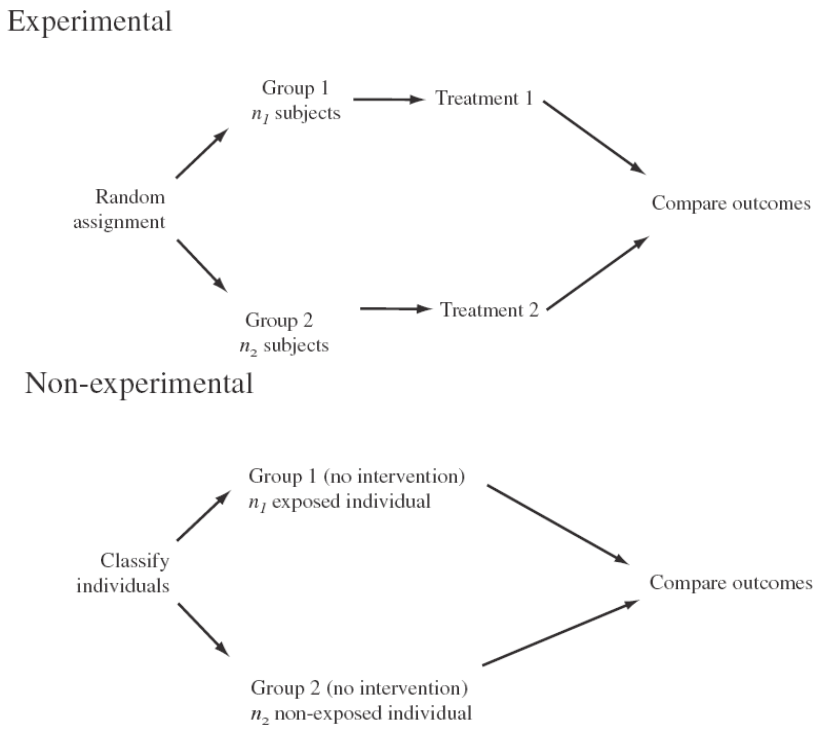
Formulas: Basic Biostat

PART I: GENERAL CONCEPTS AND TECHNIQUES

Descriptive and Exploratory Methods

Data may be regarded as quantitative, ordinal, or categorical. Ordinal data may be analyzed as quantitative or categorical, depending on whether certain distributional conditions are met.

Survey data should be selected in such a way as to allow for generalization to the population (probability sampling, notably based on SRSs).



Data should be explored with plot (e.g., stemplot, boxplot, histogram) whenever possible.

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} \qquad \text{[root mean sum of squares]}$$

Median: Arrange data from low to high. The median has a depth of $\frac{n+1}{2}$.

Quartiles (Tukey's hinges): Spit data in half at the median. When *n* is odd, the median is included in both the low group and high group. The "median" of the low group is Q1 and the "median" of the high group is Q3.

Five-point summary: Q0 (minimum), Q1, Q2 (median), Q3, Q4 (maximum)

IQR = Q3 − Q1

Fences (do not plot): $\text{Fence}_{\text{Lower}}$ = Q1 − (1.5)IQR; $\text{Fence}_{\text{Upper}}$ = Q3 + (1.5)IQR

Probability

Basic properties of probability: (1) $0 \le \Pr(A) \le 1$; (2) $\Pr(S) = 1$; (3) $\Pr(\bar{A}) = 1 - \Pr(A)$; (4) Pr(A or B) = Pr(A) + Pr(B) for disjoint events.

More advanced properties of probability (optional in some courses): (5) Independence rule (6) General rule of addition (7) Conditional probability definition (8) General rule of multiplication (9) Total probability rule (10) Bayes' theorem

Binomial random variables: $X \sim b(n,p)$, $\Pr(X = x) = {}_nC_x p^x q^{n-x}$ where ${}_nC_x = \frac{n!}{x!(n-x)!}$

Normal random variables. To find a Normal probability: (1) State (2) Standardize $z = \frac{x-\mu}{\sigma}$ (3) Sketch (4) Use Table B. To find a value from a Normal distribution: (1) State (2) Use Table B to look up z_p (3) Sketch (4) Unstandardize: $x = \mu + z_p\sigma$

Basics of Inference

The sampling distribution of $\bar{x}$ has mean μ (unbiasedness) and standard deviation $\sigma/\sqrt{n}$ (square root law). The distribution is Normal distribution when the population distribution is Normal and when the sample is large (Central Limit Theorem).

Hypothesis testing: A. Hypothesis statements B. Test statistic C. *P*-value D. Significance level (optional).

The standard deviation of the sample mean $\bar{x}$ may be referred to as the standard error of the mean. When σ is known $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

*F*or testing H_0: $\mu = \mu_0$ (σ known, SRS, Normal population or large sample): $z_{stat} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$.

(1 − α)100% confidence for $\mu = \bar{x} \pm z_{1-\alpha/2} \cdot SE_{\bar{x}}$

PART II: QUANTITATIVE RESPONSE VARIABLE – formulas pending

PART III CATEGORICAL RESPONSE VARIABLE

One-sample situation

- The numerator of sample proportion $\hat{p}$ is the observed number of successes; the denominator is the number of independent Bernoulli trials *n*. Thus, the random number of successes X ~ b(*n*, *p*).
- To test H_0: $p = p_0$ (large sample, SRS), $z_{stat} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$. Optional continuity-correction $z_{stat} = \frac{|\hat{p} - p_0| - \frac{1}{2n}}{\sqrt{p_0 q_0 / n}}$. Use exact binomial procedure in small samples. (In large samples, the continuity-corrected z statistic provides results comparable to the exact binomial procedure and the regular z statistic provides results comparable to exact mid-P results.)
- Plus-four CI for *p* (SRS and $n \geq 10$): $\tilde{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\tilde{p}\tilde{q}/\tilde{n}}$ where $\tilde{p} = \frac{x+2}{n+4}$.
- To limit the margin of error to *m*, use $n = \frac{z^2_{1-\frac{\alpha}{2}} p^* q^*}{m^2}$.
- The determinants of power when testing H_0: $p = p_0$ are p_0, p_1, *n*, and α. Conditions may be plugged into the power or sample size formulas on p. 368. Use of a software utility such as WinPepi's Describe.exe is encouraged to allow for different assumptions.

Two-sample situation

- To test H_0: $p_1 = p_2$ (SRS, large sample) $z_{\text{stat}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ or chi-square test (below). An optional continuity-correction may be incorporated.
- Risk difference is $\hat{p}_1 - \hat{p}_2$; confidence interval for $p_1 - p_2 = (\tilde{p}_1 - \tilde{p}_2) \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\tilde{p}_1 - \tilde{p}_2}$ where $SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{\tilde{p}_1\tilde{q}_1}{\tilde{n}_1} + \frac{\tilde{p}_2\tilde{q}_2}{\tilde{n}_2}}$.
- Relative risk $\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$. Confidence interval for the $RR = e^{\ln \hat{RR} \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\ln \hat{RR}}}$ where $SE_{\ln \hat{RR}} = \sqrt{\frac{1}{a_1} - \frac{1}{n_1} + \frac{1}{a_2} - \frac{1}{n_2}}$.
- The determinants of power when testing H_0: $p_1 = p_2$ are p_1, p_2 (or their ratio p_2/p_1), n_1 and n_2 (or their ratio *r*), and α. Conditions are plugged into power or sample size formulas on p. 396 – 40. Use of a software utility (e.g., the one included in WinPepi's Compare2.exe) is encouraged to allow for trying various assumptions.

R-by-C tables

- Descriptive proportions depend on whether the sample is naturalistic, cohort, or case-control.
- For naturalistic and cohort samples, let group 1 represents the least exposed group. Relative risks are calculated by comparing each group to the least exposed group: $\hat{R}R_i = \frac{\hat{p}_i}{\hat{p}_1} = \frac{a_i / n_i}{a_1 / n_1}$. Odds ratios are calculated $\hat{O}R_i = \frac{\hat{o}_i}{\hat{o}_1} = \frac{a_i / b_i}{a_1 / b_1}$.
- Systematic sources of error may overwhelm random sources of error in observational research.
- Chi-square test of association $X^2_{\text{stat}} = \sum_{\text{all}} \left[\frac{(O_i - E_i)^2}{E_i} \right]$ where

  $E_i = \frac{\text{row total} \times \text{column total}}{\text{table total}}$. With continuity-correction,

  $X^2_{\text{stat,c}} = \sum \left[\frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i} \right]$. Use exact procedure in small samples.