Statistical Methods for Data Science
Lesson 11 - Graphical summaries

Salvatore Ruggieri

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

- Probability models govern some random phenomena
- Confronted with a new phenomenon, we want to learn about the randomness that is associated with it
- Record observations $x_1, \ldots, x_n$ (a dataset)
- Can be too many: need to condense for easy visual comprehension
- Graphical methods:
  - Univariate: histograms, kernel density estimates, empirical distribution functions
  - Multi-variate: scatter plots
Barplots

- For discrete data, barplots provide frequency counts for values
  - approximate the p.m.f. due to the law of large numbers

- For continuous data, counting distinct values do not work. Why?

See R script
Histograms provide frequency counts for ranges of values:

- Split the support to intervals, called *bins*:

\[ B_1, \ldots, B_m \]

where the length \(|B_i|\) is called the *bin width*

- Count observations in each bin and normalize them:

\[ A_i = \frac{\left\{ j \in [1, n] \mid x_j \in B_i \right\}}{n} \approx P(X \in B_i) \]

- Plot bars whose area is proportional to \( A_i \):

\[ A_i = |B_i| \cdot H_i \quad \text{where} \quad H_i = \frac{\left\{ j \in [1, n] \mid x_j \in B_i \right\}}{n|B_i|} \]

See R script
Choice of the bin width

- Bins of equal width:

\[ B_i = (r + (i - 1)b, r + ib] \quad \text{for } i \in [1, m] \]

where \( r \leq \) minimum point and \( b \) is the bin width

- Scott’s normal reference rule (minimize mean integrated square error for Normal density):

\[ b = 3.49 \cdot s \cdot n, \quad \text{where } s = \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ is the sample standard deviation} \]
Choice of the bin width

- \( b = 2 \cdot IQR(x) \cdot n \), where \( IQR(x) = Q_3(x) - Q_1(x) \)  
  \[ \text{[Freedman–Diaconis’ choice]} \]

- Variable bin width
  - Logarithmic binning in power laws

- Alternative: number of bins given equal bin width \( b \):
  - \( m = \left\lceil \frac{\max x_i - \min x_i}{b} \right\rceil \)
  - \( m = \left\lceil \sqrt{n} \right\rceil \)
  - \( m = \left\lceil \log_2 n \right\rceil + 1 \)  
  \[ \text{[Sturges’ formula]} \]

N.B. R’s \texttt{hist} method take bin width as a suggestion, then it rounds bins differently

See R script
Density estimation

- Problem with histograms: as \( m \) increases, histogram becomes unusable
- Idea: estimate density function by putting a pile (of sand) around each observation
- Kernels state the shape of the pile
  - Epanechnikov \( \frac{3}{4}(1 - u^2) \) for \(-1 \leq u \leq 1\)
  - Triweight \( \frac{35}{32}(1 - u^2)^3 \) for \(-1 \leq u \leq 1\)
  - Normal \( \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \) for \(-\infty < u < \infty\)

Fig. 15.4. Examples of well-known kernels \( K \).
Kernel density estimation (KDE)

A Kernel is a function $K : \mathbb{R} \rightarrow \mathbb{R}$ such that

- $K$ is a probability density, i.e., $K(u) \geq 0$ and $\int_{-\infty}^{\infty} K(u)du = 1$
- $K$ is symmetric, i.e., $K(-u) = K(u)$
- [sometime, it is required that] $K(u) = 0$ for $|u| > 1$

A bandwidth $h$ is a scaling factor over the support of $K$ (from $[-1, 1]$ to $[-h, h]$)

- if $X \sim K$, then $\frac{X}{h} \sim \frac{1}{h}K(\frac{u}{h})$

[Change-of-Unit rule]
Kernel density estimation (KDE)

Let \( x_1, \ldots, x_n \) be the observations

- \( K \) scaled and shifted at \( x_i \) is \( \frac{1}{h} K(\frac{u-x_i}{h}) \), with support \([x_i - h, x_i + h]\)

The kernel density estimate is defined as:

\[
f_{n,h}(u) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{u-x_i}{h})
\]

- It is a probability density!

[Prove it]

See R script
KDE vs histograms

- KDE has less variability!
Choice of the bandwidth

- **Note.** The choice of the kernel is not critical: different kernels give similar results
- **A problem.** The choice of the bandwidth $h$ is critical (and it may depend on the kernel)
- Mean Integrated Squared Error (MISE) is

$$E\left[\int_{-\infty}^{\infty} (f_{n,h}(u) - f(u))^2 du\right] = \int \int_{-\infty}^{\infty} (f_{n,h,x}(u) - f(u))^2 (f(x))^n dudx$$

where $f(x)$ is the true density function and observations are independent
- For $f(x)$ being the Normal density, the MISE is minimized for

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \cdot s \cdot n^{-\frac{1}{5}} \quad [\text{Normal reference method}]$$

See R script
Kernel density estimation (KDE)

- **A problem.** The choice of the bandwidth $h$ is critical (and it may depend on the kernel)
- Automatic selection of $h$
  - Plug-in selectors
  - Cross-validation selectors
- **Another problem.** When the support is finite, symmetric kernels give meaningless results
- Boundary kernels
  - Kernel (truncation) and renormalization
  - Linear (combination) kernel
  - Beta boundary kernels
  - Reflective kernels (density=0 at boundaries)

See R script
The empirical CDF

- Empirical cumulative distribution function (CDF):

\[ F_n(x) = \frac{\left| \{ i \in [1, n] \mid x_i \leq x \} \right|}{n} \]

- Empirical complementary cumulative distribution function (CCDF):

\[ \bar{F}_n(x) = 1 - F_n(x) \]

See R script