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 variable: \( \mathbb{1}_{X \in A}(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \)

\[ E[\mathbb{1}_{X \in A}] = \sum_x \mathbb{1}_{X \in A}(x) p_X(a) = \sum_{x \in A} p_X(a) = P_X(X \in A) = P(\mathbb{1}_{X \in A} = 1) \]

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

Markov’s inequality. For \( X \) non-negative (i.e., \( P(X < 0) = 0 \)) 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. For \( X \) non-negative, \( E[X] > 0 \) and \( k > 0 \): \( P(X \geq kE[X]) \leq \frac{1}{k} \)
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:

$$P(|Y - E[Y]| \geq a) = P((Y - E[Y])^2 \geq a^2) \leq \frac{E[(Y - E[Y])^2]}{a^2} = \frac{1}{a^2} \text{Var}(Y)$$
• “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 $\sigma^2 = Var(Y) > 0$. For $k > 0$ (and hence $a = k\sigma > 0$):

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

• 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 and much more)
Expectation and variance of an average

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 $n \to \infty$, $\frac{\sigma^2}{(n\epsilon^2)} \to 0$

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

- 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:** Let \( C = (a, b] \), and want to know \( p = P(X \in C) \)

- Run \( n \) independent measurements
- Model the results as \( X_1, \ldots, X_n \) random variables
- Define the indicator variables, for \( i = 1, \ldots, n \):
  \[
  Y_i = \mathbb{1}_{X_i \in C} = \begin{cases} 
  1 & \text{if } X_i \in C \\
  0 & \text{if } X_i \notin C
  \end{cases}
  \]

- \( Y_i \)'s are independent \([\text{Propagation of independence}]\)
- \( E[Y_i] = 1 \cdot P(X_i \in C) + 0 \cdot P(X_i \notin C) = p \)
- Defined \( \bar{Y}_n = \frac{Y_1 + \cdots + Y_n}{n} \), by the law of large numbers:
  \[
  \lim_{n \to \infty} P(\mid \bar{Y}_n - p \mid > \epsilon) = 0
  \]
- Frequency counting of \( v \in (a, b] \) (e.g., in histograms) is a probability estimation method!
Estimating conditional probability

**Objective:** estimate $P(Y = y | A = a)$ given $(a_1, y_1), \ldots, (a_t, y_t)$ with $y \in \{0, 1, \ldots, k - 1\}$

- Let $n = |\{(a_i, y_i) \mid a_i = a\}$ and $n_y = |\{(a_i, y_i) \mid a_i = a, y_i = y\}$
- Use $n_y/n$, the proportion of $Y = y$ over $A = a$: Ok for $n \rightarrow \infty$, not for $n$ small
- $m$-estimate:

$$\frac{n_y + mp_y}{n + m}$$

where $m$ is a weight factor and $p_y = t_y/t$ prior probability with $t_y = |\{(a_i, y_i) \mid y_i = y\}$

- Smoothing regularization

$$\lambda(n) \frac{n_y}{n} + (1 - \lambda(n))p_y$$

where $\lambda(n) \in [0, 1]$ is increasing with $n$
  - Interpolate $P(Y = y | A = a)$ with $P(Y = y)$
  - For $\lambda(n) = n/(n + m)$, we get the $m$-estimate

- Sample usage: target encoding of categorical attributes [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):
  $$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

**Corollary.** 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 $n \geq 1/2\epsilon^2\log^2/\delta$:

$$P(|\bar{X}_n - \mu| \leq \epsilon) \geq 1 - \delta$$

- $\epsilon$ accuracy: allowed error in estimation
- $\delta$ confidence: allowed probability of failure in achieving the accuracy

- E.g., recovering probability of an event: $P(|\bar{Y}_n - p| \leq 0.01) \geq 0.99$ for $n = 3516$
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 $\bar{X}_n$?
- We already showed that, for $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ independent:

\[ \frac{X_1 + X_2}{2} \sim N(\frac{\mu_1 + \mu_2}{2}, \frac{\sigma_1^2 + \sigma_2^2}{2^2}) \]

- Assume $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$:

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

- Interestingly, the same conclusion extends to any distribution for the $X_i$’s!
The central limit theorem

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

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

- Let $X_1, \ldots, X_n \sim \text{Exp}(2)$, for $n = 100$
  $\mu = \sigma = 1/2$
- 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
  $$F_{\bar{X}_n}(x) = F_{S_n}(n \cdot x) \quad \text{and} \quad f_{\bar{X}_n}(x) = n \cdot f_{S_n}(n \cdot x)$$
- 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 $n = 100$
  \[ \mu = \sigma = 1/2 \]
- 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

- $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ implies $\bar{X}_n = \frac{\sigma}{\sqrt{n}} Z_n + \mu \sim N(\mu, \sigma^2/n)$ for $n \to \infty$
- and then:
  \[ P(\bar{X}_n \geq 0.6) = P(\frac{\sigma}{\sqrt{n}} Z_n + \mu \geq 0.6) = P(Z_n \geq \frac{0.6 - \mu}{\sigma/\sqrt{n}}) \approx 1 - \Phi(\frac{0.6 - 0.5}{0.5/10}) = 0.0228 \]
  \[ \text{See R script} \]
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 \( \mu \)

\( \text{the myth of } n \geq 30 \)
Daniele Micci-Barreca (2001)

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