Frequentist statistics: a concise introduction

Kevin P. Murphy

Last updated September 5, 2007

1 Introduction

Whereas probability is concerned with describing the relative likelihoods of generating various kinds of data, statistics is concerned with the opposite problem: inferring the causes that generated the observed data. Indeed, statistics used to be known as **inverse probability**.

As mentioned earlier, there are two interpretations of probability: the frequentist interpretation in terms of long term frequencies of future events, and the Bayesian interpretation, in terms of modeling (subjective) uncertainty given the current data. This in turn gives rise to two approaches to statistics. The **frequentist approach** to statistics is the most widely used, and hence is sometimes called the **orthodox approach** or **classical approach**. However, the Bayesian approach is becoming increasingly popular. We will study the Bayesian approach later.

2 Point estimation

Point estimation refers to computing a single "best guess" of some quantity of interest from data. The quantity could be a parameter in a parametric model (such as the mean of a Gaussian), or a regression function, or a prediction of a future value of some random variable. We assume there is some "true" value for this quantity, which is fixed but unknown, call it θ. Our goal is to construct an **estimator**, which is some function g that takes sample data, $\mathcal{D} = (X_1, \ldots, X_N)$, and returns a point estimate $\hat{\theta}_N$:

$$\hat{\theta}_N = g(X_1, \ldots, X_N) \tag{1}$$

Since $\hat{\theta}_N$ depends on the particular observed data, $\hat{\theta} = \hat{\theta}(\mathcal{D})$, it is a random variable. (Often we omit the subscript N and just write $\hat{\theta}$.)

An example of an estimator is the **sample mean** of the data:

$$\hat{\mu} = \overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2}$$

Another is the **sample variance**:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x}) \tag{3}$$

A third example is the empirical fraction of heads in a sequence of heads (1s) and tails (0s):

$$\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{4}$$

where $X_n \sim Be(\theta)$.

2.1 Desirable properties of estimators

There are many possible estimators, so how do we know which to use? Below we describe some of the most desirable properties of an estimator.

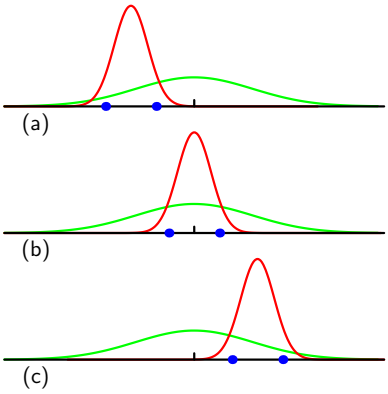

Figure 1: Graphical illustration of why $\hat{\sigma}^2_{ML}$ is biased: it underestimates the true variance because it measures spread around the empirical mean $\hat{\mu}_{ML}$ instead of around the true mean. Source: [Bis06] Figure 1.15.

2.1.1 Unbiased estimators

We define the **bias** of an estimator as

$$\text{bias}(\hat{\theta}_N) = E_{\mathcal{D}\sim\theta}(\hat{\theta}_N - \theta) \tag{5}$$

where the expectation is over data sets $\mathcal{D}$ drawn from a distribution with parameter θ. We say that $\hat{\theta}_N$ is **unbiased** if $E_\theta(\hat{\theta}_N) = \theta$. It is easy to see that $\hat{\mu}$ is an unbiased estimator:

$$E\hat{\mu} = E\frac{1}{N}\sum_{n=1}^{N} X_n = \frac{1}{N}\sum_n E[X_n] = \frac{1}{N}N\mu \tag{6}$$

However, one can show (exercise) that

$$E\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2 \tag{7}$$

Therefore it is common to use the following unbiased estimator of the variance:

$$\hat{\sigma}^2_{N-1} = \frac{N}{N-1}\hat{\sigma}^2 \tag{8}$$

In Matlab, `var(X)` returns $\hat{\sigma}^2_{N-1}$ whereas `var(X,1)` returns $\hat{\sigma}^2$.

One might ask: why is σ^2_{ML} biased? Intuitively, we "used up" one "degree of freedom" in estimating μ_{ML}, so we underestimate σ. (If we used μ instead of μ_{ML} when computing σ^2_{ML}, the result would be unbiased.) See Figure 1.

2.1.2 Bias-variance tradeoff

Being unbiased seems like a good thing, but it turns out that a little bit of bias can be useful so long as it reduces the variance of the estimator. In particular, suppose our goal is to minimize the **mean squared error** (MSE)

$$MSE = E_\theta(\hat{\theta}_N - \theta)^2 \tag{9}$$

It turns out that when minimizing MSE, there is a **bias-variance tradeoff**.

Theorem 2.1. *The MSE can be written as*

$$MSE = bias^2(\hat{\theta}) + Var_\theta(\hat{\theta}) \tag{10}$$

Proof. Let $\overline{\theta} = E_{\mathcal{D}\sim\theta}(\hat{\theta}(\mathcal{D}))$. Then

$$\begin{aligned} E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \theta)^2 &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta} + \overline{\theta} - \theta)^2 && (11)\\ &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + 2(\overline{\theta} - \theta)E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta}) + (\overline{\theta} - \theta)^2 && (12)\\ &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + (\overline{\theta} - \theta)^2 && (13)\\ &= V(\hat{\theta}) + \text{bias}^2(\hat{\theta}) && (14) \end{aligned}$$

where we have used the fact that $E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta}) = \overline{\theta} - \overline{\theta} = 0$. □

Later we will see that simple models are often biased (because they cannot represent the truth) but have low variance, whereas more complex models have lower bias but higher variance. **Bagging** is a simple way to reduce the variance of an estimator without increasing the bias: simply take weighted combinations of estimatros fit on different subsets of the data, chosen randomly with replacement.

2.1.3 Consistent estimators

Having low MSE is good, but is not enough. We would also like our estimator to converge to the true value as we collect more data. We call such an estimator **consistent**. Formally, $\hat{\theta}$ is consistent if $\hat{\theta}_N$ converges in probability to θ as $N \to \infty$. $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\sigma}^2_{N-1}$ are all consistent.

2.2 The method of moments

The k'th moment of a distribution is

$$\mu_k = E[X^k] \tag{15}$$

The k'th **sample moment** is

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{N} X_n^k \tag{16}$$

The **method of moments** is simply to equate $\mu_k = \hat{\mu_k}$ for the first few moments (the minimum number necessary) and then to solve for θ. Although these estimators are not optimal (in a sense to be defined later), they are consistent, and are simple to compute, so they can be used to initialize other methods that require iterative numerical routines. Below we give some examples.

2.2.1 Bernoulli

Since $\mu_1 = E(X) = \theta$, and $\hat{\mu}_1 = \frac{1}{N}\sum_{n=1}^{N} X_n$, we have

$$\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{17}$$

2.2.2 Univariate Gaussian

The first and second moments are

$$\begin{aligned} \mu_1 &= E[X] = \mu && (18)\\ \mu_2 &= E[X^2] = \mu^2 + \sigma^2 && (19) \end{aligned}$$

So $\sigma^2 = \mu_2 - \mu_1^2$. The corresponding estimates from the sample moments are

$$\begin{aligned} \hat{\mu} &= \overline{X} && (20)\\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \overline{X}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{X})^2 && (21) \end{aligned}$$

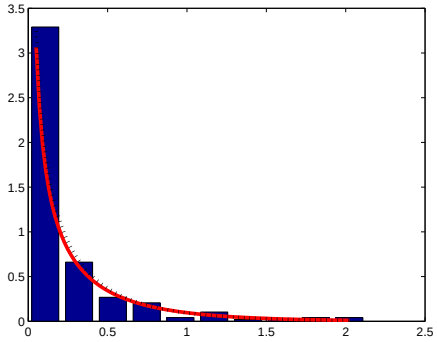

Figure 2: An empirical pdf of some rainfall data, with two Gamma distributions superimposes. Solid red line = method of moment. dotted black line = MLE. Figure generated by `rainfallDemo`.

2.2.3 Gamma distribution

Let $X \sim Ga(a, b)$. The first and second moments are

$$\mu_1 = \frac{a}{b} \tag{22}$$

$$\mu_2 = \frac{a(a+1)}{b^2} \tag{23}$$

To apply the method of moments, we must express a and b in terms of μ_1 and μ_2. From the second equation

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{b} \tag{24}$$

or

$$b = \frac{\mu_1}{\mu_2 - \mu_1^2} \tag{25}$$

Also

$$a = b\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \tag{26}$$

Since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$, we have

$$\hat{b} = \frac{\overline{x}}{\hat{\sigma}^2}, \hat{a} = \frac{\overline{x}^2}{\hat{\sigma}^2} \tag{27}$$

As an example, let us consider a data set from the book by [Ric95, p250]. The data records the average amount of rain (in inches) in southern Illinois during each storm over the years 1960 - 1964. If we plot its empirical pdf as a histogram, we get the result in Figure 2. This is well fit by a Gamma distribution, as shown by the superimposed lines. The solid line is the pdf with parameters estimated by the method of moments, and the dotted line is the pdf with parameteres estimated by maximum likelihood. Obviously the fit is very similar, even though the parameters are slightly different numerically:

$$\hat{a}_{mom} = 0.3763, \hat{b}_{mom} = 1.6768, \hat{a}_{mle} = 0.4408, \hat{b}_{mle} = 1.9644 \tag{28}$$

This was implemented using `rainfallDemo`.

2.3 Maximum likelihood estimates

A **maximum likelihood estimate** (MLE) is a setting of the parameters θ that makes the data as likely as possible:

$$\hat{\theta}_{mle} = \arg\max_{\theta} p(D|\theta) \tag{29}$$

Since the data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is iid, the likelihood factorizes

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{30}$$

It is often more convenient to work with log probabilities; this will not change $\arg\max L(\boldsymbol{\theta})$, since log is a monotonic function. Hence we define the **log likelihood** as $\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta})$. For iid data this becomes

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{31}$$

The mle then maximizes $\ell(\boldsymbol{\theta})$.

MLE enjoys various theoretical properties, such as being consistent and **asymptotically efficient**, which means that (roughly speaking) the MLE has the smallest variance of all well-behaved estimators (see [Was04, p126] for details). Therefore we will use this technique quite widely. We consider several examples below that we will use later.

2.3.1 Univariate Gaussians

Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$p(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{32}$$

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{33}$$

To find the maximum, we set the partial derivatives to 0 and solve. Starting with the mean, we have

$$\frac{\partial \ell}{\partial \mu} = -\frac{2}{2\sigma^2}\sum_{n}(x_n - \mu) = 0 \tag{34}$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=n}^{N} x_n \tag{35}$$

which is just the empirical mean. Similarly,

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2}\sigma^{-4}\sum_{n}(x_n - \hat{\mu}) - \frac{N}{2\sigma^2} = 0 \tag{36}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2 \tag{37}$$

$$= \frac{1}{N}\left[\sum_{n} x_n^2 + \sum_{n}\hat{\mu}^2 - 2\sum_{n} x_n\hat{\mu}\right] \tag{38}$$

$$= \frac{1}{N}[\sum_{n} x_n^2 + N\hat{\mu}^2 - 2N\hat{\mu}^2] = \frac{1}{N}[\sum_{n} x_n^2 + N(\frac{1}{N}\sum_{n} x_n)^2 - 2N(\frac{1}{N}\sum_{n} x_n)^2] \tag{39}$$

$$= \frac{1}{N}\sum_{n} x_n^2 - (\frac{1}{N}\sum_{n} x_n)^2 - = \frac{1}{N}\sum_{n} x_n^2 - (\hat{\mu})^2 \tag{40}$$

since $\sum_n x_n = N\hat{\mu}$. This is just the empirical variance.

2.3.2 Bernoullis

Let $X \in \{0, 1\}$. Given $\mathcal{D} = (x_1, \ldots, x_N)$, the likelihood is

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \tag{41}$$

$$= \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} \tag{42}$$

$$= \theta^{N_1}(1-\theta)^{N_2} \tag{43}$$

where $N_1 = \sum_i x_i$ is the number of heads and $N_2 = \sum_i (1 - x_i)$ is the number of tails The log-likelihood is

$$L(\theta) = \log p(D|\theta) = N_1 \log \theta + N_2 \log(1-\theta) \tag{44}$$

Solving for $\frac{dL}{d\theta} = 0$ yields

$$\theta_{ML} = \frac{N_1}{N} \tag{45}$$

the empirical fraction of heads.

Suppose we have seen 3 tails out of 3 trials. Then we predict that the probability of heads is zero:

$$\theta_{ML} = \frac{N_1}{N_1 + N_2} = \frac{0}{0+3} \tag{46}$$

This is an example of the **sparse data problem**: if we fail to see something in the training set, we predict that it can never happen in the future. Later we will consider Bayesian and MAP point estimates, which avoid this problem.

2.3.3 Multinomials

The log-likelihood is

$$\ell(\theta; D) = \log p(D|\theta) = \textstyle\sum_k N_k \log \theta_k \tag{47}$$

We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a **Lagrange multiplier**. The constrained cost function becomes

$$\tilde{\ell} = \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right) \tag{48}$$

Taking derivatives wrt θ_k yields

$$\frac{\partial \tilde{\ell}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0 \tag{49}$$

Taking derivatives wrt λ yields the original constraint:

$$\frac{\partial \tilde{\ell}}{\partial \lambda} = \left(1 - \sum_k \theta_k\right) = 0 \tag{50}$$

Using this sum-to-one constraint we have

$$N_k = \lambda \theta_k \tag{51}$$

$$\sum_k N_k = \lambda \sum_k \theta_k \tag{52}$$

$$N = \lambda \tag{53}$$

$$\hat{\theta}_k = \frac{N_k}{N} \tag{54}$$

Hence $\hat{\theta}_k$ is the fraction of times k occurs. If we did not observe $X = k$ in the training data, we set $\hat{\theta}_k = 0$, so we have the same sparse data problem as in the Bernoulli case (in fact it is worse, since K may be large, so it is quite likely that we didn't see some of the symbols, especially if our data set is small).

2.3.4 Gamma distribution

The pdf for $Ga(a, b)$ is

$$Ga(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} \tag{55}$$

So the log likelihood is

$$\begin{aligned} \ell(a, b) &= \sum_n [a \log b + (a-1) \log x_n - b x_n - \log \Gamma(a)] &(56)\\ &= N a \log b + (a-1) \sum_n \log x_n - b \sum_n x_n - N \log \Gamma(a) &(57)\end{aligned}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial a} &= N \log b + \sum_n \log x_n - N \frac{\Gamma'(a)}{\Gamma(a)} &(58)\\ \frac{\partial \ell}{\partial b} &= \frac{Na}{b} - \sum_n x_n &(59)\end{aligned}$$

where

$$\frac{\partial}{\partial a} \log \Gamma(a) \stackrel{\text{def}}{=} \psi(a) = \frac{\Gamma'(a)}{\Gamma(a)} \tag{60}$$

is the **digamma** function (which in matlab is called `psi`). Setting $\frac{\partial \ell}{\partial b} = 0$ we fund

$$\hat{b} = \frac{N \hat{a}}{\sum_n x_n} = \frac{\hat{a}}{\overline{x}} \tag{61}$$

But when we substitute this in to $\frac{\partial \ell}{\partial a} = 0$ we get a nonlinear equation for a:

$$0 = N \log \hat{a} - N \log \overline{x} + \sum_n \log x_n - N \psi(a) \tag{62}$$

This equation cannot be solved in closed form; an iterative method for finding the roots (such as Newton's method) must be used. We can start this process from the method of moments estimate.

In Matlab, if you type `type(which('gamfit'))`, you can look at its source code, and you will find that it estimates a by calling `fzero` with the following function:

$$lkeqn(a) = -\frac{1}{N} \sum_n \log X_n - \log \overline{X} - \log(a) + \psi(a) \tag{63}$$

It then substitutes $\hat{a}$ into Equation 61.

3 Sampling distributions

In addition to a point estimate, it is useful to have some measure of uncertainty. Note that the frequentist notion of uncertainty is quite different from the Bayesian. In the frequentist view, uncertainty means: how much would my estimate change if I had different data? This is called the **sampling distribution** of the estimator. In the Bayesian view, uncertainty means: how much do I believe my estimate given the current data? This is called the **posterior distribution** of the parameter. In otherwords, the frequentist is concerned with $E_{\mathcal{D}}[\hat{\theta}|\mathcal{D}]$ (and its spread), whereas the Bayesian is concerned with $E_\theta[\theta|\mathcal{D}]$ (and its spread). Sometimes these give the same answers, but not always. (The differences between Bayesian and frequentist statistics become more obvious when we consider problems such as hypothesis testing.)

The distribution of $\hat{\theta}$ is called the **sampling distribution**. Its standard deviation is called the **standard error**:

$$se(\hat{\theta}) = \sqrt{\text{Var } (\hat{\theta})} \tag{64}$$

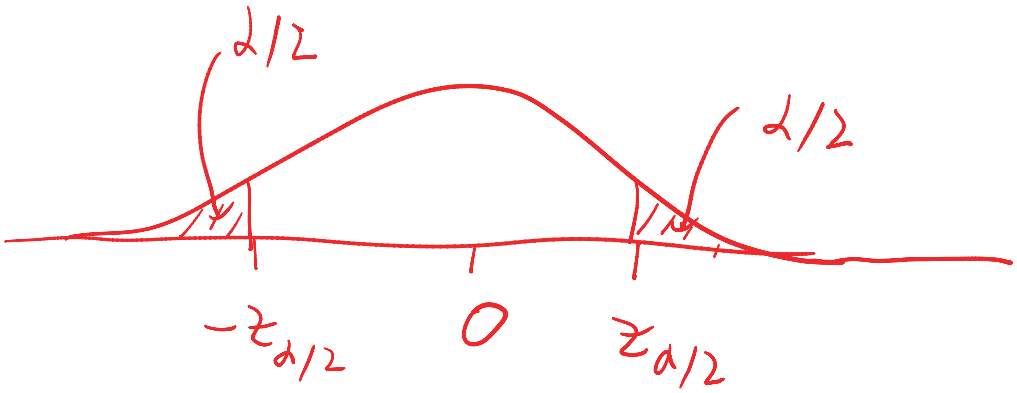

Figure 3: A $\mathcal{N}(0,1)$ distribution with the $z_{\alpha/2}$ cutoff points shown. The central non shaded area contains $1-\alpha$ of the probability mass. If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96 \approx 2$.

Often the standard error depends on the unknown distribution. In such cases, we often estimate it; we denote such estimates by $\hat{se}$.

Later we will see that many sampling distributions are approximately Gaussian as the sample size goes to infinity. More precisely, we say an estimator is **asymptotically Normal** if

$$\frac{\hat{\theta}_N - \theta}{se} \rightsquigarrow \mathcal{N}(0,1) \tag{65}$$

(where $\rightsquigarrow$ here means converges in distribution).

3.1 Confidence intervals

A $1-\alpha$ **confidence interval** is an interval $C_n = (a,b)$ where a and b are functions of the data $X_{1:N}$ such that

$$P_\theta(\theta \in C_N) \geq 1-\alpha \tag{66}$$

In other words, (a,b) traps θ with probability $1-\alpha$. Often people use 95% confidence intervals, which corresponds to $\alpha = 0.05$.

If the estimator is asymptotically normal, we can use properties of the Gaussian distribution to compute an approximate confidence interval.

Theorem 3.1. *Suppose that $\hat{\theta}_N \approx \mathcal{N}(\theta, \hat{se}^2)$. Let Φ be the cdf of a standard Normal, $Z \sim \mathcal{N}(0,1)$, and let $z_{\alpha/2} = \Phi^{-1}(1-(\frac{\alpha}{2}))$, so that $P(Z > z_{\alpha/2}) = \alpha/2$. By symmetry of the Gaussian, $P(Z < -z_{\alpha/2}) = \alpha/2$, so $P(-z_{\alpha/2} < Z < z_{\alpha/a}) = 1-\alpha$. (see Figure 3). Let*

$$C_N = (\hat{\theta}_N - z_{\alpha/2}\hat{se}, \hat{\theta}_N + z_{\alpha/2}\hat{se}) \tag{67}$$

Then

$$P(\theta \in C_N) \rightarrow 1-\alpha \tag{68}$$

Proof. Let $Z_N = (\hat{\theta} - \theta)/\hat{se}$. By assumption, $Z_n \rightsquigarrow Z$. Hence

$$\begin{aligned} P(\theta \in C_N) &= P(\hat{\theta}_N - z_{\alpha/2}\hat{se} < \theta < \hat{\theta}_N + z_{\alpha/2}\hat{se}) && (69)\\ &= P(z_{\alpha/2}\hat{se} < \frac{\hat{\theta_N} - \theta}{\hat{se}} < z_{\alpha/2}\hat{se}) && (70)\\ &\rightarrow P(z_{\alpha/2}\hat{se} < Z < z_{\alpha/2}\hat{se}) && (71)\\ &= 1-\alpha && (72) \end{aligned}$$

□

For 95% confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$, which is why people often express their results as $\hat{\theta}_N \pm 2\hat{se}$.

3.2 The counter-intuitive nature of confidence intervals

Note that saying that "$\hat{\theta}$ has a 95% confidence interval of $[a, b]$" does *not* mean $P(\hat{\theta} \in [a, b]|\mathcal{D}) = 0.95$. Rather, it means

$$p_{D' \sim P(\cdot|\theta)}(\hat{\theta} \in [a(D'), b(D')]) = 0.95 \tag{73}$$

i.e., if we were to repeat the experiment, then 95% of the time, the true parameter would be in the $[a, b]$ interval. To see that these are not the same statement, consider this example (from [Mac03, p465]). Suppose we draw two integers from

$$p(x|\theta) = \begin{cases} 0.5 & \text{if } x = \theta \\ 0.5 & \text{if } x = \theta + 1 \\ 0 & \text{otherwise} \end{cases} \tag{74}$$

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

$$(39, 39), (39, 40), (40, 39), (40, 40) \tag{75}$$

Let $m = min(x_1, x_2)$ and define a CI as

$$[a(D), b(D)] = [m, m], \tag{76}$$

For the above samples this yields

$$[39, 39], \quad [39, 39], \quad [39, 39], \quad [40, 40] \tag{77}$$

which is clearly a 75% CI. However, if $D = (39, 39)$ then $p(\theta = 39|D) = P(\theta = 38|D) = 0.5$. And if $D = (39, 40)$ then $p(\theta = 39|D) = 1.0$. Thus even if we know $\theta = 39$, we only have 75% "confidence" in this fact. Later we will see that Bayesian **credible intervals** give the more intuitively correct answer.

3.3 Sampling distribution for Bernoulli MLE

Consider estimating the parameter of a Bernoulli using $\hat{\theta} = \frac{1}{N}\sum_n X_n$. Since $X_n \sim Be(\theta)$, we have $S = \sum_n X_n \sim Binom(N, \theta)$. So the sampling distribution is

$$p(\hat{\theta}) = p(S = N\hat{\theta}) = \text{Binom}(N\hat{\theta}|N, \theta) \tag{78}$$

We can compute the mean and variance of this distribution as follows.

$$E\hat{\theta} = \frac{1}{N}E[S] = \frac{1}{N}N\theta = \theta \tag{79}$$

so we see this is an unbiased estimator. Also

$$\begin{aligned} \text{Var } \hat{\theta} &= \text{Var } [\frac{1}{N}\sum_n X_n] && (80) \\ &= \frac{1}{N^2}\sum_n \text{Var } [X_n] && (81) \\ &= \frac{1}{N^2}\sum_n \theta(1-\theta) && (82) \\ &= \frac{\theta(1-\theta)}{N} && (83) \end{aligned}$$

So

$$se = \sqrt{\theta(1-\theta)/N} \tag{84}$$

and

$$\hat{se} = \sqrt{\hat{\theta}(1-\hat{\theta})/N} \tag{85}$$

We can compute an exact confidence interval using quantiles of the Binomial distribution. However, for reasonably small N, the Binomial is well approximated by a Gaussian, so $\hat{\theta}_N \approx \mathcal{N}(\theta, \hat{se}^2)$. So an approximate $1-\alpha$ confidence interval is

$$\hat{\theta}_N \pm z_{\alpha/2}\hat{se} \tag{86}$$

3.4 Sampling distribution for the Gaussian mean

The MLE for the mean is

$$\hat{\mu} = \overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{87}$$

It can be shown (see e.g. [Ric95, p180]) that if $X_n \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\sqrt{N}(\overline{X}-\mu)}{S} \sim t_{N-1} \tag{88}$$

where

$$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2 \tag{89}$$

and t_{N-1} is the **Student** t **distibution** with $N-1$ degrees of freedom (see Section 5.1). Equivalently, we can define the distribution using the generalized T distribution

$$\hat{\mu} \sim t(N-1, \mu, S^2/N) \tag{90}$$

where

$$X \sim t(\nu, \mu, \sigma^2) \iff \frac{X-\mu}{\sigma} \sim t_\nu \tag{91}$$

Let $t_{N-1}(\alpha/2)$ be the $\alpha/2$ quantile. Since the t distribution is symmetric,

$$P(t_{N-1}(\alpha/2) \leq \frac{\sqrt{N}(\overline{X}-\mu)}{S} \leq -t_{N-1}(\alpha/2)) = 1-\alpha \tag{92}$$

So an exact $1-\alpha$ confidence interval is

$$\hat{\mu} = \overline{X} \pm \frac{S}{\sqrt{N}}\, t_{N-1}(\alpha/2) \tag{93}$$

3.5 Sampling distribution for the Gaussian variance

Now let us derive a confidence interval for σ^2. The MLE is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n}(X_n - \overline{X})^2 \tag{94}$$

It can be shown (see e.g. [Ric95, p180])

$$\frac{N\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-1} \tag{95}$$

where χ^2_{N-1} is the chi-squared distribution with $N-1$ degrees of freedom (see Section 5.2). Let $\chi^2_{N-1}(\alpha)$ denote the α quantile. Then

$$P(\chi^2_{N-1}(1-\alpha/2) \leq \frac{N\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{N-1}(\alpha/2)) = 1-\alpha \tag{96}$$

So an exact $1-\alpha$ confidence interval for σ^2 is

$$\left(\frac{N\hat{\sigma}^2}{\chi^2_{N-1}(\alpha/2)}, \frac{N\hat{\sigma}^2}{\chi^2_{N-1}(1-\alpha/2)},\right) \tag{97}$$

Note that the χ^2 distribution is not symmetric, so this interval is not of the form $\hat{\sigma}^2 \pm c$. Note also that some books write $(N-1)S^2$ in the denominator instead of $N\hat{\sigma}^2$.

3.6 Large sample theory for the MLE

Computing the exact sampling distribution of an estimator can often be difficult. Fortunately, it can be shown that, for certain models[1], as the sample size tends to infinity, the sampling distribution becomes Gaussian. We say the estimator is **asymptotically Normal**.

Define the **score function** as the derivative of the log likelihood

$$s(X, \theta) = \frac{\partial \log p(X|\theta)}{\partial \theta} \tag{98}$$

Define the **Fisher information** to be

$$\begin{aligned} I_N(\theta) &= \text{Var}\left(\sum_{n=1}^{N} s(X_n, \theta)\right) && (99) \\ &= \sum_n \text{Var}\left(s(X_n, \theta)\right) && (100) \end{aligned}$$

Intuitively, this measures the stability of the MLE wrt variations in the data set (recall that X is random and θ is fixed). It can be shown that $I_N(\theta) = NI_1(\theta)$. We will write $I(\theta)$ for $I_1(\theta)$. We can rewrite this in terms of the second derivative, which measures the curvature of the likelihood.

Theorem 3.2. *Under suitable smoothness assumptions on p, we have*

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log p(X|\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \log p(X|\theta)\right] \tag{101}$$

The proof is in Section 5.3 in the appendix. We can now prove our main theorem.

Theorem 3.3. *Under appropriate regularity conditions, the MLE is asymptotically Normal.*

$$\hat{\theta}_N \sim \mathcal{N}(\theta, se^2) \tag{102}$$

where

$$se \approx \sqrt{1/I_N(\theta)} \tag{103}$$

is the standard error of $\hat{\theta}_N$. Furthemore, this result still holds if we use the estimated standard error

$$\hat{se} \approx \sqrt{1/I_N(\hat{\theta})} \tag{104}$$

The proof is in Section 5.3 in the appendix. The basic idea is to use a Taylor series expansion of ℓ' around θ.

The intuition behind this result is as follows. The **asymptotic variance** is given by

$$\frac{1}{NI(\theta)} = -\frac{1}{E\ell''(\theta)} \tag{105}$$

so when the curvature at the MLE $|\ell''(\hat{\theta})|$, is large, then the variance is low, whereas if the curvature is nearly flat, the variance is high. (Note that $\ell''(\hat{\theta}) < 0$ since $\hat{\theta}$ is a maximum of the log likelihood.) Intuitively, the curvature is large if the parameter is "well determined".[2]

For example, consider $X_n \sim Be(\theta)$. The MLE is $\hat{\theta} = \frac{1}{N}\sum_n X_n$ and the log likelihood is

$$\log p(x|\theta) = x \log \theta + (1 - x) \log(1 - \theta) \tag{106}$$

[1] Intuively, the requirement is that each parameter in the model get to "see" an infinite amount of data.

[2] From the Bayesian standpoint, the equivalent statement is that the parameter is well determined if the posterior uncertainty is small. (Sharply peaked Gaussians have lower entropy than flat ones.) This is arguably a much more natural intepretation, since it talks about our uncertainty about θ, rather than variance induced by changing the data.

so the score function is

$$s(X,\theta) = \frac{X}{\theta} - \frac{1-X}{1-\theta} \tag{107}$$

and

$$-s'(X,\theta) = \frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2} \tag{108}$$

Hence

$$I(\theta) = E_\theta(-s'(X,\theta)) = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} \tag{109}$$

So

$$\hat{se} = \frac{1}{\sqrt{I_N(\hat{\theta}_N)}} = \frac{1}{\sqrt{NI(\hat{\theta}_N)}} = \left(\frac{\hat{\theta}(1-\hat{\theta})}{N}\right)^{\frac{1}{2}} \tag{110}$$

which is the same as the result we derived in Equation 85.

In the multivariate case, the **Fisher information matrix** is defined as

$$I_N(\theta) = -\begin{pmatrix} E_\theta H_{11} & \cdots & E_\theta H_{1p} \\ \ddots & & \\ E_\theta H_{p1} & \cdots & E_\theta H_{pp} \end{pmatrix} \tag{111}$$

where

$$H_{jk} = \frac{\partial^2 \ell_N}{\partial\theta_j \partial\theta_k} \tag{112}$$

is the Hessian of the log likelihood. The multivariate version of Theorem 3.3 is

$$\hat{\theta} \sim \mathcal{N}(\theta, I_N^{-1}(\theta)) \tag{113}$$

3.7 Parametric bootstrap

The parametric **bootstrap** is a simple **Monte Carlo** technique to approximate the sampling distribution, and thus to derive standard errors, confidence intervals, etc. This is particularly useful in cases where the estimator is a complex function of the true parameters (e.g., $\hat{b}_{mle}$ for the Gamma distribution had to be computed numerically: see Section 2.3.4). The idea is simple. If we knew the true parameters θ, we could generate many samples (say B) of size N from the true distribution, $X_n^b \sim p(X|\theta)$, for $b = 1:B$, $n = 1:N$. We could then compute our estimator from each sample, $\hat{\theta}^b = f(X^b)$. and use the empirical distribution of the resulting samples as our estimate of the sampling distribution. Since θ is unknown, the idea of the bootstrap is to generate the samples using $\hat{\theta}$ instead.

In Figure 4, we show how to use the bootstrap to approximate the sampling distribution of $\hat{a}$ and $\hat{b}$ for a Gamma distribution applied to the rainfall data. We compare the method of moments estimators with the MLE, and see that the variance of the MLE is lower. See `rainfallBootstrapDemo`.

3.8 Cramer-Rao lower bound

We showed earlier that the mean squared error of an estimator is given by the squared bias plus variance:

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = \text{Var}\,[\hat{\theta}] + (E[\hat{\theta}-\theta])^2 \tag{114}$$

If two estimators are unbiased (or have the same bias), we define the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ as

$$\text{eff}(\hat{\theta},\tilde{\theta}) = \frac{\text{Var}\,(\tilde{\theta})}{\text{Var}\,(\hat{\theta})} \tag{115}$$

If the efficiency is greater than 1, then $\hat{\theta}$ has smaller variance than $\tilde{\theta}$, so $\hat{\theta}$ would be preferred. If we approximate the variance using Var $[\hat{\theta}] \approx \frac{1}{NI(\hat{theta})}$, this is called the **asymptotic relative efficiency** (ARE). It can be shown that the MLE has the smallest asymptotic variance of any estimator. This is called **asymptotically optimal**.

It is natural to ask whether there is a lower bound on the variance of any estimator. In fact there is.

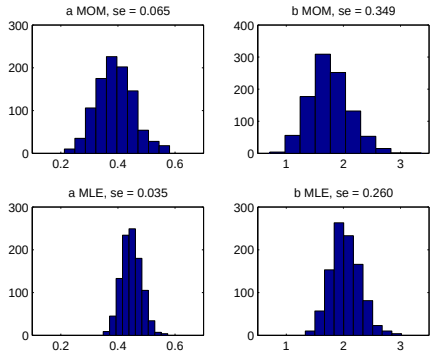

Figure 4: A bootstrap approximation to the sampling distribution of $\hat{a}$ and $\hat{b}$ for a Gamma distribution. MOM = method of moments. MLE = maximum likelihood estimation. Figure generated by `rainfallBootstrapDemo`.

	Accept	Reject
H_0 true	True detection (hit) $1-\alpha$	Type I error (false negative, miss) α
H_0 false	Type II error (false positive) β	True rejection $1-\beta$

Table 1: Summary of the various types of outcomes in a binary hypothesis test.

Theorem 3.4 (**Cramer-Rao inequality**.)**.** *Let $X_1, \ldots, X_n \sim p(X|\theta)$ and $T = T(x_1, \ldots, x_N)$ be an unbiased estimator of θ. Then, under various smoothness assumptions on $p(X|\theta)$, we have*

$$Var\,[T] \geq \frac{1}{NI(\theta)} \tag{116}$$

A proof can be found e.g., in [Ric95, p275].

Note, however, that biased estimators often have lower overall MSE than unbiased estimators.

4 Hypothesis testing

Hypothesis testing is binary decision making applied to the parameters of a probability model. Let H_0 be the **null hypothesis** that $\theta = \theta_0$. and H_1 be the **alternative hypothesis** that $\theta \neq \theta_0$ (a **two-sided test**) or $\theta > \theta_0$ (a **one-sided test**). There are two possible errors we can make, analogous to false positives and false negatives: see Table 1. The **Neyman-Pearson paradigm** fixes the **type I error** rate to α and infers the corresponding **type II error** rate, as we will see below.

4.1 Neyman-Pearson paradigm

Let x be data generated from $p(X|\theta)$. Since x may be multidimensional, we usually summarize it with a scalar quantity called a **test statistic**, $T(x)$ (e.g., a likelihood ratio, see below). We define the rejection region R in terms of a **critical value** c as follows:

$$R = \{x : T(x) > c\} \tag{117}$$

The decision rule becomes: reject H_0 iff $T(x) > c$. We choose the threshold c s.t. the probability of falsely rejecting the null hypothesis (making a **type I error**) is set to α:

$$p(\text{we reject } H_0 | H_0 \text{ is true}) = \alpha \tag{118}$$

This is called the **significance level**. We usually pick $\alpha = 0.05$.

The probability of a **type II error** (falsely accepting the null) depends on the true underyling parameter θ^*. Define

$$\beta(\theta) = p(\text{we accept } H_0 | H_1 \text{ is true and has parameter } \theta) \tag{119}$$

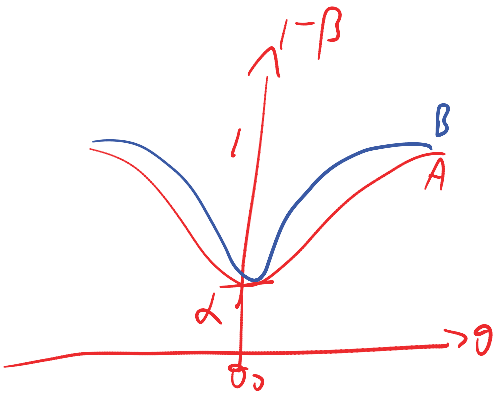

Figure 5: Two hypothetical two-sided power curves. Based on Figure 6.3.5 of [LM86].

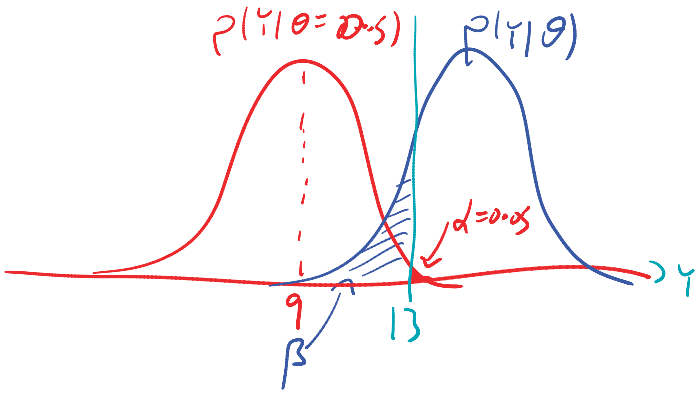

Figure 6: Two binomial distributions. Based on Figure 6.3.2 of [LM86].

Hence $1-\beta$ is the probability we reject H_0 given that H_1 is true; this is called the **power** of the test, and is the ability to correctly recognize that the null is wrong. This depends on the true θ^* and the null value θ_0:

$$1-\beta(\theta) = p(X \in R|\theta) \tag{120}$$

The least power occurs if $\theta^* = \theta_0$, since then

$$1-\beta(\theta^*) = p(X \in R|\theta^*) = p(X \in R|\theta_0) = \alpha \tag{121}$$

As the true value differs from θ_0, the power tends towards 1. See Figure 5, where method B dominates method A. A test with highest power under H_A amongst all tests with significance level α is called a **most powerful test**. Finding most powerful test is hard and, in many cases, most powerful tests don't even exist.

Note that there is duality between confidence intervals and hypothesis testing. In particular, confidence interval is the set of all values of X such that the null hypothesis $H_0 : \theta = \theta_0$ is accepted. To see this, note that if $A(\theta_0)$ is the acceptance region at significance level α, then by definition

$$P(X \in A(\theta_0)|\theta = \theta_0) = 1-\alpha \tag{122}$$

which is a $1-\alpha$ confidence interval.

4.2 Example

Let us now consider a simple example. Let $X \sim Bino(n = 18, \theta)$. We wish to test $H_0 : \theta = \theta_0 = 0.5$ vs $H_1 : \theta > 0.5$. We will use $T(X) = X$, so the **sampling distribution** of the test statistic is also binomial. We can find

	Ineffective	effective	
"Not sig"	171	4	175
"Sig"	9	16	25
	180	20	200

Table 2: Some statistics of a hypothetical clinical trial. Source: [SAM04, p74].

the critical value by solving

$$0.05 = P(X \geq c|H_0) = \sum_{x=c}^{n} \binom{n}{x} (\theta_0)^x (1-\theta_0)^{n-x} \tag{123}$$

which yields $c = 13$. See Figure 6. The corresponding type II error rate depends on θ. If $\theta = 0.7$, we have

$$\beta = P(\text{type II error}|\theta = 0.7) = P(X \leq 12|\theta = 0.7) = binocdf(12, 18, 0.7) = 0.47 \tag{124}$$

4.3 Practically vs statistically significant

When we reject H_0 we often say the result is **statistically significant** at level α. However, the result may be statistically significant but not practically significant. One way to see this is to notice that any confidence interval that excludes θ_0 corresponds to rejecting H_0, but an interval that just barely excludes θ_0 is not as practically significant as one that is far from θ_0. Thus confidence intervals are often more informative than tests.

The other thing to remember is that if $\alpha = 0.05$, the test will have false positives 1/20 times. Thus any results that declared as "significant" might be errors, and should really be viewed as noisy signals to be combined with other information. Consider this example from [SAM04, p74]. Suppose 200 clinical trials are carried out of some drug. The null hypothesis is that it is not effective. Suppose the type I error rate is $\alpha = 0.05$ and the type II error rate is $\beta = 0.2$. Suppose also that we have prior knowledge, based on past experience, that most (say 90%) drugs are ineffective; thus we should view with caution any claims of effectiveness.

Now consider the data in Table 2. We see that only 20/200=10% of trials are truly effective, even though 25 trials are claimed to be significant. Using Bayes' rule, we can see that 9/25=36% of trials with "significant results" are ineffective:

$$p(H_0|\text{'significant'}) = \frac{p(\text{'significant'}|H_0)p(H_0)}{p(\text{'significant'}|H_0)p(H_0) + p(\text{'significant'}|H_1)p(H_1)} \tag{125}$$

$$= \frac{p(\text{type I error})p(H_0)}{p(\text{type I error})p(H_0) + (1 - p(\text{type II error}))p(H_1)} \tag{126}$$

$$= \frac{\alpha p(H_0)}{\alpha p(H_0) + (1-\beta)p(H_1)} \tag{127}$$

Using $\alpha = 0.05$, $\beta = 0.2$, $p(H_0) = 0.9$ and $p(H_1) = 0.1$, we find $p(H_0|\text{'significant'}) = 0.36$.

4.4 p-values

The **p-value** of a test is defined as the smallest value of α for which the null hypothesis will be rejected:[3]

$$\text{pval}(\mathcal{D}) = \min_{\alpha} T(\mathcal{D}) \in R_\alpha \tag{128}$$

If $R_\alpha = \{T : T > t^*_\alpha\}$, we can write this as

$$\text{pval}(\mathcal{D}) = \min_{\alpha} T(\mathcal{D}) > t^*_\alpha \tag{129}$$

In other words, we decrease α as much as possible (thereby reducing the size of the rejection region), so long as the observed statistic still falls in the rejection region. In the ESP example, the p-value is $P(X \geq 7|H_0) = 0.1719$, where

[3]For a good website discussing how to interpret p-values, see `http://www.stat.duke.edu/~berger/p-values.html`.

7 is the observed statistic. Thus the p-value measures the probability (under H_0) of getting a statistic *as large or larger* than the one actually observed.

Informally, if the p-value is small, then either the null hypothesis holds and an extremely rare event has happened, or the null hypothesis is false. Thus the smaller the p-value, the less likely H_0 is to be true. However, the p-value is *not* equal to $p(H_0|\mathcal{D})$. (We will discuss how to compute such Bayesian quantities later.) Also, a large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: either H_0 is true, or H_0 is false but the test has low power.

Using p-values is often better than deciding to accept or reject a hypothesis, since it avoids thresholding at an arbitrary value of α (such as 0.05). Procedurally, instead of fixing α and determining a critical value t^*_α and then deciding whether to accept or reject by comparing with the observed t, we instead solve for α by setting the critical value to the observed value.

4.5 Example: is the eurocoin biased?

Consider this example from [Mac03]. It was observed that, when spun on edge $N = 250$ times, a Belgian one-euro coin came up heads $Y = 141$ times and tails 109. A statistician called Barry Blight commented on this event in *The Guardian* in 2002: "It looks very suspicious to me. We can reject the null hypothesis (that the coin is unbiased) with a **significance level** of 5%".

Let us compute the p-value of the data using a two-sided test:

$$\begin{aligned}
\text{pval} &= P(Y \geq 141|H_0) + P(Y \leq 109|H_0) && (130)\\
&= (1 - P(Y < 141|H_0)) + P(Y \leq 109|H_0) && (131)\\
&= (1 - P(Y \leq 140|H_0)) + P(Y \leq 109|H_0) && (132)\\
&= 1 - \text{binocdf}(140, 250, 0.5) + \text{binocdf}(109, 250, 0.5) && (133)\\
&= 0.0497 && (134)
\end{aligned}$$

Hence it seems like the coin is indeed biased. However, later we will perform a Bayesian analysis that suggest that the coin is *not* biased!

4.6 Hypothesis testing violates the likelihood principle

There is something a little odd about classical hypothesis testing. Consider, for example, the p-value computation in the coin example (Section 4.5). Why do we care about **tail probabilities**, such as

$$P(Y \geq 141|H_0) = P(Y = 141|H_0) + P(Y = 142|H_0) + \cdots \quad (135)$$

when the number of heads we observed was 141, not 142 or larger?

To understand why frequentists use tail probabilities, consider hypothesis testing for continuous quantities, e.g., testing if the mean of a Gaussian is μ given an observation x. The probability of observing exactly the same value x again in future trials is zero (since X is a continuous random variable), so a frequentist is forced to ask questions about $p(X > x|\mu)$, whereas a Bayesian would compute the more natural quantity $p(\mu|x)$, as we will see later. That is, a Bayesian conditions on the observed data, whereas classical statistics reasons in terms of future data.

The problem with p-values, and all classical statistical methods that rely on tail probabilities — including confidence intervals, hypothesis testing, and statistical learning theory — is that they violate the **likelihood principle**, which says, roughly: In order to choose between hypotheses given observed data D, one should ask how likely the observed data are under each hypothesis; one should not consider other data besides D. The famous Bayesian statistician Jeffreys has said "The use of P-values implies that a hypothesis that may be true can be rejected because it has not predicted observable results that have not actually occurred."

It can be shown that one consequence of the likelihood principle is the **stopping rule principle**, which says that the decision on when to stop collecting your data should not affect your inferences. For example, suppose we observe the data stream of $X = 3$ tails in $N = 12$ coin tossing trials

$$D = (H, H, H, T, H, H, H, H, T, H, H, T) \quad (136)$$

Is there evidence of bias, $\theta > 0.5$? What if I told you the data was generated by tossing coins until we got $X = 3$ tails? Would this change your opinion? Presumably not. Data is data, no matter how it is collected (assuming the data collection mechanism does not physically alter the coin tossing mechanism).

However, the frequentist approach violates the stopping rule principle, and therefore gives different answers depending on how the data was collected. To see this, let $X = 3$ be the random variable, and $N = 12$ be a fixed constant. Define $H_0 : P_h = 0.5$. Then, at the 5% level, there is no significant evidence of bias:

$$P(X \leq 3|N = 12, H_0) = \sum_{x=0}^{3} \binom{N}{x} \tfrac{1}{2}^N = 0.07 \tag{137}$$

However, now let $X = x = 3$ be the fixed constant and $N = n = 12$ be the random variable.

$$P(N \geq n|X = x, H_0) = \binom{n-1}{x-1} \tfrac{1}{2}^n = 0.03 \tag{138}$$

(This follows since the last trial always contains a tail (by definition), so we have to choose $n - 1$ locations for the other $x - 1$ tails.) So now there is significant evidence of bias! Later we will see that a Bayesian analysis gives the same answer in both cases, because Bayesian inference respects the likelihood principle and hence the stopping rule principle.

5 Appendix

5.1 Student T distribution

The non-central t-distribution is given as

$$t_\nu(x|\mu, \sigma^2) = c\left[1 + \frac{1}{\nu}(\frac{x-\mu}{\sigma})^2\right]^{-(\frac{\nu+1}{2})} \tag{139}$$

$$c = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}\sigma} \tag{140}$$

where c is the normalization consant. μ is the mean, $\nu > 0$ is the **degrees of freedom**, and $\sigma^2 > 0$ is the scale. See `studentTpdf`.

The distribution has the following properties:

$$\text{mean} = \mu,\ \nu > 1 \tag{141}$$

$$\text{mode} = \mu \tag{142}$$

$$\text{var} = \frac{\nu\sigma^2}{(\nu - 2)},\ \ \nu > 2 \tag{143}$$

Note: if $x \sim t(\mu, \nu, \sigma^2)$, then

$$\frac{x-\mu}{\sigma} \sim t_\nu \tag{144}$$

which corresponds to a standard t-distribution with $\mu = 0, \sigma^2 = 1$:

$$t_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2} \tag{145}$$

In Figure 7, we plot the density for different parameter values. As $\nu \to \infty$, the T approaches a Gaussian. T-distributions are like Gaussian distributions with **heavy tails**. Hence they are more robust to outliers (see Figure 8).

If $\nu = 1$, this is called a **Cauchy distribution**. This is an interesting distribution since if $X \sim Cauchy$, then $E[X]$ does not exist, since the corresponding integral diverges. Essentially this is because the tails are so heavy that samples from the distribution can get very from the center μ.

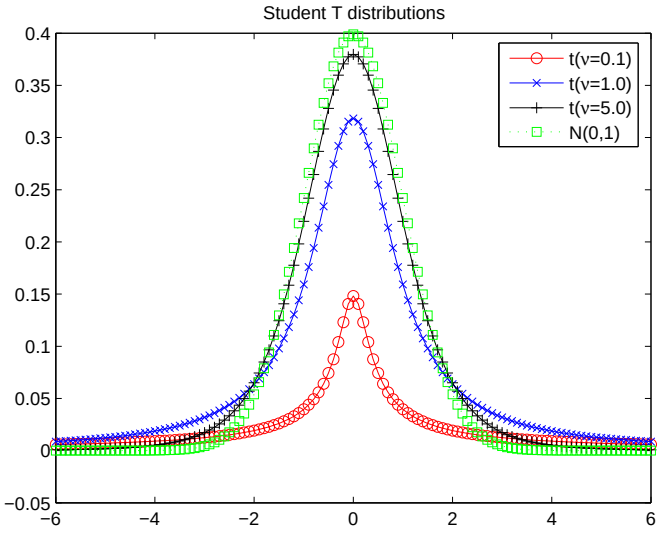

Figure 7: Student t-distributions $T(\mu, \sigma^2, \nu)$ for $\mu = 0$. The effect of σ is just to scale the horizontal axis. As $\nu \rightarrow \infty$, the distribution approaches a Gaussian. See `studentTplot`.

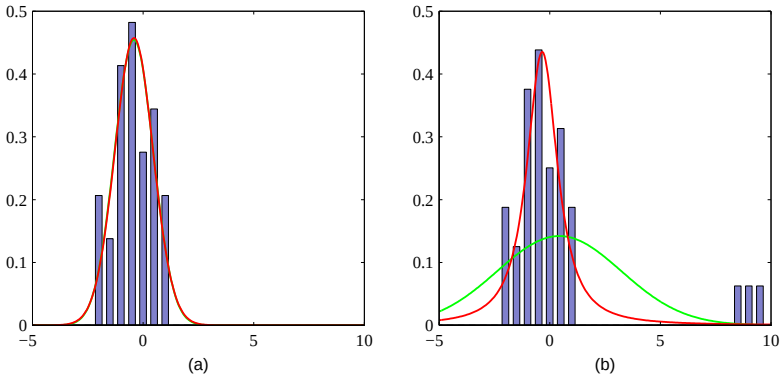

Figure 8: Fitting a Gaussian and a Student distribution to some data (left) and to some data with outliers (right). The Student distribution (red) is much less affected by outliers than the Gaussian (green). Source: [Bis06] Figure 2.16.

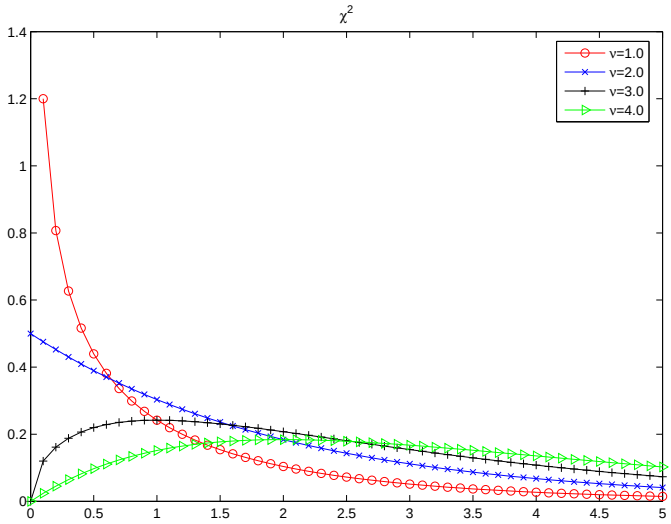

Figure 9: Some χ^2 distributions. These plots were produced by `chi2plot`.

It can be shown that the t-distribution is like an infinite sum of Gaussians, where each Gaussian has a different precision:

$$p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \tau^{-1}) Ga^{rate}(\tau|a, b) d\tau \quad (146)$$

$$= t_{2a}(x|\mu, \sigma^2 = b/a) \quad (147)$$

(See exercise 2.46 of [Bis06].)

5.2 Chi-squared distribution

The χ^2 **distribution** (pronounced "chi-squared") written $\chi^2_\nu(x)$, is just a special case of the Gamma distribution

$$\alpha = \text{shape} = \nu/2, \text{rate} = 1/2, \text{scale} = 2 \quad (148)$$

and can be written as

$$\chi^2_\nu(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} \quad (149)$$

ν is called the degrees of freedom. See Figure 9. As $\nu \to \infty$, we have $\chi^2_\nu \to \mathcal{N}(0, 1)$.

It has the following properties:

$$\text{mean} = \nu \quad (150)$$

$$\text{var} = 2\nu \quad (151)$$

The mode does not exist.

The chi-squared distribution can be derived as follows. Let $Q = X_1^2 + \cdots X_n^2$, where $X_i \sim \mathcal{N}(0, 1)$ are iid. Then $Q \sim \chi^2_n$. More generally, if $Y \sim \mathcal{N}(\mu, \Sigma)$ for $Y \in \mathbb{R}^n$, then the Mahalanobis distance $Q = (Y - \mu)^T \Sigma^{-1} (Y - \mu)$ has distribution $Q \sim \chi^2_n$. This is useful for plotting confidence ellipsoids for multivariate Gaussians, as we will see later.

5.3 Proofs for the large sample theory for MLEs

We start with a lemma.

Theorem 5.1.

$$E[s(X;\theta)] = 0. \tag{152}$$

Proof. This can be shown by noting that $1 = \int p(x|\theta)dx$. Differentiating both sides yields

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int p(x|\theta)dx = \int \frac{\partial}{\partial\theta}p(x|\theta)dx && (153)\\ &= \int \frac{\frac{\partial p(x|\theta)}{\partial\theta}}{p(x|\theta)}p(x|\theta)dx && (154)\\ &= \int \frac{\partial}{\partial\theta}[\log p(x|\theta)]p(x|\theta)dx && (155)\\ &= \int s(x,\theta)p(x|\theta)dx = E[s(X,\theta)] && (156) \end{aligned}$$

Hence

$$I(\theta) = \text{Var}\,[s(X,\theta)] = E[s^2(X,\theta)] \tag{157}$$

□

Hence we can show

Theorem 5.2. *Under suitable smoothness assumptions on p, we have*

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(X|\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\log p(X|\theta)\right] \tag{158}$$

Proof. From Equation 155, we know that

$$0 = \int \frac{\partial}{\partial\theta}[\log p(x|\theta)]p(x|\theta)dx \tag{159}$$

Taking second derivatives we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta}\log p(x|\theta)p(x|\theta)dx && (160)\\ &= \int\left[\frac{\partial^2}{\partial\theta^2}\log p(x|\theta)\right]p(x|\theta)dx + \int\left[\frac{\partial}{\partial\theta}\log p(x|\theta)\right]^2 p(x|\theta)dx && (161) \end{aligned}$$

□

We can now prove our main theorem.

Theorem 5.3. *Under appropriate regularity conditions, the MLE is asymptotically Normal.*

$$\hat{\theta}_N \sim \mathcal{N}(\theta, se^2) \tag{162}$$

where

$$se \approx \sqrt{1/I_N(\theta)} \tag{163}$$

is the standard error of $\hat{\theta}_N$. Furthemore, this result still holds if we use the estimated standard error

$$\hat{se} \approx \sqrt{1/I_N(\hat{\theta})} \tag{164}$$

Proof. Let $\ell = \sum_n \log p(x_n|\theta)$. Since $\hat{\theta}$ is an MLE, we have $\ell'(\hat{\theta}) = 0$. From a Taylor series expansion of ℓ' around θ, we have

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta) + (\hat{\theta} - \theta)\ell''(\theta) \tag{165}$$

$$(\hat{\theta} - \theta) \approx \frac{-\ell'(\theta)}{\ell''(\theta)} \tag{166}$$

$$N^{\frac{1}{2}}(\hat{\theta} - \theta) \approx \frac{N^{-\frac{1}{2}}\ell'(\theta)}{-N^{-1}\ell''(\theta)} = \frac{\text{TOP}}{\text{BOTTOM}} \tag{167}$$

First consider the TOP term. Let $Y_n = \frac{\partial}{\partial\theta} \log p(x_n|\theta)$. We have $E[Y_n] = 0$ from Theorem 5.1 so $E[\text{TOP}] = 0$. Also, Var $[Y_n] = I(\theta)$ so

$$\text{Var } [\text{TOP}] = \text{Var } [N^{-\frac{1}{2}} \sum_n Y_n] = \text{Var } [Y_n] = I(\theta) \tag{168}$$

Hence TOP, which is a sum of iid rv's, converges to a Gaussian by the central limit theorem:

$$\text{TOP} \rightsquigarrow W \sim \mathcal{N}(0, I(\theta)) \tag{169}$$

Now consider the denominator. Let $A_n = -\frac{\partial^2}{\partial\theta^2} \log p(X_n|\theta)$. By Theorem 3.2, $E[A_n] = I(\theta)$, so by the law of large numbers, BOTTOM converges to $I(\theta)$. We thus have

$$N^{\frac{1}{2}}(\hat{\theta} - \theta) \rightsquigarrow \frac{W}{I(\theta)} \sim \mathcal{N}(0, \frac{1}{I(\theta)}) \tag{170}$$

Hence

$$\hat{\theta} \sim \mathcal{N}(\theta, \frac{1}{NI(\theta)}) \tag{171}$$

□

5.4 More statistical tests

5.4.1 Likelihood ratio tests (LRT)

The previous examples have assumed the data is scalar. For vector-valued data, we can use the likelihood as a statistic. If we assume both H_0 and H_A are simple hypotheses, we define the **likelihood ratio** as

$$L = p(\mathcal{D}|H_0)/p(\mathcal{D}|H_1) \tag{172}$$

One can show the following (see [Ric95] for a proof).

Theorem 5.4 (Neyman-Pearson Lemma). *Suppose that the likelihood ratio test that rejects H_0 when $p(\mathcal{D}|H_0)/p(\mathcal{D}|H_1) < c$ has significance level α. Then any other test which has significance level $\alpha* < \alpha$ has lower power.*

If H_0 or H_A have free parameters, we define the **generalized likelihood ratio** as

$$L^* = \frac{\max_{\theta\in\Theta_0} p(\mathcal{D}|\theta)}{\max_{\theta\in\Theta_1} p(\mathcal{D}|\theta)} \tag{173}$$

For technical reasons, it is more common to allow the denominator to range over all parameters, $\Theta = \Theta_0 \cup \Theta_1$:

$$L^* = \frac{\max_{\theta\in\Theta_0} p(x|\theta)}{\max_{\theta\in\Theta} p(x|\theta)} \tag{174}$$

Another definition which is widely used is

$$\lambda = -2 \log \frac{\max_{\theta\in\Theta} p(\mathcal{D}|\theta_0)}{\max_{\theta\in\Theta_0} p(\mathcal{D}|\theta)} = 2 \log \left(\frac{\ell(\hat{\theta})}{\ell(\hat{\theta}_0)} \right) \tag{175}$$

where $\hat{\theta}$ is the MLE and $\hat{\theta}_0$ is the MLE when θ is restricted to lie in Θ_0.

Let us consider an example. Let $X_n \sim \mathcal{N}(\mu, \sigma^2)$. We wish to test $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$, where μ_0 and σ^2 are fixed. Thus $\Theta_0 = \{\mu_0\}$ and $\Theta_1 = \{\mu : \mu \neq \mu_0\}$. The likelihood under H_0 is just

$$p(\mathcal{D}|H_0) = \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\left(-\frac{1}{2\sigma^2}\sum_n (X_n - \mu_0)^2\right) \tag{176}$$

For H_1, we plug in the MLE $\hat{\mu} = \overline{X}$:

$$\max_{\mu\in\Theta} \mathcal{D}(x|H_1, \mu) = \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\left(-\frac{1}{2\sigma^2}\sum_n (X_n - \overline{X})^2\right) \tag{177}$$

So

$$\lambda = \frac{1}{\sigma^2}\left(\sum_n (X_n - \mu)^2 - \sum_n (X_n - \overline{X})^2\right) \tag{178}$$

Using the identity

$$\sum_n (X_n - \mu_0)^2 = \sum_n (X_n - \overline{X})^2 + N(\overline{X} - \mu_0)^2 \tag{179}$$

we have

$$\lambda = \frac{n(\overline{X} - \mu_0)^2}{\sigma^2} \tag{180}$$

This has the null distribution $\lambda \sim \chi^2_1$, since $\overline{X}|H_0 \sim \mathcal{N}(\mu_0, \sigma^2/N)$, so $s = \sqrt{N}(\overline{X} - \mu_0)/\sigma \sim \mathcal{N}(0, 1)$, and $\lambda = s^2 \sim \chi^2_1$ by definition of the chi-squared distribution. Hence the test rejects at level α when

$$\lambda = \frac{n(\overline{X} - \mu_0)^2}{\sigma^2} > \chi^2_1(\alpha) \tag{181}$$

Now $\sqrt{\lambda} \sim \mathcal{N}(0, 1)$, so we can reject for

$$|\overline{X} - \mu_0| \geq \frac{\sigma}{\sqrt{N}} z(\alpha/2) \tag{182}$$

Thus we see the test rejects when the empirical mean is far enough from μ_0, where the definition of "far enough" increases with σ but decreases with $\sqrt{N}$.

Often we won't know the sampling distribution of λ under the null hypothesis, but one can show the following asymptotic result.

Theorem 5.5. *Under suitable smoothness assumptions, $\lambda|H_0 \rightsquigarrow \chi^2_{d-d_0}$, where $d = |\Theta|$ are the number of free paramaters in the unrestricted space, and $d_0 = |\Theta_0|$ are the number of free parameters in the restricted space.*

In the example above, $d = |\{\mu\}| = 1$ and $d_0 = 0$, since μ_0 was fixed.

5.4.2 χ^2 tests for multinomial data

Suppose we have some count data and we want to test whether the parameters θ are equal to a particular parameter vector, say θ_0, or whether they are unrestricted (apart from the positivity and sum-to-one constraints). We can construct a generalized LRT as follows.

$$p(x|H_0) \propto \prod_k \theta_{0,k}^{x_k} \tag{183}$$

$$\arg\max_\theta p(x|H_1, \theta) \propto \prod_k \hat{\theta}_k^{x_k} \tag{184}$$

where $\hat{\theta}_k = x_k/N$ is the MLE. Then

$$\lambda = 2\log\frac{p(x|\hat{\theta})}{p(x|\theta_0)} = 2\log\prod_k\left(\frac{\hat{\theta}_k}{\theta_{0,k}}\right)^{x_k} \tag{185}$$

$$= 2\sum_k x_k \log\left(\frac{x_k}{N\theta_{0,k}}\right) = 2\sum_k O_k \log\frac{O_k}{E_k} \tag{186}$$

where $O_k = x_k$ are the observed counts and $E_k = N\theta_{0,k}$ are the expected counts (under H_0).

We will now perform a Taylor series expansion on this, to derive **Pearson's chi-squared statistic**. First rewrite $O_k = N\hat{\theta}_k$ and $E_k = N\theta_{0,k}$:

$$\lambda = 2N\sum_k \hat{\theta}_k \log\left(\frac{\hat{\theta}_k}{\theta_{0,k}}\right) \tag{187}$$

Now use the following Taylor approximation

$$f(x) = x\log\left(\frac{x}{x_0}\right) \tag{188}$$

$$\approx (x-x_0) + \tfrac{1}{2}(x-x_0)^2\frac{1}{x_0} + \cdots \tag{189}$$

Hence, using the fact that $\sum_k \hat{\theta}_k = 1$, we have

$$\lambda = 2N\sum_k[\hat{\theta}_k - \theta_{0,k}] + N\sum_k\frac{[\hat{\theta}_k - \theta_{0,k}]^2}{\theta_{0,k}} \tag{190}$$

$$= 2N\sum_k\hat{\theta}_k - 2N\sum_k\theta_{0,k} + N\sum_k\frac{[\hat{\theta}_k - \theta_{0,k}]^2}{\theta_{0,k}} \tag{191}$$

$$= N\sum_k\frac{[\hat{\theta}_k - \theta_{0,k}]^2}{\theta_{0,k}} \tag{192}$$

$$= N\sum_k\frac{[\frac{x_k - N\theta_{0,k}}{N}]^2}{\theta_{0,k}} \tag{193}$$

$$= \sum_k\frac{[x_k - N\theta_{0,k}]^2}{N\theta_{0,k}} \tag{194}$$

$$= \sum_k\frac{(O_k - E_k)^2}{E_k} \tag{195}$$

which is called **Pearson's chi-squared statistic**.

Let us examine a classic example. In one of his famous experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. According to the theory, the expected probabilites of the 4 possible types of progeny are: smooth yellow $9/16$, smooth green $3/16$, wrinled yellow $3/16$, wrinkled green $1/16$. So the expected counts are $(312.75, 104.25, 104.25, 34.75)$. The observed counts were $(315, 108, 102, 31)$. The likelihood ratio test statistic is

$$\lambda = 2\sum_{k=1}^4 O_k\log(O_k/E_k) = 0.618 \tag{196}$$

Pearson's chi-squared approximation is

$$X^2 = 2\sum_{k=1}^4(O_k - E_k)^2/E_k = 0.604 \tag{197}$$

which is pretty close. Using the result $\lambda \sim \chi^2_3$ (since H_0 has no free parameters and H_1 has 3 free parameters), we have that the p-value is $P(\chi^2_3 > \lambda) = 0.8922$, so there is no evidence against the null hypothesis. (Even if the model were correct, we would expect values this large about 90% of the time.) More formally, we can say that since `chi2inv(0.95,3) = 7.8147`, we do not reject the null hypothesis at level $\alpha = 0.05$. These numbers can be reproduced using `mendelsPeas`.

5.4.3 Permutation tests

The **permutation test** is a nonparametric method for testing whether two distributions are the same. Suppose $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ are two independent samples. We wish to test $H_0 : F_X = F_Y$ versus $H_A : F_X \neq F_Y$. Let $T(X_{1:m}, Y_{1:n})$ be some statistic. Let $N = m + n$ and consider all $N!$ permutations of the data $X_{1:n}, Y_{1:m}$. For each such permutation compute T. Under the null hypothesis, T is uniform over the $N!$ values T_j. Let t_{obs} be the observed value of T. Assuming we reject when T is large, the p-value is

$$\text{p-value} = P(T > t_{obs}|H_0) = \frac{1}{N!}\sum_{j=1}^{N!} I(T_j > t_{obs}) \tag{198}$$

Obviously this can be approximated by sampling.

5.4.4 Multiple testing

Suppose we conduct m tests at level α. Let R_i be the event that the i'th null hypothesis is falsely rejected. We have $P(R_i) = \alpha$. However, let R be the event that any null hypothesis is rejected. Then using the **union bound**

$$P(R) = P(\cup_{i=1}^m R_i) \leq \sum_{i=1}^m P(R_i) = \alpha m \tag{199}$$

So the the chance of at least one false rejection is much higher. For example, if we have six tests at confidence level 95%, the chance that we will have one false rejection is $\leq 6 \times 5 = 30\%$. This is called the **multiple testing problem**, and frequently arises when looking at high dimensional data.

The simplest solution is the **Bonferroni correction**, which is to use a significance level of α/m on each of the individual tests. But this is very conservative. An alternative appraoch, which bounds the **false discovery rate** (expected number of false rejections as a fraction of the total number of rejections), is due to Benjamini and Hochberg, and is often preferable. See [Was04, p167].

5.4.5 Goodness-of-fit tests

Often we want to test if data is likely to have come from a given model $p(x|\theta)$. Suppose the data are 1D scalars. Divide the real line into K disjoint intervals (bins) with the probability in each bin being

$$p_j(\theta) = \int_{I_j} p(x|\theta)dx \tag{200}$$

Let N_j be the observed number of data points in bin j. Now we can use a chi-squared test to see if the model is a good fit.

There are more powerful tests for special cases, such as testing for normality. Also, graphical methods, such as **quantile quantile (q-q) plots** and **probability plots**), are very useful. (In Matlab, type `qqplot`.) See [Ric95, ch9] for details.

References

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[LM86] R. Larsen and M. Marx. *An introduction to mathematical statistics and its applications*. Prentice Hall, 1986.

[Mac03] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Ric95] J. Rice. *Mathematical statistics and data analysis*. Duxbury, 1995. 2nd edition.

[SAM04] David J. Spiegelhalter, Keith R. Abrams, and Jonathan P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, 2004.

[Was04] L. Wasserman. *All of statistics. A concise course in statistical inference*. Springer, 2004.