上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

SJTU School of Biomedical Engineering

生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

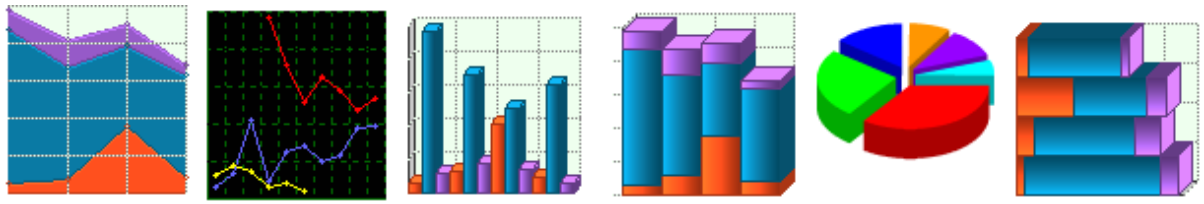
授课：林关宁

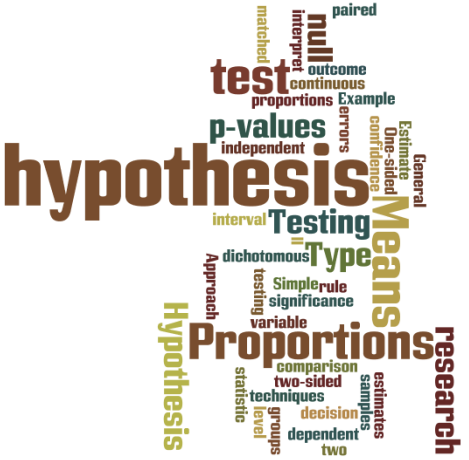
2022 春季

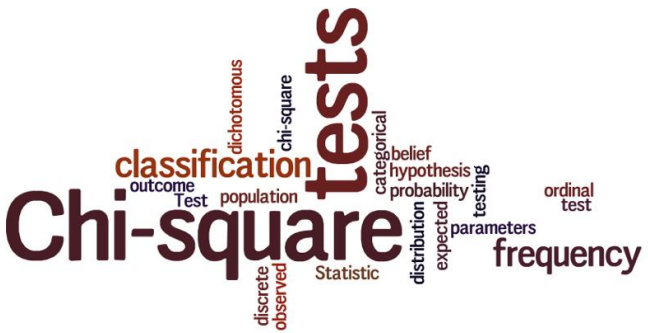
课程内容安排

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
5.7			12. 类别数据的比例比较-2： 联立表的RR，OR			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

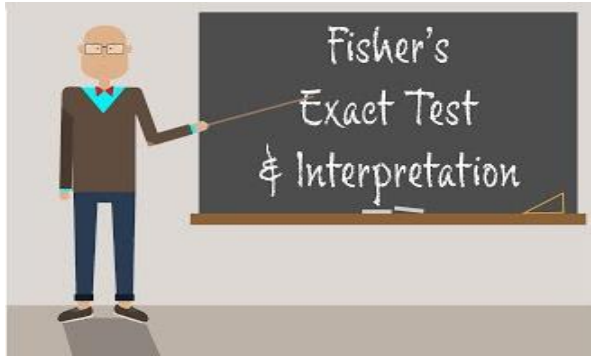






单元 3（week 10, 11, 12）

Inferential Statistics of Proportion Data



- **Sample proportion（样本的构成比）**
  - Confidence interval estimation for population proportion
  - Confidence interval estimation for difference between 2 populations
- **Contingency table（列联表）**
- **Chi-square test（卡方检验）**
  - Chi-Square Test for independence（2 variables, 1 population）
  - Chi-Square Test for goodness of fit（1 variable, 1 population）
- **Fisher's exact test（Fisher精确检验）**

（对应参考书第8章）

Confidence intervals for population proportions

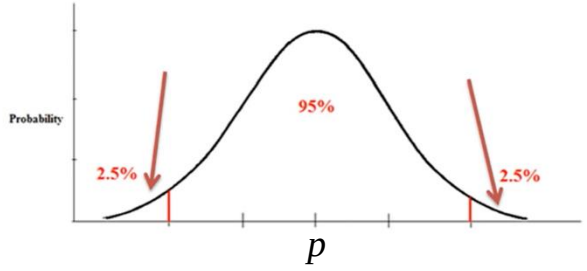
Confidence interval for Population proportion

- Each element in the population can be classified as a success or failure

  Sample proportion $\hat{p} = \dfrac{\text{number of successes}}{\text{sample size}} = \dfrac{x}{n}$
- Proportion always between 0 and 1
- For large samples the sample proportion $\hat{p}$ is approximately normal

$$CI(p)_{1-\alpha} = \left[\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



$\alpha = 0.05$ $\quad z_{0.05/2} = z_{0.025}$

Formula for Confidence Intervals and Conditions

Questions to Ask YOURSELF: How many populations?	Independent samples?	Parameter(s) of interest	Population variances?	Confidence interval	Condition(s) that MUST be satisfied. (Check to see if they are satisfied Before you use them.) These conditions also apply to corresponding test of hypothesis.
One	Not Applicable	μ	Known	$\left(\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right)$	Population standard deviation, σ_0 is **known** plus Normal population or large sample
			Unknown	$\left(\bar{X} \pm t_{(\alpha/2,\, n-1)} \cdot \frac{S}{\sqrt{n}}\right)$	Population standard deviation, σ is **unknown** plus (Almost) Normal population
		π	Unknown	$\left(p \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right)$	$np \geq 10$ and $n(1-p) \geq 10$
Two	Yes	$\mu_1 - \mu_2$	Known	$\left((\bar{X} - \bar{Y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_{X0}^2}{n_X} + \frac{\sigma_{Y0}^2}{n_Y}}\right)$	Population standard deviations, σ_{X0} and σ_{Y0} are **known** plus Normal populationS or large sampleS
	Yes		Unknown	$\left((\bar{X} - \bar{Y}) \pm t_{\alpha/2,\, df} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right)$ df = smaller of $(n_x - 1)$ and $(n_y - 1)$	Population standard deviations, σ_X and σ_Y are **unknown, unequal** plus (Almost) normal populations
	No			$\left(\bar{D} \pm t_{(\alpha/2,\, n-1)} \cdot \frac{S_D}{\sqrt{n}}\right)$	D's have a normal distribution
	Yes	$\pi_1 - \pi_2$		$\left((p_1 - p_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$	$n_1p_1 \geq 10$ **and** $n_1(1-p_1) \geq 10$ **and** $n_2p_2 \geq 10$ **and** $n_2(1-p_2) \geq 10$

One-variable Inference of Proportions

常见问题

(1) The CI of the Population
从一个样本的$\hat{p}$, 推断总体p

(2) How well does the sample fit an expected model/distribution?,

e.g. uniform/binominal/normal distribution
评价观察样本与总体期望的差异，是否符合目标分布

Hypothesis test for sample proportion （z-test）

$\hat{p}$ vs. p	confidence interval	hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Hypothesis testing for a single proportion:

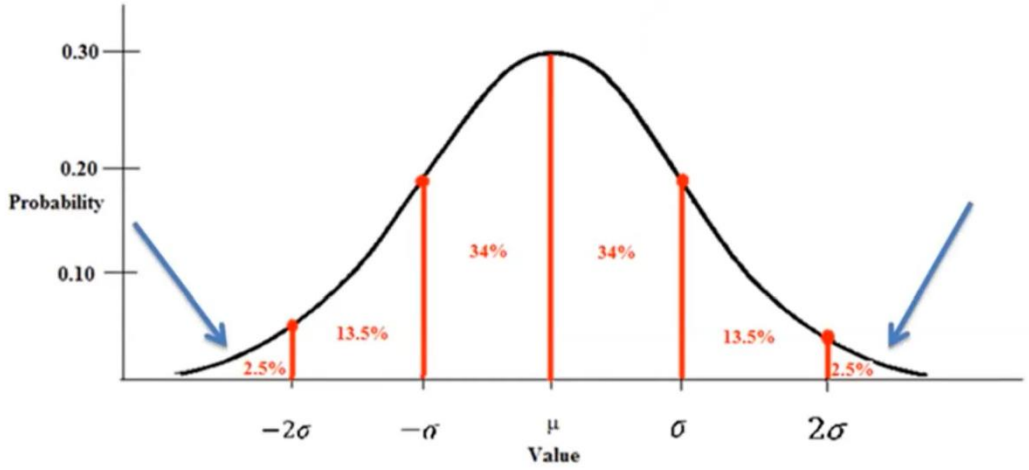
1. Set the hypotheses: $H_0 : p = null\ value$
   $H_A : p <\ or\ >\ or\ \neq\ null\ value$
2. Calculate the point estimate: $\hat{p}$
3. Check conditions: 当以下条件成立时，p^的抽样分布近似为正态分布：
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** $np \geq 10$ and $n(1-p) \geq 10$ 样本中预计至少有10次成功和10次失败
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value < α, reject H_0; the data provide convincing evidence for H_A.
   - If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.

如果H0被拒绝，只能说本样本不足以支持H0成立
如果H0不被拒绝，也并不能说明H0一定正确，只能说本数据不足以反驳H0

Hypothesis test for one sample proportion

Hypothesis test for sample proportion (z-test)

➢ One Sample z-Test for Proportions



A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is obtained. Of these 100 doctors, 82 indicate that they recommend aspirin. Is this claim accurate? Use alpha = 0.05

1. Define Null and Alternative Hypotheses

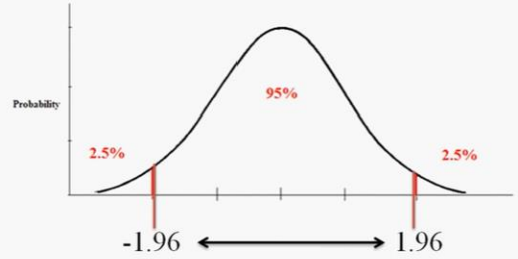
$H_0; p = .90$

$H_1; p \neq .90$

2. State Alpha

$\alpha = 0.05$

3. State Decision Rule



If z is less than -1.96 or greater than 1.96, reject the null hypothesis.

4. Calculate Test Statistic

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$\hat{p} = .82$

$p_0 = .90$

$n = 100$

$$z_0 = \frac{.82 - .90}{\sqrt{\frac{.90(1-.90)}{100}}} = \frac{-0.08}{0.03} = -2.667$$

5. State Results

Decision Rule: If Z is less than -1.96, or greater than 1.96, reject the null hypothesis.

Z = -2.667

Result: Reject H0

6. State Conclusion

The claim that 9 out of 10 doctors recommend aspirin for their patients is not accurate, z = -2.667, p < 0.05.

Hypothesis test for two sample proportions

INFERENCE WITH THE NORMAL APPROXIMATION

使用正态分布的推论

ASSUMPTIONS FOR USING THE NORMAL DISTRIBUTION

本部分内容解释了如何进行假设检验，以确定两个比例之间的差异是否显著

The test procedure, called the **two-proportion z-test**, is appropriate when the following conditions are met：

- The sampling method for each population is simple **random sampling**.
- The samples are **independent**.
- Each sample includes at least **10 successes and 10 failures**.
- Each population is at least 20 times as big as its sample.

样本中预计至少有10次成功和10次失败

$n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$

$n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

比例差异的假设检验包括四个步骤：

（1）陈述假设

（2）制定分析计划

（3）分析样本数据

（4）解释结果

State the Hypotheses

Set	Null hypothesis	Alternative hypothesis	Number of tails
1	$P_1 - P_2 = 0$	$P_1 - P_2 \neq 0$	2
2	$P_1 - P_2 \geq 0$	$P_1 - P_2 < 0$	1
3	$P_1 - P_2 \leq 0$	$P_1 - P_2 > 0$	1

$H_0: P_1 = P_2$

$H_a: P_1 \neq P_2$

Formulate an Analysis Plan

- **显著性水平：**选择的显著性水平一般是0.01、0.05或0.10；但可以使用0到1之间的任何值
- **测试方法：**使用双比例z检验来确定人口比例之间的假设差异是否与观察到的样本差异显著不同

Analyze Sample Data（分析）

通过以下步骤完成计算，以找到检验统计量及其相关的P值

- **Pooled sample proportion.** Since the null hypothesis states that $P_1=P_2$, we use a pooled sample proportion (p) to compute the standard error of the sampling distribution.

$$\hat{p} = (\hat{p}_1 * n_1 + \hat{p}_2 * n_2) / (n_1 + n_2)$$

where p_1 is the sample proportion from population 1, p_2 is the sample proportion from population 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- **Standard error.** Compute the standard error (SE) of the sampling distribution difference between two proportions.

$$SE = \text{sqrt}\{ \hat{p} * (1 - \hat{p}) * [(1/n_1) + (1/n_2)] \}$$

where p is the pooled sample proportion, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- **Test statistic.** The test statistic is a z-score (z) defined by the following equation.

$$z = (\hat{p}_1 - \hat{p}_2) / SE$$

where p_1 is the proportion from sample 1, p_2 is the proportion from sample 2, and SE is the standard error of the sampling distribution.

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the Normal Distribution Calculator to assess the probability associated with the z-score. (See sample problems at the end of this lesson for examples of how this is done.)

再将P值与显著性水平α 进行比较，如果P值小于显著性水平时，拒绝无效假设

➢ z-Test for Proportions, Two Samples

Researchers want to test the effectiveness of a new anti-anxiety medication. In clinical testing, 64 out of 200 people taking the medication report symptoms of anxiety. Of the people receiving a placebo, 92 out of 200 report symptoms of anxiety. Is the medication working any differently than the placebo? Test this claim using alpha = 0.05

1. Define Null and Alternative Hypotheses

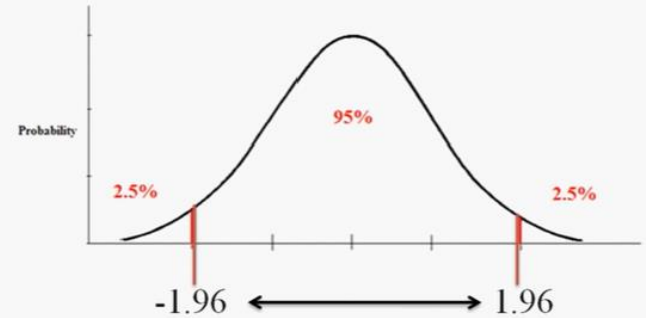
$$H_0; p_1 = p_2$$

$$H_1; p_1 \neq p_2$$

2. State Alpha

$$\alpha = 0.05$$

3. State Decision Rule



If z is less than -1.96 or greater than 1.96, reject the null hypothesis.

4. Calculate Test Statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$n_1 = 200 \qquad \hat{p}_1 = \frac{64}{200} = 0.32$$

$$n_2 = 200 \qquad \hat{p}_2 = \frac{92}{200} = 0.46$$

$$\hat{p} = 0.39$$

$$z = \frac{(0.32 - 0.46)}{\sqrt{0.39(1-0.39)}\sqrt{\frac{1}{200} + \frac{1}{200}}} = \frac{-0.14}{(.488)(0.1)} = 2.869$$

5. State Results

Decision Rule: If z is less than -1.96, or greater than 1.96, reject the null hypothesis.

z = 2.869

Result: Reject H0.

6. State Conclusion

There was a significant difference in effectiveness between the medication group and the placebo group, z = -2.869, p < 0.05.

Hypothesis test for two sample proportions

INFERENCE FOR TWO-WAY TABLES

使用列联表的推论

双向列联表（two-way tables)

列联表是医学科研中最常见的数据存储格式（或者说数据类型）之一。通常，列联表的横纵方向展示的是两个不同的分类变量，最常见的类型就是四格表（即2×2的列联表）。如下图所示，横向变量是“是否患肺癌”，纵向变量是“是否吸烟”，都是二分类变量，表格中的数据展示的则是每个分类变量水平组合下的人数（频数）

表 5.5: 列联表数据

	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

A two-way table summarizes information about the relationship between two categorical variables.

双向表总结了有关两个分类变量之间关系的信息

Testing for a difference between p_1 and p_2 is equivalent to testing for association in a two-way table that has two rows and two columns.

测试p_1和p_2之间的差异，相当于测试双向表中两行两列间的关联

Formulating hypotheses in a two-way table

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 main question of interest:

- Do the data support the claim of a difference in outcome by treatment?

问题：观察的数据是否能说明两药物治疗效果有区别?

If there is no difference in outcome by treatment, then knowing treatment provides no information about outcome; treatment assignment and outcome are *independent* (i.e., *not associated*).

- H_0: Treatment and outcome are not associated.
- H_A: Treatment and outcome are associated.
  - This is inherently a two-sided alternative.

H0：不同药物治疗和效果没关系

H1：不同药物治疗和效果有关系
（默认：双尾检测）

我们前面介绍了*z*-检验，*t*-检验，方差分析

今天我们来介绍另一种检验——**卡方检验(Chi-square test)**

- 卡方检验是一种用途很广， 卡方检验是以 χ^2 分布为基础的计数、分类资料的假设检验方法。
- 它主要是**比较两个及两个以上**样本率(构成比) 以及两个分类变量的关联性分析。
- 其根本思想就是在于比较理论频数和实际频数的**吻合程度或拟合优度**问题。

- H_0：观察频数与期望频数没有差别。H_1：观察频数与期望频数有差别。

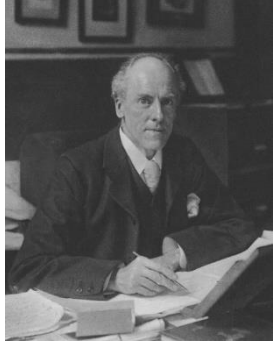
性别	肥胖程度			
	不肥胖	轻度肥胖	中/重度肥胖	Total
男	19	9	15	43
女	49	14	43	106
Total	68	23	58	149

列联表（Contingency table）

列联表

	汽车尺寸			
计数 合计百分比 列百分比 行百分比	large	Medium	Small	合计
American	36	53	26	115
	11.88	17.49	8.58	37.95
	85.71	42.74	18.98	
	31.30	46.09	22.61	
European	4	17	19	40
	1.32	5.61	6.27	13.20
	9.52	13.71	13.87	
	10.00	42.50	47.50	
Japanese	2	54	92	148
	0.66	17.82	30.36	48.84
	4.76	43.55	67.15	
	1.35	36.49	62.16	
合计	42	124	137	303
	13.86	40.92	45.21	

卡尔·皮尔逊、卡方分布（Pearson χ^2 test）



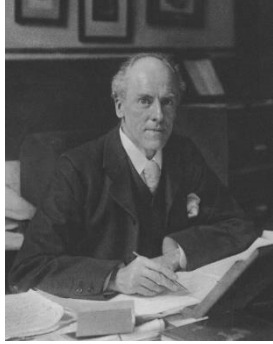
- χ^2 分布（卡方分布），所谓的卡方检验也来源于此。是由被称为数理统计学之父的卡尔·皮尔逊（Karl Pearson）于1900年提出的。
- 人们发现数据的正态性并不是一个必然的事实，促使 Pearson 研究并提出了偏态分布(偏斜分布)，并定义了四个通用的参数：均值、标准差、偏度和峰度。
- Pearson 又考虑到实际收集到的数据与理论分布总是存在一些差异，**需要一种方法来判定实际数据是否能够很好地拟合目标分布**。1900年，他提出“χ^2 拟合优度检验”的概念。
- **χ^2 拟合优度检验**的主要思想就是：
  - 针对每个指定的值，理论计算出来的频率与实际收集到的数据统计出来的频率之间总是存在一些偏差；
  - 把每一个指定值的偏差以平方的形式加起来，如果这个值比较小，则说明分布拟合得较好；
  - 如果这个值很大，则说明实际收集到的数据与目标分布并不相同，需要去寻找其它恰当的分布；
  - 为此，Pearson 引入了一个重要的统计量 —— **χ^2 统计量**

$$\chi^2 = \sum_{i=1}^{n} [\frac{(f_i - F_i)^2}{F_i}]$$

f_i 为实际值，F_i 为理论值。
χ^2 用于衡量实际值与理论值的差异程度，这也是**卡方检验的核心思想**。

χ^2 包含了以下2个信息：
- 实际值与理论值偏差的绝对大小。
- 差异程度与理论值的相对大小。

卡尔·皮尔逊、卡方分布（Pearson χ^2 test）



- χ^2 分布（卡方分布），所谓的卡方检验也来源于此。是由被称为数理统计学之父的卡尔·皮尔逊（Karl Pearson）于1900年提出的。
- 人们发现数据的正态性并不是一个必然的事实，促使 Pearson 研究并提出了偏态分布(偏斜分布)，并定义了四个通用的参数：均值、标准差、偏度和峰度。
- Pearson 又考虑到实际收集到的数据与理论分布总是存在一些差异，**需要一种方法来判定实际数据是否能够很好地拟合目标分布**。1900年，他提出“χ^2 拟合优度检验”的概念。
- **χ^2 拟合优度检验**的主要思想就是：
  - 针对每个指定的值，理论计算出来的频率与实际收集到的数据统计出来的频率之间总是存在一些偏差；
  - 把每一个指定值的偏差以平方的形式加起来，如果这个值比较小，则说明分布拟合得较好；
  - 如果这个值很大，则说明实际收集到的数据与目标分布并不相同，需要去寻找其它恰当的分布；
  - 为此，Pearson 引入了一个重要的统计量 —— **χ^2 统计量**

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - F_i)^2}{F_i}\right]$$

f_i 为实际值，F_i 为理论值。
χ^2 用于衡量实际值与理论值的差异程度，这也是**卡方检验的核心思想**。

χ^2 包含了以下2个信息：
- 实际值与理论值偏差的绝对大小。
- 差异程度与理论值的相对大小。

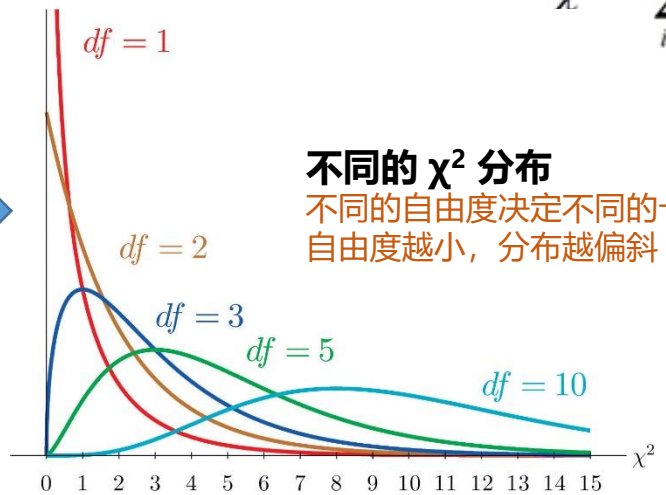
χ^2 分布

若k个相互独立的随机变量Z_1、Z_2、……、Z_k，均服从标准正态分布，则它们与均值之间偏差的平方和

$$X = \sum_{i=1}^{k}(Z_i - \bar{Z})^2 \sim \chi^2_{k-1}$$

构成一新的随机变量，其分布概率为**χ^2 分布**

χ^2 分布是由正态分布构造而成的一个新的分布，**当自由度大时，χ^2 分布近似为正态分布**



不同的 χ^2 分布
不同的自由度决定不同的卡方分布，自由度越小，分布越偏斜

The Chi-Square Test

1) Goodness-of-Fit test
2) Test of Independence
3) Test for Homogeneity

The Chi-Square Test

1) Goodness-of-Fit test 拟合度检验
2) Test of Independence
3) Test for Homogeneity

卡方检验 - 拟合度检验（Goodness of fit）

在这种类型的假设检验中，你需要确定**数据是否“符合”特定分布**。
如，你不知道未知数据是否符合二项分布。使用卡方检验（即假设检验的分布为卡方）来确定是否存在拟合度。

- **拟合度检验（Goodness of fit）**
  - 拟合度检验展示了一个观察到的频率分布是否与一个理论分布不同

    A test of goodness of fit establishes whether an observed frequency distribution differs from a theoretical distribution

以掷骰子为例，下表为投掷120次六面骰子的实际观察值：

点数	观察值
1	18
2	19
3	23
4	20
5	16
6	24

https://blog.csdn.net/zfcjhdq/article/details/83512680

拟合度/适合度检验 (Goodness of fit)

1.1 观察值和期望值

我们知道，在正常情况下，掷骰子服从二项分布$X \sim B(n,p)$，其数学期望$E[X]=np$，方差为$Var[X]=np(1-p)$。

在进行适合度检验时，我们的原假设H_0为观察到的掷骰子结果符合理论上的二项分布（Binomial distribution）。那么，我们就能得到**掷骰子的实际观察值和理论期望值如下表：**

H0: 观察到的结果符合期望值

点数	观察值	期望值
1	18	20
2	19	20
3	23	20
4	20	20
5	16	20
6	24	20

https://blog.csdn.net/zfcjhdq/article/details/83512680

适合度检验 (Goodness of fit)

1.2 χ^2值计算

根据公式 $\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i}$

$$\chi^2 = 2.3$$

自由度k \ P value （概率）	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

1.3 自由度确定

对于适合度检验，自由度的计算按照以下公式：

$$k = C - M$$

其中，C为我们观察到的类别数，此例中为6；M为我们要比较的理论分布的参数的个数，此例中为1；因此$k = 5$。

(就是约束条件个数，一般来说是1；df = k = C - 1)

1.4 H_0假设接受与拒绝

查表我们能够得到，统计量$\chi^2 = 2.3$所对应的p-value大约为0.8，与显著性水平0.05相差甚远，故我们不能拒绝原假设H_0.

https://blog.csdn.net/zfcjhdq/article/details/83512680

The Chi-Square Test

1) Goodness-of-Fit test
2) Test of Independence 独立性检验
3) Test for Homogeneity

卡方检验 - 独立性检验 (Independence)

独立性测试决定两个因素是否独立

A test of independence assesses whether unpaired observations on two variables, expressed in a contingency table, are independent of each other

独立性检验评估的是，在一个列联表中，不成对的观测对象中的两个变量是不是相互独立的

以下面的表格为例，我们来探究喝牛奶对感冒发病率有没有影响：

实际值	感冒人数	未感冒人数	合计	感冒率
喝牛奶组	43	96	139	30.94%
不喝牛奶组	28	84	112	25.00%
合计	71	180	251	28.29%

独立性测试使用所有观察（数据）值的列联表
独立性检验的检验统计量与拟合优度检验的检验统计量相似

https://blog.csdn.net/zfcjhdq/article/details/83512680

2. 独立性检验 (Independence)

2.2 卡方值计算

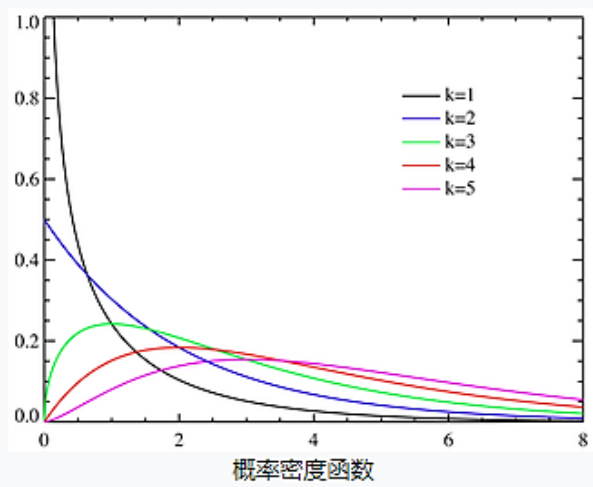
$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i}$$

其中，A_i为单元格i中的观察值，p_i为单元格i中的在H_0假设前提下的概率，k为单元格数。

上例中$\chi^2 = 1.077$

2.3 H_0假设拒绝与接受

根据得到的χ^2值，还并不能直接到的p-value。因为卡方分布根据其自由度有所不同，如下图所示：



卡方分布的概率密度函数如下：

$$f(x;k) = \begin{cases} \frac{x^{(k/2-1)}e^{-x/2}}{2^{k/2}\Gamma(\frac{k}{2})}, & x > 0 \\ 0, & otherwise \end{cases}$$

其中k为自由度。

所以，首先需要计算所研究样本的自由度

$$k = (R-1)(C-1)$$

其中R为单元格的行数，C为单元格的列数。

上述例子中，自由度$k = (2-1)(2-1) = 1$

自由度k \ P value （概率）	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

根据给定的自由度k以及 χ^2 值，计算 p-值

对于得到的p-value，与自己指定的显著性水平作比较（通常将0.05作为显著性水平），如果得到的p-value小于0.05，那我们认为样本所表现出来的结果是小概率事件，则我们有理由拒绝零假设H_0

=> 我们得到χ^2 值1.077，小于 3.84， 且接近 p=0.3 时的χ^2 值1.07，所以不能拒绝零假设

在一个志愿者中，21岁及以上的成年人每周自愿花一到九个小时与残疾老年人在一起。该项目招收社区大学生、四年制大学生和非学生。以下有一个成人志愿者的样本，以及他们每周做志愿者的小时数

Number of Hours Worked Per Week by Volunteer Type (Observed)The table contains **observed (O)** values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

社区大学生

四年制大学生

非学生

Q: Is the number of hours volunteered **independent** of the type of volunteer?

观察的数据表和末尾的问题： “志愿者的小时数是否独立于志愿者的类型？”
都告诉你这是对独立性的检验。两个因素分别是志愿者的小时数和志愿者的类型。

H_0：志愿者的小时数与志愿者的类型无关。
H_a：志愿者的时间取决于志愿者的类型。

Number of Hours Worked Per Week by Volunteer Type (Observed)The table contains **observed (O)** values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

For example, the calculation for the expected frequency for the top left cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)\,(298)}{839} = 90.57$$

Number of Hours Worked Per Week by Volunteer Type (Expected)The table contains **expected** (*E*) values (data).

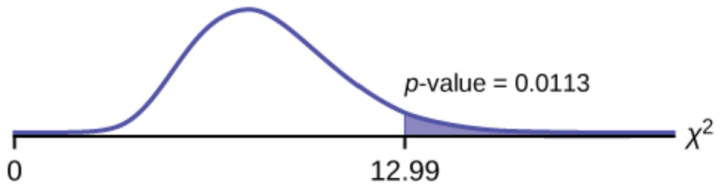
Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

df = (3 columns – 1)(3 rows – 1) = (2)(2) = 4

Graph:



{:}

Probability statement: *p*-value=$P(\chi^2 > 12.99) = 0.0113$

Compare *α* and the *p*-value: Since no *α* is given, assume *α* = 0.05. *p*-value = 0.0113. *α* > *p*-value.

Make a decision: Since *α* > *p*-value, reject H_0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

The Chi-Square Test

1) Goodness-of-Fit test
2) Test of Independence
3) Test for Homogeneity

Goodness-of-fit 拟合优度检验可以用来决定一个种群是否符合给定的分布，但它不足以决定两个种群是否遵循相同的未知分布。另一种测试称为**同质性测试**，可以用来得出两个群体是否具有相同分布的结论。我们需要计算同质性检验的检验统计量，需要**遵循与独立性检验相同的步骤。**

Hypotheses：

H_0: The distributions of the two populations are the same.

H_a: The distributions of the two populations are not the same.

Test Statistic: Use a χ2 test statistic.

It is computed in the same way as the test for independence.

Degrees of Freedom DF = (# of rows - 1)(# of columns - 1)

Requirements: All values in the table must be greater than or equal to 5.

男女大学生的生活安排是否相同？使用α=0.05的显著性级别。
假设随机选取250名男大学生和300名女大学生，被问及他们的生活安排：宿舍、公寓、与父母在一起、其他。问：男女大学生的生活安排是否相同？

Distribution of Living Arragements for College Males and College Females

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

H0：男大学生的生活安排分布与女大学生的生活安排分布相同。
Ha：男大学生的生活安排分布与女大学生的生活安排分布不一样。

Degrees of Freedom (*df*):
df = number of columns – 1 = 4 – 1 = 3

Calculate the test statistic: $\chi^2 = 10.1287$ (calculator or computer)

Probability statement: *p*-value = $P(\chi^2 > 10.1287) = 0.0175$

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students **are not the same.**

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Comparison of the Chi-Square Tests

χ^2 test statistic can be used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use：

拟合优度：使用拟合优度测试来确定未知分布的总体是否“符合”已知分布。在这种情况下，将有一个单一的定性调查问题或来自单一人群的单一实验结果。拟合优度通常用于查看总体是否一致（所有结果出现的频率相同），总体是否正常，或者总体是否与具有已知分布的另一个总体相同。零假设和替代假设是：

H0：人口符合给定的分布。
Ha：人口不符合给定的分布。

独立性：使用独立性测试来决定两个变量（因素）是独立的还是相依的。在这种情况下，将有两个定性调查问题或实验，并将构建一个列联表。目的是看看这两个变量是不相关的（独立的）还是相关的（依赖的）。零假设和替代假设是:

H0：这两个变量（因子）是独立的。
Ha：这两个变量（因素）是相关的。

同质性：使用同质性测试来确定两个分布未知的群体是否具有相同的分布。在这种情况下，将有一个单一的定性调查问题或实验给予两个不同的人群。无效假设和替代假设是：

H0：这两个种群遵循相同的分布。
Ha：这两个种群有不同的分布。

The Chi-Square Test

1) Goodness-of-Fit test
2) Test of Independence
3) Test for Homogeneity

Summary:

Goodness of Fit: used to compare a single sample proportion against a publicized model.

Homogeneity: used to examine whether things have changed or stayed the same or whether the proportions that exist between two populations are the same, or when comparing data from MULTIPLE samples.
Independence: determine if two categorical variables are associated or NOT (INDEPENDENT). The thinking is similar in objective to linear regression with quantitative variables.

Homogeneity and Independence are determined by the same chi-square test procedure. Luckily the steps for the chi-square test of independence are much the same as the steps in the chi-square test of homogeneity.

- 拟合优度检验通常用于确定数据是否符合特定分布。
- 独立性测试使用列联表来确定两个因素的独立性。
- 同质性测试确定两个群体是否来自同一分布，即使该分布未知。

卡方检验可以做特征选择（Use in machine learning）

卡方检验经常被用来做特征选择。举个例子，假设我们有一堆新闻标题，需要判断标题中包含某个词（比如某明星）是否与该条新闻的类别归属（比如娱乐）是否有关，我们只需要简单统计就可以获得这样的一个四格表：

组别	属于娱乐	不属于娱乐	合计
不包含某明星	19	24	43
包含某明星	34	10	44
合计	53	34	87

那么首先假设标题是否包含某明星与新闻是否属于娱乐是独立无关的，随机抽取1条新闻标题，属于娱乐类别的概率是：(19 + 34) / (19 + 34 + 24 +10) = 60.9%

组别	属于娱乐	不属于娱乐	合计
不包含某明星	43 * 0.609 = 26.2	43 * 0.391 = 16.8	43
包含某明星	44 * 0.609 = 26.8	44 * 0.391 = 17.2	44

显然，如果两个变量是独立无关的，那么四格表中的理论值与实际值的差异会非常小。

则χ^2值为:

$$\chi^2 = \frac{(19-26.2)^2}{26.2} + \frac{(34-26.8)^2}{26.8} + \frac{(24-16.8)^2}{16.8} + \frac{(10-17.2)^2}{17.2} = 10.00$$

标准的四格表χ^2值可以用以下方式进行计算：

$$\chi^2 = \frac{N*(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

其中，$N = A + B + C + D$

得到χ^2的值以后，怎样可以得知无关性假设是否可靠？接下来我们应该查询卡方分布的临界值表了。

首先我们明确自由度的概念：自由度v=(行数-1)*(列数-1)。
然后看卡方分布的临界概率，表如下：

自由度k \ P value （概率）	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

一般取p=0.05，也就是说两者不相关的概率为0.05时，对应的卡方值为3.84。显然10.0>3.84，那就说明包含某明星的新闻不属于娱乐的概率小于0.05。换句话说，包含某明星的新闻与娱乐新闻相关的概率大于95%!

总结一下：我们可以通过卡方值来判断特征是否与类型有关。**卡方值越大，说明关联越强，特征越需要保留。卡方值越小，说明越不相关，特征需要去除**

卡方检验的要求（assumptions）

- **卡方分布本身是连续型分布，但是在分类资料的统计分析中，显然频数只能以整数形式出现**
- **The groups being tested must be independent**
- **一般认为对于卡方检验中的每一个单元格，要求其最小期望频数均大于1；**
- **且至少有4 / 5的单元格期望频数大于5（期望频数小于5的单元格不能超过20%）；**
  - 如果有多于20%的单元格期望频数小于5，卡方统计量会变大，也容易造成假阳性（假的拒绝）的概率增大，这时可以采用似然比卡方进行修正

符合以上条件后使用卡方分布计算出的概率值才是准确的。

如果数据不符合要求，可以换成采用Fisher 精确检验（Fisher’s exact test）

Test of proportions: z-test *vs*. chi-square test

Hypothesis test with frequency (count) data and proportions

Categorical variable*:* responses or are categories or groups ("levels")

Goal: handle response variables with ***any*** number of categories and grouping variables with ***any*** number of groups ***with a single statistic***

	one variable (a response variable)	two variables (one grouping, one response)
binary variable(s)	z-test for a single proportion (1 x 2 table)	z-test for a difference in proportions (2 x 2 table)
any categorical variable(s)	chi-square test for goodness-of-fit (1 x any # table)	chi-square test for independence (any # by any # table)

小结

	One Factor	Two Factors
binary variable(s)	z-test for a single proportion (1 x 2 table) 发病率大于p0？发病率CI?	z-test for difference in proportions(2x2 table) 两个地方的发病率有差异吗?
any categorical variable(s)	chi-square test for goodness-of-fit(1 x any # table) 不同地区发病率符合假定的分布?	chi-square test for independence (any RC table) e.g. 分布与因素有关吗?

谢谢，下周见！