Chapter 10: Comparing Multiple Independent Populations
Part 1: ANOVA

In last chapter, our data consisted of samples drawn from two populations, and we saw several procedures for testing the difference between measures of central location based on samples from both populations. We now extend these tests to the more general situation of two or more populations.

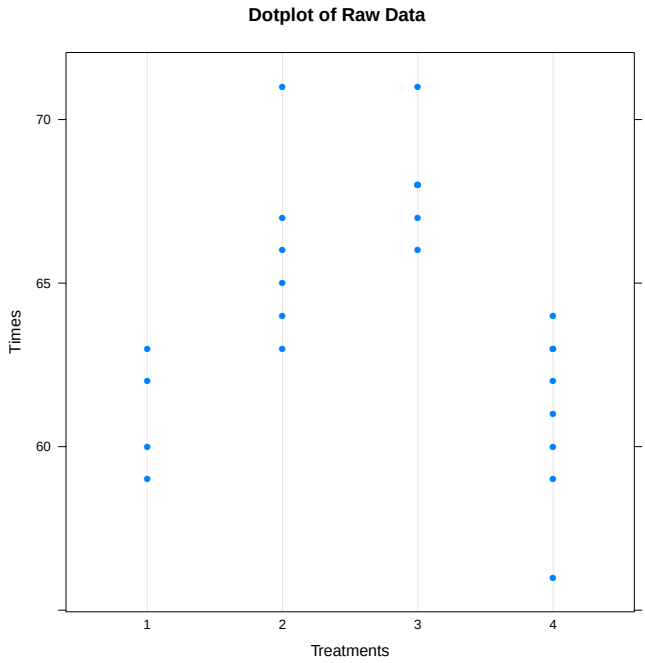
Four new formulations of rat poison are being tested, call them 1, 2, 3, and 4. All of the poisons work by thinning the blood, so the response of interest is the time it takes for the blood to coagulate. A longer blood coagulation time indicates a more effective poison. Twenty-four rats were randomly selected, and then randomized to the four poisons. They were fed the poison, and then after a specified length of time, their blood was drawn and the time to blood coagulation was measured. The data is below:

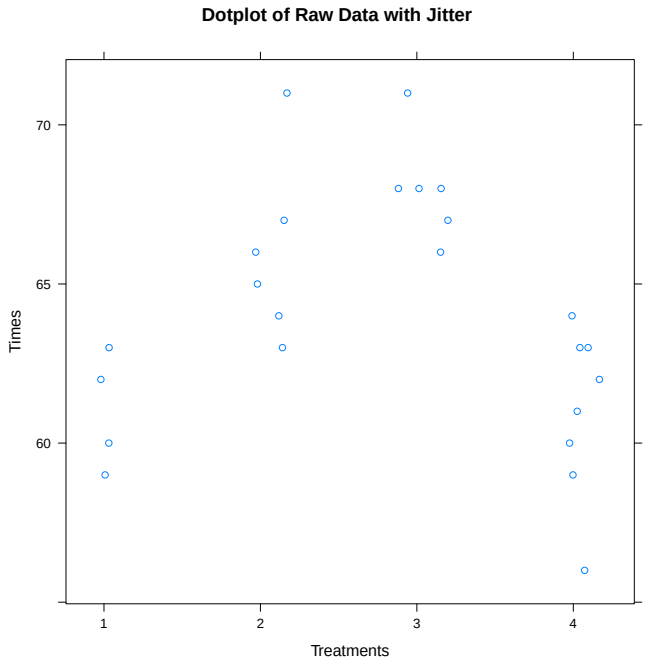
Treatment									Sample Mean
1	62	60	63	59					61
2	63	67	71	64	65	66			66
3	68	66	71	67	68	68			68
4	56	62	60	61	63	64	63	59	61

We'd like to know if any of these poisons results in a different coagulation time than any of the others. If we let μ_1 be the population mean for poison 1, μ_2 be the population mean for poison 2, etc., we can express the hypotheses as:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
vs.
H_A: At least one mean differs from one other mean.

Note that we need to be careful in our statement of H_A. The alternative isn't that the means are all different from one another, it is that at least one is different. We start with some graphing. Side-by-side dotplots seem like a good place to begin. The the first version is a "standard" dotplot, and the second version randomly jitters the points left or right a bit so you can more easily see the observations that are identical:





We can already see from this graph that the treatments seem to differ somewhat. Treatments 2 and 3 seem to have generally higher means than 1 and 4.

Up until this point, we've approached the comparison of means by looking at their differences. We were able to do this because we only had two means to compare. In this new situation, it might seem reasonable to use what we already know, and test each pair of means, using the methods that we've already developed for comparing two means. That approach does indeed work, but it is not the best we can do.

What we are instead going to do is analysis of variance(ANOVA), which is a technique of analyzing sources of

variability. In order to formalize things, we need to start with some notation:

Let t be the number of treatments. Here, $t = 4$.
Let i index the treatments.
Let n_i be the number of observations in treatment i. Here, $n_1 = 4$, $n_2 = 6$, etc.
Let $N = \sum_{i=1}^{t} n_i$ be the total sample size. Here, $N = 24$.
Let y_{ij} be observation j from treatment i. For example, $y_{11} = 62$, and $y_{12} = 60$
Let $\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ be the sample mean for treatment i. For example, $\bar{y}_{1.} = 61$.
Let $\bar{y}_{..} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}}{N}$ be the sample grand mean. Here, $\bar{y}_{..} = 64$.

Our sources of variability are going to be based on a decomposition of the original data. In words, the decompostion can be expressed as:

Observation = Grand Mean + Deviation of Treatment Mean from Grand Mean + Deviation of Observation from Treatment Mean

Or, in our notation, we could express this as:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

However we usually write this with the grand mean subtracted from the left side. Thus:

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Surprisingly, it could be shown

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

The terms in this expression are important enough that we give them names. We call them "sums of squares". In words,

Sum of Squares Total = Sum of Squares due to Treatment (i.e. between) + Sum of Squares due to Error (i.e. within)

Or using shorthand,

$$\text{SSTot} = \text{SSTrt} + \text{SSE}$$

Why do we want to square at all? The reason is that when we square, the things we're left with are very similar to variances. In fact, if you think of the sums of squares as variabilities, then the total variability in the data can be split into one part that is due to the treatments being different (between), and one part due to things we can't explain (within).

But properly speaking, the sums of squares are not quite variances. In order to make them so, we need to divide by appropriate constants. For SSTot, $N - 1$ makes sense, since $\text{SSTot}/(N-1)$ is the variance of all the data. For SSTrt, $t - 1$ is the natural choice, since it's like we're taking the variance of just the t treatment means. And for SSE, we choose $N - t$. We call these things the degrees of freedom, and we'll denote them df_{Tot}, df_{Trt}, and df_E. Sums of squares divided by their dfs are called mean squares, and will be denoted MSTot, MSTrt, and MSE.

Finally, we come to the test statistic. We call it F, and F = MSTrt/MSE. It is the ratio of the between variability to the within variability. **If the data appears to be drawn from populations with normal distributions all having the same variance, then it turns out that the distribution of F is called an F distribution, and it has two parameters, called the numerator df and denominator df**. The numerator df are df_{Trt}, and the denominator are df_E. Tables are available to find probabilites for the F curve.

Now ANOVA is as follows. If we are willing to assume:

- The data are independent within and between treatments
- The variances are the same for all treatments
- Each treatment has a normal distribution,

Then we summarize what we need in an ANOVA table:

Source	SS	df	MS	F	p-value
Treat (between)	SSTrt	$df_{Trt} = t - 1$	MSTrt = SSTrt/df_{Trt}	F = MSTrt/MSE	$p = P(F_{df_{Trt}, df_E} > F)$
Error (within)	SSE	$df_E = N - t$	MSE = SSE/df_E		
Total	SSTot	$df_{Tot} = N - 1$			

For our blood coagulation data, the table works out like this:

Source	SS	df	MS	F	p-value
Treat (between)	228	3	76	13.57	< 0.001
Error (within)	112	20	5.6		
Total	340	23			

Since the p-value is quite small, we would reject the null, and conclude that at least one poison has a different mean coagulation time than another.

We'll introduce how to check those assumptions in the next section.