Review of Statistics 101

We review some important themes from the course

1. Introduction

- *Statistics*- Set of methods for collecting/analyzing data (the art and science of learning from data). Provides methods for

- *Design* - Planning/Implementing a study
- *Description* – Graphical and numerical methods for summarizing the data
- *Inference* – Methods for making predictions about a population (total set of subjects of interest), based on a sample

2. Sampling and Measurement

- *Variable* – a characteristic that can vary in value among subjects in a sample or a population.

Types of variables

- *Categorical*
- *Quantitative*
- *Categorical* variables can be *ordinal* (ordered categories) or *nominal* (unordered categories)
- *Quantitative* variables can be *continuous* or *discrete*
- Classifications affect the analysis; e.g., for categorical variables we make inferences about proportions and for quantitative variables we make inferences about means (and use t instead of normal dist.)

Randomization – the mechanism for achieving reliable data by reducing potential bias

Simple random sample: In a sample survey, each possible sample of size *n* has same chance of being selected.

Randomization in a survey used to get a good cross-section of the population. With such *probability sampling* methods, standard errors are valid for telling us how close sample statistics tend to be to population parameters. (Otherwise, the *sampling error* is unpredictable.)

Experimental vs. observational studies

- Sample surveys are examples of **observational studies** (merely observe subjects without any experimental manipulation)
- **Experimental studies**: Researcher assigns subjects to experimental conditions.
  - Subjects should be assigned at random to the conditions (“*treatments*”)
  - Randomization “balances” treatment groups with respect to *lurking* variables that could affect response (e.g., demographic characteristics, SES), makes it easier to assess cause and effect

3. Descriptive Statistics

- Numerical descriptions of *center (mean and median), variability (standard deviation* – typical distance from mean*), position* (*quartiles, percentiles*)
- *Bivariate* description uses regression/correlation (quantitative variable), contingency table analysis such as chi-squared test (categorical variables), analyzing difference between means (quantitative response and categorical explanatory)
- Graphics include *histogram, box plot, scatterplot*

- Mean drawn toward longer tail for skewed distributions, relative to median.

- **Properties of the standard deviation *s*:**
    - *s* increases with the amount of variation around the mean
    - *s* depends on the units of the data (e.g. measure euro vs $)
    - Like mean, affected by outliers
    - *Empirical rule*: If distribution approx. bell-shaped,
    - ➢about 68% of data within 1 std. dev. of mean
    - ➢about 95% of data within 2 std. dev. of mean
    - ➢all or nearly all data within 3 std. dev. of mean

Sample statistics / Population parameters

- We distinguish between summaries of *samples* (**statistics**) and summaries of *populations* (**parameters**).

  Denote statistics by Roman letters, parameters by Greek letters:
- Population mean =μ, standard deviation = σ, proportion π are parameters. In practice, parameter values are unknown, we make inferences about their values using sample statistics.

4. Probability Distributions

Probability: With random sampling or a randomized experiment, the *probability* an observation takes a particular value is the proportion of times that outcome would occur in a long sequence of observations.

Usually corresponds to a *population proportion* (and thus falls between 0 and 1) for some real or conceptual population.

A *probability distribution* lists all the possible values and their probabilities (which add to 1.0)

Like frequency dist's, probability distributions have mean and standard deviation

$$\mu = E(Y) = \sum yP(y)$$

Standard Deviation - Measure of the "typical" distance of an outcome from the mean, denoted by σ

If a distribution is approximately normal, then:

- all or nearly all the distribution falls between
  μ - 3σ and μ + 3σ

- Probability about 0.68 falls between
  μ - σ and μ + σ

Normal distribution

- Symmetric, bell-shaped (formula in Exercise 4.56)
- Characterized by mean (μ) and standard deviation (σ), representing center and spread
- Prob. within any particular number of standard deviations of μ is same for all normal distributions
- An individual observation from an approximately normal distribution satisfies:
  - Probability 0.68 within 1 standard deviation of mean
  - 0.95 within 2 standard deviations
  - 0.997 (virtually all) within 3 standard deviations

Notes about z-scores

- z-score represents *number of standard deviations* that a value falls from mean of dist.

- A value y is $z = (y - \mu)/\sigma$ standard deviations from μ

- The **standard normal distribution** is the normal dist with $\mu = 0$, $\sigma = 1$ (used as sampling dist. for *z* test statistics in significance tests)

- In inference we use *z* to count the *number of standard errors* between a sample estimate and a null hypothesis value.

$\bar{y}$

Sampling dist. of sample mean

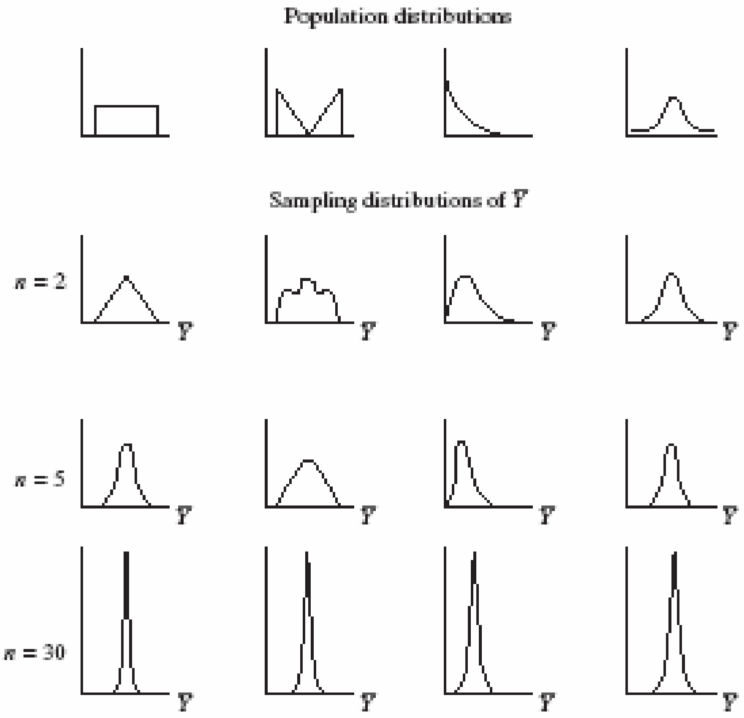
- $\bar{y}$ is a variable, its value varying from sample to sample about population mean μ. S**ampling distribution** of a statistic is the probability distribution for the possible values of the statistic
- Standard deviation of sampling dist of $\bar{y}$ is called the **standard error of** $\bar{y}$
- For random sampling, the sampling dist of $\bar{y}$

  has mean μ and standard error

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{popul. std. dev.}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem: For random sampling with "large" *n,* sampling dist of sample mean $\bar{y}$ is approximately a normal distribution

- Approx. normality applies *no matter what the shape* of the popul. dist. (Figure p. 93, next page)
- How "large" *n* needs to be depends on skew of population dist, but usually *n ≥ 30* sufficient
- Can be verified empirically, by simulating with "sampling distribution" applet at www.prenhall.com/agresti. Following figure shows how sampling dist depends on *n* and shape of population distribution.


Population distributions
Sampling distributions of $\bar{Y}$
$n = 2$
$n = 5$
$n = 30$
$\bar{Y}$

5. Statistical Inference: Estimation

Point estimate: A single statistic value that is the "best guess" for the parameter value (such as sample mean as point estimate of popul. mean)

Interval estimate: An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a ***confidence interval***.

(Based on sampling dist. of the point estimate, has form point estimate plus and minus a margin of error that is a *z* or *t* score times the standard error)

Conf. Interval for a Proportion (in a particular category)

- Sample proportion $\hat{\pi}$ is a mean when we let y=1 for observation in category of interest, y=0 otherwise
- Population prop. is mean μ of prob. dist having

$$P(1) = \pi \text{ and } P(0) = 1 - \pi$$

- The standard dev. of this prob. dist. is

$$\sigma = \sqrt{\pi(1-\pi)} \text{ (e.g., 0.50 when } \pi = 0.50)$$

- The standard error of the sample proportion is

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

Finding a CI in practice

- Complication: The true standard error

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

itself depends on the unknown parameter!

In practice, we estimate

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ by } se = \sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}}$$

and then find 95% CI using formula

$$\hat{\pi} - 1.96(se) \text{ to } \hat{\pi} + 1.96(se)$$

Confidence Interval for the Mean

- In large samples, the sample mean has approx. a normal sampling distribution with mean μ and standard error

$$\sigma_{\bar{y}} = \sigma / \sqrt{n}$$

- Thus,

$$P(\mu - 1.96\sigma_{\bar{y}} \le \bar{y} \le \mu + 1.96\sigma_{\bar{y}}) = .95$$

- We can be 95% confident that the sample mean lies within 1.96 standard errors of the (unknown) population mean

- Problem: Standard error is unknown (σ is also a parameter). It is estimated by replacing σ with its point estimate from the sample data:

$$se = \frac{s}{\sqrt{n}}$$

95% confidence interval for μ :

$$\bar{y} \pm 1.96(se), \text{ which is } \bar{y} \pm 1.96\frac{s}{\sqrt{n}}$$

This works ok for “large *n*,” because *s* then a good estimate of σ (and CLT). But for small *n*, replacing σ by its estimate *s* introduces extra error, and CI is not quite wide enough unless we replace *z*-score by a slightly larger “*t*-score.”

The *t* distribution (Student's *t*)

- Bell-shaped, symmetric about 0
- Standard deviation a bit larger than 1 (slightly thicker tails than standard normal distribution, which has mean = 0, standard deviation = 1)
- Precise shape depends on **degrees of freedom** (df). For inference about mean,

  $df = n - 1$

- More closely resembles standard normal dist. as *df* increases

  (nearly identical when $df > 30$)

- CI for mean has margin of error *t*(*se*)

CI for a population mean

- For a random sample *from a normal population distribution*, a 95% CI for μ is

$$\bar{y} \pm t_{.025}(se), \text{ with } se = s/\sqrt{n}$$

where *df* = *n-1* for the *t*-score

- Normal population assumption ensures sampling dist. has bell shape for *any n* (Recall figure on p. 93 of text and next page). Method is *robust* to violation of normal assumption, more so for large *n* because of CLT.

6. Statistical Inference: Significance Tests

A **significance test** uses data to summarize evidence about a hypothesis by comparing sample estimates of parameters to values predicted by the hypothesis.

We answer a question such as, “If the hypothesis were true, would it be unlikely to get estimates such as we obtained?”

Five Parts of a Significance Test

- **Assumptions** about type of data (quantitative, categorical), sampling method (random), population distribution (binary, normal), sample size (large?)
- **Hypotheses**:

Null hypothesis (H_0): A statement that parameter(s) take specific value(s) (Often: “no effect”)

Alternative hypothesis (H_a): states that parameter value(s) in some alternative range of values

- **Test Statistic**: Compares data to what null hypo. H_0 predicts, often by finding the number of standard errors between sample estimate and H_0 value of parameter
- ***P*-value (*P*):** A probability measure of evidence about H_0, giving the probability (under presumption that H_0 true) that the test statistic equals observed value or value even more extreme in direction predicted by H_a.
  - The smaller the *P*-value, the stronger the evidence against $H_{0.}$
- **Conclusion**:
  - If no decision needed, report and interpret P-value

- If decision needed, select a cutoff point (such as 0.05 or 0.01) and reject H_0 if P-value ≤ that value
- The most widely accepted minimum level is 0.05, and the test is said to be *significant at the .05 level* if the P-value ≤ 0.05.
- If the *P*-value is not sufficiently small, we fail to reject H_0 (not necessarily true, but plausible). We should *not* say "Accept H_0"
- The cutoff point, also called the *significance level* of the test, is also the prob. of Type I error – i.e., if null true, the probability we will incorrectly reject it.
- Can't make significance level *too* small, because then run risk that P(Type II error) = P(do not reject null) when it is false too large

Significance Test for Mean

- *Assumptions*: Randomization, quantitative variable, normal population distribution
- *Null Hypothesis*: H_0: $\mu = \mu_0$ where μ_0 is particular value for population mean (typically no effect or change from standard)
- *Alternative Hypothesis*: H_a: $\mu \neq \mu_0$ (*2-sided* alternative includes both > and <), or one-sided
- *Test Statistic*: The number of standard errors the sample mean falls from the H_0 value

$$t = \frac{\bar{y} - \mu_0}{se} \text{ where } se = s / \sqrt{n}$$

Effect of sample size on tests

- With large *n* (say, *n* > 30), assumption of normal population dist. not important because of Central Limit Theorem.
- For small *n,* the *two-sided t* test is robust against violations of that assumption. One-sided test is *not* robust.
- For a given observed sample mean and standard deviation, the larger the sample size *n*, the larger the test statistic (because *se* in denominator is smaller) and the smaller the *P*-value. (i.e., we have more evidence with more data)
- We’re more likely to reject a false H_0 when we have a larger sample size (the test then has more “power”)
- With large *n*, “statistical significance” not the same as “practical significance”. Should also find CI to see how far parameter may fall from H_0

Significance Test for a Proportion π

- Assumptions:
  - Categorical variable
  - Randomization
  - Large sample (but two-sided ok for nearly all *n*)
- Hypotheses:
  - Null hypothesis: H_0: $\pi = \pi_0$
  - Alternative hypothesis: H_a: $\pi \neq \pi_0$ (2-sided)
  - H_a: $\pi > \pi_0$ H_a: $\pi < \pi_0$ (1-sided)
  - (choose before getting the data)

- Test statistic:

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

- Note $\sigma_{\hat{\pi}} = se_0 = \sqrt{\pi_0(1-\pi_0)/n}$, not $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ as in a CI

- As in test for mean, test statistic has form

(estimate of parameter – null value)/(standard error)

= no. of standard errors estimate falls from null value

- *P*-value:

  H_a: $\pi \neq \pi_0$ *P* = 2-tail prob. from standard normal

  H_a: $\pi > \pi_0$ *P* = right-tail prob. from std. normal

  H_a: $\pi < \pi_0$ *P* = left-tail prob. from std. normal

- Conclusion: As in test for mean (e.g., reject H_0 if *P*-value ≤ α)

Error Types

- Type I Error: Reject H_0 when it is true
- Type II Error: Do not reject H_0 when it is false

Test Result –	Reject H_0	Don’t Reject H_0
True State		
H_0 True	Type I Error	Correct
H_0 False	Correct	Type II Error

Limitations of significance tests

- *Statistical significance* does not mean *practical significance*
- Significance tests don’t tell us about the *size* of the effect (like a CI does)
- Some tests may be “statistically significant” just by chance (and some journals only report “significant” results)

Example: Many medical “discoveries” are really Type I errors (and true effects are often much weaker than first reported). Read Example 6.8 on p. 165 of text.

Chap. 7. Comparing Two Groups

Distinguish between response and explanatory variables, independent and dependent samples

Comparing means is bivariate method with quantitative response variable, categorical (binary) explanatory variable

Comparing proportions is bivariate method with categorical response variable, categorical (binary) explanatory variable

se for difference between two estimates (independent samples)

- The sampling distribution of the difference between two estimates (two sample proportions or two sample means) is *approximately normal* (large n_1 and n_2) and has estimated

$$se = \sqrt{(se_1)^2 + (se_2)^2}$$

CI comparing two proportions

- Recall *se* for a sample proportion used in a CI is

$$se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

- So, the *se* for the difference between sample proportions for two independent samples is

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

- A CI for the difference between population proportions is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

(as usual, *z* depends on confidence level, 1.96 for 95% conf.)

Quantitative Responses: Comparing Means

- Parameter: μ_2-μ_1
- Estimator: $\bar{y}_2 - \bar{y}_1$
- Estimated standard error: $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - Sampling dist.: Approx. normal (large *n's,* by CLT), get approx. *t* dist. when substitute estimated std. error in *t* stat.
  - CI for independent random samples *from two normal population distributions* has form

$$\left(\bar{y}_2 - \bar{y}_1\right) \pm t(se), \text{ which is } \left(\bar{y}_2 - \bar{y}_1\right) \pm t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - Alternative approach assumes equal variability for the two groups, is special case of ANOVA for comparing means in Chapter 12

Comments about CIs for difference between two parameters

- When 0 is not in the CI, can conclude that one population parameter is higher than the other.

(e.g., if all positive values when take Group 2 – Group 1, then conclude parameter is higher for Group 2 than Group 1)

- When 0 is in the CI, it is plausible that the population parameters are identical.

Example: Suppose 95% CI for difference in population proportion between Group 2 and Group 1 is (-0.01, 0.03)

Then we can be 95% confident that the population proportion was between about 0.01 *smaller* and 0.03 *larger* for Group 2 than for Group 1.

Comparing Means with Dependent Samples

- Setting: Each sample has the same subjects (as in longitudinal studies or crossover studies) or matched pairs of subjects

- Data: y_{i} = difference in scores for subject (pair) *i*

- Treat data as single sample of difference scores, with sample mean $\bar{y}_d$ and sample standard deviation s_d and parameter μ_{d} = population mean difference score which equals difference of population means.

Chap. 8. Association between Categorical Variables

- Statistical analyses for when both response and explanatory variables are *categorical.*

- **Statistical independence (no association)**: Population conditional distributions on one variable the same for all categories of the other variable

- **Statistical dependence (association)**: Population conditional distributions are not all identical

Chi-Squared Test of Independence (Karl Pearson, 1900)

- Tests H_0: variables are statistically independent
- H_a: variables are statistically dependent
- Summarize closeness of observed cell counts $\{f_o\}$ and expected frequencies $\{f_e\}$ by

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

with sum taken over all cells in table.

- Has chi-squared distribution with *df = (r-1)(c-1)*

- For 2-by-2 tables, chi-squared test of independence (*df* = 1) is equivalent to testing H_0: $\pi_1 = \pi_2$ for comparing two population proportions.

	Proportion	
Population	Response 1	Response 2
1	π_1	$1 - \pi_1$
2	π_2	$1 - \pi_2$

H_0: $\pi_1 = \pi_2$ equivalent to

H_0: response independent of population

Then, chi-squared statistic (*df* = 1) is square of *z* test statistic,

z = (difference between sample proportions)/se_0.

Residuals: Detecting Patterns of Association

- Large chi-squared implies *strong evidence* of association but does not tell us about *nature* of assoc. We can investigate this by finding the *standardized residual* in each cell of the contingency table,

$$z = (f_o - f_e)/se,$$

Measures number of standard errors that $(f_o - f_e)$ falls from value of 0 expected when H_0 true.

- Informally inspect, with values larger than about 3 in absolute value giving evidence of *more* (positive residual) or *fewer* (negative residual) subjects in that cell than predicted by independence.

Measures of Association

- Chi-squared test answers “Is there an association?”

- Standardized residuals answer “How do data differ from what independence predicts?”

- We answer “How strong is the association?” using a measure of the strength of association, such as the difference of proportions, the relative risk = ratio of proportions, and the odds ratio, which is the ratio of odds, where

  odds = probability/(1 – probability)

Limitations of the chi-squared test

- The chi-squared test merely analyzes the extent of evidence that there is an association (through the *P*-value of the test)

- Does not tell us the *nature* of the association (standardized residuals are useful for this)

- Does not tell us the *strength* of association. (e.g., a large chi-squared test statistic and small *P*-value indicates strong evidence of assoc. but not necessarily a strong association.)

Ch. 9. Linear Regression and Correlation

Data: *y* – *a* quantitative response variable

x – a quantitative explanatory variable

We consider:

- Is there an association? (test of *independence* using slope)
- How strong is the association? (uses *correlation r* and *r*^{2})
- How can we predict *y* using *x*? (estimate a *regression equation)*

Linear *regression equation* E(*y*) = α + β *x* describes how mean of conditional distribution of *y* changes as *x* changes

Least squares estimates this and provides a sample *prediction equation* $\hat{y} = a + bx$

- The linear regression equation *E(y)* = $\alpha + \beta x$ is part of a *model.* The model has another parameter σ that describes the variability of the conditional distributions; that is, the variability of y values for all subjects having the same *x*-value.

- For an observation, difference $y - \hat{y}$ between observed value of *y* and predicted value $\hat{y}$ of *y,* is a **residual** (vertical distance on scatterplot)

- Least squares method mimimizes the sum of squared residuals (errors), which is SSE used also in r^2 and the estimate *s* of conditional standard deviation of *y*

Measuring association: The correlation and its square

- The correlation is a *standardized* slope that does not depend on units
- *Correlation r* relates to slope *b* of prediction equation by

$$r = b(s_x/s_y)$$

- $-1 \leq r \leq +1$, with *r* having same sign as *b* and *r* = 1 or -1 when all sample points fall exactly on prediction line, so *r* describes *strength of linear association*
- The larger the absolute value, the stronger the association
- Correlation implies that predictions *regress toward the mean*

- The *proportional reduction in error* in using *x* to predict *y* (via the prediction equation) instead of using sample mean of *y* to predict *y* is

$$r^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y-\bar{y})^2 - \Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$$

- Since -1 ≤ r ≤ +1, 0 ≤ r^2 ≤ 1, and r^2 = 1 when all sample points fall exactly on prediction line
- *r* and r^2 do not depend on units, or distinction between *x, y*
- The *r* and r^2 values tend to weaken when we observe *x* only over a restricted range, and they can also be highly influenced by outliers.

Inference for regression model

- Parameter: Population slope in regression model (β)
- H_0: independence is H_0: $\beta = 0$
- Test statistic $t = (b - 0)/se$, with $df = n - 2$
- A CI for β has form $b \pm t(se)$

where *t-score* has $df = n-2$ and is from *t*-table with half the error probability in each tail. (Same *se* as in test)

- In practice, CI for multiple of slope may be more relevant (find by multiplying endpoints by the relevant constant)
- CI not containing 0 equivalent to rejecting H_0 (when error probability is same for each)

Software reports SS values (SSE, regression SS, TSS = regression SS + SSE) and the test results in an ANOVA (analysis of variance) table

The *F* statistic in the ANOVA table is the square of the *t* statistic for testing H_0: $\beta = 0$, and it has the same P-value as for the two-sided test.

We need to use *F* when we have *several* parameters in H_0 , such as in testing that all β parameters in a multiple regression model = 0 (which we did in Chapter 11)

Chap. 10. Introduction to Multivariate Relationships

Bivariate analyses informative, but we usually need to take into account *many* variables.

- Many explanatory variables have an influence on any particular response variable.
- The effect of an explanatory variable on a response variable may change when we take into account other variables. (Recall admissions into Berkeley example)
- When each pair of variables is associated, then a bivariate association for two variables may differ from its "partial" association, controlling for another variable

- Association does not imply causation!

- With observational data, effect of X on Y may be partly due to association of X and Y with other *lurking variables*.

- *Experimental* studies have advantage of being able to control potential lurking variables (groups being compared should be roughly “balanced” on them).

- When X_1 and X_2 both have effects on Y but are also associated with each other, there is *confounding.* It’s difficult to determine whether either truly causes Y, because a variable’s effect could be at least partially due to its association with the other variable.

- *Simpson's paradox*: It is possible for the (bivariate) association between two variables to be positive, yet be negative at each fixed level of a third variable (or reverse)
- *Spurious association: Y* and X_1 both depend on X_2 and association disappears after controlling X_2
- *Multiple causes* more common, in which explanatory variables have associations among themselves as well as with response var. Effect of any one changes depending on what other variables controlled (statistically), often because it has a *direct* effect and also *indirect* effects.
- *Statistical interaction* – Effect of X_1 on Y changes as the level of X_2 changes.

Chap. 11. Multiple Regression

- *y* – response variable

  $x_1, x_2, \ldots, x_k$ -- set of explanatory variables

All variables assumed to be quantitative (later chapters incorporate categorical variables in model also)

Multiple regression equation (population)*:*

$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_k x_k$

- Controlling for other predictors in model, there is a linear relationship between *E(y)* and x_1 with slope $\beta_{1.}$

- Partial effects in multiple regression refer to statistically *controlling* other variables in model, so differ from effects in bivariate models, which ignore *all* other variables.

- Partial effect of a predictor in multiple regression is identical at all fixed values of other predictors in model (assumption of “no interaction”)

- Again, this is a *model.* We fit it using least squares, minimizing SSE out of all equations of the assumed form. The model may not be appropriate (e.g., if there is severe interaction).

- Graphics include scatterplot matrix (corresponding to correlation matrix), partial regression plots

Multiple correlation and *R*^{2}

- The ***multiple correlation*** *R* is the correlation between the observed *y*-values and predicted *y*-values.
- ***R*^{2}** is the proportional reduction in error from using the prediction equation (instead of sample mean) to predict *y*

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y-\overline{y})^2 - \Sigma(y-\hat{y})^2}{\Sigma(y-\overline{y})^2}$$

- $0 \leq R^2 \leq 1$ and $0 \leq R \leq 1$.
- *R*^{2} cannot decrease (and SSE cannot increase) when predictors are added to a regression model
- The numerator of *R*^{2} (namely, TSS – SSE) is the *regression sum of squares,* the variability in *y* “explained” by the regression model.

Inference for multiple regression model

- To test whether explanatory variables collectively have effect on *y*, we test

  $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

Test statistic

$$F = \frac{R^2 / k}{(1 - R^2)/[n - (k+1)]}$$

- When H_0 true, *F* values follow the *F* distribution

  $df_1 = k$ (no. of predictors in model)

  $df_2 = n - (k+1)$ (sample size – no. model parameters)

Inferences for individual regression coefficients

- To test partial effect of x_i controlling for the other explan. var's in model, test H_0: $\beta_i = 0$ using test stat.

  $$t = (b_i - 0)/se, \ df = n-(k+1)$$

- CI for β_i has form $b_i \pm t(se)$, with t-score also having

  $df = n-(k+1)$, for the desired confidence level

- Partial t test results can seem logically inconsistent with result of F test, when explanatory variables are highly correlated

Modeling interaction

The multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_k x_k$$

assumes the partial slope relating *y* to each x_i is the same at all values of other predictors

Model allowing interaction (e.g., for 2 predictors),

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2)$$

$$= (\alpha + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

is special case of multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

with $x_3 = x_1 x_2$

Chap. 12: Comparing Several Groups (ANOVA)

Classification of bivariate methods:

Response *y*	Explanatory *x* var’s	Method
Categorical	Categorical	Contingency tables (Ch. 8) (chi-squared, etc.)
Quantitative	Quantitative	Regression and correlation (Ch 9 bivariate, 11 multiple regr.)
Quantitative	Categorical	ANOVA (Ch. 12)

Ch. 12 compares the mean of *y* for the groups corresponding to the categories of the categorical explanatory variables.

Comparing means across categories of one classification (1-way ANOVA)

- The *analysis of variance* (ANOVA) is an *F* test of

  H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$
  H_a: The means are not all identical

- The *F* test statistic is large (and *P*-value is small) if variability *between* groups is large relative to variability *within* groups

- *F* statistic has mean about 1 when null true

Follow-up Comparisons of Pairs of Means

- A CI for the difference (μ_i -μ_j) is

$$\left(\bar{y}_i - \bar{y}_j\right) \pm t\,s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where *s* is square root of within-groups variance estimate.

Multiple comparisons: Obtain confidence intervals for all pairs of group mean difference, with fixed probability that *entire set* of CI's is correct.

- The *Bonferroni* approach does this by dividing the overall desired error rate by the number of comparisons to get error rate for each comparison

Regression Approach To ANOVA

- *Dummy (indicator) variable*: Equals 1 if observation from a particular group, 0 if not.
- Regression model: $E(y) = \alpha + \beta_1 z_1 + ... + \beta_{g-1} z_{g-1}$ (e.g., $z_1 = 1$ for subjects in group 1, = 0 otherwise)
- Mean for group *i* ($i = 1, ..., g - 1$): $\mu_i = \alpha + \beta_i$
- Mean for group *g*: $\mu_g = \alpha$
- Regression coefficient $\beta_i = \mu_i - \mu_g$ compares each mean to mean for last group
- 1-way ANOVA: H_0: $\mu_1 = \ldots = \mu_g$ corresponds in regression to testing H_0: $\beta_1 = ... = \beta_{g-1} = 0$.

Two-way ANOVA

- Analyzes relationship between quantitative response *y* and *two* categorical explanatory factors.
- A *main effect* hypothesis states that the means are equal across levels of one factor, within levels of the other factor.
- First test H0: *no interaction.* Testing main effects only sensible if there is no significant interaction; i.e., effect of each factor is the same at each category for the other factor.
- You should be able to give examples of population means that have no interaction and means that show a main effect without an interaction.