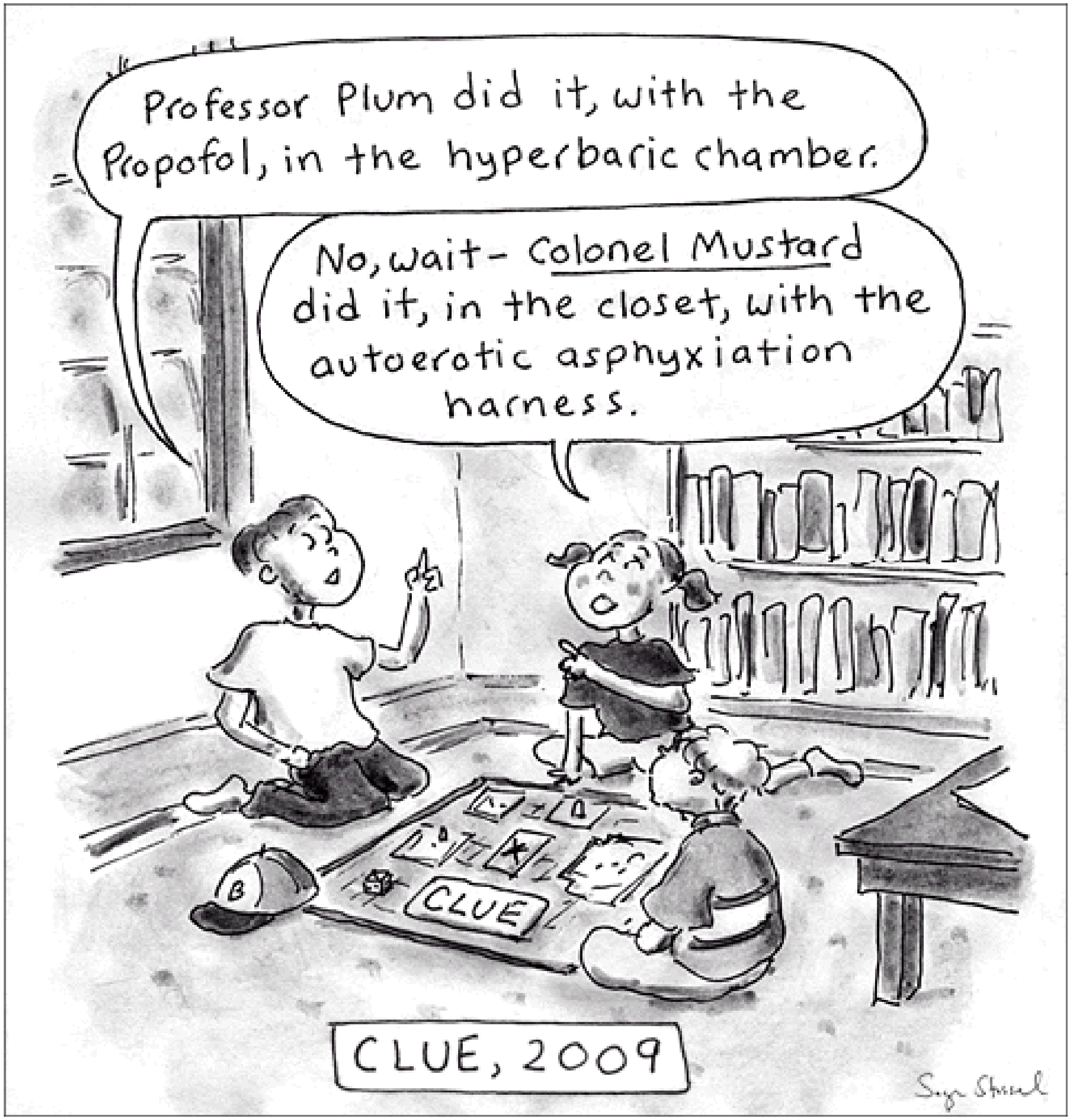
My Idea is Great, But How Do I Study It?

Plenary Session
UHMS Annual Scientific Meeting
04 June 2010

Nicole C. Close, PhD
Principal Biostatistician

Empiristat
BIOSTATISTICAL SYNERGY


Professor Plum did it, with the Propofol, in the hyperbaric chamber.
No, wait- Colonel Mustard did it, in the closet, with the autoerotic asphyxiation harness.
CLUE
CLUE, 2009

Goals of this Session

- To be introduced to various study designs
- To discuss best practices for stating your hypothesis, objectives and endpoints for your study
- To learn about sample size, power and implications to the study
- Discuss randomization methodology and techniques
- Define and acknowledge statistical terms and statistical testing
- Outline best practices for reporting research

Basic Study Designs

Empiristat
BIOSTATISTICAL SYNERGY

Scientific Method Applied

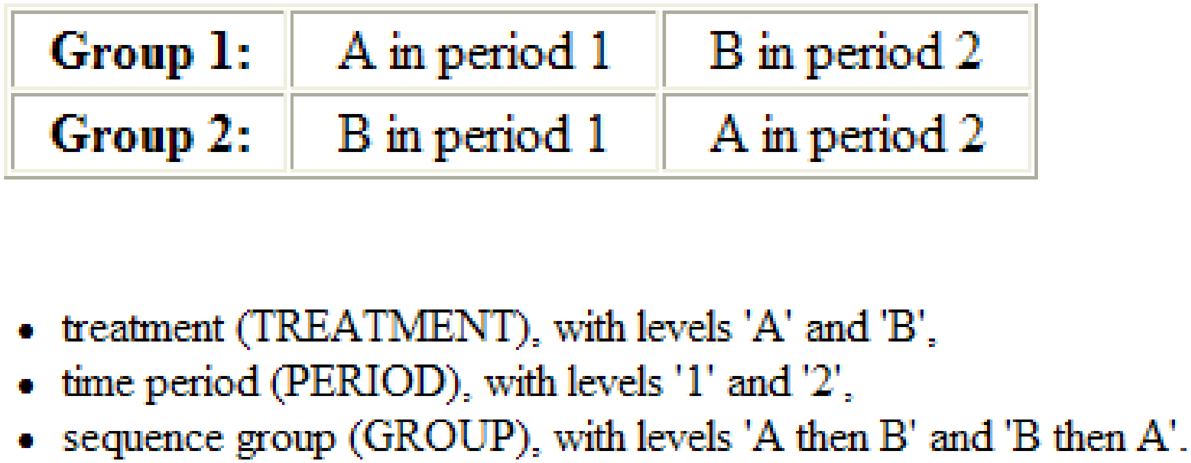
- Define the Purpose of the Study(state specific Hypothesis)
- Design the study(protocol)
- Conduct the study (good organization)
- Analyze the data (tests of hypotheses, descriptive)
- Draw Meaningful but not Overstated Conclusions (publish results)

Basic Study Designs

- Cohort
  - Prospective
  - Retrospective
  - Time Series
- Case-Control
  - Nested Case Control
- Cross-Sectional
- Ecological

- Superiority
- Non-inferiority
- Equivalence
- Cross-over
- Bioequivalence
- Pharmacokinetic
- Pharmacodynamic

Cross-Over Study



Group 1:	A in period 1	B in period 2
Group 2:	B in period 1	A in period 2

- treatment (TREATMENT), with levels 'A' and 'B',
- time period (PERIOD), with levels '1' and '2',
- sequence group (GROUP), with levels 'A then B' and 'B then A'.

Most Common Experimental Design

- Superiority: one is better than the other
  - the mean level in Group A is different than the mean level in Group B (hyp is better).

- Non-inferiority: one is no worse than the other by a certain amount
  - The mean level in Group A is no 'worse' than the mean level in Group B (within a margin of x being worse).

- Equivalence: they are approximately the same by a certain amount (in either direction, better or worse)
  - The mean level in Group A is the same as Group B (within a margin of x being better or being worse)

Hypotheses, Objectives and Endpoints

What is your question?

- Each study must have a primary question. The primary question, as well as secondary questions, should be carefully selected, clearly defined, and stated in advance.

Elements of a Good Hypothesis

- Written as a definite statement not a question
- Based on observations and knowledge
- Be testable
- Predict the anticipated results in clear terminology
- Based on an independent variable (experimental) and a dependent variable (responding, measured/observed)

Hypothesis

A hypothesis can be shown to be supported by the evidence but it can never be proved.

- One of Two Choices:
  - We reject the null hypothesis.
  - We fail to reject the null hypothesis.

Null Hypothesis

- Presumed true until statistical evidence indicates differently
- No difference exists between the two groups for the variable you are comparing

$H_0 : \mu_1 = \mu_2$
where:

H_0 = the null hypothesis
μ_1 = the mean of population 1, and
μ_2 = the mean of population 2.

What is hypothesis testing?

- Probability of observing the obtained data or data more different from the prediction of the null hypothesis, if the null hypothesis is true (significance level, p value)
- How often we would expect to observe our experimental results, or results even more extreme, if we were to take many samples from a population where there was no effect (i.e. we test against our null hypothesis)
- If we find that this happens rarely (5% of the time), we conclude that our results support our experimental prediction — we reject our null hypothesis.

Hypothesis Testing Process

- #1: State the hypothesis to be tested.
- #2: Choose the level of significance at which the test will be performed. This is called the size or level of the test. It is the probability of rejecting the null hypothesis when it is true.
- #3: Collect the data and reject the hypothesis or not depending on the observed value of the test statistic.

The questions (objectives)

- Primary:
  - Most interested in answering
  - One that is capable of being answered
  - Basis for your sample size
  - Emphasized in reporting of study results

- Secondary:
  - Related to your primary question
  - Related to subgroup hypotheses

- These define your primary and secondary objectives of the study.

Objectives

- A precise statement of the degree of benefit expected from the intervention, as well as the duration of the benefit
- Clear statements of the time frame of the study (especially in relation to how quickly benefits might occur)
- A definition of the participants for whom the benefit is sought.

Endpoints

- A measurement taken under specific conditions that reflects the response of an individual to treatment.
  - Defined and written in advance
  - Capable of being assessed in all participants
  - Participation generally ends when the endpoint occurs (unless there is a combination of primary endpoints); interested in events subsequent to primary endpoint.
  - Capable of unbiased assessment
  - Ascertained as completely as possible

Basic Types of Endpoints

- Clinical
  - Death, myocardial infarction, increase in lab value
- Surrogate
  - Measure of effect that may correlate with the clinical endpoint
  - CD4 count, viral load
- Composite
  - Stroke, MI and Revascularization

Summary

- Formulate your hypotheses to test
- Clearly define your objectives and document
- Define and describe your endpoints

Randomization

Randomization

- "Allocation concealment: A technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment. Allocation concealment prevents researchers from (unconsciously or otherwise) influencing which participants are assigned to a given intervention group."

  --CONSORT statement

The Randomization Process

- Randomization is a process where each subject/unit has the same chance of being assigned to either the intervention or the control.

- Neither subject/unit nor investigator should know what the assignment will be before the subject decides to enter the study.

BUT THEY WERE SELECTED RANDOMLY

Randomization Methods

- Common methods with advantages and disadvantages

- Can be assumed that the strategy will allocate participants to two groups, but most methods can be generalized to more than two groups.

Bias

- Allocation is predictable (selection bias)
- Unbalanced groups with respect to risk factors or prognostic covariates
- For large studies the chance of unbalanced groups are negligible.
- Report of the trial should contain brief but clear description of the method employed.

Fixed Allocation Strategies

- Simple randomization
- Blocked randomization
- Stratified randomization
- Also adaptive strategies
  - Urn (probability changes based on last assignment)
  - Play the winner (probability changes based on outcome)

Simple Randomization

- Toss an unbiased coin each time a participant is eligible for randomization
  - Random digit table (small studies)
  - Random number-producing algorithm (large studies)
- Alternating assignments is not simple randomization

Advantages/Disadvantages of Simple Randomization

- Easy to implement
- At the end each group will be in correct proportion, but at any point in the randomization, there could be a substantial imbalance.
  - Not invalid, but awkward
  - Not favored in the field

Blocked Randomization

- Participants are randomly assigned with equal probability to Group A or Group B for each block of even size (e.g. 4,6,8).
- Order in which the intervention is assigned in each block is random.

Blocked Randomization Example

Block Size	Assignment
6	ABABBA
8	BAABBABA
8	BAABABAB
6	AABABB

Advantages/Disadvantages of Blocked Randomization

- Balance between number of participants in each group is guaranteed during the course of randomization.
- If the trial is terminated before enrollment is completed, balance exists between the groups.
- If blocking factor is known and the trial is not double-blinded, assignment for the last person is known before they are randomized
  - Don't use block sizes of 2
  - Vary the block sizes, instead of using only one block size
  - Never indicate in the protocol the block sizes

Stratified Randomization

- Measure prognostic factors either before or at the time of randomization.
- If a single factor is used, it is divided into two or more strata.
- If several factors are used, a strata is formed by selecting by one subgroup from each of them.
- The randomization process involves measuring the level of the selected factors for participants, determining which stratum each belongs and performing the randomization within that stratum.
- Within each stratum it can be simple or blocked.

Stratified Randomization Example

AGE	SEX	SMOKING HX
40-49 yr	Male	Current
50-59 yr	Female	Former
60-69 yr		Never

Stratified Randomization Example

- 3x2x3 = 18 strata
- Strata 1: 40-49, Male, Current
- Strata 2: 40-49, Male, Former
- Strata 3: 40-49, Male, Never
- Strata 4: 40-49, Female, Current
- Strata 5: 40-49, Female, Former
- Strata 6: 40-49, Female, Never
- and so on ....

Randomization Consideration

- Careful attention to the method used.
- Need valid randomization
  - single center studies have an independent statistician not involved in the care or assessment of the participant
  - Multicenter studies have a coordinating center conduct randomization
- Assignments distributed:
  - Envelopes, telephone system, web based
- Assigned closest to time of beginning intervention

Factors Influencing Mechanism of Randomization

- Phase, size, scope and purpose of trial
- Who is doing the randomization and where is it taking place – study coordinator, research pharmacist; field site, clinic
- Blinded or unblinded trial
- Type of intervention (diet, capsule, tablet, IV, surgical procedure, chamber intervention)
- Point at which randomization occurs
- Cost and complexity

Once Randomized—Always Analyze

- Randomize at the closest time to intervention
  - Alprenolol Trial, Ahlmark et al (1976)
    - 393 subjects randomized two weeks before therapy
    - Only 162 subjects received intervention, 69 alprenolol & 93 placebo

Masking

Bias

- In any clinical study, bias is one of the biggest concerns.
- Systematic error: difference between the true value and that actually obtained due to all causes other than sampling variability.
- Conscious or subconscious factors or both.
- To minimize one type of bias, we can mask assigned intervention, assessment, classification and evaluation of the response variables.

Types of Masking in Studies

- Unblinded studies
- Single blind studies
- Double blind studies

Unblinded Study

- Open trial
- Participant and investigator know the intervention the participant has been assigned.
  - Exercise studies, eating habits
- Main concern is bias, reporting of symptoms and side effects, data collection.

Single Blind Study

- Only the investigators are aware of the intervention.
- Knowledge of the intervention may help the investigator care for the patient.
- Participant bias is minimized, but potential of investigator bias in administration of non-study therapy, data collection and data assessment.

Double Blind Study

- Neither participant nor investigator know the treatment assignment.
- Usually trials of drug efficacy.
- Risk of bias is reduced, any effects of actions theoretically would occur equally in the intervention and control groups.
- Certain functions of the study must be conducted by independent groups (toxicity and benefits).

Double Blind Study

- More complex and several factors must be evaluated
- Matching medications
- Coding of medications
- Unblinding trials
- Assessment of blindness

Unblinding Studies

- Accidental unblinding
- Procedural errors
- Laboratory errors
- Monitoring use of study intervention prescribed outside of the study.
- Official breaking of the blind
    - Procedures specifically defined

Assessment of Blindness

- Sometimes worthwhile to estimate the degree to which the blind was maintained.
- Patient guess
- Investigator guess
- Degree difference from 50%
- Study adherence versus what they thought they received

Sample Size and Power

Sample Size

- Studies should have sufficient statistical power to detect differences of interest between the groups.
- Sample size and power should be considered early in the planning phase.
- Sample size calculations provide only an estimate of the needed size of a study.

THE WALL STREET JOURNAL.

"This tech support person says she hopes I'm of a caste she can talk to."

Empiristat
BIOSTATISTICAL SYNERGY

Sample Size Information

- Provides Information About:
  - How many subjects should participate in the research?
  - Is the study worth conducting if only n subjects participate?

Power

- The power of a study is its ability to detect a true difference in outcome between the control arm and the intervention arm.

- By definition, a study power set at 80% accepts a likelihood of one in five (that is, 20%) of missing such a real difference.

- Power for large trials is occasionally set at 90% to reduce to 10% the possibility of a so-called "false-negative" result.

Significance Level

- The chosen level of significance sets the likelihood of detecting a treatment effect when no effect exists (leading to a so-called "false-positive" result) and defines the threshold "P value".

- Results with a P value above the threshold lead to the conclusion that an observed difference may be due to chance alone, while those with a P value below the threshold lead to rejecting chance and concluding that the intervention has a real effect.

Significance Level

- The level of significance is most commonly set at 5% (that is, p = 0.05) or 1% (p = 0.01). This means the investigator is prepared to accept a 5% (or 1%) chance of erroneously reporting a significant effect when there really isn't one.

Sample Size

- Items to review:
  - What is the primary question you want to investigate? (Ex. Effectiveness of a treatment compared to a placebo – difference)
  - What is the study design?
  - How is the question to be answered?
  - How is the data collected and analyzed?
  - How big or small of a difference of interest?
  - How much variability to this data?

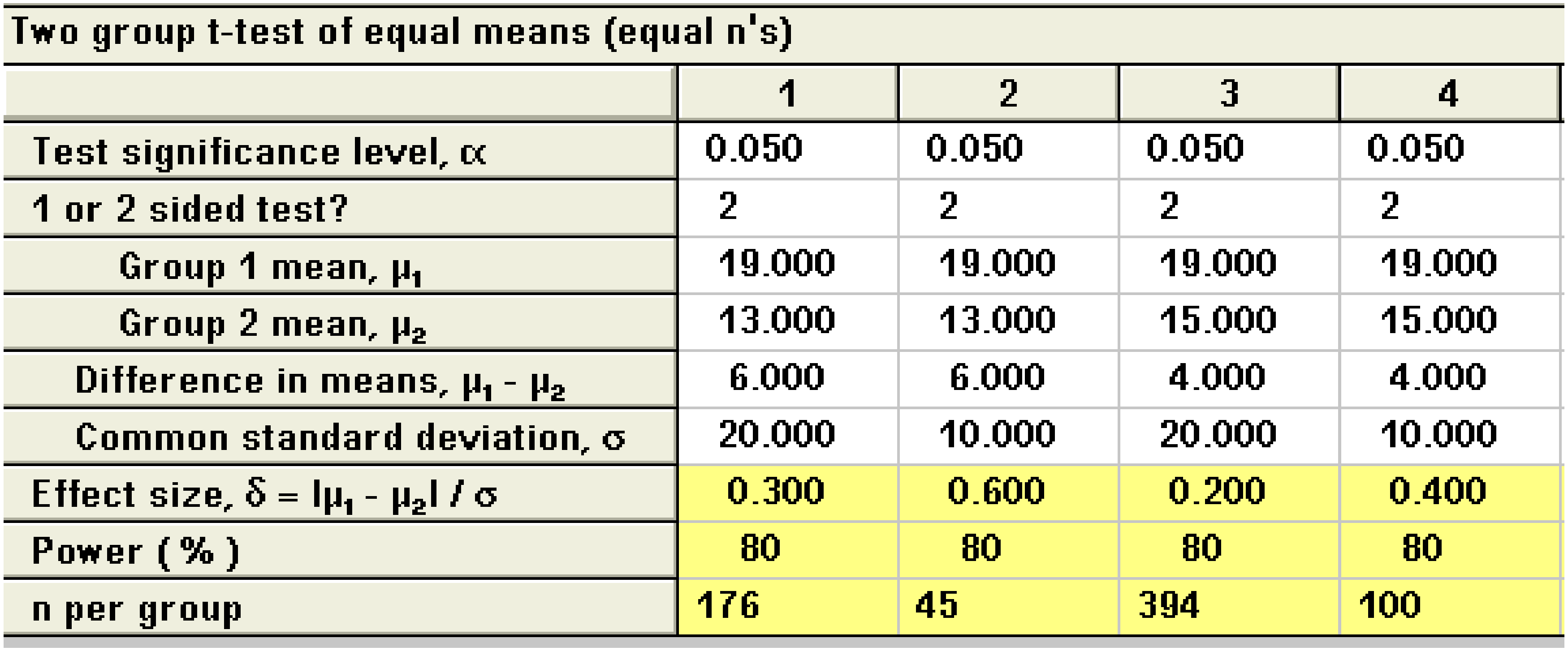
Potential Errors

Study Result	Actual Truth: Intervention Benefit	Actual Truth: No Intervention Benefit
Intervention Benefit	Correct Result	Type I error (alpha error)
No Intervention Benefit	Type II error (beta error)	Correct Result

Example 1

- Two groups, superiority design
  - The mean change from baseline in group A is different than the mean change from baseline in group B
  - Two group t-test (of equal group size)
- Clinically meaningful change in values from baseline is a reduction by 6 or more (ex SBP).
- Variability (SD=20 mmHg) previously seen in mean changes from baseline.
- P value of 0.05 and 80% power

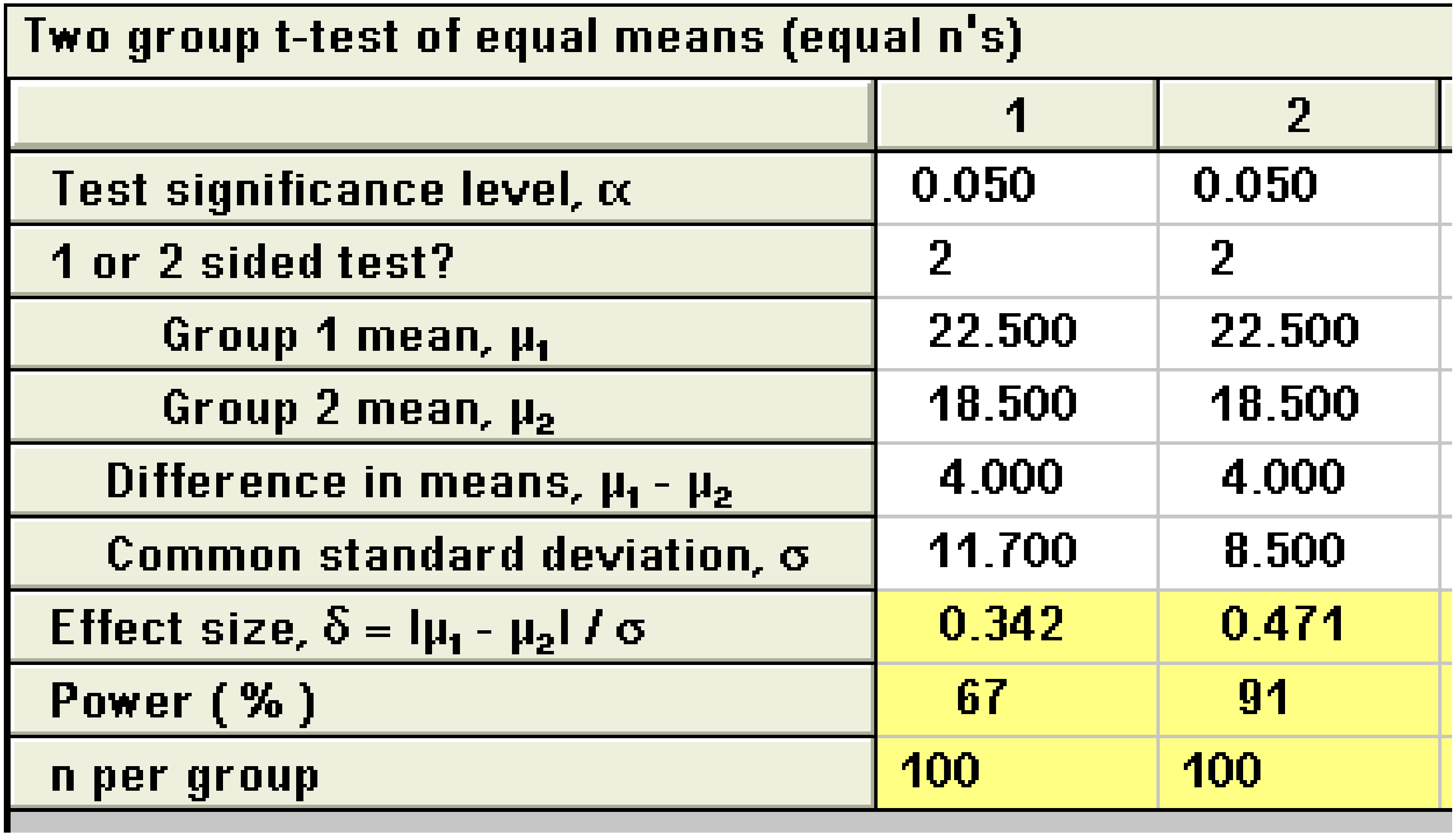
Sample Size Example 1



Two group t-test of equal means (equal n's)

	1	2	3	4
Test significance level, α	0.050	0.050	0.050	0.050
1 or 2 sided test?	2	2	2	2
Group 1 mean, μ_1	19.000	19.000	19.000	19.000
Group 2 mean, μ_2	13.000	13.000	15.000	15.000
Difference in means, $\mu_1 - \mu_2$	6.000	6.000	4.000	4.000
Common standard deviation, σ	20.000	10.000	20.000	10.000
Effect size, $\delta = \lvert\mu_1 - \mu_2\rvert / \sigma$	0.300	0.600	0.200	0.400
Power (%)	80	80	80	80
n per group	176	45	394	100

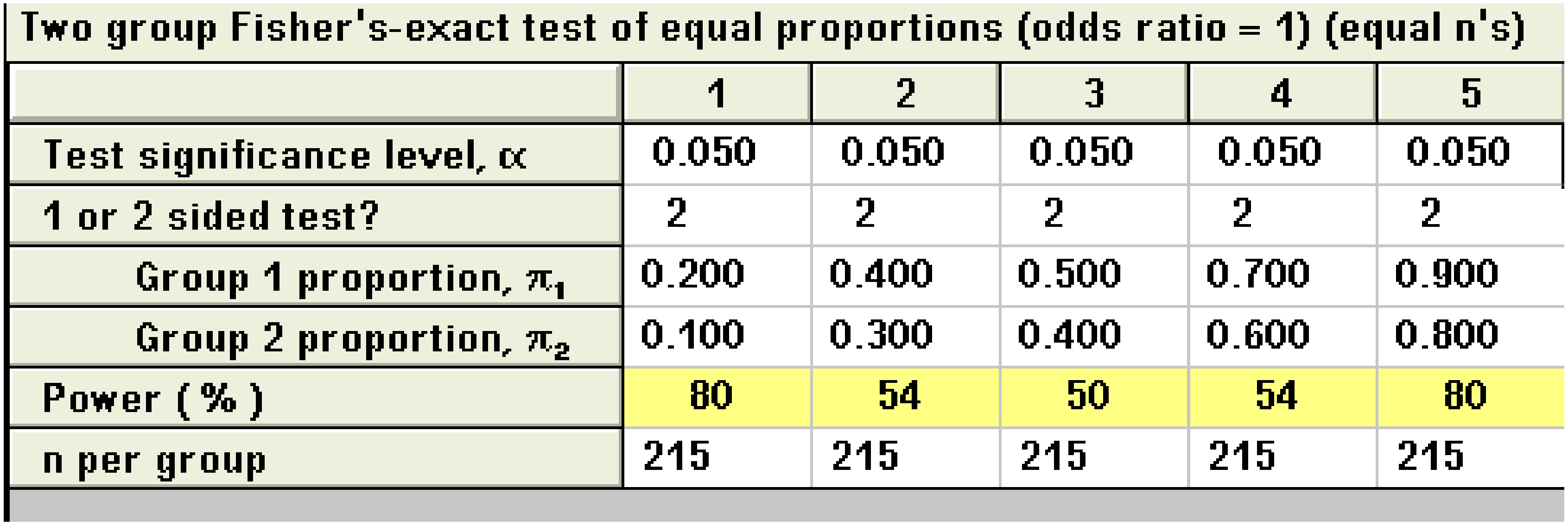
Power Post Study Example 1



Two group t-test of equal means (equal n's)

	1	2
Test significance level, α	0.050	0.050
1 or 2 sided test?	2	2
Group 1 mean, μ_1	22.500	22.500
Group 2 mean, μ_2	18.500	18.500
Difference in means, $\mu_1 - \mu_2$	4.000	4.000
Common standard deviation, σ	11.700	8.500
Effect size, $\delta = \|\mu_1 - \mu_2\| / \sigma$	0.342	0.471
Power (%)	67	91
n per group	100	100

Example 2



Two group Fisher's-exact test of equal proportions (odds ratio = 1) (equal n's)

	1	2	3	4	5
Test significance level, α	0.050	0.050	0.050	0.050	0.050
1 or 2 sided test?	2	2	2	2	2
Group 1 proportion, π_1	0.200	0.400	0.500	0.700	0.900
Group 2 proportion, π_2	0.100	0.300	0.400	0.600	0.800
Power (%)	80	54	50	54	80
n per group	215	215	215	215	215

Documentation

- The sample size calculation should be described in sufficient detail to allow its use in other protocols.
- The power, level of significance and the control and intervention event rates should be clearly documented.
- Information on the scheduled duration of the study, any adjustment for non-compliance and any other issues that formed the basis of the sample size calculation should be included.

Statistical Power

- Power is of importance when:
  - We want to have a rationale basis for establishing sample sizes for a study.
  - When we don’t want to be in a position at the end of the study, if we find no difference between our intervention and comparison group and not being able to tell whether there truly isn’t a difference or if we didn’t have a big enough sample size to detect the “true” effect.

When should SS Calculations be Performed?

- Definitely, before the study, sometimes during and sometimes after.
- In designing the study, we want to make sure:
  - What we do is worthwhile so that we get a correct answer and in the most efficient way.
  - So we can recruit enough participants to give our results adequate power but not too any that we waste time getting more data than we need.
  - When designing we may have to make assumptions about desired effect size and variance within the data.

When should SS Calculations be Performed? (cont)

- Interim power calculations are done when original ones become suspect.
- Avoid premature stopping of a study or to avoid the prolongation of a study. But should only be conducted when stated a priori in the research design.
- Assessing trials of negative results, to make sure the study was not underpowered.

Statistical Tests

- Sample size calculations indicate how the statistical tests used in the study are likely to perform.
- Type of test used affects how the sample size is calculated.
- Parametric tests versus non-parametric tests (need more participants).

Basic Statistics

© 2002 Randy Glasbergen
www.glasbergen.com

"Remember the old days when we used to eat his statistics homework?"

Descriptive Statistics

- Used to describe the main features of a set of data
  - Different from inferential statistics or hypothesis testing
  - Also called estimation
- Point estimation
- Categorical data – Frequency distribution (n, %), contingency table
- Continuous data – location, dispersion, shape

Frequency Distribution

- Continuous Variable (age) and categorized it into a ordinal variable
- Show n and percent for each category

Age Group (years)	Study Population (N=200)
<35	18 (9%)
35-45	42 (21%)
46-55	90 (45%)
56-65	38 (19%)
>65	12 (6%)

Continuous Data

- Location – measures of central tendency
  - Mean (average) – sum of all values divided by number of values
    - Most common descriptive statistic
    - Should not be used for data that does not follow a normal distribution
  - Median (50th percentile) – the value where 50% of the observations are below and 50% of the observations are above (middle value)
  - Mode – most frequently occurring value

Continuous Data

- Skewness – if mean =median, then skewness=0
  - Ex. 1, 2, 3, 4, 100 – positive skewness
  - Ex. 1, 1001, 1002, 1003, 1004 – negative skewness
- Kurtosis – measure of "peakedness" of distribution
- Not presented often in manuscripts, reports but used by statisticians to understand the data

Probability Distributions

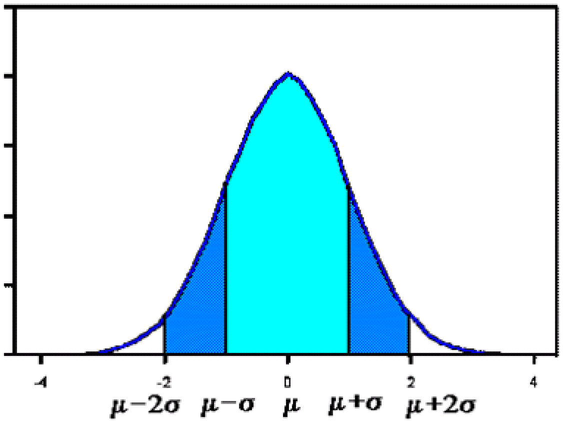
- Inferential statistical analysis is grounded in identifying the type of data being analyzed and the distribution the data follows

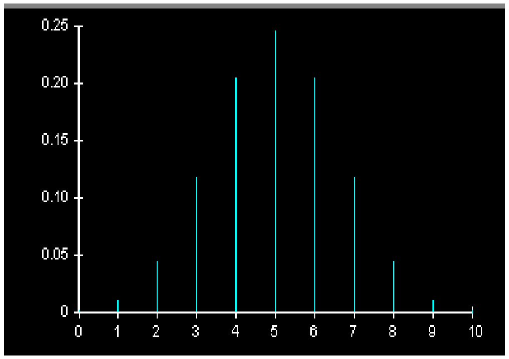
- Data that follows a distribution can be analyzed used parametric methods (normal, etc.)

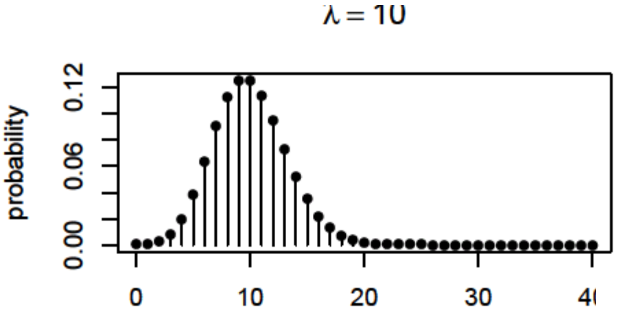
- Data that does not follow a distribution is analyzed using nonparametric methods

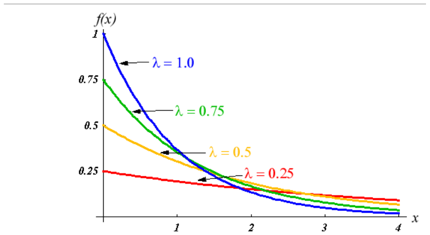
Types of Probability Distributions

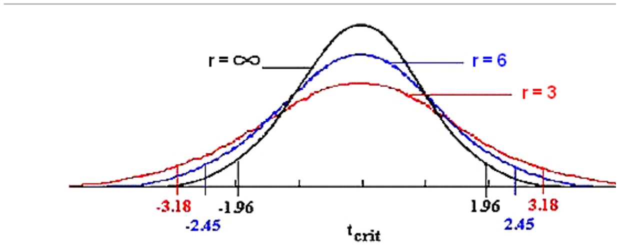
- Normal
- Binomial
- Poisson
- Student’s t
- Exponential
- Chi-Square

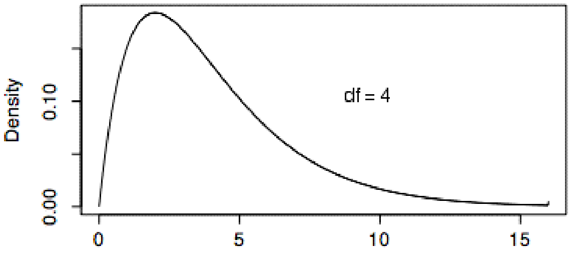


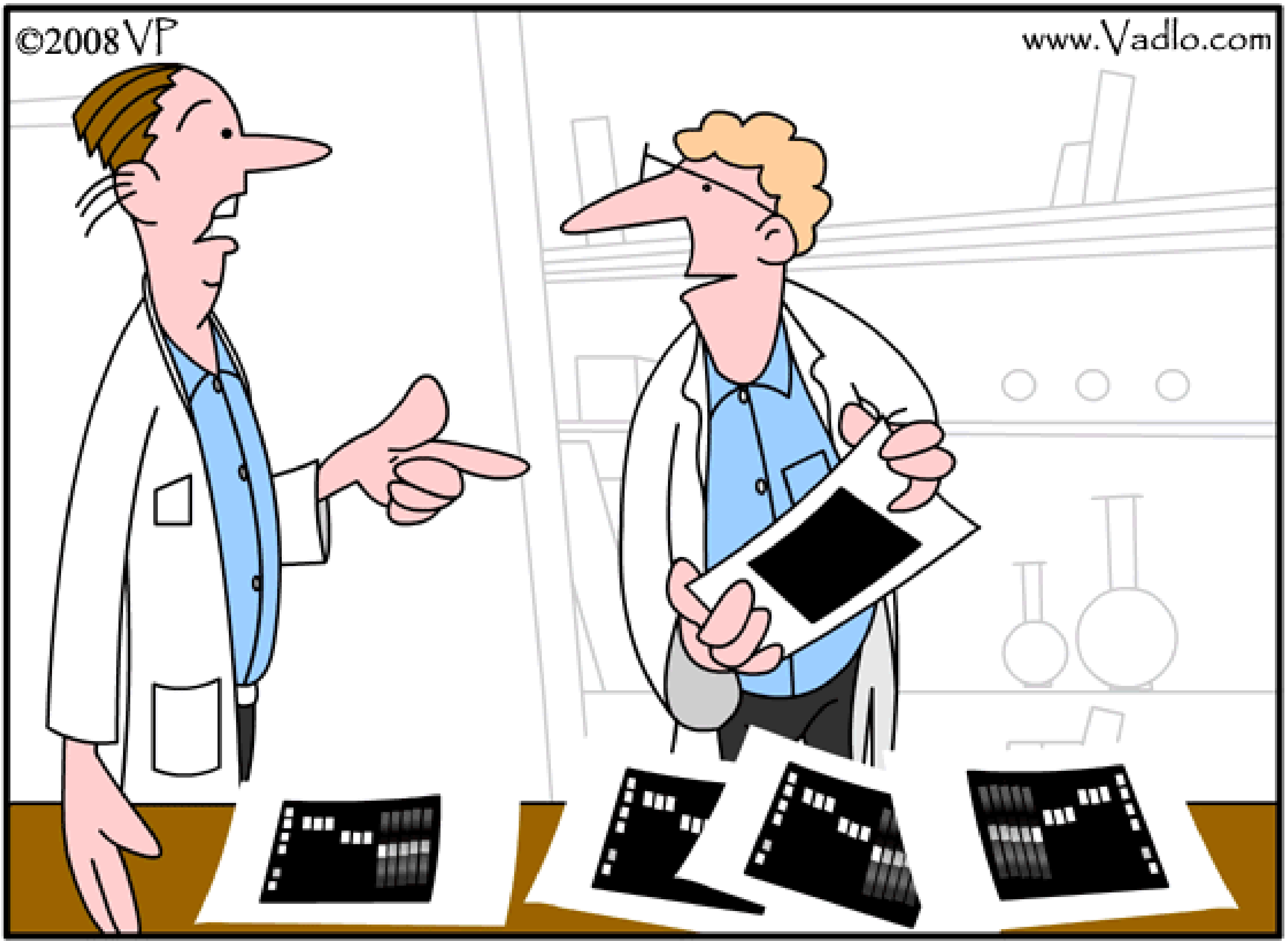












Data don't make any sense, we will have to resort to ***statistics****.*

Statistical Tests

- Assuming two groups
  - Group variable is independent variable
  - Outcome (variable being tested) is dependent variable
- How do you determine what statistical test to use?

Statistical Tests

- Are the groups independent or paired?
- What type of data do you have for the dependent variable?
  - Continuous or categorical
- How is the data distributed?
- What is the hypothesis?
- Is the variable measured across time at multiple time points?

Example: Continuous Data

- Normally distributed
  - Since it follows a distribution, use a parametric method
- Define null and alternative
  - H0: Mean birthweight in women treated with 17P = mean birthweight in women receiving placebo
  - H1: Mean birthweight in women treated with 17P ≠ mean birthweight in women receiving placebo
- Use t-test or Z-test
  - T-test will equal Z-test when n is 50 or greater
- Report mean and standard deviation at a minimum
  - Median, range, or 25th-75th percentiles often reported as well
  - Report test statistic and associated p-value

Example: Continuous Data

- What if not normally distributed
  - Ex. Gestational age at delivery

- Two options
  - Transform the data
    - Use a mathematical transformation such as log(x) – transformed data often normal
      - Difficult to interpret
  - Use a non-parametric analysis method
    - Wilcoxon Rank Sum Test is the non-parametric test analogous to a t-test
    - Report median, range or 25th-75th percentiles
    - Report test statistic and associated p-value

Example: Paired Data

- Paired data
  - Ex. Measure blood pressure prior to receipt of a an intervention and then after receipt of the intervention
  - Each subject serves as own control
  - Follows normal distribution – paired t-test
  - Not normally distributed – Wilcoxon Sign Rank test

Example: Categorical Data

- Ordinal and non-ordered categorical data can be tested using the chi-square test
  - Tests that there is a difference between groups
  - Does not indicate where the difference is

- For ordinal data, can also do a test for linearity
  - Ex. Mantel-Haenzel Chi-square test

Example: Time to Event Data

- Time to event data
  - Ex. Time to death (classic survival analysis)
  - Time to delivery
  - Time to Failure

- Special statistical methods that allow comparison of time to event data even when not all subjects have the event (data are censored)

- Survival analysis techniques
  - Most common is called Kaplan-Meier analysis

Additional Statistical Tests

- Many other statistical methods
  - More than 2 groups
    - Analysis of variance (parametric) for continuous data
    - Kruskal-Wallis (nonparametric) for continuous data
    - Chi-square for categorical data
- Controlling for covariates and modeling
  - Cochran-Mantel-Haenzel chi-square
  - Linear regression
  - Logistic regression
  - Cox regression

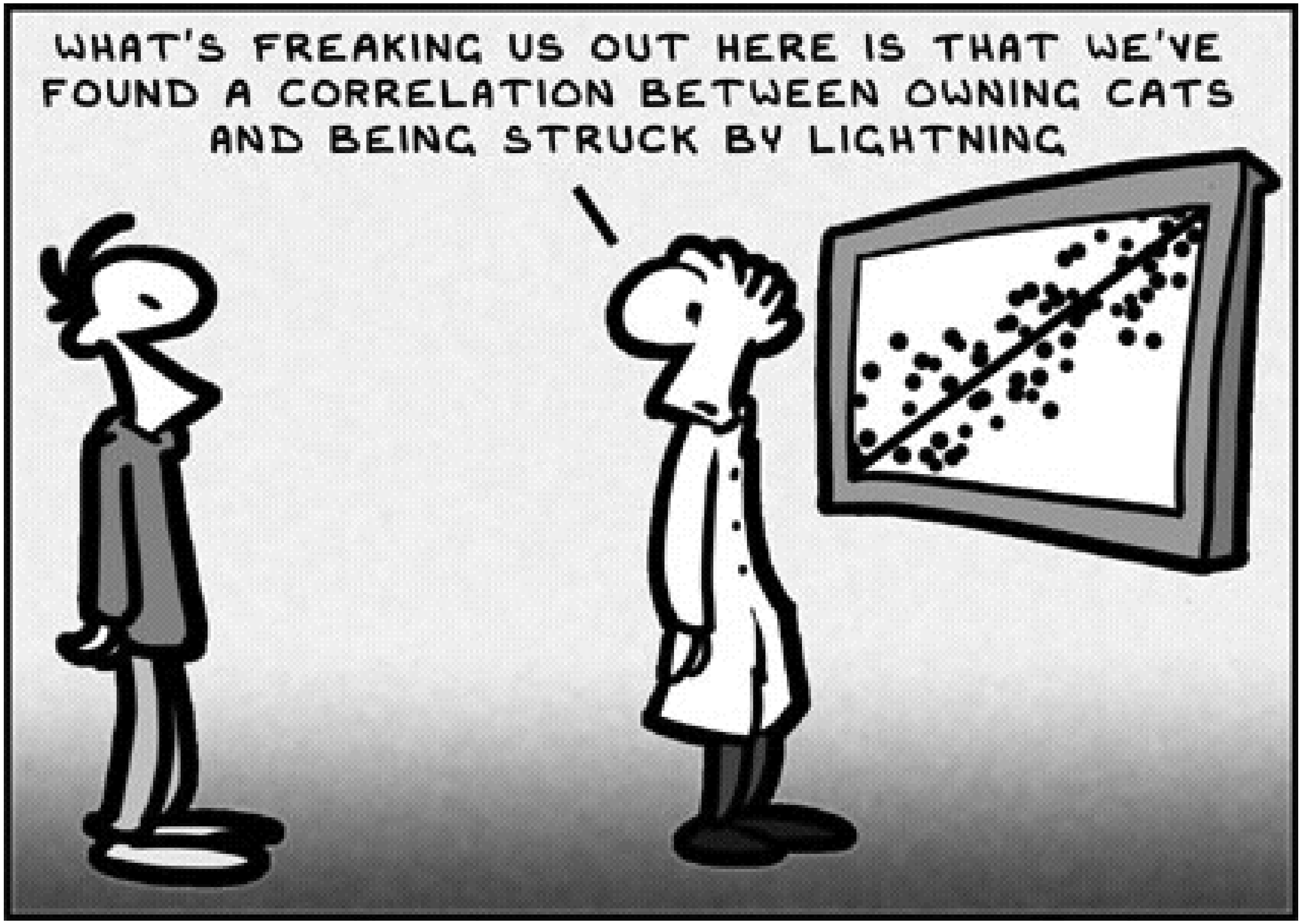
Interpretation Tips (1)

- Don’t scan for p values and interpret p values carefully
  - $0.01<=p<0.05$ – results are significant
  - $0.001<=p<0.01$ – results are highly significant
  - $p<0.001$ – results are very highly significant
  - $p>0.05$ – not statistically significant
  - $0.05<=p<0.10$ trend towards statistical significance is sometimes noted

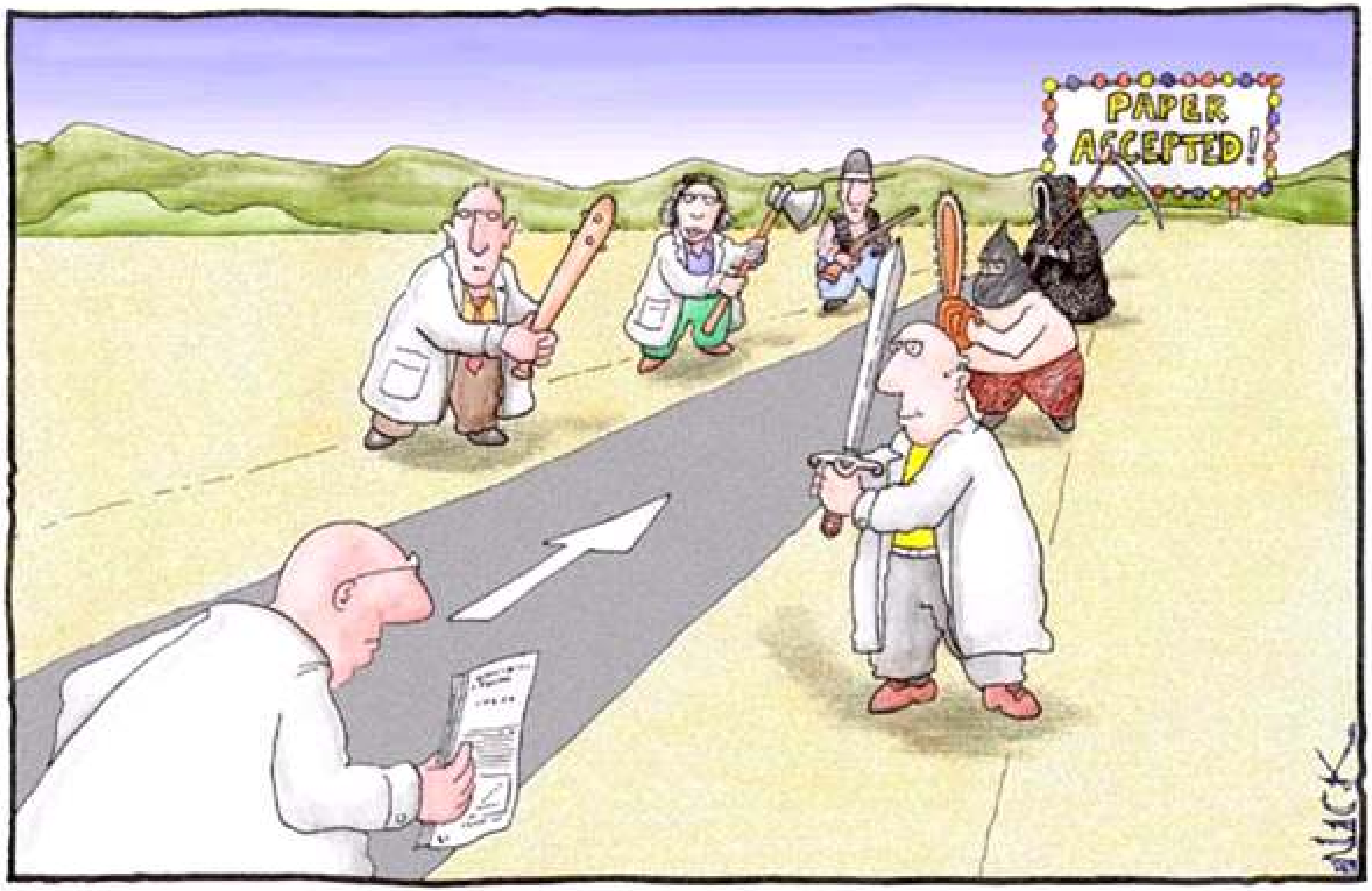
Interpretation Tips (1)

- Correlation: statistical relationship between two or more variables. Degree of correlation is measurable for linear and non-linear relationships.

- Significance: which one are you referring to? And did you interpret your results for both types?
  - Clinical
  - Statistical


WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

CONSORT



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

Background

- ***Con***solidated ***S***tandards ***o***f ***R***eporting ***T***rials
- Colleagues suggested that editors could improve the reporting of trials by providing authors with a list of items that they were expected to report
- Early in the 1990s two groups met to discuss
- Statement published for 2-group parallel design, revisions for 2010, extensions to other types of trials as well (cluster, equivalence, NI)

Background

- Trials with inadequate methodological approaches may be associated with exaggerated treatment effects

- Biased results from poorly designed and reported trials can mislead decision making in health care

- Critical appraisal of the quality is only possible if the design, conduct and analysis is described in published articles

Checklist and Updates (1)

- Example updates to the 2010 Statement:
  - Item 2b: Specific objectives or hypotheses
  - Item 3a: specify basic trial design
  - Item 5: sufficient details on the intervention to allow for replication
  - Item 6: any changes to the primary and secondary endpoints
  - Item 9: beyond "banal, assurances of concealment"—actual steps taken

Checklist and Updates (2)

- Example updates to the 2010 Statement:
  - Item 12a: statistical methods for secondary outcomes
  - Item 17b: for clinical interpretability added that for binary outcomes both relative and absolute effect size given
  - Item 24: availability of protocol
  - Item 25: funding source

Contact Information

Nicole C. Close, PhD
nclose@empiristat.com
240-744-0000
www.empiristat.com