HEADS
PhD PROGRAMME IN HEALTH DATA SCIENCE

Probability Distributions

1. Bernoulli distribution
2. Binomial distribution
3. Probability density
4. Normal distribution
5. Standard distribution

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO
FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO
MEDCIDS DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE
CINTESIS Health. Research.

Probability Distributions

- A random variable (experiment) can take any value (result) with a given probability within a set of values (sample space).

- A probability distribution shows the probability of all possible values of a variable.

Probability Distributions (Bernoulli)

In categorical variables the distribution associates a probability to each category

Example:

For the variable “gender of newborn” the probability distribution is defined by 2 probabilities (actually one because it is a dichotomous variable)

$$p(\text{gender=Male}) = p$$

$$p(\text{gender=Female}) = 1-p$$

Probability Distributions (Bernoulli)

Coding male=1 and female =0 we can use both probabilities in the same equation

$$p(gender = x) = p^x * (1-p)^{(1-x)}$$

$$p(gender = 1) = p^1 * (1-p)^0 = p$$

$$p(gender = 0) = p^0 * (1-p)^1 = 1-p$$

This function is called **Bernoulli distribution** and

represents the probability distribution of a dichotomous variable

The variable “gender of newborn” follows a **Bernoulli distribution** with probability p:

gender ~ Ber (p)

Probability Distributions (Binomial)

What is the probability that the next two newborns are boys?

Probability Distributions (Binomial)

What is the probability that the next two newborns are boys?

Each birth is independent of the other births

$$p(gender_1=m, gender_2=m) = p(gender_1=m) \times p(gender_2=m)$$

Probability Distributions (Binomial)

What is the probability that the next two newborns are boys?

Each birth is independent of the other births

$$p(gender_1=m, gender_2=m) = p(gender_1=m) \times p(gender_2=m)$$

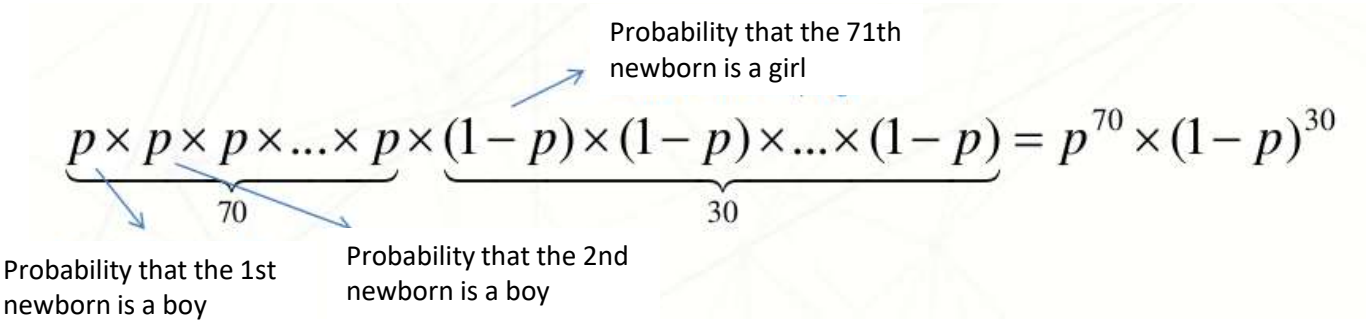
Consider a group of 100 newborns.

Which is the probability of observing exactly 70 boys?

Probability Distributions (Binomial)

Consider a group of 100 newborns. Which is the probability of observing exactly 70 boys?

Let's look the probability of the first 70 babies being boys (and, of course, the next 30 being girls)



Probability Distributions (Binomial)

But the babies may not be born in this order, first 70 being boys followed by 30 girls

We need to do all possible combinations for 70 boys and 30 girls $\binom{100}{70} = \frac{100!}{70! \times 30!}$

but the probability is always the same $p^{70} \times (1-p)^{30}$

So, what is the probability of observing exactly 70 newborn boys? $\binom{100}{70} p^{70} \times (1-p)^{30}$

Probability Distributions (Binomial)

We call it **Binomial distribution**

$$P(X = x) = \binom{n}{x} p^x \times (1-p)^{n-x}$$

If X is the number of boys of a total of n newborns and p the probability of being a boy, then X follows a binomial distribution

X ~ Bin(n,p)

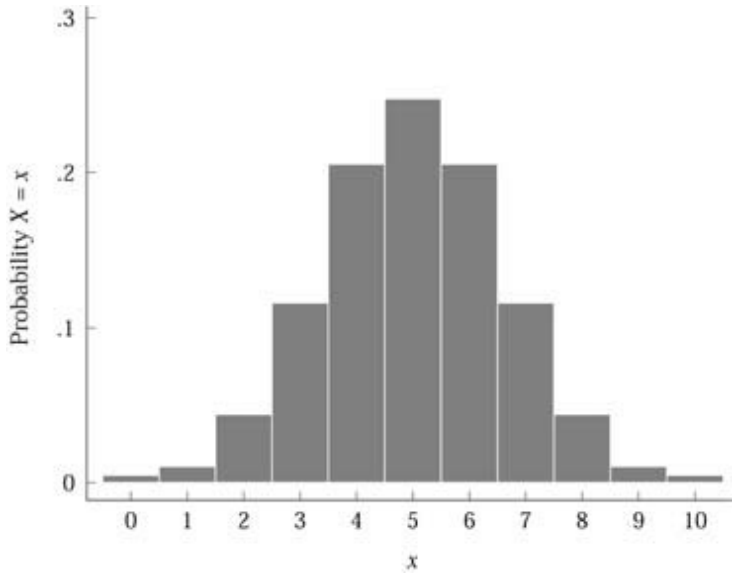
The mean of X is **np** and the variance is **np(1-p)**

Probability Distributions (Binomial)

Example of probability distribution for X ~Bin(10,0.5)

mean = 10x0.5 = 5

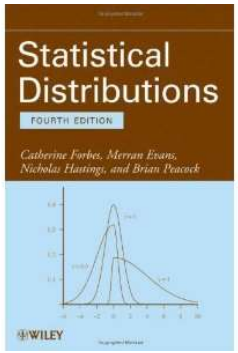
variance= 10x0.5x0.5 = 2.5



Probability Distributions (other cases)

- But, if instead of X being the number of boys in a sample of 100 babies, so a limited number of events, it was, for example, the number of epileptic seizures in a month?

- In this case, the distribution could be a Poisson



- In reality, there are a huge number of theoretical distributions and we just need to choose the one which better represents our data

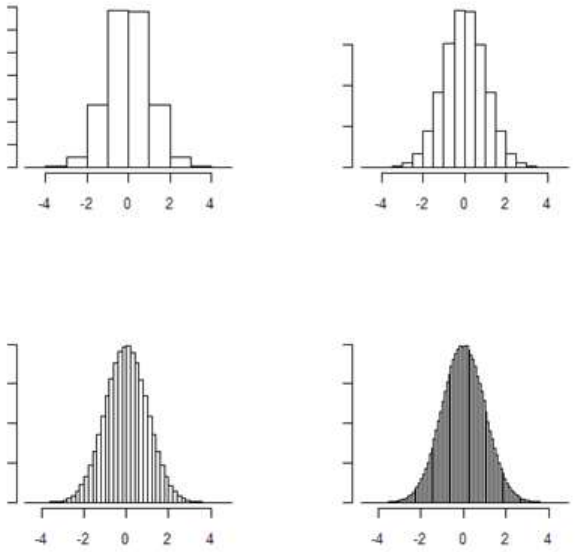
Probability Distributions (density)

For continuous variables, we can not present the probability of a given value

Instead of that, we can present the probability of an interval of values

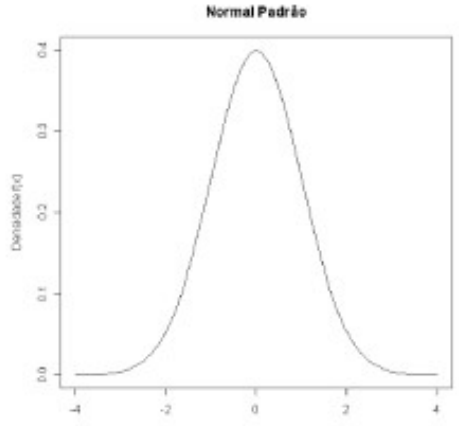
Imagine that we can choose smaller and smaller intervals

As the intervals get shorter we start seeing a curve called **probability density**



Probability Distributions (density)

The most famous is the Normal distribution, also called Gaussian curve, but there are many other

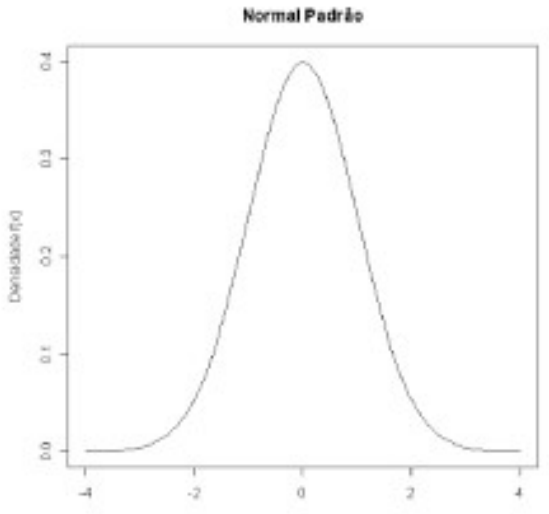


But, why do we like to fit theoretical distribution to our data?

Probability Distributions (density)

Why do we like to fit theoretical distribution to our data?

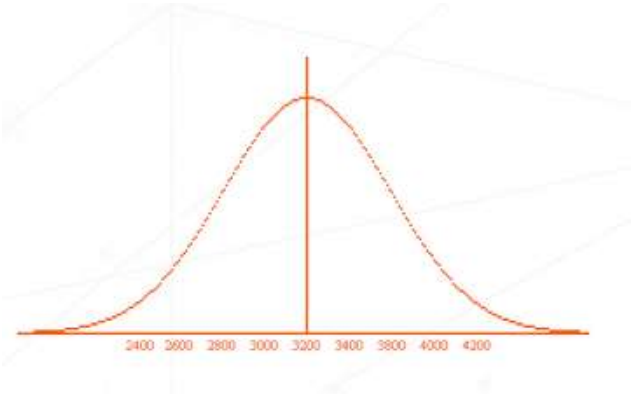
When the empirical frequency distribution of the variable that we are **studying approaches a theoretical distribution of probabilities** (distribution described with a mathematical model) **we can use the theoretical knowledge** we have about this distribution to answer data-related questions.



Probability Distributions (density)

Example:

- If we know that the newborn's weight distribution approximates the theoretical distribution represented by this probability density curve:

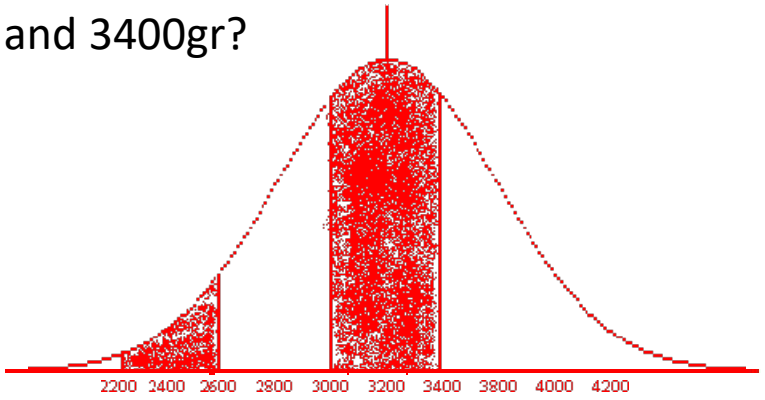


Probability Distributions (density)

Example:

- If we know that the newborn's weight distribution approximates the theoretical distribution represented by this probability density curve

- **We can ask:** it is more likely that a baby will be born weighing between 2200gr and 2600gr or between 3000gr and 3400gr?



Normal Distribution

Normal Distribution

This is the most famous of all theoretical distributions

But, **why**?

Normal Distribution

This is the most famous of all theoretical distributions

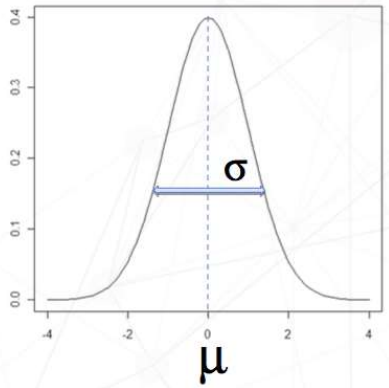
But, **why**?

- ✓ Has interesting properties
- ✓ Many biological variables have a distribution similar to a normal distribution and that allows us to use the properties and theoretical knowledge that we have about this to explore our data

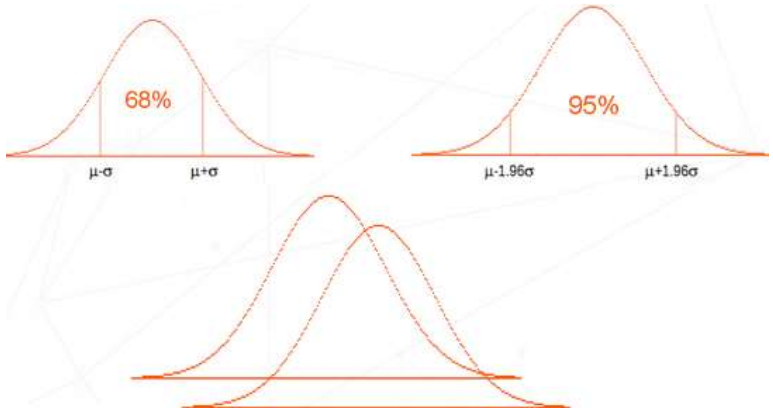
Normal Distribution

Is completely defined by the mean (μ) standard deviation (σ)

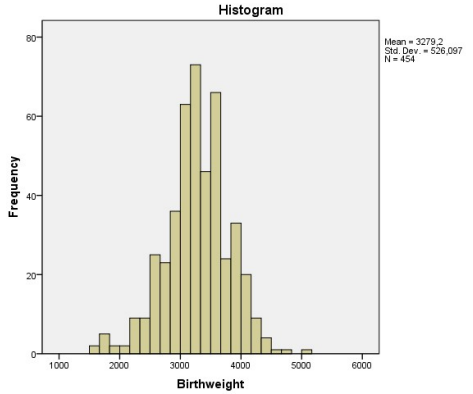
Symmetrical with respect to the mean (mean=median)



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Normal Distribution



Example: Consider the variable newborn weight

The histogram suggests that variable has a normal distribution

Assuming a normal distribution, we can see that 95% of the newborn weighing between 2227 and 4331 (mean=3279 and sd=526) is

$$3279 - \mathbf{2} \times 526 = 2227$$

$$3279 + \mathbf{2} \times 526 = 4331$$

Normal Distribution (z-Score)

We can transform any normal distribution into the standard distribution. For that, we need to subtract from each value the sample mean and divide by the standard deviation (z-score).

The standard normal distribution has mean =0 and sd=1.

Instead of saying that the baby weight was 2130 g...

We can compute the z-score (2130-3279)/526 and say that:
The baby was born with a z-score=-2.37, that is, the baby was 2.37 standard deviations below the mean.