假设检验 与 Effect Size

李宁

2020.5

1. 假设检验：H0假设【零假设，null hypothesis】与H1假设【备择假设，alternative hypothesis】

通常：H0是期望被推翻的结论；H1是期望得到的结论

- 假设检验：用于检验统计假设（属于统计学）的一种方法。
  - **假设的概念：** H1假设是研究中某种结论；H0假设是与H1假设相反的假设。
    例如：若某研究希望证明算法A和B的性能相同，即 可以令H1: A=B， 则可以 令 H0：A≠B；
  - **假设检验基本思想**：小概率反证法，通过否认H0（拒绝H0）证明H1。通常先提出H0假设，再用适当的统计方法确定H0成立的可能性大小。**如可能性小，则认为H0不成立(拒绝H0，即H1假设成立);** 但若可能性大，**则不能推翻**H0假设。
    (思想来源：全称命题只能被否认而不能被证明)
  - **常见假设检验方法：** t检验，Z检验，卡方检验，F检验等

H0: 没有充分理由不能轻易否定的命题。当发生概率很小时（近似不发生），认为可以拒绝，即证明H0不成立，H1成立。
H1: 没有把握不能轻易肯定的命题。

Parametric test：正态分布已知。
Non-parametric test：分布未知。

1. 假设检验

标记	均值	方差
样本	X	S
总体	μ	σ

- 样本与总体

由于实验是基于抽样样本进行，样本不可能绝对否定H0假设，但可以用小概率事件 ≈ 不可能发生的事件，所以 可以说：在一个约定的小概率水平（心目中的显著性水平 α，常见值：0.05或者0.01）上拒绝H0。

例如：若某研究希望证明算法A和B的性能相同，即 H1: A=B， 则可以 令 H0: A≠B；

如果统计数据表明 A≠B被误判为A=B的概率为P，P < 5% （概率很小，基本不发生，拒绝H0假设），

则 相当于 A=B （H1假设成立）。

- P值是什么?

假设检验的显著性水平： 第一类错误的最大概率α称为显著性水平。

p值：拒绝原假设H0的最小显著性水平称为检验的p值。

p值越小, 表示原假设越可疑, 从而越应拒绝原假设.

两类错误：

1. “弃真”：第一类错误, 当H0为真，却拒绝了H0的概率 P(拒绝H0|H0为真) = α 。
2. “取伪”：第二类错误, 当H0 不真, 但却接受了H0的概率，P(接受H0|H0为假) = β。

客观看待P值：https://www.guokr.com/article/438043/（统计学里“P”的故事）

1. 假设检验：H0假设【零假设，null hypothesis】与H1假设【备择假设，alternative hypothesis】

● 假设检验：举例

某车间用一台包装机包装精盐，额定标准每袋净重500g，设包装机包装出的盐每袋净重X服从正态分布。某天随机地抽取9袋，称得净重为490, 506, 508, 502, 498, 511, 510, 515, 512. 问该包装机工作是否正常。

t检验

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

假设：H0 hypothesis: true mean is equal to 500
H1 hypothesis: true mean is not equal to 500

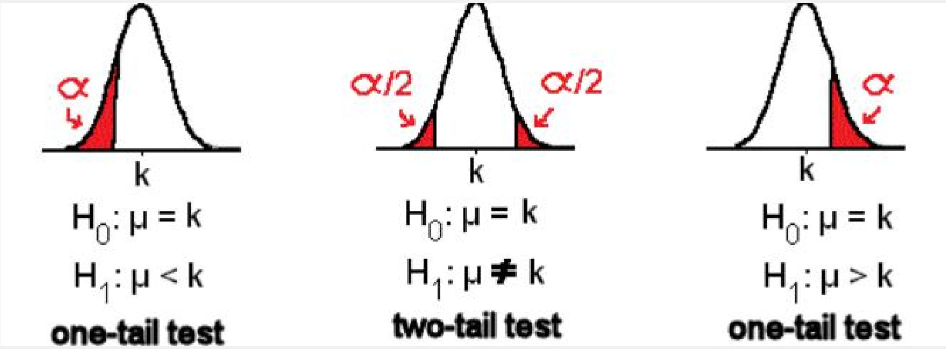
```
> salt<-c(490, 506, 508, 502, 498, 511, 510, 515 , 512)
> t.test(salt, mu=500)

        One Sample t-test

data:  salt
t = 2.1979, df = 8, p-value = 0.05919
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 499.7158 511.8398
sample estimates:
mean of x
 505.7778
```

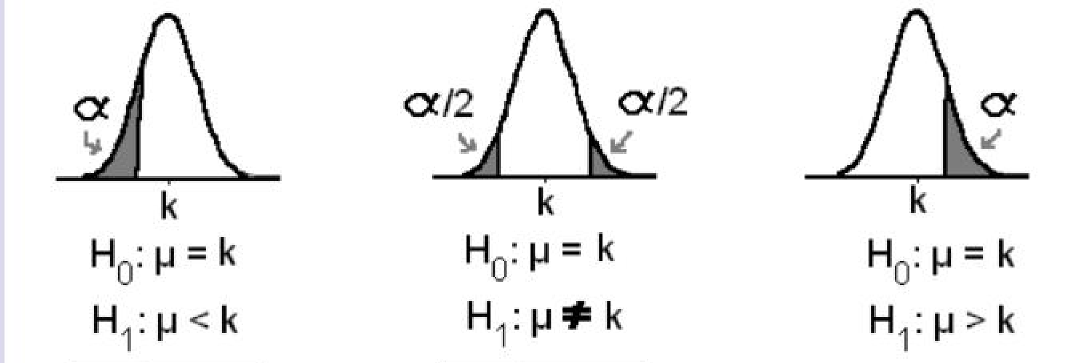
p>0.05 => 接受原假设，认为包装机正常

95%置信区间： 100次抽样中，在区间[l,u]中有95次包含真值，5次不包含真值。
错误表述：真值以95%的概率落在某个区间[l,u]之间；原因：总体的真值 是一个常数，不是一个随机变量，因此不存在多大概率属于某个区间。

1. 假设检验

双侧检验（two-tail）：H0: μ = μ0 ； H1: μ ≠ μ0
单侧检验（one-tail）：H0: μ ≥ μ0 ； H1: μ ＜ μ0 或 H0: μ ≤ μ0 ； H1: μ ＞ μ0





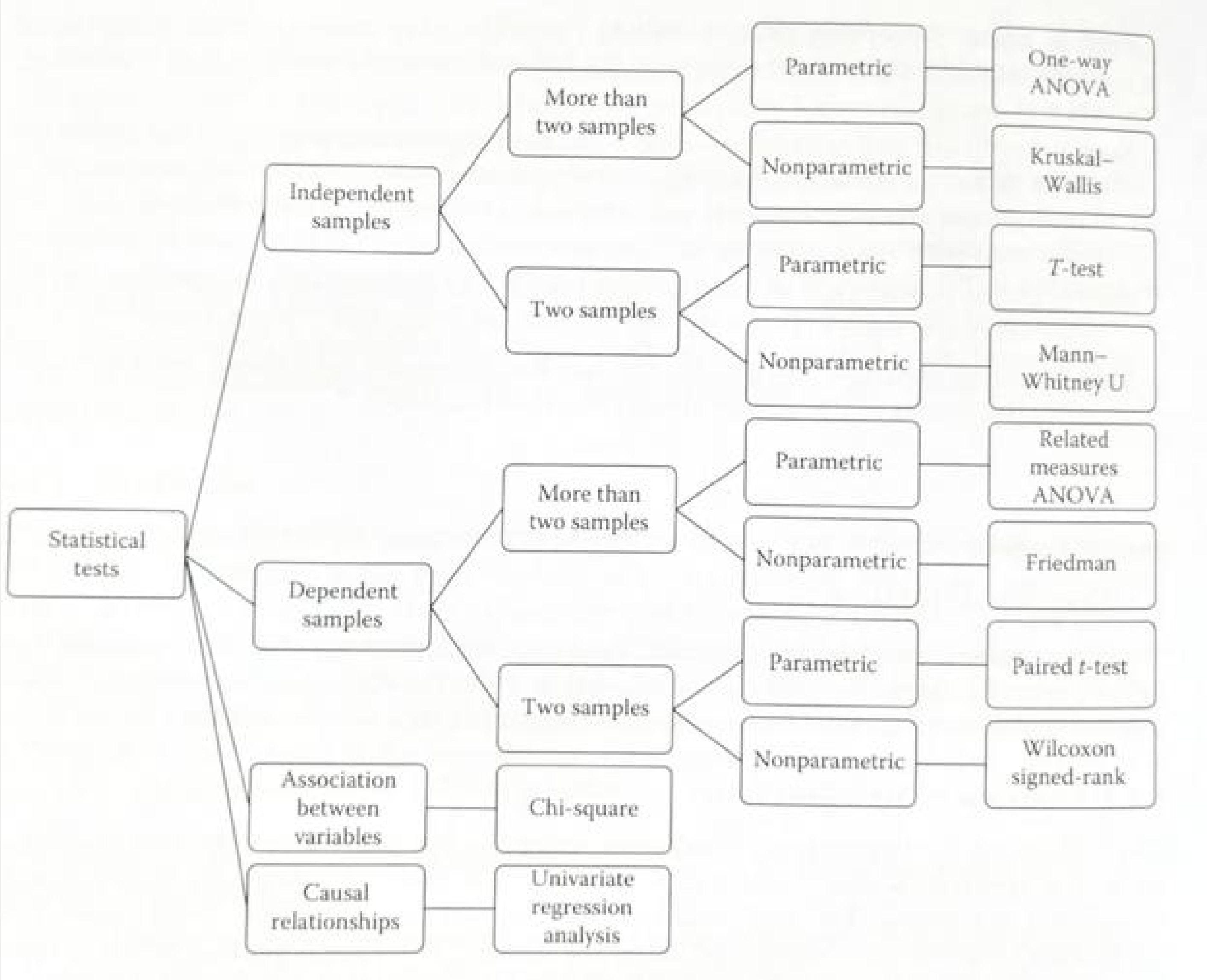
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33

α	z critical
0.10	±1.65
0.05	±1.96
0.01	±2.58

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

95%置信区间： 100次抽样中，有95次抽找到的区间包含真值，5次找到的区间不包含真值。
错误表述：真值以95%的概率落在某个区间[l,u]之间；
（原因：总体的真值 是一个常数，不是一个随机变量，因此不存在多大概率属于某个区间）

1. 假设检验 - 常见检验总结



1. 假设检验 - 常见检验总结

	参数的假设检验（服从正态分布）			非参数的假设检验
	均值(数据中心点)检验	方差(数据分散度)检验	比例检验	距离 或者考虑符号
单组样本	方差已知：Z检验 z.test 方差未知：t检验 t.test	卡方检验：chisq.var.test	样本<=30，用精确检验binom.test 样本>30，用近似检验prop.test	卡方检验：chisq.test 二项式检验
两组样本	均值比较：t检验t.test (做均值检验时，必须先做方差齐性检验[F检验]，只有在方差齐的情况下，比较均值才是有意义的，否则没有可比性)	F检验	近似检验prop.test	wilcox.test秩和检验 或者 Mann-Whitney U检验，卡方检验
两组样本成对比较	t检验t.test, paired=true			wilcox.test秩和检验 或者 sign.test符号检验
多组样本	方差分析			Kruskal-wallis秩和检验Kruskal.test 卡方检验：chisq.var.test

参考：《经验软件工程：软件工程中的实验方法》，张莉等译

1. 假设检验 - 基本步骤

通常进行假设检验的步骤:

1) 提出原假设H0与备择假设H1;

2) 选择检验统计量W并确定其分布;

- 数据满足正态分布：参数检验
- 数据非正态分布：转换为正态分布或者非参检验

3) 在给定的显著性水平下, 确定H0关于统计量W的拒绝域;

4) 算出样本点对应的检验统计量的值;

5) 判断: 若统计量的值落在拒绝域内, 则拒绝H0, 否则接受H0.

2. 假设检验 - Case Study1

```
> A <- c(0.85, 0.78,    0.99,   0.73,   0.84,   0.92,   0.95,   0.75,   0.91,   0.75)
> B <- c(0.81, 0.78,    0.99,   0.67,   0.85,   0.89,   0.91,   0.65,   0.92,   0.45)
> #Mann-Whitney-Wilcoxon Test
> wilcox.test(A, B, alternative = "two.sided", paired = TRUE, exact = FALSE, correct = FALSE)

        Wilcoxon signed rank test

data:  A and B
V = 33, p-value = 0.03524
alternative hypothesis: true location shift is not equal to 0
```

假设检验例子:

用两个算法预测10个模块中的缺陷，下表记录了对每个模块的预测召回率Recall:

模块	1	2	3	4	5	6	7	8	9	10
算法A	0.85	0.78	0.99	0.73	0.84	0.92	0.95	0.75	0.91	0.75
算法B	0.81	0.78	0.99	0.67	0.85	0.89	0.91	0.65	0.92	0.45

1) 提出原假设H0与备择假设H1;

H0: 算法A和算法B的缺陷召回率相同。

H1：算法A和算法B的缺陷召回率不同。

2）选择检验统计量W并确定其分布;

- 数据不是正态分布，所以考虑采用非参检验

- 成对数据

所以最终选择用：Mann-Whitney-Wilcoxon Test

3) 在给定的显著性水平下（0.05）, 确定H0关于统计量W的拒绝域;

4) 算出样本点对应的检验统计量的值;

5) 判断: 若统计量的值落在拒绝域内, 则拒绝H0, 否则接受H0.

拒绝H0，所以H1成立。

2. 假设检验 - Case Study2（T检验）

拒绝H0，所以H1成立。

假设检验例子:

两个项目中，不同模块的缺陷密度:

```
x <- c(3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 2.68, 4.30, 2.49, 1.54)
y <- c(3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49)
```

1) 提出原假设H0与备择假设H1;

H0：项目X与项目Y的缺陷密度相同。

H1：项目X与项目Y的缺陷密度不同。

2）选择检验统计量W并确定其分布;

- 进行正态性与等方差检验
- 成对数据

所以最终选择用：t Test

3) 在给定的显著性水平下（0.05），确定H0关于统计量W的拒绝域;

自由度：10+11-2=19， $\alpha=0.05$

对于双边检测，查表得 $t_{0.025,19} = 2.093$

4) 算出样本点对应的检验统计量的值;

$t_0 = -3.96$

5) 判断: 若统计量的值落在拒绝域内, 则拒绝H0, 否则接受H0.

$|t_0| > t_{0.025,19}$，落入拒绝域，所以拒绝H0

2. 假设检验 - Case Study2（T检验）

$$t - value = \frac{mean1 - mean2}{sp * \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

where,

$$sp = \sqrt{\frac{(n1-1) * var1^2 + (n2-1) * var2^2}{n1 + n2 - 2}}$$

where n1 and n2 are the number of records in each sample set.

and, ***Degrees of freedom = (n1 + n2 – 2)***

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	$\boxed{2.093}$	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850

```
> x <- c(3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 2.68, 4.30, 2.49, 1.54)
> y <- c(3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49)
> mean(x)
[1] 2.853
> mean(y)
[1] 4.105455
> n <- length(x);n
[1] 10
> m <- length(y);m
[1] 11
> Sp <- sqrt(((n-1)*var(x)+(m-1)*var(y))/(m+n-2));Sp
[1] 0.724303
> t0 <- (mean(x) - mean(y))/(Sp*(sqrt(1/n+1/m)));t0
[1] -3.957568
```

```
> t.test(x,y,alternative = "two.sided")

        Welch Two Sample t-test

data:  x and y
t = -3.9131, df = 17.203, p-value = 0.001098
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9271364 -0.5777727
sample estimates:
mean of x mean of y
 2.853000  4.105455
```

3. Effect Size （效应量）

- ES: 反映效应程度大小的统计量，代表变量之间的紧密或差异程度。

In statistical inference, an effect size is a measure of the strength of the relationship between two variables.

- 效应量太小，则意味着处理即使达到了显著水平，也缺乏实用价值。

P-values are designed to tell you if your result is a fluke, not if it’ s big. Because with a big enough sample size, any difference in means, no matter how small, can be statistically significant. So you also need to give some sort of effect size measure.

p是定性的，是看差异是否存在；效应量是定量的，看有多大的差异。

https://cebcp.org/practical-meta-analysis-effect-size-calculator/

3. Effect Size （效应量）

- 标准的Effect size应用场景

  - the metrics of variables being studied do not have intrinsic meaning (e.g., a score on a personality test on an arbitrary scale),
  - results from multiple studies are being combined,
  - some or all of the studies use different scales, or
  - it is desired to convey the size of an effect relative to the variability in the population.

3. Effect Size （效应量）

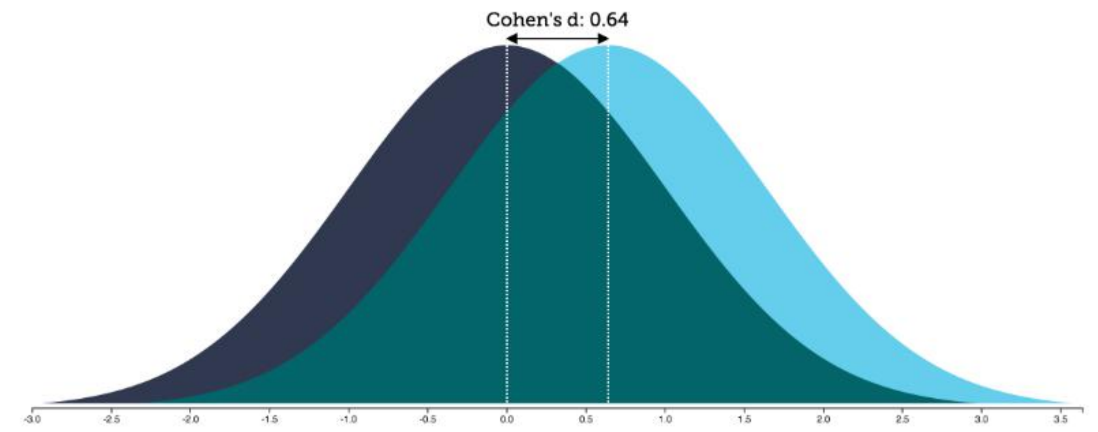
常见的Effect Size

① Standardized mean difference
- Cohen's d
- Hedge's g

② Cliff's Delta
- 适用场景：非参的ES比较
- 对异常值不敏感 (Robust)

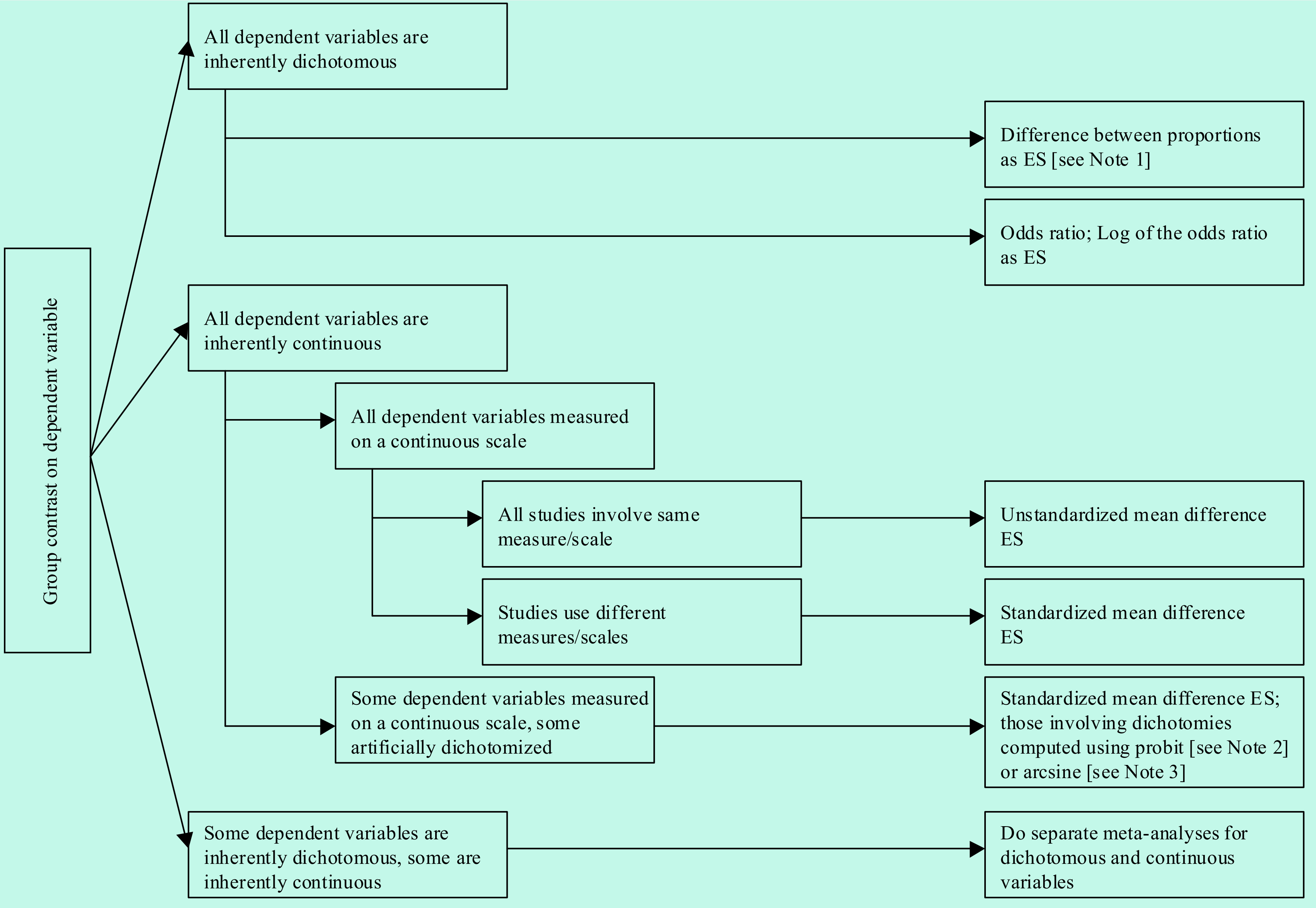
③ Correlation coefficient
- Correlation coefficient (相关系数，r)
- Coefficient of determination (决定系数，R or R2)

④ Odds-ratio
- 适用场景：研究问题关注 两个二元变量之间的关联程度



The visualization for cohen's d = 0.64

3. Effect Size （效应量）



3. Effect Size （效应量）

➤ Cohen's "Rules-of-Thumb"

- **standardized mean difference effect size**
  - small = 0.20
  - medium = 0.50
  - large = 0.80
- **correlation coefficient**
  - small = 0.10
  - medium = 0.25
  - large = 0.40
- **odds-ratio**
  - small = 1.50
  - medium = 2.50
  - large = 4.30

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

3. Effect Size （效应量）

标记	均值	方差
样本	X	S
总体	μ	σ

- Standardized mean difference

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}$$

Glass' s Δ

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{s_2}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2,$$

Cohen' s d

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where *s* can be calculated using this formula:

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

s: pooled standard deviation

Hedges' g

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

Where standard deviation can be calculated using this formula:

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

3. Effect Size （效应量）

标记	均值	方差
样本	X	S
总体	μ	σ

- μ 和 σ

Suppose that the entire population of interest was eight students in a particular class.

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16. \end{array}$$

The variance is the mean of these values:

$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{4} = 2.$$

3. Effect Size （效应量）

- Standardized mean difference (Cohen’s d)

1. Direction Calculation Method

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}$$

2. Algebraically Equivalent Formulas:

$$ES = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

independent t-test

$$ES = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}$$

two-group one-way ANOVA

3. Effect Size （效应量）

- Cliff's Delta or d: 一种非参ES，可用于两组数据比较。

  Only considers the ordinal instead of the interval properties of the data.

$$Delta = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{n_1 n_2}$$

n1, n2: 两个组各自的样本数

#(x1 > x2): X1 > X2，则为1；否则为0

Range of Cliff's Delta : [-1, +1]:

- ✓ 0.0 : group distributions overlap completely (没有差别)
- ✓ +1.0 or -1.0 : the absence of overlap between the two groups （差别很大）
- ✓ +: group1 > group2;
- ✓ -: group1 < group2

```
> treatment <- c(10,20,30,40,40,50)
> control <- c(10,20,30,40,40,50)
> cliff.delta(treatment,control, return.dm=TRUE)

Cliff's Delta

delta estimate: 0 (negligible)
95 percent confidence interval:
     lower      upper
-0.6110126  0.6110126
```

```
> treatment <- rep(6:10, 2)
> control <- rep(1:5, 2)
> treatment
 [1]  6  7  8  9 10  6  7  8  9 10
> control
 [1] 1 2 3 4 5 1 2 3 4 5
> cliff.delta(treatment,control, return.dm=TRUE)

Cliff's Delta

delta estimate: 1 (large)
95 percent confidence interval:
    lower     upper
0.9735068 1.0000000
Warning message:
In cliff.delta.default(treatment, control, return.dm = TRUE) :
  The samples are fully disjoint, using approximate Confidence Interval estimation
```

```
> treatment <- c(5,6,7,8,5)
> control <- c(1,2,3,4,5)
> treatment
[1] 5 6 7 8 5
> control
[1] 1 2 3 4 5
> cliff.delta(treatment,control, return.dm=TRUE)

Cliff's Delta

delta estimate: 0.92 (large)
95 percent confidence interval:
    lower     upper
0.5171092 0.9891504
```

3. Effect Size （效应量）

- Cliff's Delta or d: 一种非参ES，可用于两组数据比较。

Only considers the ordinal instead of the interval properties of the data.

$$Delta = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{n_1 n_2}$$

n1, n2: 两个组各自的样本数

#(x1 > x2): X1 > X2，则为1；否则为0

$$\delta_{ij} = \begin{cases} +1 \rightarrow Group1_i > Group2_j, \forall_i, \forall_j \\ -1 \rightarrow Group1_i < Group2_j, \forall_i, \forall_j \\ 0 \rightarrow Group1_i = Group2_j, \forall_i, \forall_j \end{cases}$$

$$Delta\ by\ rows = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta_{ij}$$

或

$$Delta\ by\ columns = \frac{1}{nm} \sum_{j=1}^{n} \sum_{i=1}^{m} \delta_{ij}$$

例题： 根据以下数据求d。

```
treatment <- c(5,6,7,8,5,6)
control   <- c(1,2,3,4,5)
```

与Mann–Whitney U test的统计量W值相同，只是Cliff's Delta体现了方向（正负号）

3. Effect Size （效应量）

● Odds-ratio

- The odds-ratio 通常都是基于以下二元变量的表格

	pass	fail
control group	2	1
treatment group	6	1

A study of spelling ability in exam

odds ratio
= (6/1) / (2/1)
= 6/2
= 3

The Odds-Ratio is the odds of success in the treatment group relative to the odds of success in the control group.

3. Effect Size （效应量）

● Correlation coefficient: r

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad (Eq.1)$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \operatorname{E}[(X-\mu_X)(Y-\mu_Y)],$$ [7]

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (Eq.2)$$

where:

- σ_Y and σ_X are defined as above
- μ_X is the mean of X
- μ_Y is the mean of Y
- E is the expectation.

Pearson's correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (Eq.3)$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

r 取值范围： [-1, +1]
−1 indicating a perfect negative linear relation,
 1 indicating a perfect positive linear relation,
 0 indicating no linear relation between two variables.

3. Effect Size （效应量）

- Coefficient of determination （referred to as R2 or "r-squared"）

例:

若 相关系数r 是 0.21，则决定系数 coefficient of determination 是 0.0441。

含义：4.4% of the variance of either variable is shared with the other variable.

THANKS