Master Program in *Data Science and Business Informatics*

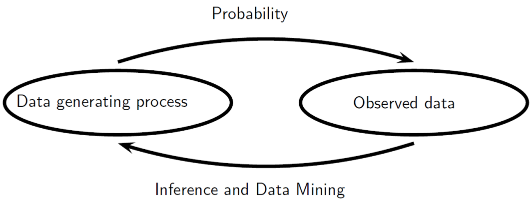
Statistics for Data Science

Lesson 15 - Graphical summaries

Salvatore Ruggieri

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

Condensed observations: graphical summaries



- Probability models governs some random phenomena
- Confronted with a new phenomenon, we want to learn about the randomness associated with it
  - Parametric (efficient) vs non-parametric (general) methods
- Record observations $x_1, \ldots, x_n$ (a dataset)
- n can be large: need to condense for easy visual comprehension
- Graphical summaries:
  - Univariate: empirical distribution functions, histograms, kernel density estimates
  - Multi-variate: kernel density estimates, scatter plots

The empirical CDF

- A r.v. X is completely characterized by its CDF F
- Record observations $x_1, \ldots, x_n$ (a dataset)
- Empirical cumulative distribution function (ECDF):

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

- Empirical complementary cumulative distribution function (ECCDF): $\bar{F}_n(x) = 1 - F_n(x)$
- Estimating F through F_n **[Glivenko-Cantelli Thm]**

$$P(\lim_{n \to \infty} \sup_x |F(x) - F_n(x)| = 0) = 1$$

allow for estimating other quantities by plugging F_n in the place of F, e.g., $E[X]$ as

$$E[X] = \sum_a a \cdot P(X = a) \approx \sum_a a \cdot \frac{|\{i \mid x_i = a\}|}{n} = \frac{1}{n} \sum_i x_i$$

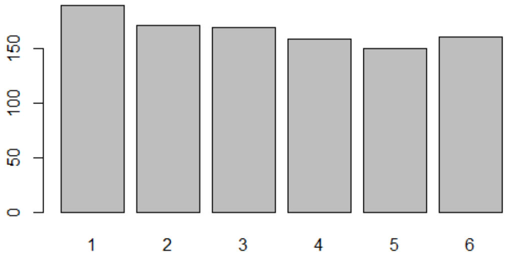
- What about p.m.f. and d.f.?

See R script

p.m.f.: Barplots

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

$$P(X = a) \approx \frac{|\{i \mid x_i = a\}|}{n}$$



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

See R script

d.f.: Histograms

- Histograms provide frequency counts for ranges of values.
- Split the support to m intervals, called *bins*:

$$B_1, \dots, B_m$$

  where the length $|B_i|$ is called the *bin width*
- Count observations in each bin and normalize them:

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

- Plot bars whose **area** is proportional to A_i

$$A_i = |B_i| \cdot H_i \qquad H_i = \frac{|\{j \in [1, n] \mid x_j \in B_i\}|}{n|B_i|}$$

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

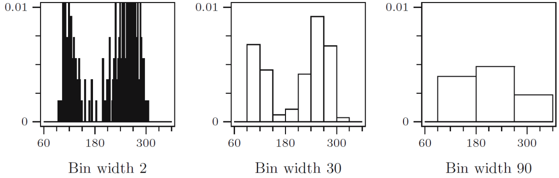


Fig. 15.2. Histograms of the Old Faithful data with different bin widths.

- Mean Integrated Square Error (MISE), for $\hat{f}$ density estimation of f:

$$MISE = E[\int (\hat{f}(t) - f(t))^2 dt] = \int \int (\hat{f}(t) - f(t))^2 f(x_1) \dots f(x_n) dt dx_1 \dots dx_n$$

- Scott's normal reference rule (minimize MISE for Normal density):

$b = 3.49 \cdot s \cdot n^{-1/3}$, where $s = \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ is the sample standard deviation

Choice of the bin width

- $b = 2 \cdot IQR \cdot n^{-1/3}$, where $IQR = Q_3 - Q_1$ *[Freedman-Diaconis' choice]*
  - It replaces $3.49 \cdot s$ in the Scott's rule by $2 \cdot IQR$ (more robust to outlier)
  - Q_3 is 75% percentile of $x_1, \dots, x_n$
  - Q_1 is 25% percentile of $x_1, \dots, x_n$
- Variable bin width
  - Logarithmic binning in power laws
- Alternative strategy: number of bins given equal bin width b: [**other methods**]
  - $m = \left\lceil \frac{\max x_i - \min x_i}{b} \right\rceil$
  - $m = \lceil \sqrt{n} \rceil$
  - Sturges's formula:
    - assume m bins: $0, 1, \dots, m-1$
    - assume normal distribution of true density
    - approximate normal density as $Bin(n, 0.5)$, hence absolute frequency of i^{th} bin is $\binom{m-1}{i}$
    - total frequency is $n = \sum_{i=0}^{m-1} \binom{m-1}{i} = 2^{m-1}$, hence $m = \lceil \log_2 n \rceil + 1$

N.B. R's `hist` method take bin width as a suggestion, then it rounds bins differently

See R script

d.f.: Kernels

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

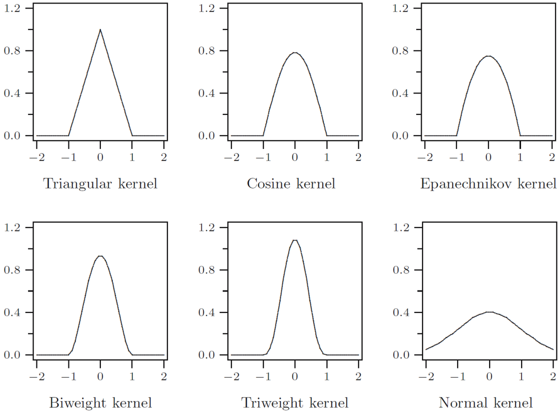


Fig. 15.4. Examples of well-known kernels K.

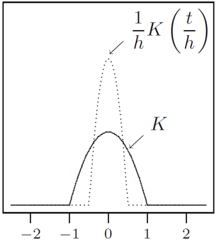
Kernel density estimation (KDE)

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

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

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

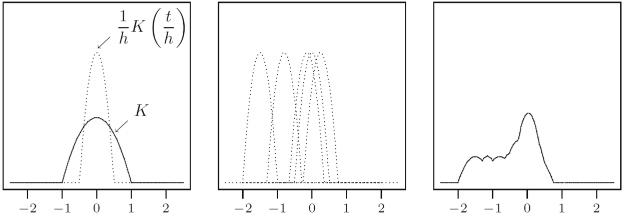
- h controls for how the probability density extends around 0
- if $X \sim K(t)$, then $hX \sim \frac{1}{h}K(\frac{t}{h})$ *[Change-of-units transformation, see Lesson 09]*



Change-of-units transformation. Let X be a continuous random variable with distribution function F_X and probability density function f_X. If we change units to $Y = rX + s$ for real numbers $r > 0$ and s, then

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right) \quad \text{and} \quad f_Y(y) = \frac{1}{r} f_X\left(\frac{y-s}{r}\right).$$

Kernel density estimation (KDE)



Let $x_1, \dots, x_n$ be the observations

- if $X \sim K$, then $hX + x_i \sim \frac{1}{h}K(\frac{t-x_i}{h})$ *[Change-of-units transformation, see Lesson 09]*
- *K scaled and shifted* at x_i, with support $[x_i - h, x_i + h]$

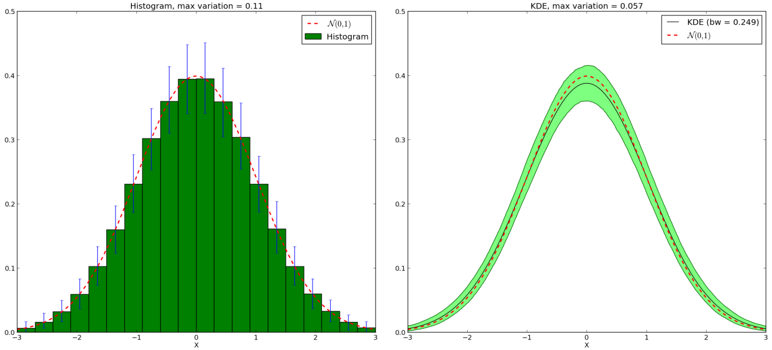
The *kernel density estimate* is defined as the mixture of scaled and shifted kernel densities:

$$f_{n,h}(t) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{t-x_i}{h})$$

- It is a probability density function! **[Prove it!]**

See R script

Histograms vs KDE



- KDE has less variability than histograms!

Choice of the bandwidth

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

$$E[\int_{-\infty}^{\infty}(f_{n,h}(t)-f(t))^2dt] = \int \int_{-\infty}^{\infty}(f_{n,h}(t)-f(t))^2 f(x_1)\dots f(x_n)dtdx_1\dots dx_n$$

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

$$h = (\frac{4}{3})^{\frac{1}{5}} \cdot s \cdot n^{-\frac{1}{5}}$$

[normal reference method]

See R script

Kernel density estimation (KDE)

- **A problem.** The choice of the bandwith h is critical (and it may depend on the kernel)
- Automatic selection of h
  - Plug-in selectors (iterative bandwith selection)
  - Cross-validation selectors (part of data for estimation and part for evaluation)
- **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 [Scott, 2015] for a complete book on KDE

See R script

Optional reference

David W. Scott (2015)
Multivariate density estimation: Theory, practice, and visualization.
John Wiley & Sons, Inc.