University of Washington

STATISTICS

Stat 425

Introduction to Nonparametric Statistics

Comparison of More Than Two Treatments

Fritz Scholz

Spring Quarter 2009*

*May 23, 2009

Comparative Studies

Often comparative studies involve more than two treatments, or treatment-control.

Or we have s samples obtained under different conditions and we wish to examine whether such differences are statistically significant or not.

If the differences appear insignificant one could pool the samples into a single sample that gains in impact through its larger sample size $N = n_1 + \ldots + n_s$.

Three Tranquilizers

We have three brands of tranquilizers A, B, C. Seven comparable mental patients are assigned randomly to these tranquilizer, two each to A and C and three to B.

After a month the seven patients are ranked w.r.t. the perceived treatment effect.

$$A : 2,4 \qquad B : 3,5,7 \qquad C : 1,6$$

Consider the hypothesis H_0 : no difference between the treatments.

Under H_0 the patients would have had the same respective rankings no matter how they were assigned to the treatments.

Under H_0 the random assignment of the 2 A's, 2 B's and 3 C's have equal chance to be assigned to the 7 patients and thus to their inherent ranks $1,2,\ldots,7$.

What is that chance for any such assignment?

Counting the Possibilities

The number of possible choices for ordered rank pairs in group *A* is $\binom{7}{2} = 21$

12	13	14	15	16	17	23
24	25	26	27	34	35	36
37	45	46	47	56	57	67

For each such choice for group *A*, there remain five ranks to choose from for *B*.

For example, when the *A*-ranks are 13, the *B*-ranks must be chosen from 24567.

See first row in the table below. As further illustrations, the second and third rows show the possible *B*-ranks when the *A*-ranks are 16 and 23, respectively.

A-ranks	Possible choices for *B*-ranks									
13	245	246	247	256	257	267	456	457	467	567
16	234	235	237	245	247	257	345	347	357	457
23	145	146	147	156	157	167	456	457	467	567

Counting the Possibilities (continued)

For each of the 21 choices of two ranks for A there are $\binom{5}{3} = 10$ choices of three ranks for B.

That amounts to $21 \times 10 = 210$ combined choices of ranks for A and B.

Once these choices have been made, there is only one choice to give the remaining two ranks to C.

Thus the total number of ordered rank allocations (2 to A, 3 to B, and 2 to C) is $21 \times 10 \times 1 = 210$.

Under H_0 the chance for each one of these allocations is $1/210$ based on our initial random assignment of 2 A's, 3 B's and 2 C's to the subjects at hand.

Generalizing

Suppose we have N subjects and s treatments.

We want to assign n_i of these subjects to treatment i, where $i = 1, \ldots, s$.

Using all subjects and each subject just once we must have $n_1 + \ldots + n_s = N$.

There are then $\binom{N}{n_1,\ldots,n_s}$ such possible assigments where

$$\begin{aligned}
\binom{N}{n_1,\ldots,n_s} &= \binom{N}{n_1} \times \binom{N-n_1}{n_2} \times \ldots \times \binom{n_{s-1}+n_s}{n_{s-1}} \\
&= \frac{N!}{n_1! \times (N-n_1)!} \times \frac{(N-n_1)!}{n_2! \times (N-n_1-n_2)!} \times \ldots \times \frac{(n_{s-1}+n_s)!}{n_{s-1}! \times n_s!} \\
&= \frac{N!}{n_1! \times \ldots \times n_s!}
\end{aligned}$$

$\binom{N}{n_1,\ldots,n_s}$ is referred to as the multinomial coefficient.

Randomization

Again we assign the subjects at random to the s treatments,

in group sizes $n_1, \ldots, n_s$, with $n_1 + \ldots + n_s = N$.

The subjects are ranked according to some measure of treatment effectiveness.

This can be subjective or be based on some numerical score or measurement.

Our hypothesis H_0: there is no difference between the s treatments.

Under H_0 all subject rankings are preordained (not influenced by the treatments).

Under H_0 each split of the ranks $1, 2, \ldots, N$ into groups of respective sizes $n_1, \ldots, n_s$

is equally likely with probability $1/\binom{N}{n_1, \ldots, n_s}$ each.

Denote the set of ordered ranks for the s groups by

$$R_{11} < \ldots < R_{1n_1}\ , \quad R_{21} < \ldots < R_{2n_2}\ , \quad \ldots\ , \quad R_{s1} < \ldots < R_{sn_s}$$

The Basic Null Distribution of Ranks

$$P_{H_0}\left(R_{1\,1}=r_{1\,1},\ldots,R_{1\,n_1}=r_{1\,n_1}, \quad \ldots, \quad R_{s\,1}=r_{s\,1},\ldots,R_{s\,n_s}=r_{s\,n_s}\right)=\frac{1}{\binom{N}{n_1,\ldots,n_s}}$$

This generalized our previous null distribution of ranks in the case of $s=2$.

This distribution generates the null distributions of all derived rank statistics.

The Growth of $\binom{N}{n_1,\ldots,n_s}$

Full enumeration of all possible rankings becomes quickly unwieldy.

$$\begin{aligned}\binom{15}{5,5,5} &= \mathtt{choose}(15,5) * \mathtt{choose}(10,5)\\ &= 3003 * 252 = 756756\\ \binom{18}{6,6,6} &= \mathtt{choose}(18,6) * \mathtt{choose}(12,6)\\ &= 18564 * 924 = 17153136\\ \binom{16}{4,4,4,4} &= \mathtt{choose}(16,4) * \mathtt{choose}(12,4) * \mathtt{choose}(8,4)\\ &= 1820 * 495 * 70 = 63063000\end{aligned}$$

What Alternatives to H_0?

When testing H_0 one should be guided by the anticipated alternatives.

In the case of $s = 2$ we focussed first on the general level of the ranks

in the two groups $\Longrightarrow$ Wilcoxon rank-sum test.

Next we focussed on changes of dispersion of ranks

$\Longrightarrow$ Siegel-Tukey test and Ansari-Bradley test.

Finally we considered all possible ways for ranks to express differences

$\Longrightarrow$ Kolmogorov-Smirnov test.

The Kruskal-Wallis Test

We will deal first with the changes in rank levels from treatment group to treatment group.

Express the rank level in each group by the average group rank

$$R_{i\bullet} = \frac{R_{i1} + \ldots + R_{in_i}}{n_i} = \frac{R_i}{n_i} \quad \text{for} \quad i = 1, 2, \ldots, s$$

If there is little variation between these average ranks they would all be close to

$$R_{\bullet\bullet} = \frac{R_{11} + \ldots + R_{1n_1} + \ldots + R_{s1} + \ldots + R_{sn_s}}{N} = \sum_{i=1}^{s} \frac{n_i}{N} R_{i\bullet} = \frac{1}{N} \sum_{i=1}^{s} R_i = \frac{N+1}{2}$$

This motivates the Kruskal-Wallis test statistic

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{s} \left(R_{i\bullet} - \frac{N+1}{2} \right)^2$$

We reject H_0 when $K \geq c$ for appropriate critical values c.

The factor $\frac{12}{N(N+1)}$ facilitates a simple large sample approximation for the null distribution of K.

Some Comments

For $s = 2$ this test is equivalent to the two-sided Wilcoxon rank-sum test.

Alternate computational expression (no longer so relevant):

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{s} \frac{R_i^2}{n_i} - 3(N+1)$$

In principle the computation of the null distribution for K is straightforward, based on the null distribution of the sets of ordered ranks, all equally likely.

“Simply” evaluate K for all splits of $1, 2, \ldots, N$ into s rank subsets of respective sizes $n_1, \ldots, n_s$.

However, the volume of these evaluations grows quickly beyond practical bounds.

KW3

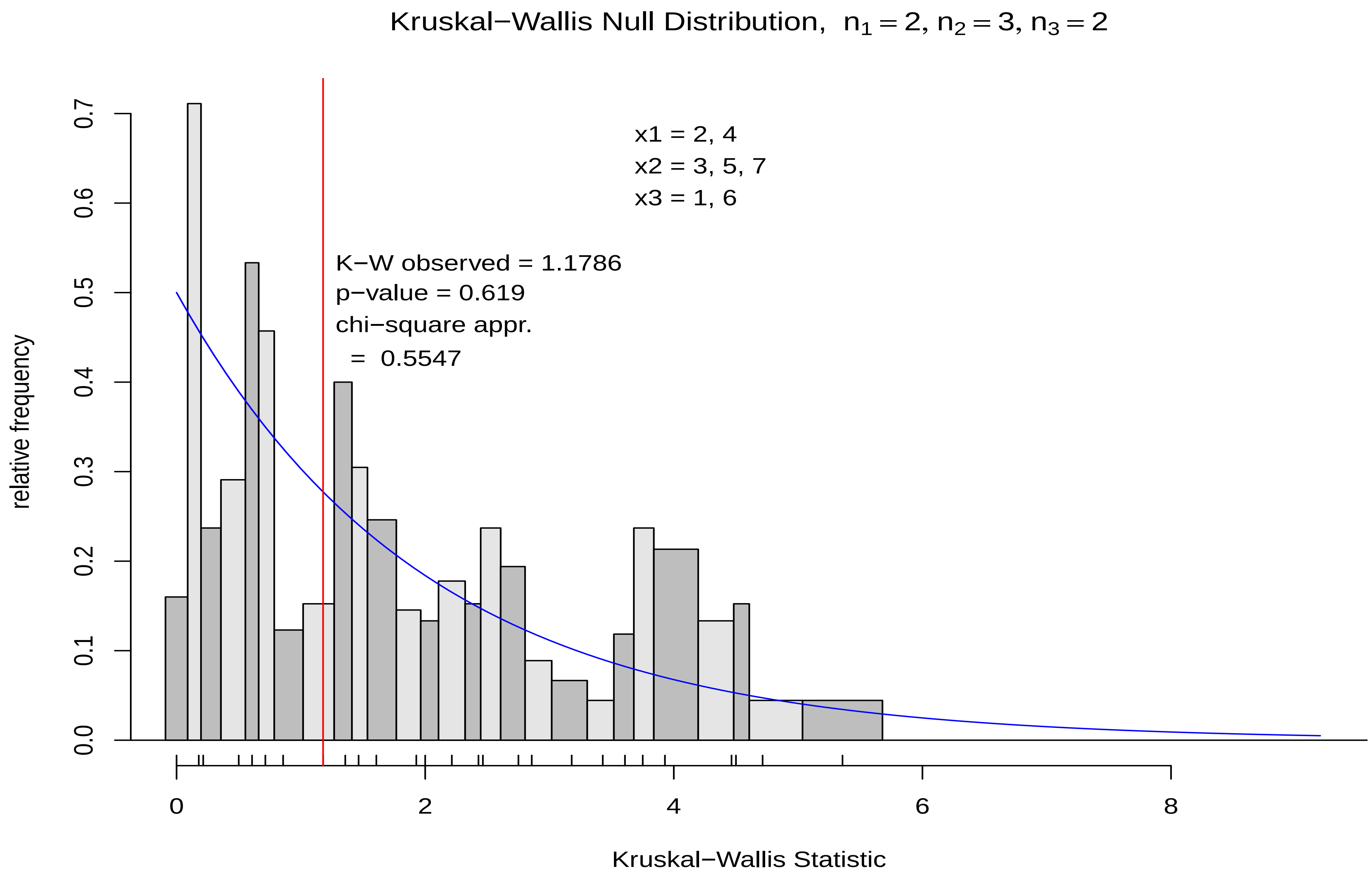
For $s=3$ treatment groups the R function `KW3` (see class web site) implements the complete enumeration of the Kruskal-Wallis test null distribution.

It either provides the exact p-value for K_{obs} or it gives the tail probability for a given critical value c and the implied group sizes of the input list of three sets of treatment group scores.

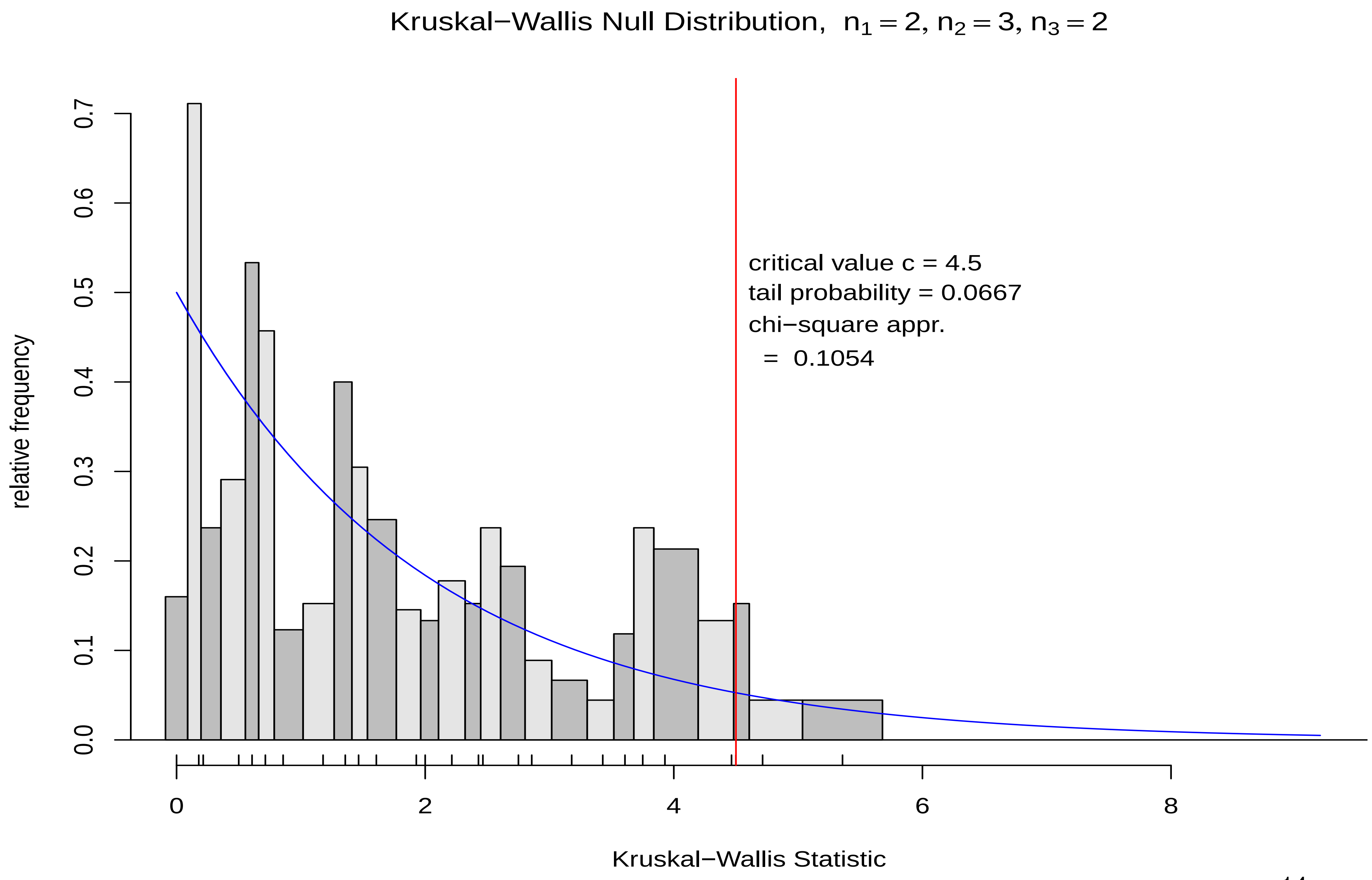
This covers and extends the territory of Table I in the Text. Table I covers tail probabilities $\le .15$ ($\le .2$ in some extreme cases) for group sizes $n_i \le 5$, $i = 1,2,3$.

For $n_1 = n_2 = n_3 = 5$ the full enumeration amounts to 756756 cases.
For $n_1 = n_2 = 5, n_3 = 6$ the full enumeration amounts to 2018016 cases.
For $n_1 = n_2 = n_3 = 6$ the full enumeration amounts to 17153136 cases.
This laptop was still able to allocate `x=rep(0,17153136)` but with much disk drive activity, i.e., it was using virtual memory (not RAM).

KW3: Three Tranquilizers



KW3: Three Tranquilizers with Critical Value



Large Group Size Approximation

For large group sizes $n_1,\ldots,n_s$ the null distribution of K becomes approximately a chi-square distribution with $s-1$ degrees of freedom, i.e., $K \approx \chi^2_{s-1}$.

For $s=3$ $(s-1=2)$ this distribution is an exponential distribution with mean 2, i.e., with density $f(x)=\exp(-x/2)/2$ for $x \geq 0$. It is overlaid in the previous two slides.

When $Z_1,\ldots,Z_f$ are i.i.d. $\sim \mathcal{N}(0,1)$ then $Z_1^2+\ldots+Z_f^2$ is said to have a chi-square distribution with f degrees of freedom.

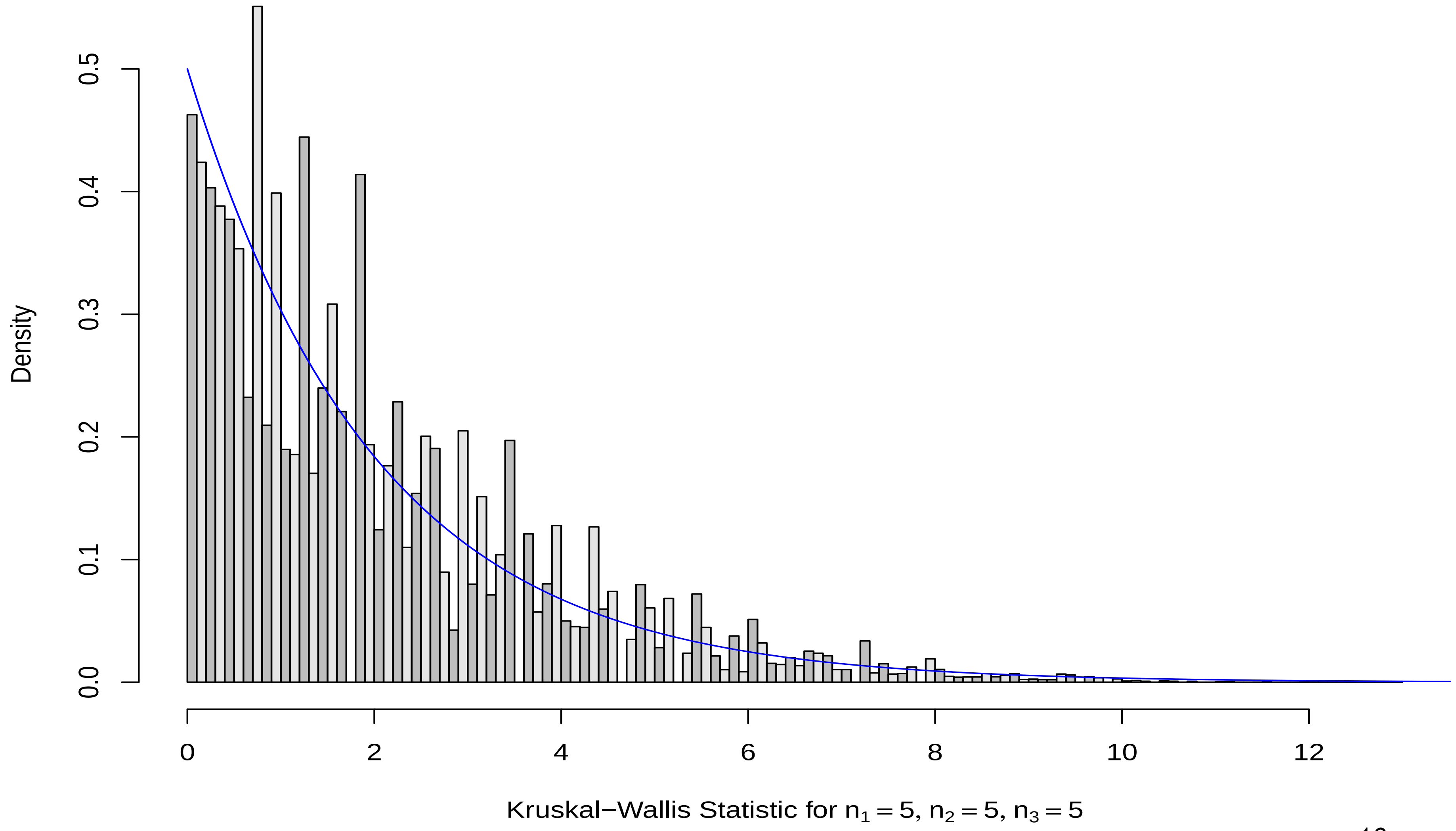
$$n_i\left(R_{i\bullet}-\frac{N+1}{2}\right)^2=\frac{1}{n_i}\left(R_i-\frac{n_i(N+1)}{2}\right)^2$$

look like squared, approximately normal random variables with zero means.

However, they are not all independent since $\sum_{i=1}^s R_i = N(N+1)/2$,

hence the loss of one degree of freedom.

Chi-Square Approximation



Ties

In case of ties we use the midrank vector in all calculations.

The formula for the Kruskal-Wallis test statistic K changes to

$$K^* = \frac{12/[N(N+1)]\sum_{i=1}^{s} R_i^{*2}/n_i - 3(N+1)}{1 - \sum_{i=1}^{e}(d_i^3 - d_i)/(N^3 - N)}$$

Here e denotes the number of distinct values in the pooled set of all N scores.

d_i is the multiplicity of the i^{th} smallest of those distinct values, $i = 1, \ldots, e$.

R_i^* is the sum of midranks for the i^{th} treatment group.

The denominator $d_{\text{fac}} = 1 - \sum_{i=1}^{e}(d_i^3 - d_i)/(N^3 - N)$ reduces to 1 when there are no ties and K^* reverts back to K.

Again we have $K^* \approx \chi^2_{s-1}$ for large $n_1, \ldots, n_s$. `KW3` works in case of ties.

Simulated Null Distribution

The null distribution is easily simulated in a loop (here just illustrated for $s = 3$)

```
z=c(x1,x2,x3)
nvec=c(length(x1),length(x2),length(x3))
rz=rank(z); N=length(rz)
out=rep(0,Nsim)
for(i in 1:Nsim){
 rzi=sample(rz,replace=F); K=0; jx=0
 for(j in 1:3){
  K=K+sum(rzi[jx+1:nvec[i]])/nvec[i]; jx=jx+nvec[i]
 }
 out[i]=12*K/(N*(N+1))-3*(N+1)}
```

here `out` is a vector of `Nsim` randomly generated `K` statistics.

Note how easily it also handles tied ranks through the midrank vector `rz`.

Of course, K still needs to be divided by d_{fac}.

`KW.sim`

Such a simulation is implemented more generally in the R function `KW.sim` using a default value `Nsim=10000` (see class web site).

Its usage is documented internally.

The basic inputs are a list of treatment score vectors and `Nsim`.

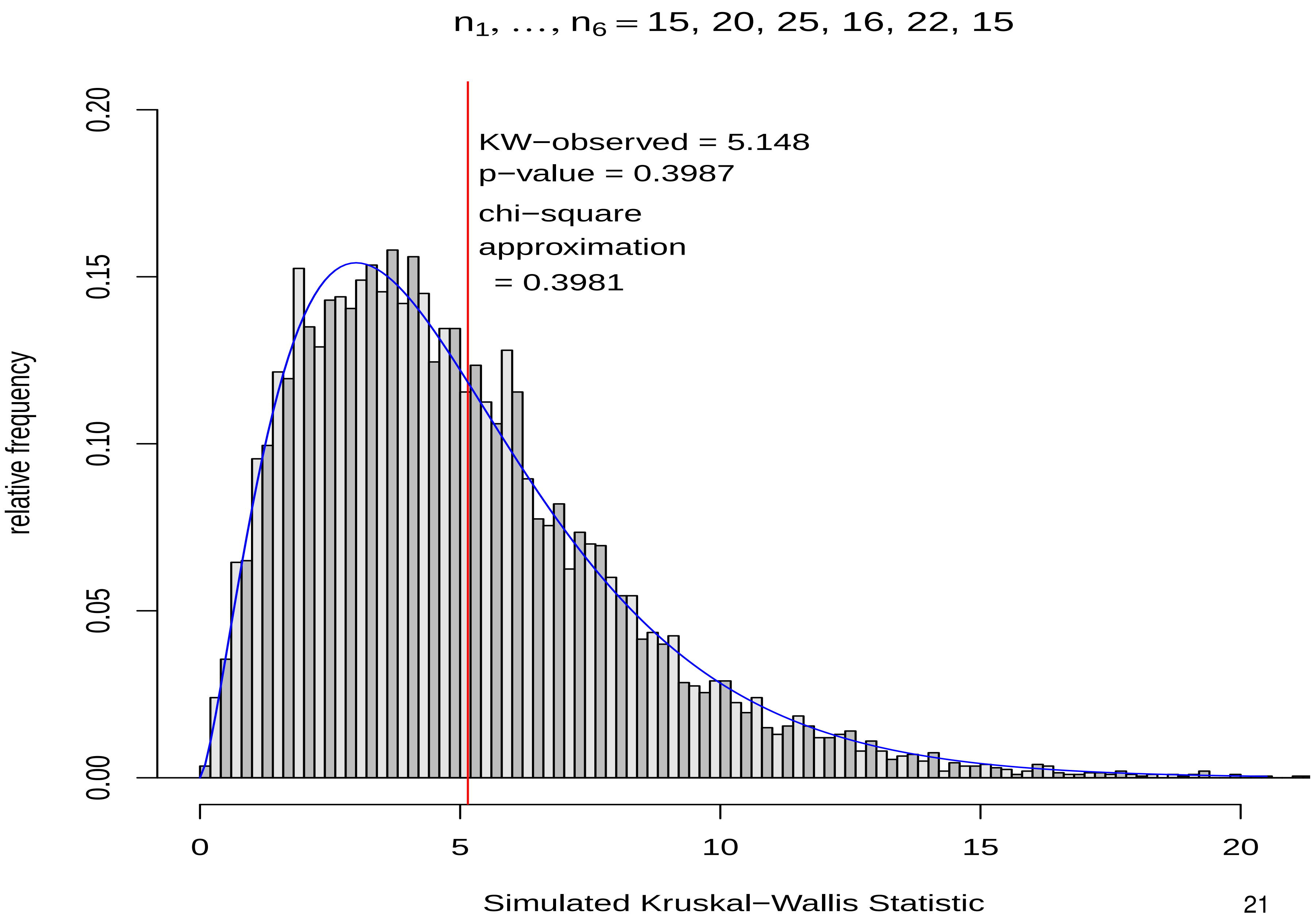
It returns the p-value for the computed K_{obs} or the upper tail probability for a given critical value c.

It also poduces the plots shown on the next slides.

The Data for the Next Slide

```
> z1=rnorm(15)
> z2=rnorm(20)
> z3=rnorm(25)
> z4=rnorm(16)
> z5=rnorm(22)
> z6=rnorm(15)
> KW.sim(list(z1,z2,z3,z4,z5,z6),PDF=T)
       KW.observed            p-value chi-square approx.
          5.148224           0.398700           0.398100
> length(c(z1,z2,z3,z4,z5,z6))
[1] 113
> length(unique(c(z1,z2,z3,z4,z5,z6)))
[1] 113  # NO TIES
```

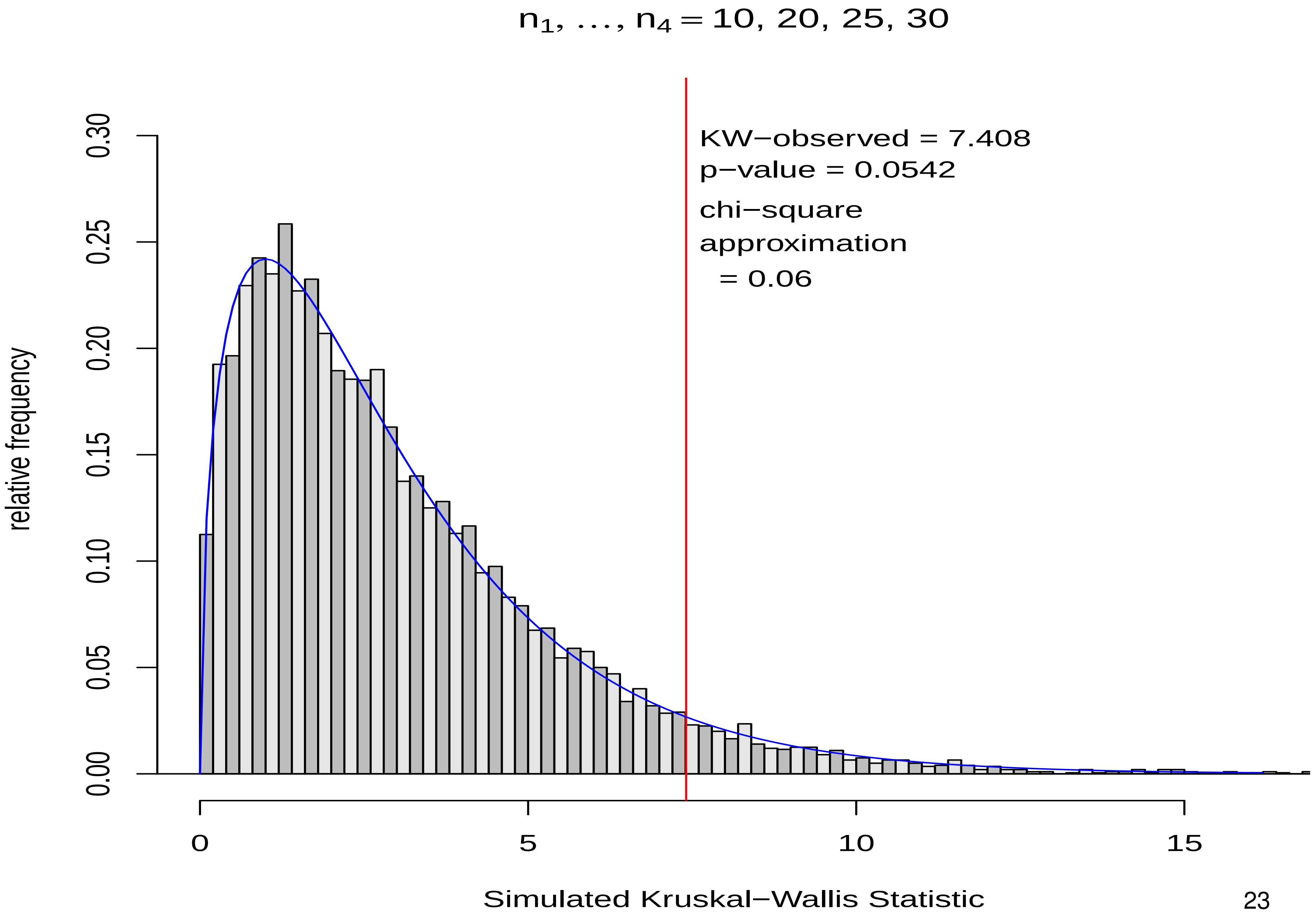
Chi-Square Approximation



The Data for the Next Slides

```
> y1=round(rnorm(10),1); sort(y1)
 [1] -2.3 -1.7 -1.1 -0.8 -0.5  0.3  0.3  0.4  0.5  0.8
> y2=round(rnorm(20),1); sort(y2)
 [1] -1.3 -1.3 -1.0 -0.9 -0.7 -0.5 -0.4 -0.4 -0.3 -0.1  0.1  0.2  0.6
[14]  0.6  0.6  0.9  0.9  1.1  1.7  1.7
> y3=round(rnorm(25,-.5),1); sort(y3)
 [1] -2.1 -2.1 -2.0 -1.7 -1.7 -1.7 -1.6 -1.5 -1.2 -0.9 -0.8 -0.7 -0.7
[14] -0.6 -0.6 -0.6 -0.4 -0.3 -0.2  0.1  0.3  0.3  0.5  0.6  0.7
> y4=round(rnorm(30),1); sort(y4)
 [1] -2.2 -1.5 -1.1 -1.1 -1.0 -1.0 -1.0 -1.0 -0.9 -0.8 -0.8 -0.6 -0.5
[14] -0.5 -0.3 -0.3 -0.2 -0.2 -0.2  0.0  0.1  0.2  0.2  0.5  0.6  0.9
[27]  1.1  1.2  1.2  1.8
> KW.sim(list(y1,y2,y3,y4),PDF=T)
      KW.observed             p-value chi-square approx.
         7.407918            0.054200            0.060000
> length(c(y1,y2,y3,y4))
[1] 85
> length(unique(c(y1,y2,y3,y4)))
[1] 34 # QUITE A FEW TIES
```

Chi-Square Approximation

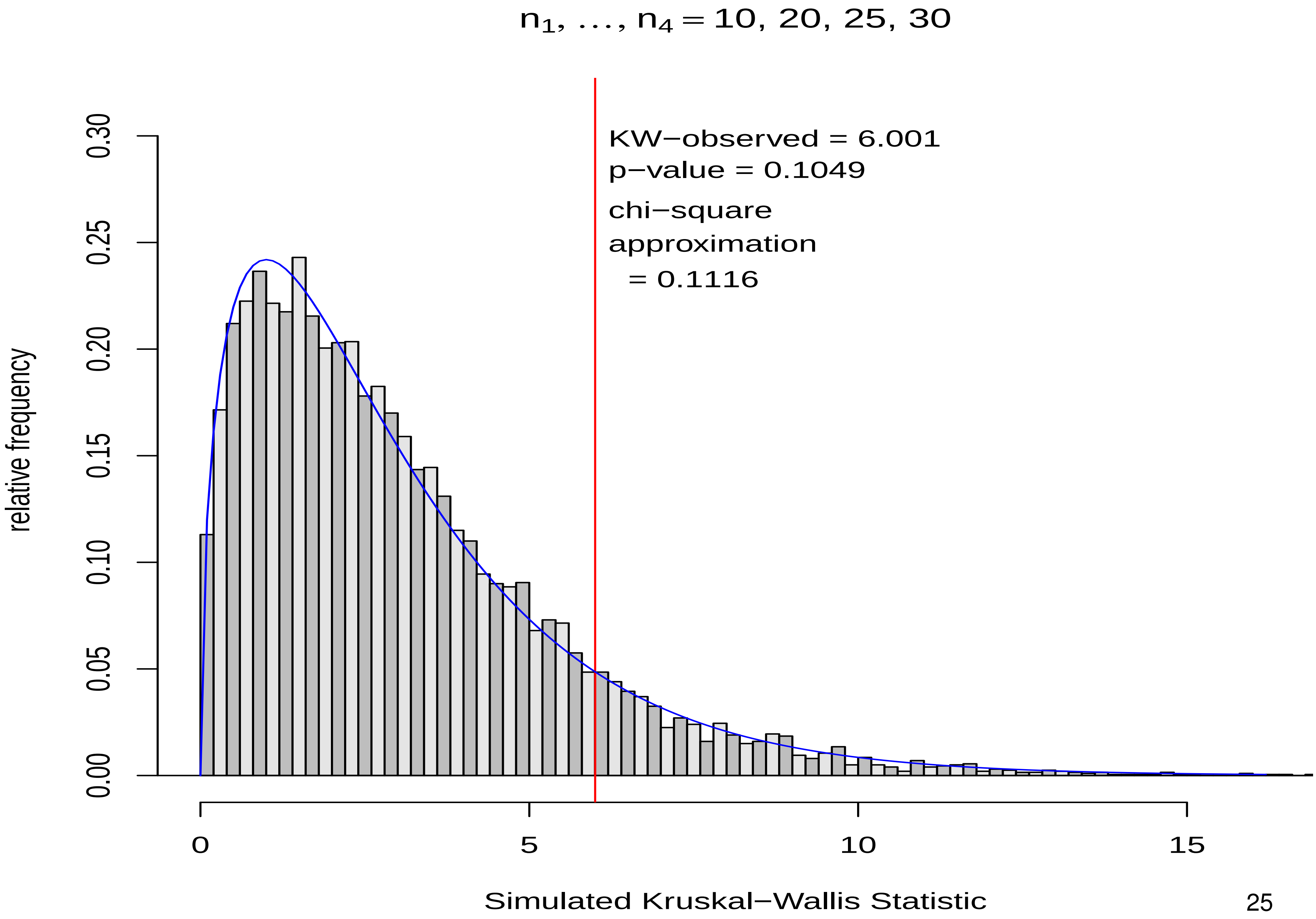


The Data for the Next Slide

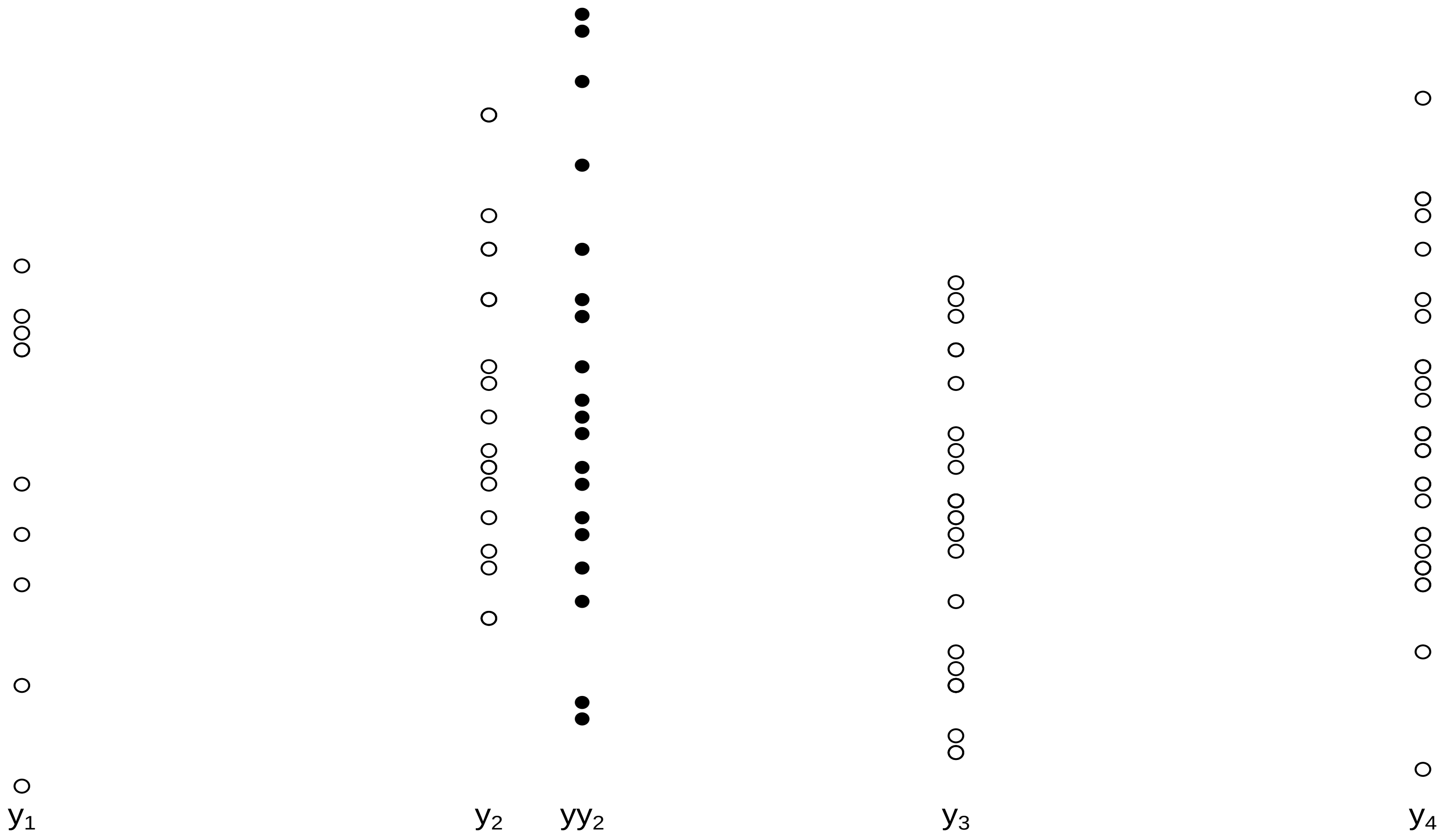
```
> yy2=round(rnorm(20,.2),1) # y1, y3, y4 as before
> sort(yy2)
 [1] -1.9 -1.8 -1.2 -1.0 -0.8 -0.7 -0.5 -0.4 -0.2 -0.1  0.0  0.2  0.5
[14]  0.5  0.6  0.9  1.4  1.9  2.2  2.3
> KW.sim(list(y1,yy2,y3,y4),PDF=T)
       KW.observed            p-value chi-square approx.
           6.00119            0.10490            0.11160
> length(unique(c(y1,yy2,y3,y4)))
[1] 38 # QUITE A FEW TIES
```

The p-value increased even though we shifted the mean of `yy2` away from zero.

Chi-Square Approximation



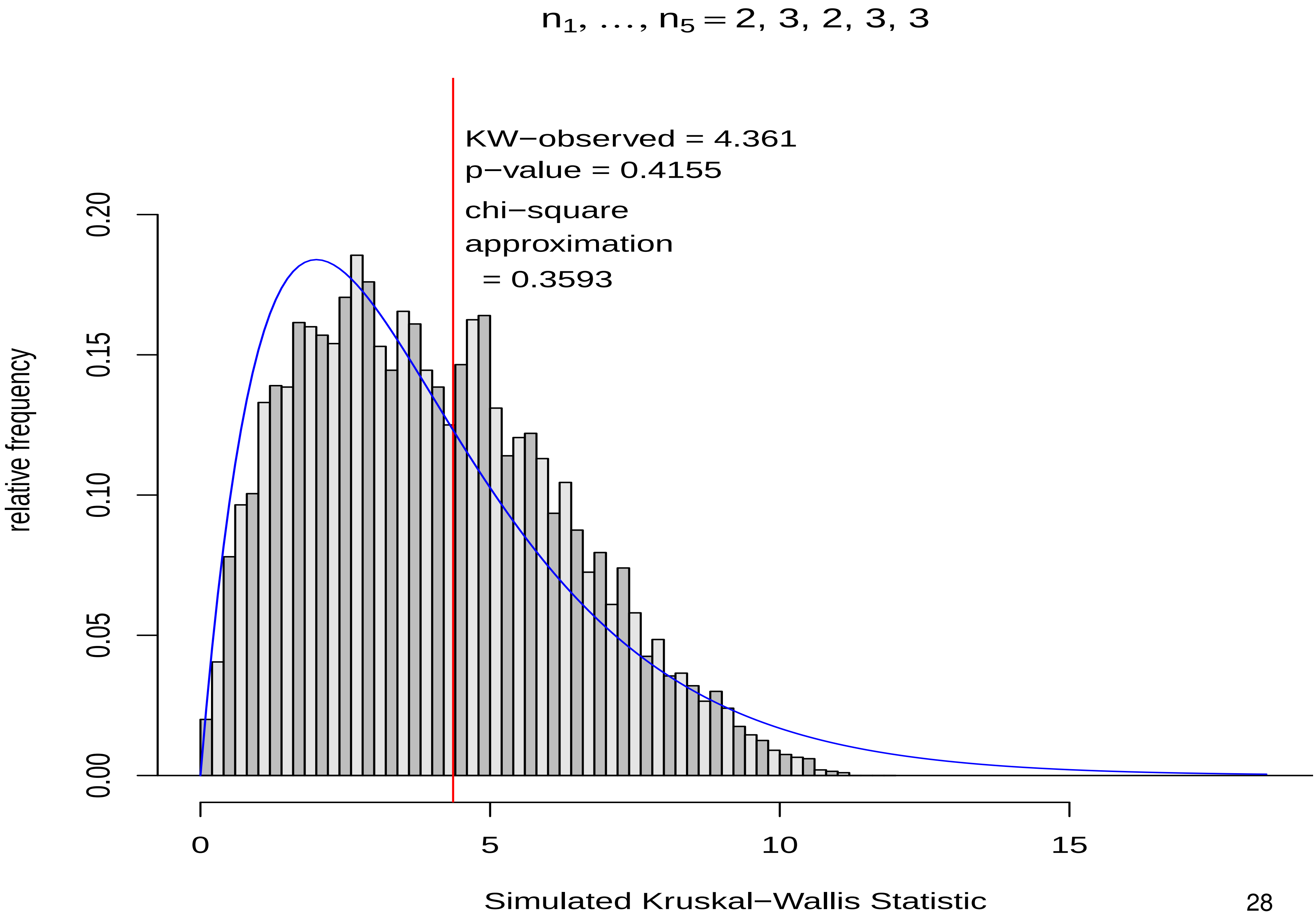
Data Plot (with Ties)



The Data for the Next Slide

```
> x1
[1] 2 4
> x2
[1] 3 5 7
> x3
[1] 1 6
> x4
[1] 2 6 8
> x5
[1] 5 8 9
> KW.sim(list(x1,x2,x3,x4,x5),PDF=T)
       KW.observed            p-value chi-square approx.
          4.361111           0.415500           0.359300
> length(unique(c(x1,x2,x3,x4,x5)))
[1] 9
```

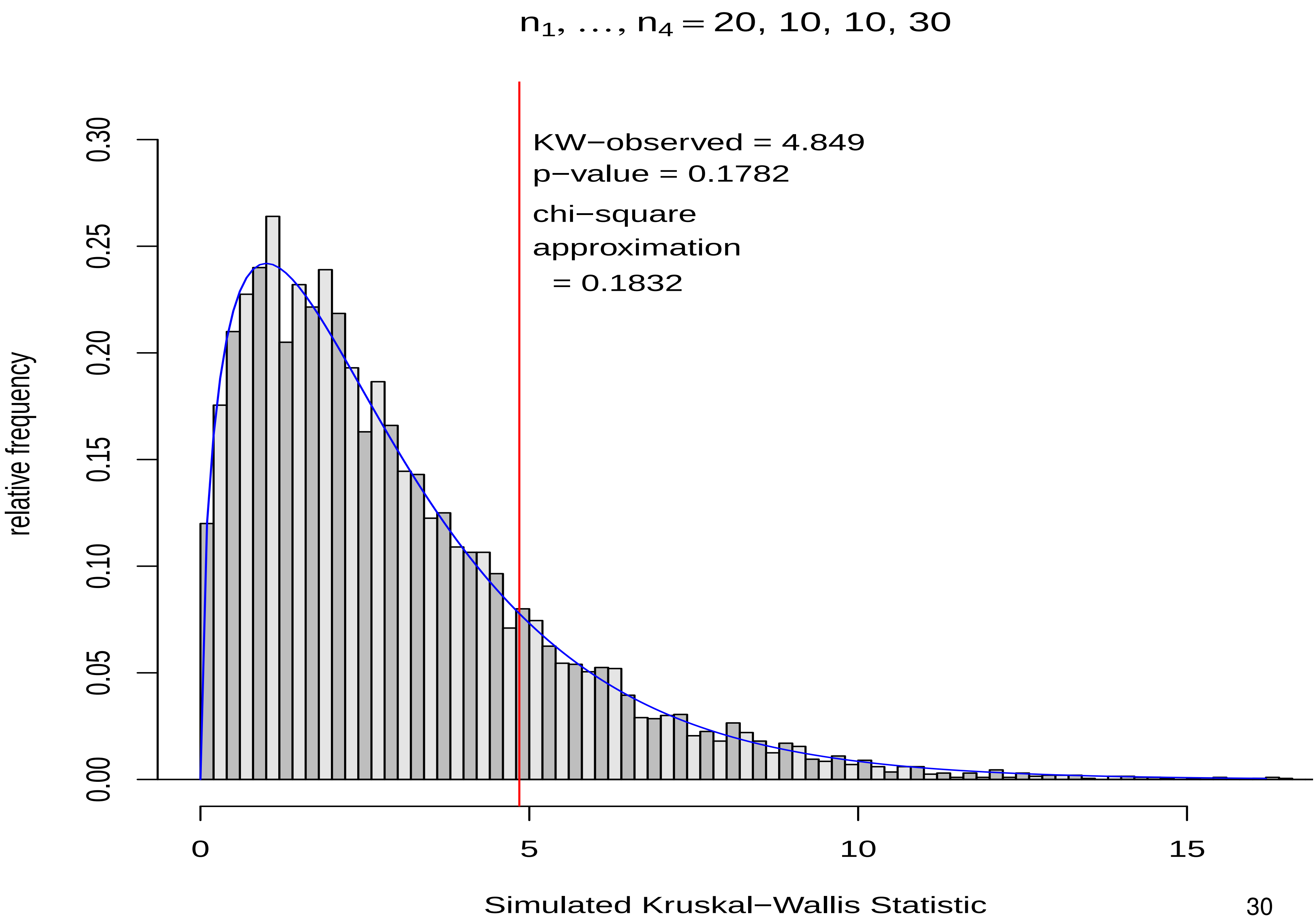
Chi-Square Approximation



The Data for the Next Slide

```
> u1=round(rnorm(20),0);  sort(u1)
 [1] -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  1  1  1  1  1  2  3
> u2=round(rnorm(10,.5),0); sort(u2)
 [1] -1  0  0  0  0  1  1  1  2  3
> u3=round(rnorm(10,-.5),0);  sort(u3)
 [1] -1 -1 -1 -1 -1  0  0  0  1  1
> u4=round(rnorm(30),0); sort(u4)
 [1] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  1
[21]  1  1  1  1  1  1  1  1  1  2
> length(unique(c(u1,u2,u3,u4)))
[1] 5
> KW.sim(list(u1,u2,u3,u4))
       KW.observed            p-value chi-square approx.
          4.848572           0.178200           0.183200
```

Chi-Square Approximation



kruskal.test

```
> kruskal.test(list(u1,u2,u3,u4))

Kruskal-Wallis rank sum test


data:  list(u1, u2, u3, u4)
Kruskal-Wallis chi-squared = 4.8486, df = 3, p-value = 0.1832
```

The intrinsic R function `kruskal.test` uses the chi-square approximation to calculate p-values.

Comments

In all the previous applications of `KW.sim` we used the default `Nsim=10000`.

The chi-square approximation appears to remain valid even for strongly tied data as long as the sample sizes are not too small.

K measures the overall discrepancy of the sample rank averages $R_{i\bullet}$ from the grand average of all ranks, i.e., $(N+1)/2$

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{s} \left(R_{i\bullet} - \frac{N+1}{2} \right)^2$$

It will be sensitive to level changes in the ranks, but not to dispersion changes.

This is the same behavior as was seen w.r.t. the Wilcoxon rank-sum test.

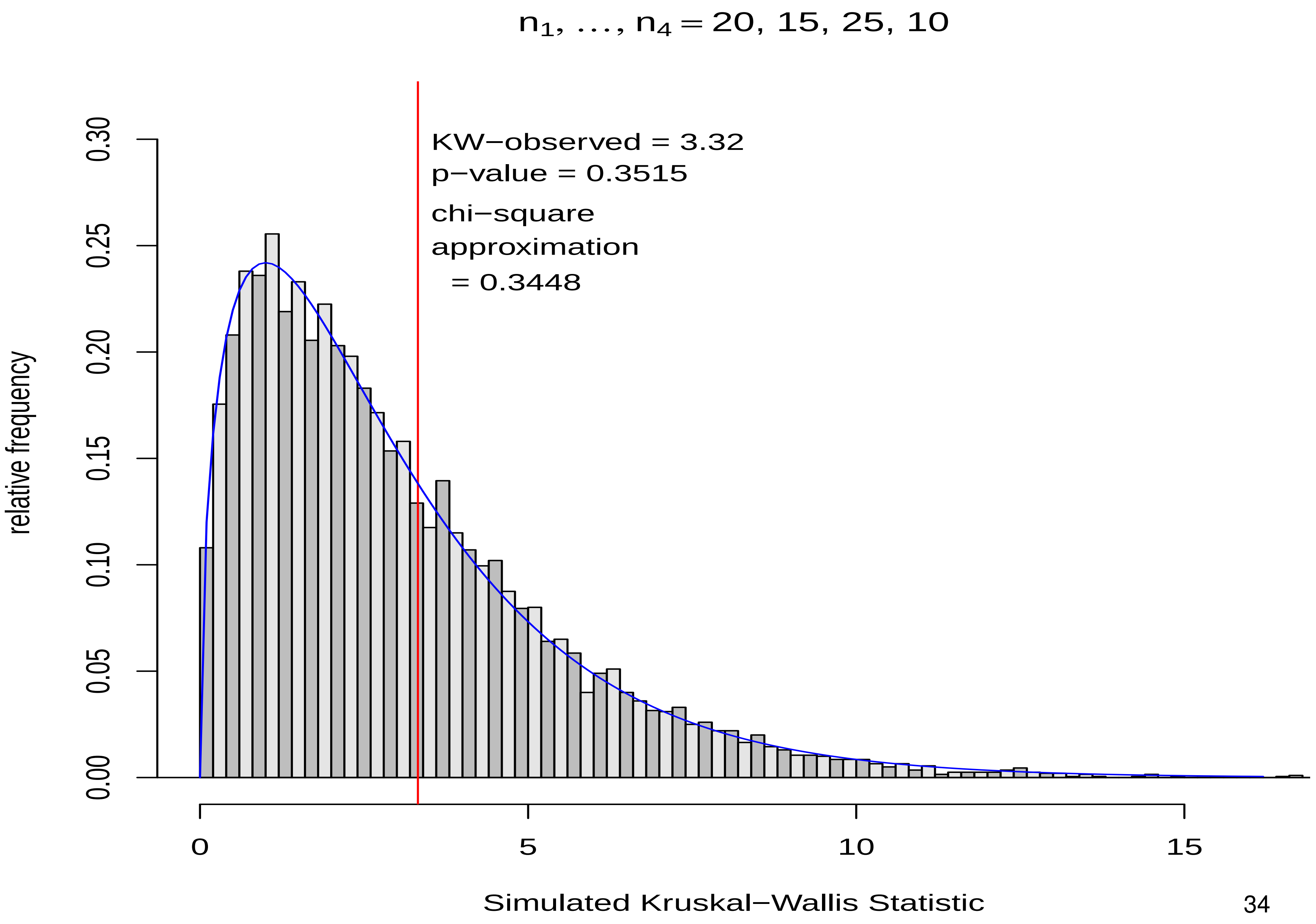
Insensitivity to Dispersion Changes

```
> s1=rnorm(20,0,1)
> s2=rnorm(15,0,3)
> s3=rnorm(25,0,2)
> s4=rnorm(10,0,1)
> KW.sim(list(s1,s2,s3,s4))
       KW.observed            p-value chi-square approx.
          3.320499           0.344100           0.344800
```

The simulated null distribution is still well approximated by the χ^2_3 distribution.

However, K_{obs} does not stand out. K is not sensitive the scale changes.

Chi-Square Approximation



KW−observed = 3.32
p−value = 0.3515
chi−square
approximation
= 0.3448

Population Model

We dealt with a limited set of N subjects and s treatments were randomly assigned to $n_1, \ldots, n_s$ of them, $n_1 + \ldots + n_s = N$. Conclusions are limited to these subjects.

Now consider s random samples from populations with respective CDF's $F_1, \ldots, F_s$.

Our null hypothesis is $H_0 : F_1 = \ldots = F_s$ without specifying the common CDF F.

In the context of s treatments we can consider a random sample from a population and randomly assign s treatments to $n_1, \ldots, n_s$ of them, $n_1 + \ldots + n_s = N$.

This is equivalent to getting independent random samples of such sizes from s distinct treatment populations with respective CDF's $F_1, \ldots, F_s$.

Distribution of Ranks

Assume a continuous population, probability of ties is zero.

Under H_0 the distribution of the ranks of the pooled observations is the same as in our randomization model.

$\Longrightarrow$ The Kruskal-Wallis test is applicable with the same null distribution.

When ties are a possibility we can enter the same discussion as in the population model for the Wilcoxon rank-sum test in case of ties.

We simply perform the KW-test conditionally given the pattern of ties.

The overall significance level $\quad \leq \quad$ maximum conditional significance level.

The Anderson-Darling k-Sample Test

Test $H_0 : F_1 = \ldots = F_k$, i.e., all k samples[*] come from a common distribution F.

Estimate $F_i(x)$ by the i^{th} sample distribution function, i.e., by its EDF $\hat{F}_i(x)$

Estimate the common cdf $F(x)$ by the EDF $\hat{F}(x)$ of all samples combined.

Under H_0 we expect that the $\hat{F}_i(x)$ should not differ much from $\hat{F}(x)$.

Compare $\hat{F}_i(x), i = 1, \ldots, k$, and $\hat{F}(x)$ via the Anderson-Darling discrepancy metric

$$AD_k = \sum_{i=1}^{k} n_i \int_B \frac{[\hat{F}_i(x) - \hat{F}(x)]^2}{\hat{F}(x)(1 - \hat{F}(x))}\, d\hat{F}(x) = \sum_{i=1}^{k} \frac{n_i}{N} \sum_{r=1}^{N-1} \frac{[\hat{F}_i(Z_r) - \hat{F}(Z_r)]^2}{\hat{F}(Z_r)(1 - \hat{F}(Z_r))}$$

where B denotes the set of all x for which $\hat{F}(x) < 1$

Assuming no ties $Z_1 < \ldots < Z_N$ denote the ordered combined sample values.

Reject H_0 for large AD_k.

[*] $k = s$ here

The AD_k Test Is a Rank Test

Assume that all N observation $Y_{i\ell}, \ell = 1, \ldots, n_i,\; i = 1, \ldots, k$ are distinct (no ties). From the second and computational form of AD_k one can see that it depends on the observations $Y_{i\ell}$ only through its ranks.

This becomes clear when looking at $\hat{F}_i(Z_r)$ which is the proportion of $Y_{i\ell}$ values that are $\le Z_r$, i.e., only the rank of the $Y_{i\ell}$ matters in such comparisons, since

$$Y_{i\ell} \le Z_r \iff \mathrm{rank}(Y_{i\ell}) \le \mathrm{rank}(Z_r) = r \iff R_{i\ell} \le r$$

Some thought makes clear that the argument stays the same in the case of ties.

The Package `adk`

For R code to carry out the AD_k test install package adk and see `?adk.test` after invoking `library(adk)` for each new R session.

`adk` uses an approximate null distribution derived under the assumption that $n_i \longrightarrow \infty$ for $i = 1, \ldots, k$. The approximation is quite reasonable when $n_i \geq 5, i = 1, \ldots, k$.

The exact null distribution (conditionally even in the case of ties) is easily estimated via simulation. However, that is not yet implemented in `adk`.

Anderson-Darling Test for Laboratory Comparisons

Comparison of four laboratories. Following are four sets of eight measurements each of the smoothness of a certain type of paper, obtained in four different laboratories:[*]

Laboratory								
A	38.7	41.5	43.8	44.5	45.5	46.0	47.7	58.0
B	39.2	39.3	39.7	41.4	41.8	42.9	43.3	45.8
C	34.0	35.0	39.0	40.0	43.0	43.0	44.0	45.0
D	34.0	34.8	34.8	35.4	37.2	37.8	41.2	42.8

Test whether there is any difference between laboratories.

[*] Part of the data from Mandel, *The Statistical Analysis of Experimental Data*, Wiley, Interscience, New York. 1964. Table 13.3.

Data Preparation and `adk.test` Call

```
> laboratory.list=list(
+ x1=c(38.7,41.5,43.8,44.5,45.5,46.0,47.7,58.0),
+ x2=c(39.2,39.3,39.7,41.4,41.8,42.9,43.3,45.8),
+ x3=c(34.0,35.0,39.0,40.0,43.0,43.0,44.0,45.0),
+ x4=c(34.0,34.8,34.8,35.4,37.2,37.8,41.2,42.8))

> adk.test(laboratory.list}
```

adk.test Output

```
Anderson-Darling k-sample test.

Number of samples:  4
Sample sizes: 8 8 8 8
Total number of values: 32
Number of unique values: 29

Mean of Anderson Darling Criterion: 3
Standard deviation of Anderson Darling Criterion: 1.20377

T = (Anderson Darling Criterion - mean)/sigma

Null Hypothesis: All samples come from a common population.

                    t.obs P-value extrapolation
not adj. for ties 4.44926 0.00236             1
adj. for ties     4.47978 0.00228             1
```

`kruskal.test` Output

```
> kruskal.test(laboratory.list)

Kruskal-Wallis rank sum test

data:  laboratory.list
Kruskal-Wallis chi-squared = 12.8757, df = 3, p-value = 0.004913
```

Based 20000 simulations the estimated p-values for `adk.test` were $.00150$ ($.00155$),

for `kruskal.test` it was $.00185$

for the randomization version of the standard F-test it was $.00165$.

In 10^6 simulations the estimated p-value of `kruskal.test` was $.002092$.