Recap: Expectation Algebra

Before we can say anything about sums and averages of random variables, we need the rules for how expectation and variance behave under linear combinations.

Fact (Expectation algebra) — For any random variables X and Y, and constants $a, b, c \in \mathbb{R}$:

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

If X and Y are **independent**, then additionally:

$$\text{Var}[aX + bY + c] = a^2 \text{Var}[X] + b^2 \text{Var}[Y]$$

Tip

The expectation result is true *always*; the variance result needs independence. Note also that the variance of $aX - bY$ is $a^2 \text{Var}[X] + b^2 \text{Var}[Y]$ — variances of independent variables always *add*, because $(-b)^2 = b^2$. Subtracting a random variable adds uncertainty; it does not remove it.

Example

X and Y are independent with $\mathbb{E}[X] = 3$, $\text{Var}[X] = 2$, $\mathbb{E}[Y] = 5$, $\text{Var}[Y] = 4$. Find the mean and variance of $2X - Y + 1$.

Linear Combinations of Normal Random Variables

Expectation algebra tells us the mean and variance of $aX + bY$, but not its *distribution*. For normal random variables we get a remarkably clean answer.

Theorem (Linear combinations of normals)

1. If X has a normal distribution then $aX + b$ has a normal distribution.
2. If X and Y have *independent* normal distributions then $aX + bY$ has a normal distribution.

The mean and variance of these are given by expectation algebra. In particular, if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent:

$$aX + bY \sim N\left(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2\right)$$

> **Remark** (Where does this come from?)**.** The slick proof uses *moment generating functions*: the MGF of a $N(\mu, \sigma^2)$ variable is $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, and for independent variables $M_{X+Y}(t) = M_X(t)M_Y(t)$, which is again of normal form. See the Moment Generating Functions notes for the details — it is a lovely two-line argument once the machinery is set up.

Sums are not multiples!

Fact (The classic trap) — Let X_1, X_2 be independent copies of X, where $\mathrm{Var}[X] = \sigma^2$. Then

$$\mathrm{Var}[X_1 + X_2] = \sigma^2 + \sigma^2 = 2\sigma^2 \qquad \text{but} \qquad \mathrm{Var}[2X] = 4\sigma^2$$

"Two different bags of sugar" is $X_1 + X_2$. "Twice the weight of one bag" is $2X$. They have the same *mean*, but different *variances*: in a sum the random errors partially cancel, whereas $2X$ simply doubles whatever error X had.

Example (Carrots)

A one-kilogram bag of carrots has mass (in kg) distributed $X \sim N(1.05, 0.02^2)$, and a three-kilogram bag has mass $Y \sim N(3.10, 0.03^2)$, independently.

(a) Find the distribution of the total mass $T = X_1 + X_2 + X_3$ of three (independent) one-kilogram bags.

(b) Explain why this is not the same as the distribution of $3X$.

(c) Find the probability that three one-kilogram bags together weigh more than one three-kilogram bag.

Example (In class)

A bolt is to fit through a nut. The internal diameter of a nut is distributed $N(10.1, 0.02^2)$ mm and the diameter of a bolt is distributed $N(10.0, 0.03^2)$ mm, independently. A randomly chosen bolt is tried in a randomly chosen nut; it fits if the nut's internal diameter exceeds the bolt's diameter. Find the probability that it fits.

Example (The freight elevator)

A large freight elevator can transport a maximum of 9800 lb. A cargo of 49 boxes needs transporting, and from experience the masses of boxes are normally distributed with mean 205 lb and standard deviation 15 lb. What is the probability that the whole cargo can be transported in one go?

Exercise. Using the carrots model above, find the probability that the mass of one three-kilogram bag is more than three times the mass of a single one-kilogram bag. (You should *not* get 0.862 again — think about which trap this is testing.)

Textbook Exercises: [CUP.S] Ch 8 §2, 4; [S2] Ch 4 §4.6

The Sample Mean of a Normal Population

Definition. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables — think of them as the n observations in a random sample. The sample mean is the random variable

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The sample mean is itself *random*: take a different sample, get a different value of $\overline{X}$. So $\overline{X}$ has its own distribution, called the sampling distribution of the mean.

Definition. The standard deviation of the sample mean, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean. It measures how far a sample mean typically strays from μ.

Theorem (Sample mean of a normal population)
If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then *exactly*:

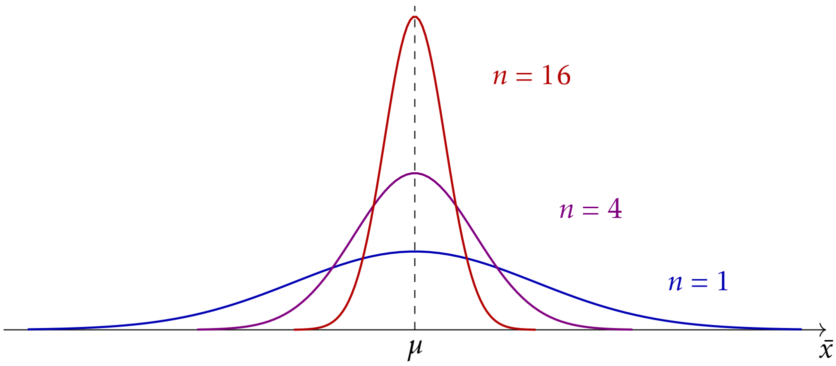
$$\sum_{i=1}^{n} X_i \sim N\left(n\mu,\ n\sigma^2\right) \qquad \text{and} \qquad \overline{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$$

Where do the parameters μ and σ^2/n come from? Derive them using expectation algebra.

> **Remark** (This is *not* the Central Limit Theorem). For a *normal* population, these distributions are **exact** for every sample size n, even $n = 2$. No approximation is involved, and no appeal to any theorem about large samples. The Central Limit Theorem (next section) is about what happens when the population is *not* normal. Examiners like to test whether you know the difference.

Tip
Both forms are worth knowing. Textbooks fixate on $\overline{X}$, but the $\sum X_i \sim N(n\mu, n\sigma^2)$ form is often more natural — as in the freight elevator example, where the constraint was on the *total* mass.

The standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$: averaging n observations shrinks the spread by a factor of $\sqrt{n}$. To halve the spread you need *four* times the data.



The sampling distribution of $\overline{X}$ narrows as n grows: $Var[\overline{X}] = \sigma^2/n$.

Example

IQ scores in a population are modelled as $N(100, 15^2)$. A random sample of 9 people is taken. Find the probability that their mean IQ exceeds 105.

Example (OCR S3, June 2007)

Two brands of car battery, 'Invincible' and 'Excelsior', have lifetimes which are normally distributed. Invincible batteries have a mean lifetime of 5 years with standard deviation 0.7 years. Excelsior batteries have a mean lifetime of 4.5 years with standard deviation 0.5 years. Random samples of 20 Invincible batteries and 25 Excelsior batteries are selected and the sample mean lifetimes are $\overline{X}_I$ years and $\overline{X}_E$ years respectively.

(i) State the distributions of $\overline{X}_I$ and $\overline{X}_E$.

(ii) Calculate $\mathbb{P}\left(\overline{X}_I - \overline{X}_E \geq 1\right)$.

Example (Edexcel S3, June 2013)

A random sample of size n is to be taken from a population that is normally distributed with mean 40 and standard deviation 3. Find the minimum sample size such that the probability of the sample mean being greater than 42 is less than 5%.

Textbook Exercises: [CUP.S] Ch 8 §2; [S3&4] S4 Ch 4 §4.4

The Central Limit Theorem

What if the population is *not* normal — a die roll, a skewed waiting time, anything at all? The expectation algebra still gives $\mathbb{E}[\overline{X}] = \mu$ and $\text{Var}[\overline{X}] = \sigma^2/n$ (no normality was used in that calculation). The astonishing fact is that the *shape* sorts itself out too.

Theorem (The Central Limit Theorem)

For any randomly and independently selected sample $X_1, \ldots, X_n$ of size n taken from *any* population with mean μ and variance σ^2:

1. $\mathbb{E}[\overline{X}] = \mu$,
2. $\text{Var}[\overline{X}] = \dfrac{\sigma^2}{n}$, and
3. $\overline{X}$ is *approximately* normally distributed when n is large (approximately $n > 25$):

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and equivalently} \qquad \sum_{i=1}^{n} X_i \approx N\left(n\mu, n\sigma^2\right)$$

Fact —

- The CLT works *irrespective of the starting distribution* — discrete or continuous, symmetric or skewed. This is what makes it so powerful: we rarely know the true population distribution, but for large samples we don't need to.
- The approximation is better for symmetric parent distributions, and needs larger n for heavily skewed ones. The OCR guideline $n > 25$ is a rule of thumb, not a law of nature.
- If the population happens to be normal, the result is exact for all n — that is the previous section, and it is *not* called the CLT.

Tip

Play with a sampling simulator (e.g. the department's spreadsheet, or an online "sample mean simulator"): start from a wildly non-normal parent — uniform, U-shaped, exponential — and watch the histogram of sample means become bell-shaped as n grows. Seeing this happen is the fastest way to believe the theorem.

Example

A fair six-sided die has $\mu = 3.5$ and $\sigma^2 = \frac{35}{12}$. The die is rolled 30 times. Estimate the probability that the mean score exceeds 4.

Remark (Continuity correction)**.** The die is discrete, so strictly we are approximating a discrete distribution by a continuous one and should apply a continuity correction: $\overline{X} > 4$ means the total exceeds 120, i.e. $\mathbb{P}(\sum X_i \geq 120.5)$ after correcting, giving 0.0606 rather than 0.0544. The correction on $\overline{X}$ itself is $\frac{1}{2n}$, which is small for large n — and it is usually ignored in practice. Be aware it exists; don't lose sleep over it.

Example

The lifetime of a certain battery has mean 120 hours and standard deviation 25 hours; the distribution of lifetimes is unknown and suspected to be skewed. A random sample of 50 batteries is tested. Estimate the probability that the sample mean lifetime is less than 115 hours.

Example (Edexcel S3, June 2006)

A report on the health and nutrition of a population stated that the mean height of three-year-old children is 90 cm and the standard deviation is 5 cm. A sample of 100 three-year-old children was chosen from the population.

(a) Write down the approximate distribution of the sample mean height. Give a reason for your answer.

(b) Hence find the probability that the sample mean height is at least 91 cm.

Example (In class)

A random sample of 36 observations is taken from a population (of unknown distribution) with mean 80 and variance 36. Estimate the probability that the sample mean lies between 79 and 81.

Example (OCR S2, June 2013)

The mean of a sample of 80 independent observations of a continuous random variable Y is denoted by $\overline{Y}$. It is given that $\mathbb{P}(\overline{Y} \leq 157.18) = 0.1$ and $\mathbb{P}(\overline{Y} \geq 164.76) = 0.7$.

(i) Calculate $\mathbb{E}[Y]$ and the standard deviation of Y.

(ii) State

(a) where in your calculations you have used the Central Limit Theorem,

(b) why it was necessary to use the Central Limit Theorem,

(c) why it was possible to use the Central Limit Theorem.

Remark (How hard is the proof?)**.** Proving the CLT is genuinely hard — far beyond A Level. One route uses the *cumulant generating function* $K_X(t) = \ln M_X(t)$: standardising the sum and letting $n \to \infty$, one shows all the cumulants κ_r for $r \geq 3$ die away, leaving exactly the cumulants of a normal distribution. The proofs of $\mathbb{E}[\overline{X}] = \mu$ and $\text{Var}[\overline{X}] = \sigma^2/n$, on the other hand, are the easy expectation algebra above.

Textbook Exercises: [CUP.S] Ch 8 §2, 4; [S2] Ch 4; [S3&4] S4 Ch 4 §4.4