Master Program in *Data Science and Business Informatics*

**Statistics for Data Science**

Lesson 14 - Law of large numbers, and the central limit theorem

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy

salvatore.ruggieri@unipi.it
Markov’s inequality

Notation. Indicator function: \( \mathbb{1}_{\varphi(x)} = \begin{cases} 1 & \text{if} \ \varphi(x) \\
0 & \text{otherwise} \end{cases} \)

- Link expectation to probability of events
- \( E[\mathbb{1}_{X \geq \alpha}] = \sum_a \mathbb{1}_{X \geq \alpha}(a)p(x(a) = \sum_{a \geq \alpha} p(x(a) = P_X(X \geq \alpha) \)

• Question: how much probability mass is near the expectation?

Markov’s inequality. Assume \( X \geq 0 \), and \( \alpha > 0 \):

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

Proof. Take expectations of \( \alpha \mathbb{1}_{X \geq \alpha} \leq X \).

• For a non-negative r.v., the probability of a large value is inversely proportional to the value

Corollary. Assume \( X \geq 0 \), \( E[X] > 0 \) and \( k > 0 \). We have: \( P(X \geq kE[X]) \leq \frac{1}{k} \)
Chebyshev’s inequality

• Question: how much probability mass is near the expectation?

Proof. Let $X = (Y - E[Y])^2$ and $\alpha = a^2$. By Markov’s inequality:

$$P(|Y - E[Y]| \geq a) = P((Y - E[Y])^2 \geq a^2) \leq \frac{E[(Y - E[Y])^2]}{a^2} = \frac{1}{a^2} \text{Var}(Y).$$
Chebyshev’s inequality

- **“μ ± a few σ” rule:** Most of the probability mass of a random variable is within a few standard deviations from its expectation!

- Let $\mu = E[Y]$ and $\sigma^2 = \text{Var}(Y) > 0$. For $k > 0$ (and hence $a = k\sigma > 0$):

  $$P(|Y - \mu| < k\sigma) = 1 - P(|Y - \mu| \geq k\sigma) \geq 1 - \frac{1}{k^2 \sigma^2 \text{Var}(Y)} = 1 - \frac{1}{k^2}$$

- For $k = 2, 3, 4$, the RHS is $\frac{3}{4}, \frac{8}{9}, \frac{15}{16}$

- Chebyshev’s inequality is sharp when nothing is known about $X$, but in general it is a large bound!

  **See R script**
Averages vary less

- Guessing the weight of a cow

- **See Francis Galton** (inventor of standard deviation, regression, and much more)
Let $X_1, X_2, \ldots, X_n$ be independent r. v. for which $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

**Expectation and variance of an average.** If $\bar{X}_n$ is the average of $n$ independent random variables with the same expectation $\mu$ and variance $\sigma^2$, then

$$E[\bar{X}_n] = \mu \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$ 

Notice that $X_1, \ldots, X_n$ are not required to be identically distributed!

See R script
The (weak) law of large numbers

- Apply Chebyshev’s inequality to $\bar{X}_n$

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{1}{\epsilon^2} \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n\epsilon^2}$$

- For $n \to \infty$, $\sigma^2/(n\epsilon^2) \to 0$

probability that $\bar{X}_n$ is far from $\mu$ tends to 0 as $n \to \infty!$ [Convergence in probability]

- It holds also if $\sigma^2$ is infinite (proof not included)

- Notice (again!) that $X_1, \ldots, X_n$ are not required to be identically distributed!
Recovering probability of an event

**Objective:** We want to know \( p = P(a < X \leq b) \)

- Run \( n \) independent measurements
- Model the results as \( X_1, \ldots, X_n \) random variables
- Define the indicator variables, for \( i = 1, \ldots, n \):

\[
Y_i = \mathbb{1}_{a<X_i\leq b} = \begin{cases} 
1 & \text{if } a < X_i \leq b \\
0 & \text{otherwise}
\end{cases}
\]

- \( Y_i \)'s are independent \[\text{by propagation of independence, see Lesson 10}\]
- \( E[Y_i] = P(a < X \leq b) = p \) and \( \text{Var}(Y_i) = p(1-p) \)
- Defined \( \bar{Y}_n = \frac{Y_1 + \ldots + Y_n}{n} \), by the law of large numbers:

\[
\lim_{n \to \infty} P\left(\left| \frac{\bar{Y}_n - p}{\epsilon} \right| > 0 \right) = 0
\]

- Frequency counting of values \((a, b]\) (e.g., in histograms) is a prob. estimation method!
Estimating conditional probability

**Objective**: estimate $p = P(C = c | A = a) = P(A = a, C = c) / P(A = a) = p_{ac} / p_a$

- Run $n$ independent measurement
- Model the results as $(A_1, C_1), \ldots, (A_n, C_n)$
- Using the approach of previous slide (but with the strong LLN):
  - for $Y_i = 1_{A_i = a, C_i = c}$: $P(\lim_{n \to \infty} \bar{Y}_n = p_{ac}) = 1$ where $p_{ac} = P(A = a, C = c)$
  - for $Z_i = 1_{A_i = a}$: $P(\lim_{n \to \infty} \bar{Z}_n = p_a) = 1$ where $p_a = P(A = a)$

- if $\bar{Z}_n \neq 0$, from previous two statements: (limit of a ratio is the ratio of the limits)
  $$P(\lim_{n \to \infty} \frac{\bar{Y}_n}{\bar{Z}_n} = \frac{p_{ac}}{p_a}) = 1$$

- Sample usage: almost everywhere in Machine Learning
- Issues when $n$ is small
  - e.g., in target encoding of rare categorical values [Micci-Barreca, 2001]

  **See R script**
Hoeffding bound

Theorem (Hoeffding bound)

If $\bar{X}_n$ is the average of $n$ independent r.v. with expectation $\mu$ and $P(a \leq X_i \leq b) = 1$, then for any $\epsilon > 0$

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

• For bounded support, a tight upper bound!
• When $a = 0, b = 1$ (e.g., Bernoulli trials):

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

• Other concentration inequalities.
The central limit theorem

• Let $X_1, X_2, \ldots, X_n$ be independent r. v. for which $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \quad E[\bar{X}_n] = \mu \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

• Can we derive the distribution of $\bar{X}_n$?

• Assume $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ known. We have:

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \quad Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

• Interestingly, the same conclusion extends to any other distribution!
The central limit theorem

The central limit theorem. Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let $\mu$ be the expected value and $\sigma^2$ the variance of each of the $X_i$. For $n \geq 1$, let $Z_n$ be defined by

$$Z_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma};$$

then for any number $a$

$$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a),$$

where $\Phi$ is the distribution function of the $N(0,1)$ distribution. In words: the distribution function of $Z_n$ converges to the distribution function $\Phi$ of the standard normal distribution.

- It extends to not identically distributed r.v.’s
- Why is it so frequent to observe a normal distribution?
  - Sometime it is the average/sum effects of other variables, e.g., as in “noise”
  - This justifies the common use of it to stand in for the effects of unobserved variables

[**Lindeberg’s condition**]

See R script and seeing-theory.brown.edu
Applications: approximating probabilities

- Let $X_1, \ldots, X_n \sim \text{Exp}(2)$, for $n = 100$ 
  
- Assume to observe realizations $x_1, \ldots, x_n$ such that $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i = 0.6$

- What is the probability $P(\bar{X}_n \geq 0.6)$ of observing such a value or a greater value?

**Option A:** Compute the distribution of $\bar{X}_n$

- $S_n = X_1 + \ldots + X_n \sim \text{Erl}(n, 2)$

- $\bar{X}_n = \frac{S_n}{n}$ hence by change-of-units transformation [See Lesson 09]

$$F_{\bar{X}_n}(x) = F_{S_n}(n \cdot x) \quad \text{and} \quad f_{\bar{X}_n}(x) = n \cdot f_{S_n}(n \cdot x)$$

- and then:

$$P(\bar{X}_n \geq 0.6) = 1 - F_{\bar{X}_n}(0.6) = 1 - F_{S_n}(n \cdot 0.6) = 1 - \text{pgamma}(60, n, 2) = 0.0279$$
Applications: approximating probabilities

- Let $X_1, \ldots, X_n \sim \text{Exp}(2)$, for $n = 100$
- Assume to observe realizations $x_1, \ldots, x_n$ such that $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i = 0.6$
- What is the probability $P(\bar{X}_n \geq 0.6)$ of observing such a value or a greater value?

**Option B:** Approximate them by using the CLT (requires $\mu$ and $\sigma$)

- Since $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ for $n \to \infty$:

$$P(\bar{X}_n \geq 0.6) = P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{0.6 - \mu}{\sigma/\sqrt{n}}\right) = P(Z_n \geq \frac{0.6 - 0.5}{0.5/\sqrt{n}}) \approx 1 - \Phi(2) = 0.0228$$

- also, notice $X_1 + \ldots + X_n = \sqrt{n}\sigma Z_n + n\mu \sim N(n\mu, n\sigma^2)$

See R script
How large should $n$ be?

- How fast is the convergence of $Z_n$ to $N(0, 1)$?
- The approximation might be poor when:
  - $n$ is small
  - $X_i$ is asymmetric, bimodal, or discrete
  - the value to test (0.6 in our example) is far from $\mu$
Target encoding of categorical features.

Daniele Micci-Barreca (2001)

A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems

SIGKDD Explor. Newsl. 3 (1), 27 – 32.