Section 1: Preliminaries

1.1 Exchangeability and de Finetti's theorem

A standard situation in statistics is to be presented with a sequence of observations, and use them to make predictions about future observations. In order to do so, we need to make certain assumptions about the nature of the statistical relationships between the sequence of observations.

A common assumption is that our data are **exchangeable**, meaning that their joint probability is invariant under permutations. More concretely, we say a sequence of N observations is finitely exchangeable if $\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_N \in A_n) = \mathbf{P}(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \ldots, X_{\sigma(N)} \in A_n)$ for any permutation of the integers 1 through N, and that an infinite sequence is infinitely exchangeable if this invariance holds for all values of N.

Exercise 1.1 *Clearly, all iid sequences are exchangeable, but not all exchangeable sequences are iid. Consider an urn, containing r red balls and b blue balls. A sequence of colors is generated by repeatedly sampling a ball from the urn, noting its color, and then returning the ball, plus another ball of the same color, to the urn. Show that the resulting sequence is exchangeable, but not iid.*

Solution:
If we consider B as a blue ball and R as a red ball, and a total of N balls, the probability of S as following:

$$P(S) = \frac{R!B!}{(R+B+1)!}$$

Therefore, the order of the ball selection is not influential to the probability (i.e., exchangeable)but the observable probability of a red or a blue ball is not always same (i.e., iid), we can explain the exchangeability. Analytically,

1.1.1 De Finetti's Theorem

Loosely speaking, de Finetti's Theorem states if a sequence of random variables is infinitely exchangeable, those random variables must be conditionally i.i.d. given some set of parameters. More formally,

Theorem 1.1 (de Finetti) *Let $(X_1, X_2, \ldots)$ be an infinite sequence of random variables in some space $\mathcal{X}$. This sequence is infinitely exchangeable if and only if there exists a probability distribution $\mathbf{Q}_\theta$, parametrized by some random parameter $\theta \sim \nu$, such that the X_i are conditionally iid given $\mathbf{Q}_\theta$ and such that*

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots) = \int_\Theta \prod_{i=1}^{\infty} \mathbf{Q}_\theta(A_i)\nu(d\theta).$$

This means we can imagine that any exchangeable sequence has been generated as a sequence of i.i.d. random variables with some unknown law. This provides a motivation for Bayesian inference: We have a

hierarchical model, where data are generated according to some distribution parametrized by a random (in the Bayesian context – i.e. unknown/uncertain) variable θ, and our uncertainty about θ is characterized by some distribution ν.

Let's consider the 0/1 form of de Finetti's theorem, for exchangeable sequences of binary variables:

Theorem 1.2 (de Finetti 0/1) *An infinite sequence $(X_1, X_2, \dots)$ of binary random variables is exchangeable if and only if its distribution can be written as*

$$\begin{aligned}\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N) &= \int_0^1 \prod_{i=1}^{N} \left\{\theta^{x_i}(1-\theta)^{1-x_i}\right\} d\nu(\theta) \\ &= \int_0^1 \theta^k (1-\theta)^{N-k} d\nu(\theta)\end{aligned}$$

where $k = \sum_i x_i$.

We will now work through (most of) a proof in the next two exercises.

Exercise 1.2 *We will start off with a finite sequence $(X_1, \dots, X_M)$. For any $N \leq M$, show that*

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s \middle| \sum_{i=1}^{M} X_i = t\right) = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}}$$

We can therefore write

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) = \binom{N}{s} \sum_{t=s}^{M-N+s} \frac{(t)_s (M-t)_{n-s}}{(M)_N} \mathbf{P}\left(\sum_{i=1}^{M} X_i = t\right), \tag{1.1}$$

where $(x)_y = x(x-1)\dots(x-y+1)$.

Let $F_M(\theta)$ be the distribution function of $\frac{1}{M}(X_1, + \dots, + X_M)$ – i.e. a step function between 0 and 1, with steps of size $\mathbf{P}(\sum_i X_i = t)$ at $t = 0, 1, \dots, M$. Then we can rewrite Equation 1.1 as

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) = \binom{N}{s} \int_0^1 \frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} dF_M(\theta)$$

Solution:
$N <= M$, $\mathbf{P}\left(\sum_{i=1}^{N} X_i = s \middle| \sum_{i=1}^{M} X_i = t\right) = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}}$,
Since $X_1, \dots X_M$ is finite sequence, we pick N items without replacement. Therefore we could get the probability

Exercise 1.3 *Show that, as $M \to \infty$, we can write*

$$\mathbf{P}\left(X_1 = x_1, \dots, X_N = x_N\right) \to \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta)$$

The proof is completed using a result (the Helly Theorem), that shows that any sequence $\{F_M(\theta); M = 1, 2, \dot{\}}$ of probability distributions on [0,1] contains a subsequence that converges to $F(\theta)$.

Solution: If $M- > \infty$, the product of

1.2 The exponential family of distributions

De Finetti's theorem can be seen as a motivation for Bayesian inference. If our data are exchangeable, we know that they are iid according to some unknown probability distribution $F_\theta(X)$, which we can think of as a **likelihood function**, and that they can be represented using an mixture of such iid sequences. As we saw from the 0/1 case, the distribution over probabilities is given by the limit of the empirical distribution function. When not working in this limit, we may choose to model this distribution over the parameters of our likelihood function using a **prior** distribution $\pi(\theta)$ – ideally one that both assigns probability mass to where we expect the empirical distribution might concentrate, and for which $\int_\Theta F_\theta(X)\pi(d\theta)$ is tractable.

The exponential family of probability distributions is the class of distributions parametrized by θ whose density can be written as

$$p(x|\theta) = h(x)\exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}$$

where

- $\eta(\theta)$ (sometimes just written as η), is a transformation of θ that is often referred to as the **natural or canonical parameter**.
- $T(X)$ is known as a **sufficient statistic** of X. We see that $p(x|\theta)$ depends only on X through $T(X)$, implying that $T(X)$ contains all the relevant information about X.
- $A(\eta(\theta)$ (or $A(\eta)$) is known as the **cumulant function** or the **log partition function** (remember, a partition function provides a normalizing constant).

Example 1.1 (The Bernoulli distribution) *A Bernoulli random variable X takes the value $X = 1$ with probability π and $X = 0$ with probability $1 - \pi$; it's density can be written:*

$$\begin{aligned} p(x|\pi) =& \pi^x(1-\pi)^{1-x} \\ =& \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \end{aligned}$$

By rewriting in this exponential family form, we see that

- $\eta = \log\left(\frac{\pi}{1-\pi}\right)$
- $T(x) = x$
- $A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$
- $h(x) = 1$

Exercise 1.4 *The Poisson random variable has PDF*

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Re-write the density of the Poisson random variable in exponential family form. What are η, $T(x)$, $A(\eta)$ and $h(x)$? What about if we have n independent samples $x_1, \ldots, x_n$?

Solution:
$p(x|\lambda) = exp(log(\frac{\lambda^x e^{-\lambda}}{x!}))$
$= \frac{1}{x!} exp(log\lambda * x - \lambda)$
Thus, $\eta = log\lambda$, $T(x) = x$, $A(\eta) = exp(loglamda) = exp(\eta)$, $h(x) = \frac{1}{x!}$
If we have n independent samples,
$\eta = log\lambda$, $T(x) = \sum_{i=1}^{N} x_i$, $A(\eta) = n * \lambda$, $h(x) = \prod_{i=1}^{n} \frac{1}{x_i!}$

Exercise 1.5 *The gamma random variable has PDF*

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

What are the natural parameters and sufficient statistics for the gamma distribution, given n observations $x_1, \ldots, x_N$?

Solution:
Let X_i $Gamma(\alpha, \beta)$ and taking exponent
$p(x|\alpha, \beta) = exp(-\beta x + (\alpha - 1)log(x) + \alpha log(\beta) - log(\Gamma(\alpha)))$
let $\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \alpha - 1 \\ \beta \end{pmatrix}$, $h(x) = 1$, $T(x) = \begin{pmatrix} \log(x) \\ x \end{pmatrix}$, $A(\eta) = log\Gamma(\eta_1 + 1) - (\eta_1 + 1)log(-\eta_2)$

1.2.1 Cumulants and moments of exponential families

We are probably most familiar with using the PDF or the CDF of a random variable to describe its distribution, but there are other representations that can be useful. The **moment generating function** $M_X(s) = \mathbb{E}[\exp(s^T x)] = \int_{\mathcal{X}} e^{s^T x} p_X(x) dx$ is the Laplace transform of the PDF $p_X(x)$. As the name suggests, we can use the moment-generating function to generate the (uncentered) moments of a random variable; the nth moment is given by

$$m_n = \frac{d^n M_X}{ds^n}\Big|_{s=0}$$

Exercise 1.6 *For exponential family random variables, we know that the sufficient statistic $T(X)$ contains all the information about X, so (for univariate X) we can write the moment generating function of the sufficient statistic as $\mathbb{E}[e^{sT(x)}|\eta]$. Show that the moment generating function for the sufficient statistic of an arbitrary exponential family random variable with natural parameter η can be written as*

$$M_{T(X)}(s) = \exp A(\eta + s) - A(\eta)$$

Solution:
since X draw from the exp(λ)
$f(x|\lambda) = \lambda e^{-\lambda x}$, $x \geq 0$
$f(x|\lambda) = \exp(\log \lambda e^{-\lambda x}) = \exp(-\lambda x + \log \lambda)$
$\eta = -\lambda$, $T(x) = x$, $A(\eta) = -\log(-\eta)$, $h(x) = 1$
$f(x|\eta) = -\eta e^{\eta x}$
$E(e^{sT(x)}|\eta) = -\int_{-\infty}^{0} e^{sx}\eta e^{\eta x} dx$
$= -\int_{-\infty}^{0} e^{sx}\eta e^{\eta x} dx = \frac{\eta}{\eta+s}\int_{-\infty}^{0} -(\eta+s)e^{(\eta+s)x} dx = \frac{\eta}{\eta+s}$
Thus, $A(\eta+s) - A(\eta) = -\log(-\eta-s) + \log(-\eta) = -\log((-\eta-s)-(-\eta)) = \log\eta - \log(\eta+s)$

A related representation is the **cumulant generating function** $C_X(s) = \log \mathbb{E}[e^{s^T x}] = \log(M_X(s))$. Clearly, for exponential families this takes the form $C_{T(X)}(s) = A(\eta+s) - A(\eta)$. This explains why $A(\eta)$ is sometimes called the cumulant function! The cumulant function can be used to generate the cumulants of a distribution as

$$\kappa_n = \frac{d^n C_X}{ds^n}\Big|_{s=0}$$

The first three cumulants are the same as the first three central moments of the distribution – meaning, the cumulant generative function is a useful tool for calculating mean, variance and the third central moment.

Exercise 1.7 *It is usually easier to calculate mean and variance using the cumulant generating function rather than the moment generating function. Starting from the exponential family representation of the Poisson distribution from Exercise 1.4, calculate the mean and variance of the Poisson using a) the moment generating function, and b) the cumulant generating function.*

1.2.2 Conjugate priors

Exponential families are very important in Bayesian statistics because, for any exponential family likelihood, we can find an conjugate exponential family prior. If our likelihood takes the form

$$f(x|\eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\}$$

then a conjugate prior is given by

$$p(\eta|\xi,\nu) = g(\xi,\nu)\exp\left\{\eta^T \xi - \nu A(\eta)\right\}$$

Below are some exercises based on common conjugate priors.

Exercise 1.8 *Suppose we have N independent observations $x_1,\ldots,x_N \overset{iid}{\sim}$ Normal(μ,σ^2). If σ^2 is known and $\mu \sim$ Normal(μ_0,σ_0^2), derive the posterior for $\mu|x_1,\ldots,x_N$*

Exercise 1.9 *Now, let's assume $x_1,\ldots,x_N \overset{iid}{\sim}$ Normal(μ,σ^2) with known mean μ but unknown variance σ^2. Let's express the likelihood in terms of the precision, $\omega = 1\sigma^2$:*

$$f(x_i|\mu,\omega) = \sqrt{\frac{\omega}{2\pi}}\exp\left\{-\frac{\omega}{2}(x_i-\mu)^2\right\}$$

Let ω have a gamma prior (this is also known as putting an inverse-gamma prior on σ^2):

$$p(\omega) = \frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha-1}e^{-\beta\omega}$$

Derive the posterior distribution for ω

Exercise 1.10 *Let's assume $x \sim Normal(0, \sigma^2)$ and that $\sigma^2 \sim InvGamma(\alpha, \beta)$ (i.e. $1/\sigma^2 \sim Gamma(\alpha, \beta)$). Show that the marginal distribution of x is given by a Student's t distribution.*

1.3 Multivariate normal distribution

So far, we have looked at univariate random variables - particularly, the univariate normal random variable, which is characterized by its mean and variance. We will often work with the multivariate normal distribution, a natural generalization characterized by a mean vector and a covariance matrix.

Exercise 1.11 (covariance matrix) *The covariance matrix Σ of a vector-valued random variable x is the matrix whose entries $\Sigma(i, j) = cov(x_i, x_j)$ are given by the covariance between the ith and jth elements of x, giving*

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right]$$

Show that a) $\Sigma = E[xx^T] - \mu\mu^T$; b) if the covariance of x is σ, then the covariance of $Ax + b$ is $A\Sigma A^T$

Exercise 1.12 (Standard multivariate normal) *The simplest multivariate normal, known as the standard multivariate normal, occurs where the entries of x are independent and have mean 0 and variance 1. a) What is the moment generating function of a univariate normal, with mean m and variance v^2? b) Express the PDF and moment generating function of the standard multivariate normal, in vector notation.*

Exercise 1.13 (Multivariate normal) *A random vector x has multivariate normal distribution if and only if every linear combination of its elements is univariate normal, i.e. if the scalar value $z = a^T x$ is normally distributed for all possible x. Prove that this implies that x is multivariate normal if and only if its moment generating function takes the form $M_X(s) = \exp\left\{a^T\mu + \frac{1}{2}a^T\Sigma a\right\}$, where μ and Σ are the mean and covariance of x. Hint: We know the moment generating function of z in terms of the mean and variance of z, from the previous question...*

Exercise 1.14 (Relationship to standard multivariate normal) *An equivalent statement is that a random vector x has multivariate normal distribution if and only if it can be written in the form*

$$x = Dz + \mu$$

for some matrix D, real-valued vector μ, and vector z distributed according to a standard multivariate normal. Express the moment generating function of x in terms of D, and uncover the relationship between D and Σ. Use this result to suggest a method for generating multivariate normal random variables, if you have a method for generating Normal(0,1) univariate random variables.

Exercise 1.15 *Use the result from the previous question to show that the PDF of a multivarite normal random vector $x \sim Normal(\mu, \Sigma)$ takes the form*

$$p(x) = \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\},$$

by using a change-of-variables from the standard multivariate normal distribution.

1.3.1 Manipulation of multivariate normals

Like its univariate counterpart, the multivariate normal distribution is closed under a number of operations, which we will explore here.

Exercise 1.16 (marginal distribution) *Let us assume that $x \sim Normal(\mu, \Sigma)$, and let us partition x into 2 components x_1 and x_2. Let us similarly partition μ and σ so that*

$$\mu = (\mu_1, \mu_2)^T \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

Derive the marginal distribution of x_1.

Exercise 1.17 (Precision matrix) *Earlier, we chose to express a univariate normal random variable in terms of its precision, to make math easier. We can also express a multivariate normal in terms of a precision matrix $\Omega = \Sigma^{-1}$. Partition Ω as*

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}$$

and express Ω_{11}, Ω_{12} and Ω_{22} in terms of Σ_{11}, Σ_{12} and Σ_{22}. Hint: You'll need the matrix inversion lemma

Exercise 1.18 (Conditional distribution) *The conditional distribution of $x_1|x_2$ is also normal, with mean $\mu_1 + \Sigma_{12}\sigma_{22}^{-1}(x_2 - \mu_2)$ and covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. Prove this for the case where μ is zero (the general case isn't really harder, just more tedious). Hint: ignore any constants that don't involve x_1. You might want to work with the log conditional density.*

1.4 Frequentist estimation and uncertainty quantification

In this section, we're going to go over basic frequentist approaches to inference, with a focus on multiple linear regression (since we're next going to look at Bayesian regression). Some of this should be familiar to you, although we will go into quite some depth. Throughout the remainder of this section, we are going to assume our data follow a linear model, of the form

$$y_i = x_i^T \beta + \epsilon_i, \qquad i = 1, \ldots, N$$

There are a number of options for estimating β. Three commonly used techniques are:

1. **Method of Moments** Select $\hat{\beta}$ so that the empirical moments of the observations match the theoretical moments.
2. **Maximum likelihood** Assume a model for generating the ϵ_i, and find the value of $\hat{\beta}$ that maximizes the likelihood.
3. **Loss function**: Construct a loss function between the y_i and $x_i^T\hat{\beta}$, and minimize that loss function.

Exercise 1.19 (method of moments) *To obtain the theoretical moments, we can assume that $E[y_i|x_i] = x_i^T\beta$, implying that the covariance between the predictors x_i and the residuals is zero. By setting the sample covariance between the x_i and the ϵ_i to zero, derive a method of moments estimator $\hat{\beta}_{MM}$*

Exercise 1.20 (maximum likelihood) *Show that, if we assume $\epsilon_i \sim Normal(0, \sigma^2)$, then the ML estimator $\hat{\beta}_{ML}$ is equivalent to the method of moments estimator.*

Exercise 1.21 (Least squares loss function) *Show that if we assume a quadratic loss function, i.e. $\hat{\beta}_{LS} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^T\beta)^2$, we recover the same estimator again.*

Exercise 1.22 (Ridge regression) *We may wish to add a regularization term to our loss term. For example, ridge regression involves adding an L2 penalty term, so that*

$$\hat{\beta}_{ridge} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{N} (y_i - x_i^T\beta)^2 \text{ s.t. } \sum_{j=1}^{p} \beta_j^2 \leq t$$

for some $t \geq 1$.

Reformulate this constrained optimization using a Lagrange multiplier, and solve to give an expression for $\hat{\beta}_{ridge}$. Comparing this with the least squares estimator, comment on why this estimator might be prefered in practice.

1.4.1 Uncertainty quantification

In a frequentist context, we typically quantify our uncertainty by looking at the sampling distribution of our estimator. Let's assume that our errors are normally distributed, i.e. (in vector notation)

$$y = X\beta + \epsilon, \qquad \epsilon \sim \text{Normal}(0, \sigma^2 I)$$

.

Exercise 1.23 *What is the sampling distribution for $\hat{\beta}_{LS}$ $(= \hat{\beta}_{MM} = \hat{\beta}_{ML})$?*

Exercise 1.24 *How about the sampling distribution for $\hat{\beta}_{ridge}$?*

Exercise 1.25 *The two exercises above assumed the residual variance σ^2 is known. This is unlikely to be the case. Propose a strategy for estimating the standard error of $\hat{\beta}_{LS}$ from data, when σ^2 is unknown. Implement it in R, and test it on the dataset `Prestige` in the R package `cars` (there's a starter script, `prestige.R` on Github). Do you get the same standard errors as the built-in function `lm`?*

1.4.2 Propogation of uncertainty

Let's now consider the general case where we have a point estimate $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_P)^T$ to some set of parameters $\theta = (\theta_1, \ldots, \theta_P)^T$, and we have an estimate $\hat{\Sigma}$ to the covariance matrix of the sampling distribution of $\hat{\theta}$. If we want to describe our uncertainty about the individual θ_i (as was the case for calculating standard errors in the regression problems above), we can look at the diagonal terms in the covariance matrix, $\hat{\Sigma}_{ii} = \hat{\sigma}_i^2$. If we care, more generally, about a *function* of the θ_i, however, the cross terms will become important.

Exercise 1.26 *Let's assume we care about $f(\theta) = \sum_i \theta_i$. What is the standard error of $f(\theta)$?*

Exercise 1.27 *How about the standard error of some arbitrary non-linear function $f(\theta)$? Hint: Try a Taylor expansion*