Introduction to Statistics for Astronomers and Physicists

Learning Objectives

Dr Angus H Wright

2022-03-24

General Statistics Notation

Understand and Explain the meanings of the following terms:

- An **observation**
- A **sample**
- A **population**
- A **variable**
- A **parameter**

Fundamentals of Data Description

Understand the meanings of different types of data:

- **Qualitative**
- **Quantitative**
- **Discrete**
- **Continuous**

Understand the meanings of different data scales:

- **Nominal scale**
- **Ordinal scale**
- **Continuous scale**

Understand concepts and uses of absolute vs relative frequency measures in data.

Understand the various methods of representing one-dimensional data graphically:

- **Bar Charts**
- **Histograms**
- **Kernel Density Estimates**
- **ECDF** and **Quantile Functions**

Be able to identify appropriate graphical representations for a given dataset, based on the datasets type and scale.

Measures of Central Tendency and Dispersion

Understand the motivations and limitations of concepts such as:

- **Arithmetic Mean**
- **Median**
- **Mode**
- **Mean Absolute Deviation**
- **Mean Squared Error**
- **Variance**
- **Standard deviation**

- **Median Absolute Deviation from Median**
- **Interquartile Range**

Understand any **useful properties** of the above.

Understand the differences between various measures of central tendency, and when they are appropriate to use.

Understand the differences between various measures of dispersion, and when they are appropriate to use.

Understand graphical methods of data comparison in one dimension, such as:

- **KDEs**
- **Box-and-whisker plots**
- **Violin plots**
- **Quantile-Quantile plots**

Summarising relationships in 2D

Understand graphical methods of exploring observations in two or more variables, such as:

- **Scatter plots**
- **KDEs**

Understand the concepts of:

- **covariance**
- **correlation**

Be able to describe the construction of a **covariance/correlation matrix**

Understand the differences between **Pearson and Spearman Correlation**, and describe the uses of each.

Understand the limitations of correlation measures, and the logical fallacy of correlation and causation.

Understand the concept of **confounding variables** and their role in correlation.

Probability

Understand the concepts of:

- **Outcomes**
- **Events**
- **Sample Space**
- **Trials**
- **Disjoint Outcomes**
- **Independence**

Understand how probability derives from the concept of relative frequency, and be able to explain this derivation with simple examples.

Understand the fundamental properties of probability.

With regards to probabilities and events, be able to explain concepts such as:

- **Independent Probabilities**
- **Joint Probabilities**
- **Conditional Probabilities**

Understand the mathematical definitions of the above.

Understand visualisations of probability, such as:

- **Urn models**
- **Probability Trees**
- **Venn Diagrams**

Be able to use the above tools to describe and analyse simple thought experiments.

Be able to use reasoning and probability to explore simple probabilistic problems.

Understand the difficulties of reasoning with conditional probability, and be able to identify flawed reasoning.

Probabilistic Distributions

Be able to identify the following probabilistic concepts/functions/distributions:

- **Bernoulli Random Variables**
- **Binomial Distribution**
- **Poisson Distribution**
- **Uniform Distribution**
- **Gaussian Distribution**
- **Beta Distribution**

Understand the distinction between a Probability Mass Function and a Probability Density Function.

Understand the term **Expectation** and the **Expectation Value**.

Understand the parameters that describe important probability distributions, such as those above.

NB: this section **does not** require students to memorize the exact forms/equations of probability distributions. However, students **should** be able to recall details of specific probability distributions. Such as:

- The parameters that govern the form of the distribution, and their interpretation;
- The fundamental properties of the distributions (expectation and variance).

Sampling and Simulation

Understand the distinction between psuedo-random and truly random numbers.

Understand the meaning of **independent and identically distributed** observations.

Describe example uses of **Simulation**.

Describe various methods of sampling, including their benefits and limitations.

Understand and demonstrate **Monte Carlo integration**.

Bayesian Statistics

Understand the concepts of:

- The **Likelihood**
- The **Prior**
- The **Posterior**

Be able to write Bayes Theorem, and understand its use.

Understand the primary differences between **classical statistics** and **bayesian statistics**.

Understand the concept of **maximum likelihood** and its uses.

Understand and demonstrate the difference that **prior belief** has on **posterior distributions**.

Understand and describe the four key steps to the Bayesian approach.

Understand and give examples of **Conjugate Prior** pairs

Describe the behaviour of priors and data in the limits of large/small data samples, and strong/weak prior information.

Understand the concept of **Jeffery's Prior**, including its limitations.

Summarising Posterior Information

Understand the concepts of:

- The **joint distribution**
- The **marginal distribution**
- The **point estimate**
- The **credible interval**
- The **predictive distribution**

Identify examples of **point estimates** and their motivations.

Posterior Simulation

Understand the concepts of:

- The **target distribution**
- The **transition probability**
- **Aperiodicity**
- **Irreducibility**
- The **candidate generation function**
- The **acceptance probability**

Understand the concept of the **Markov Chains** and its construction.

Understand the **Markov Chain Monte Carlo** sampling methodology, and its uses.

Understand and be able to explain the algorithm known as the **Gibbs Sampler**, and its uses.

Understand and be able to explain the **Metropolis-Hastings Algorithm**, and its uses.

Understand the limitations of **MCMC** and the methods that can be used to avoid biased inference.

Classical Significance

Understand the concept of:

- **Hypothesis testing**.
- **Significance**
- **Test statistic**
- **p-value**

Understand various hypothesis tests and their uses, including:

- The **One-Sample t-test**
- The **Two-Sample t-test**
- The **F-test**
- The **KS-test**

Bayesian Hypothesis Testing

Understand the concepts of:

- The **posterior odds ratio**
- The **Bayes Factor**

Understand and explain the **Jeffreys hypothesis testing criterion**, and its benefits over classical significance.

Bayesian Modelling

Understand the concepts of:

- **hierarchical model**
- **hyperparameters**

Be able to describe the difference between hierarchical and non-hierarchical models, giving examples.

Provide a basic example of a Bayesian Hierarchical Model using a **directed acyclic graph**

Demonstrate the benefits of hierarchical models over standard models.

Machine Learning

Understand the concepts of:

- **neurons**

Describe the concepts and uses of **Supervised Learning**.

Describe the concepts and uses of **Unsupervised Learning**.

Describe the concepts and uses of **Reinforcement Learning**.

Describe an **Artificial Neural Network**, its construction, and its uses.

Describe an **Self-Organising Map**, its construction, and its uses.

Understand and be able to identify **biases in machine learning**, including those caused by:

- Extrapolation
- Covariate Shift