Computer/Notes Portion: This is optional, but if you submit it, it will be graded and counted towards your final exam score. It should be very straightforward. There are four questions—make sure to complete all of them. If you decide to submit this portion, upload your completed writeup AND the SAS/R code to the D2L Dropbox folder before 8 PM Wednesday night. In all questions where you are testing a hypothesis, make sure to explicitly state the p-value and the conclusion as part of your written response. YOU MUST WORK ALONE—NO HELP!!! You are allowed to use any course materials that I have uploaded in D2L, your textbook, class notes, and even internet searches. But you can’t get help from anyone inside or outside of the class, and you can’t ask anybody questions. And clearly, don’t compare answers with anybody.

Good luck, and thanks for taking the class!

1. The data set Pima.tr (available in R after loading the MASS package, but also saved as Pima\_tr.csv in D2L) contains measurements on 200 Pima Indian women. We’ll focus on using the triceps skinfold thickness (skin in the data set) to predict body mass index (bmi). That is, the explanatory variable is skin and the response variable is bmi.

a) Test whether or not skin and bmi are associated using Spearman’s correlation. Use alpha=.01, and make sure to explicitly state your p-value and the conclusion of your test.



R code:

data <- read.csv("/Users/sbuciuma/Desktop/Final\_NP/Pima\_tr.csv")

head(data)

#test for significant association between our variables using computed spearman correlation

spearman.test <- cor.test(data$bmi, data$skin, method = "spearman", exact = T, conf.level = 0.90)

spearman.test

Conclusion: p-value < 2.2e-16, We have enough evidence to conclude that a significant correlation between BMI and skin with Spearman’s rank correlation rho exist and does point us to that is “strong” correlation between our variables.

b) Test whether or not skin and bmi are associated using Kendall’s tau. Use alpha=.01, and make sure to explicitly state your p-value and the conclusion of your test.



R code:

kendall.test <- cor.test(data$bmi, data$skin, method = "kendall", exact = T, conf.level = 0.90)

kendall.test

Conclusion: p-value < 2.2e-16, We have enough evidence that a significant strong correlation between BMI and skin with Kendall’s rank correlation tau = 0.522291.

c) Test whether or not skin and bmi are associated using Pearson’s correlation. Use alpha=.01, and make sure to explicitly state your p-value and the conclusion of your test.



Conclusion: p-value < 2.2e-16

Conclusion: p-value < 2.2e-16, We have enough evidence that a significant strong correlation between BMI and skin with Pearson cor = 0.6590356.

d) Which of the measures of association (Spearman’s correlation, Kendall’s Tau, or Pearson’s correlation) do you think is most appropriate to summarize these data? Justify your choice.

Conclusion: I think for this model the best I think will be Pearson’s correlation Pearson because is most appropriate for measurements taken from an *interval* scale, and in our case we have interval scale like BMI and skin.

e) Use **ordinary least squares** to fit the simple linear regression line. Report the estimated regression equation (i.e., y-hat=b0+b1x) below with at least three decimal places of accuracy for the intercept and for the slope:



Answer:



Conclusion: p-values < alpha so statistically the model is valid.

f) Using bivariate sampling, create a 99% confidence interval for the population slope, β1, using the BCA method.

R code:

#FOR 99% BETAS

betas <- function(indata,indices) {

tempdata <- indata[indices,]

tempfit <- lm(bmi ~ skin, data=tempdata)

tempcoef <- coef(tempfit)

return(c(b1=tempcoef[2]))

}

boots<- boot(data, betas, R=10000)

boots

boot.ci(boots,conf=0.99)



Conclusion: Ci for conf level = 99% using BCA is ( 0.1672, 0.4475 ) of the regression line by using 10000 bootstrap samples.

g) Using bivariate sampling, create a 99% confidence interval for the population slope using the percentile method.





h) Compare the confidence intervals you obtained in parts f and g. Which do you prefer, and why?

Answer: I do prefer one using BCA because is easier to compute and good interval is present. But both does provide basically same interval.

1. Fit a loess regression (local linear or local quadratic—your call) to estimate the mean BMI as a function of skin thickness and add 95% pointwise confidence bands. Include this plot here.









j) Was it appropriate to fit the straight-line relationship between skin and BMI (as in part e)? Justify your answer.

Answer: Yes, a straight-line was appropriate per our result plot.

1. Small amount of bonus: Using fixed-x sampling, create 99% confidence intervals for the population slope using the percentile and BCa methods. How do these intervals compare with those obtained in parts f and g? And do you think the fixed-x or bivariate sampling was more appropriate to use?
2. A) For the following data of 12 observations (also saved in finalQ2.csv), report a 95% residual bootstrap confidence interval for the population variance, sigma^2.

0.01 4.99 4.11 13.52 0.33 2.47 9.70 3.03 2.42 17.39 39.12 7.87

R code

a<- c(0.01,4.99,4.11,13.52,0.33,2.47,9.7,3.03,2.42,17.39,39.12,7.87)

theta.est <- var(a)

B <- 9999

# A vector to store the residuals

residthetas <- rep(NA,B)

for (i in 1:B){

residthetas[i] <- var(sample(a,replace=TRUE))-theta.est

}

residthetas

theta.est

# The interval

theta.est-sort(residthetas)[c(.975,.025)\*(B+1)]



B) For the data from part a, now report a 95% BCA confidence interval for the population variance.



c) Which of the intervals in parts a and b is more suitable to use for this data set, and why?

I will prefer to use BCA method because does provide result that variance is positive, meaning >0, part a will not be preferred because does include negative CI and is not accepted under variances.

1. Suppose that four laundry detergents are being compared to see if some are better than others in removing stains. Data are summarized in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Stain Not At All Removed | Stain Partially Removed | Stain Completely Removed |
| Detergent A | 4 | 2 | 2 |
| Detergent B | 5 | 1 | 2 |
| Detergent C | 6 | 2 | 0 |
| Detergent D | 2 | 1 | 5 |

Conduct the most appropriate analysis of these data at the .01 significance level. Explain what method you used to conduct the test, why you chose that method, what the p-value is, and what your conclusion is.

I chose for this problem the Jonckheere-Terpstra test because both of the variables are ordinal. I chose one-sided alternative hypothesis because it is believed that detergent led to stain removal.



data final4;

input detergent stain count;

Datalines;

1 1 4

1 2 2

1 3 2

2 1 5

2 2 1

2 3 2

3 1 6

3 2 2

3 3 0

4 1 2

4 2 1

4 3 5

;

proc freq data = final4;

weight count;

Tables detergent\*stain;

Exact jt;

run;

Conclusion: one-side p-value = 0.1587 failing to reject null hypothesis, not enough evidence to reject null hypothesis.

1. Conduct the appropriate test of the following (fictional) data set at alpha=.05, and report your conclusions. You are to investigate if there is a relationship between whether or not someone is a statistician and whether or not s/he is an introvert.

|  |  |  |
| --- | --- | --- |
|  | Introvert | Extrovert |
| Statistician | 4 | 0 |
| Non-statistician | 5 | 7 |



p-value = 0.07364. We do not have enough evidence to reject Ho.