Chapter 2

Probability and distributions

Exercises

Contents

2.1 Discrete random variable

|||| Exercise 2.1 Discrete random variable

a) Let X be a stochastic variable. When running `stats.binom.pmf(4, 10, 0.6)` in Python (equivalently, `dbinom(4,10,0.6)` in R) the result is 0.1115. What distribution is applied and what does 0.1115 represent?

b) Let X be the same stochastic variable as above. Given that $P(X \leq 4) = 0.1662$ and $P(X \leq 5) = 0.3669$, calculate $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

c) Let X be a stochastic variable. Running `stats.poisson.pmf(4, 3)` returns 0.168. What distribution is applied and what does 0.168 represent?

d) Let X be the same Poisson variable as above. Given that $P(X \leq 4) = 0.8153$ and $P(X \leq 5) = 0.9161$, calculate $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

Exam-style questions

e) Rain falls independently on each day of a 6-day rainy-season forecast window, with probability $p = 0.4$ on any given day. Running

```
stats.binom.pmf(2, 6, 0.4)
```

in Python returns 0.311. Which of the following correctly describes what this number represents?

1) The probability that it rains on exactly 2 out of the 6 days.
2) The probability that it rains on at most 2 out of the 6 days.

3) The probability that it rains on at least 2 out of the 6 days.
4) The probability that it rains on 2 specific, pre-chosen days (say, Monday and Tuesday) and not on any of the other 4.
5) The probability that it does not rain on any of the 6 days.

f) A support inbox receives emails at an average rate of $\lambda = 5$ per hour, following a Poisson distribution $X \sim \text{Pois}(5)$. It is known that

$$P(X \leq 3) = 0.2650 \qquad \text{and} \qquad P(X \leq 7) = 0.8666.$$

What is the probability that the inbox receives *between 4 and 7 emails (inclusive)* in a given hour?

1) 0.8666
2) 0.6016
3) 0.4261
4) 0.3984
5) 1.1316

2.2 Course passing proportions

|||| Exercise 2.2 Course passing proportions

a) If a passing proportion for a course given repeatedly is assumed to be 0.80 on average, and there are 250 students taking the exam each time, what is the expected value μ and standard deviation σ for the number of students who do *not* pass the exam for a randomly selected course?

Exam-style questions

b) A ferry company reports that, on average, 12% of its crossings are delayed due to weather. Over a season there are $n = 180$ crossings. Let Y denote the number of delayed crossings. Running

```
Y = stats.binom(180, 0.12)
print(Y.mean(), Y.std())
```

gives `21.6, 4.359`. Which statement correctly interprets this output?

1) On average 21.6 crossings are delayed per season, with a standard deviation of about 4.36 crossings.
2) On average 21.6 crossings are delayed per season, with a standard deviation of about 4.36%.
3) The probability of a delay is 21.6%, with an uncertainty of 4.36%.
4) Exactly 21.6 crossings will be delayed every season.
5) On average 4.36 crossings are delayed per season, with a standard deviation of about 21.6 crossings.

c) A dental clinic finds that, historically, 15% of scheduled appointments are cancelled at the last minute. Out of a new batch of $n = 300$ scheduled appointments, let Z denote the number of cancellations. What is the standard deviation of Z?

1) 6.19

2) 45.00
3) 6.71
4) 38.25
5) 0.15

2.3 Notes in a box

|||| Exercise 2.3 Notes in a box

A box contains 6 notes:

On 1 of the notes there is the number 1
On 2 of the notes there is the number 2
On 2 of the notes there is the number 3
On 1 of the notes there is the number 4

Two notes are drawn at random from the box without replacement. X describes the number of notes with the number 4 among the 2 drawn.

a) Find the mean and variance of X, and $P(X = 0)$.

b) The 2 notes are now drawn *with* replacement. What is the probability that none of the 2 notes has the number 1 on it?

Exam-style questions

c) A quality inspector examines a shipment of 15 circuit boards, 4 of which are faulty. She randomly selects 3 boards without replacement for testing. Let W denote the number of faulty boards among the 3 selected. What is the probability that *exactly one* of the three selected boards is faulty?

1) 0.484
2) 0.363
3) 0.145
4) 0.267
5) 0.430

d) Suppose instead the inspector samples 3 boards *with* replacement from the same batch of 15 boards (4 faulty). Running

```
print(stats.binom(3, 4/15).mean())
```

gives `0.8`. If she instead samples 3 boards *without* replacement (as in the previous question), will the expected number of faulty boards found be higher, lower, or the same as 0.8?

1) The same, because the hypergeometric mean $n \cdot a/N$ is always numerically equal to the binomial mean np when $p = a/N$ — only the variance differs between sampling with and without replacement.
2) Higher without replacement, since removing a faulty board increases the chance that the next board is also faulty.
3) Lower without replacement, since a board that has already been tested can never be found faulty again.
4) It cannot be determined without knowing the exact order in which the boards are drawn.
5) Higher without replacement, because variance is always higher when sampling without replacement.

2.4 Consumer survey

|||| Exercise 2.4 Consumer survey

In a consumer survey performed by a newspaper, 20 different groceries were purchased in a grocery store. Discrepancies between the price on the sales slip and the shelf price were found in 6 of these 20 products.

a) At the same time a customer buys 3 random (different) products within the group of 20 goods in the store. What is the probability that no discrepancies occur for this customer?

Exam-style questions

b) A student is confused by the Python call above, `stats.hypergeom.pmf(0, M=20, n=6, N=3)`: in the book's own notation, n denotes the *number of draws*, yet here `n=6` and `N=3` — seemingly the opposite of the book's $N = 20$ (population), $a = 6$ (defective items) and $n = 3$ (draws). Which statement correctly resolves this apparent conflict?

1) SciPy's `hypergeom` uses a different naming convention from the book: `M` is the population size (book's N), `n` is the number of "success" items in the population (book's a), and `N` is the number of draws (book's n). The call is correct despite the reused letters.
2) The call is wrong; it should be `stats.hypergeom.pmf(0, M=20, n=3, N=6)` to match the book's notation.
3) The call is wrong; SciPy's `n` and `N` are always interchangeable for the hypergeometric distribution.
4) The call happens to give the right numerical answer only by coincidence.
5) SciPy's `hypergeom` does not support keyword arguments; the code as written will raise an error.

c) A batch of 30 lightbulbs contains 5 defective ones. An inspector randomly selects 6 bulbs without replacement for testing. What is the probability that *at least one* of the 6 selected bulbs is defective?

1) 0.702
2) 0.298
3) 0.167
4) 1.000
5) 0.833

2.5 Hay delivery quality

Exercise 2.5 Hay delivery quality

A horse owner receives 20 bales of hay in a sealed plastic packaging. To control the hay, 3 bales are randomly selected and each checked for harmful fungal spores.

It is believed that among the 20 bales, 2 are infected with fungal spores. A random variable X describes the number of infected bales among the three selected.

a) Find the mean of X, the variance of X, and $P(X \geq 1)$.

b) Another supplier advertises that no more than 1% of his bales are infected. The horse owner buys 10 bales and decides to buy from this supplier for the rest of the season if all 10 bales are error-free. Find the probability that all 10 purchased bales are error-free if (a) 1% of bales are infected (p_1), and (b) 10% of bales are infected (p_{10}).

Exam-style questions

c) A vineyard harvests 25 crates of grapes, 3 of which have been exposed to frost damage. A wine inspector randomly selects 4 crates to test for damage. Let V denote the number of frost-damaged crates among the 4 selected. Which of the following gives the correct mean and variance of V?

1) $\mu = 0.48, \sigma^2 = 0.370$
2) $\mu = 0.48, \sigma^2 = 0.4224$
3) $\mu = 0.12, \sigma^2 = 0.370$
4) $\mu = 0.48, \sigma^2 = 0.48$
5) $\mu = 3.00, \sigma^2 = 0.370$

d) A large medical device manufacturer claims that no more than 0.5% of its syringes are defective. A hospital buys a box of 20 syringes and will switch suppliers unless all 20 are defect-free. If the manufacturer's claim is accurate ($p = 0.005$), `stats.binom.pmf(0, 20, 0.005)` returns 0.905. If the true defect rate is actually $p = 0.05$ (ten times higher than claimed), what is the probability that all 20 syringes are still defect-free, and what does comparing the two probabilities suggest?

1) $0.95^{20} = 0.358$; this large drop shows that the "all defect-free" check has decent power to catch a shift from a 0.5% to a 5% defect rate.
2) $0.95^{20} = 0.358$; since both probabilities are below 1, the test cannot distinguish between the two defect rates.
3) $0.995^{20} \times 10 = 9.05$; a nonsensical value greater than 1, so the probability must be computed a different way.
4) $0.05^{20} \approx 0$; there is essentially no chance of an all defect-free box if $p = 0.05$.
5) 0.905; the probability is unaffected by an increase in p, since only 20 syringes are tested regardless of the true rate.

2.6 Newspaper consumer survey

Exercise 2.6 Newspaper consumer survey

In a consumer survey performed by a newspaper, 20 different groceries were purchased in a grocery store. Discrepancies between the price on the sales slip and the shelf price were found in 6 of these 20 products.

a) Let X denote the number of discrepancies when purchasing 3 random (different) products within the group of 20. What is the mean and variance of X?

Exam-style questions

b) A biologist tags 15 fish in a pond containing an estimated 90 fish in total, then releases them back. A week later, she catches 12 fish at random from the pond. Let F denote the number of tagged fish among the 12 caught. Which of the following gives the correct mean and variance of F?

1) $\mu = 2.00, \sigma^2 = 1.461$
2) $\mu = 6.67, \sigma^2 = 1.71$
3) $\mu = 2.00, \sigma^2 = 1.667$
4) $\mu = 2.00, \sigma^2 = 1.389$
5) $\mu = 15.00, \sigma^2 = 1.461$

c) A biologist tags 5 out of 20 turtles in a lagoon, then a week later recaptures 4 turtles at random. What is the probability that *exactly 2* of the 4 recaptured turtles are tagged?

1) 0.217
2) 0.470
3) 0.282
4) 0.031
5) 0.211

2.7 A fully automated production

Exercise 2.7 A fully automated production

On a large fully automated production plant, items are pushed to a side band at random time points, from which they are automatically fed to a control unit. The production plant is set up so that the number of items sent to the control unit on average is 1.6 items per minute. Let X denote the number of items pushed to the side band in 1 minute. It is assumed that X follows a Poisson distribution.

a) What is the probability that more than 5 items arrive at the control unit in a given minute?

b) What is the probability that no more than 8 items arrive at the control unit within a 5-minute period?

Exam-style questions

c) A monitoring system logs server errors at an average rate of $\lambda = 2.3$ errors per hour, modelled as a Poisson process. It is given that $P(X > 4) = 1 -$ `stats.poisson.cdf(4, 2.3)` $= 0.0797$. Which of the following expressions instead gives $P(X \geq 4)$, the probability of *at least* 4 errors in an hour?

1) `1 - stats.poisson.cdf(3, 2.3)`
2) `1 - stats.poisson.cdf(4, 2.3)`
3) `stats.poisson.cdf(4, 2.3)`
4) `stats.poisson.pmf(4, 2.3)`
5) `stats.poisson.cdf(3, 2.3)`

d) Continuing the same monitoring system ($\lambda = 2.3$ errors per hour), what is the probability that *no* errors occur during a 3-hour shift?

1) $e^{-6.9} = 0.00101$
2) $e^{-2.3} = 0.1003$
3) $3e^{-2.3} = 0.301$
4) $1 - e^{-6.9} = 0.999$
5) $1/6.9 = 0.145$

2.8 Call center staff

|||| Exercise 2.8 Call center staff

The staffing for answering calls in a company is based on 180 phone calls per hour randomly distributed. If there are 20 calls or more in a period of 5 minutes, the capacity is exceeded and there will be an unwanted waiting time. Hence the capacity is 19 calls per 5 minutes.

a) What is the probability that the capacity is exceeded in a random period of 5 minutes?

b) If the probability of handling all calls without waiting time must be at least 99% for a randomly selected 5-minute period, how large should the capacity per 5 minutes be?

Exam-style questions

c) A different company's helpdesk handles 240 calls per hour, randomly distributed. If 12 or more calls arrive within a given 2-minute window, the helpdesk is considered overloaded. Let X denote the number of calls in a random 2-minute period. Which of the following Python expressions correctly computes the probability that the helpdesk is overloaded?

1) `1 - stats.poisson.cdf(11, 8)`
2) `1 - stats.poisson.cdf(12, 8)`
3) `stats.poisson.cdf(12, 8)`
4) `1 - stats.poisson.cdf(11, 4)`
5) `stats.poisson.pmf(12, 8)`

d) A ticket box office sells tickets at a mean rate of 45 per 15-minute period, modelled as a Poisson process. The manager wants to set a staffing capacity c such that the probability of *not* exceeding capacity is at least 95%

in a given 15-minute period. She finds that `stats.poisson.cdf(56, 45)` $= 0.947$ and `stats.poisson.cdf(57, 45)` $= 0.958$. What is the minimum staffing capacity c that meets the 95% requirement?

1) $c = 57$
2) $c = 56$
3) $c = 45$
4) $c = 58$
5) There is no such c; a 95% guarantee is never possible for a Poisson process.

2.9 Continuous random variable

Exercise 2.9 Continuous random variable

a) The following Python expressions evaluate to three different probabilities:

```
print(stats.norm.cdf(2))
print(stats.norm.cdf(2, 1, 1))
print(stats.norm.cdf(2, 1, 2))
```

Specify which distributions are used and explain the resulting probabilities.

b) What is the result of `stats.norm.ppf(stats.norm.cdf(2))`?

c) The following expressions return three quantiles:

```
print(stats.norm.ppf(0.975))
print(stats.norm.ppf(0.975, 1, 1))
print(stats.norm.ppf(0.975, 1, 2))
```

State what each number represents.

Exam-style questions

d) For a manufacturing process, the following three Python expressions are computed:

```
print(stats.norm.cdf(50, 45, 5))
print(stats.norm.cdf(50, 50, 5))
print(stats.norm.cdf(50, 55, 5))
```

which give, in some order, the values 0.500, 0.159 and 0.841. Which of the following correctly matches each line to its value?

1) Line 1 = 0.841, Line 2 = 0.500, Line 3 = 0.159.

2) Line 1 = 0.159, Line 2 = 0.500, Line 3 = 0.841.
3) Line 1 = 0.500, Line 2 = 0.841, Line 3 = 0.159.
4) Line 1 = 0.841, Line 2 = 0.159, Line 3 = 0.500.
5) All three lines give the same value, 0.500, since $x = 50$ is used in every call.

e) Let $Y \sim N(10, 3^2)$. Given that `stats.norm.ppf(0.90)` $= 1.282$ (the 90th percentile of the standard normal), which of the following correctly gives the 90th percentile of Y?

1) $10 + 3 \times 1.282 = 13.85$
2) 1.282
3) $10 \times 1.282 = 12.82$
4) $(10 + 1.282) \times 3 = 33.85$
5) $10 - 3 \times 1.282 = 6.15$

2.10 The normal pdf

|||| Exercise 2.10 The normal pdf

a) Which of the following statements regarding the probability density function of the normal distribution $N(1, 2^2)$ is <u>false</u>?

1. The total area under the curve is equal to 1.0
2. The mean is equal to 1^2
3. The variance is equal to 2
4. The curve is symmetric about the mean
5. The two tails of the curve extend indefinitely

Let X be normally distributed with mean 24 and variance 16.

b) Calculate the following probabilities:
   - $P(X \leq 20)$
   - $P(X > 29.5)$
   - $P(X = 23.8)$

Exam-style questions

c) For $X \sim N(3, 5^2)$, which of the following statements is <u>false</u>?
   1) The variance of X is 25.
   2) The median of X equals its mean, 3.
   3) Approximately 95% of the probability mass lies within 2 standard deviations of the mean, i.e. in the interval $[-7, 13]$.
   4) The skewness of X is positive, since $\sigma = 5$ is fairly large.
   5) $P(X < 3) = P(X > 3) = 0.5$.

d) Let Y be normally distributed with mean 52 and standard deviation 6. Which of the following Python expressions correctly computes $P(45 < Y \leq 60)$?

1) `stats.norm.cdf(60, 52, 6) - stats.norm.cdf(45, 52, 6)`
2) `stats.norm.cdf(60, 52, 6) + stats.norm.cdf(45, 52, 6)`
3) `stats.norm.cdf(45, 52, 6) - stats.norm.cdf(60, 52, 6)`
4) `stats.norm.pdf(60, 52, 6) - stats.norm.pdf(45, 52, 6)`
5) `1 - stats.norm.cdf(60, 52, 6) - stats.norm.cdf(45, 52, 6)`

2.11 Computer chip control

|||| Exercise 2.11 Computer chip control

A machine for checking computer chips uses on average 65 milliseconds per check with a standard deviation of 4 milliseconds. A newer machine, potentially to be bought, uses on average 54 milliseconds per check with a standard deviation of 3 milliseconds. Check times can be assumed normally distributed and independent.

a) What is the probability that the time savings per check using the new machine is less than 10 milliseconds?

b) What is the mean μ and standard deviation σ for the total time to check 100 chips on the new machine?

Exam-style questions

c) A courier company's Route A delivery time is $X_A \sim N(40, 6^2)$ minutes and Route B's is $X_B \sim N(35, 4^2)$ minutes, independent of each other. Let $D = X_A - X_B$ be the difference in delivery times. What is the variance of D?

1) $\sigma_D^2 = 6^2 + 4^2 = 52$
2) $\sigma_D^2 = 6^2 - 4^2 = 20$
3) $\sigma_D^2 = (6-4)^2 = 4$
4) $\sigma_D^2 = 6^2 \times 4^2 = 576$
5) $\sigma_D^2 = (6+4)^2 = 100$

d) Continuing the courier company, Route B's delivery time is $X_B \sim N(35, 4^2)$ minutes, independent across deliveries. What is the probability that the total time for 50 independent Route B deliveries exceeds 1770 minutes?

1) ≈ 0.240
2) ≈ 0.760
3) ≈ 0.106
4) ≈ 0.460
5) ≈ 0.500

2.12 Concrete items

|||| Exercise 2.12 Concrete items

A manufacturer of concrete items knows that the length L of his items is reasonably normally distributed with $\mu_L = 3000$ mm and $\sigma_L = 3$ mm. The requirement is that the length must satisfy 2993 mm $\leq L \leq$ 3007 mm.

a) What is the expected error rate in the manufacturing?

b) The concrete items are supported by beams. For the items to be supported correctly, the requirement is 90 mm $< L - L_{\text{beam}} <$ 110 mm, where L_{beam} is the (normally distributed) distance between beams with mean $\mu_{\text{beam}} = 2900$ mm. How large may the standard deviation σ_{beam} be if the requirement must be fulfilled in 99% of cases?
(Quantile: $z_{0.995} = 2.576$)

Exam-style questions

c) A bolt manufacturer requires the diameter $D \sim N(10, 0.05^2)$ mm to satisfy 9.90 mm $\leq D \leq$ 10.08 mm — note that this interval is *not* symmetric about the mean 10. What is the correct way to compute the expected proportion of bolts falling outside this specification?

1) $P(D < 9.90) + P(D > 10.08)$, with each tail computed separately.
2) $2 \times P(D < 9.90)$, since the mean lies inside the interval.
3) $2 \times P(D > 10.08)$, since defects are more likely on the high side.
4) $P(D < 9.90) \times P(D > 10.08)$
5) $P(D < 9.90) - P(D > 10.08)$

d) Window panes have height $H \sim N(1200, \sigma_H^2)$ mm and are installed above door frames with mean height $\mu_{\text{frame}} = 1150$ mm and standard deviation

$\sigma_{\text{frame}} = 8$ mm. The installation gap must satisfy 30 mm $< H - H_{\text{frame}} <$ 70 mm in 95% of installations. How large may σ_H be? *(Quantile: $z_{0.975} =$ 1.960)*

1) $\sigma_H \approx 6.33$ mm
2) $\sigma_H \approx 10.20$ mm
3) $\sigma_H \approx 12.65$ mm
4) $\sigma_H \approx 40.12$ mm
5) $\sigma_H \approx 1.96$ mm

2.13 Online statistics video views

Exercise 2.13 Online statistics video views

In 2013, there were 110,000 views of the DTU statistics videos available online. Assume that the occurrence of views through 2014 follows a Poisson process with the 2013 average: $\lambda_{365\text{days}} = 110{,}000$.

a) What is the probability that in a randomly chosen half-hour period there are no views?

b) There has just been a view. What is the probability of waiting more than 15 minutes for the next view?

Exam-style questions

c) A popular podcast received 84,000 downloads over the course of a 365-day year, modelled as a Poisson process. Which of the following Python expressions correctly computes the rate parameter (intensity) of downloads per 10-minute period?

1) `84000 / (365*24*6)`
2) `84000 / (365*24*60)`
3) `84000 * (365*24*6)`
4) `84000 / (365*6)`
5) `84000 / 365`

d) Continuing the podcast example, the mean rate is $\lambda_{10\text{min}} \approx 1.598$ downloads per 10-minute period. A listener just finished downloading an episode. Which of the following correctly computes the probability of waiting more than 20 minutes (two 10-minute periods) for the next download, and how does it relate to the exponential waiting-time distribution?

1) $P(X = 0) = e^{-2\times1.598} = e^{-3.196} = 0.0409$, using the Poisson count over 2 periods — this is exactly equal to $P(T > 20)$ from the exponential waiting-time distribution, since the two always agree for a Poisson process.
2) $P(X = 0) = e^{-1.598} = 0.202$, forgetting to scale λ up to the 20-minute window.
3) The Poisson and exponential approaches give different answers here, since they describe fundamentally different aspects of the process.
4) $1 - e^{-3.196} = 0.959$, the complement of the correct probability.
5) $2 \times e^{-1.598} = 0.404$, incorrectly scaling the probability itself instead of the rate.

2.14 Body mass index distribution

|||| Exercise 2.14 Body mass index distribution

The BMI (Body Mass Index) is defined as the weight (W) in kg divided by the squared height (H) in metres: $BMI = W/H^2$. Assume that the population distribution of BMI follows a log-normal distribution with $\alpha = 3.1$ and $\beta = 0.15$ (i.e. $\log(BMI) \sim N(3.1,\ 0.15^2)$).

a) A definition of "being obese" is a BMI value of at least 30. What proportion of the population would be obese?

Exam-style questions

b) Income in a certain population follows a log-normal distribution with $\log(\text{Income}) \sim N(10.5, 0.4^2)$ (income measured in dollars). Which of the following Python expressions correctly computes the probability that a randomly selected person earns more than $50,000?

1) `1 - stats.lognorm.cdf(50000, s=0.4, scale=np.exp(10.5))`
2) `1 - stats.lognorm.cdf(50000, s=10.5, scale=np.exp(0.4))`
3) `1 - stats.lognorm.cdf(50000, s=0.4, scale=10.5)`
4) `1 - stats.norm.cdf(50000, 10.5, 0.4)`
5) `1 - stats.lognorm.cdf(np.log(50000), s=0.4, scale=np.exp(10.5))`

c) For the income distribution above ($\log(\text{Income}) \sim N(10.5, 0.4^2)$), which of the following statements is true?

1) The median income is $e^{10.5} \approx$ \$36,320, which is *less* than the mean income $e^{10.5+0.4^2/2} \approx$ \$39,340, because the log-normal distribution is right-skewed.
2) The median and mean income are both exactly $e^{10.5} \approx$ \$36,320, since $\log(\text{Income})$ is normally (and hence symmetrically) distributed.
3) The median income is 10.5 dollars, reading the parameter off directly.

4) The mean income is *less* than the median income, since log-normal distributions are left-skewed.
5) The mean and median cannot be determined without simulating the distribution.

2.15 Bivariate normal

|||| Exercise 2.15 Bivariate normal

a) In the bivariate normal distribution, show that if Σ is a diagonal matrix then X_1 and X_2 are independent and each follows a univariate normal distribution.

b) Let Z_1 and Z_2 be independent standard normal random variables. Define:

$$\begin{aligned} X &= a_{11}Z_1 + c_1, \\ Y &= a_{12}Z_1 + a_{22}Z_2 + c_2. \end{aligned}$$

Show that an appropriate choice of $a_{11}, a_{12}, a_{22}, c_1, c_2$ can give any bivariate normal distribution for (X, Y) with given $\boldsymbol{\mu}$ and covariance matrix Σ.

c) Use the result to simulate 1000 realisations of a bivariate normal random variable with $\boldsymbol{\mu} = (1, 2)^\top$ and

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

and make a scatter plot of the two components.

Exam-style questions

d) Let (X, Y) follow a bivariate normal distribution with correlation $\rho = 0$ (i.e. $\text{Cov}(X, Y) = 0$). Which of the following statements is correct?

1) X and Y must be independent — for *jointly normal* random variables, zero correlation implies independence, a special property that does not hold for arbitrary random variables.

2) X and Y are uncorrelated but not necessarily independent, since zero correlation never implies independence for any pair of random variables.

3) X and Y must be identically distributed, since $\rho = 0$ forces $\sigma_X = \sigma_Y$.

4) X and Y cannot both be normally distributed if $\rho = 0$.

5) There is not enough information to say anything about the relationship between X and Y.

e) Using the same linear-transformation method ($X = a_{11}Z_1 + c_1, Y = a_{12}Z_1 + a_{22}Z_2 + c_2$, for independent standard normal Z_1, Z_2) as in the previous question, which choice of coefficients correctly simulates a bivariate normal with $\boldsymbol{\mu} = (0, 5)^\top$ and $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 5 \end{pmatrix}$?

1) $a_{11} = 2, a_{12} = 1, a_{22} = 2, c_1 = 0, c_2 = 5$

2) $a_{11} = 4, a_{12} = 2, a_{22} = 5, c_1 = 0, c_2 = 5$

3) $a_{11} = 2, a_{12} = 2, a_{22} = 1, c_1 = 0, c_2 = 5$

4) $a_{11} = 2, a_{12} = 1, a_{22} = 2, c_1 = 5, c_2 = 0$

5) $a_{11} = \sqrt{2}, a_{12} = 1, a_{22} = 2, c_1 = 0, c_2 = 5$

2.16 Sample distributions

|||| Exercise 2.16 Sample distributions

a) Verify by simulation that

$$\frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 \sim \chi^2(n_1 + n_2 - 2),$$

where S_p^2 is the pooled sample variance. You may use $n_1 = 5$, $n_2 = 8$, $\mu_1 = 2$, $\mu_2 = 4$, and $\sigma^2 = 2$.

b) Show that if $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$, then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Verify the result by simulation.

Exam-style questions

c) Three independent samples of sizes $n_1 = 6$, $n_2 = 9$ and $n_3 = 4$ are drawn from populations sharing the same variance σ^2, and a pooled variance estimate S_p^2 is computed by pooling all three sample variances. What are the degrees of freedom of $\frac{n_1 + n_2 + n_3 - 3}{\sigma^2} S_p^2$?

1) 16
2) 19
3) 3
4) 13
5) 17

d) A statistic is constructed as $T = \frac{V}{\sqrt{W/m}}$, where $V \sim N(0,1)$ and $W \sim \chi^2(m)$. What is the distribution of T, and what key condition is required for this to hold?

1) $T \sim t(m)$, provided V and W are independent — this is precisely the definition of the t-distribution with m degrees of freedom.
2) $T \sim N(0,1)$, since dividing by $\sqrt{W/m}$ has no effect asymptotically.
3) $T \sim \chi^2(m)$, since W is chi-squared distributed.
4) $T \sim t(m)$ regardless of whether V and W are independent.
5) $T \sim F(1,m)$, since T is built from a normal and a chi-squared variable.

2.17 Sample distributions 2

Exercise 2.17 Sample distributions 2

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent, with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$. Let S_1^2 and S_2^2 be the respective sample variances. Define:

$$Q = \frac{S_1^2}{S_2^2}.$$

a) For $n \in \{2, 4, 8, 16, 32\}$, find:
   1. $P(Q < 1)$
   2. $P(Q > 2)$
   3. $P(Q < 1/2)$
   4. $P(1/2 < Q < 2)$

b) For at least one value of n, verify the results above by direct simulation. You may use any values of μ_1, μ_2 and σ^2.

Exam-style questions

c) Two independent samples of sizes $n_1 = 10$ and $n_2 = 15$ are drawn from normal populations with equal variance σ^2. Let $Q = S_1^2/S_2^2$. What is the distribution of Q?
   1) $Q \sim F(9, 14)$
   2) $Q \sim F(10, 15)$
   3) $Q \sim F(14, 9)$
   4) $Q \sim t(23)$
   5) $Q \sim \chi^2(23)$

d) The previous exercise noted that $P(Q < 1)$ stays near 0.5 "by symmetry of the F-distribution around 1" when $n_1 = n_2 = n$ (so $\nu_1 = \nu_2 = n - 1$). If instead $n_1 = 10$ and $n_2 = 20$ (so $\nu_1 = 9 \neq \nu_2 = 19$), is $P(Q < 1)$ still exactly 0.5?

1) No — the F-distribution is only symmetric around 1 when $\nu_1 = \nu_2$; with unequal degrees of freedom, $P(Q < 1) \neq 0.5$ in general.
2) Yes — the F-distribution is always symmetric around 1, regardless of its degrees of freedom.
3) No — $P(Q < 1)$ becomes exactly 0 whenever $\nu_1 \neq \nu_2$.
4) Yes, but only because S_1^2 and S_2^2 are independent.
5) No — Q no longer follows an F-distribution at all once $n_1 \neq n_2$.

2.18 Stochastic variable

Exercise 2.18 Stochastic variable

Let X be a stochastic variable with the following probability density function ($f(x)$):

X	0	1	2	3
$f(x)$	0.1	0.3	0.4	0.2

a) What is the sample space of X? Is X a discrete or a continuous stochastic variable?

b) What is $P(X = 3)$ (the probability that X assumes the value 3)?

c) What is $P(X = 4)$ (the probability that X assumes the value 4)?

d) What is the mean of X, $\mathrm{E}[X]$ (also called the expectation value)?

e) What is the variance of X, $\mathrm{V}[X]$?
(This is question 3 from the May 2022 exam)

Exam-style questions

f) A stochastic variable Y has sample space $\{1, 2, 3, 4\}$ and probability density function $f(y)$, with $f(1) = 0.15$, $f(2) = g$, $f(3) = 0.25$, $f(4) = 0.10$, where g is unknown. What is g, and what is $P(Y \geq 3)$?

1) $g = 0.50$, $P(Y \geq 3) = 0.35$
2) $g = 0.50$, $P(Y \geq 3) = 0.10$
3) $g = 0.60$, $P(Y \geq 3) = 0.35$
4) $g = 0.50$, $P(Y \geq 3) = 0.85$
5) $g = 0.40$, $P(Y \geq 3) = 0.35$

g) Consider a stochastic variable Z with sample space $\{0, 1, 2\}$ and probability density function $f(z)$, with $f(0) = 0.2$, $f(1) = 0.5$, $f(2) = 0.3$. The following code was run in Python:

```
z = np.array([0, 1, 2])
fz = np.array([0.2, 0.5, 0.3])
print(np.sum(z * fz))

1.1

print(np.sum(z**2 * fz))

1.7
```

What is $\mathrm{V}[Z]$?

1) 0.49
2) 1.7
3) 0.6
4) 1.21
5) 2.91