Lecture 3: ANOVA Diagnostics and Remedial Measures

STA 106: Analysis of Variance

Suggested reading: ALSM Chapter 18

Xiner Zhou
Department of Statistics, University of California, Davis

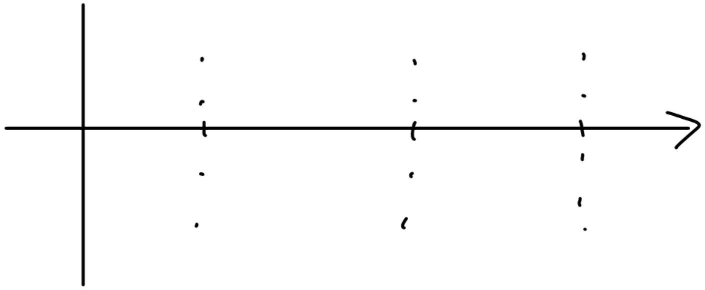
ANOVA Diagnostics and Remedial Measures

- ANOVA Model Assumptions
- Check assumptions
- Remedies and concerns about assumption violations

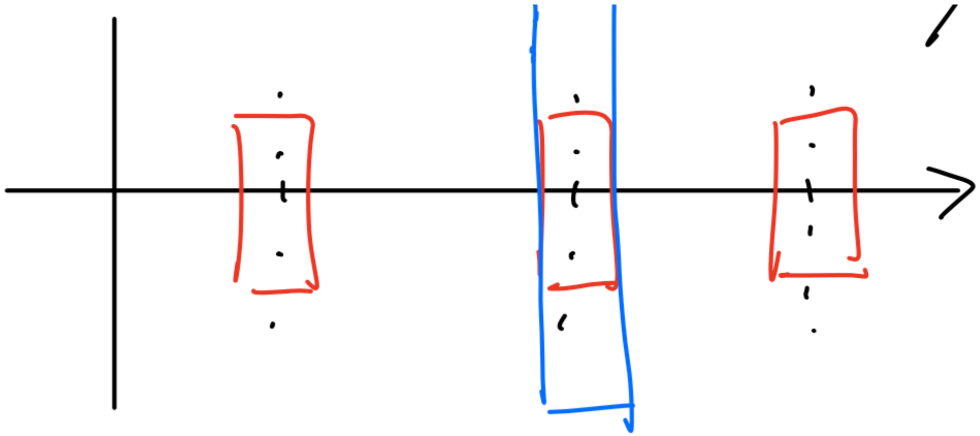
Constancy or Homogeneity of Error Variance

1. Graphical Way

Plot residuals against treatment or factor levels
If Constancy of error variance holds, then we would expect to see the same extent of scatter of residuals around zero, across all factor levels



When sample size is large, box plots of residuals arranged vertically can give a concise summary of the extent of scatters



Constancy or Homogeneity of Error Variance

2. Formal Statistical Tests

Hartley test

To test equal variance using independent samples of equal size selected from r normal populations:

$H_0: \sigma_1^2 = \cdots = \sigma_r^2$ vs H_a: not all σ_i^2 are equal

$$H^* = \frac{\max(s_i^2)}{\min(s_i^2)} \overset{H_0}{\sim} H(r, \underset{= n-1}{df})$$

The Hartley test stat is based on the largest sample variance & smallest sample...

H^* near 1 support H_0

large H^* support H_a.

where s_i^2 is sample variance of n_i obs Y_{ij} for ith factor level.

$$s_i^2 = \frac{\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}{n-1} = \frac{\sum_{j=1}^{n} e_{ij}^2}{n-1}$$

Hartley test

- simple to compute
- Only applicable if
  - sample size are equal (or almost equal) across factor levels
  - error terms are normally distributed, quite sensitive to departure from normality

Constancy or Homogeneity of Error Variance

2. Formal Statistical Tests

Brown-Forsythe test (a nonparametric test)

- more difficult to compute
- No sample size requirement
- Robust to departure from normality

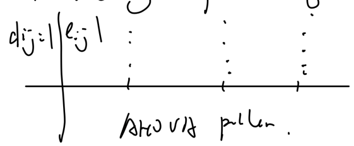
To test equal variance using independent samples from r populations:

$H_0: \sigma_1^2 = \cdots = \sigma_r^2$ vs H_a: not all σ_i^2 are equal

idea: compare absolute deviations of Y_{ij} obs

$d_{ij} = |Y_{ij} - \tilde{Y}_{i\cdot}| = |e_{ij}|$ $\quad \tilde{Y}_{i\cdot}$: median

translate original response $Y_{ij} \longrightarrow d_{ij}$ which measures the variability



ANOVA pattern.

variance = expected distance squared $\sigma_i^2 \approx E(d_{ij}^2)$

r variances σ_i^2's are equal $\Longrightarrow$ $E(d_{1j}) = \cdots = E(d_{rj})$

same exp. value of absolute deviation

not equal $\Longrightarrow$ $E(d_{ij}) \neq E(d_{rj})$

So. test equal variance is the same as F test for equal factor level means with resp. $\{d_{ij}\}$

$$\Longrightarrow F^*_{BF} = \frac{MSTR}{MSE}$$

where $MSTR = \dfrac{\sum n_i (\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{r-1}$

$$MSE = \frac{\sum_i \sum_j (d_{ij} - \bar{d}_{i\cdot})^2}{n_T - r}$$

$$\bar{d}_{i\cdot} = \frac{\sum_j d_{ij}}{n_i}$$

$$\bar{d}_{\cdot\cdot} = \frac{\sum \sum d_{ij}}{n_T}$$

$$F^*_{BF} \overset{H_0}{\sim} F(r-1, n_T - r)$$

Large F^*_{BF} value indicates: heterogeneity of error term variances across factor levels.

Example

Helicopter service.

An operations analyst in a sheriff"s department studied how frequently their emergency helicopter was used during the past year, by time of day:

shift 1: 2 A.M.-8 A.M.

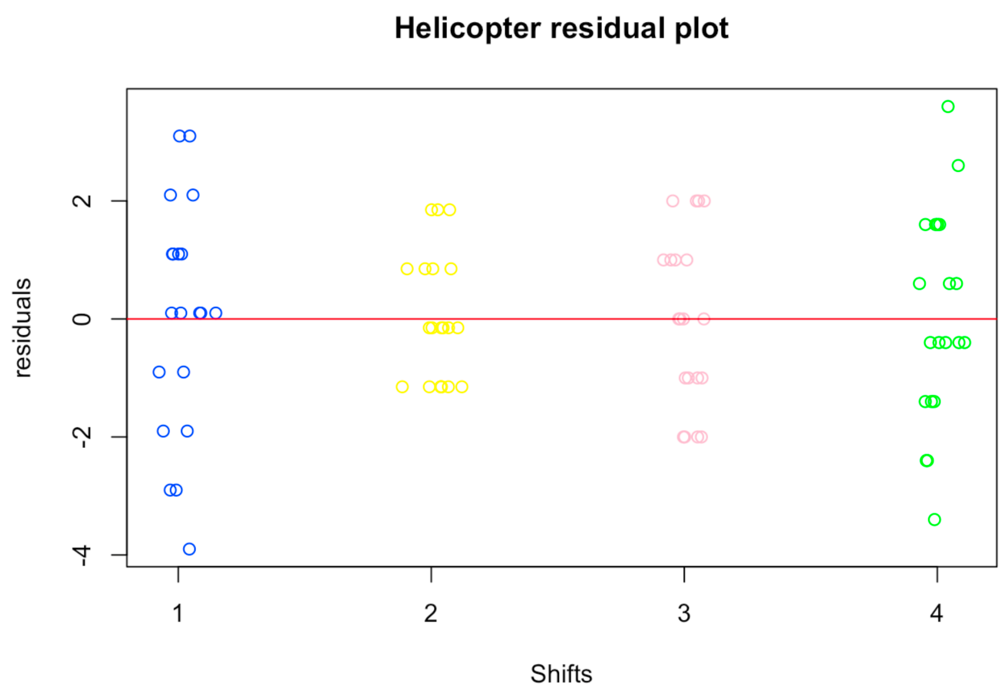
shift 2: 8 A.M.-2 P.M.

shift 3: 2 P.M.-8 P.M.

shifi 4: 8 P.M.-2 A.M..

Random Samples of size 20 for each shift were obtained. Since the data are counts. the analyst was concerned about the normality and equal variances assumptions of ANOVA model.

To study whether or not the error variances are equal in a graphical way. What are your findings?



The residual plot suggests that the error variance for the shift 2 and 3 tend to be smaller than that for shift 1 and 4 , though due to small sample sizes for each treatment, the differences might just due to randomness.

Example

To study whether or not the error variances are equal using formal tests. State the alternatives. decision rule, and conclusion. What are your findings? Are your results consistent?

Code

```
##
##  Hartley's maximum F-ratio test of homogeneity of variances
##
## data:  y by i
## F Max = 3.2727, df = 19, k = 4, p-value = 0.05991
```

Hide

```
# or implememnt by yourself
library(SuppDists)

# p-value appraoch
si2=rep(NA,4)
for(i in 1:4){
  si2[i]=var(helicopter$y[helicopter$i==i])
}

## test statistics
H=max(si2)/min(si2)
## p value
pmaxFratio(H, df=20-1, k=4, lower.tail=FALSE)
```

```
## [1] 0.05990844
```

Hide

```
# Brown-Forsythe test

# compute absolte deviations from median
Ymedian = rep(0,4)
for(i in 1:4)
{
  Ymedian[i] = median(helicopter$y[helicopter$i==i])
}

d = abs(helicopter$y-Ymedian[helicopter$i])

# ANOVA for d
mu_hat = rep(0,4)
for(i in 1:4)
{
  mu_hat[i] = mean(d[helicopter$i==i])
}

# fitted value for d
d_hat=mu_hat[helicopter$i]

MSTR = sum((d_hat-mean(d))^2)/(r-1)
MSE = sum((d-d_hat)^2)/(n_T-r)
F_BF = MSTR/MSE
## test statistic value
F_BF
```

```
## [1] 1.700657
```

Hide

```
## p value
pf(F_BF, r-1,n_T-r, lower.tail=FALSE)
```

```
## [1] 0.1739958
```

Both the Hartley and Brown-Forsythe test do not reject the null, therefore we can conclude that the data does not provide enough evidence that equal variance assumption is violated.

Normality of Error Variance

Small departures from normality do not create any serious problems
Major departure should be of concern.

Factor level sample size

- **small**
  - combine residuals for all factor levels into one. Check normality for the pooled sample
    - if no major difference in error variance is detected
      - residuals
    - if major difference in error variance is detected
      - studentized residuals

$$r_{ij} = \frac{e_{ij}}{s\{e_{ij}\}}$$

Because non-normal may be detected solely due to we combine error terms having unequal variance

- **large**
  - check normality separately for each factor level

Normality of Error Variance

Graphical Way

Distribution Plots: Histogram

- helpful for detecting serious departure
- sample size must be reasonably large for histogram to convey reliable information about the shape of the distribution

Normality of Error Variance

Graphical Way: Normal Probability Plot or Q-Q Plot

Each residual is plotted against its expected value under normality.

Suppose $e_{ij} \sim N(0, \sigma^2)$

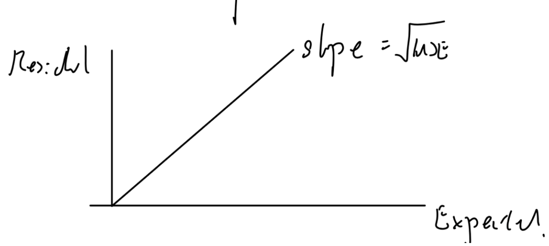
where $\sigma^2 \approx MSE$

then: order residuals $e_{(1)}, e_{(2)}, \dots e_{(k)}$

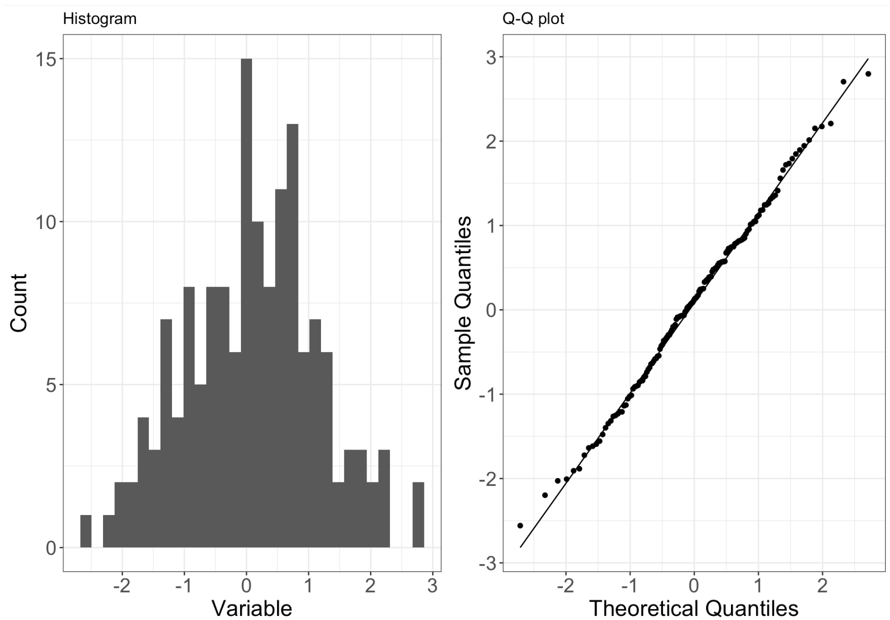
k-th smallest obs in a random sample of n

$$E[e_{(k)}] = \sqrt{MSE}\, \underbrace{z\left(\frac{k-0.375}{n+0.25}\right)}_{z(p)\ p\text{th Quantile of } N(0,1)}$$

$e_{(k)} \approx E[e_{(k)}]$ for most of obs.



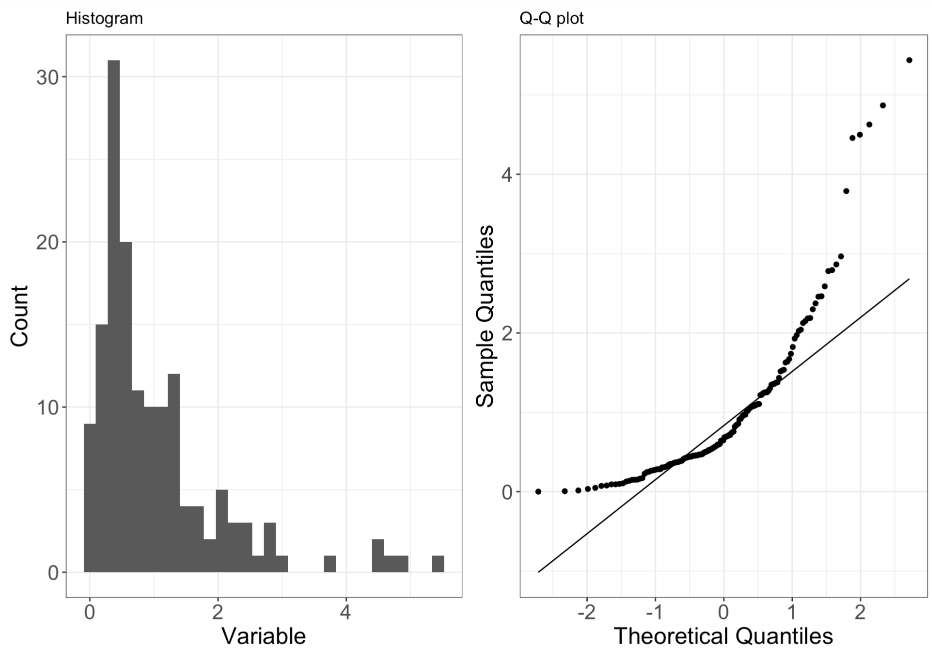
Normality of Error Variance



Normally distributed data

On a Q-Q plot, normally distributed data appears as roughly a straight line, although the ends of the Q-Q plot often start to deviate from the straight line.

Normality of Error Variance

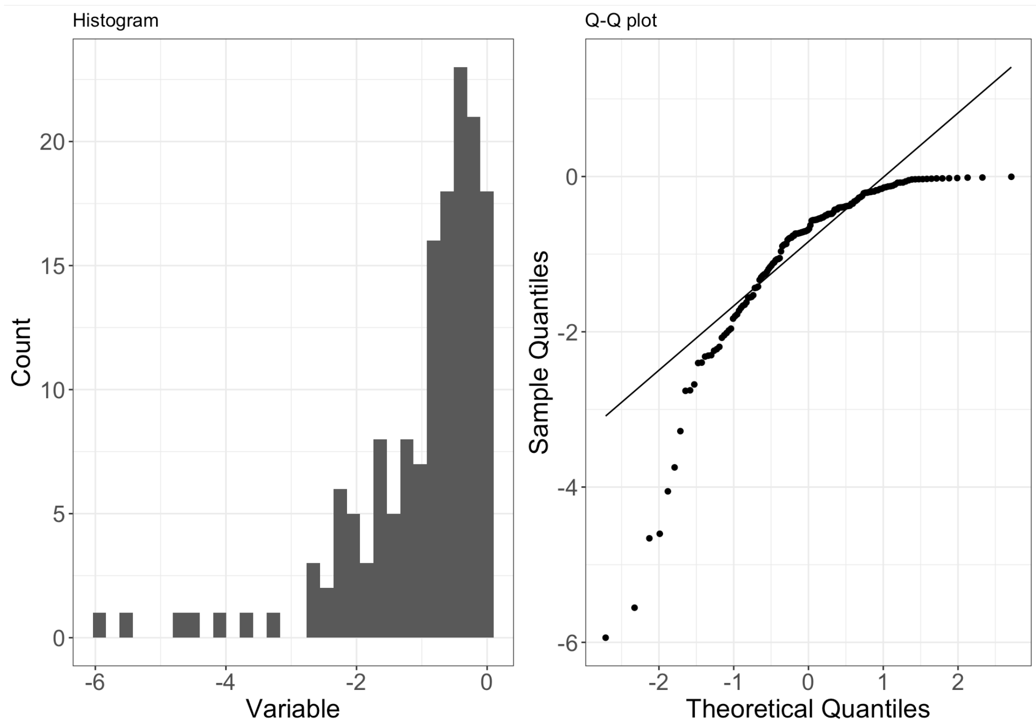


Right-skewed data

There are more data on the right tail of the distribution, than a normal distribution would.

Mostly, residual quantiles are larger than what is expected to be if it's truly normal, so the residuals are all shifted or skewed towards the right

Normality of Error Variance

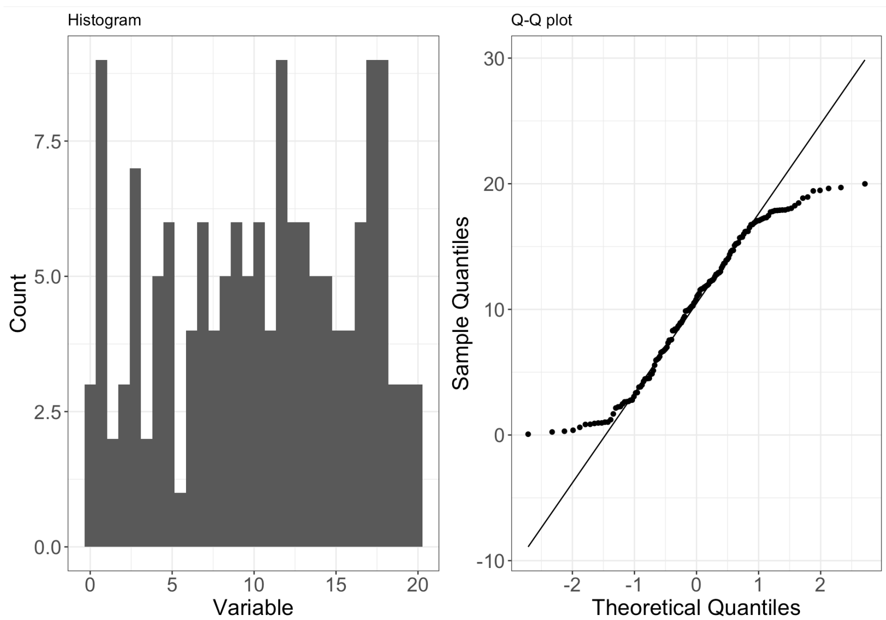


Left-skewed data

There are more data on the left tail of the distribution, than a normal distribution would.

Mostly, residual quantiles are smaller than what is expected to be if it's truly normal, so the residuals are all shifted or skewed towards the left

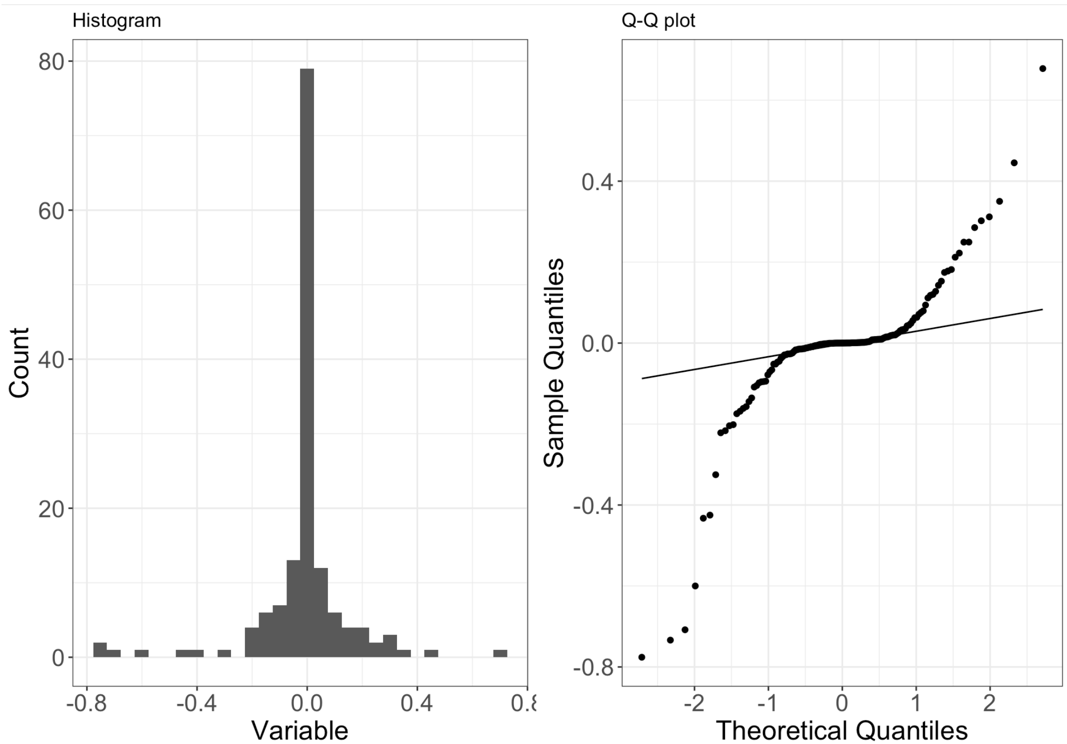
Normality of Error Variance



Under-dispersed data

It exhibits patterns resembles right-skewed data on the left tail, and left-skewed data on the right tail, so the residuals is more concentrated on the center, thus less dispersed than normally distributed data.

Normality of Error Variance



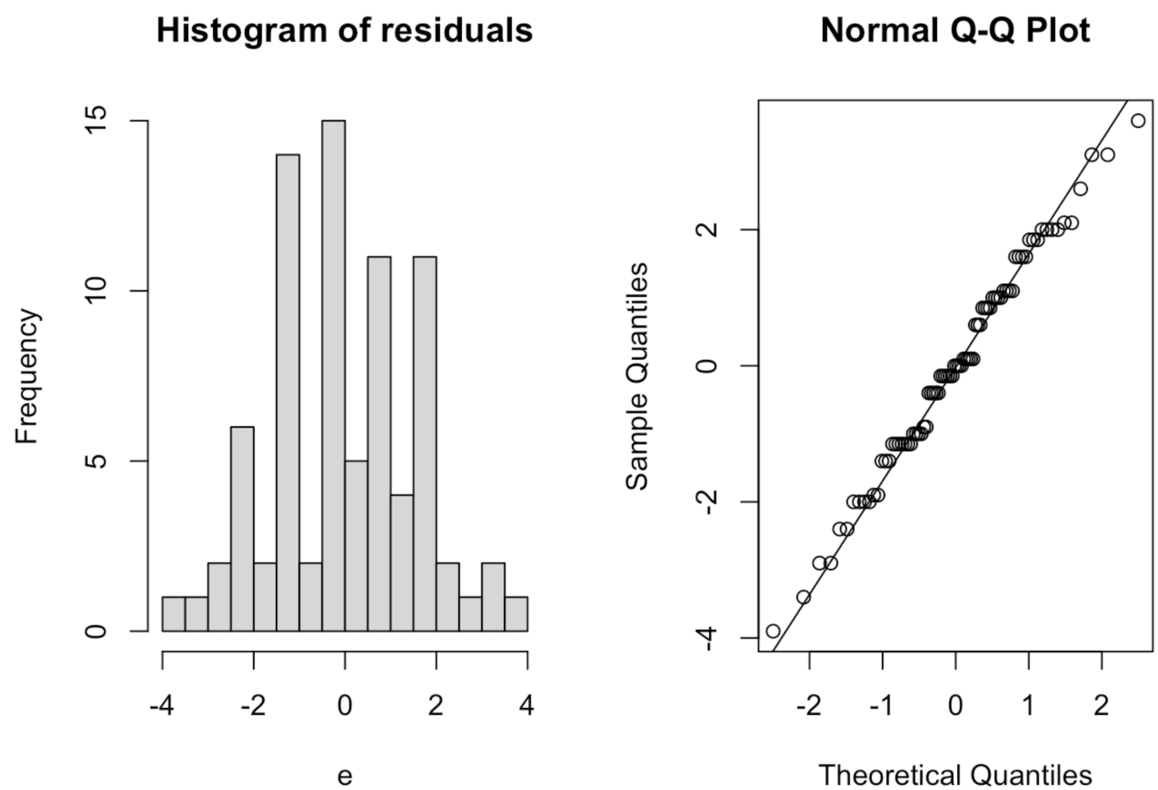
Over-dispersed data

It exhibits patterns resembles left-skewed data on the left tail, and right-skewed data on the right tail, so the residuals is fatter tails than a normal distribution, thus more dispersed than normally distributed data.

Example

To study whether or not the normality assumption is met.

Since each factor level only has 20 observations, it is considered not large, we check normality by combining the residuals.



The histogram shows the data is roughly symmetrically and bell-shaped. On the Q-Q plot, data appears as roughly a straightline. The normality assumption is met.

Example

Summarize your findings about whether the ANOVA assumptions are met.

The independence assumption is automatically satisfied by the set up of the experiemnt. No violation of he equal variance of error term assumption is detected by Hartley and Brown-Forsythe tests.

No violation of normality is detected by checking histogram and Q-Q plot.

Therefore, there is no evidence that the ANOVA assumptions are violated.

However, be aware that those tests only suggest that no violation is detected, they do not confirm these assumptions actually hold for the data at hand definitively.