4.2.5 CONDITIONAL PROBABILITY DENSITY: SCALAR CASE

When two random variables are related, the probability density of a random variable changes when the other random variable takes on a particular value.

> The probability density of a random variable when one or more other random variables are fixed is called *conditional probability density*.

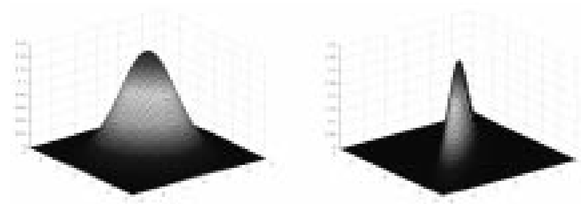
This concept is important in stochastic estimation as it can be used to develop estimates of unknown variables based on readings of other related variables.

Let x and y be random variables. Suppose $x$and y have joint probability density $\mathcal{P}(\zeta, \beta; x, y)$. One may then ask what the probability density of x is given a particular value of y (say $y = \beta$). Formally, this is called "conditional density function" of x *given* y and denoted as $\mathcal{P}(\zeta|\beta; x|y)$. $\mathcal{P}(\zeta|\beta; x|y)$ is computed as

$$\mathcal{P}(\zeta|\beta; x|y) = \frac{\lim_{\epsilon \to 0} \int_{\beta-\epsilon}^{\beta+\epsilon} \mathcal{P}(\zeta, \beta^*; x, y) d\beta^*}{\underbrace{\int_{-\infty}^{\infty} \int_{\beta-\epsilon}^{\beta+\epsilon} \mathcal{P}(\zeta, \beta^*; x, y) d\beta^* d\zeta}_{\text{normalization factor}}} \tag{4.51}$$

$$= \frac{\mathcal{P}(\zeta, \beta; x, y)}{\int_{-\infty}^{\infty} \mathcal{P}(\zeta, \beta; x, y) d\zeta} \tag{4.52}$$

$$= \frac{\mathcal{P}(\zeta, \beta; x, y)}{\mathcal{P}(\beta, y)} \tag{4.53}$$



Note:

- The above means

$$\begin{pmatrix} \text{Conditional Density} \\ \text{of } x \text{ given } y \end{pmatrix} = \frac{\text{Joint Density of } x \text{ and } y}{\text{Marginal Density of } y} \tag{4.54}$$

  This should be quite intuitive.

- Due to the normalization,

$$\int_{-\infty}^{\infty} \mathcal{P}(\zeta|\beta; x|y) \quad d\zeta = 1 \tag{4.55}$$

  which is what we want for a density function.

- $$\mathcal{P}(\zeta|\beta; x|y) = \mathcal{P}(\zeta, x) \tag{4.56}$$

  if and only if

$$\mathcal{P}(\zeta, \beta; x, y) = \mathcal{P}(\zeta, x)\mathcal{P}(\beta, y) \tag{4.57}$$

  This means that the conditional density is same as the marginal density when and only when x and y are independent.

We are interested in the conditional density, because often some of the random variables are measured while others are not. For a particular trial, if x is not measurable, but y is, we are intersted in knowing $\mathcal{P}(\zeta|\beta; x|y)$ for estimation of x.

Finally, note the distinctions among different density functions:

- $\mathcal{P}(\zeta,\beta;x,y)$: Joint Probability Density of x and y
  represents the probability density of $x=\zeta$ and $y=\beta$ simultaneously.

$$\int_{a_2}^{b_2}\int_{a_1}^{b_1}\mathcal{P}(\zeta,\beta;x,y)d\zeta d\beta=\Pr\{a_1<x\le b_1 \text{ and } a_2<y\le b_2\} \tag{4.58}$$

- $\mathcal{P}(\zeta;x)$: Marginal Probability Density of x
  represents the probability density of $x=\zeta$ NOT knowing what y is.

$$\mathcal{P}(\zeta,x)=\int_{-\infty}^{\infty}\mathcal{P}(\zeta,\beta;x,y)d\beta \tag{4.59}$$

- $\mathcal{P}(\beta;y)$: Marginal Probability Density of y
  represents the probability density of $y=\beta$ NOT knowing what x is.

$$\mathcal{P}(\beta,y)=\int_{-\infty}^{\infty}\mathcal{P}(\zeta,\beta;x,y)d\zeta \tag{4.60}$$

- $\mathcal{P}(\zeta|\beta;x|y)$: Conditional Probability Density of x given y
  represents the probability density of x when $y=\beta$.

$$\mathcal{P}(\zeta|\beta;x|y)=\frac{\mathcal{P}(\zeta,\beta;x,y)}{\mathcal{P}(\beta,y)} \tag{4.61}$$

- $\mathcal{P}(\beta|\zeta;y|x)$: Conditional Probability Density of y given x
  represents the probability density of y when $x=\zeta$.

$$\mathcal{P}(\beta|\zeta;y|x)=\frac{\mathcal{P}(\zeta,\beta;x,y)}{\mathcal{P}(\zeta,x)} \tag{4.62}$$

Baye's Rule:

Note that

$$\mathcal{P}(\zeta|\beta;x|y)=\frac{\mathcal{P}(\zeta,\beta;x,y)}{\mathcal{P}(\beta,y)} \tag{4.63}$$

$$\mathcal{P}(\beta|\zeta;y|x)=\frac{\mathcal{P}(\zeta,\beta;x,y)}{\mathcal{P}(\zeta,x)} \tag{4.64}$$

Hence, we arrive at

$$\mathcal{P}(\zeta|\beta; x|y) = \frac{\mathcal{P}(\beta|\zeta; y|x)\mathcal{P}(\zeta, x)}{\mathcal{P}(\beta, y)} \tag{4.65}$$

The above is known as the *Baye's Rule.* It essentially says

$$\begin{aligned} & \text{(Cond. Prob. of } x \text{ given } y) \times \text{(Marg. Prob. of } y) & (4.66) \\ = \; & \text{(Cond. Prob. of } y \text{ given } x) \times \text{(Marg. Prob. of } x) & (4.67) \end{aligned}$$

Baye's Rule is useful, since in many cases, we are trying to compute $\mathcal{P}(\zeta|\beta; x|y)$ and it's difficult to obtain the expression for it directly, while it may be easy to write down the expression for $\mathcal{P}(\beta|\zeta; y|x)$.

We can define the concepts of conditional expectation and conditional covariance using the conditional density. For instance, the conditional expectation of x given $y = \beta$ is defined as

$$E\{x|y\} \triangleq \int_{-\infty}^{\infty} \zeta \mathcal{P}(\zeta|\beta; x|y) d\zeta \tag{4.68}$$

Conditional variance can be defined as

$$\begin{aligned} \text{Var}\{x|y\} &\triangleq E\{(\zeta - E\{x|y\})^2\} & (4.69) \\ &= \int_{-\infty}^{\infty} (\zeta - E\{x|y\})^2 \mathcal{P}(\zeta|\beta; x|y) d\zeta & (4.70) \end{aligned}$$

Example: Jointly Normally Distributed or Gaussian Variables
Suppose that x and y have the following joint normal densities parametrized by $m_1, m_2, \sigma_1, \sigma_2, \rho$:

$$\begin{aligned} \mathcal{P}(\zeta, \beta; x, y) &= \frac{1}{2\pi\sigma^x\sigma_y(1-\rho^2)^{1/2}} & (4.71) \\ &\times \; \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\zeta - \bar{x}}{\sigma_x}\right)^2 - 2\rho\frac{(\zeta - \bar{x})(\beta - \bar{y})}{\sigma_x\sigma_y} + \left(\frac{\beta - \bar{y}}{\sigma_y}\right)^2\right]\right\} \end{aligned}$$

Some algebra yields

$$\mathcal{P}(\zeta,\beta;x,y) = \underbrace{\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left\{-\frac{1}{2}\left(\frac{\beta-\bar{y}}{\sigma_y}\right)^2\right\}}_{\text{marginal density of } y} \tag{4.72}$$

$$\times \underbrace{\frac{1}{\sqrt{2\pi\sigma_x^2(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{\zeta-\bar{x}-\rho\frac{\sigma_x}{\sigma_y}(\beta-\bar{y})}{\sigma_x\sqrt{1-\rho^2}}\right)^2\right\}}_{\text{conditional density of } x}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left\{-\frac{1}{2}\left(\frac{\zeta-\bar{x}}{\sigma_x}\right)^2\right\}}_{\text{marginal density of } x} \tag{4.73}$$

$$\times \underbrace{\frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{\beta-\bar{y}-\rho\frac{\sigma_y}{\sigma_x}(\zeta-\bar{x})}{\sigma_y\sqrt{1-\rho^2}}\right)^2\right\}}_{\text{conditional density of } y}$$

Hence,

$$\mathcal{P}(\zeta|\beta;x|y) = \frac{1}{\sqrt{2\pi\sigma_x^2(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{\zeta-\bar{x}-\rho\frac{\sigma_x}{\sigma_y}(\beta-\bar{y})}{\sigma_x\sqrt{1-\rho^2}}\right)^2\right\} \tag{4.74}$$

$$\mathcal{P}(\beta|\zeta;y|x) = \frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{\beta-\bar{y}-\rho\frac{\sigma_y}{\sigma_x}(\zeta-\bar{x})}{\sigma_y\sqrt{1-\rho^2}}\right)^2\right\} \tag{4.75}$$

Note that the above conditional densities are normal. For instance, $\mathcal{P}(\zeta|\beta;x|y)$ is a normal density with mean of $\bar{x}+\rho\frac{\sigma_x}{\sigma_y}(\beta-\bar{y})$ and variance of $\sigma_x^2(1-\rho^2)$. So,

$$E\{x|y\} = \bar{x}+\rho\frac{\sigma_x}{\sigma_y}(\beta-\bar{y}) \tag{4.76}$$

$$= \bar{x}+\frac{\rho\sigma_x\sigma_y}{\sigma_y^2}(\beta-\bar{y}) \tag{4.77}$$

$$= E\{x\}+\mathrm{Cov}\{x,y\}\mathrm{Var}^{-1}\{y\}(\beta-\bar{y}) \tag{4.78}$$

Conditional covariance of x given $y = \beta$ is:

$$
\begin{aligned}
E\{(x - E\{x|y\})^2|y\} &= \sigma_x^2(1-\rho^2) &(4.79)\\
&= \sigma_x^2 - \frac{\sigma_x^2\sigma_y^2\rho^2}{\sigma_y^2} &(4.80)\\
&= \sigma_x^2 - (\sigma_x\sigma_y\rho)\frac{1}{\sigma_y^2}(\sigma_x\sigma_y\rho) &(4.81)\\
&= \mathrm{Var}\{x\} - \mathrm{Cov}\{x,y\}\mathrm{Var}^{-1}\{y\}\mathrm{Cov}\{y,x\} &(4.82)
\end{aligned}
$$

Notice that the conditional distribution becomes a point density as $\rho \to 1$, which should be intuitively obvious.

4.2.6 CONDITIONAL PROBABILITY DENSITY: VECTOR CASE

We can extend the concept of conditional probability distribution to the vector case similarly as before.

Let x and y be n and m dimensional random vectors respectively. Then, the conditional density of x given $y = [\beta_1, \cdots, \beta_m]^T$ is defined as

$$
\begin{aligned}
&\mathcal{P}(\zeta_1, \cdots, \zeta_n|\beta_1, \cdots, \beta_m; x_1, \cdots, x_n|y_1, \cdots, y_m)\\
= &\frac{\mathcal{P}(\zeta_1, \cdots, \zeta_n, \beta_1, \cdots, \beta_m; x_1, \cdots, x_n, y_1, \cdots, y_m)}{\mathcal{P}(\beta_1, \cdots, \beta_m; y_1, \cdots, y_m)} \qquad (4.83)
\end{aligned}
$$

Baye's Rule can be stated as

$$
\begin{aligned}
&\mathcal{P}(\zeta_1, \cdots, \zeta_n|\beta_1, \cdots, \beta_m; x_1, \cdots, x_n|y_1, \cdots, y_m) \qquad (4.84)\\
= &\frac{\mathcal{P}(\beta_1, \cdots, \beta_m|\zeta_1, \cdots, \zeta_n; y_1, \cdots, y_m|x_1, \cdots, x_n)\mathcal{P}(\zeta_1, \cdots, \zeta_n; x_1, \cdots, x_n)}{\mathcal{P}(\beta_1, \cdots, \beta_m; y_1, \cdots, y_m)}
\end{aligned}
$$

The conditional expectation and covariance matrix can be defined similarly:

$$E\{x|y\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{bmatrix} \mathcal{P}(\zeta|\beta; x|y) \;\; d\zeta_1, \cdots, d\zeta_n \tag{4.85}$$

$$\text{Cov}\{x|y\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \begin{bmatrix} \zeta_1 - E\{x_1|y\} \\ \vdots \\ \zeta_n - E\{x_n|y\} \end{bmatrix} \begin{bmatrix} \zeta_1 - E\{x_1|y\} \\ \vdots \\ \zeta_n - E\{x_n|y\} \end{bmatrix}^T \mathcal{P}(\zeta|\beta; x|y) \;\; d\zeta_1, \cdots, d\zeta_n \tag{4.86}$$

Example: Gaussian or Jointly Normally Distributed Variables

Let x and y be jointly normally distributed random variable vectors of dimension n and m respectively. Let

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \tag{4.87}$$

The joint distribution takes the form of

$$\mathcal{P}(\zeta, \beta; x, y) = \frac{1}{(2\pi)^{\frac{n+m}{2}} |P_z|^{1/2}} \exp\left\{-\frac{1}{2}(\eta - \bar{z})^T P_z^{-1} (\eta - \bar{z})\right\} \tag{4.88}$$

where

$$\bar{z} = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}; \quad \eta = \begin{bmatrix} \zeta \\ \beta \end{bmatrix} \tag{4.89}$$

$$P_z = \begin{bmatrix} \text{Cov}(x) & \text{Cov}(x, y) \\ \text{Cov}(y, x) & \text{Cov}(y) \end{bmatrix} \tag{4.90}$$

Then, it can be proven that (see Theorem 2.13 in [Jaz70])

$$E\{x|y\} = \bar{x} + \text{Cov}(x, y)\text{Cov}^{-1}(y)(\beta - \bar{y}) \tag{4.91}$$

$$E\{y|x\} = \bar{y} + \text{Cov}(y, x)\text{Cov}^{-1}(x)(\zeta - \bar{x}) \tag{4.92}$$

and

$$\begin{aligned}
\text{Cov}\{x|y\} &\triangleq E\left\{(\zeta - E\{x|y\})(\zeta - E\{x|y\})^T\right\} && (4.93)\\
&= \text{Cov}\{x\} - \text{Cov}\{x,y\}\text{Cov}^{-1}\{y\}\text{Cov}\{y,x\} && (4.94)\\
\text{Cov}\{y|x\} &\triangleq E\left\{(\beta - E\{y|x\})(\beta - E\{y|x\})^T\right\} && (4.95)\\
&= \text{Cov}\{y\} - \text{Cov}\{y,x\}\text{Cov}^{-1}\{x\}\text{Cov}\{x,y\} && (4.96)
\end{aligned}$$

4.3 STATISTICS

4.3.1 PREDICTION

The first problem of statistics is prediction of the outcome of a future trial given a probabilistic model.

> Suppose $\mathcal{P}(x)$, the probability density for random variable x, is given. Predict the outcome of x for a new trial (which is about to occur).

Note that, unless $\mathcal{P}(x)$ is a point distribution, x cannot be predicted exactly.

To do optimal estimation, one must first establish a formal criterion. For example, the most likely value of x is the one that corresponds to the highest density value:

$$\hat{x} = \arg\left[\max_x \mathcal{P}(x)\right]$$

A more commonly used criterion is the following minimum variance estimate:

$$\hat{x} = \arg\left[\min_{\hat{x}} E\left\{\|x - \hat{x}\|_2^2\right\}\right]$$

The solution to the above is $\hat{x} = E\{x\}$.

Exercise: Can you prove the above?

If a related variable y (from the same trial) is given, then one should use $\hat{x} = E\{x|y\}$ instead.

4.3.2 SAMPLE MEAN AND COVARIANCE, PROBABILISTIC MODEL

The other problem of statistics is inferring a probabilistic model from collected data. The simplest of such problems is the following:

> We are given the data for random variable x from N trials. These data are labeled as $x(1), \cdots, x(N)$. Find the probability density function for x.

Often times, a certain density shape (like normal distribution) is assumed to make it a well-posed problem. If a normal density is assumed, the following sample averages can then be used as estimates for the mean and covariance:

$$\hat{\bar{x}} = \frac{1}{N}\sum_{i=1}^{N} x(i)$$

$$\hat{Rx} = \frac{1}{N}\sum_{i=1}^{N} x(i)x^T(i)$$

Note that the above estimates are *consistent* estimates of real mean and covariance $\bar{x}$ and R_x (i.e., they converge to true values as $N \to \infty$).

A slightly more general problem is:

> A random variable vector y is produced according to
>
> $$y = f(\theta, u) + x$$
>
> In the above, x is another random variable vector, u is a *known* deterministic vector (which can change from trial to trial) and θ is

an *unknown* deterministic vector (which is invariant). Given data for y from N trials, find the probability density parameters for x (e.g., $\bar{x}$, R_x) and the unknown deterministic vector θ.

This problem will be discussed later in the regression section.