Probability - II: Lecture Notes

Until Midsem

1 Covariance and Independence

Let $X \sim N(0,1)$ and $Y \sim N(0,1)$ be independent random variables. Define:

$$Z = X + Y \quad \Rightarrow \quad Z \sim N(0,2)$$

$$W = X - Y \quad \Rightarrow \quad W \sim N(0,2)$$

The covariance of Z and W is:

$$\begin{aligned}\text{cov}(Z,W) &= E[(X+Y)(X-Y)] - E[X+Y]E[X-Y] \\ &= E[X^2 - Y^2] - (E[X] + E[Y])(E[X] - E[Y]) \\ &= (E[X^2] - E[Y^2]) - (0+0)(0-0) \\ &= (1-1) - 0 = 0\end{aligned}$$

Important: $\text{cov}(X,Y) = 0 \not\Longleftrightarrow$ Independence (Not True in general)

Example 1. *Let $X \sim N(0,1)$. Define $u = X$ and $v = X^2$. Then:*

$$\begin{aligned}cov(u,v) &= E[uv] - E[u]E[v] \\ &= E[X^3] - E[X]E[X^2] \\ &= 0 - 0 \cdot E[X^2] = 0\end{aligned}$$

However, u and v are not independent.

Question: Since $Z = X + Y$ and $W = X - Y$ are both $N(0,2)$ random variables with correlation 0, are they independent?

2 Bivariate Normal Distribution

Let $X \sim N(0,1)$ with PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty$$

Let $Z \sim N(0,1)$ be independent of X. Define:

$$Y = \rho X + \sqrt{1-\rho^2}Z, \quad |\rho| \leq 1 \text{ (parameter)}$$

2.1 Properties of Y

Mean:

$$E[Y] = E[\rho X + \sqrt{1-\rho^2}Z] = \rho E[X] + \sqrt{1-\rho^2}E[Z] = 0$$

Variance:

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(\rho X + \sqrt{1-\rho^2}Z) \\ &= \rho^2 \text{Var}(X) + (1-\rho^2)\text{Var}(Z) \quad \text{(due to independence)} \\ &= \rho^2 \cdot 1 + (1-\rho^2) \cdot 1 = 1\end{aligned}$$

Thus, Y is also a normal RV with mean 0 and variance 1.

Covariance:

$$\begin{aligned}\text{cov}(Y,X) &= \text{cov}(\rho X + \sqrt{1-\rho^2}Z, X)\\ &= \rho\text{cov}(X,X) + \sqrt{1-\rho^2}\text{cov}(Z,X)\\ &= \rho\text{Var}(X) + \sqrt{1-\rho^2}\cdot 0 = \rho\end{aligned}$$

Correlation:

$$\text{corr}(Y,X) = \frac{\text{cov}(Y,X)}{\sqrt{\text{Var}(Y)\text{Var}(X)}} = \frac{\rho}{\sqrt{1\cdot 1}} = \rho$$

2.2 Finding the Joint Distribution

To find the joint distribution of (X,Y), we use a change of variables from (X,Z) to (X,Y).

Let $s = x$ and $t = \rho x + \sqrt{1-\rho^2}z$. The inverse transformation is:

$$x = s, \quad z = \frac{t-\rho s}{\sqrt{1-\rho^2}}$$

The Jacobian of the inverse transformation $(s,t) \to (x,z)$ is:

$$J = \left|\begin{matrix}\frac{\partial x}{\partial s} & \frac{\partial x}{\partial t}\\ \frac{\partial z}{\partial s} & \frac{\partial z}{\partial t}\end{matrix}\right| = \left|\begin{matrix}1 & 0\\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}}\end{matrix}\right| = \frac{1}{\sqrt{1-\rho^2}}$$

The joint PDF of (X,Z) is:

$$f_{X,Z}(x,z) = f_X(x)f_Z(z) = \frac{1}{2\pi}e^{-\frac{x^2+z^2}{2}}$$

2.3 Change of Variable Formula

General Change of Variable Formula: For a random vector $\underline{X} \in \mathbb{R}^d$ and an invertible transformation $\underline{Y} = h(\underline{X})$ where $h : \mathbb{R}^d \to \mathbb{R}^d$,

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(h^{-1}(\underline{y}))|det(J_{h^{-1}})|$$

Alternatively, for the transformation $A : \mathbb{R}^2 \to \mathbb{R}^2$, $\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}x\\ \rho x + \sqrt{1-\rho^2}z\end{pmatrix}$, the Jacobian is:

$$J_A = \left|\begin{matrix}\frac{\partial x}{\partial x} & \frac{\partial x}{\partial z}\\ \frac{\partial y}{\partial x} & \frac{\partial y}{\partial z}\end{matrix}\right| = \left|\begin{matrix}1 & 0\\ \rho & \sqrt{1-\rho^2}\end{matrix}\right| = \sqrt{1-\rho^2}$$

2.4 Joint PDF of (X,Y)

Applying the transformation formula:

$$f_{X,Y}(x,y) = f_{X,Z}\left(x, \frac{y-\rho x}{\sqrt{1-\rho^2}}\right)\left|\frac{1}{\sqrt{1-\rho^2}}\right|$$

Substituting:

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2 + \left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)^2\right)\right)\frac{1}{\sqrt{1-\rho^2}}$$

The exponent simplifies to:

$$-\frac{1}{2}\left(\frac{x^2(1-\rho^2) + y^2 - 2\rho xy + \rho^2x^2}{1-\rho^2}\right) = -\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)$$

Therefore:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right), \quad -\infty < x,y < \infty$$

This is the PDF of a **Bivariate Normal distribution** with means $(0,0)$ and correlation ρ.

Notation:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

X Y are not necessarily independent
distribution of Y is N(0,1)
if rho=0 ,then X Y are independent
Converse is always true !!!
The matrix is the covariance matrix.

2.5 General Case

Let $U = \mu_1 + \sigma_1 X$ and $V = \mu_2 + \sigma_2 Y$. Then:

- $E[U] = \mu_1$, $\text{Var}(U) = \sigma_1^2$
- $E[V] = \mu_2$, $\text{Var}(V) = \sigma_2^2$
- $\text{cov}(U, V) = E[(\sigma_1 X)(\sigma_2 Y)] = \sigma_1\sigma_2 E[XY] = \sigma_1\sigma_2 \text{cov}(X, Y) = \rho\sigma_1\sigma_2$

Mean vector: $\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

Covariance matrix: $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

3 Variance-Covariance Matrix

The matrix $\Sigma = \begin{pmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(Y, X) & \text{var}(Y) \end{pmatrix}$ is called the **Variance-Covariance Matrix**. It is symmetric.

Theorem 1. *The covariance matrix Σ is non-negative definite (n.n.d). A matrix A is n.n.d if $A = A^T$ and for any vector $\underline{x}$, $\underline{x}^T A \underline{x} \geq 0$.*

Proof. Let $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Then:

$$\begin{aligned} \underline{x}^T \Sigma \underline{x} &= (x_1, x_2) \begin{pmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(Y, X) & \text{var}(Y) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= x_1^2 \text{var}(X) + 2x_1 x_2 \text{cov}(X, Y) + x_2^2 \text{var}(Y) \\ &= \text{var}(x_1 X + x_2 Y) \geq 0 \end{aligned}$$

□

Theorem 2. *If (U, V) has a bivariate normal distribution, then U and V are independent if and only if $corr(U, V) = \rho = 0$.*

The determinant of Σ is:

$$\det(\Sigma) = \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2 = \sigma_1^2\sigma_2^2(1 - \rho^2)$$

3.1 General Form of Bivariate Normal PDF

The PDF of a bivariate normal distribution $N_2(\underline{\mu}, \Sigma)$ is:

$$f_{\underline{Z}}(\underline{z}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} (\underline{z} - \underline{\mu})^T \Sigma^{-1} (\underline{z} - \underline{\mu}) \right)$$

where $\underline{z} = \begin{pmatrix} u \\ v \end{pmatrix}$, $\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

The inverse of Σ is:

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

4 Linear Transformations

Theorem 3. *Let $\underline{X} \sim N_2(\underline{\mu}, \Sigma)$ and A be a 2×2 non-singular matrix. Let $\underline{Y} = A\underline{X}$. Then $\underline{Y}$ is also bivariate normal with:*

$$\underline{Y} \sim N_2(A\underline{\mu}, A\Sigma A^T)$$

Example 2. *Let $\underline{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Define:*

$$\underline{Y} = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

where $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

Then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$ and $\underline{Y} \sim N_2(A\underline{\mu}, A\Sigma A^T)$.

Example 3. *Let X_1, X_2 be i.i.d. $N(0,1)$. Then:*

$$\underline{X} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = N_2(\underline{0}, I_2)$$

Let $A_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ be a rotation matrix.

Define $\underline{Y} = A_\theta \underline{X}$. Since A_θ is orthogonal ($A_\theta^T A_\theta = I$), the covariance matrix is:

$$A_\theta I_2 A_\theta^T = A_\theta A_\theta^T = I_2$$

Thus, $\underline{Y} \sim N_2(\underline{0}, I_2)$. The distribution is rotationally symmetric.

5 Generating Bivariate Normal from Standard Normals

Algorithm:

1. Start with Z_1, Z_2 i.i.d. $N(0,1)$, so $\underline{Z} \sim N_2(\underline{0}, I_2)$.
2. Create correlated standard normals:

$$X_1 = Z_1, \quad X_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$$

   Then $\underline{X} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

3. Scale and shift:

$$Y_1 = \mu_1 + \sigma_1 X_1, \quad Y_2 = \mu_2 + \sigma_2 X_2$$

   Then $\underline{Y} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$.

6 Standardization

If Σ is a positive definite matrix, then there exists a matrix B such that $\Sigma = BB^T$ (e.g., from Cholesky decomposition or spectral decomposition $\Sigma = PDP^T$ with $B = P\sqrt{D}$).

Let $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$. Define:

$$\underline{Y} = B^{-1}(\underline{X} - \underline{\mu})$$

The transformed mean is:

$$E[\underline{Y}] = B^{-1}(E[\underline{X}] - \underline{\mu}) = \underline{0}$$

The transformed covariance is:

$$\mathrm{Cov}(\underline{Y}) = B^{-1}\Sigma(B^T)^{-1} = B^{-1}(BB^T)(B^T)^{-1} = I_d$$

Therefore, $\underline{Y} \sim N_d(\underline{0}, I_d)$.

Fact 1. *If* $\underline{X} \sim N_2(\underline{\mu}, \Sigma)$ *with* $\Sigma = BB^T$*, then:*

$$\underline{Y} = B^{-1}(\underline{X} - \underline{\mu}) \sim N_2(\underline{0}, I_{2\times 2})$$

equivalently:

$$\underline{Z} = B^{-1}(\underline{X} - \underline{\mu}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

7 Multivariate Normal Distribution

A random vector $\underline{X} \in \mathbb{R}^d$ is said to have a **multivariate normal distribution** with mean vector $\underline{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{d\times d}$ (symmetric and positive definite), denoted $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$, if its joint PDF is:

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1}(\underline{x} - \underline{\mu})\right), \quad \underline{x} \in \mathbb{R}^d$$

Special case: Standard multivariate normal with $\underline{\mu} = \underline{0}$ and $\Sigma = I_{d\times d}$:

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\underline{x}^T\underline{x}\right) \\ &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\underline{x}\|^2\right) \end{aligned}$$

Theorem 4. *If* $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$ *with* $\Sigma = BB^T$*, then:*

$$\underline{Y} = B^{-1}(\underline{X} - \underline{\mu}) \sim N_d(\underline{0}, I_d)$$

meaning the components of $\underline{Y}$ *are i.i.d.* $N(0, 1)$*.*

Remark 1. *There is no trivariate (or multivariate) interaction. Also, it is not possible to have multivariate normal random variables such that the components are pairwise independent but they are not independent.*

Theorem 5. *Let* $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$ *and* A *be a* $k \times d$ *matrix. Then:*

$$\underline{Y} = A\underline{X} \sim N_k(A\underline{\mu}, A\Sigma A^T)$$

7.1 Parameter Counting

Bivariate $(d = 2)$:

- 2 means (μ_1, μ_2)
- 2 variances (σ_1^2, σ_2^2)
- 1 correlation (ρ)
- Total = 5 parameters

Trivariate $(d = 3)$:

- 3 means (μ_1, μ_2, μ_3)
- 3 variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$
- 3 correlations $(\rho_{12}, \rho_{13}, \rho_{23})$
- Total = 9 parameters

Example 4 (Pairwise Independent but not Independent)**.** *Let* X, Y *be i.i.d. Bernoulli*$(1/2)$*. Define* $Z = X \oplus Y$ *(binary sum):*

$$\begin{aligned} 0 + 0 &= 0 \\ 0 + 1 = 1 + 0 &= 1 \\ 1 + 1 &= 0 \end{aligned}$$

Then X, Y, Z *are pairwise independent, but not independent.*

Theorem 6. *Suppose $X_1, \ldots, X_d$ are i.i.d. $N(0,1)$, i.e.,*

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} \sim N_d(\underline{0}, I_d)$$

Let $A_{d\times d}$ be an orthogonal matrix, i.e., $AA^T = A^TA = I_{d\times d}$. Then:

$$\underline{Y} = A\underline{X} \sim N_d(\underline{0}, I_d)$$

8 Marginal Distributions

Let $A_{k\times d}$ be a matrix with real entries and full row rank. Let $\underline{Y} = A\underline{X}$ where $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$. Then:

$$\underline{Y} \sim N_k(A\underline{\mu}, A\Sigma A^T)$$

Why? $A_{k\times d}$ is full row rank $\Rightarrow \exists$ a matrix $B = \begin{pmatrix} A \\ C \end{pmatrix}$ where C is $(d-k)\times d$ such that B is non-singular.

The rows of A are k linearly independent vectors in $\mathbb{R}^d$ and thus can be extended to a basis for $\mathbb{R}^d$ with $(d-k)$ "new" vectors which can be taken as the rows of C.

Let $\underline{Z} = B\underline{X} = \begin{pmatrix} A\underline{X} \\ C\underline{X} \end{pmatrix}$. From earlier discussion:

$$\underline{Z} \sim N_d(B\underline{\mu}, B\Sigma B^T)$$

Therefore:

$$\underline{Y} = A\underline{X} \sim N_k(A\underline{\mu}, A\Sigma A^T)$$

If $\underline{X} \sim N_d(\underline{\mu}, \Sigma)$ is partitioned as:

$$\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}, \quad \underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then $\underline{X}_1 \sim N_k(\underline{\mu}_1, \Sigma_{11})$.

8.1 Positive Definiteness of Submatrices

Σ is a p.d. matrix $\Rightarrow \Sigma_{11}$ is also p.d.

Proof. Take $\underline{x} \in \mathbb{R}^k$ and define $\underline{y} = \begin{pmatrix} \underline{x} \\ \underline{0} \end{pmatrix} \in \mathbb{R}^d$. Then:

$$\underline{x}^T\Sigma_{11}\underline{x} = \underline{y}^T\Sigma\underline{y} \geq 0$$

since Σ is p.d. □

9 Cochran's Theorem

Theorem 7 (Cochran's Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(0,1)$. Consider:*

$$T_1 = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{(sample mean)}$$

$$T_2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \textit{(sample variance)}$$

Then:

1. $T_1 \sim N(0, 1/n)$

2. T_2 *and* $(n-1)T_2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sim \chi^2_{(n-1)}$

3. T_1 *and* T_2 *are independent*

Proof Sketch. Let $\underline{X} = (X_1, \ldots, X_n)^T \sim N_n(\underline{0}, I_n)$. Let A be an $n \times n$ orthogonal matrix whose first row is $\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$.

Define $\underline{Y} = A\underline{X}$. Then $\underline{Y} \sim N_n(\underline{0}, I_n)$ because A is orthogonal, so the components $Y_1, \ldots, Y_n$ are i.i.d. $N(0, 1)$.

We have:

$$Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sqrt{n}\bar{X}_n$$

Since A is orthogonal, it preserves length:

$$\|\underline{Y}\|^2 = \|\underline{X}\|^2$$

Therefore:

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$$

This gives:

$$Y_1^2 + \sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$$

Rearranging:

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = (n-1)S_n^2$$

Now, $\bar{X}_n$ is a function of Y_1 only, and S_n^2 is a function of $Y_2, \ldots, Y_n$ only. Since $Y_1, \ldots, Y_n$ are independent, $\bar{X}_n$ and S_n^2 are independent.

Also, $\sum_{i=2}^{n} Y_i^2$ is a sum of $n-1$ squared i.i.d. $N(0, 1)$ variables, so it follows a $\chi^2_{(n-1)}$ distribution. □

Note: $\bar{X}_n$ is independent of the vector $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$. The vector of residuals has sum 0 and is thus confined to an $(n-1)$-dimensional subspace.

10 Inequalities

10.1 Markov's Inequality

Theorem 8 (Markov's Inequality)**.** *Let X be a non-negative random variable ($X \geq 0$). Then for any $\epsilon > 0$:*

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

The bound is determined only by the expected value and can be very loose.

Example 5. *Let $X \sim N(0, 1)$. Apply Markov's inequality to $|X|$:*

$$P(|X| > \epsilon) \leq \frac{E[|X|]}{\epsilon}$$

We compute:

$$E[|X|] = \int_{-\infty}^{\infty} |t| \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 2\int_{0}^{\infty} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \sqrt{\frac{2}{\pi}}$$

The exact probability is $P(|X| > \epsilon) = 2(1 - \Phi(\epsilon))$, so:

$$2(1 - \Phi(\epsilon)) \leq \frac{\sqrt{2/\pi}}{\epsilon}$$

Proof of Markov's Inequality. Let $I(X \geq \epsilon)$ be the indicator function. For $X \geq 0$:

$$\epsilon \cdot I(X \geq \epsilon) \leq X$$

Taking expectations:

$$\epsilon \cdot P(X \geq \epsilon) \leq E[X]$$

□

Digression: If $X \geq 0$ is a random variable with $E[X] = 0$, then by Markov's inequality, for any $\epsilon > 0$:

$$P(X > \epsilon) \leq \frac{E[X]}{\epsilon} = 0$$

This implies $P(X > 0) = P\left(\bigcup_{n=1}^{\infty}\{X > 1/n\}\right) \leq \sum_{n=1}^{\infty} P(X > 1/n) = 0$.
Therefore, $P(X = 0) = 1$. We say $X = 0$ **almost surely (a.s.)**.

10.2 Chebyshev's Inequality

Theorem 9 (Chebyshev's Inequality)**.** *Let X be a random variable with mean $\mu = E[X]$ and finite variance $\sigma^2 = E[(X-\mu)^2]$. Then for any $\epsilon > 0$:*

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

Proof. Let $Y = (X-\mu)^2$. Then Y is a non-negative random variable with $E[Y] = \sigma^2$. We have:

$$P(|X - \mu| \geq \epsilon) = P((X-\mu)^2 \geq \epsilon^2) = P(Y \geq \epsilon^2)$$

By Markov's inequality applied to Y:

$$P(Y \geq \epsilon^2) \leq \frac{E[Y]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

□

11 Weak Law of Large Numbers

Theorem 10 (Weak Law of Large Numbers (WLLN))**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Let:*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then $E[\bar{X}_n] = \mu$ and $Var(\bar{X}_n) = \frac{\sigma^2}{n}$.
Applying Chebyshev's inequality to $\bar{X}_n$:

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

As $n \to \infty$, $P(|\bar{X}_n - \mu| \geq \epsilon) \to 0$. This is the WLLN.

The WLLN states that the sample average converges in probability to the true mean. For Bernoulli(p) trials, it means the proportion of successes converges in probability to p. This gives a theoretical foundation for the frequentist interpretation of probability.

12 Almost Sure Convergence

12.1 Almost Surely

We say two random variables X and Y are **equal almost surely (a.s.)**, written $X = Y$ a.s., if $P(X \neq Y) = 0$.

If $X = Y$ a.s., they have the same distribution (CDF).

Example 6. *Let $X \sim N(0,1)$. Define:*

$$Y(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \neq 10 \\ 0 & \text{if } X(\omega) = 10 \end{cases}$$

Then $P(X \neq Y) = P(X = 10) = 0$ (since X is continuous), so $X = Y$ a.s.
The CDF of Y is:

$$P(Y \leq y) = P(Y \leq y, X = Y) + P(Y \leq y, X \neq Y) = P(X \leq y)$$

Thus, Y also has a standard normal distribution.

12.2 Strong Law of Large Numbers

Theorem 11 (Strong Law of Large Numbers (SLLN))**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = E[X_1]$. Then:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu \quad \text{as } n \to \infty$$

This means:

$$P\left(\left\{\omega \in \Omega \;\middle|\; \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) = \mu\right\}\right) = 1$$

SLLN implies WLLN. The SLLN requires only the existence of the mean, not the variance.

Definition 1. *A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to a random variable X* ***almost surely*** *($X_n \to X$ a.s.) if:*

$$P\left(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1$$

The event of convergence can be written as:

$$A = \{\omega \mid X_n(\omega) \to X(\omega)\} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega \mid |X_n(\omega) - X(\omega)| < 1/k\}$$

The complement event $A^c = \{\omega \mid X_n(\omega) \not\to X(\omega)\}$ means that there exists some k such that for all N, we can find an $n \geq N$ where $|X_n(\omega) - X(\omega)| \geq 1/k$.

12.3 Infinitely Often and Eventually

Definition 2. *Let $(A_n)_{n\geq 1}$ be a sequence of events.*

- *We say "A_n happens* ***infinitely often (i.o.)****" for the event:*

$$[A_n \text{ i.o.}] = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$$

- *We say "A_n happens* ***eventually****" for the event:*

$$[A_n \text{ eventually}] = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n$$

Fact: $[A_n \text{ i.o.}]^c = [A_n^c \text{ eventually}]$
With this notation, $X_n \to X$ a.s. if and only if for every $\epsilon > 0$:

$$P([|X_n - X| \geq \epsilon \text{ i.o.}]) = 0$$

13 Borel-Cantelli Lemmas

Theorem 12 (First Borel-Cantelli Lemma)**.** *Let (A_n) be a sequence of events. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.*

Theorem 13 (Second Borel-Cantelli Lemma)**.** *Let (A_n) be a sequence of* ***independent*** *events. If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \text{ i.o.}) = 1$.*

13.1 Application

Example 7. *Let X_n be i.i.d. $N(0,1)$. Show that $\frac{X_n}{n} \to 0$ a.s.*

We need to show $P\left(\left[\left|\frac{X_n}{n}\right| > \epsilon \text{ i.o.}\right]\right) = 0$ for all $\epsilon > 0$.

Let $A_n = \left[\left|\frac{X_n}{n}\right| > \epsilon\right]$. By the First Borel-Cantelli Lemma, it suffices to show $\sum_{n=1}^{\infty} P(A_n) < \infty$.

We have:

$$P(A_n) = P(|X_n| > n\epsilon) = P(|X_1| > n\epsilon)$$

since the X_n are identically distributed.

For a non-negative random variable Y, we have $E[Y] = \int_0^\infty P(Y > t)\,dt$. Let $Y = \frac{|X_1|}{\epsilon}$. Then:

$$\sum_{n=1}^{\infty} P(|X_1| > n\epsilon) = \sum_{n=1}^{\infty} P(Y > n) \leq E[Y] = \frac{1}{\epsilon} E[|X_1|]$$

Since $E[|X_1|] = \sqrt{2/\pi} < \infty$, the sum is finite. Thus, by the First Borel-Cantelli Lemma, $\frac{X_n}{n} \to 0$ a.s.

Theorem 14. *Let $X_1, X_2, \ldots$ be i.i.d. random variables. Then:*

$$\frac{X_n}{n} \to 0 \text{ a.s.} \quad \iff \quad E[|X_1|] < \infty$$

13.2 Proofs of Borel-Cantelli Lemmas

Proof of First Borel-Cantelli Lemma. Let $B_N = \bigcup_{n=N}^{\infty} A_n$. The sequence B_N is decreasing: $B_1 \supseteq B_2 \supseteq \ldots$

The event $[A_n \text{ i.o.}] = \bigcap_{N=1}^{\infty} B_N$.

By continuity of probability for decreasing events:

$$P([A_n \text{ i.o.}]) = \lim_{N\to\infty} P(B_N)$$

By the union bound:

$$P(B_N) = P\left(\bigcup_{n=N}^{\infty} A_n\right) \leq \sum_{n=N}^{\infty} P(A_n)$$

Therefore:

$$0 \leq P([A_n \text{ i.o.}]) \leq \lim_{N\to\infty} \sum_{n=N}^{\infty} P(A_n)$$

Since $\sum_{n=1}^{\infty} P(A_n) < \infty$, the tail of the series goes to zero:

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} P(A_n) = 0$$

Therefore, $P([A_n \text{ i.o.}]) = 0$. □

Proof of Second Borel-Cantelli Lemma. We want to show $P([A_n \text{ i.o.}]) = 1$, which is equivalent to $P([A_n \text{ i.o.}]^c) = 0$.

Note that $[A_n \text{ i.o.}]^c = [A_n^c \text{ eventually}]$.

Let $C_N = \bigcap_{n=N}^{\infty} A_n^c$. The sequence C_N is increasing to $[A_n^c \text{ eventually}]$, so:

$$P([A_n^c \text{ eventually}]) = \lim_{N\to\infty} P(C_N)$$

By independence:

$$P(C_N) = P\left(\bigcap_{n=N}^{\infty} A_n^c\right) = \prod_{n=N}^{\infty} P(A_n^c) = \prod_{n=N}^{\infty} (1 - P(A_n))$$

Using the inequality $1 - x \leq e^{-x}$ for $x \geq 0$:

$$P(C_N) \leq \prod_{n=N}^{\infty} e^{-P(A_n)} = \exp\left(-\sum_{n=N}^{\infty} P(A_n)\right)$$

Since $\sum_{n=1}^{\infty} P(A_n) = \infty$, the tail sum $\sum_{n=N}^{\infty} P(A_n) = \infty$ as well.

Thus, $P(C_N) \leq e^{-\infty} = 0$, which implies $\lim_{N\to\infty} P(C_N) = 0$. □

14 Conditional Distribution of Bivariate Normal

Let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$.

The conditional distribution of X_2 given $X_1 = x_1$ is also Normal with:

Conditional Mean:

$$E[X_2 \mid X_1 = x_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

This is the best linear predictor of X_2 given X_1.

Conditional Variance:

$$\mathrm{Var}(X_2 \mid X_1 = x_1) = \sigma_2^2(1 - \rho^2)$$

The variance is reduced by a factor of $(1 - \rho^2)$ and does not depend on the value x_1.

15 Summary of Key Results

- **Multivariate Normal Distribution**: Definition via PDF, mean vector $\underline{\mu}$, covariance matrix Σ.
- **Properties:**
  - Linear transformations of normal vectors are normal
  - Marginal distributions are normal
  - For jointly normal variables, zero covariance implies independence
  - Conditional distributions are normal
- **Cochran's Theorem**: For i.i.d. normal samples, the sample mean and sample variance are independent.
- **Inequalities**: Markov's and Chebyshev's.
- **Laws of Large Numbers:**
  - WLLN: $\bar{X}_n \to \mu$ in probability
  - SLLN: $\bar{X}_n \to \mu$ almost surely
- **Borel-Cantelli Lemmas**: Tools for proving a.s. convergence.
- **Convergence in Probability vs Almost Sure Convergence**: Almost sure convergence implies convergence in probability.