Solutions

Multiple-choice questions

Question	1	2	3	4	5	6	7	8
Answer	c	b	a	b	a	c	b	a

1. (c) The ECDF is a step function: $F(2.5) = F(2) = f_1 + f_2 = \frac{4+6}{20} = 0.50$. Values larger than 2 but smaller than 3 add nothing.

2. (b) Time is measured and can take any real value (continuous). Number of siblings is a count (discrete); T-shirt size is qualitative ordinal.

3. (a) For $Y = a + bX$: $s_Y^2 = b^2 s_X^2 = 2^2 \cdot 6 = 24$. The constant $a = -5$ does not change the variance.

4. (b) Pooled mean: $\bar{x} = \frac{20 \cdot 5 + 30 \cdot 10}{20+30} = \frac{400}{50} = 8$. The simple average 7.5 ignores the different group sizes.

5. (a) $\chi^2_{\max} = n\,(\min(k,h) - 1) = 60 \cdot (2-1) = 60$, hence $\tilde{\chi}^2 = 6/60 = 0.10$. Using 3 instead of $\min(2,3) = 2$ is the usual slip.

6. (c) $s^2_{X+Y} = s^2_X + s^2_Y + 2s_{X,Y} = 4 + 9 + 2 \cdot (-3) = 7$. A negative covariance lowers the variance of the sum.

7. (b) Combinations without ordering: $\binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3!} = 20$. The value 120 would count the ordered selections.

8. (a) $E[X] = 0.3 + 0.4 = 0.7$, $E[X^2] = 0.3 + 0.8 = 1.1$, so $V[X] = E[X^2] - (E[X])^2 = 1.1 - 0.49 = 0.61$.

Exercise 1

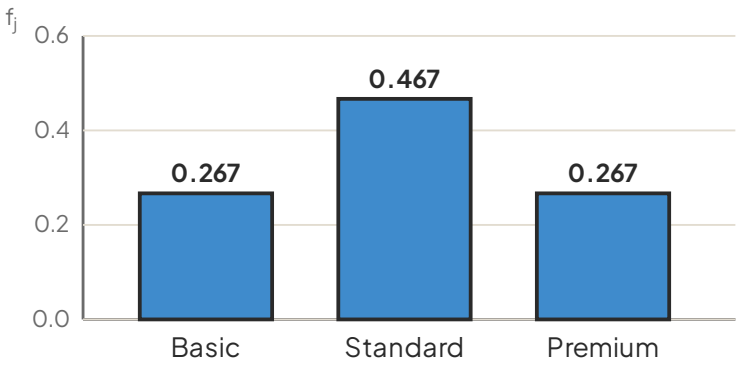
BEFORE YOU START

In R, `table(Comfort)` sorts the levels alphabetically (Basic, Premium, Standard), which is **not** the natural order. For cumulative frequencies and the median declare the order first: `Comfort = factor(Comfort, levels = c("Basic","Standard","Premium"))`.

1. Frequency table, graph and median

`Comfort` is qualitative **ordinal**: cumulative frequencies and the median make sense, and the right graph is the **bar chart**.

Comfort	n_j	f_j	F_j
Basic	8	4/15 = 0.267	0.267
Standard	14	7/15 = 0.467	0.733
Premium	8	4/15 = 0.267	1
	30	**1**	



$\tilde{x}_{0.5} = \min\{a_j : F_j \geq 0.5\} =$ **Standard** (the first value with $F_j \geq 0.5$ is Standard, $F = 0.733$).

2. Price in Turin and in Bologna

Location: arithmetic mean; variability: unbiased standard deviation and coefficient of variation $v = s/|\bar{x}|$ (the CV is free of the measurement scale, so it is the right tool to compare variability when the means differ).

	n	mean	s	v = s/|mean|
Turin	14	124.86	41.13	0.329
Bologna	16	167.00	34.94	0.209

Comment. On average a desk costs about 42 euro more in Bologna (167.00 vs 124.86). Prices are also more variable in Turin: both the standard deviation (41.13 vs 34.94) and, above all, the coefficient of variation (0.329 vs 0.209) are higher, i.e. the spread per euro of average price is larger. Turin mixes more Basic and Standard spaces with a few Premium ones.

3. Contingency table and normalized χ^2

Comfort \ City	Turin	Bologna	n_{i+}
Basic	6	2	8
Standard	6	8	14
Premium	2	6	8
n_{+j}	14	16	**30**

Expected counts $\tilde{n}_{ij} = n_{i+}n_{+j}/n$ (Basic: 3.73 and 4.27; Standard: 6.53 and 7.47; Premium: 3.73 and 4.27).

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}} = 4.17 \qquad \chi^2_{\max} = n\,(\min(3,2)-1) = 30 \qquad \tilde{\chi}^2 = \frac{4.17}{30} = 0.139$$

Comment. $\tilde{\chi}^2 \in [0,1]$ and 0.139 is close to 0: there is only a **weak association** between comfort level and city (Bologna has slightly more Premium spaces, Turin more Basic ones), and the two variables are not far from independence. R prints a warning ("Chi-squared approximation may be incorrect") because some expected counts are small: it does not affect the value of the statistic that we use here.

4. Covariance and correlation of Members and Price

$$s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 426.84 \qquad r_{X,Y} = \frac{s_{X,Y}}{s_X\,s_Y} = 0.845$$

Comment. Both are positive: spaces with more members tend to have a higher price. The covariance depends on the units (members × euro), so only the correlation can be judged: $|r| = 0.845$ is close to 1, i.e. a **strong positive linear association**.

Exercise 2

Let A, B, C be the events "the component comes from supplier A, B, C" (a partition: $P(A) = 0.5,\ P(B) = 0.3,\ P(C) = 0.2$) and D = "the component is defective", with $P(D|A) = 0.01,\ P(D|B) = 0.02,\ P(D|C) = 0.05$.

1. Law of total probability

$$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) = 0.005 + 0.006 + 0.010 = 0.021$$

2. Bayes' theorem

$$P(C|D) = \frac{P(D|C)\,P(C)}{P(D)} = \frac{0.05 \cdot 0.20}{0.021} = \frac{0.010}{0.021} = 0.476$$

Supplier C provides only 20% of the components but about 48% of the defective ones.

3. Binomial probability

Let X = "number of defective components out of 10": $X \sim \text{Bin}(n = 10,\ p = 0.021)$ (independent picks, same probability $p = P(D)$ from point 1).

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.979^{10} + 10 \cdot 0.021 \cdot 0.979^{9} = 0.8087 + 0.1735 = 0.982$$

In R: `pbinom(q = 1, size = 10, prob = 0.021)`, which gives 0.9823.

> **COMMON MISTAKE**
> In point 3 the success probability is **0.021** (probability of a defective component from the whole stock), not 0.01, 0.02 or 0.05: those are conditional on the supplier.