THE UNIVERSITY OF
SYDNEY

SSPS6001: Quantitative Methods in the Social Sciences

Lecture 7:
Normal Distribution

Lecturer: Dr. Gina Anghelescu
Department of Political Economy
School of Social and Political Sciences, FASS
University of Sydney

The following slides are based on G. Argyrous 2011 “Statistics for Research” textbook

Review

- In the previous lectures we discussed methods for describing raw data; we have raw data from research, and we agreggate that data down into graphs, tables, and/or numerical measures depending on our purpose
- This process presumes we have the raw data in front of us (such as in an SPSS data file)
- Sometimes, however, we do not have such detailed data from which we can investigate the characteristics of a given distribution.
- We may only have the final calculations, such as the mean and standard deviation, generated from these data
- In such a situation we are sometimes able to ‘work backwards’ from these descriptions in order to determine the detailed frequency breakdown of the distribution, if that is what we require

Outline

- We will see that given the following bits of information about a distribution:
  - the mean
  - the standard deviation

and provided

  - we can assume that the shape of the distribution is normal

we can infer from these statistics detailed information about the frequency distribution in which we are interested.

Example

- I am interested in whether the students' grades have improved since 1990
- To answer this broad question I decide I need to generate the following details about the frequency distribution of scores in 1990 and in 2005:
    - What percentage of students received a 'decent' grade, defined as between 60-65?
    - What percentage of students did very well, defined as a grade in excess of 65?
    - What percentage of students failed (i.e. less than 50)?
    - What range of scores did the middle 50% of students receive (i.e. what is the *IQR*)?

Summary data

- I have original data for the 2005 cohort of first year students. These are the individual grades for each student
- With these raw data I can use the techniques we have learnt in previous lectures to describe the distribution in various ways
- Given my specific research questions, I generate the following statistics:

Exam results for first year students, 2009

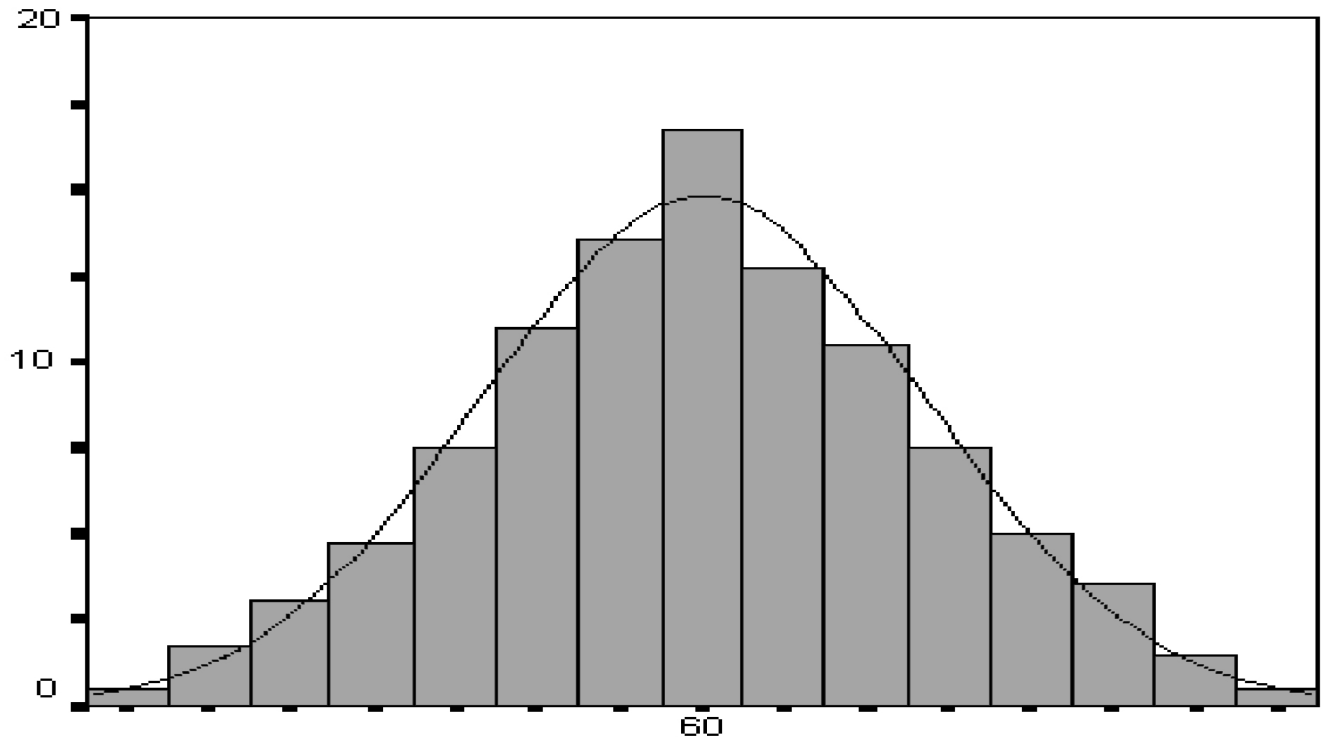
Students failing	14%
Students between 60-65	23%
Students higher than 65	34%
IQR	12 marks

Summary data cont.

- Unfortunately, I do not have the raw data for first year students in 1990
- The only information I have are the results of someone else's analysis, and the only descriptive statistics they have generated are that:
  - the mean grade for 1990 was 60
  - the standard deviation of grades in 1990 was 10
- It may seem that I can't compare the 2 distributions because I do not have the individual exam scores for 1990 that will allow me to calculate the percentage that failed, etc.
- We shall see, however, that knowledge of the mean and the standard deviation allows us to 'tease out' other aspects of a distribution that are not directly provided to us.

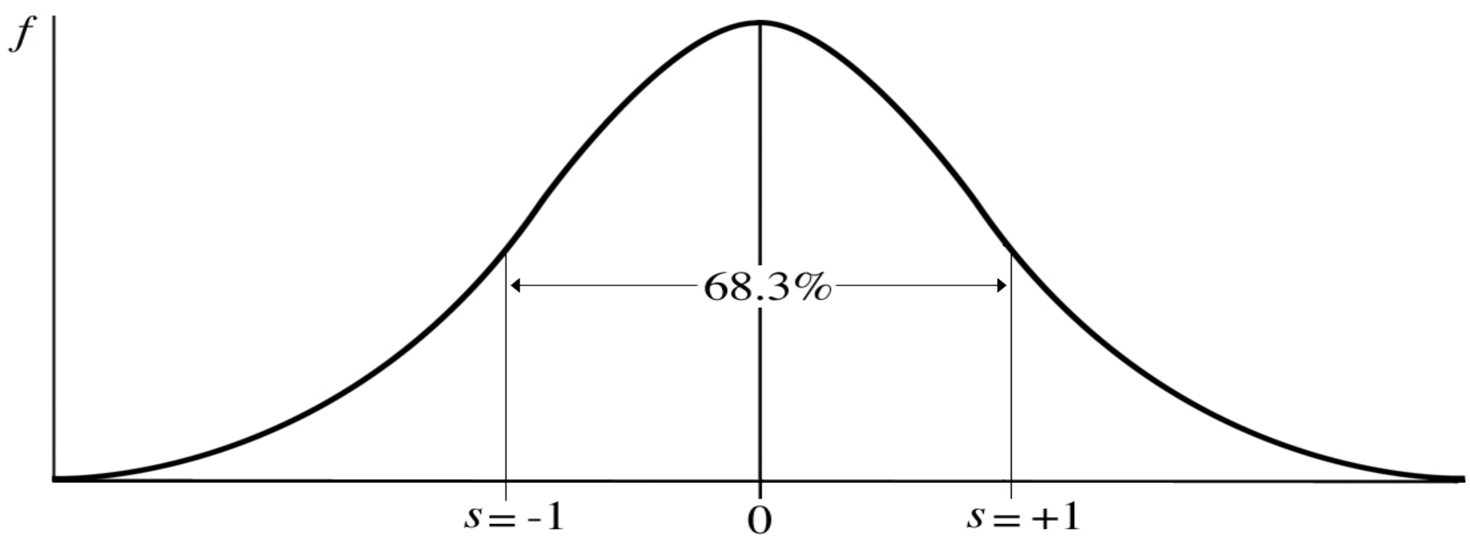
The assumption of approximate normality

- To be able to make deductions based on only 2 bits of information about a distribution (the mean and the standard deviation) I need to make an assumption.
- I need to assume that the shape of the distribution is approximately normal
- In other words I need to assume that if I graphed the exam scores for the 1990 students, it would look something like:



The normal curve

- Being able to assume normality is very useful because the properties of the normal curve are very well known
- In particular the normal curve is:
  - smooth,
  - unimodal
  - perfectly symmetrical.
  - it has 68.3% of the area under the curve within one standard deviation of the mean.



Determining the frequency of cases around the mean

- We can already make some conclusions about the 1990 distribution of marks, based on this assumption of normality
- We can conclude that approximately 68.3% of the 1990 students had grades somewhere between:

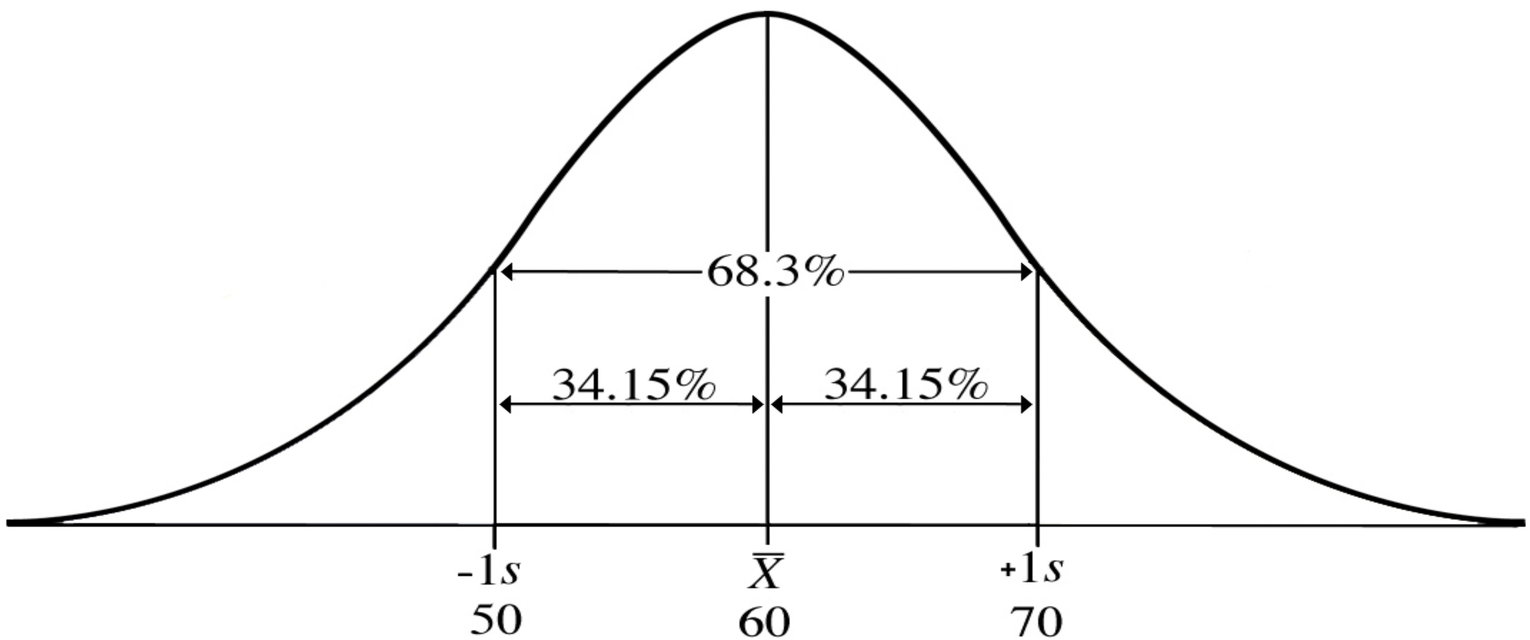
  - 60 – 10 = 50

  and

  - 60 + 10 = 70

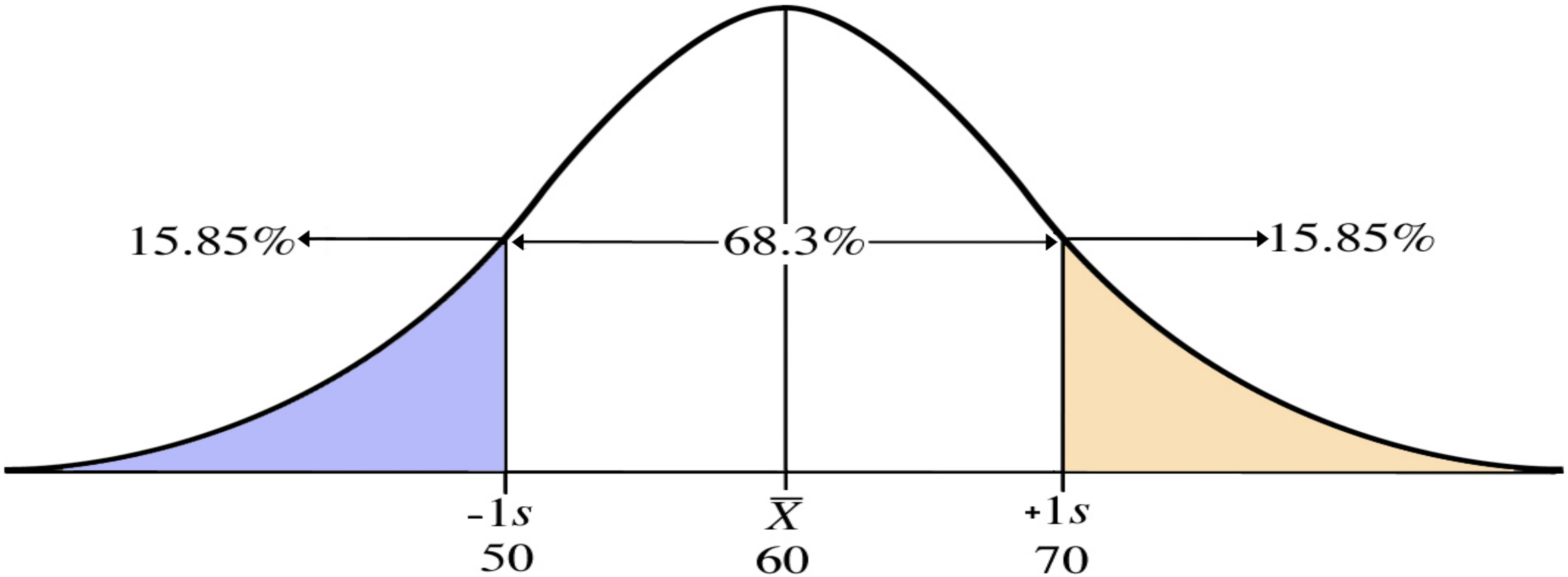
Frequency of cases between the mean and a point on the scale

- Since the curve is perfectly symmetrical we can take the logic further and deduce that:
  - 34.15% of students (half of 68.3%) have grades within one standard deviation above the mean (i.e. range from 60 to 70)
  - 34.15% of students (half of 68.3%) have grades within one standard deviation below the mean (i.e. range from 50 to 60)



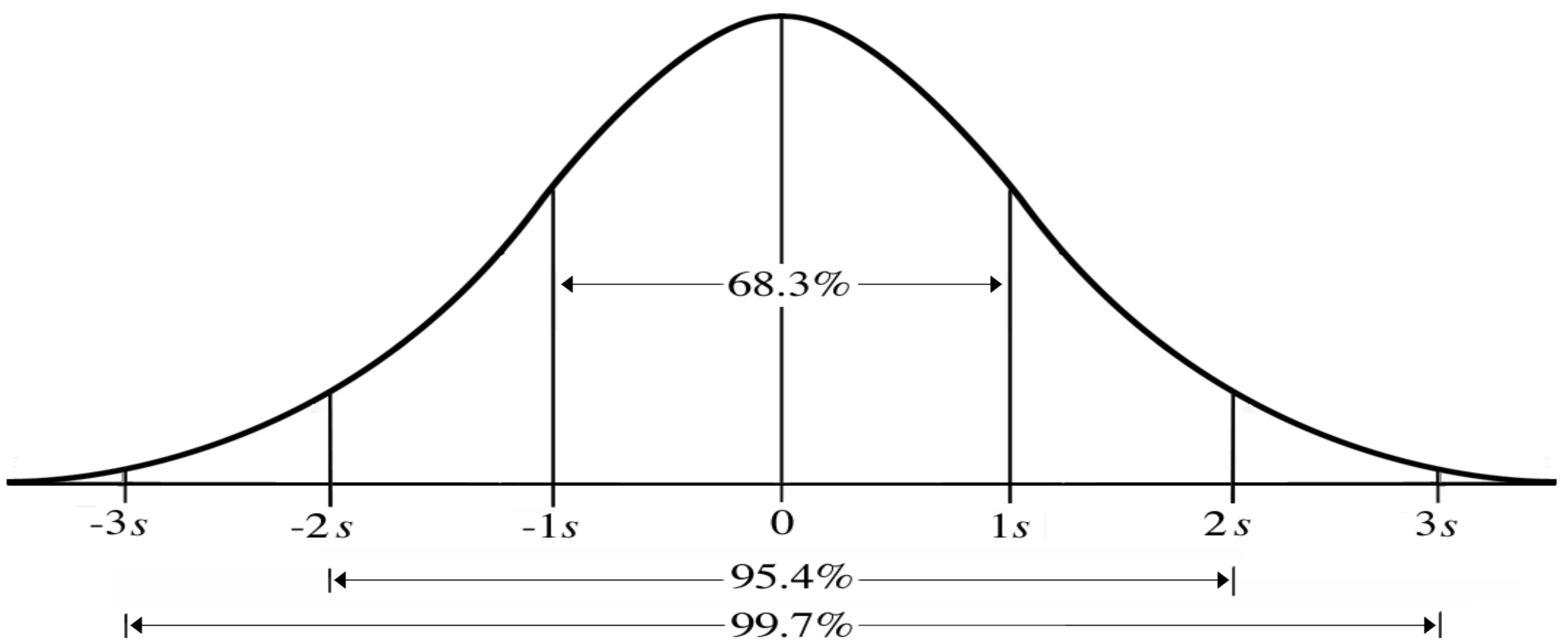
Determining the frequency of cases beyond points on the scale

- Since the whole area under the curve is 100%, we can take the logic still further and deduce that:
  - 15.85% of students have grades above one standard deviation from the mean (i.e. above 70)
  - 15.85% of students have grades below one standard deviation from the mean (i.e. below 50)



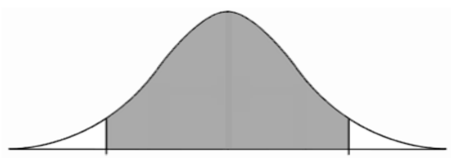
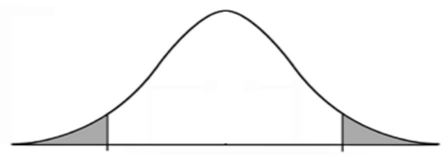
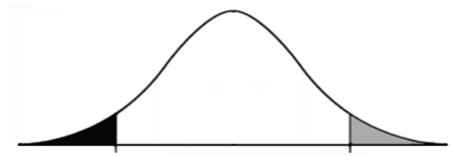
The normal curve: extending the definition

- The properties of the normal curve are actually more well-defined than this
  - Between ±1 standard deviations from the mean of a normal distribution lies 68.3 per cent of the area under the curve
  - Between ±2 standard deviations from the mean of a normal distribution lies 95.4 per cent of the area under the curve
  - Between ±3 standard deviations from the mean of a normal distribution lies 99.7 per cent of the area under the curve



The normal table: Simple version

- We can summarize this information in a table that allows us to look up the relevant areas of the curve

Standard deviations from the mean (z-scores)	Area under curve **between both** points	Area under curve **beyond both** points (two tails)	Area under curve **beyond one** point (one tail)
			
± 1	0.683	0.317	0.1585
± 2	0.954	0.046	0.0230
± 3	0.997	0.003	0.0015

- For example, if I wanted to know the percentage of cases that had a grade more than 2 standard deviations above the mean (exam grades of more than 80) I refer to the last column and the second row for ±2

- This indicates that 0.023 (2.3%) of students had grades of more than 80

The standard normal curve

- Statisticians have elaborated the definition of the normal curve and calculated the relative frequency for any range of scores that are normally distributed
- These relative frequencies, expressed as proportions, are presented in a table that is found in every statistics text
- This table is usually called *The Areas Under the Standard Normal Curve*

z-scores

- When we use the table for the areas under the standard normal curve, we don't work with original scores such as grades, but rather z-scores (number of standard deviations from the mean)
- For example, with exam grades a score of 75 is 15 marks above the mean.
- The standard deviation is 10. Therefore 15 marks represents 1.5 standard deviations above the mean, which we abbreviate to z = +1.5
- Similarly, an exam grade of 55 is –5 marks below the mean, which is –0.5 z-scores

Calculating z-scores

- We can convert any score measured in original units into a z-score by using the following formulas, where:
  - X_i is the actual value measured in original units
  - μ is the mean of the population
  - σ is the standard deviation of the population
  - X is the mean of the sample
  - s is the standard deviation of the sample

$$z = \frac{X_i - \overline{X}}{s} \quad \text{(sampl}$$

$$Z = \frac{X_i - \mu}{\sigma} \quad \text{(populatic}$$

The frequency between the mean and a point on the scale

- What percentage of 1990 students received a ‘decent’ grade, defined as between 60-65?
- The question requires us to find the frequency of cases between the mean and another score on the scale
- To answer this question we firstly convert the 65 into a *z*-score:

$$z = \frac{X_i - \bar{X}}{s} = \frac{65 - 60}{10} = 0.5$$

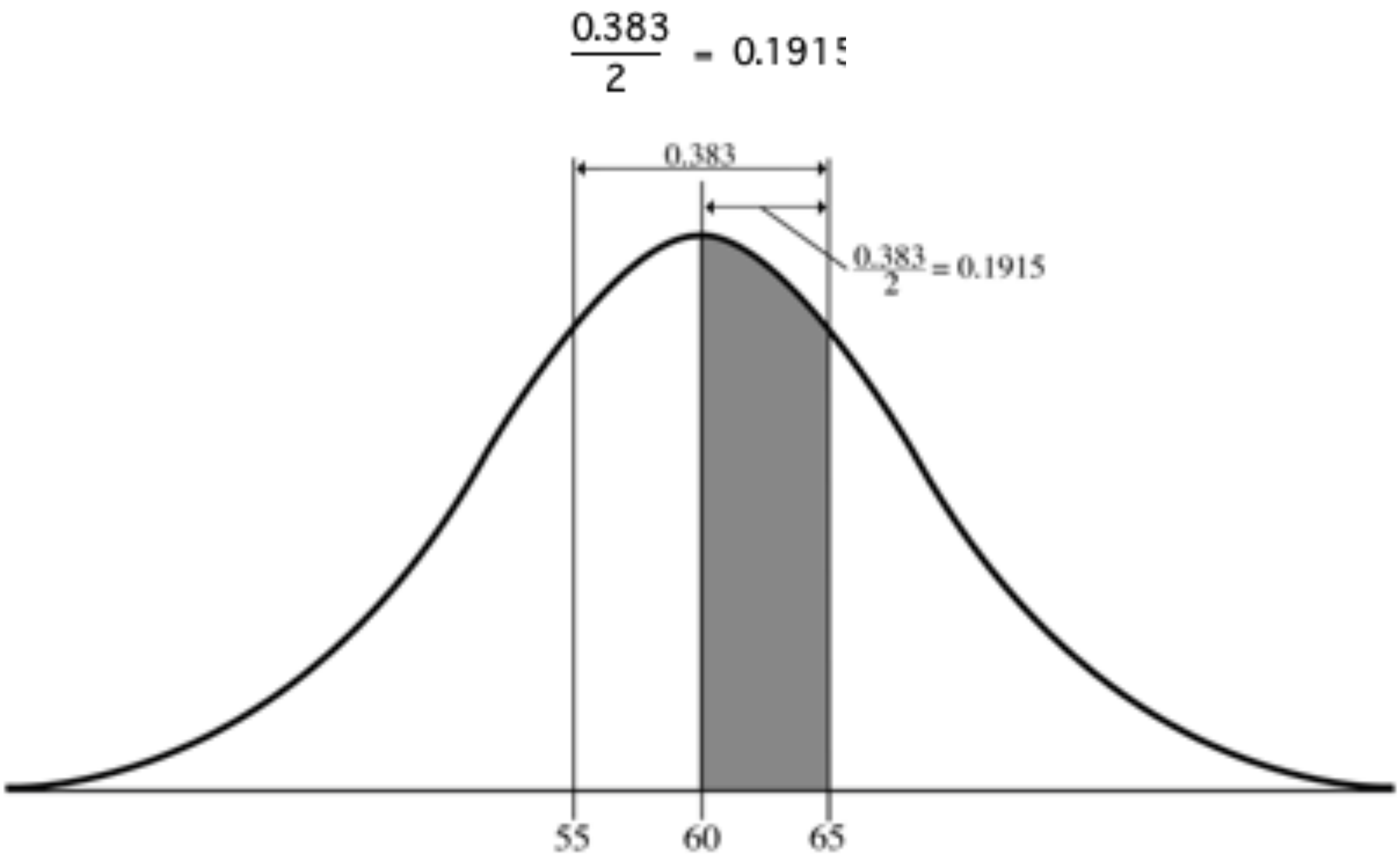
Using the normal table

- We then refer to the table and read off the relative frequency (i.e. the proportion) associated with that *z*-score
- In other words, 0.383 (38%) of all cases will have a grade of 5 marks *above or below* the mean

Standard deviations from the mean (*z*-scores)	Area under curve **between both points**	Area under curve **beyond both points** (two tails)	Area under curve **beyond one point** (one tail)
± 0.1	0.080	0.920	0. 4600
± 0.2	0.159	0.841	0.4205
± 0.3	0.236	0.764	0.3820
± 0.4	0.311	0.689	0.3445
± 0.5	**0.383**	0.617	0.3085
± 0.6	0.451	0.549	0.2745
± 0.7	0.516	0.484	0.2420
± 0.8	0.576	0.424	0.2120
± 0.9	0.632	0.368	0.1840
± 1	0.683	0.317	0.1585
…	…	…	…
± 3	0.997	0.003	0.0015

- Since we are interested in only those students that are 5 marks *above* the mean, we divide 0.383 in half

- In other words, 19.15% of students received such grades



Using z-scores to determine the frequency beyond a point

- What percentage of students did very well, defined as a grade in excess of 65?
- This question requires us to determine the frequency *beyond* a point on the distribution
- We therefore refer to the last column in the *Table for the areas under the standard normal curve* (next slide)
- From the previous example we know the z-score associated with a grade of 65 is 0.5
- This indicates that 0.3085 (30.85%) of students scored over 65

$$z = \frac{X_i - \bar{X}}{s} = \frac{65-60}{10} = 0.5$$

Using the normal table

Standard deviations from the mean (*z*-scores)	Area under curve **between both points**	Area under curve **beyond both points** (two tails)	Area under curve **beyond one point** (one tail)
±0.1	0.080	0.920	0. 4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	**0.3085**
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
…	…	…	…
±3	0.997	0.003	0.0015

Using z-scores to determine the frequency beyond a point

- What percentage of students failed (i.e. received a grade of less than 50)?
- We firstly determine the z-score for 50:

$$z = \frac{X_i - \bar{X}}{s} = \frac{50 - 60}{10} = -1$$

- The question requires us to determine the area *beyond* a point on the distribution
- We therefore refer to the last column for the *Table for the area under the standard normal curve* (next slide)
- This shows that there is 0.1586 of the curve beyond a z-score of –1, which indicates that nearly 16% of students failed

Using the normal table

Standard deviations from the mean (*z*-scores)	Area under curve **between both points**	Area under curve **beyond both points** (two tails)	Area under curve **beyond one point** (one tail)
±0.1	0.080	0.920	0. 4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	**0.1585**
…	…	…	…
±3	0.997	0.003	0.0015

The points on a scale associated with a particular frequency

- In the previous examples, we had a particular range of exam grades and we wanted to work out the frequency of students in that grade range
- Sometimes the problem is slightly different; we might have a particular section of the frequency distribution in mind, and want to determine what the range of grades for that section is
- For example, we might be interested in the grades that define the top 10% of students
- Similarly, we might be interested in the range of grades that defines the middle 50% of students (i.e. the *IQR*)
- Since we already know the relative frequency of cases we are interested in (the middle 50%) we look down the column for the **Area under curve between both points** and find the cell that has a probability of 0.5 (or the closest to it)

Using the normal table

Standard deviations from the mean (*z*-scores)	Area under curve **between both points**	Area under curve **beyond both points** (two tails)	Area under curve **beyond one point** (one tail)
±0.1	0.080	0.920	0. 4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	**0.516**	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
…			
±3	0.997	0.003	0.0015

- The closest value to 0.5 is 0.516, which is associated with *z*-scores of + 0.7 and – 0.7

- Since a *z*-score is the number of standard deviations a particular value is above/below the mean, a *z*-score of 0.7 is 0.7 standard deviations away from the mean

- The standard deviation for these exam scores is 10, so that a *z*-score of 0.7 is 7 marks away from the mean of 60 i.e. 0.7 ▫□10 = 7

- Thus *z*-scores of –0.7 and +0.7 are associated with grades of:
  - 60 – 7 = 53
  - 60 + 7 = 67

- To convert any *z*-scores into the actual units in which a given variable is measured, we use the formula:

$$X_i = \overline{X} \pm z(s$$

Limitations

- In all these examples we simply assumed that the distribution of 1990 exam scores was normal
- If we did actually get our hands on the raw data and graphed them we might find that the distribution is not exactly normal: it may be close but not exactly the same as the normal curve
- This should be expected: very few variables will be exactly normal
- This means that our calculations of frequencies may not be exactly true; for example we calculated that 16% of students failed, assuming that the distribution is normal
- If we were able to tally up the actual exam scores we may find that 18% of students actually failed
- This is because the assumption of normality may not be perfectly correct; the distribution of exam grades may only be approximately normal
- Our calculation of the percentage of students failing therefore is also only approximately correct
- Therefore whenever we derive frequencies on the assumption of normality we use such as "the percentage of students failing is approximately 16%"