Introduction to Probability

Dr. M.M. Mweu,

Level V MBChB Biostatistics,

31 October, 2019

Introduction

- Probability is the *possibility* that an event will occur
- *Experiment/ trial* – any process of observation/measurement e.g. checking if switch is turned off or counting no. of wounds on a patient
- *Outcomes* – the results of the experiment e.g. counts, yes/no answers etc.
- *Sample space* – list of all possible outcomes
- *Event* - the subset of the sample space

Example: One flip of a fair coin = experiment

Outcome = Head/Tail; *Sample space* = H, T

If we flip a fair coin twice, what is the probability of getting at least one head?

Sample space = HH, HT, TH, TT

Event (A) = at least one head

$n = 3; N = 4$

$$p(A) = \frac{3}{4} = 0.75$$

Probability rules

- Mutually exclusive events:
  - Two events are mutually exclusive if the occurrence of one event excludes the occurrence of the other
  - E.g. If a baby is male, it can't be female also or if a patient is malaria positive, he can't be malaria negative also
  - The probability of occurrence of two mutually exclusive events is the probability of either one of the two events occurring, denoted by $p(A \cup B) = p(A) + p(B)$ (add individual probabilities)
  - E.g. 200 children tested for *Entamoeba histolytica* , where 59 are positive & 141 are negative
    - Probability of *E. histolytica* positivity = $\frac{59}{200} = 0.295$
    - Probability of *E. histolytica* negativity = $\frac{141}{200} = 0.705$
    - Probability of being positive or negative for *E. histolytica* $= 0.295 + 0.705 = 1$
- Independent events:
  - Two events are independent if the occurrence of one event does not influence the occurrence of the other – denoted by $p(A \cap B)$
  - E.g. if the firstborn child is male, this doesn't influence the sex of the second born (can be male or female)

Probability rules

- Independent events:
  - The probability of two independent events is obtained by multiplying individual probabilities of the events

$$p(A \cap B) = p(A) \times p(B)$$

E.g. in a blood bank the following distribution of blood groups was observed

Blood group	n	%
O	45	45
A	29	29
B	21	21
AB	5	5
Total	100	100

What is the probability that the next 2 persons will be in blood group O?

$$p(A \cap B) = 0.45 \times 0.45 = 0.2025 \cong \mathbf{0.20}$$

Important probability distributions

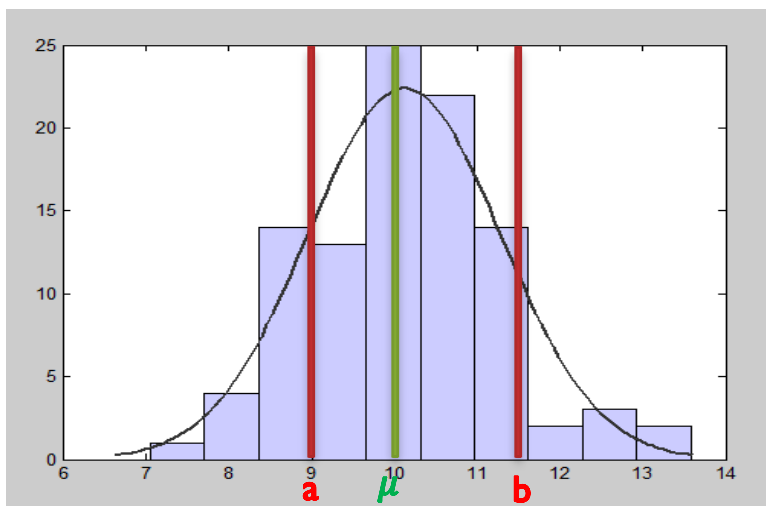
- Probabilities can be assigned to each likely outcome for a variable
- These probabilities usually follow a mathematical formula called a *probability distribution*
- The probability distribution of a variable depends on the type of variable:
  - Continuous variables – temperature, weight, height – follow a **Normal** distribution
  - Binary (yes/no) variables – sex, disease, death – follow a **Binomial** distribution
  - Discrete or rare occurrences – death, rare diseases, counts – follow a **Poisson** distribution. *Poisson* approaches *Binomial* as events become more common and/or the population becomes smaller

Normal probability distribution

- All normally distributed variables are continuous but not vice-versa
- For continuous variables all values within a certain range are observable and the probability associated with one such observation is negligibly small – on a continuous scale a zero probability is assigned to individual points
- However, we can calculate the probability that a variable will take a value betwn two points e.g. a and b.
- For continuous variables a *probability density* is defined – mathematical function such that area under curve betwn 2 points a and b is equal to the probability that the variable will take a value betwn these 2 points

Normal probability distribution

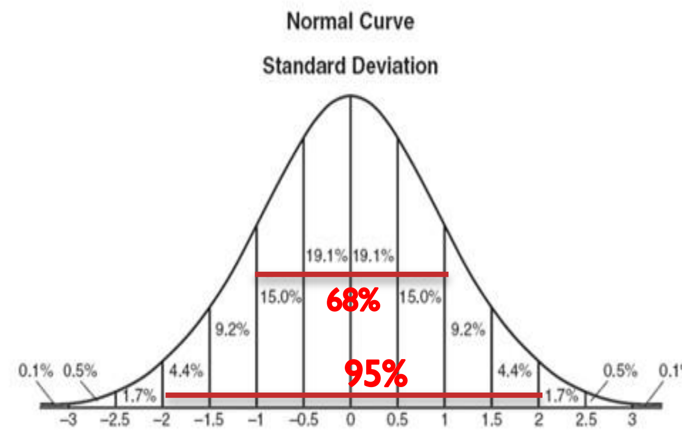
- Two important parameters are needed to calculate probabilities from the normal probability density:
  - Mean (μ) – locates the central position
  - Variance (σ^2) or its square root, the standard deviation (σ) measures the degree of spread about the mean
- If the μ and σ of a normally distributed variable are known, we can determine the probability that such a variable will take a value that lies betwn 2 points a and b.



Normal probability distribution

- To estimate the probability for a normally distributed variable, we normally standardise the variable into a standard normal deviate Z which has a mean 0 and variance

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



Such that: $p[\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}]$

Example

- The variable birth weight is known to be normally distributed with a mean of $3100g$ and variance of $2500g^2$. What is the probability that a baby will weigh greater than $3000g$? **(First draw the probability distribution)**

$$|Z| = \frac{3000 \ - 3100}{\sqrt{2500}} = \frac{100}{50} = 2$$

$$p(bwt > 3000g) = 0.02275 \ (at\ the\ tails)$$

$$= \frac{1 - (0.02275 \times 2)}{2} = 0.47725 + 0.5 = \mathbf{0.97725}$$

Binomial probability distribution

- Important for calculating the probability of d'se
- The binomial distribution describes the behaviour of a random variable X if the following conditions apply:
  - The no. of observations n is fixed
  - Each observation is independent
  - Each observation represents one of two outcomes ("success [Y]" or "failure")
  - The probability of success π is the same for each outcome

$$p(Y = y) = \binom{n}{y} \pi^{y} (1 - \pi)^{(n-y)}$$

Where: $\binom{n}{y} = \frac{n!}{y!(n-y)!}$

and: n = no. of trials/observations

π = probability of success on a single trial

y = no. of successes after n trials

- **If the conditions described above are met then X is said to have a binomial distribution with parameters π and n**

Binomial probability distribution

Example

- According to CDC, 22% of adults in the United States smoke. Suppose we take a sample of 10 people.
  - What is the probability that 5 of them will smoke?

$$p(Y = y) = \binom{n}{y} \pi^{y}(1 - \pi)^{(n-y)}$$

$$p(Y = 5) = \binom{10}{5} 0.22^{5}(1 - 0.22)^{(10-5)}$$

$$= 252 \times 0.000515 \times 0.289 = \mathbf{0.0375}\ \boldsymbol{or}\ \mathbf{3.75\%}$$

  - What is the probability that *2 or less* will be smokers?

$$p(Y \leq 2) = p(Y = 0) + p(Y = 1) + p(Y = 2)$$

$$\binom{10}{0} 0.22^{0}(1 - 0.22)^{(10-0)} + \binom{10}{1} 0.22^{1}(1 - 0.22)^{(10-1)} + \binom{10}{2} 0.22^{2}(1 - 0.22)^{(10-2)} = 0.083 + 0.235 + 0.298 = \mathbf{0.616}\ \boldsymbol{or}\ \mathbf{61.6\%}$$

Binomial probability distribution

Example

- What is the probability that *at least one* will smoke?

Probability of at least one being a smoker = $1 - p(Y = 0)$

$$= 1 - \binom{10}{0} 0.22^{0} (1 - 0.22)^{(10-0)}$$

$$= 1 - 0.083 = \mathbf{0.917 \text{ or } 91.7\%}$$

- The mean and variance of a binomial distribution can be shown to be:

$$\mu = n\pi$$

$$\sigma^2 = n\pi(1 - \pi)$$

- For large values of n the distribution of the binomial variable X and the proportion π are approximately normal. Hence, a *normal approximation to binomial distribution* is possible when:

$$n\pi \geq 5$$

and

$$n(1 - \pi) \geq 5$$

Binomial probability distribution

Example

- If the probability of a certain disease is thought to be 0.2. What is the probability that in a sample of 50 individuals, 2 or more will get the disease?

$$\pi = 0.2$$

$$n = 50$$

$$n\pi = 50 \times 0.2 = 10 \; \boldsymbol{and}$$

$$n(1 - \pi) = 50 \times (1 - 0.2) = 40$$

Hence normal approximation to Binomial distribution is possible

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

$$where \; \mu = 0.2 \times 50 = \mathbf{10} \; and \; \sigma^2 = 50 \times 0.2 \times 0.8 = \mathbf{8}$$

$$p(Y \geq 2) = \frac{2 - 10}{\sqrt{8}} (\boldsymbol{need\ to\ draw\ to\ see\ area})$$

$$|Z| = 2.83$$

$$\frac{1 - (0.00233 \times 2)}{2} = 0.4977 + 0.5 = \mathbf{0.9977}$$

Poisson probability distribution

- The Poisson distribution is a discrete probability distribution for the *counts* (or *rates*) of events that occur **randomly** (can't predict when they will happen e.g. time to next phone call) in a given interval of time or space
- If Y = the no. of events in a given interval and if the *mean* no. of events per interval is λ, the probability of observing y events in a given interval is given by:

$$p(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \; (NB: e \text{ is a constant} - \text{see } e^x \text{ in calculator})$$

$$\boldsymbol{where\ y = 0, 1, 2, 3\ etc}$$

Example

- Births in a hospital occur randomly at an average of 1.8 births per hour. What is the probability of observing 4 births in a given hour at the hospital?

$$p(Y = 4) = \frac{1.8^4 e^{-1.8}}{4!} = \mathbf{0.0723}$$