

Chapter 1

Introduction to Statistics and Data Analysis

吳明龍 助理教授
成大資訊系/醫資所
minglong.wu@csie.ncku.edu.tw
辦公室:雲平大樓東棟 415 室

1.1 Overview

- Success of Statistical Methods
- Statistical Inference
 - Samples 樣本
 - Populations 母體
 - Experimental Design
- Statistics and Data Analysis

Success of Statistical Methods

- Beginning in the 1980s
- The use of statistical methods in
 - Manufacturing
 - Development of food products
 - Computer software
 - Pharmaceuticals
 - Many other areas
- Examples
 - Improvement of quality in American industry
 - Japanese “industrial miracle”: High-quality products

Inferential Statistics

- Collecting scientific (commercial) data or information in a systematic way with planning. Data provide understanding of scientific phenomena.
- Information is gathered in the form of samples, or collections of observations. Samples are collected from populations.
- Scientists or engineers often focus only on certain properties of objects in the population, and seeks to learn about the population.
- Example
 - An engineer may need to study the effect of process conditions, temperature, humidity, amount of a particular ingredient, ...

Statistics and Data Analysis

變異度

- Data analysis: center of location, variability, and general nature of the distribution of observations in the sample.
- Offspring of inferential statistics: a large of “Toolbox” of statistical methods which are used to make scientific judgments in face of uncertainty and variation.
- Modern statistical software packages allow for computation of means, medians, standard deviations.
- Graphical methods include histograms, stem and leaf plots, dot plots, and box plots.

Role of Probability

機率可加強統計的可信度

- Concepts in probability form a major component that supplements statistical methods and help **estimate the strength of the statistical inference**.
- Example 1.1
 - In a manufacturing process, 100 items are sampled and 10 are found to be defective.
 - However, in the long run, the company can only tolerate 5% defective in the process.
 - Suppose we learn that if the process is acceptable, i.e., if it does produce items 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process.
 - The **small probability** suggests that the process indeed have a **long-run defective exceeding 5%**.
 - **Probability aids in translation of sample information into conclusions.** 機率：利用抽樣結果，對母體概況下結論

Role of Probability

- Example 1.2
 - Study the development of a relationship between the roots of trees and the action of a fungus (真菌).
 - Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus.
 - Does the use of nitrogen influence stem weight?
 - Experimental Design: Two samples of 10 northern red oak seedlings (幼苗) are planted in a greenhouse, one containing seedlings treated with nitrogen and one containing no nitrogen.
 - All other environmental conditions are held constant.

Role of Probability

- Example 1.2

- The stem weights in grams were recorded after the end of 140 days.
- 4 nitrogen observations are larger than any of the no-nitrogen observations.
- Most of the no-nitrogen observations appear to be below the center of the data.
- Would the data set indicate that nitrogen is effective?
- How can this can be quantified or summarized in some sense ?

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

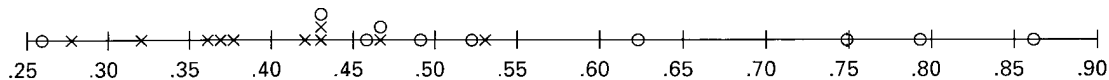


Figure 1.1 Stem weight data.

Role of Probability

推論統計：利用機率，得到資料分佈、特性，推導結論

- For a statistical problem, the sample along with inferential statistics allow us to draw conclusions about the population, with inferential statistics making clear use of elements of probability.
- Problems in probability allow us to draw conclusions about characteristics of hypothetical (假設的) data taken from the population based on known features of the population.

1.2 Sampling Procedures

- The importance of **proper sampling** revolves around the degree of confidence with which the analyst is able to answer the questions being asked.
- Simple **random sampling** implies that any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
- **Biased sample**: Simple random sampling is not always proper.
 - Example: A sample is chosen to answer certain questions regarding political preferences in a certain state in the U.S. Now, suppose that all or nearly all of the 1,000 sampling families chosen live in urban (vs. rural) areas.
- To characterize and quantify **measures of variability** is very important in inferential statistics. 資料不確定性

1.3 Measures of Location: Sample Mean and Median

- Location measures in a data set are designed to provide some quantitative measure of where the data center is in a sample.
 - The observations in a sample are x_1, x_2, \dots, x_n .
 - The sample mean is $\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$
樣本平均值
 - The sample median is $\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd.} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{if } n \text{ is even.} \end{cases}$
樣本中位數
- Example: If the data set is the following: 1.7, 2.2, 3.11, 3.9, and 14.7, then $\bar{x} = 5.72$, $\tilde{x} = 3.11$
- The computation of \bar{x} is the basis of an **estimate of the population mean** in statistical inference.

藉由樣本平均值推導出母體平均值

Other Measures of Locations

- A trimmed mean is computed by “trimming away” a certain percent of both the largest and smallest set of values. 去除極端值
- Example
 - The 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.
 - So, for the with nitrogen group the 10% trimmed mean is

No nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

$$\begin{aligned}\bar{x}_{tr(10)} &= \frac{0.43 + 0.47 + 0.49 + 0 + 0.75 + 0.79 + 0.62 + 0.46}{8} \\ &= 0.56625\end{aligned}$$

1.4 Measure of Variability

資料零散程度

- Process and product variability is a fact of life in engineering and scientific systems: **the control or reduction of process variability** is often a source of major difficulty.
- The sample range, $X_{max} - X_{min}$, has applications in the area of **statistical quality control**.
- The sample variance is donated by $s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$
- The sample standard deviation is donated by

$$s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}}$$

樣本標準差

Measure of Variability

- Example 1.4: An engineer is interested in testing the “bias” in a PH meter. Data are collected on the meter by measuring the PH of a neutral substance (H = 7.0). A sample of size 10 is taken with results given by

7.07	7.00	7.10	6.97	7.00
7.03	7.01	7.01	6.98	7.08

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \dots + 7.08}{10} = 7.0205$$

$$s^2 = \frac{(7.07 - 7.0205)^2 + (7.00 - 7.0205)^2 + (7.10 - 7.0205)^2 + \dots + (7.08 - 7.0205)^2}{10 - 1}$$

$$= 0.001939$$

$$s = \sqrt{0.00193} = 0.0440$$

Measure of Variability

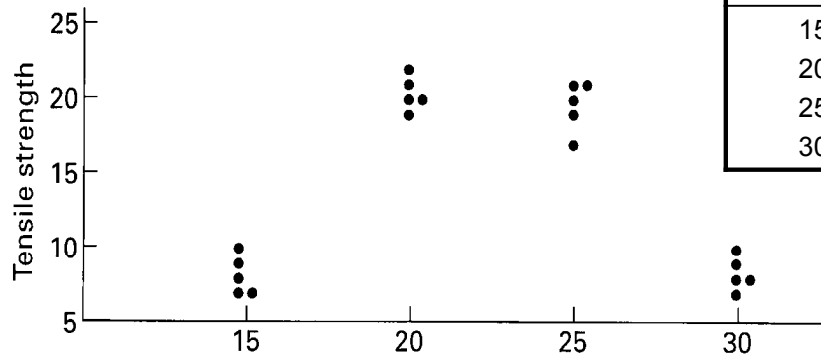
- In the context of statistical inference
 - Usually, focus on drawing conclusions about characteristics of populations, called population parameters.
 - Population mean and population variance are two important parameters. 從有限資料對未知母體評估
 - The sample mean plays an explicit role to draw inferences about the population mean.
 - The sample variance (standard deviation) plays an explicit role to draw inferences about the population variance (standard deviation).

1.6 Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

- The result of a statistical analysis is the estimation of parameters of a postulated (假設的) model. 機率分佈模型
- A statistical model is **not deterministic** but, rather, 資料的不確定性，代表其存在變異性 must entail (意味著) some probabilistic aspects.
- A model form is often the foundation of assumptions that are made by the analyst. 使用機率分佈模型，需確定資料符合分佈模型的條件
 - Example 1.2 scientists draw some distinction between “nitrogen” and “no-nitrogen” populations through the sample information.
 - The analysis may require a certain model for the data, e.g., normal (Gaussian) distributions (see Chapter 6).

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

- Some simple **graphics (plots)** can shed important light on the **clear distinction between the samples**, e.g., means and variability.
- Often, plots can illustrate 從圖表中看出資料分佈，得知其變異性 information that sometimes are not retrieved from the formal analysis.



Cotton percentage	Tensile strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

從圖表分析中可得到更多資訊，配合數學函式假設，得到更有用的結果

Figure 1.5 Plot of tensile strength and cotton percentages.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

- It is likely that the scientist anticipates the existence of a maximum population mean tensile strength.
- Here the analysis of the data should revolve around a **different type of model**, whose **structure** relating the population mean tensile strength to the cotton concentration.
 - E.g., a regression model $\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2$ where $\mu_{t,c}$ is the population mean of tensile strength, which varies with the amount of cotton in the product C .
 - The use of an **empirical model** is accompanied by **estimation theory**, where $\beta_0, \beta_1, \beta_2$ are estimated by the data.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

- The **type of model** used to describe the data often depends on the **goal of the experiment**.
- The **structure** of the model should take advantage of nonstatistical scientific input.
- A **selection of a model** represents a **fundamental assumption** upon which the resulting **statistical inference** is based.
- Often, **plots (graphics)** can illustrate information that allows the results of the **formal statistical inference** to be better communicated to the scientist or engineer, and teach the analyst something not retrieved from the formal analysis.

Graphical Methods and Data Description

- Characterizing or **summarizing** the nature of collections of data is important.
- A summary of a collection of data via **a graphical display can provide insight** regarding the system from which the data were taken.
- Example

Table 1.4 Car Battery Life							
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
.....							

Table 1.5 Stem and Leaf Plot of Battery Life		
Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

Table 1.6 Double-Stem and Leaf Plot of Battery Life		
Stem	Leaf	Frequency
1.	69	2
2*	2	1
2.	5669	4
3*	001111222333444	15
3.	5567778899	10
4*	11234	5
4.	577	3

2.0-2.4

2.5-2.9

Usually, we choose between 5 and 20 stems.

Table 1.7 Relative Frequency Distribution of Battery Life

Class interval	Class midpoint	Frequency, f	Relative Frequency
1.5-1.9	1.7	2	0.05
2.0-2.4	2.2	1	0.025
2.5-2.9	2.7	4	0.100
3.0-3.4	3.2	15	0.375
3.5-3.9	3.7	10	0.250
4.0-4.4	4.2	5	0.125
4.5-4.9	4.7	3	0.075

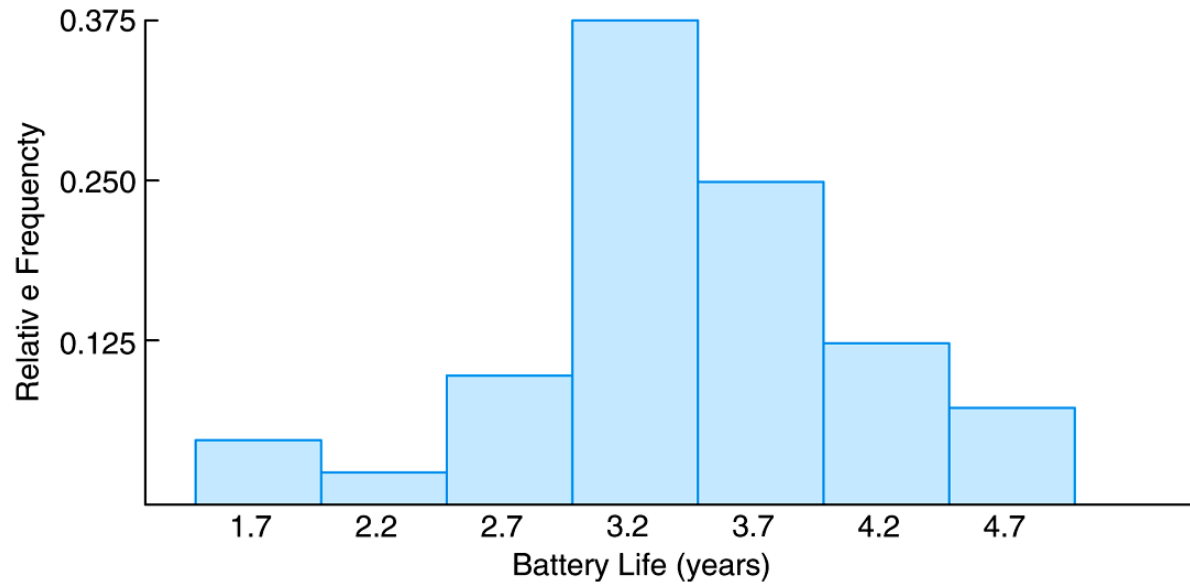


Figure 1.6 Relative frequency histogram 直方圖

- Continuous frequency distribution in Figure 1.7:
Bell-shaped curve
- **Distribution** (probability distribution) is a property of the population (Chapters 5 & 6)

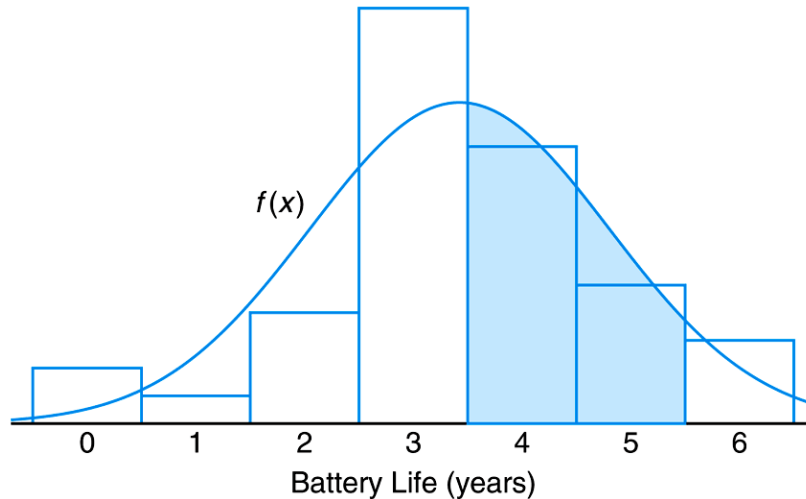


Figure 1.7 Estimating frequency distribution

- A distribution is **symmetric** if it can be folded along a vertical axis so that the two sides coincide, otherwise **skewed**.
- By rotating a stem and leaf plot counterclockwise through an angle of 90° , the resulting columns of leaves form a picture that is similar to a **histogram**.

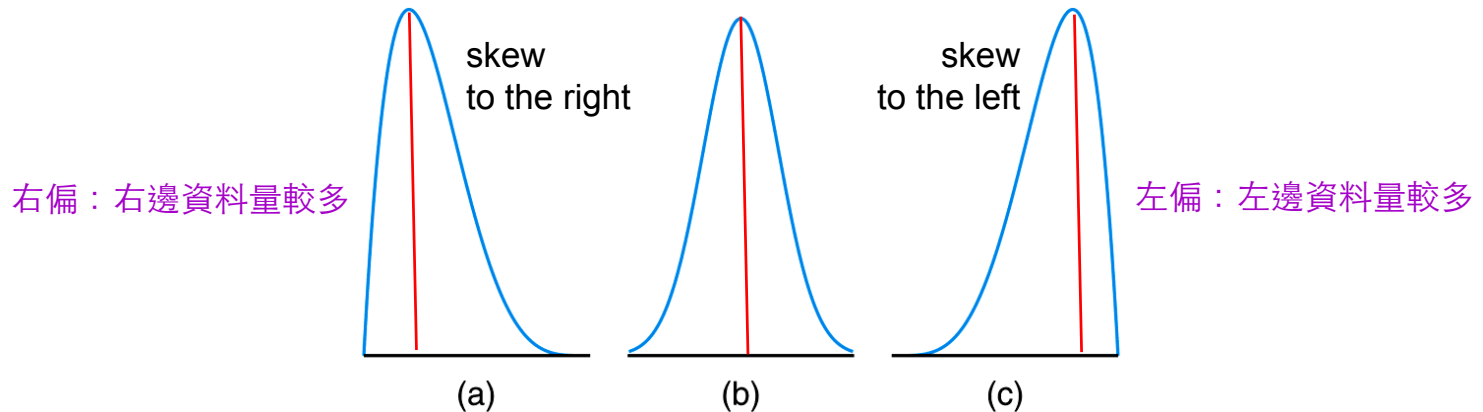


Figure 1.8 Skewness of data

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

Table 1.8 Nicotine Data for Example 1.5

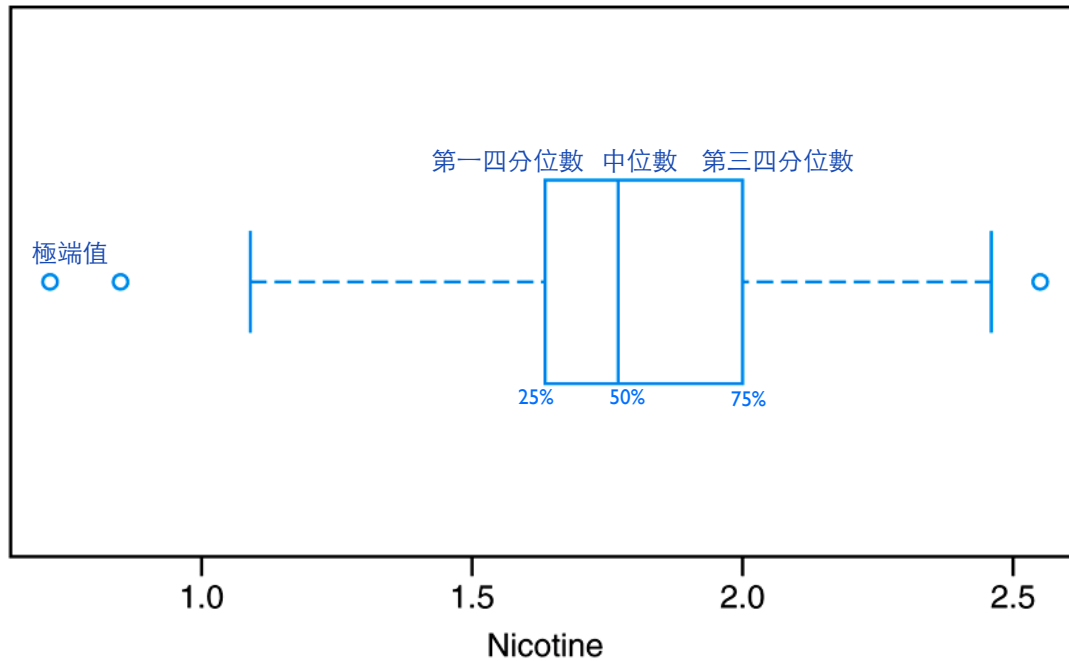


Figure 1.9 Box-and-whisker plot for Example 1.5

The decimal point is 1 digit(s) to the left of the |

7		2	小數第二位
8		5	
9			
10		9	
11			
12		4	
13		7	
14		07	
15		18	
16		3447899	
17		045599	
18		2568	
19		0237	
20		389	
21		17	
22		8	
23		17	
24		6	
25		5	

Figure 1.10 Stem-and-Leaf plot for the nicotine data

Sample	Measurements	Sample	Measurements
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

Table 1.9 Data for Example 1.6

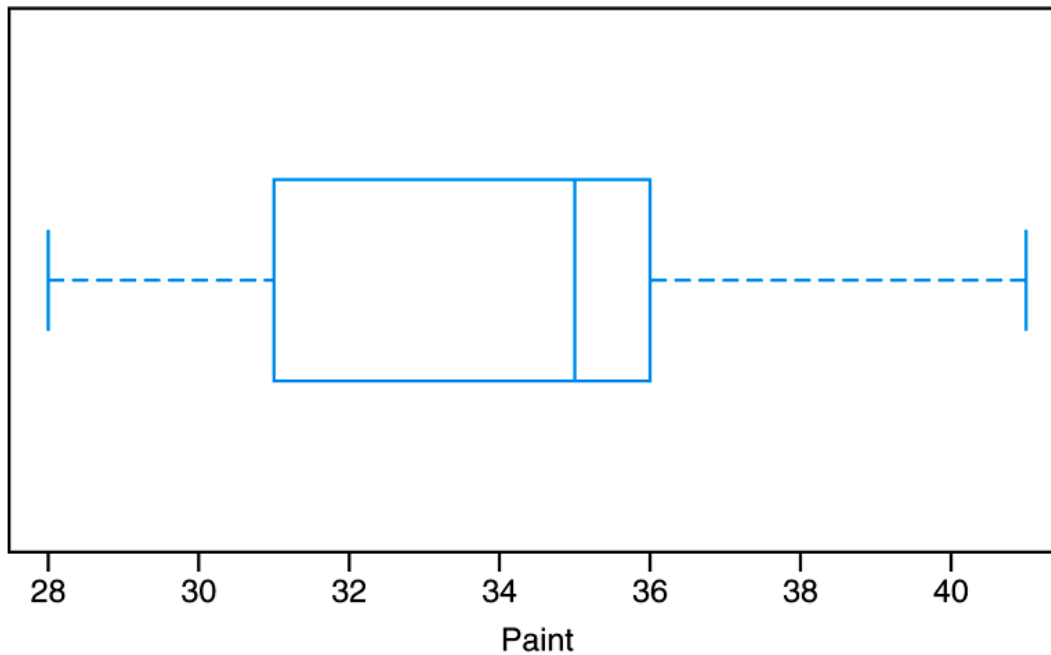


Figure 1.11 Box-and-whisker plot for thickness of paint can “ears”

Exercise

- 1.18
- 1.22 (Matlab)
 - (a) mean, std
 - (b) hist
- Due 3/11/2014