Quantitative Analysis in Ecology and Evolution

Øystein H. Opedal

## Preface

These lectures notes are written for the course BIOS11 at Lund University. They will cover what I consider to be necessary knowledge of data analysis and quantitative methods in ecology, evolutionary biology, and related fields. This document is work-in-progress and is not comprehensive. I therefore storngly encourage the reader to consult the many excellent text books available on ‘biostatistics’, R, quantitative biology etc.

What separates this document from ‘typical’ texts in biostatistics is the limited focus on statistical hypothesis testing. Though I will discuss key aspects of hypothesis testing in the traditional sense, focus will be throughout on quantification, parameter estimation, and biological interpretation of results. This choice is made in an attempt to compensate for an issue that occurs frequently in scientific writing. The presentation of results in biological publications has increasingly tended towards a focus on statistical hypothesis testing (‘significance testing’), at the cost of focus on the biological significance of any result obtained. An example is sentences on the form ‘the treatment significantly (P < 0.05) affected the growth of plants’ in place of an appropriate sentence such as ‘plants in the high-fertilizer treatment grew 40% larger than those in the low-fertilizer treatment’. There is now a growing focus on turning this trend (Wasserstein & Lazar 2016), and this must start with the way statistics are though in introductory biostatistical courses.

Beyond quantification and data analysis, the following chapters will also introduce some basic scientific programming.

## Measurement and meaning in biology

1. Measurement and meaning in Biology. Scale types. Examples of bad practice. Houle et al. + R?

The title of this section is shared with an important paper published in 2010 by evolutionary biologist David Houle and colleagues, in which they pointed out the common misconnect in biology between the measurements taken in the field or laboratory, and the biological properties they are meant to represent. This exemplifies a more general problem across biology, where the interpretation of measurements and analyses fails to focus on the biological questions that motivated the study in the first place.

### Scale types and transformations

The concept of scale types has rarely been thought in biology, yet any quantitative measurement is placed on a specific scale type, with some associated properties such as permissable transformations. The perhaps most familiar scale type is the *ratio scale*, represented for example by any linear size measurement (with units such as *mm* or *m*). Importantly, the ratio scale has a natural zero, and negative values don’t occur. In contrast, consider the measurement of days since January 1st (often used to record for example breeding dates of birds). Dates are on an *interval scale*, and lacks a natural zero (i.e., the zero is chosen arbitrarily). This has consequences, for example, for how we compare raw and proportional variances among groups (see below).

## Summary statistics

1. Summary statistics, programming, resampling, for-loops. Intro to reproducibility (set.seed etc.). Pruitt case?

Summary statistics are those that aim to describe the properties of the data. For example, the central tendency of a dataset is generally measured by the arithmetic mean (typically denoted

or

), median, or, less frequently, the mode. Which of these to choose depends in part of the distribution of the data. If the data are skewed (tendency towards large or small values), the mean can give a misleading picture of the central tendency. The same issue arises if there are extreme valeues (outliers) that will tend to affect the mean much more than they affect the median. In some cases, a good option can be to report both the mean and the median, which together will give a more complete impression of the data distribution.

Variation in the data is typically described by the variance (typically denoted

) or its square root, the standard deviation (

). A pointed out by R. A. Fisher, the advantage of the variance (mean squared deviation from the mean) is that variance component are additive, allowing us to partition the total variance in the data into different components (see section on Variance partitioning below). To understand why square values are additive, recall Pythagoras theorem (

, but

).

The standard deviation measures the mean deviation of each data point from the mean, and thus has the advantage of having the same units as the original measurements. If we have measured a trait in mm, the standard deviation gives the average deviation in mm from the mean, and is thus easy to interpret. Note that the variance and standard deviation are measures of dispersion in the data, not the certainty of an estimate. Therefore, unlike *standard errors* (

), the standard deviation should not be given with the

sign. This mistake is very frequent in papers in biology.

Very often we are interested in measuring the proportional variation in a variable. The idea is that, because larger things are more variable than small things, we may want to compare how variable different variables are, *proportional to the trait mean*. One common measure is the coefficient of variation (CV), defined as

The CV is also often miscalculated in published studies, so keep you eyes open. Because the standard deviation

is the square root of the variance

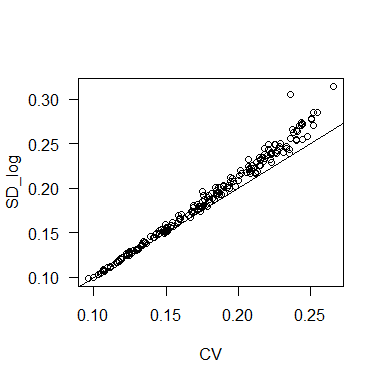
, the CV is often written as

which is sometimes misinterpreted as

.

Another method for placing data on a proportional scale is taking natural logarithms. The standard deviation of a log-transformed variable is very similar to the standard deviation on arithmetic scale divided by the trait mean (the CV), as long as the variance is much smaller than the mean.

EXERCISE: Use simulated data to show the close relationship between the SD of log-transformed data and the CV on aritmetic scale. You may need e.g. a *for*-loop to achieve this.



Another way to see the nice proportional properties of the log-scale is that log(a/b) = -log(b/a)

log(1.1/1)

## [1] 0.09531018

-log(1/1.1)

## [1] 0.09531018

Note also that when a (1.1) is 10% larger than b (1.0), the *log ratio* is ~0.1. Recall that log(a/b) = log(a) - log(b), and thus differences on log scale multiplied by 100 are rougly interpretable as difference in percent (when a and b are not very different).

## The linear model I: Introduction, t-test, ANOVA

1. The linear model I: t-test, ANOVA, fallacy of hypothesis testing. Yoccos? Muff? Confirmatory vs. exploratory analysis ANOVA by hand?

Nearly all the statistical models we will discuss in this text are forms of the linear model

The term

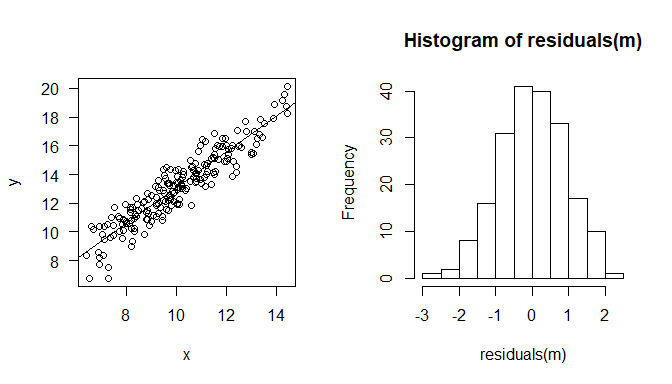
(sometimes denoted

) is the *intercept*, which in the context of a linear regression gives the value of the response variable *y* when the predictor variable *x* is zero. The

are the coefficients (‘slopes’) for the predictor variables *x*, and the

represent the *residuals*, the deviations of each data point from it’s expected value based on the fitted model. In a regression, the residual is the perpendicular distance from the fitted regression line to each data point. The linear model assumes that the residuals (not the data!) are normally distributed.

x = rnorm(200, 10, 2)  
y = 1.3\*x + rnorm(200, 0, 1)  
  
m = lm(y~x)  
  
par(mfrow=c(1,2))  
plot(x, y, las=1)  
abline(m)  
hist(residuals(m))



## 5. The linear model II: Regression and ANCOVA. E: Writing Methods and Results. Peer review.

Regression analysis is another frequent application of the linear model. The aim is to estimate the linear relationship between a response variable (or ‘dependent variable’) and one or more predictor variables (‘independent variables’). The most common form of regression analysis is so-called ordinary least-square (OLS) regression, in which the regression parameters are estimates so as to minimize the square deviations of the data points from the estimated regression line. The deviations are called *residuals*, and are assumed to be normally distributed.

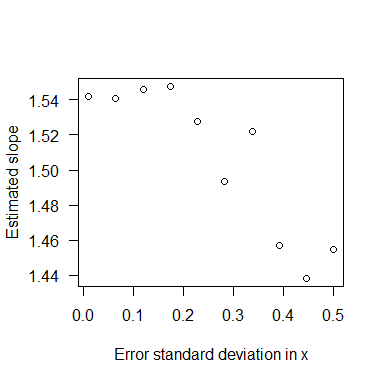
Regression example figure

Although regression slopes are very often reported withour any units, it is important to remember that the slopes in fact carry the units of both the response and predictor variables. If the response is measured in kilograms, and the predictor in mm, the slope will have units of kilograms per mm (kg/mm).

As a special case, a regression where both the response and predictor variable are natural log-transformed will have a slope interpretable as an *elasticity*, which describes the % change in the response per % change in the predictor.

### Exercise: How error in x- and y-variables affect the slope

The standard linear model assumes that the predictor variable is measured wihtout error. When there is measurement error, this can lead to a bias in the estimated slope. Simulate data with measurement error in the predictor, and produce a plot showing the effect on the estimated slope.



## 6. The linear model III: multiple regression, non-linear regression?

## 7. Generalized linear models: logistic regression, Poisson regression, overdispersion, NB reg

# Mean-variance relations for the normal, binomial, and poission distributions

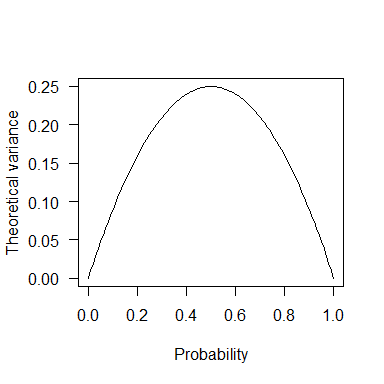
The binomial distribution has two parameters, *n* and *p*, and summarizes a set of so-called Bernouli trials with two possible outcomes, yes (1) or no (0). When we have performed more than one trial, we can compute the proportion *p* of the *n* trials with a positive outcome.

The theoretical variance of the binomial distribution is given by

rbinom(3, 10, c(0.1, 0.5, 0.9))

## [1] 3 7 8

x = seq(from=0, to=1, by=0.01)  
v\_b = x\*(1-x) #Binomial variance  
plot(x, v\_b, type="l", xlab="Probability", ylab="Theoretical variance", las=1)

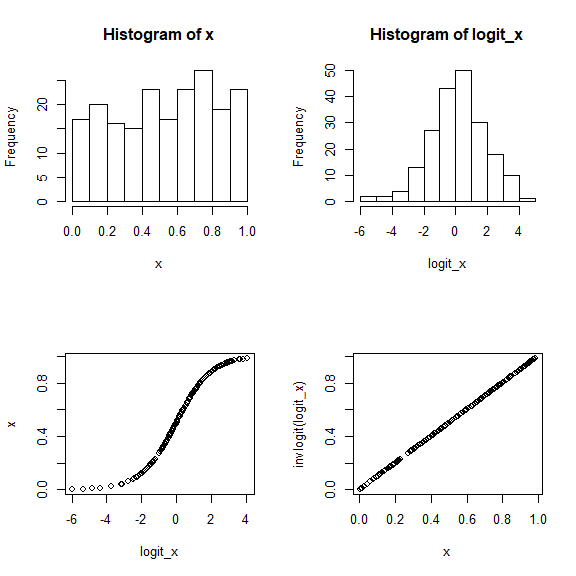


This is important to keep in mind, because it affects how we can compare the (proportional) variation of variables measures as proportions. If e.g. one population has a mean of 0.5, if will be expected to be much more variable than a second population with a mean of 0.1, just because there is less opportunity to vary. This is a now well-recognized problem e.g. in demography research, where the interest is often in comparing the extent of variation in life-history traits measured as proportions, such as germination or survival. One proposed solution is to scale the observed CV by the maximum based on the theoretical varaince, but a perhaps simpler approach is to transform the proportional data in a way that makes them approach a normal distribution.

One previously popular but now not recommended transformation is the so-called arcsin square-root transformation,

A more meaningful transformation is the logit or log odds transformation

logit = function(x) log(x/(1-x))  
invlogit = function(x) 1/(1+exp(-x))  
  
x = runif(200)  
logit\_x = logit(x)  
  
par(mfrow=c(2,2))  
hist(x)  
hist(logit\_x)  
  
plot(logit\_x, x)  
plot(x, invlogit(logit\_x))



The logit transformation is the most common *link function* in a Generalized Linear Model with binomial errors.

Generalized linear models (GLMs) adds flexibility to the linear model by allowing deviations from the usual assumption of normally distributed residuals. Briefly, a GLM consists of the familiar linar predictor of a linear model (often denoted as

),

and a link function

that places the predictor on a Gaussian (normally-distributed) scale. For a binomial GLM (‘logistic regression’), the logit is the most usual link function.

## 8. Mixed-effect models I, information criteria (AIC etc) Variance component analysis

One very common extension of the linear model is the linear mixed model. The ‘mixed’ comes from the fact that these models include two variable types: fixed effects and random effects.

Model equation

The fixed effects are the standard predictor variables of a linear model, i.e. variables for which we are interested in their (independent) effect on the response variable. For example, in an ANOVA-type analysis of a factorial experiments, the experimental factors will be treated as fixed effects.

The random effects are variables for which we are not necessarily interested in the mean value of the response for each value of the predictor, but rather the variance in these effects. A common use of random effects is to account for the non-independence of observations that arise, for example, when several measurements are taken from the same individual. Failing to account for this would lead to an artificial inflation of the degrees of freedom of the analysis. This issue is called *pseudoreplication*, because it uses non-independent data points as replicates.

Beyond modelling patterns of non-independence in the data, random-effect models are also often use to estimate variance components that may of direct interest. A typical applicaiton is in quantitative genetics, where the aim of a study can be to estimate the components of the variance in a phenotypic trait. A simple model can be

where *g* is the genetic variance component ande *e* is the environmental variance component. Estimating these variance components examplifies the general approach of *variance component analysis*.

### Variance component analysis using random-effect models

Random-effect models allow us to estimate the variance residing at multiple levels, and thus to ask for example what percentage of variation in a variable is due to differences among populations, and to differences among individuals within populations.

Consider the following simulated data.

popmeans = rnorm(10, 20, 4)

To specify a random-effect model with the lme4 package, we use the (1|pop format.

#library(lme4)  
#m = lmer(z~1+(1|pop), data=data)

## 9. Bayesian methods, simple MCMC

## 10. Multivariate stats, multivariate mixed models

## 11. Path analysis and structural equation modelling?

## Other topics to cover somewhere

Parametric vs. non-parametric methods