Attempt 5 questions from 7.

Question 1. (20 marks)

(a) An IT consultant is responsible for three software engineering projects X, Y and Z. He knows that the probability of completing project X in time is 0.99, for project Y this probability is 0.95 and for project Z it is 0.80.

i (2 marks) What assumption do you need to make in order to calculate the probability of completing all three projects in time, from the information given?

ii (3 marks) Calculate the probability of completing all three projects in time.

iii (3 marks) Calculate the probability that only projects X and Y will be completed on time.

(b) The following contingency table illustrates the number of 400 students in different departments according to gender.

	Computer Science	Statistics	Equine Science
Males	140	100	20
Females	30	80	30

i (2 marks) What is the probability that a randomly chosen person from the sample is a computer science student?

ii (2 marks) What is the probability that a randomly chosen person from the sample is both female and studying statistics?

iii (2 marks) What is the probability that a randomly chosen person from the sample is male?

iv (2 marks) Given that a student studies statistics, what is the probability that the student is female?

v (2 marks) What is the probability that a randomly chosen person from the sample is a male or a statistics student?

vi (2 marks) Given that the student is female, what is the probability that she is an equine science student?

Question 2. (20 marks)

(a) For a digital communication channel, the probability of a bit being received in error is 10%. Consider the case where 100 bits are transmitted. Answer the following questions.

- i (3 marks) What is the probability that the number of bits received in error is 10?
- ii (3 marks) What is the probability that the number of bits received in error is greater than 10?
- iii (3 marks) What is the probability that the number of bits received in error does not exceed 20?

```
> dpois(5:15,lambda=10)
 [1] 0.03783327 0.06305546 0.09007923 0.11259903 0.12511004 0.12511004
 [7] 0.11373640 0.09478033 0.07290795 0.05207710 0.03471807
>
> ppois(5:15,lambda=10)
 [1] 0.06708596 0.13014142 0.22022065 0.33281968 0.45792971 0.58303975
 [7] 0.69677615 0.79155648 0.86446442 0.91654153 0.95125960
>
> pbinom(5:15,size=100,prob=0.10)
 [1] 0.05757689 0.11715562 0.20605086 0.32087389 0.45129017 0.58315551
 [7] 0.70303310 0.80182111 0.87612321 0.92742703 0.96010947
>
> dbinom(5:15,size=100,prob=0.10)
 [1] 0.03386580 0.05957873 0.08889525 0.11482303 0.13041628 0.13186535
 [7] 0.11987759 0.09878801 0.07430209 0.05130383 0.03268244
```

(b) A computer software company which specializes in database systems sells 2 software licences every day, on average. Answer the following questions.

i (3 marks) What is the probability that the software company sells at least one licence in one particular day?

ii (4 marks) What is the probability that the software company will sell exactly one licence in one particular day?

iii (3 marks) What is the probability that the software company will sell sixteen licences or more in a five day working week?

```
> 0:5
[1] 0 1 2 3 4 5
>
> ppois(0:5,lambda=2)
[1] 0.1353353 0.4060058 0.6766764 0.8571235 0.9473470 0.9834364
>
> ppois(0:5,lambda=(1/2))
[1] 0.6065307 0.9097960 0.9856123 0.9982484 0.9998279 0.9999858
>
> dpois(0:5,lambda=2)
[1] 0.13533528 0.27067057 0.27067057 0.18044704 0.09022352 0.03608941
>
> dpois(0:5,lambda=(1/2))
[1] 0.6065306597 0.3032653299 0.0758163325 0.0126360554 0.0015795069
[6] 0.0001579507
>
> pbinom(0:5,size=2,prob=0.05)
[1] 0.9025 0.9975 1.0000 1.0000 1.0000 1.0000
```

Question 3. (20 marks)

(a) The probability distribution of discrete random variable X is tabulated below. There are 6 possible outcome of X, i.e. 1, 2, 3, 4 ,5 and 6.

x_i	1	2	3	4	5	6
$P(x_i)$	0.16	0.14	k	0.17	0.21	0.19

i (1 marks) Compute the value for k.

ii (3 marks) Determine the expected value $E(X)$.

iii (3 marks) Evaluate $E(X^2)$.

iv (3 marks) Compute the variance of random variable X.

(b) The design of an online database gives a mean time to process a query from a central server of 250 milliseconds with a standard deviation of 50 milliseconds. It can be assumed that the query times are normally distributed

i (2 Mark) What proportion of query times will be greater than 325 milliseconds?

ii (2 Mark) What proportion of query times will be less than 300 milliseconds?

iii (3 Mark) What proportion of query times will be between 150 milliseconds and 250 milliseconds?

iv (3 Mark) What is the query time above which 10% of query times will be?

```
> Zs
 [1] -2.25 -2.00 -1.75 -1.50
 [5] -1.25 -1.00 -0.75 -0.50
 [9] -0.25  0.00  0.25  0.50
[13]  0.75  1.00  1.25  1.50
[17]  1.75  2.00  2.25
>
> pnorm(Zs)
 [1] 0.01222447 0.02275013 0.04005916 0.06680720
 [5] 0.10564977 0.15865525 0.22662735 0.30853754
 [9] 0.40129367 0.50000000 0.59870633 0.69146246
[13] 0.77337265 0.84134475 0.89435023 0.93319280
[17] 0.95994084 0.97724987 0.98777553
```

Question 4. (20 marks)

(a) Answer the following questions on the theory of statistical inference.

i (3 marks) Briefly describe the central limit theorem.

ii (1 marks) Provide a brief description of the standard error.

iii (3 marks) In the context of hypothesis testing, explain what a p-value is, and how it is used. Support your answer with a simple example.

iv (4 marks) What is meant by Type I error and Type II error?

(b) In a computer hardware manufacturing plant, machine X and machine Y produce identical components. The management investigate whether or not there is a difference in the mean diameter of the components from both machines.

A random sample of 144 components from machine X had a mean of 36.38 mm and a standard deviation of 3.0 mm. A random sample of 225 components from machine Y had a mean of 36.88 mm and a standard deviation of 2.8 mm.

A hypothesis test was used to determine whether or not the means are significantly different. A 5% significance level was used.

i (2 marks) What is the null and alternative hypothesis?

ii (3 marks) Compute the test statistic.

iii (2 marks) For the test statistic, determine the corresponding p-value.

iv (2 marks) What is your conclusion for this procedure? Justify your answer.

```
        Two Sample t-test

data:  X and Y
t = -1.576, df = 367, p-value = 0.1159
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
 -1.1073443  0.1220474
sample estimates:
mean of x mean of y
 36.38613  36.87877
```

Question 5. (20 marks)

(a) The typing speeds for one group of eight computer science students were recorded both at the beginning of year l of their studies and at the end of year 4. The results (in words per minute) are given below:

Subject	A	B	C	D	E	F	G	H
Before	170	180	184	183	186	184	169	130
After	190	194	200	199	197	200	185	145

A study was carried out to determine whether students improve in terms of typing speed over the four years of their university studies? A significance level of 5% is used.

i (3 marks) Compute the case-wise differences, the mean of the case-wise differences, and the standard deviation of the case-wise differences

ii (2 marks) Formally state the null and alternative hypothesis.

iii (3 marks) Compute the test statistic.

iv (3 marks) What is your conclusion for this procedure? Justify your answer.

```
        Paired t-test

data:  After and Before
t = -17.4864, df = 7, p-value = 4.924e-07
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
 -17.59602 -13.40398
sample estimates:
mean of the differences
                  -15.5
```

(b) A study finds that a percentage of 40% of IT users out of a random sample of 400 in a large community preferred one web browser to all others. In another large community, 30% of IT users out of a random sample of 300 prefer the same web browser.

i (2 marks) Compute the point estimate for the difference in proportions of IT users who prefer this particular web browser.

ii (3 marks) Compute a 95% confidence interval for this difference in proportions.

iii (3 marks) Based on this confidence interval, test the hypothesis that the proportion of IT users using this web browser is the same for both communities. State your null and alternative hypothesis clearly.

```
> prop.test(90,300,0.4)

        1-sample proportions test

data:  90 out of 300, null probability 0.4
X-squared = 12.0868, df = 1, p-value = 0.0005078
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
 0.2493751 0.3558434
sample estimates:
  p
0.3
```

```
> prop.test(160,400,0.3)

        1-sample proportions test

data:  160 out of 400, null probability 0.3
X-squared = 18.5744, df = 1, p-value = 1.634e-05
alternative hypothesis: true p is not equal to 0.3
95 percent confidence interval:
 0.3519482 0.4500035
sample estimates:
  p
0.4
```

```
> prop.test(c(90,160),c(300,400))

        2-sample test for equality of proportions

data:  c(90, 160) out of c(300, 400)
X-squared = 7.0375, df = 1, p-value = 0.007982
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.17358417 -0.02641583
sample estimates:
prop 1 prop 2
   0.3    0.4
```

Question 6. (20 marks)

(a) The input source to a noisy communication channel is a random variable X over the four symbols $\{a, b, c, d\}$. The output from this channel is a random variable Y over these same four symbols. The joint distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d
y=b	1/8	1/8	0	1/4
y=b	0	0	1/8	0
y=c	0	1/8	1/8	0
y=d	1/8	0	0	0

i (2 marks) Write down the marginal distribution for X and compute the marginal entropy $H(X)$.

ii (2 marks) Write down the marginal distribution for Y and compute the marginal entropy $H(Y)$.

iii (4 marks) What is the joint entropy $H(X, Y)$ of the two random variables?

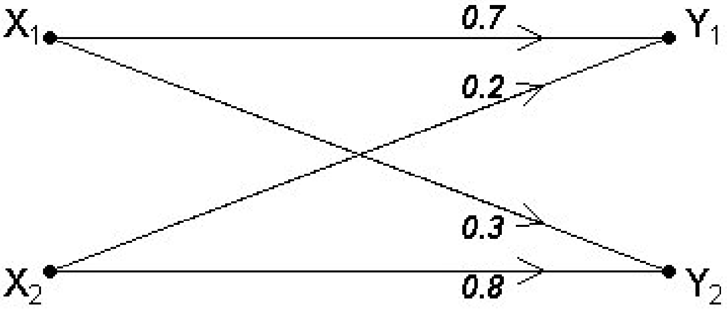
iv (4 marks) What is the conditional entropy $H(Y|X)$?

v (4 marks) What is the conditional entropy $H(X|Y)$?

vi (4 marks) What is the mutual information $I(X;Y)$ between the two random variables?

Question 7. (20 marks)

(a) Consider the binary channel in the figure below.



i (3 marks) Determine the channel matrix of the channel

ii (3 marks) Find $P(Y_1)$ and $P(Y_2)$ when $P(X_1) = 0.6$ and $P(X_2) = 0.4$

iii (3 marks) Find the joint probabilities $P(X_1, Y_1)$ and $P(X_2, Y_2)$.

(b) A discrete memoryless source X has five symbols $\{x_1, x_2, x_3, x_4, x_5\}$ with probabilities $P(x_1) = 0.42$, $P(x_2) = 0.22$, $P(x_3) = 0.17$, $P(x_4) = 0.10$ and $P(x_5) = 0.09$.

i (5 marks) Construct a Huffman code for X.

ii (4 marks) Calculate the efficiency of the code.

iii (2 marks) Calculate the redundancy of the code.

Formulae

Probability

- Conditional probability:
$$P(B|A) = \frac{P\,(A \text{ and } B)}{P\,(A)}.$$
- Bayes' Theorem:
$$P(B|A) = \frac{P\,(A|B) \times P(B)}{P\,(A)}.$$
- Binomial probability distribution:
$$P(X = k) =^n C_k \times p^k \times (1-p)^{n-k} \qquad \left(\text{where} \qquad {}^nC_k = \frac{n!}{k!\,(n-k)!}.\right)$$
- Poisson probability distribution:
$$P(X = k) = \frac{m^k \mathrm{e}^{-m}}{k!}.$$

Information Theory

- $I(p) = -log_2(p) = log_2(1/p)$
- $I(pq) = I(p) + I(q)$
- $H = -\sum_{i=1}^{m} p_i \; log_2(p_i)$
- $E(L) = \sum_{i=1}^{m} l_i p_i$
- Efficiency $= H/E(L)$
- $I(X;Y) = H(X) - H(X|Y)$
- $P(C[r]) = \sum_{j=1}^{m} P(C[r]|Y = d_j)P(Y = d_j)$

Confidence Intervals

One sample

$$S.E.(\bar{X}) \;=\; \frac{\sigma}{\sqrt{n}}.$$
$$S.E.(\hat{P}) \;=\; \sqrt{\frac{\hat{p} \times (100 - \hat{p})}{n}}.$$

Two samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

$$S.E.(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1 \times (100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}.$$

Hypothesis tests

One sample

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

$$S.E.(\pi) = \sqrt{\frac{\pi \times (100 - \pi)}{n}}$$

Two large independent samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

$$S.E.(\hat{P}_1 - \hat{P}_2) = \sqrt{(\bar{p} \times (100 - \bar{p}))\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Two small independent samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Paired sample

$$S.E.(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

Standard deviation of case-wise differences

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n - 1}}.$$