UNIVERSITY OF TORONTO

An Introduction to Sample Size Estimation

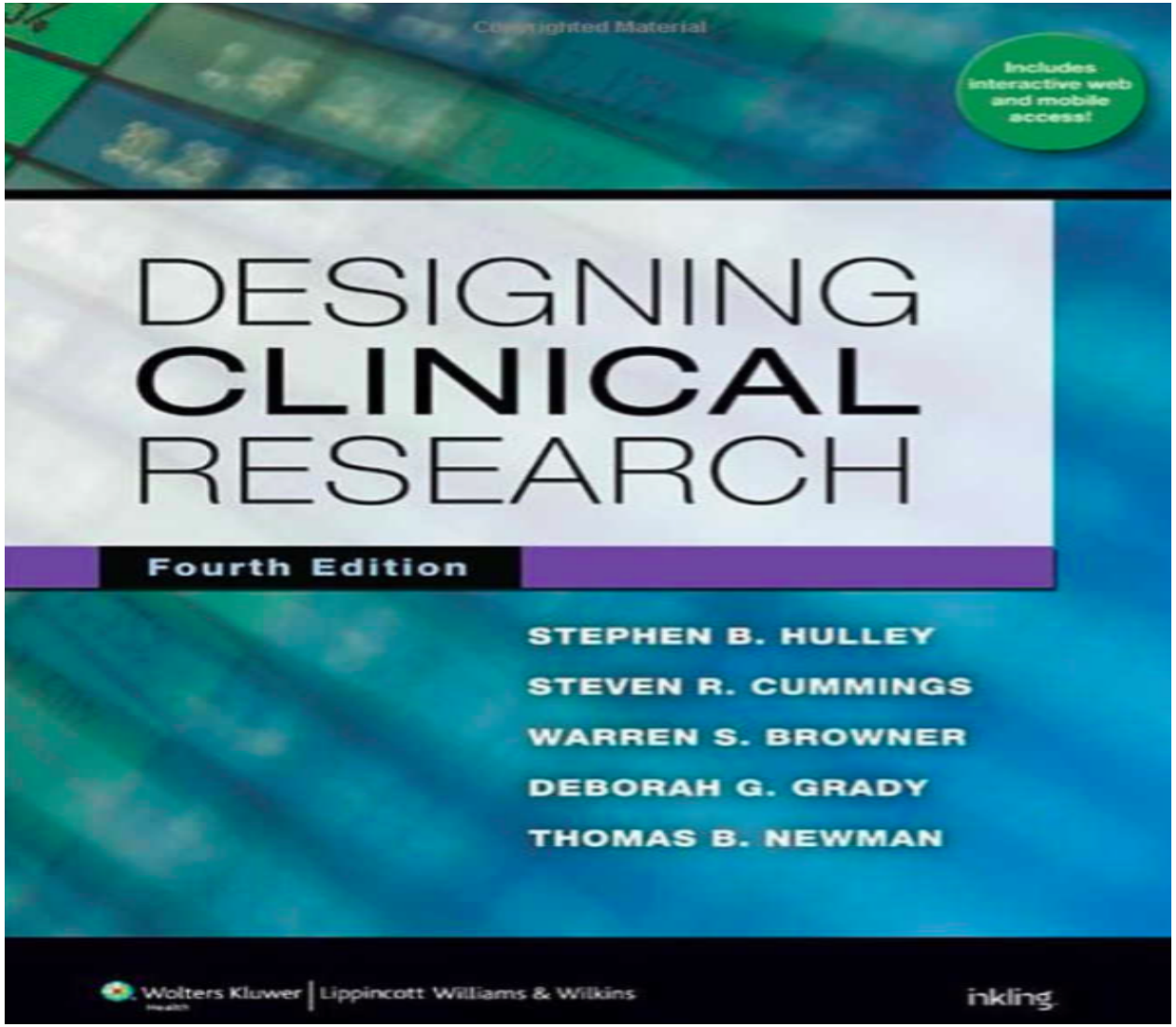
Theory and Applications

Christopher Meaney

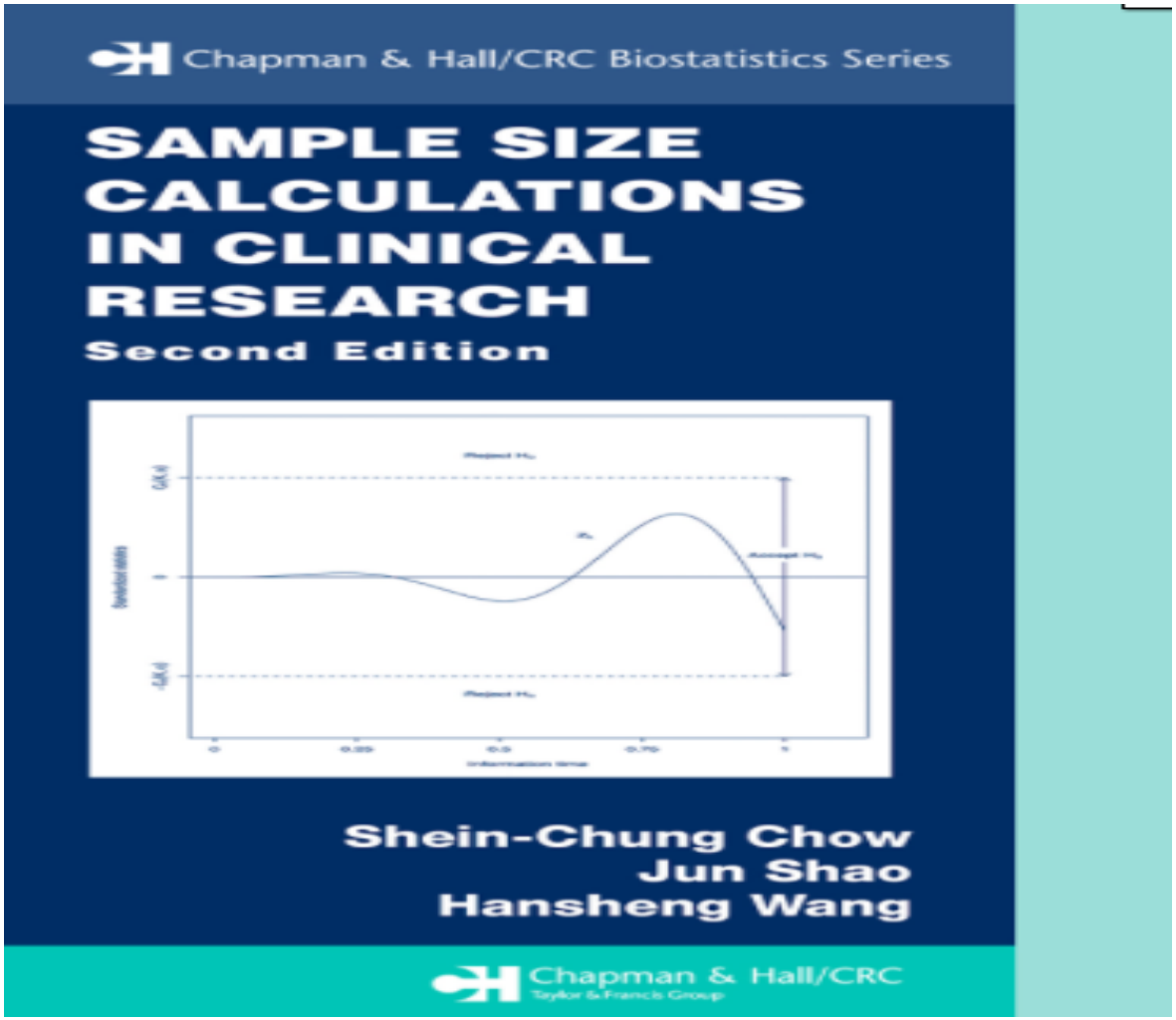
Biostatistician
Department of Family and Community Medicine
University of Toronto

October 5, 2022

UNIVERSITY OF TORONTO



UNIVERSITY OF TORONTO



UNIVERSITY OF TORONTO

Research Questions/Statements

The development of a good research question/statement is a crucial aspect of any study.

Research projects can/will fail if the study objective is poorly defined or underdeveloped.

Some aspects of good research questions/statements:

- Precisely/succinctly stated.
- Formulated early in design stages of study.
- Grounded in science, possibly explaining clinical relevance of problem.
- Implicit/explicit specification of study design.
- Clear delineation of study sample and target population for inference.
- Specify main outcome variable(s) and predictor variable(s) (if applicable).

Various frameworks exist to help investigators refine research question:

- PICO/PICOT
- FINER

Research Questions/Statements

A dichotomous framework for research studies: **descriptive** versus **analytic**.

Descriptive studies:

- Use information in sample to estimate some quantity.
- Describe uncertainty about population quantity using confidence intervals.

Analytic studies:

- Use information in sample to investigate association between treatment, exposure, intervention, etc. and outcome variable.
- Emphasis is on statistical hypothesis testing.

Different approach to sample size estimation for descriptive vs. analytic studies.

UNIVERSITY OF TORONTO

Examples of Descriptive Research Statements:

Example 1: Alcohol Use Amongst Teens

Using information from the Canadian Community Health Survey (CCHS), the purpose of this study is to estimate the proportion of Canadian teenagers (aged 13-19 years) who have drank alcohol at least once in the past 12 months.

Example 2: Breast Cancer Prevalence in Adult Women

This study aims to provide an estimate of breast cancer prevalence (in 2015), amongst women aged 40-65, using registry information from CCO.

Example 3: Mean BMI Amongst Elderly Diabetic Females

Using information from CPCSSN, the objective of this study is to estimate the mean BMI level of Canadian females (aged 65+), with diagnosed diabetes.

UNIVERSITY OF TORONTO

Examples of Analytic Research Questions/Statements:

Example 1: Thiazide Diuretics and Blood Pressure Control

This randomized controlled trial aims to investigate the impact of thiazide diuretics in reducing mean arterial pressure at 1 month post-randomization, in a sample of adults (aged 18-65) with clinically diagnosed hypertension.

Example 2: Occupational Asbestos Exposure and Lung Cancer Development

Using linked information from CCO and IWH, this matched case-control study aims to investigate differences in the likelihood of developing lung cancer amongst those with/without previous occupational exposure to asbestos.

UNIVERSITY OF TORONTO

Sample Size Planning, Estimation and Justification

Why bother conducting a sample size calculation?

- Ensures that inferences can be made with some statistical certainty.
- Ethically, it ensures scarce resources are not wasted.

Sample Size Calculations in Clinical Research; Chow, Shao & Wang (2003)

To provide an accurate and reliable sample size calculation, an appropriate statistical procedure for the research objective of interest must be derived under the study design.

Precision Analysis

- Choose sample size such that the study yields an interval estimate with desired precision/width for a fixed confidence level (α).

Power Analysis

- Choose sample size such that the study yields desired power ($1 - \beta$) for detecting a clinically meaningful effect at a fixed type-1 error rate (α).

UNIVERSITY OF TORONTO

Sample Size Planning, Estimation and Justification

The Art & Science of Sample Size Calculation:

- Closed form analytic formulas for **SOME** sample size problems.
- Sample size and power functions take many inputs/parameters.
  - Effect size: minimal clinically important difference (MCID).
  - Variability in response.
  - Acceptable type-1 error rate (α).
  - Desired power ($1 - \beta$), where β is type-2 error rate.
- Often scientists **DO NOT** have good estimates of all inputs:
  - Review literature.
  - Conduct pilot studies.
- Sensitivity analyze to determine how parameters impact sample size.

Myth: There is a Single Approach to Sample Size Estimation for All Problems

Sample size calculations need be tailored to your specific research problem!!

UNIVERSITY OF TORONTO

Statistical Inference

Given sample data, how do we make inferences about population parameters?

Mathematically: given $(x_1, x_2, \cdots x_n) \sim \mathcal{F}(\theta)$.

- Inference concerns development/evaluation of procedures to determine θ.

Example: Assume BMI follows a normal distribution.

- Estimate the mean (μ) and variance (s^2) of the BMI distribution.

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi s^2}} \exp\Big(-\frac{1}{2} \frac{(x-\mu)^2}{s^2} \Big)$$

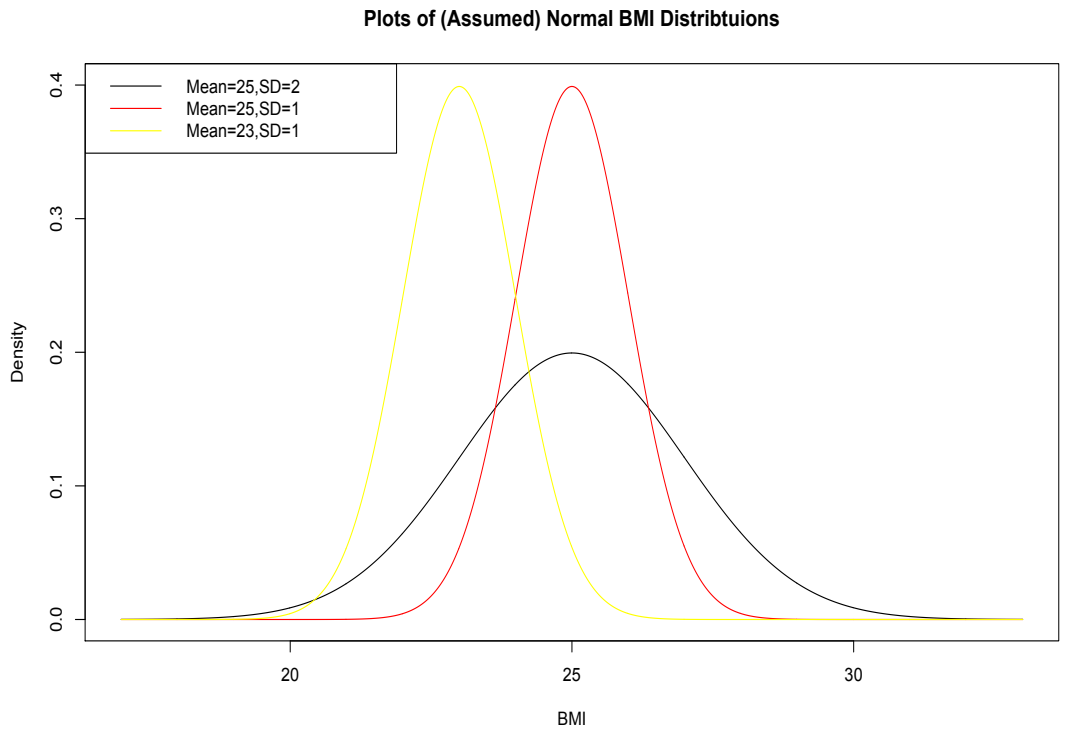
Parametric family described by two parameters: $\theta \in \{\mu, s^2\}$.

- Mean (μ): used to describe location/centrality of distribution.
- Variance (s^2): used to describe dispersion/spread of distribution.

UNIVERSITY OF TORONTO

The Univariate Normal/Gaussian Distribution

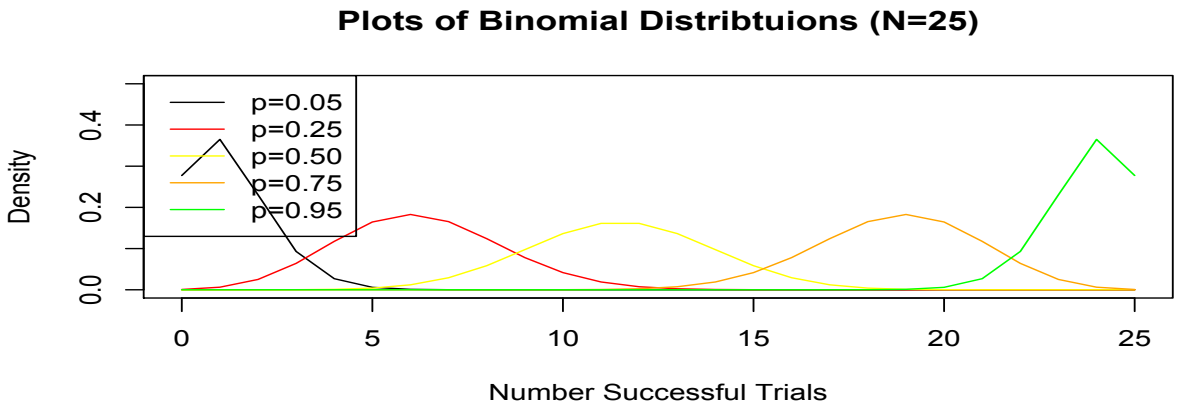
- $f(x;\mu,s^2) = \frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{s^2}\right)$
- Model for unbounded symmetrically distributed continuous data.
- Statistical Modelling: t-test, ANOVA, linear regression, etc.

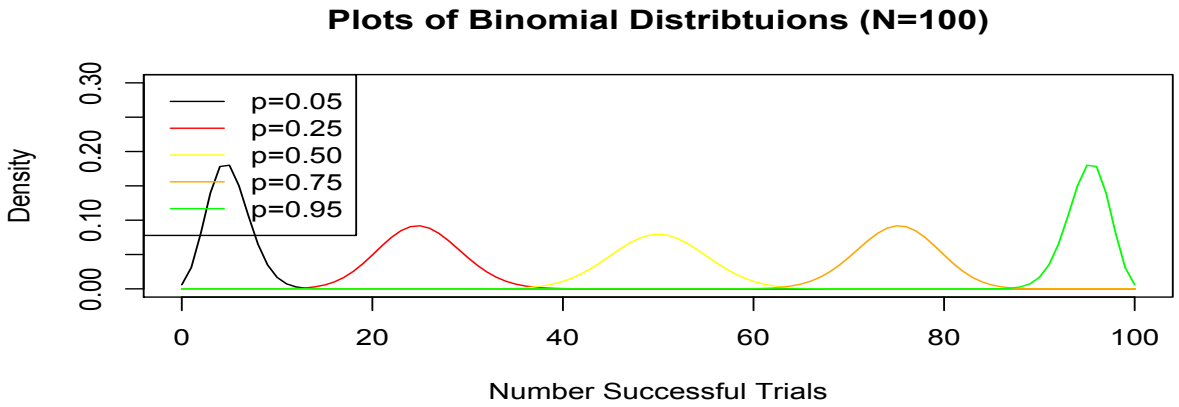


UNIVERSITY OF TORONTO

Binomial Distribution

- $f(x;p) \sim \binom{n}{x} p^x (1-p)^{n-x}$
- Statistical Modelling: Prevalence studies, sensitivity/specificity, etc.

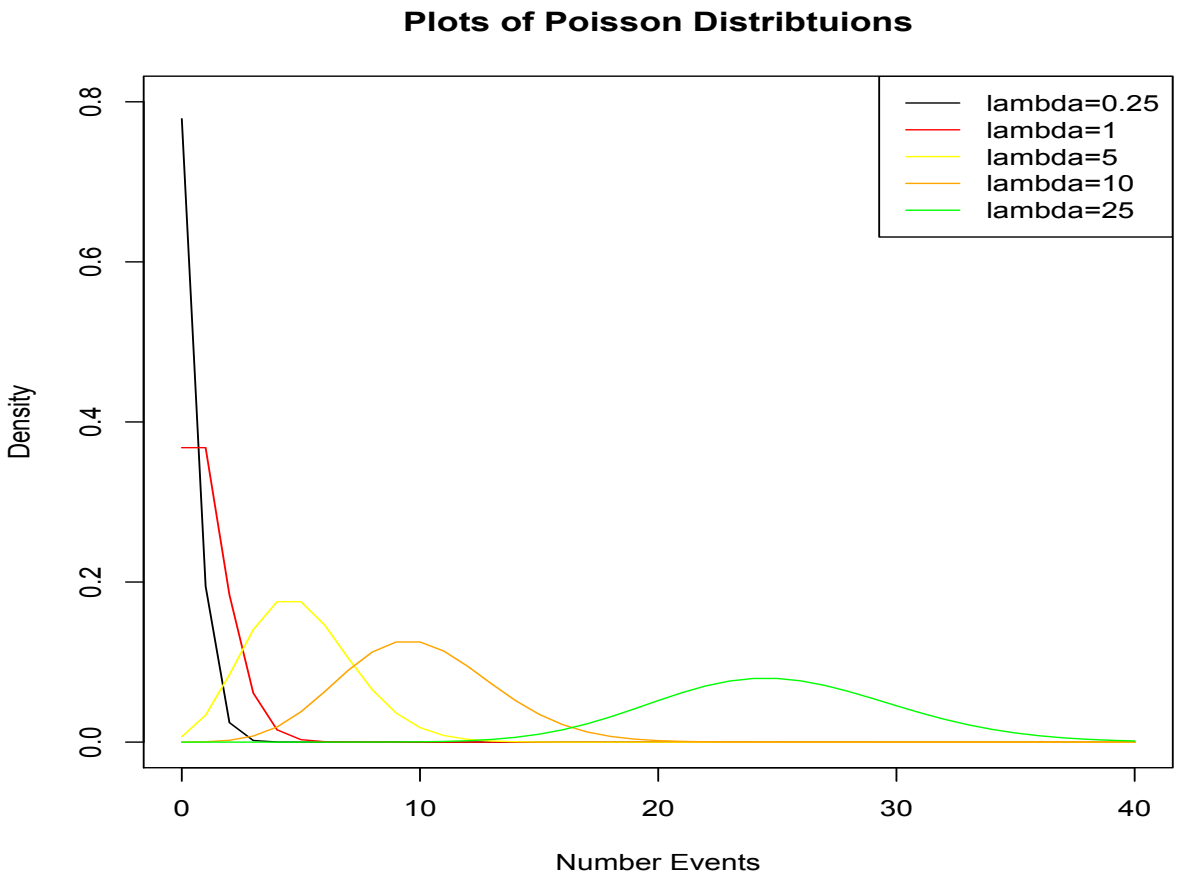




UNIVERSITY OF TORONTO

Poisson Distribution

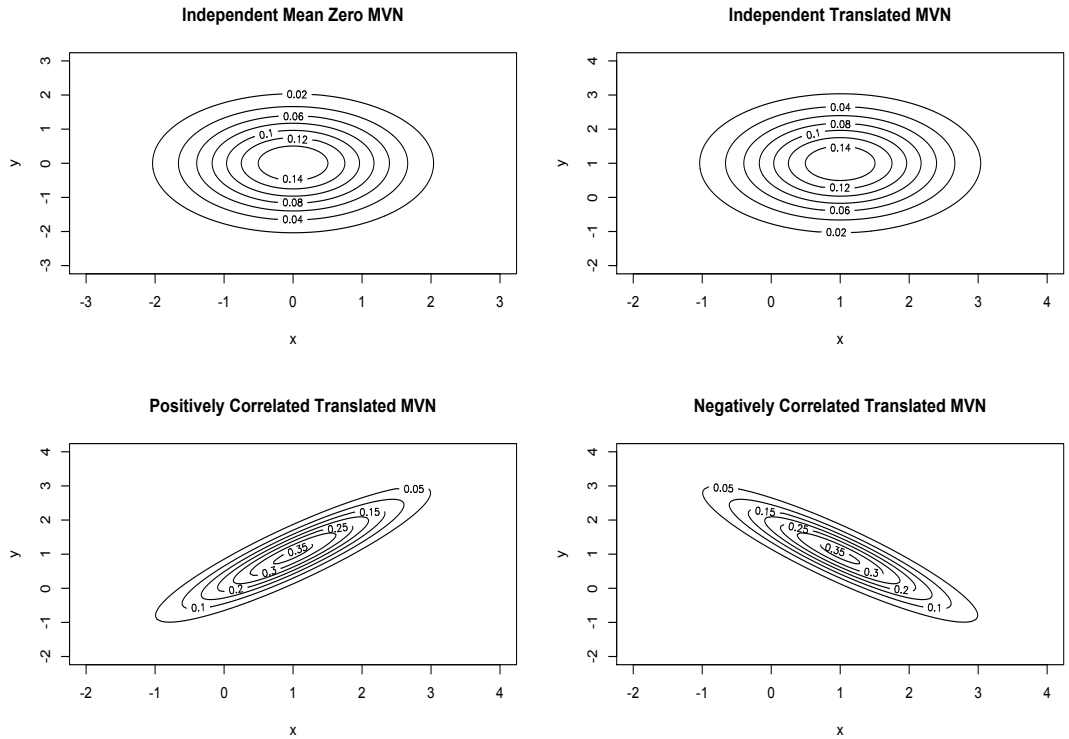
- $f(x;\lambda) \sim \frac{e^{\lambda}\lambda^{x}}{x!}$
- Statistical Modelling: Incidence studies, Poisson regression, etc.



UNIVERSITY OF TORONTO

Bivariate/Multivariate Normal Distribution

- $f(x;\mu,\Sigma) \sim \frac{1}{(2\pi)^{\frac{p}{2}}} \frac{1}{|\Sigma|} \exp(-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu))$
- Statistical Modelling: Dependence between two continuous variables.



UNIVERSITY OF TORONTO

Point Estimation

Use sample data to provide single estimate of unknown population parameter.

$$\hat{\theta} = g(x_1, x_2, \cdots, x_n)$$

Estimators specific to statistical models:

- Proportions for Bernoulli/Binomial data.
- Rates for Poisson data.
- Means/Variances for Continuous Normal data.
- Means/Variances/Covariances for Continuous Multivariate Normal Data.

Sampling Distribution

- Describes variability of an estimator around true population parameter.
- Quantified using measures such as variance and standard error.

Interval Estimation

- A random interval, such that on repeated sampling $100 * (1 - \alpha)$ percent of such intervals would contain the true population quantity.
- Yields more information than a single point estimate about value true population parameter.

UNIVERSITY OF TORONTO

Hypothesis Testing

- Framework for testing that population parameter equals some value.
- Does sample data provide evidence enabling falsifiability H_0.

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \neq \theta_0$$

Hypothesis testing: a binary decision framework.

	H_0 False	H_0 True
Test Reject H_0	Correct Decision	Type-1 Error (α)
Test Maintains H_0	Type-2 Error (β)	Correct Decision

UNIVERSITY OF TORONTO

Hypothesis Testing Terminology:

- **Null Hypothesis (H_0):** Simple proposition no effect/association.
- **Alternative Hypothesis (H_a):** Proposition effect/association exists.
- **Effect Size (δ):** Magnitude of the effect/association (clinically relevant).
- **Test Statistic (T):** Function data/parameters; $T = g(x_1, x_2, \cdots x_n; \theta)$.
- **Rejection/Critical Region (c_α):** Value such that reject H_0 if $T_{\text{obs}} > c_\alpha$.
- **P-value:** Probability, under H_0, observing T as extreme as t_{obs}.
- **Type-1 Error (α):** Falsely reject true null (too complex model).
- **Type-2 Error (β):** Fail to reject null when true effect (too simple model).
- **Power ($1 - \beta$):** Probability reject null, given true effect exists.

UNIVERSITY OF TORONTO

Hypothesis Testing

A process for statistical hypothesis testing:

- Begin with initial research statement/objective.
- Determine null hypothesis (H_0) and alternative hypothesis (H_a).
- Decide appropriate test statistic (T).
  - Consider design, sampling, measurement outcome/predictor, etc.
- Derive distribution of test statistic (T) under null hypothesis.
- Select tolerable type-1 error rate (α); determines significance level.
- Determine appropriate critical value (c_α), such that reject H_0 if $t_{\text{obs}} > c_\alpha$.
- Compute observed value of test statistic (t_{obs}).
- Decide whether evidence to reject H_0, if $t_{\text{obs}} > c_\alpha$.

A slightly more modern framework for statistical hypothesis testing:

- Compute observe value test statistic (t_{obs}).
- Calculate p-value (probability under H_0 observe test stat as extreme t_{obs}).
- Reject H_0 if p-value less than significance level (α); typically $p < 0.05$.

Note: the two approaches are equivalent.

UNIVERSITY OF TORONTO

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

UNIVERSITY OF TORONTO

TABLE 5.1 The Analogy between Jury Decisions and Statistical Tests

Jury Decision	Statistical Test
Innocence: The defendant did not counterfeit money.	**Null hypothesis:** There is no association between dietary carotene and the incidence of colon cancer in the population.
Guilt: The defendant did counterfeit money.	**Alternative hypothesis:** There is an association between dietary carotene and the incidence of colon cancer.
Standard for rejecting innocence: Beyond a reasonable doubt.	**Standard for rejecting null hypothesis:** Level of statistical significance (α).
Correct judgment: Convict a counterfeiter.	**Correct inference:** Conclude that there is an association between dietary carotene and colon cancer when one does exist in the population.
Correct judgment: Acquit an innocent person.	**Correct inference:** Conclude that there is no association between carotene and colon cancer when one does not exist.
Incorrect judgment: Convict an innocent person.	**Incorrect inference (type I error):** Conclude that there is an association between dietary carotene and colon cancer when there actually is none.
Incorrect judgment: Acquit a counterfeiter.	**Incorrect inference (type II error):** Conclude that there is no association between dietary carotene and colon cancer when there actually is one.

UNIVERSITY OF TORONTO

Estimating a Confidence Interval for a Continuous Mean

Setting: Descriptive study with single continuous outcome.

Goal: Estimate the mean and provide an estimate of uncertainty (e.g. 95% CI).

Interval Estimation Procedure for Continuous Mean:

- Estimate the sample mean ($\bar{x}$): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- If unknown, estimate the sample variance (s^2): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Determine appropriate confidence level (α).
- Determine critical value of t-distribution: $t_{\frac{\alpha}{2}}(r)$ with $r = n - 1$ DF.
- Construct CI: $\left(\bar{x} - t_{\frac{\alpha}{2}}(r) * \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(r) * \frac{s}{\sqrt{n}}\right)$

Interpretation: On repeated sampling, $100 * (1 - \alpha)$ percent of CI's generated using above procedure will capture true population parameter.

Interpretation: High confidence parameter is between upper/lower limits CI.

UNIVERSITY OF TORONTO

Estimating a Confidence Interval for a Binomial Proportion

Setting: Descriptive study with single Bernoulli/Binomial outcome.

Goal: Estimate the proportion and provide uncertainty estimate (e.g. 95% CI).

Interval Estimation Procedure for Bernoulli/Binomial Proportion:

- Estimate the sample mean ($\hat{p}$): $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Estimate standard error of $\hat{p}$: $se(\hat{p}) = \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}$
- Determine appropriate confidence level (α).
- Determine critical value of Z-distribution (standard normal): $Z_{\frac{\alpha}{2}}$.
- Construct CI: $\left(\hat{p} - Z_{\frac{\alpha}{2}} * se(\hat{p}), \hat{p} + Z_{\frac{\alpha}{2}} * se(\hat{p})\right)$

Interpretation: On repeated sampling, $100 * (1 - \alpha)$ percent of CI's generated using above procedure will capture true population parameter.

Interpretation: High confidence parameter is between upper/lower limits CI.

UNIVERSITY OF TORONTO

Testing the Equality of Means from Two Continuous Distributions

Specify null hypothesis (H_0) and alternative hypothesis (H_a):

$$\begin{array}{ll} H_0 & \mu_{\text{treatment}} = \mu_{\text{placebo}} \\ H_a & \mu_{\text{treatment}} \neq \mu_{\text{placebo}} \end{array}$$

Express problem/hypotheses in terms effect size ($\delta = \mu_{\text{treatment}} - \mu_{\text{placebo}}$):

$$\begin{array}{ll} H_0 & \delta = 0 \\ H_a & \delta \neq 0 \end{array}$$

Construct test statistic (determine distribution under null hypothesis):

$$T = \frac{\bar{X}_{\text{treatment}} - \bar{X}_{\text{placebo}}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Pooled standard deviation: $s_p = \sqrt{\frac{(n_1-1)*s^2_{\text{treatment}} + (n_2-1)*s^2_{\text{placebo}}}{n_1+n_2-2}}$

s^2 represent sample variance. n_1 and n_2 represent group specific sample sizes.

UNIVERSITY OF TORONTO

Testing the Equality of Means from Two Continuous Distributions

Null Distribution: Under the null hypothesis $T \sim \mathcal{T}_{n_1+n_2-2}$.

- i.e. Test statistic follows Student's t-distribution with $(n_1 + n_2 - 2)$ DF.

Significance Level: Biomedical research typically uses $\alpha = 0.05$ as threshold.

- Can determine rejection region as quantile of $\mathcal{T}_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$ distribution.

Main Ingredients for Test Statistic:

- Effect size (difference in sample means): $\bar{X}_{\text{treatment}} - \bar{X}_{\text{placebo}}$.
- Variability in measured effect size (pooled standard deviation): s_p.
- Group specific sample sizes: n_1 and n_2.
- Significance level (α). Determines threshold value c_α.

Assumptions: Independence, normality, homoskedasticity, etc.

Inference: Reject H_0 at significance level α if $|t_{\text{obs}}| > c_\alpha$. Else fail to reject H_0.

UNIVERSITY OF TORONTO

Testing the Equality of Two Binomial Proportions

Specify null hypothesis (H_0) and alternative hypothesis (H_a):

$$\begin{array}{ll} H_0 & p_{\text{treatment}} = p_{\text{placebo}} \\ H_a & p_{\text{treatment}} \neq p_{\text{placebo}} \end{array}$$

Express problem/hypotheses in terms effect size ($\delta = p_{\text{treatment}} - p_{\text{placebo}}$):

$$\begin{array}{ll} H_0 & \delta = 0 \\ H_a & \delta \neq 0 \end{array}$$

Construct test statistic (determine distribution under null hypothesis):

$$Z = \frac{\hat{p}_{\text{treatment}} - \hat{p}_{\text{placebo}}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Proportion success combined sample: $\bar{p} = \frac{n_1 * \hat{p}_{\text{treatment}} + n_2 * \hat{p}_{\text{placebo}}}{n_1 + n_2} = \frac{x_{\text{treatment}} + x_{\text{placebo}}}{n_1 + n_2}$

n_1 and n_2 represent group specific sample sizes.

UNIVERSITY OF TORONTO

Testing the Equality of Two Binomial Proportions

Null Distribution: Under the null hypothesis $Z \sim \mathcal{N}(0, 1)$.

- i.e. Test statistic follows standard Normal distribution (mean=0, SD=1).

Significance Level: Biomedical research typically uses $\alpha = 0.05$ as threshold.

- Can determine rejection region as quantile of $\mathcal{N}_{\frac{\alpha}{2}}(0, 1)$ distribution.

Main Ingredients for Test Statistic:

- Effect size (difference in sample proportions): $\hat{p}_{\text{treatment}} - \hat{p}_{\text{placebo}}$.
- Response variability (pooled standard deviation): $\sqrt{\bar{p}(1 - \bar{p})}$.
- Group specific sample sizes: n_1 and n_2.
- Significance level (α). Determines threshold value c_α.

Assumptions: Independence, probability event not too small, etc.

Inference: Reject H_0 at significance level α if $|z_{\text{obs}}| > c_\alpha$. Else fail to reject H_0.

UNIVERSITY OF TORONTO

Testing for an Association Between Two Continuous Measures

Bivariate Data: (x_1, y_1), (x_2, y_2),...,(x_n, y_n).

Preliminaries: Estimate correlation between continuous measures (x_i, y_i).

Pearsons Correlation Coefficient (ρ): Measure linear dependence.

- Range Pearson correlation coefficient: (-1,1).

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Fisher Z-transformation: New RV follows standard normal distribution.

$$F(\rho) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = \text{arctanh}(\rho)$$

Distribution Transformed RV: $F(\rho) \sim \mathcal{N}(\text{arctanh}(\rho), \frac{1}{n-3})$.

Specify null hypothesis (H_0) and alternative hypothesis (H_a):

$$\begin{array}{ll} H_0 & \rho = 0 \\ H_a & \rho \neq 0 \end{array}$$

Construct test statistic:

$$Z = \left(F(\hat{\rho}) - F(\rho_0)\right) * \sqrt{n-3}$$

Null Distribution: Under the null hypothesis $Z \sim \mathcal{N}(0,1)$.

- Can determine rejection region as quantile of $\mathcal{N}_{\frac{\alpha}{2}}(0,1)$ distribution.

Main Ingredients for Test Statistic:

- Effect size (magnitude association): Is $|\hat{\rho}|$ near zero vs. one.
- Sample size: number independent (x_i, y_i) observations.
- Significance level (α). Determines threshold value c_α.

Assumptions: Independence, data from bivariate normal distribution, etc.

Inference: Reject H_0 at significance level α if $|z_{\text{obs}}| > c_\alpha$. Else fail to reject H_0.

UNIVERSITY OF TORONTO

Sample Size Estimation for a Confidence Interval of a Continuous Mean

Example 6.4: Mean Hemoglobin Levels Among Urban 3rd Graders (p. 63)

"The investigator seeks to determine the mean hemoglobin level among third graders in an urban area with a 95% confidence interval, with half-width ±0.3g/dL. A previous study found that the SD of hemoglobin was 1g/dL."

Required parameters/inputs to Sample Size Calculation:

- Standard deviation response variable (s): 1g/dL.
- Total width of CI (W): 0.6g/dL.
- Confidence level ($Z_{\frac{\alpha}{2}}$): 95%.

Sample Size Formula: Invert procedure for estimating CI (solving for n).

$$N = \frac{4 * Z^2_{\frac{\alpha}{2}} * s^2}{W^2} = \frac{4 * (1.96)^2 * (1)^2}{0.6^2} = 42.684$$

Solution: A sample size of N=43 3rd graders is required to estimate a 95% CI with total width of 0.6g/dL, assuming the SD of the response is 1g/dL.

UNIVERSITY OF TORONTO

Sample Size Estimation for a Confidence Interval of a Binomial Proportion

Example 6.5: Sensitivity of a Diagnostic Test for Pancreatic Cancer (p. 64)

"The investigator wishes to determine the sensitivity of a new diagnostic test for pancreatic cancer. Based on a pilot study, she expects 80% of patients with pancreatic cancer will have a positive test result. How many patients are needed to estimate a 95% for the tests sensitivity with half width ± 0.05 units?"

Required parameters/inputs to Sample Size Calculation:

- Expected proportion positive responses: $p = 0.80$.
- Total width of CI (W): 0.10 units.
- Confidence level ($Z_{\frac{\alpha}{2}}$): 95%.

Sample Size Formula: Invert procedure for estimating CI (solving for n).

$$N = \frac{4 * Z^2_{\frac{\alpha}{2}} * (p) * (1-p)}{W^2} = \frac{4 * (1.96)^2 * (0.80) * (0.20)}{0.10^2} = 245.862$$

Solution: A sample size of N=246 cancer patients is required to estimate a 95% CI with total width of 0.10 units, assuming a sensitivity of 0.80.

UNIVERSITY OF TORONTO

Sample Size Estimation for a Two-Sample T-test

Example 6.1: Albuterol vs. Ipratropium for FEV in Asthma Patients (p. 57)

"The researcher plans a 2-arm RCT to investigate of the effect of Albuterol vs. Ipratropium on FEV1, measured two weeks post-randomization. A literature review suggests the mean FEV1 in persons with treated asthma was 2.0L, with a standard deviation of 1.0L. The investigator would like to be able to detect a difference of 10% or more in mean FEV1 between the two treatment groups. How many patients are required in each group if we assume a $\alpha = 0.05$ significance level and want the study to have 80% power?"

Required parameters/inputs to Sample Size Calculation:

- Null Hypothesis (H_0): $\mu_{\text{albuterol}} = \mu_{\text{ipratropium}}$
- Alternative Hypothesis (H_a): $\mu_{\text{albuterol}} \neq \mu_{\text{ipratropium}}$
- Effect Size (δ): clinically important mean difference ($10\% * 2.0L = 0.2L$)
- Standard Deviation Response (s): 1.0L
- Significance level (α): 0.05
- Power ($1 - \beta$): $0.80 \Rightarrow \beta = 1 - 0.80 = 0.20$

UNIVERSITY OF TORONTO

Sample Size Estimation for a Two-Sample T-test

Sample Size Formula: Invert test statistic for 2-sample t-test (solving for n).

$$N = \frac{(\frac{1}{q_1} + \frac{1}{q_2}) * s^2 * (Z_\alpha + Z_\beta)^2}{\delta^2} = \frac{4 * 1 * (1.96 + 0.84)^2}{0.2^2} = 392.44$$

Solution: A sample of N=393 asthma patients is needed in both the albuterol and ipratropium arms, in order for the study to have 80% to detect a true effect/difference of 0.2L in FEV1 at a 5% signficance level.

UNIVERSITY OF TORONTO

Sample Size Estimation for a Two-Sample Z-test for Binomial Proportions

Example 6.2: Incidence of Back Pain in Tai-Chi Practitioner vs. Joggers (p. 58)

"The researcher plans a 2-arm prospective cohort study to investigate of the effect of Tai-Chi vs. jogging on the risk of developing back pain at 2-year FUP. Research suggests that 30% of joggers develop back pain; whereas, only 20% of Tai-Chi participants develop back pain. What sample size is needed for the study to have 80% power to detect a 10% difference in the incidence of developing back pain, at a 5% significance level?"

Required parameters/inputs to Sample Size Calculation:

- Null Hypothesis (H_0): $p_{\text{Tai-Chi}} = p_{\text{jogging}}$
- Alternative Hypothesis (H_a): $p_{\text{Tai-Chi}} \neq p_{\text{jogging}}$
- Effect Size (δ): clinically important risk difference ($30\% - 20\% = 10\%$)
- Significance level (α): 0.05
- Power ($1 - \beta$): $0.80 \Rightarrow \beta = 1 - 0.80 = 0.20$

UNIVERSITY OF TORONTO

Sample Size Estimation for a Two-Sample Z-test for Binomial Proportions

Sample Size Formula: Invert test statistic for 2-sample Z-test (solving for n).

$$N = \frac{\left(Z_\alpha * \sqrt{\tilde{p} * (1 - \tilde{p} * (\frac{1}{q_2} + \frac{1}{q_1})} + Z_\beta * \sqrt{p_1 * (1 - p_1) * (\frac{1}{q_1}) + p_2 * (1 - p_2) * (\frac{1}{q_2})} \right)^2}{(p_1 - p_2)^2} = 293.15$$

Note on Terminology:

- $\tilde{p} = q_1 * p_1 + q_2 * p_2$
- q_1: proportion subjects in group 1.
- q_2: proportion subjects in group 2.
- N represents number of subjects per group.

Solution: A sample of N=294 Tai-Chi subjects and N=294 joggers are needed in order for the study to have 80% to detect a true effect/difference of 10% in back pain incidence at a 5% signficance level.

UNIVERSITY OF TORONTO

Sample Size Estimation for a One-Sample Correlation Coefficient Test

Example 6.3: Urinary Cotinine and Bone Density in Smokers (p. 60)

"The investigator seeks to investigate the association between urinary cotinine levels and bone density in a sample of smokers. A previous study indicates urinary cotinine and bone density are negatively correlation ($\rho = -0.3$). How many smokers will need to be enrolled to detect an association of magnitude 0.3 with 90% power at 5% significance level?"

Required parameters/inputs to Sample Size Calculation:

- Null Hypothesis (H_0): $\rho = 0$
- Alternative Hypothesis (H_a): $\rho \neq 0$
- Effect Size (ρ): clinically important correlation ($|\rho| = 0.3$)
- Significance level (α): 0.05
- Power ($1 - \beta$): $0.90 \Rightarrow \beta = 1 - 0.90 = 0.10$

Sample Size Formula: Invert Z-test for correlation coefficient (solving for n).

$$N = \left(\frac{Z_\alpha + Z_\beta}{\frac{1}{2} \ln \frac{1+\rho}{1-\rho}} \right)^2 + 3 = 113$$

Solution: A sample size of N=113 smokers will be needed to reject the null of no correlation with 90% power at a 5% significance level.

UNIVERSITY OF TORONTO

Additional Considerations Related to Sample Size Planning:

- Modifying The Simple Two-Sided Hypothesis Test
  - One-sided vs. two-sided hypothesis tests
  - Superiority, Equivalence and Non-inferiority designs
- Primary and secondary hypotheses
- Multiple Hypothesis Testing
  - Prior probability that hypothesis is correct
  - Control family wise error rate, Bonferroni, control FDR, etc.
- Post-hoc power calculations
- Dropouts, attrition, lost to follow-up, etc.
- Increasingly Complex Statistical Models and Sample Size Planning:
  - Sample size for regression models (multivariate adjustment)
  - Ordinal/Categorical data (multinomial distribution)
  - Survival outcomes, time-to-event outcomes
  - Reliability/Agreement as an outcome (ICC, Kappa, ROC, etc.)
  - Other epidemiological measures effect: OR, RR, RD, HR, etc.
  - Models for non-IID data
- Design Aspects for Reducing Variability Estimates

UNIVERSITY OF TORONTO

Two-Sided Hypothesis Tests

- No a priori specification of direction of effect.
- Will reject H_0 if large effect occurs in positive/negative direction.
- Hence, must split α level mass in two tails.

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_a & \theta \neq \theta_0 \end{array}$$

One-Sided Hypothesis Tests

- Increased power (place all α mass in one tail).
- Innappropriate if uncertain about direction, and impact unexpected effect.

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_a & \theta > \theta_0 \end{array}$$

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_a & \theta < \theta_0 \end{array}$$

UNIVERSITY OF TORONTO

Statistics Notes

One and two sided tests of significance

J Martin Bland, Douglas G Altman

In some comparisons—for example, between two means or two proportions—there is a choice between two sided or one sided tests of significance (all comparisons of three or more groups are two sided).

When we use a test of significance to compare two groups we usually start with the null hypothesis that there is no difference between the populations from which the data come. If this hypothesis is not true the alternative hypothesis must be true—that there is a difference. Since the null hypothesis specifies no direction for the difference nor does the alternative hypothesis, and so we have a two sided test. In a one sided test the alternative hypothesis does specify a direction—for example, that an active treatment is better than a placebo. This is sometimes justified by saying that we are not interested in the possibility that the active treatment is worse than no treatment. This possibility is still part of the test; it is part of the null hypothesis, which now states that the difference in the population is zero or in favour of the placebo.

in th
betw
sided
speci
alter
decre
supp
The
not
incre
and t
are w
are n
shoul
to te
does
week
chan
mont
have

UNIVERSITY OF TORONTO

Superiority Designs

- Goal is to show a new treatment is superior to an existing treatment.

$$\begin{array}{ll} H_0 & \mu_T - \mu_P \leq \Delta \\ H_a & \mu_T - \mu_P > \Delta \end{array}$$

Non-Inferioirity Designs

- Goal is to show new treatment is no-worse than existing treatment.

$$\begin{array}{ll} H_0 & \mu_T - \mu_P \geq \Delta \\ H_a & \mu_T - \mu_P < \Delta \end{array}$$

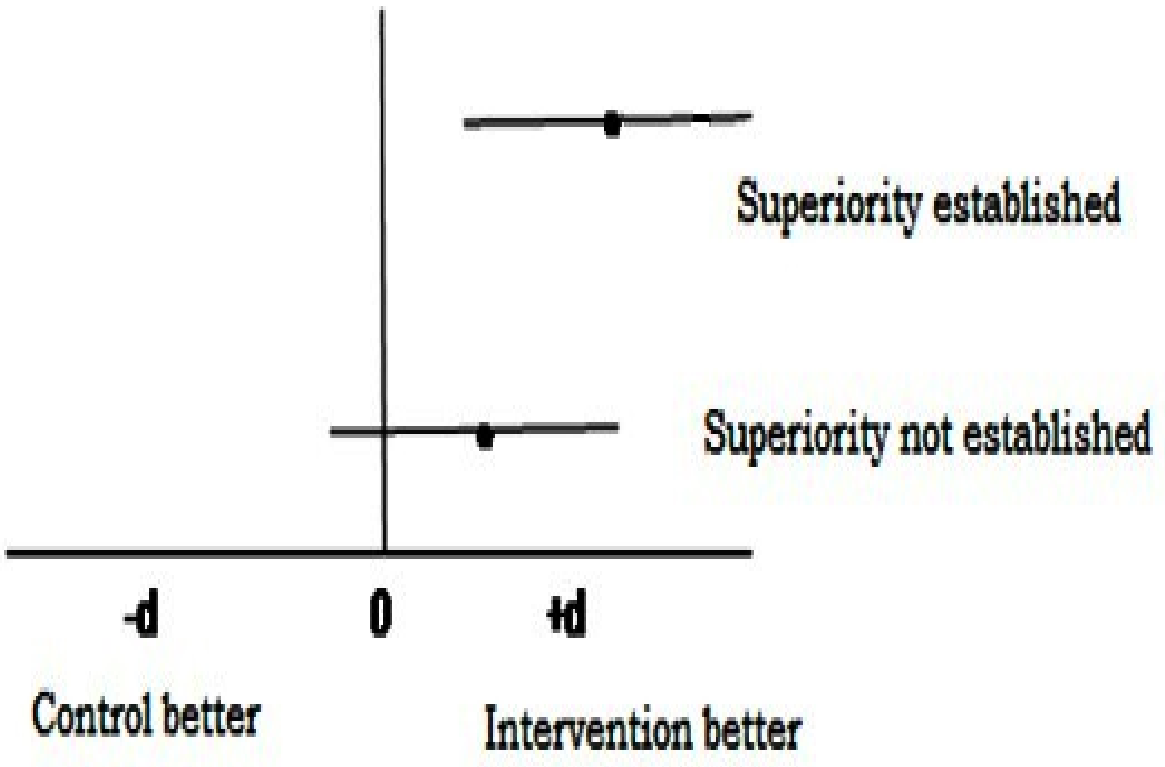
Equivalence Designs

- Goal is to show equivalence of two new treatment and existing treatment.

$$\begin{array}{ll} H_0 & |\mu_T - \mu_P| \geq \Delta \\ H_a & |\mu_T - \mu_P| < \Delta \end{array}$$

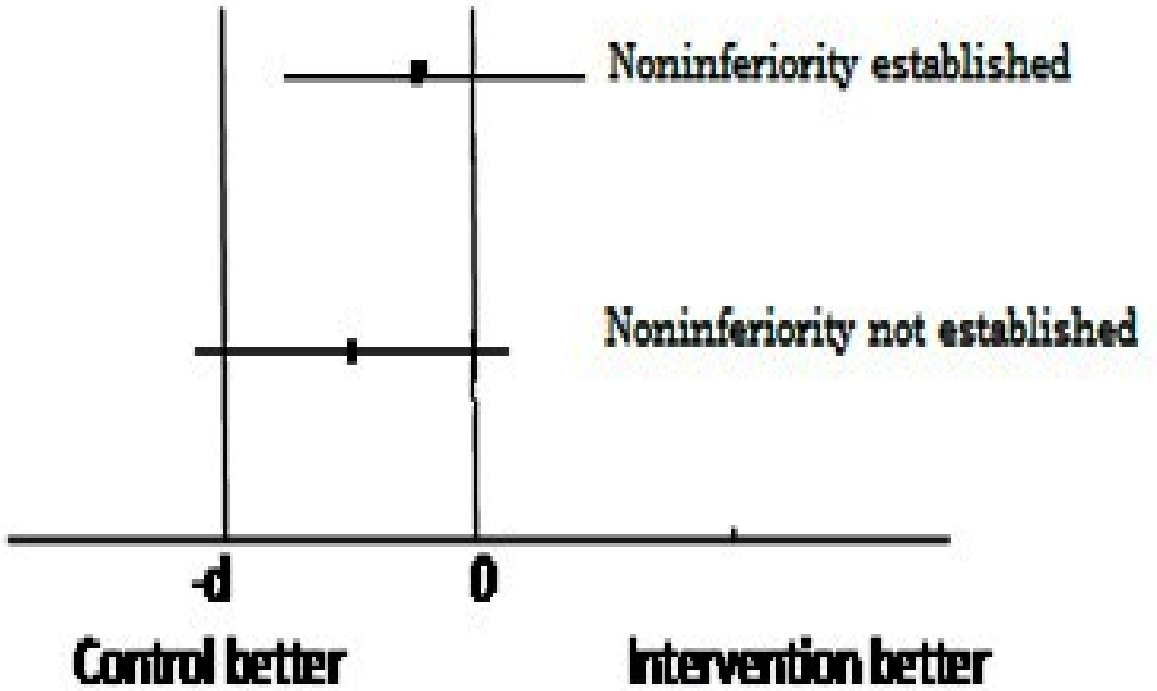
UNIVERSITY OF TORONTO

Superiority Design



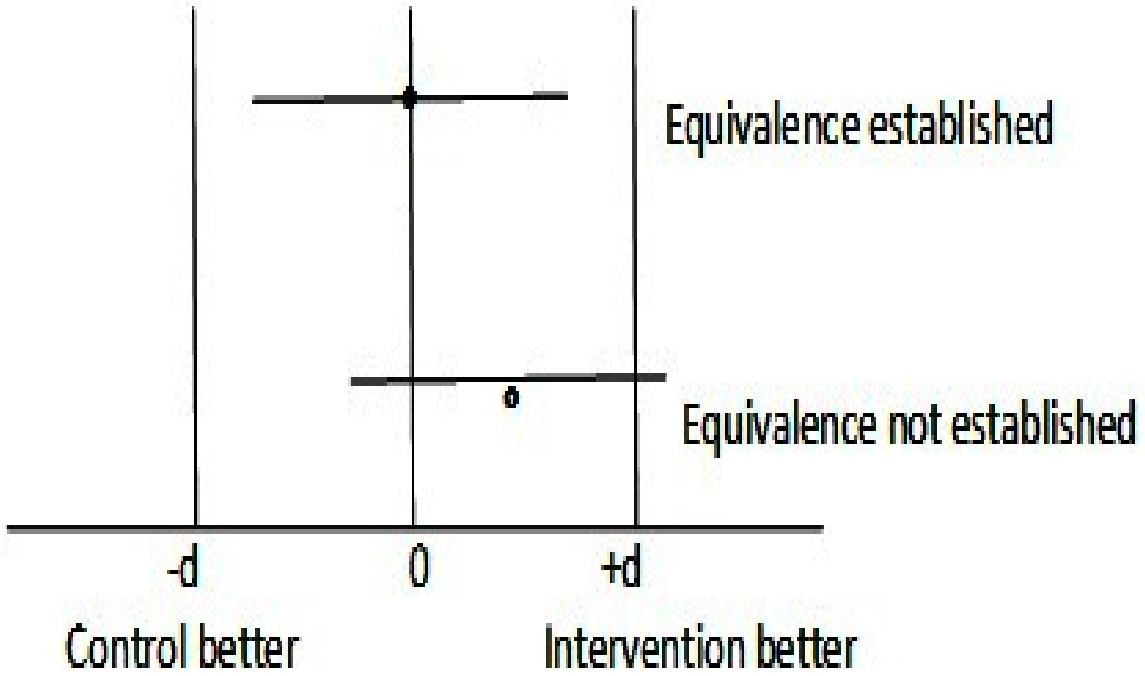
UNIVERSITY OF TORONTO

Non-Inferiority Design



UNIVERSITY OF TORONTO

Equivalence Design



UNIVERSITY OF TORONTO

Trials to assess equivalence: the importance of rigorous methods

BMJ 1996; 313:36-39 (6 July) **B Jones**, professor of medical statistics, [1] **P Jarvis**, senior lecturer in medical statistics, [1] **J A Lewis**, visiting professor in medical statistics, [1] **A F Ebbutt**, director of European clinical statistics [2]

The aim of an equivalence trial is to show the therapeutic equivalence of two treatments, usually a new drug under development and an existing drug for the same disease used as a standard active comparator. Unfortunately the principles that govern the design, conduct, and analysis of equivalence trials are not as well understood as they should be. Consequently such trials often include too few patients or have intrinsic design biases which tend towards the conclusion of no difference. In addition the application of hypothesis testing in analysing and interpreting data from such trials sometimes compounds the drawing of inappropriate conclusions, and the inclusion and exclusion of patients from analysis may be poorly managed.

UNIVERSITY OF TORONTO

A Multiplicity Problem in Hypothesis Testing (Multiple Comparisons)

- Hypothesis testing associated with risks of committing inferential errors:
  - Type-1 error
  - Type-2 error
- Family-wise error rate (FWER): risk at least one type-1 error in study.
- FWER increases as the number of hypotheses tested increases.
- Bonferroni adjustment control FWER at level α
  - Increased risk of a type-2 error.

Consider controlling FWER for m tests, using Bonferroni procedure:

$$\text{FWER} = P\left(\bigcup_{i=1}^{m_0}(p_i \leq \frac{\alpha}{m})\right) \leq \sum_{i=1}^{m_0}\left(P(p_i \leq \frac{\alpha}{m})\right) = m_0 * \frac{\alpha}{m} \leq m * \frac{\alpha}{m} = \alpha$$

UNIVERSITY OF TORONTO

Multiple Comparisons

Statistical Science
1991, Vol. 6, No. 1, 100–116

The Philosophy of Multiple Comparisons

John W. Tukey

Abstract. This paper is based on the 1989 Miller Memorial Lecture at Stanford University. The topic was chosen because of Rupert Miller's long involvement and significant contributions to multiple comparison procedures and theory. Our emphasis will be on the major questions that have received relatively little attention—on what one wants multiple comparisons to do, on why one wants to do that, and on how one can communicate the results. Very little attention will be given to how the results can be calculated—after all, there are books about that (e.g., Miller, 1966, 1981; Hochberg and Tamhane, 1987).

Key words and phrases: Confident directions, confidence directions, multiplicity, priced-out data, split multiplicity, recombining value splittings, studentized birange double differences.

UNIVERSITY OF TORONTO

Multiple Comparisons

No Adjustments Are Needed for Multiple Comparisons

Kenneth J. Rothman

Adjustments for making multiple comparisons in large bodies of data are recommended to avoid rejecting the null hypothesis too readily. Unfortunately, reducing the type I error for null associations increases the type II error for those associations that are not null. The theoretical basis for advocating a routine adjustment for multiple comparisons is the "universal null hypothesis" that "chance" serves as the first-order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations. A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not random numbers but actual observations on nature. Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings. (Epidemiology 1990;1:43–46)

Keywords: multiple comparisons, null hypothesis, significance testing, statistics.

UNIVERSITY OF TORONTO

Multiple Comparisons

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

UNIVERSITY OF TORONTO

Multiple Comparisons

Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and ‡Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved May 30, 2003 (received for review January 28, 2003)

With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in a genomewide data set are tested against some null hypothesis, where a number of features are expected to be significant. Here we propose an approach to measuring statistical significance in these genomewide studies based on the concept of the false discovery rate. This approach offers a sensible balance between the number of true and false positives that is automatically calibrated and easily interpreted. In doing so, a measure of statistical significance called the *q* value is associated with each tested feature. The *q* value is similar to the well known *p* value, except it is a measure of significance in terms of the false discovery rate rather than the false positive rate. Our approach avoids a flood of false positive results, while offering a more liberal criterion than what has been used in genome scans for linkage.

false discovery rates | genomics | multiple hypothesis testing | *q* values

to the method in ref. 5 under certain assumptions. Also, ideas similar to FDRs have appeared in the genetics literature (1, 13).

Similarly to the *p* value, the *q* value gives each feature its own individual measure of significance. Whereas the *p* value is a measure of significance in terms of the false positive rate, the *q* value is a measure in terms of the FDR. The false positive rate and FDR are often mistakenly equated, but their difference is actually very important. Given a rule for calling features significant, the false positive rate is the rate that truly null features are called significant. The FDR is the rate that significant features are truly null. For example, a false positive rate of 5% means that on average 5% of the truly null features in the study will be called significant. A FDR of 5% means that among all features called significant, 5% of these are truly null on average.

The *q* value provides a measure of each feature's significance, automatically taking into account the fact that thousands are simultaneously being tested. Suppose that features with *q* values ≤5% are called significant in some genomewide test of significance. This results in a FDR of 5% among the significant features. A

UNIVERSITY OF TORONTO

Post-Hoc Power Calculations

The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis

John M. HOENIG and Dennis M. HEISEY

It is well known that statistical power calculations can be valuable in planning an experiment. There is also a large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result. Advocates of such post-experiment power calculations claim the calculations should be used to aid in the interpretation of the experimental results. This approach, which appears in various forms, is fundamentally flawed. We document that the problem is extensive and present arguments to demonstrate the flaw in the logic.

KEY WORDS: Bioequivalence testing; Burden of proof; Observed power; Retrospective power analysis; Statistical power; Type II error.

illustrate the flawed approach (e.g., Rosner 1990; Winer, Brown, and Michels 1991; Zar 1996). Researchers need to be made aware of the shortcomings of power calculations as data analytic tools and taught more appropriate methodology.

It is important to understand the motivation of applied scientists for using power analysis to interpret hypothesis tests with nonsignificant results. The traditional, widely accepted standard has been to protect the investigator from falsely concluding that some treatment has an effect when indeed it has none. However, there is increasing recognition that a "reversal of the usual scientific burden of proof" (e.g., Dayton 1998) is preferred in many areas of scientific inference. Areas where this is a particular concern include making decisions about environmental impacts, product safety, and public welfare where some people want to be protected from failing to reject a null hypothesis of no impact when a serious (e.g., harmful or dangerous) effect exists. We be-

UNIVERSITY OF TORONTO

Sample Size Calculation - Linear Regression

STATISTICS IN MEDICINE
Statist. Med. 17, 1623–1634 (1998)

A SIMPLE METHOD OF SAMPLE SIZE CALCULATION FOR LINEAR AND LOGISTIC REGRESSION

F. Y. HSIEH[1]*, DANIEL A. BLOCH[2] AND MICHAEL D. LARSEN[3]

[1] *CSPCC, Department of Veterans Affairs, Palo Alto Health Care System (151-K), Palo Alto, California 94304, U.S.A.*
[2] *Division of Biostatistics, Department of Health Research and Policy, Stanford University, Stanford, California 94305, U.S.A.*
[3] *Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.*

SUMMARY

A sample size calculation for logistic regression involves complicated formulae. This paper suggests use of sample size formulae for comparing means or for comparing proportions in order to calculate the required sample size for a simple logistic regression model. One can then adjust the required sample size for a multiple logistic regression model by a variance inflation factor. This method requires no assumption of low response probability in the logistic model as in a previous publication. One can similarly calculate the sample size for linear regression models. This paper also compares the accuracy of some existing sample-size software for logistic regression with computer power simulations. An example illustrates the methods. © 1998 John Wiley & Sons, Ltd.

UNIVERSITY OF TORONTO

Sample Size Calculation - Logistic Regression

J Clin Epidemiol Vol. 49, No. 12, pp. 1373–1379, 1996

ELSEVIER

S0895-4356/96/$15.00
PII S0895-4356(96)00236-3

A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis

Peter Peduzzi,[1,4,*] *John Concato*,[2,3] *Elizabeth Kemper*,[1,4] *Theodore R. Holford*,[4] *and Alvan R. Feinstein*[2,3,4]

[1]COOPERATIVE STUDIES PROGRAM COORDINATING CENTER AND THE [2]MEDICAL SERVICE, VETERANS AFFAIRS MEDICAL CENTER, WEST HAVEN CONNECTICUT 06516; AND THE DEPARTMENTS OF [3]MEDICINE (CLINICAL EPIDEMIOLOGY UNIT) AND [4]EPIDEMIOLOGY AND PUBLIC HEALTH, YALE UNIVERSITY SCHOOL OF MEDICINE, NEW HAVEN, CONNECTICUT 06510

ABSTRACT. We performed a Monte Carlo study to evaluate the effect of the number of events per variable (EPV) analyzed in logistic regression analysis. The simulations were based on data from a cardiac trial of 673 patients in which 252 deaths occurred and seven variables were cogent predictors of mortality; the number of events per predictive variable was (252/7=) 36 for the full sample. For the simulations, at values of EPV = 2, 5, 10, 15, 20, and 25, we randomly generated 500 samples of the 673 patients, chosen with replacement, according to a logistic model derived from the full sample. Simulation results for the regression coefficients for each variable in each group of 500 samples were compared for bias, precision, and significance testing against the results of the model fitted to the original sample.

For EPV values of 10 or greater, no major problems occurred. For EPV values less than 10, however, the regression coefficients were biased in both positive and negative directions; the large sample variance estimates from the logistic model both overestimated and underestimated the sample variance of the regression coefficients; the 90% confidence limits about the estimated values did not have proper coverage; the Wald statistic was conservative under the null hypothesis; and paradoxical associations (significance in the wrong direction) were increased. Although other factors (such as the total number of events, or sample size) may influence the validity of the logistic model, our findings indicate that low EPV can lead to major problems. *Copyright © 1996 Elsevier Science Inc.* J CLIN EPIDEMIOL 49;12:1373–1379, 1996.

KEY WORDS. Monte Carlo, bias, precision, significance testing

UNIVERSITY OF TORONTO

Sample Size Calculation - Logistic Regression

American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Vol. 165, No. 6
DOI: 10.1093/aje/kwk052
Advance Access publication December 20, 2006

Original Contribution

Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression

Eric Vittinghoff and Charles E. McCulloch

From the Department of Epidemiology and Biostatistics, University of California, San Francisco, CA.

Received for publication March 15, 2006; accepted for publication August 15, 2006.

The rule of thumb that logistic and Cox models should be used with a minimum of 10 outcome events per predictor variable (EPV), based on two simulation studies, may be too conservative. The authors conducted a large simulation study of other influences on confidence interval coverage, type I error, relative bias, and other model performance measures. They found a range of circumstances in which coverage and bias were within acceptable levels despite less than 10 EPV, as well as other factors that were as influential as or more influential than EPV. They conclude that this rule can be relaxed, in particular for sensitivity analyses undertaken to demonstrate adequate control of confounding.

bias (epidemiology); coverage probability; event history analysis; model adequacy; type I error; variable selection

Abbreviation: EPV; events per predictor variable.

UNIVERSITY OF TORONTO

Sample Size Calculation - Logistic Regression

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 76 (2016) 175–182

Journal of Clinical Epidemiology

Adequate sample size for developing prediction models is not simply related to events per variable

Emmanuel O. Ogundimu*, Douglas G. Altman, Gary S. Collins

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Diseases, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, UK

Accepted 29 February 2016; Published online 8 March 2016

Abstract

Objectives: The choice of an adequate sample size for a Cox regression analysis is generally based on the rule of thumb derived from simulation studies of a minimum of 10 events per variable (EPV). One simulation study suggested scenarios in which the 10 EPV rule can be relaxed. The effect of a range of binary predictors with varying prevalence, reflecting clinical practice, has not yet been fully investigated.

Study Design and Setting: We conducted an extended resampling study using a large general-practice data set, comprising over 2 million anonymized patient records, to examine the EPV requirements for prediction models with low-prevalence binary predictors developed using Cox regression. The performance of the models was then evaluated using an independent external validation data set. We investigated both fully specified models and models derived using variable selection.

Results: Our results indicated that an EPV rule of thumb should be data driven and that EPV ≥ 20 generally eliminates bias in regression coefficients when many low-prevalence predictors are included in a Cox model.

Conclusion: Higher EPV is needed when low-prevalence predictors are present in a model to eliminate bias in regression coefficients and improve predictive accuracy.

Keywords: Events per variable; Cox model; External validation; Predictive modeling; Sample size; Resampling study

UNIVERSITY OF TORONTO

Sample Size Calculation - Survival and Time to Event Outcomes

Introduction to Sample Size Determination and Power Analysis for Clinical Trials

John M. Lachin

From the Biostatistics Center, George Washington University, Bethesda, Maryland

ABSTRACT: The importance of sample size evaluation in clinical trials is reviewed and a general method is presented from which specific equations are derived for sample size determination or the analysis of power for a wide variety of statistical procedures. The method is discussed and illustrated in relation to the t test, tests for proportions, tests of survival time, and tests for correlations as they commonly occur in clinical trials. Most of the specific equations reduce to a simple general form for which tables are presented.

KEY WORDS: *sample size determination, statistical power, survival analysis, tests for correlations, tests for proportions, t tests*

UNIVERSITY OF TORONTO

Sample Size Calculation - Survival and Time to Event Outcomes

BIOMETRICS 42, 507–519
September 1986

Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification

John M. Lachin

The Biostatistics Center, Department of Statistics/Computer & Information Systems, George Washington University, 7979 Old Georgetown Road, Bethesda, Maryland 20814, U.S.A.

and

Mary A. Foulkes

Biometry and Field Studies Branch, National Institute of Neurological and Communicative Disorders and Strokes, Bethesda, Maryland 20892, U.S.A.

SUMMARY

When designing a clinical trial to test the equality of survival distributions for two treatment groups, the usual assumptions are exponential survival, uniform patient entry, full compliance, and censoring only administratively at the end of the trial. Various authors have presented methods for estimation of sample size or power under these assumptions, some of which allow for an R-year accrual period with T total years of study, $T > R$. The method of Lachin (1981, *Controlled Clinical Trials* **2**, 93–113) is extended to allow for cases where patients enter the trial in a nonuniform manner over time, patients may exit from the trial due to loss to follow-up (other than administrative), other patients may continue follow-up although failing to comply with the treatment regimen, and a stratified analysis may be planned according to one or more prognostic covariates.

UNIVERSITY OF TORONTO

Sample Size Calculation - Survival and Time to Event Outcomes

BIOMETRICS 39, 499–503
June 1983

Sample-Size Formula for the Proportional-Hazards Regression Model

David A. Schoenfeld

Sidney Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, U.S.A.

SUMMARY

A formula is derived for determining the number of observations necessary to test the equality of two survival distributions when concomitant information is incorporated. This formula should be useful in designing clinical trials with a heterogeneous patient population. Schoenfeld (1981, *Biometrika* **68**, 316–319) derived the asymptotic power of a class of statistics used to test the equality of two survival distributions. That result is extended to the case where concomitant information is available for each individual and where the proportional-hazards model holds. The loss of efficiency caused by ignoring concomitant variables is also computed.

UNIVERSITY OF TORONTO

Sample Size Calculation - Reliability Studies

STATISTICS IN MEDICINE, VOL. 17, 101–110 (1998)

SAMPLE SIZE AND OPTIMAL DESIGNS FOR RELIABILITY STUDIES

S. D. WALTER,[1]* M. ELIASZIW[2,3] AND A. DONNER[3]

[1] *Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5*
[2] *The John P. Robarts Research Institute, London, Ontario, Canada N6A 5K8*
[3] *Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1*

SUMMARY

A method is developed to calculate the required number of subjects k in a reliability study, where reliability is measured using the intraclass correlation ρ. The method is based on a functional approximation to earlier exact results. The approximation is shown to have excellent agreement with the exact results and one can use it easily without intensive numerical computation. Optimal design configurations are also discussed; for reliability values of about 40 per cent or higher, use of two or three observations per subject will minimize the total number of observations required. © 1998 John Wiley & Sons, Ltd.

Statist. Med., **17**, 101–110 (1998)

UNIVERSITY OF TORONTO

Sample Size Calculation - Reliability Studies

Statistical Methods in Medical Research 2004; **13**: 251–271

Sample size requirements for the design of reliability study: review and new results

MM Shoukri Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada and Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Centre, Riyadh, Kingdom of Saudi Arabia, **MH Asyali** Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Centre, Riyadh, Kingdom of Saudi Arabia and **A Donner** Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada

The reliability of continuous or binary outcome measures is usually assessed by estimation of the intraclass correlation coefficient (ICC). A crucial step for this purpose is the determination of the required sample size. In this review, we discuss the contributions made in this regard and derive the optimal allocation for the number of subjects k and the number of repeated measurements n that minimize the variance of the estimated ICC. Cost constraints are discussed for both normally and non-normally distributed responses, with emphasis on the case of dichotomous assessments. Tables showing optimal choices of k and n are given along with the guidelines for the efficient design of reliability studies.

UNIVERSITY OF TORONTO

Sample Size Calculation - 2-Rater Agreement Studies

PSYCHOMETRIKA—VOL. 53, NO. 3, 321–325
SEPTEMBER 1988

SAMPLE SIZE DETERMINATIONS FOR THE TWO RATER KAPPA STATISTIC

V. F. FLACK, A. A. AFIFI, P. A. LACHENBRUCH

DIVISION OF BIOSTATISTICS
U.C.L.A. SCHOOL OF PUBLIC HEALTH

H. J. A. SCHOUTEN

DEPARTMENT OF MEDICAL INFORMATICS AND STATISTICS
UNIVERSITY OF LIMBURG

This paper gives a method for determining a sample size that will achieve a prespecified bound on confidence interval width for the interrater agreement measure, κ. The same results can be used when a prespecified power is desired for testing hypotheses about the value of kappa. An example from the literature is used to illustrate the methods proposed here.

Key words: interrater agreement, kappa statistic.

UNIVERSITY OF TORONTO

Sample Size Calculation - Studies with a ROC Outcome

STATISTICS IN MEDICINE, VOL. 16, 1529–1542 (1997)

SAMPLE SIZE DETERMINATION FOR DIAGNOSTIC ACCURACY STUDIES INVOLVING BINORMAL ROC CURVE INDICES

NANCY A. OBUCHOWSKI[1,*] AND DONNA KATZMAN McCLISH[2]

[1]*Department of Biostatistics and Epidemiology, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195-5196, U.S.A.*

[2]*Department of Biostatistics, Virginia Commonwealth University, 1101 E. Marshall St., Sanger Hall B1-066, Richmond, VA 23298-0032, U.S.A.*

SUMMARY

Receiver operating characteristic (ROC) curves and their associated indices are valuable tools for the assessment of the accuracy of diagnostic tests. The area under the ROC curve is a popular summary measure of the accuracy of a test. The full area under the ROC curve, however, has been criticized because it gives equal weight to all false positive error rates. Alternative indices include the area under the ROC curve in a particular range of false positive rates ('partial' area) and the sensitivity of the test for a single fixed false positive rate (FPR). We present a unified approach for computing sample size for binormal ROC curves and their indices. Our method uses Taylor series expansions to derive approximate large-sample estimates of the variance and covariance of binormal ROC curve parameters. Several examples from diagnostic radiology illustrate the proposed method.

UNIVERSITY OF TORONTO

Sample Size Calculation - Cluster Randomized Controlled Trial

© International Epidemiological Association 1999 Printed in Great Britain

International Journal of Epidemiology 1999;**28**:319–326

Simple sample size calculation for cluster-randomized trials

RJ Hayes and S Bennett

Background	Cluster-randomized trials, in which health interventions are allocated randomly to intact clusters or communities rather than to individual subjects, are increasingly being used to evaluate disease control strategies both in industrialized and in developing countries. Sample size computations for such trials need to take into account between-cluster variation, but field epidemiologists find it difficult to obtain simple guidance on such procedures.
Methods	In this paper, we provide simple formulae for sample size determination for both unmatched and pair-matched trials. Outcomes considered include rates per person-year, proportions and means. For simplicity, formulae are expressed in terms of the coefficient of variation (SD/mean) of cluster rates, proportions or means. Guidance is also given on the estimation of this value, with or without the use of prior data on between-cluster variation.
Case studies	The methods are illustrated using two case studies: an unmatched trial of the impact of impregnated bednets on child mortality in Kenya, and a pair-matched trial of improved sexually-transmitted disease (STD) treatment services for HIV prevention in Tanzania.
Keywords	Sample size, randomized controlled trials, cluster randomization, community randomization, between-cluster variation
Accepted	17 July 1998

UNIVERSITY OF TORONTO

Sample Size Calculation - Cluster Randomized Controlled Trial

Psychological Methods
1997, Vol. 2, No. 2, 173–185

1082-989X/97/$3.00

Statistical Analysis and Optimal Design for Cluster Randomized Trials

Stephen W. Raudenbush
Michigan State University

In many intervention studies, therapy outcome evaluations, and educational field trials, random treatment assignment of clusters rather than persons is desirable for political feasibility, logistics, or ecological validity. However, cluster randomized designs are widely regarded as lacking statistical precision. This article considers when and to what extent using a pretreatment covariate can increase experimental precision. To answer this question, the author first optimizes allocation of resources within and between clusters for the no-covariate case. Optimal sample sizes at each level depend on variation within and between clusters and on the cost of sampling at each level. Next, the author considers optimal allocation when a covariate is added. In this case, the explanatory power of the covariate at each level becomes highly relevant for choosing optimal sample sizes. A key conclusion is that statistical analysis that fully uses information about the covariate–outcome relationship can substantially increase the efficiency of the cluster randomized trial, especially when the cost of sampling clusters is high and the covariate accounts for substantial variation between clusters. Recent multilevel studies indicate that these conditions are common.

UNIVERSITY OF TORONTO

Sample Size Calculation - Multilevel Modelling Studies

Sufficient Sample Sizes for Multilevel Modeling

Cora J. M. Maas and Joop J. Hox

Utrecht University, The Netherlands

Abstract. An important problem in multilevel modeling is what constitutes a sufficient sample size for accurate estimation. In multilevel analysis, the major restriction is often the higher-level sample size. In this paper, a simulation study is used to determine the influence of different sample sizes at the group level on the accuracy of the estimates (regression coefficients and variances) and their standard errors. In addition, the influence of other factors, such as the lowest-level sample size and different variance distributions between the levels (different intraclass correlations), is examined. The results show that only a small sample size at level two (meaning a sample of 50 or less) leads to biased estimates of the second-level standard errors. In all of the other simulated conditions the estimates of the regression coefficients, the variance components, and the standard errors are unbiased and accurate.

Keywords: multilevel modeling, hierarchical linear model, sample size, cluster sampling

UNIVERSITY OF TORONTO

Design Strategies for Reducing Variability

- Consider measurement scale of response (and it's distribution).
- Consider precision/variability in measurement instrument.
- More efficient designs: paired, repeated measures, matching, etc.
- Consider unequal group sizes.

Control of Variability: Efficiency, increased power, reduced sample sizes...

UNIVERSITY OF TORONTO

Sample Size Estimation is an Important Aspect of Study Design

- **Goal**: Estimate appropriate number of participants for your study.
  - Ensures study has enough power to answer question of interest.
  - Ensures subjects not unnecessarily exposed to risk.
  - Makes study easier to conduct (not enrolling too many people).

Sample Size Calculations in Clinical Research; Chow, Shao & Wang (2003)

To provide an accurate and reliable sample size calculation, an appropriate statistical procedure for the research objective of interest must be derived under the study design.

Sample size calculation is a challenging task.

- Conduct early in study planning; allow for potential design modifications.
- Try to get input from clinical scientists, epidemiologists, and statisticians.

UNIVERSITY OF TORONTO