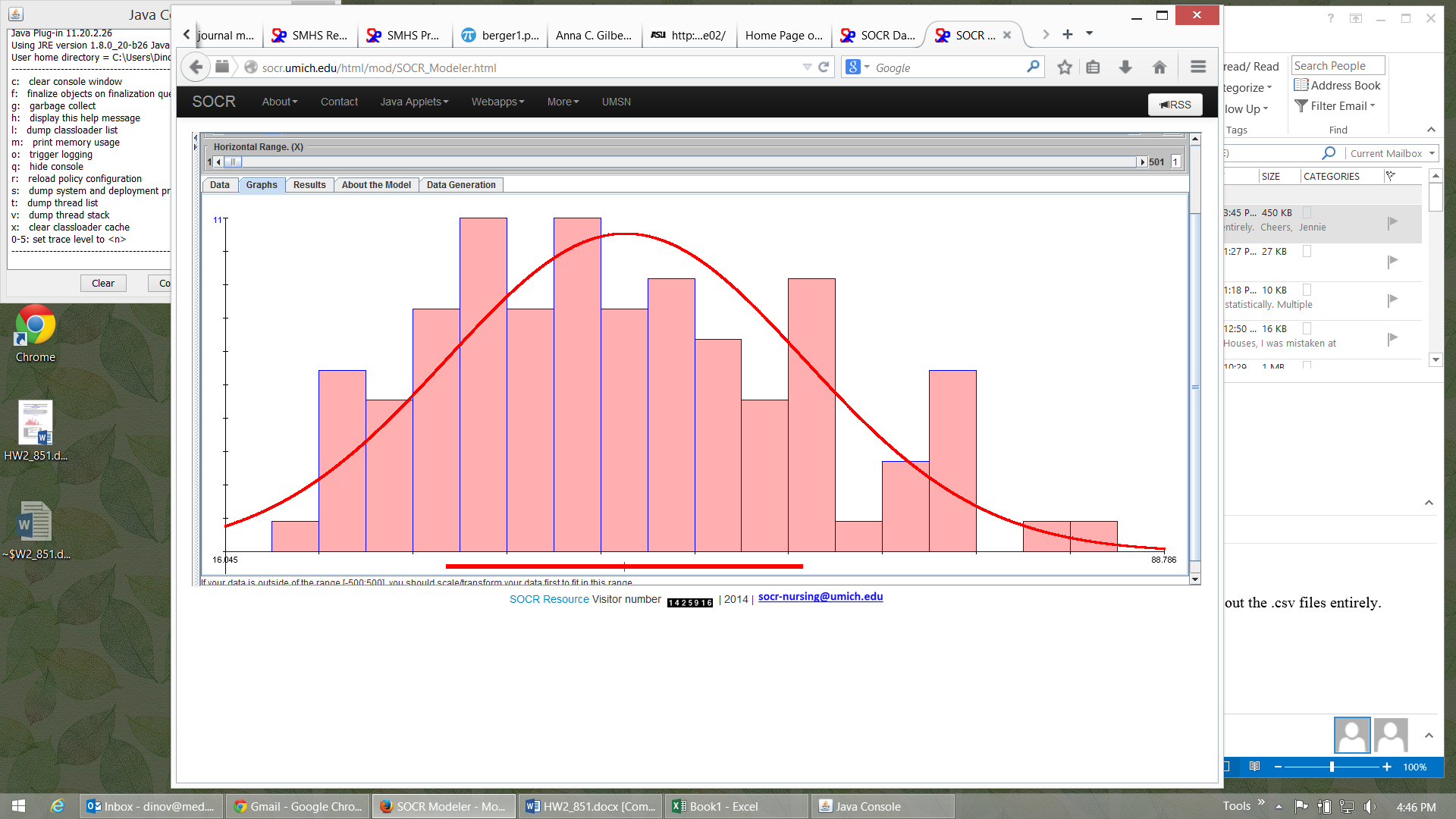
**Scientific Methods for Health Sciences: Applied Inference (HS851): Fall 2014**

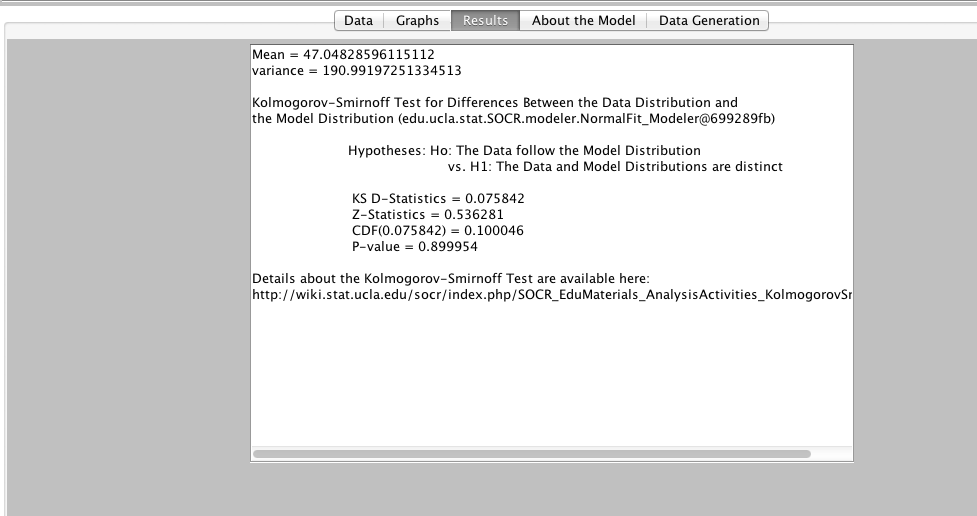
[**http://www.socr.umich.edu/people/dinov/2014/Fall/HS851**](http://www.socr.umich.edu/people/dinov/2014/Fall/HS851)

**Homework 2 Solutions**

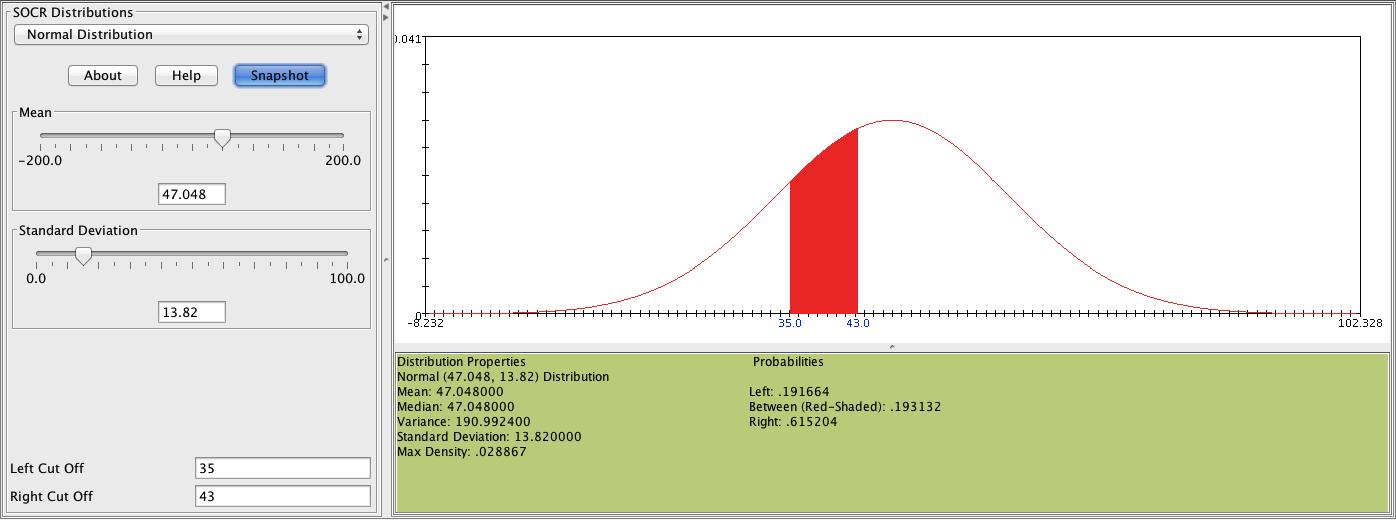
 **Problem 1**: There is great interest in comparing different countries in the world based on a variety of factors reflecting the country's internal and external international ranking. Use the [Political, Economic, Health, and Quality-of-Life Data](http://wiki.socr.umich.edu/index.php/SOCR_Data_2008_World_CountriesRankings) of 100 Countries to estimate the probabilities below. Let ED=Economic Dynamism of a Country, which is an index of productive growth in US dollars. Use the [SOCR Modeler](http://socr.umich.edu/html/mod) to fit a Normal Distribution Model to the ED variable (column) in this dataset (see [this Help page](http://www.socr.umich.edu/people/dinov/2014/Fall/HS851/HWs/wiki.socr.umich.edu/index.php/Help_pages_for_SOCR_Modeler)). Once you obtain estimates for the *mean* and *standard deviation* of the normal model (see the Results tab in the Modeler) use the [SOCR Normal Distribution Calculator](http://socr.ucla.edu/htmls/dist/Normal_Distribution.html) to estimate the likelihoods of these events:

Histogram of ED data



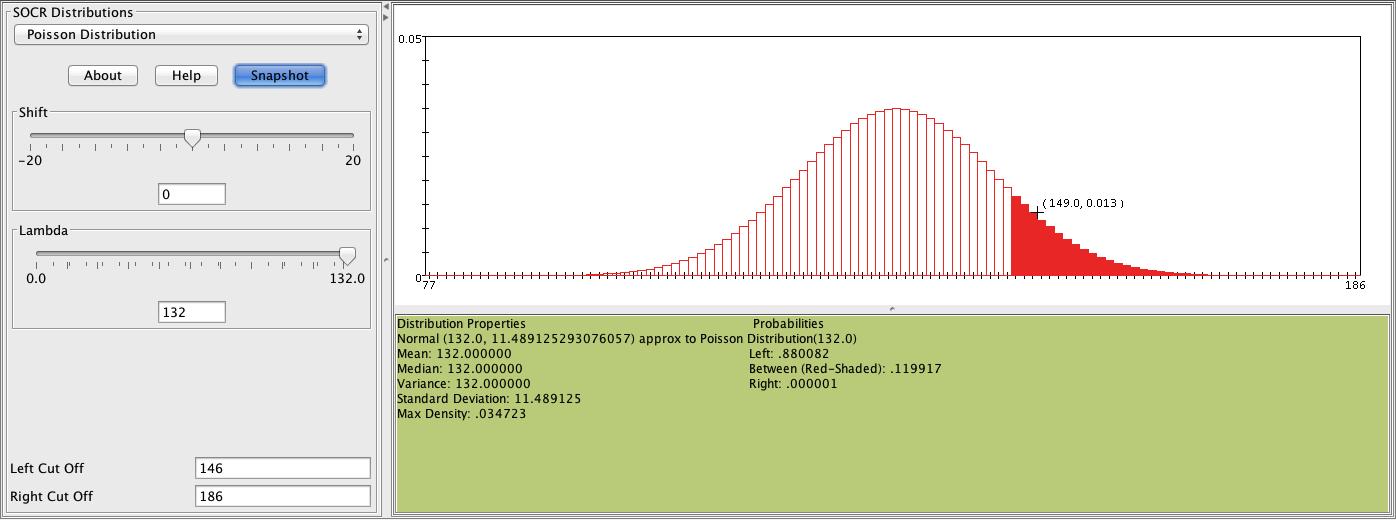


* Sample snapshot for calculating P(35 ≤ ED ≤ 43) = 0.1931296

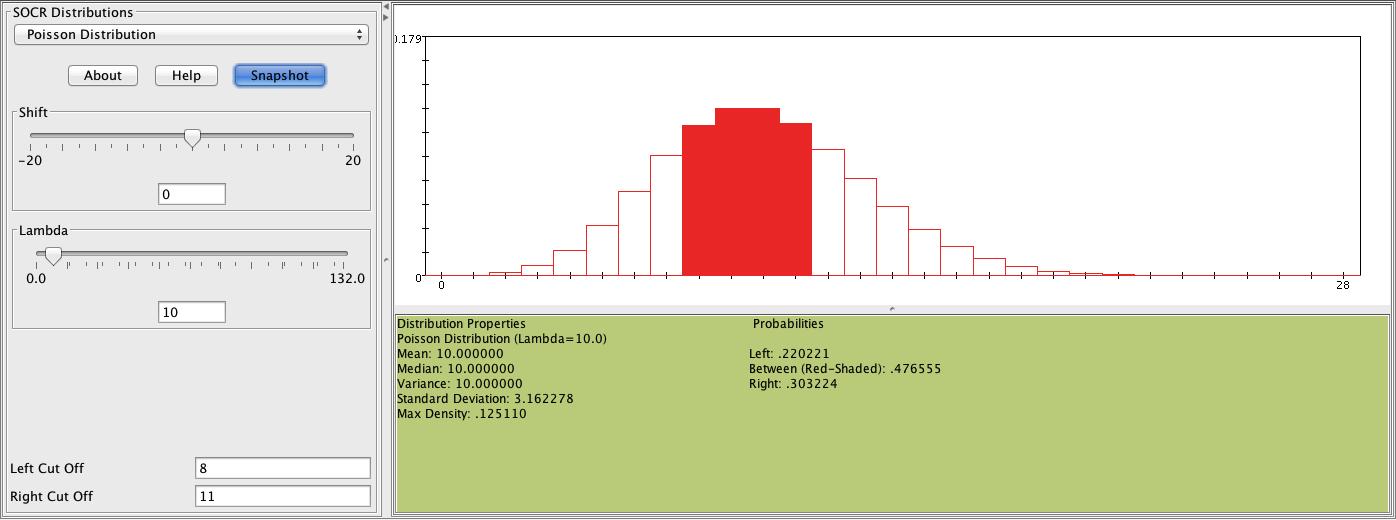


* P(ED ≤ 46) = 0.47
* P(35 ≤ ED ≤ 43) = 0.193
* P(48 ≤ ED) = 0.47
* P(53 ≤ ED) = 0.33
* P(47 ≤ ED ≤ 87) = 0.50
* P(15 ≤ ED ≤ 51) = 0.60

**Problem 2**: During a typical 24-hour shift in the ER, the healthcare providers (doctors, nurses, staff) expect to see 132 emergency visits including 10 traumatic brain injuries (TBIs). Find the probability that the ER team will see over 145 cases in total and the probability that there will be between 8 and 11 TBIs within a given 24-hour period. Recall that the [Poisson Distribution](http://wiki.socr.umich.edu/index.php/AP_Statistics_Curriculum_2007_Distrib_Poisson) can be used as a model.



P(# of patients>145) = 0.12



P(8 <= # TBI <= 11) = 0.48

 **Problem 3**: Many clinical and translational studies involve multiple variables (or events) that may be independent of one another or closely associated. Identifying and untangling data dependencies is critical in such situations. We can use the [SOCR Coin Die Experiment](http://socr.ucla.edu/htmls/exp/Coin_Die_Experiment.html) to *simulate* dependence between clinical variables. Suppose we have 2 discrete clinical variables, for example, X={stage of melanoma} (categorical) and Y={gender} (dichotomous). We can simulate this situation, specifically simulate event independence between the outcome of a die (event B, representing the cancer stage) and the outcome of a coin (event A, representing the patient gender). In the [SOCR Coin Die Experiment](http://socr.ucla.edu/htmls/exp/Coin_Die_Experiment.html) set the probabilities of both dice to be identical. Run 100 experiments and argue that the observed data implies independence between the events A={Coin=Head, say male} and B={Die=3, say stage 3 melanoma}, i.e., *P(AB) = P(A) P(B)*, approximately. You basically need to count the proportion of times each of the tree events (A, B and C={A∩B})of interest occur in the 100 experiments and validate (or disprove) the equality above. Also, try this with a larger number of experiments (e.g., n=10,000). Next, make the probability distributions of the two dice different (by clicking on the dice and manually changing the die probabilities). Show empirically the dependence of the probabilities, A={Coin=Head} and B={Die=3}. Do we have evidence of independence or association in the outcomes?

Results with two fair dice

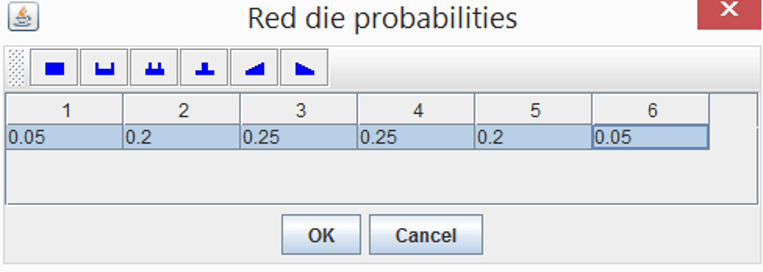
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Head | 4 | 6 | 4 | 10 | 12 | 7 | 43 |
| Tail | 6 | 10 | 10 | 12 | 12 | 7 | 57 |
| Total | 10 | 16 | 14 | 22 | 24 | 14 | 100 |

P({Coin=head} ⋂ {Die=3}) = 4/100 = 0.04

P(Coin=head) × P(Die=3) = (43/100) × (14/100) = 0.06

The numbers are similar, so the events are likely independent.

Results with one unfair die.



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Head | 9 | 6 | 7 | 7 | 8 | 11 | 48 |
| Tail | 8 | 9 | 9 | 8 | 9 | 9 | 52 |
| Total | 17 | 15 | 16 | 15 | 17 | 20 | 100 |

P({Coin=head} ⋂ {Die=3}) = 7/100 = 0.07

P(Coin=head) × P(Die=3) = (48/100) \* (16/100) = 0.08

These numbers are not far off, so there does not seem to be strong evidence for independence. Now let’s try these with larger sample sizes (i.e., 10,000 instead of 100 runs).

With larger sample size and fair dice:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Head | 827 | 857 | 812 | 848 | 811 | 887 | 5042 |
| Tail | 839 | 825 | 762 | 827 | 841 | 864 | 4958 |
| Total | 1666 | 1682 | 1574 | 1675 | 1652 | 1751 | 10000 |

P({Coin=head} ⋂ {Die=3}) = 812/10000 = 0.08

P(Coin=head) × P(Die=3) = (5042/10000)\*(1574/10000) = 0.08

These probabilities are equal and the sample size is large, so it seems likely the events are independent.

With larger sample size and one unfair die

(Probabilities of 1-6 after flipping a **tail**, respectively, 0.05 0.20 **0.25** 0.25 0.20 0.05)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |  |
| Head | 830 | 797 | 815 | 833 | 812 | 812 | 4899 |
| Tail | 244 | 1051 | 1323 | 1224 | 1008 | 251 | 5101 |
|  | 1074 | 1848 | 2138 | 2057 | 1820 | 1063 | 10000 |

P({Coin=head} ⋂ {Die=3}) = 815/10000 = 0.08

P(Coin=head) × P(Die=3) = (4899/10000) \* (2138/10000) = 0.1

The probabilities are different this time! So, the events seem dependent.

Let’s see how the results compare to their expectations, given their exact probabilities:

P({Coin=head} ⋂ {Die=3})= P(3 | head) \* P(head) = (1/6)\*0.5 = 0.08

P(Coin=head) × P(Die=3) = 0.5\*(0.5\*(1/6) + 0.5\*(1/4)) = 0.1,

where P(3) = P(3|head)P(head) + P(3|tail)P(tail), and a head coin outcome requires

tossing the fair die, whereas a tail coin outcome requires a toss of the loaded die.

**Problem 4**: Using the [SOCR Clinical, Genetic and Imaging Data of Alzheimer’s Disease](http://wiki.socr.umich.edu/index.php/SOCR_Data_AD_BiomedBigMetadata):

Part 1: Using these 2 groups: Group0={CDGLOBAL=0} vs. Group1={CDGLOBAL=1}, compute the correlations between systolic and diastolic blood pressure (VSBPSYS and VSBPDIA) within each group (ro and r1). Then test a [hypothesis for the equivalence of these correlations](http://wiki.socr.umich.edu/index.php/SMHS_SLR#Statistical_inference_on_correlation_coefficient).

Correlation for CDGLOBAL=0: 0.44 (r.0)

Correlation for CDGLOBAL=1: 0.41 (r.1)

Test for difference between correlation coefficients:

Fisher-transformed r00 = 0.5\*log(abs((1+r.0)/(1-r.0))) = 0.47

Fisher-transformed r11 = 0.5\*log(abs((1+r.1)/(1-r.1))) = 0.43

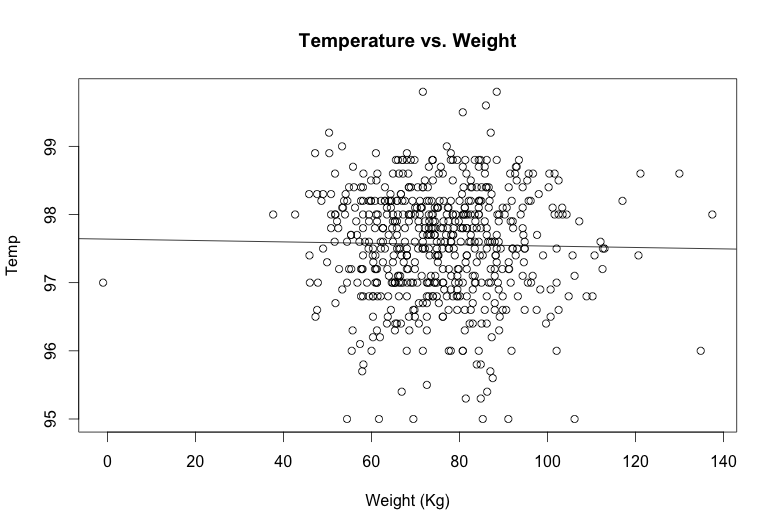
Z-score = (r00-r11)/sqrt((1/(695-3))+(1/(48-3))) = 0.27

We can now use this Z-score to assess the probability of seeing a result this extreme or more extreme by chance under the Null hypothesis (given that the two correlations are in fact equal) by using the 2-tailed probability in the normal distribution past 0.27 SDs from 0 on either side. We get p=0.78. This means, if the two means were in fact equal, we would expect a difference between correlation coefficients this great or greater by chance 78% of the time. We therefore **fail to reject the null hypothesis** that the two correlation coefficients are equal. The interpretation of this results is that the data does not support evidence suggesting the correlations between systolic and diastolic blood pressure are different between the 2 cohorts, Group0={CDGLOBAL=0} vs. Group1={CDGLOBAL=1}.

Part 2: Fit a [simple linear model](http://wiki.socr.umich.edu/index.php/SMHS_SLR) for VSTEMP and Weight\_Kg. Formulate and assess a hypothesis about trivial slope of the regression curve on these two variables. Elaborate on your findings.

H0: Temperature and weight are independent and the slope of the best fit line equals 0.

Ha: Temperature and weight are correlated and the slope of the best fit line is significantly different from 0.



Slope: -0.001

Standard error: 0.002

Zero is within one standard error of the estimated slope, so we are not at all confident that the slope is significantly different from 0. We fail to reject the null hypothesis, H0.

**Appendix: R Code**

###################################

# Jennie Lavine

# 9/17/2014

# HW #2, HS851

##################################

setwd('~/hw2\_851')

###############################

#Problem 1

###############################

world.dat <- read.csv('world\_data.csv')

ED<-world.dat$ED

hist(ED)

mean(ED)

sd(ED)

var(ED)

#pnorm is the R function that computes the CDF,

#that is, it computes the area under the curve to the left of your cut-off value (q)

# P(ED ≤ 46)

pnorm(q=46,mean=mean(ED), sd=sd(ED))

# P(35 ≤ ED ≤ 43)

pnorm(q=43,mean=mean(ED), sd=sd(ED)) - pnorm(q=35,mean=mean(ED), sd=sd(ED))

# P(48 ≤ ED)

1 - pnorm(q=48,mean=mean(ED), sd=sd(ED))

# P(53 ≤ ED)

1 - pnorm(q=53,mean=mean(ED), sd=sd(ED))

# P(47 ≤ ED ≤ 87)

pnorm(q=87,mean=mean(ED), sd=sd(ED)) - pnorm(q=47,mean=mean(ED), sd=sd(ED))

# P(15 ≤ ED ≤ 51)

pnorm(q=51,mean=mean(ED), sd=sd(ED)) - pnorm(q=15,mean=mean(ED), sd=sd(ED))

################################

#Problem 2

##############################

# We assume the clinic entries and TBI are Poisson distributed with parameter lambda = 132 and 10, respectively.

1 - ppois(q=145, lambda=132)

ppois(q=12, lambda=10) - ppois(q=8, lambda=10)

##############################

#Problem 3

#############################

#simulate 100 coinflips followed by 100 die rolls

#Define the number of repetitions of the experiment, N

N=10000

#Set up an empty matrix to put the results in

sim.res<-matrix(NA, nrow=N, ncol=2)

colnames(sim.res)<-c('coin','die')

sim.res<-as.data.frame(sim.res)

#Define the probabilities of each value on the dice for die 1 and 2

#This example shows one die with different probabilities.

die1.probs<-rep(1/6, 6)

die2.probs<-c(1,4,5,5,4,1)

die2.probs<-die2.probs/sum(die2.probs)

#Simulate coin flips using the function 'sample'

sim.res[,'coin']<-sample(c(0,1), size=N, replace=T)

#Simulate die rolls from the fair die (die1)

#if a head is flipped (i.e., coin==0)

sim.res[sim.res[,'coin']==0,'die']<-sample(1:6, sum(sim.res[,'coin']==0),

replace=T, prob=die1.probs)

#Simulate die rolls from the unfair die (die2)

#if a tail is flipped (i.e., coin==1)

sim.res[sim.res[,'coin']==1,'die']<-sample(1:6, sum(sim.res[,'coin']==1),

replace=T, prob=die2.probs)

#Tabulate the results

table(sim.res)

#Calculate the margins

apply(table(sim.res), 1, sum)

apply(table(sim.res), 2, sum)

###########################

#Problem 4

###########################

#Read in the data

alz.dat<-read.csv('alzheimers.csv')

#Obtain subsets of the data that correspond to the rows

#for which the variable CDGLOBAL takes on

#the value of interest (i.e., 0 for one set and 1 for the other)

cdglob.0 <- subset(alz.dat, alz.dat$CDGLOBAL==0)

dim(cdglob.0)

cdglob.1 <- subset(alz.dat, alz.dat$CDGLOBAL==1)

dim(cdglob.1)

#Plot the data

par(mfrow=c(1,2))

plot(cdglob.0[,'VSBPSYS'], cdglob.0[,'VSBPDIA'], main='Diastolic vs Systolic

BP with no dementia', xlab='Systolic BP, mmHG', ylab='Diastolic BP, mmHG')

plot(cdglob.1[,'VSBPSYS'], cdglob.1[,'VSBPDIA'], main='Diastolic vs Systolic

BP with dementia', xlab='Systolic BP, mmHG', ylab='Diastolic BP, mmHG')

#Use the 'cor' function' to calculate the correlations

r.0 <- cor(cdglob.0$VSBPDIA, cdglob.0$VSBPSYS)

r.1 <- cor(cdglob.1$VSBPDIA, cdglob.1$VSBPSYS)

#Show that the correlation can be calculated as follows

#using cdglob.0 as an example.

x=cdglob.0$VSBPDIA

y=cdglob.0$VSBPSYS

n=length(x)

cor.xy <- 1/(n-1)\*sum(((x-mean(x))/sd(x))\*((y-mean(y))/sd(y)))

#this checks out, we get the same answer as using the 'cor' function

####Testing for equivalence of correlations

r00 <- 0.5\*log(abs((1+r.0)/(1-r.0)))

r11 <- 0.5\*log(abs((1+r.1)/(1-r.1)))

n.0 <-nrow(cdglob.0)

n.1 <- nrow(cdglob.1)

z.val <- (r00-r11)/sqrt((1/(n.0-3))+(1/(n.1-3)))

2\*(1-pnorm(z.val))

#########################

#Problem 4 part 2: simple linear model

#########################

y=alz.dat$VSTEMP

x=alz.dat$Weight\_Kg

far.ind<-which(y>70)

y<-y[far.ind]

x<-x[far.ind]

fit <- lm(y~x)

par(mfrow=c(1,1))

plot(x,y,main='Temperature vs. Weight',xlab='Weight (Kg)',ylab='Temp')

abline(fit)

summary(fit)