

**Social Science Statistics Module I**  
**Gwilym Pryce**

**Lecture 1**  
**Density curves and the CLT**

Slides available from *Statistics & SPSS* page of [www.gpryce.com](http://www.gpryce.com)



## Reading:

- Pryce, G. (2005) *Inference and Statistics in SPSS* (wire comb binding)
  - Available from G. Pryce for £10 -- **all profits to charity** (also in library; borrow from previous students)
- Or (and):
  - Moore, D.S. and McCabe, G.P. *Introduction to the Practice of Statistics*, 4th Ed., San Francisco: Freeman, £39.99 from Amazon. (Avail in library; 5<sup>th</sup> Ed. £34.48 from Amazon; )
  - Andy Field *Discovering Statistics using SPSS for Windows*; 2<sup>nd</sup> Edition 2005, £23 from Amazon. 800+ pages

## Expectations & Support:

### 1. Independent learning:

- this is a PG course and a degree of independent learning is assumed.
- do the reading, attend labs, review the lectures, make use of the computer labs/online help in your own time.

### 2. Lab Overview & Feedback:

- Please feedback to the tutors & Class Reps how you think that is going, how it could be improved.
- Tutors and Class Reps will then report back to me how things are going each week.

### 3. Talk to tutors if you are struggling:

- Let the tutors know if you are struggling (assuming you have done the reading, attended labs etc.)
- Tutors cannot guarantee extra support, but it might be possible to arrange extra tutorials etc.

#### 4. Departmental Support:

- Struggling students should enquire whether their own dept has support to offer.
- All the grad school courses are only intended to constitute a generic training component;
- Individual depts & supervisors should supplement with additional training & support as necessary.

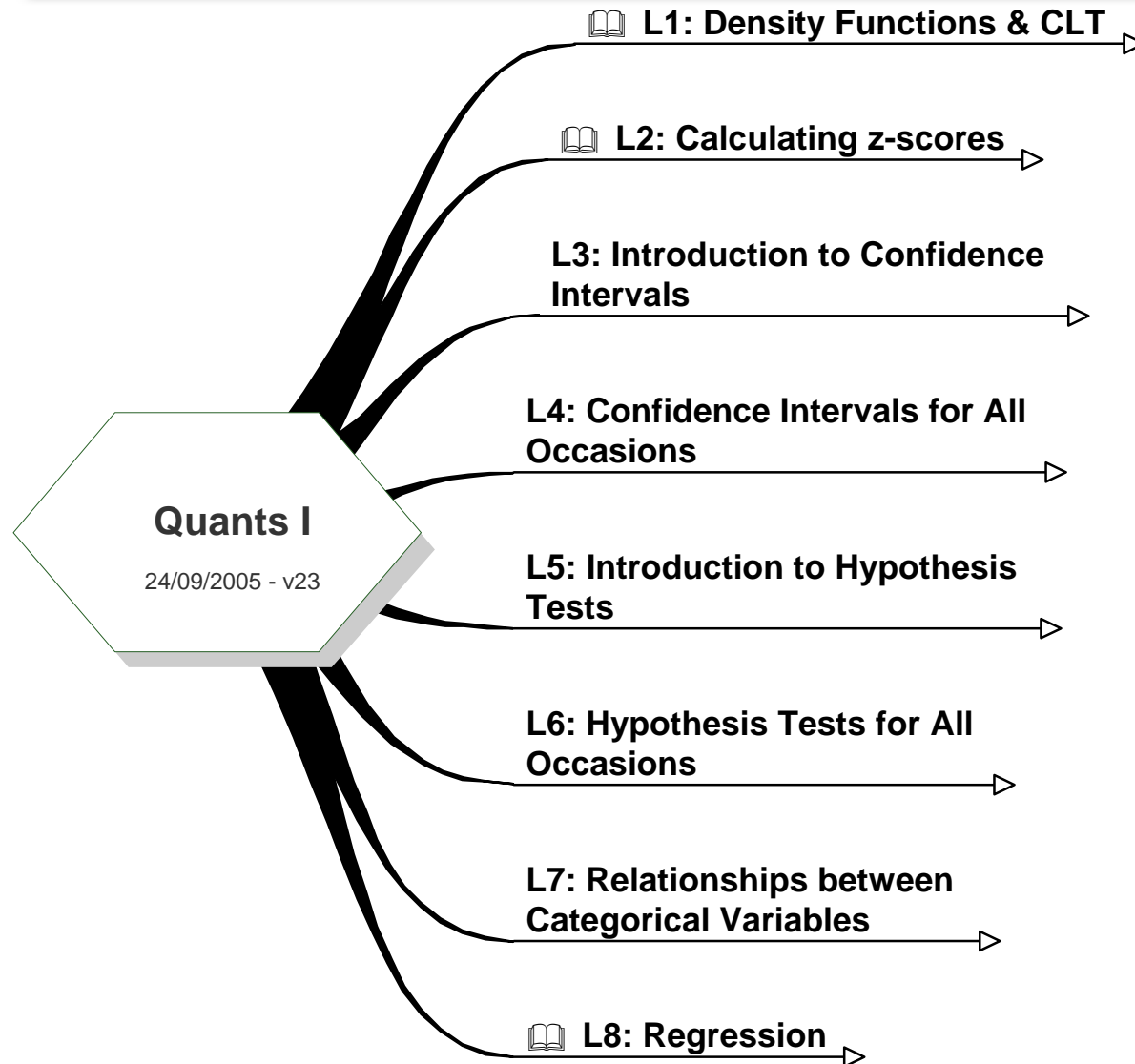
#### 5. Support from Maths Advisor Shazia Ahmed, University's Maths Adviser:

- If you have gone through steps 1 to 4, Shazia has agreed to run one-on-one sessions with students that are struggling with particular mathematical or statistical concepts (though she has made it clear that she cannot advise on SPSS problems, nor will she do the assignment for you).
- Students who have particular problems in this regard can contact her directly: **Shazia Ahmed**, Maths Adviser, Student Learning Service, McMillan Reading Room, Tel: 330 5631 Fax: 330 8063

#### 6. Tutor of Last Resort:

- Students who have gone through steps 1 to 5 above, and who still feel they are not receiving enough support, can email me directly
- I will try to arrange individual or small group meetings for people who have tried all other avenues.
  - You will need to demonstrate that you have gone through steps 1 to 5.

## Introduction & Overview:



## Aims & Objectives

- Aim
  - the aim of this lecture is to introduce the concepts that underpin statistical inference
- Objectives
  - by the end of this lecture students should be able to:
    - Understand what a density curve is
    - understand the principles that allow us to make inferences about the population from samples

## Plan

- 1. Measures of Central Tendency and Dispersion
- 2. Density curves & Symmetrical Distributions
- 3. Normal Distribution
- 4. Central Limit Theorem

## 1. Measures of Central Tendency and Dispersion

- How might you measure the typical value of a continuous variable such as income, IQ, age?
- How might you measure the variability of a continuous variable?

## Mean

- sum of values divided by no. of values:
  - e.g. mean of six numbers: 1, 3, 8, 7, 5, 3  
$$= (1 + 3 + 8 + 7 + 5 + 3) / 6 = \underline{4.5}$$
- Algebraic abbreviation:
  - abbreviation for sample mean is  $\bar{x}$
  - abbreviation for sum is capital sigma
  - abbreviation for any six observations (numbers) is  $x_1, x_2, x_3, x_4, x_5, x_6$
  - this can be abbreviated further as  $x_i = x_1, \dots, x_n$  where  $n = 6$ .
- Q/ How would we write the algebraic formula for the mean?

$$\text{mean} = \text{average} = \frac{1 + 3 + 8 + 7 + 5 + 3}{6}$$

$$= \frac{\sum x_i}{6} \text{ where, } x_i = 1, 3, 8, 7, 5, 3$$

$$\bar{x} = \frac{\sum x_i}{n}$$

- sample mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

- Population mean:

$$\mu = \frac{\sum X_i}{N}$$

- Q/ How can we measure the variability of a variable?

## Variance

- Based on the mean:
  - sum of all squared deviations from the mean divided by the number of observations
  - “average squared deviation from the average”
  - denoted by “ $s^2$ ”

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

- Q/ Why not simply take the average deviation?
  - I.e. why do we square the deviations first?

- A/ sum of deviations from mean always = 0
  - positive deviations cancel out negative deviations.
  - By squaring the deviations, negative values become positive, so we get a measure of the size of deviation of each value from the mean value.
- Q/ What's the main problem with the variance? How can we solve that problem?

## Standard Deviation

- Problem with the variance is that it's value is not in the same scale as the original variable.
  - Difficult to interpret.
- This problem is overcome by taking the square root of the variance:


$$\text{Standard Deviation} = s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$


- Degrees of freedom:
  - measure of the number of independent pieces of information on which the precision of a parameter estimate is based.
  - Calculated as the number of observations (values) minus the number of additional parameters estimated for that calculation.

Observation	Price per night of 3 star London hotels							
	$x_i$		$\bar{x}$		$x_i - \bar{x}$		$(x_i - \bar{x})^2$	
1	£ 67.20		£ 71.01		-£ 3.81		£ 14.50	
2	£ 70.49		£ 71.01		-£ 0.52		£ 0.27	
3	£ 78.26		£ 71.01		£ 7.25		£ 52.59	
4	£ 65.80		£ 71.01		-£ 5.21		£ 27.12	
5	£ 70.56		£ 71.01		-£ 0.45		£ 0.20	
6	£ 83.23		£ 71.01		£ 12.22		£ 149.38	
7	£ 80.01		£ 71.01		£ 9.00		£ 81.04	
8	£ 66.99		£ 71.01		-£ 4.02		£ 16.14	
9	£ 73.15		£ 71.01		£ 2.14		£ 4.59	
10	£ 54.39		£ 71.01		-£ 16.62		£ 276.16	
			Sum:		£ 0.00		£ 621.99	
			Sum / (n-1)		0		£ 69.11 (Variance)	
			$\sqrt{\text{sum} / (n-1)}$				8.31 (Standard Deviation)	

## Summary so far:

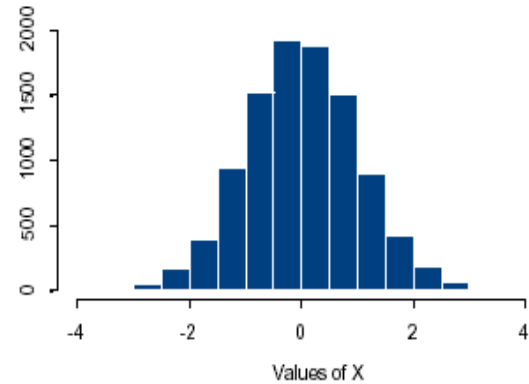
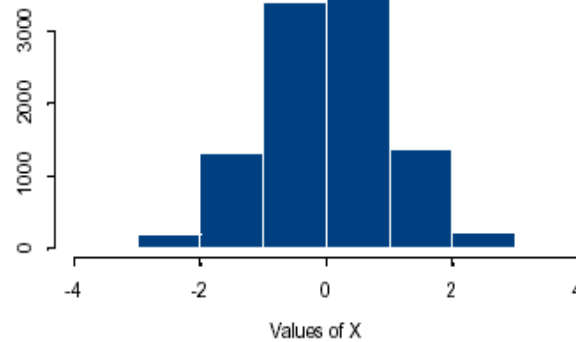
- 1. Measures of **Central Tendency**
- 2. Measures of **Spread**
  - range, **standard deviation**


$$\bar{x} = \frac{\sum x_i}{n}$$

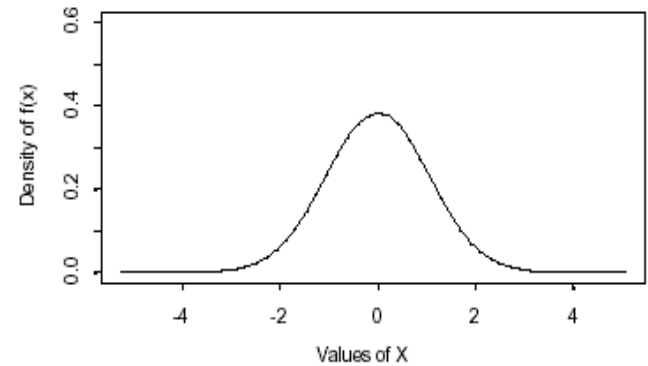
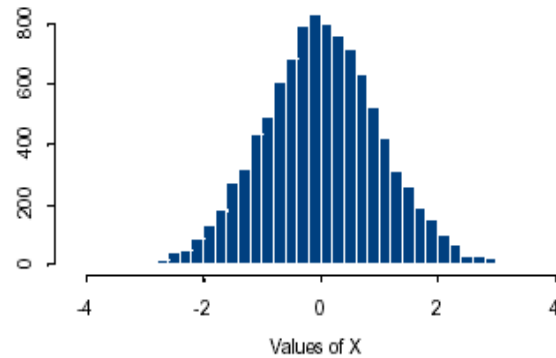

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

- But we would like a more complete way to represent the distribution of a variable...

## 2. Density curves: idealised histograms (rescaled so that area sums to one)



Standard deviation = 1, Mean = 0



## Properties of a density curve

- Vertical axis indicates relative frequency over values of the variable  $X$ 
  - Entire area under the curve is 1
  - The density curve can be described by an equation
  - Density curves for theoretical probability models have known properties

## Area under density curves:

- The area under a density curve that lies between two numbers = the proportion of the data that lies between these two numbers:
  - e.g. if area between two numbers  $x_1$  and  $x_2 = 0.6$ , then this means 60% of  $x_i$  lies between  $x_1$  and  $x_2$
  - when the density curve is symmetrical, we make use of the fact that areas under the curve will also be symmetrical

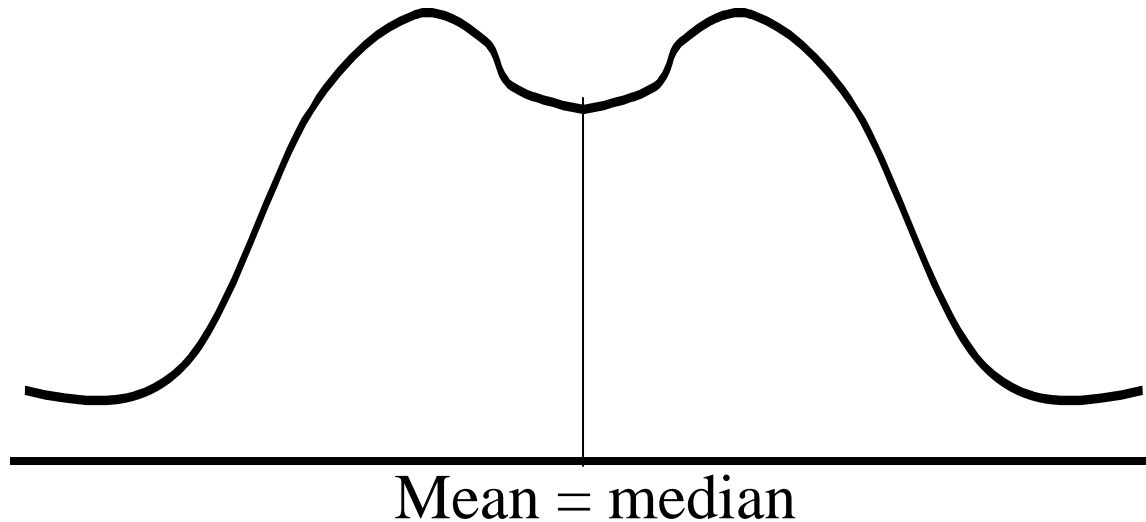
## Symmetrical Distributions

Mean = median

Areas of segments symmetrical

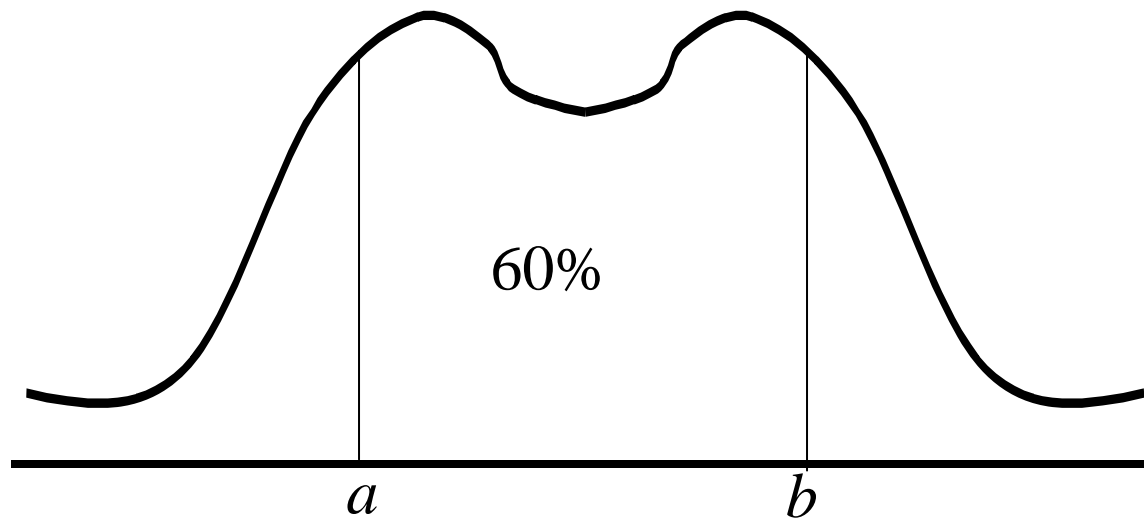
50% of sample  $<$  mean

50% of sample  $>$  mean

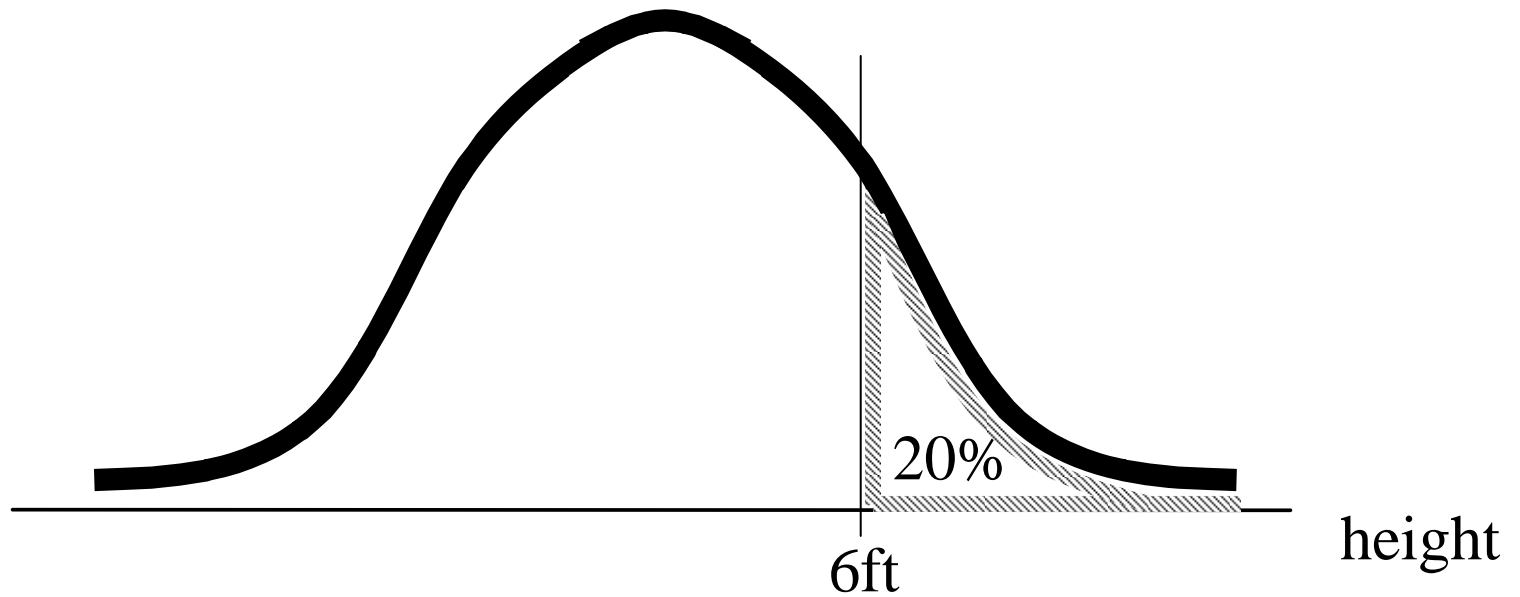


## Symmetrical Distributions

- If 60% of observations of variable  $x$  falls between  $a$  and  $b$ , what % of values of  $x$  are greater than  $b$ ?
- What's the probability of randomly choosing an observation with a value of  $x$  less than  $a$ ?

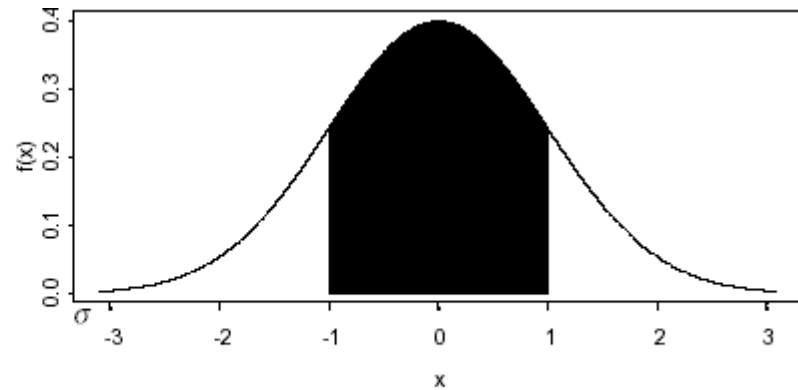


What's the probability of being less than 6ft tall?

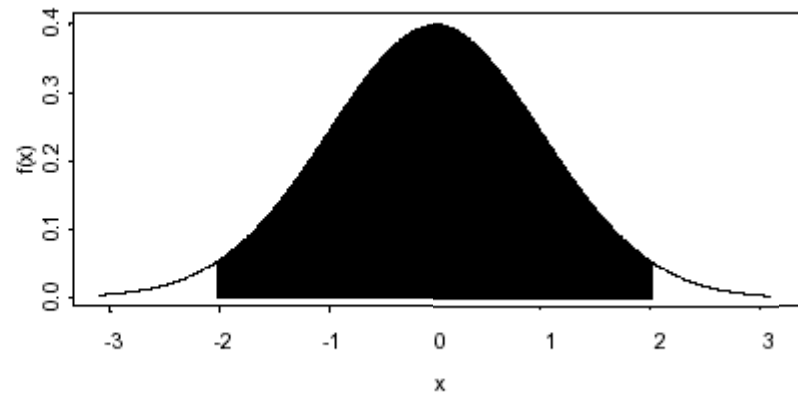


### 3. Normal distribution:

68 percent of the data in a normal distribution fall within one standard deviation from the mean

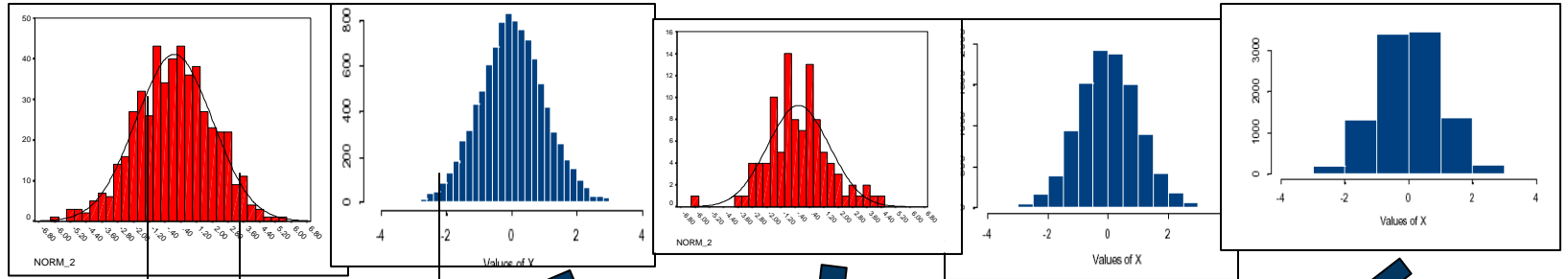


95 percent of the data in a normal distribution fall within two standard deviations from the mean



## Normal Curves are all related

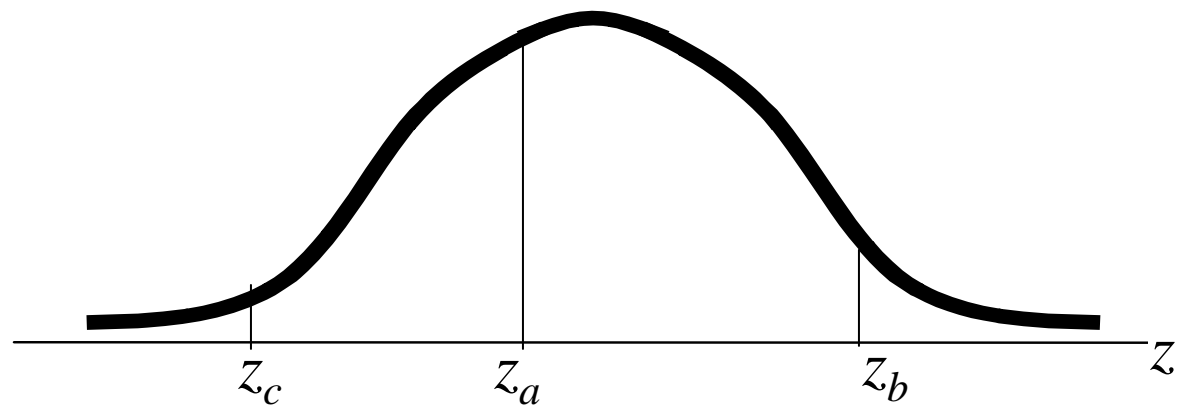
- Infinite number of poss. normal distributions
  - but they vary only by *mean* and *S.D.*
    - so they are all related -- just scaled versions of each other
- a baseline normal distribution has been invented:
  - called the **standard normal distribution**
  - has zero mean and one standard deviation



*a* *b*

*c*

Standardise



## Standard Normal Curve

- we can standardise any observation from a normal distribution
  - I.e. show where it fits on the standard normal distribution:
  - All we need is a simple conversion formula
    - A bit like converting lots of different currencies to a baseline common currency (e.g. the dollar)
  - This “currency conversion” uses a v. simple formula:
    - subtract the mean from each value and dividing the result by the standard deviation.
    - This is called the **z-score** = standardised value of any normally distributed observation.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where  $\mu$  = population mean

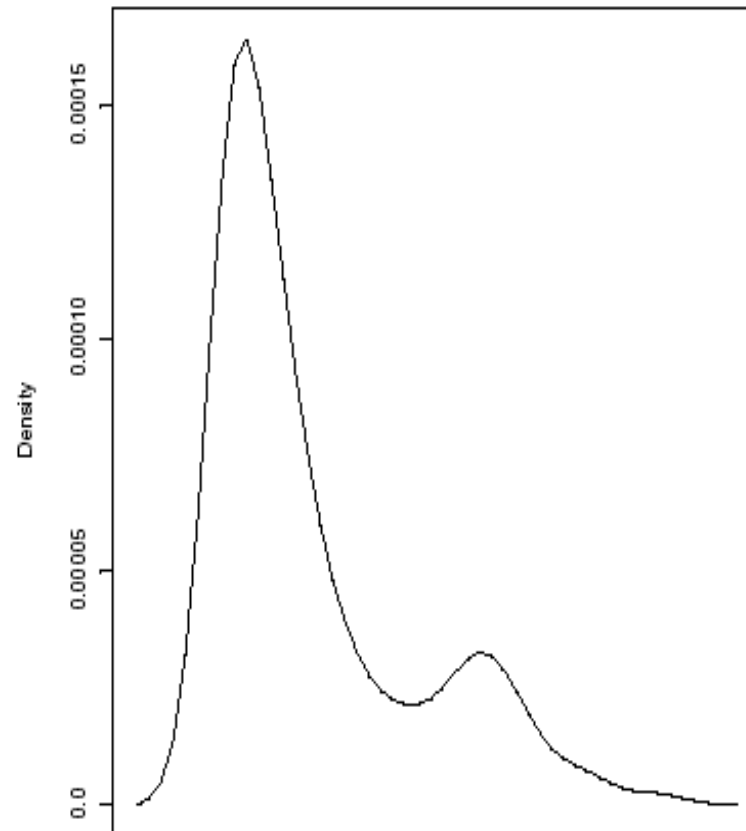
$\sigma$  = population S.D.

- Areas under the standard normal curve between different z-scores are equal to areas between corresponding values on any normal distribution
- Tables of areas have been calculated for each z-score,
  - so if you standardise your observation, you can find out the area above or below it.
- But we saw earlier that areas under density functions correspond to probabilities:
  - so if you standardise your observation, you can find out the probability of other observations lying above or below it.

## 4. Distribution of means from repeated samples

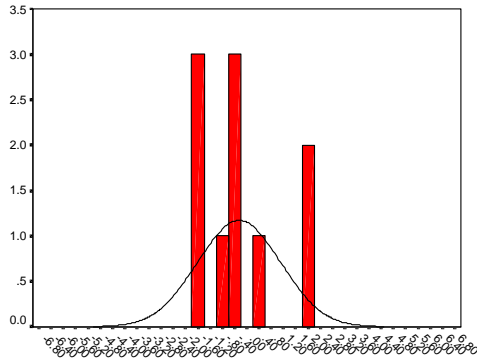
- We have looked at how to calculate the sample mean.
- What distribution of means do we get if we take repeated samples?
  - E.g. 1000 samples of 500 people. In each sample of 500 people you calculate the mean income. What would the histogram of the sample mean incomes look like?

E.g. Suppose the distribution of income in the population looks like this:

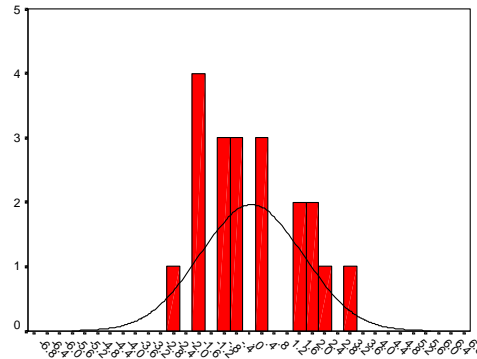


- Then suppose we ask a random sample of people what their income is.
  - This sample will probably have a similar distribution of income as the population
    - Positive skew: mean is “pulled-up” by the incomes of fat-cat, bourgeois capitalists.
    - Since the median is a “resistant measure”, the mean is greater than the median
- Then suppose we take a second sample, and then a third; and then compute the mean income of each sample:
  - Sample 1: mean income = £20,500
  - Sample 2: mean income = £18,006
  - Sample 3: mean income = £21,230

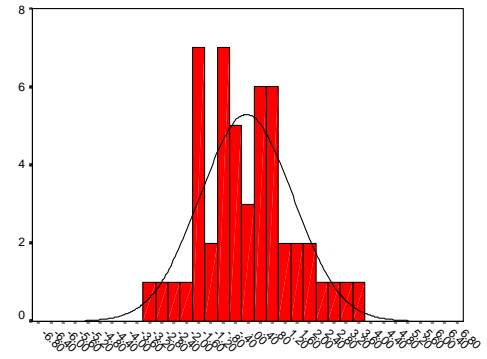
As more samples are taken, normal distribution of mean emerges



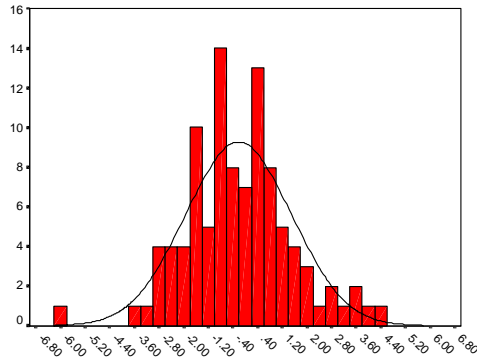
NORM\_2



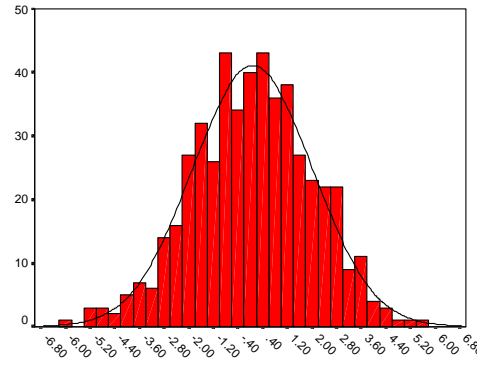
NORM\_2



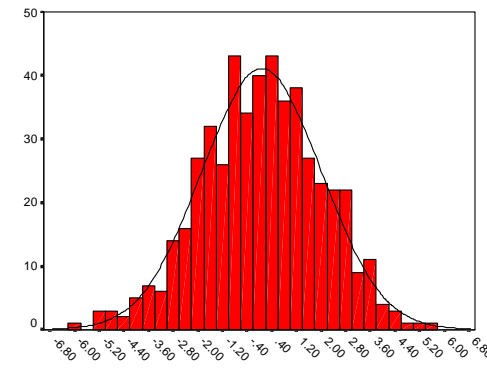
NORM\_2



NORM\_2



NORM\_2



NORM\_2

## Why the normal distribution is useful:

- Even if a variable is not normally distributed, its sampling distribution of means **will be** normally distributed, provided  $n$  is large (i.e.  $> 30$ )
  - I.e. some samples will have a mean that is way out of line from population mean, but most will be reasonably close.
  - **“Central Limit Theorem”**

- “The Central Limit Theorem is the fundamental sampling theorem. It is because of this theorem (and variations thereof), and not because of nature’s questionable tendency to normalcy, that the normal distribution plays such a key role in our work”

(Bradley & South)

- Why....?

## The standard error of the mean...

- When we are looking at the distribution of the sample mean, the standard deviation of this distribution is called the standard error of the mean
  - I.e. SE = standard deviation of the sampling distribution.
- but we don't usually know this
  - I.e. if we don't know the population mean (I.e. mean of all possible sample means), we are unlikely to know the standard error of sample means
- so what can we do?

## CLT: What about Proportions?

- What proportion of 10 catchers were female?
- What happens if I repeat the experiment?
  - What would the distribution of sample proportions look like?

## Summary

- 1. Measures of Central Tendency and Dispersion
  - Mean
    - Typical value
  - Standard deviation
    - average deviation from the mean
- 2. Density curves & Symmetrical Distributions
  - Idealised histogram, area under which = 1
- 3. Normal Distribution
  - Symmetrical with well known properties
- 4. Central Limit Theorem
  - If sample size is large, sampling distribution of the mean will be normal
  - Standard deviation of means from repeated samples is given a special name: Standard Error of the Mean

## Reading:

- **Pryce I&S in SPSS**
  - \*Pryce, Sections 1.3, 1.5, 2.4
  - \*Pryce, Section 2.6
  - \*Pryce, Section 2.5
  - Pryce, rest of Chapter 1.
- **M&M 4th Ed.**
  - Section 1.3; Chapter 5.

## Editing syntax files:

### 1. Start with an asterix:

- Use `*blah blah blah.` to put headings in syntax
  - anything after “`*`” is ignored by SPSS.
  - Important way of keeping your syntax files in order
  - e.g.

```
*Descriptive Statistics on Income.  
*-----.
```

### 2. Forward slash and an asterix:

- Use `/*blah blah blah */` to comment on lines
  - Anything between `/*` and `*/` is ignored by SPSS.
  - E.g.

```
COMPUTE z = x + y. /*Compute total income*/
```