Statistics 2010: Elements of Statistics for Business and Economics

Winter 2005
Prof. Jaimie Kwon
Dept of Statistics
Cal State East Bay

Logistics

- Grades
  - Midterm(s), final
  - Quiz
  - Homeworks
- Computer (MS excel, occasional Minitab and R)

Objectives

- Understand the key paradigms of statistics; Keywords include:
  - Population, parameter
  - Sample, statistic
  - 100(1-α)% Confidence interval
  - Statistical hypothesis testing
  - P-value
  - Linear regression
- Learn to:
  - Identify the relevant technique(s) to use for the given data or question
  - Compute basic statistics and run simple tests
  - Interpret the results

Chapter 1. What is statistics?

- Statistics
  - Descriptive statistics
  - Inferential statistics
- Population vs. sample paradigm
  - Parameter vs. statistic

Chapter 2. Graphical and tabular descriptive statistics

- Variable
  - Interval (quantitative, numerical)
  - Nominal (categorical)
  - Ordinal
- Nominal data
  - Frequency
  - Relative frequency
- Interval data
  - Histogram: class intervals; Sturges' formula
  - # of Peaks? Skewness?
  - Stem-and-leaf plot

Graphical and tabular descriptive statistics (continued)

- Bivariate variables
  - Two nominal variables
    - Contingency table
  - Two interval variables
    - Scatter diagrams
    - Linear relationship (direction and strength)
- Time series data

Chapter 4. Numerical descriptive techniques

- Measures of central location
  - Mean, median, mode
- Measures of variability
  - Range, variance, sd
- Percentiles and quartiles
- Outliers
- Boxplot
- Measures of linear relationship
  - Covariance
  - Correlation
  - Least squares line (slope and intercept)

Chapter 5. Data collection and sampling

- Data collection methods
  - Observational
  - Experimental
  - Survey
- Sampling
  - Simple random sample
  - And others...

Chapter 6. Probability

- Random experiment
- Outcomes
- Sample space
- Events
- Three approaches to assigning probabilities
  - The classical approach
  - The relative frequency approach
  - The subjective approach
- Joint, marginal, conditional probability
- Independence of events
- Union, intersection, complement
- Multiplication rule, addition rule, Bayes' rule

Chapter 7. Random variables and discrete probability distributions

- Random variable
  - Discrete
  - Continuous
- Probability distribution P(x)
- E(X), V(X),...
  - E(cX+d)=...
- Bivariate distribution P(x,y)
- Cov(X,Y), ρ,...
- Cumulative probability
- Binomial experiment and bin(n,p)
- Poisson distribution

Chapter 8. Continuous probability distributions

- Probability density functions
- Uniform(a,b) distribution
- $X \sim N(\mu, \sigma^2)$
- $Z=(X-\mu)/\sigma \sim N(0,1)$: "standard normal"
  - z_α such that $P(Z>z_\alpha)=\alpha$: "(upper) critical value"
- Student t distribution
  - $T \sim t(\nu)$ (ν = "degrees of freedom")
  - $t_\alpha(\nu)$ such that $P(T> t_\alpha(\nu))$: "(upper) critical value"
- Misc. distributions
  - Chi-squared distribution
  - F distribution

Chapter 9. Sampling distributions

- X_i~some distribution with
  - Mean μ_X
  - Variance σ_X^2
- Sampling distribution of the sample mean of a random sample of size n
  - Mean: same as μ_X
  - Variance: σ_X^2 divided by n
  - Approximately normal if n is large, say n≥30
  - Exactly normal if X_i is normally distributed
- Sampling distribution of the sample proportion (p-hat)
- Sampling distribution of the difference between two means (x1-bar – x2-bar)

Chapter 10. Estimation

- Point vs. interval estimator
- $Xi \sim N(\mu,\sigma^2)$ for known σ
  - $100(1-\alpha)\%$ Confidence interval for μ:
  - confidence level $(1-\alpha)$
  - Interpretation?
- Determining the sample size for the target accuracy W

Chapter 11. Hypothesis testing

- Null (H_0) and alternative (H_1) hypotheses
- Type I and type II error
- Test statistic
- "Rejection region" for "significance level" α
- P-value of a test: "the probability of observing a test statistic at least as extreme as the one observed, given that H_0 is true"
  - Better than simply reporting "accept/reject at α"
- Power of a test = 1-P(Type II error)

Hypothesis testing for μ, when σ is known (z-test)

- $Xi \sim N(\mu,\sigma^2)$ for known σ
  - Hypothesis
    - Case1: $H_0:\mu \le \mu_0$ vs. $H_1:\mu > \mu_0$
    - Case2: $H_0:\mu \ge \mu_0$ vs. $H_1:\mu < \mu_0$
    - Case3: $H_0:\mu = \mu_0$ vs. $H_1:\mu \ne \mu_0$
  - Test Statistic: z=...
  - Test of significance level α rejects H_0 if
    - Case1: $z>z_\alpha$
    - Case2: $z<-z_\alpha$
    - Case3: $|z|>z_{\alpha/2}$
  - P-value
    - Case1: $P(Z \ge z)$
    - Case2: $P(Z \le z)$
    - Case3: $2\,P(Z \ge |z|)$

Hypothesis testing for μ, when σ is unknown (t-test; chapter 12)

- $Xi \sim N(\mu,\sigma^2)$ for unknown σ
  - Hypothesis
    - Case1: $H_0:\mu \le \mu_0$ vs. $H_1:\mu > \mu_0$
    - Case2: $H_0:\mu \ge \mu_0$ vs. $H_1:\mu < \mu_0$
    - Case3: $H_0:\mu = \mu_0$ vs. $H_1:\mu \ne \mu_0$
  - Test Statistic: t=...
  - Test of significance level α rejects H_0 if
    - Case1: $t>t_\alpha(n-1)$:
    - Case2: $t<-t_\alpha(n-1)$
    - Case3: $|t|>t_{\alpha/2}(n-1)$
  - P-value
    - Case1: $P(T(n-1) \ge t)$
    - Case2: $P(T(n-1) \le t)$
    - Case3: $2\,P(T(n-1) \ge |t|)$

Chapter 12. Inference about a population

- Inference about μ when σ is unknown
  - $100(1-\alpha)\%$ Confidence Interval for μ
  - The t-test
- The t-test is a "robust" method
  - Checking assumptions
- Inference about σ: skip
- Inference about p:
  - $100(1-\alpha)\%$ Confidence Interval for p
    - Sample size issue
  - Essentially a z-test

Chapter 13. Inference about comparing two populations

- Inference about $\mu_1-\mu_2$
  - Independent samples ("two-sample t-test")
    - Equal variance case (relatively easier)
    - Unequal variance case
    - Checking assumptions
  - Matched pairs experiment ("matched t-test")
- Inference about p_1-p_2
  - z-test and estimator
- Observational and experimental data, revisited
  - "Confounding"

Chapter 15. Analysis of variance

- ANOVA compares means of more than two populations
  - Or, more generally, it studies the effect of categorical predictor variable(s), called factor(s), on the interval response variable
- One-way ANOVA
- Two-way ANOVA

Chapter 16. Chi-squared tests

- Recall the contingency table
- Chi-squared tests are used to answer whether two categorical variables are related (or "dependent")

Chapter 17. Simple linear regression and correlation

- Least squares line coefficients (regression line)
  - Dependent variable x
  - Independent variable y
- Conditions on error variable ε
- SSE (sum of squares for error)
- Testing whether the slope =0 (t-test and t confidence interval)
  - Standard error of estimate
- Coefficient of determination R^2
- Cause-and-effect relationship?

Simple linear regression (continued)

- Using the regression equation
  - Prediction of a particular y for a given x
  - Estimating E(y) for a given x
- Regression diagnostics (Checking assumptions for regression)
  - Residual analysis
  - Non-normality
  - Heteroscedasticity
  - Non-independence
  - outliers

Chapter 18. Multiple regression

- More than one independent (interval) variables
- Model and assumptions
- Estimating and testing
- Regression diagnostics

Remaining chapters

- Chapter 19. Model building
- Chapter 20. Time series analysis and forecasting
- Chapter 21. Nonparametric statistics
- Chapter 22. Statistical process control
- Chapter 23. Decision analysis

Course summary

- Recap of the course objectives
- Have you achieved the course objectives?