Lecture 3 - Ahmed's Submission

Ahmed Nadeem

2024-10-03

I'm not sure why my R code is going over the text, but here is my github repo for this course, the file for this is called "testingForLec3Again.Rmd"

We're going to simulate some data that can be used in each of the following tests.

The response variable is normally distributed, and then we're going to back calculate some predictor variables.

NOTE: There are 9 questions below to answer. Do your best to answer them in full sentences. This is part of what we're practicing in this class.

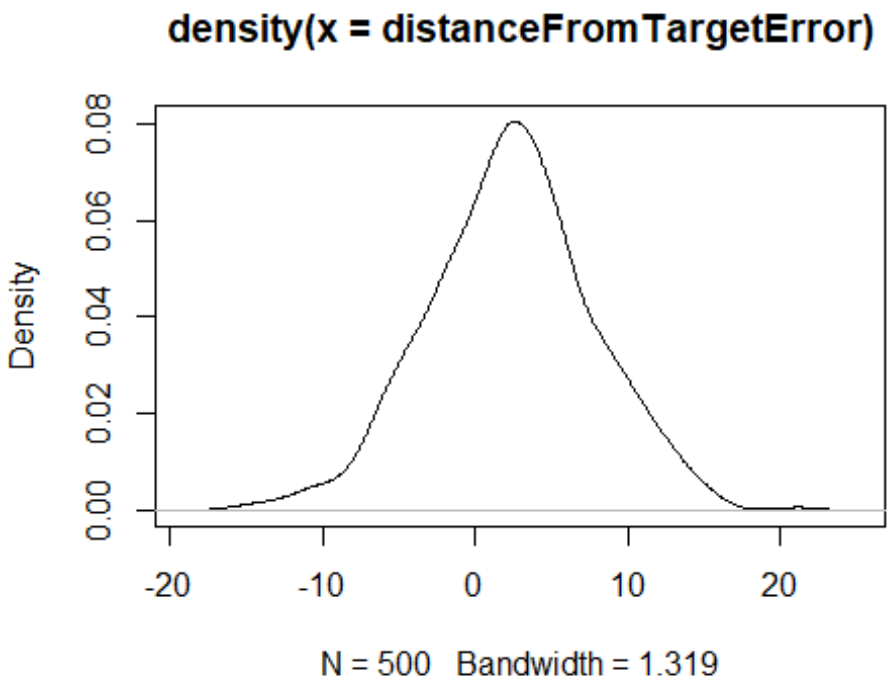
QUESTION 1: Describe the trait that you're simulating. This can be a real trait that you are working on, or hope to work on, or it can be completely made up. - Answer: I am working on the perception of space, so that means I want to see how well people are able at understanding what they look at and how well do they interact and move in that place that they look at. So for the later part, I literally measure how well you can place an item down at a given location (given some handicap or sensation cue). And to see if there are biases in the way you live/interact in the world, I have to measure how "off" you were from going to a location. And that is what `distanceFromTargetError` means in the little snippet below!

Pay special attention to the N that you're expecting, the mean, and the sd around the mean.

```
rnorm(500, 2.5, 5)->distanceFromTargetError ### change this to be a trait
that you're actually interested in, with an appropriate distribution!

#

plot(density(distanceFromTargetError))
```




```
min(distanceFromTargetError)

## [1] -15.19582

max(distanceFromTargetError)

## [1] 21.16281
```

QUESTION 2: Describe the predictor variable. What does this mean biologically to your trait of interest. How did you decide on the numbers in yes on line 33? - Answer: Biologically speaking, this `distanceFromTargetError` is me measuring the error from a given location and the location chosen by a participant. In biological terms, I am (hopefully) measuring the brain's internal representation of space based on the cues it used to understand space AND THEN I will also be measuring how well the motor response was at recreating and adjusting/using that internal representation. ANd how did I decide it lies on a line? well, if space itself lies on a line (3 lines cuz 3 axes) so that means it is natural for me to say that the error also lies on a line (error = actual position - position placed by participant)

```
yes<-sample(distanceFromTargetError, size = 30, replace=FALSE, prob =
ifelse(test = distanceFromTargetError>2.5, yes = 0.95,no =  0.15)) #### play
with this line! Is the test statistically significant. When is it not?
predictor_t.test<-(distanceFromTargetError %in% yes)

cbind.data.frame(distanceFromTargetError, predictor_t.test)->data

mean(data[which(data$predictor_t.test==TRUE),1])
```

```
## [1] 4.719865
```

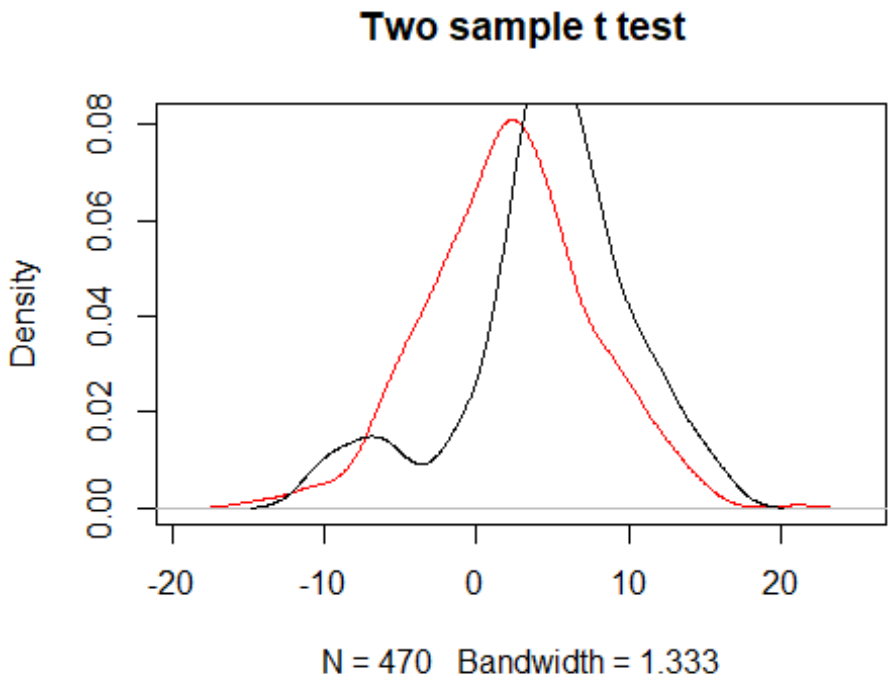
```
mean(data[which(data$predictor_t.test==FALSE),1])
```

```
## [1] 2.185718
```

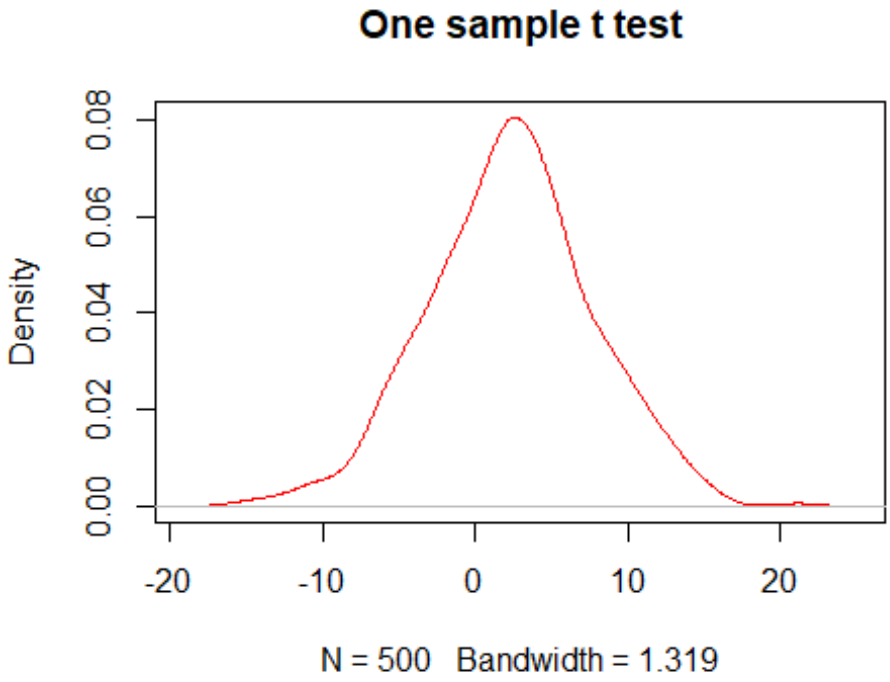
```
t.test(distanceFromTargetError~predictor_t.test, data=data) ### this does a two sample t-test. What would a one sample t test be testing? How would you do that?
```

```
##
##  Welch Two Sample t-test
##
## data:  distanceFromTargetError by predictor_t.test
## t = -2.4126, df = 32.473, p-value = 0.02166
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -4.6725159 -0.3957776
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2.185718            4.719865
```

```
### plots our two samples for distribution
plot(density(data[which(data$predictor_t.test==FALSE),1]), col="red", main="Two sample t test")
lines(density(data[which(data$predictor_t.test==TRUE),1]), ylim=c(0, 0.1), xlim=c(-20,20), main="Two Sample T test")
```



```r
###plot one sample distribution
plot(density(data$distanceFromTargetError), col="red", main="One sample t
test")
```




```r
t.test(data$distanceFromTargetError) ### what is this test doing?;
t.test(data$distanceFromTargetError, mu = mean(distanceFromTargetError)) ###
what is this test doing?

##
##  One Sample t-test
##
## data:  data$distanceFromTargetError
## t = 9.7022, df = 499, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.864363 2.811171
## sample estimates:
## mean of x
##  2.337767
```

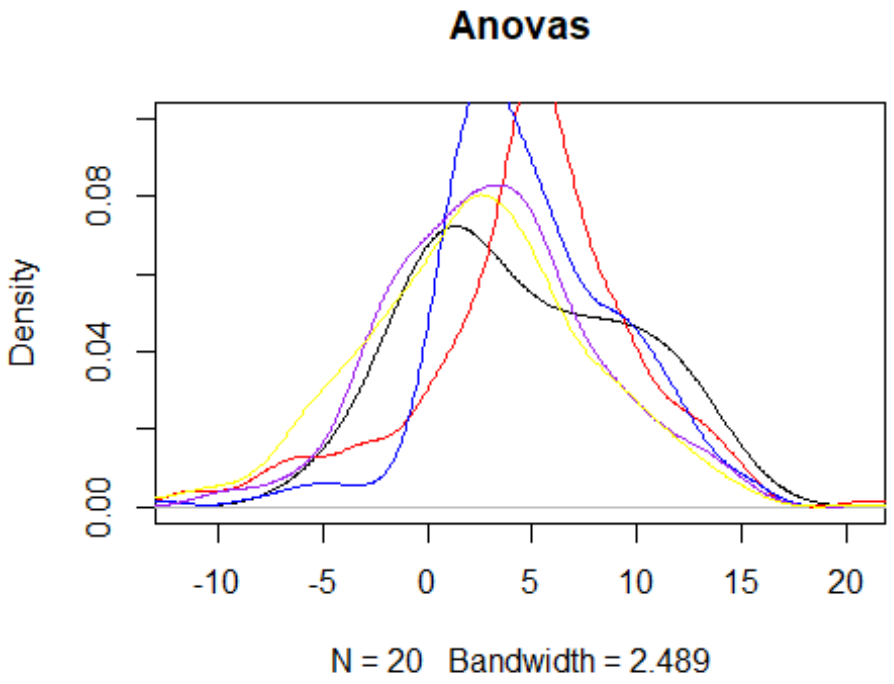
QUESTION 3: Write one sentence where you report your t.text. - Answer: if I had a sample of 30 participants, where i picked these samples based on that prob function(took me a while to understand LOL). then i would not have statstically signifcant data cuz my SD from the first code block is large. making the 2 samples VERY big, however, the one sample t test, comparing to a mean of 0 (which is not my mean of my data set.)

Next we're going to move to Anova. So, the first thing we'll do is break our response variable (same one!) into 5 different categories, just as we did for the t-tests.

QUESTION 4: Describe the predictor variable. What does this mean biologically to your trait of interest. How did you decide on the numbers in lines 60, 61, 62, 63? - Answer: I picked the sample size numbers based on how many participants I usually get in my studies (around 30-50).

```
trait_of_interest = distanceFromTargetError
test1<-sample(trait_of_interest, 20, replace=FALSE, prob =
ifelse(trait_of_interest>7, 0.95, 0.15))
test2<-sample(trait_of_interest[which(trait_of_interest %in% test1 ==
FALSE)], 200, replace=FALSE, prob =
ifelse(trait_of_interest[which(trait_of_interest %in% test1 == FALSE)] >4,
0.95, 0.15))
test3<-sample(trait_of_interest[which(trait_of_interest %in% test2 == FALSE |
trait_of_interest %in% test1 ==FALSE)], 200, replace=FALSE, prob =
ifelse(trait_of_interest[which(trait_of_interest %in% test2 == FALSE |
trait_of_interest %in% test1 == FALSE)] >0, 0.95, 0.15))
test4<-sample(trait_of_interest[which(trait_of_interest %in% test2 == FALSE |
trait_of_interest %in% test1 ==FALSE | trait_of_interest %in% test3 ==
FALSE)], 200, replace=FALSE, prob =
ifelse(trait_of_interest[which(trait_of_interest %in% test2 == FALSE |
trait_of_interest %in% test1 ==FALSE | trait_of_interest %in% test3 ==
FALSE)] >-4, 0.95, 0.15))
test5<-trait_of_interest[which(trait_of_interest %in% test1 == FALSE|
trait_of_interest %in% test2 == FALSE | trait_of_interest %in% test3 == FALSE
| trait_of_interest %in% test4 == FALSE)]


plot(density(test1), ylim=c(0, 0.1), main="Anovas")
lines(density(test2), col="red")
lines(density(test3), col="blue")
lines(density(test4), col="purple")
lines(density(test5), col="yellow")
```




```
anova_predictor<-data.frame(ifelse(trait_of_interest %in% test1 == TRUE,
"group1", ifelse(trait_of_interest %in% test2 == TRUE, "group2",
ifelse(trait_of_interest %in% test3 == TRUE, "group3",
ifelse(trait_of_interest %in% test4 == TRUE, "group4", "group5")))))

data2<-cbind.data.frame(data, anova_predictor)
names(data2)<-c("trait_of_interest", "predictor_t.test", "anova_predictor")


anova(aov(trait_of_interest~anova_predictor, data=data2)) ### what does this
do?

## Analysis of Variance Table
##
## Response: trait_of_interest
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## anova_predictor   4   3382  845.50  37.693 < 2.2e-16 ***
## Residuals       495  11103   22.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(lm(trait_of_interest~anova_predictor, data=data2)) ### what does this
do?

## Analysis of Variance Table
##
## Response: trait_of_interest
```

```
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## anova_predictor    4   3382  845.50  37.693 < 2.2e-16 ***
## Residuals        495  11103   22.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(trait_of_interest~anova_predictor, data=data2)) ### what does this do? What do you notice about the last three tests?
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## anova_predictor    4   3382   845.5   37.69 <2e-16 ***
## Residuals        495  11103    22.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### what information is missing here that you wished you had to understand your study better?
analysis_of_variance<-aov(trait_of_interest~anova_predictor, data=data2)
##name the model to keep it for downstream
TukeyHSD(analysis_of_variance, conf.level = 0.95) ### what does this do, and where are the differences?
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = trait_of_interest ~ anova_predictor, data = data2)
##
## $anova_predictor
##                       diff       lwr         upr     p adj
## group2-group1  0.2846406 -2.756362  3.32564349 0.9990487
## group3-group1 -1.3120424 -4.527574  1.90348898 0.7974956
## group4-group1 -5.0753232 -8.409830 -1.74081667 0.0003499
## group5-group1 -5.5490699 -8.662028 -2.43611157 0.0000141
## group3-group2 -1.5966830 -3.262019  0.06865276 0.0674157
## group4-group2 -5.3599638 -7.244807 -3.47512034 0.0000000
## group5-group2 -5.8337106 -7.291178 -4.37624321 0.0000000
## group4-group3 -3.7632808 -5.918409 -1.60815241 0.0000227
## group5-group3 -4.2370276 -6.030392 -2.44366351 0.0000000
## group5-group4 -0.4737468 -2.472608  1.52511448 0.9668117
```

QUESTION 5: Write one sentence where you report your ANOVA and Tukey tests. What did you find, and how do you report this? - Answer: The ANOVA test shows me that there is a lot of variablity that is not explained by the groups (lots of noise). This is because the resdiuals value is large. Then there is als oa large F value, which shows me that there is a large diff b/w the groups relative to their variablity.

Again, our simulations aren't to be actually causal, I'm simulating predictor variables to fit a response I've already made. Normally, we would simulate the response variables from more thoughtful predictor variables.

QUESTION 6:what is the difference between the assumed distributions for the prior predictor variables, and this one? - Answer: the difference is that for the 5 categories made for the anova are based on the data set itself, but the last 2 were just based on random chance (if i understand the code correctly).

QUESTION 7: Describe the predictor variable. What does this mean biologically to your trait of interest. How did you decide on the numbers in line 104? - Answer: IT would show me an aspect that is related to the motor inaccuracy in humans, such as distance or percentage of a cue shown to an individual/participant.

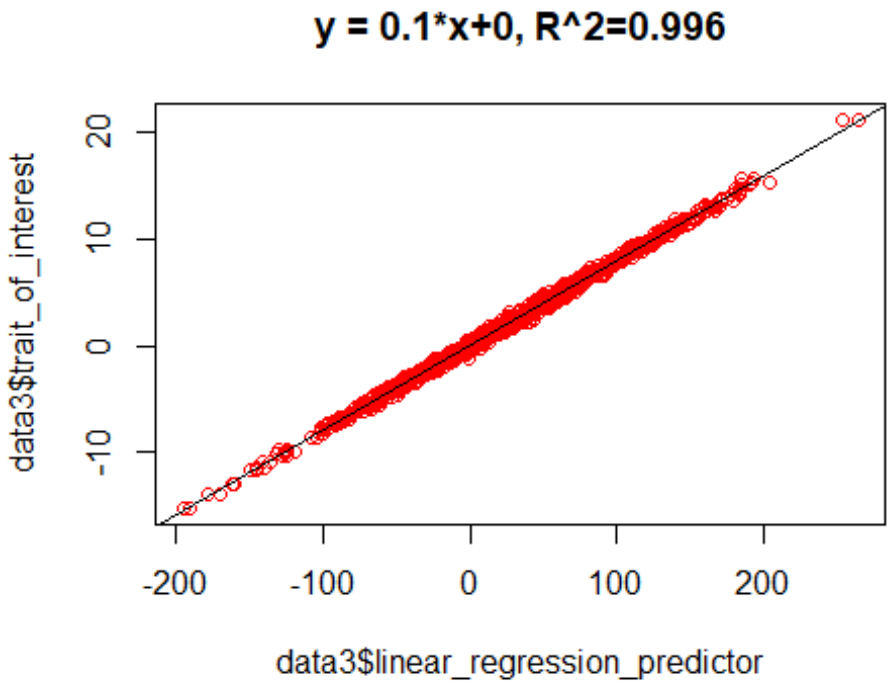
QUESTION 8: What is the difference between a regression and a correlation? When would you use each? How does the test stat from the correlation compare to the effect size from the regression? - Answer: 1. Regresssion is used to tell how associated two variables are, and it tells me how to predict them with a line of best fit 2. Correlation are used to tell me what direction and how tightly the two variables "travel" together so to speak (strong or weak) and if it going up, or down on a graph (positive and negative correlations) 3. I would use a correlation when I want to know if there is a trend in any data set, i would use a regression to predict one variable against another. 4. The correlation (0.977) and r^2 of 0.9954 show me that there is a strong correlation and predictive power in my data set.

```
linear_regression_predictor<-12.5*trait_of_interest+rnorm(1000, 0, 4) ### change these numbers!! Remember that this is backwards from how we did this on day 1, so the slope should go the other way!
data3<-cbind(data2, linear_regression_predictor)

lm<-lm(trait_of_interest~linear_regression_predictor, data=data3)
summary(lm)### what is the output here? What are we interested in understanding from this model? How do we get m? How do we get the intercept?

##
## Call:
## lm(formula = trait_of_interest ~ linear_regression_predictor,
##     data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2143 -0.2075 -0.0151  0.2346  0.9668
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.0174154  0.0110254    1.58    0.115
## linear_regression_predictor 0.0794673  0.0001497  530.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3201 on 998 degrees of freedom
## Multiple R-squared:  0.9965, Adjusted R-squared:  0.9965
## F-statistic: 2.818e+05 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
eq = paste0("y = ", round(lm$coefficients[2],1), "*x",
"+",round(lm$coefficients[1],1), ", R^2=", round(summary(lm)$adj.r.squared,
digits=3))
plot(data3$linear_regression_predictor, data3$trait_of_interest, col="red",
main=eq)
abline(lm, col="black")
```




```r
### with the same data:
cor.test(data3$trait_of_interest, data3$linear_regression_predictor) ###
compare the sample estimate to the 1) simulated effect sizes and 2) to the
estimated effect size

##
##  Pearson's product-moment correlation
##
## data:  data3$trait_of_interest and data3$linear_regression_predictor
## t = 530.83, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9980006 0.9984398
## sample estimates:
##       cor
## 0.9982338

### how does the correlation estimate change when you change line 104?
```

QUESTION 9: Report your regression and correlation in a sentence. Differentiate between them and what you report for each. - Answer: The linear regression model indicated that for every unit increase in linear_regression_predictor, the trait_of_interest increases by 0.079 units (R^2 = 0.9954, p < 2e-16). The correlation between the two variables was 0.9977 (p < 2e-16), indicating a strong positive relationship.