Review of Probability and Statistics

Juergen Meinecke

Roadmap

- Univariate Probability
  - Random Variables, Probability Distributions

Definition

The mutually exclusive potential results of a random process are called **outcomes**.

Definition

The set of all possible outcomes is called **sample space**.

Definition

An **event** is a subset of the sample space.

Example:

random process *rolling a die*

- outcomes: e.g., rolling ‘five dots’
- sample space: {one dot, two dots, …, six dots}
- event: e.g., {three dots, five dots}

Example:

random process

number of kangaroos spotted during my morning run

(in my local nature reserve)

- outcomes: e.g., five kangaroos
- sample space: {one kangaroo, two kangaroos, ..., fifty kangaroos}
  (this one's tricky, what's the upper limit?)
- example of an event: more than six kangaroos

Definition

A **random variable** Y is the numerical representation of an outcome in a random process.

Rolling a die example

- the outcome 'one dot' is represented by the number 1
- the outcome 'two dots' is represented by the number 2 and so forth

Note: outcomes can be represented by any number

For instance, the outcome 'one dot' could also be represented by the number 247

I picked the obvious and sensible candidates

Random variables save us a lot of notation

Consider the event

not less than four but fewer than ten kangaroos

(sounds clumsy, doesn't it?)

Using random variables, this can be concisely summarized mathematically as

$$4 \leqslant Y < 10$$

Definition

The **probability distribution** of a random variable Y is the full characterization of probabilities for all possible outcomes of a random process.

(this applies to *discrete* random variables; the definition for *continuous* random variables would be slightly different)

Example

- age of EMET2007 students
- suppose ages vary between 18 and 26
  (just to keep things simple; sorry if you are older!)

Example: probability distribution of age

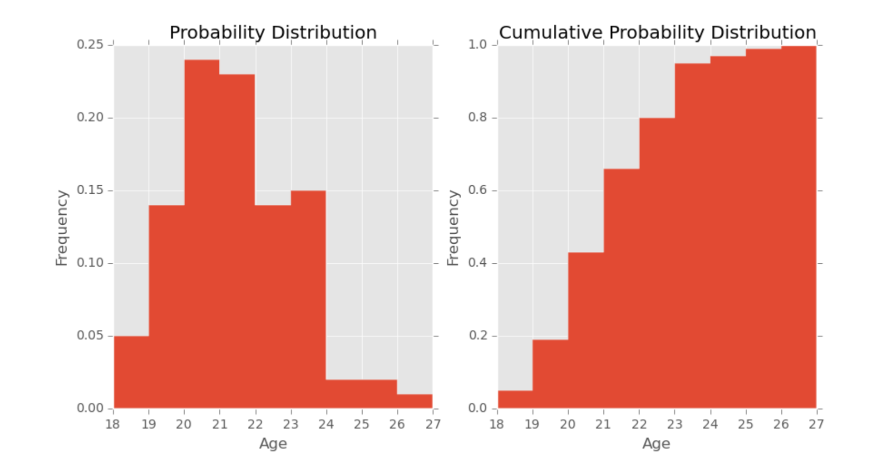
$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \\ 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases}$$

Note: little y is called the *realization* of the random variable, it's merely a placeholder for a number between 18 and 26

Example: **cumulative** probability distribution of heights

$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \\ 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases} \qquad \Pr(Y \leqslant y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.19 & \text{if } y = 19 \\ 0.43 & \text{if } y = 20 \\ 0.66 & \text{if } y = 21 \\ 0.80 & \text{if } y = 22 \\ 0.95 & \text{if } y = 23 \\ 0.97 & \text{if } y = 24 \\ 0.99 & \text{if } y = 25 \\ 1.00 & \text{if } y = 26 \end{cases}$$

Frequency plot (histogram)



Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Expected Value, Standard Deviation, and Variance

Definition

Suppose the random variable Y takes on k possible values $y_1, \dots, y_k$. The **expected value** is given by

$$E[Y] := \sum_{j=1}^{k} y_j \cdot \Pr(Y = y_j) \tag{1}$$

Occasionally we also call this the **population mean** or simply the **mean** or the **expectation**.

Often times, the expected value is also denoted μ_Y.

Example: age distribution

Recall

$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \end{cases} \qquad \Pr(Y = y) = \begin{cases} 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases}$$

We have $y_1 = 18, y_2 = 19, \dots, y_9 = 26$, therefore

$$\begin{aligned} E[Y] &= \sum_{j=1}^{9} y_j \cdot \Pr(Y = y_j) \\ &= 18 \cdot 0.05 + 19 \cdot 0.14 + \ldots + 26 \cdot 0.01 \\ &= 20.96 \end{aligned}$$

Properties of the expected value

1. Let c be a constant, then $E[c] = c$
2. Let c be a constant and Y be a random variable, then

   $$\mathrm{E}[c + Y] = c + \mathrm{E}[Y]$$

   $$\mathrm{E}[c \cdot Y] = c \cdot \mathrm{E}[Y]$$

   It follows that for two constants c and d,

   $$E[c + d \cdot Y] = c + d \cdot \mathrm{E}[Y]$$

3. Let X and Y be random variables, then

   $$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

   $$\mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y]$$

(Can you prove all of these?)

Definition

The r^{th} **moment** of a random variable Y is given by

$$m_r(Y) := E[Y^r], \qquad \text{for } r = 1, 2, 3, \ldots$$

It is obivous that the first moment and the expected value are the same

Definition

The **population variance** is defined by

$$\text{Var}\,[Y] := \sum_{j=1}^{k} (y_j - \mu_y)^2 \cdot \Pr(Y = y_j)$$

Often times, the variance is denoted by σ_Y^2.

Definition

The **population standard deviation** is defined by

$$\text{StD}[Y] := \sqrt{\text{Var}\,[Y]}$$

It follows immediately that the population standard deviation is simply σ_Y.

Example: age distribution

We have $y_1 = 18, y_2 = 19, \ldots, y_9 = 26$

Doing the math

$$\text{Var}\,[Y] = \sum_{j=1}^{9} (y_j - \mu_y)^2 \cdot \Pr(Y = y_j)$$

$$= (18 - 20.96)^2 \cdot 0.05 + (19 - 20.96)^2 \cdot 0.14 + \ldots$$

$$(26 - 20.96)^2 \cdot 0.01$$

$$= 2.74$$

Therefore

$$\text{StD}[Y] = 1.66$$

Properties of the variance

1. Let c be a constant, then $\text{Var}\left[c\right] = 0$
2. Let c be a constant and Y be a random variable, then
$$\text{Var}\left[c + Y\right] = \text{Var}\left[Y\right]$$
$$\text{Var}\left[c \cdot Y\right] = c^2 \cdot \text{Var}\left[Y\right]$$
3. Let X and Y be random variables, then
$$\text{Var}\left[X + Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] + 2 \cdot \text{Cov}(X, Y)$$
$$\text{Var}\left[X - Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] - 2 \cdot \text{Cov}(X, Y)$$

(Can you prove all of these?)

We haven't yet defined what we mean by '$\text{Cov}(X, Y)$', we'll do this later when we discuss bivariate probability

Review of Probability and Statistics

Juergen Meinecke

Roadmap

- Univariate Probability
  - Population versus Sample

Definition

A **population** is a well defined group of subjects.

The population contains all the information on the underlying probability distribution

Subjects don't need to be people only

Examples

- Australian citizens
- kangaroos in Tidbinbilla
- leukocytes in the bloodstream
- protons in an atom
- lactobacilli in yogurt

Definition

The **population size** N is the number of subjects in the population.

We typically think that N is ‘very large’

In fact, it is so large that observing the entire population becomes impossible

Mathematically, we think that $N = \infty$, even though in many applications this is clearly not the case

Setting $N = \infty$ merely symbolizes that we are not able to observe the entire population

Example: population of Australian citizens

Clearly, $N = 27,045,358$
(last checked afternoon 19 February 2024)

For all practical purposes it is so large that it might as well have been $N = \infty$

Example: kangaroos in Tidbinbilla

I have no idea how many kangaroos live in Tidbinbilla
(therefore, I do not know the actual population size)

I could ask the park ranger, but suppose she also doesn't know

We treat the population size as unimaginable: $N = \infty$

The point is:
for some reason we are not able to observe the entire population
(too difficult, too big, too costly)

Instead, we only have a random sample of the population

Definition

In a **random sample**, n subjects are selected (without replacement) at random from the population.

Each subject of the population is equally likely to be included in the random sample.

Typically, n is much smaller than N

Most important, $n < N \leqslant \infty$

The random variable for the i-th randomly drawn subject is denoted Y_i

Definition

Because each subject is equally likely to be drawn and the distribution is the same for all i, the random variables $Y_1, \ldots, Y_n$ are **independently and identically distributed (i.i.d.)** with mean μ_Y and variance σ_Y^2.

We write $Y_i \sim \text{i.i.d.}(\mu_Y, \sigma_Y^2)$.

Given a random sample, we observe the n realizations $y_1, \ldots, y_n$ of the i.i.d. random variables $Y_1, \ldots, Y_n$

What do we do with a random sample of i.i.d. data?

Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Sample Average

In analogy to the mean of a population,
we define the mean of a subset of the population:

Definition

The **sample average** is the average outcome in the sample:

$$\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Sometimes we call the sample average also the sample mean.

It should be obvious that this is a sensible definition

Let’s say we are interested in learning about the weights of kangaroos in Tidbinbilla

We drive to Tidbinbilla and somehow randomly collect 30 roos and measure their weights

This will give us a random sample of size 30 of kangaroo weights

It’s easy to calculate the average weight of these 30 roos

Suppose we obtain a sample average of 52kg

There is a huge difference between the population mean and the sample mean

There is only one population, therefore there is only one population mean

But there are many different random subsets (samples) of the population, each of which results in a (potentially) different sample average

Let’s say we drive to Tidbinbilla for a second time, again randomly collect 30 roos and measure their weights

Should we expect to obtain a sample average of 52kg?

It is unlikely that the second time around we collect exactly the same 30 roos (while it is possible, it is not probable)

If we collect a different subset of 30 kangaroos, chances are that we come up with a different sample average

Suppose we obtain a sample average of 49kg

And now we collect a third random sample ...

...and obtain a sample average of 55kg

And so forth ...

This illustrates that the sample average itself is a random variable!

Random variables have statistical distributions

What distribution does the sample average have?

- what is its expected value?
- what is its variance?
- what is its standard deviation?
- what is its shape?

Let $Y_i \sim$ i.i.d.(μ_Y, σ_Y^2) for all i

We don't know exactly which distribution generates the Y_i, but at least we know its expected value and its variance
(turns out this is all we need to know!)

Each random variable Y_i has

- population mean μ_Y
- variance σ_Y^2

Expected value

$$\begin{aligned}
\mathrm{E}[\bar{Y}] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_Y \\
&= \frac{1}{n} n\mu_Y \\
&= \mu_Y
\end{aligned}$$

(all of this follows by the properties of expected values)

Variance

$$
\begin{aligned}
\text{Var}\left[\bar{Y}\right] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[Y_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 \\
&= \frac{1}{n^2} n\sigma_Y^2 \\
&= \sigma_Y^2/n
\end{aligned}
$$

(all of this follows by the properties of variances, and realizing that $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ (why?))

Standard deviation

$$\mathrm{StD}(\bar{Y}) = \sigma_Y/\sqrt{n}$$

(that's an easy one, given that we know the variance)