Lecture 1
Distributions and Normal Random Variables

1 Random variables

1.1 Basic Definitions

Given a random variable X, we define a *cumulative distribution function (cdf)*, $F_X : \mathbb{R} \to [0,1]$, such that $F_X(t) = P\{X \le t\}$ for all $t \in \mathbb{R}$. Here $P\{X \le t\}$ denotes the probability that $X \le t$. To emphasize that random variable X has cdf F_X, we write $X \sim F_X$. Note that $F_X(t)$ is a nondecreasing function of t.

There are 3 types of random variables: discrete, continuous, and mixed.

Discrete random variable, X, is characterized by a list of possible values, $\mathcal{X} = \{x_1, ..., x_n\}$, and their probabilities, $p = \{p_1, ..., p_n\}$, where p_i denotes the probability that X will take value x_i, i.e. $p_i = P\{X = x_i\}$ for all $i = 1, ..., n$. Note that $p_1 + ... + p_n = 1$ and $p_i \ge 0$ for all $i = 1, ..., n$ by definition of probability. Then the cdf of X is given by $F_X(t) = \sum_{j=1,...,n:\, x_j \le t} p_j$.

Continuous random variable, Y, is characterized by its probability density function (pdf), $f_Y : \mathbb{R} \to \mathbb{R}$, such that $P\{a < Y \le b\} = \int_a^b f_Y(s)ds$. Note that $\int_{-\infty}^{+\infty} f_Y(s)ds = 1$ and $f_Y(s) \ge 0$ for all $s \in \mathbb{R}$ by definition of probability. Then the cdf of Y is given by $F_Y(t) = \int_{-\infty}^{t} f_Y(s)ds$. By the Fundamental Theorem of Calculus, $f_Y(t) = dF_Y(t)/dt$.

A random variable is referred to as *mixed* if it is not discrete and not continuous.

If cdf F of some random variable X is strictly increasing and continuous then it has inverse, $q(x) = F^{-1}(x)$. It is defined for all $x \in (0,1)$. Note that

$$P\{X \le q(x)\} = P\{X \le F^{-1}(x)\} = F(F^{-1}(x)) = x$$

for all $x \in (0,1)$. Therefore $q(x)$ is called the *x-quantile* of X. It is such a number that random variable X takes a value smaller or equal to this number with probability x. If F is not strictly increasing or continuous, then we define $q(x)$ as a generalized inverse of F, i.e. $q(x) = \inf\{t \in \mathbb{R} : F(t) \ge x\}$ for all $x \in (0,1)$. In other words, $q(x)$ is a number such that $F(q(x) + \varepsilon) \ge x$ and $F(q(x) - \varepsilon) < x$ for any $\varepsilon > 0$. As an exercise, check that $P\{X \le q(x)\} \ge x$.

1.2 Functions of Random Variables

Suppose we have random variable X and function $g : \mathbb{R} \to \mathbb{R}$. Then we can define another random variable $Y = g(X)$. The cdf of Y can be calculated as follows

$$F_Y(t) = P\{Y \le t\} = P\{g(X) \le t\} = P\{X \in g^{-1}(-\infty, t]\},$$

where g^{-1} may be the set-valued inverse of g. The set $g^{-1}(-\infty, t]$ consists of all $s \in \mathbb{R}$ such that $g(s) \in (-\infty, t]$, i.e. $g(s) \le t$. If g is strictly increasing and continuously differentiable then it has strictly increasing and continuously differentiable inverse g^{-1} defined on set $g(\mathbb{R})$. In this case $P\{X \in g^{-1}(-\infty, t]\} = P\{X \le g^{-1}(t)\} = F_X(g^{-1}(t))$ for all $t \in g(\mathbb{R})$. If, in addition, X is a continuous random variable, then

$$f_Y(t) = \frac{dF_Y(t)}{dt} = \frac{dF_X(g^{-1}(t))}{dt} = \left(\frac{dF_X(s)}{ds}\right)_{s=g^{-1}(t)} \left(\frac{dg(s)}{ds}\right)^{-1}_{s=g^{-1}(t)} = f_X(g^{-1}(t)) \left(\frac{dg(s)}{ds}\right)^{-1}_{s=g^{-1}(t)}$$

for all $t \in g(\mathbb{R})$. If $t \notin g(\mathbb{R})$, then $f_Y(t) = 0$.

One important type of function is a linear transformation. If $Y = X - a$ for some $a \in \mathbb{R}$, then

$$F_Y(t) = P\{Y \le t\} = P\{X - a \le t\} = P\{X \le t + a\} = F_X(t + a).$$

In particular, if X is continuous, then Y is also continuous with $f_Y(t) = f_X(t + a)$. If $Y = bX$ with $b > 0$, then

$$F_Y(t) = P\{bX \le t\} = P\{X \le t/b\} = F_X(t/b).$$

In particular, if X is continuous, then Y is also continuous with $f_Y(t) = f_X(t/b)/b$.

1.3 Expected Value

Informally, the expected value of some random variable can be interpreted as its average. Formally, if X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is some function, then, by definition,

$$E[g(X)] = \sum_i g(x_i)p_i$$

for discrete random variables and

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)dx$$

for continuous random variables.

Expected values for some functions g deserve special names:

- mean: $g(x) = x$, $E[X]$
- second moment: $g(x) = x^2$, $E[X^2]$
- variance: $g(x) = (x - E[X])^2$, $E[(X - E[X])^2]$

- k-th moment: $g(x) = x^k$, $E[X^k]$
- k-th central moment: $E[(X - EX)^k]$

The variance of random variable X is commonly denoted by $V(X)$.

1.3.1 Properties of expectation

1) For any constant a (non-random), $E[a] = a$.

2) The most useful property of an expectation is its linearity: if X and Y are two random variables and a and b are two constants, then $E[aX + bY] = aE[X] + bE[Y]$.

3)If X is a random variable, then $V(X) = E[X^2] - (E[X])^2$. Indeed,

$$\begin{aligned} V(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - E[2XE[X]] + E[(E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2. \end{aligned}$$

4) If X is a random variable and a is a constant, then $V(aX) = a^2V(X)$ and $V(X + a) = V(X)$.

1.4 Examples of Random Variables

Discrete random variables:

- *Bernoulli*(p): random variable X has Bernoully(p) distribution if it takes values from $\mathcal{X} = \{0, 1\}$, $P\{X = 0\} = 1-p$ and $P\{X = 1\} = p$. Its expectation $E[X] = 1 \cdot p + 0 \cdot (1-p) = p$. Its second moment $E[X^2] = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$. Thus, its variance $V(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$. Notation: $X \sim \text{Bernoulli}(p)$.
- *Poisson*(λ): random variable X has a Poisson(λ) distribution if it takes values from $\mathcal{X} = \{0, 1, 2, ...\}$ and $P\{X = j\} = e^{-\lambda}\lambda^j/j!$. As an exercise, check that $E[X] = \lambda$ and $V(X) = \lambda$. Notation: $X \sim \text{Poisson}(\lambda\}$.

Continuous random variables:

- *Uniform*(a, b): random variable X has a Uniform(a, b) distribution if its density $f_X(x) = 1/(b - a)$ for $x \in (a, b)$ and $f_X(x) = 0$ otherwise. Notation: $X \sim U(a, b)$.
- *Normal*(μ, σ^2): random variable X has a Normal(μ, σ^2) distribution if its density $f_X(x) = \exp(-(x - \mu)^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$ for all $x \in \mathbb{R}$. Its expectation $E[X] = \mu$ and its variance $V(X) = \sigma^2$. Notation: $X \sim N(\mu, \sigma^2)$. As an exercise, check that if $X \sim N(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma \sim N(0, 1)$. Y is said to have a standard normal distribution. It is known that the cdf of $N(\mu, \sigma^2)$ is not analytical, i.e. it can not be written as a composition of simple functions. However, there exist tables that give

its approximate values. The cdf of a standard normal distribution is commonly denoted by Φ, i.e. if $Y \sim N(0,1)$, then $F_Y(t) = P\{Y \leq t\} = \Phi(t)$.

2 Bivariate (multivariate) distributions

2.1 Joint, marginal, conditional

If X and Y are two random variables, then $F_{X,Y}(x,y) = P\{X \leq x, Y \leq y\}$ denotes their joint cdf. X and Y are said to have *joint* pdf $f_{X,Y}$ if $f_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$ and $F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dtds$. Under some mild regularity conditions (for example, if $f_{X,Y}(x,y)$ is continuous),

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

From the joint pdf $f_{X,Y}$ one can calculate the pdf of, say, X. Indeed,

$$F_X(x) = P\{X \leq x\} = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(s,t)dtds$$

Therefore $f_X(s) = \int_{-\infty}^{+\infty} f(s,t)dt$. The pdf of X is called *marginal* to emphasize that it comes from a joint pdf of X and Y.

If X and Y have a joint pdf, then we can define a *conditional* pdf of Y given $X = x$ (for x such that $f_X(x) > 0$): $f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)$. Conditional probability is a full characterization of how Y is distributed for any given given $X = x$. The probability that $Y \in A$ for some set A given that $X = x$ can be calculated as $P\{Y \in A|X = x\} = \int_A f_{Y|X}(y|x)dy$. In a similar manner we can calculate the conditional expectation of Y given $X = x$: $E[Y|X = x] = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)dy$. As an exercise, think how we can define the conditional distribution of Y given $X = x$ if X and Y are discrete random variables.

Two extremely useful properties of a conditional expectation are: for any random variables X and Y,

- $E[f(X)Y|X = x] = f(x)E[Y|X = x]$;
- *the law of iterated expectations*: $E[E[Y|X = x]] = E[Y]$.

2.2 Independence

Random variables X and Y are said to be *independent* if $f_{Y|X}(y|x) = f_Y(y)$ for all $x \in \mathbb{R}$, i.e. if the marginal pdf of Y equals conditional pdf Y given $X = x$ for all $x \in \mathbb{R}$. Note that $f_{Y|X}(y|x) = f_Y(y)$ if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. If X and Y are independent, then $g(X)$ and $f(Y)$ are also independent for any functions $g : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$. In addition, if X and Y are independent, then $E[XY] = E[X]E[Y]$.

Indeed,

$$
\begin{aligned}
E[XY] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{X,Y}(x,y)dxdy \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_X(x)f_Y(y)dxdy \\
&= \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy \\
&= E[X]E[Y]
\end{aligned}
$$

2.3 Covariance

For any two random variables X and Y we can define covariance as

$$\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])].$$

As an exercise, check that $\mathrm{cov}(X,Y) = E[XY] - E[X]E[Y]$.

Covariances have several useful properties:

1. $\mathrm{cov}(X,Y) = 0$ whenever X and Y are independent
2. $\mathrm{cov}(aX,bY) = ab\mathrm{cov}(X,Y)$ for any random variables X and Y and any constants a and b
3. $\mathrm{cov}(X+a,Y) = \mathrm{cov}(X,Y)$ for any random variables X and Y and any constant a
4. $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$ for any random variables X and Y
5. $|\mathrm{cov}(X,Y)| \leq \sqrt{V(X)V(Y)}$ for any random variables X and Y
6. $V(X+Y) = V(X) + V(Y) + 2\mathrm{cov}(X,Y)$ for any random variables X and Y
7. $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i)$ whenever $X_1, ..., X_n$ are independent

To prove property 5, consider random variable $X - aY$ with $a = \mathrm{cov}(X,Y)/V(Y)$. On the one hand, its variance $V(X - aY) \geq 0$. On the other hand,

$$
\begin{aligned}
V(X - aY) &= V(X) - 2a\mathrm{cov}(X,Y) + a^2V(Y) \\
&= V(X) - 2(\mathrm{cov}(X,Y))^2/V(Y) + (\mathrm{cov}(X,Y)^2/V(Y)
\end{aligned}
$$

Thus, the last expression is nonnegative as well. Multiplying it by $V(Y)$ yields the result.

The correlation of two random variables X and Y is defined by $\mathrm{corr}(X,Y) = \mathrm{cov}(X,Y)/\sqrt{V(X)V(Y)}$. By property 5 as before, $|\mathrm{corr}(X,Y)| \leq 1$. If $|\mathrm{corr}(X,Y)| = 1$, then X and Y are linearly dependent, i.e. there exist constants a and b such that $X = a + bY$.

3 Normal Random Variables

Let us begin with the definition of a *multivariate normal distribution*. Let Σ be a positive definite $n \times n$ matrix. Remember that the $n \times n$ matrix Σ is positive definite if $a^T \Sigma a > 0$ for any non-zero $n \times 1$ vector a. Here superindex T denotes transposition. Let μ be $n \times 1$ vector. Then $X \sim N(\mu, \Sigma)$ if X is continuous and its pdf is given by

$$f_X(x) = \frac{\exp(-(x-\mu)^T \Sigma^{-1}(x-\mu)/2)}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}$$

for any $n \times 1$ vector x.

A normal distribution has several useful properties:

1. if $X \sim N(\mu, \Sigma)$, then $\Sigma_{ij} = \text{cov}(X_i, X_j)$ for any $i, j = 1, ..., n$ where $X = (X_1, ..., X_n)^T$
2. if $X \sim N(\mu, \Sigma)$, then $\mu_i = E[X_i]$ for any $i = 1, ..., n$
3. if $X \sim N(\mu, \Sigma)$, then any subset of components of X is normal as well. In particular, $X_i \sim N(\mu_i, \Sigma_{ii})$
4. if X and Y are uncorrelated normal random variables, then X and Y are independent. As an exercise, check this statement
5. if $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, and X and Y are independent, then $X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$
6. Any linear combination of normals is normal. That is, if $X \sim N(\mu, \Sigma)$ is an $n \times 1$ dimensional normal vector, and A is a fixed $k \times n$ full-rank matrix with $k \leq n$, then $Y = AX$ is a normal $k \times 1$ vector: $Y \sim N(A\mu, A\Sigma A^T)$.

3.1 Conditional distribution

Another useful property of a normal distribution is that its conditional distribution is normal as well. If

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

then $X_1|X_2 = x_2 \sim N(\tilde{\mu}, \tilde{\Sigma})$ with $\tilde{\mu} = \mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$ and $\tilde{\Sigma} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. If X_1 and X_2 are both random variables (as opposed to random vectors), then $E[X_1|X_2 = x_2] = \mu_1 + \text{cov}(X_1, X_2)(x_2 - \mu_2)/V(X_2)$. Let us prove the last statement. Let

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

be the covariance matrix of 2×1 normal random vector $X = (X_1, X_2)^T$ with mean $\mu = (\mu_1, \mu_2)^T$. Note that $\Sigma_{12} = \Sigma_{21} = \sigma_{12}$ since $\text{cov}(X_1, X_2) = \text{cov}(X_1, X_2)$. From linear algebra, we know that $\det(\Sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}^2$ and

$$\Sigma^{-1} = \frac{1}{\det(\Sigma)} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}.$$

Thus the pdf of X is

$$f_X(x_1, x_2) = \frac{\exp\{-[(x_1-\mu_1)^2\sigma_{22} + (x_2-\mu_2)^2\sigma_{11} - 2(x_1-\mu_1)(x_2-\mu_2)\sigma_{12}]/(2\det(\Sigma)\}}{2\pi\sqrt{\det(\Sigma)}},$$

and the pdf of X_2 is

$$f_{X_2}(x_2) = \frac{\exp\{-(x_2-\mu_2)^2/(2\sigma_{22})\}}{\sqrt{2\pi\sigma_{22}}}.$$

Note that

$$\frac{\sigma_{11}}{\det(\Sigma)} - \frac{1}{\sigma_{22}} = \frac{\sigma_{11}\sigma_{22} - (\sigma_{11}\sigma_{22} - \sigma_{12}^2)}{\det(\Sigma)\sigma_{22}} = \frac{\sigma_{12}^2}{\det(\Sigma)\sigma_{22}}.$$

Therefore the conditional pdf of X_1, given $X_2 = x_2$, is

$$\begin{aligned}
f_{X_1|X_2}(x_1|X_2 = x_2) &= \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)} \\
&= \frac{\exp\{-[(x_1-\mu_1)^2\sigma_{22} + (x_2-\mu_2)^2\sigma_{12}^2/\sigma_{22} - 2(x_1-\mu_1)(x_2-\mu_2)\sigma_{12}]/(2\det(\Sigma))\}}{\sqrt{2\pi}\sqrt{\det(\Sigma)/\sigma_{22}}} \\
&= \frac{\exp\{-[(x_1-\mu_1)^2 + (x_2-\mu_2)^2\sigma_{12}^2/\sigma_{22}^2 - 2(x_1-\mu_1)(x_2-\mu_2)\sigma_{12}/\sigma_{22}]/(2\det(\Sigma)/\sigma_{22})\}}{\sqrt{2\pi}\sqrt{\det(\Sigma)/\sigma_{22}}} \\
&= \frac{\exp\{-[x_1-\mu_1-(x_2-\mu_2)\sigma_{12}/\sigma_{22}]^2/(2\det(\Sigma)/\sigma_{22})\}}{\sqrt{2\pi}\sqrt{\det(\Sigma)/\sigma_{22}}} \\
&= \frac{\exp\{-(x_1-\tilde{\mu})^2/(2\tilde{\sigma})\}}{\sqrt{2\pi}\sqrt{\tilde{\sigma}}},
\end{aligned}$$

where $\tilde{\mu} = \mu_1 + (x_2-\mu_2)\sigma_{12}/\sigma_{22}$ and $\tilde{\sigma} = \det(\Sigma)/\sigma_{22}$. Note, that the last expression equals the pdf of a normal random variable with mean $\tilde{\mu}$ and variance $\tilde{\sigma}$ yields the result.