

# Introduction to Probability and Statistics

## Literature

Raj Jain: The Art of Computer Systems  
Performance Analysis, John Wiley

Schickinger, Steger: Diskrete Strukturen Band 2, Springer

David Lilja: Measuring Computer Performance: A Practitioner's  
Guide, Cambridge University Press

1

## Goals

- Provide intuitive conceptual background for some standard statistical methods
  - Draw meaningful conclusions in presence of noisy measurements
  - Learn how to apply techniques in new situations
- Don't simply plug and crank from a formula
  
- Present techniques for aggregating large quantities of data
  - Obtain a big-picture view of your results
  - Obtain new insights from complex measurement and simulation results
- E.g., how does a new feature impact the overall system?

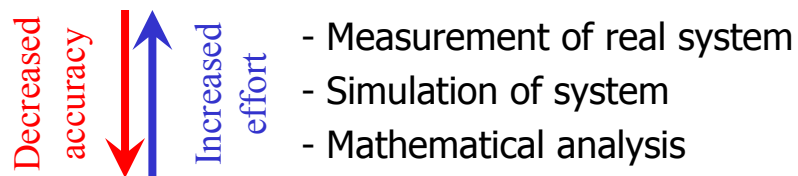
2

## Analytical performance evaluation

- Problem: How to
  - Predict system performance without implementation
  - Evaluate effects of design alternatives
  - Explain unexpected behavior
  
- Performance measures:
  - Waiting time
  - Throughput
  - Number of jobs in system
  - Utilization

3

## Performance evaluation techniques



- Model
  - Abstraction of real system
  - Extraction of essential details (essential for behavior of system)

4

## Basic definitions

- Probability as modeling an experiment
- Set of possible outcomes of experiment:  
sample space  $S$  (the universe)
- E.g.: Classic „experiment“: Tossing a die

$$S = \{1,2,3,4,5,6\}$$

- Any subset  $A$  of  $S$  is an event, e.g.,

$$A = \{the\ outcome\ is\ even\} = \{2,4,6\}$$

5

## Basic operations on events

- For any two events  $A, B$ :

$$\bar{A} = A \text{ complement} = \{\text{all outcomes not in } A\}$$

$$A \cup B = A \text{ union } B = \{\text{all outcomes in } A \text{ or } B \text{ or both}\}$$

$$A \cap B = A \text{ intersect } B = \{\text{all outcomes in both } A \text{ and } B\}$$

$$(AB = A \cap B)$$

- The empty set:  $\emptyset \Rightarrow \bar{\bar{S}} = \emptyset$
- $A$  and  $B$  are mutually exclusive  $\Leftrightarrow AB = \emptyset$

6

## Probability on events

**Probability mass function** P maps each event A into real number P(A) with

- $1 \geq P(A) \geq 0$  for every event  $A \subseteq S$
- $P(S) = 1$
- If A and B mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

- Conditional probability

$$P(A|B) = \frac{P(AB)}{P(B)} \Rightarrow P(AB) = P(B)P(A|B)$$

7

## Basic probability / Statistics

### **Independent events**

- Two events are independent
  - Event 1 occurs with no influence on prob. of event 2
  - Knowing of event 1 has no change in estimate of probability of event 2

$$P(AB) = P(A)P(B)$$

### **Random variable**

- Specified set of values with specified probabilities

8

## Random variable: Example

- Fair coin tossed 3 times (Tail: T, Head: H)
- $S = \{ (TTT), (TTH), (THT), (THT), (HTT), (HTH), (HHT), (HHH) \}$
- Random var  $X$  # of heads tossed (3 tries)
  - $X(TTT) =$                        $X(HTT) =$
  - $X(TTH) =$                        $X(HTH) =$
  - $X(THT) =$                        $X(HHT) =$
  - $X(THT) =$                        $X(HHH) =$
- Probability for variable  $X$ 
  - $P(X = 0) =$                        $P(X = 1) =$
  - $P(X = 2) =$                        $P(X = 3) =$

9

## Random variable as measurement

### Examples of complicated experiments

- A chemical reaction
- A laser emitting photons
- A packet arriving to router

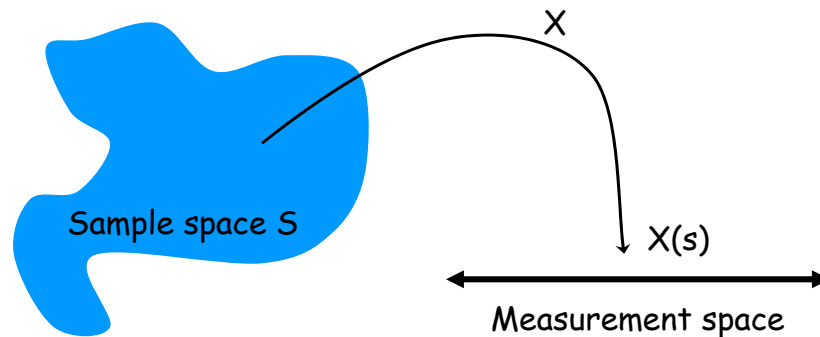
### Problem

- Difficult to exactly describe the sample space
- But we can describe specific measurements
  - Temperature change
  - Number of photons emitted in one millisecond
  - Time of arrival of packet

10

## Random variable as measurement (2)

**Random variable:** Measurement on experiment



11

## Prob. mass func. of a random var.

**Probability mass function (PMF)** of X is:

$$P_X(x) = P(X = x) = P(\{s \in S \mid X(s) = x\})$$

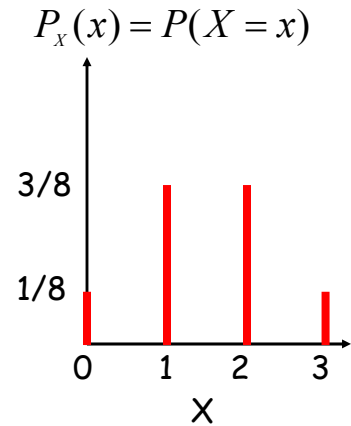
$$1 \geq P_X(x) \geq 0 \quad \text{for } -\infty < x < \infty$$

□ For (discrete-valued) random variable X

$$\sum_{x=-\infty}^{\infty} P_X(x) = 1$$

12

## PMF: 3 coin toss example



13

## Cumulative distribution function

**Cumulative distribution function (CDF)** of X is:

$$F_X(x) = P(X \leq x) = P(\{s \in S \mid X(s) \leq x\})$$

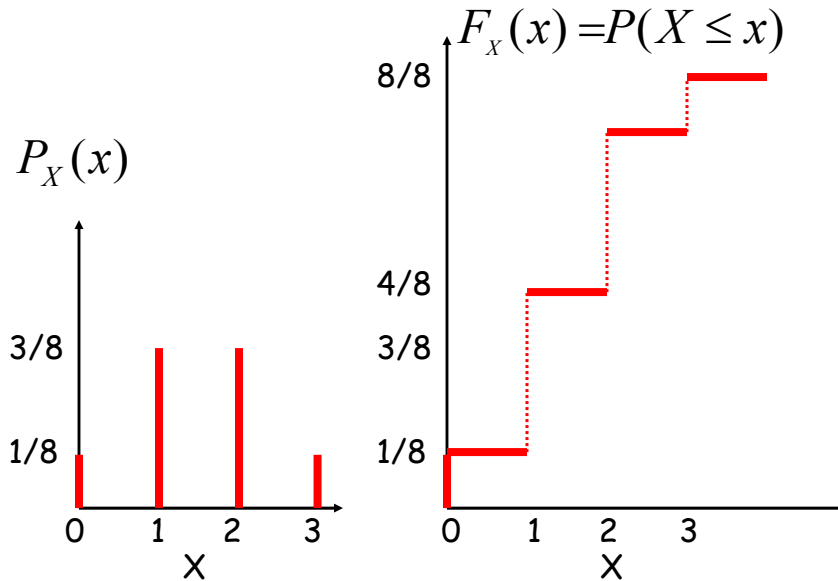
□ Note that  $F_X(x)$  is non-decreasing in  $x$ , i.e.,

$$x_1 \leq x_2 \quad \Rightarrow \quad F_X(x_1) \leq F_X(x_2)$$

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} F_X(x) = 1$$

14

## PMF, CDF: 3 coin toss example



15

## Expectation of a random variable

Expectation (average) of a random variable  $X$ :

$$\bar{X} = E(X) = \sum_{x=-\infty}^{\infty} x P(X = x) = \sum_{x=-\infty}^{\infty} x P_X(x)$$

- The expected value is also called the first moment
- Three coins example:

$$E(X) = \sum_{x=0}^3 x P_X(x) = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8} = 1.5$$

16



## Quantile

**$\alpha$ -quantile:**  $x_\alpha$  value where CDF takes a value  $\alpha$

$$F_X(x_\alpha) = P(X \leq x_\alpha) = \alpha$$

**Median:** 50-percentile

informal: one half of the values are smaller than  $X$   
one half of the values are larger than  $X$

17

## Statistics: Why do we need it?

**1. Aggregate data into meaningful information.**

445 446 397 226  
388 3445 188 1002  
47762 432 54 12  
98 345 2245 8839  
77492 472 565 999  
1 34 882 545 4022  
827 572 597 364

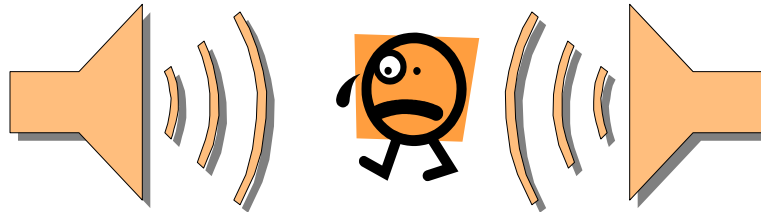


$$\bar{x} = \dots$$

18

## Statistics: Why do we need it? (2.)

### 2. Noise, noise, noise, noise, noise!



OK – not really this type of noise

19

## What is a statistic?

- "A quantity that is computed from a sample [of data]."

Merriam-Webster

→ A single number used to summarize a larger collection of values

## What are statistics?

- "A branch of mathematics dealing with the collection, **analysis, interpretation,** and **presentation** of masses of numerical data."

Merriam-Webster

→ We are most interested in analysis and interpretation here

- "Lies, damn lies, and statistics!"

20

## The simplest statistic: a mean?

- ❑ Reduce performance to a single number
- ❑ But what do these means mean?
- ❑ Indices of central tendency
  - Sample mean
  - Sample median
  - Sample mode
- ❑ Other means
  - Arithmetic
  - Harmonic
  - Geometric
- ❑ Quantifying variability

21

## The problem with means

- ❑ Performance is multidimensional
  - CPU or I/O time
  - Network time
  - Interactions of various components
  - ...
- ❑ Systems are often specialized
  - Performs great on application type X
  - Performs lousy on anything else
- ❑ Potentially a wide range of execution times on one system using different benchmark programs

22

## The problem with means (2)

- ❑ Nevertheless, people still want a single number answer!
- ❑ *How to (correctly) summarize a wide range of measurements with a single value?*

23

## Index of central tendency

- ❑ Tries to capture “center” of a distribution of values
- ❑ Use this “center” to summarize overall behavior
- ❑ You will be pressured to provide “mean” value
  - Understand how to choose the best type for the circumstance
  - Be able to detect bad results from others
- ❑ Examples
  - Sample mean: “Average” value
  - Sample median:  $\frac{1}{2}$  of the values are above,  $\frac{1}{2}$  below
  - Sample mode: Most common value

24

## Indices of central tendency (2.)

- "Sample" implies
  - Values are measured from a discrete random variable X
- Value computed is only an approximation of true mean value of underlying process
- True mean value cannot actually be known
  - Would require infinite number of measurements

25

## Sample mean

- Expected value of X =  $E[X]$ 
  - First moment of X
  - $x_i$  = values measured ( $i \in \{1, \dots, n\}$ )
  - $p_i$  =  $P(X = x_i) = P(\text{we measure } x_i)$

$$E[X] = \sum_{i=1}^n x_i p_i$$

26

## Sample mean (2)

- Without additional information, assume
  - $p_i = \text{constant} = 1/n$  (Laplace principle)
  - $n = \text{number of measurements}$
- **Arithmetic mean**
  - Common "average"

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

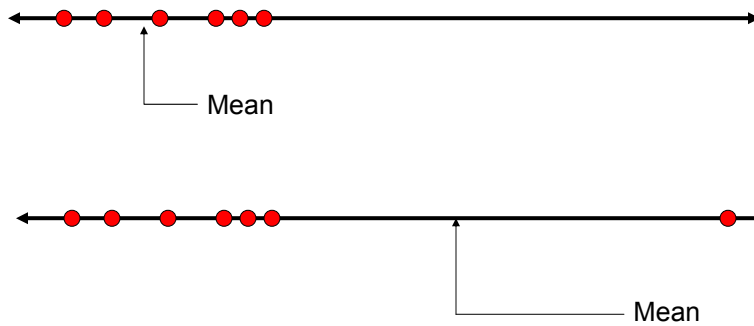
27

## Potential problem with means

- Sample mean gives equal weight to all measurements
- **Outliers** can have a large influence on the computed mean value
- Distorts our intuition about the **central tendency** of the measured values

28

## Potential problem with means (2.)



29

## Median

- Index of central tendency with
  - $\frac{1}{2}$  of the values larger,  $\frac{1}{2}$  smaller
  - Algorithm
    - Sort  $n$  measurements
    - If  $n$  is odd
      - Median = middle value
      - Else, median = mean of two middle values
- Reduces skewing effect of outliers

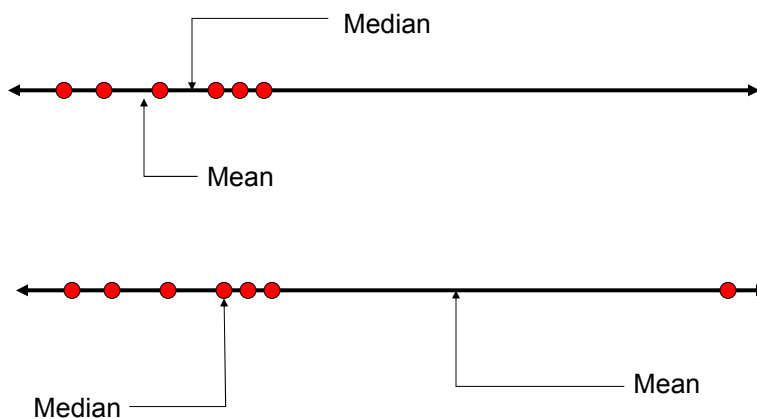
30

## Example

- Measured values: 10, 20, 15, 18, 16
  - Mean = 15.8
  - Median = 16
- Obtain one more measurement: 200
  - Mean = 46.5
  - Median =  $\frac{1}{2} (16 + 18) = 17$
- Median gives more intuitive sense of central tendency

31

## Potential problem with means (3.)



32



## Mode

- ❑ Value that occurs most often
- ❑ May not exist
- ❑ May not be unique == multiple modes
  - E.g., “bi-modal” distribution
    - Two values occur with same frequency

33

## Mean, median, or mode?

- ❑ Mean
  - If the sum of all values is meaningful
  - Incorporates all available information
- ❑ Median
  - Intuitive sense of central tendency with outliers
  - What is “typical” of a set of values?
- ❑ Mode
  - When data can be grouped into distinct types, categories (categorical data)

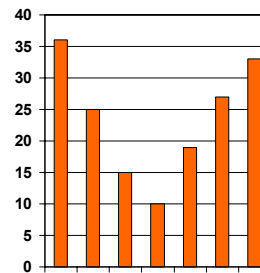
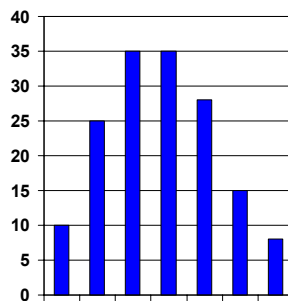
34

## Quantifying variability

- How “spread out” are the values?
  - How much spread relative to the mean?
  - What is the shape of the distribution of values?
- => A mean hides information about **variability!**

35

## Histograms



- Similar mean values
- Widely different distributions
- How to capture this variability in one number?

36

## Index of dispersion

Quantifies how “spread out” measurements are

- Range
  - (max value) – (min value)
- 10- and 90- percentiles
- Maximum distance from the mean
  - Max of  $|x_i - \text{mean}|$
- Neither efficiently incorporates all available information

37

## Sample variance

$$\begin{aligned}\text{var} = s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \\ &= \frac{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}\end{aligned}$$

- **Variance:** *second moment of random variable  $X$* 
  - Second form good for calculating “on-the-fly”
    - One pass through data
- Gives “units-squared”
  - Hard to compare to mean
- **Standard deviation:**  $s$ 
  - $s$  = square root of variance
  - Units = same as mean

38

## Coefficient of variation (COV)

- Dimensionless
- Compares relative size of variation to mean value

$$COV = \frac{s}{\bar{x}}$$

39

## How to determine the distribution of data?

- Plot a histogram
  - Count of observations within a cell or bucket
- Compare to known distributions

40

## Random variables: Bernoulli

- Simplest possible measurement on experiment

- Success (X = 1)
- Failure (X = 0)

- Notation:

$$P_X(1) = P(X = 1) = p \qquad P_X(0) = P(X = 0) = 1 - p$$

41

## Random variables: Binomial

X = # of success in n independent Bernoulli trials

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- Mean:  $np$

- Variance:  $np(1 - p)$

42

## Random variables: Geometric

X = # of independent Bernoulli trials until success

$$P(X = x) = (1 - p)^{x-1} p$$

□ Mean:  $1/p$

□ Variance:  $(1 - p)/p^2$

43

## Random variables: Poisson

Limiting form of binomial distribution

$$P(X = x) = \lambda^x \frac{e^{-\lambda}}{x!}$$

□ Mean:  $\lambda$

□ Variance:  $\lambda$

Models arrivals from large numbers of independent sources

44

## Continuous-valued random variables

**Discrete random variables:**  $X(s)$  is integer

□ Examples

- # of arrivals in one second
- # of attempts until success

**Continuous random variables:**

$X(s)$  ranges from  $-\infty$  to  $\infty$  as  $s$  varies

□ Examples

- Time of arrival event
- Time between arrivals

45

## Continuous-valued random variables 2

□ CDF  $F_X(x)$ : continuous slopes (some)

□ **Probability density function (PDF):**

$$f_X(x) = F'_X(x) = \frac{dF_X(x)}{dx}$$

□ Since the CDF is non-decreasing

$$f_X(x) \geq 0 \quad \text{for all } x$$

(may get larger than 1)

46

## Random variables: Exponential

Used to represent time, e.g., until next arrival

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } 0 \leq x$$

□ Mean:  $1/\lambda$

□ Variance:  $1/\lambda^2$

47

## Random variables: Exponential

CDF

$$F_X(x) = \int_0^x f_X(\bar{x}) d\bar{x} = \int_0^x \lambda e^{-\lambda \bar{x}} d\bar{x} = 1 - e^{-\lambda x}$$

Complementary cumulative distribution function  
(CCDF)

$$F_X^c(x) = 1 - F_X(x) = e^{-\lambda x}$$

48



## Exponential: Memoryless property

### Memoryless:

„the future is independent of the past“

„remembering the time since the last event does not help predicting the time till the next event“

### □ Mathematical:

$$P(X > s + t | X > t) = P(X > s) \quad \text{for } s, t > 0$$

49

## Proof: Exp. dist. is memoryless

$$\begin{aligned} P(X > s + t | X > t) &= \frac{P(X > s + t, X > t)}{P(X > t)} \\ &= \frac{P(X > s + t)}{P(X > t)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \\ &= P(X > s) \end{aligned}$$

50

## Exponential and Poisson dist.

- $T_1, T_2, \dots$  sequence of independent random variables with exponential probability density func.

$$p_{T_i}(t) = \sigma^{-1} e^{-t/\sigma}$$

- Consider the random variable N

$$T_1 + T_2 + \dots + T_N \leq \tau \leq T_1 + T_2 + \dots + T_{N+1}$$

N has Poisson dist. with expected value:

$$\tau / \sigma$$

51

## Normal (Gauss) distribution

$$\text{PDF: } f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } -\infty \leq x \leq \infty$$

Mean:  $\mu$

Variance :  $\sigma^2$

Sum of n independent normal variables is normal

**Central limit theorem:** Sum of a large number of independent observations from any distribution tends to a normal distribution

52

## Pareto distribution

PDF:  $f_X(x) = \frac{ak^a}{x^{(a+1)}} \quad \text{for } 0 < k \leq x$

CDF:  $F_X(x) = 1 - \left(\frac{k}{x}\right)^a \quad \text{for } 0 < k \leq x$

Mean:  $\frac{a}{a-1} \quad a > 1$

Variance :  $\frac{a}{(a-1)^2(a-2)} \quad a > 2$

**Heavy tail:** Tail of probability distribution decays like a power:  
power-law distribution

- More small events, more large events

53

## Determine the distribution of data?

- Plot a histogram
  - Count of observations within a cell or bucket
- Problem
  - How to determine cell size?
    - Small cells => large variations in # of obs per cell
    - Large cells => details are lost
    - Guideline: if any cell has less than five obs. increase cell size or use variable cell histogram
  - How to determine cell spacing?
    - Linear
    - Logarithmic

54

## Determine the distribution of data(2)?

- Plot a scatter plot
  - For each value: X vs. Y
- Problem
  - Too many points on top of each other ?
    - Large dots => hard to distinguish points
    - Small dots => hard to see outliers
  - Use two-dimensional histograms
  - Use densities
  - Which scale?
    - Linear
    - Logarithmic

55

## Determine the distribution of data(3)?

- Plot an empirical CDF
  - Concentrate 1/n probability at each of the n numbers in a sample

$$F_n(x) = 1/n \sum_{i=1}^n I(X_i \leq x)$$

- Problem
  - Tail of interest => plot CCDF

56

## Determine the distribution of data(4)?

- Plot a density
  - Smoothed normalized counts of observations
- Problem
  - How to determine cell size?
  - How to do the smoothing
  - How to determine cell spacing?
    - Linear
    - Logarithmic

57

## Sources of Experimental Errors Accuracy, precision, resolution



58

## Experimental errors

- ❑ Errors → noise in measured values
  - ❑ **Systematic** errors
    - Result of an experimental “mistake”
    - Typically produce constant or slowly varying bias
- Controlled through skill of experimenter
- Examples
    - Temperature change causes clock drift
    - Forget to clear cache before timing run

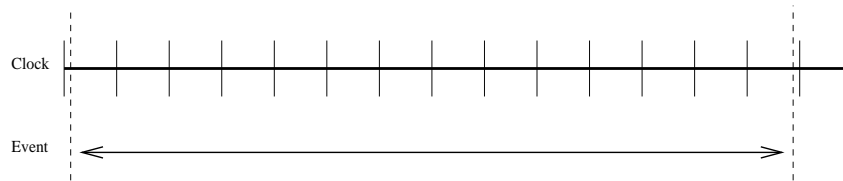
59

## Experimental errors

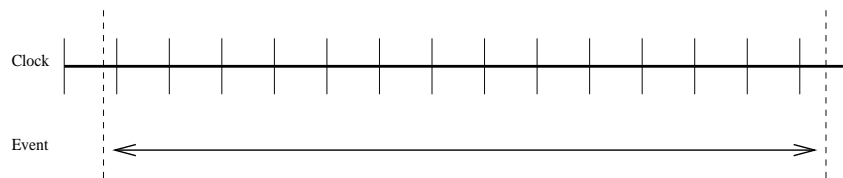
- ❑ **Random** errors
  - Unpredictable, non-deterministic
  - Unbiased → equal probability of increasing or decreasing measured value
- ❑ Result of
  - Limitations of measuring tool
  - Observer reading output of tool
  - Random processes within system
- ❑ Typically cannot be controlled
  - Use statistical tools to characterize and quantify

60

## Example: Quantization → Random error



(a) Interval timer reports event duration of  $n = 13$  clock ticks.



(b) Interval timer reports event duration of  $n = 14$  clock ticks.

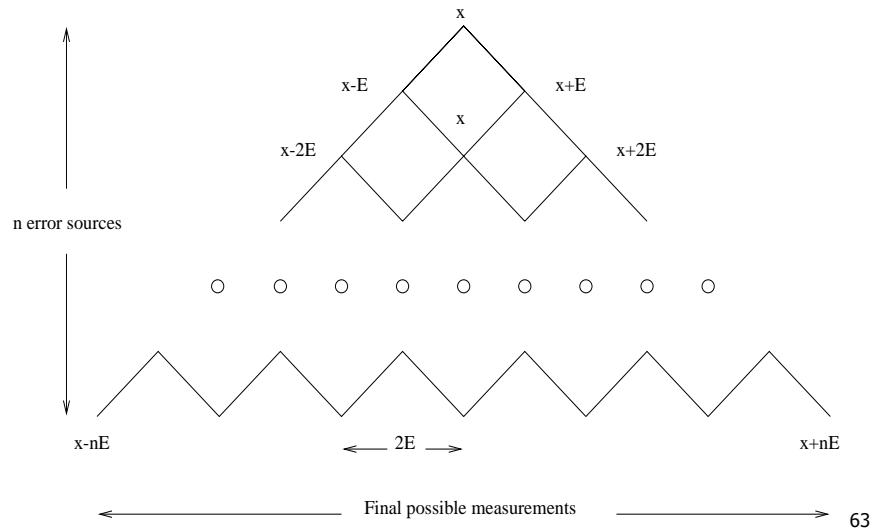
61

## Quantization error

- Timer resolution  
→ quantization error
- Repeated measurements  
 $X \pm \Delta$   
Completely unpredictable

62

## Probability of obtaining a specific measured value

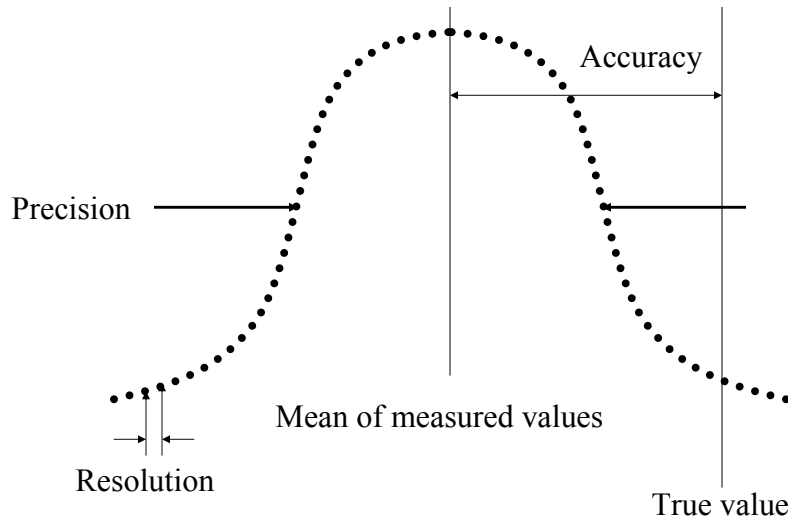


## A model of errors

- $P(X=x_j) = P(\text{to measure } x_j)$   
corresponds to the "number of possible paths"
- $P(X=x_j) \sim$  binomial distribution
- As number of error sources becomes large
  - $n \rightarrow \infty$ ,
  - Binomial  $\rightarrow$  Gaussian (Normal)
- Thus, the **bell curve**



## Frequency of measuring specific value



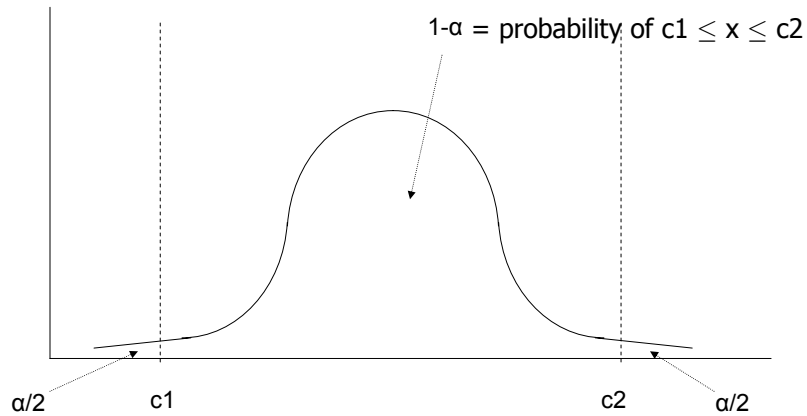
65

## Accuracy, precision, resolution

- Systematic errors → **accuracy**
  - How close mean of measured values is to true value
  - Hard to determine true accuracy
  - Relative to a predefined standard
    - E.g. definition of a "second"
- Random errors → **precision**
  - Repeatability of measurements
  - Dependent on tools
- Characteristics of tools → **resolution**
  - Smallest increment between measured values
  - Quantify amount of *imprecision* using statistical tools

66

## Confidence interval for the mean



67

## Normalize $x$

$$z = \frac{\bar{x} - x}{s/\sqrt{n}}$$

$n = \text{number of measurements}$

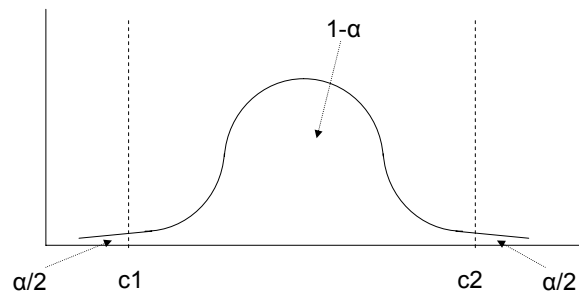
$$\bar{x} = \text{mean} = \sum_{i=1}^n x_i$$

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

68

## Confidence interval for the mean (2)

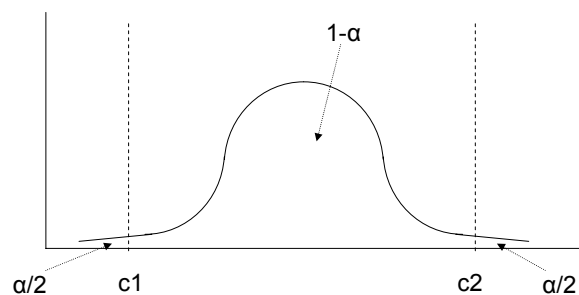
- Normalized  $z$  follows the Student's  $t$  distribution
  - $(n-1)$  degrees of freedom
  - Area left of  $c_2 = 1 - \alpha/2$
  - Tabulated values for  $t$



69

## Confidence interval for the mean (2)

- As  $n \rightarrow \infty$ , normalized distribution becomes Gaussian (normal)



70

## Confidence interval for the mean (4)

$$c_1 = \bar{x} - t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

$$c_2 = \bar{x} + t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

Then,

$$\Pr(c_1 \leq x \leq c_2) = 1 - \alpha$$

□ t-distribution:  
Values available via  
standard tables

71

## An example

Experiment	Measured value
1	8.0 s
2	7.0 s
3	5.0 s
4	9.0 s
5	9.5 s
6	11.3 s
7	5.2 s
8	8.5 s

72

## An example (2)

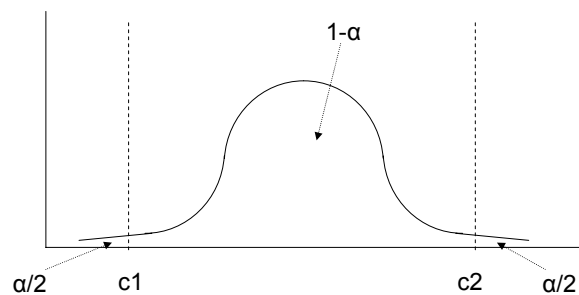
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 7.94$$

$s$  = sample standard deviation = 2.14

73

## An example (3.)

- 90% CI → 90% chance that the measured value is in the interval
- 90% CI →  $\alpha = 0.10$



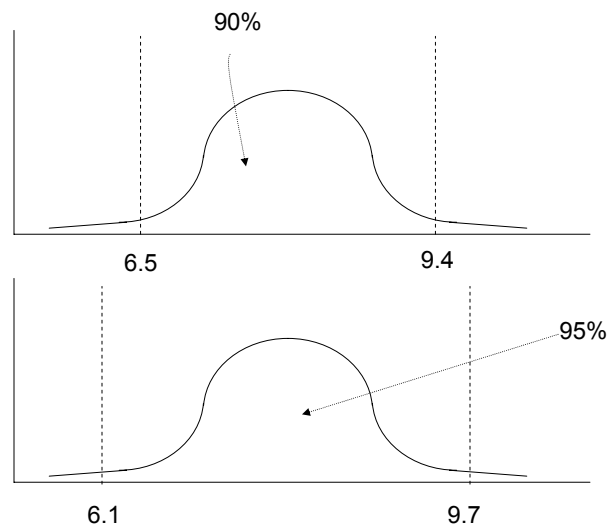
74

## An example (4.)

- 90% CI = [6.5, 9.4]
  - 90% chance value is between 6.5, 9.4
- 95% CI = [6.1, 9.7]
  - 95% chance value is between 6.1, 9.7
- Why is interval wider when we are more confident?

75

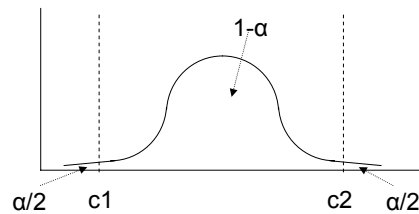
## Higher confidence → Wider interval?



76

## Key assumption

- ❑ Measurement errors are Normally distributed.
- ❑ Is this true for most measurements on real systems?



77

## Key assumption (2)

- ❑ Saved by the **Central Limit Theorem**  
*Sum of a "large number" of values from any distribution will be Normally (Gaussian) distributed.*
- ❑ What is a "large number?"
  - Typically assumed to be  $\gg 6$  or 7
  - But in our case often millions or billions

78

## How many measurements?

- Width of interval inversely proportional to  $\sqrt{n}$
- Want to minimize number of measurements
- Find confidence interval for mean, such that:
  - $P(\text{actual mean in interval}) = (1 - \alpha)$

79

## How many measurements (2)?

- But  $n$  depends on knowing mean and standard deviation!
- Estimate  $s$  with small number of measurements
- Use this  $s$  to find  $n$  needed for desired interval width

80