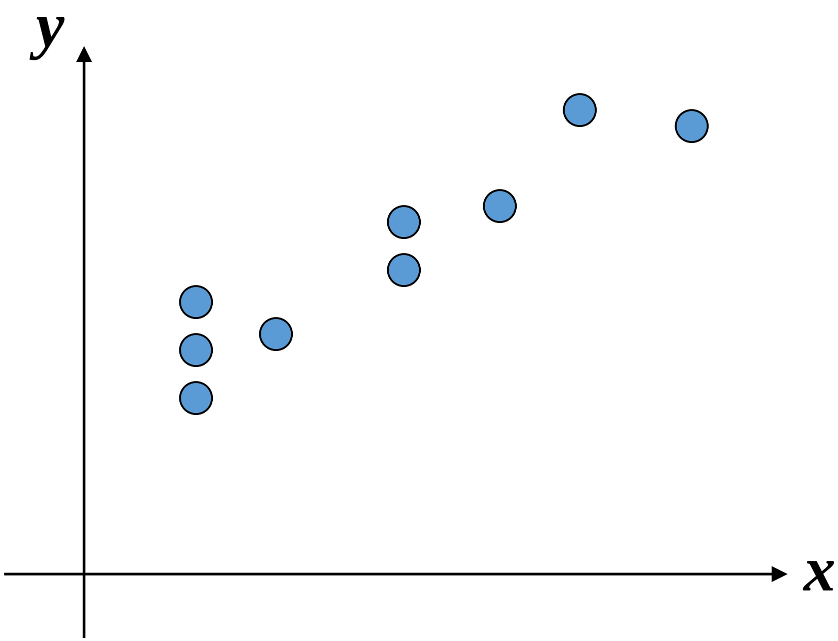
第三节 卡方拟合、独立性、齐一性检验

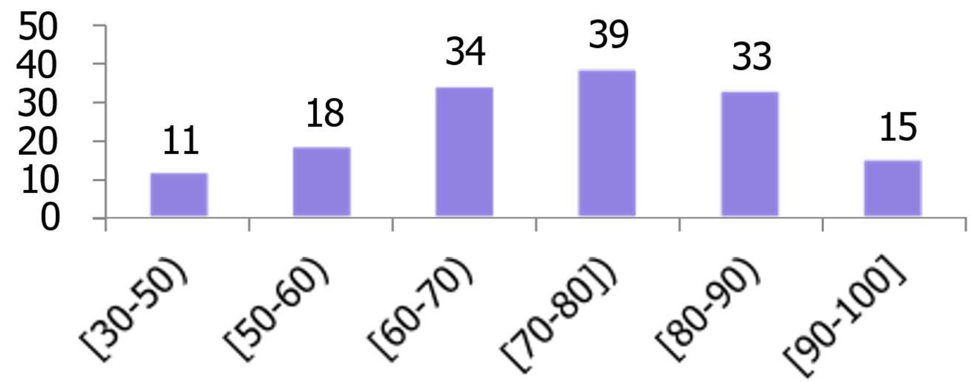
前面介绍的各种检验都是在总体服从正态分布前提下，对参数进行假设检验。

实际中可能遇到这样的情形：**总体服从何种理论分布并不知道**。

那么，如何得到分布函数形式，并验证其合理性？

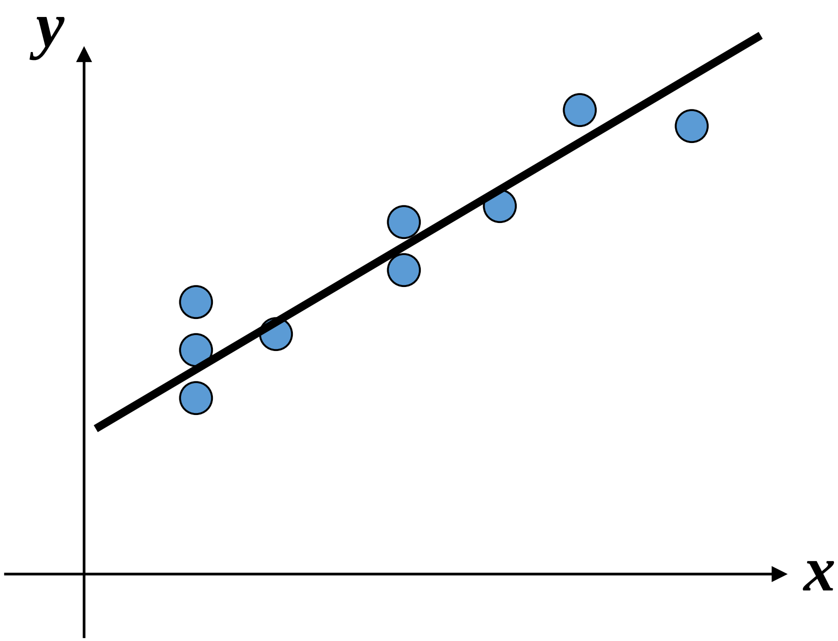
一、曲线拟合

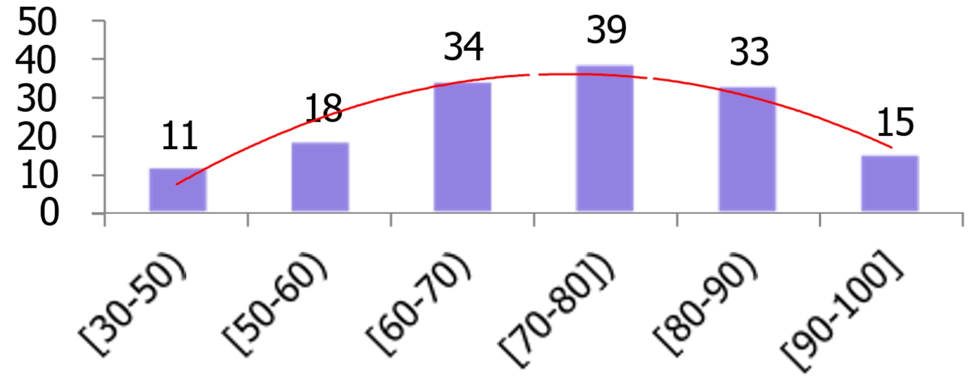




如何根据一组样本点，或是其频率分布得到数据分布规律（函数关系式）？

一、曲线拟合





曲线拟合：如图所示，常常需要从一组获得的数据点中，寻找变量与变量之间的变化规律。用几何方法来解释，就是用已知平面内的一组点，来确定一条曲线，使该曲线能在整体上刻画这组点的变化趋势而不需通过每个点，所求出的曲线称为拟合曲线。

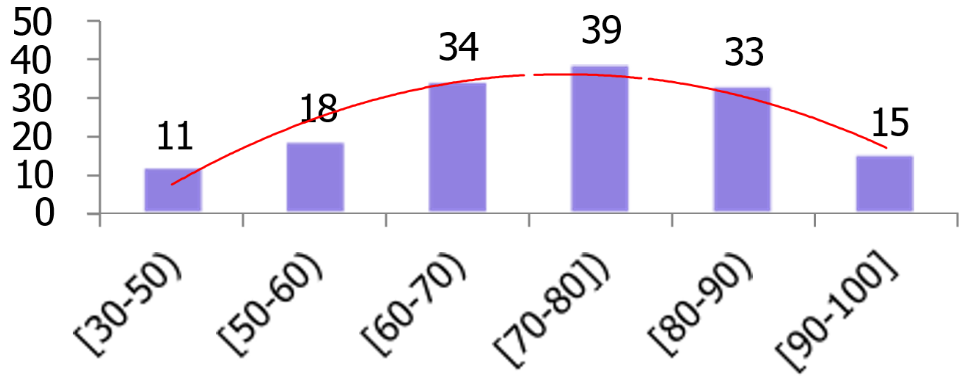
一、曲线拟合

曲线拟合方法：

- 拟合曲线：多项式拟合、直线拟合、指数拟合等；
- 参数确定：设有一组数据对$(x_i, y_i)(i = 1, 2, \cdots, m)$，常选择使得拟合曲线取值和实际数据点误差平方和最小的参数作为拟合曲线参数，即最小二乘法。

二、χ^2检验

引例1 重庆大学某次概率论与数理统计课程考试后学生成绩分类统计如下



问：能否用正态分布函数拟合学生成绩分布?

二、χ^2检验

引例2 一淘宝店主搜集了一年中每天的订单数X，除去春节期间及双十一前后外，按330天计：

订单数X	0	1	2	3	4	5	6	7
天数	3	6	21	46	48	61	52	42

订单数X	8	9	10	11	12	13	16
天数	27	11	6	4	1	1	1

通常认为每天的订单数服从泊松分布，以上的数据是否支持这个结论？

二、χ^2检验

- 对样本的频数分布所来自的总体分布是否服从某种理论分布或某种假设分布所作的假设检验，即根据样本的频数分布来推断总体的分布。
- χ^2检验用于对点计而来的离散型数据资料进行假设检验，对总体的分布不做要求，也不对总体参数进行推论, 因此属于自由分布的非参数检验。

二、χ^2检验

- 拟合检验
- ◆ 用于检验观测数据是否与理论分布或分布族相符
- ◆ 假设前提：1、观测数据来自于特定的总体或总体分布；2、预期的理论分布已知
- ◆ 应用场景：检验观测数据是否符合正态分布、泊松分布等特定的理论分布
- 独立性检验
- 齐一性检验

二、χ^2检验

- 拟合检验
- 独立性检验

◆ 用于检验两个随机变量之间是否存在相互独立的关系

◆ 假设前提：1、观测数据来自于一个大的总体；2、两个变量之间不存在相互依赖关系

◆ 应用场景：如检验商品价格与销量之间的关系等

- 齐一性检验

二、χ^2检验

- 拟合检验
- 独立性检验
- 齐一性检验

◆ 用于检验两个或多个总体分布是否相同，或多个总体是否具有相似的特征

◆ 假设前提：1. 观测数据来自于两个或多个总体或总体分布；2. 这些总体之间的差异主要是由随机因素引起的

◆ 应用场景：比较不同城市的平均气温；检验不同批次某一产品的质量是否相同

三、χ^2拟合检验

> 一般步骤：

1.在H_0下，总体X取值的全体分成k个两两不相交的子集$A_1,\ldots,A_k$；

2. 以$n_i\,(i=1,\ldots,k)$记样本观察值$x_1,\ldots,x_n$中落在A_i的个数（实际频数）；

三、χ^2拟合检验

- 一般步骤：

3.当H_0为真且$F_0(x)$完全已知时，计算事件A_i发生的概率$p_i = P_{F_0}(A_i), i = 1,\dots,k$;

当$F_0(x)$含有r个未知参数时，先利用极大似然法估计r个未知参数，然后求得p_i的估计$\hat{p}_i$.

此时称np_i（或$n\hat{p}_i$）为理论频数；

三、χ^2拟合检验

➢ 一般步骤：

4.检验统计量$\sum_{i=1}^{k} h_i(n_i - np_i)^2$， $h_i = ?$

检验的拒绝域形式为： $\sum_{i=1}^{k} h_i(n_i - np_i)^2 \geq c$

三、χ^2拟合检验

➢ 一般步骤：

定理： 若n充分大，则当H_0为真时，统计量

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\overset{\text{近似}}{\sim}\chi^2(k-1)$$

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}\overset{\text{近似}}{\sim}\chi^2(k-r-1)$$

其中k为分类数，r为$F_0(x)$中被估未知参数个数

三、χ^2拟合检验

$$\chi^2 = \boxed{\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}} = \sum_{i=1}^{k} \frac{n_i^2 - 2n_i np_i + n^2 p_i^2}{np_i} = \boxed{\sum_{i=1}^{k} \frac{n_i^2}{np_i} - n}$$

（教材P203）

三、χ^2拟合检验

➢ 一般步骤：

所以在显著水平α下拒绝域为

$$\chi^2=\sum_{i=1}^{k}\frac{n_i^{\ 2}}{np_i}-n\geq\chi_\alpha^{\ 2}(k-1),$$（没有参数需要估计）

$$\chi^2=\sum_{i=1}^{k}\frac{n_i^{\ 2}}{n\hat{p}_i}-n\geq\chi_\alpha^{\ 2}(k-r-1),$$（有r个参数需要估计）

三、χ^2拟合检验

> 一般步骤：

注：χ^2拟合检验使用时必须注意：

n要足够大，np_i(或$n\hat{p}_i$)不能太小。

根据实践，要求$n \geq 50$，np_i（或$n\hat{p}_i$）≥ 5，

否则应适当合并相邻的类，以满足要求。

三、χ^2拟合检验

> 基本思想：

- 在原假设H_0成立的条件下，实际频率与理论概率相差应该不大；
- 可用统计量$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$度量该差异，若$n$充分大（$n\geq 50$），则当$H_0$为真时，该统计量近似服从$\chi^2(k-1)$分布；
- 因此，若一次抽样得到的χ^2值超过一定显著性水平α对应的临界值$\chi_\alpha^2(k-1)$，则拒绝原假设H_0。

三、χ^2拟合检验

例1 某超市统计了某三个品牌纯净水的购买人数，统计结果如下：

品牌	1	2	3
购买人数	61	53	36

150

检验： 这三个品牌水的购买人数分布是否有差异？（$\alpha=0.05$）

三、χ^2拟合检验

解： $H_0: p_1 = p_2 = p_3 = \frac{1}{3}$， H_1:至少一个$p_i \neq \frac{1}{3}$

品牌	1	2	3
实际频数	61	53	36
理论频数	50	50	50
$(n_i - np_i)^2$	121	9	196

$$\chi^2 = \frac{121}{50} + \frac{9}{50} + \frac{196}{50} = 6.52 > \chi^2_{0.05}(2) = 5.992$$

结论： 拒绝H_0， 顾客对三种水的喜爱存在显著差异。

三、χ^2拟合检验

例2 一淘宝店主搜集了一年中每天的订单数X， 除去春节期间及双十一前后外，按330天计：

订单数X	0	1	2	3	4	5	6	7
天数	3	6	21	46	48	61	52	42

订单数X	8	9	10	11	12	13	16
天数	27	11	6	4	1	1	1

问：通常认为每天的订单数服从泊松分布，以上的数据是否支持这个结论？

三、χ^2拟合检验

解：$H_0: X \sim \pi(\lambda)$，λ 未知，总订单数为1749，

所以，平均每天订单数$\hat{\lambda} = \bar{x} = \frac{1749}{330} = 5.3$

概率估计（大于11的订单次数较小，所以将大于等于11的合并）：

$$\hat{p}_i = \frac{\hat{\lambda}^i e^{-\hat{\lambda}}}{i!}, i = 0,1,\ldots,10, \quad \hat{p}_{11} = \sum_{j=11}^{\infty} \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{j!} = 1 - \sum_{i=0}^{10} \hat{p}_i.$$

理论频数：$n\hat{p}_i, i = 0,1,\ldots,10,11, n\hat{p}_0 = 1.65 < 5$，将$x=0$与$x=1$合并（$n$=330）。

三、χ^2拟合检验

订单数X	0	1	2	3	4	5
天数	3	6	21	46	48	61
概率估计	0.005	0.026	0.070	0.124	0.164	0.174
理论频数	1.65	8.58	23.1	40.87	54.16	57.41

10.23

订单数X	6	7	8	9	10	≥ 11
天数	52	42	27	11	6	7
概率估计	0.154	0.116	0.077	0.045	0.024	0.021
理论频数	50.71	38.39	25.44	14.98	7.94	6.60

三、χ^2拟合检验

检验统计量的值为

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i^2}{n\hat{p}_i} - n = \sum_{i=1}^{11} \frac{n_i^2}{n\hat{p}_i} - 330 = 3.97$$

即在显著性水平$\alpha = 0.05$下

$$\chi_\alpha^2(k-r-1) = \chi_{0.05}^2(11\text{-}1\text{-}1) = 16.92$$

于是，$3.97 < 16.92$,接受原假设。

四、χ^2独立性检验

- 拟合检验
- 独立性检验
  - 用于检验两个随机变量之间是否存在相互独立的关系
  - 假设前提：1、观测数据来自于一个大的总体；2、两个变量之间不存在相互依赖关系
  - 应用场景：如检验商品价格与销量之间的关系等
- 齐一性检验

四、χ^2独立性检验

例3 为了研究吸烟对患肺癌是否有影响，随机调查了9965人，调查结果如下：

	未患肺癌	患肺癌	合计
不吸烟	7775	42	7817
吸烟	2099	49	2148
合计	9874	91	9965

问： 吸烟是否对患肺癌有影响？（$\alpha = 0.01$）

四、χ^2独立性检验

	未患肺癌B	患肺癌$\overline{B}$	合计
不吸烟A	7775	42	7817
吸 烟$\overline{A}$	2099	49	2148
合计	9874	91	9965

假设吸烟对是否患肺癌没有影响，即A与B独立。

$P(AB)=P(A)P(B)$　$P(\overline{A}B)=P(\overline{A})P(B)$

$P(A\overline{B})=P(A)P(\overline{B})$　$P(\overline{A}\overline{B})=P(\overline{A})P(\overline{B})$

事件AB、$\overline{A}B$、$A\overline{B}$和$\overline{A}\overline{B}$发生的理论频数？

四、χ^2独立性检验

	未患肺癌B	患肺癌$\overline{B}$	合计
不吸烟A	7775	42	7817
吸 烟$\overline{A}$	2099	49	2148
合计	9874	91	9965

事件AB、$\overline{A}B$、$A\overline{B}$和$\overline{A}\overline{B}$发生的理论频数？

$$nP(AB)=nP(A)P(B)=9965\times\frac{7817}{9965}\times\frac{9874}{9965}=7745.6$$

$$nP(A\overline{B})=nP(A)P(\overline{B})=9965\times\frac{7817}{9965}\times\frac{91}{9965}=71.4$$

$$nP(\overline{A}B)=nP(\overline{A})P(B)=9965\times\frac{2148}{9965}\times\frac{9874}{9965}=2128.4$$

$$nP(\overline{A}\overline{B})=nP(\overline{A})P(\overline{B})=9965\times\frac{2148}{9965}\times\frac{91}{9965}=19.6$$

四、χ^2独立性检验

	未患肺癌B	患肺癌$\overline{B}$	合计
不吸烟A	7775(7745.6)	42(71.4)	7817
吸烟$\overline{A}$	2099　(2128.4)	49(19.6)	2148
合计	9874	91	9965

$$\chi^2=\frac{(7775-7745.6)^2}{7745.6}+\frac{(42-71.4)^2}{71.4}+\frac{(2099-2128.4)^2}{2128.4}+\frac{(49-19.6)^2}{19.6}$$

$$=0.1112+12.1059+0.4061+44.1000\approx 56.72.$$

四、χ^2独立性检验

$$\chi^2=\frac{(7775-7745.6)^2}{7745.6}+\frac{(42-71.4)^2}{71.4}+\frac{(2099-2128.4)^2}{2128.4}+\frac{(49-19.6)^2}{19.6}$$

$$=0.1112+12.1059+0.4061+44.1000\approx 56.72.$$

在检验独立性时，自由度f=(行数-1)×（列数-1）=1

所以，$\chi^2=56.72>\chi^2_{0.01}(1)=6.637$

结论：拒绝原假设H_0，认为吸烟对患肺癌有显著影响。

五、χ^2齐一性检验

- 拟合检验
- 独立性检验
- 齐一性检验

◆ 用于检验两个或多个总体分布是否相同，或多个总体是否具有相似的特征

◆ 假设前提：1. 观测数据来自于两个或多个总体或总体分布；2. 这些总体之间的差异主要是由随机因素引起的

◆ 应用场景：比较不同城市的平均气温；检验不同批次某一产品的质量是否相同

五、χ^2齐一性检验

两个总体分布的齐一性检验

比较两个总体的分布函数$F_1(X)$和$F_2(Y)$是否一致？

假设检验： H_0： $F_1(X)=F_2(Y)$； H_1： $F_1(X)\neq F_2(Y)$。

- 对这两个总体进行独立抽样，分别获得$F_1(X)$和$F_2(Y)$的独立样本
- 这两个总体变量的值域应该一致。我们把该值域分成s段$A_1,\ldots,A_s$，比较$F_1(X)$和$F_2(Y)$在$A_1,\ldots,A_s$上的分布或比例是否一致。

五、χ^2齐一性检验

- **数据结构：**

总体分类	A_1	……	A_s	合计
X 频数	n_{11}	……	n_{1s}	n_1
Y 频数	n_{21}	……	n_{2s}	n_2
合计	n_{*1}	……	n_{*s}	n

这里

$$n = n_1 + n_2,\ \ n_{*j} = n_{1j} + n_{2j},\ (j = 1, \cdots, s)$$

五、χ^2齐一性检验

总体分类	A_1	……	A_s	合计
X 频数	n_{11}	……	n_{1s}	n_1
Y 频数	n_{21}	……	n_{2s}	n_2
合计	n_{*1}	……	n_{*s}	n

- H_0：$F_1(X)=F_2(Y)$成立时，意味着 $X_1,\cdots,X_{n_1}$和$Y_1,\cdots,Y_{n_2}$来自同一个总体，且 $P(X\in A_i)=P(Y\in A_i)$，$(i=1,\cdots,s)$
所以N_{1j}和N_{2j}的理论估计值为

$$\hat{N}_{1j}=n_1\frac{n_{*j}}{n} \text{ 和 } \hat{N}_{2j}=n_2\frac{n_{*j}}{n},\ (j=1,\cdots,s)$$

五、χ^2齐一性检验

由此得到检验统计量：

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{s}\frac{(n_{ij}-n_i n_{*j}/n)^2}{n_i n_{*j}/n}=\sum_{i=1}^{2}\sum_{j=1}^{s}\frac{(nn_{ij}-n_i n_{*j})^2}{n\cdot n_i n_{*j}}\dot{\sim}\chi^2(s-1)$$

检验方法：等价于两个总体分布的独立性检验

五、χ^2齐一性检验

	未患肺癌B	患肺癌$\bar{B}$	合计
不吸烟A	7775(7745.6)	42(71.4)	7817
吸烟$\bar{A}$	2099　(2128.4)	49(19.6)	2148
合计	9874	91	9965

独立性检验：若吸烟和患肺癌不相关，所以未患癌理论概率在两组人群（不吸烟和吸烟）中应该一致；

齐一性检验：假设两组人群（不吸烟和吸烟）分布相同，即未患癌理论概率在两组人群（不吸烟和吸烟）中应该一致。

五、χ^2齐一性检验

总结

➢χ^2独立性检验和χ^2齐一性检验，两者的区别在于原假设的不同：

◆卡方齐性检验的原假设是：分布相同

◆卡方独立性检验的原假设是：P(AB)=P(A)P(B)

➢两者都是用相同的卡方检验的拒绝规则：

$$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

其中，自由度=（I-1）×(J-1)

第八章作业（教材第五版）：

P215：1、2、3、4
P216： 6、7、8
P217： 11、14、15
P218： 18、22

注：作业不得抄袭；写上姓名、班级、学号和页码（如1/5），待第九章讲授结束后，与第九章作业一起提交至教学云平台。