Probability theory refresher

Kevin P. Murphy

Last updated August 30, 2007

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

1 Introduction

We are all familiar with the phrase "the probability that a coin will land heads is 0.5". But what does this mean? There are actually two different interpretations of probability. One is called the **frequentist interpretation of probability**. In this view, probabilities represent long run frequencies of events. For example, the above statement means that, if we flip the coin many times, we expect it to land heads about half the time. The other interpretation is called the **Bayesian interpretation of probability**. In this view, probabilities represent measures of **uncertainty** or **degrees of belief** [Jay03]. In the Bayesian view, the above statement means we think the coin is equally likely to land heads or tails on the next toss.

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might want to compute the probability that the polar ice cap will melt by 2020AD. This event will happen zero or one times, but cannot happen repeatedly. Nevertheless, we ought to be able to quantify our uncertainty about this event; based on how probable we think this event is, we will make appropriate decisions/ take appropriate actions. We might also be interested in computing the probability of **counterfactual events**, such as the probability that the ice cap would have melted in 2000AD if the Kyoto protocol had not been ratified.

Despite the two different philosophical interpretations, the mathematics of probability theory remains the same. We assume the reader is already familiar with the basic notions of probability and random variables, and simple descriptive statistics, such as the mean and variance. (For an excellent introduction, see e.g., [Was04, Ric95]). However, below we give a quick refresher. We then introduce a variety of probability distributions that will be used later.

2 Basics

2.1 Sample space and events

To formally define probability, we start with the notion of a **sample space** Ω, which is a set of possible outcomes. Subsets of Ω are called **events**. For example, if we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$. Another example is measuring the temperature. Here the event space is $\mathbb{R} = (-\infty, \infty)$. (Arguably the lower bound should be finite, but there is usually no harm in making the sample space larger than needed.) The event that the observed temperature is larger than 10 but less than or equal to 23 is $A = (10, 23]$.

A function $P()$ that assigns a real number to each event A is **probability distribution** if it satisfies the following 3 axioms:

1. $P(A) \geq 0$ for every A
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots$ are disjoint, then

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{1}$$

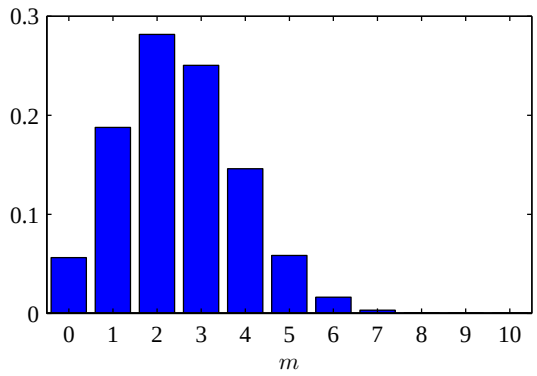

Figure 1: Illustration of the binomial distribution $Bi(n = 10, \theta = 0.25)$. Source: [Bis06] Figure 2.1.

We assume the reader is familiar with the usual rules for manipulating probabilities (e.g., $p(A^c) = 1 - p(A)$, where $A^c = \Omega \setminus A$ is the complement of A).

2.2 Random variables

A **random variable** (rv) is a mapping

$$X : \Omega \rightarrow \mathbb{R} \tag{2}$$

that assigns a real number $X(\omega)$ to every outcome ω. For example, let $X(\omega)$ be the number of heads in sequence ω. Then if $\omega = HHTHHTHHTT$, then $X(\omega) = 6$. Often we talk about random variables directly, and ignore the underlying sample space. If $X(\omega)$ can take on a countable set of values, then X is called a **discrete random variable**, otherwise it is called a **continuous random variable**. We shall initially focusing on discrete rv's for simplicity.

Consider a binary random variable (rv) X with two possible states, $X \in \{0, 1\}$. For example, $X = 1$ could represent the event that we tossed heads, and $Y = 0$ could represent the event that we tossed tails. Or $X = 1$ could represent the fact that someone is female, and $X = 0$ represents male. We say that X is **binary** and has a **Bernoulli** distribution:

$$p(X|\theta) = \text{Be}(X|\theta) = \theta^X (1 - \theta)^{1-X} \tag{3}$$

where $p(X = 1) = \theta$. The generalization of this to multiple coin tosses, where $X \in \{0, \ldots, n\}$ is the number of heads in n trials, yields the **Binomial** distribution (Figure 1):

$$p(x|n) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} = Bi(\theta, n) \tag{4}$$

where

$$\binom{n}{x} = \frac{n!}{(n - x)!x!} \tag{5}$$

is the number of ways to choose x items from n.

This can be generalized to K-ary random variables, e.g., imagine spinning a dice N times, and let X_j be the number of times face j showed up, for $j = 1, \ldots, K = 6$. We say that X has a **multinomial** distribution:

$$p(\mathbf{x}|n, \theta) = \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \tag{6}$$

where θ_j is the probability of face j showing up. (For a fair dice, $\theta_j = 1/6$.) We have $\sum_{j=1}^{K} x_j = n$. If the sample size is $n = 1$, this simplifies to

$$p(\mathbf{x}|\theta) = \text{Mu}(\mathbf{x}|1, \theta) = \prod_{j=1}^{K} \theta_j^{x_j} \tag{7}$$

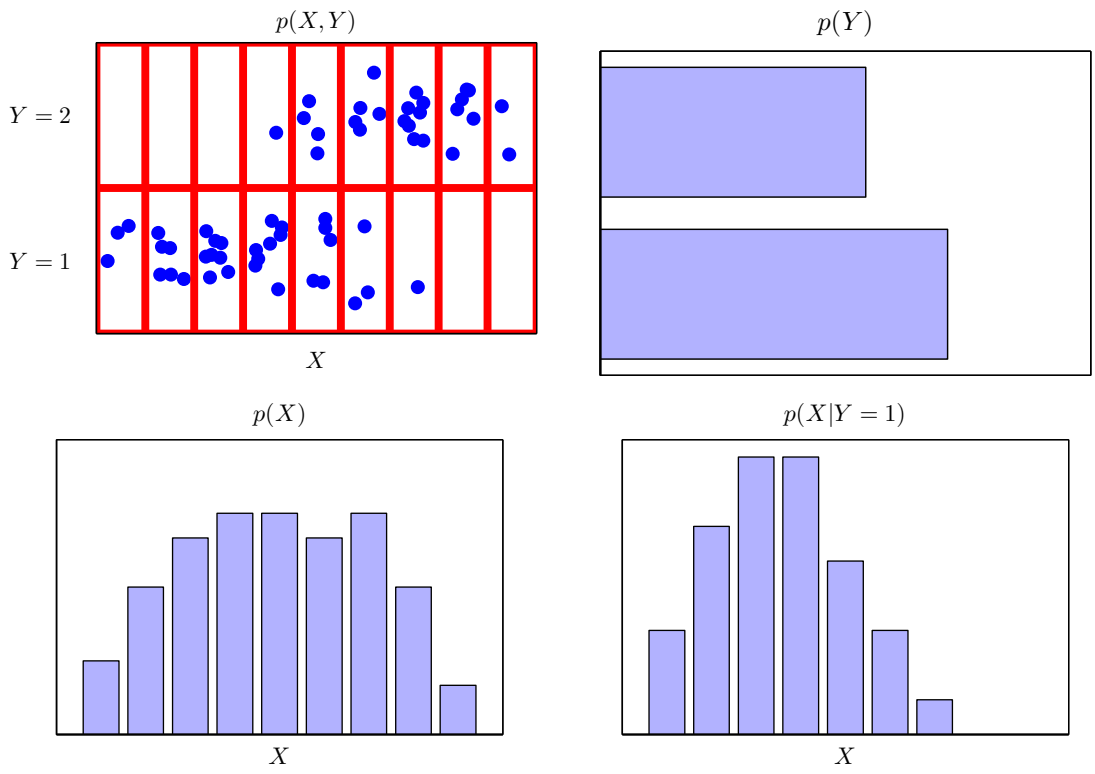

Figure 2: An example of a joint probability distribution $p(X, Y)$, where $Y \in \{1, 2\}$ and $X \in \{1, \ldots, 9\}$. Top left shows some samples from the joint, where the x coordinates have been jittered within each bin so that they can be seen. Top right: the marginal $p(Y)$. Bottom left: the marginal $p(X)$. Bottom right: the conditional $p(X|Y = 1)$. Source: [Bis06] Figure 1.11.

where $x_j \in \{0, 1\}$ and only one bit is turned on. In this case, we can think of X as a K-state random variable, and $\theta_j = p(X = j)$ is the probability of being in state j. We will write this as

$$p(X|\theta) = \prod_{j=1}^{K} \theta_j^{I(X=j)} \tag{8}$$

where $I(\cdot)$ is the **indicator function** which is 1 if its argument is true and 0 otherwise.

2.3 Joint, marginal and conditional distributions

Now suppose $X \in \{1, 2\}$ and $Y \in \{1, \ldots, 9\}$. We can represent the **joint probability distribution** $p(X, Y)$ as a 2×9 table of numbers.[1] If we **sample** from this distribution to get a dataset (x_i, y_i) for $i = 1 : n$ (where n is the sample size), we expect to get more points in the bins that have higher joint probability. See Figure 2 for an example.

The **sum rule** specifies how to compute a **marginal distribution** from a **joint distribution**:

$$p(X = i) = \sum_{j=1}^{K} p(X = i, Y = j) \tag{9}$$

This amounts to summing up along the Y dimension: see Figure 2. We can also do this with more than two variables. For example, Figure 3 illustrates computing $p(x, y) = \sum_z p(x, y, z)$.

[1] We use lower-case p to denote either a probability density function (for continuous rv's) or a probability mass function (for discrete rv's). Also, we follow the standard convention that random variables are denoted by upper case letters, and values of random variables are denoted by lower case letters. However, when we start treating parameters as random variables we will use usually use lower-case greek letters for both the variable and its value.

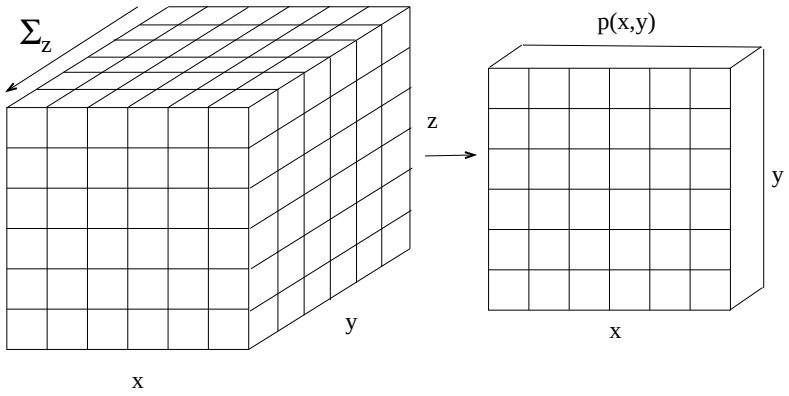

Figure 3: Computing $p(x, y) = \sum_z p(x, y, z)$ by marginalizing over dimension Z. Source: Sam Roweis.

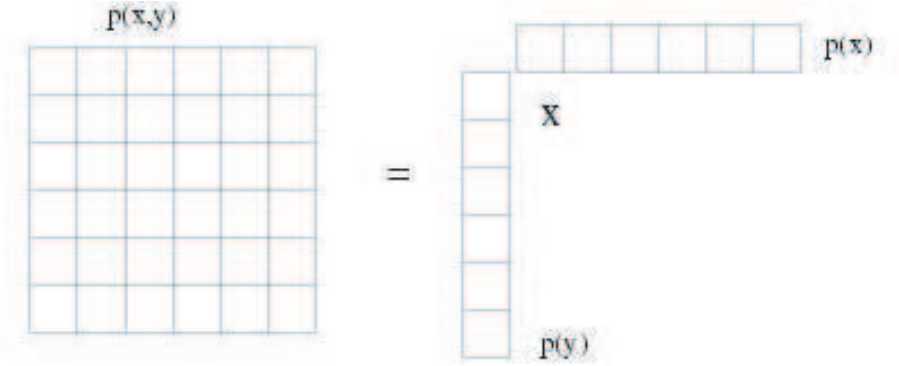

Figure 4: Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$. Source: Sam Roweis.

We define the **conditional probability distribution** as

$$p(X = i|Y = j) = \frac{p(X = i, Y = j)}{p(Y = j)} \tag{10}$$

provided $p(Y = j) > 0$, where $p(Y = j) = \sum_i p(X = i, Y = j)$ by the sum rule. The **product rule** specifies how to compute a joint distribution from the product of a marginal and a conditional distribution:

$$p(X = i, Y = j) = p(X = i|Y = j)p(Y = j) = p(Y = j|X = i)p(X = i) \tag{11}$$

The product rule can be applied multiple times to yield the **chain rule of probability**:

$$p(X_1, \ldots, X_d) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1) \ldots p(X_n|X_{1:d-1}) \tag{12}$$

where we introduce the notation $1 : d - 1$ to denote $\{1, 2, \ldots, d - 1\}$.

2.4 Conditional independence

We can simplify the chain rule by making **conditional independence assumptions**. We say Z and Y are conditionally independent given X, written as $Z \perp Y|X$, **if and only if (iff)** $p(Z, Y|X) = p(Z|X)p(Y|X)$.

If X and Y are unconditionally or **marginally independent**, $X \perp Y$, we can write $p(X, Y) = p(X)p(Y)$. Hence we can represent the joint as an outer product of the two marginals: see Figure 4.

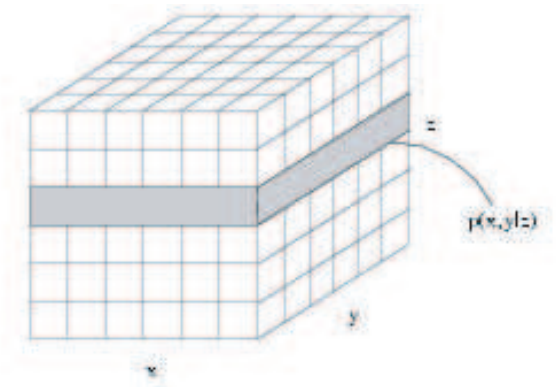

Figure 5: Computing $p(x, y|z)$ by extracting the slice from $p(x, y, z)$ corresponding to $Z = z$ and then renormalizing. Source: Sam Roweis.

	HIV-	HIV+	
Test-	97,902	5	97,907
Test+	1,998	95	2,093
	99,900	100	100,000

Table 1: Some statistics on a hypothetical HIV test.

2.5 Bayes rule

Combining the definition of conditional probability with the product and sum rules yields **Bayes rule**:

$$p(X = i|Y = j) = \frac{p(Y = j|X = i)p(X = i)}{p(Y = j)} \tag{13}$$

where

$$p(Y = j) = \sum_i p(Y = j|X = i)p(X = i) \tag{14}$$

is just the sum over the numerator.

We can think of $p(X|Y = j)$ as extracting the row from $p(X, Y)$ that corresponds to $Y = j$ and renormalizing. For example, Figure 2 shows $p(X|Y = 1)$ and Figure 5 shows $p(X, Y|z)$. However, we can also think of it as a way of "inverting" $p(Y|X)$ to get $p(X|Y)$. This is illustrated in the example below.

Let us consider an example of Bayes rule from [SAM04, p52]. Suppose a new home HIV test has 95% **sensitivity** and 98% **specifity**, and is to be used in a population of size 100,000 with an HIV prevalence of 1/1000. We expect $\frac{1}{1000} \times 100,000 = 100$ people to be truly HIV positive, of whom 95% (95) will test positive. Of the 99,900 negative individuals, we expect 2% (1998) to test positive. Thus of the 1998+5=2003 who test positive, only 95/2093 = 4.5% are truly HIV positive. Hence over 95% of the people testing positive will not in fact have HIV! This is illustrated in Table 1.

Let us see how this conclusion follows from Bayes' rule. Let H_1 be the hypothesis that the person is truly negative and H_0 be the hypothesis that the person is truly positive. x is the event the person tests positive. Then the model becomes

$$P(H_0) = 0.001, \;\; p(x|H_0) = 0.95, \;\; p(x|H_1) = 0.02 \tag{15}$$

Hence

$$P(H_0|x) = \frac{p(x|H_0)p(H_0)}{p(x)} \tag{16}$$

$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} = 0.045 \tag{17}$$

Later we show how to use Bayes rule to classify items (such as email) into different groups (such as spam or not-spam) using a **naive Bayes classifier**.

2.6 Functions of a random variable

Suppose $X \in 1:10$ has a uniform distribution, so $p(X = i) = 0.1$ if $i \in 1:10$. Consider another rv $Y = g(X)$, where g is some function. We can compute $p(Y = y)$ by simply summing up the probability mass for all the x's such that $g(x) = y$. For example, if $g(X) = 1$ if X is even and $g(X) = 0$ otherwise, then

$$p(Y = 1) = \sum_{x \in \{2,4,6,8,10\}} p(x) = 5/10 \tag{18}$$

and $p(Y = 0) = 0.5$ similarly. Note that in this example, g is a many-to-one function.

3 Discrete random variables

3.1 Bernoulli distribution

Let $X \in \{0, 1\}$ be a binary random variable (e.g., a coin toss). Suppose $p(X = 1) = \theta$. Then

$$p(X|\theta) = \text{Be}(X|\theta) = \theta^X (1-\theta)^{1-X} \tag{19}$$

is called a **Bernoulli distribution**. It is easy to show that

$$\begin{aligned} \text{mean} &= p(X = 1) = \theta & (20)\\ \text{mode} &= I(\theta > 0.5) & (21)\\ \text{var} &= \theta(1-\theta) & (22) \end{aligned}$$

The mode is the most probable value, and is 1 if $\theta > 0.5$ and is 0 otherwise.

3.2 Binomial distribution

Let X be the *number* of heads out of n trials. It has distribution

$$Bin(X|n, \theta) = \binom{n}{X} \theta^X (1-\theta)^{n-X} \tag{23}$$

where

$$\binom{N}{X} = \frac{N!}{(N-X)!X!} \tag{24}$$

is the number of ways to choose X items from N. This is called a **binomial distribution**. See Figure 1 for an example. The distribution has the following properties:

$$\begin{aligned} \text{mean} &= n\theta & (25)\\ \text{mode} &= \lfloor (n+1)\theta \rfloor & (26)\\ \text{var} &= n\theta(1-\theta) & (27) \end{aligned}$$

Note that the mode cannot be found by differentiation, since this is a discrete distribution. But since the distribution increases monotonically and then decreases monotonically, a mode does exist, and is the integer m such that

$$(n+1)\theta - 1 < m \leq (n+1)\theta \tag{28}$$

3.3 Multinomial distribution

The multivariate version of a binomial is called a multinomial. Suppose we have an urn with contains balls with colors $1, 2, \ldots, K$. Let the probability that we draw a ball of color j be θ_j. Suppose we draw n balls in total. Let

$X = (X_1, \ldots, X_K)$ be a random vector, where X_j is the number of balls of color j. Then X has a multinomial distribution with parameters n, θ, written $X \sim Mu(n, \theta)$. The pmf is

$$p(X|n,\theta) = \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \tag{29}$$

where

$$\binom{n}{x_1 \ldots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!} \tag{30}$$

and $n = \sum_{k=1}^{K} x_k$. The distribution has the following properties:

$$\begin{aligned} \text{mean} &= n\theta & (31)\\ \text{cov} &= \Sigma & (32)\\ \Sigma_{ii} &= n\theta_i(1-\theta_i) & (33)\\ \Sigma_{ij} &= -n\theta_i\theta_j & (34) \end{aligned}$$

It can be shown that the marginals are binomial, i.e.,

$$p(X_j|n,\theta) = \sum_{X_{-j}} Mu(X|n,\theta) = Bin(n,\theta_j) \tag{35}$$

where the sum is over all the variables except X_j.

Often we will consider the case where $n = 1$, so only one $X_j = 1$ and the rest are 0. In this case, we can think of X as being a **categorical** random variable with K states (values). If $X = j$, we represent it as a binary vector with only the j'th bit on; this is called a **1-of-K** encoding. In this case, the pmf becomes

$$p(X|\theta) = Mu(X|1,\theta) = \prod_{j=1}^{K} \theta_j^{I(x_j=1)} \tag{36}$$

where $I(x = j) = 1$ if $x = j$ and $I(x = j) = 0$ otherwise.

3.4 Poisson distribution

The **Poisson** distribution with parameter $\lambda > 0$ is defined by

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2 \ldots \tag{37}$$

Since $e^{\lambda} = \sum_{k=0}^{\infty} (\lambda^k/k!)$, it follows that this sums to 1. See Figure 6 for some examples. The Poisson distribution can be derived as the limit of a binomial distribution as $N \to \infty$ and $p \to 0$ such that $Np = \lambda$.

4 Continuous random variables

In the examples above, $p(X = i)$ is the probability that X takes on value i; this **probability mass function (pmf)** satisfies $\sum_i p(X = i) = 1$. If X is a continuous random variable, e.g., $X \in \mathbb{R}$ or $X \in \mathbb{R}^+$, then we use a **probability density function (pdf)** which satisfifes $\int_S p(X = x)dx = 1$, where we integrate over the **support** S of the distribution (the set of points with non zero probability). It is called a density because we must multiply it by an interval of size dx to find the probability of being in that interval:

$$p(x)dx \approx P(x \leq X \leq x + dx) \tag{38}$$

See Figure 7. We require $p(x) \geq 0$, but it is possible for $p(x) > 1$ for any given x, so long as the density integrates to 1: $\int_S p(x)dx = 1$. The **Gaussian** or **normal** distribution is an example that you are probably already familiar with: see Section 4.1.

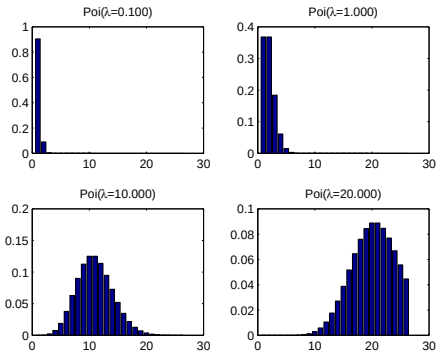

Figure 6: Some Poisson pmf's. Made with `poissonDemo`.

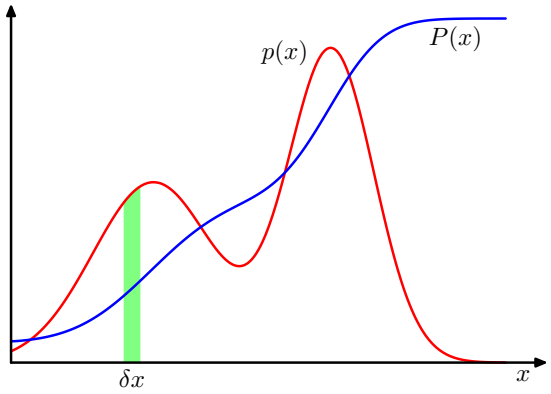

Figure 7: A probability density function $p(x)$ on a continuous random variable x, together with its cumulative distribution function $P(x)$. Source: [Bis06] Figure 1.12.

The probability that X lies in an interval (a, b) is given by

$$p(a \leq X \leq b) = \int_a^b p(x)dx \tag{39}$$

The probability that X lies in the interval $(-\infty, x)$ is given by

$$p(X \leq x) \stackrel{\text{def}}{=} P(x) = \int_{-\infty}^{x} p(x')dx' \tag{40}$$

$P(x)$ is called the **cumulative distributions function (cdf)**. Clearly $p(x) = \frac{d}{dx}P(x)$.

The **quantiles** of a distribution are defined as follows. Let f be a pdf, and let $f(\alpha)$ be the point beyond which f has probability α:

$$p_f(x > f(\alpha)) = \alpha \tag{41}$$

The quantity $f(\alpha)$ is called the α quantile of distribution f. For example, if $\alpha = 0.5$, then $f(\alpha)$ is the **median**. We can compute a $1 - \alpha$ **confidence interval** for X as follows:

$$P(f(1 - \alpha/2) \leq X \leq f(\alpha/2)) = 1 - \alpha \tag{42}$$

If we set $\alpha = 0.05$, then we get a 95% confidence interval. See Figure 8 for an example.

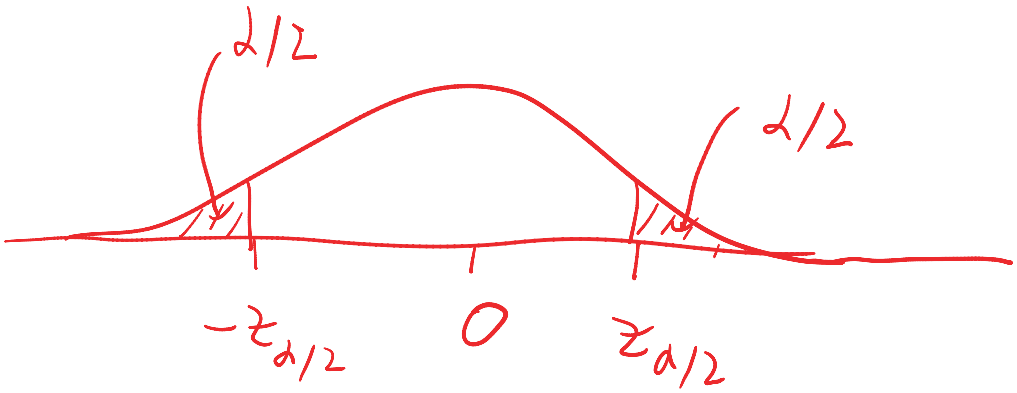

Figure 8: A $\mathcal{N}(0,1)$ distribution with the $z_{\alpha/2}$ cutoff points shown. The central non shaded area contains $1-\alpha$ of the probability mass. If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96 \approx 2$.

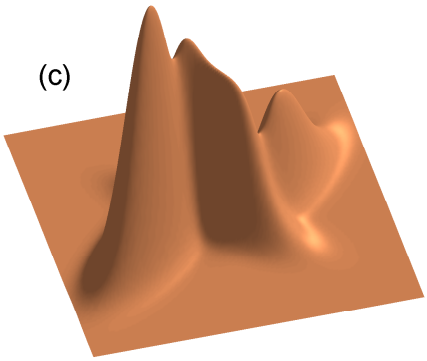

Figure 9: An example of a complex 2D pdf. Source: [Bis06] Figure 2.23.

Note that we can define joint pdf's on two or more variables. For example, Figure 9 shows a complex 2D pdf. Given a joint $p(x,y)$, we can define marginals and conditionals by analogy to the discrete case, just replacing sums with integrals:

$$p(x) = \int p(x,y)dy \tag{43}$$

$$p(x|y) = \frac{p(x,y)}{\int p(x,y)dx} \tag{44}$$

Of course, evaluating these integrals may be difficult, especially in high dimensional spaces (i.e., where there are many variables in the joint). We will consider a variety of computational techniques for efficient approximate integration later.

4.1 Univariate Gaussian distribution

The **Gaussian** or **normal** distribution gives rise to the famous **bell-shaped curve** in Figure 10. For one-dimensional variables, this is defined as

$$\mathcal{N}(x|\mu,\sigma) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{45}$$

μ is the mean and σ^2 is the variance. Note that sometimes we will use the **precision** parameter $\lambda = 1/\sigma^2$ instead of the variance, which we shall write as

$$\mathcal{N}_\lambda(x|\mu,\lambda) \tag{46}$$

Function	Matlab	R
density	`normpdf`	`dnorm`
cdf	`normcdf`	`pnorm`
inverse cdf (quantiles)	`norminv`	`qnorm`
sampling	`randn`	`rnorm`

Table 2: Translation between Matlab and R for common functions related to univariate gaussians.

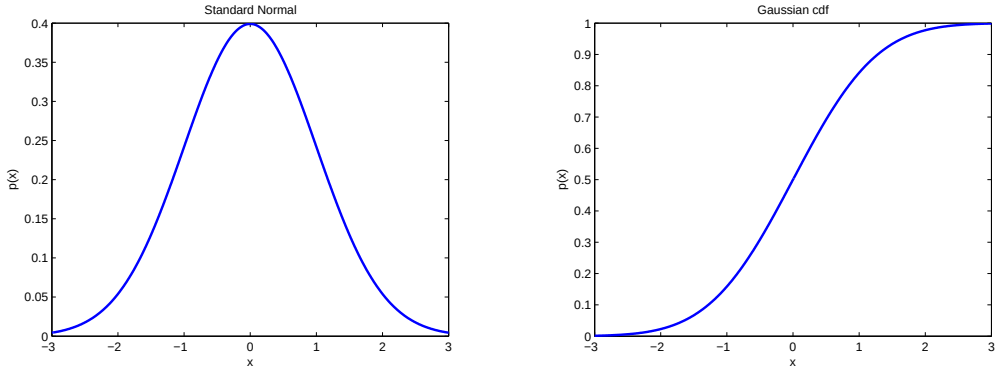

Figure 10: A standard normal pdf and cdf. The matlab code used to produce these plots is `xs=-3:0.01:3; plot(xs,normpdf(xs,mu,sigma)); plot(xs,normcdf(xs,mu,sigma));` , where $xs = [-3, -2.99, -2.98, \ldots, 2.99, 3.0]$ is a vector of points at which the density is evaluated.

Since the distribution is symmetric, μ is also the mode. In other words, the distribution has these properties

$$\text{mean} = \mu \tag{47}$$
$$\text{mode} = \mu \tag{48}$$
$$\text{var} = \sigma^2 \tag{49}$$

See Table 2 for some useful functions in Matlab/ R for manipulating Gaussians.

If $Z \sim \mathcal{N}(0, 1)$, we say Z follows a **standard normal** distribution. Its **cumulative distribution function (cdf)** is defined as

$$\Phi(x) = \int_{-\infty}^{x} p(z)dz \tag{50}$$

which is called the **probit distribution**. This has no closed form expression, but is built in to most software packages (eg. **normcdf** in the matlab statistics toolbox). In particular, we can compute it in terms of the **error (erf) function**

$$\Phi(x; \mu, \sigma) = \frac{1}{2}[1 + erf(z/\sqrt{2})] \tag{51}$$

where $z = (x - \mu)/\sigma$ and

$$erf(x) \stackrel{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{52}$$

Let us see how we can use the cdf to compute how much probability mass is contained in the interval $\mu \pm 2\sigma$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. The amount of mass contained inside the 2σ interval is given by

$$p(a < X < b) = p(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) \tag{53}$$
$$= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}) \tag{54}$$

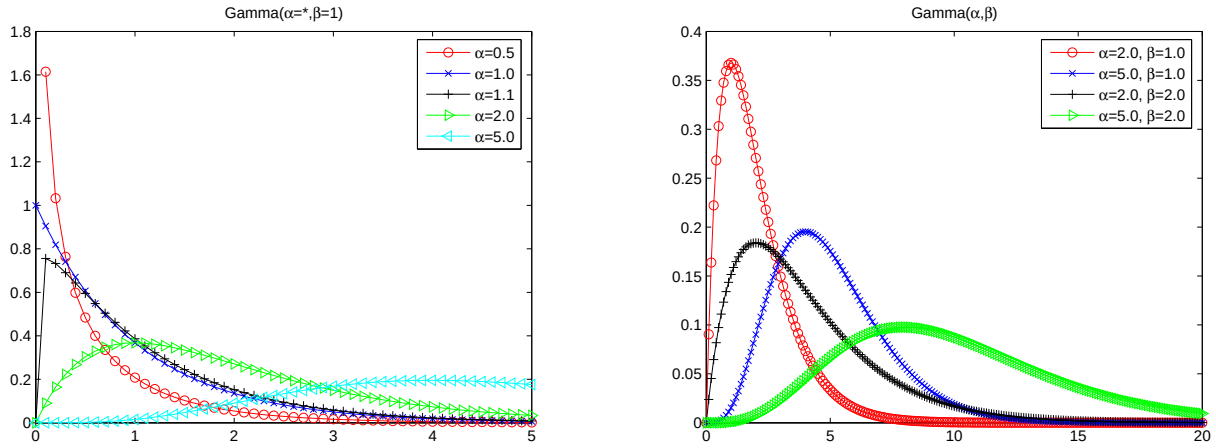

Figure 11: Some $Ga^{scale}(\alpha, \beta)$ distributions. Left: we fix the scale $\beta = 1$. If $\alpha \leq 1$, the maximum is at 0, and the distribution decays faster the smaller α gets. If $\alpha > 1$, the distribution has a maximum that is not at 0, and this peak shifts to the right the larger α gets. Right: to illustrate the effect of changing scale, we plot some pdfs for $\beta = 1$ and $\beta = 2$. We see that increasing β "stretches" the curves, and therefore decreases the height of the peak. Figures generated by `gammaDistPlot`.

Since

$$p(Z \leq -1.96) = \text{normcdf}(-1.96) = 0.025 \tag{55}$$

we have

$$p(-1.96\sigma < X - \mu < 1.96\sigma) = 1 - 2 \times 0.025 = 0.95 \tag{56}$$

Often we approximate this by replacing 1.96 with 2, and saying that the interval $\mu \pm 2\sigma$ contains 0.95 mass.

The α **quantile** is given by the value $z(\alpha)$ such that $P(Z \geq z(\alpha)) = \alpha$, or $P(Z \leq z(\alpha)) = 1 - \alpha$:

$$z(\alpha) = \Phi^{-1}(1 - \alpha) \tag{57}$$

For example, if $\alpha = 0.025$, we find

$$z(0.025) \quad = \quad \text{norminv}(1 - 0.025) = 1.96 \tag{58}$$

4.2 Gamma distribution

The gamma distribution is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters. There are two common parameterizations. This is the one used by Bishop [Bis06] (and many other authors):

$$Ga^{rate}(x|\text{shape} = a, \text{rate} = b) \quad = \quad \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}, \quad x, a, b > 0 \tag{59}$$

The second parameterization (and the one used by Matlab) is

$$Ga^{scale}(x|\text{shape} = \alpha, \text{scale} = \beta) \quad = \quad \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} = Ga(x|\alpha, 1/\beta) \tag{60}$$

Note that the shape parameter controls the shape; the scale parameter merely defines the measurement scale (the horizontal axis). The rate parameter is just inverse scale. See Figure 11 for some examples. This distribution has the

following properties (using the rate parameterization):

$$\text{mean} = \frac{a}{b} \tag{61}$$

$$\text{mode} = \frac{a-1}{b} \text{ for } a \geq 1 \tag{62}$$

$$\text{var} = \frac{a}{b^2} \tag{63}$$

4.3 Change of variables formula

Let X be an rv with pdf $p_x(x)$, and let $Y = g(X)$ for some function g. What is $p_y(y)$? This is harder to answer than in the discrete case, because p_x and p_y are probability densities, so instead we should work with the cdf's.

Let us consider an example, where $g(X) = aX + b$ for $a > 0$. Then

$$P_Y(y) = P(Y \leq y) \tag{64}$$

$$= P(aX + b \leq y) \tag{65}$$

$$= p(X \leq \frac{y-b}{a}) \tag{66}$$

$$= P_X(\frac{y-b}{a}) \tag{67}$$

Hence

$$p_Y(y) = \frac{d}{dy} P_X(\frac{y-b}{a}) \tag{68}$$

$$= \frac{1}{a} p_x(\frac{y-b}{a}) \tag{69}$$

The case for $a < 0$ can be analyzed similarly.

In general, we have the following result, which is called the **change of variables** formula.

Theorem Let X be a continuous rv with density $p_x(x)$ and let $Y = g(X)$, where g is strictly monotonic (so $x = g^{-1}(y)$ exists) and g is differentiable on some interval I. Suppose that $p_x(x) = 0$ for $x \notin I$. Then Y has the density

$$p_y(y) = p_x(x)|\frac{dx}{dy}| \tag{70}$$

where $x = g^{-1}(y)$. We set $p_y(y) = 0$ if $y \neq x$ for any $x \in I$.

Using the example above, we have

$$g^{-1}(y) = \frac{y-b}{a} \tag{71}$$

$$\frac{d}{dy} g^{-1}(y) = \frac{1}{a} \tag{72}$$

$$p_y(y) = \frac{1}{|a|} p_x(\frac{y-b}{a}) \tag{73}$$

The term $|\frac{dx}{dy}|$ is called the **Jacobian**. We use the absolute value of the derivative because we are only interested in how the unit measure changes in magnitude. (For example, we don't care if $a > 0$ or $a < 0$.)

We can understand this result more intuitively as follows. Observations falling in the range $(x, x + \delta x)$ will get transformed into $(y, y + \delta y)$, where $p_x(x)\delta x \approx p_y(y)\delta_y$. Hence $p_y(y) = p_x(x)|\frac{dx}{dy}|$.

One consequence of this is that the maximum of a pdf depends on the parameterization (choice of variable).

We can extend the above analysis to joint distributions. Suppose x_1, x_2 have joint distribution $p_x(x_1, x_2)$ and let $(y_1, y_2) = g(x_1, x_2)$, where g is an invertible transform. Then

$$p_y(y_1, y_2) = p_x(x_1, x_2)|J_{x/y}| = p_x(x_1, x_2)|J_{y/x}^{-1}| \tag{74}$$

where J is the **Jacobian** (how much the unit volume changes), defined as

$$J_{x/y} = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \stackrel{\text{def}}{=} \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} \tag{75}$$

where $\det$ is the determinant (since we use $|J|$ to denote absolute value). Thus J is a scalar. More mnemonically , we can write this as

$$p_{new} = p_{old}|J_{old/new}| = p_{old}|J^{-1}_{new/old}| \tag{76}$$

As an example, consider transforming a density from polar (r, θ) to Cartesian (x, y) coordinates:

$$(r, \theta) \rightarrow (x = r\cos\theta, y = r\sin\theta) \tag{77}$$

Then

$$\begin{aligned} J_{new/old} &= \frac{\partial(x, y)}{\partial(r, \theta)} && (78) \\ &= \det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} && (79) \\ &= \det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} && (80) \\ &= -r\sin^2\theta - r\cos^2\theta && (81) \\ &= = -r && (82) \end{aligned}$$

Hence

$$p_{X,Y}(x, y) = p_{R,\Theta}(r, \theta)|J^{-1}_{new/old}| = p_{R,\Theta}(r, \theta)\frac{1}{r} \tag{83}$$

To see this geometrically, notice that

$$p_{R,\Theta}(r, \theta)drd\theta = P(r \leq R \leq r + dr, \theta \leq \Theta \leq \theta + d\theta) \tag{84}$$

is the area of the shaded patch in Figure 12, which is clearly $rdrd\theta$, times the density at the center of the patch. Hence

$$P(r \leq R \leq r + dr, \theta \leq \Theta \leq \theta + d\theta) = p_{X,Y}(r\cos\theta, r\sin\theta)r\ dr\ d\theta \tag{85}$$

Hence

$$p_{R,\Theta}(r, \theta) = p_{X,Y}(r\cos\theta, r\sin\theta)r \tag{86}$$

4.4 Central limit theorems

There are a large number of limit theorems in statistics which describe the behavior of sums (and other functions) of independent random variables as the number of summands tends to infinity. The details are beyond the scope of this chapter. Here we just informally state the most famous of such theorems, the **central limit theorem**.

Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightarrow \mathcal{N}(0, 1) \tag{87}$$

i.e., sums of iid rv's converge (in distribution) to a Gaussian (normal) distribution. See Figure 13 for an example.

One reason this is useful is the following. If we have a variable that is subject to a large number of additive random effects, then rather than modeling each factor separately, we can model their net effect, which is to add Gaussian noise to the variable. Thus we use a Gaussian to summarize our **ignorance** of the true causes of the output. (The Gaussian can also be motivated by the fact that it is the unique distribution which maximizes the **entropy** subject to first and second moment constraints. We discuss this later.)

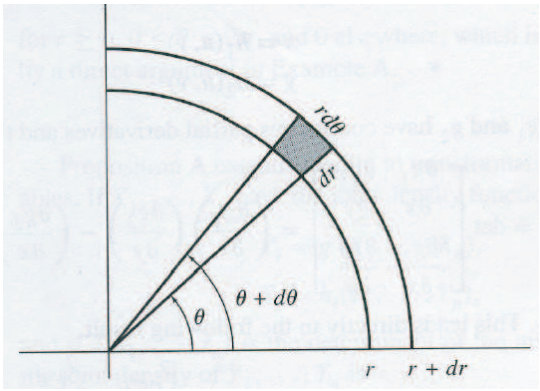

Figure 12: Change of variables from polar to Cartesian. The area of the shaded patch is $rdrd\theta$. Source: [Ric95] Figure 3.16.

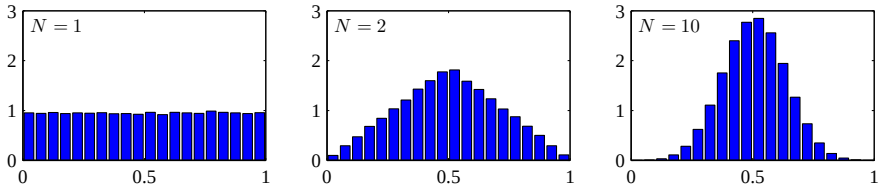

Figure 13: The central limit theorem in pictures. We plot a histogram of $\frac{1}{M}\sum_{i=1}^{M} x_i$, where $x_i \sim U(0,1)$. As $M \to \infty$, the distribution tends towards a Gaussian. Source: [Bis06] Figure 2.6.

5 Moments of a distribution

Since probability distributions can be quite complex, we often characterize them in terms of some simple scalar quantities, which capture the basic shape of the distribution. We consider some of the most important quantities below.

5.1 Expectation

We define the **expected value** of an RV X to be

$$\mu_X \stackrel{\text{def}}{=} E\,X \stackrel{\text{def}}{=} \sum_x xp(X=x) \tag{88}$$

We replace the sum by an integral if X is continuous. By **linearity of expectation**, we can push E inside $\sum$:

$$E\left(\sum_i a_i X_i\right) = \sum_i a_i E\left(X_i\right) \tag{89}$$

where the a_i are constants. If X is a **random vector**,

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \tag{90}$$

then its mean is denoted by

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \tag{91}$$

Note that we will often just write μ instead of $\vec{\mu}$. If a is a vector and A a matrix, we have the following two important results (which follow from linearity of expectation):

$$E\left(a^T X\right) = a^T \mu \tag{92}$$
$$E\left(AX\right) = A\mu \tag{93}$$

In particular, for any two random variables X,Y, whether independent or not, we have

$$E\left[aX + bY + c\right] = aE\ X + bE\ Y + c \tag{94}$$

Also, if X,Y are independent,

$$E\left[XY\right] = \left[E\ X\right]\left[E\ Y\right] \tag{95}$$

The **conditional expectation** is defined as

$$E\left(X|Y = y\right) \stackrel{\text{def}}{=} \sum_x x\ p(x|y) \tag{96}$$

Note that whereas $E\left(X\right)$ is a number, $E\left(X|Y\right)$ is a function of Y. The important **rule of iterated expectations** is

$$E\left[E\left(Y|X\right)\right] = E\left(Y\right) \tag{97}$$

This is easy to prove:

$$\begin{aligned}
E\left[E\left(Y|X\right)\right] &= \sum_x E\left(Y|X = x\right)p(X = x) && (98)\\
&= \sum_x \sum_y yp(Y = y|X = x)p(X = x) && (99)\\
&= \sum_y y\left[\sum_x p(Y = y, X = x)\right] && (100)\\
&= \sum_y yp(Y = y) && (101)\\
&= E\ Y && (102)
\end{aligned}$$

5.1.1 Trace trick

We will frequently have to compute the expected value of a weighted inner product, $E\left[\mathbf{x}^T A\mathbf{x}\right]$. The way to do this is to realise that $\mathbf{x}^T A\mathbf{x}$ is a scalar, and hence $\mathbf{x}^T A\mathbf{x} = \text{tr}(\mathbf{x}^T A\mathbf{x})$, where $\text{tr}(A) = \sum_{ii} A_{ii}$ is the **trace** of a matrix. Now using the **cyclic permutation property** of the trace operator

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA) \tag{103}$$

we get

$$\mathbf{x}^T A\mathbf{x} = \text{tr}(\mathbf{x}^T A\mathbf{x}) = \text{tr}(A\mathbf{x}\mathbf{x}^T) \tag{104}$$

This is called the **trace trick**. Hence

$$\begin{aligned}
E\left[\mathbf{x}^T A\mathbf{x}\right] &= E\left[\text{tr}(\mathbf{x}^T A\mathbf{x})\right] = E\left[\text{tr}(A\mathbf{x}\mathbf{x}^T)\right] && (105)\\
&= \text{tr}(AE\left[\mathbf{x}\mathbf{x}^T\right]) = \text{tr}(A(\Sigma + \mathbf{m}\mathbf{m}^T)) && (106)\\
&= \text{tr}(A\Sigma) + \mathbf{m}^T A\mathbf{m} && (107)
\end{aligned}$$

where $E\ \mathbf{x} = \mathbf{m}$ and $\text{Cov}\mathbf{x} = \Sigma$.

5.2 Variance

The **variance** is a measure of spread:

$$\begin{aligned}\sigma^2 &\stackrel{\text{def}}{=} \text{Var}X = E(X-\mu)^2 && (108)\\ &= \int (x-\mu)^2 p(x)dx && (109)\\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx && (110)\\ &= E[X^2] - \mu^2 && (111)\end{aligned}$$

from which we infer the useful result $E[X^2] = \mu^2 + \sigma^2$. The **standard deviation** is defined as

$$\sigma_X \stackrel{\text{def}}{=} \sqrt{\text{Var } X} \tag{112}$$

It is easy to show

$$\text{Var } (aX+b) = a^2 \text{Var } (X) \tag{113}$$

where a and b are constants.

The variance of a sum is

$$\text{Var } [X+Y] = \text{Var } X + \text{Var } Y + 2\text{Cov}(X,Y) \tag{114}$$

The **conditional variance** is defined as

$$\text{Var } (Y|X=x) \stackrel{\text{def}}{=} \sum_y (y - E\ (Y|x))^2\ p(y|x) \tag{115}$$

The **rule of iterated variance** is

$$\text{Var } (Y) = E \text{ Var } (Y|X) + \text{Var } E\ (Y|X) \tag{116}$$

This can be proved as follows. Let $\mu = E[Y|X]$. Then

$$\begin{aligned}E \text{ Var } (Y|X) + \text{Var } E\ (Y|X) &= E\left[E(Y^2|X) - \mu^2\right] + E[\mu^2] - [E^2\mu] && (117)\\ &= E(Y^2) - E[\mu^2] + E[\mu^2] - E^2(\mu) && (118)\\ &= E(Y^2) - (E(E[Y|X]))^2 && (119)\\ &= E(Y^2) - (E^2 Y) && (120)\\ &= \text{Var } Y && (121)\end{aligned}$$

5.3 Covariance

The **covariance** between two RVs X and Y is defined as

$$\begin{aligned}\text{Cov}(X,Y) &\stackrel{\text{def}}{=} E\ ((X-\mu_X)(Y-\mu_Y)) && (122)\\ &= E\ (XY) - E\ (X)E\ (Y) && (123)\end{aligned}$$

If X is a random vector, its **covariance matrix** is defined to be

$$\text{Var } (X) = \Sigma \stackrel{\text{def}}{=} E\ [(X - E\ X)(X - E\ X)'] = \begin{pmatrix} \text{Var } (X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_p, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var } (X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var } (X_p) \end{pmatrix} \tag{124}$$

If a is a vector and A a matrix, we have the following two important results:

$$\begin{aligned}\text{Var } (a^T X) &= a^T \Sigma a && (125)\\ \text{Var } (AX) &= A\Sigma A^T && (126)\end{aligned}$$

The **conditional covariance** is defined as

$$\text{Cov}(X, Y|Z = z) \stackrel{\text{def}}{=} \sum_{x,y} p(x, y|z)(x - E\,(x|z))(y - E\,(Y|z)) \tag{127}$$

which is a function of Z.

5.3.1 Correlation

The correlation is defined as

$$\rho(X, Y) \stackrel{\text{def}}{=} \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{128}$$

We can show $-1 \leq \rho(X, Y) \leq 1$ as follows.

$$\begin{aligned} 0 &\leq \text{Var}\,(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}) && (129)\\ &= \text{Var}\,(\frac{X}{\sigma_X}) + \text{Var}\,(\frac{Y}{\sigma_Y}) + 2\text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) && (130)\\ &= \frac{\text{Var}\,X}{\sigma_X} + \frac{\text{Var}\,Y}{\sigma_Y} + 2\text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) && (131)\\ &= 1 + 1 + 2\rho && (132) \end{aligned}$$

Hence $\rho \geq -1$. Similarly,

$$0 \leq \text{Var}\,(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}) = 2(1 - \rho) \tag{133}$$

so $\rho \leq 1$.

If $Y = aX + b$, then $\rho(X, Y) = 1$ if $a > 0$ and $\rho(X, Y) = -1$ if $a < 0$. Thus **correlation only measures linear relationships** between RVs. If X and Y are independent, then $\text{Cov}(X, Y) = \rho = 0$; however, the converse is not true, as we see below.

The **partial correlation coefficient** is defined as

$$r_{XY|Z} \stackrel{\text{def}}{=} \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}} \tag{134}$$

and measures the linear dependence of X and Y when Z is fixed.

5.3.2 Uncorrelated does not necessarily imply independent

.

Consider two RVs $X, Y \in \{-1, 0, 1\}$ with the following joint distribution:

$$p(X, Y) = \begin{pmatrix} X/Y| & 0 & -1 & 1 \\ \hline -1| & 0.25 & 0 & 0 \\ 0| & 0 & 0.25 & 0.25 \\ 1| & 0.25 & 0 & 0 \end{pmatrix} \tag{135}$$

The marginal distributions are clearly $p(X) = p(Y) = (0.25, 0.5, 0.25)$. We will first show that X and Y are uncorrelated. We have

$$\begin{aligned} E\,(X, Y) &= \sum_{x \in \{-1,0,1\}} \sum_{y \in \{-1,0,1\}} x\ y\ p(x, y) && (136)\\ &= -1 \cdot 0 \cdot 0.25 + 0 \cdot -1 \cdot 0.25 + 0 \cdot 1 \cdot 0.25 + 1 \cdot 0 \cdot 0.25 = 0 && (137) \end{aligned}$$

and

$$E\ X = \sum_{x \in \{-1,0,1\}} xp(x) = -1 \cdot 0.25 + 0 \cdot 0.5 + 1 \cdot 0.25 = 0 \tag{138}$$

Similarly $E\ Y = 0$. Hence

$$\text{Cov}(X, Y) = E\ (X, Y) - E\ (X)E\ (Y) = 0 - 0 \tag{139}$$

However, it is easy to see that X and Y are not independent: i.e., $p(X, Y) \neq p(X)p(Y)$. We can simply multiply out the two marginals, c.f., Figure 4.

$$\begin{pmatrix} 0.25 \\ 0.5 \\ 0.25 \end{pmatrix} \begin{pmatrix} 0.25 & 0.5 & 0.25 \end{pmatrix} = \begin{pmatrix} 0.0625 & 0.1250 & 0.0625 \\ 0.1250 & 0.2500 & 0.1250 \\ 0.0625 & 0.1250 & 0.0625 \end{pmatrix} \tag{140}$$

5.4 Moment generating functions

The **moment generating function** of an rv X is $M(t) = E[e^{tX}]$ if the expectation is defined. In the continuous case, this is

$$M(t) = \int_{-\infty}^{\infty} e^{tx} p(x) dx \tag{141}$$

The derivative is

$$M'(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} p(x) dx = \int_{-\infty}^{\infty} x e^{tx} p(x) dx \tag{142}$$

Hence

$$M'(0) = \int_{-\infty}^{\infty} p(x) dx = E[X] \tag{143}$$

Differentiating r times, we find

$$M^{(r)}(0) = E[X^r] \tag{144}$$

Let us consider the Gamma distribution as an example. The mgf is

$$M(t) = \int_0^{\infty} e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \tag{145}$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{x(t-\lambda)} dx \tag{146}$$

This integral is equivalent to an unnormalized Gamma density with parameters α and $\lambda - t$, and hence is equal to the normalizing constant

$$\int_0^{\infty} x^{\alpha-1} e^{x(t-\lambda)} dx = \frac{\Gamma(\alpha)}{(\lambda - t)^\alpha} \tag{147}$$

Hence

$$M(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda - t)^\alpha} = \left(\frac{\lambda}{\lambda - t} \right)^\alpha \tag{148}$$

Differentiating we find

$$M'(0) = E[X] = \frac{\alpha}{\lambda} \tag{149}$$

$$M''(0) = E[X^2] = \frac{\alpha(\alpha + 1)}{\lambda^2} \tag{150}$$

Hence

$$\text{Var}\ [X] = E[X^2] - (E[X])^2 = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2} \tag{151}$$

6 Alternatives to probability theory

[This section, which is based on [RN02, sec 14.7], will not be on the exam, and is included only for background interest.]

Various alternatives to probability theory have been explored by philosophers and some members of the AI (artificial intelligence) community. We will see that all of them are unnecessary, and many are inconsistent.

Fuzzy logic is sometimes touted as an alternative to probability theory. However, it addresses a different sort of uncertainty. Probability theory assumes that an event either is true or is not, but we are uncertain of its state; this is called **epistemological uncertainty**. Fuzzy logic assumes that an event can be true to different degrees, but we know this degree; this is called **ontological uncertainty** (vagueness). For example, the statement "Nate is tall" is not obviously true or false; one might want to say "sort of". So we treat $Tall(Nate)$ as a number between 0 and 1, representing degree of membership of the **fuzzy set** of tall people. **Possibility theory** adds epistemological uncertainty on top of ontological uncertainty.

In fuzzy logic, the truth of an atomic proposition $T(A)$ is the degree to which A belongs to the fuzzy set A. The rules for evaluating the fuzzy truth T of a logical sentence are

$$T(A \wedge B) = \min(T(A), T(B)) \tag{152}$$

$$T(A \vee B) = \max(T(A), T(B)) \tag{153}$$

$$T(\neg A) = 1 - T(A) \tag{154}$$

This is called a **truth-functional** system, since the truth of a sentence is a function of the truth of its parts. Unfortunately, this is a serious problem. For example, suppose $T(Tall(Nate)) = 0.6$ and $T(Heavy(Nate)) = 0.4$. Then $T(Tall(Nate) \wedge Heavy(Nate)) = 0.4$, which seems reasonable, but we also get the result $T(Tall(Nate) \wedge \neg Tall(Nate)) = 0.4$ which does not seem reasonable. (It violates the **law of excluded middle**, which says Nate is either tall or is not.) The problem arises because truth-functional systems cannot handle the correlations between the component propositions.

A much better approach to epistemological uncertainty is to use standard (Bayesian) probability theory, where we model our uncertainty about the definition of the word "tall". This can be represented as uncertainty over the members of the set of tall people. We will see similar examples when we look at Bayesian concept learning, and in particular the "number game".

Dempster-Shafer theory is designed to deal with the distinction between uncertainty and ignorance. Rather than computing the probability of a proposition, it computes the probability that the evidence supports the proposition. This measure of belief is called a **belief function**, and is written $Bel(X)$. For example, suppose you want to predict if a coin will come up heads or tails. Since you have no evidence either way, your belief function is $Bel(Heads) = 0$ and also $Bel(Tails) = 0$. Of course, this means you cannot decide what action to take! If an expert tells you he is 90% sure the coin is fair, then $Bel(Heads) = 0.9 \times 0.5 = 0.45$ and likewise $Bel(Tails) = 0.45$; there is still a 10% gap not accounted for by the evidence.

A much better approach to modeling this problem is to use standard (Bayesian) probability theory, and to express your ignorance as uncertainty over the parameter, $p(\theta)$. If you have no evidence, you can use a flat or uniformative prior, $p(\theta) = U(0, 1)$. After getting evidence from the expert, your prior gets updated so that $p(\theta = 0.5|D) = 0.9$, with the remaining 0.1 mass spread over all other values.

References

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

[Jay03] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

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

[RN02] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2002. 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.