分布拟合检验

主讲人：刘宏志

liuhz@ss.pku.edu.cn

北京大学

假设检验

假设检验 { 参数假设检验

非参数假设检验

总体分布已知，检验关于未知参数的某个假设

总体分布未知时的假设检验问题

假设检验

- 根据样本信息检验关于总体的某个假设是否正确
- 逻辑上运用反证法，统计上依据小概率原理
- 小概率：
  - 在一次试验中，一个几乎不可能发生的事件发生的概率
  - 概率是0～1之间的一个数，因此小概率就是接近0的一个数
  - R. Fisher 把1/20作为标准，即比0.05小的概率被认为是小概率
- 小概率原理：
  - 小概率事件在一次实验中几乎是不可能发生的

引例：战争频率统计

- 从1500到1931年，每年爆发战争的次数可看作一个随机变量
- 椐统计，这432年间共爆发了299次战争，数据如下:

战争次数X	发生X次战争的年数
0	223
1	142
2	48
3	15
4	4

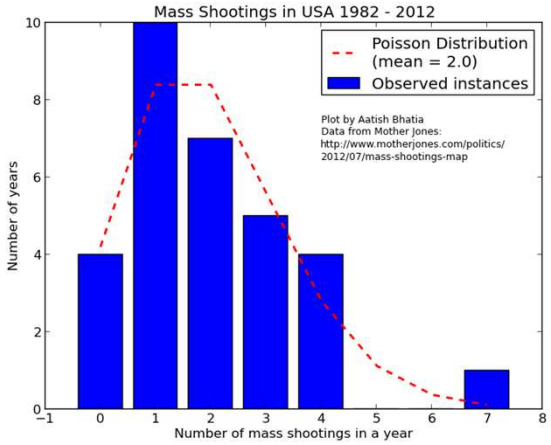
根据先验知识，这一随机变量应该服从什么分布？

泊松分布：描述单位时间内随机事件发生的次数

问题：以上数据能否证实X具有泊松分布？

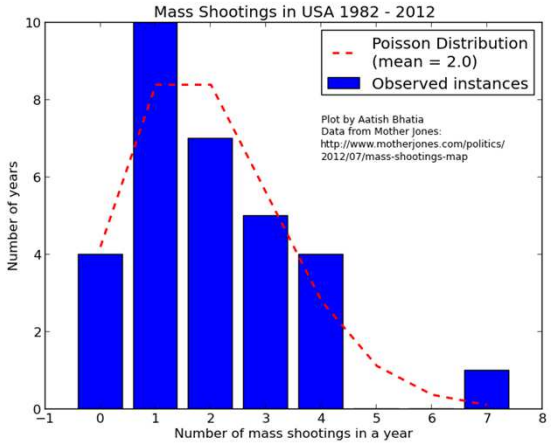
χ^2检验

- 基本思想：
  - 总体分布未知，根据样本检验关于总体分布的假设
  - 先提出原假设: H_0：总体X的分布函数为$F(x)$
  - 然后根据样本的经验分布和所假设的理论分布之间的吻合程度来决定是否接受原假设
- 这种检验通常被称作拟合优度检验，是一种非参数检验
- 由统计学家K.皮尔逊在1900年提出，被视为近代统计学的开端



χ^2检验：基本原理和步骤

1. 提出原假设: H_0：总体X的分布函数为$F(x)$
2. 将总体X的取值范围划分为k个互不重迭的区间, $(a_0,a_1],\ldots,(a_{k-1},a_k]$,记作$A_1, A_2, \ldots, A_k$
3. 把落入第i个区间A_i的样本值的个数记作f_i，称为实测频数
4. 根据所假设的理论分布算出总体X的值落入每个A_i的概率p_i,则np_i为落入A_i的样本值的理论频数



实测频数

理论频数

$$f_i - np_i$$

皮尔逊引进如下统计量表示经验分布与理论分布间的差异:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

在理论分布已知的条件下，np_i是常量

或 $$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i}\left(\frac{f_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{f_i^2}{np_i} - n$$

统计量 χ^2 的分布是什么?

皮尔逊定理

- 定理1：若原假设中的理论分布$F(x)$已经完全给定，那么当n充分大($n\geq 50$)时，统计量χ^2近似服从自由度为k-1的χ^2分布

- 定理2：如果理论分布$F(x)$中有r个未知参数需用相应估计量来代替，则当n充分大时，统计量χ^2近似服从自由度为 $(k-r-1)$ 的χ^2分布.

- 皮尔逊定理是在n无限增大时推导出来的，在使用时要注意n要足够大，以及np_i不太小这两个条件.

- 根据计算实践，要求n不小于50，以及np_i都不小于5. 否则应适当合并区间，使np_i满足这个要求 .

DoF are the measurements of the number of values in the statistic that are free to vary without influencing the result of the statistic.

北京大学

5. 对给定的显著性水平α，通过查χ^2分布表确定l值，使$P\{\chi^2 > l\} = \alpha$，得到拒绝域：$\chi^2 > l$

6. 根据所给的样本值$x_1, x_2, \cdots, x_n$计算统计量χ^2

7. 若χ^2的实测值落入拒绝域，则拒绝原假设H_0，否则就认为差异不显著而接受原假设H_0

示例：战争频率统计

- 提出假设H_0：X服从参数为λ的泊松分布
- 根据观察结果,得到参数为λ的最大似然估计：

$$\hat{\lambda} = \bar{x} = 0.69$$

- 按照参数为0.69的泊松分布，计算事件$X=i$的概率p_i的估计：$\hat{p}_i = e^{-0.69}\,0.69^i / i!$

战争次数 x	0	1	2	3	4	
实测频数 f_i	223	142	48	15	4	
$\hat{p}_i$	0.58	0.31	0.18	0.01	0.02	
$n\hat{p}_i$	216.7	149.5	51.6	12.0	2.16	
$(f_i - n\hat{p}_i)^2 / n\hat{p}_i$	0.183	0.376	0.251	1.623		$\Sigma = 2.433$

将$n\hat{p}_i < 5$的组予以合并，即将发生3次和4次战争的组归并为一组

示例：战争频率统计

- 因H_0所假设的理论分布中有一个未知参数,故自由度为：4-1-1=2

- 按α=0.05，自由度为2 查χ^2分布表得

$$\chi^2_{0.05}(2) = 5.991$$

- 因统计量χ^2的观察值χ^2=2.433<5.991，未落入拒绝域，故认为每年发生战争的次数X服从参数为0.69的泊松分布

示例：骰子检查

- 将一颗骰子掷120次，所得数据如下表：

点数 i	1	2	3	4	5	6
出现次数 f_i	23	26	21	20	15	15

问这颗骰子是否均匀、对称（取α=0.05）？

- 解：若这颗骰子是均匀、对称的,则1~6点中每点出现的可能性相同,都为1/6.如果用A_i表示第i点出现，则待检假设：

$$H_0: P(A_i)=1/6,\ i=1,2,\ldots,6$$

- 在H_0成立的条件下，理论概率$p_i=p(A_i)=1/6$
- 由$n=120$得频率 $np_i=20$

i	f_i	p_i	np_i	$(f_i - np_i)^2/(np_i)$
1	23	1/6	20	9/20
2	26	1/6	20	36/20
3	21	1/6	20	1/20
4	20	1/6	20	0
5	15	1/6	20	25/20
6	15	1/6	20	25/20
合计	120			4.8

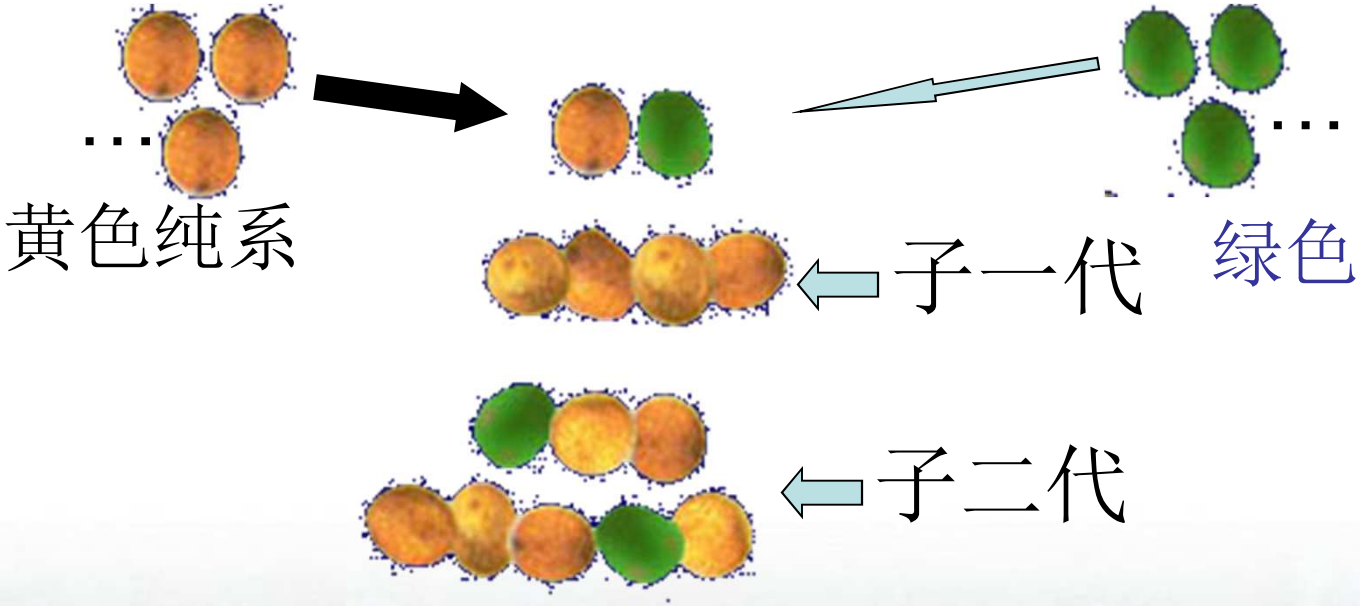
此分布不含未知参数，k=6，α=0.05，查表得：$\chi_\alpha^2(k-1) = \chi_{0.05}^2(5) = 11.071$

上表：$\chi^2 = \sum_{i=1}^{6} \frac{(f_i - np_i)^2}{np_i} = 4.8 < 11.071$，故接受$H_0$，即骰子是均匀对称的

北京大学

示例：遗传规律

- 生物学家孟德尔进行了八年的豌豆杂交试验, 并根据试验结果,运用他的数理知识, 发现了遗传的基本规律
- 理论：黄、绿豌豆杂交，子二代中，黄、绿之比为3:1
- 一组观察结果为： 黄：70、绿：27，是否符合理论？



孟德尔

示例：遗传规律

由于随机性，观察结果与3:1 总有些差距，因此有必要考察某一大小的差异是否已构成否定3:1理论的充分根据.

检验孟德尔的3:1理论:

提出假设H_0: $p_1=3/4$, $p_2=1/4$

这里，$n=70+27=97$, $k=2$,

理论频数为： $np_1=72.75$, $np_2=24.25$

实测频数为：70，27.

统计量 $\chi^2 = \sum_{i=1}^{2} \frac{(f_i - np_i)^2}{np_i} \sim \chi^2(1)$

自由度为 k-1=1

按 α=0.05，自由度为1，查χ^2 分布表得

$\chi^2_{0.05}(1)=3.841$

由于统计量 χ^2 的实测值

$\chi^2=0.4158<3.841$，

未落入拒绝域.

故认为试验结果符合孟德尔的3:1理论.

北京大学

Kolmogorov–Smirnov检验

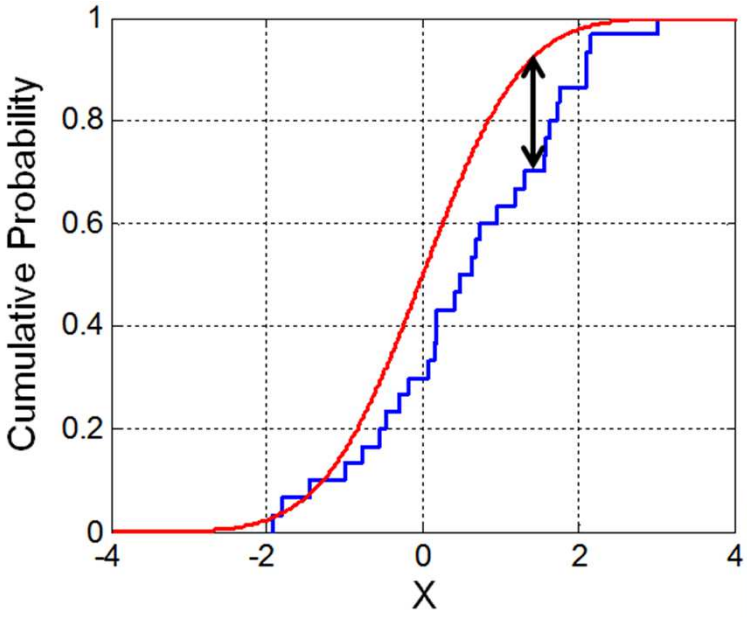
- 简写为K-S检验，亦称D检验法
- 一种拟合优度检验法
- 检验一组样本数据的实际分布与某一指定的理论分布是否相符
- 基本原理：
  - 将理论分布下的累积分布与观察到的累积分布相比较，找出它们间最大的差异点，并参照抽样分布，定出这样大的差异是否处于偶然

Kolmogorov–Smirnov检验

1. 提出原假设H_0:总体X的分布函数为$F(x)$
2. 计算样本累计频率与理论分布累计概率的绝对差，令最大的绝对差为D_n:

$$D_n = \max_{1 \le i \le n} \{|F(x) - F_n(x)|\}$$

3. 用样本容量n和显著水平α查表得临界值D_n^{α}
4. 通过D_n与D_n^{α}的比较做出判断，若$D_n < D_n^{\alpha}$，则认为拟合是满意的



北京大学

Each table entry is the value of a Kolmogorov-Smirnov one-sample statistic D_n for sample size n such that its right-tail probability is the value given on the top row.

n	.200	.100	.050	.020	.010	n	.200	.100	.050	.020	.010
1	.900	.950	.975	.990	.995	21	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
3	.565	.636	.780	.785	.829	23	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
17	.250	.286	.318	.355	.381	37	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252

For $n > 40$, right-tail critical values based on the asymptotic distribution can be calculated as follows:

.200	.100	.050	.020	.010
$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

Source: Adapted from L. H. Miller (1956), Table of percentage points of Kolmogorov statistics, *Journal of the American Statistical Association*, **51**, 111–121, with permission.

实例：正态拟合

- 某织布厂工人执行的生产定额（织机每小时生产织物的米数）情况如下表，试检验这些样本数据能否服从正态分布？

按定额执行情况分组	工人数
3.75～4.25	**20**
4.25～4.75	**372**
4.75～5.25	**498**
5.25～5.75	**103**
5.75～6.25	**7**
	1000

实例：正态拟合

- 首先，计算样本均值和标准差：$\bar{x}$=4.85；s=0.352, 作为总体均值和标准差的估计
- 建立假设：
  - H_0：样本数据服从均值为4.85，标准差为0.352的正态分布

按定额执行情况分组	工人数
3.75～4.25	**20**
4.25～4.75	**372**
4.75～5.25	**498**
5.25～5.75	**103**
5.75～6.25	**7**
	1000

正态拟合计算表

X的组限	标准化	标准正态概率	累计概率(理论概率)	累计工人数	实际累计频率	(2)-(4)的绝 对值
甲	乙	（1）	（2）	（3）	（4）	（5）
不足4.25	-∞～-1.70	0. 045	0. 045	20	0. 020	0. 025
4.25-4.75	-1.70～-0.28	0.345	0. 390	392	0. 392	0. 002
4.75-5.25	-0.28～1.14	0. 483	0. 873	890	0. 890	0. 017
5.25-5.75	1.14～2.56	0. 122	0. 995	993	0. 993	0. 002
5.75-6.25	2.56-+∞	0. 005	1. 000	1000	1. 000	0. 000

$$D_{(1000,0.05)} = \frac{1.36}{\sqrt{1000}} = 0.043 > 0.025$$

无法拒绝H_0

北京大学

A-D检验 和 CM准则

$$n\int_{-\infty}^{\infty}[Fn(x)-F(x)]^2\,\omega\,(x)\mathrm{d}F(x)$$

- **Cramér–von Mises** 准则：$w(x)=1$

$$w^2=n\int_{-\infty}^{\infty}[Fn(x)-F(x)]^2\,\mathrm{d}F(x)$$

- **Anderson–Darling** 检验：$\omega\,(x)=F(x)(1-F(x))$

$$A=n\int_{-\infty}^{\infty}\frac{[Fn(x)-F(x)]^2}{F(x)(1-F(x))}\mathrm{d}F(x)$$

$$A^2=-n-S,$$

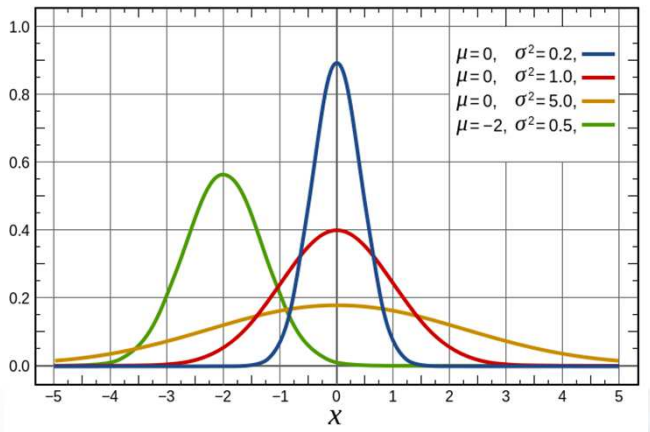
$$S=\sum_{i=1}^{n}\frac{(2i-1)}{n}[lnF(x_i)+\ln\left(1-F(x_{n+1-i})\right)]$$

正态性检验：J-B检验

- 正态分布的性质：
  - 偏度（三阶中心矩）：S=0； 峰度（四阶中心矩）：K=3
- 基本思想：
  - 若样本来自正态总体，则其偏度和峰度应该在0, 3附近
- J-B统计量： $JB = \frac{n}{6}[S^2 + \frac{(K-3)^2}{4}]$， n为样本容量

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{2}}$$



在正态分布的假设下，JB统计量渐进地服从自由度为2的卡方分布