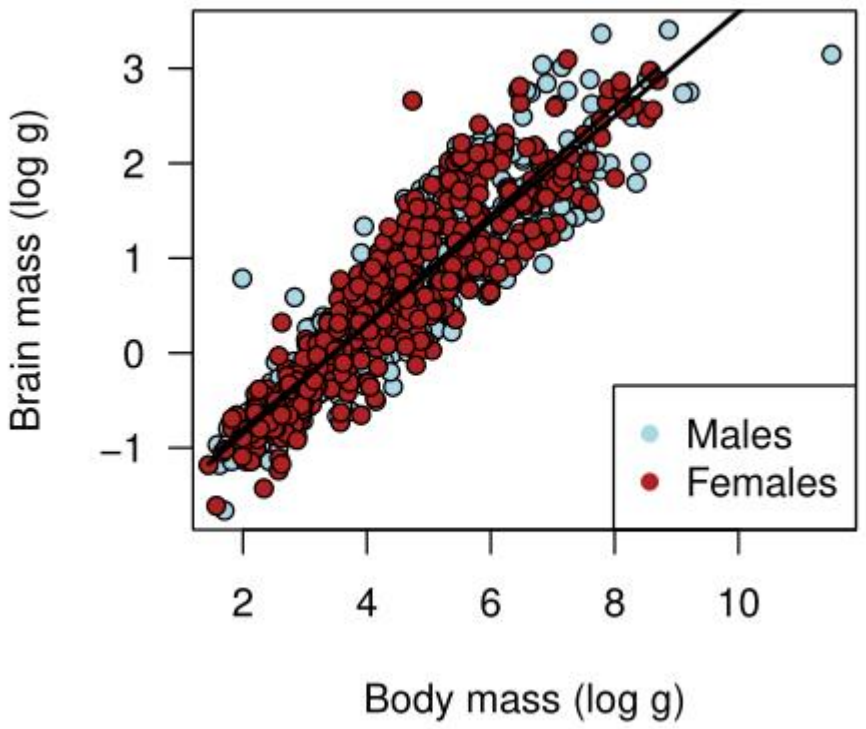
Discussion of exercise 2

- Allometric analyses
- Attenuation bias

- Linear model fitted to log-transformed data
- The allometric intercept is sometimes of interest, but mostly related to the general size of the focal organ
- The allometric slope gives the scaling exponent, i.e. the proportional change in brain size per proportional change in body size

```
##
## Call:
## lm(formula = log(brain_mass) ~ log(body_mass), data = males)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27532 -0.24778 -0.06226  0.19858  1.59801
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.905598   0.045493  -41.89   <2e-16 ***
## log(body_mass)  0.550206   0.009898   55.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4144 on 613 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8342
## F-statistic:  3090 on 1 and 613 DF,  p-value: < 2.2e-16
```

- The allometric slope is slightly steeper in females



Methods and Results

Analysis Methods We expected brain size to scale with body size according to a power-law relationship on the form $brainmass = a \times bodymass^b$. We linearized the expected power relationship through the logarithmic transformation $log(brainmass) = log(a) + b \times log(bodymass)$, and then fitted a linear regression model to the data. To assess whether the allometric slope (b) differs between the sexes, we analysed data for males and females separately.

Results Brain size scaled allometrically with body size (Fig. 1). In males, brain size increased by 5.5% per 10% increase in body mass (allometric slope $= 0.55 \pm 0.010$), and body mass explained 83.5% of the variance in brain mass. The allometric slope was slightly steeper in females (0.57 ± 0.012).

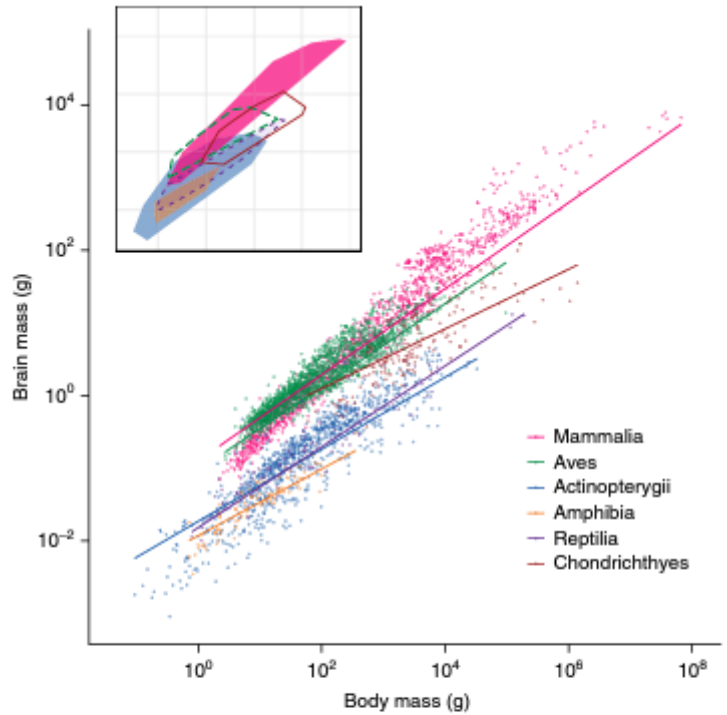


Fig. 1 | Brain–body evolutionary allometry of six vertebrate classes. Class-level brain–body allometries of six major vertebrate lineages are shown in different colours (*x* and *y* axes are in $\log_{10}$ scales). Points represent species means and unbroken lines are least square regressions accounting for phylogenetic relatedness among species. The inset shows minimum convex polygons of the morphospace occupied by Actinopterygii ($N=963$), Amphibia ($N=86$), Aves ($N=1902$), Chondrichthyes ($N=147$), Mammalia ($N=1409$) and non-avian reptiles (Reptilia, $N=79$).

Exploring attenuation bias

- 1. Define true relationship
- 2. Add some error to the measurements
- 3. Save relative errors for later correction
- 4. Fit models based on measurements

```r
```{r, fig.width=4, fig.height=4, echo=F}
x = rnorm(500, 10, 2)
y = 1.5*x + rnorm(500, 0, 1)

slope_est = NULL
errors = seq(0.01, 0.5, length.out=10)

relerrors = (errors^2)/var(x)

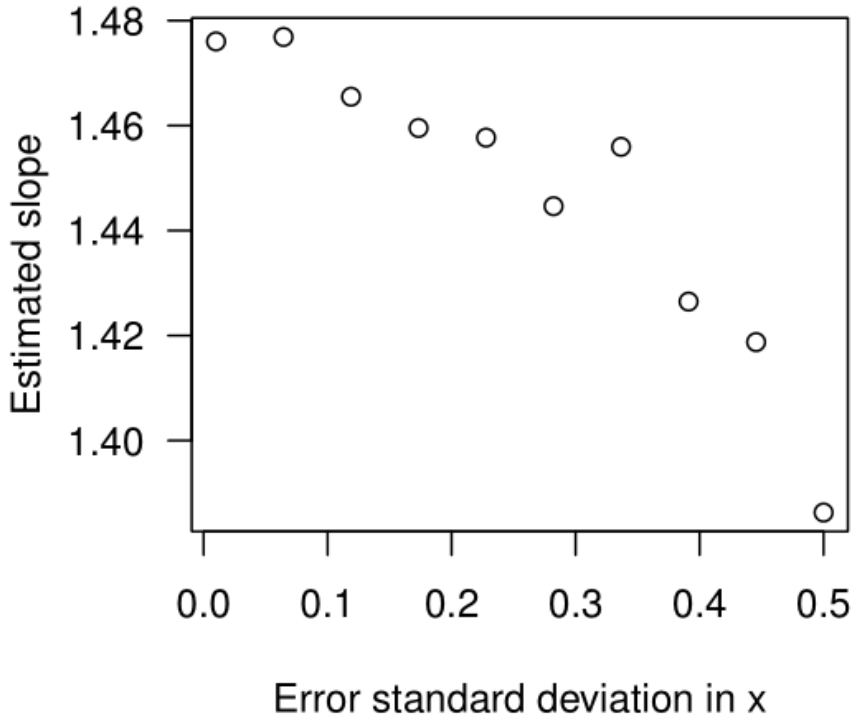
for(i in 1:10){
x_obs = x + rnorm(500, 0, errors[i])

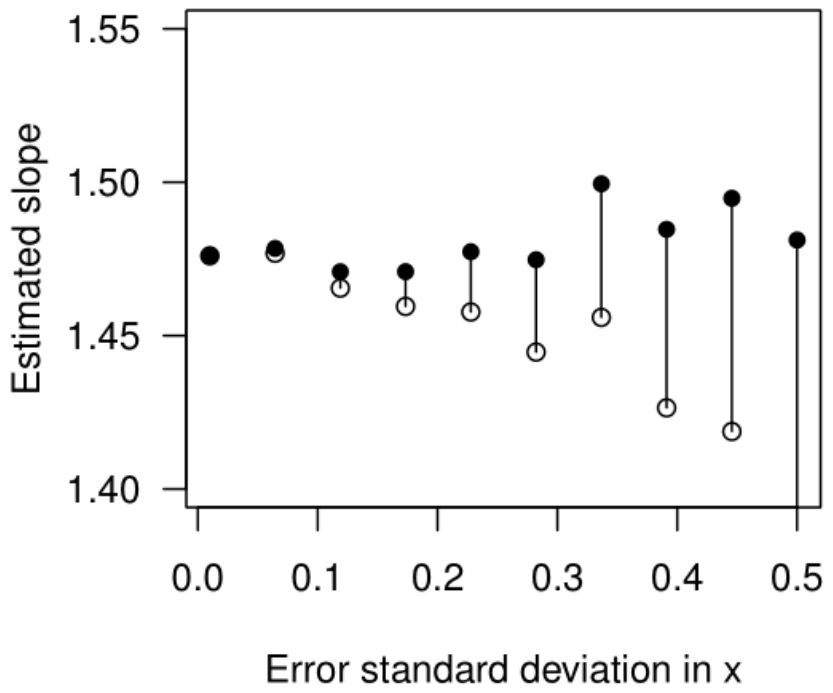
m1 = lm(y~x_obs)
slope_est[i] = summary(m1)$coef[2,1]
}

plot(errors, slope_est,
     las=1,
     xlab="Error standard deviation in x",
     ylab="Estimated slope")
```
```

- corrslope = slope_est/(1-relerrors)

$$K = 1 - \frac{\sigma_{me}}{\sigma_x} \qquad \beta' = \frac{\beta}{K}$$





Processing and Analysis of Biological Data

BIOS14 2023

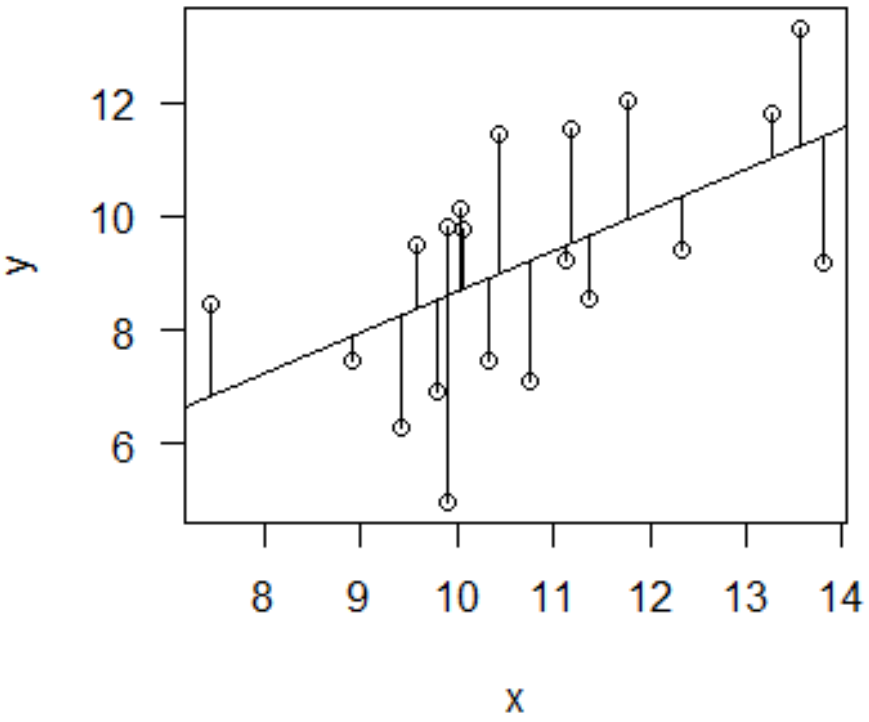
Lecture 3. ANOVA

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i$$

Øystein H. Opedal

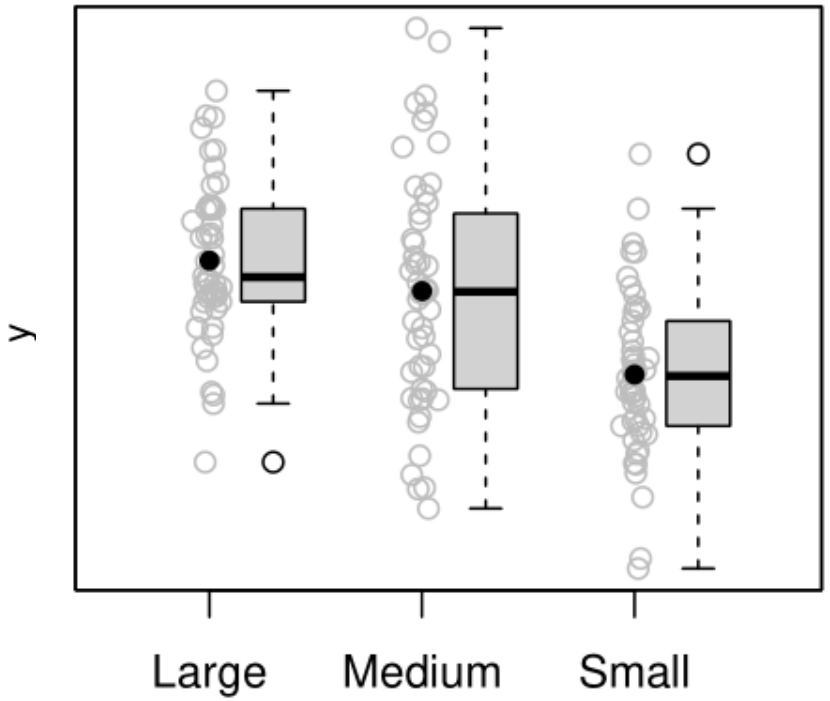
The linear model

- Most of the models we will work with in this course are linear models, that describe how a linear set of predictors relate to a response variable
- A key element of the model is the so-called linear predictor:
- $y_i = \beta_0 + \sum_j \beta_j \, x_{ij} + \varepsilon_i$, $\varepsilon \sim N(0, \sigma^2)$
- The term $\varepsilon \sim N(0, \sigma^2)$ means that the residuals (epsilon) are assumed to follow a normal distribution



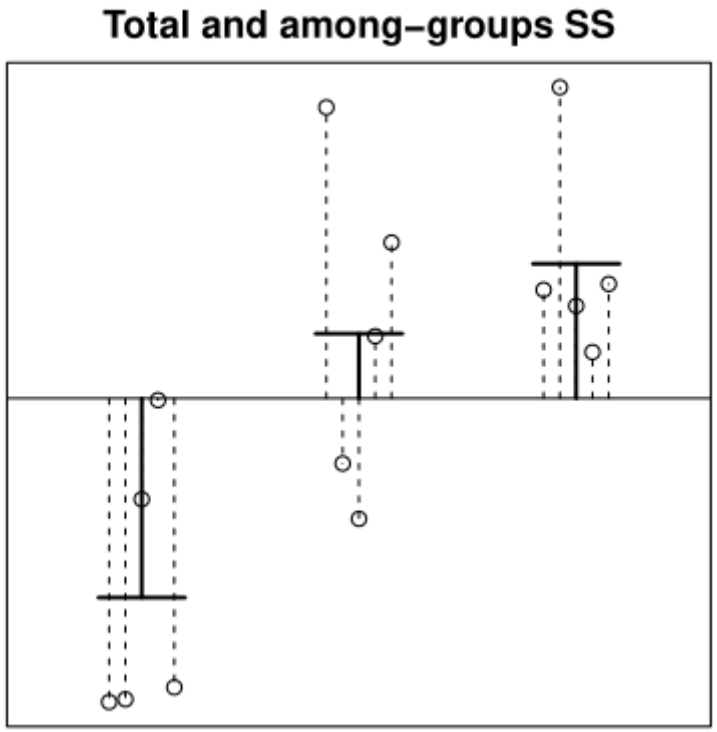
Analysis of Variance (ANOVA)

- Analyses of variance are linear models with one or more categorical predictors
- The predictor variables x are now factors, and the model parameters estimate contrasts between factor levels (e.g. treatments).
- Residuals are assumed to be normally distributed within each group, and variances are assumed to be equal



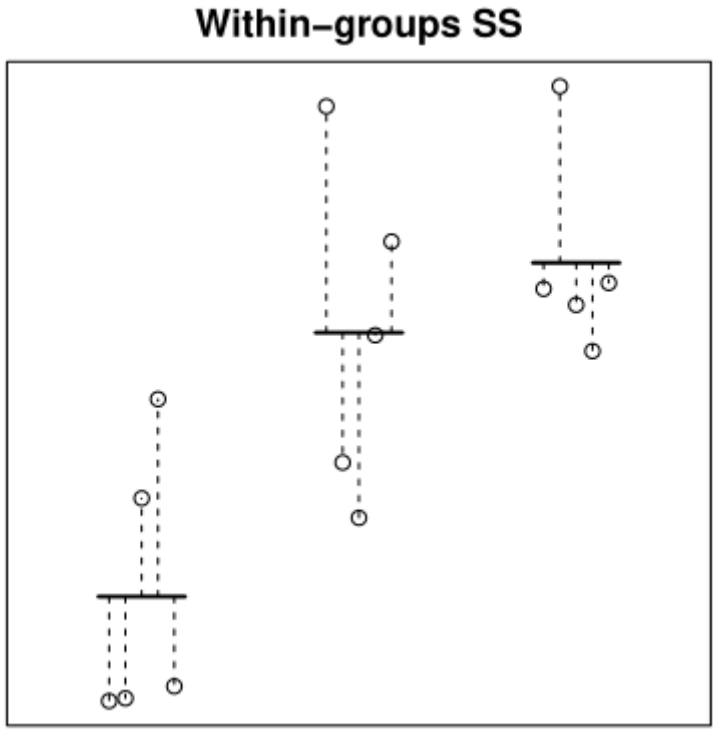
Analysis of Variance (ANOVA)

- Analyses of variance are based on partitioning variance into among- and within-group components
- The variances are the sums of the squared deviations from the corresponding mean



Analysis of Variance (ANOVA)

- Analyses of variance are based on partitioning variance into among- and within-group components
- The variances are the sums of the squared deviations from the corresponding mean



The “first” ANOVA

- Developed by Ronald Fisher for partitioning variance in phenotypic traits into genetic and environmental components
- Standard model for variance partitioning in phenotypic traits,
- $V_P = V_A + V_D + V_I + V_E + V_{G\times E}$

ANOVA model in R

- The ANOVA table gives the **sums of squares** associated with each factor, i.e. the sum of square deviations from the mean.
- We can use the sums of squares to obtain the proportion of variance explained by the grouping variable.

```
m = lm(x~groups)
anova(m)
```

```
## Analysis of Variance Table
##
## Response: x
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## groups      2  319.97 159.985  19.591 2.866e-08 ***
## Residuals 147 1200.43   8.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SS_T = 319.97+1200.43
SS_T/(150-1)
```

```
## [1] 10.20403
```

```
319.97/SS_T
```

```
## [1] 0.2104512
```

```
var(x)
```

```
## [1] 10.20403
```

ANOVA model in R

- As for all linear models, the summary table gives the parameter estimates, their standard errors, and other model statistics
- With a categorical predictor, the intercept gives the mean of the reference level (first level of the factor “groups”), and “groupsMedium” gives the contrast to the reference level

```
##
## Call:
## lm(formula = x ~ groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5887 -1.5596 -0.0987  1.6274  7.9729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.7006     0.4041  33.901  < 2e-16 ***
## groupsMedium  -0.9277     0.5715  -1.623    0.107
## groupsSmall   -3.4561     0.5715  -6.047 1.16e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.858 on 147 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.1997
## F-statistic: 19.59 on 2 and 147 DF,  p-value: 2.866e-08
```

ANOVA model in R

- As for all linear models, the summary table gives the parameter estimates, their standard errors, and other model statistics
- With a categorical predictor, the intercept gives the mean of the reference level (first level of the factor “groups”), and “groupsMedium” gives the contrast to the reference level.
- We can change the reference level to get other contrasts

```
groups = factor(groups, levels=c("Small", "Medium", "Large"))
m = lm(x~groups)
summary(m)
```

```
##
## Call:
## lm(formula = x ~ groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5887 -1.5596 -0.0987  1.6274  7.9729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2445     0.4041  25.349  < 2e-16 ***
## groupsMedium   2.5284     0.5715   4.424 1.88e-05 ***
## groupsLarge    3.4561     0.5715   6.047 1.16e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’
##
## Residual standard error: 2.858 on 147 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.1997
## F-statistic: 19.59 on 2 and 147 DF,  p-value: 2.866e-08
```

Two-way ANOVA

- With two factors, we refer to a two-way ANOVA (and so on for more factors).
- Typical example is factorial experiments (e.g. a 2×2 factorial as in today´s exercise).
- An interaction means that the effect of factor A depends on the level of factor B.
- In R-syntax, A*B means both main effects and their interaction.

TABLE 1. Potential sources of confusion in an experiment and means for minimizing their effect.

Source of confusion	Features of an experimental design that reduce or eliminate confusion
1. Temporal change	Control treatments
2. Procedure effects	Control treatments
3. Experimenter bias	Randomized assignment of experimental units to treatments Randomization in conduct of other procedures "Blind" procedures*
4. Experimenter-generated variability (random error)	Replication of treatments
5. Initial or inherent variability among experimental units	Replication of treatments Interspersion of treatments Concomitant observations
6. Nondemonic intrusion†	Replication of treatments Interspersion of treatments
7. Demonic intrusion	Eternal vigilance, exorcism, human sacrifices, etc.

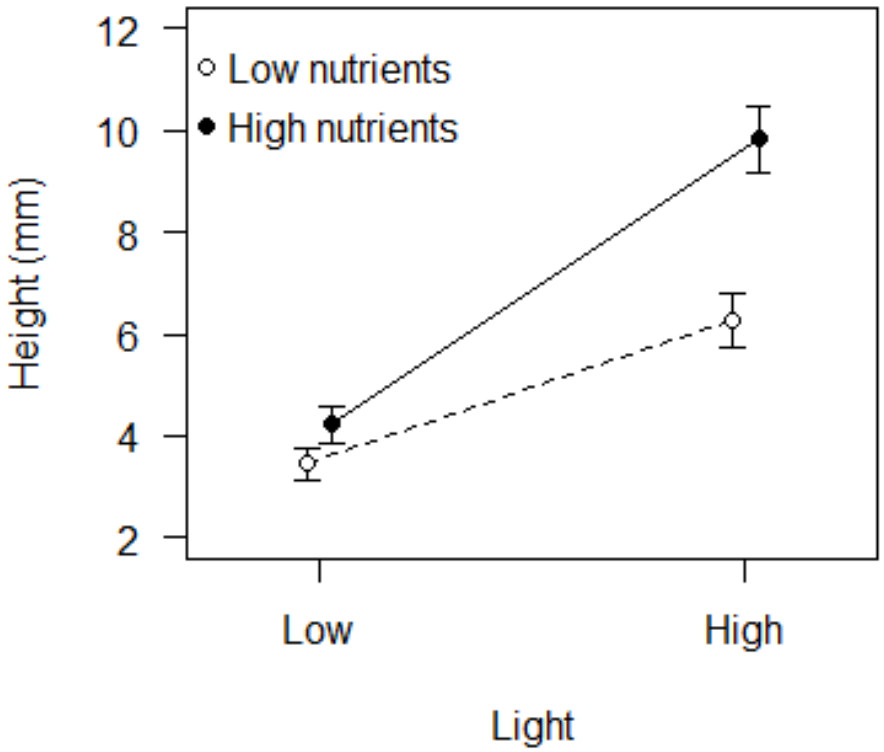
* Usually employed only where measurement involves a large subjective element.

† Nondemonic intrusion is defined as the impingement of chance events on an experiment in progress.

Table from Hurlbert 1984 Ecol Monographs

Two-way ANOVA

- With two factors, we refer to a two-way ANOVA (and so on for more factors).
- Typical example is factorial experiments (e.g. a 2×2 factorial as in today´s exercise).
- An interaction means that the effect of factor A depends on the level of factor B.
- In R-syntax, A*B means both main effects and their interaction.



General intro to exercises

- You can work with any dataset you may have!
- We will provide one or more potential datasets
- All exercises will involve formulating a research question, choosing analysis methods, performing the analysis, interpreting the results, and writing relevant Methods and Results
- The mid-term exercise and the exam will be similar
- Towards the end of each exercise session, we will show some examples of how the data could be analysed and the results presented
- No written feedback on the exercises (except mid-term), but we strongly encourage you to give feedback on each other's work.

Writing (analysis) methods

- Normally a specific section at the end of the Methods (but sometimes integrated throughout)
- Focus on the aim of the analysis before technical details
- For models, list terms in words, sometimes include model equation
- R syntax is increasingly acceptable

Examples of model description

We analysed seed set per blossom by fitting linear mixed effect models with timing of pollination, pollen type and population, as well as their interactions, as fixed factors, and maternal individual as a random factor. Although seed number is a count variable, the distribution of residuals was close to normal, allowing for the use of a linear model. To analyse variation in seed mass, we fitted linear mixed effect models with timing of pollination, pollen type and population, as well as their interactions, as fixed factors. We further included seed set and peduncle diameter as covariates to account for a possible trade-off between seed mass and seed number within seed set, and possible blossom size effects, respectively. Blossom identity nested

To investigate the relationship between microclimate and plot richness, we fitted a mixed-effects Poisson regression model with species richness of the sample plots as the dependent variable, and microclimate variables as possible explanatory variables. To account for the structure of the data (sites nested within areas), site and area were entered as random factors. At the among-site scale (using each site

bruster et al. 2009b). We modeled pollen arrival during the female phase, $P_F(z')$, in units of pollen grains, as a function of bract area, gland area, and gland-stigma distance through a log link with a negative-binomial error distribution. The minimal model describing the predicted pollen load on the last day of the female phase was

$$P_F(z') = e^{a_2 + b_{21} UBA + b_{22} GA + b_{23} GSD}. \quad (2)$$

Writing results

- Start with summary statistics/patterns of variation

Database description

The updated database (Table S3) contains 792 evolvability estimates for 54 taxa representing 27 families. Among the 72 studies included, 68.6% were conducted on populations originating in North America (Fig. S1). Most studies were conducted in glasshouses or other controlled environments.

Writing results

- Start with summary statistics/patterns of variation
- Focus on biology over statistics (explain the result in biological terms)
- Quantify and exemplify (*y* increased by 0.5 mm per mm increase in *x*)

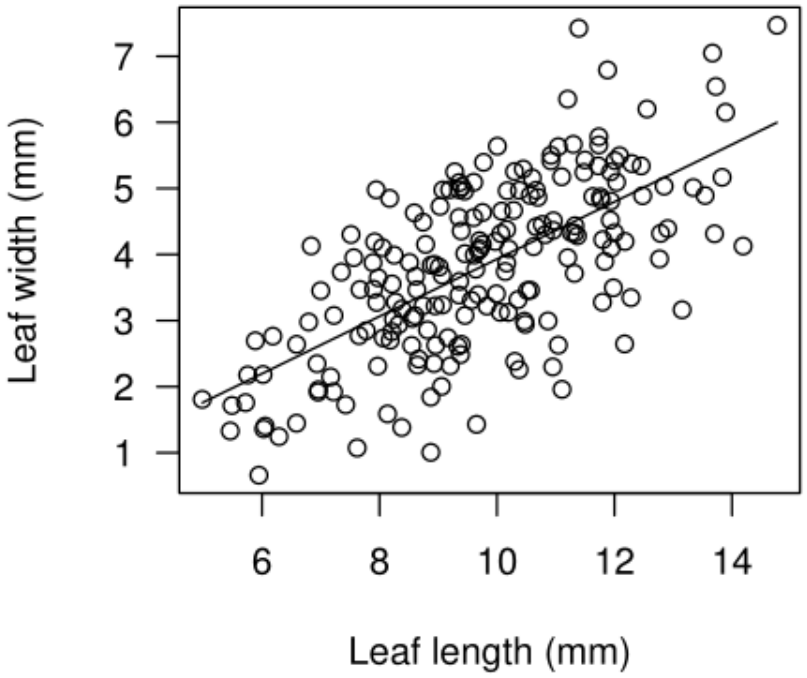
Overall, blossom development tended to be more rapid under dry conditions. In the wet treatment, the total receptive period of the blossoms (from the first day of opening to the abscision of the male cymule) lasted for 6.87 ± 0.09 d in the two large-glanded populations and 5.80 ± 0.10 d in the two small-glanded populations (fig. 4). In the dry treatment, the length of this period was moderately reduced by, on average, 5.4% across all four populations (fig. 4; table 4; see table A3 for model-selection results).

Writing results

- Start with summary statistics/patterns of variation
- Focus on biology over statistics (explain the result in biological terms)
- Quantify and exemplify (*y* increased by 0.5 mm per mm increase in *x*)
- No discussion (normally)
- Avoid unnecessary introductory sentences like “Population means are reported in Table 1”.
- Refer to tables and figures after a statement (“Body size varied substantially among populations (range = X.X to Y.Y, Table 1)”.

Making nice figures

- Units!
- Keep regression lines within the data range
- If several symbols/colors: legend inside the graph



Making nice tables

- No vertical lines
- All parameters explained in legend

Table 5

Inaccuracy Statistics under Wet and Dry Experimental Conditions

Taxon, population, treatment	Bias²	Male variance	Female variance	Joint inaccuracy (95% CI)	Mean-scaled inaccuracy (95% CI)
Large-glanded species:					
Tulum:					
Wet	.14 (16.1%)	.25 (28.8%)	.48 (55.1%)	.87 (.62, 1.17)	.03 (.02, .04)
Dry	.05 (3.0%)	1.00 (53.8%)	.80 (43.2%)	1.85 (1.42, 2.33)	.11 (.08, .14)
Puerto Morelos:					
Wet	.53 (25.8%)	.94 (46.0%)	.58 (28.2%)	2.05 (1.56, 2.66)	.07 (.05, .09)
Dry	1.42 (56.7%)	.69 (27.4%)	.40 (15.9%)	2.51 (1.94, 3.09)	.14 (.11, .17)
Small-glanded species:					
Cozumel:					
Wet	1.83 (63.7%)	.15 (5.1%)	.9 (31.3%)	2.87 (2.10, 3.71)	.22 (.17, .28)
Dry	.93 (50.2%)	.19 (10.5%)	.73 (39.4%)	1.85 (1.21, 2.57)	.20 (.14, .27)
Valladolid:					
Wet	.55 (38.8%)	.33 (23.5%)	.53 (37.7%)	1.41 (.82, 2.20)	.12 (.07, .19)
Dry	.29 (37.7%)	.24 (30.9%)	.24 (31.4%)	.77 (.43, 1.23)	.08 (.04, .13)

Note. Bias² is the mean squared deviation from the hypothesized adaptive optimum. The reported percentages are the proportion of the joint inaccuracy explained by each component. To obtain the mean-scaled inaccuracy, the joint inaccuracy was scaled by the product of the male and female trait means. Ninety-five percent confidence intervals (95% CIs) were obtained from 1000 nonparametric bootstrap estimates of the joint and mean-scaled inaccuracies, respectively.