Chapter 10: Comparing multiple independent populations

(Ott & Longnecker Sections: 14.2 and 14.5)

Duzhe Wang

https://dzwang91.github.io/stat371/

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Example

- 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.
- 24 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

Example continued

- We'd like to know if any of these poisons results in a different coagulation time than any of the others.
- Let μ_1 be the population mean for poison 1, μ_2 be the population mean for poison 2, etc.,

Example continued

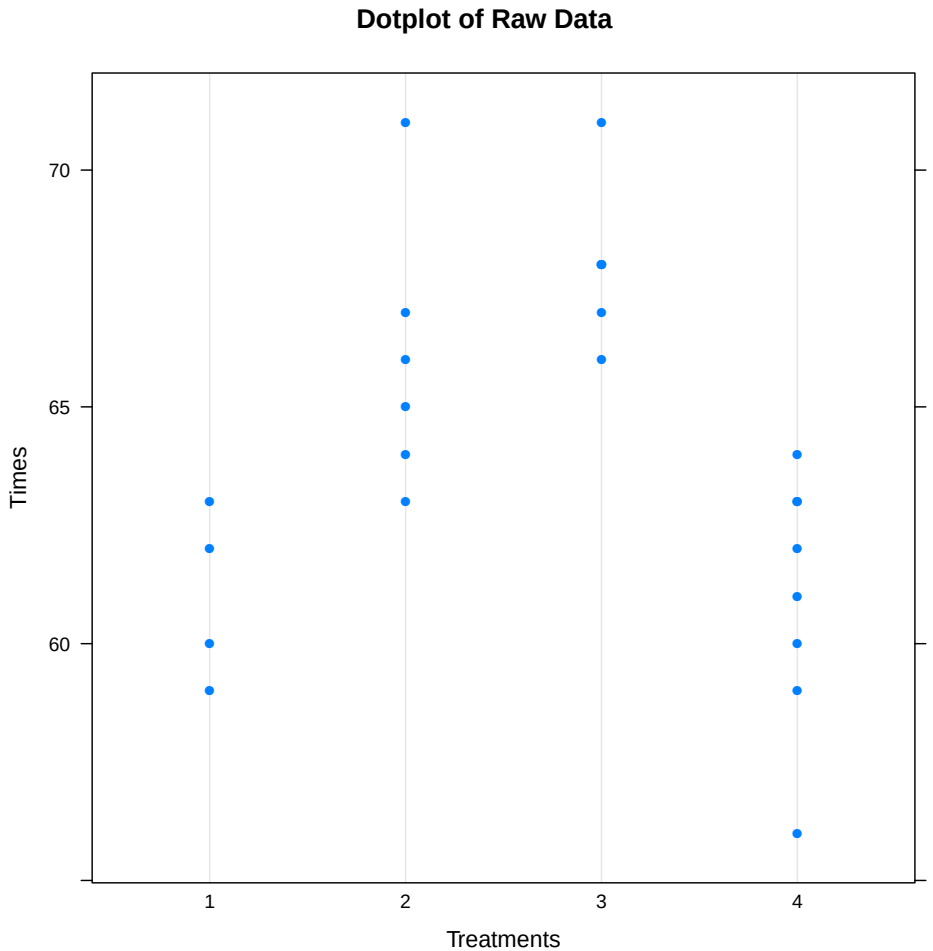
- We'd like to know if any of these poisons results in a different coagulation time than any of the others.
- Let μ_1 be the population mean for poison 1, μ_2 be the population mean for poison 2, etc.,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

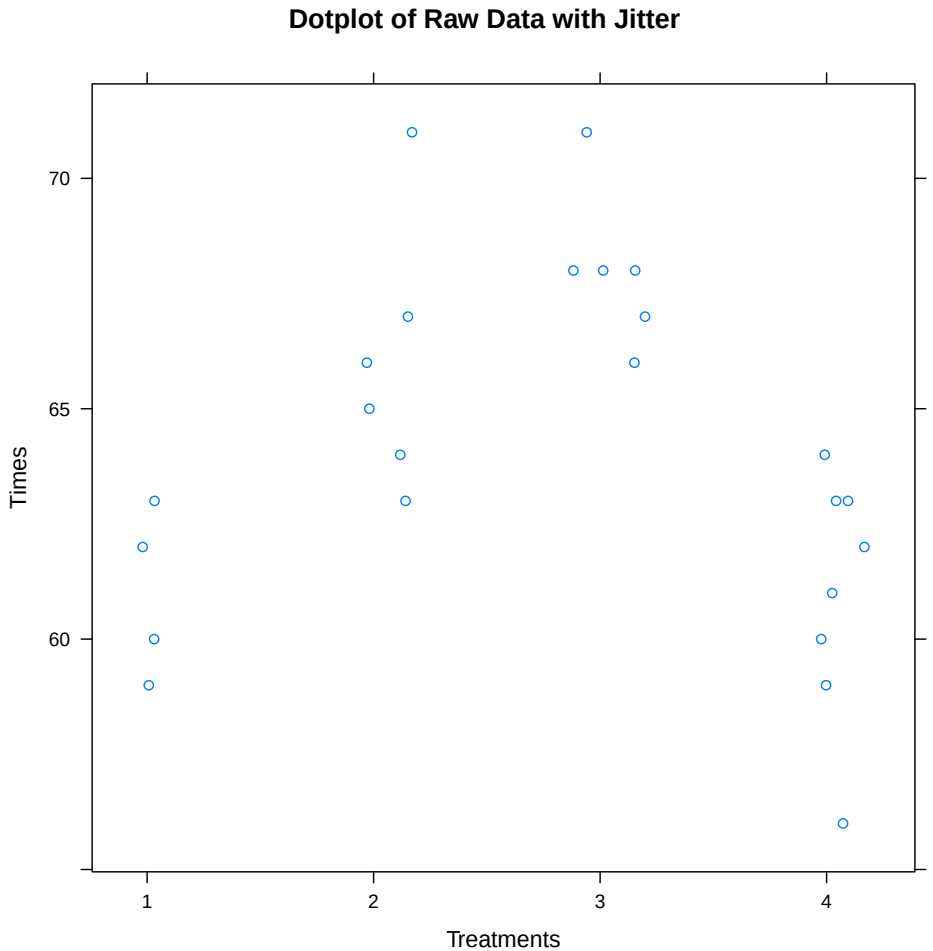
H_A: At least one mean differs from one other mean.

- We start with some graphing.

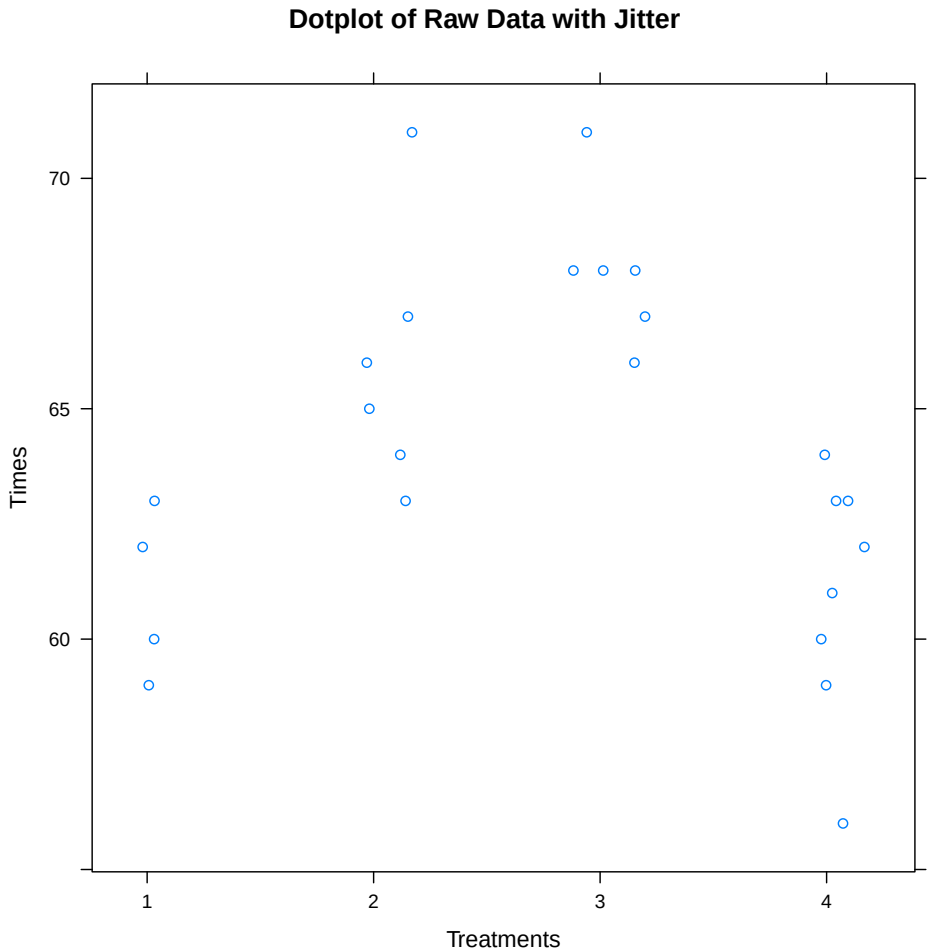
Example continued



Example continued

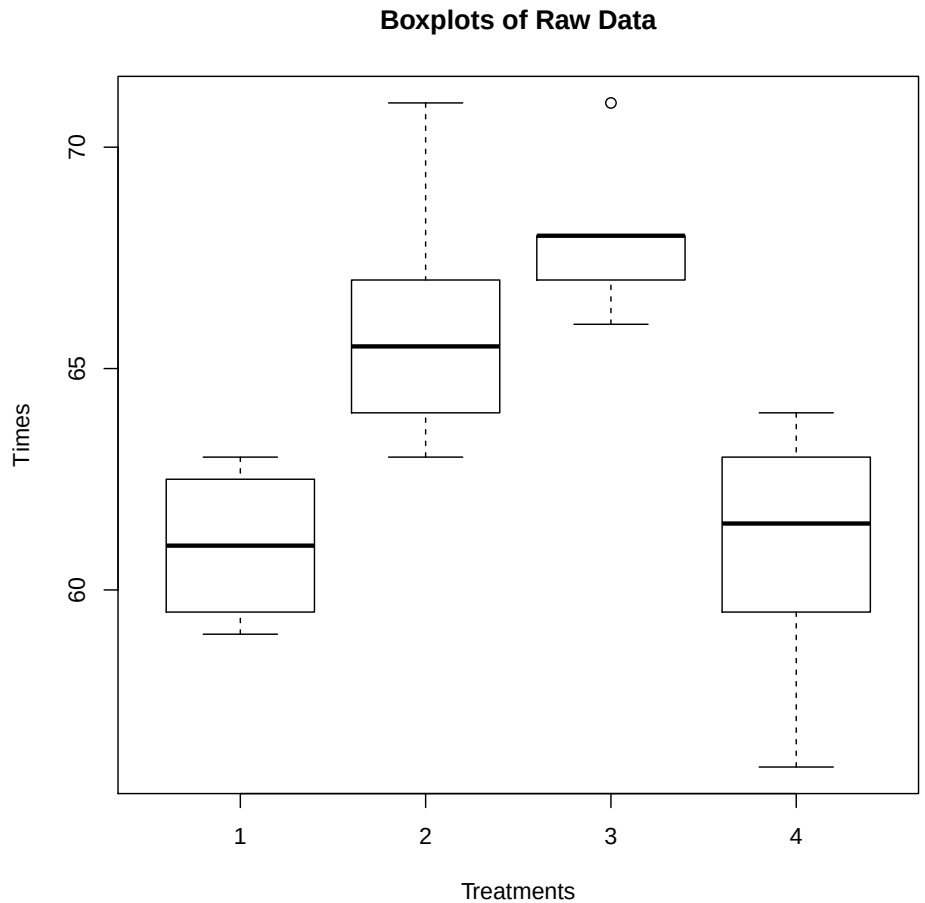


Example continued



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

Example continued



It is clear that treatment 3 might be slightly higher than treatment 2.

Example continued

- Can we use the paired T test in this example? If we can, how do we implement the test?

Example continued

- Can we use the paired T test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.

Example continued

- Can we use the paired T test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.
- A new technique: analysis of variance (ANOVA).

Notations

- 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, $n_1 = 4$, $n_2 = 6$, $n_3 = 6$, $n_4 = 8$.
- 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 $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ be the sample mean for treatment i. For example, $\bar{y}_{1\cdot} = 61$.
- Let $\bar{y}_{\cdot\cdot} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}}{N}$ be the sample grand mean. Here, $\bar{y}_{\cdot\cdot} = 64$.

The key idea of ANOVA

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

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

- In notation,

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

- Thus,

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

The key idea of ANOVA

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

  Observation = Grand Mean + Deviation of Treatment Mean from Grand Mean + Deviation of Observation from Treatment Mean
- In notation,

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

- Thus,

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

- It could be shown

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

Sum of squares

- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$: sum of squares total
- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$: sum of squares treatment
- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$: sum of squares error

Sum of squares

- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$: sum of squares total
- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$: sum of squares treatment
- $\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$: sum of squares error

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

- Why do we square?

Sum of squares

- $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$: sum of squares total
- $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2$: sum of squares treatment
- $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$: sum of squares error

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

- Why do we square?
  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).

Degrees of freedom and mean squares

- But are sum of squares the exact variances?

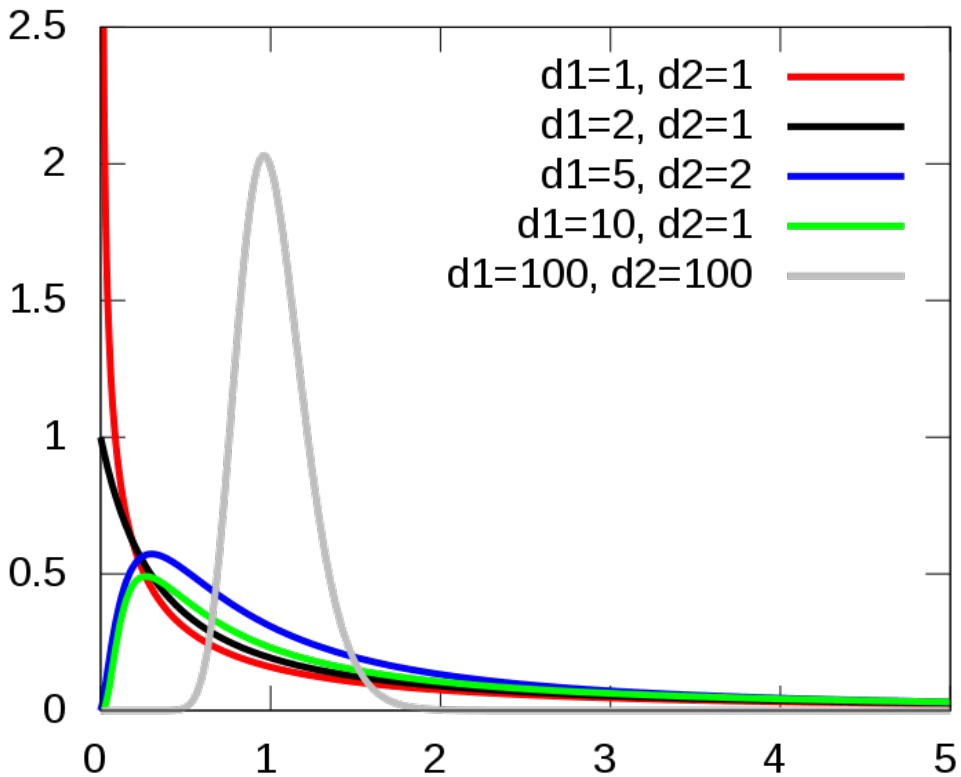
Degrees of freedom and mean squares

- But are sum of squares the exact variances?
  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 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.
  - 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.

Test statistic

- $F = \frac{MSTrt}{MSE}$.
- 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 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.

Pdf of F distribution



Here $d1$ is the df_{Trt}, $d2$ is the df_E.

ANOVA table

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

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

Example

For our blood coagulation data, the ANOVA table is:

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.

Next lecture

We'll introduce how to check assumptions in ANOVA next lecture.