Stat 205: Introduction to Nonparametric Statistics

Lecture 03: Nonparametric Inference Continued

Instructor David Donoho; TA: Yu Wang

Questions for Today

- Examples
- Inference Discussion
- Tests and estimates
- Confidence Statements
- Robustness
- Rank-estimation Linear Model

Example 3.2.1 Esophageal Cancer, 1

(Breslow et al. 1980)

- Case-control study of esophageal cancer.
- Null Hypothesis:
  Alcohol consumption same in the two groups.
- Dataset

```
library(datasets); data(esoph); head(esoph)
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
## 2 25-34 0-39g/day    10-19      0        10
## 3 25-34 0-39g/day    20-29      0         6
## 4 25-34 0-39g/day      30+      0         5
## 5 25-34     40-79 0-9g/day      0        27
## 6 25-34     40-79    10-19      0         7
```

Example 3.2.1 Esophageal Cancer, 2

```
> tail(esoph)
   agegp  alcgp     tobgp ncases ncontrols
83   75+  40-79     20-29      0         3
84   75+  40-79       30+      1         1
85   75+ 80-119 0-9g/day      1         1
86   75+ 80-119     10-19      1         1
87   75+   120+ 0-9g/day      2         2
88   75+   120+     10-19      1         1
```

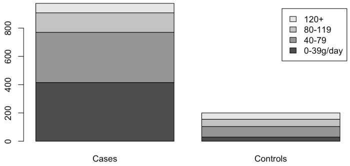
Data Wrangling: `x <- rep(esoph$alcgp,esoph$ncases)`

```
> x
 [1] 120+      0-39g/day 40-79     40-79     40-79     40-79     120+      120+
[17] 40-79     40-79     40-79     40-79     40-79     40-79     40-79     40-79
[33] 80-119    80-119    80-119    80-119    80-119    80-119    80-119    80-119
[49] 120+      120+      120+      120+      120+      120+      120+      120+
```

Data Wrangling: `x <- as.numeric(x)`

```
> as.numeric(x)
  [1] 4 1 2 2 2 2 4 4 4 4 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [82] 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[163] 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 1
```

Example 3.2.1 Esophageal Cancer, 3




```
> y<- rep(esoph$alcgp,esoph$ncontrols)
> x<- rep(esoph$alcgp,esoph$ncases)
> wilcox.test(as.numeric(x),as.numeric(y))

	Wilcoxon rank sum test with continuity correction

data:  as.numeric(x) and as.numeric(y)
W = 135612, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Reject H_0 alcohol consumption is the same

Example 3.2.1 Esophageal Cancer, 4

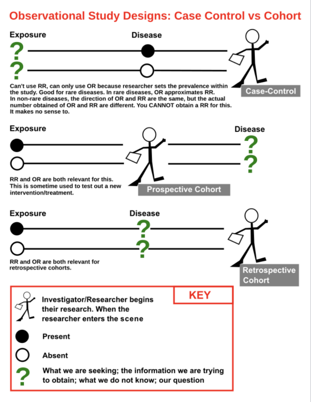
Case–control study

From Wikipedia, the free encyclopedia

A **case–control study** (also known as **case–referent study**) is a type of observational study in which two existing groups differing in outcome are identified and compared on the basis of some supposed causal attribute. Case–control studies are often used to identify factors that may contribute to a medical condition by comparing subjects who have that condition/disease (the "cases") with patients who do not have the condition/disease but are otherwise similar (the "controls").[1] They require fewer resources but provide less evidence for causal inference than a randomized controlled trial. A case–control study produces only an odds ratio, which is an inferior measure of strength of association compared to relative risk.

Contents [hide]

1 Definition
 1.1 Control group selection
 1.2 Prospective vs. retrospective cohort studies
2 Strengths and weaknesses
3 Examples
4 Analysis
5 Impact on longevity and public health
6 See also
7 References
8 Further reading
9 External links



Case–control study versus cohort on a timeline. "OR" stands for "odds ratio" and "RR" stands for "relative risk".

Example 3.2.1 Esophageal Cancer, 5

Definition [edit]

The case–control is a type of epidemiological observational study. An observational study is a study in which subjects are not randomized to the exposed or unexposed groups, rather the subjects are *observed* in order to determine both their exposure and their outcome status and the exposure status is thus not determined by the researcher.

Porta's *Dictionary of Epidemiology* defines the case–control study as: an observational epidemiological study of persons with the disease (or another outcome variable) of interest and a suitable control group of persons without the disease (comparison group, reference group).[2] The potential relationship of a suspected risk factor or an attribute to the disease is examined by comparing the diseased and nondiseased subjects with regard to how frequently the factor or attribute is present (or, if quantitative, the levels of the attribute) in each of the groups (diseased and nondiseased)."[2]

For example, in a study trying to show that people who smoke (the *attribute*) are more likely to be diagnosed with lung cancer (the *outcome*), the *cases* would be persons with lung cancer, the *controls* would be persons without lung cancer (not necessarily healthy), and some of each group would be smokers. If a larger proportion of the cases smoke than the controls, that suggests, but does not conclusively show, that the hypothesis is valid.

Examples [edit]

One of the most significant triumphs of the case–control study was the demonstration of the link between tobacco smoking and lung cancer, by Richard Doll and Bradford Hill. They showed a statistically significant association in a large case–control study.[10] Opponents argued for many years that this type of study cannot prove causation, but the eventual results of cohort studies confirmed the causal link which the case–control studies suggested,[11][12] and it is now accepted that tobacco smoking is the cause of about 87% of all lung cancer mortality in the US.

From Last Lecture
Examples in Chapter 3
About p-values and Inference
Tests and estimates
Robustness
Linear Model

Example 3.1: Melanoma Mortality

```
trying URL 'http://lib.stat.cmu.edu/R/CRAN/bin/macosx/contrib/4.0/HSAUR2_1.1-18.tgz'
Content type 'application/x-gzip' length 3111882 bytes (3.0 MB)
==================================================
downloaded 3.0 MB


The downloaded binary packages are in
/var/folders/n7/9qc2q4sn2qb6jwkcchc6636h0000gp/T//Rtmp9kqFtk/downloaded_packages
> library(HSAUR2)
Loading required package: tools
> head(USmelanoma)
            mortality latitude longitude ocean
Alabama           219     33.0      87.0   yes
Arizona           160     34.5     112.0    no
Arkansas          170     35.0      92.5    no
California        182     37.5     119.5   yes
Colorado          149     39.0     105.5    no
Connecticut       159     41.8      72.8   yes
```

From Last Lecture
Examples in Chapter 3
About p-values and Inference
Tests and estimates
Robustness
Linear Model

Example 3.1 Melanoma Mortality

Exploratory Analysis (Boxplot)

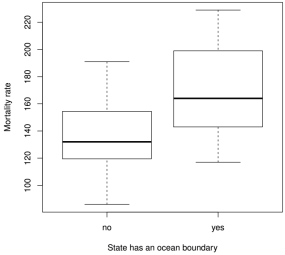


FIGURE 3.1
Mortality rates for white males due to malignant melanoma in the United States.

Formal Analysis

TABLE 3.1
Estimates of Increase in Mortality Due to Malignant Melanoma in White Males in the United States.

	Test	p-value	Estimate	Std Error
Least Squares	3.60	0.00	31.49	8.55
Wilcoxon	3.27	0.00	31.00	9.26

Q: What do such p-values mean?

One-Sample Problem: Tests and Estimates, 1

Test	Estimate
Sign Test	Median
Wilcoxon Signed-Rank	Hodges-Lehmann
Normal Scores Test	Normal Scores Estimate

Sign Test Example:

- Apply Sign Test to *shifted* sample $\{X_i - \Delta\}$.
- Record dependence of Sign Test statistic on the shift parameter.

$$S(\Delta) \equiv \sum_{i=1}^{n} \mathrm{sign}(X_i - \Delta)$$

- Let $M_n \equiv M_n(\{X_i\}) = Median(\{X_i\})$.
- Suppose no ties, n odd. Then
$$S(M_n) = 0.$$
Indeed,
$$\#\{i : X_i \leq M_n\} = \#\{i : X_i \geq M_n\}.$$

One-Sample Problem: Tests and Estimates, 2

Test	Estimate
Sign Test	Median
Wilcoxon Signed-Rank Test	Hodges-Lehmann
Normal Scores Rank Test	Normal Scores Estimate

Wilcoxon Test Example:

- Apply the Wilcoxon Rank-Sum Test to *shifted* sample $\{X_i - \Delta\}$.
- Record dependence of Wilcoxon signed-rank-sum statistic on the shift parameter.

$$W(\Delta) = \sum_{i=1}^{n} \mathrm{sign}(X_i - \Delta) R|X_i - \Delta|$$

- Let $H_n \equiv H_n(\{X_i\}) = Median_{i,j}(\{(X_i + X_j)/2\})$. *Hodges-Lehmann* aka *Pseudomedian*.
- Suppose no ties among the pairwise 'Walsh averages' $A_{i,j} \equiv (X_i + X_j)/2$. Define $\mathrm{sign}(0) = 0$.

$$W(H_n) = 0.$$

One-Sample Problem: Confidence Statements, 1

Estimate	$(1-\alpha)$ Confidence Statement
Median	$(X_{(c_1+1)}, X_{(n-c_1)})$
Hodges-Lehmann	$(A_{(c_2+1)}, A_{(n_2-c_2)})$

- Median
  - $X_{(i)}$ are the **order statistics**:
    $$X_{(1)} \leq X_{(2)} \leq \dots X_{(n)};$$
    `orstats = sort(x)` in R.
  - c_1 is the $\alpha/2$-quantile of the binomial distribution `bin(n,1/2)`
    `c.1 = qbinom(alpha/2,n,0.5)`
- Hodges-Lehmann
  - $A_{(i)}$ are the **ordered Walsh Averages**:
    $$A_{(1)} \leq A_{(2)} \leq \dots A_{(n_2)};$$
    `orwalsh = sort(as.vector(outer(xx,FUN=function(x,y)(x+y)/2)))` in R.
  - $n_2 = n(n+1)/2 = \#\{(i,j) : 1 \leq i,j \leq n\}$.
  - c_2 denotes the $\alpha/2$-quantile of the Wilcoxon signed-rank W^+
    `c.2 = qsignrank(alpha/2,n)`

One-Sample Problem: Confidence Statements, 2

Estimate	$(1-\alpha)$ Confidence Statement
Median	$(X_{(c_1+1)}, X_{(n-c_1)})$

c_1 is the $\alpha/2$-quantile of the binomial distribution `bin(n,1/2)`
`c.1 = qbinom(alpha/2,n,0.5)` Asymptotically,

$$c_1 \approx \frac{n}{2} + 2 \cdot \mathfrak{z}_{\alpha/2}\sqrt{n}, \qquad n \to \infty.$$

Theorem: *Distribution-Free Coverage probability of Confidence Statements:*

- Suppose
  - $X_i =_{iid} \Delta + Z_i$, $i = 1, \ldots, n$
  - $Z_i \sim F$ symmetric.

$$P\{Z_i < -t\} = P\{Z_i > t\}, \qquad \forall t \in \mathbf{R}.$$

- Conclude:

$$P\{X_{(c_1)} \leq median(\{X_i\}) \leq X_{(n-c_1)}\} \geq 1 - \alpha.$$

- If $\alpha/2$ is an exact lower tail probability, `alpha = 2 * pbinom(c.1,n,1/2)` , and if F is a continuous increasing CDF on its support,

$$P\{X_{(c_1)} \leq median(\{X_i\}) \leq X_{(n-c_1)}\} = 1 - \alpha.$$

so-called *exact coverage probability*. These conclusions are true for all F: *distribution-free.*

Similar statements are available for Hodges-Lehmann, Normal Scores, and other rank tests.
Would-be statements like these for t-test, in case F is not $N(0, \sigma^2)$, would not be true.

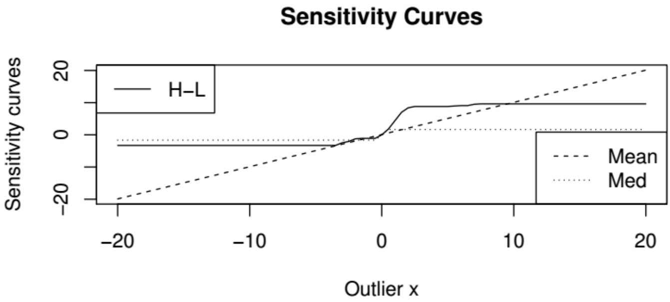
Robust Estimation Concepts. 1. Sensitivity Curve

- Statistic $\hat{\theta}(\mathbf{x})$
- Our sample $\mathbf{x}_n = (x_1, \ldots, x_n)^T$
- Augmented sample $\mathbf{x}_{n+1}(z) = (x_1, \ldots, x_n, z)^T$
- Sensitivity curve:

$$S(z; \hat{\theta}) \equiv \frac{\hat{\theta}(\mathbf{x}_{n+1}(z) - \hat{\theta}(\mathbf{x}_n)}{1/(n+1)}$$

From Last Lecture
Examples in Chapter 3
About p-values and Inference
Tests and estimates
Robustness
Linear Model

Robust Estimation Concepts. 2. Sensitivity Curve



Property	Mean	Median
Shape of Sensitivity Curve	z	sign(z)
Max of Sensitivity Curve	Unbounded	Bounded
Robust	No	Yes

Robust Estimation Concepts. 2 Breakdown Point

- Statistic $\hat{\theta} \equiv \hat{\theta}(\cdot)$
- Our sample $\mathbf{x}_n = (x_1, \ldots, x_n)^T$
- Augmented sample $\mathbf{x}_{n+k}(z_1, \ldots, z_k) = (x_1, \ldots, x_n, z_1, \ldots, z_k)^T$
- $\hat{\theta}$ *breaks down* at $\mathbf{x}_n$ under contamination with k points if

$$+\infty = \max_{(z_1, \ldots, z_k)} |\hat{\theta}[\mathbf{x}_{n+k}(z_1, \ldots, z_k)] - \hat{\theta}(\mathbf{x})|$$

- Breakdown point: $k^*(\mathbf{x}_n, \hat{\theta}) = \min_k \hat{\theta}$ breaks down at$\mathbf{x}_n$

$$\epsilon^*(\mathbf{x}_n, \hat{\theta}) = \frac{k^*}{k^* + n}.$$

Property	Mean	Median	
Number to break down	$k^* = 1$	$k^* = n$	$k^* \approx (n + k^*) \cdot (1 - \sqrt{2})$
Breakdown Point	$\varepsilon^* = \frac{1}{n+1}$	$\varepsilon^* = 1/2$	$\varepsilon^* = 0.29$

The Geometry of Linear Models

- Setup the following linear model (for $i = 1, \dots, n$)

$$Y_i = x_i^T \boldsymbol{\beta} + e_i^*$$

where $\boldsymbol{\beta}$ is a $1 \times p$ vector of unkown parameters

- $\boldsymbol{\beta}$ are the parameter of interst
- Center (usually using the median $T(e_i^*) = \alpha$) the errors $e_i = e_i^* - \alpha$

$$Y_i = \alpha + \boldsymbol{x}_i \boldsymbol{\beta} + e_i$$

- Let $f(t)$ be the pdf of the erros e_i
- Assumption: $f(t)$ can be either asymmetric or symmetric depending on whether signs or ranks are used
- The intercept α is independent of the slope $\boldsymbol{\beta}$

The Geometry of Linear Models

- Let $Y = (Y_1, \dots, Y_n)^T$ denote the $n \times 1$ vector of observations
- Let X denote the $n \times p$ matrix with rows x_i^T
- Then we can write the linear model in matrix form:

$$\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

- X is centered (that's fine since we have α in the model), and assume X is full column rank
- Let Ω_F be the column space spanned by coluns of X
- So we can rewrite the linear model as (coordinate-free because not restricited to any specific basis vectors)

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{e}$$

with $\boldsymbol{\eta} = \Omega_F$

The Geometry of Estimation

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{e} \qquad \text{with} \qquad \boldsymbol{\eta} = \Omega_F$$

- Task is to minimize some distance between $\boldsymbol{Y}$ and subspace Ω_F
- Think of $\boldsymbol{\eta}$ as a hyperplane and the task as projecting $\boldsymbol{Y}$ onto it
- For the projection we need to define a distance
- Instead of using the usual Euclidean distance, we use a distance based on signs and ranks

$$\|v_i\|_\varphi = \sum_{i=1}^{n} a(R(v_i))v_i$$

- with scores $a(1) \leq a(2) \leq \cdots \leq a(n)$ and score function $a(i) = \varphi(i/(n+1))$
- φ is nondecreasing, centered, standardized and defined on the interval $(0, 1)$

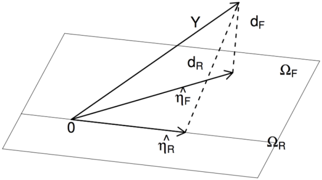
The Geometry of Estimation

- $\|\boldsymbol{v}\|_\varphi$ is a pseudo-norm:
  - triangle inequality, non-negative, $\|\alpha \boldsymbol{v}\|_\varphi = |\alpha| \|\boldsymbol{v}\|_\varphi$, and
  - additionally $\|\boldsymbol{v}\|_\varphi = 0$ if and only if $v_1 = \cdots = v_n$
- By setting $\varphi_R(u) = \sqrt{12}(u - 1/2)$, we get the Wilcoxon pseudo-norm
- By setting $\varphi_S(u) = \operatorname{sgn}(u - 1/2)$, we get the sign pseudo-norm (equivalent to using the L_1 norm)
- In general

$$D(\boldsymbol{Y}, \Omega_F) = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_\varphi\|_\varphi = \min_{\boldsymbol{\eta} \in \Omega_F} \|\boldsymbol{Y} - \boldsymbol{\eta}\|_\varphi$$

The Geometry of Estimation

$$\widehat{\eta} = D(Y, \Omega_F) = \|Y - \widehat{Y}_\varphi\|_\varphi = \min_{\eta \in \Omega_F} \|Y - \eta\|_\varphi$$



Source: Hettmansperger & McKean (2011)

- Estimate $\widehat{\eta}_\varphi$
- Distance between Y and the space Ω_F is d_F
- Reduced model subspace $\Omega_R \subset \Omega_F$

From Last Lecture
Examples in Chapter 3
About p-values and Inference
Tests and estimates
Robustness
Linear Model

Questions for Today

- Examples
- Inference Discussion
- Tests and estimates
- Confidence Statements
- Robustness
- Rank-estimation Linear Model