Elementary maths for GMT

Probability and Statistics

Part 1: Descriptive Statistics

Terminology

- Value (observation, measurement): one item of data
- Data (data set): a collection of values
- Sample: a subset of the data

- Measure vs. measurement (as nouns!)
  - When you measure (verb!) something, you get a measurement
  - When you choose a way to quantify something, you are choosing a measure (*e.g.* the body mass index is a measure that indicates that you may be overweight)

Types of statistics

- Descriptive statistics: the statistical procedures for describing and interpreting a data set
- Inferential statistics: the statistical procedures for generalizing to a population from a sample
- Predictive statistics: the statistical procedures for extrapolating data (observations) over time

Applications in computer science

- Some algorithms use random numbers and can be proven to be efficient (on any input) in an expected sense
  - Examples: QuickSort, hashing, smallest enclosing disk, genetic algorithms, optimization of a fitness function …
- To perform data mining, discovering patterns or dependencies of events, *e.g.* in processing of user studies
- To design decision support systems or expert systems
- For all sorts of experiments, *e.g.* in hypothesis testing
  - Hypothesis: The appreciation of this game decreases with the age of the player
  - Descriptive statistic: The mean duration of that adventure game is 4 hours, with a standard deviation of 47 minutes

Value

- Not only numeric values
  - Example: you can observe “land use” at eight locations and get: urban, nature, farmland, urban, urban, orchard, industrial, commercial

    ➔ these are non-numeric values that come in categories

- Most statistics assume numeric values
  - What would be the average of urban, nature and farmland?

Descriptive Measures

- Measures of central tendency
  - Give a center around which the measurements in the data are distributed
  - *E.g.* mean, median, mode
- Measures of variability
  - Describe data spread or how far away the measurements are from the center
  - *E.g.* range, variance, standard deviation
- Measures of relative standing
  - Describe the relative position of specific measurements in the data
  - *E.g.* percentile, standard score

Measures of Central Tendency

- Mode
  - the most frequent measurement(s) in the data
- Median
  - the measurement such that at most half of the measurements are below it and at most half of the measurements are above it
- Mean
  - the sum of all measurements divided by the number of measurements

Mode

- Here mode = 3

- The mode is the only measure of central tendency for non-numeric observations (land-use, color, ...)

Measurements
3
5
1
1
4
7
3
8
3

Mode

- It is possible for a data set to have more than one mode

- In this case the data has two modes: 5 and 7, both measurements occurring twice

Measurements
3
5
5
1
7
2
6
7
0
4

Median

- Value such that at most 50% of the values is smaller and 50% of the values is larger
- Another name for 50th percentile
- Appropriate for describing measurement data
- Robust to outliers, that is, not affected much by unusual values like outliers or gross measurement errors

Median

- Median: any number in the interval $[4, 5]$
- Note that only the two central values are used in the computation
- The median is not sensitive to extreme values

- Q: Can the median be an interval of values if the number of values is odd?

Measurements	Measurements ranked
3	0
5	1
5	2
1	3
7	4
2	5
6	5
7	6
0	7
4	7

Mean

- Another name for average
- When describing a population it is denoted μ, the Greek letter "mu"
- When describing a sample it is denoted $\bar{X}$, pronounced "x-bar"

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

- Appropriate for describing measurement data
- Seriously affected by extreme values

Mean

- Mean = 40/10 = 4
- Notice that the sum of the ‘deviations’ is 0
- Every single value contributes to the computation of the mean

Measurements	Deviation
3	-1
5	1
5	1
1	-3
7	3
2	-2
6	2
7	3
0	-4
4	0

Measures of Variability

- Range
  - Difference between the largest and smallest value
- Variance
  - Average squared difference of the values from the mean, denoted by σ^2 (pronounced sigma-squared)
- Standard deviation
  - Square root of the variance, denoted by σ

Range

- Largest value = 7
- Smallest value = 1
- Range = 7 - 1 = 6

Measurements
3
5
5
1
7
2
6
7
1
4

Variance

- Average squared difference of the values from the mean

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

Variance

- Variance = 54/10 = 5.4
- Measure of spread
- The larger the deviations (positive or negative), the larger the variance

Measurements	Deviation	Square of deviation
3	-1	1
5	1	1
5	1	1
1	-3	9
7	3	9
2	-2	4
6	2	4
7	3	9
0	-4	16
4	0	0
40	**0**	**54**

Standard Deviation

- Square root of the variance

$$\sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}} = \sigma$$

- In the previous example, variance $\sigma^2 = 5.4$ and standard deviation $\sigma = \sqrt{5.4} = 2.32$

Standard Deviation

- It is the typical (standard) difference (deviation) of a value from the mean
- The standard deviation has the same units as the values themselves (unlike variance)
  - *E.g.* if the values are weight measurements in kg, then the standard deviation is also in kg, while the variance is in kg^2

Tchebichev’s Rule

- At least 75% of the measurements differ from the mean less than *twice* the standard deviation
- At least 89% of the measurements differ from the mean less than *three* times the standard deviation
- This is Tchebichev’s Rule: At least $1 - 1/k^2$ of the observations falls within k standard deviations from the mean
- True for every data set

Tchebichev's Rule

- Suppose that for certain data
  - mean = 20
  - standard deviation = 3
- Then
  - at least 75% of the measurements are between 14 and 26
  - at least 89% of the measurements are between 11 and 29

Further Notes

- When the mean is greater than the median, the data distribution is skewed to the right (positive skew)
- When the median is greater than the mean, the data distribution is skewed to the left (negative skew)
- When mean and median are very close to each other, the data distribution is approximately symmetric

- Q: Which case applies with the following set?
  {2, 3, 5, 6, 9, 9, 10, 10}

Measures of Relative Standing

- The place of a particular measurement in a population
- Standard score
  - Difference between the observation and the mean normalized by the standard deviation
- Percentile
  - divides the data or population into two parts: % before and % after the sample

Standard Score

- Standardized score relative to the position and spread of the sample

$$\frac{x - \mu}{\sigma}$$

- This is useful to normalize the sample values according to the distribution

Percentile

- The *p*-th percentile is a number such that at most *p* percentile of the measurements are smaller and at most 100 – *p* percentile of the data are larger
  - Example: in a certain data set the 85th percentile is 340
    - At most 15% of the measurements are **>** 340
    - At most 85% of the measurements are **<** 340
- Note that the median is the 50th percentile

Probability

- Probability is a numerical measure of the likelihood that a specific event will occur
- If there are *n* equally likely outcomes (events) and *s* are favorable ("success") then the probability of a success is s/n

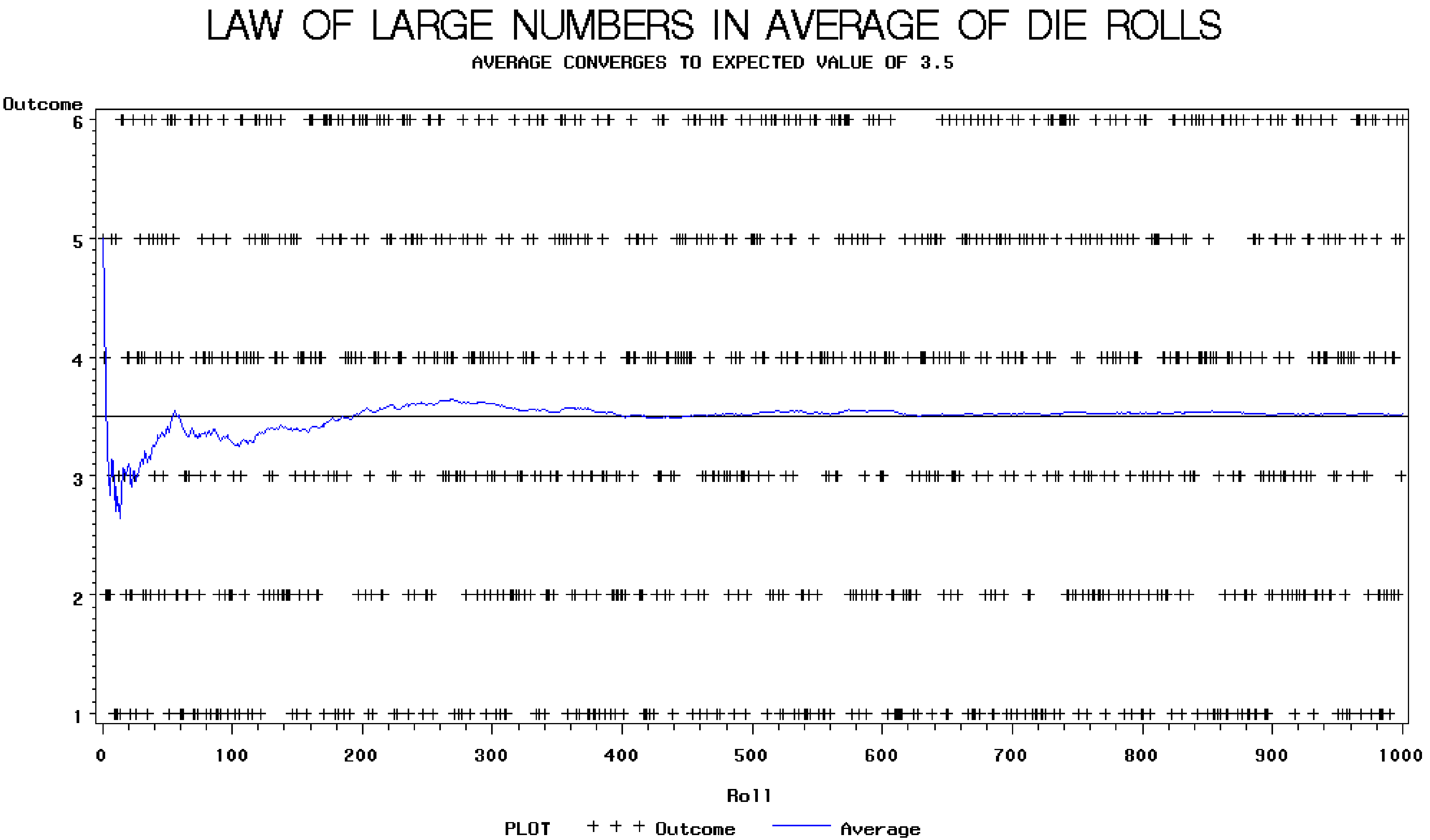
Frequency Interpretation of Probability

- The probability of an event is the proportion of the time that events of the same kind will occur in the long run
- If an experiment is repeated n times and an event A is observed f times, then, according to the relative frequency concept of probability

approximate probability: $P(A) = \frac{f}{n}$

Law of Large Numbers

- The average of the results obtained from a large number of trials should be close to the expected value, and will tend to get closer as more trials are performed


LAW OF LARGE NUMBERS IN AVERAGE OF DIE ROLLS
AVERAGE CONVERGES TO EXPECTED VALUE OF 3.5
Outcome
6
5
4
3
2
1
0
100
200
300
400
500
600
700
800
900
1000
Roll
PLOT
Outcome
Average

Law of Large Numbers

- The LLN was proven by Bernouilli
  - Bernouilli trial: single experiment with two possible outcomes: success or failure
- The LLN is important because it ‘guarantees’ stable long-term results for random events
- Crucial if you want to run a casino

Expectation

- General definition:

  If the probability of obtaining amounts

  $a_1, a_2, \dots, a_k$ are (respectively)

  $p_1, p_2, \dots, p_k$

  then the expectation (expected amount obtained) is

$$E = a_1 p_1 + a_2 p_2 + \cdots + a_k p_k = \sum_{i=1}^{k} a_i p_i$$

Expectation: example

- What is the mathematical expectation if we win € 6 when a die comes up 1 or 2, and lose € 3 when the die comes up 3, 4, 5, or 6?
- Solution
  - Amounts: $a_1 = 6$ and $a_2 = -3$
  - Assuming a balanced die, randomly rolled, the probabilities of rolling the values are

    $p_1 = 2/6 = 1/3$ and $p_2 = 4/6 = 2/3$
  - So the mathematical expectation is

$$E = 6 \times \frac{1}{3} + (-3) \times \frac{2}{3} = 0$$

Variations and combinations

- Basic experiment of selecting k items from a pool of n different items

- Variations: ordering is important
- Combinations: ordering is not important

A few basic experiments

- Variations
  - selecting samples without replacement, ordering is important
- Repeating variations
  - selecting samples with replacement, ordering is important
- Combinations
  - selecting samples without replacement, ordering is not important
- Repeating combinations
  - selecting samples with replacement, ordering is not important

Example 1: Variations

- Number of possibilities: $\frac{n!}{(n-k)!}$

- Example: Jack, Joe, Jill, and Jennifer organize a meeting and they rent a room with 25 seats. They can sit anywhere they want. How many possibilities are there?
  - Answer: 25! / (25 – 4)! = 25*24*23*22 = 303,600
    - Replacement: A seat can contain just one person, so no
    - Order: It matters who sits where, so yes

- If $k = n$, the variation becomes a permutation: n!
  - For example the number of outcomes when shuffling a deck of cards

Example 2: Repeating Variations

- Number of possibilities: n^k
- Example: Suppose a sushi restaurant has conveyor belts with little plates that contain pieces of sushi. You can take a plate, eat the sushi, and take the next plate. There are 12 different types of sushi. For € 15 you can take 8 plates. How many different menus can we create?
  - Answer: 12^8, so 429,981,696 possible menus
    - Replacement: We can choose a sushi plate that we have chosen before, so yes
    - Order: We eat the sushi in sequence, so yes
(tuna maki, salmon sashimi ≠ salmon sashimi, tuna maki)

Example 3: Combinations

- Select k items from a set of n items, final ordering is not important, number of combinations:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

- Example: Suppose a tapas place has 20 items and we choose 5 different ones. How many possibilities are there?
  - Answer: 20! / (5! · (20 – 5)!) = 15,504
    - Replacement: We choose different tapas, so no
    - Order: The 5 tapas are chosen together, so no

Example 4: Repeating Combinations

- Number of possibilities: $\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!(n-1)!}$ when we select k samples from a population of size n

- Example: Suppose we have 5 eggs that each needs to be colored either red, blue or yellow. How many different colorings of the 5 eggs are there?
  - Answer: (3+5-1)! / (3!(5-1)!) = 35
    - Replacement: a color can be used again, so yes
    - Order: it does not matter in which order we color, so no

Seating of eggs

- Instead of Jack, Joe, Jill, and Jennifer, we place 4 uncolored eggs on the 25 seats of the meeting room, but we do not place more than one egg on any chair

- Question: What type of experiment is this?

- Question: How many possibilities are there?

Seating of eggs

- Now it does not matter which egg sits where, so order is not important
- Question: What type of experiment is this?

Combination

- Question: How many possibilities are there?

25! / (4! (25-4)!)

Sorting Experiment

- Suppose we wish to analyze an implementation of a sorting algorithm on small arrays. We choose to analyze it on arrays of length 30, and use a random number generator to put the 30 values in a random order. We run 10 such experiments.
- Question: What type of experiment is this?
- Question: How many possibilities are there?

Sorting Experiment

- There are 30! different arrays (random orders) that can be generated using the random number generator
- The 10 experiments may use the same or different arrays
- The 10 experiments will be performed in a sequence, but the order is irrelevant for the experiment
- So: Repeating combination with $n = 30!$ and $k = 10$

$$\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!\,(n-1)!}$$

Conditional Probability

- Probability of event A given that event B has already occurred, or will occur
- Notation: $P(A|B)$
  - Read: P of A given B

- Example 1: A = it will rain tomorrow
  B = today is October 9
  $P(A|B) = ?$
- Example 2: A = a die roll gives a 3
  B = a die roll gives an odd number
  $P(A) = 1/6$, while $P(A|B) = 1/3$

Independence

- A is independent of B if $P(A) = P(A|B)$
- In such a case $P(A \wedge B) = P(A) \times P(B)$

  and $P(A \vee B) = P(A) + P(B)$
- Otherwise

$$
\begin{aligned}
P(A \vee B) &= 1 - P(\neg A \wedge \neg B) \\
&= 1 - (1 - P(A)) \times (1 - P(B)) \\
&= P(A) + P(B) - P(A) \times P(B)
\end{aligned}
$$

and ...

Bayes' Theorem / Rule

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \qquad P(B \mid A) = \frac{P(B \wedge A)}{P(A)}$$

$$P(A \wedge B) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

$$\Rightarrow \quad P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$