10 October 2007

In this lecture we’ll learn the following:

1. what joint probability distributions are;
2. visualizing multiple variables/joint probability distributions;
3. marginalization;
4. what covariance and correlation are;
5. a bit more about variance.

1 Joint probability distributions

Recall that a basic probability distribution is defined over a random variable, and a random variable maps from the sample space to the real numbers (\( \mathbb{R} \)). What about when you are interested in the outcome of an event that is not naturally characterizable as a single real-valued number, such as the two formants of a vowel?

The answer is really quite simple: probability distributions can be generalized over multiple random variables at once, in which case they are called joint probability distributions (jpd’s). If a jpd is over \( N \) random variables at once then it maps from the sample space to \( \mathbb{R}^N \), which is short-hand for real-valued vectors of dimension \( N \). Notationally, for random variables \( X_1, X_2, \cdots, X_N \), the joint probability density function is written as
\[ p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_n) \]

or simply

\[ p(x_1, x_2, \ldots, x_n) \]

for short.

Whereas for a single r.v., the cumulative distribution function is used to indicate the probability of the outcome falling on a segment of the real number line, the joint cumulative probability distribution function indicates the probability of the outcome falling in a region of \( N \)-dimensional space. The joint cpd, which is sometimes notated as \( F(x_1, \cdots, x_n) \) is defined as the probability of the set of random variables all falling at or below the specified values of \( X_i \):

\[ F(x_1, \cdots, x_n) \triangleq P(X_1 \leq x_1, \cdots, X_N \leq x_n) \]

The natural thing to do is to use the joint cpd to describe the probabilities of rectangular volumes. For example, suppose \( X \) is the \( f_1 \) formant and \( Y \) is the \( f_2 \) formant of a given utterance of a vowel. The probability that the vowel will lie in the region \( 480\text{Hz} \leq f_1 \leq 530\text{Hz}, 940\text{Hz} \leq f_2 \leq 1020\text{Hz} \) is given below:

\[
\begin{align*}
P(480\text{Hz} \leq f_1 \leq 530\text{Hz}, 940\text{Hz} \leq f_2 \leq 1020\text{Hz}) &= F(530\text{Hz}, 1020\text{Hz}) - F(530\text{Hz}, 940\text{Hz}) - F(480\text{Hz}, 1020\text{Hz}) + F(480\text{Hz}, 940\text{Hz}) \end{align*}
\]

and visualized in Figure 1 using the code below.

\[1\text{Technically, the definition of the multivariate cpd is then}
\]

\[ F(x_1, \cdots, x_n) \triangleq P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \sum_{\vec{x} \leq (x_1, \cdots, x_N)} p(\vec{x}) \quad \text{[Discrete]} \tag{1} \]

\[ F(x_1, \cdots, x_n) \triangleq P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} p(\vec{x}) \, dN \cdots dx_1 \quad \text{[Continuous]} \tag{2} \]
Figure 1: The probability of the formants of a vowel landing in the grey rectangle can be calculated using the joint cumulative distribution function.

```
plot(c(),c(),xlim=c(200,800),ylim=c(500,2500),xlab="f1",ylab="f2")
rect(480,940,530,1020,col=8)
```

1.1 Multinomial distributions as jpd’s

We touched on the multinomial distribution very briefly in a previous lecture. To refresh your memory, a multinomial can be thought of as a random set of \( n \) outcomes into \( r \) distinct classes—that is, as \( n \) rolls of an \( r \)-sided die where the tabulated outcome is the number of rolls that came up in each of the \( r \) classes. This outcome can be represented as a set of random variables \( X_1, \ldots, X_r \), or equivalently an \( r \)-dimensional real-valued vector. The \( r \)-class multinomial distribution is characterized by \( r - 1 \) parameters, \( p_1, p_2, \ldots, p_{r-1} \), which are the probabilities of each die roll coming out as each class. The probability of the die roll coming out in the \( r \)th class is \( 1 - \sum_{i=1}^{r-1} p_i \), which is sometimes called \( p_r \) but is not a true parameter of the model. The probability mass function looks like this:

\[
p(n_1, \ldots, n_r) = \binom{n}{n_1 \ldots n_r} \prod_{i=1}^{r} p_i
\]
2 Visualizing joint probability distributions

We'll look at two examples of visualizing jpd’s. First is a discrete distribution one out of the dative dataset in languageR. We'll look at the distribution of animacy of theme and recipient in this dataset, shown in Figure 2 using the code below:

```r
> xtabs(~ AnimacyOfRec + AnimacyOfTheme, data=dative)
AnimacyOfTheme
    animate inanimate
animate 68 2956
inanimate 6 233
```

The second example is of the joint distribution of frequency and length in the Brown corpus. First we'll put together an estimate of the joint distribution.

```r
persp(kde2d(x$Length,log(x$Count),n=50),theta=210,phi=15,
     xlab="Word Length",ylab="Word Frequency",zlab="p(Length,Freq)"
)
```

This gives a perspective plot of the joint distribution of length and frequency. Another way of visualizing the joint distribution is going along the length-axis slice by slice in a COПLOT:
Figure 3: Perspective plot and coplot (i.e., slice-by-slice histogram plot) of the relationship between length and log-frequency in the Brown corpus

```r
histogram(~ log(x$Count) | x$Length)
```

Both resulting plots are shown in Figure 3.

## 3 Marginalization

Often we have direct access to a joint density function but we are more interested in the probability of an outcome of a subset of the random variables in the joint density. Obtaining this probability is called **marginalization**, and it involves taking a weighted sum\(^2\) over the possible outcomes of the r.v.’s that are not of interest. For two variables \(X, Y\):

\[
P(X = x) = \sum_y P(x, y)
\]

\[
= \sum_y P(X = x | Y = y)P(y)
\]

In this case \(P(X)\) is often called a **marginal probability** and the process of calculating it from the joint density \(P(X, Y)\) is known as **marginalization.**

\(\text{or integral in the continuous case}\)
4 Covariance

The covariance between two random variables $X$ and $Y$ is defined as follows:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Simple example:

<table>
<thead>
<tr>
<th>Coding for $X$</th>
<th>Coding for $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Pronoun</td>
</tr>
<tr>
<td>1</td>
<td>Not Pronoun</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coding for $X$</th>
<th>Coding for $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Each of $X$ and $Y$ can be treated as a Bernoulli random variable with arbitrary codings of 1 for Postverbal and Not Pronoun, and 0 for the others. As a result, we have $\mu_X = 0.121$, $\mu_Y = 0.762$. The covariance between the two is:

$$(0 - .121) \times (0 - .762) \times .224 \quad (0,0)$$
$$(1 - .121) \times (0 - .762) \times 0.014 \quad (1,0)$$
$$(0 - .121) \times (1 - .762) \times 0.0655 \quad (0,1)$$
$$(1 - .121) \times (1 - .762) \times 0.107 \quad (1,1)$$

$$= 0.0148$$

In R, we can use the `cov()` function to get the covariance between two random variables, such as word length versus frequency across the English lexicon:

```r
> cov(x$Length, x$Count)
[1] -42.44823
> cov(x$Length, log(x$Count))
[1] -0.9333345
```

The covariance in both cases is negative, indicating that longer words tend to be less frequent. If we shuffle one of the covariates around, it eliminates this covariance:

```r
order() plus runif() give a nice way of randomizing a vector.
```
The covariance is essentially zero now.

An important aside: the variance of a random variable $X$ is just its covariance with itself:

$$\text{Var}(X) = \text{Cov}(X, X)$$

### 4.1 Covariance and scaling random variables

What happens to $\text{Cov}(X, Y)$ when you scale $X$? Let $Z = a + bX$. It turns out that the covariance with $Y$ increases by $b$:

$$\text{Cov}(Z, Y) = b\text{Cov}(X, Y)$$

As an important consequence of this, rescaling a random variable by $Z = a + bX$ rescales the variance by $b^2$: $\text{Var}(Z) = b^2\text{Var}(X)$.

### 4.2 Correlation

We just saw that the covariance of word length with frequency was much higher than with log frequency. However, the covariance cannot be compared directly across different pairs of random variables, because we also saw that random variables on different scales (e.g., those with larger versus smaller ranges) have different covariances due to the scale. For this reason, it is common to use the correlation $\rho$ as a standardized form of covariance:

---

The reason for this is as follows. By linearity of expectation, $E(Z) = a + bE(X)$. This gives us:

$$\text{Cov}(Z, Y) = E[(Z - a + bE(X))(Y - E(Y))]$$

$$= E[((bX - bE(X))(Y - E(Y))]$$

$$= E[b(X - E(X))(Y - E(Y))]$$

$$= bE[(X - E(X))(Y - E(Y))]$$

$$= b\text{Cov}(X, Y)$$

[by linearity of expectation]

---

*Linguistics 251 lecture 4 notes, page 7  Roger Levy, Fall 2007*
\[ \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \]

If \( X \) and \( Y \) are independent, then their covariance (and hence correlation) is zero.

### 4.3 Multivariate normal distributions

We’re now ready to deal with the **multivariate normal distribution**. Here we’ll just work with a 2-dimensional, or bivariate, distribution. Whereas the univariate normal distribution was characterized by two parameters—mean \( \mu \) and variance \( \sigma^2 \)—the bivariate normal distribution is characterized by two mean parameters (\( \mu_X, \mu_Y \)), two variance terms (one for the \( X \) axis and one for the \( Y \) axis), and one **covariance term** showing the tendency for \( X \) and \( Y \) to go together. The three variance and covariance terms are often grouped together into a symmetric **covariance matrix** as follows:

\[
\begin{bmatrix}
\sigma_{XX}^2 & \sigma_{XY}^2 \\
\sigma_{XY}^2 & \sigma_{YY}^2
\end{bmatrix}
\]

Note that the terms \( \sigma_{XX}^2 \) and \( \sigma_{YY}^2 \) are simply the variances in the \( X \) and \( Y \) axes (the subscripts appear doubled, \( XX \), for notational consistency). The term \( \sigma_{XY}^2 \) is the covariance between the two axes.

```r
library(mvtnorm)
sigma.xx <- 4
sigma.yy <- 1
sigma.xy <- 0  # no covariance
sigma <- matrix(c(sigma.xx,sigma.xy,sigma.xy,sigma.yy),ncol=2)
old.par <- par(mfrow=c(1,2))
x <- seq(-5,5,by=0.25)
y <- x
f <- function(x,y) {
    #cat("X: ", x,"\n")
    #cat("Y: ", y," \n")
    xy <- cbind(x,y)
    #cat("XY: ", xy,"\n")
    dmvnorm(xy,c(0,0),sigma)
}
```

---

*Linguistics 251 lecture 4 notes, page 8  Roger Levy, Fall 2007*
5 A bit more about variance, the binomial distribution, and the normal distribution

With covariance in hand, we can now express the variance of a sum of random variables. If $Z = X + Y$ then

$$Var(Z) = Var(X) + Var(Y) + 2Cov(X, Y)$$

As a result, if two random variables are independent then the variance of their sum is the sum of their variances.

As an application of this fact, we can now calculate the variance of a binomially distributed random variable $X$ with parameters $n, p$. $X$ can be expressed as the sum of $n$ identically distributed Bernoulli random variables $X_i$, each with parameter $p$. We’ve already established that the variance of each $X_i$ is $p(1-p)$. So the variance of $X$ must be $np(1-p)$.

5.1 Normal approximation to the binomial distribution

Finally, we can combine these facts with the central limit theorem to show how the binomial distribution can be approximated with the normal distribution. The central limit theorem says that the sum of many independent random variables tends toward a normal distribution with appropriate mean and variance. In our case, the binomial distribution has mean $np$ and variance $np(1-p)$. Thus, when $n$ is large, the binomial distribution should look approximately normal. We can test this:

```r
p <- 0.3
```
Figure 4: Approximation of binomial distribution with normal distribution for $p = 0.4$ (left) and $p = 0.05$ (right).

# then try with p = 0.05
par(old.par)