|  |
| --- |
| title: "Statistical Foundations for Data Science - Midterm " |
| author: "Travis Deason" |
| date: "Oct 14th 2017" |
| output: word\_document |

# MSDS 6371 Midterm

## Meloma study: Higher counts indicate more / worse disease. There have been several drugs that have been developed to fight this type of cancer, the effects of which are measured in the reduction of the Lambda and Kappa protein levels. Four of these drugs are as follows: Revlimid, Velcade, Dex, and Pomolyst. \* a researcher randomly selected the records of 35 patients who had been taking one of the drugs above. To be clear, there were 140 patients’ records in the study, 35 patients that took Revlimid, 35 that took Velcade, 35 that took Dex, and 35 that took Pomolyst. The researcher recorded the percentage drop in Lambda protein during the standard 15-month treatment of taking the drug.

meloma <- read.csv('data/myeloma.csv')  
names(meloma) <- sapply(names(meloma), tolower)  
boxplot(lambdadrop~drug, data=meloma, main='Lambda drop')

|  |
| --- |
|  |

boxplot(kappadrop~drug, data=meloma, main='Kappa drop')

dex <- subset(meloma, drug == 'Dex')  
pomo <- subset(meloma, drug == 'Pomolyst')  
revl <- subset(meloma, drug == 'Revlimid')  
velc <- subset(meloma, drug == 'Velcade')

#### A. The principal question of interest for the researchers was if there were significant differences in the mean or median percent drop of Lambda protein between the drugs and if there are, an estimate of the magnitude of the difference(s). Provide analysis that will best answer the principle question of interest above. Please address all assumptions needed to conduct your analysis, and provide a scope of inference with your findings. Assume the researchers are interested in maintaining a family wise error rate of 5% (alpha family = .05).

* The null hypothesis in this test is that there is no diffirence in the median drop of Lambda protient levels between the drugs used in the study. Our alternative hypothesis is that there is a diffirence in the median of Lambda protien drops between the groups.
* To test this assumption we will use Anova. With the Lambda protien drop data, we can use Anova becuase, while the data is not normally distributed, there is a large enough sample size (35 in each group), the group sizes are all equal, and the group variances are similar. Since the testing was done on the same patients in a before/after case, we cannot assume independence, but this violation of indepndence is acceptable for our testing

lambda\_aov <- aov(lambdadrop~drug, data= meloma)  
summary(lambda\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## drug 3 193.5 64.5 2.892 0.0377 \*  
## Residuals 136 3033.3 22.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Our critical F\_value for a siginificence level of .95 and a degrees of freedom of 3 for the reduced and 136 for the residuals is 2.67
* The F-value obtained from the data is 2.892, which is greater then 2.67 and results in a p-values of .0377, so we will reject the null hypothesis at a .95 significence level.
* There is sufficent evidence to show that there is a diffirence in the response betweent he four groups of drugs being tested in this expirement. Since the paitents were chosen at random, we can infer that this relationship is reflective of the population of paitents with Meloma. Since the paitents were not randomly assigned a drug group, we cannot assume that the change in Lambda drops is causal.

#### B. A second question of interest centered on comparing Type A drugs to Type B drugs. It turns out that Revlimid and Velcade are Type A drugs while Dex and Pomolyst are both Type B drugs. Test the claim that the Type A drugs have a greater mean percent drop of Lambda protein than the Type B drugs by comparing the mean percent drop of Revlimid and Velcade to the mean percent drop of Dex and Pomolyst using a contrast. For this question, you may assume all the assumptions are met to run a contrast, but do show all 6 steps of the hypothesis test (t-test) and provide a 95% confidence interval for the difference as well. Test at the alpha = .01 level of significance.

meloma['typea'] <- (meloma['drug'] == 'Revlimid' | meloma['drug'] == 'Velvade')  
boxplot(lambdadrop~typea, meloma, main='Is Type A')

|  |  |  |
| --- | --- | --- |
|  | |  |
|  |  |  |

typea = subset(meloma, typea == TRUE)  
typeb = subset(meloma, typea == FALSE)  
qqnorm(typea$lambdadrop, main='Type A')

qqnorm(typeb$lambdadrop, main='Type B')

hist(typea$lambdadrop, main='Type A')

hist(typeb$lambdadrop, main= 'Type B')

* In investigating the difference in means between type A and type B test. Then null hypothesis is that lambda protien drop in type A drugs is not larger then in type B drugs. Our alternative hypotheis is that there is a There is a larger drop in lambda protien among paitents taking type A drugs then there is with paitents taking type B drugs.
* The data is not normal, but since the median and mean values are similar, and the sample sizes are large (70 paitents in each group), it is safe to assume the willcentral limit theorm will approximate normality in this case. Looking at boxplots for the two data sets, varience is larger for the Type\_B group then the Type\_A group, so we will use a one sided welch's t-test to compare the data.

t.test(typea$lambdadrop, typeb$lambdadrop, conf.level=.90, alternative='greater')

##   
## Welch Two Sample t-test  
##   
## data: typea$lambdadrop and typeb$lambdadrop  
## t = 2.1538, df = 66.91, p-value = 0.01743  
## alternative hypothesis: true difference in means is greater than 0  
## 90 percent confidence interval:  
## 0.7425525 Inf  
## sample estimates:  
## mean of x mean of y   
## 81.00538 79.14456

pt(.80, 66.91)

## [1] 0.786729

* Based on a Welch's two sample T-test with 66.91 degrees of freedom, the critical value to reject the null hypothesis at a one sided .90 confidence level is .7867
* The t-score for our two sample one sided t-test is 2.1538 and our p-value is .01743. Based on a significence level of .90, we will reject the null hypothesis that type b drugs have a at least as large of an effect on lambda protien drop as type a drugs. The 90% confidence interval for the diffirence in means is 1.1182675 - 1.86082.
* Based on the data, there is sufficent evidence to support that paitents taking type-a drugs have a larger drop in lambda protien levels then paitents taking type-b drugs. Because the paitents were selected at random, we can infeer this result to the remainer of paitents with Meloma. Because paitents were not randomly assisgned to each drug group, we cannot infer that this relationship is causal.

### a. BONUS (5 pts.) KAPPA PROTEIN ANALYSIS: The researchers thought that Revlimid would be more effective than the Velcade in reducing the Kappa protein levels. One problem they encountered was that the Kappa percent reduction was missing for many of the Revlimid and Velcade patients (as can be seen in the data set). Conduct a complete analysis that will test the claim that the Revlimid has a greater mean or median percent drop in Kappa protein levels than Velcade. (State the problem, address the assumptions, conduct the 6-step test, and provide a scope of inference.) Use an alpha = .01 level of significance and provide confidence intervals with your analysis.

boxplot(kappadrop~drug, data=meloma, main='Kappa drop')

|  |  |
| --- | --- |
|  | |
|  |  |

mel\_kappa <- na.omit(meloma)  
revl <- subset(mel\_kappa, drug == 'Revlimid')  
not\_revl <- subset(mel\_kappa, drug != 'Revlimid')  
hist(revl$kappadrop)

hist(not\_revl$kappadrop)

* The null hypothesis is that the mean drop in Kappa protiens for paitents who took the Revlimid drug is not higher then the mean drop for the other three drugs. The alternative hypothesis is that the mean drop for the Revlimid drug is higher then the other drugs in the study.
* In this observational study, the data is non normal, the sample sizes are significently diffirent, and the variance is diffirent between the two groups. We can assume indepndence since the paitents were chosen at random, but the data is paired, so it is independent enough for our testing. Because we have not met the assunmptions for a t-test or welche's test, we will be performing a Wilcoxon rank sum test on the null hypothesis.

wilcox.test(revl$kappadrop, not\_revl$kappadrop, conf.level= .01, alternative='greater')

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: revl$kappadrop and not\_revl$kappadrop  
## W = 331, p-value = 0.4341  
## alternative hypothesis: true location shift is greater than 0

* Since we are using a non-parametric testing method, we will skip the critical value for this test
* The p-value for this expirement is .4341, so we fail to reject the null hypothesis based on a .90 percent sigificence level on a one sided Wilcoxon Rank Sum test.
* There is not sufficent evidence at the .90 significence level to show the median drop in kappa protien levels is higher for the Revlimid drug is higher then the other drugs tested in this observational study. Since we have not rejected the null hypothesis, and this is an observational study, we have not identified a causal relationship in this expirement,.