12.11: Removing the Normality Assumption

Now that we've seen how to check for normality, we are led naturally to ask what we can do to address violations of normality. In the context of a one-way ANOVA, the easiest solution is probably to switch to a non-parametric test (i.e., one that doesn't rely on any particular assumption about the kind of distribution involved). We've seen non-parametric tests before, in Chapter 13: when you only have two groups, the Wilcoxon test provides the non-parametric alternative that you need. When you've got three or more groups, you can use the ***Kruskal-Wallis rank sum test*** (Kruskal and Wallis 1952). So that's the test we'll talk about next.

12.11.1 logic behind the Kruskal-Wallis test

The Kruskal-Wallis test is surprisingly similar to ANOVA, in some ways. In ANOVA, we started with Y_{ik}, the value of the outcome variable for the ith person in the kth group. For the Kruskal-Wallis test, what we'll do is rank order all of these Y_{ik} values, and conduct our analysis on the ranked data. So let's let R_{ik} refer to the ranking given to the ith member of the kth group. Now, let's calculate $\bar{R}_k$, the average rank given to observations in the kth group:

$$\bar{R}_k = \frac{1}{N_K} \sum_i R_{ik}$$

and let's also calculate $\bar{R}$, the grand mean rank:

$$\bar{R} = \frac{1}{N} \sum_i \sum_k R_{ik}$$

Now that we've done this, we can calculate the squared deviations from the grand mean rank $\bar{R}$. When we do this for the individual scores – i.e., if we calculate $(R_{ik} - \bar{R})^2$ – what we have is a "nonparametric" measure of how far the ik-th observation deviates from the grand mean rank. When we calculate the squared deviation of the group means from the grand means – i.e., if we calculate $(\bar{R}_k - \bar{R})^2$ – then what we have is a nonparametric measure of how much the *group* deviates from the grand mean rank. With this in mind, let's follow the same logic that we did with ANOVA, and define our *ranked* sums of squares measures in much the same way that we did earlier. First, we have our "total ranked sums of squares":

$$\text{RSS}_{tot} = \sum_k \sum_i (R_{ik} - R)^2$$

and we can define the "between groups ranked sums of squares" like this:

$$\begin{aligned} \text{RSS}_b &= \sum_k \sum_i \left(\bar{R}_k - \bar{R}\right)^2 \\ &= \sum_k N_k \left(\bar{R}_k - \bar{R}\right)^2 \end{aligned}$$

So, if the null hypothesis is true and there are no true group differences at all, you'd expect the between group rank sums RSS_b to be very small, much smaller than the total rank sums RSS_{tot}. Qualitatively this is very much the same as what we found when we went about constructing the ANOVA F-statistic; but for technical reasons the Kruskal-Wallis test statistic, usually denoted K, is constructed in a slightly different way:

$$K = (N-1)\ x\ \frac{RSS_b}{RSS_{tot}}$$

and, if the null hypothesis is true, then the sampling distribution of K is *approximately* chi-square with G−1 degrees of freedom (where G is the number of groups). The larger the value of K, the less consistent the data are with null hypothesis, so this is a one-sided test: we reject H_0 when K is sufficiently large.

12.11.2 Additional details

The description in the previous section illustrates the logic behind the Kruskal-Wallis test. At a conceptual level, this is the right way to think about how the test works. However, from a purely mathematical perspective it's needlessly complicated. I won't show you the derivation, but you can use a bit of algebraic jiggery-pokery[213] to show that the equation for K can be rewritten as

$$K = \frac{12}{N(N-1)} \sum_k N_k \bar{R}_k^2 - 3(N+1)$$

It's this last equation that you sometimes see given for K. This is way easier to calculate than the version I described in the previous section, it's just that it's totally meaningless to actual humans. It's probably best to think of K the way I described it earlier… as an

analogue of ANOVA based on ranks. But keep in mind that the test statistic that gets calculated ends up with a rather different look to it than the one we used for our original ANOVA.

But wait, there's more! Dear lord, why is there always *more*? The story I've told so far is only actually true when there are no ties in the raw data. That is, if there are no two observations that have exactly the same value. If there *are* ties, then we have to introduce a correction factor to these calculations. At this point I'm assuming that even the most diligent reader has stopped caring (or at least formed the opinion that the tie-correction factor is something that doesn't require their immediate attention). So I'll very quickly tell you how it's calculated, and omit the tedious details about *why* it's done this way. Suppose we construct a frequency table for the raw data, and let f_j be the number of observations that have the j-th unique value. This might sound a bit abstract, so here's the R code showing a concrete example:

```
f <- table( clin.trial$mood.gain )   # frequency table for mood gain
print(f)   # we have some ties
```

```
##
## 0.1 0.2 0.3 0.4 0.5 0.6 0.8 0.9 1.1 1.2 1.3 1.4 1.7 1.8
##   1   1   2   1   1   2   1   1   1   1   2   2   1   1
```

Looking at this table, notice that the third entry in the frequency table has a value of 2. Since this corresponds to a `mood.gain` of 0.3, this table is telling us that two people's mood increased by 0.3. More to the point, note that we can say that `f[3]` has a value of `2`. Or, in the mathematical notation I introduced above, this is telling us that f_3=2. Yay. So, now that we know this, the tie correction factor (TCF) is:

$$\text{TCF} = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

The tie-corrected value of the Kruskal-Wallis statistic obtained by dividing the value of K by this quantity: it is this tie-corrected version that R calculates. And at long last, we're actually finished with the theory of the Kruskal-Wallis test. I'm sure you're all terribly relieved that I've cured you of the existential anxiety that naturally arises when you realise that you *don't* know how to calculate the tie-correction factor for the Kruskal-Wallis test. Right?

12.11.3 run the Kruskal-Wallis test in R

Despite the horror that we've gone through in trying to understand what the Kruskal-Wallis test actually does, it turns out that running the test is pretty painless, since R has a function called `kruskal.test()`. The function is pretty flexible, and allows you to input your data in a few different ways. Most of the time you'll have data like the `clin.trial` data set, in which you have your outcome variable `mood.gain`, and a grouping variable `drug`. If so, you can call the `kruskal.test()` function by specifying a formula, and a data frame:

```
kruskal.test(mood.gain ~ drug, data = clin.trial)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  mood.gain by drug
## Kruskal-Wallis chi-squared = 12.076, df = 2, p-value = 0.002386
```

A second way of using the `kruskal.test()` function, which you probably won't have much reason to use, is to directly specify the outcome variable and the grouping variable as separate input arguments, `x` and `g`:

```
kruskal.test(x = clin.trial$mood.gain, g = clin.trial$drug)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  clin.trial$mood.gain and clin.trial$drug
## Kruskal-Wallis chi-squared = 12.076, df = 2, p-value = 0.002386
```

This isn't very interesting, since it's just plain easier to specify a formula. However, sometimes it can be useful to specify `x` as a list. What I mean is this. Suppose you actually had data as three separate variables, `placebo`, `anxifree` and `joyzepam`. If that's the format that your data are in, then it's convenient to know that you can bundle all three together as a list:

```
mood.gain <- list( placebo, joyzepam, anxifree )
kruskal.test( x = mood.gain )
```

And again, this would give you exactly the same results as the command we tried originally.
