上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

2022 春季

上海交通大学生物医学工程学院
SJTU School of Biomedical Engineering

生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

授课老师：林关宁

课程内容安排

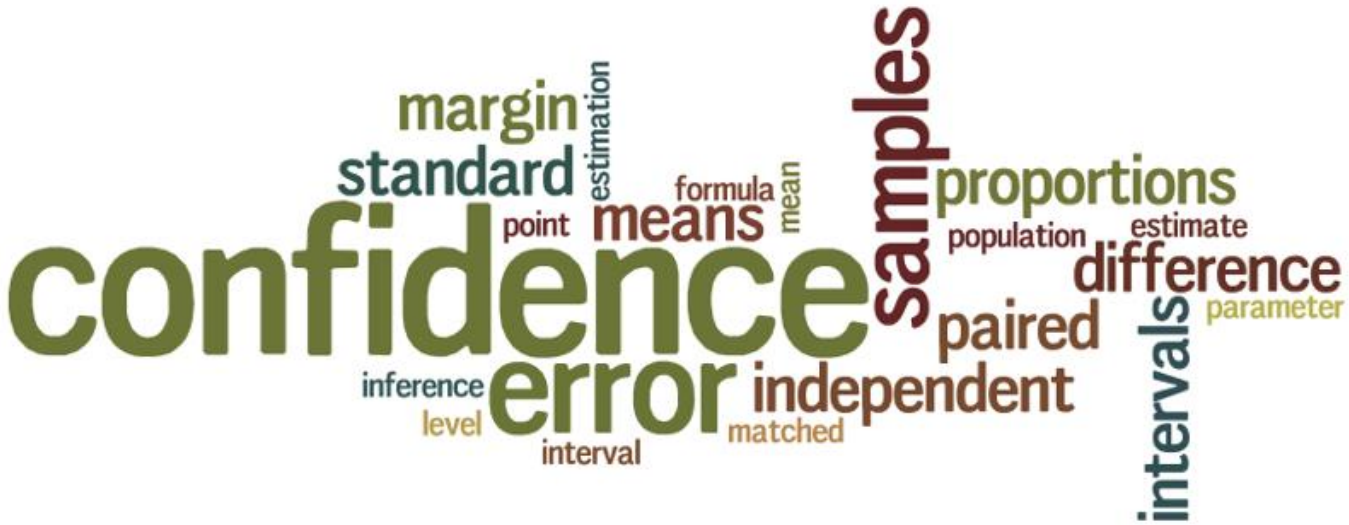
上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1（8%）	测试1（8%）	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 常用概率分布			2
3.9			4. 极限定理			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2（12%）	测试2（12%）	2
3.23			6. 统计推断基础-2:假设检验 Hypothesis Test			2
3.30			7. 数值数据的均值比较-1： 单样本t-检验			2
4.6			8. 数值数据的均值比较-2： 独立双样本t-检验，配对样本t-检验			2
4.13			9. 数值数据的均值比较-3：One-Way ANOVA			2
4.20			10. 数值数据的均值比较-4：Two-way ANOVA			2
4.27	3	比例差异检验	11. 类别数据的比例比较-1： 单样本比例推断 *	作业3（4%）	测试3（4%）	2
5.7（调）			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 （6%）	测试4 （6%）	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

* 随堂测试

单元2内容（week 4-9）

(对应参考书第4章内容)

- **总体、样本**
  - 参数、统计量
- **置信区间**
  - 区间估计、总体方差
- **均值差异**
  - One-sample 均值估计
  - Two-sample 均值差异
- **零假设检验**
  - H_0 vs H_a，Type I & II errors，P-value，One-sample vs Two sample test
- **方差分析**
  - ANOVA test（one-way ANOVA, two-way ANOVA）



什么是统计学?

- Statistics is the science of collection, analysis, interpretation, and presentation of data.

  统计学是一门收集、分析、解释和呈现数据的科学

- Descriptive statistics are numerical estimates that organize, sum up or present the data.

  描述性统计：是组织、总结或呈现一组数据的重要特征（表格，图形）

  (均值，中位数，众数，标准差)

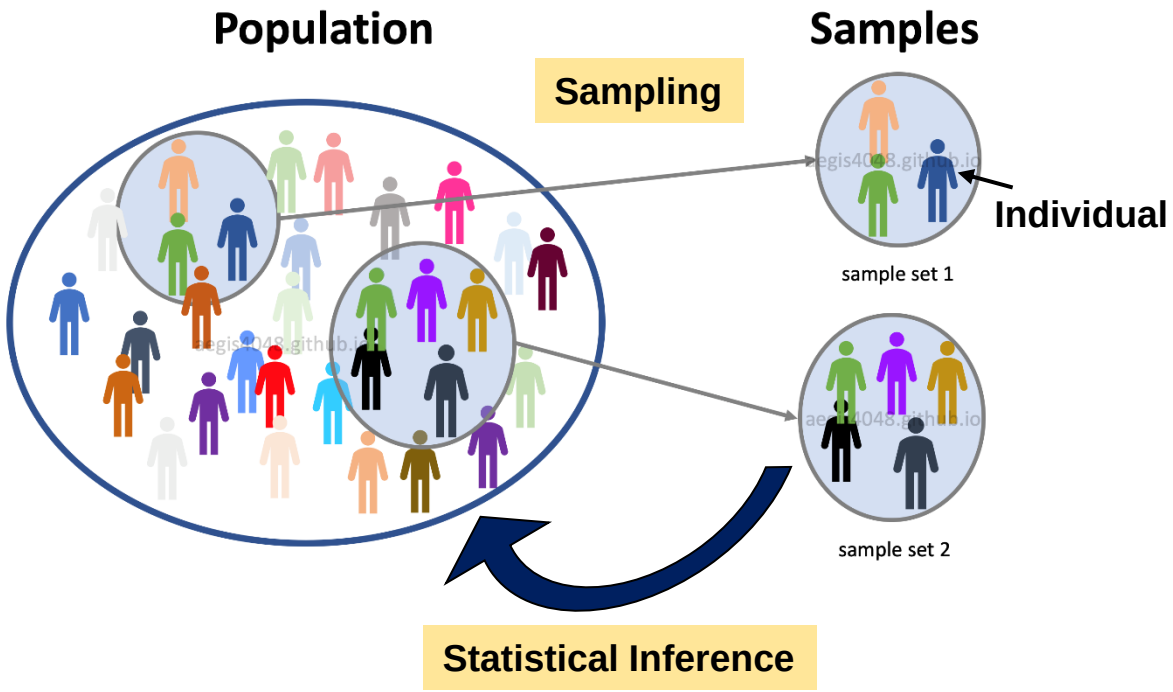
- Inferential statistics is the process of inferring from a sample to the population.

  推论统计： 利用样本信息对总体进行估计，预测或推断的一个过程（假设检验）

  (抽样，统计显著性)

总体和样本

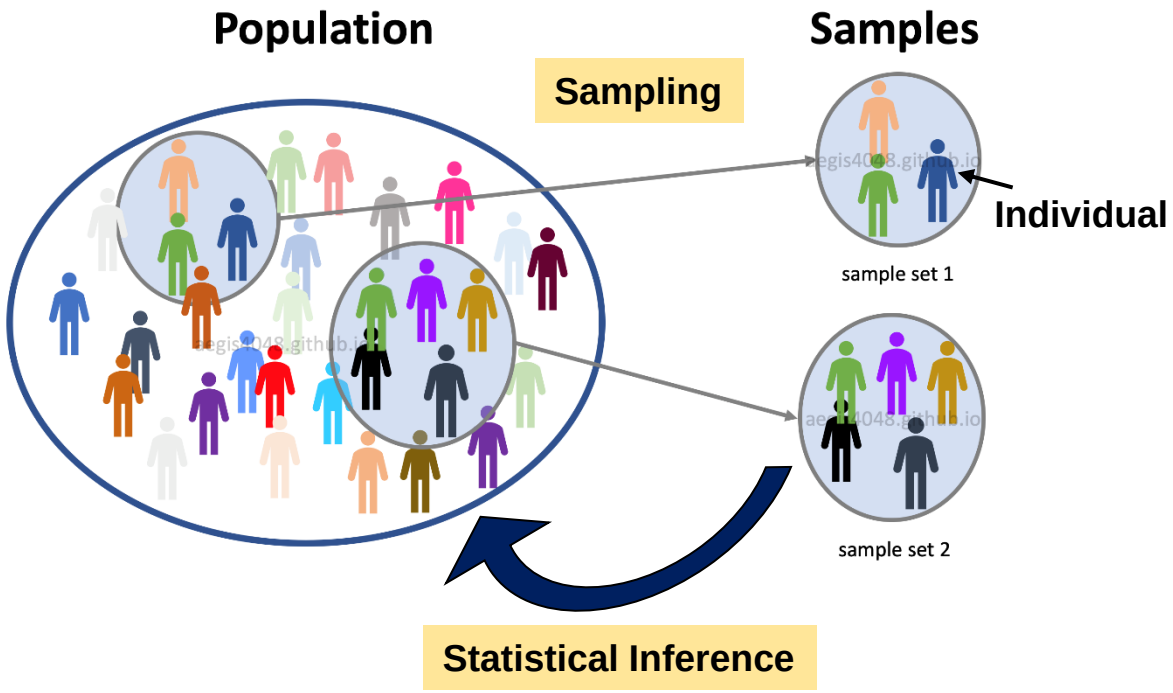
总体：包含指定组的所有成员的数据集。例如：所有居住在中国的人。
样本：包含人口一部分或一部分的数据集。例如：居住在中国的某些人。



我们使用样本统计量进行统计推断的方法称为估算/估计（estimation）

总体和样本

总体：包含指定组的所有成员的数据集。例如：所有居住在中国的人。
样本：包含人口一部分或一部分的数据集。例如：居住在中国的某些人。



我们通过 使用 样本统计量 (statistic) 估计 总体参数 (parameter) 来进行 统计推断

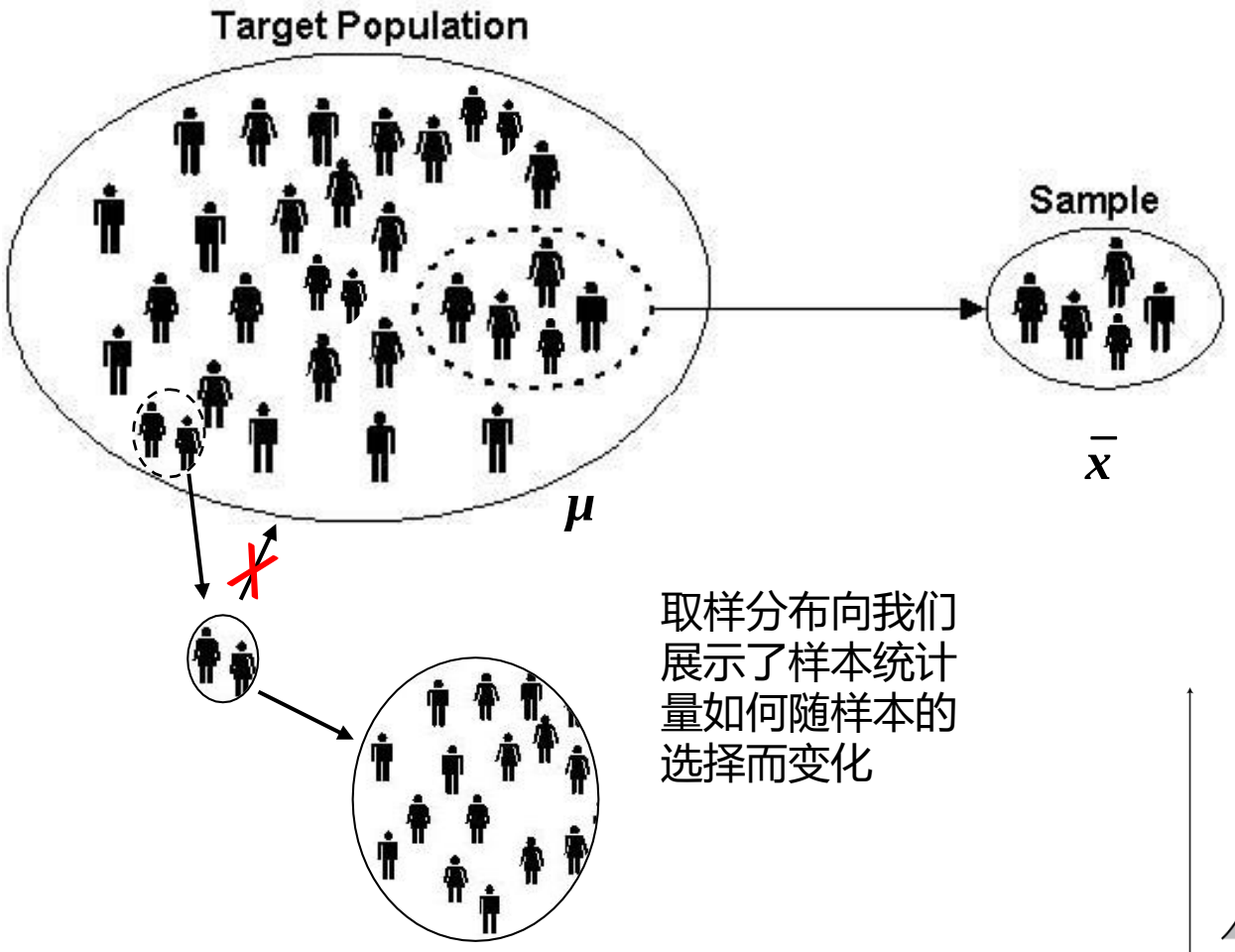
Population versus Sample

	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	

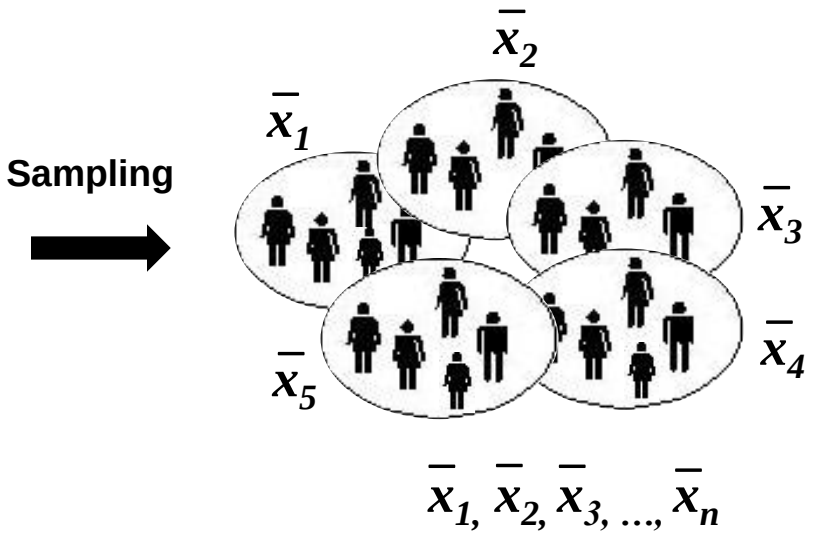
我们使用样本统计量进行统计推断的方法称为估算/估计（estimation）

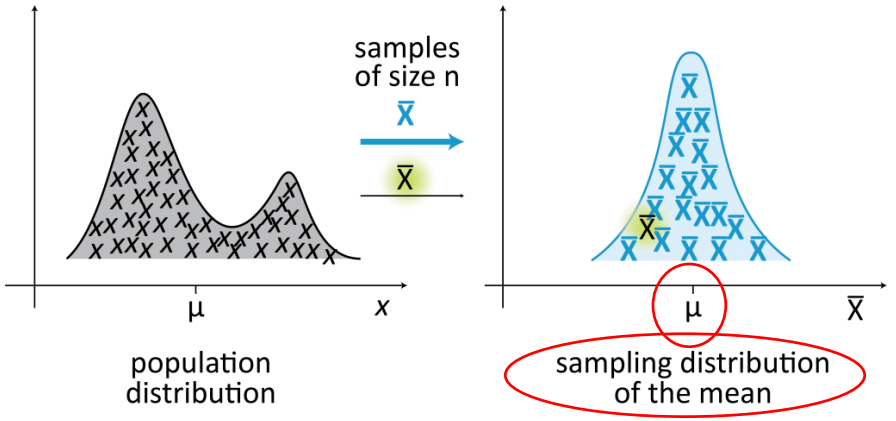
取样分布（Sampling distribution）

	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	

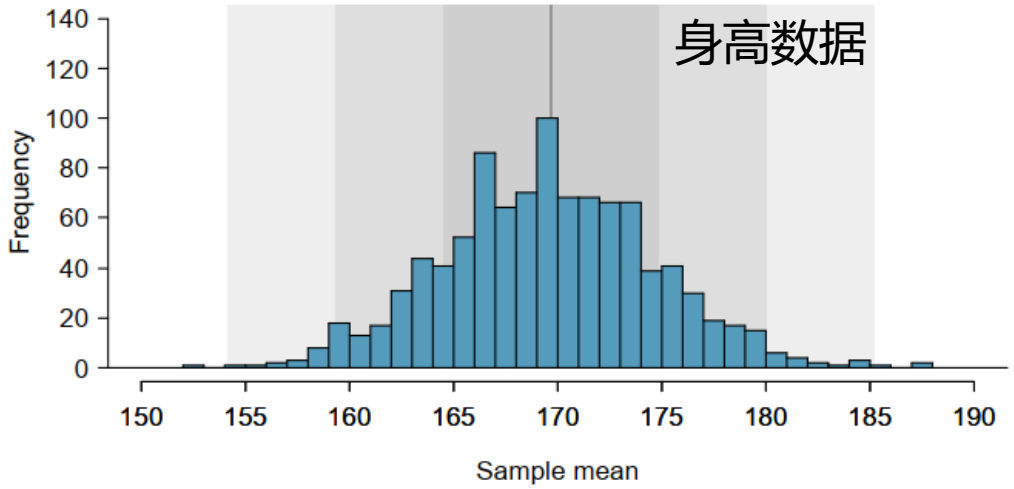


取样分布向我们展示了样本统计量如何随样本的选择而变化

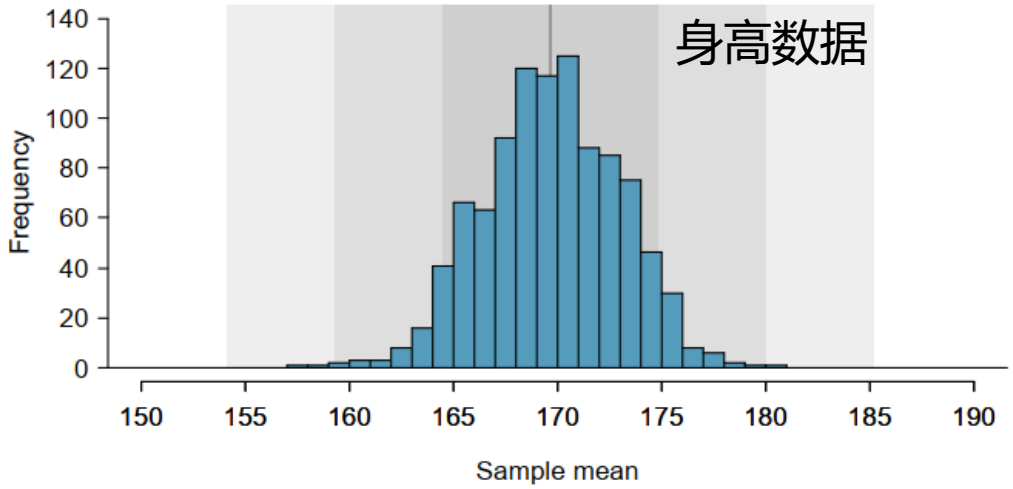




不管总体的分布是怎样的，二项分布，伯努利分布，指数分布，还是什么。均值的取样分布将近似于一个正态分布
(CLT)



(a)



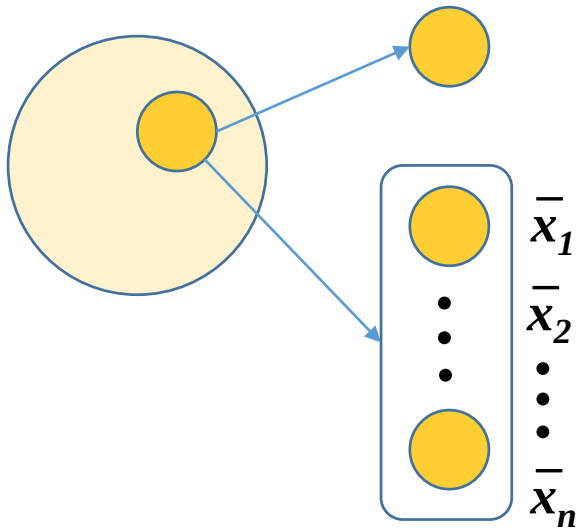
(b)

图（a）显示了来自美国cdc的1000组大小为60的随机样本的样本均值直方图。对于大小为60的样本，直方图提供了X的理论采样分布的近似值。

图（b）显示了来自美国cdc的1000组大小为200的随机样本的样本均值直方图。对于大小为200的样本，直方图提供了X的理论采样分布的近似值。

样本量越大，抽样分布的可变性就越小。如图所示，增加样本量会导致 x 的分布更紧密地聚集在总体平均值 μ 周围，从而可以从单个样本中更准确地估计 μ 。当样本量较大时，任何特定样本的平均值更可能接近总体平均值。

标准误差（standard error）



抽取一个样本，来推断总体。这时可以依据抽取的样本信息，计算出样本的均值与标准差

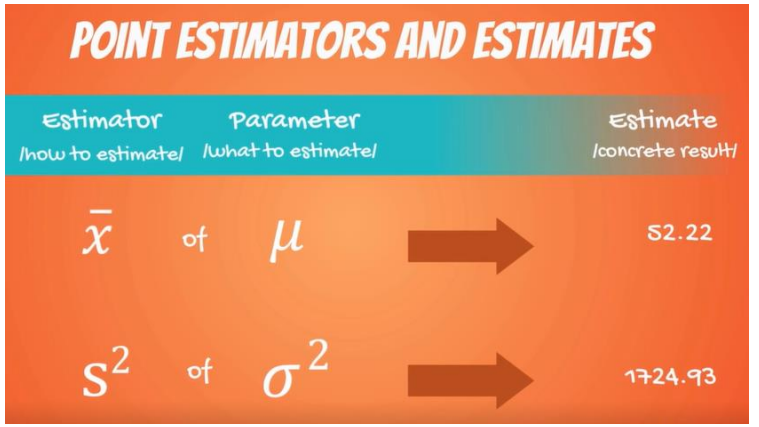
n次取样，n个均值 -> 计算这10个均值的标准差，此时的标准差就是**标准误差**

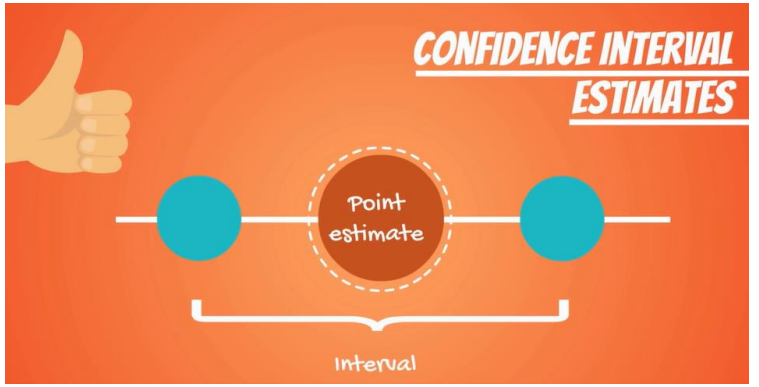
统计学中很常用的，来显示取样的均值离理论均值有多远

$$SE = \frac{\sigma}{\sqrt{n}}$$

估计（Estimate）

- **估计量（Estimator），估计（Estimate）**
  - estimator是取决于样本的一个估计变量
  - estimate是根据某个确定的样本值而得到的一个确切地估计值（不是变量）
  - estimator是一个法则，estimate指用这个法则而算出来的具体值
- **点估计（Point estimates）**
  - 平均值：比如抽样鸡腿的平均重量为150克
- **区间估计（Confidence interval estimates）**
  - 比例：比如抽样鸡腿的卫生合格率为99.9%





通俗地讲：区间估计是在点估计的基础上，给一个合理取值范围

比如：抽样鸡腿的平均重量为150克，是一个点估计值。抽样鸡腿的平均重量为145克到155克之间，是一个区间估计。
其中，145到155称为置信区间。这很符合人们的常规理解：东西很难100%准确，有个范围也是可以理解的。

但这个范围有多大可信度呢？人们用置信水平来衡量，即：“我们有多大把握，总体参数的真实值在某个置信区间内”

置信区间

- A plausible range of values for the population parameter is called a *confidence interval*. **总体参数的合理值范围称为置信区间**
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

如果仅使用一个样本的统计数据来估计参数就像用矛在浑浊的湖中捕鱼，而**使用置信区间就像用网**捕鱼

We can throw **a spear** where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

如果我们报告一个点估计，我们可能不会获得确切的总体参数。
如果我们报告一个范围内所有合理的值，我们就有机会捕捉到参数。

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny

如何理解95%置信区间?

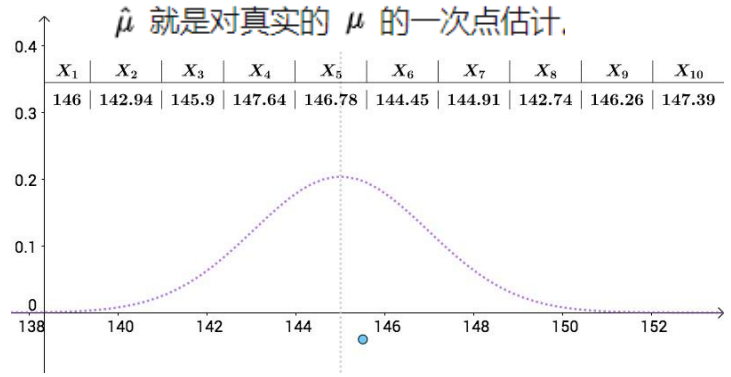
如何理解95%置信区间?

95%置信区间（Confidence Interval, CI）：

当给出某个估计值的95%置信区间为【a,b】时，可以理解为我们有95%的信心（Confidence）可以说样本的平均值介于a到b之间，而发生错误的概率为5%

区间估计

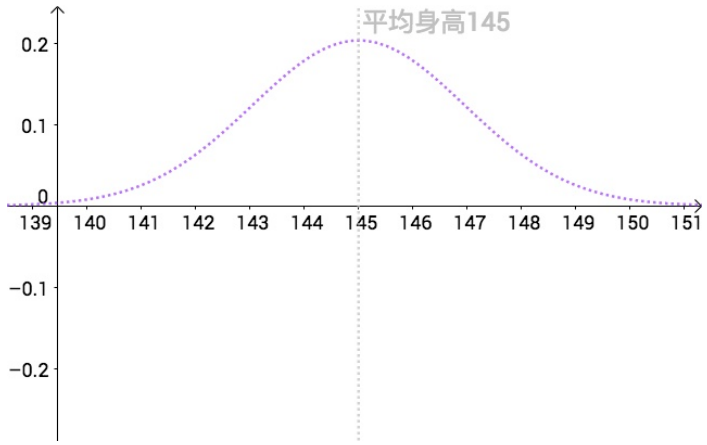
- 在一定概率保证下指出总体参数的可能范围，叫区间估计
- 所给出的可能范围叫**置信区间**（confidence interval）
- 给出的概率保证称为**置信度或置信概率**（confidence probability）



例子：人类的平均身高

假设人类的身高分布服从（$\mu = 145, \sigma = 1.4$）如下正态分布

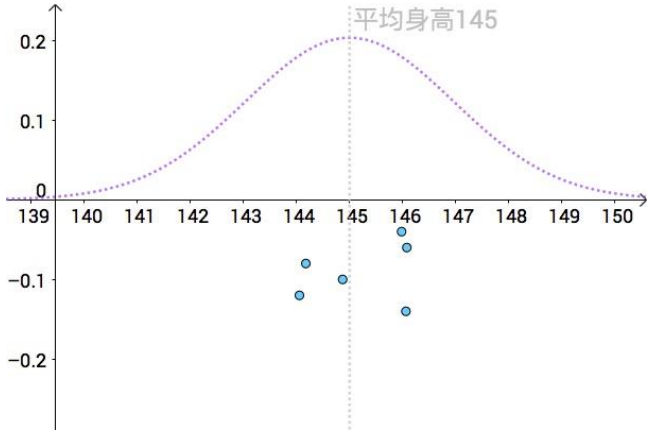
$$X \sim N(145, 1.4^2)$$

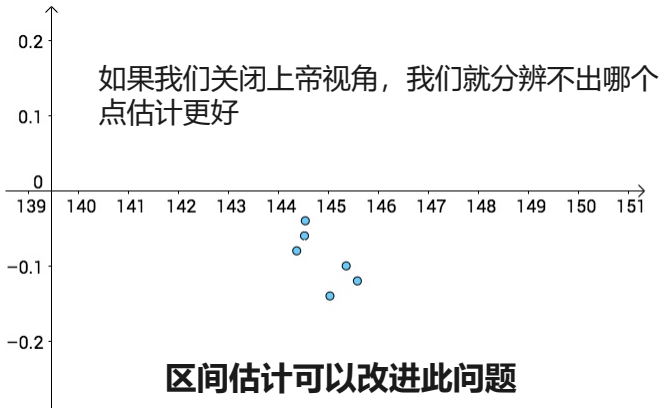


”上帝视角“

作为人类，我们只能在人群中抽样统计

通过一次次的抽样，我们可以算出不同的身高均值的点估计：





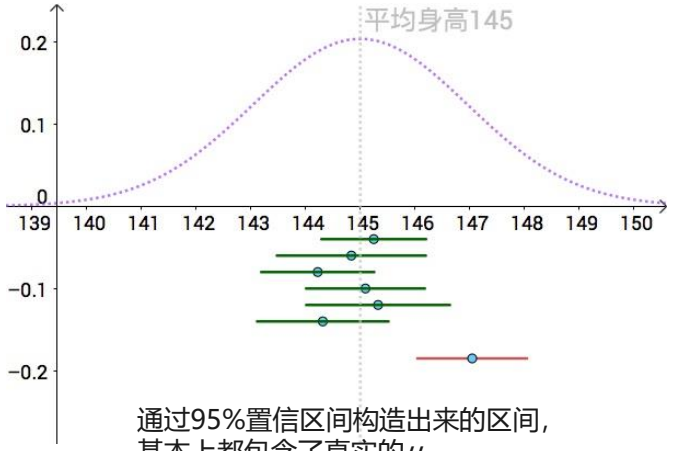
如果我们关闭上帝视角，我们就分辨不出哪个点估计更好

区间估计可以改进此问题

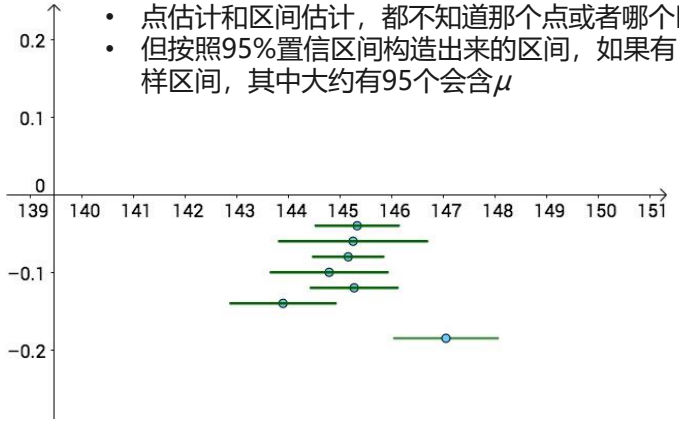
置信区间

置信区间，提供了一种区间估计的方法

下面采用95%置信区间来构造区间估计



通过95%置信区间构造出来的区间，
基本上都包含了真实的μ



- 点估计和区间估计，都不知道那个点或者哪个区间更好
- 但按照95%置信区间构造出来的区间，如果有100个这样区间，其中大约有95个会含μ

95% 置信区间

以上面的统计身高为例，假设全国人民的身高服从正态分布：

$$X \sim N(\mu, \sigma^2)$$

不断进行采样，假设样本的大小为n，则样本的均值为：

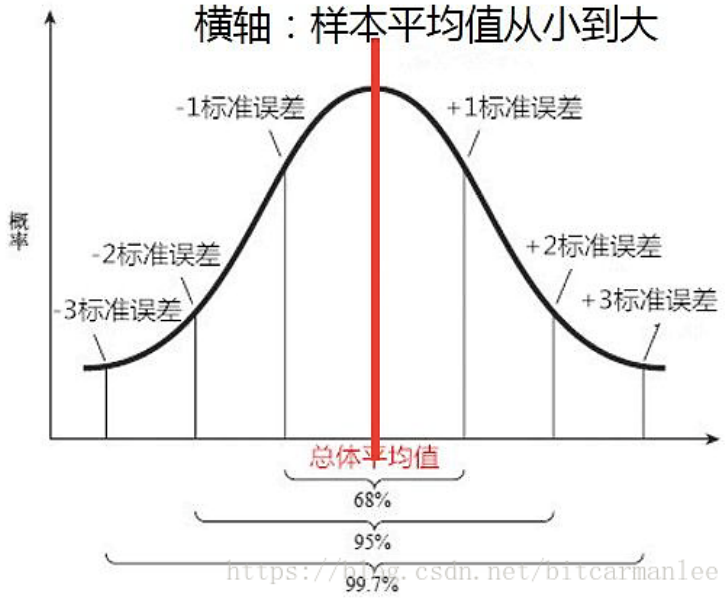
$$M = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

由大数定理与中心极限定理：

$$M \sim N(\mu, \sigma_1^2)$$

注意σ_1的计算方法为 ⟹ 标准误差!

为什么常用95%的置信水平：



用一句简单的话概括就是：
有95%的样本均值会落在2个（比较精确的值是1.96）标准误差范围内

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < M < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

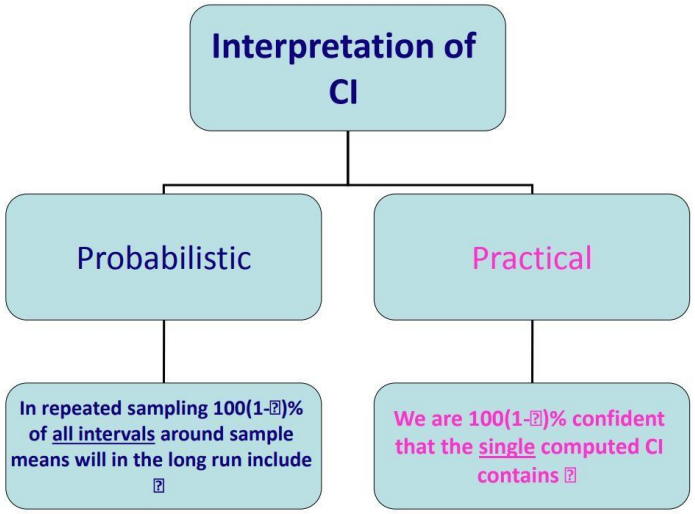
那么，只有一个问题了，我们不知道、并且永远都不会知道真实的 μ 是多少

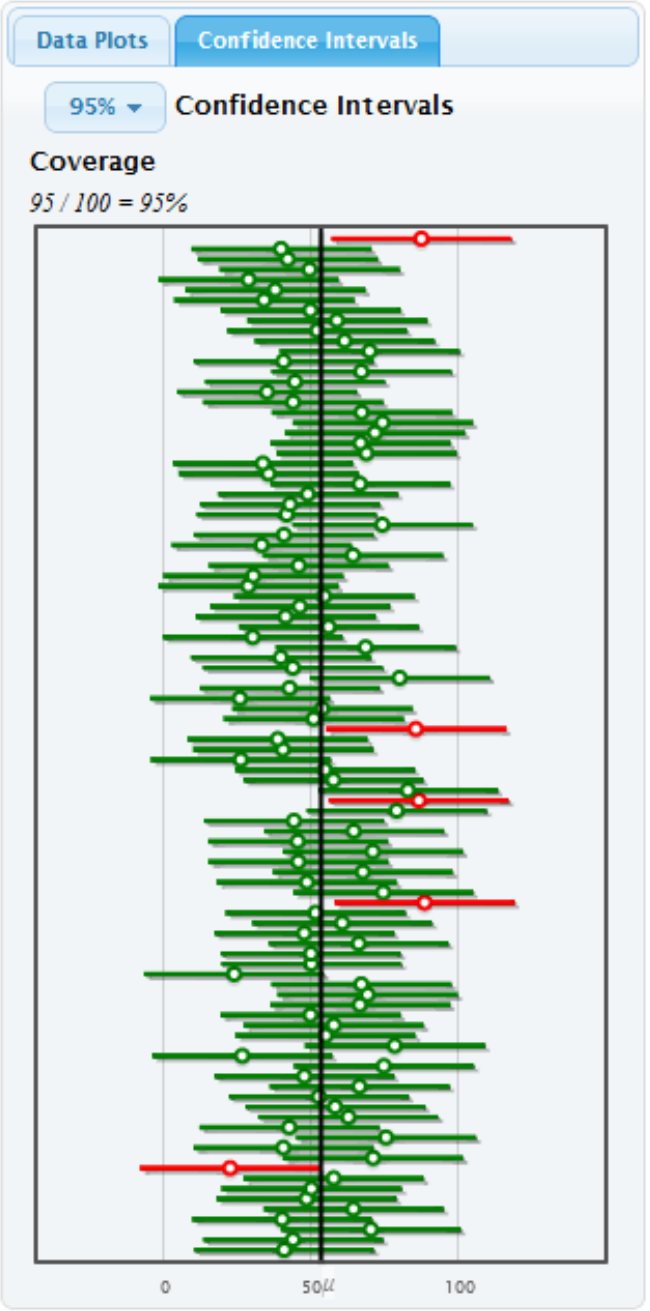
我们就只有用 $\hat{\mu}$ 来代替 μ ：

$$P(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \le M \le \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

置信区间

- **It is not the probability of samples falling into this interval!!!**
- 95%的置信区间：100次抽样，有95次的置信区间包含了总体均值





- 总体参数是固定的
- 样本统计量是 random（根据抽取的样本）
- 区间也是 random （根据样本统计量）
- 用中括号 [a,b] 表示样本估计总体平均值误差范围的区间。a、b的具体数值取决于你对于“该区间包含总体均值”这一结果的可信程度，因此 [a,b] 被称为置信区间

正态分布的置信区间 Confidence interval of normal distribution

- **统计的置信区间的计算取决于两个因素：**
  - **统计类型 Types of statistics**
    - **均值 mean**
      - 均值的置信区间 **interval of mean**
      - 均值差的置信区间 **confidence interval of difference in means**
    - **方差 variance**
    - 标准差 Standard deviation
  - **样本分布类型 Type of sample distributions**
    - **normal**

先来看下单样本的统计估计 (one-sample inference)

- How well the sample parameter(s) can represent the population 样本在多大程度上能代表总体?

点估计 • Point estimation

自信区间估计 • Confidence interval estimation

假设检验 • Hypothesis testing

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

s

center

Shape spread

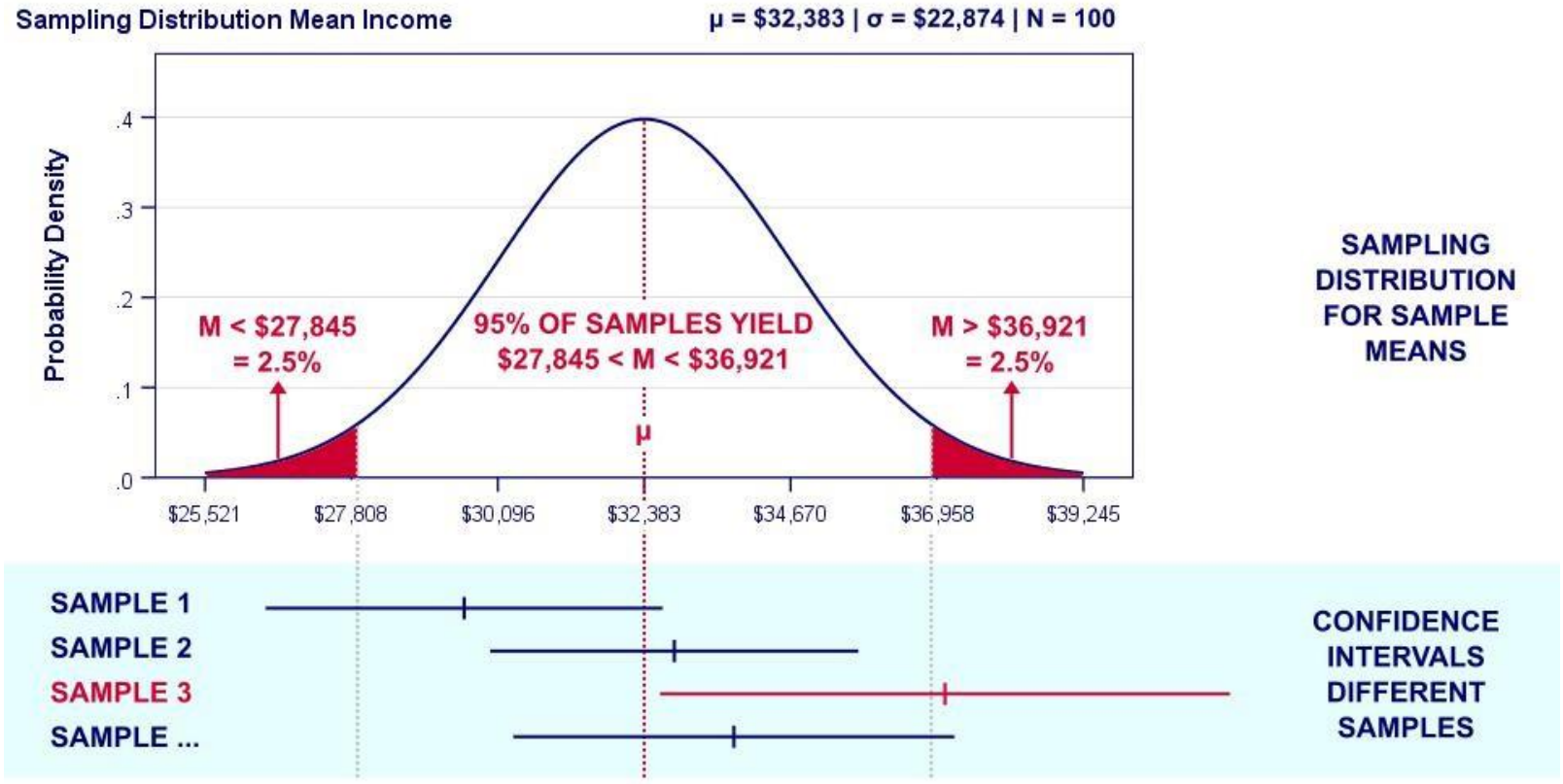
Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

Basic statistics workshop organised by The

单样本均值置信区间 （one-sample CI）的使用

- 你有一个已知的样本，想知道这个样本反映的**总体均值的置信区间是多少**；
- 你有一个已知的样本，并且也知道了总体的均值，你想知道你的**样本的均值是否不同于这个总体的均值**



* 计算置信区间的一般套路

1.首先明确要求解的问题。比如我们的例子，就是想通过样本来估计全国人民身高的平均值。
2.求抽样样本的平均值与标准误差(standard error)。注意标准误差与标准差(standard deviation)不一样。
3.确定需要的置信水平。比如常用的95%的置信水平，这样可以保证样本的均值会落在总体平均值2个标准差得范围内。
4.查z表，求z值。
5.计算置信区间
a = 样本均值 - z*标准误差
b = 样本均值 + z*标准误差

用公式表示置信区间：

$$\overline{x} \pm z\frac{s}{\sqrt{n}}$$

其中，$\overline{x}$表示样本的均值，z值表示有多少标准差，s为样本的方差。

Table - Z-Scores for Commonly Used Confidence Intervals

Desired Confidence Interval	Z Score
90%	1.645
95%	1.96
99%	2.576

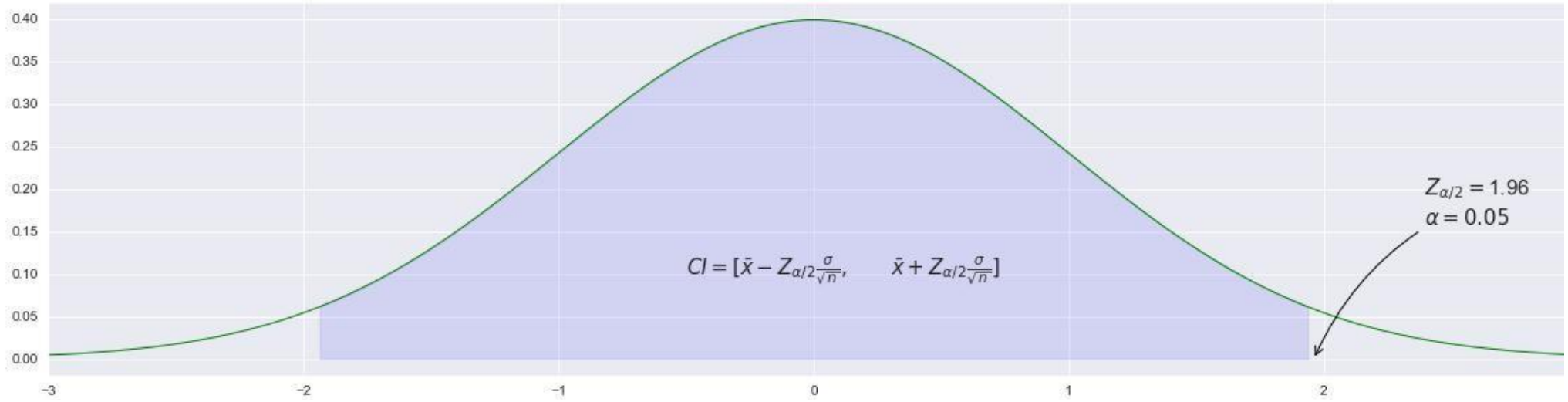
单样本的CI估算总体均值μ（σ 标准差已知，用 z-dist)

使用条件:

- The data are continuous (not discrete).
- The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.
- The data follow the normal probability distribution. (Dependent on the sample size)
- The population standard deviation σ is known.

方法： Z-Distribution，样本的 $(100-\alpha)\%$ 置信区间定义，其中$Z_{\alpha/2}$是对应的Z-score.

$$CI = \bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



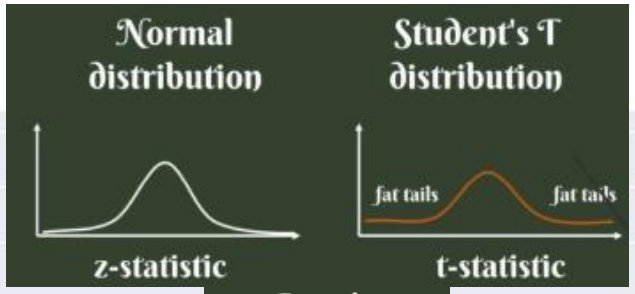
单样本的CI估算总体均值μ（σ 未知，用 t-dist)

William Gosset (威廉·戈塞, 1876-1937, 英国统计学家/化学家, 发明学生t-检验)

使用条件:

- The data are continuous (not discrete).
- The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.
- The data follow the normal probability distribution (Dependent on the sample size)
- The population standard deviation σ is unknown.

方法： t-Distribution，样本的 $(100-\alpha)\%$ 置信区间定义, 其中$t_{\frac{\alpha}{2},df}$ 可以查表或者python计算



$$CI = \bar{x} \pm t_{\frac{\alpha}{2},df}\frac{s}{\sqrt{n}}$$

- **T-统计很实用，比如利用小样本去推断统计；还有如果总体的方差未知，等等。**
- **T 比 Z 的分布要更散一点，就是有更宽的尾巴（这样就可以允许更大的方差值和未知）**

* Python Examples

```
import pandas as pd
import numpy as np
from scipy import stats

#读取数据
df = pd.read_excel("Data_normtemp.xlsx")
df.head()

alpha = 0.05                          # significance level = 5%
data = df.体温
size = len(data)

# 一个正态总体方差已知，均值的区间估计，使用的是正态分布:
conf_intveral_z = stats.norm.interval(1-alpha, loc=np.mean(data), scale=stats.sem(data))
print(conf_intveral_z)

############################
print("############")

# 一个正态总体，方差未知，均值的区间估计，使用的是 t 分布:
t = stats.t.ppf(1 - alpha/2, size-1)    # t-critical value for 95% CI = 2.093
s = np.std(data, ddof=1)                # sample standard deviation = 2.502
lower = np.mean(data) - (t * s / np.sqrt(size))
upper = np.mean(data) + (t * s / np.sqrt(size))
print(lower, upper)

conf_intveral_t = stats.t.interval(1-alpha, size-1 , loc=np.mean(data), scale=stats.sem(data))
print(conf_intveral_t)
```

```
(98.12319642818164, 98.37526511027988)
############
98.12200290560803 98.3764586328535
(98.12200290560803, 98.3764586328535)
```

现在来看下多样本的统计估计

正态分布的置信区间 Confidence interval of normal distribution

- 统计的置信区间的计算取决于两个因素：
  - 统计类型 Types of statistics
    - 均值 mean
      - 均值的置信区间 interval of mean
      - 均值差的置信区间 confidence interval of difference in means
    - 方差 variance
    - 标准差 Standard deviation
  - 样本分布类型 Type of sample distributions
    - normal

均值差的置信区间 confidence interval of difference in means

- **Dependent samples（配对样本）**
  - Before – after
  - Cause - effect
- **Independent samples（独立样本）**
  - Known population variances　　总体**方差已知**
  - Unknown population variances, but assumed to be equal　　总体**方差未知**，相等
  - Unknown population variances, but assumed to be different　　总体**方差未知**，不等

(配对样本)

Confidence interval for difference of two means (dependent/paired samples)

假设你在研制可以增加镁在血液中浓度的药，以下是患者用药前后镁的浓度（mg/dl）

Patient	1	2	3	4	5	6	7	8	9	10
Before	2.00	1.40	1.30	1.10	1.80	1.60	1.50	0.70	0.90	1.50
After	1.70	1.70	1.80	1.30	1.70	1.50	1.60	1.70	1.70	2.40
Difference	-0.30	0.30	0.50	0.20	-0.10	-0.10	0.10	1.00	0.80	0.90

问题： 这个两组镁的浓度差异的95%CI是多少

方法：t-Distribution

$$CI = \bar{x} \pm t_{\frac{\alpha}{2},df} \frac{s}{\sqrt{n}}$$

Dependent samples：

- **Before vs. after**
- **Cause vs. effect**

$$\bar{x} = \sum x_i / n, \quad \alpha = 0.05,$$
$$n = 12$$
$$df = n - 1 = 11$$
$$t_{\frac{\alpha}{2},df} = stats.t.isf(0.025, 11)$$

(独立样本1)　整体方差（σ_1^2，σ_2^2）已知：Z-distribution

假设你在调查一所美国大学的入学考试成绩，以下是俩个学院学生的成绩的统计表

	Engineering	Management	Difference
Student #	100	70	?
Grade mean	58	65	-7.00
Population grade std	10	5	1.16

问题：　这两学院学生的入学考试成绩差异的95%CI是多少?

Considerations:

- Samples are big (normally distributed)
- Population variances are known
- The sample sizes are different
- Populations are assumed to follow the Normal distribution

如果两个样本的大小分别 n_1, n_2,那么样本均值 $\bar{x}_1, \bar{x}_2$的分布分别是 $N(\mu_1, \sigma_1/\sqrt{n_1})$, $N(\mu_2, \sigma_2/\sqrt{n_2})$,
因此 $\bar{x}_1$-$\bar{x}_2$的方差为： $\sigma^2=\left(\sigma_1/\sqrt{n_1}\right)^2+\left(\sigma_2/\sqrt{n_2}\right)^2$,

所以$\bar{x}_1-\bar{x}_2$ 的标准差（标准误）为：$\text{sem}=\sigma=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$,

Variance of the difference

$$\sigma_{diff}^2=\frac{\sigma_e^2}{n_e}+\frac{\sigma_m^2}{n_m}$$

$$\sigma_{diff}^2=\frac{10^2}{100}+\frac{5^2}{70}=1.36$$

那么μ_1- μ_2的1- α置信区间是:

$$(\bar{x}-\bar{y})\pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

difference point estimator　test statistic　standard error

=(-9.28,-4.72)
95% confidence interval

(独立样本2) 整体方差σ_1^2，σ_2^2未知，但$\sigma_1 = \sigma_2$： t-分布

Student's t-interval
等方差 t-区间

问题： 这两城市的苹果价格差异的95%CI是多少

苹果价钱

北京	上海
$ 3.80	$ 3.02
$ 3.76	$ 3.22
$ 3.87	$ 3.24
$ 3.99	$ 3.02
$ 4.02	$ 3.06
$ 4.25	$ 3.15
$ 4.13	$ 3.81
$ 3.98	$ 3.44
$ 3.99	
$ 3.62	

	北京	上海
Mean	$3.94	$3.25
Std. deviation	$0.18	$0.27
Sample size	10	8

如果两个样本的大小分别n_1, n_2，那么样本均值$\bar{x}_1, \bar{x}_2$的分布分别是$N(\mu_1, \sigma_1/\sqrt{n_1})$，$N(\mu_2, \sigma_2/\sqrt{n_2})$，如果$\sigma_1 = \sigma_2$， 那么样本1和2的平均/$pooled$ 样本标准差为：

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$。 因此， $\bar{x}_1 - \bar{x}_2$的标准误差为$\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

因此，总体μ_1- μ_2的1- α置信区间是依据t-分布：

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2},(n_1+n_2-2)s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= (3.94 - 3.25) \pm 2.12\sqrt{\frac{0.05}{10}+\frac{0.05}{8}}$$

$$CI_{95\%} = (0.47, 0.92)$$

当你有理由相信俩个样本的方差几乎相等时，你可以使用student t-检验来检验均值的差异是否显著不同。要注意的是，只要样本大小相等或接近相等，并且样本量不小，学生t检验即使在方差不相等的情况下效果也ok的。但如果你绝对确定总体方差几乎相等，就用学生t检验。

(独立样本3) 整体方差σ_1^2，σ_2^2未知，且$\sigma_1 \neq \sigma_2$ t-分布

Welch's t-interval
不等方差 t-区间

问题：这个两城市的苹果价格差异的95%CI是多少

苹果价钱

北京	上海
$ 3.80	$ 3.02
$ 3.76	$ 3.22
$ 3.87	$ 3.24
$ 3.99	$ 3.02
$ 4.02	$ 3.06
$ 4.25	$ 3.15
$ 4.13	$ 3.81
$ 3.98	$ 3.44
$ 3.99	
$ 3.62	

	北京	上海
Mean	$3.94	$3.25
Std. deviation	$0.18	$0.27
Sample size	10	8

如果$\sigma_1 \neq \sigma_2$ 那么总体μ_1- μ_2的1- α置信区间为:

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2},v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1-1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2-1)}\left(\frac{s_2^2}{n_2}\right)^2}$$

(degree of freedom)

Note:
- 在比较俩个正态分布的中心趋势时，建议始终使用韦尔奇 *t*-检验，它的假设就是样本的方差不相等。
- 当总体方差不同时，等方差的student *t*-检验不具有鲁棒性，但如果总体方差相等，不等方差的welch *t*-检验确也能去检验等方差。

Formulas for Confidence Intervals

# populations	Population variance	Samples	Statistic	Variance	Formula
One	known	-	z	σ^2	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
One	unknown	-	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$
Two	-	dependent	t	$s^2_{difference}$	$\bar{d} \pm t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}$
Two	Known	independent	z	σ_x^2 , σ_y^2	$(\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$	$(\bar{x}-\bar{y}) \pm t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2 , s_y^2	$(\bar{x}-\bar{y}) \pm t_{\upsilon,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$

Key takeaways

- **置信区间量化了统计估计的不确定性**
  - 置信区间量化了与统计估计相关的不确定性，以缓解总体与样本的问题
- **置信区间是参数假设检验的基础**
  - 例如，t-检验使用均值差的置信区间来计算其p值
- **参数化的方法在非正态性条件下缺乏精度度**
- **95%的置信区间并不意味着95%的样本数据在区间内**
- **尽可能的使用t-score而不是z-score**
  - 当样本量较小时，t-分布比正态分布能容纳更大的不确定性；当样本量大于30时，t-分布也能收敛并趋向于正态分布
- **样本量越大，置信区间越窄**
  - 样本越多，不确定性越小
- **均值并不总等同于中心倾向**
  - 当样本不是正态分布时，它们的均值并不能很好地度量它们的中心趋势。例如，如果你用t-检验比较两个非正态数据集的平均值来得出它们是否来自同一人群的结论，你的方法就是错误的

谢谢，下周见！