

## Unit 3: Distributions

Statistics 102 Teaching Team

February 17, 2020

Random Variables

The Binomial distribution

The Normal distribution

The Poisson distribution

# Random Variables

## MAIN IDEAS THIS SECTION

- Definition of random variables
- Distributions of random variables
- Mean, variance, and standard deviation for random variables


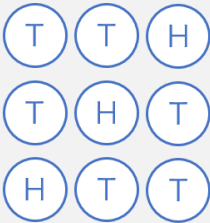
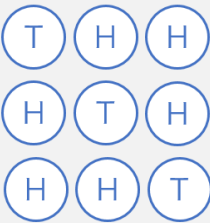

## DEFINITION OF A RANDOM VARIABLE

A *random variable* is a function that maps each event in a sample space to a number.

- A *discrete random variable* takes on a finite number of values.

Suppose  $X$  is the number of heads in 3 tosses of a fair coin.

- $X$  can take on the values 0, 1, 2, 3.

			
$X = 0$	$X = 1$	$X = 2$	$X = 3$

# DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

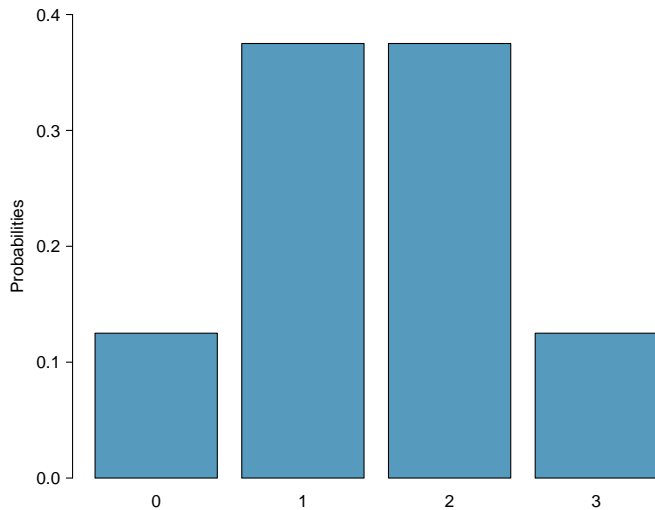
The distribution of a discrete random variable is the collection of its values and the probabilities associated with those values.

The probability distribution for  $X$  is as follows:

$x_i$	0	1	2	3
$P(X = x_i)$	1/8	3/8	3/8	1/8

$$\sum_{x=0}^3 P(X = x_i) = 1$$

## BAR GRAPH SHOWING A DISTRIBUTION



# USING A SIMULATION TO CONSTRUCT A PROBABILITY DISTRIBUTION

Distributions of random variables that arise in science can be more complex.

In lab, we will run a simulation to view the distribution of good responses in a clinical trial with 8 participants, under the assumption that the probability of a good response is 0.15.



## EXPECTATION OF A RANDOM VARIABLE

If  $X$  has outcomes  $x_1, \dots, x_k$  with probabilities  $P(X = x_1), \dots, P(X = x_k)$ , the expected value of  $X$  is the sum of each outcome multiplied by its corresponding probability:

$$E(X) = x_1P(X = x_1) + \dots + x_kP(X = x_k) = \sum_{i=1}^k x_iP(X = x_i)$$

The Greek letter  $\mu$  may be used in place of the notation  $E(X)$  and is sometimes written  $\mu_X$ .

## EXPECTATION...

In the coin tossing example,

$$\begin{aligned} E(X) &= 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 3P(X = 3) \\ &= (0)(1/8) + (1)(3/8) + (2)(3/8) + (3)(1/8) \\ &= 12/8 \\ &= 1.5 \end{aligned}$$

## VARIANCE AND SD OF A RANDOM VARIABLE

If  $X$  takes on outcomes  $x_1, \dots, x_k$  with probabilities  $P(X = x_1), \dots, P(X = x_k)$  and expected value  $\mu = E(X)$ , then the variance of  $X$ , denoted by  $\text{Var}(X)$  or  $\sigma^2$ , is

$$\begin{aligned}\text{Var}(X) &= (x_1 - \mu)^2 P(X = x_1) + \dots + (x_k - \mu)^2 P(X = x_k) \\ &= \sum_{j=1}^k (x_j - \mu)^2 P(X = x_j)\end{aligned}$$

The standard deviation of  $X$ , written as  $\text{SD}(X)$  or  $\sigma$ , is the square root of the variance. It is sometimes written  $\sigma_X$ .

## VARIANCE AND SD...

In the coin tossing example,

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 P(X = x_1) + \cdots + (x_4 - \mu)^2 P(X = x_4) \\ &= (0 - 1.5)^2(1/8) + (1 - 1.5)^2(3/8) + (2 - 1.5)^2(3/8) + (3 - 1.5)^2(1/8) \\ &= 3/4\end{aligned}$$

The standard deviation is  $\sqrt{3/4} = \sqrt{3}/2 = 0.866$ .

## The Binomial distribution

# BINOMIAL RANDOM VARIABLES

One specific type of discrete random variable is a binomial random variable.

$X$  is a binomial random variable if it represents the number of successes in  $n$  independent replications of an experiment where

- Each replicate has two possible outcomes: either success or failure
- The probability of success  $p$  in each replicate is constant

A binomial random variable takes on values  $0, 1, 2, \dots, n$ .

The number of heads in 3 tosses of a fair coin is a binomial random variable with parameters  $n = 3$  and  $p = 0.5$ .

# THE BINOMIAL COEFFICIENT

The binomial coefficient  $\binom{n}{x}$  is the number of ways to choose  $x$  items from a set of size  $n$ , where the order of the choice is ignored.

Mathematically,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- $n = 1, 2, \dots$
- $x = 0, 1, 2, \dots, n$
- For any integer  $m$ ,  $m! = (m)(m-1)(m-2) \cdots (1)$

# FORMULA FOR THE BINOMIAL DISTRIBUTION

Let  $x$  = number of successes in  $n$  trials

$$P(x \text{ successes}) = \binom{\# \text{ of trials}}{\# \text{ of successes}} p^{\# \text{ of successes}} (1 - p)^{\# \text{ of trials} - \# \text{ of successes}}$$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

*Parameters of the distribution:*

- $n$  = number of trials
- $p$  = probability of success

Shorthand notation:  $X \sim \text{Bin}(n, p)$



## MEAN AND SD FOR A BINOMIAL RANDOM VARIABLE

For a binomial distribution with parameters  $n$  and  $p$ , it can be shown that:

- Mean =  $np$
- Standard Deviation =  $\sqrt{np(1-p)}$

The derivation is not shown here nor in the text; it will not be asked for on a problem set or exam.

## CALCULATING BINOMIAL PROBABILITIES IN R

The function `dbinom()` is used to calculate  $P(X = k)$ .

- `dbinom(k, n, p)`:  $P(X = k)$

The function `pbinom()` is used to calculate  $P(X \leq k)$  or  $P(X > k)$ .

- `pbinom(k, n, p)`:  $P(X \leq k)$
- `pbinom(k, n, p, lower.tail = FALSE)`:  $P(X > k)$

## The Normal distribution

# CONTINUOUS RANDOM VARIABLES

A discrete random variable takes on a finite number of values.

- Number of heads in a set of coin tosses
- Number of people who have had chicken pox in a random sample

A continuous random variable can take on any real value in an interval.

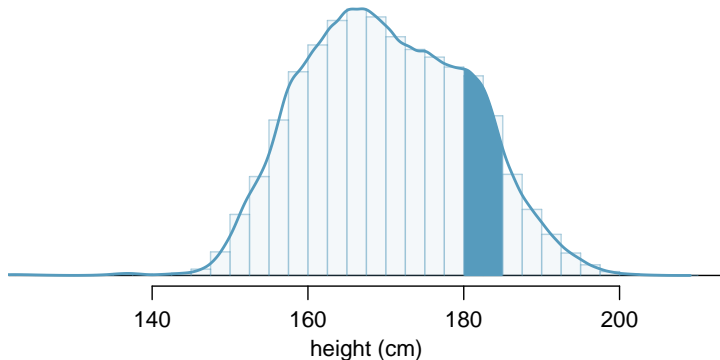
- Height in a population
- Blood pressure in a population

A general distinction to keep in mind: discrete random variables are *counted*, but continuous random variables are *measured*.

# PROBABILITIES FOR CONTINUOUS DISTRIBUTIONS

Two important features of continuous distributions:

- The total area under the density curve is 1.
- The probability that a variable has a value within a specified interval is the area under the curve over that interval.



## PROBABILITIES FOR CONTINUOUS DISTRIBUTIONS...

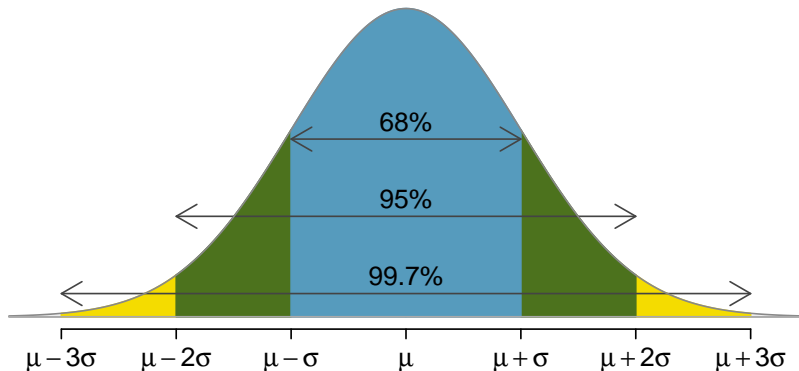
When working with continuous random variables, probability is found for intervals of values rather than individual values.

- The probability that a continuous r.v.  $X$  takes on any single individual value is 0. That is,  $P(X = x) = 0$ .
- Thus,  $P(a < X < b)$  is equivalent to  $P(a \leq X \leq b)$ .

# THE NORMAL DISTRIBUTION

According to the Empirical Rule, for any normal distribution,

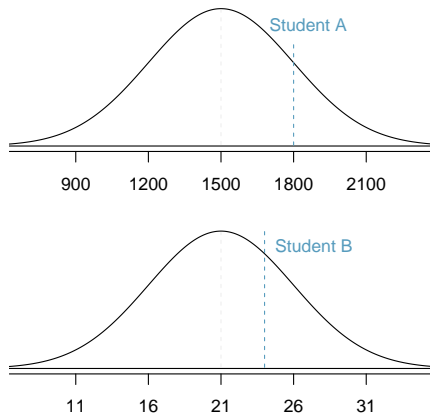
- approximately 68% of the data are within 1 SD of the mean
- approximately 95% of the data are within 2 SDs of the mean
- approximately 99.7% of the data are within 3 SDs of the mean



## A NORMAL EXAMPLE

The distribution of test scores on the SAT and the ACT are both nearly normal.

Suppose that one student scores an 1800 on the SAT (Student A) and another student scores a 24 on the ACT (Student B). Which student performed better?





# STANDARD NORMAL DISTRIBUTION

A *standard normal* distribution is defined as a normal distribution with mean 0 and variance 1. It is often denoted as  $Z \sim N(0, 1)$ .

Any normal random variable  $X$  can be transformed into a standard normal random variable  $Z$ .

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

## A NORMAL EXAMPLE...

- SAT scores are  $N(1500, 300)$ . ACT scores are  $N(21, 5)$ .
- $x_A$  represents the score of Student A;  $x_B$  represents the score of Student B.

$$Z_A = \frac{x_A - \mu_{SAT}}{\sigma_{SAT}} = \frac{1800 - 1500}{300} = 1$$

$$Z_B = \frac{x_B - \mu_{ACT}}{\sigma_{ACT}} = \frac{24 - 21}{5} = 0.6$$

## CALCULATING NORMAL PROBABILITIES (I)

What is the percentile rank for a student who scores an 1800 on the SAT for a year in which the scores are  $N(1500, 300)$ ?

1. Calculate a Z-score. If  $X$  is a normal random variable with mean  $\mu$  and standard deviation  $\sigma$ ,

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

is a standard normal random variable ( $\mu = 0, \sigma = 1$ ).

2. Calculate the normal probability.

- `pnorm(z)` calculates the area (i.e., probability) to the left of  $z$

```
pnorm(1)
```

```
## [1] 0.8413447
```

3. Alternatively, let R do the work ...

```
pnorm(1800, 1500, 300)
```

```
## [1] 0.8413447
```

## CALCULATING NORMAL PROBABILITIES (II)

What score on the SAT would put a student in the 99<sup>th</sup> percentile?

1. Identify the Z-value. `qnorm(p)` calculates the value  $z$  such that for a standard normal variable  $Z$ ,  $p = P(Z \leq z)$ .

```
qnorm(0.99)
```

```
## [1] 2.326348
```

2. Calculate the score,  $X$ . If  $Z$  is distributed standard Normal, then

$$X = \sigma Z + \mu,$$

is Normal with mean  $\mu$  and standard deviation  $\sigma$ .

$$X = \sigma Z + \mu = 300(2.33) + 1500 = 2199$$

3. Alternatively, let R do the work ...

```
qnorm(0.99, 1500, 300)
```

```
## [1] 2197.904
```

## The Poisson distribution

# INTRODUCTION TO THE POISSON DISTRIBUTION

The Poisson distribution is used to calculate probabilities for rare events that accumulate over time.

It is used most often in settings where events happen at a rate  $\lambda$  per unit of population and per unit time, such as the annual incidence of a disease in a population.

- Typical example: for children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in 2010.
- Always take care to note and understand the units.

## EXAMPLE: OUTBREAKS OF CHILDHOOD LEUKEMIA

Fortunately, childhood cancers are rare.

For children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in the decade from 2000 - 2010. Approximately 20% of the US population are in this age range.

- What is the incidence rate over a 5 year period?
- In a small city of 75,000 people, what is the probability of observing exactly 8 cases of ALL over a 5 year period?
- In the small city, what is the probability of observing 8 or more cases over a 5 year period?

# POISSON DISTRIBUTION

Suppose events occur over time in such a way that

1. The probability an event occurs in an interval is proportional to the length of the interval.
2. Events occur independently at a rate  $\lambda$  per unit of time.

Then the probability of exactly  $x$  events in one unit of time is

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

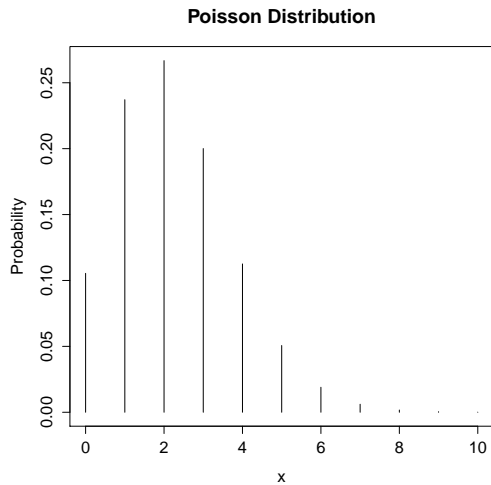
The probability of exactly  $x$  events  $t$  units of time is

$$P(X = x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots$$

Derivation given in more theoretical courses, such as Stat 110.



## POISSON DISTRIBUTION WITH $\lambda = 2.25$



# POISSON MEAN AND STANDARD DEVIATION

For the Poisson distribution modeling the number of events in one unit of time:

- The mean is  $\lambda$ .
- The standard deviation is  $\sqrt{\lambda}$ .

In  $t$  units of time, the mean and standard deviation are, respectively,  $\lambda t$  and  $\sqrt{\lambda t}$ .

## CHILDHOOD LEUKEMIA CASES (*Ol Biostat*, EXAMPLE 3.37)

The incidence rate of ALL in a year is 30 per 1,000,000 children:

- $\frac{30}{1,000,000} = 0.00003 = 3 \times 10^{-5}.$

The incidence rate over a 5-year period is (5)(30) per 1,000,000 children:

- $\frac{150}{1,000,000} = 0.00015 = 1.5 \times 10^{-4}.$

## WHAT ABOUT A CITY OF SIZE 75,000?

In a city of 75,000 people, approximately  $(75,000)(0.20) = 15,000$  children will be age 0 - 14 (from slide 31).

The five-year rate of new cases for the city would be:

$$(1.5 \times 10^{-4})(15,000) = 2.25$$

## WHAT IS THE PROBABILITY OF 8 CASES OVER 5 YEARS?

$$P(X = 8) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{(-2.25)} (2.25)^8}{8!}$$

Easiest to calculate this in R ...

Suppose  $X$  has a Poisson distribution with parameter  $\lambda$ .

- `dpois(k, lambda)`:  $P(X = k)$

```
dpois(8, lambda = 2.25)
```

```
## [1] 0.001717027
```

## WHAT IS THE PROBABILITY OF 8 OR MORE CASES?

Would 8 or more cases be a rare event?

- Calculate  $P(X \geq 8) = 1 - P(X \leq 7)$ .

Suppose  $X$  has a Poisson distribution with parameter  $\lambda$ .

- `ppois(k, lambda):  $P(X \leq k)$`

```
1 - ppois(7, lambda = 2.25)
```

```
## [1] 0.002267088
```

- `ppois(k, lambda, lower.tail = FALSE):  $P(X > k)$`

```
ppois(7, lambda = 2.25, lower.tail = FALSE)
```

```
## [1] 0.002267088
```

## DISTRIBUTIONS SUMMARY TABLE

	Binomial	Normal	Poisson
Parameters	$n, p$	$\mu, \sigma$	$\lambda$
Possible values	$0, 1, \dots, n$	$(-\infty, \infty)$	$0, 1, \dots, \infty$
Mean	$np$	$\mu$	$\lambda$
Standard Deviation	$\sqrt{np(1-p)}$	$\sigma$	$\sqrt{\lambda}$