Nonparametric Statistics

Yingxing Li

Office Hour: Tue 10:00-12:00
Econ Building B305
Email:amyli999@hotmail.com

Course Outline:

0. Introduction and Review
1. One-Sample Methods
2. Two-Sample Methods
3. Multiple-Sample Methods
4. Paired Comparisons & Block Designs
5. Tests for Trends and Association
6. Inference with Dichotomous Responses
7. Nonparametric Bootstrap Methods
8. Nonparametric Regression

Some References:

- Higgins. *Introduction to Modern Nonparametric Statistics* (required).
- Hollander and Wolfe. *Nonparametric Statistical Methods*.
- Sprent and Smeeton. *Applied Nonparametric Statistical Methods*.
- Paradis. *R for Beginners*.

Contents

0 Introduction and Review

0.1 Parametric and Nonparametric Statistics

Parametric Statistics

- **Parameter**: constant (usually unknown) that characterizes a population distribution.
- **Statistic**: a function of random variables (observations) that does not depend on the unknown parameters.
- **Parametric methods**: estimation and inference are based on some assumption on the form of the distribution.
- For example, suppose we assume IQ scores $X_i \sim N(\mu, 10^2)$. We observe 10 IQ scores: 121, 98, 95, 94, 102, 106, 112, 120, 108, 109. Question: is the mean IQ significantly greater than 100?
  - Null hypothesis: $H_0 : \mu = 100$

- Alternative hypothesis: $H_a : \mu > 100$ (upper-tailed test)
- Test procedure: z-test based on the normality assumption

Nonparametric Statistics

- **Nonparametric statistics**:
  - the form of the joint distribution is not assumed.
  - test and estimation procedures require relatively fewer assumptions about the population distribution.
  - "nonparametric" is a misnomer. We will estimate and test hypotheses about parameters;
  - but the form of the distribution is not assumed. Often we only assume that the random variables are independent and identically distributed ($i.i.d.$);
  - more accurate term: **distribution-free** statistics
- For example, suppose we only assume IQ scores X_i are $i.i.d.$. We

observe 10 IQ scores: 121, 98, 95, 94, 102, 106, 112, 120, 108, 109. Question: is the median IQ significantly greater than 100? Note here we do not assume X_i follow Normal distributions.

- **Why study nonparametric statistics?**
  - In many applications, there is no prior knowledge of the underlying distributions.
  - If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
  - For studies with small sample size, normal approximation does not work well.
- Therefore, we need statistical methods that
  - require very little model/distributional assumptions;
  - or those that are robust/insensitive to the model/distributional assumptions;

– insensitive to outliers in the data.

0.2 Review of Probability Theory

0.2.1 Normal Distribution (Continuous)

- A popular bell-shaped continuous distribution.
- The probability density function of $X \sim N(\mu, \sigma^2)$:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}, -\infty < x < +\infty,$$

where (μ, σ^2) are **parameters.**

- For $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$ and $V(X) = \sigma^2$.
- $N(0, 1)$: standard normal distribution.

- Standardization: if $X \sim N(\mu, \sigma^2)$, then

$$Z = (X - \mu)/\sigma \sim N(0, 1).$$

- For any random variable X, $F(x) = P(X \leq x)$ is the **cumulative distribution function** (CDF).

- For $Z \sim N(0, 1)$, $F_Z(z) = P(Z \leq z) = \Phi(z)$, which can be found from the normal table.

- For $X \sim N(\mu, \sigma^2)$, $F_X(x) = P(X \leq x) = P\{Z = (X - \mu)/\sigma \leq (x - \mu)/\sigma\} = \Phi\{(x - \mu)/\sigma\}$.

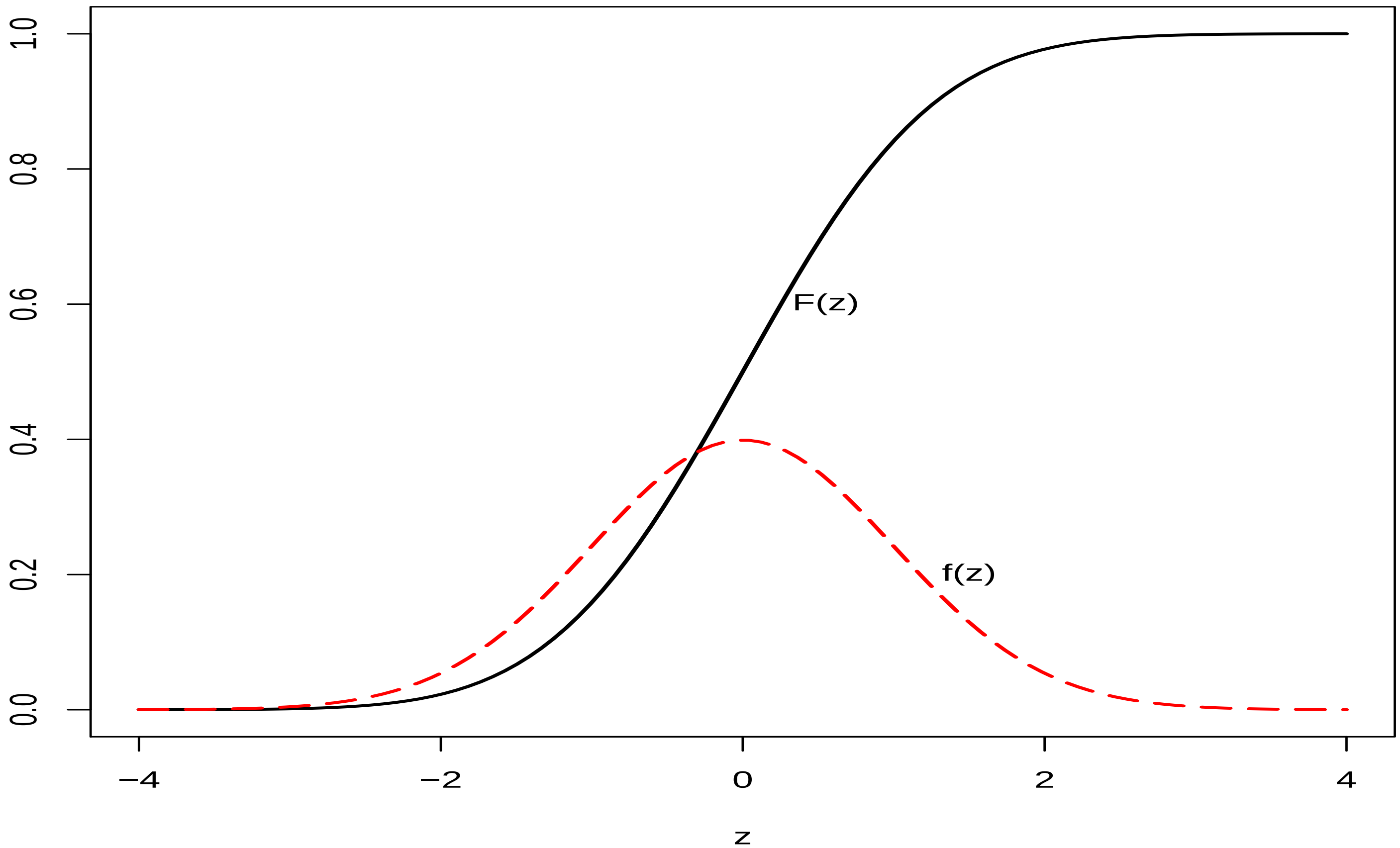

Figure 1: PDF and CDF of standard normal distribution.

Notations of Percentage Point and Quantile

Suppose a random variable Z follows a continuous distribution. The αth quantile of Z is defined as

$$Q_\alpha = \{z : P(Z \le z) = \alpha\}, \quad i.e.\ P(Z < Q_\alpha) = \alpha.$$

The αth percentage point of Z is defined as

$$z_\alpha = \{z : P(Z > z) = \alpha\}, \quad i.e.\ P(Z > z_\alpha) = \alpha.$$

By the definitions,

$$z_\alpha = Q_{1-\alpha}.$$

For example, if $Z \sim N(0,1)$, $z_{0.05} = Q_{0.95} = 1.645$.

The definitions are similar for discrete distributions.

Note that these notations are consistent with most statistical books but different from Higgins (2004), where z_α is used to denote the αth quantile.

Example 0.2.1 1. *Suppose* $Z \sim N(0,1)$. *Use Table A2 to find* $P(Z > 1.28)$ *and* $P(Z < -1.96)$.

2. *Find the 99th percentile of* $N(0,1)$.

3. *Find* $z_{0.025}$.

4. *Suppose* $X \sim N(2,9)$. *Calculate* $P(X > 6)$.

0.2.2 Binomial Distribution (Discrete)

Binomial experiment

- Consists of a known number n of Bernoulli trials.
- Each trial has only two possible outcomes, success (S) or failure (F).
- Probability of success $P(S) = p$ is the same on each trial, where $0 \leq p \leq 1$ is a **parameter**.
- Trials are independent.
- The Binomial random variable of is $X =$ Number of successes in n trials. Denote $X \sim Binomial(n, p)$.

Some properties of $Binomial(n, p)$

- Probability mass function (pmf):

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, \cdots, n.$$

- CDF:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}.$$

- Mean $E(X) = \sum_{all\ x} xp(x) = np$.

- Variance $V(X) = np(1-p)$.

0.2.3 Mean and Variance

- Mean of a random variable $E(X)$: measures the center/location of its distribution.
- Variance $V(X)$: measures the variation/dispersion of its distribution.
- For discrete random variable X:

$$E(X) = \sum_{all\ x} xP(X = x), \quad E(X^2) = \sum_{all\ x} x^2 P(X = x).$$

- For continuous random variable X:

$$E(X) = \int x f(x) dx, \quad E(X^2) = \int x^2 f(x) dx.$$

- For any random variable X:

$$V(X) = E\{X - E(X)\}^2 = E(X^2) - \{E(X)\}^2.$$

- $E(a + bX) = a + bE(X)$.
- $V(a + bX) = b^2 V(X)$.
- **Linear combination of independent random variables**. Suppose $X = a_1 X_1 + \cdots + a_n X_n$, where X_i are independent random variables, and $a_1, \cdots, a_n$ are known constants. Then

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n)$$

$$V(a_1 X_1 + \cdots + a_n X_n) = a_1^2 V(X_1) + \cdots + a_n^2 V(X_n)$$

Sample Mean and Sample Variance. Suppose X_i are $i.i.d.$ (independent and identically distributed) with (population) mean μ and variance σ^2. In reality, μ and σ^2 are unknown **parameters**. How can we estimate μ and σ^2 based on the sample $\{X_1, \cdots, X_n\}$?

- **Sample mean**: $\bar{X} = n^{-1} \sum_{i=1}^n X_i$
- **Sample variance**: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

- Other measurements of the location: sample median, mode.
- Other measurements of dispersion: range, median absolute deviation, interquartile range.

Properties of sample mean and sample variance. Suppose $X_1, \cdots, X_n$ be $i.i.d.$ with mean μ and variance σ^2. Then

- $E(\bar{X}) = \mu$. Prove.
- $V(\bar{X}) = \sigma^2/n$. Prove.
- $E(S^2) = \sigma^2$.

0.3 Review of Statistical Inference

0.3.1 One-sample Z-Test

Example 0.3.1 *Suppose 5 people participated in a weight loss program. After 4 weeks, their weight losses are: -2 20 12 11 14. Is the program effective? Suppose the weight losses* $X_i \sim N(\mu, 11^2)$.

One-Sample Z-test:

- **Assumption**: The random sample $X_1, \cdots, X_n$ are $i.i.d.$ from $N(\mu, \sigma^2)$, where μ is the unknown mean, and σ^2 is the variance and is **known**.
- Null hypothesis $H_0 : \mu = \mu_0$ (the distribution is centered at μ_0, a prespecified null value).
- **Test statistic**
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$
where $\bar{X} = 1/n \sum_{i=1}^{n} X_i$. Under H_0, $Z \sim N(0, 1)$.
- When should we reject H_0?
- Need specify a significance level α, i.e. we want to find a rejection rule such that
$$\text{Type I error} = P(\text{Reject } H_0 | H_0 \text{ is True}) \leq \alpha.$$

- **Rejection region**: the region of z_{obs} such that H_0 will be rejected at the sig. level α. That is, reject H_0 when $z_{obs} \in RR$, and do not reject H_0 otherwise.
  - (lower-tailed test) $H_a: \mu < \mu_0$, $RR = \{z_{obs} : z_{obs} < -z_\alpha\}$
  - (upper-tailed test) $H_a: \mu > \mu_0$, $RR = \{z_{obs} : z_{obs} > z_\alpha\}$
  - (two-tailed test) $H_a: \mu \neq \mu_0$, $RR = \left\{z_{obs} : |z_{obs}| > z_{\alpha/2}\right\}$
- **P-value**: The ***p*-value** is the probability of obtaining a test statistic value as extreme as the observed value, calculated assuming H_0 is true.
  - Lower-tailed test $H_a: \mu < \mu_0$, p-value$=P(Z < z_{obs})$
  - Upper-tailed test $H_a: \mu > \mu_0$, p-value$=P(Z > z_{obs})$
  - 2-tailed $H_a: \mu \neq \mu_0$, p-value$=P(Z < -|z_{obs}|\ OR\ Z > |z_{obs}|)$

  The more extreme observed test statistic value
  $\Longleftrightarrow$ smaller p-value $\Longleftrightarrow$ more evidence to reject H_0

Decision based on p-value: $\boxed{\textit{Reject } H_0 \textit{ if p-value} < \alpha.}$

Note: In order to calculate p-value, we need know what is the distribution of the test statistic when H_0 is true—**Null Distribution**. For the one-sample Z-test, $Z \sim N(0, 1)$ under H_0, so Table A2 can be used to calculate p-value.

Example 0.3.2 *Revisit Example 0.3.1. State your conclusion with significance level $\alpha = 0.1, 0.05$ and 0.01.*

0.3.2 Central Limit Theorem

Central Limit Theorem: suppose $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables with finite mean μ and variance $\sigma^2 > 0$. Then for large n,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ or } Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1),$$

approximately.

Recall

$$\boxed{E(\bar{X}) = \mu \text{ and } V(\bar{X}) = \sigma^2/n}$$

$$\boxed{E(\textstyle\sum_{i=1}^{n} X_i) = n\mu \text{ and } V(\textstyle\sum_{i=1}^{n} X_i) = n\sigma^2.}$$

Approximation of Binomial with Normal:

Suppose $X \sim Binomial(n, p)$.

We can write $X = X_1 + X_2 + \cdots + X_n$, where X_i are $i.i.d.$ Bernoulli random variables that takes value 1/0 if the ith trial is a success/failure. By CLT, for large n approximately we have

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

or

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1), \ \hat{p} = X/n.$$

Example 0.3.3 *Among 100 people who participated in a weight loss program, 75 people lost weight after 4 weeks. Test* $H_0 : p = 0.5$ *versus* $H_a : p > 0.5$, *where* p *is the weight loss success rate.*

Example 0.3.4 *Suppose* $X_1, \cdots, X_n$ *are* $i.i.d.$ *from* $Uniform(0, \theta)$.

1. *Use CLT to obtain the asymptotic distribution of* $\sqrt{n}(\bar{X} - \theta/2)$. *[Hint: the mean and variance of* $Uniform(0, \theta)$ *are* $\theta/2$ *and* $\theta^2/12$, *respectively.]*

2. *Suppose the waiting time for wolfline Rt 1 follows* $Uniform(0, \theta)$. *Among 40 students surveyed, the mean waiting time is 8.5 mins. Test* $H_0 : \theta = 20$ *versus* $H_a : \theta \neq 20$. *Use significance level* $\alpha = 0.01$.

0.4 Introduction to R

- R: a free open source software with a lot of statistical packages. Can be downloaded from: http://cran.r-project.org/.
- Use "#" to add comments
- Basic R operations and functions

```
##
#### vector and matrix
##
# define a vector
x = c(121, 98, 95, 94, 102, 106, 112, 120, 108, 109)
#character vector
w = c("F","M","M","F","F","M","M","F","M","M")
w
# print the data
x
#define y to be the log transformation of x
y=log(x)
#construct a 10 by 2 matrix, with columns x and y
mat = cbind(x, y)  # cbind: combine by columns
```

```
mat
#add another row to the data set
new = c(100, log(100))
mat2 = rbind(mat, new)
mat2
#check the dimension of a matrix
dim(mat)
dim(mat2)
ncol(mat2)  # the number of columns of a matrix
nrow(mat2)  # the number of rows of a matrix

mat[2,3] # value in the second row and the third column of mat
mat[1:3,]  # the first three rows of mat
mat[,2]  # the second column of mat
mat[-1,] # exclude the first row
mat[,-1] # exclude the first column

####access the documentation of a certain function
?cbind
help(cbind)

##
#### Arithmetics and simple functions
##
```

```
x + y
z = x - y
z
x*y

sum(x)  # summation of all the elements in x
# summation of the 2nd, 3rd and the 5th elements of x
sum(x[c(2,3,5)])
# product of all the elements in x
prod(x)
# product of all the elements in x except the first two
prod(x[-c(1,2)])

#logic operation
sum(x) > 10
1*(x[1]>10) #returns 1 if true and 0 if false
sum(w=="F")  #number of "F" in the vector w

##
#### graphical analysis
##
par(mfrow=c(2,2)) # will result in 4 plots, 2 rows & 2 columns
plot(x, main="plot of x")  # scatter plot of x
plot(y~x, main="y versus x")  # plot y versus x
```

```
hist(x, main="hist of x") # show the histogram of x
boxplot(x, main="boxplot of x") # plot the boxplot of x

# some summary statistics
# calculate the sample mean (average)
mean(x)
# calculate the sample standard deviation
sd(x)
# calculate the correlation of y and x
cor(y,x)
# compute the median of x
median(x)
# compute the 90th percentile of x
quantile(x, 0.9)
# compute the 0th, 25th, 50th, 75th, 100th percentiles of x
quantile(x)

##
#### Importing and Exporting Data
##
# you can change the directory to where the data is saved
score = read.csv("Ex1.csv") #check the dimension of the matrix
dim(score)
```

```
#print the first five rows
score[1:5,]
#check the correlation of Q1 and Q2
cor(score[,2], score[,3])

#add another column log(Q1)
logQ1 = log(score[,2])
score = cbind(score, logQ1)

#use 1 to denote Female and 0 for male
gender = 1*(score[,8]=="F")
score = cbind(score, gender)

#take a look at the first 5 rows
score[1:5,]

#export the new modified score
write.table(score, "Ex1-new.csv", sep=",",
col.names=TRUE, row.names=FALSE)

##
#### Random number generation
##
# generate 100 data points from N(mu=0, sigma=2)
```

```
x = rnorm(100, 0, 2)
mean(x)
max(x)
quantile(x, 0.9)
mean(x) > 1
hist(x)

# generate 100 data points from exponential distribution
x = rexp(100, 2)
mean(x)
max(x)
quantile(x, 0.9)
1*(mean(x) > 1)
hist(x)
abline(v=quantile(x,0.9),col="red")

# other distribution: runif, rchisq, rt, etc
```