STAT1005 Foundations of Data Science

Lecture (6): sampling and confidence interval

Yuanhua Huang (黃淵華)

Office: Rm 1-05E, 1/F, JCBIR, 5 Sassoon Road (Medical Campus)

Q&A contact hours: Wed 3-5pm

Email: yuanhua@hku.hk | Web: https://web.hku.hk/~yuanhua

18/10/2021

Department of 統計及精算學系 Statistics & Actuarial Science THE UNIVERSITY OF HONG KONG

HKU Med LKS Faculty of Medicine School of Biomedical Sciences 香港大學生物醫學學院

What to learn in next four weeks

Introduction to Inferential Statistics & Machine learning

- Week 7: Sampling & confidence interval
- Week 8: Hypothesis testing & statistical decision
- Week 9: Regression & Prediction
- Week 10: Classification: Naïve Bayes & Logistic regression

Objectives today

1. Recall Probability and normal distribution
2. Population, sample and sampling methods (bias)
3. Distribution of sampling mean and central limit theorem
4. Confidence interval
5. Bootstrapping

6. Brief introduction of scipy.stats

Recall: probability & statistical model

- **Probability:** numerical descriptions of how likely an event is to occur. The probability of an event is a number between 0 and 1.

- **Random variable**: a variable whose values depend on outcomes of a random phenomenon (a random experiment), with specific probability.

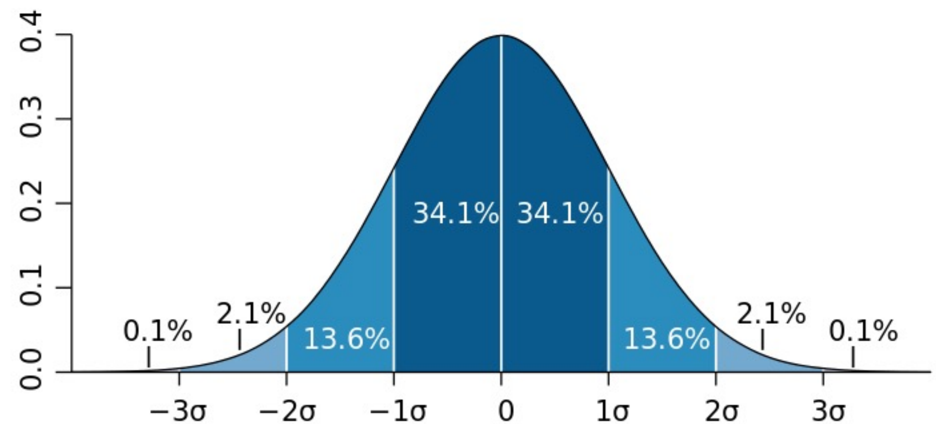
- **Statistical model**: represent, often in considerably idealized form, the data-generating process, e.g., normal distribution.

Recall: Normal distribution

- Normal (or Gaussian) distribution is a type of continuous probability distribution, with a Bell-shaped probability density function (PDF).
- Mean: μ; standard deviation: σ

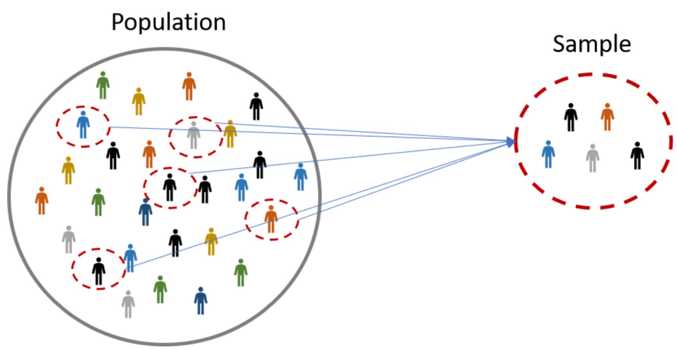
Probability Density Function

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



Sampling | Population and sample

- Population: the whole instances, generally large or even infinite
  - Population mean: μ (unknown; we want to estimate)
  - Population standard deviation: σ (usually unknown)
- Sample: a subset of the instances in the whole population; sample size *n*
  - Sample mean: $\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$
  - Sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$
- Example: the salary of an individual in Hong Kong

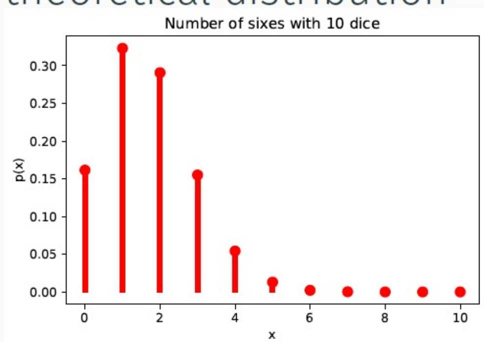


Sampling | Sample statistics

Example

- Number of sixes with 10 dice

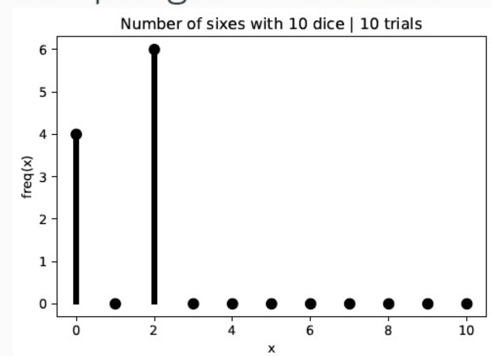
theoretical distribution



$\mu = 1.6667$

$\sigma^2 = 1.3889$

sampling $n = 10$ trials



$\bar{x} = 1.200$

$s^2 = 0.960$

Theory: the random variable X for number of sixes follows binomial distribution Binomial(n=10, p=1/6).

Some contents are from: https://github.com/johnpinney/sampling_and_hypothesis_testing

Sampling | Law of large numbers

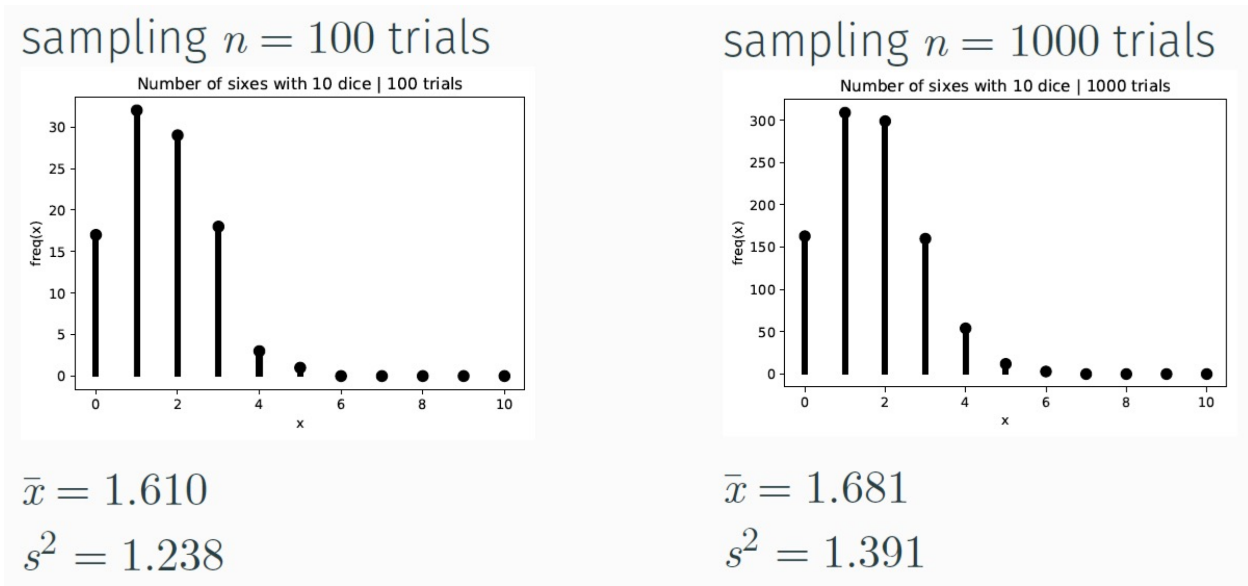
- The law of large numbers states that as we take larger and larger samples of a random variable, the **sample mean** $\bar{x}$ gets closer to the the **population** (or **theoretical**) mean, μ.

- This also implies that the **sample variance** s^2 approaches the **population variance** σ^2 as *n* increases.

Sampling | Law of large numbers

Example

- Number of sixes with 10 dice



Population mean: μ = 1.67; population variance: σ^2 = 1.389

Sampling | Distribution of sample mean

- When we only have access to a finite sample with size n, it is helpful to know how precise our estimate of the population mean will be.

- The observed sample mean, $\bar{x}$ behaves as if it is drawn from a continuous random variable $\bar{X}$ with mean μ and a variance that decreases as n increases.

- $\bar{X}$ is called the sampling distribution of the mean. You could think there are many independent sample sets, e.g., through repeats.

Sampling | Distribution of sample mean

- $\bar{x}$ becomes a more precise estimate of μ as we gather more data.

- We can see this by repeating the sampling process many times and plotting histograms of $\bar{x}$.

- Example: Number of sixes with 10 dice | 1000 replicates of sample

Sampling | Central limit theorem

- For a sample of size *n*, the central limit theorem states that $\bar{X}$ converges to a Normal distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right); \quad \text{for large } n$$

- Note that this is true regardless of the distribution of *X* itself.

- The central limit theorem is the theoretical justification for many statistical procedures.

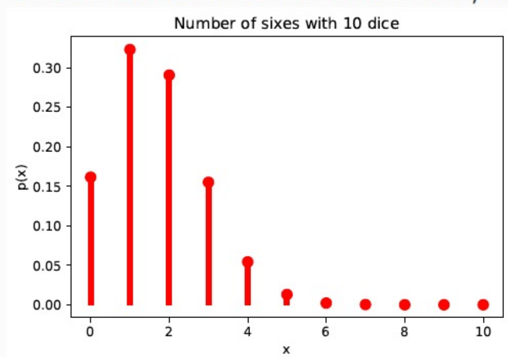
This theorem gives an explanation why many real-word quantities can be approximated by Normal distribution. They may be averaged by many instances.

Sampling | Central limit theorem

Example

- Number of sixes with 10 dice | n = 1000 trials

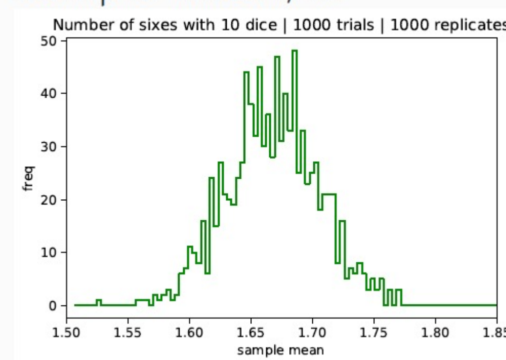
theoretical distribution, X



$$\mu = 1.6667$$

$$\sigma^2 = 1.3889$$

sample mean, $\bar{X}$



$$\text{mean} = 1.6680 \approx \mu$$

$$\text{variance} = 0.0014 \approx \frac{\sigma^2}{n}$$

Sampling | Standard deviation vs. standard error

- The population standard deviation: σ (usually unknown)
- The sample standard deviation: s (calculated from observed data)
- The standard error of the mean: $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$, for large *n*

Sampling | Unbiased estimator for population Var

- When n is small (say n < 75), the sample variance s^2 is not a good approximation for the population variance.
- In fact, it is a biased estimator, which tends to consistently under-predict the value of σ^2.
- We can improve our estimate by using the unbiased sample variance:

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

- Sample variance (divided by *n*): `numpy.var(x, ddof=0)`
- Unbiased estimate of population variance (divided by *n-1*): `numpy.var(x, ddof=1)`

See derivations: https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance

Sampling | Sampling methods

- In many practical applications, the population of interest is not infinite, just very large (e.g., the population of the Hong Kong).

- There are a variety of ways to obtain a representative sample of a finite population, so that the conclusions from the sample are generalisable to the whole population.

Sampling | Random sampling

- *Simple random*: Each individual is chosen randomly and entirely by chance.
- *Systematic*: Every k^{th} individual is sampled from a randomly ordered list.
- *Stratified*: Partition population into heterogenous subpopulations and draw a sample from each one.
- *Cluster*: Total population is split into homogenous clusters, and a subset of clusters is sampled.

Sampling | Non-random sampling

- *Quota*: Interviewers told to sample a certain number of a targeted population.
- *Convenience*: The sample is drawn from the most accessible part of the population.
- *Snowball*: Existing study subjects recruit future subjects from their acquaintances.
- *Voluntary*: Study subjects are self-selected.

Parameter Estimation

Parameter Estimation | Point estimates

- We have seen how to derive an estimated mean and variance for a population, based on a sample.

$$\hat{\mu} = \bar{x} = \frac{\sum x}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

- These are examples of point estimates, where we quote a single value for a population parameter without an associated uncertainty.

Parameter Estimation | Confidence interval (*CI*)

- However, it is often more helpful to give a plausible range of values for a parameter, based on the data collected.

- This is known as a confidence interval. It is particularly important when the sample size is small, as it has higher variability of the sample mean.

- Terms
  - Confidence level: the percentage of confidence intervals, e.g., 95% confidence level refers to the range 0.025 to 0.975.
  - Interval endpoints (low or high): the top or bottom of the confidence interval

Point estimate vs Confidence interval (*CI*)

Image adapted from https://moderndive.com/8-confidence-intervals.html#ci-build-up

Parameter Estimation | *CI* for *large* sample size

- Given only one sample set, how to estimate the confidence interval for the population mean?
- The central limit theorem can be used to derive an approximation of the confidence intervals for the population mean, through Normal distribution.

$$\bar{x} \sim N\big(\mu,\ Var(\bar{x})\big); \quad Var(\bar{x}) = \sigma^2/n; \quad \text{for large } n$$

- Equivalent:
  - $\frac{\bar{x}-\mu}{\sqrt{Var(\bar{x})}} \sim N(0,1)$
  - $\mu \sim N(\bar{x},\ Var(\bar{x}))$

We can use sample variance s^2 to approximate population variance σ^2:
$$Var(\bar{x}) = \sigma^2/n \approx s^2/n$$

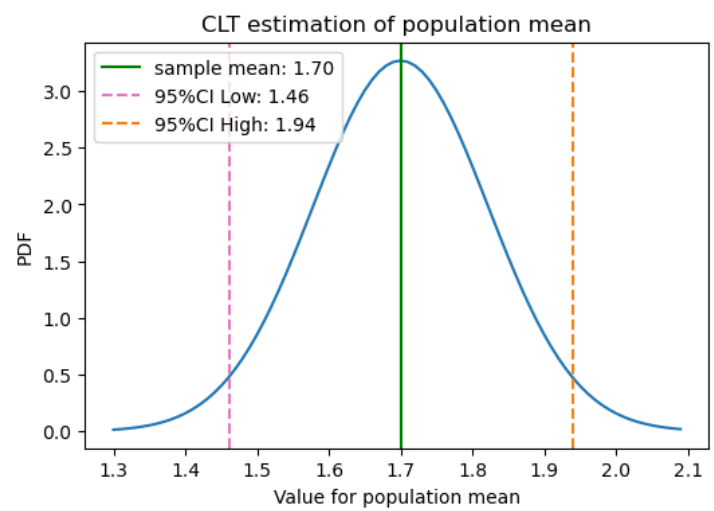
Parameter Estimation | *CI* for *large* sample size

- x: number of sixes in 10 dices. Observed sample (100 observations):

```
[1, 3, 1, 3, 2, 2, 2, 2, 1, 1, 0, 2, 1, 0, 3, 1,
1, 1, 2, 2, 2, 1, 1, 1, 1, 0, 1, 4, 4, 1, 0, 1, 2,
3, 0, 2, 0, 2, 2, 5, 1, 3, 3, 3, 0, 1, 4, 1, 3, 2,
2, 2, 0, 0, 0, 1, 4, 2, 3, 2, 2, 3, 3, 3, 3, 0, 1,
0, 2, 3, 0, 1, 1, 1, 3, 1, 2, 2, 2, 4, 4, 0, 3, 1,
0, 1, 2, 4, 1, 0, 3, 0, 1, 3, 2, 1, 3, 0, 2, 1]
```

Sample: n=100;
$\bar{x}$ =1.7; $s^2 = 1.22^2$; Var($\bar{x}$)= 0.122^2

$$\bar{X} \sim N(1.7, 0.122^2)$$



Parameter Estimation | *CI* for *small* sample size

- The distribution of $\bar{x}$ is very complicated when the sample size *n* is small.
- If the population follows a normal distribution (this is a strong condition!!), we can estimate the confidence interval analytically.

- If we know the variance of the population, σ^2, the population mean can be estimated from a Normal distribution:

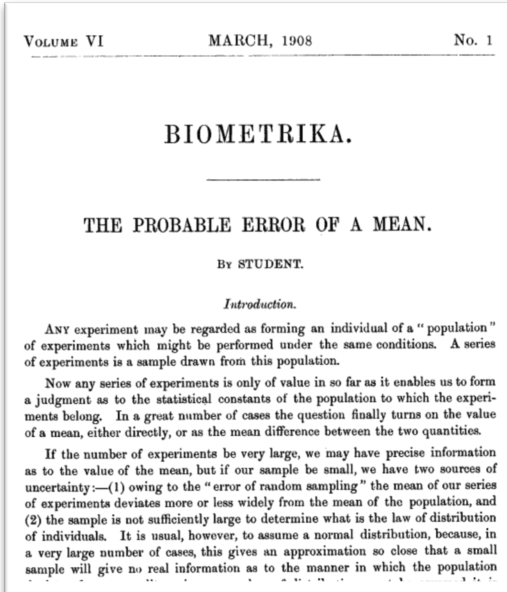
$$\frac{\bar{x}-\mu}{\sqrt{Var(\bar{x})}} \sim N(0,1); \quad Var(\bar{x}\,) = \sigma^2/n$$

- If we don't know the variance of the population, σ^2, we can approximate it with sample variance. The population mean can be estimated from a ***t* distribution** with *n-1* degrees of freedom.

$$\frac{\bar{x}-\mu}{\sqrt{\widehat{Var}(\bar{x}\,)}} \sim t(df = n-1); \quad \widehat{Var}(\bar{x}) = s^2/n$$

Parameter Estimation | Student's *t* distribution

William Sealy Gosset



VOLUME VI — MARCH, 1908 — No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

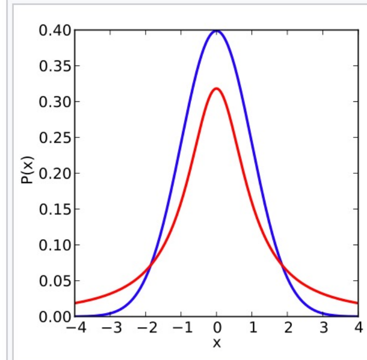
Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

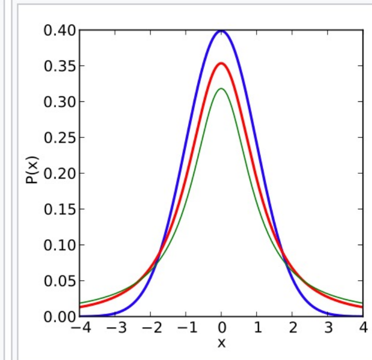
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population

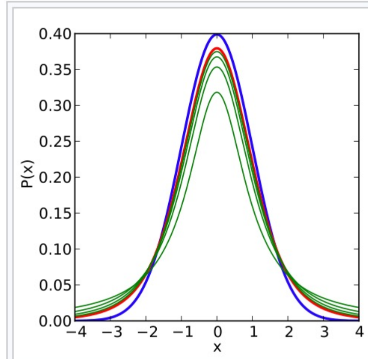
t distribution: estimate the mean of a normally-distributed population in situations where the sample size is small, and the population's standard deviation is unknown.



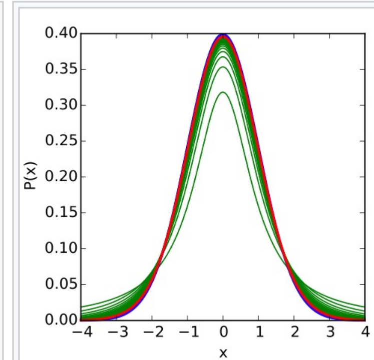
1 degree of freedom



2 degrees of freedom



5 degrees of freedom



30 degrees of freedom

Blue: Normal distribution; Red: *t* distribution; Green: previous *t* distributions

t distribution is close to Normal when df>=30

https://en.wikipedia.org/wiki/Student%27s_t-distribution

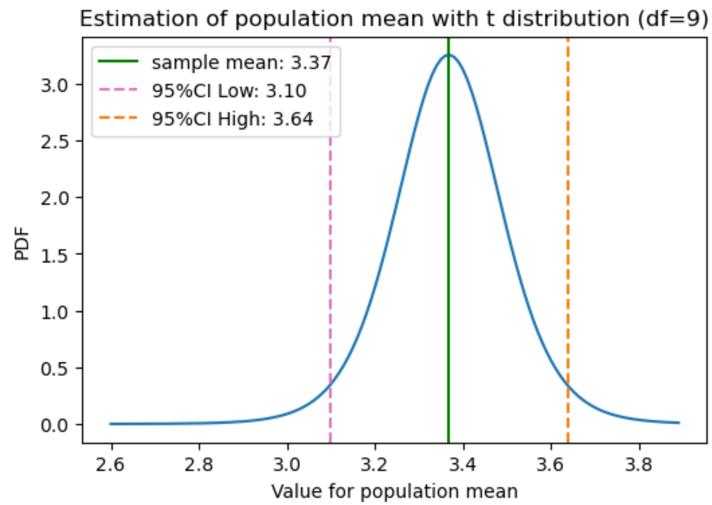
Parameter Estimation | *CI* for *small* sample size

- X: the weight of newborn baby in Hong Kong.
- Observed sample (10 observations):

```
[3.77, 3.24, 3.46, 3.95, 3.81,
2.7, 3.45, 3.02, 3.04, 3.24]
```

Sample: n=10;
$\bar{x}$ =3.37; $s^2 = 0.377^2$; $\text{Var}(\bar{x}) = 0.119^2$

$$\bar{X} \sim t(3.37, 0.119^2; df = 9)$$



Parameter Estimation | Sample size for *CI*

These approximate intervals above are good when n is large (because of the Central Limit Theorem), or when the observations x_1, x_2, ..., x_n are normal.

Sample size (rule of thumb)

- $n \geq 30$: we consider the sample size to be large and by Central Limit Theorem, $\bar{x}$ will be normal even if the sample does not come from a Normal distribution.

- $8 \leq n \leq 29$: check if the data follows a **Normal** distribution first. If it does not violate the normal distribution, then we can go ahead and use the t-interval.

- $n \leq 7$: difficult to check if it follows a Normal distribution. You may consider non-parametric methods rather than t-interval

Bootstrapping

Bootstrapping | definition and principle

Bootstrapping: random resampling with replacement.

Algorithm of Bootstrapping:

- Step1: generate a bootstrap sample
  - Draw a sample value, record it, and then put it back
  - Repeat *n* times
- Step2: calculate the mean (or other statistics, e.g., median) on bootstrap sample
- Step3: Repeat Steps 1 & 2 for *R* times (*R* bootstrapping iterations)

Then, we can obtain an empirical distribution of the sample mean from the *R* bootstrap means.

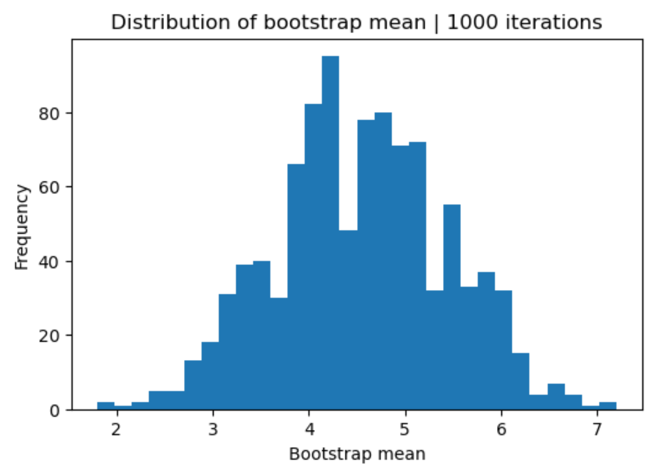
Now, we can:

a) Calculate their standard deviation (to estimate sample mean standard error);
b) Find a confidence interval
c) Visualize the empirical distribution, e.g., by histogram or boxplot

Bootstrapping | Simple example

```
# Generate one bootstrap sample

In [1]: X = np.arange(10)

In [2]: X
Out[2]: array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

In [3]: np.random.choice(X, replace=True, size=10)
Out[3]: array([6, 1, 7, 6, 4, 3, 5, 6, 6, 6])
```

Repeat 1,000 times →
1,000 bootstrap means

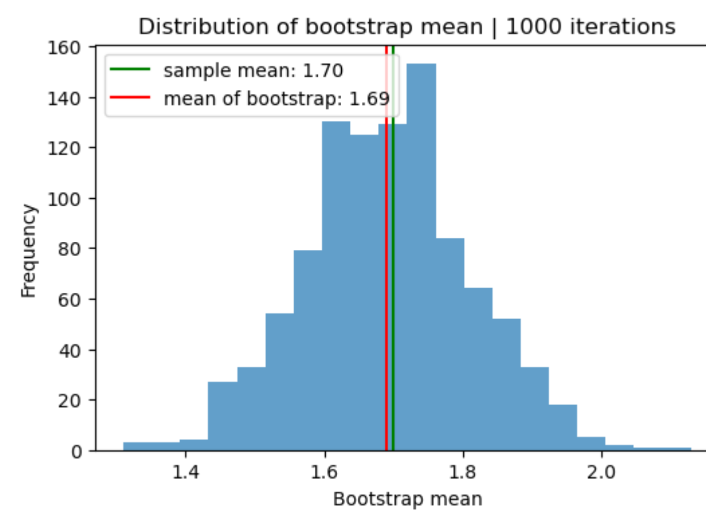


Bootstrapping | Example on dice sixes

- *x*: number of sixes in 10 dices. Observed sample (100 observations):

```
[1, 3, 1, 3, 2, 2, 2, 2, 1, 1, 0, 2, 1, 0, 3, 1,
1, 1, 2, 2, 2, 1, 1, 1, 1, 0, 1, 4, 4, 1, 0, 1, 2,
3, 0, 2, 0, 2, 2, 5, 1, 3, 3, 3, 0, 1, 4, 1, 3, 2,
2, 2, 0, 0, 0, 1, 4, 2, 3, 2, 2, 3, 3, 3, 3, 0, 1,
0, 2, 3, 0, 1, 1, 1, 3, 1, 2, 2, 2, 4, 4, 0, 3, 1,
0, 1, 2, 4, 1, 0, 3, 0, 1, 3, 2, 1, 3, 0, 2, 1]
```

- Distribution of bootstrap means:
  - $n = 100$ (sample size)
  - $R = 1000$ (Bootstrap iterations)

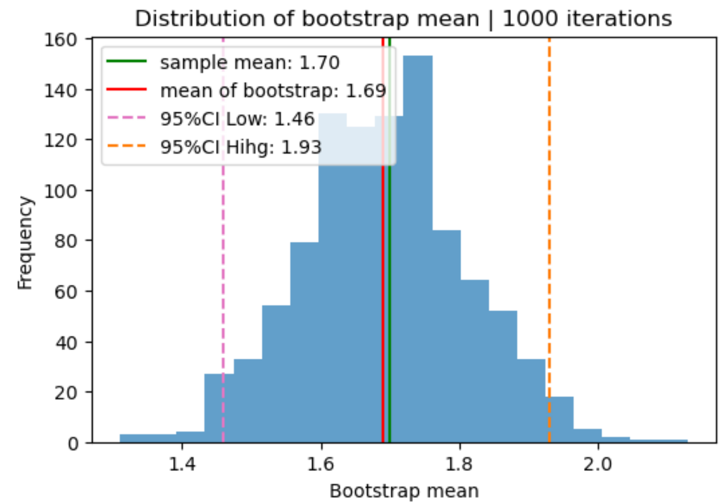


Bootstrapping | Confidence interval

Calculate confidence interval by bootstrapping distributions

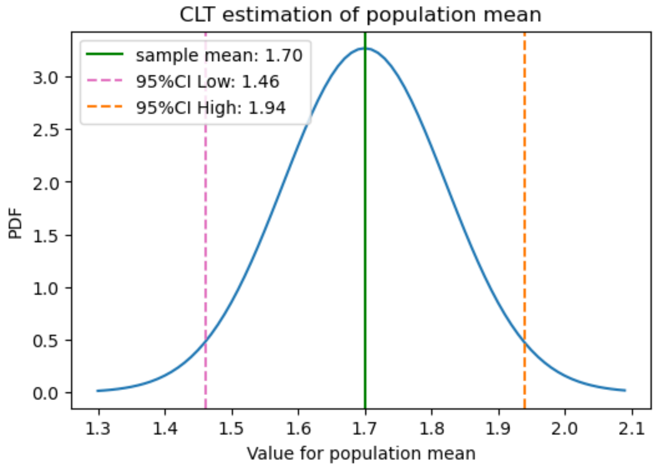
- Define confidence level α (e.g., 95%)
- Obtain quantile for $(1-\alpha)/2$
- Obtain quantile for $(1+\alpha)/2$

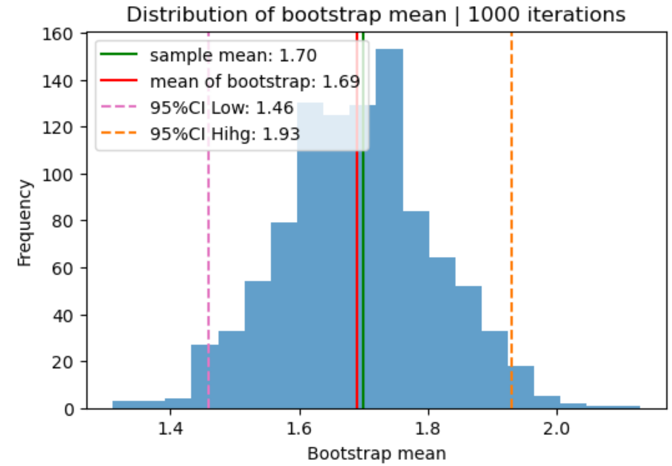
```
# Calculate the bootstrap confidence interval
In [1]: np.quantile(X_bs, q=0.025)
Out[1]: 1.46

In [1]: np.quantile(X_bs, q=0.975)
Out[2]: 1.93
```



Comparison | confidence intervals

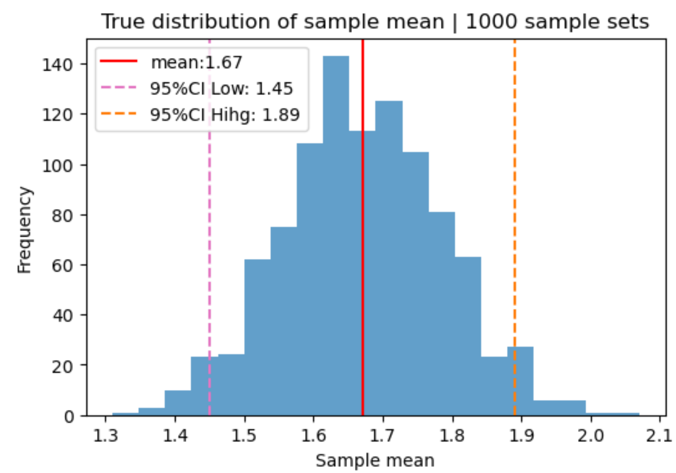
Normal distribution via central limit theorem





Bootstrap distribution

True distribution by repeating 1,000 sample sets (difficult in reality)



Summary

- We often use a sample set to estimate the statistics of a population
  - The Population mean, population standard deviation;
  - The Sample mean, sample standard deviation;
  - The Standard error of the mean.

- Point estimate vs confidence interval

- Methods for estimating confidence interval
  - Large sample size: normal distribution guaranteed by central limit theorem.
  - Small sample size from a population of normal distribution: t distribution (population variance is unknown)
  - Bootstrapping: a non-parametric way to approximate confidence interval

Python `scipy.stats` and `numpy.random`

- Distribution and useful functions:
  - `from scipy import stats`
  - Normal distribution:
    - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html
    - `stats.norm.pdf()`
    - `stats.norm.ppf()`
  - Student *t* distribution:
    - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html
    - `stats.t.pdf()`
    - `stats.t.ppf()`
- Generating random numbers following a certain distribution:
  - https://numpy.org/doc/stable/reference/random/generated/numpy.random.choice.html
  - https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html
  - https://numpy.org/doc/stable/reference/generated/numpy.var.html

Resources & Acknowledgement

- IPython Notebook for this lecture note: In Moodle

Other reference resources with acknowledgement:

- Chapter 2, Bruces & Gedeck, Practical Statistics for Data Science
- UPenn State course: Sampling Theory and Methods. Lessons 1 & 2: https://online.stat.psu.edu/stat506/
- Imperial College course: Introduction to Sampling & Hypothesis Testing (by Dr John Pinney) https://github.com/johnpinney/sampling_and_hypothesis_testing