Statistical Methods

Lecture 6 – Normal Random Variables

Luke Dickens

Autumn 2020

Continuous RVs

A continuous random variable (RV) is one whose set of possible values is an interval. Put another way, it can take any value in some interval.

Examples:

- The time taken to complete a scientific experiment
- The weight of a person at some point in time

Discrete RVs with a large sample space can sometimes be approximated by continuous RVs. E.g. The number of:

- decays from a radioactive sample in a a given time
- people voting for a given candidate in an election
- cars passing a busy intersection in an hour

Probability Density Functions

A continuous RV, X, with $S \subseteq \mathbb{R}$, has an associated **probability density function**, say p_X, where

- The probability X takes value between a and b

$$\Pr(a \leq X \leq b) = \text{area under curve } p_X \text{ between a and b}$$

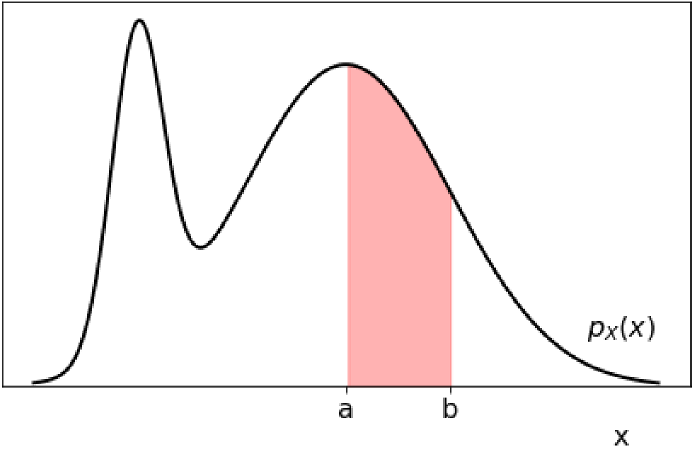
- We restrict to well-behaved functions so

$$\Pr(a \leq X \leq b) = \Pr(a < X < b) = \int_a^b p_X(x)dx$$

Probability that X takes a precise value is vanishingly small

- **Constraints:** $p_X(x) > 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p_X(x)dx = 1$

Probability density functions illustrated



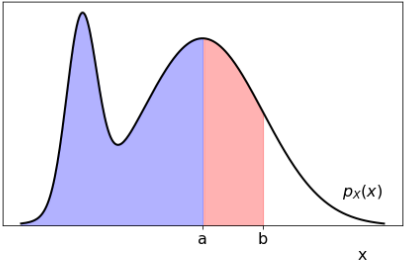
$\Pr(a \leq X \leq b) =$ area of red shaded region

Cummulative Distribution Functions

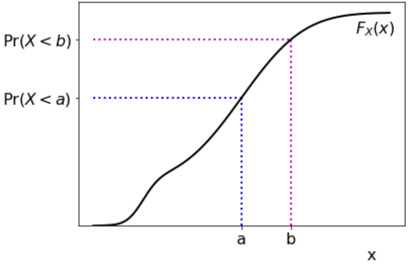
A continuous RV, X, is also associated with a cummulative distribution function, F_X, where

- $F_X(z)$ gives probability that X takes value less than z

$$F_X(z) = \Pr(X < z) = \int_{-\infty}^{z} p_X(x)dx$$



Probability density function, p_X



Cummulative distribution function, F_X

Normal RVs

Normal RVs, X, have the following properties:

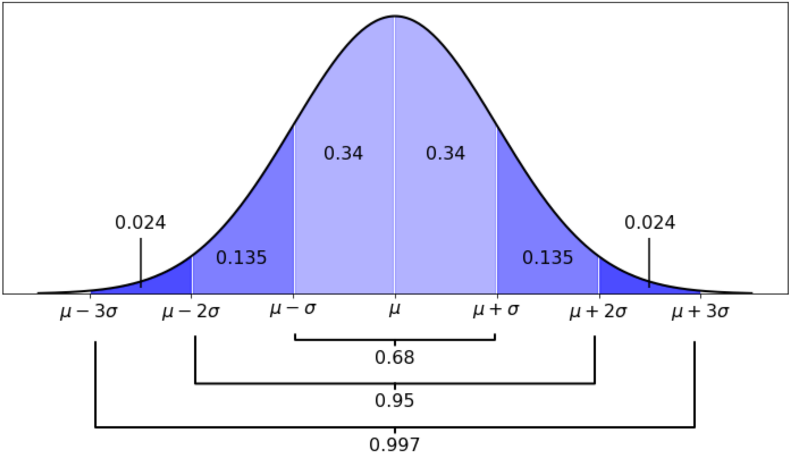
- determined by two parameters

$$\mu = \mathbb{E}[X] \quad \text{and} \quad \sigma = SD(X)$$

- **symmetric:** $\Pr(X < \mu) = \Pr(X > \mu) = \frac{1}{2}$
- **bell shaped** probability density function defined by

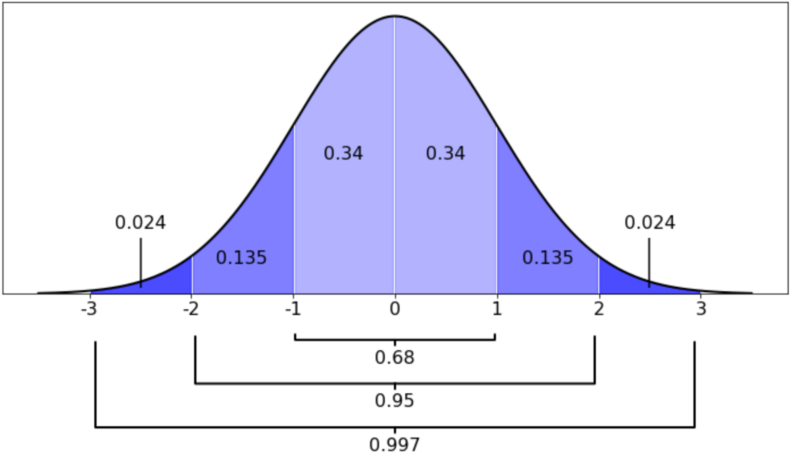
$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 / 2\sigma^2}$$

The Normal Distribution



These proportions are the basis of the empirical rule from lecture 3.

The Standard Normal Distribution



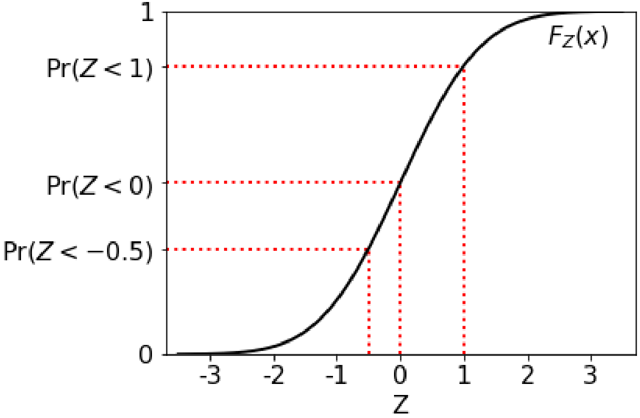
A standard normal RV, Z, is a normal RV with $\mu = 0$ and $\sigma = 1$.

The textbook [Ros17, Tbl. 6.1] shows probabilities of the form $\Pr(Z < x)$ for various values of x.

x	**0.00**	**0.01**	**0.02**	**...**
0.0	0.5000	0.5040	0.5080	...
0.1	0.5398	0.5438	0.5478	...
0.2	0.5793	0.5832	0.5871	...
0.3	0.6179	0.6217	0.6255	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

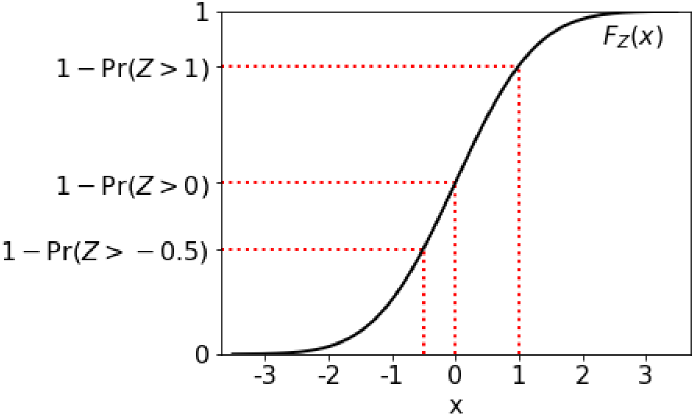
Look these up for $x = a.bc$ by looking in row **a.b** and column **0.0c**

Standard normal cummulative probabilities



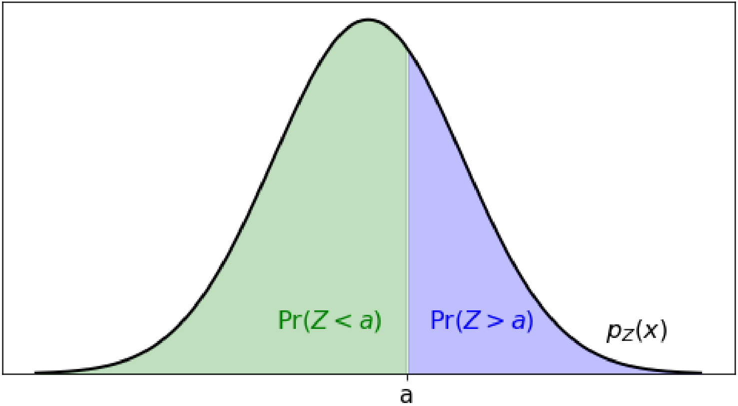
Values in the table correspond to the cummulative distribution function F_Z for the standard normal. **Recall:** $F_Z(x) = \Pr(Z < x)$

Standard normal cummulative probabilities



Can calculate complementary probabilities using formula
$\Pr(Z < x) = 1 - \Pr(Z \geq x) = 1 - \Pr(Z > x)$

Alternative view



The total area (green + blue) is 1.

Hence: $\Pr(Z < a) + \Pr(Z > a) = 1$

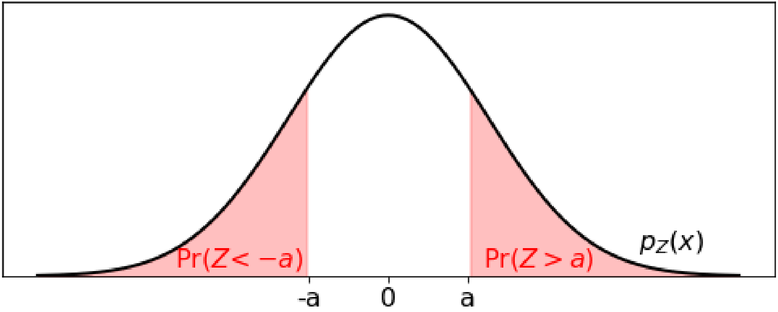
And so: $\Pr(Z > a) = 1 - \Pr(Z < a)$

Negative values and symmetry

Cummulative probabilities in [Ros17, Tbl. 6.1] are only for positive values. We can derive results for negative values.

For some $a > 0$ (a positive value), from symmetry:

$$\Pr(Z < -a) = \Pr(Z > a)$$



Converting normal to standard normal

Consider a Normal RV, X, with mean $\mu = \mathbb{E}[X]$ and standard deviation $\sigma = \text{SD}(X)$. Define new RV:

$$Z = \frac{X - \mu}{\sigma}$$

- $\mathbb{E}[Z] = \mathbb{E}[X] - \mu = 0$
- $\text{SD}(Z) = \frac{\text{SD}(X)}{\sigma} = 1$
- Z is also normal (symmetric, bell-shaped, etc)

Therefore, Z is a standard normal RV.

Cummulative probabilities of Normal RV

Consider again, the Normal RV, X, with mean μ and SD σ. We want to know $\Pr(X < a)$, and so reason as follows:

- $X < a$ equivalent to $\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma}$
- $\frac{X-\mu}{\sigma}$ equivalent to standard normal RV Z
- And so:

$$\Pr(X < a) = \Pr\left(Z < \frac{a-\mu}{\sigma}\right)$$

Example adapted from [Ros17, p. 282, q. 3]

A hair-dryer manufacturer knows that the length of time their hair-dryers function before breaking is normally distributed with mean 40 months and SD 8 months. The guarantee is for 3 years. (a) What proportion of dryers fail the guarantee? (b) What proportion of dryers fail between 24 and 56 months?

(a) 3 years is 36 months, if our RV is X, $\mu = 40$ and $\sigma = 8$ then

$$\Pr(X < 36) = \Pr\left(Z < \frac{36-40}{8}\right) = \Pr(Z < -0.5)$$
$$= \Pr(Z > 0.5) = 1 - \Pr(Z < 0.5) = 0.3085$$

(b) Standardising the two values $\frac{24-\mu}{\sigma} = -2$ and $\frac{56-\mu}{\sigma} = 2$, so

$$\Pr(24 < X < 56) = \Pr(-2 < Z < 2)$$

which by the approximation rule is 0.95

Adding two Normal RVs

Important property: The sum of two normal RVs is a normal RV

More precisely:
Consider two independent normal RVs X and Y with respective means μ_x and μ_y and respective SDs σ_x and σ_y.

The RV given by $X + Y$:

- is also normal
- has mean $\mathbb{E}[X + Y] = \mu_x + \mu_y$
- and has standard deviation $SD(X + Y) = \sqrt{\sigma_x^2 + \sigma_y^2}$

Note: A similar rule applies for subtraction, e.g. $X - Y$ which is useful to predict whether $X < Y$ will hold (or vice versa).

Weekly customer demand for a milk delivery firm is roughly normally distributed with mean 1000 litres and SD 100. Assuming demand each week is independent (a) What is the probability that demand this week and next is each below 1100l? (b) What is the probability that total demand over the two weeks is below 2200l?

(a) If X_i is demand in week i then

$$\Pr(X_1 < 1100) = \Pr(X_2 < 1100) = \Pr(Z < 1) = 0.8413$$
$$\Pr((X_1 < 1100) \text{ and } (X_2 < 1100)) = 0.8413^2 = 0.7071$$

(b) RV $X_1 + X_2$ has mean 2000 & SD $\sqrt{2 \cdot 100^2} \approx 141$

$$\Pr(X_1 + X_2 < 2200) = \Pr\left(Z < \frac{2200 - 2000}{141}\right)$$
$$= \Pr(Z < 1.41) = 0.9207 \text{ (from table)}$$

Example: Two students

A university course in General Relativity takes students from Physics and Maths. Physics and maths students' marks are both approximately normally distributed with respective means 63.5 and 66.2 and SDs 10 and 8. Marks below 50 are failing. One physics and one maths student are chosen randomly. Assuming the marks are continuous and independent, with what probability:

(a) will each student fail the module?

(b) will the physics student score more than the maths student?

Answers on next slide.

Is it reasonable to assume normality and independence?

Example: Two students - solution

(a) Probability each student fails the module: Model scores by RV P and M respectively.

$$\Pr(P < 50) = \Pr\left(Z < \frac{50 - 63.5}{10}\right) = \Pr(Z < -1.35)$$
$$= 1 - \Pr(Z < 1.35) = 0.0885$$
$$\Pr(M < 50) = \Pr\left(Z < \frac{50 - 66.2}{8}\right) = 0.0214$$

(b) Probability Physics student scores more. Create new RV $P - M$, with mean -2.7 and SD $\sqrt{10^2 + 8^2} = 12.8$. Ask:

$$\Pr(P - M > 0) = \Pr\left(Z > \frac{0 - (-2.7)}{12.8}\right) = \Pr(Z > 0.21)$$
$$= 1 - \Pr(Z < 0.21) = 0.4165$$

Percentiles of Standard Normal RV

For any $0 < \alpha < 1$, z_α is defined as the value where

$$\Pr(Z > z_\alpha) = \alpha$$

Equally: $\Pr(Z < z_\alpha) = 1 - \alpha$

z_α is the $100(1-\alpha)$ percentile of the standard normal distribution.

From zero mean and symmetry: $\Pr(Z < -z_\alpha) = \alpha$

Percentiles of Normal RVs

Given a normal RV, X, with mean μ and SD σ, we can determine percentiles by inverse of standardisation.

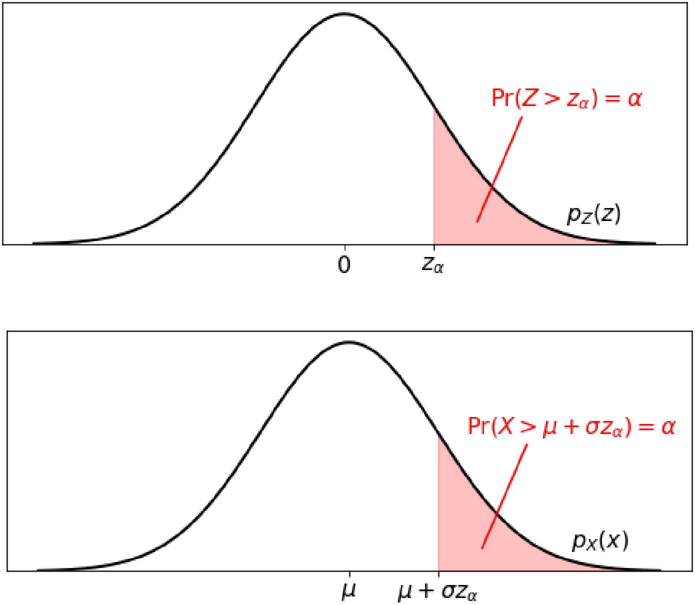
Standardisation: $Z = \dfrac{X - \mu}{\sigma}$

Inverse standardisation: $X = \mu + \sigma Z$

So $100(1-\alpha)$ percentile of X is:

$$x_\alpha = \mu + \sigma z_\alpha$$

Illustration of $100(1-\alpha)$ percentiles



Example: Men's height in the UK

In the UK, Men's heights are roughly normally distributed with mean 175.3*cm* and standard deviation 7.5*cm*, what height are:

- 10% of men taller than (and 90% shorter than):

$$x_{0.9} = 173.5 + 7.5z_{0.9} = 183.1cm$$

  using approximation $z_{0.9} \approx 1.28 + \frac{0.9-0.8997}{0.9015-0.8997}0.01$

  (Finding values either side of 0.9 *in [Ros17, Tbl. 6.1].)*

- 5% of men shorter than (and 95% taller than):

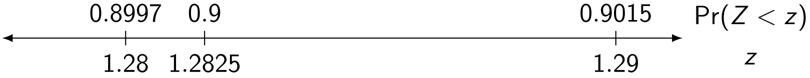
$$x_{0.05} = 173.5 + 7.5z_{0.05} = 161.2cm$$

  using approximation $z_{0.05} \approx -(1.64 + \frac{0.95-0.9495}{0.9505-0.9495}0.01)$

  (We calculate $z_{0.95}$ by finding values either side of 0.95 *in [Ros17, Tbl. 6.1] then use $z_{1-\alpha} = -z_\alpha$.)*

Why the fractions in the previous slide?

We are assuming for small changes in probability mass, there is a roughly linear relationship with changes in the *z* value:



Where we are estimating:

$$z_{0.9} \approx 1.28 + \frac{0.9 - 0.8997}{0.9015 - 0.8997} 0.01 = 1.2825$$

Although, for large changes the relationship is not linear.

References I

[Ros17] Sheldon M. Ross, *Introductory Statistics*, 4 ed., Academic Press, 2017.