Engineering Mathematics 2B

Module 16: Introduction to Probability

Nick Polydorides

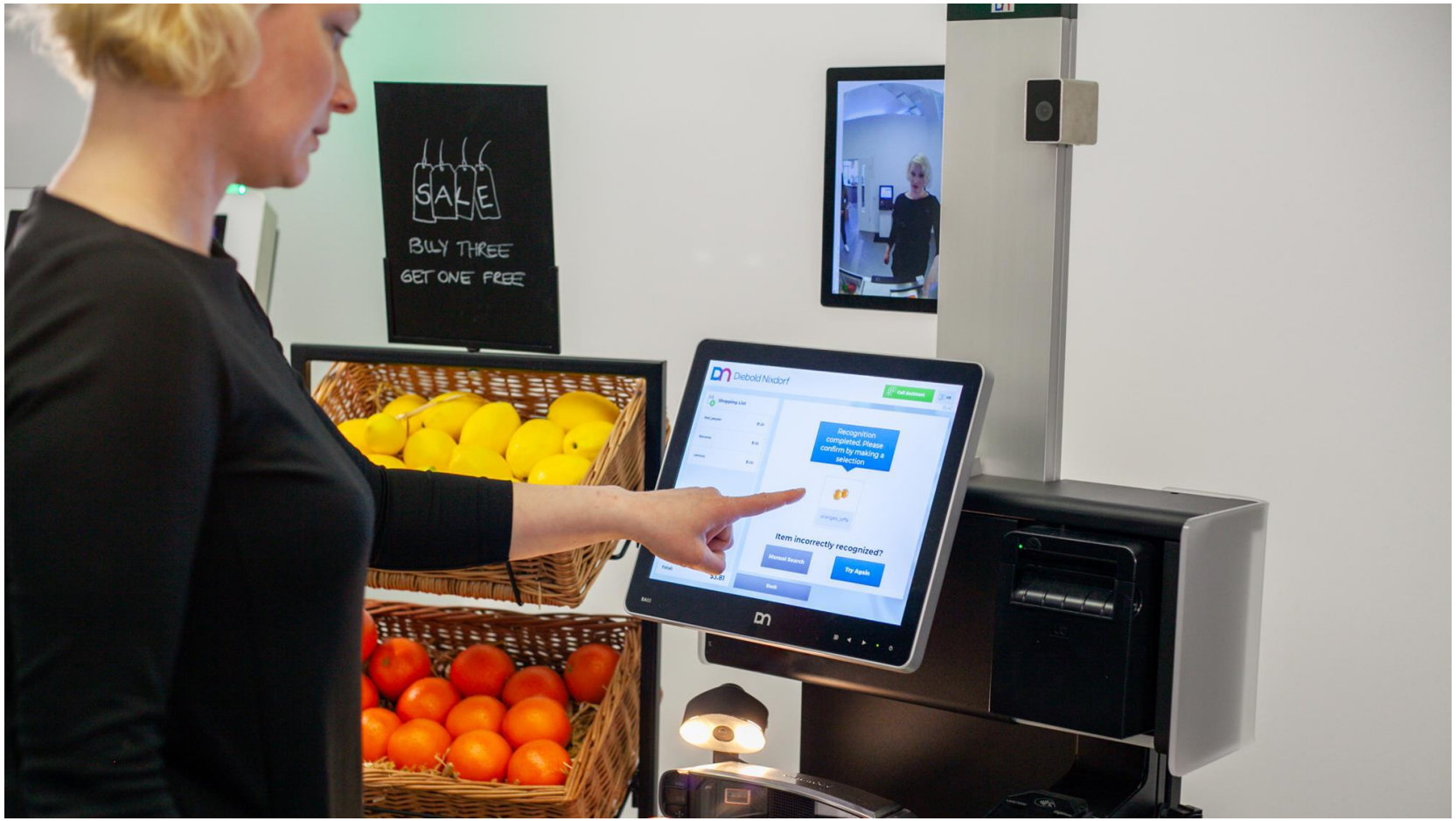
School of Engineering

THE UNIVERSITY *of* EDINBURGH

Module 16 contents

Motivation: Self-checkout Artificial Intelligence

What orange size is considered "normal"?



How does the machine know that you are buying 3 oranges?

Normality preserving transformations

Linear operations (addition, subtraction, scaling) between two independent normal variables yield a normal variable.

Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, then for a and b nonzero real values let

$$Y = aX + b.$$

Then Y is also normal, i.e. $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, with

$$\mu_Y = \mathbb{E}[Y] = a\mu_X + b, \quad \sigma_Y^2 = \mathrm{Var}(Y) = a^2\sigma_X^2.$$

Let $W \sim \mathcal{N}(\mu_W, \sigma_W^2)$ be independent of X, then

$$X \pm aW = \mathcal{N}(\mu_X \pm a\mu_W, \sigma_X^2 + a^2\sigma_W^2)$$

Sums of random variables

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ independent then

$$X + Y = Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2),$$

(linear operations preserve normality)

If $X \sim p_X(x)$, $Y \sim p_Y(y)$ independent then $X + Y = Z \sim p_Z(z)$ where

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x) p_Y(z - x) \mathrm{d}x$$

(convolution - non examinable in EM2B)

If we don't know $p_X(x)$, $p_Y(y)$ but only their means μ_X, μ_Y and variances σ_X^2 and σ_Y^2 what can we say about Z?

Sums of random variables

Without proof, for $Z = X + Y$ (not restricted to just normal variables)

1. $\mu_Z = \mu_X + \mu_Y$,
2. $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\text{Cov}(X, Y)$ (in general)
3. $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$ (when X, Y independent)

Note however, that for $Z = X - Y$

1. $\mu_Z = \mu_X - \mu_Y$
2. $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2\text{Cov}(X, Y)$ (in general)

Sum of n iid random variables

Suppose we have n iid random variables from an unknown PDF with unknown mean μ and variance σ^2.

Consider the random variable

$$\text{“sum of } n \text{ iid”} : \text{Sum}_n = X_1 + X_2 + \ldots + X_n$$

what can be said about Sum_n as n grows?

As $\{X_1, \ldots, X_n\}$ are iid (independent with *same* mean and variance)

$$\mathbb{E}[\text{Sum}_n] = n\mu, \quad \text{Var}(\text{Sum}_n) = n\sigma^2$$

As n increases the variance of Sum_n increases. (very loose estimator of the sum)

Sample mean of n iid random variables

Instead of $\text{Sum}_n = \sum_{i=1}^n X_i$ consider the random variable sample mean of n iid

$$\bar{X}_n = \frac{1}{n}\text{Sum}_n$$

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\mathbb{E}[\text{Sum}_n] = \mu, \quad \text{Var}(\bar{X}_n) = \frac{1}{n^2}\text{Var}(\text{Sum}_n) = \frac{\sigma^2}{n}$$

Notice, that the variance of the sample mean reduces to zero as n increases. (sharp estimator of the mean)

Remark: For iid variables of mean μ and variance σ^2, as $n \to \infty$

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \text{Var}(\bar{X}_n) \to 0$$

indicating that the **sample mean converges to the population mean** μ with perfect precision.

The power of more data

Self-checkout tills typically operate with

$$\mathbb{P}(\bar{X}_n \geq \text{weight/number of items})$$

as it is more precise, though practically the number of items tends to be small.

We can predict the average weight of oranges with greater accuracy by computing the sample mean of 300 oranges rather than the sample mean of 3 oranges.

This is the reason that data-driven, e.g., deep learning, models/networks enhance their performance when fed with larger training data sets.

The mean is not everything!

For an increasing number n of iid $X_1, \ldots, X_n$, where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, the **sample mean** $\bar{X}_n$ converges to the mean μ of the unknown $p_{X_i}(x_i)$ with zero variance.

Valuable as this information is, it does not allow to compute probabilities like

$$\mathbb{P}(\bar{X}_n \geq \mu + \frac{a}{\sqrt{n}}) = ? \quad \text{or} \quad \mathbb{P}(\mathrm{Sum}_n \leq n\mu + b\sqrt{n}) = ?$$

for some constants a and b, such as the mean-deviation terms $\frac{a}{\sqrt{n}}$ and $b\sqrt{n}$ aren't as big as μ or $n\mu$ respectively.

For these probabilities we need the probability density functions of the random variables Sum_n and $\bar{X}_n$.

Standardising the Sum_n variable

As we have seen with normal random variables, for an arbitrary type random variable, subtracting its mean and dividing with its standard deviation yields a standard random variable (not necessarily normal). For Sum_n we have

$$Z_n = \frac{\text{Sum}_n - \mathbb{E}[\text{Sum}_n]}{\sqrt{\text{Var}(\text{Sum}_n)}} = \frac{\text{Sum}_n - n\mu}{\sqrt{n}\sigma}.$$

We can easily show

$$\mathbb{E}[Z_n] = \mathbb{E}\Big[\frac{\text{Sum}_n - n\mu}{\sqrt{n}\sigma}\Big] = \frac{1}{\sqrt{n}\sigma}\Big(\mathbb{E}[\text{Sum}_n] - n\mu\Big) = 0,$$

$$\begin{aligned}\text{Var}(Z_n) &= \text{Var}\Big(\frac{\text{Sum}_n}{\sqrt{n}\sigma}\Big) + \text{Var}\Big(\frac{n\mu}{\sqrt{n}\sigma}\Big) \\ &= \frac{1}{n^2\sigma}\text{Var}(\text{Sum}_n) + 0 = 1\end{aligned}$$

Central Limit Theorem

The **Central Limit Theorem** (CLT) asserts that for **large** n **samples** of iid random variables, Z_n becomes **normally distributed** with mean 0 and variance 1, irrespective of what the unknown distribution of the $X_1, \ldots, X_n$ samples is.

In these circumstances Z_n is *approximately* a standard normal. This allows to compute probability integrals not too far away from the mean via the CDF (or its table)

$$F_{Z_n}(z) \approx \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t$$

In corollary,

$$\frac{1}{n}\mathrm{Sum}_n = \bar{X}_n \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2).$$

Normal approximation based on CLT

Let $\text{Sum}_n = X_1 + X_2 + \ldots + X_n$ the sum of n iid with mean μ and variance σ^2. If n is **large**, find the probability

$$\mathbb{P}(\text{Sum}_n \leq c) = ?$$

This requires integrating over the PDF of Sum_n which we don't know, hence we do:

1. Assume the PDF of Sum_n is normal by virtue of the CLT.
2. Using $\mathbb{E}[\text{Sum}_n] = n\mu$ and $\text{Var}(\text{Sum}_n) = n\sigma^2$ compute the integral limit that corresponds to c as

$$z = \frac{c - n\mu}{\sqrt{n}\sigma}$$

3. Use

$$\mathbb{P}(\text{Sum}_n \leq c) = \mathbb{P}(Z_n \leq z) \approx \Phi(z)$$

4. The CLT approximation is accurate enough for $|z| \leq 2$

Example

We load an aircraft with 100 packages whose weights are independent random variables uniformly distributed between 5 and 50 lb.

Find the probability that the total weight will exceed 3000 lb, i.e.

$$\mathbb{P}(\text{Sum}_{100} > 3000) = ?$$

and then verify that this is the same as the probability of the 100-package average weight to be more than 30, i.e.

$$\mathbb{P}(\text{Sum}_{100} > 3000) = \mathbb{P}(\bar{X}_{100} > 30)$$

Example

- We want to calculate

$$\mathbb{P}(\text{Sum}_{100} > 3000) = 1 - \mathbb{P}(\text{Sum}_{100} \leq 3000),$$

where Sum_{100} is the random variable: sum of the weights of 100 packages.

- As we know the PDF of the individual weights $\mathcal{U}[5, 50]$ but not that of their sum, since $n = 100$ is large we can invoke the CLT.

- The mean and the variance of the weight of a single package can be deduced from the uniform distribution as

$$\mu = \frac{5 + 50}{2} = 27.5 \text{ lb}, \qquad \sigma^2 = \frac{(50 - 5)^2}{12} = 168.75 \text{ lb}^2$$

Example cont

- Standardising we get

$$z = \frac{c - n\mu}{\sqrt{n}\sigma} = \frac{3000 - 100 \cdot 27.5}{10 \cdot \sqrt{168.75}} = 1.92$$

- From the standard normal CDF we get

$$\mathbb{P}(Z_{100} \leq 1.92) \approx \Phi(1.92) = 0.9726 = \mathbb{P}(\text{Sum}_{100} < 3000).$$

- $\mathbb{P}(\text{Sum}_{100} > 3000) = 1 - \mathbb{P}(\text{Sum}_{100} \leq 3000) = 1 - 0.9726 = 0.0274 \approx 3\%$

Finding $\Phi(1.92)$ from the table

z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.0
0.0	0.50000	0.50399		0.51197	0.515
0.1	0.53983	0.54380	0.54776	0.55172	0.555
0.2	0.57926	0.58317	0.58706	0.59095	0.594
0.3	0.61791	0.62172	0.62552	0.62930	0.633
0.4	0.65542	0.65910	0.66276	0.66640	0.670
0.5	0.69146	0.69497	0.69847	0.70194	0.705
0.6	0.72575	0.72907	0.73237	0.73565	0.738
0.7	0.75804	0.76115	0.76424	0.76730	0.770
0.8	0.78814	0.79103	0.79389	0.79673	0.799
0.9	0.81594	0.81859	0.82121	0.82381	0.826
1.0	0.84134	0.84375	0.84614	0.84849	0.850
1.1	0.86433	0.86650	0.86864	0.87076	0.872
1.2	0.88493	0.88686	0.88877	0.89065	0.892
1.3	0.90320	0.90490	0.90658	0.90824	0.909
1.4	0.91924	0.92073	0.92220	0.92364	0.925
1.5	0.93319	0.93448	0.93574	0.93699	0.938
1.6	0.94520	0.94630	0.94738	0.94845	0.949
1.7	0.95543	0.95637		0.95818	0.959
	96407	0.964		96638	0.967
1.9	7128	0.971	0.97257	7320	0.973
	97725	0.977		97882	0.979
2.1	0.98214	0.98257		0.98341	0.983

Example cont.

Since 'average of n = sum of n divided by n', it is not difficult to see that

$$\mathbb{P}(\text{Sum}_{100} > 3000) = \mathbb{P}(\bar{X}_{100} > 30)$$

Recall that by virtue of the CLT,

$$\frac{\text{Sum}_n - n\mu}{\sigma\sqrt{n}} = Z_n \sim \mathcal{N}(0,1), \quad \text{and} \quad \bar{X}_n \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$$

then by converting the normal $\bar{X}_n \sim \mathcal{N}(\mu, (\sigma/\sqrt{n})^2)$ into a standard normal $\mathcal{N}(0,1)$ with

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{30 - 27.5}{\sqrt{168.75}/10} = 1.92,$$

hence $\mathbb{P}(\bar{X}_{100} > 30) = 1 - \mathbb{P}(\bar{X}_{100} \leq 30) = 1 - \mathbb{P}(Z \leq 1.92) = 0.0274$.

Example part 2

Further compute the probability that 150 randomly selected packages will weight between 4100 and 4500 lb?

$$\mathbb{P}(4100 \leq \text{Sum}_{150} \leq 4500) =$$

$$= \mathbb{P}\Big(\frac{4100 - 150 \cdot 27.5}{\sqrt{168.75}\sqrt{150}} \leq Z_{150} \leq \frac{4500 - 150 \cdot 27.5}{\sqrt{168.75}\sqrt{150}}\Big)$$

$$= \mathbb{P}(-0.1571 \leq Z_{150} \leq 2.357)$$

$$= \Phi(2.357) - \Phi(-0.1571)$$
$$= \Phi(2.357) - \Big(1 - \Phi(0.1571)\Big)$$
$$\approx 0.9908 - 0.4364 = 0.5544$$

Finding $\Phi(-0.1571)$ and $\Phi(0.1571)$ from the table.

−0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
−0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
−0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
−0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
z	**−0.00**	**−0.01**	**−0.02**	**−0.03**	**−0.04**	**−0.05**	**−0.06**	**−0.07**	**−0.08**	**−0.09**

z	**+ 0.00**	**+ 0.01**	**+ 0.02**	**+ 0.03**	**+ 0.04**	**+ 0.05**	**+ 0.06**	**+ 0.07**	**+ 0.08**	**+ 0.09**
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793

Notice the table has $\Phi(\pm 0.15)$ and $\Phi(\pm 0.16)$ so I opted for the second as a more accurate approximation of $\Phi(\pm 0.1571)$

Formulas

Let $\text{Sum}_n = X_1 + X_2 + \ldots + X_n$ the sum of a large n iid variables with mean μ and variance σ^2. Further let $\bar{X}_n = \frac{1}{n}\text{Sum}_n$

- $\bar{X}_n \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$
- For a c such that $z = \frac{c - n\mu}{\sqrt{n}\sigma}$ is small, then by virtue of the CLT

$$\mathbb{P}(\text{Sum}_n \leq c) = \mathbb{P}(Z_n \leq z) \approx \Phi(z)$$

Main outcomes of module 16

You MUST know:

1. The properties of sums and averages of n iid random variables.
2. The central limit theorem says that the distribution of the average of a large number of idd samples converges to the normal/Gaussian.
3. To compute probabilities of sample means and sums involving large numbers of iid variables.