> "I am still too often confronted by problems, even in my own research, to which I cannot confidently offer a solution, ever to be tempted to imply that finality has been reached..."
>
> RONALD A. FISHER
> *Preface to Contributions to Mathematical Statistics (1950)*

Statistical inference

RICHARD D. MOREY

Contents

Scientific work involves a variety of kinds of inferences. One can reason deductively from a theory to predictions, or from observations to falsification of a theory. Inductive inference is often used when a theory is argued to be supported by many experiments. Informal inferences play an important role in deciding what lines of research might be fruitful, or whether an experiment has produced useful data.

This draft was compiled 17/04/2017 at 17:58:46 .

Statistical inferences are among the most common inferences in science. Over the past century, statistical inference has grown to be an indispensable tool for scientists seeking to make sense of data. It is perhaps surprising, then, that throughout this century there has been, and continues to be, debate about the nature of statistical inference and how it should be performed. In this chapter we present the dominant perspectives in broad strokes, each on their own terms.

What is statistical inference?

Statistical inference, most broadly conceived, is concerned with drawing an inference in the presence of variability or uncertainty. Take, for instance, the problem of establishing the Stroop effect (Stroop, 1935), the relative slowdown when indicating color of a printed stimulus when the stimulus is a color word printed in a different color, versus simply indicating the color of a printed square. Suppose we ask a participant to respond only twice: once to a trial with stimuli in which the word and color conflict (for instance, the the word "red" printed in blue ink; an *incongruent* trial), and once to trial in which the stimulus is simply a colored square (a *nonword* trial). Further suppose that the participant takes 1.1s to respond correctly to the nonword trial, and 1.2s to respond to the incongruent trial. Without a concept of "variability" one might naively assume the settled; the nonword trial was faster than the incongruent trial.

Upon reflection, however, it is obvious that these two trials are insufficient for a scientifically interesting inference. We are not just interested in the difference between two specific trials, for a given participant, at a given time. If we had asked the participant to do another two trials, or chosen different colors for the two trials, or tested a different participant, we would have obtained different response times. Response times *vary* from trial to trial, and from person to person. Particular trials are not of interest; rather, we are interested in understanding, in some sense, the properties of the set of all possible trials for all relevant people.

A statistical inference is any inference about a hypothetical collection of all possible data (called a *population*) using a limited amount of observed data (called a *sample*). The most basic form of statistical inference is informal: the so-called "interocular traumatic test" – that is, the conclusion hits one between the eyes (Berkson, 1958, as cited in W. Edwards, Lindman, & Savage, 1963, p. 217). Consider the data from Experiment 2 reported by Stroop (1935), shown in Figure 1.

The so-called "Stroop effect" — the tendency for responding to be slower to naming the color of ink when it spells out an incongruent color than when it spells out a congruent color — is one of the most robust effects in experimental psychology. The distributions of average response times to incongruently-colored words and congruently-colored words seem very different; in fact, in this case one might forego statistical inference altogether and simply present the response time distributions.

In many cases that face experimental psychologists, however, the results are less clear; as W. Edwards et al. (1963) put it, "the enthusiast's interocular trauma may be the skeptic's random error." Data may be noisy and may require substantial reduction to obtain statistics that are appropriate for inference. Simply "looking at"

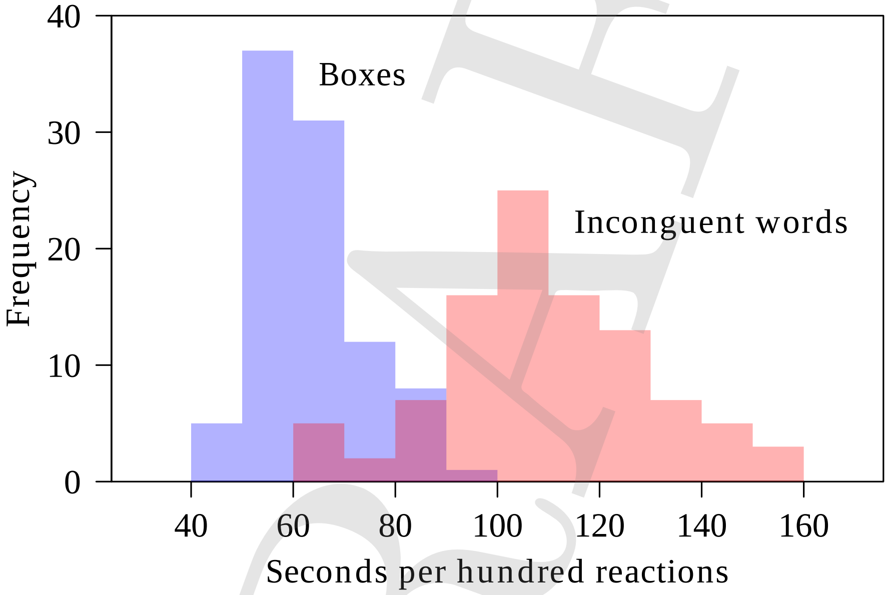

Figure 1: Distribution of seconds taken to read 100 Stroop trials in two Stroop conditions. Each talleyed unit is an average response time for a participant in that condition. Frequencies were approximated from Figure 1 of Stroop (1935).

the data is often not sufficient to draw an inference; it is in these cases that formal statistical inference is needed.

Populations and parameters

A *population* is an idealized representation of the true[1] but unknown distribution of some quantity. Suppose we are interested in the speed of a particular person's responses to a briefly-presented stimulus. The speed of these responses will vary from trial to trial. However, we can ask questions about the nature of this variability: what proportion of responses will be less than 200 milliseconds? What proportion will be between 750 milliseconds and 1 second? There are an infinite number of such questions; the population is the characterization of the answer to all of them.

Statistical populations will vary from one another. For instance, a person's response time distribution in one condition might be different than their response time distribution in another condition. *Parameters* of a population are quantities that describe how populations vary. Figure 2A shows how populations vary when a so-called *location* parameter is varied: the distributions merely shift from lesser to greater values, on average. The mean parameter (μ) of the normal distribution is an example of a location parameter. Changing a *scale parameter*, on the other hand (Figure 2B), will "stretch" the population but will not change its shape. The standard deviation of the normal distribution is another example of a scale parameter. A *shape* parameter is any parameter whose effect cannot be described as a shifting or stretching of the population (Figure 2C).

We now move to making inferences about such parameters. We will approach inference from three different perspectives: frequentist, likelihood, and Bayesian inference. Space constraints prevent addressing other approaches (e.g., information theoretic), but the three approaches described here account for a large majority of statistical inferences drawn in the scientific literature.

Frequentist approaches

Frequentism is a philosophy of statistical inference that is, broadly speaking, concerned with properties of procedures in repeated sampling. In estimation, this might present itself as a concern for bias or minimum error in repeated samples: that is, we might prefer an estimate that – on average – is equal to its true value, or that – on average – deviates from the true value as little as possible. In hypothesis testing, the concern for properties in repeated sampling manifests itself as a concern for error rates in decisions: for instance, how often one might reject a particular hypothesis, if it is indeed true. More comprehensive treatments of frequentist methods can be found elsewhere (David R. Cox, 2006; de Groot & Schervish, 2012; Hogg & Craig, 1978; Rice, 1998); here, I present a summary of the main concepts.

[1] We assume that "truth" here is a sort of useful approximation. Real populations are nothing like statistical populations, but statistical populations may serve as appropriate representations of specific aspects of a real population.

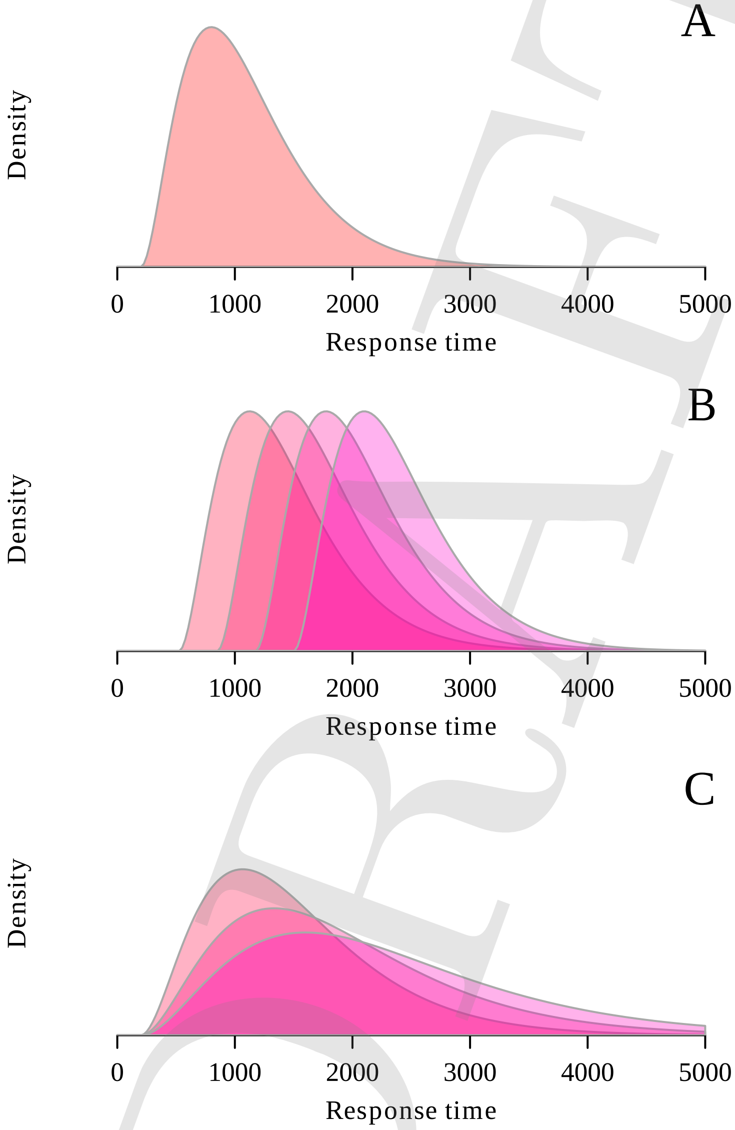

Figure 2: Examples of changes in various kinds of parameters. See text.

Point estimation

A point estimate of a parameter is a single value that, in some sense, is the "best" estimate for a parameter given the data we observed. How one defines "best" will depend on the statistical philosophy one applies to the problem. In frequentist statistics, point estimates are identified with procedures for arriving at the estimates called *estimators*. For instance, the sample mean $\sum x_i/N$ is a common estimator for the expected value of a population. Estimators may have good or bad properties. The following properties are considered important in frequentist statistics.

- *Consistency.* A consistent estimator is one that converges to the true value of the parameter as the sample size increases. Consistency is often considered to be a minimum requirement for an estimator.

- *Bias.* Bias is the difference between the expected value of an estimator and the corresponding true parameter value. If the expected value of an estimator is equal to the true value of the parameter, the estimator is called *unbiased.*

- *Variance.* The expected squared difference between an estimator and its mean (not necessarily the true value) is called the variance of the estimator. If a particular unbiased estimator achieves the smallest variance that can be achieved by any unbiased estimator, it is called a *minimum-variance unbiased estimator* (MVUE).

- *Frequentist loss and risk.* A loss function is a function that is designed to penalize "worse" parameter estimates, typically in the sense of estimates that are more distant from the true value. For instance, a quadratic loss function $L_q(\hat{\theta},\theta)$ penalizes estimates $\hat{\theta}$ in proportion to their squared distance from the true value: $L_q(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$. Frequentist risk is the expected loss, where the expectation is taken over the sampling distribution of the estimator. A higher-risk estimator is one from which we would expect a higher loss, on average.

- *Efficiency.* An efficient estimator is one that is, in some sense, expected to be "close to" the true value. For instance, the root-mean-squared-error ($RMSE(\hat{\theta},\theta)$) of an estimator can be used to assess efficiency:

  $$RMSE(\hat{\theta},\theta) = \sqrt{E\left(\hat{\theta}-\theta\right)^2}.$$

  The lower the RMSE, the more efficient the estimator. Another way to think about efficiency is the opposite of risk: low-risk estimators are high-efficiency estimators. The link between risk and efficiency can be seen in the RMSE, which is the square root of the frequentist risk assuming a quadratic loss function.

A simple example will suffice to demonstrate the properties of estimators. Consider the problem of estimating the standard deviation of a normal population, given a sample of size N, $y_1,\ldots,y_N$. The typical estimate s is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N-1}}$$

which is the square root of the unbiased estimate of σ^2. The estimator s is a non-linear function of an unbiased estimator, it will not be an unbiased estimate of σ, due to the fact as s^2 gets larger, the square-root transformation is more and more extreme. In fact, the expected value of s is

$$E(s) \quad = \quad \sigma \times \left(\frac{n-1}{2}\right)^{-\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}, \tag{1}$$

which is obviously not σ. s is a biased estimator.

In the box below the expected value of s is computed when $\sigma = 1$ and $N = 10$. On the left, the result is computed exactly; on the right the result is estimated by sampling, using R code (R Core Team, 2013). Code in subsequent Demostration boxes will also be R code.

Demonstration

$$\begin{aligned} E(s) &= \sigma \times \left(\frac{n-1}{2}\right)^{-\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \\ &= 1 \times \frac{1}{\sqrt{4.5}} \times \frac{24}{11.632} \\ &= 0.973 \end{aligned}$$

```
n = 10
sigma = 1
sds = replicate(n=10000, {
  sd(rnorm(n, 0, sigma))
})
mean(sds)

## [1] 0.973
```

(see Cureton, 1968b, 1968a). For $n = 10$, the average underestimation of s when estimating $\sigma = 1$ is about 3%, as one can see in the Demonstration box above. To mitigate this underestimation, we might consider a second estimator of σ denoted s_u (the u for *unbiased*):

$$s_u = s \times \left(\frac{n-1}{2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Both estimators s and s_u are consistent by the continuous mapping theorem and Slutsky's theorem (Casella & Berger, 2002), respectively, owing to the consistency of s^2.

Figure 3A shows the sampling distributions — that is, the distributions over repeated samples — of s and s_u when $\sigma = 1$ and $n = 10$. They both appear to be approximately centered around $\sigma = 1$, but s is somewhat lower on average than s_u. The expected values of the two distributions are shown as vertical gray lines (s_u as dashed). The expected value of s_u is exactly σ: it is an unbiased estimator. The expected value of s — computed via Eq. 1 — is 0.973. Because $\sigma = 1$, the bias in s is -0.027. Figure 3B shows the bias in the two estimators as a function of n; it is apparent that the bias in s falls dramatically as the sample size increases.

It is also apparent from Figure 3A that s_u is slightly more spread out than s, and hence more variable. Because the correction factor by which we multiplied s by to

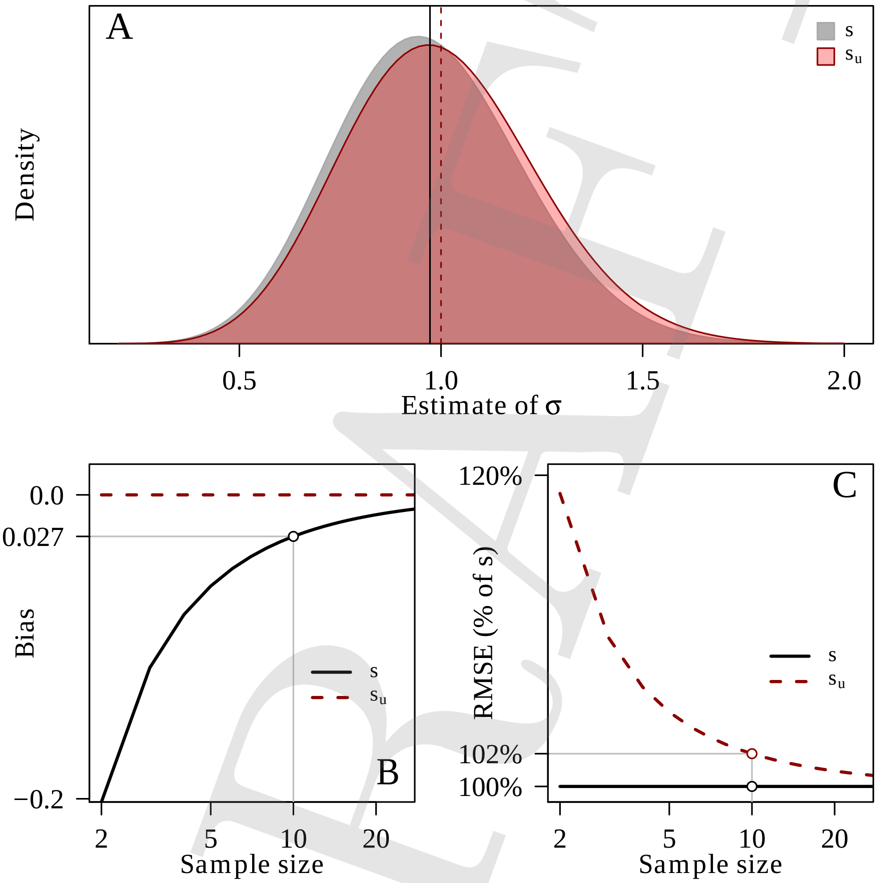

Figure 3: (A) The sampling distributions of two estimators (see text) of a normal population's standard deviation σ^2, when the true $\sigma = 1$. (B) The amount of bias in these two estimators as the sample size N varies. (C) The RMSE of the estimators relative to the RMSE in the estimator s, as N varies.

obtain s_u was greater than 1, the variance of s_u (here, 0.057) must be greater than that of s (here, 0.054). But note that this represents the variance *around the estimators' respective expected values*, not the true value; in order to determine which estimator would be preferable in practice, we need to assess the estimators' risks with respect to a loss function — or, alternatively, their efficiencies.

As a measure of efficiency, we use the RMSE, the square root of the risk with respect to quadratic loss. A better estimator will have lower RMSE. In order to compute the mean squared error (MSE), we make use of the identity

$$E\left(\left(\hat{\theta}-\theta\right)^2\right)=V(\hat{\theta})+\left(E\left(\hat{\theta}-\theta\right)\right)^2,$$

or, in words, the MSE of an estimator is equal to the sum of the variance of the estimator and the square of the estimator's bias. For the usual estimator of the standard deviation, $MSE(s,\sigma=1)=0.054+(-0.027)^2=0.055$. For the unbiased estimator, $MSE(s_u,\sigma=1)=0.057+0=0.057$. Taking the square roots yields the RMSEs, which are 0.234 and 0.239, respectively. The RMSE for s_u is about 2% higher than that of s; the usual estimator of the standard deviation is the more efficient one, in spite of being biased. Figure 3C shows the RMSE of both estimators relative to $RMSE(s,\sigma=1)$ as a function of n. The usual estimator s is always more efficient, but this advantage diminishes as the sample size grows.

Demonstration

The easiest way of estimating the RMSE is often Monte Carlo simulation. The R code to the right estimates the MSE and RMSE of s for $n=10$ and $\sigma=1$.

```
n = 10
sigma = 1
SqEr = replicate(n=10000, {
  x = rnorm(n, 0, sigma)
  (sd(x) - sigma)^2
})
mean(SqEr)

## [1] 0.0558

sqrt(mean(SqEr))

## [1] 0.236
```

This example shows an important fact about estimators: choosing an appropriate estimator involves a tradeoff between efficiency and bias. Making an estimator unbiased will often reduce its efficiency, and if one is willing to give up some bias, one can gain some efficiency.

Hypothesis testing

It is often the case that a single parameter value is seen as "privileged" — that is, it expresses something important that we might want to test, in a way that other parameters do not. Take, for example, the hypothesis that two population means are the same. If the difference between the two means is δ, we might wish to test the hypothesis that $\delta = 0$. Another example of such a hypothesis is that there is no relationship between two random variables: they are statistically independent. These privileged hypotheses are called “null” hypotheses, both because they often express nullity (no effect, no difference, no relationship) and because they serve as a “default” hypothesis to be rejected.

Typical approaches to frequentist hypothesis testing use a principle of evidence involving decision error rates: a statistic provides evidence for a hypothesis to the extent that we would not often be fooled into deciding in favor of a false hypothesis. “Often” here is taken in the sense of rates over repeated sampling.

Null hypotheses and test statistics

One of the prototypical null hypotheses that we might like to test is that two populations have the same mean. Suppose that we are sampling from two normal populations, each with variance σ^2. We draw n_x random samples $x_1, \ldots, x_{n_x}$ from the first population (with true mean μ_x) and similarly for the second sample $(y_1, \ldots, y_{n_y})$. We can thus write our statistical model as

$$\begin{aligned} x_i &\sim \text{Normal}(\mu_x, \sigma^2),\ i = 1, \ldots, n_x \\ y_j &\sim \text{Normal}(\mu_y, \sigma^2),\ j = 1, \ldots, n_y \end{aligned}$$

In order to create a hypothesis test, we must decide on a *test statistic* that captures the evidence in the data with respect to the null hypothesis $\mathscr{H}_0 : \mu_x = \mu_y$. A natural test statistic for assessing the difference between μ_x and μ_y is $d_{\bar{x}\bar{y}} = \bar{x} - \bar{y}$. Larger, positive values of $d_{\bar{x}\bar{y}}$ tend to support $\mu_x > \mu_y$; larger negative values of $d_{\bar{x}\bar{y}}$ tend to support $\mu_y > \mu_x$; and values around 0 are equivocal.

We face an immediate problem in making an inference about $d_{\bar{x}\bar{y}}$: the sampling distribution of $d_{\bar{x}\bar{y}}$ is unknown, because it depends on the unknown variance σ^2. It is therefore impossible to assess the extremeness of any particular observed value of $d_{\bar{x}\bar{y}}$.

We can, however, estimate the *standard error* of $d_{\bar{x}\bar{y}}$, which is the standard deviation of the estimate in repeated samples. The standard error is

$$s_d = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} \Big/ n_{eff}}$$

where the effective sample size $n_{eff} = n_x n_y / (n_x + n_y)$ and s_x^2, s_y^2 are the variances of their respective samples. The effective sample size accounts for the fact that the uncertainty in the difference between the two means is dependent on the uncertainty in the two group means separately; for instance, 50 observations in each group would not yield the same amount of information as 98 observations in one group and 2 in the other.

If the standard error is an estimate of the degree to which $d_{\bar{x}\bar{y}}$ would be expected to vary from sample to sample, we can construct a statistic that quantifies the deviation of $d_{\bar{x}\bar{y}}$ from 0 *relative to its standard error*:

$$t = \frac{d_{\bar{x}\bar{y}}}{s_d}$$

The t statistic can be interpreted as the difference between the two means, relative to the variability of that difference as estimated by the standard error. Larger values of $|t|$ are less consistent with the null hypothesis $\mathscr{H}_0$, and hence would cause us to doubt $\mathscr{H}_0$ more.

In this context, the t statistic is called a test statistic, because it carries the evidence in the data regarding the quantity of interest. Assuming that the two populations are normal and have the same variance under the null hypothesis the t statistic has a known distribution: Student's t distribution, an example of which is shown in Figure 4A.

The sampling distribution of a test statistic assuming that the null hypothesis is true will depend on known factors. For instance, the shape of Student's t distribution depends on its so-called degrees of freedom through the sample size. The distribution in Figure 4A is Student's t distribution with 78 degrees of freedom, appropriate for testing the differences between the means of normally-distributed, equally-variable populations with 40 pre-planned, independent samples from each population.[2] Other designs and models will have different test statistics, but all frequentist hypothesis tests use some knowledge of the sampling distribution of a test statistic under a particular, assumed hypothesis.

If one knows the sampling distribution of the test statistic, one can compute the probability of observing a test statistic as extreme, assuming that the null hypothesis is true. Suppose one obtained a sample of participants in the two-group design mentioned above with 40 participants in group, and the observed t statistic with 78 degrees of freedom was $t_{78} = 1.4$ (Figure 4A, vertical line). The probability of obtaining a t statistic more discrepant with the null hypothesis of no difference between the two groups is indicated by the red shaded region: 0.17. This probability is called the *p value*. A lower p value represents a *more discrepant* observation from the tested hypothesis.

Because the p value is calculable from any test statistic with a known sampling under the null hypothesis, it serves as a convenient summary of the evidence for any design. Although the test statistic and sampling distribution might differ from design to design and model to model, the interpretation of the p value remains the same: it is the probability of test statistic as deviant as the one found, assuming the null hypothesis were true.

The definition of a p value suggests how a frequentist might use it within a statistical test of the hypothesis. In a probabilistic approximation to the *modus tollens*

[2] The relevance of the pre-planning of the sample size is that a different sampling distribution for t will be obtained if one, say, continues to sample until a particular t statistic is reached. This important sensitivity to the stopping rule is explained later in this section.

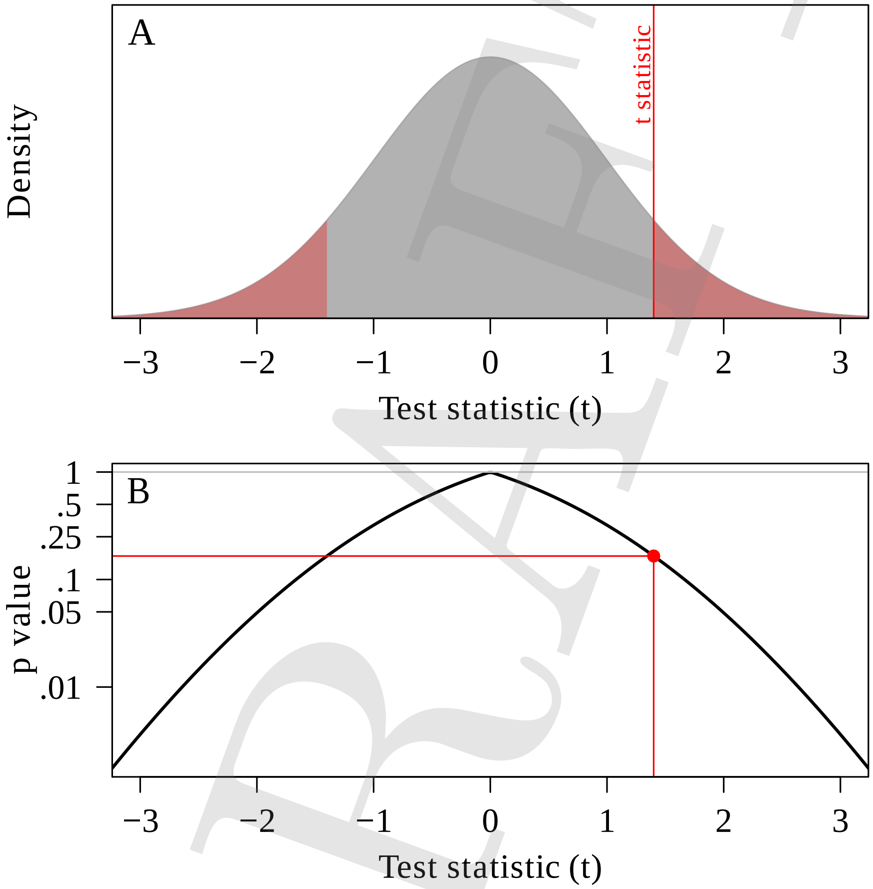

Figure 4: (A) Student's t distribution with 78 degrees of freedom. The vertical red line represents a hypothetical observed *t* statistic, for which the (two-tailed) *p* value is $p = .17$ (red shaded area). (B) The *p* values for *t* statistics between -3 and 3 for a Student's *t* distribution with 78 degrees of freedom.

from propositional logic[3], we can say that if the null hypothesis were true, we would very likely (with probability .95) have observed less evidence against the null, but we did not. The high probability of less extreme evidence under the hypothesis is what warrants the rejection of the hypothesis (see also Mayo & Cox, 2006). We might also choose a criterion α and reject a hypothesis — or call the effect "statistically significant" — whenever $p \leq \alpha$, then we will be in error in α of the cases when the hypothesis is true. This error rate is called the test's *Type I error rate*; alternatively, α is called the *level* of the test.

The result of a low p value is the tentative rejection of the hypothesis being tested. Fisher (1966) made clear that demonstrating an effect required *experimental control* of statistical significance: that is, "we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (p. 14). The result of a nonsignificant result, however, is *not* an affirmation of the hypothesis. In case of a failure to find a significant effect, we can infer that either our design was not sufficiently sensitive to whatever effect exists, or that the null hypothesis is true. This is often called a tentative *retention* of the null hypothesis, rather than an "acceptance" of it.

The inadequacy of focusing solely on the Type I error rate, however, can be seen by asking the question *what test statistic should we use to compute the p value?* Why not, for instance, take the *least discrepant* 5% of test statistics as evidence to reject a hypothesis? We would be in error in exactly the same proportion of experiments. To a scientist, this would be absurd: a scientist would choose a test statistic precisely for the property that greater test statistics would lead to greater evidence against the hypothesis, and hence rejecting on the basis of less discrepant observations makes little sense. Consistent with this view, Senn (2001) suggests that it was Fisher's view that test statistics were a sort of primitive on which a statistical test was built.

Neyman (1952), however, pointed out one way of eliminating the (seeming) arbitrary choice of the test statistic: consideration of the *power* of the test, a measure of its sensitivity to deviations from the hypothesis to be tested.

Sensitivity and statistical power

When designing an experiment to test for deviations from a hypothesis, we would often like to know how sensitive a test is for this purpose: would we often find discrepant results, if the hypothesis were, indeed, false?

Sensitivity is closely related to the value of an experiment. Fisher (1966) said that "the value of the experiment is increased whenever it permits the null hypothesis to be readily disproved [if it is false]" (p. 22) . Neyman (1977) said that "Obviously, an experiment designed [with low power] is not worth performing" (p. 107). Although Fisher did not formally define sensitivity, Neyman and Pearson (1933) did formally define power and showed that certain test statistics could be preferred over others because they yield tests that have a higher probability of rejecting a hypothesis if it is false.

[3] *Modus tollens is the Latin name for the logic reflected in: A implies B, but B is false; therefore, A cannot be true.*

In order to define power, we choose a way of quantifying how the null hypothesis might be false. One might object that in testing a hypothesis, one does not necessarily know *how* it will be false (Senn, 2001); however, in many cases we can define a *effect size* measure that will quantify how a null hypothesis might be false on a continuum of interest: for instance, in the two-group case, we might define a standardized true difference between the means,

$$\delta = \frac{\mu_y - \mu_x}{\sigma},$$

where σ is the (assumed common) standard deviation of the two populations of interest. When $\delta = 0$, the means are the same; hence, null hypothesis we tested with a p value is *nested* within the space of all possible hypotheses about δ. Notice, however, that committing to an effect size measure means committing to one of an infinite number of ways that the null hypothesis might be false.

Once the effect size measure has been chosen and design have been defined, one can compute the probability that a particular test — e.g., "reject the hypothesis $H_0 : \delta = 0$ when $p < \alpha$" — rejects the null hypothesis in the presence of a true effect, $\delta \neq 0$. In order to compute this probability, one must know the sampling distribution of the test statistic under hypotheses other than the null hypothesis $\delta = 0$, which can be determined analytically or by simulation. This probability is called the power. In formal notation, the power as a function of the effect size δ is:

$$\text{power}(\delta_1) = \Pr(p < \alpha; \delta = \delta_1).$$

Note that this is not a single probability, but a curve of probabilities for hypothetical effect sizes, called a *power curve* (e.g., Figure 5B).

Figure 5A shows how the sampling distribution of the t statistic in the two-group design changes as a function of δ: the t statistics become larger, on average. The hatched region $t \geq 1.66$ shows the values of the t statistic that would lead to a rejection of the one-sided hypothesis $\delta \leq 0$. The probability of rejecting this hypothesis when it is true must be less than 0.05, since the hatched area under the sampling distribution of t when $\delta = 0$ is 0.05. Thus, the Type I error rate of the test is $\alpha = 0.05$. Figure 5B shows how the power varies with δ.

When $\delta > 0$, however, the probability of (correctly) rejecting the hypothesis grows. When $\delta = .25$ — that is, the difference between the two populations is one-quarter of the population standard deviation — the power is about 0.3 or 30%. Given that we are more likely to "miss" the effect than detect it, we would say that the power to detect the effect $\delta = 0.25$ is quite low. If we were interested in effects as small as this, we would need to increase the sample size. When $\delta = 0.75$, the power is over 95%; if we were interested only in effect sizes as large as $\delta = 0.75$, we would likely be happy with the high sensitivity of our design with 40 participants per group. 80% power is often considered the target"acceptable" power.

In practice, it is important to state a minimum effect size of interest — an effect size that one would not like to miss, if it is there — and power the design for this effect size (Senn, 2007). Insensitive designs lead to noisy experiments that have very little *a priori* ability to detect effects and will tend to yield imprecise estimates of the

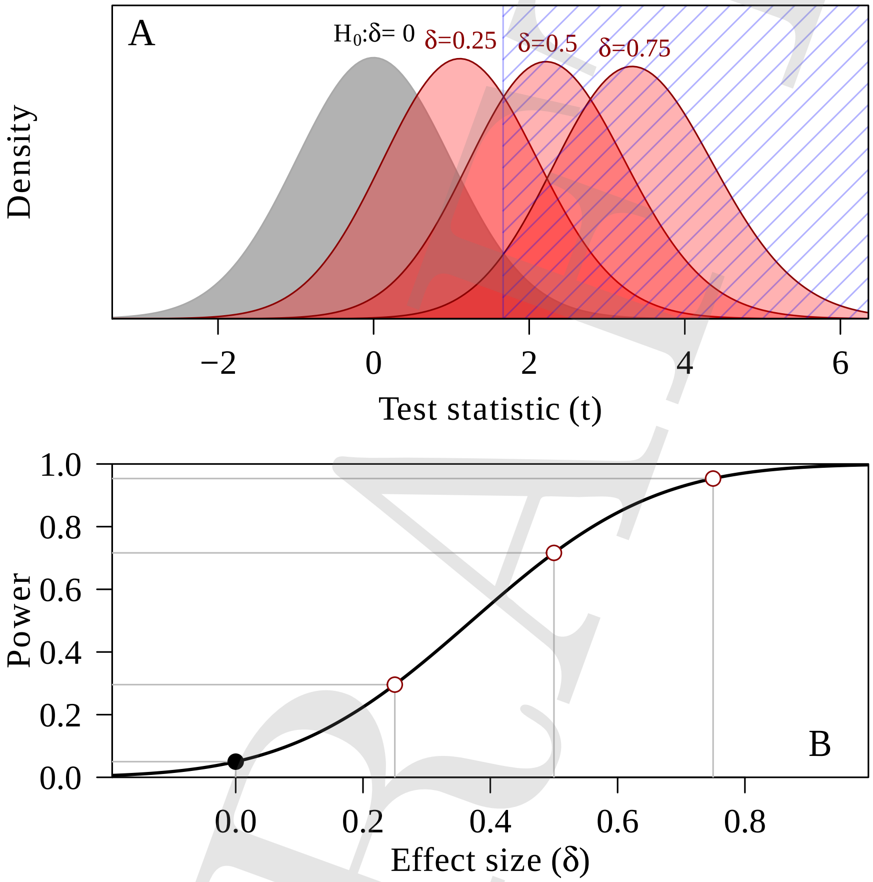

Figure 5: (A) Sampling distributions of the t test statistic for various true effect sizes. The critical t statistic for the one-sided test to reject $\delta \leq 0$ is shown as the vertical line; differences that would lead to a rejection of that hypothesis lie to the right of that line. The power of the test for the effect sizes shown is equal to the area of the sampling distribution to the right of the critical t value. (B) The power of the test as the effect size varies. Points shown represent power to the alternatives shown in panel A.

effect size of interest.

Demonstration

We can easily compute the power curve for a two-sample design under the assumptions of the t test in R. The sampling distribution of the t statistic when $\delta = \delta_1$ will have a noncentral t distribution with noncentrality parameter

$$\lambda = \delta \times \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

and degrees of freedom $n_x + n_y - 2$. The noncentral t distribution is built into R. The final figure is a version of Figure 5B.

```
n_x = 40
n_y = 40
delta = seq(0, 2,
            len = 100)
df = n_x + n_y - 2
ncp = delta *
  sqrt(n_x*n_y/(n_x+n_y))

# One-sided alpha=.05
alpha = 0.05
crit.t = qt(1 - alpha,
            df = df)

pow = 1 - pt(crit.t,
             df = df,
             ncp = ncp)

# plot the power
plot(delta, pow,
     ty='l',
     ylim = c(0,1))
```

Confidence Procedures

Neyman (1937) devised a method of parameter estimation that relies strongly on the logic of hypothesis testing developed above. We previously developed point estimates of a parameter, but point estimates do not account for the uncertainty in the estimate. Suppose we are interested in devising an “interval estimate” for a parameter, say, δ, defined above.

One way to proceed once data have been collected is to determine all hypotheses $\delta = \delta_1$ that would *not* be rejected by a certain test with Type I error rate α. This procedure will lead to a set of hypotheses that, on repeated sampling, will contain the true value $100(1-\alpha)\%$ of the time. This rate is called the *confidence coefficient* of the procedure. In many cases, the set of non-rejected hypotheses will form an interval, and an observed interval is called a *confidence interval*.

As an example, consider a α-level statistical t test statistic for our two-group design, which rejects when

$$t = \frac{d_{\bar{x}\bar{y}} - d_0}{s_d} \geq t^*$$

where d_0 is a hypothesized true difference, t^* is the appropriate two-sided test statistic for testing $\mu_x - \mu_y = d_0$ at level α. This test will fail to reject whenever d_0 is in the interval

$$d_{\bar{x}\bar{y}} \pm t^* s_d$$

and hence this interval represents a $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$.

A good confidence interval will contain false values of the parameter at lower rates than worse confidence intervals (Neyman, 1952). This property of confidence intervals is analogous to the corresponding test's power.

Relevance of stopping rules

Consider the problem of testing whether a participant is performing at chance in a two-alternative forced-choice (2AFC) task, such as the those used in studies of subliminal perception (e.g., Rouder, Morey, Speckman, & Pratte, 2007) and extrasensory perception (e.g., Bem, 2011). Suppose participants observe a screen onto which a digit from the set $\{2,3,4,6,7,8\}$ is very briefly flashed, and they must respond either "less than five" or "greater than five." If the number is flashed briefly enough, the participant will not be able to respond correctly at a rate better than chance, or 50% correct; the question of interest is whether responding is better than chance or not.

We assume that the trials are independent and all with equal probability of being correct, which we denote θ. Our null hypothesis is $\mathscr{H}_0 : \theta = 0.5$. In testing this null hypothesis, we will perform multiple trials. We denote the outcome of trial i as x_i, and each x_i has probability mass function

$$p(x_i;\theta) = \begin{cases} 1-\theta & x_i = 0 \\ \theta & x_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let n denote the total number of trials attempted by the participant. The relevant test statistic is then $y = \sum_{i=1}^{n} x_i$. Because y is the number of correct responses, larger values of y provide more evidence against $\mathscr{H}_0$.

There are multiple ways that we could perform an experiment to test $\mathscr{H}_0$. First — and most straightforwardly — we could set n before the experiment and conduct a fixed number of total trials. We might also decide to continue conducting trials until a fixed number of successes or failures, are reached. Under such a scheme, n is random and its distribution depends on θ. Suppose we learn from a fellow researcher that they performed $n = 25$ trials and observed $y = 17$ correct responses. What can we infer?

The frequentist inference will depend on the particular experimental scheme chosen. The distribution of y under the fixed-n scheme is shown in Figure 6 (top). For this scheme, one might have obtained any number of correct responses from 0 to 25. Our colleague observed $y = 17$. The shaded region in the figure highlights all observations that are at least as extreme as $y = 17$. The total probability of these observations is $p = 0.054$.

Under a scheme in which trials were run until a fixed number of failures were obtained — in this case, stopping after $n - y = 8$ failures — the distribution of y is

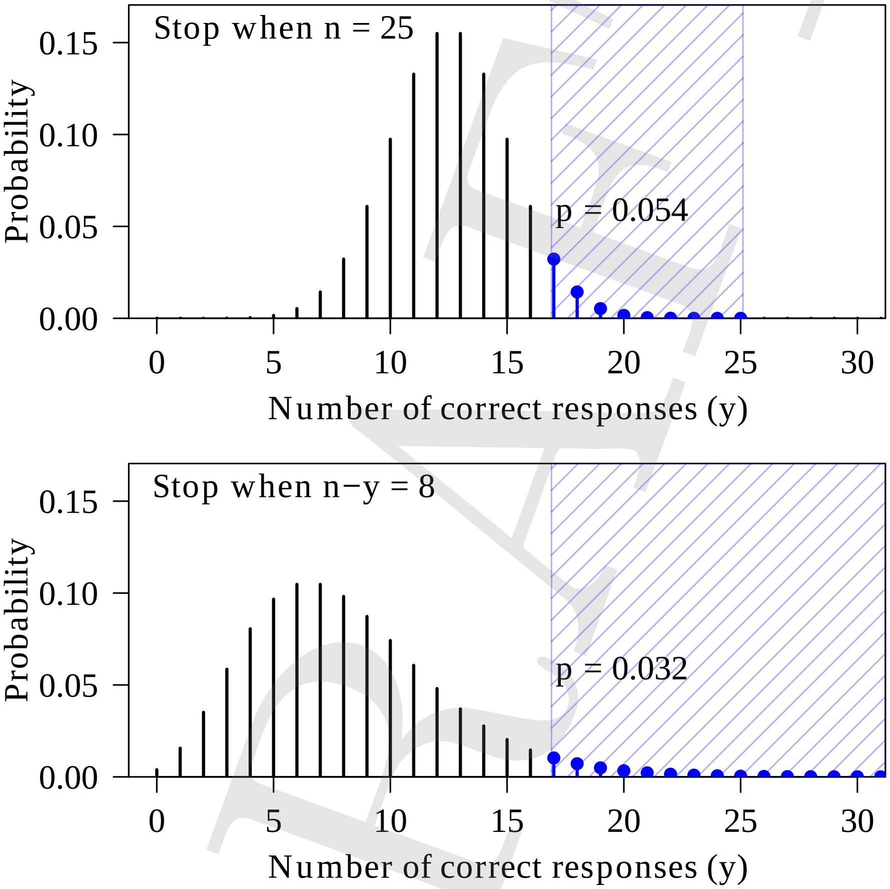

Figure 6: Distributions of the number of correct responses under two stopping rules. (A) Stop when the total number of trials $n = 25$. (B) Stop when the $n - y = 8$ incorrect responses are obtained. The blue shaded region represents the p value under the respective stopping rule when $y = 17$.

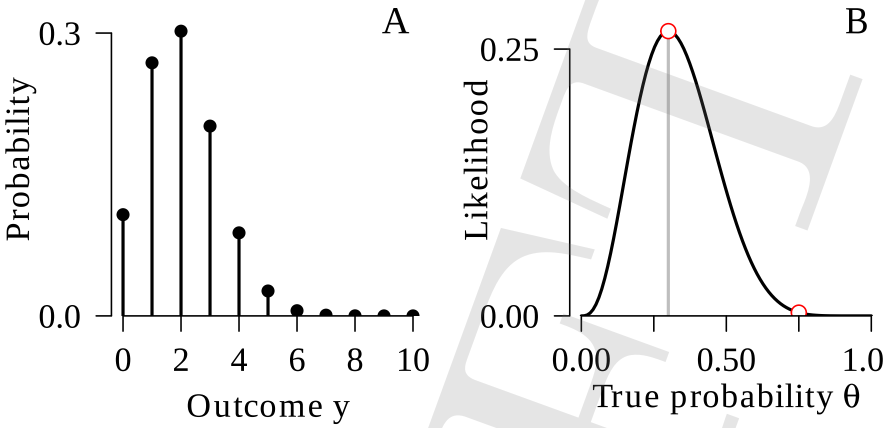

Figure 7: (A) The binomial probability density function $n = 10$ and $\theta = .2$. (B) The likelihood for the binomial parameter θ for $n = 10$ and $y = 3$. The red circles indicate the likelihood at $\theta = .3$ and $\theta = .75$.

different from that in the fixed-n scheme. Figure 6 (bottom) shows the distribution of y under a scheme in which the experiment was stopped after 8 failures. Notice that under this scheme, it is possible to obtain any number of successes ≥ 0. The shaded region in the figure highlights observations that are at least as extreme as $y = 17$. The total probability of these observations is $p = 0.032$.

The p values tell us that $y = 17$ is more extreme under the null hypothesis $\mathscr{H}_0$ under the fixed-failure scheme than under the fixed-n scheme. This is confirmed by a glance at the probability mass functions in Figure 6. Thus, in order to form a frequentist statistical inference, it is necessary to know the sampling scheme — including the stopping rule — under which the data were produced. Otherwise, it is possible to determine neither the data values that would be more extreme than the one observed, nor their probabilities. This dependence on the sampling scheme is a natural consequence of the focus on error probabilities in frequentist inference.

The likelihood approach

In the parametric inference situations discussed in this chapter, the goal is to make an inference regarding the parameter (θ, which may be a vector), having observed some data y. Our chosen parametric model specifies a family of probability distributions $p(y;\theta)$ that may have generated the observed data; we would like to determine, as closely as possible, the θ values that correspond to the probability distribution that actually generated the data.

The *likelihood* approach in statistical inference is characterized by a particular viewpoint: the evidence in the data regarding parameters can be found in the *likelihood function L*. The likelihood function is the probability mass or density function viewed as a function of unknown parameters for fixed, observed data. Consider, for instance, a binomial random variable and suppose we are interested in making an inference about θ. The probability mass function is

$$p(y;\theta,n) = \binom{n}{y}\theta^y(1-\theta)^{n-y},\ y = 0,\ldots,n.$$

If we have $n = 10$, we can use the probability mass function to determine the probability of possible outcomes $y = 0,\ldots,10$ for a given hypothetical true value of θ. Figure 7A shows the binomial distribution for $n = 10$ and $\theta = .2$, which is

$$p(y;\theta = .2, n = 10) = \binom{10}{y}.2^y.8^{10-y},\ y = 0,\ldots,n.$$

But suppose now that θ is unknown and we observe $y = 3$ out of $n = 10$. The likelihood function is

$$L(\theta;y = 3, n = 10) = \binom{10}{3}\theta^3(1-\theta)^7,\ \theta \in [0,1].$$

This likelihood function is shown in Figure 7B, and the box below shows how you can create the plots in Figure 7.

Demonstration

Plotting the probability mass function

```
n = 10
# Fixed theta
theta = .2
# vary y
y = 0:n
plot(y,
     dbinom(y, n, theta),
     ty='h')
```

Plotting the likelihood function

```
n = 10
# Fixed y
y = 3
# vary theta
theta = seq(0, 1, len = 100)
plot(theta,
     dbinom(y, n, theta),
     ty='l')
```

The so-called “law of likelihood” (A. Edwards, 1992; Hacking, 1965; Royall, 1997) summarizes the fundamental tenet of the likelihood approach.

The law of likelihood

“Within the framework of a statistical model, a particular set of data *supports* one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis.” (as expressed by A. Edwards, 1992, p. 30; emphasis in original)

Applying the law of likelihood to the example in Figure 7B, it is clear that $\theta = .3$ is better supported than $\theta = .75$ because the likelihood is greater for $\theta = .3$. If $\theta = .3$, the probability of $y = 3$ would be $p(y = 3; \theta = .3, n = 10) = 0.267$; if $\theta = .75$, the probability of $y = 3$ would be much lower, at $p(y = 3; \theta = .75, n = 10) = 0.003$.

The idea behind the law of likelihood is intuitive: if a particular hypothesis specifies that a datum has high probability, while another assigns the datum low probability, then the former is supported to a greater extent than the latter. This basic idea is several centuries old, going back to at least Bernoulli in the 18th century (Kendall, Bernoulli, Allen, & Euler, 1961), and formalized by Fisher (1935). To make the idea more concrete, we outline the simple example given by Bernoulli.

Consider an archer who is firing arrows at a target with a bull's eye at an unknown location (μ_x, μ_y). Suppose that the archer's shots are unbiased in both x and y directions the error of her shots have a standard deviation of 1 both directions, and the errors in each direction are independent. Further suppose that the errors are normally distributed. The bivariate probability distribution representing the archer's shots is shown in Figure 8A. The unknown center of the distribution of the archer's shots, (μ_x, μ_y) — also the location of the bull's eye — is at the center of the bivariate distribution.

Now suppose that we observed the $n = 10$ arrow holes shown in Figure 8B. What can we infer? It seems reasonable to think that the best supported value is somewhere in the middle of the points. The law of likelihood formalizes this idea by offering a specific way to assess the support: through the probability density of the observed data under a hypothesized (μ_x, μ_y). Consider the three possible probability distributions for the arrows (a, b, c) shown in Figure 8C. The true joint probability distribution of the data is

$$\begin{aligned} p\left((x_1, y_1), \ldots, (x_{10}, y_{10}); \mu_x, \mu_y\right) \quad &= \quad \prod_{i=1}^{10} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(x_i - \mu_x\right)\right\} \\ & \qquad \times (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(y_i - \mu_y\right)\right\} \end{aligned}$$

where (x_i, y_i) is the location of the ith arrow hole. The equation above simply the joint density function for the independent normal observations centered at (μ_x, μ_y) with $\sigma = 1$. If the distribution denoted a in Figure 8C were the true distribution, the probability density of the observed data would be 1.8×10^{-96}. This seems like a low number, but remember that the law of likelihood specifies that support is *relative*; we need to compare this very low number to the probability density under another alternative.

Suppose now that the true probability density function of the data were that shown as b in Figure 8C. The observed data would be substantially more probable than under a: the probability density under b is 1.102×10^{-42}. We can now estimate the relative support provided by the data for b versus a: it is the ratio $1.102 \times 10^{-42} / 1.8 \times 10^{-96} = 6.121 \times 10^{53}$. If distribution c represented the true distribution of arrow shots, then the probability density of the observed arrow shots would be 3.053×10^{-15}. Relative to distribution b, the support for c is $3.053 \times 10^{-15} / 1.102 \times 10^{-42} = 2.772 \times 10^{27}$.

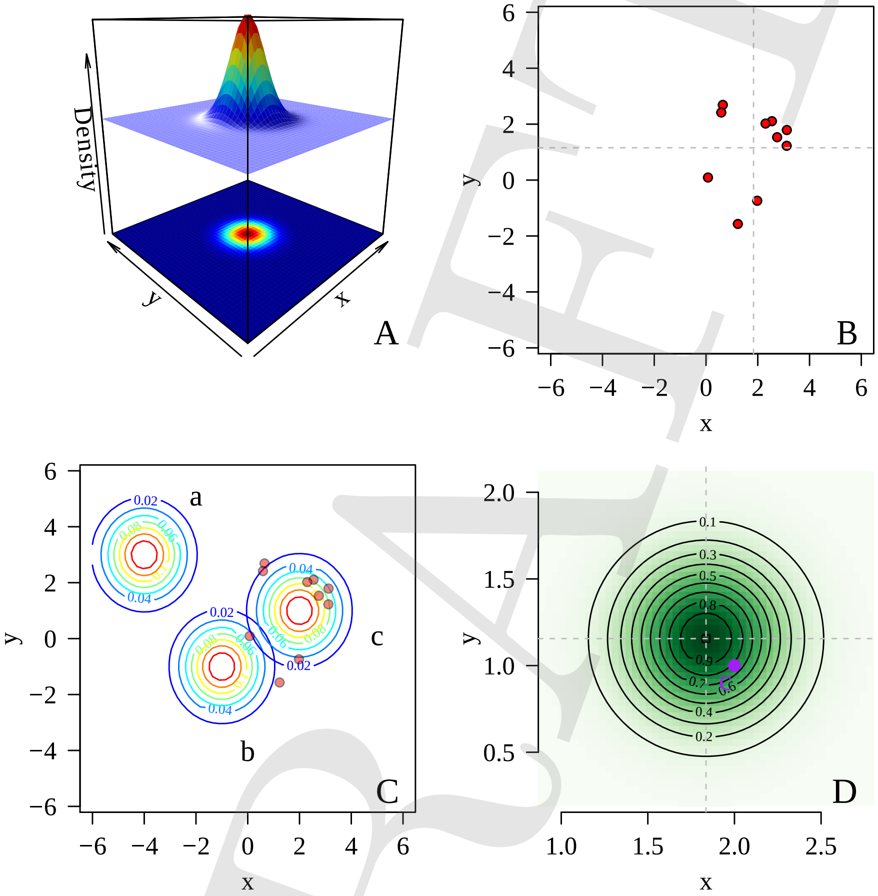

Figure 8: (A) The bivariate probability distribution function for the archer's shots in the x and y direction. The highest point is the bull's eye. (B) A set of n attempts by the archer, with the Bull's eye's location unknown to us. The vertical and horizontal dashed lines represent the mean location of the attempts in the x and y direction, respectively. (C) Three hypothetical probability distributions for the archer's shots. The center of each represents a potential Bull's eye location. (D) The likelihood function for the Bull's eye location, given the shots in panel B. The point represents likelihood for proposal c in panel C. The vertical and horizontal lines represent the maximum likelihood value in the x and y directions respectively.

Figure 8C shows the ratio of a range of locations in the vicinity of $(\bar{x}, \bar{y})$, which is in fact the location with the highest likelihood[4] (likelihood: 3.944×10^{-15}). The contours show the likelihood of the values along the contour relative to the likelihood for $(\bar{x}, \bar{y})$.

Parameter estimation

From the likelihoodist's perspective, the complete likelihood function contains the evidential support offered by the observed data for the parameters. However, the entire likelihood function will be unwieldy to work with, particularly with multidimensional parameters. It will therefore be profitable to reduce the the likelihood function to a summary representing a point estimate and some sort of indication of the precision or uncertainty with which the parameter is estimated.

The most obvious point estimate from the likelihood function is the *maximum likelihood estimator* (MLE). This is the value of the parameter that has the largest likelihood, which, in the likelihood paradigm, is the most strongly-supported value of the parameter. The MLE is typically found by either finding a unique solution to

$$\frac{\partial}{\partial\theta}\log L(\theta; y) \quad = \quad 0,$$

or numerically finding a maximum using a computer. Note that we typically use the natural logarithm of the likelihood function, because it behaves well (it is usually approximately quadratic), avoids numerically-problematic small numbers that are typical of probability densities, and has good theoretical properties which we will describe later. In most cases that applied researchers will encounter, the MLE will be unique, but it is important to note that the MLE may not be unique. A non-unique MLE often indicates a problem with the model, such as parameter non-identifiability.

The MLE has a number of properties that make it useful to statisticians from all schools of inference.[5]

- **Consistency.** MLEs will converge to the true value as sample size grows.
- **Asymptotic unbiasedness.** MLEs are often biased, but this bias will shrink to 0 as sample size grows.
- **Asymptotic normality.** The sampling distribution of the MLE $\hat{\theta}$ approaches

$$\hat{\theta} \sim \text{Normal}\left(\theta, I_n(\theta)^{-1}\right)$$

[4] One must be careful not to interpret the word "likelihood" in this context in the informal sense of "likely" as being "probable". Saying that parameters have the *highest likelihood* here means simply that these parameter values maximize the probability density of the observed data. A likelihoodist would then, by the law of likelihood, assign the highest support to these parameters.

[5] There are a number of weak regularity conditions that are needed for these properties of MLEs. We will not discuss these here, but rather refer the reader to a textbook on statistical inference, such as Casella and Berger (2002) or Rice (1998).

where the Fisher information matrix I_n is the frequentist expectation

$$I_n(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta; y_1, \ldots, y_n)\right)$$

as sample size grows (with a corresponding multivariate normal result for multiparameter model). The asymptotic precision $I_n(\theta)$ is particularly important because it establishes a relationship between the sampling distribution and the curvature of the logarithm of the likelihood function $\frac{\partial^2}{\partial\theta^2}\log L$.

- **Asymptotic efficiency.** The MLE achieves the lowest possible RMSE of any estimator as sample size grows.
- **Asymptotically quadratic log likelihood.** As the sample size grows, the logarithm of the likelihood function will approach a quadratic function. This is useful for approximating likelihood intervals.
- **Invariance to transformation.** If $\hat{\theta}$ is the MLE for θ and g some function of θ, then the MLE for $g(\theta)$ is $g(\hat{\theta})$. For instance, the MLE for the variance σ^2 of a normal random variable is $\sum(y_i - \bar{y})^2/n$. The MLE for the standard deviation $\sigma = \sqrt{\sigma^2}$ is $\sqrt{\sum(y_i - \bar{y})^2/n}$.
- **Relationship with Bayesian inference.** The MLE is the maximum *a posteriori* (as defined in the section on Bayesian inference) value for θ under a uniform prior.

Note that most of these properties are actually *not* about likelihoodist support *per se*, but are rather about frequentist properties of the MLE. A statistician may justify the use of the maximum likelihood estimator in several different ways, depending on their philosophical outlook.

Due to the likelihoodist focus on support, however, the likelihoodist is restricted to interpreting the likelihood function itself. In order to summarize the likelihood function we might view all likelihoods relative to the MLE's likelihood, and report intervals or areas within which all values have at least, say, 1/8 the support of the MLE.

Figure 9A,B shows this principle applied to the two examples so far in this section. Upon observing $y = 3$ out of $n = 10$ binomial observations, the MLE for the true probability θ is $\hat{\theta} = 3/10 = .3$. The likelihood is shown in Figure 9A, with the maximum indicated by a vertical dashed line. The horizontal line shows the likelihood at 1/8 of the maximum. The values of θ that have at least 1/8 the support of the maximum value are in the interval $(0.079, 0.619)$. We must be careful, however, not to interpret these values as necessarily well-supported. Likelihood is an inherently relative concept, which means that we cannot describe the MLE as being "supported" without reference to what it is supported *against*. Likewise, the values just inside the edge of this interval are supported only slightly more than those outside; the 1/8 likelihood interval serves only as a rough description of the likelihood, not as a hard boundary between supported and unsupported values. The valid likelihoodist interpretation of the likelihood interval is that there are no alternatives that are more than 8 times *better* supported than values in the interval.

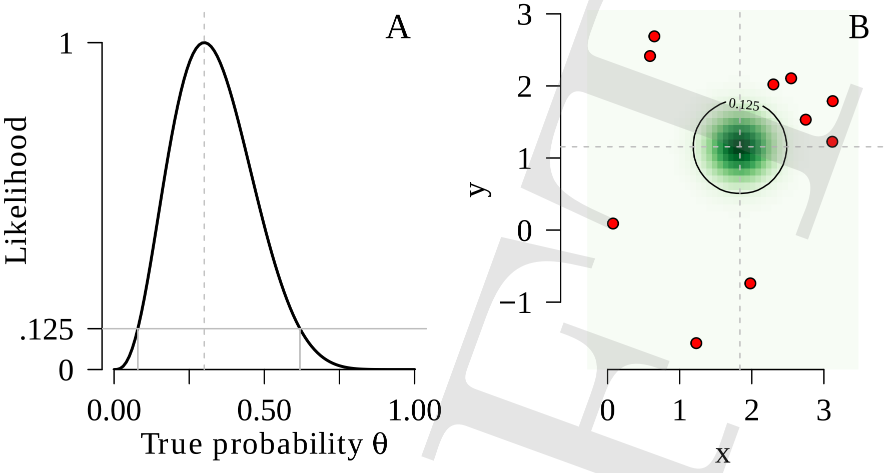

Figure 9: Likelihoods and 1/8 likelihood intervals for the binomial example (A) and the archer example (B).

Figure 9B shows the likelihood function for the archer example, along with the observed arrow holes. The cross formed by the horizontal and vertical dashed lines indicates the MLE at $(\bar{x}, \bar{y})$, and the circle indicates the contour where the likelihood is 1/8. The circle is centered at $(\bar{x}, \bar{y})$ and has radius 0.645.

A psychological example

We can demonstrate the usefulness of likelihood — and some of the limitations of the likelihoodist approach — through an example with a common psychological model, the *signal detection* model (Green & Swets, 1966, ; see also the chapter by Kellen and Klauer in this volume). Consider a task in which a participant must detect a stimulus, such as a faint tone, within a noisy background. If a participant hears the tone, they respond “tone present”; if they only hear noise, they respond “tone absent.” On some proportion of trials the tone is presented with the noise; on other trials, only the noise is presented.

If the stimulus is weak relative to the noise, participants will make errors. There are two kinds of errors: a participant can respond “tone present” when no stimulus was presented — an error called a “false alarm” — or a participant can respond “tone absent” when a tone was presented — an error called a “miss.” When a participant correctly identifies a stimulus, the response is called a “hit,” and when they correctly identify a noise-only trial, the response is called a “correct rejection.”

In signal detection theory, a stimulus is assumed to give rise to a continuous latent “strength” (denoted s_j for trial j) which then gives rise to the response. A trial eliciting a latent stimulus strength greater than a criterion c is judged by the participant to be tone-present trials; trials eliciting a weaker strength is judged to be tone-absent. The latent stimulus strengths s are assumed to be Normal random

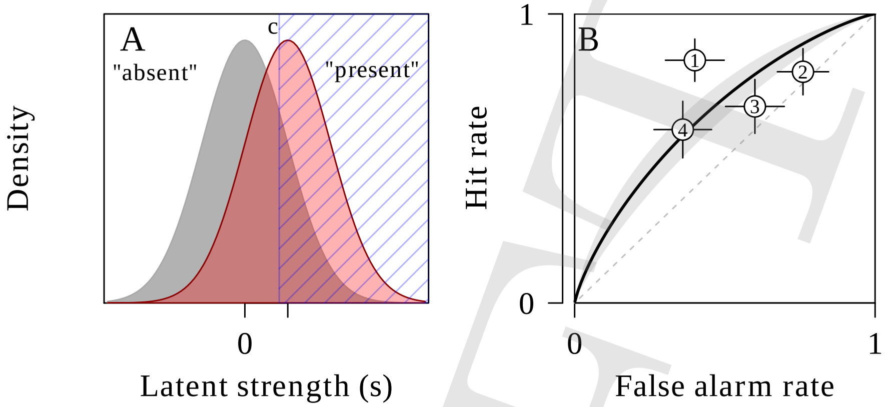

Figure 10: (A) A graphical depiction of the signal detection model, assuming $d' = 1$ and $c = .8$. (B) A receiver operating characteristic (ROC) plot showing the hypothetical data from four conditions (points with standard errors; see text), the predicted ROC curve for the maximum likelihood estimate of d' from the data (black line), and standard errors on the ROC-curve based on the Hessian matrix (gray shaded region).

variables:

$$s_j \sim \begin{cases} \text{Normal}(0,1) & \text{if trial } j \text{ is tone-absent,} \\ \text{Normal}(d',1) & \text{if trial } j \text{ is tone-present.} \end{cases}$$

The model is shown graphically in Figure 10.[6]

The probabilities of hits and false alarms can be derived using the cumulative distribution function of the Normal, Φ:

$$p_f = 1 - \Phi(c) \tag{2}$$
$$p_h = 1 - \Phi(c - d') \tag{3}$$

where p_f and p_h are the true probabilities of a false alarm and a hit, respectively. Eqs. 2 and 3 show the relationship between the model parameters and the probabilities of hits and false alarms for a single participant in a single condition; if we have experimental conditions that affect parameters d' or c, we would fit several of these parameters simultaneously, as we show now show.

Consider an experiment in which a participant performs signal detection trials in four different conditions. The four conditions differ only in the rewards or penalties we offer the participant. For a correct response, we always reward them with €0.50. For incorrect responses, we penalize them an amount that depends on the condition and error type. In conditions 1-4 we penalize them €0.10, €0.30, €0.50,

[6] The signal detection model is conceptually close to frequentist hypothesis testing. This is not an accident; Wald's statistical decision theory (Wald, 1939) inspired the development of signal detection theory (Swets, Tanner, & Birdsall, 1961). The signal detection model also has a close relationship to the ordered probit model for ordinal data.

€0.70 for false alarms, respectively, and €0.70, €0.50, €0.30, €0.10 for misses. Notice that this reward structure rewards liberal responding in condition 1 because the cost of a false alarm is low, but rewards conservative responding in condition 4. This would be expected to change the participants' criteria for responding.

In order to estimate the effect of the experimental manipulation, we build a signal detection model with one d' parameter — because the manipulation is simply of the reward, and should have no effect on the sensitivity of the participant to the stimulus — and four criteria $c_j,\ j = 1,\ldots,J$, one for each condition. Hypothetical data for a single participant are shown in Table 1 and (as proportions) in Figure 10B.

	Condition							
Response	tone-present				tone-absent			
	1	2	3	4	1	2	3	4
"tone-present"	14	17	8	6	8	7	11	2
"tone-absent"	11	8	17	19	17	18	14	23
Total	25	25	25	25	25	25	25	25

Table 1: Hypothetical frequency data for the signal detection example. Rows indicate responses; columns represent conditions.

In order to perform likelihood estimation, we need to define the likelihood. This is simply the product of the individual likelihood functions for the eight conditions (four reward conditions × two stimulus presentation conditions, present or absent):

$$\begin{aligned} L(d', c_1, \ldots, c_4; h_1, \ldots, h_4, f_1, \ldots, f_4) \quad &= \quad \prod_{j=1}^{4} \left[\binom{n}{h_j} \left(1 - \Phi(c_j - d')\right)^{h_j} \Phi(c_j - d')^{n-h_j} \right. \\ &\qquad \left. \times \binom{n}{f_j} \left(1 - \Phi(c_j)\right)^{f_j} \Phi(c_j)^{n-f_j} \right] \end{aligned} \tag{4}$$

where h_j and f_j are the observed numbers of hits and false alarms in reward condition j, and n is the number of trials in each condition. Note that we could leave out the choose terms $\binom{n}{h_j}$ and $\binom{n}{f_j}$ without changing the analysis, because they are constant. They are left in to emphasize the fact that the likelihood arises from the binomial probability mass function.

It is obvious perusing the likelihood in Eq. 4 that it would be tedious to solve for the maximum likelihood estimate by hand. It often easier, faster, and less error-prone to use a computer to maximize the likelihood through *numerical optimization*, though in more complicated models numerical optimization can pose difficulties that are beyond the scope of this chapter. The code in the box below uses the `optim` function in R to find a maximum in the log likelihood function.[7]

Demonstration

The R code in the right-hand panel of this box fits the signal detection model described above. At `A`, the data are defined. The `log.L` function at `B` defines the logarithm of the likelihood function, which will be maximized.

Within `log.L`, the code at `C` maps the full parameter vector (1 d' parameter and 4 criteria) into variables for clarity. At `D`, true probabilities are computed from the model parameters (see Eqs. 2 and 3), and at `E` the function computes (and returns) the logarithm of the likelihood.

The variable `start` contains starting values for the optimization function `optim`, which finds the parameters that maximize the (log) likelihood `log.L`. The `fnscale=-1` control argument specifies that `optim` should maximize instead of minimizing. The results are stored in the variable `mle`.

For details about the meaning of specific arguments to `optim`, see R help file for the function (`?optim` at the R prompt).

```
# A
h.fq = c(16, 14, 15, 14)
f.fq = c(15, 12, 9, 8)
n = 25

# B
log.L =
function(pars, h, f, n){
 # C
 d = pars[1]
 c = pars[-1]
 # D
 p.f = 1-pnorm(c)
 p.h = 1-pnorm(c - d)
 # E
 sum(
  dbinom(h,n,p.h,log=TRUE) +
  dbinom(f,n,p.f,log=TRUE)
 )
}

start = c(d=1, c1=-.5,
          c2=0, c3=.5, c4=1)
mle = optim(par = start,
 fn = log.L,
 control = list(fnscale=-1),
 hessian = TRUE,
 h = h.fq,
 f = f.fq,
 n = n)
```

The two important values returned by the `optim` call above — now contained in the `mle` variable — are the maximum likelihood estimates of the parameters and the Hessian matrix, which gives an indication of the curvature of the likelihood function at the maximum (and hence the precision of the estimate; the Hessian is the negative

[7] Numerical optimization routines can sometimes settle on so-called "local" maxima, which are not actually the globally maximum value that the function can take. Doing the analysis several times using varied starting values can help reveal this problem. When there are many parameters to be optimized and the likelihood is ill-behaved, alternative approaches to optimization may be used; e.g., simulated annealing (Bélisle, 1992) or particle swarm optimization (Kennedy & Eberhart, 1995)). Because this is a relatively simple example, we will assume that the `optim` function has found the global maximum.

Table 2: The Hessian matrix H returned from the `optim` call in the signal detection example. This represents an estimate of the curvature of the log likelihood function at the found maximum.

H	d'	c_1	c_2	c_3	c_4
d'	−62.10	14.80	15.67	15.73	15.89
c_1	14.80	−30.57	0.00	0.00	0.00
c_2	15.67	0.00	−31.53	0.00	0.00
c_3	15.73	0.00	0.00	−31.19	0.00
c_4	15.89	0.00	0.00	0.00	−30.96

of the Fisher information matrix computed at the MLE). The MLE for d' is $\hat{d}' = 0.383$ and the MLEs for the four criteria are $\hat{c}_1 = -0.117$, $\hat{c}_2 = 0.141$, $\hat{c}_3 = 0.243$, and $\hat{c}_4 = 0.346$. Table 2 lists the Hessian matrix H evaluated at the maximum, for which the i,jth element e_{ij} is

$$e_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L$$

where θ_i is the ith parameter in the parameter vector. The Hessian is closely related to the asymptotic (co)variance of the MLE.

In order to assess the fit of the maximum likelihood estimates, we can use a receiver operating characteristic (ROC; Green & Swets, 1966) (Figure 10B), which plots the hits rates against the false alarm rates (both actual and predicted). The data are shown as points, and the solid black line shows the relationship between the hit rate and false alarm rate when d' is equal to the MLE. This curved line can be thought of as the line of "best fit" — from the likelihood perspective — under the signal detection model.

At this point some of the difficulties inherent in performing a likelihood analysis become apparent. Although the law of likelihood specifies that the likelihood function itself can be interpreted as the relative evidence for various parameter values, it is unclear how to use the law of likelihood to characterize the results in the signal detection example. The likelihood function is a curve in five dimensions and is not amenable to visualization. The analogue to the one-dimensional likelihood interval with five parameters is a curve in five-dimensional space. Many common statistical models have hundreds or thousands of parameters; using the likelihood for drawing an inference in these cases seems practically impossible.

In our example, the criteria parameters $c_1, \ldots, c_4$ might be considered nuisance parameters. We may wish to characterize the evidence for only the sensitivity parameter d', simplifying the problem to a single dimension. There is, however, no unique way of treating nuisance parameters in the likelihood paradigm. Here we summarize the various ways that Royall (1997) suggests to deal with nuisance parameters in multidimensional parameter spaces. For details see Royall (1997).

- **Reparameterize to orthogonal parameters.** Sometimes the likelihood can be factored into two parts, one part containing only the parameter of interest θ, and one containing the vector of nuisance parameters $\boldsymbol{\lambda}$:

$$L(\theta, \boldsymbol{\lambda}) = L_\theta(\theta) \times L_\lambda(\boldsymbol{\lambda}).$$

In this case, $L_\theta(\theta)$ can be treated as the likelihood function for θ by itself. If the likelihood function does not factor nicely into two parts, a reparametrization of the model might yield a likelihood function that does.

- **Marginal likelihood.** In cases where a statistic offers "most" of the information about the parameter of interest, and its distribution is not dependent on the nuisance parameters, one might use the distribution of the statistic to perform a likelihood analysis, rather than the full likelihood. For a normal population, for example, the marginal distribution of s^2 is a scaled-χ^2_{n-1} distribution free of μ. If we knew only the value of s^2, and not the individual sample values, we would simply use its marginal distribution to build a likelihood. To use marginal likelihood, we act as though we only knew s^2. Marginal likelihood throws away information in the sample pertaining to the nuisance parameters in order to obtain a simple expression for the likelihood of a parameter of interest.

- **Conditional likelihood.** It is sometimes possible to condition on an observed statistic to obtain a probability density that depends only on a parameter of interest. For instance, in the analysis of two binomial rates whose observed values are X and Y one might condition on the sum of the two observed rates $Z = X + Y$. The distribution $p(X \mid Z)$ depends only on the true odds ratio $\psi = p_X(1 - p_Y)/(p_y(1 - p_X))$. Although the two-dimensional joint likelihood of p_X and p_Y is unwieldy, the likelihood of ψ is one-dimensional, easier to understand, and includes the information of interest.

- **Estimated likelihood.** If none of the above can be used in the problem at hand — and for our signal detection theory example, none of them are appropriate — we might simply look at the likelihood function for the parameter of interest while setting the nuisance parameters to their maximum likelihood estimates. The resulting curve is called the "estimated likelihood." For the signal detection theory example, the estimated likelihood for d' is shown by the solid lines in Figure 11. The estimated likelihood can be thought of as a "slice" through the full likelihood function at the MLE for $c_1, \ldots, c_4$.

Using likelihood for frequentist inference

Intuitively, the likelihood function appears intimately connected with the evidence in the data; the law of likelihood axiomatizes this intuition, whereas in Bayesian statistics the connection between the likelihood and evidence arises from other principles, as we shall see.

As Neyman (1977) points out, likelihood does not have any special status in frequentist statistics, but likelihood may still serve as a convenient way of generating frequentist parameter estimates and tests. We have previously mentioned the

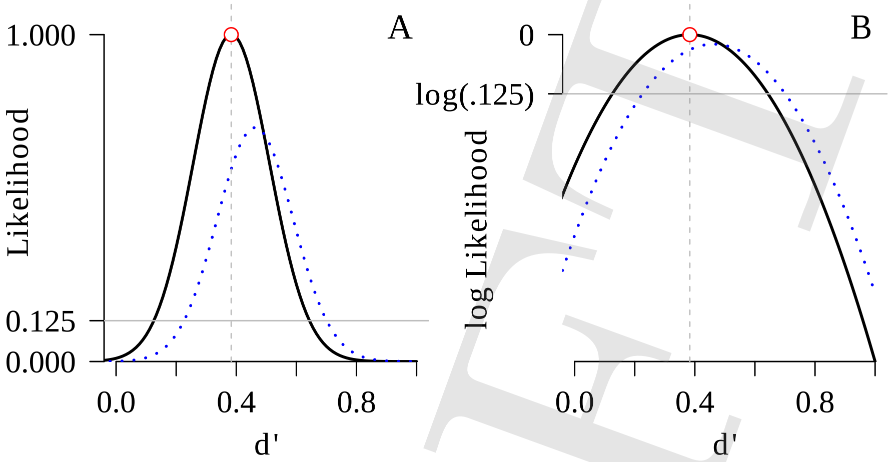

Figure 11: A: The estimated likelihood for d' (solid curve) and the likelihood curve assuming an alternate set of criteria (dashed curve). B: same as A, but showing the logarithm of the likelihood.

asymptotic normality of the MLE; we now use that fact to generate standard errors and confidence intervals.

We continue the signal detection example. We computed the Hessian matrix H, which is the negative of the observed Fisher observed information matrix I_n, computed at the MLE. Since we know that, asymptotically,

$$\hat{\theta} \sim \text{Normal}\left(\theta, I_n(\theta)^{-1}\right),$$

this suggests that we can obtain approximate standard errors by computing $-H^{-1}$, which will yield a matrix of covariances, and then taking the square root of the appropriate diagonal element. An approximate standard error for d', for instance, is $\sqrt{-(H^{-1})_{11}}$. Because the MLE is asymptotically normal, we can use the the $\alpha/2$ and $(1-\alpha)/2$ critical values from the normal distribution to build a $100(1-\alpha)\%$ confidence interval.

See the box below for R code to compute standard errors and confidence intervals for the signal detection example, which are shown in Figure 12.

Demonstration

We continue the signal detection here by computing standard errors for the parameters using the Hessian matrix computed by the `optim()` function.

We begin by computing the observed Fisher information, which is simply the negative of the Hessian.

We then invert this matrix to obtain an approximation for the asympotic covariance matrix for the MLE.

The diagonal elements of this covariance matrix provide the asymptotic variances; we take the square root to obtain standard errors.

```
# previous example continues

# Obs. Fisher info.
In = -mle$hessian

# approx. covariances
covMat = solve(In)

# approx SEs for
# all parameters
SErrs = sqrt(diag(covMat))

# 95% confidence intervals
CIlow = mle$par - 1.96*SErrs
CIupp = mle$par + 1.96*SErrs
```

A frequentist hypothesis test can also be performed using the likelihood. Suppose we are interested in testing a null hypothesis $\theta = \theta_0$ that corresponds to a model that is nested within a model for which we have maximum likelihood estimates.[8] Suppose the statistic D represents

$$D = -2\left(\log L(\hat{\theta}_0) - \log L(\hat{\theta})\right),$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are MLEs under the nested and nesting model, respectively. The statistic D captures the degree of the "lack of fit" induced by restricting to the nested model. A convenient approximation due to Wilks (1938) states that, asymptotically,

$$D \sim \chi^2_\nu$$

where the degrees of freedom ν is the difference in the number of parameters in the nesting and nested model.

In the signal detection model, we might wish to test whether the payoff manipulation was effective. If the payoff manipulation were ineffective, this would mean that $c_1 = c_2 = c_3 = c_4$; hence, there is only a single criterion, and all conditions have

[8] Although we are expressing the test in terms of a nested point null, the more general result holds for two composite hypotheses nested within a model. See Casella and Berger (2002) for theoretical details.

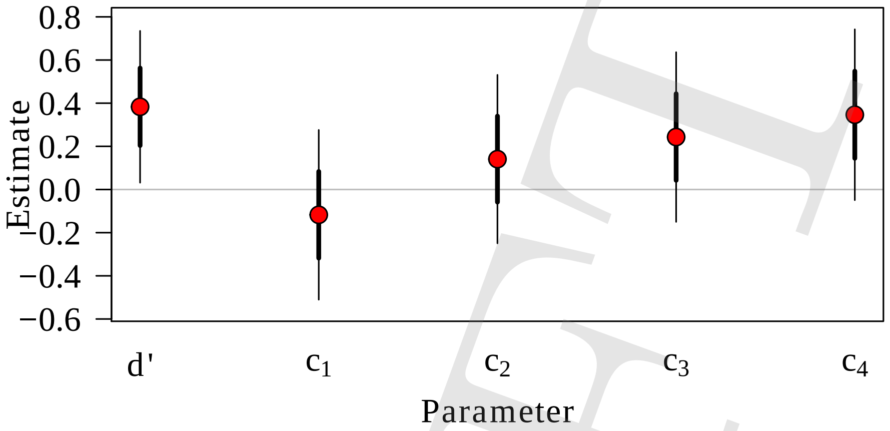

Figure 12: Estimates of the d' and the four criteria from the data in the example. Thick lines show approximate standard errors fom the Hessian; thin lines show approximate 95% confidence intervals.

identical true hit and false alarm rates. Under these conditions we combine the data from all conditions and find the MLEs for the parameters under the nested model, d'_0 and c_0. These maximum likelihood estimates will be

$$\hat{c}_0 = \Phi^{-1}\left(\frac{\sum_{i=1}^{4} f_i}{100}\right), \tag{5}$$

$$\hat{d}'_0 = \Phi^{-1}\left(\frac{\sum_{i=1}^{4} h_i}{100}\right) + c_0. \tag{6}$$

See Macmillan and Creelman (2005) for details. These estimates are $\hat{c}_0 = 0.15$ and $\hat{d}'_0 = 0.38$. If we set $c_1 = c_2 = c_3 = c_4 = c_0$, we can use the function `log.L` defined above to compute D, as shown in the box below.

Demonstration

We continue the signal detection here by testing the hypothesis that $c_1 = c_2 = c_3 = c_4$; that is, that the payoff manipulation was (in)effective.

```
# previous example continues

# Combined hit rate,
# FA rate, and n
h0.fq = sum(h.fq)
f0.fq = sum(f.fq)
n0= n*length(f.fq)

# MLEs assuming identical
# conditions
c0 = -qnorm(f0.fq/n0)
d0 = qnorm(h0.fq/n0) + c0

# log likelihood under
# nested model
pars = c(d0, rep(c0,4))
nested.L = log.L(pars,
               h = h.fq,
               f = f.fq,
               n = n)

# log likelihood under
# nesting model was
# already computed!
nesting.L = mle$value

# Wilk's theorem
D = -2*(nested.L -
          nesting.L)

# p value
p = 1 - pchisq(D,
               df = 5 - 2)
```

For the data in the example, $D = 3.678$. The probability of obtaining a more extreme value from a χ^2_3 distribution is $p = 0.298$, which would not lead a frequentist to reject the null hypothesis that the manipulation had no effect on the criteria, at the typical $\alpha = 0.05$. This is consistent with the likelihoodist estimates in Figure 12, which show only small differences between the estimates of the criteria relative to the uncertainty.

The likelihood principle

One may view likelihood inference in a variety of ways. Frequentists might view likelihood as a useful way of arriving at frequentist procedures. The properties of such procedures would still have to be tested for good frequentist properties, but if one is unsure how to derive an estimate or test, basing an estimate or test on the likelihood function would be a good place to start. Bayesian statisticians view likelihood as important for inference, but justify its use by more fundamental principles.

Some statisticians, however, take a much stronger stance: that the likelihood function offers *the only* approach to statistical evidence. This viewpoint is captured in the so-called "likelihood principle":

> **The likelihood principle**
>
> "All the information about [a parameter] θ obtainable from an experiment is contained in the likelihood function for θ given [data]. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another." (as expressed by J. O. Berger & Wolpert, 1988)

There are several reasons why one might accept such a strong principle. It arises naturally in Bayesian inference (see the next section). Birnbaum (1962) proved that it is a consequence of two seemingly uncontroversial frequentist principles. However, this proof has been called into question (see Evans, 2013; Mayo, 2014).

One of the consequences of the use of likelihood in general and of the likelihood principle more specifically is the *irrelevance* of stopping rules. Recall from the previous section on frequentist inference that different stopping rules imply different sampling distributions on a statistic, and hence different inferences, because the probability of obtaining a *more extreme test* statistic is important.

In contrast, for a likelihoodist only the likelihood function itself is important, and the likelihood function depends only on the probability (or density) of the observed test statistic, not on the probability of test statistics more extreme. Consider the example presented at the end of the previous section, concerning stopping rules in a 2AFC task. Suppose we ran one of the two designs and obtained 17 out of 25 successes. To a likelihoodist, the stopping rule is irrelevant, because the likelihoods are proportional, as shown in Figure 13. All of the methods for drawing inferences (e.g., the maximum likelihood, the Hessian) in this section will be identical.

This lack of dependence on the stopping rule chosen is seen by frequentists as a weakness, because it implies that procedures have no frequentist error guarantees. By likelihoodists, however, this lack of dependence is seen as a strength, because inferences will be invariant to seemingly irrelevant considerations like *why* an experimenter chose to end an experiment. For an accessible discussion from a likelihoodist perspective, see J. O. Berger and Berry (1988).

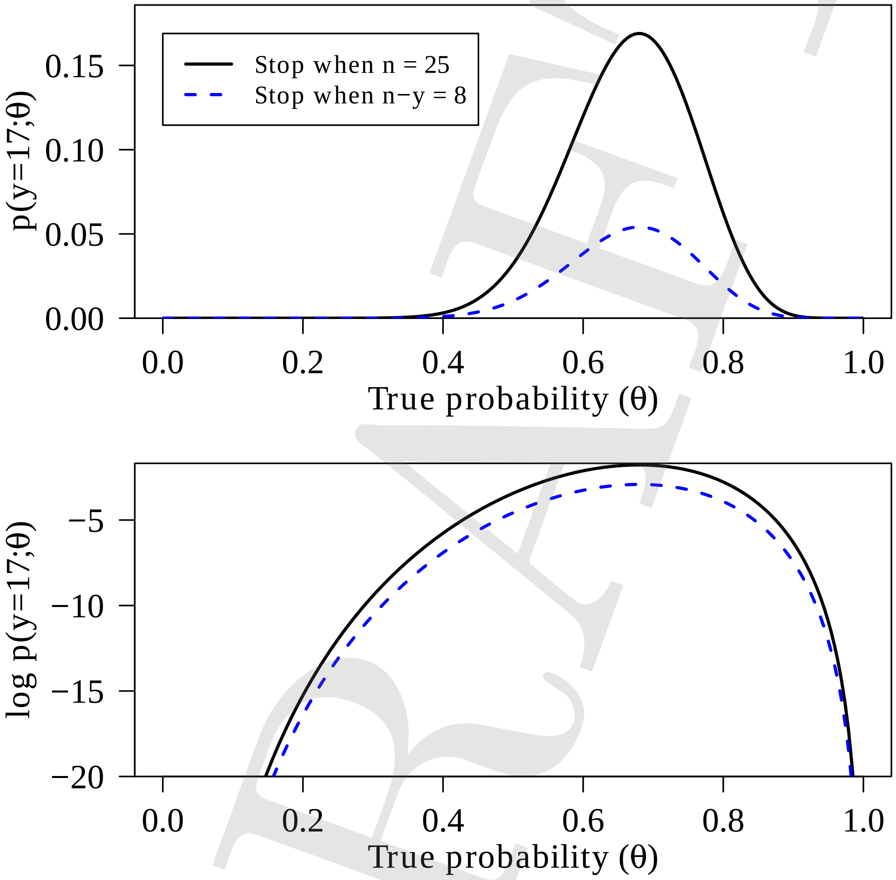

Figure 13: Likelihood (A) and log-likelihood (B) functions the two designs presented at the end of the frequentist inference section, assuming $Y = 17$ successes out of $N = 25$ attempts. Notice that the likelihood functions in panel A under the two sampling plans are proportional to one another, meaning that the inference is insensitive to the sampling plan.

Bayesian approaches

The Bayesian approach to statistical inference is concerned with drawing "rational" inference from a sample, where "rational" is defined by an application of Bayes' theorem to formalized beliefs about a parameter or other unknown quantity.

There are three basic foundational ideas underlying Bayesian statistics.

- **Numerical representation of belief/plausibility.** Belief about the plausibility of a proposition can be expressed as a probability. Propositions that are known to be false have probability 0; propositions that are known to be true have probability 1; we can express varying levels of uncertainty through probability values between 0 and 1.
- **Bayesian rationality/coherence.** Numerical belief/plausibility is modeled using probability theory, ensuring mutual coherence of statements of plausibility (de Finetti, 1937) and adherence with basic logical axioms (R. T. Cox, 1946).
- **Bayesian conditionalization.** The probability distribution representing the belief/plausibility *after* observing data y is the same as what we would have obtained through conditionalizing on y before having observed y. Bayesian conditionalization enables the *updating* of beliefs in a principled manner using known facts (e.g., the data).

Note that although the formalization of belief plays a central role in Bayesian statistics, these beliefs need not be the literal beliefs of any person. Like statistical "errors" in frequentist statistics, formalized beliefs are used as a way of drawing statistical inferences; an analyst's substantive beliefs are informed by statistical inferences but are not determined them. Indeed, human beliefs are not numerical at all, and need not be mutually consistent.

Bayes' theorem (general)

Bayes' theorem — a simple consequence of the definition of conditional probability — gives a relation between the unconditional probability of an event A and the probability of the same event, conditioned on B:

$$P(A \mid B) = \frac{P(B \mid A)}{P(B)} P(A).$$

Bayes' theorem (Bayesian statistics)

The form of Bayes theorem that serves as the foundation of Bayesian statistics gives a relation between the prior probability distribution of θ and the posterior probability distribution of θ conditioned on the data y:

$$p(\theta \mid y) = \frac{p(y \mid \theta)}{p(y)} p(\theta).$$

Because $p(y)$ is a constant that is not a function of θ, Bayes' theorem is often written as

$$p(\theta \mid y) \propto p(y \mid \theta) \times p(\theta),$$

or, "the posterior distribution is proportional to the likelihood times the prior."

When Bayes' theorem is combined with Bayesian probability (probability as belief) and Bayesian conditionalization (updating beliefs), the result is Bayesian statistics. Each of the terms in Bayes' theorem has a name.

- $p(\theta)$ is called the *prior* distribution. It expresses the uncertainty in the parameter θ before observing the data.
- $p(y \mid \theta)$ is called the *likelihood*. Although numerically the same as the classical likelihood, it is conceptually different. In Bayesian statistics the likelihood is a conditional probability distribution, because uncertainty in both y and θ are treated using (Bayesian) probability. Hence, they can be described as having a joint probability distribution.

  In frequentist and likelihoodist statistics, there is no such joint probability distribution, because uncertainty in θ is not treated using probability theory. The likelihood function is simply seen as a function of the parameter(s), not a conditional probability distribution.
- $p(y)$ is called the *marginal likelihood* if y has been observed and thus $p(y)$ simply serves as an unknown constant that normalizes the posterior distribution so that it integrates to 1. If y is unobserved, $p(y)$ is called the *prior predictive distribution*, because it is the distribution of the data marginalized — averaged over — the uncertainty in the prior. It can be expressed as an expectation:

  $$p(y) = E_{p(\theta)}\left[p(y \mid \theta)\right];$$

  that is, $p(y)$ is the average likelihood of the data y, when the average is taken with respect to the prior.
- $p(\theta \mid y)$ is the *posterior distribution*. It expresses the uncertainty in the parameter θ after seeing the data.

Bayes' theorem can be described as a method of using the data to express one's knowledge about a parameter, assuming a model and some sort of prior knowledge encapsulated in the prior distribution.

From prior to posterior

Consider the 2AFC example discussed in the Frequentist and Likelihood sections. We observe $n = 25$ independent Bernoulli trials in an ESP experiment, and we wish to estimate the true probability θ that a participant provides a correct response. A certain participant has provided a correct response in $y = 17$ of the 25 trials.

In Bayesian statistics, the posterior distribution for θ contains the information needed for the inferences about θ. In order to compute the posterior, we need to apply Bayes theorem, which combines the prior and the likelihood to obtain the posterior distribution. We begin with a prior distribution that assumes that any ESP effect will be small, and hence the true probability will be around θ. We defer discussion of the practicalities of choosing a prior until later; our goal here is to demonstrate the use of Bayes' theorem in Bayesian statistics, and hence we simply choose a convenient prior:

$$\begin{aligned} \theta &\sim \text{Beta}(10,10)\text{, that is,} \\ p(\theta) &= \frac{\theta^9(1-\theta)^9}{\text{Be}(10,10)}, \end{aligned}$$

where the function Be is the incomplete beta function (Abramowitz & Stegun, 1965) and merely serves to ensure that the distribution integrates to 1. The prior distribution is shown in Figure 14A.

Because the prior $p(\theta)$ is a probability distribution, we can use probability theory to say a number of things about θ *a priori*:

- **Prior mean/expectation.** The mean of the prior distribution is $\theta = 0.5$.

- **Prior mode, median, and quartiles.** The mode and median of the prior distribution is 0.5; the lower and upper quartiles are 0.424 and 0.576, respectively.

- **Prior variance/standard deviation.** The variance of the prior distribution is 0.012, and its standard deviation is 0.109.

- **Prior probability of parameter ranges.** The *a priori* probability that $p(\theta > .6)$ is 0.186; the *a priori* probability that $p(\theta < .25)$ is 0.009. Because the lower and upper quartiles are 0.424 and 0.576, respectively, $p(0.424 < \theta < 0.576) = .5$.

Each of these mutually statements about θ is a result of the choice of prior, aiding in its interpretation.

In order to obtain the posterior, we must multiply the prior by the term $p(y \mid \theta)/p(y)$. The numerator $p(y \mid \theta)$ is numerically identical with the likelihood function used in likelihoodist inference. The denominator $p(y)$ is defined as

$$p(y) = \int_{\Theta} p(y \mid \theta)\, p(\theta)\, d\theta,$$

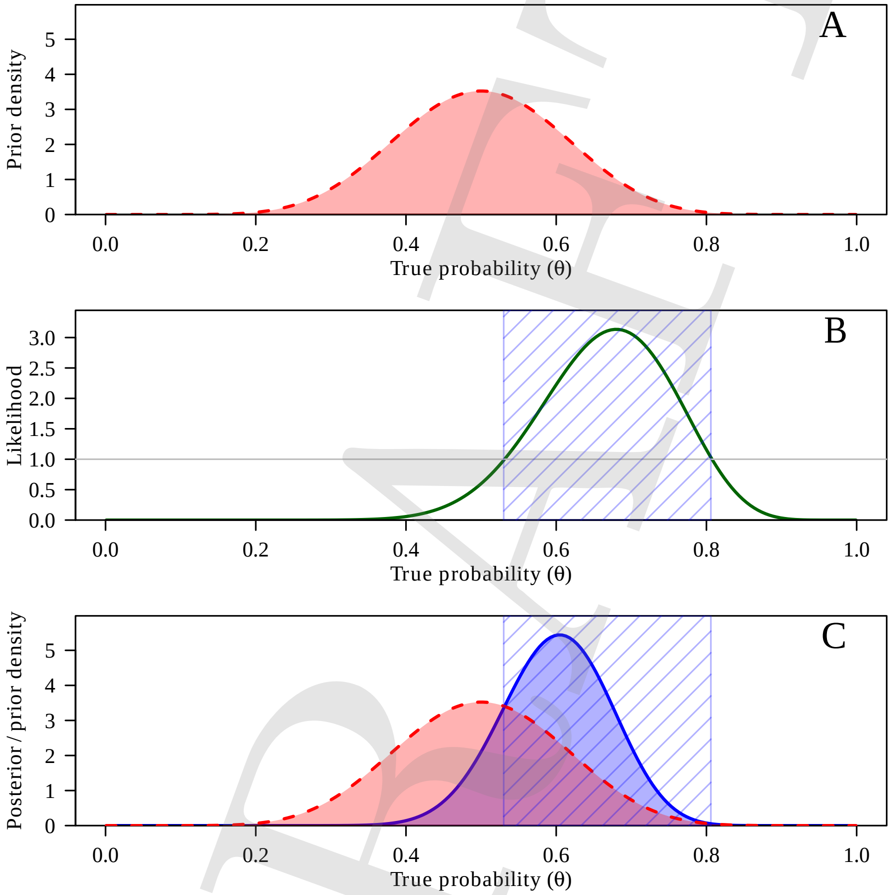

Figure 14: (A) A prior distribution for θ, the binomial probability parameter. (B) The evidence-scaled likelihood function, assuming the prior in panel A and observing 17 out of 25 successes. The shaded region represents parameter values for which the evidence-scaled likelihood is greater than 1. (C) The prior (red, dashed) and posterior (blue, solid). The shaded region represents parameter values for which the posterior density is higher than the prior density.

where Θ is the parameter space of θ. If y has not yet been observed and hence does not have a numerical value, this term is called the *prior predictive* distribution. The prior predictive distribution is the marginal distribution for the data, averaged across the prior. The prior predictive distribution represents the prior uncertainty about future, unobserved data. If y has been observed and hence y has a numerical value, the value of this term is the prior predictive density of the observed data, which is constant and is called the *marginal likelihood*. The marginal likelihood — also sometimes called the "evidence" for reasons that will be clear subsequently — represents the "average" likelihood of the observed data y under the prior $p(\theta)$.

The term $p(y \mid \theta)/p(y)$, which we will call the "evidence-scaled" likelihood, represents the impact of the data y on the Bayesian's beliefs. The evidence-scaled likelihood is represented in Figure 14B. This curve will be multiplied by the prior in Figure 14A to obtain the posterior; thus, the values of the evidence-scaled likelihood represent the degree of change in a particular region of the prior distribution. In the area around $\theta = 0.68$, the prior distribution will be increased by a factor of 3.1. In contrast, in the area around $\theta = .9$, the prior distribution will be decreased by a factor of 30. The function of the evidence-scaled likelihood is to "update" the prior distribution in light of the data.

Figure 14C (solid-outlined distribution) shows the resulting posterior distribution. Note that it is concentrated closer to the parameter value suggested by the data, $\theta = 17/25$. Also note that the posterior appears to be a sort of "compromise" between the data and the prior; the maximum value *a posteriori* (0.6) is between the maximum *a priori* value ($\theta = .5$) and the maximum likelihood estimate $\theta = 0.68$. The degree of the compromise depends on the relative concentrations of the prior distribution and the likelihood.

Demonstration

For a single parameter, posteriors are easy to compute with R. The following code will reproduce the the result in Figure 14.

After defining the prior parameters and data, we create a function pypt that computes $p(y \mid \theta)p(\theta)$ for any θ and fixed y and n. We then integrate this function to find $p(y)$, the marginal likelihood.

By Bayes' theorem, the posterior is $p(y \mid \theta)p(\theta)/p(y)$, which is plotted along with the prior distribution.

The second and final plot shows $p(y \mid \theta)/p(y)$, the evidence-scaled likelihood. Note the maximum at $y/n = 17/25 = .68$.

```
# Prior
a = 10; b = 10

# Data
y = 17; n = 25

# p(y|theta)p(theta)
pypt = function(theta){
  dbinom(y,n,theta) *
    dbeta(theta, a, b)
}

# p(y)
py = integrate(pypt,
      lower = 0,
      upper = 1)$value

# plot posterior
# p(y|theta)p(theta)/p(y)
thetas = seq(0,1,len=100)
plot(thetas,
     pypt(thetas)/py,
     ty='l',col="blue")

# Add prior
lines(thetas,
      dbeta(thetas,a,b),
      ty='l',col="red")

# Evid.-scaled likelihood
plot(thetas,
     dbinom(y,n,thetas)/py,
     ty='l',col="darkgreen")
abline(h=1)
```

Multiple parameters

Application of Bayes' theorem when there are more than one parameter is not conceptually more difficult than when there is only a single parameter. Consider an example where we would like to estimate the mean intelligent quotient of a particular population of children. We will assume the scores are normal, and that we wish to estimate the mean μ of the subpopulation as well as the variance, σ^2. If we obtain a simple random sample of n children from the subpopulation, $y_1, \ldots, y_n$, Bayes' theorem now takes the following form:

$$p(\mu, \sigma^2 \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \mu, \sigma^2)}{p(\boldsymbol{y})} p(\mu, \sigma^2)$$

where $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is the collection of all the children's IQ scores. The prior, likelihood, and posterior are now functions of both μ and σ^2. The *joint* prior and posterior, in particular, are bivariate probability distributions.

We will discuss this example in more detail later, but first, we consider how a Bayesian will proceed when faced with a multivariate posterior distribution. We could take the joint posterior itself as the inference. This will not be too difficult in the bivariate case; one can simply plot bivariate posteriors. Using the joint posterior distribution will be generally unsatisfactory for several reasons. First, the parameters are often not of equal interest. Many or most of the parameters in a model are nuisance. We are often interested in a single parameter by itself. Second, the joint posterior becomes very difficult to describe when it has more than two dimensions. Consider even a simple linear regression: there are three parameters (intercept, slope, and error variance). It is not clear how to describe the joint posterior in a way that is intuitive.

The typical approach in Bayesian statistics is *marginalization*: that is, averaging or integrating over the uncertainty in all but a single parameter of interest, and then focusing on the so-called marginal posterior for that parameter. In our IQ example,

$$p(\mu \mid \boldsymbol{y}) = \int_0^\infty p(\mu, \sigma^2 \mid \boldsymbol{y})\, d\sigma^2.$$

In simple cases this integration can be done analytically; in almost all practical analyses, the integration is performed by numerical methods such Markov chain Monte Carlo.

Informing the choice of prior

Within Bayesian statistics, there are a variety of ways of approaching the prior distribution, each distinguishing a somewhat different viewpoint within Bayesian statistics. In fact, one analyst might approach the problem of generating a prior from any one of several directions, depending on the specifics of the problem. In this way it is similar to the bias/efficiency tradeoff discussed in frequentist estimation. We note several approaches — neither mutually exclusive nor exhaustive — here.

- **Substantive prior information.** One of the oft-touted benefits to Bayesian statistics is the ability to use priors to inject information drawn from sources outside the data itself into the analysis. For instance, if one is estimating the effect of an experimental manipulation in a task that takes, on average, 500 ms to complete, it is highly unlikely that the manipulation will cause an average effect of more than 100 ms. A prior distribution might include this information, favoring effects of less than 100 ms much more than those greater than 100 ms. A prior generated with these sorts of concerns in mind is often called "informative."

- **Conjugacy/Computational tractability.** Computationally intractable, difficult, expensive, or very time-consuming analyses are of little use to a researcher who must solve a practical statistical inference problem. The choice of a prior distribution, or a family of prior distributions, must in some sense be guided by practical considerations. Use of conjugate families of priors, for instance, can simplify and speed an analysis dramatically through the use of known properties of the family of distributions.

- **"Objective" properties.** So-called "objective" priors, which include Jeffreys priors and reference priors, are chosen for their special properties, not for their correspondence to any actual degrees of belief. Jeffreys' rule (Jeffreys, 1946, 1961) for creating priors is a well-known rule for generating priors that are supposed to be "noninformative," due to the fact that applying the rule for any parameterization yields an equivalent prior (unlike, for instance, choosing a "flat" prior). Reference priors (J. O. Berger & Bernardo, 1992) are an improvement on Jeffreys' approach. Very often these priors are not proper probability distributions; that is, they do not integrate to 1.

- **Frequentist properties.** The choice of certain priors will lead to a numerical correspondence between Bayesian and frequentist inferences. Interestingly, these "frequentist-friendly" priors are often ones considered objective as well (Bayarri & Berger, 2004).

- **Correspondence to theory.** Prior choice can be dictated by theories under consideration, particularly when performing model comparison or model selection. For instance, if one theory predicts that an effect size should be negative, and another that it should be positive, then one might choose priors that attempt to quantify what would be predicted under the two models separately, then compare the quality of the model fit quantitatively. See also Vanpaemel (2010).

Further discussion about the practicalities of prior choice can be found in A. Gelman, Carlin, Stern, and Rubin (2004), Goldstein (2006), and Bernardo and Smith (2000).

Parameter estimation

In Bayesian parameter estimation, the posterior distribution is the target of inference, as it is an expression of the uncertainty in the parameters under the model assumptions and prior.[9]

[9] Andrew Gelman and Shalizi (2013) note that in practice, Bayesian statistical analysis often proceeds by iterative rounds of model fitting and parameter estimation. Nothing we say here is meant to contradict this; the posterior distribution can be used in a variety of ways.

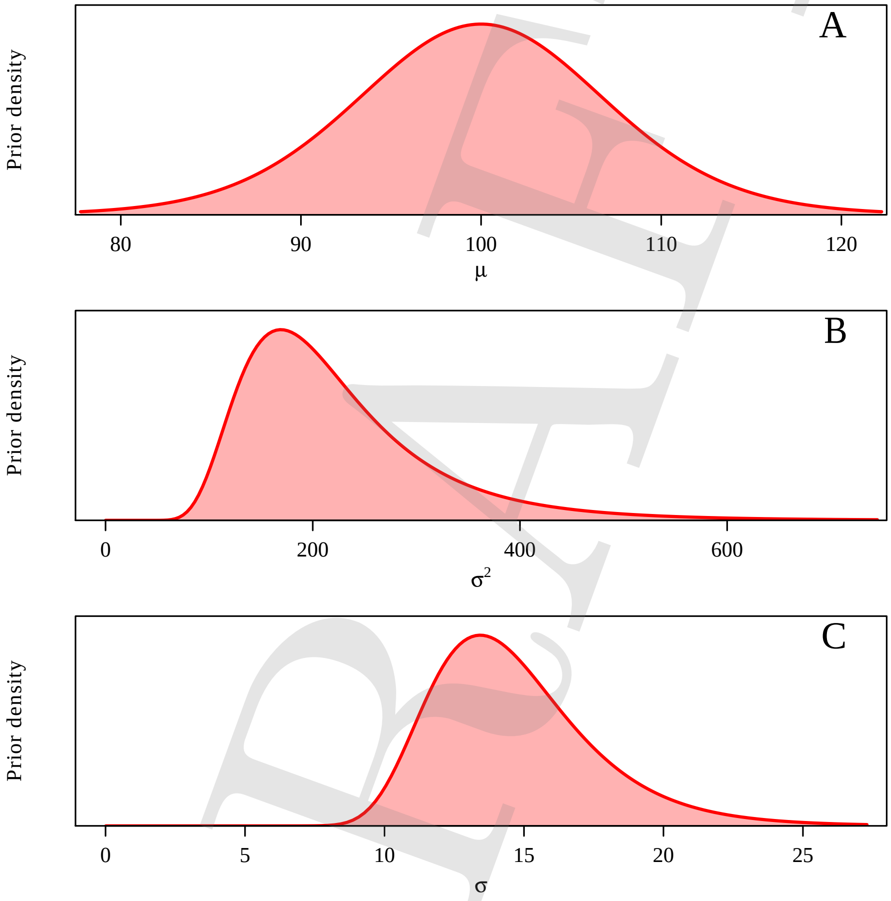

Figure 15: Marginal prior distributions for each of the parameters computed from prior p_1.

We now continue our IQ example from the previous section. The starting point of a Bayesian analysis is a prior distribution. We define several here so that the reader can compare the logic and the resulting inferences. To define an informed prior, consider that substantive information about IQ helps a great deal in this case. In the entire population, IQ is normed so that it has a mean of 100 and a standard deviation of 15; any particular sub-population may deviate from this for various reasons. Deviations are unlikely to be extreme, unless the subpopulation is abnormal in some way. For the sake of the example, we place the following informative priors on μ and σ^2

$$\begin{aligned} p_1(\mu,\sigma^2) &= p_1(\mu \mid \sigma^2)\, p_1(\sigma^2) \\ \mu \mid \sigma^2 &\sim \text{Normal}(100, \sigma^2/4) \\ \sigma^2 &\sim \text{Inverse Gamma}(7, 1364) \end{aligned}$$

The marginal informed priors are shown in Figure 15. The conditional prior on μ, which depends on σ^2, makes explicit the assumption that the uncertainty we have in the mean depends what we know about the variance of the population; if the population is highly variable, then μ is more free to deviate from its prior mean. The prior for σ^2 were chosen by manipulating the parameters of the inverse gamma distribution. This joint prior is proper, because it represents a valid joint probability distribution.

We can also choose an objective reference prior for this problem. Bernardo and Smith (2000) give the reference prior as

$$p_2(\mu,\sigma^2) \propto \left(\sigma^2\right)^{-1}.$$

Notice that this is not a proper probability distribution, because it does not integrate to 1. One can think of this prior as being completely flat over the parameterization $(\mu, \log\sigma^2)$. Because p_2 is not proper, we omit visualization of this prior.

We can also define third prior that is informed, but in a different way than p_1. Suppose that we change the parameters of p_1 such that the marginal prior for μ is different:

$$\mu \mid \sigma^2 \sim \text{Normal}(80, \sigma^2/4).$$

We have thus changed the marginal prior on μ to have a substantially lower mean, but the prior on σ^2 remains the same. We denote the resulting joint prior distribution p_3. This third prior may, for instance, be a way of incorporating the information that the population is intellectually disabled. The marginal prior for μ will appear as in Figure 15A, but shifted so that it is centered around 80.

Now that we have defined our three priors, suppose we collect a sample of $n = 100$ children from our subpopulation of interest. We then compute our summary statistics, which are $\bar{y} = 106$ and $s = 12$. A classical analysis might proceed by offering an estimate of the average IQ μ, with standard error, as 106 ± 1.2, or perhaps a 95% confidence interval of [103.62,108.38]. One might also desire to test the hypothesis that the mean is 100 using a t test, which yields $t_{99} = 5$, for a two-tailed $p < 0.001$.

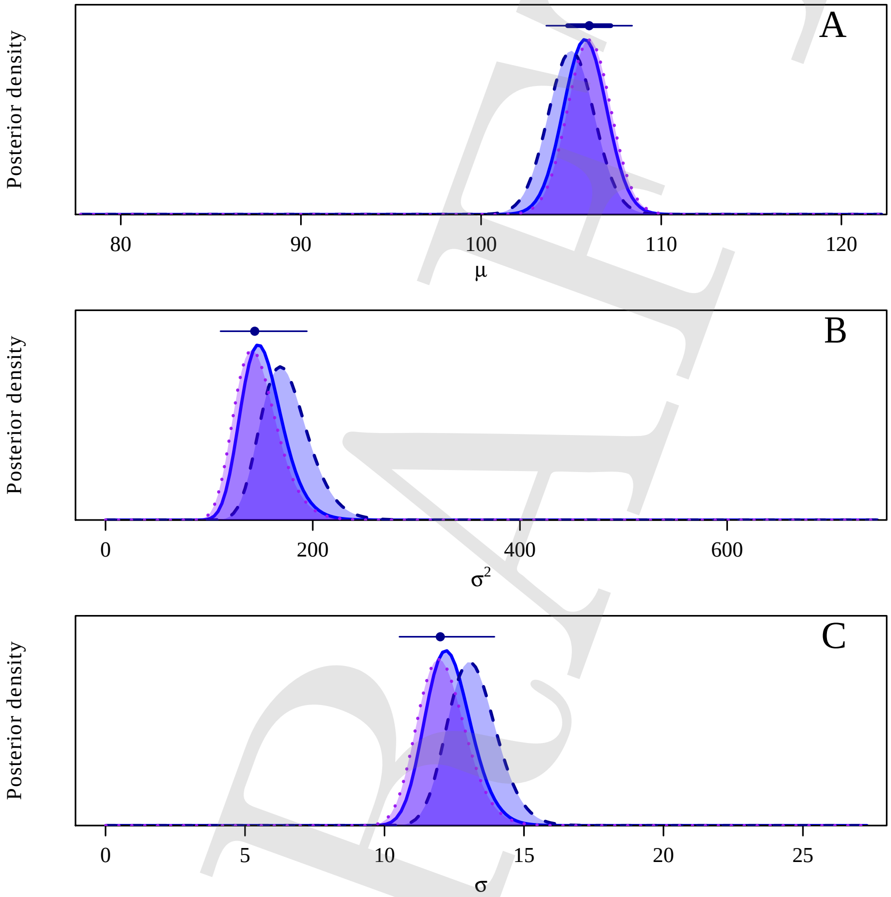

Figure 16: Three marginal posteriors for each parameter, obtained through the use three prior distributions described in the text. The classical point estimates (dot), 95% CIs (thin lines), and standard error (thick line) are also shown at the top of each graph. The solid, dotted, and dashed posterior distributions are from priors p_1, p_2 (reference prior), and p_3, respectively.

The three Bayesian posteriors corresponding to the three priors are shown in Figure 16. The marginal posterior distributions for μ are shown in the top panel, those for σ^2 in the middle panel, and the corresponding marginal posteriors for σ in the bottom panel.

Once we have obtained a posterior distribution, there are various ways of summarizing it.

- **Point estimate.** Common point estimates include the posterior mean, the posterior median, and the posterior mode. If we define the *posterior loss* $\ell(\theta, \theta_0)$, where θ_0 is some point estimate, we can then define the expected loss as

  $$E_{p(\theta|y)}\left[\ell(\theta,\theta_0)\right],$$

  where the expectation is taken with respect to the posterior distribution. The various point estimates will each minimize the expected posterior loss with respect to a different loss function; the posterior mean minimizes the expected loss with respect to the quadratic loss $(\theta - \theta_0)^2$; the posterior median, with respect to the absolute loss $|\theta - \theta_0|$; and the posterior mode with respect to the so-called 0-1 loss $I_{\theta=\theta_0}(\theta)$ where I is an indicator function equal to 1 when $\theta = \theta_0$, and 0 otherwise (Bernardo & Smith, 2000).

  The posterior mean is by far the most commonly-used point estimate, but also reporting the posterior median or mode may be useful for especially skewed posteriors. The marginal posterior means for μ under the three posteriors are 105.77, 106, and 105 respectively. Notice that the posterior mean for the reference prior is exactly $\bar{y}$, and that the other two priors are pulled slightly toward their respective prior means (100 and 80).

- **Spread.** Although the posterior distribution itself is the uncertainty in the estimate of the parameter, it sometimes helps to have a single number to summarize this uncertainty, rather than an entire distribution. Typical measures of uncertainty include the posterior standard deviation and the posterior variance. In many cases, the posterior standard deviation can be interpreted in much the same way as the standard error is interpreted in classical analysis, and will often be numerically similar.

  The marginal posterior standard deviations for μ under the three priors are 1.22, 1.21, and 1.31 respectively.

- **Interval estimates.** In Bayesian statistics, the dominant interval estimate is the *credible interval*, which is an interval whose collective posterior probability is X% of the posterior. A 95% credible interval will thus have a 95% posterior probability of containing the parameter. Note that there are typically uncountably many credible intervals, all mutually consistent. The most common credible intervals are the "central" credible interval, which has equal probability in each tail (e.g., 2.5% on the left and 2.5% on the right for a 95% central credible interval), and the "highest posterior density" (HPD) interval, which is the X% interval that contains only the highest posterior density values of θ. For approximately symmetric posteriors, the central and HPD intervals will be about the same.

The marginal posterior central 95% credible intervals under the three priors are $[103.39, 108.15]$, $[103.62, 108.38]$, and $[102.46, 107.54]$. Notice that under the reference prior, the credible interval is the same as the frequentist confidence interval. Note also that these credible intervals are also HPDs, because the marginal posteriors are symmetric.

Hypothesis testing

Hypothesis testing in Bayesian inference is typically performed in one of two closely-related ways: through the Bayes factor, or through posterior probabilities or odds.

Consider a set of possible *models* $\mathscr{M}_i$, $i = 1, \ldots, M$. A model in this context is the combination of a likelihood and a prior. Together, the likelihood and the prior must a proper prior predictive distribution $p(\mathbf{y})$; that is, all possible observations must have a definite, valid probability or density under the two models. This requirement will typically rule out the use of improper priors.

We can write Bayes' theorem as

$$p(\mathscr{M}_i \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathscr{M}_i)}{p(\mathbf{y})} p(\mathscr{M}_i),$$

which implies that

$$\frac{p(\mathscr{M}_i \mid \mathbf{y})}{p(\mathscr{M}_j \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid \mathscr{M}_i)}{p(\mathbf{y} \mid \mathscr{M}_j)} \times \frac{p(\mathscr{M}_i)}{p(\mathscr{M}_j)}.$$

Bayes factor

Given two models $\mathscr{M}_0$ and $\mathscr{M}_1$, the strength of the Bayesian evidence offered by the data $\mathbf{y}$ regarding these two models is the *Bayes factor*:

$$B_{01} = \frac{p(\mathbf{y} \mid \mathscr{M}_0)}{p(\mathbf{y} \mid \mathscr{M}_1)}.$$

The Bayes factor represents the (multiplicative) degree to which the odds for $\mathscr{M}_0$ against $\mathscr{M}_1$ change in response to the data.
The Bayes factor can be multiplied by *prior odds* to obtain *posterior odds*:

$$\frac{p(\mathscr{M}_0 \mid \mathbf{y})}{p(\mathscr{M}_1 \mid \mathbf{y})} = B_{01} \times \frac{p(\mathscr{M}_0)}{p(\mathscr{M}_1)}.$$

Properties of the Bayes factor include:

- **Representation of equivocal evidence.** If the Bayes factor is around 1, the data are as likely under one hypothesis as the other. The evidence is therefore equivocal; it does not favor either hypothesis.

- **Symmetry of hypotheses.** If a hypothesis can be represented as a model with a proper prior predictive distribution, it can be tested. Often commentators incorrectly confuse hypothesis testing using Bayes factors with testing point hypotheses. Bayes factors can be used with ranges of hypotheses as well, and the results will, necessarily, be consistent with those obtained using a parameter-estimation-focused approach.

- **Transitivity.** Because Bayes factors are ratios, for any three models i, j, and k, knowing any two of the pairwise Bayes factors will yield the third:

  $$B_{ij} = B_{ik}/B_{jk}.$$

  A related fact is that that $B_{01} = 1/B_{10}$.

- **Model/prior dependency.** Bayes factors are dependent on the prior distributions that help to define the models insomuch as they help to determine the marginal probability of the data $p(\mathbf{y})$.

- **Link between evidence and prediction.** The evidence between two models is precisely the ratio of the probability (or density) of the data under the two models. Models which predict that the observed data have high probability will have more evidential support than those that do not.

Example: the JZS Bayes factor t test

Consider again the problem of testing whether two normal populations have the same mean: that is, that $\mu_x - \mu_y = 0$ where μ_x and μ_y are the population means of the two populations, respectively. We will draw samples of size n_x and n_y from the two populations, we and take as the parameter of interest the effect size

$$\delta = \frac{\mu_x - \mu_y}{\sigma}$$

where σ is standard deviation of the two populations (assumed to be the same). When $\mu_x = \mu_y$, the standardized effect size will be $\delta = 0$. It will be convenient to summarize the evidence in the data with the classical t statistic:

$$t = d\sqrt{n_{eff}}$$

where the observed effect size d is

$$d \quad = \quad \frac{\bar{x} - \bar{y}}{s},$$

and the effective sample size is as defined in the section on frequentist inference.

The t statistic has a noncentral $t_{n_x+n_y-2}$ distribution with noncentrality parameter $\delta\sqrt{n_{eff}}$. Up to this point, the exposition has been consistent with frequentist or likelihoodist inference; the point of departure for a Bayesian hypothesis test using t is to place priors on the unknown parameter δ. The priors will correspond to various hypotheses of interest, and are combined with the likelihood to obtain a prior predictive distribution. A single combination of likelihood and prior are collectively referred to as a model.

The first model we consider is one corresponding to the null hypothesis that $\delta = 0$, depicted as the “spike” in Figure 17A. We call $\mathcal{M}_0$, and its marginal likelihood

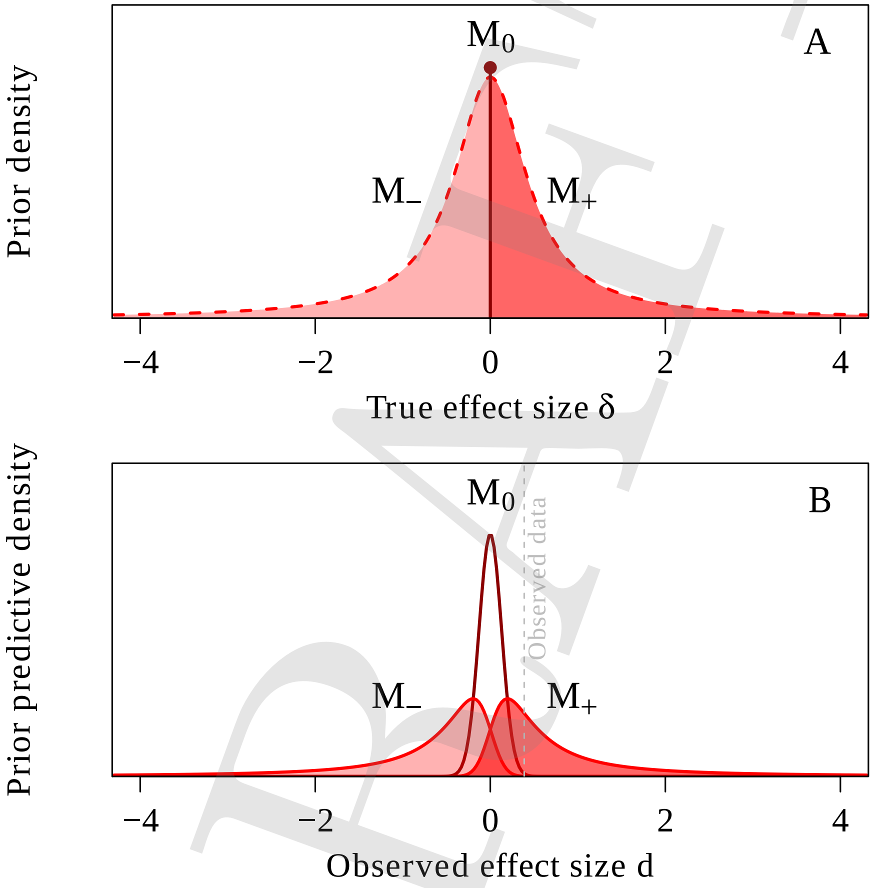

Figure 17: Top: Several hypotheses within a JZS t test. The null hypothesis, $\delta = 0$, is represented by the "spike". Bottom: The prior predictive distributions for the observed effect size d under the three hypotheses. The observed effect size in the example is represented by the vertical dashed line.

is simply the probability density of the observed t statistic from Student's t distribution, which a special case of the noncentral t distribution when the noncentrality parameter is 0.

For models corresponding to various alternatives, such as $\delta < 0$, $\delta > 0$, or $\delta \neq 0$, we must define prior distributions over δ. Rouder, Speckman, Sun, Morey, and Iverson (2009) and Morey and Rouder (2011) consider Cauchy priors (also known as Student's t_1 priors) priors on δ. We follow these authors and select the family of prior distributions that we can create by scaling a Cauchy distribution by a parameter r and then selecting ranges of hypothetical δ parameters. The shaded regions in Figure 17A show two models; to the left of 0 is the model $\mathscr{M}_-$, a negative half-Cauchy scaled by $r = 1/2$, and to the right of 0 is $\mathscr{M}_+$, a positive half-Cauchy scaled by $r = 1/2$. The models $\mathscr{M}_-$ and $\mathscr{M}_+$ are representations of one-sided hypotheses. The parameter r could be increased or decreased to express the expectation that the true effect size lies somewhere farther, or closer, to 0, respectively.

Suppose we obtained a sample of size $n = 25$ and observed $d = 0.39$. We can compare the hypotheses by comparing how probable the observed data are under the hypotheses, as shown in Figure 17B. The prior predictive density at $d = 0.39$ is 22 times higher under $\mathcal{M}_+$ than it is under $\mathcal{M}_0$. The Bayes factor is thus $B_{+0} = 22$.

Demonstration

We can compute the marginal likelihoods to compute the Bayes factor B_{+0} using R.

We first define the prior and data. The prior scale for the Cauchy distribution will be set at $r = 1/2$. We assume we obtain $n = 120$ observations from each population, and observe $t_{238} = 3, d = .387$.

Under the null model $\mathcal{M}_0$, the prior predictive distribution for t is simply the familiar Student's t distribution. We compute the marginal likelihood under the null model using the R function `dt`.

For the model $\mathcal{M}_+$, we use the `integrate` function to marginalize over all the possible values of δ under the model, weighted by their prior density. This integral is the marginal likelihood for $\mathcal{M}_+$.

The ratio of these two marginal likelihoods yields the Bayes factor `Bp0`.

```
# Prior
r.scale = 1/2

# Data
d = .387
n_x = 120
n_y = 120
df = n_x + n_y - 2
n_eff = n_x*n_y/(n_x+n_y)
t.stat = d * sqrt(n_eff)

# M0: p(y)
py_M0 = dt(t.stat, df = df)

# M+: p(y|delta)p(delta)
pypd = function(delta){
  dt( t.stat, df,
      delta*sqrt(n_eff) ) *
    2*dcauchy( delta,
      scale = r.scale )
}

# M+: p(y)
py_Mp = integrate(pypd,
      lower = 0,
      upper = Inf)$value

# Bayes factor (about 22)
Bp0 = py_Mp / py_M0
```

Although it is impossible to see in the figure due to the very low density of the data under the hypotheses, the prior predictive density at $d = 0.39$ is 0.05 times higher under $\mathcal{M}_-$ than it is under $\mathcal{M}_0$. The Bayes factor favoring $\mathcal{M}_-$ over $\mathcal{M}_0$ is thus $B_{-0} = 0.05$. Note that this is below 1; the evidence actually favors the model in the denominator, $\mathcal{M}_0$. The favoring of $\mathcal{M}_0$ over $\mathcal{M}_-$ natural due to the fact that the effect size was positive. We can say, equivalently, that $B_{0-} = 20$, in order to express the Bayes factor as a number greater than 1.

From these two Bayes factors, we can compute another Bayes factor representing the evidence favoring $\mathcal{M}_+$ over $\mathcal{M}_-$. The data were 22 times more likely under $\mathcal{M}_+$ than $\mathcal{M}_0$; and they are 22 times more likely under $\mathcal{M}_0$ than $\mathcal{M}_-$. The Bayes factor

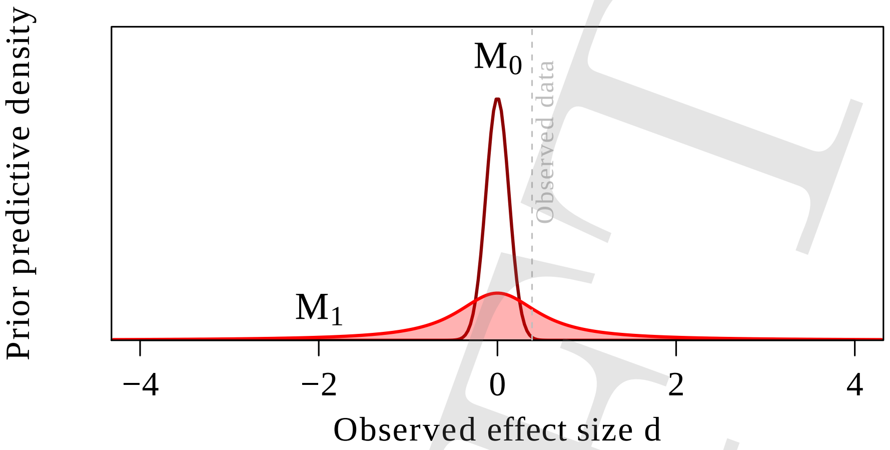

Figure 18: Prior predictive distributions for the observed effect size d under the null model $\mathcal{M}_0$ (dark red, unshaded) and the two-sided alternative model $\mathcal{M}_1$ (red, shaded).

comparing $\mathcal{M}_+$ to $\mathcal{M}_-$ is thus

$$B_{+-} = 22 \times 20 = 436.$$

Finally, we can obtain a two-tailed test by assuming that, for instance, the alternative hypothesis is the entire Cauchy distribution shown in Figure 17A if we weight the positive and negative hypotheses equally. The resulting prior distribution corresponds to a two-tailed alternative distribution. We denote the two-tailed model as $\mathcal{M}_1$, and note that its marginal likelihood is

$$p(d \mid \mathcal{M}_1) = \frac{1}{2} p(d \mid \mathcal{M}_-) + \frac{1}{2} p(d \mid \mathcal{M}_+)$$

If we divide both sides by $p(d \mid \mathcal{M}_0)$, the marginal likelihood for the null model, we can obtain the two tailed Bayes factor:

$$B_{10} = \frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)} = \frac{1}{2} \frac{p(d \mid \mathcal{M}_-)}{p(d \mid \mathcal{M}_0)} + \frac{1}{2} \frac{p(d \mid \mathcal{M}_+)}{p(d \mid \mathcal{M}_0)} = \frac{B_{+0} + B_{-0}}{2}$$

The two-sided Bayes factor quantifying the evidence for the two-sided Cauchy alternative to to the point $\delta = 0$ is thus:

$$B_{10} = \frac{21.99 + 0.05}{2} = 11.02$$

The Bayes factor can also be computed my means of the respective prior predictive distributions, as shown in Figure 18. The observed data are 11.02 times as likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$.

Connection with parameter estimation

Recall that in classical statistics, there is a close connection between parameter estimation using confidence intervals and hypothesis testing: the $100(1-\alpha)\%$ confidence interval contains all the parameter values that would not be rejected by an α-level test, were they tested as the null hypothesis. To determine whether a parameter value θ_0 would be rejected by an α-level test, one can simply look at the corresponding confidence interval; if θ_0 is contained within the interval, than it would not be rejected.

It is tempting to apply the same logic in Bayesian estimation: check to see whether a parameter value θ_0 is located in a credible interval, and then, if so, to reject the value as "not credible". However, Bayesian statistics does not have the same symmetry between hypothesis testing and parameter estimation as classical statistics does, and trying to adopt the classical approach within Bayesian statistics is, as J. O. Berger (2006) puts it, "simply wrong" (p. 383). It has no justification from Bayes' theorem.

This is not, however, to say that there is no symmetry between parameter estimation and hypothesis testing in Bayesian statistics: There is a symmetry, but it is a different symmetry than the one in classical statistics. In order to see the Bayesian symmetry, we now show that we can obtain the same Bayes factors we computed above through a parameter-estimation approach. Suppose that we assumed a full Cauchy prior on the effect size estimate parameter δ:

$$\delta \sim \text{Cauchy}(\text{scale} = 1/2).$$

The likelihood, which contains the evidence that determines how prior is updated, is

$$\frac{p(d \mid \delta)}{\int_{-\infty}^{\infty} p(d \mid \delta) p(\delta)\, d\delta}.$$

This function of δ is shown in Figure 19A. For all values of the function above 1.0, the posterior will be above the prior, and vice versa for values of the function below 1.0. One interesting point to examine is the likelihood at $\delta = 0$, which has the value 0.09. This value is precisely B_{01}, the Bayes factor for the test of the point null hypothesis that $\delta = 0$ against the alternative corresponding to the prior used for the parameter estimation. The Bayesian updating in the parameter estimate is intimately connected with the hypothesis test.

Figure 19B shows the prior and posterior, where the posterior is obtained by multiplying the prior (dashed, red curve) with the likelihood (solid green curve in Figure 19A). Because the prior is symmetric, the prior probability that $\delta > 0$ is exactly $1/2$, and hence the prior odds are

$$\frac{Pr(\delta > 0)}{Pr(\delta \leq 0)} = 1.$$

The posterior probability that $\delta > 0$ is 0.998, as shown by the proportion of the posterior distribution in 19B that is to the right of 0. The corresponding posterior odds are thus 436. The data have changed the odds that $\delta > 0$ by a factor of $436/1$, which is precisely the Bayes factor B_{+-} we obtained previously using the prior predictive distributions.

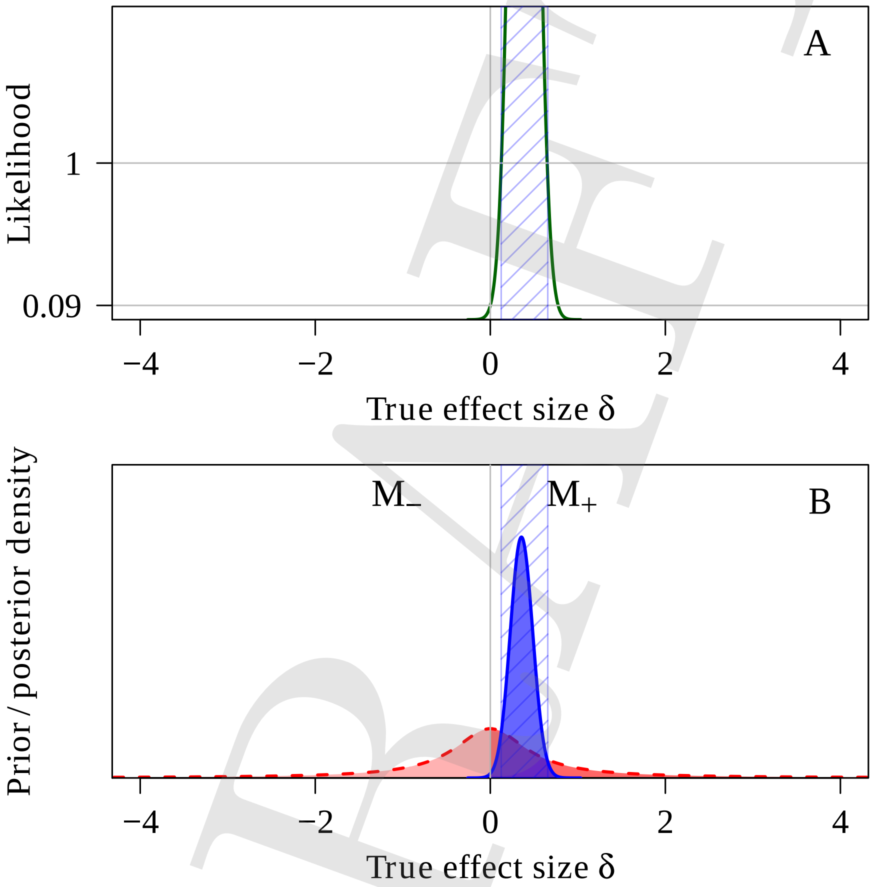

Figure 19: (A) The evidence-scaled likelihood (see text) under the two-sided JZS hypothesis $\mathscr{M}_1$ upon observing $d = 0.39$. The vertical line shows $\delta = 0$, for which the scaled-evidence likelihood is 0.09. (B). The prior (solid,red) and posterior (dashed, blue). Sub-models $\mathscr{M}_-$ (light shading, left of $\delta = 0$) and $\mathscr{M}_+$ (dark shading, right of $\delta = 0$) are also shown.

Dependency on models chosen

It is often pointed out, correctly, out that the Bayes factor is dependent on the models chosen to compare. This dependence can be seen as positive or negative, depending on the perspective. For those who view it as positive, it is natural that the relative evidence for various propositions will depend on the identity of those propositions. For those who view it as negative, the dependence shows that seemingly arbitrary decisions about priors can change the result.

As a demonstration, we now explore how sensitive the Bayes factor can be to the prior specification. We will examine the sensitivity of the two-sided JZS Bayes factor B_{10} discussed in the previous chapter by manipulating the prior scale r through several doublings. Keep in mind that the prior scale quantifies how large or small we expect effect sizes to be; thus, increasing the prior scale is the same as expressing an expectation that large prior scales will be found.

Figure 20A shows three alternative hypotheses that span the range against which the null hypothesis $\mathscr{M}_0$ will be tested. The widest alternative hypothesis, a Cauchy with $r = 2$, gives 50% *a priori* probability that $|d| > 2$. In most research fields, $|d| = 2$ would represent a very large effect. The narrowest hypothesis, a Cauchy with $r = .125$, gives 50% *a priori* probability that $|d| < .125$. In the middle is the $r = 1/2$ hypothesis that was used throughout this section, which gives 50% *a priori* probability that $|d| > .5$. This range of prior scales represents four doublings from the smallest to the largest.

The corresponding Bayes factors are shown in Figure 20B. These Bayes factors are all between between 4 and 12. This represents a much smaller change than what might be expected on the basis of the large change in the prior. If one continued to increase the prior scale to ∞, the Bayes factor would continue to approach 0: that is, as one expects larger and larger effect sizes, the small effect size looks increasingly null. As the prior scale approaches 0, the Bayes factor will approach 1 because the a Cauchy with a prior scale of 0 is indistinguishable from a point null.

We should note, however, that it is not always intuitive how the Bayes factor will depend on the prior in models more complex than the one parameter model presented here. Nuisance parameters and hierarchical models can pose difficulties with Bayes factors, because a Bayes factor is a test of the entire model jointly. Care must be taken to specify a model that allows a faithful test of the hypothesis in question, not merely a test of ancillary assumptions.

Broader considerations

There are a number of considerations underlying statistical inference that are broader than any single inferential philosophy. We address several here: parametric assumptions, and model checking.

Parametric and non-parametric inference

The exposition in chapter was based on the idea of estimating parameters and testing parametric hypotheses. This was largely for ease of exposition and interface with

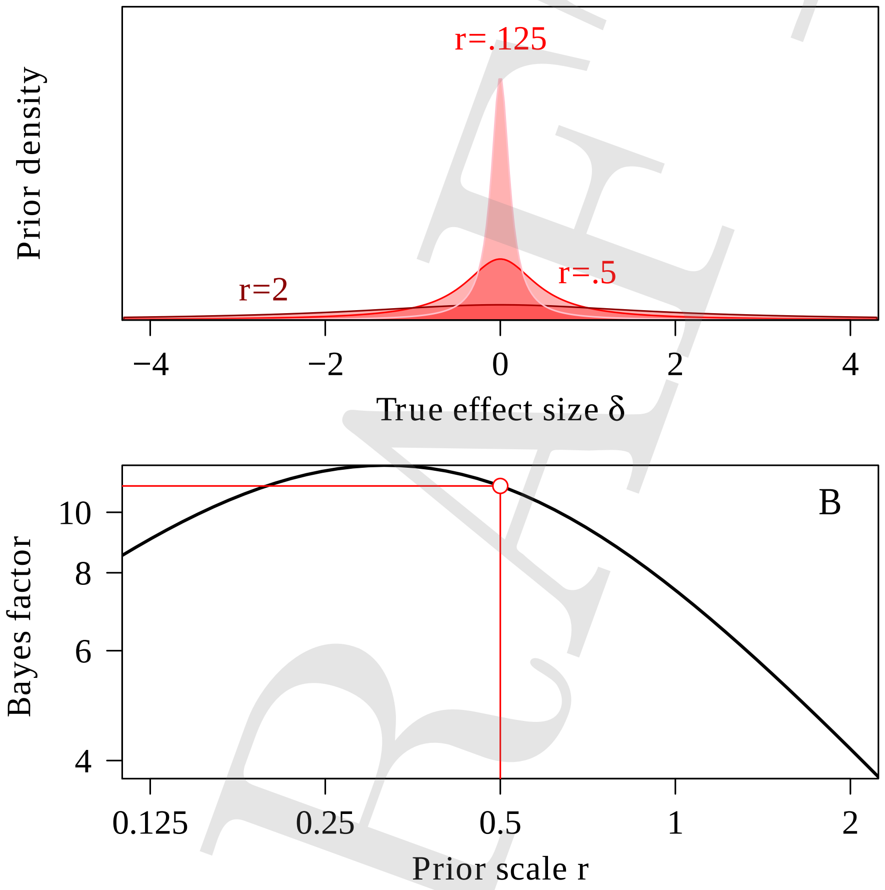

Figure 20: The sensitivity of the JZS Bayes factor to the prior scale. (A) Three Cauchy prior distributions on the effect size δ. Each prior distribution is labeled with its respective prior scale r. (B) The JZS Bayes factor as a function of the prior scale r. The point represents $r = 0.5$, as used in the text.

the most commonly-used statistical techniques.

However, in some cases one may not know what parametric assumptions to make about one's data, if any. In these cases, non-parametric inferential methods exist to avoid making strong assumptions. For an introduction to frequentist nonparametric inference, see Conover (1971); for Bayesian nonparametrics, see Hjort, Holmes, Müller, and Walker (2010).

If parametric inference can be undertaken, it will often lead to more interpretable, more powerful inferences; however, nonparametric inference serves as a viable alternative if parametric the viability of parametric inference is in doubt. Model checking is a common way of choosing whether a particular parametric approach, or nonparametric approach, appropriate.

Model checking

All statistical inferences are based on a number of assumptions. These may range from very strong assumptions for the sort of parametric inferences covered in this chapter, to the weaker assumptions of nonparametric techniques. A key step in statistical inference is checking the assumptions underlying the inference.

How model checking proceeds is largely informal (Andrew Gelman & Shalizi, 2013; Morey, Romeijn, & Rouder, 2013). It may involve tests of assumptions (e.g., Levene's test for equal variances) or visualizations designed to make plain violations of key assumptions (e.g. residual plots in regression). Wilkinson and the Task Force on Statistical Inference (1999) recommend that visualizations — as opposed to formal statistical tests — play a primary role in model checking. Using visualizations allows one to be sensitive to a wider range of possible violations than can be quantified by the small number of statistical tests that are typically used.

Conclusion

Statistical inference is an indispensable tool for scientists. There are a number of approaches to statistical inference, each with its own atomic concepts (frequency and error, likelihood, and belief-as-probability, for frequentism, likelihoodism, and Bayesian statistics respectively) and inferential philosophy. Thus whether one prefers the inferential techniques from one philosophy or another must be guided by the aims of the analyst.

It would be an understatement to say that statistical philosophy has been a rhetorical battleground for decades. Each philosophy has its ardent defenders, and reading the work by these authors will give one a sense of the issues at stake. Accessible accounts of various approaches include David R. Cox (2006), Neyman (1977), and Mayo and Cox (2006) from the frequentist perspective; A. Edwards (1992) and Royall (1997) from the likelihoodist perspective; and W. Edwards et al. (1963), A. Gelman et al. (2004), and Lee (2004) from the Bayesian perspective. A firm grounding in the philosophies of statistical inference will benefit even experienced data analysts.

References

Abramowitz, M. & Stegun, I. A. (1965). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. New York: Dover.

Bayarri, M. J. & Berger, J. O. (2004, February). The interplay of bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58–80.

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on r^d. *Journal of Applied Probability*, *29*(4), 885–895.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Second edition, Vol. 1, pp. 378–386). Hoboken, New Jersey: John Wiley & Sons.

Berger, J. O. & Bernardo, J. M. (1992). On the development of reference priors. In *Bayesian statistics 4. Proceedings of the Fourth Valencia International Meeting* (pp. 35–49).

Berger, J. O. & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)* Hayward, CA: Institute of Mathematical Statistics.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian theory*. Chichester, England: John Wiley & Sons.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*, 269–326.

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Conover, W. J. (1971). *Practical nonparametric statistics*. New York: Wiley.

Cox, D. R. [David R.]. (2006). *Principles of statistical inference*. Cambridge University Press.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cureton, E. E. (1968a). Priority correction to "unbiased estimation of the standard deviation". *The American Statistician*, *22*(3), 27–27.

Cureton, E. E. (1968b). Unbiased estimation of the standard deviation. *The American Statistician*, *22*(1), 22–22.

de Finetti, B. (1937). Foresight: its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in subjective probability*. New York: Wiley.

de Groot, M. H. & Schervish, M. J. (2012). *Probability and statistics* (fourth edition). Addison-Wesley.

Edwards, A. (1992). *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference* (Expanded edition). London: The John Hopkins University Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Evans, M. (2013). What does the proof of Birnbaum's theorem prove? *Electronic Journal of Statistics*, *7*, 2645–2655.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*, 39–82.

Fisher, R. A. (1966). *The design of experiments* (8th edition). New York: Hafner Publishing Company.

Gelman, A. [A.], Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman and Hall.

Gelman, A. [Andrew] & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 57–64.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, *1*, 403–420.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge, England: Cambridge University Press.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics*. Cambridge series in statistical and probabilistic mathematics. Cambridge, UK: Cambridge University Press.

Hogg, R. V. & Craig, A. T. (1978). *Introduction to mathematical statistics*. New York: MacMillan.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*(1007), 453–461.

Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.

Kendall, M. G., Bernoulli, D., Allen, C. G., & Euler, L. (1961). Studies in the history of probability and statistics: XI. Daniel Bernoulli on maximum likelihood. *Biometrika*, *48*(1/2), 1–18.

Kennedy, J. & Eberhart, R. (1995, November). Particle swarm optimization. In *IEEE International Conference on Neural Networks, 1995. Proceedings* (Vol. 4, 1942–1948 vol.4).

Lee, P. M. (2004). *Bayesian statistics: An introduction (3rd ed.)* New York: Wiley.

Macmillan, N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd). Mahwah, N.J.: Lawrence Erlbaum Associates.

Mayo, D. G. (2014). On the birnbaum argument for the strong likelihood principle (with discussion & rejoinder). *Statistical Science*, *29*, 227–266.

Mayo, D. G. & Cox, D. R. [D. R.]. (2006). Frequentist statistics as a theory of inductive inference. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, *49*, 77–97.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68–75.

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*, 333–380.

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. Washington, D.C.: Graduate School, U.S. Department of Agriculture.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231*, 289–337.

R Core Team. (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rice, J. (1998). *Mathematical statistics and data analysis*. Monterey, CA: Brooks/Cole.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597–605.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237.

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. New York: CRC Press.

Senn, S. (2001). Two cheers for P-values? *Journal of Epidemiology and Biostatistics*, *6*(2), 193–204.

Senn, S. (2007). *Statistical issues in drug development* (2nd ed.). Chichester, UK: John Wiley & Sons.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*(5), 301–340.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: an apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, *10*(4), 299–326.

Wilkinson, L. & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*, 60–62.