Master Program in *Data Science and Business Informatics*

**Statistics for Data Science**

Lesson 09 - Expectation and variance. Computations with random variables

Salvatore Ruggieri

Department of Computer Science  
University of Pisa, Italy  
[salvatore.ruggieri@unipi.it](mailto:salvatore.ruggieri@unipi.it)
Expectation of a discrete random variable

- Buy lottery ticket every week, \( p = \frac{1}{10000} \), what is probability of winning at \( k^{th} \) week?
  \[
  X \sim \text{Geo}(p) \quad P(X = k) = (1 - p)^{k-1} \cdot p \quad \text{for} \quad k = 1, 2, \ldots
  \]

- What is the average number of weeks to wait (expected) before winning?
  \[
  E[X] = \sum_{k=1}^{\infty} k \cdot (1 - p)^{k-1} \cdot p = \frac{1}{p}
  \]
  because \( \sum_{k=1}^{\infty} k \cdot x^{k-1} = \frac{1}{(1-x)^2} \)

**Definition.** The *expectation* of a discrete random variable \( X \) taking the values \( a_1, a_2, \ldots \) and with probability mass function \( p \) is the number

\[
E[X] = \sum_{i} a_i P(X = a_i) = \sum_{i} a_ip(a_i).
\]

- Expected value, mean value (weighted by probability of occurrence), center of gravity

*See seeing-theory.brown.edu*
Expected value may be infinite or may not exist!

- Fair coin: win $2^k$ euros if first $H$ appears at $k^{th}$ toss
  - $X$ with p.m.f. $p(2^k) = 2^{-k}$ for $k = 1, 2, \ldots$
  - $p()$ is a p.m.f. since $\sum_{k=1}^{\infty} 2^{-k} = 1$
  - Expected win (fair value to enter the game):
    
    $$E[X] = \sum_{k=1}^{\infty} 2^k \cdot 2^{-k} = \sum_{k=1}^{\infty} 1 = \infty$$

- Expectation does not exist when $\sum_i a_i p(a_i)$ does not converge
  - $X$ with p.m.f. $p(2^k) = p(-2^k) = 2^{-k}$ for $k = 2, 3, \ldots$
  - $E[X] = \sum_{k=2}^{\infty} (2^k \cdot 2^{-k} - 2^k \cdot 2^{-k}) = \sum_{k=2}^{\infty} (1 - 1) = 0$ wrong!
  - $E[X] = \sum_{k=2}^{\infty} 2^k \cdot 2^{-k} - \sum_{k=2}^{\infty} 2^k \cdot 2^{-k} = \infty - \infty$ undefined
  - $E[X]$ is finite if $\sum_i |a_i| p(a_i) < \infty$
  - In the case above, $\sum_{k=2}^{\infty} (|2^k| \cdot 2^{-k} + |-2^k| \cdot 2^{-k}) = \infty$

[St. Petersburg paradox] using $\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}$ for $|a| < 1$
• Expectation of some other discrete distributions

  ▶ \( X \sim U(m, M) \quad E[X] = \frac{m+M}{2} \)

  □ \[ \sum_{i=m}^{M} \frac{i}{M-m+1} = \frac{1}{M-m+1} \sum_{i=0}^{M-m}(m + i) = m + \frac{(M - m)/2}{2} = \frac{m + M}{2} \]

  ▶ \( X \sim Ber(p) \quad E[X] = p \)

  □ 0 \cdot (1 - p) + 1 \cdot p = p

  ▶ \( X \sim Bin(n, p) \quad E[X] = n \cdot p \)

  □ Because . . . we’ll see later

  ▶ \( X \sim NBin(n, p) \quad E[X] = \frac{np}{1-p} \)

  □ Because . . . we’ll see later

  ▶ \( X \sim Poi(\mu) \quad E[X] = \mu \)

  □ Because, when \( n \to \infty: Bin(n, \mu/n) \to Poi(\mu) \)
Expectation of a continuous random variable

**Definition.** The *expectation* of a continuous random variable \( X \) with probability density function \( f \) is the number

\[
E[X] = \int_{-\infty}^{\infty} x f(x) \, dx.
\]

- **Expectation of some continuous distributions**
  - \( X \sim U(\alpha, \beta) \quad E[X] = (\alpha + \beta)/2 \)
  - \( X \sim \text{Exp}(\lambda) \quad E[X] = 1/\lambda \)
    - Because \( \int_{0}^{\infty} x \lambda e^{-\lambda x} \, dx = [-e^{-\lambda x}(x + 1/\lambda)]_{0}^{\infty} = e^0(0 + 1/\lambda) \) \[[See Lesson 06]\]
  - \( X \sim N(\mu, \sigma^2) \quad E[X] = \mu \)
    - Because: \( \int_{-\infty}^{\infty} x \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \, dx = \mu + \int_{-\infty}^{\infty} (x - \mu) \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \, dx = z \frac{\mu - \mu}{\sigma} \)
    - Because . . . we’ll see later
  - \( X \sim \text{Erl}(n, \lambda) \quad E[X] = n/\lambda \)
    - Because . . . we’ll see later
Expected value may not exists!

- Cauchy distribution (distribution of the ratio of two standard normals)

\[ f(x) = \frac{1}{\pi(1 + x^2)} \]

- \( X_1, X_2 \sim N(0, 1) \) i.i.d., \( X = X_1/X_2 \sim \text{Cau}(0, 1) \)

\[ E[X] = \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx \]

- \( \int_{-\infty}^{0} xf(x)dx = [\frac{1}{2\pi} \log(1 + x^2)]_{-\infty}^{0} = -\infty \)

- \( \int_{0}^{\infty} xf(x)dx = [\frac{1}{2\pi} \log(1 + x^2)]_{0}^{\infty} = \infty \)

\[ E[X] = -\infty + \infty \]

- \( E[X] \) is finite if \( \int_{-\infty}^{\infty} |x|f(x)dx < \infty \)

Mean value does not always make sense in your data analytics project!
Recall that velocity = space/time, and then time = space/velocity!

Vector $v$ of speed (Km/h) to reach school and probabilities $p$ using feet, bike, bus, train:

\[ v = c(5, 10, 20, 30) \quad p = c(0.1, 0.4, 0.25, 0.25) \]

Distance house-schools is 2 Km

What is the average time to reach school?

- \[ \frac{2}{\text{sum}(v*p)} \] i.e., space/$E[velocity]$, or
- \[ \text{sum}(\frac{2}{v*p}) \] i.e., $E[space/velocity]$

\[ X = \text{velocity}, \quad g(X) = \frac{2}{X} \text{ time to reach school} \]

\[ E[g(X)] \neq g(E[X]) \]

$E[g(X)]$ mean time

$g(E[X])$ time at mean velocity
The change of variable formula (or rule of the lazy statistician)

- $X \sim U(0, 10)$, width of a square field, $E[X] = 5$
- $g(X) = X^2$ is the area of the field, $E[g(X)] = ?$
- $F_g(a) = P(g(X) \leq a) = P(X \leq \sqrt{a}) = \sqrt{a}/10$ for $0 \leq a \leq 100$
- Hence, $f_g(a) = dF_g(a)/da = 1/20\sqrt{a}$
- $E[g(X)] = \frac{1}{20} \int_0^{100} \frac{x}{\sqrt{x}} \, dx = \frac{1}{20} \frac{2}{3} \left[ x^{3/2} \right]_0^{100} = 100/3$
- A more direct way:

$$E[g(X)] = \int_0^{10} x^2 \frac{1}{10} \, dx = \frac{1}{10} \frac{1}{3} \left[ x^3 \right]_0^{10} = 100/3$$

See R script
Theorem (Change of units)

\[ E[rX + s] = rE[X] + s \]

- Example: for \( Y = 1.8X + 32 \), we have \( E[Y] = 1.8E[X] + 32 \) \[\text{[Celsius to Fahrenheit]}\]

Corollary.

\[ E[X - E[X]] = E[X] - E[X] = 0 \]

**Theorem.** Expectation minimizes the square error, i.e., for \( a \in \mathbb{R} \):

\[ E[(X - E[X])^2] \leq E[(X - a)^2] \]

- Proof. (sketch) set \( \frac{d}{da} \int_{-\infty}^{\infty} (x - a)^2 f(x) dx = 0 \)
Computation with discrete random variables

**Theorem**
For a discrete random variable $X$, the p.m.f. of $Y = g(X)$ is:

$$P_Y(Y = y) = \sum_{g(x) = y} P_X(X = x) = \sum_{x \in g^{-1}(y)} P_X(X = x)$$

**Proof.** $\{Y = y\} = \{g(X) = y\} = \{x \in g^{-1}(y)\}$

**Corollary** (the change-of-variable formula):

$$E[g(X)] = \sum_y y P_Y(Y = y) = \sum_y y \sum_{g(x) = y} P_X(X = x) = \sum_x g(x) P_X(X = x)$$
Example

- \( X \sim U(1, 200) \) number of tickets sold
- Capacity is 150
- \( Y = \max\{X - 150, 0\} \) overbooked tickets

\[
P_Y(Y = y) = \begin{cases} 
\frac{150}{200} & \text{if } y = 0 \\
\frac{1}{200} & \text{if } 1 \leq y \leq 50 
\end{cases}
\]

- \( g^{-1}(0) = \{1, \ldots, 150\} \)
- \( g^{-1}(y) = \{y + 150\} \)

- Hence:

\[
E[Y] = 0 \cdot \frac{150}{200} + \frac{1}{200} \cdot \sum_{y=1}^{50} y = 6.375
\]

or using the change-of-variable formula:

\[
E[Y] = \frac{1}{200} \cdot \sum_{x=1}^{200} \max\{X - 150, 0\} = \frac{1}{200} \cdot \sum_{x=151}^{200} (X - 150) = 6.375
\]
Theorem

For a continuous random variable $X$, the density functions of $Y = g(X)$ when $g()$ is increasing/decreasing are:

$$F_Y(y) = F_X(g^{-1}(y)) \quad f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

**Proof. (for $g()$ increasing)** Since $g()$ is invertible and $g(x) \leq y$ iff $x \leq g^{-1}(y)$:

$$F_Y(y) = P_Y(g(X) \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

and then:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1} \frac{dy}{dy}} = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$$

**Exercise at home:** show the case $g()$ decreasing!
Change of units

For $X \sim N(\mu, \sigma^2)$, how is $Z = \frac{1}{\sigma}X + \frac{-\mu}{\sigma} = \frac{X-\mu}{\sigma}$ distributed?

- $f_Z(z) = \sigma f_X(\sigma y + \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$
- Hence, $Z \sim N(0, 1)$
- In particular, for $X \sim N(\mu, \sigma^2)$, we have:

$$P(X \leq a) = P(Z \leq \frac{a-\mu}{\sigma}) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

[See Lesson 08]
Example: $\Lambda(\mu, \sigma^2)$

**Log-normal distribution** $Y = e^X$ for $X \sim N(\mu, \sigma^2)$, i.e., $\log(Y) \sim N(\mu, \sigma^2)$

- $Y = g(X) = e^X$
- $g(x) = e^x$ is increasing, and $g^{-1}(y) = \log y$, and $\frac{dg^{-1}(y)}{dy} = \frac{1}{y}$

$$F_Y(y) = F_X(g^{-1}(y)) = \Phi\left(\frac{\log y - \mu}{\sigma}\right) \quad f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = \frac{1}{y} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\log y - \mu}{\sigma}\right)^2}$$

- $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{-\infty}^{\infty} y f_Y(y) dy = e^{\mu+\sigma^2/2}$

Plausible and empirically adequate model for long-tailed distributions:

- length of comments in posts, dwell time reading online articles, length of chess games, ...
- size of living tissue, number of hospitalized cases in epidemics, blood pressure, ...
- income of 97%-99% of the population, the number of citations, log of city size, ...
- times to repair a maintainable system, size of audio-video files, amount of internet traffic per unit time, ...

See R script
Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R

Example

- $X \sim U(0, 1)$ radius $f_X(x) = 1 \quad F_X(x) = x$ for $x \in [0, 1]$
- $Y = g(X) = \pi \cdot X^2$

$g(x) = \pi x^2$ is increasing, and $g^{-1}(y) = \sqrt{\frac{y}{\pi}}$, and $\frac{dg^{-1}(y)}{dy} = \frac{1}{2\sqrt{\pi y}}$

\[
F_Y(y) = F_X(g^{-1}(y)) = \sqrt{\frac{y}{\pi}} \quad f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = \frac{1}{2\sqrt{\pi y}}
\]

Do not lift distributions from a data column to a derived column in your data analytics project!

See R script

- Notice that: $g(E[X]) = \pi/4 \leq E[g(X)] = \int_0^1 g(x)f_X(x)dx = \int_0^\pi yf_Y(y)dy = \frac{\pi}{3}$
Jensen’s inequality

Jensen’s inequality. Let $g$ be a convex function, and let $X$ be a random variable. Then

$$g(E[X]) \leq E[g(X)].$$

- $f()$ is convex if $f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$ for $t \in [0, 1]$

- if $f''(x) \geq 0$ then $f()$ is convex, e.g., $g(x) = \pi x^2$ or $g(x) = 1/x$ for $x \geq 0$

**Corollary [T, Ex. 8.11].** For a concave function $g$, namely $g''(x) \leq 0$: $g(E[X]) \geq E[g(X)]$
Variance

- **Investment A.** \( P(X = 450) = 0.5 \quad P(X = 550) = 0.5 \quad E[X] = 500 \)
- **Investment B.** \( P(X = 0) = 0.5 \quad P(X = 1000) = 0.5 \quad E[X] = 500 \)

Spread around the mean is important!

### Variance and standard deviations

The **variance** \( \text{Var}(X) \) of a random variable \( X \) is the number:

\[
\text{Var}(X) = E[(X - E[X])^2]
\]

\( \sigma_X = \sqrt{\text{Var}(X)} \) is called the **standard deviation** of \( X \).

- The standard deviation has the same dimension as \( E[X] \) (and as \( X \))
- For \( X \) discrete, \( \text{Var}(X) = \sum_i (a_i - E[X])^2 p(a_i) \)
- **Investment A.** \( \text{Var}(X) = 50^2 \) and \( \sigma_X = 50 \)
- **Investment B.** \( \text{Var}(X) = 500^2 \) and \( \sigma_X = 500 \)
Examples

- For $a \in \mathbb{R}$:
  \[
  E[|X - a|] \leq \sqrt{E[(X - a)^2]}
  \]

  ▶ Apply Jensen’s ineq. for $g(y) = y^2$ convex on the r.v. $Y = |X - a|

- Median minimizes absolute deviation, i.e., for any $a \in \mathbb{R}$:
  \[
  E[|X - m_X|] \leq E[|X - a|]
  \]

  ▶ **Prove it!** (for continuous functions) Hint: $\frac{d}{dx} |x| = x/|x|

- Maximum distance between expectation and median:
  \[
  |E[X] - m_X| \leq E[|X - m_X|] \leq E[|X - E[X]|] \leq \sqrt{E[(X - E[X])^2]} = \sigma_X
  \]

  ▶ Jensen’s ineq. for $g(y) = |y|$ convex on the r.v. $Y = X - m_X$ plus the two results above
Mode

• For discrete r.v. $X$ with p.m.f. $p()$: the values $a$ such that $p(a)$ is maximum, i.e.:
  \[ \arg \max_a p(a) \]
  - Can be more than one, e.g., in $\text{Ber}(0.5)$

• For continuous r.v. $X$ with d.f. $f()$: the values $x$ such that $f(x)$ is a local maximum, e.g.:
  \[ f'(x) = 0 \quad \text{and} \quad f''(x) < 0 \]
  - Notice: local maximum!

• Unimodal distribution = that have only one mode
Variance

**Theorem**

\[ \text{Var}(X) = E[X^2] - E[X]^2 \]

**Proof.**

\[
\text{Var}(X) = E[(X - E[X])(X - E[X])]
= E[X^2] + E[X]^2 - 2XE[X]
= E[X^2] + E[X]^2 - E[2XE[X]]
\]

- \(E[X^2]\) is called the *second moment* of \(X\) for continuous r.v.’s: \(\int_{-\infty}^{\infty} x^2 f(x) dx\)

**Corollary.**

\[ \text{Var}(rX + s) = r^2 \text{Var}(X) \]

**Prove it!**

- Variance insensitive to shift \(s\)! 

\[ 21 / 25 \]
Variance may be infinite or may not exist!

Standard deviation $\sigma_X$ is a measure of the margin of error around a predicted value

- E.g., temperature “20 ± 1.5”

An infinite or non-existent margin of error is no prediction at all.

- Variance may not exists!
  - If expectation does not exist!
  - Also in cases when expectation exists: we’ll see later *Power laws*.

- Variance can be infinite
  - Distributions have fat upper tails that decrease at an extremely slow rate.
  - The slow decay of probability increases the odds of very extreme values (*outliers*)
  - E.g., $e^X$ for $X \sim \text{Cau}(0,1)$

![Log-Cauchy distribution](image)
Variance

- Variance of some discrete distributions
  - $X \sim U(m, M)$ \quad $\mathbb{E}[X] = \frac{(m+M)}{2}$ \quad $\text{Var}(X) = \frac{(M-m+1)^2-1}{12}$
    - use $\text{Var}(X) = \text{Var}(X - m)$, call $n = M - m + 1$ and \( \sum_{i=1}^{n-1} i^2 = \frac{(n-1)n(2n-1)}{6} \)
  - $X \sim \text{Ber}(p)$ \quad $\mathbb{E}[X] = p$ \quad $\text{Var}(X) = p^2(1 - p) + (1 - p)^2p = p(1 - p)$
  - $X \sim \text{Bin}(n, p)$ \quad $\mathbb{E}[X] = np$ \quad $\text{Var}(X) = np(1 - p)$
    - Because . . . we’ll see later
  - $X \sim \text{Geo}(p)$ \quad $\mathbb{E}[X] = \frac{1}{p}$ \quad $\text{Var}(X) = \frac{1-p}{p^2}$
    - Hint: use $\text{Var}(X) = E[X^2] - E[X]^2$ and $\sum_{k=1}^{\infty} k^2 \cdot x^{k-1} = \frac{1+x}{(1-x)^3}$
  - $X \sim \text{NBin}(n, p)$ \quad $\mathbb{E}[X] = \frac{n \cdot p}{1-p}$ \quad $\text{Var}(X) = n \frac{1-p}{p^2}$
    - Because . . . we’ll see later
  - $X \sim \text{Poi}(\mu)$ \quad $\mathbb{E}[X] = \mu$ \quad $\text{Var}(X) = \mu$
    - Because, when $n \to \infty$: $\text{Bin}(n, \frac{\mu}{n}) \to \text{Poi}(\mu)$

See seeing-theory.brown.edu
Variance

- Variance of some continuous distributions
  - $X \sim U(\alpha, \beta)$ \hspace{1cm} $E[X] = (\alpha + \beta)/2$ \hspace{1cm} $Var(X) = (\beta - \alpha)^2/12$
    - **Prove it!** Recall that $f(x) = 1/(\beta - \alpha)$
  - $X \sim \text{Exp}(\lambda)$ \hspace{1cm} $E[X] = 1/\lambda$ \hspace{1cm} $Var(X) = 1/\lambda^2$
    - **Prove it!** Recall that $f(x) = \lambda e^{-\lambda x}$
  - $X \sim N(\mu, \sigma^2)$ \hspace{1cm} $E[X] = \mu$ \hspace{1cm} $Var(X) = \sigma^2$
    - **Prove it!** Hint: use $z = \frac{x - \mu}{\sigma}$ and integration by parts.
  - $X \sim \text{Erl}(n, \lambda)$ \hspace{1cm} $E[X] = n/\lambda$ \hspace{1cm} $Var(X) = n/\lambda^2$
    - Because . . . we’ll see later
**E[] and Var() of random variables with bounded support**

- Assume $a \leq X \leq b$, or more generally $P(a \leq X \leq b) = 1$ [almost surely or a.s.]
- It turns out that expectation and variance are finite!

(A) $a \leq E[X] \leq b$
  
  - E.g., for $X$ continuous, $E[X] = \int_a^b xf(x)dx \leq \int_a^b bf(x)dx = b$

(B) $0 \leq Var(X) \leq \frac{(b-a)^2}{4}$

**Proof.**

- From (A), since $0 \leq (X - E[X])^2$, we have $0 \leq E[(X - E[X])^2] = Var(X)$
- For any $\gamma \in \mathbb{R}$, we have $E[(X - E[X])^2] \leq E[(X - \gamma)^2]$ [See slide 9]
  
  □  Thus, $E[(X - E[X])^2] = Var(X) \leq E[(X - \gamma)^2]$
- For $\gamma = (a+b)/2$, we have $(X - \gamma)^2 \leq (b - \gamma)^2$, and then by (A):

  \[ Var(X) \leq E[(X - \gamma)^2] \leq (b - \gamma)^2 = \left(b - \frac{(a+b)}{2}\right)^2 = \frac{(b-a)^2}{4} \]

- **Exercise at home:** show that the bound $(b-a)^2/4$ can be reached for some distribution.