Chapter 6
Model Assumption Checking and Remedies

Yibi Huang

THREE ASSUMPTIONS We Need to Check

In the means model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

we make 3 assumptions about the error term ε_{ij}'s.:

1. the errors ε_{ij} are **independent**, randomly distributed
2. the errors ε_{ij} have **constant variance across treatments**
3. the errors ε_{ij} follow a **normal distribution**

As the 3 assumptions are all related to errors ε_{ij}, most of the model diagnostic methods are based on the *residuals*

$$\text{residual } e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y}_{i\bullet}.$$

In the real world, data never exactly conform to these assumptions. Thankfully, the analysis in Ch3&4 work reasonably well if the reality doesn't deviate from the assumptions too much.

> "All models are wrong, but some are useful." — George P.E. Box

Standardized Residuals = Internally Studentized Residuals

- The error ε_{ij} in the means model has mean 0 and SD σ
- The SD of $e_{ij} = y_{ij} - \overline{y}_{i\bullet}$ is actually $\sigma\sqrt{1 - \frac{1}{n_i}}$., not σ.
- We hence standardize the ith raw residuals as follows, called the standardized residual or the internally Studentized residual

$$\text{standardized residual } s_{ij} = \frac{e_{ij}}{\sqrt{\text{MSE}(1 - \frac{1}{n_i})}}.$$

- If the errors ε_{ij} are normal, s_{ij} is approximately $N(0, 1)$.
- Observations with $|s_{ij}| > 3$ are potential **outliers**.

Studentized Residuals = Externally Studentized Residuals

For each externally studentized residuals, we use the MSE obtained from the model that uses all the data EXCEPT that observation, denoted as $\text{MSE}_{(ij)}$.

- Subscript "(ij)" means "all but the jth observation in the ith group".

Studentized residuals or externally studentized residuals are defined as:

$$t_{ij} = \frac{e_{ij}}{\sqrt{\text{MSE}_{(ij)}(1 - \frac{1}{n_i})}}$$

- e_{ij} is still computed using all the data but $\text{MSE}_{(ij)}$ is computed excluding the jth observation in the ith group".
- t_{ij} has a t-distribution with $N - g - 1$ d.f. but s_{ij} does not have a t-distribution.

Which Residuals Should We Use?

Internally and externally studentized residuals are related as follows

$$t_{ij} = s_{ij}\sqrt{\frac{N-g-1}{N-g-1-s_{ij}^2}}$$

If an observation is not an outlier, $t_{ij} \approx s_{ij}$. It makes little difference which one we used.
For potential outliers, it's **better using the externally studentized residuals**.

Example: Hodgkin's Disease

Hodgkin's disease is a type of lymphoma, which is a cancer originating from white blood cells called lymphocytes.

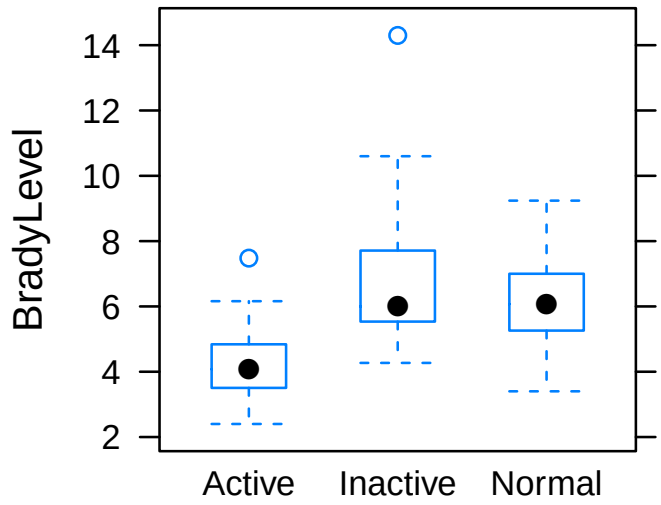
The data file `Hodgkins.txt` contains plasma bradykininogen levels (in μg of bradykininogen per ml of plasma) in 3 types of subjects

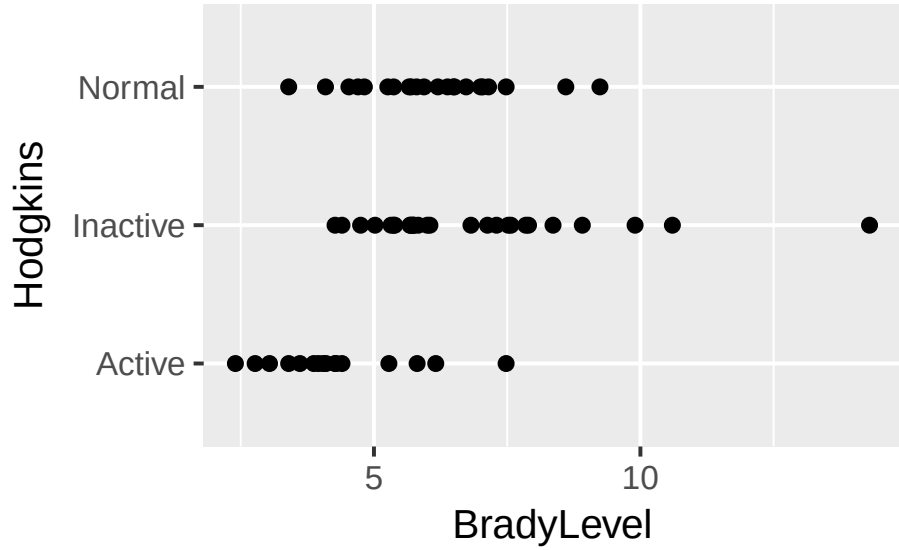
- normal,
- in patients with active Hodgkin's disease, and
- in patients with inactive Hodgkin's disease.

Response: The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

- Is this an experiment?
- We can use ANOVA to compare means of several samples in an observational study.

```
> hodgkins = read.table("Hodgkins.txt", header=T)
> library(hodgkins)
> bwplot(BradyLevel~Hodgkins, data=hodgkins)
> qplot(BradyLevel, Hodgkins, data=hodgkins)
```





The distribution within each group looks right skewed. Let's fit the ANOVA model anyway and take a look at the residuals.

```
> brady1 = lm(BradyLevel ~ Hodgkins, data=hodgkins)
> anova(brady1)
          Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins   2  65.893  32.946   10.67 0.0001042 ***
Residuals 62 191.449   3.088
```

Residuals

$$\text{Residual } e_{ij} = y_{ij} - \overline{y}_{i\bullet}$$

```
> round(brady1$res,2)
    1     2     3     4     5     6     7     8     9    10    11
-0.72 -0.29 -1.40 -0.40 -2.69  2.50  1.38  1.06  0.40 -2.01 -0.15
   12    13    14    15    16    17    18    19    20    21    22
 0.28  3.15 -0.44 -1.57  0.41  0.90  0.10  0.94 -1.28  0.64 -0.84
   23    24    25    26    27    28    29    30    31    32    33
-0.35 -1.27  0.97 -0.91 -0.21 -0.70  1.85  3.17 -0.44 -0.04 -0.26
   34    35    36    37    38    39    40    41    42    43    44
-1.91  1.50 -0.02 -1.54  0.09 -1.49  3.74 -1.84  7.44  3.04 -2.59
   45    46    47    48    49    50    51    52    53    54    55
-1.11 -1.12  0.99 -0.04  1.04  1.50 -1.14 -0.86 -2.11 -1.03  0.44
   56    57    58    59    60    61    62    63    64    65
 0.66 -1.54 -0.81 -1.18  0.71 -1.18  2.05 -1.47 -2.46  0.27
```

Observation #42 has the largest residual 7.44 (NOT standardized). Is it an outlier?

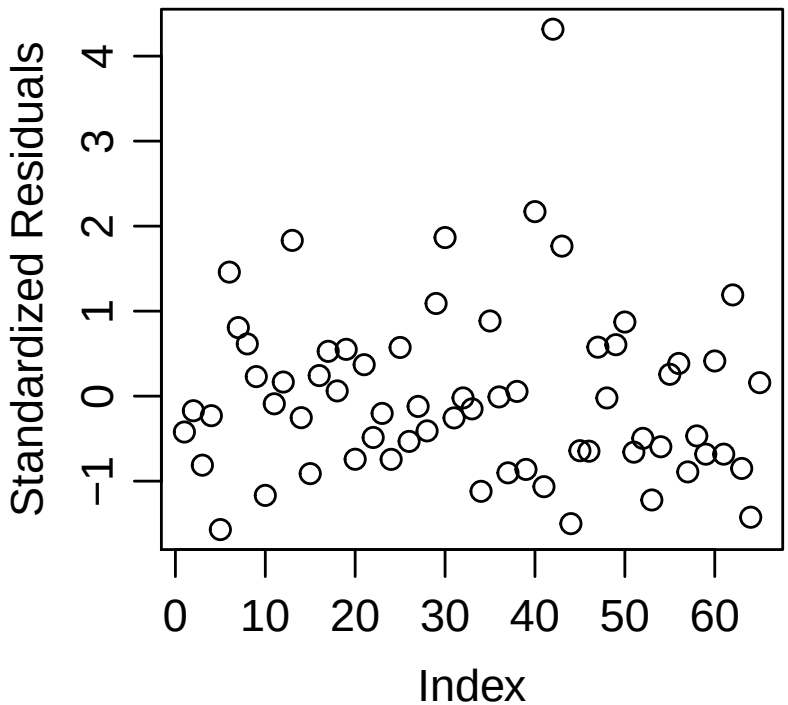
Standardized and Studentized Residuals in R

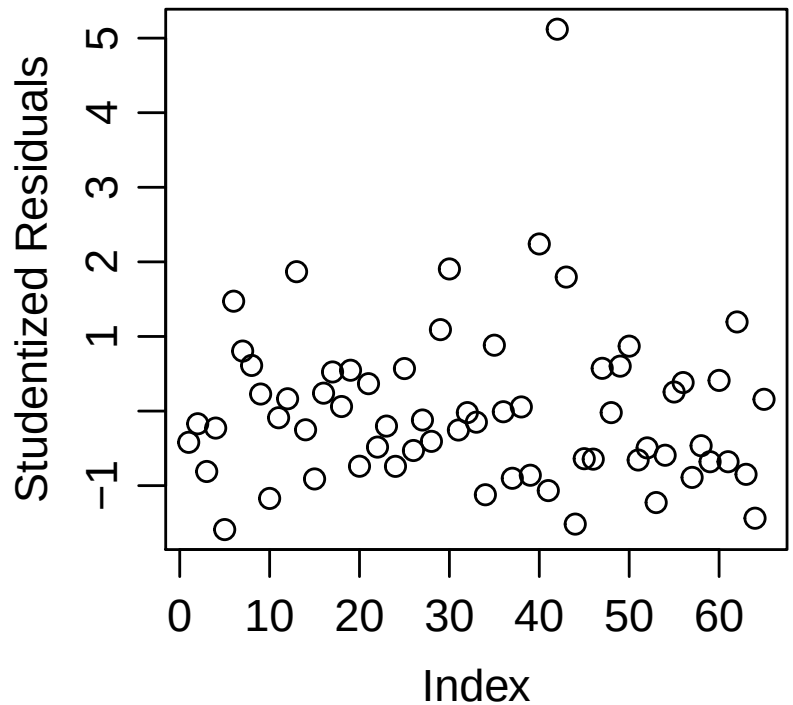
The `rstandard()` and `rstudent()` command can produce the standardized and studentized residuals in R

```
> rstandard(brady1)          # Standardized residuals
         1          2          3          4          5          6
-0.4222885 -0.1718277 -0.8125414 -0.2300744 -1.5697484  1.4590797
(... some output is omitted ...)
        61         62         63         64         65
-0.6823674  1.1907600 -0.8505429 -1.4246593  0.1585103

> rstudent(brady1)          # (externally) studentized residual
         1          2          3          4          5          6
-0.4194728 -0.1704770 -0.8102878 -0.2283089 -1.5889328  1.4727714
(... some output is omitted ...)
        61         62         63         64         65
-0.6793980  1.1948600 -0.8486212 -1.4368375  0.1572586
```

```
> plot(rstandard(brady1), ylab="Standardized Residuals")
> plot(rstudent(brady1), ylab="Studentized Residuals")
```





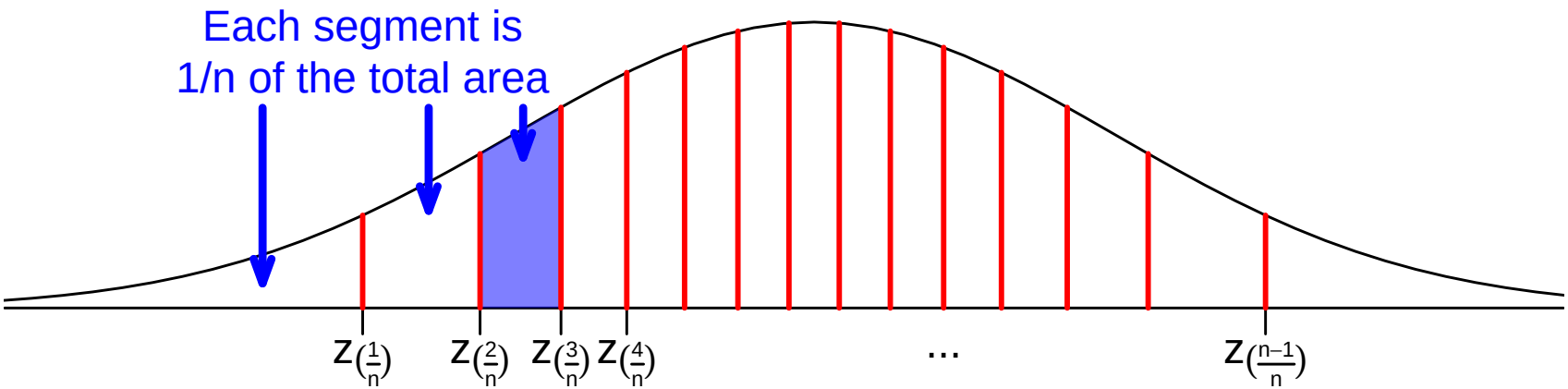
There is a potential outlier with a studentized residual > 5.

3.4.1 How to Check the Normality Assumption

- Histogram of the residuals: if normal, should be bell-shaped
  - Pros: simple, easy to understand
  - Cons: for a small sample, histogram may not be bell-shaped even though the sample is from a normal distribution
- Normal probability plot of the residuals
  - aka. normal QQ plot,
    QQ stands for "quantile-quantile"
  - the best method to assess normality
  - See next slide for details

Ideas Behind the Normal Probability Plot (1)

- Data $y_1, y_2, \ldots, y_n$
- Sorted Data: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, call the **Sample Quantiles**
- **Theoretical Quantiles** of the N(0, 1): $z_{(\frac{1}{n})}$, $z_{(\frac{2}{n})}$, $\ldots$, $z_{(\frac{n-1}{n})}$, where, $z_{(\frac{k}{n})}$ is a value such that $\mathrm{P}(Z \leq z_{(\frac{k}{n})}) = k/n$ for $Z \sim N(0, 1)$.



Ideas Behind the Normal Probability Plot (2)

- If $Y \sim N(\mu, \sigma^2)$, then

$$P(Y \leq \mu + \sigma z_{(\frac{k}{n})}) = P\left(\underbrace{\frac{Y-\mu}{\sigma}}_{\sim N(0,1)} \leq z_{(\frac{k}{n})} \right) = k/n$$

  We expected k/n of the observations to be $\leq \mu + \sigma z_{(\frac{k}{n})}$

- We observe k/n of the observations are $\leq y_{(k)}$.
- If the data are indeed $N(\mu, \sigma^2)$, we expect

$$y_{(k)} \approx \mu + \sigma z_{(\frac{k}{n})}$$

- If one plots the Sample Quantiles $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ against the Theoretical Quantiles $z_{(\frac{1}{n})}$, $z_{(\frac{2}{n})}$, $\ldots$, $z_{(\frac{n-1}{n})}$, the points would fall on the straight line

$$y = \mu + \sigma z.$$

  if the data follow $N(\mu, \sigma^2)$

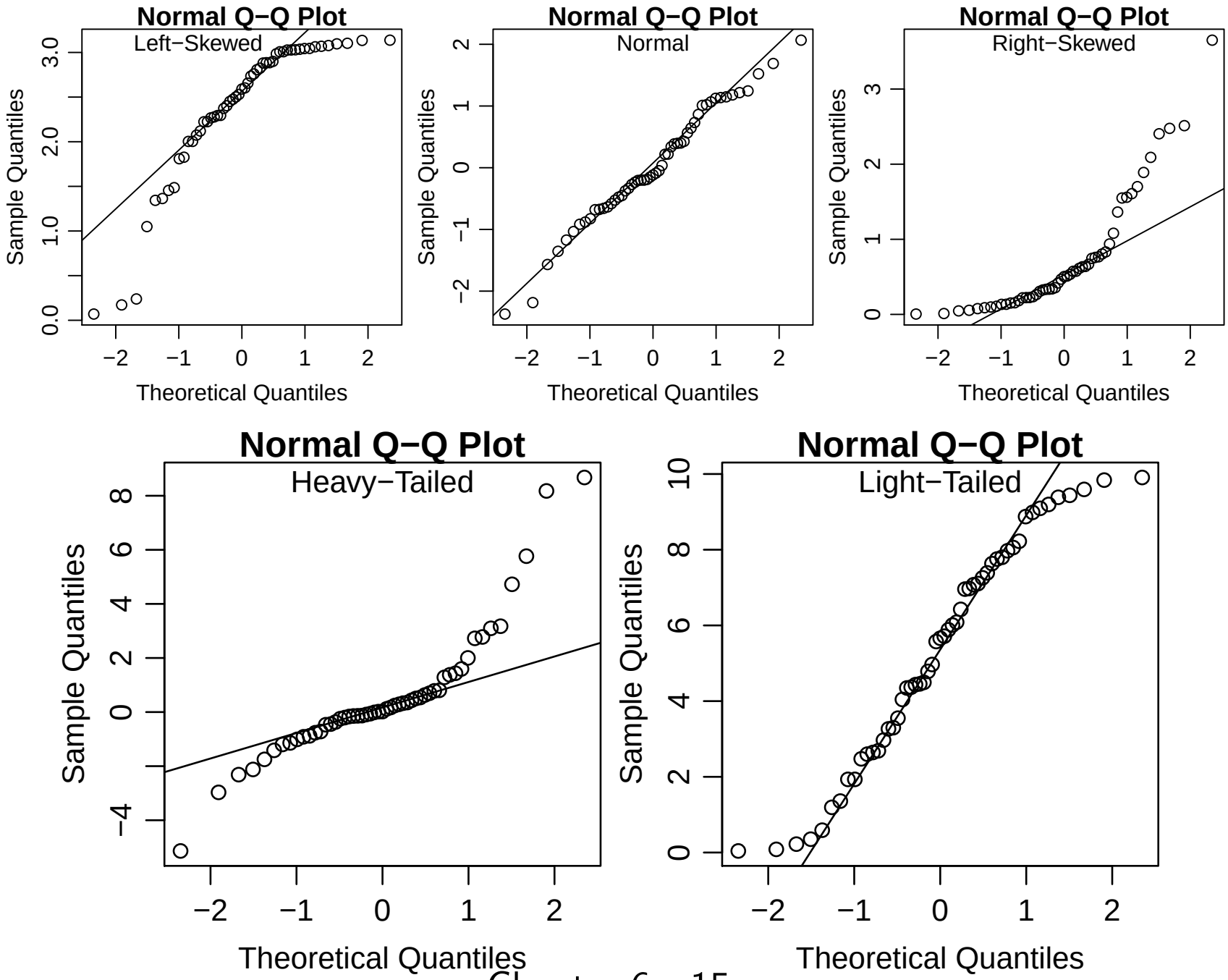
A Technical Remark

As $z_{(n/n)} = \infty$, R actually uses the Theoretical Quantiles:

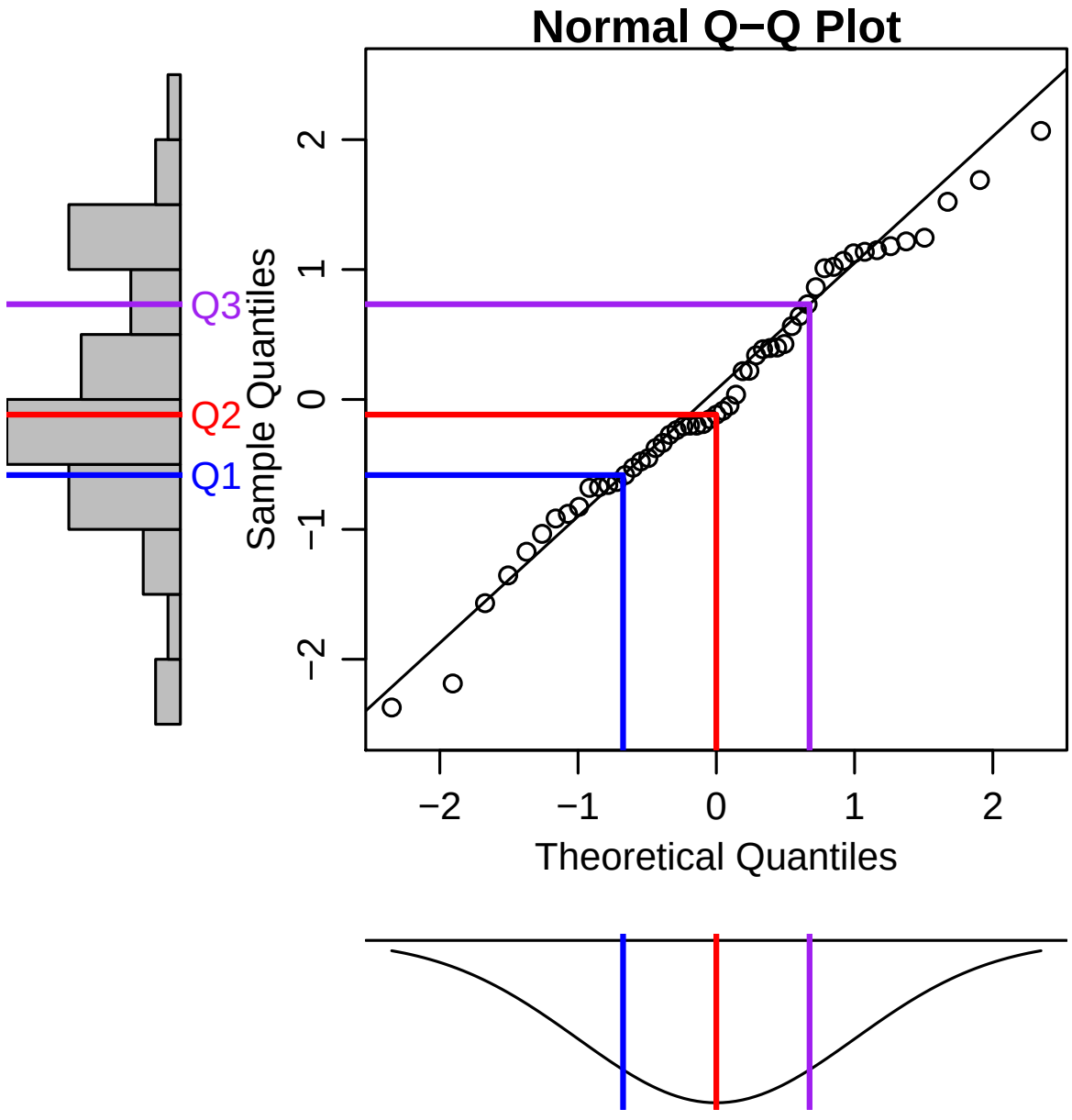
$$z_{(\frac{1-0.5}{n})}, \quad z_{(\frac{2-0.5}{n})}, \quad z_{(\frac{3-0.5}{n})}, \ \cdots, \ z_{(\frac{n-0.5}{n})}$$

instead of

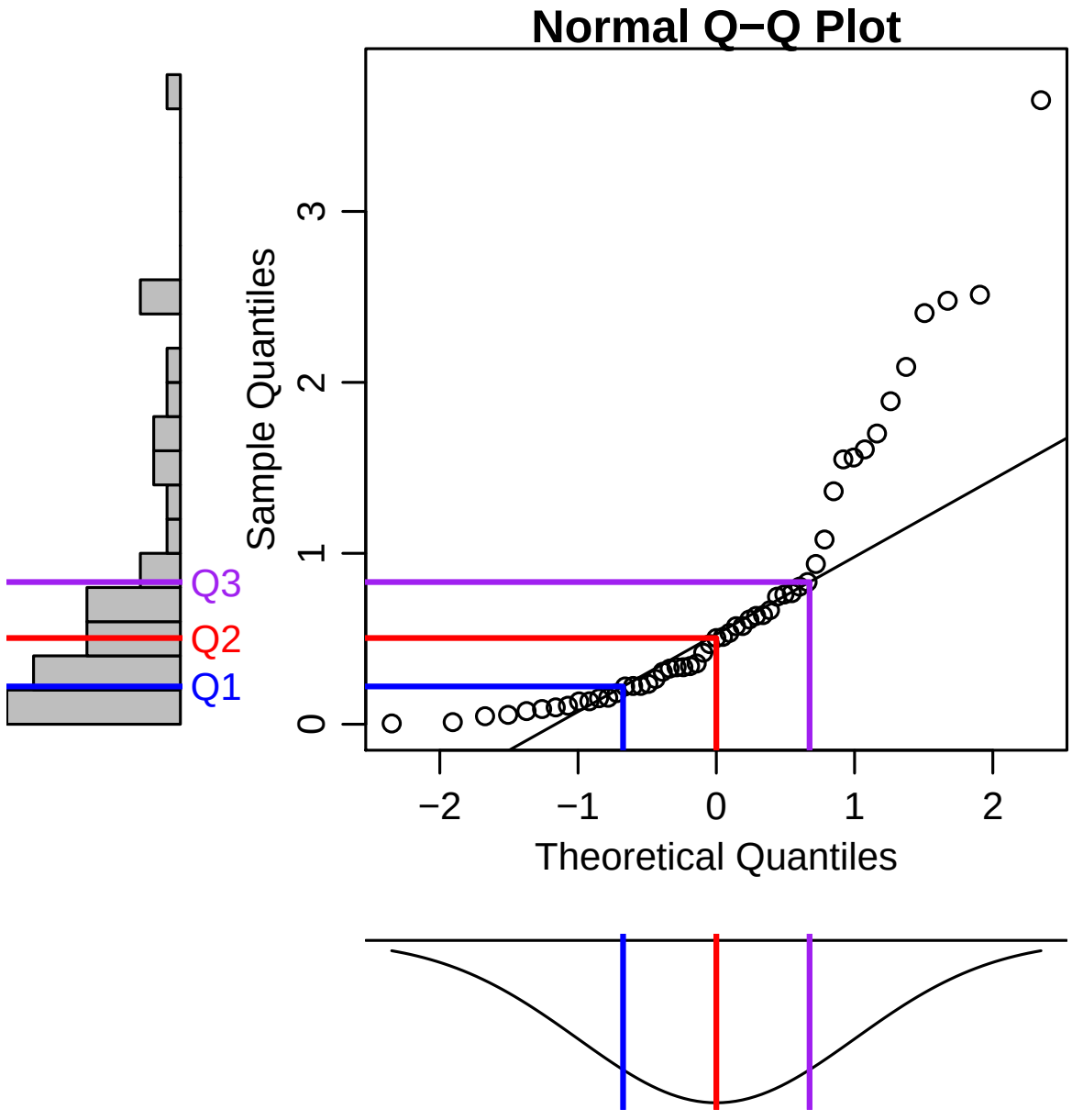
$$z_{(\frac{1}{n})}, \quad z_{(\frac{2}{n})}, \cdots, z_{(\frac{n-1}{n})}, z_{(\frac{n}{n})}.$$


Normal Q−Q Plot
Left−Skewed
Normal Q−Q Plot
Normal
Normal Q−Q Plot
Right−Skewed
Normal Q−Q Plot
Heavy−Tailed
Normal Q−Q Plot
Light−Tailed
Sample Quantiles
Theoretical Quantiles

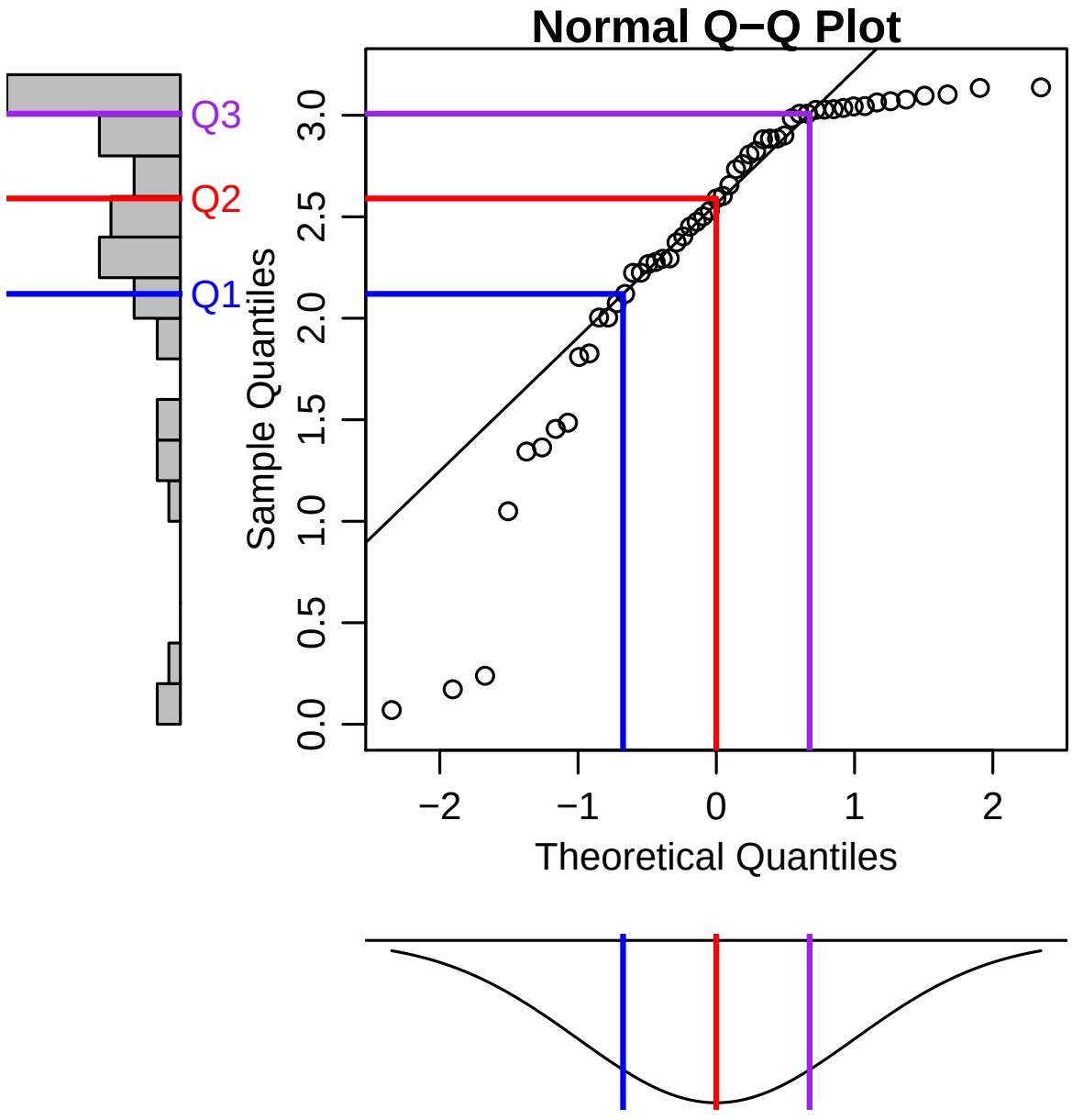
Normal QQ Plot — Normal Data



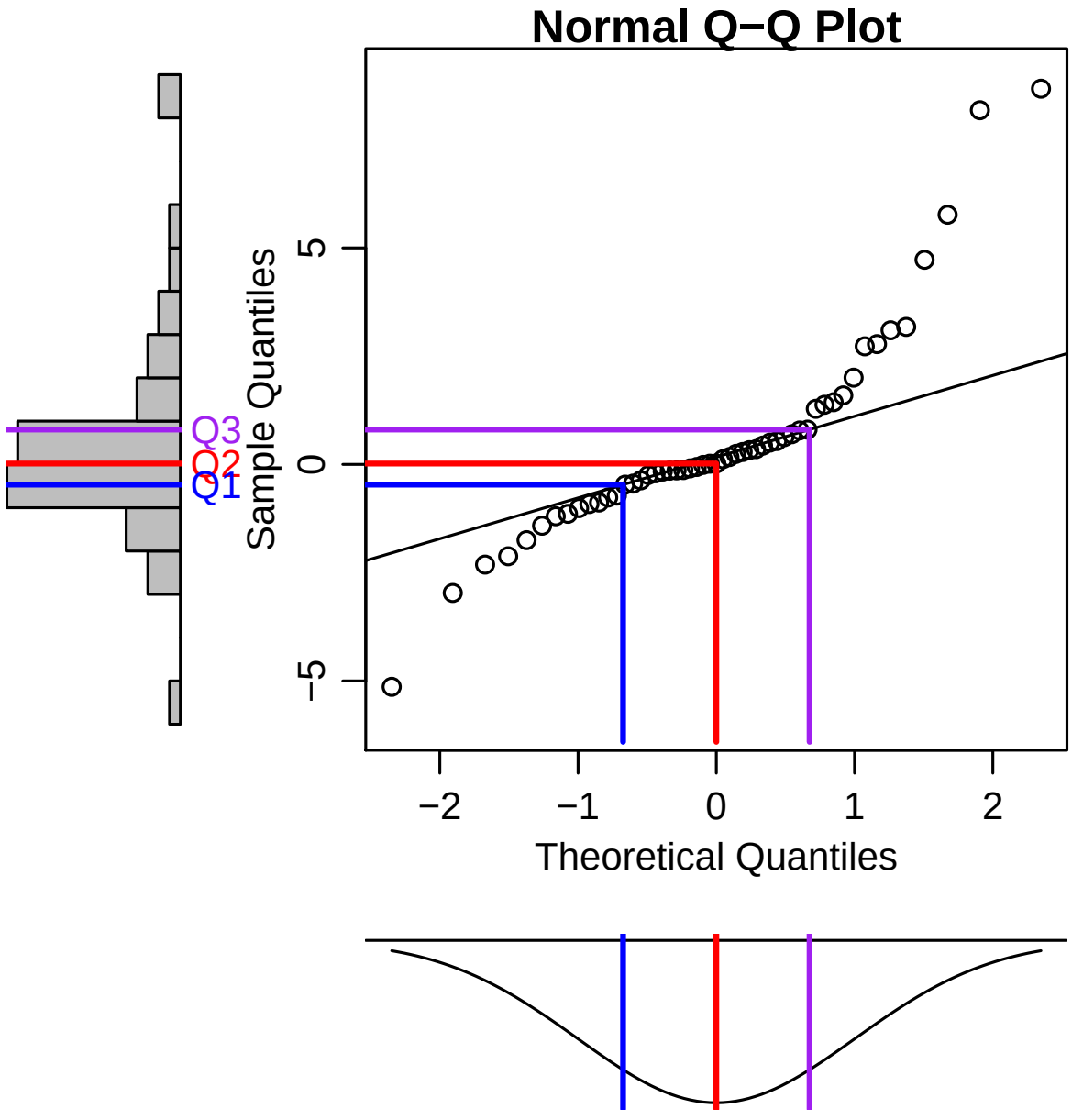
Normal QQ Plot — Right-Skewed Data



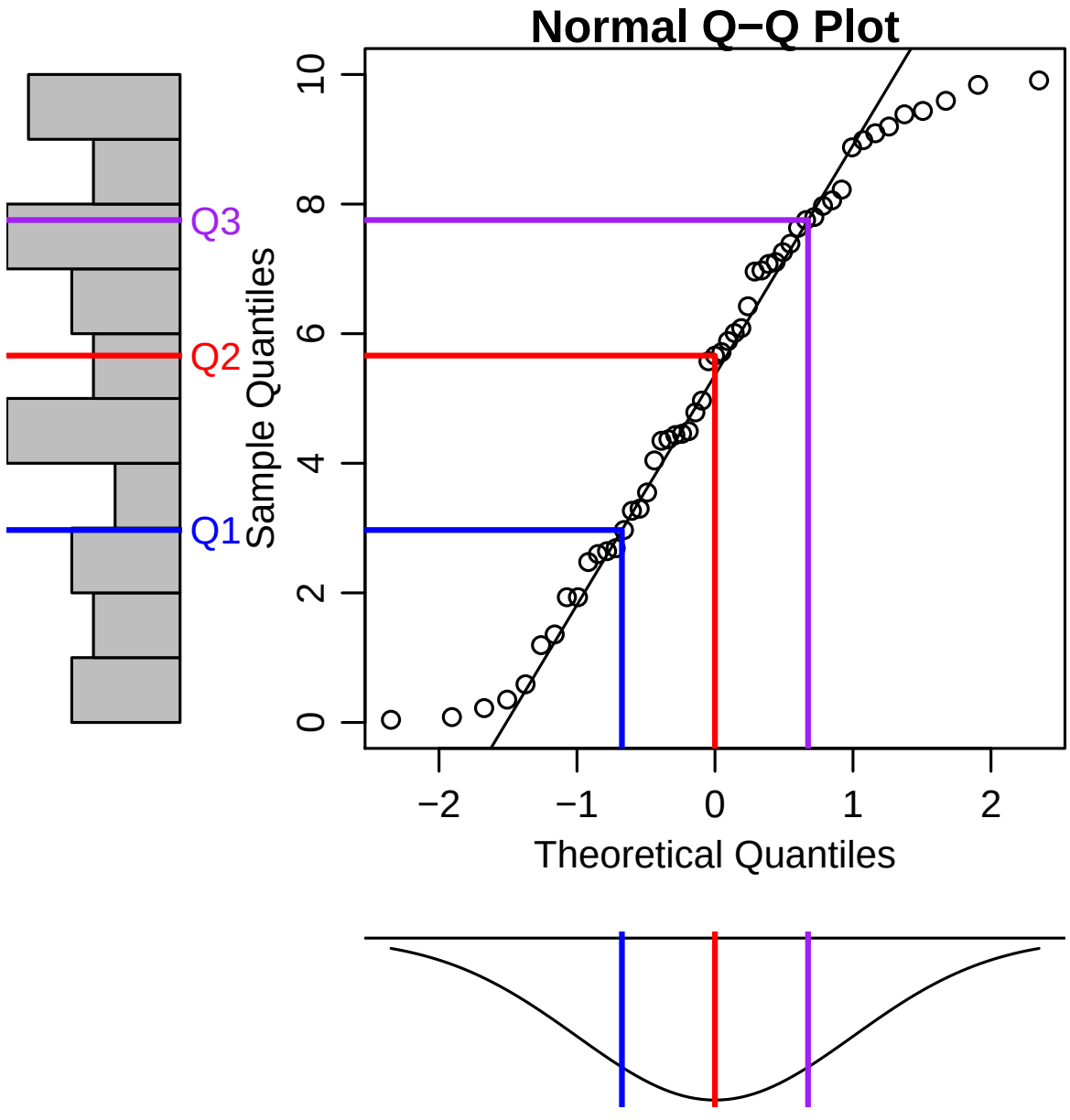
Normal QQ Plot — Left-Skewed Data



Normal QQ Plot — Heavy-Tailed Data



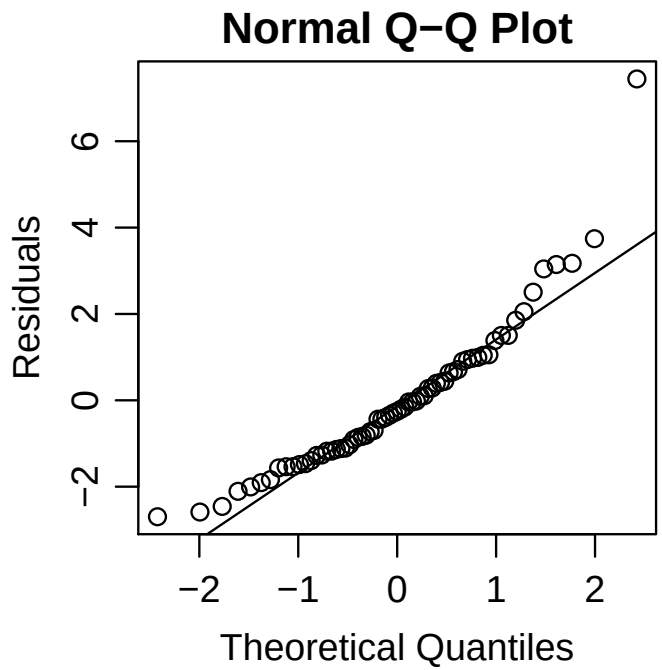
Normal QQ Plot — Light-Tailed Data

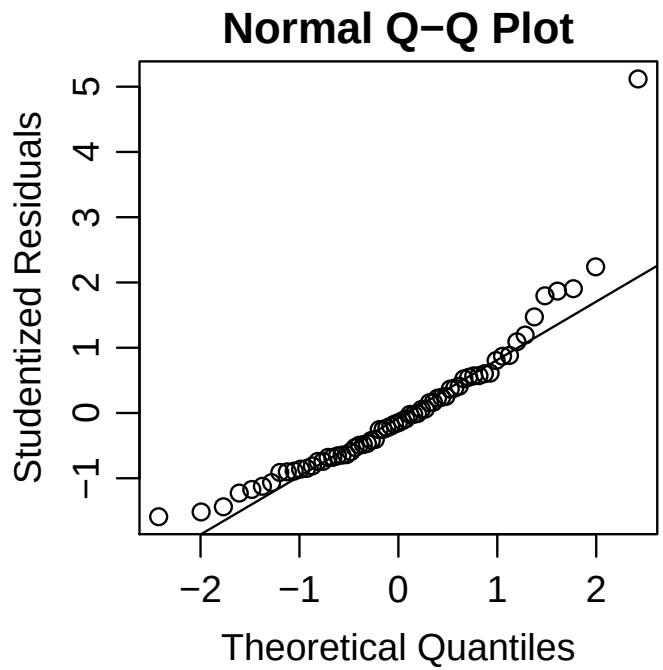


Normal QQ Plot for The Hodgkin Data

```
qqnorm(brady1$res, ylab="Residuals")
qqline(brady1$res)

library(MASS)
qqnorm(rstudent(brady1), ylab="Studentized Residuals")
qqline(rstudent(brady1))
```





Does the distribution of the residuals look normal?

Somewhat right-skewed, a potential outlier with $s_{ij} > 5$.

Remedies for Non-Normality

Skewness can often be ameliorated by **transforming the response** — often a power transformation.

$$f_\lambda(y) = \begin{cases} y^\lambda, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0. \end{cases}$$

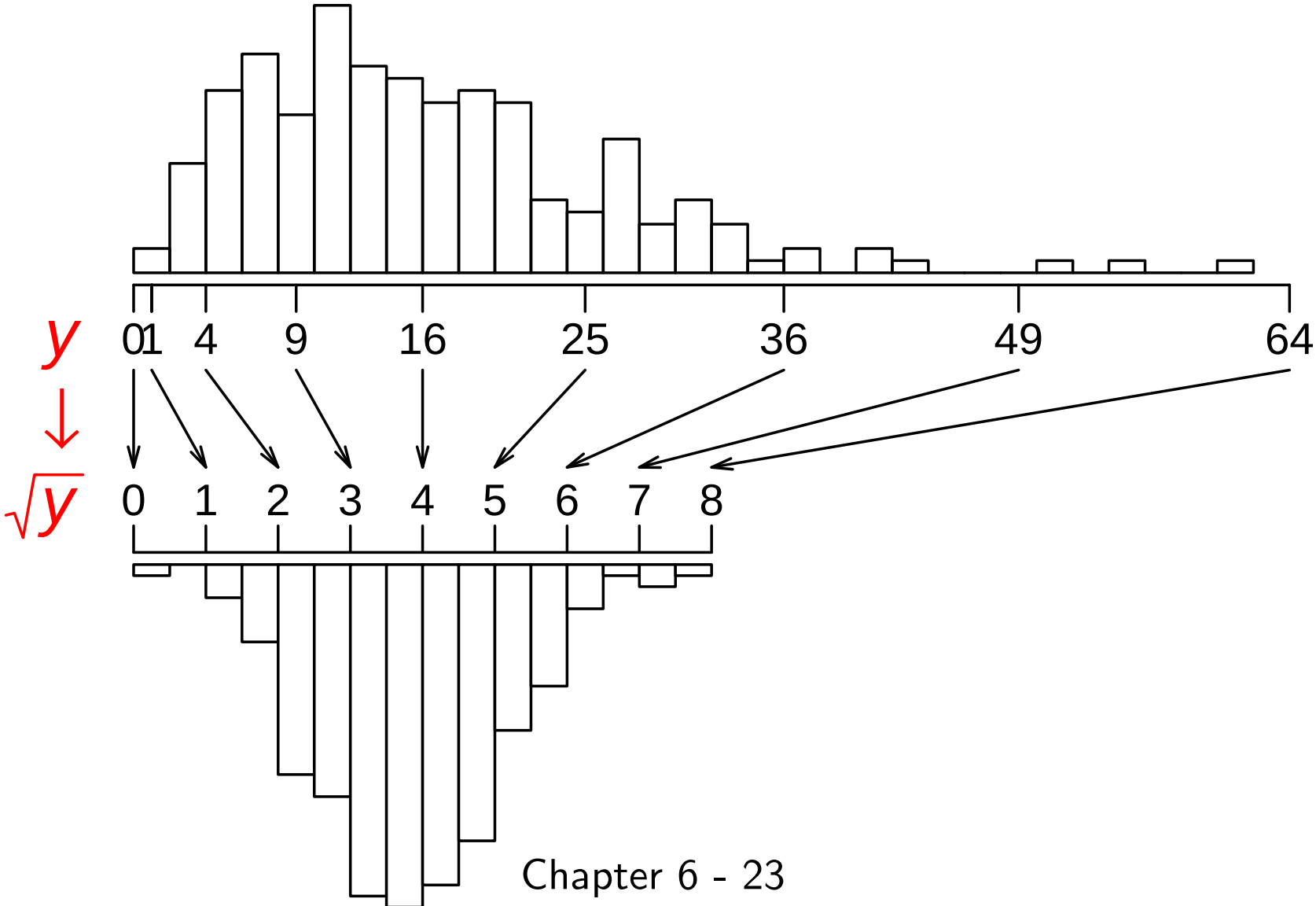
1. If **right-skewned**, try taking square root, logarithm, or other powers $\lambda < 1$

$$y \longrightarrow 1/y, \log(y), \sqrt{y}, \text{or } y^\lambda \text{ with } \lambda < 1$$

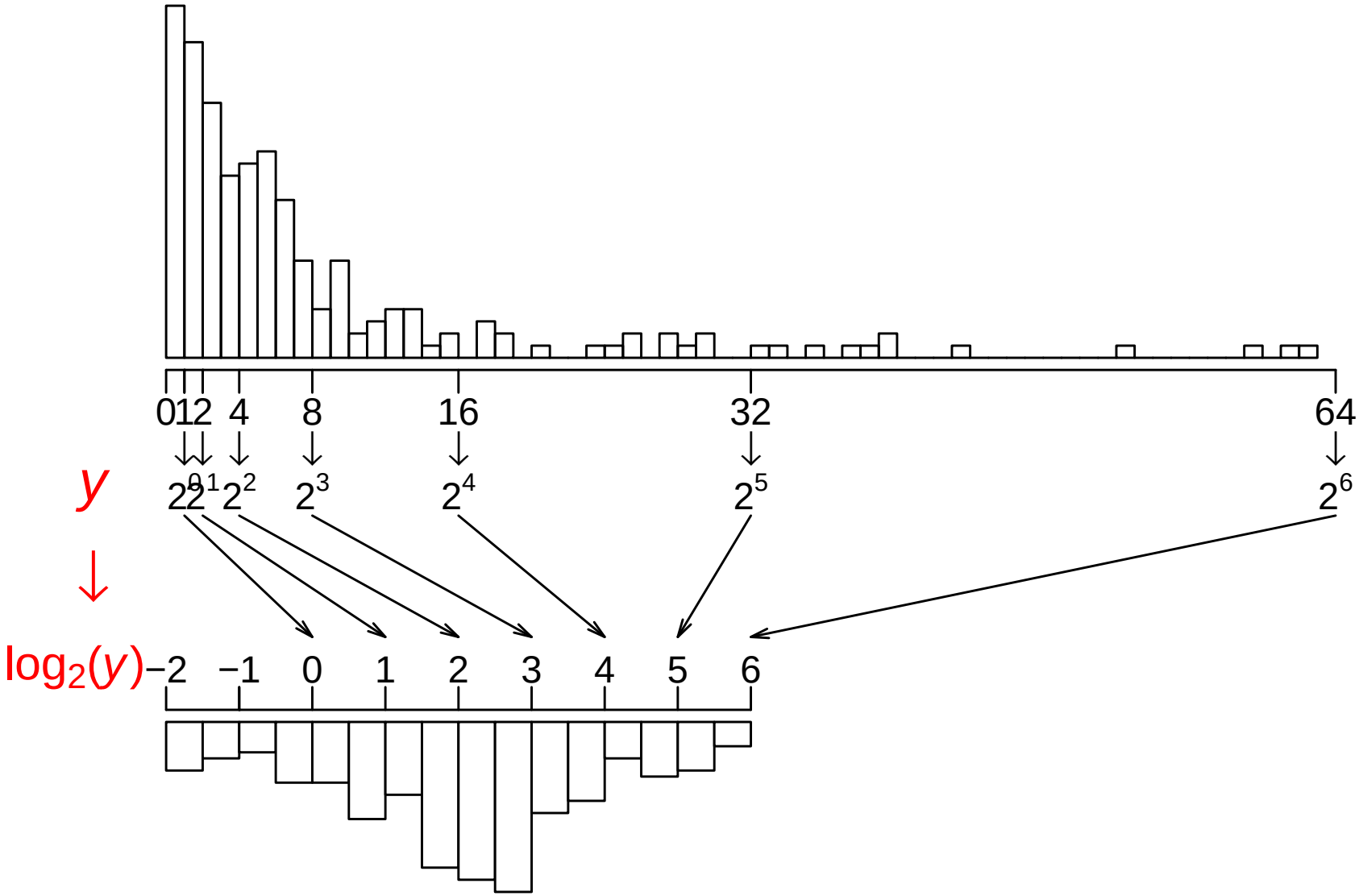
2. If **left-skewned**, try squaring, cubing, or other powers $\lambda > 1$

$$y \longrightarrow y^2, y^3, \text{or } y^\lambda \text{ with } \lambda > 1$$

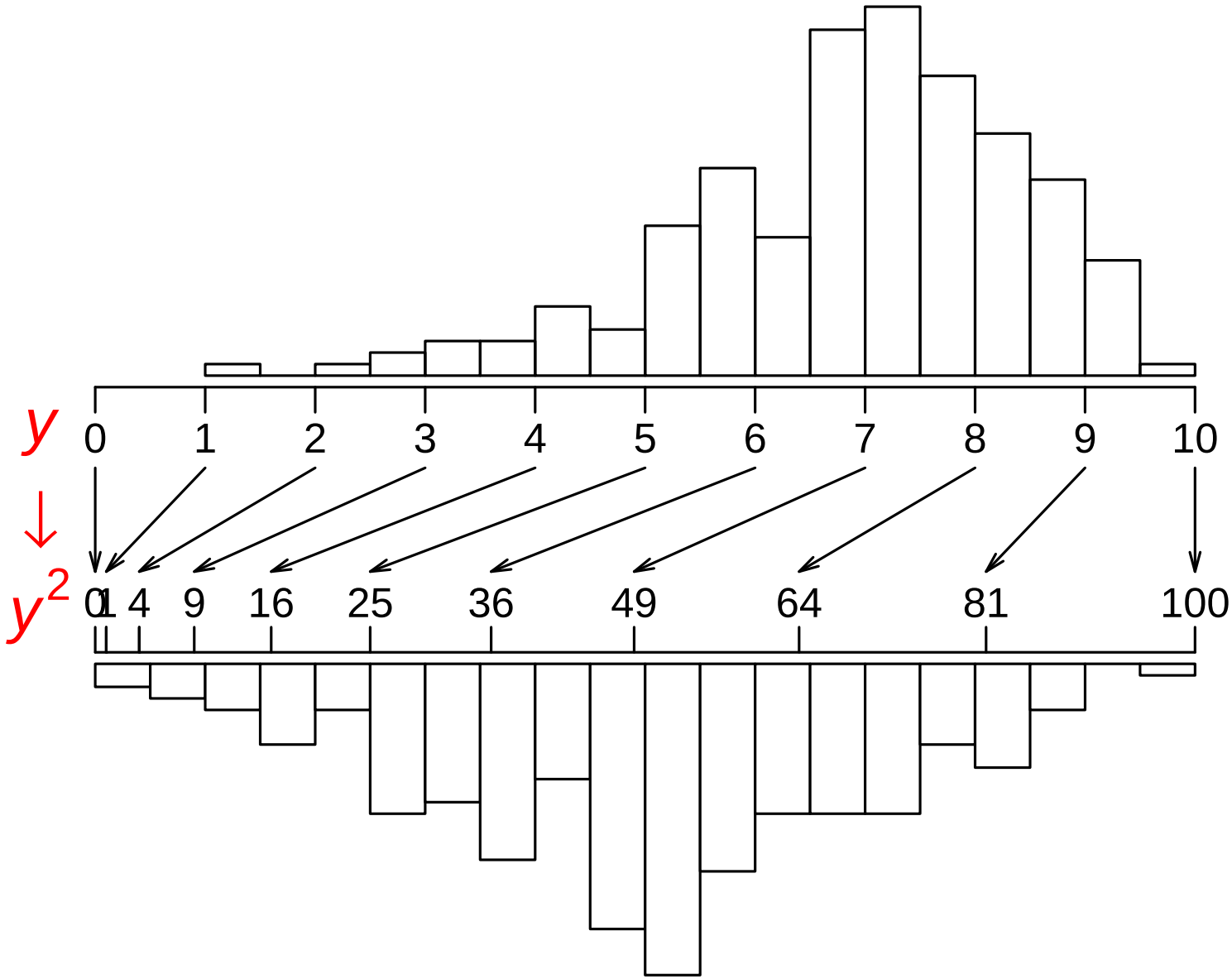
Square-Root Transformation Shrinks the Upper Tail and Extends the Lower Tail, and Hence Reduces Right-Skewness



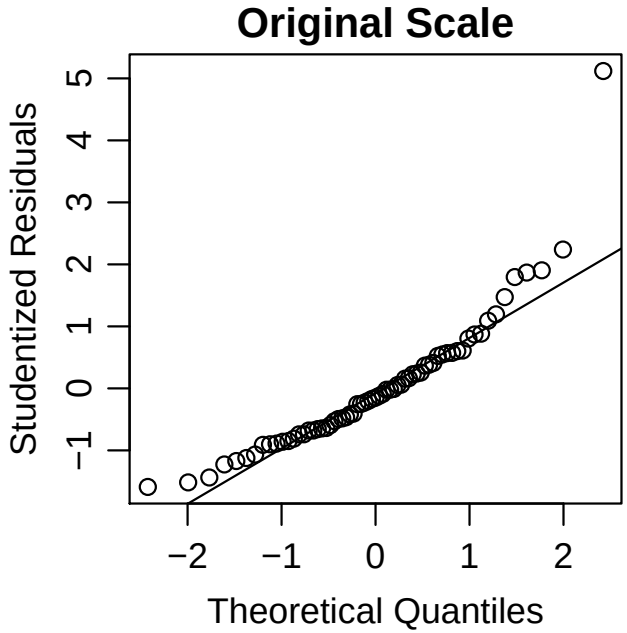
Log Transformation Shrinks the Upper Tail and Extends the Lower Tail Even More!



$y \rightarrow y^2$ Extends the Upper Tail and Shrinks the Lower Tail, and Hence Reduces Left-Skewness

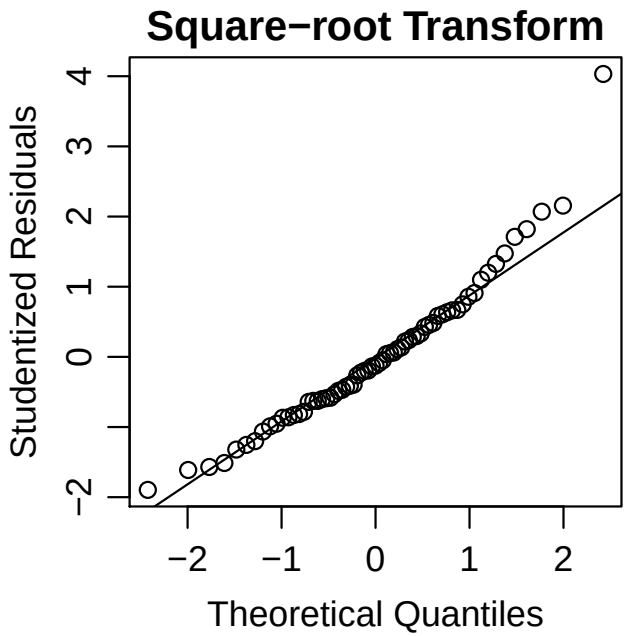


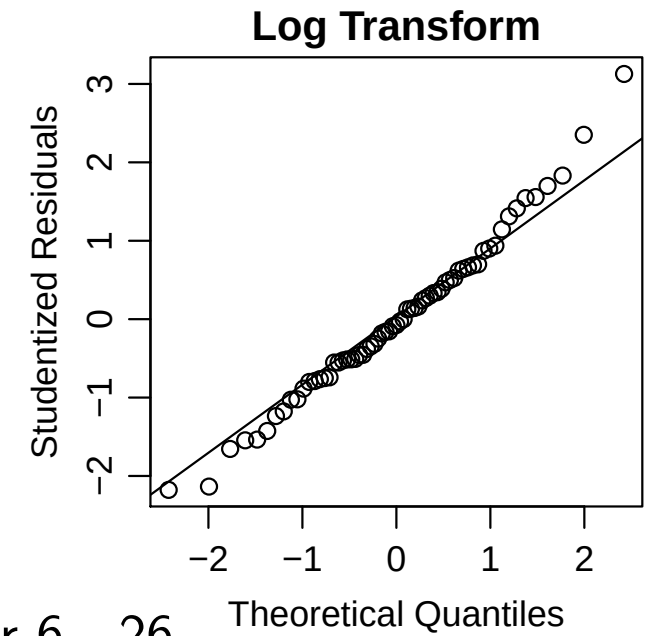
Example: Hodgkin's Disease – QQ Plots



Log-transformation makes residuals less R-skewed, and the outlier less extreme.

The square-root transformation also reduces R-skewness but not as much as the log transformation.





R-Codes for Making the Plots on the Previous Slide

```
brady1 = lm(BradyLevel ~ Hodgkins, data=hodgkins)
brady2 = lm(sqrt(BradyLevel) ~ Hodgkins, data=hodgkins)
brady3 = lm(log(BradyLevel) ~ Hodgkins, data=hodgkins)

library(MASS)

qqnorm(rstudent(brady1), main="Original Scale")
qqline(rstudent(brady1))

qqnorm(rstudent(brady2), main="Square-root Transform")
qqline(rstudent(brady2))

qqnorm(rstudent(brady3), main="Log Transform")
qqline(rstudent(brady3))
```

Box-Cox Method

Box-Cox method is an automatic procedure to select the "best" power λ that make the residuals of the model

$$y_{ij}^{\lambda} = \mu_i + \varepsilon_{ij}$$

closest to normal.

- We usually round the optimal λ to a convenient power like

$$-1,\ -\frac{1}{2},\ -\frac{1}{3},\ 0,\ \frac{1}{3},\ \frac{1}{2},\ 1,\ 2, \ldots$$

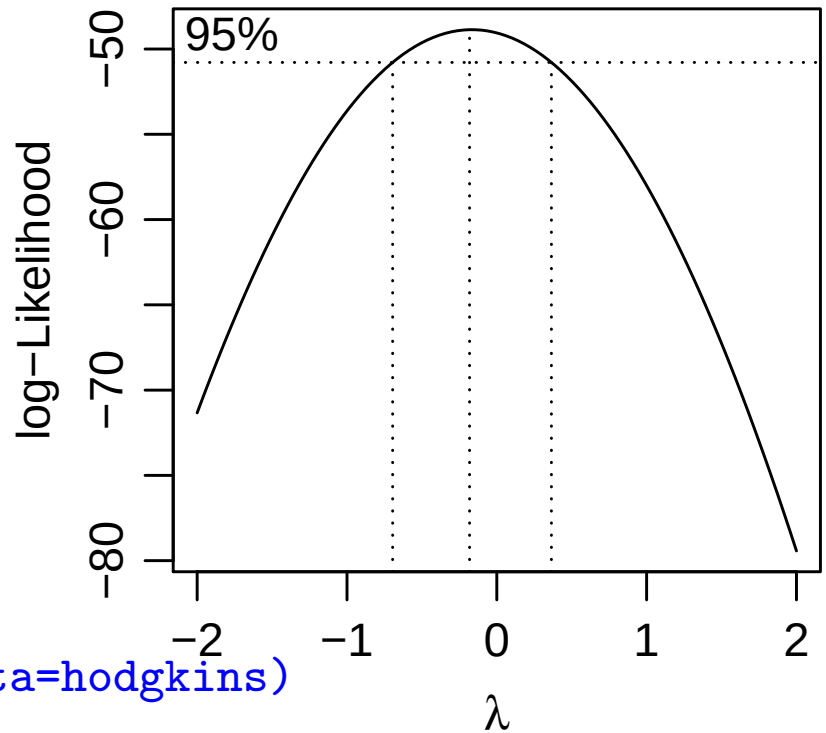
since the practical difference of $y^{0.5827}$ and $y^{0.5}$ is usually small, but the square-root transformation is much easier to interpret.

- A confidence interval for the optimal λ can also be obtained. See Oehlert, p.129 for details.
We usually select a convenient power λ^* in this C.I.

Example: Hodgkin's Disease – Box-Cox

In R, one must first load the `MASS` library to use `boxcox()`.

The argument of the `boxcox()` function can be a model formula, an `lm` model.




```
library(MASS)
boxcox(brady1)
boxcox(BradyLevel ~ Hodgkins, data=hodgkins)
```

The middle dash line marks the optimal λ, the right and left dash line mark the 95% C.I. for the optimal λ.

For the plot, we see the optimal λ is around -0.2, and the 95% C.I. contains 0. For simplicity, we use the log-transformed `BradyLevel` as our response.

Example: Hodgkin's Disease

From the 2 ANOVA tables below, we see the differences of the 3 group of patients become more significant after a log transformation, because the outlier become less extreme and do not inflate the MSE as much.

```
Response: BradyLevel
           Df  Sum Sq Mean Sq F value    Pr(>F)
Hodgkins    2  65.893  32.946   10.67 0.0001042 ***
Residuals  62 191.449   3.088

Response: log(BradyLevel)
           Df Sum Sq Mean Sq F value    Pr(>F)
Hodgkins    2 2.2526 1.12631  15.436 3.628e-06 ***
Residuals  62 4.5238 0.07297
```

Log-Scale is Commonly Used for Concentrations

Log-scale is also commonly used for variables that are about money.h

In fact, for measurements of concentration, the log scale is more commonly used than the original scale. For example,

- The concentrations 10.1 and 10.001 are nearly indistinguishable
- However, there is a huge difference between concentration of 0.1 and 0.001 since 0.1 is 100 times higher than 0.001.
- In the original scale, (10.1, 10.001) and (0.1, 0.001) differ by the same amount, 0.099.
- In log scale,

$$\log_{10} 0.1 - \log_{10} 0.001 = 2 \quad \text{far greater than}$$
$$\log_{10} 10.1 - \log_{10} 10.001 \approx 0.0043$$

Non-Parametric Tests

- Transformation does not always fix non-normality.
  For example, it helps little for symmetric but heavy-tailed (many outliers) distributions.

- The ANOVA F-test is *robust* to non-normality, but it is not resistent to *outliers*.
- If outliers are unavoidable, and cannot be removed, try non-parametric tests, like permutation test in Chapter 2, and Kruskal-Wallis test in Section 3.11.1, that doesn't rely on normality assumption, which we will introduce soon after we finish Chapter 6.

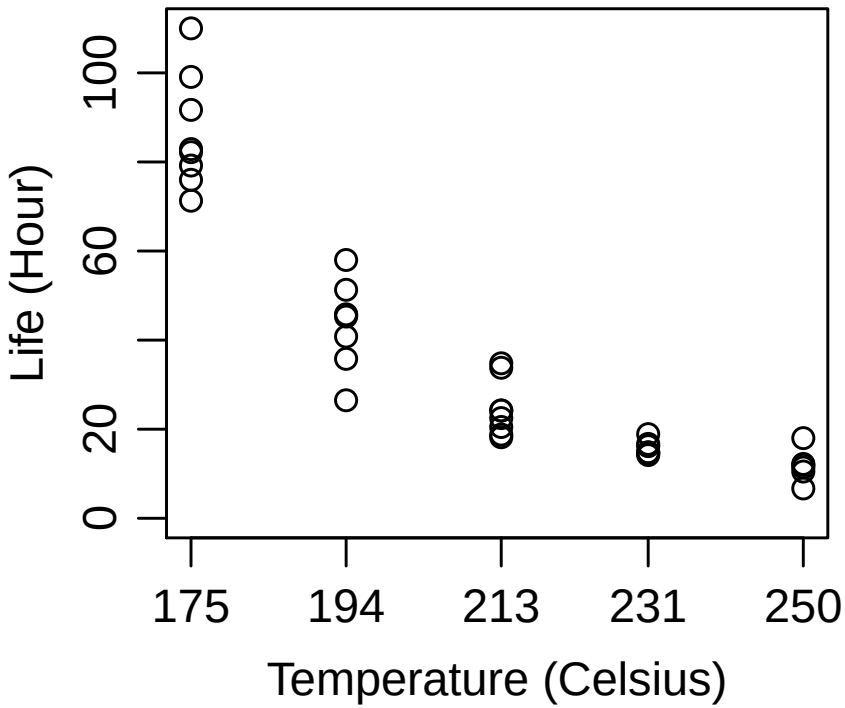
Part II: Constant Variance Assumption

Outline:

- Why Is Non-Constant Variance a Problem?
- Tools for checking nonconstant variance — Residual Plots
  - Residuals v.s. Fitted Values
  - Residuals v.s. Treatments
  - Residuals v.s. Other Variables
- Remedies
  - Transforming the Response — Variance-Stabilizing Transformation
  - Brown-Forsythe Modified F-test — an alternative to ANOVA F-test
  - Welch Test for Contrasts w/o Constant Variance Assumption

Example: Resin Glue Failure Time

In previous lectures, the response of the resin glue experiment is `log10(lifetime)`.
In the original data, the response is simply the life time of the glue without the log transformation.



Temperature	Failure Time in Hours							
175°C	110.0	82.2	99.1	82.9	71.3	91.7	76.0	79.2
194°C	45.8	51.3	26.5	58.0	45.3	40.8	35.8	45.6
213°C	33.8	34.8	24.2	20.5	22.5	18.8	18.2	24.2
231°C	14.2	16.7	14.8	14.6	16.2	18.9	14.8	
250°C	18.0	6.7	12.0	10.5	12.2	11.4		

Data file: `resinlife.txt`

Example — Milk Pasteurization (Exercise 6.2 on p.143)

- Goal: to compare 3 pasteurization methods for milk
- Design: 15 samples of milk randomly assigned to the 3 trts
- Response: the bacterial load in each sample after treatment, determined via serial dilution plating
- Data: http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2

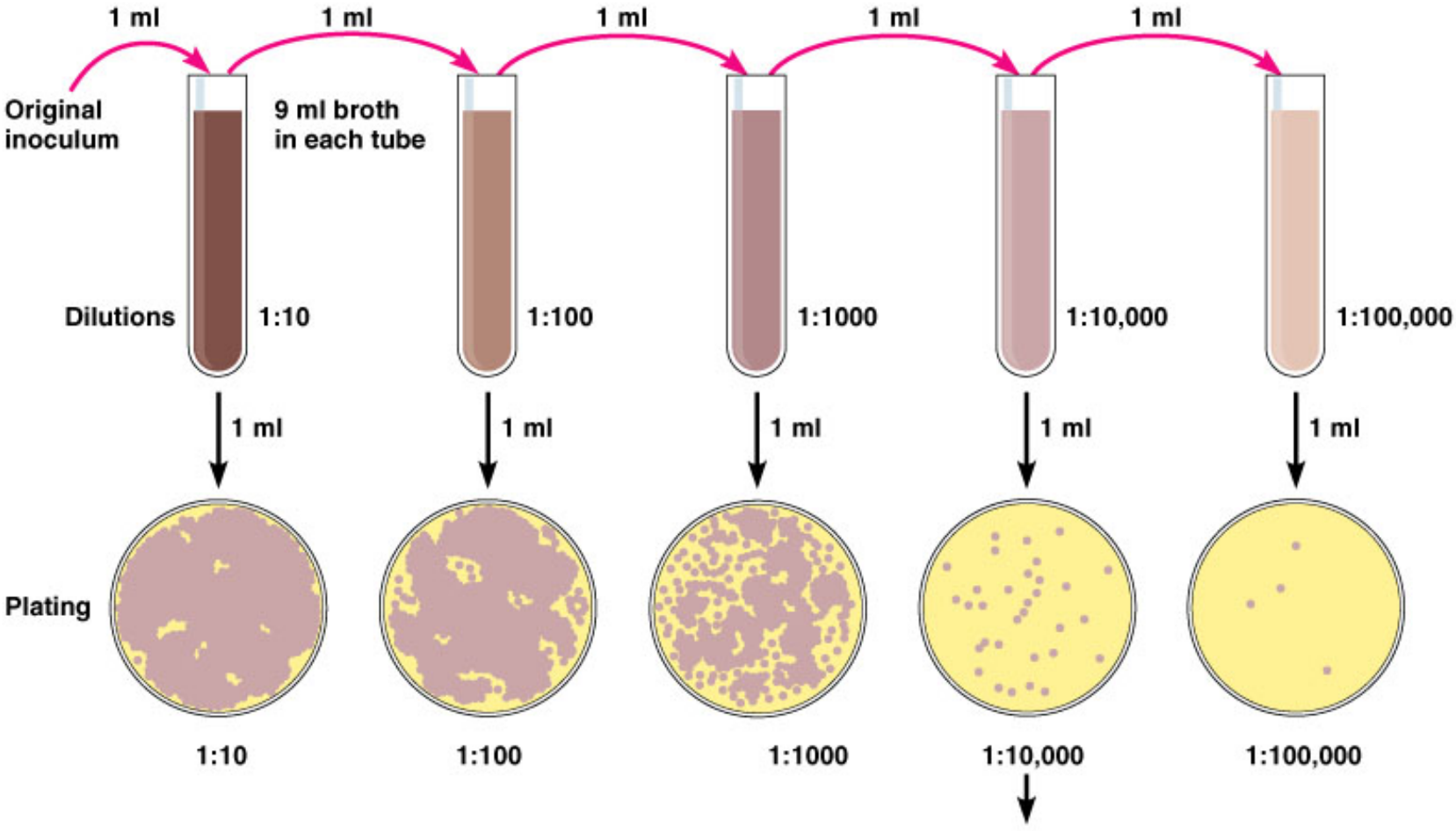
	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656
Size of Noise	100's	1000's	100,000's

$\sqrt{\text{MSE}} = 309900$

$\Rightarrow$ Unequal Variability

How to Count the Number of Bacteria in a Cup of Milk?

Serial Dilution Plating



Calculation: Number of colonies on plate × reciprocal of dilution of sample = number of bacteria/ml
(For example, if 32 colonies are on a plate of $^{1}/_{10,000}$ dilution, then the count is 32 × 10,000 = 320,000/ml in sample.)

Why Non-Constant Variability Causes Problems?

	Method 1	Method 2	Method 3
	26×10^2	35×10^3	29×10^5
	29×10^2	23×10^3	23×10^5
	20×10^2	20×10^3	17×10^5
	22×10^2	30×10^3	29×10^5
	32×10^2	27×10^3	20×10^5
Mean	25.8×10^2	27×10^3	23.6×10^5
SD	492	5874	536656

$\sqrt{\text{MSE}} = 309900$

95%C.I. for the mean of Method 1:

$$\begin{aligned}\overline{y}_{1\bullet} \pm t_{0.025,15-3}\frac{\sqrt{\text{MSE}}}{\sqrt{n_1}} &= 2580 \pm 2.179\frac{309900}{\sqrt{5}} \\ &= 2580 \pm 301965 \\ &= (-299385, 304545)\end{aligned}$$

which is far greater than the range of 5 observations for method 1 (2000-3200). What happened?

Check the Constant Variance Assumption Using Residual Plots

- Residuals v.s. Fitted Values
- Residuals v.s. Treatments
- Residuals v.s. Other Variables

If the constant variance assumption is true, residuals will evenly spread around the zero line.

Rule of thumb: ANOVA F-tests for CRD can tolerate non-constant variance to some extent, so do tests for contrasts. It usually fine as long as

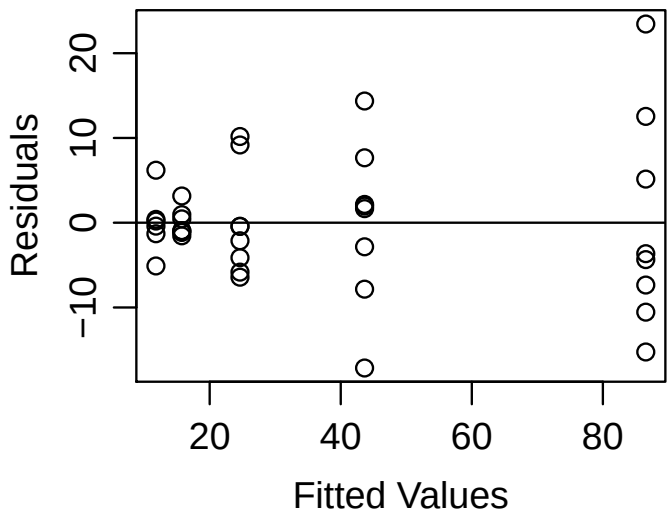
$$\frac{\max\{\widehat{\sigma}_i\}}{\min\{\widehat{\sigma}_i\}} \leq 2,\ 3 \text{ or even } 4,$$

especially when the group sizes n_i are (roughly) equal.

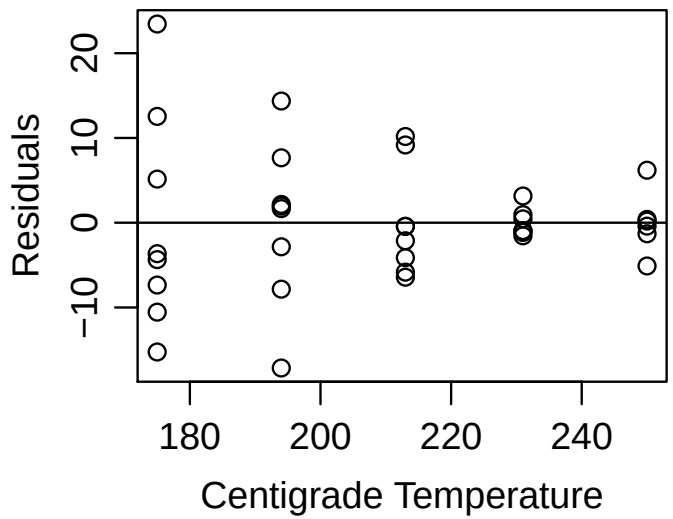
Example: Resin Glue Data

```
resinlife = read.table(
    "http://www.stat.uchicago.edu/~yibi/s222/resinlife.txt", h=T)
lmorig = lm(life ~ as.factor(temp), data=resinlife)
plot(lmorig$fit,lmorig$res,ylab="Residuals", xlab="Fitted Values")
abline(h=0)       # adding a zero line
plot(temp, lmorig$res,ylab="Residuals",xlab="Centigrade Temperature")
abline(h=0)       # adding a zero line
```

Residuals v.s. Fitted Values



Residuals v.s. Treatments



- Why do the points line up vertically?
- Variability of residuals increases with the fitted value, but decreases with the temperature

Remedy 1: Variance-Stabilizing Transformation

If the SD σ (the spread of residuals) changes the mean μ (the fitted values), you can try a *variance-stabilizing transformation* of the response to make the variance (closer to) constant.

- if the SD is proportional to the fitted value, then

$$y \to \log(y)$$

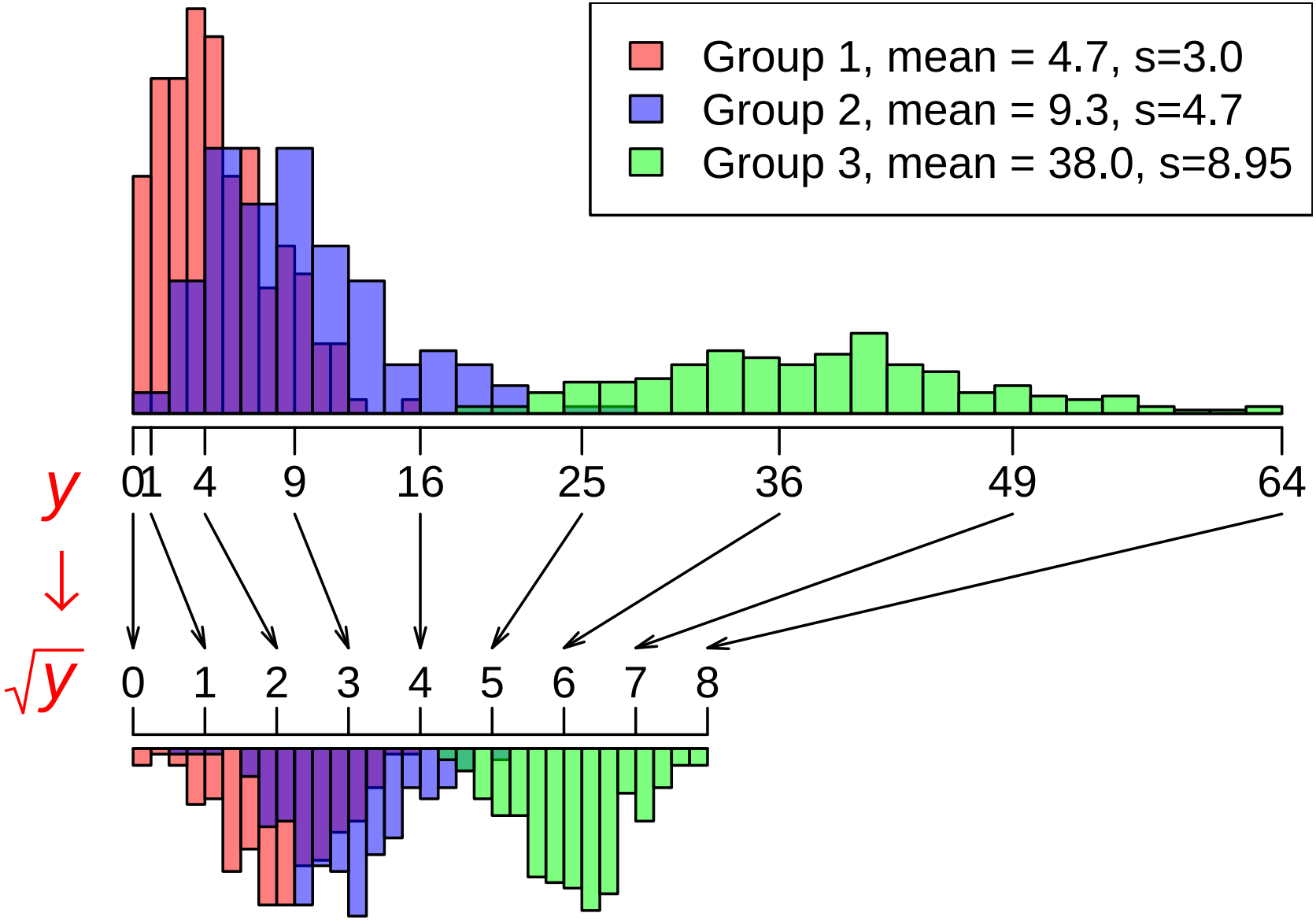
- if the SD is proportional to $\sqrt{\text{the fitted value}}$, i.e., the variance is proportional to the fitted value, then

$$y \to \sqrt{y}$$

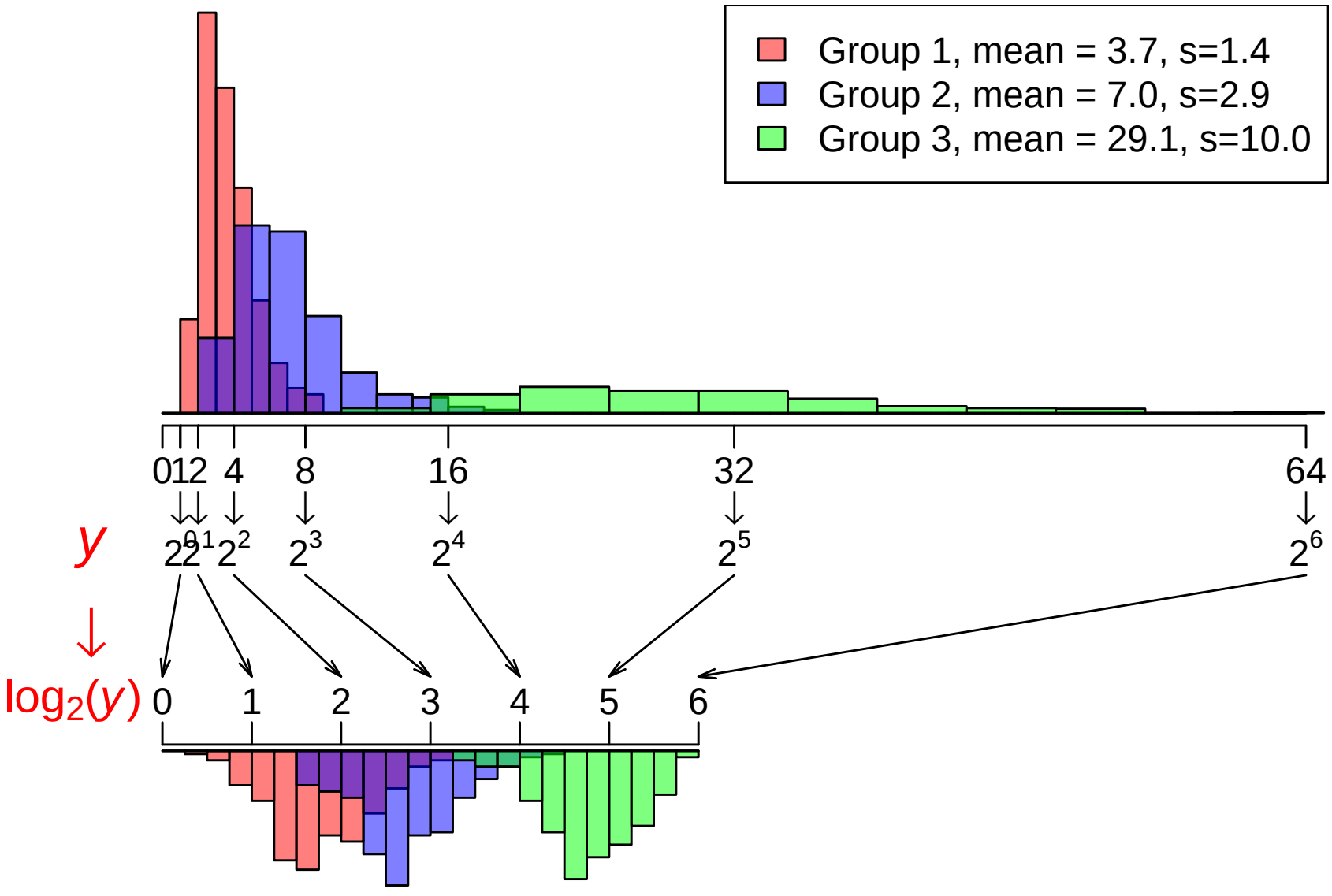
- In general, if the SD σ is proportional to (the fitted values)$^{\alpha}$, then the variance-stabilizing transformation is

$$y \to \begin{cases} y^{1-\alpha} & \text{for } \alpha \neq 1 \\ \log(y) & \text{for } \alpha = 1 \end{cases}$$

How Variance-Stabilizing Transformation Work? (1)



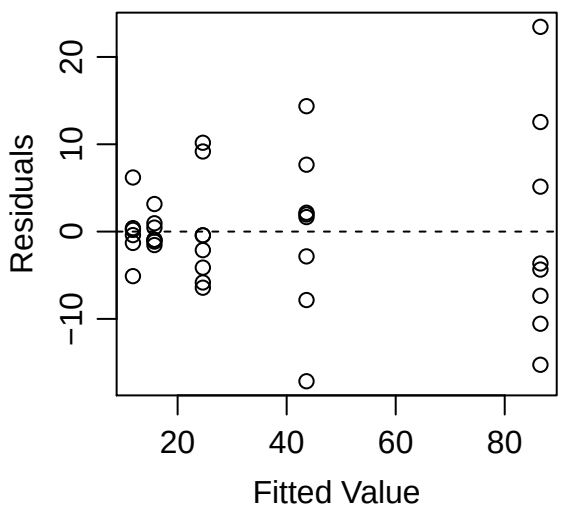
How Variance-Stabilizing Transformation Work? (2)



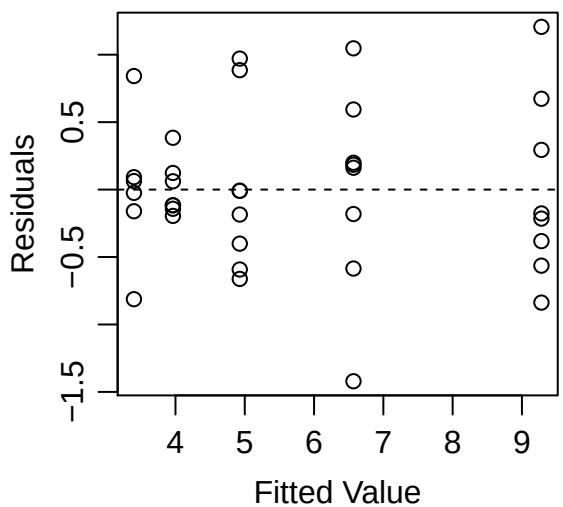
Example: Resin Glue

```
lm1 = lm(life ~ as.factor(temp))
lm2 = lm(sqrt(life) ~ as.factor(temp))
lm3 = lm(log(life) ~ as.factor(temp))
plot(lm1$fit, lm1$res, xlab="Fitted Value", ylab="Residuals")
abline(h=0,lty=2)
plot(lm2$fit, lm2$res, xlab="Fitted Value", ylab="Residuals")
abline(h=0,lty=2)
plot(lm3$fit, lm3$res, xlab="Fitted Value", ylab="Residuals")
abline(h=0,lty=2)
```

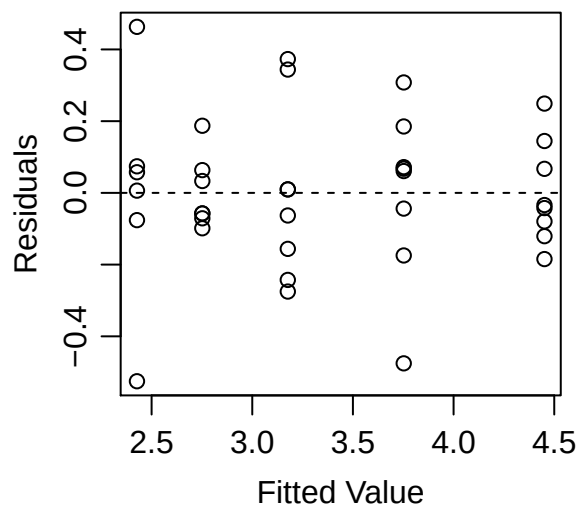
Untransformed



Square-rooted



Log Transform



Box-Cox Again

In many cases, we don't have a good idea what is the value of α. We can still try power transformation of the response.

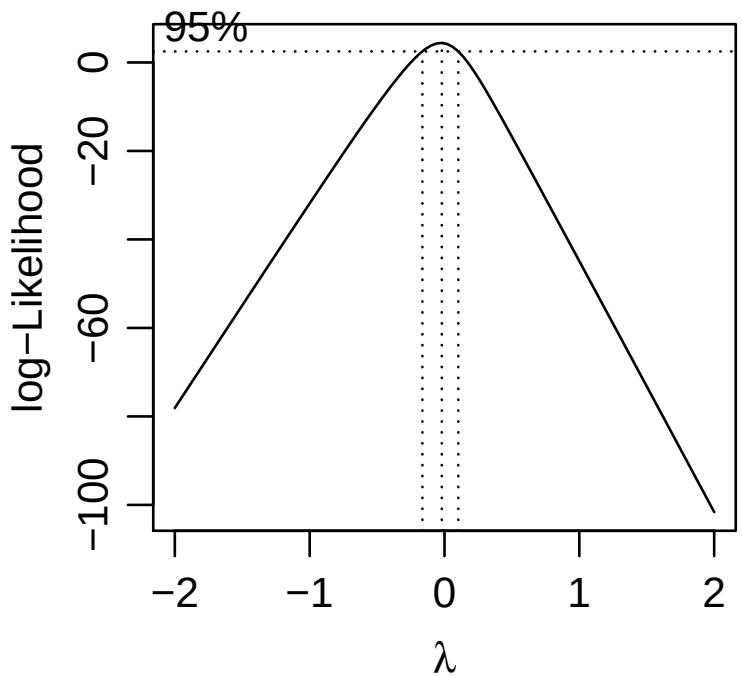
$$f_\lambda(y) = \begin{cases} y^\lambda, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0. \end{cases}$$

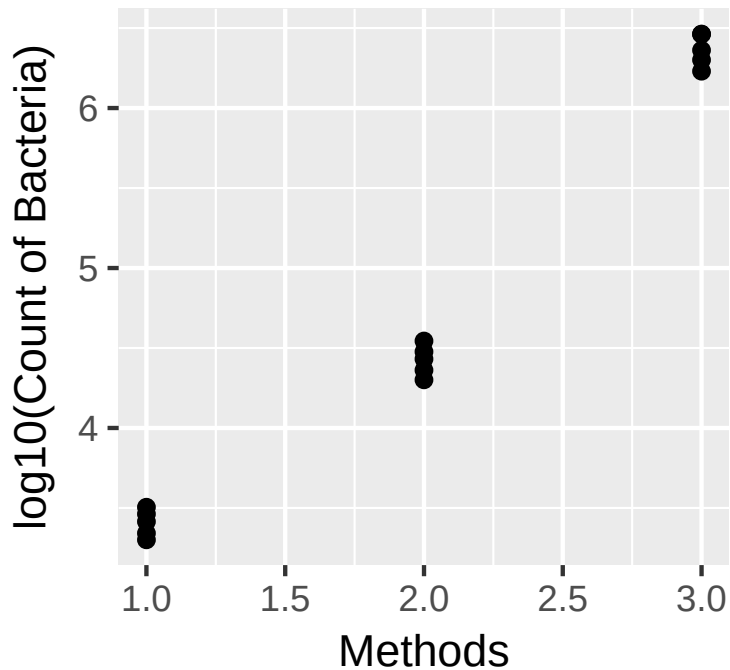
How to select λ?

- Trial and error: try convenient power like $-1, -1/2, -1/3, 0, 1/3, 1/2, 2, \ldots$ and then check residual plots for each of them for the constant variance.
- **Box-Cox method**:
  Though Box-Cox is developed to select a power transformation making the residuals as normal as possible, it's been shown that the optimal λ is often close to the variance-stabilizing λ.

Example – Count of Bacteria Revisit

```
ex6.2 = read.table(
        "http://users.stat.umn.edu/~gary/book/fcdae.data/ex6.2", h=T)
library(MASS)
boxcox(count ~ as.factor(method), data=ex6.2)
plot(ex6.2$method, log10(ex6.2$count),
     xlab="Methods", ylab="log10(Count of Bacteria)")
```

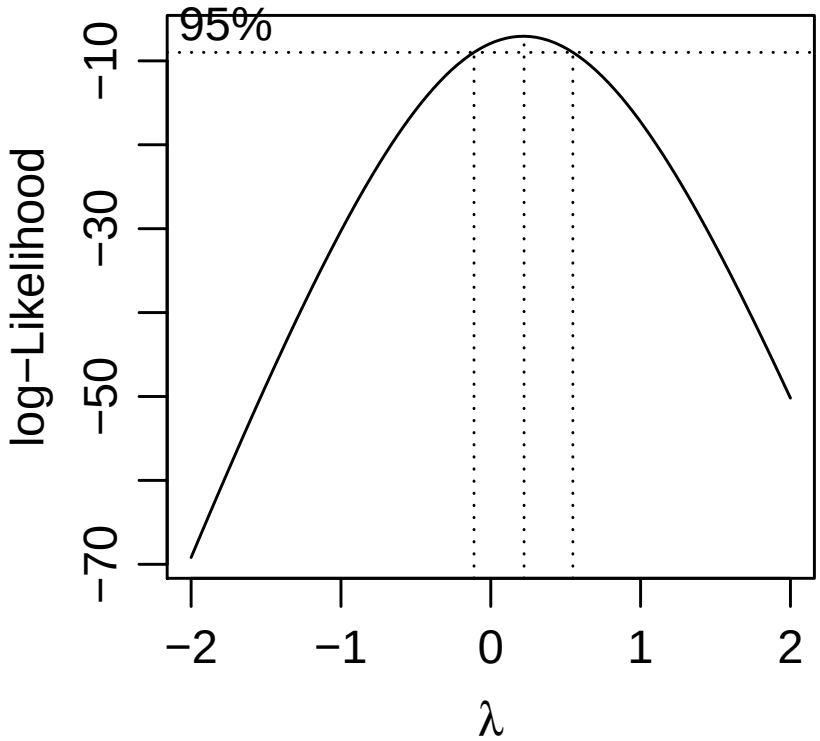




After log transformation, the 3 groups look even in variability.

Example: Resin Glue — Box-Cox

```
library(MASS)
boxcox(life ~ as.factor(temp), data=resinlife)
```



The 95% C.I. for λ contains both 0 and $1/2$. As $\lambda = 1/2$ is very close to the boundary of the C.I, $\lambda = 0$ seems to be a better choice, which is consistent with the Arrhenius Law.

Drawbacks of Transformation

- Except for a few special transformation (log, $\sqrt{\ }$, inverse), the transformed response usually lacks natural interpretation (How to interpret $y^{0.1}$?)

- Unless having a good interpretation on the transformed response, think again before making transformations

Remember that ANOVA tests have some tolerance for non-constant variance. If

$$\frac{\max\{\widehat{\sigma}_i\}}{\min\{\widehat{\sigma}_i\}} \leq 2, 3, \text{ or even } 4,$$

don't worry too much about non-constant variance.
In that case, it is fine to leave the response untransformed.

```
> library(mosaic)
> sd(life ~ temp, data=resinlife)
      175       194       213       231       250
12.895514  9.557785  6.378703  1.661181  3.646917

> sd(sqrt(life) ~ temp, data=resinlife)
      175       194       213       231       250
0.6802037 0.7505140 0.6223047 0.2048698 0.5305554

> sd(log(life) ~ temp, data=resinlife)
      175       194       213       231       250
0.1440872 0.2393055 0.2451210 0.1012540 0.3177677
```

After the log transformation, the ratio of the largest and smallest SD is 2.2, which is acceptable

Brown-Forsythe Modified F-test

If one cannot find an appropriate transformation, **Brown-Forsythe modified F-test** is an alternative of the ANOVA F-test that doesn't rely on the constant variance assumption. The BF test statistic is

$$BF = \frac{\sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)} = \frac{\text{SS}_{trt}}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)}$$

in which s_i^2 is the sample variance in treatment i. Under the null hypothesis of equal treatment means, BF is approximately distributed as an F-distribution with $g - 1$ and ν degrees of freedom, where

$$\nu = \frac{(\sum_{i=1}^{g} d_i)^2}{\sum_{i=1}^{g} d_i^2/(n_i - 1)} \quad \text{in which} \quad d_i = s_i^2(1 - n_i/N).$$

Welch Tests for Contrasts w/o the Constant Variability Assumption

If one cannot find an appropriate transformation, try **Welch test for a contrast** $\sum_{i=1}^{g} \omega_i \mu_i$, which doesn't rely on the constant variability assumption. The test statistic is

$$t = \frac{\sum_{i=1}^{g} \omega_i \overline{y}_{i\bullet}}{\sqrt{\sum_{i=1}^{g} \omega_i^2 s_i^2 / n_i}}$$

which is approximately a t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\sum_{i=1}^{g} \omega_i^2 s_i^2 / n_i\right)^2}{\sum_{i=1}^{g} \frac{1}{n_i - 1} \frac{\omega_i^4 s_i^4}{n_i^2}}$$

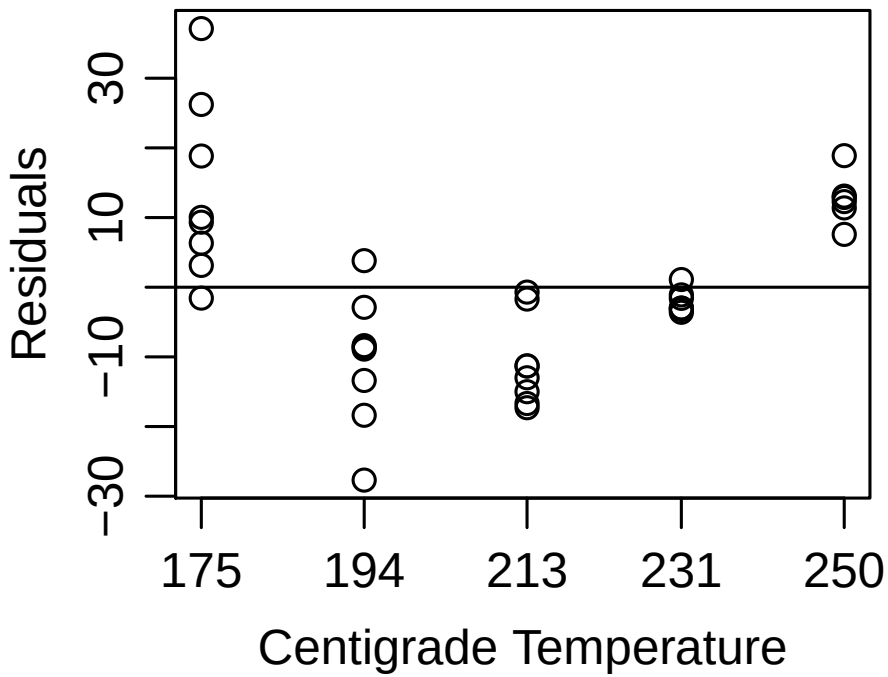
This is a generalization of the two-sample test without the equal variance assumption.

More on Residuals Plots

Residuals plots can be used to check many other things, like non-linearity.

E.g., for the resin glue data, one can check if the linear model $y_{ij} = \beta_0 + \beta_1 T + \varepsilon_{ij}$ is appropriate by checking the residual plot

```
lmorig1 = lm(life ~ temp, data=resinlife)
plot(temp,lmorig1$res,ylab="Residuals",xlab="Centigrade Temperature")
abline(h=0)
```



From the residual plot, we see

- non-constant variance across temperature
- lifetime is *curved*, not linear with temperature

Plot of Residuals v.s. Variables Not in the Model

If there exist other variables that might affect the response, but are not included in the model, then one should check the plots of residuals versus these variables. For example,

- if experimental units come from different batches, then plot residuals v.s. batches
- if measurements are made by several operators, then plot residuals should v.s. operators

Patterns in such residual plots suggest these variables should

- either be included in the analysis
  (but note one CANNOT claim these variables have a *causal-effect* on the response, since they are not controlled in advance)
- or be controlled more carefully, e.g., by a block design, in future experiments

Part III — Checking for Dependent Errors

- Among the 3 assumptions, violation of the independence assumption causes severest problem. Most of our the analysis (ANOVA, test of contrasts, multiple comparisons, etc.) have little tolerance on dependence of errors

- There are various forms of dependence, *serial dependence* and spatial dependence are two common ones

- **Remedies for Dependence**
  - There isn't much we can do about dependence using our current machinery, since no simple transformation can remove dependence.
  - Analysis of dependent data requires tools like **time series** or **spatial statistics**, which is beyond the scope of this class

Check Time Dependence By Plotting Residuals Against Time

- If there is a **time order** that the data are collected, please plot the residuals against the time order.
- If the time-plot of the residuals exhibit any pattern, ...
- Better **keep track of the time order** the data are collected.

Example: Balloon Experiment (Meily Li 1985)

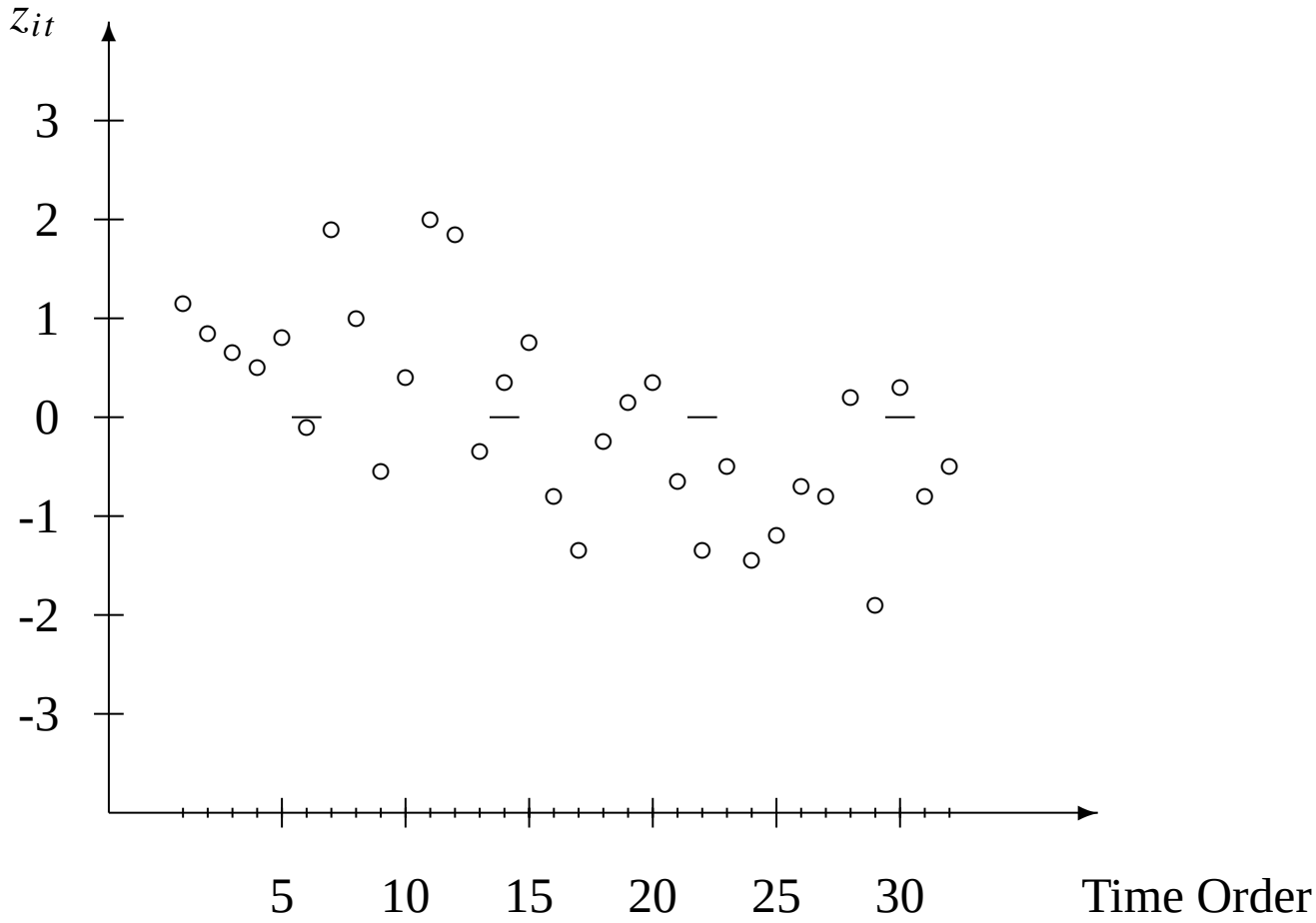
- Goal: To determine whether balloons of different colors are similar in terms of the time taken for inflation to a diameter of 7 inches.
- Four colors were selected from a single manufacturer.
- An assistant blew up the balloons and the experimenter recorded the times (to the nearest 1/10 second) with a stop watch. The data, in the order collected, are given in next page where the codes 1, 2, 3, 4 denote the colors pink, yellow, orange, blue, respectively.
- Any flaw in the design of the experiment?

Example: Balloon Experiment — Data

Times (in seconds) for the balloon experiment

Time Order	1	2	3	4	5	6	7	8
Coded color	1	3	1	4	3	2	2	2
Inflation Time	22.4	24.6	20.3	19.8	24.3	22.2	28.5	25.7
Time Order	9	10	11	12	13	14	15	16
Coded color	3	1	2	4	4	4	3	1
Inflation Time	20.2	19.6	28.8	24.0	17.1	19.3	24.2	15.8
Time Order	17	18	19	20	21	22	23	24
Coded color	2	1	4	3	1	4	4	2
Inflation Time	18.3	17.5	18.7	22.9	16.3	14.0	16.6	18.1
Time Order	25	26	27	28	29	30	31	32
Coded color	2	4	2	3	3	1	1	3
Inflation Time	18.9	16.0	20.1	22.5	16.0	19.3	15.9	20.3

Balloon Experiment — Residual Time Plot



- a clear downward drift in the residuals as time progresses
- a single assistant blew up all the balloons in the experiment, and has become more skillful (required less time to inflate the balloon to the required size) as the he blew up more balloons

Example: Standard Gravity

The National Bureau of Standards performed 8 series of experiments in 1924-1935 to determine g, the standard gravity.

The data are given in the table below (in deviations from $9.8 m/s^2 \times 10^5$, e.g., the first measurement of g is 9.80076 m/s^2), with series 1 representing the earliest set of experiments and series 8 the last.

Series	Measurements												
1	76	82	83	54	35	46	87	68					
2	87	95	98	100	109	109	100	81	75	68	67		
3	105	83	76	75	51	76	93	75	62				
4	95	90	76	76	87	79	77	71					
5	76	76	78	79	72	68	75	78					
6	78	78	78	86	87	81	73	67	75	82	83		
7	82	79	81	79	77	79	79	78	79	82	76	73	64
8	84	86	85	82	77	76	77	80	83	81	78	78	78

Weird ANOVA F-Test

```
> g = c(76,82,83,54,35,46,87,68,87,95,98,100,109,109,100,81,75,68,67,
  105,83,76,75,51,76,93,75,62,95,90,76,76,87,79,77,71,
   76,76,78,79,72,68,75,78,78,78,78,86,87,81,73,67,75,82,83,
   82,79,81,79,77,79,79,78,79,82,76,73,64,
   84,86,85,82,77,76,77,80,83,81,78,78,78)
> series = c(rep(1,8),rep(2,11),rep(3,9),rep(4,8),rep(5,8),rep(6,11),
             rep(7,13),rep(8,13))
> lmg = lm(g ~ as.factor(series))
> anova(lmg)
Response: g
                  Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(series)  7 2818.6  402.66  3.5675 0.002357 **
Residuals         73 8239.4  112.87
```

ANOVA rejects the H_0 of the 8 series having equal means. What does this mean? Will you conclude that

(a) g had changed in the 8 series of measurements, or

(b) the ANOVA F-test failed?

If your answer is (a), how do you explain the change of g?
If your answer is (b), why the ANOVA F-test failed?

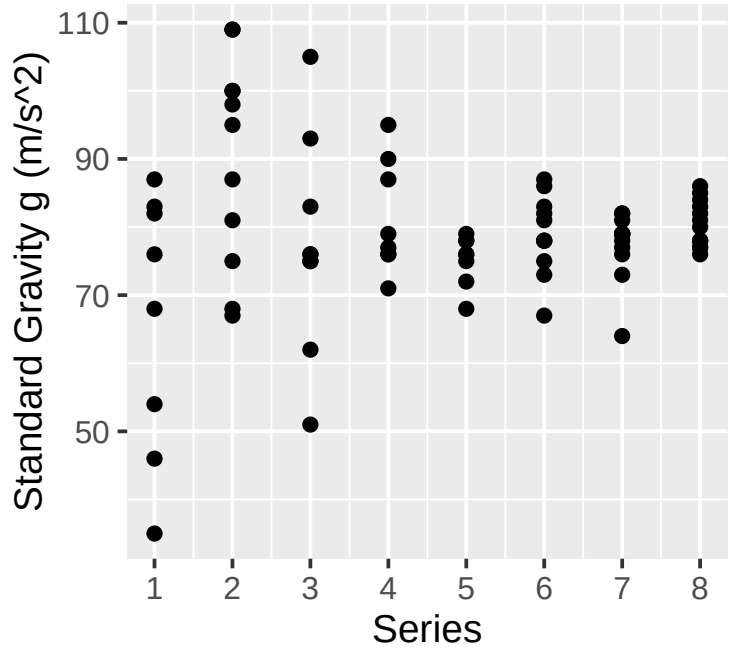
Example: Standard Gravity

The National Bureau of Standards (NBS) (now called the National Institute of Standards and Technology (NIST)) is the government agency that measures things. The following statement is taken from the NIST website:

> *Founded in 1901, NIST is a non-regulatory federal agency within the U.S. Department of Commerce. NIST mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.*

Thus, it is safe to assume that the NBS scientists were trying hard to measure the same quantity g (e.g., all experiments were done in the same location) throughout all 8 series of experiments.

```
> qplot(series, g, xlab="Series",
  ylab="Standard Gravity g (m/s^2)")+
  scale_x_continuous(breaks=1:8)
```



Variance decreases with series, which makes sense since the accuracy of measurement improved as time went by.

The ANOVA F-test here may not be reliable because at least the constant variance assumption is not met

```
> round(sd(g ~ series), 2)
    1     2     3     4     5     6     7     8
19.25 15.29 15.76  8.30  3.65  5.84  4.74  3.36
```

Will a transformation work here? Box-Cox?

No. The variance-stabilizing transformation works only when the variability increases or decreases with the mean. Here the means of the 8 series are nearly the same.

Brown-Forsythe Modified F-test

In view of the nonconstant variability, let's try the Brown-Forsythe modified F-test

$$BF = \frac{SS_{Trt}}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)}$$

The numerator $SS_{Trt} = 2819$ can be found in the ANOVA table above. The denominator is found using R (see the codes below) to be 888.5747

```
> sds = sd(g ~ series); sds
       1        2        3        4        5        6        7        8
19.24977 15.29349 15.75595  8.29694  3.65474  5.83874  4.73665  3.35506
> ni = c(8, 11, 9, 8, 8, 11, 13, 13)
> di  = sds^2*(1-ni/sum(ni))
> BFbottom = sum(di)
> BFbottom
[1] 888.5747
> BF = 2819/BFbottom
```

The BF-statistic is thus BF $= \dfrac{2819}{888.5747} = 3.1725.$

Under the null hypothesis of equal group means, BF has an approx. F-distribution with df1$= g - 1$ and df2 given below

$$\text{df2} = \frac{(\sum_{i=1}^{g} d_i)^2}{\sum_{i=1}^{g} d_i^2/(n_i - 1)} \quad \text{in which} \quad d_i = s_i^2(1 - n_i/N)$$

where the second degrees of freedom df2 is calculated as 29.46 in the R code below.

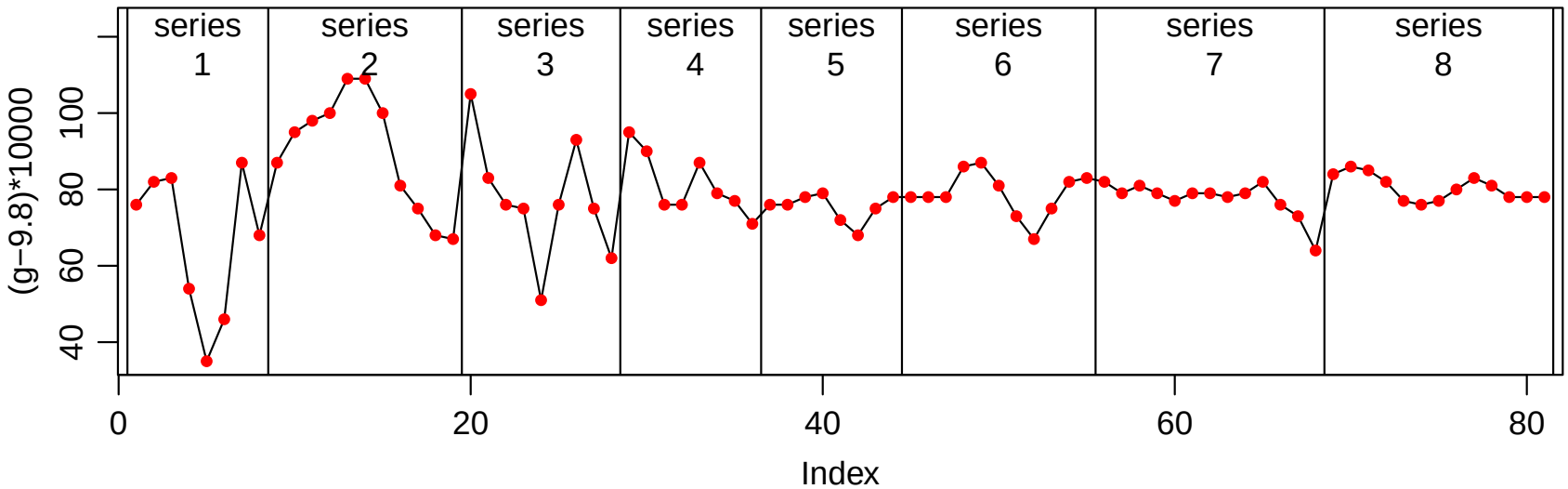
```
> df2 = (BFbottom)^2/sum(di^2/(ni-1))
> df2
[1] 29.46249
> pf(BF, 8-1, df2, lower.tail=F)        # P-value of the BF-test
[1] 0.01263341
```

However, the BF-test, not relying on the constant variance assumption, also rejects the null hypothesis of equal mean at a P-value 0.0126. Why the BF-test also failed?

Tools for Checking Serial Dependence

1. **Time plot**: a plot of residuals v.s. the order they are measured)
   - It's better to keep track of the order units are measured.
   - A smooth time-plot is a sign of positive serial dependence, since a smooth time plot means successive residuals are too close together
2. **Autocorrelation Plots**
   - Lag 1 autocorrelation plot: plotting $(e_1, \ldots, e_{n-1})$ v.s. $(e_2, \ldots, e_n)$
   - Lag k autocorrelation plot: plotting $(e_1, \ldots, e_{n-k})$ v.s. $(e_{1+k}, \ldots, e_n)$
   - Any trend in the autocorrelation plot is a sign of serial dependence.
3. **Autocorrelation**
   - the Lag-k *autocorrelation coefficient* is the correlation coefficient of $(e_1, \ldots, e_{n-k})$ v.s. $(e_{1+k}, \ldots, e_n)$, $k = 1, 2, 3, \ldots$.

The observations in each series are in fact given in time order taken. We can thus make a time plot.

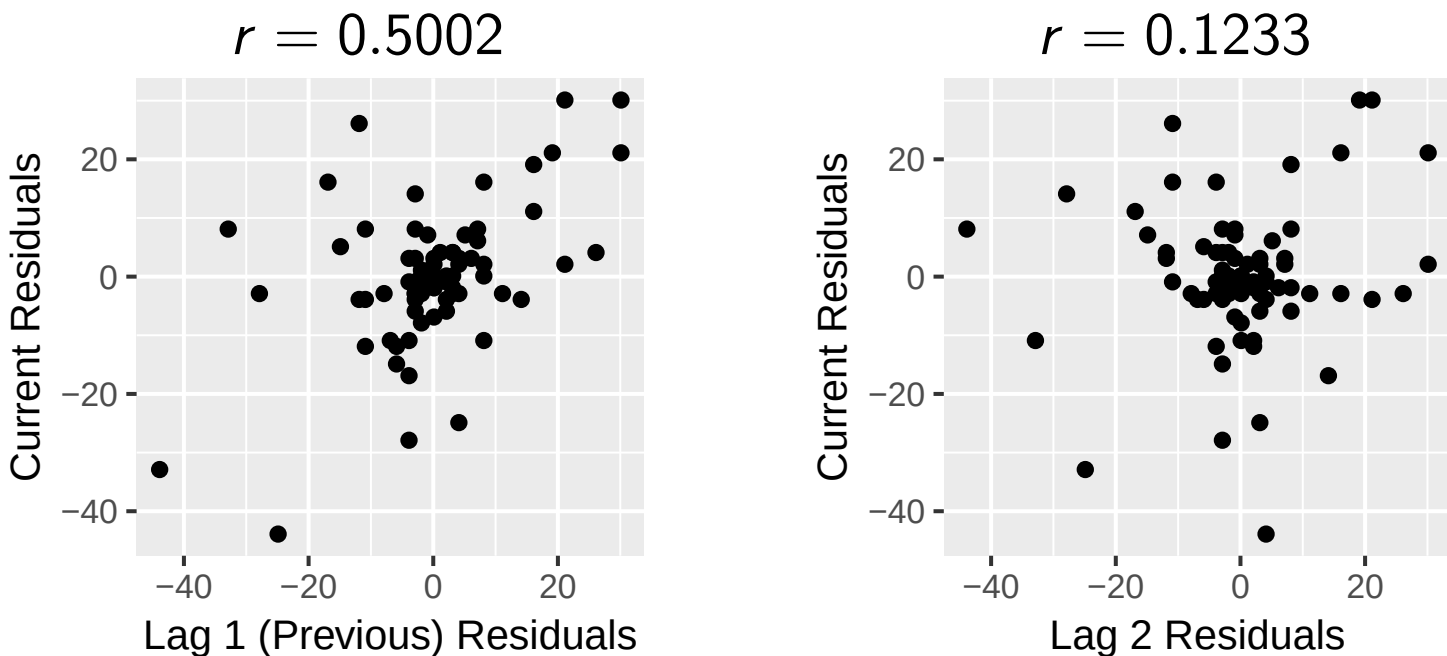


We can see a lot of measurements are close to the previous measurements, which indicates a positive serial correlation.

It's not surprising that scientists in NIST might unconsciously match their results with the previous measurement, which was often regarded as the most accurate one till then.

Autocorrelation Plots

```
> lm0 = lm(g ~ 1)   # Null model: All series had the same mean
> res = lm0$res     # Residual of null model
> qplot(res[2:81],res[1:80], xlab="Current Residual",
   ylab="Lag 1 (Previous) Residual")
> cor(res[2:81],res[1:80])
[1] 0.5002454
> qplot(res[1:79], res[3:81], ylab="Current Residuals",
      xlab="Lag 2 Residuals")
> cor(res[3:81],res[1:79])
[1] 0.1232966
```



Effect of Dependent Errors

- When $Y_1, Y_2, \ldots, Y_n$ are dependent,

  $$\mathrm{Var}(Y_1 + Y_2 + \ldots + Y_n) = \sum_{j=1}^{n} \mathrm{Var}(Y_j) + \sum_{j,k:j\neq k} \mathrm{Cov}(Y_j, Y_k).$$

  Suppose $\mathrm{Var}(Y_j) = \sigma^2$ and $\mathrm{Cov}(Y_j, Y_k) = \rho_{jk}$.

  $$\mathrm{Var}(Y_1 + Y_2 + \ldots + Y_n) = n\sigma^2 + \sum_{j,k:j\neq k} \rho_{jk}$$

  $$\Rightarrow \mathrm{Var}(\overline{Y}) = \frac{\sigma^2}{n} + \frac{\sum_{j,k:j\neq k} \rho_{jk}}{n^2} \neq \frac{\sigma^2}{n}$$

- If one-way ANOVA data $\{y_{ij}\}$ have stronger positive within group correlations than between group correlation

  $$\mathrm{Var}(\overline{y}_{i\bullet} - \overline{y}_{j\bullet}) > \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j},$$

  the SE $= \sqrt{\mathrm{MSE}(\frac{1}{n_i} + \frac{1}{n_j})}$ calculated assuming independence would underestimate the actual SD of $\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$.Hence, $t = \frac{\overline{y}_{i\bullet} - \overline{y}_{j\bullet}}{\mathrm{SE}}$ tends to be too large, making it more likely to reject H_0: $\mu_i = \mu_j$.

- The usual ANOVA F-test will be affected for similar reasons.

Spatial Dependence

Spatial dependence can arise when the experimental units are arranged in space, like plants in a farm. Spatial dependence occurs when units that are closer together are more similar than units farther apart.