Statistical Tests (Data Analysis Methods)

Research Methodology, Biostatistics and Epidemiology (BMS 4430)

S. M. Munsaka, BSc., MSc., PhD

Department of Biomedical Sciences

School of Health Sciences

University of Zambia

A statistician is someone who, with his head in an oven and his feet in a bucket of ice water, when asked how he feels, responds: "On the average, I feel fine."

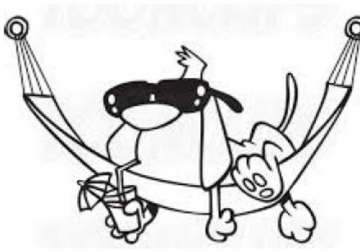
Why measurement

- Measurements supply us with numbers used in data analysis.

- No matter how profound are the theoretical formulation, how sophisticated the experimental designs are, or how elegant the analytical techniques, you **cannot compensate for poor measures**.

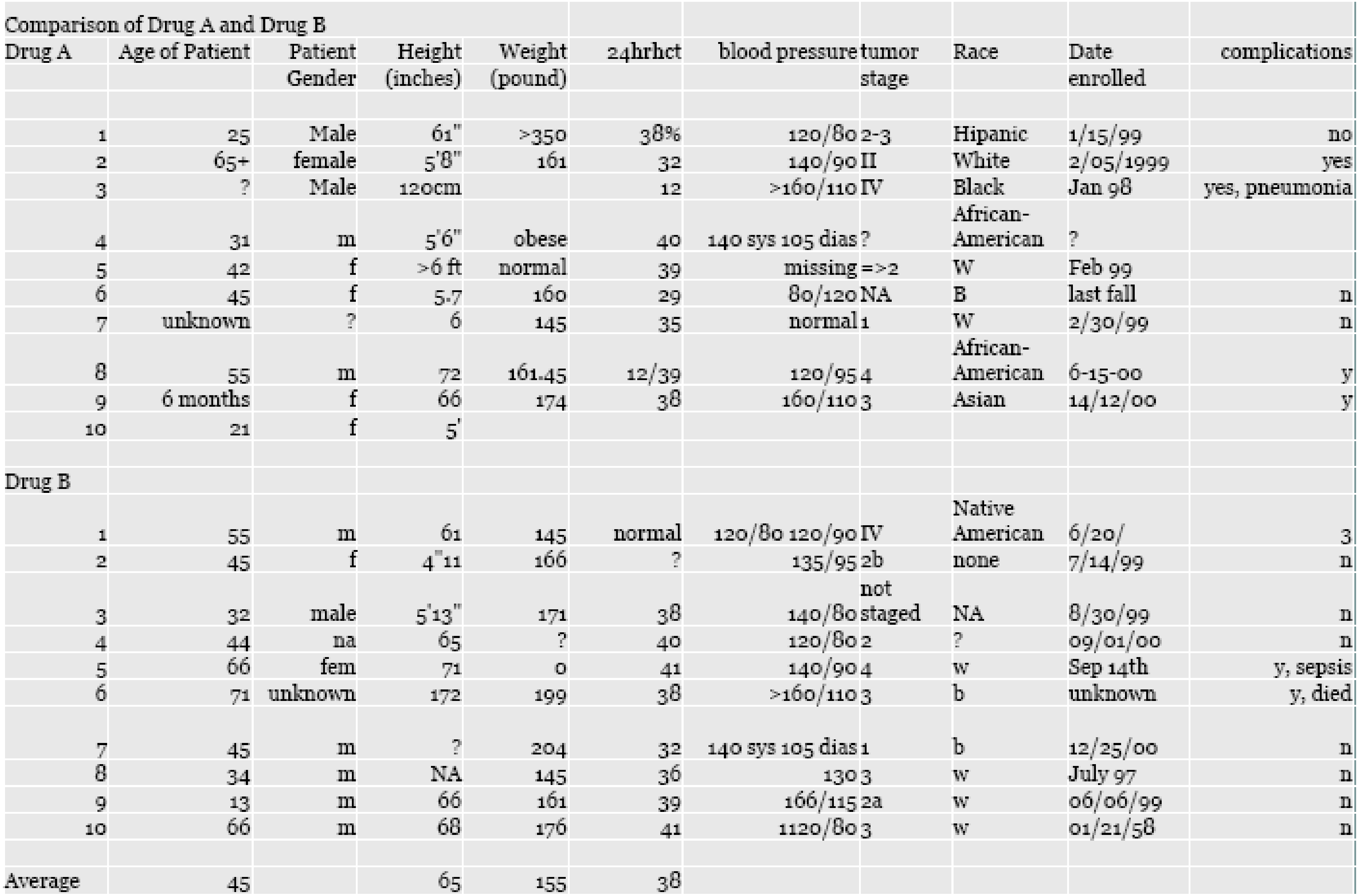
27°F

- Units of measurement are equally important



27°C

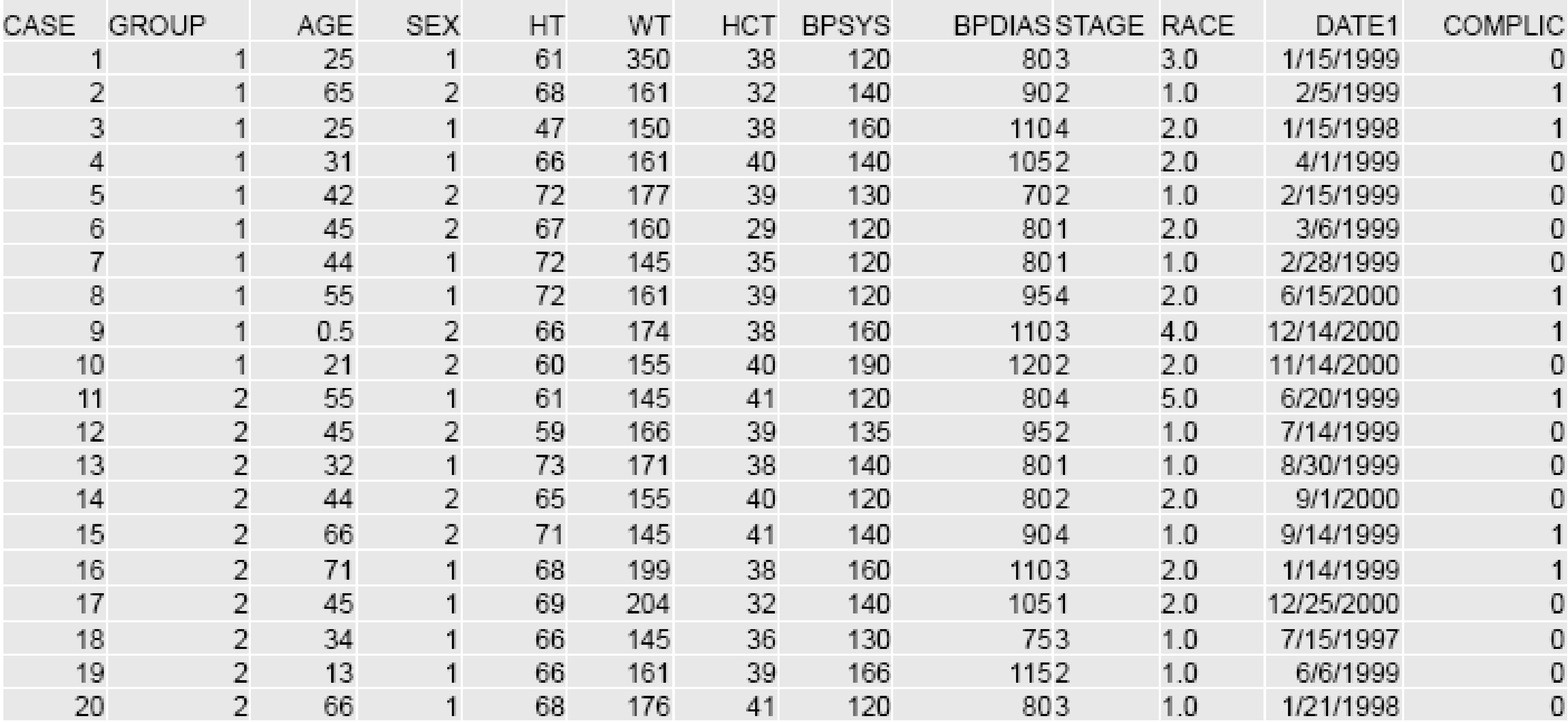
Spreadsheet from HELL



Comparison of Drug A and Drug B

Drug A	Age of Patient	Patient	Height	Weight	24hrhct	blood pressure	tumor	Race	Date	complications
		Gender	(inches)	(pound)			stage		enrolled	
1	25	Male	61"	>350	38%	120/80	2-3	Hipanic	1/15/99	no
2	65+	female	5'8"	161	32	140/90	II	White	2/05/1999	yes
3	?	Male	120cm		12	>160/110	IV	Black	Jan 98	yes, pneumonia
4	31	m	5'6"	obese	40	140 sys 105 dias	?	African-American	?	
5	42	f	>6 ft	normal	39	missing	=>2	W	Feb 99	
6	45	f	5.7	160	29	80/120	NA	B	last fall	n
7	unknown	?	6	145	35	normal	1	W	2/30/99	n
8	55	m	72	161.45	12/39	120/95	4	African-American	6-15-00	y
9	6 months	f	66	174	38	160/110	3	Asian	14/12/00	y
10	21	f	5'							
Drug B										
1	55	m	61	145	normal	120/80 120/90	IV	Native American	6/20/	3
2	45	f	4"11	166	?	135/95	2b	none	7/14/99	n
3	32	male	5'13"	171	38	140/80	not staged	NA	8/30/99	n
4	44	na	65	?	40	120/80	2	?	09/01/00	n
5	66	fem	71	0	41	140/90	4	w	Sep 14th	y, sepsis
6	71	unknown	172	199	38	>160/110	3	b	unknown	y, died
7	45	m	?	204	32	140 sys 105 dias	1	b	12/25/00	n
8	34	m	NA	145	36	130	3	w	July 97	n
9	13	m	66	161	39	166/115	2a	w	06/06/99	n
10	66	m	68	176	41	1120/80	3	w	01/21/58	n
Average	45		65	155	38					

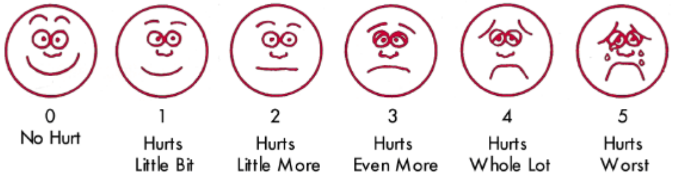
Spreadsheet from *Heaven*



CASE	GROUP	AGE	SEX	HT	WT	HCT	BPSYS	BPDIAS	STAGE	RACE	DATE1	COMPLIC
1	1	25	1	61	350	38	120	80	3	3.0	1/15/1999	0
2	1	65	2	68	161	32	140	90	2	1.0	2/5/1999	1
3	1	25	1	47	150	38	160	110	4	2.0	1/15/1998	1
4	1	31	1	66	161	40	140	105	2	2.0	4/1/1999	0
5	1	42	2	72	177	39	130	70	2	1.0	2/15/1999	0
6	1	45	2	67	160	29	120	80	1	2.0	3/6/1999	0
7	1	44	1	72	145	35	120	80	1	1.0	2/28/1999	0
8	1	55	1	72	161	39	120	95	4	2.0	6/15/2000	1
9	1	0.5	2	66	174	38	160	110	3	4.0	12/14/2000	1
10	1	21	2	60	155	40	190	120	2	2.0	11/14/2000	0
11	2	55	1	61	145	41	120	80	4	5.0	6/20/1999	1
12	2	45	2	59	166	39	135	95	2	1.0	7/14/1999	0
13	2	32	1	73	171	38	140	80	1	1.0	8/30/1999	0
14	2	44	2	65	155	40	120	80	2	2.0	9/1/2000	0
15	2	66	2	71	145	41	140	90	4	1.0	9/14/1999	1
16	2	71	1	68	199	38	160	110	3	2.0	1/14/1999	1
17	2	45	1	69	204	32	140	105	1	2.0	12/25/2000	0
18	2	34	1	66	145	36	130	75	3	1.0	7/15/1997	0
19	2	13	1	66	161	39	166	115	2	1.0	6/6/1999	0
20	2	66	1	68	176	41	120	80	3	1.0	1/21/1998	0

Types of Measurements (Variables)

- **Categorical (Nominal)**
  - Mutually exclusive and exhaustive
  - No particular order
  - E.g. profession, Disease condition, Sex
    - Binary (dichotomous): two outcomes
      - E.g. Sex; male and female, HIV status; Positive and negative
- **Ordinal**
  - Categorical measurements (variables) that can be ordered
    - E.g. severity of symptoms, age ranges
- **Continuous (interval)**
  - Numerical variable with many possible val
  - Has the most statistical information
    - Egg age, blood pressure, blood glucose levels, CD4, viral load etc



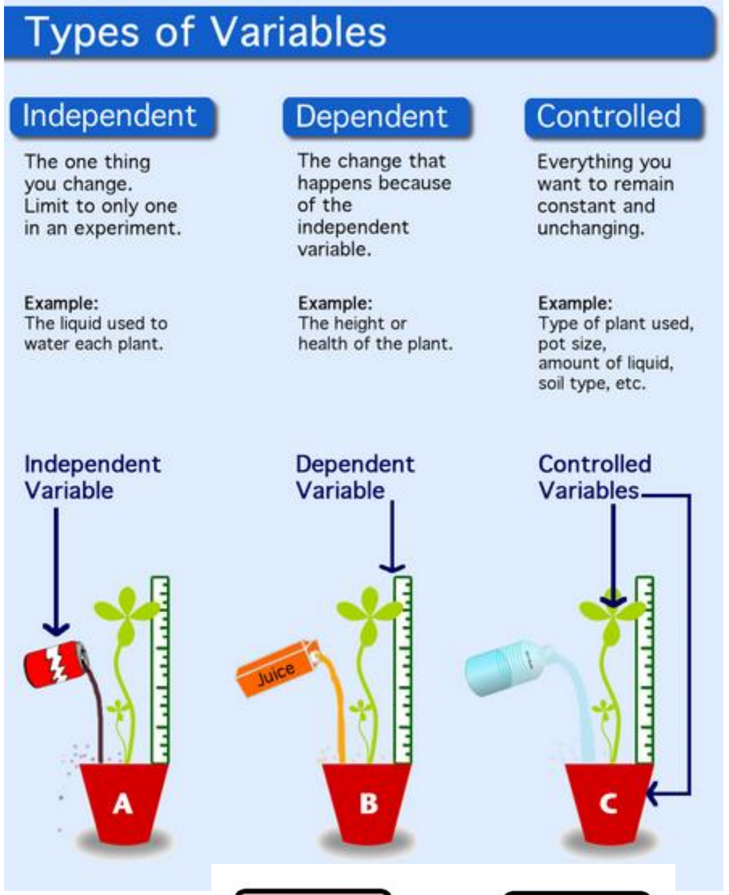
- **Discrete:**
  - Order and magnitude matter, but possible values can be listed (Number of Seizures)

Types of Measurements (Variables)

- Dependent variable (Experimental or Response variable)
  - The variable that you measure
  - The **dependent variable** responds to the **independent variable**. It is called **dependent** because it "depends" on the **independent variable**.
  - E.g. outcomes; death, CD4, viral load, blood glucose

- Independent variable (predictor or descriptor variable)
  - Not controlled by experimenter
  - E.g. race, sex, HIV status

- Adjustment variable (cofounder)
  - A variable that can affect both the dependent and independent variable



Types of Variables

Independent

The one thing you change. Limit to only one in an experiment.

Example: The liquid used to water each plant.

Dependent

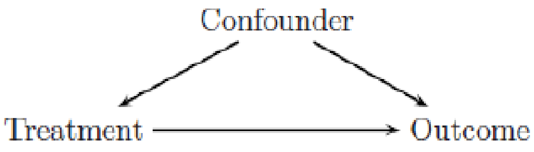
The change that happens because of the independent variable.

Example: The height or health of the plant.

Controlled

Everything you want to remain constant and unchanging.

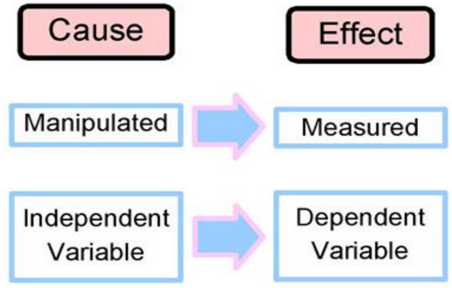
Example: Type of plant used, pot size, amount of liquid, soil type, etc.



Confounder

Treatment

Outcome



Cause

Effect

Manipulated

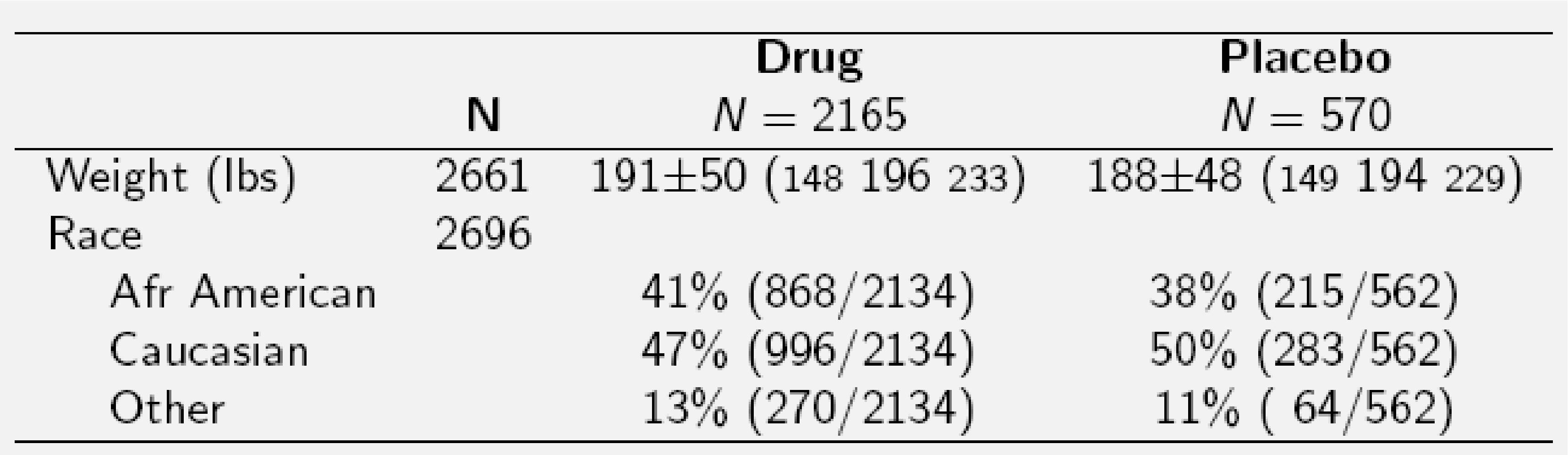
Measured

Independent Variable

Dependent Variable

Goals of Data Analysis

- 1. To describe (summarizes) the population of interest
  - Description of what was observed (measured) in the study/sample population
  - Known as Descriptive Statistics
    - Summarizes continuous variables using
      - Means, standard deviations, percentiles, medians, ranges, mode etc
    - Summarizes categorical variables using
      - Raw or relative frequencies, percentages etc



		Drug	**Placebo**
	N	$N = 2165$	$N = 570$
Weight (lbs)	2661	191±50 (148 196 233)	188±48 (149 194 229)
Race	2696		
Afr American		41% (868/2134)	38% (215/562)
Caucasian		47% (996/2134)	50% (283/562)
Other		13% (270/2134)	11% (64/562)

Data Analysis

- Start with the specific aim
  - What is your research question or what hypothesis are you testing?

- The wording in specific aims convey what statistical test to be used
  - To determine……the relationship (association)…..
  - To compare…
  - To predict…the probability of survival with the first year of…..
  - To evaluate…
  - To describe….
  - To determine whether there is a statistical difference/change between

Specific
Measurable
Achievable
Realistic
Time bound

- E.g. 1. To compare at baseline, and months 2, 4, and 6 of intervention, parameters of menstrual cycle regularity for 6-months, between women taking *Gymnema s*. vs. placebo.

- E.g. 2. To determine the relationship between hirsutism, free testosterone, FSH/LH ratio, fasting glucose, insulin, SHBG, HgA1c, estradiol, progesterone, and cortisol, in women taking *Gymnema s*. vs. placebo for 6-months.

Data Analysis Methods

- Tests for association
  - To determine of two variables are independent (two continuous or one continuous and the other categorical)
    - Two variables are associated if one affects the value/distribution of the other
      - E.g. association between race (categorical predictor) and disease (categorical outcome) or Age (continuous predictor) and HDL cholesterol (continuous outcome)
  - To determine a difference in the distribution of a variable between (two) or across groups (more than 2) or testing for an association between group (predictor) variable and an outcome variable
    - E.g. to determine the difference in the distribution of cholesterol (continuous variable) between genders (categorical predictors)
    - Includes paired observations e.g. To determine differences in test scores before and after a didactic class

Data Analysis Methods

- Tests for association
  - Hypothesis testing
    - *Null hypothesis*: proposes that two variables are independent
      - E.g. There is no difference in the frequency of drinking well water and between patients who develop peptic ulcers and those who do not
    - *Alternate hypothesis*: proposes that there is an association between variables
      - Can be one sided e.g. Drinking well water is more common in people who develop peptic ulcer disease.
      - Can be two sided e.g. The frequency of drinking well water is different between those who develop peptic ulcers and those who do not.
    - If tests p value is less than 0.05 we conclude that the variables are significantly associated and we fail to reject the null hypothesis.
      - But note that statistical significance does not always mean clinical significance or association does not always equal to causation.
    - If tests p value is greater than 0.05 we reject the null hypothesis. However, rejecting the null hypothesis does not mean that there is no association between the two variables in the population; you just failed to find this association in your sample.

Tests for Association

Purpose	Dependent variable	Independent variable	Test to use
Compare means of two independent groups	Continuous	Categorical/ nominal	Students T test [Mann-Whitney U test or Wilcoxon Rank-sum Test]
Compare means of more than two independent groups	Continuous	Categorical/ nominal	1-Way Analysis of Variance (ANOVA) [Kruskal-Wallis Test]
Compared means of paired responses	Continuous	Categorical/ nominal	Paired T test [Wilcoxon Signed-Rank Test]
Compare means of more than 3 measurements of the same subject	Continuous	Time variable	Repeated measures ANOVA [Friedman Test]

Non-parametric equivalents are given in parentheses [ ]

Tests for Association

Purpose	Dependent variable	Independent variable	Test to use
Determine association between two independent groups	Categorical/ nominal	Categorical/ nominal	Chi-square test Fisher's Exact Test (when counts are <5)
Determine an association between two continuous variables	Continuous	Continuous	Pearson's Correlation [Spearman's Correlation] also used for ordinal data
Predicting a value from another measured variable (predictor)	Continuous	any	Simple linear regression (multiple linear regression) if from several measured variables
Predicting a value from another measured variable (predictor)	Continuous	Categorical (binomial) two outcomes	Simple logistic regression (multiple logistic regression), proportional odds (ordinal), Cox proportional hazards (time to event)

Student’s T-Test

- Used to compare means of two group
  - Assumptions
    - Data is continuous
    - Groups are independent
    - Data is normally distributed

- Hypothesis being tested:
  - The difference between the means is not due to sampling error (randomness)

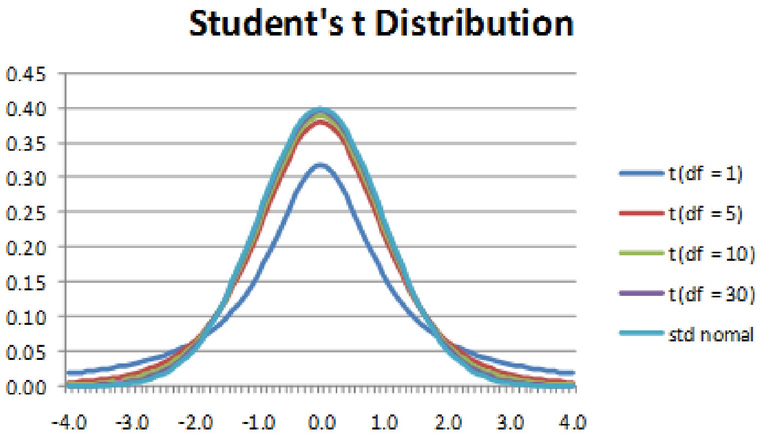
Difference in means

- Formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

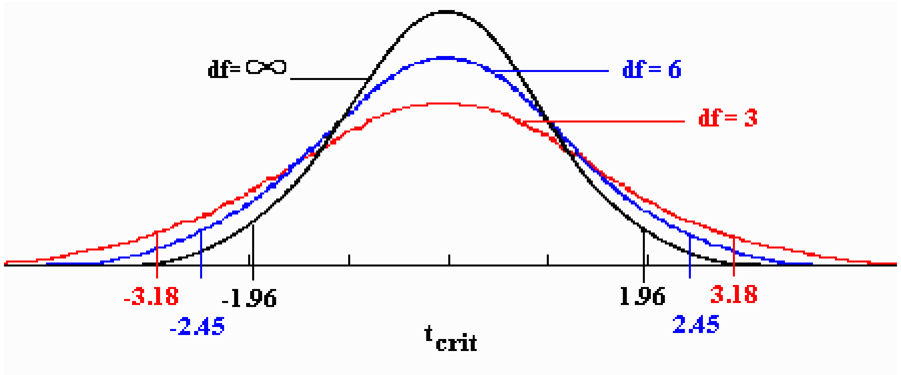
Variance (variability)

T-Distribution



Example:
Group A: n=20, mean=75, squared variance =16.25

Group B: n=20, mean=71, squared variance =18.5



95% of the data are bound within 1.96 standard deviations and
99% of data are within 2.58 standard deviations

Z-test

- When Z test is applied to sampling variability where samples are larger than 30.
- Means of samples in population will also follow normal distribution.
- So, observer difference is calculated in terms of SE instead of SD.

$$Z = \frac{\bar{X} - \mu}{SE\ of\ mean}$$

$$= \frac{Observed\ difference\ between\ sample\ mean\ and\ population\ mean}{SE\ of\ mean}$$

Z-TEST

- Formula to find the value of Z (z-test) Is:

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

- x̄ = mean of sample
- μ_0 = mean of population
- σ = standard deviation of population
- n = no. of observations

Assumptions of the *z*-Test

1. We have randomly selected one sample
2. The dependent variable is at least approximately normally distributed in the population and involves an interval or ratio scale
3. We know the mean of the population of raw scores under another condition of the independent variable
4. We know the true standard deviation of the population (σ_X) described by the null hypothesis

ANOVA

- Used to assess differences between more than two sample means

- Uses the F statistic

- Assumptions:
  - Populations are normally distributed
  - Populations have equal variances (similar variability)
  - Samples are independent and are selected randomly

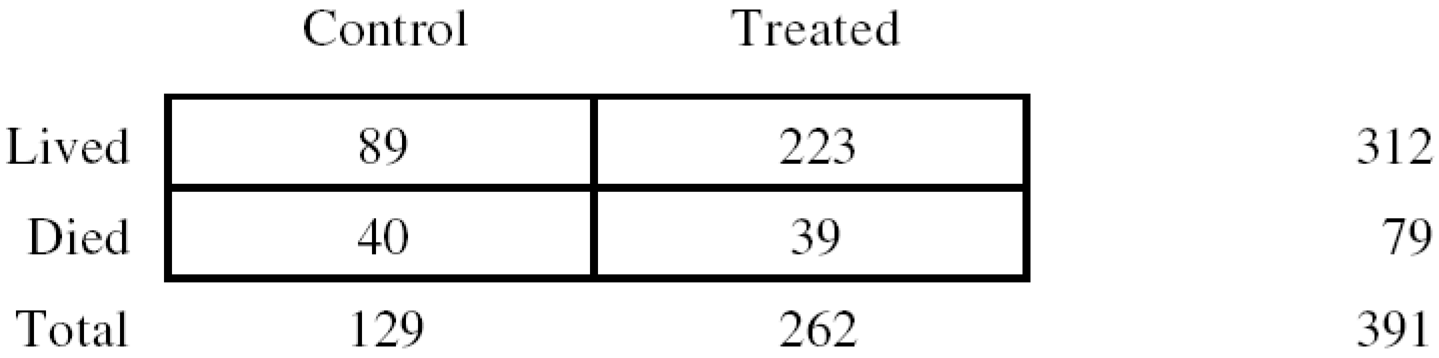
- F= Vb/Vw where Vb=total group variance – within group variance
- Or F=Msb/Msw where Msb=mean squared, between group variance and Msw=mean squared within group variance

Chi-square Test

- A statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance

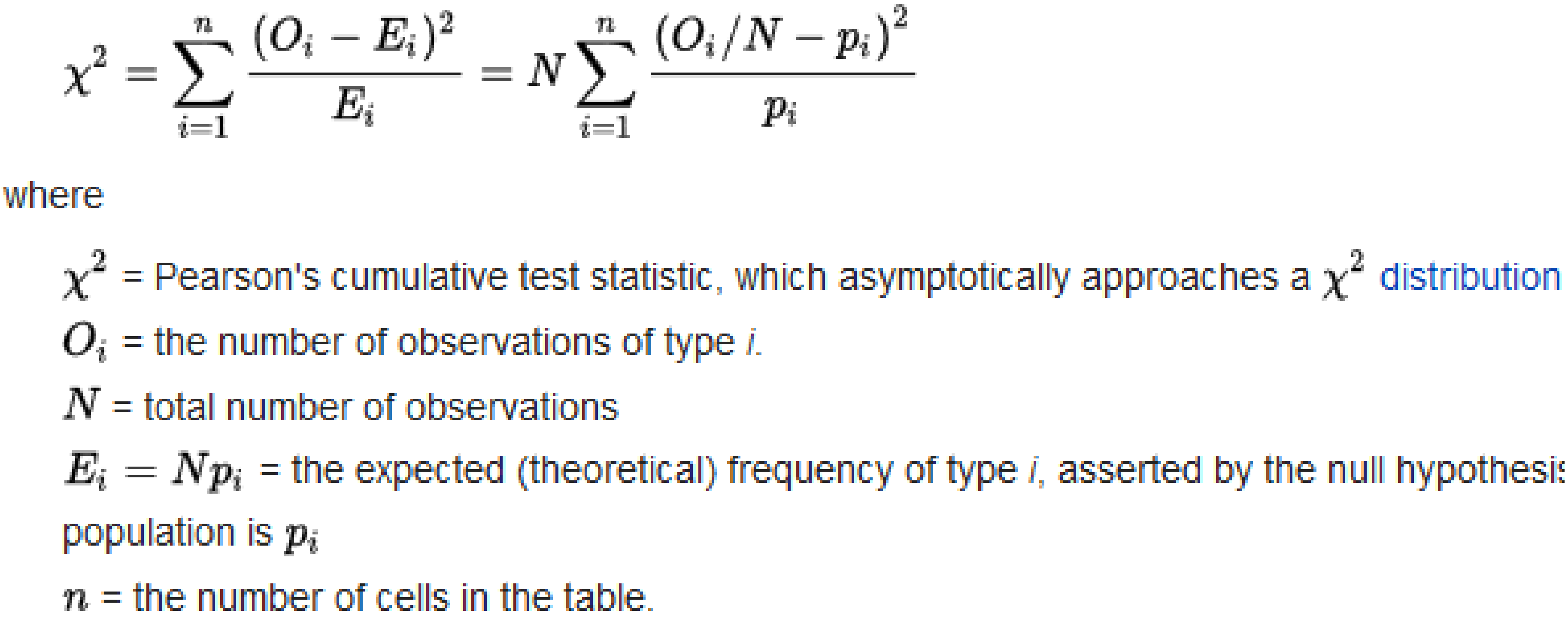
- Also used to compare proportions in two groups

- Data often presented in a 2 x 2 table



	Control	Treated	
Lived	89	223	312
Died	40	39	79
Total	129	262	391

Chi-square calculation



$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where

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

O_i = the number of observations of type *i*.

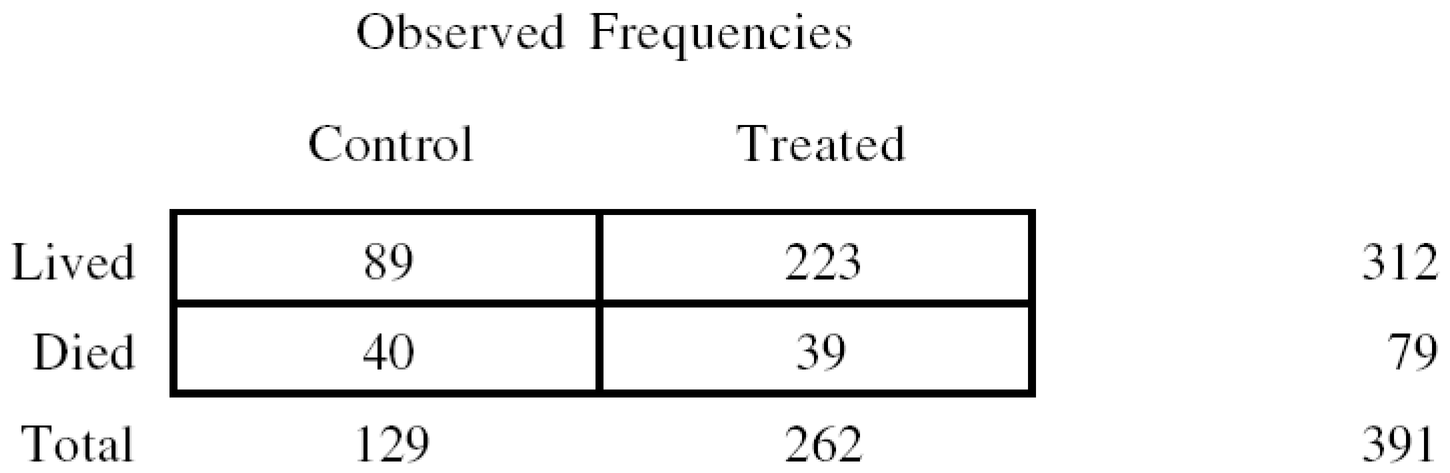
N = total number of observations

$E_i = Np_i$ = the expected (theoretical) frequency of type *i*, asserted by the null hypothesis

population is p_i

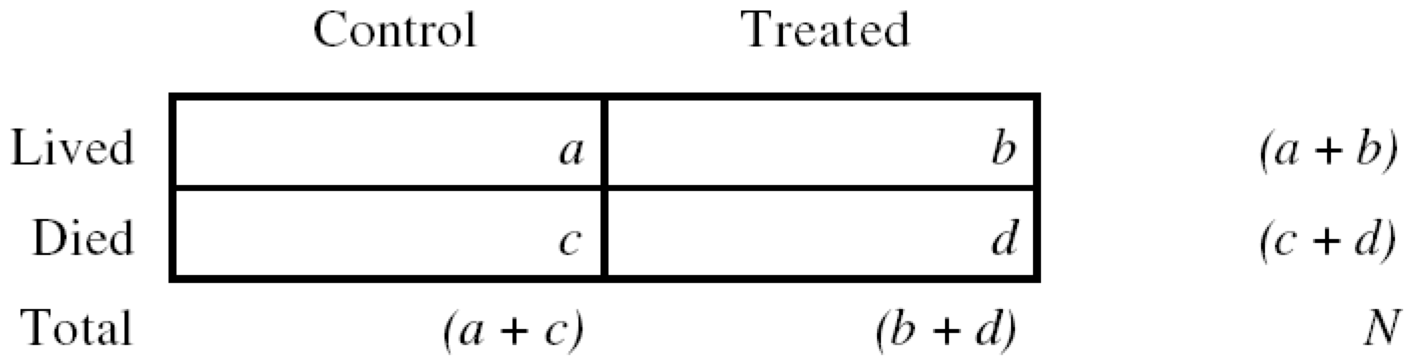
n = the number of cells in the table.

Chi-square example



Observed Frequencies

	Control	Treated	
Lived	89	223	312
Died	40	39	79
Total	129	262	391

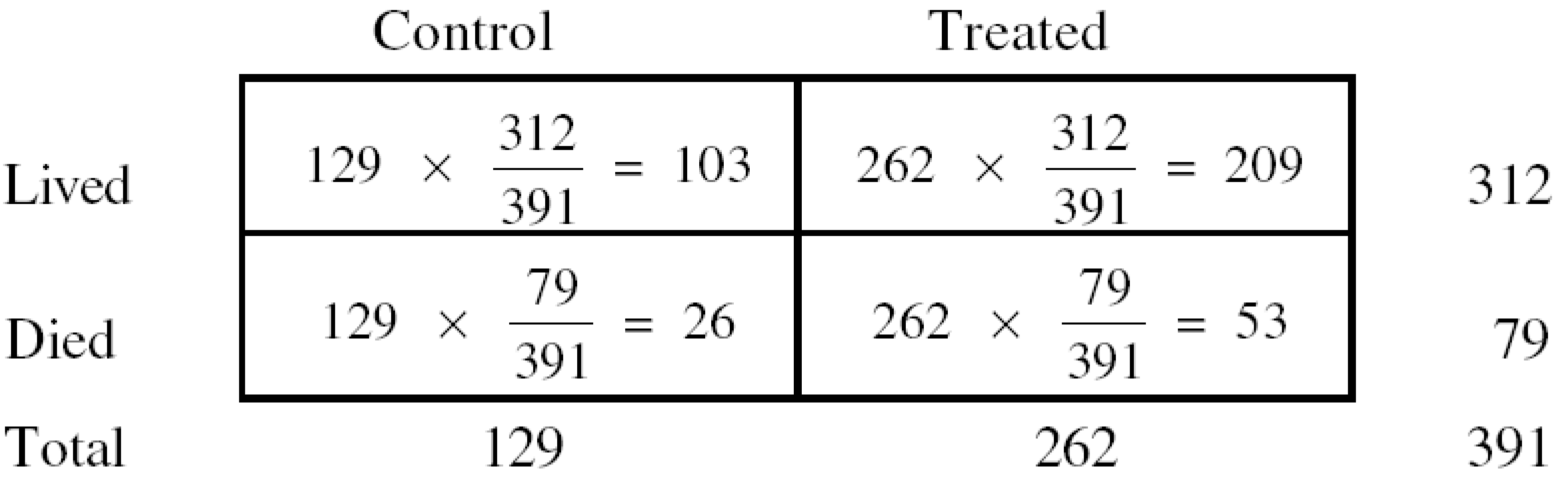


	Control	Treated	
Lived	a	b	$(a + b)$
Died	c	d	$(c + d)$
Total	$(a + c)$	$(b + d)$	N

Chi-square example

$$N \times P(\mathit{control\ and\ lived}) = N \times P(\mathit{control}) \times P(\mathit{lived})$$

$$= N\left[\frac{(a+c)}{N} \times \frac{(a+b)}{N}\right] = (a+c) \times \frac{(a+b)}{N}$$



	Control	Treated	
Lived	$129 \times \frac{312}{391} = 103$	$262 \times \frac{312}{391} = 209$	312
Died	$129 \times \frac{79}{391} = 26$	$262 \times \frac{79}{391} = 53$	79
Total	129	262	391

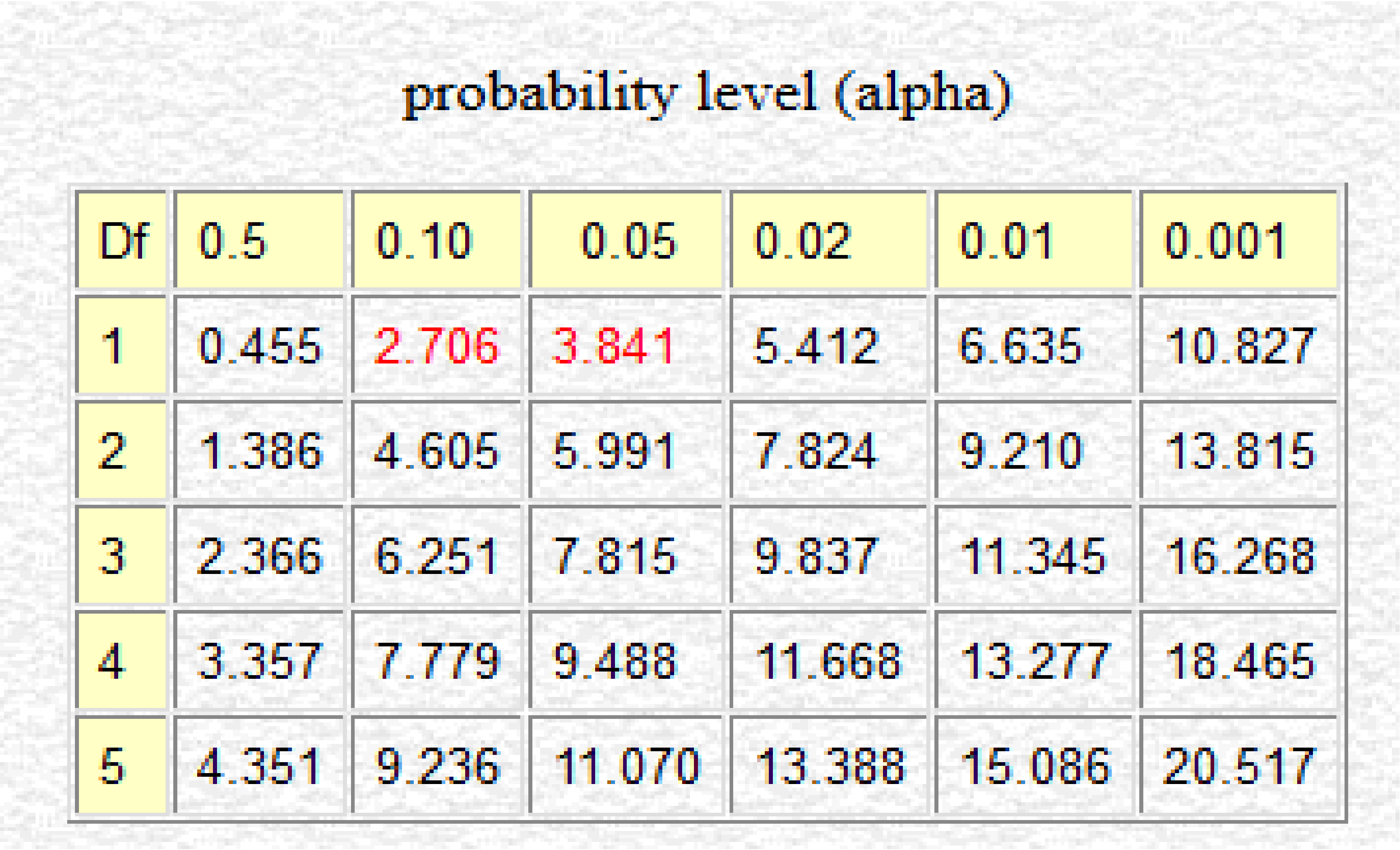
E = row total x column total / N

Chi-square example

- Calculate the chi square statistic x^2 by completing the following steps:
  - For each *observed* number in the table subtract the corresponding *expected* number (*O* — *E*).
  - Square the difference [(O —E)^2].
  - Divide the squares obtained for each cell in the table by the *expected* number for that cell [(O - E)^2 / E].
  - Sum all the values for (O - E)^2 / E. This is the chi square statistic.

	Observed	**Expected**	**O - E**	**(O-E)^2**	**(O-E)^2 /E**
Category 1					
Category 2					
Totals					

Chi-Square Table



probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

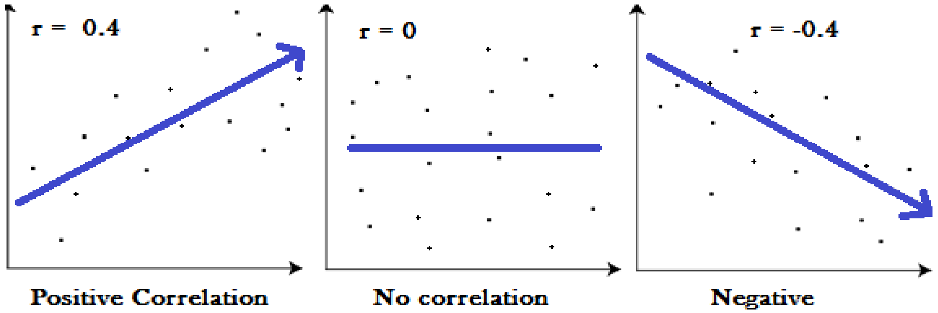
Correlation Analysis

- A quantitative measure of **association** between two continuous variables
  - The degree to which they change together

- **Pearson correlation** describes the strength (p value) and direction (r value) of a linear relationship
  - Usual between -1 to +1
  - Uses interval and ratio scale data for computation (Parametric)

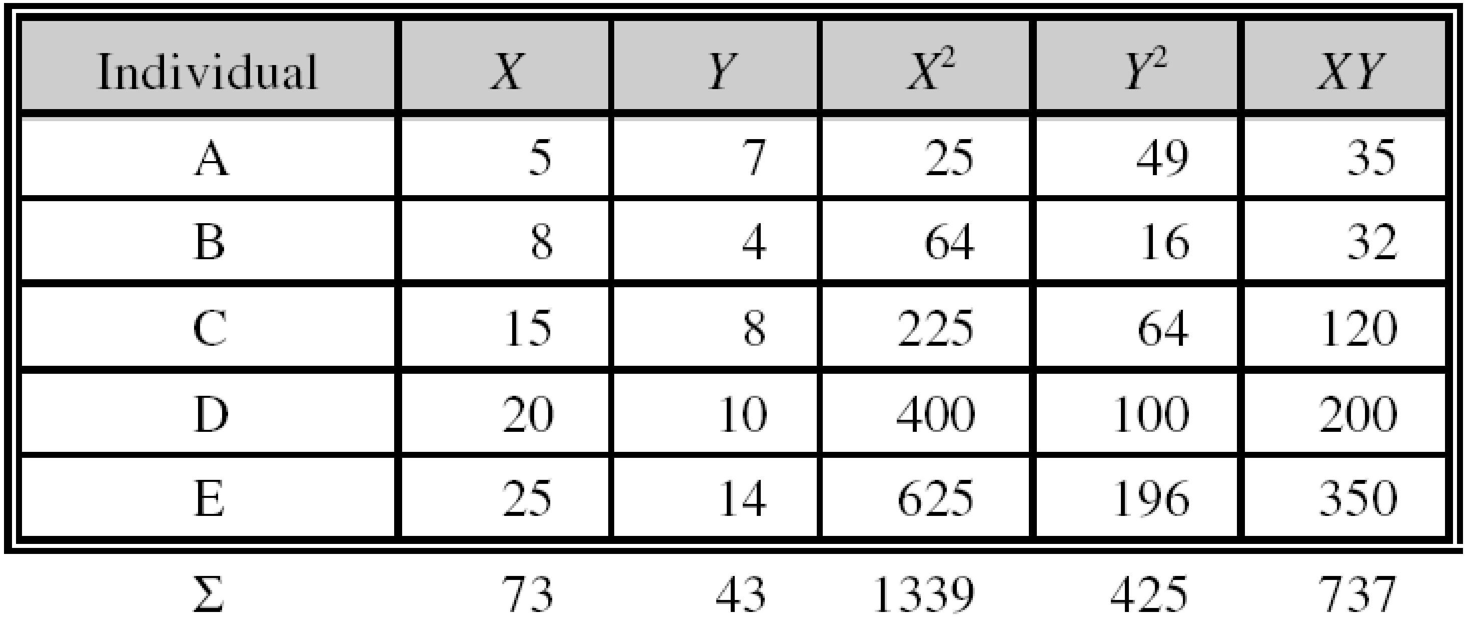
- **Spearman rank correlation** does not assume a linear relationship
  - Uses ordinal or ranked data (Non-parametric)

- Note:
  - Correlation ≠ estimate of slope
  - Correlation ≠ causation
  - Correlation ≠ agreement



Calculation of correlation coefficient

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$



Individual	X	Y	X^2	Y^2	XY
A	5	7	25	49	35
B	8	4	64	16	32
C	15	8	225	64	120
D	20	10	400	100	200
E	25	14	625	196	350
Σ	73	43	1339	425	737

Individual	X	Y	X^2	Y^2	XY
A	5	7	25	49	35
B	8	4	64	16	32
C	15	8	225	64	120
D	20	10	400	100	200
E	25	14	625	196	350
Σ	73	43	1339	425	737

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5(737) - (73)(43)}{\sqrt{5(1339) - (73)^2}\ \sqrt{5(425) - (43)^2}} = \frac{3685 - 3139}{\sqrt{1366}\ \sqrt{276}}$$

$$= \frac{546}{(37)\,(16.6)} = \frac{546}{614} = .89$$

Regression Analysis

- Determining relationships and making predictions
  - An extension of associations; use assumptions
  - Allows to estimate significance, direction/shape and magnitude of the effect
    - Determine if outcome is significantly affected by the predictors
    - Allows you to incorporate confounders (confounder-outcome relationships)
    - You can adjust predictor-outcome association for predictor-confounder and confounder-outcome relationships
  - Results may be used to predict the outcome of subjects that were not sampled but are from the same population
  - Specific regressions used based on outcome:
    - Simple/Multiple linear (continuous)
    - Logistic regression-categorical variables or binary logistic regression-two outcomes
    - Proportional odds (ordinal)
    - Cox/Proportional Hazards (time to event)

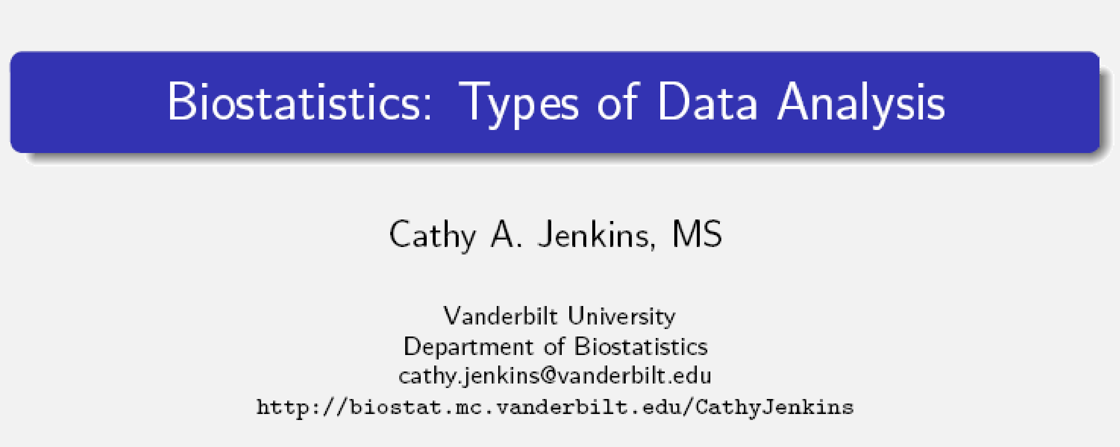
Things to avoid and pitfalls

- Having missing data
- Categorizing continuous variables
- Not using clinical knowledge to specify model
  - Statistical significance but no clinical significance
  - Clinical significance but no statistical significance

- Inappropriate linear assumptions
- Lack of model validation

- Report only favorable results
- Delete outliers based on observed responses
- Non-reproducible analyses/results

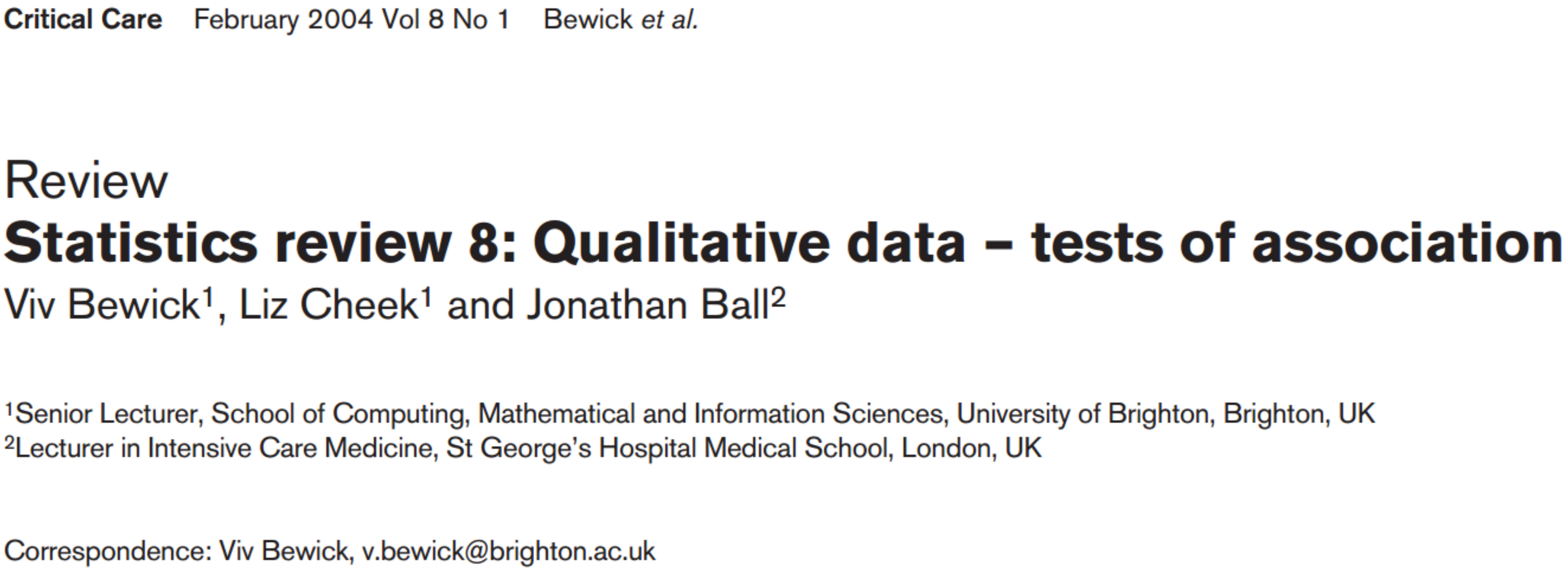
References and Acknowledgements



Biostatistics: Types of Data Analysis

Cathy A. Jenkins, MS

Vanderbilt University
Department of Biostatistics
cathy.jenkins@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/CathyJenkins



Critical Care February 2004 Vol 8 No 1 Bewick *et al.*

Review

Statistics review 8: Qualitative data – tests of association

Viv Bewick[1], Liz Cheek[1] and Jonathan Ball[2]

[1]Senior Lecturer, School of Computing, Mathematical and Information Sciences, University of Brighton, Brighton, UK
[2]Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Viv Bewick, v.bewick@brighton.ac.uk