LECTURE 4: COMMON FAMILIES OF DISTRIBUTIONS

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
SEPTEMBER 30, 2020

1. Some important inequalities

1.1. Jensen's Inequality

A function $g(x)$ is convex if and only if $\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y)$ for $0 < \lambda < 1$. Graphically, a straight line connecting any two points of the convex function lies above the function.

Jensen's Inequality: For any random variable X, if $g(X)$ is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.

For example: take $g(X) = X^2$, then $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, which implies that $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$.

1.2. Markov and Chebychev's Inequalities

Chebyshev's inequality. Let X be a random variable and $g(X)$ be a non-negative function. Then for any $\epsilon > 0$,

$$P(g(X) \geq \epsilon) \leq \frac{\mathbb{E}[g(X)]}{\epsilon}$$

Proof:

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\ &\geq \int_{x:g(x)\geq\epsilon}^{\infty} g(x)f(x)\,dx \\ &\geq \int_{x:g(x)\geq\epsilon}^{\infty} \epsilon f(x)\,dx \\ &= \epsilon P(g(X)\geq\epsilon)\end{aligned}$$

Markov's inequality is just $P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$.

Now let $g(x) = \frac{(x-\mu)^2}{\sigma^2} \geq 0$, where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}(X)$. Note that g is always positive. By the Chebyshev's inequality,

$$\begin{aligned}P(g(X) \geq \epsilon^2) &\leq \frac{\mathbb{E}[g(X)]}{\epsilon^2} \\ P(\frac{(X-\mu)^2}{\sigma^2} \geq \epsilon^2) &\leq \frac{\mathbb{E}[\frac{(X-\mu)^2}{\sigma^2}]}{\epsilon^2} \\ P(\frac{(X-\mu)^2}{\sigma^2} \geq \epsilon^2) &\leq \frac{1}{\epsilon^2} \\ P(|X-\mu| \geq \epsilon\sigma) &\leq \frac{1}{\epsilon^2} \qquad (1)\end{aligned}$$

If we take $\epsilon = 2$, then $P(|x-\mu| \geq 2\sigma) \leq 0.25$ or $P(|x-\mu| < 2\sigma) > 0.75$. That is, there is at least 75% chance that a random variable (any random variable!) will be within 2 standard deviation of its mean.

In general, the Chebyshev's inequality can be used to show that as $\text{Var}(X_n) \to 0$, $P(|X_n - \mu| \geq \epsilon) \to 0$, by taking $g(X) = (X-\mu)^2$.

As such, Chebyshev's inequality can be used to prove the Weak Law of Large Numbers. Let $X_1, \ldots, X_n$ be n independent random variables, each with the same density f. Define the sample mean as the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that $\bar{X}$ has expectation $\mathbb{E}[X] \equiv \mu$, and variance $\frac{\text{Var}(X)}{n} \equiv \frac{\sigma^2}{n}$.

By the inequality in (1), we have:

$$P(|\bar{X} - \mu| \geq \epsilon\frac{\sigma}{\sqrt{n}}) \leq \frac{1}{\epsilon^2}$$

Now if we let $\epsilon = v\frac{\sqrt{n}}{\sigma}$,

$$P(|\bar{X} - \mu| \geq v) \leq \frac{\sigma^2}{nv^2}$$

Therefore, as $n \to \infty$, $P(|\bar{X} - \mu| \geq v) = 0$ for any $v > 0$, which is the Weak Law of Large Numbers.

2. Common families of statistical distributions

2.1. Multivariate Normal

We are already familiar with the one-dimensional Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, which has the pdf $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.

The k-dimensional Gaussian random variable is described as:

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

$\boldsymbol{X}$ is a k-dimensional random vector. $\boldsymbol{\mu}$ is a k-dimensional vector, Σ is a k-by-k symmetric matrix called the variance-covariance matrix. A matrix Σ is symmetric if $\Sigma^T = \Sigma$, as such Σ has $k + (k^2 - k)/2 = (k^2 + k)/2$ number of parameters. Intuitively, k diagonal terms of Σ describe the variances of each individual random variable, and $(k^2 - k)/2$ off-diagonal terms of Σ describe the pairwise correlations between each of the variable.[1]

Therefore a k-dimensional Gaussian variable has $\frac{3k+k^2}{2}$ number of parameters. For example, a 2-dimensional multivariate Gaussian has 5 parameters.

For the bivariate Normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right]$$

The pdf of (X, Y) is:

(2)

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right] \right)$$

Check that the marginal pdf of X is just the univariate Normal pdf:

[1] In additional Σ also has to be positive semi-definite, that is, $\boldsymbol{x}^T\Sigma\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^k$.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

Hence, the moments of (X, Y) are described by the parameters of the pdf, i.e. $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\text{Var}(X) = \sigma_X^2$, $\text{Var}(Y) = \sigma_Y^2$.

In addition, we can compute $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ from the joint pdf, which turns out to be $\rho\sigma_X\sigma_Y$. As such the correlation of X and Y is just ρ.

If we set $\rho = 0$, i.e. zero correlation between X and Y, then:

$$f(x, y) = f_X(x) f_Y(y)$$

Hence, for Multivariate Normals, zero correlation implies independence. Also, if X and Y are independent with univariate Normal distributions, then (X, Y) trivially has a bivariate Normal distribution.

However, if two random variables X and Y are univariate Normals, it is not true that (X, Y) has a bivariate Normal distribution. Can you work out an example?

The conditional distribution of Y given $X = x$ is:

$$(Y|X = x) \sim \mathcal{N}\Big(\mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(x - \mathbb{E}[X]), (1 - \rho^2)\sigma_Y^2\Big) \tag{3}$$

This implies that the conditional expectation of Y given X is:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}[X])$$

It is a **linear** function of X and has a normal pdf. The fact that $\mathbb{E}[Y|X]$ is linear in X is a very subtle but powerful result. It means that the best prediction of Y using X is some linear function of X. That is, we can't do better than a linear regression of Y on X if (Y, X) is a bivariate Normal.

The conditional variance of Y given X is $\text{Var}[Y|X] = (1 - \rho^2)\sigma_Y^2$, which does not depend on X.

In general, the joint density of a k-th dimensional multivariate Normal distribution is:

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

Where $\boldsymbol{\Sigma}$ is a k-by-k variance-covariance matrix of $\boldsymbol{X}$, and $\boldsymbol{\mu}$ is a k-dimensional vector. We say that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2.2. Sampling from a multivariate Normal

To sample from a scalar random variable, we learned how to use the probability integral transform. We can use the conditional distribution to sample from a multivariate distribution. For instance, to sample from a bivariate Normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right]$$

First, we sample from the marginal of X, which is just $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$.

Recall the conditional distribution of Y given $X = x$ is:

$$(Y|X=x) \sim \mathcal{N}\Big(\mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(x - \mathbb{E}[X]), (1-\rho^2)\sigma_Y^2\Big)$$

For every draw of x_i from the marginal distribution $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, we then sample y_i from $Y|X = x_i$. The sample $(x_i, y_i)_{i=1}^n$ will be a valid sample from the the bivariate Normal distribution.

This approach is called *Gibbs Sampling*.[2] More generally, to sample from a trivariate distribution $f(x, y, z)$, we first draw x_i from the marginal of X, then draw y_i from $Y|X = x_i$, then finally, draw z_i from $Z|Y = y_i, X = x_i$. Now, the density of $Z|Y, X$ can be derived as $f(x, y, z)/f(x, y)$.

Let's try to implement Gibbs sampling using R or Python.

[2] More specifically, this is the Collapsed Gibbs Sampling

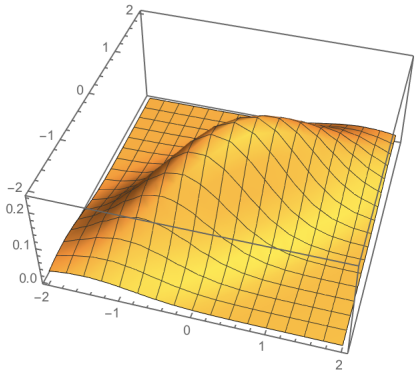
2.3. Example

For example, let $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$ in the joint pdf of Bivariate Normal (Equation 4). The location parameters μ_X and μ_Y merely shift the center of the distribution around. Then we have:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) \tag{4}$$

Visualize this joint pdf at various values of ρ as in Figure 1.

```
Plot3D[ReplaceAll[1/(2 π √(1 - ρ^2)) Exp[-(x^2 + y^2 - 2 ρ x y)/(2 (1 - ρ^2))], ρ → 0.75], {x, -2, 2},
 {y, -2, 2}]
```




```
Plot3D[ReplaceAll[1/(2 π √(1 - ρ^2)) Exp[-(x^2 + y^2 - 2 ρ x y)/(2 (1 - ρ^2))], ρ → 0], {x, -2, 2}, {y, -2, 2}]
```

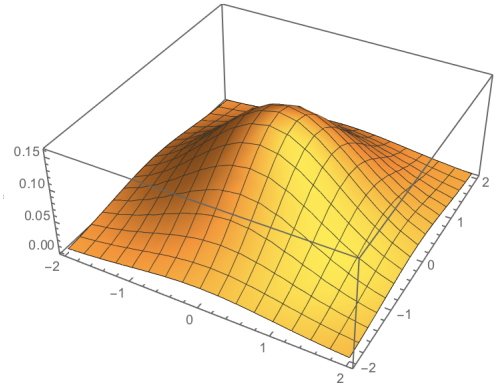


FIGURE 1

Now we derive the conditional distribution of X given Y.

$$\begin{aligned}
f(x|y) &= \frac{f(x,y)}{f(y)} \\
&= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right)}{\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}+\frac{y^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left(-\frac{x^2+\rho^2y^2-2\rho xy}{2(1-\rho^2)}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right)
\end{aligned}$$

The last line is the pdf of a univariate Normal distribution with mean ρy and variance $1-\rho^2$. Therefore,

$$X|Y=y \sim N(\rho y, 1-\rho^2)$$

Which is consistent with Equation 3. We can also verify this using the FullSimplify command in Mathematica.

Let's also check whether the joint pdf integrates to the marginal pdfs (which can be evaluated analytically by completing the squares):

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{-y^2}{2(1-\rho^2}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2-2\rho xy}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{\rho^2y^2-y^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\
&= \frac{\exp\left(\frac{-y^2}{2}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}
\end{aligned}
$$

2.4. Beta distribution

Beta distribution is used to model random variables that lie within the unit interval $[0, 1]$. For example, if we want to model fractions or probabilities, then we use the Beta distribution.

The Beta distribution is controlled by two parameters $\alpha > 0$ and $\beta > 0$, that is, $X \sim Beta(\alpha, \beta)$.

The pdf is $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$ for $x \in [0, 1]$. The constant of proportionality is $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where Γ is the Gamma function.

The Beta distribution is a very flexible class of distributions that can generate distributions that are positively or negatively skewed, varying modes and medians. The mean is given by $\frac{\alpha}{\alpha+\beta}$.

The Dirichlet distribution generalizes the Beta distribution to multiple dimensions:

$$f\left(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K\right) \propto \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

Where $\{x_k\}_{k=1}^{k=K}$ belong to the standard $K-1$ simplex, or in other words: $\sum_{i=1}^{K} x_i = 1$ and $x_i \geq 0$ for all $i \in \{1, \ldots, K\}$.

The normalizing constant is the multivariate beta function, which can be expressed in terms of the gamma function

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

2.5. Gamma distribution

The Gamma distribution is used to model random variables that takes positive values. It is a general form of the Exponential distribution. It is also used in Bayesian statistics as conjugate priors. Moreover, it is used in the frequentist setting for hypothesis testing.

$X \sim \text{Gamma}(\alpha, 1) \equiv \Gamma(\alpha, 1)$ if X has the density:

$$f(x) \propto x^{\alpha-1} e^{-x}$$

Where the constant of proportionality is $\Gamma(\alpha)$, the Gamma function, and the density is defined for $\alpha, x > 0$. By letting $Y = \beta X$, we obtain $Y \sim \text{Gamma}(\alpha, \beta)$, which has the pdf $f(y) = \frac{\beta^{-\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}}}{\Gamma(\alpha)}$. α is the shape parameter, while β is the scale parameter. The Gamma distribution gives the duration it takes until α number of successes, where the rate of a success is given by $\frac{1}{\beta}$.

The Gamma function is an interesting function. It is defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. The Gamma function satisfies the following recurrence relation: $\Gamma(z) = (z-1)\Gamma(z-1)$. As such, when z is an integer, $\Gamma(z) = (z-1)!$. We can think of the Gamma function as an extension of the factorial function to non-negative real numbers. For non-integers $z > 1$, it must be that $\Gamma(z) = (z-1)(z-2)\ldots\delta\Gamma(\delta)$ where $0 < \delta < 1$.

If $X \sim \text{Gamma}(1, \lambda)$, then X has an exponential distribution with rate parameter $\frac{1}{\lambda}$. If $X \sim \text{Gamma}(v/2, 1/2)$, then X is identical to $\chi(v)$, the chi-squared distribution with v degrees of freedom.

Visualize the pdf of the Gamma distribution using the Mathematica command `Plot[PDF[GammaDistribution[2, 2], x], {x, 0, 20}]`.

2.6. Bernoulli and Binomial Distribution

X is a Bernoulli distribution with parameter p if $X = 1$ with probability p, and $X = 0$ with probability $1 - p$.

Let $X_1, X_2, \ldots, X_n$ be n independent Bernoulli random variables with parameter p. $Y = \sum_{i=1}^{n} X_i$ is a Binomial distribution with parameters (n, p).

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Y is the number of successes in n independent trials, where p is the probability of a success in a trial. The mean of Y is np, and the variance of Y is $np(1-p)$, can you prove this?

3. A note on truncated random variables

Consider the random variables (X, Y) which are joint uniformly distributed on the unit square. That is, $f(x, y) = 1$ for $0 < x < 1$ and $0 < y < 1$.

(1) Show that $\mathbb{E}[X|Y > X] = \frac{1}{3}$. Note that $Y > X$ is an event, not a random variable. As such, the formula to compute this conditional expectation is $\mathbb{E}[X|Y > X] = \frac{\mathbb{E}[X \mathbb{1}_{\{Y>X\}}]}{P(Y>X)}$, and NOT $\mathbb{E}[Y|X = x] = \int y \frac{f(x,y)}{f(x)}\, dy$, which is the formula when conditioning on a random variable

(2) What is $\mathbb{E}[X|X > a]$? Again $X > a$ is an event, not a random variable. So we have:

$$\begin{aligned}
\mathbb{E}[X|X > a] &= \frac{\mathbb{E}[X \mathbb{1}_{\{X>a\}}]}{P(X > a)} \\
&= \frac{\int_a^\infty x f_X(x)\, dx}{1 - F_X(a)} \\
&= \frac{\int_a^1 x\, dx}{1 - a} \\
&= \frac{\int_a^1 x\, dx}{1 - a} \\
&= \frac{a+1}{2} \quad \text{for } a \in (0, 1)
\end{aligned}$$

For instance, if $X \sim \mathcal{N}(0, \sigma^2)$, then we can use the above formula to show that $\mathbb{E}[X|X > 0] \approx 0.7978\sigma$.

As a side-note, another interpretation of the Law of Iterated Expectation is:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x)\, dx$$