BMI 713 / GEN 212

Lecture 7: Correlations

- Pearson correlation coefficient
- Spearman correlation coefficient
- Parametric vs nonparametric

October 21, 2010

Correlation Analysis

- Previously we focused on measures of the strength of association between two dichotomous random variables
- We can also look at the relationship between two continuous variables
- One technique often used to measure association is called correlation analysis
- Correlation is defined as the quantification of the degree to which two continuous variables are related**, provided that the relationship is linear**

Pearson Correlation Coefficient

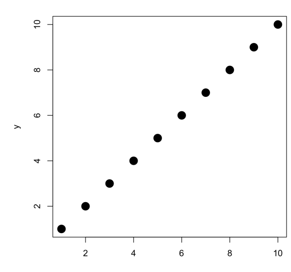
- Denote the true underlying population correlation between *X* and *Y* by ρ (rho)
- The correlation ρ quantifies the strength of the linear relationship between *X* and *Y*
- The population correlation can be estimated from a sample of data using the **Pearson correlation coefficient** *r* (the "product-moment" correlation coefficient)
- The correlation coefficient is calculated as

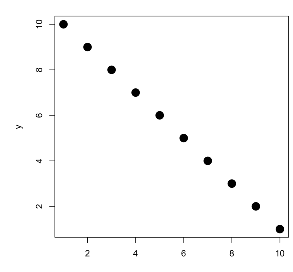
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}}$$

Pearson Correlation Coefficient

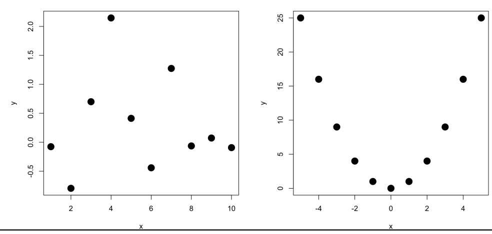
- The correlation coefficient has no units of measurement
- It can assume values from −1 to +1
- The values $r = 1$ and $r = -1$ imply a perfect linear relationship between the variables — the points (x_i, y_i) all lie on a straight line





Linearity

- If $r > 0$ the two variables are said to be positively correlated; if $r < 0$ they are negatively correlated
- If $r = 0$ there is no linear relationship at all
- r does not depend on the units of measurement
- A nonlinear relationship may exist
- Therefore, if two variables are uncorrelated, that **does not mean they're independent!**



Hypothesis Testing

- We can make inference about the unknown population correlation ρ using the sample correlation coefficient r
- Most often we want to test whether X and Y are linearly associated, i.e.,

$$H_0 : \rho = \rho 0 = 0$$

- We need to find the probability of obtaining a sample correlation as extreme or more extreme than r given that H_0 is true
- The estimated standard error of r is

$$\widehat{\text{se}}(r) = \sqrt{\frac{1-r^2}{n-2}},$$

Hypothesis Testing

- we use the test statistic

$$\text{t} = \frac{r-0}{\sqrt{(1-r^2)/(n-2)}} = r\sqrt{\frac{n-2}{1-r^2}}$$

- If X and Y are both normally distributed, the statistic has a t distribution with $n–2$ df
- This procedure is valid for testing $\rho = 0$ only
- Another method: Fisher's z-transformation

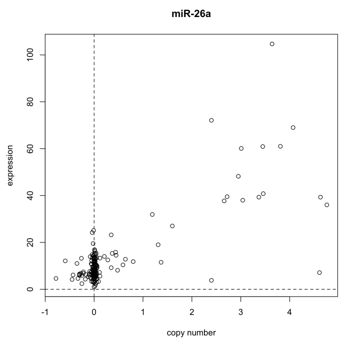
$$Z_r = \frac{1}{2}\ln\left(\frac{r+1}{r-1}\right)$$

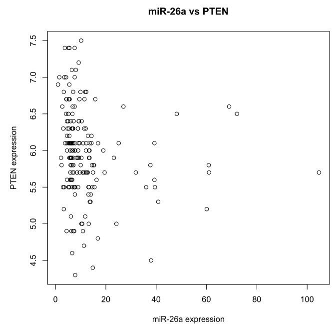
- This follows a normal distr with standard error 1/sqrt(N-3)
- Also the cosine of the angle between two vectors after centering

Example

- Correlation between miR-26a copy number and expression in glioblastoma
- Huse et al, The PTEN-regulating microRNA miR-26a is amplified in high-grade glioma and facilitates gliomagenesis in vivo, *G&D*, 2009

```
X = as.matrix(read.table("GBM_miR26a.txt",header=T,row.names=1))
> X[1:5,]
  miR26a_ACGH miR26a_EXPR PTEN_ACGH PTEN_EXPR
1        0.28        12.5      0.00       5.7
2        0.12        13.3     -0.15       6.0
3        3.81        61.0     -0.83       5.7
6        3.01        60.1     -1.28       5.2
7        0.02        12.1     -0.80       6.0
```






```
> cor(X[,1],X[,2])
[1] 0.7815552
> round(cor(X[,1],X[,2]),3)
[1] 0.782
```

```
> cor(X[,2],X[,4])
[1] -0.1445774
> round(cor(X[,2],X[,4]),3)
[1] -0.145
```

Correlation Coeffficients in R

```
> cor(X[,1],X[,2])
[1] 0.7815552
> cor.test(X[,1],X[,2])

        Pearson's product-moment correlation

data:  X[, 1] and X[, 2]
t = 16.855, df = 181, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7178850 0.8322588
sample estimates:
      cor
0.7815552
```

- What is the proper way to write this p-value?

Limitations of the correlation coefficient

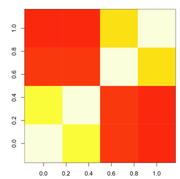
- It quantifies only the strength of the linear relationship between two variables
- **It is very sensitive to outlying values, and thus can sometimes be misleading**
- It cannot be extrapolated beyond the observed ranges of the variables
- **A high correlation does not imply a cause-and-effect relationship**

Correlation Matrix

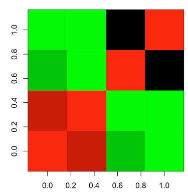
- When analyzing multiple variables, it is common to display all pairwise sample correlations at once in a **correlation matrix**

```
> cor(X)
            miR26a_ACGH miR26a_EXPR   PTEN_ACGH   PTEN_EXPR
miR26a_ACGH  1.00000000  0.78155519 -0.02290024 -0.08262843
miR26a_EXPR  0.78155519  1.00000000 -0.04655677 -0.14457737
PTEN_ACGH   -0.02290024 -0.04655677  1.00000000  0.53645667
PTEN_EXPR   -0.08262843 -0.14457737  0.53645667  1.00000000
```

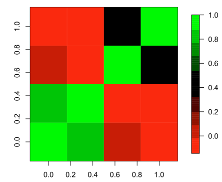
```
library(gplots)
library(fields)
image(cor(X))
```



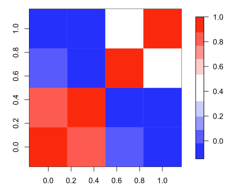
```
image(cor(X),col=greenred(10))
```



```
image.plot(cor(X),col=redgreen(10))
```



```
image.plot(cor(X),col=colorpanel(10,low
  ="blue",mid="white",high="red"))
```



Spearman Correlation Coefficient

- If the variables are not normally distributed or if there are any outliers in the data, then Spearman's rank correlation coefficient is a more robust measure of association
- It is a **nonparametric** technique
- Spearman's rank correlation coefficient is denoted by r_s and is simply Pearson's r calculated for the ranked values of x and y
- Therefore

$$r_s = \frac{\sum_{i=1}^{n}(x_{ri}-\bar{x}_r)(y_{ri}-\bar{y}_r)}{\sqrt{[\sum_{i=1}^{n}(x_{ri}-\bar{x}_r)^2][\sum_{i=1}^{n}(y_{ri}-\bar{y}_r)^2]}}$$

where x_{ri} and y_{ri} are the ranks associated with the ith subject rather than the actual observations

Spearman Correlation Coefficient

- Spearman's rank correlation may also be thought of as a measure of the concordance of the ranks for the outcomes x and y
- The Spearman rank correlation takes on values between −1 and +1; values close to the extremes indicate a high degree of correlation
- If all measurements are ranked in the same order for each variable, then $r_s = 1$
- If the ranking of the first variable is the inverse of the ranking of the second, $r_s = -1$

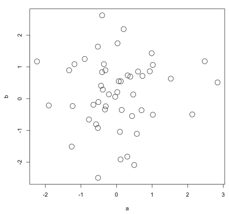
Hypothesis Testing

- The rank correlation coefficient can also be used to test the null hypothesis $H_0 : \rho = 0$
- If the sample size is not too small ($n \geq 10$), we use the same procedure that we used for Pearson's r

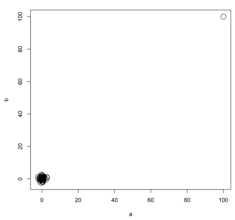
- Like other nonparametric techniques, **Spearman's rank correlation is less sensitive to outlying values and the assumption of normality than the Pearson correlation**
- In addition, the rank correlation can be used when one or both of the variables are ordinal

Effect of Outliers

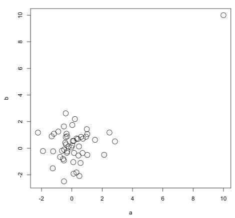
```
a=rnorm(50)
b=rnorm(50)
plot(a,b,cex=2)
cor(a,b)
```



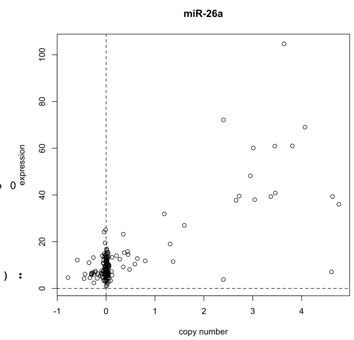
```
a[51]=100
b[51]=100
plot(a,b,cex=2)
cor(a,b)
cor(a,b,m="sp")
```



```
a[51]=10
b[51]=10
plot(a,b,cex=2)
cor(a,b,method=
"spearman")
```



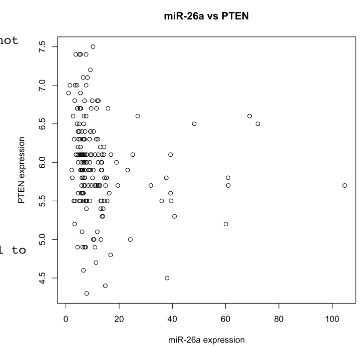
```
> cor.test(X[,1],X[,2])

        Pearson's product-moment correlation

data:  X[, 1] and X[, 2]
t = 16.855, df = 181, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 0.7178850 0.8322588
sample estimates:
      cor
0.7815552

> cor.test(X[,1],X[,2],method="spearman")

        Spearman's rank correlation rho

data:  X[, 1] and X[, 2]
S = 633403.7, p-value = 1.136e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3798575

Warning message:
In cor.test.default(X[, 1], X[, 2], method = "s") :
  Cannot compute exact p-values with ties
```



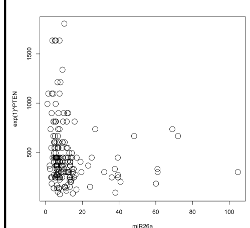
What do you get with the following?

```
cor.test(rank(X[,1]),rank(X[,2]))
```

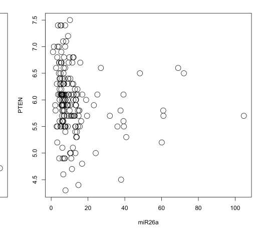
```
> cor.test(X[,2],X[,4])

        Pearson's product-moment correlation

data:  X[, 2] and X[, 4]
t = -1.9657, df = 181, p-value = 0.05086
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 -0.2836846104  0.0004895475
sample estimates:
       cor
-0.1445774

> cor.test(X[,2],X[,4],method="sp")

        Spearman's rank correlation rho

data:  X[, 2] and X[, 4]
S = 1321283, p-value = 5.482e-05
alternative hypothesis: true rho is not equal to
0
sample estimates:
       rho
-0.2936198

Warning message:
In cor.test.default(X[, 2], X[, 4], method =
"sp") :
  Cannot compute exact p-values with ties
```



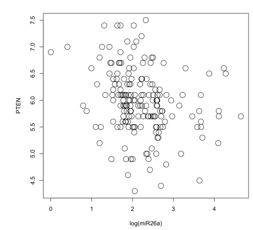
Taking logarithm



Original data:
(both unlogged)
PCC = -.149



Data as used in the paper
(take log of PTEN exp)
PCC = -.145



Take log of both
PCC = -.255