

# Introduction to 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

# 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

# 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



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

# 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!”

# The simplest statistic: a mean?

- ❑ Reduces dataset to a single number
- ❑ But what does this mean mean?
- ❑ Indices of central tendency
  - Sample mean
  - Sample median
  - Sample mode
- ❑ Other means
  - Arithmetic
  - Harmonic
  - Geometric
- ❑ Quantifying dispersion

# The problem with means

- ❑ Performance is multidimensional
  - CPU or I/O time
  - Network delay
  - 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

## 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?*

# 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



# 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

# Sample mean

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

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

## Sample mean (2)

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

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

# 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

# Potential problem with means (2.)



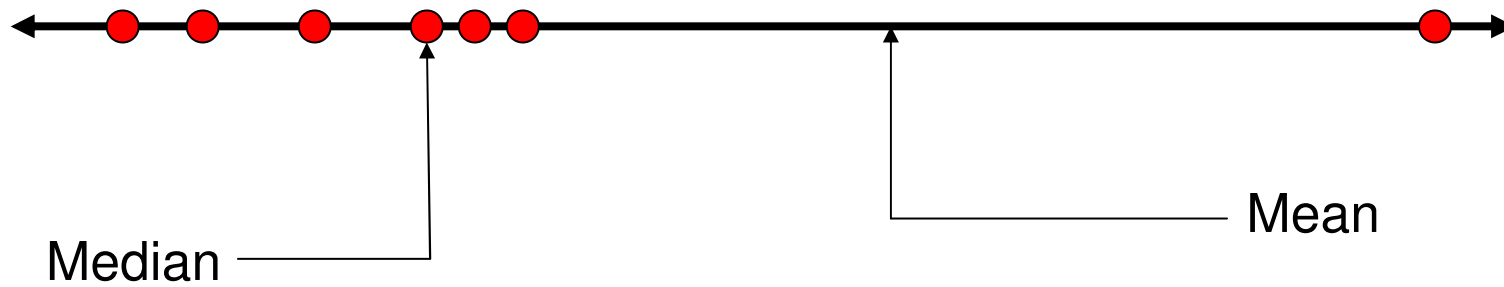
# Median

- Index of central tendency with
  - 1/2 of the values larger, 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

# 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

# Potential problem with means (3.)





# 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

# 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?

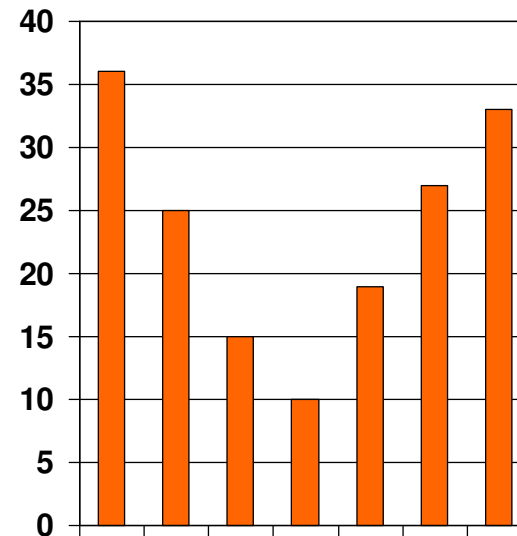
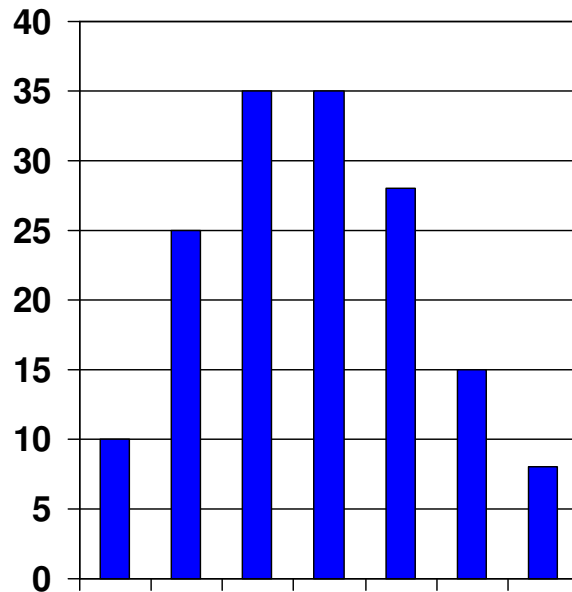
## □ Mode

- When data can be grouped into distinct types, categories (categorical data)

# Quantifying dispersion

- 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!**

# Histograms



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

# 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

# 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

# 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

# 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



# 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

# Sources of Experimental Errors

Accuracy, precision, resolution



# 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

# 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

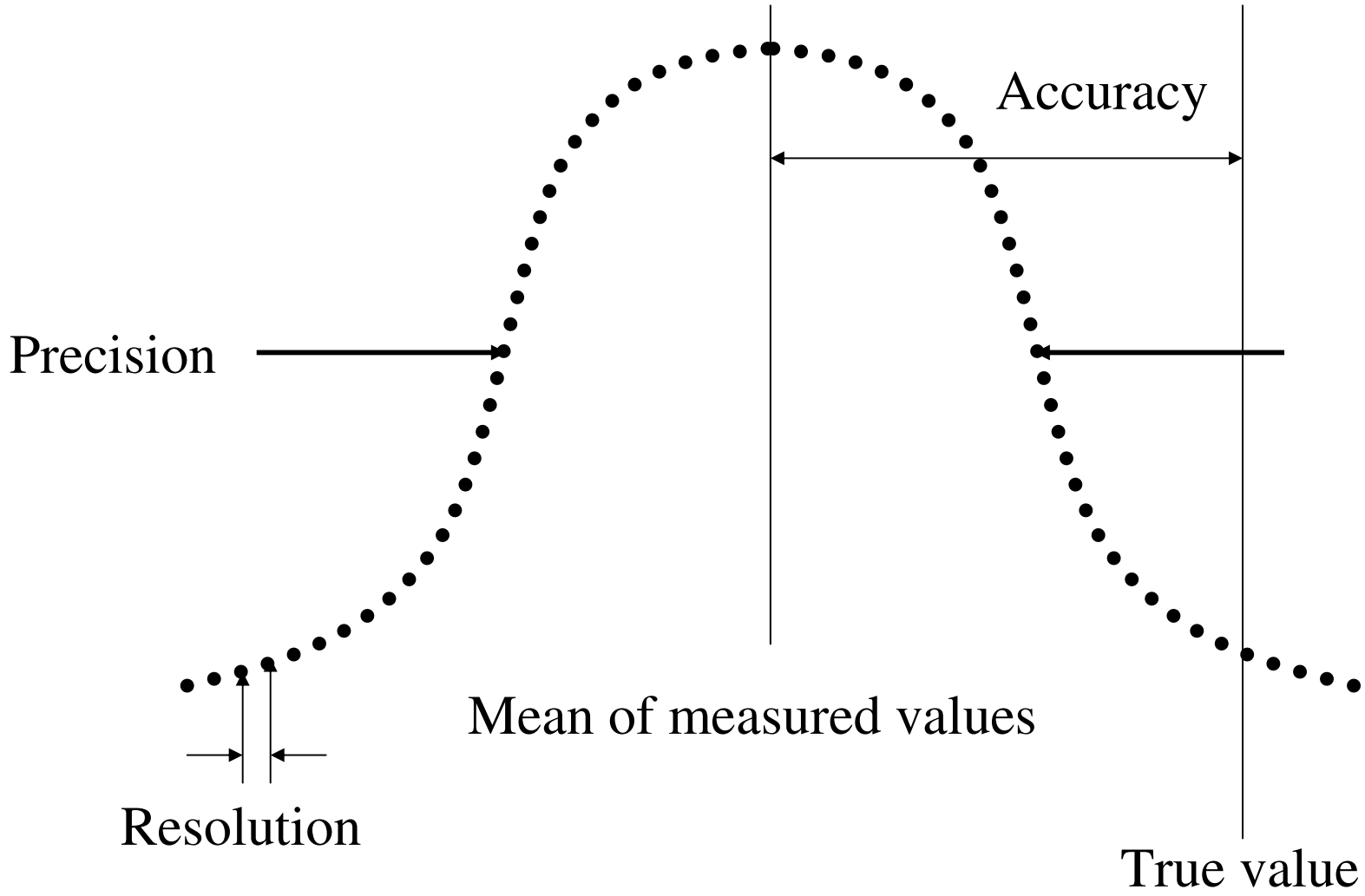
## □ Typically cannot be controlled

- Use statistical tools to characterize and quantify

# A model of errors

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

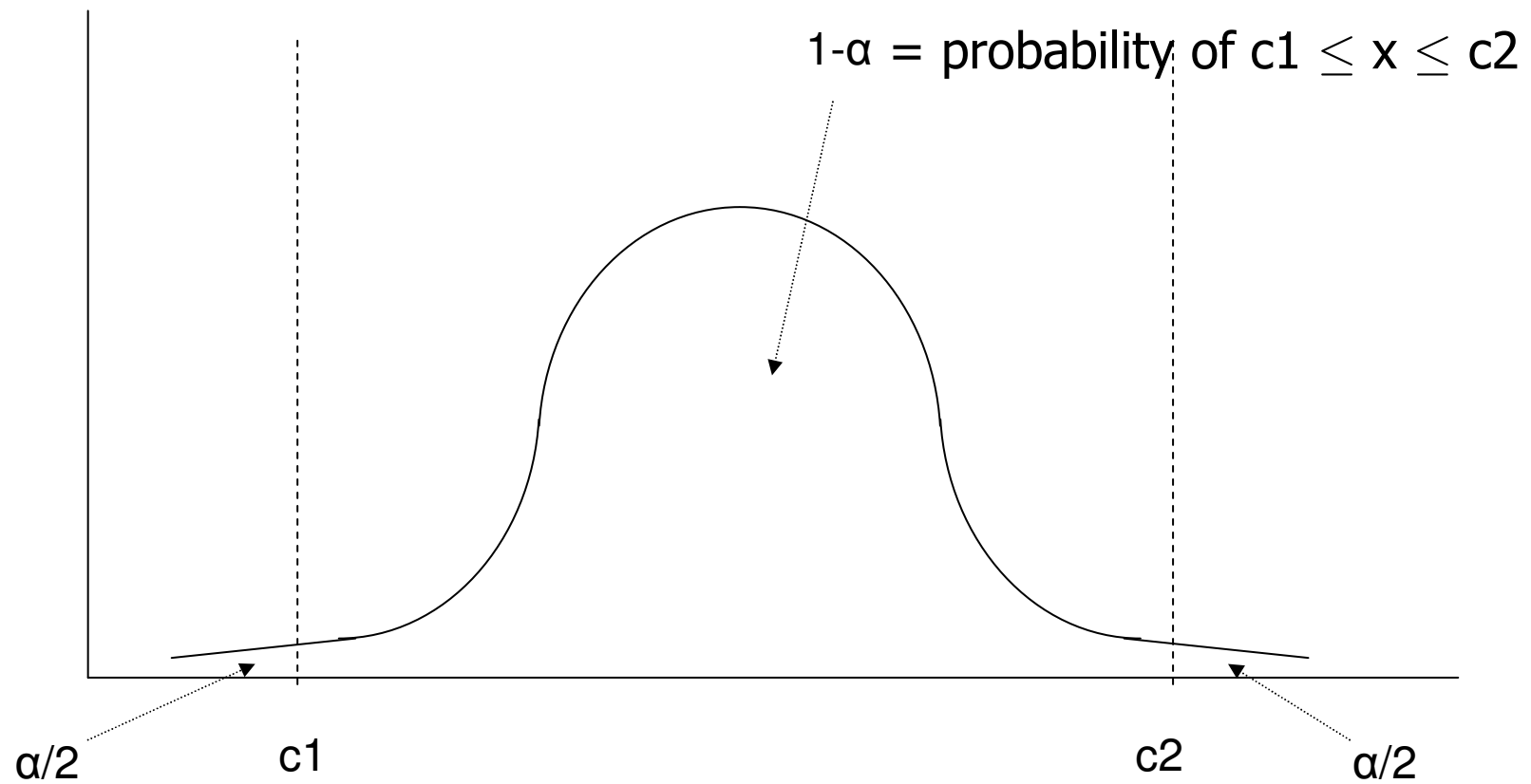
# Frequency of measuring specific value



# 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

# Confidence interval for the mean





# Normalize $x$

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

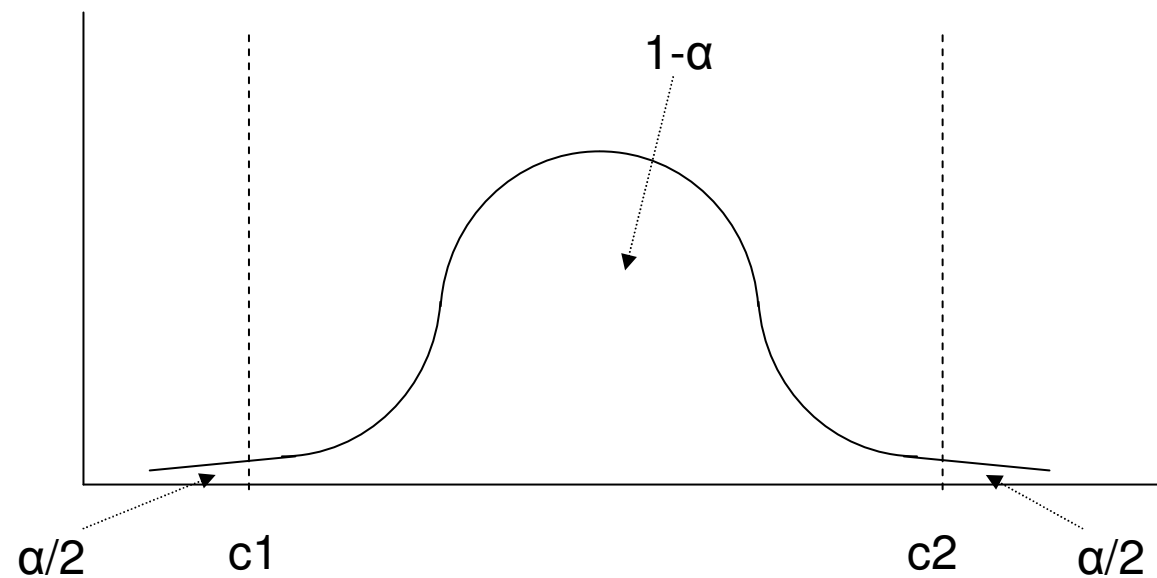
$n$  = number of measurements

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

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

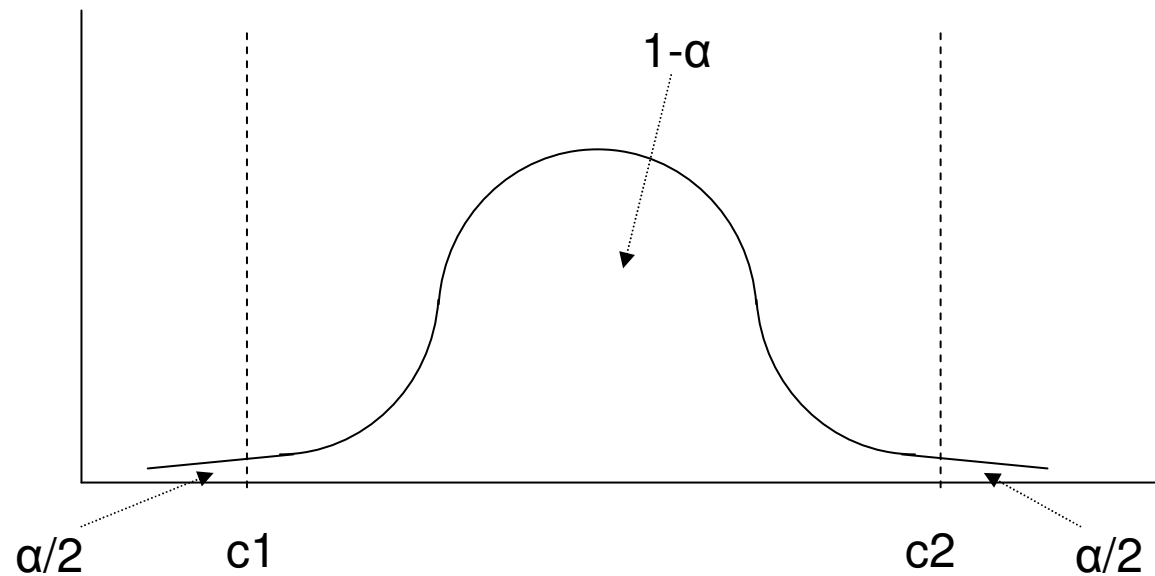
# Confidence interval for the mean (2)

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



# Confidence interval for the mean (2)

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



# An example

<b>Experiment</b>	<b>Measured value</b>
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

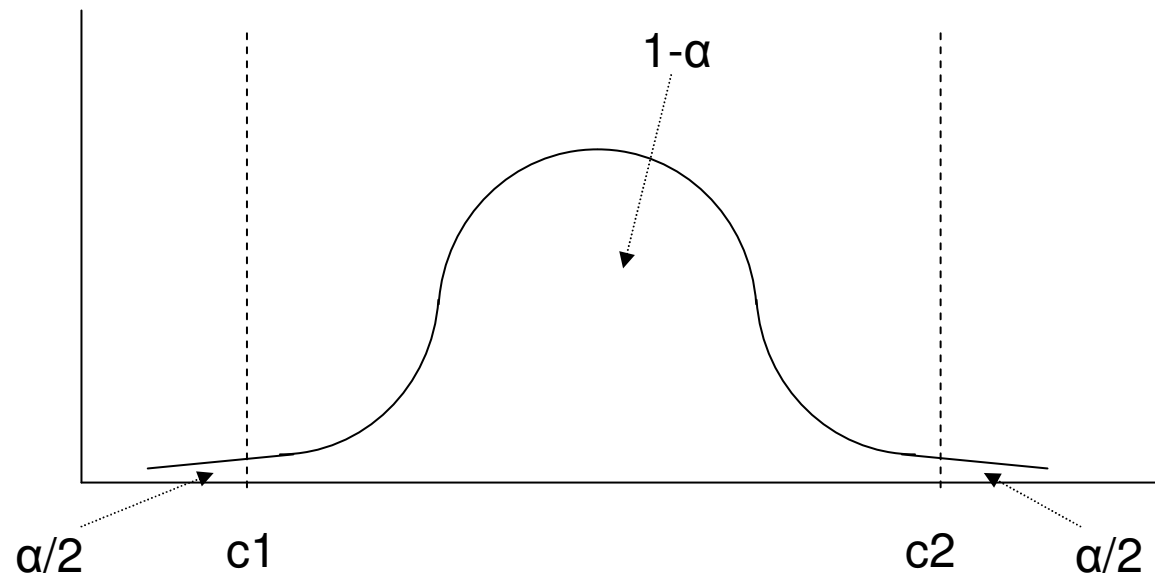
## An example (2)

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

$s$  = sample standard deviation = 2.14

## An example (3)

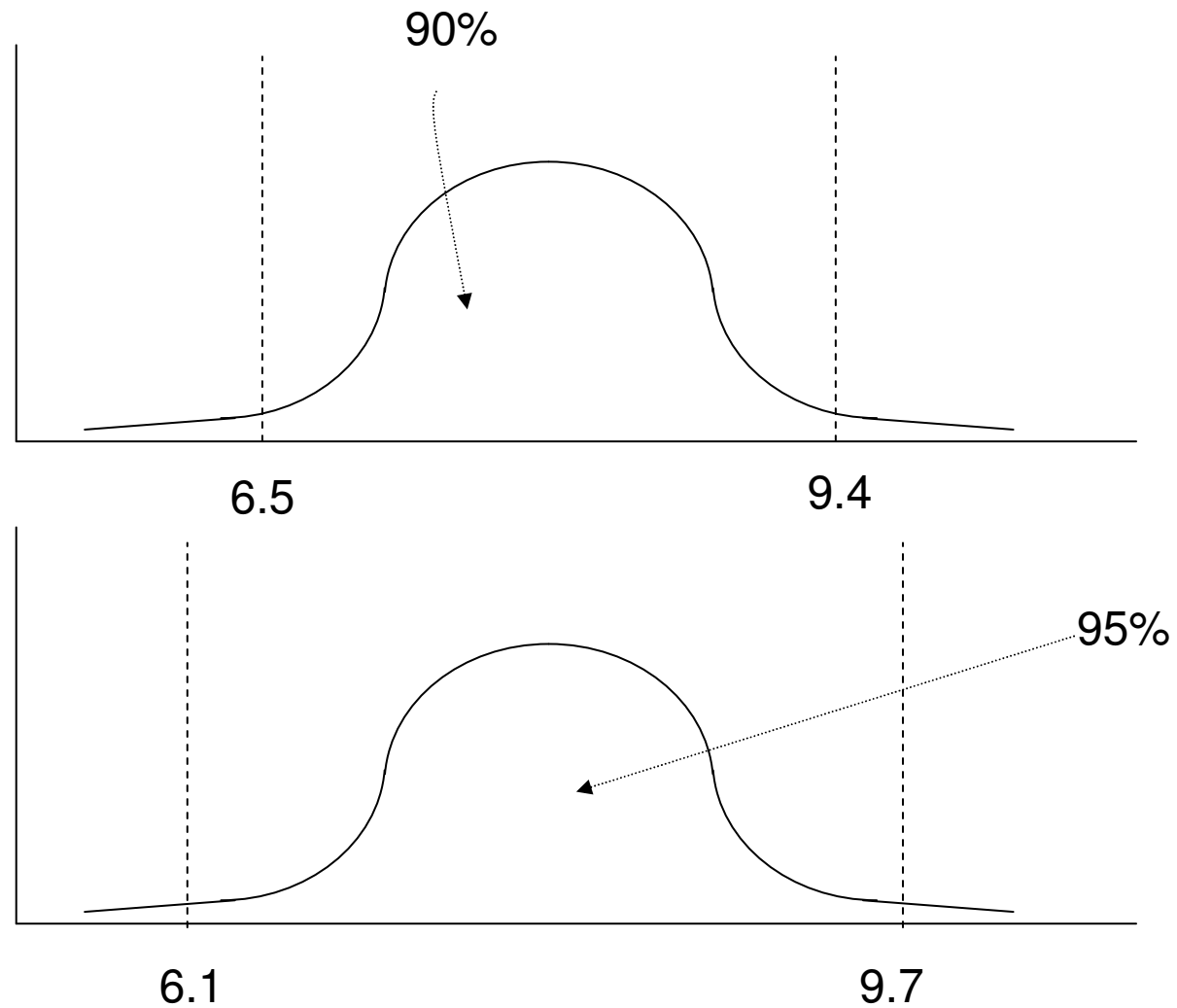
- 90% CI  $\rightarrow$  90% chance that the measured value is in the interval
- 90% CI  $\rightarrow \alpha = 0.10$



## 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?

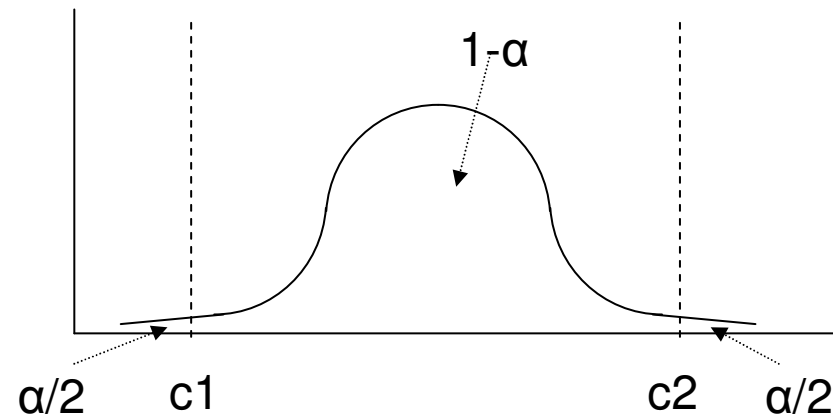
# Higher confidence → Wider interval?





# Key assumption

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



## 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  $>\approx 6$  or  $7$
  - But in our case often millions or billions

# How many measurements?

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

## 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