Review

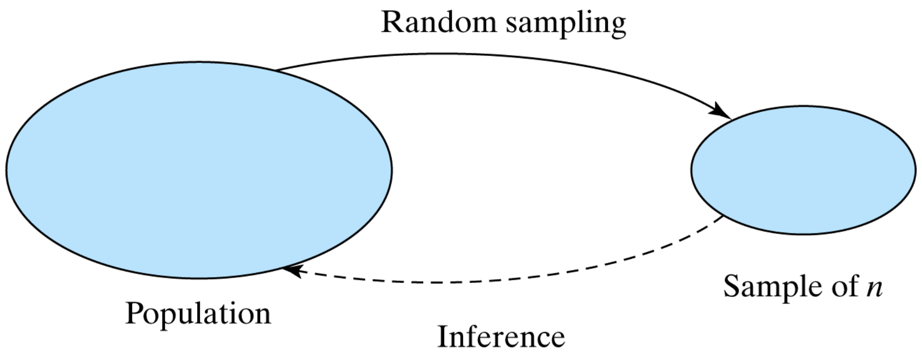
Chapter 1. Introduction

1.2 Types of Evidence

- Anecdotal evidence/observational study/experiment (blinding)

1.3 Random Sampling

- Population/sample;
- A simple random sample;
- Sampling/non-sampling error



Chapter 2. Description of Samples and Populations

2.1 Introduction

- Variable： categorical / numeric(continuous/discrete) variables

2.3 Descriptive Statistics: Measures of Center

- Median/mean

2.4 Boxplots

- Quartiles: Q1 , Q3 , interquartile range (IQR)
- Outliers: lower fence = Q1 - 1.5 x IQR, upper fence upper fence = Q_3 + 1.5 x IQR
- Boxplots for data without/with outliers

2.6 Measures of Dispersion

- Range
- The standard deviation s = $\sqrt{\frac{\sum_{i=1}^{n}(yi-\bar{y})^2}{n-1}}$
- Variance s^2

Chapter 3. Probability and Binomial Distribution

3.2 Introduction to Probability

- Probability trees

3.3 Probability Rules (optional)

- Basic Rules/Additional Rules/Multiplication Rules
- Conditional probability

3.4 Density Curves

- Interpretation of density curve

3.5 Random Variables

- Mean of a Random Variable: E(Y) = $\mu_Y = \sum y_i \Pr(Y = yi)$
- Variance of a Random Variable: $\sigma^2{}_Y = \sum (y_i - \mu_Y)^2 \Pr(Y = yi)$

Chapter 3. Probability and Binomial Distribution

3.6 The Binomial Distribution

- Binomial Random Variable
- The Binomial Distribution Formula

$$\Pr\{j\ successes\} = \Pr(Y = j\} = {}_nC_j p^j (1-p)^{n-j}$$

  - For a binomial random variable Y, the probability that the n trials result in j successes (and n - j failures) is given by the above formula.
  - Binomial coefficient ${}_n\boldsymbol{C}_j$ is given in Table 2.
- Mean of a Binomial: *np*
- Standard deviation of a Binomial: $\sqrt{np\ (1-p)}$

Chapter 4. The Normal Distribution

4.2 The Normal Curves

- The normal curve is a symmetric “bell-shaped” curve
- Y ~ N(μ, σ): a variable Y follows a normal distribution with mean μ and standard deviation σ.

4.3 Areas Under a Normal Curve

- Standardization Formula: $Z = (Y – \mu) / \sigma$; the Z scale is referred to as a standardized scale.
- The variable Z is referred to as the standard normal and its distribution follows a normal curve with mean = 0 and standard deviation = 1.
- Use Table 3 find areas under the standard normal curve, below a specified value of z.

4.4 Assessing Normality

- 68%-95%-99.7% rule (± 1/2/3 SD)
- Normal quantile plot

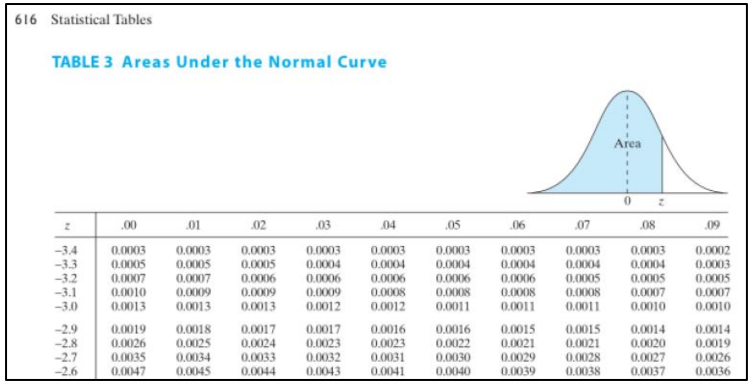

616 Statistical Tables

TABLE 3 Areas Under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
–3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
–3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
–3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
–3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
–2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
–2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
–2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
–2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036

Chapter 5. Sampling Distribution

5.2 The Sample Mean

- Theorem 5.2.1: The sampling distribution of $\bar{Y}$: "How close to μ is $\bar{Y}$ *likely* to be?"
  1. Mean: The mean of the sampling distribution of $\bar{Y}$ is equal to the population mean.
     - In symbols, $\mu_{\bar{Y}} = \mu$
  2. Standard deviation: The standard deviation of the sampling distribution of $\bar{Y}$ is equal to the population standard deviation divided by the square root of the sample size.
     - In symbols, $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$
  3. Shape
     1) If the population distribution of Y is normal, then the sampling distribution of $\bar{Y}$ is normal, regardless of the sample size n.

     2) **Central Limit Theorem**: If n is large, then the sampling distribution of $\bar{Y}$ is approximately normal, even if the population distribution of Y is not normal.

Chapter 6. Confidence Intervals

6.2 Standard Error of the Mean

- Standard error of the mean is defined as $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$
- SE is an estimate of $\sigma_{\bar{Y}}$.

6.3 Confidence Interval for μ

- Two-sided Confidence Interval
  - A 95% confidence interval for μ: $\bar{y} \pm t_{0.025} \times s/\sqrt{n}$
    - — s is sample SD
    - — $t_{0.025}$ is determined from Student's t distribution (Table 4) with df = n-1
    - — $t_{0.025}$ is called the "two-tailed 5% critical value" of Student's t distribution
  - We can be 95% confident that the (population) mean of XX is between A and B.
- One-sided Confidence Interval
  - A one-sided 95% (lower) confidence interval for μ: $\bar{y} - t_{0.05} \times s/\sqrt{n}$
  - A one-sided 95% (upper) confidence interval for μ: $\bar{y} + t_{0.05} \times s/\sqrt{n}$

Chapter 6. Confidence Intervals

6.6 Comparing Two Means

- Standard Error of ($\bar{Y}_1$ -$\bar{Y}_2$)： $SE_{(\bar{Y}1 - \bar{Y}2)} = \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}} = \sqrt{SE_1{}^2 + SE_2{}^2}$

6.7 Confidence Interval for (μ_1 - μ_2)

- A 95% confidence interval for μ_1 - μ_2： ($\bar{y}_1$ - $\bar{y}_2$) $\pm\, t_{0.025} \times SE_{(\bar{Y}1 - \bar{Y}2)}$
  - — The critical value $t_{0.025}$ is determined from Student’s t distribution (Table 4) using degrees of freedom* given as

$$\text{df} = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)}$$

Chapter 7. Hypothesis Testing

7.2 Hypothesis Testing: The t Test

- Formulate hypothesis：
  - Null hypothesis H_0: μ1 = μ2 → μ1 - μ2 = 0 - population 1 and 2 are the same
  - Alternative hypothesis H_A: μ1 ≠ μ2 → μ1 - μ2 ≠ 0 - population 1 and 2 are NOT the same
- Calculate P-value by test statistic t_s: $t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{(\bar{Y}_1 - \bar{Y}_2)}}$
- Select significance level α, and compare P-value with α to accept/not accept H_A
  - If the P-value ≤ α, accept H_A; H_0 is rejected.
  - If the P-value > α, NOT accept H_A; H_0 is NOT rejected.

7.3 Further Discussion of the t Test

- Type I and Type II errors

Table 7.3.2 Possible outcomes of testing H_0

		True situation	
		H_0 true	H_A true
OUR DECISION	Lack of significant evidence for H_A	Correct	Type II error
	Significant evidence for H_A	Type I error	Correct

Chapter 7. Hypothesis Testing

7.8 Summary of t Test Mechanics

t Test

$$H_0: \mu_1 = \mu_2$$
$$H_A: \mu_1 \neq \mu_2 \text{ (nondirectional)}$$
$$H_A: \mu_1 < \mu_2 \text{ (directional)}$$
$$H_A: \mu_1 > \mu_2 \text{ (directional)}$$

$$\text{Test statistic: } t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\mathrm{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

$$P\text{-value} = \text{tail area under Student's } t \text{ curve with}$$
$$\mathrm{df} = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1 - 1) + \mathrm{SE}_2^4/(n_2 - 1)}$$

Nondirectional H_A: P-value $=$ two-tailed area beyond t_s and $-t_s$

Directional H_A: Step 1. Check directionality.

Step 2. P-value $=$ single-tail area beyond t_s

Decision: Significant evidence for H_A if P-value $\leq \alpha$

Chapter 8. Comparison of Paired Samples

Summary of Formulas

Standard Error of $\overline{D}$

$$\mathrm{SE}_{\overline{D}} = \frac{s_D}{\sqrt{n_D}}$$

***t* Test**

$$H_0\colon \mu_D = 0$$

$$t_s = \frac{\bar{d} - 0}{\mathrm{SE}_{\overline{D}}}$$

95% Confidence Interval for μ_d

$$\bar{d} \pm t_{0.025}\mathrm{SE}_{\overline{D}}$$

Intervals with other confidence levels (e.g., 90%, 99%) are constructed analogously (e.g., using $t_{0.05}$, $t_{0.005}$).

Chapter 9. Categorical Data: One-Sample Distribution

9.1 Dichotomous Observations

- Dichotomous categorical variable: only two possible values

9.2 Confidence Interval for a Population Proportion

- Standard error of $\tilde{P}$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$, where $\tilde{p}$ = (y+2) / (n+4)

- 95% Confidence interval for p

$$\tilde{p} \pm t_{0.025}\, SE_{\tilde{p}} \quad \rightarrow \quad \tilde{p} \pm 1.96\, SE_{\tilde{p}}$$

- Planning a study to estimate p

Desired SE = $\sqrt{\frac{(Guessed\ \tilde{p})(1-Guessed\ \tilde{p})}{n+4}}$, where Guessed $\tilde{p}$ = (y+2) / (n+4)

Chapter 9. Categorical Data: One-Sample Distribution

9.4 The Chi-Square Goodness-of-Fit Test

Goodness-of-fit test

Data:

$$o_i = \text{the observed frequency of category } i$$

Null hypothesis:

H_0 specifies the probability of each category.*

Calculation of expected frequencies:

$$e_i = n \times \text{Probability specified for category } i \text{ by } H_0$$

Test statistic:

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Null distribution (approximate):

$$\chi^2 \text{ distribution with df} = k - 1$$

where k = the number of categories

This approximation is adequate if $e_i \geq 5$ for every category.

Chapter 10. Categorical Data: Relationships

Summary of Chi-Square Test for a Contingency Table

Null hypothesis:

H_0: Row variable and column variable are independent

Calculation of expected frequencies:

$$e_i = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Test statistic:

$$\chi_s^2 = \sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}$$

Null distribution (approximate):

$$\chi^2 \text{ distribution with df} = (r - 1)(k - 1)$$

where r is the number of rows and k is the number of columns in the contingency table. This approximation is adequate if $e_i \geq 5$ for every cell. If r and k are large, the condition that $e_i \geq 5$ is less critical and the χ^2 approximation is adequate if the average expected frequency is at least 5, and no expected frequency is less than 1.

The observations must be independent of one another. If paired data are collected for a 2×2 table, then McNemar's test is appropriate (Section 10.8).

Chapter 11. Comparing the Means of Many Independent Samples

11.2 The Basic One-Way Analysis of Variance

- The global null hypothesis is H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
  - H_0: all the population means are equal (no difference)
- The nondirectional alternative hypothesis H_A: The μi's are not all equal
  - H_A : at least one pair of the population means are NOT equal (differ)

11.4 The Global F Test

- The F statistic: $F_s = \dfrac{\text{MS(between)}}{\text{MS(within)}}$
- F distribution depends on two parameters
  - Numerator df = df(between)
  - Denominator df = df(within)
- Critical values for the F distribution are given in Table 10.

ANOVA Table

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between groups	$I - 1$	$\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Within groups	$n. - I$	$\sum_{i=1}^{I} (n_i - 1)s_i^2$	SS/df
Total	$n. - 1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	

— The total number of observations

$n. = \sum_{i=1}^{I} n_i$

— The grand mean

$\bar{\bar{y}} = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}}{n.}$

Chapter 11. Comparing the Means of Many Independent Samples

11.6 One-Way Randomized Blocks Design

- The global null hypothesis is H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
- The nondirectional alternative hypothesis H_A: The μi's are not all equal
- The F statistic: F_s = MS(treatments)/MS(within)

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between treatments	$I - 1$	$\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Between blocks	$J - 1$	$\sum_{j=1}^{J} m_j(\bar{y}_{\cdot j} - \bar{\bar{y}})^2$	SS/df
Within groups	$n. - I - J + 1$	SS(within) = SS(total) – SS(treatment) – SS(blocks)	SS/df
Total	$n. - 1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2$	

Chapter 11. Comparing the Means of Many Independent Samples

11.7 Two-Way ANOVA

- The global null hypothesis is H_0: $\gamma_{11} = \gamma_{12} = \ldots = \gamma_{IJ} = 0$
- The nondirectional alternative hypothesis H_A: The γ_{ij}'s are not all equal
- The F statistic: F_s = MS(**interaction**)/MS(within)

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between i treatments	$I - 1$	$\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Between j treatments	$J - 1$	$\sum_{j=1}^{J} m_j(\bar{y}_j - \bar{\bar{y}})^2$	SS/df
Interaction	$(I - 1) \times (J - 1)$		SS/df
Within groups	$n. - IJ$	SS(within) = SS(total) – SS(treatment) – SS(interaction)	SS/df
Total	$n. - 1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2$	

Chapter 12. Linear Regression and Correlation

12.2 The Correlation Coefficient

Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Fact 12.3.1: $$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Bivariate Random Sampling Model:

We regard each pair (x_i, y_i) as having been sampled at random from a population of (x, y) pairs.

- R testing the hypothesis H_0: ρ = 0 (population correlation ρ = 0)
  - H_0: There is no linear relationship between X and Y.
- A t test
  - the test statistic : $t_s = r\sqrt{\frac{n-2}{1-r^2}}$
  - Critical values are obtained from Student's t-distribution with df = n – 2

Chapter 12. Linear Regression and Correlation

12.3 The Fitted Regression Line

12.5 Statistical Inference Concerning β1

Fitted Regression Line

$$\hat{y} = b_0 + b_1 x$$

where

$$b_1 = r \times \left(\frac{s_y}{s_x}\right)$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Residuals:

$$y_i - \hat{y}_i \quad \text{where} \quad \hat{y}_i = b_0 + b_1 x_i$$

Residual Sum of Squares:

$$\text{SS(resid)} = \sum (y_i - \hat{y}_i)^2$$

Residual Standard Deviation:

$$s_e = \sqrt{\frac{\text{SS(resid)}}{n - 2}}$$

Inference

Standard Error of b_1:

$$\text{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

95% confidence interval for β_1:

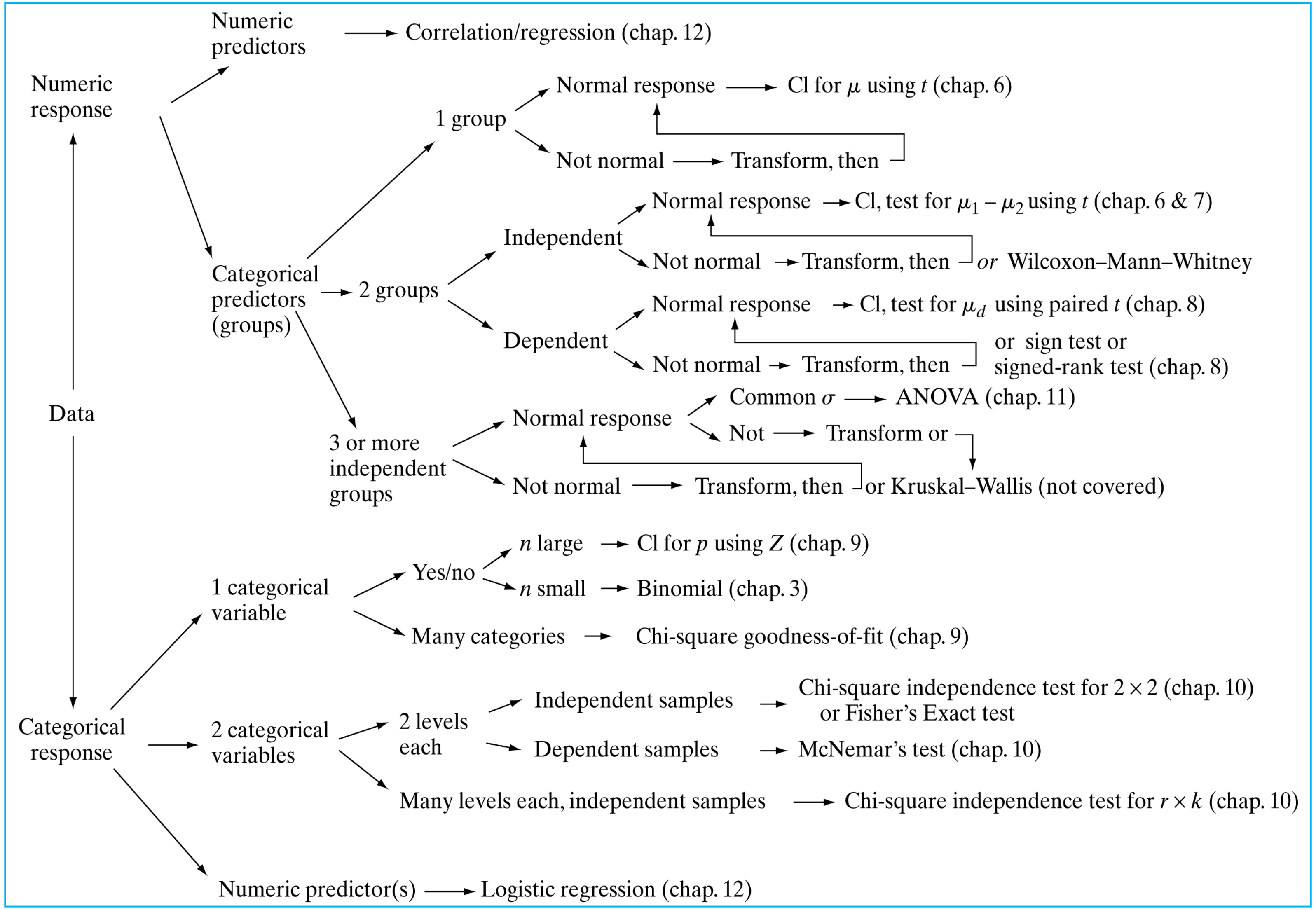
$$b_1 \pm t_{0.025}\text{SE}_{b_1}$$

Test of H_0: $\beta_1 = 0$ or H_0: $\rho = 0$:

$$t_s = \frac{b_1}{\text{SE}_{b_1}} = r\sqrt{\frac{n-2}{1-r^2}}$$

Critical values for the test and confidence interval are determined from Student's t distribution with df $= n - 2$.

Chapter 13. A Summary of Inference Methods



Thank you!