# 6 *Frequentist statistics*

## 6.1 Introduction

The approach to statistical inference that we described in Chapter 5 is known as Bayesian statistics. Perhaps surprisingly, this is considered controversial by some people, whereas the application of Bayes rule to non-statistical problems — such as medical diagnosis (Section 2.2.3.1), spam filtering (Section 3.4.4.1), or airplane tracking (Section 18.2.1) — is not controversial. The reason for the objection has to do with a misguided distinction between parameters of a statistical model and other kinds of unknown quantities.[1]

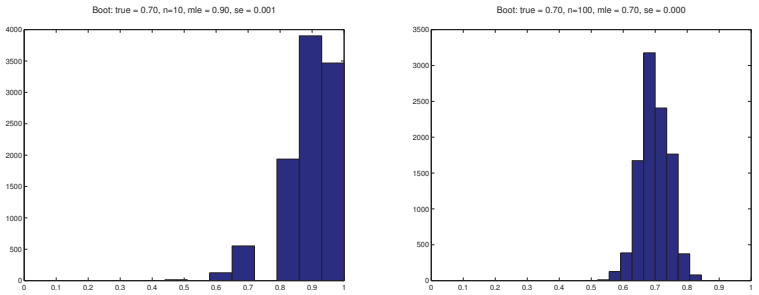
Attempts have been made to devise approaches to statistical inference that avoid treating parameters like random variables, and which thus avoid the use of priors and Bayes rule. Such approaches are known as **frequentist statistics**, **classical statistics** or **orthodox statistics**. Instead of being based on the posterior distribution, they are based on the concept of a sampling distribution. This is the distribution that an estimator has when applied to multiple data sets sampled from the true but unknown distribution; see Section 6.2 for details. It is this notion of variation across repeated trials that forms the basis for modeling uncertainty used by the frequentist approach.

By contrast, in the Bayesian approach, we only ever condition on the actually observed data; there is no notion of repeated trials. This allows the Bayesian to compute the probability of one-off events, as we discussed in Section 2.1. Perhaps more importantly, the Bayesian approach avoids certain paradoxes that plague the frequentist approach (see Section 6.6). Nevertheless, it is important to be familiar with frequentist statistics (especially Section 6.5), since it is widely used in machine learning.

## 6.2 Sampling distribution of an estimator

In frequentist statistics, a parameter estimate $\hat{\boldsymbol{\theta}}$ is computed by applying an **estimator** $\delta$ to some data $\mathcal{D}$, so $\hat{\boldsymbol{\theta}} = \delta(\mathcal{D})$. The parameter is viewed as fixed and the data as random, which is the exact opposite of the Bayesian approach. The uncertainty in the parameter estimate can be measured by computing the **sampling distribution** of the estimator. To understand this

1. Parameters are sometimes considered to represent true (but unknown) physical quantities, which are therefore not random. However, we have seen that it is perfectly reasonable to use a probability distribution to represent one's uncertainty about an unknown constant.




(a) (b)

**Figure 6.1** A bootstrap approximation to the sampling distribution of $\hat{\theta}$ for a Bernoulli distribution. We use $B = 10,000$ bootstrap samples. The $N$ datacases were generated from $\text{Ber}(\theta = 0.7)$. (a) MLE with $N = 10$. (b) MLE with $N = 100$. Figure generated by `bootstrapDemoBer`.

concept, imagine sampling many different data sets $\mathcal{D}^{(s)}$ from some true model, $p(\cdot|\theta^*)$, i.e., let $\mathcal{D}^{(s)} = \{x_i^{(s)}\}_{i=1}^N$, where $x_i^s \sim p(\cdot|\theta^*)$, and $\theta^*$ is the true parameter. Here $s = 1 : S$ indexes the sampled data set, and $N$ is the size of each such dataset. Now apply the estimator $\hat{\theta}(\cdot)$ to each $\mathcal{D}^{(s)}$ to get a set of estimates, $\{\hat{\theta}(\mathcal{D}^{(s)})\}$. As we let $S \to \infty$, the distribution induced on $\hat{\theta}(\cdot)$ is the sampling distribution of the estimator. We will discuss various ways to use the sampling distribution in later sections. But first we sketch two approaches for computing the sampling distribution itself.

### 6.2.1 Bootstrap

The **bootstrap** is a simple Monte Carlo technique to approximate the sampling distribution. This is particularly useful in cases where the estimator is a complex function of the true parameters.

The idea is simple. If we knew the true parameters $\theta^*$, we could generate many (say $S$) fake datasets, each of size $N$, from the true distribution, $x_i^s \sim p(\cdot|\theta^*)$, for $s = 1 : S$, $i = 1 : N$. We could then compute our estimator from each sample, $\hat{\theta}^s = f(x_{1:N}^s)$ and use the empirical distribution of the resulting samples as our estimate of the sampling distribution. Since $\theta$ is unknown, the idea of the **parametric bootstrap** is to generate the samples using $\hat{\theta}(\mathcal{D})$ instead. An alternative, called the **non-parametric bootstrap**, is to sample the $x_i^s$ (with replacement) from the original data $\mathcal{D}$, and then compute the induced distribution as before. Some methods for speeding up the bootstrap when applied to massive data sets are discussed in (Kleiner et al. 2011).

Figure 6.1 shows an example where we compute the sampling distribution of the MLE for a Bernoulli using the parametric bootstrap. (Results using the non-parametric bootstrap are essentially the same.) We see that the sampling distribution is asymmetric, and therefore quite far from Gaussian, when $N = 10$; when $N = 100$, the distribution looks more Gaussian, as theory suggests (see below).

A natural question is: what is the connection between the parameter estimates $\hat{\theta}^s = \hat{\theta}(x_{1:N}^s)$ computed by the bootstrap and parameter values sampled from the posterior, $\theta^s \sim p(\cdot|\mathcal{D})$?

Conceptually they are quite different. But in the common case that that the prior is not very strong, they can be quite similar. For example, Figure 6.1(c-d) shows an example where we compute the posterior using a uniform Beta(1,1) prior, and then sample from it. We see that the posterior and the sampling distribution are quite similar. So one can think of the bootstrap distribution as a "poor man's" posterior; see (Hastie et al. 2001, p235) for details.

However, perhaps surprisingly, bootstrap can be slower than posterior sampling. The reason is that the bootstrap has to fit the model $S$ times, whereas in posterior sampling, we usually only fit the model once (to find a local mode), and then perform local exploration around the mode. Such local exploration is usually much faster than fitting a model from scratch.

### 6.2.2 Large sample theory for the MLE *

In some cases, the sampling distribution for some estimators can be computed analytically. In particular, it can be shown that, under certain conditions, as the sample size tends to infinity, the sampling distribution of the MLE becomes Gaussian. Informally, the requirement for this result to hold is that each parameter in the model gets to "see" an infinite amount of data, and that the model be identifiable. Unfortunately this excludes many of the models of interest to machine learning. Nevertheless, let us assume we are in a simple setting where the theorem holds.

The center of the Gaussian will be the MLE $\hat{\boldsymbol{\theta}}$. But what about the variance of this Gaussian? Intuitively the variance of the estimator will be (inversely) related to the amount of curvature of the likelihood surface at its peak. If the curvature is large, the peak will be "sharp", and the variance low; in this case, the estimate is "well determined". By contrast, if the curvature is small, the peak will be nearly "flat", so the variance is high.

Let us now formalize this intuition. Define the **score function** as the gradient of the log likelihood evaluated at some point $\hat{\boldsymbol{\theta}}$:

$$\mathbf{s}(\hat{\boldsymbol{\theta}}) \triangleq \nabla \log p(\mathcal{D}|\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}} \tag{6.1}$$

Define the **observed information matrix** as the gradient of the negative score function, or equivalently, the Hessian of the NLL:

$$\mathbf{J}(\hat{\boldsymbol{\theta}}(\mathcal{D})) \triangleq -\nabla \mathbf{s}(\hat{\boldsymbol{\theta}}) = -\nabla^2_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}} \tag{6.2}$$

In 1D, this becomes

$$J(\hat{\theta}(\mathcal{D})) = -\frac{d}{d\theta^2} \log p(\mathcal{D}|\theta)|_{\hat{\theta}} \tag{6.3}$$

This is just a measure of curvature of the log-likelihood function at $\hat{\theta}$.

Since we are studying the sampling distribution, $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is a set of random variables. The **Fisher information matrix** is defined to be the expected value of the observed information matrix:[2]

$$\mathbf{I}_N(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}^*) \triangleq \mathbb{E}_{\boldsymbol{\theta}^*}\left[\mathbf{J}(\hat{\boldsymbol{\theta}}|\mathcal{D})\right] \tag{6.4}$$

2. This is not the usual definition, but is equivalent to it under standard assumptions. More precisely, the standard definition is as follows (we just give the scalar case to simplify notation): $I(\hat{\theta}|\theta^*) \triangleq \mathrm{var}_{\theta^*}\left[\frac{d}{d\theta} \log p(X|\theta)|_{\hat{\theta}}\right]$, that is, the variance of the score function. If $\hat{\theta}$ is the MLE, it is easy to see that $\mathbb{E}_{\theta^*}\left[\frac{d}{d\theta} \log p(X|\theta)|_{\hat{\theta}}\right] = 0$ (since

where $\mathbb{E}_{\boldsymbol{\theta}^*}[\mathbf{f}(\mathcal{D})] \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}(\mathbf{x}_i)p(\mathbf{x}_i|\boldsymbol{\theta}^*)$ is the expected value of the function $\mathbf{f}$ when applied to data sampled from $\boldsymbol{\theta}^*$. Often $\boldsymbol{\theta}^*$, representing the "true parameter" that generated the data, is assumed known, so we just write $\mathbf{I}_N(\hat{\boldsymbol{\theta}}) \triangleq \mathbf{I}_N(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}^*)$ for short. Furthermore, it is easy to see that $\mathbf{I}_N(\hat{\boldsymbol{\theta}}) = N\mathbf{I}_1(\hat{\boldsymbol{\theta}})$, because the log-likelihood for a sample of size $N$ is just $N$ times "steeper" than the log-likelihood for a sample of size 1. So we can drop the 1 subscript and just write $\mathbf{I}(\hat{\boldsymbol{\theta}}) \triangleq \mathbf{I}_1(\hat{\boldsymbol{\theta}})$. This is the notation that is usually used.

Now let $\hat{\boldsymbol{\theta}} \triangleq \hat{\boldsymbol{\theta}}_{mle}(\mathcal{D})$ be the MLE, where $\mathcal{D} \sim \boldsymbol{\theta}^*$. It can be shown that

$$\hat{\boldsymbol{\theta}} \rightarrow \mathcal{N}(\boldsymbol{\theta}^*, \mathbf{I}_N(\boldsymbol{\theta}^*)^{-1}) \tag{6.5}$$

as $N \rightarrow \infty$ (see e.g., (Rice 1995, p265) for a proof). We say that the sampling distribution of the MLE is **asymptotically normal**.

What about the variance of the MLE, which can be used as some measure of confidence in the MLE? Unfortunately, $\boldsymbol{\theta}^*$ is unknown, so we can't evaluate the variance of the sampling distribution. However, we can approximate the sampling distribution by replacing $\boldsymbol{\theta}^*$ with $\hat{\boldsymbol{\theta}}$. Consequently, the approximate **standard errors** of $\hat{\theta}_k$ are given by

$$\hat{se}_k \triangleq \mathbf{I}_N(\hat{\boldsymbol{\theta}})_{kk}^{-\frac{1}{2}} \tag{6.6}$$

For example, from Equation 5.60 we know that the Fisher information for a binomial sampling model is

$$I(\theta) = \frac{1}{\theta(1-\theta)} \tag{6.7}$$

So the approximate standard error of the MLE is

$$\hat{se} = \frac{1}{\sqrt{I_N(\hat{\theta})}} = \frac{1}{\sqrt{NI(\hat{\theta})}} = \left(\frac{\hat{\theta}(1-\hat{\theta})}{N}\right)^{\frac{1}{2}} \tag{6.8}$$

where $\hat{\theta} = \frac{1}{N}\sum_i X_i$. Compare this to Equation 3.27, which is the posterior standard deviation under a uniform prior.

## 6.3 Frequentist decision theory

In frequentist or classical decision theory, there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic way of deriving an optimal estimator, unlike the Bayesian case. Instead, in the frequentist approach, we are free to choose any estimator or decision procedure $\delta : \mathcal{X} \rightarrow \mathcal{A}$ we want.[3]

---

the gradient must be zero at a maximum), so the variance reduces to the expected square of the score function: $I(\hat{\theta}|\theta^*) = \mathbb{E}_{\theta^*}\left[(\frac{d}{d\theta}\log p(X|\theta))^2\right]$. It can be shown (e.g., (Rice 1995, p263)) that $\mathbb{E}_{\theta^*}\left[(\frac{d}{d\theta}\log p(X|\theta))^2\right] = -\mathbb{E}_{\theta^*}\left[\frac{d^2}{d\theta^2}\log p(X|\theta)\right]$, so now the Fisher information reduces to the expected second derivative of the NLL, which is a much more intuitive quantity than the variance of the score.

3. In practice, the frequentist approach is usually only applied to one-shot statistical decision problems — such as classification, regression and parameter estimation — since its non-constructive nature makes it difficult to apply to sequential decision problems, which adapt to data online.

Having chosen an estimator, we define its expected loss or **risk** as follows:

$$R(\theta^*, \delta) \triangleq \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)}\left[L(\theta^*, \delta(\tilde{\mathcal{D}}))\right] = \int L(\theta^*, \delta(\tilde{\mathcal{D}}))p(\tilde{\mathcal{D}}|\theta^*)d\tilde{\mathcal{D}} \tag{6.9}$$

where $\tilde{\mathcal{D}}$ is data sampled from "nature's distribution", which is represented by parameter $\theta^*$. In other words, the expectation is wrt the sampling distribution of the estimator. Compare this to the Bayesian posterior expected loss:

$$\rho(a|\mathcal{D}, \pi) \triangleq \mathbb{E}_{p(\theta|\mathcal{D},\pi)}\left[L(\theta, a)\right] = \int_{\Theta} L(\theta, \mathbf{a})p(\theta|\mathcal{D}, \pi)d\theta \tag{6.10}$$

We see that the Bayesian approach averages over $\theta$ (which is unknown) and conditions on $\mathcal{D}$ (which is known), whereas the frequentist approach averages over $\tilde{\mathcal{D}}$ (thus ignoring the observed data), and conditions on $\theta^*$ (which is unknown).

Not only is the frequentist definition unnatural, it cannot even be computed, because $\theta^*$ is unknown. Consequently, we cannot compare different estimators in terms of their frequentist risk. We discuss various solutions to this below.

### 6.3.1 Bayes risk

How do we choose amongst estimators? We need some way to convert $R(\boldsymbol{\theta}^*, \delta)$ into a single measure of quality, $R(\delta)$, which does not depend on knowing $\boldsymbol{\theta}^*$. One approach is to put a prior on $\theta^*$, and then to define **Bayes risk** or **integrated risk** of an estimator as follows:

$$R_B(\delta) \triangleq \mathbb{E}_{p(\boldsymbol{\theta}^*)}\left[R(\boldsymbol{\theta}^*, \delta)\right] = \int R(\boldsymbol{\theta}^*, \delta)p(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^* \tag{6.11}$$

A **Bayes estimator** or **Bayes decision rule** is one which minimizes the expected risk:

$$\delta_B \triangleq \operatorname*{argmin}_{\delta} R_B(\delta) \tag{6.12}$$

Note that the integrated risk is also called the **preposterior risk**, since it is before we have seen the data. Minimizing this can be useful for experiment design.

We will now prove a very important theorem, that connects the Bayesian and frequentist approaches to decision theory.

**Theorem 6.3.1.** *A Bayes estimator can be obtained by minimizing the posterior expected loss for each* $\mathbf{x}$.

*Proof.* By switching the order of integration, we have

$$\begin{aligned}
R_B(\delta) &= \int \left[\sum_{\mathbf{x}}\sum_{y} L(y, \delta(\mathbf{x}))p(\mathbf{x}, y|\boldsymbol{\theta}^*)\right] p(\boldsymbol{\theta}^*)d\boldsymbol{\theta}^* && (6.13)\\
&= \sum_{\mathbf{x}}\sum_{y}\int_{\Theta} L(y, \delta(\mathbf{x}))p(\mathbf{x}, y, \boldsymbol{\theta}^*)d\boldsymbol{\theta}^* && (6.14)\\
&= \sum_{\mathbf{x}}\left[\sum_{y} L(y, \delta(\mathbf{x}))p(y|\mathbf{x})dy\right] p(\mathbf{x}) && (6.15)\\
&= \sum_{\mathbf{x}} \rho(\delta(\mathbf{x})|\mathbf{x})\ p(\mathbf{x}) && (6.16)
\end{aligned}$$

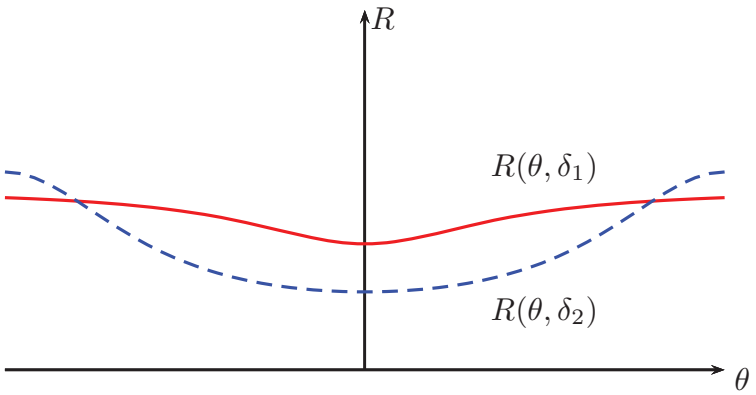



**Figure 6.2** Risk functions for two decision procedures, $\delta_1$ and $\delta_2$. Since $\delta_1$ has lower worst case risk, it is the minimax estimator, even though $\delta_2$ has lower risk for most values of $\theta$. Thus minimax estimators are overly conservative.

To minimize the overall expectation, we just minimize the term inside for each $\mathbf{x}$, so our decision rule is to pick

$$\delta_B(\mathbf{x}) = \operatorname*{argmin}_{a \in \mathcal{A}} \rho(a|\mathbf{x}) \tag{6.17}$$

□

Hence we see that the picking the optimal action on a case-by-case basis (as in the Bayesian approach) is optimal on average (as in the frequentist approach). In other words, the Bayesian approach provides a good way of achieving frequentist goals. In fact, one can go further and prove the following.

**Theorem 6.3.2** (Wald, 1950)**.** *Every admissable decision rule is a Bayes decision rule with respect to some, possibly improper, prior distribution.*

This theorem shows that *the best way to minimize frequentist risk is to be Bayesian!* See (Bernardo and Smith 1994, p448) for further discussion of this point.

### 6.3.2 Minimax risk

Obviously some frequentists dislike using Bayes risk since it requires the choice of a prior (although this is only in the evaluation of the estimator, not necessarily as part of its construction). An alternative approach is as follows. Define the **maximum risk** of an estimator as

$$R_{max}(\delta) \triangleq \max_{\boldsymbol{\theta}^*} R(\boldsymbol{\theta}^*, \delta) \tag{6.18}$$

A **minimax rule** is one which minimizes the maximum risk:

$$\delta_{MM} \triangleq \operatorname*{argmin}_{\delta} R_{max}(\delta) \tag{6.19}$$

For example, in Figure 6.2, we see that $\delta_1$ has lower worst-case risk than $\delta_2$, ranging over all possible values of $\theta^*$, so it is the minimax estimator (see Section 6.3.3.1 for an explanation of how to compute a risk function for an actual model).

Minimax estimators have a certain appeal. However, computing them can be hard. And furthermore, they are very pessimistic. In fact, one can show that all minimax estimators are equivalent to Bayes estimators under a **least favorable prior**. In most statistical situations (excluding game theoretic ones), assuming nature is an adversary is not a reasonable assumption.

### 6.3.3 Admissible estimators

The basic problem with frequentist decision theory is that it relies on knowing the true distribution $p(\cdot|\theta^*)$ in order to evaluate the risk. However, It might be the case that some estimators are worse than others regardless of the value of $\theta^*$. In particular, if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, then we say that $\delta_1$ **dominates** $\delta_2$. The domination is said to be strict if the inequality is strict for some $\theta$. An estimator is said to be **admissible** if it is not strictly dominated by any other estimator.

#### 6.3.3.1 Example

Let us give an example, based on (Bernardo and Smith 1994). Consider the problem of estimating the mean of a Gaussian. We assume the data is sampled from $x_i \sim \mathcal{N}(\theta^*, \sigma^2 = 1)$ and use quadratic loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. The corresponding risk function is the MSE. Some possible decision rules or estimators $\hat{\theta}(\mathbf{x}) = \delta(\mathbf{x})$ are as follows:

- $\delta_1(\mathbf{x}) = \overline{x}$, the sample mean
- $\delta_2(\mathbf{x}) = \tilde{\mathbf{x}}$, the sample median
- $\delta_3(\mathbf{x}) = \theta_0$, a fixed value
- $\delta_\kappa(\mathbf{x})$, the posterior mean under a $\mathcal{N}(\theta|\theta_0, \sigma^2/\kappa)$ prior:

$$\delta_\kappa(\mathbf{x}) = \frac{N}{N+\kappa}\overline{x} + \frac{\kappa}{N+\kappa}\theta_0 = w\overline{x} + (1-w)\theta_0 \tag{6.20}$$

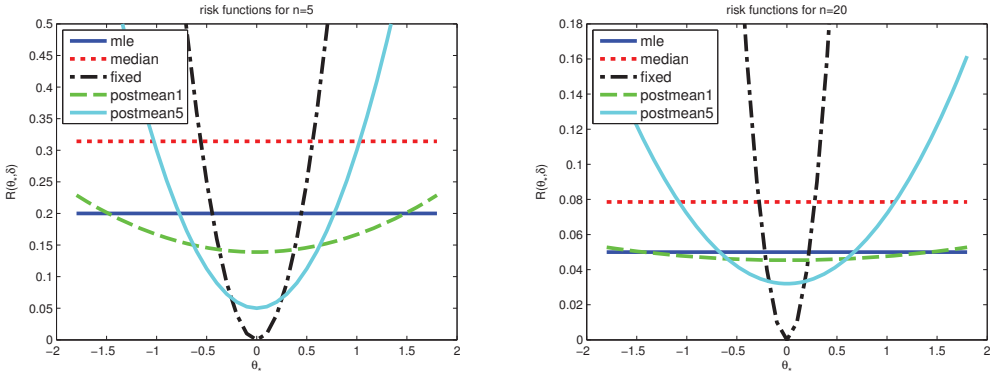
For $\delta_\kappa$, we consider a weak prior, $\kappa = 1$, and a stronger prior, $\kappa = 5$. The prior mean is $\theta_0$, some fixed value. We assume $\sigma^2$ is known. (Thus $\delta_3(\mathbf{x})$ is the same as $\delta_\kappa(\mathbf{x})$ with an infinitely strong prior, $\kappa = \infty$.)

Let us now derive the risk functions analytically. (We can do this since in this toy example, we know the true parameter $\theta^*$.) In Section 6.4.4, we show that the MSE can be decomposed into squared bias plus variance:

$$MSE(\hat{\theta}(\cdot)|\theta^*) \quad = \quad \text{var}\left[\hat{\theta}\right] + \text{bias}^2(\hat{\theta}) \tag{6.21}$$

The sample mean is unbiased, so its risk is

$$MSE(\delta_1|\theta^*) = \text{var}\left[\overline{x}\right] = \frac{\sigma^2}{N} \tag{6.22}$$




(a) (b)

**Figure 6.3** Risk functions for estimating the mean of a Gaussian using data sampled $\mathcal{N}(\theta^*, \sigma^2 = 1)$. The solid dark blue horizontal line is the MLE, the solid light blue curved line is the posterior mean when $\kappa = 5$. Left: $N = 5$ samples. Right: $N = 20$ samples. Based on Figure B.1 of (Bernardo and Smith 1994). Figure generated by `riskFnGauss`.

The sample median is also unbiased. One can show that the variance is approximately $\pi/(2N)$, so

$$MSE(\delta_2|\theta^*) = \frac{\pi}{2N} \tag{6.23}$$

For $\delta_3(\mathbf{x}) = \theta_0$, the variance is zero, so

$$MSE(\delta_3|\theta^*) = (\theta^* - \theta_0)^2 \tag{6.24}$$

Finally, for the posterior mean, we have

$$\begin{aligned} MSE(\delta_\kappa|\theta^*) &= \mathbb{E}\left[(w\overline{x} + (1-w)\theta_0 - \theta^*)^2\right] && (6.25)\\ &= \mathbb{E}\left[(w(\overline{x} - \theta^*) + (1-w)(\theta_0 - \theta^*))^2\right] && (6.26)\\ &= w^2\frac{\sigma^2}{N} + (1-w)^2(\theta_0 - \theta^*)^2 && (6.27)\\ &= \frac{1}{(N+\kappa)^2}\left(N\sigma^2 + \kappa^2(\theta_0 - \theta^*)^2\right) && (6.28) \end{aligned}$$

These functions are plotted in Figure 6.3 for $N \in \{5, 20\}$. We see that in general, the best estimator depends on the value of $\theta^*$, which is unknown. If $\theta^*$ is very close to $\theta_0$, then $\delta_3$ (which just predicts $\theta_0$) is best. If $\theta^*$ is within some reasonable range around $\theta_0$, then the posterior mean, which combines the prior guess of $\theta_0$ with the actual data, is best. If $\theta^*$ is far from $\theta_0$, the MLE is best. None of this should be suprising: a small amount of shrinkage (using the posterior mean with a weak prior) is usually desirable, assuming our prior mean is sensible.

What is more surprising is that the risk of decision rule $\delta_2$ (sample median) is always higher than that of $\delta_1$ (sample mean) for every value of $\theta^*$. Consequently the sample median is an

inadmissible estimator for this particular problem (where the data is assumed to come from a Gaussian).

In practice, the sample median is often better than the sample mean, because it is more robust to outliers. One can show (Minka 2000d) that the median is the Bayes estimator (under squared loss) if we assume the data comes from a Laplace distribution, which has heavier tails than a Gaussian. More generally, we can construct robust estimators by using flexible models of our data, such as mixture models or non-parametric density estimators (Section 14.7.2), and then computing the posterior mean or median.

#### 6.3.3.2 Stein's paradox *

Suppose we have $N$ iid random variables $X_i \sim \mathcal{N}(\theta_i, 1)$, and we want to estimate the $\theta_i$. The obvious estimator is the MLE, which in this case sets $\hat{\theta}_i = x_i$. It turns out that this is an inadmissible estimator under quadratic loss, when $N \geq 4$.

To show this, it suffices to construct an estimator that is better. The James-Stein estimator is one such estimator, and is defined as follows:

$$\hat{\theta}_i \quad = \quad \hat{B}\overline{x} + (1 - \hat{B})x_i = \overline{x} + (1 - \hat{B})(x_i - \overline{x}) \tag{6.29}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $0 < B < 1$ is some tuning constant. This estimate "shrinks" the $\theta_i$ towards the overall mean. (We derive this estimator using an empirical Bayes approach in Section 5.6.2.)

It can be shown that this shrinkage estimator has lower frequentist risk (MSE) than the MLE (sample mean) for $N \geq 4$. This is known as **Stein's paradox**. The reason it is called a paradox is illustrated by the following example. Suppose $\theta_i$ is the "true" IQ of student $i$ and $X_i$ is his test score. Why should my estimate of $\theta_i$ depend on the global mean $\overline{x}$, and hence on some other student's scores? One can create even more paradoxical examples by making the different dimensions be qualitatively different, e.g., $\theta_1$ is my IQ, $\theta_2$ is the average rainfall in Vancouver, etc.

The solution to the paradox is the following. If your goal is to estimate just $\theta_i$, you cannot do better than using $x_i$, but if the goal is to estimate the whole vector $\boldsymbol{\theta}$, and you use squared error as your loss function, then shrinkage helps. To see this, suppose we want to estimate $||\boldsymbol{\theta}||_2^2$ from a single sample $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$. A simple estimate is $||\mathbf{x}||_2^2$, but this will overestimate the result, since

$$\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] = \mathbb{E}\left[\sum_i x_i^2\right] = \sum_{i=1}^{N}\left(1 + \theta_i^2\right) = N + ||\boldsymbol{\theta}||_2^2 \tag{6.30}$$

Consequently we can reduce our risk by pooling information, even from unrelated sources, and shrinking towards the overall mean. In Section 5.6.2, we give a Bayesian explanation for this. See also (Efron and Morris 1975).

#### 6.3.3.3 Admissibility is not enough

It seems clear that we can restrict our search for good estimators to the class of admissible estimators. But in fact it is easy to construct admissible estimators, as we show in the following example.

**Theorem 6.3.3.** *Let $X \sim \mathcal{N}(\theta, 1)$, and consider estimating $\theta$ under squared loss. Let $\delta_1(x) = \theta_0$, a constant* independent of the data. *This is an admissible estimator.*

*Proof.* Suppose not. Then there is some other estimator $\delta_2$ with smaller risk, so $R(\theta^*, \delta_2) \leq R(\theta^*, \delta_1)$, where the inequality must be strict for some $\theta^*$. Suppose the true parameter is $\theta^* = \theta_0$. Then $R(\theta^*, \delta_1) = 0$, and

$$R(\theta^*, \delta_2) = \int (\delta_2(x) - \theta_0)^2 p(x|\theta_0) dx \tag{6.31}$$

Since $0 \leq R(\theta^*, \delta_2) \leq R(\theta^*, \delta_1)$ for all $\theta^*$, and $R(\theta_0, \delta_1) = 0$, we have $R(\theta_0, \delta_2) = 0$ and hence $\delta_2(x) = \theta_0 = \delta_1(x)$. Thus the only way $\delta_2$ can avoid having higher risk than $\delta_1$ at some specific point $\theta_0$ is by being equal to $\delta_1$. Hence there is no other estimator $\delta_2$ with strictly lower risk, so $\delta_2$ is admissible. □

## 6.4 Desirable properties of estimators

Since frequentist decision theory does not provide an automatic way to choose the best estimator, we need to come up with other heuristics for choosing amongst them. In this section, we discuss some properties we would like estimators to have. Unfortunately, we will see that we cannot achieve all of these properties at the same time.

### 6.4.1 Consistent estimators

An estimator is said to be **consistent** if it eventually recovers the true parameters that generated the data as the sample size goes to infinity, i.e., $\hat{\theta}(\mathcal{D}) \to \theta^*$ as $|\mathcal{D}| \to \infty$ (where the arrow denotes convergence in probability). Of course, this concept only makes sense if the data actually comes from the specified model with parameters $\theta^*$, which is not usually the case with real data. Nevertheless, it can be a useful theoretical property.

It can be shown that the MLE is a consistent estimator. The intuitive reason is that maximizing likelihood is equivalent to minimizing $\mathbb{KL}\left(p(\cdot|\boldsymbol{\theta}^*)||p(\cdot|\hat{\boldsymbol{\theta}})\right)$, where $p(\cdot|\boldsymbol{\theta}^*)$ is the true distribution and $p(\cdot|\hat{\boldsymbol{\theta}})$ is our estimate. We can achieve 0 KL divergence iff $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$.[4]

### 6.4.2 Unbiased estimators

The **bias** of an estimator is defined as

$$\text{bias}(\hat{\theta}(\cdot)) = \mathbb{E}_{p(\mathcal{D}|\theta_*)}\left[\hat{\theta}(\mathcal{D}) - \theta_*\right] \tag{6.32}$$

where $\theta_*$ is the true parameter value. If the bias is zero, the estimator is called **unbiased**. This means the sampling distribution is centered on the true parameter. For example, the MLE for a Gaussian mean is unbiased:

$$\text{bias}(\hat{\mu}) = \mathbb{E}\left[\overline{x}\right] - \mu = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] - \mu = \frac{N\mu}{N} - \mu = 0 \tag{6.33}$$

4. If the model is unidentifiable, the MLE may select a set of parameters that is different from the true parameters but for which the induced distribution, $p(\cdot|\hat{\boldsymbol{\theta}})$, is the same as the exact distribution. Such parameters are said to be likelihood equivalent.

However, the MLE for a Gaussian variance, $\hat{\sigma}^2$, is not an unbiased estimator of $\sigma^2$. In fact, one can show (Exercise 6.3) that

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{N-1}{N}\sigma^2 \tag{6.34}$$

However, the following estimator

$$\hat{\sigma}^2_{N-1} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{6.35}$$

*is* an unbiased estimator, which we can easily prove as follows:

$$\mathbb{E}\left[\hat{\sigma}^2_{N-1}\right] = \mathbb{E}\left[\frac{N}{N-1}\hat{\sigma}^2\right] = \frac{N}{N-1}\frac{N-1}{N}\sigma^2 = \sigma^2 \tag{6.36}$$

In Matlab, `var(X)` returns $\hat{\sigma}^2_{N-1}$, whereas `var(X,1)` returns $\hat{\sigma}^2$ (the MLE). For large enough $N$, the difference will be negligible.

Although the MLE may sometimes be a biased estimator, one can show that asymptotically, it is always unbiased. (This is necessary for the MLE to be a consistent estimator.)

Although being unbiased sounds like a desirable property, this is not always true. See Section 6.4.4 and (Lindley 1972) for discussion of this point.

### 6.4.3 Minimum variance estimators

It seems intuitively reasonable that we want our estimator to be unbiased (although we shall give some arguments against this claim below). However, being unbiased is not enough. For example, suppose we want to estimate the mean of a Gaussian from $\mathcal{D} = \{x_1, \ldots, x_N\}$. The estimator that just looks at the first data point, $\hat{\theta}(\mathcal{D}) = x_1$, is an unbiased estimator, but will generally be further from $\theta_*$ than the empirical mean $\overline{x}$ (which is also unbiased). So the variance of an estimator is also important.

A natural question is: how long can the variance go? A famous result, called the **Cramer-Rao lower bound**, provides a lower bound on the variance of any unbiased estimator. More precisely,

**Theorem 6.4.1** (Cramer-Rao inequality)**.** *Let $X_1, \ldots, X_n \sim p(X|\theta_0)$ and $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ be an unbiased estimator of $\theta_0$. Then, under various smoothness assumptions on $p(X|\theta_0)$, we have*

$$\text{var}\left[\hat{\theta}\right] \geq \frac{1}{nI(\theta_0)} \tag{6.37}$$

*where $I(\theta_0)$ is the Fisher information matrix (see Section 6.2.2).*

A proof can be found e.g., in (Rice 1995, p275).

It can be shown that the MLE achieves the Cramer Rao lower bound, and hence has the smallest asymptotic variance of any unbiased estimator. Thus MLE is said to be **asymptotically optimal**.

### 6.4.4 The bias-variance tradeoff

Although using an unbiased estimator seems like a good idea, this is not always the case. To see why, suppose we use quadratic loss. As we showed above, the corresponding risk is the MSE. We now derive a very useful decomposition of the MSE. (All expectations and variances are wrt the true distribution $p(\mathcal{D}|\theta^*)$, but we drop the explicit conditioning for notational brevity.) Let $\hat{\theta} = \hat{\theta}(\mathcal{D})$ denote the estimate, and $\overline{\theta} = \mathbb{E}\left[\hat{\theta}\right]$ denote the expected value of the estimate (as we vary $\mathcal{D}$). Then we have

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - \theta^*)^2\right] &= \mathbb{E}\left[\left[(\hat{\theta} - \overline{\theta}) + (\overline{\theta} - \theta^*)\right]^2\right] && (6.38)\\
&= \mathbb{E}\left[\left(\hat{\theta} - \overline{\theta}\right)^2\right] + 2(\overline{\theta} - \theta^*)\mathbb{E}\left[\hat{\theta} - \overline{\theta}\right] + (\overline{\theta} - \theta^*)^2 && (6.39)\\
&= \mathbb{E}\left[\left(\hat{\theta} - \overline{\theta}\right)^2\right] + (\overline{\theta} - \theta^*)^2 && (6.40)\\
&= \mathrm{var}\left[\hat{\theta}\right] + \mathrm{bias}^2(\hat{\theta}) && (6.41)
\end{aligned}$$

In words,

$$\boxed{\text{MSE = variance + bias}^2} \tag{6.42}$$

This is called the **bias-variance tradeoff** (see e.g., (Geman et al. 1992)). What it means is that it might be wise to use a biased estimator, so long as it reduces our variance, assuming our goal is to minimize squared error.

#### 6.4.4.1 Example: estimating a Gaussian mean

Let us give an example, based on (Hoff 2009, p79). Suppose we want to estimate the mean of a Gaussian from $\mathbf{x} = (x_1, \ldots, x_N)$. We assume the data is sampled from $x_i \sim \mathcal{N}(\theta^* = 1, \sigma^2)$. An obvious estimate is the MLE. This has a bias of 0 and a variance of

$$\mathrm{var}\left[\overline{x}|\theta^*\right] = \frac{\sigma^2}{N} \tag{6.43}$$

But we could also use a MAP estimate. In Section 4.6.1, we show that the MAP estimate under a Gaussian prior of the form $\mathcal{N}(\theta_0, \sigma^2/\kappa_0)$ is given by

$$\tilde{x} \triangleq \frac{N}{N+\kappa_0}\overline{x} + \frac{\kappa_0}{N+\kappa_0}\theta_0 = w\overline{x} + (1-w)\theta_0 \tag{6.44}$$

where $0 \leq w \leq 1$ controls how much we trust the MLE compared to our prior. (This is also the posterior mean, since the mean and mode of a Gaussian are the same.) The bias and variance are given by

$$\begin{aligned}
\mathbb{E}\left[\tilde{x}\right] - \theta^* &= w\theta_0 + (1-w)\theta_0 - \theta^* = (1-w)(\theta_0 - \theta^*) && (6.45)\\
\mathrm{var}\left[\tilde{x}\right] &= w^2\frac{\sigma^2}{N} && (6.46)
\end{aligned}$$

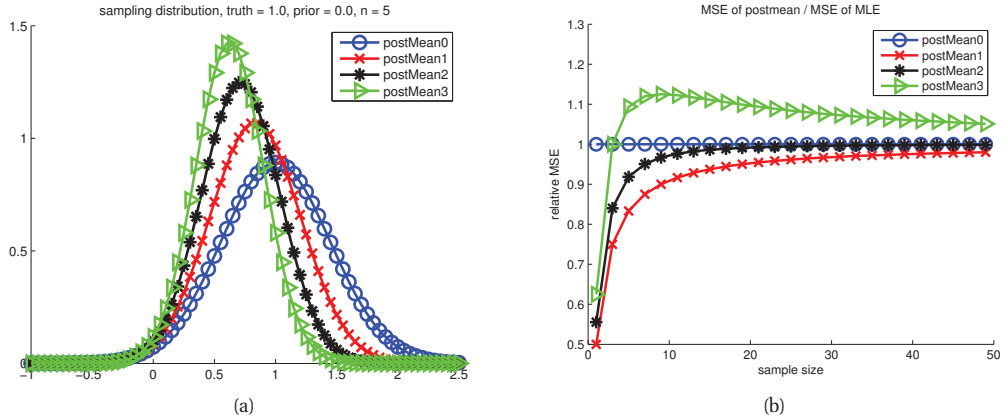



**Figure 6.4** Left: Sampling distribution of the MAP estimate with different prior strengths $\kappa_0$. (The MLE corresponds to $\kappa_0 = 0$.) Right: MSE relative to that of the MLE versus sample size. Based on Figure 5.6 of (Hoff 2009). Figure generated by `samplingDistGaussShrinkage`.

So although the MAP estimate is biased (assuming $w < 1$), it has lower variance.

Let us assume that our prior is slightly misspecified, so we use $\theta_0 = 0$, whereas the truth is $\theta^* = 1$. In Figure 6.4(a), we see that the sampling distribution of the MAP estimate for $\kappa_0 > 0$ is biased away from the truth, but has lower variance (is narrower) than that of the MLE.

In Figure 6.4(b), we plot $\text{mse}(\tilde{x})/\text{mse}(\overline{x})$ vs $N$. We see that the MAP estimate has lower MSE than the MLE, especially for small sample size, for $\kappa_0 \in \{1, 2\}$. The case $\kappa_0 = 0$ corresponds to the MLE, and the case $\kappa_0 = 3$ corresponds to a strong prior, which hurts performance because the prior mean is wrong. It is clearly important to "tune" the strength of the prior, a topic we discuss later.

#### 6.4.4.2 Example: ridge regression

Another important example of the bias variance tradeoff arises in ridge regression, which we discuss in Section 7.5. In brief, this corresponds to MAP estimation for linear regression under a Gaussian prior, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})$ The zero-mean prior encourages the weights to be small, which reduces overfitting; the precision term, $\lambda$, controls the strength of this prior. Setting $\lambda = 0$ results in the MLE; using $\lambda > 0$ results in a biased estimate. To illustrate the effect on the variance, consider a simple example. Figure 6.5 on the left plots each individual fitted curve, and on the right plots the average fitted curve. We see that as we increase the strength of the regularizer, the variance decreases, but the bias increases.

#### 6.4.4.3 Bias-variance tradeoff for classification

If we use 0-1 loss instead of squared error, the above analysis breaks down, since the frequentist risk is no longer expressible as squared bias plus variance. In fact, one can show (Exercise 7.2 of (Hastie et al. 2009)) that the bias and variance combine multiplicatively. If the estimate is on

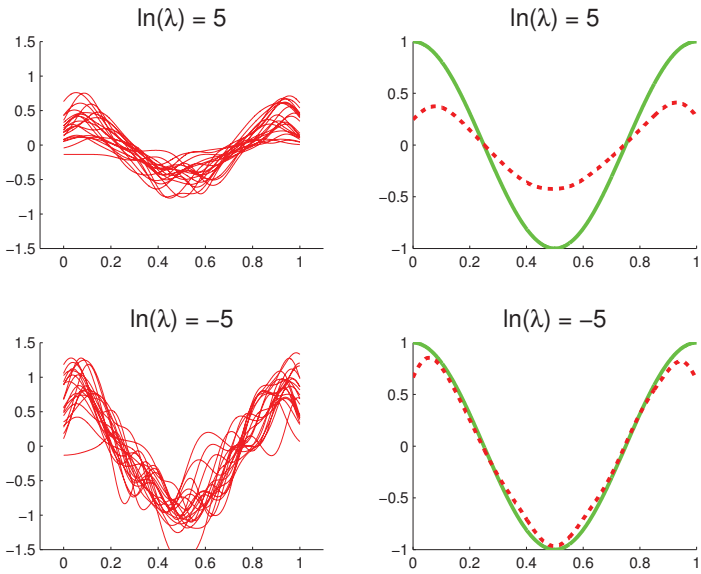



**Figure 6.5** Illustration of bias-variance tradeoff for ridge regression. We generate 100 data sets from the true function, shown in solid green. Left: we plot the regularized fit for 20 different data sets. We use linear regression with a Gaussian RBF expansion, with 25 centers evenly spread over the $[0, 1]$ interval. Right: we plot the average of the fits, averaged over all 100 datasets. Top row: strongly regularized: we see that the individual fits are similar to each other (low variance), but the average is far from the truth (high bias). Bottom row: lightly regularized: we see that the individual fits are quite different from each other (high variance), but the average is close to the truth (low bias). Based on (Bishop 2006a) Figure 3.5. Figure generated by `biasVarModelComplexity3`.

the correct side of the decision boundary, then the bias is negative, and decreasing the variance will decrease the misclassification rate. But if the estimate is on the wrong side of the decision boundary, then the bias is positive, so it pays to *increase* the variance (Friedman 1997a). This little known fact illustrates that the bias-variance tradeoff is not very useful for classification. It is better to focus on expected loss (see below), not directly on bias and variance. We can approximate the expected loss using cross validatinon, as we discuss in Section 6.5.3.

## 6.5 Empirical risk minimization

Frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. (By contrast, the Bayesian posterior expected loss can always be computed, since it conditions on the the data rather than conditioning on $\theta^*$.) However, there is one setting which avoids this problem, and that is where the task is to predict observable quantities, as opposed to estimating hidden variables or parameters. That is, instead of looking at loss functions of the form $L(\boldsymbol{\theta}, \delta(\mathcal{D}))$, where $\boldsymbol{\theta}$ is the true but unknown parameter, and $\delta(\mathcal{D})$ is our estimator, let us look at loss

functions of the form $L(y, \delta(\mathbf{x}))$, where $y$ is the true but unknown response, and $\delta(\mathbf{x})$ is our prediction given the input $\mathbf{x}$. In this case, the frequentist risk becomes

$$R(p_*, \delta) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim p_*}\left[L(y, \delta(\mathbf{x})\right] = \sum_{\mathbf{x}} \sum_{y} L(y, \delta(\mathbf{x}))p_*(\mathbf{x}, y) \tag{6.47}$$

where $p_*$ represents "nature's distribution". Of course, this distribution is unknown, but a simple approach is to use the empirical distribution, derived from some training data, to approximate $p_*$, i.e.,

$$p_*(\mathbf{x}, y) \approx p_{\text{emp}}(\mathbf{x}, y) \triangleq \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}_i}(\mathbf{x})\delta_{y_i}(y) \tag{6.48}$$

We then define the **empirical risk** as follows:

$$R_{emp}(\mathcal{D}, \mathcal{D}) \triangleq R(p_{\text{emp}}, \delta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \delta(\mathbf{x}_i)) \tag{6.49}$$

In the case of 0-1 loss, $L(y, \delta(\mathbf{x})) = \mathbb{I}(y \neq \delta(\mathbf{x}))$, this becomes the **misclassification rate**. In the case of squared error loss, $L(y, \delta(\mathbf{x})) = (y - \delta(\mathbf{x}))^2$, this becomes the **mean squared error**. We define the task of **empirical risk minimization** or **ERM** as finding a decision procedure (typically a classification rule) to minimize the empirical risk:

$$\delta_{ERM}(\mathcal{D}) = \underset{\delta}{\operatorname{argmin}}\, R_{emp}(\mathcal{D}, \delta) \tag{6.50}$$

In the unsupervised case, we eliminate all references to $y$, and replace $L(y, \delta(\mathbf{x}))$ with $L(\mathbf{x}, \delta(\mathbf{x}))$, where, for example, $L(\mathbf{x}, \delta(\mathbf{x})) = ||\mathbf{x} - \delta(\mathbf{x})||_2^2$, which measures the reconstruction error. We can define the decision rule using $\delta(\mathbf{x}) = \text{decode}(\text{encode}(\mathbf{x}))$, as in vector quantization (Section 11.4.2.6) or PCA (section 12.2). Finally, we define the empirical risk as

$$R_{emp}(\mathcal{D}, \delta) = \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x}_i, \delta(\mathbf{x}_i)) \tag{6.51}$$

Of course, we can always trivially minimize this risk by setting $\delta(\mathbf{x}) = \mathbf{x}$, so it is critical that the encoder-decoder go via some kind of bottleneck.

### 6.5.1 Regularized risk minimization

Note that the empirical risk is equal to the Bayes risk if our prior about "nature's distribution" is that it is exactly equal to the empirical distribution (Minka 2001b):

$$\mathbb{E}\left[R(p_*, \delta)|p_* = p_{\text{emp}}\right] = R_{emp}(\mathcal{D}, \delta) \tag{6.52}$$

Therefore minimizing the empirical risk will typically result in overfitting. It is therefore often necessary to add a complexity penalty to the objective function:

$$R'(\mathcal{D}, \delta) = R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta) \tag{6.53}$$

where $C(\delta)$ measures the complexity of the prediction function $\delta(\mathbf{x})$ and $\lambda$ controls the strength of the complexity penalty. This approach is known as **regularized risk minimization** (RRM). Note that if the loss function is negative log likelihood, and the regularizer is a negative log prior, this is equivalent to MAP estimation.

The two key issues in RRM are: how do we measure complexity, and how do we pick $\lambda$. For a linear model, we can define the complexity of in terms of its degrees of freedom, discussed in Section 7.5.3. For more general models, we can use the VC dimension, discussed in Section 6.5.4. To pick $\lambda$, we can use the methods discussed in Section 6.5.2.

### 6.5.2 Structural risk minimization

The regularized risk minimization principle says that we should fit the model, for a given complexity penalty, by using

$$\hat{\delta}_\lambda = \underset{\delta}{\operatorname{argmin}} \left[ R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta) \right] \tag{6.54}$$

But how should we pick $\lambda$? We cannot using the training set, since this will underestimate the true risk, a problem known as **optimism of the training error**. As an alternative, we can use the following rule, known as the **structural risk minimization** principle: (Vapnik 1998):

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \hat{R}(\hat{\delta}_\lambda) \tag{6.55}$$

where $\hat{R}(\delta)$ is an estimate of the risk. There are two widely used estimates: cross validation and theoretical upper bounds on the risk. We discuss both of these below.

### 6.5.3 Estimating the risk using cross validation

We can estimate the risk of some estimator using a validation set. If we don't have a separate validation set, we can use **cross validation** (CV), as we briefly discussed in Section 1.4.8. More precisely, CV is defined as follows. Let there be $N = |\mathcal{D}|$ data cases in the training set. Denote the data in the $k$'th test fold by $\mathcal{D}_k$ and all the other data by $\mathcal{D}_{-k}$. (In **stratified CV**, these folds are chosen so the class proportions (if discrete labels are present) are roughly equal in each fold.) Let $\mathcal{F}$ be a learning algorithm or fitting function that takes a dataset and a model index $m$ (this could a discrete index, such as the degree of a polynomial, or a continuous index, such as the strength of a regularizer) and returns a parameter vector:

$$\hat{\boldsymbol{\theta}}_m = \mathcal{F}(\mathcal{D}, m) \tag{6.56}$$

Finally, let $\mathcal{P}$ be a prediction function that takes an input and a parameter vector and returns a prediction:

$$\hat{y} = \mathcal{P}(\mathbf{x}, \hat{\boldsymbol{\theta}}) = f(\mathbf{x}, \hat{\boldsymbol{\theta}}) \tag{6.57}$$

Thus the combined **fit-predict cycle** is denoted as

$$f_m(\mathbf{x}, \mathcal{D}) = \mathcal{P}(\mathbf{x}, \mathcal{F}(\mathcal{D}, m)) \tag{6.58}$$

The $K$-fold CV estimate of the risk of $f_m$ is defined by

$$R(m, \mathcal{D}, K) \triangleq \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} L\left(y_i, \mathcal{P}(\mathbf{x}_i, \mathcal{F}(\mathcal{D}_{-k}, m))\right) \tag{6.59}$$

Note that we can call the fitting algorithm once per fold. Let $f_m^k(\mathbf{x}) = \mathcal{P}(\mathbf{x}, \mathcal{F}(\mathcal{D}_{-k}, m))$ be the function that was trained on all the data except for the test data in fold $k$. Then we can rewrite the CV estimate as

$$R(m, \mathcal{D}, K) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} L\left(y_i, f_m^k(\mathbf{x}_i)\right) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f_m^{k(i)}(\mathbf{x}_i)\right) \tag{6.60}$$

where $k(i)$ is the fold in which $i$ is used as test data. In other words, we predict $y_i$ using a model that was trained on data that does not contain $\mathbf{x}_i$.

Of $K = N$, the method is known as **leave one out cross validation** or LOOCV. In this case, n the estimated risk becomes

$$R(m, \mathcal{D}, N) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f_m^{-i}(\mathbf{x}_i)\right) \tag{6.61}$$

where $f_m^i(\mathbf{x}) = \mathcal{P}(\mathbf{x}, \mathcal{F}(\mathcal{D}_{-i}, m))$. This requires fitting the model $N$ times, where for $f_m^{-i}$ we omit the $i$'th training case. Fortunately, for some model classes and loss functions (namely linear models and quadratic loss), we can fit the model once, and analytically "remove" the effect of the $i$'th training case. This is known as **generalized cross validation** or GCV.

#### 6.5.3.1 Example: using CV to pick $\lambda$ for ridge regression

As a concrete example, consider picking the strength of the $\ell_2$ regularizer in penalized linear regression. We use the following rule:

$$\hat{\lambda} = \arg \min_{\lambda \in [\lambda_{min}, \lambda_{max}]} R(\lambda, \mathcal{D}_{\text{train}}, K) \tag{6.62}$$

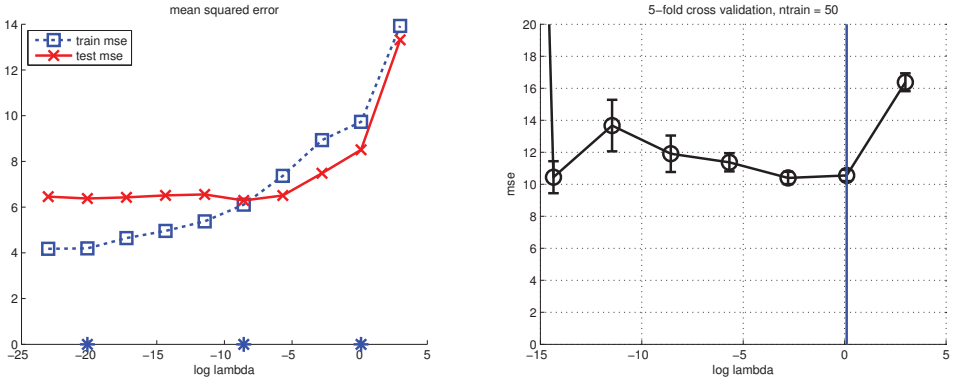
where $[\lambda_{min}, \lambda_{max}]$ is a finite range of $\lambda$ values that we search over, and $R(\lambda, \mathcal{D}_{\text{train}}, K)$ is the $K$-fold CV estimate of the risk of using $\lambda$, given by

$$R(\lambda, \mathcal{D}_{\text{train}}, K) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} L(y_i, f_\lambda^k(\mathbf{x}_i)) \tag{6.63}$$

where $f_\lambda^k(\mathbf{x}) = \mathbf{x}^T \hat{\mathbf{w}}_\lambda(\mathcal{D}_{-k})$ is the prediction function trained on data excluding fold $k$, and $\hat{\mathbf{w}}_\lambda(\mathcal{D}) = \arg\min_{\mathbf{w}} NLL(\mathbf{w}, \mathcal{D}) + \lambda ||\mathbf{w}||_2^2$ is the MAP estimate. Figure 6.6(b) gives an example of a CV estimate of the risk vs $\log(\lambda)$, where the loss function is squared error.

When performing classification, we usually use 0-1 loss. In this case, we optimize a convex upper bound on the empirical risk to estimate $\mathbf{w}_\lambda$m but we optimize (the CV estimate of) the risk itself to estimate $\lambda$. We can handle the non-smooth 0-1 loss function when estimating $\lambda$ because we are using brute-force search over the entire (one-dimensional) space.

When we have more than one or two tuning parameters, this approach becomes infeasible. In such cases, one can use empirical Bayes, which allows one to optimize large numbers of hyper-parameters using gradient-based optimizers instead of brute-force search. See Section 5.6 for details.




(a) (b)

**Figure 6.6** (a) Mean squared error for $\ell_2$ penalized degree 14 polynomial regression vs log regularizer. Same as in Figures 7.8, except now we have $N = 50$ training points instead of 21. The stars correspond to the values used to plot the functions in Figure 7.7. (b) CV estimate. The vertical scale is truncated for clarity. The blue line corresponds to the value chosen by the one standard error rule. Figure generated by `linregPolyVsRegDemo`.

#### 6.5.3.2 The one standard error rule

The above procedure estimates the risk, but does not give any measure of uncertainty. A standard frequentist measure of uncertainty of an estimate is the standard error of the mean, defined by

$$se = \frac{\hat{\sigma}}{\sqrt{N}} = \sqrt{\frac{\hat{\sigma}^2}{N}} \tag{6.64}$$

where $\hat{\sigma}^2$ is an estimate of the variance of the loss:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(L_i - \overline{L})^2, \quad L_i = L(y_i, f_m^{k(i)}(\mathbf{x}_i)) \quad \overline{L} = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{6.65}$$

Note that $\sigma$ measures the intrinsic variability of $L_i$ across samples, whereas $se$ measures our uncertainty about the mean $\overline{L}$.

Suppose we apply CV to a set of models and compute the mean and se of their estimated risks. A common heuristic for picking a model from these noisy estimates is to pick the value which corresponds to the simplest model whose risk is no more than one standard error above the risk of the best model; this is called the **one-standard error rule** (Hastie et al. 2001, p216). For example, in Figure 6.6, we see that this heuristic does not choose the lowest point on the curve, but one that is slightly to its right, since that corresponds to a more heavily regularized model with essentially the same empirical performance.

#### 6.5.3.3 CV for model selection in non-probabilistic unsupervised learning

If we are performing unsupervised learning, we must use a loss function such as $L(\mathbf{x}, \delta(\mathbf{x})) = ||\mathbf{x} - \delta(\mathbf{x})||_2$, which measures reconstruction error. Here $\delta(\mathbf{x})$ is some encode-decode scheme. However, as we discussed in Section 11.5.2, we cannot use CV to determine the complexity of $\delta$, since we will always get lower loss with a more complex model, even if evaluated on the test set. This is because more complex models will compress the data less, and induce less distortion. Consequently, we must either use probabilistic models, or invent other heuristics.

### 6.5.4 Upper bounding the risk using statistical learning theory *

The principle problem with cross validation is that it is slow, since we have to fit the model multiple times. This motivates the desire to compute analytic approximations or bounds to the generalization error. This is the studied in the field of **statistical learning theory** (SLT). More precisely, SLT tries to bound the risk $R(p_*, h)$ for any data distribution $p_*$ and hypothesis $h \in \mathcal{H}$ in terms of the empirical risk $R_{emp}(\mathcal{D}, h)$, the sample size $N = |\mathcal{D}|$, and the size of the hypothesis space $\mathcal{H}$.

Let us initially consider the case where the hypothesis space is finite, with size $\dim(\mathcal{H}) = |\mathcal{H}|$. In other words, we are selecting a model/ hypothesis from a finite list, rather than optimizing real-valued parameters, Then we can prove the following.

**Theorem 6.5.1.** *For any data distribution $p_*$, and any dataset $\mathcal{D}$ of size $N$ drawn from $p_*$, the probability that our estimate of the error rate will be more than $\epsilon$ wrong, in the worst case, is upper bounded as follows:*

$$P\left(\max_{h \in \mathcal{H}} |R_{emp}(\mathcal{D}, h) - R(p_*, h)| > \epsilon\right) \leq 2\dim(\mathcal{H}) e^{-2N\epsilon^2} \tag{6.66}$$

*Proof.* To prove this, we need two useful results. First, **Hoeffding's inequality**, which states that if $X_1, \ldots, X_N \sim \text{Ber}(\theta)$, then, for any $\epsilon > 0$,

$$P(|\overline{x} - \theta| > \epsilon) \leq 2e^{-2N\epsilon^2} \tag{6.67}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Second, the **union bound**, which says that if $A_1, \ldots, A_d$ are a set of events, then $P(\cup_{i=1}^{d} A_i) \leq \sum_{i=1}^{d} P(A_i)$.

Finally, for notational brevity, let $R(h) = R(h, p_*)$ be the true risk, and $\hat{R}_N(h) = R_{emp}(\mathcal{D}, h)$ be the empirical risk.

Using these results we have

$$\begin{aligned} P\left(\max_{h \in \mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon\right) &= P\left(\bigcup_{h \in \mathcal{H}} |\hat{R}_N(h) - R(h)| > \epsilon\right) && (6.68) \\ &\leq \sum_{h \in \mathcal{H}} P\left(|\hat{R}_N(h) - R(h)| > \epsilon\right) && (6.69) \\ &\leq \sum_{h \in \mathcal{H}} 2e^{-2N\epsilon^2} = 2\dim(\mathcal{H}) e^{-2N\epsilon^2} && (6.70) \end{aligned}$$

□

Ths bound tells us that the optimism of the training error increases with $\dim(\mathcal{H})$ but decreases with $N = |\mathcal{D}|$, as is to be expected.

If the hypothesis space $\mathcal{H}$ is infinite (e.g., we have real-valued parameters), we cannot use $\dim(\mathcal{H}) = |\mathcal{H}|$. Instead, we can use a quantity called the **Vapnik-Chervonenkis** or **VC** dimension of the hypothesis class. See (Vapnik 1998) for details.

Stepping back from all the theory, the key intuition behind statistical learning theory is quite simple. Suppose we find a model with low empirical risk. If the hypothesis space $\mathcal{H}$ is very big, relative to the data size, then it is quite likely that we just got "lucky" and were given a data set that is well-modeled by our chosen function by chance. However, this does not mean that such a function will have low generalization error. But if the hypothesis class is sufficiently constrained in size, and/or the training set is sufficiently large, then we are unlikely to get lucky in this way, so a low empirical risk is evidence of a low true risk.

Note that optimism of the training error does not necessarily increase with model complexity, but it does increase with the number of different models that are being searched over.

The advantage of statistical learning theory compared to CV is that the bounds on the risk are quicker to compute than using CV. The disadvantage is that it is hard to compute the VC dimension for many interesting models, and the upper bounds are usually very loose (although see (Kaariainen and Langford 2005)).

One can extend statistical learning theory by taking computational complexity of the learner into account. This field is called **computational learning theory** or **COLT**. Most of this work focuses on the case where $h$ is a binary classifier, and the loss function is 0-1 loss. If we observe a low empirical risk, and the hypothesis space is suitably "small", then we can say that our estimated function is **probably approximately correct** or **PAC**. A hypothesis space is said to be **efficiently PAC-learnable** if there is a polynomial time algorithm that can identify a function that is PAC. See (Kearns and Vazirani 1994) for details.

### 6.5.5 Surrogate loss functions

Minimizing the loss in the ERM/ RRM framework is not always easy. For example, we might want to optimize the AUC or F1 scores. Or more simply, we might just want to minimize the 0-1 loss, as is common in classification. Unfortunately, the 0-1 risk is a very non-smooth objective and hence is hard to optimize. One alternative is to use maximum likelihood estimation instead, since log-likelihood is a smooth convex upper bound on the 0-1 risk, as we show below.

To see this, consider binary logistic regression, and let $y_i \in \{-1, +1\}$. Suppose our decision function computes the log-odds ratio,

$$f(\mathbf{x}_i) = \log \frac{p(y=1|\mathbf{x}_i, \mathbf{w})}{p(y=-1|\mathbf{x}_i, \mathbf{w})} = \mathbf{w}^T \mathbf{x}_i = \eta_i \tag{6.71}$$

Then the corresponding probability distribution on the output label is

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \text{sigm}(y_i \eta_i) \tag{6.72}$$

Let us define the **log-loss** as as

$$L_{\text{nll}}(y, \eta) = -\log p(y|\mathbf{x}, \mathbf{w}) = \log(1 + e^{-y\eta}) \tag{6.73}$$

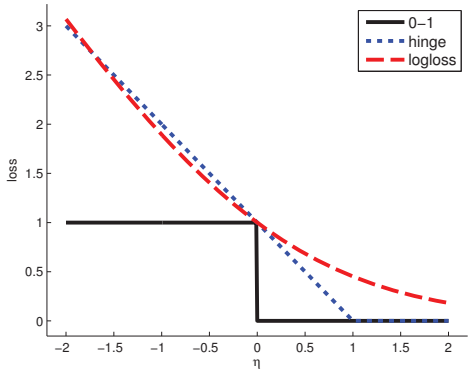



**Figure 6.7** Illustration of various loss functions for binary classification. The horizontal axis is the margin $y\eta$, the vertical axis is the loss. The log loss uses log base 2. Figure generated by `hingeLossPlot`.

It is clear that minimizing the average log-loss is equivalent to maximizing the likelihood.

Now consider computing the most probable label, which is equivalent to using $\hat{y} = -1$ if $\eta_i < 0$ and $\hat{y} = +1$ if $\eta_i \geq 0$. The 0-1 loss of our function becomes

$$L_{01}(y, \eta) = \mathbb{I}(y \neq \hat{y}) = \mathbb{I}(y\eta < 0) \tag{6.74}$$

Figure 6.7 plots these two loss functions. We see that the NLL is indeed an upper bound on the 0-1 loss.

Log-loss is an example of a **surrogate loss function**. Another example is the **hinge loss**:

$$L_{\text{hinge}}(y, \eta) = \max(0, 1 - y\eta) \tag{6.75}$$

See Figure 6.7 for a plot. We see that the function looks like a door hinge, hence its name. This loss function forms the basis of a popular classification method known as support vector machines (SVM), which we will discuss in Section 14.5.

The surrogate is usually chosen to be a convex upper bound, since convex functions are easy to minimize. See e.g., (Bartlett et al. 2006) for more information.

## 6.6 Pathologies of frequentist statistics *

> I believe that it would be very difficult to persuade an intelligent person that current [frequentist] statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem. — George Box, 1962.

Frequentist statistics exhibits various forms of weird and undesirable behaviors, known as **pathologies**. We give a few examples below, in order to caution the reader; these and other examples are explained in more detail in (Lindley 1972; Lindley and Phillips 1976; Lindley 1982; Berger 1985; Jaynes 2003; Minka 1999).

### 6.6.1 Counter-intuitive behavior of confidence intervals

A **confidence interval** is an interval derived from the sampling distribution of an estimator (whereas a Bayesian credible interval is derived from the posterior of a parameter, as we discussed in Section 5.2.2). More precisely, a frequentist confidence interval for some parameter $\theta$ is defined by the following (rather un-natural) expression:

$$C'_\alpha(\theta) = (\ell, u) : P(\ell(\tilde{\mathcal{D}}) \leq \theta \leq u(\tilde{\mathcal{D}})|\tilde{\mathcal{D}} \sim \theta) = 1 - \alpha \tag{6.76}$$

That is, if we sample hypothetical future data $\tilde{\mathcal{D}}$ from $\theta$, then $(\ell(\tilde{\mathcal{D}}), u(\tilde{\mathcal{D}}))$, is a confidence interval if the parameter $\theta$ lies inside this interval $1 - \alpha$ percent of the time.

Let us step back for a moment and think about what is going on. In Bayesian statistics, we condition on what is known — namely the observed data, $\mathcal{D}$ — and average over what is not known, namely the parameter $\theta$. In frequentist statistics, we do exactly the opposite: we condition on what is unknown — namely the true parameter value $\theta$ — and average over hypothetical future data sets $\tilde{\mathcal{D}}$.

This counter-intuitive definition of confidence intervals can lead to bizarre results. Consider the following example from (Berger 1985, p11). Suppose we draw two integers $\mathcal{D} = (x_1, x_2)$ from

$$p(x|\theta) = \begin{cases} 0.5 & \text{if } x = \theta \\ 0.5 & \text{if } x = \theta + 1 \\ 0 & \text{otherwise} \end{cases} \tag{6.77}$$

If $\theta = 39$, we would expect the following outcomes each with probability 0.25:

$$(39, 39), (39, 40), (40, 39), (40, 40) \tag{6.78}$$

Let $m = \min(x_1, x_2)$ and define the following confidence interval:

$$[\ell(\mathcal{D}), u(\mathcal{D})] = [m, m] \tag{6.79}$$

For the above samples this yields

$$[39, 39], \quad [39, 39], \quad [39, 39], \quad [40, 40] \tag{6.80}$$

Hence Equation 6.79 is clearly a 75% CI, since 39 is contained in 3/4 of these intervals. However, if $\mathcal{D} = (39, 40)$ then $p(\theta = 39|\mathcal{D}) = 1.0$, so we know that $\theta$ must be 39, yet we only have 75% "confidence" in this fact.

Another, less contrived example, is as follows. Suppose we want to estimate the parameter $\theta$ of a Bernoulli distribution. Let $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the sample mean. The MLE is $\hat{\theta} = \overline{x}$. An approximate 95% confidence interval for a Bernoulli parameter is $\overline{x} \pm 1.96\sqrt{\overline{x}(1-\overline{x})/N}$ (this is called a **Wald interval** and is based on a Gaussian approximation to the Binomial distribution; compare to Equation 3.27). Now consider a single trial, where $N = 1$ and $x_1 = 0$. The MLE is 0, which overfits, as we saw in Section 3.3.4.1. But our 95% confidence interval is also $(0, 0)$, which seems even worse. It can be argued that the above flaw is because we approximated the true sampling distribution with a Gaussian, or because the sample size was to small, or the parameter "too extreme". However, the Wald interval can behave badly even for large $N$, and non-extreme parameters (Brown et al. 2001).

### 6.6.2 p-values considered harmful

Suppose we want to decide whether to accept or reject some baseline model, which we will call the **null hypothesis**. We need to define some decision rule. In frequentist statistics, it is standard to first compute a quantity called the **p-value**, which is defined as the probability (under the null) of observing some **test statistic** $f(\mathcal{D})$ (such as the chi-squared statistic) that is as large *or larger* than that actually observed:[5]

$$\text{pvalue}(\mathcal{D}) \triangleq P(f(\tilde{\mathcal{D}}) \geq f(\mathcal{D})|\tilde{\mathcal{D}} \sim H_0) \tag{6.81}$$

This quantity relies on computing a **tail area probability** of the sampling distribution; we give an example of how to do this below.

Given the p-value, we define our decision rule as follows: we reject the null hypothesis iff the p-value is less than some threshold, such as $\alpha = 0.05$. If we do reject it, we say the difference between the observed test statistic and the expected test statistic is **statistically significant** at level $\alpha$. This approach is known as **null hypothesis significance testing**, or **NHST**.

This procedure guarantees that our expected type I (false positive) error rate is at most $\alpha$. This is sometimes interpreted as saying that frequentist hypothesis testing is very conservative, since it is unlikely to accidently reject the null hypothesis. But in fact the opposite is the case: because this method only worries about trying to reject the null, it can never gather evidence in favor of the null, no matter how large the sample size. Because of this, p-values tend to overstate the evidence against the null, and are thus very "trigger happy".

In general there can be huge differences between p-values and the quantity that we really care about, which is the posterior probability of the null hypothesis given the data, $p(H_0|\mathcal{D})$. In particular, Sellke et al. (2001) show that even if the p-value is as slow as 0.05, the posterior probability of $H_0$ is at least 30%, and often much higher. So frequentists often claim to have "significant" evidence of an effect that cannot be explained by the null hypothesis, whereas Bayesians are usually more conservative in their claims. For example, p-values have been used to "prove" that ESP (extra-sensory perception) is real (Wagenmakers et al. 2011), even though ESP is clearly very improbable. For this reason, p-values have been banned from certain medical journals (Matthews 1998).

Another problem with p-values is that their computation depends on decisions you make about when to stop collecting data, even if these decisions don't change the data you actually observed. For example, suppose I toss a coin $n = 12$ times and observe $s = 9$ successes (heads) and $f = 3$ failures (tails), so $n = s + f$. In this case, $n$ is fixed and $s$ (and hence $f$) is random. The relevant sampling model is the binomial

$$\text{Bin}(s|n, \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s} \tag{6.82}$$

Let the null hypothesis be that the coin is fair, $\theta = 0.5$, where $\theta$ is the probability of success (heads). The one-sided p-value, using test statistic $t(s) = s$, is

$$p_1 = P(S \geq 9|H_0) = \sum_{s=9}^{12} \text{Bin}(s|12, 0.5) = \sum_{s=9}^{12} \binom{12}{s} 0.5^{12} = 0.073 \tag{6.83}$$

5. The reason we cannot just compute the probability of the observed value of the test statistic is that this will have probability zero under a pdf. The p-value is defined in terms of the cdf, so is always a number between 0 and 1.

The two-sided p-value is

$$p_2 = \sum_{s=9}^{12} \text{Bin}(s|12, 0.5) + \sum_{s=0}^{3} \text{Bin}(s|12, 0.5) = 0.073 + 0.073 = 0.146 \tag{6.84}$$

In either case, the p-value is larger than the magical 5% threshold, so a frequentist would not reject the null hypothesis.

Now suppose I told you that I actually kept tossing the coin until I observed $f = 3$ tails. In this case, $f$ is fixed and $n$ (and hence $s = n - f$) is random. The probability model becomes the **negative binomial distribution**, given by

$$\text{NegBinom}(s|f, \theta) = \binom{s+f-1}{f-1} \theta^s (1-\theta)^f \tag{6.85}$$

where $f = n - s$.

Note that the term which depends on $\theta$ is the same in Equations 6.82 and 6.85, so the posterior over $\theta$ would be the same in both cases. However, these two interpretations of the same data give different p-values. In particular, under the negative binomial model we get

$$p_3 = P(S \geq 9|H_0) = \sum_{s=9}^{\infty} \binom{3+s-1}{2} (1/2)^s (1/2)^3 = 0.0327 \tag{6.86}$$

So the p-value is 3%, and suddenly there seems to be significant evidence of bias in the coin! Obviously this is ridiculous: the data is the same, so our inferences about the coin should be the same. After all, I could have chosen the experimental protocol at random. It is the outcome of the experiment that matters, not the details of how I decided which one to run.

Although this might seem like just a mathematical curiosity, this also has significant practical implications. In particular, the fact that the **stopping rule** affects the computation of the p-value means that frequentists often do not terminate experiments early, even when it is obvious what the conclusions are, lest it adversely affect their statistical analysis. If the experiments are costly or harmful to people, this is obviously a bad idea. Perhaps it is not surprising, then, that the US Food and Drug Administration (FDA), which regulates clinical trials of new drugs, has recently become supportive of Bayesian methods[6], since Bayesian methods are not affected by the stopping rule.

### 6.6.3 The likelihood principle

The fundamental reason for many of these pathologies is that frequentist inference violates the **likelihood principle**, which says that inference should be based on the likelihood of the observed data, not based on hypothetical future data that you have not observed. Bayes obviously satisfies the likelihood principle, and consequently does not suffer from these pathologies.

A compelling argument in favor of the likelihood principle was presented in (Birnbaum 1962), who showed that it followed automatically from two simpler principles. The first of these is the **sufficiency principle**, which says that a sufficient statistic contains all the relevant information

6. See http://yamlb.wordpress.com/2006/06/19/the-us-fda-is-becoming-progressively-more-bayesian/.

about an unknown parameter (arguably this is true by definition). The second principle is known as **weak conditionality**, which says that inferences should be based on the events that happened, not which might have happened. To motivate this, consider an example from (Berger 1985). Suppose we need to analyse a substance, and can send it either to a laboratory in New York or in California. The two labs seem equally good, so a fair coin is used to decide between them. The coin comes up heads, so the California lab is chosen. When the results come back, should it be taken into account that the coin could have come up tails and thus the New York lab could have been used? Most people would argue that the New York lab is irrelevant, since the tails event didn't happen. This is an example of weak conditionality. Given this principle, one can show that all inferences should only be based on what was observed, which is in contrast to standard frequentist procedures. See (Berger and Wolpert 1988) for further details on the likelihood principle.

### 6.6.4 Why isn't everyone a Bayesian?

Given these fundamental flaws of frequentist statistics, and the fact that Bayesian methods do not have such flaws, an obvious question to ask is: "Why isn't everyone a Bayesian?" The (frequentist) statistician Bradley Efron wrote a paper with exactly this title (Efron 1986). His short paper is well worth reading for anyone interested in this topic. Below we quote his opening section:

> The title is a reasonable question to ask on at least two counts. First of all, everone used to be a Bayesian. Laplace wholeheatedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit. This included Gauss, whose statistical work is usually presented in frequentist terms.
>
> A second and more important point is the cogency of the Bayesian argument. Modern statisticians, following the lead of Savage and de Finetti, have advanced powerful theoretical arguments for preferring Bayesian inference. A byproduct of this work is a disturbing catalogue of inconsistencies in the frequentist point of view.
>
> Nevertheless, everyone is not a Bayesian. The current era (1986) is the first century in which statistics has been widely used for scientific reporting, and in fact, 20th-century statistics is mainly non-Bayesian. However, Lindley (1975) predicts a change for the 21st century.

Time will tell whether Lindley was right....

## Exercises

**Exercise 6.1** Pessimism of LOOCV

(Source: Witten05, p152.). Suppose we have a completely random labeled dataset (i.e., the features $\mathbf{x}$ tell us nothing about the class labels $y$) with $N_1$ examples of class 1, and $N_2$ examples of class 2, where $N_1 = N_2$. What is the best misclassification rate any method can achieve? What is the estimated misclassification rate of the same method using LOOCV?

**Exercise 6.2** James Stein estimator for Gaussian means

Consider the 2 stage model $Y_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$ and $\theta_i|\mu \sim \mathcal{N}(m_0, \tau_0^2)$. Suppose $\sigma^2 = 500$ is known and we observe the following 6 data points, $i = 1 : 6$:

```
1505, 1528, 1564, 1498, 1600, 1470
```

a. Find the ML-II estimates of $m_0$ and $\tau_0^2$.

b. Find the posterior estimates $\mathbb{E}[\theta_i|y_i, m_0, \tau_0]$ and $\text{var}[\theta_i|y_i, m_0, \tau_0]$ for $i = 1$. (The other terms, $i = 2:6$, are computed similarly.)

c. Give a 95% credible interval for $p(\theta_i|y_i, m_0, \tau_0)$ for $i = 1$. Do you trust this interval (assuming the Gaussian assumption is reasonable)? i.e. is it likely to be too large or too small, or just right?

d. What do you expect would happen to your estimates if $\sigma^2$ were much smaller (say $\sigma^2 = 1$)? You do not need to compute the numerical answer; just briefly explain what would happen qualitatively, and why.

**Exercise 6.3** $\hat{\sigma}^2_{MLE}$ is biased

Show that $\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$ is a biased estimator of $\sigma^2$, i.e., show

$$\mathbf{E}_{\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n}\sim\mathcal{N}(\mu,\sigma)}[\hat{\sigma}^2(\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n}) \neq \sigma^2$$

Hint: note that $X_1, \ldots, X_N$ are independent, and use the fact that the expectation of a product of independent random variables is the product of the expectations.

**Exercise 6.4** Estimation of $\sigma^2$ when $\mu$ is known

Suppose we sample $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu$ is a *known* constant. Derive an expression for the MLE for $\sigma^2$ in this case. Is it unbiased?