# Probability and Statistics

Omid Safarzadeh

January 30, 2022

# Table of contents



*Acknowledgement: This slide is prepared based on Casella and Berger, 2002

# Moments definition

## Definition 1.1

For each of integer n, the $n^{th}$ moment of $X$ is

$$\mu_n' = E[X^n].$$

The $n^{th}$central moment of $X$, $\mu_n$, is

$$\mu_n = E[(X - \mu)^n],$$

where $\mu = \mu_1' = E[X]$.

# Expected Value

- Recall that"average" is an arithmetic average where all available observations are weighted equally.
- The expected value, on the other hand, is the average of all possible values a random variable can take, weighted by the probability distribution.
- The question is,which value would we expect the random variable to take on, on average.

# Expected Value

## Definition 1.2

The expected value or mean of a random variable $g(X)$, denoted by $E[g(X)]$, is

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } \ X \text{ is continuous} \\ \\ \sum_{x\in\mathcal{X}} g(x)f_X(x) = \sum_{x\in\mathcal{X}} g(x)P(X=x) & \text{if } \ X \text{ is discrete} \end{cases}$$

If $E[g(X)] = \infty$, we say that $E[g(X)]$ does not exist.

- we are taking the average of $g(x)$ over all of its possible values ($x \in \mathcal{X}$), where these values are weighted by the respective value of the pdf, $f_X(x)$.

# Expected Value

## Example 1.1

Suppose $X$ has an exponential($\lambda$) distribution, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x < \infty \quad \lambda > 0.$$

Then,

$$E[X] = \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda} dx = -xe^{-x/\lambda}|_0^\infty + \int_0^\infty e^{-x/\lambda} dx \tag{1}$$

$$= \int_0^\infty e^{-2/\lambda} dx = \lambda. \tag{2}$$

- To obtain this result, we use a method called integration by parts. This is based on

$$\int u dv = uv - \int v du.$$

# Expected Value

- A very useful property of the expectation operator is that it is a linear operator.
- For example, consider some $X$ such that $E[X] = \mu$.
- Then, for two constants $a$ and $b$,

$$E[a + Xb] = a + E[Xb] = a + bE[x] = a + b\mu.$$

- Notice that, clearly, the expectation of a constant is equal to itself.

# Expected Value

## Theorem 1.1

Let $X$ be a random variable and let $a$, $b$ and $c$ be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

- $E[ag_1(X) + bg_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c$.
- If $g_1(x) \geq 0$ for all $x$, then $E[g_1(X)] \geq 0$.
- If $g_1(x) \geq g_2(x)$ for all $x$, then $E[g_1(X)] \geq E[g_2(X)]$.
- If $a \leq g_1(x) \leq b$ for all $x$, then $a \leq E[g(X)] \leq b$.

**Proof**: Exercise!

## Example 1.2

Let $X$ have a uniform distribution, such that

$$f_X(x) = \begin{cases} 1 & \text{if } \; 0 \leq x \leq 1 \\ 0 & \text{if } \; \text{otherwise} \end{cases}$$

Define $g(X) = -\log X$. Then,

$$E[g(X)] = E[-\log X] = \int_0^1 -\log x dx = (-x \log x + x)|_0^1 = 1,$$

where we use integration by parts.

# Variance

- variance measures the variation/dispersion/spread of the random variable around expectation.
- While the expectation is usually denoted by $\mu$, $\sigma^2$ is generally used for variance.
- Variance is a second-order moment.

# Variance

## Definition 1.3

The variance of a random variable $X$ is its second central moment,

$$Var(X) = E[(X - \mu)^2],$$

while $\sqrt{Var(X)}$ is known as the standard deviation of $X$.

- Importantly,
$$Var(X) = E[(X - \mu)^2] = E[X^2] - \mu^2.$$

## covariance

- When it exists, the covariance of two random variables $X$ and $Y$ is defined as

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

# Covariance and Correlation

- Let $X$ and $Y$ be two random variables. To keep notation concise, we will use the following notation.

$$E[X] = \mu_X, \quad E[Y] = \mu_Y, \quad Var(X) = \sigma_X^2 \quad and \quad Var(Y) = \sigma_Y^2.$$

## Definition 2.1

The covariance of $X$ and $Y$ is the number defined by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

## Definition 2.2

The correlation of $X$ and $Y$ is the number defined by

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_y},$$

which is also called the correlation coefficient.

# Covariance and Correlation

- If large(small) values of $X$ tend to be observed with large(small) values of $Y$, then will be positive.
- Why so? Within the above setting, when $X > \mu_X$ then $Y > \mu_Y$ is likely to be true whereas when $X < \mu_X$ then $Y < \mu_Y$ is likely to be true. Hence

$$E[(X - \mu_X)(Y - \mu_Y)] > 0.$$

- Similarly, if large(small) values of $X$ tend to be observed with small(large) values of $Y$, then $Cov(X, Y)$ will be negative.

# Covariance and Correlation

- Correlation normalises covariance by the standard deviations and is, therefore, a more informative measure.
- If Cov($X, Y$)=50 while Cov($W, Z$)=0.9, this does not necessarily mean that there is a much stringer relationship between $X$ and $Y$. For example, if Car($X$)=Var($Y$)=100 while Var($W$)=Var($Z$)=1, then

$$\rho_{XY} = 0.5 \quad \rho_{WZ} = 0.9.$$

**Theorem 2.1**

For any random variables $X$ and $Y$,

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

- **Proof**:

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y. \end{aligned}$$

**Theorem 2.2**

If $X \perp\!\!\!\perp Y$, then Cov$(X, Y)$= $\rho_{XY} = 0$.

- **Proof**: Since $X \perp\!\!\!\perp Y$, by Theorem (2.1), Then

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y = \mu_X \mu_Y - \mu_X \mu_Y = 0,$$

  and consequently,

$$\rho_{XY} = \frac{Cov(X, Y)}{\sum_X \sum_Y} = 0.$$

- It is crucial to note that although $X \perp\!\!\!\perp Y$ implies that $Cov(X, Y) = \rho_{XY} = 0$, the relationship does not necessarily hold in the reverse direction.

**Theorem 2.3**

If $X$ and $Y$ are any two random variables and $a$ and $b$ are any two constants, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

If $X$ and $Y$ are independent random variables, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

- **Proof**: Exercise!

# Covariance and Correlation

- Note that if two random variables, $X$ and $Y$, are positively correlated, then

$$Var(X+Y) > Var(X) + Var(Y),$$

whereas if $X$ and $Y$ are negatively correlated, then

$$Var(X+Y) < Var(X) + Var(Y).$$

- For positively correlated random variables, large values in one tend to be accompanied by large values in the other. Therefore, the total variance is magnified.
- Similarly, for negatively correlated random variables, large values in one tend to be accompanied by small values in the other. Hence, the variance of the sum is dampened.

# third and fourth moments

- third and fourth moments are concerned with how symmetric and fat-tailed the underlying distribution is.

# Moment Generating Functions

- moment generating function can be used to obtain moments of a random variable.

# Moments and Moment Generating Functions

## Definition 3.1

Let $X$ be a random variable with cdf $F_X$.The moment generating function (mgf)of $X$ (or $F_X$), denoted by $M_X(t)$, is

$$M_X(t) = E[e^{tX}],$$

provided that the expectation exists for $t$ in some neighbourhood of 0. That is, there is an $h > 0$ such that, for all $t$ in $-h < t < h$, $E[e^{tX}]$ exists.If the expectation does not exist in a neighbourhood of 0, we say that the mgf does not exist.

- We can write the mgf of $X$ as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \quad \text{if } X \text{ is continuous,}$$

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

# Moment Generating Functions

**Theorem 3.1**

If $X$ has mgf $M_X(t)$, then

$$E[X^n] = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

That is, the $n^{th}$ moment is equal to the $n^{th}$ derivative of $M_X(t)$ evaluated at t=0.

# Normal mgf

- Now consider the pdf for $X\ N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2], \qquad -\infty < x < \infty.$$

- The mgf is given by

$$M_X(t) = E[e^{Xt}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx.$$

# Normal mgf

- Then

$$M_X(t) = exp(\mu t + \frac{\sigma^2 t^2}{2}).$$

- Clearly,

$$E[X] = \frac{d}{dt} M_X(t)|_{t=0} = (\mu + \sigma^2 t) exp(\mu t + \frac{\sigma^2 t^2}{2})|_{t=0} = \mu,$$

$$\begin{aligned} E[X^2] &= \frac{d^2}{dt^2} M_X(t)|_{t=0} = \sigma^2 exp(\mu t + \frac{\sigma^2 t^2}{2}|_{t=0} \\ &+ (\mu + \sigma^2 t)^2 exp(\mu t + \frac{\sigma^2 t^2}{2})^2|_{t=0} \\ &= \sigma^2 + \mu^2, \\ Var(X) &= E[X^2] - \{E[X]\}^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2. \end{aligned}$$

# Discrete Uniform Distribution

- A random variable $X$ has a discrete *uniform*(1,N) distribution if

$$P(X = x|N) = \frac{1}{N}, \qquad x = 1, 2, ..., N,$$

where $N$ is a specified integer. This distribution puts equal mass on each of the outcomes $1, 2, ..., N$.

# Binomial Distribution

- This is based on a Bernoulli trial (after James Bernoulli which is an experiment with two, and only, two, possible outcomes.
- A random variable $X$ has Bernoulli($p$) distribution if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \qquad 0 \leq p \leq 1.$$

- $X = 1$ is often termed as "success" and $p$ is, accordingly, the probability of success. Similarly, $X = 0$ is termed a "failure".
- Now,

$$E[X] = 1 * p + 0 * (1-p) = p,$$
$$\text{and} \quad Var(X) = (1-p)^2 p + (0-p)^2 (1-p) = p(1-p).$$

- $E[X] = np$ (**Proof**: Exercise!)
- $Var(X) = np(1-p)$ (**Proof**: Exercise!)
- Examples:
  1. Tossing a coin ($p$ =probability of a head, $X = 1$ if heads)
  2. Roulette ($X = 1$ if red occurs, $p$ =probability of red)
  3. Election polls ($X = 1$ if candidate $A$ gets vote)
  4. Incidence of disease ($p$ =probability that a random person gets infected)

# Binomial Distribution

- We can extend the scope to a collection of many independent trials.
- Define

$$A_i = \{X = 1 \text{ on the } i^{th} \text{ trial}\}, \qquad i = 1, 2, ..., n.$$

- Assuming that $A_1, ..., A_n$ are independent events, we can derive the distribution of the total number of successes in $n$ trials.Define $Y =$"total number of successes in $n$ trials".
- The event $\{Y = y\}$ means that out of $n$ trials, $y$ resulted as success. Therefore, $n - y$ trials have been unsuccessful.
- In other words, exactly $y$ of $A_1, ..., A_n$ must have occurred.
- There are many possible orderings of the events that would lead to this outcome. Any particular such ordering has probability

$$p^y(1-p)^{n-y}.$$

- Since there are $\binom{n}{y}$ such sequences, we have

$$P(Y = y|n, p) = \binom{n}{y} p^y(1-p)^{n-y}, \qquad y = 0, 1, 2, ..., n,$$

and $Y$ is called a *binomial(n,p)* random variable.

# Binomial Distribution

## Example 4.1

**If you flip a fair coin 10 times, what is the probability of getting all tails?**

- Let's first calculate the probability of getting tail on fair coin when you flip it one time.

$$P(1) = \frac{1}{2} = 50\%(\text{Because coin has two sides, H \& T})$$

- Since all the trails are independent, probability of getting head on *nth* turn is also $1/2$.
- Then,

$$P(10) = \frac{1}{2} * \frac{1}{2} * ... * \frac{1}{2} \quad (10 \text{ times})$$
$$= (\frac{1}{2})^{10}.$$

# Poisson Distribution

- In modelling a phenomenon in which we are waiting for an occurrence (such as waiting for a bus), the number of occurrence in a given time interval can be modelled by the Poisson distribution.
- The basic assumption is as follows: for small time intervals, the probability of an arrival is proportional to the length of waiting time.
- If we are waiting for the bus, the probability that a bus will arrive within the next hour is higher than the probability that it will arrive within 5 minutes.
- Other possible applications are distribution of bomb hits in an area or distribution of fish in a lake.
- The only parameter is $\lambda$, also sometimes called the "intensity parameter."

# Poisson Distribution

- $E[X] = \lambda$
- $Var(X) = \lambda$
- **Proof**: Exercise!

# Poisson Distribution

## Example 4.2

As an example of a waiting-for-occurrence application, consider a telephone operator who, on average, handles fie calls every 3 minutes. What is the probability that there will be no calls in the next minute? At leas two calls? If we let $X$ =number of calls in a minute, then $X$ has a Poisson distribution with $E[X] = \lambda = 5/3$. So,

$$P(\text{no calls in the next minute}) = P(X = 0)$$

$$= \frac{e^{-5/3}(5/3)^0}{0!} = e^{-5/3} = 0.189$$

and

$$P(\text{at least two calls in the next minute}) = P(X \geq 2)$$

$$= 1 - P(X = 0) - P(X = 1)$$

$$= 1 - 0.189 - \frac{e^{-5/3}(5/3)^1}{1!}$$

$$= 0.496.$$

# Uniform Distribution

- The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } \; x \in [a,b] \\ 0 & \text{if } \; \text{otherwise} \end{cases}.$$

- One can easily show that

$$\int_a^b f(x)dx = 1,$$

$$E[X] = \frac{b+a}{2},$$

$$Var(X) = \frac{(b-a)^2}{12}.$$

- In many cases, when people say Uniform distribution, they implictly mean $(a, b) = (0, 1)$.

# Exponential Distribution

- Now consider, $\alpha = 1$ :

$$f(x|a,\beta) = f(x|1,\beta) = \frac{1}{\beta}e^{-e/\beta}, \quad 0 < x < \infty.$$

- Again, using our previous results, for $X$ exponential $(\beta)$ we have

$$E[X] = \beta \quad \text{and} \quad Var(X) = \beta^2$$

- A peculiar feature of this distribution is that it has no memory.

# Exponential Distribution

- If $X$ exponential($\beta$), then, for $s > t \geq 0$,

$$P(X > s|X > t) = \frac{P(X > s, X > t)}{P(X > t)} = \frac{P(X > s)}{P(X > t)}$$
$$= \frac{\int_s^\infty \frac{1}{\beta} e^{-x/\beta} dx}{\int_t^\infty \frac{1}{\beta} e^{-x/\beta} dx} = \frac{e^{-s/\beta}}{e^{-t/\beta}}$$
$$= e^{-(s-t)/\beta} = P(X > s - t).$$

- This is because,

$$\int_{s-t}^\infty \frac{1}{\beta} e^{-x/\beta} dx = -e^{-x/\beta}|_{x=s-t}^\infty = e^{-(s-t)/\beta}.$$

- What does this mean? When calculating $P(X > s|X > t)$, what matters is not whether $X$ has passed a threshold or not. What matters is the distance between the threshold and the value to be reached.
- If Mr X has been fired more than 10 times, what is the probability that he will be fired more than 12 times? It is not different from the probability that a person, who has been fired once, will be fired more than two times. History does not matter.

# Normal Distribution

- We now consider the normal distribution or the Gaussian distribution.
- Why is this distribution so popular?
  1. Analytical tractability
  2. Bell shaped or symmetric
  3. It is central to Central Limit Theorem; this type of results guarantee that, under (mild) conditions, the normal distribution can be used to approximate a large variety of distribution in large samples.
- The distribution has two parameters: mean and variance, denoted by $\mu$ and $\sigma^2$, respectively.
- The pdf is given by,

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp[-1/2 \frac{(x-\mu)^2}{\sigma^2}].$$

# Normal Distribution

- This distribution is usually denoted as $N(\mu, \sigma^2)$.
- A very useful result is that for $X\ N(\mu, \sigma^2)$,

$$Z = \frac{X - \mu}{\sigma}\ N(0, 1).$$

- $N(0, 1)$ is known as the standard normal distribution.
- To see this, consider the following:

$$P(Z \leq z) = P(\frac{(X - \mu)/\sigma \leq z}{)}$$

$$= P(X \leq z\sigma + \mu)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2}/2\sigma^2 dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt,$$

where we substitute $t = (x - \mu)/\mu$. Notice that this implies that $dt/dx = 1/\sigma$. This shows that $P(Z \leq z)$ is the standard normal cdf.

# Normal Distribution

- Then, we can do all calculations for the standard normal variable and then convert these results for whatever normal random variable we have in mind.
- Consider, for $Z$ $N(0, 1)$, the following:

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-z^2/2} dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} |_{-\infty}^{\infty} = 0.$$

- Then, to find $E[X]$ for $X$ $N(\mu, \sigma^2)$, we can use $X = \mu + Z\sigma$:

$$E[X] = E[\mu + Z\sigma] = \mu + \sigma E[Z] = \mu + \sigma * 0 = \mu.$$

- Similarly,

$$Var(X) = Var(\mu + Z\sigma) = \sigma^2 Var(Z) = \sigma^2.$$

- What about

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \stackrel{?}{=} 1.$$

# Lognormal Distribution

- Let $X$ be a random variable such that

$$\log X \ N(\mu, \sigma^2).$$

  Then, $X$ is said to have a lognormal distribution.

- By using a transformation argument (Theorem (1.2)), the pdf of $X$ is given by,

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} exp[-\frac{(\log x - \mu)^2}{2\sigma^2}],$$

  where $0 < x < \infty$, $-\infty < \mu, \infty$, and $\sigma > 0$.

- How? Take $W = \log X$. We start from distribution of $W$ and want to find the distribution of $X = expW$. Then, $g(W) = exp(W)$ and $g^{-1}(X) = log(X)$. The rest follows by using Theorem (1.2).

# Laplace distribution

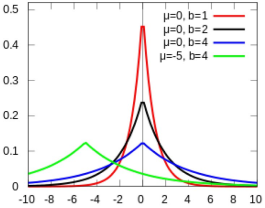
Laplace prior for $\beta$ :

$$\beta_j \sim Lap(0, \frac{2\sigma^2}{\lambda} \Rightarrow p(\beta_j) = \frac{\lambda}{4\sigma^2} exp(-\frac{\lambda}{2\sigma^2}|\beta_j|)$$

where $\beta_j, j = 1, ..., p$ are i.i.d

- If $Z \sim Lap(\mu, b)$, then $E[Z] = \mu$, $Var(Z) = 2b^2$,

$$p(Z) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b})$$




- **Likelihood**: Gaussian likelihood

# References

Casella, G., & Berger, R. (2002). *Statistical inference*. Cengage Learning.
https://books.google.fr/books?id=FAUVEAAAQBAJ