Lecture 5.1

Common Probability Distributions: Normal Distribution

Normal Distribution

A random variable X is said to have a **normal** or **Gaussian** distribution with parameters μ (expectation) and σ^2 (variance) if its pdf is given by

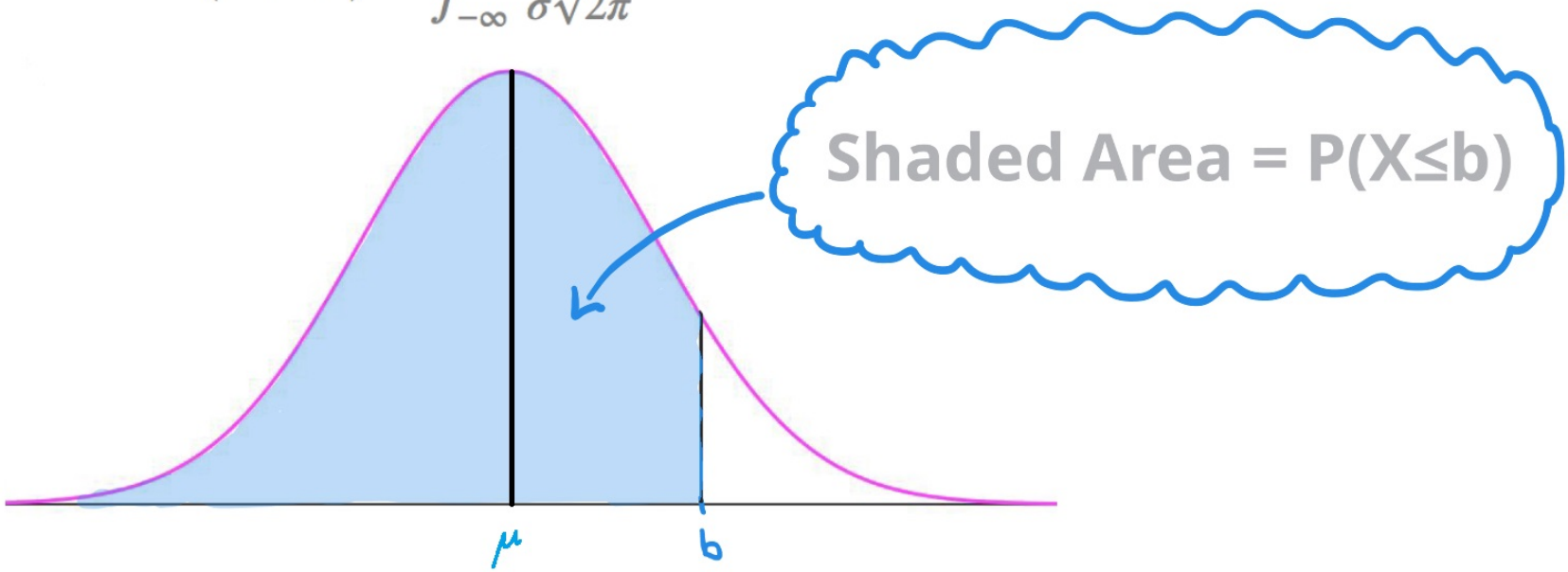
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Note that the pdf is symmetric about μ. As σ increases, the pdf becomes flatter and more spread out.

Normal Distribution

$$P(X \leq b) = \int_{-\infty}^{b} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$



Normal Distribution

The normal distribution has two important roles in data analysis.

- Modelling: Data arising in many contexts can be well represented by the normal distribution, e.g., heights, weights, and other physical characteristics[1], measurement errors in scientific experiments, reaction times in psychological experiments, measurements of intelligence and aptitude, scores on various tests, and numerous economic measures and indicators. In addition, even when individual variables themselves are not normally distributed, sums and averages of the variables will under suitable conditions have approximately a normal distribution.
- Inference: When constructing confidence intervals and hypothesis tests, we need to know the distribution of the estimator. The distribution of many estimators is (at least approximately) normal.

[1] The famous 1903 Biometrika article "On the Laws of Inheritance in Man" discussed many examples of this sort.

Linear transformations

Suppose $a > 0$ and $b \in \mathbb{R}$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $Y = aX + b$ has a $\mathcal{N}(a\mu + b, a^2\sigma^2)$ distribution.

As Y is a linear transformation of X:

$$F_Y(y) = \mathbb{P}(aX + b \leq y) = \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Differentiating the cdf of Y:

$$\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{a} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\frac{(y-b)}{a} - \mu}{\sigma}\right)^2}$$

$$= \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y - a\mu - b}{a\sigma}\right)^2},$$

which is the pdf of the $\mathcal{N}(a\mu + b, a^2\sigma^2)$ distribution.

Standard normal distribution

If $\mu = 0$ and $\sigma = 1$ then the distribution is known as the **standard normal** distribution.

If Z has standard normal distribution, then $X = \mu + \sigma Z$ has a $\mathcal{N}(\mu, \sigma^2)$ distribution. Conversely, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the **standardised** random variable

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

has a standard normal distribution.

This can be used to derive the mean, variance and MGF of the normal distribution from those of the standard normal distribution.

Properties of the Normal Distribution

If $Z \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}(Z) = 0, \quad \text{Var}(Z) = 1, \quad \text{and} \quad M_Z(s) = \exp(s^2/2).$$

Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Then,

1. $\mathbb{E}X = \mu$
2. $\text{Var}(X) = \sigma^2$
3. $M(s) = \exp(\mu s + \frac{\sigma^2}{2} s^2) \quad s \in \mathbb{R}$

$$M_X(s) = \mathbb{E}(e^{sX}) = \mathbb{E}(e^{s(\mu+\sigma Z)}) = e^{s\mu}\mathbb{E}(e^{s\sigma Z}) = e^{s\mu} M_Z(s\sigma).$$

Normal Probabilities

Suppose the heights of students have a Normal distribution with mean 172 cm and standard deviation 9.8 cm. What is the probability a student is shorter than 180cm?

$$\mathbb{P}(X \leq 180) = \mathbb{P}\left(\frac{X - 172}{9.8} \leq \frac{180 - 172}{9.8}\right) = \mathbb{P}(Z \leq 0.816)$$

The cdf of the normal distribution does not have a simple expression but can be evaluated numerically in `R` using `pnorm`.

```
> pnorm((180-172)/9.8)
[1] 0.7928433
```

The probability a student is taller than 180cm is $1 - 0.793 = 0.207$.

Normal Probabilities

Suppose the heights of students have a Normal distribution with mean 172 cm and standard deviation 9.8 cm. What is the probability a student is between 170cm and 180cm?

Let $X \sim \mathcal{N}(172, 9.8^2)$. Then

$$\begin{aligned}\mathbb{P}(170 < X < 180) &= \mathbb{P}\left(\frac{170-172}{9.8} < Z < \frac{180-172}{9.8}\right) \\ &= \mathbb{P}\left(-0.204 < Z < 0.816\right) \\ &= \mathbb{P}(Z \leq 0.816) - \mathbb{P}(Z \leq -0.204)\end{aligned}$$

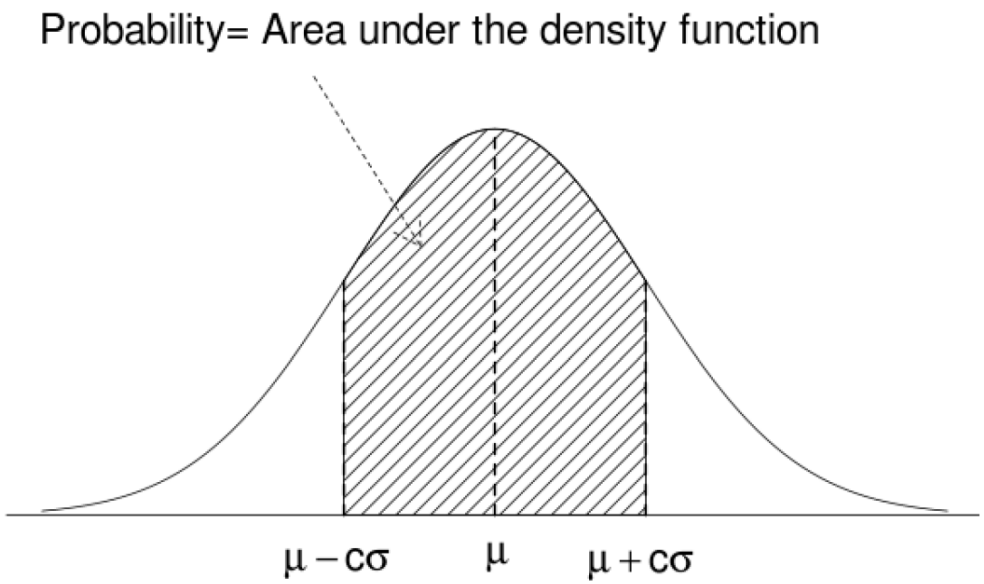
```
> pnorm((180-172)/9.8) - pnorm((170-172)/9.8)
[1] 0.3736984
```

Normal Probabilities

The `pnorm` function evaluates the cdf of any normal distribution not just the standard normal. For example

```
> pnorm(180, mean=172, sd=9.8)
[1] 0.7928433
```

68–95–99.7 Rule



There is no simple formula for finding areas under the normal pdf. However, as a rough empirical rule for $X \sim \mathcal{N}(\mu, \sigma^2)$:

- area within $c = 1$ standard deviation of the mean is $\approx 68\%$
- area within $c = 2$ standard deviations of the mean is $\approx 95\%$
- area within $c = 3$ standard deviations of the mean is $\approx 99.7\%$

Normal Quantiles

Recall that the p-quantile of a continuous random variable with cdf F is the smallest value x such that $F(x) = p$. For normal distribution, since the cdf is strictly increasing, there is a unique x such that $F(x) = p$.

Suppose z_p is the p-quantile of the standard normal, i.e., $F_Z(z_p) = p$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$p = F_Z(z_p) = \mathbb{P}(Z \leq z_p) = \mathbb{P}(\mu + \sigma Z \leq \mu + \sigma z_p) = \mathbb{P}(X \leq \mu + z_p\sigma).$$

So $\mu + \sigma z_p$ is the p-quantile of the $\mathcal{N}(\mu, \sigma^2)$ distribution.

So, in particular we have

$$Q1 = \mu + \sigma Q1_z, \quad Q3 = \mu + \sigma Q3_z, \quad \text{and} \quad IQR = \sigma IQR_z.$$

$1.5 \times$ IQR Rule

Assuming our data is from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, how often do we flag observations as unusual using the $1.5 \times$ IQR rule for box plots? We would flag any observation x_i such that

$$x_i \leq Q1 - 1.5 \times IQR, \quad \text{or} \quad x_i \geq Q3 + 1.5 \times IQR.$$

Equivalently, x_i s flagged as unusual if

$$\frac{x_i - \mu}{\sigma} \leq \frac{Q1 - \mu}{\sigma} - 1.5 \times \frac{IQR}{\sigma}, \quad \text{or} \quad \frac{x_i - \mu}{\sigma} \geq \frac{Q3 - \mu}{\sigma} + 1.5 \times \frac{IQR}{\sigma},$$

which gives

$$z_i \leq Q1_z - 1.5 \times IQR_z, \quad \text{or} \quad z_i \geq Q3_z + 1.5 \times IQR_z.$$

For the standard normal distribution $Q1_z = -0.674$ and $Q3_z = 0.674$ so $IQR_z = 1.349$. Hence, $\mathbb{P}(Z \leq -2.698 \text{ or } Z \geq 2.698) = 0.007$. So about 0.7% of observations would be flagged.

Question

Suppose the lymphocyte count (per litre) from a blood test has a $\mathcal{N}(2.5 \times 10^9, (0.765 \times 10^9)^2)$ distribution. What is the probability that a randomly chosen blood test will have a lymphocyte count per litre between 2.3×10^9 and 2.9×10^9?

Question

Suppose the copper level (in μmol/L) from a blood test has a $\mathcal{N}(18.5, 3.827^2)$ distribution. What is the lowest copper level that would put a blood test result in the highest 1%?