Lecture 3: Expectations, moments, and distributions

STATS 101: Foundations of Statistics

Linh Tran
ThetaHat.AI@gmail.com

December 5, 2019

Announcements

- Review survey
- Question: How long did last homework take?
- Next assignment will be posted tonight
  - Due 12/11 @ 11:59pm
- Classes will be recorded going forward

Outline

Expectations, moments, and distributions

- Expected value
- Moments
- Moment generating functions
- Distributions

Expectation

The *expected value* of rv X is defined as

$$\mathbb{E}[X] = \begin{cases} \sum_x x f_X(x) & \text{if x is discrete} \\ \int x f_X(x) dx & \text{if x is continuous} \end{cases} \tag{1}$$

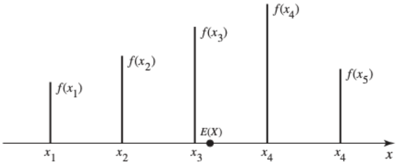
For functions g of X,

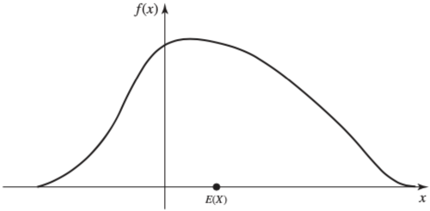
$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) f_X(x) & \text{if x is discrete} \\ \int g(x) f_X(x) dx & \text{if x is continuous} \end{cases} \tag{2}$$

n.b. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$

Expectation

Examples:





Expectation

Important: Expectations might not exist!

Example: Suppose $f_X(x) = \frac{1}{x^2}$, defined on $[1, \infty]$. Then

$$\mathbb{E}[X] = \int x f_X(x) dx = \int x \frac{1}{x^2} dx = \int \frac{1}{x} dx = \infty \tag{3}$$

Expectation

Important: Expectations might not exist!

Example: Suppose $f_X(x) = \frac{1}{x^2}$, defined on $[1, \infty]$. Then

$$\mathbb{E}[X] = \int x f_X(x) dx = \int x \frac{1}{x^2} dx = \int \frac{1}{x} dx = \infty \tag{3}$$

Some properties of expectations:

- Linearity: $\mathbb{E}[ag(X) + bh(X)] = \mathbb{E}[ag(X)] + \mathbb{E}[bh(X)]$
- Order preserving:
  $g(X) \leq h(X), \forall x \in \mathbb{R} \Rightarrow \mathbb{E}[g(X)] \leq \mathbb{E}[h(X)]$

Variance

The *variance* of rv X is defined as

$$var(X) = \mathbb{E}[(X-\mu)^2] : \mu = \mathbb{E}[X] \tag{4}$$

Variance

The *variance* of rv X is defined as

$$var(X) = \mathbb{E}[(X-\mu)^2] : \mu = \mathbb{E}[X] \tag{4}$$

Some notes:

- If $\mathbb{E}[X]$ doesn't exist then $var(X)$ doesn't exist.
- $var(X)$ can be infinite.
- The standard deviation σ of X is $\sqrt{var(X)}$.

Variance

With some algebra, we see that

$$
\begin{aligned}
var(X) &= \mathbb{E}[(X-\mu)^2] && (5)\\
&= \mathbb{E}[X^2 - 2X\mu - \mu^2] && (6)\\
&= \mathbb{E}[X^2] - \mathbb{E}[2X\mu] - \mathbb{E}[\mu^2] && (7)\\
&= \mathbb{E}[X^2] - \mu^2 && (8)\\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (9)
\end{aligned}
$$

Variance

Some properties:

- If X is bounded, then $var(X)$ exists and is finite.
- $var(X) = 0 \iff P(X = c) = 1$ for some constant c.
- $var(cX) = c^2 var(X)$ for some constant c.
- variance is linear, i.e. $var(X_1 + X_2) = var(X_1) + var(X_2)$.

Moments

The *k^{th} moment* of rv X is defined as

$$\mathbb{E}[X^k] = \mu'_k : k \in \mathbb{N} \tag{10}$$

The *k^{th} central/centered moment* of rv X is defined as

$$\mathbb{E}[(X - \mu)^k] = \mu_k : k \in \mathbb{N} \tag{11}$$

Moments

The *k^{th} moment* of rv X is defined as

$$\mathbb{E}[X^k] = \mu'_k : k \in \mathbb{N} \tag{10}$$

The *k^{th} central/centered moment* of rv X is defined as

$$\mathbb{E}[(X - \mu)^k] = \mu_k : k \in \mathbb{N} \tag{11}$$

Notes:

- μ'_k exists if and only if $\mathbb{E}[|X|^k] < \infty$.
- If μ'_k exists, then for all $j < k$, μ'_j also exists.
- Variance is μ_2.
- *Skewness* is μ_3/σ^2.
- *Kurtosis* is μ_4/σ^4.

Moments

Example: Suppose $X \sim N(0,1) \ni f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

$$\mu_1' = \mathbb{E}[X] = \int x f_X(x) dx = f_X(x)|_{-\infty}^{\infty} = 0 \tag{12}$$

n.b. For the normal distribution, $x f_X(x) = -\frac{\partial}{\partial x} f_X(x)$.

Moments

Example: Suppose $X \sim N(0,1) \ni f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

$$\mu_1' = \mathbb{E}[X] = \int x f_X(x) dx = f_X(x)|_{-\infty}^{\infty} = 0 \tag{12}$$

n.b. For the normal distribution, $x f_X(x) = -\frac{\partial}{\partial x} f_X(x)$.

$$\mu_2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[(X-0)^2] = \mathbb{E}[X^2] = \int x^2 f_X(x) dx \tag{13}$$

using integration by parts, we get

$$x^2 f_X(x) dx = \underbrace{-x f_X(x)|_{-\infty}^{\infty}}_{=0} + \underbrace{f_X(x)|_{-\infty}^{\infty}}_{=1} \tag{14}$$

Moment generating function

Moment generating functions (mgf) are used to calculate the moments of a rv. The mgf of a rv X is a function $M_X : \mathbb{R} \Rightarrow \mathbb{R}_+$ such that

$$M_X(t) = \mathbb{E}[e^{tX}] : t \in \mathbb{R} \tag{15}$$

Moment generating function

Moment generating functions (mgf) are used to calculate the moments of a rv. The mgf of a rv X is a function $M_X : \mathbb{R} \Rightarrow \mathbb{R}_+$ such that

$$M_X(t) = \mathbb{E}[e^{tX}] : t \in \mathbb{R} \tag{15}$$

Notes:

- The mgf is a function of t; X is integrated out by $\mathbb{E}$.
- The mgf only applies if the moments of the rv exists.
- If two rv X, Y have the same mgf (i.e. $M_X(t) = M_Y(t)$), then they have the same distribution.
- Even if a rv has moments, the mgf may yield infinity (e.g. log-normal distribution).

Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \tag{16}$$

What happens when $t = 0$?

Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \tag{16}$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \tag{17}$$

Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \tag{16}$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \tag{17}$$

What happens when $t = 0$ for the k^{th} derivative?

Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \tag{16}$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \tag{17}$$

What happens when $t = 0$ for the k^{th} derivative?

$$\frac{\partial}{\partial t^k} M_X(t) = \int x^k \cdot e^{tx} f_X(x) dx \tag{18}$$

At $t = 0$, we get $\frac{\partial}{\partial t^k} M_X(t)|_{t=0} = \mathbb{E}[X^k]$

Evaluating the k^{th} derivative at $t = 0$ gives us the k^{th} moment of X.

Moment generating function

Example: The standard normal distribution

$$
\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] = \int e^{tX} f_X(x) dx & (19)\\
&= \int e^{tX} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx & (20)\\
&= \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) \exp\left(\frac{t^2}{2}\right) dx & (21)\\
&= \exp\left(\frac{t^2}{2}\right) \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) dx & (22)\\
&= \exp\left(\frac{t^2}{2}\right) & (23)
\end{aligned}
$$

Moment generating function

The mgf for *affine transformations* is straight forward,
e.g. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.

Example: Let $X = \mu + \sigma Z : Z \sim N(0, 1)$. Then

$$M_X(t) = M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \quad (24)$$

Moment generating function

The mgf for *affine transformations* is straight forward,
e.g. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.

Example: Let $X = \mu + \sigma Z : Z \sim N(0, 1)$. Then

$$M_X(t) = M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \quad (24)$$

Another example:
Let $X_1, \ldots, X_n \overset{iid}{\sim} P_0$ and $Y = \sum_{i=1}^n X_i$. Then

$$\begin{aligned} M_Y(t) &= \mathbb{E}[e^{tY}] = \mathbb{E}[e^{t(X_1 + \cdots + X_n)}] = \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right] && (25) \\ &= \prod_{i=1}^n \mathbb{E}\left[e^{tX_i}\right] = \prod_{i=1}^n M_{X_i}(t) && (26) \end{aligned}$$

Distributions

Most useful distributions have names, e.g.

- Normal distribution
- Uniform distribution
- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Gamma distribution

Normal distribution

A rv X follows a *Normal distribution*, denoted as $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2, if X is continuous with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : x \in \mathbb{R} \tag{27}$$

Note:

If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. It follows that

- $\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu$.
- $var(X) = var(\mu + \sigma Z) = \sigma^2 var(Z) = \sigma^2$.

Normal distribution

A rv X follows a *Normal distribution*, denoted as $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2, if X is continuous with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : x \in \mathbb{R} \tag{27}$$

Note:

If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. It follows that

- $\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu$.
- $var(X) = var(\mu + \sigma Z) = \sigma^2 var(Z) = \sigma^2$.

Most well known distribution due to:

1. Good mathematical properties
2. Often (approximately) observed in the real world (e.g. heights, weights, etc.)
3. Central limit theorem

Uniform distribution

A rv X follows a Uniform distribution $U(a, b)$ if X is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Under $U(a, b)$, all observations are "*equally likely*"
$\mathbb{E}[X] = \frac{a+b}{2}$, $var(X) = \frac{(b-a)^2}{12}$, and $M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$.

Uniform distribution

A rv X follows a Uniform distribution $U(a, b)$ if X is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Under $U(a, b)$, all observations are "*equally likely*"
$\mathbb{E}[X] = \frac{a+b}{2}$, $var(X) = \frac{(b-a)^2}{12}$, and $M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$.

Note: if $X \sim U(a, b)$, then $X = (b-a)\tilde{X} + a : \tilde{X} \sim U(0, 1)$ and

$$f_{\tilde{X}}(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

Bernoulli distribution

A rv X follows a Bernoulli distribution $Ber(p)$ if X is discrete with pmf

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

$\mathbb{E}[X] = p$, $var(X) = p(1-p)$, and $M_X(t) = e^t p + (1-p)$.

Binomial distribution

A rv X follows a Binomial distribution $Bin(n, p)$ if X is discrete with pmf

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-1} & \text{if } x \in \{0, 1, ..., n\} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

$\mathbb{E}[X] = np$, $var(X) = np(1-p)$, and $M_X(t) = (e^t p + (1-p))^n$.

If $X_1, ..., X_n \overset{iid}{\sim} Ber(p)$, then $Y = X_1 + \cdots + X_n$ follows $B(n, p)$.

Negative Binomial distribution

A rv X follows a Negative Binomial distribution $NB(r,p)$ if X is discrete with pmf

$$f_X(x) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x & \text{if } x \in \{0, 1, ..., n\} \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

$\mathbb{E}[X] = \frac{r(1-p)}{p}$, $var(X) = \frac{r(1-p)}{p^2}$, and
$M_X(t) = \left(\frac{p}{1-qe^t}\right)^r : t < \log\left(\frac{1}{q}\right)$.

When $r = 1$, we refer to it as the *Geometric distribution*.

- It has a *memoryless* property.

Poisson distribution

A rv X follows a Poisson distribution $Pois(\lambda)$ if X is discrete with pmf

$$f_X(x) = \begin{cases} e^{\lambda} \frac{\lambda^x}{x!} & x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

$\mathbb{E}[X] = \lambda$, $var(X) = \lambda$, and $M_X(t) = e^{\lambda(e^t - 1)}$.

Some notes:

- $Bin(n, p) \approx Pois(np)$ when n is large and np is small.
- "*Poisson Processes*" are typically used to model rates, e.g. mortality rates
  1. The number of events in each fixed time interval t has a Poisson distribution with mean λt.
  2. The number of events in each time interval is independent.

Gamma distribution

A rv X follows a Gamma distribution $Gamma(\alpha, \beta)$ if X is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

where $\Gamma(x) = \int_0^\infty t^{\alpha-1} e^{-t} dt : \alpha > 0$.

$\mathbb{E}[X] = \alpha\beta$, $var(X) = \alpha\beta^2$, and
$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} : t < \beta$.

Gamma distribution

A rv X follows a Gamma distribution $Gamma(\alpha, \beta)$ if X is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

where $\Gamma(x) = \int_0^\infty t^{\alpha-1} e^{-t} dt : \alpha > 0$.

$\mathbb{E}[X] = \alpha\beta$, $var(X) = \alpha\beta^2$, and
$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} : t < \beta$.

Notes:

- $\frac{1}{\Gamma(\alpha)\beta^\alpha}$ is often referred to as the '*normalizing constant*'.
- When $\alpha = 1$, we get the exponential distribution.

Beta distribution

A rv X follows a Beta distribution $Beta(\alpha, \beta)$ if X is continuous with pdf

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (35)$$

$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$, $var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, and
$M_X(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1} dx$.

n.b. Very popular distribution in Bayesian statistics.

Multinomial distribution

Suppose rv $\mathbf{X} = (X_1, ..., X_k)$ represents counts of k different classes. Then it follows a Multinomial distribution $Multi(p_1, ..., p_k)$ if it has pdf

$$f_X(x) = \begin{cases} \binom{n}{x_1, ..., x_k} p_1^{x_1} \cdots p_k^{x_k} & x_1 \geq 0, ..., x_k \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

where $n = \sum_{i=1}^{k} X_i$.

$\mathbb{E}[X_i] = np$, $var(X_i) = np_i(1 - p_i)$, and
$Cov(X_i, X_j) = -np_i p_j$.

Dirac delta function

While not technically a pdf, often used for e.g. mixture of discrete distributions

The Dirac delta function is defined as $\delta : \mathbb{R} \to \mathbb{R} \cup \infty \ \ni$

$$\delta(x) = \begin{cases} +\infty & x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

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

The sifting property:

$$\int f(x)\delta(x-a)dx = f(a) \tag{38}$$

Dirac delta function

Example: Let

$$Y = \begin{cases} 1 & \text{w.p. } \alpha \\ U(0,1) & \text{w.p. } 1-\alpha \end{cases} \tag{39}$$

Then $f_Y(y) = \alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1])$

Dirac delta function

Example: Let

$$Y = \begin{cases} 1 & \text{w.p. } \alpha \\ U(0,1) & \text{w.p. } 1-\alpha \end{cases} \tag{39}$$

Then $f_Y(y) = \alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1])$

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{\infty}^{\infty} y(\alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1]))dy & (40) \\
&= \alpha \int_{\infty}^{\infty} y(\delta(y-1)dy + (1-\alpha)\int_0^1 ydy & (41) \\
&= \alpha + (1-\alpha)\frac{y^2}{2}|_0^1 & (42) \\
&= \alpha + \frac{1-\alpha}{2} & (43) \\
&= \frac{1+\alpha}{2} & (44)
\end{aligned}$$

References

- DeGroot & Schervish Chapters 4.1-4.5,5.1-5.9
- Grinstead & Snell Chapters 5,6