Training course of Research & Biostatistics for Medical Professionals

Dr. Wong Kai Choi

MBChB, BSc, MEd, GradStat, AMIMA, FHKAM(Psychiatry), FHKCPsych

30th July 2011

Dr. Wong Kai Choi

- Bachelor of Medicine and Bachelor of Surgery (*CUHK*)
- Bachelor of Science with First Class Honors in Mathematics (*OUHK*)
- Master of Education (*OUHK*)
- Graduate Diploma in Statistics (*Royal Statistical Society*)
- Fellow of Hong Kong Academy of Medicine (Psychiatry)
- Fellow of Hong Kong College of Psychiatrists
- Associate Member of Institute of Mathematics and its Application (*UK*)
- Graduate Statistician of Royal Statistical Society
- Specialist in Psychiatry

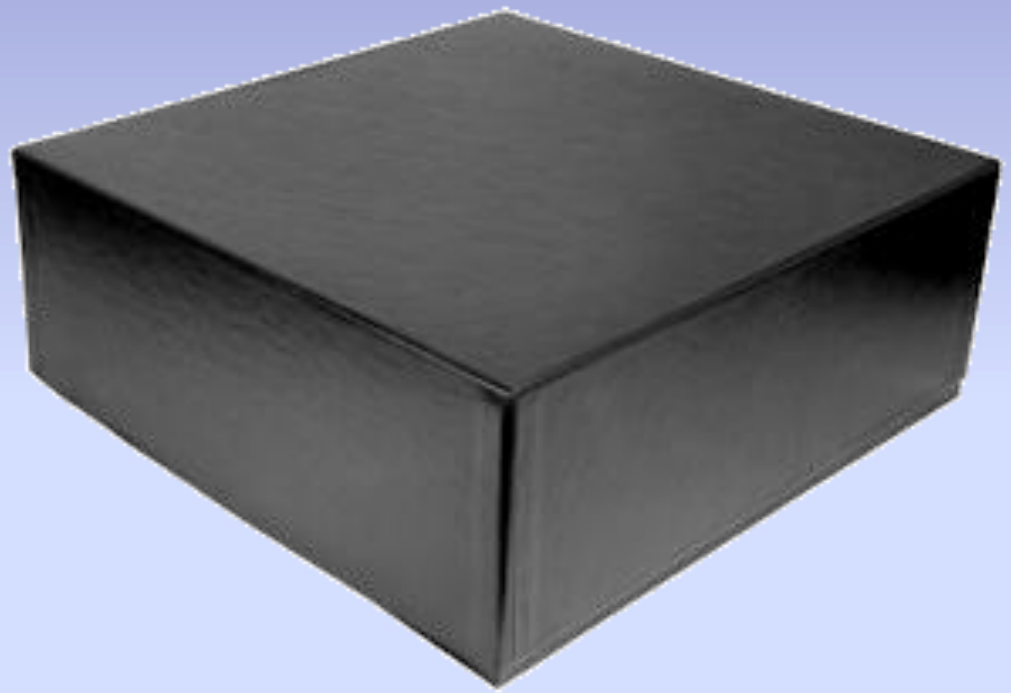
WHAT IS STATISTICS

Fact and Truth

Great Philosopher said:

“ If you want to learn more about FACT, please ask STATISTICIAN;

If you want to learn more about TRUTH, please ask MATHEMATICIAN.”



Role of Medical Statistician

- Should be involved in the beginning of study
- Advise on study design
- Calculation of sample size (if appropriate)
- Choice of appropriate statistical test
- Data analysis
- Data management
- Interpretation of statistical result

PRE-TEST

Variable	OCD (n = 50)	Schizophrenia (n = 47)	X^2 (degrees of freedom), p value
Mean (standard deviation) age (years)	29 (9)	36 (11)	514 (506), 0.38
Sex			
Male	37 (74)	38 (81)	1.29 (2), 0.52
Female	13 (26)	9 (19)	
Marital status			
Single	14 (28)	10 (21)	
Married	34 (68)	36 (77)	1.66 (4), 0.79
Widowed	2 (4)	-	
Separated	-	1 (2)	
Occupation			
Professional	13 (26)	6 (13)	
Clerical / shop owner	3 (6)	6 (13)	
Farmer	2 (4)	1 (2)	
Skilled worker	10 (20)	9 (19)	
Semi-skilled / unskilled worker	1 (2)	14 (30)	38.8 (48), 0.82
Unemployed	10 (20)	6 (13)	
Housewife	-	1 (2)	
Retired	11 (22)	4 (9)	

TABLE 3. Univariate analysis of potential factors affecting preferences for developing a Health Call Centre

Characteristic	Chi squared test			
	Agree to develop a Call Centre % (No.)	Disagree to develop a Call Centre % (No.)	X^2	P value
Gender				
Male	49.3 (168)	51.7 (31)	0.118	0.780
Female	50.7 (173)	48.3 (29)		
Age (years)				
18-30	6.8 (23)	3.3 (2)	36.269	<0.001
31-64	78.5 (266)	48.3 (29)		
≥65	14.7 (50)	48.3 (29)		
Education				
Primary school or below	28.1 (96)	61.0 (36)	25.784	<0.001
Secondary school	58.2 (199)	27.1 (16)		
University or above	13.7 (47)	11.9 (7)		
Marital status				
Married	82.5 (282)	81.7 (49)	0.022	0.856
Others	17.5 (60)	18.3 (11)		
Working status*				
Working	48.2 (164)	31.7 (19)	5.641	0.024
Not working	51.8 (176)	68.3 (41)		
Individual monthly income				
No income ($0)	38.9 (128)	63.2 (36)	11.694	0.001
Received income	61.1 (201)	36.8 (21)		

Health status				
No difference – much better	68.0 (230)	56.9 (33)	2.760	0.100
Worse – much worse	32.0 (108)	43.1 (25)		
Being cared for by other people				
Yes	9.6 (33)	20.0 (12)	5.501	0.026
No	90.4 (309)	80.0 (48)		
Caregiver				
Yes	19.7 (67)	8.5 (5)	4.288	0.043
No	80.3 (273)	91.5 (54)		
Health insurance coverage				
Yes	32.9 (112)	18.3 (11)	5.111	0.023
None	67.1 (228)	81.7 (49)		
Fee waiver				
Yes	27.5 (94)	28.3 (17)	0.018	0.877
None	72.5 (248)	71.7 (43)		

Types of fee waiver[†]				
CSSA/disability allowance	26.7 (24)	56.3 (9)	5.545	0.037
Civil servant/HA benefits	73.3 (66)	43.8 (7)		
Presence of chronic disease				
Yes	41.3 (141)	31.7 (19)	1.995	0.198
None	58.7 (200)	68.3 (41)		
Treatment on health problem last time				
Yes	2.9 (10)	6.7 (4)	2.127	0.141
No treatment	97.1 (332)	93.3 (56)		
Methods of solving health problem last time				
Ask health professional/others	80.1 (274)	83.3 (50)	0.781	0.677
Self (internet/books/others)	7.6 (26)	8.3 (5)		
Both	12.3 (42)	8.3 (5)		
Used health hotline before				
Yes	2.7 (9)	1.7 (1)	0.204	1.000
No	97.3 (330)	98.3 (59)		

HISTORY OF STATISTICS

- 1532 – First weekly data on deaths in London (Sir W. Petty)
- 1539 – start of data collection on birth, marriages and deaths in France
- 1662 – First published demographic study based on bills of mortality (J. Braunt)
- Publication of *Ars Conjectandi* (J Bernoulli)
- 1834 – establishment of Royal Statistical Society

- 1839 – Establishment of American Statistical Association (Boston)
- 1889 - Publication of *Natural Inheritance* (F. Galton)
- 1900 – development of chi-squared test (K Pearson)
- 1901 – publication of first issue of *Biometrika* (F. Galton)
- 1903 – development of Principal Component Analysis (K Pearson)

- 1908 – publication of *The Probable error of a mean* (“Student”)
- 1935 – Publication of *The Design of Experiments* (R A Fisher)
- 1972 – Publication of *Regression models and life tables* (D R Cox)
- 1979 – Publication of *Bootstrap methods: another look at the jackknife* (B Efron)

CONTENT

- Sample size calculation
- Sampling method
- Descriptive statistics
- Type of data and data management
- Statistical test and significant test
- Interpretation of statistical result
- Common statistical software

SAMPLE SIZE CALCULATION

Sample

- Subset of population
- Rarely use the entire population
- May be imperfect representation of population
- If sample is unbiased and large enough, it will give useful information

If sample size too small

- 95% confidence interval will not be narrow enough to give precise estimate (mean, median or proportion)
- In a comparative test, a true difference may be missed
- Unethical to the subject involved and the society (if the intervention is beneficial with minimal side effect)

If the sample size too large

- Wasting of resource, including time, money, manpower, etc.
- In naturalistic study, positive result (in comparative test) with low clinical significant will be detected.
- Unethical to society and subject (if intervention with no real effect but have serious side effect)

Sample size for means estimation

- Information required
  - Standard deviation (SD) of the measure being estimated (can be from previous studies)
  - The desired width of the confidence interval (d)
  - The confidence interval (CI) (95%, 99% , 90%, etc)

Sample size for means estimation

- $n = (CI)^2 \times 4\,(SD)^2 / d^2$
  - For 95% confidence interval, CI = 1.96
  - For 90% confidence interval, CI = 1.64
  - For 99% confidence interval, CI = 2.58

Sample size for proportion estimation

- Information required
  - The expected population proportion, p
  - The desired width of the confidence interval (d)
  - The confidence interval (CI)

Sample size for proportion estimation

- $n = (CI)^2 \times 4\, p(1-p)/ d^2$
  - For 95% confidence interval, CI = 1.96
  - For 90% confidence interval, CI = 1.64
  - For 99% confidence interval, CI = 2.58

Sample size for comparative studies

- Sample size calculation aim at minimize the type I and type II error
- Type I error – we conclude that there is a difference between groups in the target population when in fact there is not.
- Type II error – We conclude that there is no difference between the groups in the target population when in fact there is a real difference.

Sample size for comparative studies

- Clinically important difference – the minimum clinically important difference is determined by physician who would not want the study miss it. It should be clinically meaningful but not statistically meaningful
- Significance level (α) – it is the probability of type I error
- The power of the test ($1 - \beta$) – where β is the probability of type II error

Sample size for comparing means

- The information required:
  - The standard deviation (SD) of the measure being compared
  - The minimum difference (d) that is clinically important
  - The significance level (α)
  - The power of the test ($1 - \beta$)
  - n is the number of sample in each group
- $n = 2K(SD)^2 / d^2$

Sample size for comparing proportions

- The information required:
  - The expected population proportions in group 1, P_1
  - The expected population proportions in group 2, P_2
  - The significance level (α)
  - The power of the test ($1 - \beta$)
  - n is the number of sample in each group
- $n = K[P_1(1- P_1) + P_2(1- P_2)] / (P_1 - P_2)^2$

Value of K corresponding to α and β

Power $(1 - \beta)$	Significance level (α)		
	5%	1%	0.1%
80%	*7.8*	*11.7*	*17.1*
90%	*10.5*	*14.9*	*20.9*
95%	*13.0*	*17.8*	*24.3*
99%	*18.4*	*24.1*	*31.6*

Assumption for sample size calculation

- There was no sample lost till the end of study once the samples was selected.
- There are equal sample size in each group of comparison
- Samples are simple random sample selected at individual level
- It is not applied to cluster sampling and regression
- The samples are large enough (> 50 in each group) to use large sample methods for analysis

Other issues

- We can select more sample if we expected that there are sample lost during study
- Software can be used to calculate sample size
- Is calculation always needed:
  - Not if the study is a qualitative study
  - Not always for a small survey
  - Not always for a pilot study

SAMPLING METHODS

Different types of sampling

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling
- Quota sampling
- Convenience sampling

Simple Random Sampling

- You should have whole list of subjects in sample space
- Label each subject with a number
- If you want n sample from sample space, select n random number form computer or random number table
- Select the sample label with the random number selected.

Simple Random Sampling

- Advantage
  - Gold standard of sampling
  - No detail of each sample is needed
  - Each individual has equal chance of being selected
- Disadvantage
  - You need to know the whole list in sampling frame

Systematic sampling

- We got whole list in sampling frame and label the subject by natural number
- Calculate interval *I* = population size / sample size
- Randomly choose a number less than *I* from computer program or random number table and we select the first sample which label by the random number selected
- Then, select another sample by interval *I* till the end of the list.

Systematic sampling

- Advantage
  - Easy to perform
  - If the list ranked in a group followed by another, it got the effect of stratified sampling
- Disadvantage
  - Cause sampling error if the list order in particular way (say, there must be a man followed by a woman)

Stratified Sampling

- Put each subject into one of the strata in which each stratum must be mutually exclusive and all strata must include all subject.
- The characteristics of subjects in particular stratum should be homogenous.
- From each stratum, select a proportion of sample by simple random sampling or systematic sampling and form subsample (the proportion = sample size / population size)
- Combined all subsample.

Stratified Sampling

- Advantage
  - We will not miss sample from minority group
  - Minimal sampling error
- Disadvantage
  - Time and manpower consuming
  - Some characteristic of each individual must be known
  - Resulting sample size may be larger than expected if there are too many strata, or there are some strata with few subjects only

Cluster Sampling

- Divided the population into several clusters (the difference in characteristics of subjects in each cluster should be minimal)
- Randomly selected few clusters
- From each cluster, collect the list of sampling frame and selected certain number of sample from each cluster by simple random sampling, systematic sampling or stratified sampling.
- Combine all subsample

Cluster Sampling

- Advantage
  - Whole list of the population is not necessary
  - Decreased traffic time
- Disadvantage
  - Increased sampling error
  - Different method of sample size calculation

Quota Sampling

- Each interviewer assigned certain number of sample to be interviewed (location of selecting the sample may also be fixed)
- Certain proportion of characteristics are required (say, half should be female)
- If the number of sample achieved, the sampling process ended
- There was no randomization

Convenience Sampling

- There was no limitation
- You can select any sample you like and agree to be interviewed or join the test
- Sampling error
- Convenience sampling should be avoided if possible

Randomization

- Randomization reduces bias by equalising so-called factors (independent variables) that have not been accounted for in the experimental design
- It can be achieved by random number table, or generated by computer (like Excel)
- Pseudo-random number generated by computer

DESCRIPTIVE STATISTICS

Descriptive Statistics

- It described the main features of a collection of data quantitatively
- It aimed at summarize the dataset
- There are 4 degrees:
  - Location
  - Spread
  - Skewness
  - Kurtosis

Location

- It is the first degree
- Mean: the arithmetic means or expected value of random variables
- Median: the value separating the higher half of a sample from the lower half
- Mode: The most common value among the group
- In normal distribution: mean = median = mode

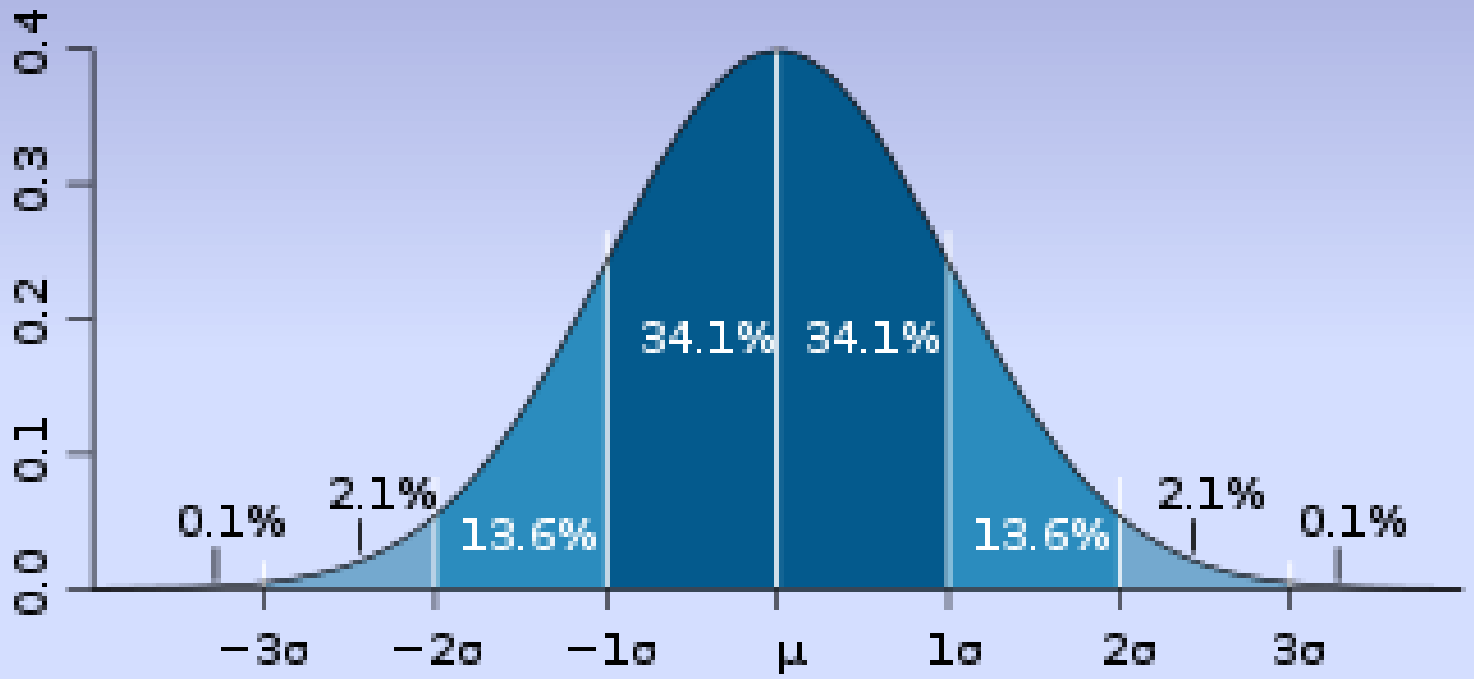
Spread

- It is a second degree
- Standard deviation: related to mean

$$\sigma = \sqrt{\mathrm{E}\left[(X - \mu)^2\right]}.$$

- Range and Percentile
- Interquartile range: related to median
  - It contain half of the sample inside the range
  - Upper quartile: separate the higher ¼ and lower ¾
  - Lower quartile: separate the lower ¼ and higher ¾

Standard Deviation



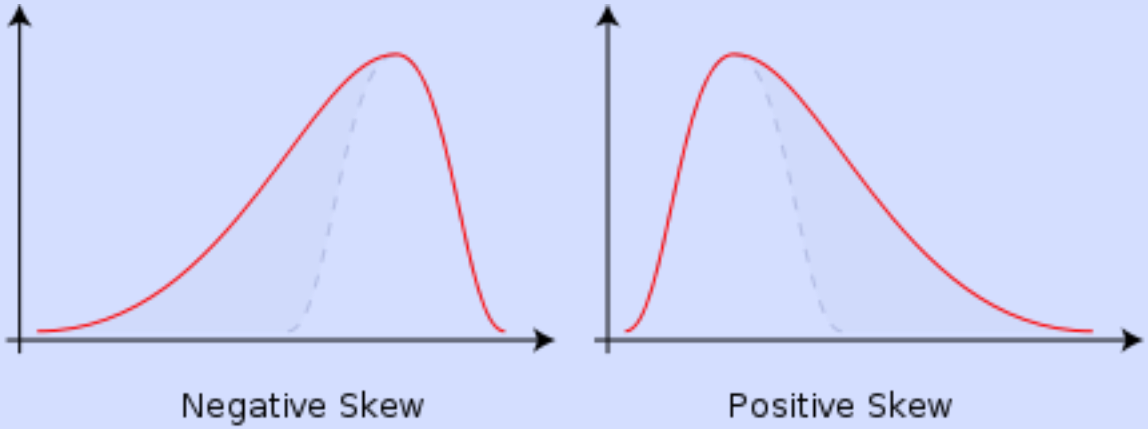
Skewness

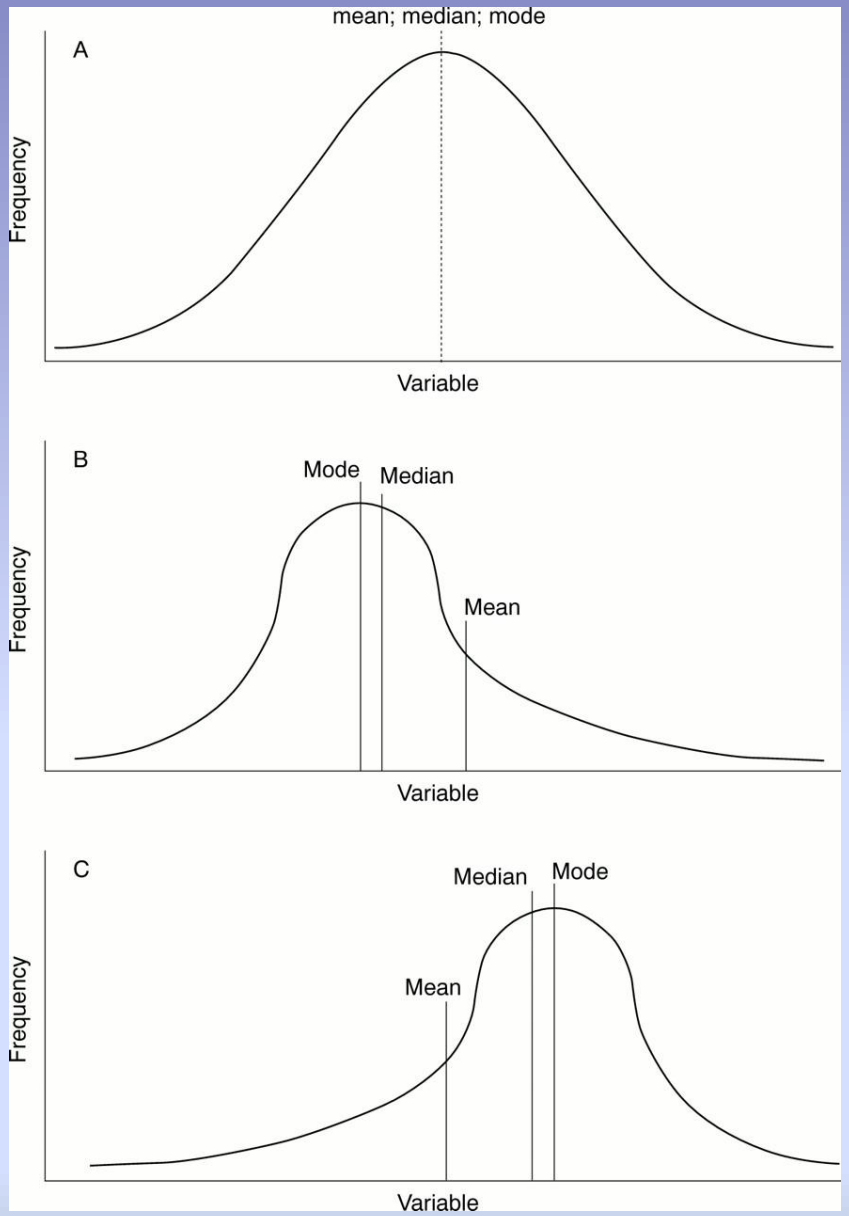
- It is the third degree
- It measure the asymmetry of the distribution
- It is calculated by

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

Skewness

- Looked at the tail of the distribution
- Positive (right) skewed (mean > median > mode)
- Negative (left) skewed (mean < median < mode)

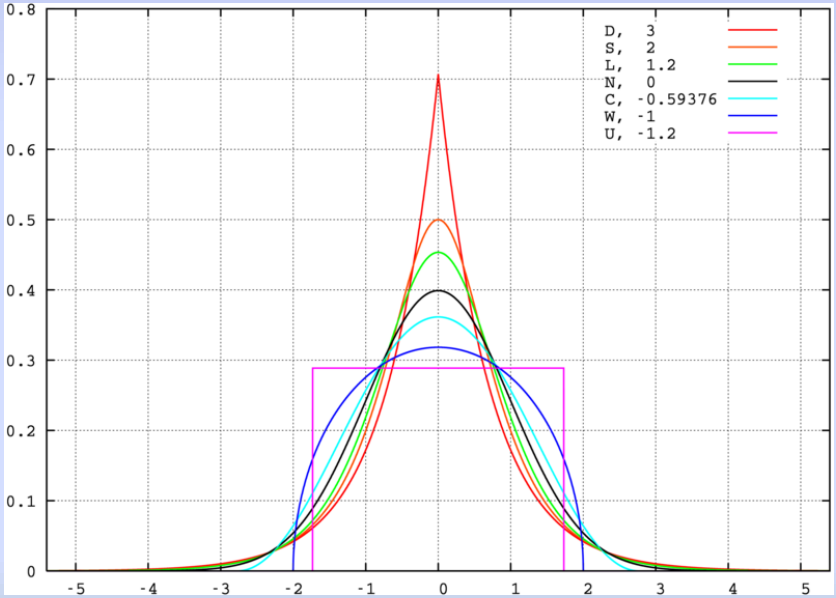



A
mean; median; mode
Frequency
Variable
B
Mode
Median
Mean
Frequency
Variable
C
Median
Mode
Mean
Frequency
Variable

Kurtosis

- It is the fourth degree
- It is a measure of "peakedness" of the distribution

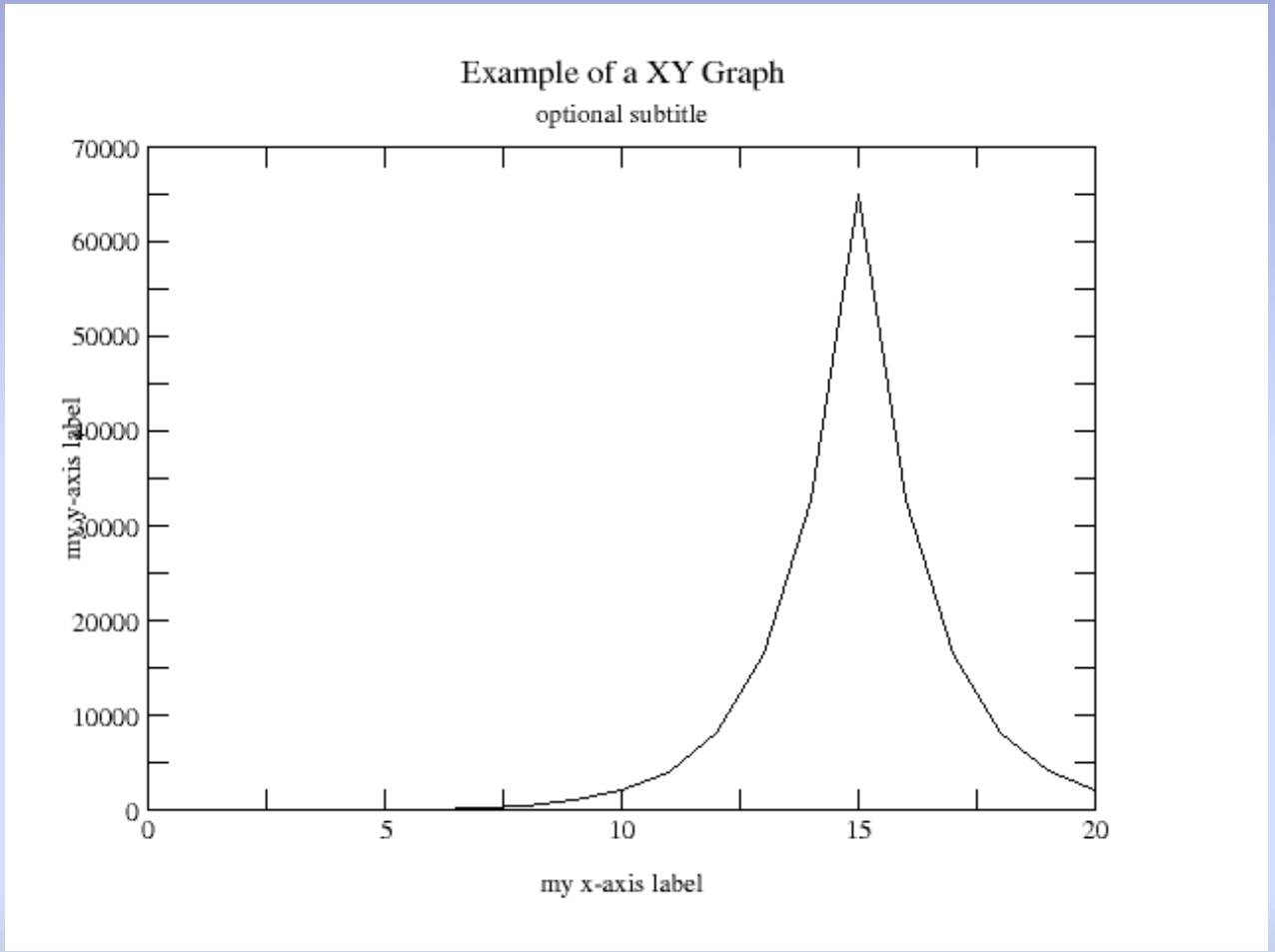
$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$



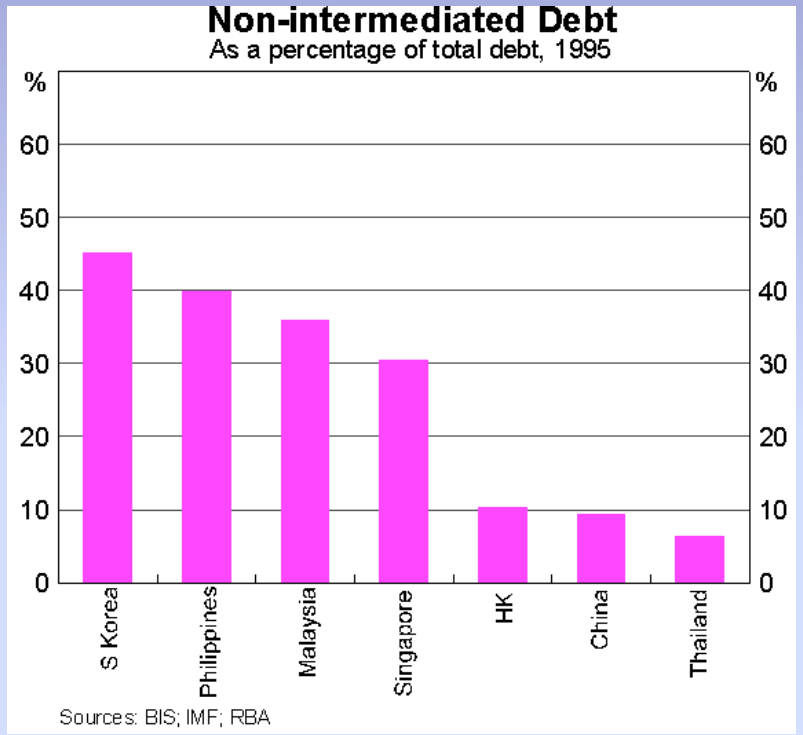
Visualize the data

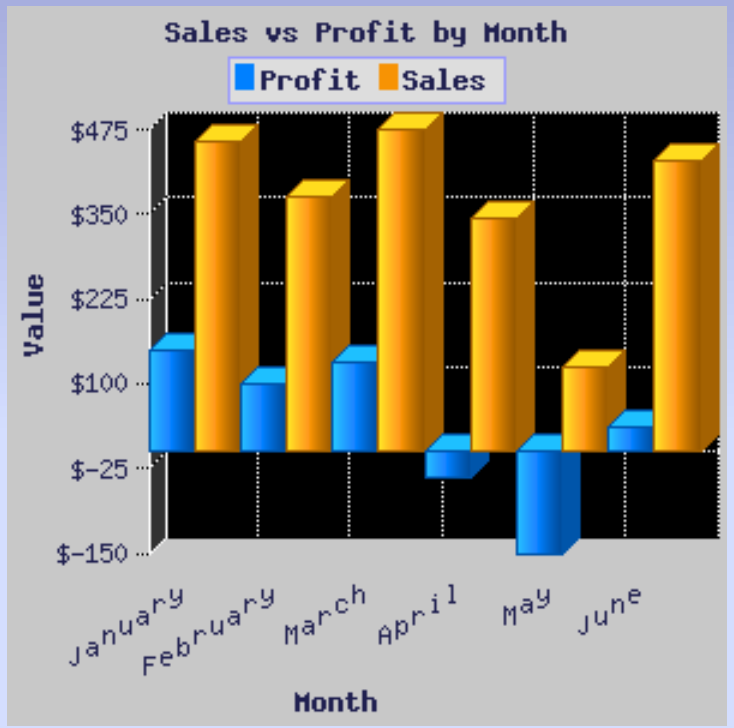
- Graph
- Histogram
- Boxplot
- Stemplot
- Scattergram

Graph

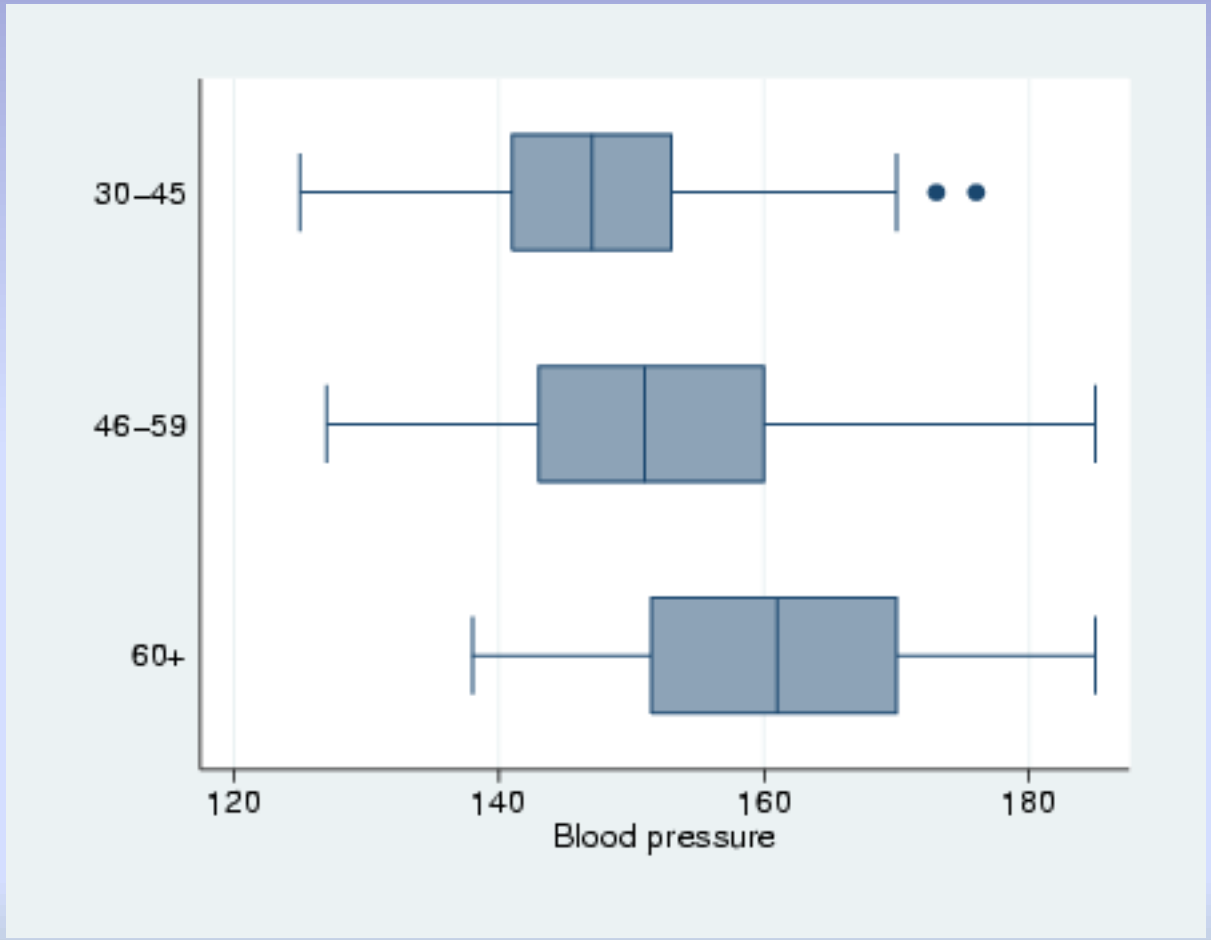

Example of a XY Graph
optional subtitle
70000
60000
50000
40000
30000
20000
10000
0
my y-axis label
0
5
10
15
20
my x-axis label

Histogram


Non-intermediated Debt
As a percentage of total debt, 1995
%
60
50
40
30
20
10
0
S Korea
Philippines
Malaysia
Singapore
HK
China
Thailand
Sources: BIS; IMF; RBA


Sales vs Profit by Month
Profit
Sales
$475
$350
$225
$100
$-25
$-150
Value
January
February
March
April
May
June
Month

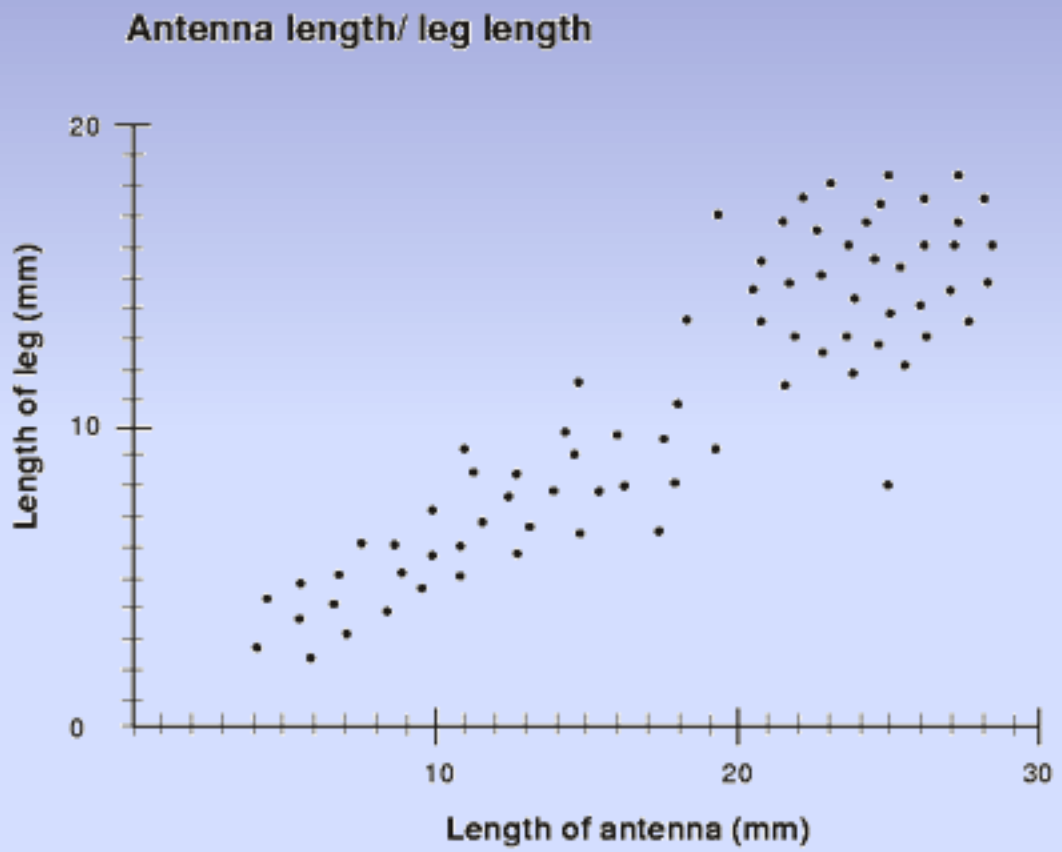
Boxplot



Stemplot

72	48	
54	59	
9	50	
9	51	
9	52	6
321	53	7
986	54	5
77755443221	55	556
977	56	024789
96220	57	23466679
766544422211000	58	00034444556678889
99988877643333210	59	01223355567788899
743211110	60	0001223344445677789
887655544322110	61	22224444555566777789
999988875544322110	62	00112223334444667777889999
98665543333321000	63	0011223455566888999
9888877665543332211110	64	000001122245566689
985544200	65	12334677789
99866422220	66	00478
77664	67	00136
9331	68	8
8210	69	
	70	
9	71	
	72	
3	73	

Scattergram



Normal Distribution

- How to realize the data is normally distributed or not from descriptive statistics
  - Mean = median = mode
  - ~~Mean = median~~ (It can occur in bimodal symmetrical distribution)
  - Mean and standard deviation (if 2 standard deviation from the mean is out of the normal range of the variable, it is unlikely to be normally distributed)
  - histogram

DATA TYPES AND DATA MANAGEMENT

Types of Data

- Either quantitative or categorical
- A variable is a quantity that is measured or observed in an individual and which varies from person to person
- A statistics (test statistics) is any quantity that is calculated from a set of data

Quantitative data

- Can be measured numerically
- Continuous data – lie on a continuum and can take any value between two limits
- Discrete data – can take certain value (count)
- Interval scale – differences between values at different points of the scale have the same meaning
- Ratio scale – ratio of two measures has same meaning (temperature scale)
- Ordinal data – data values can be arranged in a numerical order from the smallest to the largest.

Categorical data

- Individuals fall into a number of separate categories or classes
- Number was assigned for coding purpose, it can be ordinal but not interval data because the “distance” has no real measure
- Distinguish between ordered and non-ordered data
- Dichotomous data – only 2 classes
- Categorizing continuous data
  - Dichotomizing – potentially very problematic
  - Into several groups – less effect on statistical power when compared with dichotomizing

Data Entry

- Give a unique identifier to each subject
- Code the data before entry (remember the meaning of the code)
- Excel or statistical program
- Can be directly transfer form electronic data form
- Double entry can reduce the rate of error during data transfer

Data storing and transporting

- Paper forms
  - File systematically to allow easy access
  - Stored securely
  - Store identifying detail separately from data form
  - Keeping a copy in one location if they are being transported
  - If data are valuable, a copy should always be kept in a different place.

Data storing and transporting

- Electronic files
  - Identifying details should always be removed when transporting the file
  - Password protected and encrypted
  - Keep more than one copy of the data in two separate locations
  - Useful to use file names that show the version or date where files are updated

Data Checking

- Check early
- Range checks
- Consistency
- Original forms
- Missing data
- Snowballing errors
- Digit preference (round up problem)
- Scatter plot
- Statistical analysis

Correcting errors

- Check the original data form wherever possible to identify the source of potential data error
- Not to make assumptions or guesses where data values look unusual or are missing
- An outlying value should not be deleted simply because it is unusual.
- Keep a record of any changes that are made to the dataset and keep dated copies of datasets as changes are made.

STATISTICAL TEST AND ITS INTERPRETATION

HYPOTHESIS

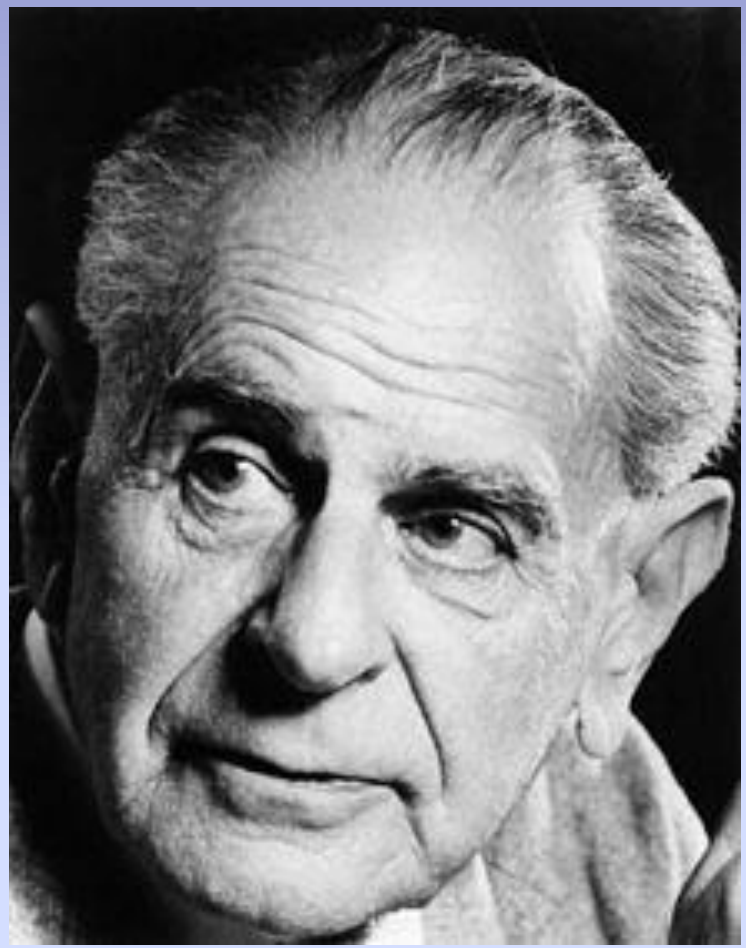
Hypothesis test

- We decide that we should “reject” the hypothesis or not.
- If we want to know whether A is true
- We set a null hypothesis – A
- Then, by means of rejecting null hypothesis to prove A is true
- Why?

Hypothesis test

- We cannot check all the case in the world to prove the hypothesis is true
- But once there is a contra-example, we can disprove a statement or hypothesis
- So, it is easier to reject a null hypothesis rather than test all the case in the world to check a hypothesis
- It is the concept of Falsifiability or Refutability.

- ***Sir Karl Raimund Popper*** (1902 – 1994)
- From 1930 to 1936, he taught secondary school. Popper published his first book, *Logik der Forschung* (*The Logic of Scientific Discovery*) in 1934, in which he introduce the concept of Falsifiability.



Null / Alternative Hypothesis

- The **null hypothesis** typically corresponds to a general or default position, that are capable of being proven false using a test of observed data.
- It is typically paired with a second hypothesis, the alternative hypothesis, which asserts a particular relationship between the phenomena.
- It is important to understand that the ***null hypothesis can never be proven***. Your data can only **reject** a null hypothesis or **fail to reject it**.

Hypothesis of one sided test

$$H_0 : M_A \leqq M_B$$
$$H_1 : M_A > M_B$$

Where H_0 is null hypothesis

and H_1 is alternative hypothesis

Hypothesis of one sided paired test

$$H_0 : M_{A-B} \leqq 0$$
$$H_1 : M_{A-B} > 0$$

Where H_0 is null hypothesis

and H_1 is alternative hypothesis

Hypothesis of two sided test

$$H_0 : M_A = M_B$$
$$H_1 : M_A \neq M_B$$

Where H_0 is null hypothesis

and H_1 is alternative hypothesis

Hypothesis of two sided paired test

$$H_0 : M_{A-B} = 0$$
$$H_1 : M_{A-B} \neq 0$$

Where H_0 is null hypothesis

and H_1 is alternative hypothesis

Reject the hypothesis

- The hypothesis is rejected if a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis were true.
- In case of one way testing, 5% in one side
- In case of two way testing, 2.5% in each side.

Why 5% is chosen?

- Fisher played a major role in the canonization of the 5% level as a criterion for statistical significance.
- In 1925, in his book “Statistical Methods for Research Workers” he fixed 5% as the only significance level in Table VI (F-distribution).
- 5% is arbitrary , as Fisher knew, but fulfils a general social purpose.

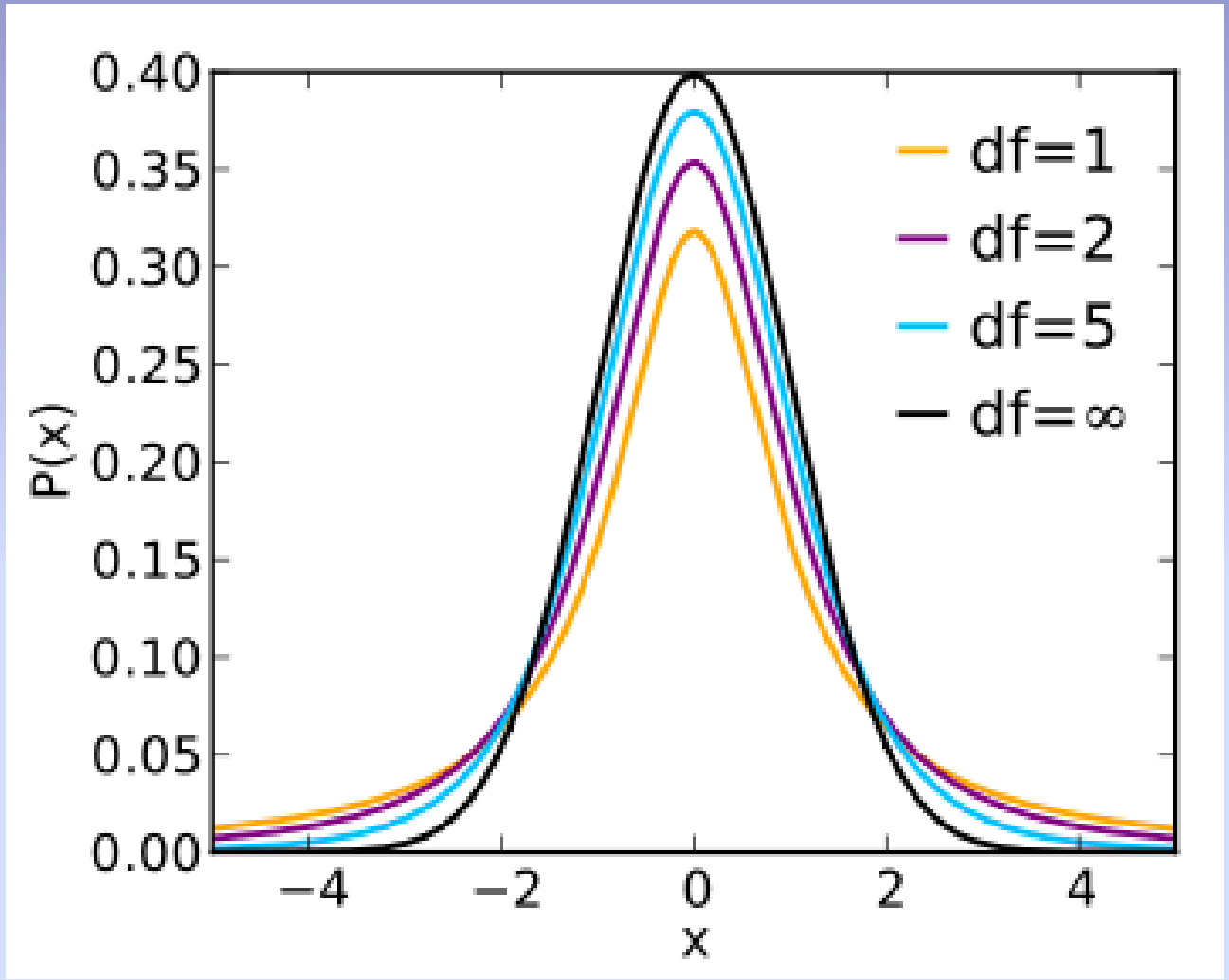
P value

- P value is the probability of observing data as extreme or more extreme than that observed, if the null hypothesis is true
- $P < 0.05$ is usually regarded as statistically significant
- Not significant does not mean “there is no difference” or “there is no effect”. It means there is insufficient evidence for a difference or effect
- Exact p values should be given with estimates and confidence intervals wherever possible.

T TEST

History

- The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working in Dublin, Ireland ("Student" was his pen name).
- He published the test in *Biometrika* in 1908


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
P(x)
−4
−2
0
2
4
x
df=1
df=2
df=5
df=∞

T-test for 2 independent means

- Details of the test
  - Compares means from 2 independent sample
  - Based on sampling distribution of difference of two samples
  - Allow calculate a difference and confidence interval of the difference
  - Can be calculated by formula or statistical program

T-test for 2 independent means

- Null hypothesis
  - Two samples come from population with same means
- Assumptions of test
  - Continuous data with normal distribution
  - Variances are the same

T-test for 2 independent means

- If assumptions do not hold
  - The statistical test is dubious and the p value may be wrong
  - Try transformation of the data
  - It is robust to slight skewness (2 samples with same size) but is less robust if variances are clearly different
  - Skewness and different variance can be corrected by transformation.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

Where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Degree of freedom

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

T test for paired (matched) data

- Also called one sample t-test
- It analyses mean difference in paired sample
- Null hypothesis: means difference is zero
- Assumption
  - differences follow a normal distribution
  - variance are constant
- If assumption do not hold – transform the raw data (not the difference)

$$t = \frac{\overline{X}_D - \mu_0}{s_D / \sqrt{n}}.$$

Where X_D and s_D is the average and standard deviation of the differences

The degree of freedom is *n-1*

Z TEST

Z test

- Compares proportions from two independent samples
- Based on sampling distribution of the difference of proportions
- Allow calculation of differences and a confidence interval for the difference
- Is equivalent to chi-squared test
- Can be calculated by formula and statistical program

Z test

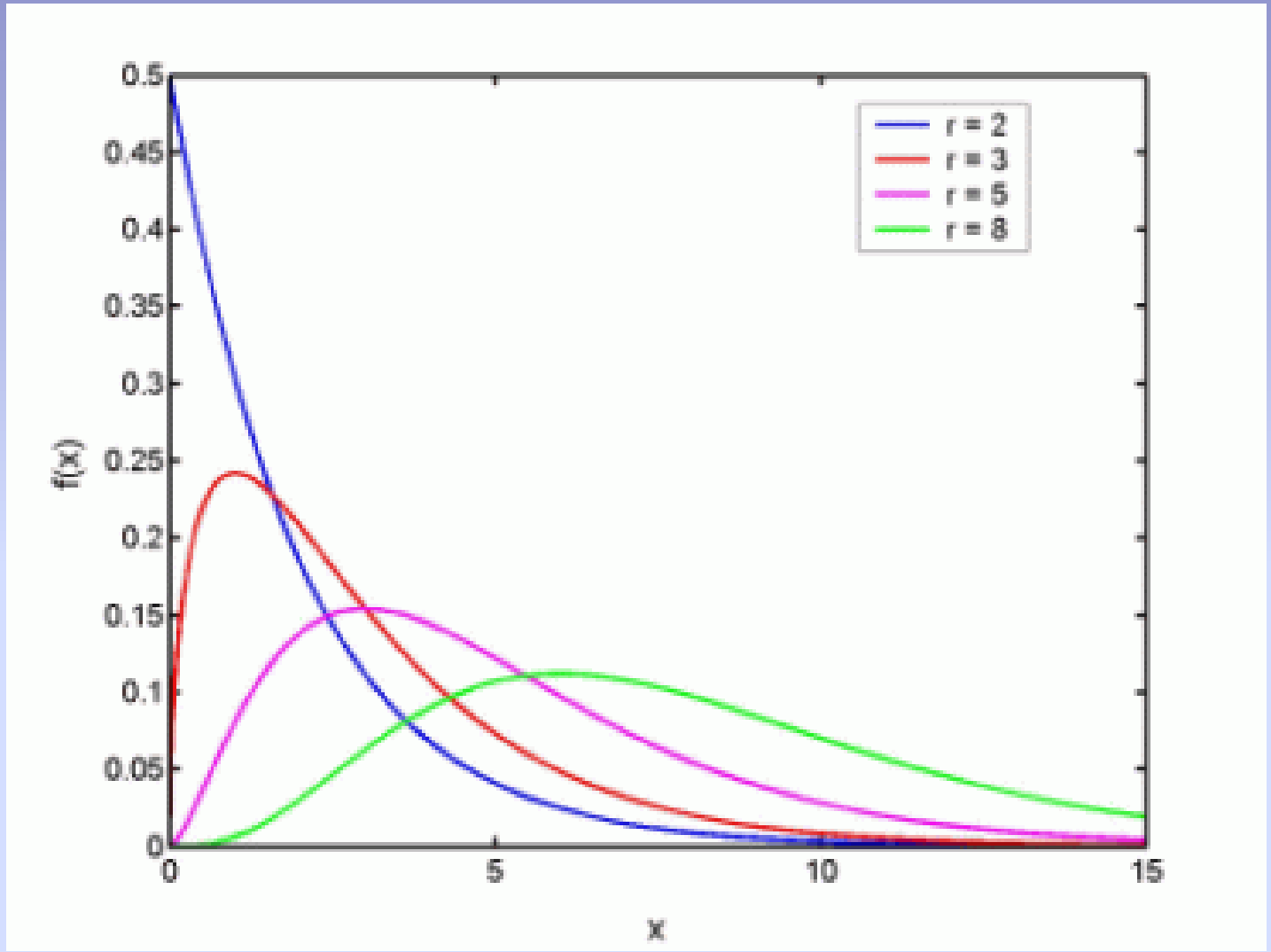
- Null hypothesis
  - Two samples come from populations with the same proportions
- Assumption of the test
  - Binary data
  - Sample is large. Both subjects with or without characteristics should be greater than 5 in both groups.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

CHI SQUARED TEST

History

- **Pearson's chi-square test** is the best-known of several chi-square tests – statistical procedures whose results are evaluated by reference to the chi-square distribution.
- Its properties were first investigated by Karl Pearson in 1900.


r = 2
r = 3
r = 5
r = 8
f(x)
x
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0
5
10
15

Chi-squared test

- Tests for association between two categorical variables
- When each variables has only 2 categories, it is equivalent to z test
- Based on the chi-squared distribution with n degree of freedom
- *n = (no. of row – 1) x (no. of column – 1)*
- It gives p value but not direct estimate or confidence interval as z test

Chi-squared test

- Rationale of test
  - Calculates the frequencies that would be expected if there was no association
  - It compares the observed frequencies and expected values
  - It they are very different, this provides evidence that there is an association
  - The test uses a formula based on chi-squared distribution to give p value

Chi-squared test

- Null hypothesis
  - There is no association between the two variables in the population form which the samples come
- Assumptions of test
  - Large sample size
  - At least 80% of expected frequencies must be greater than 5

Chi-squared test

- If assumption do not hold
  - Collapsing the table
  - Continuity correction (Yates' correction)
  - Fisher's exact test
- Doing chi-squared test
  - Always use with frequencies, never use percentage
  - The formula works with all size tables
  - Can be done by computer program

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where

X^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

O_i = an observed frequency;

E_i = an expected (theoretical) frequency, asserted by the null hypothesis;

n = the number of cells in the table.

Yates' Correction

- Chi-squared test based on frequencies (discrete) whilst the chi-squared distribution is continuous.
- The fit is not good in small sample size
- Yates' correction modified the chi-squared formula to make better fit
- Corrected p value (slightly bigger) should be reported

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where:

O_i = an observed frequency

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

N = number of distinct events

FISHER'S EXACT TEST

History

- Fisher is said to have devised the test following a comment from Muriel Bristol, who claimed to be able to detect whether the tea or the milk was added first to her cup in 1922

Fisher's Exact test

- Useful for small samples where chi-squared test is invalid
- Tests for an association between 2 categorical variables
- Normally used for 2 x 2 tables, but computer program allow for bigger tables
- Evaluating the probabilities associated with all possible tables which have the same row and column totals as the observed data, assuming the null hypothesis is true

Fisher's Exact test

- Based on exact probabilities, it is computationally intensive and may be slow or fail for large sample size.
- Give p values but not direct estimate or confidence interval

Fisher's Exact test

- Null hypothesis
  - No association between the two variables in the population from which the samples come
  - Same null hypothesis as the chi-squared test
- Assumptions of test
  - none

Fisher's exact test

- Always use with frequencies, never use percentages for calculation
- No simple formula, statistical program needed
- Unless with good reason, use the two-sided p value
- It gives p values at least as big as the chi-squared test. For large sample size, p values are similar
- If in doubt about the sample size, use Fisher's exact test instead of chi-squared test.

a	b	a + b
c	d	c + d
a + c	b + d	a + b + c + d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

CORRELATION

Pearson's correlation

- It investigate the strength of a linear relationship between two continuous variables
- It is used when neither variable can be assumed to predict the other
- It gives an estimate, the correlation coefficient and a p value
- A confidence interval can be calculated

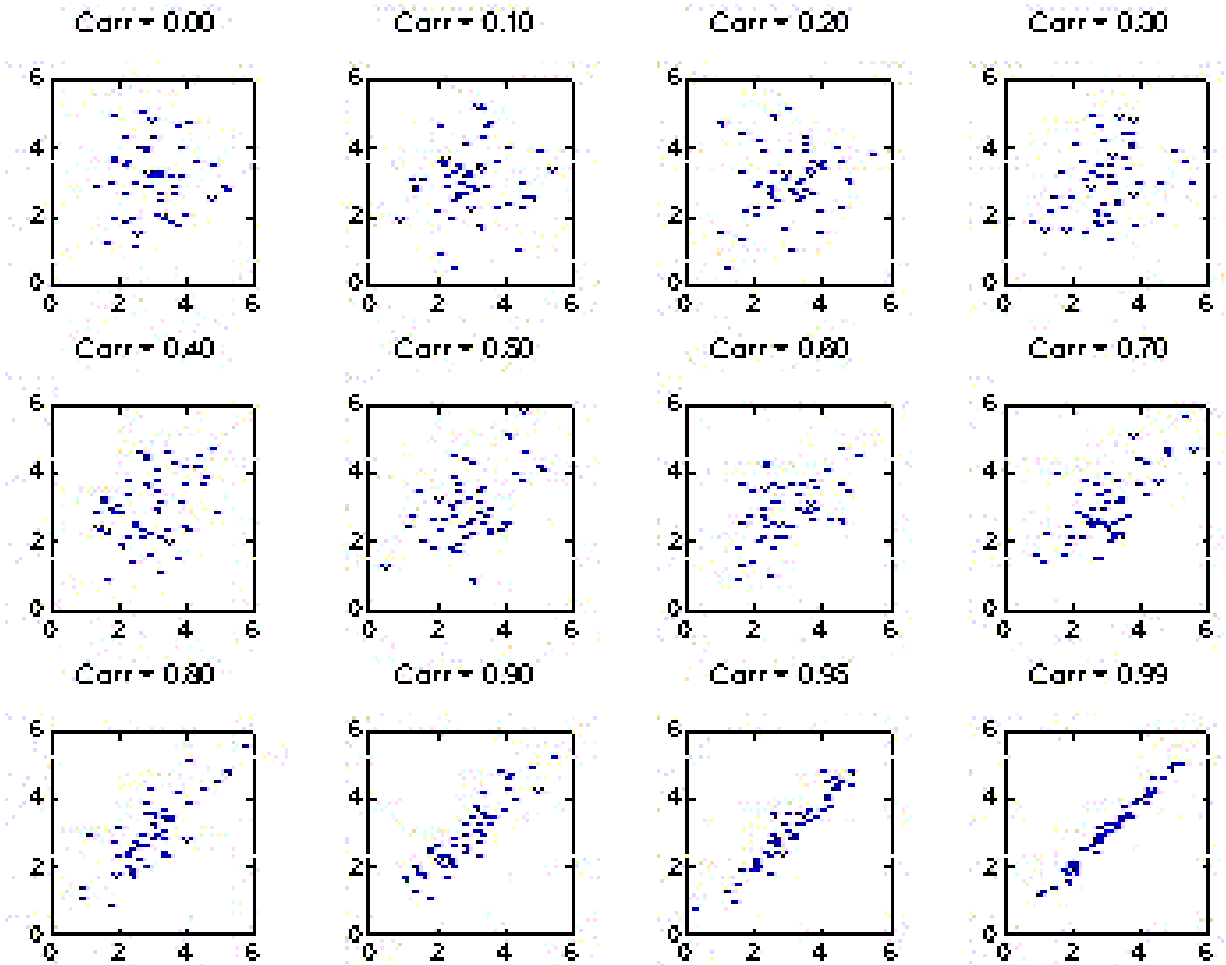
Pearson's correlation

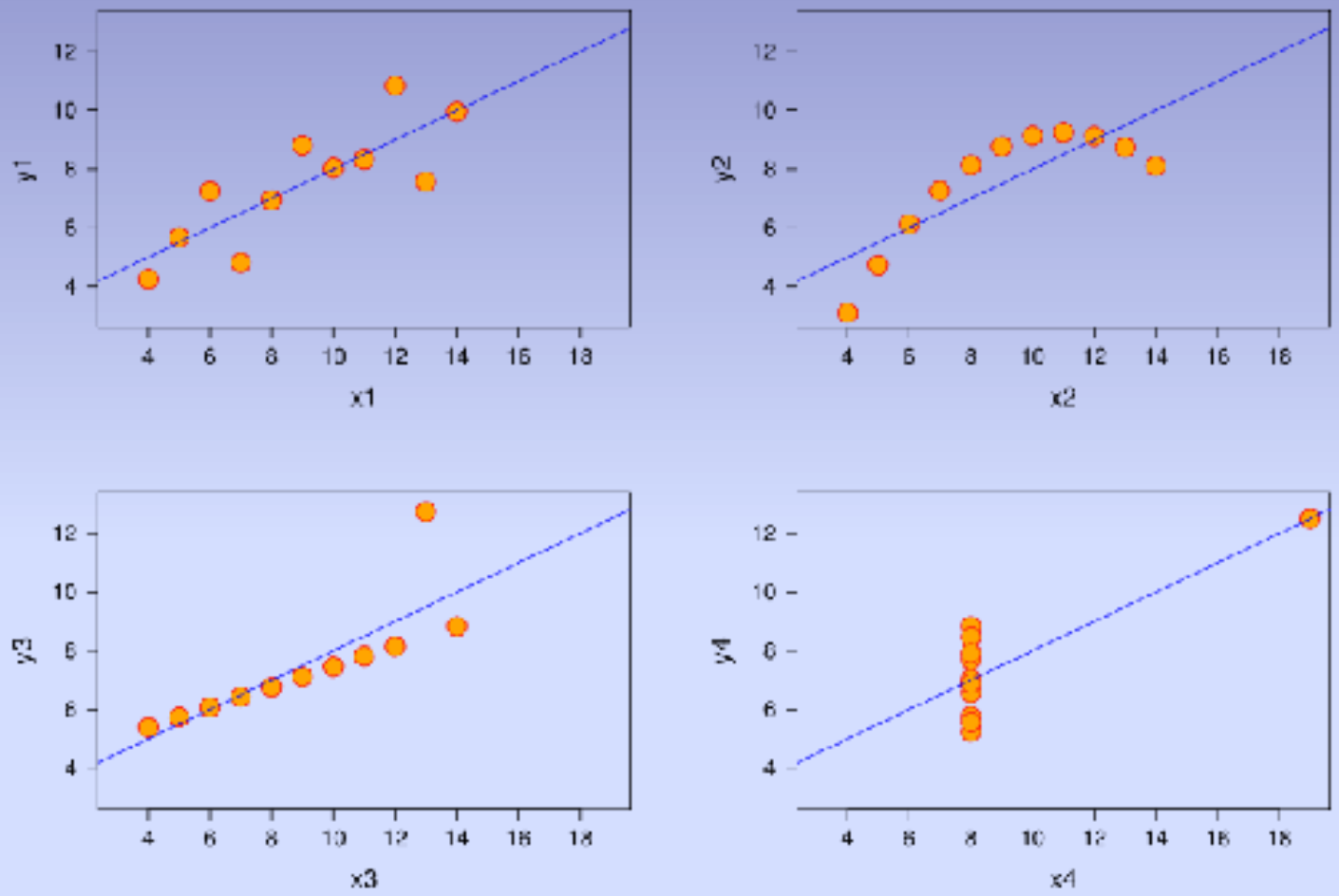
- Assumption
  - The relationship is linear
  - Normal distribution
    - For significant test – at least one variable to be normally disturbed
    - For confidence intervals – both variables should be normally distributed
  - A random sample within the range of interest

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Interpretation of *r*

- *r* tells us how close is the linear relationship between two variables
- It lies between +1 and -1
- Negative (positive) values indicate negative (positive) linear relationship
- $r = 0$ indicate that is no linear relationship
- The closer the value +1 or -1, the stronger relationship between two variables

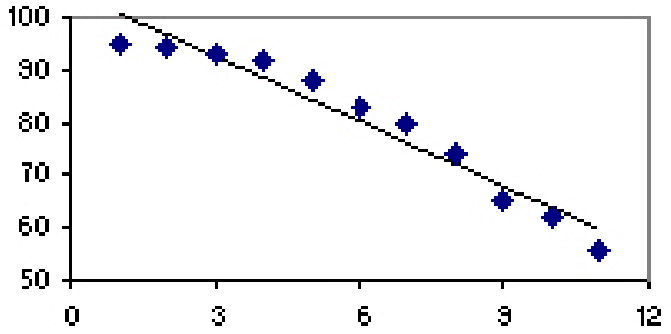

Corr = 0.00
Corr = 0.10
Corr = 0.20
Corr = 0.30
Corr = 0.40
Corr = 0.50
Corr = 0.60
Corr = 0.70
Corr = 0.80
Corr = 0.90
Corr = 0.95
Corr = 0.99


y1
x1
y2
x2
y3
x3
y4
x4

Outlier

- If outlier is removed, r is closer to +1 or -1

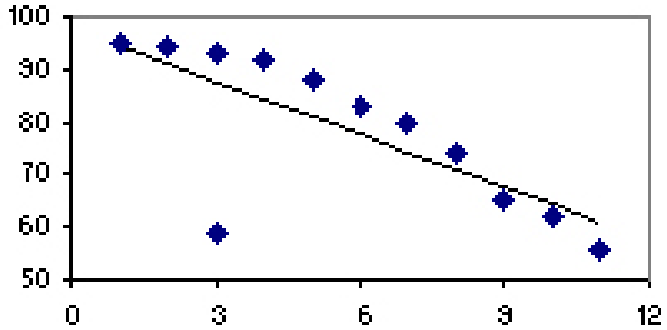
Without Outlier



Regression equation: $\hat{y}$ = 104.78 - 4.10x

Coefficient of determination: R^2 = 0.94

With Outlier

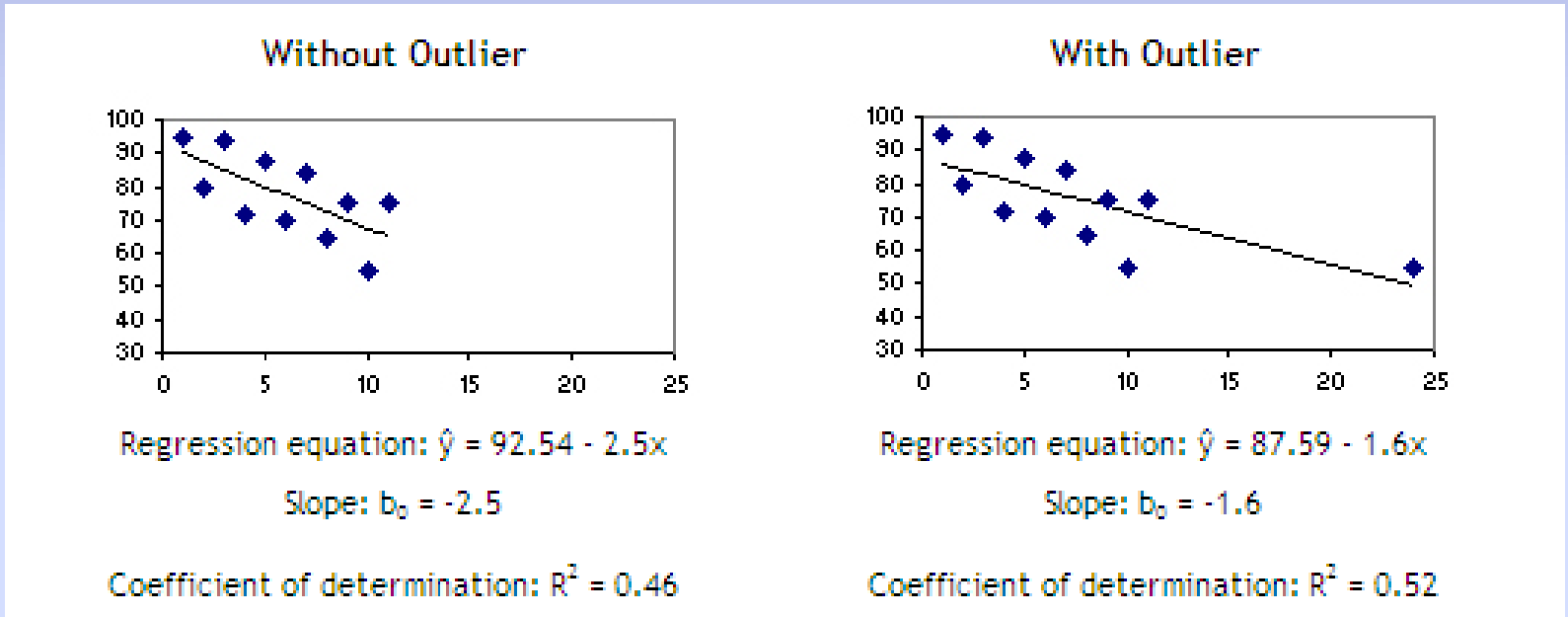


Regression equation: $\hat{y}$ = 97.51 - 3.32x

Coefficient of determination: R^2 = 0.55

Influential point

- If influential point is removed, *r* is closer to 0



Test and estimate of *r*

- A significant test can be done with null hypothesis that $r = 0$
- A confidence interval of *r* can be calculated
- Statistical significance of *r* directly related to sample size
  - If sample size is large, it may be statistically significant even the relationship is weak

LETTER TO THE EDITOR

Interpretation of correlation coefficients

A recently published article by Ko et al[1] illustrates a commonly encountered problem with the interpretation of correlation coefficients in medical research. The authors stated the correlation coefficient ranged from -0.002 to 0.15. If the P value was greater than 0.05, they interpreted the findings as indicating no correlation. Otherwise, there was weak correlation. The *t* value of the correlation coefficient was given by:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where *r* is the correlation coefficient and *n* the sample size.[2]

From this we can infer that if the sample size or correlation coefficient increases, so will the *t* value, and hence P value will decrease and thus tend to increase statistical significance. Thus, if there is a huge sample size, the result will be significant even when there is very weak association. However, if the sample size is small, the result will remain non-significant even though the correlation coefficient is large. Theoretically, if the sample size approaches infinity, the test will always be highly significant (P<0.0001) though the correlation coefficient tends to zero.

So the P value for the correlation coefficient just tells us whether the coefficient is or is not reliable, it cannot determine whether there is correlation between two factors. For example, in the article referred to, for *r*=0.15 and P=0.003, we should interpret as "the correlation coefficient is reliable and there is weak correlation between two factors". In another example with *r*= –0.002 and P=0.97, we should interpret as "there is no adequate evidence to show whether two factors are correlated or not" rather than "they are not correlated" as stated in the article.

KC Wong, FHKAM (Psychiatry)
Email: wongkc05@hotmail.com
Department of Psychiatry
Pamela Youde Nethersole Eastern Hospital
Chai Wan
Hong Kong

References

1. Ko FW, Chan AM, Chan HS, et al. Are Hong Kong doctors following the Global Initiative for Asthma guidelines: a questionnaire "Survey on Asthma Management"? Hong Kong Med J 2010;16:86-93.
2. Armitage P, Berry G, Matthews JN. Statistical methods in medical research. 4th ed. Massachusetts: Blackwell Publishing; 2002: 199.

Spearman's correlation

- It is used when none of the 2 variables follows a normal distribution – it is assumption for Pearson's correlation.
- Null hypothesis
  - There is no tendency for one variable either to increase or to decrease as the other increases
- Assumption
  - The variables can be ranked
  - Monotonic relationship between 2 variables

Spearman's correlation

- This is calculated using same formula as for Pearson's correlation but uses the ranks of the data rather than the data values themselves
- It gives a values between -1 and +1
- p value can be obtained from statistical program

MANN WHITNEY U TEST (WILCOXON TWO SAMPLE SIGNED RANK TEST)

History

- It is also called as Wilcoxon two sample signed ranked test.
- It was proposed initially by Frank Wilcoxon in 1945, for equal sample sizes, and extended to arbitrary sample sizes and in other ways by Henry Mann and his student Donald Ransom Whitney in 1947

Mann-Whitney U test

- It is based on the rank of the data
- It gives p value but no estimate
- Given a table of cut-offs, the test is easy to do by hand
- If the sample is very small (both smaller than four observations), then statistical significant is impossible.

Mann-Whitney U test

- The samples should be unrelated (independent)
- The test can be applied to ordinal data
- The population distributions should have the same shape (it tests the whole distribution)
- Tied values in the data
  - If number of ties is small, it still satisfactory
  - If data are ordinal and fall into only 3 or 4 categories, it is better to use chi-squared test

WILCOXON MATCHED PAIRS TEST

History

- It also called Wilcoxon signed rank test
- The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945)

Wilcoxon matched pairs test

- Based on the signs of the differences in the pairs and the relative sizes of differences rather than the actual values
- It gives p values but no estimate
- Given a table of cut-offs, the test is easy to do by hand
- If the sample is smaller than 6, then statistically significance is impossible

Wilcoxon matched pairs test

- The data should be paired
- The data should consist of numerical measurements (interval data)
- The distribution of the differences should be symmetrical. It is difficult to know in practice whether the distribution is symmetrical or not, but we can apply it with small sample size
- The zero difference (tie) should be omitted

SIMPLE LINEAR REGRESSION

Simple Linear Regression

- Estimate the nature of linear relationship between two continuous variables
  - Dependent (response) variable
  - Independent (explanatory) variables
- The calculation based on least squared method – It minimize the sum of the squares of these residual (= observed value – fitted values)

Simple Linear Regression

- Slope or regression coefficient is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{Cov}[x, y]}{\mathrm{Var}[x]} = r_{xy}\frac{s_y}{s_x},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

- The line goes through the mean point: (x, y)
- Therefore the intercept is given by: $a = y - bx$

Interpretation of the equation

- Regression coefficient gives the change in the outcome (y) for a unit change in the predictor variable (x)
- The intercept gives the value of y when x is 0
- The line gives the mean or expected value of y for each value of x

Simple Linear Regression

- If there is no relationship between *x* and *y*, the true regression coefficient *b* will be 0
- Can be tested using a form of *t* test
- The regression coefficient *b* is a useful summary to show how the two variables related
- 95% confidence interval can be calculated for *b*
- The equation of the line can be used for prediction

Simple Linear Regression

- Assumption
  - The relationship is linear
  - The distribution of the residuals is normal
  - The variance of the outcome *y* is constant over *x*
- Predication
  - Within sample prediction
  - Prediction outside the sample

OTHERS

Bonferroni Correction

- It is a simple method to correct the cut-off for statistical significant for multiple testing
- For $\alpha = 0.05$, if the test repeated for 20 times, there should be 1 false positive
- If there is n repeated testing, we need $(1 - \alpha)^n = 0.95$, which approximately equal to $1 - n\alpha$. So, α is approximately equal to $0.05 / n$
- Disadvantage – too conservative

Estimate for test of proportion

- Risk difference ($p_1 - p_2$)
  - Use if actual size of difference is of interest
  - Most straight forward and useful for survey
- Relative risk (p_1 / p_2)
  - Use if relative risk is of interest
  - Useful when comparing the size of effect for several factor
  - Easier to interpret
  - Do not use for case control test

Estimate for test of proportion

- Odds ratio (*ad / bc* where *a*, *b*, *c* and *d* are 2x2 tables frequencies)
  - Use for case-control studies
  - Approximately equal to relative risk when the outcome is rare
  - Can be misinterpreted when the outcome is common
  - Can adjust for others factors using logistics regression

Confidence Interval

- In statistics, a **confidence interval** (**CI**) is a particular kind of interval estimate of a population parameter and is used to indicate the ***reliability*** of an estimate. It is an observed interval, in principle ***different from sample to sample***, that frequently includes the parameter of interest, if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the **confidence level** or **confidence coefficient**.

Others

- Multiple regression analysis
- Analysis of variance
- Logistics regression
- Poisson Regression
- Kappa (inter-rater reliability)
- McNamer test
- Epidemiology
- Principle component analysis
- …….

COMMON STATISTICAL SOFTWARE

Choosing a package

- Cost
- Support
- Your institution
- Usability
- Data management
- Type of analysis
- Specialized method
- Graphics
- Size of datasets
- Transferring between packages
- Testing
- Operating system
- License versus perpetual copy
- Upgrades
- Discounted versions

Commonly used package

Free

- Epi-info (CDC)
- OpenEpi (CDC)
- R (Programming language)
- WinBUGS (for Bayesian analysis)

Commercial

- Microsoft Excel
- SPSS
- STATA
- Minitab
- SAS
- STATISTICA

POST TEST

Variable	OCD (n = 50)	Schizophrenia (n = 47)	X^2 (degrees of freedom), p value
Mean (standard deviation) age (years)	29 (9)	36 (11)	514 (506), 0.38
Sex			
Male	37 (74)	38 (81)	1.29 (2), 0.52
Female	13 (26)	9 (19)	
Marital status			
Single	14 (28)	10 (21)	
Married	34 (68)	36 (77)	1.66 (4), 0.79
Widowed	2 (4)	-	
Separated	-	1 (2)	
Occupation			
Professional	13 (26)	6 (13)	
Clerical / shop owner	3 (6)	6 (13)	
Farmer	2 (4)	1 (2)	
Skilled worker	10 (20)	9 (19)	
Semi-skilled / unskilled worker	1 (2)	14 (30)	38.8 (48), 0.82
Unemployed	10 (20)	6 (13)	
Housewife	-	1 (2)	
Retired	11 (22)	4 (9)	

TABLE 3. Univariate analysis of potential factors affecting preferences for developing a Health Call Centre

Characteristic	Chi squared test			
	Agree to develop a Call Centre % (No.)	Disagree to develop a Call Centre % (No.)	X^2	P value
Gender				
Male	49.3 (168)	51.7 (31)	0.118	0.780
Female	50.7 (173)	48.3 (29)		
Age (years)				
18-30	6.8 (23)	3.3 (2)	36.269	<0.001
31-64	78.5 (266)	48.3 (29)		
≥65	14.7 (50)	48.3 (29)		
Education				
Primary school or below	28.1 (96)	61.0 (36)	25.784	<0.001
Secondary school	58.2 (199)	27.1 (16)		
University or above	13.7 (47)	11.9 (7)		
Marital status				
Married	82.5 (282)	81.7 (49)	0.022	0.856
Others	17.5 (60)	18.3 (11)		
Working status*				
Working	48.2 (164)	31.7 (19)	5.641	0.024
Not working	51.8 (176)	68.3 (41)		
Individual monthly income				
No income ($0)	38.9 (128)	63.2 (36)	11.694	0.001
Received income	61.1 (201)	36.8 (21)		

Health status				
No difference – much better	68.0 (230)	56.9 (33)	2.760	0.100
Worse – much worse	32.0 (108)	43.1 (25)		
Being cared for by other people				
Yes	9.6 (33)	20.0 (12)	5.501	0.026
No	90.4 (309)	80.0 (48)		
Caregiver				
Yes	19.7 (67)	8.5 (5)	4.288	0.043
No	80.3 (273)	91.5 (54)		
Health insurance coverage				
Yes	32.9 (112)	18.3 (11)	5.111	0.023
None	67.1 (228)	81.7 (49)		
Fee waiver				
Yes	27.5 (94)	28.3 (17)	0.018	0.877
None	72.5 (248)	71.7 (43)		

Types of fee waiver[†]				
CSSA/disability allowance	26.7 (24)	56.3 (9)	5.545	0.037
Civil servant/HA benefits	73.3 (66)	43.8 (7)		
Presence of chronic disease				
Yes	41.3 (141)	31.7 (19)	1.995	0.198
None	58.7 (200)	68.3 (41)		
Treatment on health problem last time				
Yes	2.9 (10)	6.7 (4)	2.127	0.141
No treatment	97.1 (332)	93.3 (56)		
Methods of solving health problem last time				
Ask health professional/others	80.1 (274)	83.3 (50)	0.781	0.677
Self (internet/books/others)	7.6 (26)	8.3 (5)		
Both	12.3 (42)	8.3 (5)		
Used health hotline before				
Yes	2.7 (9)	1.7 (1)	0.204	1.000
No	97.3 (330)	98.3 (59)		

TO SUM UP

You should learn…

- Basic concept of hypothesis test
- Interpretation of the test result
- Sampling method
- Identify and avoid common mistake in medical statistics
- At least, you should know when, how and where to seek advice

- ***Oxford Handbook of Medical Statistics*** JL Peacock, PJ Peacock. Oxford University Press. 2011
- ***Statistical Methods in Medical Research*** 4th edition. P Armitage, G Berry, LNS Matthews. Blackwell Publishing. 2002
- ***Elementary Statistics Tables.*** HR Neave. Routledge. 1989
- http://www.cct.cuhk.edu.hk/stat/
  - Sample size estimation (CUHK)
- http://udel.edu/~mcdonald/statintro.html
  - Handbook of Biological Statistics
- http://www.york.ac.uk/depts/maths/histstat/
  - History of Statistics (The University of York)

To sum up

- Now, you may forget half of the content
- You will forget 90% of the talk within 1 week
- Learn it and revise it if you use it
- Consult statistician if necessary (better at the beginning of the study)
- Compared with mathematics, medical statistics is a “kid”.
- My website: www.kcclinic.com
- My e-mail: kc@drkcwong.com

Thank You

Q &A