Journal of Apicultural Research 52(4): (2013)
DOI 10.3896/IBRA.1.52.4.13

REVIEW ARTICLE

Statistical guidelines for *Apis mellifera* research

Christian W W Pirk[1*], Joachim R de Miranda[2], Matthew Kramer[3], Tomàs E Murray[4], Francesco Nazzi[5], Dave Shutler[6], Jozef J M van der Steen[7] and Coby van Dooremalen[7]

[1]Social Insect Research Group, Department of Zoology & Entomology, University of Pretoria, Private Bag X20 Hatfield 0028, Pretoria, South Africa.
[2]Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, 750 07, Uppsala, Sweden.
[3]Biometrical Consulting Service, Agricultural Research Service/USDA, Beltsville, MD, 20705, USA.
[4]Institute of Biology, Martin Luther University Halle-Wittenberg, Hoher Weg 8, Halle (Saale) 06120, Germany.
[5]Dipartimento di Scienze Agrarie e Ambientali, Università di Udine, via delle Scienze 206, 33100 Udine, Italia.
[6]Department of Biology, Acadia University, Wolfville, Nova Scotia, B4P 2R6, Canada.
[7]Bees@wur, Bio-Interactions and Plant Health, Plant Research International, Wageningen UR, Droevendaalsesteeg 1, 6708 PB Wageningen, Netherlands.

Received 13 May 2013, accepted subject to revision 15 June 2013, accepted for publication 17 July 2013.

*Corresponding author: Email: cwwpirk@zoology.up.ac.za

Summary

In this article we provide guidelines on statistical design and analysis of data for all kinds of honey bee research. Guidelines and selection of different methods presented are, at least partly, based on experience. This article can be used: to identify the most suitable analysis for the type of data collected; to optimise one's experimental design based on the experimental factors to be investigated, samples to be analysed, and the type of data produced; to determine how, where, and when to sample bees from colonies; or just to inspire. Also included are guidelines on presentation and reporting of data, as well as where to find help and which types of software could be useful.

Guia estadistica para estudios en *Apis mellifera*

Resumen

En este trabajo se proporcionan directrices sobre el diseño estadístico y el análisis de datos para todo tipo de investigación sobre abejas. Tanto las directrices como la selección de los diferentes métodos que se presentan están basadas, al menos en parte, en la experiencia. Este artículo se puede utilizar: para identificar el análisis más adecuado para el tipo de datos recogidos; para optimizar el diseño experimental basado en los factores experimentales a ser investigados, las muestras a analizar, y el tipo de datos que se producen; para determinar cómo, dónde , y cuando muestras abejas de las colonias, o simplemente para inspirar. También se incluyen directrices para la presentación y comunicación de los datos, así como dónde encontrar ayuda y distintos software que puedan ser útiles.

西方蜜蜂研究的统计指南

摘要

在本文中，我们提供了针对蜜蜂所有研究的统计设计和数据分析指南。这些指南和方法的选择至少部分基于我们的经验。本文也可用于：针对收集到的数据类型选择最优分析方法；基于所研究的实验因素、待分析的样本和获得的数据类型优化实验设计；确定从蜂群中采集蜜蜂样本的地点、时间和方式；或者仅为实验提供参考。另外，也包含展示和报告数据时的指南，以及如何寻求帮助和选用何种软件。

Keywords: COLOSS, *BEEBOOK*, honey bees, sampling, sample size, GLMM, robust statistics, resampling, PCA, Power, rule of thumb

Footnote: PIRK, C W W; DE MIRANDA, J R; KRAMER, M; MURRAY, T; NAZZI, F; SHUTLER, D; VAN DER STEEN, J J M; VAN DOOREMALEN, C (2013) Statistical guidelines for *Apis mellifera* research. In *V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS* BEEBOOK*, Volume I: standard methods for* Apis mellifera *research. Journal of Apicultural Research* 52(4): http://dx.doi.org/10.3896/IBRA.1.52.4.13

Table of Contents

1. Introduction

Bees are organisms and, as such, are inherently variable at the molecular, individual, and population levels. This intrinsic variability means that a researcher needs to separate the various sources of variability contained in the measurements, whether obtained by observational or experimental research, into signal and noise. The former may be due to treatments received, bee age, or innate differences in resistance. The latter is largely due to the genetic background (and its phenotypic expression) that characterises individual living organisms. Statistics is the branch of mathematics we use to isolate and quantify the signal and determine its importance, relative to the inherent noise. For the researcher, with an eye toward the statistical analysis to come, and before data collection starts, one should ask:

1) Which variables (VIM, 2008) am I going to measure and what kind of data will those variables generate?
2) What degree of accuracy do I want to achieve and what is the corresponding sample size required?
3) Which statistical analysis will help me to answer my research question? This is related to the question. What kind of underlying process produces data like those I will be collecting?
4) From what population do I want to sample? (What is the statistical population/ statistical universe?) For example, do I want to make inferences about the local, national, continental, or worldwide population?

One function of statistics is to summarise information to make it more usable and easier to grasp. A second is inductive, where one makes generalisations based on a subset of a population or based on repeated observations (through replication or repeated over time). For example, if 50 workers randomly sampled from 20 colonies all produce 10-hydroxydecanoic acid (10-HDAA, one of the major components in the mandibular gland secretion, especially in workers; Crewe, 1982; Pirk *et al.*, 2011), one could infer that all workers produce 10-HDAA. An example of inferring a general pattern from repeated observations would be: If an experiment is repeated 5 times and

yields the same result each time, one makes a generalisation based on this limited number of experiments. One should keep in mind that, if one is measuring a quantitative variable, irrespective of how precise measuring instruments are, each experimental unit/replicate produces a unique data value. A third function of statistics is based on deductive reasoning and might involve statistical modelling, in the classical or Bayesian paradigm, to understand the basic processes that produced the measurements, possibly by incorporating prior information (e.g. predicting species distributions or phylogenetic relationships/trees; see Kaeker and Jones, 2003). In this article we will cover, albeit only cursorily, all three functions of statistics. We have largely focused on research with bee pathogens, in part because these are of intense practical and theoretical interest, and in part because of our own backgrounds. However, bee biology rightly includes a much greater spectrum of research, and for much of it there are specialised statistical tools. Some of the ones we discuss are broadly applicable but, by necessity, this section can only provide an uneven treatment of current statistical methods that might be used in bee research. In particular, we do not discuss multivariate methods (other than principal component analysis); Bayesian approaches, and touch only lightly on simulation and resampling methods. All are current fields of investigation in statistics. Molecular, and in particular, genomic research has spawned substantial new statistical methods, also not covered here. These areas of statistics will be included in the next edition of the *BEEBOOK*.

Furthermore, we restrict ourselves here to providing guidelines on statistics for certain kinds of honey bee research, as mentioned above, with reference to more detailed sources of information. Fortunately, there are excellent statistical tools available, the most important of which is a good statistician.

The statistics we describe can be roughly grouped into two main areas, one having to do with sampling to estimate population characteristics (e.g. for pathogen prevalence = proportion of infected bees in an apiary or a colony), and the other having to do with experiments (e.g. comparing treatments, one of which may be a control). Due to the complex social structure of a bee hive, and the peculiar developmental and environmental aspects of bee biology, sampling in this discipline has more components to consider than in most biological fields. Some statistical topics are relevant to both sampling and experimental studies, such as sample size and power. Others are primarily of concern for just one of the areas. For example, when sampling for pathogen prevalence, primary issues include representativeness, and how or when to sample. For experiments, they include hypothesis formulation and development of appropriate statistical models for the processes (which includes testing and assumptions of models). Of course, good experiments require representative samples, and also require a good understanding of sampling. Both areas are important for data acquisition and analysis. We start with statistical issues related to sampling.

1.1. Types of data

There are several points to consider in selecting a statistical analysis including sample size, distribution of the data, and type of data. These points and the statistical analysis in general should be considered **before** conducting an experiment or collecting data. One should know beforehand what kind of measurement and what type of data one is collecting. The dependent variable is the variable that may be affected by which treatment a subject is given (e.g. control *vs.* treated, an ANOVA framework), or as a function of some other measured variable (e.g. age, in a regression framework). Data normally include all measured quantities of an experiment (dependent and independent/predictor/ factor variables). The dependent variable can be one of several types: nominal, ordinal, interval or ratio, or combinations thereof. An example of nominal data is categorical (e.g. bee location A/B/C, where the location of a bee is influenced by some explanatory variables, such as age) or dichotomous responses (yes/no). Ordinal data are also categorical, but which can be ordered sequentially. For example, the five stages of ovarian activation (Hess, 1942; Schäfer *et al.*, 2006; Pirk *et al.*, 2010; Carreck *et al.*, 2013) are ordinal data because undeveloped ovaries are smaller than intermediate ovaries, which are smaller than fully developed. However, one cannot say intermediate is half of fully developed. If one assigned numbers to ranked categories, one could calculate a mean, but it would be most likely a biologically meaningless value. The third and fourth data types are interval and ratio; both carry information about the order of data points and the size of intervals between values. For example, temperature in Celsius is on an interval scale, but temperature in Kelvin is on a ratio scale. The difference is that the former has an arbitrary "zero point" and negative values are used, whereas the latter has an absolute origin of zero. Other examples of data with an absolute zero point are length, mass, angle, and duration.

The type of dependent variable data is important because it will determine the type of statistical analysis that can or cannot be used. For example, a common linear regression analysis would not be appropriate if the dependent variable is categorical. (Note: In such a case a logistic regression, discussed below, may work).

1.2. Confidence level, Type I and Type II errors, and Power

For experiments, once we know what kind of data we have, we should consider the desired confidence level of the statistical test. This confidence is expressed as α; it gives one the probability of making a Type I error (Table 1) which occurs when one rejects a **true** null hypothesis. Typically that level for α is set at 0.05, meaning that we are 95% confident ($1 - \alpha = 0.95$) that we will not make a Type I error, i.e. 95% confident that we will *not* reject a true null hypothesis. For many commonly used statistical tests, the p-value is the probability that the test statistic calculated from the observed data occurred by chance, given that the null hypothesis is true. If $p < \alpha$ we reject the null hypothesis; if $p \geqq \alpha$ we do not reject the null hypothesis.

Table 1. The different types of errors in hypothesis-based statistics.

	The null hypothesis (H_0) is	
Statistical result	True	False
Reject null hypothesis	**Type I error,** α value = probability of falsely rejecting H_0	Probability of correctly rejecting H_0: (1 - ß) = power
Accept null hypothesis	Probability of correctly accepting H_0 : (1 - α)	**Type II error,** ß value = probability of falsely accepting H_0

A Type II error, expressed as the probability ß occurs when one fails to reject a **false** null hypothesis. Unlike α, the value of ß is determined by properties of the experimental design and data, as well as how different results need to be from those stipulated under the null hypothesis to make one believe the alternative hypothesis is true. Note that the null hypothesis is, for all intents and purposes, rarely true. By this we mean that, even if a treatment has very little effect, it has some small effect, and given a sufficient sample size, its effect could be detected. However, our interest is more often in biologically important effects and those with practical importance. For example, a treatment for parasites that is only marginally better than no treatment, even if it could be shown to be statistically significant with a sufficiently large sample size, may be of no practical importance to a beekeeper. This should be kept in mind in subsequent discussions of sample size and effect size.

The power or the sensitivity of a test can be used to determine sample size (see section 3.2.) or minimum effect size (see section 3.1.3.). Power is the probability of correctly rejecting the null hypothesis when it is false (power = 1 – ß), i.e. power is the probability of not committing a Type II error (when the null hypothesis is false) and hence the probability that one will identify a significant effect when such an effect exists. As power increases, the chance of a Type II error decreases. A power of 80% (90% in some fields) or higher seems generally acceptable. As a general comment the words "power", "sensitivity", "precision", "probability of detection" are / can be used synonymously.

2. Sampling

2.1. Where and when to sample a colony

Colony heterogeneity in time and space are important aspects to consider when sampling honey bees and brood. For example, the presence and prevalence of pathogens both depend on the age class of bees and brood, physiological status of bees, and/or the presence of brood. Note that pathogens in a colony have their own biology and that their presence and prevalence can also vary over space and time. The relation between a pathogen and particular features of a colony should be taken into account when deciding where and when samples are taken, including the marked seasonality of many pathogen infections.

2.1.1. When to sample?

A honey bee colony is a complex superorganism with changing features in response to (local) seasonal changes in the environment. Average age increases, for example, in colonies in the autumn in temperate regions, because of the transition to winter bees. Age-related tasks are highly plastic (Huang and Robinson, 1996), but after a major change of a colony's organisation it can take some time before the division of labour is restored (Johnson, 2005). Immediately after a colony has produced a swarm, for example, bees remaining in the nest will have a large proportion of individuals younger than 21 days, lowering the average age of bees in the colonies. Over time, these bees will become older and the average age of bees in the colony will increase again. Therefore, it is recommended that if the aim is to have an average / normal / representative sample with respect to age structure, one should only sample established colonies that have not recently swarmed. The same is true for recently caught swarms, because brood will not have had enough time to develop, and one could expect rather an over-aged structure. Age polyethism in honey bees and its implications for the physiology, behaviour, and pheromones is discussed in detail elsewhere (Lindauer, 1952; Ribbands, 1952; Lindauer, 1953; Jassim *et al.*, 2000; Crewe *et al.*, 2004; Moritz *et al.*, 2004).

Furthermore, physiological variables in individual bees (and in pooled samples from a colony) can change over time when these parameters are, for example, related to age of bees or presence of brood. Moreover, build-up of vitellogenin takes place in the first 8-10 days of a bee's adult life and then decreases, but is much faster in summer than in winter when no brood is present and bees are on average older (Amdam and Omholt, 2002), affecting averages of individual bees of the same age, but also averages of pooled samples. *Nosema apis*, *Paenibacillus larvae*, and *Melissococcus plutonius* are examples of organisms with bee age-related prevalence in colonies. *N. apis* infections are not microscopically detectable in young bees; after oral infection it takes three to five days before spores are released from infected cells (Kellner, 1981). *P. larvae* and *M. plutonius* can be detected in and on young bees that clean cells (Bailey and Ball, 1991; Fries *et al.*, 2006). Depending on the disease, higher prevalences can be found in colonies with relatively old and young bees, respectively. Furthermore, seasonal variation in pathogen and parasite loads may also affect when to sample. For example, screening for brood pathogens during brood-less periods (e.g. winter, in temperate climates) is less likely to return positive samples than screening when brood is present.

2.1.2. Where to sample?

To determine proper locations for sampling inside a beehive, one must consider colony heterogeneity in time and space. Feeding brood, and capping and trimming of cells takes place in the brood nest. Other activities such as cleaning, feeding and grooming, honey-storing, and

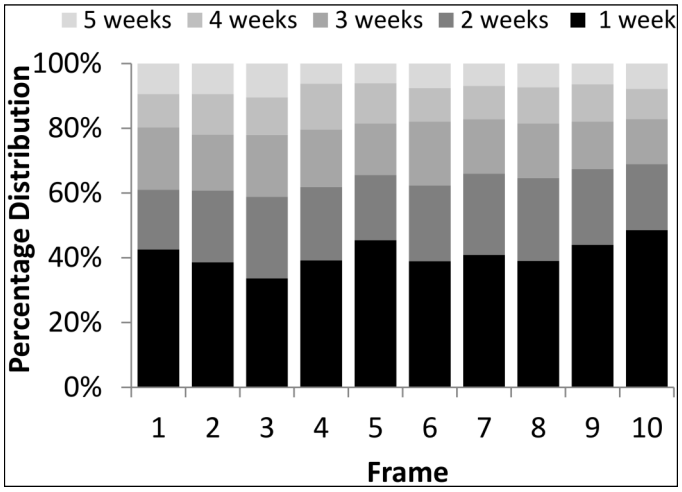

Fig. 1. The percentage distribution of age classes recorded between 24 August and 20 September for pooled colonies of *Apis mellifera* in the Netherlands. The different shades represent different age classes. The distribution of age classes did not differ among frames (p = 0.99). There was also no difference between the mean number of bees per frame (p = 0.94). Adapted from: van der Steen *et al*. (2012).

shaping combs take place all over frames (Seeley, 1985). Free (1960) showed an equal distribution of bees of successive age classes on combs containing eggs, young larvae, and sealed brood, although there were proportionally more young bees (4-5 days old) on brood combs and more old bees (> 24 days) on storage combs. Older bees were overrepresented among returning bees at colony entrances. This was supported by findings of van der Steen *et al*. (2012), who also reported that age classes are distributed in approximately the same ratio over frames containing brood (Fig. 1).

2.2. Probability of pathogen detection in a honey bee colony

For diagnosis or surveys of pathogen prevalence, the more bees that are sampled, the higher the probability of detecting a pathogen, which is particularly important for low levels of infection. An insufficient sample size could lead to a false negative result (apparent absence of a pathogen when it is actually present but at a low prevalence). Historically, 20-30 bees per colony have been suggested as an adequate sample size (Doull and Cellier, 1961) when the experimental unit is a colony. However, based on binomial probability theory, such small sample sizes will only detect a 5% true prevalence in an infected colony with a probability of 65% (20 bees) or 78% (30 bees). If only high infection prevalence is of interest for detection, then small sample sizes may be acceptable, as long as other sampling issues (such as representativeness, see above) have been adequately handled.

In general, sample size should be based on the objectives of the study and a specified level of precision (Fries *et al*., 1984; Table 2). If the objective is to detect a prevalence of 5% or more (5% of bees infected) with 95% probability, then a sample of 59 bees per colony is needed. If the objective is to detect prevalence as low as 1% with 99% probability, then 459 bees per colony are required. Above are tabulated sample sizes (number of bees) needed based on such requirements, provided that every infected bee is detected with 100% efficiency. If detection efficiency is less than 100%, this is the equivalent (for sample size determination) of trying to detect a lower prevalence. For example, if only 80% of bees actually carrying a pathogen are detected as positive using the diagnostic test, then the parameter P below needs to be adjusted (by multiplying P by the proportion of true positives that are detected, e.g. use 0.8*P instead of P if the test flags 80% of true positives as positive). Sample size needed for various probability requirements and infection levels can be calculated from Equation I (Equations adapted from Colton, 1974).

Table 2. Example of sample sizes needed to detect different infection levels with different levels of probability (from Equation I.).

Proportion of infected bees, **P**	Required probability of detection, **D**	Sample size needed, **N**
0.25	0.95	11
0.25	0.99	16
0.10	0.95	29
0.10	0.99	44
0.05	0.95	59
0.05	0.99	90
0.01	0.95	298
0.01	0.99	459

Equation I.

$$N = \ln(1-D) / \ln(1-P)$$

where:

N = sample size (number of bees)
ln = the natural logarithm
D = the probability (power) of detection in the colony
P = minimal proportion of infected bees (infection prevalence), which can be detected with the required power D by a random sample of N bees (e.g. detect an infection rate of 5% or more).

Because the prevalence of many pathogens varies over space and time (Bailey *et al*., 1981; Bailey and Ball, 1991; Higes *et al*., 2008; Runckel *et al*., 2011), it is important, prior to sampling, to specify the minimum prevalence (P) that needs to be detected and the power (D). Colony-to-colony (and apiary-to-apiary) heterogeneity exists and needs to be taken into consideration in sampling designs. For example, a large French virus survey in 2002 (Tentcheva *et al*., 2004; Gauthier *et al*., 2007) showed that for nearly all virus infections there

were considerable differences among colonies in an apiary. This suggests that pooling colonies is a poor strategy for understanding the distribution of disease in an apiary, and that sample size should be sufficient to detect low pathogen prevalence, because the probability of finding no infected bees in a small sample is high if the pathogen prevalence is low, as it may be in some colonies. For a colony with low pathogen prevalence, one might have falsely concluded that the hive is pathogen-free due to low power (D) to detect the pathogen.

For *Nosema* spp. infection in adult bees, the infection intensity (spores per bee) as well as prevalence may change rapidly, particularly in the spring, when young bees rapidly replace older nest mates. To understand such temporal effects on infection intensity or prevalence, sample size must be adequate at each sampling period to detect the desired degree of change (i.e. larger samples are necessary to detect smaller changes). Note that sampling to detect a change in prevalence requires a different mathematical model than simple sampling for prevalence because of the uncertainty associated with each prevalence estimate at different sampling periods. Because, for a binomial distribution, variances are a direct function of sample sizes n_1, n_2, n_3, ..., one can use a rule of thumb which is based on the fact that the variance of a difference of two samples will have twice the variance of each individual sample. Thus, doubling the sample size for each period's sample should roughly offset the increased uncertainty when taking the difference of prevalence estimates of two samples. For determining prevalence, limitations due to laboratory capacity are obviously a concern if only low levels of false negative results can be accepted.

Equation I gives the sample size needed to find a pre-determined infection level (P) with a specified probability level (D) in a sample or, in the case of honey bees, in individual colonies. If we want to monitor a population of colonies and describe their health status, or prevalence in this population, we first have to decide with what precision we want to achieve detection within colonies. For example, for composite samples for *Nosema* spp. or virus detection in which many bees from the same colony are pooled (one yes/no or value per colony), this is not a major concern because we can easily increase the power by simply adding more bees to the pool to be examined. For situations in which individual honey bees from a colony are examined to determine prevalence in that colony, we may not want to increase the power because of the labour involved. But if the objective is to describe prevalence in a population of honey bee colonies, not in the individual colony, we can still have poor precision in the estimates if we do not increase the number of colonies we sample. There could be a trade-off between costs in terms of labour and finance and the precision of estimates of the prevalence in each individual. However, if one decreases the power at the individual level one can compensate by an increase in colonies sampled. The more expensive, or labour intensive, the method for diagnosis of the pathogen is, the more cost effective it becomes to lower the precision of estimates of prevalence in each individual colony, but increase the number of colonies sampled.

2.2.1. Probability of pathogen detection in a colony based on a known sample size

Instead of focusing on sample size, one can calculate the resulting probability of detection of a disease organism using a specific sample size. This probability can be calculated (for an individual colony) using Equation I, but solving for D, as given in Equation II, below.

Equation II.

$$D = 1-(1-P)^N$$

where

D, P, and N are defined as in Equation I above.

For example, within a colony, if the pathogen prevalence in worker bees is 10% (90% of bees are not infected), then the probability of detecting the pathogen in the colony using a sample size of one bee is 0.10, much lower than that for 30 bees (probability is 0.96). A lower prevalence will lower the probability of detection for the same sample size (Fig. 2).

Based on Equation II, it is also possible to calculate the number of bees that need to be tested (sample size) to detect at least one infected bee as a function of the probability, e.g. at a probability of detection (D), of 95% or 99% (Fig. 3). The number of bees to be tested to detect at least one infected bee is higher if one needs a higher probability of detection (D), i.e. when one needs to be able to detect low prevalence.

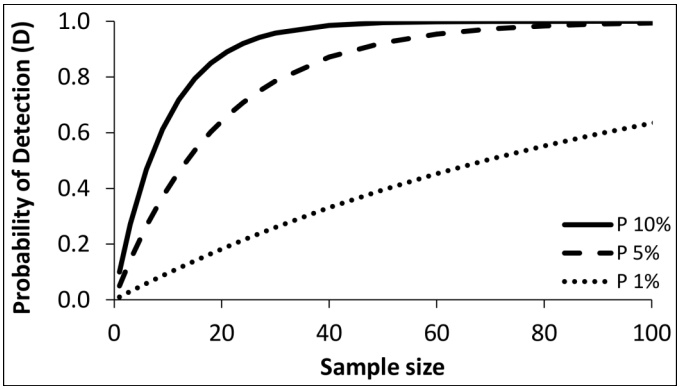

Fig. 2. The probability of detecting a pathogen in a colony (D) as a function of the sample size of bees from that colony, where bees are a completely random sample from the colony. The minimal (true) infection prevalences (P) are 10% (solid line), 5% (dashed line), and 1% (dotted line).

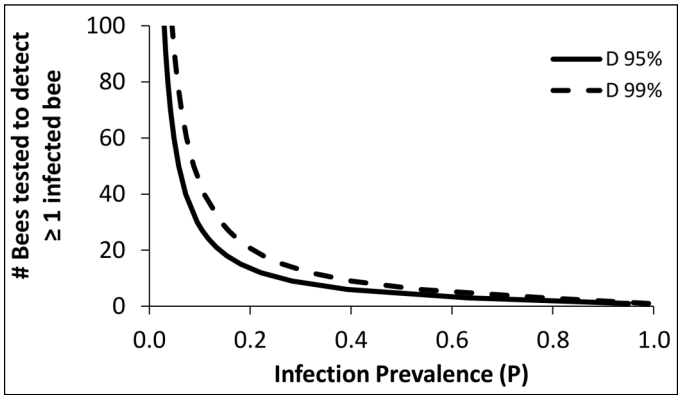

Fig. 3. The number of bees that need to be tested (sample size) to detect at least one infected bee as a function of the prevalence (P), e.g. at a probability of detection (D) of 95% (solid line) or 99% (striped line).

2.2.2. Probability of pathogen detection in a population of colonies

If one wants to calculate the probability of detection of a pathogen in a population of colonies using a known number of colonies, this probability can be calculated (for a population of colonies) according to Equation III.

Equation III.

$E = 1-(1-P*D)^N$

where:

E = the probability of detection (in the population)
D and P are defined as above in Equation I and II, N is the sample size, in this case number of colonies

If one wants to determine if a pathogen is present in a population and the probability of pathogen detection in individual colonies is known (see Equation II above), then one can calculate how many colonies need to be sampled in order to detect that pathogen using Equation IV. The computation now calculates the probability of *at least* one positive recording in two-stage sampling situations (the probability of detection in individual colonies and the probability of detection in the population). The probability of pathogen detection in the population can be calculated using Equation III, or it can be set to 0.95 or 0.99, depending on the power required for the investigation.

Equation IV.

$N = \ln(1-E) / \ln(P * D)$

where:

N, ln, E, P, and D are defined as above in Equations I, II and III

Equation I, II, III, and IV can easily be entered into a spread sheet for calculation of sample sizes needed for different purposes and desired probabilities of detection.

2.2.3. Extrapolating from sample to colony

A confidence interval of a statistical population parameter, for example, the mean detection rate in brood or the prevalence in the population/colony, can be estimated in a variety of ways (Reiczigel, 2003), most of which can be found in modern statistical software. We do not recommend using the (asymptotic) normal approximation to the binomial method; it gives unreasonable results for low and high prevalence. We show here Wilson's score method (Reiczigel, 2003), defined as:

Equation V.

$(2N\hat{p} + z^2 \pm z\sqrt{\{z^2 + 4N\hat{p}\ (1 - \hat{p})\}}) / 2(N + z^2)$,

where N is the sample size; $\hat{p}$ is the observed proportion as used by Reiczigel (2003) to indicate that it is an estimated quantity; and z is the $1 - \alpha/2$ quantile, which can be defined as a critical value/threshold, from the standard normal distribution. A shortcoming for all the methods, not only Wilson's method, is that they assume bees in a sample are independent of each other (i.e. there is no over-dispersion, discussed below section 5.2.), which is typically not true, especially given the transmission routes of bee parasites and pathogens (for a detailed discussion of the shortcoming of all methods of confidence interval calculation, see Reiczigel, (2003)).

If the degree of over-dispersion can be estimated, it can be used to adjust confidence limits, most easily by replacing the actual sample size with the effective sample size (if bees are not independent, then the effective sample size is smaller than the actual sample size). One calculates the effective sample size by dividing the actual sample size by the over-dispersion parameter (see section 5.2.3., design effect or *deff* and see Madden and Hughes (1999) for a complete explanation). The latter can be estimated as a parameter assuming the data are beta-binomial distributed, but more easily using software by assuming the distribution is quasi-binomial. The beta-binomial distribution is a true statistical distribution, the quasi-binomial is not, but the theoretical differences are probably of less importance to practitioners than the practical differences using software.

Estimating the parameters of the stochastic model and / or the distribution which will be used to fit the data, based on a beta-binomial distribution (simultaneously estimating the linear predictor, such as regression type effects and treatment type effects, and the other parameters characterising the distribution), is typically difficult in today's software. On the other hand, there are standard algorithms for estimating these quantities if one assumes the data are generated by a quasi-binomial distribution. Essentially, the latter includes a multiplier (not a true parameter) that brings the theoretical variance, as determined by a function of the linear predictor, to the observed variance. This multiplier may be labelled the over-dispersion parameter in software output.

The quasi-binomial distribution is typically in the part of the software that estimates generalised linear models, and requires

having bees grouped in logical categories (e.g. based on age or location in a colony), and there must be replication (e.g. two groups that get treatment A, two that get treatment B, etc.). In this kind of analysis, for the dependent variable one gives the number of positive bees and the total number of bees for each category (for some software, e.g. in R, one gives the number of positive bees and the number of negative bees for each category).

Prevalence $\hat{p}$ (estimated proportion positive in the population, as in section 2.2.1. and 2.2.2.) and a 95% confidence interval based on Wilson's score method is given in Fig. 4 for sample sizes (N) of 15, 30, and 60 bees. Note that, for the usual sample size of 30, there is still considerable uncertainty about the true infection prevalence (close to 30% if half the bees are estimated to be infected).

3. Experimental design

There are five components to an experiment: hypothesis, experimental design, execution of the experiment, statistical analysis, and interpretation (Hurlbert, 1984). To be able to analyse data in an appropriate manner, it is important to consider one's statistical analyses at the experimental design stage before data collection, a point which cannot be emphasised enough.

Critical features of experimental design include: controls, replication, and randomisation; the latter two components will be dealt with in the next section (3.1.). In terms of a 'control' in an experiment: a negative control group is a standard against which one contrasts treatment effects (untreated or sham-treated control), whereas a positive control group is also often included usually as a "standard" with an established effect (i.e. dimethoate in the case of toxicological studies, see the *BEEBOOK* paper on toxicology by Medrzycki *et al.*, 2013). Additionally, experiments conducted blind or double blind avoid biases from the experimenter or observer. If that is not possible, one should control for the biases of observers by randomly assigning several different observers to different experimental units or by comparing results from one observer with previous observers to quantify the bias so one can account for it statistically when interpreting results of analyses.

3.1. Factors influencing sample size

A fundamental design element for correct analysis is the choice of the sample size used when obtaining the data of interest. Several factors influencing sample size in an experimental setting are considered below.

3.1.1. Laboratory constraints

Laboratory constraints, such as limitations of space and resources, limit sample size. However, one should not proceed if constraints preclude good science (see the *BEEBOOK* paper on maintaining adult workers in cages, Williams *et al.*, 2013).

3.1.2. Independence of observation and pseudo-replication

A second factor in deciding on sample size, and a fundamental aspect of good experimental design, is independence of observations; what happens to one experimental unit should be independent of what happens to other experimental units before results of statistical analyses can be trusted. The experimental unit is the unit (subject, plant, pot, animal) that is randomly assigned to a treatment. Replication is the repetition of the experimental situation by replicating the experimental unit (Casella, 2008). Where observations are not independent, i.e. there are no true replicates within an experiment, we call this pseudo-replication or technical replication. Pseudo-replication can either be: i) temporal, involving repeated measures over time from the same bee, cage, hive, or apiary; or ii) spatial, involving several measurements from the same vicinity. Pseudo-replication is a problem because one of the most important assumptions of standard statistical analysis is independence.

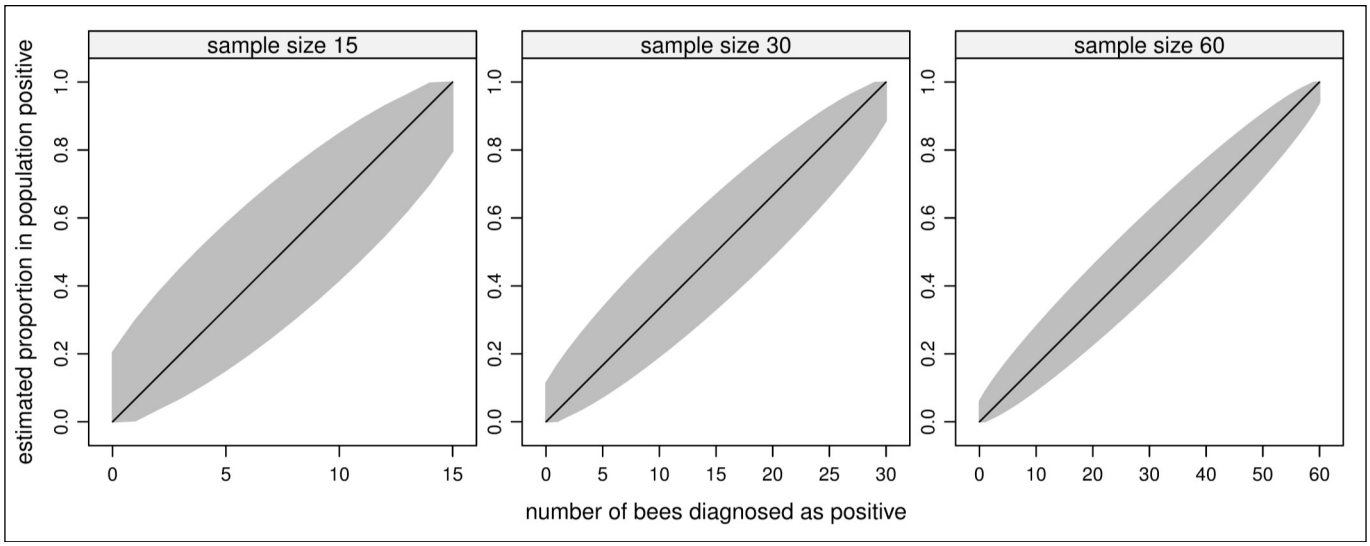

Fig. 4. Estimated proportion of infected bees in a population as a function of the number of bees diagnosed as positive ($\hat{p}$) for various sample sizes (N = 15, 30, 60). Lower and upper limits for a 95% confidence interval are based on Wilson's score method.

Repeated measures through time on the same experimental unit will have non-independent (= dependent) errors because peculiarities of individuals will be reflected in all measurements made on them. Similarly, samples taken from the same vicinity may not have non-independent errors because peculiarities of locations will be common to all samples. For example, honey bees within the same cage might not be independent because measurements taken from one individual can be dependent on the state (behaviour, infection status, etc.) of another bee within the same cage (= spatial pseudo-replication), so each cage becomes the minimum unit to analyse statistically (i.e. the experimental unit). An alternative solution is to try estimating the covariance structure of the bees within a cage, i.e. allow for correlation within a cage in the statistical modelling.

But, are honey bees in *different* cages independent? This and similar issues have to be considered and were too often neglected in the past. Potential non-independence can be addressed by including cage, colony, and any other potentially confounding factors as random effects (or fixed effects in certain cases) in a more complex model (i.e. model the covariance structure imposed by cages, colonies, etc.). If pseudo-replication is not desired and is an unavoidable component of the experimental design, then it should be accounted for using the appropriate statistical tools, such as (generalised) linear mixed models ((G)LMM; see section 5.2.).

Some examples may clarify issues about independence of observations.

Example 1: A researcher observes that the average number of *Nosema* spores per bee in a treated cage is significantly higher than in a control cage; one cannot rule out whether the observed effect was caused by the treatment or the cage.

Possible solution 1: Take cage as the experimental unit and pool the observations per cage; including more cages is statistically preferred (yields more power) to including more bees per cage.

Example 2: Relative to cages containing control bees, experimental cages were housed closer to a fan in the lab, resulting in higher levels of desiccation for the experimental cages and, in turn, higher mortality under constant airflow. In this case, the statistical difference between treatments would be confounded by the experimental design.

Possible solution 2: A rotation system could be included to ensure all cages are exposed to the same environmental conditions i.e. placed at identical distances from the fan and for the same periods of time.

Example 3: Honey bees from treated colonies had high levels of a virus and were *A. mellifera mellifera*, whereas control honey bees from untreated colonies that had low levels of a virus were *A. mellifera ligustica*. In such a case the statistical differences could be due to colony differences and/or to subspecies differences and/or due to the treatment and/or due to interactions.

Possible solution 3: Design the experiment using a factorial design with colony as the experimental unit. For half of the colonies in a treatment, use *A. mellifera mellifera* bees and for the other half use *A. mellifera ligustica* bees. Equal numbers of colonies of both subspecies should then be present in the treatment and control groups. Although equal numbers is not a requirement, it is nevertheless preferable to have a completely balanced design (equal numbers in each group or cell) for several reasons (e.g. highest power, efficiency, ease of parameter interpretation, especially interactions). It is, however, also possible to estimate and test with unbalanced designs. In a balanced design the differences between colonies, subspecies, and treatments (and their interactions!) can be properly quantified.

In essence, there are both environmental and genetic factors (which can also interact) that can profoundly affect independence and hence reliability of statistical inference. The preceding examples illustrate, among other things, the importance of randomising experimental units among different treatments. The final solutions of the experimental design are of course highly dependent on the research question and the variables measured.

In summary, randomisation and replication have two separate functions in an experiment. Variables that influence experimental units may be known or unknown, and random assignment of treatments to cages of honey bees is the safest way to avoid pitfalls of extraneous variables biasing results. Larger sample sizes (i.e. replication: number of colonies, cages, or bees per cage) improve the precision of an estimate (e.g. infection rate, mortality, etc.) and reduce the probability of uncontrolled factors producing spurious statistical insignificance or significance. Researchers should use as many honey bee colony sources from unrelated stock as possible if they want their results to be representative, and hence generalisable. One should also not be too cavalier about randomising honey bees to experimental treatments, or about arranging experimental treatments in any setting, including honey bee cage experiments; sound experimental design at this stage is critical to good science; more details are provided below.

3.1.3. Effect size

A third factor affecting decisions about sample size in experimental design is referred to as effect size (Cohen, 1988). As an illustration, if experimental treatments with a pesticide decrease honey bee food intake to 90% that of controls, more replication is needed to achieve

statistical significance than if food intake is reduced to 10% that of controls (note that one's objective should be to find biologically meaningful results rather than statistical significance). This is because treatment has a greater effect size in the latter situation. Effect size and statistical significance are substantially intertwined, and there are equations, called power analyses (see section 3.2.1.), for calculating sample sizes needed for statistical significance once effect size is known.

Without preliminary trials, effect size, and also statistical power, may be impossible to know in advance. If one's objective is statistical significance, and one knows effect size, one can continue to sample until significance is achieved. However, this approach is biased in favour of a preferred result. Moreover, it introduces the environmental influence of time; results one achieves in spring may not be replicated in summer e.g. Scheiner *et al.* (2003) reported seasonal variation in proboscis extension responses (previously called proboscis extension reflexes; also see Frost *et al.*, 2012). Removing the influence of time requires that one decides in advance of replication, and accepts results one obtains. Without preliminary trials, it will always be preferable to maintain as many properly randomised cages as possible. A related factor that will influence sample size is mortality rate of honey bees in cages; if control group mortality rates are 20% for individual bees, one will want to increase the number of bees by at least 20%, and even more if variability in mortality rates is high. Alternatively, without knowledge of effect size, one should design an experiment with sufficient replicates such that an effect size of biological relevance can be measured.

3.2. Sample size determination

There are many online sample size calculators available on the internet that differ in the parameters required to calculate sample size for experiments. Some are based on the effect size or minimal detectable difference (see section 3.1.3.); for others input on the estimated mean (μ) and standard deviation (δ) for the different treatment groups is required. Fundamentally, the design of the experiment, the required power, the allowed α and the expected effect size dictate the required sample size. The following two sections (3.2.1. and 3.2.2.) suggest strategies for determining sample size.

3.2.1. Power analyses and rules of thumb

Power (1-β) of a statistical test is its ability to detect an effect of a particular size (see section 3.1.3.), and this is intrinsically linked with sample size (N) and the error probability level (α) at which we accept an effect as being statistically significant (see section 1., Table 1). Once we know two of these values, it is possible to calculate the remaining one; in this case for a given α and β, what is N? Power analyses can incorporate a variety of data distributions (normal, Poisson, binomial, etc.), but the computations are beyond the scope of this paper. Fortunately, there are many freely available computer programs that can conduct these calculations (e.g. G*Power; Faul *et al.*, 2007, the R-packages "pwr" and "sample size" online programs can be found at www.statpages.org/#Power) and all major commercial packages also have routines for calculating power and required sample sizes.

A variety of 'rules of thumb' exist regarding minimum sample sizes, the most common being that you should have at least 10-15 data points per predictor parameter in a model; e.g. with three predictors such as location, colony and infection intensity, you would need 30 to 45 experimental units (Field *et al.*, 2012). For regression models (ANOVA, GLM, etc.), where you have k predictors, the recommended minimum sample size should be $50 + 8k$ to adequately test the overall model, and $104 + k$ to adequately test each predictor of a model (Green, 1991). Alternatively, with a high level of statistical power (using Cohen's (1988) benchmark of 0.8), and with three predictors in a regression model: i) a large effect size (> 0.5) requires a minimum sample size of 40 experimental units; ii) a medium effect size (of ca. 0.3) requires a sample size of 80; iii) a small effect size (of ca. 0.1) requires a sample size of 600 (Miles and Shevlin, 2001; Field *et al.*, 2012).

These numbers need to be considerably larger when there are random effects in the model (or temporal or spatial correlations due to some kind of repeated measures, which decreases effective sample size). Random effects introduce additional parameters to the model, which need to be estimated, but also inflate standard errors of fixed parameters. The fewer the levels of the random effects (e.g. only three colonies used as blocks in the experiment), the larger the inflation will be. Because random factors are estimated as additional variance parameters, and one needs approximately 30 units to estimate a variance well, increasing the number of levels for each random effect will lessen effects on fixed parameter standard errors. That will also help accomplish the goals set in the first place by including random effects in a designed experiment: increased inference space and a more realistic partitioning of the sources of variation. We recommend increasing the number of blocks (up to 30), with fewer experimental units in each block (i.e. more, smaller blocks), as a general principle to improve the experimental design. Three (or the more common five) blocks is too few. Fortunately, there are open source (R packages "pamm" and "longpower") and a few commercial products (software NCSS PASS, SPSS, STATISTICA) which could be helpful with estimating sample sizes for experiments that include random effects (or temporally or spatially correlated data).

If random effects are considered to be fixed effects and one uses the methods described above for sample size estimation or power, required sample sized will be seriously underestimated and power seriously overestimated. The exemplary data set method (illustrated for GLMMs and in SAS code in Stroup (2013), though easily ported to other software that estimates GLMMs) and use of Monte-Carlo methods (simulation, example explained below, though it is not for a model with random effects) are current recommendations. For count data (binomial, Poisson distributed), one should always assume there will be over-dispersion (see section 5.2.3.).

3.2.2. Simulation approaches

Simulation or 'Monte-Carlo' methods (Manly, 1997) can be used to work out the best combination of "bees/cage" x "number of cages/ group" given expected impact of a certain treatment. Given a certain average life span and standard deviation for bees of a control group and a certain effect of a treatment (in terms of percentage reduction of the life span of bees), one can simulate a population of virtual bees each with a given life span. Then a program can test the difference between the treated group and the control group using increasing numbers of bees (from 5 to 20) and increasing numbers of cages (from 3 to 10). The procedure can be repeated (e.g. 100 times) and a table produced with the percentage of times a significant difference was achieved using any combination of bees/cage x number of cages/ group.

A program using a *t*-test to determine these parameters is given as online supplementary material (http://www.ibra.org.uk/ downloads/20130812/download). It is assumed that the dependent variables in the bee population are normally distributed. The simulation can be run another 100 times simply by moving the mouse from one cell to another. Alternatively, automatic recalculation can be disabled in the excel preferences.

3.2.3. Sample size and individual infection rates

Common topics in honey bee research are pathogens. Prevalence of pathogens can be determined in a colony or at a population level (see section 2.2.). Most likely, the data will be based on whether in the smallest tested unit the pathogen is present or not: a binomial distribution. Hence, sample size will be largely dependent on detection probability of a pathogen. However, with viruses (and possibly other pathogens), concentration of virus particles is measured on a logarithmic scale (Gauthier *et al.*, 2007; Brunetto *et al.*, 2009). This means, for example, that the virus titre of a pooled sample is disproportionately determined by the one bee with the highest individual titre. For the assumption of normality in many parametric analyses we suggest a power-transformation of these data (Box and Cox, 1964; Bickel and Doksum, 1981). For further reading on sample-size determination for log-normal distributed variables, see Wolfe and Carlin (1999).

In summary, a minimum sample of 30 independent observations per treatment (and the lowest level of independence will almost always be cages) may be desirable, but constraints and large effect sizes will lower this quantity, especially for experiments using groups of caged honey bees. Because of this, development of methods for maintaining workers individually in cages for a number of weeks should be investigated. This would be an advantage because depending on the experimental question, each honey bee could be considered to be an independent experimental unit. The same principles of experimental design that apply to the recommended number of cages also apply to other levels of experimental design, such as honey bees per cage, with smaller effect sizes and more complex questions, recommended sample sizes necessarily increase (in other words the more variables/factors included, the greater the sample size has to be). Researchers must think about, and be able to justify, how many of their replicates are truly independent; 30 replicates is a reasonable starting point to aim for when effect sizes are unknown, but again, this may not be realistic. In the context of wax producing and comb building, colony size and queen status play a role. For example, comb construction only takes place in the presence of a queen and at least 51 workers, and egg-laying occurs only if a mated queen is surrounded by at least 800 workers (reviewed in Hepburn, 1986; page 156). Additionally, novel experiments on new sets of variables means uncertainty in outcomes, but more importantly means uninformed experimental designs that may be less than optimal. Designs should always be scrutinised and constantly improved by including preliminary trials, which could, for example, provide a better idea of prevalence resulting in a better estimate of the required sample size.

4. A worked example

Although a single recommended experimental design, including sample size, may be difficult to find consensus on given the factors mentioned above, we provide below a recommendation for experimental design when using groups of caged honey bees to understand the impact of a certain factor (e.g. parasite or pesticide) on honey bees. For our example, imagine the focus is the impact of the gut parasite *Nosema ceranae* and black queen cell virus on honey bees. One should consider the following:

1. Each cage should contain the same number of honey bees, and be exposed to the same environmental conditions (e.g., temperature, humidity, feeding regime, see the *BEEBOOK* paper on maintaining adult workers in cages, Williams *et al.*, 2013). Each cage of treated honey bees is a single experimental unit, or unit of replication. Because there is no other restriction on randomisation, other than the systematic sampling from different colonies for each replicate (see below), this is a completely randomised design. If one instead put only bees from one colony in a cage, but made sure that all treatments were evenly represented for each colony (e.g. 5 cages from colony A get treatment 1, 5 cages from colony A get treatment 2, etc.), then we would have a randomised complete block design.

2. We recommend 4-9 replicate cages per treatment. Honey bees should be drawn from 6-9 different colonies to constitute each replicate and equal numbers of honey bees from each source colony should be placed in each cage (i.e., in all cages,

including controls and treatments) to eliminate effects of colony; this makes only cage a random factor. For example, if one draws honey bees from 6 source colonies and wants cages to contain 24 honey bees each, then one must randomly select 4 honey bees from each colony for each cage. If one wants to keep colony and cage both as random effects (e.g. to estimate effects of a pathogen on bees from a population of colonies, only some of which were sampled), one should not mix bees from different colonies in the cages. Note that the minimum number of bees also depends on the experimental design. Darchen (1956, 1957) showed that comb construction only started with a minimum number of 51 -75 workers and a queen; in cases of a dead queen, 201-300 workers were needed (summarised in Table 14.1 in Hepburn, 1986). Furthermore, cage design itself can influence behaviour (Köhler *et al.,* 2013) therefore identical cages should be used for all replicates. A group size of 15 workers ensures that the impact of experimentally administered *Nosema* and black queen cell virus in honey bees in general is measured, as opposed to impacts of these parasites on a specific honey bee colony. It also ensures that chance stochastic events, such as all the honey bees dying in a specific treatment cage, do not unduly affect the analysis and interpretation of results. Low numbers of source colonies (i.e. low numbers of replicates) could lead to an over- or under-estimation of the impact of the studied factor(s). A computer simulation based on Monte-Carlo methods (see section 3.2.3.) and parametric statistics supports the appropriateness of the proposed values. Experiments across replicate colonies must be conducted at approximately the same time, because effects such as day length and seasonality can introduce additional sources of error (see section 2.1.1. and section 3. for relationships between model complexity and sources of error).

5. Statistical analyses

5.1. How to choose a simple statistical test

Before addressing the question of how to choose a test, we describe differences between parametric and non-parametric statistics. As stated in the introduction, one has to know what kind of data one has or will obtain. In the discussion below, we use a traditional definition of "parametric" versus "non-parametric tests". In all statistical tests, parameters of one kind or another (means, medians, etc.) are estimated. The distinction has grown murkier over the years as more and more statistical distributions become available for use in contexts where previously only the normal distribution was allowed (e.g. regression, ANOVA). "Parametric" tests assume (1) models where the residuals (the variation that is not explained by the explanatory variables one is testing, i.e. inherent biological variation of the experimental units), following fitting a linear predictor of some kind, are normally distributed, or that the data follow a (2) Poisson, multinomial, or hypergeometric distribution. This definition holds for simple models only; parametric models are actually a large class of models where all essential attributes of the data can be captured by a finite number of parameters (estimated from the data), so include many distributions and both linear and non-linear models, but the distribution(s) must be specified when analysing the data. The complete definition is quite mathematical. A non-parametric test does not require that the data be samples from any particular distribution (i.e. they are distribution-free). This is the feature that makes them so popular.

For models based on the normal distribution, this *does not* mean that the dependent variable is normally distributed; in fact one hopes it is multimodal, with a different mode for each different treatment. However, if one subtracts (or conditions on) the linear predictor (e.g. subtract each treatment mean from its group of observations), the distribution of each resulting group (and all groups combined) follows the same normal distribution. Also, the discussion below pertains only to "simple" statistical tests and where observations are independent.

Note that chi-square and related tests are often considered "non-parametric" tests. This is incorrect; they are very distribution dependent (data must be drawn from Poisson, multinomial, or hypergeometric distributions), and observations must be independent. Whereas "non-parametric" tests may not require that one samples from a particular distribution, they do require that each set of samples come from the same general distribution. That is, one sample cannot come from a right-skewed distribution and the other from a left-skewed distribution; both must have the same degree of skew and in the same direction. Note that when one has dichotomous (Yes/No) or categorical data, non-parametric tests will be required if we stay in the realm of "simple" statistical tests (Fig. 4). For parametric statistics based on the normal distribution, an important second assumption is that the variance among groups of residuals is similar (homogeneous variances, also called homoscedasticity) (as shown in Fig. 5a) and not heterogeneous variances (heteroscedasticity, Fig. 5b). If only one assumption is violated, a parametric statistic is not applicable. The alternative in such a case would be to either transform the data (see Table 4 and section 5.2.), so that the transformed data no longer violate assumptions, or to conduct non-parametric statistics. The advantage of non-parametric statistics is that they do not assume a specific distribution of the data; the disadvantage is that the power (1-ß, see section 1.) is lower compared to their parametric counterparts (Wasserman, 2006), though the differences may not be great. Power itself is not of such great concern because biologically relevant effects shall be detected with a large enough effect size in a **well-designed experiment**. Table 3 provides a comparison between parametric and non-parametric statistics.

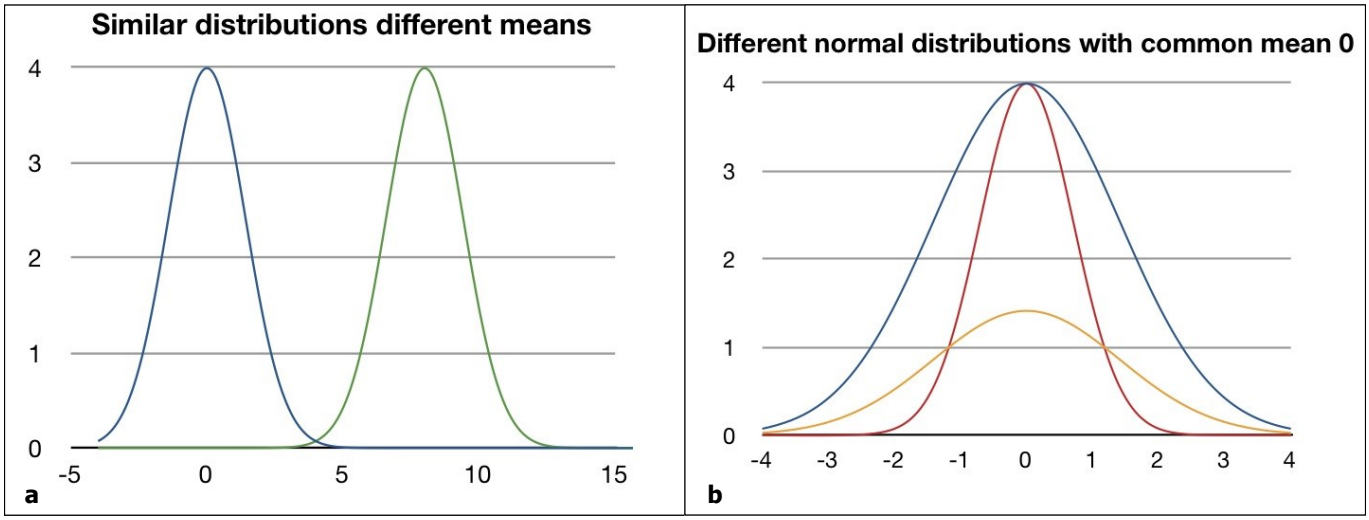

Fig. 5. **a.** Two similar distributions with different means, where variances of the two groups are homogeneous; **b.** shows three different distributions where the means are the same but the variances of three groups are heterogeneous.

Table 3. Comparison between parametric and non-parametric statistics.

	Parametric	**Non-parametric**
Distribution	Normal	Any
Variance	Homogenous	Any
General data type	Interval or ratio (continuous)	Interval, ratio, ordinal or nominal
Power	Higher	Lower
Example Tests		
Correlation	Pearson	Spearman
Independent data	t-test for independent samples	Mann-Whitney U test
Independent data more than 2 groups	One way ANOVA	Kruskal Wallis ANOVA
Two repeated measures, 2 groups	Matched pair t-test	Wilcoxon paired test
Two repeated measures, > 2 groups	Repeated measures ANOVA	Friedman ANOVA

5.1.1 Tests for normality and homogeneity of variances

The flow diagram in Fig. 6 gives a simple decision tree to choose the right test; for more examples, see Table 5. Starting at the top, one has to make a decision based on what kind of data one has. If two variables are categorical, then a chi-square test could be applicable. When investigating the relationship between two continuous variables, a correlation will be suitable. In the event one wants to compare two or more groups and test if they are different, one follows the pathway "difference". The next question to answer is how many variables one wants to compare. Is it one variable (for example the effect of a new varroa treatment on brood development in a honey bee colony), or is it the effect of varroa treatment and supplementary feeding on brood development? For the latter, one could conduct a 2-way ANOVA or an even more complex model depending on the actual data set. For the former, the next question would be "how many treatments?"; sticking with the example, does the experiment consist of two groups (control and treatment) or more (control and different dosages of the treatment)? In both cases, the next decision would be based on if the data sets are independent or dependent. Relating back to the example, one could design the experiment where some of the colonies are in the treatment group and some in the control, in which case one could say that the groups are independent. However, one could as well compare before and after the application of the varroa agent, in which case all colonies would be in the before (control) and after (treatment) group. In this case it is easy to see that the before might affect the after or that the two groups are not independent. A classical example of dependent data is weight loss in humans before and after the start of diet; clearly weight loss depends on starting weight.

To arrive at an informed decision about the extent of non-normality or heterogeneity of variances in your data, a critical first step is to plot your data: i) for correlational analyses as in regression, use a scatterplot ii) for 'groups' (e.g. levels of a treatment factor), use a histogram or box plot; it provides an immediate indication of your data's distribution, especially whether variances are homogeneous. The next step would be to objectively test for departures from normality and homoscedasticity. Shapiro-Wilks W, particularly for sample sizes < 50, or Lilliefors test, can be used to test for normality, and the Anderson-Darling test is of similar if not better value (Stephens, 1974). Similarly, for groups of data, Levene's test tests the null hypothesis that different groups have equal variances. If tests are significant, assumptions that a distribution is normal or its variances are equal **must** be rejected and either the data has to be transformed, a non-parametric test or generalised linear model applied.

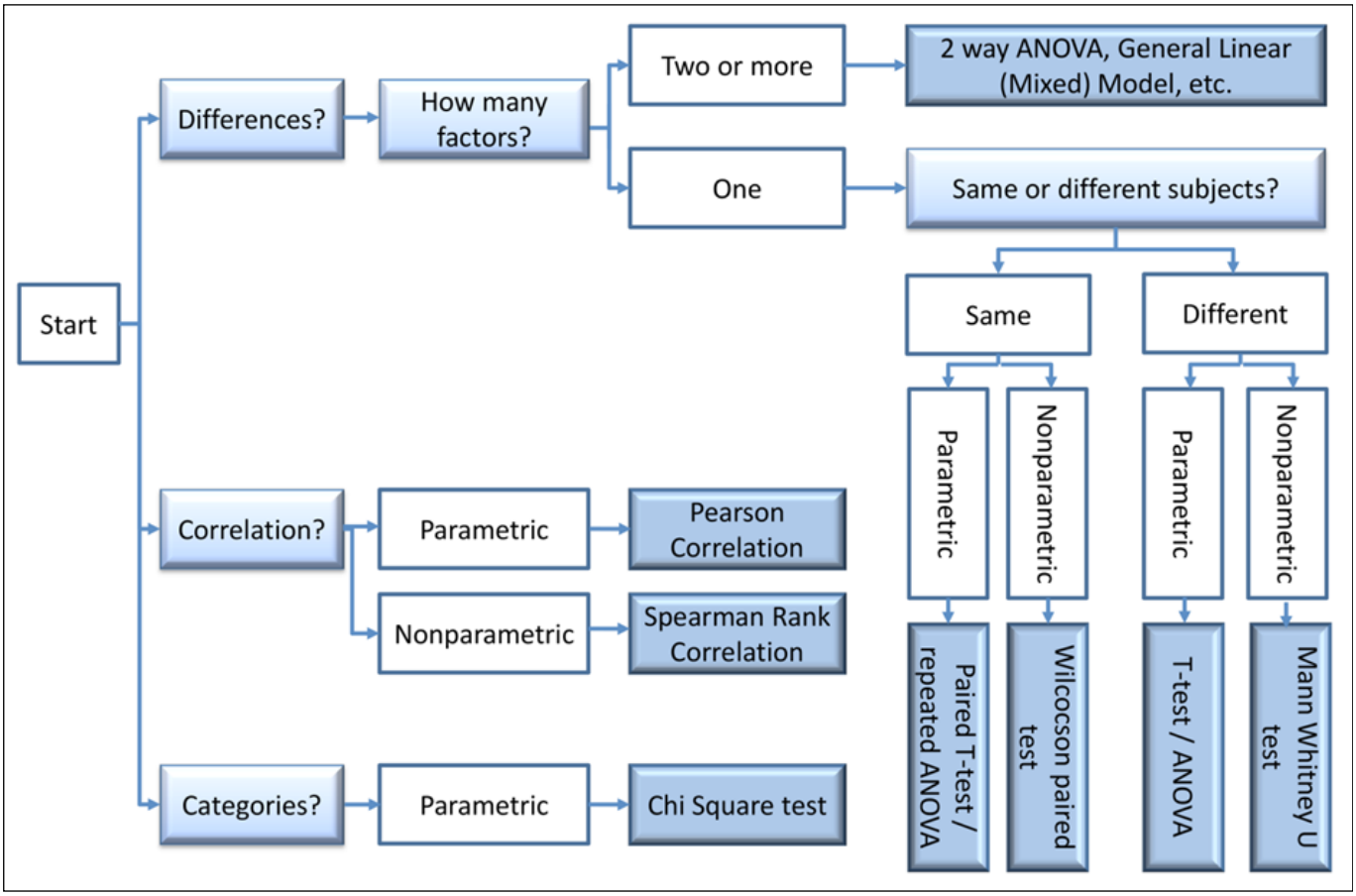

Fig. 6. A basic decision tree on how to select the appropriate statistical test is shown.

Table 4. Common underlying distributions for generalised linear models and their canonical link functions.

Distribution	Canonical Link
Gaussian	identity (no transformation)
Poisson	log
Binomial	logit
Gamma	inverse

5.2. Generalised Linear Mixed Models (GLMM)

A central dogma in statistical analyses is always to apply the simplest statistical test to your data, but ensure it is applied correctly (Zuur *et al.*, 2009). Yes, you could apply an ANOVA or linear regression to your data, but in the **vast majority of cases**, the series of assumptions upon which these techniques are based are violated by 'real world' data and experimental designs, which often include blocking or some kind of repeated measures. The assumptions typically violated are: i) normality; ii) homogeneity; and iii) independence of data.

1. Normality

Although some statistical tests are robust to minor violations of normality (Sokal and Rohlf, 1995; Sokal and Rohlf, 2012), where your dependent variable/data (i.e. the residuals, see section 5.1.) are clearly not normal (positively/negatively skewed, binary data, etc.), a better approach would be to account for this distribution within your model, rather than ignore it and settle for models that poorly fit your data. As an obvious example, a process that produces counts will not generate data values less than zero, but the normal distribution ranges from -∞ to +∞.

2. Homogeneity of variances

As stated above, minor violations of normality can be tolerated in some cases, and the same could be said for heterogeneous dependent variable/data (non-homogenous variance across levels of a predictor in a model, also called heteroscedasticity). However, marked heterogeneity fundamentally violates underlying assumptions for linear regression models, thereby falsely applying the results and conclusions of a parametric model, making results of statistical tests invalid.

3. Independence of data

See section 3.1.2. Simply, if your experimental design is hierarchical (e.g. bees are in cages, cages from colonies, colonies from apiaries) or involves repeated measures of experimental units, your data strongly violate the assumption of independence and invalidate important tests such as the *F*-test and *t*-test; these tests will be too liberal (i.e. true null hypotheses will be rejected too often).

Table 5. Guideline to statistical analyses in honey bee research including examples/ suggestions for tests and graphical representation. Blank fields indicate that a wide variety of options are possible and all have pros and cons.

Subject	**Variable**	**Short description**	**Fields of research where it is used**	**Synthetic representation**	**Measure of dispersion**	**Statistical test**	**Graphical representation**	**Notes**
Honey bee	Morphometric variables (e.g. fore-wing angles)	Measures related to body size. Other data can be included here such as, for example, cuticular hydrocarbons	Taxonomic studies	Average	Standard deviation	Parametric tests such as ANOVA. Multivariate analysis such as PCA and DA	Bar charts for single variables, scatterplots for PC, DA	Please note that some morphometric data are ratios; consider possible deviations from normality
	Physiological parameters (e.g. concentration of a certain compound in the haemolymph)	Measures related to the functioning of honey bee systems		Average	Standard deviation		Bar charts or lines	
	Survival			Median	Range	Kaplan Meyer Cox hazard	Bar charts or lines scatterplots	
Pathogens (e.g. DWV, Nosema)	Prevalence	Proportion of infected individuals	Epidemiological studies	Average	Standard deviation can be used but transfor-mation is necessary due to non-normal distri-bution	Fisher exact solution or Chi square according to sample size	Bar charts, pie charts	
	Infection level	Number of pathogens (e.g. viral particles)	Epidemiological studies, studies on bee-parasite interaction	Average		Parametric tests (e.g. t test/ ANOVA) can be used after log transformation otherwise non parametric tests can be used (e.g. Mann-Whitney/Kruskal-Wallis)		
Parasites (e.g. *Varroa destructor*)	Fertility	Proportion of reproducing females	Factors of tolerance, biology of parasites	Average	Range	Fisher exact solution or chi square according to sample size		
	Fecundity	Number of offspring per female	Factors of tolerance, biology of parasites	Average	Standard deviation			

GLMMs are a superset of linear models, they allow for the dependent variable to be samples from non-normal distributions (allowed distributions have to be members of the one and two parameter exponential distribution family; this includes the normal distribution, but also many others). For distributions other than the normal, the statistical model produces heterogeneous variances, which is a desired result if they match the heterogeneous variances seen in the dependent variable. The 'generalised' part of GLMM means that, unlike in linear regression, the experimenter can choose the kind of distribution they believe underlies the process generating their data. The 'mixed' part of GLMM allows for random effects and some degree of non-independence among observations. Ultimately, this level of flexibility within GLMM approaches allows a researcher to apply more rigorous, but biologically more realistic, statistical models to their data.

One pays a price for this advantage. The basic one is that the state of statistical knowledge in this area, especially computational issues, lags behind that for models based on the normal distribution. This translates into software that is buggy, which can result in many kinds of model estimation problems. Also, there are now far more choices to be made, such as which estimation algorithm to use (e.g. the Laplace and quadrature methods do not allow for correlation among observations), and which link function to use. The link function "links" the data scale to the model scale (Table 4). For example, if dependent variable is assumed to be generated by a Poisson process, the typical link function is the log, i.e. $\log(E(\mathbf{\mu})) = \mathbf{X\beta} + \mathbf{ZU}$; in words, the natural log of the expected value of the mean is modelled as a sum of fixed and random effects). Tests are based on asymptotic behaviours of various quantities, which can give quite biased results for small samples. One is simultaneously working on two scales: the data scale and the model scale; the two are linked, but model estimates and hypothesis tests are done on the model scale, and so are less easily interpretable (i.e. a change in unit value of a predictor variable has different effects on the data scale depending on whether one is looking at low values or high values). One parameter and two parameter members of the exponential family have to be handled quite differently.

Over-dispersion (see section 5.2.3.) cannot be handled using a quasi-likelihood approach (e.g. using a quasi-binomial distribution); instead, appropriate random effects need to be added (e.g. one for every observation), which can lead to models with many parameters (Note: Over-dispersion means that one has a greater variability than expected based on the theoretical statistical distribution; for example the expected variance of a Poisson distribution is its mean - if the observed variance is larger than the estimated mean, then there is over-dispersion). For some one-parameter members of the exponential distribution (e.g. Poisson, binomial), one can try the analogous two-parameter member (e.g. for a Poisson distribution, it is the negative binomial distribution; for the binomial it is the beta-binomial). Model diagnosis is in its infancy. While we encourage researchers to explore the use of these models, we also caution that considerable training is necessary for both the understanding of the theoretical underpinnings of these models and for using the software. A recent book using GLMM methodology is Stroup (2013), which developed from experience with researchers in agriculture and covering both analyses and design of experiments. He discusses in detail what we can only allude to superficially; a shortcoming is that the worked examples only use the SAS software.

5.2.1. General advice for using GLMMs

If the response variable to be measured (i.e. the phenotype of interest that may change with treatment) is a quantitative or a qualitative (i.e. yes-diseased/no-not diseased) trait and the experiment is hierarchical (e.g. bees in cages, cages from colonies, colonies from locations), repeated over years, or has some other random effects, then a generalised linear mixed model (GLMM; as provided in the statistical software R, Minitab, or SAS) can be used to analyse the results. The treatment (control, *Nosema*, black queen cell virus) is a 'fixed effect' parameter (Crawley, 2005; Bolker *et al.*, 2009). Several fixed and random effect parameters can be estimated in the same statistical model. The distinction between what is a fixed or a random effect can be difficult to make because it can be highly context-dependent, but in most experiments it should be obvious. To help clarify the distinction between the two, Crawley (2013) suggests that fixed effects influence the mean of your response variable and random effects influence the variance or correlation structure of your response variable, or is a restriction on randomisation (e.g. a block effect). A list of fixed effects would include: treatment, caste, wet *vs.* dry, light *vs.* shade, high *vs.* low, etc. i.e. treatments imposed by the researcher or inherent characteristics of the subjects (e.g. age). A list of random effects would include: cage, colony, apiary, region, genotype (if genotypes were sampled at random, not if the design was to compare two or more specific genotypes), block within a field, plot, subject measured repeatedly.

Example:

The experimenter must consider the structure of the GLMM by addressing two questions, as follows:

- Which underlying distribution?
  Gaussian, useful for data where one expects residuals to follow a 'normal' distribution
  Poisson, useful for count data (e.g. number of mites per bee)
  Binomial, useful for data on proportions based on counts (*y* out of *n*) or binary data
  Gamma, useful for data showing a constant coefficient of variation

- What link function to use?
  The link function maps the expected values of the data, conditioned on the random effects, to the linear predictor. Again, this means that the linear predictor and data reside on different scales. Canonical link functions are the most commonly used link functions associated with each 'family' of distributions (Table 4). The term "canonical" refers to the form taken of one of the parameters in the mathematical definition of each distribution.

If two or more experimental cages used in the same treatment group are drawn from the same colony of honey bees (Table 6), then a GLMM with 'source colony' as a random effect parameter should also be included, as described above. This random effect accounts for the hierarchical experimental design whereby, for the same treatment level, variation between two cages of honey bees drawn from the same colony may not be the same as the variation between two cages drawn from two separate colonies. This statistical approach can account for the problem of pseudo-replication in the experimental design.

Finally, if the factor 'cage' and 'source colony' are not significant, the experimenter may be tempted to treat individual bees from the same cage as independent samples; i.e. ignore 'cage'. However, individual bees drawn from the same cage might not truly be independent samples and therefore it would inflate the degrees of freedom to treat individual bees and individual replicates. Because there are currently no good tests to determine if a random effect is 'significant', we suggest retaining any random effects that place restrictions on randomisation - cage and source colony are two such examples - even if variance estimates are small. This point requires further attention by statisticians. The experimenter should consider using a nested experimental design in which 'individual bee' is nested within a random effect, 'cage', as presented above (see section 5.).

5.2.2. GLMM where the response variable is mortality

If survival of honey bees is the response variable of interest, then each cage should contain a minimum of 30 bees so as to provide a more robust estimate of their survival function. A typical survival analysis then needs to be undertaken on the data, e.g. the non-parametric Kaplan-Meier survival analysis for 'censored' data (so-called right-censored data in which bees are sampled from the cage during the experiment) or the semi-parametric Cox proportional hazards model (Cox model) for analysing effects of two or more 'covariates', or predictor variables such as *N. ceranae* or black queen cell virus (Collett, 2003; Zuur *et al.*, 2009; Hendriksma *et al.*, 2011). Note: these models do not only allow for random effects, if the design includes random effects then a GLMM (see section 5.2.) could be an alternative (including some function of time is a predictor variable in the GLMM).

5.2.3. Over-dispersion in GLMM

Over-dispersion is "the polite statistician's version of Murphy's law: if something can go wrong, it will" (Crawley, 2013). It is particularly relevant when working with count or proportion data where variation of a response variable does not strictly conform to the Poisson or binomial distribution, respectively. Fundamentally, over-dispersion causes poor model fitting where the difference between observed and predicted values from the tested model are larger than what would be predicted by the error structure. To identify possible over-dispersion in the data for a given model, divide the deviance (−2 times the log-likelihood ratio of the reduced model, e.g. a model with only a term for the intercept, compared to the full model; see McCullagh and Nelder,1989) by its degrees of freedom: this is called the dispersion parameter. If the deviance is reasonably close to the degrees of freedom (i.e. the dispersion or scale parameter = 1) then evidence of over-dispersion is lacking.

Table 6. Experimental design for studying the impact of *Nosema ceranae* and black queen cell virus (BQCV) on caged honey bees. †Notation represents individual cages (Treatment, Colony 1, Cage 1 = T1_1; and Control, Colony 1, Cage 1 = C1_1), each containing equal number of honey bees (e.g. 30) exposed to the same conditions (except experimental treatment differences). Two replicate cages within treatments drawn from the same colony are displayed (T1_1 and T1_2), and more could be used (T1_3, T1_4, etc.). Additional control colonies would then also be required. 'Colony' should be used as a random effect in such cases. But, it is statistically more powerful to maximise inter- as opposed to intra-colony replication; that is, favour the use of replicate cages between colonies, rather than repeated sets of cages per treatment drawn from the same colony. Thus we recommend one set of treatment and control cages per colony of source honey bees rather than repeated sets of cages per treatment and control drawn from a single colony i.e. T1_**1**, T2_**1**, T3_**1** …. T9_**1** and C1_**1**, C2_**1**, C3_**1** …. C9_**1** would be a far superior design compared to T1_**1**, T1_**2**, T1_**3** …. T1_**9** and C1_**1**, C1_**2**, C1_**3** …. C1_**9**.

Treatment	**Colony**								
	1	2	3	4	5	6	7	8	9
N. ceranae & BQCV	T1_1†, T1_2	T2_1, T2_2	T3_1, T3_2	T4_1, T4_2	T5_1, T5_2	T6_1, T6_2	T7_1, T7_2	T8_1, T8_2	T9_1, T9_2
control	C1_1, C1_2	C2_1, C2_2	C3_1, C3_2	C4_1, C4_2	C5_1, C5_2	C6_1, C6_2	C7_1, C7_2	C8_1, C8_2	C9_1, C9_2

Causes of over-dispersion can be apparent or real. Apparent over-dispersion is due to model misspecification, i.e. missing covariates or interactions, outliers in the response variable, non-linear effects of covariates entered as linear effects, the wrong link function, etc. Real over-dispersion occurs when model misspecifications can be ruled out, and variation in the data is real due to too many zeros, clustering of observations, or correlation between observations (Zuur *et al.*, 2009). Solutions to over-dispersion can include: i) adding covariates or interactions, ii) including individual-level random effects, e.g. using bee as a random effect, where multiple bees are observed per cage, iii) using alternative distributions: if there is no random effect included in the model consider quasi-binomial and quasi-Poisson; if there are, consider replacing Poisson with negative-binomial, and iv) using a zero-inflated GLMM (a model that allows for numerous zeros in your dataset, the frequency of the number zero is inflated) if appropriate. Over-dispersion cannot occur for normally distributed response variables because the variance is estimated independently from the mean. However, residuals often have "heavy tails", i.e. more outlying observations than expected for a normal distribution, which nevertheless can be addressed by some software packages.

5.3. Accounting for multiple comparisons

Thus far, we have assumed that we are investigating two categories of an explanatory variable or experimental treatment (i.e. comparing a treatment group with a control group). However, the objective may instead be to compare multiple levels of an explanatory variable (e.g. different concentrations of a pesticide) or multiple independent kinds of the same sort of explanatory variable (e.g. competing manufacturers of protein substitutes). In addition, one may be interested in testing multiple explanatory variables at the same time (e.g. effects of three different humidity levels and honey bee age on susceptibility to the tracheal mite *Acarapis woodi*). More complex statistical models warrant increased sample sizes for all treatments. Consider the case where one has one control and one treatment group; there is a single comparison possible. Yet if one has one control and 9 treatments groups, there are 9 + 8 +…+ 1 = 55 possible comparisons. If one rigorously follows the cut-off of $P = 0.05$, one could obtain 0.05 * 55 = 2.8 significant results by chance or in other words the probability of at least one significant by chance alone is $1 - 0.95^{55} = 0.9405$, so one is likely to incorrectly declare significance at least once (in general, 5% of statistical results will have $p \geq 0.05$ if there are no true differences among treatments, this is what setting $\alpha = 0.05$ represents). *Post hoc* tests or *a posteriori* testing, such as Bonferroni corrections, attempt to account for this excessive testing, but in so doing can become very conservative, and potentially significant results may be overlooked (i.e. correctly control for Type I error, but have inflated Type II errors; Rothman, 1990; Nakagawa, 2004). Less conservative corrections, such as the False Discovery Rate, are now typically favoured as they represent a balance between controlling for Type I and Type II errors (Benjamini and Hochberg, 1995). Other ways to avoid or minimise this problem include increasing sample size and simplifying experimental design by reducing the number of treatments and variables.

5.4. Principal components to reduce the number of explanatory variables

With an increasing number of explanatory variables (related or not-related, similar or dissimilar units) in one experiment, multivariate statistics may be of interest. Multivariate statistics are widely used in ecology (Leps and Smilauer, 2003), but less often in bee research. Multivariate statistics can be used to reduce the number of response variables without losing information in the response variables (van Dooremalen and Ellers, 2010), or to reduce the number of explanatory variables (especially valuable if they are correlated). A Principle Component Analysis (PCA) can be used to examine, for example, morphometric or physiological variables (such as protein content of different bee body parts or several volatile compounds in the head space of bee brood cells). The PCA is usually used to obtain only the first principal component that forms one new PC variable (the axis explaining most variation in your variables). The correlations between the original variables and the new PC variable will show the relative variation explained by the original variables compared to each other and their reciprocal correlation. The new PC variable can then be used to investigate effects of different treatments (and/or covariates) using statistics as explained above in section 5. For an example in springtails see van Dooremalen *et al.* (2011), or in host-parasite interactions see Nash *et al.* (2008). Note that the new PC variables are uncorrelated with each other, which improves their statistical properties. Unfortunately, it is also easy to lose track of what they represent or how to interpret them. However, by reducing dimensionality and dealing with uncorrelated variables one can transform a data set with a great many explanatory and response variables into one with only a few of each, and ones which capture most of the variability (i.e. the underlying processes) in the data set. Related procedures are factor analysis, partial least squares, non-metric multidimensional scaling (NMDS), and PC regression.

5.5. Robust statistics

Robust statistics were developed because empirical data that considered samples from normal distributions often displayed clearly non-normal characteristics, which invalidates the analyses if one assumes normality. They are usually introduced early on in discussions of measures of central tendency. For example, medians are far more resistant to the influence of outliers (observations that are deemed to deviate for reasons that may include measurement error, mistakes in data entry, etc.) than are means, so the former are considered more robust. Even a small number of outliers (as few as one) may adversely affect a mean, whereas a median can be resistant when up to 50% of

observations are outliers. On the other hand, screening for outliers for removal may be subjective and difficult for highly structured data, where a response variable may be functionally related to many independent variables. If "outliers" are removed, resulting variance estimates are often too small, resulting in overly liberal testing (i.e. *p* values are too small).

What are the alternatives when one cannot assume that data are generated by typical parametric models (e.g. normal, Poisson, binomial distributions)? This may be a result of contamination (e.g. most of the data comes from a normal distribution with mean μ and variance σ_1^2 but a small percentage comes from a normal distribution with mean μ and variance σ_2^2, where $\sigma_2^2 >> \sigma_1^2$), a symmetric distribution with heavy tails, such as a *t* distribution with few degrees of freedom, or some highly skewed distribution (especially common when there is a hard limit, such as no negative values, typical of count data and also the results of analytic procedures estimation; e.g. titres). Robust statistics are generally applicable when a sampling distribution from which data are drawn is symmetric. "Non-parametric" statistics are typically based on ordering observations by their magnitude, and are thus more general, but have lower power than either typical parametric models or robust statistical models. However, robust statistics never "caught on" to any great degree in the biological sciences; they should be used far more often (perhaps in most cases where the normal distribution is assumed).

Most statistics packages have some procedures based on robust statistics; R has particularly good representation (e.g. the MASS package). All typical statistical models (e.g. regression, ANOVA, multivariate procedures) have counterparts using robust statistics. Estimating these models used to be considered difficult (involving iterative solutions, maximisation, etc.), but these models are now quickly estimated. The generalised linear class of models (GLM) has some overlap with robust statistics, because one can base models on, e.g. heavy-tailed distributions in some software, but the approach is different. In general, robust statistics try to diminish effects of "influential" observations (i.e. outliers). GLMs, once a sampling distribution is specified (theoretical sampling distributions include highly skewed or heavy-tailed ones, though what is actually available depends on the software package) consider all observations to be legitimate samples from that distribution. We recommend analysing data in several different ways if possible. If they all agree, then one might choose the analysis which best matches the theory (the sampling distribution best reflecting our knowledge of the underlying process) of how the data arose. When methods disagree, one must then determine why they differ and make an educated choice on which to use. For example, if assuming a normal distribution results in different confidence limits around means than those obtained using robust statistics, it is likely that there is serious contamination from outliers that is ignored by assuming a normal distribution. A recent reference on robust statistics is Maronna *et al.* (2006), while the classic one is Huber (1981).

5.6. Resampling techniques

Statistical methodology has benefited enormously from fast and ubiquitous computing power, with the two largest beneficiaries being methods that rely on numerical techniques, such as estimating parameters in GLMMs, and methods that rely on sampling, either from known distributions (such as most Bayesian methods, often called "Monte-Carlo" methods) or from the data (resampling or "bootstrapping"). Resampling techniques are essentially non-parametric, the only assumption is that the data are representative of the population you want to make inferences from. The data set must also be large enough to resample from, following the rules stated earlier for sample sizes for parametric models, (*i.e.* at least 10 observations per "parameter", so a difference between two medians would require at least 20 observations).

As a simple example, if we want to estimate a 95% confidence interval around a median, based on 30 observations, we can draw 100 random resampled data sets (*with replacement*) from the original data set, each of size 30, calculate the median for each of these resampled data sets, and rank those values. The 95% confidence interval is then the interval from the 5th to the 95th calculated median. Even though the original data set and the resampled data sets are the same size (n = 30), they are likely not identical because we are sampling with replacement, meaning that there will be duplicates (or even triplicates) of some of the original values in each resampled data set, and others will be missing.

Resampling can be used for statistical testing in a similar way. For example, if we want to know if the difference in medians between two data sets (each of size 30) is significant at $\alpha = 0.05$, we could use the following approach. Take a random sample (with replacement) of size 30 from data set 1 and calculate its median, do the same for data set 2. Subtract the sample 2 median from the sample 1 median and store the value. Repeat this until you have 1,000 differences. Rank the differences. If the interval between the 50th and 950th difference does not contain zero, the difference in medians is statistically significant.

This general method can be applied to many common statistical problems, and can be shown to have good power (often better than a parametric technique if an underlying assumption of the parametric technique is even slightly violated). It can be used for both quantitative and qualitative (e.g. categorical) data, for example for testing the robustness of phylogenetic trees derived from nucleotide or amino acid sequence alignments, and is also useful as an independent method to check the results of statistical testing using other techniques. It does require either some programming skills or use of a statistical package that implements resampling techniques.

If one writes a program, three parts are required. The first is used for creating a sample by extracting objects from the original data set, based on their position in the data file, using a random number generator. As a simple example, if there are five values, a random

number generator (sampling with replacement) might select the values in positions (4, 3, 3, 2, 4). Note that some positions are repeated, others are missing. That is fine because this process will be repeated 10,000 times, and, on average, all data values will have equal representation. The second part is used for calculating the parameters of interest, for example, the median, and is also run 10,000 times. More complicated statistics take longer, and that will affect how long the program takes to complete. The third part stores the results of the second part, and may be a vector of length 10,000 (or a matrix with 10,000 rows, if several statistics are calculated from each resampled data set). Finally, summary statistics or confidence intervals are created, based on the third part. For example, if medians were calculated, one could calculate 90%, 95%, and 99% confidence intervals after ranking the medians and selected appropriate endpoints of the intervals. In general, 10,000 resampled data sets are considered to be a minimum to use for published results, though 500 are usually adequate for preliminary work (and that number is also useful for estimating how long it will take 10,000 to run).

All the major statistical software packages have resampling routines, and some rely almost exclusively on it (e.g. PASS, in the NCSS statistical software). We recommend the **boot** package in the R software, which is very flexible and allows one to estimate many of the quantities of interest for biologists (e.g. differences of means or medians, regression parameters). The classic book is Efron and Tibshirani (1993); Bradley Efron is the developer of the technique. A recent, less technical book is by Good (2013). A related technique is "jack-knifing", where one draws all possible subsamples *without replacement*, typically of size $n - 1$, where n is the original sample size.

6. Presentation and reporting of data

Presentation depends on the data collected and what the authors wants to emphasise. For example, to present the mean when one has done a non-parametric test is not meaningful, though a median is (consider using boxplots). The mean is a valid descriptive representation of the location parameter if the distribution is symmetric. The best way to summarise descriptively and represent graphically a given data set depends on both the empirical distribution of the data and the purpose of the statistics and graphs. There are excellent references on this topic such as those by Cleveland (1993) and Tufte (2001), whereas the classic book by Tukey (1977) has a decidedly statistical slant.

Standard error or Standard deviation - the former indicates uncertainty around a calculated mean; the latter is a measurement of the variability of the observed data around the mean. We believe that the standard deviation is the better metric to convey characteristics of the data because the standard error, which is also a function of sample size, can be made arbitrarily small by including more observations.

Presentation of data might be overlaid with statistics one has applied, such as regression lines or mean separation letters. If data were transformed for the analysis, data on the original scale should be presented, but any means fit from a statistical model back-transformed to the original scale (even though this will create curves in a "straight" line model, like a linear regression). Back-transformed confidence intervals on means should replace standard error bars.

7. Which software to use for statistical analyses?

Statistical programmes, such as the freeware R and its packages, as well as other packages such as Minitab, SPSS, and SAS, can handle the analyses described in this paper. There are several sites comparing the different packages: http://en.wikipedia.org/wiki/Comparison_of_statistical_packages, http://en.wikipedia.org/wiki/List_of_statistical_packages
Although spreadsheet software has improved and many statistical tests are available, they often lack good diagnostics on the model fit and checks for the appropriateness of the statistical test.

8. Where to find help with statistics

A statistician, preferably with an understanding of biology, remains the best solution to get one's statistics right. Given the importance of sample size for analyses, it is important to contact one as early as the design stage of an experiment or survey. If your university or institute does not offer the service of a statistician, there are freelance professionals as well as numerous forums on the internet where questions can be posted. Examples of such sites can be found on the support sites for R and commercial programmes. Most maths departments offer some kind of introduction to basic statistics.

9. Conclusion

Guidelines and the selection of the different methods presented are, at least partly, based on experience and we cannot cover all statistical methods available, for example we have not discussed resampling methods like jackknife in detail (for further reading see Good, 2006). More details on designing specific experiments and performing statistical analyses on the ensuing data can be found in respective chapters of the COLOSS *BEEBOOK* (e.g. in the toxicology chapter, Medrzycki *et al.*, 2013).

Experimenters need to use statistical tests to take (or to help take) a decision. A statistical analysis can be conducted only if its assumptions are met, which largely depends on how the experiment was designed, defined during the drafting of the study protocol. Without some effort at the *a priori* conception stage and input from those knowledgeable in statistics and/or experimental design, the resulting analyses are frequently poor and the conclusions can be biased or flat-out wrong. Why spend a year or more collecting data and then realise that, due to poor design, it is not suitable for its original purpose: to test the hypotheses of interest. The most important point to understand about statistics is that one should think about the statistical analysis before collecting data or conducting the experiment.

10. Acknowledgements

We are more than grateful to Ingemar Fries of the Swedish University of Agricultural Sciences for comments on and contributions to earlier versions of the chapter. We thank K L Crous and A A Yusuf for comments on an earlier version of the manuscript and we also thank Werner Luginbühl for a very thoughtful and thorough review of the original submission. The University of Pretoria, the National Research Foundation of South Africa and the Department of Science and Technology of South Africa (CWWP) granted financial support. The COLOSS (Prevention of honey bee COlony LOSSes) network aims to explain and prevent massive honey bee colony losses. It was funded through the COST Action FA0803. COST (European Cooperation in Science and Technology) is a unique means for European researchers to jointly develop their own ideas and new initiatives across all scientific disciplines through trans-European networking of nationally funded research activities. Based on a pan-European intergovernmental framework for cooperation in science and technology, COST has contributed since its creation more than 40 years ago to closing the gap between science, policy makers and society throughout Europe and beyond. COST is supported by the EU Seventh Framework Programme for research, technological development and demonstration activities (Official Journal L 412, 30 December 2006). The European Science Foundation as implementing agent of COST provides the COST Office through an EC Grant Agreement. The Council of the European Union provides the COST Secretariat. The COLOSS network is now supported by the Ricola Foundation - Nature & Culture.

11. References

AMDAM, G V; OMHOLT, S W (2002) The regulatory anatomy of honey bee lifespan. *Journal of Theoretical Biology* 216: 209-228.

BAILEY, L; BALL, B V; PERRY, J N (1981) The prevalence of viruses of honey bees in Britain. *Annals of Applied Biology* 97: 109-118. http://dx.doi.org/10.1111/J.1744-7348.1981.Tb02999.X

BAILEY, L, BALL; B V (1991) *Honey bee pathology*. Academic Press; London, UK.

BENJAMINI, Y; HOCHBERG, Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289-300. http://dx.doi.org/10.2307/2346101

BICKEL, P J; DOKSUM, K A (1981) An analysis of transformations revisited. *Journal of the American Statistics Association* 76:

BOLKER, B M; BROOKS, M E; CLARK, C J; GEANGE, S W; POULSEN, J R; STEVENS, M H H; WHITE, J S S (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24: 127-135. http://dx.doi.org/10.1016/J.Tree.2008.10.008

BOX, G E P; COX, D R (1964) An analysis of transformations. *Journal of the Royal Statistical Society B* 26: 211-252.

BRUNETTO, M R; COLOMBATTO, P; BONINO, F (2009) Bio-mathematical models of viral dynamics to tailor antiviral therapy in chronic viral hepatitis. *World Journal of Gastroenterology* 15: 531-537. http://dx.doi.org/10.3748/Wjg.15.531

CARRECK, N L; ANDREE, M; BRENT, C S; COX-FOSTER, D; DADE, H A; ELLIS, J D; HATJINA, F; VANENGELSDORP, D (2013) Standard methods for *Apis mellifera* anatomy and dissection. In *V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS* BEEBOOK*, Volume I: standard methods for* Apis mellifera *research*. *Journal of Apicultural Research* 52(4): http://dx.doi.org/10.3896/IBRA.1.52.4.03

CASELLA, G (2008) *Statistical design*. Springer; Berlin, Germany.

CLEVELAND, W S (1993) *Visualizing data*. Hobart Press; Summit, USA.

COHEN, J (1988) *Statistical Power Analysis for the behavioral sciences (2nd Ed.).* Lawrence Erlbaum Associates.

COLLETT, D (2003) *Modelling survival data in medical research (2nd Ed.)*. Chapman & Hall/CRC.

COLTON, T (1974) *Statistics in medicine*. Little, Brown and Co.; Boston, USA.

CRAWLEY, M J (2005) *Statistics: an introduction using R*. Wiley & Sons; Chichester, UK.

CRAWLEY, M J (2013) *The R Book (2nd Ed.).* Wiley & Sons; Chichester, UK.

CREWE, R M (1982) Compositional variability: The key to the social signals produced by honey bee mandibular glands. *The Biology of Social Insects*: 318-322.

CREWE, R M; MORITZ, R F A; LATTORFF, H M (2004) Trapping pheromonal components with silicone rubber tubes: fatty acid secretions in honey bees (*Apis mellifera*). *Chemoecology* 14: 77-79.

DARCHEN, R (1956) La construction sociale chez *Apis mellifica*. *Insectes Sociaux* 3: 293-301. http://dx.doi.org/10.1007/bf02224312

DARCHEN, R (1957) La reine d'*Apis mellifica* les ouvrières pondeuses et les constructions cirières. *Insectes Sociaux* 4: 321-325. http://dx.doi.org/10.1007/bf02224152

DOULL, K M; CELLIER, K M (1961) A survey of incidence of Nosema disease (*Nosema apis* Zander) of the honey bee in South Australia. *Journal of Insect Pathology* 3: 280.

EFRON, B; TIBSHIRANI, R J (1993) *An introduction to the bootstrap.* Chapman & Hall; London, UK.

FAUL, F; ERDFELDER, E; LANG, A-G; BUCHNER, A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175-191.

FIELD, A; MILES, J; FIELD, Z (2012) *Discovering statistics using R.* SAGE Publications Ltd.; London, UK.

FREE, J B (1960) The distribution of bees in a honey-bee (*Apis mellifera*. L) colony. *Proceedings of the Royal Entomological Society of London* A35: 141-141.

FRIES, I; EKBOHM, G; VILLUMSTAD, E (1984) *Nosema apis*, sampling techniques and honey yield. *Journal of Apicultural Research* 23: 102-105.

FRIES, I; LINDSTROM, A; KORPELA, S (2006) Vertical transmission of American foulbrood (*Paenibacillus larvae*) in honey bees (*Apis mellifera*). *Veterinary Microbiology* 114: 269-274. http://dx.doi.org/10.1016/J.Vetmic.2005.11.068

FROST, E H; SHUTLER, D; HILLIER, N K (2012) The proboscis extension reflex to evaluate learning and memory in honey bees (*Apis mellifera*): some caveats. *Naturwissenschaften* 99: 677-686. http://dx.doi.org/10.1007/S00114-012-0955-8

GAUTHIER, L; TENTCHEVA, D; TOURNAIRE, M; DAINAT, B; COUSSERANS, F; COLIN, M E; BERGOIN, M (2007) Viral load estimation in asymptomatic honey bee colonies using the quantitative RT-PCR technique. *Apidologie* 38: 426-U7. http://dx.doi.org/10.1051/Apido:2007026

GOOD, P I (2006) *Resampling methods (3rd Ed.).* Springer; Berlin, Germany.

GOOD, P I (2013) *Introduction to statistics through resampling methods and R (2nd Ed.).* John Wiley & Sons, Inc.; Hoboken, USA.

GREEN, S B (1991) How many subjects does it take to do a regression -analysis. *Multivariate Behavioral Research* 26: 499-510. ttp://dx.doi.org/10.1207/S15327906mbr2603_7

HENDRIKSMA, H P; HARTEL, S; STEFFAN-DEWENTER, I (2011) Honey bee risk assessment: new approaches for *in vitro* larvae rearing and data analyses. *Methods in Ecology and Evolution* 2: 509-517. http://dx.doi.org/10.1111/J.2041-210x.2011.00099.X

HEPBURN, H R (1986) *Honey bees and wax: an experimental natural history*. Springer Verlag; Berlin, Germany.

HESS, G (1942) Über den Einfluß der Weisellosigkeit und des Fruchtbarkeitsvitamins E auf die Ovarien der Bienenarbeiterin Ein Beitrag zur Frage der Regulationen im Bienenstaat. *Beihefte zur Schweizerischen Bienen-Zeitung* 2: 33-111.

HIGES, M; MARTIN-HERNANDEZ, R; BOTIAS, C; BAILON, E G; GONZALEZ-PORTO, A V; BARRIOS, L; DEL NOZAL, M J; BERNAL, J L; JIMENEZ, J J; PALENCIA, P G; MEANA, A (2008) How natural infection by *Nosema ceranae* causes honey bee colony collapse. *Environmental Microbiology* 10: 2659-2669. http://dx.doi.org/10.1111/J.1462-2920.2008.01687.X

HUANG, Z-Y; ROBINSON, G E (1996) Regulation of honey bee division of labor by colony age demography. *Behavioral Ecology and Sociobiology* 39: 147-158.

HUBER, P J (1981) *Robust statistics*. Wiley; New York, USA.

HURLBERT, S H (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211. http://dx.doi.org/10.2307/1942661

JASSIM, O; HUANG, Z Y; ROBINSON, G E (2000) Juvenile hormone profiles of worker honey bees, *Apis mellifera*, during normal and accelerated behavioural development. *Journal of Insect Physiology* 46: 243-249.

JOHNSON, B R (2005) Limited flexibility in the temporal caste system of the honey bee. *Behavioral Ecology and Sociobiology* 58: 219-226. http://dx.doi.org/10.1007/S00265-005-0949-Z

KAEKER, R; JONES, A (2003) On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent. *Metrologia* 40: 235-248.

KELLNER, N (1981) Studie van de levenscyclus van *Nosema apis* Zander in de honingbij (*Apis mellifera*). PhD thesis, Rijksuniversiteit Gent, Belgium.

KÖHLER, A; NICOLSON, S W; PIRK, C W W (2013) A new design for honey bee hoarding cages for laboratory experiments. *Journal of Apicultural Research* 52(2): 12-14. http://dx.doi.org/10.3896/IBRA.1.52.2.03

LEPS, J; SMILAUER, P (2003) *Multivariate analysis of ecological data using CANOCO*. Cambridge University Press; Cambridge, UK.

LINDAUER, M (1952) Ein Beitrag zur Frage der Arbeitsteilung im Bienenstaat. *Zeitschrift für vergleichende Physiologie* 34: 299-345.

LINDAUER, M (1953) Division of labour in the honey bee colony. *Bee World* 34: 63-90.

MADDEN, L V; HUGHES, G (1999) An effective sample size for predicting plant disease incidence in a spatial hierarchy. *Phytopathology* 89: 770-781. http://dx.doi.org/10.1094/Phyto.1999.89.9.770

MANLY, B F J (1997) *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman and Hall; London, UK.

MARONNA, R A; MARTIN, R D; YOHAI, V J (2006) *Robust statistics: theory and Methods*. Wiley; New York, USA.

MCCULLAGH, P; NELDER, J (1989) *Generalized Linear Models.* Chapman & Hall/CRC Press.

MEDRZYCKI, P; GIFFARD, H; AUPINEL, P; BELZUNCES, L P; CHAUZAT, M-P; CLAßEN, C; COLIN, M E; DUPONT, T; GIROLAMI, V; JOHNSON, R; LECONTE, Y; LÜCKMANN, J; MARZARO, M; PISTORIUS, J; PORRINI, C; SCHUR, A; SGOLASTRA, F; SIMON DELSO, N; VAN DER STEEN, J J F; WALLNER, K; ALAUX, C; BIRON, D G; BLOT, N; BOGO, G; BRUNET, J-L; DELBAC, F; DIOGON, M; EL ALAOUI, H; PROVOST, B; TOSI, S; VIDAU, C (2013) Standard methods for toxicology research in *Apis mellifera*. In *V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS* BEEBOOK*, Volume I: standard methods for* Apis mellifera *research*. *Journal of Apicultural Research* 52(4) http://dx.doi.org/10.3896/IBRA.1.52.4.14

MILES, J; SHEVLIN, M (2001) *Applying regression and correlation: a guide for students and researchers*. SAGE Publications Ltd.; London, UK.

MORITZ, R; LATTORFF, H; CREWE, R (2004) Honey bee workers (*Apis mellifera capensis*) compete for producing queen-like pheromone signals. *Proceedings of the Royal Society B: Biological Sciences* 271: S98-S100.

NAKAGAWA, S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology* 15: 1044 -1045. http://dx.doi.org/10.1093/Beheco/Arh107

NASH, D R; ALS, T D; MAILE, R; JONES, G R; BOOMSMA, J J (2008) A mosaic of chemical coevolution in a large blue butterfly. *Science* 319: 88-90. http://dx.doi.org/10.1126/Science.1149180

PIRK, C W W; BOODHOO, C; HUMAN, H; NICOLSON, S W (2010) The importance of protein type and protein to carbohydrate ratio for survival and ovarian activation of caged honey bees (*Apis mellifera scutellata*). *Apidologie* 41: 62-72. http://dx.doi.org/10.1051/Apido/2009055

PIRK, C W W; SOLE, C L; CREWE, R M (2011) Pheromones. In *H R HEPBURN; S E RADLOFF (Eds). Honey bees of Asia.* Springer; Berlin Heidelberg, Germany. pp 207-214.

REICZIGEL, J (2003) Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine* 22: 611-621. http://dx.doi.org/10.1002/Sim.1320

RIBBANDS, C R (1952) Division of labour in the honey bee community. *Proceedings of the Royal Society B: Biological Sciences* 140: 32-43.

ROTHMAN, K J (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1: 43-46.

RUNCKEL, C; FLENNIKEN, M L; ENGEL, J C; RUBY, J G; GANEM, D; ANDINO, R; DERISI, J L (2011) Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, *Nosema*, and *Crithidia. PLoS ONE* 6: http://dx.doi.org/10.1371/journal.pone.0020656

SCHÄFER, M O; DIETEMANN, V; PIRK, C W W; NEUMANN, P; CREWE, R M; HEPBURN, H R; TAUTZ, J; CRAILSHEIM, K (2006) Individual versus social pathway to honey bee worker reproduction (*Apis mellifera*): pollen or jelly as protein source for oogenesis? *Journal of Comparative Physiology* A192: 761-768.

SCHEINER, R; BARNERT, M; ERBER, J (2003) Variation in water and sucrose responsiveness during the foraging season affects proboscis extension learning in honey bees. *Apidologie* 34: 67-72. http://dx.doi.org/10.1051/Apido:2002050

SEELEY, T D (1985) *Honey bee ecology: a study of adaptation in social life*. Princeton University Press; Princeton, USA.

SOKAL, R R; ROHLF, F J (1995) *Biometry: the principles and practice of statistics in biological research (1st Ed.).* W H Freeman and Co.; New York, USA.

SOKAL, R R; ROHLF, F J (2012) *Biometry: the principles and practice of statistics in biological research (4th Ed.).* W H Freeman and Co.; New York, USA.

STEPHENS, M A (1974) EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69: 730-737. http://dx.doi.org/10.1080/01621459.1974.10480196

STROUP, W W (2013) *Generalized Linear Mixed Models: modern concepts, methods and applications.* CRC Press; Boca Raton, Florida, USA.

TENTCHEVA, D; GAUTHIER, L; ZAPPULLA, N; DAINAT, B; COUSSERANS, F; COLIN, M E; BERGOIN, M (2004) Prevalence and seasonal variations of six bee viruses in *Apis mellifera* L. and *Varroa destructor* mite populations in France. *Applied and environmental microbiology* 70: 7185-7191. http://dx.doi.org/10.1128/Aem.70.12.7185-7191.2004

TUFTE, E R (2001) *The visual display of quantitative information*. Graphics Press; Cheshire, USA.

TUKEY, J W (1977) *Exploratory data analysis*. Addison-Wesley Publishing Company; Reading, MA, USA.

VAN DER STEEN, J J M; CORNELISSEN, B; DONDERS, J; BLACQUIÈRE, T; VAN DOOREMALEN, C (2012) How honey bees of successive age classes are distributed over a one storey, ten frame hive. *Journal of Apicultural Research* 51(2): 174-178. http://dx.doi.org/10.3896/IBRA.1.51.2.05

VAN DOOREMALEN, C; ELLERS, J (2010) A moderate change in temperature induces changes in fatty acid composition of storage and membrane lipids in a soil arthropod. *Journal of Insect Physiology* 56: 178-184. http://dx.doi.org/10.1016/J.Jinsphys.2009.10.002

VAN DOOREMALEN, C; SURING, W; ELLERS, J (2011) Fatty acid composition and extreme temperature tolerance following exposure to fluctuating temperatures in a soil arthropod. *Journal of Insect Physiology* 57: 1267-1273. http://dx.doi.org/10.1016/j.jinsphys.2011.05.017

VIM (2008) International vocabulary of metrology - Basic and general concepts and associated terms (VIM). http://www.iso.org/sites/JCGM/VIM/JCGM_200e.html

WASSERMAN, L (2006) *All of nonparametric statistics*. Springer; Berlin, Germany.

WILLIAMS, G R; ALAUX, C; COSTA, C; CSÁKI, T; DOUBLET, V; EISENHARDT, D; FRIES, I; KUHN, R; MCMAHON, D P; MEDRZYCKI, P; MURRAY, T E; NATSOPOULOU, M E; NEUMANN, P; OLIVER, R; PAXTON, R J; PERNAL, S F; SHUTLER, D; TANNER, G; VAN DER STEEN, J J M; BRODSCHNEIDER, R (2013) Standard methods for maintaining adult *Apis mellifera* in cages under *in vitro* laboratory conditions. In *V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS* BEEBOOK*, Volume I: standard methods for* Apis mellifera *research*. *Journal of Apicultural Research* 52(1): http://dx.doi.org/10.3896/IBRA.1.52.1.04

WOLFE, R; CARLIN, J B (1999) Sample-size calculation for a log-transformed outcome measure. *Controlled Clinical Trials* 20: 547-554. http://dx.doi.org/10.1016/S0197-2456(99)00032-X

ZUUR, A L E N; WALKER, N; SAVELIEV, A A; SMITH, G M (2009) *Mixed effects models and extensions in ecology with R. Vol 1*. Springer; Berlin, Germany.