ALGONQUIN COLLEGE

CST8502
MACHINE LEARNING

Statistics (Self-study Material)

Professor: Dr. Anu Thomas
Email: thomasa@algonquincollege.com

Random Number Generators

- The numbers seem random but they're not. They're pseudo-random
- The sequence of numbers generated depends on the starting "seed"
- ***Two number generators will produce the same numbers if they have the same seed.***
- Create two Random Objects with the constructor: Random(int seed)
- call nextInt() on both. They will produce the same sequence.
- To create a new sequence every time, use: System.currentTimeMillis() as the seed

Statistics

- Given a set of data (Numbers), there are several things we can compute:
- Mean: What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$
- Median: What is the middle item? Array[size/2]
- Mode: What item appears the most often:
- 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5 ?
- Order Statistics: What is the 3rd largest number? What is (n-2)nd largest number?

Moving Mean/Moving Average

- Moving average: What is the average of values in the last 5 days?
- If N is the total sum, and the day changes, you don’t need to recompute the mean. Instead, add the new values, subtract the expired values, and divide by the new number of items, N.

Computing Moving Mean

- If the mean, N, represents the average of 5 days:

  [1, 2],[3, 4], [4, 6], [5, 7], [9, 11]

- $N = (1 + 2 + 3 + 4 + 4 + 6 + 5 + 7 + 9 + 11)/10 = 5.2$
- Now new data are ready to be added: 10, 12
- Imagine that [1, 2] are now out of date and no longer part of your computation. Just add the new data, subtract the old data, and divide by the number of data:
- $N' = N + [(10+12) - (1+2)]/10$

  $= 5.2 + 1.9$

  $= 7.1$

- Verify: $(3 + 4 + 4 + 6 + 5 + 7 + 9 + 11 + 10 + 12)/10 = 7.1$

Weighted Average

- Suppose you want more recent data to be worth more than older data, like for predicting gas prices.
- Decide the comparative weights of the components:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}})$$

- Now divide by the weighted number of elements:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}}) \\ / (5 + 3 + 2 + 1)$$

Sample vs Population

- A population data set contains all members of a specified group (the entire list of possible data values)
  - Example: all people in Ottawa

- A sample data set contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken.
  - Example: some people in Ottawa

Standard Deviation

- Let X be an array: X = {21, 37, 13, 25, 32, 8 }
- What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$ = 22.6667
- **Standard deviation**: N is the number of elements
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/N}$. This formula is used when you have measured the entire population. In Excel, this is stdev.p() = 10.07748
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/(N-1)}$. This formula is used when you have only part of the data. In Excel, this is stdev.s() = 11.03932

Standard Deviation

	x	x-Mean	$(x\text{-Mean})^2$
	21	-1.6667	2.7778
	37	14.3333	205.4444
	13	-9.6667	93.4444
	25	2.3333	5.4444
	32	9.3333	87.1111
	8	-14.6667	215.1111
Mean	22.6667		

Sum of $(x\text{-Mean})^2$	609.3333
Sum / 6	101.5556
sqrt (sum/6)	10.0775

Mean and Variance

- Difference of means is the difference between the averages of two samples
- Variance is just standard deviation squared.

- https://www.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/v/range-variance-and-standard-deviation-as-measures-of-dispersion

- https://www.khanacademy.org/math/statistics-probability/displaying-describing-data/sample-standard-deviation/v/statistics-sample-variance

Distributions

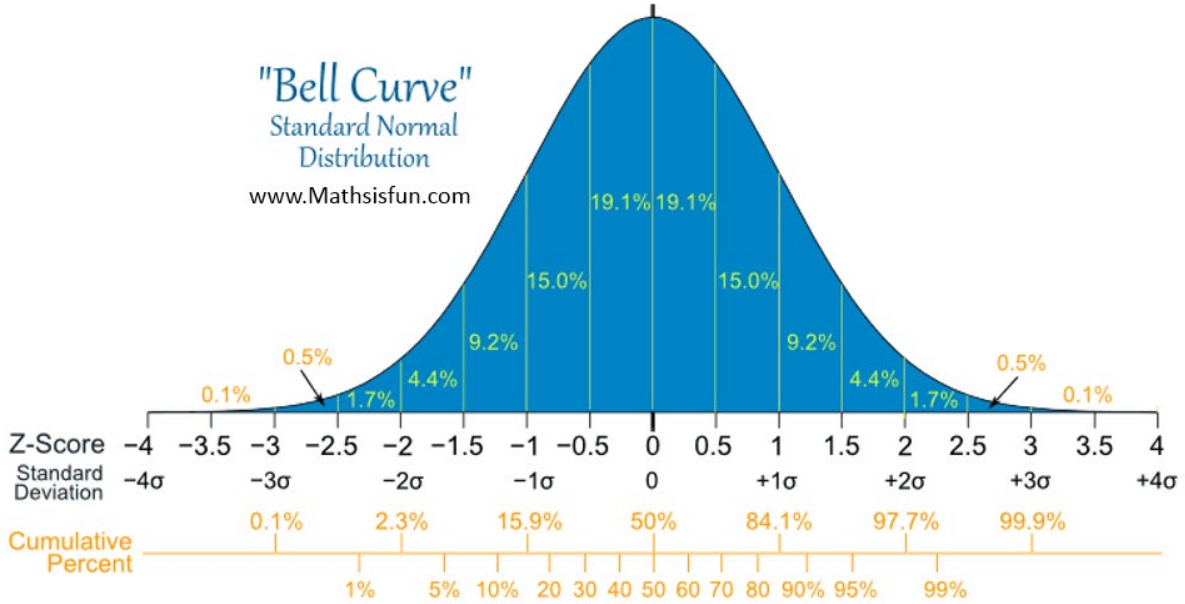
- Uniform – The probability of an event is equal (uniform). The probability of getting tails for flipping a coin, or rolling a 1 with a die.
- Gaussian (Normal) – The values are centered around a midpoint (mean), but decrease as you get farther from the mean: grades on a test.
- Geometric, Poisson, Exponential – These are other distributions that exist, but we don't have time to cover.

Normal / Gaussian

- Also called a bell curve, with midpoint of μ (pronounced me-you), and standard deviation of σ (pronounced sigma).
- The density of an event $(x \mid \mu, \sigma^2)$ is: $\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- The variance is σ^2
- If $\mu = 0$ and $\sigma^2 = 1$ then this is called Standard Normal Distribution

Normal / Gaussian

- The area under the curve must add up to 1. Probabilities are calculated by a number being less than a number.

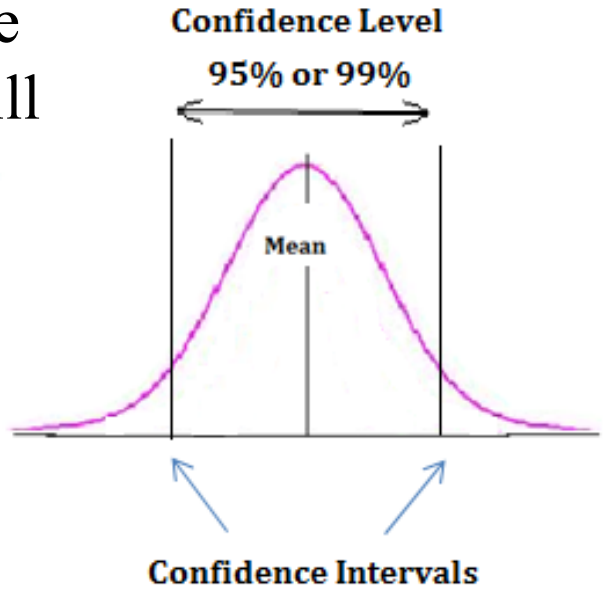


Rank Statistics

- Rank statistics compute where a number compares to the rest of the data, for instance to 5%, bottom 15%, etc.
- They are described in percentiles, meaning how much of the data is less than the number. 95% percentile means that 95% of the numbers are less. The median is the 50% percentile.

Confidence Intervals

- Confidence intervals describe the uncertainty of a parameter. If you repeatedly take a small sample to measure, your values will always be different. The mean you repeatedly sample will also follow a normal distribution.
- The confidence intervals calculates the probability that the actual mean falls close to your measured number.



Confidence Intervals

- The formula for calculating the confidence interval is: $\mu \pm z_c \left(\frac{\sigma}{\sqrt{n}}\right)$
- Z_c is the "critical value" for difference confidence levels: 90% is 1.645, 95% is 1.96, 99% is 2.575.
- This computes the limits for 90, 95 or 99% of the data.

- The 90% confidence interval says there is a 90% chance that the true mean falls within the range you have measured +/- some error
- The 95% confidence interval says there is 95% chance that it falls within a larger range.