Statistical Computation Math 475

Jimin Ding

Department of Mathematics
Washington University in St. Louis
www.math.wustl.edu/ jmding/math475/index.html

August 29, 2013

Mean and Variance

Assume $X_1, X_2, \cdots, X_n$ are independent identically distributed (i.i.d) random variables (r.v.) with probability distribution function (pdf) $f(x)$.

- Population mean:

$$E(X) = \int xf(x)dx.$$

- Population variance:

$$Var(X) = \int (x - E(X))^2 f(x)dx.$$

Mean and Variance

- Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

  Note sample mean $\bar{X}$ is a random variable, which follows a different distribution than $f(x)$.

- Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- Sample standard deviation: s.

A Statistic

A statistic is a function of data. For example, sample mean ($\bar{X}$) and sample variance (s^2). As a random variable, a statistic can be described by its distribution function (df) or probability distribution function (pdf) or probability mass function (pmf). It is usually used to estimate a population characteristic of a r.v. or construct a hypothesis test.

A Statistic

For example:

- If $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, 2, \cdots, n$, then

$$\bar{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2).$$

- If n is large enough (≥ 30), $X_i's$ are i.i.d. with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, then based on central limit theorem (CLT)

$$\bar{X} \quad app. \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2).$$

Hence the statistic $\bar{X}$ can be used to estimate the population mean μ and s is to estimated the population standard deviation σ.

"Standard error" is usually an estimated standard deviation of a statistic. (Q: What is the standard error of the sample mean?)

Summary Statistics

- Mean, Variance
- Sample size
- Min, Max, Median (Q2), other quantiles (Q1, Q3)
- Coefficient of Variation: $CV = s/\bar{X}$
  A measure of dispersion of a probability distribution.
  Noise-to-signal ratio. Example: exponential.
- Test statistics: z-score
- Pearson correlation coefficient: $r = \frac{s_{XY}}{s_X s_Y}$
  A measure of linear correlation between two samples.
- Vector statistics: Ranks

Describing Data

Besides Summary Statistics,

- For categorical Data: gender, color, region, grade,
  - Frequency table;
  - Bar/Pie chart;
- For continuous Data: height, weight, income, GPA
  - Histogram;
  - QQplot;

For example: see SAS output.

Hypothesis test

Example: Student's t-test for population mean:

- $H_0 : \mu = c$ *v.s.* $H_a : \mu \neq c$ (two-sided)

Hypothesis test

Example: Student's t-test for population mean:

- $H_0 : \mu = c$ *v.s.* $H_a : \mu \neq c$ (two-sided)
- Choose a statistical test and calculate the test statistic(s):

$$t = \frac{\bar{X} - c}{s/\sqrt{n}} \sim t_{n-1}.$$

Hypothesis test

Example: Student's t-test for population mean:

- $H_0 : \mu = c$ *v.s.* $H_a : \mu \neq c$ (two-sided)
- Choose a statistical test and calculate the test statistic(s):

$$t = \frac{\bar{X} - c}{s/\sqrt{n}} \sim t_{n-1}.$$

- P-value:
  $= P($given H_0, observe a "worse" $t)$
  $= P(|T| > t | H_0$ is true),
  where T is a random variable with t_{n-1} distribution.

Hypothesis test

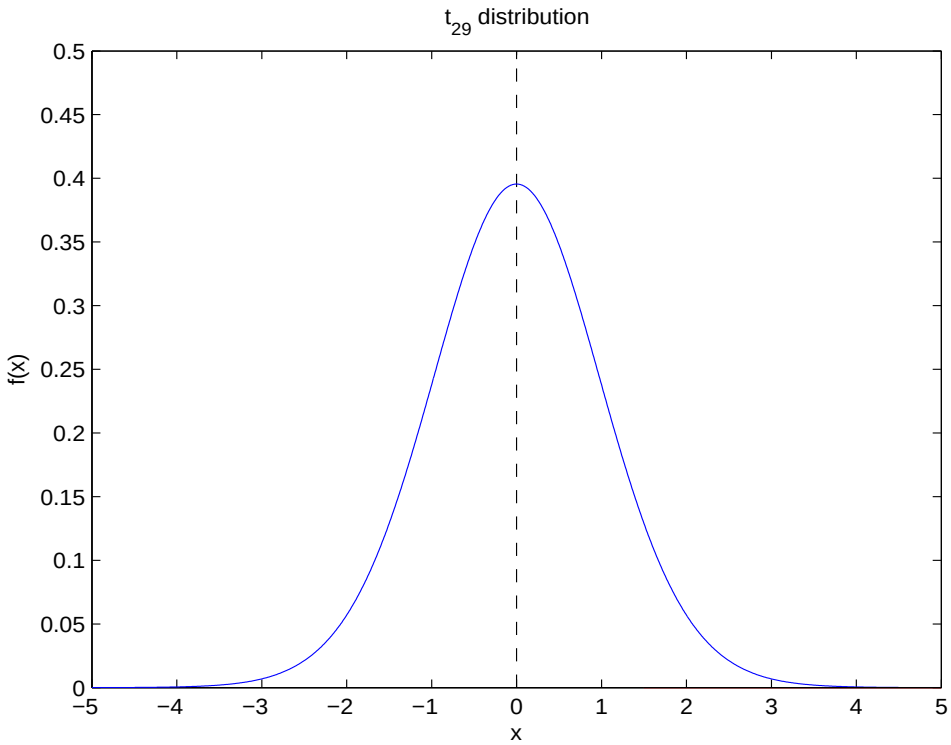
Example: Student's t-test for population mean:

- $H_0 : \mu = c$ *v.s.* $H_a : \mu \neq c$ (two-sided)
- Choose a statistical test and calculate the test statistic(s):

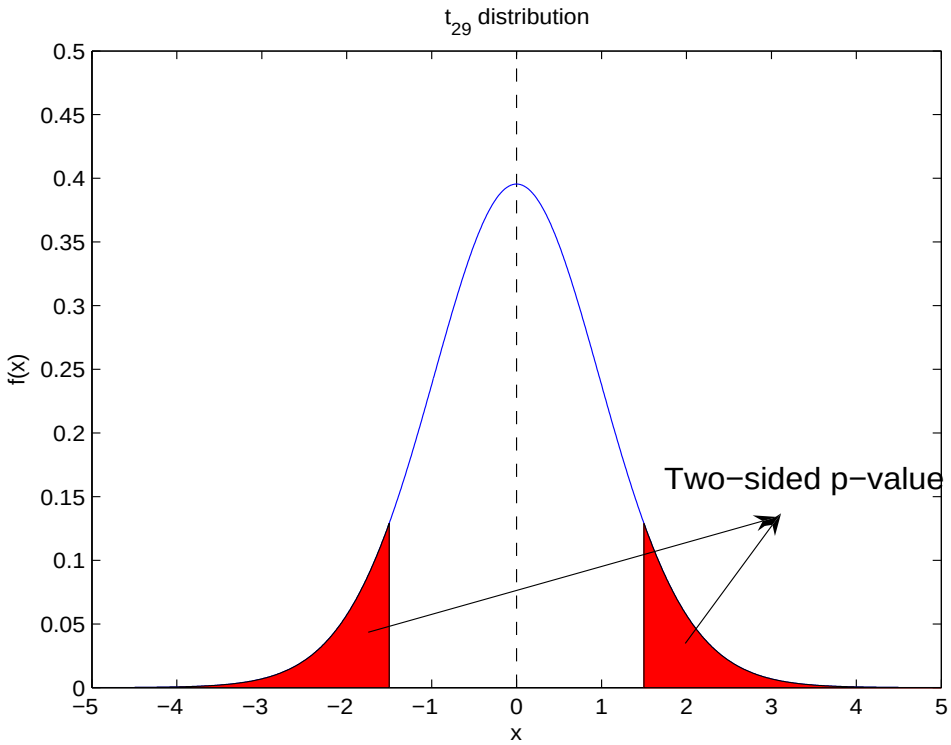
$$t = \frac{\bar{X} - c}{s/\sqrt{n}} \sim t_{n-1}.$$

- P-value:
  $= P$(given H_0, observe a "worse" t)
  $= P(|T| > t | H_0$ is true),
  where T is a random variable with t_{n-1} distribution.
- Conclusion:
  At significance level of α, we reject the null hypothesis if p-value$< \alpha$. (Otherwise, we fail to reject the null hypothesis.)

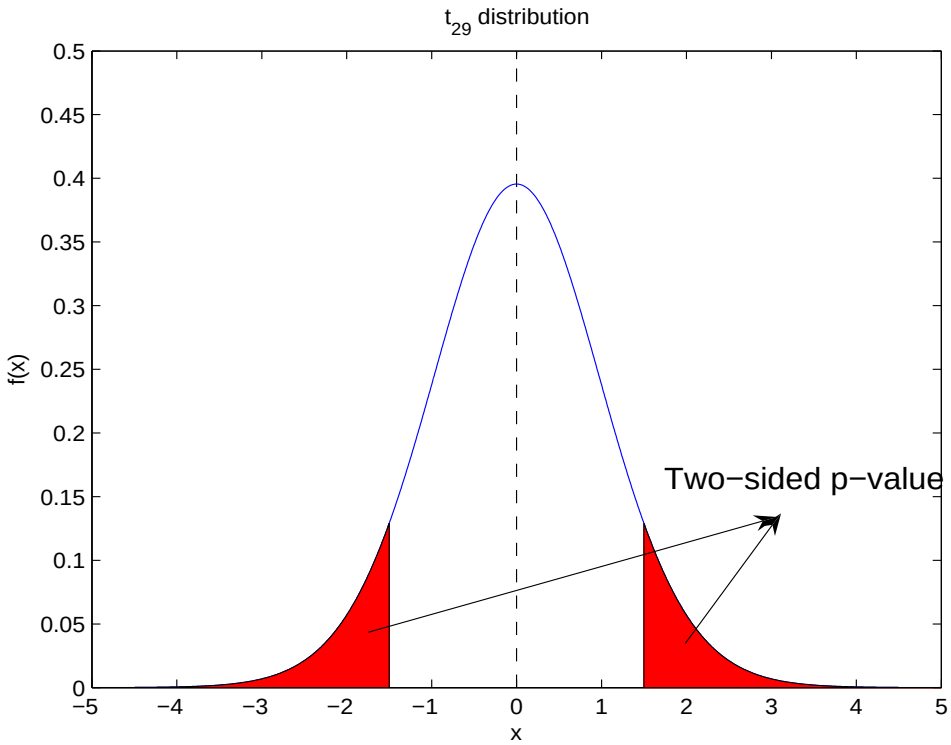
t_{n-1} distribution



t_{n-1} distribution



t_{n-1} distribution



Note: A hypothesis test is always based on population characteristics.
It is NEVER VALID to test: ~~$H_0 : \bar{X} = 0$~~
but should test: $H_0 : \mu = 0$.

Confidence Intervals :

Example: $100(1-\alpha)\%$ confidence interval (CI) for the population mean:

$$\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Confidence Intervals :

Example: $100(1-\alpha)\%$ confidence interval (CI) for the population mean:

$$\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Note: A CI can be always constructed as:
point estimation $\pm$ critical value $\times$ standard error

Confidence Intervals :

Example: $100(1-\alpha)\%$ confidence interval (CI) for the population mean:

$$\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Note: A CI can be always constructed as:
point estimation $\pm$ critical value $\times$ standard error

CI and hypothesis test are both referred as "inference" in statistics and involve calculation of variance of estimation.

Test for Normality

H_0 : The r.v.s are normally distributed.
H_a : The r.v.s are not normally distributed.

- Kolmogorov-Sminov: not good for practice.
  It is based on

$$D = \sup_x |F_n(x) - F_0(x)|, \quad \text{where} F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1[X_i \leq x].$$

  The $F_n(x)$ is called the Empirical Distribution Function, which is an estimation of df $F(x)$. KS test can be also used to test distributions other than normal.
- Anderson-Darling test (Stephens,1974):
  An extension from KS test, which puts more weights at the tail. The critical value depends on the $F_0(x)$, and is hence a more sensitive test.

Test for Normality

- Shapiro-Wilk test (1965):

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

where a_i's are constants generated from means, variance and covariance of the order statistics of a sample size of n, $\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$'s.
The critical value is selected based on Monte Carlo simulations. This test has a very good practical performance.

Test for Normality: Graphic Tools

- Boxplot
- Stem-and-leaf plot
- Histogram plot
- QQ plot/Normality Probability Plot:
  A plot that is nearly linear suggests normal distribution. Plot the *ith* smallest observation in a random sample of size n on y-axis, and plot $sz(\frac{i-0.375}{n+0.25}) + \bar{x}$ on x-axis. Under normality assumption, this value is an approximation of the expected value and should be close to the observed value if data are from a normal random sample.

  In SAS, normality probability plots have normal percentiles marked on on x-axis, and QQ plots have normal quantiles. But the plots are same.

Basic SAS

- SAS command is case insensitive
- Semicolon (;) is required at the end of each statement (a command line)
- Comments in SAS:

```
/* my comments */
 * my comments;
```

- SAS programs contain two parts: data management and statistical analysis
- Data step in SAS: create SAS datasets
  `DATALINES` (`CARDS`): type raw data directly in the SAS program
  `INFILE`: read raw data from an external file

Basic SAS

- SAS/STAT procedures: `PROC XXX;`
  Standard build in statistical analysis, which requires very rigid structure and commands.
- End of a paragraph in SAS: `RUN;` (`QUIT;`)
  The SAS keywords required to finish each block of program codes (data step, proc xxx).
  You still need to click on the running man icon to process the whole (or highlighted part of) program.
- Formatting plain text output:
  `OPTIONS`: controls the line size, page size, page number, date and so on.
  `TITLE`: creates informative titles in SAS output.

SAS Programs

- DATA step
- PROC MEAN
- PROC UNIVARIATE
- PROC FREQ
- PROC SORT

Reading Assignment

Textbook: Applied Statistics and the SAS Programming Language,
Chap 1 and 2, P1-P64

Probability Joke

Three roommates slept through their midterm statistics exam on Monday morning. Since they had returned together by car from the same hometown late Sunday evening, they decided on a great little falsehood. The three met with the instructor Monday afternoon and told him that an ill-timed flat tire had delayed their arrival until noon.The instructor, while somewhat skeptical, agreed to give them a makeup exam on Tuesday. When they arrived the instructor issued them the same makeup exam and ushered each to a different classroom. The first student sat down and noticed that the exam was divided into Parts I and II weighted 10% and 90% respectively. Thinking nothing of this disparity, he answered the questions in Part I, which was rather easy, and moved confidently to Part II on the next page, which had only one short and pointed question..... "Which tire was it?"

Source: http://my.ilstu.edu/ gcramsey/ChanceProb.html