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?

|||| Solution

The binomial distribution with $n = 10$ and $p = 0.6$. The value 0.1115 is $P(X = 4)$: the probability of exactly 4 successes in 10 independent draws, each with success probability 60%.

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 $P(X = 5)$.

|||| Solution

$$
\begin{aligned}
P(X \leq 5) &= 0.3669, \\
P(X < 5) = P(X \leq 4) &= 0.1662, \\
P(X > 4) = 1 - P(X \leq 4) &= 1 - 0.1662 = 0.8338, \\
P(X = 5) = P(X \leq 5) - P(X \leq 4) &= 0.3669 - 0.1662 = 0.2007.
\end{aligned}
$$

```
X = stats.binom(10, 0.6)
print(X.cdf(5), X.cdf(4), 1-X.cdf(4), X.pmf(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?

|||| Solution

The Poisson distribution with $\lambda = 3$. The value 0.168 is $P(X = 4)$: the probability of observing exactly 4 events per interval when the average rate is $\lambda = 3$ events per interval.

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 $P(X = 5)$.

|||| Solution

$$\begin{aligned}
P(X \leq 5) &= 0.9161, \\
P(X < 5) = P(X \leq 4) &= 0.8153, \\
P(X > 4) = 1 - P(X \leq 4) &= 1 - 0.8153 = 0.1847, \\
P(X = 5) = P(X \leq 5) - P(X \leq 4) &= 0.9161 - 0.8153 = 0.1008.
\end{aligned}$$

```
X = stats.poisson(3)
print(X.cdf(5), X.cdf(4), 1-X.cdf(4), X.pmf(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.

|||| Solution

Correct answer: **1)**. `stats.binom.pmf(k, n, p)` evaluates the binomial probability mass function $P(X = k)$ for $X \sim \text{Bin}(n, p)$, so here it gives $P(X = 2)$ for $X \sim \text{Bin}(6, 0.4)$: the probability of *exactly* 2 rainy days among the 6, summing over all $\binom{6}{2} = 15$ ways those 2 rainy days could be arranged among the 6. Option 2) describes $P(X \leq 2)$ (the cdf) rather than the pmf. Option 3) describes $P(X \geq 2)$. Option 4) omits the $\binom{6}{2} = 15$ factor: it corresponds to $p^2(1-p)^4 = 0.4^2 \cdot 0.6^4 = 0.0207$, the probability of rain on two *specific* named days and no others, not on any 2 of the 6. Option 5) describes $P(X = 0)$.

```
X = stats.binom(6, 0.4)
print(X.pmf(2))
```

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

|||| Solution

Correct answer: **2)**. The event $\{4 \leq X \leq 7\}$ is the same as $\{X \leq 7\}$ with $\{X \leq 3\}$ removed, so

$$P(4 \leq X \leq 7) = P(X \leq 7) - P(X \leq 3) = 0.8666 - 0.2650 = 0.6016.$$

Option 1) forgets to subtract the lower tail $P(X \leq 3)$. Option 3) makes an off-by-one error, subtracting $P(X \leq 4) = 0.4405$ instead of $P(X \leq 3)$: $0.8666 - 0.4405 = 0.4261$. Option 4) incorrectly computes the complement $1 - 0.6016$ instead of the interval probability itself. Option 5) incorrectly adds the two given probabilities instead of subtracting them.

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?

|||| Solution

Let X be the number of students not passing. Then $X \sim Bin(n = 250,\ p = 0.20)$, and the mean and variance formulas for the binomial give:

$$\mu = np = 0.20 \times 250 = 50,$$
$$\sigma^2 = np(1-p) = 250 \times 0.20 \times 0.80 = 40, \quad \sigma = \sqrt{40} = 6.32.$$

So $\mu = 50$ and $\sigma = 6.32$.

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.

|||| Solution

Correct answer: **1)**. With $Y \sim \text{Bin}(180, 0.12)$: $\mu = np = 180 \times 0.12 = 21.6$ and $\sigma = \sqrt{np(1-p)} = \sqrt{180 \times 0.12 \times 0.88} = \sqrt{19.008} = 4.359$.

Both the mean and standard deviation are counts of crossings, not percentages (rules out 2) and 3)). The mean is an average, not a guaranteed outcome, so it need not be an integer and does not mean exactly 21.6 crossings are delayed every season (rules out 4)). Option 5) swaps the mean and standard deviation.

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

|||| Solution

Correct answer: **1)**. With $Z \sim \text{Bin}(300, 0.15)$:

$$\sigma = \sqrt{np(1-p)} = \sqrt{300 \times 0.15 \times 0.85} = \sqrt{38.25} = 6.19.$$

Option 2) is the mean $np = 45$, not the standard deviation.

Option 3) makes the common mistake of using $\sqrt{np} = \sqrt{45} = 6.71$ (a Poisson-style approximation), forgetting the $(1-p)$ factor.

Option 4) is the variance $np(1-p) = 38.25$ before taking the square root.

Option 5) is just the success probability p.

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 $P(X = 0)$.

|||| Solution

X follows the hypergeometric distribution with $N = 6$, $a = 1$ (one "4"-note), and $n = 2$ draws. The mean and variance are:

$$\mu_X = n\frac{a}{N} = \frac{2}{6} = \frac{1}{3},$$
$$\sigma_X^2 = n\frac{a(N-a)}{N^2} \cdot \frac{N-n}{N-1} = 2 \cdot \frac{1 \times 5}{36} \cdot \frac{4}{5} = \frac{8}{36} = \frac{2}{9}.$$

The probability formula gives:

$$P(X = 0) = \frac{\binom{1}{0}\binom{5}{2}}{\binom{6}{2}} = \frac{10}{15} = \frac{2}{3}.$$

So $\mu_X = 1/3$, $\sigma_X^2 = 2/9$ and $P(X = 0) = 2/3$.

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?

|||| Solution

With replacement, each draw is independent with success probability $p = 1/6$. The binomial pdf gives:

$$P(X=0) = \binom{2}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^2 = \frac{25}{36} = 0.694.$$

```
print(stats.binom.pmf(0, 2, 1/6))
```

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

|||| Solution

Correct answer: **1)**. W follows the hypergeometric distribution with $N = 15$, $a = 4$, $n = 3$:

$$P(W=1) = \frac{\binom{4}{1}\binom{11}{2}}{\binom{15}{3}} = \frac{4 \times 55}{455} = 0.484.$$

Option 2) is $P(W=0) = \binom{11}{3}/\binom{15}{3} = 0.363$: no faulty boards, not exactly one. Option 3) is $P(W=2) = \binom{4}{2}\binom{11}{1}/\binom{15}{3} = 0.145$: two faulty boards, not one. Option 4) is the naive proportion $a/N = 4/15 = 0.267$, which ignores the combinatorics of choosing 3 boards altogether. Option 5) wrongly treats the draws as independent (i.e. *with* replacement), giving the binomial value $\binom{3}{1}(4/15)^1(11/15)^2 = 0.430$ instead of the hypergeometric one.

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.

|||| Solution

Correct answer: **1)**. For the hypergeometric distribution, $\mu = n\,a/N = 3 \times 4/15 = 0.8$, exactly matching the binomial mean $np = 3 \times (4/15) = 0.8$. Sampling without replacement changes the *variance* (via the finite population correction factor $\frac{N-n}{N-1}$, which makes the hypergeometric variance *smaller* than the binomial one, the opposite of what option 5) claims) but never the mean. Options 2) and 3) describe plausible-sounding but incorrect intuitions about how removal without replacement would shift the average.

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?

|||| Solution

Let X denote the number of discrepancies when buying 3 different products from the 20. Since products are bought without replacement, X follows the hypergeometric distribution with $N = 20$, $a = 6$, $n = 3$:

$$P(X = 0) = \frac{\binom{6}{0}\binom{14}{3}}{\binom{20}{3}} = \frac{1 \times 364}{1140} = 0.319.$$

With Python the answer would look like this:

```
print(stats.hypergeom.pmf(0, M=20, n=6, N=3))
0.319
```

Hence the probability that none of the 3 products have a discrepancy is 0.319.

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.

|||| Solution

Correct answer: **1)**. SciPy documents `hypergeom.pmf(k, M, n, N)` with `M` = total population size, `n` = number of objects in the population carrying the property of interest, and `N` = number of draws — i.e. SciPy's `n` plays the role of the book's a, and SciPy's `N` plays the role of the book's n. This is a genuine, common source of confusion (same letters, different roles), not an error in the code. Option 2) swaps the two arguments the wrong way and would compute a different (incorrect) quantity. Options 3), 4) and 5) are all false statements about how SciPy's function works.

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

|||| Solution

Correct answer: **1)**. Let X be the number of defective bulbs among the 6 selected, $X \sim \text{Hypergeometric}(N = 30, a = 5, n = 6)$. Using the complement rule:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{5}{0}\binom{25}{6}}{\binom{30}{6}} = 1 - \frac{177100}{593775} = 1 - 0.298 = 0.702.$$

Option 2) is $P(X = 0)$ itself — the answer to “no defective bulbs”, not “at least one”, i.e. forgetting to take the complement. Option 3) is the naive proportion $5/30 = 0.167$, ignoring the combinatorics of selecting 6 bulbs. Option 4) would only be correct if it were certain that a defective bulb must be drawn, which it is not.

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 $P(X \geq 1)$.

|||| Solution

X follows the hypergeometric distribution (sampling without replacement) with $N = 20$, $a = 2$, $n = 3$:

$$\mu_X = n\frac{a}{N} = 3 \cdot \frac{2}{20} = 0.3,$$
$$\sigma_X^2 = n\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = 3 \cdot \frac{2}{20} \cdot \frac{18}{20} \cdot \frac{17}{19} = 0.2416.$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{18}{3}}{\binom{20}{3}} = 1 - \frac{816}{1140} = 0.2842.$$

So $\mu_X = 0.3$, $\sigma_X^2 = 0.242$ and $P(X \geq 1) = 0.2842$.

```
print(stats.hypergeom(M=20, n=2, N=3).mean(),
      stats.hypergeom(M=20, n=2, N=3).var(),
      1 - stats.hypergeom.pmf(0, M=20, n=2, N=3))
```

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}).

Solution

Since the supplier has a very large stock, we use the binomial distribution (sampling effectively with replacement from a large population):

$$p_1 = P(X = 0 \mid p = 0.01) = 0.99^{10} = 0.9044,$$
$$p_{10} = P(X = 0 \mid p = 0.10) = 0.90^{10} = 0.3487.$$

```
print(stats.binom.pmf(0, 10, 0.01), stats.binom.pmf(0, 10,
    0.10))
```

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) $\mu = 3.00, \sigma^2 = 0.370$

Solution

Correct answer: **1)**. With $V \sim \text{Hypergeometric}(N = 25, a = 3, n = 4)$:

$$\mu_V = n\frac{a}{N} = 4 \times \frac{3}{25} = 0.48,$$
$$\sigma_V^2 = n\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = 4 \times 0.12 \times 0.88 \times \frac{21}{24} = 0.370.$$

Option 2) omits the finite population correction factor $\frac{N-n}{N-1} = \frac{21}{24}$, using the binomial-style variance $np(1-p) = 0.4224$ instead. Option 3) uses the naive proportion $a/N =$

0.12 as the mean, instead of $n \cdot a/N$. Option 4) wrongly assumes variance equals the mean (a Poisson-style mistake). Option 5) confuses $a = 3$ itself for the mean.

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.

|||| Solution

Correct answer: **1)**. $P(\text{all defect-free}) = (1-p)^{20}$. For $p = 0.05$: $0.95^{20} = 0.358$, notably lower than 0.905 for $p = 0.005$: a supplier secretly shipping 10 times the claimed defect rate would still often pass the all-defect-free check, but far less often than a supplier truly meeting the claim — so the check does carry some (imperfect) power to detect the higher defect rate. Option 4) mistakenly raises p itself to the 20th power instead of $(1-p)$. Option 5) incorrectly claims the probability does not depend on p.

```
print(stats.binom.pmf(0, 20, 0.05))
```

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?

Solution

X follows the hypergeometric distribution (without replacement) with $N = 20$, $a = 6$, $n = 3$:

$$\mu_X = n \cdot \frac{a}{N} = 3 \cdot \frac{6}{20} = 0.90,$$
$$\sigma_X^2 = n \cdot \frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = 3 \cdot \frac{6}{20} \cdot \frac{14}{20} \cdot \frac{17}{19} = 0.564.$$

With Python the answer would look like this:

```
print(stats.hypergeom.mean(M=20, n=6, N=3))
0.90

print(stats.hypergeom.var(M=20, n=6, N=3))
0.564
```

So $\mu_X = 0.90$ and $\sigma_X^2 = 0.56$.

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) $\mu = 15.00, \sigma^2 = 1.461$

|||| Solution

Correct answer: **1)**. With $F \sim \text{Hypergeometric}(N = 90, a = 15, n = 12)$:

$$\mu_F = n\frac{a}{N} = 12 \times \frac{15}{90} = 2.00,$$
$$\sigma_F^2 = n\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = 12 \times \frac{1}{6} \times \frac{5}{6} \times \frac{78}{89} = 1.461.$$

Option 2) mistakenly uses $N = 27$ (the sum $15 + 12$) instead of the true pond population $N = 90$.

Option 3) omits the finite population correction $\frac{N-n}{N-1}$.

Option 4) wrongly treats F as binomial (i.e. as if fish were caught *with* replacement), giving $np(1-p) = 1.389$ instead of the hypergeometric variance.

Option 5) confuses $a = 15$ (the number tagged) with the mean.

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

|||| Solution

Correct answer: **1)**. With $X \sim \text{Hypergeometric}(N = 20, a = 5, n = 4)$:

$$P(X = 2) = \frac{\binom{5}{2}\binom{15}{2}}{\binom{20}{4}} = \frac{10 \times 105}{4845} = 0.217.$$

With Python the answer would look like this:

```
print(stats.hypergeom.pmf(2, M=20, n=5, N=4))
0.217
```

Option 2) is $P(X = 1) = 0.470$

Option 3) is $P(X = 0) = 0.282$

Option 4) is $P(X = 3) = 0.031$ – all correct pmf values but for the wrong count of tagged turtles.

Option 5) instead computes the binomial (with-replacement) value $\binom{4}{2}(5/20)^2(15/20)^2 = 0.211$, ignoring that recapturing is without replacement.

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?

Solution

With $\lambda = 1.6$:

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.9940 = 0.006.$$

With Python the answer would look like this:

```
print(1 - stats.poisson.cdf(5, 1.6))
0.006
```

About 0.6% of minutes will have more than 5 items.

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

|||| Solution

Over 5 minutes the rate scales to $\lambda_{5\text{min}} = 5 \times 1.6 = 8$:

$$P(X \leq 8) = 0.593.$$

With Python the answer would look like this:

```
print(stats.poisson.cdf(8, 8))
0.593
```

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)`

|||| Solution

Correct answer: **1)**. $P(X \geq 4) = 1 - P(X \leq 3) = 1 - $ `stats.poisson.cdf(3, 2.3)`.

Option 2) is the given expression for $P(X > 4)$, one boundary off from what is asked.

Option 3) computes $P(X \leq 4)$.

Option 4) computes only $P(X = 4)$.

Option 5) computes $P(X \leq 3)$, forgetting to take the complement.

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) $1/6.9 = 0.145$

|||| Solution

Correct answer: **1)**. Over a 3-hour shift the rate scales to $\lambda_{3\mathrm{h}} = 3 \times 2.3 = 6.9$, so $P(X = 0) = e^{-\lambda_{3\mathrm{h}}} = e^{-6.9} = 0.00101$.

Option 2) forgets to scale λ up to the 3-hour window and uses the 1-hour rate instead.

Option 3) makes the equally common (but equally wrong) mistake of scaling the *probability* by 3 rather than scaling λ before exponentiating.

Option 4) computes the complement, $P(X \geq 1)$, instead of $P(X = 0)$.

Option 5) is not based on the Poisson pmf at all.

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?

|||| Solution

The 5-minute mean is $\lambda_{5\text{min}} = 180/12 = 15$. Let $X \sim \text{Po}(15)$:

$$P(X \geq 20) = 1 - P(X \leq 19) = 0.125.$$

With Python the answer would look like this:

```
print(1 - stats.poisson.cdf(19, 15))
0.125
```

There is a 12.5% chance the capacity is exceeded in a random 5-minute period.

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?

Solution

We need the smallest c such that $P(X \leq c) \geq 0.99$ for $X \sim \text{Po}(15)$. Evaluating for $c = 22, 23, \ldots, 26$:

c	22	23	24	25	26
$P(X \leq c)$	0.9615	0.9762	0.9857	0.9915	0.9950

With Python the answer would look like this:

```
for c in range(22, 27):
    print(c, stats.poisson.cdf(c, 15))
22 0.9615
23 0.9762
24 0.9857
25 0.9915
26 0.9950
```

The first (smallest) capacity level achieving this is $c = 25$. The capacity must be at least 25 calls per 5 minutes.

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)`

|||| Solution

Correct answer: **1)**. First convert the rate to the 2-minute window: $\lambda = 240$ calls/hour $\times \frac{2}{60}$ hour $= 8$. "Overloaded" means $X \geq 12$, so $P(X \geq 12) = 1 - P(X \leq 11) = $ `1 - stats.poisson.cdf(11, 8)`.

Option 2) is off by one, computing $P(X > 12) = P(X \geq 13)$ instead.

Option 3) computes $P(X \leq 12)$, the wrong direction entirely.

Option 4) uses the correct boundary but the wrong rate $\lambda = 240/60 = 4$, forgetting to multiply by the 2-minute window length.

Option 5) computes only the single point probability $P(X = 12)$.

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.

|||| Solution

Correct answer: **1)**. We need the smallest c with $P(X \leq c) \geq 0.95$. Since $P(X \leq 56) = 0.947 < 0.95$ but $P(X \leq 57) = 0.958 \geq 0.95$, the minimum capacity is $c = 57$.

Option 2) picks $c = 56$, which does *not* meet the requirement.

Option 3) confuses the mean rate $\lambda = 45$ with the required capacity.

Option 4) meets the requirement but is not the *minimum* capacity needed.

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.

|||| Solution

The three calls all use the normal cumulative distribution function (CDF). They return $P(X \leq 2)$ for:

- $X \sim N(\mu = 0,\ \sigma^2 = 1)$: $P(X \leq 2) = 0.9772$,
- $X \sim N(\mu = 1,\ \sigma^2 = 1)$: $P(X \leq 2) = 0.8413$,
- $X \sim N(\mu = 1,\ \sigma^2 = 4)$: $P(X \leq 2) = 0.6915$.

For the standard normal, $P(Z \leq 2) \approx 0.977$ represents the shaded area to the left of $z = 2$ under the bell curve.

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

|||| Solution

`ppf` (percent-point function) and `cdf` are inverses of each other, so the result is 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.

|||| Solution

Each number is the 97.5% percentile (i.e. the value x such that $P(X \leq x) = 0.975$) for:

- $N(\mu = 0,\ \sigma^2 = 1)$: $z_{0.975} = 1.960$,
- $N(\mu = 1,\ \sigma^2 = 1)$: 2.960 (shifted right by 1),
- $N(\mu = 1,\ \sigma^2 = 4)$: 4.920 (shifted right by 1 and stretched by 2).

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.

|||| Solution

Correct answer: **1)**. Each call computes $P(X \leq 50)$ for a normal distribution with standard deviation 5 but a different mean:

$$\begin{aligned}
&\text{Line 1: } X \sim N(45, 5^2), \quad z = \tfrac{50-45}{5} = 1, \quad \Phi(1) = 0.841, \\
&\text{Line 2: } X \sim N(50, 5^2), \quad z = \tfrac{50-50}{5} = 0, \quad \Phi(0) = 0.500, \\
&\text{Line 3: } X \sim N(55, 5^2), \quad z = \tfrac{50-55}{5} = -1, \quad \Phi(-1) = 0.159.
\end{aligned}$$

As the mean moves further above $x = 50$, the probability of being *below* 50 decreases — the opposite of what option 2) claims. Option 5) ignores that the mean also affects the probability, not just x.

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) $10 - 3 \times 1.282 = 6.15$

|||| Solution

Correct answer: **1)**. Quantiles of $N(\mu, \sigma^2)$ are obtained from standard-normal quantiles via $x = \mu + \sigma z$, so the 90th percentile of Y is $10 + 3 \times 1.282 = 13.85$. Option 2) forgets to shift and scale the standard-normal quantile at all. Option 3) incorrectly scales the mean instead of the quantile. Option 4) applies the shift and the scaling in the wrong order. Option 5) uses the wrong sign, which would instead give the 10th percentile (since the normal distribution is symmetric, $z_{0.10} = -z_{0.90}$).

```
print(stats.norm.ppf(0.90, 10, 3))
```

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 false?

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

|||| Solution

1. True — the total area under any pdf is 1.
2. True — the mean is 1, and indeed $1^2 = 1$.
3. **False** — for $N(1, 2^2)$ the variance is $2^2 = 4$, not 2.
4. True — the normal distribution is symmetric about its mean.
5. True — the normal density is defined on $(-\infty, \infty)$.

The correct (false) answer is **3**.

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

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

|||| Solution

$X \sim N(24,\, 4^2)$:

$$P(X \le 20) = \Phi\left(\frac{20-24}{4}\right) = \Phi(-1) = 0.1587,$$
$$P(X > 29.5) = 1 - \Phi\left(\frac{29.5-24}{4}\right) = 1 - \Phi(1.375) = 0.0845.$$

For a continuous random variable, $P(X = 23.8) = 0$.

```
print(stats.norm.cdf(20, 24, 4))
print(1 - stats.norm.cdf(29.5, 24, 4))
```

Exam-style questions

c) For $X \sim N(3, 5^2)$, which of the following statements is false?

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) $P(X < 3) = P(X > 3) = 0.5$.

|||| Solution

Correct answer: **4)**. Every normal distribution is perfectly symmetric and has skewness exactly 0, *regardless of how large σ is* — a larger standard deviation only spreads the distribution out, it never introduces skew. The other statements are all true: $\sigma^2 = 5^2 = 25$; the normal distribution's median always equals its mean; about 95% of the mass lies within $\mu \pm 2\sigma = 3 \pm 10 = [-7, 13]$; and by symmetry $P(X < 3) = P(X > 3) = 0.5$.

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

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)`

|||| Solution

Correct answer: **1)**. For any continuous distribution, $P(a < Y \leq b) = P(Y \leq b) - P(Y \leq a) = F(b) - F(a)$, so here `cdf(60,...)` $-$ `cdf(45,...)` ≈ 0.787. Option 2) adds instead of subtracting. Option 3) subtracts in the wrong order, giving a negative number. Option 4) mistakenly uses the density (`pdf`) instead of the cumulative distribution function. Option 5) is not a meaningful combination of the two probabilities.

```
print(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?

|||| Solution

Let $X_{\text{old}} \sim N(65, 4^2)$ and $X_{\text{new}} \sim N(54, 3^2)$. The time saving $U = X_{\text{old}} - X_{\text{new}}$ follows:

$$\mathrm{E}(U) = 65 - 54 = 11,$$
$$\mathrm{V}(U) = 4^2 + 3^2 = 25, \quad U \sim N(11,\ 5^2).$$

$$P(U < 10) = P\left(Z < \frac{10-11}{5}\right) = P(Z < -0.2) = 0.4207.$$

```
print(stats.norm.cdf(10, 11, 5))
```

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

|||| Solution

Let $U = \sum_{i=1}^{100} X_i$ with each $X_i \sim N(54, 3^2)$:

$$\mu = \mathrm{E}(U) = \sum_{i=1}^{100} 54 = 5400 \text{ ms},$$
$$\sigma^2 = \mathrm{V}(U) = \sum_{i=1}^{100} 9 = 900, \quad \sigma = 30 \text{ ms}.$$

So $\mu = 5400$ ms and $\sigma = 30$ ms.

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) $\sigma_D^2 = (6+4)^2 = 100$

|||| Solution

Correct answer: **1)**. For independent random variables, variances always *add*, even when the variables themselves are subtracted: $\mathrm{V}(X_A - X_B) = \mathrm{V}(X_A) + \mathrm{V}(X_B) = 6^2 + 4^2 = 52$. Option 2) mistakenly subtracts the variances. Options 3) and 5) incorrectly combine the standard deviations first (by subtracting or adding) before squaring. Option 4) multiplies the variances instead of adding them.

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

|||| Solution

Correct answer: **1)**. Let $U = \sum_{i=1}^{50} X_i$. Then $\mu_U = 50 \times 35 = 1750$ and $\sigma_U^2 = 50 \times 4^2 = 800$, so $\sigma_U = \sqrt{800} = 28.28$.

$$P(U > 1770) = P\left(Z > \frac{1770 - 1750}{28.28}\right) = P(Z > 0.707) \approx 0.240.$$

Option 2) is the complement, $P(Z \leq 0.707)$. Option 3) mistakenly uses the single-delivery variance 16 directly as σ, i.e. $z = (1770 - 1750)/16 = 1.25$. Option 4) mistakenly scales the standard deviation linearly with n (using $\sigma = 50 \times 4 = 200$) instead of by $\sqrt{n}$. Option 5) wrongly assumes the total is centred at exactly 1770.

```
print(1 - stats.norm.cdf(1770, 50*35, np.sqrt(50*4**2)))
```

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?

|||| Solution

The expected error rate is:

$$P(L < 2993) + P(L > 3007) = 2 \cdot P(L < 2993)$$
$$= 2 \cdot \Phi\left(\frac{2993 - 3000}{3}\right) = 2 \cdot \Phi(-7/3) = 0.0196.$$

About 2% of items will not meet the specification.

```
print(stats.norm.cdf(2993,3000,3) + 1 - stats.norm.cdf
   (3007,3000,3))
```

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: $z_{0.995} = 2.576$)

|||| Solution

We need:

$$P(90 < L - L_{\text{beam}} < 110) = 0.99.$$

Since $\mathrm{E}(L - L_{\text{beam}}) = 3000 - 2900 = 100$ and $\mathrm{V}(L - L_{\text{beam}}) = 9 + \sigma^2_{\text{beam}}$, we need:

$$P(-z_{0.995} < Z < z_{0.995}) = 0.99, \quad \text{where } z_{0.995} = \frac{10}{\sqrt{9 + \sigma^2_{\text{beam}}}} = 2.576.$$

Solving:

$$\sigma_{\text{beam}} = \sqrt{\left(\frac{10}{2.576}\right)^2 - 9} = 2.464 \text{ mm}.$$

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) $P(D < 9.90) - P(D > 10.08)$

|||| Solution

Correct answer: **1)**. Doubling one tail (as is valid when an interval is symmetric about the mean, as in the question above) only works when the interval *is* symmetric. Here $10.08 - 10 = 0.08 \neq 10 - 9.90 = 0.10$, so the two tails must be found separately and added:

$$P(D < 9.90) = \Phi(-2.0) = 0.0228, \qquad P(D > 10.08) = 1 - \Phi(1.6) = 0.0548,$$

giving a total defect proportion of $0.0228 + 0.0548 = 0.0776$, about 7.8%. Options 2) and 3) wrongly assume symmetry. Option 4) combines the tail probabilities as if they needed multiplying (an independent-events-style error), and option 5) subtracts them instead of adding.

```
print(stats.norm.cdf(9.90, 10, 0.05) + 1 - stats.norm.cdf
    (10.08, 10, 0.05))
```

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 $30 \text{ mm} < H - H_{\text{frame}} < 70$ 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

|||| Solution

Correct answer: **1)**. The interval $[30, 70]$ is symmetric about $\mathrm{E}(H - H_{\text{frame}}) = 1200 - 1150 = 50$, with half-width 20. We need:

$$z_{0.975} = \frac{20}{\sqrt{\sigma_H^2 + \sigma_{\text{frame}}^2}} = 1.960 \implies \sigma_H = \sqrt{\left(\frac{20}{1.960}\right)^2 - 8^2} = \sqrt{104.12 - 64} = 6.33 \text{ mm}.$$

Option 2) forgets to account for σ_{frame} at all, using $20/1.960 = 10.20$ directly. Option 3) uses the 99% quantile $z_{0.995} = 2.576$ instead of the correct 95% quantile $z_{0.975} = 1.960$. Option 4) reports the variance $\sigma_H^2 = 40.12$ instead of taking the square root. Option 5) mistakes the quantile itself for the answer.

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?

Solution

The half-hour intensity is:

$$\lambda_{30\text{min}} = \frac{110{,}000}{365 \times 24 \times 2} = 6.28.$$

Let $X \sim Po(6.28)$:

$$P(X = 0) = e^{-6.28} = 0.00187.$$

```
lam = 110000 / (365*24*2)
print(stats.poisson.pmf(0, lam))
```

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

Solution

The 15-minute rate is $\lambda_{15\text{min}} = 6.28/2 = 3.14$. The probability of zero views in 15 minutes is:

$$P(X = 0) = e^{-3.14} = 0.043.$$

Equivalently, using the exponential distribution with mean waiting time $\beta = 365 \times 24 \times 60/110{,}000 = 4.778$ minutes:

$$P(T > 15) = e^{-15/\beta} = e^{-3.14} = 0.043.$$

```
beta = 365*24*60/110000
print(1 - stats.expon.cdf(15, scale=beta))
```

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`

|||| Solution

Correct answer: **1)**. There are $24 \times 6 = 144$ ten-minute periods per day (6 per hour), so $\lambda_{10\text{min}} = 84000/(365 \times 24 \times 6) \approx 1.598$. Option 2) divides by minutes instead of 10-minute periods, giving the per-minute rate instead. Option 3) multiplies instead of dividing. Option 4) forgets the 24 hours per day. Option 5) only converts to a daily rate, not a per-10-minute rate.

```
print(84000 / (365*24*6))
```

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) $P(X = 0) = e^{-2\times 1.598} = e^{-3.196} = 0.0409$, 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.

|||| Solution

Correct answer: **1)**. For a Poisson process, "no events in the next t minutes" (a Poisson count of 0 over rate λt) and "the waiting time until the next event exceeds t" (an exponential tail probability) are the same event, so the two calculations always agree — as also seen in the previous question of this exercise, just here confirmed as a general fact rather than a coincidence. Option 2) forgets to scale λ to the 20-minute window. Option 3) is false — the two approaches are mathematically equivalent. Option 4) computes the complementary probability instead. Option 5) scales the final probability instead of the rate parameter, which is not a valid operation.

```
print(stats.poisson.pmf(0, 2*1.598))
```

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?

|||| Solution

$$P(BMI > 30) = P(\log(BMI) > \log(30)) = P\left(Z > \frac{\log(30) - 3.1}{0.15}\right) = P(Z > 2.008) = 0.0223.$$

About 2.23% of the population would be obese.

```
print(1 - stats.lognorm.cdf(30, s=0.15, scale=np.exp(3.1)))
```

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))`

|||| Solution

Correct answer: **1)**. SciPy's `lognorm` expects the *shape* parameter `s`$= \sigma$ (here 0.4) and `scale`$= e^{\mu}$ (here $e^{10.5}$); it already applies the log-transform internally, so the raw value (50000, not its log) is passed in. Option 2) swaps μ and σ. Option 3) forgets to exponentiate μ for the scale parameter. Option 4) wrongly applies a plain normal distribution directly to the untransformed income value. Option 5) redundantly takes $\log(50000)$ *in addition to* using `lognorm`, effectively log-transforming twice.

```
print(1 - stats.lognorm.cdf(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.

|||| Solution

Correct answer: **1)**. For $\log(X) \sim N(\mu, \sigma^2)$, the median of X is e^{μ}, but the mean is $e^{\mu+\sigma^2/2}$ — strictly larger than the median whenever $\sigma > 0$. This reflects the right skew of the log-normal distribution: symmetry of the underlying normal distribution for $\log(X)$ does *not* carry over to symmetry of X itself, since the exponential transform stretches the upper tail. Option 4) has the direction of skew backwards.

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.

|||| Solution

If $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, then $|\Sigma| = \sigma_1^2\sigma_2^2$ and $\Sigma^{-1} = \mathrm{diag}(1/\sigma_1^2, 1/\sigma_2^2)$. The joint density becomes:

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}}_{f_{X_1}(x_1)} \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}}_{f_{X_2}(x_2)}. \end{aligned}$$

Since the joint density factors into the product of two univariate normal densities, X_1 and X_2 are independent.

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.

|||| Solution

Matching means and covariances:

$$
\begin{aligned}
c_1 &= \mu_X, \quad c_2 = \mu_Y, \\
a_{11} &= \sqrt{\Sigma_{11}}, \\
a_{12} &= \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}}, \\
a_{22} &= \sqrt{\Sigma_{22} - \frac{\Sigma_{12}^2}{\Sigma_{11}}}.
\end{aligned}
$$

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.

|||| Solution

Since all questions involving simulation produce plots, Python code is omitted here; the simulation follows directly by substituting the values above into the formulas of the previous question.

With $\Sigma_{11} = 1$, $\Sigma_{12} = 1$, $\Sigma_{22} = 2$: $a_{11} = 1$, $a_{12} = 1$, $a_{22} = 1$, $c_1 = 1$, $c_2 = 2$. The scatter plot should show a positive correlation between X and Y.

Exam-style questions

d) Let (X, Y) follow a bivariate normal distribution with correlation $\rho = 0$ (i.e. $\mathrm{Cov}(X, Y) = 0$). 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.

|||| Solution

Correct answer: **1)**. This follows directly from the first question of this exercise: when Σ is diagonal (which is exactly what $\rho = 0$ means), the joint density factors into the product of the two marginal densities, which is the definition of independence. In general (for arbitrary, not necessarily normal, random variables) zero correlation only rules out a *linear* relationship and does *not* imply independence — but the bivariate/multivariate normal distribution is a special case where it does. Option 2) states the general-case fact but wrongly applies it here, ignoring the joint-normality assumption.

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

|||| Solution

Correct answer: **1)**. Using the formulas derived in the previous question:

$$
\begin{aligned}
a_{11} &= \sqrt{\Sigma_{11}} = \sqrt{4} = 2, \\
a_{12} &= \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}} = \frac{2}{2} = 1, \\
a_{22} &= \sqrt{\Sigma_{22} - \Sigma_{12}^2/\Sigma_{11}} = \sqrt{5 - 4/4} = \sqrt{4} = 2,
\end{aligned}
$$

with $c_1 = \mu_X = 0$ and $c_2 = \mu_Y = 5$. Option 2) mistakenly plugs the raw entries of Σ in directly, without applying the square-root/division formulas at all. Option 3) swaps a_{12} and a_{22}. Option 4) swaps the two means. Option 5) uses $\sqrt{\Sigma_{12}}$ in place of $\sqrt{\Sigma_{11}}$ for a_{11}.

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 $\sigma^2 = 2$.

|||| Solution

Simulate k independent pairs of samples, compute the pooled variance each time, scale by $(n_1 + n_2 - 2)/\sigma^2$, and compare the histogram to a $\chi^2(11)$ density. The two should match closely for large k. (Python code omitted as the exercise primarily produces diagnostic plots.)

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.

|||| Solution

From the previous question, $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$. Also,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1).$$

Dividing the standard normal by the square root of the (scaled) chi-squared divided by its degrees of freedom yields a $t(n_1 + n_2 - 2)$ distribution by definition.

Simulation verification: compute the t-statistic for k simulated pairs of samples and compare the histogram to the $t(11)$ density. (Python code omitted as the exercise primarily produces diagnostic plots.)

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

|||| Solution

Correct answer: **1)**. Each sample contributes $n_i - 1$ degrees of freedom (one is "used up" estimating that sample's own mean), so pooling three samples gives $(n_1-1)+(n_2-1)+(n_3-1) = n_1+n_2+n_3-3 = 6+9+4-3 = 16$ — the same principle as the two-sample case in the previous question of this exercise (n_1+n_2-2), just generalised to three samples. Option 2) forgets to subtract anything. Option 3) subtracts a fixed 3 regardless of the sample sizes, rather than 1 per sample. Options 4) and 5) subtract the wrong constant (as if pooling 1 or 2 samples, respectively).

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.

|||| Solution

Correct answer: **1)**. This is exactly the mechanism used (implicitly) in the previous question: a standard normal divided by the square root of an independent chi-squared variable (divided by its degrees of freedom) follows a t-distribution with those degrees of freedom, by definition. Independence of V and W is essential — without it, the resulting distribution is generally not $t(m)$, so option 4) is false. Option 5) is a plausible-looking distractor: it is actually T^2 (not T itself) that would follow an $F(1, m)$ distribution.

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. $P(1/2 < Q < 2)$

Solution

Q follows an F-distribution with degrees of freedom $\nu_1 = \nu_2 = n - 1$:

```
for n in [2, 4, 8, 16, 32]:
    F = stats.f(n-1, n-1)
    print(n, F.cdf(1), 1-F.cdf(2), F.cdf(0.5), F.cdf(2)-F.
       cdf(0.5))
```

As n increases, $P(Q < 1)$ stays near 0.5 (by symmetry of the F-distribution around 1), while $P(Q > 2)$ and $P(Q < 1/2)$ decrease (the ratio concentrates near 1).

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.

|||| Solution

Simulate k pairs of samples of size n, compute $Q = S_1^2/S_2^2$ for each pair, and count how often each event occurs. The empirical proportions should match the theoretical F-distribution probabilities for large k. (Python code omitted as the exercise primarily validates via simulation plots.)

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) $Q \sim \chi^2(23)$

|||| Solution

Correct answer: **1)**. The ratio of two independent sample variances follows an F-distribution with degrees of freedom equal to $(n_1 - 1, n_2 - 1)$ for the numerator and denominator samples respectively: $Q = S_1^2/S_2^2 \sim F(9, 14)$. Option 2) uses the sample sizes directly instead of $n_i - 1$. Option 3) swaps the numerator and denominator degrees of freedom (these are *not* interchangeable, unlike for some symmetric distributions). Options 4) and 5) confuse Q with a t- or χ^2-statistic.

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 $n_1 \neq n_2$.

|||| Solution

Correct answer: **1)**. The exact symmetry $P(Q < 1) = P(Q > 1) = 0.5$ relied on $\nu_1 = \nu_2$ in the earlier calculation; it is not a general property of every F-distribution. With $\nu_1 = 9$ and $\nu_2 = 19$, the distribution of Q is skewed and $P(Q < 1)$ will differ from 0.5 (it remains true more generally that if $Q \sim F(\nu_1, \nu_2)$ then $1/Q \sim F(\nu_2, \nu_1)$, but this alone does not force $P(Q < 1) = 0.5$ unless $\nu_1 = \nu_2$). Q still follows an F-distribution regardless of whether $n_1 = n_2$ (ruling out option 5), just not a symmetric one.

```
print(stats.f(9, 19).cdf(1))
```

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?

|||| Solution

The sample space is $\{0, 1, 2, 3\}$. Since X can only take a finite, countable set of values (rather than any value in an interval), X is a *discrete* stochastic variable.

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

|||| Solution

Read directly from the table: $P(X = 3) = f(3) = 0.2$.

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

|||| Solution

The value 4 is not part of the sample space of X, so $P(X = 4) = 0$.

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

|||| Solution

By Definition 2.13 (mean value),

$$\mathrm{E}[X] = \sum_x x \cdot f(x).$$

Applying this to the distribution of X:

$$\mathrm{E}[X] = \sum_x x \cdot f(x) = 0(0.1) + 1(0.3) + 2(0.4) + 3(0.2) = 1.7.$$

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

|||| Solution

By Definition 2.16,

$$\mathrm{V}[X] = \mathrm{E}[(X - \mu)^2].$$

We compute:

$$\mathrm{V}[X] = (0 - 1.7)^2(0.1) + (1 - 1.7)^2(0.3) + (2 - 1.7)^2(0.4) + (3 - 1.7)^2(0.2) = 0.81,$$

Alternatively, use the fundamental identity $\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ (which is Eq. 4-5 in the textbook. If you have not encountered this identity before, it is a worthwhile exercise to prove it directly from the definition of variance).

$$\mathrm{E}[X^2] = 0^2(0.1) + 1^2(0.3) + 2^2(0.4) + 3^2(0.2) = 3.7,$$

$$\mathrm{V}[X] = 3.7 - 1.7^2 = 3.7 - 2.89 = 0.81.$$

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) $g = 0.40$, $P(Y \geq 3) = 0.35$

|||| Solution

Correct answer: **1)**.
Since probabilities must sum to 1: $0.15 + g + 0.25 + 0.10 = 1 \Rightarrow g = 0.50$.
Also we compute: $P(Y \geq 3) = f(3) + f(4) = 0.25 + 0.10 = 0.35$.

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 $V[Z]$?

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

|||| Solution

Correct answer: **1)**. The two printed values are $E[Z] = 1.1$ and $E[Z^2] = 1.7$, so — using the same $V[Z] = E[Z^2] - (E[Z])^2$ formula as in the variance question above —

$$V[Z] = 1.7 - 1.1^2 = 1.7 - 1.21 = 0.49.$$

Note that it is also possible to solve this questions by using Definition 2.16,

$$V[X] = E[(X - \mu)^2].$$

We compute:

$$\mu = E[X] = 0(0.2) + 1(0.5) + 2(0.3) = 1.1,$$
$$V[X] = (0 - 1.1)^2(0.2) + (1 - 1.1)^2(0.5) + (2 - 1.1)^2(0.3) = 0.49,$$