Introduction to Statistics

Chris Parrish

August 25, 2013

Contents

Chapter 1

data

Minard's graphic of Napoleon's invasion of Russia ... Edward Tufte says it is "probably the best statistical graphic ever drawn" ... See also Minard's graphic of Hannibal's invasion of Italy

A large part of the scientific enterprise is based on an empirical process of discovery :

identify a problem → collect relevant data → analyze the data → form conclusions.

Statistics deals with stages two through four, and we begin with stage two by constructing a formal framework for collecting and organizing data.

Atlantic hurricanes, 1940-2010

Exit polls, simulation

1.1 exploratory data analysis

Hubble, recession of galaxies, 1929

Statistics seeks to organize and learn from data. The first step is to organize the data, on paper and in our computers.

- cases (observations), variables
- tables (in text), data frames (in computers)
- types of variables : numerical (discrete, continuous), categorical

Variables are classified as *categorical* or *numerical*, and numerical variables are further distinguished as *discrete* or *continuous*. Discrete variables take on only integer values. Continuous variables can take any real number as a value. At this early stage of collecting and organizing data, the distinction between discrete and continuous numerical data might not be so apparent, especially if all measurements are reported with a certain fixed number of decimal places, but later on, when we come to the analysis of data, the mathematical methods available to us will depend very much on that distinction.

Michelson, speed of light, 1879

OECD (2013), Education at a Glance 2013 : OECD Indicators

OECD (2011), Health at a Glance 2011 : OECD Indicators

1.2 numerical data

Old Faithful, bimodal distribution

Numerical values:

- parameters (characteristic of the population, e.g. μ, σ)
- statistics (characteristic of the sample, e.g. $\overline{x}, s$)

Summarizing numerical data:

- shape (symmetry: symmetric, right skewed, left skewed; modality: unimodal, bimodal, multimodal, uniform)
- center (mean, median, mode)
- spread (standard deviation, quartiles, IQR, range)
- outliers (1.5 × IQR rule)

Statistics vary in their sensitivity to changes in the data. Some statistics are more *robust* than others (median, quartiles).

For symmetric data, the mean and standard deviation are often cited. For skewed data, the median, quartiles, and IQR might better represent the distribution of the data. In either case, any exceptional values (*outliers*) should be noted.

Displaying numerical data:

- dot plot (one numerical variable)
- box plot (one numerical variable)
- histogram (density)
- plot (scatter plot, graph of a function)
- density map (geographical distribution)
- time series (temporal distribution)

Arbuthnot, baptismal records in London, 1629 to 1710, Arbuthnot in HistData

Pearson and Lee, heights of parents and children, 1903

Best actor and actress Oscar winners, 1929-2012, skewed distributions

Collecting prizes in cereal boxes, simulation, skewed distributions

Runs in basketball, simulation, skewed distributions

1.3 categorical data

Florence Nightingale, Crimean War, 1854-1856, rose diagrams

Summarizing categorical data:

- counts
- percentages

Displaying categorical data:

- frequency table (one set of categories)
- bar plot
- contingency table, $r \times c$ (two sets of categories)
- mosaic plot
- side-by-side bar plots

A *mosaic plot* is an effective representation of an $r \times c$ contingency table. In effect, the mosaic plot transposes the matrix. The first variable splits a big square into W-E blocks. The second variable splits those blocks into N-S sub-blocks. The dimensions of the blocks carry the information (not the areas of the rectangles). Designated colors will denote the levels of the last (response) variable.

Contingency tables, mosaic plots

1.4 design of studies

Salk, polio vaccine field trials, 1954 (Friedman, et al., Statistics, Fourth edition, pp.3-6).

Interpretation of variables:

- explanatory variables
- response variables

In some studies, the values of some of the variables, the *explanatory variables*, might be thought to affect the values of some other variables, the *response variables.*

If the values taken on by a variable are unaffected by the values taken on by another variable, then the variables are *independent*; otherwise, the variables are *dependent.* Dependent variables might exhibit a *linear* relationship. If an approximating line has a positive slope, then a linear relationship is *positive.* If an approximating line has a negative slope, then the linear relationship is *negative.*

Simpson's paradox. Imagine a large cloud of data points drooping slightly to the southeast. Ah, one thinks, a negative relationship. But then it is discovered that the large cloud is comprised of two subclouds, one on top of the other, and both rising to the northeast. Within those two groups, the association is positive. Aggregating data reversed the apparent trend.

A *hierarchy of evidence*:

- anecdote
- observational evidence (sample of convenience, retrospective study, volunteers)
- experiment (controlled, randomized, double-blind)

In an *observational study*, the researcher observes the values of the explanatory and response variables in the study sample.

In an *experiment*, the researcher randomly assigns the subjects in the study sample to the treatments and then observes the values of the response variables.

Sampling methods include taking *simple random samples*, *stratified samples*, and *cluster samples.*

Survey *bias* may take the form of *undercoverage* or *selection bias* (nonrandom sampling), *nonresponse bias* (participants are unavailable or fail to answer), or *response bias* (reporting is untruthful or influenced by leading questions).

Blocking during random assignment is analogous to stratified sampling during random sampling.

Outline for design of studies : research question, sample, variables, conclusions

1.5 inference

Literary Digest, the poll that changed polling : Landon vs. FDR, 1936

What is required to support *inference from a sample to a population*? The basic requirement is that the sample must be representative of the population. A formal structure to achieve this is a *simple random sample* (each subset of size n of the population is equally likely to be chosen as the sample). Thus, studies performed with volunteers, retrospective studies, and samples of convenience do not support inference from a sample to a population. Conclusions from such studies apply only to the participants in the study.

What is required to support *inference of causality* (this causes that)? The basic requirement is that members of the treatment groups should differ only in their treatments. The formal requirement to support this is an experiment with *random assignment* of the study subjects to the treatment groups. The gold standard in medicine is the *controlled, randomized, double-blind experiment.*

A famous saying in statistics : *correlation does not imply causation. Lurking variables* may be present. Think of the correlation of drowning deaths and ice cream sales along the famous Gold Coast near Brisbane, Australia. Both of them happen mostly in the summertime, and neither one causes the other.

Random sampling supports inference from the sample to the population.

Random assignment (to treatment groups) supports inference of causality (this causes that).

Correlation does not imply causation.

Outline for design of studies : research question, sample, variables, conclusions

Outline for inference : contingency table displaying random sample vs. random assignment, and the type of inference which is supported or not supported in each case (Çetinkaya-Rundel lecture slides, Stat 2, S2013, Duke University)

1.6 outlines

Outline for exploratory data analysis (Probability and Statistics, Open Learning Initiative, CMU) :

- distributions (one variable)
  - categorical
    - visual displays - barplot
    - numerical measures - proportions
  - quantitative
    - visual displays - dotplot, barplot, histogram, scatter plot, graph
    - numerical measures - center (mean, median), spread (standard deviation, IQR, range)
- relationships (two variables)
  - categorical × categorical
    - visual displays - barplot, mosaic plot
    - numerical measures - contingency table, proportions, conditional proportions
  - quantitative × categorical
    - visual displays - boxplots, scatter plots
    - numerical measures - quantiles
  - quantitative × quantitative
    - visual displays - scatter plot, graph, regression
    - numerical measures - center (mean, median), spread (standard deviation, IQR, range)

Outline for design of studies : research question, sample, variables, conclusions

Outline for inference : contingency table, random sample, random assignment, type of inference supported or not supported

Chapter 2

probability

2.1 probability functions

Sample space, experiment, outcome, event.

Probability. Venn diagrams. Mutually exclusive events. $P(A \cup B)$. Complements. $P(A^c)$.

Independent events. $P(A \cap B)$ when A and B are independent.

2.2 random variables

Definition and first examples

Expected value of a random variable, $E(X)$

Variance of a random variable, $Var(X)$

Linear combinations of random variables.

Expected value and variance of a linear combination of random variables.

2.3 distributions

2.3.1 joint, marginal, and conditional distributions

Law of total probability (sum of conditional probabilities)

Tree diagrams.

2.3.2 Bayes' theorem

2.3.3 continuous distributions

Calculation of probability using a continuous distribution

Chapter 3

distributions of random variables

3.1 discrete distributions

3.1.1 Bernoulli

Bernoulli random variable, $X \sim Bernoulli(p)$. Success or failure. Density function. Expected value. Variance.

3.1.2 Binomial

Binomial random variable, $X \sim Binomial(n, p)$. The probability of k successes in n trials. Expected value. Variance. Normal approximation to a binomial distribution.

3.2 continuous distributions

3.2.1 Normal

Normal random variable, $X \sim N(\mu, \sigma)$. Standardized normal random variable, $Z \sim N(0, 1)$. Areas of regions under a normal distribution curve. Quantiles. The 68-95-99.7% rule. Q-Q plots.

3.2.2 Student's t, Chi-Square, F

Student's t, Chi-Square, and F distributions play key roles in the sequel. All of them are families of continuous distributions. Student's t distributions resemble Normal distributions but they have fatter tails. Chi-Square and F distributions have domains the half line $[0, \infty)$, so neither one is symmetric.

Chapter 4

inference

4.1 distribution of the sample mean

Two scenarios : (1) If we are studying a quantitative variable and we know the population parameters μ and σ, what can be said of the sample statistics? (2) If we know the sample statistics $\overline{x}$ and s, what can be said of the population parameters? The first question is easiest to answer, but is rarely the case. The second question leads to much of contemporary statistics.

We might be studying a quantitative variable in a population (normal or not) with population parameters μ and σ, and we wish to know the sampling distribution of the sample mean $\overline{x}$. Or we might be studying a categorical variable in a population (normal or not) with proportion p, and we wish to know the sampling distribution of the sample proportion $\hat{p}$. The following table summarizes the sample distributions in both cases (Probability and Statistics, Open Learning Initiative, CMU).

Variable	Statistic	Shape	Center	Standard Error	Conditions
quantitative (σ known)	$\overline{x}$	Normal	μ	$\sigma/\sqrt{n}$	$n \geq 30$ or approx. normal
quantitative (σ unknown)	$\overline{x}$	t	μ	$s/\sqrt{n}$	$n \geq 30$ or approx. normal
categorical	$\hat{p}$	Normal	p	$\sqrt{\frac{p(1-p)}{n}}$	$\min(np, n(1-p)) \geq 10$

The standard deviation of a statistic is its *standard error*:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

When they are known, these values indicate the accuracy of $\overline{X}$ and $\hat{p}$. When they are not known, which is often the case, we use the *estimated standard errors* $s_{\overline{X}}$ and $s_{\hat{p}}$ in their place (Rice, p.213). They are calculated from the data.

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \quad \text{and} \quad s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Sample mean, point estimate, variation of such statistics, distribution of the sample mean. Standard error, $SE = \sigma/\sqrt{n}$ or $SE = s/\sqrt{n}$.

An important contrast : the population distribution, with its mean and its standard deviation, vs. the distribution of the sample mean, $\overline{X}$, and its SE.

4.2 LLN and CLT

LNN : $\overline{X}_n \to \mu$ as n gets larger.

CLT : the distribution of $\overline{X}_n \to Normal\ Distribution$ as n gets larger.

Both theorems require certain conditions to be satisfied for the theorem to be applicable. The CLT requires

- independent sample elements (from less than 10% of the population)
- relatively large sample size (for instance, at least 30 elements)
- data which are not highly skewed and without extreme outliers

in order to conclude that the sample mean is approximately normally distributed with standard deviation approximately SE (OIS, p.168).

4.3 confidence intervals for a single mean

Three approaches to understanding data : *point estimates, confidence intervals, hypothesis tests.*

Confidence interval, confidence level, 95% confidence interval, *point estimate* $\pm z^*SE$.

The *margin of error* is the half width of the confidence interval, $ME = z^*SE$.

Interpretation of a confidence interval.

4.4 hypothesis tests for a single mean

Hypothesis tests, null hypothesis, alternative hypothesis, one-sided and two-sided alternative hypotheses, sample statistic, normalized sample statistic, p-value, significance level α, conclusion of an hypothesis test, Type I and Type II errors, framework for statistical inference using hypothesis tests.

Note that statistical significance and practical relevance are two distinct issues.

There are four steps in the formal process of using hypothesis tests for statistical inference (Probability and Statistics, Open Learning Initiative, CMU):

- *Hypotheses.* Formulate the null and alternative hypotheses.
- *Data and sample statistic.* Collect relevant data from a random sample and summarize them using an appropriate sample statistic. Verify the conditions which determine the distribution of the sample statistic.
- *p value.* Calculate the associated p value, the probability of obtaining the observed sample statistic if the null hypothesis is true.
- *Conclusion.* Decide whether or not there is enough evidence to reject H_0 and accept H_A, and state the conclusion in context.

Chapter 5

numerical variables

5.1 simulation

5.1.1 bootstrap

Bradley Efron, Stanford, 1979, “Bootstrap Methods: Another Look at the Jackknife”

Suppose that we would like to construct a confidence interval for a small data set. Recall that the CLT requires a sample size of at least 30 elements and data which are not highly skewed for us to conclude that the distribution of the sample mean is nearly normal. What can be done if the sample size is less than 30 elements?

One recourse is the *bootstrap*. Suppose the sample size is n. Choose a random sample of the same size, n, *with replacement from the sample data*, and calculate the statistic of interest, in this instance the mean. Repeat this experiment many times (say, 1000 times). Make a histogram of the results. This is close to the distribution of your statistic. To create a 95% confidence interval for the statistic, choose the middle 95% of the bootstrap distribution.

5.1.2 randomization

To calculate a p value for a statistic taken from a small data set, shift the bootstrap distribution to the null value and calculate the probability of getting that statistic from that distribution (see Çetinkaya-Rundel lecture slides, S2013, unit 4, lecture 1).

The basic idea is that the original sample represents the population, so resampling from the original sample should also resemble (to a certain extent) sampling from the population.

5.2 difference of two means for paired data

To analyze *paired data*, examine the difference in outcomes for each pair, e.g. $X_{diff} = X_{before} - X_{after}$.

Assumptions and conditions for the analysis of paired data (see Çetinkaya-Rundel lecture slides, S2013, unit 4, lecture 1).

Confidence interval for difference of two means.

Formula for SE for difference of two means.

Hypothesis testing for difference of two means, e.g.

$$H_0 : X_{before} - X_{after} = 0,$$
$$H_A : X_{before} - X_{after} \neq 0.$$

Test statistic, Z value, p value, and conclusion for difference of two means.

5.3 difference of two means for small samples, Student's t

Student, Gosset, Guinness brewery in Dublin, 1908, Student's t, Gosset's letter to RA Fisher, "You may be the only person ..."

If the population mean μ is unknown (as is often the case) we approximate it with the sample mean $\overline{x}$.

If the population standard deviation is unknown (as is often the case) we approximate it with the sample standard deviation s.

Test statistic for small samples : $T \sim t(df)$

Conditions for using a t-distribution (see Çetinkaya-Rundel lecture slides, S2013, unit 4, lecture 2).

Hypothesis tests for difference of two small sample means

Confidence intervals for difference of two small sample means

Test statistic, t value, p value, and conclusion for difference of two small sample means.

Protocol : Inference for difference of two small sample means (see Çetinkaya-Rundel lecture slides, S2013, unit 4, lecture 2).

Wetsuits and swimming speed (Lock, p.427-431)

5.4 many means

One-way ANOVA

Multiple comparisons : Bonferroni method, Tukey method

Which means differ, and by how much? Fisher method

ANOVA table for comparing means and ANOVA table for regression

Two-way ANOVA

Main effects, interaction

Outline for inference : theoretical, simulation, conclusion in context

Chapter 6

categorical variables

6.1 one proportion

Population proportion, p, sample proportion, $\hat{p}$, standard error of the proportion.

$SE = \sqrt{p(1-p)/n}$ when p is known, and $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ when p is unknown.

CLT for proportions : $\hat{p} \sim N(mean = p, SE = \sqrt{p(1-p)/n})$ provided that certain conditions are satisfied.

Conditions for the CLT for proportions : (1) sample observations are independent, (2) sample size is sufficiently large, which is determined by the *success-failure conditions* ($pn \geq 10$ and $(1-p)n \geq 10$)

Confidence intervals for proportions.

Test statistics for proportions.

SE for confidence intervals for proportions, $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ where $\hat{p}$ is the sample proportion.

SE for test statistics for proportions, $SE = \sqrt{p_0(1-p_0)/n}$ where p_0 is the null value.

6.2 two proportions

Standard error for the difference of two proportions for confidence intervals and for null hypotheses of the form $H_0 : p_1 - p_2 \neq 0$:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Standard error for the difference of two proportions for null hypotheses of the form $H_0 : p_1 - p_2 = 0$:

$$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$$

where

$$\hat{p}_{pool} = \frac{\text{number of successes in both groups combined}}{n_1 + n_2}$$

6.3 many proportions

Chi-Square test for *goodness of fit*.

Chi-Square test for independence.

Chapter 7

modeling (numerical response)

7.1 one explanatory variable

7.1.1 linear model

Explanatory (predictor) and response (predicted) variables.

Fitting $y \sim x$ results in a *linear model*, $y = \beta_0 + \beta_1 x$

Point estimates for β_0 and β_1 are determined from the sample and are denoted b_0 and b_1

An association is characterized by its direction (positive or negative), form (linear or non-linear) and strength (which for linear relationships is measured by the correlation)

Correlation coefficient, ρ

Residual, $e_i = y_i - \hat{y}_i$

7.1.2 least squares line

Conditions for least squares : (1) nearly linear relationship, (2) nearly normal residuals, (3) with nearly constant variability.

Calculate b_0 and b_1

The center of mass of the sample lies on the least squares line.

Use a least squares line to predict y from x : $\hat{y} = b_0 + b_1 x^*$

R^2 is the percent of the response variable explained by the explanatory variable.

7.2 many explanatory variables

7.2.1 multiple linear regression

Multiple linear regression creates a linear model with k explanatory variables.

Highly correlated explanatory variables are to be avoided.

Either of two algorithms (forward or backward) can be used to systematically add uncorrelated variables or delete highly correlated variables to or from the developing model, with either p or R^2 ordering the addition or deletion of variables at each step

Conditions for multiple linear regression

Acorn size and oak tree range, eesee

Chapter 8

nonparametric statistics

8.1 Wilcoxon test

The *Wilcoxon test* compares two groups by ranking all of the individuals in the sample and then analyzing how those rankings are distributed between the two groups. If the response variable is quantitative, we can calculate point estimates and confidence intervals for the difference between the population medians.

8.2 Kruskal-Wallis test

The *Kruskal-Wallis test* compares the mean rankings of several groups, not just two. If the Kruskal-Wallis test indicates that the groups do not share the same distribution, then the Wilcoxon test can be used to compare the mean rankings of pairs of groups to indicate the extent to which they differ.

8.3 Sign test

The *Sign test* quantifies differences in matched pairs. For each pair, we record whether the member of the first group is somehow better (positive) or worse (negative) than the second. The Sign test summarizes the distribution of the positive and negative scores between the groups.

8.4 Wilcoxon Signed-Ranks test

The *Wilcoxon Signed-Ranks test* quantifies differences in matched pairs for which we can measure the degree of difference between the two members of each pair. For each pair, we will have a sign indicating that the first member of the pair is somehow better or worse than the second (positive or negative score) and a ranking of just how different they are.

Chapter 9

retrospective

9.1 data

Variable	Display
quantitative	dotplot, boxplot, beanplot, histogram, graph
categorical	barchart
quantitative × quantitative	scatterplot, Q-Q plot
categorical × quantitative	side-by-side dotplots, boxplots, histograms
categorical × categorical	mosaic plot, barchart (segmented or side-by-side)

Technique	Inference
random sample	supports inference from sample to population
random assignment	supports inference of causation

9.2 numerical and categorical variables

9.2.1 one mean or proportion

Inference for one variable. If the variable of interest is quantitative, we may wish to infer about the population mean, μ. If the variable of interest is categorical, we may wish to infer about the population proportion, p.

For quantitative variables, if the population parameter σ is known, and the CLT theorem applies, then the distribution of $\overline{x}$ is approximately normal with mean μ and standard deviation given by $SE = \sigma/\sqrt{n}$. Confidence multipliers are of the form z^*, and confidence intervals are given by $\overline{x} \pm ME = \overline{x} \pm z^* SE$.

For quantitative variables, if the population parameter σ is not known, but the sample is still random, then we substitute s for σ, and the CLT theorem no longer applies. In this new situation, $\overline{x}$ has a t distribution with $n-1$ degrees of freedom and with mean μ and standard deviation given by $SE = s/\sqrt{n}$. Confidence multipliers are of the form t^*, and confidence intervals are given by $\overline{x} \pm ME = \overline{x} \pm t^* SE$.

For categorical variables, the CLT theorem applies and the distribution of the sample proportion $\hat{p}$ is approximately normal with mean p and standard deviation given by $SE = \sqrt{\frac{p(1-p)}{n}}$.

Variable	Statistic	Shape	Center	Standard Error	Conditions
quantitative (σ known)	$\overline{x}$	Normal	μ	$\sigma/\sqrt{n}$	$n \geq 30$ or approx. normal
quantitative (σ unknown)	$\overline{x}$	t	μ	$s/\sqrt{n}$	$n \geq 30$ or approx. normal
categorical	$\hat{p}$	Normal	p	$\sqrt{\frac{p(1-p)}{n}}$	$\min(np, n(1-p)) \geq 10$

9.2.2 two means or proportions

Inference for two variables. For quantitative variables, the method extends to paired differences and to differences of means. For categorical variables, the method extends to differences of proportions.

Inference	Distribution	Conditions	Standard Error
mean	$t, df = n-1$	$n \geq 30$ or approx. normal	$s/\sqrt{n}$
paired difference in means	$t, df = n_d - 1$	$n_d \geq 30$ or approx. normal	$s_d/\sqrt{n_d}$
difference in means	$t, df = \min(n_1, n_2) - 1$	$\min(n_1, n_2) \geq 30$ or approx. normal	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	Normal	$\min(np, n(1-p)) \geq 10$	$\sqrt{\frac{p(1-p)}{n}}$
difference in proportions	Normal	$\min(n_i p_i, n_i(1-p_i)) \geq 10$	$\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$

Inference	Confidence Interval	Test Statistic
	sample statistic $\pm$ *multiplier* $\cdot$ *SE*	(sample statistic - null) / SE
mean	$\overline{x} \pm t^* \cdot s/\sqrt{n}$	$\big(\overline{x} - 0\big)/\big(s/\sqrt{n}\big)$
paired difference in means	$\overline{x}_d \pm t^* \cdot s_d/\sqrt{n_d}$	$\big(\overline{x}_d - 0\big)/\big(s_d/\sqrt{n_d}\big)$
difference in means	$(\overline{x}_1 - \overline{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\big((\overline{x}_1 - \overline{x}_2) - 0\big)/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$(\hat{p} - p_0)/\sqrt{\frac{p_0(1-p_0)}{n}}$
difference in proportions	$(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$\big((\hat{p}_1 - \hat{p}_2) - 0\big)/\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$

9.2.3 many means

ANOVA tests for a difference in means among the levels of a categorical variable

If ANOVA leads us to conclude that there is a difference in means among the levels of a categorical variable (that the means are not all the same), then which means are different?

If the categorical variable has only two levels, then ANOVA is equivalent to a test for difference in means.

9.2.4 many proportions

one categorical variable : chi-square test for goodness of fit

two categorical variables : chi-square test for independence

If there is one categorical variable having only two levels, then the chi-square test for goodness of fit is equivalent to a test for a single proportion.

If there are two categorical variables having only two levels each, then the chi-square test for independence is equivalent to a test for a difference in two proportions.

9.3 simulation

9.3.1 bootstrap

9.3.2 randomization

9.4 modeling (numerical response)

9.4.1 one explanatory variable : linear regression

9.4.2 many explanatory variables : multiple regression

Bibliography

[1] Alan Agresti, Christine Franklin, *Statistics, The Art and Science of Learning from Data*, 3/e, Pearson, 2013.

[2] David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel, *OpenIntro Statistics*, 2/e, OpenIntro, 2012.

[3] David Freedman, Robert Pisani, Roger Purves, *Statistics*, 4/e, W. W. Norton and Company, New York, 2007.

[4] David S. Moore, George P. McCabe, Bruce A. Craig, *Exploring the Practice of Statistics*, W. H. Freeman and Company, New York, 2014.

Index