Advanced Systems Lab
Tutorial III
Statistics and Analysis

G. Alonso
Systems Group

http://www.systems.ethz.ch
Reading assignment

• Read chapters 10, 11, 12, and 13
• Read chapters 17 to 22
Basic statistics

• Not a course on statistics
  – You have done that already
  – We assume familiarity with the basics

• Focus on experimental aspects
  – What and when to measure
  – Side effects and different performance patterns
  – Data distributions
  – Sampling
  – Mean, Average, Outliers, deviation, plotting
  – Confidence intervals
Accuracy vs. Precision

Accuracy = how close to the real value (often unknown)
Precision = similarity of the results of repeated experiments

http://en.wikipedia.org/wiki/Accuracy_and_precision
What and when to measure

• Decide on the parameters to measure:
  – Throughput, response time, latency, etc.

• Design your experiment
  – Configuration, data, load generators, instrumentation, hypothesis

• Run the experiment and start measuring:
  – When to measure (life cycle of an experiment)
  – What to measure (sampling)
Life cycle of an experiment:

- **Warm up**
- **Steady state**
- **Cool down**

Measure only here
Warm up phase

• Warm up phase
  – Time until clients are all up, caches full (warm), data in main memory, etc.
  – Throughput lower than steady state throughput
  – Response time better than in steady state

• Detect by watching measured parameter changing with time

• Measure only in steady state
Cool down phase

• Cool down phase
  – Clients start finishing, resulting in less load in the system
  – Throughput is lower than in steady state
  – Response time better than in steady state

• Detect by observing when measured parameter suddenly changes behavior

• Stop measuring when clients no longer generate a steady load
Patterns to watch for - glitches

QUESTIONS TO ASK:
- IS THIS A BUG OR A FEATURE?
- IS IT MY SYSTEM OR AN INTERFERENCE?
- SHOULD BE INCLUDED IN MEASUREMENTS OR EXCLUDED?

ASSUME STEADY STATE MEASUREMENTS
Patterns to watch for - trends

QUESTIONS TO ASK:
• IS THE PARAMETER SUPPOSED TO GROW IN TIME?
• IS IT MY SYSTEM OR AN INTERFERENCE?
• SHOULD BE COMPENSATED IN THE RESULTS?

ASSUME STEADY STATE MEASUREMENTS
Patterns to watch for - periodic

QUESTIONs TO ASK:
• WHERE DOES THE PERIOD COME FROM?
• IS IT MY SYSTEM OR THE LOAD GENERATORS?
• I AM SEEING EVERYTHING?

ASSUME STEADY STATE MEASUREMENTS
Why are these pattern relevant?

• Too few measurements and too short experiments are meaningless
  – May not capture system behavior
  – May not show pathological behavior
  – May not reflect real values

• Statistics are a way to address some of these issues by providing more information from the data and a better idea of the system behavior
  – but applying statistics to the wrong data will not help!
Data distributions
What are we measuring?

• When measuring, we are trying to estimate the value of a given parameter
• The value of the parameter is often determined by a complex combination of many effects and is typically not a constant
• Thus, the parameter we are trying to measure can be seen as a RANDOM VARIABLE
• The assumption is that this random variable has a NORMAL (GAUSSIAN) DISTRIBUTION
Central limit theorem

• Let $X_1, X_2, X_3, \ldots X_n$ be a sequence of independently and identically distributed random variables with finite values of
  – Expectation ($\mu$)
  – Variance ($\sigma^2$)

as the sample size $n$ increases, the distribution of the sample average of the $n$ random variables approaches the normal distribution with a mean $\mu$ and variance $\sigma^2/n$ regardless of the shape of the original distribution.
How does it work?

Normal or Gaussian distribution

http://en.wikipedia.org/wiki/Normal_distribution
Meaning of standard deviation

• The standard deviation gives an idea of how close are the measurements to the mean
  – important in defining service guarantees
  – important to understand real behavior

Mean is 50 but standard deviation is 20, indicating that there is a large variation in the value of the parameter
Mean of a sample

• To interpret a given measurement, we need to provide complete information
  – The mean
  – The standard deviation around the mean

Mean and standard deviation

• The standard deviation defines margins around the mean:
  – 64% of the values are within $\mu \pm \sigma$
  – 95% of the values are within $\mu \pm 2\sigma$
  – 99.7% of the values are within $\mu \pm 3\sigma$

• For a real system is very important to understand what happens when the values go beyond those margins (delays, overload, thrashing, system crash, etc.)
Calculating the standard deviation

• Mean and standard deviation:

\[ \mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \]

• In practice, use:

\[ s = \sqrt{\frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \]
Comparisons

- What is better?

**Deterministic behavior is often more important than good performance**
In practice

• In many systems, the standard deviation is almost more important than the mean:
  – 90% of the queries need to be answered in less than $X$ seconds
  – No web request can take longer than 5 seconds
  – Changes have to be propagated in less than 10 seconds
  – Guaranteed bandwidth higher than $X$ 90% of the time

• Achieving determinism is often done at the cost of performance
Confidence intervals
Background

• When measuring in software system, we typically do not know neither the value of the parameter we are measuring (µ) nor its standard deviation (σ)
• Instead, we work with mean of the sample \( \bar{x} \) and the estimated standard deviation (s)
  – the result is no longer a normal distribution but a t-distribution
  – The t-distribution depends on n, the amount of samples
  – For large n, the t.distribution tends to a normal distribution
Confidence interval

- Since typically we are not measuring an absolute value (unlike in the natural sciences), the notion of confidence interval is particularly useful in computer science.
- A confidence interval is a range (typically around the mean) where we can say that if we repeat the experiment 100 times, the value observed will be within the confidence interval \( m \) times (e.g., \( m=95 \), leading to a 95% confidence interval).
Calculation

• The confidence interval is calculated as follows:

\[ CI = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}} \]

Where \( s \) is the sample standard deviation, \( n \) the number of samples and \( t \) the critical value of the \( t \)-distribution.
## T in a table

Look up the value for the desired confidence interval and \((n-1)\)

<table>
<thead>
<tr>
<th></th>
<th>One Sided</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75%</td>
<td>80%</td>
<td>85%</td>
<td>90%</td>
<td>95%</td>
<td>97.5%</td>
<td>99%</td>
<td>99.5%</td>
<td>99.75%</td>
<td>99.9%</td>
<td>99.95%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>80%</td>
<td>85%</td>
<td>90%</td>
<td>95%</td>
<td>97.5%</td>
<td>99%</td>
<td>99.5%</td>
<td>99.75%</td>
<td>99.9%</td>
<td>99.95%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>60%</td>
<td>70%</td>
<td>80%</td>
<td>90%</td>
<td>95%</td>
<td>98%</td>
<td>99%</td>
<td>99.5%</td>
<td>99.8%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Two Sided</td>
<td>1.000</td>
<td>1.376</td>
<td>1.963</td>
<td>3.078</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>127.3</td>
<td>318.3</td>
<td>636.6</td>
</tr>
<tr>
<td></td>
<td>0.816</td>
<td>1.061</td>
<td>1.386</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>14.09</td>
<td>22.33</td>
<td>31.60</td>
</tr>
<tr>
<td></td>
<td>0.765</td>
<td>0.978</td>
<td>1.250</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>7.453</td>
<td>10.21</td>
<td>12.92</td>
</tr>
<tr>
<td></td>
<td>0.741</td>
<td>0.941</td>
<td>1.190</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>5.598</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td></td>
<td>0.727</td>
<td>0.920</td>
<td>1.156</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>4.773</td>
<td>5.893</td>
<td>6.869</td>
</tr>
<tr>
<td></td>
<td>0.718</td>
<td>0.906</td>
<td>1.134</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>4.317</td>
<td>5.208</td>
<td>5.959</td>
</tr>
<tr>
<td></td>
<td>0.711</td>
<td>0.896</td>
<td>1.119</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.029</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td></td>
<td>0.706</td>
<td>0.889</td>
<td>1.108</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>3.833</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td></td>
<td>0.703</td>
<td>0.883</td>
<td>1.100</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>3.690</td>
<td>4.297</td>
<td>4.781</td>
</tr>
</tbody>
</table>
Some observations

• For a fixed \( n \)
  – Increasing the confidence (100\%(1-\( \alpha \))) implies to extend the confidence interval

• To reduce the confidence interval
  – we decrease the confidence or,
  – we increase the number of examples

• For experiments, fix a target (typically 95% confidence in a 5-10% interval around the mean) and repeat the experiments until the level of confidence is reached –if ever ...
Example

- Mean = 122
- $s = 9$
- $n = 30$
- $t$ (two sided, 29, 95%) = 2.045
- CI = $122 \pm (2.045 \cdot 9/30^-2)$
- In 95 out of 100 runs, the mean will be between 119 and 125
Putting it all together
Look at all the data

• Make sure you are looking at the complete picture of the experiment and your measurements do not include side effects (warm up, cool down, repetition effects)

• Once you are sure you have identified the valid data and that it looks reasonable, then apply statistics to it
Standard deviation

• All measurements and graphs have to be accompanied by the standard deviation, otherwise they are meaningless
  – Provides an idea of the precision
  – Provides an idea of what will happen in practice
  – Provides an idea of how predictable performance is

• Repeat the experiments until you get a reasonable standard deviation (enough values are close enough to the mean)
How long to run?

• Until you reach a reasonable confidence level that the value lies within a reasonable margin of the mean

• Confidence intervals are the way to quantify how often the reported result is going to be observed in practice
  
  — “we repeated the experiments until we reached a 95% level confidence for an interval 5% around the mean”
Advice

• It is a good idea to run a long experiment to make sure you have seen all possible behavior of the system:
  – Glitches only every 3 hours
  – Memory leaks after 1 M transactions

• In reality, tests have to resemble how the system will be used in practice
Designing an experiment
Experiments, but which ones?

• What does it mean to design an experiment?
• Performance is affected by a large number of factors
  – Workloads
  – Systems
  – Knobs
• We are interested in:
  – Which ones are the most important?
  – Which ones are related?
• Goal: get the most information with least effort (minimum number of experiments)
Definitions

- A **response variable** is the outcome of an experiment - typically the measured performance of the system (e.g., throughput, response time).
- A **factor** is any variable that affects the response, and which has several alternatives (amount of memory, number of cores, data sizes).
- **Levels** are the values that a given factor can assume – the alternatives for a factor.
- **Primary factors** are those whose effects need to be quantified.
- **Secondary factors** are those that impact performance but whose effect we are not interested in quantifying.
- **Replications** are the number of times each experiment is to be repeated with particular levels for each factor.
An experiment

• An experimental design consists of:
  – the number of different experiments
  – the factor level combinations for each experiment
  – the number of replications of each experiment

• An experimental unit is any entity used for the experiment
Interaction

- Two factors interact if the effect of one depends on the level of the other.
- Interaction considerably complicates the business of interpreting experimental results.

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>$B_2$</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>$B_2$</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>
Avoid mistakes

• Try to avoid the following:
  – Ignoring the variation due to experimental errors
  – Not controlling important parameters (secondary factors)
  – Not isolating the effects of different factors
  – Overly simple (and very inefficient designs)
  – Ignoring interactions between factors
  – Conducting too many experiments
    • Take it slowly!
    • Break up the project into steps
Exploring the space

• Given a number of factors, what to do?
• Bad idea:
  – Vary one factor at a time
  – Find best value, fix it
  – Repeat for each factor
• Why is this a bad idea?: too many experiments, will get stuck in local minimum
2^k Factorial Designs

• Experimental technique to find the relative weight of different factors
  – Pick K factors
  – Pick two levels for each factor
  – Behavior of factors must be unidirectional or monotonic in the range explored (!)
Example $2^2$ Factorial Design

Observation: can update book examples by multiplying by 1000!

<table>
<thead>
<tr>
<th>Cache size (MB)</th>
<th>Memory size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4GB</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

Define variables $x_A$ and $x_B$ to represent levels for each factor:

$$x_A = \begin{cases} 
-1 & \text{if 4 GB main memory;} \\
1 & \text{if 16 GB main memory.} 
\end{cases}$$

$$x_B = \begin{cases} 
-1 & \text{if 1 MB cache,} \\
1 & \text{if 2 MB cache.} 
\end{cases}$$
Solving the model

A useful fiction: non-linear regression model for performance:

\[ y = q_0 + q_A \times A + q_B \times B + q_{AB} \times A \times B \]

This means we can write:

\[ 15 = q_0 - q_A - q_B + q_{AB} \]
\[ 45 = q_0 + q_A - q_B - q_{AB} \]
\[ 25 = q_0 - q_A + q_B - q_{AB} \]
\[ 75 = q_0 + q_A + q_B + q_{AB} \]
Relative weights on response variable

Solving:

\[ y = 40 + 20x_A + 10x_B + 5x_Ax_B \]

What does this mean?

- Mean performance is 40
- Effect of memory is 20
- Effect of cache is 10
- Interaction between the two accounts for 5
In the form of a table

The same calculation can be done using a sign table:

<table>
<thead>
<tr>
<th>l</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>45</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>75</td>
</tr>
<tr>
<td>160</td>
<td>80</td>
<td>40</td>
<td>20</td>
<td>Total</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>Total/4</td>
</tr>
</tbody>
</table>
Repetitions and errors

• Look in the book
  – How to allocate variation
  – How to consider repetitions of the experiments to look for errors

• For the milestone and analysis, please use repetitions to get meaningful results.