Data Valorization:
Data Gathering

Lionel Fillatre

fillatre@unice.fr

Topics

- Statistical Inference
- Sampling
- Kernel Density Estimation
- Random Variate Generation
- Conclusion

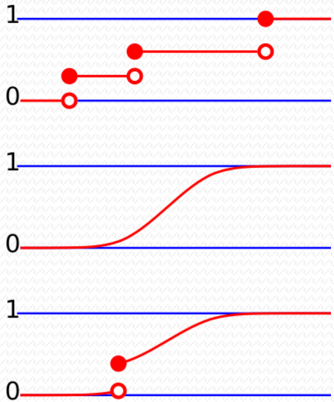
1 Statistical Inference

Recall on Random Variables (rv)

- Cumulative Distribution Function (cdf):

$$F_X(x) = \mathbb{P}(X \le x), \forall x \in \mathbb{R}$$

  - $F_X(x)$: discrete in $x \rightarrow$ discrete rv's
  - $F_X(x)$: continuous function of $x \rightarrow X$ is a continuous rv's.
  - $F_X(x)$: piecewise continuous $\rightarrow$ mixed rv's



- Basic properties
  - $0 \le F_X(x) \le 1, \forall x \in \mathbb{R}$
  - $F_X(x)$ monotonically increasing function of x
  - $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$
  - If $F_X(x)$ is continuous, $\mathbb{P}(X = x) = 0$

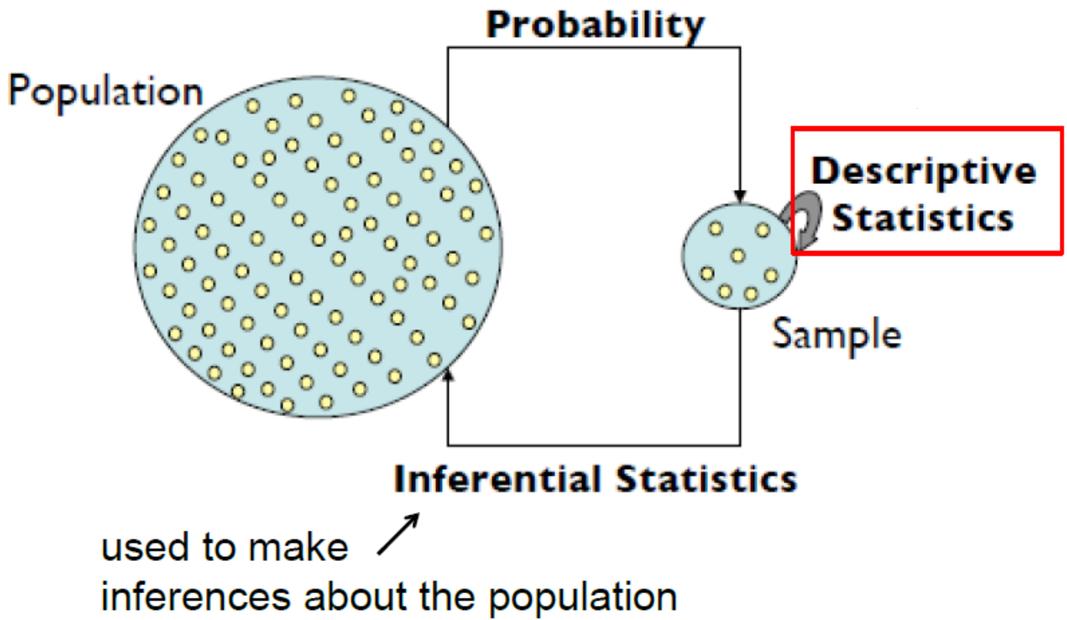
Recall on Probability Density Function (pdf)

- Pdf of continuous rv:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt\,, \forall x \in \mathbb{R}$$

- For continuous rv, $f_X(x) = \frac{dF_X(x)}{dx}$
- Positivity: $f_X(x) \geq 0, \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f_X(t)dt = 1$
- $\mathbb{P}(a < X \leq b)$=$F_X(b) - F_X(a) = \int_a^b f_X(t)dt$

Probability and Inference



Population

- **Population**: A collection of objects for study

- Generally, a (multivariate) distribution $F(x)$ describes the objects considered as a random variable X (or a random vector)

- Example 1:
  - Goal: Study the efficacy of a new malaria vaccine
  - Population: Individuals prone to malarial infection

- Example 2:
  - Goal: Study the pattern of spam mail in Gmail
  - Population: All the possible spam mail that are (and will be) in Google’s servers
  - Note: Objects in the population may not exist! (for example, a new kind of spam mail)

Sample

- Often, we can’t take measurements for every single object in the population
  - Expensive, morally unjustified, etc.
  - May not even exist yet!

- **Sample**: A manageable subset of the population that is representative of the population
  - Measurements from sample denoted as $X_1, \ldots, X_n$

Parameters

- **Parameters**: numerical features/descriptions/characteristics of the population, usually unknown
  - From example 1 (malaria vaccine efficacy):
    - **Distribution** of body temperature for all individuals after vaccination
    - **Average difference** in parasite levels for all individuals before and after vaccination
  - From example 2 (Gmail spam pattern):
    - **Average** word count in spam
    - **Frequency** of spam for each day of the week

- **Formally:** generally, the population is assumed to be infinite. Each object of the population X follows the cumulative distribution function $F_\theta(x)$ where $\theta \in \mathbb{R}^p$ is the parameter describing the population.

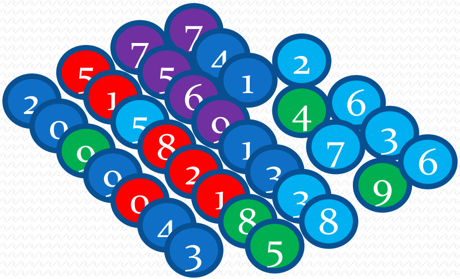
Statistic

- **Statistic**: a function of the sample that is used to estimate/infer about the unknown **parameters**!
  - Examples: sample mean, sample variance, empirical distribution/frequency, etc.

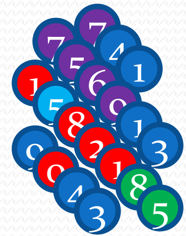
- Generally a statistic is denoted as $T(X_1, \dots, X_n)$. It is a function of the samples!

Population/Parameter and Sample/Statistic

Population



Sample



Features of the population (parameters)

Mean: $\mu = 4.6364$

Distribution:

Red	DBlue	LBlue	Green	Purple
6	9	8	5	5

Estimates of the features (statistics)

Mean: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = 4.619048$

Empirical Distribution/Frequency

Red	DBlue	LBlue	Green	Purple
5	7	1	2	5

2 Sampling

Methods of Collecting Data

- There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:
  - Direct Observation
  - Surveys (pre-election polls, marketing surveys, etc.)
  - Experiments (the main source for Big Data)

Examples of Sampling Strategies

- **Simple Random Sampling (SRS)**: randomly sample n objects from the population
  - Any n-subset of the population is equally likely
  - If objects are randomly sampled with replacement or if the population size is infinite, it is i.i.d. (independent and identically)

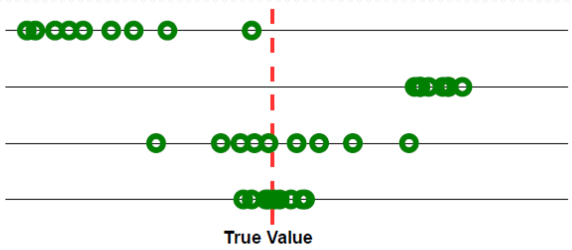
- **Stratified Sampling**: divide the population into K homogenous groups and perform SRS on each group
  - Example 1: Efficacy of malaria vaccine
  - Divide the population into children and adults.

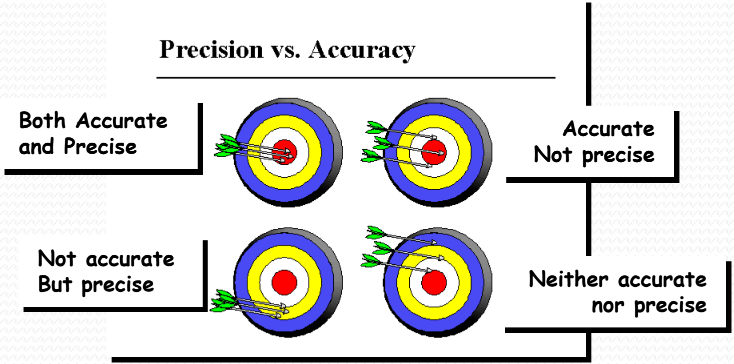
- There exists a numerous number of sampling strategies…

Measurement quality

Assume we make repeated measurements of the same underlying quantity and use this set of values to calculate a mean value (average) that serves as our estimate of the true value.

- **Precision:** The closeness of repeated measurements (of the same quantity) to one another (often measured by standard deviation)
- **Accuracy:** The closeness of measurements to the true value of the quantity being measured.





Analysis of the samples

- Use the summaries presented in the previous lecture:
  - Empirical mean
  - Empirical variance
  - Empirical median
  - Etc.

- An additional tool that represents the distribution of the samples: the empirical cumulative distribution function

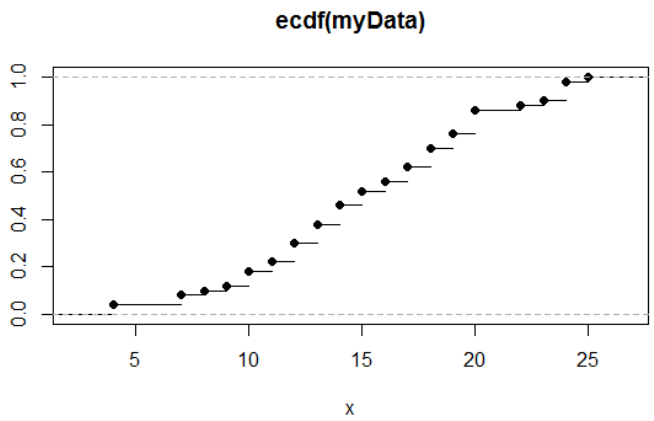
- **Note:** all these quantities are some statistics $T(X_1, \dots, X_n)$, i.e. some functions of the samples.

Distrubution of samples: ecdf

- Measurements: suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed from $F(x) = P(X \leq x)$.
- Empirical cumulative distribution function (ecdf) $\hat{F}_n(x)$:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq x\}$$

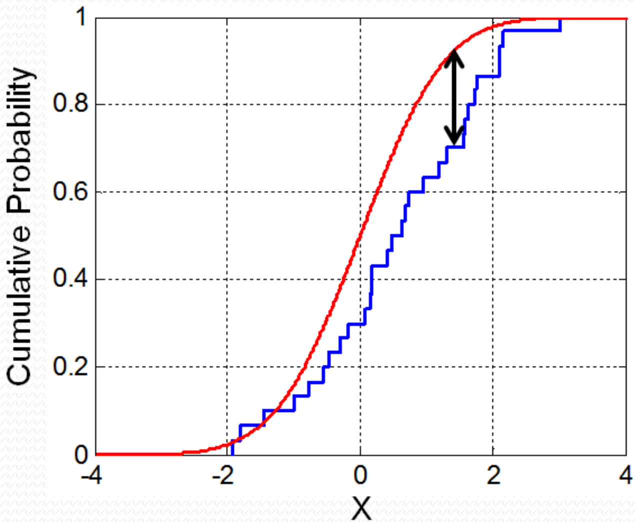
$$1\{X_i \leq x\} = \begin{cases} 1 \text{ if } X_i \leq x \\ 0 \text{ otherwise} \end{cases}$$



Ecdf and cdf

- Dvoretzky–Kiefer–Wolfowitz (DKW) inequality:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F(x)\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \forall\varepsilon > 0$$



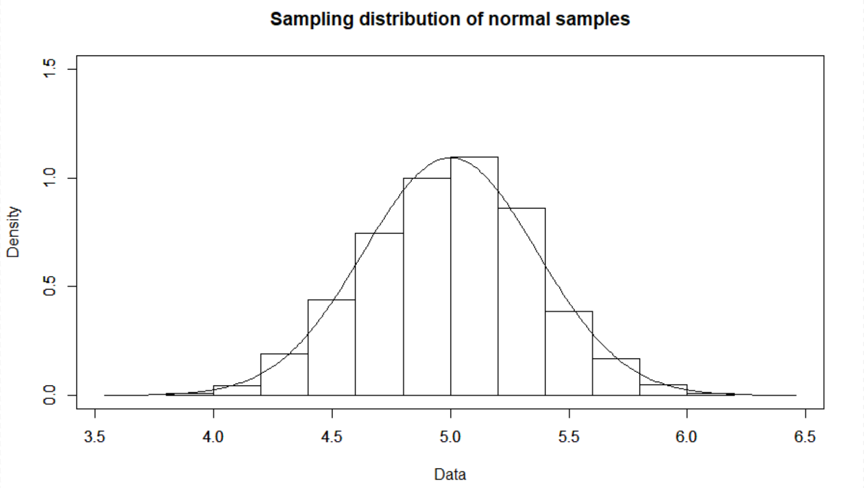
3 Kernel Density Estimation

Kernel Density Esimation (KDE)

- Kernel Density Estimation (KDE) is a non-parametric technique for density estimation
- Low restrictive assumptions about data and underlying probability distributions
- The kernel function is averaged across the observed data points to create a smooth approximation.

Problems with histogram

- The histogram is not continuous athough $f(x)$ is
- The approximation is not accurate



The Naive estimator

- Probability of an interval centered at x:

$$P\big(X \in (x-h, x+h)\big) = \int_{x-h}^{x+h} f(t)dt \approx f(x) \int_{x-h}^{x+h} dt = 2h\, f(x) \qquad \text{(if } h \text{ is small enough)}$$

- Assume we have some samples $X_1, X_2, \cdots, X_n$
- Basic estimation $\hat{P}$:

$$\hat{P}\big(X \in (x-h, x+h)\big) = \frac{[\text{number of } X_i \text{ in } (x-h, x+h)]}{n}$$

- A naive estimator of $f(t)$ around x is

$$\hat{f}(x) = \frac{[\text{number of } X_i \text{ in } (x-h, x+h)]}{n \times 2h}$$

- Rewritten as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \text{ with } K(x) = \begin{cases} \frac{1}{2}, & \text{if } |x| < 1 \\ 0, & \text{otherwise} \end{cases}$$

Improve the Naive estimator

- The naive estimator is easy to compute but
  - It is still not continuous
  - $\hat{f}'(x) = \begin{cases} 0, & \text{if } \hat{f}(x) \text{ is flat} \\ \infty, & \text{if } \hat{f}(x) \text{jumps} \end{cases}$

- This is due to the fact that the uniform kernel $K(x)$ is not smooth.

- Improvement: change the kernel function $K(x)$!

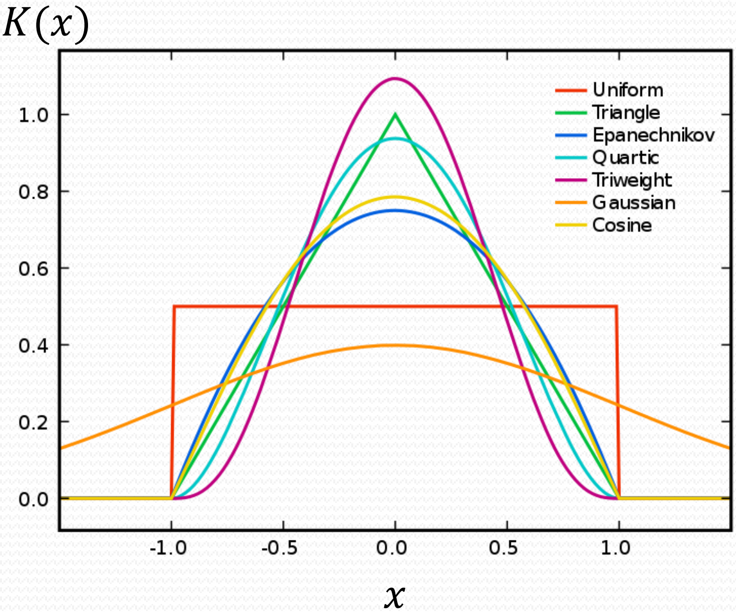
Definition of a Kernel function

- Let K be a non-negative real-valued function such that

1. $\int_{-\infty}^{+\infty} K(x)dx = 1$
2. $\int_{-\infty}^{+\infty} xK(x)dx = 0$
3. $\int_{-\infty}^{+\infty} x^2K(x)dx = \sigma_K^2 > 0$

- K is called a kernel function.
- It is generally a symmetric function.

Some examples



- Triangle:

$$K(x) = 1 - |x| \text{ for } |x| \leq 1$$

- Epanechnikov:

$$K(x) = \frac{3}{4}(1 - x^2) \text{ for } |x| \leq 1$$

- Gaussian:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The Kernel Density Estimator (KDE)

- The KDE is defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

- $K_h(x) = \frac{K\left(\frac{x}{h}\right)}{h}$ is called the scaled kernel.

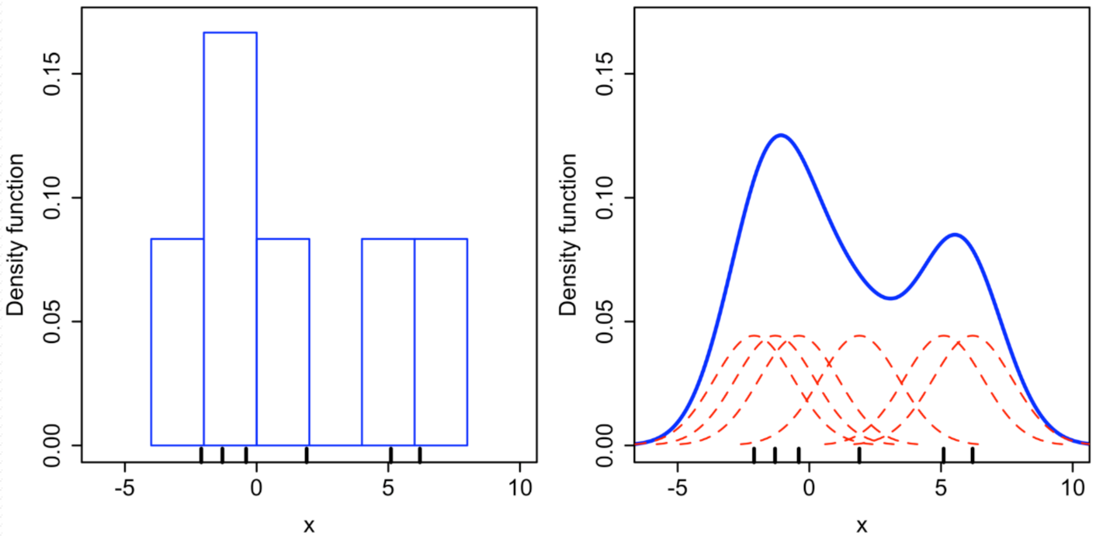
- h is called the bandwidth: it controls the amount of smoothness in the fitted density estimate

- K plays a lesser role (it can be shown) that h in determining the performance of $\hat{f}(x)$. The differentiability of $\hat{f}(x)$ depends on K.

Comparison with histogram

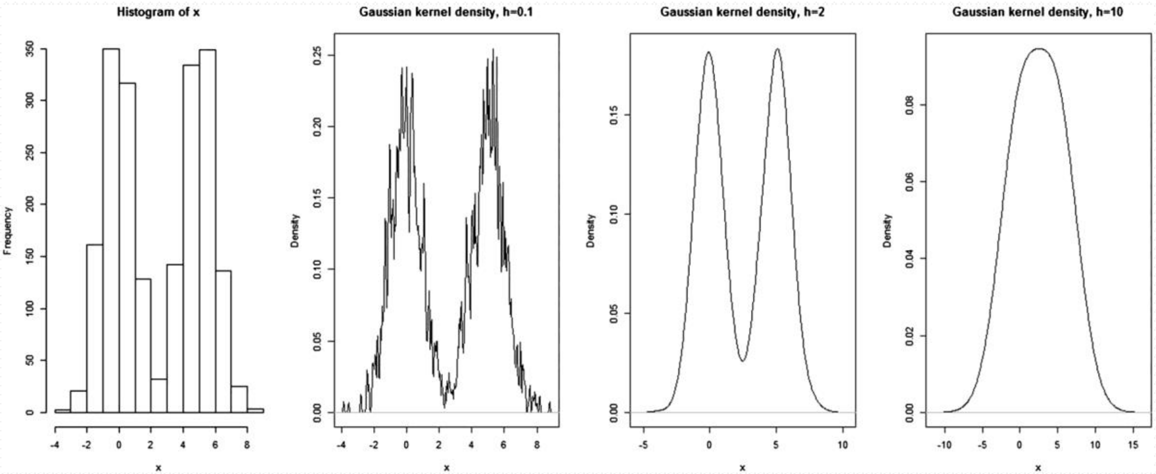
- Assume 6 data points:

$$x_1 = -2.1, x_2 = -1.3, x_3 = -0.4, x_4 = 1.9, x_5 = 5.1, x_6 = 6.2$$



Influence of the bandwidth

- From left to right:
  - the first plot shows simulated data from a mixture of two normal distributions.
  - The second, third, and fourth plots show the Gaussian kernel density estimates using bandwidth values $h = 0.1, h = 2$, and $h = 10$.



Measures of performance

- Mean Squared Error (MSE) – a local measure at x

$$\mathrm{MSE}\left(\hat{f}(x)\right) = E\left(\left(\hat{f}(x) - f(x)\right)^2\right)$$

- Mean Integrated Squared Error (MISE) – a global measure

$$\mathrm{MISE}\left(\hat{f}\right) = \int_a^b \mathrm{MSE}\left(\hat{f}(x)\right) dx$$

4 Random Variate Generation

Random Variate Generation

- It is assumed that a distribution is completely specified and we wish to generate samples from this distribution as input to a simulation model.

- Very useful:
  - Study the distribution numerically
  - Test a code (remove bugs) with simulated datasets
  - Study the reliability of a method
  - Add data to a real dataset (be cautious!)

- Many techniques
  - Inverse Transformation (see in this lecture)
  - Acceptance-Rejection
  - Composition of distributions
  - Etc.

Uniform Source

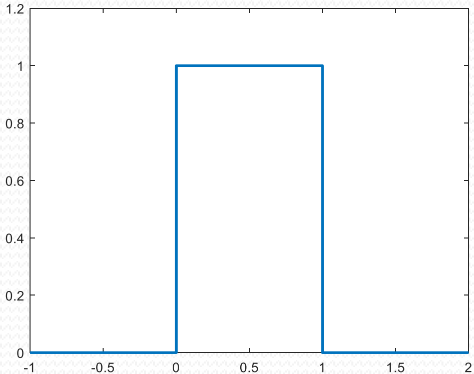
- All these techniques assume that a source of uniform (0,1) random numbers is available: R_1, R_2,…, where each R_i has:

- Pdf: $f_R(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$

- Cdf: $F_R(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$

Uniform Source

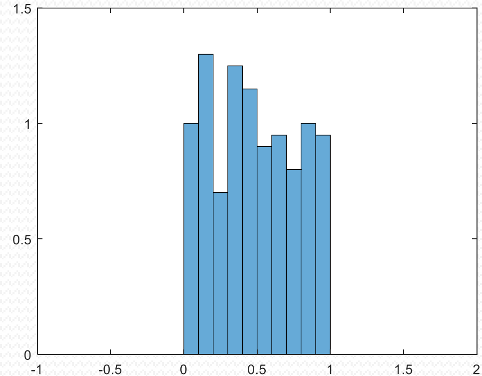
$f_R(x)$



x

Theoretical uniform density on (0,1)

$\hat{f}_R(x)$



x

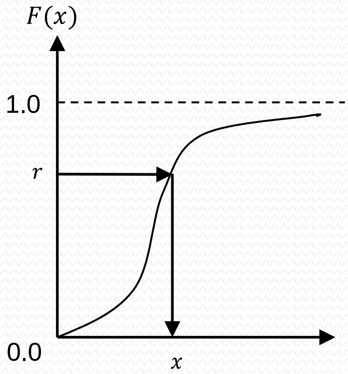
Empirical histogram of 200 uniform random numbers (normalised by the width of the bin and the number of samples)

Inverse-transform Technique

- The concept for generating one sample from cdf $F(x)$:
  - For cdf function: $r = F(x)$
  - Generate r from uniform $(0,1)$
  - Find x such that

$$x = F^{-1}(r)$$

- Following this concept,
  - Generate (as needed) uniform random numbers $r_1, r_2, r_3, \ldots$
  - Compute the many random variates from cdf $F(x)$ by $x_i = F^{-1}(r_i)$



Does X have the desired distribution?

- Check: does the random variable X generated through transformation have the desired distribution?
  - R is uniformly distributed on (0,1)
  - $X = F^{-1}(R)$
  - Remember that $F_R(r) = \mathbb{P}(R \leq r) = r$ for $0 \leq r \leq 1$
  - $F_X(x) = \mathbb{P}(X \leq x)$=$\mathbb{P}(F^{-1}(R) \leq x)$=$\mathbb{P}(R \leq F(x)) = F(x)$

Example: Exponential Distribution

- Exponential Distribution:
  - Exponential pdf: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0, f(x) = 0$ for $x < 0$
  - Exponential cdf: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$

- To generate $X_1, X_2, X_3, \ldots,$ compute

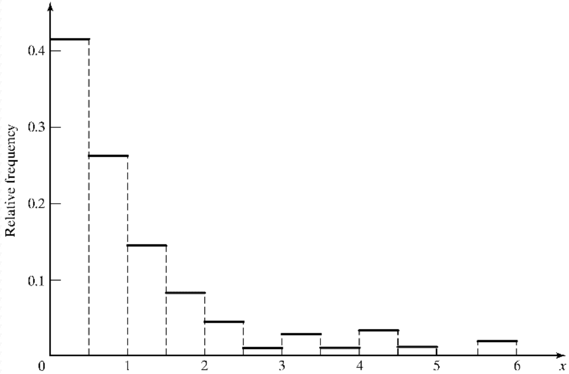
$$X_i = F^{-1}(R_i) = -\frac{1}{\lambda}\log(1 - R_i)$$

- Note that both R_i and $1 - R_i$ are uniformly distributed on $(0,1)$. One simplification is to replace $1 - R_i$ with R_i

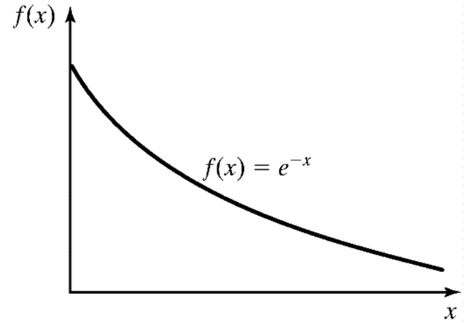
$$X_i = -\frac{1}{\lambda}\log(R_i)$$

Example: Empirical Exponential Distribution

- Generate 200 variates x_i with exponential distribution $\lambda = 1$
- Generate 200 r_i with $U(0,1)$ and compute $x_i = -\frac{1}{\lambda}\log(r_i)$
- The histogram of x_i becomes:



Empirical exponential pdf

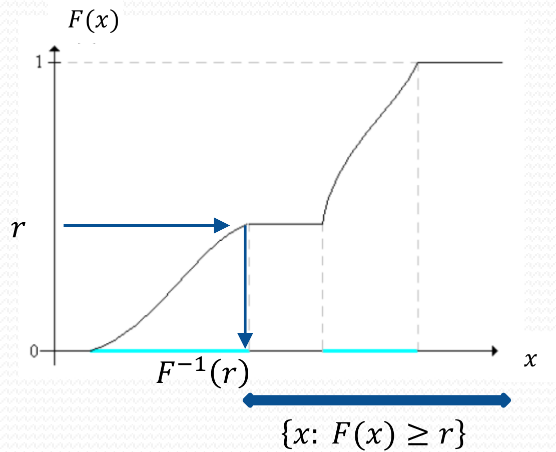


Exponential pdf

Comment on F^{-1}(r)

- $F(x)$ is a non-decreasing function over $\mathbb{R}$
  - If the function is strictly increasing, it is a one-to-one mapping so $F^{-1}(r)$ exists
  - If the function is not strictly increasing, $F^{-1}(r)$ is the generalized inverse function defined by

$$F^{-1}(r) = \inf\{x:\ F(x) \geq r\}$$



Application to Discrete Distribution

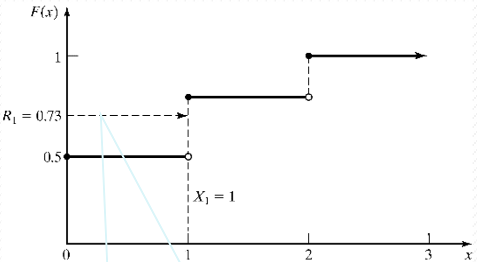
- All discrete distributions can be generated via inverse-transform technique, either numerically through a table-lookup procedure, or algebraically using a formula

- Examples:
  - Empirical
  - Discrete uniform
  - Geometric
  - Gamma

Example: Empirical Discrete Distribution

- Example: Suppose the result of a football team is either 0 (loss), 1 (tie), or 2 (win)
  - Data - Probability distribution:

x	$p(x)$	$F(x)$
0	0.50	0.50
1	0.30	0.80
2	0.20	1.00



  - Method - Given R, the generation scheme becomes:

$$x = \begin{cases} 0, & \text{if } R \leq 0.5 \\ 1, & \text{if } 0.5 < R \leq 0.8 \\ 2, & \text{if } 0.8 < R \leq 1.0 \end{cases}$$

In general, assume discrete values $x_0, x_1, x_2, \ldots$
- generate R
- if $F(x_{i-1}) < R \leq F(x_i)$, set $X = x_i$

Consider $R_1 = 0.73$:
$F(x_0) < R_1 \leq F(x_1)$
Hence, $x_1 = 1$

Example: Geometric Distribution

- Consider the Geometric distribution with probability mass function (pmf)

$$p(x) = p(1-p)^x, \qquad x = 0,1,2,\ldots$$

- It's cdf is given by

$$F(x) = \sum_{j=0}^{x} p(1-p)^x = p\frac{(1-(1-p)^{x+1})}{1-(1-p)} = 1-(1-p)^{x+1}, \qquad x = 0,1,2,\ldots$$

- Using the inverse transform technique, Geometric RV assume the value *x* whenever,

$$F(x-1) = 1-(1-p)^x < R \le (1-p)^{x+1} = F(x)$$
$$\Rightarrow (1-p)^{x+1} \le 1-R < (1-p)^x$$
$$\Rightarrow (x+1)\ln(1-p) \le \ln(1-R) < x\ln(1-p)$$
$$\Rightarrow \frac{\ln(1-R)}{\ln(1-p)} - 1 \le x < \frac{\ln(1-R)}{\ln(1-p)}$$
$$\rightarrow \text{using the round-up function, } X = \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil$$

Distributions with no closed-form inverse

- A number of useful continuous distributions do not have a closed form expression for their cdf or inverse
- Examples are: Normal, Gamma, Beta
- Approximations are possible to inverse cdf
- Other methods exist: acceptance/rejection, composition of distributions,…

5 Conclusion

Conclusion

- Statistical inference:
  - from data to probability
  - from probability to information/knowledge
- Data must be sampled carefully
- Density estimation should be carefully tuned
- Simulated data is important to understand the data, to clean the data, to debug a code and to test algorithms