Notice that the CC species are, relative to the overall average, over-represented in the WD classification, the CR species are over-represented in the D classification, while the RC species are under-represented in D and WD and over-represented in W.

Interpretation issues

Having found an association in a contingency table, what does it mean? The interpretation will differ depending on the context. The incidence of gastric cancer is relatively high in Japan and China. Do screening programs help? Here are two ways in which the problem has been studied:

- In a long term follow-up study, patients who had surgery for gastric cancer may be classified into two groups – a "screened" group whose cancer was detected by mass screening, and an "unscreened" group who presented at a clinic or hospital with gastric cancer. The five-year mortality may be around 58% in the unscreened group, compared with 72% in the screened group, out of approximately 300 patients in each group.
- In a prospective cohort study, two populations – a screened population and an unscreened population – may be compared. The death rates in the two populations over a ten-year period may then be compared. For example, the annual death rate may be of the order of 60 per 100 000 for the unscreened group, compared with 30 per 100 000 for the screened group, in populations of several thousand individuals.

In the long term follow-up study, the process that led to the detection of cancer was different between the screened and unscreened groups. The screening may lead to surgery for some cancers that would otherwise lie dormant long enough that they would never attract clinical attention. It is necessary, as in the prospective cohort study, to compare all patients in a screened group with all patients in an unscreened group. As patients were not divided randomly between the two groups, results are even so not conclusive.

Modeling approaches

Modeling approaches typically work with data that record information on each case separately. Data where there is a binary (yes/no) outcome, and where logistic regression may be appropriate, are an important special case. Chapter 5 gives further details.

2.4 A critique of *p*-value methodology

Statistical summary information has to be judged in context, as one component of the evidence. If results from the experiment are to be made the basis for a recommendation for changes to farming practice or to medical treatment, the evidence will need to be strong. If the interest is in deciding which effects merit further experimental or other investigation, relatively weak evidence may be acceptable. There should be a careful balancing of the likely costs and benefits of any decision.

What constitutes evidence? What do p-values contribute? The discussion of this section is intended as a relatively non-technical answer to this question. There are however subtle technicalities that, in the attempt to get a clear understanding of what inferences can and cannot reasonably be drawn from p-values, cannot be avoided.

Subsection 2.4.1 will discuss the common paradigm that sets a significance threshold α, commonly with $\alpha = 0.05$, and treats all $p \leq \alpha$ as "significant". Subsection 2.4.3 will consider the evidential value of the specific p-value.

2.4.1 *From $p \leq \alpha$ to the probability of an effect?*

In cases where the interest is in a confidence interval for a mean or for a difference μ, and where there is no formally defined null hypothesis, it can be reasonable to expect that plausible values for μ will follow a roughly bell-shaped distribution. Treating a 95% confidence interval as a statement about probable values of μ, although wrong, may then be not be too seriously misleading!

Where there is a precise null, p-values can more readily be seriously misleading. In this case, the interest is commonly in testing the null hypothesis $\mathrm{H}_0: \mu = 0$, against the alternative $\mu \neq 0$ or $\mu > 0$. For simplicity, the alternative $\mu \neq 0$ will be assumed.

At one extreme, consider an experiment that is designed to test the effect of a drug on the length of time that mice sleep. If an inactive substance rather than the drug is added to the food, the prior odds of finding a difference should be zero. On average, 5% of experiments will return $p <= 0.05$. In 2000 experiments 100 will, on average, return $p <= 0.05$.

More generally, suppose that among 2000 drugs tested, 20 have an effect that is in principle detectable, while 1980 are inactive. The (usually unknown) prior odds R are, in this thought experiment, 20:1980 = 1:99 .

At this point, the idea of "power" is needed. The probability that a true positive will be detected as such has the name "power", denoted in the sequel as P_w. Suppose for example that the power is $P_w = 0.8$, relative to the $\alpha = 0.05$ cutoff. (This is high relative to the standards of much published work.) Then results will divide up as follows:

```
True positives False positives
   0.8*20 = 16  0.05*1980 = 99
```

True positives are then much less likely that false positives, by a ratio of 16:99.

As the prior odds increase from 0, the probability that $p <= 0.05$ will indicate a real effect will, unless $P_w = 0$, increase from zero. If $P_w = 0$ (the effect would be that of tossing a 20-sided die, and calling $p < 0.05$ if face 1 comes up), 5% of the false positives will be detected as real. The probability that $p <= 0.05$ will indicate a real effect nevertheless remains zero.

In an article whose primary focus is medical research, Ioannidis (2005) gave reasons why "most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds.". Given low pre-study odds R, any notion that $p \leq 0.05$ provides assurance that a claimed effect is real has to be abandoned. Depending on how small R is, a very small p-value, or repeated small p-values in a series of well-powered trials, may be required to shift the weight of evidence so that the posterior odds for a real effect are substantially greater than 1:1. The effect of reducing the threshold for significance to $\alpha = 0.01$ or lower is to insist on a larger shift in the weight of evidence.

One experiment with $p \leq 0.05$ is not conclusive!

Here note what Fisher (1926), who introduced the use of *p*-values, had to say:

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely* fails to give this level [0.05 or 0.02] of significance.

In other words, *p*-values serve as a screening device, to identify results that may merit further investigation. This is very different from the way that *p*-values are commonly used in most current scientific discourse. A *p*-value is a measure of change in the weight of evidence, not a measure of the absolute weight of evidence. Subsection 2.4.2 will flesh out the argument in more mathematical detail.

Subsection 1.1.2 drew attention to published work that gives an indication of the extent to which much published work in laboratory science is not reproducible. A major contributor to the problem is the use of statistical analysis as a substitute for, rather than as a complement to, direct checks on reproducibility. Replication, independently in another laboratory, provides an indispensible check on all the processes involved. Where mistakes have been made, it is unlikely that another research group will repeat the same mistakes. If the paper and supplementary material does not describe work with sufficient accuracy, or in sufficient detail, to allow work to be reproduced, this becomes immediately obvious.

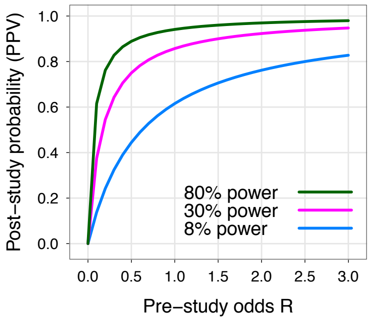
The use of p-values in scientific papers

Reliance on multiple significance test results adds to the problems already noted. It can be difficult to know what to make of the combined results. Significance tests should be closely tied to points that relate to the main thrust of the paper. Try to assess prior probabilities, at least to the extent of distinguishing between studies with "low prior odds" and those with prior odds that may be of the order of 1:1 or higher. Once it seems clear that an effect is real, the focus of interest should shift to its pattern and magnitude, and to its scientific significance.

It is a poor use of the evidence in the data to perform tests for each comparison between treatments when the real interest is (or should be) in the overall pattern of response. Where the response depends on a continuous variable, it may be pertinent to ask whether, e.g., the response keeps on rising (falling), or whether it rises (falls) to a maximum (minimum) and then falls (rises).

2.4.2*A more detailed analysis — Positive Predictive Values

The discussion that follows will have elements of a Bayesian perspective on *p*-value methodology and to that extent foreshadows Subsection 2.10.2. In the argument in Subsection 2.4.1, the prior odds R was an actual but unknown ratio of drugs with a potentially detectable effect to those with no effect, giving a prior probability is $\frac{R}{R+1}$. The reasoning of Subsection 2.4.1 carries through in the same way if R is a guess at a prior odds ratio, perhaps based on previous experience.



Example: With R = 1:15 (e.g. 100 true +ves to 1500 true −ves), $\alpha = 0.05$, and $P_w = 0.3$, the posterior odds are:

$$\frac{RP_w}{\alpha} = \frac{0.3}{15 \times 0.05} = 0.4$$

$$\text{PPV} = \frac{0.4}{1 + 0.4} \simeq 0.29$$

Figure 2.5: Post-study probability (PPV)), as a function of the pre-study odds, for different levels of statistical power.

The effect size (ES) for identifying a detectable effect may be chosen as the smallest change that is of interest. For instance, in an experiment to compare drugs that induce sleep (as in the dataset `datasets::sleep`), a minimum effect size of ES = 30 min might be reasonable.

The (posterior) probability of a genuine effect of the given size is known as the *Positive Predictive Value* or PPV. Examples that will be given shortly should make the idea clear. The *False Discovery Rate* or FDR is related to the PPV thus:

$$\text{FDR} = 1 - \text{PPV}$$

Given a single p-value, estimation of the PPV or FDR requires contextual information that typically has to be guessed. Given a large number of p-values that relate to the same experiment or experimental program, the FDR provides a helpful summary. With $\alpha = 0.05$, 5% of tests are expected, under the null hypothesis, to show $p <= \alpha$. Subsection 2.8.3 will show how the proportion that exceed this threshold can be used to estimate the FDR.

Given that a decision has been made in advance to regard as *significant* any p-value that is less than α, the power P_w is the probability of detecting an effect of the specified size. (Note that $\beta = 1 - P_w$ has the name "Type II error".)

Thus, given that the ratio of positives to negatives is $R : 1$, the odds that an apparent positive will be a true positive is:

$$\frac{RP_w}{\alpha} = R\frac{P_w}{\alpha} \tag{2.2}$$

i.e., multiply the prior odds R by $\frac{P_w}{\alpha}$ to get an odds ratio that accounts for the new evidence. Ioannidis (2005) has a modification of Equation 2.2 that allows for bias.

The positive predictive value (PPV), or posterior probability, is then:

$$PPV = \frac{RP_w}{RP_w + \alpha} \tag{2.3}$$

Figure 2.5 illustrates what the formula means in practice.

Button et al. (2013) report, based on what has been reported in a large number of different meta-analyses, on estimates of the power in a large number of neuroscience studies. A power of 0.8 is at the high end of the range, relative to what those authors report. Very low power values are common. Based on eight journal articles that were published between 2006 and 2009, Button et al. (2013) report an astonishing median power of 0.08 across 461 individual studies of brain volume anormalities. Results were better, but still not encouraging, in neuroscience more generally. Based on meta-analysis reports published in 2011, Button et al. (2013) found a median power of 0.21 for 730 individual primary studies.

Prinz et al. (2011) reported a success rate of around 30% in their efforts to reproduce the main result in 67 published studies. Two scenarios that, in the absence of other faults, would on average reproduce this approximate success rate are:

- R = 1:15, $\alpha = 0.05$, $P_w = 0.3$
- R = 1:4, $\alpha = 0.05$, $P_w = 0.08$

Calculations for the first of these are shown to the right of Figure 2.5. These scenarios ignore the likely contributions of design, data, execution, and presentation faults. Such faults will increase the risk of finding a spurious effect, the risk of failing to detect a genuine effect, and the risk of biases that distort the result.

Small studies with low power compromise the use of p-values for any purpose – whether as a screening device, or to confirm an earlier "result". As Button et al. (2013) note, there are other problems – the magnitude of a true effect, when found, is on average exaggerated. Small studies may not be conducted with the same care as larger studies that require more careful organization, and the smaller datasets that result are more susceptible to minor changes in the analysis process.

The issues to which Equation 2.3 and Figure 2.5 relate are of central importance for large areas of laboratory science. See in particular Ioannidis (2005). In practical use, for estimating the PPV, these formulae provide only broad guidelines. The estimate of the power P_w, and an assessment of what is a reasonable effect size, is often based on pilot study information from a sample that has been chosen for convenience rather than randomly selected, and is susceptible to selection bias. Estimates that are accurate enough to be a good basis for design may be available when a trial is the latest in a series of comparable trials. Gelman and Carlin (2014) argue for the use of information external to the specific study as a basis for choosing the effect size. A literature review will often provide useful leads.

2.4.3 Interpreting the specific p-value

For the apparent positives in the account just given, p is from a distribution in the range $p \leq 0.05$; thus $p = 0.01$, $p = 0.005$, and $p = 0.05$ all add just 1 to the count. Clearly however, $p = 0.005$ is stronger evidence against the null than $p = 0.05$.

A result given in Sellke et al. (2001) suggests that the following, with the same prior odds of 1:25 for an effect in each case, may give roughly the same PPV = 0.3:

- With Power = 0.5, treat as statistically significant any $p \leq 0.05$
- The experiment has returned $p = 0.007$

According to the Sellke et al. (2001) result, given p and assuming $p < e^{-1}$ the posterior odds are no greater than:

$$\frac{R}{-eplog(p)} \tag{2.4}$$

For $p = 0.05$, the odds ratio given by Equation 2.4 is slightly less than $2.5R$. We leave the reader to verify that Equation 2.4 implies that $p = 0.007$ shifts odds of 1:25 for an effect (a positive) to, at most, slightly more than 3:7. Thus, PPV $\simeq 0.3$, close to the figure that Prinz et al. (2011) found when considering the main results from the studies that they investigated. As noted, this account of circumstances that might give a PPV of 0.3 is suggestive only. It neglects other likely contributing factors.

Increasing the power shifts the distribution of p-values in a direction that favors obtaining more small p-values. The bound given in Equation 2.4 still applies, but for a p from a distribution of p-values in which there are relatively more small values. The odds ratio given by Equation 2.2 is a composite over the different values $p <= 0.05$ that result for studies with the specified power P_w.

The take-home message is that a p-value is a measure of *the extent to which experimental results shift the weight of evidence.* It is not an absolute measure of the weight of evidence.

2.4.4 Reproducibility studies

Concerns about the reproducibility of published results are the motivation for several major projects. Two major studies, conducted under the auspices of the Center for Open Science (`http://centerforopenscience.org`), and on which results have been reported, are:

- Klein et al. (2014), involving 36 independent samples with a total of 6344 subjects in the US and internationally, checked the reproducibility of 13 classic and contemporary psychology studies. Replications were successful for 10 of the 13 studies, weakly successful in one case, and unsuccessful in two cases. The data are available online (at `https://osf.io/ebmf8/`). Plots that show differences between the different samples are much as expected. See also Yong (2012).
- The "Reproducibility: Psychology" project replicated 100 studies, published in 2008 in one of three journals. Using a simplistic $p \leq 0.05$, around 40% of the studies were successfully reproduced. Data and R code used for analyses in the paper Open Science Collaboration (2015), are available from `https://osf.io/fgjvw/`.

A $1.3 million grant from the Laura and John Arnold Foundation is funding replication of the 50 "most impactful" cancer biology studies from 2010-2012. At the time of writing, this study is still in progress.

It is important that replications are carried out independently in different laboratories. This limits the influence of circumstances that are specific to the individual trial or laboratory. Those circumstances may include deficiencies in design and/or execution.

Results from replication studies are important in establishing the extent to which publications (even in very reputable journals) can be relied on. They provide important evidence, in the areas covered, on the extent of problems with current scientific processes. The tightening of reproducibility standards has to be managed so that it leaves room for imaginative exploration.