Section 9: Summary

$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ if A and B are mutually exclusive outcomes.

$\mathbf{P}(A \cap B) = \mathbf{P}(A) \times \mathbf{P}(B)$ provided A and B are independent.

$\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\,\mathbf{P}(B)$

$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$

The number of different orders in which n unique objects can be placed is $n!$

Permutations: ${}_nP_r = \frac{n!}{(n-r)!}$ is the number of ways of choosing r items from n when the order of the chosen items matters.

Combinations: ${}_nC_r = \frac{n!}{(n-r)!r!}$ is the number of ways of choosing r items from n when the order of the chosen items does not matter.

Section 10

Statistics

In this section we summarise the key issues in pages 14–20 of the basic probability teach-yourself document. This presentation is intended to be reinforced by the examples in the teach-yourself document and questions 13 and 14 in examples paper 10.

The main focus is on the mean and standard deviation of a probability distribution. We also explain how to calculate a range within which we are (say) 95% sure that the true value of an experimental reading will lie.

Mean

The mean μ, of a population of values x_i (where i goes from 1 to N), is defined.

$$\mu = \frac{\text{Sum of all the values}}{\text{Number of values}} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Imagine a pack of cards with all the jokers and picture cards removed. We are only concerned with the numerical value of the cards. We have four each of all the numbers from one to ten so $N = 40$.

$$\text{Arithmetic mean: } \mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{220}{40} = 5.5$$

The mean is a measure of the *central tendency* or *location* of the population.

Variance & Standard Deviation

Variance and standard deviation are measures of the *spread* of the distribution. The variance is the average squared difference between each value and the mean. The population variance is usually given the symbol σ^2.

$$\text{Variance: } \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

The standard deviation (SD) is the square root of the variance. The population standard deviation is usually given the symbol σ.

$$\text{Standard Deviation: } \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

We can work out the variance and standard deviation of the values on our set of forty cards.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{330}{40} = 8.25$$

$$\sigma = \sqrt{\frac{330}{40}} = 2.8723$$

Sample

Until now we have assumed that we can see all the cards at once. Now we are going to change the game. Imagine that someone else is holding the cards and allowing us to pick one at random, note its value and then replace it. Using this pick-and-replace process we can view a *sample* of the cards. This sample can be of any size as the cards are picked at random and replaced. Assume that the sample size is n.

The challenge is to estimate the mean and standard deviation of the original numbers on the cards based only on what we see in the sample. Here are the formulae that enable us to do this.

Estimate of Mean (based on sample):

$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

Estimate of Standard Deviation (based on sample):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}}$$

Convenient Formula for s

In the literature, s, the standard deviation of the underlying population estimated from a sample is called the *"sample standard deviation"*.

There is a convenient formula for calculating s.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}}$$

$$= \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}}$$

Standard Deviation from a Sample

Ten cards are selected individually from our special reduced pack of 40 cards (described on slide 147), noted and replaced in the pack. This gives a sample size $n = 10$. The values of the cards are:

10	3	4	3	5
4	1	5	8	5

We wish to estimate the mean m, and standard deviation s, of the values on all the cards, based only on knowledge of this sample.

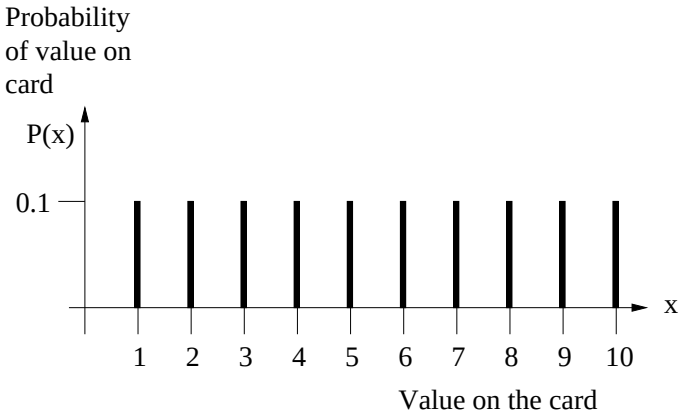
$$m = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{48}{10} = 4.8$$

$$s = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}}$$

$$= \sqrt{\frac{(290) - \frac{1}{10}(48)^2}{9}} = 2.5734$$

Discrete Probability Distribution

Consider picking a card (from the pack described on slide 147), noting its value and then replacing it in the pack. We can compute the probability of picking each of the possible values.



This is a probability distribution. In this case it is a *discrete* distribution because the cards can only carry certain integer values. Notice that the sum of all the histogram bars is $10 \times 0.1 = 1$. There are ten possible outcomes and they each have a probability of $1/10$. This is called a *uniform* distribution.

Mean and SD from the Distribution

The probability distribution is a property of the population of the numbers on the cards. Knowing the complete probability distribution enables us to calculate the mean μ and the standard deviation σ exactly.

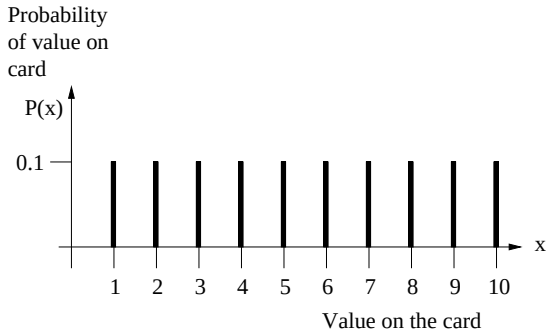
Let x_j represent each of the *different* values that are printed on the cards and M equal the number of these different values. In our example $M = 10$.

$$\text{Arithmetic mean: } \mu = \sum_{j=1}^{M} x_j P(x_j)$$

$$\text{Variance: } \sigma^2 = \sum_{j=1}^{M} (x_j - \mu)^2 P(x_j)$$

$$\text{Standard Deviation: } \sigma = \sqrt{\sum_{j=1}^{M} (x_j - \mu)^2 P(x_j)}$$

Example



We can see from the histogram that $P(x_j) = 0.1$ for all the values on the cards (i.e. for all j). In this particular case, the values x_j are the same numerically as the index j, so we can substitute $x_j = j$. Hence

$$\mu = \sum_{j=1}^{M} x_j P(x_j) = \sum_{j=1}^{10} j \times 0.1 = 5.5$$

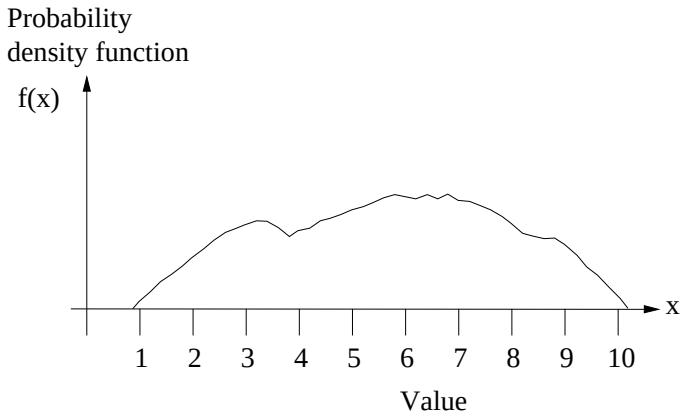
Now we use this value of μ in the formula for standard deviation.

$$\sigma = \sqrt{\sum_{j=1}^{M} (x_j - \mu)^2 P(x_j)}$$

$$= \sqrt{\sum_{j=1}^{10} (j - 5.5)^2 \times 0.1} = 2.8723$$

Continuous Probability Distribution

If you have an outcome that can take any real value (rather than a finite number of discrete values) this can be described by a probability density function (PDF).



Here the total area under the curve must be 1 and the probability of x taking a value in the range from (say) 6 to 7 is given by the integral (i.e. area) between 6 and 7. More generally:

$$\text{The probability of } (a < x < b) = \int_a^b f(x)\,dx$$

Examples of continuous random variables: the weight of a sample, the time for a physical process to complete, an output voltage.

Mean and SD from PDF

It is also possible to calculate the mean (μ) and standard deviation (σ) of a distribution from its probability density function.

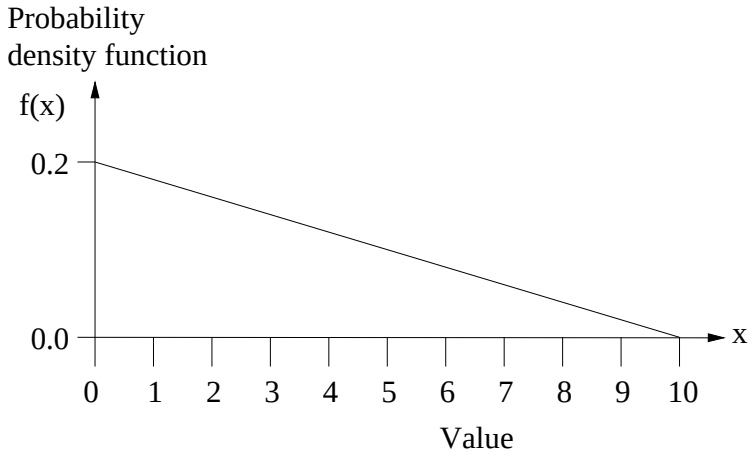
$$\mu = \int_{-\infty}^{+\infty} x f(x)\,dx$$

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx}$$

Knowing the probability density function enables us to calculate the mean μ and the standard deviation σ exactly.

Example using a PDF

Consider a machine than makes widgets which are supposed to be a particular length. Unfortunately, the machine often makes widgets that are slightly too long; it never makes widgets that are too short. The graph below shows the probability density function for the number of millimetres that a widget is too long.



Widget Example 1

1. What is the probability that a widget is less than 1 mm too long?

For $0 \leq x \leq 10$ we can see that $f(x) = 0.2 - 0.02x$, hence:

$$\mathrm{P}(0 < x < 1) = \int_0^1 f(x)\,dx$$

$$= \int_0^1 0.2 - 0.02x\,dx = 0.19$$

So the probability of a widget being less than 1 mm too long is 0.19.

2. Calculate the mean and standard deviation of the distribution of excess lengths.

$$\mu = \int_0^{10} x(0.2 - 0.02x)\,dx = 3.3333$$

$$\sigma = \sqrt{\int_0^{10} (x - 10/3)^2 (0.2 - 0.02x)\,dx} = 2.3570$$

Widget Example 2

3. What is the probability that a widget is produced with an excess length within one standard deviation from the mean excess length?

We wish to calculate the probability of an excess length in the range $3.33 - 2.36$ to $3.33 + 2.36$, which is given by:

$$\begin{aligned} \text{P}(x \text{ within one } \sigma \text{ of } \mu) &= \int_{0.9763}^{5.6904} 0.2 - 0.02x \, dx \\ &= 0.6285 \end{aligned}$$

4. The manufacturer wants to quote an excess length that he is sure 95% of the widgets produced will be shorter than. What should it be?

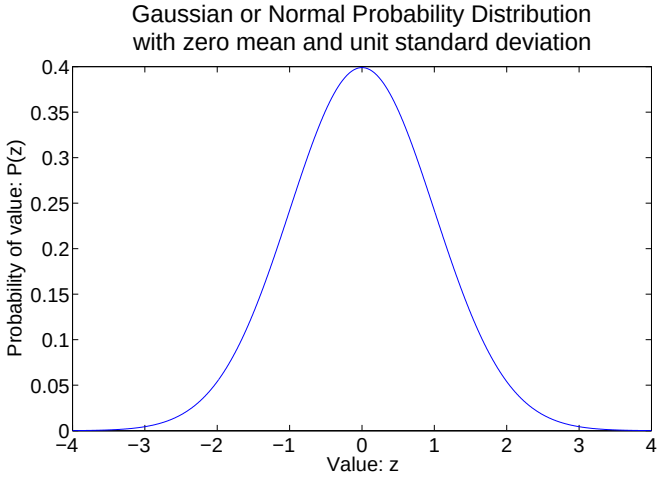
We need to solve for d in:

$$0.95 = \int_0^d 0.2 - 0.02x \, dx$$

$$\Rightarrow 0.95 = 0.2d - 0.01d^2$$

So $d = 10 - \sqrt{5} = 7.7639$ mm

Normal Probability Distribution



The Normal distribution is a symmetric distribution with two parameters, its mean μ and standard deviation σ.

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

If you add together a sufficient number (say 30), of independent random variables that are identically distributed, the sum will conform to the Normal distribution. This is called the “central limit theorem”.

Normal Distribution Example

Consider a laboratory experiment that results in a single real output x, each time that we perform it. In theory, it should produce the same output each time but in practice x varies slightly because of noise in the measurement system.

If we repeat the experiment at least 30 times and average the result ($\overline{x} = m = \sum_{i=1}^{n} x_i/n$) then we can say the following things about the way that $\overline{x}$ is distributed.

- Provided $n > 30$ it is reasonable to assume that $\overline{x}$ is Normally distributed.
- The standard deviation of $\overline{x}$ will be a factor of $\sqrt{n}$ less than the standard deviation of the original experimental data. Hence:

Estimate of standard deviation of $\overline{x}$:

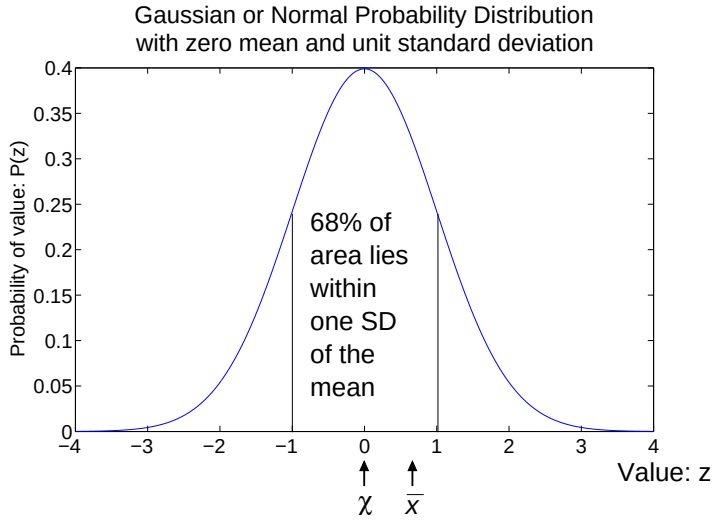
$$s_{(\overline{x})} = \frac{1}{\sqrt{n}} \times \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}}$$

Normal Distribution Example 2

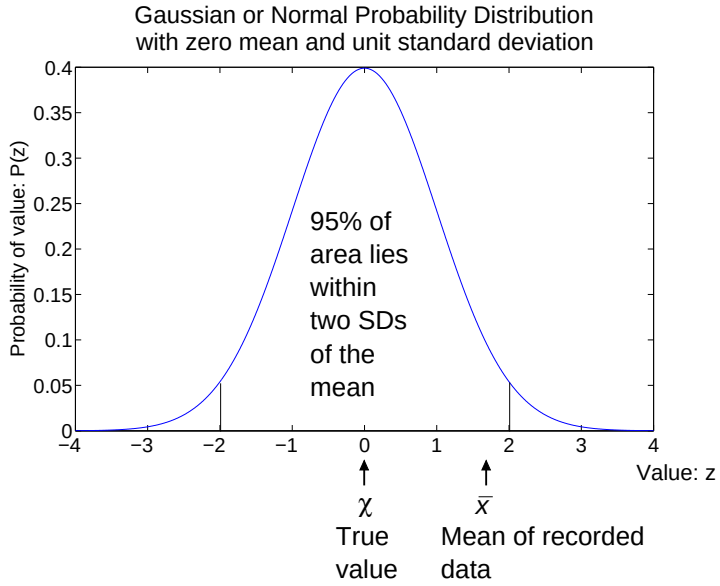
If it is fair to assume that the error in the original experimental data is unbiased, then the standard deviation of $\overline{x}$ gives us useful information about the error it is likely to contain.

Let $s_{(\overline{x})}$ be our estimate of the standard deviation of $\overline{x}$ and let $\mathcal{X}$ be the (unknown) true value of the thing we are intending to measure.

50%	of the time, $\overline{x}$ will lie within	$0.67 s_{(\overline{x})}$	of $\mathcal{X}$
68%	of the time, $\overline{x}$ will lie within	$s_{(\overline{x})}$	of $\mathcal{X}$
95%	of the time, $\overline{x}$ will lie within	$2 s_{(\overline{x})}$	of $\mathcal{X}$
99.73%	of the time, $\overline{x}$ will lie within	$3 s_{(\overline{x})}$	of $\mathcal{X}$



Normal Distribution Example 3



If we repeat an experiment 30 times and

- the average result $\overline{x} = 3.0279$,
- we calculate an estimate of the standard deviation of the experimental error as 0.2036
- then our estimate of the standard deviation of $\overline{x}$ will be $0.2036/\sqrt{30} = 0.0372$.

The true experimental result will have a 95% chance of lying within two standard deviations from the mean, i.e. in the range from 2.9536 to 3.1023.

Section 10: Summary

PDF	probability density function
SD	standard deviation
μ	mean of a distribution (or from a PDF)
σ^2	variance of a distribution (or from a PDF)
σ	SD of a distribution (or from a PDF)
$m = \overline{x}$	estimate of μ based on a sample of x values
s	estimate of σ based on a sample of x values
$s_{(\overline{x})}$	estimate of SD of $\overline{x}$ based on sample of x
	$s_{(\overline{x})} = s/\sqrt{n}$ where n is the sample size.

If we are prepared to do an experiment at least 30 times and believe our results to be unbiased, we can use the "central limit theorem" together with the shape of the Normal distribution to calculate a range within which the true result is likely to lie. We can make a statement of the form: "There is a probability of *blah* that the true result lies between *blah* and *blah*." This is called a confidence interval.

Acknowledgements

These notes are inspired by Prof Woodhouse's notes for the same course and incorporate several suggestions from Drs Gee and Treece. I am very grateful to Martin Weber and Naveed Ahmad for checking the algebra in the examples.