CAMBRIDGE

Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information

# 1 Introduction

*To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
Sir Ronald A. Fisher, FRS

*Many experiments fail because the data collectors have not been properly trained and many statisticians have their own horror stories to illustrate this.*
John A. Nelder, FRS

*Modeling is sometimes regarded as primarily a task for subject matter specialists, but in most fields requisite knowledge and understanding of statistics remains thinly spread.*
Arthur P. Dempster

This chapter begins by defining reproducibility and discussing non-statistical – mainly psychological – sources of experimental bias. The next section assesses the quality of the published literature and discusses statistical sources of bias. The discussion will be familiar if you have been following the 'reproducibility crisis' over the past several years. The above topics are included to stimulate reflection about your own research practices and to bring together ideas that have been discussed in separate disciplines. The chapter ends with a refresher on statistical inference and a discussion on statistics software.

## 1.1 What is reproducibility?

An experiment is reproducible when subsequent experiments, by the same or different scientists, confirm the results. The terms *repeatability* and *replicability* are sometimes used interchangeably or with related meanings, but we will use reproducibility as an all-encompassing term. Reproducibility can occur at several levels.[1]

**Analytical:** Analytical or computational reproducibility refers to obtaining the same results using the original data and a description of the analysis. This is a minimum standard but is impossible to achieve when the data are unavailable. Even if the data are provided in the supplementary material or in public databases, reproducing the results may be hard if the description of the analysis is incomplete [173]. A minimum

[1] Adapted from a report on reproducibiliy by the American Society for Cell Biology: `http://www.ascb.org/reproducibility`



© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



requirement for analytical reproducibility is to provide the data underlying the results and the scripts that produced them. This is simple when using R because the code can be integrated into documents such as reports and publications. For example, large portions of this book have embedded R code, which is evaluated, and then the outputs are inserted into the text document. The `knitr` and `rmarkdown` R packages make this process straightforward [402].

**Direct:** Direct reproducibility refers to obtaining the same results using the same experimental conditions, materials, and methods as the original experiment. The aim is to make the second experiment as similar as possible to the original, which requires an adequate description of how the original experiment was conducted. Direct replication is the focus of this book, but it may not be immediately clear how better experimental designs can improve direct reproducibility. The brief answer is that a well-designed experiment (1) can isolate the effects of interest from other factors that may influence the outcome, (2) replicates the right aspect of the experiment, and (3) can generalise the results to other times, places, conditions, and samples.

**Systematic:** Systematic reproducibility refers to obtaining the same results, but under different conditions; for example, using another cell line or mouse strain, or inhibiting a gene pharmacologically instead of genetically. Reasons for a lack of systematic reproducibility are harder to determine because the cell lines might be dissimilar, and what works in one will not work in another. This should not be taken as evidence of poor research practices, and one function of subsequent studies is to find the conditions under which an initial finding holds. Experimental design can help here too, as initial studies can be designed to address the question of generalisability early on.

**Conceptual:** Conceptual reproducibility refers to obtaining the same general results under diverse conditions, where the aim is to demonstrate the validity of a concept or a finding using another paradigm. The general concept or hypothesis might be 'stress inhibits memory formation', which could be tested in one experiment where people memorise word pairs with loud music playing and in another experiment where rats learn the location of food pellets after a corticosterone injection (a stress hormone). There are many valid reasons why some experiments support the hypothesis and others not – maybe corticosterone, while part of the stress response, is irrelevant for learning. Discrepancies between the results of such experiments do not necessarily indicate poor reproducibility.

A reproducible result was defined above as one that is confirmed by subsequent experiments, but what does *confirmed* mean? One idea is that if the original experiment has a $p$-value below 0.05, then the experiment is confirmed if the subsequent experiment also has a significant $p$-value. Although this criterion seems plausible, it has several problems. First, a study with a $p$-value of 0.03 would be considered irreproducible if the subsequent experiment had a $p$-value of 0.08. But for all practical purposes the studies may have the same effect sizes and their two confidence intervals (CIs) may overlap substantially. This relationship is shown in Figure 1.1 between the original experiment and the second experiment, New 1. A second problem is that this approach ignores the sample size and power of the experiments. Suppose that a power analysis was conducted based on the results of

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



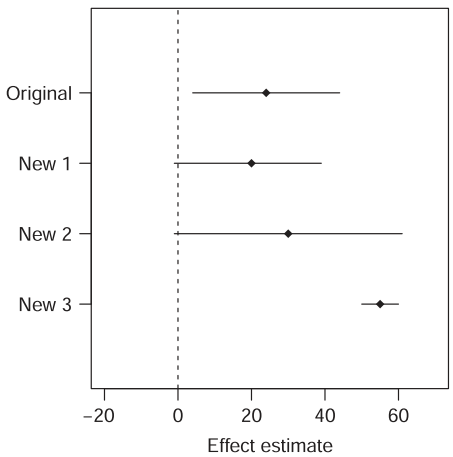


**Fig. 1.1** Effect sizes and 95% CI for an original experiment and three follow-up experiments. Using significance as a criterion for reproducibility, only New 3 would be considered to reproduce the original finding, despite the different effect size.

the original experiment and the follow-up experiment uses a slightly smaller sample size and therefore the confidence intervals will be slightly wider, assuming everything else is constant (New 2 in Figure 1.1). Even though the effect size for New 2 is larger than the original, New 2 would not have reproduced the original findings by this criterion. A third problem is that a follow-up study may have a different effect size than the original but would be considered to have successfully reproduced the original if the *p*-value is significant. This situation is shown for experiment New 3, where the 95% CIs do not overlap with those of the original experiment. There is no agreed criteria for when one experiment can be said to reproduce another, but within a field, scientists 'know it when they see it'.

## 1.2 The psychology of scientific discovery

It is uncommon for a book on experimental design to discuss psychological aspects of research, but scientific investigations are not conducted in a vacuum; they take place in the context of previous research, are conducted by people who prefer certain outcomes over others, and are constrained by standards and conventions used by research groups and the wider scientific community. Expectations and desires of the researcher and external pressures to publish and to demonstrate creativity and innovation influence how data are analysed, interpreted, and reported. This needs to be acknowledged and discussed because improving reproducibility and 'making more published research true' [172] – one of the aims of this book – cannot be achieved by only improving scientists' maths skills.

Some of the topics discussed below fall within the 'heuristics and biases' field of psychological research. Cognitive biases or cognitive illusions are deviations from true or optimal answers or responses when making estimations, inferences, decisions, conclusions, or judgements [312]. They are cognitive in that they result from perceptual or cognitive

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



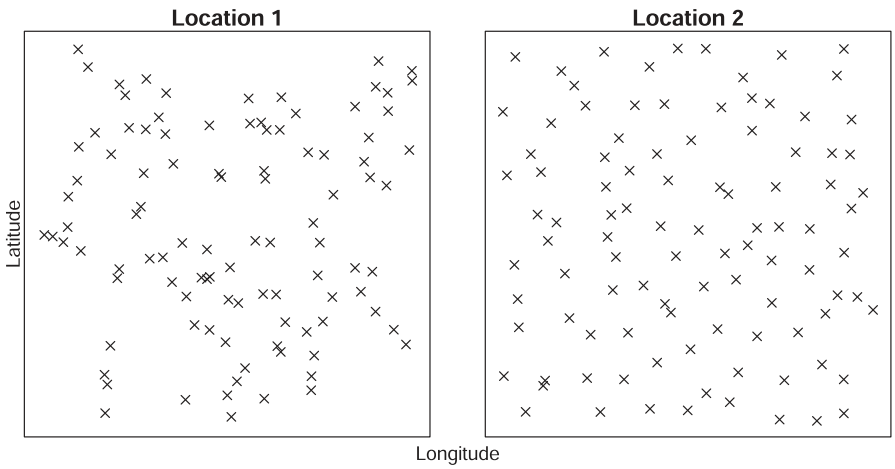


Fig. 1.2 Positions of bombs dropped for two geographic locations. Which location represents the uniform random bombing strategy?

processes instead of, for example, an uncalibrated measuring device. They are also systematic, meaning that the deviations tend to be in a certain direction. They are also hard to avoid. Cognitive biases can influence the design, analysis, interpretation, and reporting of biological experiments and are therefore relevant for scientific investigations. Several such biases are described below, with an emphasis on how they apply to experimentation and statistical inference. Methods for avoiding them are also suggested.

### 1.2.1 Seeing patterns in randomness

People often see patterns where none exist, including clusters, associations between variables, and sequences of similar values. A fictitious example is given in Figure 1.2. The positions (latitude and longitude) of 100 bombs dropped during a World War II bombing campaign are shown for two different geographic locations. The General wants to know if the enemy is dropping bombs at random, or if they are targeting certain positions more heavily, and he asks you to investigate. Intelligence from the front line indicates that *if* the enemy is using a random strategy, they will randomly sample a pair of latitude and longitude coordinates with equal probability anywhere within the bombing region – known as a uniform random bombing strategy.[2] Was a uniform random bombing strategy used at either of the locations in Figure 1.2? If so, which one? Furthermore, is there evidence at either location for certain positions to be bombed more heavily while others are avoided, possibly reflecting the strategic importance of the positions?

Many would say that the distribution of points at Location 2 represents the strategy of randomly picking a latitude and longitude from a uniform distribution. The positions for

[2] The name arises from sampling latitude and longitude positions from a uniform distribution, where all values between an upper and lower boundary have an equal probability of being chosen.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



Location 2 were instead generated by selecting a 10-by-10 grid of equally spaced positions, and then adding some noise to these values. This makes the bomb positions evenly spaced. The random uniform strategy is only used at Location 1. This appears counter-intuitive because there are large regions with no bombs, while other regions have a denser clustering. Such clustering and empty regions are to be expected under a uniform random strategy. Intuition about what randomness looks like does not come easily or naturally.

### 1.2.2 Not wanting to miss anything

Potentially meaningful patterns like the above example can be formally tested with a statistical analyses, but it is important to avoid using the same data to first find an interesting pattern (such as the lower left empty region of Location 1 in Figure 1.2), and then to statistically test for this pattern. For example, we might try to calculate the probability of seeing no bombs dropped in an area the size of the empty lower left region. Random data – especially if there is a lot of it – will have local regularities and patterns. Picking out one such pattern that catches our attention and then performing a statistical test has implicitly performed many informal tests, in that all of the patterns that *could have been* interesting were examined and discarded without a formal test. For example, it does not appear that more bombs were dropped at higher latitudes compared with lower latitudes (comparing the top versus the bottom half of Location 1). If such a pattern did appear to exist, then we would test that instead. The key principle is: *if a hypothesis is derived from the data, then the ability of the data to support that hypothesis is diminished.* The ability of the data to support a hypothesis can also be compromised by what others do. For example, a PhD student is the first person to analyse a data set and explores it thoroughly. He finds a relationship but is unsure of the appropriate statistical test and so brings it to the principal investigator's (PI) attention. The PI then conducts only one analysis and feels confident that the *p*-value is valid, because she is unaware of how the data were used to discover this relationship.

Even when a visual-driven inspection of the data is not so pronounced, people want to make the most of the data and to avoid missing anything interesting. This desire is likely greater when the primary result is not significant and then we have to see 'what else we can get out of the data'. One might begin to look for correlations between variables, then again after normalising or correcting for other variables. Then checking for differences between sexes, or the old versus the young, or the less severely affected compared to the most affected, and so on until there are enough interesting findings to report. On the one hand, it seems foolish not to thoroughly examine the data, given all of the work that went into generating it. On the other hand, such a search process can generate many false positives.

There are two approaches to limit the number of false positive results that arise from data-driven discoveries. The first is to divide the analyses into confirmatory and exploratory parts. The confirmatory analysis specifies everything in advance (before seeing the data), including the hypothesis to be tested, the main outcome variable, and the analysis that will be used. The subsequent exploratory analysis allows for greater flexibility to find other relationships of interest, but with the knowledge that the findings carry less weight and are

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



less convincing because they were not predicted in advance, even if attempts have been made to correct for multiple testing. The second approach is to validate the findings, either by conducting a subsequent experiment, or by dividing the data into two parts. Once the experiment is complete, but before any analysis, about 20–30% of the data are removed and locked away. The remaining data are used to find interesting relationships. Once the analysis is complete, the data that were locked away are used to confirm the findings. This is a common approach in the data mining, machine learning, and predictive modelling fields, but it does require enough samples to split into two sets.

People differ in how easily they detect signals in pure noise, find patterns in randomness, or meaning in the coincidental. A sign of 'inferential maturity' is to know where you lie on the spectrum. If you find anything vaguely resembling an association or effect interesting and tend to believe that it is 'real', then pay attention to controlling false positives. If instead you are sceptical and find only large associations or effects convincing, then you risk not further exploring small but true findings.

### 1.2.3 Psychological cliff at $p = 0.05$

One criticism of $p$-values is that they encourage dichotomous thinking – the effect or relationship is either significant or it is not – even though evidence is continuous.[3] In the 1960s, Rosenthal and Gaito showed that such a psychological effect exists. They asked psychology researchers and graduate students to rate their degree of belief or confidence in a research hypothesis with $p$-values ranging from 0.001 to 0.90. They found a 'psychological cliff' at $p = 0.05$ – an abrupt jump in confidence just below 0.05 [331, 332]. A replication experiment by Poitevineau and Lecoutre provided a more nuanced view [313]. They found a strong cliff effect, but only in a subset of subjects; others had a linear or exponentially decreasing confidence as the $p$-values increased (Figure 1.3).

The cliff effect likely contributes to another misinterpretation of statistical results, which Gelman and Stern have phrased as 'The difference between "significant" and "not significant" is not itself statistically significant' [132]. They are referring to a situation where, for example, Group A is significantly different from the control group, Group B is not significantly different from the control group, and then an incorrect conclusion is made that Group A is significantly different from Group B. If differences between Group A and B are of interest, then they need to be compared directly against each other.

To the extent that a small $p$-value provides evidence for a research hypothesis, there is no sharp evidential distinction between 0.04 and 0.06. An obvious question is 'What is the correct relationship between a $p$-value and evidence for a hypothesis?' The short answer is that there is no correct relationship because a $p$-value says nothing about hypotheses and so the question makes no sense. If you are interested in evidence or the probability that a hypothesis is correct, then likelihood or Bayesian methods are required. These are beyond the scope of this book but good introductions can be found in references [50, 95, 139, 201–204, 269].

[3] Technically, a $p$-value does *not* provide evidence against a hypothesis. Informally, however, a smaller $p$-value suggests that an effect is present. The interpretation of a $p$-value is discussed in Section 1.4.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



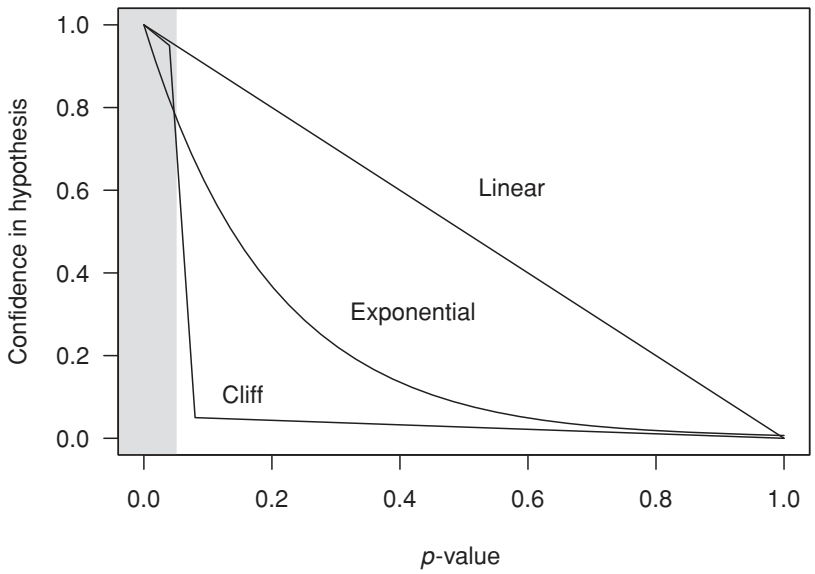


**Fig. 1.3** Confidence from *p*-values. A schematic diagram from Poitevineau and Lecoutre [313]. The confidence in a hypothesis drops rapidly just past $p = 0.05$ for some people. Shaded area is $p < 0.05$.

The key point is that there is nothing special about 0.05, or values on either side, that indicates an abrupt change in what the data have to say about a hypothesis. Gelman and Loken raise two related points about interpreting statistical results [130]. The first is that effects cannot be divided into those that are 'real' and those that are 'not real', based on a *p*-value. The presence and magnitude of effects and associations are conditional on (1) the sample material used, (2) background variables and conditions (such as laboratory equipment and experimenter), (3) the experimental design (was a blocking factor incorporated), and (4) data preprocessing and the statistical analysis. Since the (true) magnitude of an effect or association is always conditional on so many factors it makes sense to consider how the effect or association varies across diverse situations. Under some conditions the effect may be smaller and the *p*-value above 0.05, and this does not imply a lack of reproducibility.

Their second point is that the statistical analysis does not determine whether an effect is 'real', just as microscopes do not determine whether bacteria are real, but both microscopes and statistics can help one see things that are not obvious with the naked eye. Effects are determined by the biological process under investigation, the experiment used to probe it, and the data derived from it. Occasionally, effects are so large and clear that no statistical analysis is necessary. When the experiment is more complex and the results less obvious, a statistical analysis only helps one to interpret what is already there. Interpretation of the results may differ depending on the analysis, but so too may a conclusion about a phenomenon depending on the microscope (e.g. light, confocal, or electron). Do not believe a result just because 'the statistics said. . . '.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



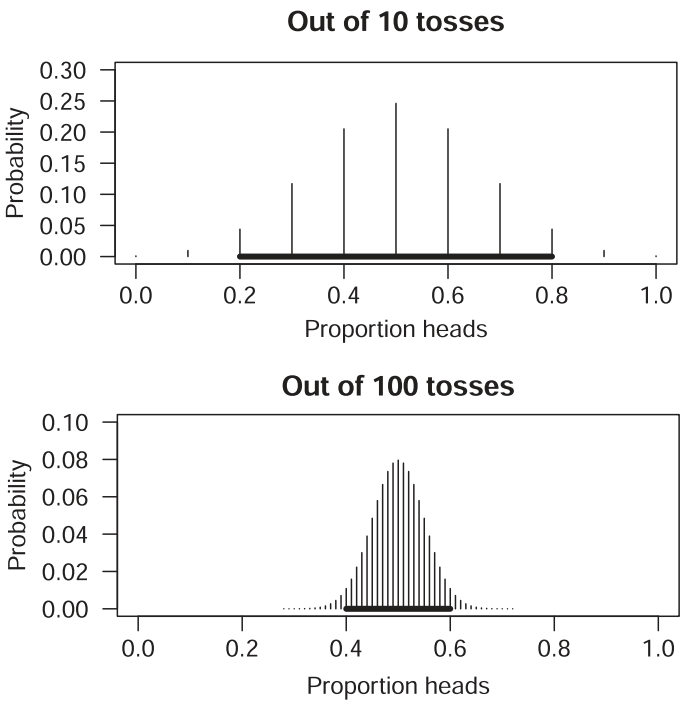


**Fig. 1.4** Sampling variability and its dependence on sample size. When tossing a coin 10 times, we expect the proportion of heads to be 0.5, but it would not be unusual to obtain a proportion between 0.2 and 0.8 (thick black line on the *x*-axis). When tossing a coin 100 times (larger sample size), the range of likely values for the proportion is narrower, between 0.4 and 0.6.

### 1.2.4 Neglect of sampling variability

Sampling variability is the reason that the outcome of a random process differs from run to run. If a fair coin is tossed 10 times, we would expect, on average, five heads and five tails. On any given trial we may get more or less heads, but we would expect most tosses to have between two and eight heads. Rephrasing, the expected proportion of heads is 0.5, with the majority between 0.2 and 0.8. This is sampling variability: we do not always get five heads, and is illustrated in Figure 1.4. Furthermore, as the sample size increases, the variability in the outcome decreases. When the sample size is increased to 100 tosses, we still expect the proportion of heads to be 0.5, but now the majority will lie between 0.4 and 0.6 – a narrower interval. This is the dependence on sample size: the larger the sample size, the narrower the interval of values that we are likely to see. As the sample size increases, we converge to the true proportion of heads when tossing a fair coin. These simple ideas appear often and can lead to incorrect inferences and conclusions if not taken into account. Some examples are discussed below.

Wainer provides two real-world examples [384]. In his first example he shows a map of kidney cancer death rates in the USA, with the top (worst) 10% of counties highlighted.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



Counties with high death rates tend to be rural and in the west or midwest, and one might speculate that people in these regions have unhealthy lifestyles or less access to high quality medical care, leading to higher death rates. The interesting twist is that if the counties with the *lowest* 10% of death rates are plotted, they also tend to be in the west or midwest. Counties with the highest and lowest rates are often right beside each other! What is going on? Counties in the west and midwest are sparsely populated, so the addition or subtraction of a few cases will have a larger influence on the cancer rate than in a larger population. Just as in the coin example above, with a small sample size the value of whatever is calculated will fluctuate more widely around the true value. Thus, sparsely populated counties are over-represented at both ends of the death rate distribution.

Wainer's second example discusses the billions of dollars wasted on supporting smaller schools in the USA by educational charities. Smaller schools tend to outperform larger schools on student achievement, and so a logical conclusion is that if larger schools were split into several smaller ones, then student achievement would increase. However, smaller schools are expected to be over-represented in both the top and bottom of the achievement distribution, and that is what Wainer found. Again, with a small sample size (number of students), achievement results will have a wider spread and therefore the tails of the distributions will mostly contain smaller schools.

This phenomenon of larger variances with smaller sample sizes is also relevant for biological experiments. The following situation is often seen in high-throughput screening experiments. The data are simulated, but the example is from a real experiment. Suppose 5000 compounds are tested in a cellular assay and the goal is to find compounds that increase the number of cells expressing a key protein marker. Cells are plated in high-density microtitre plates and images are taken after treatment with the compounds. The total number of cells and the number of cells positive for the marker (which is a subset of the total cell count) are obtained from the images. There are no active compounds in this simulation and so the results are what would be expected from random fluctuations. The distribution of total cell counts across all 5000 wells is shown in Figure 1.5A. The average number of cells per well is 122, but it ranges from 3 to 425. The distribution of positive cell counts across the 5000 wells is shown in 1.5B. The number of positive cells is related to the total number of cells (Figure 1.5C), and the proportion (or percentage) of positive cells is calculated by dividing the number of positive cells by the total number of cells (Figure 1.5D). The proportion was calculated because it supposedly removes the dependence on total cell count. The mean proportion is 0.25 and the range is 0 to 0.8 (the few high values are hard to see in this graph). A cut-off is made based on some criterion such as three standard deviations above the mean of the distribution (dashed line in Figure 1.5D), and all compounds above that are considered 'hits' and will be tested in further experiments. The criterion or threshold used to determine a hit is not important for this example.

Figure 1.5E shows how the *variation* in the proportion of positive cells is dependent on the total cell count, even though the mean no longer is. When cell counts are low, variation is high, and vice versa. The horizontal dashed line is the threshold for hit calling, and all of the hits (high proportion of positive cells) are from wells with few cells. One compound

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-107-07429-3 — Experimental Design for Laboratory Biologists
Stanley E. Lazic
Excerpt
More Information



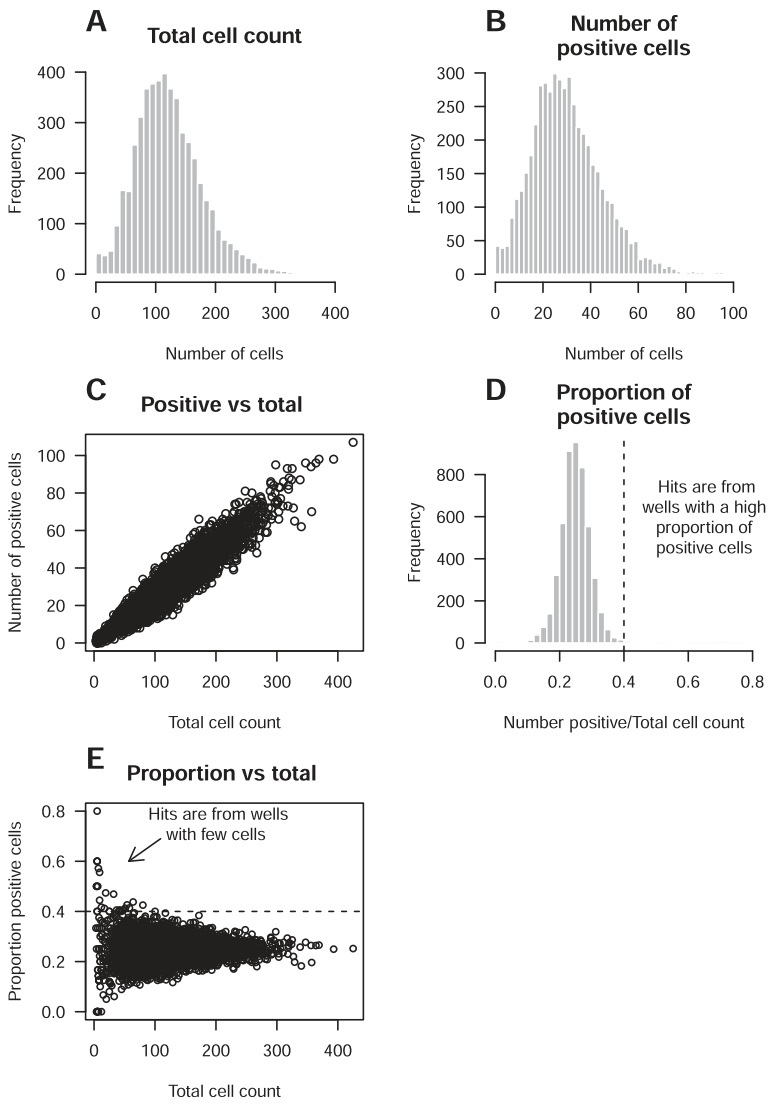


**Fig. 1.5** Sensitivity to sample size. The two measured variables are the total cell count and the number of positive cells (A, B). There is a correlation between these variables because the positive cells are a subset of the total number of cells (C). A common strategy to remove the dependence on total count is to take the ratio of positive to total cells, and to look for high ratios (D). However, wells with few cells have the highest ratios (E), which are statistical artefacts.

has a very high proportion of 0.8. This is a statistical artefact but is routinely seen in real experiments and can be mistaken for a true hit. Resources could then be wasted in trying to validate it. But why does this occur? Think of each cell as having a probability of being positive, determined by the flip of an unbalanced coin. Each cell has a 25% chance of being positive (coin lands heads) and a 75% chance of being negative (coin lands tails). If there are only three coins, it is not unusual that all three of them land heads (proportion = 1),

© in this web service Cambridge University Press www.cambridge.org