ST3009 – Statistics
2017 Exam Solutions

Question 1

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

   (a) 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]

   (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]

   (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]

i)

Probability of red = 0.5
Probability of black = 0.5

$$\binom{5}{2}(0.5)^2(0.5)^3$$

$$10 * (0.25) * (0.125)$$

$$\boldsymbol{P(X) = 0.3125}$$

ii)

There are $\binom{10}{5}$ ways of drawing 5 balls from 10.

There are $\binom{5}{2}$ ways of drawing 2 red balls from 5.

There are $\binom{5}{3}$ ways of drawing 3 black balls from 5.

Therefore, the probability is:

$$\boldsymbol{P(X)} = \frac{\binom{5}{2} * \binom{5}{3}}{\binom{10}{5}} = \boldsymbol{0.3968}$$

iii) 4 floors
3 guests
P(floorX) = 0.25

First guest has 4 floors to choose from, second has 3, third has 2.

Total combinations = 4 * 3 * 2 = 24

There are 4^3 ways of 3 guests choosing from 4 floors, therefore probability is:

$$\boldsymbol{P(X) = \frac{4 * 3 * 2}{4^3} = 0.375}$$

Question 2

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

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

(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]

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

(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]

i)

Random Event: A random event is a subset of the sample space. Consider the tossing of two coins. The event {H, H} is a random event which is a subset of the sample space.

Random Variable: A random variable maps a random event to a real number. Consider the tossing of a six-sided die. Let the random variable X denote the number tossed by the die. X can take the values [1,6].

ii)

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

$$Var(X) = E[X^2] - E[X]^2$$

iii)

$$E[X] = 1(0.2) + 2(0.3) + 3(0.5)$$

$$\boldsymbol{E[X] = 2.3}$$

$$Var(X) = E[X^2] - E[X]^2$$

$$Var(X) = 5.9 - 5.29$$

$$\boldsymbol{Var(X) = 0.61}$$

iv) Two random variables X and Y are independent iff:

$$P(X = x \text{ } and \text{ } Y = y) = P(x = x)P(Y = y)$$

Holds for all values of x and y that variables X and Y can take.

v) X and Y can take values {1, 2, 3}
P(X/Y = 1/2/3) = p
V = XY

To verify they are dependent consider the example P(V=1 and X=2).
P(V=1) = P(X=1 and Y=1) = (1/3)(1/3) = 1/9
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. Therefore V and X are not independent.

Question 3

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]

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

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

Therefore, it is clear that E[X/n] = 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]

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

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i)P(X = x_i \; and \; Y = y_j) + \sum_{i=1}^{n}\sum_{j=1}^{n}(y_j)P(X = x_i \; and \; Y = y_j)$$

$$= \sum_{i=1}^{n}(x_i)P(X = x_i) + \sum_{j=1}^{n}(y_j)P(Y = y_j)$$

$$= \boldsymbol{E}[\boldsymbol{X}] + \boldsymbol{E}[\boldsymbol{Y}]$$

(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]

$$E[{}^{Z}/_{N}] = E\left[\frac{1}{n} * \sum_{i=1}^{n} Y_i\right]$$

$$= \frac{1}{n} * \sum_{i=1}^{n} E[Y_i]$$

$$= \frac{1}{n} * \; n * E[Y_i]$$

$$= E[Y_i]$$

$$= \mu$$

(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]

First, calculate the variance. Since our Yi's are independent we can represent that variance as:

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

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

$$= \frac{1}{n}Var(Y)$$

$$Var(Y) = E[Y^2] - E[Y]^2$$

Y_i represents a bit being an error or not, thus is an indicator variable and can only take values 0 or 1. As a result:

$$E[Y^2] = 0^2 * P(Y = 0) + 1^2 * P(Y = 1)$$

$$= 1^2 * P(Y = 1)$$
$$= P(Y = 1)$$
$$= \mu$$

Using this, we can then calculate Var(Yi):

$$Var(Y) = \mu - \mu^2$$

$$Var(Z) = \frac{(\mu - \mu^2)}{n}$$

Then plugging this into Chebyshev's inequality:

$$P(|X - \mu| \geq k) \leq \frac{Var(X)}{k^2}$$
$$P\left(\left|\frac{Z}{n} - \mu\right| \geq k\right) \leq \frac{Var\left(\frac{Z}{n}\right)}{k^2}$$

$$P\left(\left|\frac{Z}{n} - \mu\right| \geq k\right) \leq \frac{(\mu - \mu^2)}{nk^2}$$

From this, we can see that as n goes to infinity, $P\left(\left|\frac{Z}{n} - \mu\right| \geq k\right)$ goes to 0.

(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]

A confidence interval is typically a statement of the form:

$$P(a \ \leq X \leq b) \geq c$$

Where c represents the confidence that a random variable X lies between a and b. For example c might be 0.95.

$P\left(\left|\frac{Z}{n} - \mu\right| \geq k\right) \leq c$ is an example of the following confidence interval:

$$P\left(\mu - k \ \leq \frac{Z}{n} \leq \mu + k\right) \geq c$$

Bootstrapping can be used to estimate a confidence interval as follows. Suppose we have observed n values of Y_i. In bootstrapping we re-sample (with replacement) from these observed values. Letting S be the indices of the values sampled, we then calculate:

$$\frac{\hat{Z}}{n} = \sum_{i \in S} Y_{\frac{i}{n}}$$

Repeating this we obtain a sequence of estimates for $\frac{\hat{Z}}{n}$ from which we can estimate the distribution of $\frac{\hat{Z}}{n}$ (from 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.

Question 4

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

For random events E and F, Bayes Rule states:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

- $P(F|E)$ is the Likelihood
- $P(E)$ is the Prior
- $P(E|F)$ is the Posterior

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

A MAP maximises the posterior P(E|F)

A ML maximises the likelihood P(F|E)

(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]

We model each value as the sum of an underlying linear function θx_i plus zero-mean Gaussian noise i.e as the following (where ni is Gaussian noise):

$$y_i = \theta x_i + n_i$$

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]

We can change the model to be as:

$$y_i = \theta\sqrt{x_i} + n_i$$

And now select θ that maximises:

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