Course Notes on Probability and Statistics

Max Kasperowski

January 3, 2018

Contents

1 Miscellaneous Foundations

1.1 Permutations and Combinations

The number of arrangements of n elements equals $n!$.
When taking r elements from a set of n elements, the number arrangements retaining order (permutations) is represented as:

$$P_r^n = \frac{n!}{(n-r)!}$$

When not regarding the order, but merely which elements have been chosen:

$$C_r^n = \frac{n!}{r! \times (n-r)!}$$

1.2 Binomial Formula

The most basic binomial formula is to the second power and goes as follows.

$$(a+b)^2 = a^2 + 2ab + b^2$$

When expanded to the n^{th} power, this is the formula.

$$(a+b)^n = \sum_{k=0}^{n} C_r^n \times a^k \times b^{n-k}$$

1.3 Sequences & Series

A series is the summation of the terms of an infinite sequence.

2 Probability Fundamentals

2.1 Axioms of Probability

The three axioms of probability are:

$$P(S) = 1$$
$$0 \leq P(E) \leq 1$$
$$E_1 \cap E_2 = \Phi \Leftrightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

The following equalities can be derived from these:

$$P(\phi) = 0$$
$$P(\bar{E}) = 1 - P(E)$$
$$E_1 \subseteq E_2 \Leftrightarrow P(E_1) \leq P(E_2)$$

2.2 Addition Rule

The addition rule is used to calculate the probability of the union of two or more events.
For mutually exclusive events:

$$\text{If } E_1 \cup E_2 = \phi, \text{ then } P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

General rule for two events:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Expanded for three events:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P((A \cap B) \cup (A \cap C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C) \end{aligned}$$

2.3 Conditional Probability

The conditional probability represents the probability of an event A given that an event B has happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0$$

2.4 Multiplication Rule

From the definition of the conditional probability we can get the multiplication rule for calculating the probability of two or more intersecting events.

$$P(A \cap B) = P(A|B) \times P(B)$$

For the intersection of three events we can expand the formula as follows:

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C) \times P(B \cap C) \\ &= P(A|B \cap C) \times P(B|C) \times P(C) \end{aligned}$$

2.5 Total Probability Rule

For an event A divided by the event B and its complement:

$$P(A) = P(A|B) \times P(B) + P(A|\bar{B}) \times P(\bar{B})$$

For an event A divided by a group of events $B_1, B_2 \ldots B_n$,
where $B_1 \cap B_2 \cap \ldots \cap B_n = \phi$:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \times P(B_i)$$

2.6 Independence

Two events A and B are independent if and only if their joint probability equals the product of their probabilities.

$$A \perp B \Leftrightarrow P(A|B) = P(A) \times P(B)$$

This also implies that $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

2.7 Bayes' Theorem

Bayes' Theorem comes from the definition of conditional probability. It can be used to calculate $P(A|B)$, when $P(B|A)$ is known.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

We can use the rule of total probability to determine $P(B)$. This is useful when there are two possible outcomes which affect B.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\bar{A}) \times P(\bar{A})}$$

The extended form for multiple mutually exclusive events, spanning the sample space is:

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{\sum_j P(B|A_j) \times P(A_j)}$$

3 Discrete Random Variables

3.1 Probability Mass Function

The probability mass function of a random variable X is function, that given x yields the probability of $P(X = x)$. More generally, a probability mass function for the discrete random variable X with the values $x_1, x_2, \ldots, x_n$, fulfils the following properties:

$$f(x_i) \geq 0$$

$$\sum_{i=1}^{n} f(x_i) = 1$$

$$f(x_i) = P(X = x_i)$$

3.2 Cumulative Distribution Function

The cumulative distribution function $F(x)$ for a discrete random variable X is the the sum of probability mass functions for $x_i \leq x$ and the probability for $X \leq x$. It has the following properties:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$0 \leq F(x) \leq 1$$

$$x \leq y \rightarrow F(x) \leq F(y)$$

Relations of Probability Mass Functions and Cumulative Distribution Functions Solutions according to the definition: $F(x) = P(X \leq x)$

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) \end{aligned}$$

$$\begin{aligned} P(a < X \leq b) &= P(X < b) - P(X < a) \\ &= F^{-}(b) - F^{-}(a) \end{aligned}$$

$$\begin{aligned} P(X > a) &= 1 - P(X \leq a) \\ &= 1 - F(a) \end{aligned}$$

3.3 Mean and Variance

The mean of a discrete random variable X, is the expected value of X and is represented by μ or $E(X)$. It is the sum of the possible values for X weighted their probability $P(X = x)$.

$$\mu = E(X) = \sum_{x} x \times f(x)$$

The variance provides a measure of the dispersion of the possible values of X. It is represented by σ^2 or $V(X)$.

$$\begin{aligned}\sigma^2 = V(X) = E(X-\mu)^2 &= \sum_x (x-\mu)^2 \times f(x) \\ &= \sum_x x^2 \times f(x) - \mu^2 \\ &= E(X^2) - (E(X))^2\end{aligned}$$

The standard deviation is the square root of the variance.

$$\sigma = \sqrt{\sigma^2}$$

3.3.1 The Linear Operation $E(X)$

Given the random variable X and the real numbers a, b.

$$E(a) = a$$

$$E(a \times X + b) = a \times E(X) + b$$

$$E(X - E(X)) = E(X) - E(EX) = 0$$

Due to this property, $X - E(X)$, is known as a centred random variable.

Properties of the variance derived from the properties of the mean and the definition $V(X) = E(X - E(X))^2$.

$$V(a) = 0$$

$$V(a \times X) = a^2 \times V(X)$$

$$V(X - a) = V(X)$$

These properties can be used to show, that

$$V\left(\frac{X - E(X)}{\sqrt{V(X)}}\right) = 1$$

which gives us the standard random variable X^*.

$$X^* = \frac{X - E(X)}{\sqrt{V(X)}}$$

$$E(X^*) = 0$$

$$V(X^*) = 1$$

3.4 Distributions

3.4.1 Discrete Uniform Distribution

The simplest form of distribution, where all values for X in range $x_1, x_2, \ldots, x_n$ have an equal probability. The probability mass function for this kind of distribution is:

$$f(x_i) = \frac{1}{n}$$

3.4.2 Binomial Distribution

A chain of events, known as a *Bernoulli Trial*, where the probability doesn't change and where the outcome is always either the event or its complement has a binomial distribution, when X represents the number successful (or unsuccessful outcomes). In this case the probability mass function is made up of a combination of elements multiplied by p taken to the power of the number of successful outcomes and the complement of p to the power of unsuccessful events.

For $x = 0, 1, 2, \ldots, n$

$$f(x_i) = C^n_{x_i} \times p^{x_i} \times (1-p)^{n-x_i}$$

It can also be expressed more simply as:

$$X \sim B(n, p)$$

Mean and Variance of a binomial distribution

$$\mu = E(X) = n \times p$$

$$\sigma^2 = V(X) = n \times p \times (1-p)$$

3.4.3 Geometric Distribution

A series of *Bernoulli Trials* where only the k^{th} trial has a positive outcome and all previous trials do not. So X represents the first time an event occurs.

For $x = 1, 2, \ldots, n$

$$f(x_i) = (1-p)^{x_i - 1} \times p$$

$$X \sim G(p)$$

Expected value of a geometric distribution

$$\mu = E(X) = \frac{1}{p}$$

No memory property Whenever a success happens it doesn't affect future successes, the counter "resets" to zero and there is no memory of the first success.

$$P(X \geq k_1 + k_2 | X \geq k_1) = P(X \geq k_2)$$

3.4.4 Hypergeometric Distribution

Given N elements, in which there are contained K elements, which are considered successes. If we take n elements, what is the probability, that we take x elements from K. This is represented by a hypergeometric distribution.

$$f(x_i) = \frac{\binom{K}{x} \times \binom{N-K}{n-x}}{\binom{N}{n}}$$

$$X \sim H(N, K, n)$$

Expected value and approximation as binomial distribution

$$E(X) = \frac{n \times K}{N}$$

As N and K approach infinity the ratio of the two changes less and less when we take elements away, which lets us use a binomial distribution as an approximation when dealing with large sets of elements.

3.4.5 Poisson Distribution

An interval is divided into small intervals, so that each interval can contain only a single success and the probability of a success is equal for all intervals.

$$f(x) = \frac{e^{-\lambda} \times \lambda^x}{x!}$$

$$X \sim P(\lambda)$$

The mean $E(X)$ and variance $V(X)$ are both equal to the frequency λ.

3.5 Approximations

Given distributions with certain properties, the following approximations can be applied.

If N and K are large

$$X \sim H(N, K, n) \to X \sim B\left(n, \frac{K}{N}\right)$$

If n is large and p is small

$$X \sim B(n, p) \to X \sim P(np)$$

4 Continuous Random Variables

4.1 Probability Density Function

$$f(x_i) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\mathrm{d}x$$

Due to the properties of integration, the following is true.

$$P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$$

4.2 Cumulative Distribution Function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

4.3 Mean and Variance

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$$

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x - \mu^2$$

4.4 Distributions

4.4.1 Continuous Uniform Distribution

Pretty much the same as for discrete variables only that we have a continuous distribution.

$$X \sim U(a, b)$$

$$a \leq x \leq b$$

$$f(x) = \frac{1}{(b-a)}$$

$$E(X) = \frac{(a+b)}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

4.4.2 Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

$$-\infty \leq x \leq \infty$$

$$f(x) = \frac{1}{\sqrt{2\pi} \times \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu$$

$$V(X) = \sigma^2$$

Standard Normal Random Variable The standard normal random variable Z has a distribution with a mean of 0 and a variance of 1. This lets us use a reference table to find the probability.

$$Z \sim N(0, 1)$$

$$\Phi(z) = P(Z \leq z)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Standardising can be used to determine the probabilities of non-standard random variables that have a normal distribution.
Given the random variable $X \sim N(\mu, \sigma^2)$. We create the random variable $Y = \frac{X-\mu}{\sigma}$, which is a standard normal random variable. $Y \sim N(0, 1)$

4.4.3 Approximations for Binomial and Poisson Distributions to Normal Distributions

Given $X \sim B(n,p)$, $Z = \frac{X-np}{\sqrt{np(1-p)}}$ is approximately a standard normal variable. This is good for $np > 5$ and $n(1-p) > 5$. Then we can approximate X as $X \sim N(np, np(1-p))$. Given $X \sim P(\lambda)$, $Z = \frac{X-\lambda}{\sqrt{\lambda}}$ is approximately a standard normal variable. This is good for $\lambda > 5$. Then we can approximate X as $X \sim N(\lambda, \lambda)$.

4.4.4 Exponential Distribution

An exponential distribution describes the probability of the duration of time until an event, where λ is the frequency of the event in some time frame.

$$X \sim Exp(\lambda)$$

$$0 \leq x \leq \infty$$

$$f(x) = \lambda e^{-\lambda x}$$

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

Lack of memory property

$$P(X > t_1 + t_2 | X > t_1) = P(X > t_2)$$

4.4.5 Erlang Distribution

An Erlang distribution is series of Poisson events, with an Exponential distribution. Simply it the probability of r counts of an exponential distribution.

$$x > 0 \, and \, r = 1, 2, \ldots$$

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}$$

$$E(X) = \frac{r}{\lambda}$$

$$V(X) = \frac{r}{\lambda^2}$$

5 Joint Probability

Covers the probability of joint events. It is composed of the marginal probabilities. Given the events X_1, X_2, the marginal probabilities are $f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_1$ and $f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{d}x_2$. If $X_1 \perp X_2$ $f(x_1, x_2) = f_1 f_2$ else $f(x_1, x_2) = \iint\limits_{\sigma} f_1 f_2 \sigma$.

Covariance

$$cov(X_1, X_2) = E((X_1 - E(X_2)) \times (X_2 - E(X_1)))$$

Correlation

$$\rho = \frac{cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

6 Statistics

6.1 Random Sampling and Data Description

Sample Mean

$$\bar{X} = \frac{\sum x_i}{n}$$

Sample Variance

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Sample range

$$r = max(x_i) - min(x_i)$$

Population Totality of observation with which we are concerned. X is the population. Analogous to a random variable in probability.

Sample Subset of observations selected from the population. $X_1, X_2, \ldots, X_n$ are a random sample of size n if all X_i are independent and they have the same probability distribution. That means they are independent identically distributed (i.i.d.).

Statistic Any function of observations in a random sample. For example $\bar{X}$, S^2, $min(X)$ and $max(X)$.

6.2 Point Estimation of Parameters

A point estimator $\hat{\Theta}$ is a statistic of a random sample used to estimate the parameter θ. $\hat{\theta}$ is then called the point estimate and is a value.

Unbiased Estimator An estimator is unbiased if $E(\hat{\Theta}) = \theta$. If this does not hold true, then it is a biased estimator and $E(\hat{\Theta}) - \theta$ is the bias of $\hat{\Theta}$.

Minimum Variance Unbiased Estimator The MVUE is the unbiased estimator with the smallest variance. This is the best kind of unbiased estimator. One example is the sample mean $\bar{X}$. It is the MVUE for μ.

Standard Error Similar to the standard deviation. It is a measure of how accurate an estimator $\hat{\Theta}$ is.

$$\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$$

Mean Square Error To measure the suitability of a biased estimator we can use the MSE. We can also use it to compare biased and unbiased estimators.

$$\begin{aligned} MSE(\hat{\Theta}) &= E(\hat{\Theta} - \theta)^2 \\ &= E\left[\hat{\Theta} - E(\hat{\Theta})\right]^2 + E\left[\theta - E(\hat{\Theta})\right]^2 \\ &= V(\hat{\Theta}) + (\text{bias})^2 \end{aligned}$$

6.2.1 Method of Moments

k is the order of the moment. For the calculations we want k to be as small as possible.

Population Moment

$$E(X^k)$$

Sample Moment

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k$$

To find the point estimator, we make the population moment equal the sample moment and solve for the parameter whose estimator we are trying to determine.

6.2.2 Method of Maximum Likelihood

The likelihood function gives the probability of a parameter given the sample. So when the probability is high the parameter is more likely.

Likelihood Function

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Solving the Likelihood Function $X \sim Exp(\lambda)$ so the probability mass function is $f(x|\lambda) = \lambda e^{-\lambda x}$

$$\begin{aligned} L(\lambda|x_1 \dots x_n) &= \prod \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda(\sum x_i)} \\ \ln\left(L(\lambda)\right) &= n \ln \lambda - \lambda \left(\sum x_i\right) \\ \frac{\mathrm{d} \ln L(\lambda)}{\mathrm{d}\lambda} &= n\frac{n}{\lambda} - \sum x_i = 0 \\ \hat{\lambda}_{MLE} &= \frac{1}{\bar{x}} \end{aligned}$$

To solve the likelihood function for the optimum estimator means finding the maximum of it. So typically the procedure involves taking the natural logarithm to eliminate exponents and then finding the derivative and solving for 0. If there are are more than one parameter, it is necessary to do partial derivation to find the maxima for all parameters.