Chapter 4 : Expectation and Moments

Dr. Salim El Rouayheb *Scribe: Serge Kas Hanna, Lu Liu, Ghadir Ayache*

1 Expected Value of a Random Variable

Example 1.

Definition 1. *The expected or average value of a random variable X is defined by,*

1. $E[X] = \mu_X = \sum_i x_i P_X(x_i)$, *if X is discrete.*
2. $E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$, *if X is continuous.*

Example 2. *Let $X \sim Poisson(\lambda)$. What is the expected value of X?*
The PMF of X is given by,

$$Pr(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, \dots, \ .$$

Therefore,

$$\begin{aligned} E[X] &= \sum_{k=0}^{+\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{+\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{+\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda. \end{aligned}$$

Theorem 1. *(Linearity of Expectation)*
Let X and Y be any two random variables and let a and b be constants, then,

$$E[aX + bY] = aE[X] + bE[Y].$$

Example 3. *Let $X \sim Binomial(n, p)$. What is the expected value of X?*
The PMF of X is given by,

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \dots, n.$$

Therefore,

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

Rather than evaluating this sum, an easier way to calculate $E[X]$ is to express X as the sum of n independent Bernoulli random variables and apply Theorem 1. In fact,

$$X = X_1 + X_2 + \ldots + X_n.$$

Where $X_i \sim Bernoulli(p)$, for all $i = 1, \ldots, n$. Hence,

$$E[X] = E[X_1 + \ldots + X_n].$$

$$X_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}.$$

Therefore,

$$E[X_i] = 1 \times p + 0 \times (1-p) = p.$$

By linearity of expectation (Theorem 1),

$$\begin{aligned} E[X] &= E[X_1] + \ldots + E[X_n] \\ &= np. \end{aligned}$$

Theorem 2. *(Expected value of a function of a RV)*
Let X be a RV. For a function of a RV, that is, $Y = g(X)$, the expected value of Y can be computed from,

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

Example 4. *Let $X \sim N(\mu, \sigma^2)$ and $Y = X^2$. What is the expected value of Y?*
Rather than calculating the pdf of Y and afterwards computing $E[Y]$, we apply Theorem 2:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \mu^2 + \sigma^2. \end{aligned}$$

2 Conditional Expectations

Definition 2. *The conditional expectation of X given that the event B was observed is:*

- ***For X discrete:*** $E[X|B] = \sum_i x_i P_{X|B}(x_i|B)$.
- ***For X continuous:*** $E[X|B] = \int_{-\infty}^{+\infty} x f_{X|B}(x|B) dx.$.

Intuition: Think of $E[X|Y = y]$ as the best estimate (guess) of X given that you observed Y.

Example 5. *A fair die is tossed twice. Let X be the number observed after the first toss, and Y be the number observed after the second toss.*
Let $Z = X + Y$.

1. *Calculate $E[X]$.*

$$E[X] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

2. *Calculate* $E[X|Y=3]$.

$$E[X|Y=3] = E[X] = 3.5.$$

3. *Calculate* $E[X|Z]$.

Z	Pr (Z = 2)	E[X/Z]
2	1/36	1
3	2/36	1.5
4	3/36	2
5	4/36	2.5
6	5/36	3
7	6/36	3.5
8	5/35	4
9	4/36	4.5
10	3/36	5
11	2/36	5.5
12	1/36	6

Observation: $E[X|Z]$ is a random variable.

Theorem 3. *(Towering Property of Conditional Expectation)*
Let X and Y be two random variables, then,

$$E_Y\left[E_X[X|Y]\right] = E_X[X].$$

Proof:

$$\begin{aligned}
E_Y\left[E_X\left[X/Y\right]\right] &= \sum_{Z=z} P\left(Z=z\right) \sum_{X=x} x.P\left(X=x/Z=z\right) \\
&= \sum_{Z=z} \sum_{X=x} xP\left(Z=z\right) P\left(X=x/Z=z\right) \\
&= \sum_{X=x} \sum_{Z=z} xP\left(X=x,\, Z=z\right) \\
&= \sum_{X=x} x \sum_{Z=z} P\left(X=x,\, Z=z\right) \\
&= \sum_{X=x} xP\left(X=x\right) \\
&= E\left[X\right]
\end{aligned}$$

Example 6. *Let X and Y be two zero mean jointly gaussian random variables, that is,*

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

Where $|\rho| \leq 1$. Calculate $E[X|Y=y]$.

$$\begin{aligned}
E[X|Y=y] &= \int_{-\infty}^{+\infty} x f_{X|Y}(x|Y=y)dx. \\
f_{X|Y}(X|Y=y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
&= \frac{\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left[-\frac{x^2+y^2-2\rho xy}{2\sigma^2(1-\rho^2)}\right]}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2}\right]} \\
&= \frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left[-\frac{x^2+y^2-2\rho xy-(1-\rho^2)y^2}{2\sigma^2(1-\rho^2)}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left[-\frac{(x-\rho y)^2}{2\sigma^2(1-\rho^2)}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
E[X|Y=y] &= \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left[-\frac{(x-\rho y)^2}{2\sigma^2(1-\rho^2)}\right] dx \\
&= \rho y.
\end{aligned}$$

Remark 1. *If $\rho = 0 \Rightarrow X$ and Y are independent,*

$$E[X|Y=y] = E[X] = 0.$$

Remark 2. *For gaussian random variables,*

$$\begin{aligned}
E[X|Y] = \rho Y. \\
E[Y|X] = \rho X.
\end{aligned}$$

Example 7. *A movie, of size N bits, is downloaded through a binary erasure channel. Where $N \sim Poisson(\lambda)$. Let K be the number of received bits.*

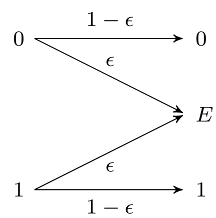


Figure 1: Binary Erasure Channel

1. *Calculate $E[K]$.*
   Intuitively $E[K] = \lambda(1-\epsilon)$. Now we will prove it mathematically by conditioning on N as a first step. For a given number of bits $N = n$, K is a binomial random variable ($K \sim Binomial(n, 1-\epsilon)$). Therefore,

$$E[K|N = n] = n(1-\epsilon).$$
$$E[K|N] = N(1-\epsilon).$$

   Applying the towering property of conditional expectation,

$$\begin{aligned} E[K] &= E\left[E[K|N]\right] \\ &= E[N(1-\epsilon)] \\ &= (1-\epsilon)E[N] \\ &= \lambda(1-\epsilon). \end{aligned}$$

2. *Calculate $E[N|K]$.*
   Intuitively $E[N|K] = k + \lambda\epsilon$. Now we will prove it mathematically,

$$\begin{aligned} E[N|K = k] &= \sum_{n=0}^{+\infty} nPr(N|K = k). \\ Pr(N|K = k) &= \frac{Pr(N = n, K = k)}{Pr(K = k)} \\ &= \frac{Pr(N = n)Pr(K = k|N = n)}{Pr(K = k)}. \\ Pr(N = n) &= \frac{\lambda^n e^{-\lambda}}{n!}. \\ Pr(K = k|N = n) &= \binom{n}{k}(1-\epsilon)^k \epsilon^{n-k}. \\ Pr(K = k) &= \sum_{n=k}^{+\infty} Pr(N = n)Pr(K = k|N = n) \\ &= \sum_{n=k}^{+\infty} \frac{e^{-\lambda}\lambda^n}{n!}\binom{n}{k}(1-\epsilon)^k \epsilon^{n-k}. \end{aligned}$$

$$
\begin{aligned}
Pr(N=n|K=k) &= \frac{\frac{e^{-\lambda}\lambda^n}{n!}\frac{n!}{k!(n-k)!}(1-\epsilon)^k\epsilon^{n-k}}{\sum_{n=k}^{+\infty}\frac{e^{-\lambda}\lambda^n}{n!}\frac{n!}{k!(n-k)!}(1-\epsilon)^k\epsilon^{n-k}} \\
&= \frac{\frac{\lambda^n\epsilon^n}{(n-k)!}}{\lambda\epsilon\sum_{n=k}^{+\infty}\frac{(\lambda\epsilon)^{n-k}}{(n-k)!}} \\
&= \frac{(\lambda\epsilon)^n}{(n-k)!}\frac{1}{(\lambda\epsilon)^k e^{\lambda\epsilon}} \\
&= \frac{(\lambda\epsilon)^{n-k}e^{-\lambda\epsilon}}{(n-k)!}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E[N|K=k] &= \sum_{n=k}^{+\infty} n\frac{(\lambda\epsilon)^{n-k}e^{-\lambda\epsilon}}{(n-k)!} \\
&= \sum_{n=k}^{+\infty}(n-k+k)\frac{(\lambda\epsilon)^{n-k}e^{-\lambda\epsilon}}{(n-k)!} \\
&= \underbrace{\sum_{n=k}^{+\infty}\frac{(n-k)(\lambda\epsilon)^{n-k}e^{-\lambda\epsilon}}{(n-k)!}}_{\lambda\epsilon} + k\underbrace{\sum_{n=k}^{+\infty}\frac{(\lambda\epsilon)^{n-k}e^{-\lambda\epsilon}}{(n-k)!}}_{1} \\
&= k+\lambda\epsilon.
\end{aligned}
$$

3 Moments of Random Variables

Definition 3. *The r^{th} moment, $r = 0, 1, \dots,$ of a RV X is defined by,*

1. $E[X^r] = m_r = \sum_i x_i^r P_X(x_i)$, *if X is discrete.*
2. $E[X^r] = m_r = \int_{-\infty}^{+\infty} x^r f_X(x)dx$, *if X is continuous.*

Remark 3. *Note that $m_0 = 1$ for any X, and $m_1 = E[X] = \mu$ (the mean).*

Definition 4. *The r^{th} central moment, $r = 0, 1, \dots,$ of a RV X is defined as,*

1. $E[(X-\mu)^r] = c_r = \sum_i (x_i-\mu)^r P_X(x_i)$, *if X is discrete.*
2. $E[(X-\mu)^r] = c_r = \int_{-\infty}^{+\infty}(x-\mu)^r f_X(x)dx$, *if X is continuous.*

Remark 4. *Note that for any RV X,*

1. $c_0 = 1$.
2. $c_1 = E[X-\mu] = E[X]-\mu = \mu-\mu = 0$.

3. $c_2 = E[(X-\mu)^2] = \sigma^2 = Var[X]$ *(the variance). In fact,*

$$\sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2.$$

σ *is called the standard deviation.*

Example 8. *Let $X \sim Binomial(n,p)$. What is the variance of X?*
The PMF of X is given by,

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

$E[X] = np$ *(from Example 2). Therefore,*

$$\sigma^2 = E[X^2] - E[X]^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - n^2p^2.$$

Check the textbook for the calculation of this sum. Here we will calculate σ^2 using the same idea of Example 2, i.e. expressing X as the sum of n independent Bernoulli random variables. In fact,

$$X = X_1 + X_2 + \ldots + X_n.$$

Where $X_i \sim Bernoulli(p)$, for all $i = 1, \ldots, n$. Hence,

$$\begin{aligned} E[X^2] &= E\left[(X_1 + \ldots + X_n)^2\right] \\ &= E[X_1^2 + \ldots + X_n^2 + \sum_i \sum_{\substack{j \\ j \neq i}} X_i X_j] \\ &= nE[X_i^2] + n(n-1)E[X_iX_j] \\ &= nE[X_i^2] + n(n-1)E[X_i][X_j]. \end{aligned}$$

Where,

$$\begin{aligned} E[X_i] &= p. \\ E[X_i^2] &= 1^2 \times p + 0^2 \times (1-p) = p. \end{aligned}$$

Hence,

$$E[X^2] = np + n(n-1)p^2 = n^2p^2 - np^2 + np.$$

Therefore,

$$\begin{aligned} \sigma^2 &= E[X^2] - E[X]^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p). \end{aligned}$$

Example 9. *Let $X \sim Geometric(p)$. What is the variance of X?*
The PMF of X is given by,

$$Pr(X = k) = (1-p)^{k-1}p, \ k = 1, 2, \ldots$$

The variance of X is given by,

$$\sigma^2 = E[X^2] - E[X]^2.$$

$$E[X] = \sum_{k=1}^{+\infty} k(1-p)^{k-1}p.$$

$$E[X^2] = \sum_{k=1}^{+\infty} k^2(1-p)^{k-1}p.$$

To calculate these sums we use the following facts, for $|x| < 1$,

$$\sum_{i=0}^{+\infty} x^i = \frac{1}{1-x}$$

Deriving both sides with respect to x,

$$\sum_{i=1}^{+\infty} ix^{i-1} = \frac{1}{(1-x)^2}$$

Deriving both sides with respect to x,

$$\sum_{i=2}^{+\infty} i(i-1)x^{i-2} = \frac{2}{(1-x)^3}$$

Hence,

$$\begin{aligned}
\sum_{i=2}^{+\infty} i^2x^{i-2} - \sum_{i=2}^{+\infty} ix^{i-2} &= \frac{2}{(1-x)^3} \\
\sum_{i=2}^{+\infty} i^2x^{i-2} &= \sum_{i=2}^{+\infty} ix^{i-2} + \frac{2}{(1-x)^3} \\
\sum_{i=1}^{+\infty}(i+1)^2x^{i-1} &= \sum_{i=1}^{+\infty}(i+1)x^{i-1} + \frac{2}{(1-x)^3} \\
\sum_{i=1}^{+\infty} i^2x^{i-1} &= \frac{2}{(1-x)^3} - \sum_{i=1}^{+\infty} ix^{i-1} \\
&= \frac{2}{(1-x)^3} - \frac{1}{(1-x)^2}.
\end{aligned}$$

Hence,

$$\sum_{i=1}^{+\infty} i^2x^{i-1} = \frac{1+x}{(1-x)^3}.$$

Therefore,

$$E[X] = \frac{p}{(1-(1-p))^2} = \frac{1}{p}.$$

$$E[X^2] = \frac{p\,(1+(1-p))}{(1-(1-p))^3} = \frac{2-p}{p^2}.$$

Therefore,

$$\begin{aligned}\sigma^2 &= E[X^2] - E[X]^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}.\end{aligned}$$

Definition 5. *The covariance of two random variables X and Y is defined by,*

$$cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - \mu_X\mu_Y.$$

Definition 6. *The correlation coefficient of two random variables X and Y is defined by,*

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

If $\rho_{X,Y} = 0$, then $cov(X,Y) = 0$ and X and Y are said to be uncorrelated.

Lemma 1. *If X and Y are independent $\Rightarrow$ X and Y are uncorrelated.*

Remark 5. *There could be two RVs which are uncorrelated but dependent.*

Example 10. *(discrete case: uncorrelated $\not\Rightarrow$ independent)*
Consider two random variables X and Y with joint PMF $P_{X,Y}(x_i, y_j)$ as shown.

	$x_1 = -1$	$x_2 = 0$	$x_3 = +1$
$y_1 = 0$	0	$\frac{1}{3}$	0
$y_2 = 1$	$\frac{1}{3}$	0	$\frac{1}{3}$

Figure 2: Values of $P_{X,Y}(x_i, y_j)$.

$$\begin{aligned}\mu_X &= -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0. \\ \mu_Y &= 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}.\end{aligned}$$

$$XY = \begin{cases} -1 & \textit{with probability } 1/3, \\ 0 & \textit{with probability } 1/3, \\ 1 & \textit{with probability } 1/3. \end{cases}$$

Hence,

$$E[XY] = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0.$$

Therefore,

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E[XY] - \mu_X\mu_Y}{\sigma_X\sigma_Y} = \frac{0}{\sigma_X\sigma_Y} = 0 \Rightarrow X \textit{ and } Y \textit{ are uncorrelated.}$$

However, X and Y are dependent. For example $P(X = -1|Y = 0) = 0 \neq P(X = -1) = 1/3$.

Example 11. *(continuous case: uncorrelated $\nRightarrow$ independent)*
Consider a RV Θ uniformly distributed on $[0, 2\pi]$. Let $X = \cos\Theta$ and $Y = \sin\Theta$.
X and Y are obviously dependent, in fact,

$$X^2 + Y^2 = 1.$$

$$\begin{aligned} E[X] = E[\cos\Theta] &= \frac{1}{2\pi}\int_0^{2\pi} \cos\theta d\theta = 0. \\ E[Y] = E[\sin\Theta] &= \frac{1}{2\pi}\int_0^{2\pi} \sin\theta d\theta = 0. \\ E[XY] = E[\sin\Theta\cos\Theta] &= \frac{1}{2\pi}\int_0^{2\pi} \sin\theta\cos\theta d\theta = \frac{1}{4\pi}\int_0^{2\pi} \sin 2\theta d\theta = 0. \end{aligned}$$

Hence,

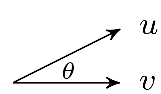
$$\rho_{X,Y} = cov(X, Y) = 0.$$

Therefore, X and Y are uncorrelated although they are dependent.

Theorem 4. *Given two random variables X and Y, $|\rho_{X,Y}| \leq 1$.*

Proof. We will prove this theorem using the Cauchy-Schwarz inequality by showing that,

$$|cov(X, Y)| \leq \sigma_X \sigma_Y.$$



$$< u, v >= ||u||.||v|| \cos\theta \Rightarrow |< u, v >| \leq ||u||.||v|| \ .$$

Using the same idea on random variables, let $Z = Y - aX$ where $a \in \mathbb{R}$. Assume that X and Y have zero mean. Consider the following 2 cases:

1. $Z \neq 0$ for all $a \in \mathbb{R}$. Hence,
$$E[Z^2] = E\left[(Y - aX)^2\right] > 0.$$
Where,
$$E\left[(Y - aX)^2\right] = E[Y^2 - 2aXY + a^2X^2] = E[X^2]a^2 - 2E[XY]a + E[Y^2].$$
Since $E[Z^2] > 0$ and $a^2 \geq 0$,
$$\Delta = 4E[XY]^2 - 4E[X^2]E[Y^2] < 0.$$
Hence,
$$\begin{aligned} &E[XY]^2 < E[X^2]E[Y^2] = \sigma_X^2\sigma_Y^2, \\ &|E[XY]| < \sigma_X\sigma_Y. \end{aligned}$$
Therefore,
$$|cov(X, Y)| < \sigma_X\sigma_Y.$$

2. There exists $a_0 \in \mathbb{R}$ such that $Z = Y - a_0 X = 0$.
   $Y = a_0 X$, hence,

$$\begin{aligned} E[XY] &= E[a_0 X^2] \\ &= a_0 E[X^2] \\ &= a_0 \sigma_X^2. \\ Var[Y] &= Var[a_0 X] \\ &= a_0^2 Var[X] \\ &= a_0^2 \sigma_X^2 \\ &= \sigma_Y^2. \end{aligned}$$

   Therefore,

$$E[XY] = a_0 \sigma_X \sigma_X = \sigma_X \sigma_Y.$$

   Therefore there is equality in this case,

$$|cov(X, Y)| = \sigma_X \sigma_Y.$$

Combining the results of the 2 cases,

$$|cov(X, Y)| \leq \sigma_X \sigma_Y.$$

And therefore,

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

□

Lemma 2. $Var[X + Y] = Var[X] + Var[Y] + 2cov(X, Y)$.

Proof.

$$\begin{aligned} Var[X + Y] &= E[(X + Y)^2] - E[(X + Y)]^2 \\ &= E[(X + Y)^2] - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ &= Var[X] + Var[Y] + 2cov(X, Y). \end{aligned}$$

□

Lemma 3. *If X and Y are uncorrelated, then $Var[X + Y] = Var[X] + Var[Y]$.*

Proof. X and Y are correlated $\implies cov(X, Y) = 0$. Result then follows for Lemma 2. □

4 Jointly Gaussian Random Variables

Definition 7. *Two random variable X and Y are jointly gaussian if,*

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

Properties:

1. If X and Y are jointly gaussian random variables,

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}.$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_y)^2}{2\sigma_Y^2}}.$$

2. If X and Y are jointly gaussian uncorrelated random variables $\Rightarrow$ X and Y are independent.

3. If X and Y are jointly gaussian random variables then any linear combination $Z = aX + bY$ is a gaussian random variable.

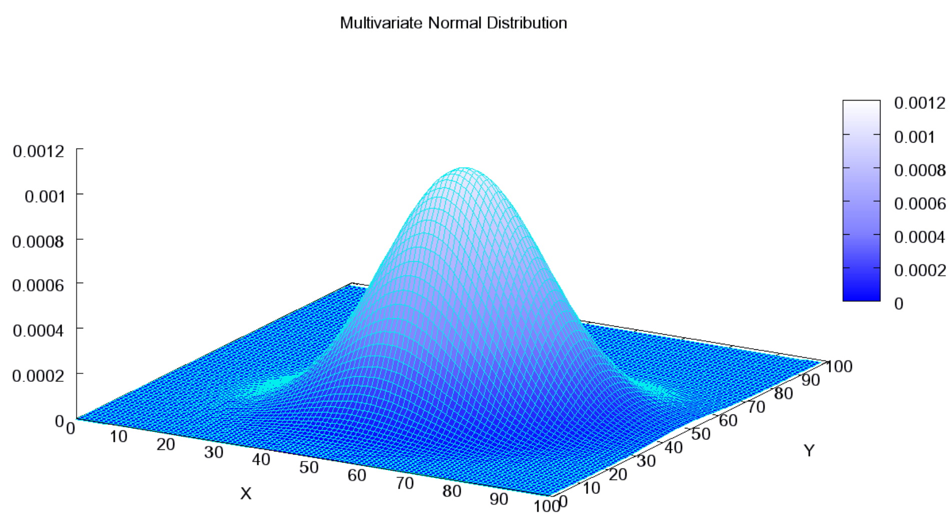


Figure 3: Two-variable joint gaussian distribution (from Wikipedia).

5 Bounds

Theorem 5. *(Chebyshev's Bound)*
Let X be an arbitrary random variable with mean μ and finite variance σ^2. Then for any $\delta > 0$,

$$Pr(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}.$$

Proof.

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx \geq \int_{|x-\mu|\geq\delta} (x-\mu)^2 f_X(x)dx \geq \delta^2 \int_{|x-\mu|\geq\delta} f_X(x)dx = \delta^2 Pr(|X-\mu| \geq \delta).$$

□

Corollary 1.

$$Pr(|X - \mu| < \delta) \geq 1 - \frac{\sigma^2}{\delta^2}.$$

Corollary 2.

$$Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Example 12. *A fair coin is flipped n times, let X be the number of heads observed. Determine a bound for $Pr(X \geq 75\%\ n)$.*

$$E[X] = np = \frac{n}{2}.$$
$$V[X] = np(1-p) = \frac{n}{4}.$$

By applying Chebyshev's inequality,

$$Pr(X \geq \frac{3}{4}n) = Pr(X - \frac{n}{2} \geq \frac{n}{4}) = \frac{1}{2}Pr(|X - \frac{n}{2}| \geq \frac{n}{4}) \leq \frac{1}{2}\frac{n/4}{n^2/4^2} = \frac{2}{n}.$$

Theorem 6. *(Markov Inequality)*
Consider a RV X for which $f_X(x) = 0$ for $x < 0$. Then X is called a nonnegative RV and the Markov inequality applies:

$$Pr(X \geq \delta) \leq \frac{E[X]}{\delta}.$$

Proof.

$$E[X] = \int_0^{+\infty} x f_X(x)dx \geq \int_\delta^{+\infty} x f_X(x)dx \geq \delta \int_\delta^{+\infty} x f_X(x)dx = \delta Pr(X \geq \delta).$$

□

Example 13. *Same setting as Example* **??**. *According to Markov inequality,*

$$Pr(X \geq \frac{3}{4}n) \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

which is not dependent on n. The bound from Chebyshev's inequality is much tighter.

Theorem 7. *(Law of Large Numbers)*
Let $X_1, X_2, \ldots, X_n$ be n iid RVs with mean μ and variance σ^2. Consider the sample mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

$$\lim_{n\to+\infty} Pr(|\hat{\mu}-\mu| \geq \delta) = 0 \ \forall \delta > 0.$$

We say that $\hat{\mu}$ converges in probability to μ,

$$\hat{\mu} \xrightarrow{\text{in probability}} \mu.$$

Proof.

$$E[\hat{\mu}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = E[X_i] = \mu.$$

Since X_i's are iid,

$$V[\hat{\mu}] = V[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n}V[X_i] = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality,

$$Pr(|\hat{\mu}-\mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2} \to 0 \text{ as } n \to +\infty \text{ for } \delta \leq \frac{1}{\sqrt{n}}.$$

□

6 Moment Generating Functions (MGF)

Definition 8. *The moment-generating function (MGF), if it exists, of an RV X is defined by*

$$\mathcal{M}(t) \triangleq E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\mathrm{d}x,$$

where t is a complex variable.

For discrete RVs, we can define $\mathcal{M}(t)$ using the PMF as

$$\mathcal{M}(t) = E[e^{tX}] = \sum_i e^{tx_i} P_X(x_i).$$

Example: Let $X \sim \text{Poisson}(\lambda)$, $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, 2, \ldots$.

1. Find $\mathcal{M}_x(t)$.

$$\mathcal{M}_X(t) = E(e^{t\lambda}) = \sum_{k=0}^{\infty} e^{tk} P(X=k)$$
$$\mathcal{M}_X(t) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!}$$
$$\mathcal{M}_X(t) = e^{-\lambda}e^{\lambda e^t}.$$

2. Find $E(X)$ from $\mathcal{M}_x(t)$.

$$E(X) = \frac{\partial \mathcal{M}_X(t)}{\partial t} \ |_{t=0} = \lambda e^t e^{\lambda(e^t-1)} \ |_{t=0} = \lambda.$$

Example: Let $X \sim N(\mu, \sigma^2)$, find $\mathcal{M}_x(t)$.

$$\begin{aligned} \mathcal{M}_X(t) &= \int_{-\infty}^{\infty} e^{xt} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ &= \cdots = e^{\mu t + \frac{\sigma^2 t^2}{2}} \end{aligned}$$

Lemma 4. *If $\mathcal{M}(t)$ exists, moments $m_k = E[X^k]$ can be obtained by*

$$m_k = \mathcal{M}^{(k)}(0) = \frac{d^k}{dt^k}(\mathcal{M}(t))\bigg|_{t=0}, \quad k = 0, 1, \ldots.$$

Proof.

$$\begin{aligned} \mathcal{M}_X(t) = E[e^{tX}] &= E\left[1 + tX + \frac{(tX)^2}{2!} + \cdots + \frac{(tX)^n}{n!} + \ldots\right] \\ &= 1 + E[X] + \frac{t^2}{2!}E[X^2] + \cdots + \frac{t^n}{n!}E[X^n] + \ldots \\ m_k = E[X^k] &= \frac{d^k}{dt^k}(\mathcal{M}(t))\bigg|_{t=0} \end{aligned}$$

□

Remark 6. *The MGF doesn't always exist. It is the case of Cauchy distribution for example where there is no closed form solution for the integral*

$$f_x(x) = \frac{\alpha . x}{(\alpha^2 + x^2)}$$

7 Chernoff Bound

In this section, we introduce the Chernoff bound. Recall that to use Markov's inequality X must be positive.

Theorem 8. *(Chernoff's bound) For any RV X,*

$$P(X \geq a) \leq e^{-at}\mathcal{M}_X(t) \ \forall \ t > 0.$$

In particular,

$$P(X \geq a) \leq \min_t e^{-at}\mathcal{M}_X(t).$$

Proof. Apply Markov on $Y = e^{tX}$, but first recall that $P(X \geq a) = P(e^{tX} \geq e^{ta}) = P(Y \geq e^{ta})$, by Markov we get

$$
\begin{aligned}
P(Y \geq e^{ta}) &\leq \frac{E(Y)}{e^{ta}} = e^{-ta}E(Y) \\
P(X \geq a) &\leq e^{-ta}\mathcal{M}_X(t)
\end{aligned}
$$

$\square$

Example: Consider $X \sim N(\mu, \sigma)$ and try to bound $P(X \geq a)$ using Chernoff bound, this is an artificial example because we know the distribution of X.

From last lecture $\mathcal{M}_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$ hence

$$P(X) \leq \min_t e^{-at} e^{\mu t + \frac{\sigma^2 t^2}{2}} = \min_t e^{(\mu - a)t + \frac{\sigma^2 t^2}{2}}$$

Remark: You can check at home how the parameter t can affect the outer bound. For example pick $\mu = 0, \sigma = 1$ and change t; for $t = 0$ you will get the trivial bound $P \leq 1$ and for $t \to \infty$ you will get $P \leq \infty$. See how it varies.

$$
\begin{aligned}
\min_t e^{(\mu - a)t + \frac{\sigma^2 t^2}{2}} &\Rightarrow \frac{\partial f(t)}{\partial t} = 0 \\
&\Rightarrow (\sigma^2 t + \mu - a)e = 0 \\
&\Rightarrow t^* = \frac{a - \mu}{\sigma^2}
\end{aligned}
$$

Which gives us the following:

$$
\begin{aligned}
P(X \geq a) &\leq e^{(\mu - a)t^* + \frac{\sigma^2 t^*}{2}} \\
P(X \geq a) &\leq e^{\frac{-(a-\mu)(\mu - a)}{\sigma^2} + \frac{\sigma^2 (a-\mu)^2}{2\sigma^4}} \\
P(X \geq a) &\leq e^{\frac{-(a-\mu)^2}{2\sigma^2}}
\end{aligned}
$$

We can compare this result with the reality where we know that $P(X \geq a) = \int_a^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$.

8 Characteristic Function

In this section, we define a characteristic function and give some examples. The characteristic function of a RV is similar to a Fourrier transform of a function without the '-'.

Definition 9. *X is a RV,*

$$\Phi_X(w) = E(e^{jwX}) = \int_{-\infty}^{+\infty} f_X(x) e^{jwx} dx, \tag{1}$$

is called the characteristic function of X where j is the complex number $j^2 = -1$.

Example: Find the characteristic function of $X \sim \exp(\lambda)$. Recall that for $\lambda \geq 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \Phi_X(w) &= \int_0^\infty \lambda e^{-\lambda x} e^{jwx} dx \\ &= \lambda \int_0^\infty e^{(jw-\lambda)x} dx \\ &= \frac{\lambda}{jw-\lambda} \left[e^{(jw-\lambda)x} \right]_0^\infty \end{aligned}$$

Since $\lambda \geq 0$ and jw is a unit quantity $\Rightarrow (jw - \lambda) \leq 0$ therefore $\lim_{x\to\infty} e^{(jw-\lambda)x} = 0$. Which results in

$$\begin{aligned} \Phi_X(w) &= \frac{\lambda}{jw-\lambda}(0-1) \\ \Phi_X(w) &= \frac{\lambda}{\lambda - jw} \end{aligned}$$

Lemma 5. *If $\Phi_X(w)$ exists, moments $m_n = E[X^n]$ can be obtained by*

$$m_n = \frac{1}{j^n} \Phi_X^{(n)}(0)$$

where

$$\Phi_X^{(n)}(0) = \frac{d^n}{dw^n} \Phi_X^{(w)} \bigg|_{w=0}.$$

Proof.

$$\begin{aligned} \Phi_X(w) &= E[e^{jwX}] \\ &= \sum_{n=0}^{\infty} \frac{(jw)^n}{n!} m_n \\ m_n &= \frac{1}{j^n} \left(\frac{d^n}{dw^n} \Phi_X^{(w)} \bigg|_{w=0} \right). \end{aligned}$$

□

Lemma 6. *if X, Y are two independent RV and $Z = X + Y$ then $\Phi_Z(w) = \Phi_X(w)\Phi_Y(w)$ and $\mathcal{M}_Z(t) = \mathcal{M}_X(t)\mathcal{M}_Y(t)$*

Remark: To find the distribution of $Z = X + Y$ it could be easier to find $\Phi_X(w)$, $\Phi_Y(w)$, multiply them and then invert the from "Fourrier" domain by integrating or by using tables of Fourrier inverse.

Example: Consider the example of problem 9 of homework 3:

Question: Let X_1 and X_2 be two independent RV such that $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$ and let $X = aX_1 + bX_2$. Find the distribution of X.

Answer: Let $X_1' = aX_1$, $X_2' = bX_2$ it is clear that $X_1' \sim N(a\mu_1, a\sigma_1)$ and $X_2' \sim N(b\mu_2, b\sigma_2)$ and that $\Phi_X(w) = \Phi_{X_1'}(w)\Phi_{X_2'}(w)$.

$$\Phi_{X_1'}(w) = \ldots = e^{a\mu_1 jw - \frac{a^2\sigma_1^2 w^2}{2}}$$
$$\Phi_X(w) = e^{a\mu_1 jw - \frac{a^2\sigma_1^2 w^2}{2}} e^{b\mu_2 jw - \frac{b^2\sigma_2^2 w^2}{2}}$$
$$\Phi_X(w) = e^{j(a\mu_1 + b\mu_2)w - (a^2\sigma_1^2 + b^2\sigma_2^2)\frac{w^2}{2}}$$

Which implies that $X \sim N(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2})$.

Fact 1. *A linear combination of two independent Gaussian RV is a Gaussian RV.*

9 Central Limit Theorem

In this section we state the central limit theorem and give a rigourous proof.

Theorem 9. *Let $X_1, X_2, \ldots, X_n$ be n independent RVs with $\mu_{X_i} = 0$ and $V(X_i) = 1$ $\forall i$ then*

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \underset{n\to\infty}{\to} N(0,1)$$

In other words

$$\lim_{n\to\infty} P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$$

This is for example a way to convert flipping a coin n times to a Gaussian RV (fig. **??**,fig. **??**).

$$X_i = \begin{cases} 0 & \text{if a tail is observed with } p = \frac{1}{2} \\ 1 & \text{if a head is observed with } 1 - p = \frac{1}{2} \end{cases}$$

And set $S_n = \frac{\sum_{i=0}^{n} X_i}{\sqrt{n}}$, notice that $S_n \in \{0, 1, \ldots, \frac{n}{\sqrt{n}}\}$ and according to CLT $S_n \underset{n\to\infty}{\sim} N(0,1)$.

CLT: says that no matter how far you are from the mean, the probability of $X = x$ being outside $|x - \mu| \leq \sqrt{n}$ decreases exponentially with n.

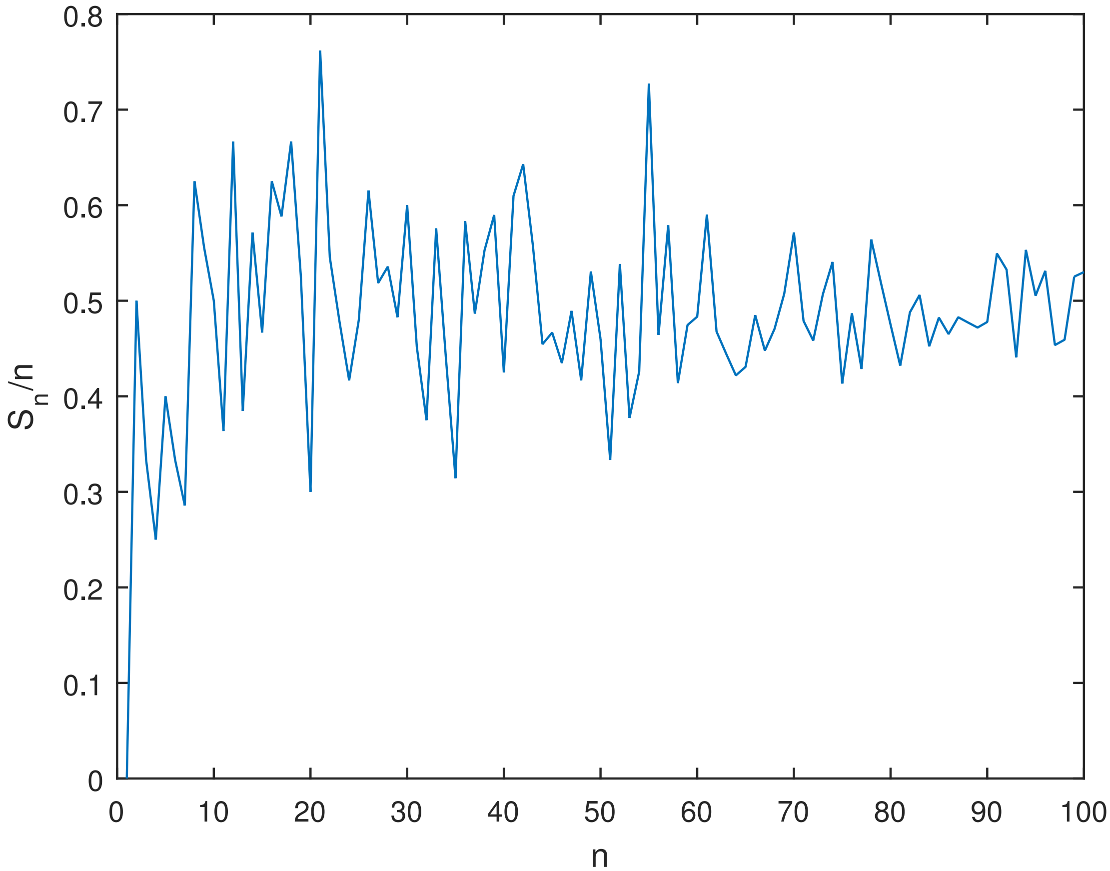


Figure 4: This is $\frac{S_n}{n}$ as a function of n, we can clearly see that when n grows $\frac{S_n}{n}$ goes to 0.5 for the equation of the example below for n goes to 100. Refer to section 6 for detailed code.

Remark: The RVs X_i have to be independent because if for example $X_i = X_1$ for $i \in \{2, 3, \dots, n\}$ then

$$S_n = nX_i = \begin{cases} \sqrt{n} & \text{if } X_1 = 1 \\ 0 & \text{if } X_1 = 0 \end{cases}$$

which does not converge to a Gaussian distribution when $n \to \infty$.

Proof. (of theorem **??**)

$$\lim_{n\to\infty} \Phi_{Z_n}(w) = e^{-\frac{w^2}{2}} \Rightarrow f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

where this form of $\Phi_{Z_n}(w)$ is the characteristic function of a $N(0, 1)$ RV.

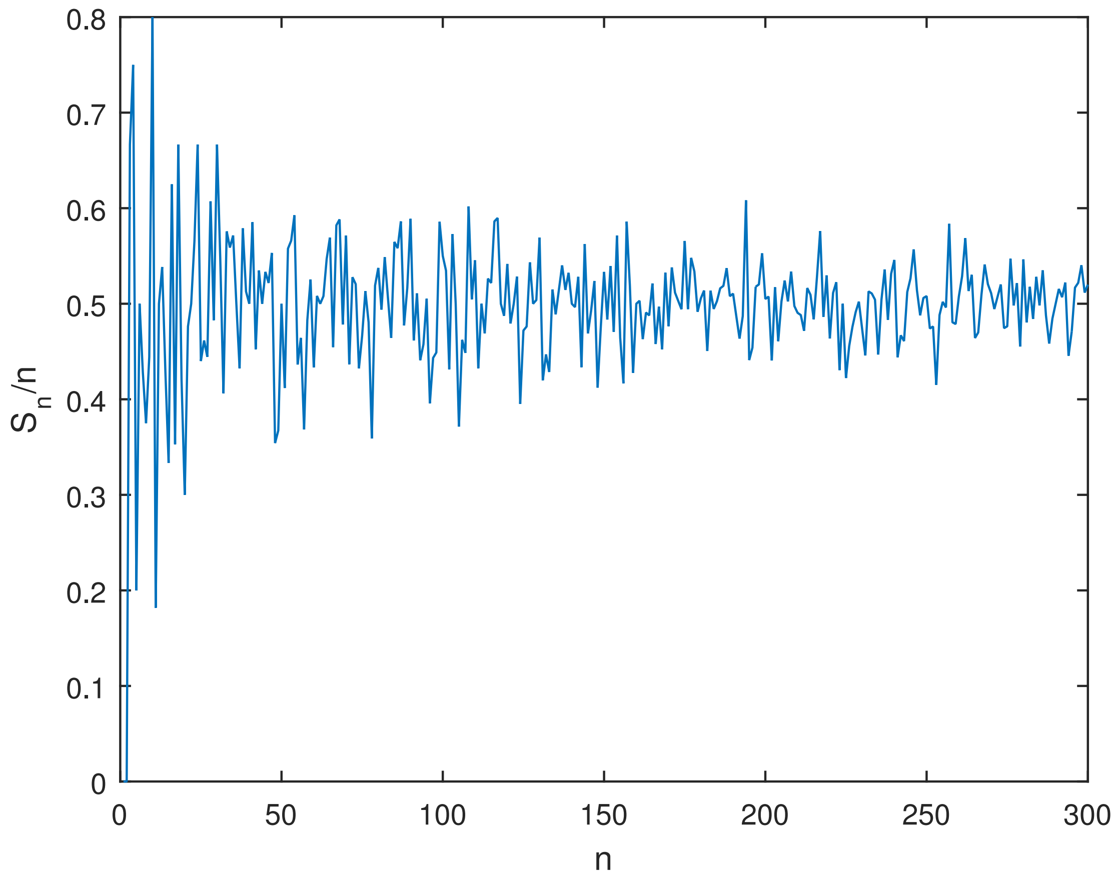


Figure 5: This is $\frac{S_n}{n}$ as a function of n, we can clearly see that when n grows $\frac{S_n}{n}$ goes to 0.5 for the equation of the example below for n goes to 300. Refer to section 6 for detailed code.

$$Z_n = \underbrace{\frac{X_1}{\sqrt{n}}}_{W_1} + \underbrace{\frac{X_2}{\sqrt{n}}}_{W_2} + \cdots + \underbrace{\frac{X_n}{\sqrt{n}}}_{W_n}$$

$$\Phi_{Z_n}(w) = \Phi_{W_1}(w)\Phi_{W_2}(w)\Phi_{W_3}(w)\ldots\Phi_{W_n}(w) = [\Phi_{W_1}(w)]^n$$

$$\Phi_{W_1}(w) = E(e^{wjW_1}) = E(e^{\frac{jwX_1}{\sqrt{n}}}) = \Phi_{X_1}(\frac{w}{\sqrt{n}})$$

Taylor expansion: Using the Taylor expansion of $\Phi_{W_1}(w)$ around 0 we get,

$$\Phi_{W_1}(w) = \Phi_{W_1}(0) + \Phi'_{W_1}(0) + \frac{\Phi''_{W_1}(0)}{2!} + \ldots$$

.

1. Find the value of $\Phi_{W_1}(0)$.

$$\begin{aligned}\Phi_{W_1}(0) &= E(e^{\frac{jwX}{\sqrt{n}}})\mid_{w=0} \\ &= \int_{-\infty}^{+\infty} e^{\frac{j0X}{\sqrt{n}}} f_X(x)dx \\ &= \int_{-\infty}^{+\infty} f_X(x)dx \\ &= 1\end{aligned}$$

2. Find the value of $\Phi'_{W_1}(0)$.

$$\begin{aligned}\Phi'_{W_1}(0) &= \int_{-\infty}^{+\infty} \frac{jx}{\sqrt{n}} e^{\frac{j0X}{\sqrt{n}}} f_X(x)dx \\ &= \frac{j}{\sqrt{n}} \int_{-\infty}^{+\infty} x f_X(x)dx \\ &= \frac{j}{\sqrt{n}} E(X_1) \\ &= 0\end{aligned}$$

3. Find the value of $\Phi''_{W_1}(0)$.

$$\begin{aligned}\Phi''_{W_1}(0) &= \frac{d^2\Phi_{W_1}(w)}{dw^2} \\ &= \int_{-\infty}^{+\infty} (\frac{jx}{\sqrt{n}})^2 e^{\frac{j0X}{\sqrt{n}}} f_X(x)dx \\ &= \frac{-1}{n} \int_{-\infty}^{+\infty} x^2 f_X(x)dx \\ &= \frac{-1}{n}(\underbrace{V(X)}_{1} + \underbrace{E^2(X)}_{0}) \\ &= \frac{-1}{n}\end{aligned}$$

Hence, using these results and Taylor's expansion, $\Phi_{W_1}(w) = 1 - \frac{w^2}{2n}$. Therefore $\Phi_{Z_n}(w) = [1 - \frac{w^2}{2n}]^n$.

Recall that $\log(1-\epsilon) \simeq -\epsilon$, then

$$\begin{aligned}
\log \Phi_{Z_n} &= n \log(1 - \frac{w^2}{2n}) \\
\log \Phi_{Z_n} &\simeq n(-\frac{w^2}{2n}) \\
\log \Phi_{Z_n} &\simeq -\frac{w^2}{2} \\
\Phi_{Z_n} &= e^{-\frac{w^2}{2}}
\end{aligned}$$

□

10 MATLAB Code generating the figures

In this section we give the MATLAB code used to generate fig. **??** and fig **??**.

```
A=[];B=[];    % generate two empty vectors
for i=1:100   % in this loop i stands for the number of times the coin is flipped
    A=[A,binornd(i,0.5)/i]; % at each iteration generate a binomial random number with
end                         % parameters n=i, p=0.5 and divide it by n to have (S_n)/n
for n=1:300
B=[B,binornd(n,0.5)/n];     % same as previous but repeat it 300 times
end

x1=[1:i];x2=[1:n];          % x1 and x2 are used to represent n in each figure

figure(1)
plot(x1,A);
hold on
plot(x1,0.5,'r','linewidth',2);
xlabel('n');
ylabel('S_n/n');

figure(2)
plot(x2,B);
hold on
plot(x2,0.5,'r','linewidth',2);
xlabel('n');
ylabel('S_n/n');
```