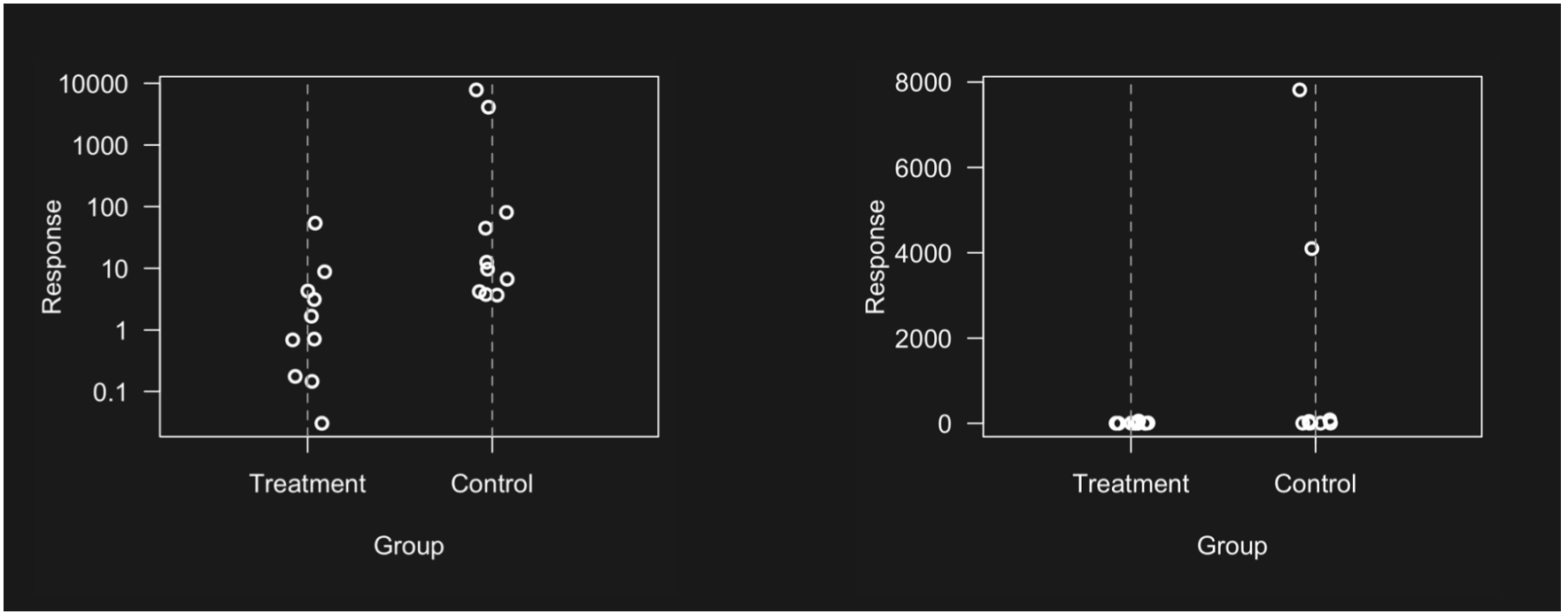
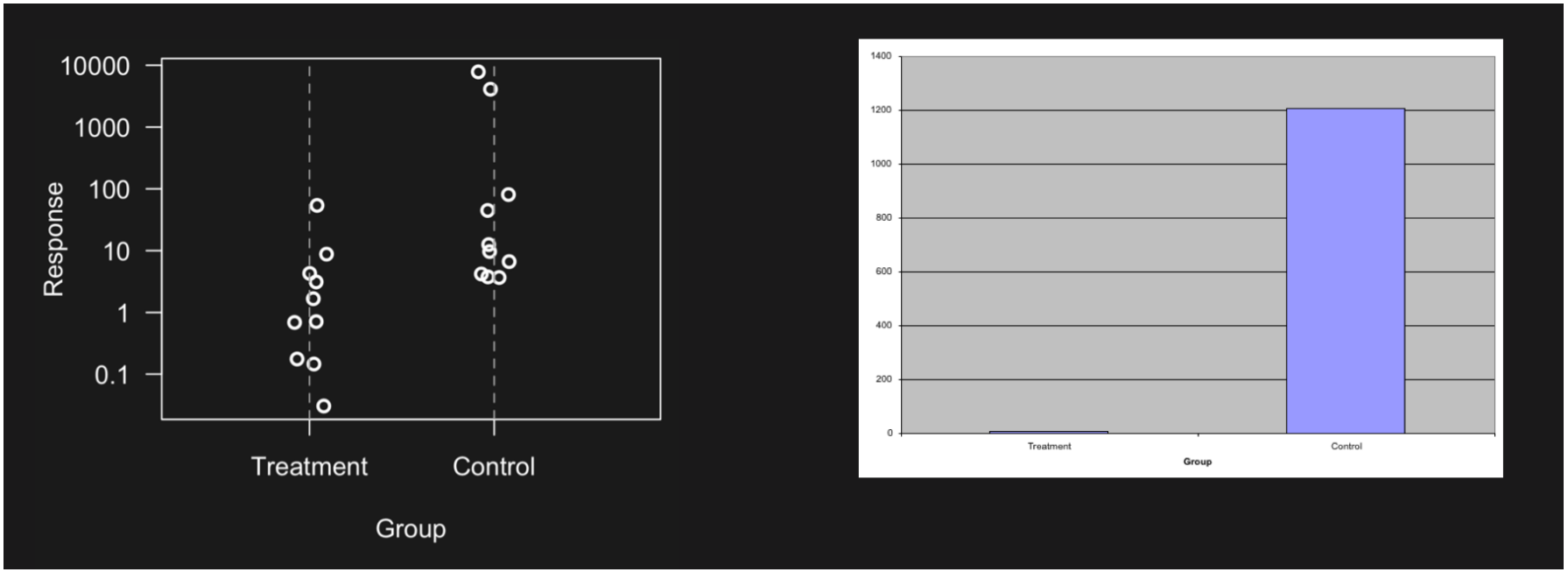
Consider data transformations

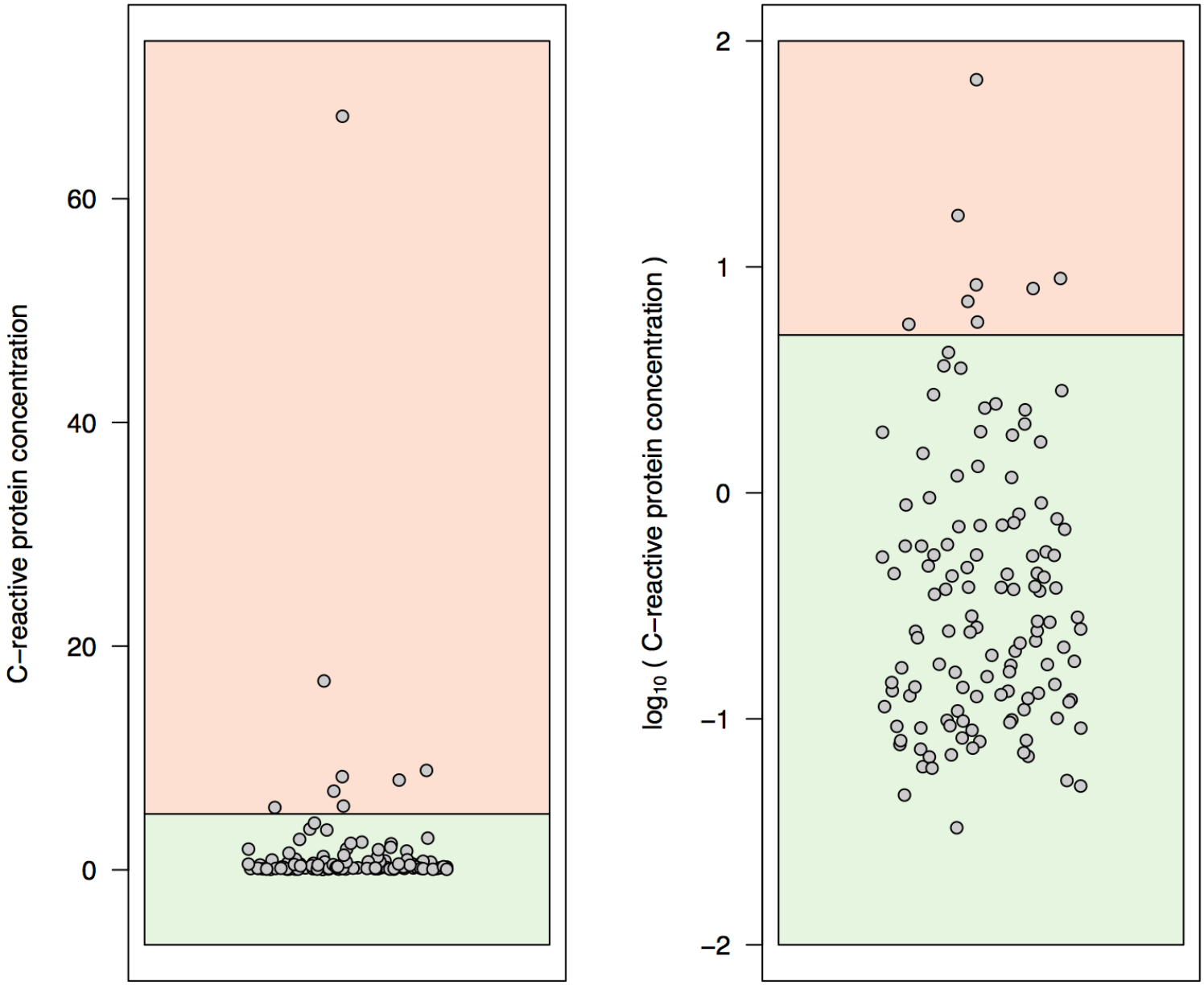


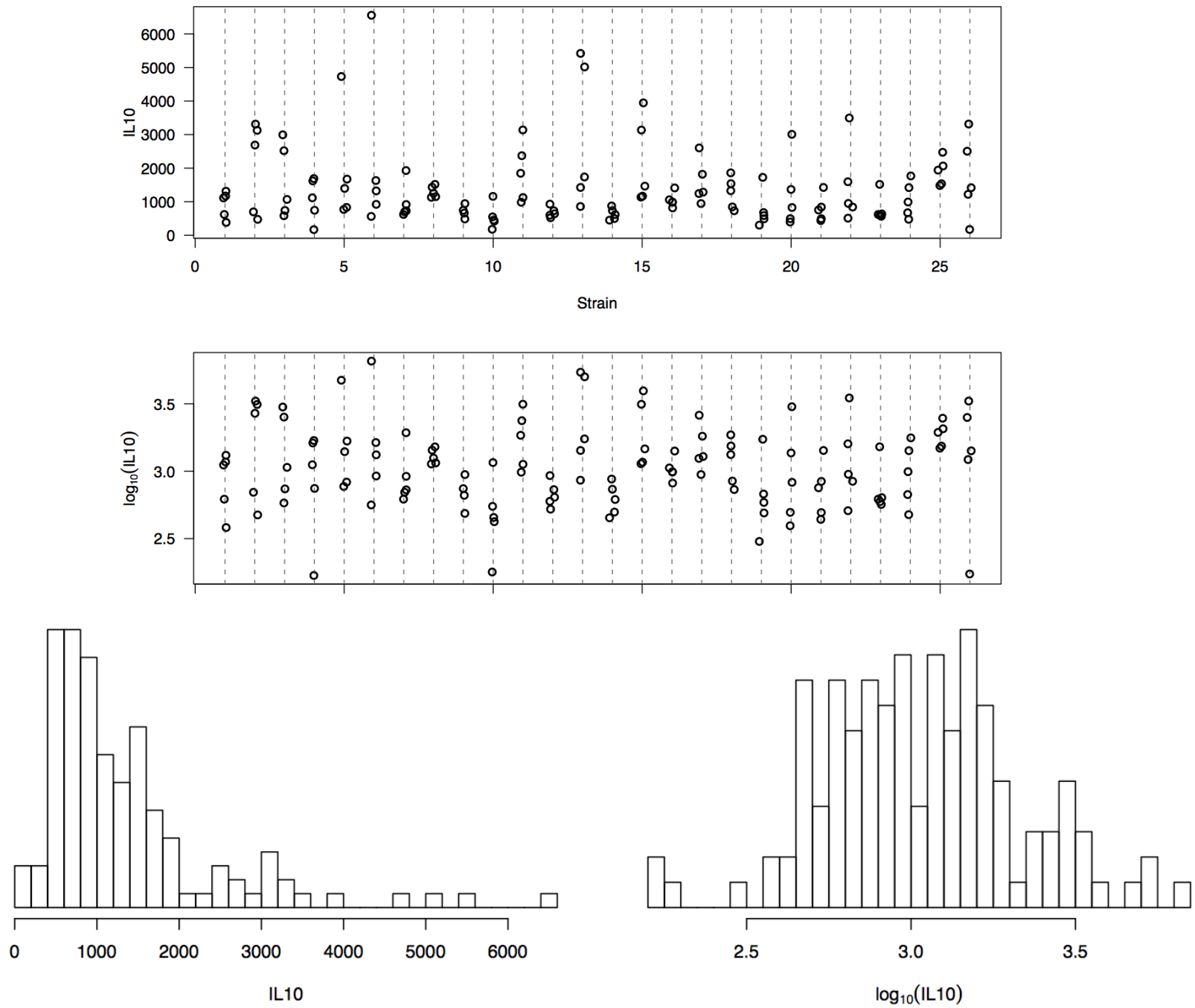
[KB]

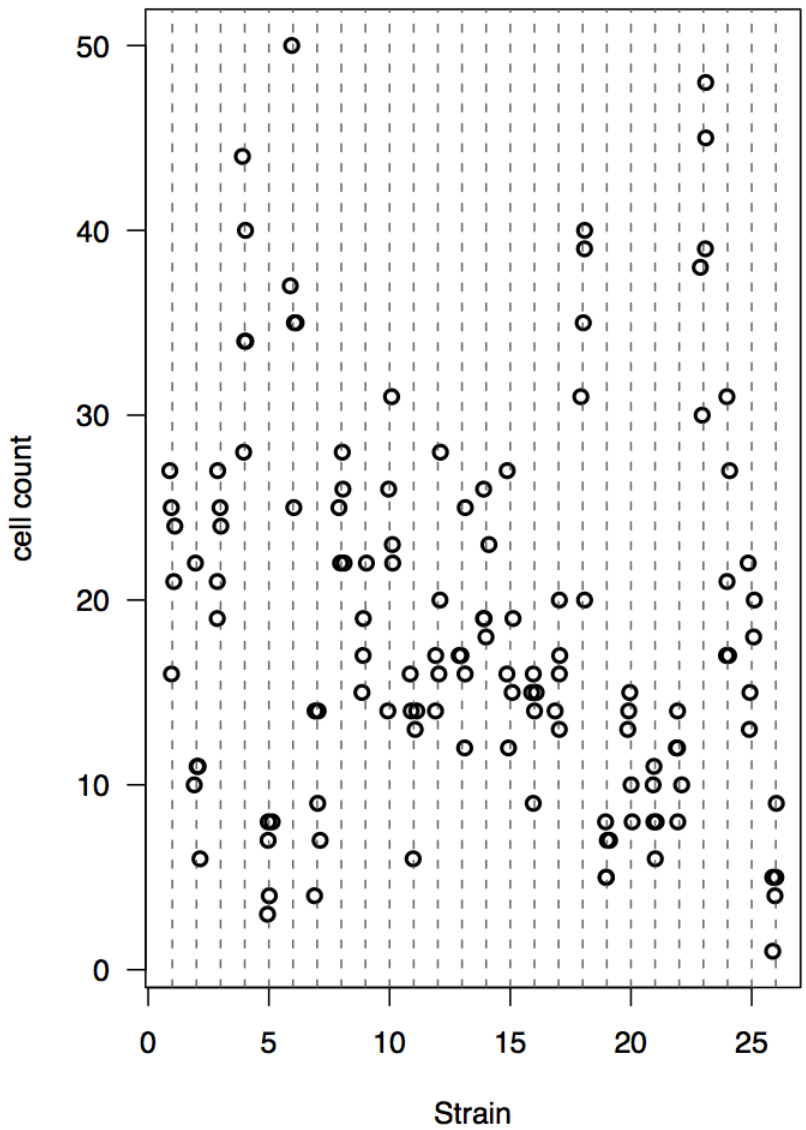
Consider data transformations

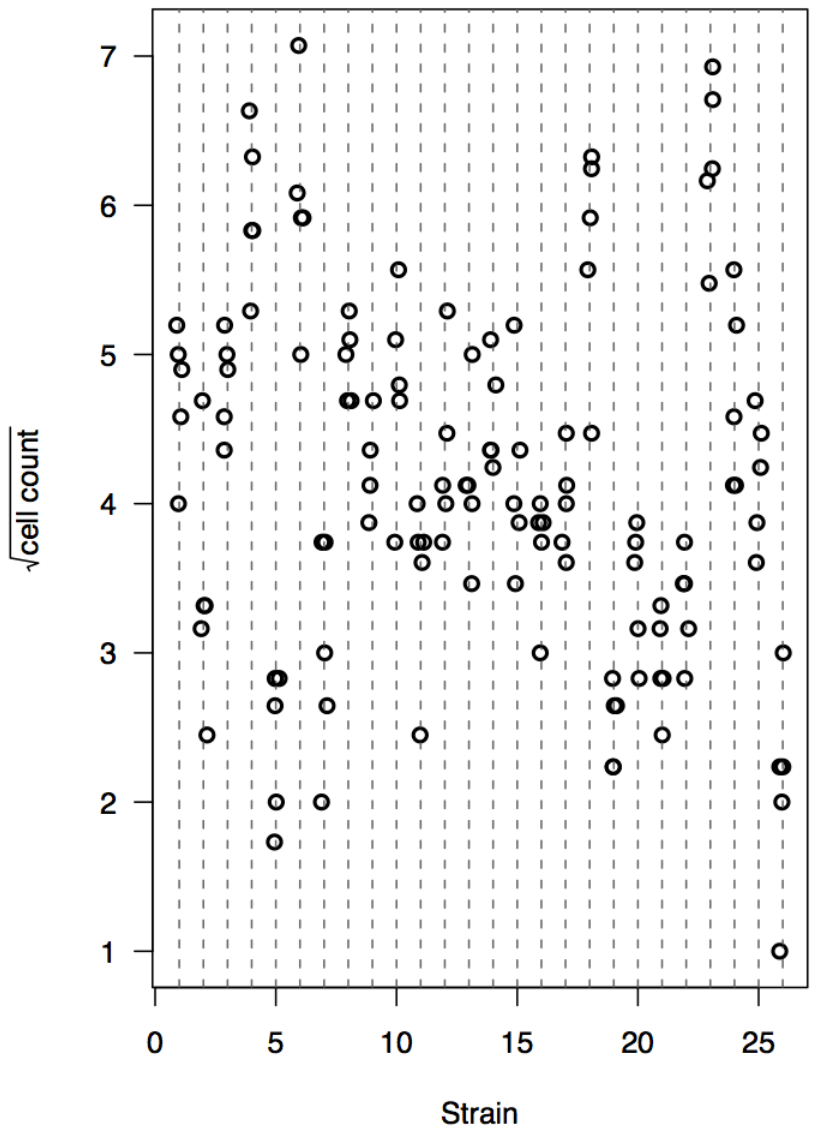


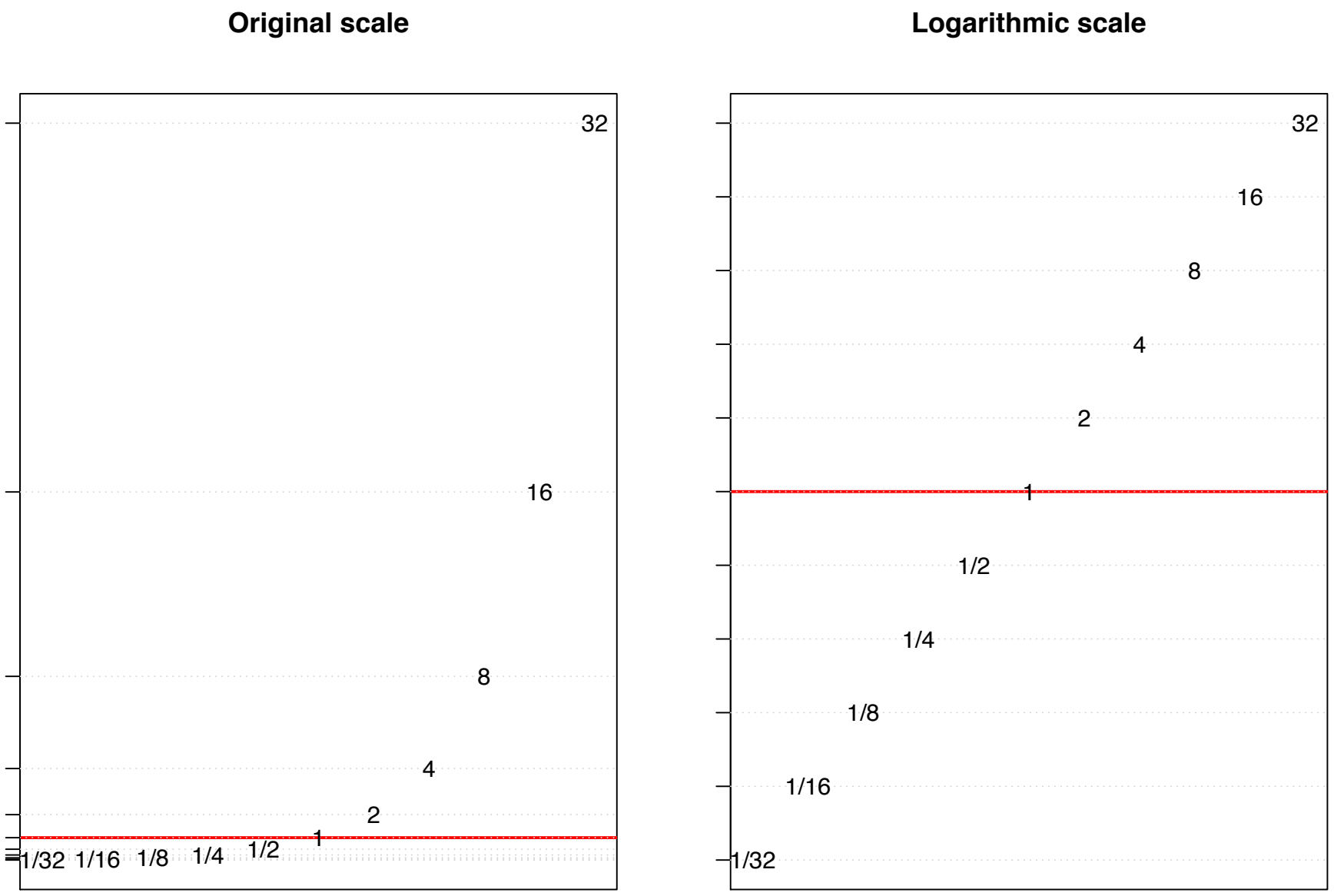
[KB]


C−reactive protein concentration
60
40
20
0
log10 (C−reactive protein concentration)
2
1
0
−1
−2

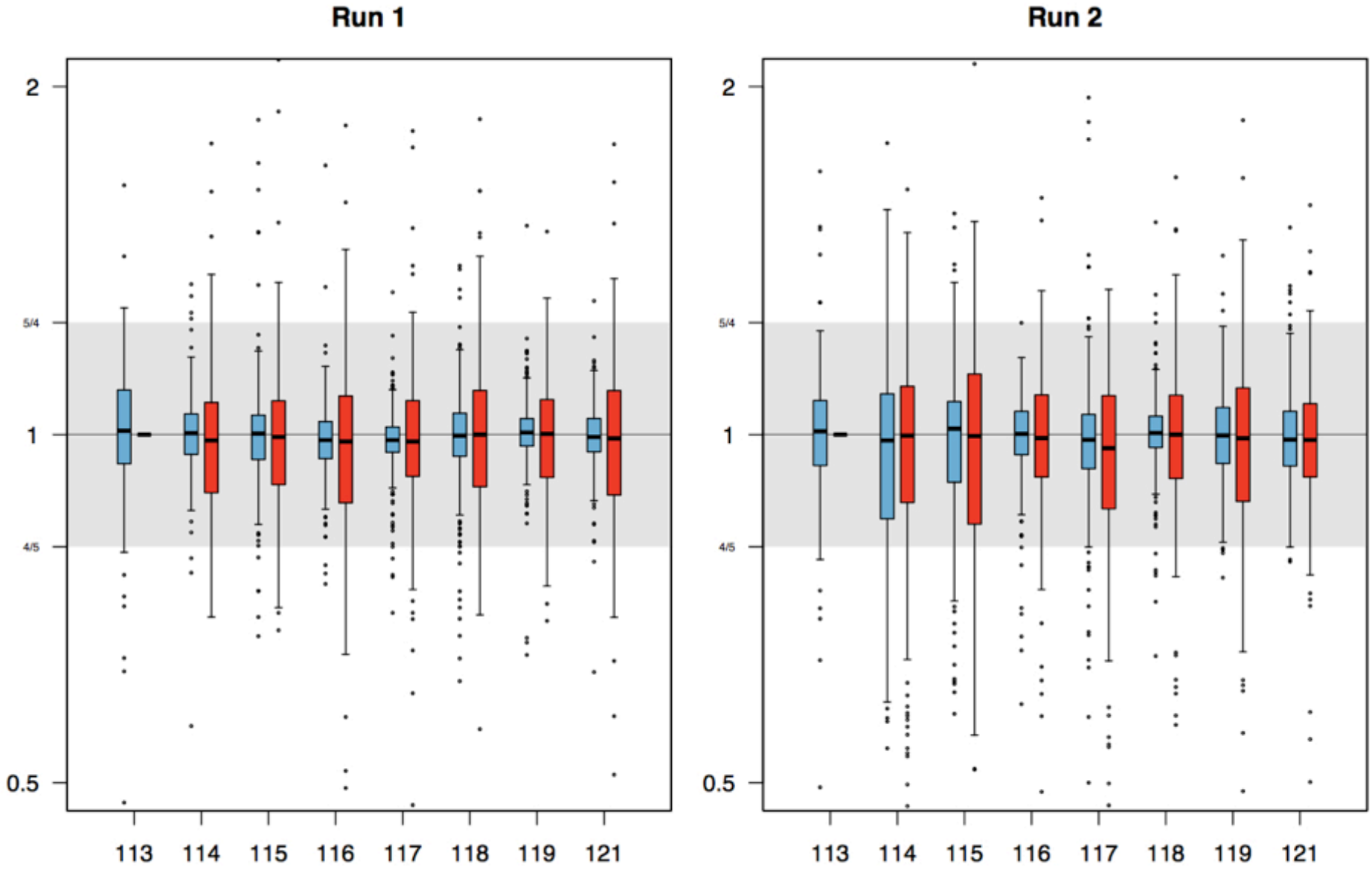

IL10
Strain
log10(IL10)
IL10
log10(IL10)


cell count
0
10
20
30
40
50
0
5
10
15
20
25
Strain


√cell count
1
2
3
4
5
6
7
0
5
10
15
20
25
Strain



$$\log(X/Y) = \log(X) - \log(Y) = -(\log(Y) - \log(X)) = -\log(Y/X)$$


Run 1
Run 2
2
5/4
1
4/5
0.5
113
114
115
116
117
118
119
121

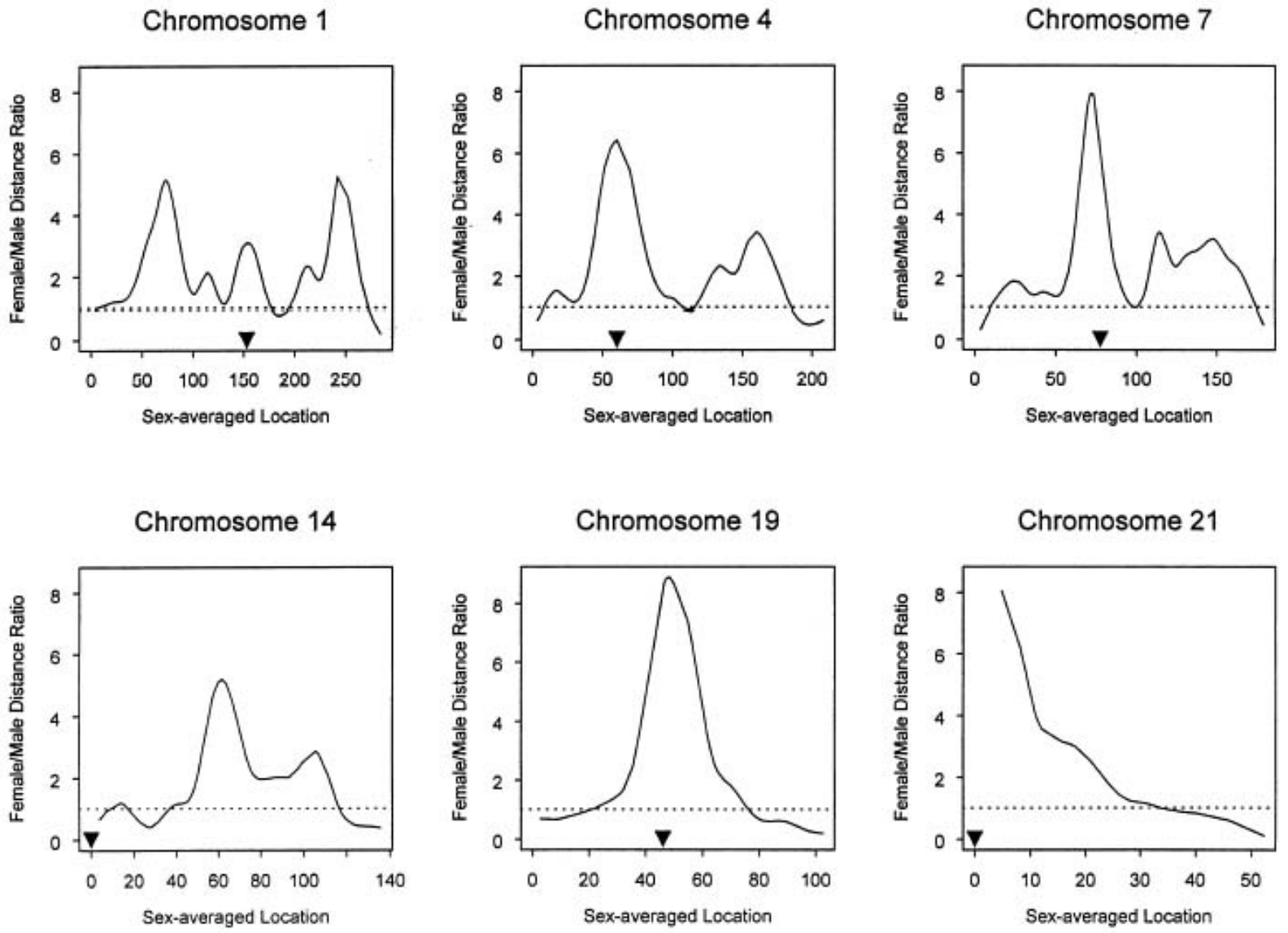


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

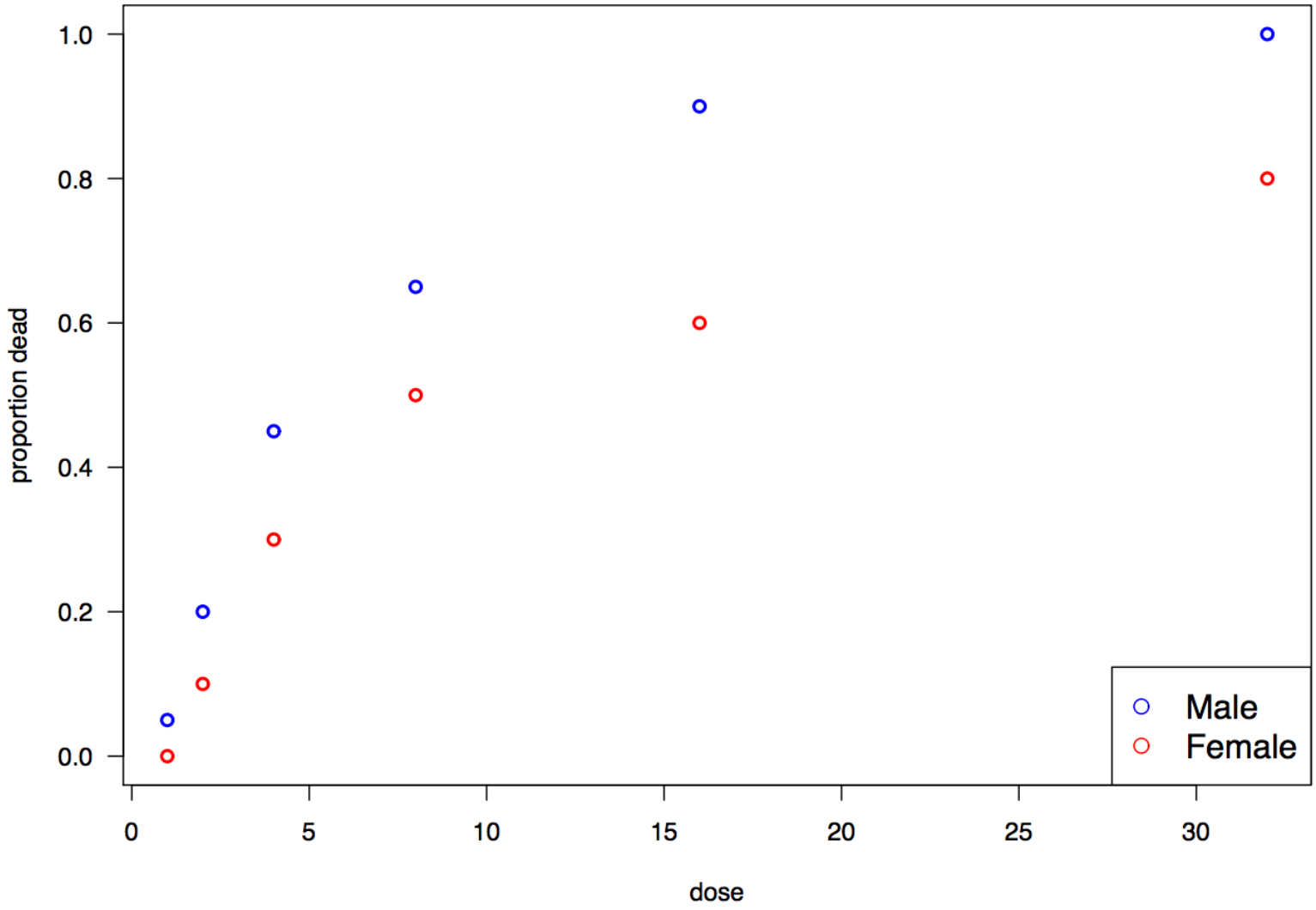
PMID 9718341

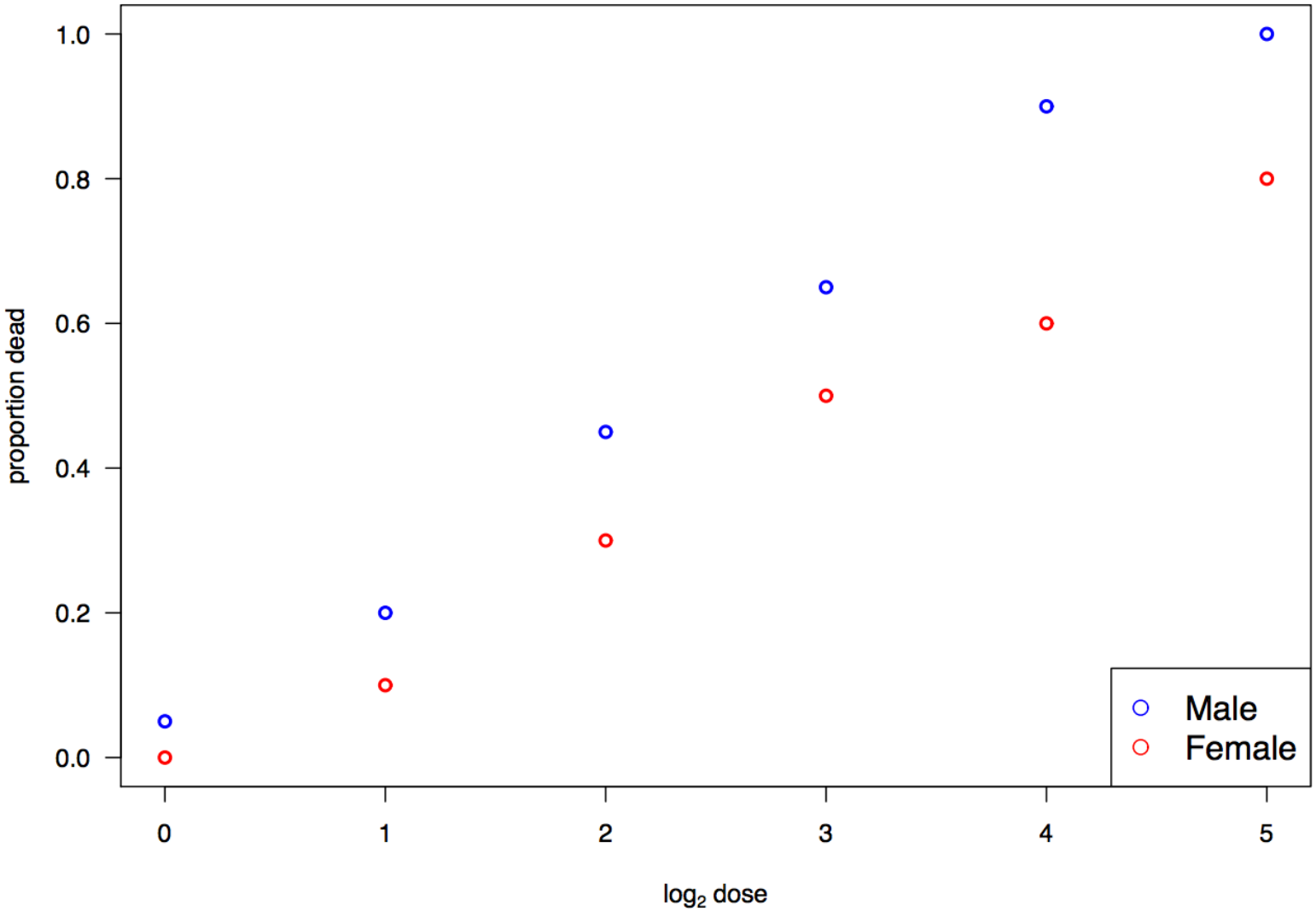
Tobacco budworm, *Heliothis virescens*

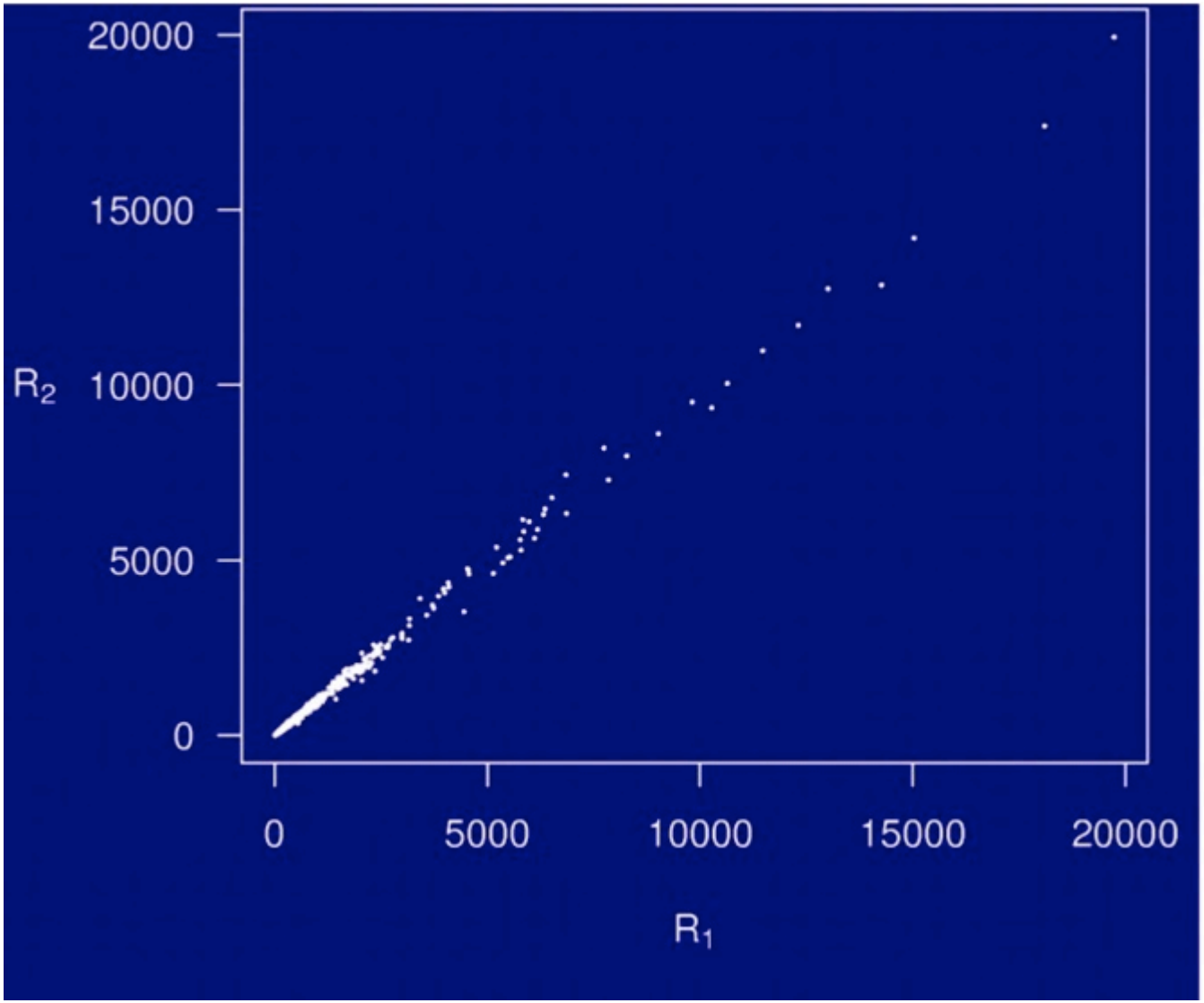
Batches of 20 male and 20 female worms were given a 3-day dose of pyrethroid *trans*-cypermethrin

The no. dead or “knocked down” in each batch was noted.

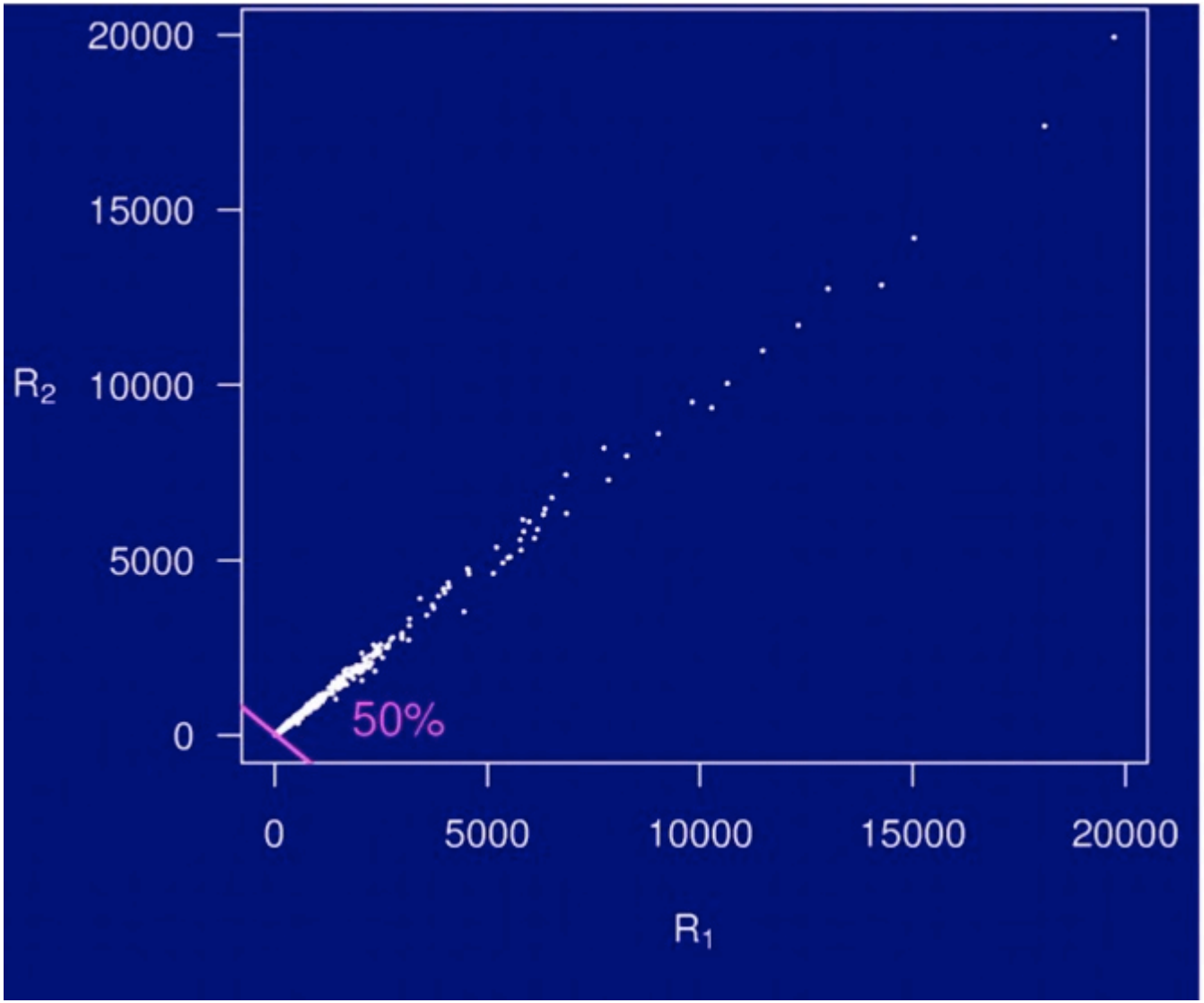
	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16


1.0
0.8
0.6
0.4
0.2
0.0
proportion dead
0
5
10
15
20
25
30
dose
Male
Female

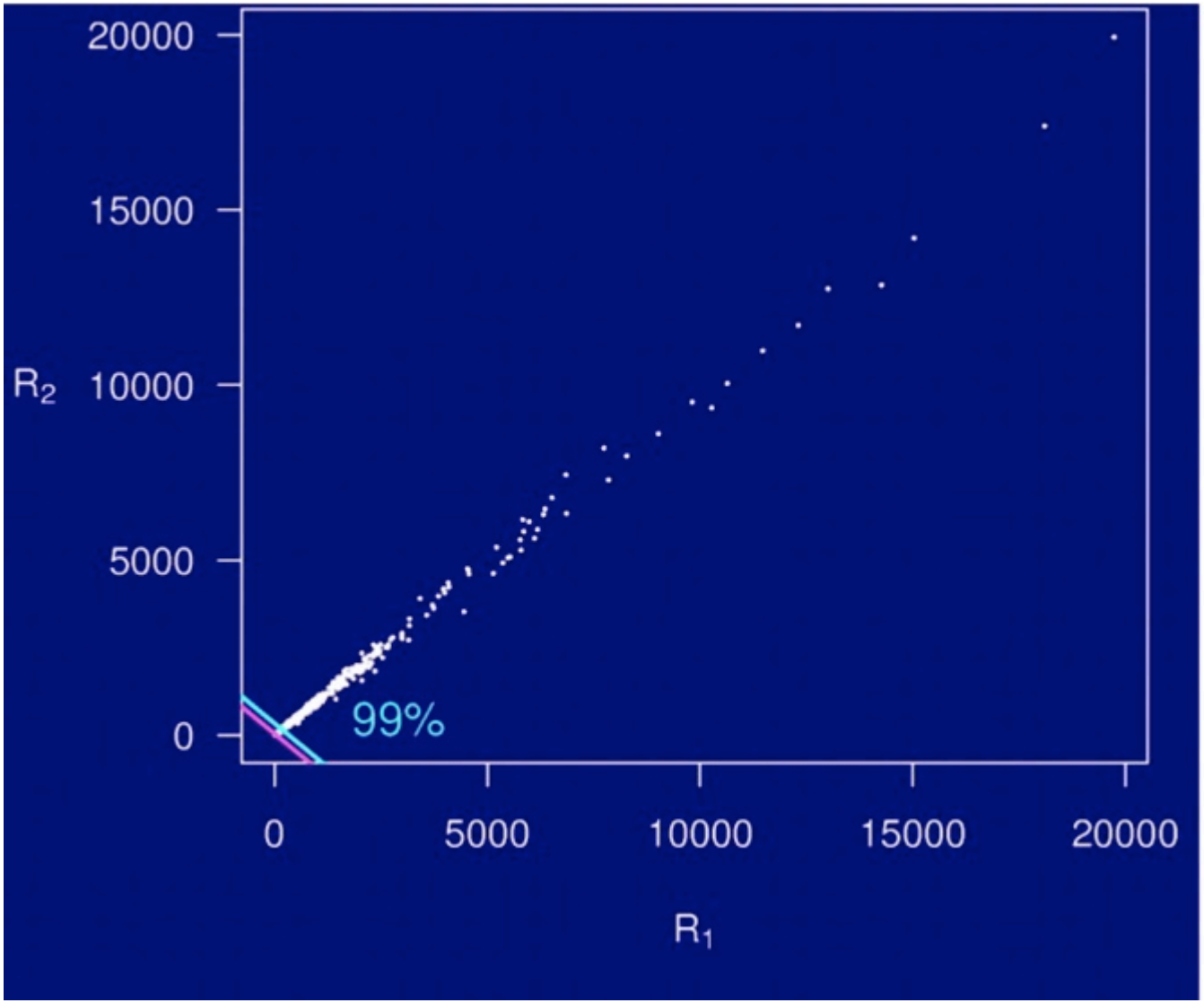

1.0
0.8
0.6
0.4
0.2
0.0
proportion dead
0
1
2
3
4
5
log2 dose
Male
Female



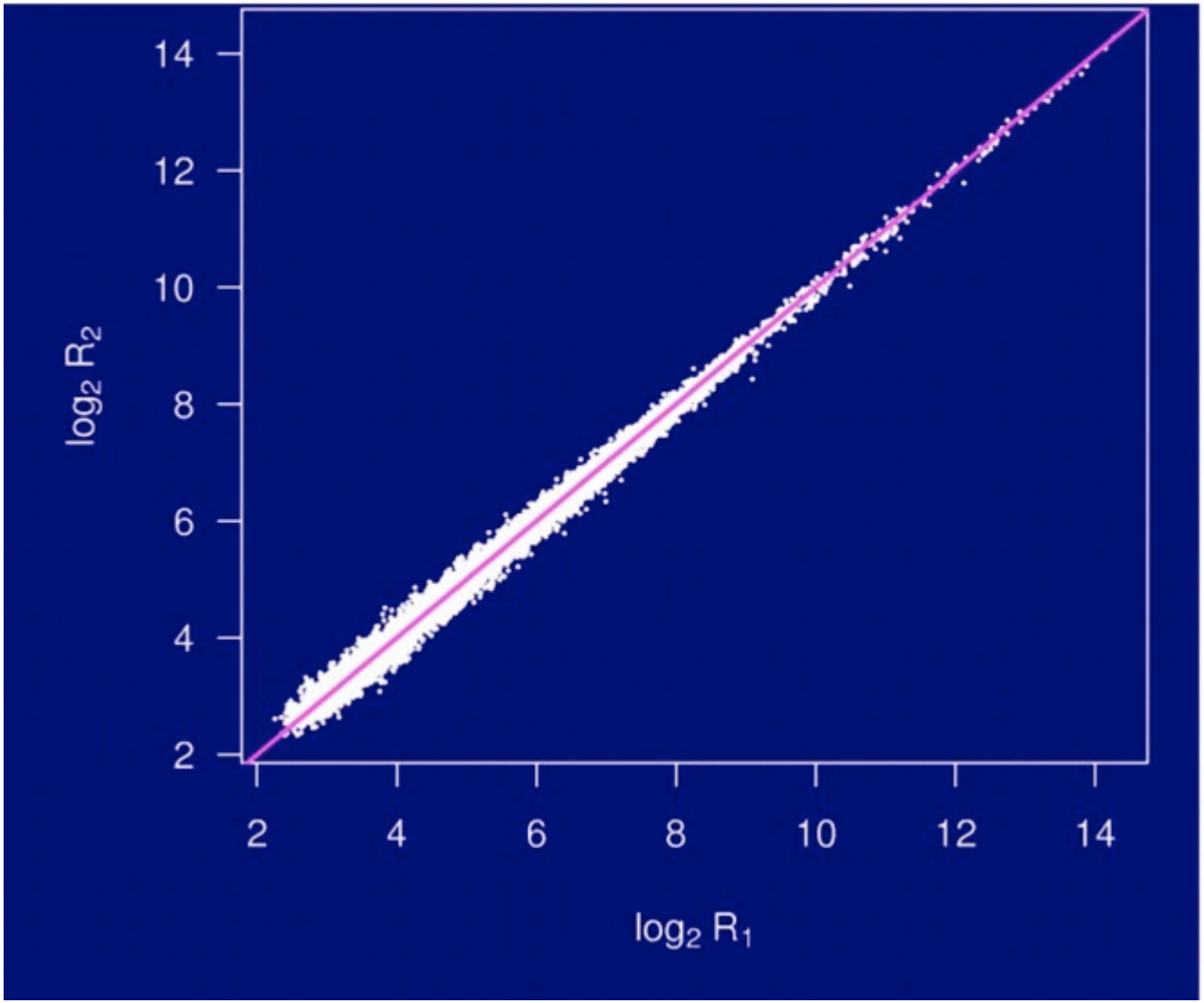
[KB]



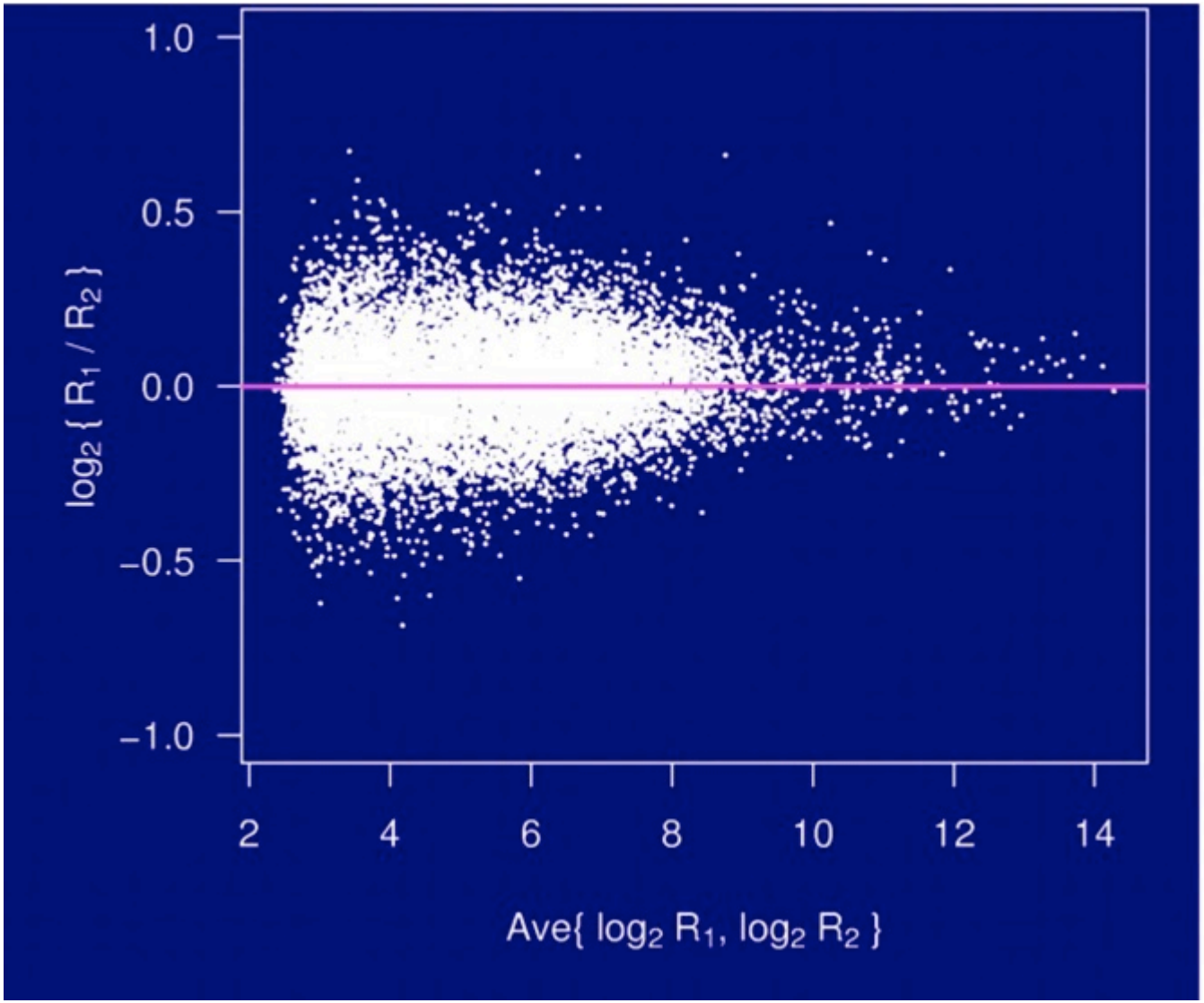
[KB]



[KB]



[KB]



[KB]

Summary statistics

Location / Center

- mean (average)
- median
- mode
- geometric mean
- harmonic mean

Scale

- standard deviation (SD)
- inter-quartile range (IQR)
- range

Other

- quantile
- quartile
- quintile

Summary statistics

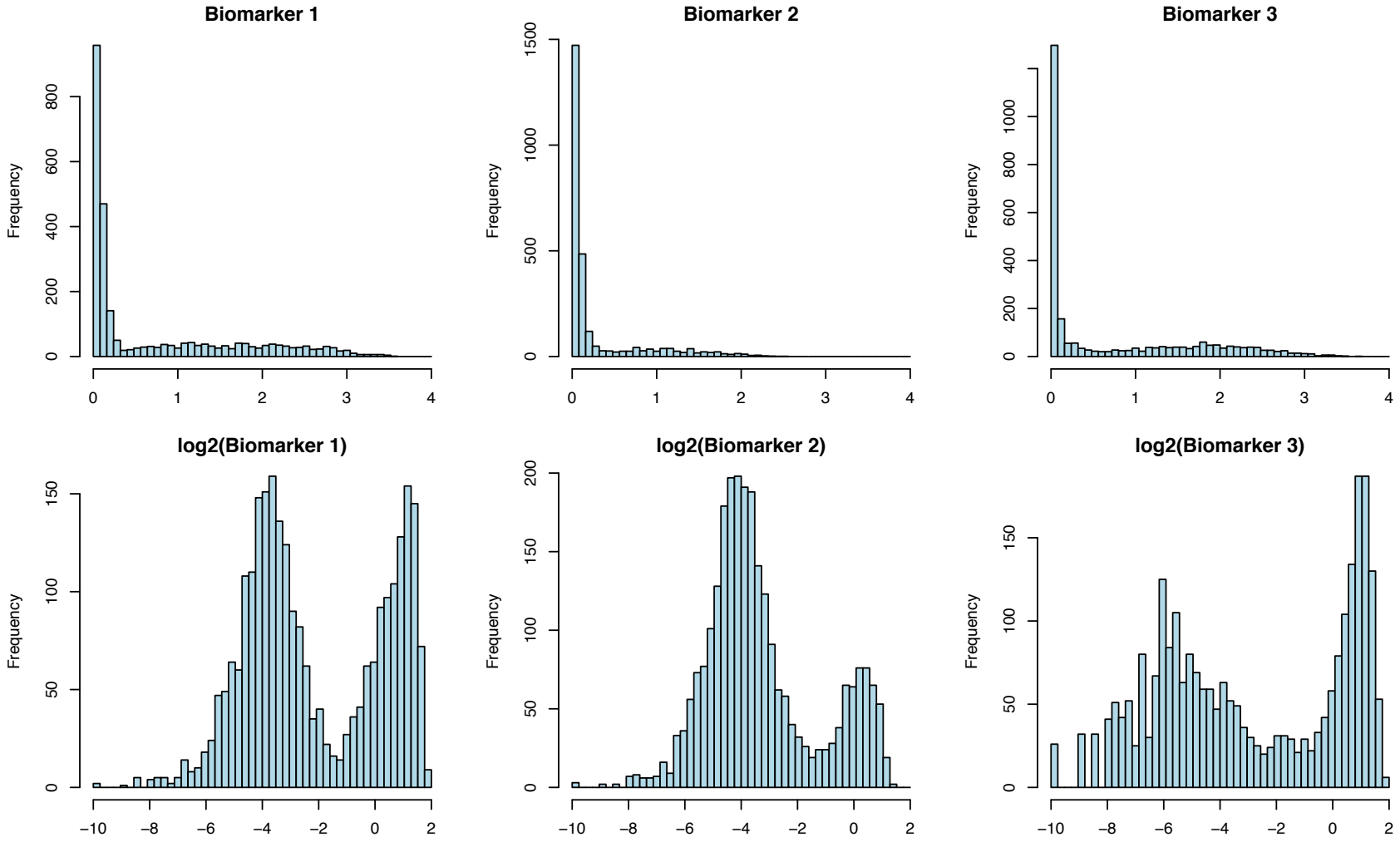
$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + x_2 + \ldots + x_n)/n$$

$$\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log x_i\right\}$$

$$\text{harmonic mean} = 1/\left\{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)\right\}$$

$\longrightarrow$ Note: these are all sample means.

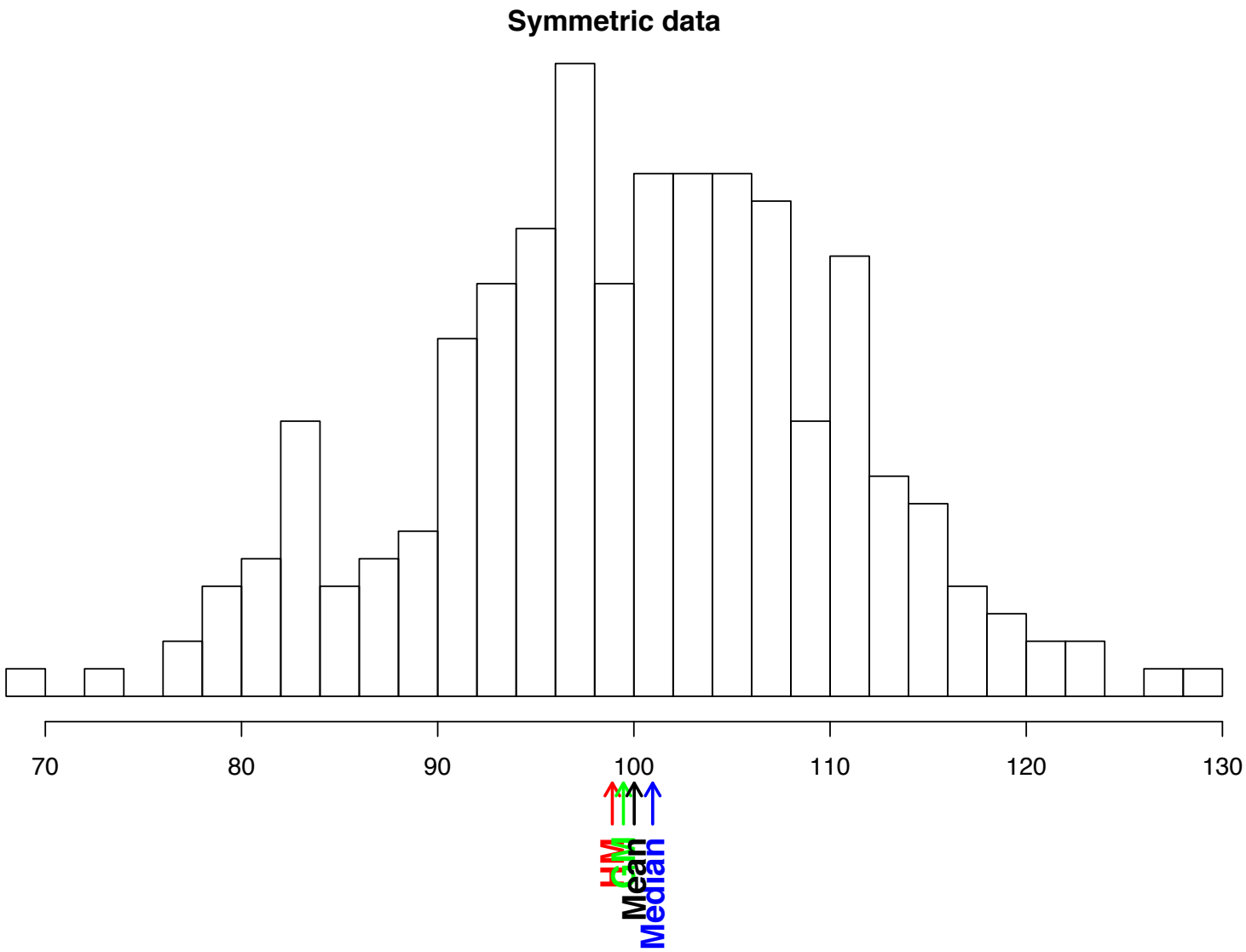
Measures of location / center



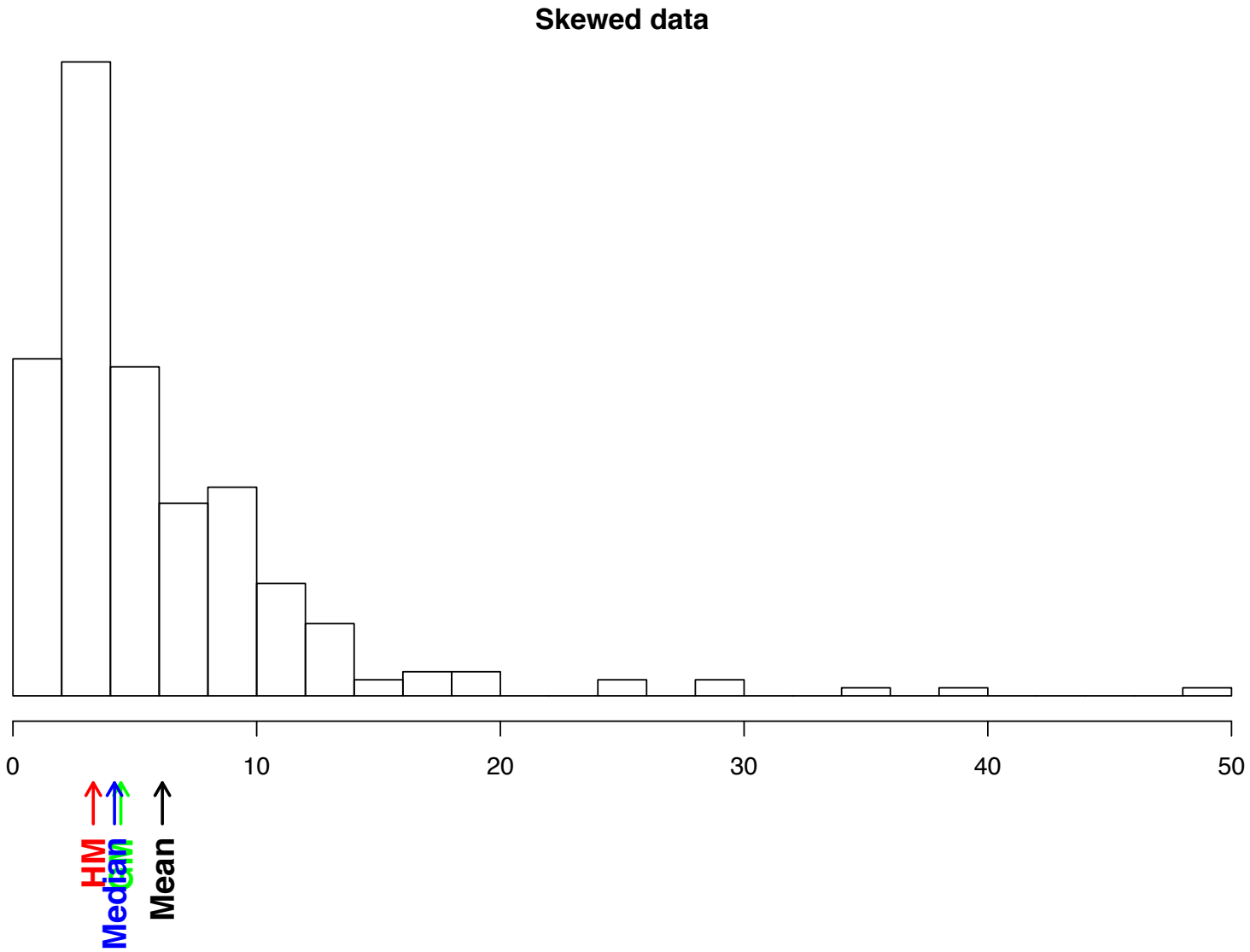
Measures of location / center

- Forget about the mode.
- The mean is sensitive to outliers.
- The median is resistant to outliers.
- The geometric mean is used when a logarithmic transformation is appropriate (for example, when the distribution has a long right tail).
- The harmonic mean may be used when a reciprocal transformation is appropriate (very seldom).

Measures of location / center



Measures of location / center



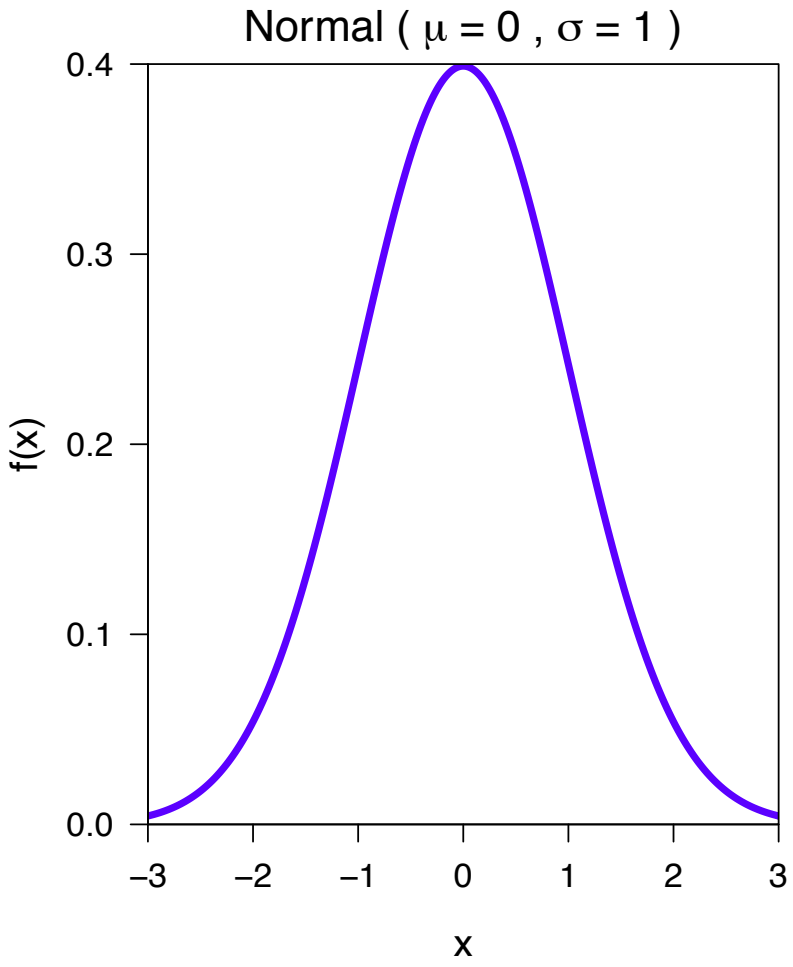
A key point

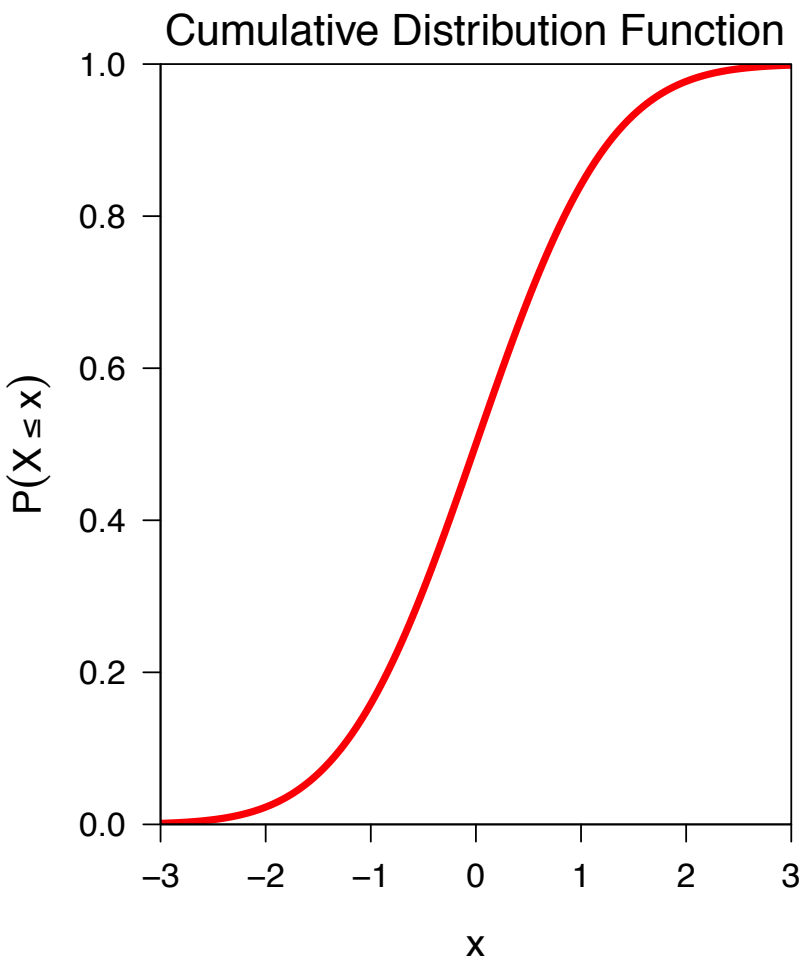
The different possible measures of the ”center” of the distribution are all allowable.

You should consider the following though:

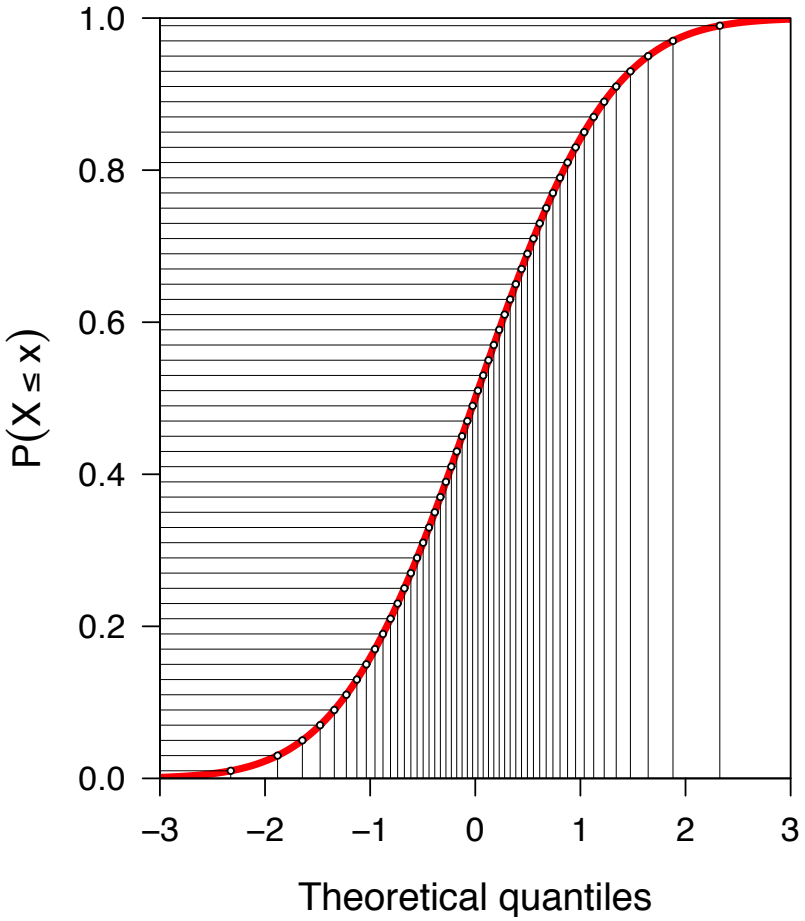
- $\longrightarrow$ Which is the best measure of the ”typical” value in your particular setting?
- $\longrightarrow$ Be sure to make clear which ”average” you use.

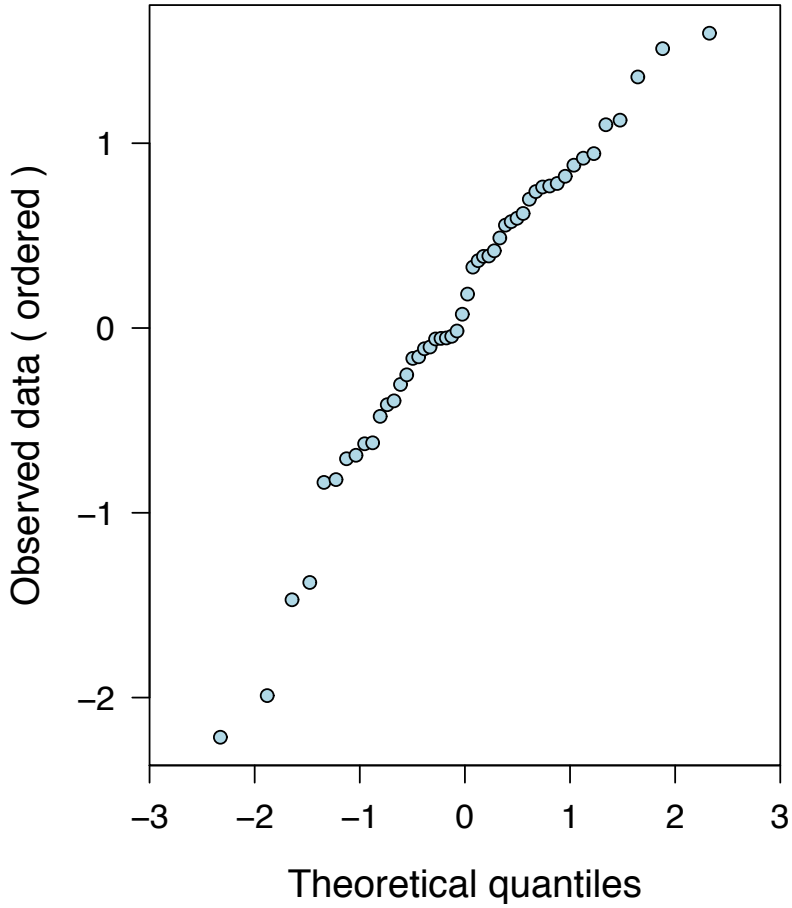
QQ-plots


Normal (μ = 0 , σ = 1)
f(x)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

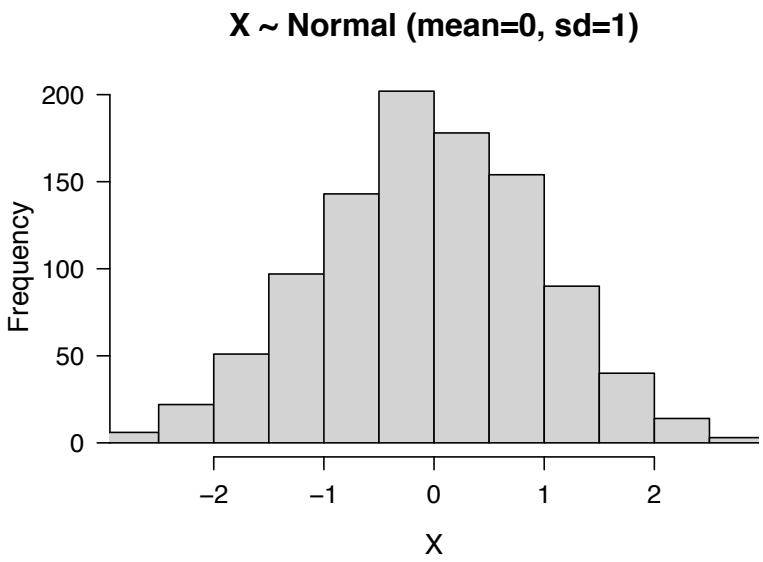

Cumulative Distribution Function
P(X ≤ x)
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3
x

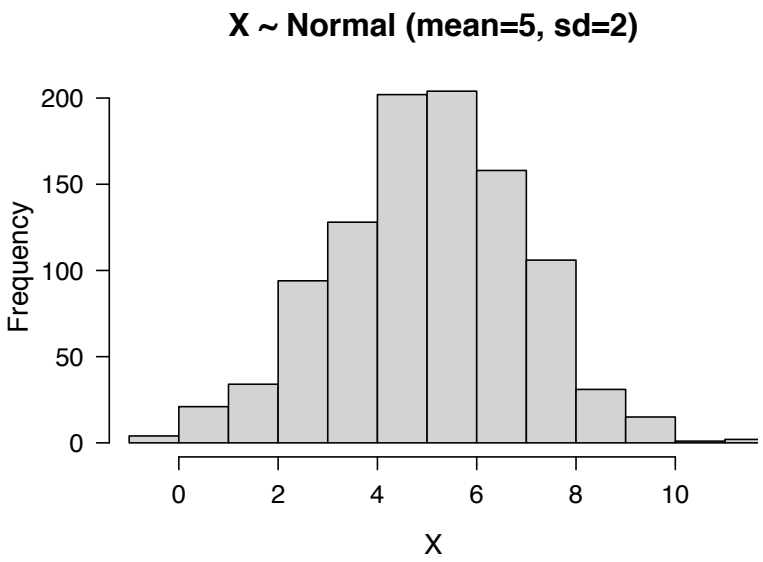
QQ-plots

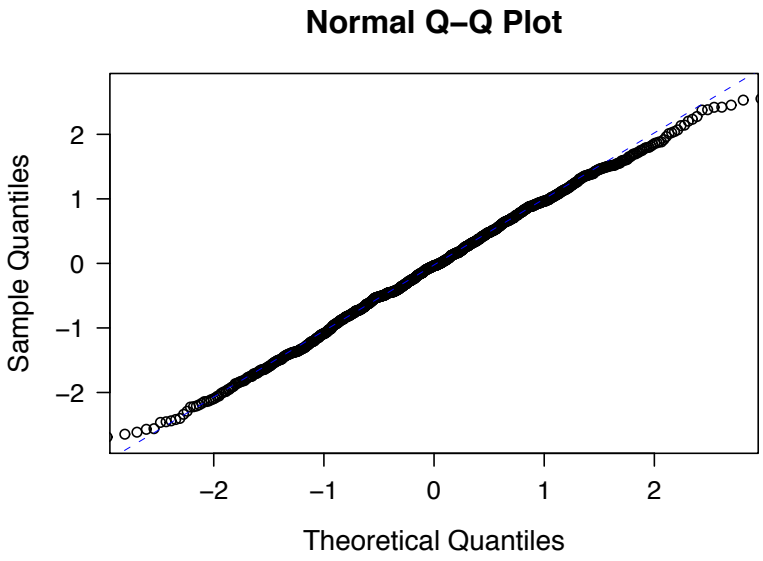

1.0
0.8
0.6
0.4
0.2
0.0
P(X ≤ x)
−3
−2
−1
0
1
2
3
Theoretical quantiles

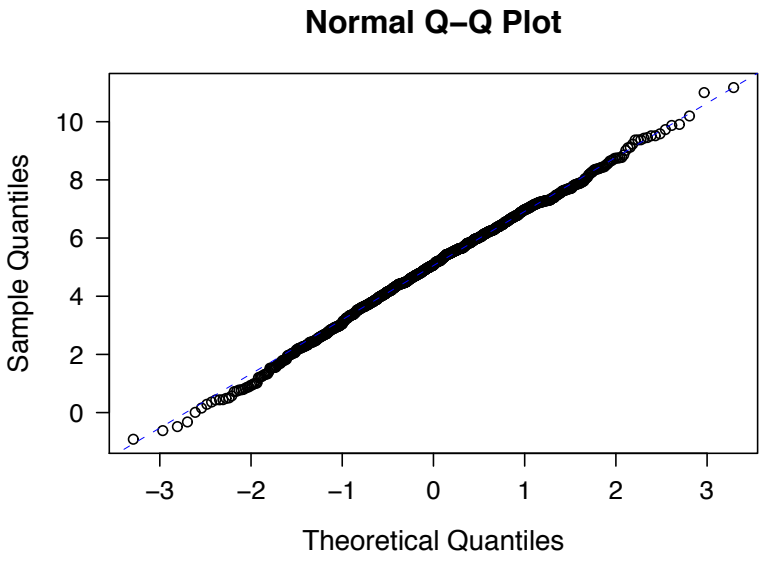

1
0
−1
−2
Observed data (ordered)
−3
−2
−1
0
1
2
3
Theoretical quantiles

QQ-plots

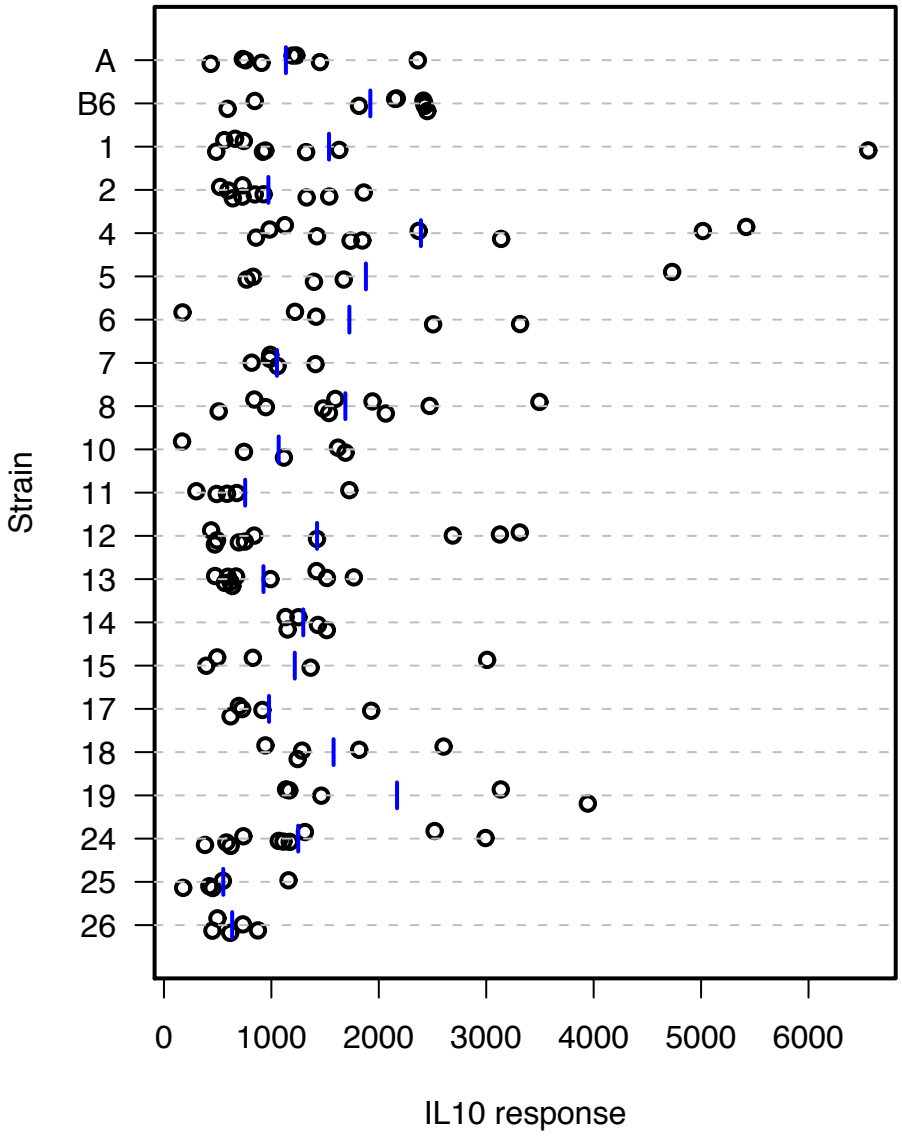

X ~ Normal (mean=0, sd=1)
Frequency
200
150
100
50
0
−2
−1
0
1
2
X

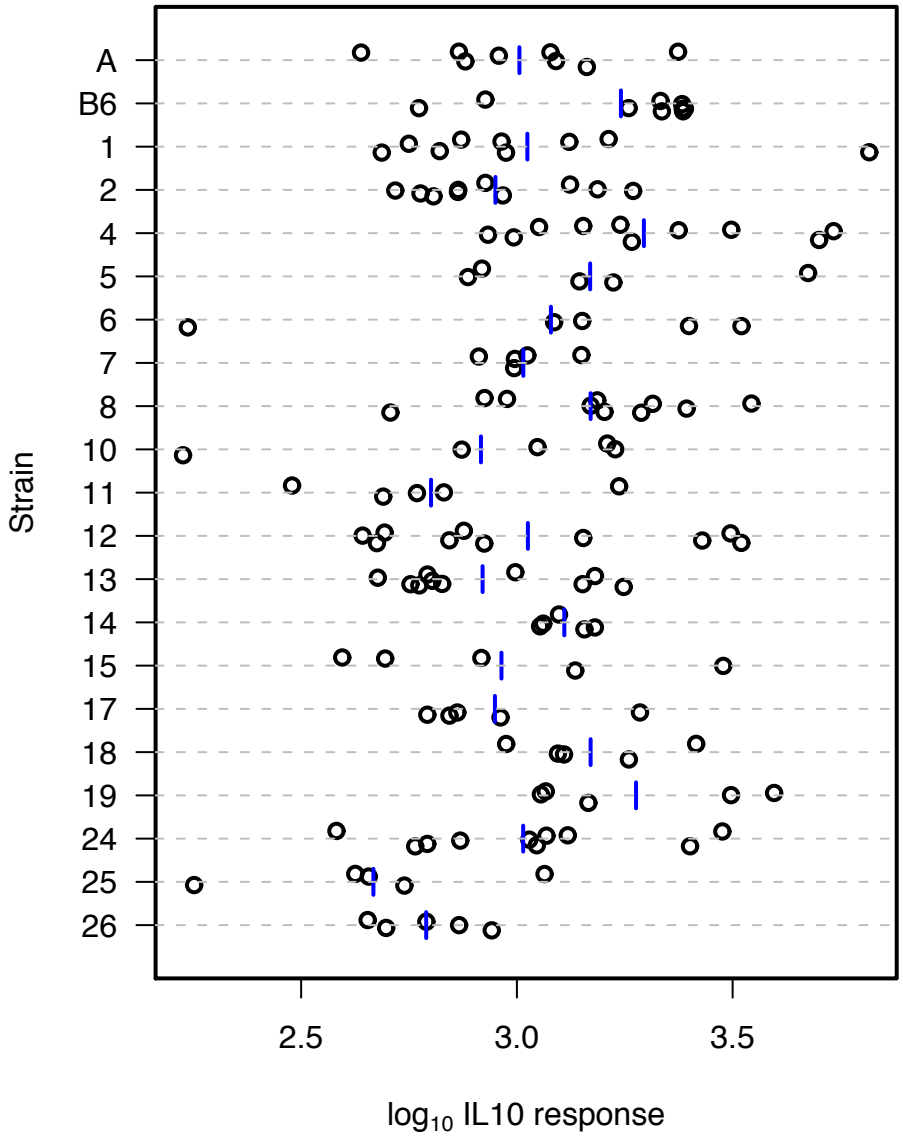

X ~ Normal (mean=5, sd=2)
Frequency
200
150
100
50
0
0
2
4
6
8
10
X


Normal Q–Q Plot
Sample Quantiles
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles

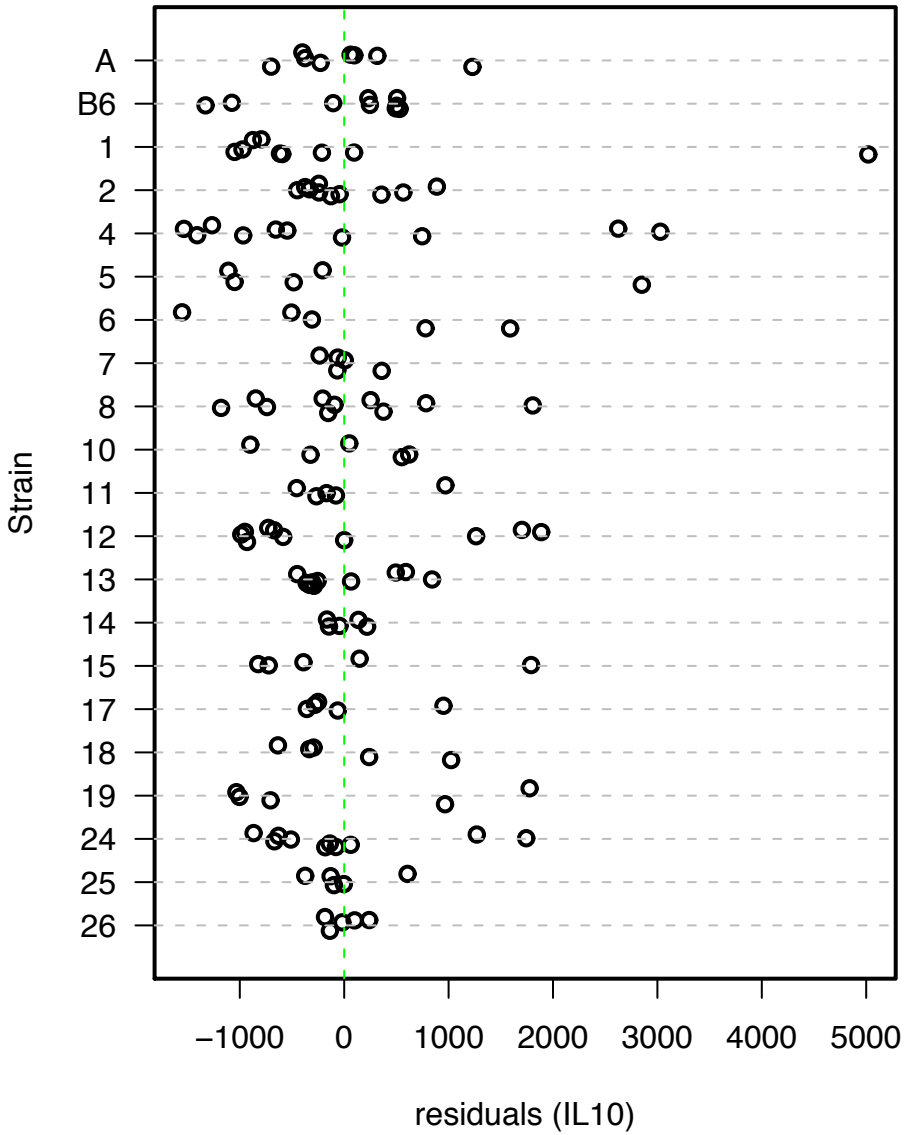

Normal Q–Q Plot
Sample Quantiles
10
8
6
4
2
0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

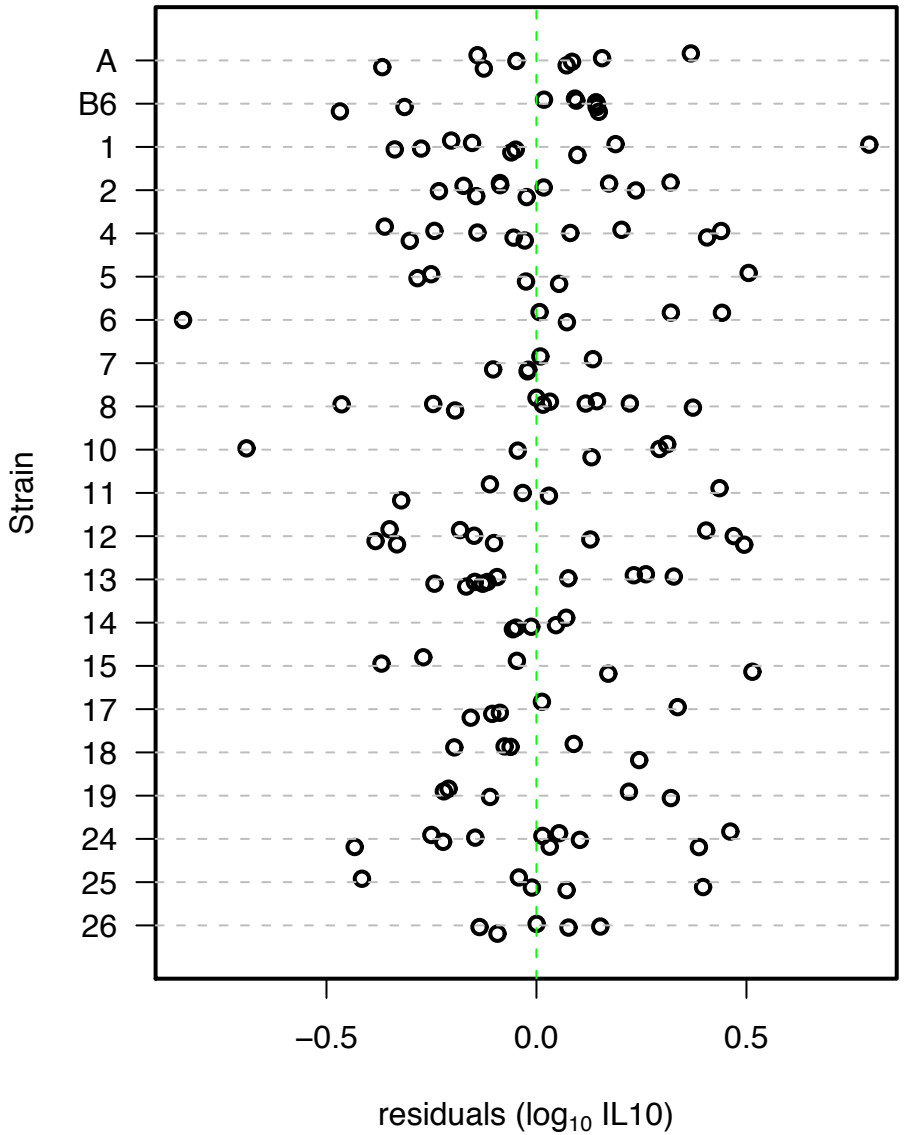
Example


Strain
A
B6
1
2
4
5
6
7
8
10
11
12
13
14
15
17
18
19
24
25
26
0
1000
2000
3000
4000
5000
6000
IL10 response

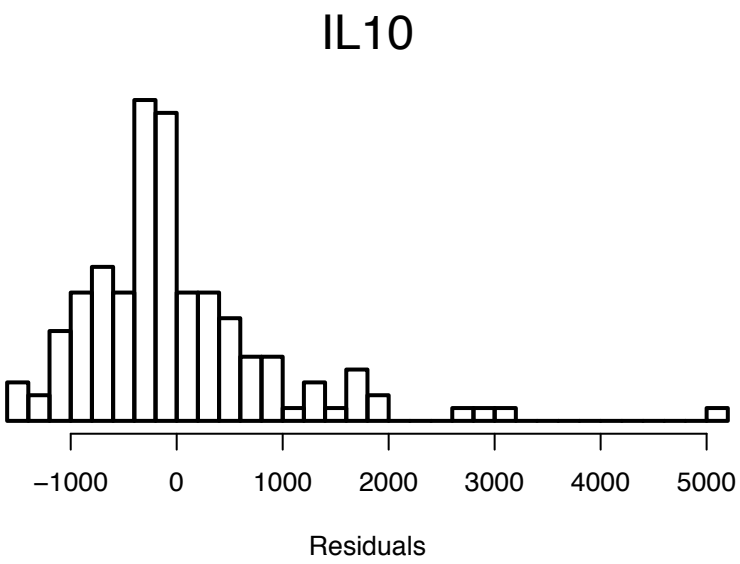

Strain
A
B6
1
2
4
5
6
7
8
10
11
12
13
14
15
17
18
19
24
25
26
2.5
3.0
3.5
log10 IL10 response

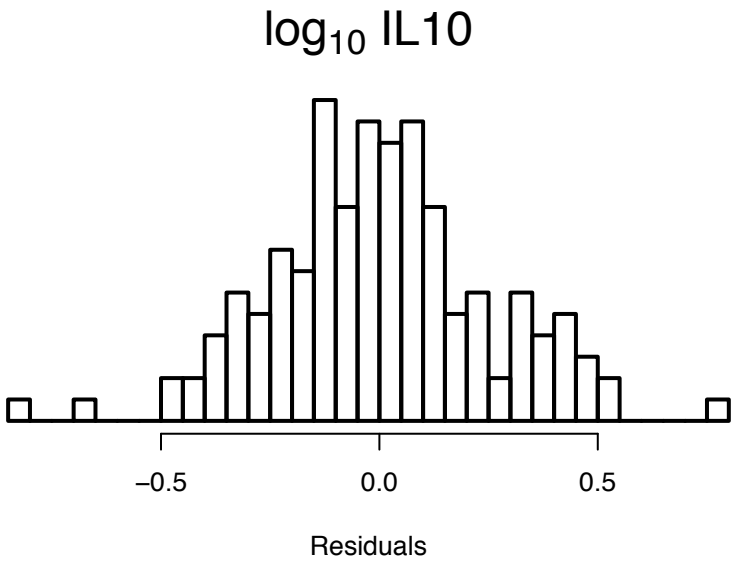
Residuals

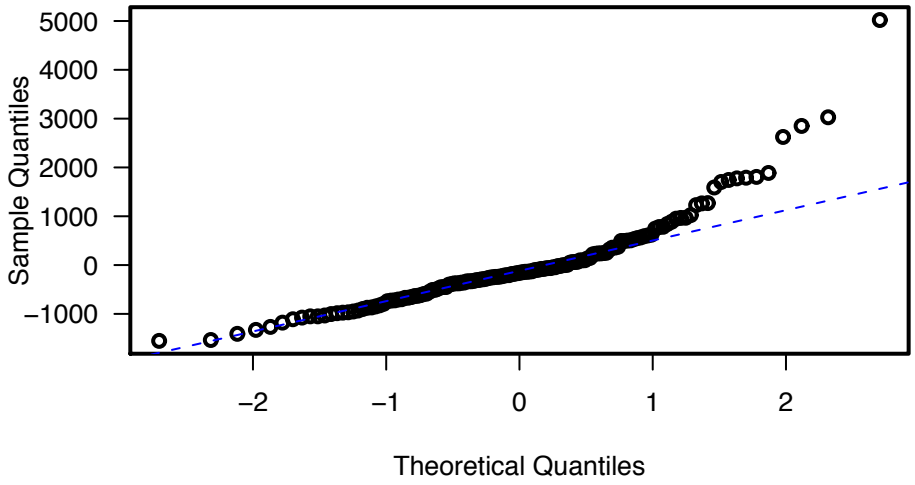

A
B6
1
2
4
5
6
7
8
10
11
12
13
14
15
17
18
19
24
25
26
Strain
−1000
0
1000
2000
3000
4000
5000
residuals (IL10)

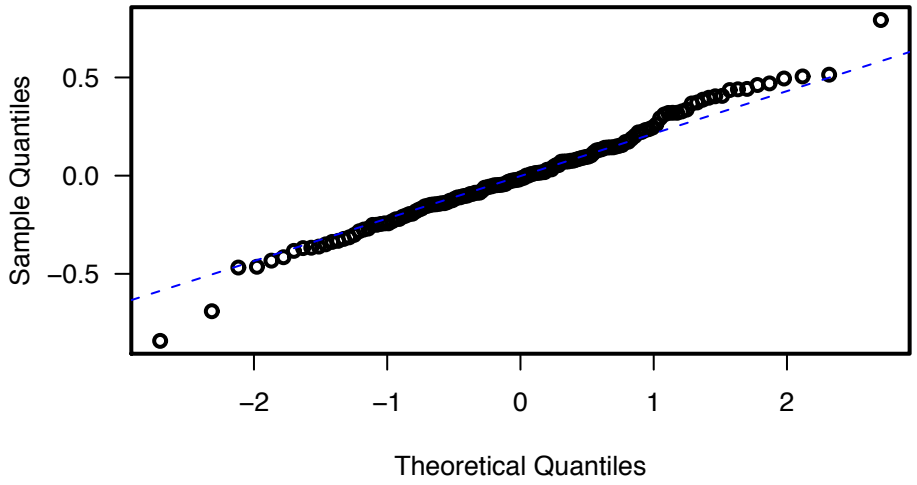

A
B6
1
2
4
5
6
7
8
10
11
12
13
14
15
17
18
19
24
25
26
Strain
−0.5
0.0
0.5
residuals (log10 IL10)

QQ plots of all residuals

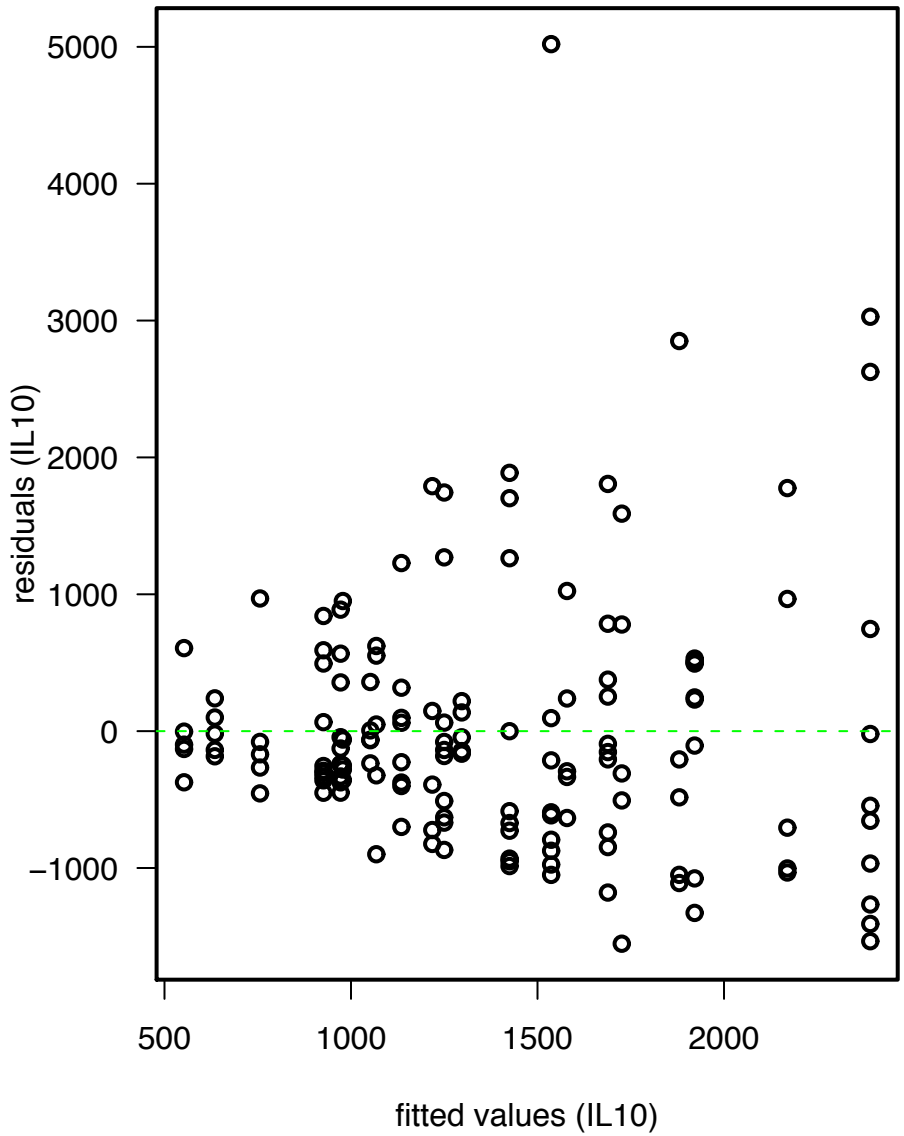


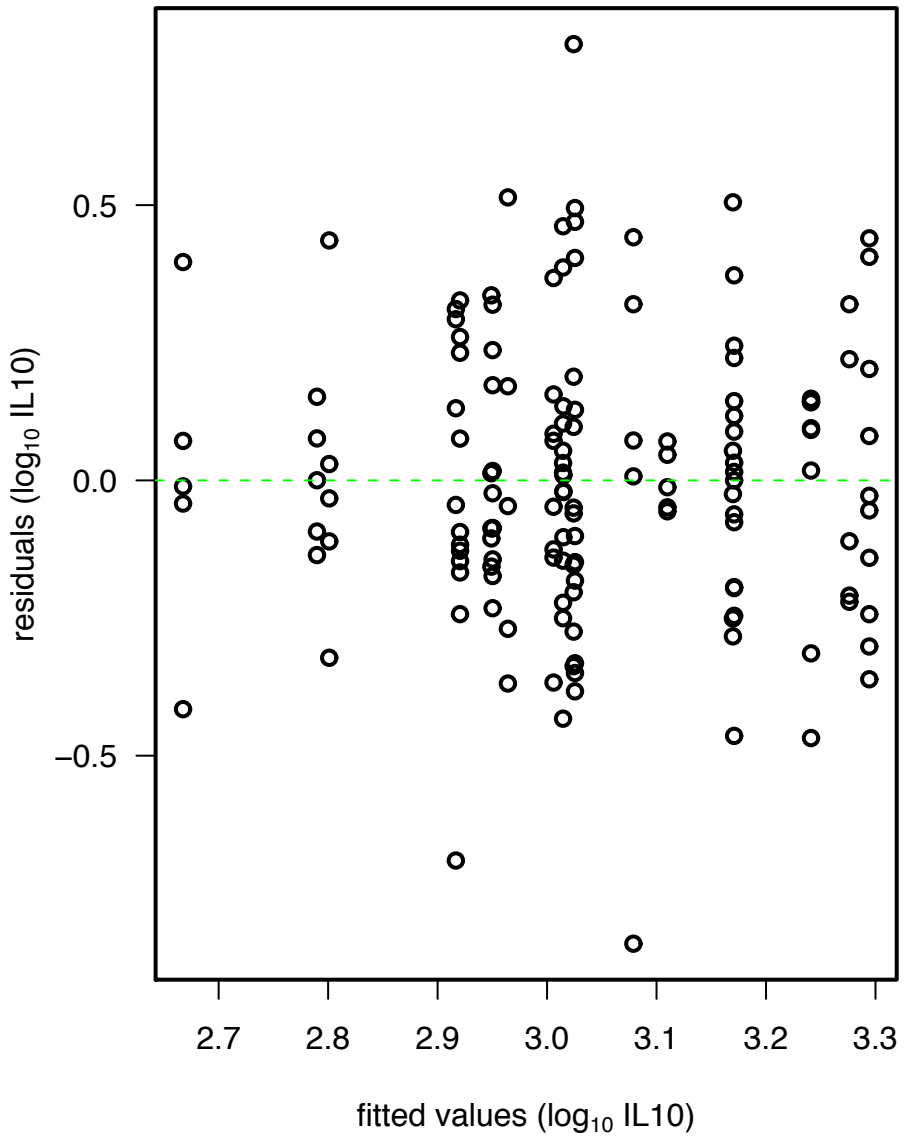




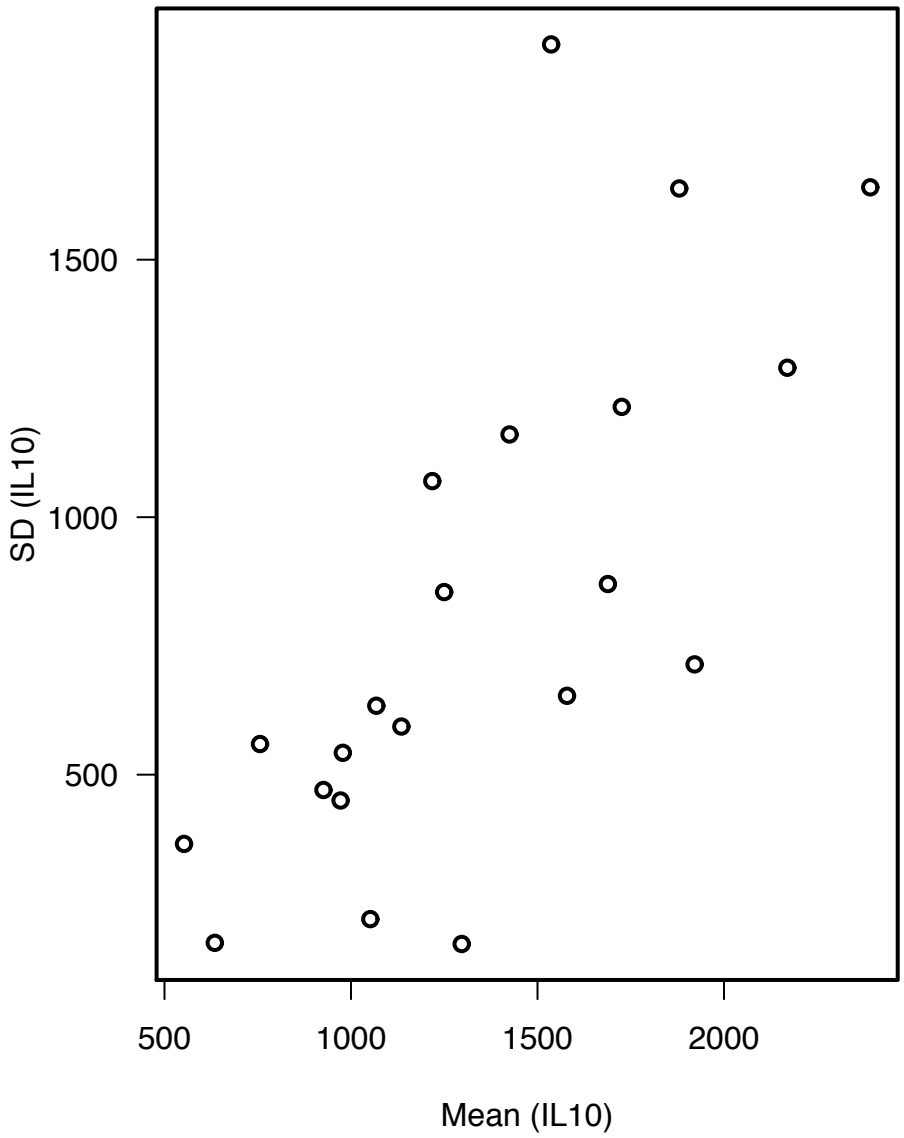


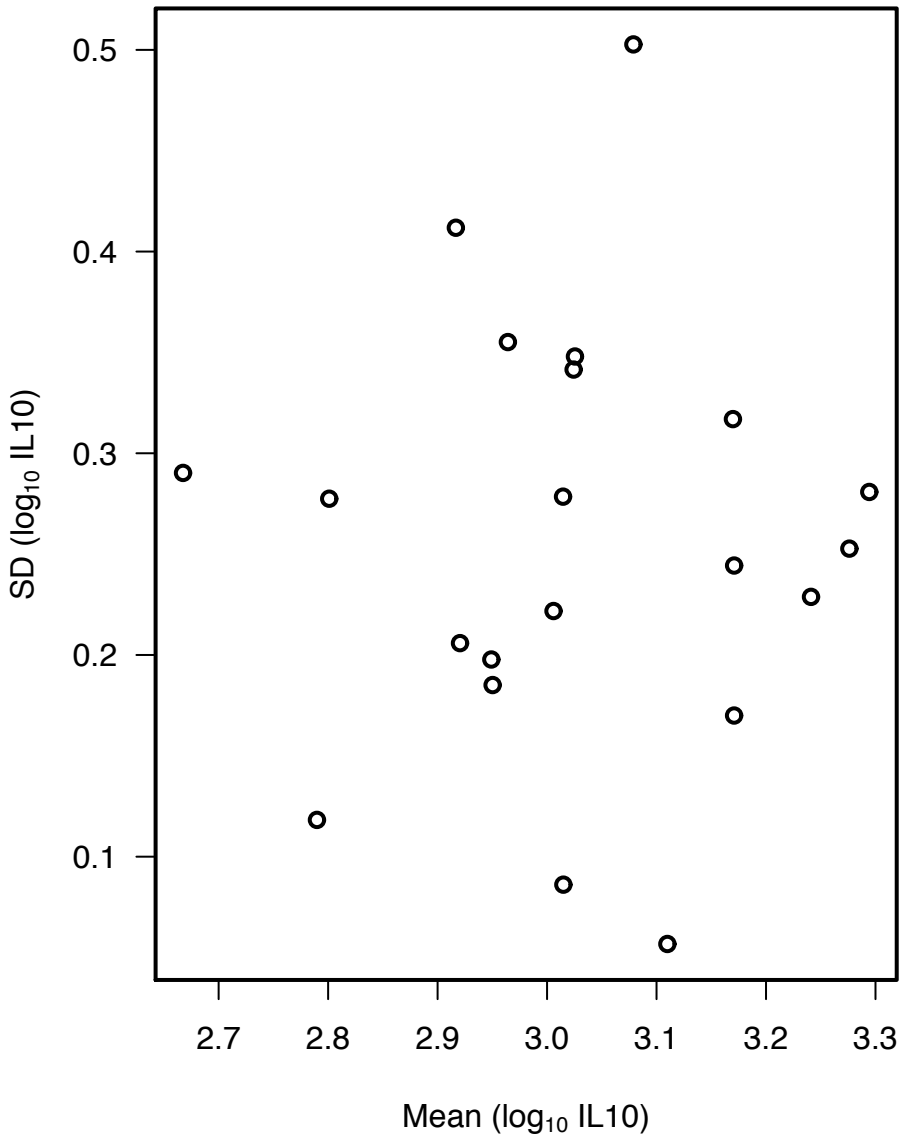
Residuals vs fitted values





SDs vs means


SD (IL10)
1500
1000
500
500
1000
1500
2000
Mean (IL10)


SD (log10 IL10)
0.5
0.4
0.3
0.2
0.1
2.7
2.8
2.9
3.0
3.1
3.2
3.3
Mean (log10 IL10)