APPLIED ANALYTICAL STATISTICS

PROBLEM SET FOR LECTURE 2

A Mahdi, GY Qian, AE Zarebski

Set theory

1. Let A and B be events with probability $P(A) = \frac{2}{3}$ and $P(B) = \frac{2}{5}$. Show that $\frac{1}{15} \leq P(A \cap B) \leq \frac{2}{5}$. Find the corresponding bounds for $P(A \cup B)$.

2. If A, B and C are subsets of S, show that:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C)$$
$$-P(A \cap C) + P(A \cap B \cap C)$$

Random variables

3. To prepare them for marking students' test papers, some volunteer practised by marking a paper that the examiners had originally graded at 97.0%. The random distribution of marks given by these volunteers are as follows:

Mark (%)	96.8	96.9	97.0	97.1	97.2
Proportion	0.03	0.07	0.80	0.07	0.03

   What are the mean and standard deviation of the marks? Give your answer to 3 significant figures.

4. The following year, another group of markers practised on that same paper, producing a mean mark of 97.1% with a standard deviation of 0.15%. What are the mean and standard deviation of the average of two marks, one from each group of markers?

Binomial distribution

5. What is a binomial random variable? Derive its mean and variance.
6. At a local constituency, the probability that residents will spoil their votes is 0.03. What is the probability that, in a group of 10 residents living in the area:
   (a) Nobody spoils the vote?
   (b) Exactly two people spoils their votes?
7. In a group of ten people, what is the most likely number of people who will spoil their votes?

8. Use the Central Limit Theorem to estimate the probability that no more than 30 people out of 1000 will spoil their votes. Use the continuity correction for this problem.

Poisson distribution

9. What is the Poisson random variable? Derive its mean and variance.
10. The number of tweets sent out from the social media account of a large charity organisation in any time interval of length T minutes is a Poisson random variable with mean $\frac{T}{2}$. What is the probability that in a 10 minute period:
    (a) There are no tweets.
    (b) There are 3 or more tweets.
11. Being an avid follower of this charity, Nora does not want to be late in reading a single tweet. Find the maximum length of time that Nora can be distracted so that the probability of being late to read any number of tweets is to be less than 10%.

Normal distribution

12. A busy politician typically asks two PR firms, which we will call companies A and B (to protect their identities), to write the party's speeches. Company A writes 500 speeches a year with the length of these speeches being a normally distributed random variable with mean 8.01 minutes and standard deviation 0.1 minutes, while Company B writes 1000 per month with their lengths being a normally distributed random variable with mean 7.95 minutes and standard deviation 0.08 minutes. It is found that 5% of Company A's speeches must be sent back for editing and that 8% of speeches from Company B require editing. The politician cannot give a speech if it requires editing or if its length is outside the range $(7.85, 8.15)$ minutes.
    (a) What proportion of speeches cannot be given?
    (b) Suppose a random chosen speech cannot be given. What is the probability that Company A wrote this speech?
    (c) What is the probability of the length of two randomly chosen speeches differing by more than 0.3 minutes if: i) they both come from Company A or ii) one comes from Company A, while the other from B?
    (d) So concerned is the politician that the speeches be of the right length of time, that they are routinely timed during test runs. These test runs aim to estimate the mean with a confidence interval smaller or equal to ± 0.1 minutes, with 99% confidence, and are applied separately to companies A and B. How many speeches from each company do we need to time, assuming that the variances do not vary from their original values?

Confidence intervals

13. We want to estimate the mean household income in a certain area, for which we take a sample of 1000 households and obtain a sample average of £20,000 with a sample standard deviation of £4,000. We assume this corresponds to a normal distribution. Use a 95% confidence level for all the following questions:

(a) Assuming that the variance of the distribution is equal to the sample variance, what is the confidence interval for the mean income?
(b) What would be the confidence interval if no strong assumptions are made about the variance?
(c) What would be the answers to the two questions above if the values were to come from a sample of only 15 households?
(d) During recent years, the mean income in the area grew, while the median income remained stable. How would you explain this?

Hypothesis testing

14. A new mobile phone producer is trying to break into the market. Its point of difference is that the phones can last an average of 100 hours, with a 5 hour standard deviation, on a single battery charge. Before the stores on Cornmarket Street can accept them, however, samples of size $n = 10$ are periodically checked to ensure that the mean remains equal to or greater than 100 hours, with a significance level $\alpha = 1\%$. If the significance level is not reached, the production is stopped for inspection.

(a) Assuming that the standard deviation remains constant, what is the minimum sample average that can be accepted?
(b) If we take samples once a week, what is the probability that the production will stop for an unnecessary inspection (i.e. that the test will detect an error even though the mean battery life is still equal to or greater than 100 hours) during the 10 weeks?
(c) If a production error makes the mean of the battery life drop to 99.5 hours, with no change in variance, what is the probability that this will not be detected in the first test after the error? What is the probability of detecting it in the first 10 tests?
(d) What are the alternatives to reducing: i) Type I errors; ii) Type II errors, iii) both Type I and Type II errors? Discuss the factors that would be considered to carry out an ideal or optimal test.

Answers

1. $\frac{2}{3} \leq P(A \cup B) \leq 1$
2. Proof question
3. $\mu = 97.0\%, \sigma = 0.0616\%$
4. $\mu = 97.05\%, \sigma = 0.0811\%$
5. $\mu = np, \sigma^2 = npq$
6. (a) 0.737
   (b) 0.032
7. 0
8. 53.7%
9. $\mu = \lambda, \sigma^2 = \lambda$
10. (a) 0.0067
    (b) 0.875
11. 0.21 minutes
12. (a) 18.15%
    (b) 32.8%
    (c) i) 3.4% ii) 3.3%
    (d) 7 speeches from Company A, 5 from B
13. (a) $\$19,752 \leq \mu \leq \$20,248$
    (b) $\$19,752 \leq \mu \leq \$20,248$
    (c) $17,976 \leq \mu \leq 22,024$; $\$17,976 \leq \mu \leq \$22,215$
14. (a) 96.32 hours
    (b) 9.56%
    (c) 97.8%; 20.1%