the requirement for the size to be equal to  $\alpha$ . Tables exist (check Table K in the set of tables given to you) for the exact distribution of the statistic  $H$  but since both  $K$  and all  $n_i$  are entries, these tables are too cumbersome to work with. Therefore asymptotic approximations are preferred in most situations. Asymptotically, the following result is true:

**5.2.2. Asymptotic distribution of the  $H$  statistic.** When  $H_0$  is true, the statistic  $H$  has, as  $\min(n_1, n_2, \dots, n_K) \rightarrow \infty$ , as a limiting distribution  $\chi^2_{K-1}$ . The rejection region is then  $\{H : H \geq \chi^2_{K-1, \alpha}\}$ .

**“Proof:”** We shall only indicate why the result can be expected to hold. Under the null hypothesis, the  $n_i$  entries of the ranks in the  $i$ th population are randomly selected from the set  $\{1, 2, \dots, n\}$ . These entries are a random sample without replacement from the **finite** population of  $n$  integers. The mean of this population is

$$\mu = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2}$$

and its variance (per **definition** of a finite population's variance is

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n [i - \frac{n+1}{2}]^2 = \dots = \frac{n(n+1)}{12}.$$

The average rank sum  $\bar{R}_i$  has a distribution of a sample mean from a **finite** population and as for any sample mean from finite population, enjoys  $E(\bar{R}_i) = \mu$  (the population mean) and  $Var(\bar{R}_i) = (1 - \frac{n_i}{n}) \frac{\sigma^2}{n_i}$  (the factor  $(1 - \frac{n_i}{n})$  being called **fpc** (finite population correction factor)). Here then we have  $E(\bar{R}_i) = \frac{n+1}{2}$ ,  $Var(\bar{R}_i) = \frac{(n+1)(n-n_i)}{12n_i}$ .

A variant of the CLT (Central Limit Theorem) then allows us to approximate (if  $n_i$  is large) the distribution of

$$Z_i = \frac{\bar{R}_i - \frac{n+1}{2}}{\sqrt{(n+1)(n-n_i)/(12n_i)}}$$

$$\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij} \quad Var(\bar{R}_i) = (1 - \frac{n_i}{n}) \frac{\sigma^2}{n_i}$$

by the standard normal and hence  $Z_i^2$  by a  $\chi^2$  with one degree of freedom. This holds for  $i = 1, 2, \dots, K$  **but** the  $Z_i$ 's are clearly **not** independent since  $\sum_{i=1}^K n_i \bar{R}_i = \frac{n(n+1)}{2}$  holds.

It is easy to see (check it (!)) that

$$E[H] = K - \frac{\sum_{i=1}^K n_i}{n} = K - 1$$

$$\sum_{i=1}^K (1 - \frac{n_i}{n}) Z_i^2 = H$$

$$Z_i^2 = \frac{\sum_{j=1}^K 12 n_i (\bar{R}_i - \frac{n+1}{2})^2}{(n+1)(n-n_i)} = \frac{(n+1)}{n} = H \text{ statistic}$$

$\sum_{i=1}^K n_i \bar{R}_i = \frac{n(n+1)}{2}$  !!! They are dependent! Cannot not say that they are chi sq

holds (i.e. the statistic  $H$  equals the fpc-corrected sum of squares of the  $Z_i$ 's). Then it is no wonder that in the limit the expected value under  $H_0$  is  $K - 1 = \sum_{i=1}^K (1 - \frac{n_i}{n})$ . But it was Kruskal who first dealt successfully with the dependence of the  $Z_i$ 's and showed rigorously that the limiting distribution is  $\chi^2_{K-1}$ .

**5.3. The case of  $K$  related samples-Friedman's Test for a Randomized Block Experiment.** The Friedman test to be presented below, can be considered as a generalisation of the Sign test. It should be noted that there exists also the **Quade** test which can be considered as a generalization of the Wilcoxon signed rank test. Historically, the Friedman test is the better-known test of the two and it also requires fewer assumptions (although it suffers from a lack of power when there are only  $K = 3$  treatments, just as the sign test has less power than the Wilcoxon signed test when there are only two treatments). When  $K > 3$ , the Friedman test gains power and is the preferred test for large values of  $K$ . The problem of several **related samples** arises in experiments where the main purpose is to detect eventual differences between  $K$  different treatments where

the observations are arranged in blocks, each block being a group of  $K$  experimental units similar to each other (think of  $K$  puppies that are littermates and therefore may tend to respond to a particular stimulus more similarly than would randomly selected puppies from various litters). The  $K$  experimental units within a block are matched randomly with the  $K$  treatments being scrutinized, so that each treatment is administered once and only once within each block. In this way, the treatments may be compared in a fair way with each other without an excess of unwanted effect confusing the results of the experiment. This type of experiment is called *randomized complete block design* and is widely used in Psychology, Medicine, Agriculture etc. Suppose now that

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where  $\alpha_i$  can be interpreted as  $i$ th treatment effect,  $\beta_j$  is the  $j$ th block effect and the  $\epsilon_{ij}$  are drawn independently from a continuous (but **not** necessarily normal) distribution. One observation only is available for each combination of the indices  $i = 1, 2, \dots, K$  and  $j = 1, 2, \dots, l$ , ie.  $n = l * K$  holds. To test  $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_K = 0$  ie. the null hypothesis of no treatment effect, the observations are first ranked separately from 1 to  $K$  within each of the  $l$  blocks, then the rank average  $\bar{R}_i$  is computed for each of the  $K$  treatments. Under  $H_0$ , the  $\bar{R}_i$  values should be close to each other, since within each block the  $K!$  assignments of ranks are equally likely. **Friedman's test statistic** is defined then as

$$F = \frac{12l}{K(K+1)} \sum_{i=1}^K (\bar{R}_i - \frac{K+1}{2})^2 = \frac{12}{lK(K+1)} \sum_{i=1}^K R_i^2 - 3l(K+1)$$

*correction*      *easier to calculate*

As with the Kruskal-Wallis, Friedman's test rejects  $H_0$  when the computed value of the test-statistic is large enough. Exact distribution is tabulated in some more comprehensive textbooks. Alternatively, even for moderate values of  $K$ , the test-statistic tends in distribution to  $\chi^2_{K-1}$ . The rejection region is then  $\{F : F \geq \chi^2_{K-1, \alpha}\}$ .

#### 5.4. Examples. At lectures/tutorials.

	BL 1	BL 2	BL 3	...	BL	
Tr=1						$\bar{R}_{1.}$
Tr=2						$\bar{R}_{2.}$
⋮						⋮
Tr=l						$\bar{R}_{l.}$

*do it for each treatment.*

$(1+2+\dots+K) = \frac{K(K+1)}{2}$

$K=4$