WARWICK
THE UNIVERSITY OF WARWICK

MA930 Data Analysis & Machine Learning

Lecture 2: Basic Statistics

Haoran Ni

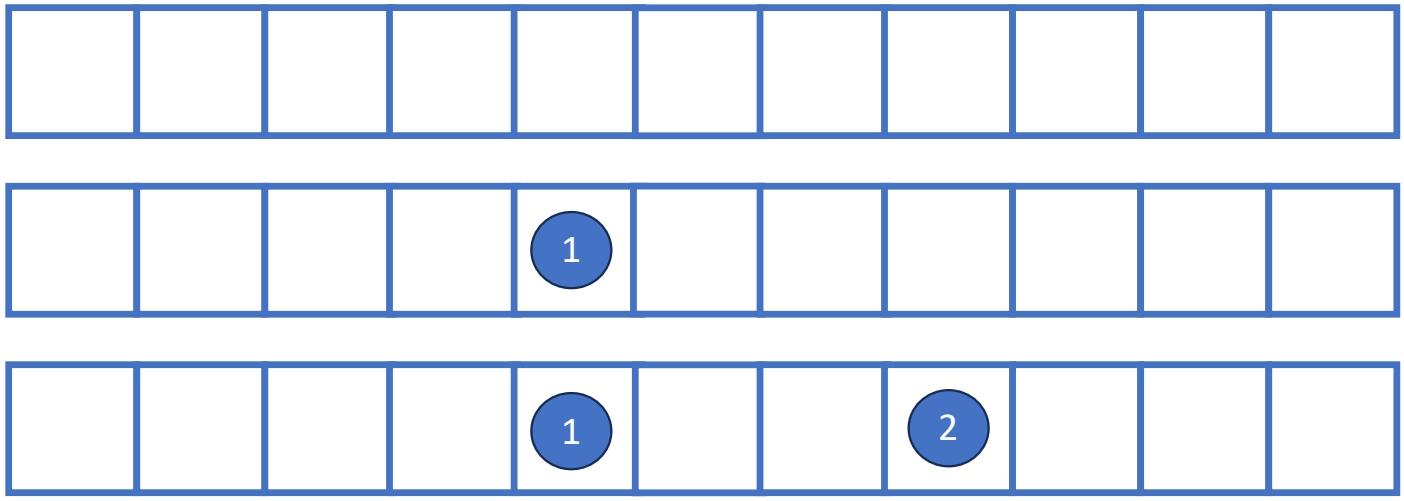
Outline

WARWICK
THE UNIVERSITY OF WARWICK

- Arrangements
- Multivariate distributions
- Sums of random numbers
- Sample statistics
- Distribution of variance
- Additional Questions

Arrangements

- The number of arrangements of n distinguishable objects (permutations) is $n!$
- This can be seen by considering the number of available positions for the first object, then second, etc.



n positions

n -1 positions

n -2 positions

Arrangements

- If m of these objects are the same, and we want to know the number of arrangements, then we note that each of the previous arrangements belongs to a group of size $m!$ that differ only in the arrangements of the identical objects. That means there are now $n!/m!$ arrangements.

- For example, there are e.g. 4!/3! arrangements of (a a a b)
- More generally, there are e.g. 6!/(3!2!) arrangements of (a a a b b c)

Exercise 2.2. Find the number of non-negative integer solutions of the equation

$$x_1 + x_2 + \cdots + x_m = n$$

Outline

WARWICK
THE UNIVERSITY OF WARWICK

- Arrangements
- Multivariate distributions
- Sums of random numbers
- Sample statistics
- Distribution of variance
- Additional Questions

Multinomial distribution

WARWICK
THE UNIVERSITY OF WARWICK

- Outcomes of multiple independent binary Bernoulli events results in a binomial distribution
- Analogously, outcomes of multiple multinoulli events results in a multinomial distribution
- A multinomial distribution is what you get if you throw a *k* sided die *n* times, and record the number of times each of the *k* numbers come up
- Then, $P(x_1, x_2, x_3, \dots, x_k)$ represents the probability of getting x_1 1's, x_2 2's, etc. [where $x_1 + x_2 + x_3 + \cdots + x_k = n$]

Multinomial distribution

In general, if the *k* sides of the die have probabilities $p_1, p_2, p_3, \ldots$ of coming up each time, then

$$P(x_1, x_2, x_3, \ldots, x_k) = \frac{n!}{x_1!\, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

- When $n = 1$, this is a multinoulli distribution
- When $n = 1$ and $k = 2$, this is a Bernoulli distribution

Multinomial distribution

WARWICK
THE UNIVERSITY OF WARWICK

Example 2.1. (board) Awareness of infection history.

In a large population with widespread antibody testing, 20% of individuals know they have been infected, 35% know they were not infected, 45% are unsure (they have not received an antibody test). What is the probability, if six individuals are selected at random from the population, that there will be one person who knows they have been infected, two who know they have not, and three who are unsure?

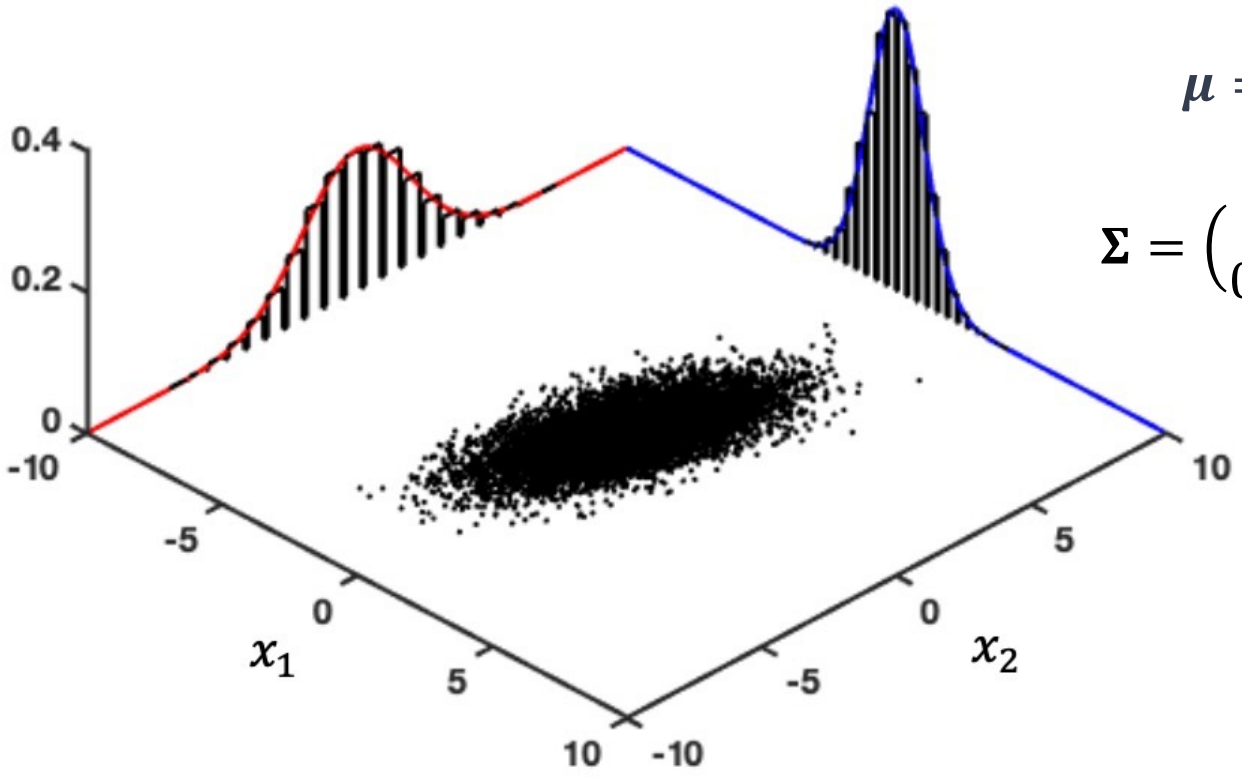
Multivariate normal (Gaussian) distribution

- Distribution for multiple Gaussian variables that allows for correlations

- For n Gaussian variables, where $\boldsymbol{\mu}$ is the vector of means and $\boldsymbol{\Sigma}$ is the covariance matrix, the distribution is

$$P(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(\frac{-(x-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})}{2}\right)$$

Multivariate normal (Gaussian) distribution



$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 4 \end{pmatrix}$$

Outline

WARWICK
THE UNIVERSITY OF WARWICK

- Arrangements
- Multivariate distributions
- Sums of random numbers
- Sample statistics
- Distribution of variance
- Additional Questions

Sums of random numbers

- Consider repeated samples from the same distribution:

$$X_1, X_2, X_3, \ldots, X_n$$

- A ***statistic*** is some function of these measurements

  E.g. The sample mean: $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

- This is a random number with some distribution, for repeated draws

- What are the convergence properties of this quantity when many samples are taken?

Sums of random numbers

- (Weak) Law of large numbers (LLN)

- Given a large enough number of samples, the sample mean converges to the true mean
- $\overline{X_n} \to \mu$

Example 2.2. (board) Use characteristic functions and Maclaurin series to prove the LLN.

Sums of random numbers

- Central limit theorem

Exercise 2.3. In the previous example, instead go to second order in the MacLaurin series. Compare the result to the list of characteristic functions from lecture 1. Infer the distribution.

- The sample mean approaches a normal distribution with variance: $\frac{\sigma^2}{n}$

Sums of random numbers

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 2.4. Consider a gamma distribution with shape and scale parameters 5 and 0.2.

i. What is the mean?

ii. Plot the pdf

iii. Consider a sample of size n for calculating the sample mean. Plot a histogram of sample means for N repeated experiments (try a few values of N)

iv. Plot the normal distribution corresponding to the CLT on the same graphs

Sample statistics

- Consider repeated samples from the same distribution: $X_1, X_2, X_3, \ldots, X_n$
- The sample mean can be used to approximate the true mean

Sample statistics

- Consider repeated samples from the same distribution: $X_1, X_2, X_3, \ldots, X_n$
- The sample mean can be used to approximate the true mean
- The sample mean is an unbiased estimator (it has the correct expected value)
- What about the sample variance?

Sample statistics

- The population variance is $\sigma^2 = \mathbf{E}\left((X-\mu)^{\mathbf{2}}\right) = \mathbf{E}\left(X^{\mathbf{2}}\right) - \mathbf{E}(X)^{\mathbf{2}}$
- The sample variance is $s^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^{\mathbf{2}}$

- Does $\mathbf{E}\left(s^{\mathbf{2}}\right) = \sigma^2$?

WARWICK
THE UNIVERSITY OF WARWICK

Sample statistics

- The population variance is $\sigma^2 = \mathbf{E}\left((X-\mu)^{\mathbf{2}}\right) = \mathbf{E}\left(X^{\mathbf{2}}\right) - \mathbf{E}(X)^{\mathbf{2}}$
- he sample variance is $s^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^{\mathbf{2}}$

- Does $\mathbf{E}\left(s^{\mathbf{2}}\right) = \sigma^2$?

- No! $\mathbf{E}\left(s^{\mathbf{2}}\right) = \frac{n-1}{n}\sigma^2$.

- $s^{\mathbf{2}}$ is a biased estimator of σ^2 (it underestimates the true variance)
- $\frac{n-1}{n}s^2$ is an unbiased estimator of σ^2

(Take care: the literature is a mess, and its not always clear what estimator is being used)

Sample statistics

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 2.5. Generate some uniform random numbers on [0,1].

i. Calculate the sample mean.

ii. Calculate the biased sample variance and the unbiased sample variance, and compare this to the true variance.

iii. Test whether the var command in Matlab or Julia (or equivalent in other languages) uses the biased or unbiased variance.

Distribution of variance

WARWICK
THE UNIVERSITY OF WARWICK

- As previously explained, the sample mean, and unbiased sample variance is $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ and $s_u^2 = \frac{n}{n-1}s^2$
- The sample mean for large numbers of samples approaches normal with mean μ and variance $\frac{\sigma^2}{n}$
- For normal distributions, the distribution of the sample variance can be calculated (Cochran's theorem – beyond the scope of this course)
- Result is: $\frac{s_u^2(n-1)}{\sigma^2}$ follows a χ_{n-1}^2 distribution

Distribution of variance

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 2.6. Sample variance numerics.

Draw n samples and calculate the sample mean and unbiased sample variance for normally distributed random numbers.

Repeat this many times and plot the histograms of the sample means and unbiased sample variances, along with their theoretical distributions.

Outline

- Arrangements
- Multivariate distributions
- Sums of random numbers
- Sample statistics
- Distribution of variance
- Additional Questions

Additional questions

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 2.7. Sample mean of UK house prices in 2020.
Go to https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Scroll to 2020 and open the data file "the complete yearly data as a CSV file (complete)", which contains UK house price data from 2020.
Extract all prices less than or equal to £2M and calculate the mean and standard deviation.
Draw a sample of size n and calculate the sample mean.
Using repeated samples, show (by comparing the empirical and theoretical distributions by eye) that the sample mean follows a normal distribution (for sufficiently large numbers of samples).

Verify that the parameters of this normal distribution are given by the true mean and $\frac{\sigma^2}{n}$

Additional questions

WARWICK
THE UNIVERSITY OF WARWICK

Exercise 2.8 (Optional). Detailed analysis of UK house prices in 2020

Explore the UK house price data and make a note of any interesting patterns you find.

For example, can you find a distribution that fits the distribution of house prices (excluding outliers, e.g. houses > £2M)? Are there any noticeable features in the data?