# Notes 11: Inference For Centers of 2+ Independent Samples

In Notes 9, our data consisted of two independent samples, and we saw several procedures for testing the difference between the population mea­­ns. We now extend these tests to the more general situation of two or more independent samples.

## Writing Hypotheses for two or more population means

***Rat Poison Example (a).*** *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. We would like to know if any of the poisons result in a different coagulation time than any of the others.*

***Write hypotheses*** *about the mean blood coagulation times that would answer our question of interest.*

***average blood coagulation time for treatment group [population] i***

***Ho: Ha: at least one pop mean is different***

1 mean describes the mean response we need more than 1 means to describe the

for all 4 poisons 4 population means

**Levels of a factor** distinguish the different populations and their sample measurements.

\*In experiments, the different levels are often combined to make different **treatments**

\*We will be looking at **one-factor** experiments in these notes.

\*We can use the R function as.factor() to specify that a factor is categorical and save it as such.

* *E.g., An experiment to study the effect of three different fertilizer mixtures (type A, type B, type C) on crop yields.* ***Factor:*** *fertilizer mixture* ***Levels****: type A, type B, type C*
* *E.g., An experiment to study the effect of three different categories of social media use (low, moderate, high) on hours of sleep.* ***Factor:*** *social media usage* ***Levels****: low, moderate, high*

## Collect and Summarize Sample Data

***Rat Poison Example (b)*** *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 can be accessed in poison.csv. Read the data into R. Identify the factor levels. Resave your data as needed so your data is stored correctly.*

Chart

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Treatment Level | Observations | Sample Mean | Sample Sd | Sample Size |
| 1 | 62, 60, 63, 59 | 61 | 1.826 | 4 |
| 2 | 63, 67, 71, 64, 65, 66 | 66 | 2.828 | 6 |
| 3 | 68, 66, 71, 67, 68, 68 | 68 | 1.673 | 6 |
| 4 | 56, 62, 60, 61, 63, 64, 63, 59 | 61 | 2.619 | 8 |

***Rat Poison Example (c).*** *Summaries for the treatments are given at right. Describe what you see.*

We notice the sample means are not all the same (there is some variability in observed gp means) look graph.

We notice the sample standard deviations are not all the same, but they are close (2.828/1.673=1.69), within 2.

(It is reasonable to use 1 pooled estimate for a common variance)

## Assumptions for One-Way Anova

1. **We have independent simple random samples from two or more populations**.

\* the study design/implementation much be closely evaluated

1. **Each of the populations must be approximately normal.**

\*To check normality of each population, we can look at qqnorm plots of each sample data. More

commonly with small sample sizes in each group, we look at a QQplot of all *residuals* (since an

equivalent assumption is that the population of errors is normally distributed).

1. **The populations must all have the same variance, which we will denote**

\* To check equality of variance assumption we can look at dotplots of data by group or dot plots of observed within-group error (observed residuals).

Unequal variance assumption

“funnel pattern”

\*We should see similar spread within each of the groups.

Chart, box and whisker chart

Description automatically generated\*We can look at the ratio of SDs: assumption is reasonable if smallest sample SD and largest sample SD are within factor of 2.

\*We often plot the residuals vs fitted values because it is not uncommon to see greater variability when the predicted values are higher. When this is observed, often a transformation of the observed data can rectify the situation.

(*The residual plots at right are examples where unequal variance within the groups is observed.)*

***Rat Poison Example (d). Evaluate whether the assumptions for One-Way (1 factor – poison type) ANOVA appear to be well met in the rat poison experiment.***

1. *Timeline, box and whisker chart

   Description automatically generated***We have independent simple random samples from two or more populations *?***

*We have 4 treatments randomly assigned, so we will consider those 4 random samples from the population of (wild? Lab-raised?) rats*

*We need to worry about independence within groups in experimental design before data collection– cohousing can complicate this*

1. **Each of the populations must be approximately normal.**

*The sample histograms don’t give strong evidence that any of the 4 populations are non-normal. However, these very small samples are often inconclusive. Instead, we will look at the whole collection of residuals to access this assumption (later).*

1. **The populations must all have the same variance, which we will denote**

*Since the sample sds are within a factor of 2, this is a reasonable assumption.*

## Building a Test Statistic by Comparing within to Between Group Variability

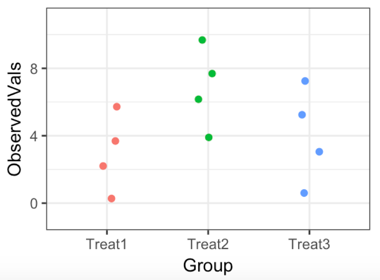
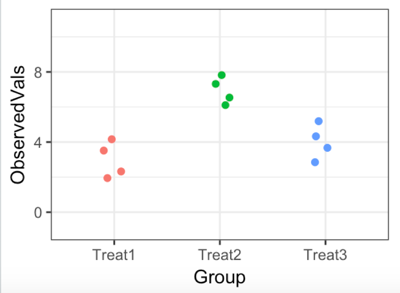
The **one-way** **analysis of variance (ANOVA)** technique compares two sources of variability

\*The “between-treatment” variability is the amount of variability between the treatment means

\*The “within-treatment” variability is the common variability within each treatment group, (pooled over all treatments).

\*The larger the *between-treatment* variability is compared to the *within-treatment* variability, the larger the value of F=the more evidence we have against and more evidence of a treatment effect.

\* ANOVA is a more general case of our 2 sample pooled T test

***Evidence against pictorially example.*** *Consider the two graphs at right. The treatment means are the same in both* *graphs.*

Experiment 2

Experiment 1

*The 3 groups in Experiment 1 have (lower/equal/higher)* ***within group*** *variability compared to the 3 groups in Experiment 2.*



*The more compact circle around values in Exp1 show that groups are more tightly clustered around respective means*

*The 3 groups in Experiment 1 have (lower/equal/higher)* ***between group*** *variability compared to the 3 groups in Experiment 2.*



*The straight horizontal lines between both groups equal.*

*The 3 groups in Experiment 1 have (lower/equal/higher) F=(****between group variability)/(within group variability)*** *compared to the 3 groups in Experiment 2. This means the 3 groups in Experiment 1 have (weaker/equal/stronger) evidence against a null of* ***.***



When we have more than two random variables in our experiment, we abandon the use of and to differentiate the two samples in favor of a single letter *Y (or X)* with two subscripts and .

\*Let *t* be the number of levels/treatments *E.g. t=4 in the poison study*

\* *i* is used to index the levels/treatments: *E.g. i goes from 1: 4 in the poison study*

\* number of observations in level/treatment i, *E.g.*

*\**  is total sample size*, E.g.*

\* is the *jth* observation in theithgroup *E.g.*

\* sample mean of the ith group *E.g.*

\* is the sample standard deviation of the ith group *E.g.*

*\**  sample grand mean *E.g.*

*\*s* is the sample standard deviation for all N values *E.g. s=* *3.845*

### Decomposing Observations

We compute the variability between treatment means and within treatment groups by decomposing each observation into a sum of pieces

Each observation’s value can be thought of as a sum of pieces.

Observation = Grand Mean + Deviation due to Treatment + Deviation of observation

(“Treatment effect”) within group (“Residual”)

*The deviation of an individual observation from the overall grand mean ( ) could come from differences between the factor levels and from random variation in measurements of subjects within the same group.*

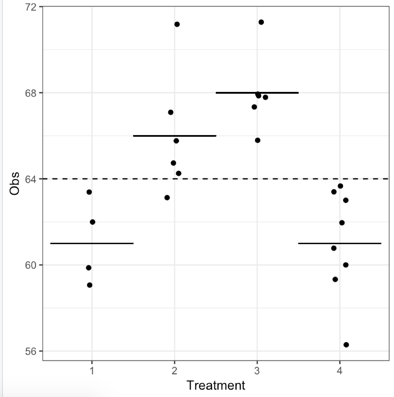
An [observed] **residual** is the difference between an observation’s value and the value that is predicted by a statistical model given the factor level or other predictor variable combinations.

\* In an ANOVA (groups’ mean) model, =

\*The collection of observed residuals is a sample of the population of errors.

***Rat Poison Example (e).*** *Decompose the observation according to the method above. Show the decomposed pieces graphically. Identify the observed residual when comparing to the group 4 mean.*

|  |  |  |  |
| --- | --- | --- | --- |
| Treat | Observations | Sample Mean | Size |
| 1 | 62, 60, 63, 59 | 61 | 4 |
| 2 | 63, 67, 71, 64, 65, 66 | 66 | 6 |
| 3 | 68, 66, 71, 67, 68, 68 | 68 | 6 |
| 4 | 56, 62, 60, 61, 63, 64, 63, 59 | 61 | 8 |
| Overall |  | 64 | 24 |



*Observed residual:*



*overall mean + “treatment effect” + “random error”*

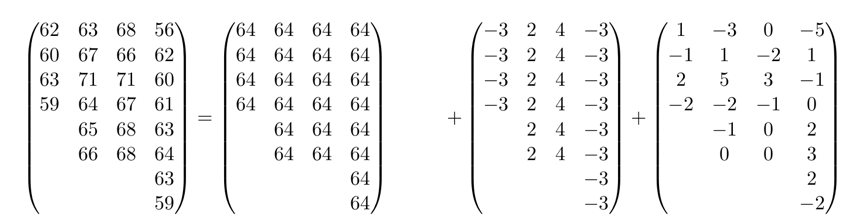


*has an observed residual of +1 (using the ANOVA model)*

***Rat Poison Example (f).*** *Compute the [observed] residual for the value compared to the group 2 mean. What does the sign of the residual indicate about the relative position of the value?*

*has a residual of 63-66=-3 [Using the ANOVA model]*

*Which tells us this observed value is 3 units below the treatment 2 sample mean.*

**

***Rat Poison Example (g).*** *Each observation within the rat poison experiment can be broken down in a similar way. In these decompositions, observations within a single treatment group only differ in how far they are from their treatment group mean. This value is called the observation’s* ***residual.***



### Getting Variance Estimates from Sample Data

The total squared deviations of all values (SSTotal) can be partitioned into the sum of squared deviations comparing group means to the overall mean (SSTreat) and the sum of the squared residuals (SSError)



Total Sum of Squares = Treatment sum of Squares + Error sum of Squares

**We combine the squared deviations with an appropriate degrees of freedom to get variance estimates**

**\*Total Variability**

**Toal Sum of Squares: SStotal=**

\*gives the total sum of squared deviations across treatments

**Total df = N-1**

**Total Variability:**

**\*Between Group Variability**

**Treatment Sum of Squares: SSTreat=**

\* gives the sum of squared deviations comparing the sample means the sample

Grand mean

\*notice the means for the larger samples count more since each squared

deviation is multiplied by sample size

**Treatment df =t-1**

**Mean Square Treatment: MSTreat=**

**\*Within Group (Unexplained) Variability**

**Error Sum of Squares: SSError=**

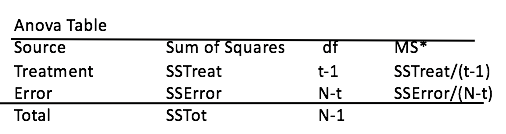
**\*** gives the sum of squared deviations comparing the individual sample points to

their respective group means

**Error df=N-t**

**Mean Square Error: MSError=**

Summarizing Sources of Variability in an Anova Table



The **Sums of Squares, Degrees of Freedom, and Mean Squares** are often summarized in an ANOVA table

***Rat Poison Example (h).*** *Compute the Sums of Squares for Total, Treatment and Error. Confirm the statement that SSTot=SSTreat+SSError. Also compute the degrees of freedom for Total, Treatment, and Error and the resulting mean square values. Summarize your values in an ANOVA table.*

|  |  |  |  |
| --- | --- | --- | --- |
| Treatment Level | Sample Mean | Sample Sd | Sample Size |
| 1 | 61 | 1.826 | 4 |
| 2 | 66 | 2.828 | 6 |
| 3 | 68 | 1.673 | 6 |
| 4 | 61 | 2.619 | 8 |
| Overall | 64 | 3.845 | 24 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Source* | *Sum of Squares* | *Degrees of Freedom* | *Mean Square* |
| *Treatment* | *228* | *4-1=3* | *228/3=76* |
| *Error* | *112* | *20* |  |
| *Total* | *340* | *23* | *XXXX* |

*For pooled varienace = sqrt((3\*1.826^2 + 5\**2.828^2 +5\*1.673^2 +7\*2.619^2)/20)

**Total Sum of Squares: SStotal=**

**Total DF: 24-1=23 Overall Variability:**

**Treatment Sum of Squares**:

SSTrt=

**Treatment DF: MSTreat:**

**(between Gp variance)**

**Error Sum of Squares: SSError=**

**SSError=**

**Error DF: MSError:**

*(pooled within group variance)*

### MSTreat and MSError under null of equal population means (Variances)

If in fact the population means across the n groups are the same and each population has the same variability, then MSTreatment and MSError are estimating the same variability and should have similar value. If the group means differ markedly, then the treatment mean square would be larger and give us stronger evidence against the null of no difference in population means.

ANOVA Test Statistic:

has an distribution with df: (t-1, N-t). is called numerator df and is called denominator df.



Chart, line chart

Description automatically generated

The **F Distributions** are a collection of density functions determined by two degrees of freedom d1 and d2.



\*They are defined from 0 to infinity and right skewed.

In R:

pf(q, df1, df2,lower.tail = TRUE)

qf(p, df1, df2,lower.tail = TRUE)

**F Test for Equality of Means in t groups,** also called **ANOVA: one-way analysis of variance**

\* To test for: vs

At least one of the t population means is different from the rest

\* If we can assume:

1. The t sets of measurements are independent random samples from their respective populations.
2. Each of the t populations has a normal distribution
3. The variances of the t populations are equal, that is

\***Test Statistic:** has an distribution with df: (t-1, N-t). Where is called numerator df and is called denominator df.

\***p value:**

\*The F test is more powerful than regular two-sample t tests with more than 2 samples because if treatment variances are actually equal, we get a much more accurate estimate of the shared variance

\*The F test is **robust** in the sense that small or moderate departures from the normality and constant variance do not seriously affect its performance. The test is less robust when groups have unequal sample sizes.

In R:

Model<-aov(formula, data = NULL,…)

summary(Model)

or

Model2<-lm(formula, data=NULL,…)

anova(Model2)

Plantanova<-aov(Data~Plant, data=Plantdata) or Plantanova2<-lm(Data~Plant, data=Plantdata)

summary(Plantanova) anova(Plantanova2)

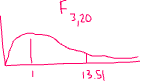
|  |  |
| --- | --- |
| *Source* | *Mean Square* |
| *Treatment* | *228/3=76* |
| *Error* | *112/20=5.6* |
| *Total* |  |

***Rat Poison Example(i).*** *Compute the test statistic and p value for the F test for the hypotheses: vs at least one mean differs from the others where is the population mean time to coagulation for blood treated with treatment i using the information in the ANOVA table (part h) and then check your computations in R.*

***Test Statistic:***

***If the 4 groups are drawn from the same population (or 4 populations with the same mean and variances), we expect to see F test statistics around 1.***

***p-value:***



***pf(13.571, 3, 20, lower.tail=FALSE)=***



***4.658 e-05***

***We have enough evidence to reject the null and suggest that at least one poison***

***mean time to coagulate is different from the others.***

***check computations in R:***

*#AOV function willl make ANOVA table for you*

*#notice we need to have our data in long-format*

*#aov (Response ~ Predictor)*

*aov\_rat=aov(Obs ~ Treatment, data=poison)*

*summary(aov\_rat) #ANOVA table*

***rechecking ANOVA assumptions by looking at Residual Plots:***

|  |  |
| --- | --- |
| *A graph with numbers and lines  Description automatically generated* | *A graph of a number of numbers and a line  Description automatically generated with medium confidence* |
| *This first graph is displaying the fitted values (treatment sample means) on the x axis and residuals on the y axis.*  *This tells me the equal variance assumption we made earlier is reasonable.* | *This second graph is displaying a qqnorm plot of of [standardized] residuals.*  *This tells me a normality assumption for the population of residuals seems reasonable. This is consistent with our assumption that all 4 populations we are drawing from are normally distributed.* |

*Chart, scatter chart

Description automatically generated*

***Powerplant example:*** *In the article “Review of Development and Application of CRSTER and MPTER Models” (R.Wilson Atmospheric Environment), several measurements of the maximum hourly concentrations of S02 are presented for each of four power plants. The paper was interested in evaluating whether there is evidence that all four plants have similar average hourly concentration of SO2. The S02 measurements and a graphical summary for the four plants are given at right. (data available in Plantdata.csv).*

***Powerplant example (a) :*** *We can construct an ANOVA table to summarize variability within and between the four groups.*

**Table

Description automatically generated**

*We can compute the total sum of squared deviations using the overall sample standard deviation of the combined data set:*

***.***

*We can compute the SSTreat by comparing the group means to the overall mean for as many replicates are in each group:*

***.***

*We can compute all within group deviations SSError by comparing each observation to its the group mean via the sample standard deviations.*

*We then compute MSTreat and MSError to get estimates for the between group and within group variability, respectively.*

***MSTreat=SSTreat/dfTreat=***

***MSError=SSError/dfError=***

*We can then organize these into an ANOVA table:*

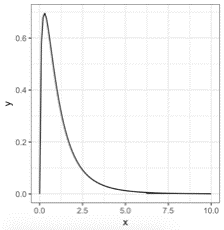
**ANOVA table for Power Plant SO2 Data:**

***Source Sum of Squares d.f. Mean Square \_\_***

***Treatment 4-1=3***

***Error 19-4=15***

***Total 19-1=18 X***

***Powerplant example (b):*** *For the hypothesis that all four plants have similar average hourly concentration of SO2, that is and At least one of the plants’ average hourly concentration of SO2 is different from the rest, we will use an F test for equality of means (after checking that the assumptions necessary are reasonable – more on that next).*

*We’ve computed and , so*

*. Comparing the test statistic to results in a p value: . We have very strong evidence against the null. Evidence suggests at least one of the plants’ average hourly concentration of SO2 is different from the rest.*

*We can check in R:*

> Plantanova<-aov(Data~Plant, data=PPlant)

> summary(Plantanova)

Df Sum Sq Mean Sq F value Pr(>F)

Plant 3 378610 126203 6.21 0.00592 \*\*

Residuals 15 304838 20323

Chart, histogram

Description automatically generated*Chart, scatter chart

Description automatically generated****Powerplant example (c):*** *Power plant data residual graphs given at right show a heavy left tail, but not too off what we might see from a sample of size 19 from a normal population.*

*Code used: qqnorm(Plantanova$residuals); qqline(Plantanova$residuals); hist(Plantanova$residuals) or plot(Plantanova)*

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated***Powerplant example (d):*** *The variability within the four plant groups’ residuals looks to be similar. Additionally, the smallest group SD: 122.7 and the largest group SD: 185.3 are within a factor of 2.*

*R code: plot(Plantanova).*

## Steps to Take After Overall F Test for Equality of Means

If the F test **is not significant**, we do not reject the null and we often stop there.

If the F test **is significant**, we can only say at least two of the groups are significantly different.

\*One way to clarify which group means are significantly different is with **post-hoc pairwise** tests

or CIs that use the pooled variance.

\*We often use a lettering system to show groups that are[not] significantly different from one another by giving a list of the group means in ascending order and assigning the same letter to groups that are not significantly different from one another.

## Multiple Comparisons with Post-Hoc Pairwise Tests

With 2 treatment groups, we have 1 comparison we are making . When we have 3 treatment groups, we are making 3 comparisons ()(THE MORE TEST WE DO THE HIGER THE TYPE 1 ERROR RATE GETS).. We can use the general binomial formula to determine how many comparisons need to be made with treatment groups.

## Adjusting Significance for Multiple Comparisons

We often make significance level adjustments to control our Type 1 error rates when we are doing more than one comparison

\*The probability of a Type 1 error of any single test is called the **Comparison-Wise Error Rate, or CWER.** Suppose the CW rate is .

\*Doing a large collection of m independent tests greatly inflates our probability of making a Type 1 error somewhere in the collection-the **Family-Wise Error Rate** or **FWER.** The probability of a Type I error in at least one of the m independents test is:

***Family-Wide Error Rate Example:*** *At a CWER rate of , the FWER increases quickly. For m=1 test, . For m=5 tests, , and for m=10,*

## Options for Pairwise Comparisons with CIs following significant ANOVA

**Fisher’s method** has the most power but does not account for making more than one comparison for and



\* Construct:

**Bonferroni’s method** can be used for any collection of *m* tests. It is very easy, but loses power quickly with a large number of pairwise comparisons. Set the individual test comparison rate to 1/m th the desired family error rate. .



\*Construct:



\* Compute the p value as before for each comparison, then **multiplying each p value by m** to get a Bonferoni-adjusted p values.

**Tukey-Kramer method** for finding multiple comparison adjusted CI’s is more work, but has higher power if we want to look at all pairwise tests.



\* Construct: Where q comes from Q: the Studentized Range Distribution.



In R: To get

OR TukeyHSD(aovmodel)

|  |  |  |  |
| --- | --- | --- | --- |
| Treatment Level | Sample Mean | Sample Sd | Sample Size |
| 1 | 61 | 1.826 | 4 |
| 2 | 66 | 2.828 | 6 |
| 3 | 68 | 1.673 | 6 |
| 4 | 61 | 2.619 | 8 |
| Overall | 64 | 3.845 | 24 |

***Rat Poison Example(j). Confidence Intervals for post-hoc comparisons.*** *Since the overall F test for equality of means was significant, we will use Cis to do all post-hoc comparisons.*

*How many post-hoc comparisons need to be made for the 4 treatment groups?*

*1-2, 1-3, 1-4, 2-3, 2-4, 3-4*

*6 different post-hoc comparisons*

|  |  |
| --- | --- |
| *Source* | *Mean Square* |
| *Treatment (poison)* | *228/3=76* |
| *Error* | *112/20=5.6* |
| *Total* |  |

***Rat Poison Example(k). Confidence Intervals 3 different ways*** *compute the confidence interval comparison between the rat poison treatment* ***groups 1 and 2*** *(****)*** *first using Fisher’s method, then Bonferroni’s, and then Tukey’s method. A family-wide overall 5% significance level is chosen.*

*Compare the conclusions that should be drawn with the 3 methods.*

|  |
| --- |
| *Fisher’s Method:*    *CV:*  *SE for*  *=* ***(-8.196, -1.814) \*We notice the interval does not contain 0, so we have significant finding. Enough evidence to suggests mean blood coagulation for T1 and T2 are different*** |
| *Bonferroni’s Method: (This is often seen as too conservative)*  *\*because we have 6 pair-wise comparisons with 4 treatment gps\**  *Point estimate :*  *CV: =* *qt(.05/12, df=20, lower.tail=FALSE)*  *2.927119*  *SE for*  *= (-9.47, -0.53) (We notice the BF interval is wider than Fisher’s.) However, the interval still does not contain 0. Even with Bonferonni adjustment, we have enough evidence to suggest mean blood coagulation for T1 and T2 are different.* |
| *Tukey’s method:*  ***(we notice Tukey-Kramer CI is wider than Fisher’s, but narrower than Bonferoni.*** |

***Rat Poison Example(l).*** *Construct the confidence interval for* *using Fisher’s method, Bonferroni, and Tukey-Kramer. Compare components and widths of the three confidence intervals.*

|  |
| --- |
| *Fisher’s Method: (-10.19, -3.81)*    *\*Notice, because gp 2 and 3 had the same sample sizes, the ME for mu1-mu2 and mu1-m3 CI were the same* |
| *Bonferroni’s Method: (-11.47, -2.53)* |
| *Tukey’s method: (-11.28, -2.72)* |

***Rat Poison Example(m).*** *Use the completed confidence interval tables below to do the comparisons for each set of 2 independent samples. Summarize your findings in a table,* ***assigning the same letter to groups that are not significantly different from one another.***

***Groups that share a letter code not found significantly different using Fisher’s method.***

***used for each comparison.***

***Fisher’s Method (C.I. NOT INCLUDE ZERO)***

|  |  |  |
| --- | --- | --- |
| ***Estimated Difference*** | ***Confidence Interval*** | ***Conclusion for*** |
|  | ***(-8.186, -1.814)*** | ***Reject (significant finding)*** |
|  | ***(-10.186, -3.814)*** | ***Reject*** |
|  | ***(-3.023, 3.023)*** | ***Fail to Reject (No significant finding)*** |
|  | ***(-4.850, 0.850)*** | ***Fail to Reject*** |
|  | ***(2.334, 7.666)*** | ***Reject*** |
|  | ***(4.334, 9.666)*** | ***Reject*** |

|  |  |  |
| --- | --- | --- |
| ***Treatment Group*** | ***Sample Mean*** | ***Group Code*** |
| ***4*** | ***61*** | ***A*** |
| ***1*** | ***61*** | ***A*** |
| ***2*** | ***66*** | ***B*** |
| ***3*** | ***68*** | ***B*** |

***Bonferroni’s Method***

***Groups that share a letter code not found significantly different using Bonferroni’s method.***

***used for each comparison.***

|  |  |  |
| --- | --- | --- |
| ***Estimated Difference*** | ***Confidence Interval*** | ***Conclusion for ?*** |
|  |  | ***Reject*** |
|  |  | ***Reject*** |
|  |  | ***Fail to reject*** |
|  |  | ***Fail to reject*** |
|  |  | ***reject*** |
|  |  | ***Reject*** |

|  |  |  |
| --- | --- | --- |
| ***Treatment Group*** | ***Sample Mean*** | ***Group Code*** |
| ***4*** | ***61*** | ***A*** |
| ***1*** | ***61*** | ***A*** |
| ***2*** | ***66*** | ***B*** |
| ***3*** | ***68*** | ***B*** |

***Tukey-Kramer’s Method***

***Groups that share a letter code not found significantly different using Tukey-Kramer method. An overall was used.***

|  |  |  |
| --- | --- | --- |
| ***Estimated Difference*** | ***Confidence Interval*** | ***Conclusion for ?*** |
|  |  | ***Reject*** |
|  |  | ***Reject*** |
|  |  | ***Fail to Reject*** |
|  |  | ***Fail to Reject*** |
|  |  | ***Reject*** |
|  |  | ***Reject*** |

|  |  |  |
| --- | --- | --- |
| ***Treatment Group*** | ***Sample Mean*** | ***Group Code*** |
| ***4*** | ***61*** | ***A*** |
| ***1*** | ***61*** | ***A*** |
| ***2*** | ***66*** | ***B*** |
| ***3*** | ***68*** | ***B*** |

***Powerplant example (e):*** *An overall F test gave evidence that at least one of the plants’ average hourly concentration of SO2 is different from the rest, so we should do the 6 pairwise comparisons to better understand which group means may be different. We compute CIs and use the fact that if 0 falls into the CI, we do not have sufficient evidence to suggest the population means differ. (only 3 of the 6 Cis computed here to save space)*

*Fisher’s CI method: and MSE=20322.54, for a 95% CI , t(15, 0.25) = 2.13 so*

*For:*

*For: )*

*For: ) , etc.*

*.*

*We summarize our findings and those not fully shown below, organizing the sample means in increasing order and giving groups that are not significantly different the same lettering.*

*Treatment Sample Mean Group (\*Based on Fisher’s CI with for each CI)*

*Plant 1 606.8 A*

*Plant 4 777.7 AB*

*Plant 3 919.0 BC*

*Plant 2 992.0 C*

*Bonferroni’s CI method: and MSE=20322.54, for a FWER of We use the t multiplier: .*

*For*

*For )*

*For ), etc*

*Tukey-Kramer CI method*

***Powerplant example (f):*** *With the plant data: , MSE=20322.54, for a FWER of*

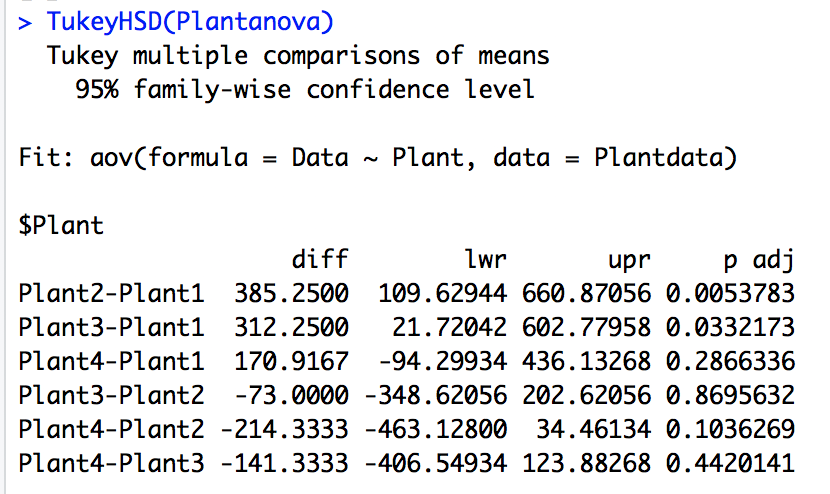
*We use the multiplier:*

*For*

*For )*

*For )*

In R:



Grouping Summary based on Tukey:

*Treatment Sample Mean Group*

*Plant 1 606.8 A*

*Plant 4 777.7 AB*

*Plant 3 919.0 B*

*Plant 2 992.0 B*

## Method Not assuming the populations are normally distributed

The **Kruskal-Wallis Test** does an overall test of difference between independent groups if the residuals give strong evidence the groups’ populations are not normally distributed (or if we just prefer to not make a normality assumption). Less Powerful than before.

\*Test statistic uses ranks of the data instead of values.

\*Assumptions:

(1) Each sample of observations is randomly selected from their respective populations.

(2) The populations are identical except for a possible difference in shift parameter. (have same shape and spread)

\*TS is complicated to compute by hand, so will rely on R:

In R: kruskal.test(response ~ group)

\*If overall Kruskal-Wallis is significant, we have evidence that at least 1 populationis shifted from the other populations. Then move to Wilcoxon Rank Sum pairwise comparisons and do a Bonferroni adjustment as desired.

***Rat Poison Kruskal-Wallis Example (n).*** *Perform an overall Kruskal-Wallis test for a difference between rat poison treatment time to coagulation distributions. Perform follow-up comparisons if appropriate. Summarize your findings in a table and compare the conclusions with that given by the F test and its follow up pairwise methods.*

***Overall Kruskal-Wallis:***

***Kruskal-Wallis***

***chi-squared = 17.015, df = 3, p-value= 0.0007016***

*since p value < 0.05, there is evidence of a shift in at least one of the 4 populations from the others.*

***Post-Hoc Comparisons***

*wilcox.test(treat1,treat2, paried=FALSE)*

|  |  |  |  |
| --- | --- | --- | --- |
| ***Comparison*** | ***Wilcoxon Rank Sum***  ***Test Stat, P-value***  ***(Notes 9)*** | ***Bonferroni Adjusted P-value*** | ***Conclusion for shifted from***  ***(comparing BF adjusted p value to*** |
| ***Pop 1 and Pop 2*** | ***W = 0.5, p-value = 0.01866*** | ***0.01866\*6=0.11196*** | ***Fail to reject that there is difference*** |
| ***Pop 1 and Pop 3*** | ***W = 0, p-value = 0.01306*** | ***0.07836*** | ***Fail to reject*** |
| ***Pop 1 and Pop 4*** | *W = 14.5,*  *p-value = 0.8635* | *0.8635\*6=1* | ***Fail to Reject*** |
| ***Pop 2 and Pop 3*** | *W = 7.5,*  *p-value = 0.105* | *0.105\*6=0.63* | ***Fail to Reject*** |
| ***Pop 2 and Pop 4*** | *W = 45.5,*  *p-value =0.006409* | *0.006409\*6=*  *0.038454* | ***Reject*** |
| ***Pop 3 and Pop 4*** | *W = 48,*  *p-value =0.002284* | *0.002284\*6=*  *0.013704* | ***Reject*** |

|  |  |  |
| --- | --- | --- |
| ***Treatment Group*** | ***Sample Mean*** | ***Group Code*** |
| ***1*** | ***61*** | ***A B*** |
| ***4*** | ***61*** | ***A*** |
| ***2*** | ***66*** | ***B*** |
| ***3*** | ***68*** | ***B*** |

***Groups that share a letter code not found significantly different using Wilcoxon Rank Sum and Bonferroni Adjustment. was used.***

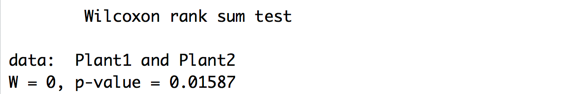
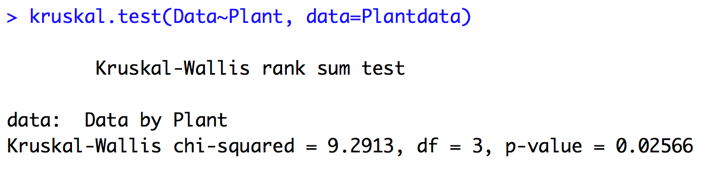
A screen shot of a graph

Description automatically generated

***Why would Wilcoxon Rank Sum w/ Bonfronni adjustment not find Tr1 significantly different from Tr2 and T3?***

*We notice the sample size of T1 group is only 4 (compared to n=8 in Tr 4). Wilcoxon Rank Sum doesn’t have high power to find significance when we have small sample sizes.*

***Powerplant example (g):***  *If we do not want to make the assumption that the 4 populations of S02 emittants would be approximately normal, we can instead perform a Kruskal-Wallis test to test whether there is evidence that one plant’s distribution is shifted from the others. We’ll do this in R:*

**



**



## *The overall Kruskal-Wallis suggests one population is shifted from the others. We can multiply the p value in each of the follow up Wilcoxon Rank Sum Test pairwise comparisons by 6 to do a Bonferoni adjustment. That, however, leaves no significant differences at the 5% level between the power plants.*