Module 2

1 Joint Random Variables

To motivate the idea of joint random variables, let's go back to the experiment where we toss a fair coin three times. Remember what the sample space is, $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$. Let $X =$ the number of heads on the first toss, and let $Y =$ the total number of heads. We say these are jointly distributed random variables and follow the joint probability mass function $p(x, y)$ where $p(x, y) = P(X = x, Y = y)$. In this case, $(X, Y) \sim p(x, y)$ where $p(x, y)$ is given by the table below. Prove to yourself that the probabilities given below are correct and make sense.

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

And with this table, we can calculate the probabilities given below:

- $P(X = 0 \;\&\; 0 \leq Y \leq 1) = P(X = 0 \;\&\; Y = 0) + P(X = 0 \;\&\; Y = 1) = \frac{1}{8} + \frac{2}{8} = \frac{3}{8}$
- $P(X = 0) = P(X = 0 \;\&\; Y = 0) + P(X = 0 \;\&\; Y = 1) + P(X = 0 \;\&\; Y = 2) = \frac{1}{8} + \frac{2}{8} + \frac{1}{8} = \frac{1}{2}$
- $P(Y = 1) = P(X = 0 \;\&\; Y = 1) + P(X = 1 \;\&\; Y = 1) = \frac{2}{8} + \frac{1}{8} = \frac{3}{8}$

The example above hopefully illustrates some intuitive results and rules regarding jointly distributed random variables. Below are some general rules about joint random variables. If $(X, Y) \sim p(x, y)$, then

1. $P\{(X, Y) \in A\} = \sum_{(x,y)\in A} p(x, y)$.
2. The marginal distribution of X can be calculated as $p_X(x) = \sum_{\text{all } y} p(x, y)$, and (equivalently) the marginal distribution of Y can be calculated as $p_Y(y) = \sum_{\text{all } x} p(x, y)$.
3. The conditional distribution of X given $Y = y$ is $p(x|y) = p(x, y)/\, p(y)$.
4. Two discrete random variables are independent if $p(x, y) = p(x)p(y)$.

In the example above, we can calculate the conditional probability of x given $y = 1$ as

$$p(x|y = 1) = \begin{cases} P(X = 0 \;\&\; Y = 1)/\, P(Y = 1) = \frac{2}{8}/\,\frac{3}{8} & x = 0 \\ \\ P(X = 1 \;\&\; Y = 1)/\, P(Y = 1) = \frac{1}{8}/\,\frac{3}{8} & x = 1 \end{cases}$$

We can extend these results and methods to continuous random variables. Let's assume that the two random variables X and Y have joint distribution $(X, Y) \sim f(x, y)$. In this case, the rules are

1. $P\{(X, Y) \in A\} = \displaystyle\int_A f(x, y)dxdy$
2. The marginal distribution of X can be calculated as $f_X(x) = \displaystyle\int_{-\infty}^{\infty} f(x, y)dy$, and (equivalently) the marginal distribution of Y can be calculated as $f_Y(y) = \displaystyle\int_{-\infty}^{\infty} f(x, y)dx$.
3. The conditional density of x given y is $f(x|y) = f(x, y)/\, f_Y(y)$.
4. The random variables X and Y are independent if $f(x, y) = f_X(x)f_Y(y)$.

The covariance of two random variables is denoted as $\text{Cov}(X,Y)$ and it is calculated as

$$\text{Cov}(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right].$$

$\text{Cov}(X,Y)$ measures how the two random variables co-vary (obviously). To be more specific, it measures how they co-vary about their means. If $X > \mu_X$ when, on average, $Y > \mu_Y$, then $\text{Cov}(X,Y) > 0$. If $X > \mu_X$ when, on average, $Y < \mu_Y$, then $\text{Cov}(X,Y) < 0$. Below is an example.

Example 1 Let $f(x,y) = k(x-y) \;\; 0 \le y \le x \le 1$. Draw this region out to make sure you understand the limits of integration.

- Find k: We know that k must be the value such that $\displaystyle\int_0^1 \int_y^1 k(x-y)dxdy = 1$. Let's do this integral.

$$\int_0^1 \left\{\frac{kx^2}{2} - kyx\right\}_y^1 dy \Longrightarrow \int_0^1 \left\{\frac{k}{2} - ky - \frac{ky^2}{2} + ky^2\right\} dy = \int_0^1 \left(\frac{ky^2}{2} - ky - \frac{k}{2}\right) dy = \frac{k}{6} \Longrightarrow k = 6$$

- Find $f_X(x)$. $f_X(x) = \displaystyle\int_0^x 6(x-y)dy = 6xy - 3y^2\big|_0^x = 6x^2 - 3x^2 = 3x^2$.
- Find $f_Y(y)$. $f_Y(y) = \displaystyle\int_y^1 6(x-y)dx = 3x^2 - 6xy\big|_y^1 = 3 - 6y - 3y^2 + 6y^2 = 3y^2 - 6y + 3$.
- Find $f(x|y) = \frac{6(x-y)}{3y^2-6y+3}$

1.1 Random Vectors

We've spoken about methods necessary for two random variables, but what about 3, or 4, or n random variables? These are called random vectors and they are the subject of this section. Assume $\mathbf{X}$ is an $n-$dimensional random vector. That is $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$. The mean of the random vector is $E(\mathbf{X}) = [E(X_1), E(X_2), \ldots, E(X_n)]^T = [\mu_1, \mu_2, \ldots, \mu_n]^T = \boldsymbol{\mu}$ and the covariance of the random vector is $\text{Cov}(\mathbf{X}) = \Sigma$. Σ is an $n \times n$ positive definite matrix such that

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_n) \\ \text{Cov}(X_2,X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2,X_n) \\ & & \ddots & \vdots \\ \text{Cov}(X_n,X_1) & \text{Cov}(X_n,X_2) & \cdots & \text{Var}(X_n) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_n^2 \end{bmatrix}.$$

In many cases, a linear transformation is applied to such random vectors. For example, let $\mathbf{Y}$ be an $m \times 1$ random vector such that $\mathbf{Y} = \mathbf{AX}$ where $\mathbf{A}$ is any $m \times n$ matrix. In this case, $E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X})$ and $\text{Cov}(\mathbf{Y}) = \mathbf{A}\Sigma\mathbf{A}^T$.

A commonly used multivariate density is the multivariate normal density. The random vector $\mathbf{X}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ (denoted $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$) if $\mathbf{X}$ has density

$$f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Properties of the multivariate normal random vector $\mathbf{X}$ include:

1. The marginal distribution of X_i is normal. That is $X_i \sim N\left(\mu_i, \sigma_i^2\right)$.
2. For any $i \neq j$, $\begin{pmatrix} X_i \\ X_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \sigma_{i,j} \\ \sigma_{j,i} & \sigma_j^2 \end{pmatrix}\right)$.
3. If $\text{Cov}(X_i, X_j) = 0$, then X_i and X_j are independent, and vice-versa

Example 2 (from Casella & Berger) Assume $X_1, X_2, X_3, X_4 \sim f(x_1, x_2, x_3, x_4)$ where

$$f(x_1,x_2,x_3,x_4) = \frac{3}{4}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right) \qquad 0 \le x_i \le 1, \; i = 1,2,3,4.$$

- $P\left(X_1 < \frac{1}{2}, X_2 \leq \frac{3}{4}, X_4 > \frac{1}{2}\right) = \int_{\frac{1}{2}}^{1}\int_{0}^{1}\int_{0}^{\frac{3}{4}}\int_{0}^{1} \frac{3}{4}\left(x_1^2+x_2^2+x_3^2+x_4^2\right) dx_1 dx_2 dx_3 dx_4 = \frac{3}{256}$. (verify this for yourself)
- The marginal distribution of (X_1, X_2) can be calculated as

$$f(x_1,x_2) = \int_0^1\int_0^1 \frac{3}{4}\left(x_1^2+x_2^2+x_3^2+x_4^2\right) dx_3 dx_4 = \frac{3}{4}\left(x_1^2+x_2^2\right) + \frac{1}{2}.$$

  Again....calculate this integral yourself and make sure you get the same thing.

Example 3 Assume that $\mathbf{X}$ is a multivariate distribution such that

$$\mathbf{X} \sim N\left(\left(\begin{array}{c} -3 \\ 1 \\ 4 \end{array}\right), \left(\begin{array}{ccc} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{array}\right)\right).$$

Are $\frac{X_1+X_2}{2}$ and X_3 independent? To answer this question, let's construct the random vector $\mathbf{Y}$, where

$$\mathbf{Y} = \left(\begin{array}{c} \frac{X_1+X_2}{2} \\ \\ X_3 \end{array}\right)$$

by identifying the appropriate transformation, then calculate the covariance matrix of $\mathbf{Y}$, and see if the off-diagonal elements are 0. The appropriate transformation is $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A} = \left(\begin{array}{ccc} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{array}\right)$. Remember that $\mathbf{Y} \sim N\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}'\right)$. $\mathbf{A}\boldsymbol{\mu} = \left(\begin{array}{c} -1 \\ 4 \end{array}\right)$ and $\mathbf{A}\Sigma\mathbf{A}' = \left(\begin{array}{cc} 0.5 & 0 \\ 0 & 2 \end{array}\right)$. Since the diagonal elements are 0, the variables $Y_1 = \frac{X_1+X_2}{2}$ and $Y_2 = X_3$ are independent.

2 Limit Theorems

The limit theorems we discuss show, mostly, how sample averages behave the larger the sample sizes become. The two we discuss here are the Law of Large Numbers and the Central Limit Theorem.

Law of Large Numbers : Generally speaking, this states that the more random variables you average over in your sample, the more likely you are to get closer to the population mean. The actual theorem is given below.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Then for any $\epsilon > 0$,

$$P\left\{\left(|\overline{X}_n - \mu| > \epsilon\right)\right\} \longrightarrow 0.$$

So if you pick an ϵ to be really small (say $\epsilon = .000001$) this theorem states that the probability that your sample average, $\overline{X}$, will **not** be within .000001 of the true population average, μ, will get closer and closer to 0 the more numbers you average over.

Central Limit Theorem : Let $X_1, X_2, \ldots, X_n$ be i.i.d. from some distribution $f_X(x)$. Note that $f_X(x)$ doesn't have to be normal. It doesn't even have to be continuous . Also assume that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Then let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\text{Var}(\overline{X}) = \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}(X_j) = \frac{\sigma^2}{n}$. If all of this is true, then

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow N(0,1).$$

A fancy way to say this is that $\overline{X} - \mu \big/ \frac{\sigma}{\sqrt{n}}$ converges in distribution to a standard normal random variable.

$$\lim_{n\to\infty}\left[P\left\{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\le x\right\}\right]=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2}\right\}dz=\Phi(x),$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal random distribution evaluated at x.

The Central Limit Theorem states that if you average over enough random variables, then the *scaled* distribution of the average is normal regardless of how the random variables are originally distributed. It's critical to understand what I mean by "scaled." In the Law of Large numbers, we stated that $\overline{X}-\mu$ got closer and closer to 0 the more numbers you average over, but with the Central Limit Theorem, keep in mind that we're looking at the same value, $\overline{X}-\mu$ scaled up by $\frac{\sqrt{n}}{\sigma}$. So the larger the value of n, the larger the quantity $(\overline{X}-\mu)$ gets scaled. So $\left(\overline{X}-\mu\right)/\frac{\sigma}{\sqrt{n}}$ doesn't converge to a number, it converges to a distribution (the standard normal distribution).

Example 4 (from Rice's book) A certain type of particle is emitted at a rate of 900 per hour. What is the probability that more than 950 particles will be emitted in a given hour if the counts form a Poisson process?

To answer this, let X be a Poisson random variable with mean and variance 900 (for a Poisson random variable, this means that $\lambda = 900$). We will use the central limit theorem to approximate $P(X > 950)$.

$$P(X>950)=P\left(\frac{X-900}{\sqrt{900}}>\frac{950-900}{\sqrt{900}}\right)\approx 1-P\left(Z>\frac{50}{30}\right)=.04779,$$

where Z is a standard normal random variable.

Example 5 (from Casella's Book). A negative binomial random variable counts the number of failures before the r^{th} success in a sequence of success/no-success trials. Assume the probability of a success in one trial is p. If this is the case, and $X =$ the number of failures before the r^{th} success, then

$$p_X(x)=\left(\begin{array}{c} r+x-1 \\ x \end{array}\right)p^r(1-p)^x \quad x=0,1,\ldots.$$

For a negative binomial random variable, $E(X) = r(1-p)/p$ and $\text{Var}(X) = r(1-p)/p^2$. Now suppose that $X_1, X_2, \ldots, X_n$ are a random sample fro ma negative binomial distribution with $r = 10$, $p = \frac{1}{2}$, and $n = 30$. What is $P\left(\overline{X}\le 11\right)$? You can calculate this exactly (using the distribution of the negative binomial random variable), but this is not a trivial calculation at all. It's much easier to do it using the Central Limit Theorem. The Central Limit Theorem tells us that, approximately

$$\frac{\sqrt{n}\left(\overline{X}-r(1-p)/p\right)}{\sqrt{r(1-p)/p^2}}\sim N(0,1).$$

And with this we get that

$$P\left(\overline{X}\le 11\right)\approx P\left(\frac{\sqrt{30}(\overline{X}-10)}{\sqrt{20}}\le\frac{\sqrt{30}(11-10)}{\sqrt{20}}\right)=P\left(Z\le 1.2247\right)=.89.$$

3 Maximum Likelihood Estimation

The theory discussed about joint random variables can be used to explain one of the most common methods of parameter estimation, maximum likelihood estimation. Maximum likelihood estimation uses the joint probability density function to estimate a parameter of interest.

Assume you collect n random variables, $X_1, X_2, \ldots, X_n$, that are independent and identically distributed with probability density function $f(x|\theta)$, where θ is some unknown parameter. Since they are independent and identically distributed, the distribution of the data is

$$f(x_1,x_2,\ldots,x_n|\theta)=f(\mathbf{x}|\theta)=\prod_{j=1}^{n}f(x_j|\theta).$$

The principle of the maximum likelihood estimator is to find the value of θ which maximizes the probability of the observed data. Finding the maximum likelihood estimator thus involves maximizing the likelihood function, $L(\theta|\mathbf{x})$, where

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta),$$

and the observed values of $x_1, x_2, \ldots, x_n$ are plugged in for $\mathbf{x}$. The maximum likelihood estimate of θ, $\hat{\theta}_{\text{MLE}}$, is that value which maximizes $L\left(\theta|\mathbf{x}\right)$, i.e.,

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta}\left\{L\left(\theta|\mathbf{x}\right)\right\}.$$

The maximum likelihood estimator is typically found by maximizing the log-likelihood $l(\theta) = \log\left[L(\theta|\mathbf{x})\right]$ (maximizing $l(\theta)$ is equivalent to maximizing $L(\theta)$ since the *log* function is monotonically increasing and one-to-one.)

Example 6 . Assume you collect/observe observations $X_1, X_2, \ldots, X_n$ which come from a Poisson distribution, i.e., $f(x_j|\theta) = \exp\left(-\theta\right)\theta^{x_j} / \, x_j!$. What is the maximum likelihood estimator for θ?

$$\begin{aligned}
L(\theta) &= \prod_{j=1}^{n} \frac{\exp(-\theta)\theta^{-x_j}}{x_j!} = \frac{\exp(-n\theta)\cdot\theta^{\sum_j x_j}}{\prod_{j=1}^{n} x_j!} \\
\Longrightarrow l(\theta) &= -n\theta + \log(\theta)\sum_j x_j - \log\left(\prod_{j=1}^{n} x_j!\right) \\
\Longrightarrow \frac{\partial l}{\partial\theta} &= -n + \frac{\sum_j x_j}{\theta}.
\end{aligned}$$

Setting the above equation equal to 0, $\hat{\theta}_{\text{MLE}} = \sum_j x_j/n$. This is no surprise. The maximum likelihood estimate for the rate parameter of a Poisson process is the average.

Example 7 Assume you randomly sample n values of X, $X_1, X_2, \ldots, X_n$ which are i.i.d. and come from a normal distribution with mean μ and variance σ^2. What are the maximum likelihood estimates of μ and σ^2? In this case,

$$\begin{aligned}
f(x_1, x_2, \ldots, x_n|\mu, \sigma^2) &= \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right\} \\
\Longrightarrow \text{Lik}\left(\mu, \sigma^2\right) &= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_j \left(x_j - \mu\right)^2\right\} \\
l(\mu, \sigma^2) &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_j \left(x_j - \mu\right)^2.
\end{aligned}$$

Taking the partial derivative of this with respect to μ and setting it to 0, we get

$$\begin{aligned}
\frac{\partial l}{\partial\mu} &= -\frac{1}{2\sigma^2}\sum_j \frac{\partial(x_j - \mu)^2}{\partial\mu} = \frac{1}{\sigma^2}\sum_j \left(x_j - \mu\right) = 0 \\
\Longrightarrow & \sum_j \left(x_j - \mu\right) = 0 \Longrightarrow \hat{\mu}_{\text{MLE}} = \overline{X}.
\end{aligned}$$

Taking the partial derivative with respect to σ and setting it equal to 0, we get

$$\begin{aligned}
\frac{\partial l}{\partial\sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_j \left(x_j - \mu\right)^2 = 0 \\
\Longrightarrow n\hat{\sigma}^2_{\text{MLE}} &= \sum_j \left(x_j - \hat{\mu}_{\text{MLE}}\right)^2 \\
\Longrightarrow \hat{\sigma}^2_{\text{MLE}} &= \frac{1}{n}\sum_j (x_j - \overline{x})^2
\end{aligned}$$