Topic 8

Correlation analysis

Sergey Mastitsky ©
Klaipeda, 28-30 September 2011

Correlation

- Correlation is the strength of association between two variables
- A correlation coefficient is calculated to quantitatively assess this association
- $-1 \leq r \leq +1$
- (!) Be aware of the incorrect use of correlation coefficients

Author: Sergey Mastitsky

8. Correlation analysis

8.1. Pearson correlation

Pearson correlation

- Assumes that both variables are normally distributed
- Assumes that the relationship between the variables is linear
- Pearson *r* is based on the calculation of covariance between two variables:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

An example: effect of pH on Dreissena

```
> setwd("~/Introductory R
   Course/R_Course_Datasets")
```

- In **RStudio:**

Workspace -> Load Workspace… -> …

`pH_experiment.rda`

Calculating Pearson correlation in R

```
> logL <- log(LWdata$Length)
> logW <- log(LWdata$Weight)
> cor(logL, logW)
[1] 0.9807
# If there were missing data, the command would've
  looked as:
> cor(logL, logW,
      use = "complete.obs")
# To calculate r between all variables in a data frame,
  just provide its name as argument for cor()
```

Testing the significance of *r*, i.e. whether it is significantly different from 0

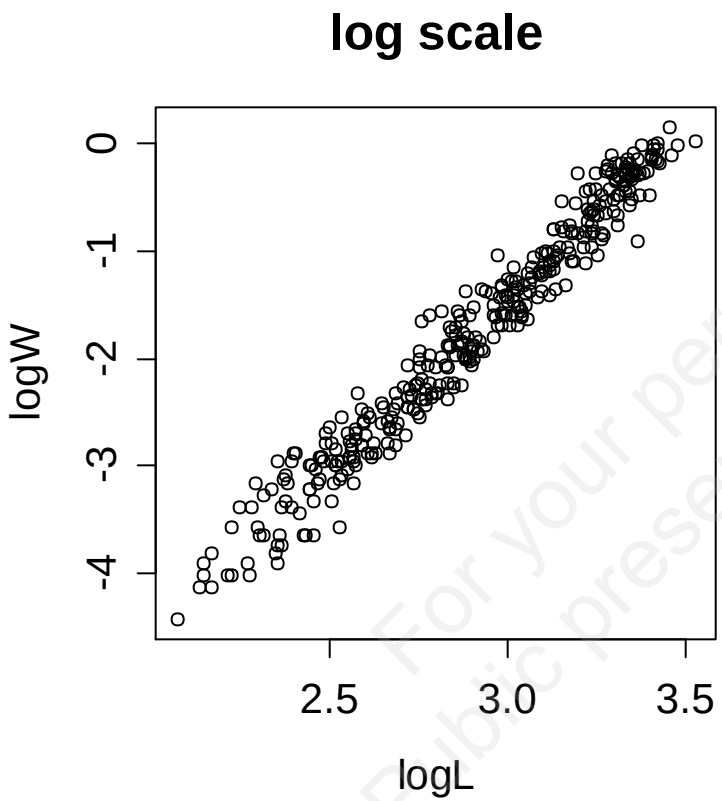
It's possible to transform *r* into a t-distributed variable, and then check its significance

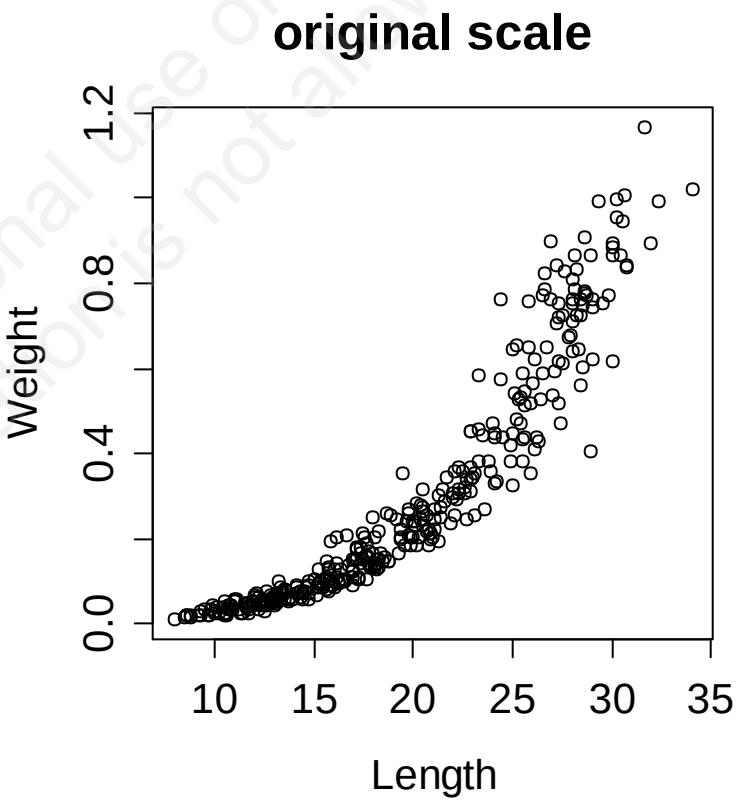
```
> cor.test(logL, logW)

 Pearson's product-moment correlation

data:  logL and logW
t = 95.0486, df = 359, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9763151 0.9842866
sample estimates:
      cor
0.9807043
```

8. Correlation analysis

8.2. Spearman correlation (ρ)

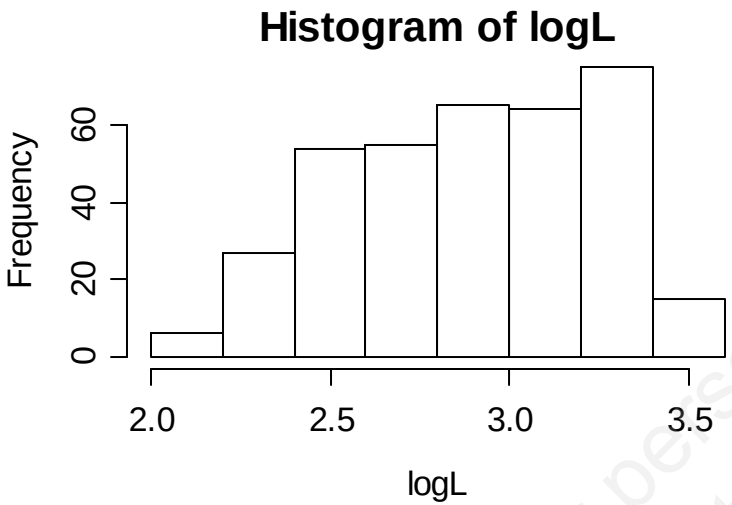
On the original scale, the relationship between L and W is not linear

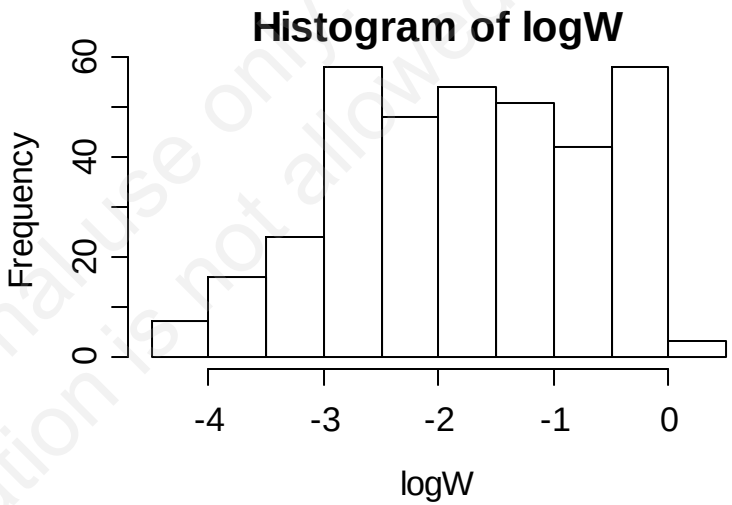


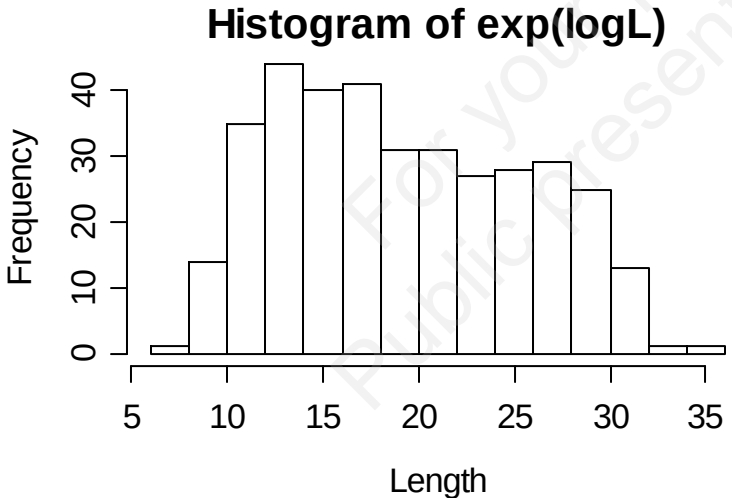


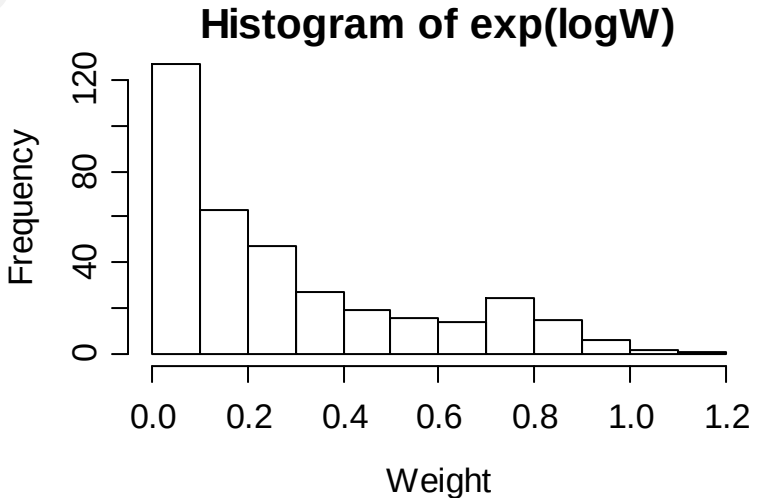
Author: Sergey Mastitsky

Data are not normally distributed at neither of the scales









Spearman correlation

- Assumptions-free measure of association
- Instead of the original values, makes use of their ranks to calculate the correlation (same formula though)
- (!) Interpretation is not as straightforward as with the Pearson correlation (relationships can be nonlinear)

Calculating Spearman correlation in R

```
> cor(logL, logW,
      method = "spearman")
[1] 0.98196
```

Testing the significance of ρ in R

```
> cor.test(logL, logW,
           method = "spearman")

 Spearman's rank correlation rho

data:  logL and logW
S = 141415.2, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9819645

Warning message:
In cor.test.default(logL, logW, method = "spearman") :
  Cannot compute exact p-values with ties
```

8. Correlation analysis

8.3. Kendall's τ

Kendall's τ

- Suppose we have a set of joint observation for two variables, i.e.
  $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$
- Any pair of observations (x_i, y_i) and (x_j, y_j) is said to be *concordant* if the ranks for both elements agree, i.e.

if both $x_i > x_j$ and $y_i > y_j$
or if both $x_i < x_j$ and $y_i < y_j$

$$\tau = \frac{(n_{conc.pairs}) - (n_{dicord.pairs})}{0.5n(n-1)}$$

Calculating Kendall's τ in R

```
> cor(logL, logW,
      method = "kendall")
[1] 0.8827
```

Testing the significance of Kendall correlation

```
> cor.test(logL, logW,
           method = "kendall")
```

```
 Kendall's rank correlation tau

data:  logL and logW
z = 24.9778, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8827185
```