NORMAL DISTRIBUTION AND NORMAL STANDARDIZED DISTRIBUTION.

Week 5

!!!

- Mean, median, and mode measure the central tendency of a variable.
- Measures of dispersion include variance, standard deviation, range, and interquartile range (IQR).
- We can draw a histogram, a stem-and-leaf plot, or a box plot to see how a variable is distributed.

Interval/cardinal/continous variables

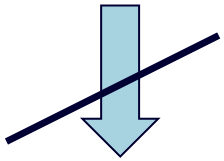
- We run various statistical tests to check to what extent our data corresponds to a certain model.
- To do it… we need normally distributed variables.
- Normal distribution ⇔ bell curve shape (Frederich Gausse 18.-19. century).

Normal distribution

- It is typical for a large number of biological or physical phenomena.
- It can also characterize some social phenomena.

COMMON ASSUMPTION

<u>A RANDOM VARIABLE IS NORMALLY DISTRIBUTED!!!</u>



<u>INTERPRETATION AND INFERENCE MAY NOT BE RELIABLE OR VALID</u>

Figure 1. Normal Distribution Curve and its basic characteristics (σ)

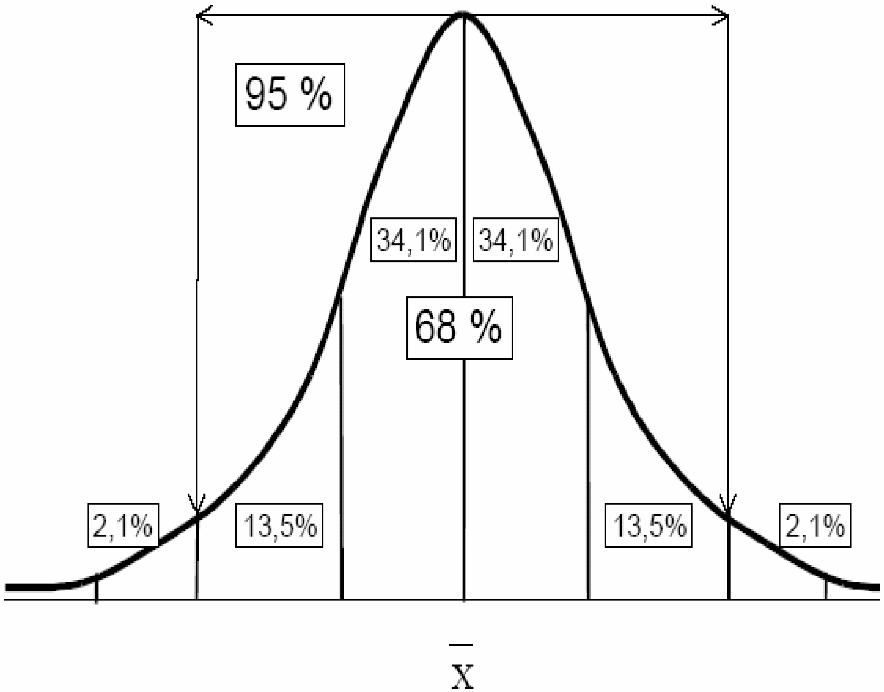
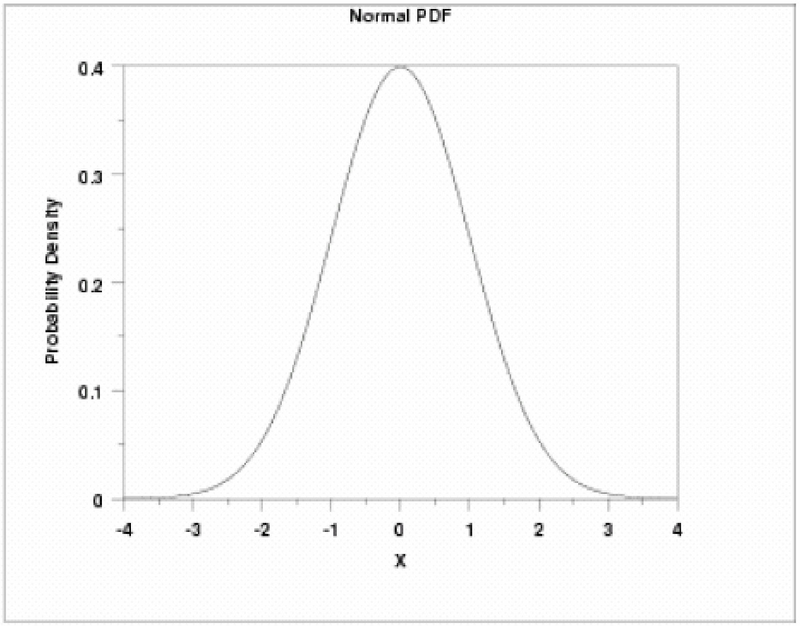


Figure 2. Normal standardized distribution



Why is important for statistical analysis?

- Majority of values are found around the average and are symmetrically distributed ⇨ average = median = mode
- It has one peak only.
- We can calculate the percentage of certain values found within a certain interval around the average.
- It is just a model and instrument of help. It is a mathematical ideal.
- If we find that our variables are very close to be normally distributed, than we are lucky ☺

PARAMETRIC DATA ⇨

- Normally distributed data – it is assumed that data are from a normally distributed population.
- Homogeneity of variance – the variance should not change systematically throughout the data.
- Interval data – it should be measured at least at the interval level.
- Independence – data from different subjects are independent.

How to tell if a distribution is normal?

STEP 1 - Run a histogram with a normal curve and see if your variable is normally distributed.

ANALYZE

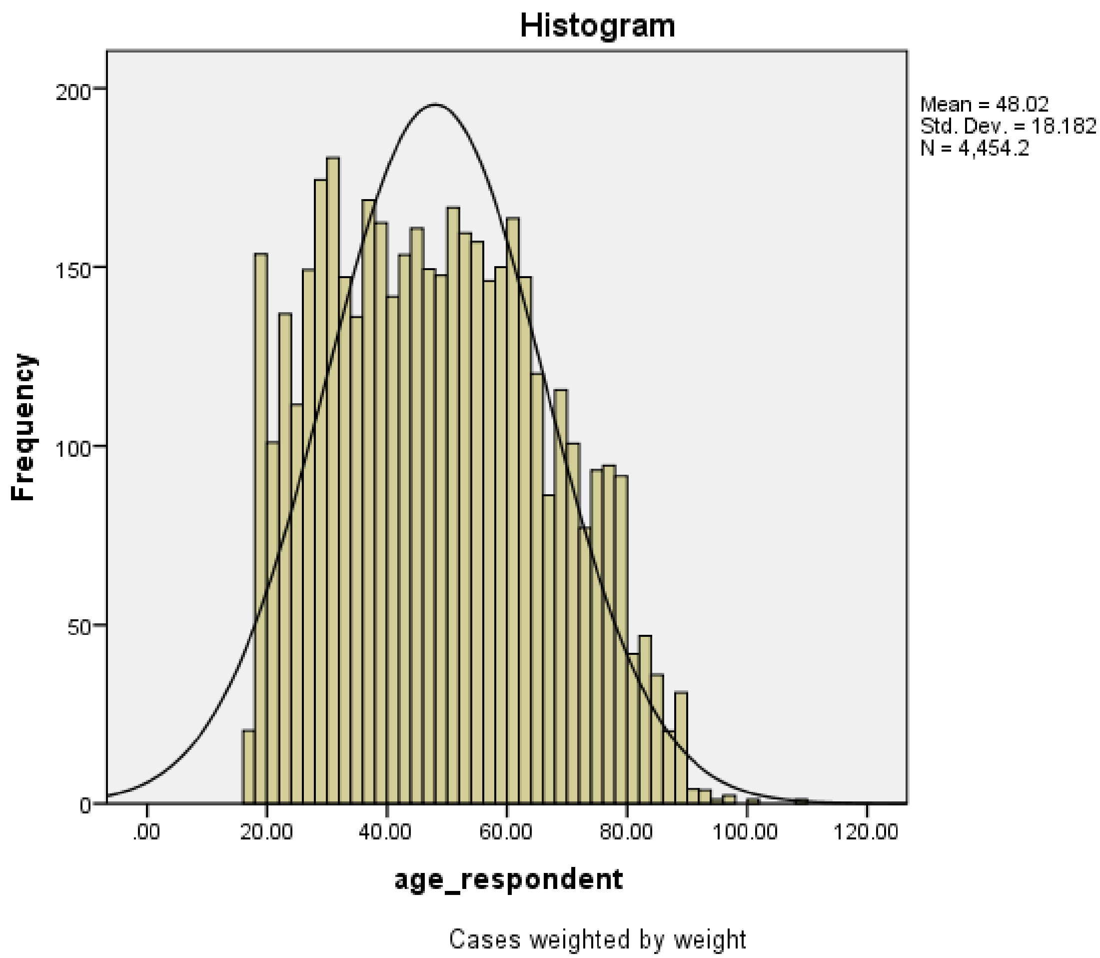
DESCRIPTIVE STATISTICS

FREQUENCIES (please do not display *frequency tables*)

CHARTS

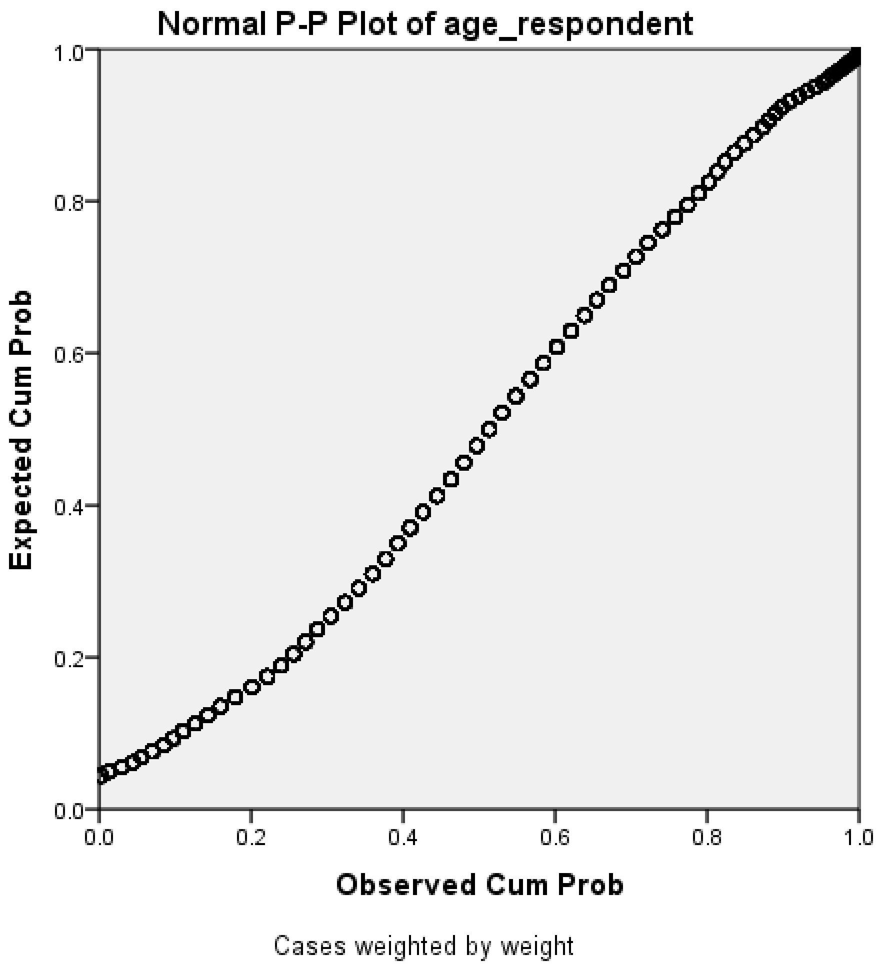
HISTOGRAMS (with normal curve)

Example ☞ dataset EVS, variable age



OR use P-P plots

- *Analyze-Descriptives-P-P plots*



STEP 2 - We have to examine the skewness and kurtosis statistics for the distribution. A normal distribution is symmetrical.

1. If a distribution meets the criteria of zero kurtosis and zero skewness it will have a normal distribution.

2. If skewness higher than 1, than it is not normally distributed.

Figure 3. Probability distribution with different Kurtosis

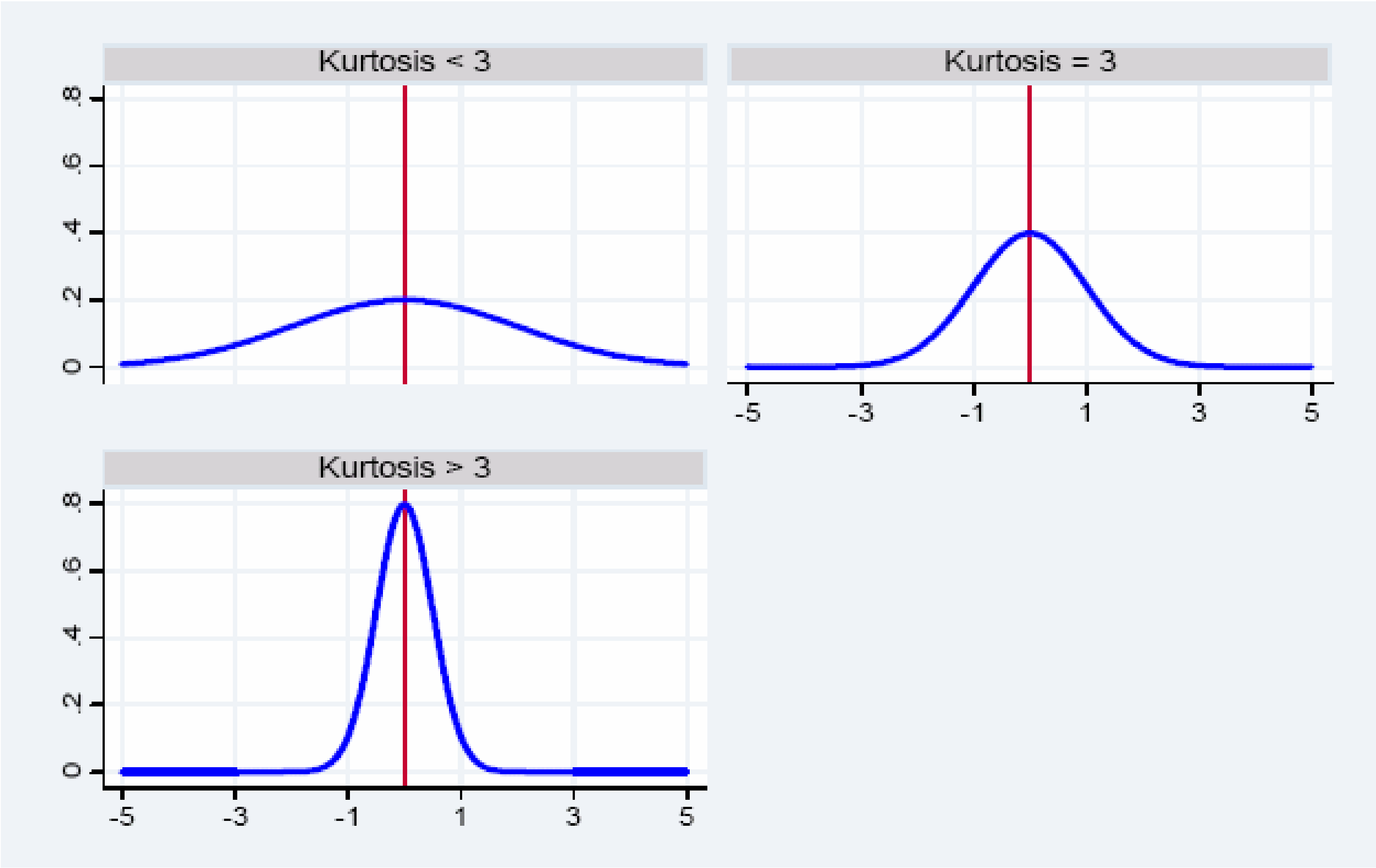


Table 1 shows the relevant statistics for variable age

Statistics

age_respondent

N	Valid	4454
	Missing	0
Mean		48.0203
Median		47.0000
Mode		29.00
Std. Deviation		18.18223
Skewness		.233
Std. Error of Skewness		.037
Kurtosis		-.848
Std. Error of Kurtosis		.073

!!!

- If we have N>>200 $\Rightarrow$ we get statistically significant values even when we have low deviation from normality

- Criteria for asymmetry not to be used when we have large samples (e.g. Field 2009, p.139)

STEP 3 - we use **Kolmogorov-Smirnov Z test**

If the Kolmogorov-Smirnov Z test indicates a significance level of less than 0.05 it means that the distribution is probably not normal.

ANALYZE

Descriptive statistics

Explore

Plots

Normality plots with tests

Table shows the results of the test

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
age_respondent	.061	4454	.000

a. Lilliefors Significance Correction

The Kolmogorov-Smirnov Z test indicates that this distribution is <u>not normal</u>.

But… remember…

- No criteria should be applied in case we have large samples (N>200).

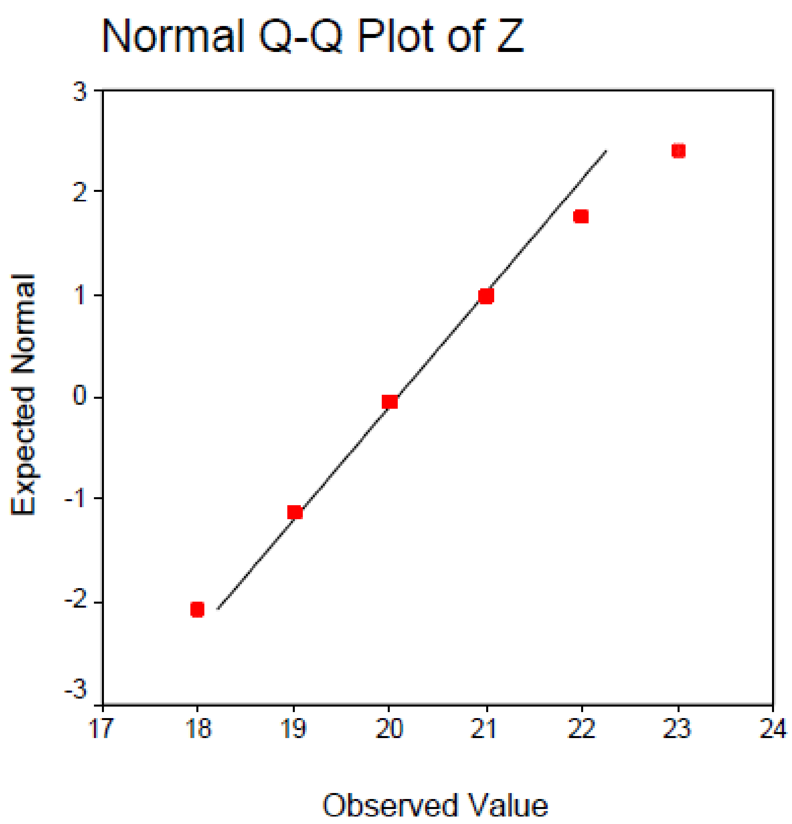
- When we work with large samples, statistical significant values are obtained even for very small deviation from normality!!!

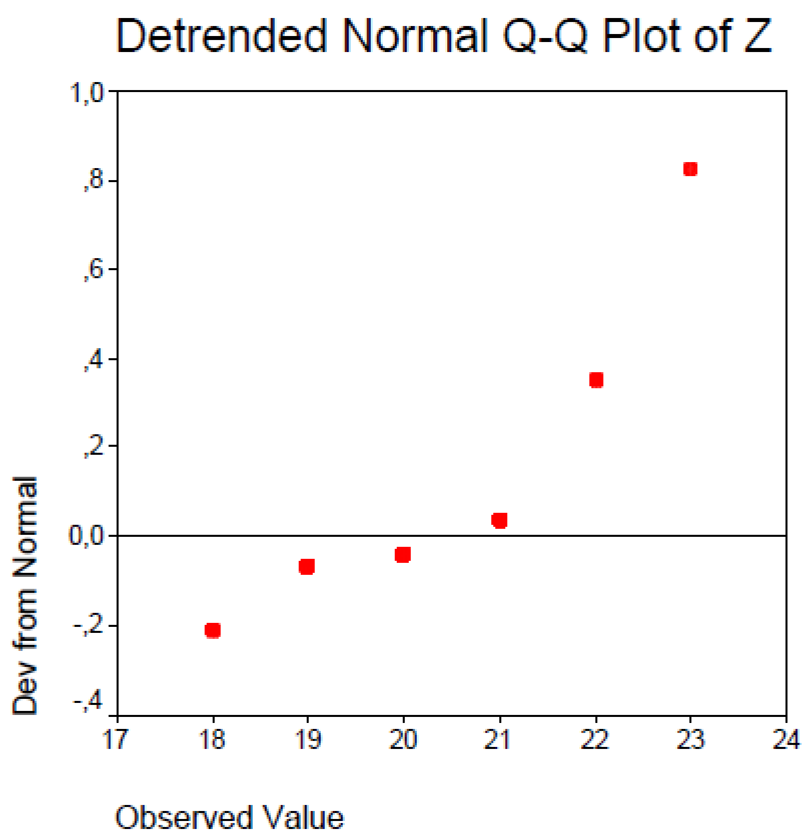
If N<50 than…

- Use Shaphiro-Wilk test

Graphical options: Normal Q-Q Plots a Detrended Normal Q-Q Plots

Explore – Plots – Normality plots with test





What to do when variables are not normally distributed?

1) Use non-parametric statistics – to be discussed later

2) Transform variables – by use of mathematical fucntions - e.g. **log function**

3) Decide to ignore it when working with big enough sample sizes – at least 100/200 cases

STANDARDIZED NORMAL DISTRIBUTION AND Z-SCORES – HOW TO CALCULATE AND USE THEM

Why *z-scores* are important?

- How do we compare bananas and oranges?
- Are you as good a student of French as you are in Sociology?
- How many people did better or worse than you on a test?

- When you analyze data ➡ to compare scores within a sample or across variables.

You may be asked:

- What percentage of people falls below a given score?
- What is the relative standing of a score in one distribution versus another?
- What score or scores can be used to define an extreme or deviant situation?

Example

- Test results SOC758 – Student 1 = 66 points, but we do not know what does mean…
- If we know the mean, than we can say whether student 1 result is better or worse than average…
- If we also know the results for another student, than we can calculate the position of these two students related to the total distribution of the results.
- For this… we need Z-scores!!!!
- To calculate… we need also SD.
- Value Z-score tells us how many SD above or bellow the average is a certain case.

Example…

- Student A = 66 points
- Student B = 81 points
- Mean = 70 points, SD = 5

Calculating the Standard Score (Z-Score)

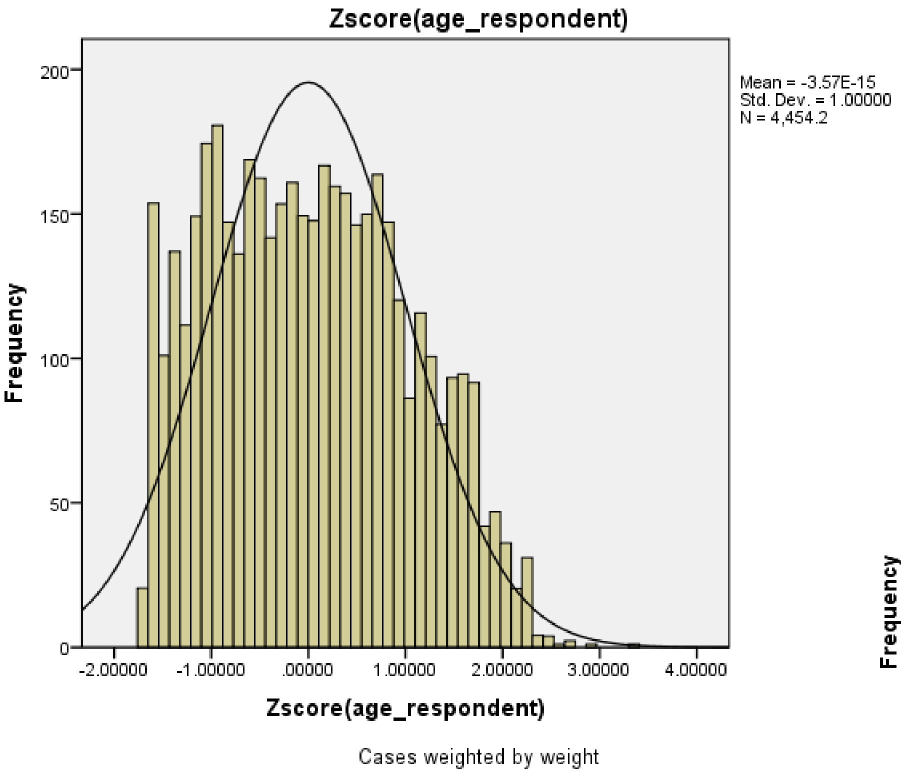
Standard Score, $z = \frac{X - \mu}{\sigma}$

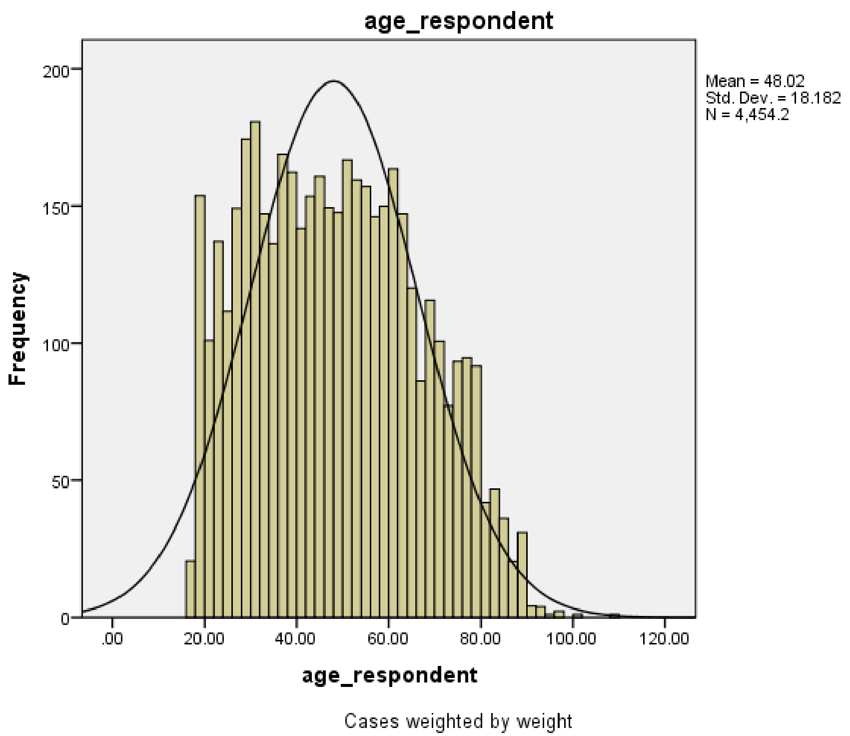
TERMS:

μ = mean (pronounced 'mu')
X = score
σ = standard deviation (pronounced 'sigma')

- Student A = (66-70)/5 = -0.8
- Student B = (81-70)/5 = 2.2

Analyze-Descriptive – Save standardized values as variables





Why do we need z-scores?

- Attributes are often measured using items with difference upper and lower limits.
- The measures have a different number of categories.

↓

- It is difficult to compare across these variables!!!
- When creating multi-item scales, items that have different lower and upper points will contribute differently to the final score!!!

How to solve these problems?

- Convert each scale to have the same lower and upper levels

OR

- Standardize the variables and express scores as standard deviation units: z-scores

1. Convert each scale to have the same lower and upper levels

- Formula:

$$Y = [(X - X_{min})/X_{range}]*n$$

Y – new adjusted variable

X – old variable to be adjusted

X_{min} *– the minimum observed value on the original variable*

X_{range} *– the difference between the maximum and minimum observed on the original variable*

n – the upper limit of the adjusted variable

Example: political implication/orientation

- 4 variables:

- **V186** – measured on 4-point
- **V193** – measured on 10-point scale
- **V222** – measured on 4-point
- **V224** – measured on 10-point

We want to convert them to a scale of 1-10.

It will help us to compare scores and averages across them!!!

2. Standardize the variables and express scores as standard deviation units: z-scores

- It gives each person's score in terms of the number of standard deviations it lies from the mean!
- A ***z-score*** reflects how many standard deviations above or below the population mean a score is.
- A normal distribution that is standardized is called the standard normal distribution or *the normal distribution of z-scores*.
- **It has a mean of 0 and a SD of 1.**

How to calculate Z-scores?

Here are the formulas for z-scores, z-skewness and z-kurtosis:

Calculating the Standard Score (Z-Score)

Standard Score, $z = \frac{X - \mu}{\sigma}$

TERMS:
μ = mean (pronounced 'mu')
X = score
σ = standard deviation (pronounced 'sigma')

$$\mathbf{Z_{skewness} = (S\text{-}0) \,/\, SE_{skewness}}$$

$$\mathbf{Z_{kurtosis} = \sqrt{(K\text{-}0)/SE_{kurtosis}}}$$

Sx= *standard deviation,*

$SE_{skewness}$ = *standard deviation for Skewness*

$SE_{kurtosis}$ = *standard deviation for Kurtosis*

Things to know about the **Z-Score**:

- The Z-score can be positive or negative.
- Positive is above the mean.
- Negative is below the mean.
- The mean of the Z-scores is always zero.
- The SD of the Z distribution = 1.

Does it matter if my dependent variable is normally distributed?

YES

↓

When running a t-test or ANOVA, the assumption is that the distribution of the sample means are normally distributed.