STA305/1004-Class 16

Nov. 12, 2019

Today's Class

Assessing significance in unreplicated factorial designs:

- Normal plots
- half-Normal plots
- Lenth's method

ANOVA:

- Multiple comparisons
- Sample size for ANOVA

Assessing Significance in Unreplicated Factorial Designs

How can significance be assessed in unreplicated factorial designs?

Quantile-Quantile Plots

- Quantile-quantile (Q-Q) plots are useful for comparing distribution functions.
- If X is a continuous random variable with strictly increasing distribution function $F(x)$ then the *pth* quantile of the distribution is the value of x_p such that,

$$F(x_p) = p$$

or

$$x_p = F^{-1}(p).$$

- In a Q-Q plot, the quantiles of one distribution are plotted against another distribution.
- Q-Q plots can be used to investigate if a set of numbers follows a certain distribution.

Quantile-Quantile Plots

- Suppose that we have independent observations $X_1, X_2, ..., X_n$ from a uniform distribution on $[0, 1]$ or Unif[0,1].
- The ordered sample values (also called the order statistics) are the values $X_{(j)}$ such that

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

.

- It can be shown that

$$E\left(X_{(j)}\right) = \frac{j}{n+1}.$$

- This suggests that if we plot

$$X_{(j)} \quad \text{vs.} \quad \frac{j}{n+1}$$

then if the underlying distribution is Unif[0,1] then the plot should be roughly linear.

Quantile-Quantile Plots

- A continuous random variable with strictly increasing CDF F_X can be transformed to a Unif[0,1] by defining a new random variable $Y = F_X(X)$.
- Suppose that it's hypothesized that X follows a certain distribution function with CDF F.
- Given a sample $X_1, X_2, ..., X_n$ plot

$$F(X_{(k)}) \quad \text{vs.} \quad \frac{k}{n+1}$$

or equivalently

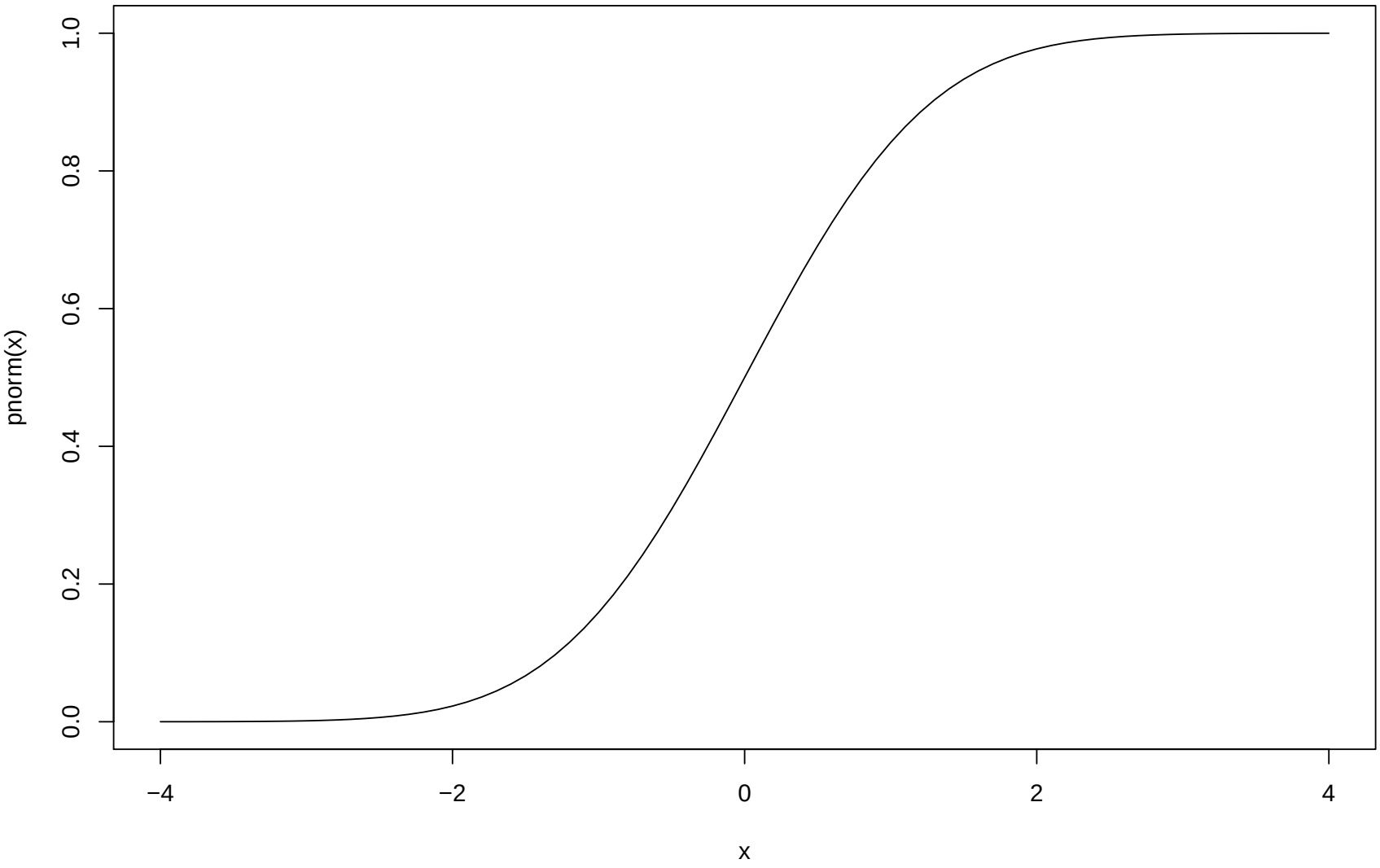
$$X_{(k)} \quad \text{vs.} \quad F^{-1}\left(\frac{k}{n+1}\right)$$

- $X_{(k)}$ can be thought of as empirical quantiles and $F^{-1}\left(\frac{k}{n+1}\right)$ as the hypothesized quantiles.
- The quantile assigned to $X_{(k)}$ is not unique.
- Instead of assigning it $\frac{k}{n+1}$ it is often assigned $\frac{k-0.5}{n}$. In practice it makes little difference which definition is used.

Normal Quantile-Quantile Plots

The cumulative distribution function (CDF) of the normal has an S-shape.

```
x <- seq(-4,4,by=0.1)
plot(x,pnorm(x),type="l")
```



Normal Quantile-Quantile Plots

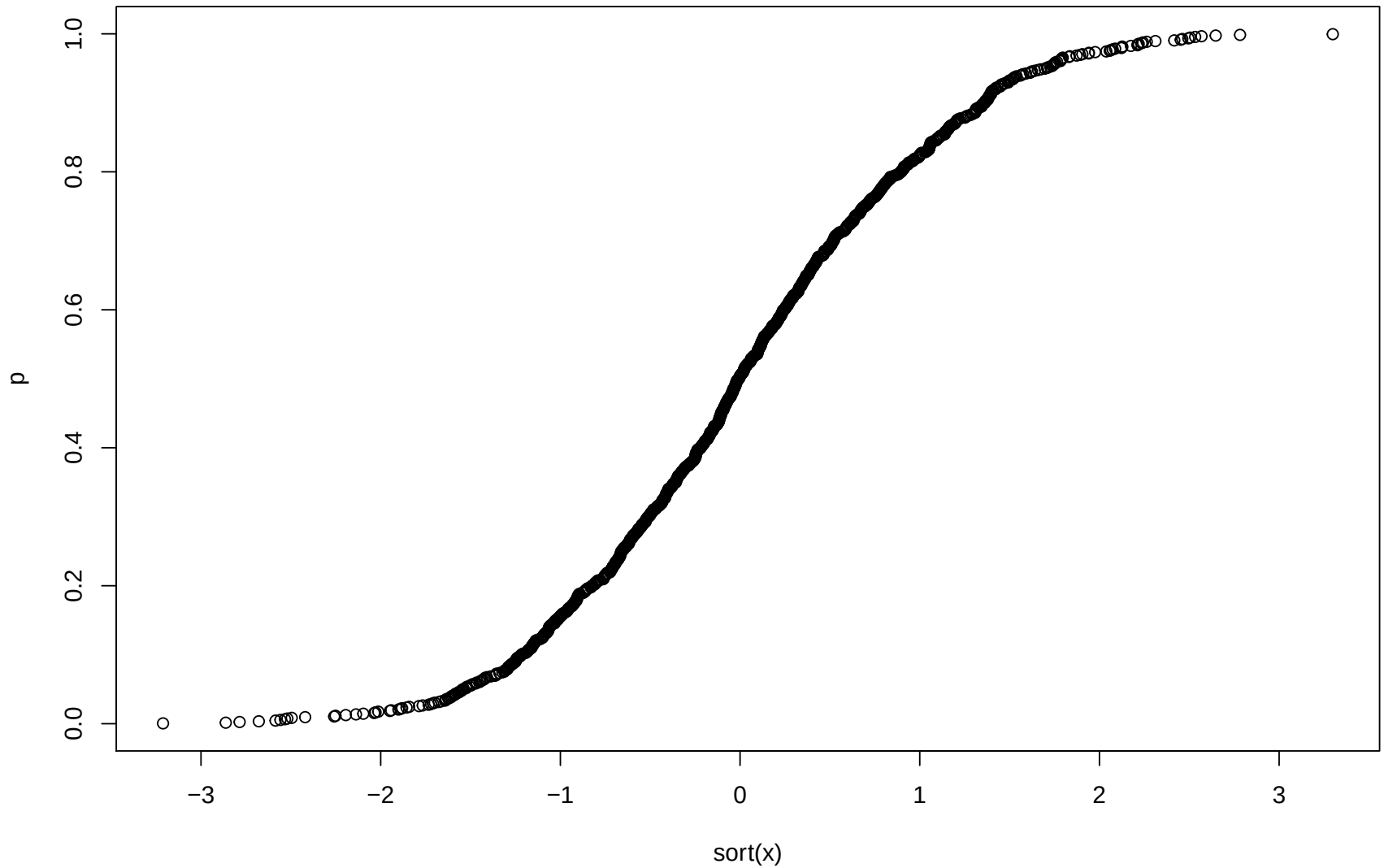
The normality of a set of data can be assessed by the following method.

- Let $r_{(1)} < ... < r_{(N)}$ denote the ordered values of $r_1, ..., r_N$.
- A test of normality for a set of data is to plot the ordered values $r_{(i)}$ of the data versus $p_i = (i - 0.5)/N$.
- If the plot has the same S-shape as the normal CDF then this is evidence that the data come from a normal distribution.

Normal Quantile-Quantile Plots

- A plot of $r_{(i)}$ vs. $p_i = (i - 0.5)/N, i = 1, ..., N$ for a random sample of 1000 simulated from a $N(0, 1)$.

```
N <- 1000;x <- rnorm(N);p <- ((1:N)-0.5)/N
plot(sort(x),p)
```



Normal Quantile-Quantile Plots

- It can be shown that $\Phi(r_i)$ has a uniform distribution on $[0, 1]$.
- This implies that $E(\Phi(r_{(i)})) = i/(N+1)$ (this is the expected value of the *jth* order statistic from a uniform distribution over $[0, 1]$.
- This implies that the N points $(p_i, \Phi(r_{(i)}))$ should fall on a straight line.
- Now apply the Φ^{-1} transformation to the horizontal and vertical scales. The N points

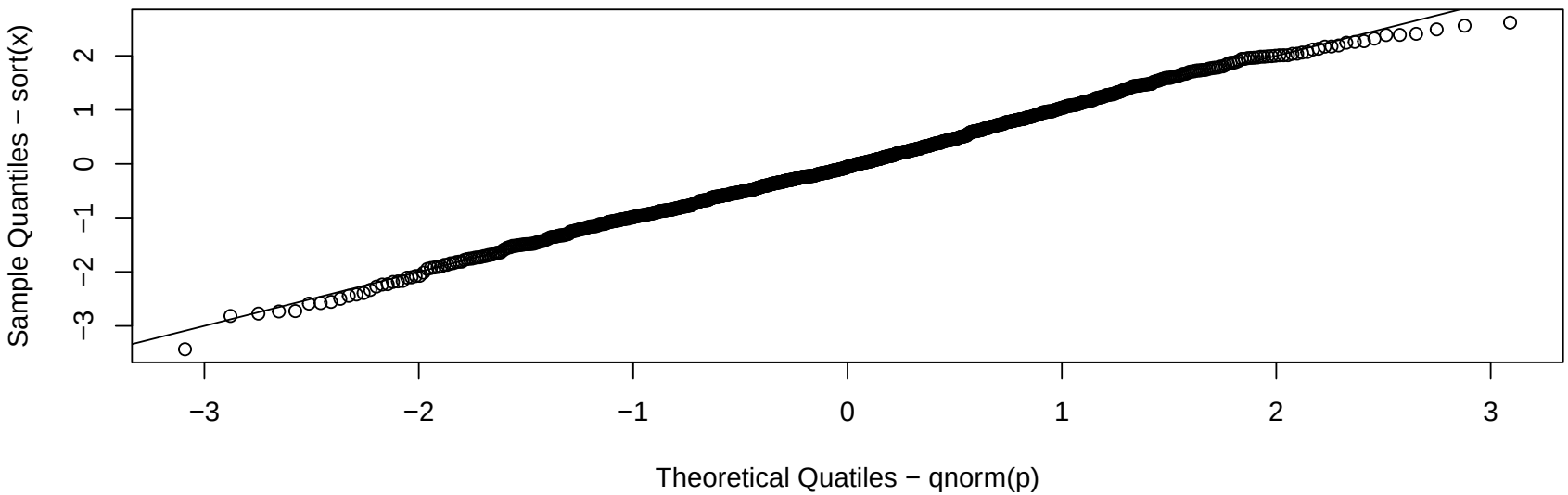
$$\left(\Phi^{-1}(p_i), r_{(i)}\right),$$

  form the normal probability plot of $r_1, ..., r_N$.
- If $r_1, ..., r_N$ are generated from a normal distribution then a plot of the points $\left(\Phi^{-1}(p_i), r_{(i)}\right), i = 1, ..., N$ should be a straight line.

Normal Quantile-Quantile Plots

In R `qnorm()` is Φ^{-1}.

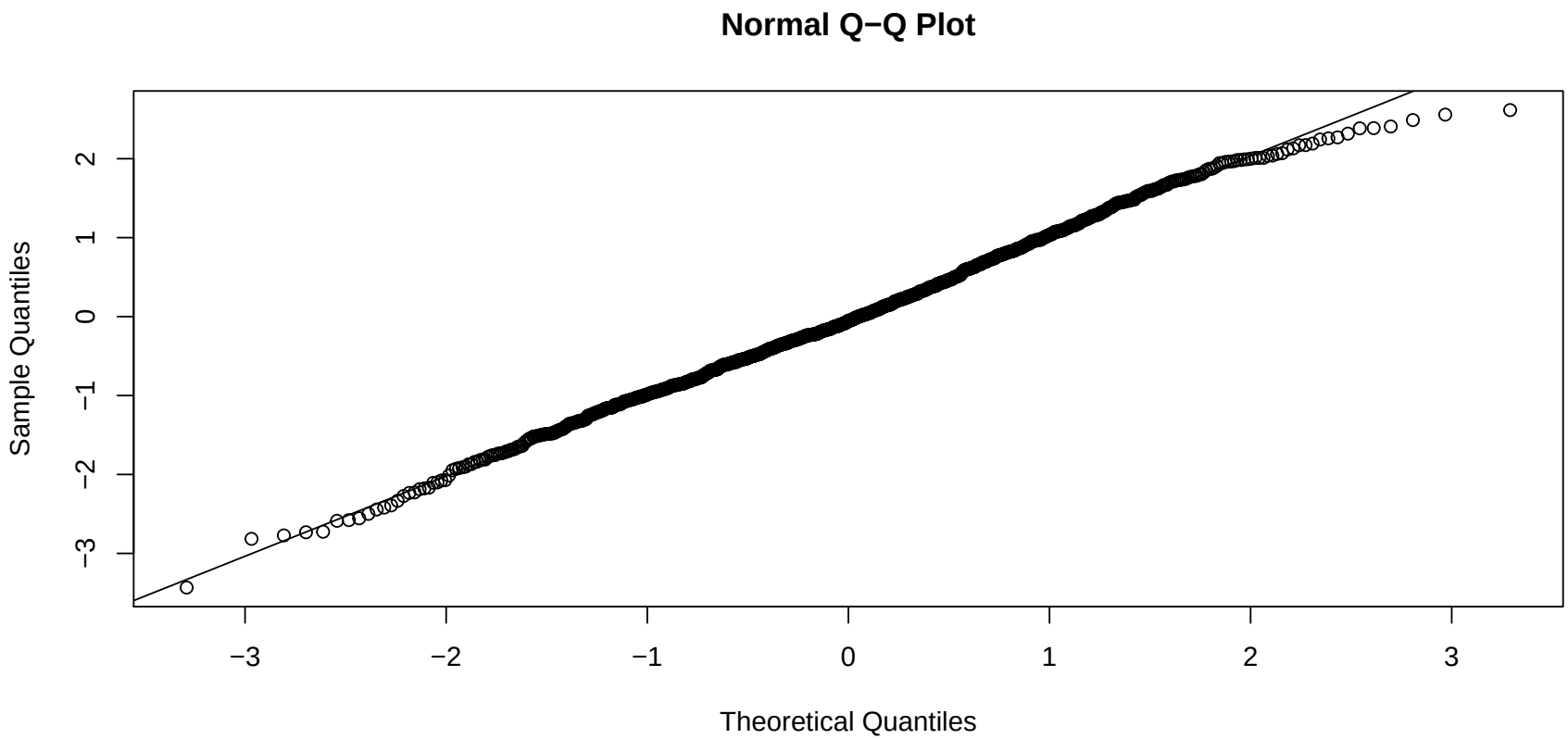
```
set.seed(2503)
N <- 1000
x <- rnorm(N)
p <- (1:N) / (N + 1) # the vector i/(N+1), i = 1,...,N.
plot(qnorm(p), sort(x),
     ylab = "Sample Quantiles - sort(x)",
     xlab = "Theoretical Quatiles - qnorm(p)")
abline(a = 0, b = 1) # add the line y=x
```



Normal Quantile-Quantile Plots

We usually use the built in function `qqnorm()` (and `qqline()` to add a straight line for comparison) to generate normal Q-Q plots. Note that R uses a slightly more general version of quantile ($p_i = (1-a)/(N+(1-a)-a)$, where $a = 3/8$, if $N \leq 10$, $a = 1/2$, if $N > 10$.

```
qqnorm(x);qqline(x)
```



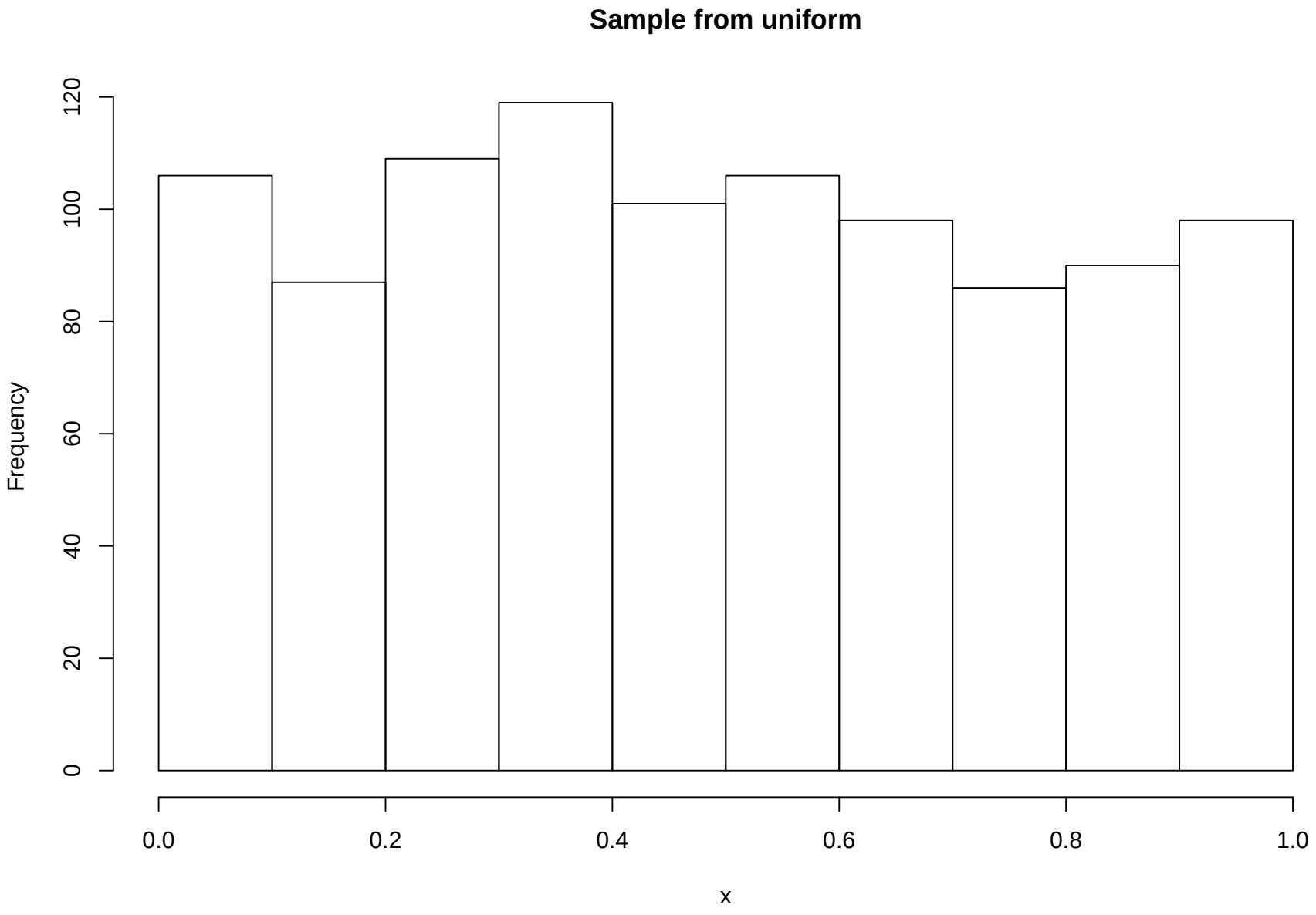
Normal Quantile-Quantile Plots

A marked (systematic) deviation of the plot from the straight line would indicate that:

1. The normality assumption does not hold.
2. The variance is not constant.

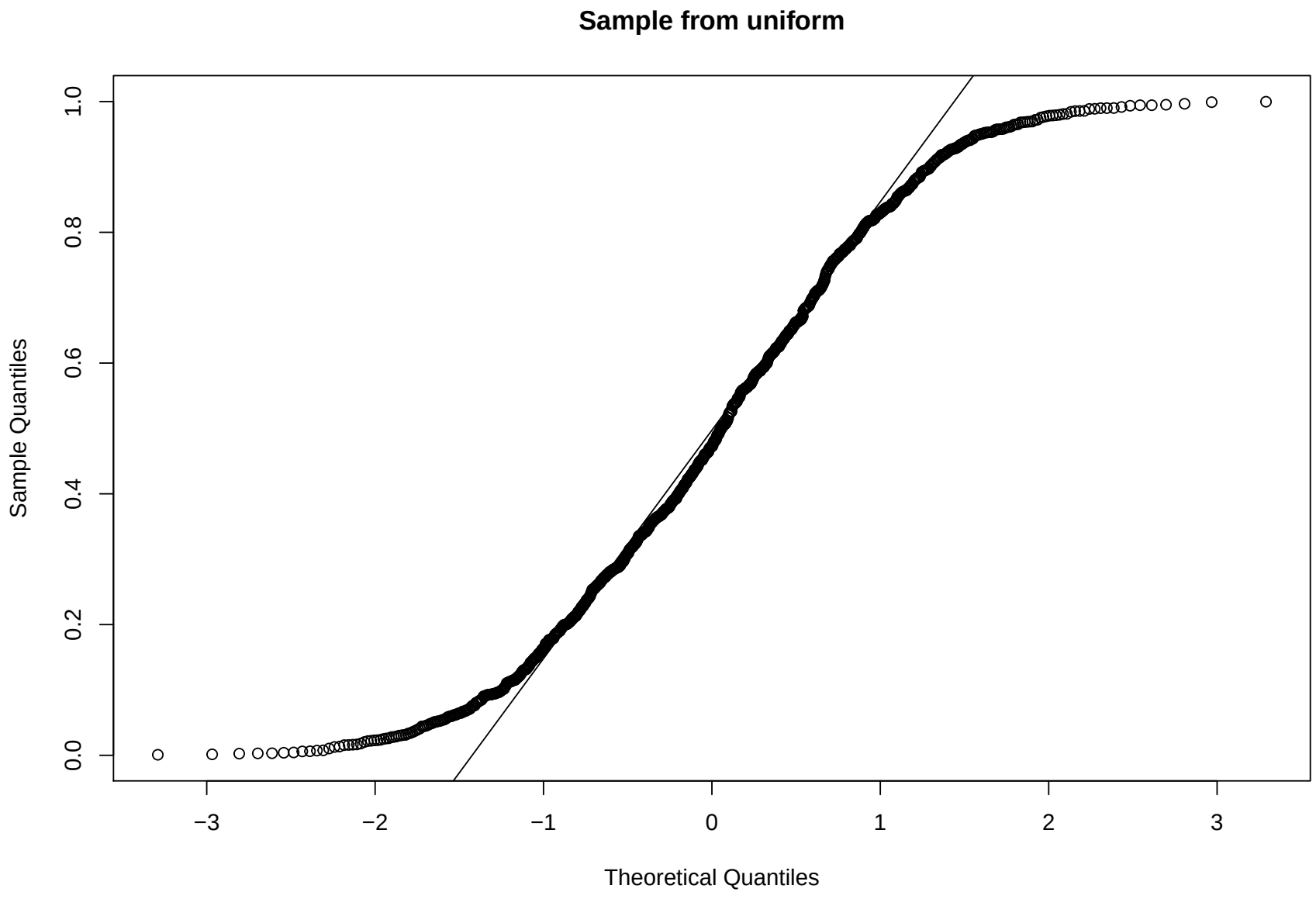
Normal Quantile-Quantile Plots

```
x <- runif(1000)
hist(x,main = "Sample from uniform")
```



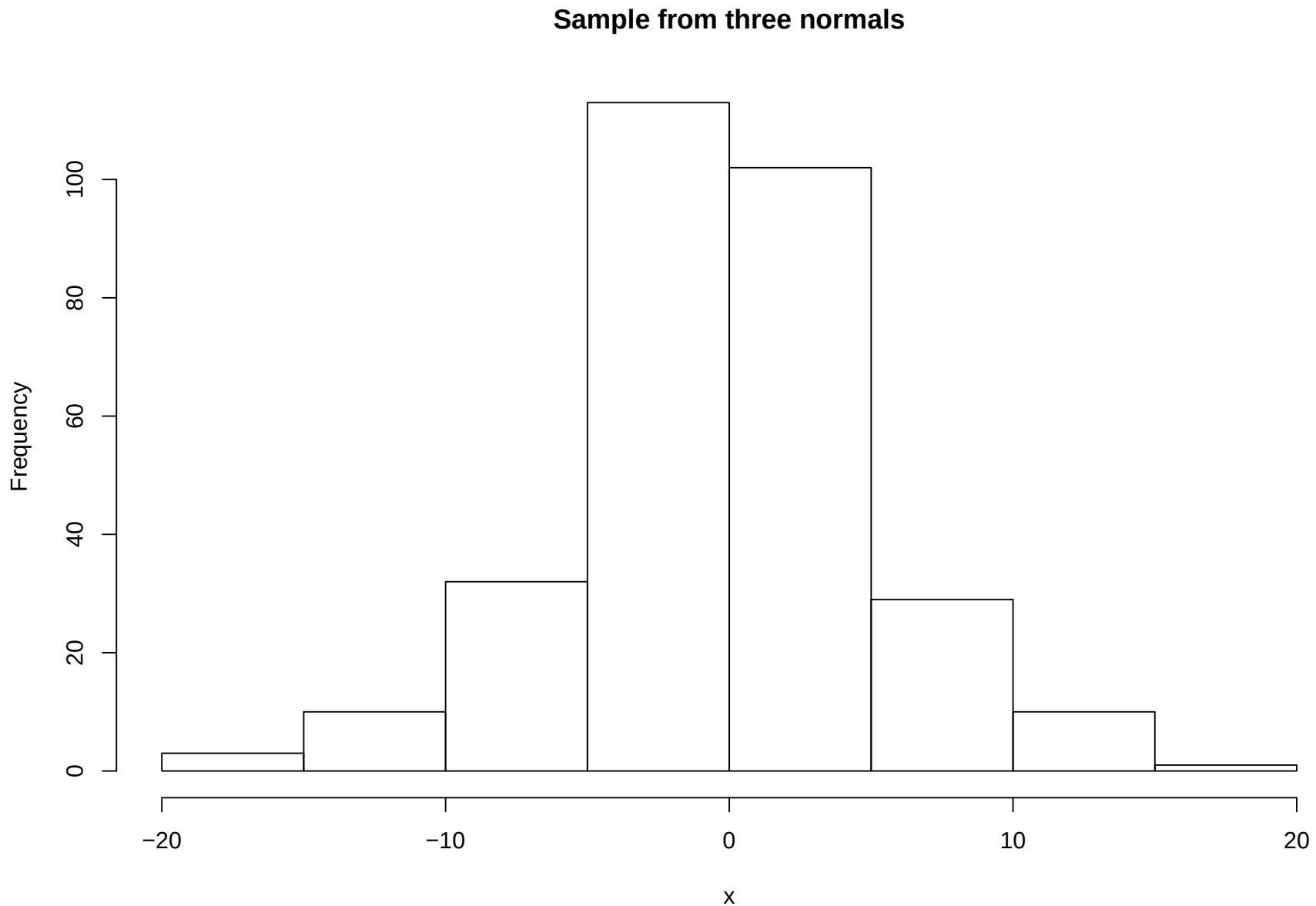
Normal Quantile-Quantile Plots

```
qqnorm(x,main = "Sample from uniform");qqline(x)
```



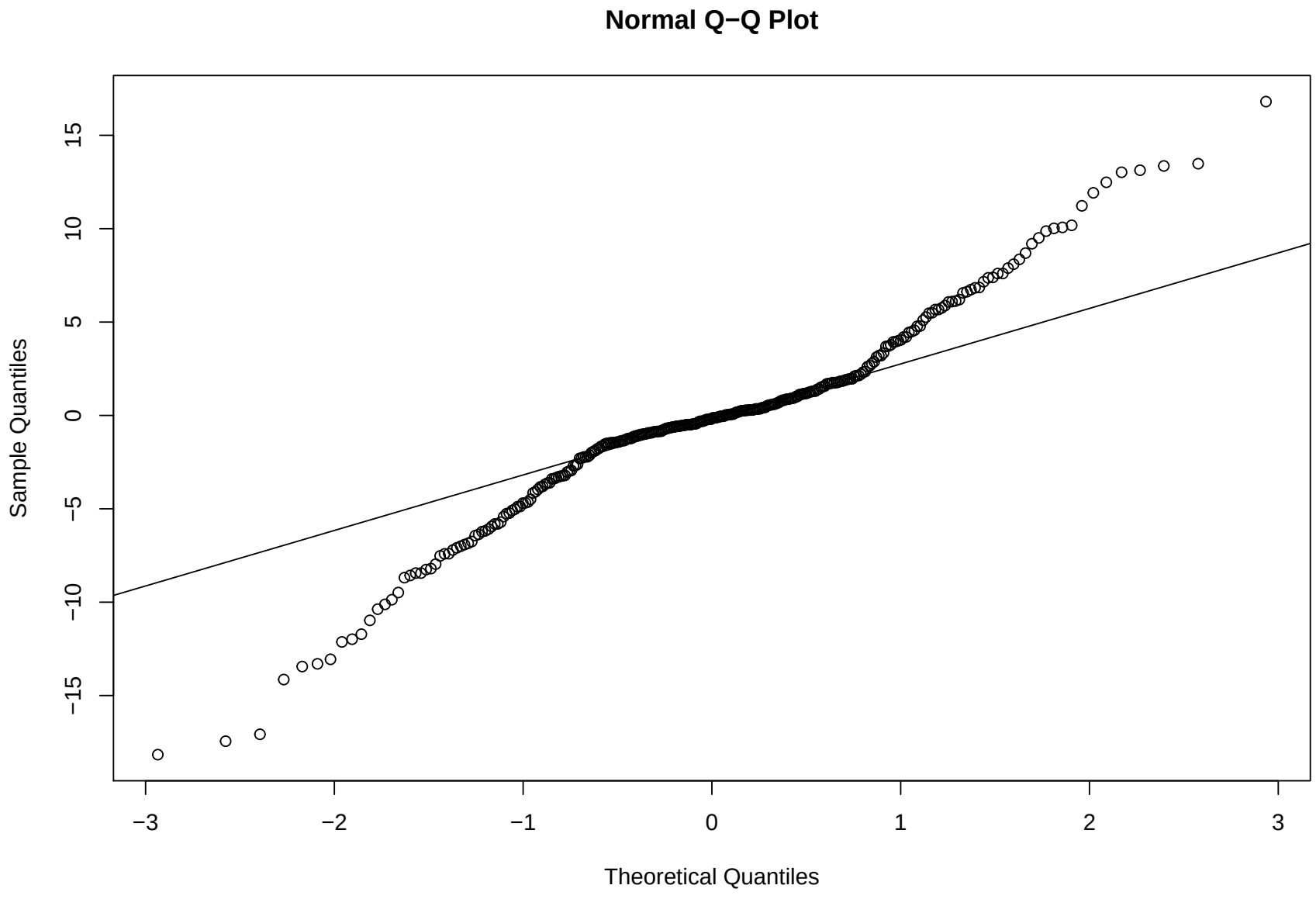
Normal Quantile-Quantile Plots

```
x1 <- rnorm(100,mean = 0,sd = 1);x2 <- rnorm(100,mean = 0,sd = 5)
x3 <- rnorm(100,mean = 0,sd = 8); x <- c(x1,x2,x3)
hist(x,main = "Sample from three normals")
```



Normal Quantile-Quantile Plots

```
qqnorm(x);qqline(x)
```



Normal plots in factorial experiments

- A major application is in factorial designs where the $r(i)$ are replaced by ordered factorial effects.
- Let $\hat{\theta}_{(1)} < \hat{\theta}_{(2)} < \cdots < \hat{\theta}_{(N)}$ be N ordered factorial estimates.
- If we plot

$$\hat{\theta}_{(i)} \text{ vs. } \Phi^{-1}(p_i).\ i = 1, ..., N.$$

then factorial effects $\hat{\theta}_i$ that are close to 0 will fall along a straight line. Therefore, points that fall off the straight line will be declared significant.

Normal plots in factorial experiments

The rationale is as follows:

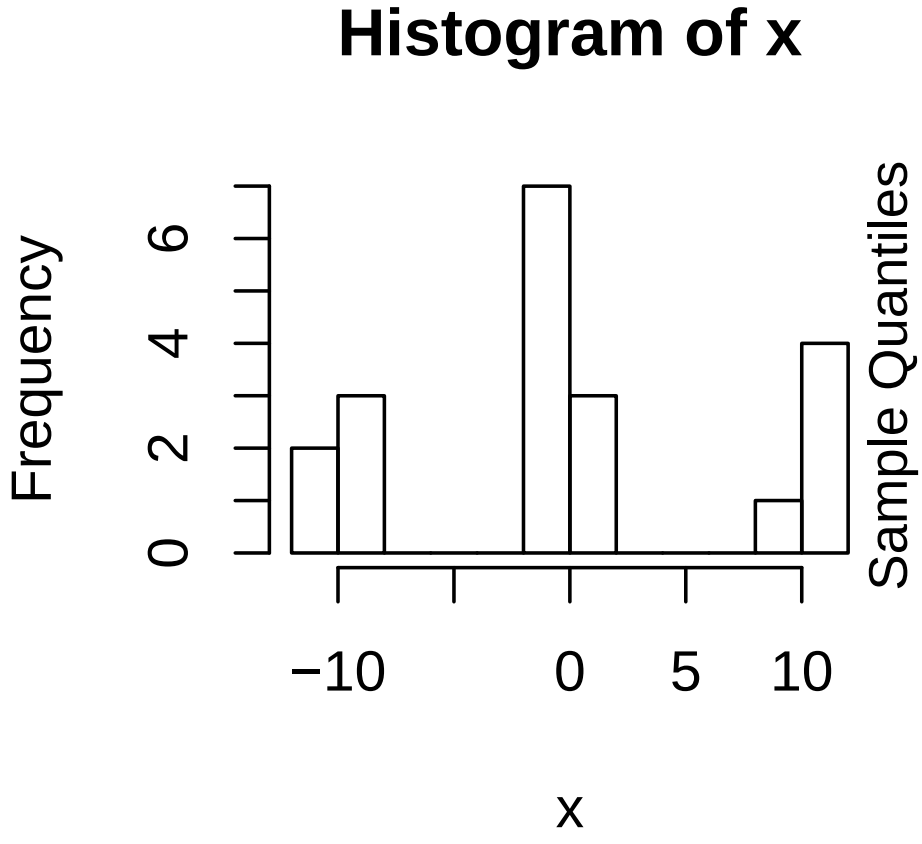
1. Assume that the estimated effects $\hat{\theta}_i$ are $N(\theta, \sigma)$ (estimated effects involve averaging of N observations and CLT ensures averages are nearly normal for N as small as 8).
2. If $H_0 : \theta_i = 0$, $i = 1, ..., N$ is true then all the estimated effects will be zero.
3. The resulting normal probability plot of the estimated effects will be a straight line.
4. Therefore, the normal probability plot is testing whether all of the estimated effects have the same distribution (i.e. same means).

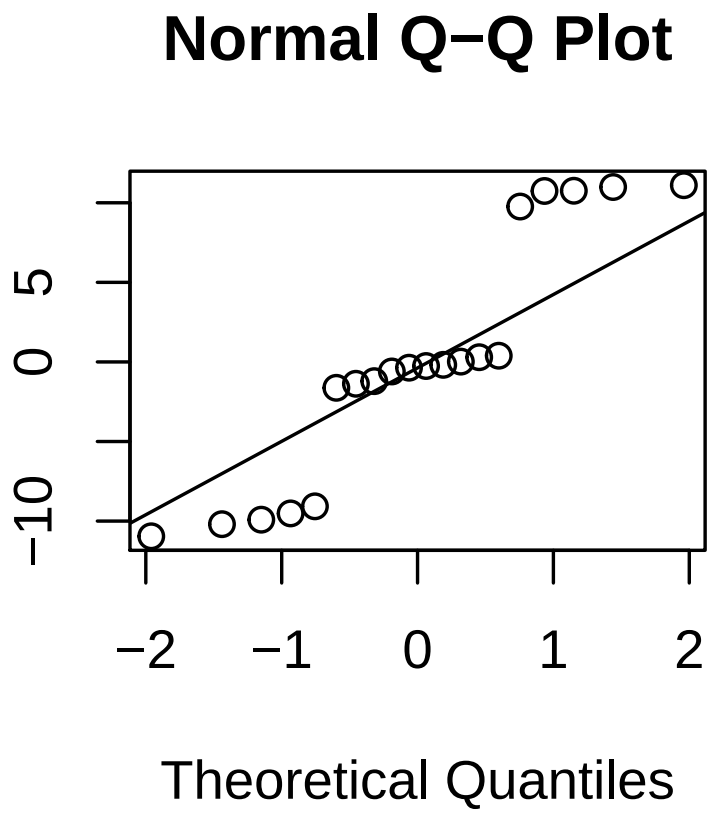
- When some of the effects are nonzero the corresponding estimated effects will tend to be larger and fall off the straight line.

Normal plots in factorial experiments

Positive effects fall above the line and negative effects fall below the line.

```
set.seed(10)
x1 <- rnorm(10,0,1); x2 <- rnorm(5,10,1); x3 <- rnorm(5,-10,1)
x <- c(x1,x2,x3)
hist(x, breaks = 10)
qqnorm(x); qqline(x)
```





Example - 2^3 design for studying a chemical reaction

A process development experiment studied four factors in a 2^4 factorial design.

- amount of catalyst charge **1**,
- temperature **2**,
- pressure **3**,
- concentration of one of the reactants **4**.
- The response y is the percent conversion at each of the 16 run conditions. The design is shown below.

Example - 2^4 design for studying a chemical reaction

x1	x2	x3	x4	conversion
-1	-1	-1	-1	70
1	-1	-1	-1	60
-1	1	-1	-1	89
1	1	-1	-1	81
-1	-1	1	-1	69
1	-1	1	-1	62
-1	1	1	-1	88
1	1	1	-1	81
-1	-1	-1	1	60
1	-1	-1	1	49
-1	1	-1	1	88
1	1	-1	1	82
-1	-1	1	1	60
1	-1	1	1	52
-1	1	1	1	86
1	1	1	1	79

The design is not replicated so it's not possible to estimate the standard errors of the factorial effects.

Example - 2^4 design for studying a chemical reaction

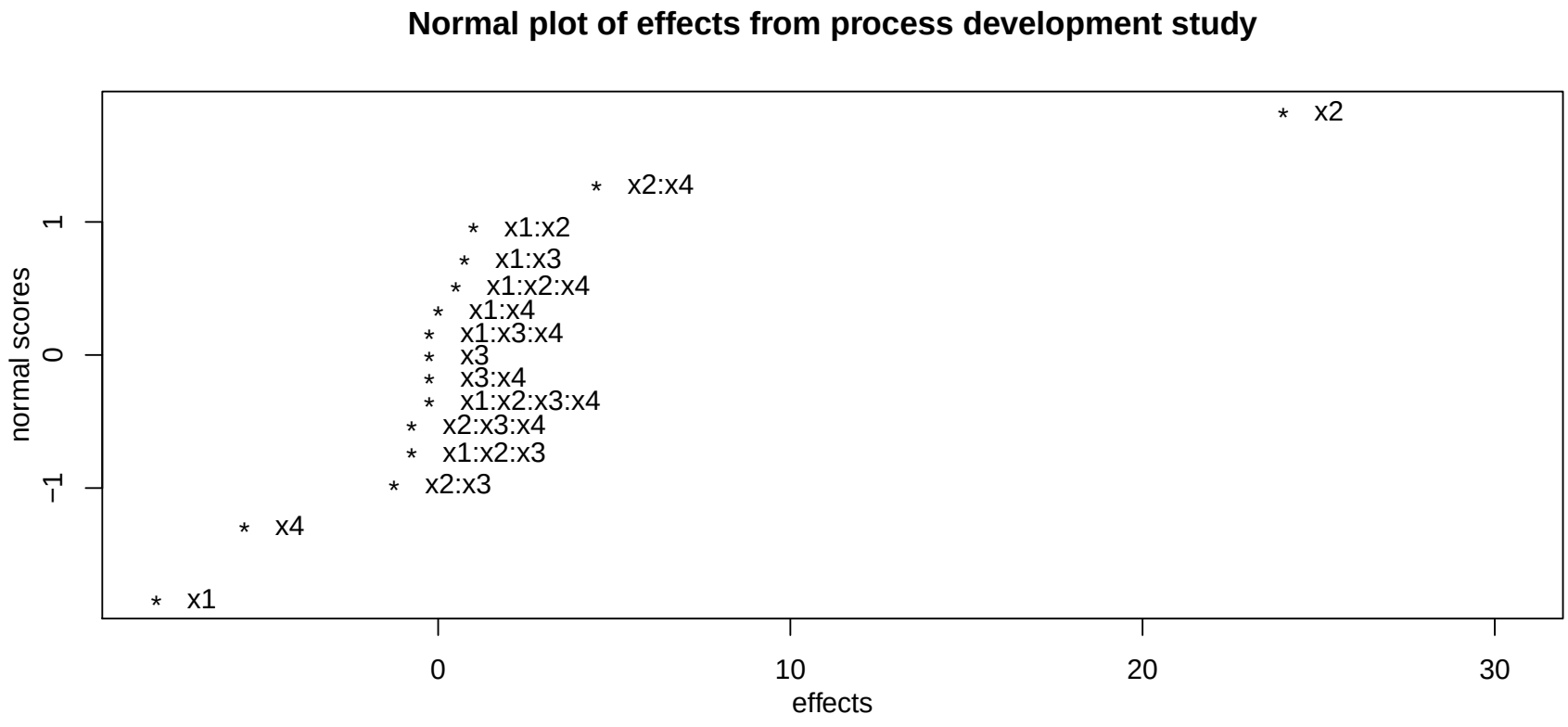
```
fact1 <- lm(conversion~x1*x2*x3*x4,data=tab0510a)
round(2*fact1$coefficients,2)
```

```
(Intercept)          x1          x2          x3          x4       x1:x2
     144.50       -8.00       24.00       -0.25       -5.50        1.00
      x1:x3       x2:x3       x1:x4       x2:x4       x3:x4    x1:x2:x3
       0.75       -1.25        0.00        4.50       -0.25       -0.75
   x1:x2:x4    x1:x3:x4    x2:x3:x4 x1:x2:x3:x4
       0.50       -0.25       -0.75       -0.25
```

Example - 2^4 design for studying a chemical reaction

A normal plot of the factorial effects is obtained by using the function `DanielPlot()` in the `FrF2` library.

```
library(FrF2)
DanielPlot(fact1, autolab=F,
           main="Normal plot of effects from process development study")
```



Example 1

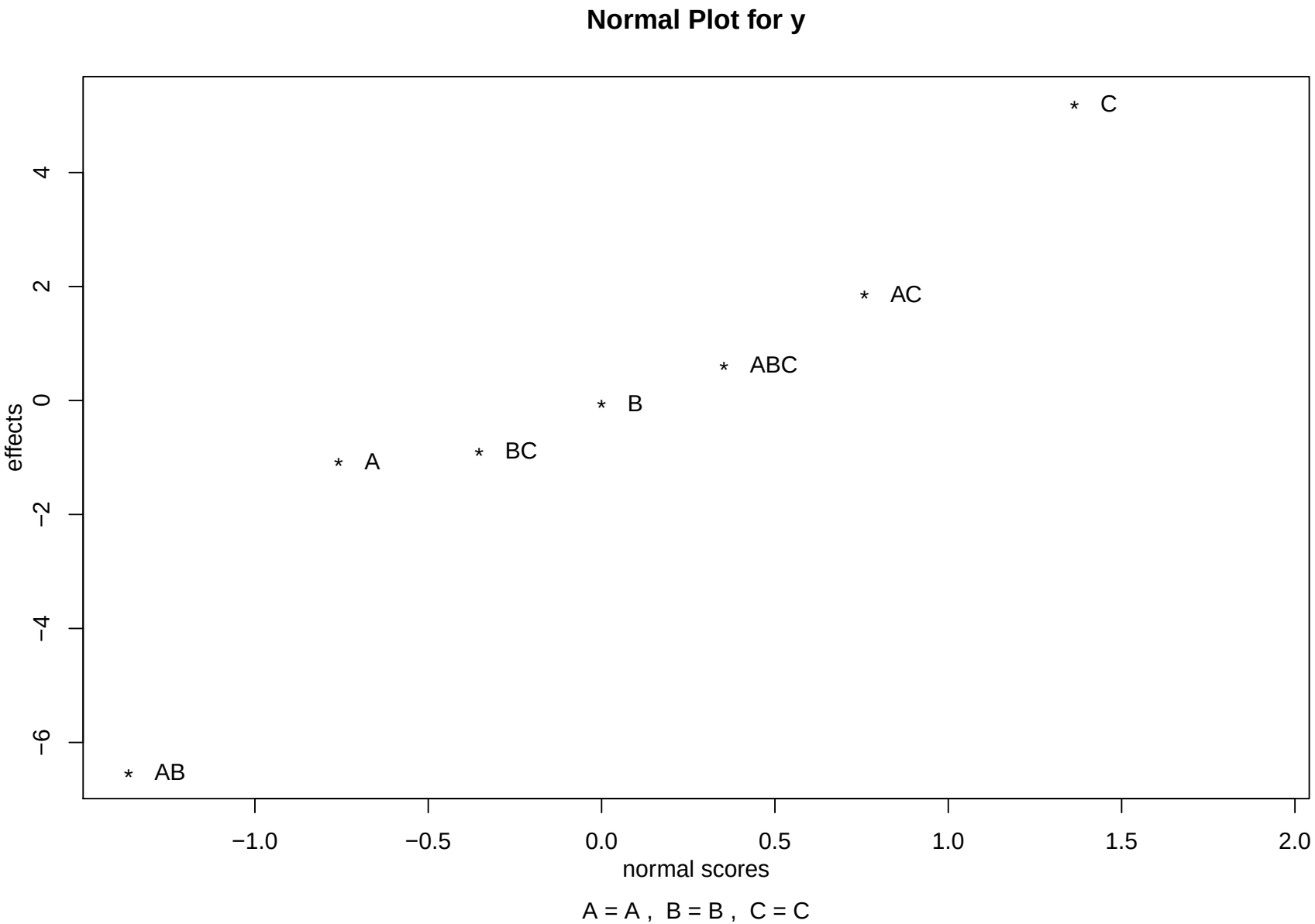
Which effects are not explained by chance?

```
##
## Call:
## lm.default(formula = y ~ A * B * C, data = dat)
##
## Residuals:
## ALL 8 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90361         NA      NA       NA
## A1          -0.52770         NA      NA       NA
## B1          -0.01836         NA      NA       NA
## C1           2.60717         NA      NA       NA
## A1:B1       -3.25821         NA      NA       NA
## A1:C1        0.93739         NA      NA       NA
## B1:C1       -0.43695         NA      NA       NA
## A1:B1:C1     0.31787         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

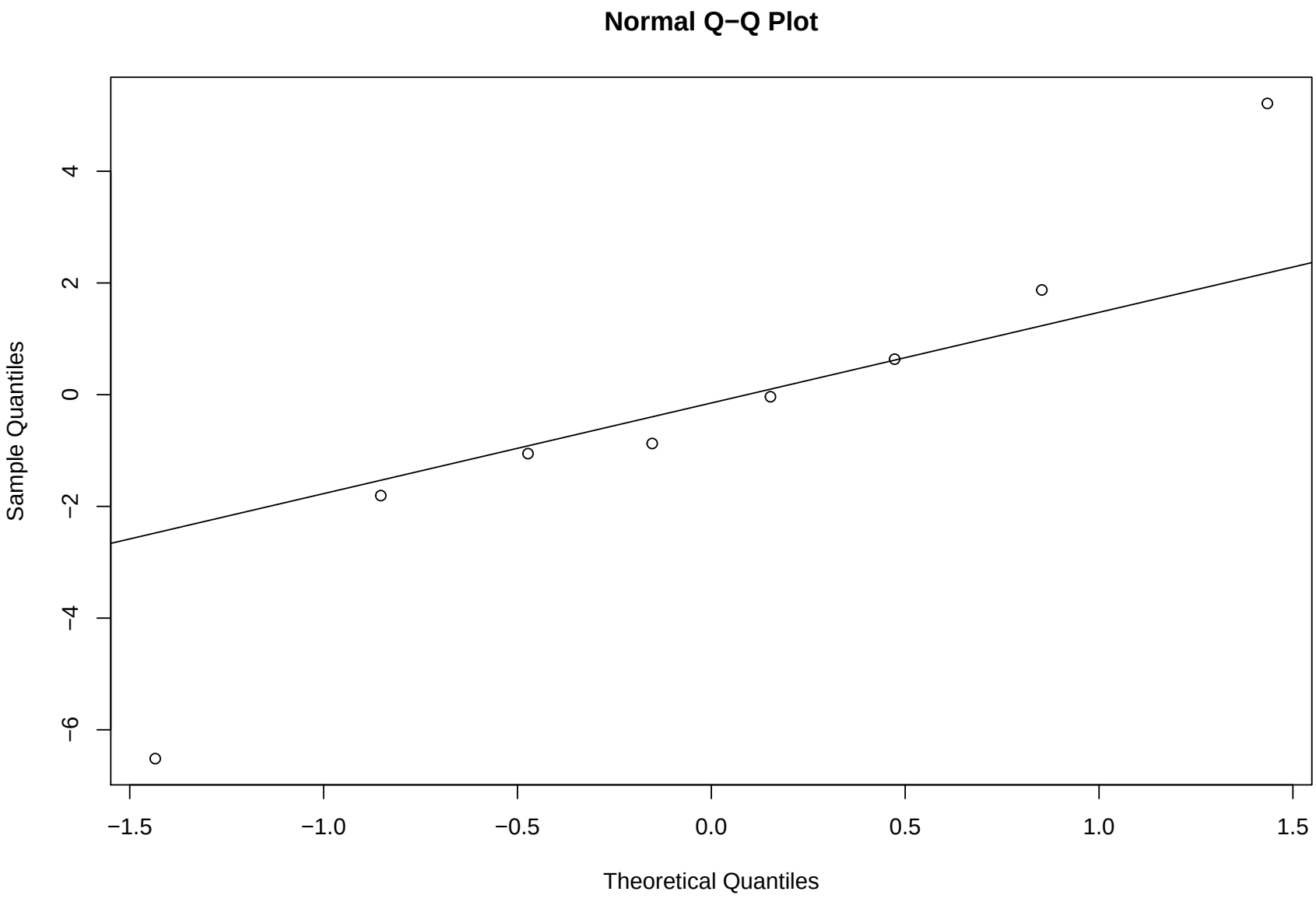
Example 1

Which effects are not explained by chance according to the normal plot?

```
FrF2::DanielPlot(mod1,code=TRUE,autolab=F,datax=F)
```



Example 1


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−6
−4
−2
0
2
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

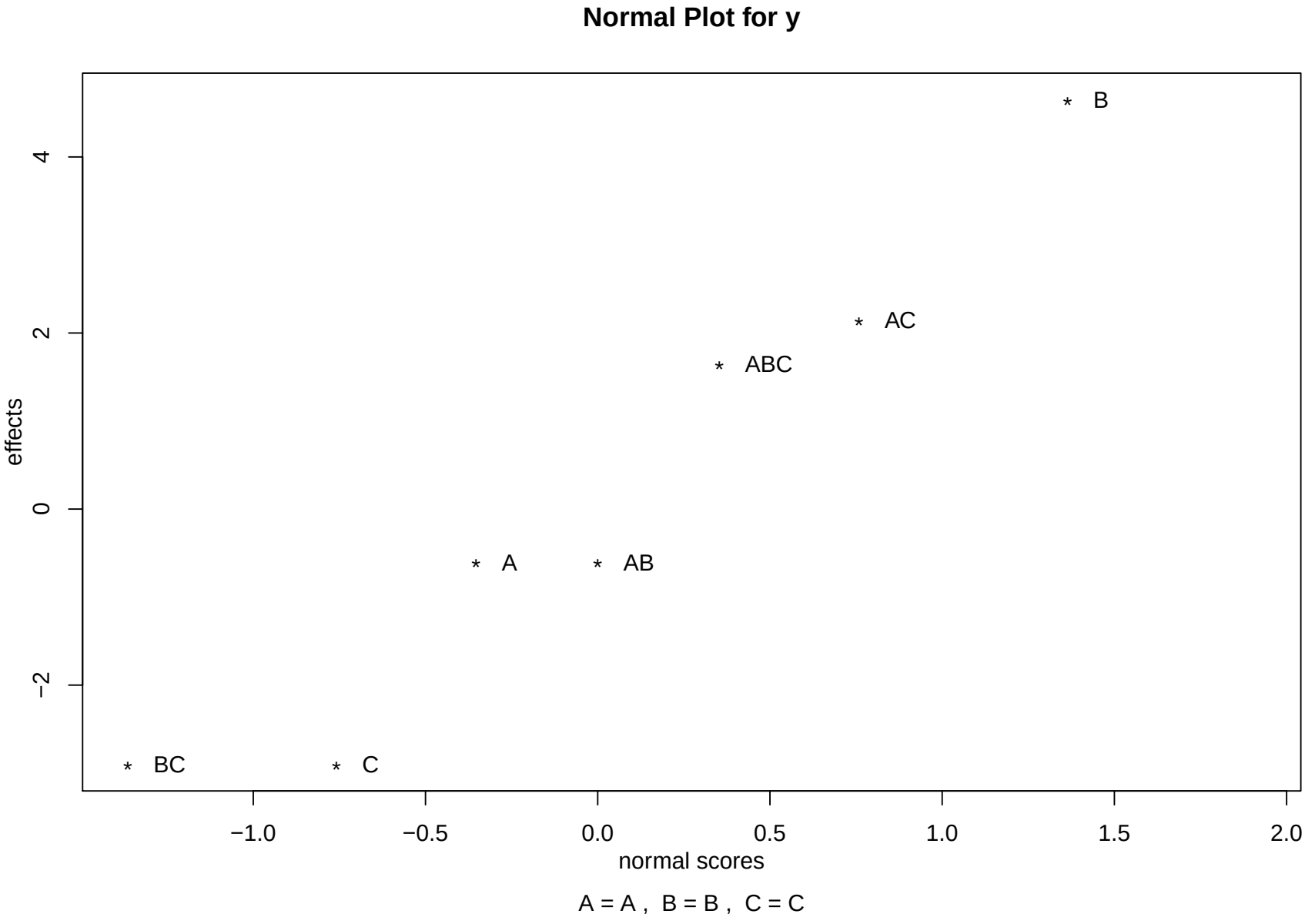
Example 2

Which effects are not explained by chance?

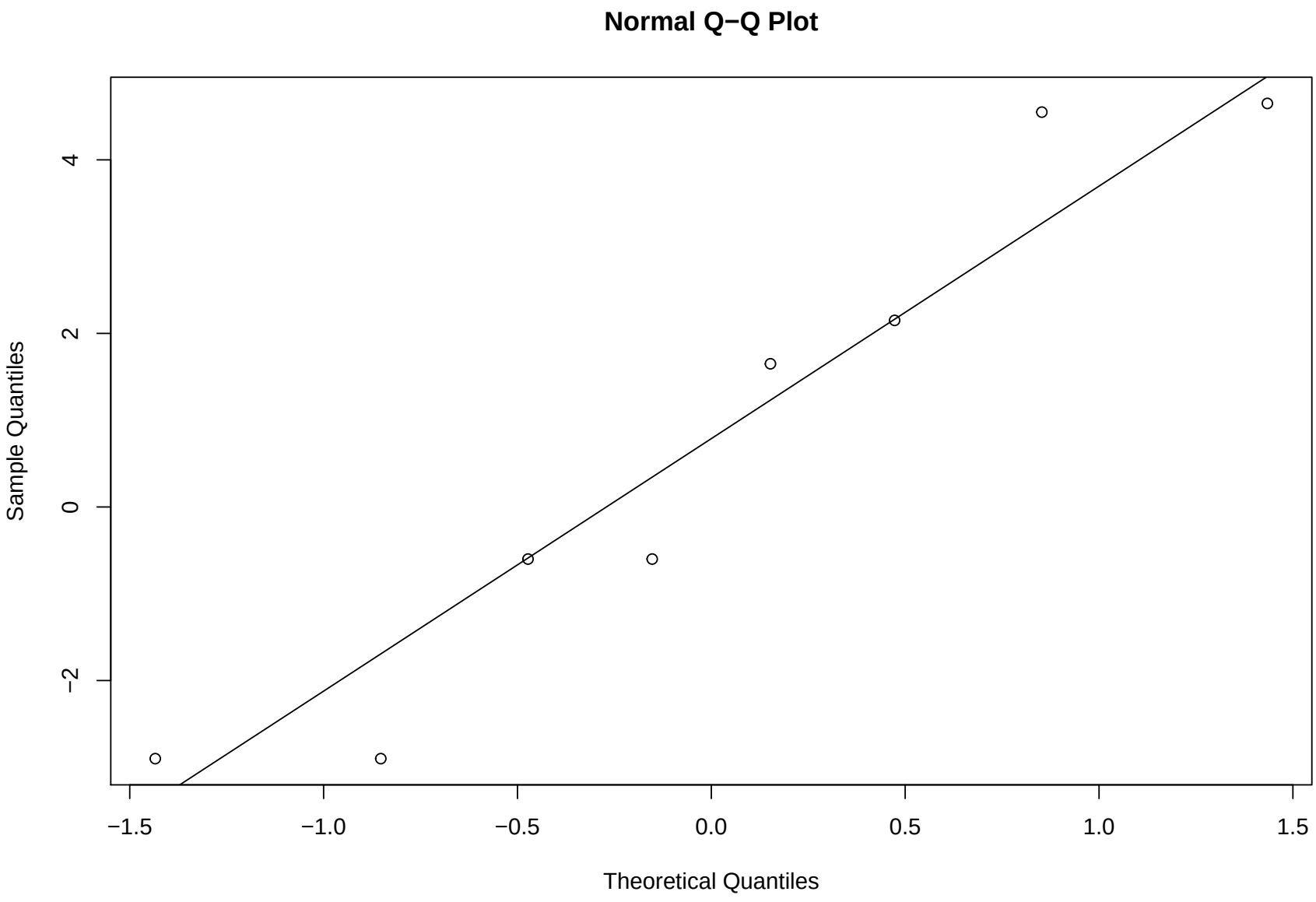
```
##
## Call:
## lm.default(formula = y ~ A * B * C, data = dat)
##
## Residuals:
## ALL 8 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.275         NA      NA       NA
## A1            -2.150         NA      NA       NA
## B1             0.300         NA      NA       NA
## C1            -0.950         NA      NA       NA
## A1:B1         -0.125         NA      NA       NA
## A1:C1          1.125         NA      NA       NA
## B1:C1         -1.575         NA      NA       NA
## A1:B1:C1       1.500         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

Example 2

Which effects are not explained by chance according to the normal plot?



Example 2


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
0
2
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Half-Normal Plots

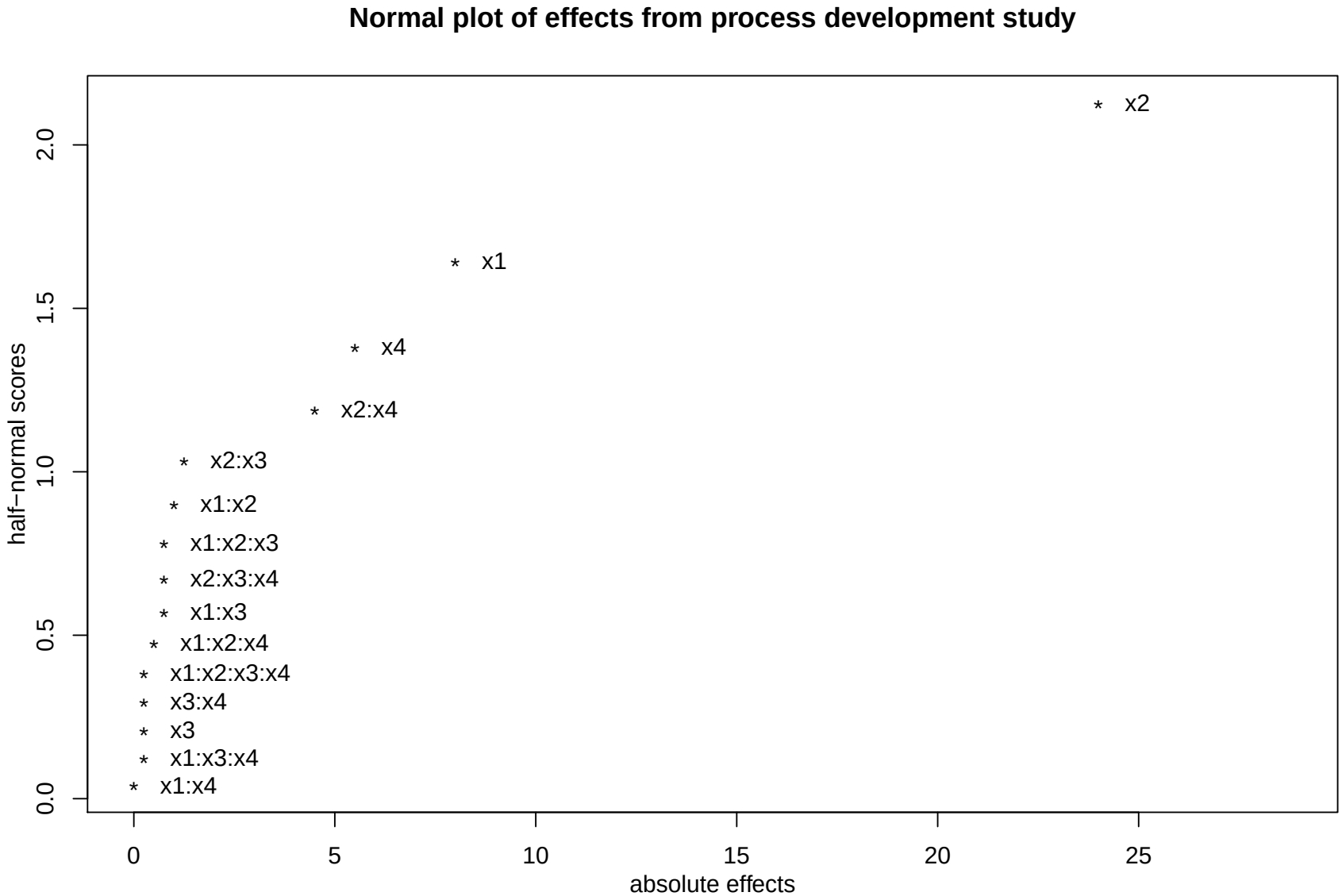
- A related graphical method is called the half-normal probability plot.
- Let
$$\left|\hat{\theta}\right|_{(1)} < \left|\hat{\theta}\right|_{(2)} < \cdots < \left|\hat{\theta}\right|_{(N)}.$$
denote the ordered values of the unsigned factorial effect estimates.
- Plot them against the coordinates based on the half-normal distribution - the absolute value of a normal random variable has a half-normal distribution.
- The half-normal probability plot consists of the points
$$\left|\hat{\theta}\right|_{(i)} \text{ vs. } \Phi^{-1}(0.5 + 0.5[i - 0.5]/N).\ i = 1, ..., N.$$

Half-Normal Plots

- An advantage of this plot is that all the large estimated effects appear in the upper right hand corner and fall above the line.
- The half-normal plot for the effects in the process development example is can be obtained with `DanielPlot()` with the option `half=TRUE`.

Half-Normal Plots - 2^4 design for studying a chemical reaction

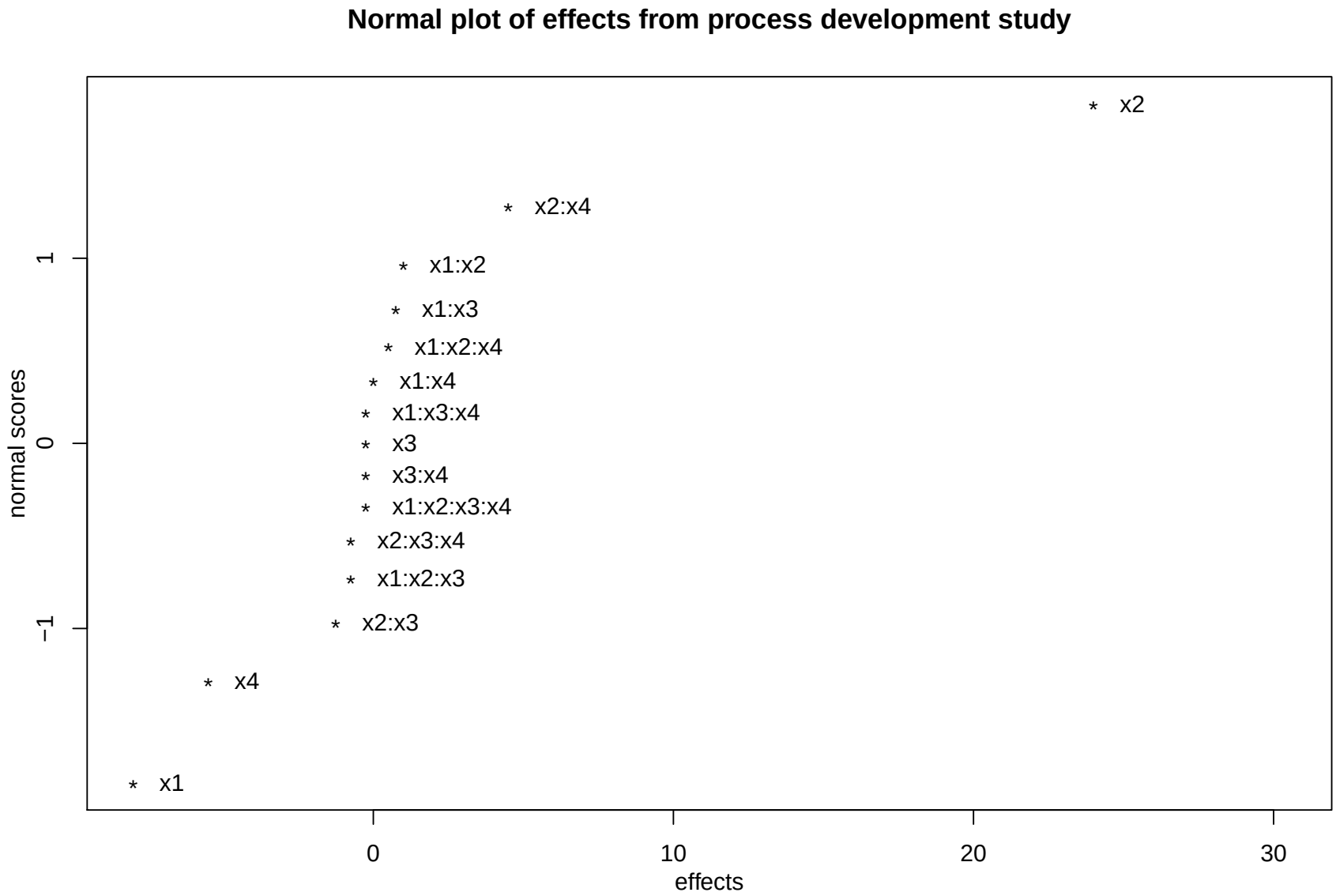
```
library(FrF2)
DanielPlot(fact1,half=TRUE,autolab=F,
           main="Normal plot of effects from process development study")
```



Half-Normal Plots - 2^4 design for studying a chemical reaction

Compare with full Normal plot.

```
library(FrF2)
DanielPlot(fact1,half=F,autolab=F,
           main="Normal plot of effects from process development study")
```



Multiple Comparisons

Suppose that experimental units were randomly assigned to three treatment groups. The hypothesis of intrest is:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_1 : \mu_i \neq \mu_j.$$

Now, suppose that we reject H_0 at level α. Which pairs of means are significantly different from each other at level α? There are $\binom{3}{2} = 3$ possibilites.

1. $\mu_1 \neq \mu_2$
2. $\mu_1 \neq \mu_3$
3. $\mu_2 \neq \mu_3$

Multiple Comparisons

Suppose that $k = 3$ separate (independent) hypothesis tests at level α tests are conducted:

$$H_{0_k} : \mu_i = \mu_j \text{ vs. } H_{1_k} : \mu_i \neq \mu_j,$$

When H_0 is true, $P(\text{reject } H_0) = \alpha \Rightarrow 1 - P(\text{do not reject } H_0) = 1 - (1 - \alpha)$.

So, if H_0 is true then

$$P\left(\text{reject at least one } H_{0_k}\right) = 1 - P\left(\text{do not reject any } H_{0_k}\right)$$

This is the same as

$$1 - P(\text{do not reject } H_{0_1} \text{ and do not reject } H_{0_2} \text{ and do not reject } H_{0_3})$$

or since the hypotheses are independent

$$1 - P(\text{do not reject } H_{0_1})\, P(\text{do not reject } H_{0_2})\, P(\text{do not reject } H_{0_3}) = 1 - (1 - \alpha)^3$$

If $\alpha = 0.05$ then the probability that at least one H_0 will be falsely rejected is $1 - (1 - .05)^3 = 0.14$, which is almost three times the type I error rate.

Multiple Comparisons

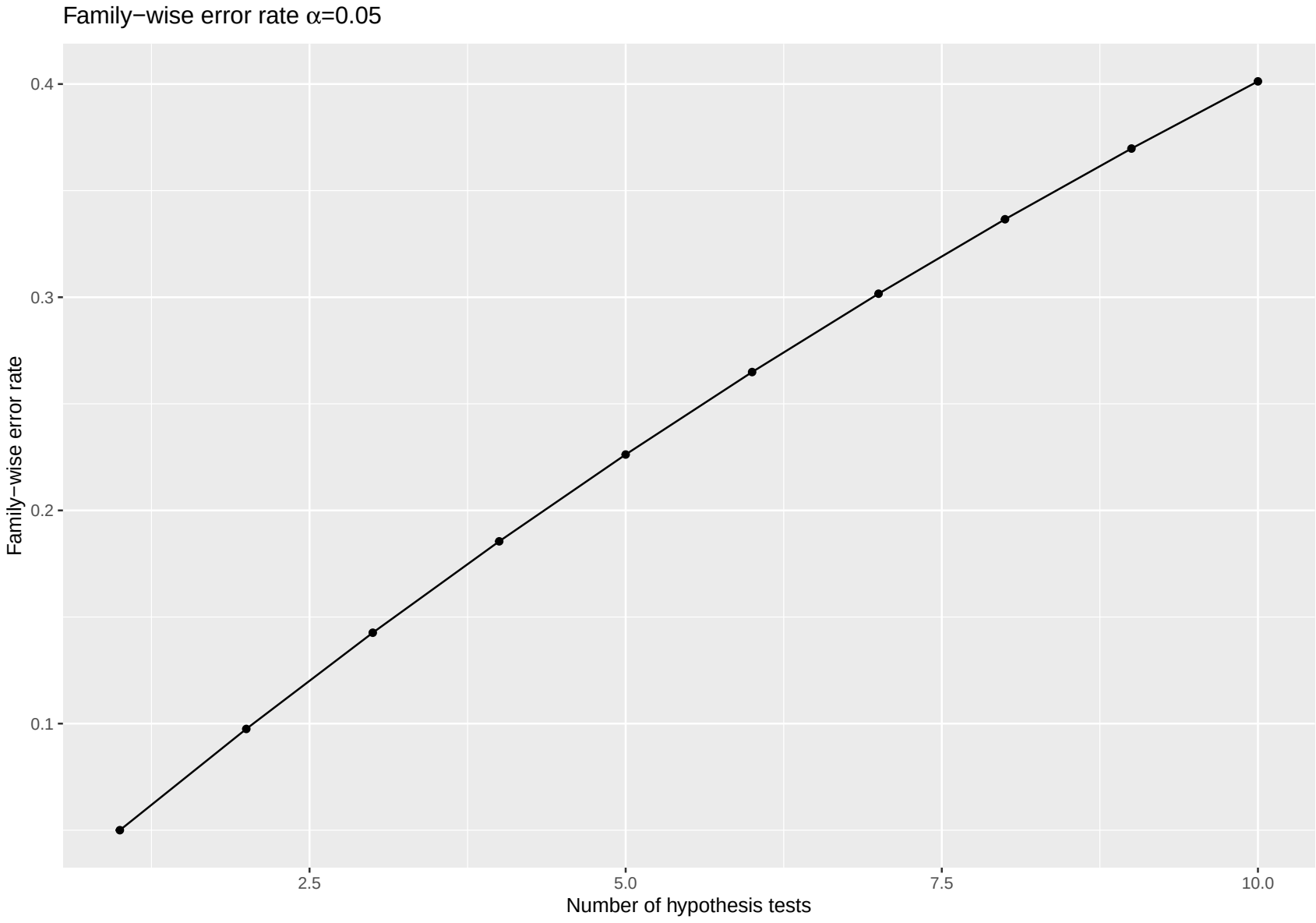
A clinical trial comparing four treatment means using an ANOVA model at the 5% level found a significant F test. If all pairs of treatment means are compared then the probability of falsely declaring that at least one pair of treatment means is significantly different is:

Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A or B**

less than or equal to 0.05	**A**
greater than 0.05	**B**

Multiple Comparisons



Multiple Comparisons

In general if

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs. } H_1 : \mu_i \neq \mu_j.$$

If c independent hypotheses are conducted then the probability

$$P\left(\text{reject at least one } H_{0_k}\right) = 1 - (1 - \alpha)^c$$

is called the **family-wise error rate**.

The **pairwise error rate** is $P\left(\text{reject } H_{0_k}\right) = \alpha$ for any c.

The Multiple Comparisons Problem

- The multiple comparison problem is that multiple hypotheses are tested level α which increases the probability that at least one of the hypotheses will be falsely rejected (family-wise error rate).
- If treatment means are significantly different from the ANOVA F test then researchers will usually want to explore where the differences lie.
- Is it appropriate to test for differences looking at all pairwise comparisons?
- Testing all possible pairs increases the type I error rate.
- This means the chance that there is a higher probability, beyond the pre-stated type I error rate (e.g. 0.05), that that a significant difference is detected when the truth is that no difference exists.

Example

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

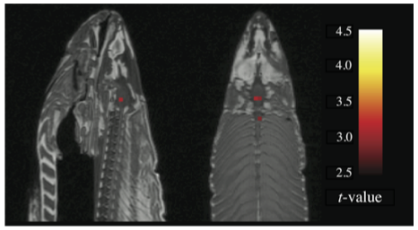
Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

GLM RESULTS

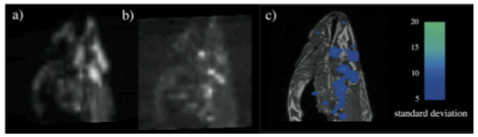


A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

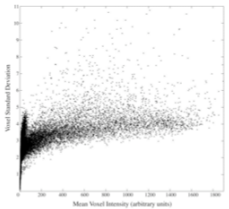
VOXELWISE VARIABILITY



To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.



The Bonferroni Method

To test for the difference between the *ith* and *jth* treatments, it is common to use the two-sample t test. The two-sample t statistic is

$$t_{ij} = \frac{\bar{y_j}. - \bar{y_i}.}{\hat{\sigma}\sqrt{1/n_j + 1/n_i}},$$

where $\bar{y_j}.$ is the average of the n_i observations for treatment j and $\hat{\sigma}$ is $\sqrt{MS_E}$ from the ANOVA table.

Treatments i and j are declared significantly different at level α if

$$|t_{ij}| > t_{N-k,\alpha/2},$$

where $t_{N-k,\alpha/2}$ is the upper $\alpha/2$ percentile of a t_{N-k}.

The Bonferroni Method

The total number of pairs of treatment means that can be tested is

$$c = \binom{k}{2} = \frac{k(k-1)}{2}.$$

The Bonferroni method for testing $H_0 : \mu_i = \mu_j$ vs. $H_0 : \mu_i \neq \mu_j$ rejects H_0 at level α if

$$|t_{ij}| > t_{N-k,\alpha/2c},$$

where c denotes the number of pairs being tested.

The Bonferroni Method

In R the function `pairwise.t.test()` can be used to compute Bonferroni adjusted p-values.

This is illustrated below for the blood coagualtion study.

```
pairwise.t.test(tab0401$y,tab0401$diets,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  tab0401$y and tab0401$diets
##
##   A       B       C
## B 0.00934 -       -
## C 0.00031 0.95266 -
## D 1.00000 0.00934 0.00031
##
## P value adjustment method: bonferroni
```

There are signifciant differences at the 5% level between diets A and B, A and C, B and D, and C and D using the Bonferroni method.

The Bonferroni Method

For comparison the unadjusted p-values are also calculated.

```
pairwise.t.test(tab0401$y,tab0401$diets,p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  tab0401$y and tab0401$diets
##
##   A       B      C
## B 0.0016  -      -
## C 5.2e-05 0.1588 -
## D 1.0000  0.0016 5.2e-05
##
## P value adjustment method: none
```

The significant differences are the same using the unadjusted p-values but the p-values are larger then the p-values adjusted using the Bonferroni method.

The Bonferroni Method

A $100(1-\alpha)\%$ simultaneous confidence interval for c pairs $\mu_i - \mu_j$ is

$$\bar{y_j}. - \bar{y_i}. \pm t_{N-k,\alpha/2c}\hat{\sigma}\sqrt{1/n_j + 1/n_i}.$$

After identifying which pairs are different, the confidence interval quantifies the range of plausible values for the differences.

The Bonferroni Method - coagulation study

The treatment means can be obtained from the table below.

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

The Bonferroni Method - coagulation study

$\hat{\sigma} = \sqrt{MS_E}$ can be obtained from the ANOVA table.

```
anova(lm(y~diets,data=tab0401))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diets      3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The upper $.05/(2 \cdot 6) = 0.004$ percentile of the t_{24-4} can be obtained with the t quantile function in R `qt()`.

```
qt(p = 1-0.004,df = 20)
```

```
## [1] 2.945349
```

The Bonferroni Method - coagulation study

Plugging in these values to the confidence interval formula we can obtain a Bonferroni adjusted 95% confidence interval for $\mu_B - \mu_A$:

$$66 - 61 \pm 2.95\sqrt{5.6}\sqrt{1/6 + 1/6}$$

The lower and upper limits can be calculated in R.

```
66-61 - qt(p = 1-0.004,df = 20)*sqrt(5.6)*sqrt(1/6+1/6) # lower limit
```

```
## [1] 0.9758869
```

```
66-61 + qt(p = 1-0.004,df = 20)*sqrt(5.6)*sqrt(1/6+1/6) # upper limit
```

```
## [1] 9.024113
```

The 95% confidence interval for $\mu_B - \mu_A$ is (0.98, 9.02).

The Tukey Method

- The only difference between the Tukey and Bonferroni methods is in the choice of the critical value.
- Treatments i and j are declared significantly different at level α if

$$|t_{ij}| > \frac{1}{\sqrt{2}} q_{k,N-k,\alpha},$$

- t_{ij} is the observed value of the two-sample t-statistic
- $q_{k,N-k,\alpha}$ is the upper α percentile of the Studentized range distribution with parameters k and $N-k$ degrees of freedom.
- The CDF and inverse CDF of the Studentized Range Distribution is available in R via the functions `ptukey()` and `qtukey()` respectively.

The Tukey Method

A $100(1-\alpha)\%$ simultaneous confidence interval for c pairs $\mu_i - \mu_j$ is

$$\bar{y_j}. - \bar{y_i}. \pm \frac{1}{\sqrt{2}} q_{k,N-k,\alpha} \hat{\sigma} \sqrt{1/n_j + 1/n_i}.$$

The Bonferroni method is more conservative than Tukey's method. In other words, the simutaneous confidence intervals based on the Tukey method are shorter.

The Tukey Method

- In the coagualtion study $N = 24, k = 4$ so the 5% critical value of the Studentized range distribution is obtained using the the inverse CDF function `qtukey()` for this distribution.
- The argument `lower.tail=FALSE` is used so we obtain the upper percentile of the distribution (i.e., the value of x such that $P(X > x) = 0.05$).

```
qtukey(p = .05,nmeans = 4,df = 20,lower.tail = FALSE)
```

```
## [1] 3.958293
```

The Tukey Method

- Let's obtain the Tukey p-value and confidence interval for $\mu_B - \mu_A$.
- The observed value of the test statistic is

$$q^{obs} = \sqrt{2}|t_{AB}|,$$

where

$$t_{AB} = \frac{\bar{y_A}. - \bar{y_B}.}{\hat{\sigma}\sqrt{1/n_A + 1/n_B}}.$$

```
(sqrt(2)*(66-61))/(sqrt(5.6)*sqrt(1/6+1/6))
```

```
## [1] 5.175492
```

The Tukey Method

The p-value

$$P\left(q_{4,20} > q^{obs}\right)$$

is then obtained using the CDF of the Studentized range distribution

```
1-ptukey(q = sqrt(2)*5/sqrt(2*5.6/6),nmeans = 4,df = 20)
```

```
## [1] 0.007797788
```

The Tukey Method

The 95% limits of the Tukey confidence interval for $\mu_B - \mu_A$ is

```
tuk.crit <- qtukey(p=.05,nmeans=4,df=20,lower.tail=FALSE)
#lower limit
round(5-(1/sqrt(2))*tuk.crit*sqrt(5.6)*sqrt(1/6+1/6),2)
```

```
## [1] 1.18
```

```
#upper limit
round(5+(1/sqrt(2))*tuk.crit*sqrt(5.6)*sqrt(1/6+1/6),2)
```

```
## [1] 8.82
```

The Tukey Method

The width of the Tukey confidence interval for $\mu_B - \mu_A$ is

```
round((1/sqrt(2))*tuk.crit*sqrt(5.6)*sqrt(1/6+1/6),2)
```

```
## [1] 3.82
```

The width of Bonferroni $\mu_B - \mu_A$ is

```
round(qt(p = 1-0.004,df = 20)*sqrt(5.6)*sqrt(1/6+1/6),2)
```

```
## [1] 4.02
```

The Tukey Method

- This shows that the Tukey confidence interval is shorter than Bonferroni confidence intervals.
- The command `TukeyHSD()` can be used to obtain all the Tukey confidence intervals and p-values for an ANOVA.

The Tukey Method

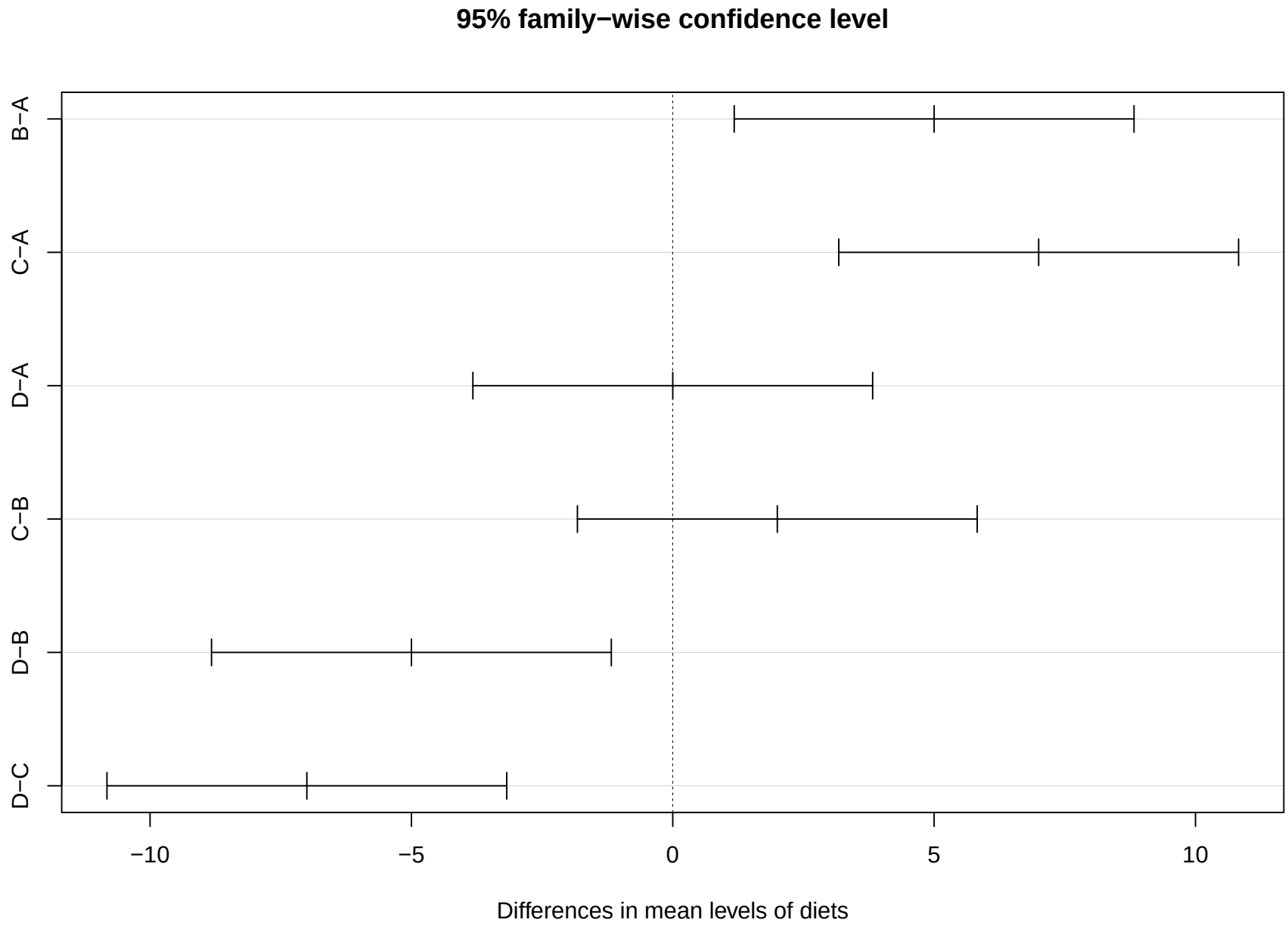
```
TukeyHSD(aov(y~diets,data=tab0401))
```

```
round(TukeyHSD(aov(y~diets,data=tab0401))$diets,2)
```

```
##      diff    lwr   upr p adj
## B-A     5   1.18  8.82  0.01
## C-A     7   3.18 10.82  0.00
## D-A     0  -3.82  3.82  1.00
## C-B     2  -1.82  5.82  0.48
## D-B    -5  -8.82 -1.18  0.01
## D-C    -7 -10.82 -3.18  0.00
```

The Tukey Method

```
plot(TukeyHSD(aov(y~diets,data=tab0401)))
```



Sample size for ANOVA - Designing a study to compare more than two treatments

- Consider the hypothesis that k means are equal vs. the alternative that at least two differ.
- What is the probability that the test rejects if at least two means differ?
- Power $= 1 - P(\text{Type II error})$ is this probability.

Sample size for ANOVA - Designing a study to compare more than two treatments

The null and alternative hypotheses are:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ vs. } H_1 : \mu_i \neq \mu_j.$$

The test rejects at level α if

$$MS_{Treat}/MS_E \geq F_{k-1,N-K,\alpha}.$$

The power of the test is

$$1 - \beta = P\left(MS_{Treat}/MS_E \geq F_{k-1,N-K,\alpha}\right),$$

when H_0 is false.

Sample size for ANOVA - Designing a study to compare more than two treatments

When H_0 is false it can be shown that:

- MS_{Treat}/σ^2 has a non-central Chi-square distribution with $k-1$ degrees of freedom and non-centrality parameter δ.
- MS_{Treat}/MS_E has a non-central F distribution with the numerator and denominator degrees of freedom $k-1$ and $N-k$ respectively, and non-centrality parameter

$$\delta = \frac{\sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu}\right)^2}{\sigma^2},$$

where n_i is the number of observations in group i, $\bar{\mu} = \sum_{i=1}^{k} \mu_i/k$, and σ^2 is the within group error variance .

This is dentoted by $F_{k-1,N-k}(\delta)$.

Direct calculation of Power

- The power of the test is

$$P\left(F_{k-1,N-k}(\delta) > F_{k-1,N-K,\alpha}\right).$$

- The power is an increasing function δ
- The power depends on the true values of the treatment means μ_i, the error variance σ^2, and sample size n_i.
- If the experimentor has some prior idea about the treament means and error variance, and the sample size (number of replications) the formula above will calculate the power of the test.

Blood coagulation example - sample size

Suppose that an investigator would like to replicate the blood coagulation study with only 3 animals per diet. In this case $k = 4, n_i = 3$. The treatment means from the initial study are:

Diet	A	B	C	D
Average	61	66	68	61

```
lm.diets <- lm(y~diets,data=tab0401);round(summary(lm.diets)$coefficients,2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)       61       0.97   63.14        0
## dietsB             5       1.37    3.66        0
## dietsC             7       1.37    5.12        0
## dietsD             0       1.37    0.00        1
```

```
anova(lm.diets)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diets      3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Blood coagulation example - sample size

- $\mu_1 = 61$, $\mu_2 = 66$, $\mu_3 = 68$, $\mu_4 = 61$.
- The error variance σ^2 was estimated as $MS_E = 5.6$.
- Assuming that the estimated values are the true values of the parameters, the non-centrality parameter of the F distribution is

$$\delta = 3 \times \left((61-64)^2 + (66-64)^2 + (68-64)^2 + (61-64)^2\right)/5.6 = 20.35714$$

Blood coagulation example - sample size

If we choose $\alpha = 0.05$ as the significance level then $F_{3,20,0.05} = 3.0983912$. The power of the test is then

$$P\left(F_{3,20}(20.36) > 3.10\right) = 0.94.$$

This was calculated using the CDF for the F distribution in R `pf()`.

```
1-pf(q = 3.10,df1 = 3,df2 = 20,ncp = 20.36)
```

```
## [1] 0.9435208
```

Calculating power and sample size using the pwr library

- There are several libraries in R which can calculate power and sample size for statistical tests. The library `pwr()` has a function
- `pwr.anova.test(k = NULL, n = NULL, f = NULL, sig.level = 0.05, power = NULL)`

for computing power and sample size.

- k Number of groups
- n Number of observations (per group)
- f Effect size
- The effect size is the square root of the non-centrality parameter of the non-central F distribution.

$$f = \sqrt{\frac{\sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu}\right)^2}{\sigma^2}}.$$

where n_i is the number of observations in group i, $\bar{\mu} = \sum_{i=1}^{k} \mu_i / k$, and σ^2 is the within group error variance.

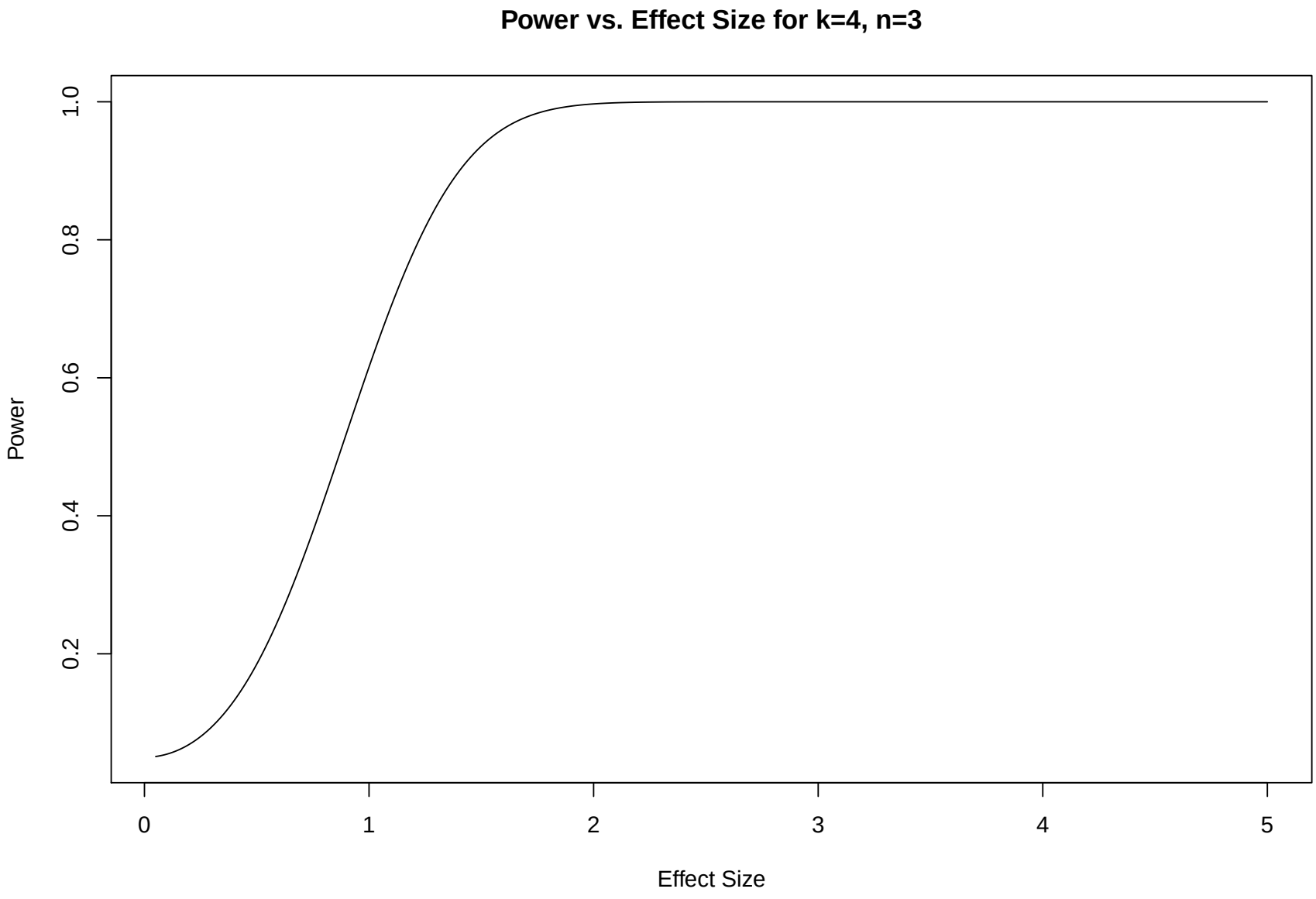
Calculating power and sample size using the pwr library

In the previous example $\delta = 20.35714$ so $f = \sqrt{20.35714} = 4.5118887$.

```
library(pwr)
pwr.anova.test(k = 4,n = 3,f = 4.5)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 3
##               f = 4.5
##       sig.level = 0.05
##           power = 1
##
## NOTE: n is number in each group
```

Calculating power and sample size using the pwr library



Calculating power using simulation

The general procedure for simulating power is:

1. Use the underlying model to generate random data with (a) specified sample sizes, (b) parameter values that one is trying to detect with the hypothesis test, and (c) nuisance parameters such as variances.
2. Run the estimation program (e.g., `t.test()`,`lm()`) on these randomly generated data.
3. Calculate the test statistic and p-value.
4. Do Steps 1–3 many times, say, N, and save the p-values. The estimated power for a level alpha test is the proportion of observations (out of N) for which the p-value is less than alpha.

Calculating power using simulation

One of the advantages of calculating power via simulation is that we can investigate what happens to power if, say, some of the assumptions behind one-way ANOVA are violated.

Calculating power using simulation - R program

```r
#Simulate power of ANOVA for three groups

NSIM <- 1000 # number of simulations
res <- numeric(NSIM) # store p-values in res

mu1 <- 2; mu2 <- 2.5;mu3 <- 2  # true mean values of treatment groups
sigma1 <- 1; sigma2 <- 1; sigma3 <- 1 #variances in each group
n1 <- 40; n2 <- 40; n3 <- 40 #sample size in each group

for (i in 1:NSIM) # do the calculations below N times
  {
# generate sample of size n1 from N(mu1,sigma1^2)
y1 <- rnorm(n = n1,mean = mu1,sd = sigma1)
# generate sample of size n2 from N(mu2,sigma2^2)
y2 <- rnorm(n = n2,mean = mu2,sd = sigma2)
# generate sample of size n3 from N(mu3,sigma3^2)
y3 <- rnorm(n = n3,mean = mu3,sd = sigma3)
y <- c(y1,y2,y3) # store all the values from the groups
# generate the treatment assignment for each group
trt <- as.factor(c(rep(1,n1),rep(2,n2),rep(3,n3)))
m <- lm(y~trt) # calculate the ANOVA
res[i] <- anova(m)[1,5] # p-value of F test
}
sum(res<=0.05)/NSIM # calculate p-value
```

```
## [1] 0.598
```