Paper Code

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

Integrated Computer Science Programme **Hilary Term 2017**
Year 3

ST3009: Statistical Methods for Computer Science

DD MMM YYYY **Venue** **00.00 – 00.00**

Doug Leith

Instructions to Candidates:

Attempt **all** questions.

You may not start this examination until you are instructed to do so by the invigilator.

Materials Permitted for this examination:

Non-programmable calculators are permitted for this examination – please indicate the make and model of your calculator on each answer book used.

1. (i) A bag contains 10 balls, of which 5 are red and the other 5 black.

   1. Suppose you take out 5 balls from this bag, with replacement. What is the probability that among the 5 balls in this sample exactly 2 are red and 3 are black? [5 marks]

Solution

$\binom{5}{2}$(5/10)^2(5/10)^3

(b) Now suppose that the balls are taken out of the bag without replacement. What is the probability that out of 5 balls exactly 2 are red and 3 are black? [10 marks]

Solution

We can take 5 balls out of 10 in $\binom{10}{5}$ ways. Picking 2 red balls can be done in $\binom{5}{2}$ ways and picking 3 black balls can be done in $\binom{5}{3}$ ways. So the probability is $\frac{\binom{5}{2}\binom{5}{3}}{\binom{10}{5}}$.

(iii) Three people get into an elevator at the ground floor of a hotel which has four upper floors. Assuming each person gets off at a floor independently and is equally likely to choose each of these four floors, what is the probability that no two people get off at the same floor? [10 marks]

Solution

The first person has 4 floors to choose from, the second person has 3 floors to choose from and so on. So the number of combinations is 4.3.2. The total number of ways to for 3 people to choose from 4 floors is 4^3. So the probability is $4.3.2/4^3$

2. (i) Define the terms "random event" and "random variable" and give an example of each. [5 marks]

Solution

A random event is a subset of the sample space. A random variable maps from random events to a real number.

(ii) For a random variable X, define E[X] and var(X). [5 marks]

Solution

Assuming X takes values x_1, x_2,…,x_n:

$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$, $Var(x) = E[X^2] - E[X]^2$

(iii) A random variable X has P(X=1)=0.2, P(X=2)=0.3, P(X=3)=0.5 and P(X=x)=0 for all values of x other than 1,2 or 3. What is the mean and variance of X? [5 marks]

Solution

Mean is 1x0.2+2x0.3+3x0.5=2.3, Variance = 1^2x0.2+2^2x0.3+3^2x0.5-2.3^2=0.61

(iv) Define what it means for two random variables to be independent. [5 marks]

Solution

Random variables X and Y are independent if

P(X=x and Y=y)=P(X=x)P(Y=y)

holds for all values x and y that the two RVs can take.

(v) Let X and Y be independent random variables that take values in the set {1,2,3}. Assume that X and Y are uniformly distributed on {1, 2, 3} i.e. the probability of each value occurring is the same. Let V = XY. Are V and X independent? Explain. [5 marks]

Solution

They are not independent. To verify this, consider for example P(V=1 and X=2). P(V=1)=P(X=1 and Y=1)=(1/3)(1/3). P(X=2)=1/3. P(V=1 and X=2)=0 since there is no value of Y for which V=XY=1 when X=2.

3. (i) Write down expressions for E[X] and E[X/n] for random variable X and n≠0. Show that E[X/n]=E[X]/n. [5 marks]

Solution

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)\,, \qquad E[X/n] = \sum_{i=1}^{n} \frac{x_i}{nP(X = x_i)} = E[X]/n,$$

(ii) Give a proof that the expected value is linear i.e. E[X+Y]=E[X]+E[Y] for random variables X and Y. [5 marks]

Solution

$$E[X+Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i + y_j)P(X = x_i \text{ } and \text{ } Y = y_j)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i P\big(X = x_i \text{ } and \text{ } Y = y_j\big)$$
$$+ \sum_{i=1}^{n}\sum_{j=1}^{n} y_i P\big(X = x_i \text{ } and \text{ } Y = y_j\big)$$
$$= \sum_{i=1}^{n} x_i P(X = x_i \text{ } a) + \sum_{j=1}^{n} y_i P\big(\text{ } Y = y_j\big) = E[X] + E[Y]$$

A sequence of n bits is sent across a wireless link. Let random variable Y_i take value 1 when the i'th bit is received without error and 0 otherwise. Suppose the random variables Y_i i=1,2,..,n are independent and identically distributed with $E[Y_i]= \mu$.

(iii) Let random variable $Z = \sum_{i=1}^{n} Y_i$ be the number of bits received without error. Show that E[Z/n] = μ. Hint: use the linearity of the expected value. [5 marks]

Solution

$E\left[\frac{Z}{n}\right]$=$E[\frac{1}{n}\sum_{i=1}^{n} Y_i] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n} n\mu = \mu$

(iv) Using Chebyshev's inequality explain the weak law of large numbers and the behaviour of |Z/n − μ | as n becomes large. Recall that for random variable X Chebyshev's inequality is: $P(|X - \mu|\geq k) \leq E[(X- \mu)^2]/k^2$ for an k and μ. [5 marks]

Solution

Since the Y_i's are independent,

$$Var[\frac{1}{n}\sum_{i=1}^{n} Y_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i] = \frac{1}{n}Var(Y)$$

Also, $E(Y_i^2)= 1^2P(Y_i= 1)= P(Y_i= 1)=\mu$ and so $Var(Y_i)=E[Y_i^2]-E[Y_i]^2= \mu- \mu^2$. So by Chebyshev, $P(|Z/n - \mu|\geq k) \leq \mu (1- \mu)/(nk^2)$ and as n goes to infinity $P(|Z/n - \mu|\geq k)$ goes to zero. That is the estimate Z/n concentrated around the true value μ with probability 1.

(v) Explain what a confidence interval is, using Z/n as an estimate of μ as an example. Describe how to use bootstrapping to estimate a confidence interval.

[5 marks]

Solution

A confidence interval is typically a statement of the form $P(a \leq X \leq b)\geq c$, where c might for example have a value of 0.95. $P(|Z/n - \mu|\leq k))\geq c$ is an example of the confidence interval $P(\mu -k\leq Z/n\leq \mu+k)\geq c$. Suppose we have observed n values Y_i. In bootstrapping we resample (with replacement) from these observed values. Letting S be the indices of the values sampled, we then calculate $\widehat{Z/n}$=$\sum_{i \in S} Y_{i/n}$. Repeating this we obtain a sequence of estimates $\widehat{Z/n}$ from which we can estimate the distribution of $\widehat{Z/n}$ (form the fraction of times each value appears). Using this estimated distribution we can now either calculate the value c for a confidence interval by just summing up the fraction of values lying in the interval of interest or for a specified value of c we can calculate an interval over which the sum of the fractions is greater than or equal to c.

4. (i) With reference to Bayes Rule explain what is meant by the likelihood, prior and posterior. [5 marks]

Solution

Bayes rule states that P(F|E)=P(E|F)P(F)/P(E) for random events E and F. P(E|F) is referred to as the likelihood, P(F) as the prior and P(F|E) as the posterior.

(ii) Explain how the maximum a posteriori (MAP) estimate of a parameter differs from the maximum likelihood estimate. [5 marks]

Solution

A MAP estimate maximises P(F|E) whereas as ML estimate maximises P(E|F).

(iii) We observe data (x_i,y_i), i=1,2,…,*n* from *n* people, where x_i is the person's height and y_i is the person's weight.

(a) Explain how to construct a linear regression model for this data. [10 marks]

Solution

We model each value as the sum of an underyling linear function θx_i plus zero-mean gaussian noise i.e. as $y_i = \theta x_i + n_i$ where n_i is gaussian noise. We then typically select the value for θ that maximises the likelihood, or equivalently maximises the log-likelihood

$$-\sum_{i=1}^{n}(y_i - \theta x_i)^2$$

(b) Suppose we suspect that the weight of a person is not linearly related to their height but rather is related to the square root of their height. Explain how to modify the linear regression model to accommodate this. [5 marks]

Solution

We can change the model to be as $y_i = \theta\sqrt{x_i} + n_i$ and now select θ that maximises $-\sum_{i=1}^{n}(y_i - \theta\sqrt{x_i})^2$