Chapter 5: Summarizing Data

Mathematical Statistics

UIC

May 15, 2024

Overview

Summarizing Data I

- Methods Based on the CDF
  - The Empirical CDF
    - Example: Data from Uniform Distribution
    - Example: Data from Normal Distribution
  - Statistical Properties of the eCDF
  - The Survival Function
    - Example: Data from Exponential Distribution
  - The Hazard Function
    - Example: The Hazard Function for the Exponential Distribution
- Summary

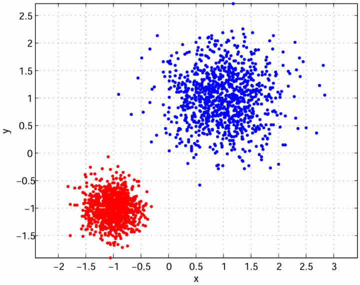
Describing Data

In the next few Lectures we will discuss methods for describing and summarizing data that are in the form of one or more samples. These methods are useful for revealing the structure of data that are initially in the form of numbers.

Example: the arithmetic mean $\bar{x} = (x_1 + \ldots + x_n)/n$ is often used as a summary of a collection of numbers $x_1, \ldots, x_n$: it indicates a "typical value".

Example:

- $x = (1.5147, 1.7223, 1.063, 1.4916, \ldots)$,
  $y = (0.7353, 0.0781, 0.276, 1.5666, \ldots)$



Empirical CDF

Suppose that $x_1, \ldots, x_n$ is a batch of numbers.
Remark: We use the word

- "sample" when $X_1, \ldots, X_n$ is a collection of random variables.
- "batch" when $x_1, \ldots, x_n$ are fixed numbers (realization of sample).

Definition 5.1.1 (Empirical Cumulative Distribution Function)

The **empirical cumulative distribution function** (eCDF) is defined as

$$\boxed{F_n(x) = \frac{1}{n}\left(\# x_i \leq x\right)}$$

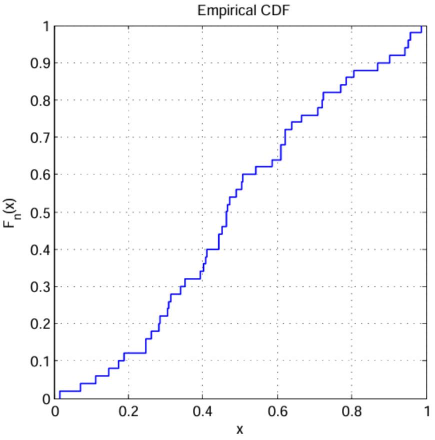
Denote the ordered batch of numbers by $x_{(1)}, \ldots, x_{(n)}$.

- If $x < x_{(1)}$, then $F_n(x) = 0$
- If $x_{(1)} \leq x < x_{(2)}$, then $F_n(x) = 1/n$
- If $x_{(k)} \leq x < x_{(k+1)}$, then $F_n(x) = k/n$

The eCDF is the "data analogue" of the CDF of a random variable

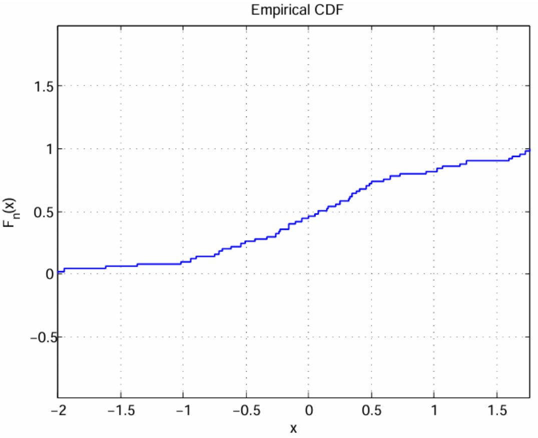
Example: Data from Uniform Distribution

- Let $(X_1, \dots, X_n) \sim U[0,1]$
- Let $(x_1, \dots, x_n)$ is a particular realization of $(X_1, \dots, X_n)$, $n = 50$
  - $(x_1, \dots, x_n) = (0.24733, 0.3527, 0.18786, 0.49064, \dots)$



Example: Data from Normal Distribution

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (-0.23573, 0.45952, -0.93808, -0.62162, \ldots)$



Statistical Properties of the eCDF

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution F.
Then the eCDF can be written as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}\left(X_i\right),$$

where

$$I_{(-\infty,x]}\left(X_i\right) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}$$

The random variables $I_{(-\infty,x)}\left(X_1\right), \ldots, I_{(-\infty,x)}\left(X_n\right)$ are independent Bernoulli random variables:

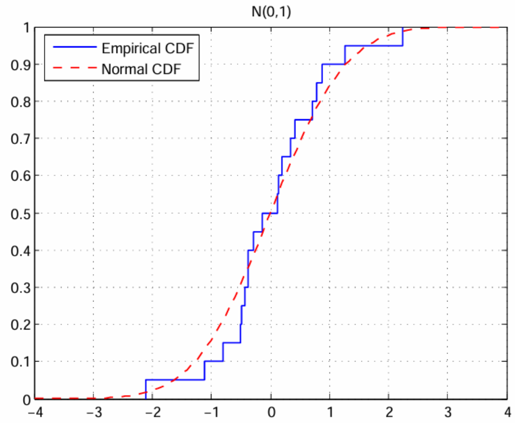
$$I_{(-\infty,x)}\left(X_i\right) = \begin{cases} 1, & \text{with probability } F(x) \\ 0, & \text{with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable: $nF_n(x) \sim \text{Bin}(n, F(x))$

- $\mathbb{E}\left[F_n(x)\right] = F(x)$
- $\mathbb{V}\left[F_n(x)\right] = \frac{1}{n} F(x)(1 - F(x))$
- $\mathbb{V}\left[F_n(x)\right] \to 0$, as $n \to \infty$

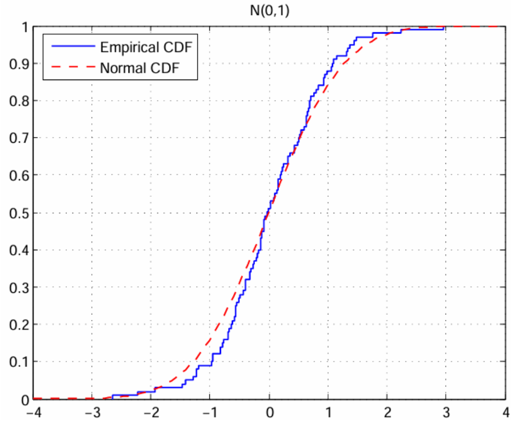
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 20$



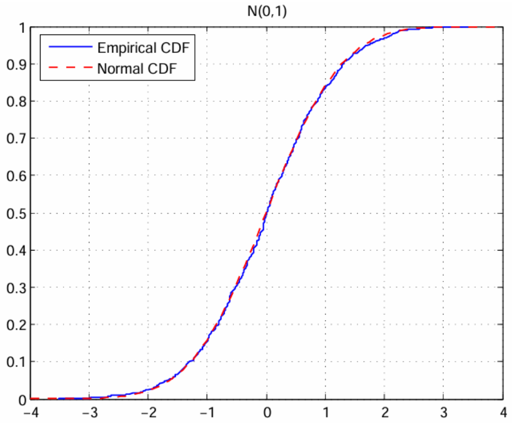
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 100$



Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 1000$



The Survival Function

The **survival function** is equivalent to the CDF and is defined as

$$\boxed{S(t) = \mathbb{P}(T > t) = 1 - F(t)}$$

In applications where the data consists of times until failure or death (and are thus nonnegative), it is often customary to work with the survival function rather than the CDF, although the two give equivalent information.
Data of this type occur in

- medical studies
- reliability studies

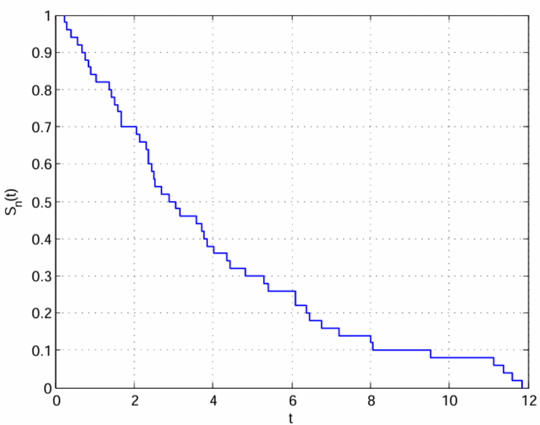
$$\boxed{S(t) = \text{Probability that the lifetime will be longer than } t}$$

The data analogue of $S(t)$ is the **empirical survival function**:

$$\boxed{S_n(t) = 1 - F_n(t)}$$

Example: Data from Exponential Distribution

- Let $(X_1, \ldots, X_n) \sim \text{Exp}(\beta), \beta = 5$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (4.4356, 1.684, 11.376, 4.8357, \ldots)$



The Hazard Function

Let T be a random variable (time) with the CDF F and PDF f.

Definition 5.1.2

The **hazard function** is defined as

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}}$$

- The **hazard function** may be interpreted as the instantaneous death rate for individuals who have survived up to a given time: if an individual is alive at time t, the probability that individual will die in the time interval $(t, t + \epsilon)$ is

$$\mathbb{P}(t \leq T \leq t + \epsilon \mid T \geq t) \approx \frac{\epsilon f(t)}{1 - F(t)}$$

- If T is the lifetime of a manufactured component, it maybe natural to think of $h(t)$ as the age-specific failure rate. It may also be expressed as

$$h(t) = -\frac{d}{dt} \log S(t)$$

Example: Hazard Function for the Exponential Distribution

Let $T \sim \text{Exp}(\beta)$, then

- $f(t) = \beta e^{-\beta t}$
- $F(t) = 1 - e^{-\beta t}$
- $S(t) = e^{-\beta t}$
- $h(t) = \beta$

The instantaneous death rate is constant.

If the exponential distribution were used as a model for the lifetime of a component, it would imply that the probability of the component failing did not depend on its age.

Typically, a hazard function is U-shaped:

- the rate of failure is high for very new components because of flaws in the manufacturing process that show up very quickly,
- the rate of failure is relatively low for components of intermediate age,
- the rate of failure increases for older components as they wear out.

Summary

- The empirical cumulative distribution function (eCDF) is

$$\boxed{F_n(x) = \frac{1}{n}\left(\#x_i \leq x\right)}$$

- The survival function is equivalent to the CDF and is defined as

$$\boxed{S(t) = \mathbb{P}(T > t) = 1 - F(t)}$$

- The data analogue of $S(t)$ is the empirical survival function:

$$\boxed{S_n(t) = 1 - F_n(t)}$$

- The hazard function is

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}}$$

  - may be interpreted as the instantaneous death rate for individuals who have survived up to a given time

Summaring Data II

- Quantile-Quantile Plots
- Histograms
- Kernel Probability Density Estimate
- Summary

Quantile-Quantile Plots

Quantile-Quantile (Q-Q) plots are used for comparing two probability distributions.

Suppose that X is a continuous random variable with a strictly increasing CDF F.

Definition 5.2.3

The p^{th} **quantile** of F is that value x_p such that

$$F(x_p) = p \quad \text{or} \quad \boxed{x_p = F^{-1}(p)}$$

Suppose we want to compare two CDF: F and G.

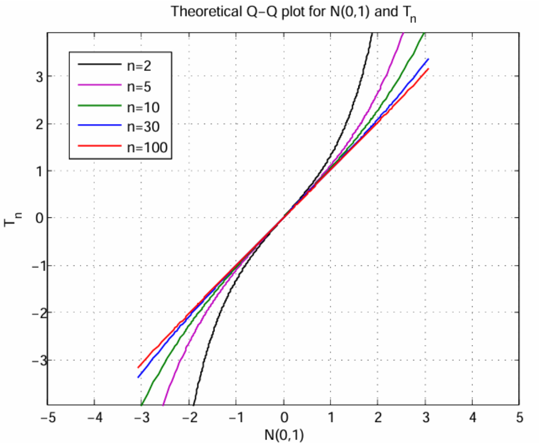
Definition 5.2.4

The **theoretical Q-Q plot** is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$, that is the graph $\left[F^{-1}(p), G^{-1}(p)\right]$ for $p \in (0,1)$.

- If the two CDFs are identical, the theoretical Q-Q plot will be the line $y = x$.

Example of a Theoretical Q-Q plot

- $F = \mathcal{N}(0,1)$
- $G = T_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi_n^2/n}}$, t-distribution with n degrees of freedom.
- We know that $T_n \to \mathcal{N}(0,1)$ as $n \to \infty$.



Properties Q-Q plots

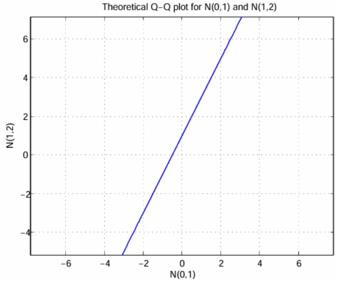
Theorem 5.2.5

If $G(x) = F\left(\frac{x-\mu}{\sigma}\right)$ *for some constants* μ *and* $\sigma \neq 0$*, then*

$$\boxed{y_p = \mu + \sigma x_p}$$

- Thus, if two distributions differ only in location and/or scale, the theoretical Q-Q plot will be a straight line with slope σ and intercept μ.

Example: Let $F = \mathcal{N}(0,1)$ and $G = \mathcal{N}(1,2)$, then $G(x) = F\left(\frac{x-1}{\sqrt{2}}\right)$.



Empirical Q-Q plots

In practice, a typical scenario is the following:

- $F(x) = F_0(x)$ is a specified CDF which is a theoretical model for data $X_1, \ldots, X_n$.
- $G(x)$ is the empirical CDF for $x_1, \ldots, x_n$, a realization of $X_1, \ldots, X_n$ (actually observed data).
- We want to compare the model $F(x)$ with the observation $G(x)$.

Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered batch. Then

Definition 5.2.6

The **empirical Q-Q plot** is the plot of $F_0^{-1}(i/n)$ on the horizontal axis versus $G^{-1}(i/n) = x_{(i)}$ on the vertical axis, for $i = 1, \ldots, n$.

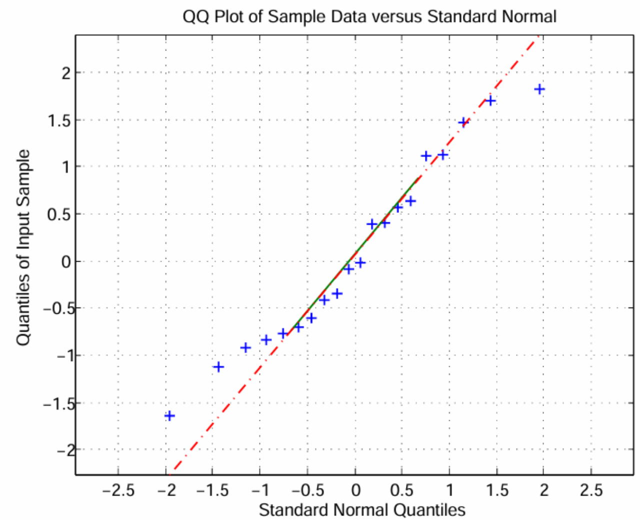
Remarks:

- The quantities $p_i = i/n$ are called plotting positions
- At $i = n$, there is a technical problem since $F_0^{-1}(1) \to \infty$.
- Many software packages graph the following as the empirical Q-Q plot:

$$\left\{ \left(F_0^{-1}\left(\frac{i - 0.375}{n + 0.25}\right), x_{(i)} \right) \right\}, \quad i = 1, \ldots, n$$

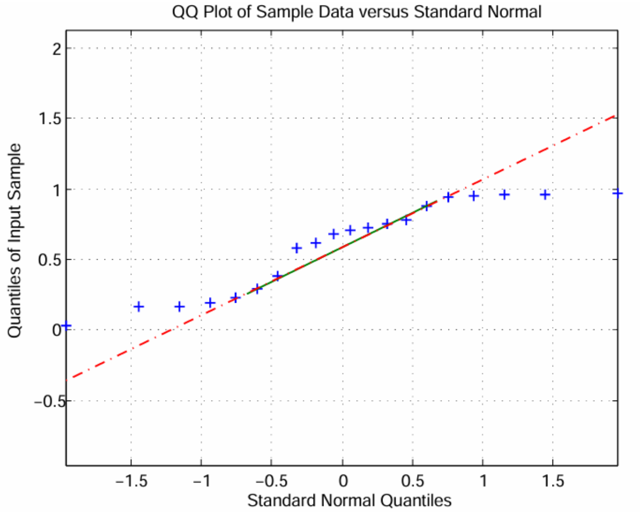
Example of an Empirical Q-Q plot

- $F_0 = \mathcal{N}(0,1)$, a model.
- $X_1, \ldots, X_{20} \sim \mathcal{N}(0,1)$.



Example of an Empirical Q-Q plot

- $F_0 = \mathcal{N}(0, 1)$, a model.
- $X_1, \ldots, X_{20} \sim U[0, 1]$.



Histograms

Histogram displays the shape of the distribution of data values.

Histograms are constructed in the following way:

1. The range of data $x_1, \ldots, x_n$ is divided into several intervals, called bins
2. The number of the observations falling in each bin is then plotted.

Remarks:

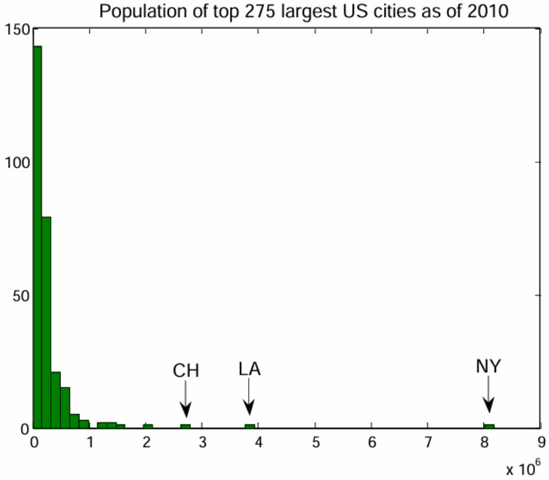
- The total area of the histogram is equal to the sample size n.
- A histogram may also be normalized displaying the proportion of observations falling in each bin. In this case, the area under the histogram is 1.

Applications:

- Histograms are frequently used to display data for which there is no assumption of any probability model. For example, populations of US cities.
- If the data are modeled as a random sample from some continuous distribution, then the normalized histogram may be also viewed as an estimate of the PDF.

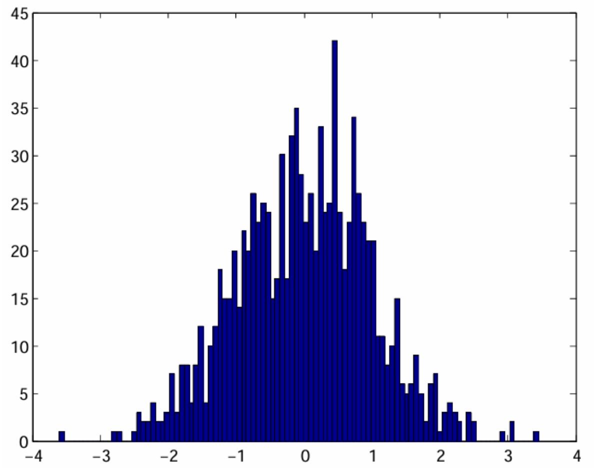
Example: Populations of US Cities

- Data $x_1, \ldots, x_{275}$ are populations of the top 275 largest US cities.
- Data source: wikipedia.org
- Number of bins: 50



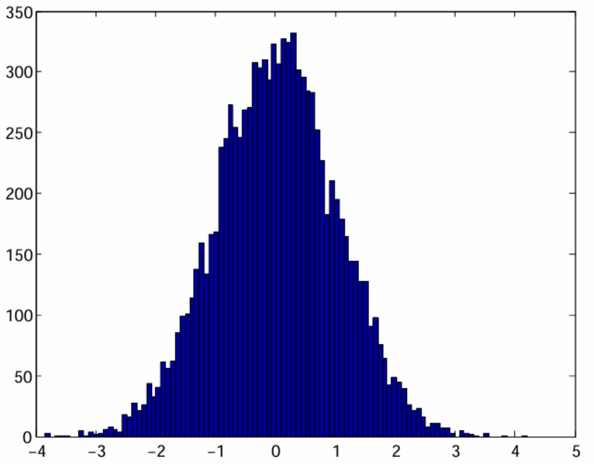
Example: Histogram Approximates PDF

- \- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^3$
- Number of bins: 100



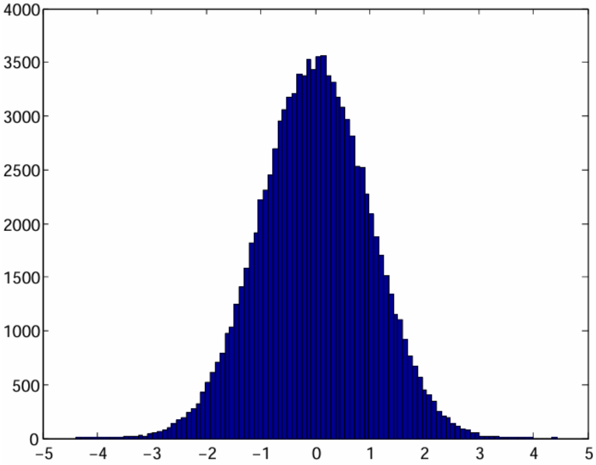
Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0, 1)$, $n = 10^4$
- Number of bins: 100



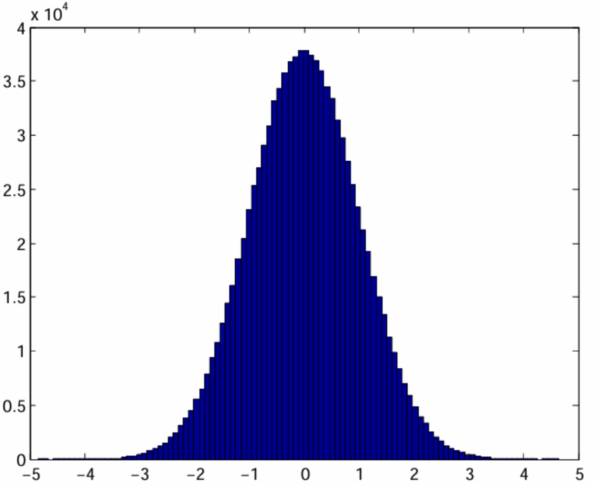
Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 10^5$
- Number of bins: 100



Example: Histogram Approximates PDF

- $X_1, \ldots, X_n \sim \mathcal{N}(0, 1)$, $n = 10^6$
- Number of bins: 100



The main drawback of estimating PDFs by histograms is that these estimates are not smooth. A smooth probability density estimate can be constructed in the following way. Let $w(x)$ be a nonnegative, symmetric and smooth weight function, centered at zero and integrating to 1. For example, $w(x) = \mathcal{N}(x \mid 0, 1)$ – pdf of $\mathcal{N}(0, 1)$. The function

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

is a re-scaled version of $w(x)$.

- As $h \to 0$, $w_h(x)$ becomes more concentrated and peaked about zero.
- As $h \to \infty$, $w_h(x)$ becomes more spread out and flatter.
- If $w(x) = \mathcal{N}(x \mid 0, 1)$, then $w_h(x) = \mathcal{N}\left(x \mid 0, h^2\right)$ – pdf of $\mathcal{N}(0, h^2)$

Definition 5.2.7 (Kernel Probability Density Estimate)

If $X_1, \ldots, X_n \sim \pi$, then an estimate of π is

$$\boxed{\pi_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h\left(x - X_i\right)}$$

This estimate is called a **kernel probability density estimate**.

Kernel Probability Density Estimate

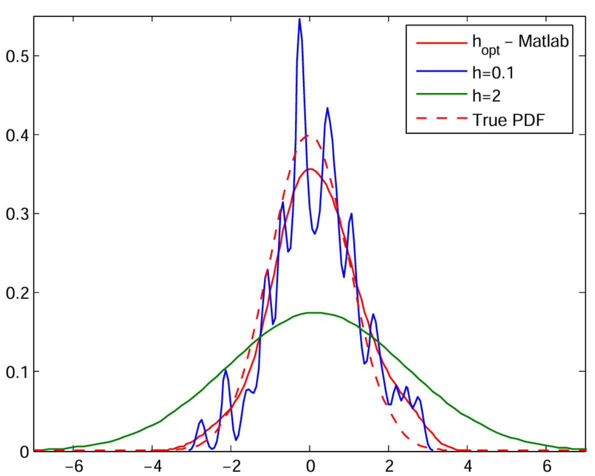
$$\boxed{\pi_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h\left(x - X_i\right)}$$

Remarks:

- $\pi_h(x)$ consists of the superposition of "hills" centered on the observations.
- If $w(x) = \mathcal{N}(x \mid 0, 1)$, then $w_h\left(x - X_i\right) = \mathcal{N}\left(x \mid X_i, h^2\right)$ – pdf of $\mathcal{N}(X_i, h^2)$.
- The parameter h is called the bandwidth. It controls the smoothness of $\pi_h(x)$ and corresponds to the bin width of the histogram:
  - if h is too small, then $\pi_h(x)$ is too rough,
  - if h is too large, then the shape of $\pi_h(x)$ is smeared out too much.

Example

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, $n = 100$
- $w(x) = \mathcal{N}(x \mid 0, 1) \quad \Rightarrow \quad w_h(x - X_i) = \mathcal{N}\left(x \mid X_i, h^2\right)$



Summary

- Quantile-Quantile (Q-Q) plots are used for comparing two distributions.
  - The p^{th} quantile x_p of the CDF F is $x_p = F^{-1}(p)$
  - The theoretical Q-Q plot is the graph of the quantiles of a the CDF F, $x_p = F^{-1}(p)$, versus the corresponding quantiles of the CDF G, $y_p = G^{-1}(p)$.
  - If $F = G$, then the theoretical $Q - Q$ plot will be the line $y = x$.
  - If $G(x) = F\left(\frac{x-\mu}{\sigma}\right)$ for some constants μ and $\sigma \neq 0$, then $y_p = \mu + \sigma x_p$.
  - The empirical Q-Q plot is the plot of $F_0^{-1}(i/n)$ on the horizontal axis versus $x_{(i)}$ on the vertical axis.
- Histogram displays the shape of the distribution of data values.
  - Histograms are frequently used to display data for which there is no assumption of any probability model.
  - Normalized histogram may be also viewed as a non-smooth estimate of PDF.
- Kernel Probability Density Estimate: If $X_1, \ldots, X_n \sim \pi$, then an estimate of π is

$$\pi_h(x) = \frac{1}{n} \sum_{i=1}^{n} w_h\left(x - X_i\right)$$

  - If $w(x) = \mathcal{N}(x \mid 0, 1)$, then $w_h\left(x - X_i\right) = \mathcal{N}\left(x \mid X_i, h^2\right)$.
  - h is the bandwidth.

Summarizing Data III

- Measures of Location
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates
- Measures of Dispersion
  - Sample Standard Deviation
  - Interquartile Range (IQR)
  - Median Absolute Deviation (MAD)
- Boxplots
- Summary

Measures of Location

In the lectures before, we discussed data analogues of the CDFs and PDFs, which convey visual information about the shape of the distribution of the data.

Next Goal: to discuss simple numerical summaries of data that are useful when there is not enough data for construction of an eCDF, or when a more concise summary is needed.

- A **measure of location** is a measure of the center of a batch of numbers.
  - Arithmetic Mean
  - Median
  - Trimmed Mean
  - M Estimates

Example: If the numbers result from different measurement of the same quantity, a measure of location is often used in the hope that it is more accurate than any single measurement.

The Arithmetic Mean

The most commonly used measure of location is the arithmetic mean,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A common statistical model for the variability of a measurement process is the following:

$$x_i = \mu + \varepsilon_i$$

- x_i is the value of the i^{th} measurement
- μ is the true value of the quantity
- ε_i is the random error, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$

The arithmetic mean is then:

$$\bar{x} = \mu + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i, \quad \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

The Median

The main drawback of the arithmetic mean is it is sensitive to outliers. If fact, by changing a single number, the arithmetic mean of a batch of numbers can be made arbitrary large or small. For this reason, measures of location that are robust, or insensitive to outliers, are important.

Definition 5.3.8

i) If the batch size is an odd number, $x_1, \ldots, x_{2n-1}$, then the **median** $\tilde{x}$ is defined to be the middle value of the ordered batch values:

$$x_1, \ldots, x_{2n-1} \rightsquigarrow x_{(1)} < \ldots < x_{(2n-1)}, \quad \boxed{\tilde{x} = x_{(n)}}$$

ii) If the batch size is even, the median is the average of the two middle values.

Important Remark:

Moving the extreme observations does not affect the sample median at all, so the median is quite robust.

The Trimmed Mean

Another simple and robust measure of location is the **trimmed mean** or **truncated mean**.

Definition 5.3.9

The $100\alpha\%$ trimmed mean is defined as follows:

1. Order the data: $x_1, \ldots, x_n \rightsquigarrow x_{(1)} < \ldots < x_{(n)}$
2. Discard the lowest $100\alpha\%$ and the highest $100\alpha\%$
3. Take the arithmetic mean of the remaining data:

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

where $[s]$ denotes the greatest integer less than or equal to s.

Remark:

- It is generally recommended to use $\alpha \in [0.1, 0.2]$.

M Estimates

Let $x_1, \ldots, x_n$ be a batch of numbers. It is easy to show that

- The mean
$$\bar{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} (x_i - y)^2$$
Outliers have a great effect on mean, since the deviation of y from x_i is measured by the square of their difference.
- The median
$$\tilde{x} = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} |x_i - y|$$
Here, large deviations are not weighted as heavily, that is exactly why the median is robust.

In general, consider the following function:
$$f(y) = \sum_{i=1}^{n} \Psi(x_i, y),$$
where Ψ is called the weight function. M **estimate** is the minimizer of f:
$$y^* = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$$

Measures of Dispersion

A measure of dispersion, or scale, gives a numerical characteristic of the "scatteredness" of a batch of numbers. The most commonly used measure is the sample standard deviation s, which is the square root of the sample variance,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Q: Why $\frac{1}{n-1}$ instead of $\frac{1}{n}$?

A: s^2 is an unbiased estimate of the population variance σ^2. If n is large, then it makes little difference whether $\frac{1}{n-1}$ or $\frac{1}{n}$ is used.

Like the mean, the standard deviation s is sensitive to outliers.

Measures of Dispersion

Two simple robust measures of dispersion are the interquartile range (IQR) and the median absolute deviation (MAD).

- IQR is the difference between the two sample quartiles:

$$\mathrm{IQR} = Q_3 - Q_1$$

  - Q_1 is the first (lower) quartile, splits lowest 25% of batch
  - $Q_2 = \tilde{x}$, cuts batch in half
  - Q_3 is the third (upper) quartile, splits highest 75% of batch

  How to compute the quartile values (one possible method):

  1. Find the median. It divides the ordered batch into two halves. Do not include the median into the halves.
  2. Q_1 is the median of the lower half of the data. Q_3 is the median of the upper half of the data.

- **MAD** is the median of the numbers $|x_i - \tilde{x}|$.

Example

Let the ordered batch be $\{x_i\} = \{1, 2, 5, 6, 9, 11, 19\}$

- $Q_2 = \tilde{x} = 6$
- $Q_1 = 2$
- $Q_3 = 11$

$$\mathrm{IQR} = 9$$

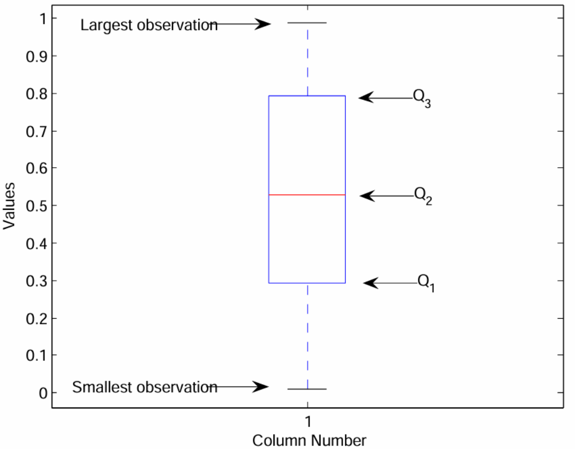
- $\{|x_i - \tilde{x}|\} = \{5, 4, 1, 0, 3, 5, 13\}$

$$\mathrm{MAD} = 4$$

Boxplots

A boxplot is a graphical display of numerical data that is based on five-number summaries: the smallest observation, lower quartile (Q_1), median (Q_2), upper quartile (Q_3), and largest observation.

Example: $x_1, \ldots, x_n \sim U[0,1], n = 100$



Summary

- Measures of Location
  - Arithmetic Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers)
  - Median: the middle value of the ordered batch values $\tilde{x} = Q_2$
  - Trimmed Mean:

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

  - M estimate: $y^* = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^{n} \Psi(x_i, y)$
    - if $\Psi(x_i, y) = (x_i - y)^2$, then $y^* = \bar{x}$
    - it $\Psi(x_i, y) = |x_i - y|$, then $y^* = \tilde{x}$
- Measures of Dispersion
  - Sample Standard Deviation (sensitive to outliers):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$
  - Median Absolute Deviation: MAD = median of the numbers $|x_i - \tilde{x}|$
- Boxplots are useful graphical displays.