Non-parametric Statistics: Notes 5
Correlation

Gregory Matthews [1]

[1]Department of Mathematics and Statistics
Loyola University Chicago

Fall 2014

Outline

- When you think of correlation you are probably actually thinking of **Pearson correlation**.
- Pearson correlation measure the linear association between two variables.

$$\rho = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2 \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

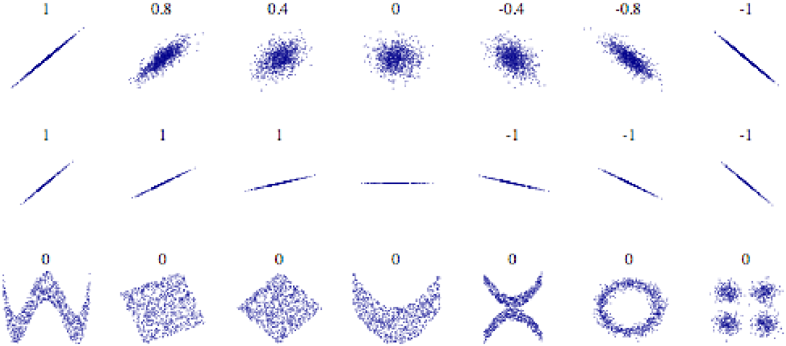


Figure : Pearson Correlation

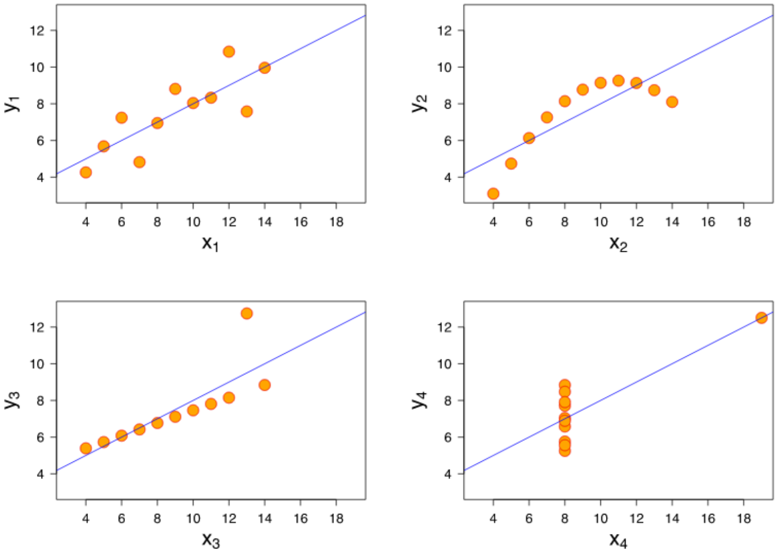
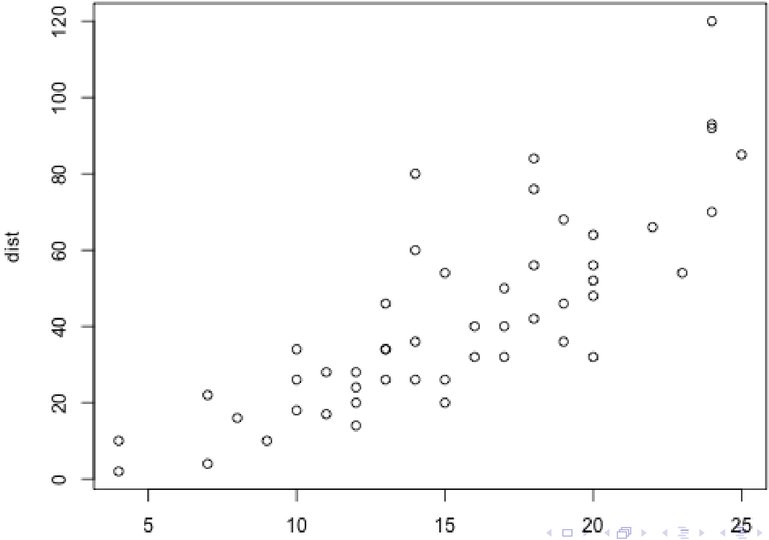


Figure : Pearson Correlation: $\rho = 0.816$ in each graph

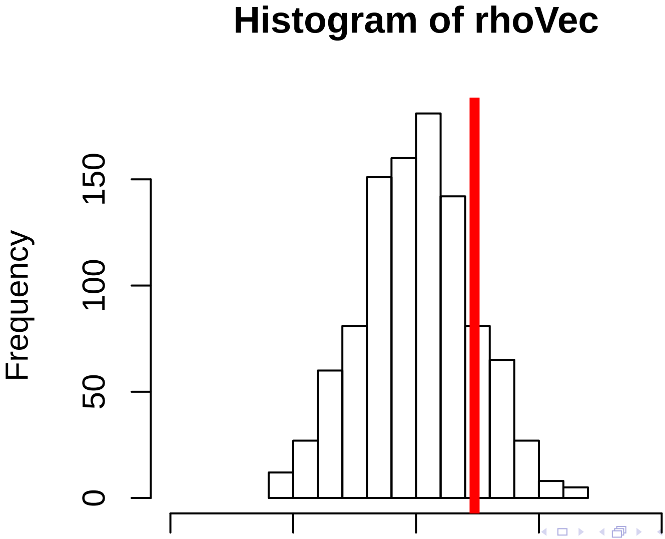
```r
plot(x,y)
```



```
#Calucalte the Pearson correlation coefficient using funct
(rho<-cor(x,y))

## [1] 0.2382

#Correlation test
#What are we testing here?
cor.test(x,y)

##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 1.041, df = 18, p-value = 0.3118
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.2283  0.6158
## sample estimates:
##    cor
## 0.2382
```

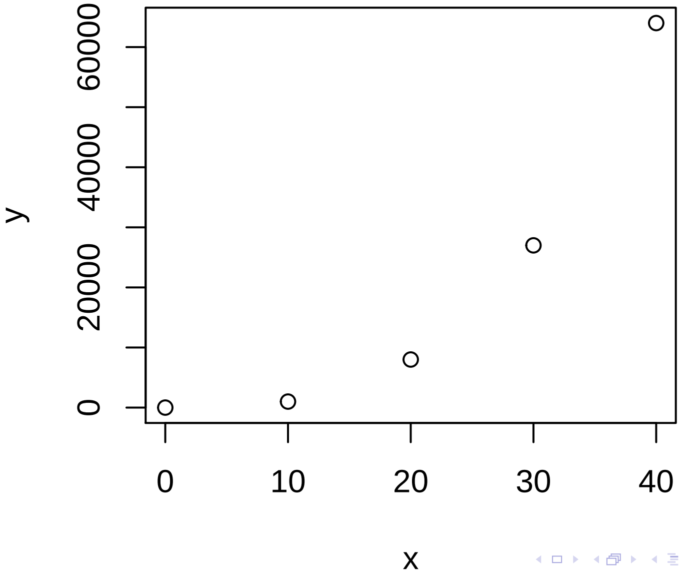
```
#permutation test!
nsim<-1000
rhoVec<-rep(NA,nsim)
for (i in 1:nsim){
yPerm<-sample(y,length(y),replace=FALSE)
rhoVec[i]<-cor(x,yPerm)
}
```

```r
hist(rhoVec,xlim=c(-1,1))
abline(v=rho,col="red",lwd=5)
```



```r
#two sided p-value
sum(rhoVec>=abs(rho))/nsim+sum(rhoVec<=-abs(rho))/nsim

## [1] 0.301

cor.test(x,y)

##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 1.041, df = 18, p-value = 0.3118
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.2283  0.6158
## sample estimates:
##    cor
## 0.2382
```

- What if there is a relationship between two variables, but the relationship is not linear?
- For instance, what is the relationship is strictly monotonic?
- Let's take a look at that situation.


0
20000
40000
60000
y
0
10
20
30
40
x

```
#If we look at the Pearson correlation we get:
cor(x,y)

## [1] 0.9059

#But in some sense these two variables are perfectly corre
#What if we computed correlation based on the ranks?
xR<-rank(x);yR<-rank(y)
cor(xR,yR)

## [1] 1

#This is Spearman correlations
cor(x,y,method="spearman")

## [1] 1
```

- Spearman correlation is Pearson correlation using ranks of the data.
- Spearman correlation measure the strength of the monotonic relationship between two variables.

$$\rho = \frac{\Sigma_{i=1}^{n}(R_i^X - \bar{R}^{X]})(R_i^Y - \bar{R}^Y)}{\sqrt{\Sigma_{i=1}^{n}(R_i^X - \bar{R}^X)^2 \Sigma_{i=1}^{n}(R_i^Y - \bar{R}^Y)^2}} = \frac{cov(R^X, R^Y)}{\sigma_{R^X}\sigma_{R^Y}}$$

```
#Test the null hypothesis that the Spearman correlation is

#In other words there is no association between X and Y
cor.test(x,y,method="spearman")

##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 0, p-value = 0.01667
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
##   1

#This p-value of this test is equal to
2*(1/120)

## [1] 0.01667
```

```r
library(gtools)
perms<-permutations(5,5)
corNull<-rep(NA,dim(perms)[1])
for (i in 1:dim(perms)[1]){
  corNull[i]<-cor(x,y[perms[i,]],method="spearman")
}
sort(corNull)

##   [1] -1.0 -0.9 -0.9 -0.9 -0.9 -0.8 -0.8 -0.8 -0.7 -0.7
##  [15] -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.5 -0.5 -0.5
##  [29] -0.4 -0.4 -0.4 -0.3 -0.3 -0.3 -0.3 -0.3 -0.3 -0.3
##  [43] -0.2 -0.2 -0.2 -0.2 -0.2 -0.1 -0.1 -0.1 -0.1 -0.1
##  [57] -0.1  0.0  0.0  0.0  0.0  0.0  0.0  0.1  0.1  0.1
##  [71]  0.1  0.1  0.1  0.2  0.2  0.2  0.2  0.2  0.2  0.3
##  [85]  0.3  0.3  0.3  0.3  0.3  0.4  0.4  0.4  0.4  0.5
##  [99]  0.5  0.6  0.6  0.6  0.6  0.6  0.6  0.6  0.7  0.7
## [113]  0.8  0.8  0.8  0.9  0.9  0.9  0.9  1.0
```

Two quick notes

- When there are ties, simply take the average of the ranks and proceed as usual.
- Careful using Spearman correlation (or any correlation). If observations are correlated, Spearman may result in incorrect results. **Only use when observations are uncorrelated.**

Kendall's τ

- First we need to talk about **concordance** and **discordance**.
- Two pairs of data (X_i, Y_i) and (X_j, Y_j) are condorance if $X_i < X_j$ implies that $Y_i < Y_j$.
- Two pairs of data (X_i, Y_i) and (X_j, Y_j) are condorance if $X_i < X_j$ implies that $Y_i > Y_j$.
- Concordance means that the changes in the X's and Y's are in the same direction.

Kendall's τ

- If pairs of observations are more likely to be concordant, then we say there is a positive relationship.
- If pairs of observations are more likely to be disconcordant, then we say there is a negative relationship.
- Based on the definition of concordance, we can also express concordance as $(X_i - X_j)(Y_i - Y_j) > 0$. (WHY?)
- And we can write $\tau = 2P[(X_i - X_j)(Y_i - Y_j) > 0] - 1$
- $P[(X_i - X_j)(Y_i - Y_j) > 0]$ is the probability of concordance. Then we rescale this number to range from -1 to 1 like the other measures of correlation.

Kendall's τ

- How do we estimate Kendall's τ (assuming no ties.)

Define:

$$U_{ij} = \begin{cases} 1 & : (X_i - X_j)(Y_i - Y_j) > 0 \\ 0 & : (X_i - X_j)(Y_i - Y_j) < 0 \end{cases}$$

Then let:

$$V_i = \Sigma_{j=i+1}^{n} U_{ij}$$

V_i is the number of pairs that are concordant with the i-th pair for $j \geq i + 1$ Then:

$$r_\tau = 2\frac{\Sigma_{i=1}^{n-1} V_i}{\binom{n}{2}} - 1$$

```
#let's compute Kendall's Tau by hand (ONCE!)
(dat<-data.frame(x=c(1,2,3,4),y=c(2,4,3,1)))

##   x y
## 1 1 2
## 2 2 4
## 3 3 3
## 4 4 1

#Concordant pairs (1,2),(1,3)
#Disconcordant pairs (1,4),(2,3),(2,4),(3,4)
#2*(numConcordPairs/TotalNumberOfPairs)-1
2*(2/6)-1

## [1] -0.3333

#Or just use the R function
cor(dat$x,dat$y,method="kendall")

## [1] -0.3333
```

```
#Null hypothesis that there is no association
#Versus alternative that there is some association
cor.test(dat$x,dat$y,method="kendall")

##
##  Kendall's rank correlation tau
##
## data:  dat$x and dat$y
## T = 2, p-value = 0.75
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##     tau
## -0.3333
```

Contigency Tables

- When we compare two categorical variables to each other, we often us two way contingency tables.
- We say there are R levels of the first variable and C levels of the second variable, then we can create an R by C table.
- The elements of the contingency table are counts of the individuals who fall into both categories corresponding to a cell in the table.

```
#A contigency table
dat<-data.frame(x=c(1,1,1,2,2,2,3,3,3,4,4,4),y=c(1,3,2,4,2,
table(dat)

##    y
## x   1 2 3 4
##   1 1 1 1 0
##   2 0 1 0 2
##   3 1 0 1 1
##   4 1 1 1 0
```

Contigency Tables

- What hypothesis are we interested in testing here? It depends!
- Consider two cases:
  1. All n individuals are selected at random and cross classified according to row and column characteristic.
  2. A fixed numnber $n_{i\cdot}$ for $i = 1, 2, \ldots, r$ are sampled and classified according to column characteristic.

- First define: $p_{ij} = \frac{E[n_{ij}]}{n}$
- Case 1: $H_0 : p_{ij} = p_{i.}p_{.j}$ vs $H_a : p_{ij} \neq p_{i.}p_{.j}$
- Case 2: $H_0 : p_{j|i} = p_{j|i'}$ vs $H_a : p_{j|i} \neq p_{j|i'}$ where $p_{j|i} = \frac{p_{ij}}{p_{i.}}$
- The null hypotheses in both Case 1 and Case 2 are equivalent EVEN THOUGH the sampling situations are different.

- How do we test these hypotheses?
- Use the following test statistic. (Why is this a good test statistic?)

Test Statistic:

$$\chi^2 = \Sigma_{i=1}^{r} \Sigma_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

where $e_{ij} = \frac{n_{i.} n_{.j}}{n}$

- If all $e_{ij} > 5$, the $\chi^2 \sim \chi^2_{df=(r-1)(c-1)}$
- However, when expected cell frequencies are small, the approximation may not be appropriate.
- This is especially true when many cells have no responses.
- What can we do in that situation? Permutation test!

```
#Table 5.4.2 in the book
dat<-data.frame(who=factor(c(rep("Physician-Prescribed",4)
table(dat)

##                           satisfaction
## who                         Not Somewhat Very
##   Physician-Prescribed        2        2    0
##   Self-Administered           0        1    2
```

```
(XsqObs<-chisq.test(table(dat))$statistic)

## Warning:  Chi-squared approximation may be
incorrect

## X-squared
##     4.278
```

```r
library(gtools)
perms<-combinations(7,4)
XsqPerms<-rep(NA,35)
for (i in 1:dim(perms)[1]){
  datTemp<-dat
  ind<-c(perms[i,],setdiff(c(1:7),perms[i,]) )
  datTemp$satisfaction<-datTemp$satisfaction[ind]
  XsqPerms[i]<-chisq.test(table(datTemp))$statistic
}
```

```
## Warning:  Chi-squared approximation may be
incorrect
## Warning:  Chi-squared approximation may be
incorrect
## Warning:  Chi-squared approximation may be
incorrect
## Warning:  Chi-squared approximation may be
incorrect
## Warning:  Chi-squared approximation may be
```

```
XsqObs

## X-squared
##     4.278

table(round(XsqPerms,2))

##
## 0.19 2.24 4.28 4.96    7
##   12   12    6    4    1

sum(XsqPerms>=XsqObs)/35

## [1] 0.3143
```

- Real quick. The special case of the permutation test for a 2 by 2 contingency table is called the Fisher's exact test.
- This is implemented in R using the function fisher.test
- Example follows:

```
## Agresti (1990, p. 61f; 2002, p. 91) Fisher's Tea Drinker
## A British woman claimed to be able to distinguish whether milk or
##  tea was added to the cup first.  To test, she was given 8 cups of
##  tea, in four of which milk was added first.  The null hypothesis
##  is that there is no association between the true order of pouring
##  and the woman's guess, the alternative that there is a positive
##  association (that the odds ratio is greater than 1).
```

```
TeaTasting <-
matrix(c(3, 1, 1, 3),
       nrow = 2,
       dimnames = list(Guess = c("Milk", "Tea"),
                       Truth = c("Milk", "Tea")))
TeaTasting

##        Truth
## Guess  Milk Tea
##   Milk    3   1
##   Tea     1   3
```

```
#Why do we want one-sided here?
fisher.test(TeaTasting, alternative = "greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  TeaTasting
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than
## 95 percent confidence interval:
##  0.3136    Inf
## sample estimates:
## odds ratio
##      6.408

## => p = 0.2429, association could not be established
```

McNemar’s test

- McNemar’s test is used for paired comparison studies when the responses are dichotomous.
- The classic example of this is to ask people which canidate they prefer before a political debate, and then ask them again after the debate.
- If there are two candidates (A and B) there are four possibilities: (A,A) (A,B) (B,A) (B,B).
- We are interested in testing the null hypothesis that the probability of A before is the same as the probability of A after.
- $H_0 : P_{AA} + P_{AB} = P_{AA} + P_{BA} = P_{AB} = P_{BA}$

McNemar's test

- If we let $n = X_{AB} + X_{BA}$ and $P_{AB} = P_{BA}$, then there is an equal chance of switching from A to B or B to A.
- In this way the distribution of $X_{AB}|n \sim$ *binomial* with mean $0.5n$ and variance $0.25n$. WHY?
- We can then use this to calculate a p-value.

```r
## Agresti (1990), p. 350.
## Presidential Approval Ratings.
##  Approval of the President's performance in office in two surveys,
##  one month apart, for a random sample of 1600 voting-age Americans.
Performance <-
matrix(c(794, 86, 150, 570),
       nrow = 2,
       dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                       "2nd Survey" = c("Approve", "Disapprove")))
Performance
```

```
##             2nd Survey
## 1st Survey   Approve Disapprove
##   Approve        794        150
##   Disapprove      86        570
```

```
(n<-Performance[1,2]+Performance[2,1])

## [1] 236

#calculate the p-value
2*(1-pbinom(Performance[1,2]-1,n,.5))

## [1] 3.716e-05

2*(pbinom(Performance[2,1],n,.5))

## [1] 3.716e-05
```

```r
#approximate the p-value
T4<-((Performance[1,2]-0.5*n)/sqrt(0.25*n))^2
1-pchisq(T4,1)
```

```
## [1] 3.099e-05
```

```r
mcnemar.test(Performance,correct=FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  Performance
## McNemar's chi-squared = 17.36, df = 1, p-value = 3.099e-05
```

```
mcnemar.test(Performance,correct=TRUE)

##
##  McNemar's Chi-squared test with continuity correction
##
## data:  Performance
## McNemar's chi-squared = 16.82, df = 1, p-value = 4.115e-05

## => significant change (in fact, drop) in approval ratings
```