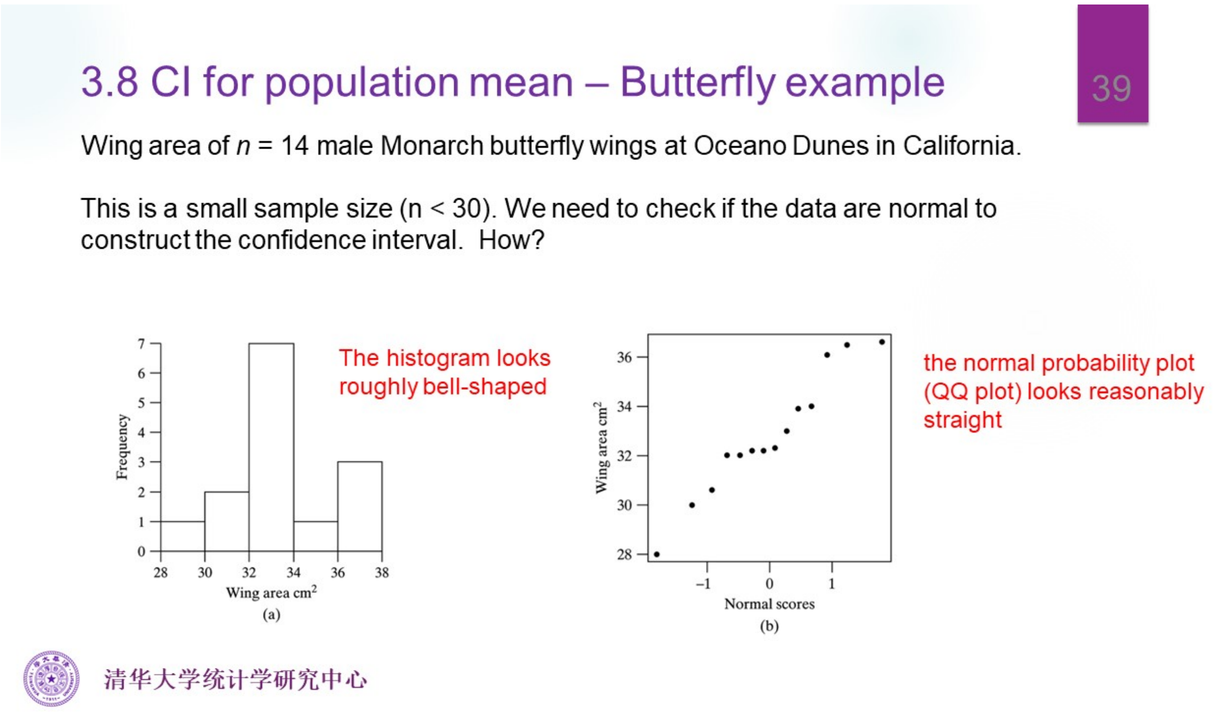
CI

总体方差未知的时候只能得到总体分布是t分布
n足够大→t分布近似正态分布

无论原本是什么分布如果感兴趣的是mean
当n足够大（*>30*）的时候根据CLT，**sample mean的分布**总是近似为正态分布√
样本量够大所以总体也是正态分布×
如果是点估计，用X-bar估计μ并匹配se，需要知道x-bar服从什么分布才能计算se
然后就可以求出ci
一定要注意样本量是否够大

样本量不够大： 判断是否正态

- 画直方图
- 画 **qq plot** （更方便看）the normal probility plot(qq plot) looks reasonably straight



3.8 CI for population mean – Butterfly example

39

Wing area of *n* = 14 male Monarch butterfly wings at Oceano Dunes in California.

This is a small sample size (n < 30). We need to check if the data are normal to construct the confidence interval. How?

The histogram looks roughly bell-shaped

the normal probability plot (QQ plot) looks reasonably straight

清华大学统计学研究中心

如果曲线呈现弧度（下凸、上凸ry）就不是正态
中间可能有离群值

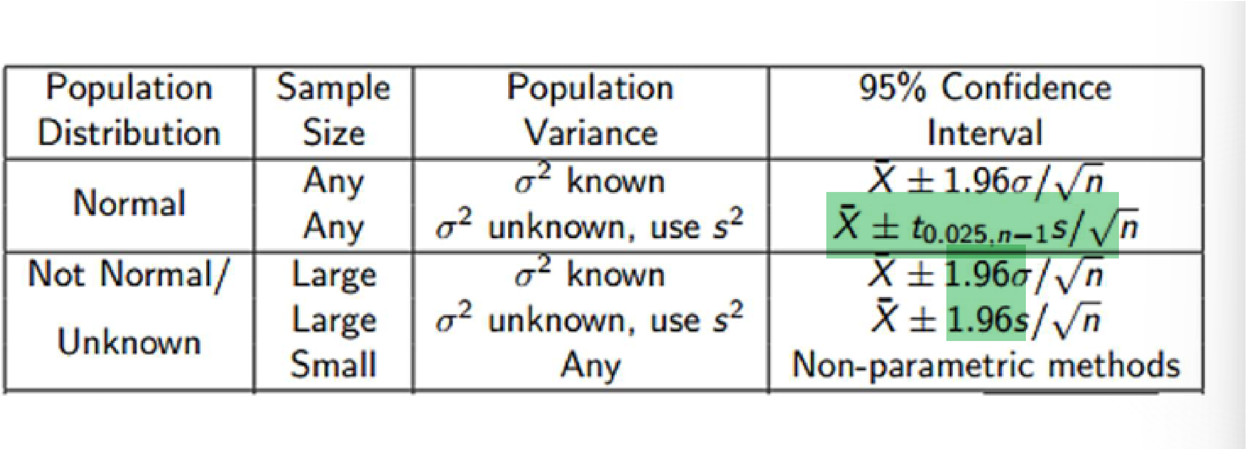
```
x=c(1,2,123,54356,234, ... )
qqnorm(x)
```

此时：
是正态分布→t检验

```
t.test(x)
```

不是→不能用

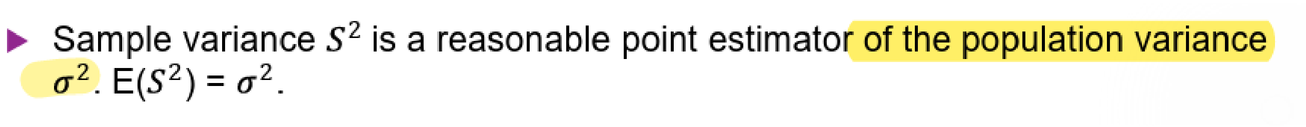
Summary



Population Distribution	Sample Size	Population Variance	95% Confidence Interval
Normal	Any	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Any	σ^2 unknown, use s^2	$\bar{X} \pm t_{0.025,n-1}s/\sqrt{n}$
Not Normal/ Unknown	Large	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Large	σ^2 unknown, use s^2	$\bar{X} \pm 1.96s/\sqrt{n}$
	Small	Any	Non-parametric methods

t分布的分位数/ 已经采用CLT，正态分布的分位数

σ^2的点估计



- Sample variance S^2 is a reasonable point estimator of the population variance σ^2. E(S^2) = σ^2.

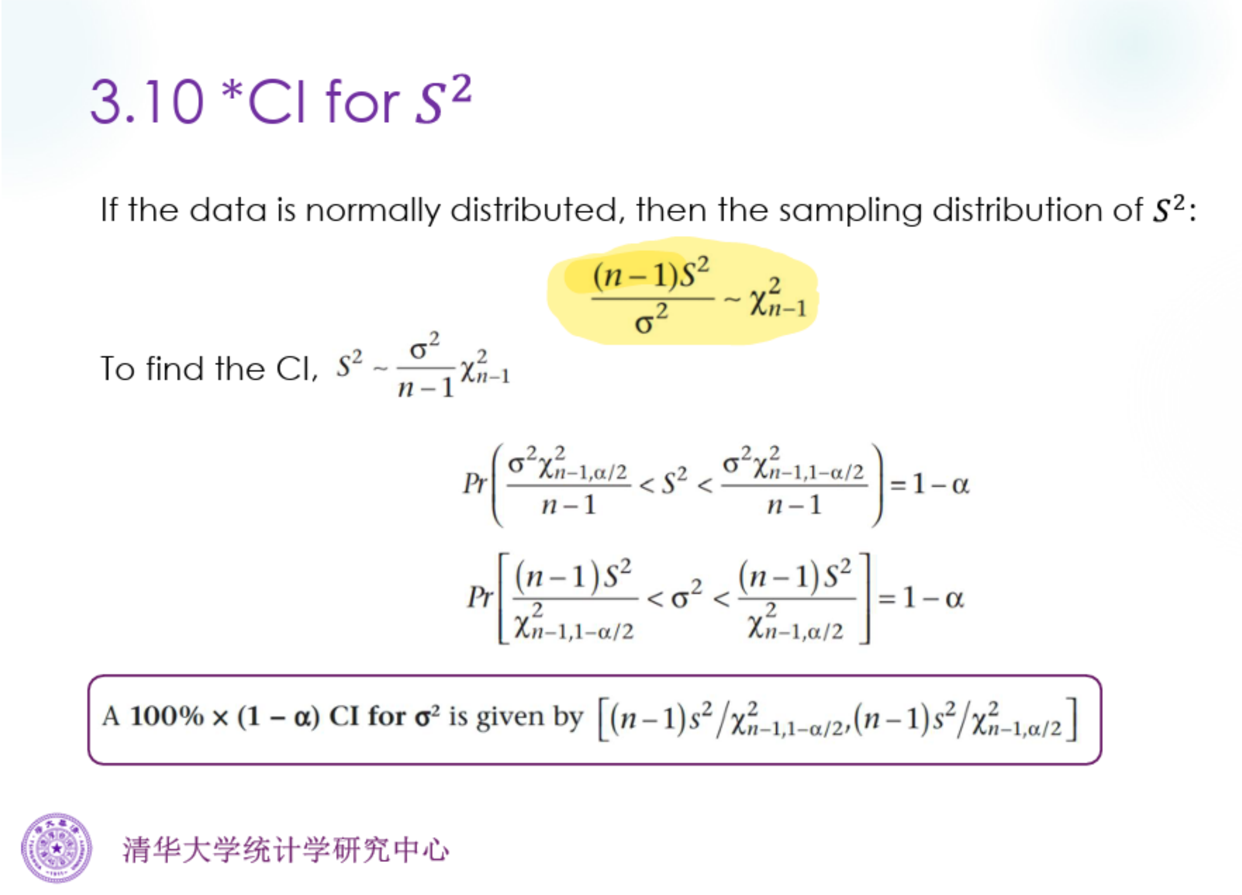
S^2是一个合适的估计 **因为** 它无偏

$S^2=\Sigma(xi-xbar)^2/(n-1)$

$\sigma^2=\Sigma(xi-xbar)^2/n$

σ-hat^2=S^2 无偏估计，但如何判断估计的准确性?

为了求出SE(S^2)，需要知道sampling dist：Chi-square（来源于标准正态分布的 **平方**，自由度是决定它形状的唯一参数）



3.10 *CI for S^2

If the data is normally distributed, then the sampling distribution of S^2:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

To find the CI, $S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$

$$Pr\left(\frac{\sigma^2\chi^2_{n-1,\alpha/2}}{n-1} < S^2 < \frac{\sigma^2\chi^2_{n-1,1-\alpha/2}}{n-1}\right) = 1-\alpha$$

$$Pr\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right] = 1-\alpha$$

A **100% × (1 − α) CI for σ²** is given by $\left[(n-1)s^2/\chi^2_{n-1,1-\alpha/2}, (n-1)s^2/\chi^2_{n-1,\alpha/2}\right]$

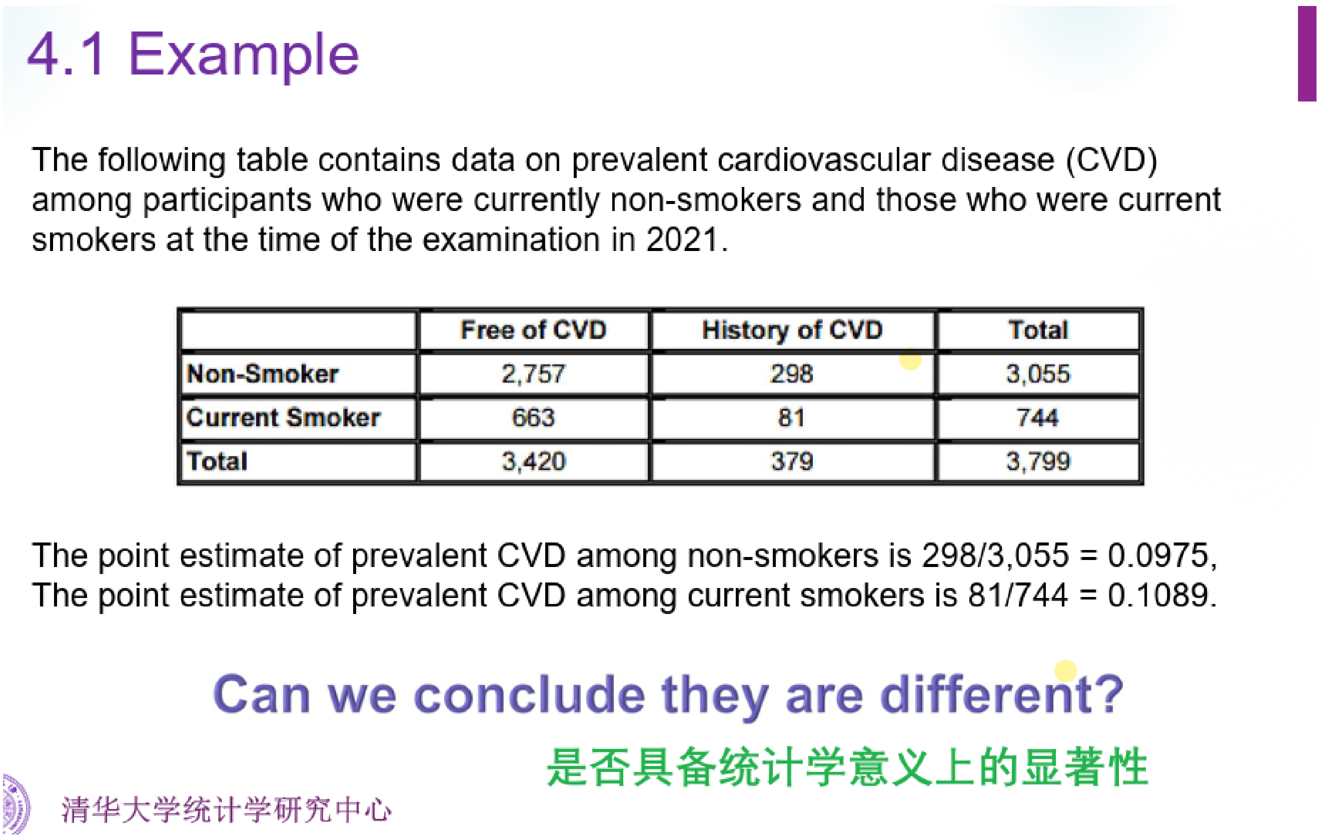
清华大学统计学研究中心

得到ci的估计

> SE depends on sample size

two sample

两个样本来自于不同的总体-比较这两个总体
eg两个不同的疗法



4.1 Example

The following table contains data on prevalent cardiovascular disease (CVD) among participants who were currently non-smokers and those who were current smokers at the time of the examination in 2021.

	Free of CVD	History of CVD	Total
Non-Smoker	2,757	298	3,055
Current Smoker	663	81	744
Total	3,420	379	3,799

The point estimate of prevalent CVD among non-smokers is 298/3,055 = 0.0975,
The point estimate of prevalent CVD among current smokers is 81/744 = 0.1089.

Can we conclude they are different?

是否具备统计学意义上的显著性

清华大学统计学研究中心

在样本量足够大的情况下，任何统计学意义上的显著性都能实现；但 ***医学意义上****的显著性（可能有其他标准ry 比如标准更高，统计学上无论有多显著，医学意义上都不具有显著性，反之统计学上没有显著性但医学意义上已经具有显著性*

目标：两重意义都具有显著性

条件：样本独立、样本量足够大（np、n（1-p） 均>5，两个样本都如此）

那么 **p1hat-p2hat~N(μ,V)** difference:作差

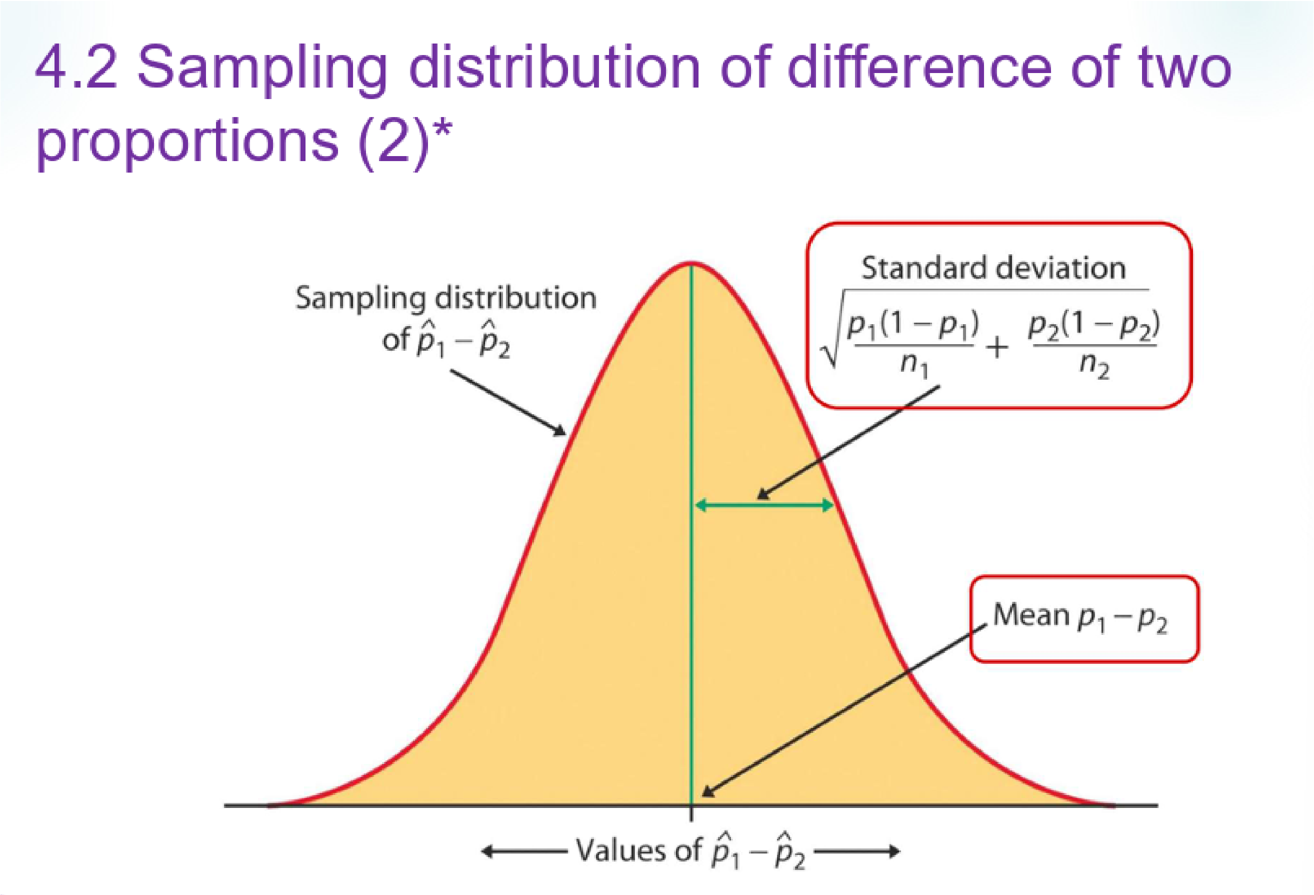
μ=p1-p2

V=p1(1-p1)/n1+p2(1-p2)/n2

已知sampling dist是正态分布，计算se和ci

①p1hat是p1为中心，p1(1-p1)/n1为方差的正态分布，p2同理

②因为样本独立，所以方差和mean相加减



4.2 Sampling distribution of difference of two proportions (2)*

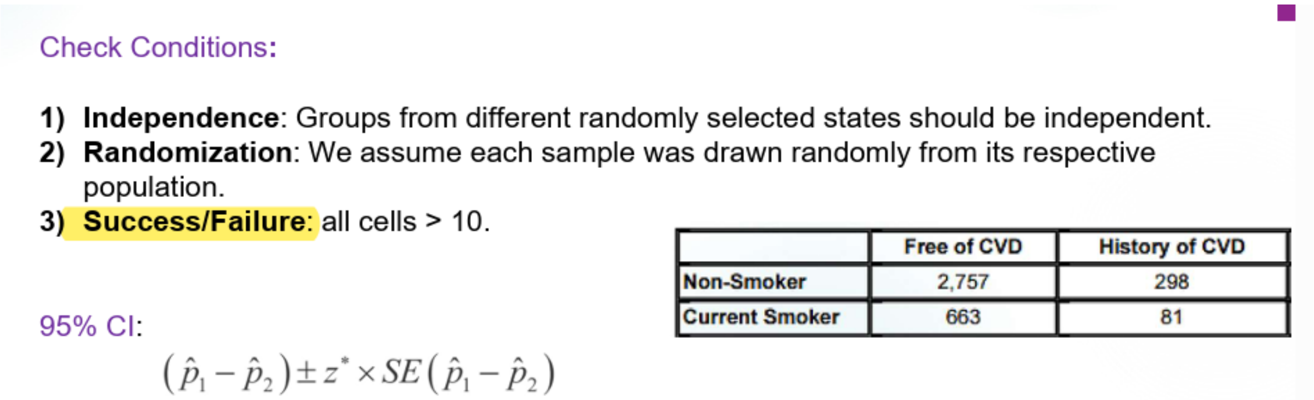
Sampling distribution of $\hat{p}_1 - \hat{p}_2$
Standard deviation
$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Mean $p_1 - p_2$
Values of $\hat{p}_1 - \hat{p}_2$

那么

The confidence interval is $\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \times SE\left(\hat{p}_1 - \hat{p}_2\right)$

where $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$

The critical value z* depends on the particular confidence level, e.g., 95%, 99%, etc.



Check Conditions:

1) **Independence**: Groups from different randomly selected states should be independent.
2) **Randomization**: We assume each sample was drawn randomly from its respective population.
3) **Success/Failure**: all cells > 10.

	Free of CVD	History of CVD
Non-Smoker	2,757	298
Current Smoker	663	81

95% CI:

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \times SE\left(\hat{p}_1 - \hat{p}_2\right)$$

这四个格子就是np所以需要和10check

$$0.0114 \pm 1.96 * 0.0126 = (-0.0133, 0.0361)$$

CI计算结果：因为 **0也落在范围之内**所以 **无法得出** 有差异 **的结论**注意对ci的理解

对于一个2x2列联表，N>40的情况下就不需要校正，否则需要进行连续性校正

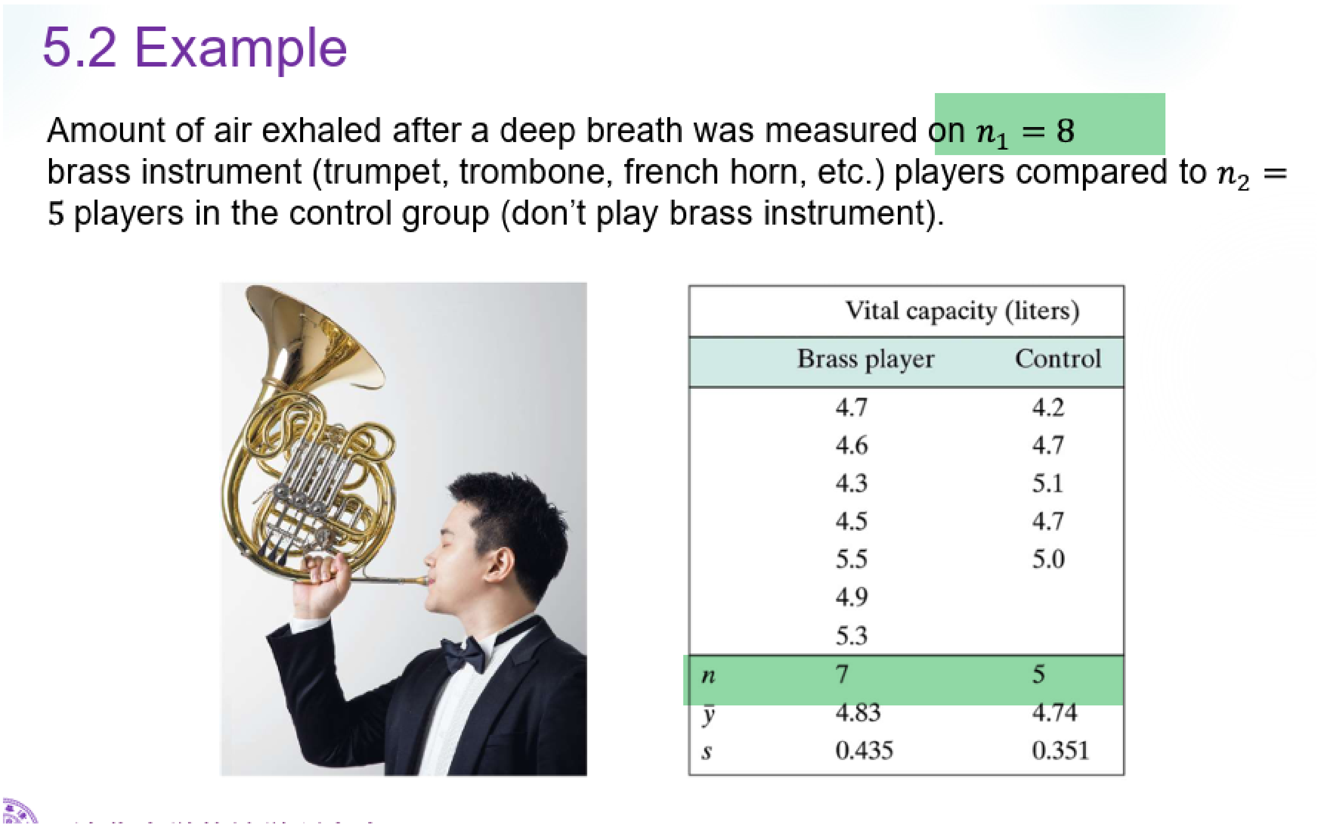
上面的应该是离散型变量得出的列联表（？？）

对于连续型变量，关注总体均值之间的差异性（两个总体独立）

- Sample sizes (n_1, n_2), means $(\bar{x}_1, \bar{x}_2)$, standard deviations of the sample (s_1, s_2).
- The point estimate of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$

作差

we want to compare μ1-μ2
接下来需要找sampling dist和se

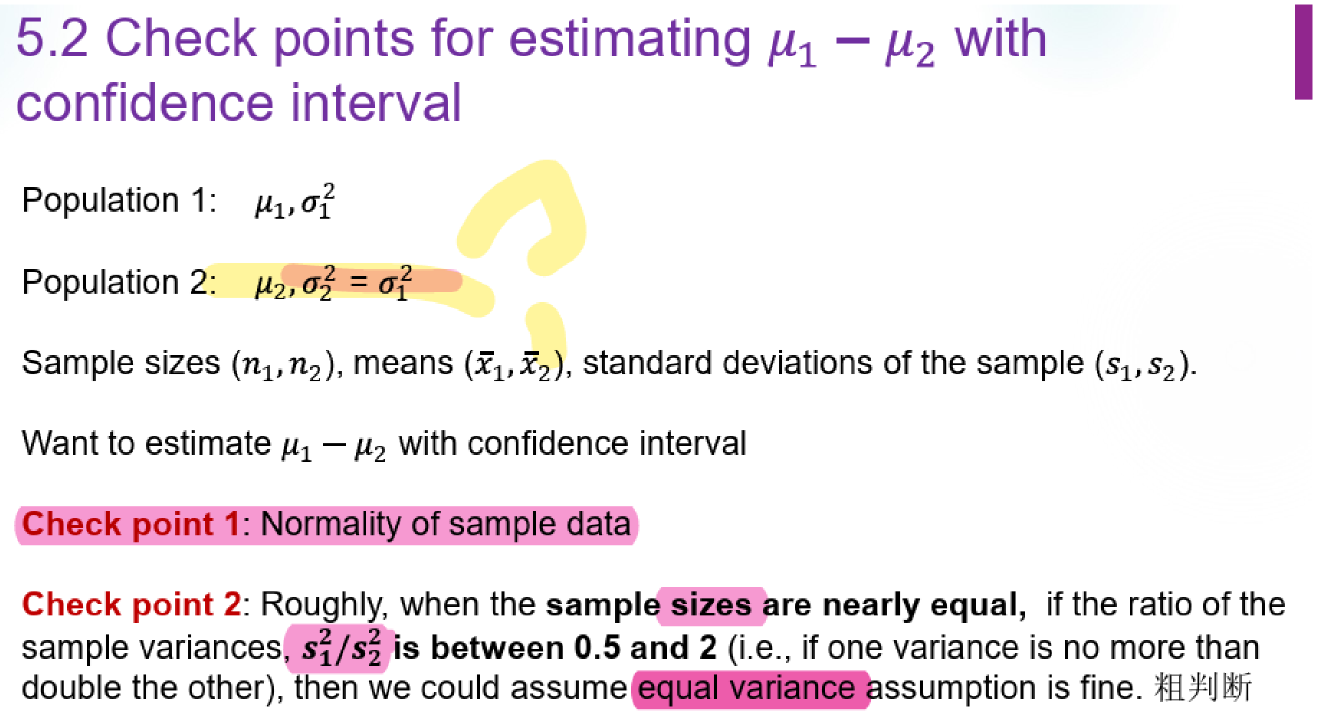


5.2 Example

Amount of air exhaled after a deep breath was measured on $n_1 = 8$ brass instrument (trumpet, trombone, french horn, etc.) players compared to $n_2 =$ 5 players in the control group (don't play brass instrument).

	Vital capacity (liters)	
	Brass player	Control
	4.7	4.2
	4.6	4.7
	4.3	5.1
	4.5	4.7
	5.5	5.0
	4.9	
	5.3	
n	7	5
$\bar{y}$	4.83	4.74
s	0.435	0.351

n1=8但是只拿到了七个人的数据
点估计得出： ybar1比ybar2更大，差异是**稳定 显著**的吗？
进行区间估计：



5.2 Check points for estimating $\mu_1 - \mu_2$ with confidence interval

Population 1: μ_1, σ_1^2

Population 2: $\mu_2, \sigma_2^2 = \sigma_1^2$

Sample sizes (n_1, n_2), means $(\bar{x}_1, \bar{x}_2)$, standard deviations of the sample (s_1, s_2).

Want to estimate $\mu_1 - \mu_2$ with confidence interval

Check point 1: Normality of sample data

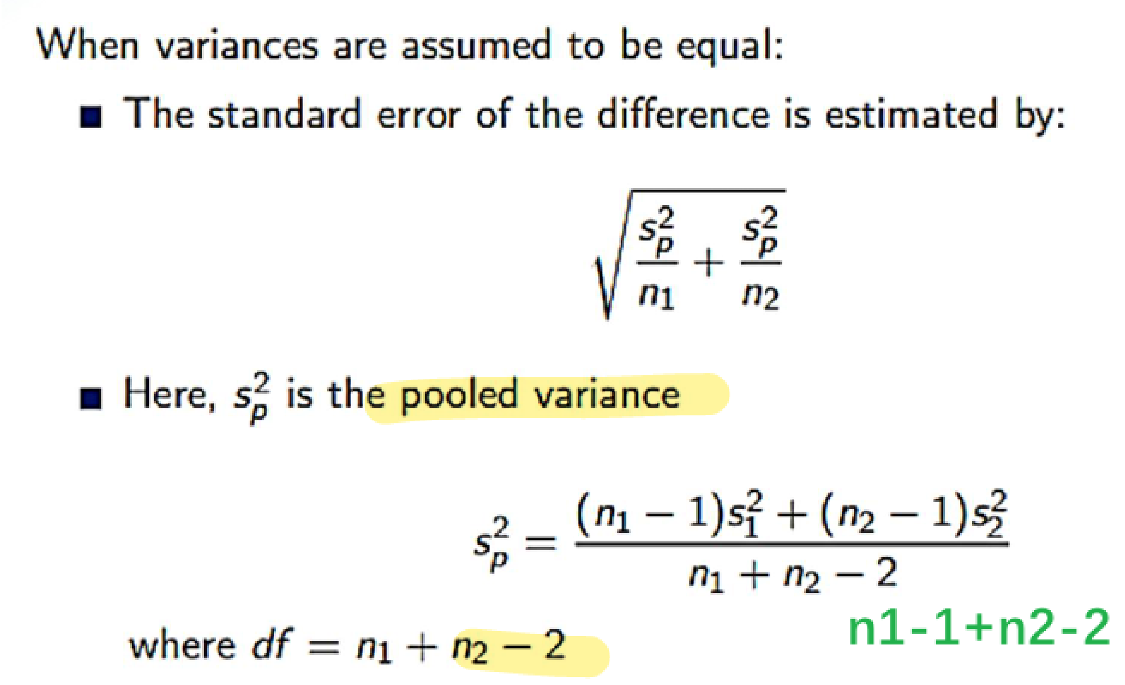
Check point 2: Roughly, when the **sample sizes are nearly equal,** if the ratio of the sample variances, s_1^2/s_2^2 **is between 0.5 and 2** (i.e., if one variance is no more than double the other), then we could assume equal variance assumption is fine. 粗判断

1. **check**样本数据是否正态（比如画qqplot）
   x1bar~N(μ1,σ^2) x2bar~N(μ2,σ^2)（见下一步，等方差）

等方差

2. 从数值上判断，当样本量量级差不多的时候可以计算sample variances ratio，在两倍之内就可以近似认为equal variance assumption is fine等方差

问题：σ取决于什么？（以及se是什么？）
pooled variance：



When variances are assumed to be equal:

- The standard error of the difference is estimated by:

$$\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}$$

- Here, s_p^2 is the pooled variance

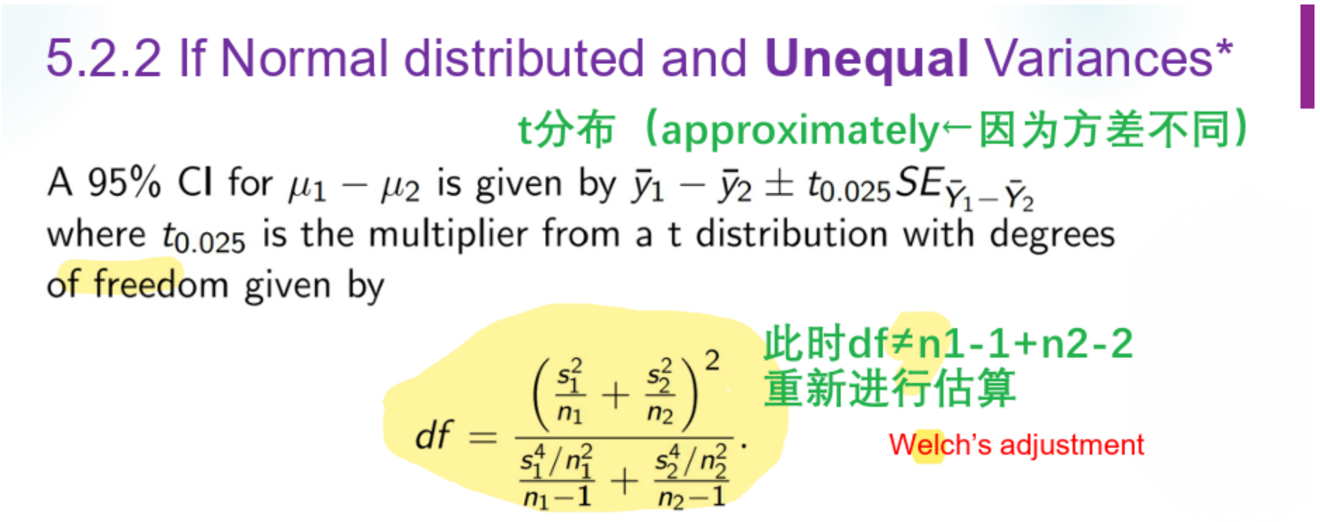
$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

where $df = n_1 + n_2 - 2$

n1-1+n2-2

如果有更多组样本，**都等方差**，计算方式同理
在 **等方差**的情况下：estimation x1bar-x2bar ←sampling dist?
x1bar-x2bar~N(μ1-μ2,σ^2/n1+σ^2/n2)（前提：满足正态性）

不等方差



5.2.2 If Normal distributed and **Unequal** Variances*

t分布（approximately←因为方差不同）

A 95% CI for $\mu_1-\mu_2$ is given by $\bar{y}_1-\bar{y}_2 \pm t_{0.025}SE_{\bar{Y}_1-\bar{Y}_2}$ where $t_{0.025}$ is the multiplier from a t distribution with degrees of freedom given by

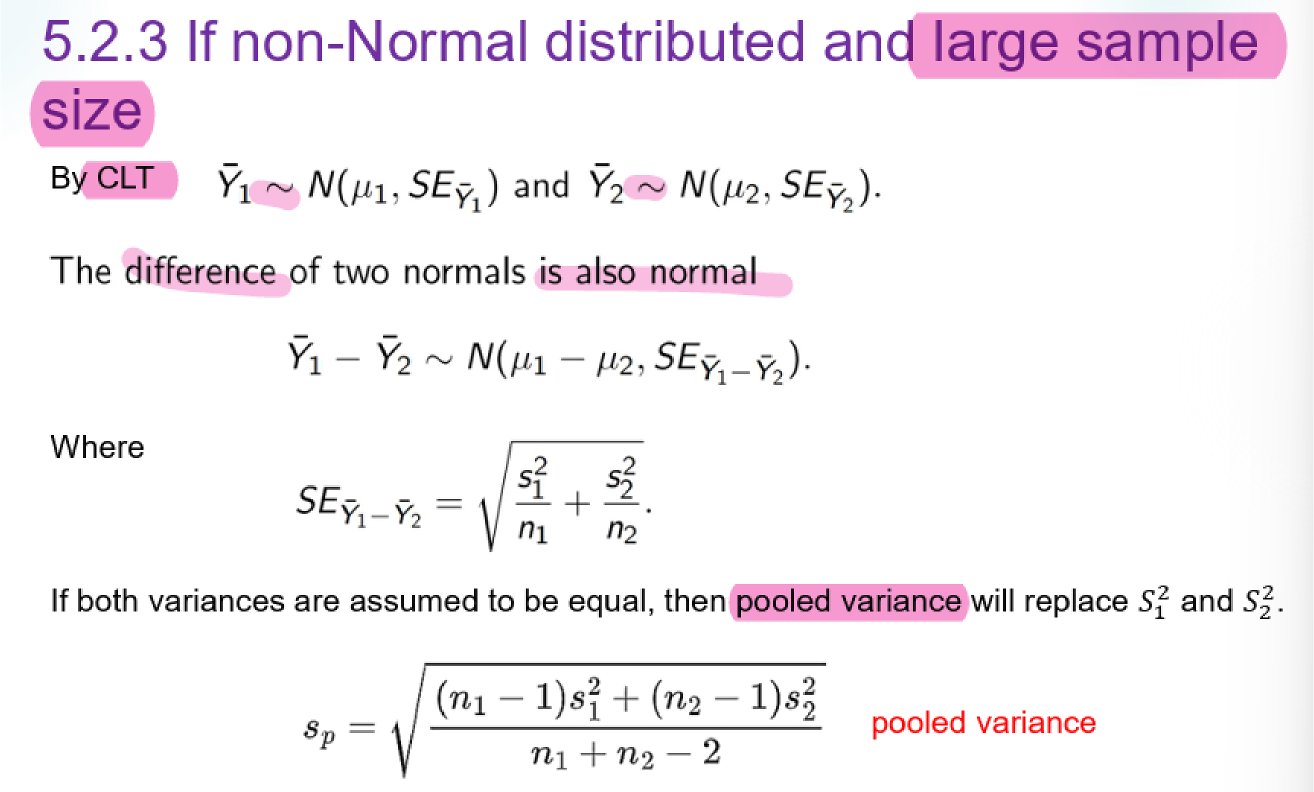
$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4/n_1^2}{n_1-1}+\frac{s_2^4/n_2^2}{n_2-1}}.$$

此时df≠n1-1+n2-2
重新进行估算

Welch's adjustment

This *df* formula is due to Welch (1947) and Satterthwaite (1946). It doesn't give an integer; people generally round down.

非正态大样本-clt



5.2.3 If non-Normal distributed and large sample size

By CLT $\bar{Y}_1 \sim N(\mu_1, SE_{\bar{Y}_1})$ and $\bar{Y}_2 \sim N(\mu_2, SE_{\bar{Y}_2})$.

The difference of two normals is also normal

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, SE_{\bar{Y}_1 - \bar{Y}_2}).$$

Where

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

If both variances are assumed to be equal, then pooled variance will replace S_1^2 and S_2^2.

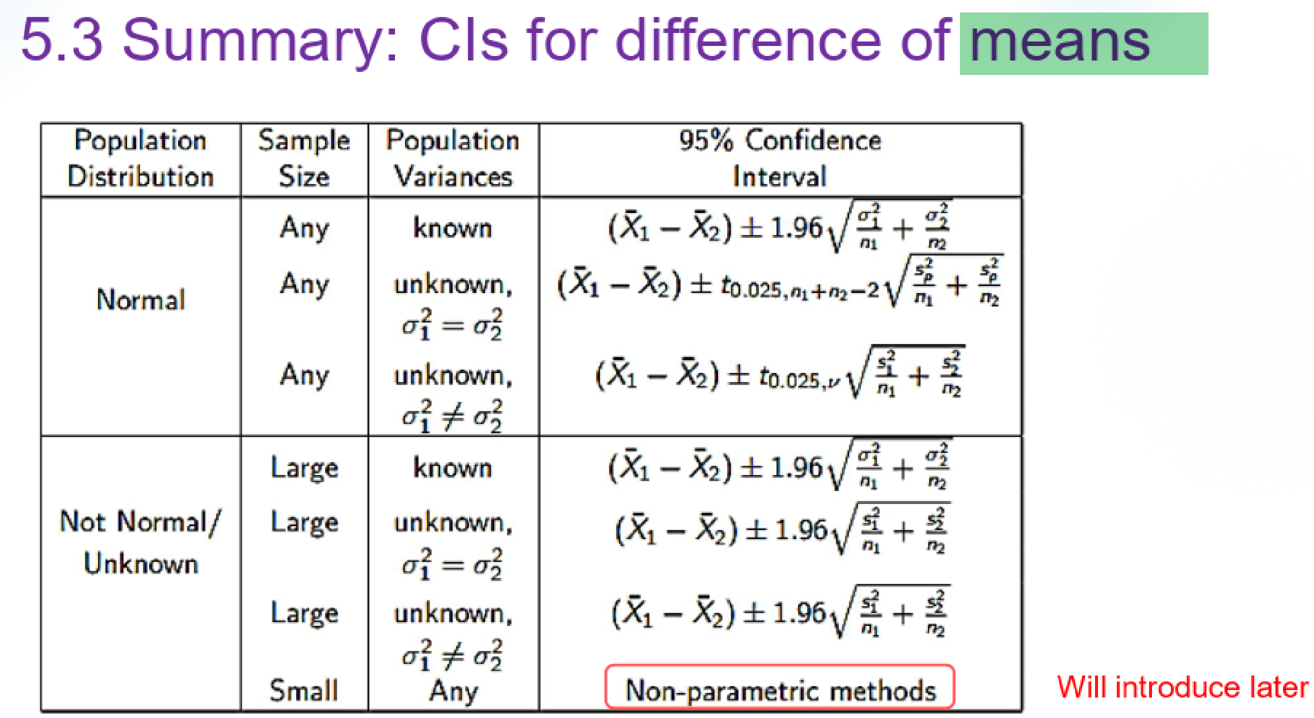
$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

pooled variance

没有 **等方差**，直接把方差的估计值代进去

正态分布写的到底是var还是sv

Summary



5.3 Summary: CIs for difference of means

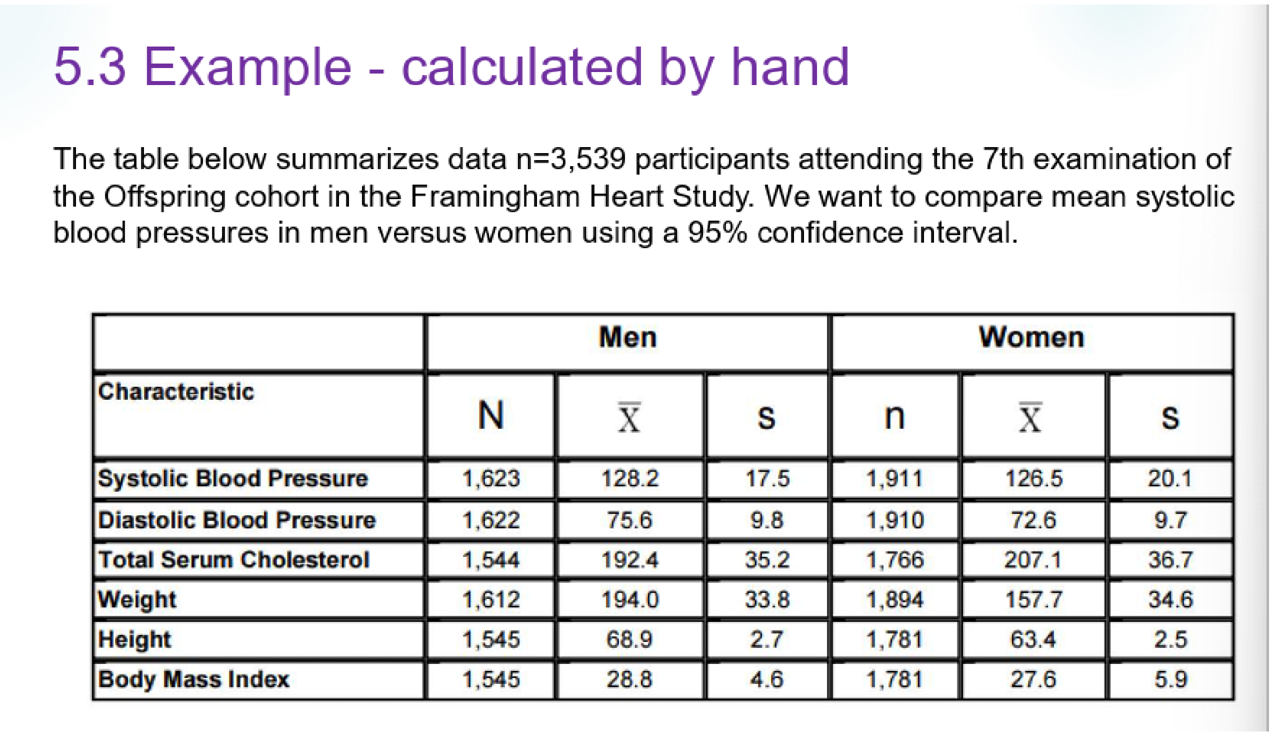
Population Distribution	Sample Size	Population Variances	95% Confidence Interval
Normal	Any	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Any	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025, n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
	Any	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025, \nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Not Normal/ Unknown	Large	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Large	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
	Large	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
	Small	Any	Non-parametric methods

Will introduce later

下一行，已经使用了clt所以都是根据正态分布用1.96（已经经过clt近似，不特别追求精确性）

非参数方法：撇开分布，从数据上考察

考试题型（可能）

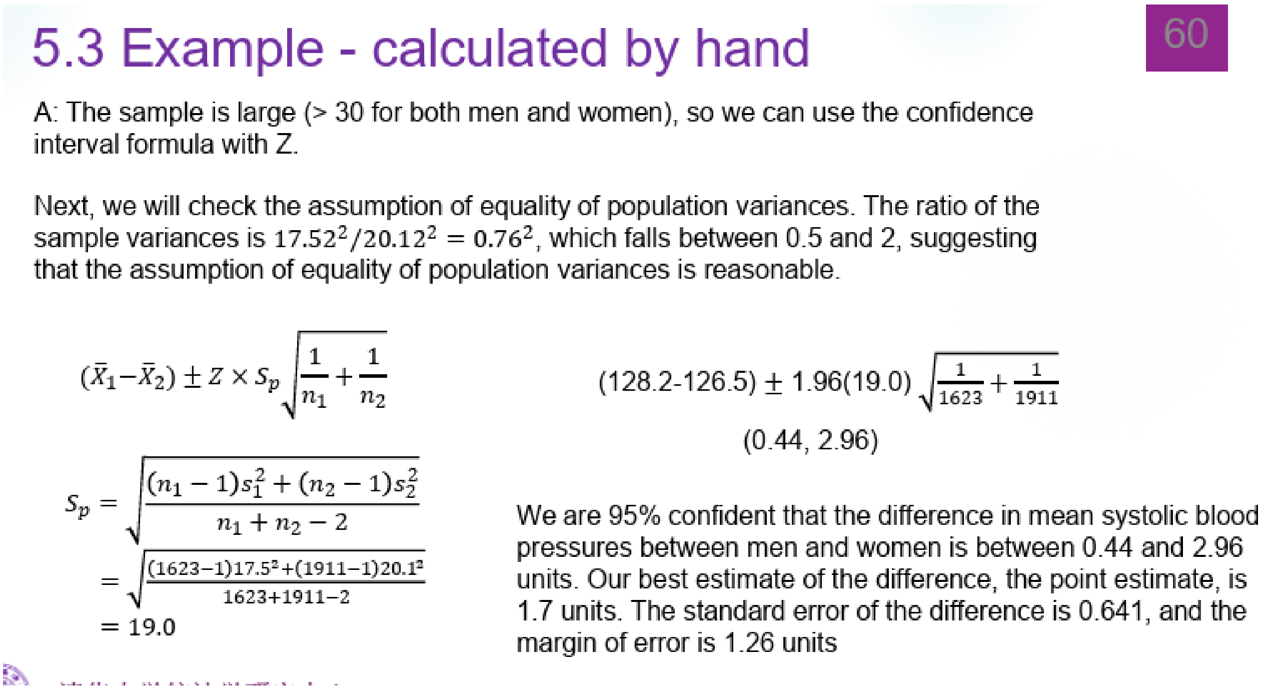


5.3 Example - calculated by hand

The table below summarizes data n=3,539 participants attending the 7th examination of the Offspring cohort in the Framingham Heart Study. We want to compare mean systolic blood pressures in men versus women using a 95% confidence interval.

	Men			Women		
Characteristic	N	$\bar{X}$	s	n	$\bar{X}$	s
Systolic Blood Pressure	1,623	128.2	17.5	1,911	126.5	20.1
Diastolic Blood Pressure	1,622	75.6	9.8	1,910	72.6	9.7
Total Serum Cholesterol	1,544	192.4	35.2	1,766	207.1	36.7
Weight	1,612	194.0	33.8	1,894	157.7	34.6
Height	1,545	68.9	2.7	1,781	63.4	2.5
Body Mass Index	1,545	28.8	4.6	1,781	27.6	5.9

随便找两个比较差异



5.3 Example - calculated by hand

60

A: The sample is large (> 30 for both men and women), so we can use the confidence interval formula with Z.

Next, we will check the assumption of equality of population variances. The ratio of the sample variances is $17.52^2/20.12^2 = 0.76^2$, which falls between 0.5 and 2, suggesting that the assumption of equality of population variances is reasonable.

$$(\bar{X}_1 - \bar{X}_2) \pm Z \times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$(128.2\text{-}126.5) \pm 1.96(19.0)\sqrt{\frac{1}{1623}+\frac{1}{1911}}$$

(0.44, 2.96)

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(1623-1)17.5^2+(1911-1)20.1^2}{1623+1911-2}}$$

$$= 19.0$$

We are 95% confident that the difference in mean systolic blood pressures between men and women is between 0.44 and 2.96 units. Our best estimate of the difference, the point estimate, is 1.7 units. The standard error of the difference is 0.641, and the margin of error is 1.26 units

图中的z改成t也一样（n足够大的情况下两个值几乎相等）

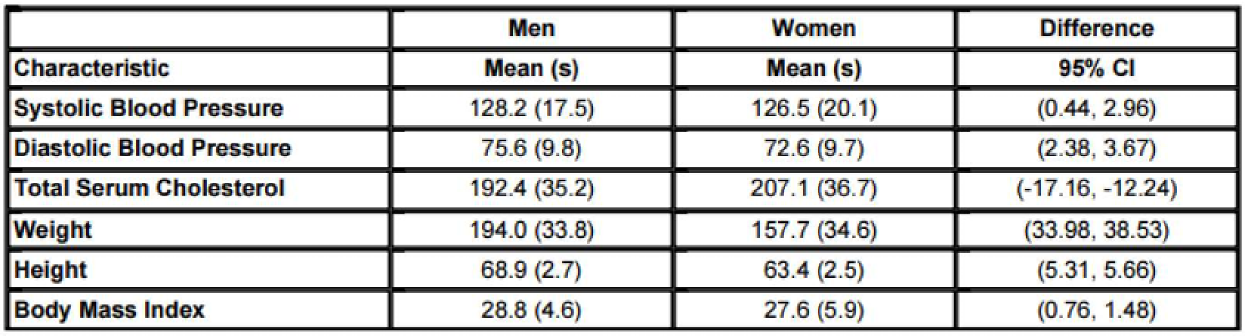
①check样本量

足够大→clt，正态

②等方差?

采用pooled variance

发现ci不包含0，差异显著（即稳定）

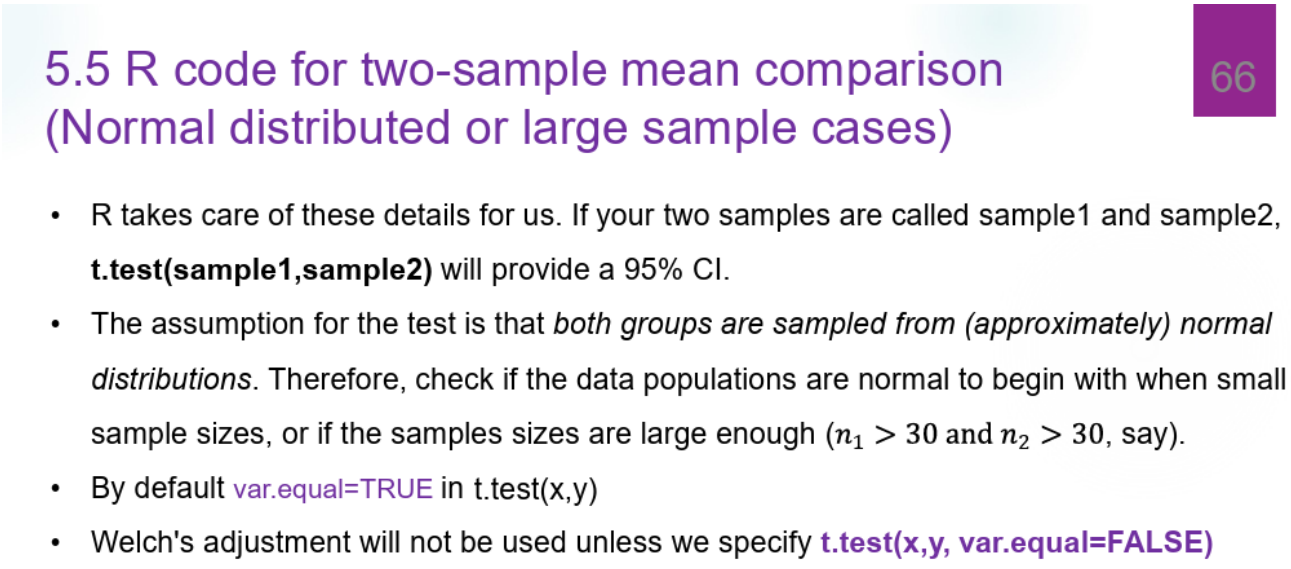


	Men	Women	Difference
Characteristic	Mean (s)	Mean (s)	95% CI
Systolic Blood Pressure	128.2 (17.5)	126.5 (20.1)	(0.44, 2.96)
Diastolic Blood Pressure	75.6 (9.8)	72.6 (9.7)	(2.38, 3.67)
Total Serum Cholesterol	192.4 (35.2)	207.1 (36.7)	(-17.16, -12.24)
Weight	194.0 (33.8)	157.7 (34.6)	(33.98, 38.53)
Height	68.9 (2.7)	63.4 (2.5)	(5.31, 5.66)
Body Mass Index	28.8 (4.6)	27.6 (5.9)	(0.76, 1.48)

配对数据（非独立样本）

eg 服药前-服药后
考虑 **差异本身**（把差异看成 一个样本）（原本的 *两*个 样本不重要）
差异性处理-回到one sample问题

R code 采用t test



5.5 R code for two-sample mean comparison (Normal distributed or large sample cases)

66

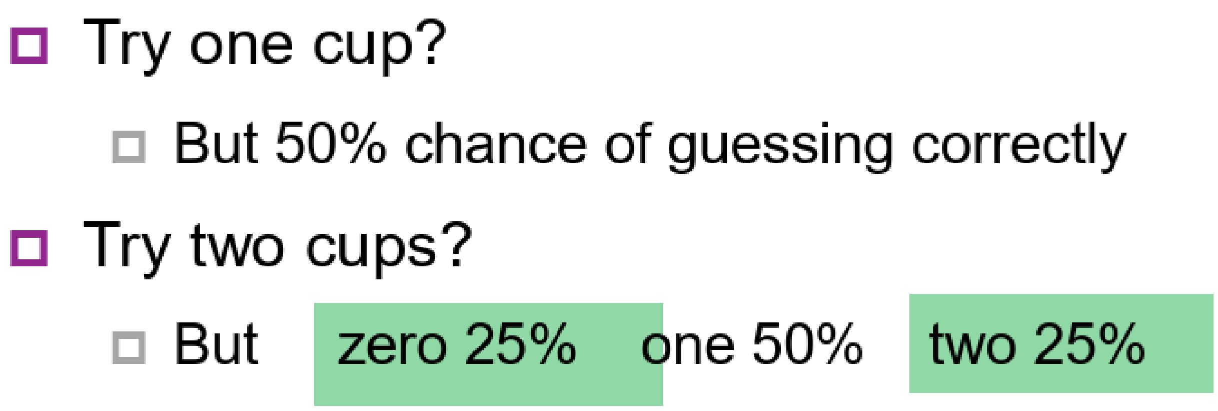
- R takes care of these details for us. If your two samples are called sample1 and sample2, **t.test(sample1,sample2)** will provide a 95% CI.
- The assumption for the test is that *both groups are sampled from (approximately) normal distributions*. Therefore, check if the data populations are normal to begin with when small sample sizes, or if the samples sizes are large enough ($n_1 > 30$ and $n_2 > 30$, say).
- By default var.equal=TRUE in t.test(x,y)
- Welch's adjustment will not be used unless we specify **t.test(x,y, var.equal=FALSE)**

前提是数据是正态分布（的样式），特别是小样本
默认等方差，否则要改参数

Hypothesis testing假设检验

假设 statement
假设检验：验证称述是否正确

General concepts



- □ Try one cup?
  - □ But 50% chance of guessing correctly
- □ Try two cups?
  - □ But zero 25% one 50% two 25%

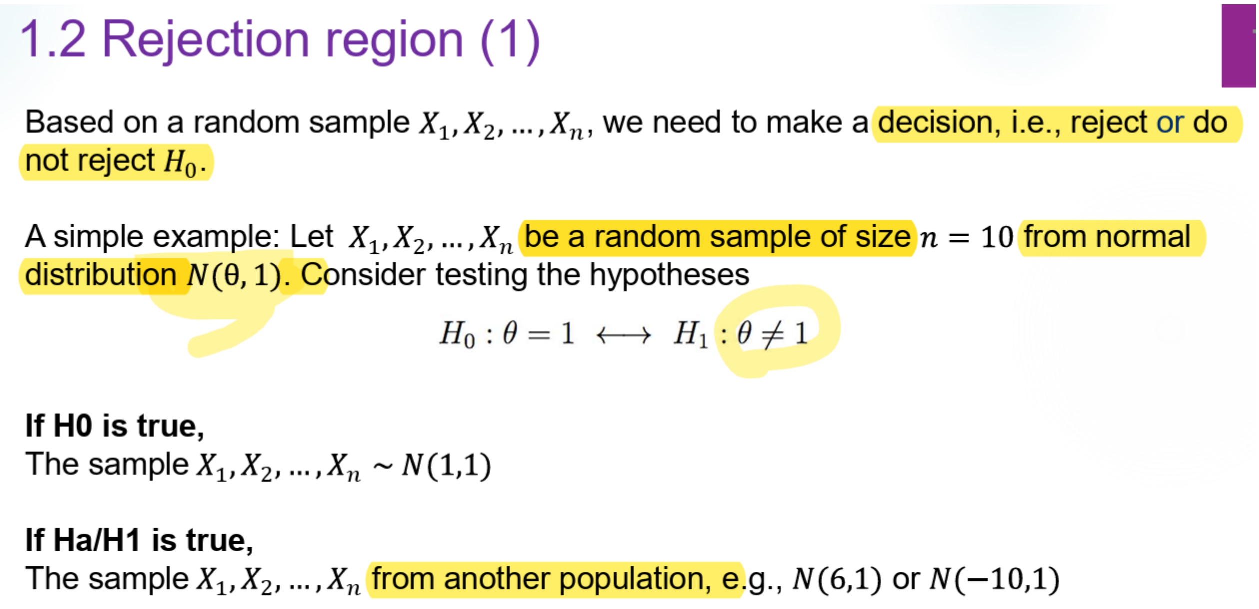
假设

直接验证： esitimation 差异非常显著
反证法（间接）： 与claim相反
claim：r不能抽到卡 Ha 备择假设 aka H1 （what we're trying to show is true)
假设：r能抽到 **H0** 原假设/无效假设 (what we're trying to **disprove**)
无法拒绝H0的可能性但希望Ha是真的

判断范围——rejection region

哪些数据呈现出来的时候会拒绝假设?
假设：四月份气温在15°C左右
预期（H0=T）：14、16、15、13、15...（正常范围：10~20）
数据：30、31、32、30、29...异常范围：30+ry

- 针对样本能呈现的数据：eg-∞~+∞
- acceptance region 合理的，容错
- rejection region 拒绝原假设
  小概率事件在一次实验观测中不可能发生
  如果H0是真的：气温在15左右
  如果H0不是真的：气温 *可能* 是任何值 但不是15)



1.2 Rejection region (1)

Based on a random sample $X_1, X_2, \ldots, X_n$, we need to make a decision, i.e., reject or do not reject H_0.

A simple example: Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n = 10$ from normal distribution $N(\theta, 1)$. Consider testing the hypotheses

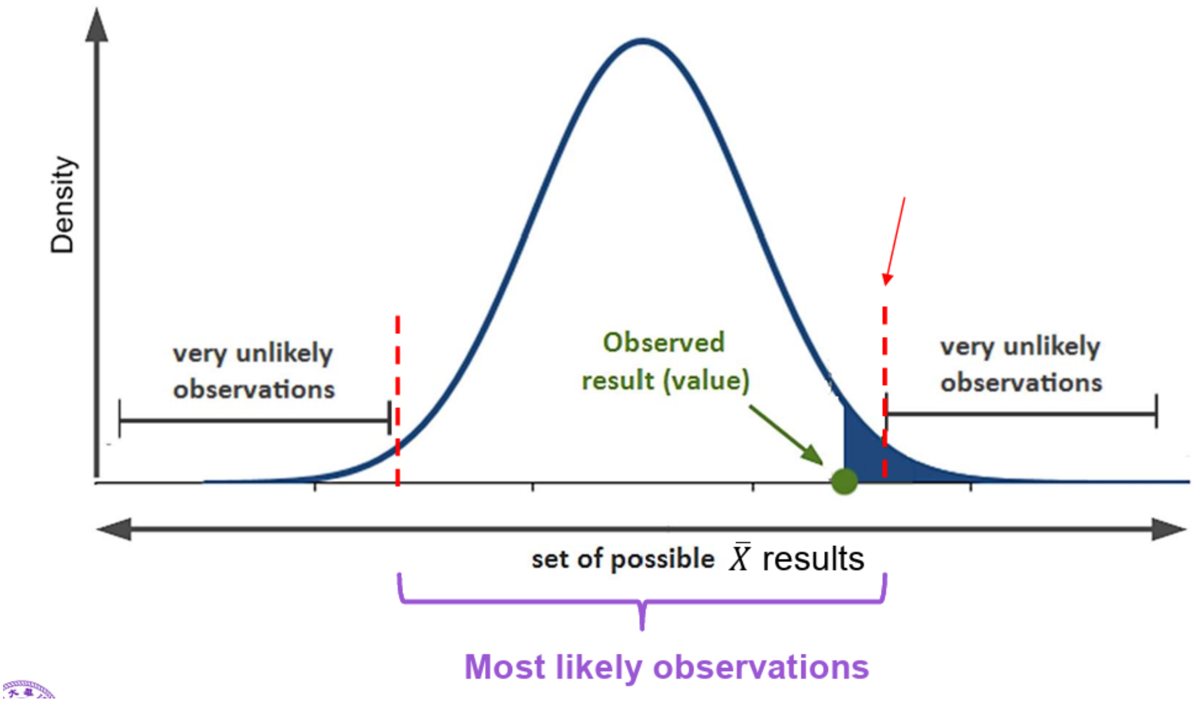
$$H_0 : \theta = 1 \longleftrightarrow H_1 : \theta \neq 1$$

If H0 is true,
The sample $X_1, X_2, \ldots, X_n \sim N(1,1)$

If Ha/H1 is true,
The sample $X_1, X_2, \ldots, X_n$ from another population, e.g., $N(6,1)$ or $N(-10,1)$

H0=T：Xbar~N(1,1/10)
比如观测到100，需要一个 *指标*： 用以上正态分布中100~∞下的概率（面积）表示它的 *位置*，与 *quantile*有关——**p-value**

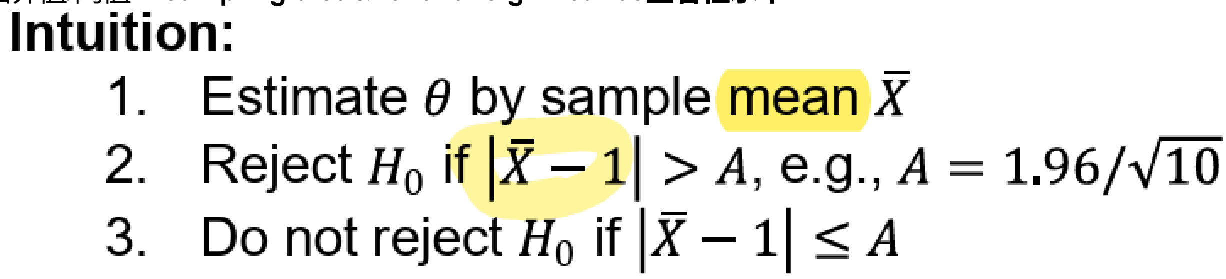
正态分布图像：x轴为xbar可能出现的值



被分为两部分：接受域；拒绝域 一旦知道统计量对应的分布类型，这两部分的分界是事先定好的 抽样分布的类型决定了估计值出现在不同区域的可能性
想用xbar估计θ（点估计），背后的抽样分布是N（参数），然后就可以划分范围
所以如何划分?

临界值——critical value

- 检验统计量 xbar
- 临界值/阈值 A **sampling dist & level of significance显著性水平**

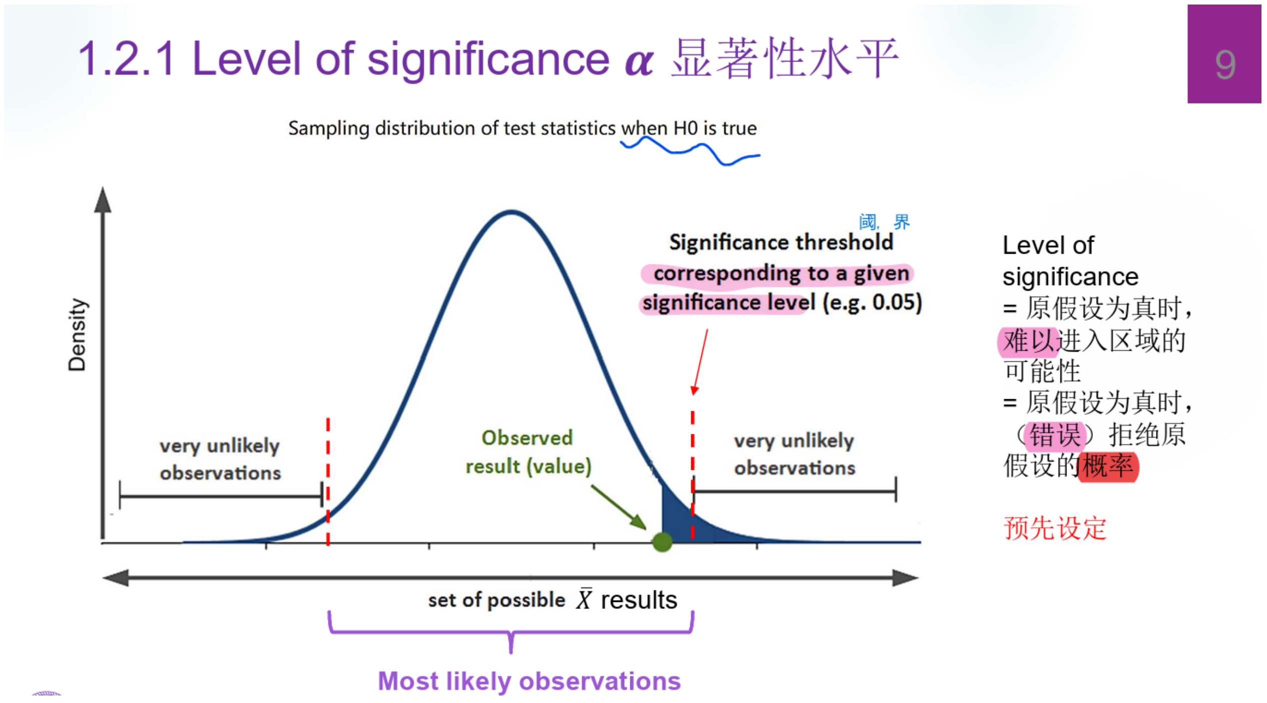


Intuition:

1. Estimate θ by sample mean $\bar{X}$
2. Reject H_0 if $|\bar{X}-1| > A$, e.g., $A = 1.96/\sqrt{10}$
3. Do not reject H_0 if $|\bar{X}-1| \leq A$

显著性水平：

理解①当H0为真时，数据难以落入区域的可能性



1.2.1 Level of significance α 显著性水平

Level of significance
= 原假设为真时，难以进入区域的可能性
= 原假设为真时，（错误）拒绝原假设的概率

预先设定

显著性水平↑，拒绝的数变多，结论可能也会随之改变
理解②H0为真时，错误地拒绝了原假设的概率 ← 犯错的可能性 **最多**是这么大
很小的可能性依然存在
容错范围

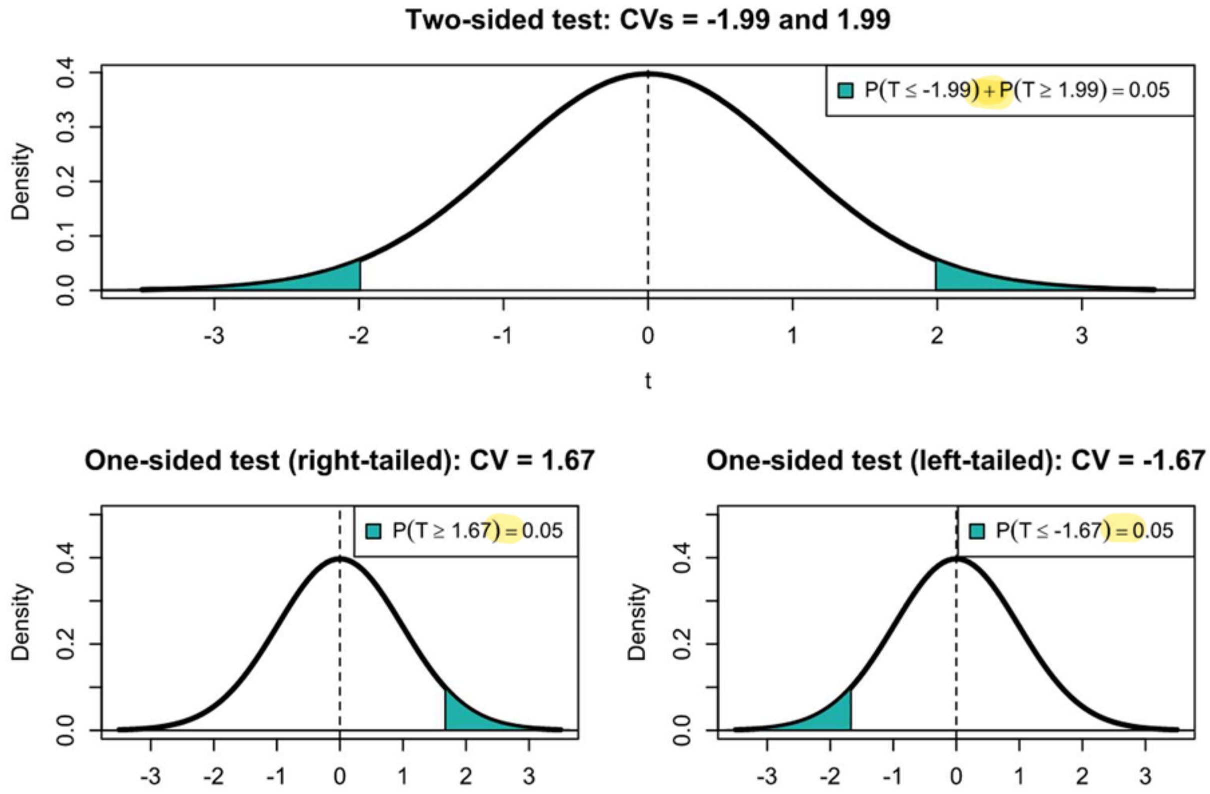
续-rejection region

- Rejection/critical region $D = \{\boldsymbol{X} = (X_1, \cdots, X_n): |\bar{X} - 1| > A = 0.62\}$
- Acceptance region $D^c = \{\boldsymbol{X} = (X_1, \cdots, X_n): |\bar{X} - 1| \leq A = 0.62\}$

凡是出现在拒绝域中的都拒绝

单边检验 双边检验

two-sided test = vs ≠
one-sided test = vs > or <
方向性

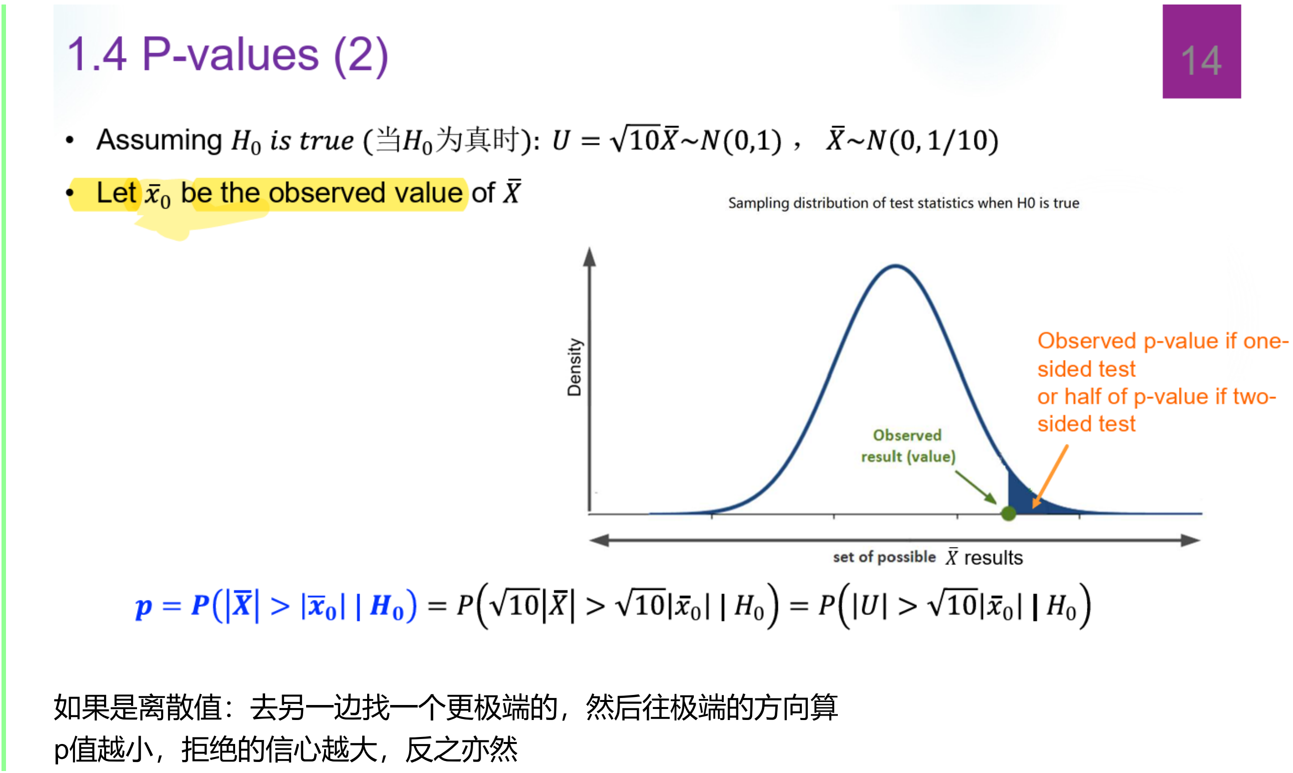


容错范围随之改变，单侧所有结果都可以被接受
显著性水平对应着临界值的大小（在某一sampling dist下）

p-values

同样落在拒绝域，哪种拒绝的把握度更大?
太靠近不被拒绝的范围 把握就变小了eg胆固醇正常值60~90，一个人59，另一个人20
p-value是用于衡量相对位置的指标

> **要求：假设H0为真**
> →得到抽样分布，根据数据得到观测值x0bar，
> p-value：检验统计量比观测值还要极端的可能性
> 极端的方向：拒绝域出现的方向
> 如果是双边：对称（如果不对称分布的话另算）



1.4 P-values (2)

- Assuming H_0 *is true* (当H_0为真时): $U=\sqrt{10}\bar{X} \sim N(0,1)$， $\bar{X} \sim N(0,1/10)$
- Let $\bar{x}_0$ be the observed value of $\bar{X}$

Sampling distribution of test statistics when H0 is true
Density
Observed result (value)
Observed p-value if one-sided test
or half of p-value if two-sided test
set of possible $\bar{X}$ results

$$\boldsymbol{p=P(|\bar{X}|>|\bar{x}_0|\mid H_0)}=P\left(\sqrt{10}|\bar{X}|>\sqrt{10}|\bar{x}_0|\mid H_0\right)=P\left(|U|>\sqrt{10}|\bar{x}_0|\mid H_0\right)$$

如果是离散值：去另一边找一个更极端的，然后往极端的方向算

p值越小，拒绝的信心越大，反之亦然

One-sample proportion test

Two-sample proportion test(large)

small sample proportions

one-sample mean test

two-sample mean test