HST 190: Introduction to Biostatistics and Epidemiology

James Diao

Last updated: 28 August 2018

Contents

Lecture 1: Probability (7/30/18)

Definitions

1. Deductive reasoning: general to specific
2. Inductive reasoning: specific to general
3. Sample space: the set of all possible outcomes
4. Event: any subset of the sample space

Properties

1. Probabilities are additive (with inclusion-exclusion principle)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2. Law of total probability: if S can be divided into mutually exclusive events $B_1, B_2, \ldots, B_n$

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cup B_3) \ldots \\ &= \sum_{i=1}^{n} P(A \cup B_i) \\ P(A) &= P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + P(B_3)P(A \mid B_3) \ldots \\ &= \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \end{aligned}$$

3. Conditional probability

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

4. Independence

$$P(A \cap B) = P(A)P(B)$$

If B is not the null event, the following also applies:

$$P(A \mid B) = P(A)$$

5. Bayes's Theorem

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)}$$

Diagnostic testing

Measure definitions

$$
\begin{aligned}
\text{Prevalence} &= P(D^+) \\
\text{Sensitivity} &= P(T^+ \mid D^+) \\
\text{Specificity} &= P(T^- \mid D^-)
\end{aligned}
$$

$$
\begin{aligned}
\text{Positive Predictive Value (PPV, PV}^+) &= P(D^+ \mid T^+) = \frac{P(D^+)P(T^+ \mid D^+)}{P(T^+)} \\
&= \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1 - \text{specificity})(1 - \text{prevalence})}
\end{aligned}
$$

$$
\begin{aligned}
\text{Negative Predictive Value (NPV, PV}^-) &= P(D^- \mid T^-) = \frac{P(D^-)P(T^- \mid D^-)}{P(T^-)} \\
&= \frac{(\text{specificity})(1 - \text{prevalence})}{(1 - \text{sensitivity})(\text{prevalence}) - (\text{specificity})(1 - \text{prevalence})}
\end{aligned}
$$

Receiver operating characteristic (ROC)

- ROC curves plot TPR (sensitivity) vs. FPR (1-specificity) as the prediction threshold is changed.
- The area under the curve (AUC) is a useful summary of a classifier's predictive power.

Random variables and distributions

Distribution properties

- Probability distribution: assigns a probability to each outcome in the sample space.
- Expected value:

$$E(X) = \sum_i x_i P(X = x_i) = \int x f(x) dx$$

- Variance:

$$Var(X) = \sum_i (x_i - \mu_x)^2 P(X = x_i) = \int (x - \mu_x)^2 f(x) dx$$

Binomial: $X \sim Bin(n, p)$

$$P(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Normal: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma^2}\right)$$

- Rule of thumb: ~68.25% within 1 sd, ~95.50% within 2 sd, ~99.75 within 3 sd.
- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

Lecture 2: CTL and One-Sample Inference (8/1/18)

Central limit theorem

For large n:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

- The distribution converges to normality more quickly when the population distribution is also normal.
- Standardization of the sample mean is the first step for computing p-values and confidence intervals. - If sample size (n) is unknown, we have to approximate σ with $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}$. As a result, the sampling distribution actually converges to the t_{n-1} distribution.

$$E(s^2) = \sigma^2$$

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

Confidence intervals

To compute a 95% confidence interval, we start with:

$$\begin{aligned}
0.95 &= P\left(-1.96 \leq \frac{\bar{x}-\mu_x}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) \\
&= P\left(-1.96\left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x} \leq -\mu_x \leq 1.96\left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x}\right) \\
&= P\left(\bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \geq \mu_x \geq \bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right)
\end{aligned}$$

The confidence interval is:

$$\bar{x} \pm z_{1-\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{OR} \quad \bar{x} \pm t_{n-1,1-\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

- If confidence = 90%, $z = 1.65$
- If confidence = 95%, $z = 1.96$
- If confidence = 99%, $z = 2.58$

One-sample z-test

Assumptions

1. Random sampling
2. Independence
3. Approximately normal (or large sample)

Procedure

1. Define hypotheses

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

2. Standardize the observed sample mean

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

3. Collect the p-value: probability of sample estimate as extreme or higher, if H_0 is true. Reject H_0 if p-value $<$ significance level α.

Critical value

- The critical value is the test statistic for which p-value $= \alpha$
- It determines the acceptance and rejection regions (exactly like a confidence interval).

Lecture 3: Two-Sample Inference and Power (8/13/18)

Maximum likelihood estimation (MLE)

Let X_n denote a vector of n independent observations and θ_k denote k parameters to be estimated

$$\begin{aligned} L(X_n \mid \theta_k) &= P(x_1 \mid \theta_k) P(x_2 \mid \theta_k) \ldots P(x_n \mid \theta_k) \\ &= \prod_{i=1}^{n} P(x_i \mid \theta_k) \end{aligned}$$

The maximum likelihood estimator is:

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \text{arg max}_\theta \left[L(X_n \mid \theta_k) \right] \\ &= \text{arg max}_\theta \left[\log L(X_n \mid \theta_k) \right] \end{aligned}$$

Rationale

- MLE chooses the parameter values that maximize the probability of observing the given data.
- MLE is consistent: $\hat{\theta}_{\text{MLE}} \to \theta_0$ as $n \to \infty$.
- MLE is asymptotically normal: $\hat{\theta}_{\text{MLE}} \to \mathcal{N}(\theta_0, \sigma^2)$ as $n \to \infty$.

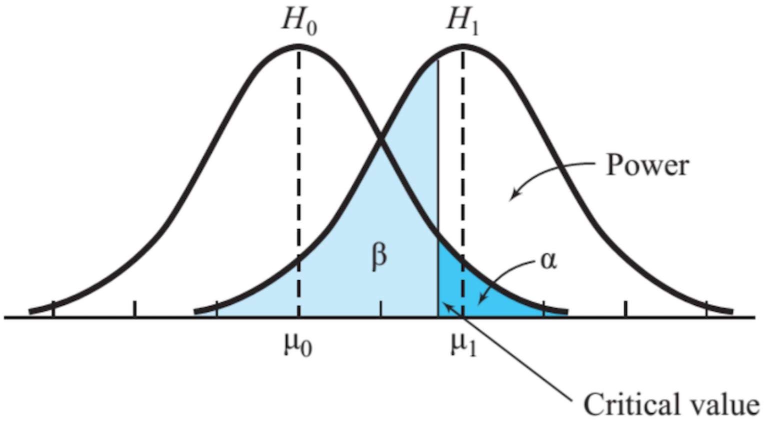


Figure 1: Statistical Power Graphic

Power and error

Definitions

$$\begin{aligned}\text{Type I Error}(\alpha) &= P(H_0 \text{ true and falsely reject } H_0) = \text{false alarm}\\ \text{Type II Error}(\beta) &= P(H_0 \text{ false and fail to reject } H_0) = \text{alarm failure}\\ \text{Power } (1-\beta) &= P(\text{reject } H_0 \mid H_0 \text{ false})\end{aligned}$$

Power of a z-test

$$1-\beta = \Phi\left(-z + (\mu_1 - \mu_0)\frac{\sqrt{n}}{\sigma}\right)$$

- One-sided: $z = z_{1-\alpha}$
- Two-sided: $z = z_{1-\alpha/2}$

Factors affecting power

1. Significance level (α)
2. Effect size ($\mu_1 - \mu_0$)
3. Sample size (n)
4. Population standard deviation (σ)

Proof:

$$\begin{aligned}
\text{Power} &= P(\text{Reject } H_0 \mid H_1) \\
&= P(Z > z_{1-\alpha} \mid \mu = \mu_1) \\
&= P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} \,\middle|\, \mu = \mu_1 \right) \\
&= P\left(\bar{X} > z_{1-\alpha} \frac{\sigma}{\sqrt{n}} + \mu_0 \,\middle|\, \mu = \mu_1 \right) \\
&= P\left(\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} + (\mu_0 - \mu_1) \frac{\sqrt{n}}{\sigma} \right) \\
&= 1 - \Phi\left(z_{1-\alpha} + (\mu_0 - \mu_1) \frac{\sqrt{n}}{\sigma} \right) \\
\text{Power} &= \Phi\left(-z_{1-\alpha} + (\mu_1 - \mu_0) \frac{\sqrt{n}}{\sigma} \right) \quad \text{if } \mu_1 > \mu_0 \\
\text{Power} &= \Phi\left(z_{1-\alpha} - (\mu_1 - \mu_0) \frac{\sqrt{n}}{\sigma} \right) \quad \text{if } \mu_1 < \mu_0
\end{aligned}$$

Required sample size for desired power

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_1 - \mu_0)^2}$$

Proof:

$$\begin{aligned}
1 - \beta &= \Phi\left(-z_{1-\alpha} + (\mu_1 - \mu_0) \frac{\sqrt{n}}{\sigma} \right) \\
z_{1-\beta} &= -z_{1-\alpha} + (\mu_1 - \mu_0) \frac{\sqrt{n}}{\sigma} \\
\frac{\sqrt{n}}{\sigma} &= \frac{z_{1-\beta} + z_{1-\alpha}}{\mu_1 - \mu_0} \\
n &= \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_1 - \mu_0)^2}
\end{aligned}$$

Two-sample t-test

Paired data

- Each data point on one sample is related to a unique data point in the other sample.
- This is actually one sample (of differences) in disguise.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

Unpaired data

Pooled variance estimator (if population variances are equal):

s_p^2 is the df-weighted average of s_1^2 and s_2^2

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

s_d^2 is the sum of the variances of the sample averages

$$\begin{aligned} t &= \frac{(\bar{x}_2-\bar{x}_1)-(\mu_2-\mu_1)}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}} \sim t_{n_1+n_2-2} \\ (\bar{x}_2-\bar{x}_1) &\pm t_{df,1-\alpha/2}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}} \\ df &= n_1+n_2-2 \end{aligned}$$

Separate variance estimators (if population variances are unequal):

$$\begin{aligned} t &= \frac{(\bar{x}_2-\bar{x}_1)-(\mu_2-\mu_1)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{df} \\ df &= \left\lfloor \frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}} \right\rfloor \\ (\bar{x}_2-\bar{x}_1) &\pm t_{df,1-\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \end{aligned}$$

Lecture 4: Nonparametric Tests and Clinical Trials (8/13/2018)

Nonparametric tests

- Agnostic to the underlying population distribution.
  - More robust to non-normality, small sample sizes, outliers, ordinary data, and measurement error.
  - Loss of power from reduced information.
- Ranked methods only apply to hypotheses about population medians, not means.

Wilcoxon signed-rank test

Rationale

- Replaces observations with signed ranks (robust to outliers and retains relative magnitudes)
- Used in lieu of the 1-sample t-test
- Tests whether the population median is equal to a value (usually 0)

Procedure

1. Rank the differences
   - Arrange the differences d_i in order of absolute value.
   - Count the number of differences with the same absolute value.
   - Ignore the observations where $d_i = 0$ and rank the remaining observations as 1-n for low-high.
   - If there is a group of several observations with the same absolute value, then assign the average rank for the whole group.
2. Compute the rank sum R_1 of the positive differences and the corresponding t-statistic and p-value based on the underlying distribution (boxed area only necessary for $g > 0$ number of ties):

$$R \sim \mathcal{N}\left(\mu = \frac{n(n+1)}{4},\ \ \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24 - \boxed{\sum_{i=1}^{g}(t_i^3 - t_i)/48}}}\right)$$

3. Apply the continuity correction (boxed) and compute T:

$$T = \frac{|R_1 - \mu_R| - \boxed{\frac{1}{2}}}{\sigma_R} \sim \mathcal{N}(0,1)$$

Wilcoxon rank sum test (Mann-Whitney)

- Replaces observations with ranks in lieu of the unpaired two-sample t-test
- Assumes that the two distributions have the same shape

Procedure

1. Rank the differences
   - Same procedure as before
2. Compute the rank sum R_1 from the first sample and the corresponding t-statistic and p-value based on the underlying distribution (boxed area only necessary for $g > 0$ number of ties):

$$R \sim \mathcal{N}\left(\mu = \frac{n_1(n_1+n_2+1)}{2},\ \ \sigma = \sqrt{\frac{n_1 n_2}{12}\left[n_1 + n_2 + 1 - \boxed{\frac{1}{(n_1+n_2)(n_1+n_2-1)}\sum_{i=1}^{g} t_i^3 - t_i}\right]}\right)$$

3. Apply the continuity correction (boxed) and compute T:

$$T = \frac{|R_1 - \mu_R| - \boxed{\frac{1}{2}}}{\sigma_R} \sim \mathcal{N}(0,1)$$

Clinical trial study design

Design process

1. Define questions → aims → endpoints
2. Define study populations (eligibility)
3. Design and plan study: concepts → details (protocol)
4. Implement and monitor study (randomization)
5. Analyze and interpret interim and final data

Phase	Objective	Sample Size
I	safety, dosage	~15-30
II	safety, efficacy	~100
III	safety, efficacy	~100-1000s
IV	post-marketing surveillance	Depends

Design factors

1. Choosing a target population (easy to accrue, compliant, likely treatment benefit)
2. Hypotheses
   - Superiority and non-inferiority
   - Predefine as: hazard(A)/hazard(B) < H, or $\mu_A - \mu_B > D$
3. Endpoints
   - Primary endpoint: type I error determines power/sample size
   - Secondary endpoints: may be accounted for in powering, but not always
4. Randomization
   - Simple random sampling: can be inefficient
   - Stratification : institution, gender, severity, past exposure
   - Blocking: treatment assignment
   - Adaptive: based on responses (play the winner)
5. Blinding:
   - Single-blind: subject does not know which group they're in
   - Double-blind: researchers also don't know which groups the subjects are in
   - Triple-bind: monitoring committee also doesn't know which groups the subjects are in
6. Interim monitoring: safety, efficacy
   - Data Safety Monitoring Board (DSMB) regularly reviews study conduct and data
   - Problem: loses info on secondary endpoints
7. Statistical details:
   - Sample size, type I error, power, desired effect size
   - Stopping rule, number of interim analyses, drop-out rate
8. Statistical Model
   - Translate data and hypotheses into a statistical model; fit the model and interpret results
   - Intention-to-treat: unbiased and conservative with non-compliance/drop-out
   - Per-protocol: biased and optimistic (higher power) with non-compliance/drop-out
9. Analyses
   - Missing data: last observation carried forward (LOCF) and multiple imputation.
   - Report all subgroup analyses; settles multiplicity problem.
   - Adjust for subgroup heterogeneity using an interaction test
   - Adjust for dependencies (longitudinal/repeated measurements, adjacent anatomical locations, adjacent genetic loci)

Lecture 5: Linear Regression (8/15/2018)

Correlation

Given paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the Pearson product-moment correlation is:

$$\begin{aligned}\rho &= E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \\ &= \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}\end{aligned}$$

The sample correlation coefficient is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{X}}{s_X}\right)\left(\frac{y_i-\bar{Y}}{s_Y}\right)$$

- r is sensitive to outliers and highly non-normal distributions
- An alternative is Spearman's rank correlation, which replaces the data values with their relative ranks.
- To test the null hypothesis $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, we compute the test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

Simple linear regression

Basic model

$$Y = \beta_0 + \beta_1 X + e, \quad e \sim \mathcal{N}(0, \sigma_e^2)$$

$$\begin{aligned}Y \mid X &\sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma_e^2) \\ E[Y \mid X] &= \beta_0 + \beta_1 X \\ Var[Y \mid X] &= \sigma_e^2\end{aligned}$$

Assumptions for residuals (LINE):

1. Linearity
2. Independence
3. Normal errors
4. Equal variance

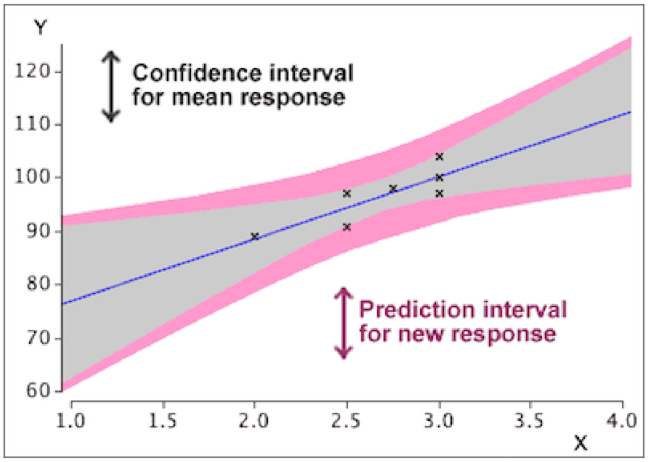


Figure 2: Confidence and Prediction Intervals

Fitting the model

The optimal parameters $(\hat{\beta}_0, \hat{\beta}_1)$ will minimize the squared error. This can be solved using differential calculus (taking partial derivatives), linear algebra (solving the normal equations), or MLE (algebraically or by gradient descent).

$$\text{Squared error} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{arg min}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\beta_1 = r\left(\frac{s_Y}{s_X}\right)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Predictions and intervals

1. Confidence interval: estimates the mean response:

$$\hat{y} \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2 / s_x^2}{(n-k)}}$$

where:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}$$

2. Prediction interval: estimates a new response:

$$\hat{y} \pm t_{n-k,1-\alpha/2} \times \hat{\sigma}\sqrt{\boxed{1+}\frac{1}{n} + \frac{(x_{new} - \bar{x})^2 / s_x^2}{(n-k)}}$$

This accounts for the additional uncertainty of the error term itself.
Caution: do not extrapolate to x-values beyond where you have data!

Inference about the slope (β_1)

$$\text{S.E. } (\beta_1) = \frac{\hat{\sigma}}{\sqrt{s_x^2(n-1)}}$$

$$b \pm t_{n-k,1-\alpha/2} \times \text{S.E. } (\beta_1)$$

$$t = \frac{\beta_1}{\text{S.E. } (\beta_1)} \sim t_{n-k}$$

Multiple linear regression

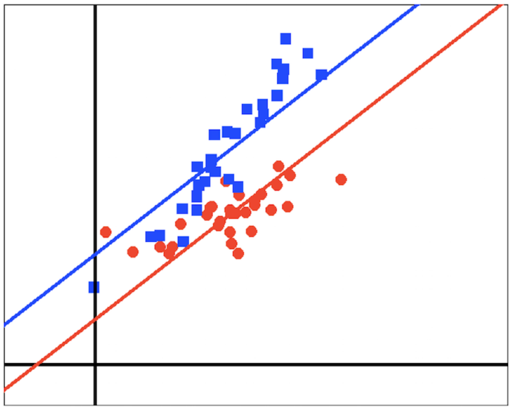
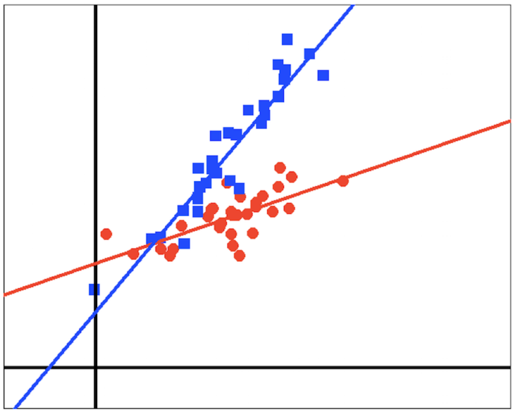
Basic model

$$Y = \beta_0 + \beta_1 X + \beta_2 x_2 + \cdots + \beta_k x_k + e, \quad e \sim \mathcal{N}(0, \sigma_e^2)$$

Categorical variables

- Indicator/dummy variables are created to capture categories. You need $k-1$ indicators for k variables.
- The associated regression parameters represent constant differences between each category and the baseline category.
- Interaction terms involve the product of predictor variables.
- This leads to different slopes when conditioning on a given x_k; the divergence in slopes is given by β_k.
- When fitting a linear model computationally, t-statistics and p-values for β_j are computed with respect to $H_0 : \beta_j = 0$ and all other $\beta_{\neq j}$ are fixed as their point estimates.

$$\frac{\beta_j}{\text{S.E. } (\beta_j)} \sim t_{n-k-1}$$

Main effect term: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 G$	Interaction term: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$
	

Lecture 6: Multiple Linear Regression Cont. (8/17/18)

Describing variation

Sums of squares (SS)

$$\begin{aligned} SS_{total} &= SS_{regression} + SS_{residual} \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 \\ \text{Total variation} &= \text{From model} + \text{From noise/other} \end{aligned}$$

Coefficient of determination (R^2)

Definition:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{resid}}{SS_{total}} \\ &= \text{proportion of variation explained by the model} \\ &= \text{correlation coefficient squared } (r^2) \end{aligned}$$

Adjusted R^2:

$$\begin{aligned} R^2_{adj} &= 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2) \\ \lim_{n\to\infty} R^2_{adj} &= R^2 \end{aligned}$$

- n = number of data points and k = number of predictor variables
- R^2 always increases when a new variable is added; does not account for model complexity

F-test

For the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Test H_0 (intercept only) vs. H_1 (full model), where:

$$\begin{aligned} H_0 :& \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \ldots \text{ and } \beta_k = 0 \\ H_1 :& \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \ldots \text{ or } \beta_j \neq 0 \end{aligned}$$

Under H_0, the F-statistic follows the F-distribution:

$$\begin{aligned}
MS_{reg} &= \frac{RSS_{reg}}{k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2}{k} \sim \chi^2_k \\
MS_{resid} &= \frac{RSS_{resid}}{n-k-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{ij} - \bar{Y}_i)^2}{n-k-1} \sim \chi^2_{n-k-1} \\
F &= \frac{MS_{reg}}{MS_{resid}} \\
&= \frac{\text{explained variance}}{\text{unexplained variance}} = \frac{\text{between-group variability}}{\text{within-group variability}} \\
F &\sim F_{k,n-k-1}
\end{aligned}$$

Reject H_0 if $F > F_{n-k-1,1-\alpha}$

ANOVA

Suppose we had k different populations, each roughly normal with common variance σ^2, and we wanted to test for equality:

$$\begin{aligned}
H_0 &: \ \mu_1 = \mu_2 = \cdots = \mu_k \\
H_1 &: \ \text{at least one } \mu \text{ is different from the others}
\end{aligned}$$

Assumptions

- Homoscedasticity (equal variance σ^2)
- Units in k samples are independent (within and between samples)
- Populations are approximately normal

Basic model

$$\begin{aligned}
y &= \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \cdots + \beta_k I_k + e \\
\beta_0 &= \bar{y}_0 \\
\beta_{i|i\neq 0} &= \bar{y}_{i|i\neq 0} - \bar{y}_0
\end{aligned}$$

Non-parametric model: apply ANOVA to the sample ranks (Kruskal-Wallis)

$$\begin{aligned}
H_0 &: \ \text{median}(y_1) = \text{median}(y_2) = \cdots = \text{median}(y_k) \\
H_1 &: \ \text{at least one median is different from the others} \\
&\text{The final test statistic is } KW \sim \chi^2_{k-1}
\end{aligned}$$

Multiple comparisons

When comparing 2 of k groups, use the pooled variance estimator if all group variances are considered equal:

$$s_p^2 = \frac{SS_{resid}}{n-k} = \frac{1}{n-k}\sum_{i=0}^{k}(n_i - 1)s_i^2$$

This gives the t-statistic:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n-k}$$

Bonferroni correction

$$\alpha^* = P(\text{at least 1 type I error}) \leq \frac{\alpha}{\text{No. of tests}}$$

False discovery rate

Controls the expected proportion of incorrectly rejected null hypotheses: P(reject each H_0 | all H_0 true).

Procedure

1. Rank tests by p-value ($p_1 \leq p_2 \leq \cdots \leq p_k$)
2. Define $q_i = \frac{k}{i}p_i$
3. Define $FDR_i = \min(q_i, \ldots q_k)$. These will be ranked in increasing order.
4. Reject all hypotheses with $FDR_i < FDR^*$.

Example (for $n = 50$)

i (rank)	Test	p-value	q_i	FDR$_i$
1	#31	0.0001	0.0001(50/1) = 0.0050	0.0050
2	#21	0.0015	0.0015(50/2) = 0.0375	0.0317
3	#49	0.0019	0.0019(50/3) = 0.0317	0.0317
4	#50	0.0170	0.0170(50/4) = 0.2125	0.1800
5	#4	0.0180	0.0180(50/5) = 0.1800	0.1800

Lecture 7: Inference for Categorical Data (8/19/18)

One-proportion inference

Suppose that we count X successes and $N - X$ failures from a sample size of N.

$$\begin{array}{ll} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{array} \quad \text{or} \quad \begin{array}{ll} H_0: & p > p_0 \\ H_1: & p \leq p_0 \end{array}$$

$$\hat{p} = X/N$$

$$X \sim Bin(p_0, N)$$

$$P(X = k) = \binom{N}{k} p_0^k (1 - p_0)^{N-k}$$

Binomial exact method:

$$\text{If } \hat{p} \leq p_0: \; p\text{-value} = 2 \sum_{k=0}^{X} P(X = k)$$

$$\text{If } \hat{p} > p_0: \; p\text{-value} = 2 \sum_{k=X}^{N} P(X = k)$$

Normal approximation method (valid when $\text{Var}(X) = np(1 - p) \geq 5$):

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \sim \mathcal{N}(0, 1)$$

$$n = \frac{p_0(1 - p_0)\left(z_{1-\alpha/2} + z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{(p_1 - p_0)^2}$$

Two-proportion comparisons

$$H_0: \; p_1 = p_2$$

$$H_1: \; p_1 \neq p_2$$

$$\text{If } p_1 = p_2: \; \hat{p}_{pooled} = \frac{X_1 + X_2}{N_1 + N_2}$$

$$p_1 - p_2 \sim \mathcal{N}\left(0, \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}\right)$$

If variances are assumed equal, replace $\hat{p}_1$ and $\hat{p}_2$ with the pooled proportion estimate $\hat{p}$

Chi-square test

Contingency table

Data is cross-classified according to discrete/categorical variables:

	Positive	Negative	Total
Sharing needles	12	28	40
Not sharing needles	11	49	60
Total	23	77	100

Pearson's chi-square test

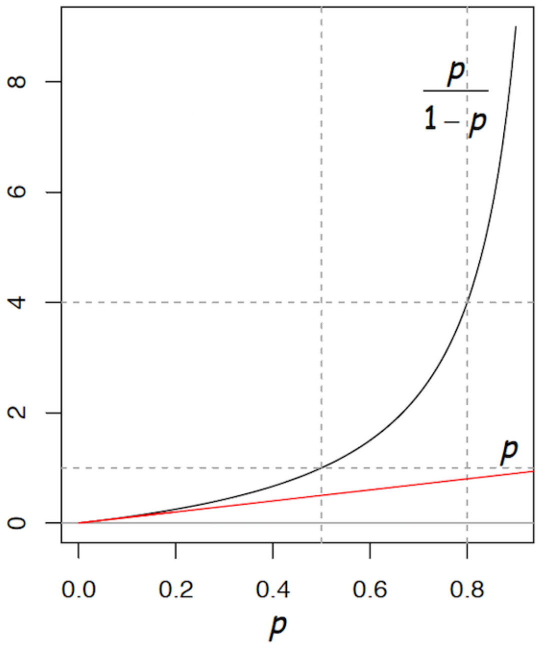
Tests association between 2 categorical variables:

- H_0: variables are not associated (joint = product of marginals)
- H_1: variables are associated

$$\begin{aligned}
E_{ij} &= \frac{O_i O_j}{N} \\
X^2 &= \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(\text{observed}_{ij} - \text{expected}_{ij} \boxed{-0.5})^2}{\text{expected}_{ij}} \\
&= \frac{N\left(|ad - bc| \boxed{-\frac{N}{2}}\right)^2}{(a+b)(c+d)(a+c)(b+d)} \\
X^2 &\sim \chi^2_{df,\,1-\alpha} \\
df &= (R-1)(C-1)
\end{aligned}$$

Details

- 2x2 is valid only if all $E_{ij} \geq 5$
- RxC is valid only if all $E_{ij} \geq 1$ and at least 80% of $E_{ij} \geq 5$
- Subtract 0.5 for the Yates continuity correction
- Right-tail integral gives the p-value for a 2-sided alternative



$$\frac{p}{1-p} \approx p \text{ for small } p$$

Figure 3: Odds Ratio

Odds ratios and relative risk

	Disease$^+$	Disease$^-$	Total
Exposure$^+$	A	B	A+B
Exposure$^-$	C	D	C+D
Total	A+C	B+D	N

Odds ratios and relative risk measure the *magnitude* of association between 2 categorical variables.

$$\begin{aligned} p_1 &= P(\text{disease} \mid \text{exposed}) \\ p_2 &= P(\text{disease} \mid \text{NOT exposed}) \\ \text{Risk Difference } &= p_1 - p_2 \\ \text{Risk Ratio} &= \frac{p_1}{p_2} \\ \text{Odds ratio} &= \frac{p_1}{1-p_1} \bigg/ \frac{p_2}{1-p_2} \\ &= ad/bc \end{aligned}$$

Case-control (retrospective) study:

- To compute $\hat{p}_1$ and $\hat{p}_2$, we need to sample patients on exposure and classify on disease
- Instead, a case-control study samples patients on disease status and classifies on exposure.
- The case-control odds-ratio is the same as sampling by exposure and taking the ratio of the odds for $\hat{p}_1$ and $\hat{p}_2$.
- If $p \ll 1$ (low prevalence), $\frac{p}{1-p} \approx p$ and odds-ratio $\approx$ risk ratio

Odds ratio

$$\begin{aligned}
OR > 1 &: \text{ exposure} \rightarrow \text{higher disease risk} \\
OR < 1 &: \text{ exposure} \rightarrow \text{lower disease risk} \\
OR = 1 &: \text{ no association between exposure and disease risk} \\
\ln(\widehat{OR}) &\sim \mathcal{N}\left(\ln(OR), \sqrt{\text{Var}(\ln \widehat{OR})}\right) \\
\text{Var}(\ln \widehat{OR}) &\approx \frac{1}{n_1\hat{p}_1(1-\hat{p}_1)} + \frac{1}{n_2\hat{p}_2(1-\hat{p}_2)} \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \\
\text{ln(OR) CI: } & \ln(\widehat{OR}) \pm z_{1-\alpha/2}\sqrt{\text{Var}(\ln \widehat{OR})} = [c_{\text{lower}}, c_{\text{upper}}] \\
\text{OR CI: } & [e^{c_{\text{lower}}}, e^{c_{\text{upper}}}]
\end{aligned}$$

Note: the OR confidence interval is NOT symmetric about the point estimate.

Mantel-Haenszel method

- Confounding: stratifying results by a confounding variable may affect disease-exposure association
- Simpson's paradox:
  - A factor associated with both treatment assignment and outcome may reverse the direction of association
  - Example: compared to open procedures, percutaneous procedures are associated with higher success rate overall (OR > 1), but lower success rate (OR < 1) when outcomes are stratified by small stones and large stones

Chi-square test for homogeneity:

1. Stratify your data into k strata (RxC tables)
2. Compute the statistic X^2_{homo}

$$\begin{aligned}
H_0 &: \; OR_1 = OR_2 = \cdots = OR_k \text{ (homogeneity)} \\
H_1 &: \text{ at least one } OR \text{ is different (heterogeneity)} \\
X^2_{homo} &= \sum_{i=1}^{k} w_i \left(\ln \widehat{OR}_i - \ln \overline{OR}\right)^2 \sim \chi^2_{k-1} \\
w_i &= \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)^{-1}
\end{aligned}$$

$$\ln \widehat{OR}_i = \ln\left(\frac{a_i d_i}{b_i c_i}\right)$$
$$\ln \overline{OR} = \sum_{i=1}^{k} w_i \ln \widehat{OR}_i \Bigg/ \sum_{i=1}^{k} w_i$$

3. If you conclude homogeneity, compute the Mantel-Haenszel estimator of the common odds ratio:

$$\widehat{OR}_{MH} = \sum_{i=1}^{k} \frac{a_i d_i}{n_i} \Bigg/ \sum_{i=1}^{k} \frac{b_i c_i}{n_i}$$

4. Compute the confidence interval
   - Check the following assumptions:

$$\sum_{i=1}^{k} \frac{(a_i + c_i)(a_i + b_i)}{n_i} \geq 5, \qquad \sum_{i=1}^{k} \frac{(a_i + c_i)(c_i + d_i)}{n_i} \geq 5$$

$$\sum_{i=1}^{k} \frac{(b_i + d_i)(a_i + b_i)}{n_i} \geq 5, \qquad \sum_{i=1}^{k} \frac{(b_i + d_i)(c_i + d_i)}{n_i} \geq 5$$

   - Compute the CI as:

$$\ln OR_{MH} \text{ CI: } \ln \widehat{OR}_M H \pm z_{1-\alpha/2}\left(\frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}\right) = [c_{\text{lower}}, c_{\text{upper}}]$$
$$OR_{MH} \text{ CI: } [e^{c_{\text{lower}}}, e^{c_{\text{upper}}}]$$

5. Perform a hypothesis test on:

$$H_0 : \ OR = 1$$
$$H_1 : \ OR \neq 1$$

   Compute:

$$X^2_{MH} = \frac{(|O - E| - 0.5)^2}{V}$$
$$O = \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} a_i$$
$$E = \sum_{i=1}^{k} E_i = \sum_{i=1}^{k} \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$
$$V = \sum_{i=1}^{k} V_i = \sum_{i=1}^{k} \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

   If H_0 is true and $V \geq 5$: $X^2_{MH} \sim \chi^2_{(R-1)(C-1)}$

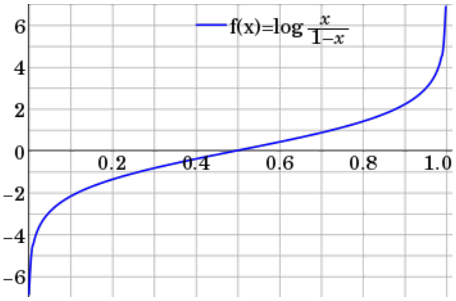


Figure 4: The logit function

Lecture 8: Logistic Regression and Survival Analysis (8/22/18)

Logistic regression

$$\begin{aligned}\text{logit}(p) &= \ln\left(\frac{p}{1-p}\right)\\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\\ p &= \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}\end{aligned}$$

- The β_j coefficients are log-odds (an increase of 1 unit means that the odds increase by a factor of e^1).
- In other words, $\ln\widehat{OR} \sim \mathcal{N}\left(\hat{\beta}_j, \text{S.E.}(\hat{\beta}_j)\right)$.

Recall:

$$\text{S.E.}(\beta_k) = \frac{\hat{\sigma}}{\sqrt{s_x^2(n-1)}}, \quad \hat{\sigma} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k}$$

Additional assumption for logistic regression:

$$Y \sim Bern\left(p = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right)$$

Survival analysis

- Survival analysis covers “time-to-event.”
- Often described by a survival curve $S(t)$, which plots 1−CDF(failure).
- Censored data: when you stop receiving data on an subject before failure.
  - You only know they survived at least as long as they did.
  - Assume that censoring is noninformative and unbiased, i.e., that being lost is unrelated to prognosis.

Kaplan-Meier estimation

The Kaplan-Meier product-limit estimator: $\hat{S}(t)$

$$\begin{aligned} S(t_i) = &P(\text{alive at } t_i \mid \text{alive at } t_{i-1}) \\ &\times P(\text{alive at } t_1 \mid \text{alive at } t_{i-2}) \\ &\dots \\ &\times P(\text{alive at } t_2 \mid \text{alive at } t_1) \\ &\times P(\text{alive at } t_1) \end{aligned}$$

Estimate each "step survival" probability as:

$$\begin{aligned} P(S_{i-1,\,i}(t)) &= \frac{N_{\text{alive and not censored at } t_i}}{N_{\text{alive and not censored at } t_{i-1}}} \\ &= 1 - \frac{N_{\text{died at } t_i}}{N_{\text{alive and not censored at } t_{i-1}}} \end{aligned}$$

That gives us $\hat{S}(t)$:

$$\hat{S}_{KM}(t_i) = \left(1 - \frac{d_1}{S_0}\right) \times \left(1 - \frac{d_2}{S_1}\right) \times \cdots \times \left(1 - \frac{d_i}{S_{i-1}}\right) = \prod_{j=1}^{i} 1 - \frac{d_j}{S_{j-1}}$$

This estimator jumps at event times only; goes to zero if there are no more events.

Example:

Year (t)	Failed (d_i)	Censored (I_i)	Survived (S_i)	Total (S_{i-1})	$\hat{S}(t)$
2	7	2	$100 - 7 - 2 = 91$	100	$\hat{S}(2) = 1 - \frac{7}{100}$
4	16	5	$91 - 16 - 5 = 70$	91	$\hat{S}(4) = \left(1 - \frac{16}{91}\right) \hat{S}(2)$
6	19	8	$70 - 19 - 8 = 43$	70	$\hat{S}(6) = \left(1 - \frac{19}{70}\right) \hat{S}(4)$

Confidence intervals:

$$\ln \hat{S}(t_i) \sim \mathcal{N}\left(\ln S(t_i),\ \sum_{j=1}^{i} \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}\right)$$

$$\text{CI for } \ln \hat{S}(t_i) :\ \ln \hat{S}(t_i) \pm z_{1-\alpha/2} \sqrt{\sum_{j=1}^{i} \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}} = [c_{\text{lower}}, c_{\text{upper}}]$$

$$\text{CI for } \hat{S}(t_i) :\ [e^{c_{\text{lower}}}, e^{c_{\text{upper}}}]$$

Hazard function

- The hazard function can be considered to be:
  - An instantaneous conditional death rate
  - The probability of an event at time t given no event up to time t
- $h(t)$ often represents survival distributions; constant for the exponential distribution

$$h(t) = \frac{\lim_{\Delta t \to 0} \left(\frac{S(t) - S(t+\Delta t)}{\Delta t} \right)}{S(t)} = \frac{\text{instantaneous death rate}}{\text{fraction of individuals still alive}}$$

Log-rank test

The log-rank test compares survival functions with the following hypotheses:

- H_0: $h_1(t) = h_2(t)$ for all t in the study (or $h_1(t)/h_2(t) = 1$)
- H_1: $h_1(t) \neq h_2(t)$ for all t in the study

The log-rank test is a direct application of the Mantel-Haenszel test:

- Divide the study period into k intervals
- Create a 2x2 table for each interval (group 1/2 vs. death/survival)

	Death	Survived or Censored	Total
Group 1	a_i	b_i	$n_{i,1}$
Group 2	c_i	d_i	$n_{i,2}$
Total	$a_i + c_i$	$b_i + d_i$	n_i

- Compute the $\chi^2_{LR} \sim \chi^2_1$ test statistic (see chi-square test for homogeneity on page 22)

Modeling survival with regression

Cox model (proportional-hazards model) is semiparametric and fits diverse survival distributions.

- Models $h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$
- In other words: $\ln h(t) = \ln h_0(t) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

Interpreting the Cox Model

- $h_0(t)$ is the "baseline hazard rate" = $h(t)$ when all $x_i = 0$.
- A unit incrase of any covariate will scale $h(t)$ by $\exp(\Delta\beta_i)$
- Use the hypothesis H_0: $\beta_i = 0$ to test whether a covariate affects survival time
- Make sure the final KM curves are proportional (not converging or diverging)

Lecture 9: Optimizing Linear Models (8/24/2018)

Mixed-effects model

Basic model:

$$
\begin{aligned}
y_{ij} &= \alpha_i + \beta_0 + \beta_1 x_j + e_{ij} \\
y_{ij} &= \text{response at time } j \text{ for person } i \\
x_j &= \text{time point} \\
\alpha_i &= \text{random effect: intercept adjustment} \\
&\sim \mathcal{N}(0, \sigma_A^2) \\
\beta_1 &= \text{fixed effect: slope} \\
e_{ij} &\sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

Examples

- Longitudinal studies (group by same subject, same time point, etc.)
- Multi-stage sampling (counties → households → individuals)

Rationale

- Accounts for correlations/dependencies without adding i parameters for i subjects (as fixed effects does)
- Properly indicates smaller coefficient S.E. values
- Gives an estimate for σ_A; how much subject intercepts differ

Alternative

- Sandwich estimator: estimates variance of $\hat{\beta}$ as a function of the covariates $(x_1, \ldots x_k)$ and $\text{Var}(y_i)$.

Model checking

1. Linearity
   - Diagnose using pairwise scatterplots. Take vertical slices and look for (a) means in straight line and (b) SDs approximately equal
   - Consider transformations (log, inverse, square, exp, or interaction terms)
   - Use splines, polynomial regression, or generalized additive models
   - Coefficients are biased
2. Independence of Errors
   - Diagnose on plot of residuals vs. each X
   - Problems occur with interacting units, spatial/temporal proximity, common data source, or clustering effects
   - Consider modeling dependencies with random-effects or time-series models
   - Coefficients are unbiased, but standard errors are affected.
   - If residuals are positively correlated, we have less information and our confidence intervals will be optimistic
3. Normality of errors
   - Diagnose using QQ-plot
   - Consider conducting regression with t-distributed errors
4. Equal variance of residuals
   - Diagnose on plot of residuals vs. each X
   - Consider “squashing” transformations or weighted regression (observations weighted by 1/variance)
   - Heteroscedasticity does not bias coefficients, but standard errors are affected

Variable selection

- Methods: forward, backward, stepwise, all subsets
- Criteria:
  - General form: $f(\hat{\sigma}^2) + g(p)$
  - Adjusted-R^2 (minimize residual variance): $1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$
  - AIC: $n \ln(SS_{resid}/n) + 2p$
  - BIC: $n \ln(SS_{resid}/n) + p \ln n$
- Also include:
  - Predictors that are significant or with the expected sign
  - Interaction terms for predictors with large effect sizes

Cross-validation

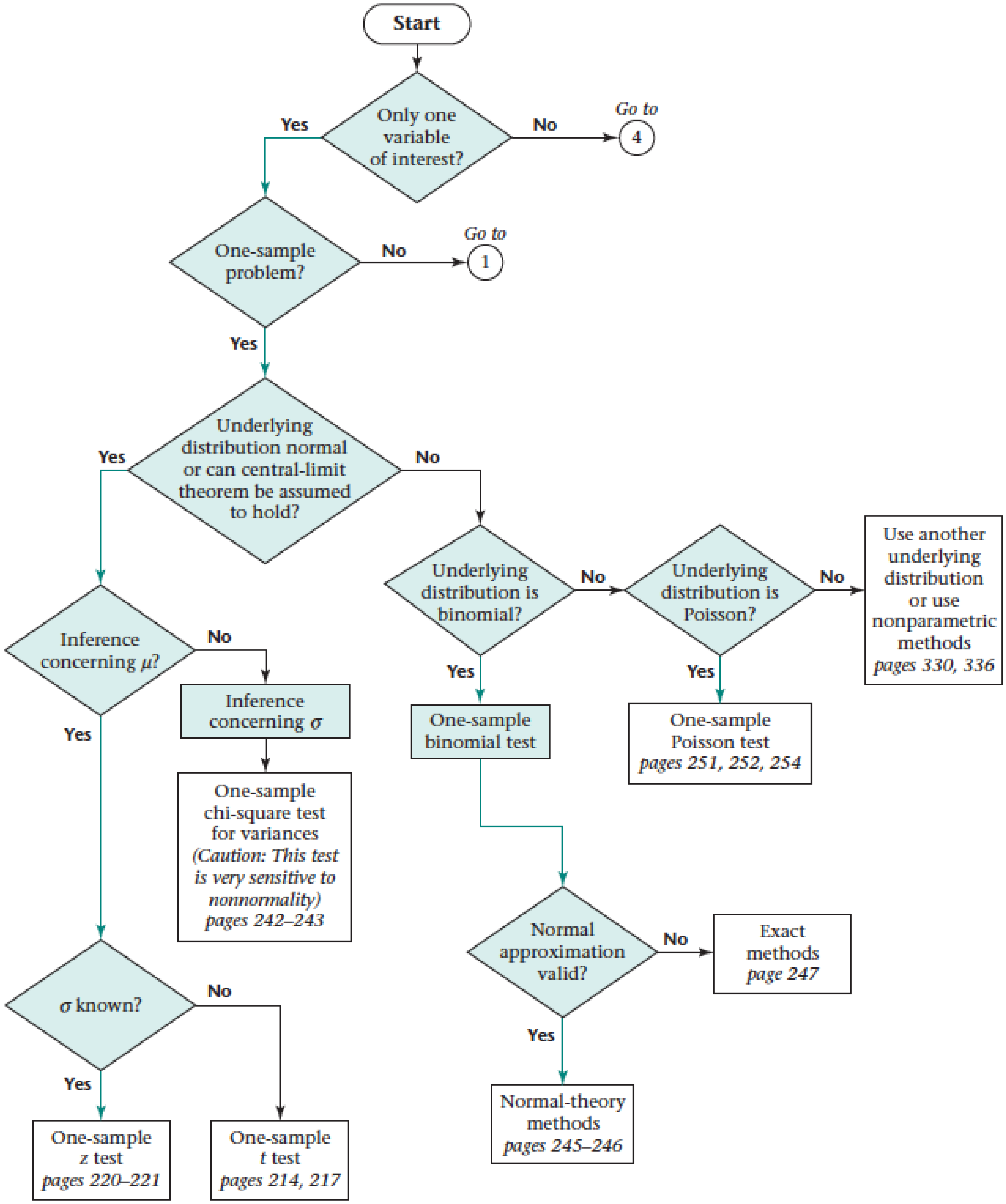
Divide the data into three parts:

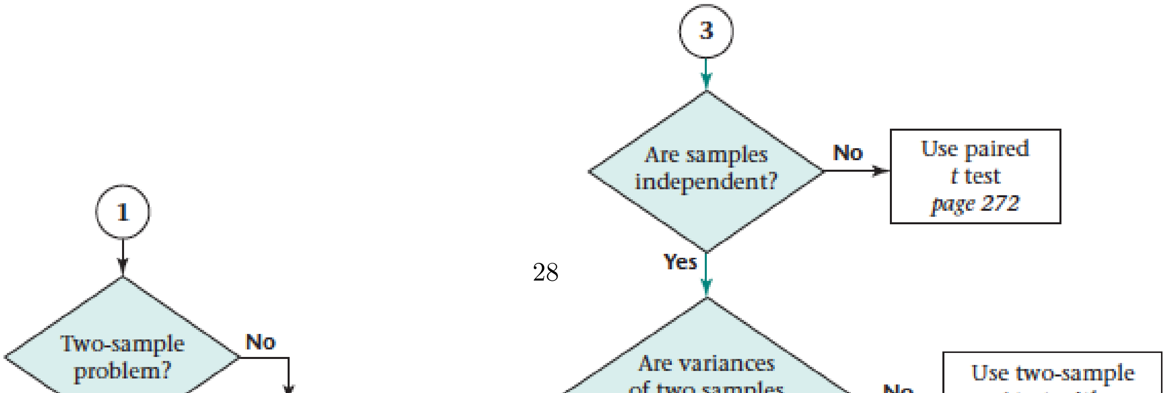
1. Training set: to fit the model
2. Validation set: to estimate hyperparameters and refine the model
3. Test set: to provide an unbiased estimate of predictive capacity

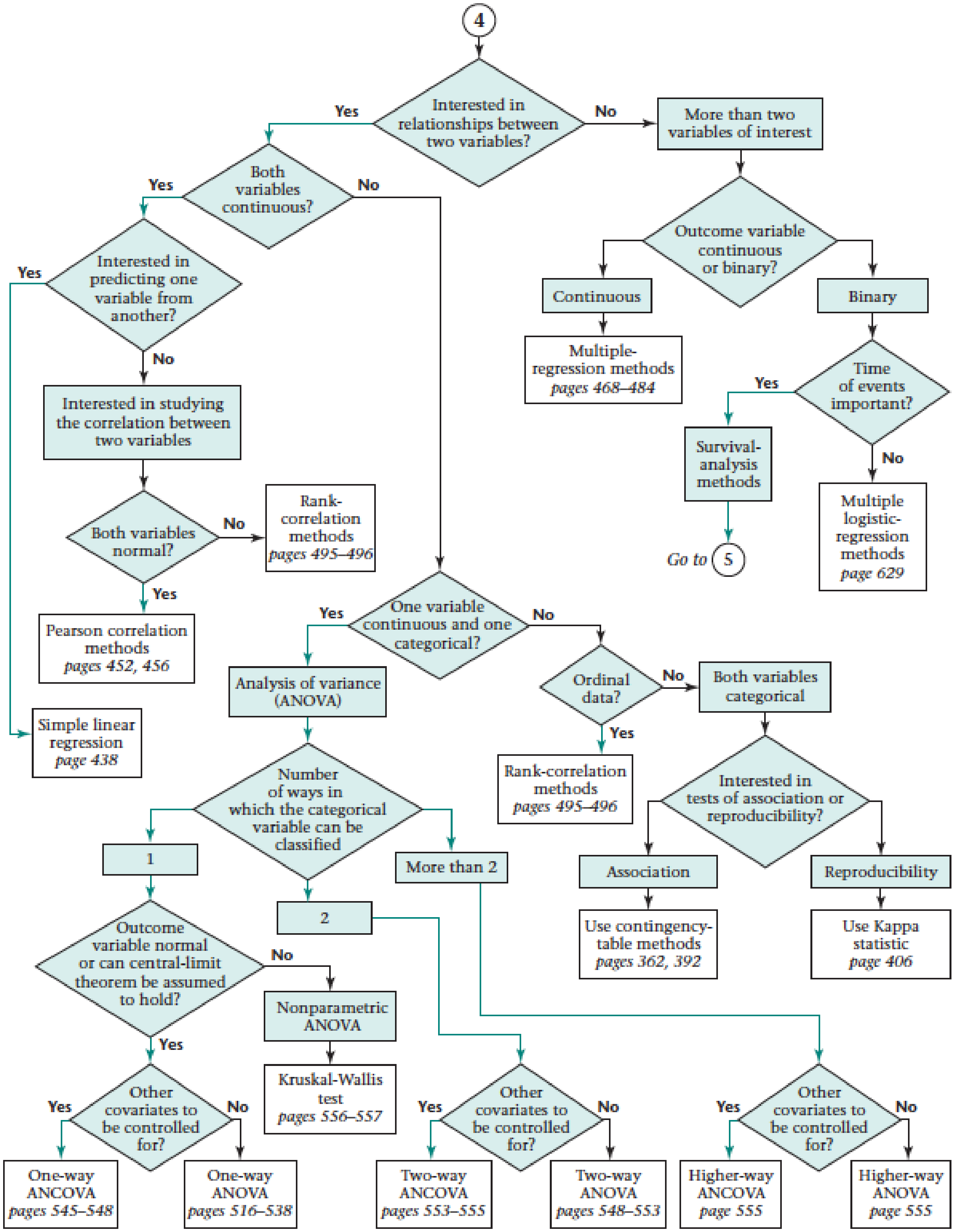
K-fold cross-validation

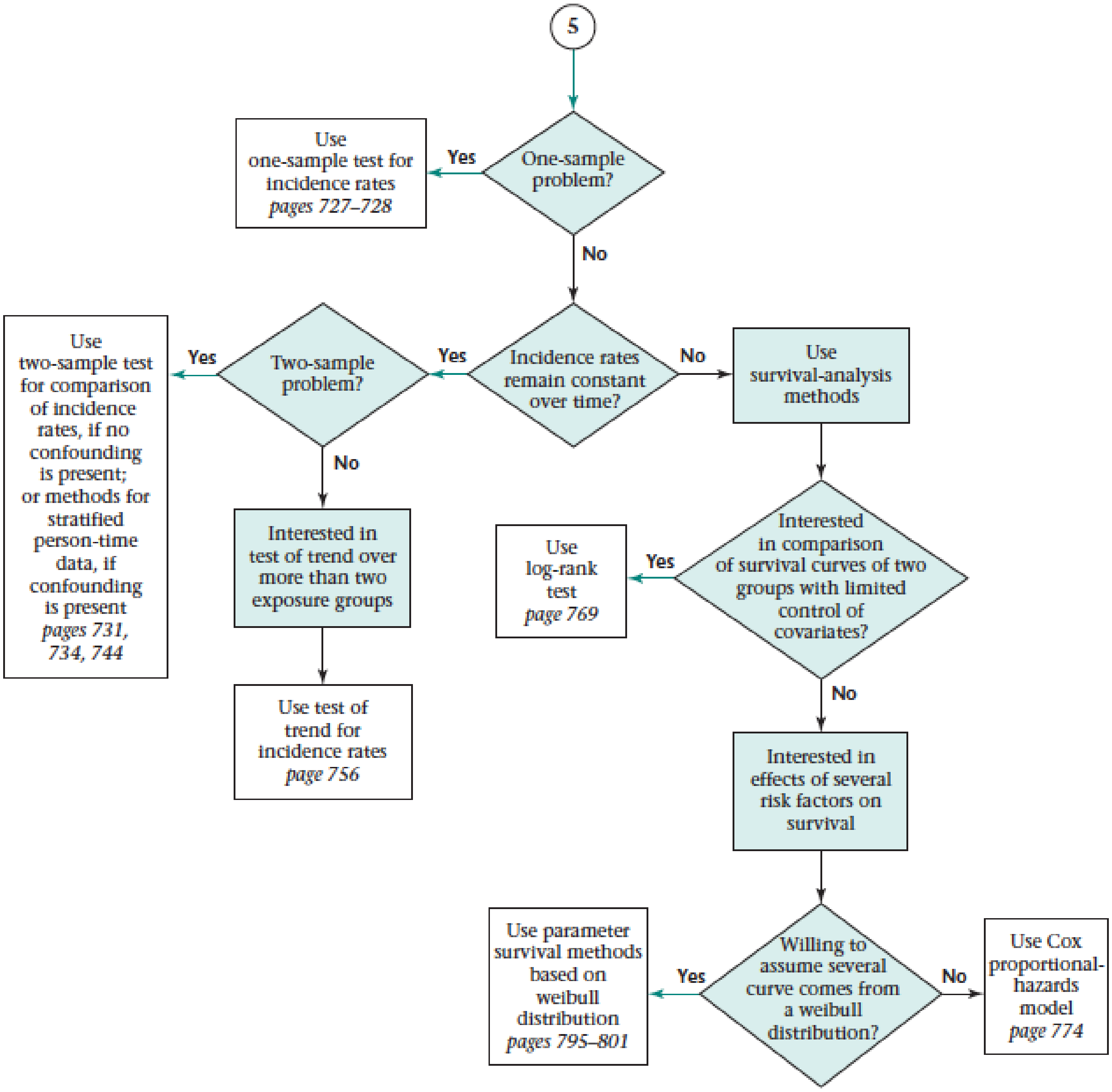
- Split the data into k components
- Conduct cross-validation k times, with a different component as the validation set each time.
- This gives k error measures, which can be combined into mean and variance estimates.

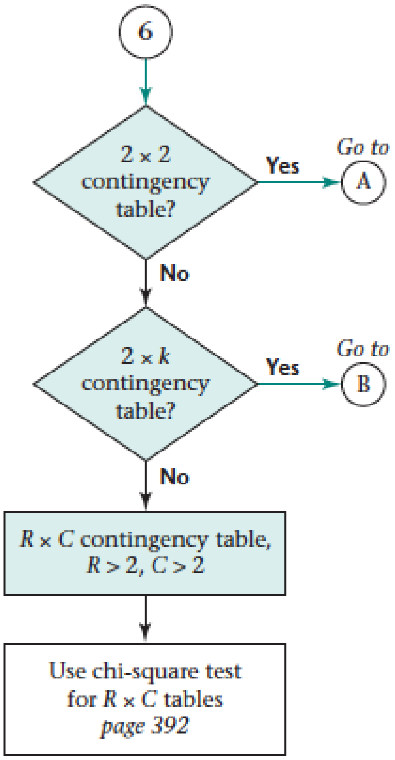
Flowcharts from Fundamentals of Biostatistics (Rosner 7th ed.)

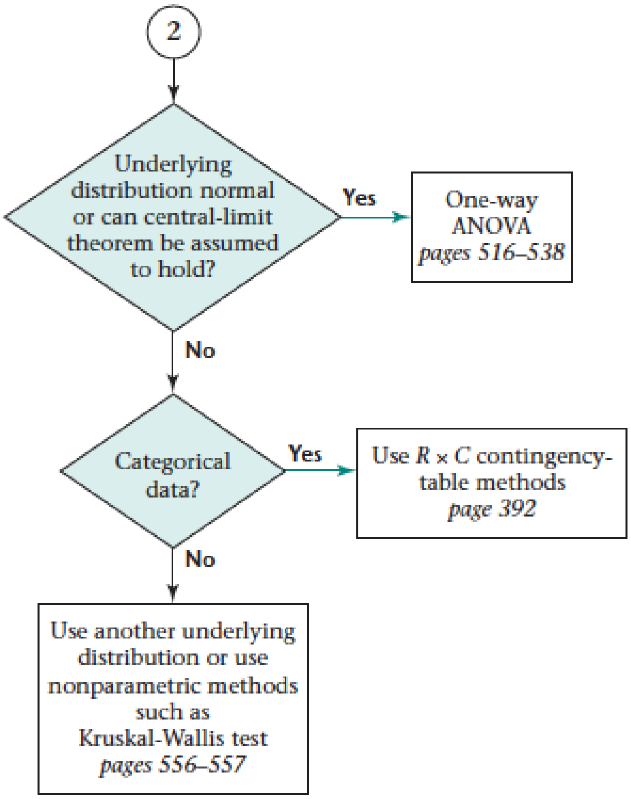

Start
Only one variable of interest?
Yes
No
Go to
4
One-sample problem?
No
Go to
1
Yes
Underlying distribution normal or can central-limit theorem be assumed to hold?
Yes
No
Underlying distribution is binomial?
No
Underlying distribution is Poisson?
No
Use another underlying distribution or use nonparametric methods pages 330, 336
Inference concerning μ?
No
Inference concerning σ
One-sample chi-square test for variances (Caution: This test is very sensitive to nonnormality) pages 242–243
Yes
Yes
One-sample binomial test
Yes
One-sample Poisson test pages 251, 252, 254
Normal approximation valid?
No
Exact methods page 247
Yes
Normal-theory methods pages 245–246
σ known?
No
Yes
One-sample z test pages 220–221
One-sample t test pages 214, 217

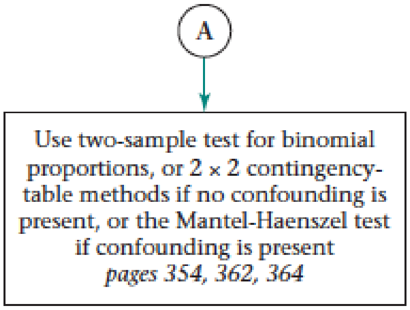


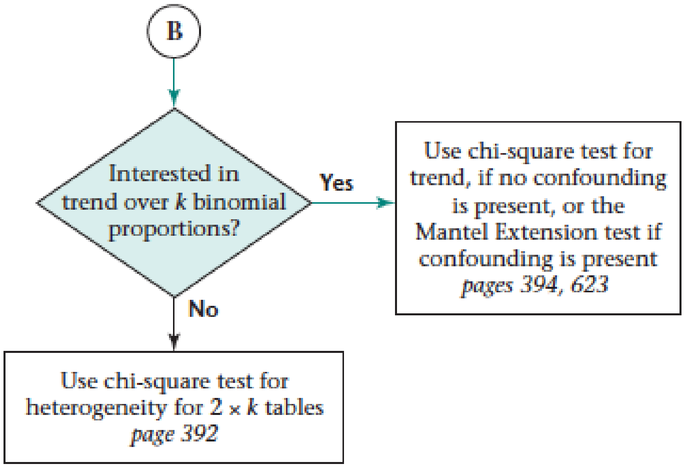

4
Interested in relationships between two variables?
Yes
No
More than two variables of interest
Both variables continuous?
Yes
No
Outcome variable continuous or binary?
Continuous
Binary
Multiple-regression methods pages 468–484
Interested in predicting one variable from another?
Yes
No
Time of events important?
Yes
No
Survival-analysis methods
Go to 5
Multiple logistic-regression methods page 629
Interested in studying the correlation between two variables
Both variables normal?
No
Yes
Rank-correlation methods pages 495–496
Pearson correlation methods pages 452, 456
Simple linear regression page 438
One variable continuous and one categorical?
Yes
No
Analysis of variance (ANOVA)
Ordinal data?
No
Yes
Both variables categorical
Rank-correlation methods pages 495–496
Interested in tests of association or reproducibility?
Association
Reproducibility
Use contingency-table methods pages 362, 392
Use Kappa statistic page 406
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
Yes
Nonparametric ANOVA
Kruskal-Wallis test pages 556–557
Other covariates to be controlled for?
Yes
No
One-way ANCOVA pages 545–548
One-way ANOVA pages 516–538
Other covariates to be controlled for?
Yes
No
Two-way ANCOVA pages 553–555
Two-way ANOVA pages 548–553
Other covariates to be controlled for?
Yes
No
Higher-way ANCOVA page 555
Higher-way ANOVA page 555


5
One-sample problem?
Yes
Use one-sample test for incidence rates *pages 727–728*
No
Incidence rates remain constant over time?
Yes
Two-sample problem?
Yes
Use two-sample test for comparison of incidence rates, if no confounding is present; or methods for stratified person-time data, if confounding is present *pages 731, 734, 744*
No
Interested in test of trend over more than two exposure groups
Use test of trend for incidence rates *page 756*
No
Use survival-analysis methods
Interested in comparison of survival curves of two groups with limited control of covariates?
Yes
Use log-rank test *page 769*
No
Interested in effects of several risk factors on survival
Willing to assume several curve comes from a weibull distribution?
Yes
Use parameter survival methods based on weibull distribution *pages 795–801*
No
Use Cox proportional-hazards model *page 774*


6
2 × 2 contingency table?
Yes
Go to
A
No
2 × k contingency table?
Yes
Go to
B
No
R × C contingency table, R > 2, C > 2
Use chi-square test for R × C tables page 392


2
Underlying distribution normal or can central-limit theorem be assumed to hold?
Yes
One-way ANOVA pages 516–538
No
Categorical data?
Yes
Use R × C contingency-table methods page 392
No
Use another underlying distribution or use nonparametric methods such as Kruskal-Wallis test pages 556–557


A
Use two-sample test for binomial proportions, or 2 × 2 contingency-table methods if no confounding is present, or the Mantel-Haenszel test if confounding is present pages 354, 362, 364


B
Interested in trend over k binomial proportions?
Yes
Use chi-square test for trend, if no confounding is present, or the Mantel Extension test if confounding is present pages 394, 623
No
Use chi-square test for heterogeneity for 2 × k tables page 392