Master of Space Big Data Analytics
Course MSDZ06

Class 1
Statistical Analysis and Inference

Min Ding 丁忞
Tuesday, Jan. 19, 2021

Intended Learning Outcomes

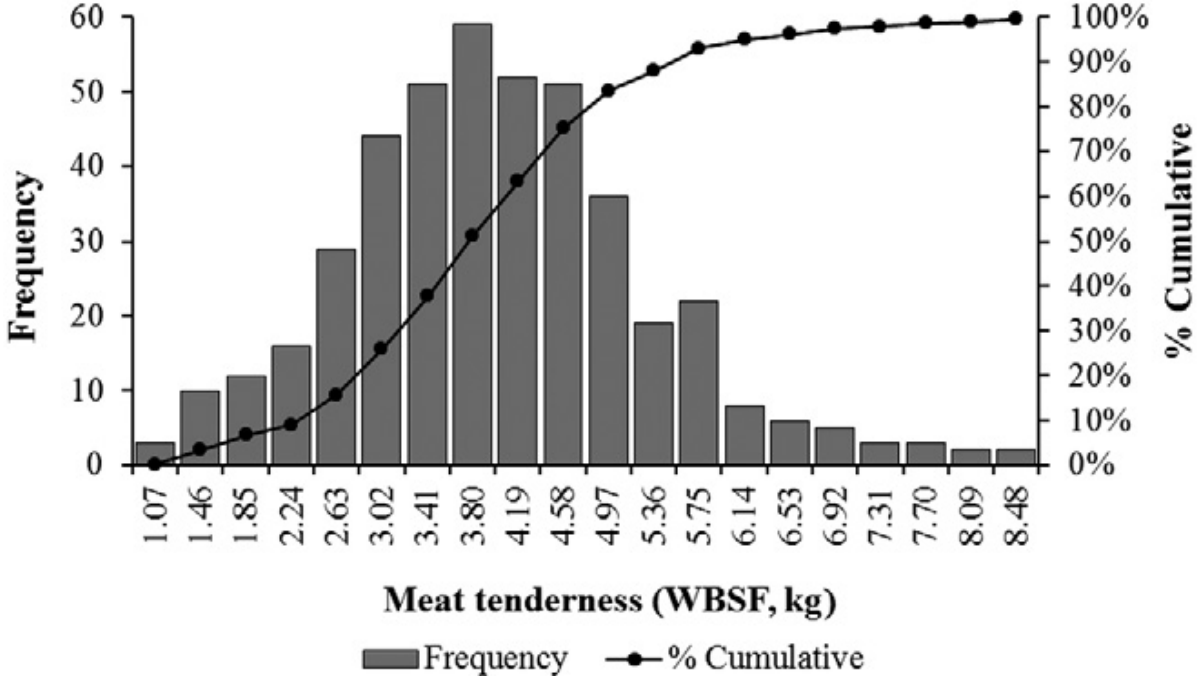
At the end of this class, you will be able to conduct/estimate for a given dataset using Matlab:

1. Visualization and Graphical Exploration
2. Parameter Estimation
3. Confidence Interval and Hypothesis Test
4. Bootstrapping Confidence Intervals

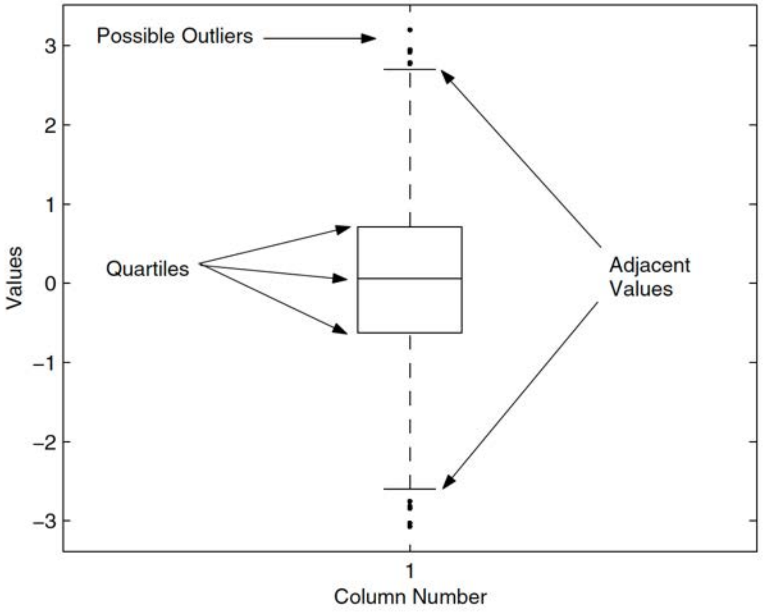
Prerequisite course: Probability and statistics 概率论与数理统计

1. Visualization and Graphical Exploration: Univariate Data

hist()

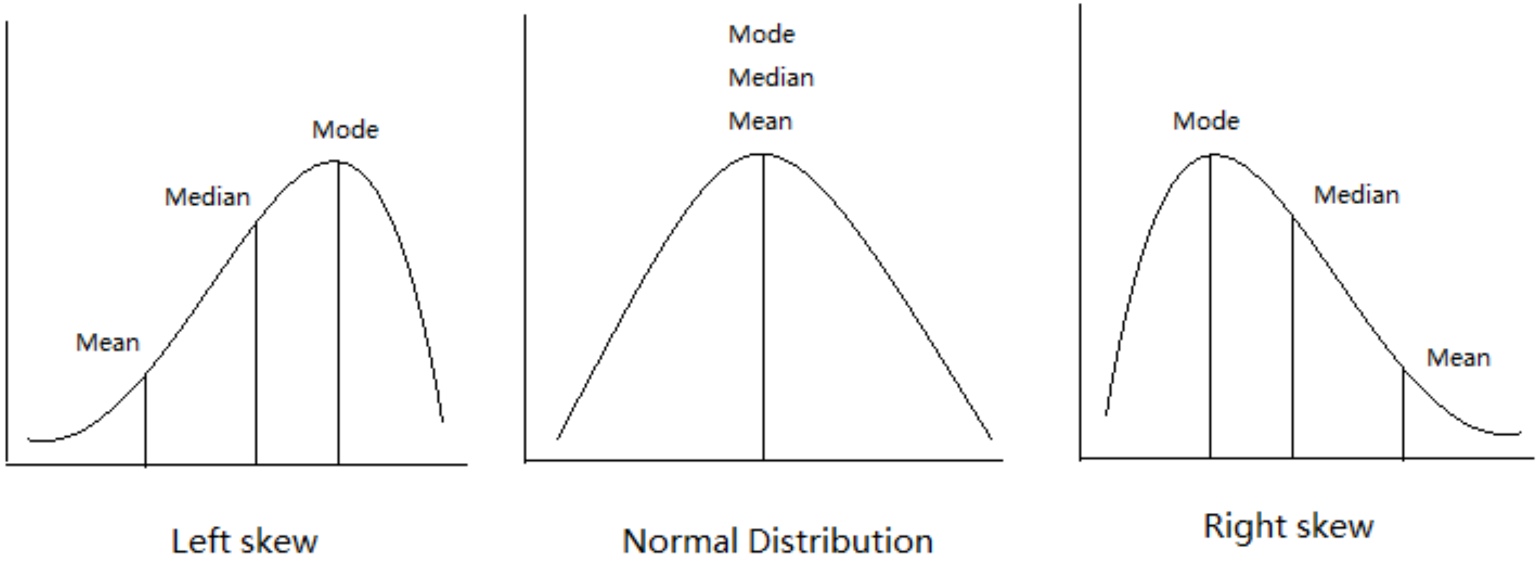


boxplot()



Descriptive Statistics

- Location: Mean, Median, Mode
- Variability: Range, Standard Variance/Deviation
- Shape: Skewness (symmetric? 偏度), Kurtosis (flat? 峰度)
- Percentile & quantile (Q1-Q4)



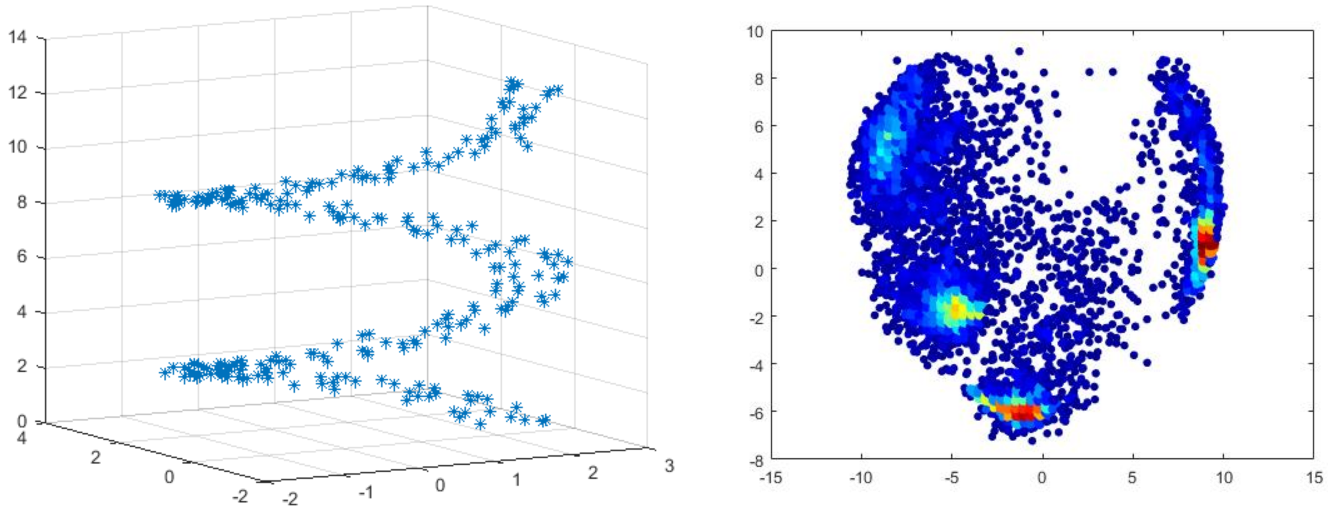
Left skew

Normal Distribution

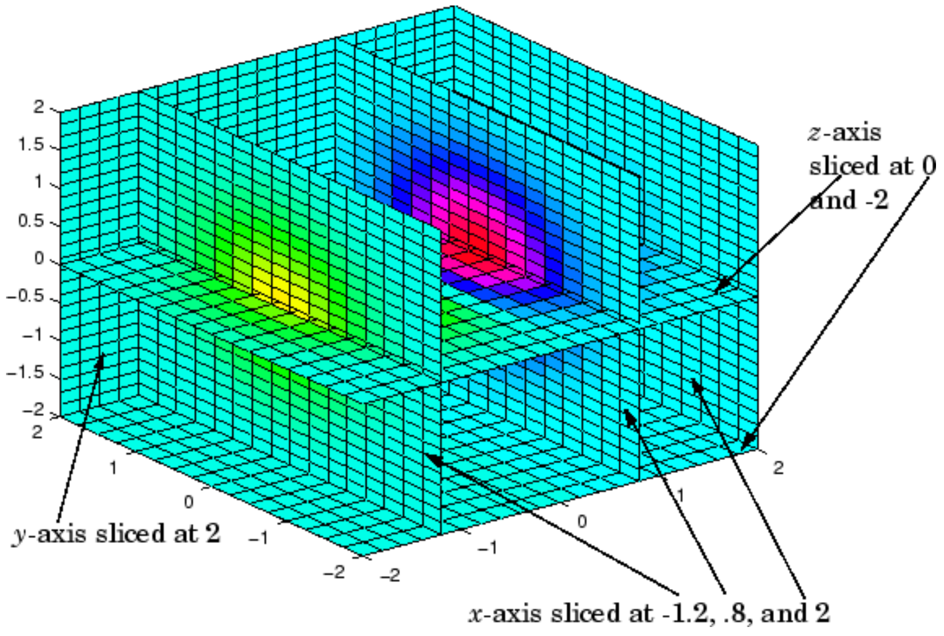
Right skew

1. Visualization and Graphical Exploration: 2D/3D Data

scatter(); scatter3()

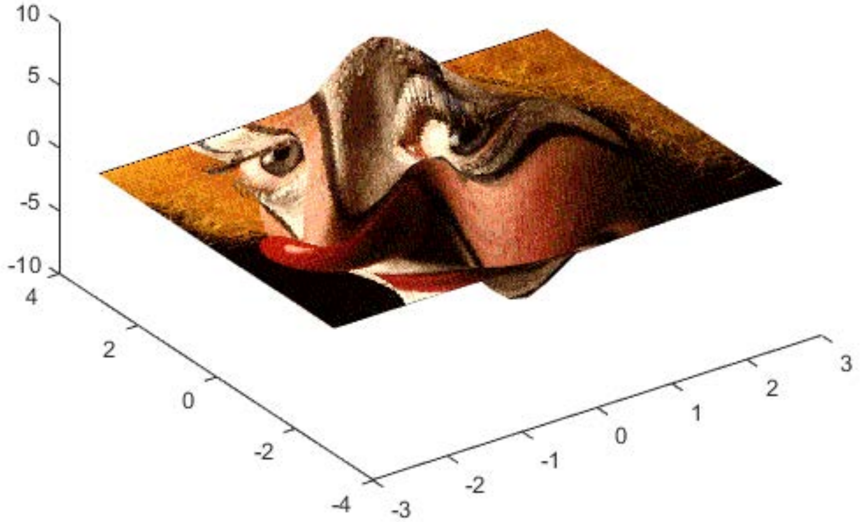


slice()

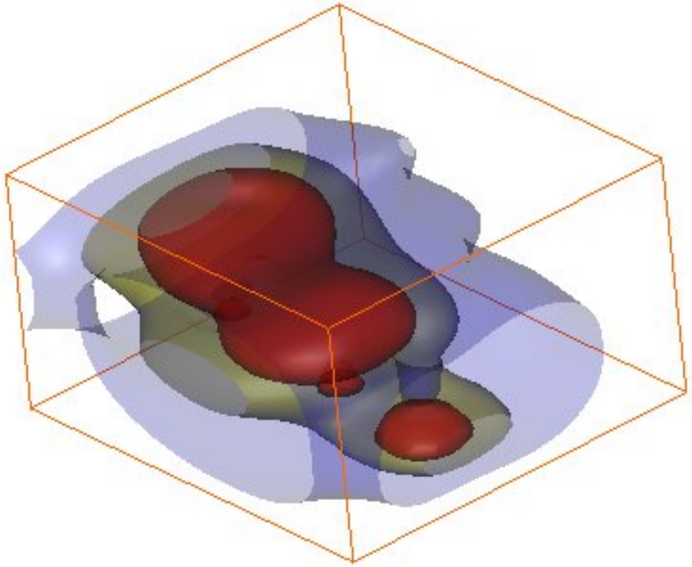


1. Visualization and Graphical Exploration: 2D/3D Data

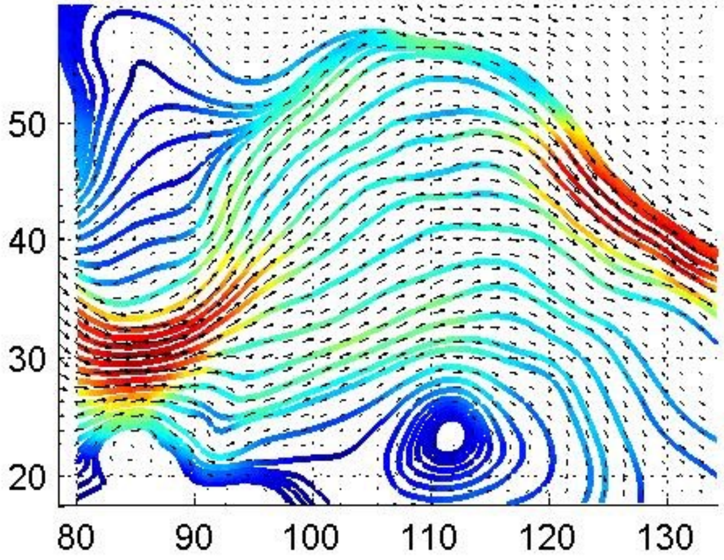
surface()



isosurface()

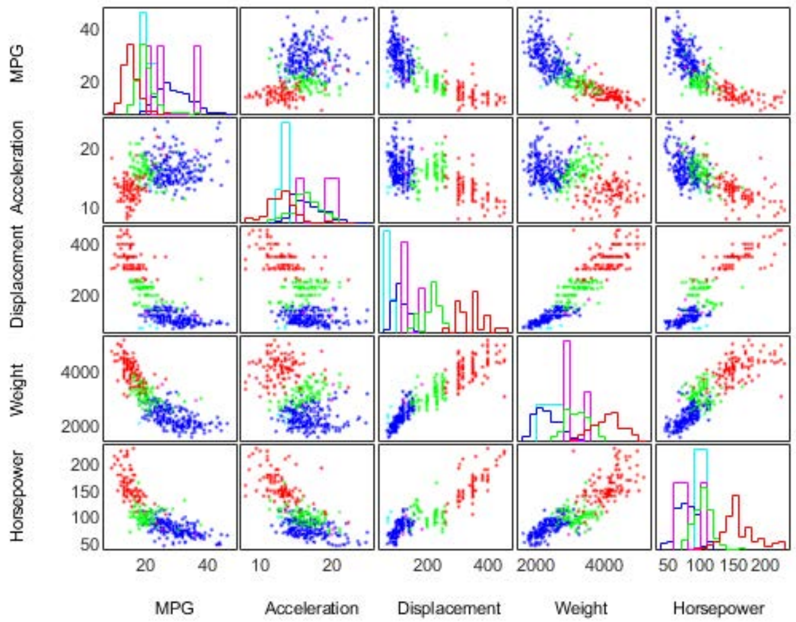


stream2()+quiver()

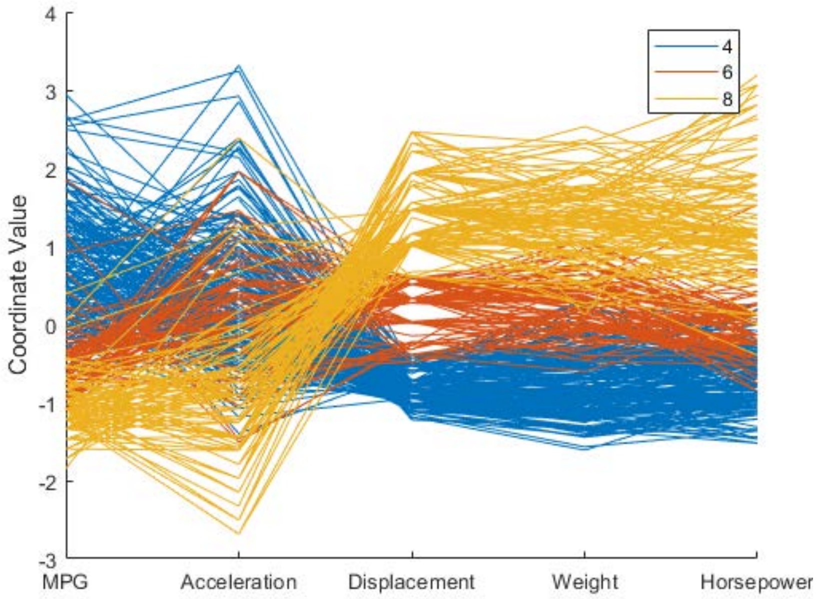


1. Visualization and Graphical Exploration: Multidimensional Data

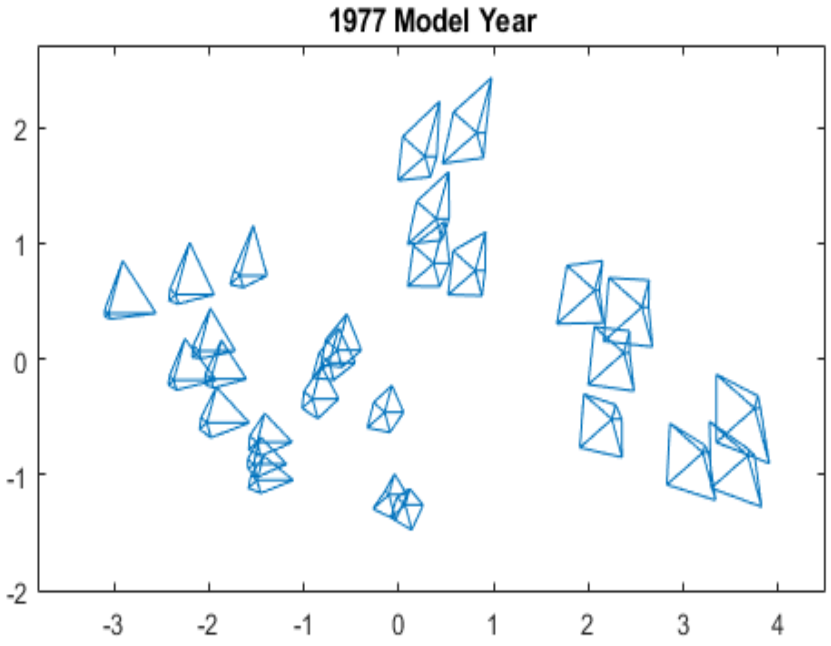
plotmatrix()



parallelcoords()

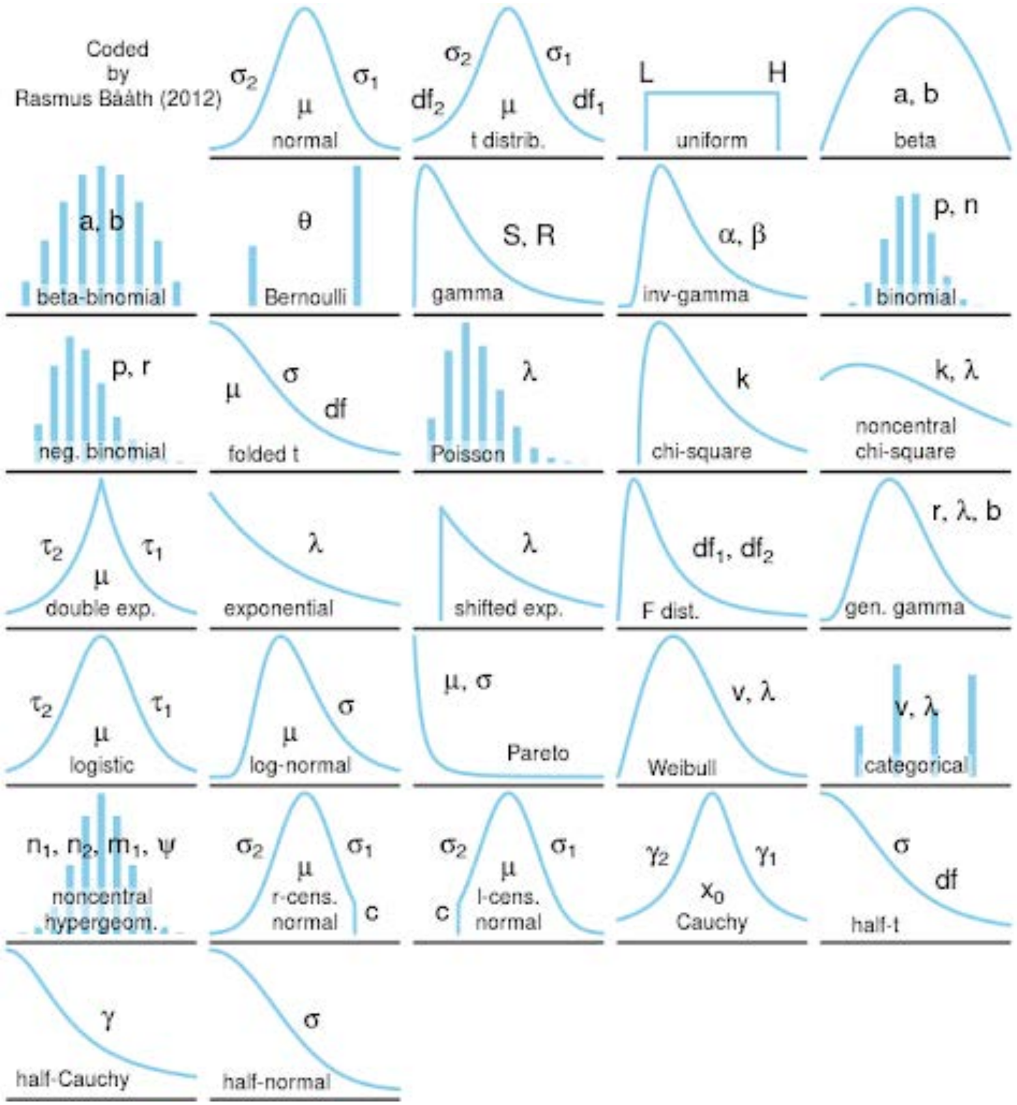


glyphplot()



2. Statistical Parameter Estimation

- For common probability distribution:



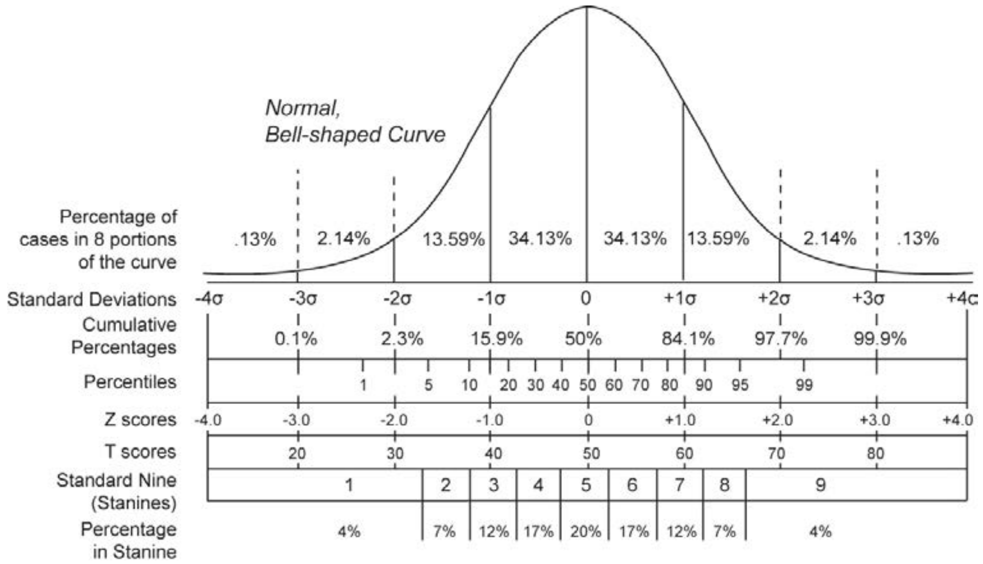
2. Statistical Parameter Estimation

Goal: Use **samples** (X_1,X_2, …, X_N) to estimate the **parameters of the population** from where the samples are drawn from.

Example: Use **maximum likelihood estimation (MLE)** to infer the mean (μ) and standard error (σ) of a Gaussian (normal) distribution. 最大似然估计

Normal Distribution Probability Density Function (PDF):

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Example 3.3

In this example, we derive the maximum likelihood estimators for the parameters of the normal distribution. We start off with the likelihood function for a random sample of size n given by

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Since this has the exponential function in it, we will take the logarithm to obtain

$$\ln[L(\theta)] = \ln\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\right] + \ln\left[\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\right].$$

This simplifies to

$$\ln[L(\theta)] = -\frac{n}{2}\ln[2\pi] - \frac{n}{2}\ln[\sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{3.25}$$

with $\sigma > 0$ and $-\infty < \mu < \infty$. The next step is to take the partial derivative of Equation 3.25 with respect to μ and σ^2. These derivatives are

$$\frac{\partial}{\partial\mu}\ln L = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \tag{3.26}$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{3.27}$$

We then set Equations 3.26 and 3.27 equal to zero and solve for μ and σ^2. Solving the first equation for μ, we get the familiar sample mean for the estimator.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0,$$

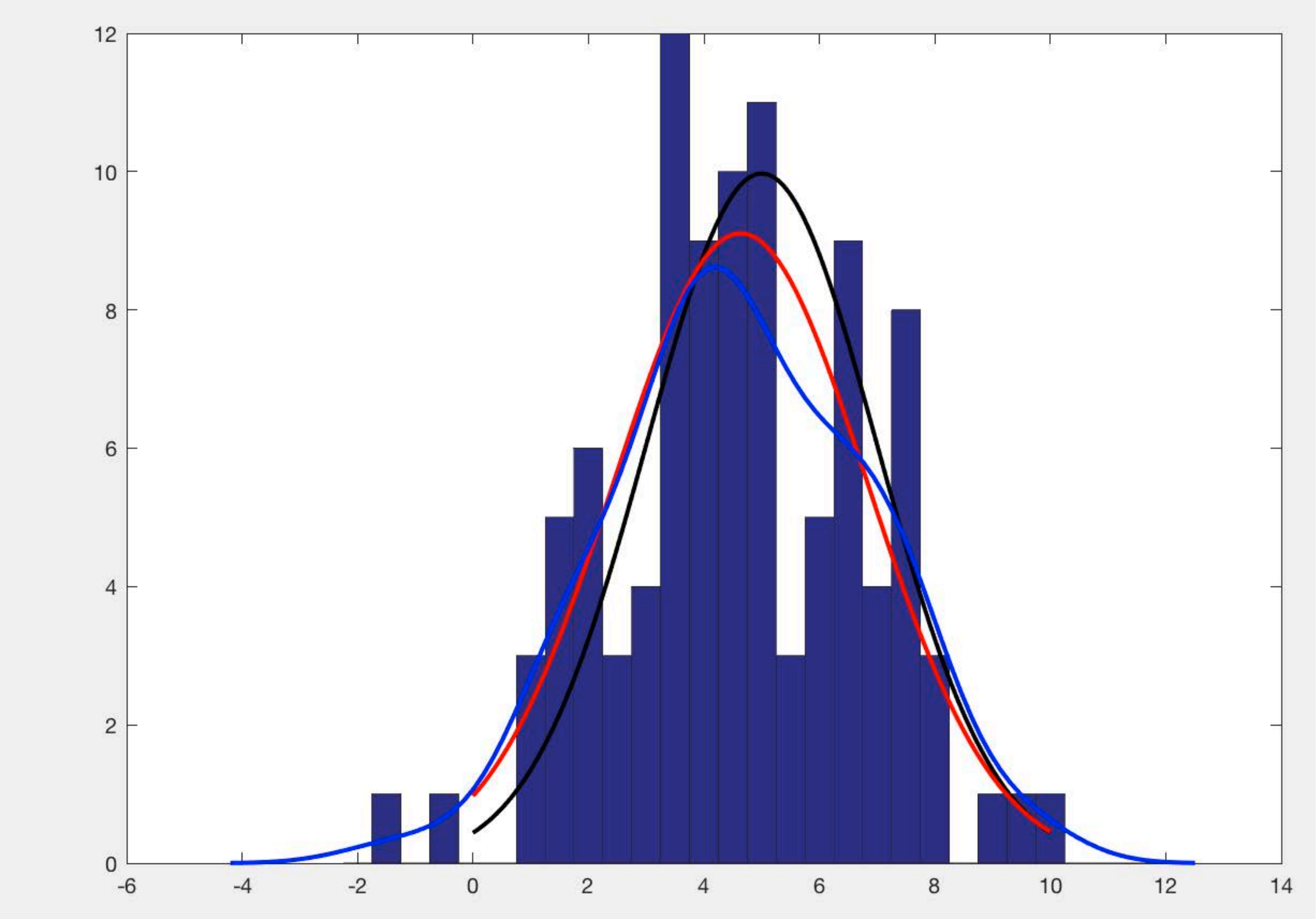
$$\sum_{i=1}^{n} x_i = n\mu,$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Substituting $\hat{\mu} = \bar{x}$ into Equation 3.27, setting it equal to zero, and solving for the variance, we get

$$\begin{aligned} -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= 0 \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned} \tag{3.28}$$

EXAMPLE 1

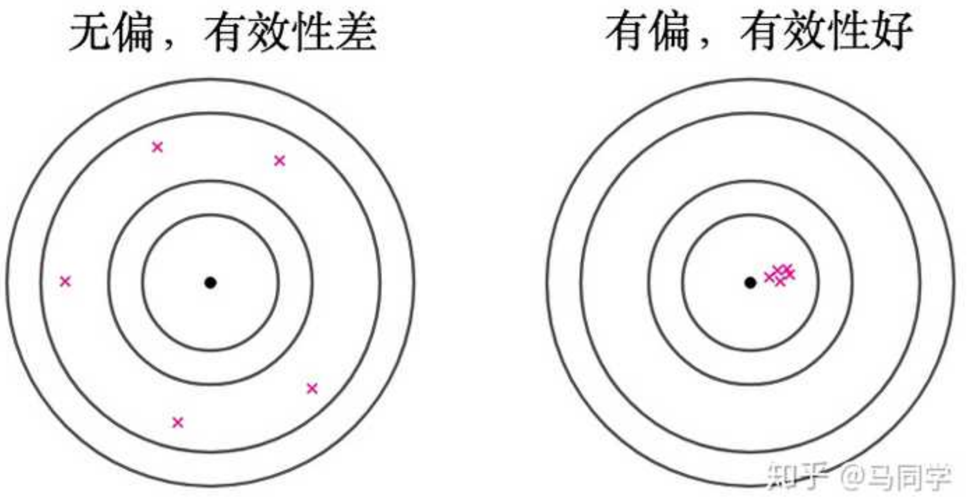


MLE vs MVUE

无偏估计

The *minimum variance unbiased estimator* (MVUE) is commonly used to estimate the parameters of the normal distribution. The MVUE is the estimator that has the minimum variance of all unbiased estimators of a parameter.

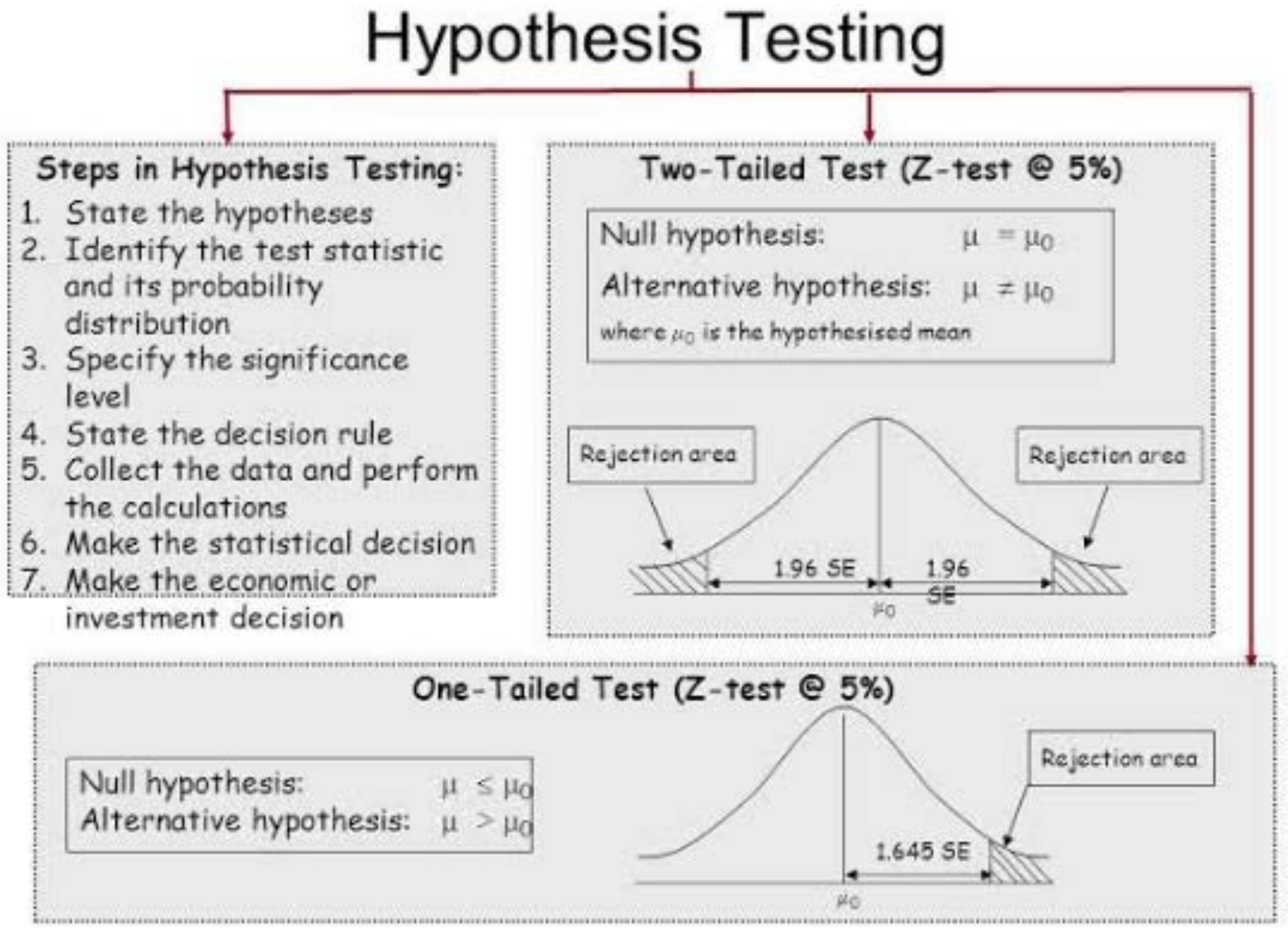
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}},$$



为什么样本方差（sample variance）的分母是 n-1？

3. Hypothesis Test and Confidence Interval

- Goal: Address how **reliable** our estimator is.



Hypothesis Testing

Steps in Hypothesis Testing:

1. State the hypotheses
2. Identify the test statistic and its probability distribution
3. Specify the significance level
4. State the decision rule
5. Collect the data and perform the calculations
6. Make the statistical decision
7. Make the economic or investment decision

Two-Tailed Test (Z-test @ 5%)

Null hypothesis: $\mu = \mu_0$

Alternative hypothesis: $\mu \neq \mu_0$

where μ_0 is the hypothesised mean

One-Tailed Test (Z-test @ 5%)

Null hypothesis: $\mu \leq \mu_0$

Alternative hypothesis: $\mu > \mu_0$

Example 2: Two-Tailed Hypothesis Test for Population Mean

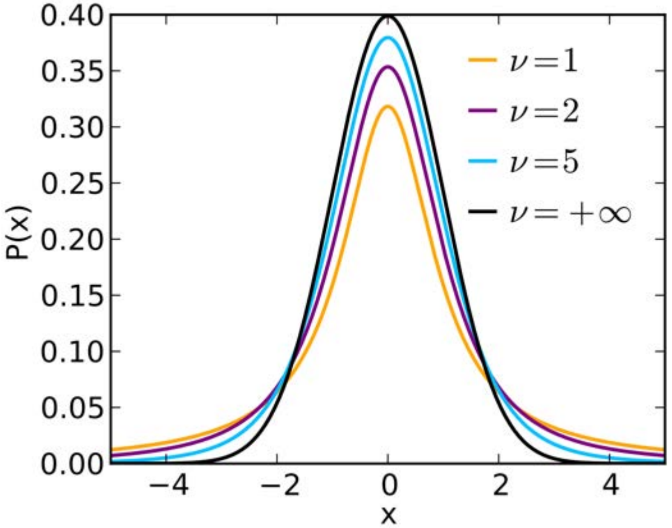
Assume a normal distribution with a mean value of 5. Randomly draw a sample of size 30 from the normal distribution, determine whether this sample is typical.

1. State the hypothesis:

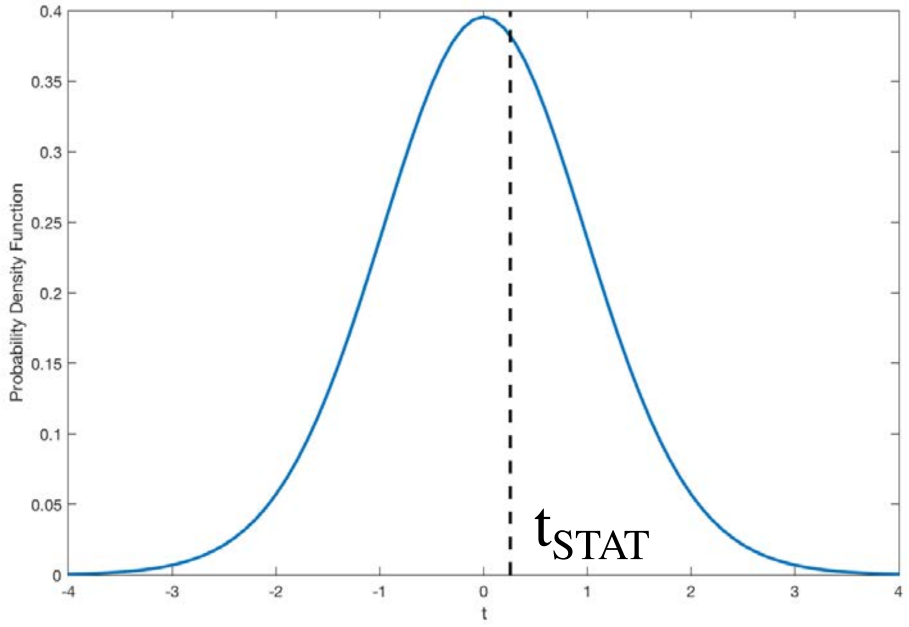
- Null hypothesis H_0: μ = 5 (this sample is typical)
- H1: μ ≠ 5 (this sample is not typical; two-tailed test)

2. Identify the test statistic and its distribution

- For unknown population standard deviation, $t_{STAT} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = 2.3172$ follows a Student's t-distribution with a degree of freedom $\upsilon = $ n-1 = 29.
- Here population mean μ = 5, sample size n = 30, sample mean $\bar{x}$= 6.1, and sample standard deviation s = 2.6

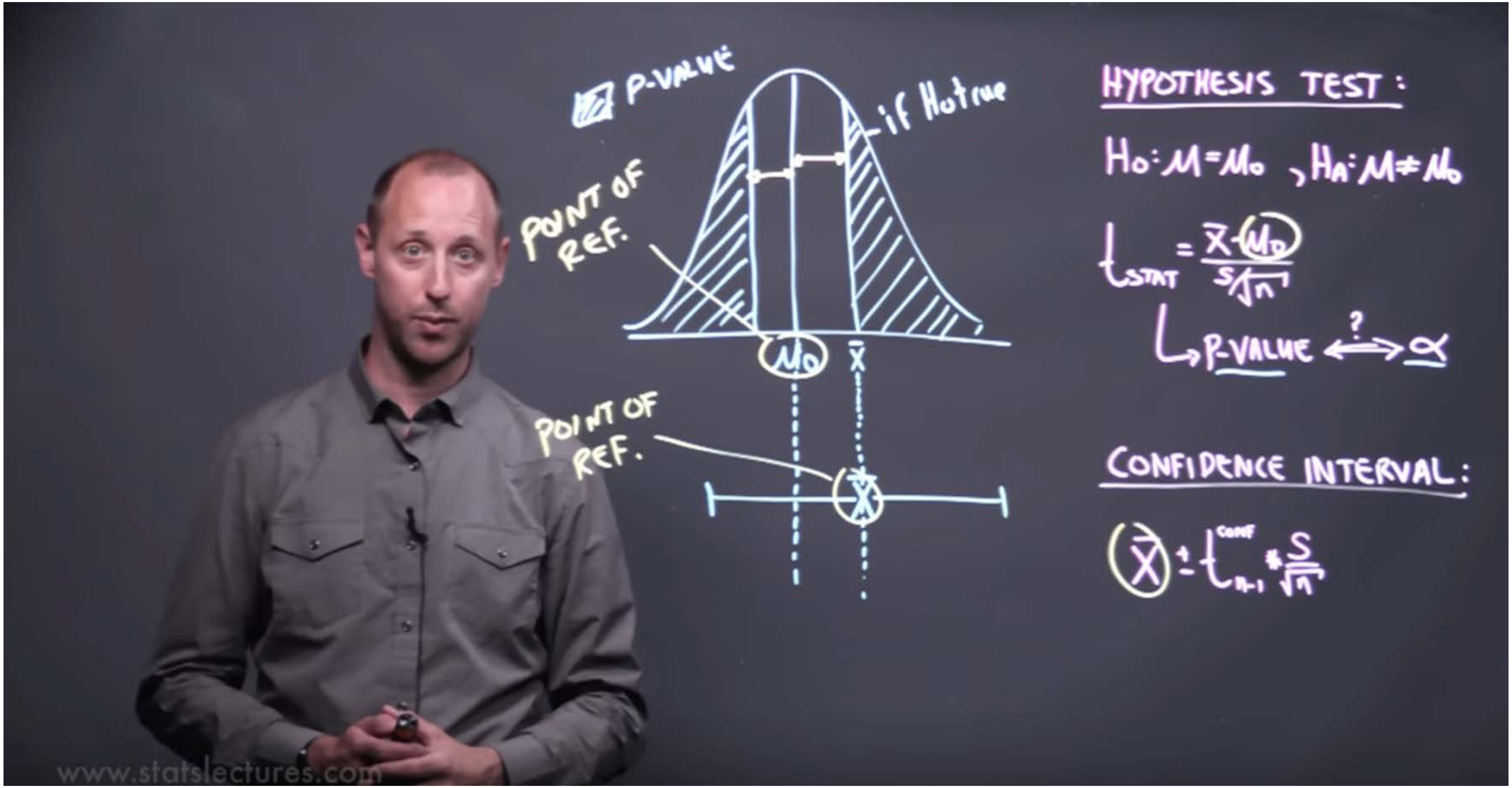


3. Make decision based on significance level: H_0 will be ejected if the test average is either significantly higher or lower than the expected μ = 5, or t_{STAT} significantly deviates from 0. It is not the case here, so we decide can accept the hypothesis (as expected). In Matlab, this process can be done usir



- We usually use ***p-value*** to quantify the probability of obtaining test results as extreme as the results actually observed, under the assumption that the null hypothesis is correct. In this case p-value = 0.79.
- We choose ***a significance level*** α of 0.1 (usually 0.1, 0.05, 0.01) to compare with p-value. Since p-value is larger than α, we conclude that the sample is typical.

Hypothesis Test ⇔ Confidence Interval


P-VALUE
if Ho true
POINT OF REF.
μ_0
$\bar{X}$
POINT OF REF.
$\bar{X}$
HYPOTHESIS TEST:
$H_0: \mu = \mu_0$, $H_A: \mu \neq \mu_0$
$t_{STAT} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
P-VALUE $\overset{?}{\longleftrightarrow}$ α
CONFIDENCE INTERVAL:
$\bar{X} \pm t^{conf}_{n-1} * \frac{s}{\sqrt{n}}$
www.statslectures.com

- https://www.youtube.com/watch?v=J-yMiTaai4c

Example 2: Continued

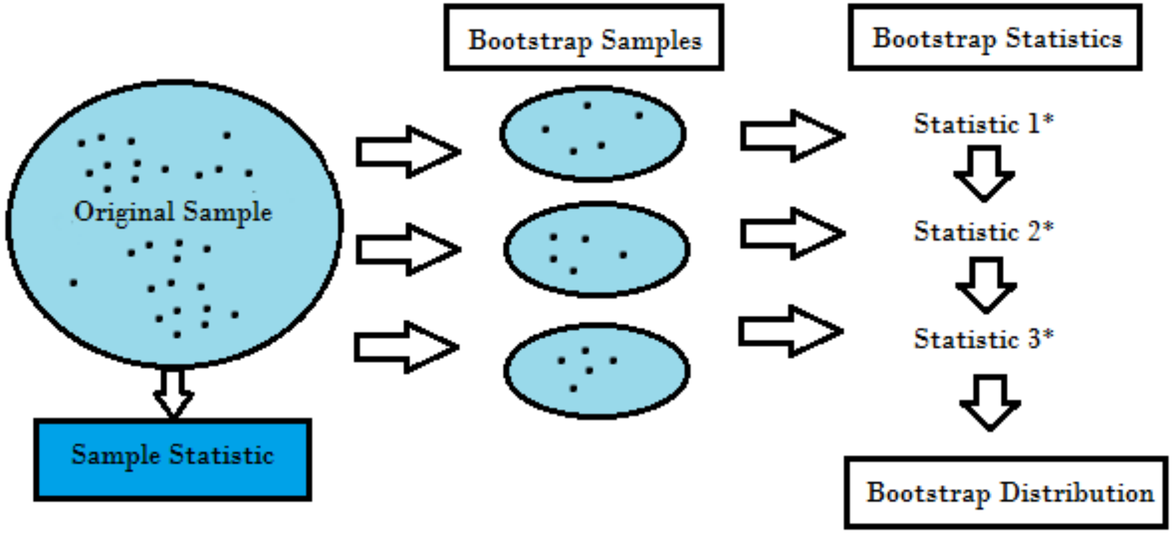
- Calculate the (1- α) confidence interval (CI) for the true population mean using ttest(): 4.17–5.70.
- True population mean is 5, within the CI.
- So the CI and hypothesis test results are consistent.

4. Bootstrapping Confidence Intervals

- It creates multiple resamples (with replacement) from a single set of observations, the bootstrap resamples of the effect size can then be used to determine the CI.

中心极限定理

- The resampling distribution of the difference in means approaches a normal distribution. This is due to the Central Limit Theorem: a large number of independent random samples will approach a normal distribution even if the underlying population is not normally distributed.



Example 4: Use bootstrapping method to estimate the 95% CI for the mean of the sample.

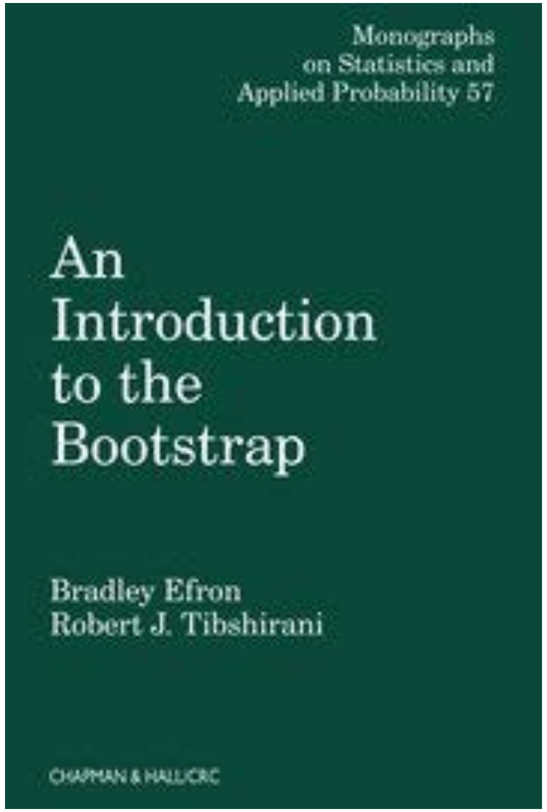
Solution: In matlab, type ci = bootci(10000,@mean,x) to obtain [4.38 5.63] as the CI.

Bootstrap, Jackknife, and Monte Carlo

- Bootstrap (自助法;自助抽樣法;拔靴法) is the most popular resampling method today. It uses sampling with replacement to estimate the sampling distribution for a desired estimator. The main purpose for this particular method is to evaluate the variance of an estimator.
- Jackknife (刀切法) works by sequentially deleting one observation in the data set, then recomputing the desired statistic. It is simpler than bootstrap and a linear approximation of bootstrap method.

- Both of them are resampling method, a way to reuse data to generate new, hypothetical samples (called resamples) that are representative of an underlying population. It's used when:
  - You don't know the underlying distribution for the population,
  - Traditional formulas are difficult or impossible to apply,
  - As a substitute for traditional methods.

- Both of them can be considered Monte Carlo method (蒙特卡洛统计模拟方法): Use random sampling to solve problems.



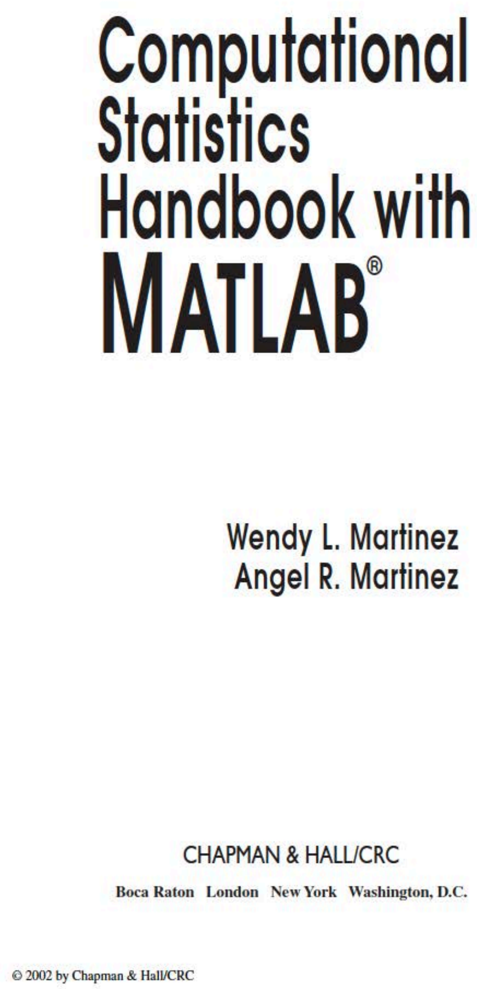
Monographs
on Statistics and
Applied Probability 57

An
Introduction
to the
Bootstrap

Bradley Efron
Robert J. Tibshirani

CHAPMAN & HALL/CRC

References

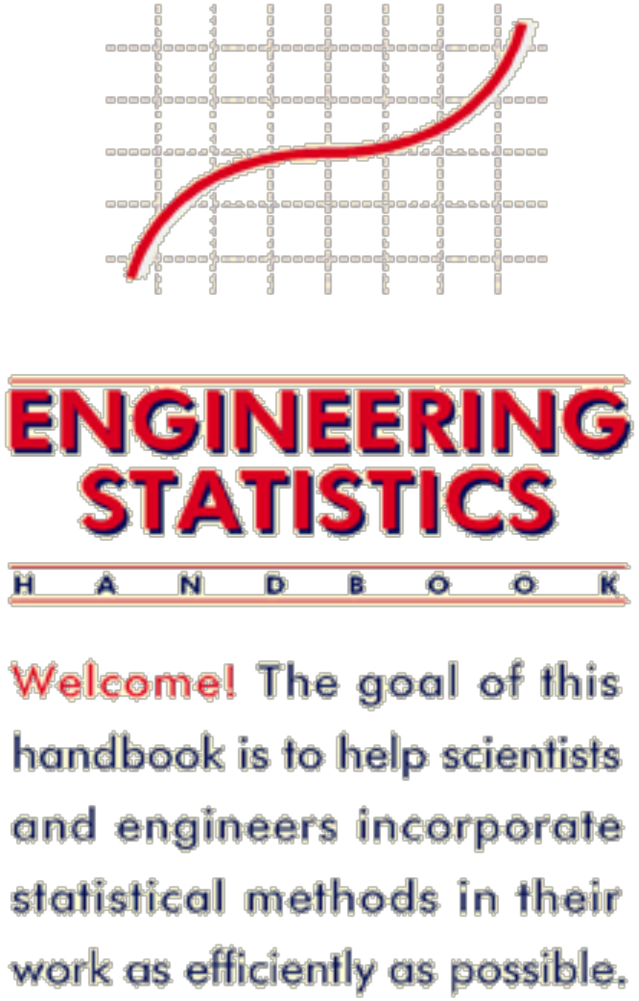


Computational Statistics Handbook with MATLAB®

Wendy L. Martinez
Angel R. Martinez

CHAPMAN & HALL/CRC
Boca Raton London New York Washington, D.C.

© 2002 by Chapman & Hall/CRC



ENGINEERING STATISTICS HANDBOOK

Welcome! The goal of this handbook is to help scientists and engineers incorporate statistical methods in their work as efficiently as possible.

https://www.itl.nist.gov/div898/handbook/index.htm