Effective Use of Advanced Statistical Methods in Research I

KEHINDE OLUWADIYA
Professor of Surgery (Orthopaedics)
Ekiti State University, Ado-Ekiti
CEO
POSK Educational Consult
www.oluwadiya.com

EKSU
EKITI STATE UNIVERSITY

POSK

- Kudos to the College for organizing this workshop
- Thank you for making me a part of it

Preamble

This is the first of a 3-part lecture on the use of advanced statistical methods in medical research

Objectives

PART I

1. Why do you need to know statistics?
2. What you need for effective use of statistics
3. Data transformation

PART II

1. Limitations of P-value
2. Statistics for comparing 2 or more groups with continuous data
3. Regressions and Correlation

PART III

1. Risk Ratios and Odds Ratios
2. Survival Analysis
3. Sensitivity, Specificity and ROC Curves
4. Finding the right test for specific data

Objective of the lecture series

- To provide a 3-hour overview (including demonstrations) of the important practical information that a clinical investigator needs to know about biostatistics to be successful.

Introduction

WHY DO YOU NEED TO KNOW STATISTICS?

BECAUSE STATISTICS, WHEN MISUSED, CAN BE DANGEROUS

Statistics can be made to lie!

“THERE ARE THREE KINDS OF LIES:

LIES, DAMN LIES AND STATISTICS”

Benjamin Disraeli

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."

Statistics can contain errors!

There is an increasing number of publications on the flaws and errors in much of published medical literature:

- The scandal of poor medical research- **DG Altman, 1994**
- Statistical errors in medical research, a chronic disease?- **J Young, 2007**
- Improved reporting of statistical design and analysis: guidelines, education, and editorial policies. **Mazumdar M et al 2010**
- "Why most published research findings are false"- **John Iaonnidis**

And many more…

Medical Lies?

In 2005, PLoS Medicine published an article by John Ioannidis that has been downloaded over 100,000 times and has won the author many prizes and accolades..

The title of the article?

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

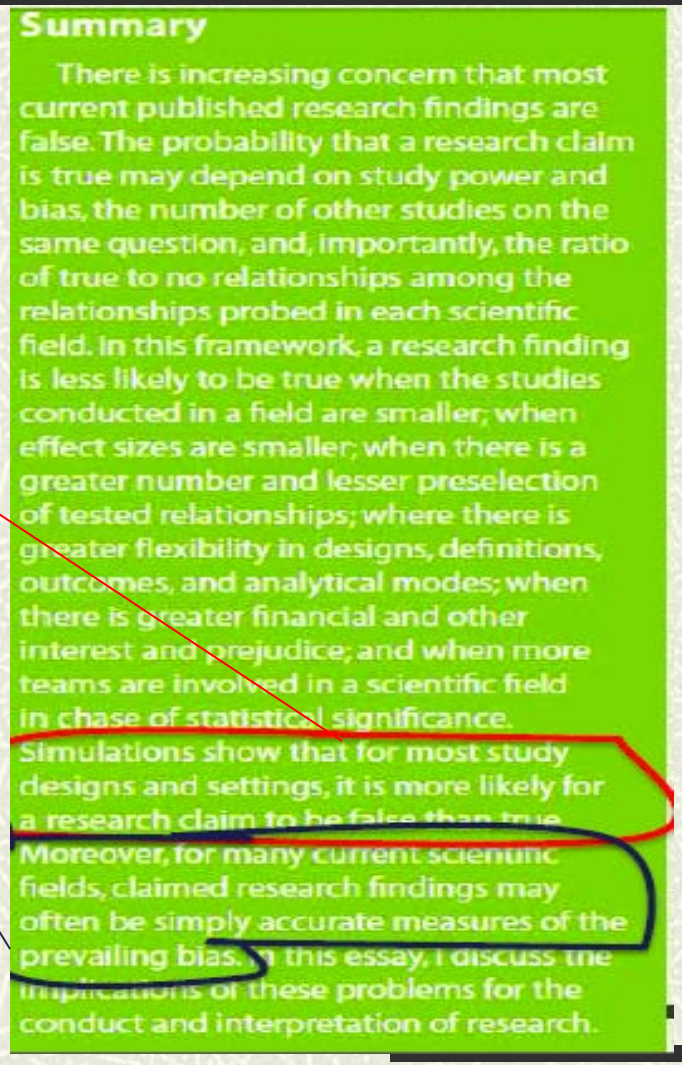
Summary

factors that influence this problem and is characteristic of the field and can

Why most published research findings are false…. **John P. A. Ioannidis**

…..Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias…..



Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Medical lies?

ORIGINAL CONTRIBUTION

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

CLINICAL RESEARCH ON IMPOR-

John P. A. Ioannidis

His Methodology

- Examined all original clinical-research studies that were published in three major non-specialty journals (*New England Journal of Medicine*, *JAMA*, and the *Lancet*) and "high-impact-factor" specialty journals between 1990 and 2003 that were cited more than 1000 times in the literature.
- He compared the results of these highly cited studies with those from subsequent studies of comparable or larger sample sizes and similar or better designs.

His Findings

- **Hall of shame: % Contradicted by later studies:**

- **80% of non-randomized studies were wrong**
- **25% of supposedly gold-standard! randomized trials were contradicted!**
- **10% of large randomized trials were contradicted!**

Excerpts from DG Altman……

- *"We need less research, better research, and research done for the right reasons"*
- *"We need not be experts in statistics, but we should understand the principles of sound methods of research. If we can also analyze our own data, so much the better.* ***Amazingly, it is widely considered acceptable for (medical) researchers to be ignorant of statistics. Many are not ashamed (and some seem proud) to admit that they don't know anything about statistics"***

The scandal of poor medical research- DG Altman, 1994

As scientists and medical researchers, it is our sacred duty to uphold the standards of research in our specialties

To do this, we should empower ourselves, understand the underlying principles, and be ready to stand by them…

Functions of statistics…..1

1. **To reduce data.** This is done:
   I. Graphically by compiling charts, tables, graphs, histograms, frequency polygons etc.,
   II. Univariate analysis (mean, median, standard deviations, range etc.)

- Aim is to determine trends and summaries of variables.

Simply speaking…..

Statistics is a tool for converting ***DATA*** into ***INFORMATION***

Data into Information

Data → Information

age	age2	MAC
10	13	15.00
5	8	10.00
6	9	11.00
7	10	12.00
4	7	9.00
9	12	14.00
15	18	20.00
10	13	15.00
5	8	10.00
6	9	11.00
7	10	12.00
4	7	9.00
9	12	14.00
15	18	20.00
10	13	15.00
5	8	10.00
6	9	11.00
7	10	12.00
4	7	9.00
9	12	14.00
15	18	20.00
10	13	15.00

Frequencies

Statistics

		age	age2	MAC
N	Valid	22	22	22
	Missing	6	6	6
Mean		8.09	11.09	13.0909
Median		7.00	10.00	12.0000
Mode		10	13	15.00
Std. Deviation		3.490	3.490	3.49025

Functions of statistics….2

2. To provide methods of applying tests of significance.

- Tests of significance are used to separate real differences from those due to chance. In general, the level of significance is arbitrarily set at 5% (p = probability = 0.05).

Functions of statistics…..3

3. To provide a sound basis for experimental design

- Experiments must be carefully designed because a good design may mean the difference between a sound, scientific research and worthless data which yield little or no information.
- In many instances, more information are obtained with the same amount of work if the researcher has a knowledge of statistical methods and plans his experiment accordingly.

To effectively use statistics:

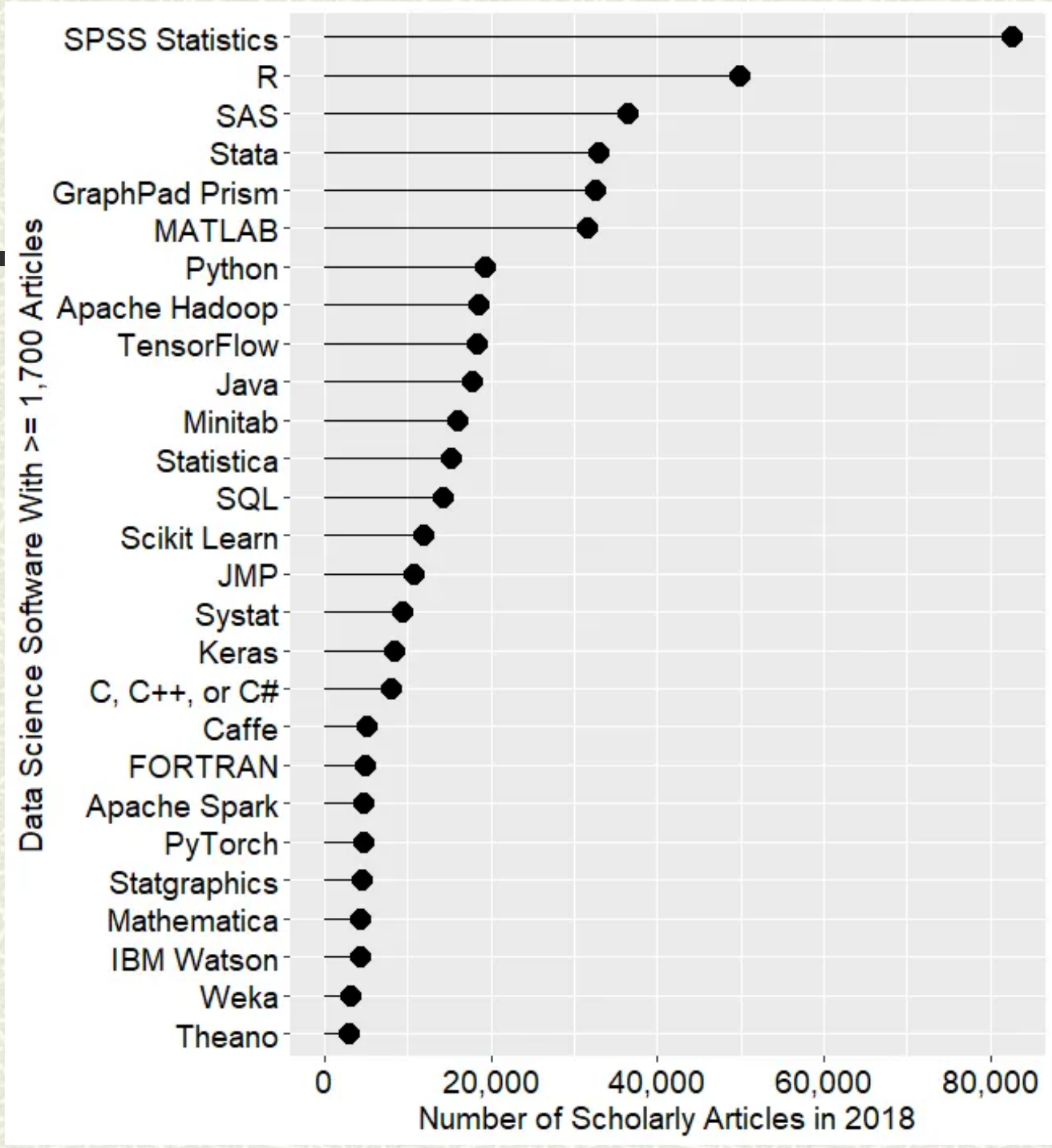
1.YOU NEED THE RIGHT TOOLS

Install a powerful, yet easy to use, statistical software package on your computer.

- You don’t need to do the math
- Many comes with explanation of their outputs
- There are numerous help sources to get you going
- I recommend SPSS.

Why SPSS?

- By far, SPSS is the most popular package.
- It nicely balance between power and ease-of-use
- It is much easier to use than R, SAS or Stata



Bottom line

- Get to know whatever software you are using.
- Learn to use the correct statistics correctly!
- Learn to interpret the outputs correctly

To effectively use statistics:

2.YOU NEED DATA THAT WAS COLLECTED CORRECTLY

CORRECTLY COLLECTED DATA

While this topic has been covered extensively by some of the previous speakers, I just want to add the following:

Surveys can be conducted using smartphones & Tablets

And they have many advantages over traditional methods of surveys

- They’re portable
- Come with an on-board GPS receiver (helps to guide against fudging)
- Have on-board cameras
- Automatically record time taken to enter data (helps to guide against fudging)
- No need to enter data separately after collection
- Can connect to wireless networks
- Access to the internet
- Email is available
- There’s an app for it!

Electronic data collection using smartphones

Different (Free) Apps:

- Open Data Kit (ODK) (https://www.google.com.au/intl/en/earth/outreach/tutorials/odk_gettingstarted.html)
- Epicollect (http://www.epicollect.net/instructions/)
- Epicollect+ (http://www.epicollect.net/plus_Instructions/default.html)

What about online survey tools?

- All the advantages of questionnaire plus the convenience of online tools
- Free versions available e.g., Google Forms, Survey Monkey
- Reach and scalability is much more than traditional questionnaire
- Cheap
- Less time consuming for the researcher
- More accurate data entering
- Quicker
- Ensures better anonymity and therefore improves confidentiality

TRANSFORMING DATA

Why Transform Data?

- The assumptions of most parametric methods include:
  - Homogeneity of variance (Homoscedasticity)
  - Normality
  - Linearity
- Data transformation is used to make your data conform to the assumptions of the statistical methods

Normal vs Skewed Data

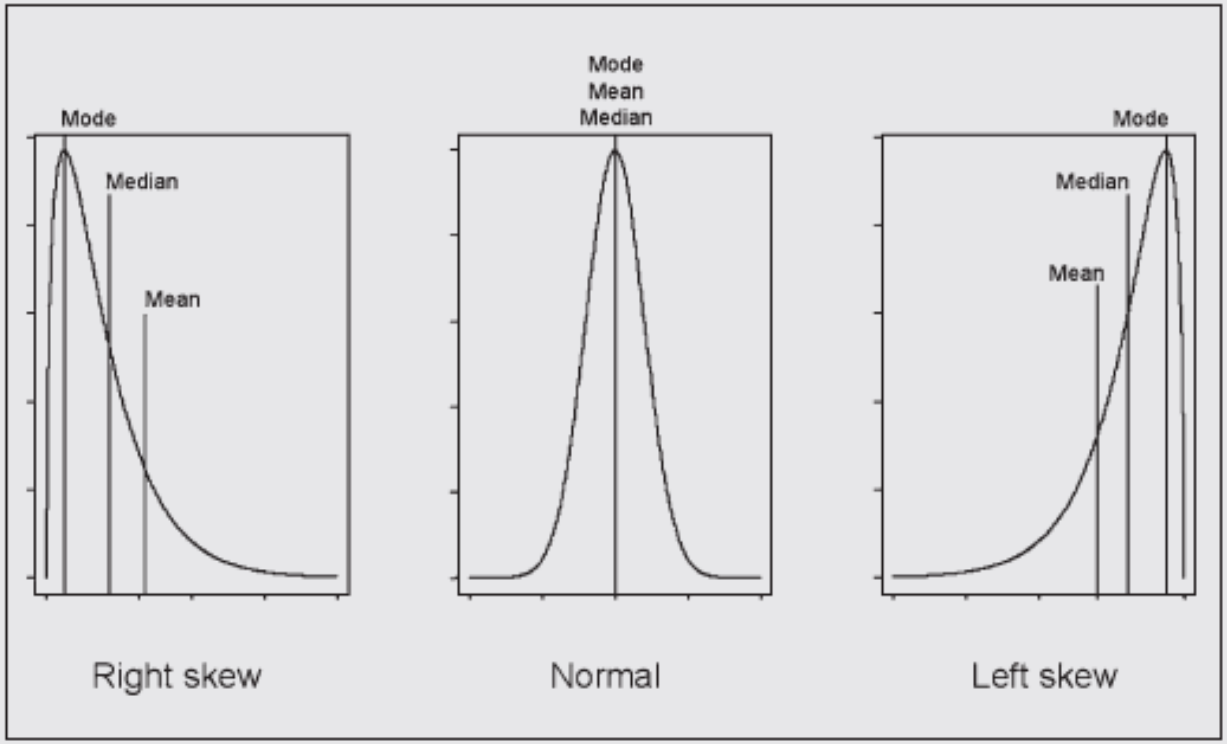
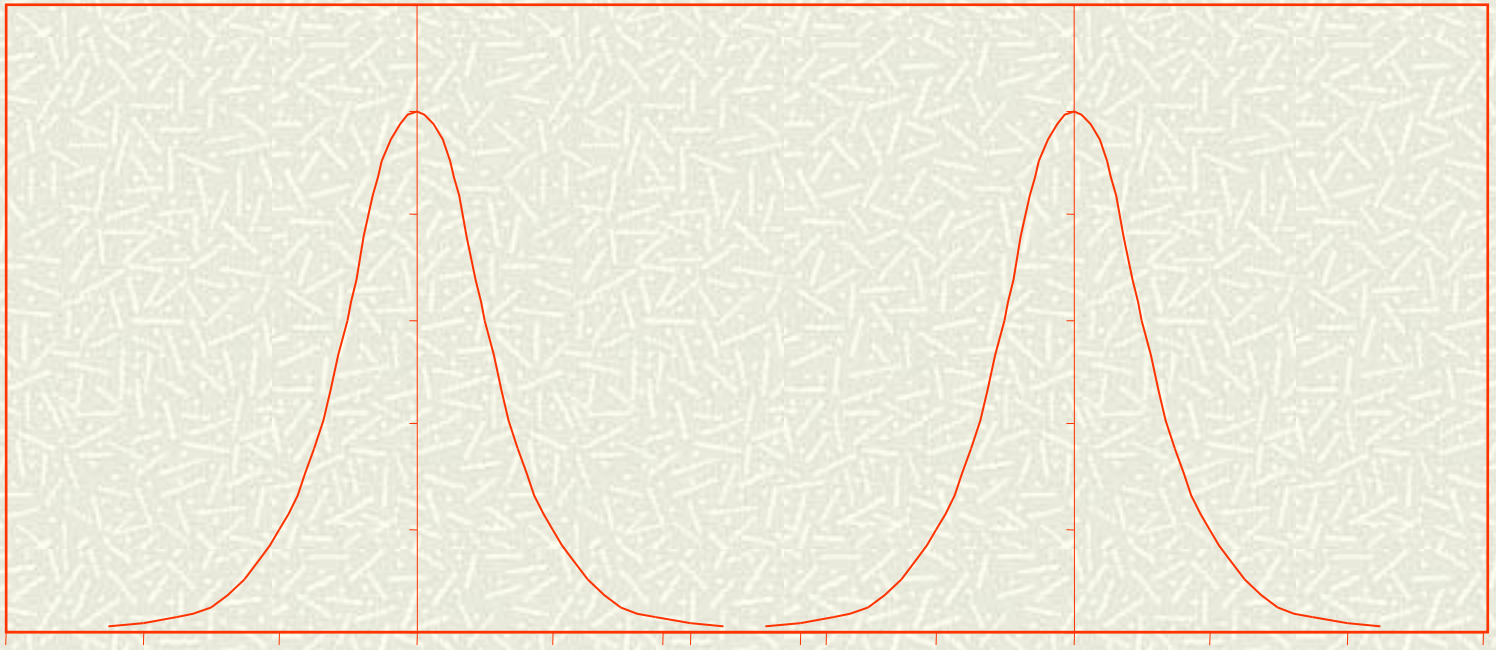


Fig. 2 Distributions of Quantitative Data.

Homoscedasticity

Homo: Same **Scedastiscity**: to scatter



This two groups are both normal and have equal scatter (variance)

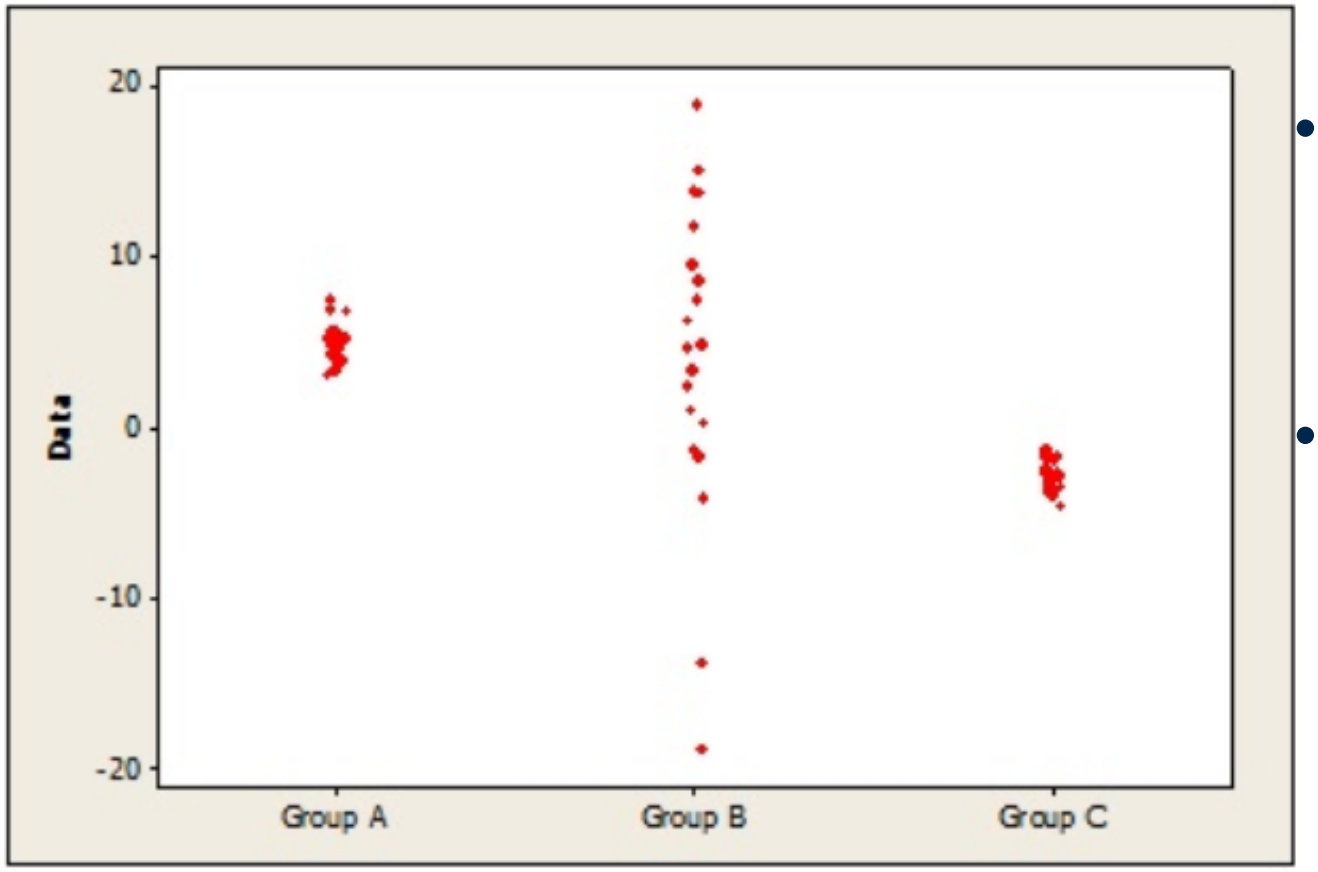
Won't it be nice if we would make the previous data look this way?

Heteroscedasticity

The two groups have unequal variance

The **Barlett test** or the **Levene test** are used to determine if the variances of groups defer

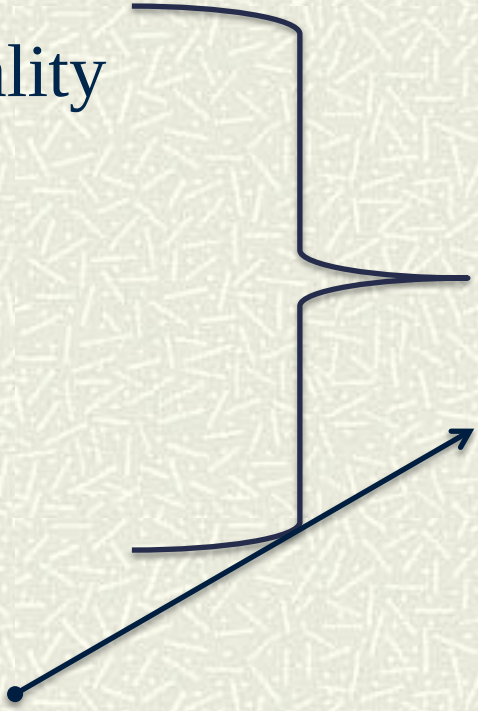
More on scatter



- **Group A and C exhibits homoscedasticity**
- **Group A and B exhibit heteroscedasticity**

Determining normality of data

- Graphical method
  - Histograms and Normality plots
  - Boxplots
  - Normal Q-Q plots and detrended Q-Q plots
- Statistical method
  - Skewness and kurtosis
  - Smolgrov-Smirnov statistics



In SPSS, use the EXPLORE procedure to obtain these parameters

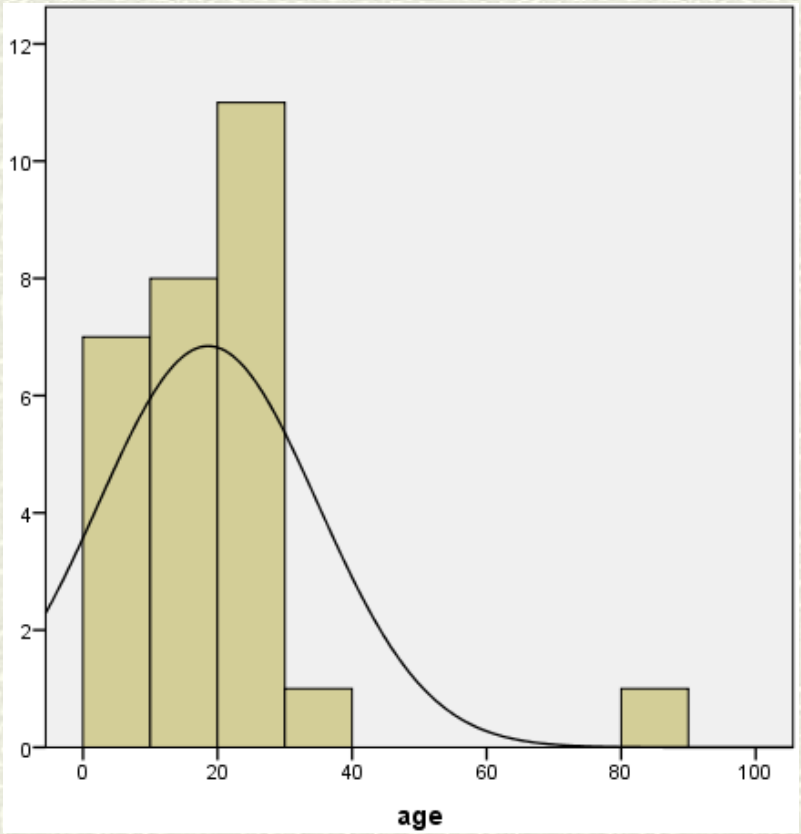
Determining normality of data

Which of the two variables has a normal distribution?

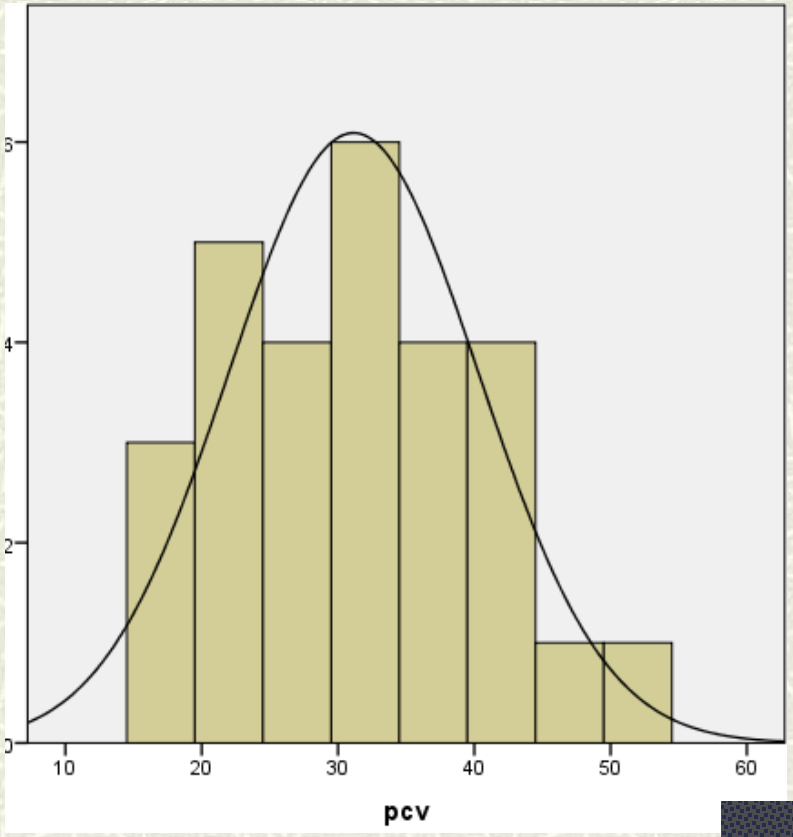
AGE	PCV
4	35
5	34
6	34
6	30
10	25
12	43
5	34
4	38
3	43
2	30
17	29
18	40
18	25
20	21
20	99
21	26
22	50
22	32
22	17
22	19
26	40
26	20
29	47
29	21
32	38
8	19
10	23
11	36
12	99
89	23

Determining normality of data: Normality Curve

Age

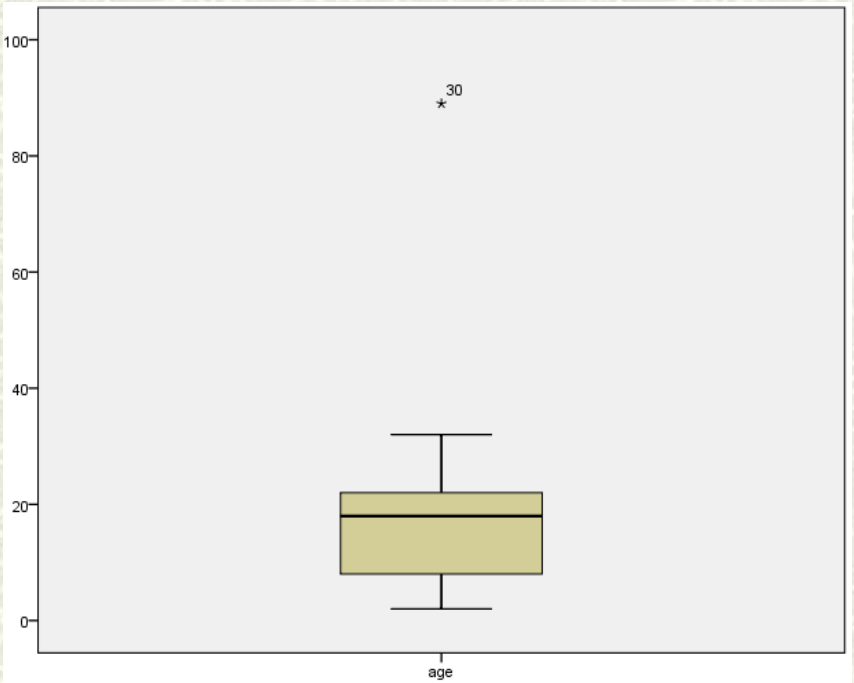


PCV

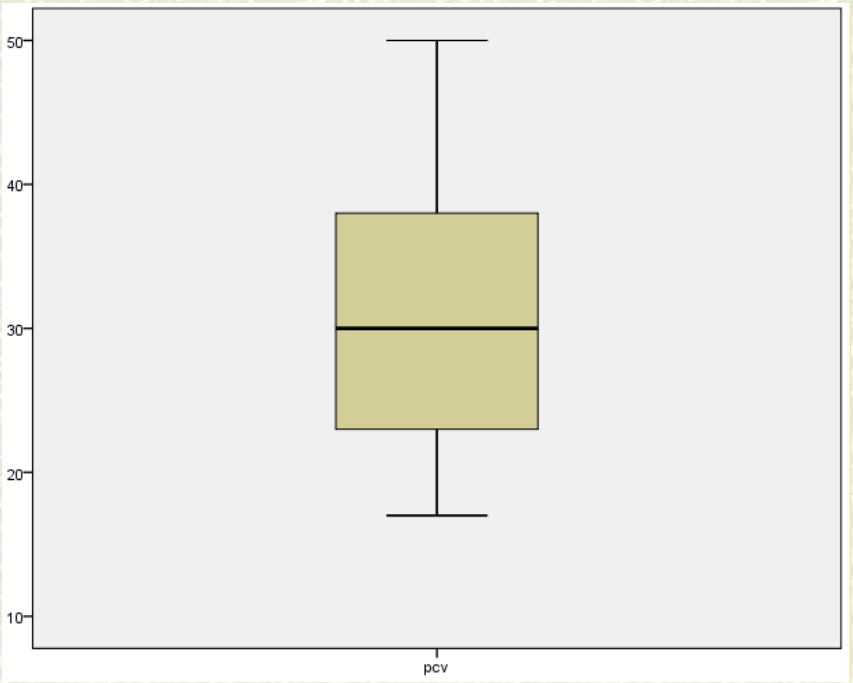


Determining normality of data: Box plot

AGE



PCV



Determining normality of data: The Explore procedure

AGE

Descriptives

			Statistic	Std. Error
age	Mean		18.81	3.315
	95% Confidence Interval for Mean	Lower Bound	11.98	
		Upper Bound	25.63	
	5% Trimmed Mean		16.56	
	Median		18.00	
	Variance		285.682	
	Std. Deviation		16.902	
	Minimum		2	
	Maximum		89	
	Range		87	
	Interquartile Range		16	
	Skewness		2.946	.456
	Kurtosis		11.971	.887

PCV

pcv	Mean		30.92	1.861
	95% Confidence Interval for Mean	Lower Bound	27.09	
		Upper Bound	34.76	
	5% Trimmed Mean		30.65	
	Median		30.00	
	Variance		90.074	
	Std. Deviation		9.491	
	Minimum		17	
	Maximum		50	
	Range		33	
	Interquartile Range		16	
	Skewness		.313	.456
	Kurtosis		-.989	.887

Determining normality of data: The K-S procedure

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
age	.196	26	.011	.705	26	.000
pcv	.121	26	.200*	.952	26	.260

Decision Time

- We should always check the assumptions that data follow a normal distribution with uniform variance (homoskedasticity):
  i. If the data meet the assumptions, we can analyze the raw data as described.
  ii. If the assumptions are not met, we have **three** possible strategies:

What if the variable is not normally distributed?

1. We can use a method which does not require these assumptions, such as a rank-based (non-parametric) method.
2. Thanks to the Central Limit Theory, if you have a large enough sample size (Taken to be at least 30), you may go ahead, and use a parametric technic even if your data is skewed!
3. We can transform the data mathematically to make them fit the assumptions more closely before analysis.

Methods of data transformation

In healthcare research, there are three commonly used transformations for quantitative data:

1. Logarithmic transformation,
2. Square root transformation
3. Inverse (reciprocal) transformation.

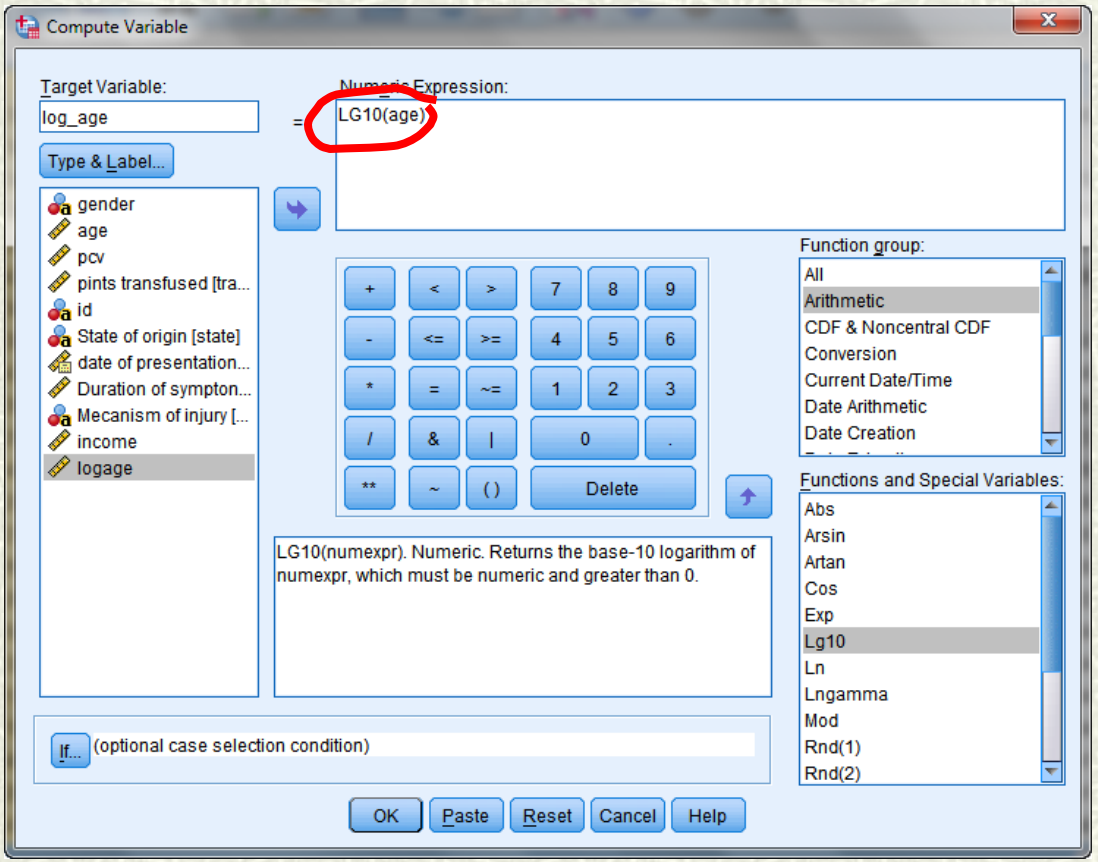
Normalizing data in SPSS

- We know that Age is not normally distributed
- We are going to normalize Age:
- Log transformation: use SPSS Compute Sub menu:

Transform

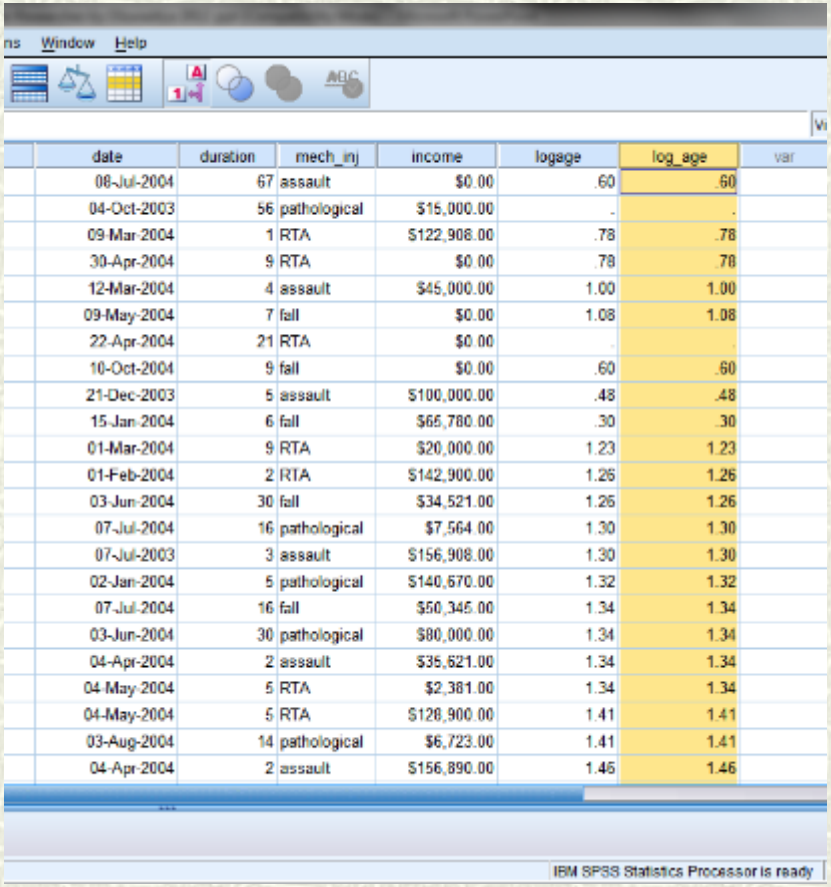
↓

Compute



Normalizing data in SPSS

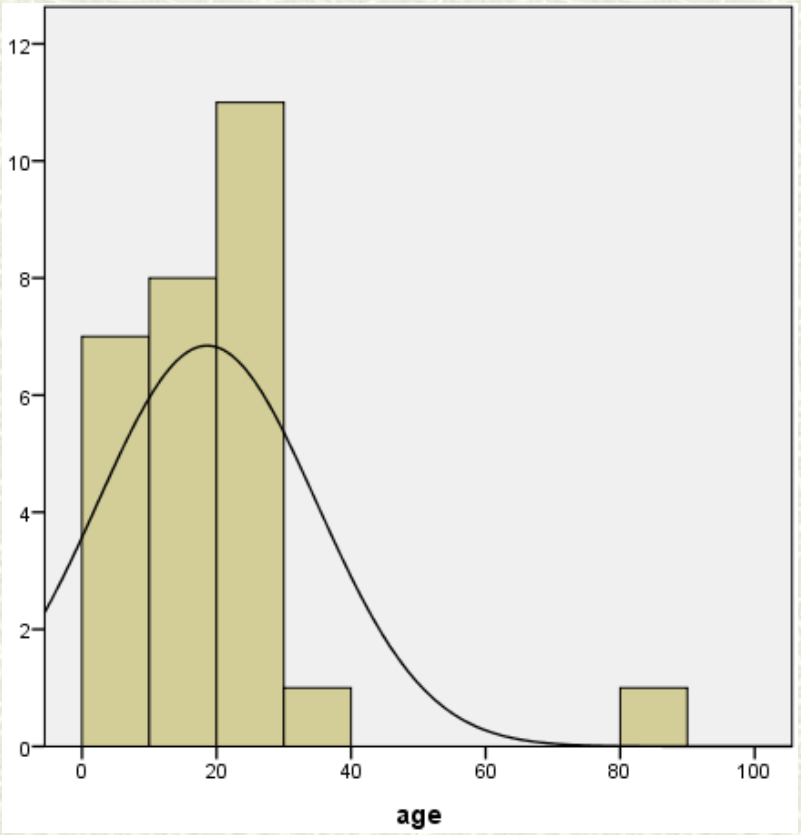
- The transformed variable (log_age) which we asked SPSS to create has been created



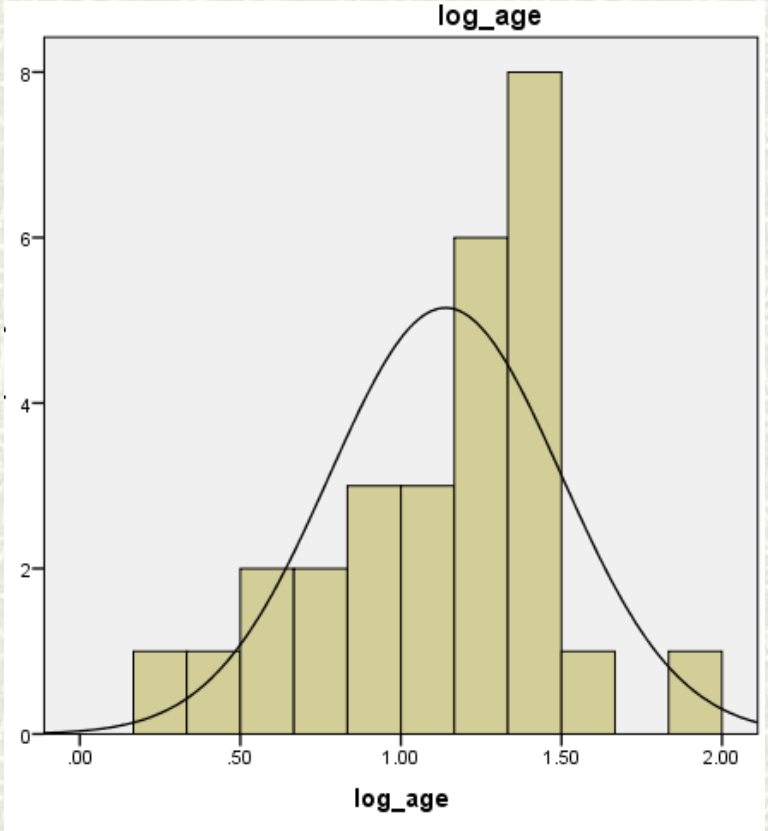
date	duration	mech_inj	income	logage	log_age	var
08-Jul-2004	67	assault	$0.00	.60	.60	
04-Oct-2003	56	pathological	$15,000.00	.	.	
09-Mar-2004	1	RTA	$122,908.00	.78	.78	
30-Apr-2004	9	RTA	$0.00	.78	.78	
12-Mar-2004	4	assault	$45,000.00	1.00	1.00	
09-May-2004	7	fall	$0.00	1.08	1.08	
22-Apr-2004	21	RTA	$0.00	.		
10-Oct-2004	9	fall	$0.00	.60	.60	
21-Dec-2003	5	assault	$100,000.00	.48	.48	
15-Jan-2004	6	fall	$65,780.00	.30	.30	
01-Mar-2004	9	RTA	$20,000.00	1.23	1.23	
01-Feb-2004	2	RTA	$142,900.00	1.26	1.26	
03-Jun-2004	30	fall	$34,521.00	1.26	1.26	
07-Jul-2004	16	pathological	$7,564.00	1.30	1.30	
07-Jul-2003	3	assault	$156,908.00	1.30	1.30	
02-Jan-2004	5	pathological	$140,670.00	1.32	1.32	
07-Jul-2004	16	fall	$50,345.00	1.34	1.34	
03-Jun-2004	30	pathological	$80,000.00	1.34	1.34	
04-Apr-2004	2	assault	$35,621.00	1.34	1.34	
04-May-2004	5	RTA	$2,381.00	1.34	1.34	
04-May-2004	5	RTA	$128,900.00	1.41	1.41	
03-Aug-2004	14	pathological	$6,723.00	1.41	1.41	
04-Apr-2004	2	assault	$156,890.00	1.46	1.46	

Normalize Age: Result

Original Data (Age)



Transformed data



Normalize PCV: Result

Log_Age

logage	Mean		1.1388	.06831
	95% Confidence Interval for Mean	Lower Bound	.9986	
		Upper Bound	1.2789	
	5% Trimmed Mean		1.1445	
	Median		1.2553	
	Variance		.131	
	Std. Deviation		.36146	
	Minimum		.30	
	Maximum		1.95	
	Range		1.65	
	Interquartile Range		.42	
	Skewness		-.470	.441
	Kurtosis		.379	.858

Age

Descriptives

			Statistic	Std. Error
age	Mean		18.81	3.315
	95% Confidence Interval for Mean	Lower Bound	11.98	
		Upper Bound	25.63	
	5% Trimmed Mean		16.56	
	Median		18.00	
	Variance		285.682	
	Std. Deviation		16.902	
	Minimum		2	
	Maximum		89	
	Range		87	
	Interquartile Range		16	
	Skewness		2.946	.456
	Kurtosis		11.971	.887

THIS BRINGS US TO THE END OF PART I

About Me

Oluwadiya Kehinde

- Professor of Surgery at the Ekiti State University, Ado-Ekiti
- Author of "**Getting to Know SPSS**", the best selling book on SPSS in Nigeria
- CEO of **POSK Educational Consult**, Consultancy Firm for Training in Statistical and Health Education

www.Oluwadiya.com

Thanks for your attention

To ask questions, please join the forum at www.oluwadiya.com