COMP 1215: Foundations of Computer Science
Introduction to Probability and Statistics

Kate Farrahi

2017/8
Tipli and Pikli are a married couple (don't ask me who he is and who she is)
They have two children, one of the children is a boy. Assume that the probability of each gender is 1/2. What is the probability that the other child is also a boy?
Hint: It is not 1/2 as you would first think.
Teaser Solution

Answer: \( \frac{1}{3} \)

The following are possible combinations of two children that form a sample space in any earthly family:
{Boy - Girl, Girl - Boy, Boy - Boy, Girl - Girl}

Since we know one of the children is a boy, we will drop the girl-girl possibility from the sample space. This leaves only three possibilities, one of which is two boys. Hence the probability is 1/3


**Definitions: Probability and Statistics**

**Probability** is the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible.

**Statistics** is a collection of methods for collecting, displaying, analyzing, and drawing conclusions from data.
Probability and Statistics in the Real World!

A basic understanding of probability and statistics makes it possible for us to understand many important events in the real world.

- Weather forecasting
- The lottery
- Finance
- Sports
- Medicine
- Are we likely to get struck by lightning?
- Artificial intelligence and Machine Learning!
Textbook

References

- Modern Probability Theory and Its Applications by Emanuel Parzen. Published by Wiley.

- Miller Freund’s Probability and Statistics: for Engineers by Richard Johnson. Published by Pearson.

- Probability and Statistics by DeGroot and Schervish. Published by Pearson.


Descriptive Statistics
Suppose 30 students in a statistics class took a test and made the following scores:

86  80  25  77  73  76  100  90  69  93
90  83  70  73  73  70  90  83  71  95
40  58  68  69  100  78  87  97  92  74

How did the class do on the test? A quick glance at the set of numbers does not immediately give a clear answer.
## Ordered Stem and Leaf Diagram

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>
Frequency Histograms
Relative Frequence Histograms

![Relative Frequency Histogram](image)

- Score
- Relative Frequency
Sample Sizes

(a) Small Sample

(b) Medium Sample

(c) Large Sample

(d) Very Large Sample
Measures of Central Location

The **mean** or **arithmetic mean** of a set of $N$ sample data $x_1, x_2, ..., x_N$ is a measure of central location and is defined by the formula $\bar{x} = \frac{1}{N} \left( \sum_{i=1}^{N} x_i \right)$.

The **median**, $\tilde{x}$, is the middle measurement when the data are arranged in numerical order. The median of a set of sample data for which there are an even number of measurements is the mean of the middle two measurements when the data are arranged in numerical order.

The **mode** of a set of sample data is the most frequently occurring value.
Skewness of Relative Frequency Histograms

(a) $\bar{x} = \tilde{x}$

(b) $\bar{x} = \tilde{x}$

(c) $\bar{x} > \tilde{x}$

(d) $\bar{x} < \tilde{x}$
When the distribution is symmetric, as in panels (a) and (b) of slide 12, the mean and the median are equal.

When the distribution is as shown in panel (c), it is said to be skewed right. The mean has been pulled to the right of the median by the long right tail of the distribution, the few relatively large data values.

When the distribution is as shown in panel (d), it is said to be skewed left. The mean has been pulled to the left of the median by the long left tail of the distribution, the few relatively small data values.
Question

Begin with the following set of data, call it Data Set I.
5 -2 6 14 -3 0 1 4 3 2 5

- Compute the mean, median, and mode.
- Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the mean, median, and mode of Data Set II.
- Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the mean, median, and mode of Data Set III.

Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.
Solution

- $\bar{x} = 3.18, \tilde{x} = 3, \text{mode} = 5$
- $\bar{x} = 6.18, \tilde{x} = 6, \text{mode} = 8$
- $\bar{x} = -2.18, \tilde{x} = -2, \text{mode} = -1$

If a number is added to every measurement in a data set, then the mean, median, and mode all change by that number.
Measures of Variability

Given the two datasets below, each have a mean, median, and mode of 40.

Data Set I  40 38 42 40 39 39 43 40 39 40
Data Set II 46 37 40 33 42 36 40 47 34 45

Data Set I has measurements that vary only slightly from the center, while Data Set II has measurements that vary greatly from the center. Measures of variability quantify how the data scatter away from the center.
Measures of Variability

The **range** of a data set is the number $R$ defined by the formula $R = x_{\text{max}} - x_{\text{min}}$ where $x_{\text{max}}$ is the largest measurement of the data set and $x_{\text{min}}$ is the smallest.

The **variance** of a set of $N$ sample data is the number $\sigma^2$ defined by the formula $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$

The **standard deviation** of a set of $N$ sample data is the square root of the variance, $\sigma = \sqrt{\left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}\right)}$
Measures of Variability - Range

Data Set I 40 38 42 40 39 39 43 40 39 40
Data Set II 46 37 40 33 42 36 40 47 34 45

The range of Data Set I is \( R = 43 - 38 = 5 \)

The range of Data Set II is \( R = 47 - 33 = 14 \)

The range indicates the size of the interval over which the data points are distributed.
Measures of Variability - Variance and Standard Deviation

Data Set I 40 38 42 40 39 39 43 40 39 40
Data Set II 46 37 40 33 42 36 40 47 34 45

Homework: Compute the variance and standard deviation of Data Set I and II.
Relative Position of Data - Percentiles and Quartiles

In general, the significance of one observed value in a data set strongly depends on how that value compares to the other observed values in a data set. Therefore we wish to attach to each observed value a number that measures its relative position.
Given an observed value $x$ in a data set, $x$ is the $P^{th}$ percentile of the data if the percentage of the data that are less than or equal to $x$ is $P$. The number $P$ is the percentile rank of $x$. 
Percentiles and Quartiles

**Question** Given the following data set,  
1.39, 1.76, 1.90, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00.

1. What percentile is the value 1.39?  
2. What percentile is the value 3.33?
Solution

1. The only data value that is less than or equal to 1.39 is 1.39 itself. Since 1 is \( \frac{1}{10} = 0.10 \) or 10% of 10, the value 1.39 is the 10th percentile.

2. Eight data values are less than or equal to 3.33. Since 8 is \( \frac{8}{10} = 0.80 \) or 80% of 10, the value 3.33 is the 80th percentile.
Quartiles

The $P$th percentile cuts the data set in two so that approximately $P\%$ of the data lie below it and $(100 - P)\%$ of the data lie above it. In particular, the three percentiles that cut the data into fourths are called the **quartiles**.
Quartiles

Definition

For any data set:

1. The second quartile $Q_2$ of the data set is its median.
2. Define two subsets:
   - the lower set: all observations that are strictly less than $Q_2$;
   - the upper set: all observations that are strictly greater than $Q_2$.
3. The first quartile $Q_1$ of the data set is the median of the lower set.
4. The third quartile $Q_3$ of the data set is the median of the upper set.
Question Find the quartiles of the following ordered dataset: 1.39, 1.76, 1.9, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00
**Solution**  \( N = 10 \) observations. The median is the mean of the two middle observations:  \( \bar{x} = \frac{(2.53+2.71)}{2} = 2.62 \). Thus the second quartile is \( Q_2 = 2.62 \). The lower and upper subsets are

Lower: \( L = \{1.39, 1.76, 1.90, 2.12, 2.53\} \)

Upper: \( U = \{2.71, 3.00, 3.33, 3.71, 4.00\} \)

Each has an odd number of elements, so the median of each is its middle observation. Thus \( Q_1 = 1.90 \), and \( Q_3 = 3.33 \).
Question Add 3.88 to the previous dataset and find the quartiles:
1.39, 1.76, 1.9, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 3.88, 4.00
Quartiles

**Solution** This data set has $N = 11$ observations. $\bar{x} = 2.71$. Thus the second quartile is $Q_2 = 2.71$. The lower and upper subsets are

Lower: $L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$

Upper: $U = \{3.00, 3.33, 3.71, 3.88, 4.00\}$

$Q_1 = 1.90$ and $Q_3 = 3.71$. 
In the box plot, each of the numbers of the five-number summary, $\{x_{min}, Q_1, Q_2, Q_3, x_{max}\}$, is represented by a vertical line segment. A box is formed using the line segments at $Q_1$ and $Q_3$. Note, there are other types of box plots that differ from this configuration.
Interquartile Range

The interquartile range (IQR) is the quantity $IQR = Q_3 - Q_1$. 
Question
Construct a box plot for the following dataset: 1.39, 1.76, 1.9, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00
Box Plot Example

Solution

\[ x_{\text{min}} = 1.39, \quad Q_1 = 1.90, \quad Q_2 = 2.62, \quad Q_3 = 3.33, \quad x_{\text{max}} = 4.00. \]

The \( IQR = 3.33 - 1.90 = 1.43 \)
The Empirical Rule

If a data set has an approximately bell-shaped relative frequency histogram, then

- approximately 68% of the data lie within one standard deviation of the mean.
- approximately 95% of the data lie within two standard deviations of the mean; and
- approximately 99.7% of the data lies within three standard deviations of the mean.
The Empirical Rule - Example

Heights of 100 men

68.7  72.3  71.3  72.5  70.6  68.2  70.1  68.4  68.6  70.6
73.7  70.5  71.0  70.9  69.3  69.4  69.7  69.1  71.5  68.6
70.9  70  70.4  68.9  69.4  69.4  69.2  70.7  70.5  69.9
69.8  69.8  68.6  69.5  71.6  66.2  72.4  70.7  67.7  69.1
68.8  69.3  68.9  74.8  68.0  71.2  68.3  70.2  71.9  70.4
71.9  72.2  70.0  68.7  67.9  71.1  69.0  70.8  67.3  71.8
70.3  68.8  67.2  73.0  70.4  67.8  70.0  69.5  70.1  72.0
72.2  67.6  67.0  70.3  71.2  65.6  68.1  70.8  71.4  70.2
70.1  67.5  71.3  71.5  71.0  69.1  69.5  71.1  66.8  71.8
69.6  72.7  72.8  69.6  65.9  68.0  69.7  68.7  69.8  69.7
The mean of the heights data is $\bar{x} = 69.92$ and the standard deviation is $\sigma = 1.7$.

If we count the number of observations that are within one standard deviation of the mean, there are $69.96 - 1.7 = 68.22$ and $69.92 + 1.7 = 71.62$, there are 69 of them.

If we count the number of observations that are within two standard deviations of the mean, there are $69.96 - 2(1.7) = 66.52$ and $69.92 + 2(1.7) = 73.32$, there are 95 of them.

All of the measurements are within three standard deviations.
Chebyshev’s Theorem

For any numerical data set,

1. at least $\frac{3}{4}$ of the data lie within two standard deviations of the mean;
2. at least $\frac{8}{9}$ of the data lie within three standard deviations of the mean;
3. at least $1 - \frac{1}{k^2}$ of the data lie within $k$ standard deviations of the mean, where $k$ is any positive whole number that is greater than 1.
Basic Concepts of Probability
A random phenomenon is an empirical phenomenon characterized by the property that its observation under a given set of circumstances does not always lead to the same observed outcome but rather to different outcomes in such a way that there is statistical regularity.

Probability theory is the study of mathematical models of random phenomena.
A random event is one whose relative frequency of occurrence, in a very long sequence of observations of randomly selected situations in which the event may occur, approaches a stable limit value as the number of observations is increased to infinity.

The limit value of the relative frequency is called the probability of the event.

Given the following sequence of observations, H, H, T, H, T, T, H, H, T, H, T, T, T, what are the random events?
Let us examine an urn containing six balls, 4 of which are white and 2 are red. Let a ball be drawn and its color noted.

- What will be the color of a ball drawn from the urn?
- What can we conclude about the color of the ball drawn from the urn?
Question

The statistical abstract of the US reports that among the several million babies born in the US the number of boys born per 1000 girls was as follows for the years listed:

<table>
<thead>
<tr>
<th>Year</th>
<th># Male Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>1935</td>
<td>1053</td>
</tr>
<tr>
<td>1940</td>
<td>1054</td>
</tr>
<tr>
<td>1945</td>
<td>1055</td>
</tr>
<tr>
<td>1950</td>
<td>1054</td>
</tr>
<tr>
<td>1951</td>
<td>1052</td>
</tr>
<tr>
<td>1952</td>
<td>1051</td>
</tr>
<tr>
<td>1953</td>
<td>1053</td>
</tr>
<tr>
<td>1954</td>
<td>1051</td>
</tr>
<tr>
<td>1955</td>
<td>1051</td>
</tr>
</tbody>
</table>

What is the probability of this random event?
Sample Spaces

- How can one formulate postulates concerning a random phenomenon?
- This is done by introducing the **sample description space (or sample space)** of the random phenomenon.
- The sample space of a random phenomenon, usually denoted by $S$, is the space of all possible outcomes of the phenomenon.
- A sample is often the outcome of an experiment or observation.
- The sample space consists of a **set**.
Generally, sample spaces are classified according to the number of elements that they contain.

The **size** of the sample set is defined as the number of members of $S$.

A sample space is defined to be **finite** if its size is one of the finite numbers $1, 2, 3, \ldots$. 
Example of a Sample Space

- Let us consider an urn containing six balls, numbered 1 to 6.
- What is the sample space $S$
- $S = \{1, 2, 3, 4, 5, 6\}$
- Now suppose that one draws two balls from an urn containing six balls, numbered 1 to 6. Suppose that the first ball is *not* returned to the urn before drawing the second ball.
- Suppose that the first ball drawn is 5 and the second ball drawn is 3. The outcome of the experiment is $(5, 3)$. Note $(3, 5)$ would have represented a different outcome.
- What is the sample space of drawing two balls and how many members does it have?
Example Sample Space

\[ S = \{ (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\
(2, 1), (2, 3), (2, 4), (2, 5), (2, 6), \\
(3, 1), (3, 2), (3, 4), (3, 5), (3, 6), \\
(4, 1), (4, 2), (4, 3), (4, 5), (4, 6), \\
(5, 1), (5, 2), (5, 3), (5, 4), (5, 6), \\
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5) \} \]
Example Sample Space - Marriage

Suppose one is observing the ages (in years) of couples who apply for marriage licenses in a certain city. For notation, if we observe a man and woman aged 24 and 22, respectively, getting married, we record this observation as 2-tuple \((24, 22)\). For simplicity, we assume the age at which a man or woman may get married to be any number from 1-200. We can describe the sample space as follows:

\[
S = \{2 \text{-tuples}(x, y) : x \text{ is any integer, 1 to 200,} \quad y \text{ is any integer, 1 to 200}\}
\]
A sample description space can be finite or infinite. An example of an infinite sample space.

- If we are measuring the time (microseconds) between two neighbouring peaks on an electrocardiogram, then we might take the set $\mathcal{S} = \{ \text{real numbers } x : 0 < x < \infty \}$.
The Notion of Events

Given the experiment of drawing two balls from an urn with balls numbered from 1 to 6. Now assume that two of the six balls are white (numbered 1 and 2). Some possible events are

1. the event that the ball drawn on the first draw is white,
2. the event that the ball drawn on the second draw is white,
3. the event that both of the balls drawn are white,
4. the event that the sum of the numbers on the balls drawn is 7,
5. the event that the sum of the numbers on the balls drawn is less than or equal to 4.
We define an event $E$ as any subset of the sample space $\mathcal{S}$. 
Relations and Operations on Events

- $E^c$ denotes the complement of $E$ and is the event that $E$ does not occur.
- Consider two events, $E$ and $F$. We may ask whether $E$ and $F$ both occurred or whether at least one of them occurred.
- The intersection $E \cap F$ is said to occur if and only if both $E$ and $F$ occur.
- The union $E \cup F$ is said to occur if and only if either $E$ or $F$ occurs.
- The events are said to be equal, written $E = F$, if every description in one event belongs to the other. $E = F$ if and only if $E \subset F$ and $F \subset E$.
- The impossible event, denoted $\emptyset$ is defined as the event that is empty and therefore cannot occur.
Example

Let us consider the experiment of drawing a ball from an urn containing twelve balls, numbered 1 to 12. Then $S = \{1, 2, \ldots, 12\}$.

Consider the events $E = \{1, 2, 3, 4, 5, 6\}$ and $F = \{4, 5, 6, 7, 8, 9\}$.

What is $E^c$, $E \cap F$ and $E \cup F$?

$E^c = \{7, 8, 9, 10, 11, 12\}$

$E \cap F = \{4, 5, 6\}$

$E \cup F = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
The Definition of Probability as a Function of Events on a Sample Space

Given a random situation, which is described by a sample space $S$, probability of a function $P[.]$ that every event $E$ assigns a nonnegative real number, denoted by $P[E]$ and called by the probability of the event $E$. The probability function must satisfy three axioms:

- $P[E] \geq 0$ for every event $E$,
- $P[S] = 1$ for the certain event $S$,
- $P[E \cup F] = P[E] + P[F]$, if $E \cap F = \emptyset$ (The probability of the union of two mutually exclusive events is the sum of their probabilities.)
Finite Sample Spaces

Consider a finite sample space $S$ of size $N$. We may list the members of $S = \{D_1, D_2, D_3, D_4\}$, where $D_1 = (H, H)$, $D_2 = (H, T)$, $D_3 = (T, H)$, $D_4 = (T, T)$. $2^N$ possible events may be defined on a sample space of finite size $N$.

Then there are sixteen possible events that may be defined; namely $S$, $\emptyset$, $\{D_1\}$, $\{D_2\}$, $\{D_3\}$, $\{D_4\}$, $\{D_1, D_2\}$, $\{D_1, D_3\}$, $\{D_1, D_4\}$, $\{D_2, D_3\}$, $\{D_2, D_4\}$, $\{D_3, D_4\}$, $\{D_1, D_2, D_3\}$, $\{D_1, D_2, D_4\}$, $\{D_1, D_3, D_4\}$, $\{D_2, D_3, D_4\}$.

Note, a single-member event is an event that contains exactly one description.
Calculating the Probabilities of Events When the Sample Space is Finite

Let $E$ be any event on a finite sample space $S = \{D_1, D_2, \ldots, D_N\}$. Then the probability $P[E]$ of the event $E$ is the sum, over all descriptions $D_i$ that are members of $E$, of the probabilities $P[\{D_i\}]$. For example, if $E = \{D_{i1}, D_{i2}, \ldots, D_{ik}\}$ then

$$P[E] = P[\{D_{i1}\}] + P[\{D_{i2}\}] + \ldots + P[\{D_{ik}\}]$$
Finite Sample Description Spaces with Equality Likely Descriptions

We define a sample space $S = \{D_1, D_2, ... D_N\}$ as having equally likely descriptions if all of the single-member events on $S$ have equal probabilities, so that

$$P[\{D_1\}] = P[\{D_2\}] = ... = P[\{D_N\}] = \frac{1}{N}$$

Then we can define the probability of an event as follows

$$P[D] = \frac{N[E]}{N[S]} = \frac{\text{sizeof} E}{\text{sizeof} S}$$