Basic Course on R:
Hypothesis Testing and Confidence Intervals 2
Practical Answers

Elizabeth Ribble*

18-24 May 2017

Contents

*emcclel3@msudenver.edu

1 Baby Data

1. Read in the data "R_data_January2015.csv" with a header and row names from the first column. Assign it to the object `babydata` and allow strings be converted to factors. Attach the data to the environment.

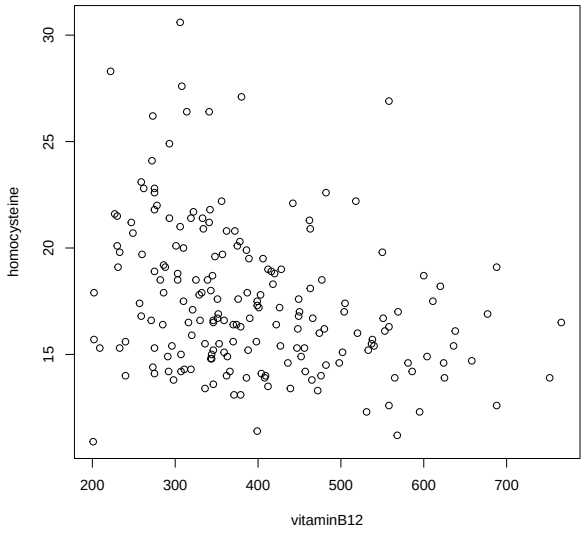
```
babydata <- read.csv("R_data_January2015.csv",header=T,row.names=1)
attach(babydata)
```

2. Answer the following questions pertaining to the variables `vitaminB12` and `homocysteine`:

   (a) What are the Pearson and Spearman correlations between `vitaminB12` and `homocysteine`? Are they similar? Formulate a hypothesis, do a test, and make a decision as to whether either the Pearson or Spearman correlation is statistically significantly. Include a scatterplot of `homocysteine` versus `vitaminB12` to support your findings.

```
cor(vitaminB12, homocysteine)

## [1] -0.3024089

cor(vitaminB12, homocysteine, method="spearman")

## [1] -0.3218201

cor.test(vitaminB12, homocysteine)

##
##  Pearson's product-moment correlation
##
## data:  vitaminB12 and homocysteine
## t = -4.3501, df = 188, p-value = 2.229e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4264059 -0.1672558
## sample estimates:
##        cor
## -0.3024089

cor.test(vitaminB12, homocysteine, method="spearman")

## Warning in cor.test.default(vitaminB12, homocysteine, method =
      "spearman"):  Cannot compute exact p-value with ties
```

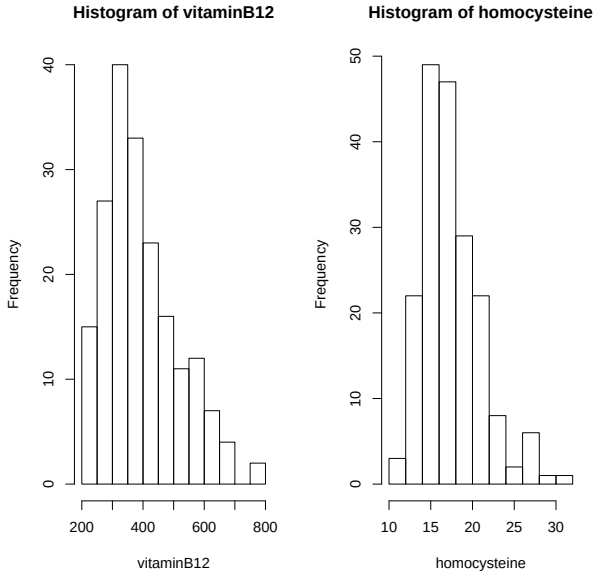
```
##
##  Spearman's rank correlation rho
##
## data:  vitaminB12 and homocysteine
## S = 1511000, p-value = 5.962e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.3218201

plot(vitaminB12,homocysteine)
```



There does appear to be a linear association between the two variables. Our null hypothesis is that there is no association between the two measures, and our alternative hypothesis is that there is an association. The tests in both cases are statistically significant (p-value ¡ 0.05 and CI doesn't contain 0) indicating there is an association between the two variables.

(b) Plot a histogram of each variable to decide whether the Pearson correlation is appropriate to use. Is it?

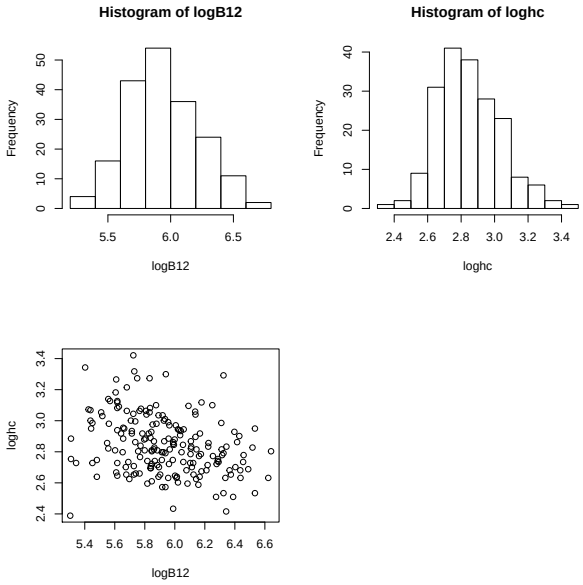
```
par(mfrow=c(1,2))
hist(vitaminB12)
hist(homocysteine)
```



The distributions are not normal but skewed, so it is more appropriate to use the Spearman correlation.

(c) Does the correlation improve after a log transformation of both variables? Make plots and do a test on the appropriate (Spearman or Pearson - depends on distribution!) correlation to answer this question.

```
logB12 <- log(vitaminB12)
loghc <- log(homocysteine)
par(mfrow=c(2,2))
hist(logB12)
hist(loghc)
plot(logB12, loghc)
cor(logB12, loghc)

## [1] -0.3031782
```



The log-transformed variables are normally distributed, and the relationship looks linear, so we can use the Pearson correlation. Our null hypothesis is that there is no association between the two log-transformed measures, and our alternative hypothesis is that there is an association. Note that the correlation didn't change much from the unlogged data.

```
cor.test(logB12, loghc)

##
##  Pearson's product-moment correlation
##
## data:  logB12 and loghc
## t = -4.3623, df = 188, p-value = 2.119e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4270986 -0.1680789
## sample estimates:
##       cor
## -0.3031782
```

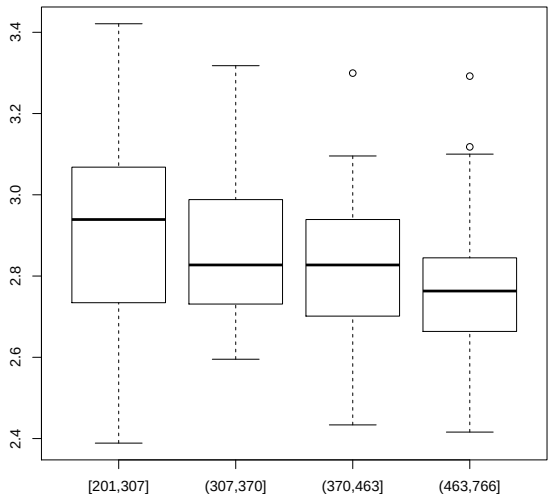
The test is statistically significant (p-value < 0.05 and CI doesn't contain 0) indicating there is an association between the two log-transformed variables.

(d) Let's see what happens when we "categorize" a continuous variable. Cut `vitaminB12` into 4 groups, where the breaks are the 5 quantile points of `vitaminB12`. Make sure you include the lowest breakpoint by specifying "`incl=TRUE`". Assign the output to `catB12`. What are the levels of this new variable?

```
catB12 <- cut(vitaminB12, breaks=quantile(vitaminB12), incl=TRUE)
levels(catB12)

## [1] "[201,307]" "(307,370]" "(370,463]" "(463,766]"
```

(e) Using the log-transformed variable from part (c), assess how the log of `homocysteine` and `catB12` relate. Make a boxplot of log-`homocysteine` for each level of `catB12`.

```
boxplot(loghc~catB12)
```



There appears to be a trend: the higher levels of `B12` correspond to lower levels of log-transformed `homocysteine`.

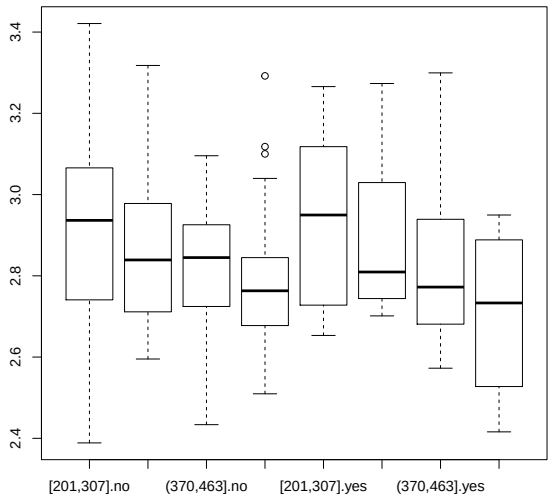
(f) Are the means of log-`homocysteine` equal across all levels of `catB12`? Formulate a hypothesis, test it, and make a decision for statistical significance.
Our null hypothesis is that the mean log-`homocysteine` level of all categories of `catB12` are equal. The alternative is that at least one of the group means differs from the other 3.

```
summary(aov(loghc~catB12))

##              Df Sum Sq Mean Sq F value   Pr(>F)
## catB12        3  0.626 0.20880   6.286 0.000438 ***
## Residuals   186  6.179 0.03322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value from an ANOVA is less than 0.05, so we conclude that there is a statistically significant difference between the mean log-homocysteine of at least one of the groups and the others. But note that this categorization of `vitaminB12` leads to a p-value that is larger than when we directly used the continuous values. This is because we are reducing the information used in the test: from 190 (nearly unique) individual data points to 190 data points that can only be 1 of 4 values/categories. It's actually recommended that you never categorize continuous variables for this exact reason!

(g) Now let's see if log-`homocysteine` varies on both `smoking` and `catB12` levels. Make a boxplot of log-`homocysteine` for all combinations of the 2 categories. Then formulate a hypothesis, test it, and make a decision for statistical significance on both categorical variables.

```
boxplot(loghc ~ catB12 + smoking)
```



Our null hypothesis is that the mean log-`homocysteine` level of all combinations of categories of `catB12` and `smoking` are equal. The alternative is that at least one of the group means differs from the others.

```
summary(aov(loghc~catB12 * smoking))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## catB12           3  0.626 0.20880   6.171 0.000512 ***
## smoking          1  0.000 0.00006   0.002 0.965727
## catB12:smoking   3  0.020 0.00675   0.199 0.896732
## Residuals      182  6.158 0.03384
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since there is no statistically significant interaction between `catB12` and `smoking` we will actually proceed after removing the interaction:

```
summary(aov(loghc~catB12 + smoking))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## catB12        3  0.626 0.20880   6.252 0.000458 ***
## smoking       1  0.000 0.00006   0.002 0.965502
## Residuals   185  6.178 0.03340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of homocysteine from a 2-way ANOVA test is less than 0.05, so we conclude that there is a statistically significant difference between the mean log-homocysteine of at least one of the `catB12` groups and the others (at any level of smoking). `smoking` is not significant, so there is no difference in the means of `homocysteine` in smokers and non-smokers (at any level of B12). You could now just use the model without `smoking` (see part (f)).

3. We suspect that people who drink alcohol (`alcohol` is yes) might also be smokers (`smoking` is yes). Formulate a hypothesis, test it using the appropriate test, and make a decision about statistical significance. [Hint: use `table`.]
The null hypothesis is that alcohol and smoking are independent; the alternative is that they are dependent.

```
tas <- table(alcohol, smoking)
chisq.test(tas)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tas
## X-squared = 7.8407, df = 1, p-value = 0.005108
```

The p-value is less than 0.05 so we conclude that there is a statistically significant relationship between smoking and drinking.