

Interpreting Statistical data: making sense of the numbers

Dr David Chinn,
Formerly of NHS Fife
Research & Development Department,
Queen Margaret Hospital,
Dunfermline

davidchinn@nhs.net

djc20@st-andrews.ac.uk

Learning outcomes

- Distinguish between absolute and relative measures
- Explain the distinction between confidence intervals and P-values
- Interpret the results of one-sample, unmatched and paired t-tests
- Make sense of data presented in graphs
- Interpret the chi-square test for comparing proportions
- Interpret and make sense of the results from a randomised controlled trial
- Interpret an odds ratio

Sample paper: Sickness absence

Statements

- Mortality in Group A was 60% higher than that in Group B
- The mean length of stay was 4.3 days but the median was only 1 day
- Mean age of disease onset was 38.2 years in men and 43.3 years in women (mean difference 5.1 years, 95% confidence interval 3.5 to 6.7, $P=0.009$, unmatched t-test)
- The average reduction in diastolic blood pressure was 9 mm Hg (95% CI 4.5 to 13.1, $P<0.01$, paired t-test)
- The number needed to treat with drug A was 38 whereas that for drug B was 12

Statistics:

The science of assembling
and interpreting
numerical data

Concerned with 'estimation' and
'describing uncertainty'

Confidence interval

CI = A range of values in which the true mean for a population is likely to lie.

It usually has a proportion assigned to it (e.g. 95%)

It is derived from the data

The more observations (data) you have the smaller the confidence interval

What's the average age of onset of Huntington's Disease?

- Random sample 30 patients
- Youngest = 30, oldest = 70
- Hence, range = 30 - 70 years
- Mean age of onset = 48 years, SD = 7 years
- 95%CI of the average age = 45.4 to 50.6.

Hence, we are 95% confident that the true average age of onset of HD lies between 45 and 51 years.

Confidence intervals and P-values

CI: A range of values in which the true mean for a population is likely to lie. A narrow CI means your estimate is precise, a wide CI means your estimate is imprecise, due to either,

*small number of observations, or
simply a lot of variation in the data*

P-Value: used in hypothesis testing

The probability of occurrence of a result as extreme or more extreme than that observed if the null hypothesis were true.

Interpretation

P assesses how likely it is to observe such an effect in a sample when there is no such difference in the population from which the sample was drawn

$P < 0.05$ A statistically significant result !

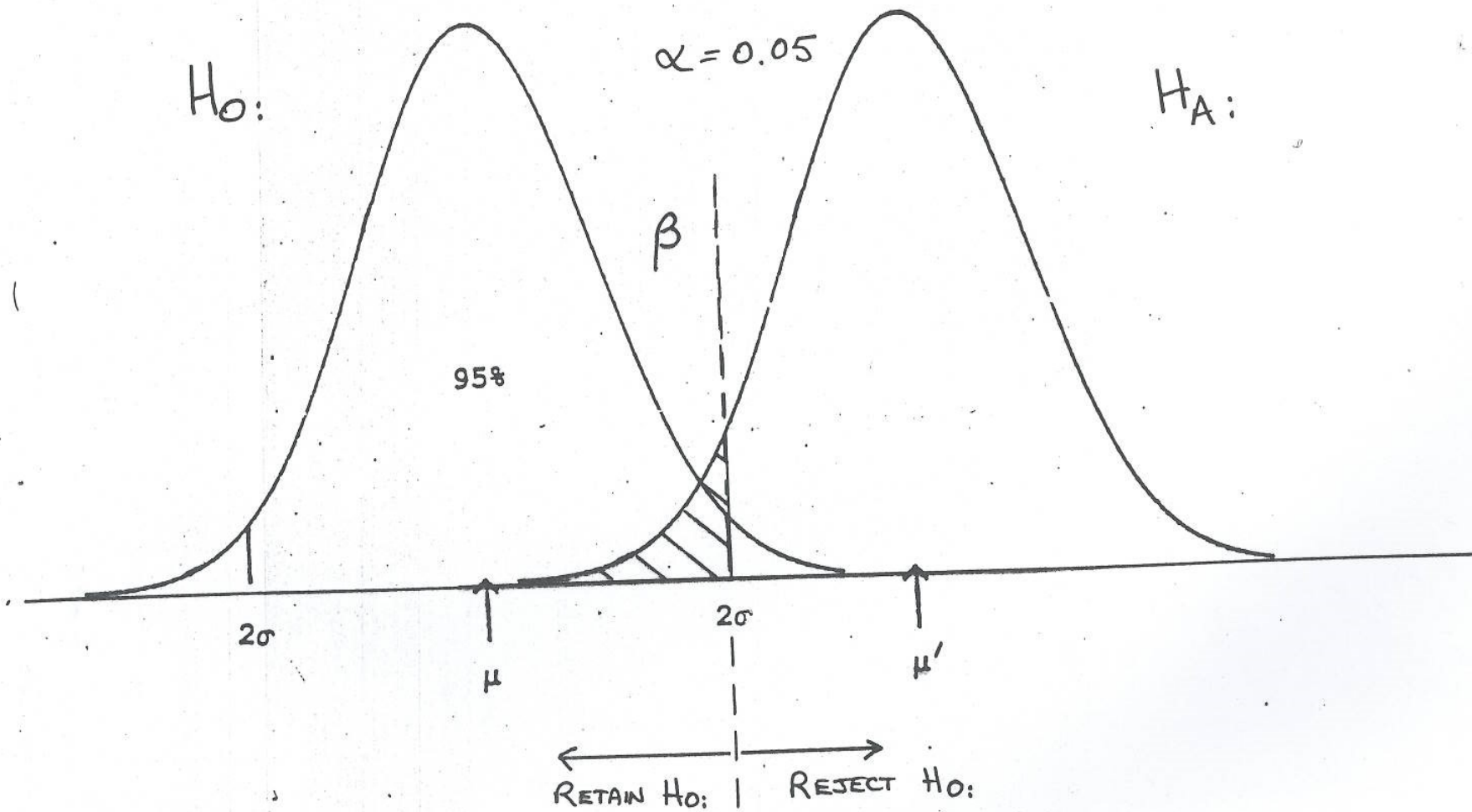
(the smaller the P-value the stronger the evidence against the null hypothesis of no difference between the groups)

($P = 0.05$ means we are prepared to be wrong 1 in 20, or 5%)

$P > 0.05$ "the data failed to reach statistical significance"

Meaning:

- (1) There is no difference between the groups, or
- (2) There were too few participants to demonstrate the difference between the groups if one truly exists



Example: Mortality in pts with heart failure

Drug A: 33% pts died; Drug B: 38% pts died

Difference = 5%, $P=0.07$, 95% CI= -1% to +12%

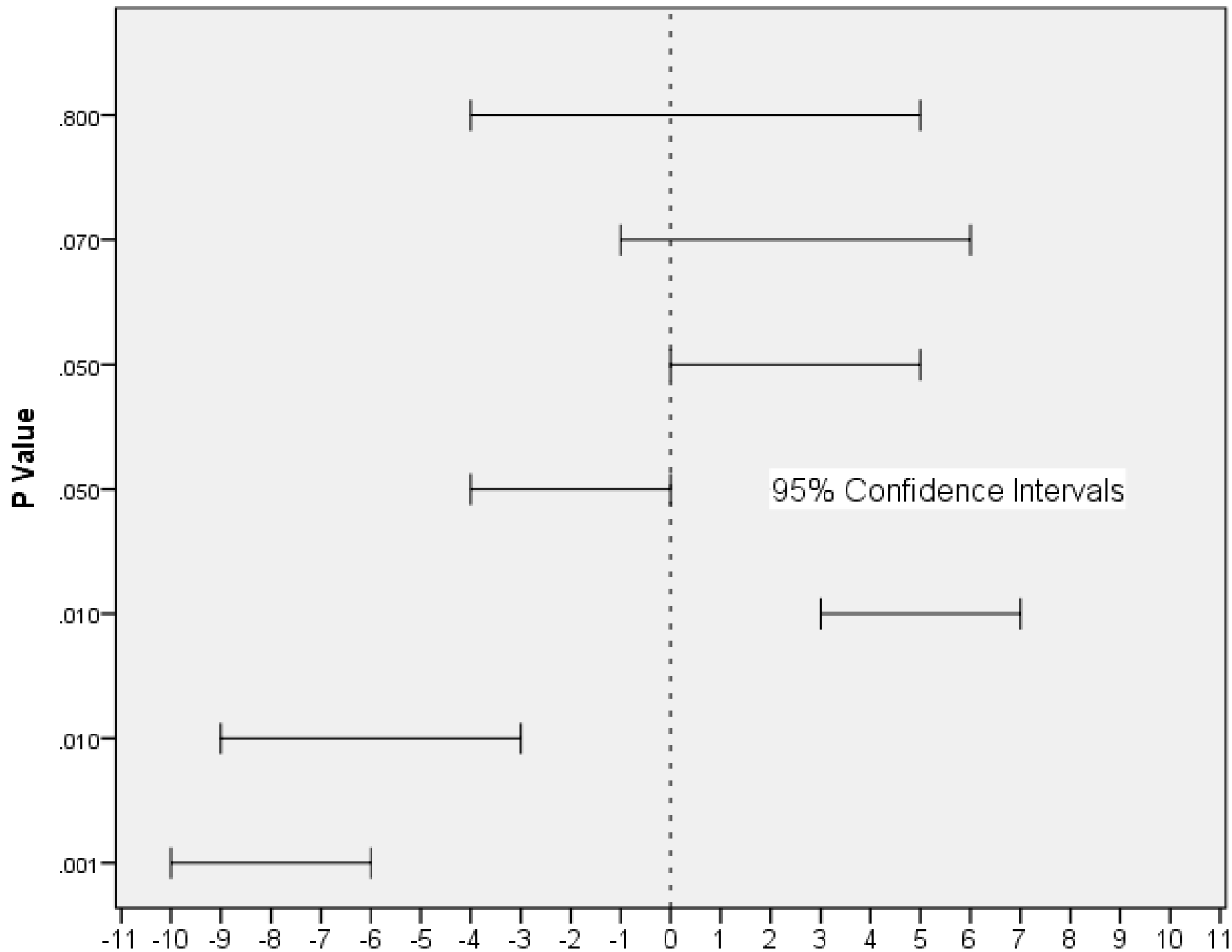
The 95% CI includes the value zero so is
not significantly different from zero at the 5% level
($P=0.07$)

But, we are 95% confident that the true difference in mortality between drugs A & B is between:

-1% (A worse than B), or

+12% (A better than B)

Conclusion: Drug A is likely to be better for pts with HF than Drug B but the evidence underpinning the inference is weak.



Confidence intervals or P-values ?

P-value will tell you whether or not there is a statistically significant difference between two groups

Confidence interval will give information about the **SIZE** of the difference and the strength of the evidence

Confidence intervals provide more information than a P-value alone and should always be cited

Beware of trials described as 'negative'

Lack of evidence of an effect is not the same
as evidence of no effect

Streptokinase and acute MI

24 randomised controlled trials:

19 'negative' ($P > 0.05$)

Streptokinase ineffective ?

But, a meta-analysis of data from all 24 trials showed a 22% increase in survival (95%CI +12% to +32%) ($P < 0.001$ so highly significant, and clinically worthwhile)

The 'negative' trials had led to a long delay in adopting the treatment !

Statistical versus clinical significance

Beware:

In very large studies small effects can be
statistically significant but have very little
clinical meaning

The t-test:

A parametric test used to compare data

Assumptions:

Data come from a normal (Gaussian) distribution

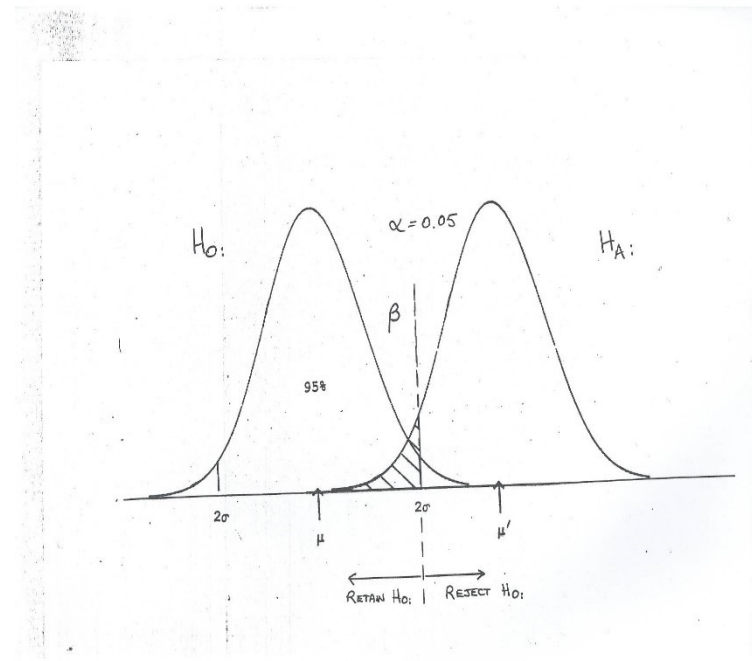
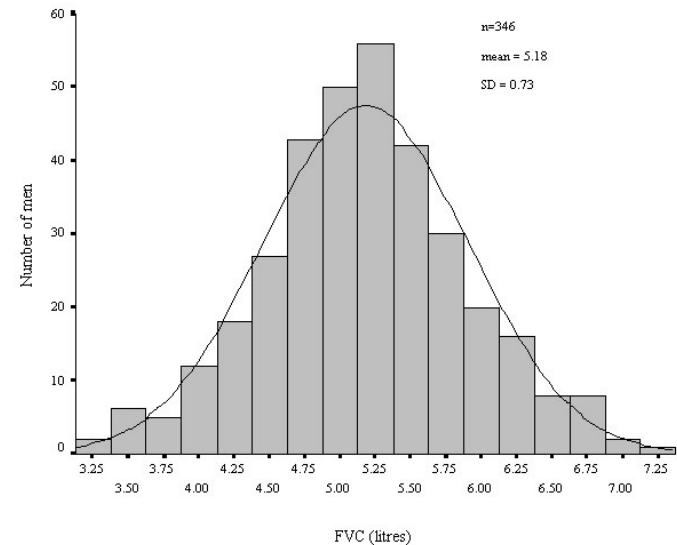
Samples are not too small

Samples do not contain outliers

(can be a big problem for small samples)

For comparison of 2 samples:

- samples of equal or nearly equal size
- variances equal or approximately so (not critical)

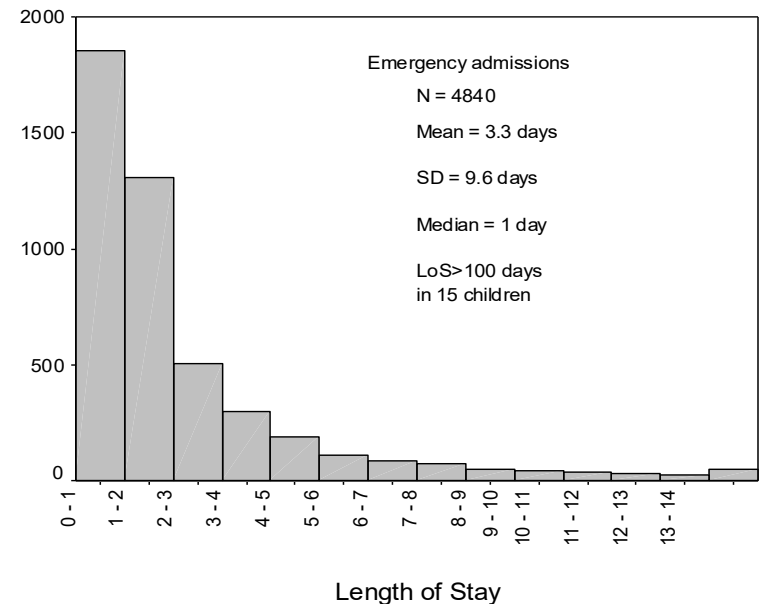
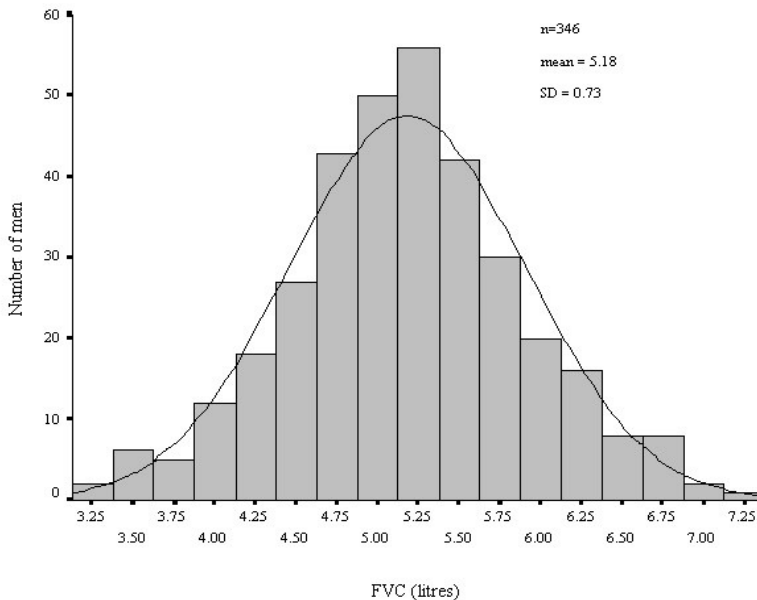


How to check if a distribution is 'normal'

- Check the distribution of data - histogram
- Compare Mean & Median
- Check mean & SD: (*variables that are positive*):
if $\text{Mean} < 2 \times \text{SD} \Rightarrow$ data likely skewed

Mean = 5.18, SD = 0.73

Mean = 3.3, SD = 9.6



One sample t-test

When you think you know what the true mean should be

Example haemoglobin:

Expected value for women = 13.6 g/dl.

58 women with Colorectal cancer:

Mean Hb on presentation=12.0, SD 1.9 g/dl

Mean (individual Hb - 13.6) = -1.6

95%CI = -2.1 to -1.1, $P < 0.001$

Comparison of 2 groups: unmatched t-test

Two independent samples

Example: haemoglobin in 58 women
with Colorectal cancer

Qu: Does the degree of anaemia differ between
those with left-sided and right-sided disease?

Mean Hb on presentation = 12.0, SD 1.9 g/dl

Mean Hb: (L-sided, n=36) = 12.9, SD 1.5

(R-sided, n=22) = 10.7, SD 1.6

Mean difference = -2.2 g/dl

95%CI = -3.0 to -1.4, $P < 0.001$

Paired t-test

Similar to one-sample t-test

One group of subjects - two measurements on each subject (e.g. before and after a drug)

Null hypothesis, mean difference = 0

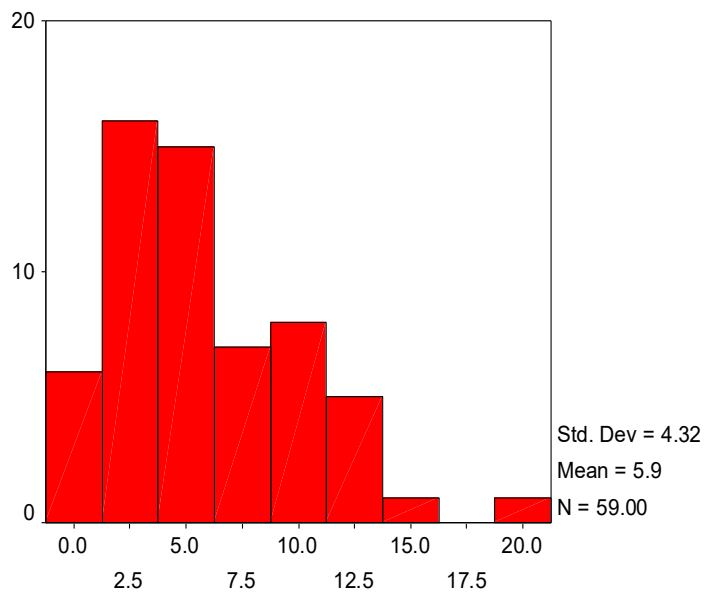
Example -change in disease activity in Ulcerative colitis over 3 months (n=59):

Baseline score, mean 5.9 SD 4.3

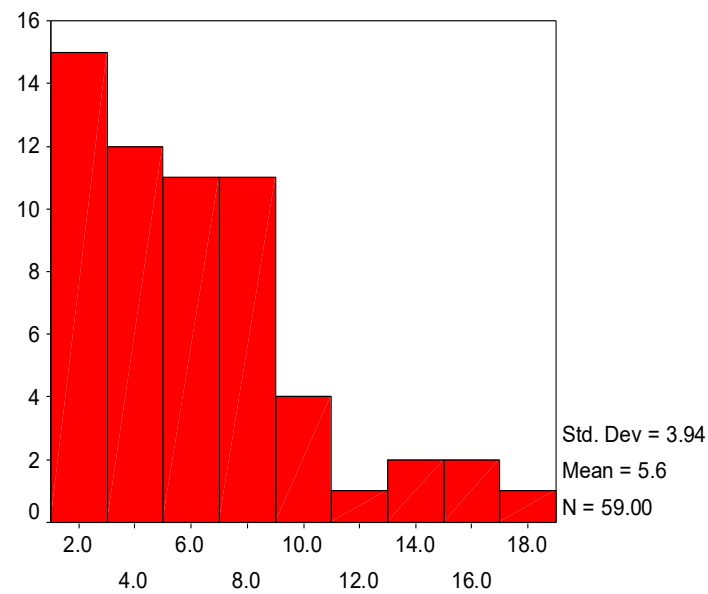
Follow-up score, mean 5.6 SD 3.9

Difference in score, mean -0.3 (SD2.7)

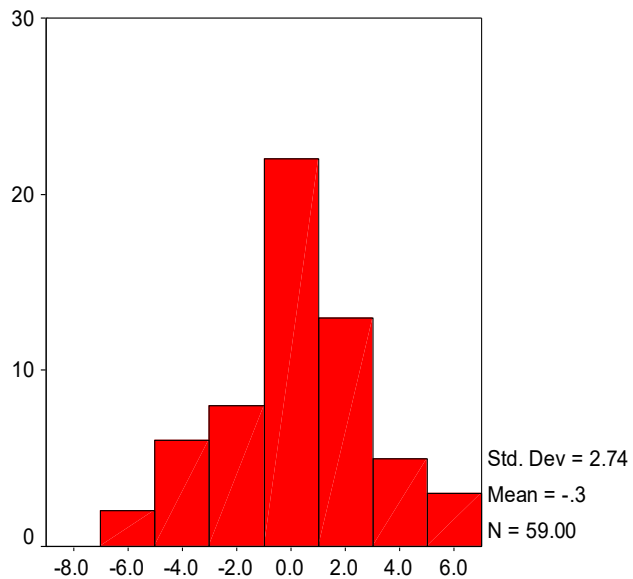
95% CI -1.0 to + 0.4, P=0.4



UC-Disease activity score (baseline)



UC-Disease activity score (follow-up)



DIFFERENCE DISEASE ACTIVITY

Could we compare baseline and follow-up scores using the Unmatched t-test ?

No !

Because the two samples are not independent

One aspect of critical appraisal is to decide if the correct stats tests have been used

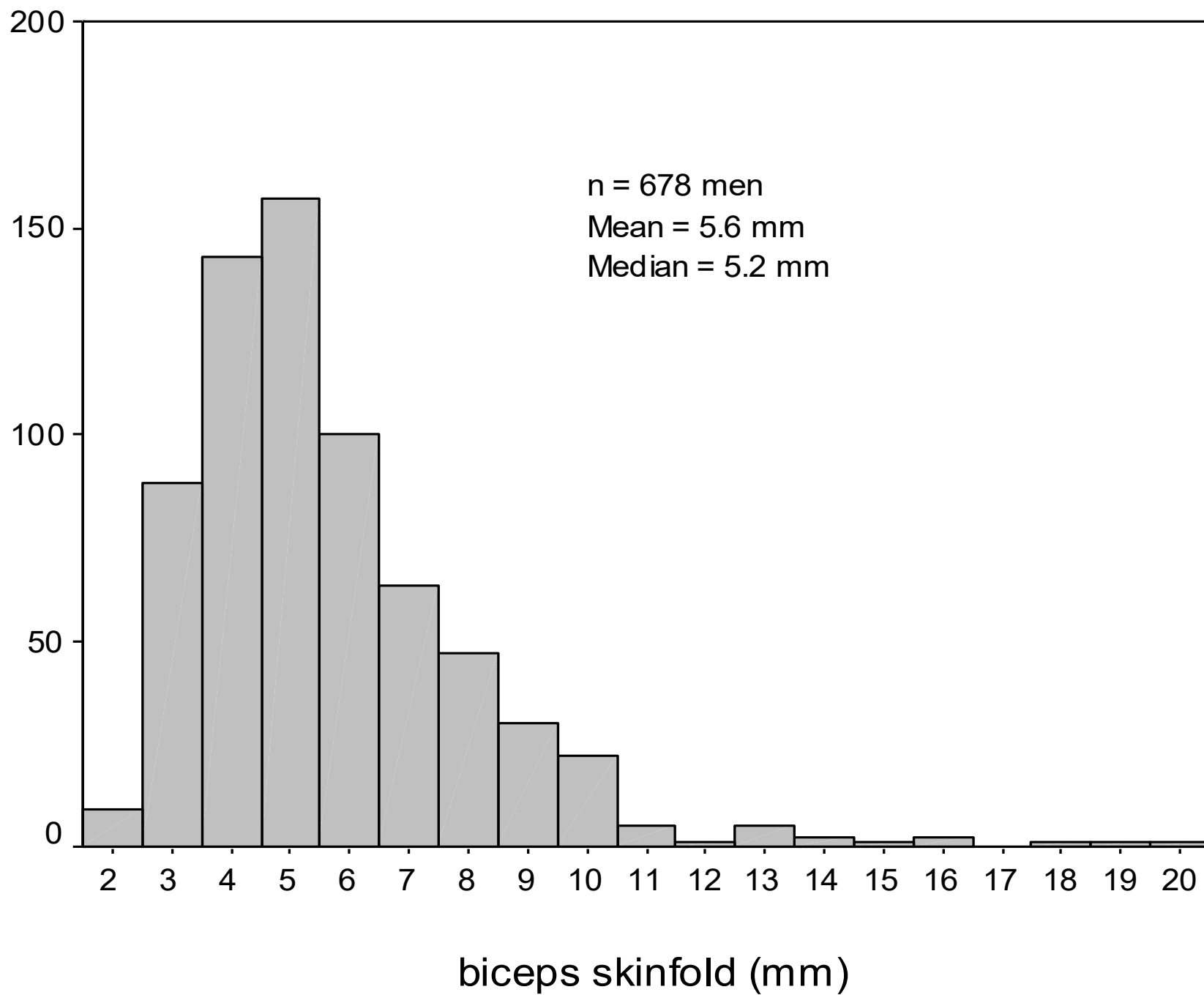
What if the assumptions are not met?

(1) Use non-parametric (distribution-free) tests which make no assumptions about the distribution of the data.

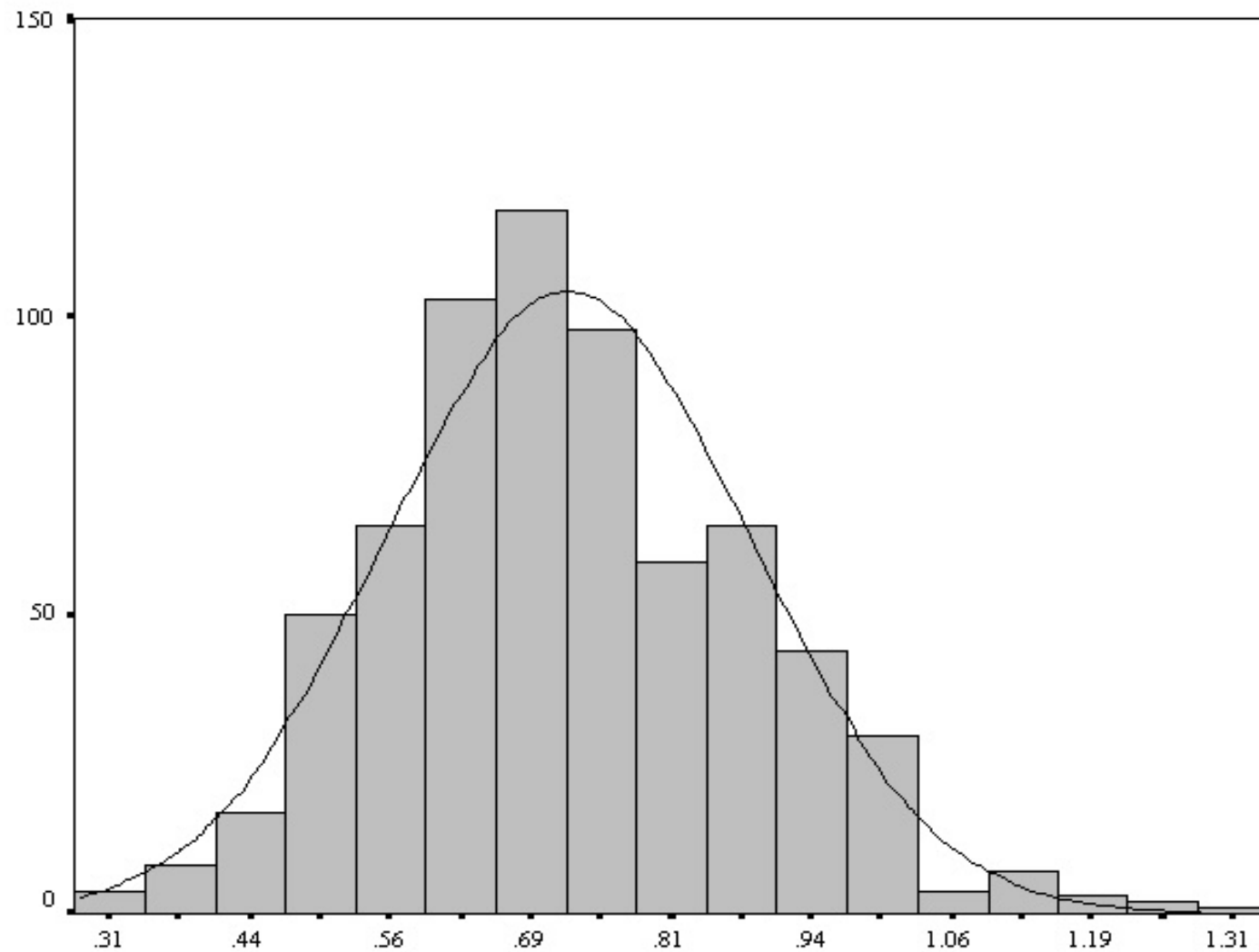
- *Effectively compares medians, not means*
- *Will tell you whether 2 groups differ statistically but not by how much*

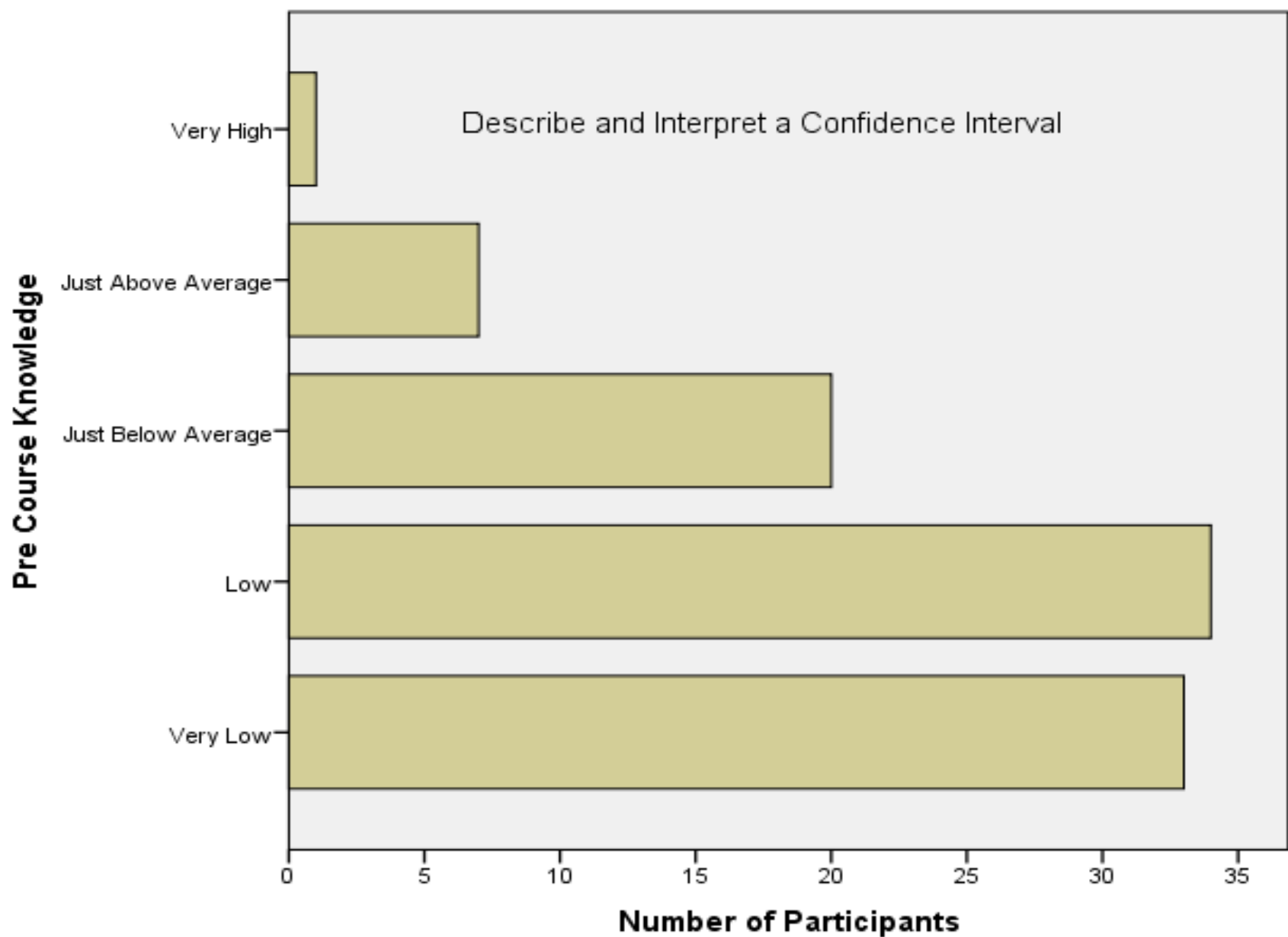
(2) Transform the data so they do meet the criteria

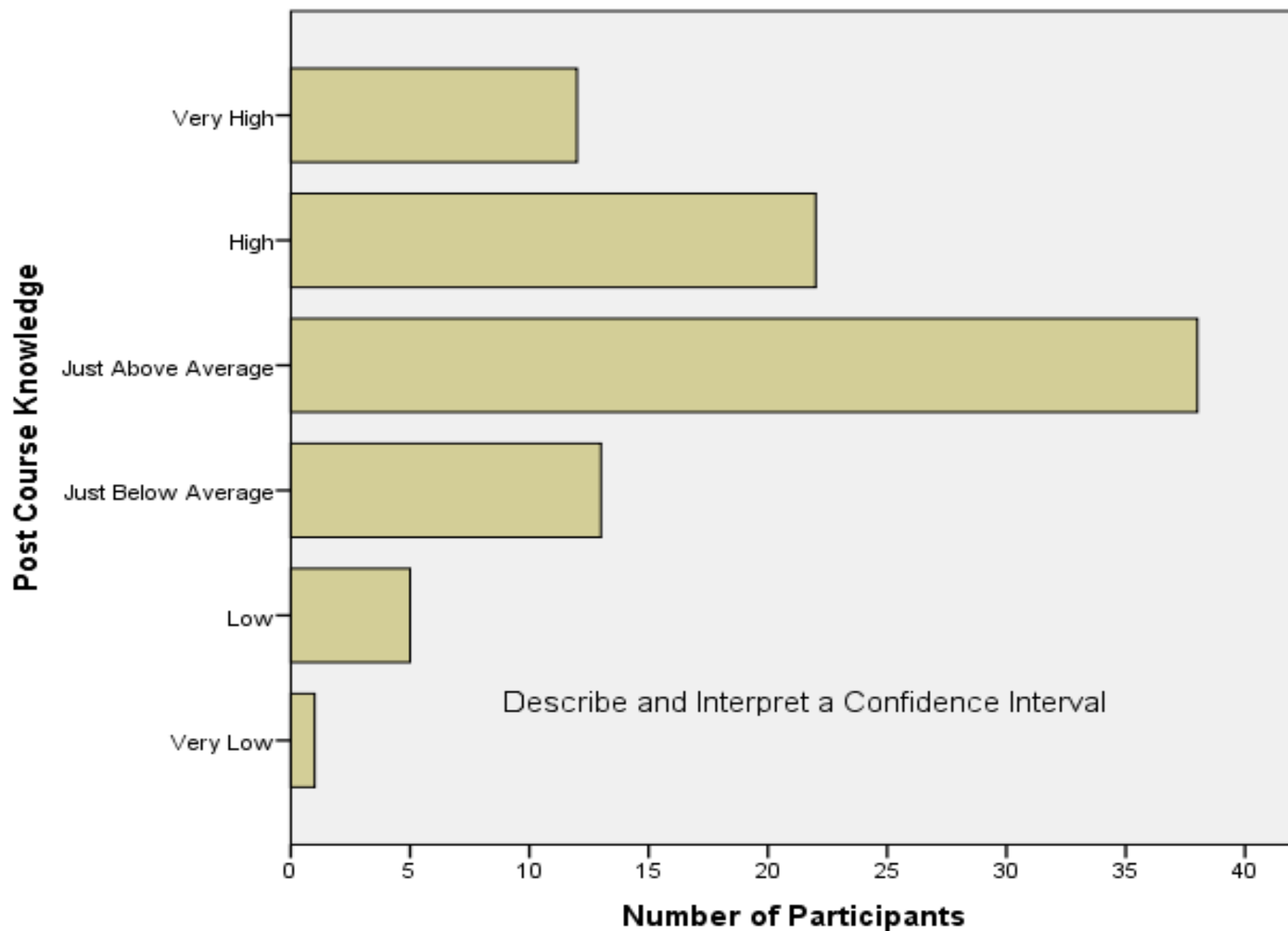
- $\text{Log}_{10}(x)$
- $1 / x$
- x^2, x^3

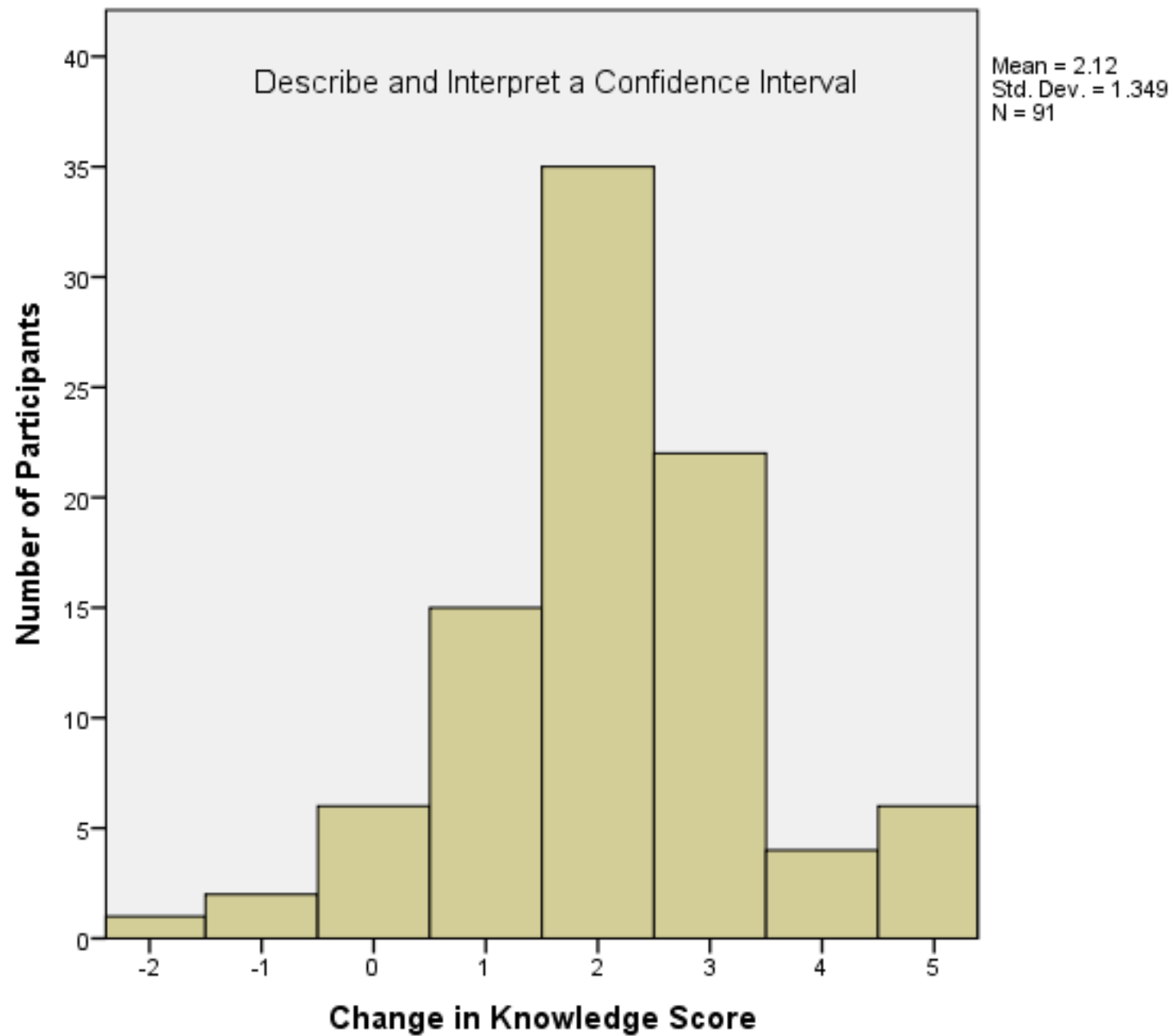


Log_{10} of biceps









Consequences of using the wrong test

Blood loss during surgery (t-test, $P = 0.108$)

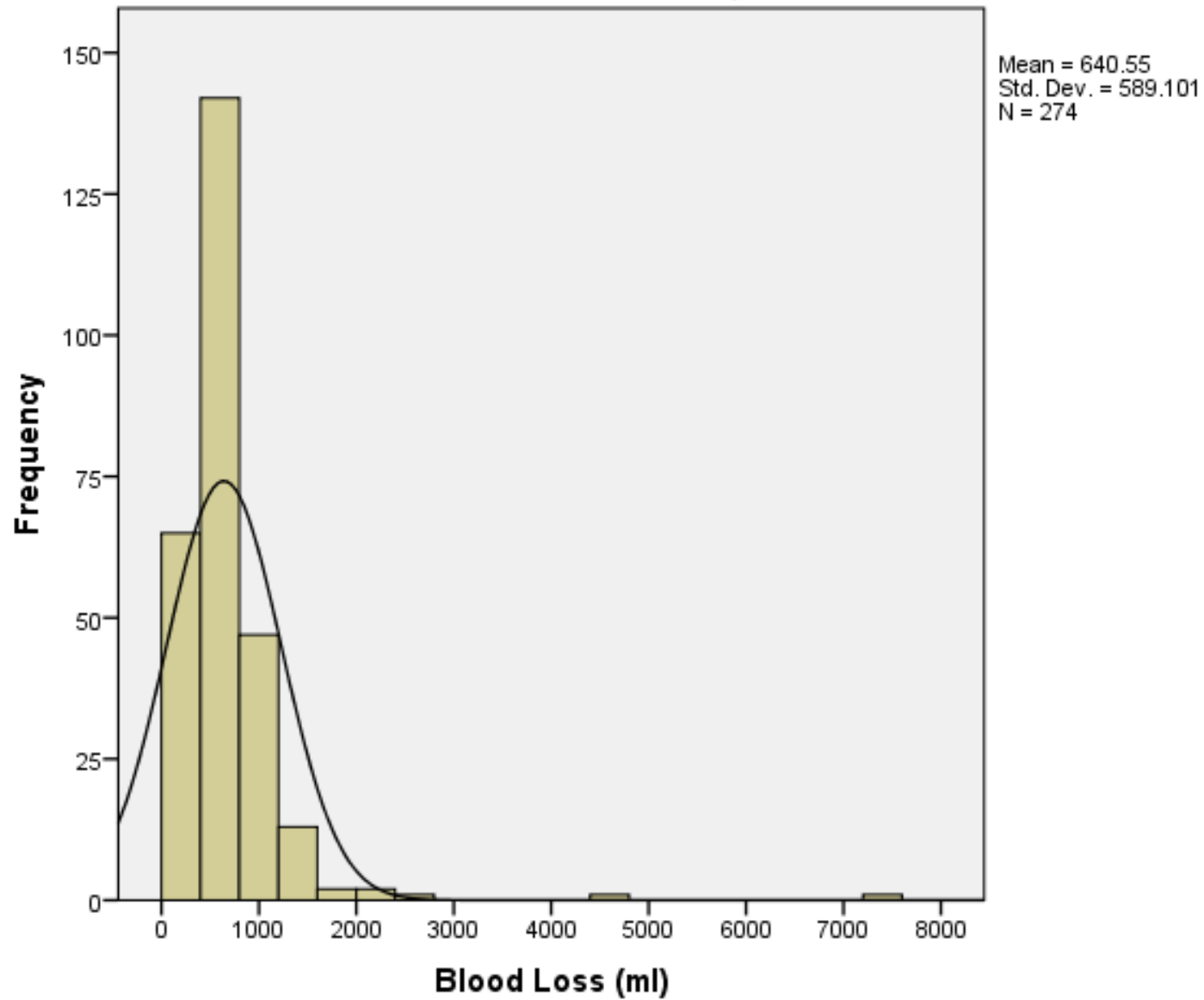
Blood loss (ml)	Technique 1	Technique 2
n	274	45
Mean	640	789
SD	589	444

Consequences of using the wrong test

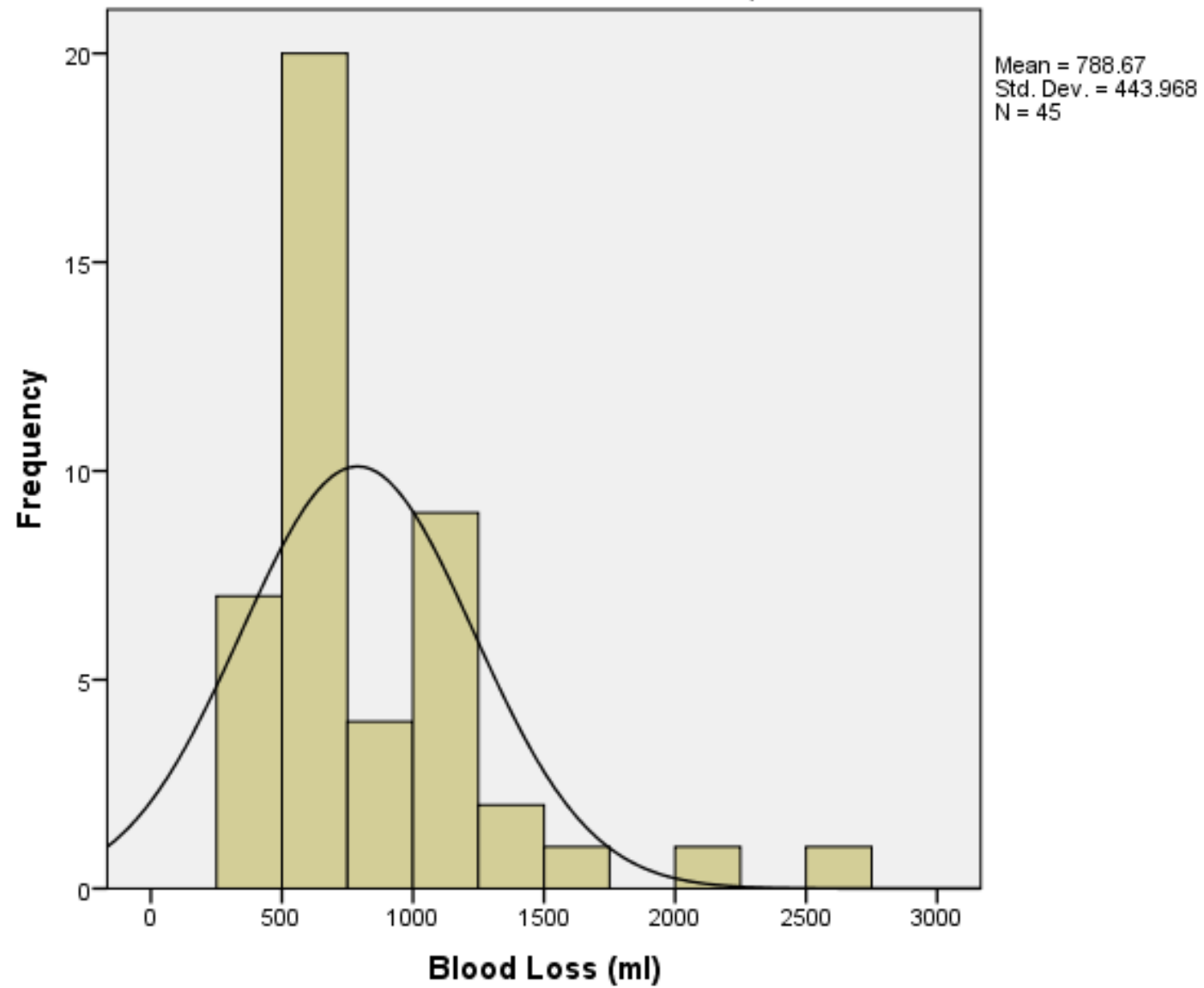
Blood loss during surgery (t-test, $P = 0.108$)

Blood loss (ml)	Technique 1	Technique 2
n	274	45
Mean	640	789
Median	500	700
SD	589	444
Minimum	100	250
Maximum	7500	2500

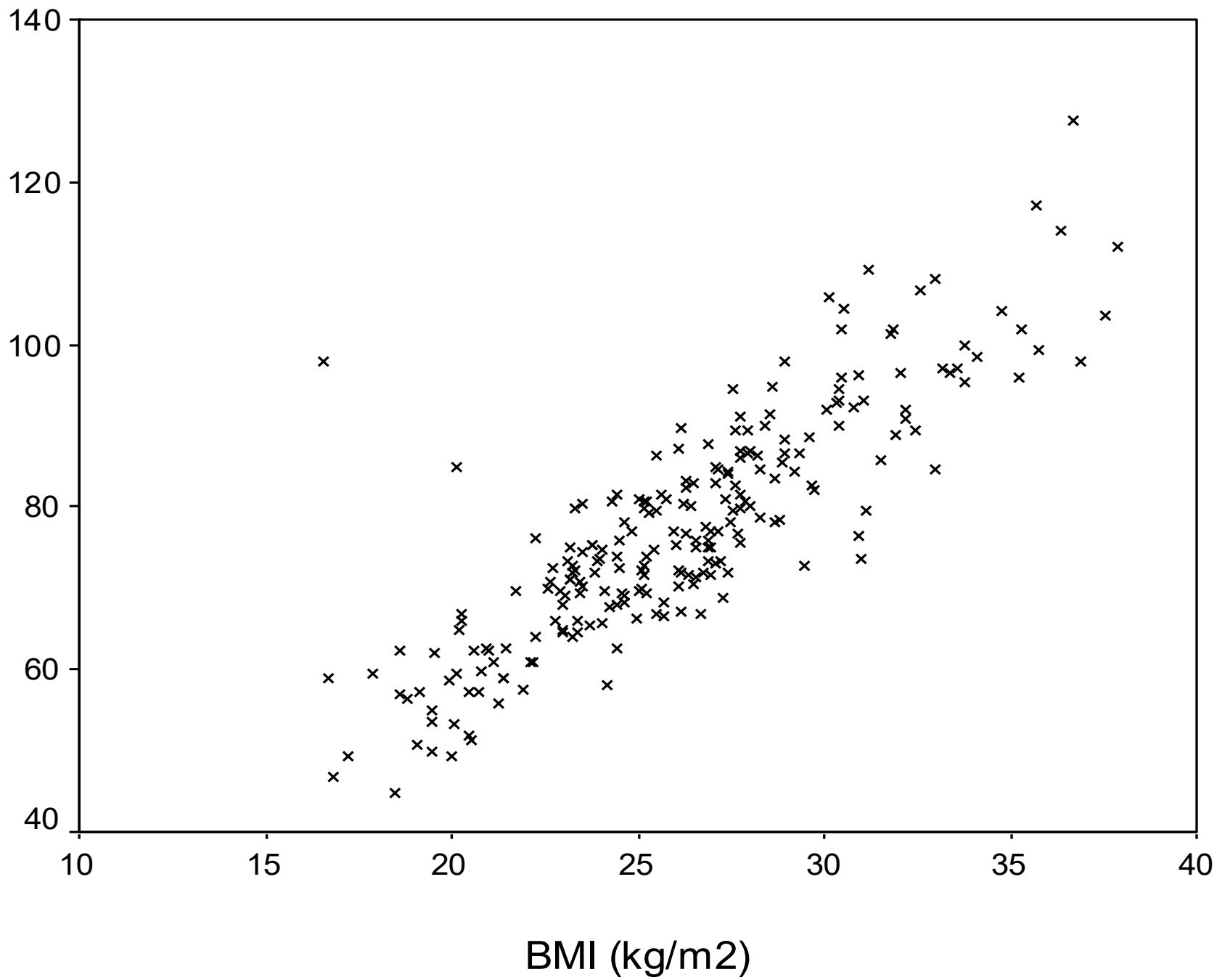
SURGICAL TECHNIQUE 1



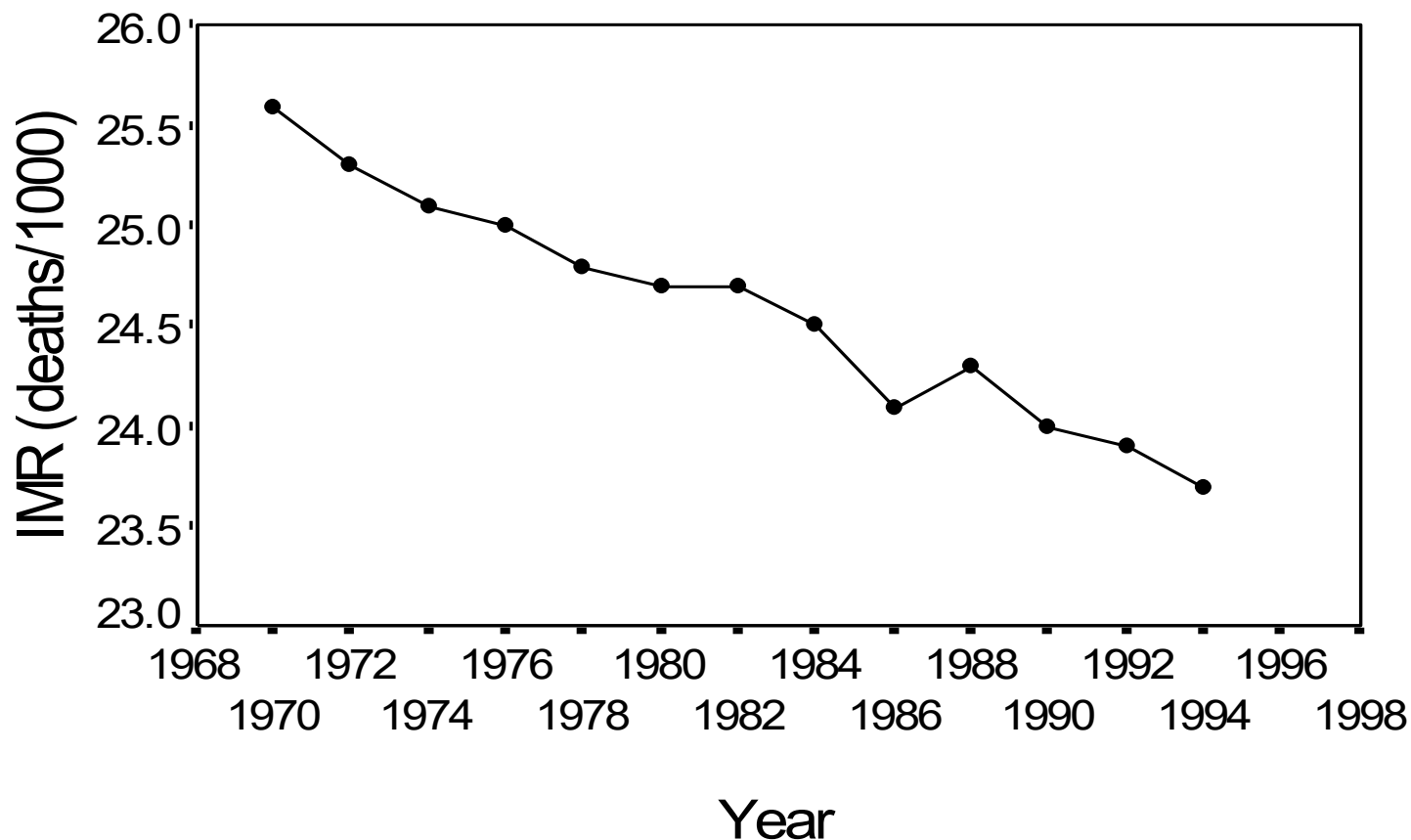
SURGICAL TECHNIQUE 2



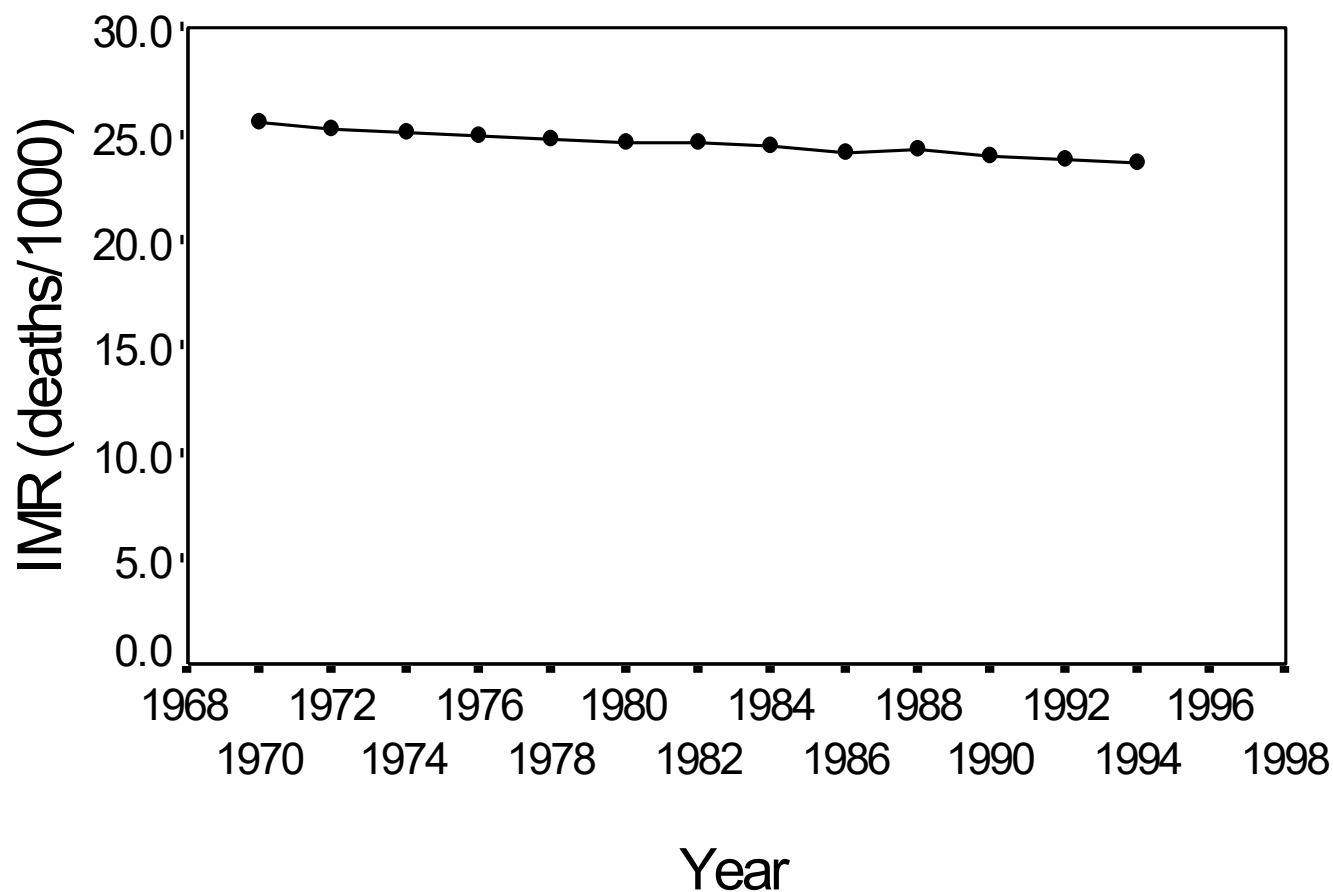
Blood loss (ml)	Technique 1	Technique 2	Statistical test	
n	274	45	t-test P = 0.108	Mann-Whitney U P = 0.002
Mean	640	789		
Median	500	700		
SD	589	444		
Minimum	100	250		
Maximum	7500	2500		
25 th percentile	400	500		
75 th percentile	750	1000		

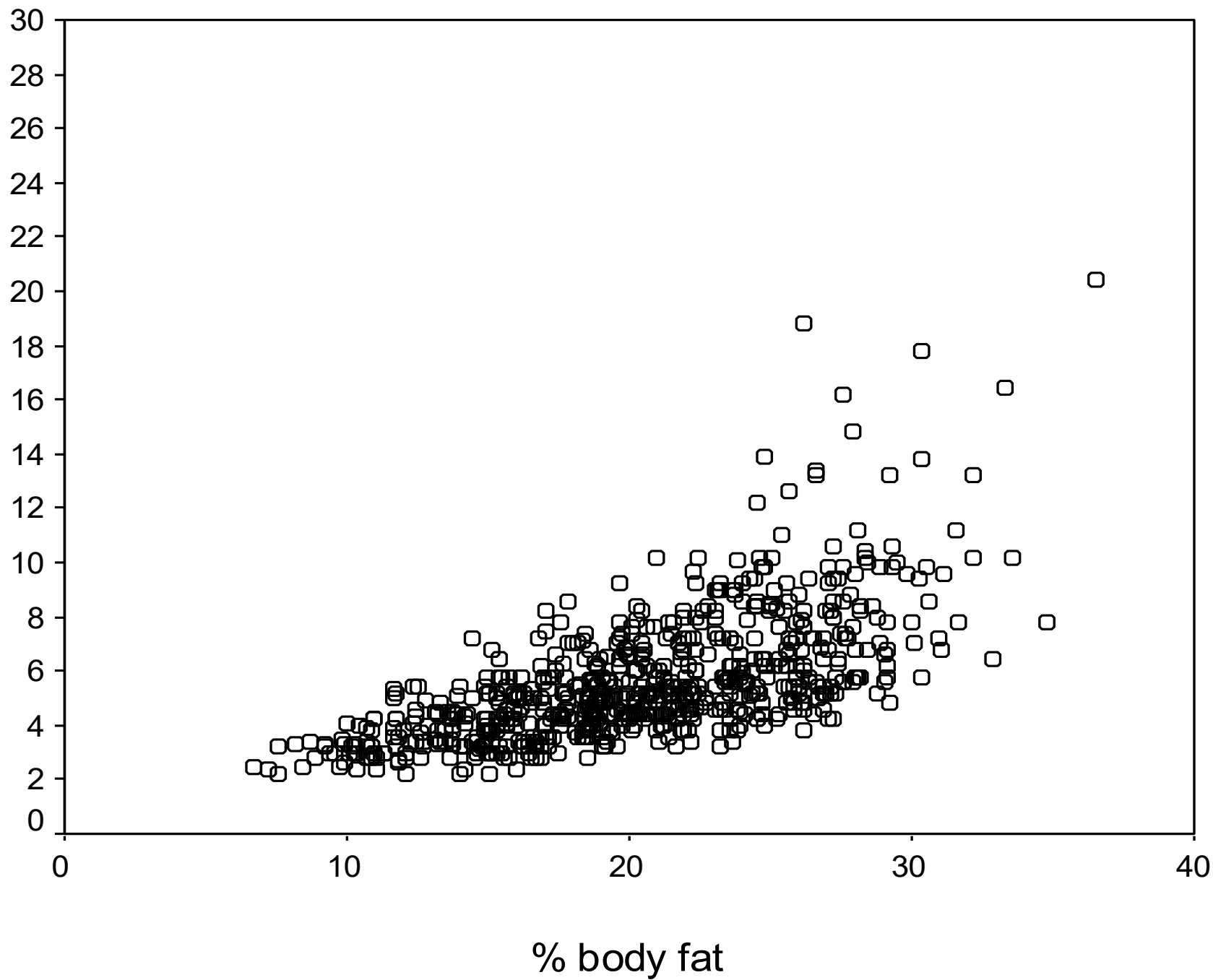


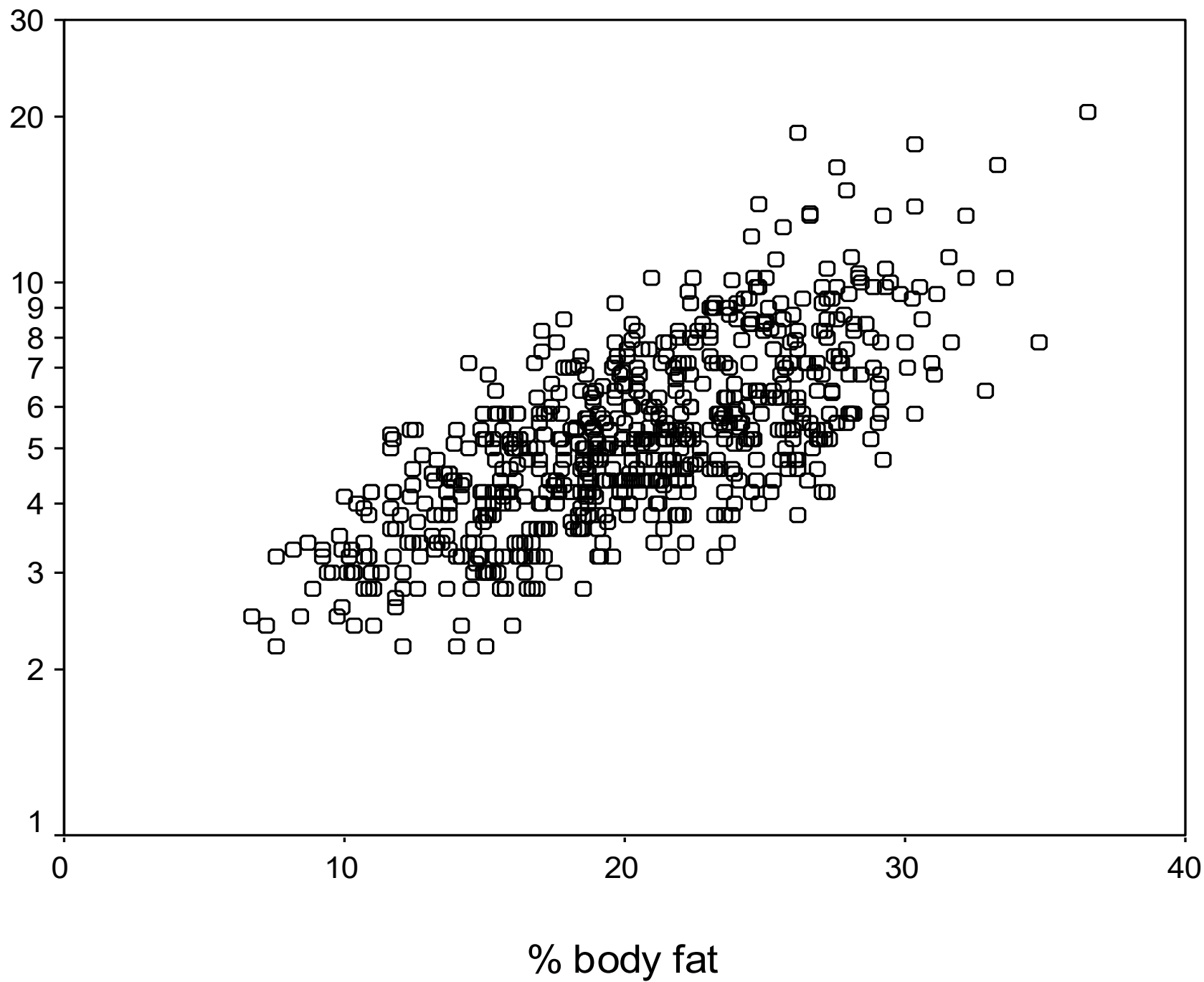
Infant Mortality Rate (deaths / 1000 live births)



Infant Mortality Rate (deaths / 1000 live births)







Chi-square test

A common test to compare frequencies in 2 or more groups.

Research question: In women eligible for breast screening does a personalised letter from the GP improve uptake?

RCT, n= 470 women, 2 groups

Attended for mammography	+ letter (intervention)	- letter (control)	N
Yes	134	120	254
No	102	114	216
N	236	234	470

254/470 (54%) attended
+ letter 134/236 (57%) attended
- letter 120/234 (51%) attended

Attended for mammography	+ letter (intervention)	- letter (control)	N
Yes	134 (127.5)	120 (126.4)	254
No	102 (108.5)	114 (107.6)	216
N	236	234	470

$$= \sum (\text{observed} - \text{expected})^2 / \text{expected}$$

$$\text{chi-square} = 1.42$$

$$P=0.23$$

What is the interpretation?

In this trial there is no evidence that a personalised letter from the GP would improve the uptake of breast screening among the population of women from which the sample was drawn.

But,

Lack of evidence of an effect is not the same as evidence of no effect

95% CI of the difference =

-3.5% (GP letter reduced uptake)

to +14.5% (GP letter improved uptake)

Interpretation of clinical trial data

Qu: is aspirin effective in reducing the incidence of MI ?

RCT, n=22,071 men, random allocation, to get either aspirin or placebo (1 tablet / day)

Outcome = MI attack rate over 1 year

Null Hypothesis: Attack rate same in both groups

Aspirin: $MI = 139 / 11037 = 0.0126$ (1.26%)

Placebo: $MI = 239 / 11034 = 0.0217$ (2.17%)

difference = 0.0091 (0.91%)

95% CI = 0.0057 to 0.0125

sig test, $P = 0.00001$

Aspirin and MI: interpretation

Relative risk:

$$\frac{\text{attack rate placebo group}}{\text{attack rate aspirin group}} \\ = 0.0217 / 0.0126 = 1.7$$

So, members of the placebo group are 1.7 times more likely to have an MI than those in the aspirin group.

Aspirin and MI: interpretation

Relative risk reduction (RRR)

The difference in attack rates, ignoring the sign, divided by the attack rate in the placebo group

$$= | 0.0126 - 0.0217 | / 0.0217 = 0.419 \text{ or } \sim 42\%$$

So, aspirin is associated with a 42% reduction in adverse outcome.

Aspirin and MI: interpretation

Absolute risk reduction (ARR):

The difference in attack rates, ignoring the sign:

$$= | 0.0126 - 0.0217 | = 0.0091 \text{ or } \sim 0.9\%$$

Number needed to treat (NNT)

$$\text{NNT} = 1 / \text{ARR} = 1 / 0.0091 = 110$$

So, we need to treat 110 pts with aspirin for a year to prevent one MI

What else ?

NNH - adverse events e.g. GI bleed

4 Rehab programmes: which one should we fund?

Prog 1 - absolute ↓ in deaths of 3%

Prog 2 - ↑ survival from 84% to 87%

Prog 3 - ↓ death rates by 19%

Prog 4 - 33 pts needed to avoid 1 death

The evidence:

RCT rehab prog vs no rehab

Death rate rehab = 13% (survival 87%)

Death rate control = 16% (survival 84%)

Reduction in death rate = 3%

Proportional reduction in deaths = $3\% / 16\% = 19\%$

NNT = $1 / 0.03$ (or $100 / 3$) = 33

See: Fahey et al BMJ 1995; 311: 1056-1059.

140 board members responded but only 3 identified the summary stats were from the same programme

Odds & odds ratio

Odds = ratio of the number of times an event occurs to the number of times it does not occur from a given number of chances. Used to quantify the 'risk' of something happening.

Odds ratio = a comparison of odds between two groups to quantify 'relative risk'. If the odds are the same in the two groups the odds ratio is 1.

Road accident statistics, Scotland, 2007

	Casualties	Fatalities	% deaths
Car	9953	160	1.6
Pedestrian	2682	61	2.3
Motorcycle	1039	40	3.8
Other	??	21	
Totals	13,674 ++	282	

Road accident statistics, Scotland, 2007

	Fatal	Non-fatal	Casualties
Motorcycle	40	999	1039
Car	160	9793	9953
Totals	200	10792	10992

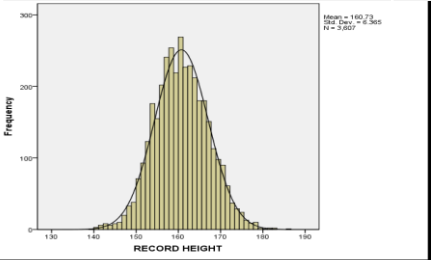
$$\begin{aligned}\text{Ratio of odds} \\ &= (40/999) / \\ &\quad (160/9793) \\ &= 2.45\end{aligned}$$

$$\begin{aligned}95\%CI \\ &= 1.72 \text{ to } 3.48\end{aligned}$$

Interpretation ?

Which statistical test ?

Categorical Data				
2 Categories (investigating proportions)				>2 Categories
1 group	2 groups		>2 groups	
	Paired	Independent		
z-test for a proportion	McNemar's test	Chi-squared test	Chi-squared test	Chi-squared test
Sign test		Fisher's Exact test	Chi-squared trend test	

	Numerical Data			
	1 group	2 groups		> 2 groups
		Paired	Independent	Independent
Parametric	One sample t-test 	Paired t-test	Unpaired t-test	One-way ANOVA
Non-parametric				
	Sign test	Sign test	Wilcoxon rank sum test	Kruskal-Wallis test
		Wilcoxon signed ranks test	Mann-Whitney U-test	

No amount of clever statistics can salvage a badly designed study

Be circumspect when interpreting data

Think beyond the numbers

Awareness of numbers

Confidence intervals

P-values

T-tests

Presentation & interpretation of data in graphs

Interpretation of the Chi-square test &
comparison of proportions

Interpretation of clinical trial data

Interpretation of the odds ratio

Sample Paper

The impact of a work-based fitness program on
sickness absence of female employees

Recommended texts

- Interpreting Statistical Findings: A guide for health professionals and students. Jan Walker & Palo Almond, 2010, Open University Press.
- Medical Statistics at a Glance. 3rd ed. Aviva Petrie & Caroline Sabin, 2009, Blackwell Publishing
- Essential Statistics for Medical Examinations. 2nd ed. Brian Faragher and Chris Marguerie, 2005, PASTEST
- Practical Statistics for Medical Research. Douglas G Altman, 1991, Chapman and Hall. (new ed due 2011)
- An Introduction to Medical Statistics. 3rd ed. Martin Bland, 2000, Oxford Medical Publications.
- Essential Medical Statistics. 2nd ed. Betty Kirkwood & Jonathan Sterne, 2003, Blackwell Scientific Publications.
- A-Z of Medical Statistics. F. Pereira Maxwell, 1998, Arnold.
- Statistical Questions in Evidence-Based Medicine. Martin Bland & Janet Peacock, 2000, Oxford Medical Publications.

Prevalence of diabetes

/1000 persons

Region	Scheme	Males	Females
England	(GPRD)	21	18
Northern RHA	(GPRD)	20	17
Northumberland	(MEDICS)	20	16
Yorkshire	(GPRD)	22	20
Yorkshire	(WAPPCHIP)	23	17
Somerset		19	17
Teesside		19	16

Figure 26 Prevalence of diabetes

