Basic Biostatistics by B. Burt Gerstman
Summary Points and Objectives

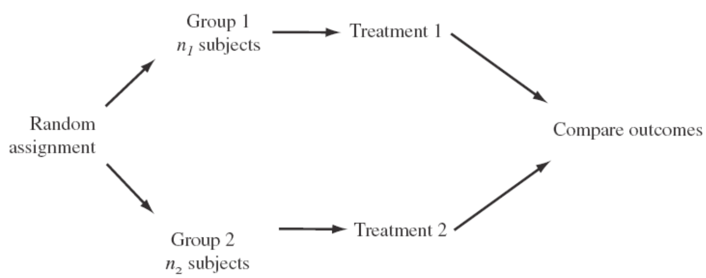
Chapter 1: Measurement

- Biostatistics is more than a compilation of computational techniques!
- Identify the main types of measurement scales: quantitative, ordinal, and categorical.
- Understand the layout of a data table (observations, variables, values)
- Appreciate the essential nature of data quality (GIGO principle).

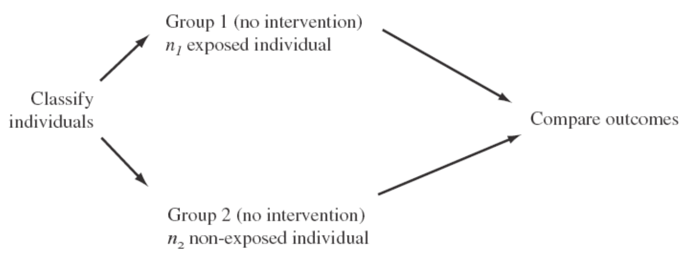
Chapter 2: Types of Studies

- Understand the difference between experimental and non-experimental ("observational") designs

Experimental



Non-experimental



- Understand the procedure for a simple random sample
- Understand the procedure for randomizing a treatment
- Define "confounding" and "lurking variable"
- List preconditions for confounding

Chapter 3: Frequency Distributions

- Create and interpret stemplots
- Describe distributional shape, location, and spread; check for outliers
- Create frequency tables containing frequency, relative frequency, cumulative frequency using uniform or non-uniform class intervals

Chapter 4: Summary Statistics

- Appreciate that great care must be taken in *interpreting* and *reporting* statistics!
- Sample mean: $\bar{x} = \frac{1}{n}\sum x_i$
- Median: Form an ordered array. The median is the value with a depth of $\frac{n+1}{2}$; when n is odd, average the two middle values.

- Quartiles (Tukey's hinges): Divide the ordered array at the median; when *n* is odd, the median belongs to both the low group and the high group. Q1 is median of the low group. Q3 is the median of the high group.
- Five-point summary: minimum, Q1, median, Q3, maximum
- IQR = Q3 − Q1
- Boxplot: plot median and quartiles (box); determine upper and lower fences: F_L = Q1 − 1.5·IQR, F_U = Q3 + 1.5·IQR; plot outside values; draw whiskers from hinges to inside values
- Understand the strengths and limitations of the mean, median, and mode
- Sample variance: $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$
- Sample standard deviation: $s = \sqrt{s^2}$; direct formula $s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$
- Select descriptive statistics suitable for distributional shape

Chapters 5: Probability Concepts

- Understand and use in practice these basics rules for probabilities:
  (1) 0 ≤ Pr(A) ≤ 1
  (2) Pr(S) = 1
  (3) Pr(Ā) = 1 − Pr(A)
  (4) Pr(A or B) = Pr(A) + Pr(B) for disjoint events
- Use probability mass function (*pmf*s) to find probabilities for discrete random variables
- Use probability density function *pdf*s to find probabilities for continuous random variables
- *Optional*: Understand the more advances rules for probabilities: (5) Independence rule (6) General rule of addition (7) Conditional probability definition (8) General rule of multiplication (9) Total probability rule (10) Bayes' theorem

Chapter 6: Binomial Distributions

- Identify a binomial random variable and its parameters: X~b(n,p)
- Calculate and interpret binomial probabilities: $\Pr(X = x) = {}_nC_x\, p^x q^{n-x}$ where ${}_nC_x = \frac{n!}{x!(n-x)!}$
- Calculate and interpret expected values (mean) and standard deviation for binomial random variables: $\mu = np$ and $\sigma = \sqrt{npq}$ where $q = 1 - p$.

Chapter 7: Normal Distributions

- Characterize and sketch, Normal distributions with parameters μ and σ: $X \sim N(\mu, \sigma)$
- Use the 68–95–99.7 rule to determine approximate probabilities for Normal random variables
- Characterize and sketch Standard Normal random variable Z ~ N(0,1); and understanding Table B
- Finding Normal probabilities (1) State (2) Standardize $z = \frac{x-\mu}{\sigma}$ (3) Sketch (4) Table B
- Finding percentile values on a Normal distribution: (1) State (2) Sketch (3) Table B (4) Unstandardize: $x = \mu + z_p\sigma$

Chapter 8: Introduction to Statistical Inference

- Define statistical inference; list the two primary forms of statistical inference
- Distinguish parameters from statistics!
- Understand the method of simulating a sampling distribution of a mean
- Characterize the sampling distribution of $\bar{x}$ from a Normal population: $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Understand the standard error of $\bar{x}$ in relation to *the square root law*: $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- Appreciate that *the central limit theorem* assures $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ when the sample size is moderate to large
- Know that *the law of large numbers* assures that $\bar{x}$ approaches μ as the sample gets large

Chapter 9: Basics of Hypothesis Testing

- Appreciate that hypothesis testing looks for evidence against the claim of H_0 and understand the meaning of *each* step of the procedure:
  Step A. H_0 and H_a
  Step B. Test statistic
  Step C. *P*-value
  Step D. *Optional*: Significance level
- See how hypothesis testing relates to the sampling distribution of $\bar{x}$
- Conduct one sample tests of means when σ is known:
  Conditions: SRS, Normal population or moderate to large sample size.
  (A.) H_0: $\mu = \mu_0$ (B.) $z = \dfrac{\bar{x} - \mu_0}{SE_{\bar{x}}}$ where $SE_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ (C.) *P*-value and interpretation
- Define: type I error; type II error; beta, power
- Determine the power and sample size requirements of a test (these objective are covered / reviewed under the Chapter 11 objectives)

Chapter 10: Basics of Confidence Intervals

- Appreciate how a confidence interval seek to locate a *parameter* with given margin of error
- See how confidence intervals estimation relates to the sampling distribution of $\bar{x}$
- Calculate and interpret confidence intervals for μ at various levels of confidence when σ is known:
  Conditions: SRS, Normal population or moderate to large sample size.
  Formula: $\bar{x} \pm z_{1-\alpha/2} \cdot SE_{\bar{x}}$ where $SE_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
- Determine sample size requirements for estimating μ with given level of confidence and margin of error (see Chap 11 for formula)
- Understand the relationship between confidence interval location and hypothesis testing

PART II: QUANTITATIVE RESPONSE VARIABLE

Chapter 11: Inference about a Mean

- Quantitative response variable, no explanatory variable *per se* (single sample or paired samples)
- Understand when to use *t* procedures
- Sketch *t* distributions; use Table C to look up *t* values and associated probabilities
- Conduct one-sample and paired-sample *t* tests (conditions: SRS, population Normal or large sample):
  (A.) H_0: $\mu = \mu_0$ (B.) $t_{stat} = \dfrac{\bar{x} - \mu_0}{SE_{\bar{x}}}$ where $SE_{\bar{x}} = \dfrac{s}{\sqrt{n}}$ with $n - 1$ *df* C. *P*-value and interpretation
- Calculate and interpret one-sample and paired-sample confidence interval for μ:
  Formula: $\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot SE_{\bar{x}}$
- Recognize paired samples and adapt the one-sample *t* procedures to paired samples
- Evaluate the Normality assumption in small, medium, and large samples
- Conduct sample size and power analyses:
  - to limit margin of error *m* when estimating μ, use $n = \left(z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{m}\right)^2$

- to detect a difference of Δ with stated power and α, use $n = \dfrac{\sigma^2\left(z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)^2}{\Delta^2}$
- to determine the power of a test to detect Δ , $1-\beta = \Phi\left(-z_{1-\frac{\alpha}{2}} + \dfrac{|\Delta|\sqrt{n}}{\sigma}\right)$

Chapter 12: Comparing Independent Means

- Quantitative response variable, binary explanatory variable (two independent samples)
- Compare group means, standard deviations, sample sizes
- Compare group distributions graphically (e.g., side-by-side boxplots, side-by-side stemplots)
- Conduct independent *t* test: (conditions: independent samples and Normality or large samples)

  (A.) H_0: $\mu_1 = \mu_2$ (B.) $t_{\textbf{stat}} = \dfrac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1-\bar{x}_2}}$ where $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ with $df_{\text{conservative}}$ = smaller of $(n_1 - 1)$ or $(n_2 - 1)$ [use df_{Welch} when working with a computer] (C.) *P*-value and interpretation
- Calculate and interpret $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$:
  Formula: $(\bar{x}_1 - \bar{x}_2) \pm (t_{df,1-\frac{\alpha}{2}})(SE_{\bar{x}_1-\bar{x}_2})$
- *Optional*: Be aware and understand the historical relevance of equal variance ("pooled") *t* procedures where $SE = \sqrt{s_{\text{pooled}}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ where $s_{\text{pooled}}^2 = \dfrac{df_1 \cdot s_1^2 + df_2 \cdot s_2^2}{df_1 + df_2}$ and $df = (n_1 - 1) + (n_2 - 1)$
- Power and sample size
  - To estimate $\mu_1 - \mu_2$ with margin of error *m*, use $n = \dfrac{2\sigma^2 z_{1-\frac{\alpha}{2}}^2}{m^2}$ in each group
  - To test H_0: $\mu_1 = \mu_2$ to detect Δ at given $(1-\beta)$ and α: use $n = \dfrac{2\sigma^2\left(z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)^2}{\Delta^2}$ in each group
  - If it is not possible to study groups of equal size, then determine *n* by the above formulas, fix the size of n_1, and have $n_2 = \dfrac{nn_1}{2n_1 - n}$.

Chapter 13: ANOVA

- Quantitative response variable, categorical explanatory variable (*k* independent samples)
- Always start with descriptive and exploratory comparisons!
- ANOVA test (conditions: independent samples, normality, equal variance)
  (A.) H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ versus H_a: at least two of the population means differ
  (B.) F_{stat} with df_B and df_W from ANOVA table
  (C.) *P*-value and interpretation

Variance	Sum of Squares	df	Mean Square
Between groups	$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$df_B = k - 1$	$MSB = \dfrac{SS_B}{df_B}$
Within groups	$SS_W = \sum_{i=1}^{k} (n_i - 1)s_i^2$	$df_W = N - k$	$MSW = \dfrac{SS_W}{df_W}$
Total	$SS_T = SS_B + SS_W$	$df = df_B + df_W$	

$F_{\text{stat}} = \dfrac{MSB}{MSW}$ with df_B and df_W

- Use post-hoc procedures such as the *least squares difference method* to delineate significant differences (A.) H_0: $\mu_i = \mu_j$ for groups i and j (B.) $t_{\text{stat}} = \dfrac{\bar{x}_i - \bar{x}_j}{SE_{\bar{x}_i - \bar{x}_j}}$ where

  $SE_{\bar{x}_i - \bar{x}_j} = \sqrt{MSW\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}$ and $df = N - k$ (C.) P-value and interpretation
- Recognize the problem of multiple comparisons and use Bonferroni method to keep the the family-wise error rate in check (when appropriate): $P_{\text{Bonf}} = P_{\text{LSD}} \times c$ where c represents the number of post hoc comparisons made.
- Assess the equal variance assumption graphically, by comparing group standard deviations, and with Levene's test of H_0: $\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k$.
- Use robust non-parametric ANOVA (i.e., the Kruskal-Wallis test) when necessary.

Chapter 14: Correlation and Regression

- Quantitative explanatory variable; quantitative response variable
- Linear relations only!
- Start with a scatterplot. Describe form, direction, and strength. Also check for outliers.
- Correlation does not necessarily indicate causation; beware of lurking variables.
- Correlation coefficient r is always between −1 and 1; it quantifies the direction (positive/negative) and strength of an association. As rules of thumb: $|r| < 0.3$ suggests weak strength and $|r| > 0.7$ suggests strong strength ("grain of salt" no firm cutoffs, and best used merely as a screening tool).

  Formula: $r = \dfrac{1}{n-1}\sum z_X z_Y$

  [Use calculator or software tool to check calculations.]
- Inferences about population correlation coefficient ρ:

  To test H_0: $\rho = 0$, use $t_{\text{stat}} = \dfrac{r}{SE_r}$ where $SE_r = \sqrt{\dfrac{1-r^2}{n-2}}$ and $df = n - 2$

  Confidence interval for ρ: $LCL = \dfrac{r - \varpi}{1 - r\varpi}$ and $UCL = \dfrac{r + \varpi}{1 + r\varpi}$ where $\varpi = \sqrt{\dfrac{t^2_{df,1-\frac{\alpha}{2}}}{t^2_{df,1-\frac{\alpha}{2}} + df}}$
- Least squares regression model: $\hat{y} = a + bx$ where $b = r\dfrac{s_Y}{s_X}$ and $a = \bar{y} - b\bar{x}$.
- Slope estimate b is the key statistic in all this, representing the predicted change in Y per unit X.
- Inference about population slope β:

  Standard error of the regression $s_{Y|x} = \sqrt{\dfrac{1}{n-2}\sum \text{residuals}^2}$ with $df = n - 2$

  $(1 - \alpha)100\%$ confidence interval for $\beta = b \pm (t_{n-2,1-\alpha/2})(SE_b)$ where $SE_b = \dfrac{s_{Y|x}}{\sqrt{n-1} \cdot s_X}$

  To test H_0: $\beta = 0$, use $t_{\text{stat}} = \dfrac{b}{SE_b}$

  Optional: An ANOVA procedure can be used to test H_0: $\beta = 0$ using an F_{stat} (pp. 321–324)

Chapter 15: Multiple Regression

- Multiple regression is an extension of simple regression; students should master simple regression before moving on to multiple regression.
- The quantitative response variable Y depends on multiple explanatory variables $X_1, X_2, \ldots X_k$ via this model: $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$.

- Categorical explanatory variables can be entered into the model if coded with indicator "dummy" variables.
- The computer uses a least squares criterion to fit a regression surface by minimizing $\sum$residuals^2.
- The key statistics are the slope estimates, b_is, representing predicted changes in Y per unit X_i, adjusting for the other explanatory variables in the model.
- Interpret confidence intervals for each β_i
- Interpret t tests for each H_0: $\beta_i = 0$.
- Residuals are examined to assess linearity, independence, normality, equal variance.
- *Optional* analysis of variance derives:

	Sum of Squares	df	Mean Square
Regression	$\sum(\hat{y}_i - \bar{y})^2$	k	$\frac{\text{SS regression}}{\text{df regression}}$
Residual "error"	$\sum(y_i - \hat{y}_i)^2$	$n - k - 1$	$\frac{\text{SS residual}}{\text{df residual}}$
Total	$\sum(y_i - \bar{y})^2$	$n - 1$	

$F_{\text{stat}} = \frac{\text{MS regression}}{\text{MS residual}}$ with k and $n - k - 1$ *df*s

Model fit (of secondary concern) is quantified with $R^2 = \frac{\text{Sum of Squares Regression}}{\text{Sum of Squares Total}}$.

PART III CATEGORICAL RESPONSE VARIABLE

Chapter 16: Inference about a Proportion

- Single sample; binary outcome.
- Sample proportion $\hat{p}$ is viewed in the context of a binomial numerator (x) and constant denominator (n); inference are directed toward binomial parameter p
- $\hat{p}$ represents incidence or prevalences, depending how data are accrued
- Hypothesis test (large samples)
  (A.) H_0: $p = p_0$ (B.) $z_{stat} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$ (C.) P-value and interpretation

  Optional continuity-correction $z_{stat,c} = \frac{|\hat{p} - p_0| - \frac{1}{2n}}{\sqrt{p_0 q_0 / n}}$
- Hypothesis test (small samples, e.g., less than 5 successes)
  (A.) H_0: $p = p_0$ (B.) Observed number of success (C.) P-value from "exact" binomial calculations (computer assisted) and interpretation
- The power of the hypothesis test depends on assumed values for p_0, p_1, n, and α (p. 368)
- $(1 - \alpha)100\%$ confidence intervals for p by "plus-four" method (similar to Wilson's):
  $\tilde{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\tilde{p}\tilde{q} / \tilde{n}}$ where $\tilde{p} = \frac{x + 2}{n + 4}$ and $\tilde{q} = 1 - \tilde{p}$
- With $n < 10$ use, use exact binomial procedure (computer) for confidence interval.
- To limit the margin of error (m) when estimating p, use $n = \frac{z_{1-\frac{\alpha}{2}}^2 p^* q^*}{m^2}$.

Chapter 17: Comparing Two Proportions

- Binary response variable, binary explanatory variable (two independent groups)

	Successes	Failures	Total
Group 1	a_1	b_1	n_1
Group 2	a_2	b_2	n_2
Total	m_1	m_2	N

- $\hat{p}_1 = \frac{a_1}{n_1}$ and $\hat{p}_2 = \frac{a_2}{n_2}$. Sample proportions $\hat{p}_1$ and $\hat{p}_2$ reflect underlying parameters p_1 and p_2.
- Hypothesis test, large samples:
  (A.) H_0: $p_1 = p_2$
  (B.) $z_{\text{stat}} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ (or chi-square, next chapter)
  (C.) P-value and interpretation
- Hypothesis test, small samples, use Fisher's test (computer assisted)
- Risk difference = $\hat{p}_1$ - $\hat{p}_2$; "excess risk in absolute terms associated with exposure"
  (1 – α)100% confidence interval for $p_1 - p_2$ by plus-four method:
  $(\tilde{p}_1 - \tilde{p}_2) \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\tilde{p}_1 - \tilde{p}_2}$ where $\tilde{p}_i = \dfrac{a_i + 1}{n_i + 2}$ and $SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\dfrac{\tilde{p}_1\tilde{q}_1}{\tilde{n}_1} + \dfrac{\tilde{p}_2\tilde{q}_2}{\tilde{n}_2}}$
- Relative risk $\hat{R}R = \dfrac{\hat{p}_1}{\hat{p}_2}$; "excess risk in relative terms associated with exposure"
  (1−α)100% CI for $RR = e^{\ln \hat{R}R \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\ln \hat{R}R}}$ where $SE_{\ln \hat{R}R} = \sqrt{\dfrac{1}{a_1} - \dfrac{1}{n_1} + \dfrac{1}{a_2} - \dfrac{1}{n_2}}$
- Systematic sources of error due to selection bias, information bias, and confounding!
- The power of testing H_0: $p_1 = p_2$ depends on p_1, p_2, n_1 and n_2, and α. Use software to calculate sample size and power; encourage students to think about underlying "inputs".

Chapter 18: Cross-Tabulated Counts

- Understand that data can come from naturalistic, cohort, or case-control samples.
- Cross-tabulate counts from categorical response variable (C columns) and categorical explanatory variable (R rows). Example of R-by-2 table:

	Successes	Failures	Total
Group 1	a_1	b_1	n_1
Group 2	a_2	b_2	n_2
↕	↕	↕	↕
Group R	a_R	b_R	n_R
Total	m_1	m_2	N

- In naturalistic and cohort samples, report incidence (or prevalences) in each group: $\hat{p}_i = \dfrac{a_i}{n_i}$.
- Characteristics of chi-square probability distributions (e.g., start at 0, asymmetrical, become increasingly symmetrical as the *df* increases)
- Hypothesis test for association (large samples)
  (A.) H_0: no association in population (homogeneity of proportions)

(B.) $X^2_{\text{stat}} = \sum_{\text{all}} \left[\frac{(O_i - E_i)^2}{E_i} \right]$ where $E_i = \frac{\text{row total} \times \text{column total}}{\text{table total}}$ with $df = (R - 1)(C - 1)$

(C.) *P*-value from chi-square table or program and interpretation

- Hypothesis test (small samples): use Fisher's procedure when more than 20% of expected frequencies are less than 5 or any expected frequency is less than 1.
- In naturalistic and cohort samples, use risk difference or risk ratio as measure of association.
- Hypothesis test for trend (ordinal explanatory or response variable)
  (A.) H_0: "no trend in population" (B.) Use program to calculate Mantel trend statistic (C.) *P*-value and interpretation
- Case-control sample: population cases and random sample of population non-cases → do *not* calculate incidence or prevalences. Calculate odds ratio as estimate of population rate ratio (equivalent to the risk ratio when the outcome is rare).

  $\hat{OR} = \frac{a_1 / b_2}{a_2 / b_1}$

  $(1 - \alpha)100\%$ CI for the $OR = e^{\ln \hat{OR} \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\ln \hat{OR}}}$ where $SE_{\ln \hat{OR}} = \sqrt{\frac{1}{a_1} + \frac{1}{b_1} + \frac{1}{a_2} + \frac{1}{b_2}}$
- Matched-pairs:

	Case E+	Case E−
Control E+	*a*	*b*
Control E−	*c*	*d*

$\hat{OR} = \frac{c}{b}$; $(1 - \alpha)100\%$ confidence interval for the $OR = e^{\ln \hat{OR} \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\ln \hat{OR}}}$ where $SE_{\ln \hat{OR}} = \sqrt{\frac{1}{c} + \frac{1}{b}}$

Hypothesis test: (A.) H_0: $OR = 1$ (B.) $z_{stat} = \sqrt{\frac{(c-b)^2}{c+b}}$ (C.) *P*-value and interpretation; use exact binomial procedure when there are 5 or less discordant pairs

Chapter 19: Stratified 2-by-2 Tables

- Methods to mitigate confounding: randomization, restriction, matching, regression, stratification
- Simpson's paradox is an extreme form of confounding in which the direction of association is reversed by the confounding factor
- Strata specific *RR*s are denoted with subscripts: RR_1, RR_2, …, RR_K
- See if strata-specific *RR*s provide the same "picture" as the crude *RR*. If not, this is evidence of confounding or interaction.
- Heterogeneous strata-specific *RR*s suggest statistical interaction.
- Chi-square test for interaction. Example considers *RR*s from two strata:
  (A.) H_0: $RR_1 = RR_2$ (no interaction) (B.) Chi-square interaction statistics (various forms) (C.) *P*-value and interpretation
- There are no statistical tests for confounding.
- If there is confounding and no interaction), Mantel-Haenszel procedures are applied to summarize the *RR*s and test the association (pp. 468 – 472).