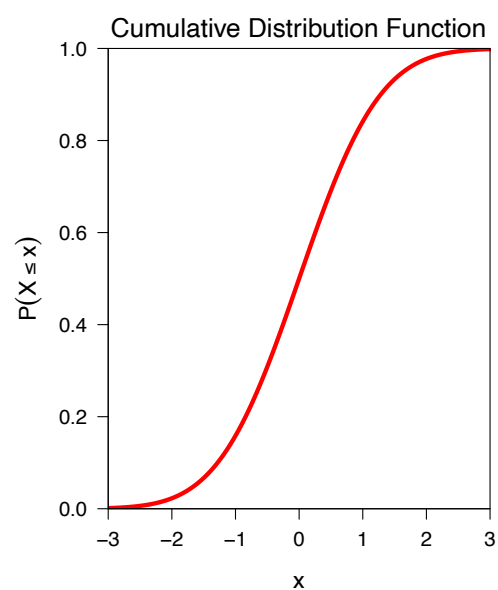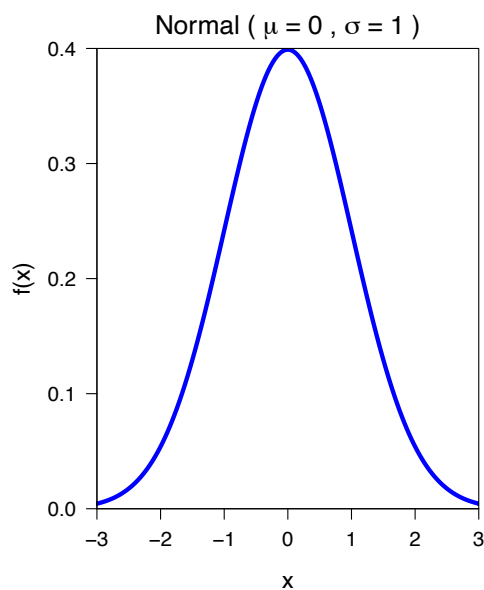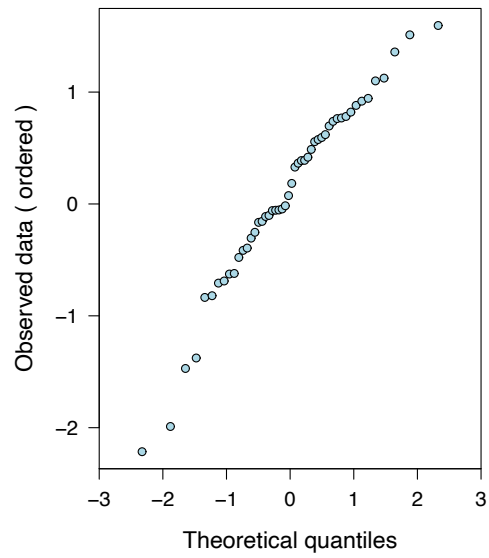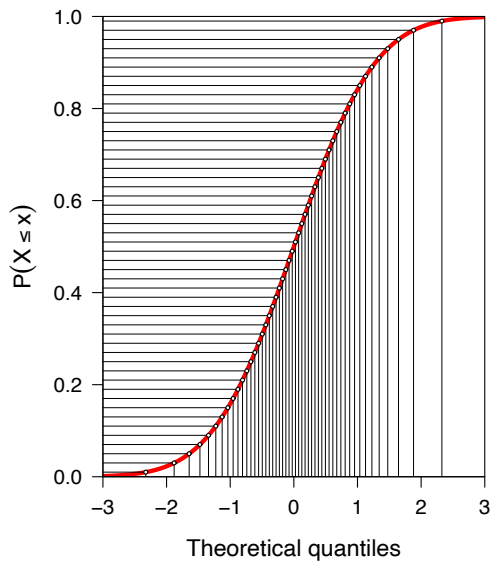# Model Assumptions and Diagnostics

QQ-plots

Normal ($\mu = 0$, $\sigma = 1$)

Cumulative Distribution Function

QQ-plots

QQ-plots

X ~ Normal (mean=0, sd=1)

X ~ Normal (mean=5, sd=2)

Normal Q–Q Plot

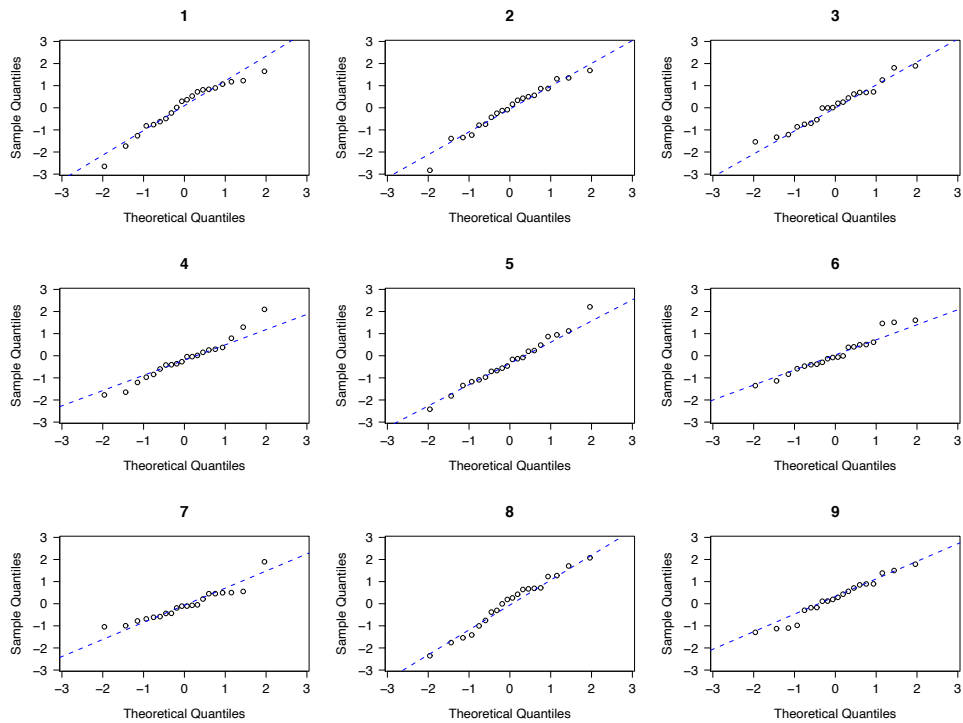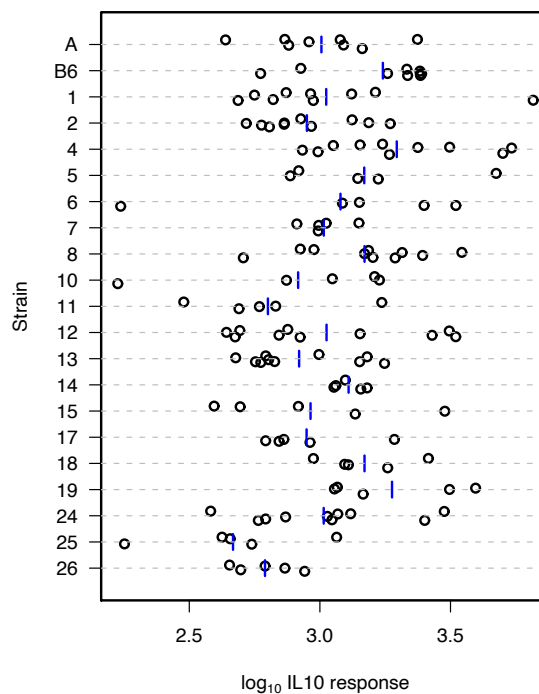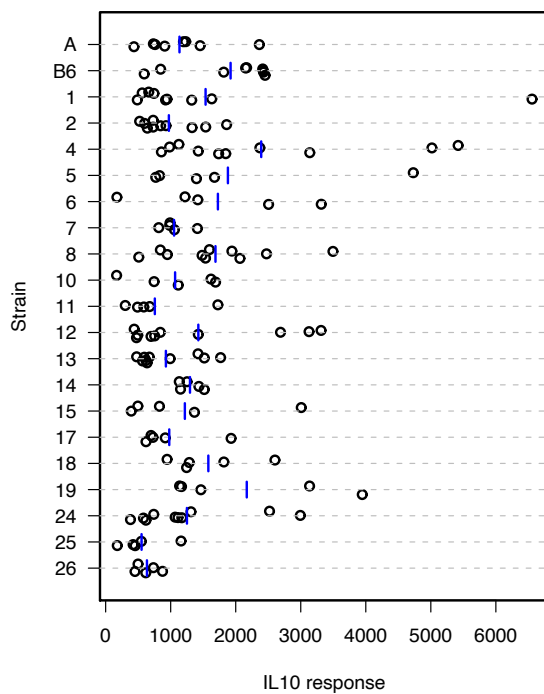**Normal Q–Q Plot**
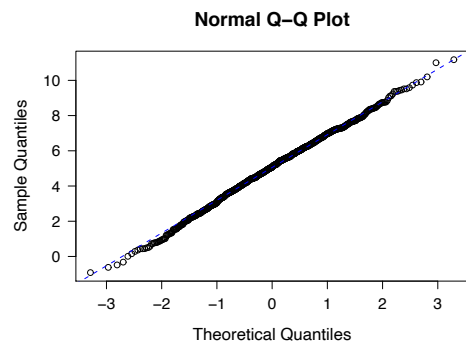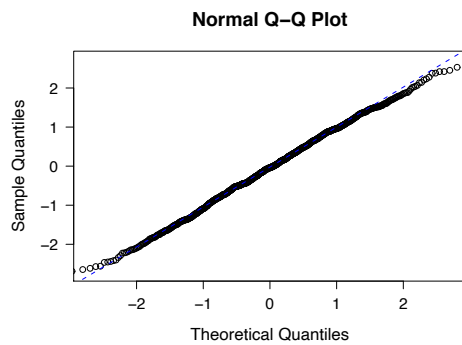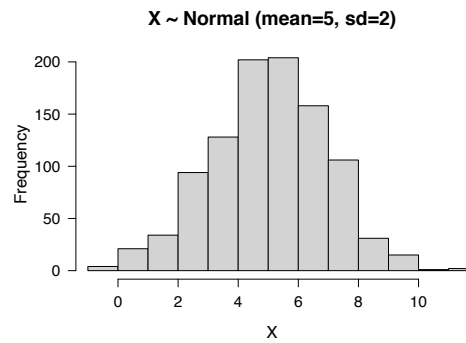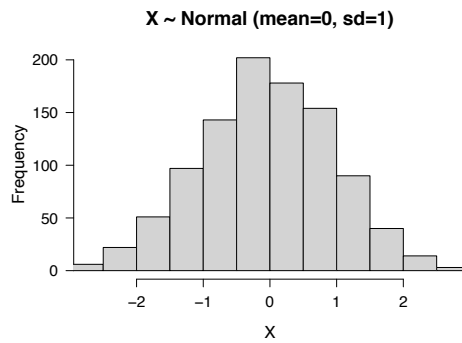
QQ-plots

Diagnostics

- QQ plot within each group

- QQ plot of all residuals, $y_{ti} - \bar{y}_{t\cdot}$

- Plot residuals, $y_{ti} - \bar{y}_{t\cdot}$, against fitted values, $\bar{y}_{t\cdot}$.

- Plot SD versus mean for each group.

- Plot the residuals against other factors.

  Order of measurements, weight or age of mouse, etc.

Example

ANOVA Tables

Original scale / 1000:

source	SS	df	MS	F	P-value
between strains	33	20	1.69	1.70	0.042
within strains	124	125	0.99		
total	157	145			

$\log_{10}$ scale:

source	SS	df	MS	F	P
between strains	3.35	20	0.167	2.25	0.0036
within strains	9.29	125	0.074		
total	12.63	145			

Residuals
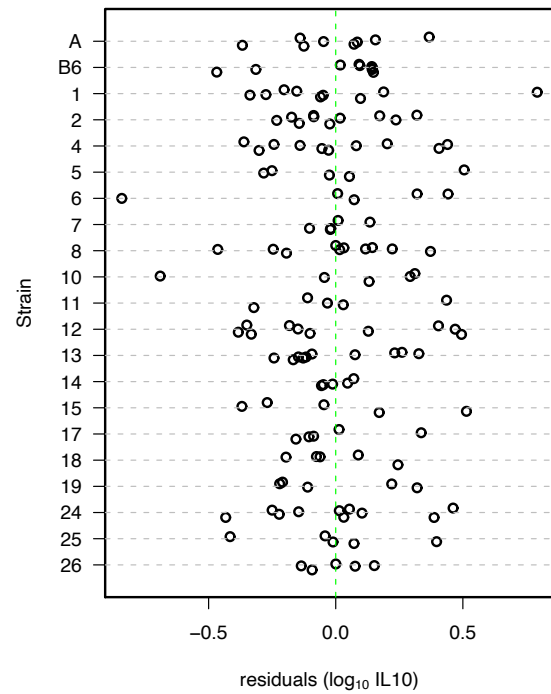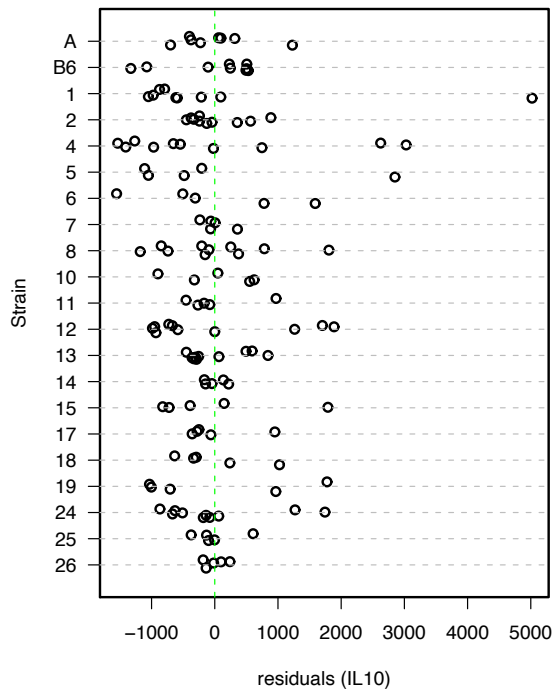
QQ plots of all residuals

IL10
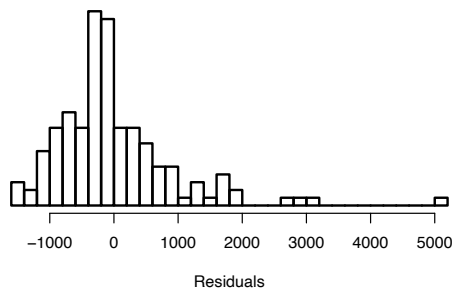
log₁₀ IL10

Residuals vs fitted values

SDs vs means

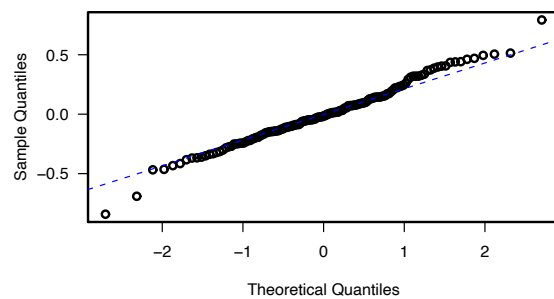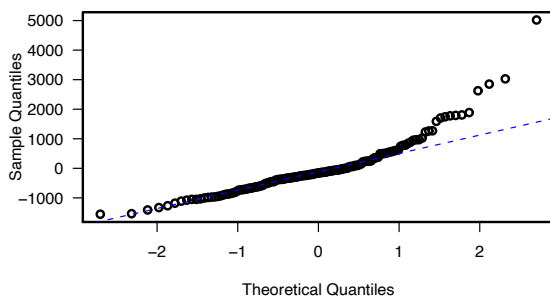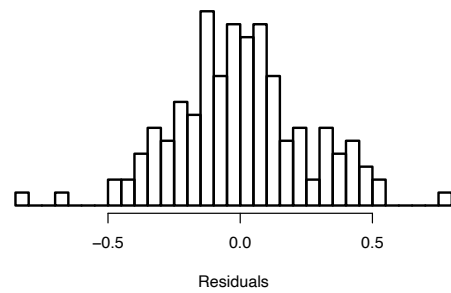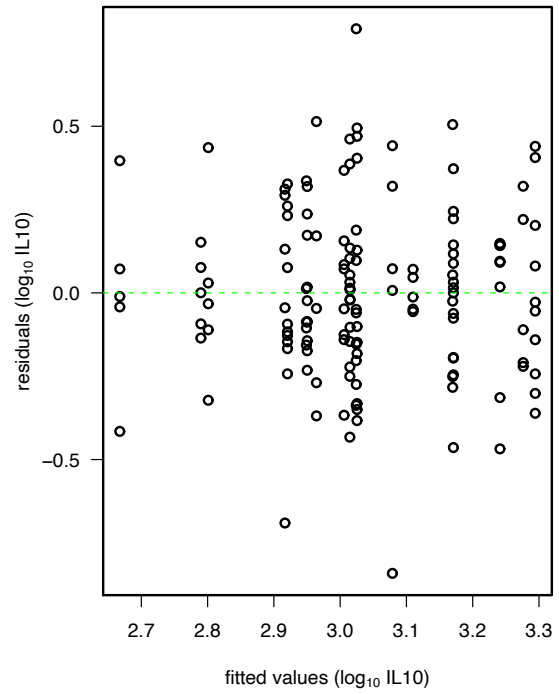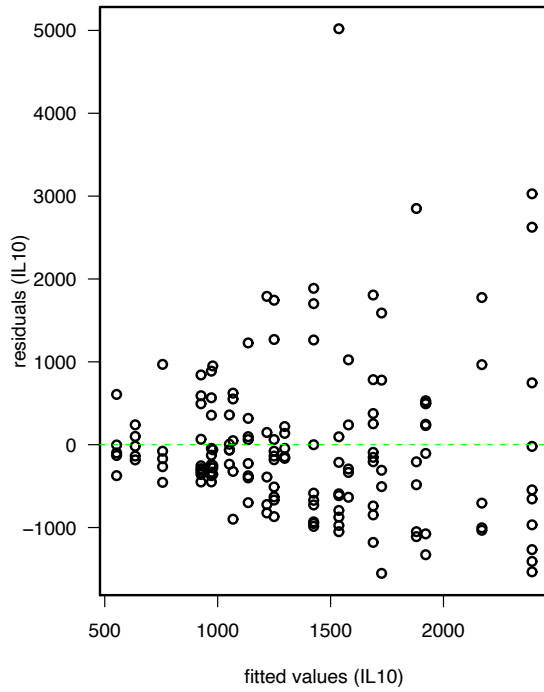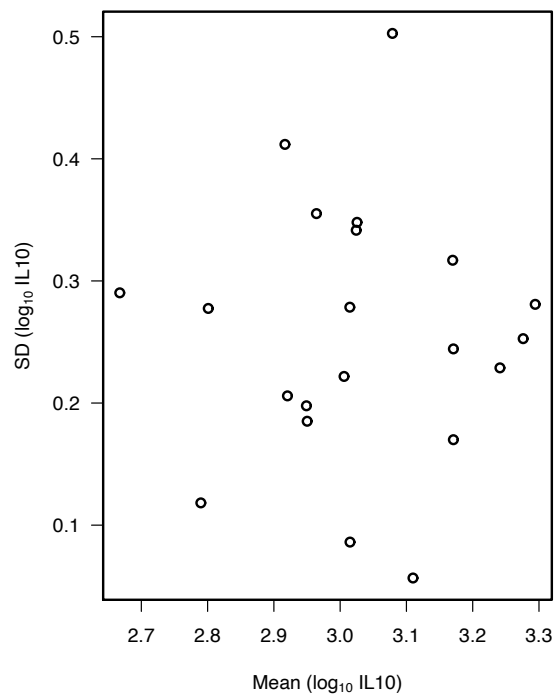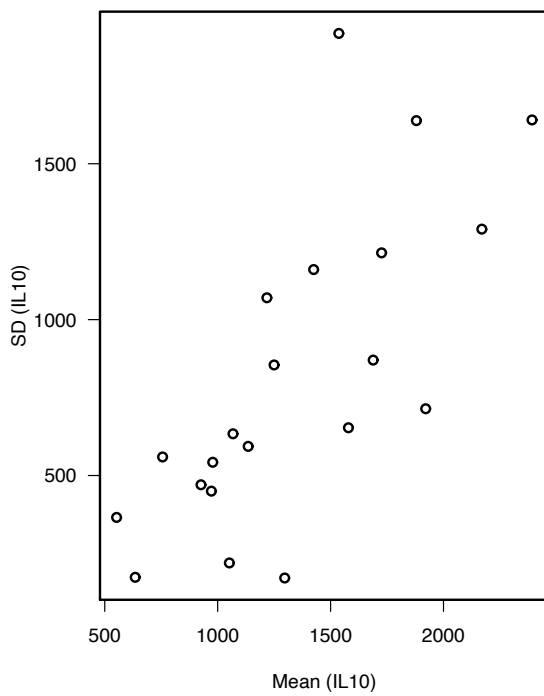# Homogeneity of variances

One of the ANOVA assumptions was homogeneity of the group variances. This can formally be tested with Bartlett's test.

Assume we have k treatment groups.

n_t number of cases in treatment group t.

N number of cases (overall).

Y_{ti} response i in treatment group t.

$\bar{Y}_{t.}$ average response in treatment group t.

S_t^2 the sample variance in treatment group t.

Bartlett's test

We want to test $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ versus $H_a : H_0$ is false.

- Calculate the pooled sample variance:

$$S^2 = \frac{\sum_t (n_t - 1) \times S_t^2}{\sum_t (n_t - 1)} = \frac{\sum_t (n_t - 1) \times S_t^2}{N - k}$$

- Calculate the test statistic

$$X^2 = (N - k) \times \log(S^2) - \sum_t (n_t - 1) \times \log(S_t^2)$$

- Calculate the following correction factor:

$$C = 1 + \frac{1}{3(k - 1)} \left[\sum_t \frac{1}{n_t - 1} - \frac{1}{\sum_t (n_t - 1)} \right]$$

If H_0 is true, then

$$X^2/C \sim \chi^2(df = k-1)$$

Example

- For the example data, there are 21 strains with between 5 and 10 observations per strain.

- The pooled sample variance on original scale / 1000 is 0.99.

- The pooled sample variance on $\log_{10}$ scale is 0.074.

- The test statistics were 79.9 and 34.0.

- The correction factor ended up being 1.07.

- Thus we look at the values 79.9 / 1.07 = 74.8 and 34.0 / 1.07 = 31.8.

- Since there are 21 strains, we refer to the χ^2(df = 20) distribution.

- We end up with P-values of 2.9×10^{-8} and 0.045.

$\longrightarrow$ The R function bartlett.test() can be used to do these calculations.