1 – Review of Probability and Statistics

List of Slides

1  Review of Probability and Statistics
5  Basic Probability Definitions
6  CDF and pdf Defined
7  CMF and PMF defined
8  Properties of the Exponential Distribution
9  The Exponential Distribution
11 Intro to Reliability
12 Intro to Reliability
13 In Class Exercise
14 Poisson Distributions and Poisson Processes
15 Poisson Distributions and Poisson Processes
16 Percentile Defined
17 Averaging and Variation defined
18 Averaging defined
19 Selecting a measure of the Average
20 Some Commonly Used Means
21 How To Choose a Mean?
22 Taking the Mean of a Ratio
24 Sample Mean
25 Variance and Sample Variance
26 COV and SIQR
27 Measuring Variation
28 Utilization and Performance
29 Performance and Throughput
30 Comparing Performance
31 Unix Measurements
32 Programs To Measure
33 Reporting Results
34  An Example
35  Some Conclusions
36  Total Execution Time Compared
37  But How to Compare Throughput?
38  Throughput Comparisons
39  Weighted Comparisons
40  Timings and Weighted Arithmetic Mean
41  Weighting Throughput
42  Normalized Run Times
43  Caveats of Normalized Run Times
44  Experimental Scientific Method Reviewed
45  Form of a Hypothesis
46  Experiment design
47  Controls and Confounding
48  Data Fitting and Linear Regression
49  How to do linear regression
50   Estimating Regression Parameters
51   An Example
52   Solution
53   Variation
Recall the following definitions:

- **Sample Space** — the set of outcomes we are selecting from.
- **Event** — an individual outcome.
- **Independent events** — Two events are independent if their outcomes do not affect each other.
- **Random Variable (rv)** — A variable is a random variable if it takes one of a specified set of values with a specified probability.
3 – CDF and pdf Defined

For Continuous Distributions we have:

- Cumulative Distribution Function (CDF) —
  For rv x the CDF is defined for continuous distributions as:

  \[ F_x(a) = \text{Prob}[x \leq a] \]  \( 1 \)

- Probability Distributions Function (pdf) —
  for rv x the pdf is defined for continuous distributions as:

  \[ f_x(a) = \text{Prob}[x = a] = \frac{d}{dx}F_x(a) \]  \( 2 \)

This gives another definition of the CDF:

\[ F_x(a) = \int_{-\infty}^{a} ayf_x(a) dy \]  \( 3 \)
4 – CMF and PMF defined

For Discrete Distributions we have:

- Probability Mass Function (pmf) — for rv $x$ the pmf is defined for discrete distributions where $x \in \{x_1, x_2, \ldots x_n\}$ and $(\forall i : 1 \leq i < n : x_i < x_{i+1})$ as:

$$f_x(x_i) = \text{Prob}[x = x_i] = p_i \quad (4)$$

- Cumulative Mass Function (CMF) — for rv $x$ is defined for discrete distributions as:

$$F_x(x_i) = \sum_{j=1}^{j \leq i} p_i \quad (5)$$

These correspond to the pdf and CDF of continuous distributions.
5 – Properties of the Exponential Distribution

Arrival and service times have an exponential distribution in an $M/M/1$ system.

The exponential distribution $T = \exp(a)$ has the properties [4]:

1. Is Memoryless: The time since the last event does not help to predict the time till the next event.

2. Parameter: $a > 0$ the scale parameter = mean

3. Range: $0 \leq t \leq \infty$

4. CDF: $F(t) = 1 - e^{-t/a} = \text{Prob}[T \leq t]$

5. pdf: $f(t) = \frac{1}{a} e^{-t/a} = \frac{d}{dt} F(t)$

6. Mean: $a$

7. Variance: $a^2$
6 – The Exponential Distribution
Figure 1: Exponential Distribution: pdf and CDF
The reliability factor at time \( t \) is defined as:

\[
R(t) = \frac{\text{The Number Surviving at instant } t}{\text{Number of items at } t = 0}
\]  
(6)

Often, the failure rate, \( \lambda \), is constant and \( N(t) \) be the number surviving at time \( t \), then:

\[
\frac{dN}{dt} = -\lambda N(t)
\]  
(7)

so integrating and substituting yields:

\[
R(t) = \frac{N(t)}{N(0)} = e^{-\lambda t}
\]  
(8)

The Mean Time to Failure (MTTF) is defined as:

\[
\text{MTTF} = \frac{1}{\lambda}
\]  
(9)
Your company sells web servers with the following components, with the following failure rates:

<table>
<thead>
<tr>
<th>Part</th>
<th>MTTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard Disk Drive</td>
<td>70 years</td>
</tr>
<tr>
<td>Network Card</td>
<td>50 years</td>
</tr>
<tr>
<td>Mother Board</td>
<td>100 years</td>
</tr>
</tbody>
</table>

The company has a 1/2 year hardware warranty for the server, but is considering using a 1 year warranty to lure in more customers. What would be the change in the reliability factor of the system during warranty.
The first step is to set up our notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>Total Web Server reliability factor</td>
</tr>
<tr>
<td>$R_H$</td>
<td>Hard Disk reliability factor</td>
</tr>
<tr>
<td>$R_N$</td>
<td>Network Card reliability factor</td>
</tr>
<tr>
<td>$R_M$</td>
<td>Motherboard reliability factor</td>
</tr>
</tbody>
</table>

Since the server would crash if any of the components failed, the server’s reliability factor is then:

$$R(t) = R_H(t) \times R_F(t) \times R_N(t) \times R_M(t) \quad (10)$$

Since we know the failure rate of each component, and the duration of the warranty get (where times are expressed in days):

$$R(0.5) = R_H(0.5) \times R_N(0.5) \times R_M(0.5) \quad (11)$$

$$= e^{-\frac{0.5}{70}} \times e^{-\frac{0.5}{50}} \times e^{-\frac{0.5}{100}} \approx 0.98 \quad (12)$$

$$R(1) = e^{-\frac{1}{70}} \times e^{-\frac{1}{50}} \times e^{-\frac{1}{100}} \approx 0.96 \quad (13)$$
10 – Poisson Distributions and Poisson Processes

Poisson processes or Poisson Streams are queueing models with IID interarrival times that are exponentially distributed. The number of arrivals, $k$, obeys a Poisson Distribution [4]:

1. $k$ is the number of arrivals of events with IID exponentially distributed interarrival times.

2. Parameters:
   (a) $\lambda = \text{mean arrival rate, mean arrival rate}$
   (b) $t, \lambda > 0$ is the of the interval during which arrivals are measured.

3. Range: $x = 0, 1, 2, \ldots, \infty$

4. pmf: $f(x) = \text{Prob}[k = x] = \lambda^x e^{-\lambda} / x!$

5. Mean: $\lambda$

6. Variance: $\lambda$
Figure 2: Graphs of the Poisson Distribution [1]
The points are the distribution but not the curves
12 – Percentile Defined

Often the percentile is used for such applications as acceptance criteria:

- *The ath percentile of* \( x \), denoted \( x_a \) is defined as satisfying:

\[
\text{Prob}[x \leq x_a] = \frac{100 - a}{100}
\]  

(14)

where \( 0 \leq a \leq 100 \) and \( x \) is an rv.
13 – Averaging and Variation defined

Jain [4] defines the *average* as the *central tendency* of the data.

*Variation* is the amount of dispersion of the data about its average value.
What does it mean to be average?

- (Arithmetic) Mean or Expected Value: is defined for rv $x$ is defined as:

$$E[x] = \begin{cases} 
\sum_{i=1}^{n} x_i p_i & \text{if } x \text{ is discrete}, \\
\int_{-\infty}^{\infty} y f_x(y) dy & \text{if } x \text{ is continuous.}
\end{cases}$$  

\hspace{1cm} (15)

- Median: The 50th percentile of a continuous rv $x$.

- Mode: The most likely value of $x$, i.e. where the pdf or pmf is maximum.
15 – Selecting a measure of the Average

To select between the (arithmetic) mean, median and mode:
Some commonly used means are:

- **(Arithmetic) Mean:** is defined for rv $x$ is defined as:

$$E[x] = \begin{cases} \sum_{i=1}^{n} x_i p_i & \text{if } x \text{ is discrete,} \\ \int_{-\infty}^{\infty} y f_x(y) dy & \text{if } x \text{ is continuous.} \end{cases}$$

(16)

- **Geometric mean:** given $n$ values, $x_1, x_2, \ldots, x_n$, the geometric mean is denoted:

$$\hat{x} = \left( \prod_{i=1}^{n} x_i \right)^{1/n} \quad (17)$$

- **Harmonic Mean:** given $n$ discrete values, $x_1, x_2, \ldots, x_n$, the harmonic mean is denoted:

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad (18)$$
17 – How To Choose a Mean?

Jain [4] recommends that when the values being averaged are NOT a ratio:

- Select the \textit{arithmetic mean} when the sum of values is of interest.
- Select the \textit{geometric mean} when the product of values is of interest.
Jain [4] recommends that when averaging a ratio:

1. If the sums of numerators and denominators both have a physical meaning, the average of the ratio is the ratio of the averages.

\[
\text{Mean} \left( \frac{a_i}{b_i} \right) = \frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} = \frac{(1/n) \sum_{i=1}^{n} a_i}{(1/n) \sum_{i=1}^{n} b_i} = \frac{E[a]}{E[b]} \tag{19}
\]

(a) If the denominator is a constant and the sum of the numerator has a physical meaning, the arithmetic mean of the ratios can be used, i.e. \( b_i = b \) for all \( i \) so:

\[
\text{Mean} \left( \frac{a_i}{b} \right) = \frac{a_1 + a_2 + \cdots + a_n}{b + b + \cdots + b} = \frac{\sum_{i=1}^{n} a_i}{nb} = \frac{E[a]}{nb} \tag{20}
\]

(b) If the numerator is a constant and the
sum of the denominator has a physical meaning, then a harmonic mean should be used, i.e. $a_i = a$ for all $i$ so:

$$\text{Mean} \left( \frac{a}{b_i} \right) = \frac{a + a + \cdots + a}{b_1 + b_2 + \cdots + b_n} = \frac{n}{b_1/a + b_2/a + \cdots b_n/a} = \frac{na}{\sum_{i=1}^{n} b_i} \quad (21)$$

2. If the numerator and denominator are expected to follow a multiplicative property such that $a_i = cb_i$, where $c$ is (approximately) a constant to estimate, then $c$ can be estimated by using the geometric mean of $a_i/b_i$:

$$c \approx \text{Mean} \left( \frac{a_i}{b_i} \right) = \left( \prod_{i=1}^{n} \frac{a_i}{b_i} \right)^{1/n} \quad (22)$$
The \textit{sample mean} averages together a set of data points, $x = \{x_1, x_2, \ldots, x_n\}$, so:

$$\hat{E}[x] = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$  \hspace{1cm} (23)

The weak law of large numbers [5, 6] (informally) states that when the number of samples is large the sample mean more closely approximates the mean of the measured distribution:

$$\lim_{n \to \infty} \hat{E}[x] = E[x]$$ \hspace{1cm} (24)
The variance of $x$ measures its dispersion about the mean:

$$\text{VAR}[x] = E[(x - E[x])^2] = E[x^2] - (E[x])^2 \quad (25)$$

The standard deviation of $x$ is defined as:

$$\sigma_x = \sqrt{\text{VAR}[x]} \quad (26)$$

The sample variance measures dispersion of measurements about the sample mean [2]:

$$\overline{\text{VAR}}[x] = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{E}[x]) \quad (27)$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{(\sum_{i=1}^{n} x_i)^2}{n(n-1)} \quad (28)$$

While form (28) appears more complex, it has the advantage that it can be computed in a single pass over the data.
The coefficient of variation (COV) is defined as:

$$\text{COV}[x] = \frac{\sigma_x}{E[x]} \quad (29)$$

Which estimates the relative error of the estimate. The \(k\)th quartile of rv \(x\) is the \(k \times 25\)th percentile, that is:

$$Q_k = x_{k \times 25} \quad (30)$$

The semi-interquartile range (SIQR) is defined as:

$$\text{SIQR}[x] = \frac{Q_3 - Q_1}{2} = \frac{x_{75} - x_{25}}{2} \quad (31)$$
Jain [4] describes how to measure variability of data using indices of dispersion:

- For Bounded Data — Use the range (just keep track of min/max values).
- Else if Unimodal Symmetric Distribution — Use Coefficient of Variation
- Otherwise — Use Percentiles (e.g. SIQR)
23 – Utilization and Performance

Expensive computing resources should be used.

1. Idle Time — How long the resources are unused

2. Utilization — The ratio: \[
\frac{\text{Time Used}}{\text{Total Time}}
\]

They call throughput *performance*, so:

$$\text{Performance}_X = \frac{1}{\text{Execution Time}_X} \quad (32)$$
25 – Comparing Performance

Saying $X$ is $n$ times faster than $Y$ means:

$$n = \frac{\text{Execution Time}_Y}{\text{Execution Time}_X} = \frac{1}{\text{Performance}_Y} = \frac{1}{\text{Performance}_X} = \frac{\text{Performance}_X}{\text{Performance}_Y}$$

(33)
Most flavors of Unix have a \textit{time} command which takes an executable and returns how long it took:

\texttt{-time lectfoils.tex}

\texttt{1.03u 0.39s 0:01.90 74.7%}

Means that this job took 1.03 seconds of user mode time (i.e. not in the kernel) and 0.39 seconds of system time (i.e. in the kernel) and \texttt{0:01.90} means it took 1.9 seconds \textit{wall clock time} and had 74.7\% processor utilization.

Try the following on a Unix box (say an SGI):

\texttt{man getrusage}
The following are candidate programs for performance measurement:

1. Real Programs — This is best, however real data sets may be prohibitive to run, and the vendor may try some more commonly used programs.

2. Kernels — Extract some small (hopefully representative) part of a real program and measure its performance (e.g. Livermore loops, Linpack).

3. Toy Benchmarks — Small programs (10-100 lines of code) produce a known result (e.g. Sieve of Erastosthenes, Puzzle, Quicksort).

4. Synthetic Benchmarks — The instructions issued during program execution are sampled and a statistically similar work load is run. This is not a good idea, since it accurate
Performance results should be *reproducible*:

1. The executable should be obtained in a specified (which version if purchased, and if compiled the source version, compiler version and compiler options used).

2. The data should be specified.

3. The machine configuration should be specified:

   (a) Operating System (including installation configuration).

   (b) Amount of memory/speed

   (c) Peripherals used (and their configuration).

4. If possible compare optimized vs. unoptimized (or *baseline*) results.
Consider the following timings (in seconds) on computers $A$, $B$, and $C$:

<table>
<thead>
<tr>
<th>Program</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>P2</td>
<td>1000</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>1001</td>
<td>110</td>
<td>40</td>
</tr>
</tbody>
</table>
So we can see:

- $A$ is 10 times faster than $B$ for P1.
- $B$ is 10 times faster than $A$ for P2.
- $A$ is 20 times faster than $C$ for P1.
- $C$ is 50 times faster than $A$ for P2.
- $B$ is 2 times faster than $C$ for P1.
- $C$ is 5 times faster than $B$ for P2.

But what does this really mean?
31 – Total Execution Time Compared

Total execution times are:

<table>
<thead>
<tr>
<th>Program</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>P2</td>
<td>1000</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>1001</td>
<td>110</td>
<td>40</td>
</tr>
</tbody>
</table>

Letting us say:

- B is 9.1 times faster than A for P1 and P2.
- C is 25 times faster than A for P1 and P2.
- C is 2.75 times faster than B for P1 and P2.
Suppose we are given the throughput as the performance measure but **NOT** the timings. Then our given for the corresponding system is:

<table>
<thead>
<tr>
<th>Program</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>$\frac{1}{10}$</td>
<td>$\frac{1}{20}$</td>
</tr>
<tr>
<td>P2</td>
<td>$\frac{1}{1000}$</td>
<td>$\frac{1}{100}$</td>
<td>$\frac{1}{20}$</td>
</tr>
</tbody>
</table>

How can we compare performance?
When comparing several throughputs, use the harmonic mean as per case 1b.

<table>
<thead>
<tr>
<th>Program</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>1/10</td>
<td>1/20</td>
</tr>
<tr>
<td>P2</td>
<td>1/1000</td>
<td>1/100</td>
<td>1/20</td>
</tr>
<tr>
<td>$\bar{P}$</td>
<td>2/1001</td>
<td>2/110</td>
<td>2/40</td>
</tr>
</tbody>
</table>

where $\bar{P} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\text{Performance}_i}}$.

Remember for execution time smaller values are better while for throughput larger values are better.
34 – Weighted Comparisons

Sometimes the times are not equal in importance (e.g., some jobs run more often). Consider the following workload distributions:

<table>
<thead>
<tr>
<th>Job</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>W(1)</th>
<th>W(2)</th>
<th>W(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>10</td>
<td>20</td>
<td>0.50</td>
<td>0.909</td>
<td>0.999</td>
</tr>
<tr>
<td>P2</td>
<td>1000</td>
<td>100</td>
<td>20</td>
<td>0.50</td>
<td>0.091</td>
<td>0.001</td>
</tr>
</tbody>
</table>

How do we take this into account?
35 – Timings and Weighted Arithmetic Mean

For weighted timing measurements use the weighted arithmetic mean:

\[ \sum i = 1^n \text{Weight}_i \times \text{Time}_i \]  

(34)

Using the means from the previous slide:

<table>
<thead>
<tr>
<th>W.A.M.</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>W(1)</td>
<td>500.50</td>
<td>55.0</td>
<td>20.0</td>
</tr>
<tr>
<td>W(2)</td>
<td>91.82</td>
<td>18.18</td>
<td>20.0</td>
</tr>
<tr>
<td>W(3)</td>
<td>2.00</td>
<td>10.09</td>
<td>20.00</td>
</tr>
</tbody>
</table>
36 – Weighting Throughput

For weighted throughput measurements use the weighted arithmetic mean:

\[
\sum i = \frac{1}{\sum i = 1^n \frac{\text{Weight}_i}{\text{Performance}_i}} = \frac{1}{\sum i = 1^n \text{Weight}_i \times \text{Time}_i}
\]

Repeating the above exercise gives:

<table>
<thead>
<tr>
<th>W.A.M.</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W(1))</td>
<td>(\frac{1}{500.50})</td>
<td>(\frac{1}{55.0})</td>
<td>(\frac{1}{20.0})</td>
</tr>
<tr>
<td>(W(2))</td>
<td>(\frac{1}{91.82})</td>
<td>(\frac{1}{18.18})</td>
<td>(\frac{1}{20.0})</td>
</tr>
<tr>
<td>(W(3))</td>
<td>(\frac{1}{2.00})</td>
<td>(\frac{1}{10.09})</td>
<td>(\frac{1}{20.00})</td>
</tr>
</tbody>
</table>
Let $X_i$ denote the run time of the $i$th job on $X$ and $Y_i$ denote the run time of the $i$th job on $Y$. The *run time of $i$ on $X$ normalized to $Y$* is $\frac{X_i}{Y_i}$.

The geometric mean is well suited to averaging normalized run times as per case 2 mentioned above.

<table>
<thead>
<tr>
<th>Job</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1.0</td>
<td>10.0</td>
<td>20.0</td>
</tr>
<tr>
<td>p2</td>
<td>1.0</td>
<td>0.1</td>
<td>0.02</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td>1.0</td>
<td>1.0</td>
<td>0.63</td>
</tr>
<tr>
<td>Total Time</td>
<td>1.0</td>
<td>0.11</td>
<td>0.04</td>
</tr>
</tbody>
</table>

The geometric means of normalized times are consistent regardless of which machine is chosen for the reference. This is NOT true of the arithmetic mean.
38 – Caveats of Normalized Run Times

The geometric means of normalized execution times do not predict run time. It masks too many features of the work load, and should not be used.
39 – Experimental Scientific Method Reviewed

Experimental computer science relies on the classic method:

1. Define a problem — Choose a phenomena
2. Create a hypothesis — Try to predict the phenomena
3. Perform an Experiment — Get controlled measures of the phenomena
4. Statistically Confirm or Refute Hypothesis

Let’s explore experiment design and statistical testing of the hypothesis.
40 – Form of a Hypothesis

Typically a (quantitative) hypothesis proposes that:

- A set of \( k \) measurable influences on the phenomena called *controls*, denoted \( \{x_1, x_2, \ldots, x_k\} \).

- A measurement of the impact of the controls \( y \).

- A mapping of controls to relations:
  \[ y = f(x_1, x_2, \ldots, x_k). \]
41 – Experiment design

The experiment can be designed as a matrix, where

- Each row in the matrix consists of an individual *trial*.
- Each column represents a particular control.
If resources permit, we want an experiment to:

- Each trial has controls which differ in at most 1 position from some other trial.
- Systematically varies each control \( x_i \) over a range of values.
  - A common design technique tries two values for each control to test for sensitivity called \( 2^k \) experiment design.
- Trials may be repeated \( r \) times for improved accuracy, giving rise to \( 2^k r \) experiment design.

If we do not try each control in isolation, 

*confounding* can occur, where it is not possible to uniquely identify the cause of the measured effect.
43 – Data Fitting and Linear Regression

Consider the following (simple) example experiment: with one control.

- There is one control, $x$.
- The hypothesis predicts $\hat{y} = f(x) = b_1 x + b_0$.
- The slope $b_1$ and $y$ intercept, $b_0$ are unknown regression parameters.
- We have a performed a set of $n$ experiments, $\{(x_i, y_i)\}$ where $1 \leq i \leq n$.
- The residual error, $e_i$, of the $i$th trial is:

$$e_i = y_i - \hat{y}_i = y_i - b_1 x_i - b_0 \quad (36)$$

We want to fit a line to the data (that is pick the line closest to the data points).
How to do linear regression

The best fit line has $b_0$ and $b_1$ such that:

- It minimizes the *Sum of Squared Errors* (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i) = e_i^2$$  \hspace{1cm} (37)

$$= \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2$$ \hspace{1cm} (38)

- and constrains the mean error to be zero:

$$E[e_i] = \overline{e_i} = \frac{1}{n} \sum_{i=1}^{n} e_i = \frac{1}{n} \sum_{i=1}^{n} (y_i - b_0 - b_i x_i) = 0$$ \hspace{1cm} (40)
45 – Estimating Regression Parameters

The following solution can be shown to satisfy the constraints:

\[
b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n(\overline{x})(\overline{y})}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \quad (41)
\]

\[
b_0 = \overline{y} - b_1 \overline{x} \quad (42)
\]

\[
\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (43)
\]

\[
\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (44)
\]
Suppose we observe (via experiment):

<table>
<thead>
<tr>
<th>Disk I/O</th>
<th>14</th>
<th>16</th>
<th>27</th>
<th>42</th>
<th>39</th>
<th>50</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Time (in msec)</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>20</td>
</tr>
</tbody>
</table>

Suppose we want to derive a linear function to predict CPU time, $y$, based on the number of disk I/O’s, $x$.

What do we do?
47 – Solution

Following the principles of linear regression:

\[ n = 7 \]  \hspace{1cm} (45)

\[ \bar{x} = \sum_{i=1}^{n} x_i = 38.71 \]  \hspace{1cm} (46)

\[ \sum_{i=1}^{n} x_i y_i = 3375 \]  \hspace{1cm} (47)

\[ \sum_{i=1}^{n} x_i^2 = 1385 \]  \hspace{1cm} (48)

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 9.43 \]  \hspace{1cm} (49)

\[ b_1 = \frac{\left( \sum_{i=1}^{n} x_i y_i \right) - n(\bar{x})(\bar{y})}{\left( \sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2} \]  \hspace{1cm} (50)

\[ = \frac{3375 - 7 \times 38.71 \times 9.43}{13855 - 7 \times 38.71^2} = 0.2438 \]

\[ b_0 = \bar{y} - b_1 \bar{x} = 9.43 - 38.71 = -0.0083 \]
The total sum of squares (SST) is:

\[
\begin{align*}
\text{SST} & = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\sum_{i=1}^{n} y_i^2) - n\bar{y}^2 \\
\text{SSY} & = \sum_{i=1}^{n} y_i^2 \\
\text{SS0} & = \sum_{i=1}^{n} \bar{y}^2 = n\bar{y}^2 \\
\text{SST} & = \text{SSY} - \text{SS0} \\
\text{SSR} & = \text{SST} - \text{SSE}
\end{align*}
\] (51-55)

The coefficient of determination, \( R^2 \) is the sample correlation between \( x \) and \( y \):

\[
R^2 = \frac{\text{SSR}}{\text{SST}}
\] (56)

Typically \( 0 \leq R^2 \leq 1 \), with larger values indicating a better fit.
References


