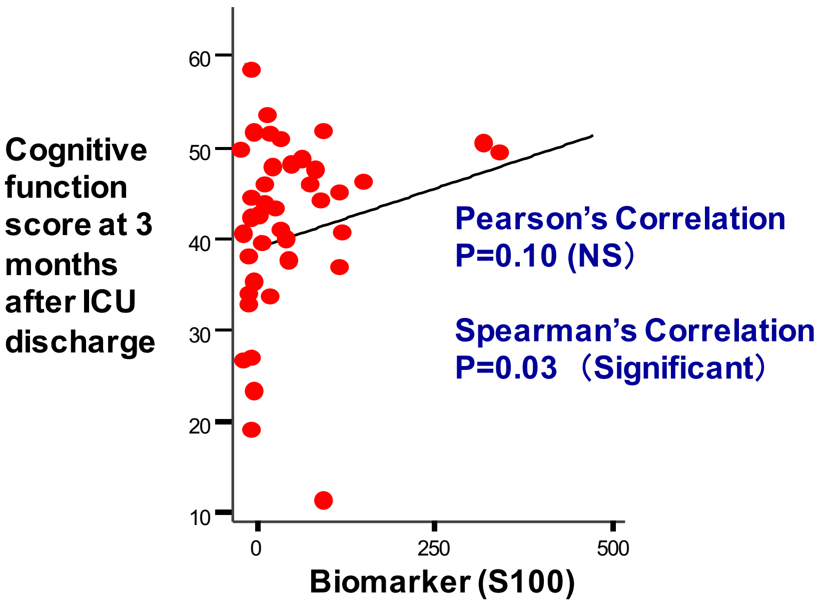
Selecting Proper Statistical Tests for Evidence Based Medicine

Overview:

1.1 Why the selection of valid statistical tests is important?
1.2 What are the factors to be considered for test selection?
1.3 Can you select now? (Tutorials)

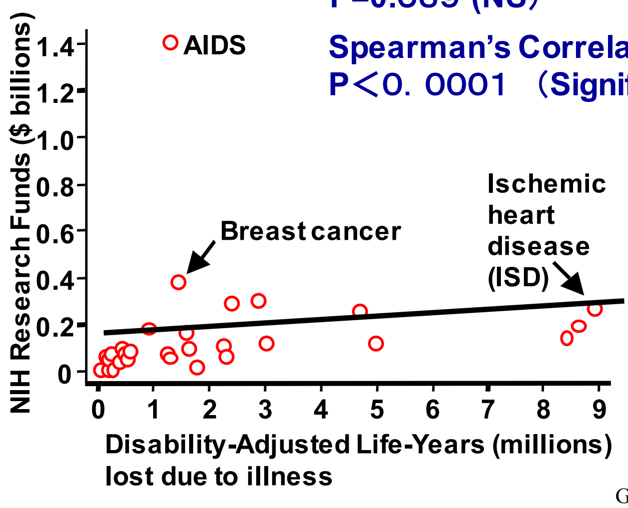
Example 1 : Different tests → Different results



Example 2: Different tests → Different results

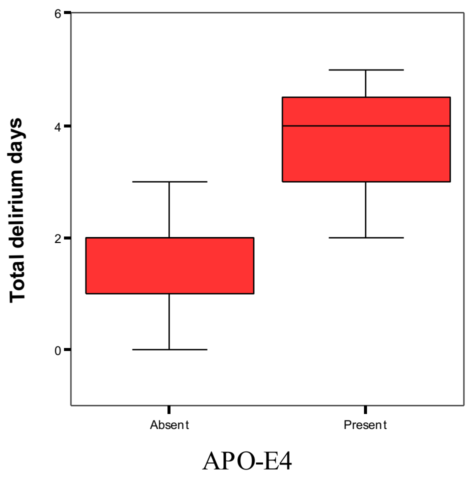
Pearson's Correlation
P=0.539 (NS)

Spearman's Correlation
P<0. 0001 (Significant)



Gross et al. (1999)

Example 3: Different tests → Different results



Student's T-test
P=0.405 (NS)

Mann-Whitney U test /
Wilcoxon rank sum test
P=0.012 (Significant)

The Scandal of Poor Medical Research

Douglas G. Altman. British Medical Journal, 1994.

What should we think about a doctor who uses the wrong treatment, either willfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behavior was unprofessional, arguably unethical, and certainly unacceptable.

What then would we think about researchers who use the wrong techniques (either willfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled. Yet numerous studies of the medical literature, in both general and specialist journals, have shown that all of the above phenomena are common. This is surely a scandal.

2 Factors to be considered for test selection

Understanding A Statistician's Mind

Flow-chart for popularly used statistical tests

Q1 Q2 Q3 Q4,5 Q6 Q7

Q1,Univariate /Mutivariable	Q2, Difference /Correlation	Q3, Paired / related	Q4, Q5 Type of outcome (Normality)		Q6, No. of groups	Q7,sample size	Valid Tests
Univariate	Difference	Indepdendent (un-paired)	Continuous (Normal)		2		Student's t-test
					>2		One-way ANOVA
			Continuous (Non-normal)/ Ordered categorical		2		Mann-Whitney U test
					>2		Kruskal-Wallis H test
			Nominal		2	<20	Fisher's exact test
					≧2	≧20	Chi-square test
			Time to Event				Log-Rank test (Kaplan-Meier plot)
		D	Continuous		2		Paied-t test
							Repeated measured ANOVA
							ixed effect Regression
			Continuous				ilcoxon signed-rank test
			Ordered cat				riedman test
			Nominal		2		McNemar's test
	Correlation		Continuous (Normal)				Pearson's correlation (r)
			Continuous (Non-normal)/ordered				Spearman's correlation (rs)
			Nominal (2 levels)		2		Spearman/Kappa (Agrreement)
Multivariable		Indepdendent (un-paired)	Continuous (Normal residulas)				Linear Regression
			Continuous (Non-normal residulas)				Linear Regression*
			Ordered categorical				Ordered Logistic Regression
			Nominal	(2 levels)			Binary Logistic Regression
				(>2)			Multinomial Logistic Regression
			Time to Event				Cox Proportional Hazard Regression
		Dependent (paired)	Continuous (Normal residulas)				Linear Mixed Effect Regression
			Continuous (Non-normal residulas)				Linear Mixed Effect Regression*
			Ordered categorical				Generalized Estimation Equation (GEE)
			Nominal (2 levels)				Generalized Estimation Equation (GEE)

*Transform outcome variables for normalizing residuals

Created based on Publishing Your Medical Research Paper, by Daniel Byrne, Williams and Wi

Flow-chart for popularly used statistical tests

Q1,Univariate /Mutivariable	Q2, Difference /Correlation	Q3, Paired / related	Q4, Q5 Type of outcome (Normality)		Q6, No. of groups	Q7,sample size	Valid Tests
Univariate	Difference	Indepdendent (un-paired)	Continuous (Normal)		2		Student's t-test
					>2		One-way ANOVA
			Continuous (Non-normal)/ Ordered categorical		2		Mann-Whitney U test
					>2		Kruskal-Wallis H test
			Nominal		2	<20	Fisher's exact test
					≧2	≧20	Chi-square test
			Time to Event				Log-Rank test (Kaplan-Meier plot)
		Dependent (paired)	Continuous (Normal)		2		Paied-t test
					>2		Repeated measured ANOVA Mixed effect Regression
			Continuous (Non-normal)/ Ordered categorical		2		Wilcoxon signed-rank test
					>2		Friedman test
			Nominal		2		McNemar's test
	Correlation		Continuous (Normal)				Pearson's correlation (r)
			Continuous (Non-normal)/ordered				Spearman's correlation (rs)
			Nominal (2 levels)		2		Spearman/Kappa (Agrreement)
Multivariable		Indepdendent (un-paired)	Continuous (Normal residulas)				Linear Regression
			Continuous (Non-normal residulas)				Linear Regression*
			Ordered categorical				Ordered Logistic Regression
			Nominal	(2 levels)			Binary Logistic Regression
				(>2)			Multinomial Logistic Regression
			Time to Event				Cox Proportional Hazard Regression
		Dependent (paired)	Continuous (Normal residulas)				Linear Mixed Effect Regression
			Continuous (Non-normal residulas)				Linear Mixed Effect Regression*
			Ordered categorical				Generalized Estimation Equation (GEE)
			Nominal (2 levels)				Generalized Estimation Equation (GEE)

*Transform outcome variables for normalizing residuals

Created based on Publishing Your Medical Research Paper, by Daniel Byrne, Williams and Wilkins (1998)

Question 1 – Univariate?

Which type of test do you need:
Univariate or Multivariable?

Univariate - Unadjusted Analysis

Multivariable - Adjusted Analysis

-Are there confounders?
- Need for adjustment?

Question 1 – Univariate? (cont'd)

Randomization prevents confounding. Thus, in general, confounders are more problematic in observations studies than RCT' s.

If you want to adjust for confounders, then you need to perform **regression analysis**.

RCT -> Probably OK with univariate analysis
Observational studies -> Need to use Regression

ORIGINAL CONTRIBUTION

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease

A Propensity Analysis

Patricia A. Gum, MD
Maran Thamilarasan, MD
Junko Watanabe, MD
Eugene H. Blackstone, MD
Michael S. Lauer, MD

Context Although aspirin has been shown to reduce cardiovascular morbidity and short-term mortality following acute myocardial infarction, the association between its use and long-term all-cause mortality has not been well defined.

Objectives To determine whether aspirin is associated with a mortality benefit in stable patients with known or suspected coronary disease and to identify patient characteristics that predict the maximum absolute mortality benefit from aspirin.

Population?
Patients with an echo for possible coronary disease

Exposure?
Use of Aspirin at the baseline visit

Control?
No use of Aspirin at the baseline visit

Outcome?
Long term mortality
(median FU of 3.1 years)

Results

Table 2. Cox Proportional Hazards Analyses of Time to Death Among Patients Using Aspirin (N = 6174)*

Model	Hazard Ratio (95% CI)	*P* Value
Unadjusted	1.08 (0.85-1.39)	.50
Adjusted for age and sex	0.75 (0.58-0.96)	.02
Adjusted for age, sex, and history of CAD	0.57 (0.44-0.74)	<.001
Multivariable adjusted†	0.67 (0.51-0.87)	.002
Adjusted for age and sex among prespecified strata		
Normal LV function	0.75 (0.56-1.01)	.06
Abnormal LV function	0.54 (0.34-0.84)	.006
No history of prior CABG surgery	0.74 (0.54-1.08)	.06
History of prior CABG surgery	0.56 (0.35-0.89)	.01

*CI indicates confidence interval; CAD, coronary artery disease; LV, left ventricular; and CABG, coronary artery bypass graft.
†Adjusted for age, sex, body mass index, resting heart rate, resting systolic blood pressure, use of antihypertensive medications, digoxin, β-blockers, lipid-lowering therapy, nitrates, angiotensin-converting enzyme inhibitors, dihydropyridine and nondihydropyridine calcium channel blockers, congestive heart failure, smoking, atrial fibrillation, left and right bundle-branch block, pathologic Q waves, prior CABG surgery, prior percutaneous coronary intervention, chronic lung disease, peripheral vascular disease, exercise capacity, chronotropic response, heart rate recovery, left ventricular ejection fraction, echocardiographic evidence of myocardial ischemia, and failure of the left ventricle to decrease in size with exercise.

- People who use aspirin had reasons to use aspirin, they were sicker and had poorer prognosis. Therefore this may mask the effect of aspirin as its effect is mixed with poorer patients prognosis. This is called "Confounding".

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	*P* Value
Demographics			
Age, mean (SD), y	62 (11)	56 (12)	<.001
Men, No. (%)	1779 (77)	2167 (56)	<.001
Clinical history			
Diabetes, No. (%)	388 (17)	432 (11)	<.001
Hypertension, No. (%)	1224 (53)	1569 (41)	<.001
Tobacco use, No. (%)	234 (10)	500 (13)	.001
Prior coronary artery disease, No. (%)	1609 (70)	778 (20)	<.001
Prior coronary artery bypass graft, No. (%)	689 (30)	240 (6)	<.001
Prior percutaneous coronary intervention, No. (%)	667 (29)	148 (4)	<.001
Prior Q-wave MI, No. (%)	369 (16)	285 (7)	<.001
Atrial fibrillation, No. (%)	27 (1)	55 (1)	.04
Congestive heart failure, No. (%)	127 (6)	178 (5)	.12
Medication use			
Digoxin use, No. (%)	171 (7)	216 (6)	.004
β-Blocker use, No. (%)	811 (35)	550 (14)	<.001
Diltiazem/verapamil use, No. (%)	452 (20)	405 (10)	<.001
Nifedipine use, No. (%)	261 (11)	283 (7)	<.001
Lipid-lowering therapy, No. (%)	775 (34)	380 (10)	<.001
ACE inhibitor use, No. (%)	349 (15)	441 (11)	<.001
Cardiovascular assessment and exercise capacity			
Body mass index, mean (SD), kg/m^2	29 (5)	30 (7)	<.001
Ejection fraction, mean (SD), %	50 (9)	53 (7)	<.001
Resting heart rate, mean (SD), beats/min	74 (13)	79 (14)	<.001
Resting blood pressure, mean (SD), mm Hg			
Systolic	141 (21)	138 (20)	<.001
Diastolic	85 (11)	86 (11)	.04
Purpose of test to evaluate chest pain, No. (%)	300 (13)	468 (12)	.31
Mayo Risk Index ≥1, No. (%)†	2021 (87)	2517 (65)	<.001
Peak exercise capacity, mean (SD), METs			
Men	8.6 (2.4)	9.1 (2.6)	<.001
Women	6.6 (2.0)	7.3 (2.1)	<.001
Heart rate recovery, mean (SD), beats/min	28 (11)	30 (12)	<.001

Question 1 – Univariate? (cont'd)

Randomization prevents confounding. Thus, in general, confounders are more problematic in observations studies than RCT's.

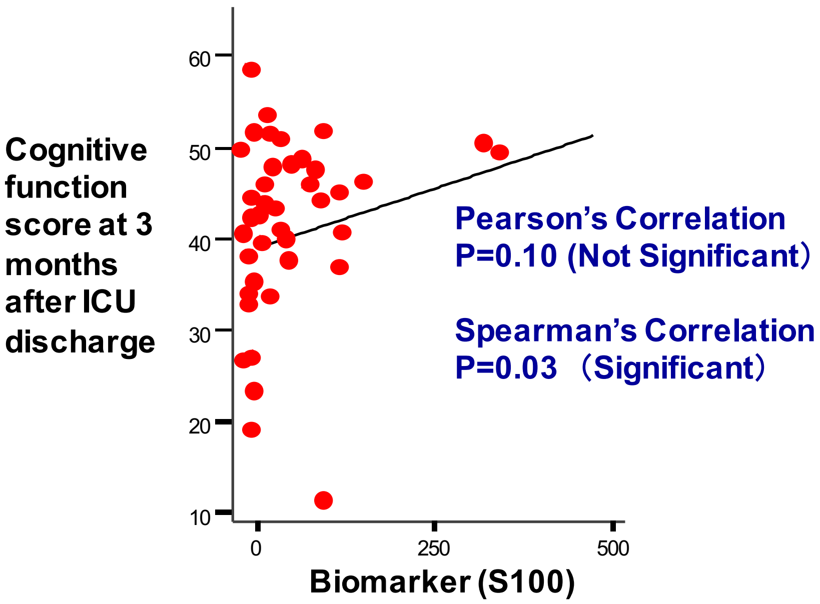
If you want to adjust for confounders, then you need to perform **regression analysis**.

Question 2 -Difference?

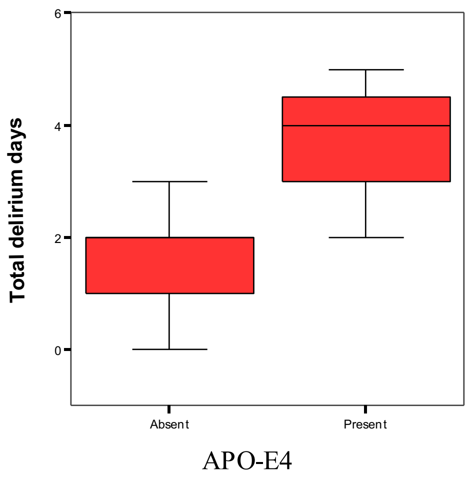
- Do you want to test for a difference between groups or want to test for correlation between variables?

–Comparing mean (or median) of two groups?
–Correlation between two variables in one group?

Example 1 : Different tests → Different results



Comparing Difference



Examples:

Student's T-test
Comparing 2 means

Mann-Whitney U test /
Wilcoxon rank sum test
Comparing 2 medians

Question 3 - Paired?

- Were the groups paired or unpaired / (dependent or independent)?

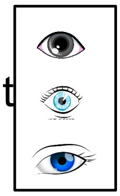
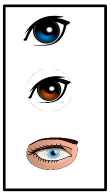
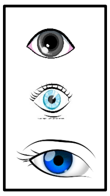
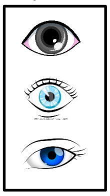
Are you measuring more than once from one sample?

Examples:

Student t-test comparing 2 independent means.
(Comparing outcome between intervention and control groups)

Paired t-test comparing 2 related means.
(Comparing outcome before and after an intervention).

Paired or Independent groups?

	New eye drop		Placebo eye drop
Independent		VS	
	New eye drop Right eye		Placebo eye drop Left eye
Paired		VS	

Question 4 - Outcome Type?

- What is the level of measurement for outcome variable?

-Continuous (Interval)? Ex. Blood pressure, BMI, Weight

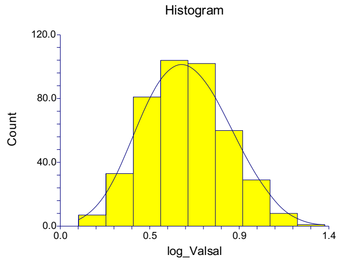
- Discrete/Categorical/Factor?

-Nominal? 2 levels (Binary, dichotomous) ex. Died / Survived
>2 levels. Ex. Disease Type (cancer, DM, cardiovascular)

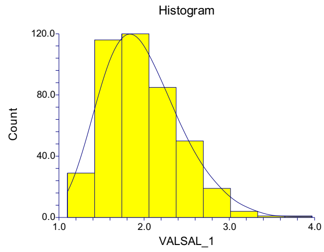
-Ordinal? > 2 levels. Ex. Disease severity (1: Mild, 2: Moderate, 3: Severe)
Disease score (0: normal, 10: abnormal)

Question 5 – Normality?

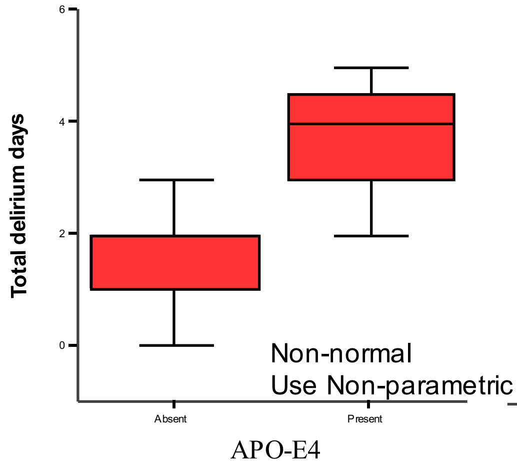
- If an outcome variable is continuous, is it normally distributed? If your histogram forms a bell-shaped curve, assume that it is normal; otherwise, assume that it is non-normal.

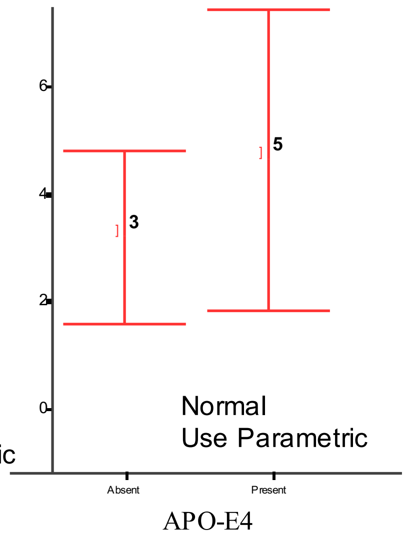


Normal



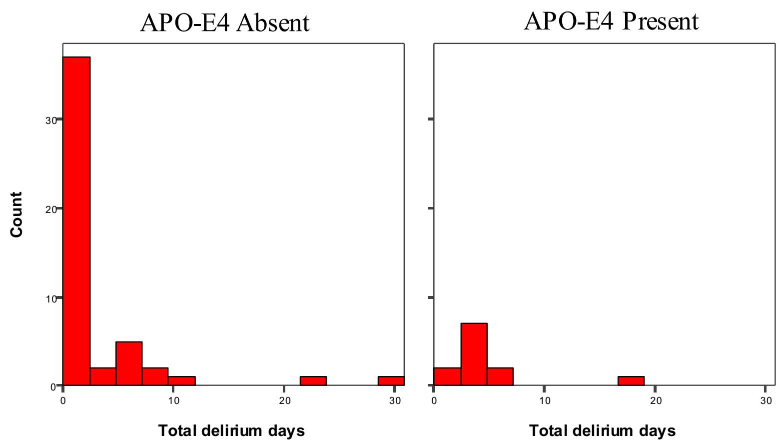
Non-normal / skewed





Mann-Whitney U test / Wilcoxon rank sum test P=0.012 (Significant)

Student's T-test P=0.405 (NS)



Usually, any variable involving "days" are highly skewed (cannot take negative value, long tail value)

Parametric Tests are valid only when…

Student t-test, ANOVA are valid only when outcome variable is normally distributed within a group.

Paired t-test (for example comparing BP before after an intervention) is valid only when within-patient difference in outcome variable (e.g.BP) is normally distributed.

Linear regression is valid only when residuals (difference between observed and predicted values) are normally distributed.

Pearson-correlation analysis is valid only when both (outcome and exposure) variables are normally distributed.

Non-Parametric Tests are always valid regardless of distribution of data

Statistical Methods Recommendation by New England Journal of Medicine

The basis for these guidelines is described in Bailar JC III, Mosteller F. Guidelines for statistical reporting in articles for medical journals: amplifications and explanations. Ann Intern Med 1988;108:266-73. **Exact methods should be used as extensively as possible in the analysis of categorical data. For analysis of measurements, nonparametric methods should be used to compare groups when the distribution of the dependent variable is not normal.**

This page can be found at
http://authors.nejm.org/Misc/NewMs.asp#statistics

Question 6 - #groups?

- How many groups are there for the independent (predictor) variable?

- 2 levels ?
- 3 or More?

Examples:

Student t-test comparing 2 group means
ANOVA comparing 3 or more group means

Question 7 - Sample Size?

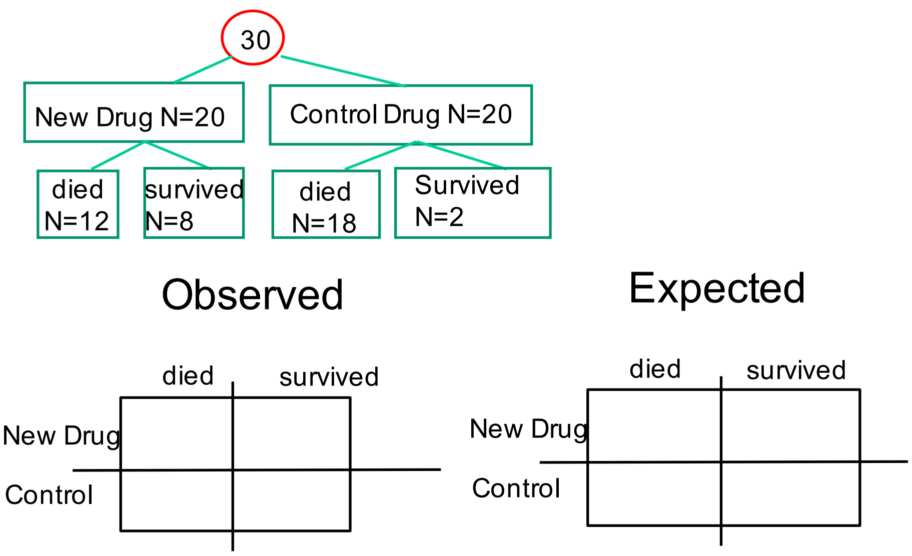
- What is the total sample size?

Examples:

Greater than total N=20, use Chi-square test

Greater than 20 and less than 40 and
an expected # in a cell < 5,
use Fisher's exact test

When to use Fisher's Exact test



Selection of Regression

Only depends on the following 2 things:
- type of outcome variable
- Whether data are paired or not (Repeated or not).

	Not repeated	Repeated
Continuous	Linear	Mixed effect
Binary	Logistic	GEE
Time to Events	Cox	MULCOX

Flow-chart for popularly used statistical tests

Q1, Univariate /Mutivariable	Q2, Difference /Correlation	Q3, Paired / related	Q4, Q5 Type of outcome (Normality)	Q6, No. of groups	Q7, sample size	Valid Tests
Univariate	Difference	Indepdendent (un-paired)	Continuous (Normal)	2		Student's t-test
				>2		One-way ANOVA
			Continuous (Non-normal)/ Ordered categorical	2		Mann-Whitney U test
				>2		Kruskal-Wallis H test
			Nominal	2	<20	Fisher's exact test
				≧2	≧20	Chi-square test
			Time to Event			Log-Rank test (Kaplan-Meier plot)
		Dependent (paired)	Continuous (Normal)	2		Paied-t test
				>2		Repeated measured ANOVA Mixed effect Regression
			Continuous (Non-normal)/ Ordered categorical	2		Wilcoxon signed-rank test
				>2		Friedman test
			Nominal	2		McNemar's test
	Correlation		Continuous (Normal)			Pearson's correlation (r)
			Continuous (Non-normal)/ordered			Spearman's correlation (rs)
			Nominal (2 levels)	2		Spearman/Kappa (Agrreement)
Multivariable		Indepdendent (un-paired)	Continuous (Normal residulas)			Linear Regression
			Continuous (Non-normal residulas)			Linear Regression*
			Ordered categorical			Ordered Logistic Regression
			Nominal (2 levels)			Binary Logistic Regression
			Nominal (>2)			Multinomial Logistic Regression
			Time to Event			Cox Proportional Hazard Regression
		Dependent (paired)	Continuous (Normal residulas)			Linear Mixed Effect Regression
			Continuous (Non-normal residulas)			Linear Mixed Effect Regression*
			Ordered categorical			Generalized Estimation Equation (GEE)
			Nominal (2 levels)			Generalized Estimation Equation (GEE)

*Transform outcome variables for normalizing residuals

Created based on Publishing Your Medical Research Paper, by Daniel Byrne, Williams and Wi

1.3 Tutorials for selecting valid statistical tests

Example 1

- Comparing ventilator free days between patients who were randomized to daily awakening and breathing trial vs daily breathing trial among ventilated patients in medical ICU: A prospective randomized study.

Q1: **(Univariate?)** ~~Univariate~~ → Mutivariable → Linear regression

Q2: **(Difference?)** Difference

Q3: **(Paired?)** Unpaired

Q4: **(Type?)** Continuous

Q5: **(Normality?)** ~~Normal~~ → ~~Non-Normal~~

Q6: **(#groups?)** 2

Q7: **(sample size?)** > 30 in each group

~~Student's T-test~~ Mann-Whitney U Test

Example 2

- Cytokine responses of peripheral blood mononuclear cells (PBMC) from HIVseronegative adults with prior extra pulmonary TB were compared with responses from persons with prior pulmonary tuberculosis and latent *M. tuberculosis* infection in a case-control study. Antas, *Journal of Allergy and Clinical Immunology.* 2006.

Q1: **(Univariate?)** ~~Univariate~~ ⟶ Multivariable → Linear regression
Q2: **(Difference?)** Difference
Q3: **(Paired?)** Unpaired
Q4: **(Type?)** Continuous
Q5: **(Normality?)** ~~Normal~~ ⟶ ~~Non-Normal~~
Q6: **(#groups?)** 3
Q7: **(sample size?)** > 15 in each group

~~1-way ANOVA~~ ~~Kruskal-Wallis H Test~~

Example 3

- We want to estimate the relationship between two numerical measures: Bio-marker value for S100 and patient's cognitive scores measured at 3 months after ICU discharge among patients in medical ICU.

Q1: **(Univariate?)** ~~Univariate~~ ⟶ Multivariable → Linear regression
Q2: **(Difference?)** Correlation
Q3: **(Paired?)** NA
Q4: **(Type?)** Continuous
Q5: **(Normality?)** ~~Normal~~ ⟶ ~~Non-Normal~~
Q6: **(#groups?)** 1 group
Q7: **(sample size?)** > 30 in each group

~~Pearson's r Correlation coeffient~~ ~~Spearman's ρ Rank Correlation coefficient~~

Example 4

- Martinez-Picado et. al. compared proportion of patients with HIV infection who had viral surge between alternation of antiretroviral drug regimens and standard regimens. A Randomized, Controlled Trial. *Annals of Internal Medicine.* 2003

Q1: (Univariate?) ~~Univariate~~ ⟶ Multivariable → Logistic regression

Q2: (Difference?) Difference

Q3: (Paired?) Unpaired

Q4: (Type?) Nominal

Q5: (Normality?) NA

Q6: (#groups?) 2

Q7: (sample size?) ~~> 20~~ → ~~Chi-square test~~ ; < 20 → Fisher's Exact test

Example 5

- A researcher wants to evaluate the effect of a new diet on weight loss by comparing patient's weight before and after the diet program.

Q1: (Univariate?) Univariate

Q2: (Difference?) Difference

Q3: (Paired?) Paired

Q4: (Type?) Continuous

Q5: (Normality?) Normal ⟶ Non-Normal

Q6: (#groups?) 2

Q7: (sample size?) > 30 in each group