STA 104 Applied Nonparametric Statistics

Chapter 4: One-Way Layout Problems: Nonparametric One-Way Analysis of Variance

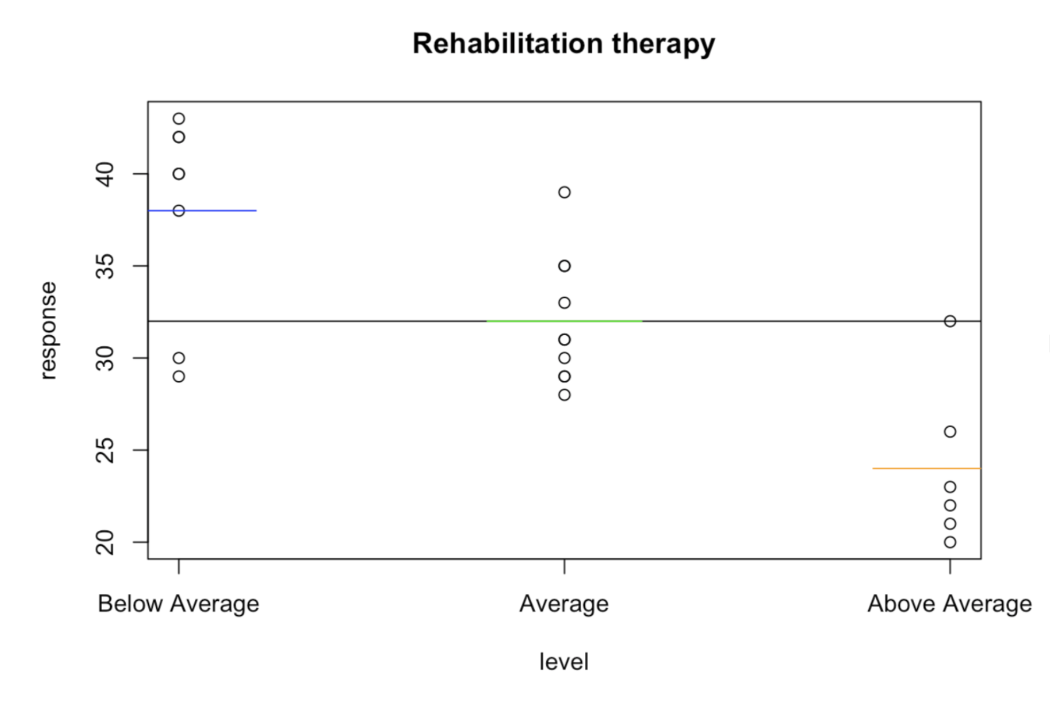
Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents

One-Way Data Layout

Treatments	Observations	Sample Sizes
1	$X_{11}, X_{12}, \ldots, X_{1n_1}$	n_1
2	$X_{11}, X_{12}, \ldots, X_{1n_2}$	n_2
...	...	...
k	$X_{11}, X_{12}, \ldots, X_{1n_k}$	n_k



Setting

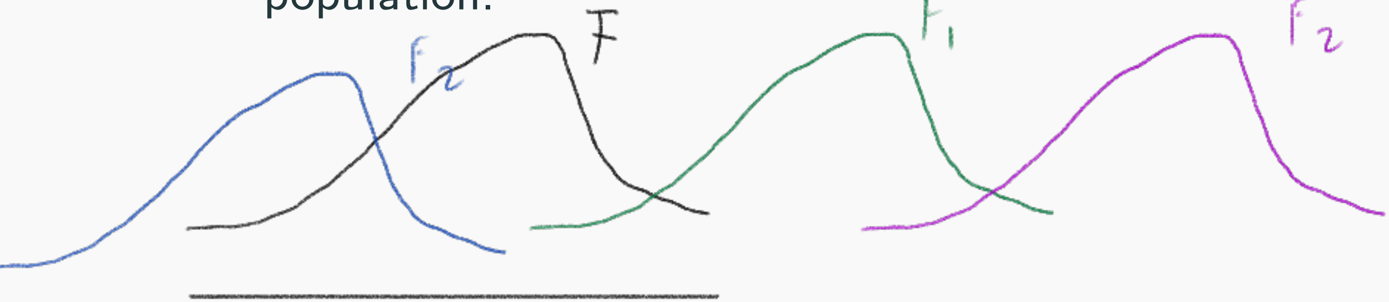
The data consist of $N = \sum_{i=1}^{k} n_j$ observations, with n_i observations from the ith treatment, $i = 1, \ldots, k$.

- For each treatment group $i \in \{1, \ldots, k\}$, the n_i observations are a random sample from a continuous distribution with distribution function F_i.
- The N observations are mutually independent.
- The distribution functions $F_1, \ldots, F_k$ are connected through the relationship

Location shift model

$$F_i(t) = F\left(t - \tau_i\right), -\infty < t < \infty,$$

for $i = 1, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ and τ_i is the unknown treatment effect for the i th population.



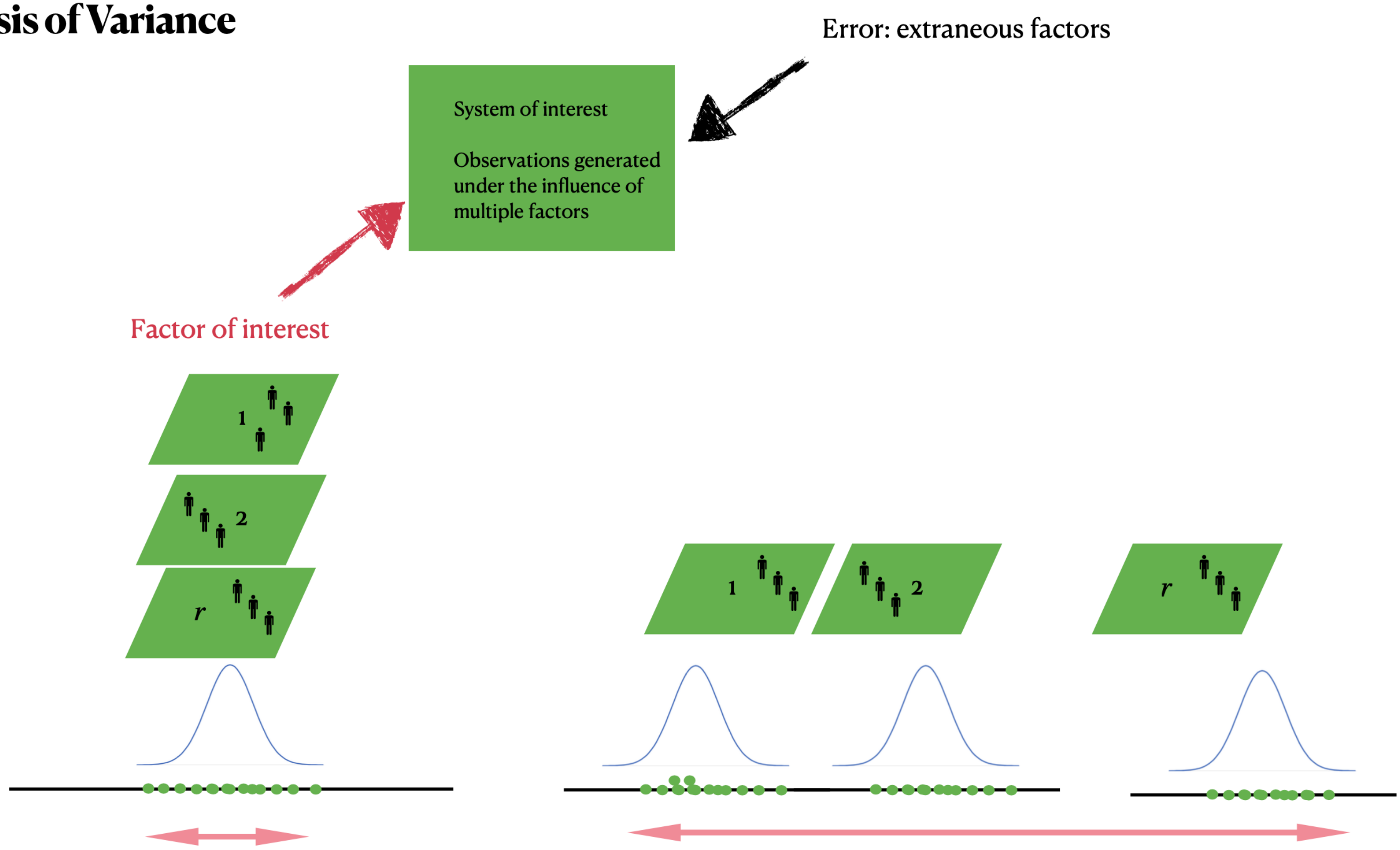
This is the usual one-way layout model: One-Way Analysis of Variance (ANOVA), commonly associated with normal assumptions:

$$X_{ij} = \theta + \tau_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

where

- θ is the overall median,
- τ_i is the treatment i effect,
- the noise e_{ij}s are a random sample from a continuous distribution with median 0. (Under the additional assumption of normality, the medians θ and 0 are, of course, also the respective means.)

Analysis of Variance



Error: extraneous factors

System of interest

Observations generated under the influence of multiple factors

Factor of interest

Without factor of interest, the observations have some natural variation due to other extraneous factors, i.e. “error variance”

If the factor of interest indeed has some effects on the system, then we would expect more volatility than a system without the factor

Hypothesis

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \qquad \underbrace{\tau_1 \ldots \tau_k \text{ not all equal}}_{\text{at least two of the treatment effects are not equal}}$$

Review of One-Way ANOVA

The sum of squares for treatments is defined as

$$\text{SST} = \sum_{i=1}^{k} n_i \left(\bar{X}_{i.} - \bar{X}_{..}\right)^2$$

where $\bar{X}$ is the mean of all the observations-namely,

$$\bar{X}_{..} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_j} X_{ij}}{N}$$

The mean squares for treatment is

$$\text{MST} = \frac{\text{SST}}{k-1}$$

The sum of squares for error is defined as

$$\text{SSE} = \cancel{\sum_{i=1}^{k} (n_i - 1) S_i^2} \qquad \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)^2$$

and the mean squares for error is

$$\text{MSE} = \frac{\text{SSE}}{N-k}$$

Review of One-Way ANOVA

if $F_i \sim$ Normal :

The F statistic is given by

$$F = \frac{\text{MST} \overset{H_0}{\sim} \chi^2_{k-1}}{\text{MSE} \sim \chi^2_{N-k}} \sim F(\cdot,\cdot)$$

In ANOVA course, we learn that: If the observations are selected at random from normally distributed populations with equal variances, then this statistic has an F-distribution with $k-1$ degrees of freedom for the numerator and $N-k$ degrees of freedom for the denominator. One may use this distribution to determine a p-value for the observed statistic and therefore conduct hypothesis testing.

However, if we are unwilling to assume that the population distributions are normal or the normally assumption is fundamentally wrong for the data at hand?

⇒ Nonparametric ANOVA

The Kruskal-Wallis Test

Hypothesis

Two-Sided Test:

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

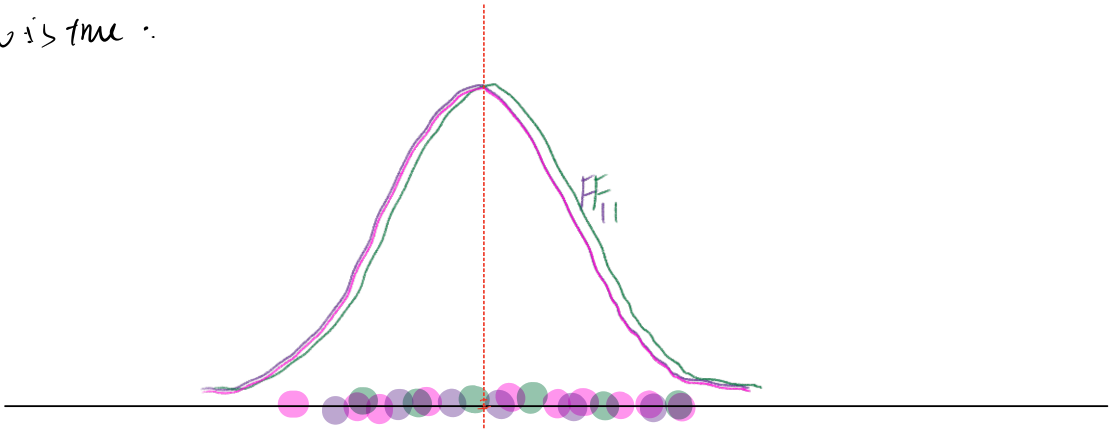
$$H_1 : \underbrace{\tau_1 \ldots \tau_k \text{ not all equal}}_{\text{at least two of the treatment effects are not equal}}$$

Motivation

A way to obtain a nonparametric rank test for comparing k treatments is to replace the original observations with ranks and then perform the permutation F-test on these ranks.

We will obtain a statistic that is equivalent to the F statistic applied to ranks, with a permutation distribution that may be approximated by the chi-square distribution with $k-1$ degrees of freedom.

if H_0 is true :

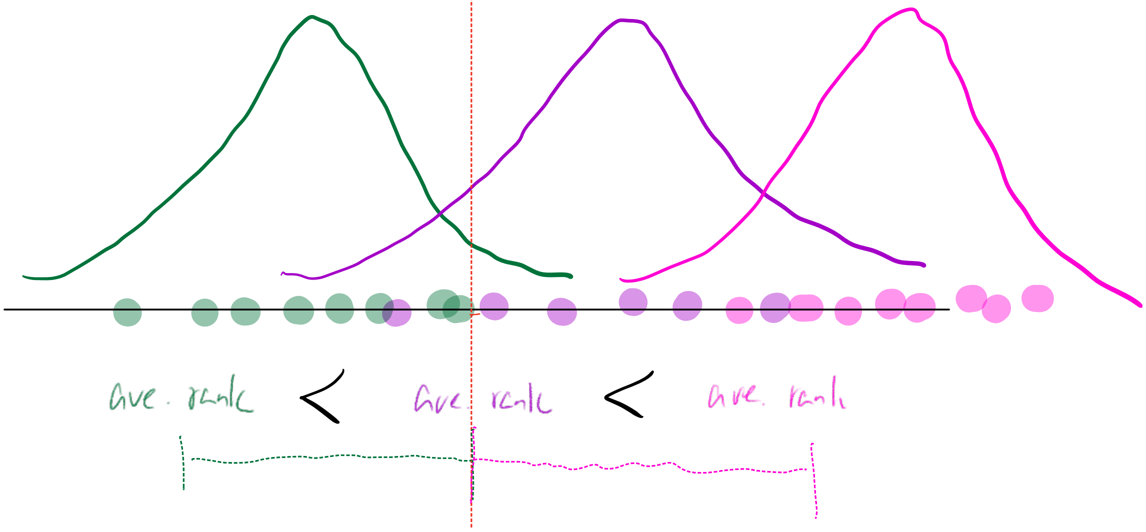


ave. rank ≈ ave. rank ≈ h

≈ common values

H small ≈ 0

if H_a is true :



ave. rank < ave. rank < ave. rank

H is large

Combine all N observations from the k samples, order them from least to greatest:

Data Layout for Ranks

Treatments	Ranks	Sample Size	Means
1	$R_{11}, R_{12}, \ldots, R_{1n_1}$	n_1	$R_{1\cdot}$ ← $\bar{X}_{1\cdot}$
2	$R_{11}, R_{12}, \ldots, R_{1n_2}$	n_2	$R_{2\cdot}$ ← $\bar{X}_{2\cdot}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
k	$R_{11}, R_{12}, \ldots, R_{1n_k}$	n_k	$R_{k\cdot}$ ← $\bar{X}_{k\cdot}$

$$\Rightarrow \quad \underbrace{R_i = \sum_{j=1}^{n_i} R_{ij}}_{\text{sum of joint ranks received by treatment i observations}}$$

$$\Rightarrow \quad \underbrace{R_{i\cdot} = \frac{R_i}{n_i}}_{\text{average of joint ranks received by treatment i observations}}$$

⇒ Under H_0, rank vector $(R_{11}, R_{12}, \dots, R_{1n_1}; \dots R_{11}, R_{12}, \dots, R_{1n_k})$ has a uniform distribution over the set of all $N!$ permutations of the rank $(1, 2, \dots N)$

⇒

$1 \dots \quad N$

$P(R_{ij} = 1) = P(R_{ij} = 2) \dots$

$= \frac{1}{N}$

$$E_0(R_i) = E_0\Big(\sum_{j=1}^{n_i} R_{ij}\Big)$$

$$= \sum_{j=1}^{n_i} E_0(R_{ij})$$

$\frac{1}{N} \times 1 + \frac{1}{N} \cdot 2 + \dots \frac{1}{N} \cdot N$

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

$= \frac{1}{\cancel{N}} \frac{(1+N)\cancel{N}}{2}$

$$= \sum_{j=1}^{n_i} \frac{\frac{N(N+1)}{2}}{N}$$

$$= \sum_{j=1}^{n_i} \frac{N+1}{2}$$

$$= n_i \frac{N+1}{2}$$

⇒

$$E_0(R_{i.}) = \frac{N+1}{2}$$

H_0 is true

exp. ave. ranks when H_0 is true

$= \frac{N+1}{2}$

Common value

we would expect average rank sum to be close to the expected value when H_0 is true.

⇒ Kruskal-Wallis statistics: [1] [2]

Scaly factor

$$H = \boxed{\frac{12}{N(N+1)}} \sum_{i=1}^{k} n_i \boxed{\left(R_{i.} - \frac{N+1}{2}\right)^2} = R_{i.}^2 - R_{i.}\frac{N+1}{2} + \left(\frac{N+1}{2}\right)^2$$

$$= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j}\right) - 3(N+1)$$

$\approx$ "SSTR"

⇒ The KS test statistic H is a constant times a weighted sum of squared differences between the observed treatment average ranks, $R_{i.}$, and their null expected values, $(N+1)/2$

- small values of H represent agreement with H_0
- When the treatment effects τ_i 's are not all equal, we would expect a portion of the associated treatment average ranks to differ from their common null expectation, with some tending to be larger and some smaller. The net result (after squaring the observed differences) would be a large value of H.
- This suggests rejecting H_0 in favor of H_1 for large values of H.

[1] The Kruskal-Wallis test can also be motivated by considering the usual analysis of variance $\mathcal{F}$ statistic calculated using the ranks, rather than the original observations. SS~~B~~T reduces to $\sum_{j=1}^{k} n_j \left(R_{.j} - (N+1)/2\right)^2$ when applied to the ranks rather than the original observations and SS~~T~~E becomes a fixed constant when calculated on the ranks. Using these facts, it can be shown that when $\mathcal{F}$ is calculated for the ranks, $\mathcal{F}$ is an increasing function of H.

[2] For the case of $k = 2$ treatments, Kruskal-Wallis test is equivalent to the two-sided Wilcoxon rank sum test.

Derivation of null distribution using permutation

$k = 3 \quad n_1 = n_2 = n_3 = 2 \;;\; N = 6$

$$\binom{6}{2}\binom{4}{2}\binom{2}{2} = \frac{6!}{2!\,2!\,2!} = \frac{6\times5\times\cancel{4}\times3\times\cancel{2}}{\cancel{2}\times\cancel{2}\times\cancel{2}} = 90$$

⇓

previously when $k = 2 \quad n_1 = n_2 = 3 \quad N = 6$

$$\binom{6}{3}\underbrace{\binom{3}{3}}_{1} = \frac{6!}{3!\times3!} = \frac{\cancel{6}\times5\times4\times\cancel{3\times2}\times1}{(\cancel{3\times2})\,1\,\cancel{3\times2}\,1} = 20$$

Large sample approximation of null distribution

$R_{i\cdot}$ = ave. ranks $\sim$ Normal( )

CLT $\left(R_{1\cdot}, R_{2\cdot} \dots R_{(k-1)\cdot}\right)$

$\sim N(\cdot\ \cdot\ \cdot)$

$R_{i\cdot} - E_0(R_{i\cdot})$ $R_{i\cdot}$

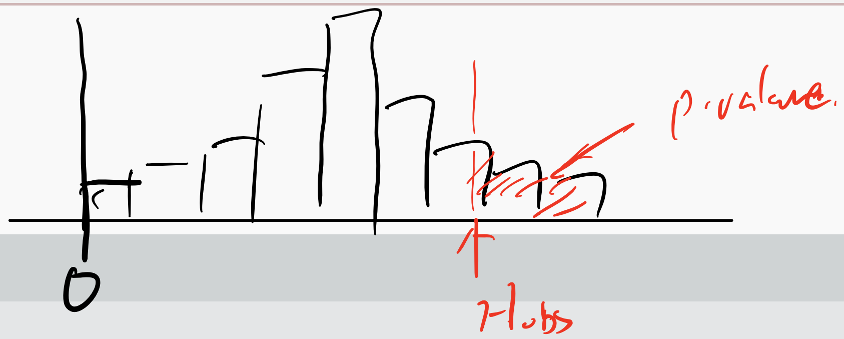
Define $T_j = R_{.j} - E_0\left(R_{.j}\right) = R_{.j} - (N+1)/2$, for $j = 1, 2, \dots, k$. As each $R_j = \sum_{i=1}^{n_j} r_{ij}/n_j$ is an average, it is not surprising (see Kruskal and Wallis (1952), e.g., for justification) that a properly standardized version of the vector $\mathbf{T}^* = \left(T_1, \dots, T_{k-1}\right)$ has an asymptotic (min $\left(n_1, \dots, n_k\right)$ tending to infinity) $(k-1)$-variate normal distribution when the null hypothesis H_0 is true.

H is a quadratic form in the variables $\left(T_1, \dots, T_{k-1}\right)$, it is therefore quite natural that H has an asymptotic (min $\left(n_1, \dots, n_k\right)$ tending to infinity) chi-square distribution with $k-1$ degrees of freedom.

$$H \sim \chi^2_{k-1}$$

in large sample.

Procedure



Permutation

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

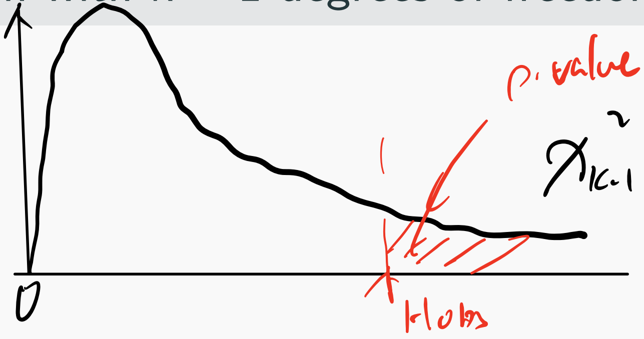
versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal }] \, ,$$

at the α level of significance, Reject H_0 if $H \geq h_\alpha$; otherwise do not reject, where the constant h_α is chosen to make the type I error probability equal to α. The constant h_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of H.

Large-sample approximation

Reject H_0 if $H \geq \chi^2_{k-1,\alpha}$; otherwise do not reject, where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k - 1$ degrees of freedom.



Example:Length of YOY Gizzard Shad

To determine the number of game fish to stock in a given system and to set appropriate catch limits, it is important for fishery managers to be able to assess potential growth and survival of game fish in that system. Such growth and survival rates are closely related to the availability of appropriately sized prey. Young-of-year (YOY) gizzard shad (Dorosoma cepedianum) are the primary food source for game fish in many Ohio environments. However, because of their fast growth rate, YOY gizzard shad can quickly become too large for predators to swallow.

Thus it is useful to know both the size structure of the resident YOY shad populations. We want to assess whether there are any differences between the median lengths for the YOY gizzard shad populations in the four 3 Kokosing Lake sites. With this in mind, Johnson (1984) sampled the YOY gizzard shad population at four different sites in Kokosing Lake (Ohio) in summer 1984.

3 subgroups

$k = 3$

$n_1 = n_2 = n_3 = 5$

$N = 15$

Site I	Site II	Site III
29(5)	60(15)	33(8)
46(13)	32(7)	26(2)
37(9)	42(10)	25(1)
31(6)	45(12)	28(4)
44(11)	52(14)	27(3)

$R_1 = 44$ $R_2 = 58$ $R_3 = 18$

$R_{1.} = \frac{44}{5}$ $R_{2.} = \frac{58}{5}$ $R_{3.} = \frac{18}{5}$

$$
\begin{aligned}
H &= \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(R_{i.} - \frac{N+1}{2}\right)^2 \\
&= \left(\frac{12}{\underset{15}{N}(\underset{15}{N}+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j {=} 5}\right) - 3(\underset{15}{N}+1) \\
&= \frac{12}{15(15+1)} \left(\frac{(44)^2}{5} + \frac{(58)^2}{5} + \frac{(18)^2}{5}\right) - 3(15+1) \\
&= 8.24 \leftarrow
\end{aligned}
$$

$$\frac{15!}{5!\,5!\,5!}$$

For the large-sample approximation:

c k-1

```
> pchisq(4.85,df=2,lower.tail = F)
[1] 0.08847812
```

> 0.05

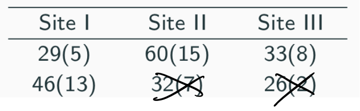
Hence, there is no ~~strong~~ evidence to reject the null hypothesis and there is no statistically significant differences between the median lengths for the YOY gizzard shad populations in the four Kokosing Lake sites.

Check with built-in function:

observation / data

groups

```
> library(NSM3)
> # permutation
> pKW(x=c(29,46,37,31,44,60,32,42,45,52,33,26,25,28,27),
+      g=c(rep(1,5),rep(2,5),rep(3,5)),method='Exact')
Group sizes: 5 5 5
Kruskal-Wallis H Statistic: 8.24
Exact upper-tail probability: 0.0077
> # large sample approximation
> pKW(x=c(29,46,37,31,44,60,32,42,45,52,33,26,25,28,27),
+      g=c(rep(1,5),rep(2,5),rep(3,5)),method='Asymptotic')
Group sizes: 5 5 5
Kruskal-Wallis H Statistic: 8.24
Asymptotic upper-tail probability: 0.0162
```



Site I	Site II	Site III
29(5)	60(15)	33(8)
46(13)	~~32(7)~~	~~26(2)~~

$n_1 : 2 \quad n_2 : 1 \quad n_3 = 1$

$k : 3 \quad N = 4$

permutation ranks : 1. 2. 3. 4 $\quad \frac{4!}{2!\,1!\,1!} = \frac{4 \times 3 \times 2}{2} = 12$

1	2	3
1 2	3	4

$R_i =$ $\Rightarrow H$

1	2	3
1 2	4	3

$R_i =$ $\Rightarrow H$

1	2	3
1 3	2	4

$R_i =$ $\Rightarrow H$

1	2	3
1 3	4	2

1	2	3
1 4	2	3

1	2	3
1 4	3	2

1	2	3
2 3	1	4

1	2	3
2 3	4	1

1	2	3
2 4	1	3

1	2	3
2 4	3	1

1	2	3
3 4	1	2

1	2	3
3 4	2	1

$\Rightarrow$

H	Prob

null dist of H