Performance of Computer System

Summarising Measured Data

Dr. Lina Xu
lina.xu@ucd.ie

School of Computer Science,
University College Dublin

November 6, 2018

Covered Topics

- Understand why
  - A measurement project may result in several hundred or millions of observations on a given variable. To present the measurements to the decision makers, it is necessary to summarise the data.
- Basic Probabilities and Statistics
- Summarising measured data/Summarising Variability
- Solve a problem using Probability Mass Function (PMF)

Basic Probabilities and Statistics

Independent Events

- When two events are said to be independent of each other, what this means is that the probability that one event occurs in no way affects the probability of the other event occurring.
- An example of two independent events is as follows; say you rolled a die and flipped a coin. The probability of getting any number face on the die in no way influences the probability of getting a head or a tail on the coin.
- If **A** and **B** are dependent events, the probability of this event happening can be calculated as shown below:
  - $P(A \bigcap B) = P(A \bigcup B) - P(A_{only}) - P(B_{only})$
- If **A** and **B** are independent events, the probability of this event happening can be calculated as shown below:
  - $P(A \bigcap B) = P(A_{only}) \times P(B_{only})$

Basic Probabilities and Statistics

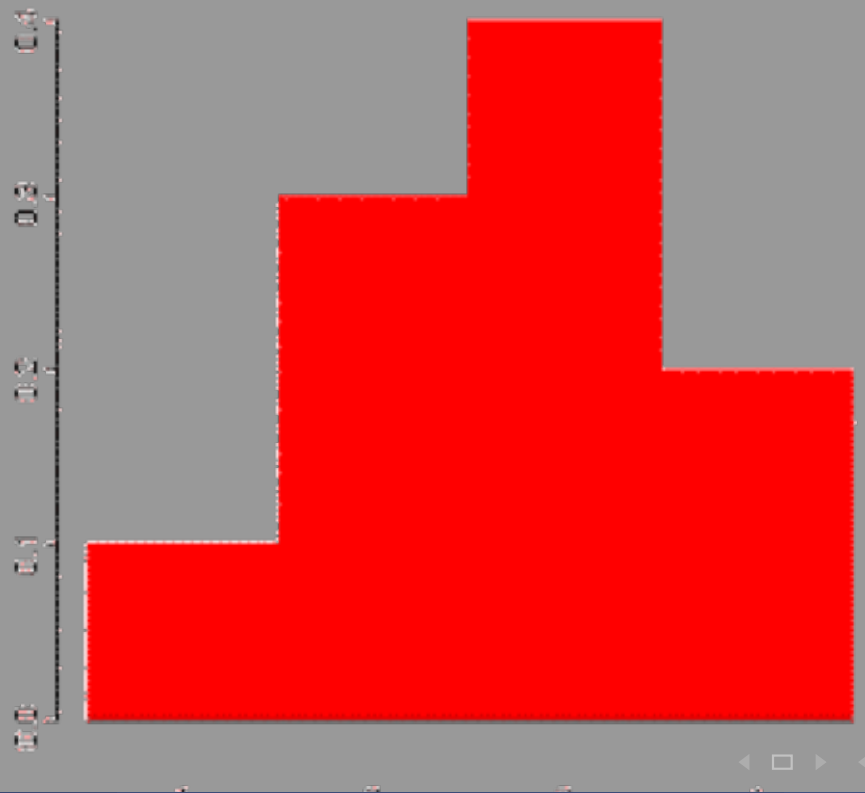
Random Variable

- Arandom variable, usually written$X$, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables,**discrete**and**continuous**.
- Discrete Random Variables
  - Adiscrete random variableis one which may take on only a countable number of distinct values such as 0,1,2,3,4,..... Discrete random variables are usually (but not necessarily) counts.
  - If a random variable can take only a finite number of distinct values, then it must be discrete.
  - Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

Basic Probabilities and Statistics

Probability Mass Function (PMF)

- Theprobability distributionof a discrete random variable is a list of probabilities associated with each of its possible values.
  - Outcome 1 2 3 4
  - Probability 0.1 0.3 0.4 0.2
- This distribution may also be described by theprobability histogram.



Basic Probabilities and Statistics

Probability Density Function (pdf)

- continuous random variableis one which takes an infinite number of possible values. Continuous random variables are usually measurements.
- Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- A continuous random variable is not defined at specific values. Instead, it is defined over anintervalof values, and is represented by thearea under a curve(in advanced mathematics, this is known as anintegral).
- The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Basic Probabilities and Statistics

- The Cumulative Distribution Function (CDF) of a random variable maps a given value a to the probability of the variable taking a value less than or equal to a:
  - $F_x(a) = P(x <= a)$

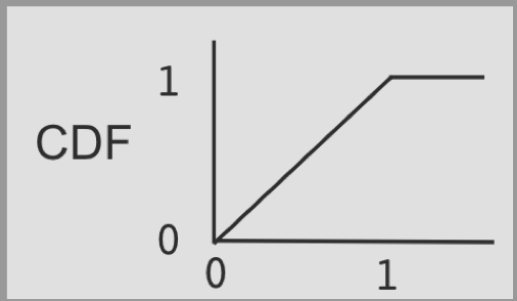
Basic Probabilities and Statistics

If the variable is continues,

- The derivative of the CDF F(x) is the probability density function (pdf) of x.
- Given a pdf f(x), the probability of x being in the interval (x1, x2) can also be computed by integration:

$$f(x) = \frac{dF(x)}{dx}$$

$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$$



Basic Probabilities and Statistics

If the variable is discrete,

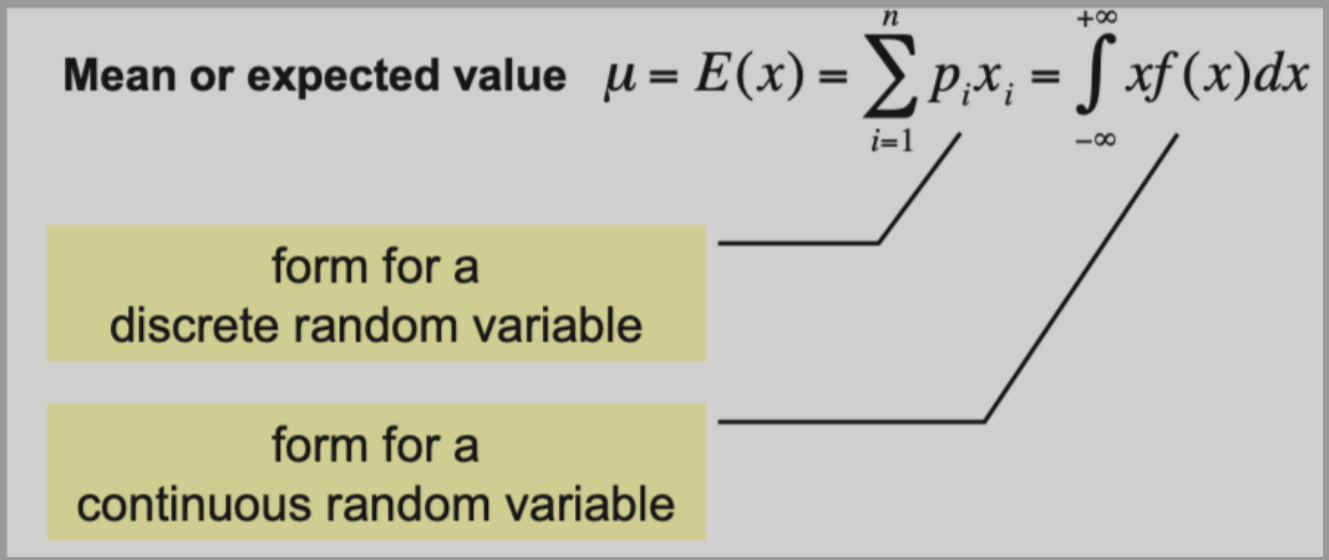
- The CDF is not continuous and, therefore, not differentiable. In such cases, the Probability Mass Function (pmf) is used in place of pdf.

$$f(x_i) = p_i$$

$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1) = \sum_{x_1 < x_i \leq x_2} p_i$$

Basic Probabilities and Statistics

Mean or Expected Value for discrete variables



Basic Probabilities and Statistics

Variance

- Discrete
  - $\sigma^2 = Var(x) = E((x-\mu)^2) = \sum_{i=1}^{n} p_i(x_i - \mu)^2$
- Continuous
  - $\sigma^2 = Var(x) = E((x-\mu)^2) = \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx$
- Coefficient of Variation
  - $C.O.V = \frac{StandardDeviation}{mean} = \frac{\sigma}{\mu}$

Basic Probabilities and Statistics

Covariance

- Given two random variables x and y with means μ_x and μ_y
  - $Cov(x, y) = \sigma_{xy}^2 = E((x - \mu_x)(y - \mu_y))) = E(xy) - E(x)E(y)$
- For independent random variables x and y, $Cov(x, y) = 0$

Correlation Coefficient

- The correlation always lies between -1 and +1
  - $Cor(x, y) = \rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$

Basic Probabilities and Statistics

Mean and Variance of Sums

- $x_1, x_2, ...x_k$ are k random variables
- $a_1, a_2, ..., a_k$ are arbitrary weights
- $E(a_1x_1 + a_2x_2 + ... + a_kx_k) = a_1E(x_1) + a_2E(x_2) + ... + a_kE(x_k)$

Variance of sums for independent variables

- $Var(a_1x_1 + a_2x_2 + ... + a_kx_k) = a_1^2Var(x_1) + a_2^2Var(x_2) + ... + a_k^2Var(x_k)$

α-Quantile

- The x value at which the CDF takes a value α
- $P(x \leqslant X_a) = F(x) = \alpha$

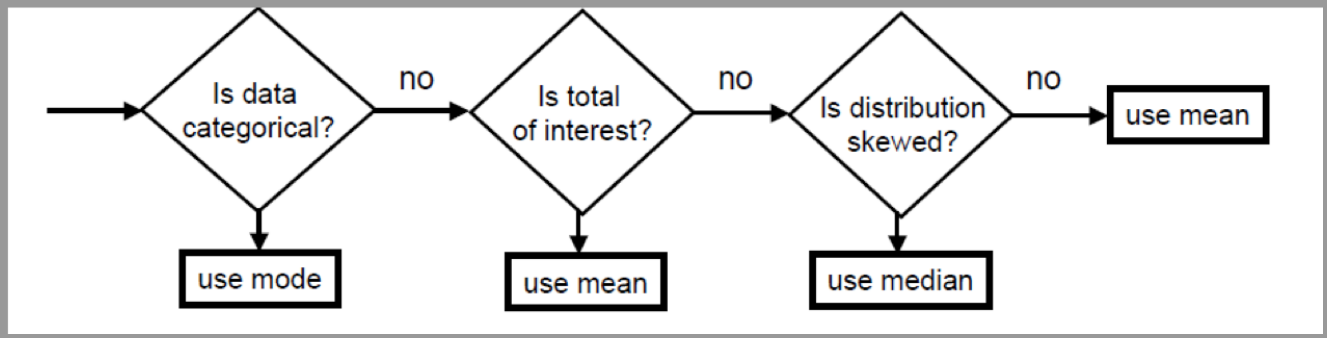
Basic Probabilities and Statistics

Median

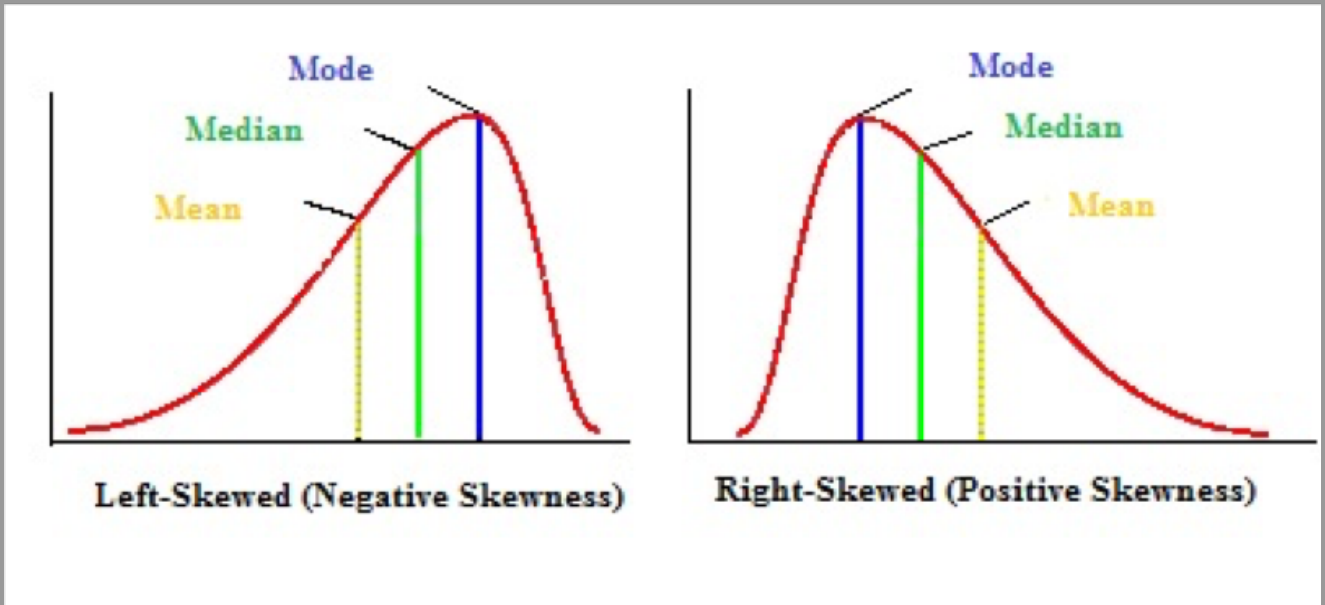
- 0.5-Quantile

Mode

- The most likely value.
- For a discrete variable, the xi that has the highest probability
- For a continuous variable, the x where pdf is maximum



Basic Probabilities and Statistics



Normal Distribution

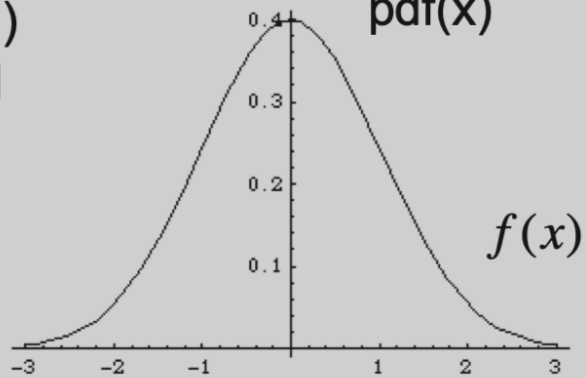
N(μ,ρ) most commonly used distribution in data analysis

pdf = $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty \le x \le \infty$

μ = mean
σ = std dev

(also known as a Gaussian distribution)

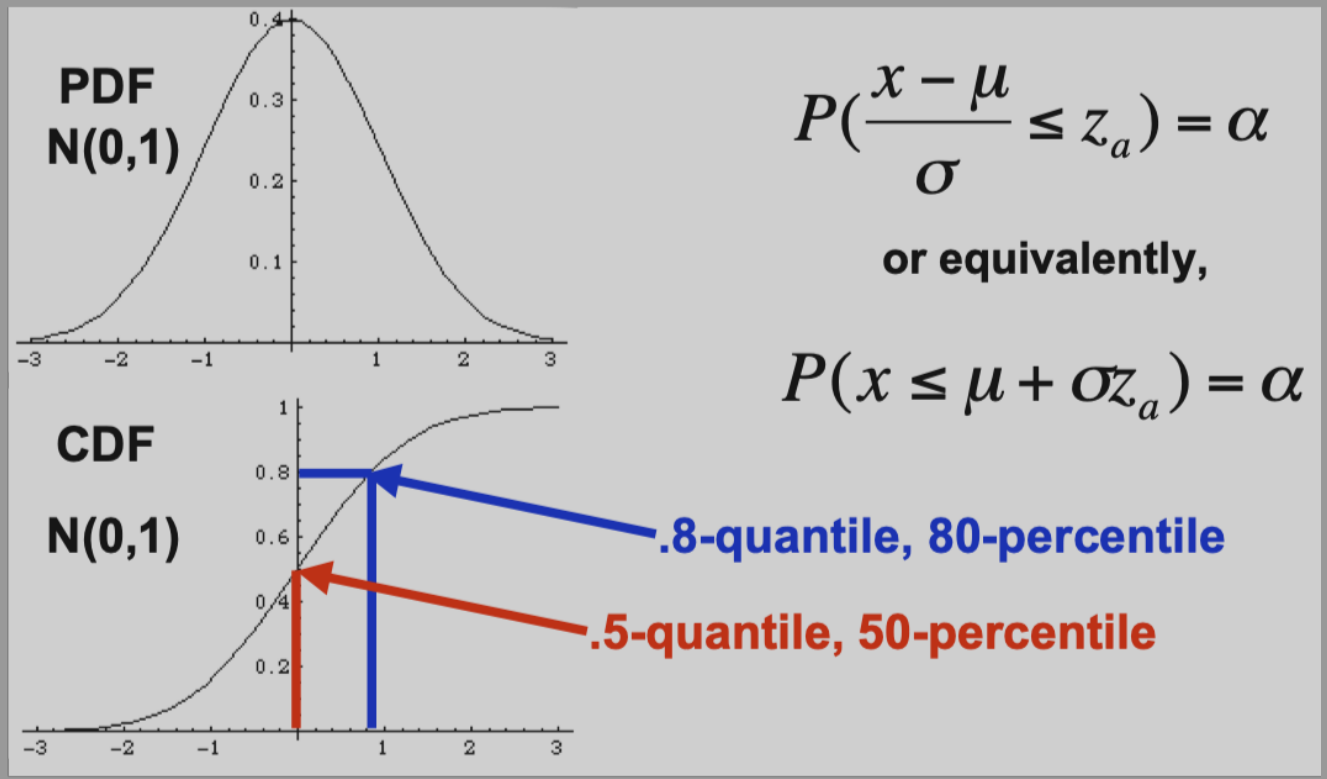
N(μ=0 , σ=1)
unit normal
distribution



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

Quantiles of the Normal Distribution

unit normal variate x N(0,1)



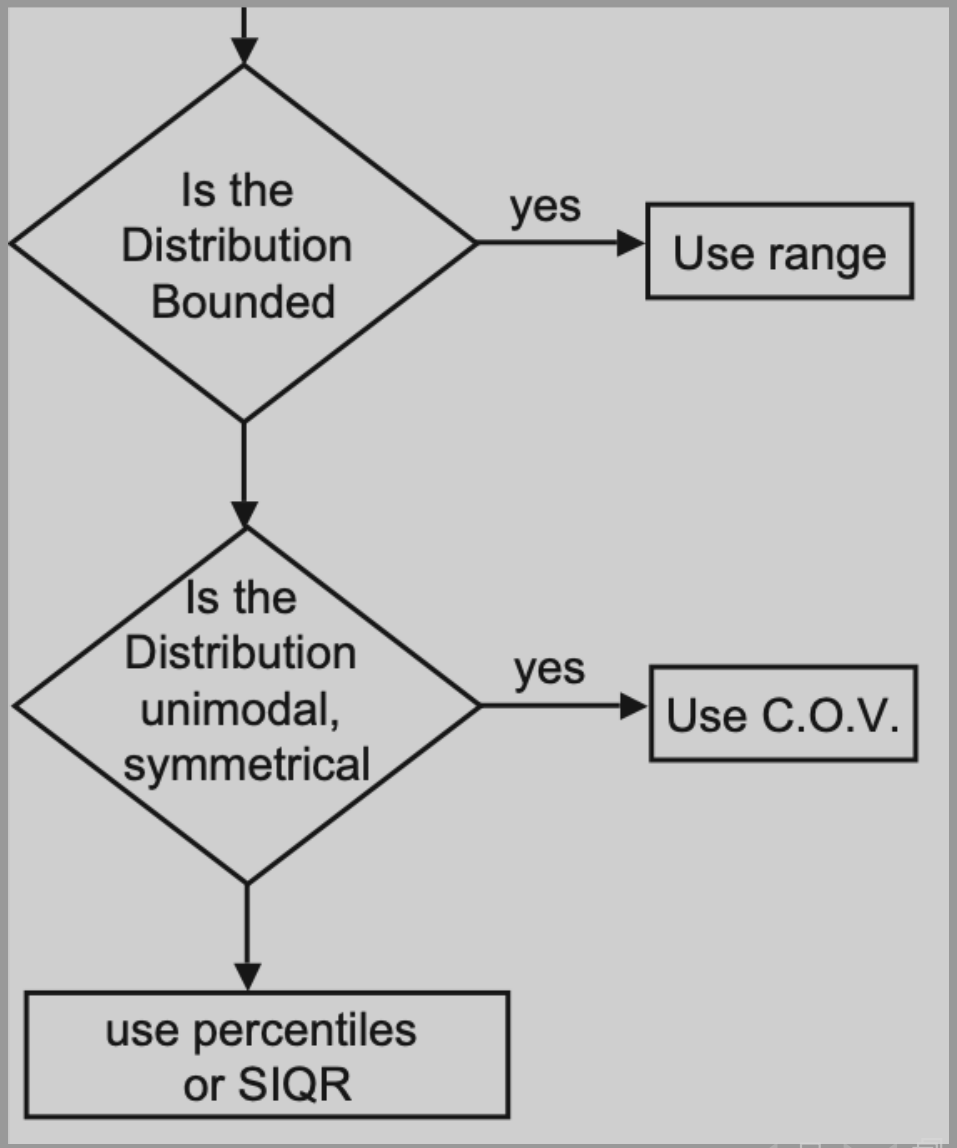
Properties of Normal Distribution

Linearity

- sum of n independent normal variates is a normal variate
- if $x_i \sim N(\mu_i, \sigma_i)$, then
  - $x = \sum_{i=1}^{n} a_i x_i$
- has a normal distribution with mean
  - $\mu = \sum_{i=1}^{n} a_i \mu_i$
- and variance
  - $\sigma^2 = \sum_{i=1}^{n} a_i^2 \mu_i^2$

Summarising Variability

Selecting the Index of Dispersion



Determining Distribution of Data

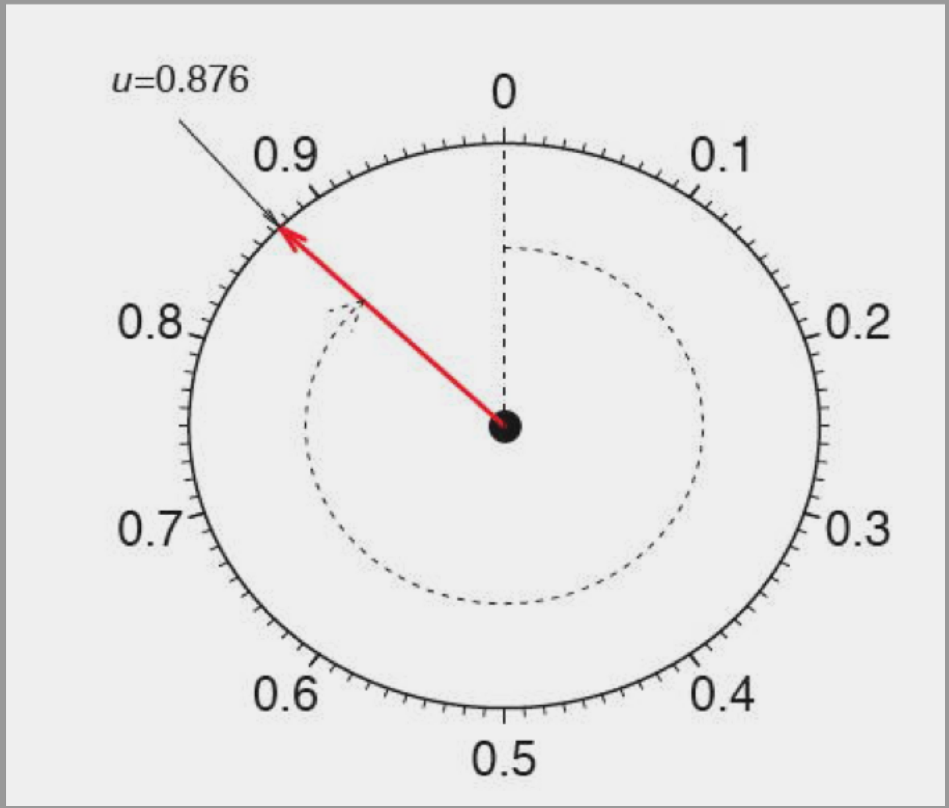
- Can summarise data by its
  - average
  - Variability
- More complete summary: type of distribution
  - e.g. number of I/O calls uniformly distributed 1-25 more meaningful than mean 13, variance is 48
- Distribution useful for simulation or analytical modeling
- How to determine distribution?
  - determine range, divide into cells, plot histogram of observations
  - quantile-quantile plot

Quantile-Quantile Plots

- A simple graphical method for comparing two sets of sample quantiles.
- Sample quantiles are based on order statistics
- Q–Q plots are commonly used to compare a data set to a theoretical model.
  - goodness of fit
  - Difference between measured and predicted values is modeling error
- Q–Q plots are also used to compare two theoretical distributions to each other.

Quantile-Quantile Plots?Example I

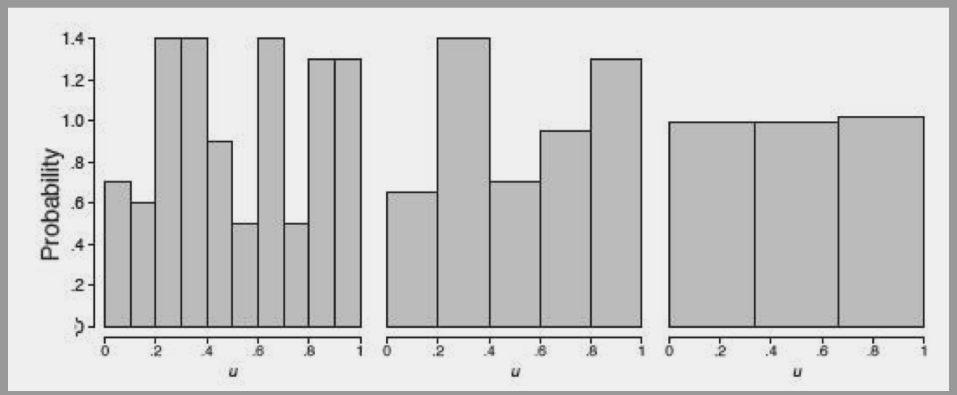
As an example, consider data measured from a physical device such as the spinner depicted. The red arrow is spun around the center, and when the arrow stops spinning, the number between 0 and 1 is recorded. Can we determine if the spinner is fair?



Quantile-Quantile Plots – Example I

To investigate whether the spinner is fair, spin the arrow n times, and record the measurements by $\{\mu_1, \mu_2, ..., \mu_n\}$. In this example, we collect n = 100 samples.

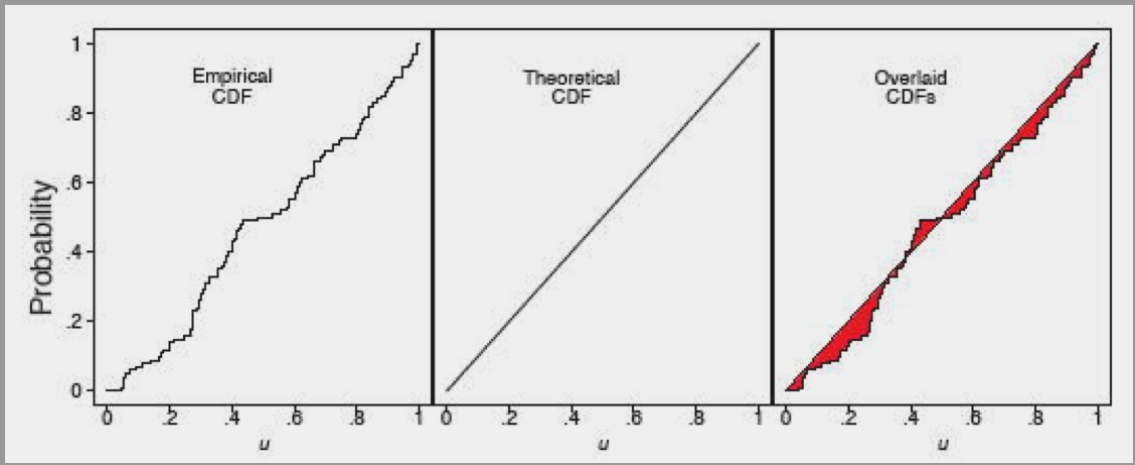
- Using histogram



Quantile-Quantile Plots – Example I

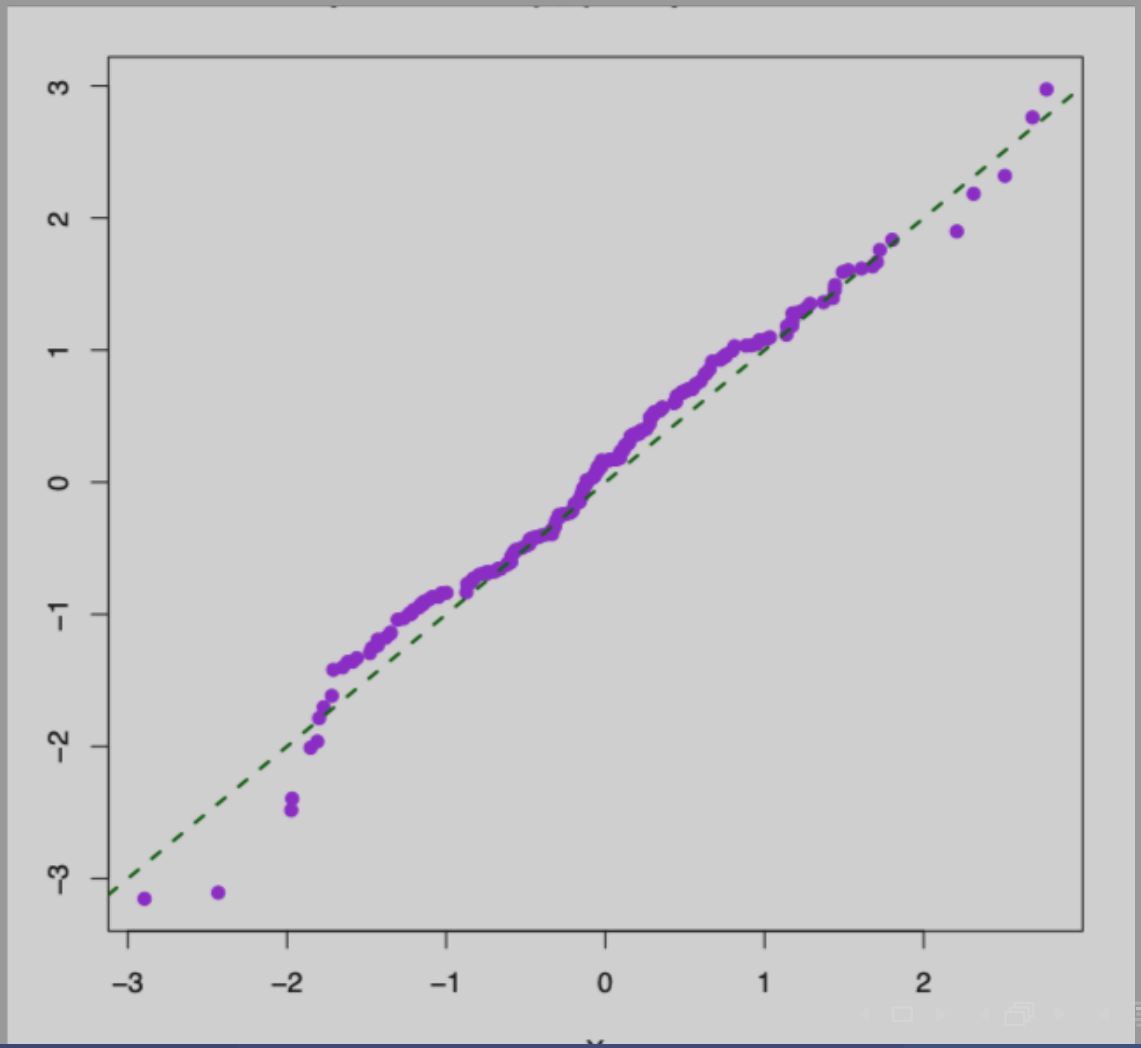
Alternatively, we might use the cumulative distribution function (CDF)

- The CDF gives the probability of the value less than or equal to μ



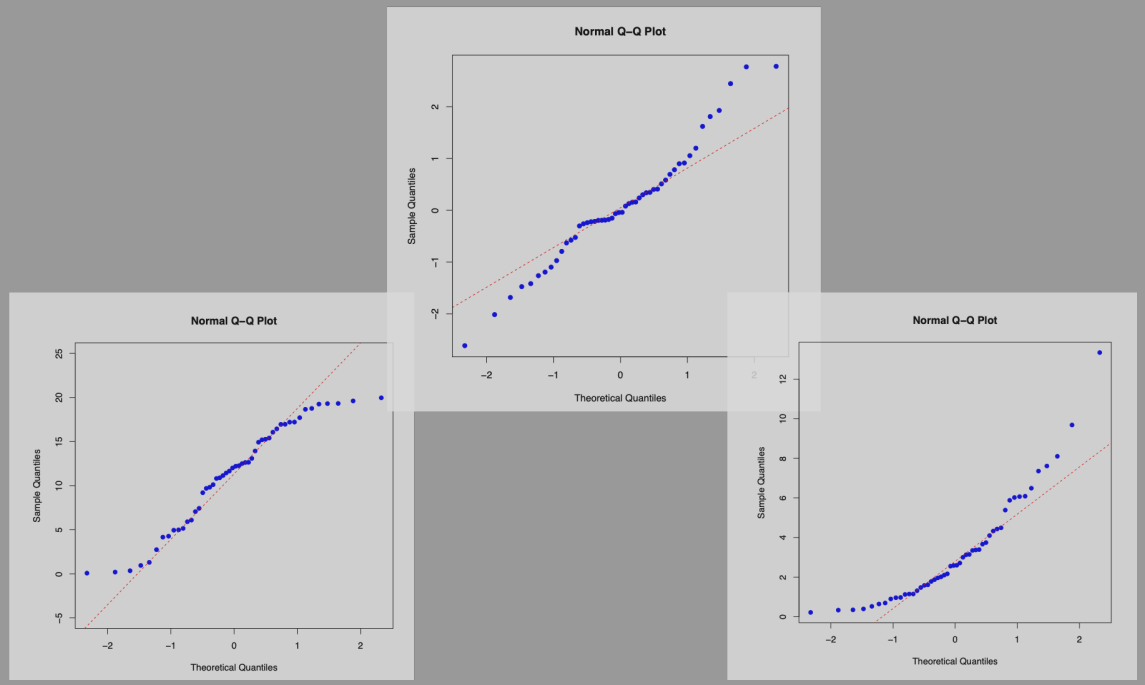
Quantile-Quantile Plots?Example II

- Suppose we have two samples of size n, $X_1, X_2, ...X_n$ and $Y_1, Y_2, ...Y_n$.
- If the two datasets come from the same distribution, the points should lie roughly on a line through the origin with slope 1.



Interpreting

Normal



Working with QQ Plot for Normal Data

Let $\{z_1, z_2, ..., z_n\}$ denote a random sample from a normal distribution with mean $\mu = 0$ and standard deviation $\rho = 1$. Let the ordered values be denoted by

- $z_1 < z_2 < z_3 ... z_{(n-1)} < z_{(n)}$

Data (z)	Rank	Middle of the i_{th} Interval	Normal (z)
-1.96	1		
-.78	2		
0.31	3		
1.15	4		
1.62	5		

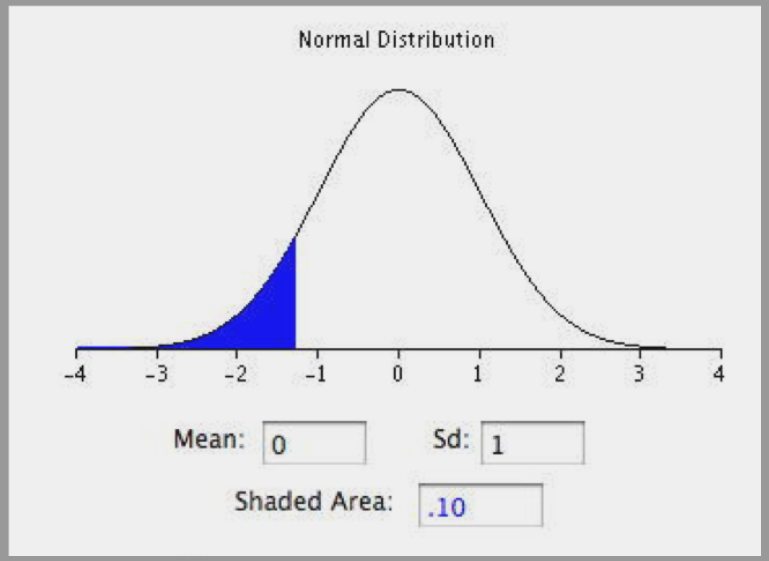
Working with QQ Plot for Normal Data

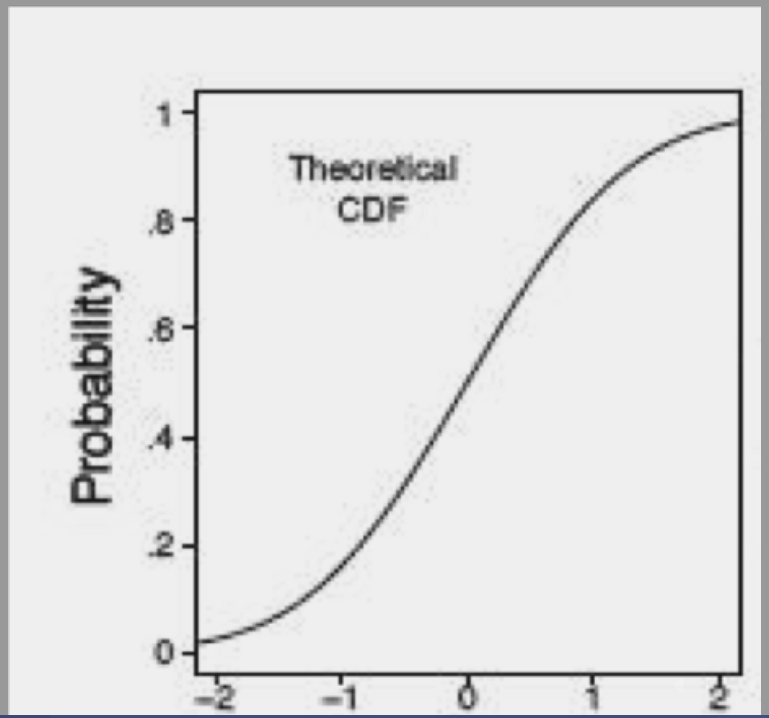
Let $\{z_1, z_2, ..., z_n\}$ denote a random sample from a normal distribution with mean $\mu = 0$ and standard deviation $\rho = 1$. Let the ordered values be denoted by

- $z_1 < z_2 < z_3 ... z_{(n-1)} < z_{(n)}$

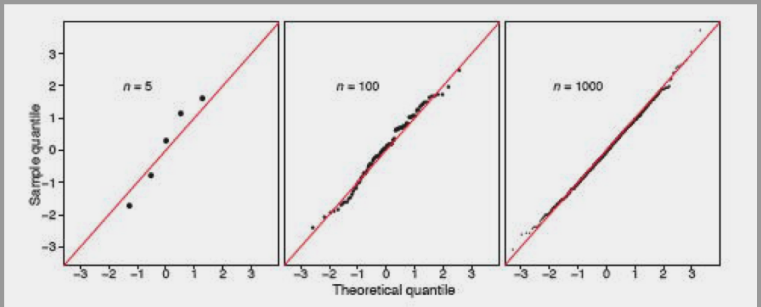
Data (z)	Rank	Middle of the i_{th} Interval	Normal (z)
-1.96	1	1	
-.78	2	3	
0.31	3	5	
1.15	4	7	
1.62	5	9	

Working with QQ Plot for Normal Data





Working with QQ Plot for Normal Data



Question 1

q-q plots are used to investigate

- A. possible non-independence.
- B. whether the sample is from a specified distribution.
- C. the distribution is bimodal.

Which one?

Question II

The advantage of q-q plots over histograms is

- A. it is not necessary to specify a bin width.
- B. lines are easier to interpret than rectangles.

Which one?

Question III

A q-q plot is used to test for normality based on 10 standardised sample values. The first interval is from 0 to 0.1 so the middle of the interval is 0.05. What is Normal (z), the theoretically expected z score for the first sample data point?