EE3211 Modelling Techniques

Final Review

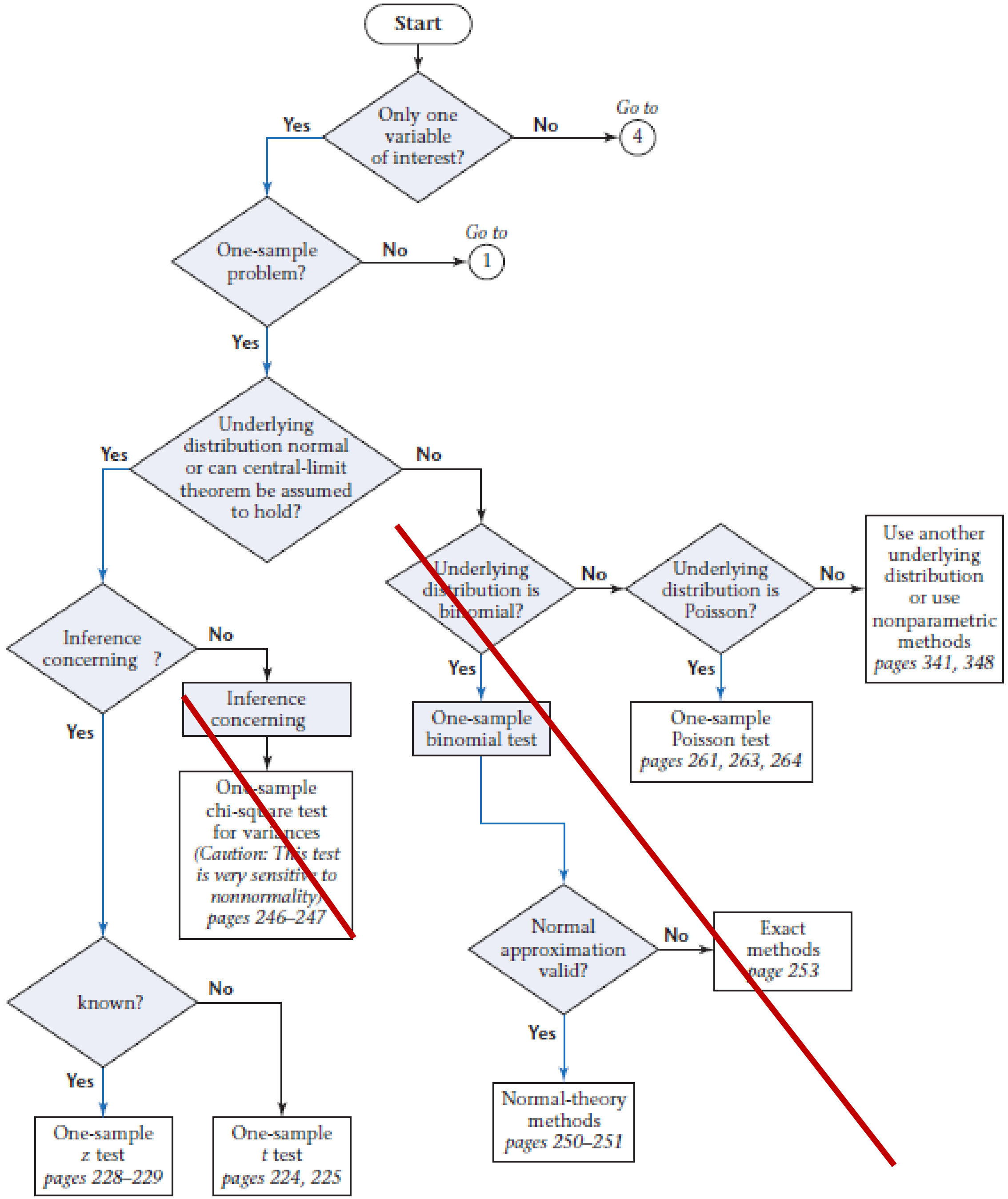
Notes for Project

- Deadline for submission: April 23rd, 2021
  - Choose one out of the three topics
  - Make sure to arrange submission earlier than the deadline
- Cite research paper (Endnote)

Exam

- 3 hours
- All topics covered in this course
- Open-book
- Format:
  - 30 questions: multiple choice, true/false, fill in blanks
  - 3 long questions (with sub questions)

CHOICE OF STATISTICAL TEST


Start
Only one variable of interest?
Yes
No
Go to
4
One-sample problem?
No
Go to
1
Yes
Underlying distribution normal or can central-limit theorem be assumed to hold?
Yes
No
Inference concerning ?
No
Inference concerning
One-sample chi-square test for variances (Caution: This test is very sensitive to nonnormality) pages 246–247
Yes
known?
No
Yes
One-sample z test pages 228–229
One-sample t test pages 224, 225
Underlying distribution is binomial?
No
Underlying distribution is Poisson?
No
Use another underlying distribution or use nonparametric methods pages 341, 348
Yes
One-sample binomial test
Yes
One-sample Poisson test pages 261, 263, 264
Normal approximation valid?
No
Exact methods page 253
Yes
Normal-theory methods pages 250–251

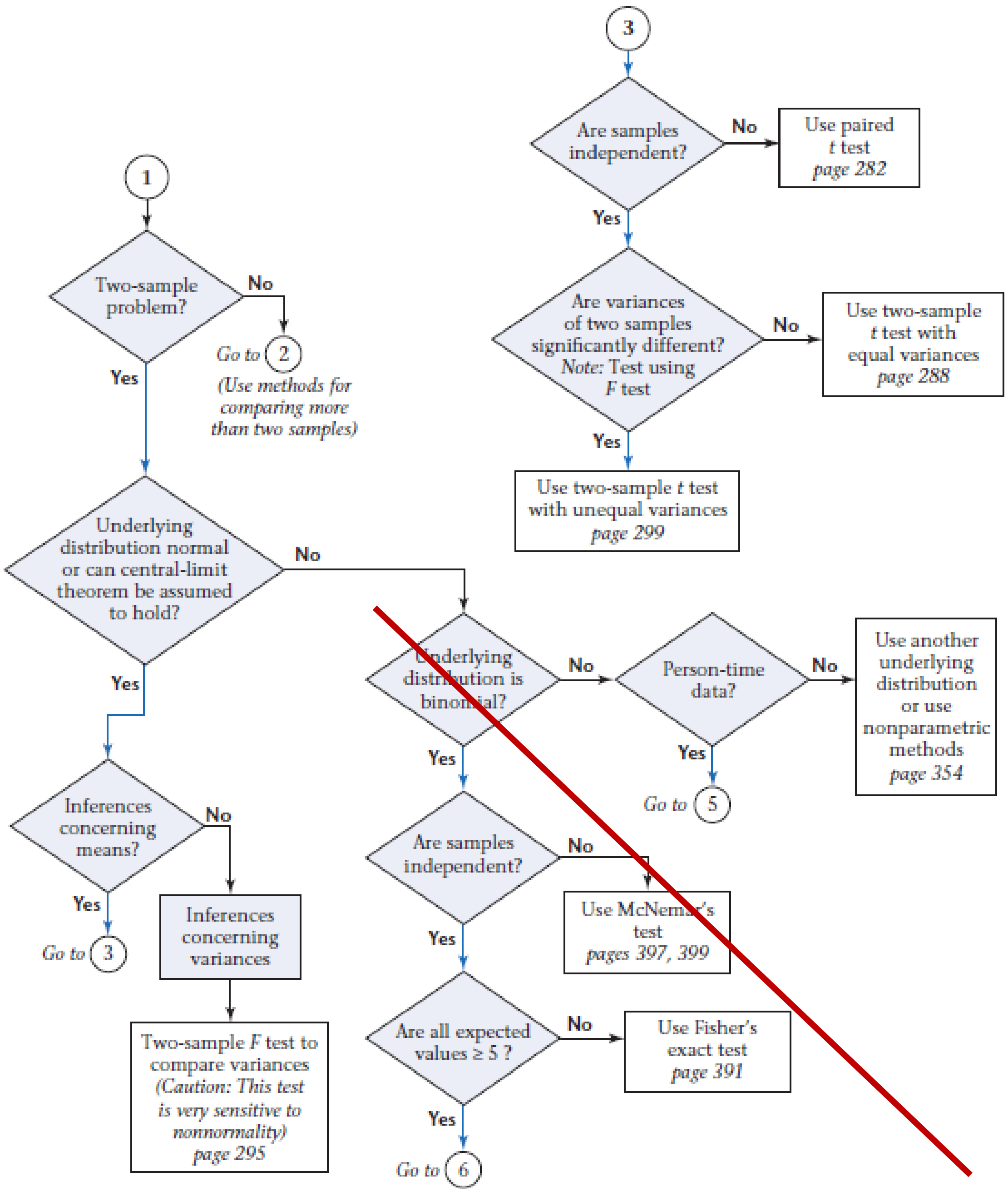
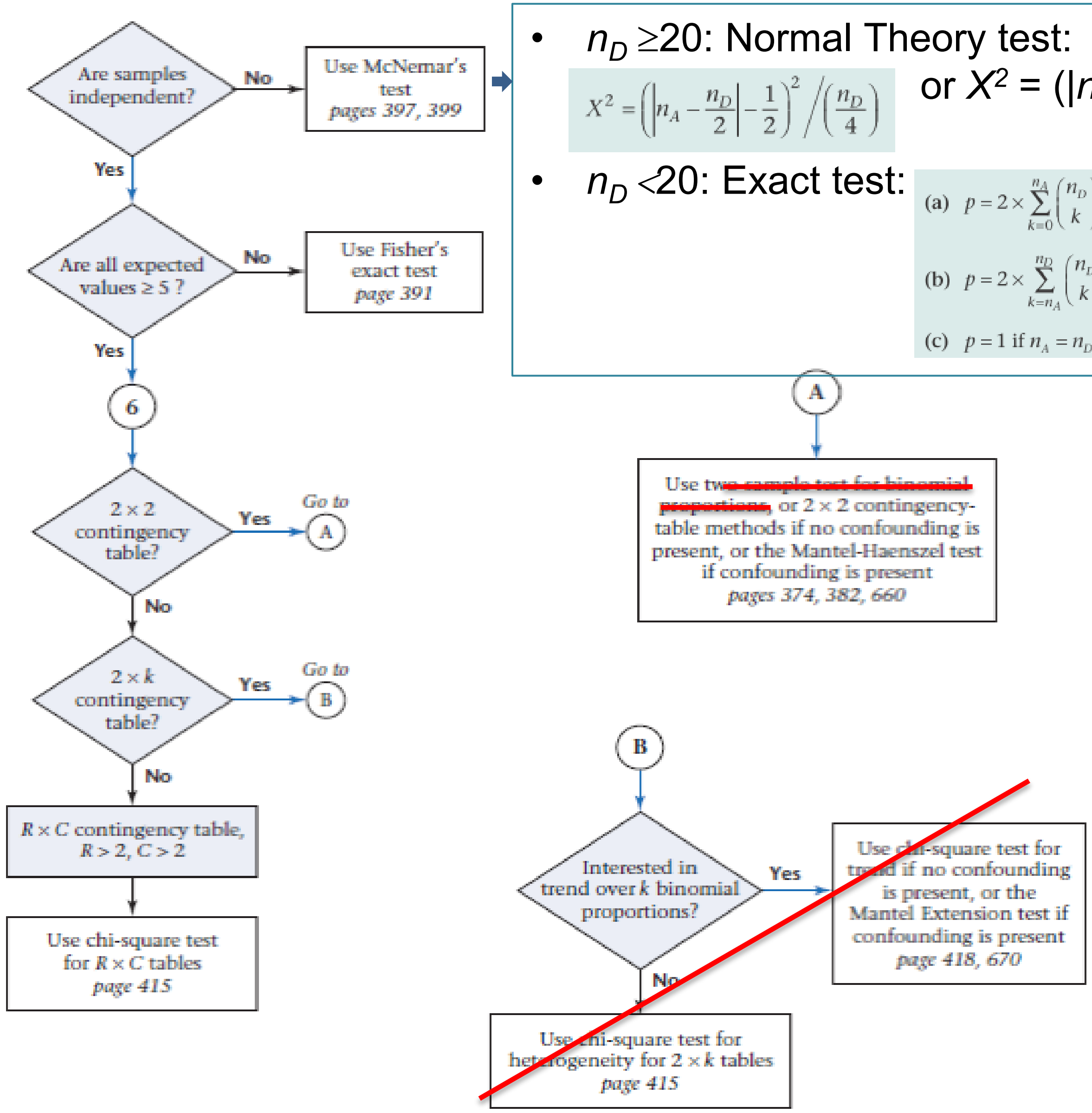

1
Two-sample problem?
No
Go to 2
(Use methods for comparing more than two samples)
Yes
Underlying distribution normal or can central-limit theorem be assumed to hold?
Yes
Inferences concerning means?
Yes
Go to 3
No
Inferences concerning variances
Two-sample F test to compare variances (Caution: This test is very sensitive to nonnormality) page 295
No
Underlying distribution is binomial?
Yes
Are samples independent?
No
Use McNemar's test pages 397, 399
Yes
Are all expected values ≥ 5 ?
No
Use Fisher's exact test page 391
Yes
Go to 6
No
Person-time data?
Yes
Go to 5
No
Use another underlying distribution or use nonparametric methods page 354
3
Are samples independent?
No
Use paired t test page 282
Yes
Are variances of two samples significantly different? Note: Test using F test
No
Use two-sample t test with equal variances page 288
Yes
Use two-sample t test with unequal variances page 299

FIGURE 10.16 Flowchart for appropriate methods of statistical inference for categorical data



- $n_D \geq 20$: Normal Theory test:

$$X^2 = \left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2 \Big/ \left(\frac{n_D}{4}\right)$$

or $X^2 = (|n_A - n_B| - 1)^2/(n_A + n_B)$

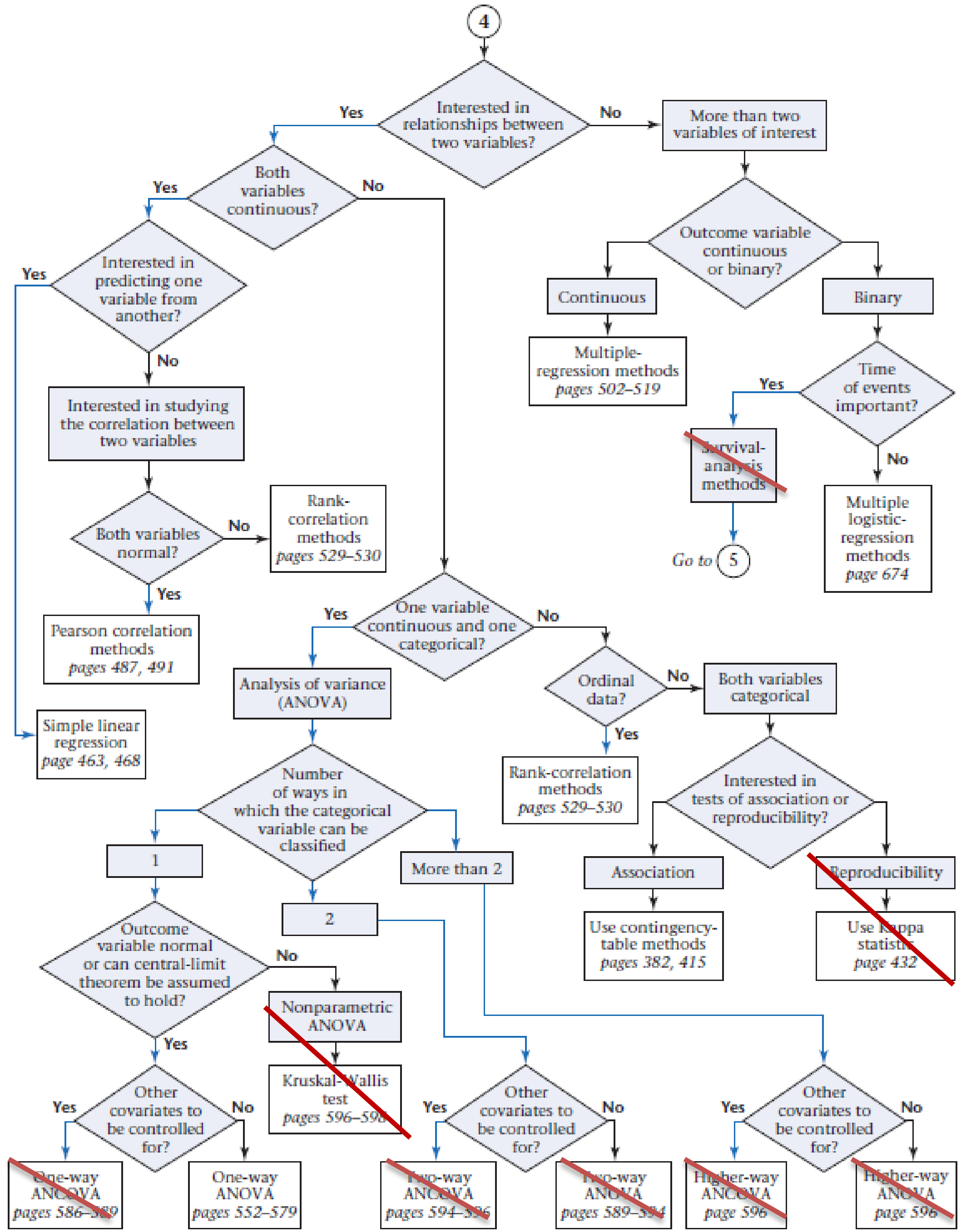
- $n_D < 20$: Exact test:

(a) $p = 2 \times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$ if $n_A < n_D/2$

(b) $p = 2 \times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$ if $n_A > n_D/2$

(c) $p = 1$ if $n_A = n_D/2$

- No more than 1/5 of the cells have expected values <5
- No cell has an expected value <1
- Continuity correction is not needed


4
Interested in relationships between two variables?
Yes
No
More than two variables of interest
Outcome variable continuous or binary?
Continuous
Multiple-regression methods pages 502–519
Binary
Time of events important?
Yes
Survival-analysis methods
Go to 5
No
Multiple logistic-regression methods page 674
Both variables continuous?
Yes
No
Interested in predicting one variable from another?
Yes
Simple linear regression page 463, 468
No
Interested in studying the correlation between two variables
Both variables normal?
No
Rank-correlation methods pages 529–530
Yes
Pearson correlation methods pages 487, 491
One variable continuous and one categorical?
Yes
Analysis of variance (ANOVA)
No
Ordinal data?
Yes
Rank-correlation methods pages 529–530
No
Both variables categorical
Interested in tests of association or reproducibility?
Association
Use contingency-table methods pages 382, 415
Reproducibility
Use Kappa statistic page 432
Number of ways in which the categorical variable can be classified
1
2
More than 2
Outcome variable normal or can central-limit theorem be assumed to hold?
No
Nonparametric ANOVA
Kruskal-Wallis test pages 596–598
Yes
Other covariates to be controlled for?
Yes
One-way ANCOVA pages 586–589
No
One-way ANOVA pages 552–579
Other covariates to be controlled for?
Yes
Two-way ANCOVA pages 594–596
No
Two-way ANOVA pages 589–594
Other covariates to be controlled for?
Yes
Higher-way ANCOVA page 596
No
Higher-way ANOVA page 596

Descriptive statistics

- Measures of location e.g. mean, median, mode
- Measures of spread e.g. standard deviation
- Graphic methods e.g. boxplot, stem-and-leaf plot
  - Symmetric distribution
  - Unsymmetric distribution (skewed to the right / left)

Probability distribution

- Discrete vs. continuous
- Measure of location
- Measure of spread
- Standardization of a normal variable
- Z-table

TABLE 3 The normal distribution

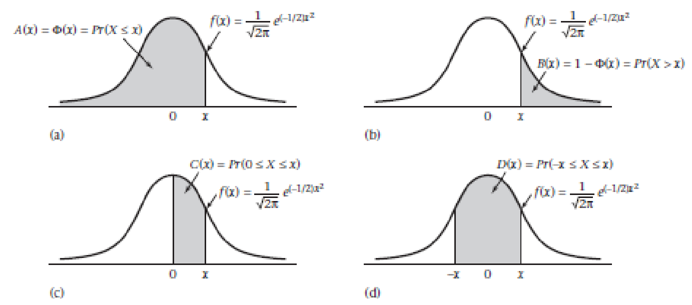


TABLE 3 The normal distribution (*continued*)

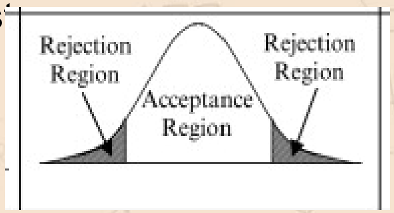
x	A[a]	B[b]	C[c]	D[d]
1.82	.9656	.0344	.4656	.9312
1.83	.9664	.0336	.4664	.9327
1.84	.9671	.0329	.4671	.9342
1.85	.9678	.0322	.4678	.9357
1.86	.9686	.0314	.4686	.9371
1.87	.9693	.0307	.4693	.9385
1.88	.9699	.0301	.4699	.9399
1.89	.9706	.0294	.4706	.9412
1.90	.9713	.0287	.4713	.9426
1.91	.9719	.0281	.4719	.9439
1.92	.9726	.0274	.4726	.9451
1.93	.9732	.0268	.4732	.9464
1.94	.9738	.0262	.4738	.9476
1.95	.9744	.0256	.4744	.9488
1.96	.9750	.0250	.4750	.9500
1.97	.9756	.0244	.4756	.9512
1.98	.9761	.0239	.4761	.9523
1.99	.9767	.0233	.4767	.9534
2.00	.9772	.0228	.4772	.9545
2.01	.9778	.0222	.4778	.9556
2.02	.9783	.0217	.4783	.9566
2.03	.9788	.0212	.4788	.9576
2.04	.9793	.0207	.4793	.9586
2.05	.9798	.0202	.4798	.9596
2.06	.9803	.0197	.4803	.9606
2.07	.9808	.0192	.4808	.9615
2.08	.9812	.0188	.4812	.9625
2.09	.9817	.0183	.4817	.9634
2.10	.9821	.0179	.4821	.9643
2.11	.9826	.0174	.4826	.9651
2.12	.9830	.0170	.4830	.9660
2.13	.9834	.0166	.4834	.9668
2.14	.9838	.0162	.4838	.9676
2.15	.9842	.0158	.4842	.9684
2.16	.9846	.0154	.4846	.9692
2.17	.9850	.0150	.4850	.9700
2.18	.9854	.0146	.4854	.9707
2.19	.9857	.0143	.4857	.9715
2.20	.9861	.0139	.4861	.9722
2.21	.9864	.0136	.4864	.9729
2.22	.9868	.0132	.4868	.9736
2.23	.9871	.0129	.4871	.9743
2.24	.9875	.0125	.4875	.9749
2.25	.9878	.0122	.4878	.9756
2.26	.9881	.0119	.4881	.9762
2.27	.9884	.0116	.4884	.9768
2.28	.9887	.0113	.4887	.9774
2.29	.9890	.0110	.4890	.9780
2.30	.9893	.0107	.4893	.9786
2.31	.9896	.0104	.4896	.9791
2.32	.9898	.0102	.4898	.9797
2.33	.9901	.0099	.4901	.9802
2.34	.9904	.0096	.4904	.9807
2.35	.9906	.0094	.4906	.9812
2.36	.9909	.0091	.4909	.9817
2.37	.9911	.0089	.4911	.9822
2.38	.9913	.0087	.4913	.9827

Hypothesis testing: one and two sample inference

- Type 1 error
- Type 2 error

Your Decision Based On a Random Sample		Given the Null Hypothesis Is: True	False
	Reject	Type I Error	Correct Decision
	Do Not Reject	Correct Decision	Type II Error

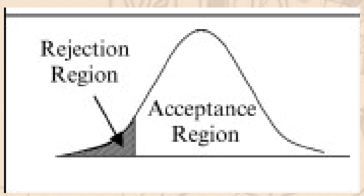
- Two-sided test

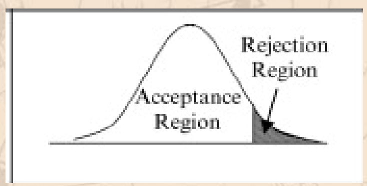


$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

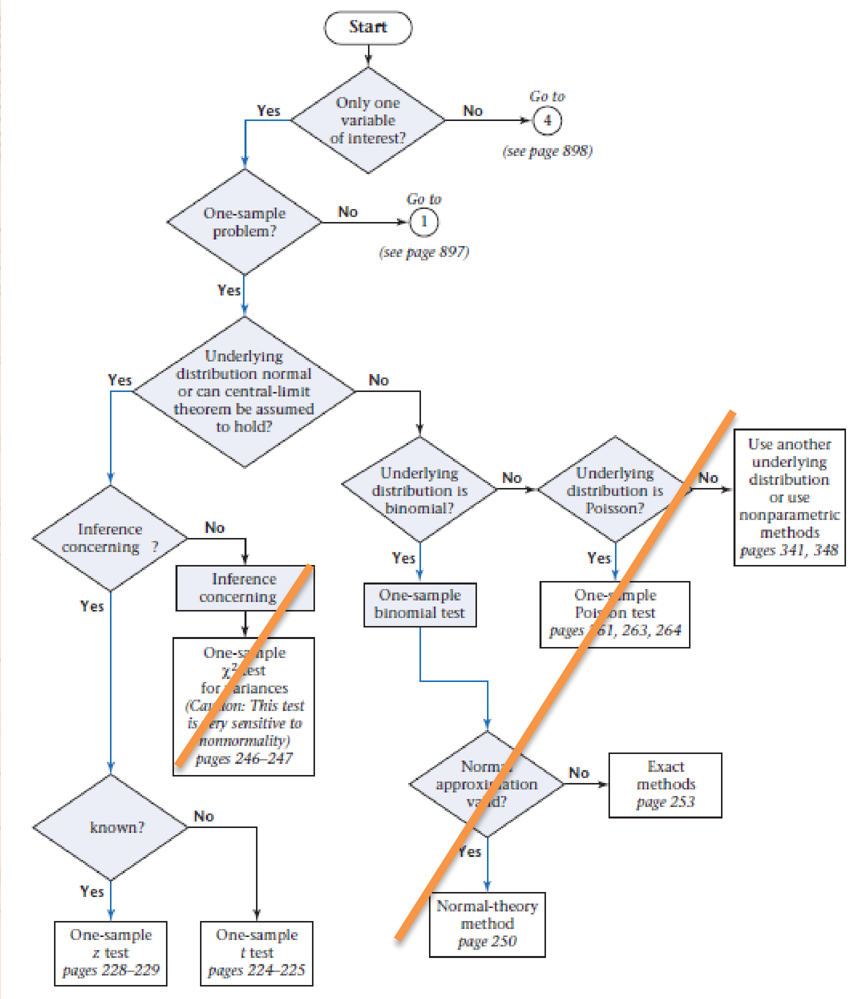
- One-sided test





- P-value
- Z-test (Normal distribution table) *when population variance is known
- T-test (t distribution table) *when population variance is unknown
- Power and sample size

FIGURE 7.18 Flowchart for appropriate methods of statistical inference



Hypothesis testing: Categorical data

- Contingency Table Approach
  - Expected number of units in the (i,j) cell (E_{ij}):

    $$E_{ij} = \frac{i\text{th row margin} \times j\text{th column margin}}{\text{grand total}}$$

  - none of the four expected values < 5
- Fisher's exact test
  - 1 of the cells with expected values <=5
- McNemar's Test

i) Normal Theory test ($n_D \geq 20$)

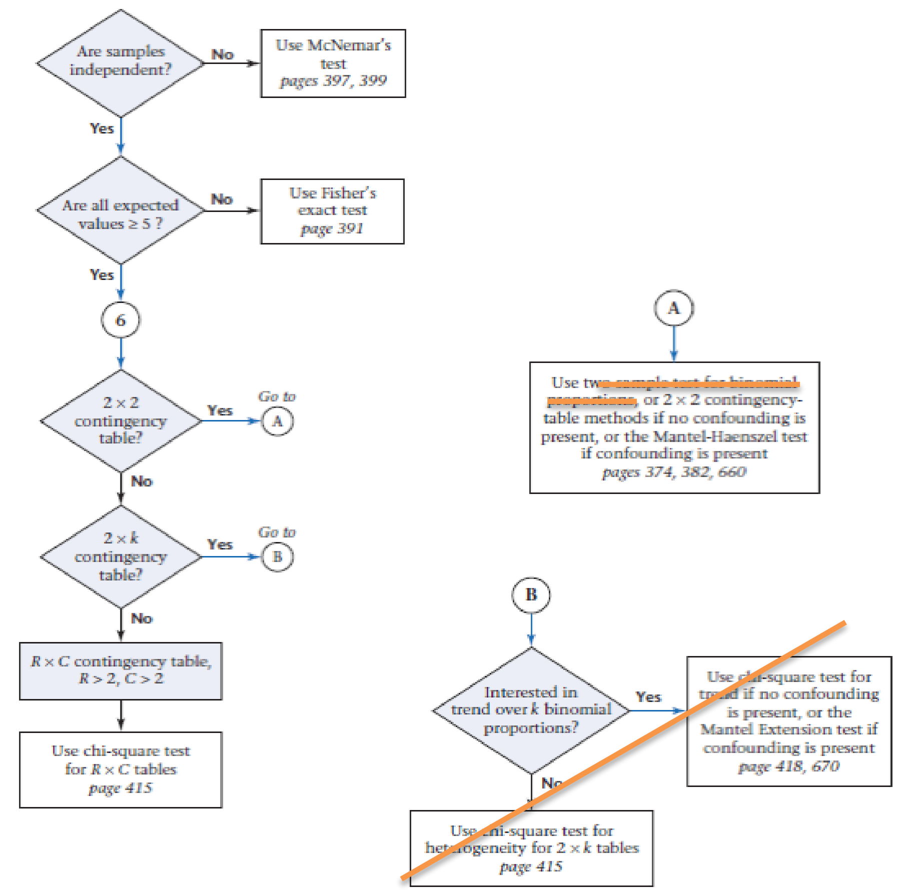
ii) Exact Method ($n_D < 20$)

- RxC Contingency Table
  - Test statistic:

    $$X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$$

  - $H_0 \sim \chi^2$ distribution with $(R - 1) \times (C - 1)$ *df*

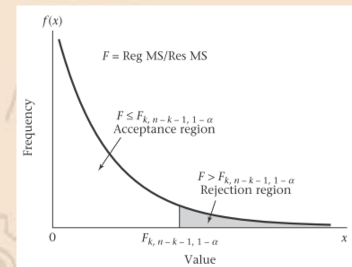
FIGURE 10.16 Flowchart for appropriate methods of statistical inference for categorical data



Regression and Correlation

- **Interpretation of regression line**
- **Correlation (Pearson's vs. Spearman ranks)**
- **Hypothesis testing for multiple regression**
  - **F test:**

    H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
    vs. *H_1*: at least one of the $\beta_j \neq 0$ in multiple linear regression



$$\text{Res SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{Reg SS} = \text{Total SS} - \text{Res SS}$$

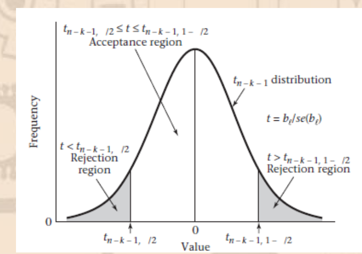
$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}$$

Test statistic:
F = Reg MS/Res MS, df = n-k-1 where n=sample size, k =no of independent

  - **T test:**

    H_0: $\beta_l = 0$, All other $\beta_j \neq 0$ vs. H_1: $\beta_l \neq 0$, all other $\beta_j \neq 0$ in multiple linear regression



- *Statistical output for multiple regression model*

Nonparametric Methods

- Parametric Methods: data of known distribution
- Non-parametric methods: data of unknown distribution, skewed / not normally distributed, ordinal

Analysis Type	Example	Parametric Procedure	Nonparametric Procedure	Note
Compare means between two distinct/independent groups	Is the mean systolic blood pressure (at baseline) for patients assigned to placebo different from the mean for patients assigned to the treatment group?	Two-sample t-test	Wilcoxon rank-sum test	1) Both n1 and n2 ⩾ 10: normal approximation method 2) n1 or n2 <10: small-sample Wilcoxon rank-sum test table (two-tailed critical values)
Compare two quantitative measurements taken from the same individual	Was there a significant change in systolic blood pressure between baseline and the six-month follow-up measurement in the treatment group?	Paired t-test	Wilcoxon signed-rank test	No. of non-zero di’s (differences of magnitudes) ⩾ 16 → normal approximation method
Estimate the degree of association between two quantitative variables	Is systolic blood pressure associated with the patient’s age?	Pearson coefficient of correlation	Spearman’s rank correlation	Pearson: actual values Spearman: rank scores

Multisample Inference

- **ANOVA test**: compare means of >2 groups

  Assumption: each group follows a normal distribution with the same variance

  – **F test:**

  H_0: $\alpha_i = 0$ for all i

  H_1: at least one $\alpha_i \neq 0$

  $$\text{Between SS} = \sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{\left(\sum_{i=1}^{k} n_i \bar{y}_i\right)^2}{n} = \sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{y_{..}^2}{n}$$

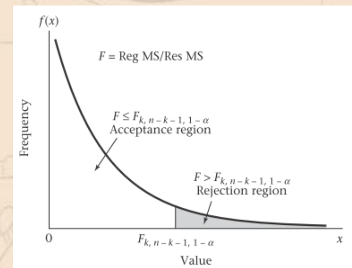
  $$\text{Within SS} = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

  Between MS = Between SS/(k-1)

  Within MS = Within SS/(n-k)

  Test statistic: *F* = Between MS/Within MS, df=k-1,n-k

  (n=sample size, k=no. of group)



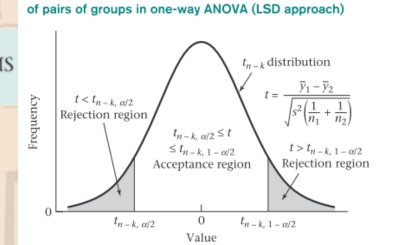
  – **T test:**

  H_0: $\alpha_1 = \alpha_2$ vs. H_1: $\alpha_1 \neq \alpha_2$

  Pooled estimate of variance = $s^2 = \sum_{i=1}^{k}(n_i-1)s_i^2 \Big/ \sum_{i=1}^{k}(n_i-1) = \left[\sum_{i=1}^{k}(n_i-1)s_i^2\right] \Big/ (n-k) = \text{Within MS}$

  Test statistic: $t = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, df=n-k

Acceptance and rejection regions for the *t* test for the comparison of pairs of groups in one-way ANOVA (LSD approach)



- Methods to adjust for multiple comparisons: **Bonferroni correction** and **false-discovery rate**

Odds ratio and logistic regression

- Main epidemiological study designs and relevant effect estimates:

Study Design	Prospective cohort study	Case and Control
Effect Estimate	Relative Risk/ Risk Ratio	Odds ratio

- **Multiple logistic regression:** $p = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

-

- OR: links exposure variable to the dependent variable: $\hat{OR} = e^{\hat{\beta}_j}$

  95% CI for OR: $$\left[e^{\hat{\beta}_j - z_{1-\alpha/2} se(\hat{\beta}_j)}, e^{\hat{\beta}_j + z_{1-\alpha/2} se(\hat{\beta}_j)}\right]$$