Chapter 5

Confidence Intervals and Hypothesis Testing

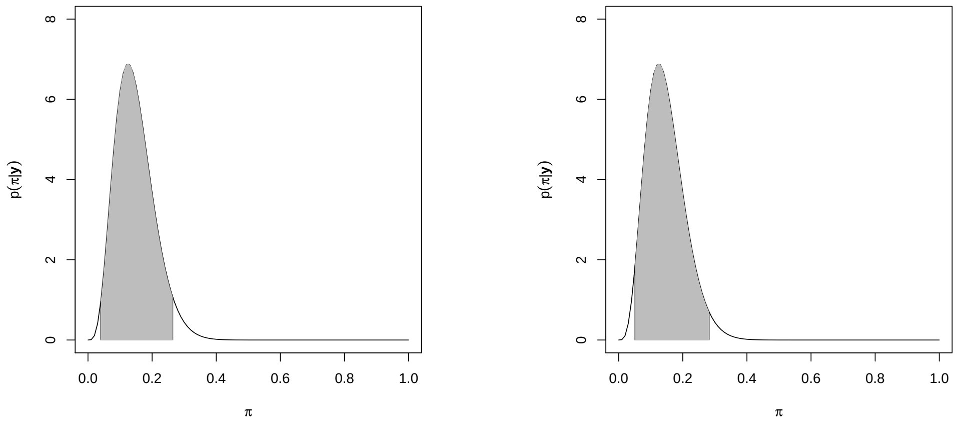
Although Chapter 4 introduced the theoretical framework for estimating the parameters of a model, it was very much situated in the context of prediction: the focus of statistical inference is on inferring the kinds of additional data that are likely to be generated by a model, on the basis of an existing set of observations. In much of scientific inquiry, however, we wish to use data to make inferences about models themselves: what plausible range can be inferred for a parameter or set of parameters within a model, or which of multiple models a given set of data most strongly supports. These are the problems of CONFIDENCE INTERVALS and HYPOTHESIS TESTING respectively. This chapter covers the fundamentals of Bayesian and frequentist approaches to these problems.

5.1 Bayesian confidence intervals

Recall from Section 4.4 that Bayesian parameter estimation simply involves placing a posterior probability distribution over the parameters θ of a model, on the basis of Bayes rule:

$$P(\theta|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\theta)P(\theta)}{P(\boldsymbol{y})} \tag{5.1}$$

In Bayesian inference, a CONFIDENCE INTERVAL over a single model parameter ϕ is simply a contiguous interval $[\phi_1, \phi_2]$ that contains a specified proportion of the posterior probability mass over ϕ. The proportion of probability mass contained in the confidence interval can be chosen depending on whether one wants a narrower or wider interval. The tightness of the interval (in frequentist as well as Bayesian statistics) is denoted by a value α that expresses the amount of probability mass *excluded* from the interval—so that $(1-\alpha)\%$ of the probability mass is within the interval. The interpretation of a $(1-\alpha)\%$ confidence interval $[\phi_1, \phi_2]$ is that **the probability that the model parameter ϕ resides in $[\phi_1, \phi_2]$ is $(1-\alpha)$.**



(a) 95% HPD confidence interval

(b) Symmetric 95% confidence interval

Figure 5.1: HPD and symmetric Bayesian confidence intervals for a posterior distributed as $Beta(5, 29)$

Of course, there is always more than one way of choosing the bounds of the interval $[\phi_1, \phi_2]$ to enclose $(1 - \alpha)\%$ of the posterior mass. There are two main conventions for determining how to choose interval boundaries:

- Choose the *shortest* possible interval enclosing $(1 - \alpha)\%$ of the posterior mass. This is called a HIGHEST POSTERIOR DENSITY (HPD) confidence interval.

- Choose interval boundaries such that an *equal amount of probability mass* is contained on either side of the interval. That is, choose $[\phi_1, \phi_2]$ such that $P(\phi < \phi_1 | \boldsymbol{y}) = P(\phi > \phi_2 | \boldsymbol{y}) = \frac{\alpha}{2}$. This is called a SYMMETRIC confidence interval.

Let us return, for example, to our American English speaker of Chapter 4, assuming that she models speaker choice in passivization as a binomial random variable (with passive voice being "success") with parameter π over which she has a Beta prior distribution with parameters $(3, 24)$, and observes five active and two passive clauses. The posterior over π has distribution $Beta(5, 29)$. Figure 5.1 shows HPD and symmetric 95% confidence intervals over π, shaded in gray, for this posterior distribution. The posterior is quite asymmetric, and for the HPD interval there is more probability mass to the right of the interval than there is to the left. The intervals themselves are, of course, qualitatively quite similar.

5.2 Bayesian hypothesis testing

In all types of statistics, hypothesis testing involves entertaining multiple candidate generative models of how observed data has been generated. The hypothesis test involves an assessment of which model is most strongly warranted by the data. Bayesian hypothesis testing in particular works just like any other type of Bayesian inference. Suppose that we have a collection of hypotheses $H_1, \dots, H_n$. Informally, a hypothesis can range over diverse ideas such as "this coin is fair", "the animacy of the agent of a clause affects the tendency of speakers to use the passive voice", "females have higher average F1 vowel formants than males regardless of the specific vowel", or "a word's frequency has no effect on naming latency". Formally, each hypothesis should specify a model that determines a probability distribution over possible observations $\boldsymbol{y}$. Furthermore, we need a prior probability over the collection of hypotheses, $P(H_i)$. Once we have observed some data $\boldsymbol{y}$, we use Bayes' rule (Section 2.4.1) to calculate the posterior probability distribution over hypotheses:

$$P(H_i|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|H_i)P(H_i)}{P(\boldsymbol{y})} \tag{5.2}$$

where $P(\boldsymbol{y})$ marginalizes (Section 3.2) over the hypotheses:

$$P(\boldsymbol{y}) = \sum_{j=1}^{n} P(\boldsymbol{y}|H_j)P(H_j) \tag{5.3}$$

As an example, let us return once more to the case of English binomials, such as *salt and pepper*. A number of constraints have been hypothesized to play a role in determining binomial ordering preferences; as an example, one hypothesized constraint is that ordered binomials of the form *A and B* should be disfavored when *B* has ultimate-syllable stress (*Bstr; Bolinger, 1962; Müller, 1997). For example, *pepper and salt* violates this constraint against ultimate-syllable stress, but its alternate *salt and pepper* does not. We can construct a simple probabilistic model of the role of *Bstr in binomial ordering preferences by assuming that every time an English binomial is produced that could potentially violate *Bstr, the binomial is produced in the satisfying order *B and A* ordering with probability π, otherwise it is produced in the violating ordering *A and B*.[1] If we observe n such English binomials, then the distribution over the number of satisfactions of *Bstr observed is (appropriately enough) the binomial distribution with parameters π and n.

Let us now entertain two hypotheses about the possible role of *Bstr in determining binomial ordering preferences. In the first hypothesis, H_1, *Bstr plays no role, hence orderings *A and B* and *B and A* are equally probable; we call this the "no-preference" hypothesis. Therefore in H_1 the binomial parameter π is 0.5. In Bayesian inference, we need to assign probability distributions to choices for model parameters, so we state H_1 as:

$$H_1 : P(\pi|H_1) = \begin{cases} 1 & \pi = 0.5 \\ 0 & \pi \neq 0.5 \end{cases}$$

[1] For now we ignore the role of multiple overlapping constraints in jointly determining ordering preferences, as well as the fact that specific binomials may have idiosyncratic ordering preferences above and beyond their constituent constraints. The tools to deal with these factors are introduced in Chapters 6 and 8 respectively.

The probability above is a prior probability on the binomial parameter π.

In our second hypothesis H_2, *Bstr *does* affect binomial ordering preferences (the "preference" hypothesis). For this hypothesis we must place a non-trivial probability distribution on π. Keep in mind have arbitrarily associated the "success" outcome with satisfaction of *Bstr. Suppose that we consider only two possibilities in H_2: that the preference is either $\frac{2}{3}$ for *A and B* or $\frac{2}{3}$ for outcome *B and A*, and let these two preferences be equally likely in H_2. This gives us:

$$H_2 : P(\pi|H_2) = \begin{cases} 0.5 & \pi = \frac{1}{3} \\ 0.5 & \pi = \frac{2}{3} \end{cases} \tag{5.4}$$

In order to complete the Bayesian inference of Equation (5.2), we need prior probabilities on the hypotheses themselves, $P(H_1)$ and $P(H_2)$. If we had strong beliefs one way or another about the binomial's ordering preference (e.g., from prior experience with other English binomials, or with experience with a semantically equivalent binomial in other languages), we might set one of these prior probabilities close to 1. For these purposes, we will use $P(H_1) = P(H_2) = 0.5$.

Now suppose we collect a dataset $\boldsymbol{y}$ of six English binomials in which two orderings violate *Bstr from a corpus:

Binomial	Constraint status (S: *Bstr satisfied, V: *Bstr violated)
salt and pepper	S
build and operate	S
follow and understand	V
harass and punish	S
ungallant and untrue	V
bold and entertaining	S

Do these data favor H_1 or H_2?

We answer this question by completing Equation (5.2). We have:

$$\begin{aligned} P(H_1) &= 0.5 \\ P(\boldsymbol{y}|H_1) &= \binom{6}{4} \pi^4 (1-\pi)^2 && = \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = 0.23 \end{aligned}$$

Now to complete the calculation of $P(\boldsymbol{y})$ in Equation (5.3), we need $P(\boldsymbol{y}|H_2)$. To get this, we need to marginalize over the possible values of π, just as we are marginalizing over H to get the probability of the data. We have:

$$
\begin{aligned}
P\left(\boldsymbol{y}|H_2\right) &= \sum_i P\left(\boldsymbol{y}|\pi_i\right) P\left(\pi_i|H_2\right) \\
&= P\left(\boldsymbol{y}|\pi = \frac{1}{3}\right) P\left(\pi = \frac{1}{3}|H_2\right) + \quad P\left(\boldsymbol{y}|\pi = \frac{2}{3}\right) P\left(\pi = \frac{2}{3}|H_2\right) \\
&= \binom{6}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 \times 0.5 + \binom{6}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^2 \times 0.5 \\
&= 0.21
\end{aligned}
$$

thus

$$
\begin{aligned}
P(\boldsymbol{y}) &= \overbrace{0.23}^{P(\boldsymbol{y}|H_1)} \times \overbrace{0.5}^{P(H_1)} + \overbrace{0.21}^{P(\boldsymbol{y}|H_2)} \times \overbrace{0.5}^{P(H_2)} && (5.5) \\
&= 0.22 && (5.6)
\end{aligned}
$$

And we have

$$
\begin{aligned}
P(H_1|\boldsymbol{y}) &= \frac{0.23 \times 0.5}{0.22} && (5.7) \\
&= 0.53 && (5.8)
\end{aligned}
$$

Note that even though the maximum-likelihood estimate of $\hat{\pi}$ from the data we observed is exactly one of the two possible values of π under H_2, our data in fact support the "preference" hypothesis H_1 – it went from prior probability $P(H_1) = 0.5$ up to posterior probability $P(H_1|\boldsymbol{y}) = 0.53$. See also Exercise 5.3.

5.2.1 More complex hypotheses

We might also want to consider more complex hypotheses than H_2 above as the "preference" hypothesis. For example, we might think all possible values of π in $[0, 1]$ are equally probable *a priori*:

$$H_3 : P(\pi|H_3) = 1 \quad 0 \leq \pi \leq 1$$

(In Hypothesis 3, the probability distribution over π is continuous, not discrete, so H_3 is still a proper probability distribution.) Let us discard H_2 and now compare H_1 against H_3.

Let us compare H_3 against H_1 for the same data. To do so, we need to calculate the likelihood $P(\boldsymbol{y}|H_3)$, and to do this, we need to marginalize over π:

Since π can take on a continuous range of values under H_3, this marginalization takes the form of an integral:

$$P(\boldsymbol{y}|H_3) = \int_\pi P(\boldsymbol{y}|\pi)P(\pi|H_3)\,d\pi = \int_0^1 \overbrace{\binom{6}{4}\pi^4(1-\pi)^2}^{P(\boldsymbol{y}|\pi)} \overbrace{1}^{P(\pi|H_3)}\,d\pi$$

We use the critical trick of recognizing this integral as a beta function (Section 4.4.2), which gives us:

$$= \binom{6}{4}B(5,3) = 0.14$$

If we plug this result back in, we find that

$$\begin{aligned} P(H_1|\boldsymbol{y}) &= \frac{\overbrace{0.23}^{P(\boldsymbol{y}|H_1)} \times \overbrace{0.5}^{P(H_1)}}{\underbrace{0.23}_{P(\boldsymbol{y}|H_1)} \times \underbrace{0.5}_{P(H_1)} + \underbrace{0.14}_{P(\boldsymbol{y}|H_3)} \times \underbrace{0.5}_{P(H_3)}} \\ &= 0.62 \end{aligned}$$

So H_3 fares even worse than H_2 against the no-preference hypothesis H_1. Correspondingly, we would find that H_2 is favored over H_3.

5.2.2 Bayes factor

Sometimes we do not have strong feelings about the prior probabilities $P(H_i)$. Nevertheless, we can quantify how much evidence a given dataset provides for one hypothesis over another. We can express the relative preference between H and H' in the face of data $\boldsymbol{y}$ in terms of the PRIOR ODDS of H versus H' combined with the LIKELIHOOD RATIO between the two hypotheses. This combination gives us the POSTERIOR ODDS:

$$\overbrace{\frac{P(H|\boldsymbol{y})}{P(H'|\boldsymbol{y})}}^{\text{Posterior odds}} = \overbrace{\frac{P(\boldsymbol{y}|H)}{P(\boldsymbol{y}|H')}}^{\text{Likelihood ratio}} \overbrace{\frac{P(H)}{P(H')}}^{\text{Prior odds}}$$

The contribution of the data $\boldsymbol{y}$ to the posterior odds is simply the likelihood ratio:

$$\frac{P(\boldsymbol{y}|H)}{P(\boldsymbol{y}|H')} \qquad (5.9)$$

which is also called the BAYES FACTOR between H and H'. A Bayes factor above 1 indicates support for H over H'; a Bayes factor below 1 indicates support for H' over H. For example, the Bayes factors for H_1 versus H_2 and H_1 versus H_3 in the preceding examples

$$\frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_2)} = \frac{0.23}{0.21} = 1.14 \qquad \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_3)} = \frac{0.23}{0.14} = 1.64$$

indicating weak support for H_1 in both cases.

5.2.3 Example: Learning contextual contingencies in sequences

One of the key tasks of a language learner is to determine which cues to attend to in learning distributional facts of the language in their environment (Saffran et al., 1996a; Aslin et al., 1998; Swingley, 2005; Goldwater et al., 2007). In many cases, this problem of cue relevance can be framed in terms of hypothesis testing or model selection.

As a simplified example, consider a length-21 sequence of syllables:

```
da ta da ta ta da da da da da ta ta ta da ta ta ta da da da da
```

Let us entertain two hypotheses. The first hypothesis H_1, is that the probability of an `da` is independent of the context. The second hypothesis, H_2, is that the probability of an `da` is dependent on the preceding token. The learner's problem is to choose between these hypotheses—that is, to decide whether immediately preceding context is relevant in estimating the probability distribution over what the next phoneme will be. How should the above data influence the learner's choice? (Before proceeding, you might want to take a moment to examine the sequence carefully and answer this question on the basis of your own intuition.)

We can make these hypotheses precise in terms of the parameters that each entails. H_1 involves only one binomial parameter $P(\texttt{da})$, which we will denote as π. H_2 involves three binomial parameters:

1. $P(\texttt{da}|\emptyset)$ (the probability that the sequence will start with `da`), which we will denote as $\pi_\emptyset$;
2. $P(\texttt{da}|\texttt{da})$ (the probability that an `da` will appear after an `da`), which we will denote as $\pi_{\texttt{da}}$;
3. $P(\texttt{da}|\texttt{ta})$ (the probability that an `da` will appear after an `ta`), which we will denote as $\pi_{\texttt{ta}}$.

(For expository purposes we will assume that the probability distribution over the number of syllables in the utterance is the same under both H_1 and H_2 and hence plays no role in the Bayes factor.) Let us assume that H_1 and H_2 are equally likely; we will be concerned with the Bayes factor between the two hypotheses. We will put a uniform prior distribution on all model parameters—recall that this can be expressed as a beta density with parameters $\alpha_1 = \alpha_2 = 1$ (Section 4.4.2).

There are 21 observations, 12 of which are da and 9 of which are ta. The likelihood of H_1 is therefore simply

$$\begin{aligned}\int_0^1 \pi^{12}(1-\pi)^9 \, d\pi &= B(13,10) \\ &= 1.55 \times 10^{-7}\end{aligned}$$

once again recognizing the integral as a beta function (see Section 4.4.2).

To calculate the likelihood of H_2 it helps to lay out the 21 events as a table of conditioning contexts and outcomes:

	Outcome	
Context	da	ta
∅	1	0
da	7	4
ta	4	5

The likelihood of H_2 is therefore

$$\begin{aligned}\int_0^1 \pi_{\emptyset}^1 \, d\pi_{\emptyset} \int_0^1 \pi_{\texttt{da}}^7 (1-\pi_{\texttt{da}})^4 \, d\pi_{\texttt{da}} \int_0^1 \pi_{\texttt{ta}}^4 (1-\pi_{\texttt{ta}})^5 \, d\pi_{\texttt{ta}} &= B(2,1)B(8,5)B(5,6) \\ &= 1 \times 10^{-7}\end{aligned}$$

This dataset provides some support for the simpler hypothesis of statistical independence—the Bayes factor is 1.55 in favor of H_1.

5.2.4 Phoneme discrimination as hypothesis testing

In order to distinguish spoken words such as *bat* and *pat* out of context, a listener must rely on acoustic cues to discriminate the sequence of phonemes that is being uttered. One particularly well-studied case of phoneme discrimination is of voicing in stop consonants. A variety of cues are available to identify voicing; here we focus on the well-studied cue of *voice onset time* (VOT)—the duration between the sound made by the burst of air when the stop is released and the onset of voicing in the subsequent segment. In English, VOT is shorter for so-called "voiced" stops (e.g., /b/,/d/,/g/) and longer for so-called "voiceless" stops (e.g., /p/,/t/,/k/), particularly word-initially, and native speakers have been shown to be sensitive to VOT in phonemic and lexical judgments (Liberman et al., 1957).

Within a probabilistic framework, phoneme categorization is well-suited to analysis as a Bayesian hypothesis test. For purposes of illustration, we dramatically simplify the problem by focusing on two-way discrimination between the voiced/voiceless stop pair /b/ and /p/. In order to determine the phoneme-discrimination inferences of a Bayesian listener, we must specify the acoustic representations that describe spoken realizations x of any phoneme,

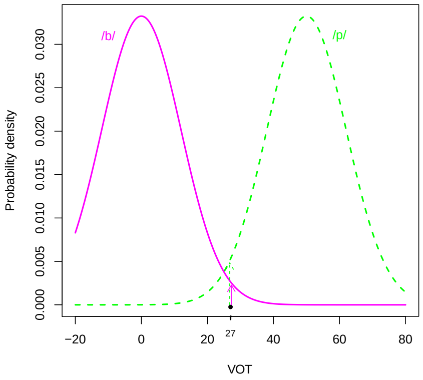


Figure 5.2: Likelihood functions for /b/–/p/ phoneme categorizations, with $\mu_{\mathrm{b}} = 0, \mu_{\mathrm{p}} = 50, \sigma_{\mathrm{b}} = \sigma_{\mathrm{p}} = 12$. For the input $x = 27$, the likelihoods favor /p/.

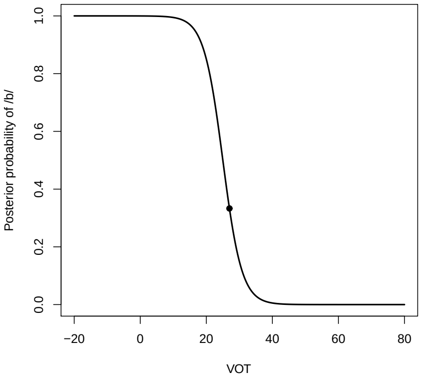


Figure 5.3: Posterior probability curve for Bayesian phoneme discrimination as a function of VOT

the conditional distributions over acoustic representations, $P_{\mathrm{b}}(x)$ and $P_{\mathrm{p}}(x)$ for /b/ and /p/ respectively (the likelihood functions), and the prior distribution over /b/ versus /p/. We further simplify the problem by characterizing any acoustic representation x as a single real-valued number representing the VOT, and the likelihood functions for /b/ and /p/ as normal density functions (Section 2.10) with means $\mu_{\mathrm{b}}, \mu_{\mathrm{p}}$ and standard deviations $\sigma_{\mathrm{b}}, \sigma_{\mathrm{p}}$ respectively.

Figure 5.2 illustrates the likelihood functions for the choices $\mu_{\mathrm{b}} = 0, \mu_{\mathrm{p}} = 50, \sigma_{\mathrm{b}} = \sigma_{\mathrm{p}} = 12$. Intuitively, the phoneme that is more likely to be realized with VOT in the vicinity of a given input is a better choice for the input, and the greater the discrepancy in the likelihoods the stronger the categorization preference. An input with non-negligible likelihood for each phoneme is close to the "categorization boundary", but may still have a preference. These intuitions are formally realized in Bayes' Rule:

$$P(/\mathrm{b}/|x) = \frac{P(x|/\mathrm{b}/)P(/\mathrm{b}/)}{P(x)} \tag{5.10}$$

and since we are considering only two alternatives, the marginal likelihood is simply the weighted sum of the likelihoods under the two phonemes: $P(x) = P(x|/\mathrm{b}/)P(/\mathrm{b}/) + P(x|/\mathrm{p}/)P(/\mathrm{p}/)$. If we plug in the normal probability density function we get

$$P(/\mathrm{b}/|x) = \frac{\frac{1}{\sqrt{2\pi\sigma_{\mathrm{b}}^2}} \exp\left[-\frac{(x-\mu_{\mathrm{b}})^2}{2\sigma_{\mathrm{b}}^2}\right] P(/\mathrm{b}/)}{\frac{1}{\sqrt{2\pi\sigma_{\mathrm{b}}^2}} \exp\left[-\frac{(x-\mu_{\mathrm{b}})^2}{2\sigma_{\mathrm{b}}^2}\right] P(/\mathrm{b}/) + \frac{1}{\sqrt{2\pi\sigma_{\mathrm{p}}^2}} \exp\left[-\frac{(x-\mu_{\mathrm{p}})^2}{2\sigma_{\mathrm{p}}^2}\right] P(/\mathrm{p}/)} \tag{5.11}$$

In the special case where $\sigma_{\mathrm{b}} = \sigma_{\mathrm{p}} = \sigma$ we can simplify this considerably by cancelling the

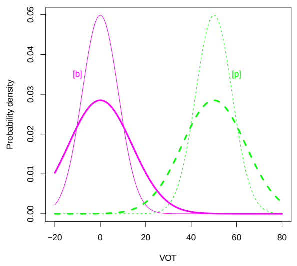


Figure 5.4: Clayards et al. (2008)'s manipulation of VOT variance for /b/–/p/ categories

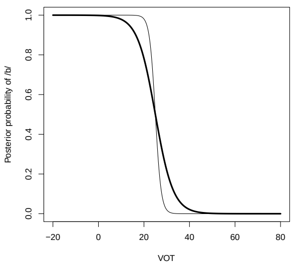


Figure 5.5: Ideal posterior distributions for narrow and wide variances

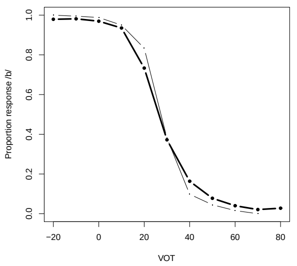


Figure 5.6: Response rates observed by Clayards et al. (2008)

constants and multiplying through by $\exp\left[\frac{(x-\mu_{\mathrm{b}})^2}{2\sigma_{\mathrm{b}}^2}\right]$:

$$P(/\mathrm{b}/|x) = \frac{P(/\mathrm{b}/)}{P(/\mathrm{b}/) + \exp\left[\frac{(x-\mu_{\mathrm{b}})^2-(x-\mu_{\mathrm{p}})^2}{2\sigma^2}\right] P(/\mathrm{p}/)} \tag{5.12}$$

Since $e^0 = 1$, when $(x - \mu_{\mathrm{b}})^2 = (x - \mu_{\mathrm{p}})^2$ the input is "on the category boundary" and the posterior probabilities of each phoneme are unchanged from the prior. When x is closer to μ_{b}, $(x - \mu_{\mathrm{b}})^2 - (x - \mu_{\mathrm{p}})^2 > 0$ and /b/ is favored; and vice versa when x is closer to μ_{p}. Figure 5.3 illustrates the phoneme categorization curve for the likelihood parameters chosen for this example and the prior $P(/\mathrm{b}/) = P(/\mathrm{p}/) = 0.5$.

This account makes clear, testable predictions about the dependence on the parameters of the VOT distribution for each sound category on the response profile. Clayards et al. (2008), for example, conducted an experiment in which native English speakers were exposed repeatedly to words with initial stops on a /b/–/p/ continuum such that either sound category would form a word (*beach–peach*, *beak–peak*, *bes–peas*). The distribution of the /b/–/p/ continuum used in the experiment was bimodal, approximating two overlapping Gaussian distributions (Section 2.10); high-variance distributions (156ms^2) were used for some experimental participants and low-variance distribution (64ms^2) for others (Figure 5.4). If these speakers were to learn the true underlying distributions to which they were exposed and use them to draw ideal Bayesian inferences about which word they heard on a given trial, then the posterior distribution as a function of VOT would be as in Figure 5.5: note that low-variance Gaussians would induce a steeper response curve than high-variance Gaussians. The actual response rates are given in Figure 5.6; although the discrepancy between the low- and high-variance conditions is smaller than predicted by ideal inference, suggesting that learning may have been incomplete, the results of Clayards et al. confirm human response curves are indeed steeper when category variances are lower, as predicted by principles of Bayesian inference.

5.3 Frequentist confidence intervals

We now move on to frequentist confidence intervals and hypothesis testing, which have been developed from a different philosophical standpoint. To a frequentist, it does not make sense to say that "the true parameter θ lies between these points x and y with probability p^*." The parameter θ is a real property of the population from which the sample was obtained and is either in between x and y, or it is not. Remember, to a frequentist, the notion of probability as reasonable belief is not admitted! Under this perspective, the Bayesian definition of a confidence interval—while intuitively appealing to many—is incoherent.

Instead, the frequentist uses more indirect means of quantifying their certainty about the estimate of θ. The issue is phrased thus: imagine that I were to repeat the same experiment—drawing a sample from my population—many times, and each time I repeated the experiment I constructed an interval I on the basis of my sample according to a fixed procedure `Proc`. Suppose it were the case that $1 - p$ percent of the intervals I thus constructed actually contained θ. Then for any given sample S, the interval I constructed by `Proc` is a $(1 - p)\%$ confidence interval for θ.

If you think that this seems like convoluted logic, well, you are not alone. **Frequentist confidence intervals are one of the most widely misunderstood constructs in statistics.** The Bayesian view is more intuitive to most people. Under some circumstances, there is a happy coincidence where Bayesian and frequentist confidence intervals look the same and you are free to misinterpret the latter as the former. In general, however, they do *not* necessarily look the same, and you need to be careful to interpret each correctly.

Here's an example, where we will explain the STANDARD ERROR OF THE MEAN. Suppose that we obtain a sample of n observations from a normal distribution $N(\mu, \sigma^2)$. It turns out that the following quantity follows the t_{n-1} distribution (Section B.5):

$$\frac{\hat{\mu} - \mu}{\sqrt{S^2/n}} \sim t_{n-1} \tag{5.13}$$

where

$$\hat{\mu} = \frac{1}{n}\sum_i X_i \qquad \text{[maximum-likelihood estimate of the mean]}$$

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \hat{\mu})^2 \qquad \text{[unbiased estimate of } \sigma^2\text{; Section 4.3.3]}$$

Let us denote the quantile function for the t_{n-1} distribution as $Q_{t_{n-1}}$. We want to choose a symmetric interval $[-a, a]$ containing $(1 - \alpha)$ of the probability mass of t_{n-1}. Since the t distribution is symmetric around 0, if we set $a = \sqrt{S^2/n}\, Q_{t_{n-1}}(1 - \alpha/2)$, we will have

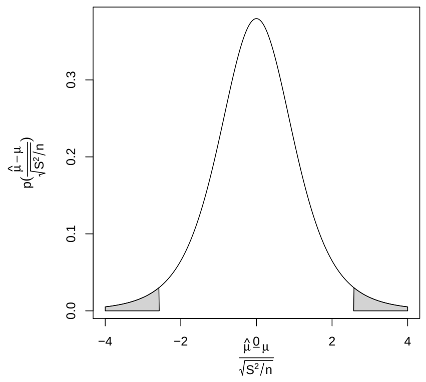


Figure 5.7: Visualizing confidence intervals with the t distribution

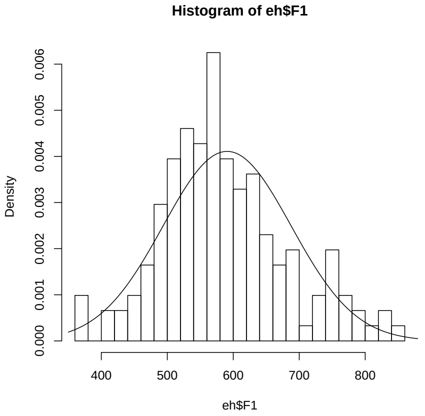


Figure 5.8: Distribution of F1 formant for [ɛ]

$$
\begin{aligned}
P(\hat{\mu} - \mu < -a) &= \frac{\alpha}{2} \\
P(\hat{\mu} - \mu > a) &= \frac{\alpha}{2}
\end{aligned}
\tag{5.14}
$$

Figure 5.7 illustrates this for $\alpha = 0.05$ (a 95% confidence interval). Most of the time, the "standardized" difference between $\hat{\mu}$ and μ is small and falls in the unshaded area. But 5% of the time, this standardized difference will fall in the shaded area—that is, the confidence interval won't contain μ.

Note that the quantity $S/\sqrt{n}$ is called the the STANDARD ERROR OF THE MEAN or simply the STANDARD ERROR. Note that this is different from the standard deviation of the sample, but related! (How?) When the number of observations n is large, the t distribution looks approximately normal, and as a rule of thumb, the symmetric 95% tail region of the normal distribution is about 2 standard errors away from the mean.

Another example: let's look at the data from a classic study of the English vowel space (Peterson and Barney, 1952). The distribution of the F1 formant for the vowel ɛ is roughly normal (see Figure 5.8). The 95% confidence interval can be calculated by looking at the quantity $S/\sqrt{n}\ Q_{t_{151}}(0.975) = 15.6$. This is half the length of the confidence interval; the confidence interval should be centered around the sample mean $\hat{\mu} = 590.7$. Therefore our 95% confidence interval for the mean F1 is $[575.1, 606.3]$.

5.4 Frequentist hypothesis testing

In most of science, including areas such as psycholinguistics and phonetics, statistical inference is most often seen in the form of hypothesis testing within the NEYMAN-PEARSON PARADIGM. This paradigm involves formulating two hypotheses, the NULL HYPOTHESIS H_0

and a more general ALTERNATIVE HYPOTHESIS H_A (sometimes denoted H_1). We then design a *decision procedure* which involves collecting some data $\boldsymbol{y}$ and computing a statistic $T(\boldsymbol{y})$, or just T for short. Before collecting the data $\boldsymbol{y}$, $T(\boldsymbol{y})$ is a random variable, though we do not know its distribution because we do not know whether H_0 is true. At this point we divide the range of possible values of T into an ACCEPTANCE REGION and a REJECTION REGION. Once we collect the data, we accept the null hypothesis H_0 if T falls into the acceptance region, and reject H_0 if T falls into the rejection region.

Now, T is a random variable that will have one distribution under H_0, and another distribution under H_A. Let us denote the probability mass in the rejection region under H_0 as α, and the mass in the same region under H_A as $1 - \beta$. There are four logically possible combinations of the truth value of H_0 and our decision once we have collected $\boldsymbol{y}$:

(1)

		Our decision	
		Accept H_0	Reject H_0
H_0 is...	True	Correct decision (prob. $1 - \alpha$)	Type I error (prob. α)
	False	Type II error (prob. β)	Correct decision (prob. $1 - \beta$)

The probabilities in each row of I sum to 1, since they represent the conditional probability of our decision given the truth/falsity of H_0.

As you can see in I, there are two sets of circumstances under which we have done the right thing:

1. The null hypothesis is true, and we accept it (probability $1 - \alpha$).

2. The null hypothesis is false, and we reject it (probability $1 - \beta$).

This leaves us with two sets of circumstances under which we have made an error:

1. The null hypothesis is true, but we reject it (probability α). This by convention is called a TYPE I ERROR.

2. The null hypothesis is false, but we accept it (probability β). This by convention is called a TYPE II ERROR.

For example, suppose that a psycholinguist uses a simple visual world paradigm to examine the time course of word recognition. She presents to participants a display on which a desk is depicted on the left, and a duck is depicted on the right. Participants start with their gaze on the center of the screen, and their eye movements are recorded as they hear the word “duck”. The question at issue is whether participants’ eye gaze fall reliably more often on the duck than on the desk in the window $200 - 250$ milliseconds after the onset of “duck”, and the researcher devises a simple rule of thumb that if there are more than twice as many fixations on the duck than on the chair within this window, the null hypothesis will be rejected. Her experimental results involve 21% fixations on the duck and 9% fixations on the chair, so she rejects the null hypothesis. However, she later finds out that her computer was miscalibrated by 300 milliseconds and the participants had not even heard the onset of the word by the end of the relevant window. The researcher had committed a Type I error. (In this type of scenario, a Type I error is often called a FALSE POSITIVE, and a Type II error is often called a FALSE NEGATIVE.)

The probability α of Type I error is referred to as the SIGNIFICANCE LEVEL of the hypothesis test. In the Neyman-Pearson paradigm, T is always chosen such that its (asymptotic) distribution can be computed. The probability $1-\beta$ of *not* committing Type II error is called the POWER of the hypothesis test. There is always a trade-off between significance level and power, but the goal is to use decision procedures that have the highest possible power for a given significance level. To calculate β and thus the power, however, we need to know the true model, so determining the optimality of a decision procedure with respect to power can be tricky.

Now we'll move on to a concrete example of hypothesis testing in which we deploy some probability theory.

5.4.1 Hypothesis testing: binomial ordering preference

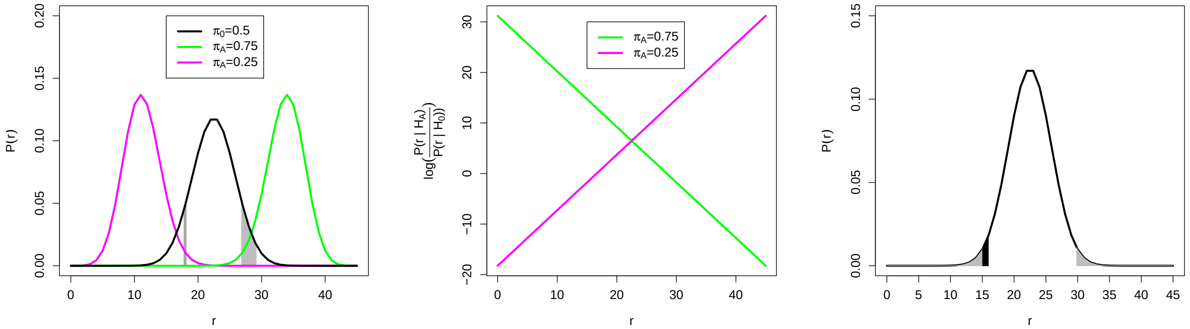
One of the simplest cases of hypothesis testing—and one that is often useful in the study of language—is the BINOMIAL TEST, which we illustrate here.

You decide to investigate the role of ultimate-syllable stress avoidance in English binomial ordering preferences by collecting from the British National Corpus 45 tokens of binomials in which *BSTR could be violated. As the test statistic T you simply choose the number of successes r in these 45 tokens. Therefore the distribution of T under the no-preference null hypothesis $H_0 : \pi = 0.5$ is simply the distribution on the number of successes r for a binomial distribution with parameters $n = 45, \pi = 0.5$. The most general natural alternative hypothesis H_A of "preference" would be that the binomial has some arbitrary preference

$$H_A : 0 \leq \pi \leq 1 \qquad (5.15)$$

Unlike the case with Bayesian hypothesis testing, we do not put a probability distribution on π under H_A. We complete our decision procedure by partitioning the possible values of T into acceptance and rejection regions. To achieve a significance level α we must choose a partitioning such that the rejection region contains probability mass of no more than α under the null hypothesis. There are many such partitions that achieve this. For example, the probability of achieving 18 successes in 45 trials is just under 5%; so is the probability of achieving at least 27 successes but not more than 29 successes. The black line in Figure 5.9a shows the probability density function for H_0, and each of the gray areas corresponds to one of these rejection regions.

However, the principle of maximizing statistical power helps us out here. Recall that when H_A is true, the power of the hypothesis test, $1-\beta$, is the probability that $T(\boldsymbol{y})$ will fall in our rejection region. The significance level α that we want to achieve, however, constrains how large our rejection region can be. To maximize the power, it therefore makes sense to choose as the rejection region that part of the range of T assigned lowest probability by H_0



(a) Probability densities over r for H_0 and two instantiations of H_A

(b) The log-ratio of probability densities under H_0 and two instantiations of H_A

(c) Standard (maximum-power) two-tailed and one-tailed rejection regions

Figure 5.9: The trade-off between significance level and power

and highest probability by H_A. Let us denote the probability mass functions for T under H_0 and H_A as $P_0(T)$ and $P_A(T)$ respectively. Figure 5.9b illustrates the tradeoff by plotting the log-ratio $\log \frac{P_A(T)}{P_0(T)}$ under two example instantiations of H_A: $\pi_A = 0.25$ and $\pi_A = 0.75$. The larger this ratio for a given possible outcome t of T, the more power is obtained by inclusion of t in the rejection region. When $\pi_A < 0.5$, the most power is obtained by filling the rejection region with the largest possible values of T. Likewise, when $\pi_A > 0.5$, the most power is obtained by filling the rejection region with the smallest possible values of T.[2] Since our H_A entertains all possible values of π, we obtain maximal power by splitting our rejection region into two symmetric halves, one on the left periphery and the other on the right periphery. In Figure 5.9c, the gray shaded area represents the largest such split region that contains less than 5% of the probability mass under H_0 (actually $\alpha = 0.0357$). If our 45 tokens do not result in at least 16 and at most 29 successes, we will reject H_0 in favor of H_A under this decision rule. This type of rule is called a TWO-TAILED TEST because the rejection region is split equally in the two tails of the distribution of T under H_0.

Another common type of alternative hypothesis would be that is a tendency to *satisfy* *BSTR. This alternative hypothesis would naturally be formulated as $H'_A : 0.5 < \pi \leq 1$. This case corresponds to the green line in Figure 5.9b; in this case we get the most power by putting our rejection region entirely on the left. The largest possible such rejection region for our example consists of the lefthand gray region plus the black region in Figure 5.9c ($\alpha = 0.0362$). This is called a ONE-TAILED TEST. The common principle which derived the form of the one-tailed and two-tailed tests alike is the idea that *one should choose the rejection region that maximizes the power of the hypothesis test if H_A is true.*

Finally, a common approach in science is not simply to choose a significance level α

[2] Although it is beyond the scope of this text to demonstrate it, this principle of maximizing statistical power leads to the same rule for constructing a rejection region regardless of the precise values entertained for π under H_A, so long as values both above and below 0.5 are entertained.

in advance and then report whether H_0 is accepted or rejected, but rather to report the lowest value of α for which H_0 would be rejected. This is what is known as the p-value of the hypothesis test. For example, if our 45 tokens resulted in 14 successes and we conducted a two-tailed test, we would compute twice the cumulative distribution function of Binom(45,0.5) at 14, which would give us an outcome of $p = 0.016$.

5.4.2 Quantifying strength of association in categorical variables

There are many situations in quantitative linguistic analysis where you will be interested in the possibility of association between two categorical variables. In this case, you will often want to represent your data as a contingency table. A 2 × 2 contingency table has the following form:

$$\begin{array}{cc|cc:c} & & \multicolumn{3}{c}{Y} \\ & & y_1 & y_2 & \\ \hline X & x_1 & n_{11} & n_{12} & n_{1*} \\ & x_2 & n_{21} & n_{22} & n_{2*} \\ \hdashline & & n_{*1} & n_{*2} & n_{**} \end{array} \tag{5.16}$$

where the n_{i*} are the marginal totals for different values of x_i across values of Y, the n_{*j} are the marginal totals for different values of y_j across values of X, and n_{**} is the grand total number of observations.

We'll illustrate the use of contingency tables with an example of quantitative syntax: the study of coordination. In traditional generative grammar, rules licensing coordination had the general form

NP → NP Conj NP

or even

X → X Conj X

encoding the intuition that many things could be coordinated with each other, but at some level every coordination should be a "combination of like categories", a constraint referred to as Conjoin Likes (Chomsky, 1965). However, this approach turned out to be of limited success in a categorical context, as demonstrated by clear violations of like-category constraints such as II below (Sag et al., 1985; Peterson, 1986):

(2) Pat is *a Republican* and *proud of it* (coordination of noun phrase with adjective phrase)

However, the preference for coordination to be between like categories is certainly strong as a *statistical tendency* (Levy, 2002). This in turn raises the question of whether the preference for coordinated constituents to be similar to one another extends to a level more fine grained than gross category structure (Levy, 2002; Dubey et al., 2008). Consider, for example, the following four coordinate noun phrases (NPs):

Example		NP1	NP2
1.	The girl and the boy	noPP	noPP
2.	[The girl from Quebec] and the boy	hasPP	noPP
3.	The girl and [the boy from Ottawa]	noPP	hasPP
4.	The girl from Quebec and the boy from Ottawa	hasPP	hasPP

Versions 1 and 4 are *parallel* in the sense that both NP conjuncts have prepositional-phrase (PP) postmodifiers; versions 2 and 3 are non-parallel. If Conjoin Likes holds at the level of NP-internal PP postmodification as a violable preference, then we might expect coordinate NPs of types 1 and 4 to be more common than would "otherwise be expected"—a notion that can be made precise through the use of contingency tables.

For example, here are patterns of PP modifications in two-NP coordinations of this type from the parsed Brown and Switchboard corpora of English, expressed as 2 × 2 contingency tables:

			NP2					NP2		
	Brown		hasPP	noPP		Switchboard		hasPP	noPP	
(3)	NP1	hasPP	95	52	147	NP1	hasPP	78	76	154
		noPP	174	946	1120		noPP	325	1230	1555
			269	998	1267			403	1306	1709

From the table you can see that in both corpora, NP1 is more likely to have a PP postmodifier when NP2 has one, and NP2 is more likely to have a PP postmodifier when NP1 has one. But we would like to go beyond that and *quantify* the strenght of the association between PP presence in NP1 on NP2. We would also like to *test for significance* of the association.

Quantifying association: odds ratios

In Section 3.3 we already saw one method of quantifying the strength of association between two binary categorical variables: COVARIANCE or CORRELATION. Another popular way way of quantifying the predictive power of a binary variable X on another binary variable Y is with the ODDS RATIO. To introduce this concept, we first introduce the overall ODDS ω^Y of y_1 versus y_2:

$$\omega^Y \stackrel{\text{def}}{=} \frac{n_{*1}}{n_{*2}} \tag{5.17}$$

Likewise, the odds ω^X of x_1 versus x_2 are $\frac{n_{1*}}{n_{2*}}$. For example, in our Brown corpus examples we have $\omega^Y = \frac{147}{1120} = 0.13$ and $\omega^X = \frac{269}{998} = 0.27$.

We further define the odds for Y if $X = x_1$ as ω_1^Y and so forth, giving us:

$$\omega_1^Y \stackrel{\text{def}}{=} \frac{n_{11}}{n_{12}} \quad \omega_2^Y \stackrel{\text{def}}{=} \frac{n_{21}}{n_{22}} \qquad\qquad \omega_1^X \stackrel{\text{def}}{=} \frac{n_{11}}{n_{21}} \quad \omega_2^X \stackrel{\text{def}}{=} \frac{n_{12}}{n_{22}}$$

If the odds of Y for $X = x_2$ are greater than the odds of Y for $X = x_1$, then the outcome of $X = x_2$ **increases** the chances of $Y = y_1$. We can express the effect of the outcome of X on the odds of Y by the **odds ratio** (which turns out to be symmetric between X, Y):

$$\mathcal{OR} \overset{\text{def}}{=} \frac{\omega_1^Y}{\omega_2^Y} = \frac{\omega_1^X}{\omega_2^X} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

An odds ratio $\mathcal{OR} = 1$ indicates no association between the variables. For the Brown and Switchboard parallelism examples:

$$\mathcal{OR}_{Brown} = \tfrac{95\times946}{52\times174} = 9.93 \quad \mathcal{OR}_{Swbd} = \tfrac{78\times1230}{325\times76} = 3.88$$

So the presence of PPs in left and right conjunct NPs seems more strongly interconnected for the Brown (written) corpus than for the Switchboard (spoken). Intuitively, this difference might be interpreted as parallelism of PP presence/absence in NPs being an aspect of stylization that is stronger in written than in spoken language.

5.4.3 Testing significance of association in categorical variables

In frequentist statistics there are several ways to test the significance of the association between variables in a two-way contingency table. Although you may not be used to thinking about these tests as the comparison of two hypotheses in form of statistical models, they are!

Fisher's exact test

Fisher's exact test applies to 2×2 contingency tables such as (5.16). It takes as H_0 the model in which all *marginal totals* are fixed, but that the individual cell totals are not—alternatively stated, that the individual outcomes of X and Y are independent. **This means that under H_0, the true underlying odds ratio $\mathcal{OR}$ is 1.** H_A is the model that the individual outcomes of X and Y are not independent. With Fisher's exact test, the test statistic T is the odds ratio $\mathcal{OR}$, which follows the HYPERGEOMETRIC DISTRIBUTION under the null hypothesis (Section B.3).

An advantage of this test is that it computes the *exact* p-value (that is, the smallest α for which H_0 would be rejected). Because of this, Fisher's exact test can be used even for very small datasets. In contrast, many of the tests we cover elsewhere in this book (including the chi-squared and likelihood-ratio tests later in this section) compute p-values that are only asymptotically correct, and are unreliable for small datasets. As an example, consider the small hypothetical parallelism dataset given in IV below:

(4)

			NP2	
		hasPP	noPP	
NP1	hasPP	3	14	26
	noPP	22	61	83
		34	75	109

The odds ratio is 2.36, and Fisher's exact test gives a p-value of 0.07. If we were to see twice the data in the exact same proportions, the odds ratio would stay the same, but the significance of Fisher's exact test for non-independence would increase.

Chi-squared test

This is probably the best-known contingency-table test. It is very general and can be applied to arbitrary N-cell tables, if you have a model with k parameters that predicts expected values E_{ij} for all cells. For the chi-squared test, the test statistic is Pearson's X^2:

$$X^2 = \sum_{ij} \frac{[n_{ij} - E_{ij}]^2}{E_{ij}} \tag{5.18}$$

In the chi-squared test, H_A is the model that each cell in the table has its own parameter p_i in one big multinomial distribution. When the expected counts in each cell are large enough (the generally agreed lower bound is ≥ 5), the X^2 statistic is approximately distributed as a chi-squared (χ^2) random variable with $N - k - 1$ degrees of freedom (Section B.4). The χ^2 distribution is asymmetric and the rejection region is always placed in the right tail of the distribution (see Section B.4), so we can calculate the p-value by subtracting from one the value of the cumulative distribution function for the observed X^2 test statistic.

The most common way of using Pearson's chi-squared test is to test for the independence of two factors in a two-way contingency table. Take a $k \times l$ two-way table of the form:

	y_1	y_2	$\cdots$	y_l	
x_1	n_{11}	n_{12}	$\cdots$	n_{1l}	n_{1*}
x_2	n_{21}	n_{22}	$\cdots$	n_{2l}	n_{2*}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_l	n_{k1}	n_{k2}	$\cdots$	n_{kl}	n_{k*}
	n_{*1}	n_{*2}	$\cdots$	n_{*l}	n

Our null hypothesis is that the x_i and y_i are independently distributed from one another. By the definition of probabilistic independence, that means that H_0 is:

$$P(x_i, y_j) = P(x_i)P(y_j)$$

In the chi-squared test we use the relative-frequency (and hence maximum-likelihood; Section 4.3.1) estimates of the marginal probability that an observation will fall in each row or column: $\hat{P}(x_i) = \frac{n_{i*}}{n}$ and $\hat{P}(y_j) = \frac{n_{*j}}{n}$. This gives us the formula for the expected counts in Equation (5.18):

$$E_{ij} = nP(x_i)P(y_j)$$

Example: For the Brown corpus data in III, we have

$$P(x_1) = \frac{147}{1267} = 0.1160 \qquad P(y_1) = \frac{269}{1267} = 0.2123 \tag{5.19}$$

$$P(x_2) = \frac{1120}{1267} = 0.8840 \qquad P(y_2) = \frac{998}{1267} = 0.7877 \tag{5.20}$$

giving us

$$E_{11} = 31.2 \quad E_{12} = 115.8 \qquad E_{21} = 237.8 \quad E_{21} = 882.2 \tag{5.21}$$

Comparing with III, we get

$$X^2 = \frac{(95-31.2)^2}{31.2} + \frac{(52-115.8)^2}{115.8} + \frac{(174-237.8)^2}{237.8} + \frac{(946-882.2)^2}{882.2} \tag{5.22}$$

$$= 187.3445 \tag{5.23}$$

We had 2 parameters in our model of independence, and there are 4 cells, so X^2 is distributed as χ^2_1 (since $4-2-1=1$). The cumulative distribution function of χ^2_1 at 187.3 is essentially 1, so the p-value is vanishingly small; by any standards, the null hypothesis can be confidently rejected.

Example with larger data table:

	NP PP	NP NP	NP other
gave	17	79	34
paid	14	4	9
passed	4	1	16

It is worth emphasizing, however, that the chi-squared test is not reliable when expected counts in some cells are very small. For the low-count table in IV, for example, the chi-squared test yields a significance level of $p = 0.038$. Fisher's exact test is the gold standard here, revealing that the chi-squared test is too aggressive in this case.

5.4.4 Likelihood ratio test

With this test, the statistic you calculate for your data $\boldsymbol{y}$ is the LIKELIHOOD RATIO

$$\Lambda^* = \frac{\max \mathrm{Lik}_{H_0}(\boldsymbol{y})}{\max \mathrm{Lik}_{H_A}(\boldsymbol{y})} \tag{5.24}$$

—that is, the ratio of the data likelihood under the MLE in H_0 to the data likelihood under the MLE in H_A. This requires that you explicitly formulate H_0 and H_A, since you need to find the MLEs and the corresponding likelihoods. The quantity

$$G^2 \overset{\text{def}}{=} -2 \log \Lambda^* \tag{5.25}$$

is sometimes called the DEVIANCE, and it is approximately chi-squared distributed (Section B.4) with degrees of freedom equal to the difference in the number of free parameters in H_A and H_0. (This test is also unreliable when expected cell counts are low, as in < 5.)

The likelihood-ratio test gives similar results to the chi-squared for contingency tables, but is more flexible because it allows the comparison of arbitrary nested models. We will see the likelihood-ratio test used repeatedly in later chapters.

Example: For the Brown corpus data above, let H_0 be the model of independence between NP1 and NP2 with respective success parameters π_1 and π_2, and H_A be the model of full non-independence, in which each complete outcome $\langle x_i, y_j \rangle$ can have its own probability π_{ij} (this is sometimes called the SATURATED MODEL). We use maximum likelihood to fit each model, giving us for H_0:

$$\pi_1 = 0.116 \qquad \qquad \pi_2 = 0.212$$

and for H_A:

$$\pi_{11} = 0.075 \quad \pi_{12} = 0.041 \qquad \qquad \pi_{21} = 0.137 \quad \pi_{22} = 0.747$$

We calculate G^2 as follows:

$$\begin{aligned}
-2 \log \Lambda^* &= -2 \log \frac{(\pi_1\pi_2)^{95}(\pi_1(1-\pi_2))^{52}((1-\pi_1)\pi_2)^{174}((1-\pi_1)(1-\pi_2))^{946}}{\pi_{11}^{95}\pi_{12}^{52}\pi_{21}^{174}\pi_{22}^{946}} \\
&= -2\left[95 \log(\pi_1\pi_2) + 52 \log(\pi_1(1-\pi_2)) + 174 \log((1-\pi_1)\pi_2) + 946 \log((1-\pi_1)(1-\pi_2))\right. \\
&\quad \left. -95\pi_{11} - 52\pi_{12} - 174\pi_{21} - 946\pi_{22}\right] \\
&= 151.6
\end{aligned}$$

H_0 has two free parameters, and H_A has three free parameters, so G^2 should be approximately distributed as χ^2_1. Once again, the cumulative distribution function of χ^2_1 at 151.6 is essentially 1, so the p-value of our hypothesis test is vanishingly small.

5.5 Exercises

Exercise 5.1

Would the Bayesian hypothesis-testing results of Section 5.2 be changed at all if we did not consider the data as summarized by the number of successes and failures, and instead used the likelihood of the specific sequence `HHTHTH` instead? Why?

Exercise 5.2

For Section 5.2, compute the posterior probabilities of H_1, H_2, and H_3 in a situation where all three hypotheses are entertained with prior probabilities $P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$.

Exercise 5.3

Recompute the Bayesian hypothesis tests, computing both posterior probabilities and Bayes Factors, of Section 5.2 (H_1 vs. H_2 and H_1 vs. H_3) for the same data replicated twice – that is, the observations `SSVSVSSSVSVS`. Are the preferred hypotheses the same as for the original computations in Section 5.2? What about for the data replicated three times?

Exercise 5.4: Phoneme discrimination for Gaussians of unequal variance and prior probabilities.

1. Plot the optimal-response phoneme discrimination curve for the /b/–/p/ VOT contrast when the VOT of each category is realized as a Gaussian and the Gaussians have equal variances $\sigma_{\mathrm{b}} = 12$, different means $\mu_{\mathrm{b}} = 0, \mu_{\mathrm{p}} = 50$, and different prior probabilities: $P(/b/) = 0.25, P(/p/) = 0.75$. How does this curve look compared with that in Figure **??**?

2. Plot the optimal-response phoneme discrimination curve for the /b/–/p/ VOT contrast when the Gaussians have equal prior probabilities but both unequal means and unequal variances: $\mu_{\mathrm{b}} = 0, \mu_{\mathrm{p}} = 50, \sigma_{\mathrm{b}} = 8, \sigma_{\mathrm{p}} = 14$.

3. Propose an experiment along the lines of Clayards et al. (2008) testing the ability of listeners to learn category-specific variances and prior probabilities and use them in phoneme discrimination.

4. It is in fact the case that naturalistic VOTs in English have larger variance /p/ than for /b/ [**TODO: get reference for this**]. For part 2 of this question, check what the model predicts as VOT extends to very large negative values (e.g., -200ms). There is some counter-intuitive behavior: what is it? What does this counter-intuitive behavior tell us about the limitations of the model we've been using?

Exercise 5.5

Use simulation to check that the theoretical confidence interval based on the t distribution for normally distributed data in Section 5.3 really works.

Exercise 5.6

For a given choice of α, is the procedure denoted in Equation (5.14) the only frequentist confidence interval that can be constructed for μ for normally distributed data?

Exercise 5.7: Hypothesis testing: philosophy.

You surreptitiously observe an obsessed linguist compulsively searching a corpus for binomials of the form *pepper and salt* (`P`) or *salt and pepper* (`S`). He collects twenty examples, obtaining the sequence

```
SPSSSPSPSSSSPPSSSSSP
```

1. The linguist's research assistant tells you that the experiment was to obtain twenty examples and record the number of `P`'s obtained. Can you reject the no-preference null hypothesis H_0 at the $\alpha = 0.05$ level?

2. The next day, the linguist tells you in class that she purposely misled the research assistant, and the actual experiment was to collect tokens from the corpus until six `P` examples were obtained and then stop. Does this new information affect the p-value with which you can reject the null hypothesis?

3. The linguist writes up her research results and sends them to a prestigious journal. The editor sends this article to two Bayesian reviewers. Both reviewers argue that this mode of hypothesis testing is ridiculous, and that a Bayesian hypothesis test should be made. Reviewer A suggests that the null hypothesis H_0 of $\pi = 0.5$ should be compared with the alternative hypothesis H_1 of $\pi = 0.25$, and the two hypotheses should be given equal prior probability.

   (a) What is the posterior probability of H_0 given the binomials data? Does the criteria by which the scientist decided how many binomials to collect affect the conclusions of a Bayesian hypothesis test? **Hint:** if $\frac{P(H_0|\vec{x})}{P(H_1|\vec{x})} = a$, then

   $$P(H_0|\vec{x}) = \frac{a}{1+a}$$

   because $P(H_0|\vec{x}) = 1 - P(H_1|\vec{x})$.

   (b) Reviewer B suggests that H_1 should be $\pi = 0.4$ instead. What is the posterior probability of H_0 under this Bayesian comparison?

Exercise 5.8: Bayesian confidence intervals.

The `binom.bayes()` function in the `binom` package permits the calculation of Bayesian confidence intervals over π for various numbers of successes x, total trials n, and a and b (specified as `prior.shape1` and `prior.shape2` respectively—but `prior.shape1`$=a-1$ and `prior.shape2`$=b-1$). Install the `binom` package with the command

```
install.packages("binom")
```

and then use the `binom.bayes()` function to plot the size of a 95% confidence interval on π as a function of the total number of trials n, ranging from 10 to 10000 (in multiples of 10), where 70% of the trials are always successes. (Hold a and b constant, as values of your choice.)

Exercise 5.9: Contextual dependency in phoneme sequences.

1. Reproduce the Bayesian hypothesis test of Section 5.2.3 for uniform priors with the following sequences, computing $P(H_1|\boldsymbol{y})$ for each sequence. One of the three sequences was generated from a context-independent distribution, whereas the other two were generated from context-dependent distributions. Which one is most strongly indicated by the Bayesian hypothesis test to be generated from the context-independent distribution?

   (a) A B A B B B B B A B B B A A A B B B B B B

   (b) B A B B A B A B A A B A B A B B A A B A B

   (c) B B B B A A A A B B B B B A A A A B B B B

2. Although we put uniform priors on all success parameters in Section 5.2.3, in the contextual-dependence model it makes more sense to have a SPARSE prior—that is, one that favors strong preferences for some phonemes over others after each type of context. A sparse beta prior is one for which at least one α_i parameter is low (< 1). Revise the model so that the prior on $\pi_\emptyset$ remains uniform, but that π_A and π_B have symmetric $\langle\alpha,\alpha\rangle$ priors (and give both π_A and π_B the same prior). Plot the posterior probabilities $P(H_1|\boldsymbol{y})$ for sequences (i-iii) as a function of α for $0 < \alpha \leq 1$. What is the value of α for which the context-independent sequence is most strongly differentiated from the context-dependent sequences (i.e. the differences in $P(H_1|\boldsymbol{y})$ between sequence pairs are greatest)?

Exercise 5.10: Phoneme categorization.

1. Plot the Bayesian phoneme discrimination curve for /b/–/p/ discrimination with $\mu_{\text{b}} = 0, \mu_{\text{p}} = 50, \sigma_{\text{b}} = 5, \sigma_{\text{p}} = 10$.

2. Write out the general formula for Bayesian phoneme discrimination when VOTs in the two categories are normally distributed with unequal variances. Use algebra to simplify it into the form $P(/\text{b}/|x) = \frac{1}{1+\dots}$. Interpret the formula you obtained.

Exercise 5.11

Frequentist confidence intervals.

In this problem you'll be calculating some frequentist confidence intervals to get a firmer sense of exactly what they are.

1. The `english` dataset in `languageR` has lexical-decision and naming reaction times (`RTlexdec` and `RTnaming` respectively) for 2197 English words. Plot histograms of the mean RT of each item. Calculate 95% frequentist confidence intervals for lexical-decision and naming times respectively, as described in Lecture 7, Section 2. Which experimental method gives tighter confidence intervals on mean RT?

2. The `t.test()` function, when applied to a set of data, returns a list whose component `conf.int` is the upper and lower bounds of a 95% confidence interval:

   ```
   > x <- rnorm(100,2)
   > t.test(x)$conf.int

   [1] 1.850809 2.241884
   attr(,"conf.level")
   [1] 0.95
   ```

   Show that the procedure used in Section 5.3 gives the same results as using `t.test()` for the confidence intervals for the English lexical decision and naming datasets.

3. Not all confidence intervals generated from an "experiment" are the same size. For "experiments" consisting of 10 observations drawn from a standard normal distribution, use `R` to calculate a histogram of lengths of 95% confidence intervals on the mean. What is the shape of the distribution of confidence interval lengths? For this problem, feel free to use `t.test()` to calculate confidence intervals.

Exercise 5.12: Comparing two samples.

In class we covered confidence intervals and the one-sample t-test. This approach allows us to test whether a dataset drawn from a(n approximately) normally distributed population departs significantly from has a particular mean. More frequently, however, we are interested in comparing two datasets $\boldsymbol{x}$ and $\boldsymbol{y}$ of sizes $n_{\boldsymbol{x}}$ and $n_{\boldsymbol{y}}$ respectively, and inferring whether or not they are drawn from the same population. For this purpose, the TWO-SAMPLE t-TEST is appropriate.[3]

1. **Statistical power.** Suppose you have two populations and you can collect n total observations from the two populations. Intuitively, how should you distribute your observations among the two populations to achieve the greatest STATISTICAL POWER in a test that the two populations follow the same distribution?

[3] For completeness, the statistic that is t-distributed for the two-sample test is:

$$\frac{\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}}{\sqrt{\sigma^2 \left(\frac{1}{n_{\boldsymbol{x}}} + \frac{1}{n_{\boldsymbol{y}}} \right)}}$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are the sample means; in general, the variance σ^2 is unknown and is estimated as $\hat{\sigma}^2 = \frac{\sum_i (x_i - \bar{\boldsymbol{x}})^2 + \sum_i (y_i - \bar{\boldsymbol{y}})^2}{N-2}$ where $N = n_{\boldsymbol{x}} + n_{\boldsymbol{y}}$.

2. Check your intuitions. Let $n = 40$ and consider all possible values of $n_{\boldsymbol{x}}$ and $n_{\boldsymbol{y}}$ (note that $n_{\boldsymbol{y}} = n - n_{\boldsymbol{x}}$). For each possible value, run 1000 experiments where the two populations are $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1,1)$. Plot the power of the two-sample t-test at the $\alpha = 0.05$ level as a function of $n_{\boldsymbol{x}}$.

3. **Paired t-tests.** Sometimes a dataset can be naturally thought of as consisting of pairs of measurements. For example, if a phonetician measured voice onset time for the syllables [ba] and [bo] for many different speakers, the data could be grouped into a matrix of the form

	Syllable	
Speaker	[ba]	[bo]
1	x_{11}	x_{12}
2	x_{21}	x_{22}
	$\vdots$	

   If we wanted to test whether the voice onset times for [ba] and [bo] came from the same distribution, we could simply perform a two-sample t-test on the data in column 1 versus the data in column 2.

   On the other hand, this doesn't take into account the systematic differences in voice-onset time that may hold across speakers. What we might really want to do is test whether the *differences* between x_{i1} and x_{i2} are clustered around zero—which would indicate that the two data vectors probably do come from the same population—or around some non-zero number. This comparison is called a PAIRED t-TEST.

   The file `spillover_word_rts` contains the average reading time (in milliseconds) of the second "spillover" word after a critical manipulation in self-paced reading experiment, for 52 sentence pairs of the form:

   > The children went outside to **play** early in the afternoon. (Expected)
   > The children went outside to **chat** early in the afternoon. (Unexpected)

   In a separate sentence completion study, 90% of participants completed the sentence

   > The children went outside to __

   with the word *play*, making this the Expected condition. In these examples, the word whose reading time (RT) is measured would be *in*, as it appears two words after the critical word (in bold).

   (a) Use paired and unpaired t-tests to test the hypothesis that mean reading times at the second spillover word differ significantly in the Expected and Unexpected conditions. Which test leads to a higher significance value?

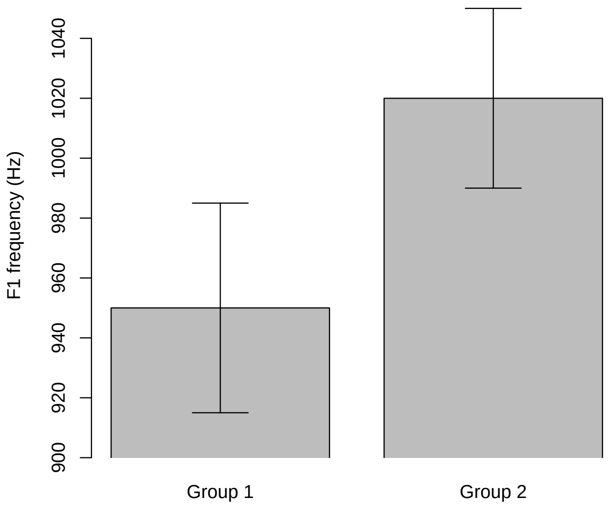


Figure 5.10: A bar plot with non-overlapping standard error bars

(b) Calculate the correlation between the RTs for the unexpected and expected conditions of each item. Intuitively, should higher correlations lead to an increase or drop in statistical power for the paired test over the unpaired test? Why?

Exercise 5.13: Non-overlapping standard errors.$^\heartsuit$

You and your colleague measure the F1 formant frequency in pronunciation of the vowel [a] for two groups of 50 native speakers of English, one measurement for each speaker. The means and standard errors of these measurements are shown in Figure 5.10. Your colleague states, "we can be fairly confident in inferring that the two groups have significantly different mean F1 formant frequencies. As a rule of thumb, when you have a reasonably large number of measurements in each group and the standard error bars between the two groups are non-overlapping, we can reject the null hypothesis that the group means are the same at the $p < 0.05$ level." Is your colleague's rule of thumb correct?

Exercise 5.14: Log odds ratio versus correlation.

Are (log) odds ratios any different from correlation coefficients? Plot the relationship between log odds ratio and correlation coefficient for a number of different 2×2 contingency tables. (If you want to take a sampling-based approach to exploring the space of possible contingency tables, you might use the Dirichlet distribution—see Section B.8—to randomly generate sets of cell probabilities).

Exercise 5.15: Contingency tables.

1. Bresnan et al. (2007) conducted a detailed analysis of the dative alternation, as in the example below:

   The actress gave **the toys** *to the children.* (Prepositional Object, PO)
   The actress gave *the children* **the toys**. (Double Object, DO)

   The analysis was based on data obtained from the parsed Switchboard corpus (Godfrey et al., 1992).[4] Irrespective of which alternate was used, it turns out that there are correlations among the properties of the theme (**the toys**) and the recipient (*the children*).

   Definiteness and animacy are often found to be correlated. Look at the relationship between animacy and definiteness of (1) the theme, and (2) the recipient within this dataset, constructing contingency tables and calculating the odds ratios in each case. For which semantic role are definiteness and animacy more strongly associated? Why do you think this might be the case? (Note that organizations, animals, intelligent machines, and vehicles were considered animate for this coding scheme (Zaenen et al., 2004)).

2. The language Warlpiri, one of the best-studied Australian Aboriginal languages, is characterized by extremely free word order and heavy use of morphological cues as to the grammatical function played by each word in the clause (i.e. case marking). Below, for example, the *ergative* case marking (ERG) on the first word of the sentence identifies it as the subject of the sentence:

Ngarrka-	ngku	ka	wawirri	panti-	rni.	(Hale, 1983)
man	ERG	AUX	kangaroo	spear	NONPAST	

   "The man is spearing the kangaroo".

   In some dialects of Warlpiri, however, using the ergative case is not obligatory. Note that there would be a semantic ambiguity if the case marking were eliminated from the first word, because neither *man* nor *kangaroo* would have case marking to indicate its grammatical relationship to the verb *spear*. O'Shannessy (2009) carried out a study of word order and case marking variation in sentences with transitive main clauses and overt subjects ("A" arguments in the terminology of Dixon, 1979) in elicited story descriptions by Warlpiri speakers. Her dataset includes annotation of speaker age, whether the transitive subject was animate, whether the transitive subject had ergative case marking, whether the sentence had an animate object (Dixon's "O" argument), whether that object was realized overtly, and whether the word order of the sentence was subject-initial.[5]

[4] The dataset can be found in `R`'s `languageR` package; there it is a data frame named `dative`.

[5] O'Shannessy's dataset can be found in `R`'s `languageR` package under the name `warlpiri`.

(a) Does *subject animacy* have a significant association (at the $\alpha = 0.05$ level) with ergative case marking? What about *word order* (whether the subject was sentence-initial)?

(b) Which of the following variables have an effect on whether subject animacy and word order have a significant association with use of ergative case marking? (For each of the below variables, split the dataset in two and do a statistical test of association on each half.)

overtness of object
age group