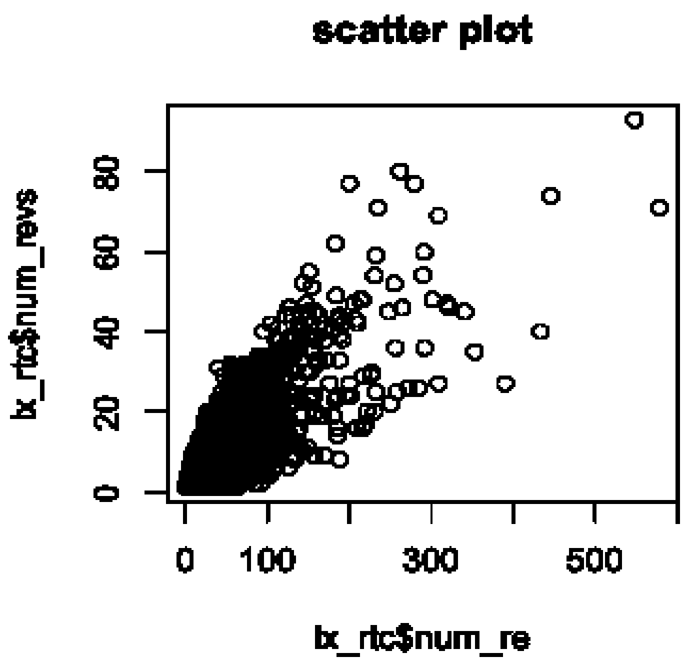
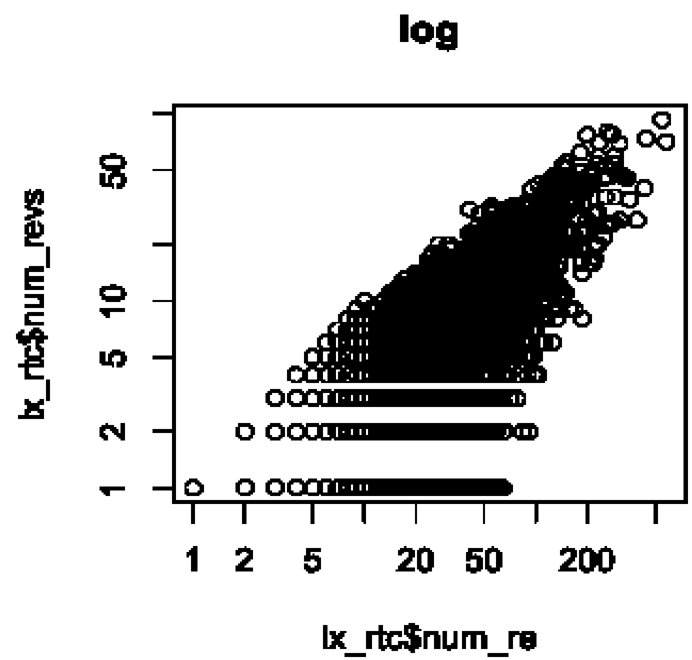
Practical Statistics

Peter C Rigby

Step 1: Take a look at the data

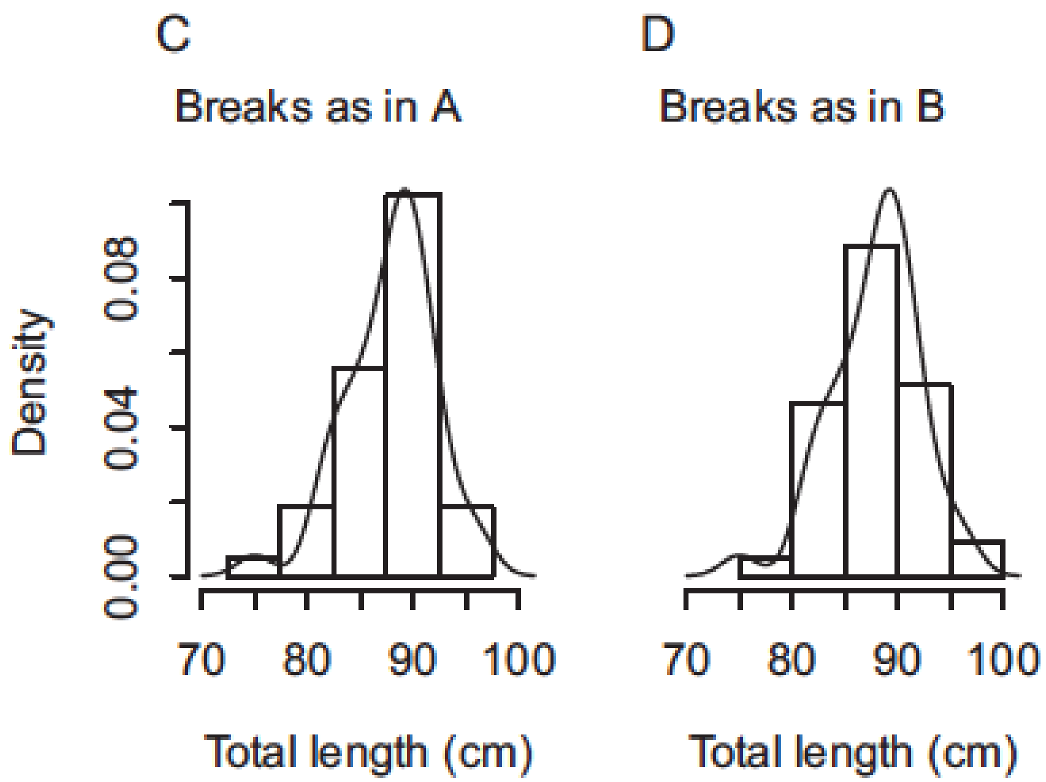
Scatter Plots





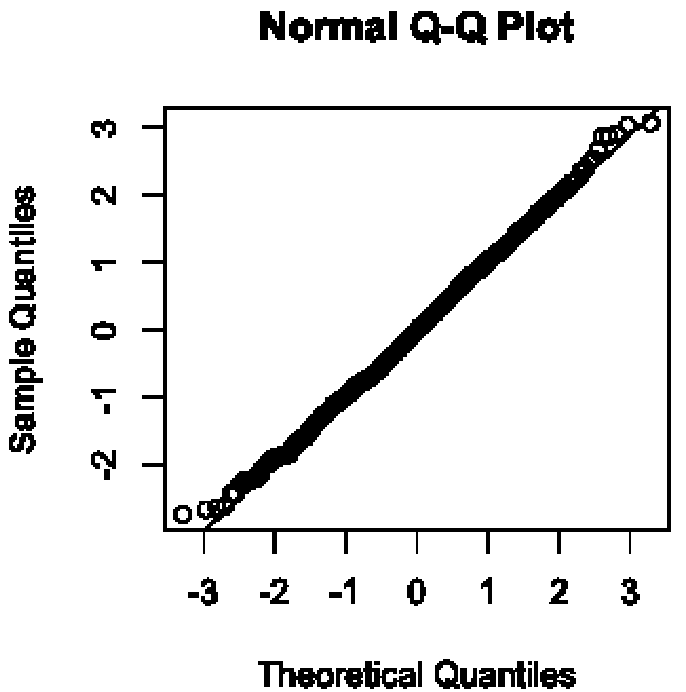
Scatter Plot is visual correlation “test”
- plot(num_revs, num_re)

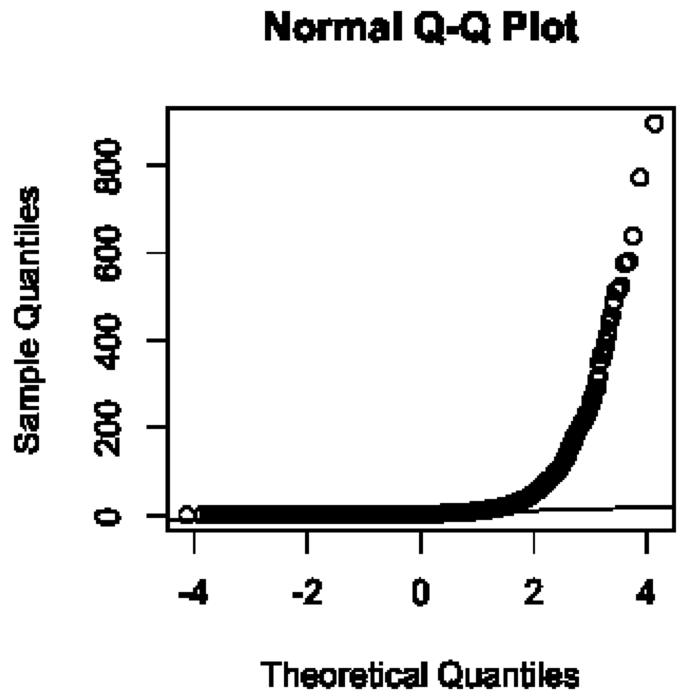
Histograms



Histograms can distort the data

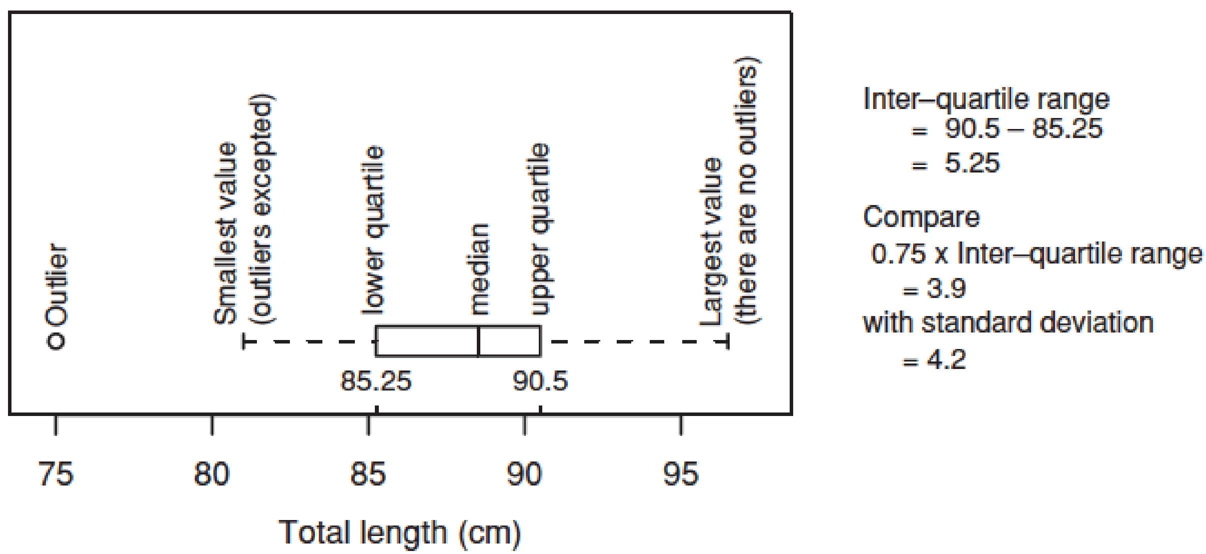
QQ Plot





Is the data normally distributed?
qqnorm(bugs)

Boxplot



Visual version of the r summary function
-boxplot(bugs, log = 'y')

Step 2: Model the data

- Things to consider
  - Are your variables count, categorical, continuous?
- Do you want to compare two or more conditions?

Linear Model

- y value, response, or dependent
- x value, predictor, or independent
- ε, error, or residuals
- b0 intercept
- b1 slope

$$y = b_0 + b_1 x + \epsilon$$

```
m <- lm(rev_interval ~ num_re + num_revs)
```

Linear Model

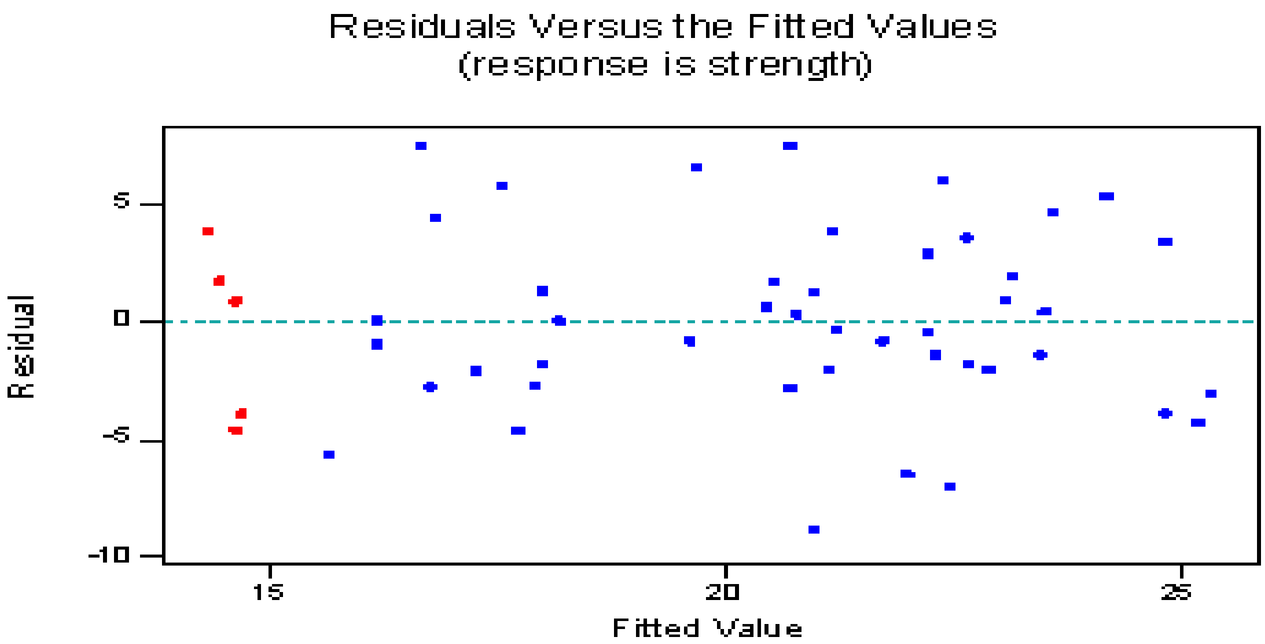
$$\hat{y}_i = b_0 + b_1 x_i$$

Model Assumptions

- Independence of observations
- Normality – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called homoscedasticity — the variance of data in groups should be the same.

Constant Variance the Error

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$



Coefficient of determination R^2

Mean squared error $$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}$$

Total sum of squares $$\mathrm{SSTO} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Proportion of the variation in y explained by x

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = \frac{\mathrm{SSTO} - \mathrm{SSE}}{\mathrm{SSTO}}$$

Generalized Linear Models

- Continuous
  - gaussian(link = “indentity”)
- Categorical
  - binomial(link = "logit")
- Count data
  - poisson(link = "log")
  - glm(bugs ~ num_re, family = 'quasipoisson')

Dispersion

- Overdispersion makes variables look more statistically significant.
  - Underdispersion has opposite effect
- Quasi function correct for dispersion
  - quasibinomial(link = "logit")
  - quasipoisson(link = "log")

Interpreting predictors

- You must apply inverse of link function to estimates interpret estimates
  - Poisson link function is log
  - Take the exp of each estimate

ANOVA

- Comparing two groups of data
  - bugs ~ as.factor(module)
  - bugs ~ as.factor(devs)
  - Bugs ~ as.factor(organization)
- The null hypothesis is that all groups are simply random samples of the same population.
- Example
  - http://www.youtube.com/watch?v=Dwd3ha0P8uw