Elements of Probability Theory

Klaus Neusser

September 30, 2015

Contents

1 Density and Distribution Function

1.1 Univariate Density Function

Definition 1. *A* random variable *is a variable whose values are determined by a probability distribution.*

Definition 2. *A* random variable *is a real valued function defined on some probability space. The random variable is denoted by capital letters whereas its realizations are denoted by small letters.*

Definition 3. *The* density function *of a continuous random variable X is a nonnegative function f such that :*

$$\mathbf{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx.$$

Remark 1.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

The *conditional distribution* of two events A and B, denoted by $\mathbf{P}(A|B)$, is defined as

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

The conditional distribution $\mathbf{P}(x_1 \leq X \leq x_2 | a \leq X \leq b)$ is therefore given by

$$\mathbf{P}(x_1 \leq X \leq x_2 | a \leq X \leq b) = \frac{\mathbf{P}(x_1 \leq X \leq x_2 \text{ and } a \leq X \leq b)}{\mathbf{P}(a \leq X \leq b)}$$

if $[x_1, x_2] \subseteq [a, b]$.

Definition 4. *The* conditional density function *is defined as:*

$$f(x|a \leq X \leq b) = \begin{cases} \frac{f(x)}{\int_a^b f(x)dx}, & a \leq x \leq b; \\ 0, & \textit{otherwise.} \end{cases}$$

For any event S with $\mathbf{P}(X \in S) > 0$ we have:

$$f(x|X \in S) = \begin{cases} \frac{f(x)}{\mathbf{P}(X \in S)}, & x \in S; \\ 0, & \textit{otherwise.} \end{cases}$$

1.2 Bivariate random variables

Definition 5. *If X and Y are two random variables, then the nonnegative function $f(x,y)$ is called the* joint density function *of X and Y if*

$$\mathbf{P}(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,y) dy dx$$

Remark 2. *It holds that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dy dx = 1$.*

Remark 3. *The order of the integration can be exchanged.*

Example

$$f(x,y) = \begin{cases} 1, & 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

$\mathbf{P}(X > Y) = \int \int_S f(x,y) dx dy = \int \int_S 1 dx dy = \frac{1}{2}$.
$\mathbf{P}(X^2 + Y^2) = \frac{\pi}{4}$

Definition 6. *The* marginal distribution *(*marginal density function*) is defined as:*

$$\mathbf{P}(x_1 \leq X \leq x_2) = \mathbf{P}(x_1 \leq X \leq x_2, -\infty \leq Y \leq \infty)$$
$$f(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

Definition 7. *: Let $f(x,y)$ be the joint density of (X,Y) and let S be an event such that $\mathbf{P}((X,Y) \in S) > 0$, then the* conditional density *is given by*

$$f(x,y|S) = \begin{cases} \frac{f(x,y)}{\mathbf{P}((X,Y) \in S)}, & (x,y) \in S; \\ 0, & \textit{otherwise.} \end{cases}$$

An important special case is given by the event $S = \{y_1 \leq Y \leq y_2\}$:

$$f(x|S) = \frac{\int_{y_1}^{y_2} f(x,y) dy}{\int_{-\infty}^{\infty} \int_{y_1}^{y_2} f(x,y) dy \, dx}$$

It is also possible to define the conditional density with respect to the event $S = \{Y = y\}$:

$$f(x|Y = y) = \frac{f(x,y)}{f(y)}.$$

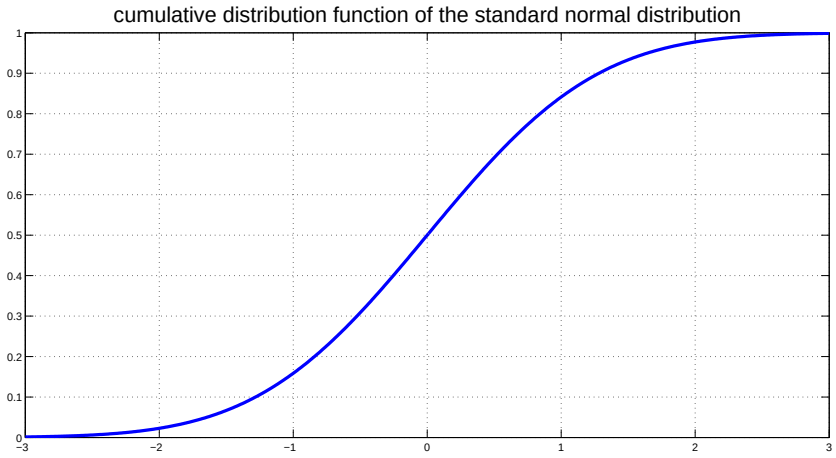

Figure 1: Cumulative distribution function of the standard normal distribution

Definition 8. *Two random variables are called* independent*, if and only if*

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

This amounts to:

$$\mathbf{P}(x_1 \leq X \leq x_2 \text{ and } y_1 \leq Y \leq y_2) = \mathbf{P}(x_1 \leq X \leq x_2)\mathbf{P}(y_1 \leq Y \leq y_2)$$

1.3 Cumulative Distribution Function

Definition 9. *The* cumulative distribution function *F of a random variable X is defined by*

$$F(x) = \mathbf{P}(X < x) = \int_{-\infty}^{x} f(t)dt$$

Remark 4. *F has the following properties:*

(i) *F is monotonically nondecreasing*

(ii) *$F(-\infty) = 0$ and $F(\infty) = 1$.*

(iii) *F is continuous to the left*

Similarly, the n-dimensional cumulative distribution function is defined as

$$F(x_1, x_2, \ldots, x_n) = \mathbf{P}(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n)$$

If X and Y are two independent random variables, we have:

$$F(x, y) = \mathbf{P}(X < x, Y < y) = \mathbf{P}(X < x)\mathbf{P}(Y < y) = F(x)F(y)$$

2 Moments of a Distribution

Definition 10. *: Let X be a real continuous random variable with density $f(x)$, then the* expected value *or* mean *of X, $\mathbb{E}X$, is defined as*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x) dx,$$

if the integral exists.

Besides the expected value, there are other statistics which determine the location of a distribution:

- The *mode* is given by the maximum of $f(x)$.
- The *median* is the value m which satisfies: $\mathbf{P}(X \leq m) = \frac{1}{2}$.

A distribution is called *positively skewed* if mode < median < expected value.

Theorem 1. *For any continuous function Φ we have:*

$$\mathbb{E}\Phi(X) = \int_{-\infty}^{\infty} \Phi(x) f(x) dx$$
$$\mathbb{E}\Phi(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(x, y) f(x, y) dx\, dy$$

This Theorem leads to the following conclusions:

(i) $\mathbb{E}\alpha = \alpha$ for all $\alpha \in \mathbb{R}$

(ii) $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}X + \beta \mathbb{E}Y$

(iii) For two independent random variables X and Y, $\mathbb{E}X\,Y = \mathbb{E}X\mathbb{E}Y$

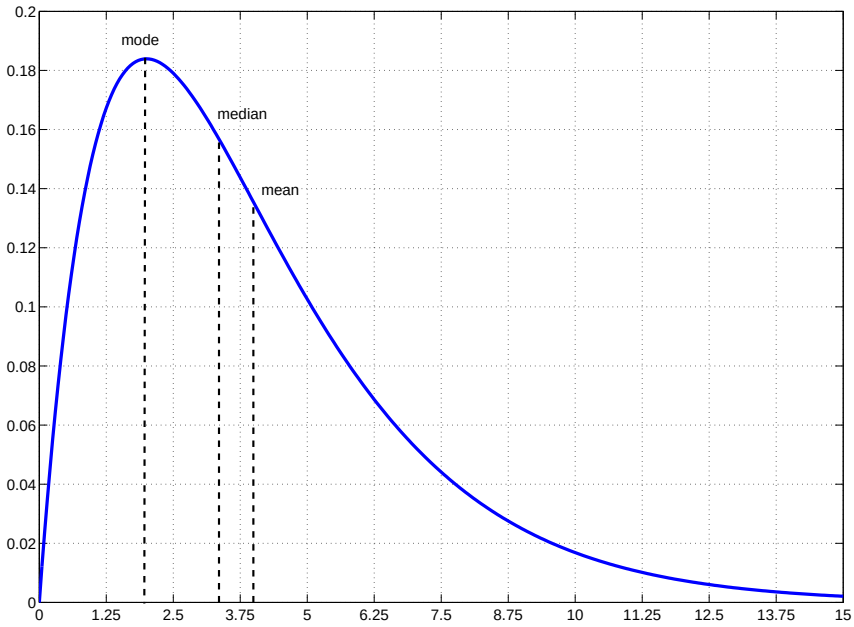

Figure 2: Location of distribution with positive skewness

In order to characterize the distribution completely it is necessary to consider moments of higher order. The k-th moment around zero is defined as $\mathbb{E}X^k$; the k-th moment around the mean $\mathbb{E}(X - \mathbb{E}X)^k$
The second moment around the mean is called the *variance*. It is the second most important quantity after the mean to characterize a distribution. It measures the spread of the variable around its mean. The variance of X is defined as:

$$\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

The variance has the following properties:

$$\begin{aligned} \mathbb{V}X &\geq 0 \\ \mathbb{V}X = 0 &\Longleftrightarrow X = \mathbb{E}X \text{ constant} \\ \mathbb{V}(\alpha X + \beta) &= \alpha^2 \mathbb{V}X \end{aligned}$$

Instead of the variance one often considers the *standard deviation*:
$\sigma_X = \sqrt{\mathbb{V}X}$.

3 Relation between Random Variables

The relation between two random variables can be measured by their *covariance*:

Definition 11. *: The* covariance *between two random variables X and Y is defined as:*

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y \\ &= \mathbb{E}((X - \mathbb{E}X)Y) = \mathbb{E}(X(Y - \mathbb{E}Y)) \end{aligned}$$

$\operatorname{cov} > 0$, if $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ show a tendency to have the same sign; $\operatorname{cov} < 0$, if $X - \mathbb{E}X$ und $Y - \mathbb{E}Y$ show a tendency to have the opposite sign.

Theorem 2. *If X and Y are two independent random variables: $cov(X, Y) = 0$.*

However, the reverse is not true! Although $cov(X,Y) = 0$, the random variables X and Y may still dependent.

Theorem 3. $\mathbb{V}(X \pm Y) = \mathbb{V}X + \mathbb{V}Y \pm 2\text{cov}(X,Y)$

Theorem 4. *If X_i, $i = 1, \ldots, n$, are pairwise independent, we have:*

$$\mathbb{V}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathbb{V}X_j$$

This Theorem remains correct if we replace the condition of independence by $\text{cov}(X_i, X_j) = 0$, $i \neq j$.
One disadvantage of the covariance as a measure of relation between two random variables is its dependence on the units of measurement. A measure which is free from the units of measurement is given by the correlation coefficient.

Definition 12. *: The* correlation coefficient *between two random variables X and Y is defined as:*

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y denote the corresponding standard deviations.

Theorem 5. *The correlation between αX and βY, $\alpha, \beta \in \mathbb{R}$, is equal to the correlation between X and Y, i.e. $\rho_{\alpha X, \beta Y} = \rho_{X,Y}$*

Theorem 6. $|\rho_{X,Y}| \leq 1$

Proof. For all $\lambda \in \mathbb{R}$ we have: $\mathbb{E}((X - \mathbb{E}X) - \lambda(Y - \mathbb{E}Y))^2 = VX + \lambda^2 VY - 2\lambda\text{cov}(X,Y) \geq 0$. Setting λ equal to $\frac{cov(X,Y)}{VY}$, we get: $\mathbb{V}X + \frac{(\text{cov}(X,Y))^2}{\mathbb{V}Y} - 2\frac{(\text{cov}(X,Y))^2}{\mathbb{V}Y} = \mathbb{V}X - \frac{(\text{cov}(X,Y))^2}{\mathbb{V}Y} \geq 0$. This implies: $\rho_{X,Y}^2 = \frac{(\text{cov}(X,Y))^2}{\mathbb{V}X\mathbb{V}Y} \leq 1$. □

We say that

- X and Y are uncorrelated if $\rho = 0$;
- X and Y are positively correlated if $\rho > 0$;
- X and Y are negatively correlated if $\rho < 0$.

4 Forecast

We are looking for the best linear predictor for Y in the class of all linear functions of X. By "best predictor" we mean to minimize the mean quadratic forecast error. Thus we want to solve the following minimization problem:

$$S(\alpha, \beta) = \mathbb{E}(Y - \alpha - \beta X)^2 \longrightarrow \min_{\alpha,\beta}.$$

S is called the mean squared error (MSE) or means squared prediction error. Differentiating the above expression with respect to α and β leads to the following first order condition for the minimum:

$$\begin{aligned}
\frac{\partial S}{\partial \alpha} &= 2\alpha - 2\mathbb{E}Y + 2\beta\mathbb{E}X = 0 \\
&= \alpha - \mathbb{E}Y + \beta\mathbb{E}X = 0 \\
\frac{\partial S}{\partial \beta} &= 2\beta\mathbb{E}X^2 + 2\alpha\mathbb{E}X - 2\mathbb{E}XY = 0 \\
&= \beta\mathbb{E}X^2 + \alpha\mathbb{E}X - \mathbb{E}XY = 0
\end{aligned}$$

These equations are also called the *normal equations* because they imply that the expected prediction error is zero and that the prediction error is orthogonal to X.
The solution of this equation system leads to:

$$\begin{aligned}
\hat{\beta} &= \frac{\operatorname{cov}(X, Y)}{\mathbb{V}X} \\
\hat{\alpha} &= \mathbb{E}Y - \hat{\beta}\mathbb{E}X = \mathbb{E}Y - \frac{\operatorname{cov}(X, Y)}{\mathbb{V}X}\mathbb{E}X
\end{aligned}$$

Theorem 7. *The best linear predictor of Y given X in the MSE sense is given by $\hat{\alpha} + \hat{\beta}X$.*

In the case of the least-squares (LS) estimator, we replace $\operatorname{cov}(X, Y)$, $\mathbb{V}X$ and the expected values by their sample counterparts. In this sense, we can speak of the LS-estimator as a natural candidate for an estimator. The forecast function is then $\widehat{Y} = \hat{\alpha} + \hat{\beta}X$. The LS-residual is defined by $U = Y - \widehat{Y}$.

The following relations hold:

$$\begin{aligned}
\mathbb{V}\widehat{Y} &= \hat{\beta}^2\mathbb{V}X = \frac{\mathrm{cov}(X,Y)^2}{(\mathbb{V}X)^2}\mathbb{V}X = \frac{\mathrm{cov}(X,Y)^2}{\mathbb{V}X\mathbb{V}Y}\mathbb{V}Y = \rho_{X,Y}^2\mathbb{V}Y \\
\mathbb{V}U &= \mathbb{V}(Y - \hat{\alpha} - \hat{\beta}X) = VY + \frac{\mathrm{cov}(X,Y)^2}{(\mathbb{V}X)^2}\mathbb{V}X - 2\frac{\mathrm{cov}(X,Y)}{\mathbb{V}X}\mathrm{cov}(X,Y) \\
&= (1-\rho_{X,Y}^2)\mathbb{V}Y \\
\mathrm{cov}(\widehat{Y},U) &= \mathrm{cov}(\widehat{Y},Y-\widehat{Y}) = \mathrm{cov}(\widehat{Y},Y) - \mathbb{V}\widehat{Y} = \hat{\beta}\mathrm{cov}(X,Y) - \hat{V}\widehat{Y} = 0
\end{aligned}$$

The variance of Y can be decomposed as follows:

$$\mathbb{V}Y = \rho_{X,Y}^2\mathbb{V}Y + (1-\rho_{X,Y}^2)\mathbb{V}Y = \mathbb{V}\widehat{Y} + \mathbb{V}U$$

The correlation coefficient is thus a measure of the linear dependence of Y and X. ρ can be small even if the two random variables are non-linearly related. For example, let X be a random variable with the properties $\mathbb{E}X = 0$ and $\mathbb{E}X^3 = 0$, and define Y as $Y = X^2$. In this case $\mathrm{cov}(X,Y) = \mathbb{E}XY = \mathbb{E}X^3 = 0$. Therefore $\rho_{X,Y} = 0$ although Y and X are perfectly related.

5 The Conditional Expectation

As conditional distributions are again ordinary distributions, it is possible to compute their expected values and their variance.

Definition 13. *: Let (X,Y) be a bivariate continuous random variable with conditional density $f(y|x)$ and let ϕ be an arbitrary function, then the* conditional expectation *of $\phi(X,Y)$ given X is defined as*

$$\mathbb{E}(\phi(X,Y)|X) = \mathbb{E}_{Y|X}\phi(X,Y) = \int \phi(x,y)f(y|X)dy.$$

Remark 5. *$\mathbb{E}(\phi(X,Y)|X) = \mathbb{E}_{Y|X}\phi(X,Y)$ is just a function of X which can be evaluated at particular values of X. $\mathbb{E}(\phi(X,Y)|X) = \mathbb{E}_{Y|X}\phi(X,Y)$ can thus be seen as a random variable in X.*

Theorem 8. $\mathbb{E}\phi(X,Y) = \mathbb{E}_X\mathbb{E}_{Y|X}\phi(X,Y)$

Proof.

$$\begin{aligned}\mathbb{E}\phi(X,Y) &= \int\int \phi(x,y)f(x,y)dxdy = \int\int \phi(x,y)f(y|x)f(x)dxdy \\ &= \int \left[\int \phi(x,y)f(y|x)dy\right] f(x)dx = \int \mathbb{E}_{Y|X}\phi(X,Y)f(x)dx \\ &= \mathbb{E}_X\mathbb{E}_{Y|X}\phi(X,Y)\end{aligned}$$

□

Remark 6. *This Theorem is called the Law of Iterated Expectations and plays a central role in the theory of rational expectations.*

Theorem 9. $\mathbb{V}\phi(X,Y) = \mathbb{E}_X\mathbb{V}_{Y|X}\phi(X,Y) + \mathbb{V}_X\mathbb{E}_{Y|X}\phi(X,Y)$

Proof. We have:

$$\mathbb{V}_{Y|X}\phi(X,Y) = \mathbb{E}_{Y|X}\phi^2(X,Y) - (\mathbb{E}_{Y|X}\phi(X,Y))^2.$$

In addition:

$$\begin{aligned}\mathbb{V}_X\mathbb{E}_{Y|X}\phi(X,Y) &= \mathbb{E}_X(\mathbb{E}_{Y|X}\phi(X,Y))^2 - (\mathbb{E}_X\mathbb{E}_{Y|X}\phi(X,Y)))^2 \\ &= \mathbb{E}_X(\mathbb{E}_{Y|X}\phi(X,Y))^2 - (\mathbb{E}\phi(X,Y))^2.\end{aligned}$$

Combining the two equations leads to:

$$\mathbb{E}_X\mathbb{V}_{Y|X}\phi(X,Y)+\mathbb{V}_X\mathbb{E}_{Y|X}\phi(X,Y) = E\phi^2(X,Y)-(\mathbb{E}\phi(X,Y))^2 = V(\phi(X,Y)$$

□

Sometimes it is easier to compute the unconditional variance or the unconditional expectation in this way.
We are again looking for the best predictor, but we do not restrict ourselves to the class of linear predictors and consider also non-linear predictors. The minimization problem then becomes:

$$\mathbb{E}(Y-\phi(X))^2 \min_{\phi}$$

Theorem 10. *The best predictor of Y given X is the conditional expectation $\mathbb{E}(Y|X)$.*

Proof.

$$\begin{aligned}\mathbb{E}(Y-\phi(X))^2 &= \mathbb{E}\{(Y-\mathbb{E}(Y|X)+(\mathbb{E}(Y|X)-\phi(X))\}^2 \\ &= \mathbb{E}(Y-\mathbb{E}(Y|X))^2+\mathbb{E}(\mathbb{E}(Y|X)-\phi(X))^2 \\ &\quad + 2\mathbb{E}(Y-\mathbb{E}(Y|X))(\mathbb{E}(Y|X)-\phi(X))\end{aligned}$$

On the other hand, the Law of Iterated Expectations implies:

$$\begin{aligned}&\mathbb{E}(Y-\mathbb{E}(Y|X))(\mathbb{E}(Y|X)-\phi(X)) = \\ &\qquad \mathbb{E}_X\mathbb{E}_{Y|X}\left((Y-\mathbb{E}(Y|X))(\mathbb{E}(Y|X)-\phi(X))\right)=0\end{aligned}$$

We must therefore choose $\phi(X)=\mathbb{E}(Y|X)$. □

6 The Normal Distribution

Univariate Normal Distribution

The normal distribution plays a pivotal role in probability theory and statistics. Its density is given by:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The normal distribution is therefore characterized by the two parameters μ and σ and we write $X\sim \mathrm{N}(\mu,\sigma^2)$.

Theorem 11. *If $X\sim \mathrm{N}(\mu,\sigma^2)$, then $\mathbb{E}X=\mu$ and $\mathbb{V}X=\sigma^2$.*

The density of the normal distribution is symmetric around μ and has the typical shape of a bell. Because $f(\mu)=\frac{1}{\sigma\sqrt{2\pi}}$, $f(x)$ becomes flatter when σ is increasing.
If $X\sim N(\mu,\sigma^2)$, then $Z=\frac{X-\mu}{\sigma}\sim \mathrm{N}(0,1)$. $\mathrm{N}(0,1)$ is called the standard normal distribution. It has a mean of zero and a variance of one.

Theorem 12. *Let $X\sim \mathrm{N}(\mu,\sigma^2)$ and $Y=a+bX$, then $Y\sim \mathrm{N}(a+b\mu,b^2\sigma^2)$.*

Bivariate Normal Distribution

Definition 14. *: The density of the* bivariate normal distribution *is*

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

Theorem 13. *If (X,Y) follow a bivariate normal distribution, then the following is true:*

(i) *The marginal densities $f(x)$ and $f(y)$ are univariate normal distributions.*

(ii) *The conditional distributions $f(x|y)$ and $f(y|x)$ are univariate normal distributions;*

(iii) $\mathbb{E}X = \mu_X$, $\mathbb{V}X = \sigma_X^2$, $\mathbb{E}Y = \mu_Y$, $\mathbb{V}Y = \sigma_Y^2$.

(iv) *The correlation coefficient between X and Y, $\rho_{X,Y}$, is equal to $\rho_{X,Y} = \rho$.*

(v) $\mathbb{E}(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$ *and* $\mathbb{V}(Y|X) = \sigma_Y^2(1-\rho^2)$.

The above properties characterize the normal distribution. It is the only distribution with these properties.

Theorem 14. *If (X,Y) is distributed as a bivariate normal distribution, then $aX+bY$ is also normally distributed.*

Remark 7. *The reverse implication is not true, even when X and Y are each normally distributed.*

Theorem 15. *Let $\{X_t\}$, $t = 1,\ldots,T$, be pairwise independently distributed normal random variables with distribution $\mathrm{N}(\mu,\sigma^2)$ then*

$$\overline{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{T}\right)$$

Theorem 16. *If X and Y are bivariate normally distributed with* $\operatorname{cov}(X,Y) = 0$ *then X and Y are independent.*

The above considerations imply that the best predictor of Y given X, $E(Y|X)$ is the linear predictor.
If X and Y are arbitrary random variables then there exists a random variable Z such that

$$Y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \sigma_Y Z$$

with $\mathbb{E}Z = 0$, $\mathbb{V}Z = 1 - \rho^2$ and $\operatorname{cov}(X,Z) = 0$. If X and Y are normally distributed then Z is also normally distributed. Because $\mathbb{E}(Z|X) = \mathbb{E}Z = 0$ and $\mathbb{V}(Z|X) = \mathbb{V}Z = 1 - \rho^2$, we have

$$\mathbb{E}(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X).$$

Multivariate Normal Distribution

Let $X = (X_1, \ldots, X_n)'$ be a n-vector such that each element X_i is a random variable. In addition let $\mathbb{E}X = (\mu_1, \ldots, \mu_n)$ and $\mathbb{V}X = \Sigma$ with

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}$$

where $\sigma_{ij} = \operatorname{cov}(X_i, X_j)$.

Definition 15. *: A n-dimensional random variable X is* multivariate normally distributed *with mean μ and variance-covariance matrix Σ, $X \sim \mathrm{N}(\mu, \Sigma)$ if its density is:*

$$f(x) = (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

Theorem 17. *Let $X \sim \mathrm{N}(\mu, \Sigma)$ and $X = (Y', Z')'$, where Y and Z are of dimensions h and k with $n = h + k$ with corresponding partition of Σ*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = \mathbb{V}Y$, $\Sigma_{22} = \mathbb{V}Z$, $\Sigma_{12} = \Sigma_{21}' = \operatorname{cov}(Y, Z) = \mathbb{E}(Y - \mathbb{E}Y)'(Z - \mathbb{E}Z)$, *then the subvectors* Y *and* Z *are again normally distributed. The conditional distribution of* Y *given* Z *(and similarly given* X*) are also normally distributed with*

$$\mathbb{E}(Y|Z) = \mathbb{E}Y + \Sigma_{12}\Sigma_{22}^{-1}(Z - \mathbb{E}Z),$$
$$\mathbb{V}(Y|Z) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Theorem 18. *Let* $X \sim \mathrm{N}(\mu, \Sigma)$ *and* A *a* $(m \times n)$ *matrix with* $m \leq n$ *and* m *linearly independent rows then we have*

$$AX \sim \mathrm{N}(A\mu, A\Sigma A')$$

Theorem 19. *Let* $X \sim \mathrm{N}(\mu, \Sigma)$ *and* Y *and* Z *be defined as above. If* $\Sigma_{12} = 0$*, then* Y *and* Z *are independent. In this case* $f(x) = f(y)f(z)$ *where* $f(y)$ *and* $f(z)$ *are the corresponding multivariate normal densities of* Y *and* Z*.*

7 Foundations of Probability Theory

Let Ω be an arbitrary set with elements ω. In probability theory Ω consists of all possible outcomes from an experiment or observation.

Definition 16. *Let* $\mathfrak{A}$ *be a collection of subsets of* Ω*, then* $\mathfrak{A}$ *is called a* σ*-algebra if and only if the following conditions hold:*

(i) $\mathfrak{A}$ *is non-empty. There is at least one* $A \subseteq \Omega$ *in* $\mathfrak{A}$*.*

(ii) $\mathfrak{A}$ *is closed under complementation.* $A \in \mathfrak{A}$ *implies* $\Omega \setminus A \in \mathfrak{A}$*.*

(iii) $\mathfrak{A}$ *is closed under countable unions. If* $A_1, A_2, \ldots \in \mathfrak{A}$ *then* $\bigcup_n A_n \in \mathfrak{A}$*.*

Corollary 20. *The above conditions imply that* $\emptyset \in \mathfrak{A}$ *and* $\Omega \in \mathfrak{A}$*.*

Definition 17. *A probability space is triplet* $(\Omega, \mathfrak{A}, \mathbf{P})$ *such that*

(i) Ω *is an arbitrary nonempty set, called the sample space.*

(ii) $\mathfrak{A}$ *is a* σ*-algebra with respect to* Ω*. The elements* $A \in \mathfrak{A}$ *are called events.*

(iii) $\mathbf{P} : \mathfrak{A} \longrightarrow [0,1]$ *assigns to each event a probability, i.e. a number in* $[0,1]$*, such that* $\mathbf{P}(\emptyset) = 0$ *and* $\mathbf{P}(\Omega) = 1$*.*

(iv) *For every countable collection* $\{A_i\}_{i=1,2,\ldots}$ *of pairwise disjoint sets in* $\mathfrak{A}$*,* $\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$

Definition 18. *Given a probability space* $(\Omega, \mathfrak{A}, \mathbf{P})$ *a real valued random variable* X *is function from* Ω *to* $\mathbb{R}$*, i. e.* $X : \Omega \longrightarrow \mathbb{R}$*, such that* $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathfrak{A}$ *for every* $B \in \mathfrak{B}$*, the Borel* σ*-algebra of real numbers.*

8 Stochastic Processes

Definition 19. *A* stochastic process $\{X_t\}$ *is a family of random variables indexed by* $t \in \mathcal{T}$ *and defined on some probability space* $(\Omega, \mathcal{F}, \mathbf{P})$*.*

Most of the time the mentioning of the probability space is suppressed. $\mathcal{T}$ is often interpreted as a time index. In this case $\mathcal{T} = \mathbb{N}$ respectively $\mathbb{Z}$ or $\mathbb{R}$. In the following we consider only discrete stochastic processes with infinite past and future, i.e. we take $\mathcal{T} = \mathbb{Z}$.

Definition 20. *If* $\{X_t\}$ *is a stochastic process with* $\mathbb{V}X_t < \infty$ *for all* $t \in \mathbb{Z}$ *then the function* $\gamma_X(t,s)$ *with* $t, s \in \mathbb{Z}$ *is called the* autocovariance function *of* $\{X_t\}$ *which is defined as follows:*

$$\gamma_X(t,s) = \operatorname{cov}(X_t, X_s) = \mathbb{E}\left[(X_t - \mathbb{E}X_t)(X_s - \mathbb{E}X_s)\right] = \mathbb{E}X_tX_s - \mathbb{E}X_t\mathbb{E}X_s.$$

8.1 Stationarity

Definition 21. *A stochastic process* $\{X_t\}$ *is called* stationary *if for all integers* r*,* s *and* t *the following properties hold:*

(i) $\mathbb{E}X_t = \mu$ *constant;*

(ii) $\mathbb{V}X_t < \infty$*;*

(iii) $\gamma_X(t,s) = \gamma_X(t+r, s+r)$.

Remark 8. *Processes with these properties are also called weakly stationary, stationary in the wide sense, covariance stationary, or stationary of second order.*

Remark 9. *If $t = s$, $\gamma_X(t,s) = \gamma_X(t,t) = \mathbb{V}X_t$. If $\{X_t\}$ is stationary then $\gamma_X(t,t) = \gamma_X(0) = \mathbb{V}X_t$ is the (unconditional) variance of the process.*

Remark 10. *If $\{X_t\}$ is stationary then the autocovariance function for $r = -s$ is given by:*

$$\gamma_X(t,s) = \gamma_X(t-s, 0).$$

The covariance $\gamma_X(t,s)$ does therefore not depend on the time periods t and s, but only on the time difference $t-s$. For stationary processes, we therefore consider the covariance function only as a function in one argument. We denote the covariance function in this case by $\gamma_X(h)$, $h \in \mathbb{Z}$. Because $\gamma_X(t,s) = \gamma_X(s,t)$, for all t and s, we get in addition

$$\gamma_X(h) = \gamma_X(-h) \text{ for all integers } h.$$

We therefore consider the covariance function (autocorrelation function) only for nonnegative integers $h = 0, 1, 2, \ldots$ whereby h is called the order *or* lag.

If one considers instead of the covariances the correlations of a stationary process, one obtains the *autocorrelation function* (ACF) defined as:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \operatorname{corr}(X_{t+h}, X_t) \text{ for all integers } h.$$

In many applications, knowledge of the first two moments is sufficient to characterize the properties of the process. However, there are situation where it is necessary to consider the whole distribution. This leads to to the concept of strong stationarity.

Definition 22. *A stochastic process is called* strictly (strongly) stationary *if and only if the joint distribution of $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_1+h}, \ldots, X_{t_n+h})$ is the same for all $h \in \mathbb{Z}$ and all $(t_1, \ldots, t_n) \in \mathcal{T}^n$, $n = 1, 2, \ldots$*

An equivalent definition is given by:

Definition 23. *A stochastic process is called* strictly (strongly) stationary *if and only if, for all integers h and $n \geq 1$, the distributions of $(X_1, \ldots, X_n)$ and $(X_{1+h}, \ldots, X_{n+h})$ are identical.*

Remark 11. *If $\{X_t\}$ is strictly stationary then X_t has the same distribution for all t (n=1). The above definition applied to $n = 2$ implies that the joint distribution of X_{t+h} and X_t is independent of t. Thus the covariances dependent on h only. A strictly stationary process with finite second moments is therefore also stationary.*

The converse is, however, not true as the following example shows:

$$X_t \sim \begin{cases} \text{exponential with mean 1 (i.e. } f(x) = \mathrm{e}^{-x}), & t \text{ uneven;} \\ \text{N(1,1)}, & t \text{ even;} \end{cases}$$

such that the X_t's are independent of each other. In this case:

- $\mathbb{E}X_t = 1$
- $\gamma_X(0) = 1$ und $\gamma_X(h) = 0$ für $h \neq 0$

The process is therefore stationary, but clearly not strictly stationary as the distribution changes according to whether t is even or uneven.

Definition 24. *A stochastic process $\{X_t\}$ is called a* Gaussian process *if and only if all finite-dimensional distributions of $\{X_t\}$ are multivariate normal.*

Remark 12. *A stationary Gaussian process is also strictly stationary. For all $n, h, t_1, \ldots, t_n$, $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_1+h}, \ldots, X_{t_n+h})$ have the same mean and the same covariance matrix.*

8.2 White Noise

The simplest process is called a White noise process or shortly White noise. It is defined as follows:

Definition 25. *The stochastic process* $\{Z_t\}$ *is called* White noise *(a White Noise process) if and only if* $\{Z_t\}$ *is stationary and*

- $\mathbb{E}Z_t = 0$
- $\gamma_Z(h) = \begin{cases} \sigma^2 & h = 0; \\ 0 & h \neq 0. \end{cases}$

Such processes are denoted by $Z_t \sim \mathrm{WN}(0, \sigma^2)$.

White noise processes are therefore processes with no autocorrelation in the time dimension, i.e. the autocorrelation function is always equal to zero, except for $h = 0$ where it is equal to one. As the autocorrelation function has no particular structure, it is impossible to conjecture the future development of the process from past observations by considering the first two moments only. A White noise process has therefore no memory.
If $\{Z_t\}$ is not only uncorrelated but also independently and identically distributed we write $Z_t \sim \mathrm{IID}(0, \sigma^2)$. Thereby IID stands for independently and identically distributed. If in addition Z_t normally distributed, we write $Z_t \sim \mathrm{IIN}(0, \sigma^2)$. A $\mathrm{IID}(0, \sigma^2)$ process is therefore always white noise. The converse, however, is not true.

8.3 Martingale

The sequence of σ-algebras $\mathcal{F}_t = \{X_t, X_{t-1}, \dots\}$ is called the information set.[1]

Definition 26. *A* martingale *is stochastic process* $\{M_t\}$ *with the following properties:*

- $\mathbb{E}|M_t| < \infty$
- $\mathbb{E}_t M_{t+h} = \mathbb{E}(M_{t+h}|\mathcal{F}_t) = M_t$ *for all* $h \geq 0$.

[1] More precisely, it is the smallest σ-algebra such that all X_{t-j}, $j = 0, 1, 2, \dots$ are measurable random variables. The sequence $\{\mathcal{F}_t\}$ is then called the natural filtration.

The classical example of a martingale is the random walk with IID increments, i.e. $X_t = X_{t-1} + Z_t$ with $Z_t \sim \text{IID}(0, \sigma^2)$.

Definition 27. $\{X_t\}$ *is called a* martingale difference *if and only if* $\mathbb{E}_t X_{t+1} = \mathbb{E}(X_{t+1}|\mathcal{F}_t) = 0$.

A martingale difference is thus a process for which the past provides no information about the future evolution. A martingale difference can be considered as something between a White noise and an IID process. An important example of a martingale difference is given by the forecast errors:

$$\varepsilon_{t+1} = X_{t+1} - \mathbb{E}_t X_{t+1}.$$

Martingale differences share the following properties:

- The Law of Iterated Expectations implies (see Theorem 8):
  $\mathbb{E}X_t = \mathbb{E}(\mathbb{E}(X_{t+1}|\mathcal{F}_t)) = 0$.
- The same argument implies that:
  $\text{cov}(X_t, X_{t+h}) = \mathbb{E}(X_t X_{t+h}) = 0$ for $h \neq 0$.

These properties do not imply that martingale differences are stationary. In particular, it can be the case that their variances depend on t. Moreover, martingale differences are not independently distributed as the following example demonstrates.

Example: $X_t = X_{t-1}\frac{\varepsilon_t}{\varepsilon_{t-2}}$ with $X_1 = \varepsilon_1$, $\varepsilon_0 = 1$ and $\varepsilon_t \sim \text{IID}(0, \sigma_t^2)$. One can immediately verify that $X_2 = X_1\frac{\varepsilon_2}{\varepsilon_0} = \varepsilon_2\varepsilon_1$ and that therefore $X_t = \varepsilon_t\varepsilon_{t-1}$.
Moreover it is possible to infer $\{\varepsilon_t\}$ perfectly from $\{X_t\}$ because $\varepsilon_2 = \frac{X_2}{X_1}\varepsilon_0 = \frac{X_2}{X_1}$ respectively $\varepsilon_3 = \frac{X_3}{X_2}\varepsilon_1 = \frac{X_2}{\varepsilon_2\varepsilon_1}\varepsilon_1 = \frac{X_3}{\varepsilon_2}$ or more generally $\varepsilon_t = \frac{X_t}{\varepsilon_{t-1}}$.
We therefore get: $\mathbb{E}(X_{t+1}|\mathcal{F}_t) = \mathbb{E}\left(\frac{X_t\varepsilon_{t+1}}{\varepsilon_{t-1}}|\mathcal{F}_t\right) = \frac{X_t}{\varepsilon_{t-1}}\mathbb{E}(\varepsilon_{t+1}|\mathcal{F}_t) = 0$. $\{X_t\}$ is therefore a martingale difference, despite the fact that there is an exact dependence between X_t and its past.

9 Stochastic Convergence

In probability theory it is often necessary to compute the limit of a sequence of random variables. There are different concepts for convergence. The most important ones are *convergence in probability*, *convergence in quadratic mean* and *convergence in distribution*.

Definition 28. *A sequence of random variables* $\{X_t\}$ *converges in* probability *to a random variable* X *if and only if for all* $\varepsilon > 0$

$$\lim_{t\to\infty} \mathbf{P}[|X_t - X| > \varepsilon] = 0.$$

This is denoted by $X_t \xrightarrow{p} X$.

Theorem 21 (Continuous Mapping Theorem)**.** *If* $\{X_t\}$ *is a sequence of random variables with realizations in* R^n *which converges in probability to a random variable* X *then* $f(X_t) \xrightarrow{\mathrm{p}} f(X)$ *for any continuous function* $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$.

Definition 29. *A sequence of random variables* $\{X_t\}$ *converges in* quadratic mean *(*mean-square convergence*) to a random variable* X *if and only if*

$$\lim_{t\to\infty} \mathbb{E}[|X_t - X|^2] = 0.$$

This is denoted by $X_t \xrightarrow{m.s.} X$.

The following inequality, the so-called Chebychev's Inequality, is a very useful tool.

Theorem 22 (Chebychev's Inequality)**.** *If* $\mathbb{E}|X|^r < \infty$ *for* $r \geq 0$ *then for any* $\varepsilon > 0$ *we have*

$$\mathbf{P}[|X| \geq \varepsilon] \leq \varepsilon^{-r}\mathbb{E}|X|^r.$$

Chebychev's Inequality immediately implies the following theorem.

Theorem 23. *If* $X_t \xrightarrow{m.s.} X$ *then* $X_t \xrightarrow{p} X$.

The reverse is not true. In addition we have:

Theorem 24. *If* $\mathbb{E}X_t \longrightarrow \mu$ *and* $\mathbb{V}X_t \longrightarrow 0$ *then* $X_t \xrightarrow{m.s.} X$ *and therefore* $X_t \xrightarrow{p} X$.

Lemma 1. *Let* $\{X_t\}$ *be a stochastic process with the property that* $\sup_t \mathbb{E}|X_t| < \infty$ *and let* $\{\psi_j : j \in \mathbb{Z}\}$ *be any sequence of real numbers such that* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ *then the infinite sum*

$$\Psi(\mathrm{L})X_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$$

converges absolutely in probability. If in addition $\sup_t \mathbb{E}|X_t|^2 < \infty$ *then the sum also converges in quadratic mean to the same limit.*

Definition 30. *A sequence of n-dimensional random variables* $\{X_t\}$ *with corresponding distribution functions* $\{F_{X_t}\}$ *converges in* distribution *if there exists a n-dimensional random variable* X *with distribution function* F_X *such that*

$$\lim_{t\to\infty} F_{X_t}(x) = F_X(x) \qquad \textit{für alle } x \in \mathcal{C},$$

where $\mathcal{C}$ *denotes the set of points for which* $F_X(x)$ *is continuous. This is denoted by* $X_t \xrightarrow{d} X$.

Convergence in distribution means that we can approximate, for large t, the distribution of X_t by the distribution of X. We also have a Continuous Mapping Theorem for convergence in distribution. In addition we have:

Theorem 25. *If* $X_t \xrightarrow{p} X$ *then* $X_t \xrightarrow{d} X$.

The reverse is not true. If the limit, however, is not a random variable but a constant n-dimensional vector x, convergence in distribution implies convergence in probability, i.e. $X_t \xrightarrow{d} x$ implies $X_t \xrightarrow{p} x$. A further useful result is:

Theorem 26. *If* $\{X_t\}$ *and* $\{Y_t\}$ *are two sequences of n-dimensional random variables which converge in distribution to* X*, respectively to a constant* c*, then we have:*

(i) $X_t + Y_t \xrightarrow{d} X + c$,

(ii) $Y_t' X_t \xrightarrow{d} c'X$.

In many instances the limit is given by the normal distribution. In this case one speaks of *asymptotic normality*.

Definition 31. *A sequence of random variables* $\{X_t\}$ *with "means"* μ_t *and "variances"* $\sigma_t^2 > 0$ *is called* asymptotically normal *if and only if*

$$\sigma_t^{-1}(X_t - \mu_t) \xrightarrow{d} X \sim \mathrm{N}(0,1).$$

The definition does not require that $\mu_t = \mathbb{E}X_t$ nor that $\sigma_t^2 = \mathbb{V}(X_t)$. In particular, asymptotic normality is achieved if X_t is an identically and independently distributed sequence of random variables with constant mean and variance. In this case the so-called *Central Limit Theorem* holds.

Theorem 27 (Central Limit Theorem)**.** *If* $\{X_t\}$ *is a sequence of identically and independently distributed random variables with constant mean* μ *and constant variance* σ^2 *then we have*

$$\sqrt{T}\,\frac{\overline{X}_T - \mu}{\sigma} \xrightarrow{d} \mathrm{N}(0,1),$$

where $\overline{X}_T = T^{-1}\sum_{t=1}^{T} X_t$.

The assumption of identically distributed random variables can be relaxed in several dimensions which leads to a bunch of Central Limit Theorems.