Lecture Notes 1

Brief Review of Basic Probability

(Casella and Berger Chapters 1-4)

1 Probability Review

Chapters 1-4 are a review. **I will assume you have read and understood Chapters 1-4.** Let us recall some of the key ideas.

1.1 Random Variables

A *random variable* is a map X from a set Ω (equipped with a probability P) to $\mathbb{R}$. We write

$$P(X \in A) = P(\{\omega \in \Omega : \ X(\omega) \in A\})$$

and we write $X \sim P$ to mean that X has distribution P. The *cumulative distribution function (cdf)* of X is

$$F_X(x) = F(x) = P(X \le x).$$

If X is discrete, its *probability mass function (pmf)* is

$$p_X(x) = p(x) = P(X = x).$$

If X is continuous, then its *probability density function function (pdf)* satisfies

$$P(X \in A) = \int_A p_X(x)dx = \int_A p(x)dx$$

and $p_X(x) = p(x) = F'(x)$. The following are all equivalent:

$$X \sim P, \quad X \sim F, \quad X \sim p.$$

Suppose that $X \sim P$ and $Y \sim Q$. We say that X and Y have the same distribution if $P(X \in A) = Q(Y \in A)$ for all A. In that case we say that X and Y are *equal in distribution* and we write $X \overset{d}{=} Y$.

It can be shown that $X \stackrel{d}{=} Y$ if and only if $F_X(t) = F_Y(t)$ for all t.

1.2 Expected Values

The *mean* or expected value of $g(X)$ is

$$\mathbb{E}\left(g(X)\right) = \int g(x)dF(x) = \int g(x)dP(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases}$$

Recall that:

1. $\mathbb{E}(\sum_{j=1}^k c_j g_j(X)) = \sum_{j=1}^k c_j \mathbb{E}(g_j(X))$.

2. If $X_1, \ldots, X_n$ are independent then
$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_i \mathbb{E}\left(X_i\right).$$

3. We often write $\mu = \mathbb{E}(X)$.

4. $\sigma^2 = \mathsf{Var}\left(X\right) = \mathbb{E}\left((X-\mu)^2\right)$ is the **Variance**.

5. $\mathsf{Var}\left(X\right) = \mathbb{E}\left(X^2\right) - \mu^2$.

6. If $X_1, \ldots, X_n$ are independent then
$$\mathsf{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_i a_i^2 \mathsf{Var}\left(X_i\right).$$

7. The covariance is
$$\mathsf{Cov}(X,Y) = \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(XY) - \mu_X \mu_Y$$
and the correlation is $\rho(X,Y) = \mathsf{Cov}(X,Y)/\sigma_x\sigma_y$. Recall that $-1 \leq \rho(X,Y) \leq 1$.

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X)$ whose value, when $X = x$ is

$$\mathbb{E}(Y|X = x) = \int y \; p(y|x) dy$$

where $p(y|x) = p(x, y)/p(x)$.

The *Law of Total Expectation* or *Law of Iterated Expectation*:

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \int \mathbb{E}(Y|X = x) p_X(x) dx.$$

The *Law of Total Variance* is

$$\mathsf{Var}(Y) = \mathsf{Var}\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[\mathsf{Var}(Y|X)\big].$$

The *moment generating function (mgf)* is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

If $M_X(t) = M_Y(t)$ for all t in an interval around 0 then $X \stackrel{d}{=} Y$.

Check that $M_X^{(n)}(t)|_{t=0} = \mathbb{E}\left(X^n\right).$

1.3 Exponential Families

A family of distributions $\{p(x; \theta) : \ \theta \in \Theta\}$ is called an *exponential family* if

$$p(x; \theta) = h(x) c(\theta) \exp\left\{\sum_{i=1}^{k} w_i(\theta) t_i(x)\right\}.$$

Example 1 $X \sim \text{Poisson}(\lambda)$ *is exponential family since*

$$p(x) = P(X = x) \quad = \quad \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!} e^{-\lambda} \exp\{\log \lambda \cdot x\}.$$

Example 2 $X \sim U(0, \theta)$ *is not an exponential family. The density is*

$$p_X(x) = \frac{1}{\theta} I_{(0,\theta)}(x)$$

where $I_A(x) = 1$ *if* $x \in A$ *and 0 otherwise.*

We can rewrite an exponential family in terms of a *natural parameterization*. For $k = 1$ we have

$$p(x;\eta) = h(x)\exp\{\eta t(x) - A(\eta)\}$$

where

$$A(\eta) = \log \int h(x)\exp\{\eta t(x)\}dx.$$

For example a Poisson can be written as

$$p(x;\eta) = \exp\{\eta x - e^{\eta}\}/x!$$

where the natural parameter is $\eta = \log\lambda$.

Let X have an exponential family distribution. Then

$$\mathbb{E}\left(t(X)\right) = A'(\eta), \quad \mathsf{Var}\left(t(X)\right) = A''(\eta).$$

Practice Problem: Prove the above result.

1.4 Transformations

Let $Y = g(X)$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) = \int_{A(y)} p_X(x)dx$$

where

$$A_y = \{x: \ g(x) \le y\}.$$

Then $p_Y(y) = F_Y'(y)$.

If g is monotonic, then

$$p_Y(y) = p_X(h(y))\left|\frac{dh(y)}{dy}\right|$$

where $h = g^{-1}$.

Example 3 *Let* $p_X(x) = e^{-x}$ *for* $x > 0$. *Hence* $F_X(x) = 1 - e^{-x}$. *Let* $Y = g(X) = \log X$. *Then*

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(\log(X) \leq y) \\ &= P(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y} \end{aligned}$$

and $p_Y(y) = e^y e^{-e^y}$ *for* $y \in \mathbb{R}$.

Example 4 Practice problem. *Let* X *be uniform on* $(-1, 2)$ *and let* $Y = X^2$. *Find the density of* Y.

Let $Z = g(X, Y)$. For exampe, $Z = X + Y$ or $Z = X/Y$. Then we find the pdf of Z as follows:

1. For each z, find the set $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find the CDF
$$F_Z(z) = P(Z \leq z) = P(g(X, Y) \leq z) = P(\{(x, y) : g(x, y) \leq z\}) = \int\int_{A_z} p_{X,Y}(x, y)dxdy.$$
3. The pdf is $p_Z(z) = F'_Z(z)$.

Example 5 Practice problem. *Let* (X, Y) *be uniform on the unit square. Let* $Z = X/Y$. *Find the density of* Z.

1.5 Independence

X and Y are *independent* if and only if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all A and B.

Theorem 6 *Let* (X, Y) *be a bivariate random vector with* $p_{X,Y}(x, y)$. X *and* Y *are independent iff* $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

$X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Thus, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$.

If $X_1, \ldots, X_n$ are independent and identically distributed we say they are iid (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P \quad \text{or} \quad X_1, \ldots, X_n \sim F \quad \text{or} \quad X_1, \ldots, X_n \sim p.$$

1.6 Important Distributions

$X \sim N(\mu, \sigma^2)$ if

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

If $X \in \mathbb{R}^d$ then $X \sim N(\mu, \Sigma)$ if

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

$X \sim \chi^2_p$ if $X = \sum_{j=1}^{p} Z_j^2$ where $Z_1, \ldots, Z_p \sim N(0, 1)$.

$X \sim \text{Bernoulli}(\theta)$ if $\mathbb{P}(X = 1) = \theta$ and $\mathbb{P}(X = 0) = 1 - \theta$ and hence

$$p(x) = \theta^x (1-\theta)^{1-x} \qquad x = 0, 1.$$

$X \sim \text{Binomial}(\theta)$ if

$$p(x) = \mathbb{P}(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad x \in \{0, \ldots, n\}.$$

$X \sim \text{Uniform}(0, \theta)$ if $p(x) = I(0 \le x \le \theta)/\theta$.

1.7 Sample Mean and Variance

The sample mean is

$$\overline{X} = \frac{1}{n}\sum_i X_i,$$

and the sample variance is

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2.$$

Let $X_1, \ldots, X_n$ be iid with $\mu = \mathbb{E}(X_i) = \mu$ and $\sigma^2 = \mathsf{Var}(X_i) = \sigma^2$. Then

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathsf{Var}(\overline{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2) = \sigma^2.$$

Theorem 7 *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then*

(a) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

(b) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

(c) $\overline{X}$ *and* S^2 *are independent*

1.8 Delta Method

If $X \sim N(\mu, \sigma^2)$, $Y = g(X)$ and σ^2 is small then

$$Y \approx N(g(\mu), \sigma^2(g'(\mu))^2).$$

To see this, note that

$$Y = g(X) = g(\mu) + (X - \mu)g'(\mu) + \frac{(X-\mu)^2}{2}g''(\xi)$$

for some ξ. Now $\mathbb{E}((X-\mu)^2) = \sigma^2$ which we are assuming is small and so

$$Y = g(X) \approx g(\mu) + (X - \mu)g'(\mu).$$

Thus

$$\mathbb{E}(Y) \approx g(\mu), \quad \mathsf{Var}(Y) \approx (g'(\mu))^2\sigma^2.$$

Hence,

$$g(X) \approx N\left(g(\mu),\ (g'(\mu))^2\sigma^2\right).$$

Appendix: Useful Facts

Facts about sums

- $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.
- $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.
- Geometric series: $a + ar + ar^2 + \ldots = \frac{a}{1-r}$, for $0 < r < 1$.
- Partial Geometric series $a + ar + ar^2 + \ldots + ar^{n-1} = \frac{a(1-r^n)}{1-r}$.
- Binomial Theorem

$$\sum_{x=0}^{n} \binom{n}{x} a^x = (1+a)^n, \quad \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n.$$

- Hypergeometric identity

$$\sum_{x=0}^{\infty} \binom{a}{x} \binom{b}{n-x} = \binom{a+b}{n}.$$

Common Distributions

Discrete

Uniform

- $X \sim U(1, \ldots, N)$
- X takes values $x = 1, 2, \ldots, N$
- $P(X = x) = 1/N$
- $\mathbb{E}(X) = \sum_x x P(X = x) = \sum_x x \frac{1}{N} = \frac{1}{N} \frac{N(N+1)}{2} = \frac{(N+1)}{2}$
- $\mathbb{E}(X^2) = \sum_x x^2 P(X = x) = \sum_x x^2 \frac{1}{N} = \frac{1}{N} \frac{N(N+1)(2N+1)}{6}$

Binomial

- $X \sim \text{Bin}(n,p)$
- X takes values $x = 0, 1, \ldots, n$
- $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Hypergeometric

- $X \sim \text{Hypergeometric}(N, M, K)$
- $P(X = x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$

Geometric

- $X \sim \text{Geom}(p)$
- $P(X = x) = (1-p)^{x-1}p$, $x = 1, 2, \ldots$
- $\mathbb{E}(X) = \sum_x x(1-p)^{x-1} = p \sum_x \frac{d}{dp}\left(-(1-p)^x\right) = p\frac{p}{p^2} = \frac{1}{p}$.

Poisson

- $X \sim \text{Poisson}(\lambda)$
- $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$
- $\mathbb{E}(X) = \mathsf{Var}(X) = \lambda$
- $M_X(t) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$.
- $\mathbb{E}(X) = \lambda e^t e^{\lambda(e^t-1)}|_{t=0} = \lambda$.
- Use mgf to show: if $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, independent then $Y = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Continuous Distributions

Normal

- $X \sim N(\mu, \sigma^2)$
- $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{-1}{2\sigma^2}(x-\mu)^2\}$, $x \in \mathcal{R}$
- mgf $M_X(t) = \exp\{\mu t + \sigma^2 t^2/2\}$.
- $E(X) = \mu$
- $\mathsf{Var}(X) = \sigma^2$.
- e.g., If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$. Show this...

Proof.

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) = E\left(e^{t(\mu+\sigma Z)}\right) = e^{t\mu}E\left(e^{t\sigma Z}\right) \\ &= e^{t\mu}M_Z(t\sigma) = e^{t\mu}e^{(t\sigma)^2/2} = e^{t\mu + t^2\sigma^2/2} \end{aligned}$$

which is the mgf of a $N(\mu, \sigma^2)$.

Alternative proof:

$$\begin{aligned} F_X(x) &= P(X \le x) = P(\mu + \sigma Z \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \\ p_X(x) &= F'_X(x) = p_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \end{aligned}$$

which is the pdf of a $N(\mu, \sigma^2)$. □

Gamma

- $X \sim \Gamma(\alpha, \beta)$.
- $p_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, x a positive real.
- $\Gamma(\alpha) = \int_0^\infty \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx$.
- Important statistical distribution: $\chi_p^2 = \Gamma(\frac{p}{2}, 2)$.
- $\chi_p^2 = \sum_{i=1}^p X_i^2$, where $X_i \sim N(0, 1)$, iid.

Exponential

- $X \sim \exp(\beta)$
- $p_X(x) = \frac{1}{\beta} e^{-x/\beta}$, x a positive real.
- $\exp(\beta) = \Gamma(1, \beta)$.
- e.g., Used to model waiting time of a Poisson Process. Suppose N is the number of phone calls in 1 hour and $N \sim Poisson(\lambda)$. Let T be the time between consecutive phone calls, then $T \sim \exp(1/\lambda)$ and $E(T) = (1/\lambda)$.
- If $X_1, \ldots, X_n$ are iid $\exp(\beta)$, then $\sum_i X_i \sim \Gamma(n, \beta)$.
- Memoryless Property: If $X \sim \exp(\beta)$, then

$$P(X > t + s | X > t) = P(X > s).$$

Linear Regression

Model the response (Y) as a linear function of the parameters and covariates (x) plus random error (ϵ).

$$Y_i = \theta(x, \beta) + \epsilon_i$$

where

$$\theta(x, \beta) = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

Generalized Linear Model

Model the natural parameters as linear functions of the the covariates.

Example: Logistic Regression.

$$P(Y = 1|X = x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}.$$

In other words, $Y|X = x \sim Bin(n, p(x))$ and

$$\eta(x) = \beta^T x$$

where

$$\eta(x) = \log\left(\frac{p(x)}{1 - p(x)}\right).$$

Logistic Regression consists of modelling the natural parameter, which is called the log odds ratio, as a linear function of covariates.

Location and Scale Families, CB 3.5

Let $p(x)$ be a pdf.

$$\text{Location family} : \ \{p(x|\mu) = p(x - \mu) : \ \mu \in \mathbb{R}\}$$

$$\text{Scale family} : \ \left\{p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) : \ \sigma > 0\right\}$$

$$\text{Location} - \text{Scale family} : \ \left\{p(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) : \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

(1) Location family. Shifts the pdf.

e.g., Uniform with $p(x) = 1$ on $(0, 1)$ and $p(x - \theta) = 1$ on $(\theta, \theta + 1)$.

e.g., Normal with standard pdf the density of a $N(0,1)$ and location family pdf $N(\theta,1)$.

(2) Scale family. Stretches the pdf.

e.g., Normal with standard pdf the density of a $N(0,1)$ and scale family pdf $N(0,\sigma^2)$.

(3) Location-Scale family. Stretches and shifts the pdf.

e.g., Normal with standard pdf the density of a $N(0,1)$ and location-scale family pdf $N(\theta,\sigma^2)$, i.e., $\frac{1}{\sigma}p(\frac{x-\mu}{\sigma})$.

Multinomial Distribution

The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn with has balls with k different colors labeled "color 1, color 2, ..., color k." Let $p = (p_1, p_2, \ldots, p_k)$ where $\sum_j p_j = 1$ and p_j is the probability of drawing color j. Draw n balls from the urn (independently and with replacement) and let $X = (X_1, X_2, \ldots, X_k)$ be the count of the number of balls of each color drawn. We say that X has a Multinomial (n,p) distribution. The pdf is

$$p(x) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}.$$

Multivariate Normal Distribution

Let $Y \in \mathbb{R}^d$. Then $Y \sim N(\mu, \Sigma)$ if

$$p(y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right).$$

Then $E(Y) = \mu$ and $\mathsf{cov}(Y) = \Sigma$. The moment generating function is

$$M(t) = \exp\left(\mu^T t + \frac{t^T\Sigma t}{2}\right).$$

Theorem 8 *(a). If $Y \sim N(\mu,\Sigma)$, then $E(Y) = \mu$, $cov(Y) = \Sigma$.*

(b). If $Y \sim N(\mu, \Sigma)$ and c is a scalar, then $cY \sim N(c\mu, c^2\Sigma)$.

(c). Let $Y \sim N(\mu, \Sigma)$. If A is $p \times n$ and b is $p \times 1$, then $AY + b \sim N(A\mu + b, A\Sigma A^T)$.

Theorem 9 *Suppose that $Y \sim N(\mu, \Sigma)$. Let*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

where Y_1 and μ_1 are $p \times 1$, and Σ_{11} is $p \times p$.

(a). $Y_1 \sim N_p(\mu_1, \Sigma_{11}), Y_2 \sim N_{n-p}(\mu_2, \Sigma_{22})$.

(b). Y_1 and Y_2 are independent if and only if $\Sigma_{12} = 0$.

(c). If $\Sigma_{22} > 0$, then the condition distribution of Y_1 given Y_2 is

$$Y_1|Y_2 \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Lemma 10 *Let $Y \sim N(\mu, \sigma^2\mathbf{I})$, where $Y^T = (Y_1, \ldots, Y_n), \mu^T = (\mu_1, \ldots, \mu_n)$ and $\sigma^2 > 0$ is a scalar. Then the Y_i are independent, $Y_i \sim N_1(\mu, \sigma^2)$ and*

$$\frac{||Y||^2}{\sigma^2} = \frac{Y^T Y}{\sigma^2} \sim \chi_n^2 \left(\frac{\mu^T \mu}{\sigma^2} \right).$$

Theorem 11 *Let $Y \sim N(\mu, \Sigma)$. Then:*

(a). $Y^T \Sigma^{-1} Y \sim \chi_n^2(\mu^T \Sigma^{-1} \mu)$.

(b). $(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi_n^2(0)$.