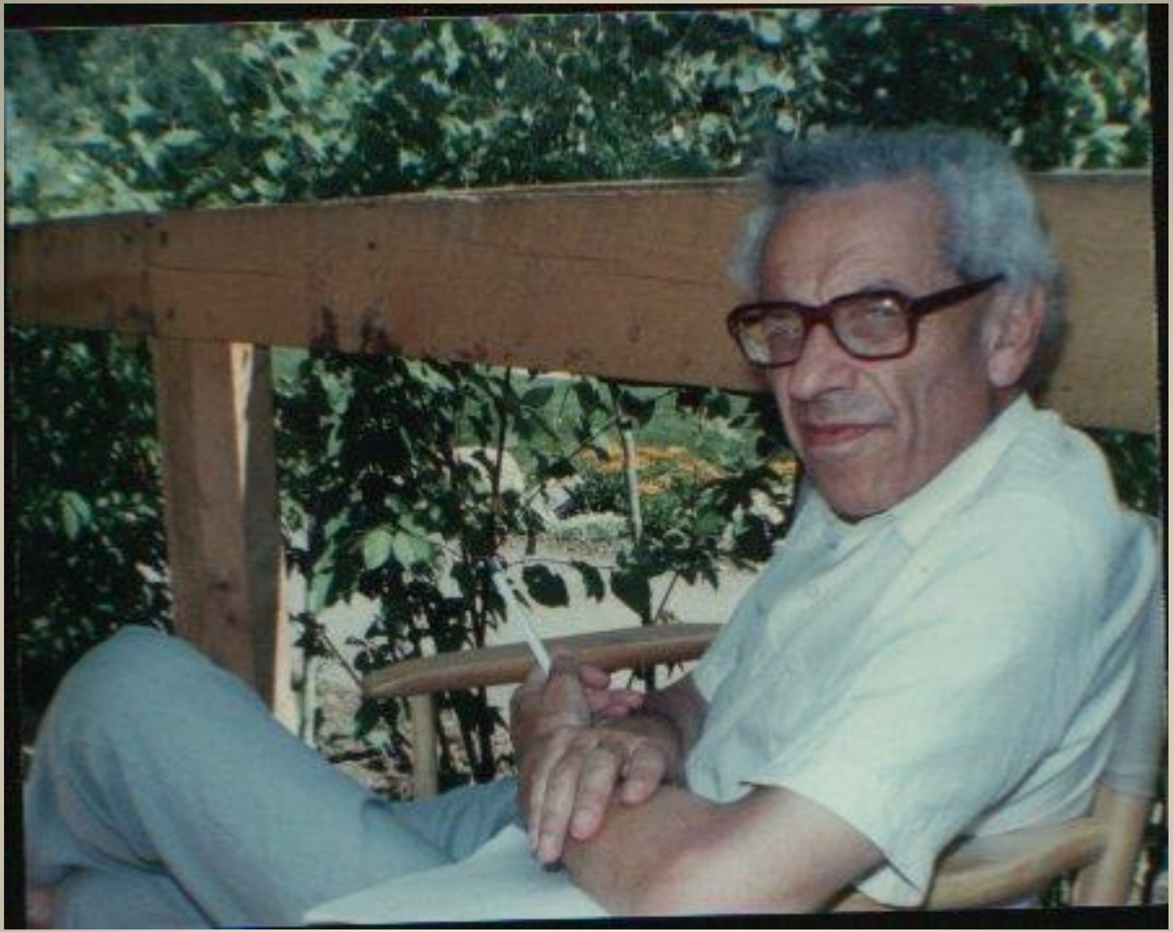
Medical Statistics relevant to Psychiatrists

Dr. Wong Kai Choi

28th October 2011



Paul Erdos (1913 – 1996)

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

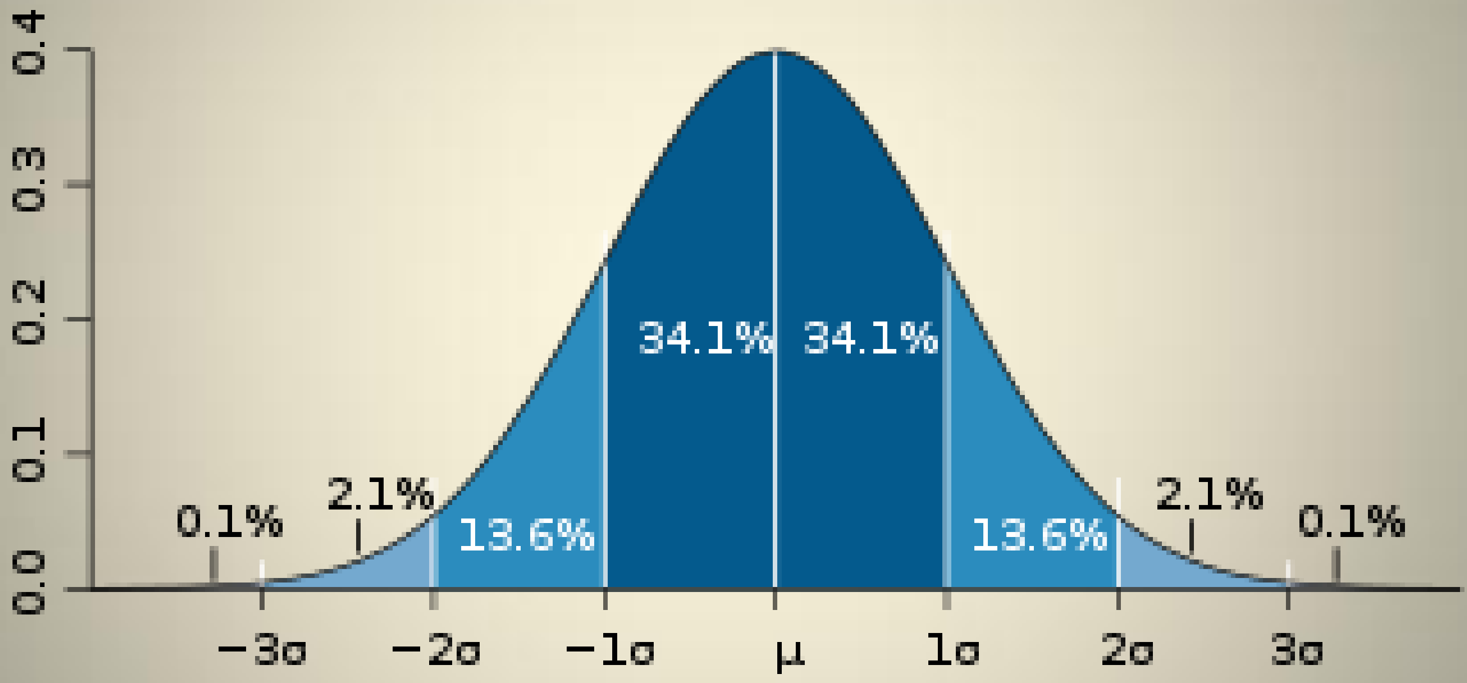
Spread

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

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

- 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



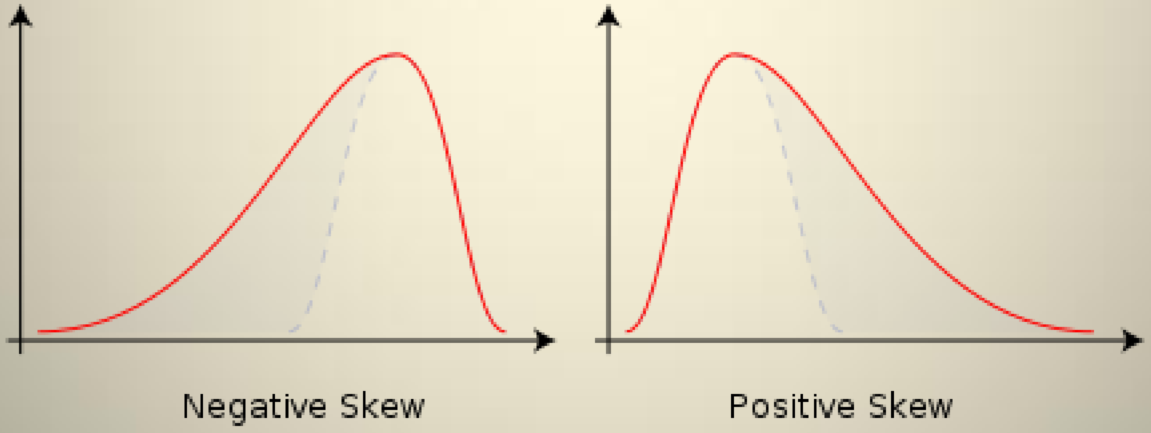
Skewness

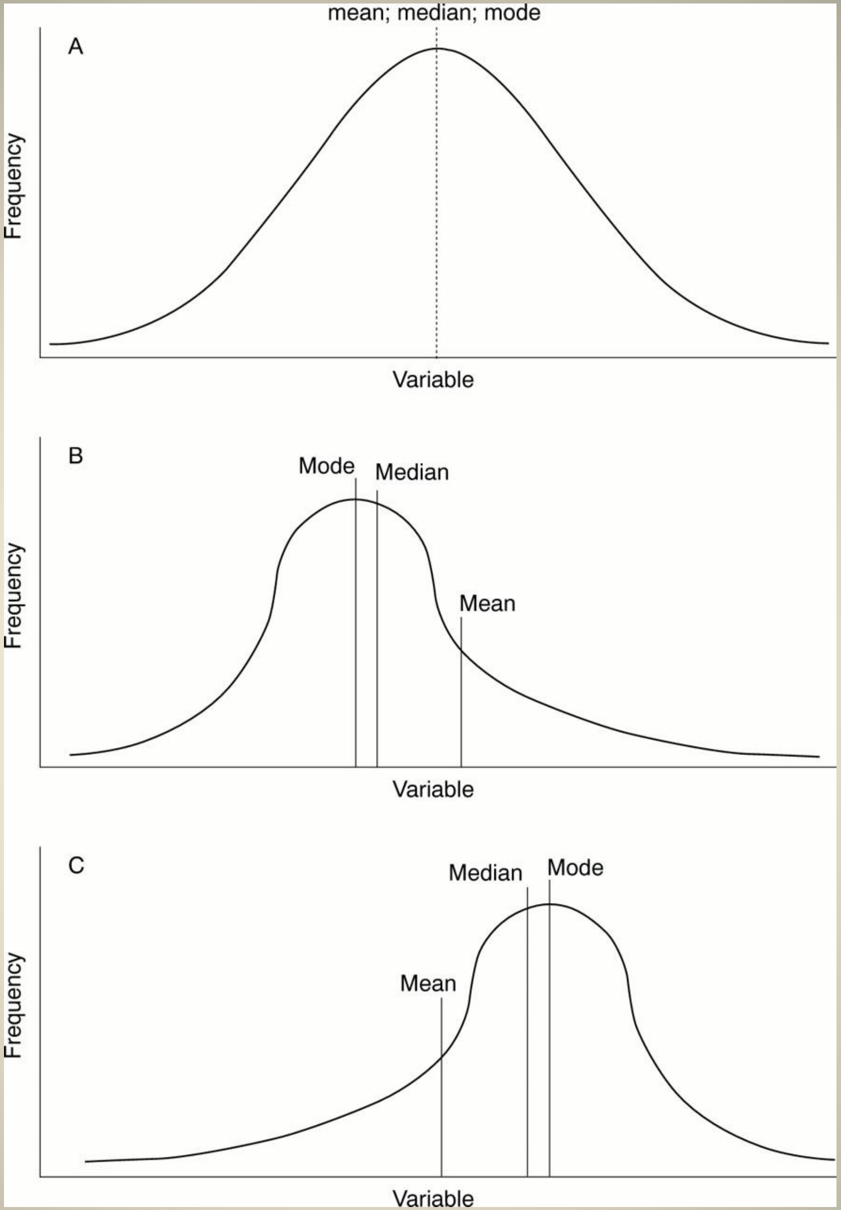
- It is the third degree
- It measure the asymmetry of the distribution
- It is calculated by

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

Skewness

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

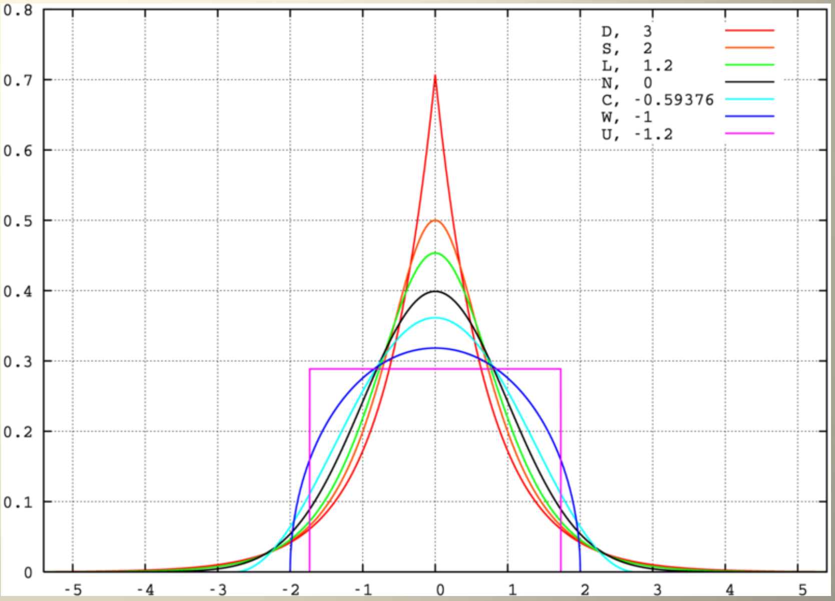



mean; median; mode
A
Frequency
Variable
B
Mode
Median
Mean
Frequency
Variable
C
Median
Mode
Mean
Frequency
Variable

Kurtosis

- It is the fourth degree
- It is a measure of “peakedness” of the distribution

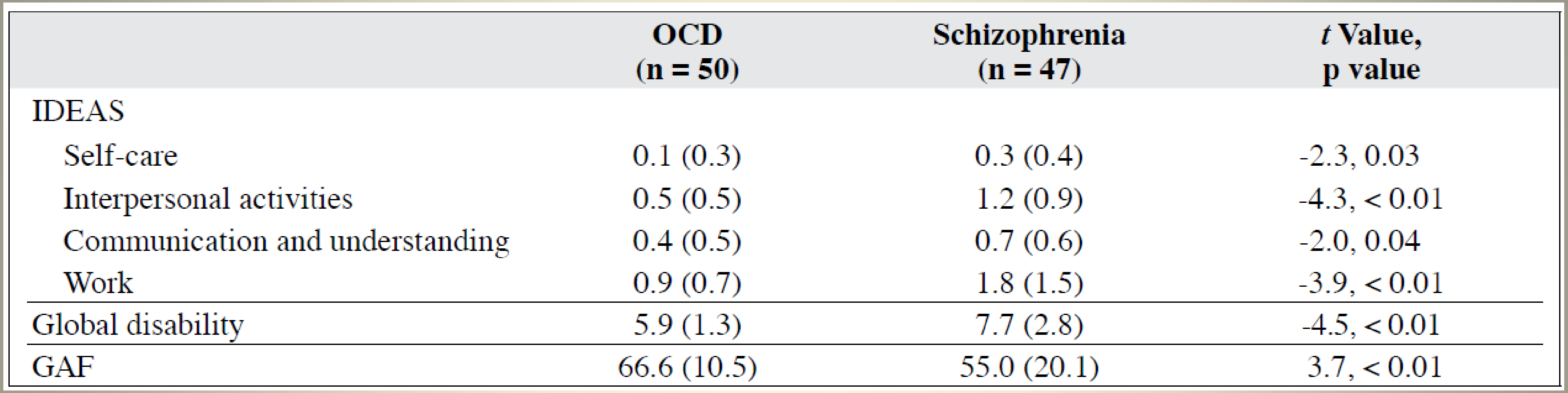
$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$



"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of"

- Sir R.A. Fisher

Indian Statistical Congress (1938)

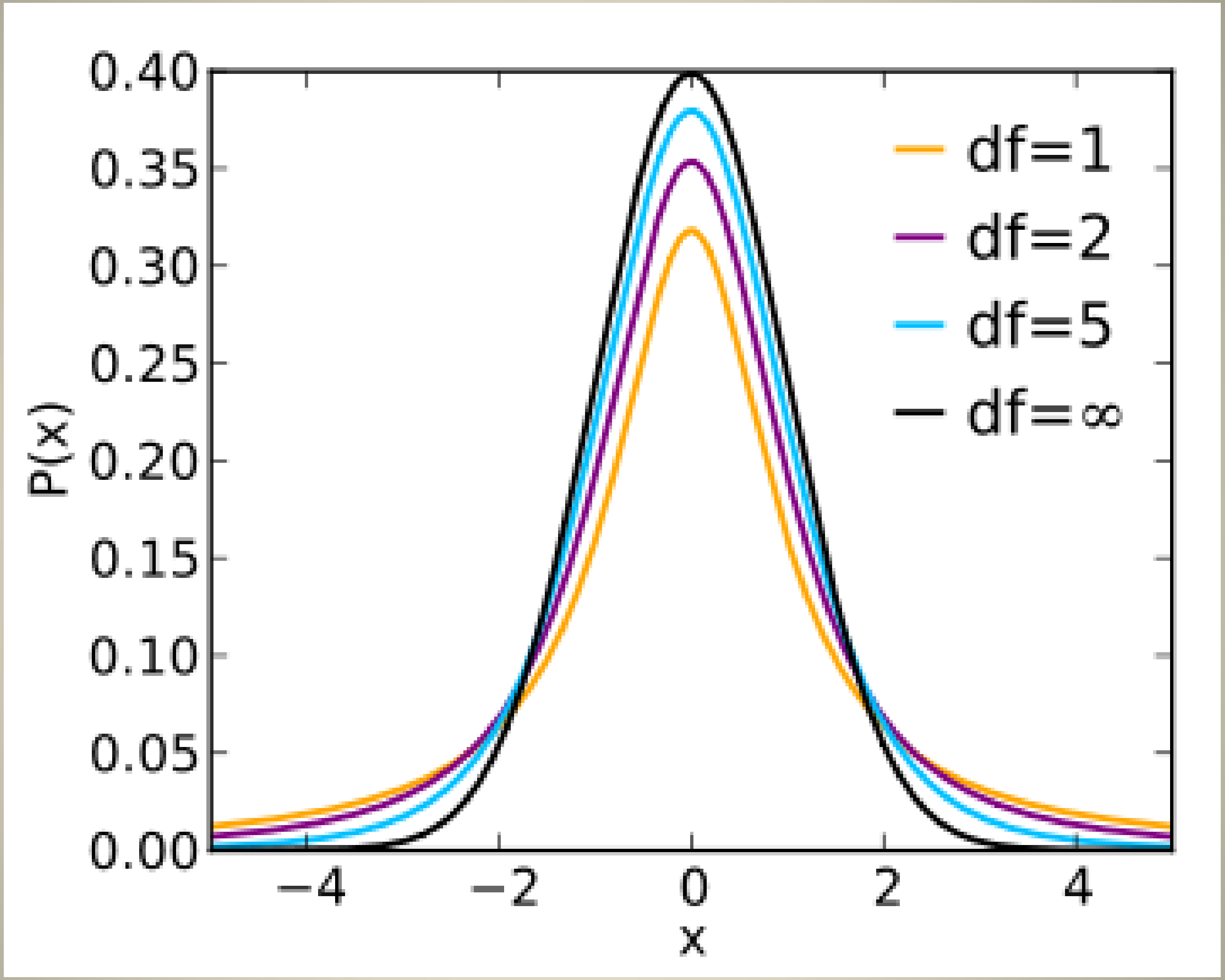


	OCD (n = 50)	Schizophrenia (n = 47)	*t* Value, p value
IDEAS			
Self-care	0.1 (0.3)	0.3 (0.4)	-2.3, 0.03
Interpersonal activities	0.5 (0.5)	1.2 (0.9)	-4.3, < 0.01
Communication and understanding	0.4 (0.5)	0.7 (0.6)	-2.0, 0.04
Work	0.9 (0.7)	1.8 (1.5)	-3.9, < 0.01
Global disability	5.9 (1.3)	7.7 (2.8)	-4.5, < 0.01
GAF	66.6 (10.5)	55.0 (20.1)	3.7, < 0.01

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

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

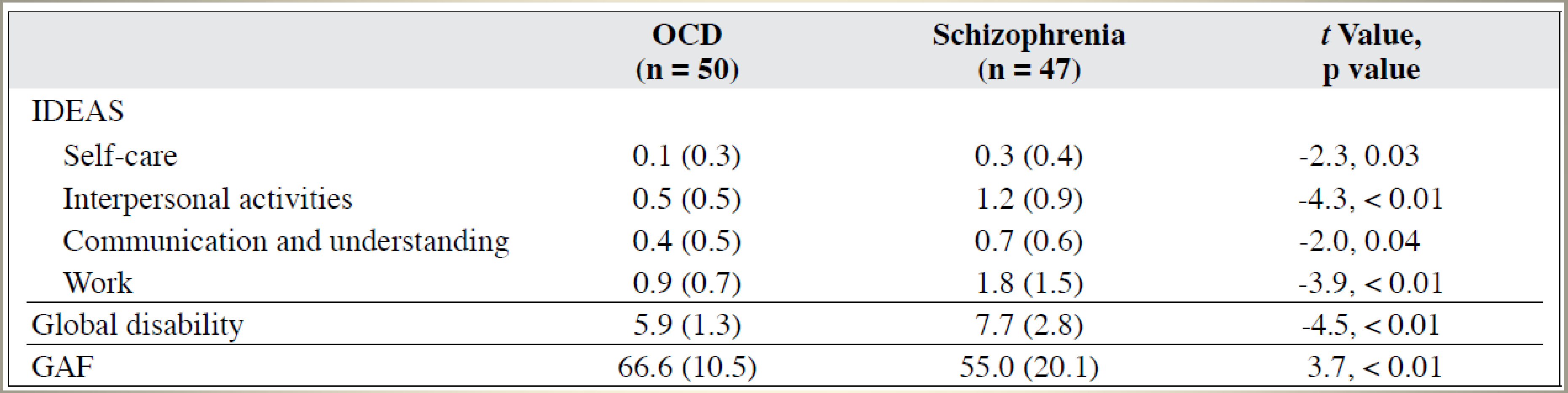
T test for paired (matched) data

- 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
- If assumption do not hold – transform the raw data (not the difference)

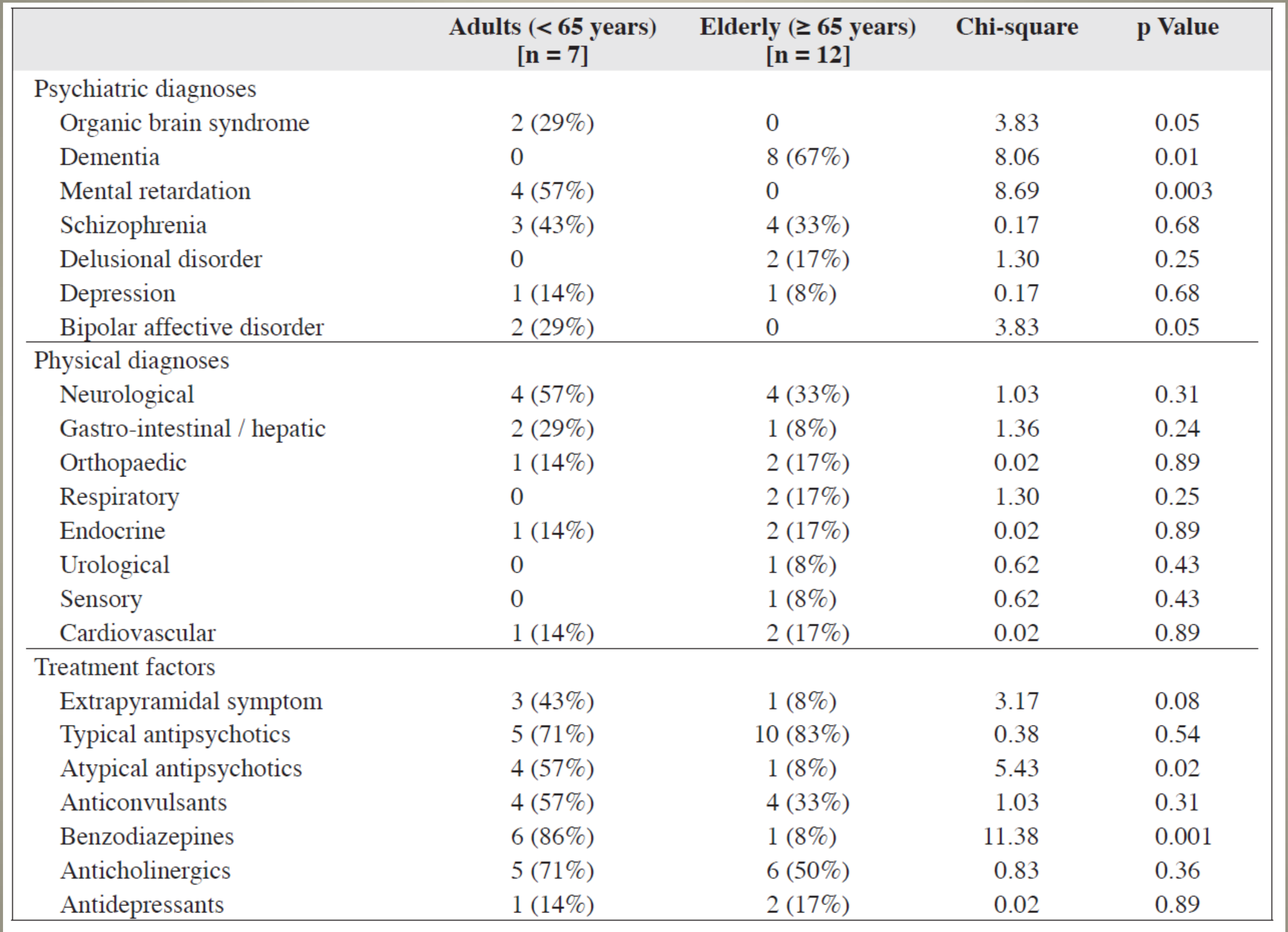
$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

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

The degree of freedom is *n-1*



	OCD (n = 50)	Schizophrenia (n = 47)	*t* Value, p value
IDEAS			
Self-care	0.1 (0.3)	0.3 (0.4)	-2.3, 0.03
Interpersonal activities	0.5 (0.5)	1.2 (0.9)	-4.3, < 0.01
Communication and understanding	0.4 (0.5)	0.7 (0.6)	-2.0, 0.04
Work	0.9 (0.7)	1.8 (1.5)	-3.9, < 0.01
Global disability	5.9 (1.3)	7.7 (2.8)	-4.5, < 0.01
GAF	66.6 (10.5)	55.0 (20.1)	3.7, < 0.01

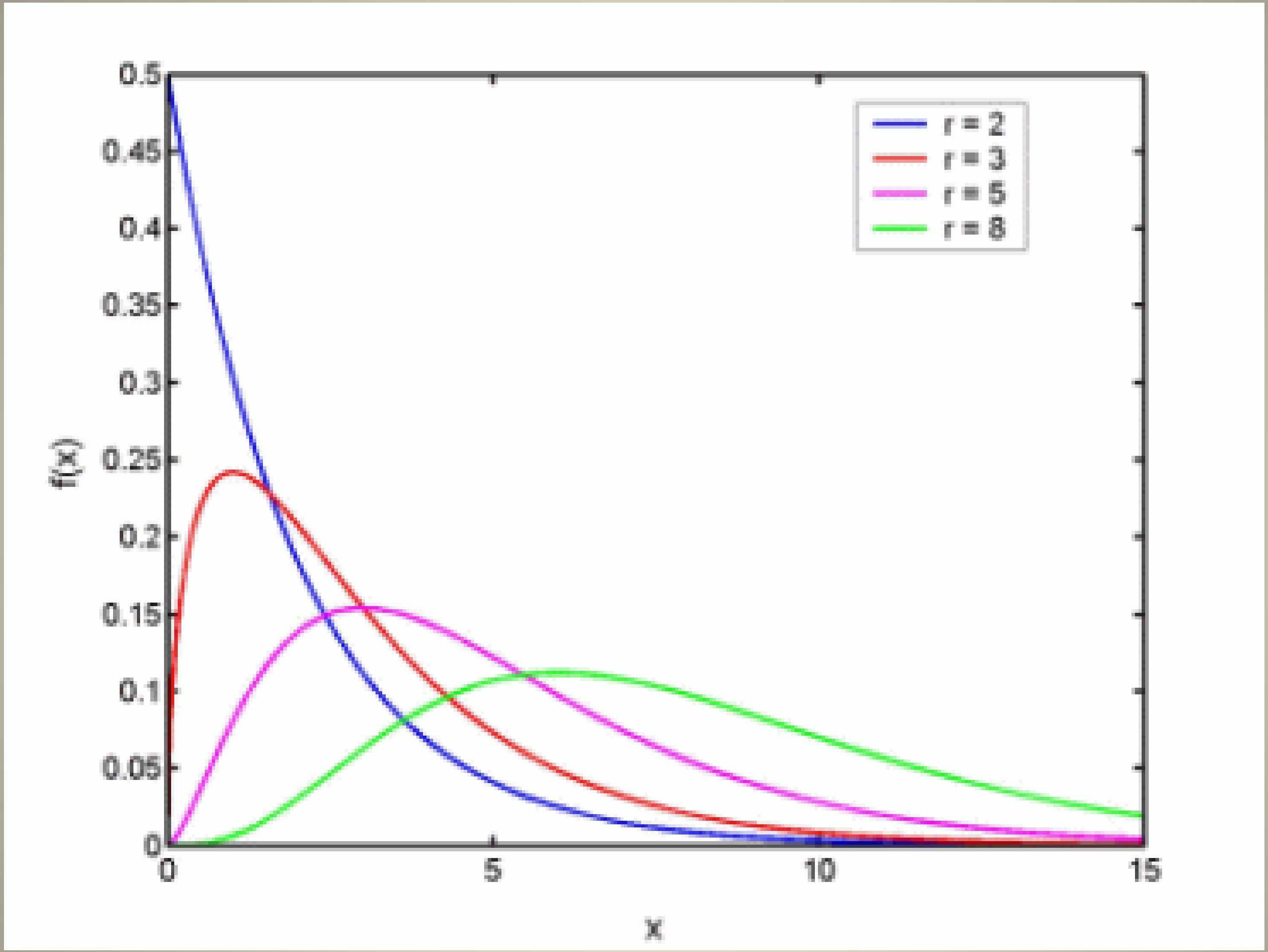


	Adults (< 65 years) [n = 7]	Elderly (≥ 65 years) [n = 12]	Chi-square	p Value
Psychiatric diagnoses				
Organic brain syndrome	2 (29%)	0	3.83	0.05
Dementia	0	8 (67%)	8.06	0.01
Mental retardation	4 (57%)	0	8.69	0.003
Schizophrenia	3 (43%)	4 (33%)	0.17	0.68
Delusional disorder	0	2 (17%)	1.30	0.25
Depression	1 (14%)	1 (8%)	0.17	0.68
Bipolar affective disorder	2 (29%)	0	3.83	0.05
Physical diagnoses				
Neurological	4 (57%)	4 (33%)	1.03	0.31
Gastro-intestinal / hepatic	2 (29%)	1 (8%)	1.36	0.24
Orthopaedic	1 (14%)	2 (17%)	0.02	0.89
Respiratory	0	2 (17%)	1.30	0.25
Endocrine	1 (14%)	2 (17%)	0.02	0.89
Urological	0	1 (8%)	0.62	0.43
Sensory	0	1 (8%)	0.62	0.43
Cardiovascular	1 (14%)	2 (17%)	0.02	0.89
Treatment factors				
Extrapyramidal symptom	3 (43%)	1 (8%)	3.17	0.08
Typical antipsychotics	5 (71%)	10 (83%)	0.38	0.54
Atypical antipsychotics	4 (57%)	1 (8%)	5.43	0.02
Anticonvulsants	4 (57%)	4 (33%)	1.03	0.31
Benzodiazepines	6 (86%)	1 (8%)	11.38	0.001
Anticholinergics	5 (71%)	6 (50%)	0.83	0.36
Antidepressants	1 (14%)	2 (17%)	0.02	0.89

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.


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

Chi-squared test

- Tests for association between two categorical variables
- Based on the chi-squared distribution with n degree of freedom
- *df = (no. of row – 1) x (no. of column – 1)*
- It gives p value but not direct estimate or confidence interval

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

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

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

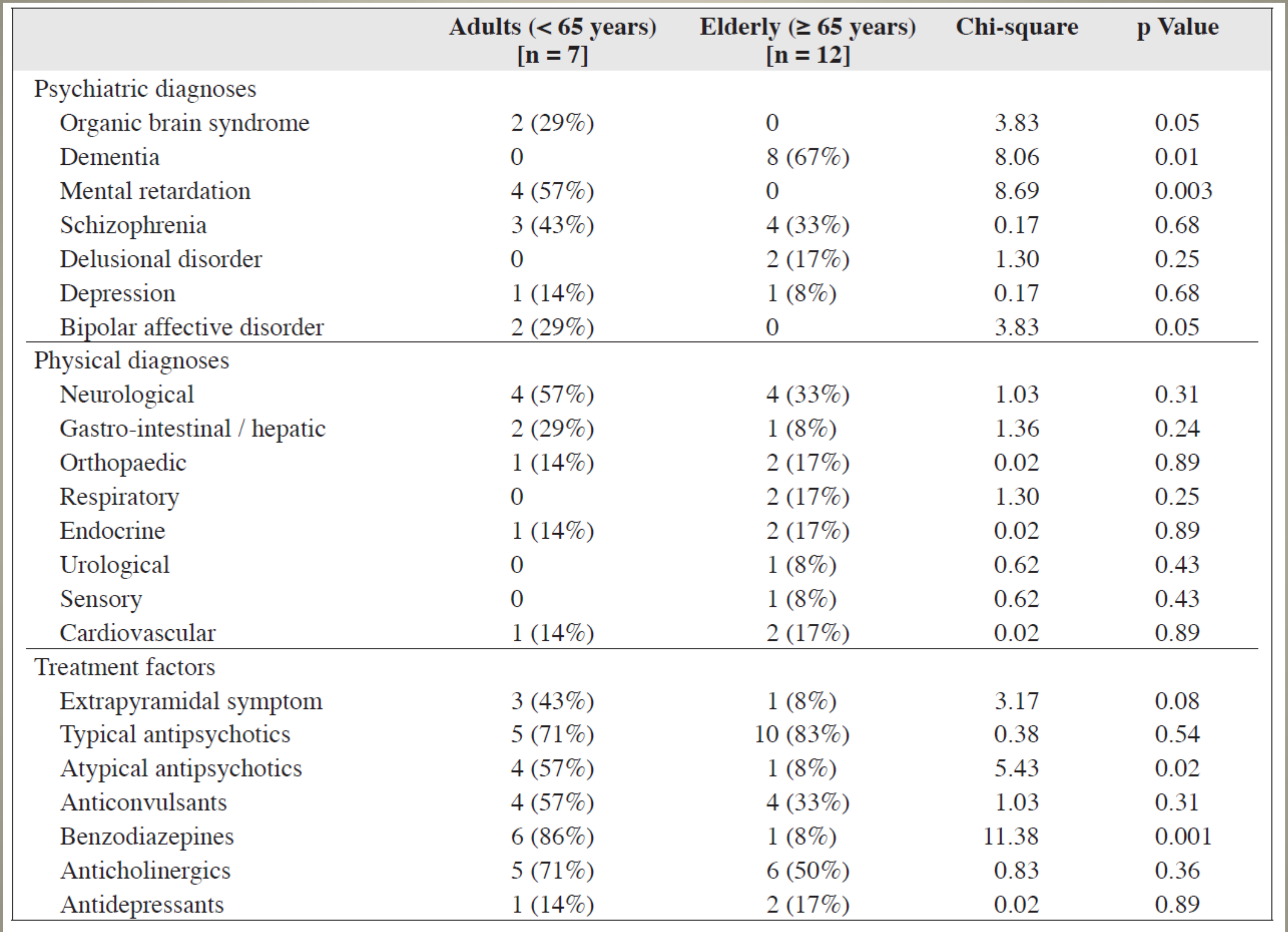
- Null hypothesis
  - No association between the two variables in the population from which the samples come
  - Same null hypothesis as the chi-squared test
- Assumptions of test
  - none

Fisher's exact test

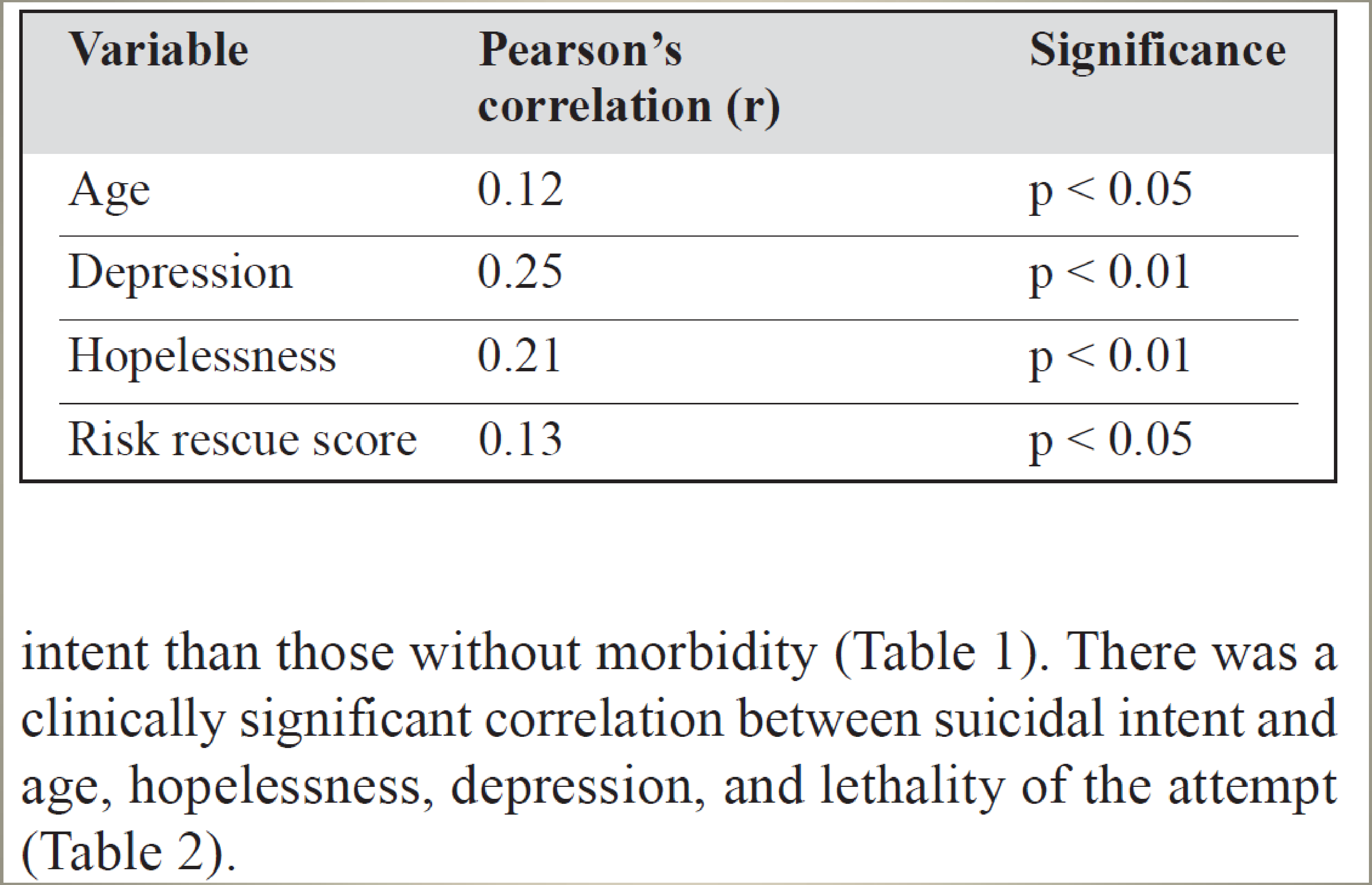
- Always use with frequencies, never use percentages for calculation
- No simple formula, statistical program needed
- Unless with good reason, use the two-sided p value
- It gives p values at least as big as the chi-squared test. For large sample size, p values are similar
- If in doubt about the sample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

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



	Adults (< 65 years) [n = 7]	Elderly (≥ 65 years) [n = 12]	Chi-square	p Value
Psychiatric diagnoses				
Organic brain syndrome	2 (29%)	0	3.83	0.05
Dementia	0	8 (67%)	8.06	0.01
Mental retardation	4 (57%)	0	8.69	0.003
Schizophrenia	3 (43%)	4 (33%)	0.17	0.68
Delusional disorder	0	2 (17%)	1.30	0.25
Depression	1 (14%)	1 (8%)	0.17	0.68
Bipolar affective disorder	2 (29%)	0	3.83	0.05
Physical diagnoses				
Neurological	4 (57%)	4 (33%)	1.03	0.31
Gastro-intestinal / hepatic	2 (29%)	1 (8%)	1.36	0.24
Orthopaedic	1 (14%)	2 (17%)	0.02	0.89
Respiratory	0	2 (17%)	1.30	0.25
Endocrine	1 (14%)	2 (17%)	0.02	0.89
Urological	0	1 (8%)	0.62	0.43
Sensory	0	1 (8%)	0.62	0.43
Cardiovascular	1 (14%)	2 (17%)	0.02	0.89
Treatment factors				
Extrapyramidal symptom	3 (43%)	1 (8%)	3.17	0.08
Typical antipsychotics	5 (71%)	10 (83%)	0.38	0.54
Atypical antipsychotics	4 (57%)	1 (8%)	5.43	0.02
Anticonvulsants	4 (57%)	4 (33%)	1.03	0.31
Benzodiazepines	6 (86%)	1 (8%)	11.38	0.001
Anticholinergics	5 (71%)	6 (50%)	0.83	0.36
Antidepressants	1 (14%)	2 (17%)	0.02	0.89



Variable	Pearson's correlation (r)	Significance
Age	0.12	$p < 0.05$
Depression	0.25	$p < 0.01$
Hopelessness	0.21	$p < 0.01$
Risk rescue score	0.13	$p < 0.05$

intent than those without morbidity (Table 1). There was a clinically significant correlation between suicidal intent and age, hopelessness, depression, and lethality of the attempt (Table 2).

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

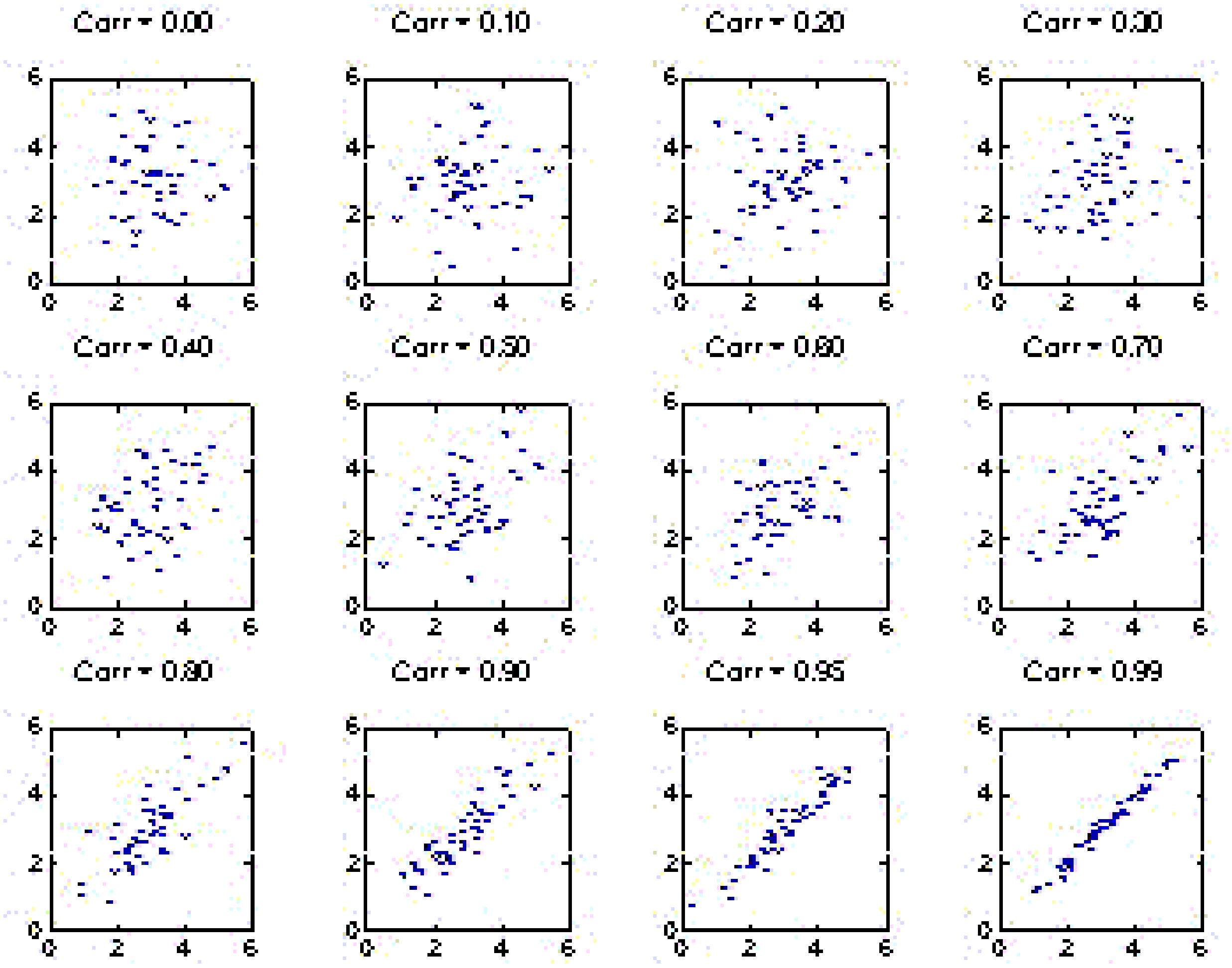
Pearson’s correlation

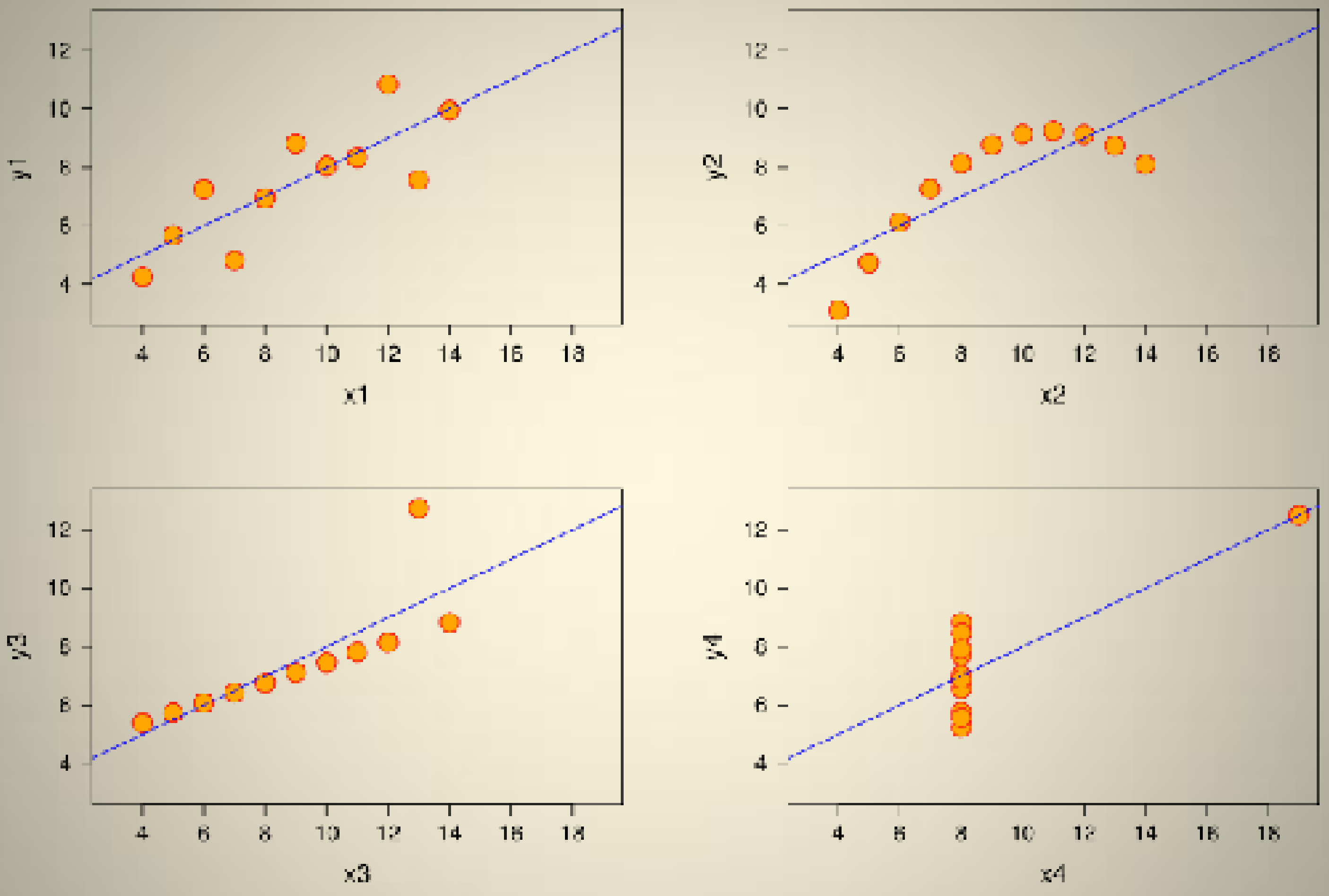
- Assumption
  - – The relationship is linear
  - – Normal distribution
    - For significant test – at least one variable to be normally disturbed
    - For confidence intervals – both variables should be normally distributed

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

Interpretation of *r*

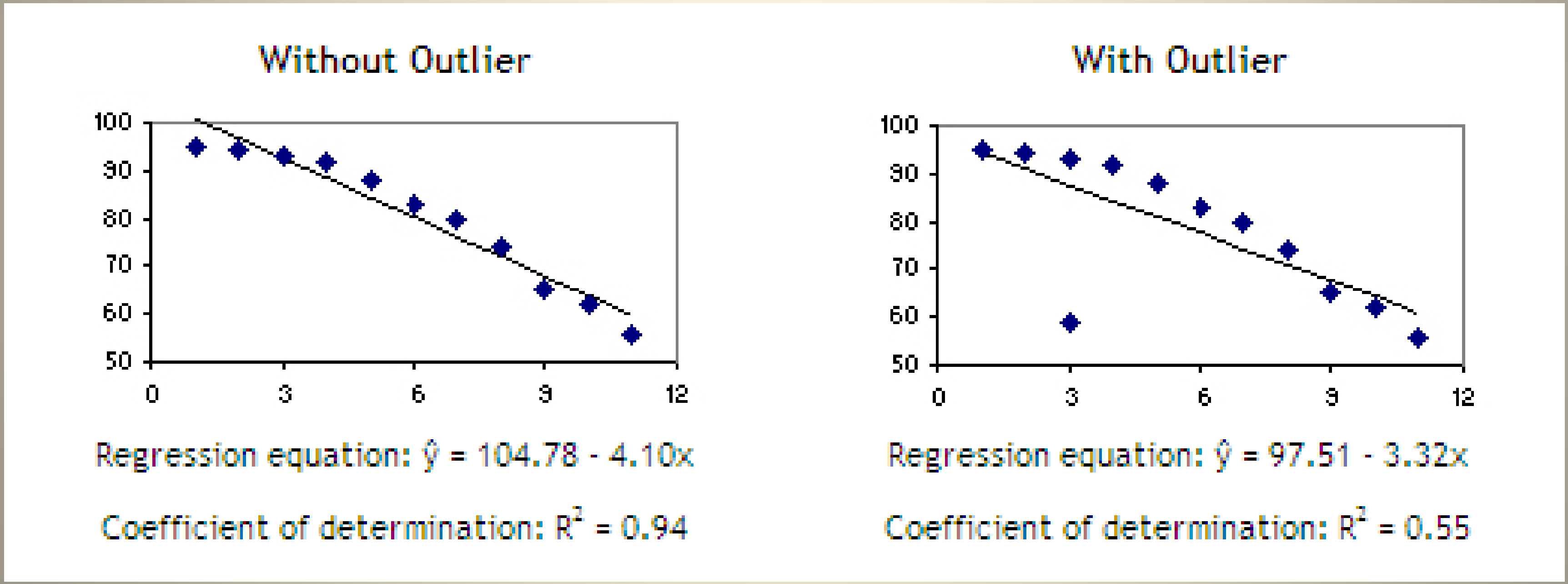
- *r* tells us how close is the linear relationship between two variables
- It lies between +1 and -1
- Negative (positive) values indicate negative (positive) linear relationship
- *r* = 0 indicate that is no linear relationship
- The closer the value +1 or -1, the stronger relationship between two 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
0
2
4
6


y1
4
6
8
10
12
4
6
8
10
12
14
16
18
x1
y2
4
6
8
10
12
4
6
8
10
12
14
16
18
x2
y3
4
6
8
10
12
4
6
8
10
12
14
16
18
x3
y4
4
6
8
10
12
4
6
8
10
12
14
16
18
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

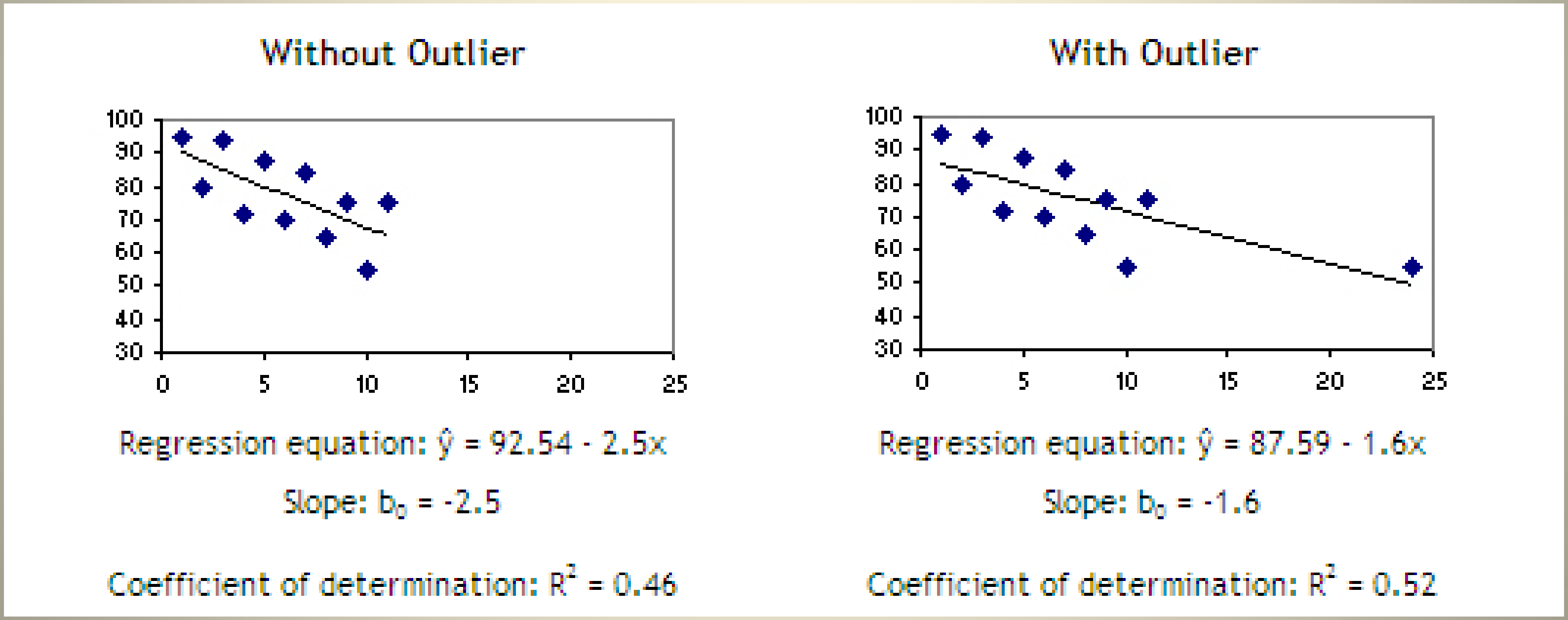
With Outlier

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



Regression equation: $\hat{y} = 92.54 - 2.5x$

Slope: $b_0 = -2.5$

Coefficient of determination: $R^2 = 0.46$

Regression equation: $\hat{y} = 87.59 - 1.6x$

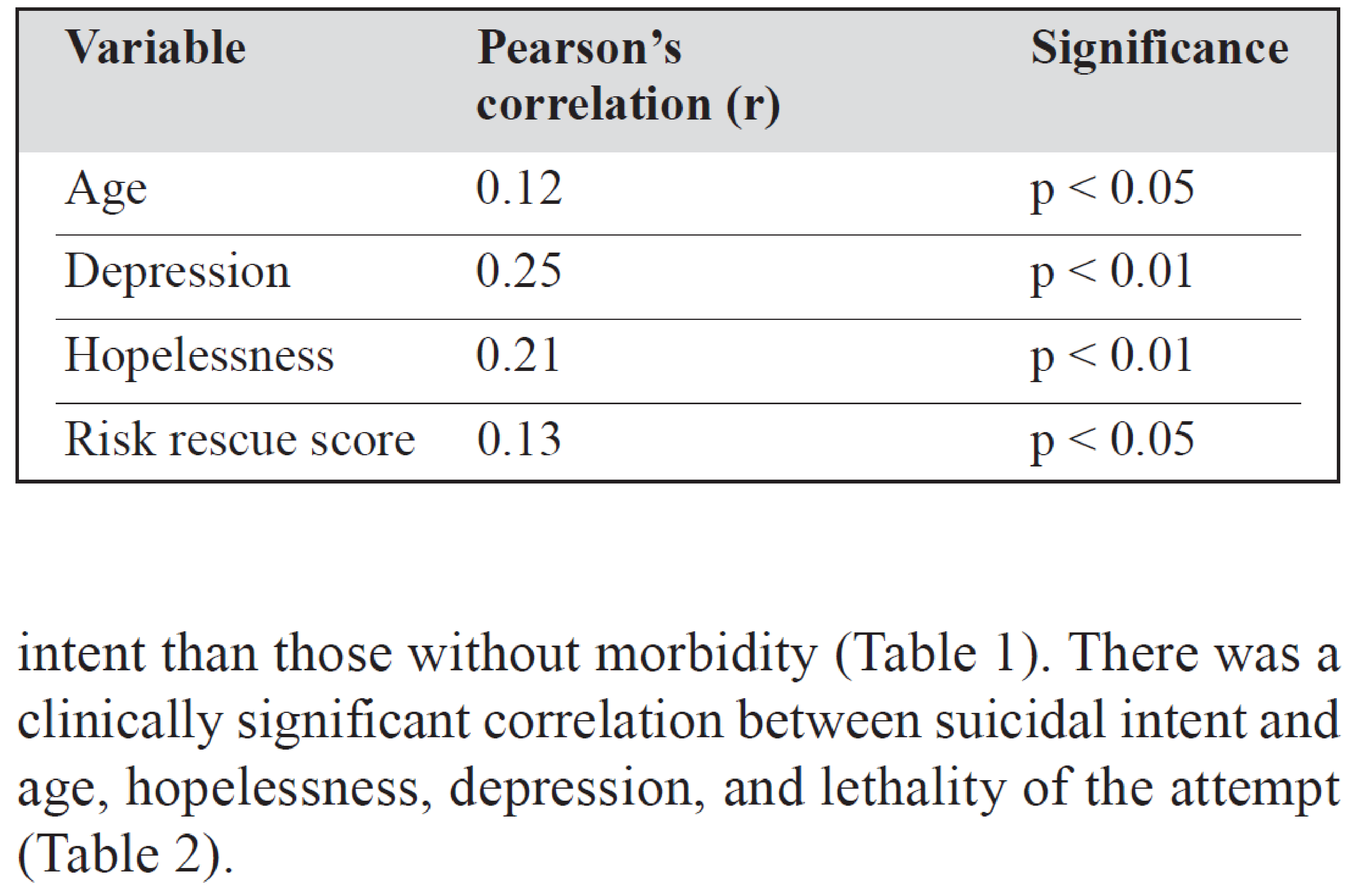
Slope: $b_0 = -1.6$

Coefficient of determination: $R^2 = 0.52$

Test and estimate of *r*

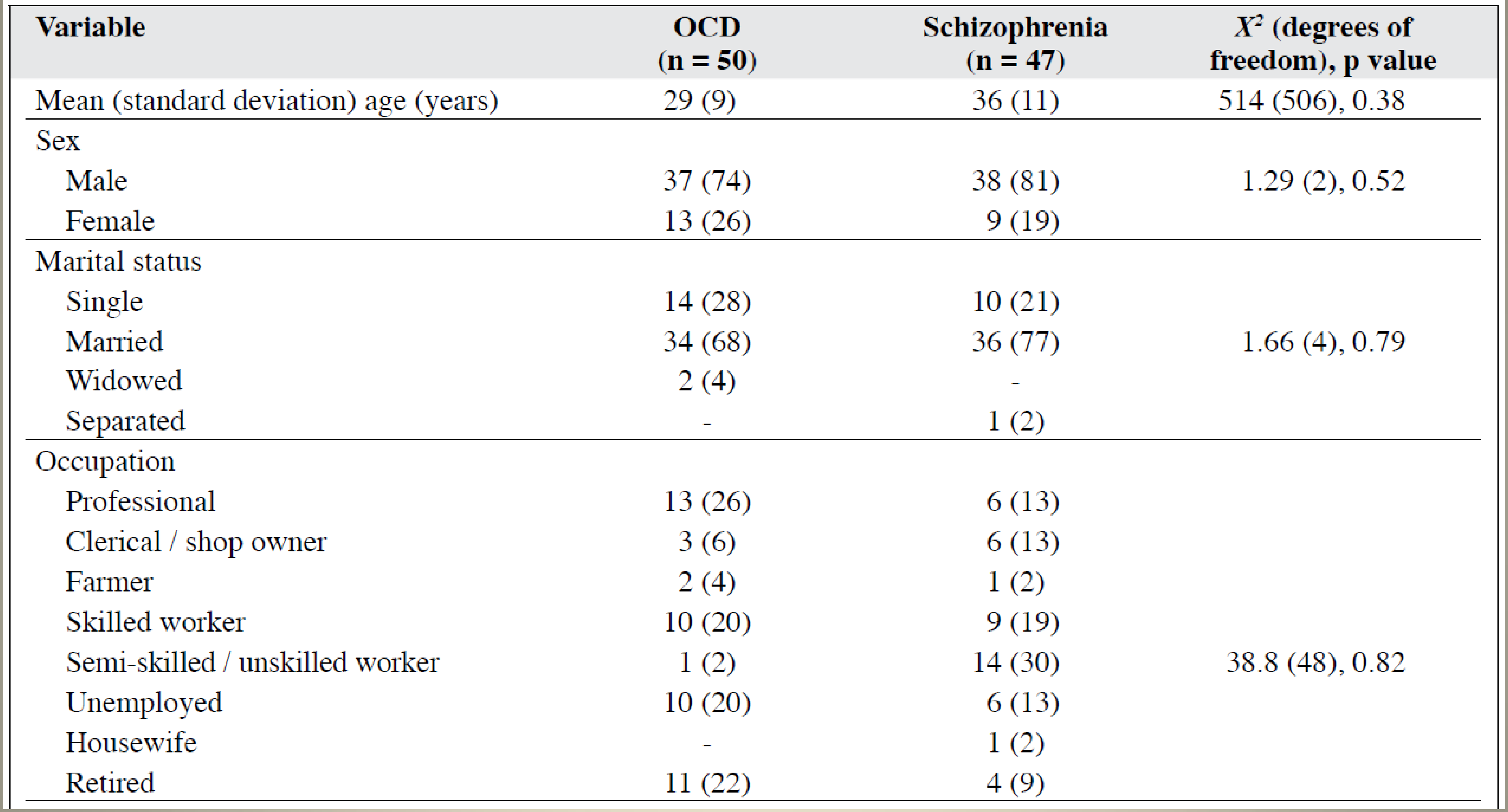
- A significant test can be done with null hypothesis that $r = 0$
- A confidence interval of *r* can be calculated
- Statistical significance of *r* directly related to sample size
  - If sample size is large, it may be statistically significant even the relationship is weak

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



Variable	Pearson's correlation (r)	Significance
Age	0.12	p < 0.05
Depression	0.25	p < 0.01
Hopelessness	0.21	p < 0.01
Risk rescue score	0.13	p < 0.05

intent than those without morbidity (Table 1). There was a clinically significant correlation between suicidal intent and age, hopelessness, depression, and lethality of the attempt (Table 2).



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)	

Hypothesis test

Hypothesis testing

- Set up hypothesis
- Find value of test statistics
- Look up critical value
- Is test statistics smaller (or greater) than critical value
- Decide reject the hypothesis or not

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?

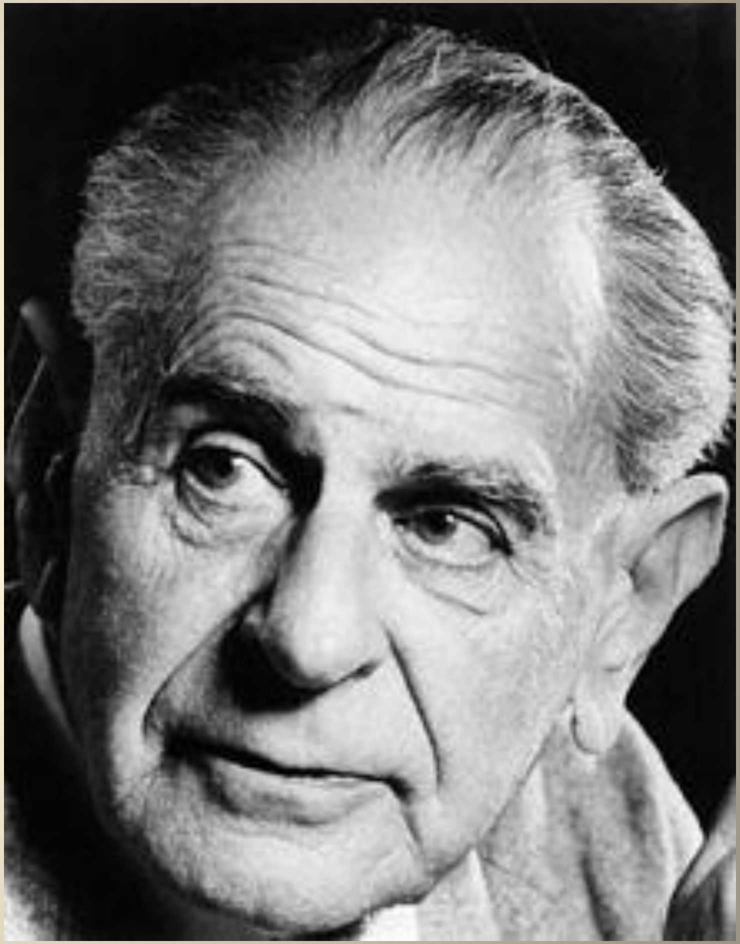
Hypothesis test

- 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.

Falsifiability

A theory can be contradicted by an observation or the outcome of a physical experiment. That something is "falsifiable" does not mean it is false; rather, that if it is false, then some observation or experiment will produce a reproducible result that is in conflict with it.

- ***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**.

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.

P value

- P value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that 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.

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”)
- 1920 – Pearson create the role of medical statistician
- 1925 – Publication of *Statistical Methods for Research Worker* (R A Fisher)
- 1935 – Publication of *The Design of Experiments* (R A Fisher)

- 1946 – first clinical trial conducted by British Medical Research Council
- 1972 – Publication of *Regression models and life tables* (D R Cox)
- 1979 – Publication of *Bootstrap methods: another look at the jackknife* (B Efron)

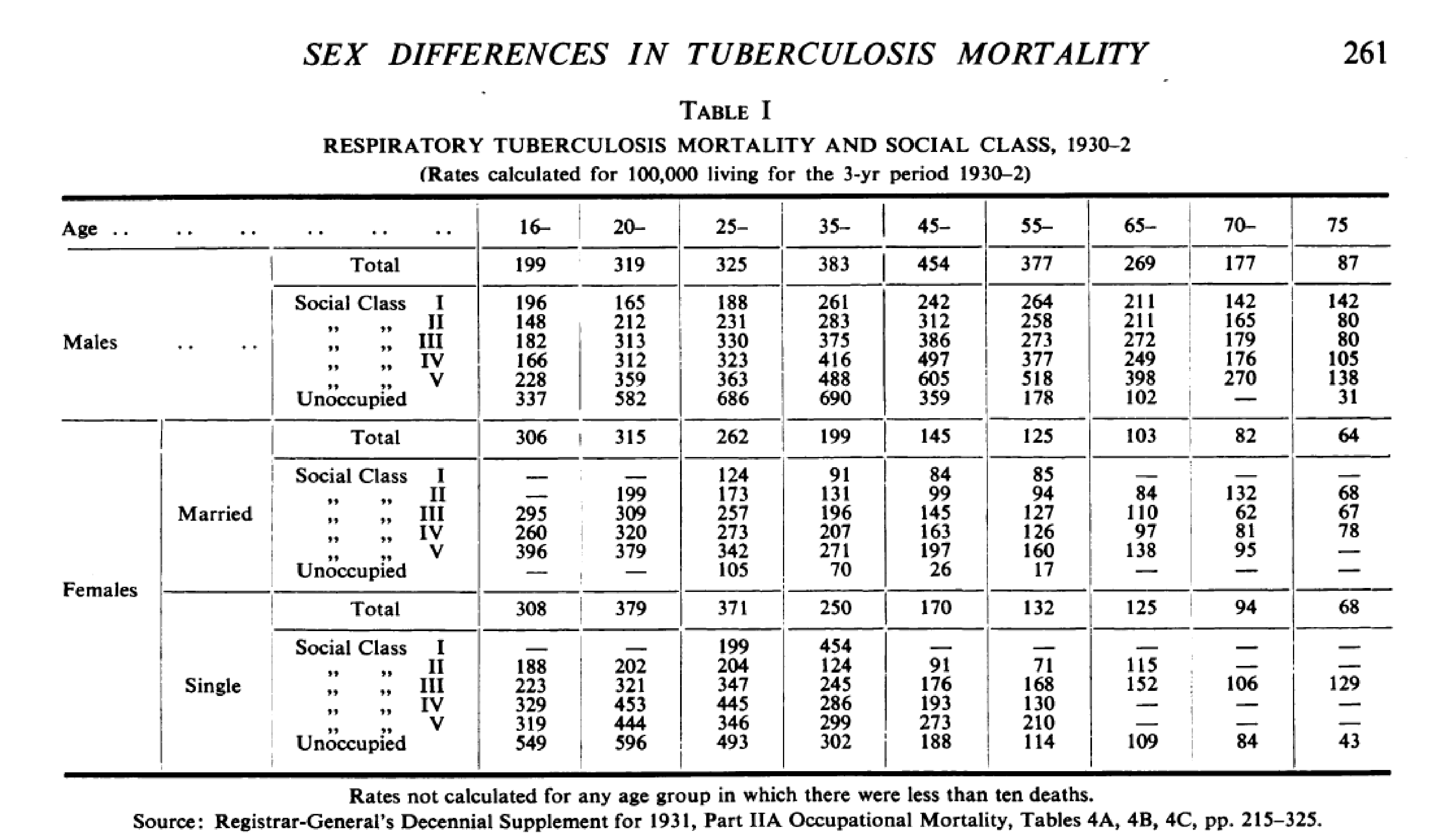


TABLE I

RESPIRATORY TUBERCULOSIS MORTALITY AND SOCIAL CLASS, 1930–2

(Rates calculated for 100,000 living for the 3-yr period 1930–2)

Age ..			16–	20–	25–	35–	45–	55–	65–	70–	75
Males		Total	199	319	325	383	454	377	269	177	87
		Social Class I	196	165	188	261	242	264	211	142	142
		,, ,, II	148	212	231	283	312	258	211	165	80
		,, ,, III	182	313	330	375	386	273	272	179	80
		,, ,, IV	166	312	323	416	497	377	249	176	105
		,, ,, V	228	359	363	488	605	518	398	270	138
		Unoccupied	337	582	686	690	359	178	102	—	31
Females	Married	Total	306	315	262	199	145	125	103	82	64
		Social Class I	—	—	124	91	84	85	—	—	—
		,, ,, II	—	199	173	131	99	94	84	132	68
		,, ,, III	295	309	257	196	145	127	110	62	67
		,, ,, IV	260	320	273	207	163	126	97	81	78
		,, ,, V	396	379	342	271	197	160	138	95	—
		Unoccupied	—	—	105	70	26	17	—	—	—
	Single	Total	308	379	371	250	170	132	125	94	68
		Social Class I	—	—	199	454	—	—	—	—	—
		,, ,, II	188	202	204	124	91	71	115	—	—
		,, ,, III	223	321	347	245	176	168	152	106	129
		,, ,, IV	329	453	445	286	193	130	—	—	—
		,, ,, V	319	444	346	299	273	210	—	—	—
		Unoccupied	549	596	493	302	188	114	109	84	43

Rates not calculated for any age group in which there were less than ten deaths.

Source: Registrar-General's Decennial Supplement for 1931, Part IIA Occupational Mortality, Tables 4A, 4B, 4C, pp. 215–325.

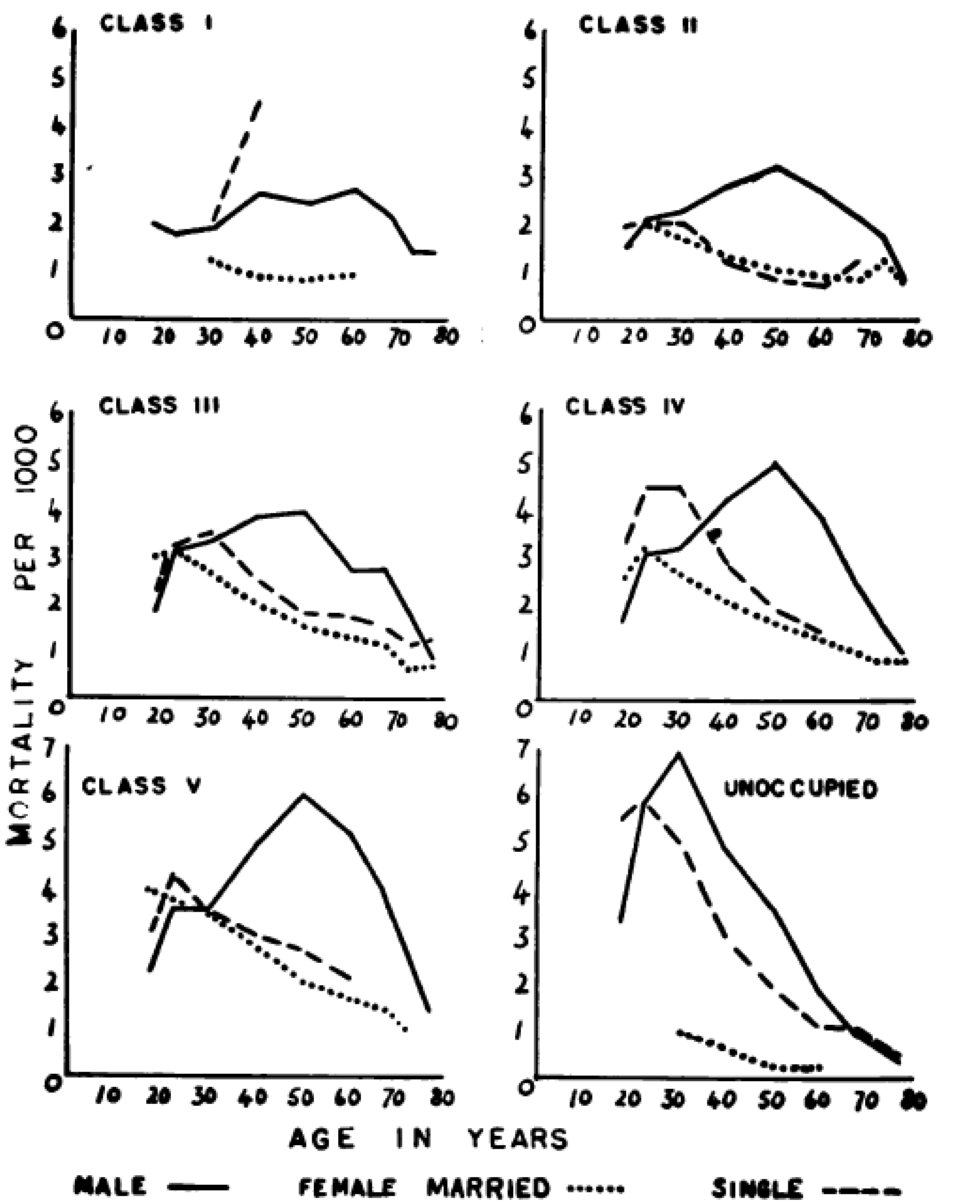


FIG. 3—Respiratory tuberculosis mortality and social class (1930–32)—based on Table I.

Drawbacks of using significant test in medical research

Can Statistical Results apply to clinical setting?

Can collective phenomena explain individual behavior?

- Macroscopically, the survival rate of a particular disease is *n*%
- Microscopically, if a patient survives, survival rate is 100%; if patient dies, survival rate is 0% - all or none
- Use of multiple regression analysis can partially solve the problem and try to individualize the treatment.

Deduction vs Induction

Deductive Reasoning

Deductive logic, is reasoning which constructs or evaluates deductive arguments. Deductive arguments are attempts to show that a conclusion necessarily follows from a set of premises. A deductive argument is valid if the conclusion must be true provided that the premises are true. A deductive argument is sound if it is valid and its premises are true.

Inductive Reasoning

Inductive logic, is a kind of reasoning that constructs or evaluates propositions that are abstractions of observations. It is commonly construed as a form of reasoning that makes generalizations based on individual instances. In this sense it is often contrasted with deductive reasoning

Clinical Inductive / Deductive Reasoning

- Inductive Reasoning – symptoms of a patient is …, and we draw a list of differential diagnoses
- Deductive Reasoning – we have a particular diagnosis in mind and compare the symptoms of the patient to see whether he fit the diagnosis or not.

Deductive Reasoning

- We can only acquire part of the truth with deductive reasoning, depends on the hypothesis we set.

5%?

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.

5% - evidence based?

- The value is fixed in 1925 when the medical statistics and clinical trial is not well established.
- 5% is reasonable for social research, as Fisher knew, but it may not reasonable in medical research
- Before we have evidence to show we should choose 5% as sigificiant level, we cannot say that we are practicing “evidence based medicine”

Points to learn

Points to learn

- Reader
  - Besides reading abstract, you can look at the table in the result
  - If problem is identified, you can put the paper in

Points to learn

- Researcher
  - Please call in your statistician as you plan your research, otherwise, what statistician can do is

Points to learn

- Academics
  - Further researches on theory of medical statistics
  - Training of medical statisticians

Q & A

kc@drkcwong.com

www.drkcwong.com