Statistical Methods for Data Science

Lesson 10 - Law of large numbers, and the central limit theorem

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Chebyshev's inequality

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

> CHEBYSHEV'S INEQUALITY. For an arbitrary random variable Y and any $a > 0$:
>
> $$\mathrm{P}(|Y - \mathrm{E}[Y]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y).$$

- **Proof.** (continuous case) Let $\mu = E[Y]$:

$$Var(Y) = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy \geq \int_{|y-\mu|\geq a} (y-\mu)^2 f(y)dy$$

$$\geq \int_{|y-\mu|\geq a} a^2 f(y)dy = a^2 P(|Y-\mu| \geq a)$$

Chebyshev's inequality

- **"$\mu\pm$ a few σ" rule:** Most of the probability mass of a random variable is within a few standard deviations from its expectation!
- Let $\sigma^2 = Var(Y)$. For $a = k\sigma$:

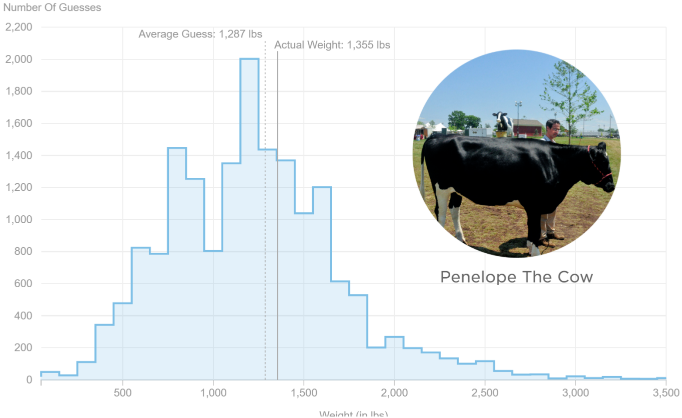
$$P(|Y-\mu| < k\sigma) = 1 - P(|Y-\mu| \geq k\sigma) \geq 1 - \frac{1}{k^2\sigma^2} Var(Y) = 1 - \frac{1}{k^2}$$

- For $k = 2, 3, 4$, the RHS is $3/4, 8/9, 15/16$
- 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 $Var(X_i) = \sigma^2$

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

EXPECTATION AND VARIANCE OF AN AVERAGE. If $\bar{X}_n$ is the average of n independent random variables with the same expectation μ and variance σ^2, then

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

- 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 $\bar{X}_n$

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

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

THE LAW OF LARGE NUMBERS. If $\bar{X}_n$ is the average of n independent random variables with expectation μ and variance σ^2, then for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathrm{P}\left(|\bar{X}_n - \mu| > \varepsilon\right) = 0.$$

- $\bar{X}_n$ converges to μ as $n \to \infty$!
- It holds also if σ^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

- 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 $i = 1, \ldots, n$:

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C \\ 0 & \text{if } X_i \notin C \end{cases}$$

- Y_i's are independent *[Propagation of independence]*
- $E[Y_i] = 1 \cdot P(X_i \in C) + 0 \cdot P(X_i \in C) = p$
- Defined $\bar{Y}_n = \frac{Y_1 + \ldots + Y_n}{n}$, by the law of large numbers:

$$\lim_{n \to \infty} P(|\bar{Y}_n - p| > \epsilon) = 0$$

- Frequency counting (e.g., in histograms) is a probability estimation method!

The central limit theorem

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

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

- Can we derive the distribution of $\bar{X}_n$?
- For $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ indepedent:
  - $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ *[the converse is also true (Levy Cramer thm)]*
  - and $\frac{Y_1+Y_2}{2} \sim N(\frac{\mu_1+\mu_2}{2}, \frac{\sigma_1^2+\sigma_2^2}{2^2})$
- Assume $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$:

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

- OK, does it generalize to any distribution? **Yes!**

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 μ be the expected value and σ^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 Φ is the distribution function of the $N(0,1)$ distribution. In words: the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

- Some generalizations get rid of the identically distributed assumption.
- Why is it so frequent to observe a normal distribution?
  - Sometime it is the average/sum effects of other variables
  - 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 $X_1, \ldots, X_n \sim Exp(2)$, for $n = 100$ $\qquad \mu = \sigma = 1/2$
- 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 $\bar{X}_n$

- $S_n = X_1 + \ldots + X_n \sim Erl(n, 2)$
- $\bar{X}_n = S_n/n$ 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:

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

Applications: approximating probabilities

- Let $X_1, \ldots, X_n \sim Exp(2)$, 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:

$$P(\bar{X}_n \geq 0.6) = P(\frac{\sigma}{\sqrt{n}} Z_n + \mu \geq 0.6) = P(Z_n \geq \frac{0.6 - \mu}{\sigma/\sqrt{n}}) \approx 1 - \Phi(\frac{0.6 - 0.5}{0.5/10}) == 0.0228$$

- 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 **the myth of $n \geq 30$**
  - X_i is asymmetric, bimodal, or discrete
  - the value to test (0.6 in our example) is far from μ