

# 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

**1** Statistics review

2 Item statistics

3 Test score statistics

4 Exercises

## Sums, products, and sets

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

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

The **product** of these constants is written

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

In this notation,  $k$  is an **index variable** which takes integer values from 1 to the number  $K$ .

A **set** is denoted as  $\mathcal{A} = \{1, 2, \dots, K\}$ . 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  $p_1, \dots, p_k$ . 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

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

The expected value is a **parameter** often denoted by  $\mu$ . 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  $X_1, \dots, X_k$ , the expectation of their sum is the sum of their expectations

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

and for constants  $a_1, \dots, a_k$

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

That is, 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  $p_1, \dots, p_I$ . The  $k$ -th moment of  $X$  is

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

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

$$\mathbb{E}(Y^k) = \int_a^b y^k f(y) dy.$$

## 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

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

## The sample mean

Let  $x_1, \dots, x_n$  denote the sample. The **sample mean**  $\bar{x}$  is computed as

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

The sample mean is often used as an **estimator** of the parameter  $\mu$ .

## Variance and standard deviation

**Variances** measure the dispersion of the data. For a R.V.  $X$  with expected value  $\mu$ ,

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}(X^2) - \mu^2.$$

The variance is a parameter often denoted by  $\sigma^2$ .

The **standard deviation**  $\sigma$  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

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

For constants  $a$  and  $b$ ,

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

## Sample variance

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

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

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

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

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

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

## Properties of covariance

From the definition of covariance, we have

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



## Covariance and independence

Since

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

if  $X$  and  $Y$  are **independent**,

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

However,  $\text{Cov}(X, Y) = 0$  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

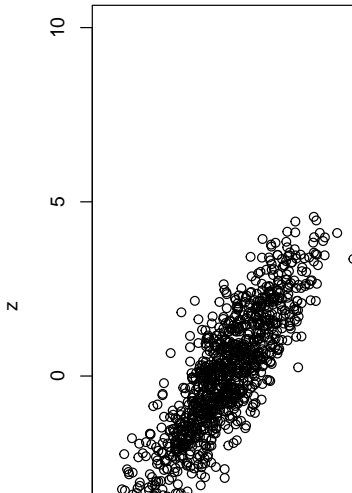
$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where  $\sigma_X$  and  $\sigma_Y$  are the standard deviations of  $X$  and  $Y$  respectively.

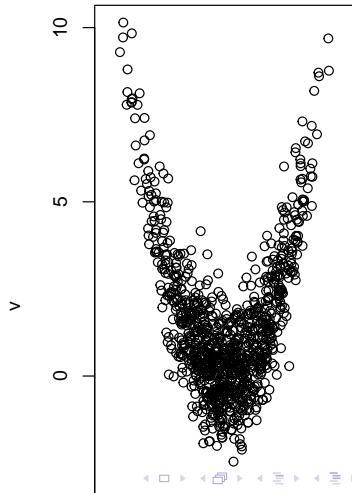
Note that  $-1 \leq \rho_{X,Y} \leq 1$ .

# Pearson correlation measures a linear relationship

**Analysis 1**



**Analysis 2**



## Pearson correlation measures a linear relationship

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

```
## [1] 0.8440646
```

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

```
## [1] -0.09752877
```

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

# Standard deviation and standard error

Note the difference in these two concepts

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

# Confidence intervals

- A 95% confidence interval  $(a, b)$  for a parameter  $\theta$  means that the parameter  $\theta$  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  $se(\bar{x}) = \sqrt{s^2/n}$ , 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  $\bar{x} = 173.4$  cm and  $\text{se}(\bar{x}) = 15.5 \text{ cm} / \sqrt{50} \approx 2.2$  cm.
- We note that the statistic  $(\bar{x} - \mu) / \text{se}(\bar{x})$  follows the  $t$ -distribution with 49 degree of freedom.
- We calculate  $t(49)_{(0.025)} \approx -2.0$  and construct the confidence interval for  $\mu$  as

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

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

# Notation

- $\mu$  — expected value
- $\mathbb{E}(X)$  — expected value of  $X$
- $\sigma^2$  — variance
- $\text{Var}(X)$  — variance of  $X$
- $\text{Cov}(X, Y)$  — covariance of  $X$  and  $Y$
- $\rho$  — correlation

## Bias, variance and mean squared error of an estimator

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

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta).$$

If  $\text{Bias}(\hat{\theta}) = 0$ , we say that the estimator  $\hat{\theta}$  is an **unbiased estimator** of  $\theta$ .

The estimator also has a variance

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

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

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

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

# Distributions

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

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

# Table of Contents

1 Statistics review

2 Item statistics

3 Test score statistics

4 Exercises



## 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  $\pi_j$  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

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

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

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

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

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

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

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

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

$$\Sigma_{X,Y} = \begin{bmatrix} 0.25 & 0.10 \\ 0.10 & 0.24 \end{bmatrix},$$

the diagonal elements of the matrix are the variances of  $X$  and  $Y$ , and the off diagonals contain the covariances.

The matrix  $\Sigma_{X,Y}$  is **symmetric** since the upper- and lower-diagonals contain identical entries.

# Table of Contents

1 Statistics review

2 Item statistics

**3 Test score statistics**

4 Exercises

# 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

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

or by

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

# 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

1 Statistics review

2 Item statistics

3 Test score statistics

4 Exercises

## 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  $\text{Var}(X)$ ?

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

## L3 Task 1: Solution

$$\mathbb{E}(X) = 1 \times 0.1 + 2 \times 0.2 + 3 \times 0.4 + 4 \times 0.3 = 2.9$$

$$\mathbb{E}(X^2) = 1^2 \times 0.1 + 2^2 \times 0.2 + 3^2 \times 0.4 + 4^2 \times 0.3 = 9.3$$

$$\text{Var}(X) = \mathbb{E}(X^2) + [\mathbb{E}(X)]^2 = 9.3 - 2.9^2 = 0.89$$

To verify  $\text{Var}(X) = \mathbb{E}(X^2) + [\mathbb{E}(X)]^2$ :

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\ &= \mathbb{E}\left\{X^2 - 2X\mathbb{E}(X) + [\mathbb{E}(X)]^2\right\} \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + [\mathbb{E}(X)]^2 \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.\end{aligned}$$

## 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  $\text{Cov}(X, Y)$ ?
- b What is  $\text{Cov}(5X, 10Y)$ ?
- c What is  $\text{Var}(5X + 10Y)$ ?



## L3 Task 2: Solution

Notice that

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \mathbb{E}(Y) = 20 - 50 = -30$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 150 - 100 = 50$$

$$\text{Var}(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = 75 - 25 = 50$$

We thus obtain

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-30}{\sqrt{50}\sqrt{50}} = \frac{-30}{50} = -0.6$$

$$\text{Cov}(5X, 10Y) = 5 \times 10 \times \text{Cov}(X, Y) = -1500$$

$$\begin{aligned}\text{Var}(5X + 10Y) &= 5^2 \text{Var}(X) + 10^2 \text{Var}(Y) + 2 \times 5 \times 10 \text{Cov}(X, Y) \\ &= 25 \times 50 + 100 \times 50 + 100 \times (-30) \\ &= 3250\end{aligned}$$

## 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 3: Solution

Per our textbook (p. 34), the difficulty parameter for binary items is the mean score of each item. We thus obtain (textbook Eq. (3.1))  $\hat{p}_1 = (20 + 16)/100 = 0.36$  and  $\hat{p}_2 = (22 + 16)/100 = 0.38$ . We also know that the variance of a sum is the sum of the variances plus all covariance pairs. We thus need variances (Eq. (3.7))

$$s_1^2 = \hat{p}_1 \times (1 - \hat{p}_1) = 0.36 \times 0.64 = 0.2304$$

$$s_2^2 = \hat{p}_2 \times (1 - \hat{p}_2) = 0.38 \times 0.62 = 0.2356,$$

and covariances (Eq. (3.14a))

$$s_{12} = s_{21} = \hat{p}_{12} - \hat{p}_1 \times \hat{p}_2 = 0.16 - 0.36 \times 0.38 = 0.0232.$$

The total variance therefore is (Eq. (3.27))

$$s_{1+2}^2 = 0.2304 + 0.2356 + 2 \times 0.0232 = 0.5124.$$

## L3 Task 4

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

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

- 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 4: Solution

The variance of a sum is the sum of the variances (main diagonals) plus all covariance pairs (off-diagonals). We hence add up all entries in the  $\Sigma_{X_1, X_2, X_3}$  matrix:

$$\text{Var}(X_1 + X_2 + X_3) = \sum_{i=1}^3 \sum_{j=1}^3 s_{ij} = 5.72.$$

$$\widehat{\text{corr}}(X_1, X_3) = \frac{\widehat{\text{Cov}}(X_1, X_3)}{\widehat{\sigma}_{X_1} \widehat{\sigma}_{X_3}} = \frac{0.22}{\sqrt{1.19} \sqrt{1.47}} \approx 0.166$$

## 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

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i.$$

## L3 Task 5: Solution

$$\begin{aligned}\mathbb{E}(\bar{x}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \\&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(x_i) \\&= \frac{1}{n} [\mathbb{E}(x_1) + \cdots + \mathbb{E}(x_n)] \\&= \frac{1}{n} [n\mu] \\&= \mu.\end{aligned}$$

Since  $\mathbb{E}(\bar{x} - \mu) = 0$ , we conclude that the sample mean  $\bar{x}$  is an unbiased estimator of the population mean  $\mu$ .

## 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

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



## L3 Task 6: Solution

$$\begin{aligned}\text{Var}(\bar{x}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \\&= \frac{1}{n^2} \left[ \sum_{i=1}^n \text{Var}(x_i) + 2 \sum_{i \neq j} \text{Cov}(x_i, x_j) \right] \\&= \frac{1}{n^2} [\text{Var}(x_1) + \cdots + \text{Var}(x_n)] \\&= \frac{1}{n^2} [n\sigma^2] \\&= \frac{\sigma^2}{n}.\end{aligned}$$

Combining Task 5 and 6, we have derived the **sampling distribution of the mean** :

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$$

## L3 Task 7

Show that

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

## L3 Task 7: Solution

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right] \\&= \mathbb{E} (\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \\&= \mathbb{E} (\hat{\theta}^2) + \theta^2 - 2\theta \mathbb{E} (\hat{\theta}) + [\mathbb{E} (\hat{\theta})]^2 - [\mathbb{E} (\hat{\theta})]^2 \\&= \mathbb{E} (\hat{\theta}^2) + [\theta - \mathbb{E} (\hat{\theta})]^2 - [\mathbb{E} (\hat{\theta})]^2 \\&= \mathbb{E} (\hat{\theta}^2) - [\mathbb{E} (\hat{\theta})]^2 + [\theta - \mathbb{E} (\hat{\theta})]^2 \\&= \text{Var} (\hat{\theta}) + [\text{Bias} (\hat{\theta})]^2\end{aligned}$$

## L3 Task 8

Derive the expected value of

$$s_{xy}^* = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{n - 1},$$

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

## L3 Task 8: Solution I

The covariance between  $X$  and  $Y$  is

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \mathbb{E}(Y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

We first focus on the term

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Since

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y},$$

## L3 Task 8: Solution II

taking expected value on both sides yields

$$\begin{aligned} & \mathbb{E} \left[ \sum (x_i - \bar{x})(y_i - \bar{y}) \right] \\ &= \mathbb{E} \left[ \sum x_i y_i \right] - \mathbb{E} \left[ \sum x_i \bar{y} \right] - \mathbb{E} \left[ \sum \bar{x} y_i \right] + \mathbb{E} \left[ \sum \bar{x} \bar{y} \right]. \end{aligned}$$

We now study each of the four terms separately:

$$\mathbb{E} \left( \sum_{i=1}^n x_i y_i \right) = \sum_{i=1}^n \mathbb{E} (x_i y_i) = n \mathbb{E} (XY),$$

$$\begin{aligned} \mathbb{E} \left[ \sum x_i \bar{y} \right] &= \mathbb{E} \left[ \sum_i x_i \left( \frac{1}{n} \sum_j y_j \right) \right] \\ &= \frac{1}{n} \left[ \sum_i \mathbb{E} \left( x_i \sum_j y_j \right) \right] \end{aligned}$$

$$= \frac{1}{n} \sum_i \sum_j \mathbb{E} (x_i) \mathbb{E} (y_j) = \frac{1}{n} \sum_i \mathbb{E} (x_i) \sum_j \mathbb{E} (y_j)$$

## L3 Task 8: Solution III

$$\begin{aligned}\mathbb{E} \left[ \sum \bar{x} y_i \right] &= \mathbb{E} \left[ \left( \frac{1}{n} \sum_j x_j \right) \sum_i y_i \right] \\ &= \frac{1}{n} \left[ \sum_i \mathbb{E} \left( y_i \sum_j x_j \right) \right] \\ &= \frac{1}{n} \sum_i \sum_{i \neq j} \mathbb{E}(y_i) \mathbb{E}(x_j) + \frac{1}{n} \sum_{i=j} \mathbb{E}(y_i x_j) \\ &= (n-1) \mathbb{E}(Y) \mathbb{E}(X) + \mathbb{E}(YX),\end{aligned}$$

## L3 Task 8: Solution IV

and lastly,

$$\begin{aligned}\mathbb{E} \left[ \sum \bar{x} \bar{y} \right] &= n \mathbb{E} \left[ \sum_i \left( \frac{1}{n} \sum_j x_j \right) \left( \frac{1}{n} \sum_k y_k \right) \right] \\&= \frac{1}{n} \mathbb{E} \left( \sum_i x_i \sum_j y_j \right) \\&= \frac{1}{n} \sum_i \mathbb{E} \left( x_i \sum_j y_j \right) \\&= \frac{1}{n} (n-1) n \mathbb{E}(X) \mathbb{E}(Y) + \frac{1}{n} n \mathbb{E}(XY) \\&= (n-1) \mathbb{E}(X) \mathbb{E}(Y) + \mathbb{E}(XY).\end{aligned}$$



## L3 Task 8: Solution V

Combine the four terms:

$$\begin{aligned}\mathbb{E} \left[ \sum (x_i - \bar{x})(y_i - \bar{y}) \right] &= n\mathbb{E}(XY) - (n-1)\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(XY) \\ &\quad - (n-1)\mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(YX) \\ &\quad + (n-1)\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(XY) \\ &= (n-1)\mathbb{E}(XY) - (n-1)\mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

The covariance of  $X$  and  $Y$  demands a factor of  $1/n$ :

$$\begin{aligned}\mathbb{E}(s_{xy}) &= \frac{n-1}{n} [\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \frac{n-1}{n} \sigma_{XY}.\end{aligned}$$

## L3 Task 8: Solution VI

Consequently, an unbiased estimator of  $\sigma_{xy}$  is

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