Probability Distribution Functions

Advanced Uncertainties, Part 1

Probability Distribution Functions

- A way to be precise about what we don't know!
- A rather difficult concept, but a crucial one for all of statistics and data analysis.
- But first - a reminder of what we've already done on uncertainty.

What do we already know?

- Look at the range of the data to estimate the uncertainty (maximum value and minimum value, in a repeated experiment).
- When comparing measurements with each other, or with theory, look at whether the ranges overlap.

Most common mistake

- Not to use them. To correctly get an answer with an uncertainty and then not do anything with it!
- For example: you are measuring g. You get a value of 9.4±0.2.
- The normal value of 9.8 is outside this range!
- So you should at least comment on this!

Really important for Jobs

- The biggest source of employment for Physics is not in physics.
- It is in data analysis and modelling (usually in the financial sector).
- Because Physicists are used to real data (unlike mathematicians) and can handle maths (unlike economists).

Really Dangerous

- The techniques and equations I'm going to give you only work in certain situations.
- This is frequently ignored.
- Ignoring this was a major cause of the Global Financial Crisis.
- So REMEMBER THE LIMITATIONS! NOT JUST THE EQUATIONS!
- (or tell me first so I can sell all my investments in your company...)

Describing anything that varies

- For example,your uncertain measurements of a quantity.
- Or anything that varies, like people's heights, stock market movements, diameter of craters on the Moon, student exam scores...

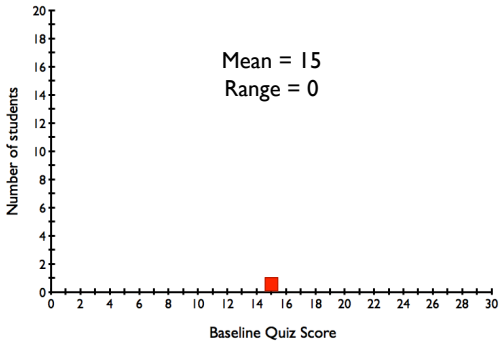
Crucial Concept

- Probability Distribution Function.
- Whenever you have anything that varies.
- Take all the measurements and create a HISTOGRAM.

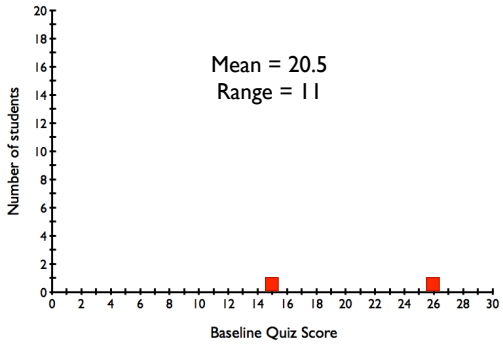
Example

- For example - here is something that varies - how well you did on the baseline quiz.
- Here are the first few students' scores:
- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27, 18, 29, 10, 14, 8, 26, 24, 14, 10, 17, 29, 19, ...

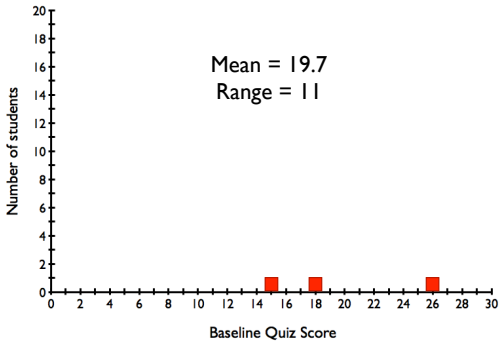
- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27 ...



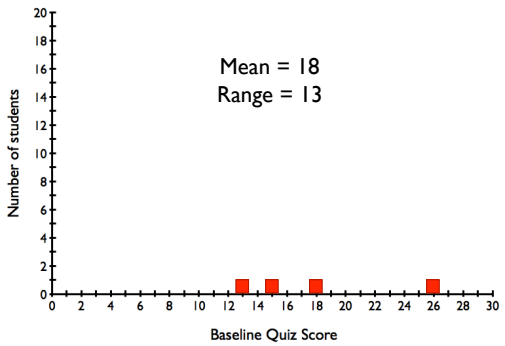
- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27...



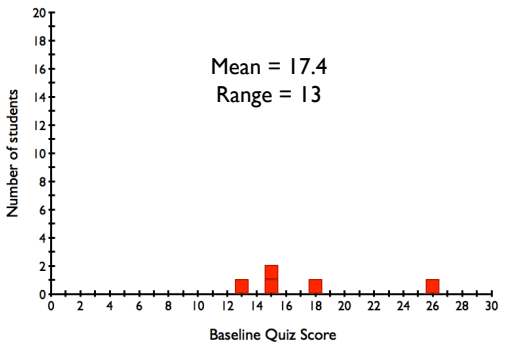
- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27...



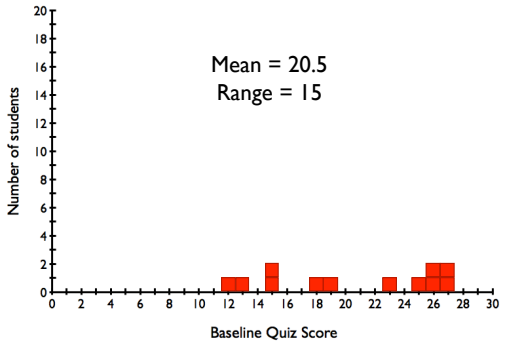
- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27...



- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27...

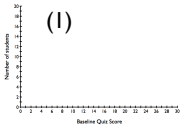


- 15, 26, 18, 13, 15, 19, 25, 27, 12, 23, 26, 27...

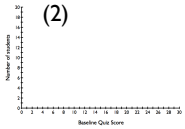


What would it look like when we have all the students listed?

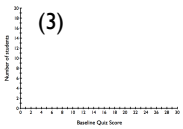
(1)



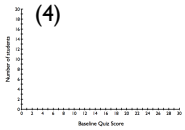
(2)

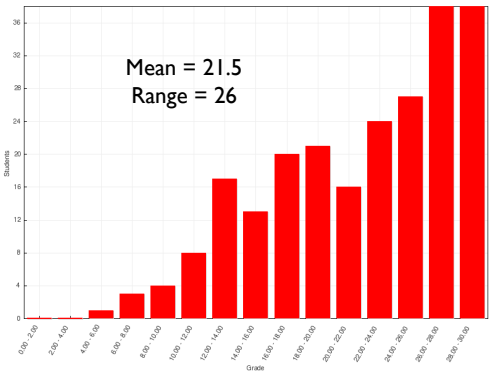


(3)



(4)





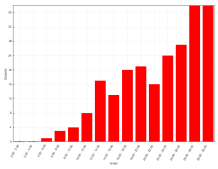
This is a Probability Distribution Function

- (PDF)
- A graph of the probability of getting a particular value.
- If you make a histogram of a sufficiently large sample, it will look increasingly like the probability distribution function.

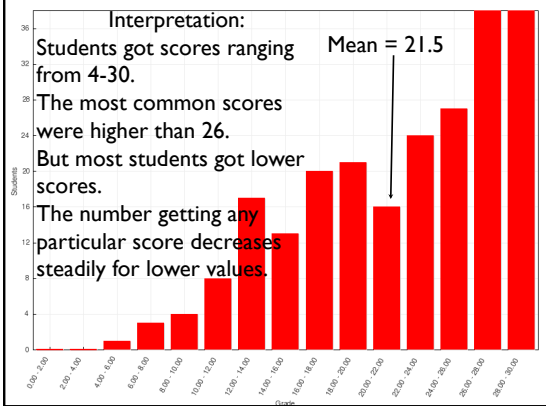
Not an x-y graph

- Note - a Probability Distribution function is not a graph of one quantity against another (like score against date).
- There is only one quantity here (in this case, score).
- You are plotting HOW FREQUENTLY different values of this quantity occur.

What's the interpretation of this graph?

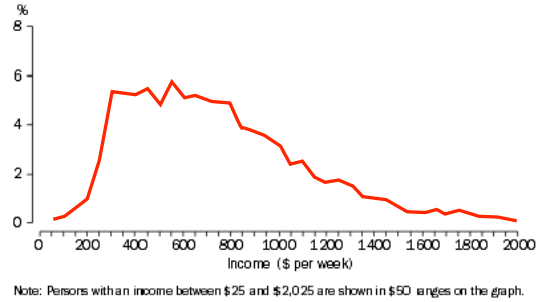


1. More intelligent students score better.
2. The more you practice the better you do.
3. Students with larger student numbers did better.
4. Most students did better than average
5. The most common scores were high ones



Personal Income distribution

- in Australia. What do you think it looks like?

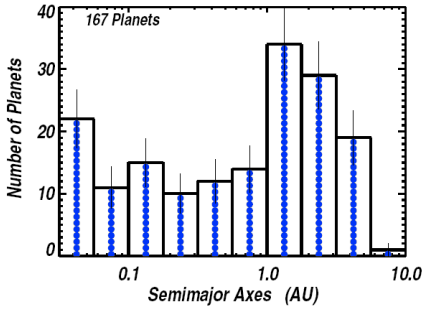


Extrasolar Planets

- There are a bunch really close in, then a smaller number at intermediate distances. Then a whole bunch around 1 AU out. Very few are currently known further out still, but that may just be due to the limitations of current detection techniques.

What does the PDF look like?

Exoplanet Orbit Radii

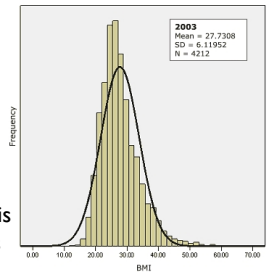


Body-Mass Index in Mississippi

Body Mass Index (BMI is a person's weight (in kg) divided by their height (in m) squared.

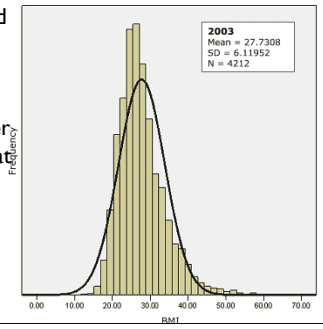
$$\mathrm{BMI} = \frac{\mathrm{Weight}}{(\mathrm{Height})^2}$$

18-25 is healthy, 25-30 is overweight and 30+ is obese



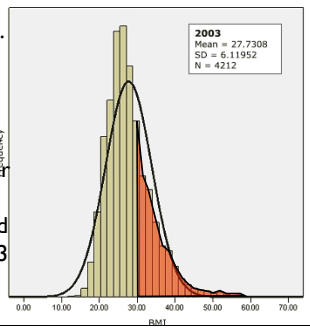
Interpreting this.

If you want to find out how many numbers lie in a particular range, find the area under the curve over that range.



Interpreting this.

e.g. BMI > 30 is classified as obese. How many obese people are there? These BMI values are not individually as common as lower ones, but collectively, they add up to more than 1/3 people.



S&P 500 Annual Returns (including dividends)

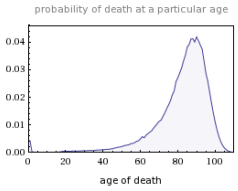
How often do you lose?

About one year in four.

How long will you live?

bottom 1%	24.1
bottom 5%	54.5
bottom 10%	64.5
bottom 25%	76.4
top 50%	84.9
top 25%	91.1
top 10%	95.5
top 5%	97.9
top 1%	102.0

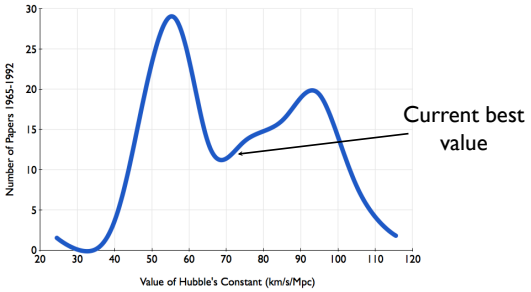
(based on death rates in 2007)



Experiment Uncertainties

- will have a probability distribution function.
- What does it look like? Could be anything, depending on what is causing the uncertainty.
- How can you find out? Do lots and lots of repeat experiments and plot a histogram.

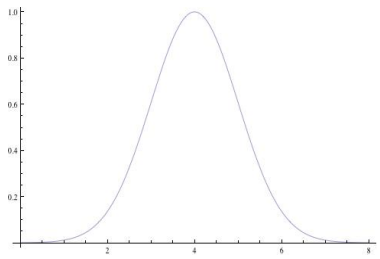
Expansion Rate of Universe



But now the dangerous bit...

- There is one particular probability distribution that is mathematically easy, and very popular with statisticians.
- It is the Gaussian Distribution.
- Also known as the Normal Distribution or the Bell Curve.

A Gaussian Distribution

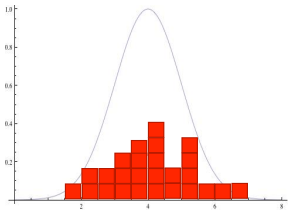


Equation

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

- where f(x) is the probability distribution function for x. μ is the mean value of x, and σ is the standard deviation.

Gaussian Numbers



Mostly close to the centre, occasional ones further out. The histogram won't look much like a Gaussian until you have lots of measurements.

4.04
4.28
5.09
5.55
3.76
2.19
2.93
3.24
6.11
4.44
3.67
...

Central limit theorem

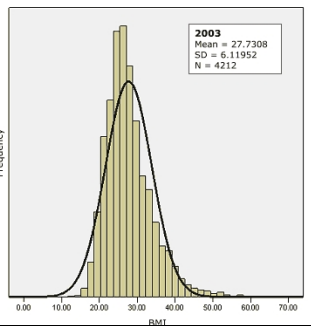
- says (paraphrasing) that if enough random stuff is going on, you will always get a Gaussian distribution.
- If your result depends on lots of random things, and
- and your final result is not dominated by only one or two of them,
- and the random things are not correlated with each other.

Does this ever actually work in practice?

- Not really.
- Lots of studies have shown that few real-world distributions are Gaussian.
- But many look fairly close (high in the middle, low at the edges and tailing away gently).

For example...

The Body Mass Index distribution. Compare in this plot to a Gaussian. Not too bad, but hardly a perfect fit.



But very widely used

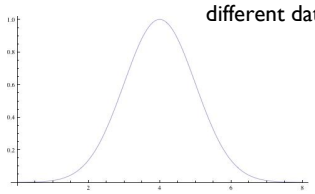
- An awful lot of statistics assumes that all probability distribution functions are Gaussian.
- Most of what's in your lab manuals about uncertainties assumes this.
- There is a growing group of people (myself included) who are on a crusade to get rid of Gaussians!
- But they are still the standard and so you have to know about them.

Global Financial Crisis

- The risk management models used by the big banks to manage their investments (for example, in Credit Default Swaps) were based on the assumption that investment returns and default rates had Gaussian distributions.
- The did until 2007, but then they didn't, and the banks lost more money than they'd made in the entire previous history of banking.

How to Measure a Gaussian

- Measure the centre by taking the Mean of the different data points.
- Measure the spread by calculating the standard deviation of the different data points.



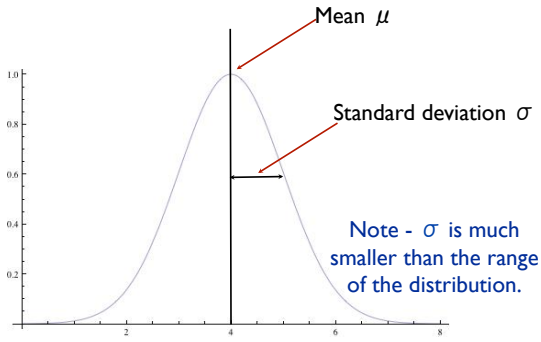
Standard deviation

- If your measurements (N of them) are x_1, x_2, x_3, ... x_n, then the standard deviation σ is given by

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

where μ is the mean.

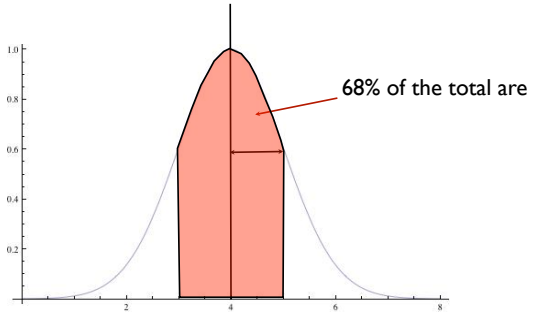
In this case, mean=4, standard deviation = 1



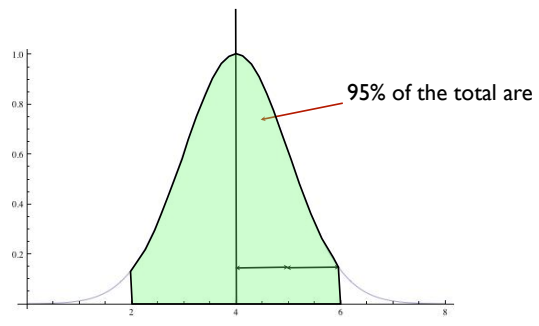
So you see that a lot of the time you are outside a standard deviation

- If the curve really is a Gaussian, then (on average) 68% of measurements should fall within one standard deviation, 95% within two and 99.7% within three.
- So if I say x = 45.3±0.2, I do NOT mean that x is always in the range 45.1 to 45.5.
- Typically it will be in this range about 68% of the time.
- It will lie between 44.9 and 45.7 95% of the time.

In this case, mean=4, standard deviation = 1



In this case, mean=4, standard deviation = 1



So comparing data with theory

- You should expect about 68% of the measurements to lie within the standard uncertainties.
- You should expect roughly equal numbers of observations to lie above and below the theory.
- And the ones that lie above or below should not all be at one end of the data.

ISO standard

- There is actually international agreement that the standard deviation of the distribution of measurements is used as a measure of the uncertainty.
- It is called the "Standard uncertainty"