Dispersion

BIOE 498/598 PJ

Spring 2022

Why replication?

1. Reduce noise effects.
2. Estimate confidence intervals for effect sizes.
3. Analyze dispersion effects.

Sample variance across replicates

If a run is replicated r times with responses y_1, y_2, . . ., y_r and mean $\bar{y}$,

$$\text{sample variance} = s^2 = \frac{\sum_i^r (y_i - \bar{y})^2}{r - 1}$$

Sample variance across replicates

If a run is replicated r times with responses y_1, y_2, ..., y_r and mean $\bar{y}$,

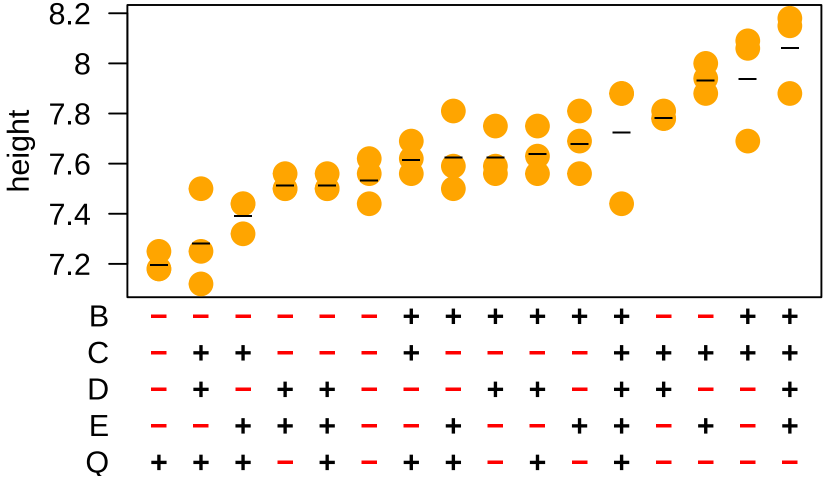
$$\text{sample variance} = s^2 = \frac{\sum_i^r (y_i - \bar{y})^2}{r - 1}$$

For a factorial design with N *unreplicated* runs ($N = 2^k$ for a full factorial or $N = 2^{k-p}$ for a fractional factorial),

$$\text{standard error of effects} = SE(\beta_i) = \sqrt{\frac{\text{mean}(s^2)}{rN}}$$

Visualizing the data

```
farplot(data, factors=c("B","C","D","E","Q"), response="height")
```

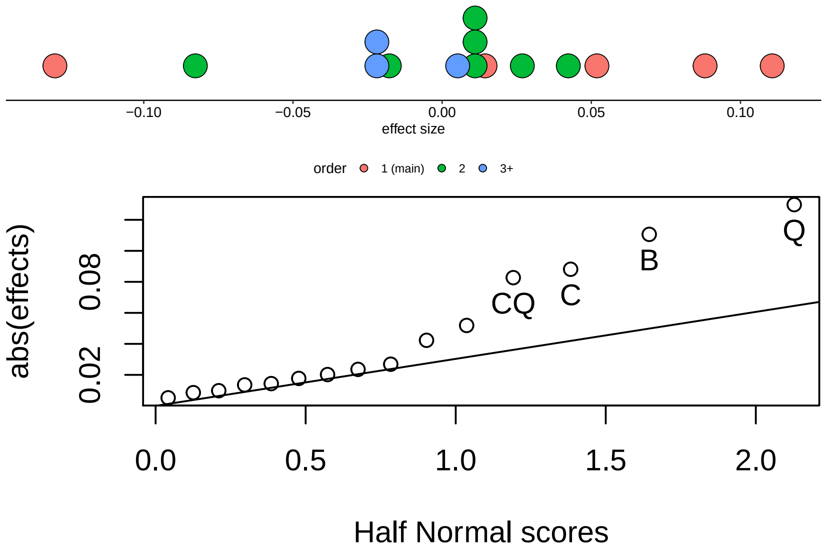


Linear models find the “best fit” effect sizes

```
model <- lm(height ~ B*C*D*E*Q, data=data)
show_model(model, n_coefs=17, show_fit=FALSE)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.636042   0.018571 411.183  < 2e-16 ***
## Q           -0.129792   0.018571  -6.989 6.42e-08 ***
## B            0.110625   0.018571   5.957 1.23e-06 ***
## C            0.088125   0.018571   4.745 4.16e-05 ***
## C:Q         -0.082708   0.018571  -4.454 9.64e-05 ***
## E            0.051875   0.018571   2.793  0.00874 **
## B:Q          0.042292   0.018571   2.277  0.02959 *
## D:Q          0.026875   0.018571   1.447  0.15758
## C:D:Q       -0.023542   0.018571  -1.268  0.21406
## B:D:Q       -0.020208   0.018571  -1.088  0.28465
## C:D         -0.017708   0.018571  -0.954  0.34746
## D            0.014375   0.018571   0.774  0.44458
## E:Q          0.013542   0.018571   0.729  0.47119
## B:D          0.009792   0.018571   0.527  0.60165
## B:C          0.008542   0.018571   0.460  0.64866
## B:C:Q        0.005208   0.018571   0.280  0.78093
## B:E                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Half-normal & dot plots — significance based only on effect size




```
## zscore= 0.0417893 0.1256613 0.2104284 0.2967378 0.3853205 0.4770404
```

Location vs. Dispersion

- Sometimes we want to study the variation in the response, not the response itself.
- **Location** describes the central tendency of a response
  - Mean, median, mode
  - All of our models so far use response = location

Location vs. Dispersion

- Sometimes we want to study the variation in the response, not the response itself.
- **Location** describes the central tendency of a response
  - Mean, median, mode
  - All of our models so far use response = location
- **Dispersion** describes the spread of a response
  - Range, inter-quartile range (IQR), variance, standard deviation

Location vs. Dispersion

- Sometimes we want to study the variation in the response, not the response itself.
- **Location** describes the central tendency of a response
  - Mean, median, mode
  - All of our models so far use response = location
- **Dispersion** describes the spread of a response
  - Range, inter-quartile range (IQR), variance, standard deviation
- Location can be studied with unreplicated or replicated designs
- Studying dispersion always requires replicates

Studying dispersion

- The variance σ^2 is the natural statistic for studying dispersion with linear models fit by least-squares
- However, the sample variance s^2 is not a good response for studying σ^2
  - s^2 is left-censored ($s^2 \geq 0$)
  - s^2 follows a χ^2 distribution, not a normal distribution

Studying dispersion

- The variance σ^2 is the natural statistic for studying dispersion with linear models fit by least-squares
- However, the sample variance s^2 is not a good response for studying σ^2
  - s^2 is left-censored ($s^2 \geq 0$)
  - s^2 follows a χ^2 distribution, not a normal distribution
- Both problems are fixed by modeling $\ln s^2$ instead of s^2

Studying dispersion

- The variance σ^2 is the natural statistic for studying dispersion with linear models fit by least-squares
- However, the sample variance s^2 is not a good response for studying σ^2
  - s^2 is left-censored ($s^2 \geq 0$)
  - s^2 follows a χ^2 distribution, not a normal distribution
- Both problems are fixed by modeling $\ln s^2$ instead of s^2
- Moreover, maximizing $-\ln s^2$ minimizes the variance, so we can keep the same maximization-based framework used for location models

Calculating ln s^2

```r
disp <- add_dispersion(data, factors=c("B","C","D","E","Q"),
                       response="height")
```

```r
head(data, n=16)
```

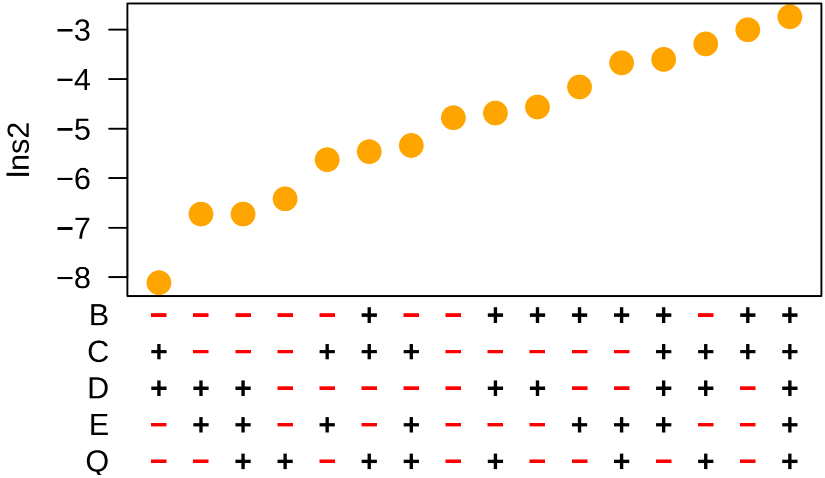
```
##     B  C  D  E  Q height
## 1  -1 -1 -1 -1 -1   7.56
## 2  -1 -1 -1 -1 -1   7.62
## 3  -1 -1 -1 -1 -1   7.44
## 4  -1 -1 -1 -1  1   7.18
## 5  -1 -1 -1 -1  1   7.18
## 6  -1 -1 -1 -1  1   7.25
## 7  -1 -1  1  1 -1   7.50
## 8  -1 -1  1  1 -1   7.56
## 9  -1 -1  1  1 -1   7.50
## 10 -1 -1  1  1  1   7.50
## 11 -1 -1  1  1  1   7.56
## 12 -1 -1  1  1  1   7.50
## 13 -1  1 -1  1 -1   7.94
## 14 -1  1 -1  1 -1   8.00
## 15 -1  1 -1  1 -1   7.88
## 16 -1  1 -1  1  1   7.32
```

```r
disp
```

```
##     B  C  D  E  Q      lns2   height
## 1  -1 -1 -1 -1 -1 -4.779524 7.540000
## 2  -1 -1 -1 -1  1 -6.417132 7.203333
## 3  -1 -1  1  1 -1 -6.725434 7.520000
## 4  -1 -1  1  1  1 -6.725434 7.520000
## 5  -1  1 -1  1 -1 -5.626821 7.940000
## 6  -1  1 -1  1  1 -5.339139 7.400000
## 7  -1  1  1 -1 -1 -8.111728 7.790000
## 8  -1  1  1 -1  1 -3.288762 7.290000
## 9   1 -1 -1  1 -1 -4.158350 7.686667
## 10  1 -1 -1  1  1 -3.671695 7.633333
## 11  1 -1  1 -1 -1 -4.562749 7.633333
## 12  1 -1  1 -1  1 -4.684935 7.646667
## 13  1  1 -1 -1 -1 -3.003093 7.946667
## 14  1  1 -1 -1  1 -5.464766 7.623333
## 15  1  1  1  1 -1 -3.600869 8.070000
## 16  1  1  1  1  1 -2.740573 7.733333
```

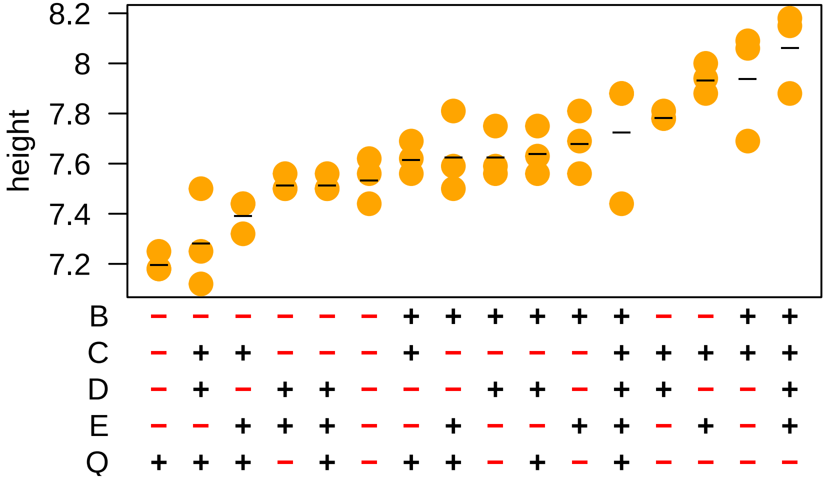
Visualizing the **dispersion**

```
farplot(disp, factors=c("B","C","D","E","Q"), response="lns2")
```



Visualizing the **data**

```
farplot(data, factors=c("B","C","D","E","Q"), response="height")
```



Building the model

```
disp_model <- lm(-lns2 ~ B+C+D+E+Q +
                         B:Q + C:Q + D:Q + E:Q + B:C + B:D + B:E,
                 data=disp)
```

Confounding in the 2^{5-1} design with I=BCDE:

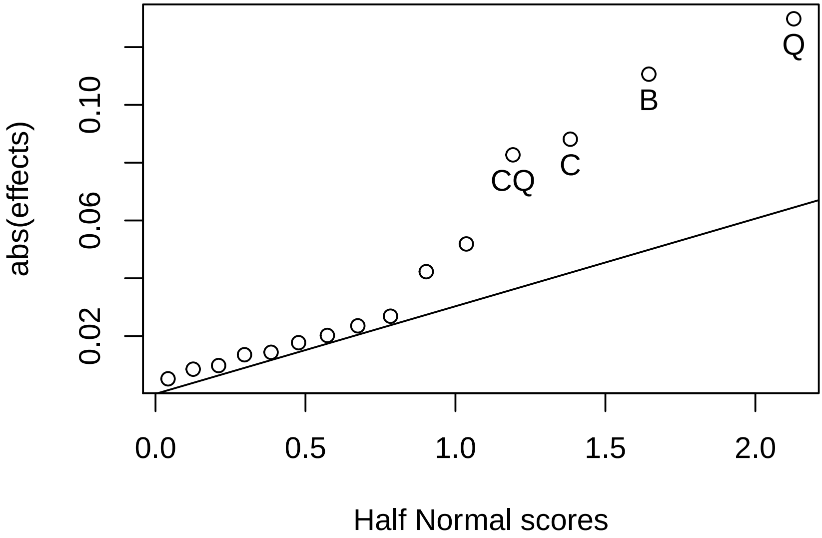
- main effects clear
- BQ
- CQ
- DQ
- EQ
- BC=DE
- BD=CE
- BE=CD

```
show_effects(disp_model, ordered="abs")
```

```
## (Intercept)    4.93131
##           B    -.94543
##         D:Q    -.55538
##         B:E    -.33523
##         C:Q    -.2989
##         B:Q     .29437
##           C    -.28434
##         B:D    -.21234
##           Q    -.13976
##           D     .12375
##           E    -.10777
##         E:Q    -.06457
##         B:C     .00079
```

Factors affecting **location** (spring height)

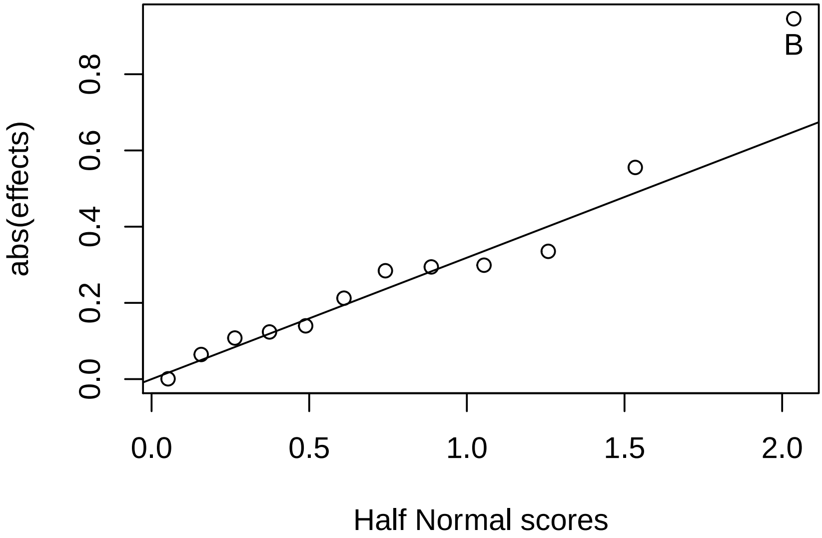
```
daewr::halfnorm(na.omit(get_effects(model)))
```




```
## zscore= 0.0417893 0.1256613 0.2104284 0.2967378 0.3853205 0.4770404
```

Factors affecting **dispersion** ($\ln s^2$)

```
daewr::halfnorm(get_effects(disp_model))
```




```
## zscore= 0.05224518 0.1573107 0.264147 0.3740954 0.4887764 0.6102946
```