Parameter estimation: point estimation and interval estimation

1.1 Population and Sample

The table below succinctly summarizes the salient differences between a population and a sample (a parameter and a statistic):

Comparison between parameters and statistics	
Statistics	*Parameters*
• Describes a sample	• Describes a population
• Always known	• Usually unknown
• Random, changes upon repeated sampling	• Fixed
• Ex: $\bar{X}, S^2, S$	• Ex: μ, σ^2, σ

Parameter estimation: **point estimation** and **interval estimation**

点估计

unbiased estimator

- 选sd小的

Accuracy: We say that $\widehat{\theta}$ is an **unbiased estimator** of θ if and only if

无偏估计量

$$E(\widehat{\theta}) = \theta$$

RESULT: When $Y_1, \ldots, Y_n$ is a random sample,

$$E(\overline{Y}) = \mu$$

$$E(S^2) = \sigma^2$$

Precision: Suppose that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of θ. We would like to pick the estimator with smaller variance, since it is more likely to produce an estimate close to the true value θ.

-

Standard error

- 上文的另一种表述

SUMMARY: We desire point estimators $\widehat{\theta}$ which are **unbiased** (perfectly accurate) and have **small variance** (highly precise).

TERMINOLOGY: The **standard error** of a point estimator $\widehat{\theta}$ is equal to

$$se(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta})}.$$

这和标准差的区别是啥

Note:

$$\text{smaller } se(\widehat{\theta}) \Longleftrightarrow \widehat{\theta} \text{ more precise.}$$

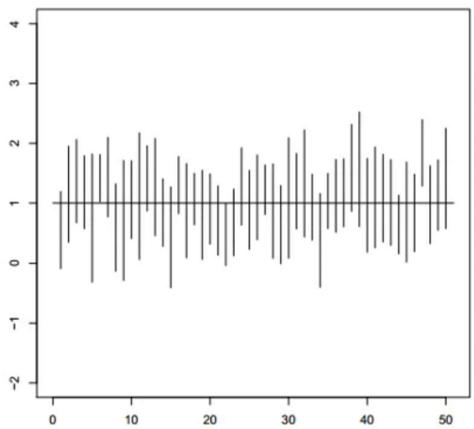
区间估计

CI=置信度=置信水平

Confidence level 置信水平: 置信区间包含总体平均值的概率

又称 **置信度**

E.g.：95%的置信水平表示：抽取的100个样本，有100个置信区间，其中有95个置信区间可能包含总体的真实平均值。



清华大学统计学研究中心

Precision 精度

- 置信区间变窄的速度不像样本量增加的速度那么快，也就是说并不是样本量增加一倍，置信区间也变窄一倍

1.4 Evaluate an estimator – interval estimator

Precision 精度：置信区间的平均长度

随着置信度的上升，置信区间的跨度也就越大，对参数估计的精度下降

Calculate an interval estimator

1.5 Calculate an interval estimator

If you want to calculate a confidence interval on your own, you need to know:

1. The point estimate you are constructing the confidence interval for
2. The critical values for the test statistic
3. The standard error of point estimate (depends on sample size)
4. The sample size

置信区间= point estimate ± $\underbrace{\text{(critical values × standard error)}}_{\text{Margin of Error}}$

Population Proportion

分类变量乃至二项分布

We can connect these binary outcomes to the *Bernoulli trails* ? assumptions for each individual in the sample:

1. Each trial results in only two possible outcomes, labeled as "success" and "failure."
2. The trials are independent.
3. The probability of a success in each trial, denoted as p, remains constant. It follows that the probability of a failure in each trial is $1 - p$.

Point Estimator of Proportion p：phat

Suppose we define Y = the number of successes out of n sampled individuals so $Y \sim \text{bin}(n,p)$. A natural point estimator for p, *the population proportion*, is

$$\widehat{p} = \frac{Y}{n},$$

the **sample proportion**. $\widehat{p}$ is read as p hat.

phat的分布

$\widehat{p}$ is a unbiased estimator of p. That is,

$$E(\widehat{p}) = p.$$

To quantify the precision of $\widehat{p}$,

$$\text{var}(\widehat{p}) = \frac{p(1-p)}{n}$$

Question: What is the (asymptotic) distribution of $\widehat{p}$?

$$\widehat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$$

Confidence Interval for phat

2.3 Confidence Interval for $\hat{p}$

Recall that $\hat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$.

Let us define z_α be the upper α percentage point of the standard normal distribution, i.e., $P(Z > z_\alpha) = \alpha$.

An approximate $100(1-\alpha)\%$ confidence interval for p is

$$\left(\hat{p} - \boxed{z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},}\ \ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

margin of error 误差幅度

Interpretation

2.4 CI - Interpretation

21

- In practice, we only perform the study once, and get CI (0.193, 0.237).
- We have no way of knowing if this interval that we calculated is one of the 95% (that covers the true parameter) or one of the 5% that does not.
- Thus **we are 95% confident that the true response rate of this new treatment in 2L NSCLC patients is between 0.193 and 0.237**
- Every time that we calculate a 95% CI, then there is a 5% chance that the CI does not cover the quantity that you are estimating.

What if the sample size is small-non-parametric testing

2.7 What if the sample size is small*

Example. As part of a demographic survey of her Statistical Consultant course, Miss Wang asks students if they have any statistical consultant experience before. The following are the data from her course:

Experience	Count
Yes	7
No	14
----------	--
Total	21

Point estimate = 7/21 = 0.33

Calculate exact binomial CI satisfying $P(p_{LB} < p < p_{UB}) = 1 - \alpha,$

(e.g., Clopper-Pearson interval)

$$\sum_{k=0}^{k} \binom{n}{k} p_{UB}{}^{k}(1-p_{UB})^{n-k} = \frac{\alpha}{2}$$

$$\sum_{k=x}^{n} \binom{n}{k} p_{LB}{}^{k}(1-p_{LB})^{n-k} = \frac{\alpha}{2}$$

Calculation demanding…

Let R help you

Summarize

2.8 Summarize: CI construction for population proportion

1. **Check all conditions** before using the sampling distribution of the sample proportion. For the confidence interval, we will use

   - $n\hat{p} > 5$ and ??
   - $n(1-\hat{p}) > 5$

   If conditions are NOT satisfied, we will use exact methods (more complicated, based on the relationship between the binomial and F-distribution).

2. **Construct the general form**

Population Distribution	Sample Size	Population Variance	95% Confidence Interval
Binomial	Large	-	$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$
	Small	-	Exact methods

population mean

群体呈正态分布+σ已知

- 那么群体均值-sd也呈正态分布但要/n
- 点估计

3.1 Point estimator of population mean

- Suppose $Y_1, Y_2, \dots, Y_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution.
- Sample mean $\bar{Y}$ is a reasonable point estimator of the population mean μ.
- Properties $\overline{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

$\overline{Y}$ is an **unbiased** estimator of μ.
$\text{se}(\overline{Y}) = \sigma/\sqrt{n}$ gives the variability of $\bar{Y}$, i.e. how "close" we expect $\bar{Y}$ to be to μ.

-
- 区间估计

3.3 CI for μ

31

- $\bar{y}$ provides an estimate of μ.
- By CLT, $\bar{Y}$ is $N(\mu; \frac{\sigma^2}{n})$. This holds perfectly when the data $Y_1; \dots; Y_n$ are normal, otherwise it's approximate.
- The 68/95/99.7 rule says that any normal random variable is within about 2 standard deviations of its mean 95% of the time.
- Estimate $\frac{\sigma}{\sqrt{n}}$ by SE of $\bar{Y}$.

-

3.4 CI for μ. Assume normality and KNOWN σ

Recall if $Y_1, \dots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **known**, then

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Similar to CI derivation in population proportion, a $100(1-\alpha)\%$ CI for μ is given by

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

-

正态分布+σ未知-t分布

- df=v=n-1
- 图像特点-更矮

However, population standard deviation σ is usually unknown. Replacing it with the **sample standard deviation** S, we get a new sampling distribution:

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a t **distribution** with degrees of freedom $\nu = n - 1$.

The t distribution was published by Gosset in 1908 & related to quality control at Guinness brewery.

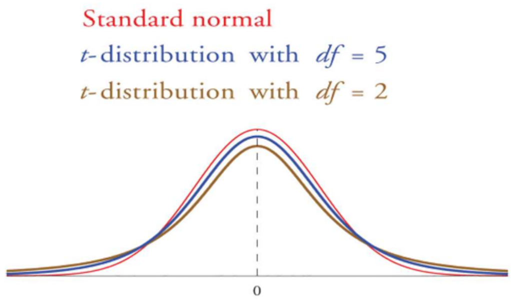
-

3.5 t distribution

34

The t distribution has the following characteristics:

1. It is continuous and symmetric about 0.
2. It is indexed by a value ν called the degrees of freedom.
3. As $\nu \longrightarrow \infty$, $t(\nu) \longrightarrow \mathcal{N}(0,1)$.
4. When compared to the standard normal distribution, the t distribution, in general, is less peaked and has more probability (area) in the tails.



-

3.6 CI for μ. Assume normality and UNKNOWN σ

Recall if $Y_1, \dots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **unknown**, then

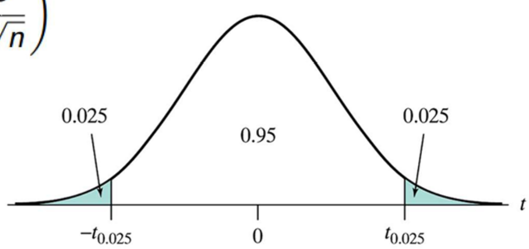
$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

A $100(1-\alpha)\%$ CI for μ is given by

We replace 1.96 (from a normal) by the equivalent t distribution value, denoted $t_{0.025}$ for 95% CI.

$$\left(\overline{y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\quad \overline{y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

where S is the sample standard deviation.

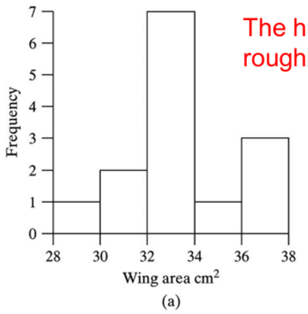


清华大学统计学研究中心

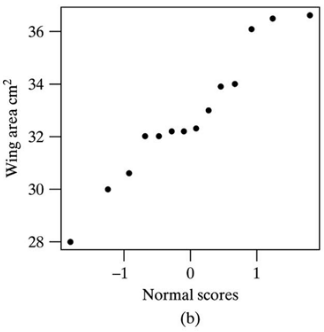
-

非正态分布-大样本-CLT

- check



The histogram looks roughly bell-shaped



the normal probability plot (QQ plot) looks reasonably straight

-

For $n < 30$ the data must be normal; check this with normal probability plot. For $n \geq 30$ don't worry about it.

-

summarize

3.9 Summarize: CI construction for population mean

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

Note:

- For n < 30 the data must be normal.
- Interpretation is important. “With 95% confidence the true mean of population characteristic is between a and b units .”

星号部分

3.10 Point estimator of population variance*

- Suppose $Y_1, Y_2, \dots, Y_n$ is a random sample from some distribution.
- Sample variance S^2 is a reasonable point estimator of the population variance σ^2. E(S^2) = σ^2.
- How can we evaluate the precision of our estimation?
  - Through sampling distribution of S^2

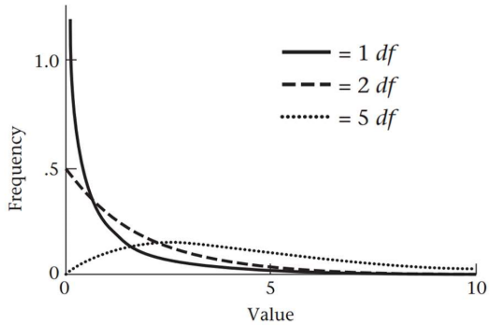
3.10.1 Chi-square distribution – Recall*

If $G = \sum_{i=1}^{n} X_i^2$

where $X_1, \ldots, X_n \sim N(0,1)$

and the X_i's are independent, then G is said to follow a **chi-square distribution with *n* degrees of freedom (*df*)**. The distribution is often denoted by χ_n^2.

General shape of various χ^2 distributions with *d df*



As *df* gets large, curve is less skewed, more normal.

3.10 *CI for S^2

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

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

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

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

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

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

Comparison two populations

星号部分

4.2 Sampling distribution of difference of two proportions (1)*

Assuming the samples are **independent**, and the samples sizes are **large enough**, the sampling distribution can be approximated normal:

The number of expected successes and failures in **BOTH** samples must be at least 5* (recommended 15).

? ?

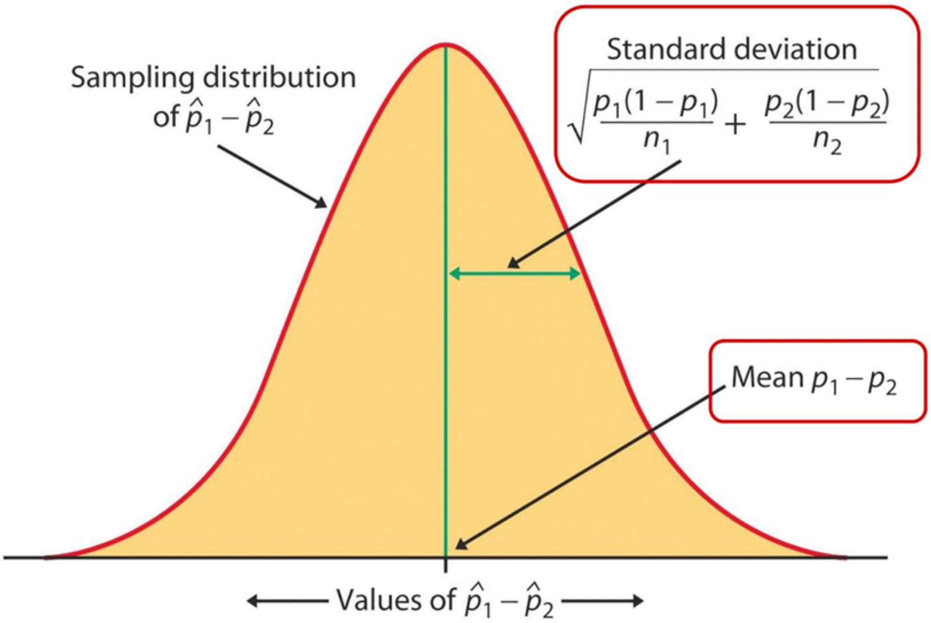
$$n_1 p_1 > 5^*,\ n_1(1-p_1) > 5^*$$
$$n_2 p_2 > 5^*,\ n_2(1-p_2) > 5^*$$

$$\hat{p}_1 - \hat{p}_2 \sim N(\mu, V)$$

作差

where $\mu = p_1$ - p_2, $V = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$

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



CIs for difference of proportions

- SE和上文计算方法有不同
- check

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

-

4.3 CIs for difference of proportions

When the conditions are met, we are ready to find the confidence interval for the difference of two proportions:

??????????????????????

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

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.

-
- 校正

Note:
Continuity Correction of Yates if the total N<40 for a 2 × 2 table

-

mean of Two independent samples

- 总体思想是作差

- Be interest to compare two groups with respect to their **mean** scores on a **continuous** outcome.
- For example, we might be interested in comparing mean systolic blood pressure in men and women, or perhaps compare BMI in smokers and non-smokers.
  - Between two **independent** groups!
- 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

正态分布

- check1 等方差

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

-
- 计算sp

5.2.1 If Normal distributed and equal variances

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 $df = n_1 + n_2 - 2$

-
- 不等方差:

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

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

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

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

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 on

Example - calculated by hand

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

,

5.3 Example - calculated by hand

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.

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

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

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

$$= 19.0$$

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

(0.44, 2.96)

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

5.3 Example - calculated by hand

	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)

Paired data (dependent sample)

- 作差，Equivalent to one-sample case
- 和独立变量的区别应该是不用考虑sp（ry

eg

5.4 Example

- In a study, participants attend clinical examinations approximately every four years. Suppose we want to compare systolic blood pressures between examinations (i.e., changes over 4 years).
- The data below are systolic blood pressures measured at the sixth and seventh examinations in a subsample of n=15 randomly selected participants.
- Since the data in the two samples (examination 6 and 7) are matched, we compute **difference scores by subtracting the blood pressure measured** at examination 7 from that measured at examination 6 or vice versa.
- If we subtract the blood pressure measured at examination 6 from that measured at examination 7, then positive differences represent increases over time and negative differences represent decreases over time.

- We now estimate the **mean difference in blood pressures** over 4 years.

Subject #	Examination 6	Examination 7	Difference
1	168	141	-27
2	111	119	8
3	139	122	-17
4	127	127	0
5	155	125	-30
6	115	123	8
7	125	113	-12
8	123	106	-17
9	130	131	1
10	137	142	5
11	130	131	1
12	129	135	6
13	112	119	7
14	141	130	-11
15	122	121	-1

Equivalent to one-sample case

- We now estimate the **mean difference in blood pressures** over 4 years.

$$X_d = \frac{\Sigma X}{n} = \frac{-79.0}{15} = -5.3$$

$$s_d = \sqrt{\frac{\Sigma(\text{Differences}-\bar{X}_d)^2}{n-1}} = \sqrt{\frac{2296.95}{14}} = \sqrt{164.07} = -12.8$$

$$\bar{X}_d = \pm t \frac{s_d}{\sqrt{n}}$$

$$-5.3 \pm 2.145 \frac{12.8}{\sqrt{15}} = -5.3 \pm 2.145\,(3.3) = -5.3 \pm 7.1$$

We are 95% confident that the mean difference in systolic blood pressures between examinations 6 and 7 (approximately 4 years apart) is between -12.4 and 1.8. The null (or no effect) value of the CI for the mean difference is zero. Therefore, based on the 95% confidence interval we can conclude that there is no statistically significant difference in blood pressures over time, because the confidence interval for the mean difference includes zero.

review

Review

- Types of parametric estimation? How to evaluate the estimations? How to interpret the CI?
- **Estimating one-sample proportion**: What is the procedure? What is the sampling distribution of sample proportion? Assumptions to check?
- **Estimating one-sample mean**: assumptions to check? What is the sampling distribution of statistics?
- **Estimating two-sample proportions**: conditions to check? What is the sampling distribution of statistics?
- **Estimating two-sample means**: conditions to check? What is the sampling distribution of statistics?