Concluding Remarks
Statistics: Principles, Methods and R (I)

Fengnan Gao[1,2]

2017.9.11

[1]School of Data Science, Fudan University

[2]Shanghai Center for Mathematical Sciences

fngao@fudan.edu.cn

Contents (Exam Version) i

1. Introduction to *R*: What is R? Installing R, help and documentation, data objects, data import and export, basic data manipulation, computing with data, organising an analysis.
2. Probability: sample space and events, probability, independent events, conditional probability, Bayes' formula.
3. Random variables and distributions: distribution functions and probability functions, mean and variance, moment generating functions, discrete random variables, continuous random variables.

Contents (Exam Version) ii

4. Multivariate distributions: bivariate distributions, marginal distributions, independent random variables, conditional distributions, multivariate distributions, IID samples, transformations of random variables.
5. Inequalities: probability inequalities, inequalities for expectations.
6. Convergence of random variables: types of convergence, law of large numbers (LLN), central limit theorem (CLT).
7. Introduction to Statistical Inference: what is statistics? parametric and nonparametric models, fundamental concepts in inference, empirical distributions.

Contents (Exam Version) iii

8. Point estimation: method of moments estimation, maximum likelihood estimation (MLE), properties of MLE.
9. Hypothesis testing I: null and alternative hypotheses, p-values, two-types of errors, the Wald test, t-tests and t-intervals.
10. Hypothesis testing II: likelihood ratio tests, Pearson's χ^2-test, goodness-of-fit tests, permutation tests.

- The first 6 chapters are basic probablity backgrounds.
- The statistical part include the statitical inference, parameter estimation, hypothesis testing (both parametric and nonparametric).

Contents (Exam Version) iv

- Please note that all contents related to bootstrap, EM algorithm, Monte-Carlo (those by Zhang Nan) have all been removed from the exam contents. For any questions in that direction, please ask Zhang Nan.
- Office hour: 14:30–17:00, 5 January 2018. Please only ask specific questions and **refrain** from asking anything about the exam.

In the final Exam

- Each student is allowed to carry one handwritten A4 paper with notes.
- Each student is allowed to carry one pocket (non-smart) calculator.
- It is not necessary to remember complicated formulas by heart, such as the densities of complicated distributions. If such things are necessary in the exam, they will be provided.
- Remembering is not understanding.
- Learn to use the statistical table. You might need to look up quantiles from the table. Each student will be provided with the statistical table in the exam.

Relations to Stats II

- There will be the course Statistics: Principles, Methods and R (II) in the next semester.
- Wang Qinwen will teach that course.
- Stats I covers the **basic** aspects of statistics, and Stats II will be more advanced and state-of-art.

Standard error or standard deviation

- Let X be an random variable with finite variance $\mathrm{Var}(X)$. The *population standard deviation* is simply the square root of the variance $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$.
- You may estimate the (population) standard deviation by the sample standard deviation

$$S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

- Suppose θ_n is an estimator of the parameter θ. $\hat{\theta}_n$ is also a random variable with standard deviation $\mathrm{SD}(\hat{\theta}_n)$, which is also called the *standard error* $\mathrm{SE}(\hat{\theta}_n)$.
- Typically $\mathrm{SE}(\hat{\theta}_n)$ depends on the unknown distribution, which we have to estimate. The *(estimated) standard error* $\widehat{\mathrm{SE}}(\hat{\theta}_n)$ is often obtained by pluging in the estimated population distribution.

Example of SD, SE and $\widehat{\text{SE}}$

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$-distributed. which is the estimated standard deviation of X_i. For estimators of μ and σ^2, we have

	μ	σ^2
estimator	$\bar{X}_n$	S_n^2
SE or SD	$\sigma/\sqrt{n}$	$\sqrt{2}\sigma^2/\sqrt{n-1}$
$\widehat{\text{SE}}$ (or SE)	$S_n/\sqrt{n}$	$\sqrt{2}S_n^2/\sqrt{n-1}$

Therefore, the asymptotic normality of S_n^2 has the form

$$\frac{S_n^2 - \sigma^2}{\sqrt{2}\sigma^2/\sqrt{n-1}} \xrightarrow{d} N(0, 1).$$

Between confidence intervals and hypothesis testing

One final note

- Consider the simple testing problem for iid data $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with known σ^2

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0.$$

- Under the null hypothesis, $\mu = 0$ $\bar{X}_n \sim N(0, \sigma^2/n)$. Thus, we reject the hypothesis at level α if $\bar{X}_n$

$$|\bar{X}_n| \geq z_{\alpha/2}\sigma/\sqrt{n}.$$

- The $(1 - \alpha)$ confidence interval for μ is

$$[\bar{X}_n \pm z_{\alpha/2}\sigma/\sqrt{n}].$$

- Rejection of the null hypothesis corresponds to

$$0 \notin [\bar{X}_n \pm z_{\alpha/2}\sigma/\sqrt{n}]$$