Lecture 3 - Statistics Review and Item Statistics

Tony Tan

University of Oslo

Friday, 21 October 2022

Today's session

- Review of concepts: expectation, variance, covariance and correlation
- Illustrate statistical principles
- Define some useful notation
- Discuss different types of statistics and their interpretation for various types of item and test data

Table of Contents

Sums, products, and sets

Let $a_1, a_2, \ldots, a_K$ be a set of constants. The sum of these constants is written

$$\sum_{k=1}^{K} a_k = a_1 + a_2 + \cdots + a_K.$$

The product of these constants is written

$$\prod_{k=1}^{K} a_k = a_1 \times a_2 \times \cdots \times a_K.$$

In this notation, k is an index variable which takes integer values from 1 to the number K.
A set is denoted as $\mathcal{A} = \{1, 2, \ldots, K\}$. We say that "3 belongs to set $\mathcal{A}$", and write $3 \in \mathcal{A}$.

Expected value

Let X be a discrete random variable (R.V.) taking k different values with probabilities $p_1, \ldots, p_k$. The expected value of X is

$$\mathbb{E}(X) = \sum_{i=1}^{k} x_i \, p_i.$$

Let Y be a continuous R.V. with support (a, b) and density $f(\cdot)$. The expected value of Y is

$$\mathbb{E}(Y) = \int_a^b y \, f(y) \, \mathrm{d}y.$$

The expected value is a parameter often denoted by μ. We can interpret it as the long-run average value of the random variable under repeated sampling. Expected values can be infinite or undefined.

Expected value: Example

Let X be a discrete R.V. which can take values $x_1 = 0$ or $x_2 = 1$ with corresponding probabilities $p_1 = 0.4$ and $p_2 = 0.6$. The expected value of X is

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=1}^{2} x_i\, p_i \\ &= 0 \times 0.4 + 1 \times 0.6 \\ &= 0.6.\end{aligned}$$

Linearity of the expected value

For a number of R.V.s $X_1, \ldots, X_k$, the expectation of their sum is the sum of their expectations

$$\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k),$$

and for constants $a_1, \ldots, a_k$

$$\mathbb{E}(a_1 X_1 + \cdots + a_k X_k) = a_1 \mathbb{E}(X_1) + \cdots + a_k \mathbb{E}(X_k).$$

That it, expectation is transparent to linear operations.

Linearity of the expected value: Example

Let X_1 and X_2 be two random variables, where $\mathbb{E}(X_1) = 0.6$ and $\mathbb{E}(X_2) = 0.4$. The expected value of their sum

$$\begin{aligned}\mathbb{E}(X_1 + X_2) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) \\ &= 0.6 + 0.4 \\ &= 1.\end{aligned}$$

Consider constants $a_1 = 1$ and $a_2 = 2$. The expected value of the linear combination

$$\begin{aligned}\mathbb{E}(a_1X_1 + a_2X_2) &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) \\ &= 1 \times 0.6 + 2 \times 0.4 \\ &= 1.4.\end{aligned}$$

The k-th moment

Let X be a discrete R.V. that can take I different values with probabilities $p_1, \ldots, p_I$. The k-th moment of X is

$$\mathbb{E}\left(X^k\right) = \sum_{i=1}^{I} x_i^k p_i.$$

Let Y be a continuous R.V. with support (a, b) and density $f(\cdot)$. The k-th moment of Y is

$$\mathbb{E}\left(Y^k\right) = \int_a^b y^k f(y)\,\mathrm{d}y.$$

The k-th moment: Example

Let X be a discrete R.V. that can take values

$$x_1 = 0, \ x_2 = 1, \ \text{or } x_3 = 2$$

with corresponding probabilities

$$p_1 = 0.2, \ p_2 = 0.3, \ \text{and } p_3 = 0.5.$$

The 4th moment of X is

$$\begin{aligned}\mathbb{E}\left(X^4\right) &= 0^4 \times 0.2 + 1^4 \times 0.3 + 2^4 \times 0.5 \\ &= 0 + 0.3 + 8 \\ &= 8.3.\end{aligned}$$

The sample mean

Let $x_1, \ldots, x_n$ denote the sample. The sample mean $\overline{x}$ is computed as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample mean is often used as an estimator of the parameter μ.

Variance and standard deviation

Variances measure the dispersion of the data. For a R.V. X with expected value μ,

$$\mathrm{Var}\,(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left(X^2\right) - \mu^2.$$

The variance is a parameter often denoted by σ^2.
The standard deviation σ is the positive square root of the variance.

Properties of variance

Let X and Y be two R.V.s. The variances of their sum and difference are

$$\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2\operatorname{Cov}(X, Y).$$

For constants a and b,

$$\operatorname{Var}(aX \pm bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) \pm 2ab\operatorname{Cov}(X, Y).$$

Sample variance

For a sample $x_1, \ldots, x_n$, we can estimate their sample variance s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Dividing by $n-1$ is required in order to obtain the unbiased sample variance.

Example: We observe

$$5, 6, 9, 3, 20.$$

The sample mean is

$$(5 + 6 + 9 + 3 + 20)/5 = 8.6.$$

The unbiased sample variance is

$$[(5-8.6)^2+(6-8.6)^2+(9-8.6)^2+(3-8.6)^2+(20-8.6)^2]/(5-1) = 45.3.$$

Covariance

Consider two R.V.s X and Y. The covariance is a measure of the degree to which X and Y are interrelated

$$\operatorname{Cov}(X, Y) = \mathbb{E}\left\{[X - \mathbb{E}(X)]\,[Y - \mathbb{E}(Y)]\right\}.$$

Covariance is a parameter that can be estimated by the sample covariance. For matching-pair samples $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)$, their sample covariance is

$$\widehat{\operatorname{Cov}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} [x_i - \overline{x}]\,[y_i - \overline{y}].$$

Properties of covariance

From the definition of covariance, we have

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] \\
&= \mathbb{E}\left[XY - X\mathbb{E}(Y) - \mathbb{E}(X)\, Y + \mathbb{E}(X)\,\mathbb{E}(Y)\right] \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) - \mathbb{E}(X)\,\mathbb{E}(Y) + \mathbb{E}(X)\,\mathbb{E}(Y) \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).
\end{aligned}$$

Covariance and independence

Since

$$\mathrm{Cov}\,(X, Y) = \mathbb{E}\,(XY) - \mathbb{E}\,(X)\,\mathbb{E}\,(Y)\,,$$

if X and Y are independent ,

$$\mathrm{Cov}\,(X, Y) = \mathbb{E}\,(XY) - \mathbb{E}\,(X)\,\mathbb{E}\,(Y) = 0.$$

However, $\mathrm{Cov}\,(X, Y) = 0$ does not necessarily imply that X and Y are independent .

Correlation

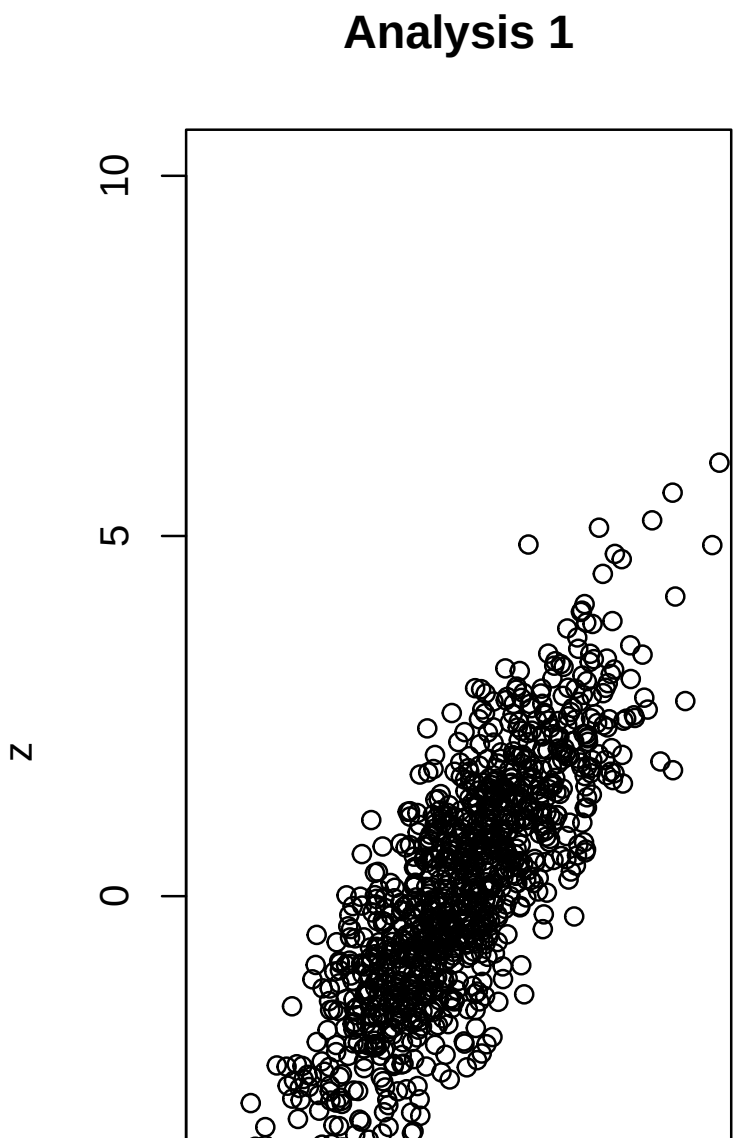
- The magnitude of the covariance is difficult to interpret on its own because its value depends on the scale of the R.V.s X and Y.
- A standardised measure, the correlation, can be used as a measure of the magnitude of the linear relationship between two R.V.s.

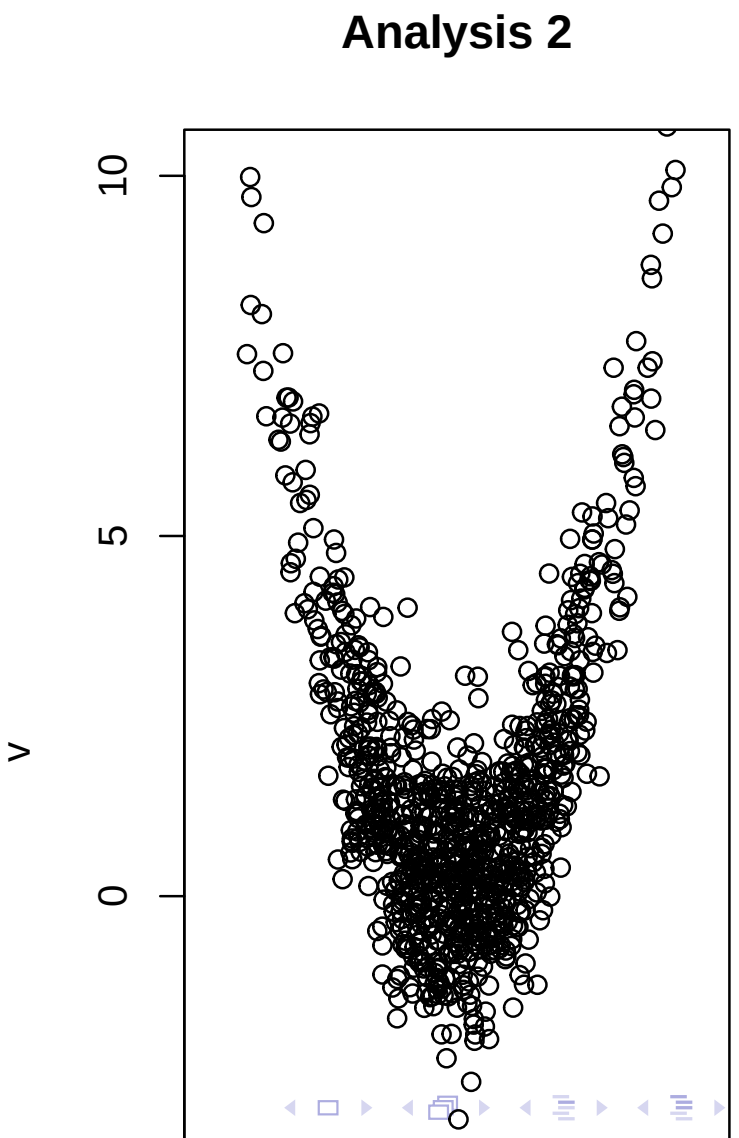
The Pearson correlation is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y respectively.
Note that $-1 \leqslant \rho_{X,Y} \leqslant 1$.

Pearson correlation measures a linear relationship





Pearson correlation measures a linear relationship

```
  cor(y, z)
```

```
## [1] 0.8218025
```

```
  cor(u, v)
```

```
## [1] 0.02514863
```

Statistics

- Statistics can be viewed as the methods by which we draw conclusions from incomplete information
- We design the data collection by considering statistical principles
- We utilize statistical methods when analysing the data
- We use statistical inference to accept or reject hypotheses or add knowledge to a body of scientific results
- All empirical research utilises statistical methods and principles

Parameters, estimators, and estimates

- The "truth": The parameter (e.g. μ)
- How we learn about the truth: The estimator (e.g. $\widehat{\mu}$)
- What we learn from the truth: The estimate (e.g. $\widehat{\mu}_{\text{obs}}$)

Standard deviation and standard error

Note the difference in these two concepts

- Standard deviation: The square root of the variance of a random variable $\sqrt{\mathrm{Var}(Y)}$ or of a population parameter $\sqrt{\mathrm{Var}(\theta)}$
- Standard error: The square root of the variance of an estimator $\sqrt{\mathrm{Var}\left(\widehat{\theta}\right)}$

Confidence intervals

- A 95% confidence interval (a, b) for a parameter θ means that the parameter θ is covered by such an interval 95% of the time if the sampling would be repeated infinitely.
- The confidence interval does not mean that the parameter has probability 0.95 of being in the interval. (cf. credible interval)
- In practice, we estimate a confidence interval and that interval either covers or does not cover the true parameter.

Exercise

- We observed the heights of 50 randomly selected UiO students.
- The sample mean was 173.4 cm and the sample standard deviation 15.5 cm.
- The standard error of the mean can be calculated by $\mathrm{se}(\bar{x}) = \sqrt{s^2/n}$, where s^2 is the sample variance and n is the sample size.

Estimate a 95% confidence interval for the population mean height and interpret the results.

Solution

- We observe $\overline{x} = 173.4$ cm and $\mathrm{se}(\overline{x}) = 15.5 \text{ cm}/\sqrt{50} \approx 2.2$ cm.
- We note that the statistic $(\overline{x} - \mu)/\mathrm{se}(\overline{x})$ follows the t-distribution with 49 degree of freedom.
- We calculate $t(49)_{(0.025)} \approx -2.0$ and construct the confidence interval for μ as

$$(\overline{x} - 2.0 \times 2.2,\ \overline{x} + 2.0 \times 2.2) = (169.0, 177.8).$$

That is, if we were to repeat the sampling infinitely many times, the true value μ would be covered by such an estimated interval 95% of the time.

Notation

- μ — expected value
- $\mathbb{E}(X)$ — expected value of X
- σ^2 — variance
- $\mathrm{Var}(X)$ — variance of X
- $\mathrm{Cov}(X, Y)$ — covariance of X and Y
- ρ — correlation

Bias, variance and mean squared error of an estimator

For an estimator $\widehat{\theta}$ of a parameter θ, the bias is defined as

$$\text{Bias}\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta} - \theta\right).$$

If $\text{Bias}\left(\widehat{\theta}\right) = 0$, we say that the estimator $\widehat{\theta}$ is an unbiased estimator of θ.

The estimator also has a variance

$$\text{Var}\left(\widehat{\theta}\right) = \mathbb{E}\left\{\left[\widehat{\theta} - \mathbb{E}\left(\widehat{\theta}\right)\right]^2\right\}.$$

We often consider the mean squared error (MSE) of an estimator

$$\text{MSE}\left(\widehat{\theta}\right) = \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right].$$

It can be shown that $\text{MSE}\left(\widehat{\theta}\right) = \text{Var}\left(\widehat{\theta}\right) + \left[\text{Bias}\left(\widehat{\theta}\right)\right]^2$.

Distributions

- $X \sim \mathcal{N}(\mu, \sigma^2)$ — X follows a normal distribution with mean μ and variance σ^2
- $X \sim t(\nu)$ — X follows a t-distribution with ν degrees of freedom
- $X \sim \chi^2(\nu)$ — X follows a χ^2-distribution with ν degrees of freedom

R.V.s and their observations

- The textbook defines upper-case letters as random variables and lower-case letters as observations from a sample.
- X_j denotes the j-th item score on a test and is a random variable
- x_{ji} denotes the score obtained on item j for an individual i and is an observation rather than a random variable

Mean vectors and covariance matrices

We can have vector-value random variables. We can consider the joint distribution of two variables X, Y.

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}(X) \\ \mathbb{E}(Y) \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \operatorname{Var}(X) & \operatorname{Cov}(X, Y) \\ \operatorname{Cov}(Y, X) & \operatorname{Var}(Y) \end{bmatrix}.$$

Table of Contents

Expected value of binary scores

- If the item score X_j can take values 0 or 1, it is a binary item
- The sample mean $\frac{1}{n}\sum_{i=1}^{n} x_{ji}$ can be used to compute an estimate of $\mu_j = \mathbb{E}(X_j)$
- Since this is a binary item, the parameter $\mu_j = \mathbb{E}(X_j)$ can be interpreted as defining the probability π_j of a randomly selected individual answering the item correctly

Variance of binary scores

For a random variable X_j defined such that

$$\mathbb{P}(X_j = 1) = \pi_j, \text{ and } \mathbb{P}(X_j = 0) = 1 - \pi_j,$$

we have

$$\mathbb{E}(X_j) = \pi_j,$$

and

$$\mathbb{E}\left(X_j^2\right) = 0^2 \times (1 - \pi_j) + 1^2 \times \pi_j = \pi_j.$$

Since $\mathrm{Var}(X_j) = \mathbb{E}\left(X_j^2\right) - [\mathbb{E}(X_j)]^2$,

$$\mathrm{Var}(X_j) = \pi_j - \pi_j^2 = \pi_j(1 - \pi_j).$$

To estimate the variance of a binary variable we can simply calculate the sample mean $\widehat{p}_j$ and obtain

$$\widehat{\mathrm{Var}}(X_j) = \widehat{p}_j(1 - \widehat{p}_j).$$

Note that this is a biased estimator of the variance, since it is divided by n instead of $n - 1$.

Covariances of binary scores

The sample covariance between two sets of observations $\{x_i\}$ and $\{y_i\}$ is

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

For binary variables X_j and X_k, this expression reduces to

$$s_{jk} = p_{jk} - p_j p_k,$$

where p_{jk} denotes the relative frequency of the event $\{X_j = 1, X_k = 1\}$, which can be estimated from the sample.

Exercise: Variance and covariance of binary scores

We observe the following frequencies from a sample $\{x_i\}$, $\{y_i\}$:

	$x_i = 0$	$x_i = 1$
$y_i = 0$	4	2
$y_i = 1$	1	3

What is s_x^2, s_y^2 and s_{xy}?

Solution: Variance and covariance of binary scores

We have

$$p_x = 0.5,$$

$$p_y = 0.4,$$

and

$$p_{xy} = 0.3.$$

We thus obtain

$$s_x^2 = 0.5 \times (1 - 0.5) = 0.25,$$

$$s_y^2 = 0.4 \times (1 - 0.4) = 0.24,$$

and

$$s_{xy} = 0.3 - 0.5 \times 0.4 = 0.1.$$

Covariance matrix

We can consider a number of item scores and calculate all their covariances. If we organise these results into a matrix, we obtain a covariance matrix .
With a two-item test

$$\mathbf{\Sigma}_{X,Y} = \begin{bmatrix} 0.25 & 0.10 \\ 0.10 & 0.24 \end{bmatrix},$$

the diagonal elements of the matrix are the variances of X and Y, and the off diagonals contain the covariances.
The matrix $\mathbf{\Sigma}_{X,Y}$ is symmetric since the upper- and lower-diagonals contain identical entries.

Table of Contents

Test scores

- A test score is typically the summation or a transformation of the individual item scores
- We may be interested in a single test score or multiple subscores
- We are interested in the same statistics as in item statistics: expected values, variances, covariances, and correlations

Total test score

If we consider m number of items for an individual i, the total test score y_i is simply the sum of the individual item scores x_{ji}:

$$y_i = \sum_{j=1}^{m} x_{ji}.$$

The mean test score is the average of the item scores

$$m_i = \frac{1}{m} \sum_{j=1}^{m} x_{ji}.$$

Sample variance of the total test score

We can calculate the sample variance of the total test score s_y^2 either from

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2,$$

or from the sum of the sample variances and covariances

$$s_y^2 = \sum_{j=1}^{m}\sum_{k=1}^{m} s_{jk}.$$

Example: Sample variance of the total test score

We can estimate the variance of the sum $X + Y$ from the previous exercise either by

$$\begin{aligned} s^2_{x+y} &= (4 \times 0^2 + 3 \times 1^2 + 3 \times 2^2)/10 - [(4 \times 0 + 3 \times 1 + 3 \times 2)/10]^2 \\ &= 1.5 - 0.81 \\ &= 0.69, \end{aligned}$$

or by

$$\begin{aligned} s^2_{x+y} &= 0.25 + 0.24 + 0.10 + 0.10 \\ &= 0.69. \end{aligned}$$

Review

- Discrete and continuous R.V.s
- Properties of expectation, variance and covariance
- Parameters, estimators and estimates
- Statistical inference
- How to estimate expectations, variances and covariances for item scores and test scores

Table of Contents

L3 Task 1

Consider a R.V. X that takes values 1, 2, 3 and 4 with corresponding probabilities 0.1, 0.2, 0.4 and 0.3.

a. What is $\mathbb{E}(X)$?

b. What is $\mathbb{E}(X^2)$?

c. What is $\mathrm{Var}(X)$?

Hint: Recall that $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

L3 Task 2

X and Y are two R.V.s such that $\mathbb{E}(X) = 10$, $\mathbb{E}(X^2) = 150$, $\mathbb{E}(Y) = 5$, $\mathbb{E}(Y^2) = 75$ and $\mathbb{E}(XY) = 20$.

a. What is $\mathrm{Cov}(X, Y)$?

b. What is $\mathrm{Cov}(5X, 10Y)$?

c. What is $\mathrm{Var}(5X + 10Y)$?

L3 Task 3

The following frequency table was observed from a two-item test where each item was scored 0 or 1.

	Item 1 = 0	Item 1 = 1
Item 2 = 0	42	20
Item 2 = 1	22	16

a Estimate the difficulty of each item.

b Estimate the variance of the total score.

L3 Task 4

The following covariance matrix was observed from a three-item test.

$$\mathbf{\Sigma}_{X_1,X_2,X_3} = \begin{bmatrix} 1.19 & 0.28 & 0.22 \\ 0.28 & 1.26 & 0.40 \\ 0.22 & 0.40 & 1.47 \end{bmatrix}.$$

a Calculate the sample variance of $X_1 + X_2 + X_3$.

b Calculate the estimated correlation between X_1 and X_3.

L3 Task 5

Show that the sample mean is an unbiased estimator of the expected value of a random variable X — that is, derive the expected value of

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

L3 Task 6

Assume that observations x_i are independent realisations of a random variable with finite first and second moments. Derive the variance of

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

L3 Task 7

Show that

$$\text{MSE}\left(\widehat{\theta}\right) = \text{Var}\left(\widehat{\theta}\right) + \left[\text{Bias}\left(\widehat{\theta}\right)\right]^2.$$

L3 Task 8

Derive the expected value of

$$s^*_{xy} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\overline{x}\,\overline{y}}{n-1},$$

where observations k and l $(k \neq l)$ are independent.