# 3B Distribution of sums, and normal approximation

## Class problems

To calculate CDF values of a normal distribution, you can use a computer (see e.g. the R command `pnorm`), or use tables such as Mellin's tables provided on the course page, or Appendix A1 in Ross's book.

**3B1** (Road salting) To keep the roads free from ice, the officials have stored enough salt for a total snowfall of 200 cm. (We assume, for simplicity, that the amount of salt needed is proportional to the snowfall.) The daily snowfalls have expectation 4.5 cm and standard deviation 2.5 cm.

(a) Using the normal approximation (central limit theorem), calculate an approximate probability that the salt stored will suffice for 50 days.

(b) What additional assumptions did you have to make to be able to solve (a)? Do you think the assumptions are realistic?

**Solution.**

(a) Let the snowfall on $i$th day be $X_i$, and the total snowfall $S = X_1 + \cdots X_{50}$. We were told that $\mathrm{E}(X_i) = 4.5$ and $\mathrm{SD}(X_i) = 2.5$. Then the expected total snowfall is

$$\mathrm{E}(S) = 50 \times 4.5 \;=\; 225.$$

However, we *cannot* calculate the standard deviation of the total snowfall without knowing (or assuming) something about the dependence between the daily snowfalls. Also, we do not know the distribution of the sum.

In order to apply the central limit theorem, let us make the following **additional assumption**.

**Assumption.** The daily snowfalls $X_1, X_2, \ldots$ are independent.

If this holds, then snowfalls have zero correlation, and their sum has standard deviation

$$\begin{aligned} \mathrm{SD}(S) &= \sqrt{\mathrm{SD}(X_1)^2 + \cdots \mathrm{SD}(X_{50})^2} \\ &= \sqrt{50 \times 2.5^2} \\ &= \sqrt{50} \times 2.5 \;\approx\; 17.68. \end{aligned}$$

Also, if the snowfalls are independent, their sum is approximately normally distributed, with the mean $\mathrm{E}(S) = 225$ and standard deviation $\mathrm{SD}(S) \approx 17.68$ we have calculated. So the probability that the total snowfall is at most 200 cm is

$$\mathrm{P}(S \le 200) \;=\; \mathrm{P}\left( \frac{S - \mathrm{E}(S)}{\mathrm{SD}(S)} \le \frac{200 - 225}{17.68} \right) \;\approx\; \mathrm{P}(Z \le -1.41),$$

where $Z$ follows the standard normal distribution. From tables, or by the R command `pnorm(-1.41)`, we find $\mathrm{P}(S \leq 200) \approx 7.9\%$.

(b) We assumed that the snowfalls of different days are stochastically independent. This may not be very realistic.

**3B2** (Winning at the casino) A simplified roulette wheel has 37 slots, numbered from 0 to 36. On every round of the game, Harry bets one euro on his personal lucky number. If the ball lands on that number, he receives 36 euros (net gain +35 because he spent 1 eur on the bet). Otherwise, he loses the euro (net gain $-1$). All slots are equally likely. We denote Harry's net gain after $n$ rounds by $S_n = X_1 + \cdots + X_n$, where $X_i$ is the net gain from round $i$.

(a) Find the mean and standard deviation of $X_i$.

(b) Find the mean and standard deviation of $S_n$.

Applying the normal approximation (central limit theorem), calculate the approximate probability that Harry's net gain is positive

(c) after 30 rounds,

(d) after 3 000 rounds,

(e) after 300 000 rounds.

The first of these probabilities can also be calculated exactly, with relative ease:

(f) Calculate the probability in (c) exactly, and compare to the approximate value.
Hint: What exactly must happen so that Harry's net gain would be zero or negative, after 30 rounds?

**Solution.**

(a) The random variable $X_i$ takes value $r = 35$ with probability $p = 1/37$, and value $-1$ with probability $1 - p$. Thus we have $\mathrm{E}(X_i) = rp + (-1)(1-p) \approx -0.027$ and $\mathrm{SD}(X_i) = (r+1)\sqrt{p(1-p)} \approx 5.84$.

(b) $\mathrm{E}(S_n) = n\,\mathrm{E}(X_1) \approx -0.027 \times n$. $\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X_1) \approx 5.84 \times \sqrt{n}$.

(c)–(e)

$$\begin{aligned} \mathrm{P}(S_n > 0) &= \mathrm{P}\left( \frac{S_n - \mathrm{E}(S_n)}{\mathrm{SD}(S_n)} > \frac{n(1-(r+1)p)}{\sqrt{n}(r+1)\sqrt{p(1-p)}} \right) \\ &\approx \mathrm{P}\left( Z > \frac{n(1-(r+1)p)}{\sqrt{n}(r+1)\sqrt{p(1-p)}} \right) \\ &\approx \mathrm{P}\left( Z > \frac{0.027n}{5.84\sqrt{n}} \right) \\ &= \mathrm{P}(Z > 0.0046\sqrt{n}), \end{aligned}$$

where $Z$ follows the standard normal distribution.

Plugging in the values of $n$, we obtain

| Number of rounds $n$ | Probability of positive gains |
|---:|---:|
| 30 | $\approx 49\%$ |
| 3 000 | $\approx 40\%$ |
| 300 000 | $\approx 0.6\%$ |

(f) If Harry wins even once during the 30 rounds, his net gain will be positive. The complement event is that he lost on *all* 30 rounds, so

$$\mathrm{P}(S_{30} > 0) = 1 - \left(\frac{36}{37}\right)^{30} \approx 56.04\%.$$

This exact probability is slightly different from what we got from the normal approximation (49%), but the approximation was not too bad. In any case it seems Harry has approximately fifty-fifty chances of having won or lost money after 30 rounds.

Note: Because Harry's exact probability for winning here is greater than 1/2, he might think that the game is advantageous to him. Not really, because the *expected* net gain from 30 rounds is negative: $\mathrm{E}(S_{30}) = -0.81$ eur.

## Home problems

**3B3** (Getting enough responses) Opinion pollsters have calculated (by some method) that they need responses from at least 100 people for their opinion poll, in order to have enough data for the statistical inference they wish to conduct. From previous experience, they estimate that when they send their poll to a person, the person has 80% probability of responding, independently from the other persons. Thus they decide to send the poll to 140 persons.

(a) Applying the binomial distribution (without normal approximation), find the probabilities that the pollsters get exactly $x$ responses, for values $x = 100$, 112 and 120. Find the density function of binomial distribution from e.g. Ross's section 5.1. To calculate it, you can use a computer.

(b) Applying the normal approximation (central limit theorem), find approximately the probability that the pollsters get enough responses (at least 100). Use tables or a computer to calculate the CDF.

**Grading.** (a) 0.5 points (b) 0.5 p for correct mean of $S$, 0.5 p for correct standard deviation, 0.5 p for applying the normal approximation correctly. Total rounded up. No penalty for small differences in rounding.

**Solution.**

(a) Let $S$ denote the number of responses. It is the sum

$$S \;=\; X_1 + X_2 + \cdots + X_{140},$$

where the indicator variable $X_i$

$$X_i \;=\; \begin{cases} 1, & \text{if the } i\text{th person responds,} \\ 0, & \text{otherwise.} \end{cases}$$

Because the indicators are independent, their sum $S$ has the binomial distribution with parameters $n = 140$ and $p = 0.8$. The probabilities for *exactly* $x$ responses are obtained from the binomial density (probability mass) function,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

by plugging in the known parameters $n$ and $p$, and the values $x = 100$, $x = 112$ and $x = 140$.

The numerical values are: $f(100) \approx 0.0040$, $f(112) \approx 0.0840$ ja $f(120) \approx 0.0204$.

(b) The indicator $X_i$ has mean

$$\mathrm{E}(X_i) \;=\; \mathrm{P}(X_i = 1) \;=\; 0.8.$$

and from a short calculation we see that also $\mathrm{E}(X_i^2) = 0.8$, thus

$$\mathrm{SD}(X_i) \;=\; \sqrt{\mathrm{E}(X_i^2) - \mathrm{E}(X_i)^2} = 0.4.$$

ex By linearity of expectation

$$\mathrm{E}(S) \;=\; 140 \times 0.8 \;=\; 112.$$

Because the indicators are independent, the standard deviation of the sum is

$$\begin{aligned} \mathrm{SD}(S) \;&=\; \sqrt{\mathrm{SD}(X_1)^2 + \cdots + \mathrm{SD}(X_{140})^2} \\ &=\; \sqrt{140 \times \mathrm{SD}(X_1)^2} \\ &=\; \sqrt{140} \times \mathrm{SD}(X_1) \;\approx\; 4.7329. \end{aligned}$$

By the central limit theorem, we approximate that the sum $S$ is normally distributed with mean $\mu = \mathrm{E}(S) = 112$ and standard deviation $\sigma = \mathrm{SD}(S) \approx 4.7329$. The probability is approximately

$$\mathrm{P}(S \geq 100) \;=\; \mathrm{P}\left(\frac{S-\mu}{\sigma} \geq \frac{100-\mu}{\sigma}\right) \;\approx\; \mathrm{P}(Z \geq -2.5355) \;=\; 1 - \mathrm{P}(Z < -2.5355),$$

where $Z$ follows the standard normal distribution. From tables, or by the R command `1-pnorm(-2.5355)`, we obtain

$$\mathrm{P}(S \geq 100) \;\approx\; 1 - \mathrm{P}(Z < -2.5355) \;\approx\; 99.4\%.$$

**3B4** (Stock portfolios) The shares of two companies, Xanadu and Ypsilon, are on the stock market. Currently both are priced at 100 euros. The next-year return on one share of Xanadu (in euros) is modelled as a random variable $X$ that follows normal distribution with mean $\mu_X = 15$ and standard deviation $\sigma_X = 10$. The return on one share of Ypsilon is modelled as $Y$, which follows normal distribution with parameters $\mu_Y = 10$ and $\sigma_Y = 10$. The returns of the two shares are assumed independent.

(a) Abel buys 200 shares of Xanadu, so he will gain $A = 200X$ during the next year. Find the distribution of $A$. What is the probability that Abel loses money (has negative gain)?

(b) Bertha buys 100 shares of Xanadu and 100 shares of Ypsilon, so she will gain $B = 100X + 100Y$. Find the distribution of $B$. What is the probability that Bertha loses money?

(c) Find the correlation of $A$ and $B$. Are they independent?

(d) Find the distribution of $A - B$. What is the probability that Abel makes more money than Bertha?

Hint. The sum of two independent normally distributed variables is also normally distributed, but what are its parameters? Recall also how scaling (multiplication by constant) works. Note also that $-Y = (-1) \cdot Y$.

**Grading.** 0.5 points per item, total rounded up. No penalty for small differences in rounding.

**Solution.**

(a) $A$ is a constant times $X$, so $A$ also has normal distribution. Its parameters are

$$\begin{aligned} \mu_A &= \mathrm{E}(A) = \mathrm{E}(200X) = 200\mu_X = 3000 \\ \sigma_A &= \mathrm{SD}(A) = \mathrm{SD}(200X) = 200\sigma_X = 2000 \end{aligned}$$

Now $Z = (A - \mu_A)/\sigma_A$ has standard normal distribution, and

$$\mathrm{P}(A < 0) = \mathrm{P}\left(Z < \frac{0 - \mu_A}{\sigma_A}\right) = F_Z\left(\frac{-3000}{2000}\right) = F_Z(-1.5) \approx 6.7\%$$

where $F_Z$ is the CDF of standard normal distribution.

(b) $B$ is the sum of two independent normal variables $100X$ and $100Y$, so $B$ also has normal distribution. Its parameters are

$$\begin{aligned} \mu_B &= \mathrm{E}(B) = \mathrm{E}(100X + 100Y) = 100\mu_X + 100\mu_Y = 2500 \\ \sigma_B &= \mathrm{SD}(B) = \mathrm{SD}(100X + 100Y) = \sqrt{\mathrm{Var}(100X + 100Y)} \\ &= \sqrt{100^2 \,\mathrm{Var}(X) + 100^2 \,\mathrm{Var}(Y)} = \sqrt{100^2 \cdot 10^2 + 100^2 \cdot 10^2} \approx 1414.2 \end{aligned}$$

Now $Z = (B - \mu_B)/\sigma_B$ has standard normal distribution, and

$$\mathrm{P}(B < 0) = \mathrm{P}\left(Z < \frac{0 - \mu_B}{\sigma_B}\right) \approx F_Z\left(\frac{-2500}{1414.2}\right) \approx F_Z(-1.7678) \approx 3.9\%$$

Note that Bertha has smaller expected gain, but also smaller standard deviation, and smaller probability of losing money than Abel, even though both are investing in 200 shares that all have the same standard deviation of 10 euros. Bertha's diversification causes her portfolio to have smaller variance.

(c) Because $A = 200X$ and $B = 100X + 100Y$, applying the linearity of covariance we have

$$\begin{aligned} \mathrm{Cov}(A, B) &= \mathrm{Cov}(200X,\ 100X + 100Y) \\ &= \mathrm{Cov}(200X, 100X) + \mathrm{Cov}(200X, 100Y) \\ &= \mathrm{Cov}(200X, 100X) \qquad \text{(note: } X \text{ and } Y \text{ independent)} \\ &= 200 \cdot 100 \cdot \mathrm{Cov}(X, X) \\ &= 20000 \cdot \mathrm{Var}(X) = 20000 \cdot 100 = 2000000. \end{aligned}$$

Then

$$\operatorname{Cor}(A,B) = \frac{\operatorname{Cov}(A,B)}{\operatorname{SD}(A)\operatorname{SD}(B)} = \frac{2000000}{2000 \cdot 1414.2} \approx 0.71.$$

The correlation is nonzero, so the returns of Abel and Bertha are dependent.

The dependence (and positive correlation) is caused by the fact that they both own some shares of Xanadu.

(d) The difference is the random variable

$$D = A - B = (200X) - (100X + 100Y) = 100X - 100Y = 100X + (-100) \cdot Y.$$

Thus

$$\begin{aligned}
\mu_D &= \operatorname{E}(D) = 100\operatorname{E}(X) - 100\operatorname{E}(Y) = 3000 - 2500 = 500, \\
\sigma_D^2 &= \operatorname{Var}(D) = \operatorname{Var}(100X - 100Y) = 100^2 \operatorname{Var}(X) + (-100)^2 \operatorname{Var}(Y) = 2000000, \\
\sigma_D &= \operatorname{SD}(D) = \sqrt{\operatorname{Var}(D)} \approx 1414.2.
\end{aligned}$$

Now $Z = (D - \mu_D)/\sigma_D$ has standard normal distribution, and

$$\begin{aligned}
\operatorname{P}(D > 0) &= 1 - \operatorname{P}(D \le 0) = 1 - \operatorname{P}\left(Z < \frac{0 - \mu_D}{\sigma_D}\right) \\
&\approx 1 - F_Z\left(\frac{-500}{1414.2}\right) \approx 1 - F_Z(-0.3536) \approx 63.82\%.
\end{aligned}$$

So Abel has 63.82% probability of making more money than Bertha.

Lookup tables usually give $F_Z$ at increments of 0.01. For example $F_Z(-0.35) \approx 0.3632$ and $F_Z(-0.36) \approx 0.3594$. From this we can deduce that $F_Z(-0.3536)$ should be something in between. Using the nearer endpoint $-0.35$ is good enough for this course; then your answer to (d) is $1 - 0.3632 = 63.68\%$.

**Extra information, not required on the course.** If you want, you can get much more precision out of the tables by observing that between $-0.36$ and $-0.35$, $F_Z$ increases by 0.0038 units, when $z$ increases by 0.01 units, so the derivative $F_Z'$ is approximately 0.38. Thus at $z = -0.3536 = -0.36 + 0.0064$ we should have approximately

$$\begin{aligned}
F_Z(z) &\approx F_z(-0.36) + 0.0064 \times 0.38 \\
&\approx 0.3594 + 0.0024 = 0.3618
\end{aligned}$$

which is actually accurate to four decimals.