IMPERIAL COLLEGE LONDON

M.Eng EXAMINATION 2024

PART II

This paper is also taken for the relevant examination for the Associateship

CIVE50008 STATISTICS

Monday 20th May 9:30am - 12:30pm

Answer the five questions

The five questions carry the marks 17, 12, 21, 25, 25 respectively.

Statistical tables are provided at the end of the paper, and a formula sheet is provided

Question 1

A company provides parking space for coaches. As this mobilises a large amount of space that could perhaps be used for more profitable purposes, the company is monitoring the number of slots occupied during working hours (a slot is considered occupied during a given hour when there is a coach parked there at some point during the hour). They carry out a survey during 25 hours per week over 4 weeks. The sample of size 100 they obtain is shown in table 1:

Number of coaches	0	1	2	3	4	5	6	7	8
Number of hours	5	15	23	22	17	11	4	2	1
cummulative freq. (to find med, q_1, q_3)	5	20	43	65	82	93	97	99	100

(a) Calculate the sample mean $\bar{x}$ and the variance s^2 of the distribution of number of coaches parked during any hour, bearing in mind that these will be used to estimate distribution parameters below. What are the median $qp(0.5)$ and the inter-quartile range of this distribution?

we unbiased estimator!

$IQR = qp(0.75) - qp(0.25)$

[7 marks]

$\bar{x} = \frac{\sum fx}{\sum f} = 2.96$ $\quad s^2 = \frac{1}{n-1}\sum f(x-\bar{x})^2 = 2.8469$

$k = 50/100\,(100+1) = 50.5$

$\text{med} = \frac{50^{th}\text{ term} + 51^{st}\text{ term}}{2} = \frac{3+3}{2} = 3$

$k\,(\text{for } q_1) = 25/100\,(100+1) = 25.25$

$q_1 = \frac{25^{th}\text{ term} + 26^{th}\text{ term}}{2} = \frac{2+2}{2} = 2$

$k\,(\text{for } q_3) = 75/100\,(100+1) = 75.75$

$q_3 = \frac{75^{th}\text{ term} + 76^{th}\text{ term}}{2} = \frac{4+4}{2} = 4$

$IQR = q_3 - q_1 = 4 - 2 = 2$

Consider a Poisson distribution given by its probability mass function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } k = 0, 1, \ldots$$

forgot $X \sim Po(\lambda)$ is discrete or continuous? hint given, PMF → discrete.

keep in mind we know $E(X)$ for $X \sim Po(\lambda) = \lambda$

(b) Find the mean *E(X)* (show your derivation).

[3 marks]

this should match...

$E(X)$, for discrete PMF: $\sum_{\forall x \in \mathbb{R}} x \underbrace{P(X=x)}_{PMF} = \sum_{k=0,1,2,\ldots} k \frac{\lambda^k e^{-\lambda}}{k!} = \underbrace{e^{-\lambda}}_{\text{not } k\text{-dependent (bring it out)}} \sum \cancel{k} \frac{\lambda^k}{\cancel{k}(k-1)!} = e^{-\lambda} \sum_{k=1,2,3,\ldots} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{\text{taylor series} = e^{\lambda}} = \lambda e^{-\lambda}(e^{\lambda}) = \lambda$

this is $(k-1)$

(realise that $(0-1)!$ doesn't exist...)

we know $Var(X)$ for $X \sim Po(\lambda) = \lambda$ too! (memorise)

(c) Find the variance *var(X)* (show your derivation).

[5 marks]

$Var(X)$ for discrete PMF $= \sum x^2 P(X=x) - E(X)^2$
$= \sum k^2 P(X=k) - \mu^2$
$= \sum_{k=0,1,2,\ldots} k^2 \left(\frac{\lambda^k e^{-\lambda}}{k!}\right) - \mu^2$

(by the way I ain't writing the full working, check notes page. 21)

(d) Fit a Poisson distribution to the number of coaches parked during any hour using the method of moments. Note: use the simplest possible type of method of moments.

number of moment equation = number of parameters of the distribution.

[2 marks]

$X \sim Po(\lambda)$ has ONE parameter,

$E(X) = \lambda = \bar{x}$

$X \sim Po(2.96)$

If question want it more accurate, do minimalisation with weightage:

$E(X) = \lambda = \bar{x}$
$\lambda = 2.96$

$Var(X) = \lambda = s^2$
$\lambda = 2.8469$

(I put equal weightage for simplification, usually, smaller moment should have higher weightage, eg. 0.65 : 0.35

$\min_{\lambda} \sum_{k=1}^{p} w_k (M_k - m_k)^2 = \min_{\lambda} \{0.5(E(X) - \bar{x})^2 + 0.5(Var(X) - s^2)^2\}$
$= \min_{\lambda} \{0.5(\lambda - 2.96)^2 + 0.5(\lambda - 2.8469)^2\}$

let $f(\lambda) = 0.5(\lambda - 2.96)^2 + 0.5(\lambda - 2.8469)^2$

$f'(\lambda) = (\lambda - 2.96) + (\lambda - 2.8469)$
$= 2\lambda - 5.8069$

let $f'(\lambda) = 0$
$\therefore \lambda = 2.90345$

$f''(\lambda) = 2$
$f''(\lambda) > 0$ (min ✓)

$X \sim Po(2.90345)$

Question 2

(a) Consider a consignment of electronic components for environmental monitoring devices. If the probability of a defective component is 5%, what is the probability that there are 2 or 3 defective components in a 5 pack?

- independent
- only two outcome (defective / not defective)
- n trials

→ Binomial dist.

$X \sim B(5, 0.05)$

$P(X=2) + P(X=3) = {}^5C_2 \times 0.05^2 \times 0.95^3 + {}^5C_3 \times 0.05^3 \times 0.95^2$

$= 0.0225625$

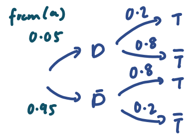
[5 marks]

(b) A test has been devised to identify non-defective components. Its ability to do so is characterised by the following probabilities:

$$P(T|D) = 0.2; P(T|\overline{D}) = 0.8$$

where T is a positive outcome for the test and D denotes that the component is defective while $\overline{D}$ denotes that it is not.

A test is carried out on one component and produces a positive result, i.e. outcome T. What is the probability that this component is not defective?



need to find $P(\bar{D}|T)$

$P(\bar{D}|T) = \dfrac{P(T|\bar{D})\,P(\bar{D})}{P(T)}$ (Bayes' Law)

$= \dfrac{0.8 \times 0.95}{0.77}$

$= 0.9870$

$P(T) = P(T|D)\,P(D) + P(T|\bar{D})\,P(\bar{D})$

$= 0.2 \times 0.05 + 0.8 \times 0.95$

$= 0.77$

[7 marks]

Question 3

The times T between two consecutive truck arrivals at a sand loading station on a worksite are approximately exponentially distributed: $T \sim Exp(\lambda)$. This means that the probability distribution function of T is given by: $f(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$.

(a) Explain the principle of maximum likelihood estimation as applied to a sample $\{t_1, t_2, \ldots t_n\}$ and why it amounts to maximising the function of λ: $\prod_{i=1}^{n} f(t_i)$. What are its advantages compared to the method of moments?

→ maximising $\prod_{i=1}^{m} f(t_i \mid \text{parameters of distributions})$

→ MOM, more subjective. depends on what order of moments

→ MLM: compare all data to PDF

[4 marks]

(b) Show that the maximum likelihood estimator of λ is: $\dfrac{n}{\sum_{i=1}^{n} t_i} = \dfrac{1}{\bar{t}}$ where $\bar{t}$ is the average of a sample of size n of inter-arrival times, $\{t_1, t_2, \ldots t_n\}$. How does this compare with what the method of moments would provide as estimator?

$\mathcal{L}(\lambda) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}$

$= \lambda^n e^{-\lambda(t_1 + t_2 + t_3 + \ldots + t_n)}$

$= \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}$

want to find $\max\{\ln(\mathcal{L}(\lambda))\}$

$\ln(\mathcal{L}(\lambda)) = \ln(\lambda^n e^{-\lambda \Sigma t})$

$= \ln \lambda^n + \ln e^{-\lambda \Sigma t}$

$= n \ln \lambda + -\lambda \Sigma t$

$\frac{d}{d\lambda}(\ln(\mathcal{L}(\lambda))) = \frac{n}{\lambda} - \Sigma t$

to find max: $\frac{n}{\lambda} - \Sigma t = 0$ $\quad \frac{n}{\lambda} = \Sigma t$ $\quad \lambda = \frac{n}{\Sigma t}$

note that $\bar{t} = \frac{\Sigma t}{n}$

$\frac{1}{\bar{t}} = \frac{n}{\Sigma t}$

and hence $\lambda = \frac{n}{\Sigma t} = \frac{1}{\bar{t}}$

method of moment:

$T \sim Exp(\lambda)$

simplest method is number of moment equation = number of parameters = 1

$\underbrace{E(T) = \frac{1}{\lambda}}_{\text{known, Exp dist. (population)}} = \underbrace{\bar{t}}_{\text{unbiased estimator (sample)}}$

$\boxed{\lambda = \frac{1}{\bar{t}}}$ ✓

[6 marks]

(c) The following sample of the times of arrival of trucks during a 90 minute measurement period was collected:

{0; 3; 11; 15; 25; 26; 28; 30; 42; 47; 50; 58; 74; 76; 77; 78} in minutes

$T = \{3, 8, 4, 10, 1, 2, 2, 12, 5, 3, 8, 16, 2, 1, 1\}$, $\bar{t} = 5.2$, $\lambda = \frac{1}{\bar{t}} = 0.1923$: $T \sim Exp(0.1923)$, $P(T > 10) = \int_{10}^{\infty} \lambda e^{-\lambda t} dt = [-e^{-\lambda t}]_{10}^{\infty} = 0 + e^{-0.1923(10)} = 0.1462$

Using this sample, calculate the maximum likelihood estimate of λ and the probability that the time between two truck arrivals is greater than 10 minutes. Is there a piece of information contained in the sampled data that is being overlooked in the calculation of the likelihood for this sample? Can you give an idea of how much difference taking this into account could make?

watch marking scheme.
I didn't figure this out.

[6 marks]

(d) Consider three consecutive truck arrivals which are assumed independent. How is the time separating the first from the third arrival distributed (show this by considering this as the sum of two random variables)?

[5 marks]

$T \sim Exp(\lambda)$

$T_1 \sim Exp(\lambda)$ $\quad$ $T_2 \sim Exp(\lambda)$

$S = T_1 + T_2 \sim ?$

Step 1: Find PDF of $S = T_1 + T_2$

Assuming independence of the added distributions:

$$K = I + J \Rightarrow p_K(k) = \sum_{j=0}^{k} p_I(k-j)p_J(j); \quad Z = X + Y \Rightarrow f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy$$

$$S = T_1 + T_2 \rightarrow f_S(s) = \int_{-\infty}^{\infty} f_{T_1}(s-y) f_{T_2}(y)\,dy$$

$s - y \geq 0$ and $y \geq 0$

$y \leq s$ and $y \geq 0$

$$= \int_0^s \lambda e^{-\lambda(s-y)} \cdot \lambda e^{-\lambda y}\,dy$$

$$= \lambda^2 \int_0^s e^{-\lambda s}\,dy$$

$$= \lambda^2 \left[y e^{-\lambda s} \right]_0^s$$

$$= \lambda^2 s e^{-\lambda s}$$

$S = T_1 + T_2 \sim Gamma\left(2, \frac{1}{\lambda}\right)$

Question 4

(a) Consider the joint probability density function given by

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4}(1+xy) & for\ |x| < 1\ and\ |y| < 1 \\ 0 & otherwise \end{cases}$$

$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$

need to first find $f_X(x)$ (marginal of x):

$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$

$= \int_{-1}^{1} \frac{1}{4}(1+xy)\,dy$

$= \left[\frac{1}{4}y + \frac{1}{8}xy^2\right]_{-1}^{1}$

$= \left[\frac{1}{4} + \cancel{\frac{1}{8}x} - \left(-\frac{1}{4} + \cancel{\frac{1}{8}x}\right)\right]$

$= \frac{1}{2}$

$f_{Y|X}(y|x) = \frac{\frac{1}{4}(1+xy)}{\frac{1}{2}}$

$= \frac{1}{2}(1+xy)$ //

(i) Find the conditional PDF of Y given X

(ii) Show that X and Y are not independent and X^2 and Y^2 are independent

$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$

$= \int_{-1}^{1} \frac{1}{4}(1+xy)\,dx$

$= \left[\frac{1}{4}x + \frac{1}{8}x^2y\right]_{-1}^{1} = \frac{1}{4} + \cancel{\frac{1}{8}y} - \left(-\frac{1}{4} + \cancel{\frac{1}{8}y}\right) = \frac{1}{2}$

$f_{Y|X}(y|x) = \frac{1}{2}(1+xy) \neq f_Y(y)$

not independent.

marking scheme. is definitely wrong!

you can't prove independence with $E(X^2Y^2) = E(X^2)E(Y^2)$!

we have to find PDF of X^2 and Y^2 and compare with joint PDF of X^2Y^2

$U = X^2,\ V = Y^2$

$F_{UV}(u,v) = P(U \le u, V \le v)$

$= P(X^2 \le u, Y^2 \le v)$

$= P(-\sqrt{u} \le X \le \sqrt{u}; -\sqrt{v} \le Y \le \sqrt{v})$

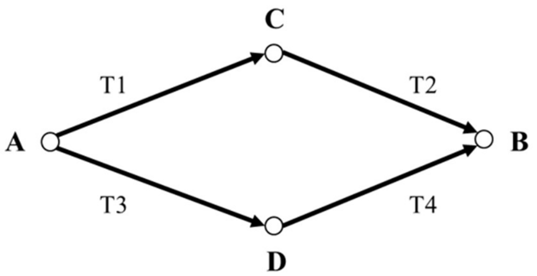
$= \int_{-\sqrt{v}}^{\sqrt{v}}\int_{-\sqrt{u}}^{\sqrt{u}} f_{XY}(x,y)\,dx\,dy = \sqrt{uv}$ $\longrightarrow$ then find $F_U(u), F_V(v)$ check if $F_{UV}(u,v) = F_U(u)F_V(v)$ ✓

$\int F_{UV}(u,v)\,dv$

$\int F_{UV}(u,v)\,du$

[5 marks]

(b) Two students (Student 1 and Student 2) are travelling from City A to City B. There are two alternative routes between City A and City B. One is the upper route through City C and the other is the lower route through City D. Student 1 decides to take the upper route and Student 2 decides to take the lower route. The road network of two routes is shown below:



ans b(ii): $P(\underbrace{T_3 + T_4}) > 10)$

find the dist. of this:

$T_3 + T_4 \sim N(\mu_{T_3} + \mu_{T_4}, \sigma_{T_3}^2 + \sigma_{T_4}^2 + 2\rho\sigma_{T_3}\sigma_{T_4})$

$T_3 + T_4 \sim N(5+4, 3^2 + 1^2 + 2 \times 0.8 \times 3 \times 1)$

$\therefore T_3 + T_4 \sim N(9, 14.8)$

$P(T_3 + T_4 > 10) = P\left(Z > \frac{10-9}{\sqrt{14.8}}\right)$

$= P(Z > 0.2599)$

$= 1 - P(Z < 0.2599)$

$= 1 - 0.6026$

$= 0.3974$

Travel times (in hours) between the cities indicated are normally distributed $N(\mu, \sigma^2)$ as: $T1 \sim N(6, 2^2)$; $T2 \sim N(4, 1^2)$; $T3 \sim N(5, 3^2)$ and $T4 \sim N(4, 1^2)$. Although the travel times can generally be assumed to be statistically independent, T3 and T4 are dependent with a correlation coefficient of 0.8.

(i) What is the probability that Student 2 will not arrive in City B within 10 hours?

$P(T_3 + T_4 > 10)$

(ii) What is the probability that Student 1 will arrive in City B earlier than Student 2 by at least 1 hour?

$P(\underbrace{T_1 + T_2} < T_3 + T_4 - 1)$

$E(X-Y) = E(X) - E(Y)$; $Var(X-Y) = Var(X) + Var(Y)$!

$T_1 + T_2 \sim N(6+4, 2^2+1^2)$

$T_1 + T_2 \sim N(10, 5)$

$(T_1 + T_2) - (T_3 + T_4) \sim N(10 - 9, 5 + 14.8)$

$\sim N(1, 19.8)$

$P(T_1 + T_2 - (T_3 + T_4) < -1)$

$= P\left(Z < \frac{-1-1}{\sqrt{19.8}}\right) = 1 - P\left(Z < \frac{2}{\sqrt{19.8}}\right)$

$= P\left(Z < \frac{-2}{\sqrt{19.8}}\right) = 1 - P(Z < 0.45)$

$= 1 - 0.6736 = 0.3264$

[8 marks]

$n=50$

(c) The average speed of vehicles on a highway is being studied. Suppose observations on 50 vehicles yielded a sample mean of 65 mph (miles per hour) and the standard deviation of vehicle speed is known to be 6 mph. $\bar{x}=65$ $\hat{S}_x = 6$

so standard deviation is known → normal distribution

(i) Assume that the population standard deviation of vehicle speed is well estimated by the sample standard deviation. Determine the two-sided 99% confidence intervals of the mean speed.

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$P(-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) = 0.99$

$P(\bar{x} - z_{\alpha/2}\frac{6}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{6}{\sqrt{n}}) = 0.99$

$P(z < z_{\alpha/2}) = 0.995$

$\therefore z_{\alpha/2} = 2.575$

$CI = 65 \pm 2.575\left(\frac{6}{\sqrt{50}}\right) = [62.815, 67.185]$

(ii) From part (i), how many additional vehicles' speeds should be observed such that the mean speed can be estimated to be within ±1 mph with 99% confidence?

this means $CI = \bar{x} \pm 1 = [64, 66]$

we know $CI = \bar{x} \pm z_{\alpha/2}\left(\frac{6}{\sqrt{n}}\right)$

$z_{\alpha/2}\frac{6}{\sqrt{n}} = 1$

$z_{\alpha/2}6 = \sqrt{n}$

$n = 2.575^2 \times 6^2 = 238.7025$

$\therefore$ additionally, $239 - 50 = 189$ //

(iii) Without the assumption that we know the population standard deviation, determine the two-sided 99% confidence intervals of the mean speed. Briefly comment on the key differences in results between (i) and (iii).

[12 marks]

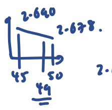
Cannot estimate σ^2 with s^2 → t-test

$\frac{\bar{x}-\mu}{S_x/\sqrt{n}} \sim t(n-1)$

$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 0.99$

$P(T < t_{\alpha/2}) = 0.995$

$t_{\alpha/2} = 2.6804$



$2.690 - \frac{2.690 - 2.678}{50 - 45} \times (49 - 45)$

$CI = \bar{x} \pm t_{\alpha/2}\frac{\hat{S}_x}{\sqrt{n}}$

$= 65 \pm 2.6804\left(\frac{6}{\sqrt{50}}\right)$

$= [62.7256, 67.2744]$ compared to CI from (i) $= [62.815, 67.185]$

→ the range is slightly bigger due to uncertainty of s.d.

Question 5

(a) Two measuring campaigns have been conducted to estimate the concentration of chemical by-products in a lake. In the first campaign, sample A with 100 measurements had a mean concentration of 2.4 mg/m^3 and a standard deviation of 0.2 mg/m^3. In the second campaign, sample B with 25 measurements had a mean concentration of 2.6 mg/m^3 and a standard deviation of 0.26 mg/m^3.

Test with 99% confidence the assumption that the two campaigns have taken samples from the same population of contaminants, assuming that the sample sizes are large enough to describe the population standard deviations.

Standard deviation is known → Normal-distribution.

both population mean is not known → symmetric case (two-sample)

(two-tailed)

$H_0: \mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0$

$H_1: \mu_A \neq \mu_B \rightarrow \mu_A - \mu_B \neq 0$

$n_A = 100,\ \bar{x}_A = 2.4,\ \sigma_A = 0.2$

$n_B = 25,\ \bar{x}_B = 2.6,\ \sigma_B = 0.26$

$$\frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0,1)$$

[5 marks]

$z = \frac{-0.2 - 0}{0.0557} = -3.5898$

$P(Z > z_{\alpha/2}) = 0.005$

$P(Z < z_{\alpha/2}) = 0.995$

$z_{\alpha/2} = 2.575 \ (-z_{\alpha/2} = -2.575)$

since $z < -z_{\alpha/2}$, reject H_0
→ not same population

(b) The marks of a recent Statistics test taken by 200 students were summarised in the following table:

$H_0: X \sim N(60, 15^2)$ is a good fit

$H_1: X \sim N(60, 15^2)$ is NOT a good fit

Mark intervals	[20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]	(80,90]	(90,100]
observed freq. O_i: NO. of students	5	15	30	51	60	23	10	6
Expected freq. E_i:	4.56	13.8	31.92	49.72	49.72	31.92	13.8	4.56

(i) Based on these 200 observations, is a normal distribution with $\mu = 60$ and $\sigma = 15$ an appropriate model? Perform a Chi-square goodness-of-fit procedure at the 1% significance level.

(ii) Briefly discuss the theoretical differences between the Chi-square test and Kolmogorov-Smirnov (KS) goodness-of-fit test.

[10 marks]

working for (i):

although first bin is $20 \le x \le 30$,
but we usually start from 0!

$E_1 = 200 P(X \le 30)$
$= 200 P(Z \le \frac{30-60}{15})$
$= 200 P(Z \le -2)$
$= 200 (1 - P(Z \le 2))$
$= 200 (1 - 0.9772)$
$= 4.56$

E_2 (second bin) $= 200 P(30 \le X \le 40)$
$= 200 P(\frac{30-60}{15} \le x \le \frac{40-60}{15})$
$= \ldots$
$= 13.8.$

$$C_n = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i}$$

$= \frac{(5-4.56)^2}{4.56} + \frac{(15-13.8)^2}{13.8} + \ldots + \frac{(6-4.56)^2}{4.56}$

$= 6.4145$

$C_n \sim \chi^2(8-1-2)$

$X \sim N(\underbrace{\mu, \sigma^2}_{2 \text{ parameters.}})$

$P(\chi^2 < \chi_\alpha^2) = 0.99$

$\chi_\alpha^2 = 15.086$

(6.4145) (15.086)

$\therefore C_n < \chi^2_{5,0.99}$ cannot reject H_0 (good fit)

(c) Ten steel bars were tested under tension in the laboratory. The purpose of the experiment was to derive a model that will describe the length increase of the steel bars under forces of varying magnitude. The recorded results are shown in the following table in standard units:

Force (N)	10	16	25	22	42	48	54	56	68	90
Length increase (mm)	1.8	2.2	2.3	3.4	3.8	2.7	6.1	5.7	6.7	7.2

(i) Estimate the parameters of a simple linear regression model with Force as an explanatory variable (variable X), and length increase as the variable to be explained (variable Y), indicating the coefficient of determination.

(ii) Find 95% confidence limits for the two regression parameters and test the hypothesis that the slope is zero. $\bar{x} = 43.1$, $\bar{y} = 4.19$

[10 marks]

(i) $\hat{\beta} = \dfrac{cov_{x,y}}{var_x} = \dfrac{46.789}{636.98} = 0.0734$

$cov_{x,y} = \dfrac{1}{10-1}\sum(x-\bar{x})(y-\bar{y}) = \dfrac{421.01}{9} = 46.789$

$var_x = 636.9\overline{8}$

should all use population parameter! (biased)

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 4.19 - 0.0734(43.1) = 1.0248$

$r^2 = (c_{x,y})^2 = \left(\dfrac{cov_{x,y}}{S_x S_y}\right)^2 = \dfrac{46.789^2}{636.98 \times 4.1699} = 0.8242$

(ii) $\widehat{Var}(Z)$ is not given and hence need to find it and use t-distribution!

Test that the slope = 0 or find confidence interval for the slope with: $T = \dfrac{\hat{\beta} - \beta}{\sqrt{\dfrac{\widehat{Var}(Z)}{n\,\mathrm{var}_x}}} \sim t(n-2)$

Find a confidence interval for the constant parameter using: $\dfrac{\hat{\alpha} - \alpha}{\sqrt{\widehat{Var}(Z)\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{n\,\mathrm{var}_x}\right]}} \sim t(n-2)$

$P(T < t_{n-2,\alpha/2}) = 0.025$

$\alpha = \hat{\alpha} \pm t_{n-2,\alpha/2}\sqrt{\widehat{Var}(Z)\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{n\,var_x}\right]}$

$\widehat{Var}(Z) = \dfrac{1}{n-2}\left(n\,\mathrm{var}_y - n\dfrac{\mathrm{cov}^2_{x,y}}{\mathrm{var}_x}\right)$

$\beta = \hat{\beta} \pm t_{n-2,\alpha/2}\sqrt{\dfrac{\widehat{Var}(Z)}{n\,var_x}}$

END OF PAPER

$H_0: \beta = 0$

$H_1: \beta \neq 0$

$T = \dfrac{\hat{\beta} - \beta}{\sqrt{\dfrac{\widehat{Var}(Z)}{n\,var_x}}}$

(c) Ten steel bars were tested under tension in the laboratory. The purpose of the experiment was to derive a model that will describe the length increase of the steel bars under forces of varying magnitude. The recorded results are shown in the following table in standard units:

Force (N) X	10	16	25	22	42	48	54	56	68	90
Length increase (mm) Y	1.8	2.2	2.3	3.4	3.8	2.7	6.1	5.7	6.7	7.2

(i) Estimate the parameters of a simple linear regression model with Force as an explanatory variable (variable X), and length increase as the variable to be explained (variable Y), indicating the coefficient of determination.

(ii) Find 95% confidence limits for the two regression parameters and test the hypothesis that the slope is zero.

sample.

[10 marks]

6. Linear Regression → α, β

For: $\boxed{Y = \alpha + \beta X} + Z$ the parameter estimates are: $\boxed{\hat{\beta} = \frac{\text{cov}_{x,y}}{\text{var}_x}}$ and $\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$

Coefficient of determination: $r^2 = \frac{\text{var}_{\hat{\alpha}+\hat{\beta}x}}{\text{var}_y}$ or $r^2 = (c_{x,y})^2$

Estimate of the error variance: $\hat{Var}(Z) = \frac{1}{n-2}\left(n\,\text{var}_y - n\frac{\text{cov}^2_{x,y}}{\text{var}_x}\right)$

$$\boxed{\begin{aligned}\hat{\alpha} &= 1.025\\ \hat{\beta} &= 0.073\\ r &= 0.908\end{aligned}}$$

0.07344

$$\bar{x} = \frac{431}{10} = 43.1, \quad \bar{y} = \frac{41.9}{10} = 4.19$$

$$\text{cov}_{x,y} = \frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y}) = 1.8424.$$

$$\hat{\beta} = \frac{\text{cov}_{x,y}}{\text{var}_x} = \frac{1.8424}{25.2386} = 0.073$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 4.19 - 0.073(43.1) = 1.025 \rightarrow 1.0247$$

END OF PAPER

NORMAL CUMULATIVE DISTRIBUTION FUNCTION

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

STUDENT'S t Table

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

CHI-SQUARED Table - 1/2

ν	0.1%	0.5%	1.0%	2.5%	5.0%	10.0%	12.5%	20.0%	25.0%	33.3%	50.0%
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335
55	28.173	31.735	33.570	36.398	38.958	42.060	43.220	46.036	47.610	49.972	54.335
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335

CHI-SQUARED Table - 2/2

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

Kolmogorov-Smirnov Test (level of significance a)

n	0.2	0.1	0.05	0.02	0.01
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

FORMULA SHEET STATISTICS CIVE50008

1. Discrete distributions pmf

Bernouilli *B(1,p)*: $P(X=k)=p^k(1-p)^{1-k}$ for $k\in\{0,1\}$; $E(X)=p; Var(x)=p(1-p)$

Binomial *B(n,p)*: $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ for $k\in\{0,...,n\}$; $E(X)=np; Var(x)=np(1-p)$

Poisson *P(μ)*: $P(X=k)=\dfrac{\mu^k e^{-\mu}}{k!}$ for $k=0,1,...$; $E(X)=\mu; Var(x)=\mu$

Geometric *Geom(p)*: $P(X=k)=p(1-p)^k$ for $k=0,1,..$; $E(X)=\dfrac{(1-p)}{p}; Var(X)=\dfrac{(1-p)}{p^2}$.

2. Continuous distributions pdf or definition

Uniform *U[a,b]*: $f_X(x)=\begin{cases}\dfrac{1}{b-a} & \text{if } a<x<b\\ 0 & \text{otherwise}\end{cases}$. Exponential *Exp(λ)*: $f_X(x)=\begin{cases}\lambda e^{-\lambda x} & \text{if } 0<x\\ 0 & \text{otherwise}\end{cases}$

Gamma *Gamma(α,β)*: $f_X(x)=\begin{cases}\dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)} & \text{if } 0<x\\ 0 & \text{otherwise}\end{cases}$. Normal $N(\mu,\sigma^2)$: $f_X(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Chi-square $\chi^2(n)$: $Y=X_1^2+X_2^2+...+X_n^2$ where $X_i \sim N(0,1)$ and independent; $Y=Gamma(n/2,2)$

Student t distribution $t(n)$: $\left.\begin{array}{r} Z\sim N(0,1)\\ Y\sim\chi^2(n)\\ Y,Z \text{ independent}\end{array}\right\}\Rightarrow \dfrac{Z}{\sqrt{Y/n}}\sim t(n)$

3. Sum of two distributions

$E(X+Y)=E(X)+E(Y);\ Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$

Assuming independence of the added distributions:

$$K=I+J\Rightarrow p_K(k)=\sum_{j=0}^{k}p_I(k-j)p_J(j);\quad Z=X+Y\Rightarrow f_Z(z)=\int_{-\infty}^{+\infty}f_X(z-y)f_Y(y)dy$$

$I\sim Poiss(\lambda)\wedge J\sim Pois(\mu)\Rightarrow I+J\sim Pois(\lambda+\mu)$

$X\sim Gamma(\alpha_1,\beta)\wedge Y\sim Gamma(\alpha_2,\beta)\Rightarrow X+Y\sim Gamma(\alpha_1+\alpha_2,\beta)$

$X\sim N(\mu_1,\sigma_1^2)\wedge Y\sim N(\mu_2,\sigma_2^2)\Rightarrow X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$

4. Confidence intervals and tests for means and variances

If X is normally distributed (n: sample size), $T=\dfrac{\overline{X}-\mu}{\hat{S}_x/\sqrt{n}}\sim t(n-1)$, & $V=\dfrac{(n-1)\hat{S}_x^2}{\sigma^2}\sim\chi^2(n-1)$

Test that μ_B is the population mean with: $Y=\dfrac{\overline{X_B}-\mu_B}{\sigma/\sqrt{n}}\sim N(0,1)$, or $T=\dfrac{\overline{X_B}-\mu_B}{\hat{S}_x/\sqrt{n}}\sim t(n-1)$

Test that $\mu_B=\mu_A$ with $Z=\dfrac{(\overline{X}_A-\overline{X}_B)-(\mu_A-\mu_B)}{\sigma_{\overline{X}_A-\overline{X}_B}}$ where: $\sigma_{\overline{X}_A-\overline{X}_B}=\sqrt{\dfrac{\sigma_A^2}{n_A}+\dfrac{\sigma_B^2}{n_B}}$ and $Z\sim N(0,1)$

or with: $T=\dfrac{(\overline{X}_A-\overline{X}_B)-(\mu_A-\mu_B)}{\hat{s}_{x_A,x_B}\sqrt{1/n_A+1/n_B}}$ where: $\hat{s}^2_{x_A,x_B}=\frac{(n_A-1)\hat{s}^2_{xA}+(n_B-1)\hat{s}^2_{xB}}{n_A+n_B-2}$ and $T\sim t(n_A+n_B-2)$

or with: $T = \dfrac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{\dfrac{\hat{s}_{x_A}^2}{n_A} + \dfrac{\hat{s}_{x_B}^2}{n_B}}}$ and $T \sim t(\nu)$ where: $\nu \approx \dfrac{\left(\dfrac{\hat{s}_{x_A}^2}{n_A} + \dfrac{\hat{s}_{x_B}^2}{n_B}\right)^2}{\dfrac{1}{n_A - 1}\left(\dfrac{\hat{s}_{x_A}^2}{n_A}\right)^2 + \dfrac{1}{n_B - 1}\left(\dfrac{\hat{s}_{x_B}^2}{n_B}\right)^2}$

5. Goodness-of-fit tests

Chi-squared: $C_n = \sum_{i=1}^{l} \dfrac{(O_i - E_i)^2}{E_i}$ (l : number of bins) ; $\lim_{n\to\infty} C_n \sim \chi^2_{l-1-k}$ (k : number of parameters)

Kolmogorov-Smirnov: $D_n = \sup_x \{F_n(x) - F_X(x)\}$; $\lim_{n\to\infty} P(\sqrt{n} D_n \le z) = \dfrac{\sqrt{2\pi}}{z} \sum_{k=1}^{\infty} \exp\left[\dfrac{-(2k-1)^2 \pi^2}{8z^2}\right]$

6. Linear Regression

For: $Y = \alpha + \beta X + Z$ the parameter estimates are: $\hat{\beta} = \dfrac{\mathrm{cov}_{x,y}}{\mathrm{var}_x}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Coefficient of determination: $r^2 = \dfrac{\mathrm{var}_{\hat{\alpha}+\hat{\beta}x}}{\mathrm{var}_y}$ or $r^2 = (c_{x,y})^2$

Estimate of the error variance: $\hat{Var}(Z) = \dfrac{1}{n-2}\left(n\,\mathrm{var}_y - n\dfrac{\mathrm{cov}_{x,y}^2}{\mathrm{var}_x}\right)$

Test that the slope = 0 or find confidence interval for the slope with: $T = \dfrac{\hat{\beta} - \beta}{\sqrt{\dfrac{\hat{Var}(Z)}{n\,\mathrm{var}_x}}} \sim t(n-2)$

Find a confidence interval for the constant parameter using: $\dfrac{\hat{\alpha} - \alpha}{\sqrt{\hat{Var}(Z)\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{n\,\mathrm{var}_x}\right]}} \sim t(n-2)$

Find a confidence interval for the mean of Y when $X=x_0$ using: $\dfrac{\hat{\alpha} + \hat{\beta} x_0 - (\alpha + \beta x_0)}{\sqrt{\hat{Var}(Z)\left[\dfrac{1}{n} + (x_0 - \bar{x})^2 \dfrac{1}{n\,\mathrm{var}_x}\right]}} \sim t(n-2)$

Find a confidence interval for a predicted value Y when $X=x_0$ using:

$$\dfrac{\hat{\alpha} + \hat{\beta} x_0 - (\alpha + \beta x_0 \mp z_0)}{\sqrt{\hat{Var}(Z)\left[\dfrac{1}{n} + (x_0 - \bar{x})^2 \dfrac{1}{n\,\mathrm{var}_x} + 1\right]}} \sim t(n-2)$$

Test that the correlation is 0 using: $V = \dfrac{\sqrt{n-3}}{2} \ln\left(\dfrac{(1 + c_{x,y})(1 - Corr(X,Y))}{(1 - c_{x,y})(1 + Corr(X,Y))}\right) \sim N(0,1)$