生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

授课：林关宁

2022 春季

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. 大数定理 & 中心极限定理			2
3.9			4. 常用概率分布			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 (10%)	测试2 (10%)	2
3.23			6. 统计推断基础-2：假设检验，I及II类错误，统计量，p-值			2
3.30			7. 数值数据的均值比较-1： 单样本及双样本t-检验，效应量，功效			2
4.6			8. 数值数据的均值比较-2：One-Way ANOVA，正态性检验			2
4.13			9. 数值数据的均值比较-3：Two-Way ANOVA			2
4.20	3	比例差异检验	10. 样本和置信区间预估 *	作业3 (6%)	测试3 (6%)	2
4.27			11. 类别数据的比例比较-1：联立表的卡方检验			2
➡ 5.11			13. 类别数据的比例比较-2： 联立表的RR，OR			2
5.18	4	协方差，相关分析，回归分析	14. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 (6%)	测试4 (6%)	2
5.25			15. 简单回归分析			2
6.1			16. 多元回归Multiple Regression			2
	5	Course Summary	17. 课程总结 *			2
			Total	30%	30%	32

*** 随堂测试**

双向列联表

以下双向列联表，例如要研究吸烟和肺癌之间的关系，行变量为是否吸烟：吸烟、不吸烟，列变量为肺癌发病：发病，不发病，如下表：

	患肺癌	未患肺癌	合计
吸烟	60	32	92
不吸烟	3	11	14
合计	63	43	106

对于这种数据，我们的统计目的是分析行列变量的**独立性**，即：肺癌发病是否与吸烟有关，可选用的方法有以下两种：

1、Pearson Chi-square Test 卡方检验：基于卡方分布，H_0为行、列变量相互独立。
四格表 使用条件：①样本总数大于40；②所有单元格理论值≥5。

2、Fisher's Exact Test 精确概率：
基于超几何分布，当数据不满足Pearson卡方检验时使用。

Here is a summary of the properties of the two tests:

Criterion	Chi-squared test	Fisher's exact test
Minimal sample size	Large	Small
Accuracy	Approximate	Exact
Contingency table	Arbitrary dimension	Usually 2x2
In		

Generally, Fisher's exact test is preferable to the chi-squared test because it is an exact test. The chi-squared test should be particularly avoided if there are few observations (e.g. less than 10) for individual cells. Since Fisher's exact test may be computationally infeasible for large sample sizes and the accuracy of the χ^2 test increases with larger number of samples, the χ^2 test is a suitable replacement in this case. Another advantage of the χ^2 test is that it is more suitable for contingency tables whose dimensionality exceeds 2×2.

Measures of association in two-by-two tables

Measures of disease frequency

- Ratios 比率
- Proportions 比例
- Prevalence流行率, incidence 发病率
- risks, rates, odds 风险

all functions of numerators (cases) and denominators (population at risk or those at risk but disease free)

函数：分子（病例）和分母（高危人群或无病高危人群）

Measures of disease frequency

- 比率 ratios: the relative magnitudes of two quantities (usually expressed as a quotient) (A/B)

- 比值 proportions: a ratio that relates the part (the numerator) to the whole (the denominator) — numerator always part of the denominator (A/A+B)

Prevalence流行率 & incidence 发病率

Prevalence proportion ≡ proportion with the characteristic or condition at a particular time

患病率 ≡ 特定时间段里**有某种病**的人的比例

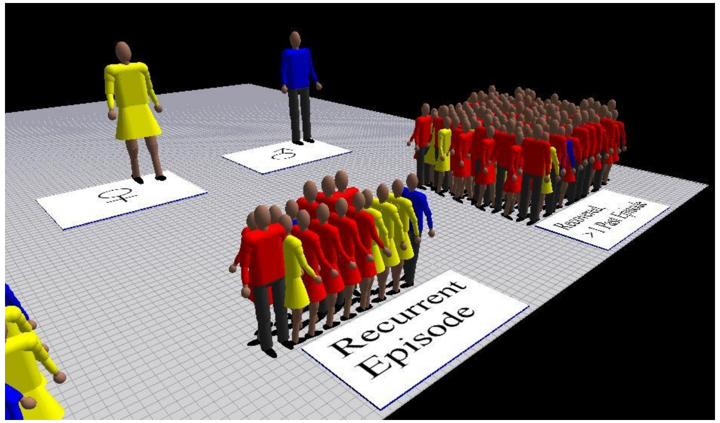
Incidence proportion ≡ proportion at risk that develop a condition over a specified period of time ≡ “average risk”

发病率 ≡ 在特定**时间段**内**从没病到产生疾病**的人的比例 ≡ “平均风险”

描述的是一种静止的状态

某一特定时间人群中该疾病的病例总数（现有病例）

=人口中的病例总数除以人群的个体数



描述的是一段时间里的快慢，变化，增减

在一个确定的时间段内，在一个风险的人群中出现的新病例总数

=新病例总数除以风险人群的总人数

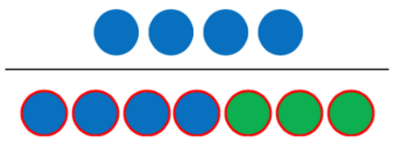
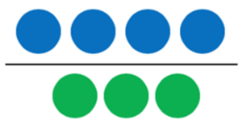
Odd & Risk of disease 得病几率 & 得病风险

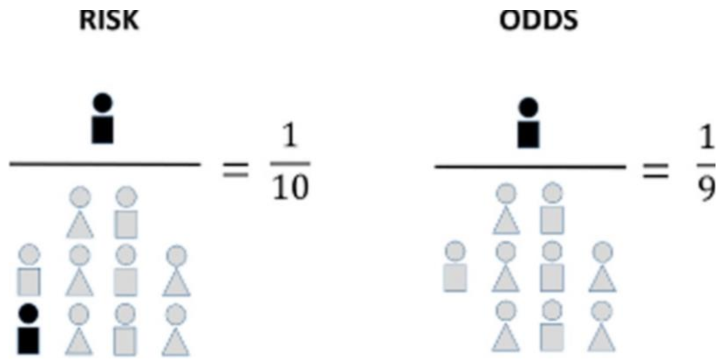
- Risk is number of events over number of possible events
- Odds is defined as the number of events to the number of non-events

Example: number of cases in exposed group 60, number of cases in unexposed group 10, odds are six to one (60/10) and risk is 86% (60/70)

The odds has properties that make it very useful in epidemiology

Probability vs Odds

	Risk	Odds
Mathematically	$P(p) = \frac{p}{p+q}$	$O(p) = \frac{\frac{p}{p+q}}{\frac{q}{p+q}} = \frac{p(p+q)}{q(p+q)} = \frac{p}{q}$
Graphically		



Odd Ratio (OR) & Risk Ratio/ Relative Risk (RR)

- 胜算比/优势比/比值比（odd ratio, OR)
  - Absolute (differences)

- 相对风险/风险比率（Relative risk, Risk ratio, RR)
  - Relative (ratios)

优势比 Odds ratio

	Number developed disease	Number disease-free	Total
Family history (exposed)	120	4880	5000
No family history (unexposed)	50	4950	5000
Total	170	9830	1000

	Case	Control
Exposed	a	b
Unexposed	c	d

Odds of a case being exposed (R_e) = a/b

Odds of a control being exposed (R_u)= c/d

Odds ratio = $R_{e/} R_u$ = (a/b)/(c/d) = ad/bc

Odds ratio = $R_{e/} R_u$

= (120/4880)/(50/4950)

= 2.4

Risk ratio 相对风险

	Number developed disease	Number disease-free	Total
Family history (exposed)	120	4880	5000
No family history (unexposed)	50	4950	5000
Total	170	9830	1000

	Case	Control
Exposed	a	b
Unexposed	c	d

Risk in exposed (R_e) = a/(a+b)

Risk in exposed (R_u)= c/(c+d)

Risk ratio = $R_{e/} R_u$

Risk ratio = $R_{e/} R_u$

= (120/5000)/(50/5000)

= 2.4

The following table summarizes the results of a 2012 study comparing NVP versus LPV in treatment of HIV-infected infants.[3] Children were randomized to receive either NVP or LPV.

	NVP	**LPV**	**Total**
Virologic Failure	60	27	87
Stable Disease	87	113	200
Total	147	140	287

The odds ratio in a 2×2 table 优势比

The **odds ratio (OR)** is a measure of the odds of a certain event occurring in one group relative to the risk of the event occurring in another group.

The odds of virologic failure among the NVP group is

$$\frac{\#\text{ in NVP group and had virologic failure}}{\#\text{ in NVP group and did not have virologic failure}} = \frac{60}{87} = 0.690$$

The odds of virologic failure among the LPV group is

$$\frac{\#\text{ in LPV group and had virologic failure}}{\#\text{ in LPV group and did not have virologic failure}} = \frac{27}{113} = 0.239$$

Thus, the odds ratio of virologic failure comparing NVP to LPV is $0.690/0.239 = 2.89$.

- The odds of virologic failure when treated with NVP are almost three times as large as the odds of virologic falure when treated with LPV.

Relative risk in a 2×2 table 相对风险

The **relative risk (RR)** is a measure of the risk of a certain event occurring in one group relative to the risk of the event occuring in another group.

The risk of virologic failure among the NVP group is

$$\frac{\#\text{ in NVP group and had virologic failure}}{\text{total }\#\text{ in NVP group}} = \frac{60}{147} = 0.408$$

The risk of virologic failure among the LPV group is

$$\frac{\#\text{ in LPV group and had virologic failure}}{\text{total }\#\text{ in LPV group}} = \frac{27}{140} = 0.193$$

Thus, the relative risk of virologic failure comparing NVP to LPV is $0.408/0.193 = 2.11$.

- Children treated with NVP are estimated to be more than twice as likely to experience virologic failure.

Odd Ratio (OR) vs. Risk Ratio/ Relative Risk (RR)

Table 1

Intervention	Outcome		Total (a + b)	Risk (a/[a + b])	Odds (a/b)
	Death (a)	Survival (b)			
I	30	70	100	30/100=0.30	30/70=0.43
II	10	90	100	10/100=0.10	10/90=0.11
III	1	99	100	1/100=0.01	1/99=0.01

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4640017/

Table 1 shows the risk and odds for different event rates. As “a” decreases with respect to “b” (probability of outcome becomes less), the odds and risk are similar. For rare events (i.e., if “a” is small and “a + b” approaches “b”), $a/(a + b) \approx a/b$ and risk approximates odds.

Therefore, though “odds” does not represent true risk, its value is close to risk when the event rates are low (typically <10%)

效应量大小评估的经验法则

χ^2 卡方检验的效应量有3种计算方式：

- ❖ Phi (φ)：只能 2 × 2 列联表
- ❖ odds ratio (OR)：只能 2 × 2 列联表
- ❖ Cramer‘s V (V)：可以 r x c 的表格

$\varphi = \chi^2 * n$

$V = \chi^2 * n * df, df = (r - 1) * (c - 1)$

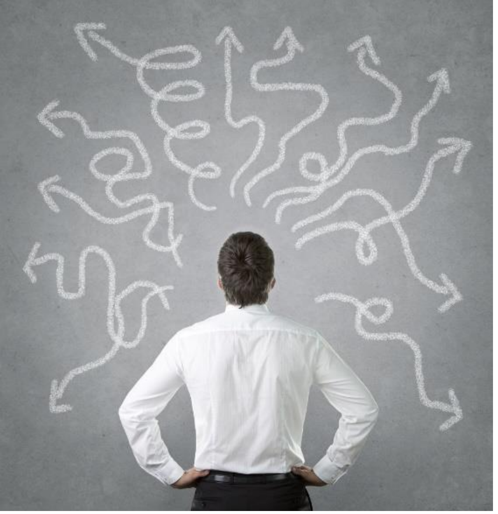
Exposure Status	Event Occurred	
	Yes	No
Exposed	a	b
Not Exposed	c	d

$$\text{Relative Risk} = \frac{a/(a+b)}{c/(c+d)}$$

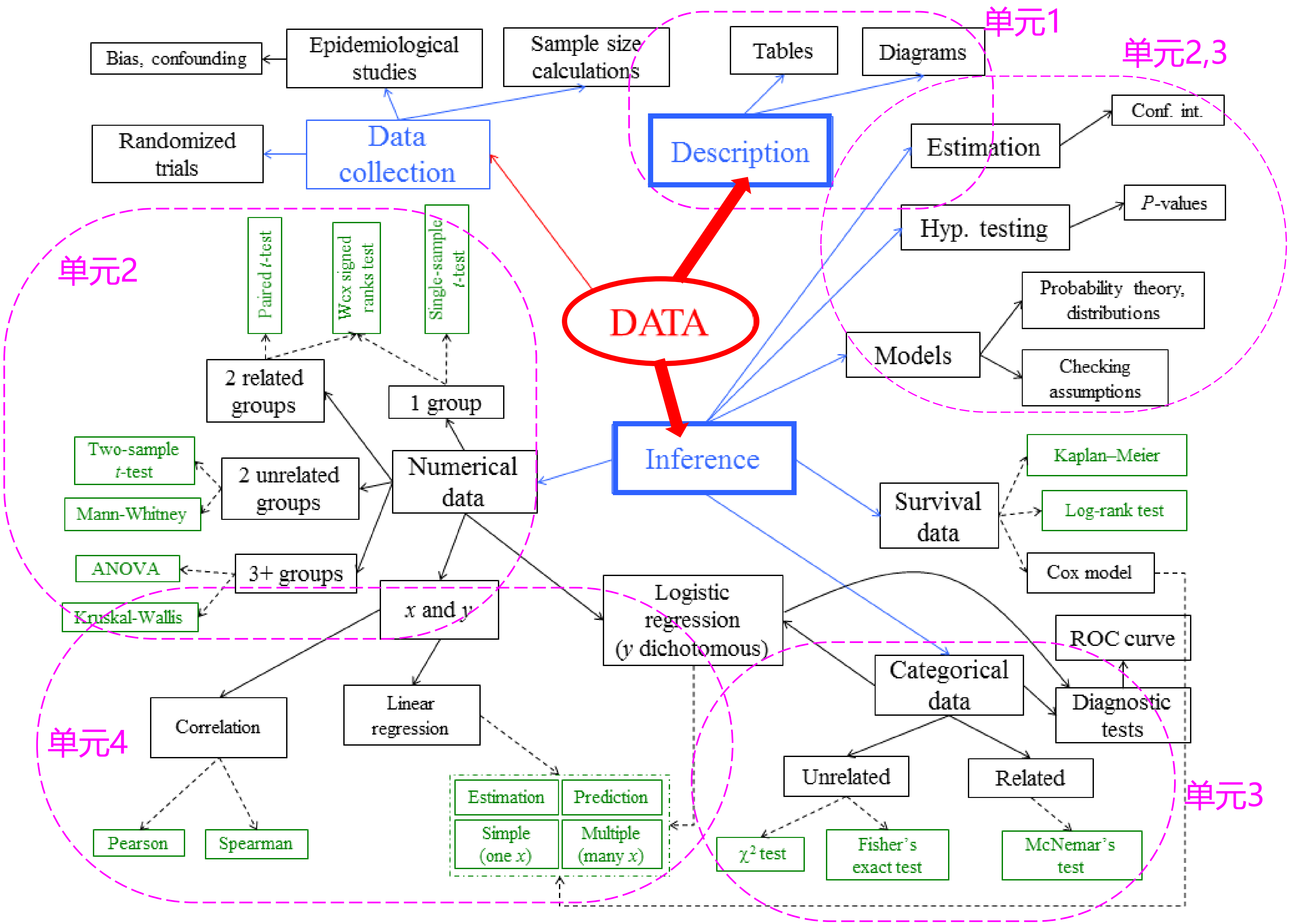
$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{cb}$$

Effect Size	Use	Small	Medium	Large
Correlation inc Phi		0.1	0.3	0.5
Cramer's V	r x c frequency tables	0.1 (Min(r-1,c-1)=1), 0.07 (Min(r-1,c-1)=2), 0.06 (Min(r-1,c-1)=3)	0.3 (Min(r-1,c-1)=1), 0.21 (Min(r-1,c-1)=2), 0.17 (Min(r-1,c-1)=3)	0.5 (Min(r-1,c-1)=1), 0.35(Min(r-1,c-1)=2), 0.29 (Min(r-1,c-1)=3)
Difference in arcsines	Comparing two proportions	0.2	0.5	0.8
η^2	Anova	0.01	0.06	0.14
omega-squared	Anova; See Field (2013)	0.01	0.06	0.14
Multivariate eta-squared	one-way MANOVA	0.01	0.06	0.14
Cohen's f	one-way an(c)ova (regression)	0.10	0.25	0.40
η^2	Multiple regression	0.02	0.13	0.26
κ^2	Mediation analysis	0.01	0.09	0.25
Cohen's f	Multiple Regression	0.14	0.39	0.59
Cohen's d	t-tests	0.2	0.5	0.8
Cohen's ω	chi-square	0.1	0.3	0.5
Odds Ratios	2 by 2 tables	1.5	3.5	9.0
Odds Ratios	p vs 0.5	0.55	0.65	0.75
Average Spearman rho	Friedman test	0.1	0.3	0.5

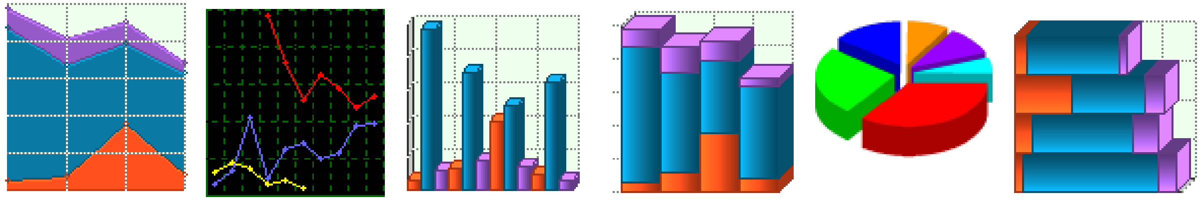
http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize



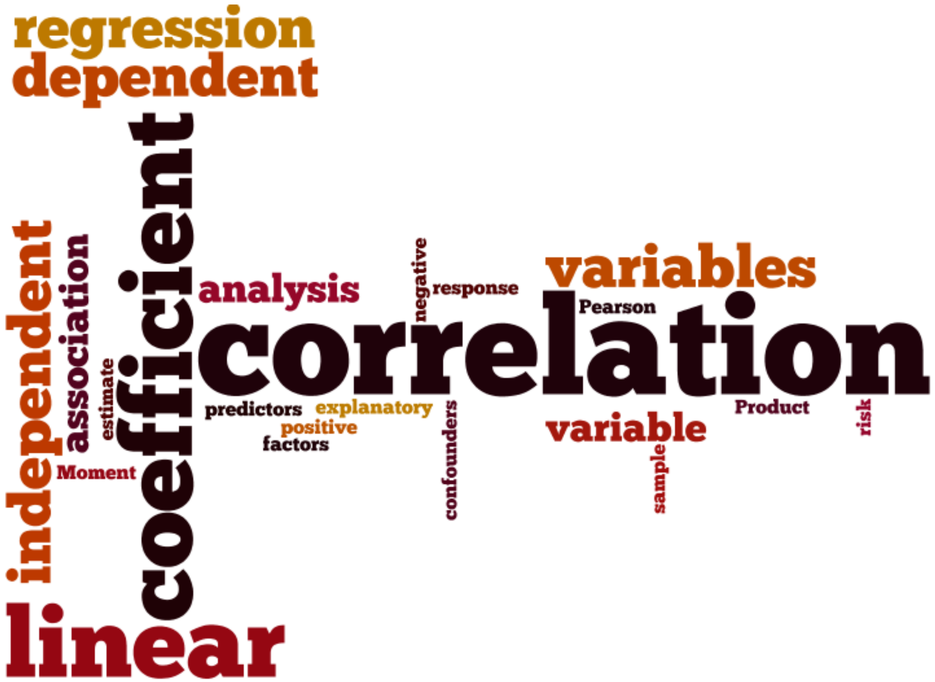
课程知识点导图



单元1
单元2,3
单元2
单元3
单元4
DATA
Data collection
Description
Inference
Epidemiological studies
Bias, confounding
Sample size calculations
Randomized trials
Tables
Diagrams
Estimation
Conf. int.
Hyp. testing
P-values
Models
Probability theory, distributions
Checking assumptions
Numerical data
Paired t-test
Wcx signed ranks test
Single-sample t-test
2 related groups
1 group
Two-sample t-test
Mann-Whitney
2 unrelated groups
ANOVA
Kruskal-Wallis
3+ groups
x and y
Correlation
Linear regression
Pearson
Spearman
Estimation
Prediction
Simple (one x)
Multiple (many x)
Logistic regression (y dichotomous)
Survival data
Kaplan–Meier
Log-rank test
Cox model
ROC curve
Categorical data
Diagnostic tests
Unrelated
Related
χ² test
Fisher's exact test
McNemar's test



Unit 4: Covariance, Correlation & Regression



回顾：离散趋势（单元2 Measure of dispersion）

- 均值（Mean）

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- 方差（Variance）

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- 标准差 (SD, STD)

$$\mathrm{S} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

1. 协方差（covariance）

- **用来度量两个随机变量关系的统计量**
  - 可以通俗的理解为：两个变量在变化过程中是同方向变化？还是反方向变化？同向或反向程度如何
    - 你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的
    - 你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的
    - 从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然

$$Cov(X,Y)=E[(X-\mu_x)(Y-\mu_y)]=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

简易版理解：
有X, Y两个变量，每个时刻的“X值和其均值之差”乘以“Y值和其均值之差”得到一个乘积后，再对这每时刻的乘积求和并求和再均值。

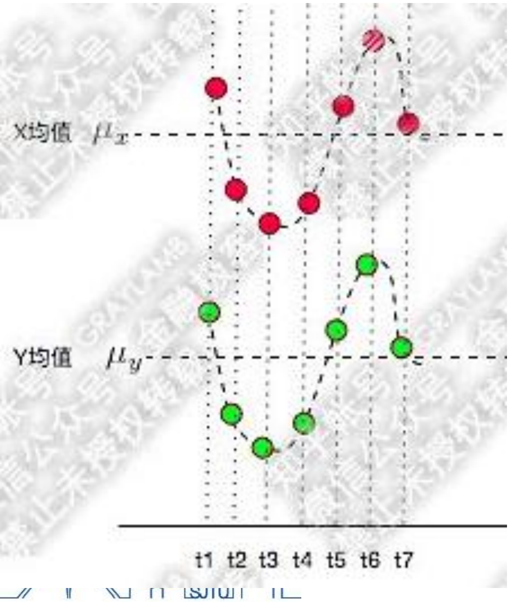
1. 协方差（covariance）

- **用来度量两个随机变量关系的统计量**
  - 可以通俗的理解为：两个变量在变化过程中是同方向变化？还是反方向变化？同向或反向程度如何
    - 你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的
    - 你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的
    - 从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然

$$Cov(X,Y)=E[(X-\mu_x)(Y-\mu_y)]=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

简易版理解：
有X, Y两个变量，每个时刻的“X值和其均值之差”乘以“Y值和其均值之差”得到一个乘积后，再对这每时刻的乘积求和并求和再均值。

举个例子：两个变量X,Y，观察 t1-t7（7个时刻）他们的变化情况



分别用红点和绿点表示X、Y，横轴是时间。可以看到X，Y均围绕各自的均值运动，并且很明显是同向变化的。

我们看到每一时刻 $(X-\mu_x)$ 的值与 $(Y-\mu_y)$ 的值的“正负号”一定相同。

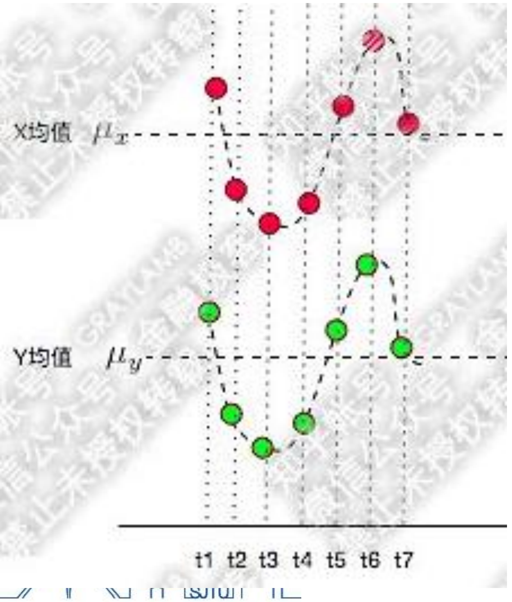
1. 协方差（covariance）

- **用来度量两个随机变量关系的统计量**
  - 可以通俗的理解为：两个变量在变化过程中是同方向变化？还是反方向变化？同向或反向程度如何
    - 你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的
    - 你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的
    - 从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

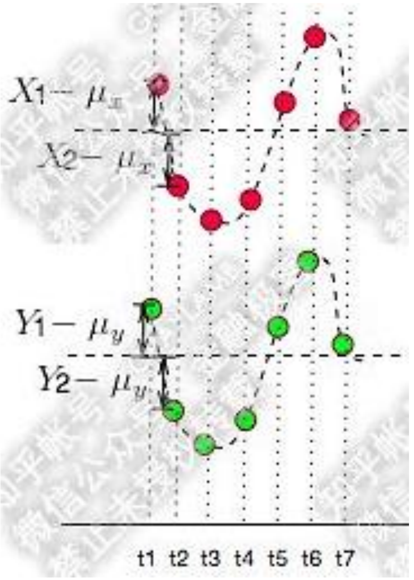
简易版理解：
有X, Y两个变量，每个时刻的“X值和其均值之差”乘以“Y值和其均值之差”得到一个乘积后，再对这每时刻的乘积求和并求和再均值。

举个例子：两个变量X,Y，观察 t1-t7（7个时刻）他们的变化情况



分别用红点和绿点表示X、Y，横轴是时间。可以看到X，Y均围绕各自的均值运动，并且很明显是同向变化的。

我们看到每一时刻 $(X-\mu_x)$ 的值与 $(Y-\mu_y)$ 的值的“正负号”一定相同。



比如t1时刻，他们同为正，t2时刻他们同为负

当他们同向变化时，$(X-\mu_x)$ 与 $(Y-\mu_y)$ 的乘积为正。这样，当你把 t1-t7 时刻 $(X-\mu_x)$ 与 $(Y-\mu_y)$ 的乘积加在一起，求平均也是**正数**了。

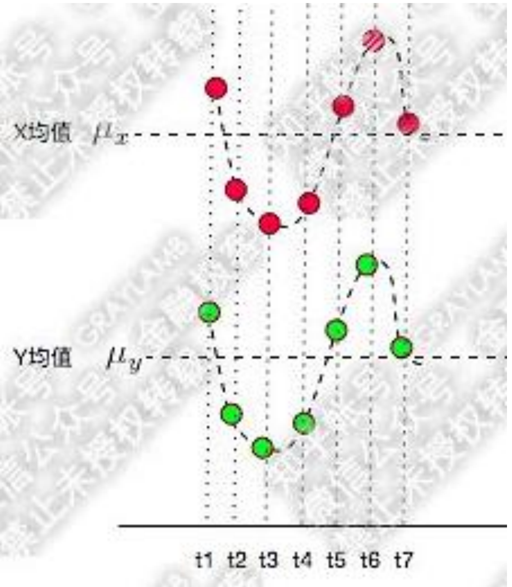
1. 协方差（covariance）

- **用来度量两个随机变量关系的统计量**
  - 可以通俗的理解为：两个变量在变化过程中是同方向变化？还是反方向变化？同向或反向程度如何
    - 你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的
    - 你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的
    - 从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

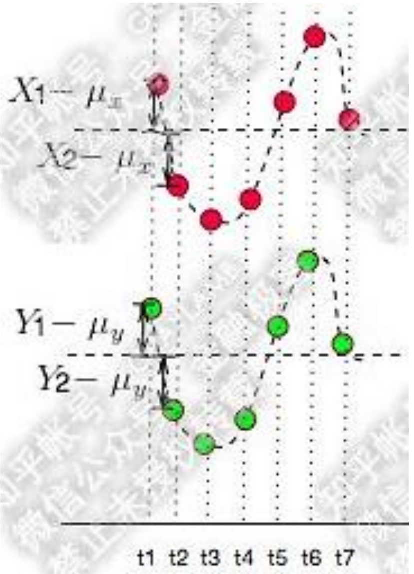
简易版理解：
有X, Y两个变量，每个时刻的“X值和其均值之差”乘以“Y值和其均值之差”得到一个乘积后，再对这每时刻的乘积求和并求和再均值。

举个例子：两个变量X,Y，观察 t1-t7（7个时刻）他们的变化情况



分别用红点和绿点表示X、Y，横轴是时间。可以看到X，Y均围绕各自的均值运动，并且很明显是同向变化的。

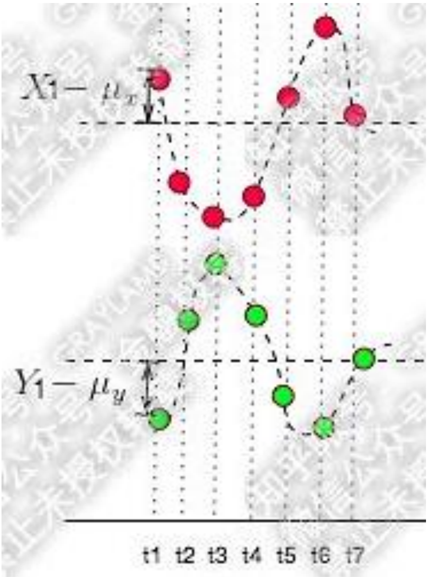
我们看到每一时刻 $(X - \mu_x)$ 的值与$(Y - \mu_y)$ 的值的“正负号”一定相同。



比如t1时刻，他们同为正，t2时刻他们同为负

当他们同向变化时，$(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积为正。这样，当你把 t1-t7 时刻 $(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积加在一起，求平均也是**正数**了。

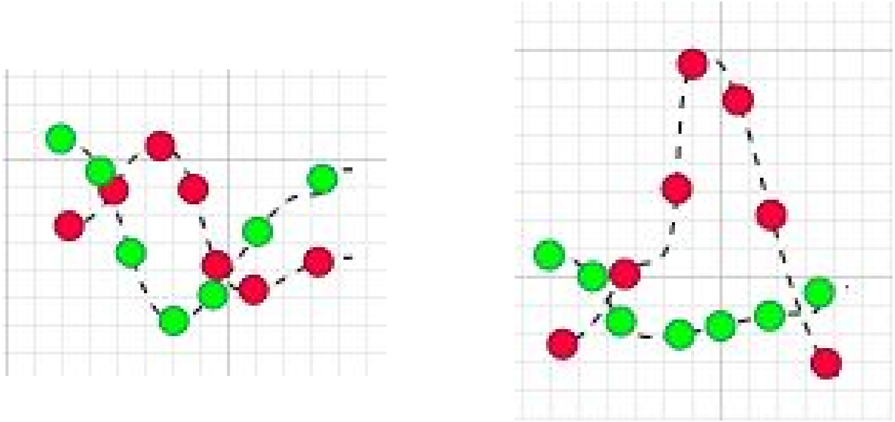
如果反向运动呢？



很明显，$(X - \mu_x)$ 的值与$(Y - \mu_y)$ 的值的“正负号”一定相反了，$(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积就是负值了。

这样，当你把 t1-t7 时刻$(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积加在一起，求平均也就是**负数**了。

但很多时候 X，Y 的运动是不规律的，比如：



这时， 很可能某一时$(X-\mu_x)$ 的值与$(Y-\mu_y)$ 的乘积为正，另一时刻$(X-\mu_x)$ 的值与$(Y-\mu_y)$ 的乘积为负。

将每一时刻 $(X-\mu_x)$ 与$(Y-\mu_y)$ 的乘积加在一起，其中的正负项就会抵消，最后求平均得出的就是**协方差**，然后通过协方差的数值大小，就可以判断这两个变量同向或反向的程度了。

所以，如果 例子里的 t1-t7时刻中， $(X-\mu_x)$ 与$(Y-\mu_y)$ 的乘积为正的越多，说明同向变化的次数越多，也即同向程度越高。反之亦然。

总结一下，如果协方差为正，说明X, Y 同向变化，协方差越大说明同向程度越高；如果协方差为负，说明X, Y 反向运动，协方差越小说明反向越高。

wait ...

那如果X，Y同向变化，但X大于均值，Y小于均值，那$(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积为负值啊？<u>这不是矛盾了吗？</u>

wait ...

那如果X，Y同向变化，但X大于均值，Y小于均值，那 $(X - \mu_x)$ 与$(Y - \mu_y)$ 的乘积为负值啊？这不是矛盾了吗？

再来，如果t1, t2, t3...t7 时刻X, Y 都在增大，而且X都比均值大，Y都比均值小，这种情况协方差不就是负的了？但是X, Y都是增大的，都是同向变化的，这又矛盾了？

这个怎么解释呢？

Python for covariance

numpy.cov

```
def cov(a, b):

    if len(a) != len(b):
        return

    a_mean = np.mean(a)
    b_mean = np.mean(b)

    sum = 0

    for i in range(0, len(a)):
        sum += ((a[i] - a_mean) * (b[i] - b_mean))

    return sum/(len(a)-1)
```

numpy.cov¶

numpy.cov(*m, y=None, rowvar=True, bias=False, ddof=None, fweights=None, aweights=None*) [source]

Estimate a covariance matrix, given data and weights.

Covariance indicates the level to which two variables vary together. If we examine N-dimensional samples, $X = [x_1, x_2, ... x_N]^T$, then the covariance matrix element C_{ij} is the covariance of x_i and x_j. The element C_{ii} is the variance of x_i.

See the notes for an outline of the algorithm.

Parameters:

m : *array_like*
A 1-D or 2-D array containing multiple variables and observations. Each row of *m* represents a variable, and each column a single observation of all those variables. Also see *rowvar* below.

y : *array_like, optional*
An additional set of variables and observations. *y* has the same form as that of *m*.

rowvar : *bool, optional*
If *rowvar* is True (default), then each row represents a variable, with observations in the columns. Otherwise, the relationship is transposed: each column represents a variable, while the rows contain observations.

bias : *bool, optional*
Default normalization (False) is by `(N - 1)`, where `N` is the number of observations given (unbiased estimate). If *bias* is True, then normalization is by `N`. These values can be overridden by using the keyword `ddof` in numpy versions >= 1.5.

ddof : *int, optional*
If not `None` the default value implied by *bias* is overridden. Note that `ddof=1` will return the unbiased estimate, even if both *fweights* and *aweights* are specified, and `ddof=0` will return the simple average. See the notes for the details. The default value is `None`.
New in version 1.5.

fweights : *array_like, int, optional*
1-D array of integer frequency weights; the number of times each observation vector should be repeated.
New in version 1.10.

aweights : *array_like, optional*
1-D array of observation vector weights. These relative weights are typically large for observations considered "important" and smaller for observations considered less "important". If `ddof=0` the array of weights can be used to assign probabilities to observation vectors.
New in version 1.10.

Returns:

out : *ndarray*
The covariance matrix of the variables.

2. 相关系数（correlation coefficient）

- 相关系数的公式为：

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]\,Var[Y]}}$$

  - ***X*，*Y* 的协方差除以 *X* 的标准差和 *Y* 的标准差**

Population versus Sample

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

- 所以，相关系数也可以看成：**一种**剔除了两个变量量纲影响、标准化后的**特殊协方差（如同变异系数是标准化后的标准差）**
- 是一种标准化后的协方差
  - 可以反映两个变量变化时是同向还是反向，如果同向变化就为正，反向变化就为负
  - 由于它是标准化后的协方差，因此它还有个重要的特性：**它消除了两个变量变化幅度的影响，只是单纯反映了两个变量每个单位变化时的相似程度，这样不同的实验之间就可以进行比较了**
  - 取值在 -1 到 1 之间
  - 通常绝对值大于0.7时认为两变量之间表现出非常强的相关关系，绝对值大于0.4时认为有着强相关关系，绝对值小于0.2时相关关系较弱。

举个例子：还是用之前的例子，变量X、Y变化的示意图（X为红点，Y为绿点），来看两种情况：

很容易可以看到图一，图二两种情况下的，X，Y都是同向变化的，而这个“同向变化”，有个显著特征：X, Y同向变化的过程，具有极高的相似度！无论是在图一还是图二的情况下，都是
t1 时刻X, Y 都大于均值，
t2 时刻都变小且小于均值，
t3 时刻X, Y 继续变小且小于均值，
t4 时刻X, Y 变大但仍小于均值，
t5 时刻X, Y 变大且大于均值。。。

可是，计算下协方差：

第一种情况下：

$$[(100-0)\times(70-0)+(-100-0)\times(-70-0)+(-200-0)\times(-200-0)\ldots]\operatorname{div}7\approx 15428.57$$

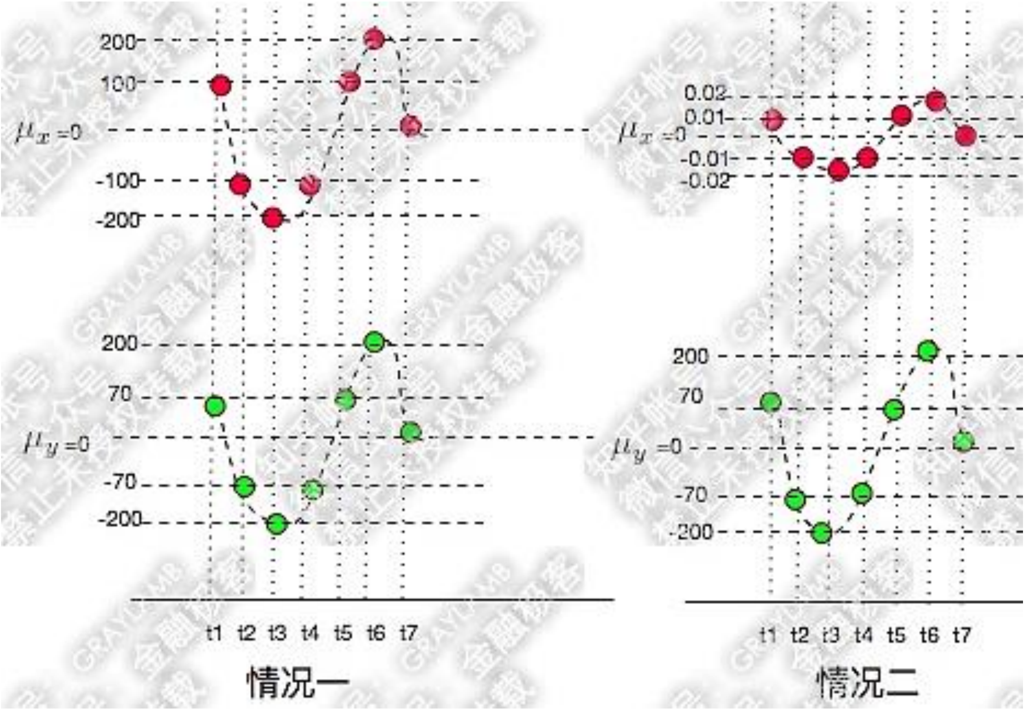
第二种情况下：

$$[(0.01-0)\times(70-0)+(-0.01-0)\times(-70-0)+(-0.02-0)\times(-200-0)\ldots]\operatorname{div}7\approx 1.542857$$

协方差差了一万倍，只能从2个协方差都是正数来判断这两种情况下的X, Y 都是同向变化，但是无法看出两种情况下X, Y 的变化是否具有相似性。

所以，为了能准确的研究两个变量在变化过程中的相似度，我们需要把变化幅度对协方差的影响，从协方差中剔除掉。于是，就有了相关系数的公式了

$$\rho=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$



第一种情况下：

X的标准差为

$$\sigma_X=\sqrt{E((X-\mu_x)^2)}=\sqrt{[(100-0)^2+(-100-0)^2\ldots]\operatorname{div}7}\approx 130.9307$$

Y的标准差为

$$\sigma_Y=\sqrt{E((Y-\mu_y)^2)}=\sqrt{[(70-0)^2+(-70-0)^2\ldots]\operatorname{div}7}\approx 119.2836$$

于是相关系数为

$$\rho=15428.57\div(130.9307\times 119.2836)\approx 0.9879$$

第二种情况下：

X的标准差为

$$\sigma_X=\sqrt{E((X-\mu_x)^2)}=\sqrt{[(0.01-0)^2+(-0.01-0)^2\ldots]\operatorname{div}7}\approx 0.01309307$$

Y的标准差为

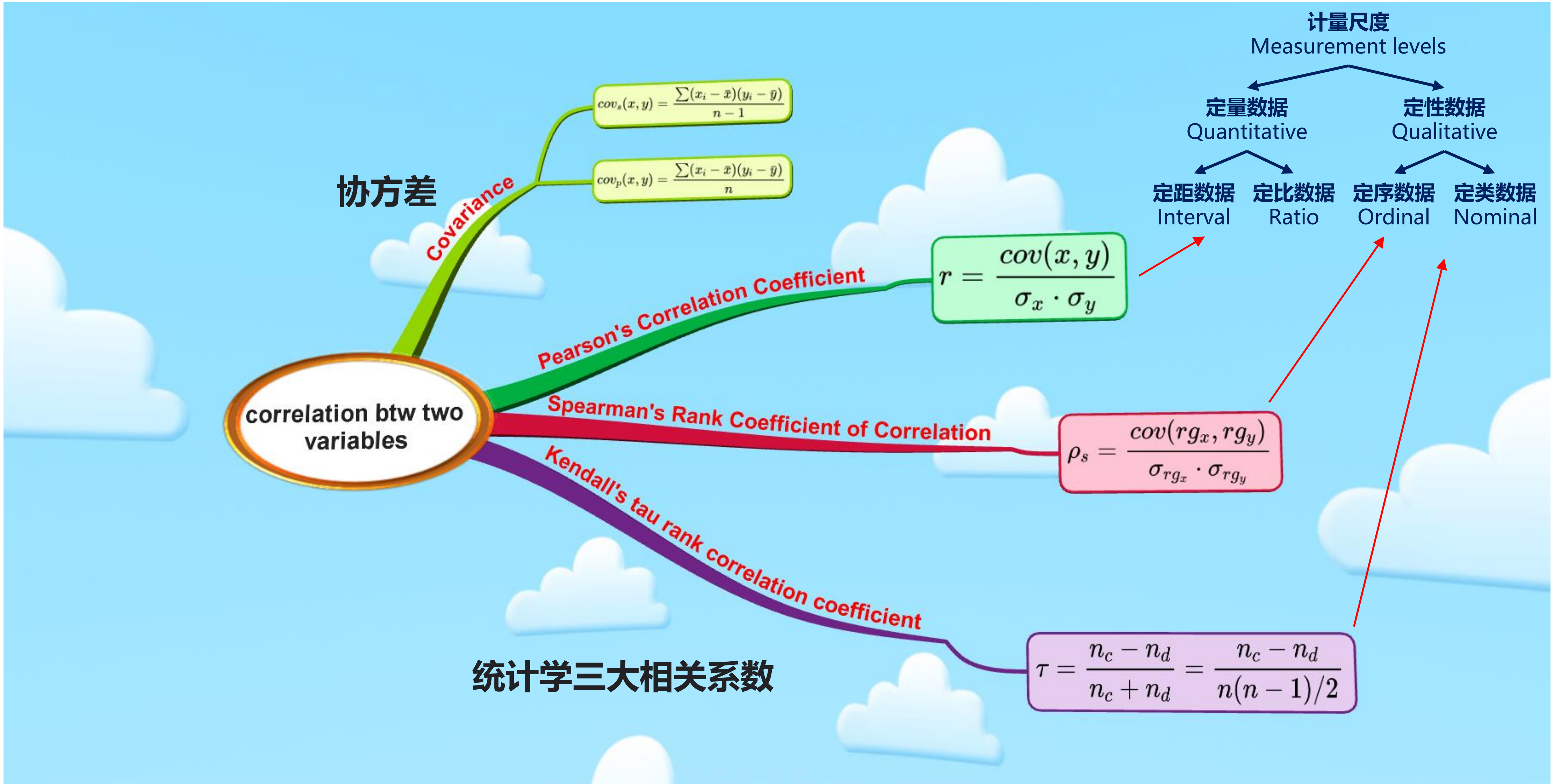
$$\sigma_Y=\sqrt{E((Y-\mu_y)^2)}=\sqrt{[(70-0)^2+(-70-0)^2\ldots]\operatorname{div}7}\approx 119.2836$$

于是相关系数为

$$\rho=1.542857\operatorname{div}(0.01309307\times 119.2836)\approx 0.9879$$

说明第二种情况下，虽然X的变化幅度比第一张情况X的变化幅度小了10000倍，但是丝毫没有改变“X的变化与Y的变化具有很高的相似度”这个结论。同时这两种情况的相关系数相等，说明有着一样的相似度。

Describing the correlation between two variables



谢谢，下周见！