Imperial College London



Course: M2SJ
Setter: Battey (Q4), Lau (Q1-Q3)
Checker: Fitz-Simon
Editor: Walden
External: Jennison
Date: March 8, 2017

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May – June 2017

# M2SJ

# Statistical Methods

| Setter's signature | Checker's signature | Editor's signature |
|---|---|---|
| .................. | .................. | .................. |

Imperial College London

BSc, MSc and MSci EXAMINATIONS (MATHEMATICS)

May – June 2017

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science.

Statistical Methods

Date: ??

Time: ??

Time Allowed: 2 Hours

This paper has *4* Questions.

Candidates should start their solutions to each question in a new main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

Formula sheets and statistical tables are available on pages 6 to 11

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers to each question as per the table below.

| Raw Mark | Up to 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| Extra Credit | 0 | $1/2$ | 1 | $1\,1/2$ | 2 | $2\,1/2$ | 3 | $3\,1/2$ | 4 |

- Each question carries equal weight.
- Calculators may be used.

1. *Choose one answer for each part. Partial credit will be awarded for working if an incorrect answer is selected. There is no negative marking.*

(i) If you were to randomly guess the answer for Question 1 parts (i) to (v) (with no working) where each correct answer is worth 1 mark and any other answer is worth 0 marks, the total expected number of marks for Question 1 is

(a) 1/6; (b) 4/5; (c) 5/6; (d) 1; (e) 6/5; (f) 3?

(ii) Suppose $X$ and $Y$ are jointly continuous random variables with joint probability density given by

$$f(x, y) = c \exp\left\{-\frac{x^2 + y^2}{4}\right\}, \quad x \in \mathbb{R},\ y \in \mathbb{R},$$

where $c$ is a constant which does not depend on $x$ or $y$. Is $c$

(a) $2\sqrt{\pi}$; (b) $\frac{1}{2\sqrt{\pi}}$; (c) $\frac{1}{\sqrt{2\pi}}$; (d) $4\pi$; (e) $\frac{1}{4\pi}$; (f) $\frac{1}{8\pi}$?

(iii) Suppose $X_1, X_2, \ldots, X_n$ are $n$ independent random variables which all follow the same distribution $P_X$ which has mean $\mu$ and variance $\sigma^2$. Let $\overline{X} = \sum_{i=1}^{n} X_i/n$ be the mean of the $n$ samples. By the Central Limit Theorem, an approximate distribution for $\frac{\overline{X} - \mu}{3}$ is

(a) $N\left(\frac{\mu}{3}, \frac{\sigma^2}{3}\right)$; (b) $N\left(\frac{n\mu}{3}, \frac{n\sigma^2}{3}\right)$; (c) $N\left(\frac{n\mu}{\sqrt{3}}, \frac{\sigma^2}{\sqrt{3}n}\right)$;
(d) $N\left(0, \frac{n\sigma^2}{3}\right)$; (e) $N\left(0, \frac{\sigma^2}{\sqrt{3}n}\right)$; (f) $N\left(0, \frac{\sigma^2}{9n}\right)$?

(iv) The power of a hypothesis test is

(a) the probability of rejecting the null hypothesis;
(b) the probability of rejecting the alternate hypothesis;
(c) the probability of not rejecting the alternate hypothesis;
(d) the probability of not rejecting the alternate hypothesis given the alternate hypothesis is true;
(e) the probability of rejecting the null hypothesis given the null hypothesis is true;
(f) None of the above.

(v) Events $A$ and $B$ have probabilities $P(A) = 0.4$, $P(B) = 0.8$ and $P(\overline{A}|B) = 0.7$. What is $P(\overline{B}|A)$?

(a) 0.3; (b) 0.4; (c) 0.5; (d) 0.6; (e) 0.7; (f) 0.8?

2. In a production factory, machines $A$, $B$ and $C$ are all producing compact discs (CDs). Machine $A$ produces $55\%$ of the CDs, machine $B$ produces $15\%$ and the rest are produced by machine $C$. Of their production of CDs, machines $A$, $B$ and $C$ produce $2\%$, $4\%$ and $5\%$ defective CDs respectively.

(i) Find the probability that a randomly selected CD is

(a) produced by machine $B$ and is defective,

(b) is defective.

(ii) Given that a randomly selected CD is defective, find the probability that is was produced by machine A.

A separate machine, $D$, produces CD racks of height $X$ centimeters where $X$ is a discrete random variable with probability mass function

$$P(X = x) = \begin{cases} 0.1 & \text{for } x = 1 \\ \alpha & \text{for } x = 2 \\ 0.2 & \text{for } x = 3 \\ \beta & \text{for } x = 4 \\ 0.3 & \text{for } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

(iii) Given that $E(X) = 3.3$, find the value of $\alpha$ and $\beta$.

(iv) Find Var$(X)$ and the skewness of $X$.

3. A company produces an energy drink called NoSleep that contains caffeine. Let the random variable $X$ represent the number of milligrams (mg) of caffeine in a single can of NoSleep. $X$ is Normally distributed with mean $80$ and standard deviation $10$. You may assume that the amounts of caffeine in different cans are independent. All cans referred to in this question are of the same size.

(i) Show that $P(X \geq 75) = 0.691$.

(ii) These energy drinks are sold in packs of $6$. Find the probability that the amounts of caffeine of exactly $4$ of the $6$ cans in a randomly chosen pack are below $75$mg.

(iii) Using a suitable approximating distribution, find the probability that the amounts of caffeine of at least $35$ out of $100$ randomly selected cans are below $75$mg.

Another company claims that the average amount of caffeine of its energy drink AlwaysUp is $100$mg in a can. A consumer organisation suspects that the true figure may be lower than this. The amounts of caffeine of a random sample of $100$ of these cans are measured. Note that we do not know if the amount of caffeine in a can of AlwaysUp is Normally distributed. A hypothesis test is then carried out to check the claim.

(iv) Write down a suitable null hypothesis and explain briefly why the alternative hypothesis should be $H_1 : \mu < 100$. State the meaning of $\mu$.

(v) Suppose we know that the standard deviation of the amount of caffeine in AlwaysUp cans is $30$mg and that the sample mean amount of caffeine of the sample of $100$ cans is $95$mg. Suppose we want to conduct the hypothesis test referred to in (iv) at the $100\alpha\%$ significance level.

(a) Clearly state the appropriate test statistic and any associated distribution. Also construct a suitable rejection region $R$ for the test.

(b) What is the conclusion of the hypothesis test using a $5\%$ significance level?

4. (i) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables each with moment generating function $M_X(t)$. Demonstrate that the moment generating function of $Z = \frac{1}{n}\sum_{i=1}^n X_i$ is $M_Z(t) = [M_X(t/n)]^n$.

(ii) Let $(X_1, X_2, \ldots, X_n)$ be a random sample of size $n$ from a $N(\mu, \sigma^2)$ distribution and let $(Y_1, Y_2, \ldots, Y_m)$ be a random sample of size $m$ from a $N(2\mu, \sigma^2)$ distribution.

(a) Assuming that $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ are independent random variables, show that $\widehat{\mu}$ is an unbiased estimator of $\mu$, where

$$\widehat{\mu} = \frac{\sum_{i=1}^n X_i + 2\sum_{j=1}^m Y_j}{n + 4m}.$$

(b) Derive the variance of $\widehat{\mu}$ and use Markov's inequality to show that $\widehat{\mu}$ is consistent, i.e. $\widehat{\mu}$ converges in probability to $\mu$.

(iii) Let $\delta > 0$, let $C$ and $D$ be finite positive constants, and let $f_n$ and $g_n$ be deterministic sequences converging to zero. Let $T_n$ and $S_n$ be two statistics such that $\mathbb{P}(|T_n| > Cf_n) < \frac{1}{2}\delta$ for all $n > n_1$ and $\mathbb{P}(|S_n| > Dg_n) < \frac{1}{2}\delta$ for all $n > n_2$

(a) Use the axioms of probability to demonstrate that

$$\mathbb{P}(|T_n| \times |S_n| > Cf_n Dg_n) < \delta \text{ for all } n > \max\{n_1, n_2\}.$$

**Hint:** Write the event $\{|T_n| \times |S_n| > Cf_nDg_n\}$ as

$$\Big\{\{|T_n| \times |S_n| > Cf_nDg_n\} \cap \{|S_n|/Dg_n > 1\}\Big\} \\ \cup \Big\{\{|T_n| \times |S_n| > Cf_nDg_n\} \cap \{|S_n|/Dg_n \leq 1\}\Big\}$$

(b) Use this result to show that $\mathbb{P}(|T_nS_n| > Cf_nDg_n) < \delta$ for all $n > \max\{n_1, n_2\}$.

# STATISTICS FORMULA SHEET

## 1. Probabilities for events

For events $A$, $B$, and $C$ $\quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$

More generally $\quad P(\bigcup A_i) = \sum P(A_i) - \sum P(A_i \cap A_j) + \sum P(A_i \cap A_j \cap A_k) - \cdots$

The odds in favour of $A$ $\quad P(A) \,/\, P(\overline{A})$

Conditional probability $\quad P(A \,|\, B) = \dfrac{P(A \cap B)}{P(B)}$ provided that $P(B) > 0$

Chain rule $\quad P(A \cap B \cap C) = P(A)\, P(B \,|\, A)\, P(C \,|\, A \cap B)$

Bayes' rule $\quad P(A \,|\, B) = \dfrac{P(A)\, P(B \,|\, A)}{P(A)\, P(B \,|\, A) \;+\; P(\overline{A})\, P(B \,|\, \overline{A})}$

$A$ and $B$ are independent if $\quad P(B \,|\, A) = P(B)$

$A$, $B$, and $C$ are independent if $\quad P(A \cap B \cap C) = P(A)P(B)P(C)$, and

$$P(A \cap B) = P(A)P(B), \quad P(B \cap C) = P(B)P(C), \quad P(C \cap A) = P(C)P(A)$$

## 2. Probability distribution, expectation and variance

The probability distribution for a discrete random variable $X$ is called the probability mass function (pmf) and is the complete set of probabilities $\{p_x\} = \{P(X = x)\}$

Expectation $\quad E(X) = \mu = \sum_x x p_x$

For function $g(x)$ of $x$, $\quad E\{g(X)\} = \sum_x g(x) p_x$, so $E(X^2) = \sum_x x^2 p_x$

Sample mean $\quad \overline{x} = \dfrac{1}{n} \sum_k x_k$ estimates $\mu$ from random sample $x_1, x_2, \ldots, x_n$

Variance $\quad \mathrm{var}\,(X) = \sigma^2 = E\{(X - \mu)^2\} = E(X^2) - \mu^2$

Sample variance $\quad s^2 = \dfrac{1}{n-1} \left\{ \sum_k x_k^2 - \dfrac{1}{n} \left( \sum_j x_j \right)^2 \right\}$ estimates $\sigma^2$

Standard deviation $\quad \mathrm{sd}\,(X) = \sigma$

If value $y$ is observed with frequency $n_y$

$$n = \sum_y n_y, \quad \sum_k x_k = \sum_y y n_y, \quad \sum_k x_k^2 = \sum_y y^2 n_y$$

Skewness $\quad \beta_1 = E\left(\dfrac{X - \mu}{\sigma}\right)^3$ is estimated by $\dfrac{1}{n-1} \sum \left(\dfrac{x_i - \overline{x}}{s}\right)^3$

Kurtosis $\quad \beta_2 = E\left(\dfrac{X - \mu}{\sigma}\right)^4 - 3$ is estimated by $\dfrac{1}{n-1} \sum \left(\dfrac{x_i - \overline{x}}{s}\right)^4 - 3$

Sample median $\quad \widetilde{x}$ or $x_{\text{med}}$. Half the sample values are smaller and half larger

If the sample values $x_1, \ldots, x_n$ are ordered as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$,
then $\widetilde{x} = x_{(\frac{n+1}{2})}$ if $n$ is odd, and $\widetilde{x} = \frac{1}{2}\,(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})})$ if $n$ is even

$\alpha$-quantile $Q(\alpha)$ is such that $P(X \le Q(\alpha)) = \alpha$

Sample $\alpha$-quantile $\widehat{Q}(\alpha)$ Proportion $\alpha$ of the data values are smaller

Lower quartile Q1 $= \widehat{Q}(0.25)$ one quarter are smaller

Upper quartile Q3 $= \widehat{Q}(0.75)$ three quarters are smaller

Sample median $\widetilde{x} = \widehat{Q}(0.5)$ estimates the population median $Q(0.5)$

## 3. Probability distribution for a continuous random variable

The cumulative distribution function (cdf) $$F(x) = P(X \le x) = \int_{x_0=-\infty}^{x} f(x_0)\mathrm{d}x_0$$

The probability density function (pdf) $$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

$$E(X) = \mu = \int_{-\infty}^{\infty} x\, f(x)\mathrm{d}x\,, \quad \mathrm{var}\,(X) = \sigma^2 = E(X^2) - \mu^2, \text{ where } E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)\mathrm{d}x$$

## 4. Discrete probability distributions

Discrete Uniform *Uniform* $(n)$

$$p_x = \frac{1}{n} \quad (x = 1, 2, \ldots, n) \qquad \mu = (n+1)/2, \;\; \sigma^2 = (n^2-1)/12$$

Binomial distribution *Binomial* $(n, \theta)$

$$p_x = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad (x = 0, 1, 2, \ldots, n) \qquad \mu = n\theta\,, \;\; \sigma^2 = n\theta(1-\theta)$$

Poisson distribution *Poisson* $(\lambda)$

$$p_x = \frac{\lambda^x e^{-\lambda}}{x!} \quad (x = 0, 1, 2, \ldots) \quad (\text{with } \lambda > 0) \qquad \mu = \lambda\,, \;\; \sigma^2 = \lambda$$

Geometric distribution *Geometric* $(\theta)$

$$p_x = (1-\theta)^{x-1}\theta \quad (x = 1, 2, 3, \ldots) \qquad \mu = \frac{1}{\theta}\,, \;\; \sigma^2 = \frac{1-\theta}{\theta^2}$$

## 5. Continuous probability distributions

Uniform distribution *Uniform* $(\alpha, \beta)$

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & (\alpha < x < \beta), \\ 0 & (\text{otherwise}). \end{cases} \qquad \mu = (\alpha+\beta)/2\,, \;\; \sigma^2 = (\beta-\alpha)^2/12$$

Exponential distribution *Exponential* $(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (0 < x < \infty), \\ 0 & (-\infty < x \le 0). \end{cases} \qquad \mu = 1/\lambda\,, \;\; \sigma^2 = 1/\lambda^2$$

Normal distribution $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} \quad (-\infty < x < \infty), \qquad E(X) = \mu\,, \quad \mathrm{var}\,(X) = \sigma^2$$

Standard normal distribution $N(0,1)$

If $X$ is $N(\mu, \sigma^2)$, then $Y = \dfrac{X-\mu}{\sigma}$ is $N(0,1)$

## 6. Reliability

For a device in continuous operation with failure time random variable $T$ having pdf $f(t)$ $(t > 0)$

| | | | |
|---|---|---|---|
| The reliability function at time $t$ | $R(t)$ | $=$ | $P(T > t)$ |
| The failure rate or hazard function | $h(t)$ | $=$ | $f(t)/R(t)$ |
| The cumulative hazard function | $H(t)$ | $=$ | $\int_0^t h(t_0)\,\mathrm{d}t_0 = -\ln\{R(t)\}$ |
| The Weibull($\alpha, \beta$) distribution has | $H(t)$ | $=$ | $\beta t^\alpha$ |

## 7. System reliability

For a system of $k$ devices, which operate independently, let

$R_i = P(D_i) = P($"device $i$ operates"$)$

The system reliability, $R$, is the probability of a path of operating devices

A system of devices in series operates only if every device operates

$$R = P(D_1 \cap D_2 \cap \cdots \cap D_k) = R_1 R_2 \cdots R_k$$

A system of devices in parallel operates if any device operates

$$R = P(D_1 \cup D_2 \cup \cdots \cup D_k) = 1 - (1 - R_1)(1 - R_2) \cdots (1 - R_k)$$

## 8. Covariance and correlation

The covariance of $X$ and $Y$ $\quad \mathrm{cov}\,(X, Y) = E(XY) - \{E(X)\}\{E(Y)\}$

From pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ $\quad S_{xy} = \sum_k x_k y_k - \frac{1}{n}\left(\sum_i x_i\right)\left(\sum_j y_j\right)$

$$S_{xx} = \sum_k x_k^2 - \frac{1}{n}\left(\sum_i x_i\right)^2, \qquad S_{yy} = \sum_k y_k^2 - \frac{1}{n}\left(\sum_j y_j\right)^2$$

| | | | |
|---|---|---|---|
| Sample covariance | $s_{xy}$ | $= \dfrac{1}{n-1} S_{xy}$ | estimates $\mathrm{cov}\,(X, Y)$ |
| Correlation coefficient | $\rho$ | $= \mathrm{corr}\,(X, Y)$ | $= \dfrac{\mathrm{cov}\,(X, Y)}{\mathrm{sd}\,(X) \cdot \mathrm{sd}\,(Y)}$ |
| Sample correlation coefficient | $r$ | $= \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ | estimates $\rho$ |

## 9. $\boxed{\text{Sums of random variables}}$

$$
\begin{aligned}
E(X+Y) &= E(X)+E(Y)\\
\operatorname{var}(X+Y) &= \operatorname{var}(X)+\operatorname{var}(Y)+2\operatorname{cov}(X,Y)\\
\operatorname{cov}(aX+bY,\ cX+dY) &= (ac)\operatorname{var}(X)+(bd)\operatorname{var}(Y)+(ad+bc)\operatorname{cov}(X,Y)
\end{aligned}
$$

If $X$ is $N(\mu_1,\sigma_1^2)$, $Y$ is $N(\mu_2,\sigma_2^2)$, and $\operatorname{cov}(X,Y)=c$, then $X+Y$ is $N(\mu_1+\mu_2,\ \sigma_1^2+\sigma_2^2+2c)$

## 10. $\boxed{\text{Bias, standard error, mean square error}}$

If $t$ estimates $\theta$ (with random variable $T$ giving $t$)

| | | | |
|---|---|---|---|
| Bias of $t$ | $\operatorname{bias}(t)$ | $=$ | $E(T)-\theta$ |
| Standard error of $t$ | $\operatorname{se}(t)$ | $=$ | $\operatorname{sd}(T)$ |
| Mean square error of $t$ | MSE$(t)$ | $=$ | $E\{(T-\theta)^2\} = \{\operatorname{se}(t)\}^2+\{\operatorname{bias}(t)\}^2$ |

If $\overline{x}$ estimates $\mu$, then $\operatorname{bias}(\overline{x})=0$, $\operatorname{se}(\overline{x})=\sigma/\sqrt{n}$, MSE$(\overline{x})=\sigma^2/n$, $\widehat{\operatorname{se}}(\overline{x})=s/\sqrt{n}$

Central limit property If $n$ is fairly large, $\overline{x}$ is from $N(\mu,\ \sigma^2/n)$ approximately

## 11. $\boxed{\text{Likelihood}}$

The likelihood is the joint probability as a function of the unknown parameter $\theta$.

For a random sample $x_1,x_2,\ldots,x_n$

$$\ell(\theta;\, x_1,x_2,\ldots,x_n) = P(X_1=x_1\,\big|\,\theta)\ \cdots\ P(X_n=x_n\,\big|\,\theta) \qquad \text{(discrete distribution)}$$

$$\ell(\theta;\, x_1,x_2,\ldots,x_n) = f(x_1\,\big|\,\theta)\,f(x_2\,\big|\,\theta)\cdots f(x_n\,\big|\,\theta) \qquad \text{(continuous distribution)}$$

The maximum likelihood estimator (MLE) is $\widehat{\theta}$ for which the likelihood is a maximum

## 12. $\boxed{\text{Confidence intervals}}$

If $x_1,x_2,\ldots,x_n$ are a random sample from $N(\mu,\sigma^2)$ and $\sigma^2$ is known, then the 95% confidence interval for $\mu$ is $(\overline{x}-1.96\dfrac{\sigma}{\sqrt{n}},\ \overline{x}+1.96\dfrac{\sigma}{\sqrt{n}})$

If $\sigma^2$ is estimated, then from the Student t table for $t_{n-1}$ we find $t_0=t_{n-1,0.05}$

The 95% confidence interval for $\mu$ is $(\overline{x}-t_0\dfrac{s}{\sqrt{n}},\ \overline{x}+t_0\dfrac{s}{\sqrt{n}})$

13. Standard normal table — Values of pdf $\phi(y) = f(y)$ and cdf $\Phi(y) = F(y)$

| $y$ | $\phi(y)$ | $\Phi(y)$ | $y$ | $\phi(y)$ | $\Phi(y)$ | $y$ | $\phi(y)$ | $\Phi(y)$ | $y$ | $\Phi(y)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .399 | .5 | .9 | .266 | .816 | 1.8 | .079 | .964 | 2.8 | .997 |
| .1 | .397 | .540 | 1.0 | .242 | .841 | 1.9 | .066 | .971 | 3.0 | .999 |
| .2 | .391 | .579 | 1.1 | .218 | .864 | 2.0 | .054 | .977 | 0.841 | .8 |
| .3 | .381 | .618 | 1.2 | .194 | .885 | 2.1 | .044 | .982 | 1.282 | .9 |
| .4 | .368 | .655 | 1.3 | .171 | .903 | 2.2 | .035 | .986 | 1.645 | .95 |
| .5 | .352 | .691 | 1.4 | .150 | .919 | 2.3 | .028 | .989 | 1.96 | .975 |
| .6 | .333 | .726 | 1.5 | .130 | .933 | 2.4 | .022 | .992 | 2.326 | .99 |
| .7 | .312 | .758 | 1.6 | .111 | .945 | 2.5 | .018 | .994 | 2.576 | .995 |
| .8 | .290 | .788 | 1.7 | .094 | .955 | 2.6 | .014 | .995 | 3.09 | .999 |

14. Student t table — Values $t_{m,p}$ of $x$ for which $P(|X| > x) = p$, when $X$ is $t_m$

| $m$ | $p=$ 0.10 | 0.05 | 0.02 | 0.01 | $m$ | $p=$ 0.10 | 0.05 | 0.02 | 0.01 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.31 | 12.71 | 31.82 | 63.66 | 9 | 1.83 | 2.26 | 2.82 | 3.25 |
| 2 | 2.92 | 4.30 | 6.96 | 9.92 | 10 | 1.81 | 2.23 | 2.76 | 3.17 |
| 3 | 2.35 | 3.18 | 4.54 | 5.84 | 12 | 1.78 | 2.18 | 2.68 | 3.05 |
| 4 | 2.13 | 2.78 | 3.75 | 4.60 | 15 | 1.75 | 2.13 | 2.60 | 2.95 |
| 5 | 2.02 | 2.57 | 3.36 | 4.03 | 20 | 1.72 | 2.09 | 2.53 | 2.85 |
| 6 | 1.94 | 2.45 | 3.14 | 3.71 | 25 | 1.71 | 2.06 | 2.48 | 2.78 |
| 7 | 1.89 | 2.36 | 3.00 | 3.50 | 40 | 1.68 | 2.02 | 2.42 | 2.70 |
| 8 | 1.86 | 2.31 | 2.90 | 3.36 | $\infty$ | 1.645 | 1.96 | 2.326 | 2.576 |

15. Chi-squared table — Values $\chi^2_{k,p}$ of $x$ for which $P(X > x) = p$, when $X$ is $\chi^2_k$ and $p = .995$, .975, *etc*

| $k$ | .995 | .975 | .05 | .025 | .01 | .005 | $k$ | .995 | .975 | .05 | .025 | .01 | .005 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .000 | .001 | 3.84 | 5.02 | 6.63 | 7.88 | 18 | 6.26 | 8.23 | 28.87 | 31.53 | 34.81 | 37.16 |
| 2 | .010 | .051 | 5.99 | 7.38 | 9.21 | 10.60 | 20 | 7.43 | 9.59 | 31.42 | 34.17 | 37.57 | 40.00 |
| 3 | .072 | .216 | 7.81 | 9.35 | 11.34 | 12.84 | 22 | 8.64 | 10.98 | 33.92 | 36.78 | 40.29 | 42.80 |
| 4 | .207 | .484 | 9.49 | 11.14 | 13.28 | 14.86 | 24 | 9.89 | 12.40 | 36.42 | 39.36 | 42.98 | 45.56 |
| 5 | .412 | .831 | 11.07 | 12.83 | 15.09 | 16.75 | 26 | 11.16 | 13.84 | 38.89 | 41.92 | 45.64 | 48.29 |
| 6 | .676 | 1.24 | 12.59 | 14.45 | 16.81 | 18.55 | 28 | 12.46 | 15.31 | 41.34 | 44.46 | 48.28 | 50.99 |
| 7 | .990 | 1.69 | 14.07 | 16.01 | 18.48 | 20.28 | 30 | 13.79 | 16.79 | 43.77 | 46.98 | 50.89 | 53.67 |
| 8 | 1.34 | 2.18 | 15.51 | 17.53 | 20.09 | 21.95 | 40 | 20.71 | 24.43 | 55.76 | 59.34 | 63.69 | 66.77 |
| 9 | 1.73 | 2.70 | 16.92 | 19.02 | 21.67 | 23.59 | 50 | 27.99 | 32.36 | 67.50 | 71.41 | 76.15 | 79.49 |
| 10 | 2.16 | 3.25 | 13.31 | 20.48 | 23.21 | 25.19 | 60 | 35.53 | 40.48 | 79.08 | 83.30 | 88.38 | 91.95 |
| 12 | 3.07 | 4.40 | 21.03 | 23.34 | 26.22 | 28.30 | 70 | 43.28 | 48.76 | 90.53 | 95.02 | 100.4 | 104.2 |
| 14 | 4.07 | 5.63 | 23.68 | 26.12 | 29.14 | 31.32 | 80 | 51.17 | 57.15 | 101.9 | 106.6 | 112.3 | 116.3 |
| 16 | 5.14 | 6.91 | 26.30 | 28.85 | 32.00 | 34.27 | 100 | 67.33 | 74.22 | 124.3 | 129.6 | 135.8 | 140.2 |

## 16. The chi-squared goodness-of-fit test

The frequencies $n_y$ are grouped so that the fitted frequency $\widehat{n}_y$ for every group exceeds about 5.

$X^2 = \sum_y \frac{(n_y - \widehat{n}_y)^2}{\widehat{n}_y}$ is referred to the table of $\chi^2_k$ with significance point $p$,

where $k$ is the number of terms summed, less one for each constraint, *eg* matching total frequency, and matching $\overline{x}$ with $\mu$

## 17. Joint probability distributions

Discrete distribution $\{p_{xy}\}$, where $p_{xy} = P(\{X = x\} \cap \{Y = y\})$.

Let $p_{x\bullet} = P(X = x)$, and $p_{\bullet y} = P(Y = y)$, then

$$p_{x\bullet} = \sum_y p_{xy} \quad \text{and} \quad P(X = x \,\big|\, Y = y) = \frac{p_{xy}}{p_{\bullet y}}$$

### Continuous distribution

Joint cdf $\quad F(x, y) = P(\{X \le x\} \cap \{Y \le y\}) = \int_{x_0=-\infty}^{x} \int_{y_0=-\infty}^{y} f(x_0\,, y_0)\,\mathrm{d}x_0\,\mathrm{d}y_0$

Joint pdf $\quad f(x, y) = \frac{\mathrm{d}^2 F(x, y)}{\mathrm{d}x\,\mathrm{d}y}$

Marginal pdf of $X$ $\quad f_X(x) = \int_{-\infty}^{\infty} f(x, y_0)\,\mathrm{d}y_0$

Conditional pdf of $X$ given $Y = y$ $\quad f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$ (provided $f_Y(y) > 0$)

## 18. Linear regression

To fit the linear regression model $y = \alpha + \beta x$ by $\widehat{y}_x = \widehat{\alpha} + \widehat{\beta} x$ from observations $(x_1, y_1), \ldots, (x_n, y_n)$, the least squares fit is $\quad \widehat{\alpha} = \overline{y} - \overline{x}\widehat{\beta}, \quad \widehat{\beta} = \frac{S_{xy}}{S_{xx}}$

The residual sum of squares $\quad \text{RSS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$\widehat{\sigma^2} = \frac{\text{RSS}}{n-2} \qquad \frac{n-2}{\sigma^2}\,\widehat{\sigma^2}$ is from $\chi^2_{n-2}$

$$E(\widehat{\alpha}) = \alpha\,, \quad E(\widehat{\beta}) = \beta\,,$$

$$\operatorname{var}(\widehat{\alpha}) = \frac{\sum x_i^2}{n\,S_{xx}}\sigma^2\,, \quad \operatorname{var}(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}}\,, \quad \operatorname{cov}(\widehat{\alpha}, \widehat{\beta}) = -\frac{\overline{x}}{S_{xx}}\,\sigma^2$$

$$\widehat{y}_x = \widehat{\alpha} + \widehat{\beta} x\,, \qquad E(\widehat{y}_x) = \alpha + \beta x\,, \qquad \operatorname{var}(\widehat{y}_x) = \left\{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right\}\sigma^2$$

$\frac{\widehat{\alpha} - \alpha}{\widehat{\text{se}}\,(\widehat{\alpha})}\,, \quad \frac{\widehat{\beta} - \beta}{\widehat{\text{se}}\,(\widehat{\beta})}\,, \quad \frac{\widehat{y}_x - \alpha - \beta\,x}{\widehat{\text{se}}\,(\widehat{y}_x)}$ are each from $t_{n-2}$

| | |
|---|---|
| Course: | M2SJ Solutions |
| Setter: | Battey (Q4), Lau (Q1-Q3) |
| Checker: | Fitz-Simon |
| Editor: | Walden |
| External: | Jennison |
| Date: | March 8, 2017 |

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May – June 2017

# M2SJ Solutions

# Statistical Methods Solutions

| Setter's signature | Checker's signature | Editor's signature |
|---|---|---|
| .................. | .................. | .................. |

 

BSc, MSc and MSci EXAMINATIONS (MATHEMATICS)

May – June 2017

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science.

Statistical Methods Solutions

Date: ??

Time: ??

Time Allowed: 2 Hours

This paper has *4* Questions.

Candidates should start their solutions to each question in a new main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

Statistical tables are provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers to each question as per the table below.

| Raw Mark | Up to 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| Extra Credit | 0 | $1/2$ | 1 | $1\,1/2$ | 2 | $2\,1/2$ | 3 | $3\,1/2$ | 4 |

- Each question carries equal weight.
- Calculators may be used.

1. **Each 4 marks**

(i) [**seen similar**] (c). Each question has a $X_i \sim Bernoulli(1/6)$ distribution of being answered correctly. Therefore, the total number of points from Question 1 is $\sum_{i=1}^{n} X_i \sim Binomial(5, 1/6)$, which has expectation $5/6$.

(ii) [**seen similar**] (e) Since $\int\int f(x,y)dxdy = 1$ we have

$$\frac{1}{c} = \left(\int_{-\infty}^{\infty} \exp(-x^2/4)\right)^2.$$

By recognising the integrand is proportion to the pdf of a $N(0,2)$ we obtain $c = \frac{1}{4\pi}$.

(iii) [**seen similar**] (f) As $\overline{X} \sim N(\mu, \sigma^2/n)$ it follows that $\frac{\overline{X}-\mu}{3} \sim N(0, \sigma^2/(9n))$

(iv) [**seen similar**] (f).

(v) [**seen similar**] (b) . First $P(\overline{B}|A) = 1 - P(B|A)$ and

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A)} \\ &= \frac{\left(1 - P(\overline{A}|B)\right)P(B)}{P(A)} \\ &= \frac{(1-0.7)(0.8)}{0.4} \\ &= 0.6 \end{aligned}$$

Finally, $P(\overline{B}|A) = 1 - 0.6 = 0.4$.

2. (i) We have $P(A) = 0.55$, $P(B) = 0.15$, $P(C) = 0.30$ and $P(D|A) = 0.02$, $P(D|B) = 0.04$, $P(D|C) = 0.05$.

(a) [**seen, 2 marks**]

$$P(B \cap D) = P(D|B)P(B) = 0.04 \times 0.15 = 0.006$$

(b) [**seen, 2 marks**]

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) \\ &= (0.02)(0.55) + (0.04)(0.15) + (0.05)(0.3) \\ &= 0.032 \end{aligned}$$

(ii) [**seen, 2 marks**]

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{(0.02)(0.55)}{0.032} = 0.34375$$

(iii) [**unseen, 6 marks**] Since the probabilities need to add up to one we have

$$0.1 + \alpha + 0.2 + \beta + 0.3 = 1 \implies \alpha + \beta = 0.4 \tag{1}$$

By the definition of expectation for a discrete random variable, we also have

$$\begin{aligned} 3.3 = E(X) = \sum_x xP(X = x) &= 0.1 + 2\alpha + 0.6 + 4\beta + 1.5 \\ \implies 3.3 &= 2\alpha + 4\beta + 2.2 \\ \implies 1.1 &= 2\alpha + 4\beta \\ \implies \alpha + 2\beta &= 0.55 \end{aligned} \tag{2}$$

Then

$$\text{Eqn } (2) - \text{Eqn } (1) \implies \beta = 0.55 - 0.4 = 0.15$$

and a simple substitution gives $\alpha = 0.25$.

(iv) [**unseen, 8 marks**] For the variance of $X$, first compute the second moment:

$$E(X^2) = \sum_x x^2 P(X = x) = 0.1 + 4(0.25) + 9(0.2) + 16(0.15) + 25(0.3) = 12.8$$

Therefore,

$$\mathsf{Var}(X) = E(X^2) - (E(X))^2 = 12.8 - (3.3)^2 = 1.91$$

Next, recall the skewness of a random variable (as noted in the formulae sheet) is given by:

$$\frac{E\left[(X-\mu)^3\right]}{\sigma^3}$$

For the given probability mass function we have $\mu = E(X) = 3.3$, hence

$$\begin{aligned} E\left[(X-\mu)^3\right] &= \sum_x (x - 3.3)^3 P(X = x) \\ &= (-2.3)^3(0.1) + (-1.3)^3(0.25) + (-0.3)^3(0.2) + (0.7)^3(0.15) + (1.7)^3(0.3) \\ &= -0.246 \end{aligned}$$

Finally, the skewness is $\dfrac{-0.246}{1.91^{3/2}} = -0.09319338$.

3. (i) [**seen, 1 mark**] We have that $X \sim N(80, 10^2)$. Let $Z \sim N(0,1)$. We then have that

$$\begin{aligned} P(X \geq 75) &= P\left(Z \geq \frac{75-80}{10}\right) \\ &= 1 - \Phi\left(-0.5\right) \\ &= \Phi\left(0.5\right) \\ &= 0.691 \end{aligned}$$

(ii) [**seen, 2 marks**] The random variable of interest is $X \sim Binomial(6, p)$ where $p = 1 - 0.691 = 0.309$. Therefore, the required probability is

$$\binom{6}{4} 0.309^4 0.691^2 = 0.0652952$$

(iii) [**seen, 5 marks**] We can approximate the $Binomial(n, p)$ distribution with a $N(np, np(1-p))$ distribution, where

$$np = (100)(0.309) = 30.9 \quad \text{and} \quad np(1-p) = (30.9)(0.691) = 21.3519$$

Therefore,

$$\begin{aligned} P(\text{at least 35 cans..}) \approx & P\left(Z \geq \frac{35-30.9}{\sqrt{21.3519}}\right) \\ &= P\left(Z \geq 0.88729\right) \\ &= 1 - \Phi(0.88729) \\ &\approx 1 - 0.816 \\ &\approx 0.184 \end{aligned}$$

(iv) [**seen, 3 marks**] The null hypothesis is $H_0 : \mu = 100$.

The alternative is of the form $H_1 : \mu < 100$ as the organisation suspects that the mean is below 100ml.

$\mu$ denotes the mean number of ml of caffeine in AlwaysUp made by the campany.

(v) (a) [**seen, 6 marks**] Although we have not been told that the amount of caffeine in AlwaysUp are individually normally distributed, by the CLT we have that the mean amount is **approximately** normally distributed. Hence the test statistic $T = \frac{X-\mu}{\sigma/\sqrt{n}}$ follows a standard normal $N(0,1)$ (**approximately**) under the null hypothesis.

Since this is a one-sided test the rejection region takes the form $R = \{t | t < z_{1-\alpha}\}$ where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$.

(b) [**seen, 3 marks**] The observed test statistic is $t = \frac{95-100}{30/\sqrt{100}} = -5/3 - 0.1666..$ and for $\alpha = 0.05$ we have $R = \{t | t < -1.645\}$. Since $t \in R$, there is sufficient evidence to reject the null at the 5% level.

4. (i) **[seen, 3 marks]**

$$M_Z(t) = \mathbb{E}\Big[\exp\{(X_1 + \cdots + X_n)(t/n)\}\Big] = \mathbb{E}\Big[\prod_{i=1}^{n} e^{(t/n)X_i}\Big] = \prod_{i=1}^{n} \mathbb{E}[e^{(t/n)X_i}] = \prod_{i=1}^{n} M_{X_i}(t),$$

where the penultimate equality follows by independence. Since $M_{X_i}(t) = M_X(t)$ for every $i \in \{1, \ldots, n\}$, $M_Z(t) = [M_X(t/n)]^n$.

(ii) (a) **[unseen, 3 marks]**

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i + 2\sum_{j=1}^{m} Y_j}{n+4m}\right] \\
&= \frac{1}{n+4m}\left\{\mathbb{E}\left[\sum_{i=1}^{n} X_i + 2\sum_{j=1}^{m} Y_j\right]\right\} = \frac{1}{n+4m}\left\{\sum_{i=1}^{n}\mathbb{E}[X_i] + 2\sum_{j=1}^{m}\mathbb{E}[Y_j]\right\} \\
&= \frac{1}{n+4m}\left[n\mu + 2m(2\mu)\right] = \frac{1}{n+4m}\left[(n+4m)\mu\right] = \mu
\end{aligned}$$

(b) **[unseen, 5 marks]** By Markov's inequality, $\Pr(|\widehat{\mu} - \mu| > \delta) \leq \frac{\mathbb{E}[|\widehat{\mu}-\mu|^r]}{\delta^r}$. Take $r = 2$, and notice that $\mathbb{E}[|\widehat{\mu} - \mu|^r] = \mathrm{Var}(\widehat{\mu})$.

$$\begin{aligned}
Var[\widehat{\mu}] &= Var\left[\frac{\sum_{i=1}^{n} X_i + 2\sum_{j=1}^{m} Y_j}{n+4m}\right] \\
&= \frac{1}{(n+4m)^2} Var\left[\sum_{i=1}^{n} X_i + 2\sum_{j=1}^{m} Y_j\right] = \frac{1}{(n+4m)^2}\left\{\sum_{i=1}^{n} Var[X_i] + 2^2\sum_{j=1}^{m} Var[Y_j]\right\} \\
&= \frac{1}{(n+4m)^2}\left[n\sigma + 4m\sigma\right] = \frac{1}{(n+4m)^2}\left[(n+4m)\sigma\right] = \sigma^2/(n+4m) \to 0
\end{aligned}$$

as $n$ or $m \to \infty$. Therefore $\widehat{\mu} \to_p \mu$, i.e. $\widehat{\mu}$ is consistent.

(iii) (a) **[unseen, 7 marks]**

$$\begin{aligned}
& \mathbb{P}\left(|T_n||S_n| > Cf_n Dg_n\right) \\
= \; & \mathbb{P}\Big(\Big\{\{|T_n||S_n| > Cf_n Dg_n\} \cap \{|S_n|/Dg_n > 1\}\Big\} \\
& \qquad \cup \Big\{\{|T_n||S_n| > Cf_n Dg_n\} \cap \{|S_n|/Dg_n \leq 1\}\Big\}\Big) \\
= \; & \mathbb{P}\Big(\Big\{|T_n||S_n| > Cf_n Dg_n\Big\} \cap \Big\{|S_n|/Dg_n > 1\Big\}\Big) \\
& \qquad + \mathbb{P}\Big(\Big\{|T_n||S_n| > Cf_n Dg_n\Big\} \cap \Big\{|S_n|/Dg_n \leq 1\Big\}\Big) \\
\leq \; & \mathbb{P}\left(|S_n| > Dg_n\right) + \mathbb{P}\left(|T_n| > Cf_n\right) < \delta \text{ for all } n > \max\{n_1, n_2\}.
\end{aligned}$$

The final line follows because the probability of the joint event $\Big\{|T_n||S_n| > Cf_n Dg_n\Big\} \cap \Big\{|S_n|/Dg_n > 1\Big\}$ must be smaller than the probability of the single event $\{|S_n| > Dg_n\}$, whilst if events $A := \Big\{|T_n||S_n| > Cf_n Dg_n\Big\}$ and $B := \Big\{|S_n|/Dg_n \leq 1\Big\}$ both occur, a fortiori (replacing $|S_n|/Dg_n$ by 1), event $E := \Big\{|T_n| > Cf_n\Big\}$ occurs, i.e. $(A \cap B) \subseteq E$, thus $\mathbb{P}(A \cap B) \leq \mathbb{P}(E)$.

(b) **[unseen, 2 marks]** Since $|T_n S_n| \leq |T_n||S_n|$, we know that $\mathbb{P}(|T_n S_n| > Cf_n Dg_n) \leq \mathbb{P}(|T_n||S_n| > Cf_n Dg_n)$, which is less than $\delta$ by part (i).

**Imperial College London**
**Department of Mathematics**

## Examiner's Comments

**Exam:** M2SJ **Session: 2016-2107**

**Question 1**

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

This was quite generally well done. In part (i), failing to note that there are six possible answers to each of five questions, so that the relevant Binomial distribution is $B(5, 1/6)$, was common. Part (iv) caused problems. many correctly noted that Power = P(reject $H_0$ | $H_0$ false), but then equated this to option (d), which is understandable, but not logically correct. In hypothesis testing the only two decisions are accept / reject null: 'not rejecting the alternative' really doesn't have meaning.

**Marker:** Alastair Young

**Signature:** G Alastair Young **Date:** 12 / 5 / 17

**Please return with exam marks (one report per marker)**

Imperial College London
Department of Mathematics

## Examiner's Comments

**Exam:** M2S1 **Session: 2016-2107**

**Question 2**

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

This question was very well done. The only part that caused any difficulty was calculation of the skewness in (iv) many people confused population and sample skewness, thus using the wrong definition, and calculation of $E(X-\mu)^3$ was badly handled by many candidates, with quite a few forgetting the cube.

**Marker:** Alastair Young

**Signature:** G Alastair Young **Date:** 12/5/17

**Please return with exam marks (one report per marker)**

Imperial College London
Department of Mathematics

## Examiner's Comments

**Exam:** M2SJ **Session: 2016-2107**

**Question 3**

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

Parts (i) – (iii) were well done, and appear to have been found routine. The meaning of $\mu$ in (iv), as the population mean number of mg, or the average number of mg in an infinite number of cans, was often badly expressed. Though the question pointed towards a one-sided test, by giving $H_1 : \mu < 100$, a lot of candidates ended up doing a two-sided test in (v), leading to the wrong final conclusion. Sound appreciation of the duality between a hypothesis test and constuction of a confidence interval was evident.

**Marker:** Alastai Young

**Signature:** G Alastai Young **Date:** 12/5/17

**Please return with exam marks (one report per marker)**

Imperial College London
Department of Mathematics

## Examiner's Comments

**Exam:** M2SJ **Session: 2016-2107**

**Question 4**

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

This question was found difficult, and there were several weak attempts. Many attempts failed to note the point where the independence assumption is needed in (i). Basic properties of variance (e.g. $var(aX) = a^2 var(X)$), def$^n$ of bias, Markov's inequality etc were often badly remembered in (ii). There were lots of bold attempts to (iii), but few that were entirely convincing.

**Marker:** Alastair Young

**Signature:** G Alastair Young **Date:** 12/5/17

**Please return with exam marks (one report per marker)**