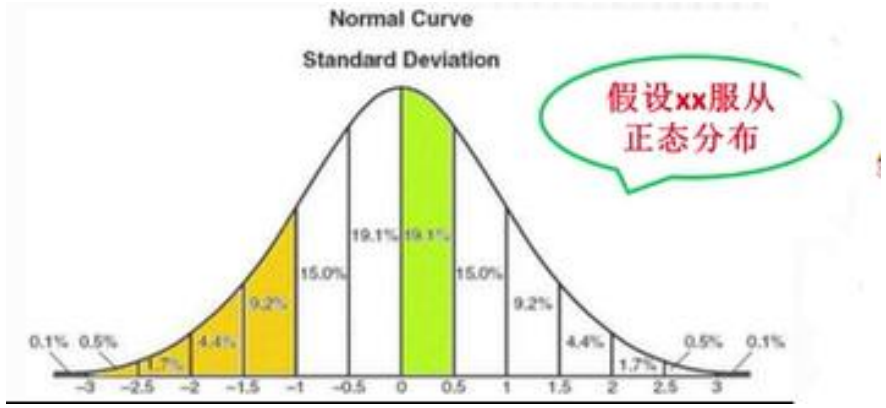
第五章 统计推断之假设检验

- 何为假设检验
- 参数假设检验
- 非参数假设检验

5.3 非参数假设检验

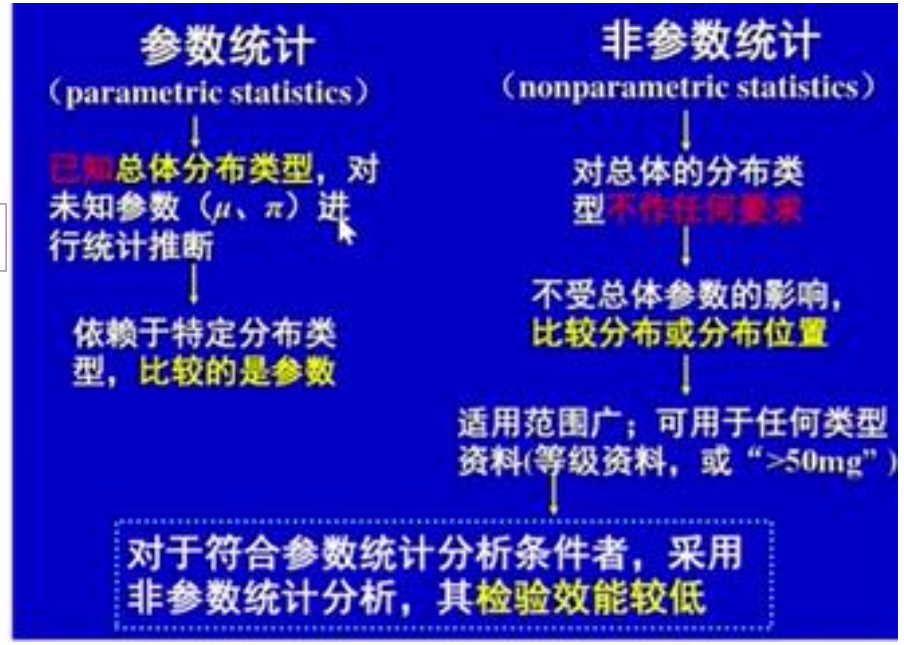
- 分布类型检验
  - 多样本独立性检验



非参数假设检验

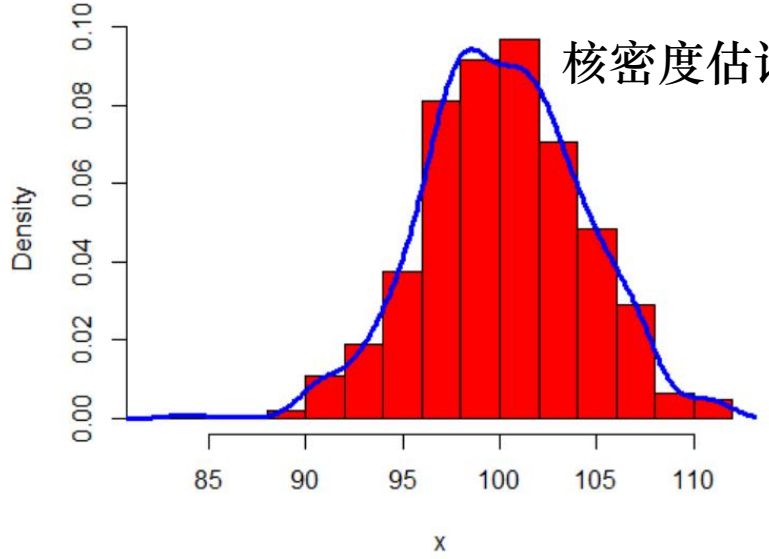
- 参数检验，通常基于知道总体的分布（比如正态分布）只需要通过样本对分布中的参数进行检验(比如均值检验)；
- 然而现实生活中很多情况并不知晓随机变量的分布类型；这种不假定总体分布的具体形式，尽量从数据本身来获得所需信息的统计方法---非参数假设检验。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$



正态分布??

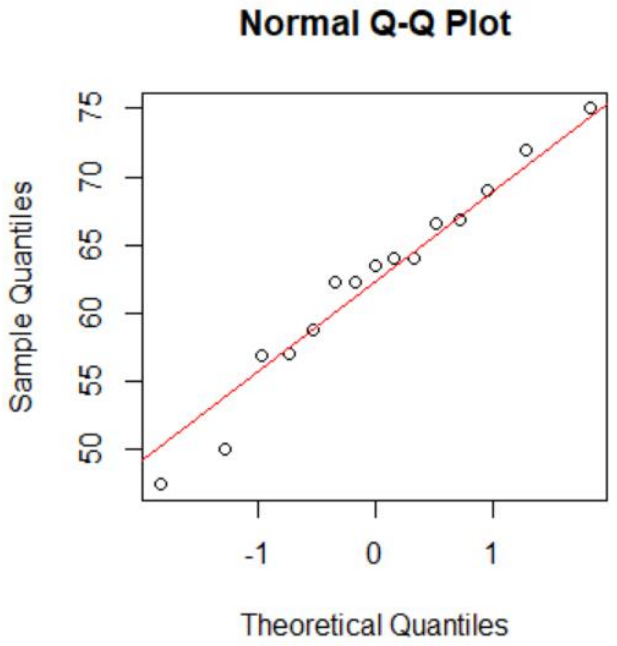
核密度估计曲线




```
density(x, bw = "nrd0",
        kernel = c("gaussian", "epanechnikov", "rectangular",
                   "triangular", "biweight", "cosine", "optcosine"),
        n = 512, from, to)
```

QQ图

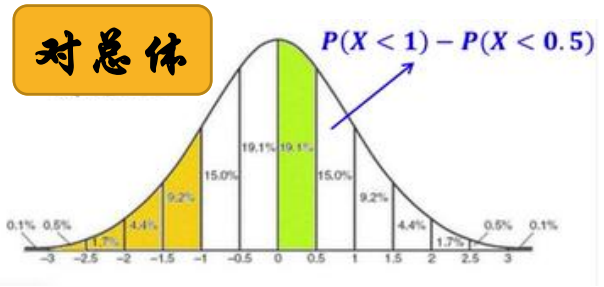
```
> qqnorm(weight)
> qqline(weight,col="red")
```



如何用统计检验?

分布检验-卡方检验

对总体



1. 理论分布 F 完全已知的情况

样本 $X_1, X_2, \cdots, X_n$

H_0 : X 具有分布 F.

H_1 : X 不具有分布 F.

在随机变量X的取值范围$[a,b]$(a可为$-\infty$, b可为∞)内选取$m-1$个实数$a = a_0 < a_1 < a_2 < \cdots < a_{m-1} < a_m = b$, 它们将$[a,b]$分为$m$个小区间$A_i = [a_{i-1}, a_i)$, 记$p_{i0} = F_0(a_i) - F_0(a_{i-1})$

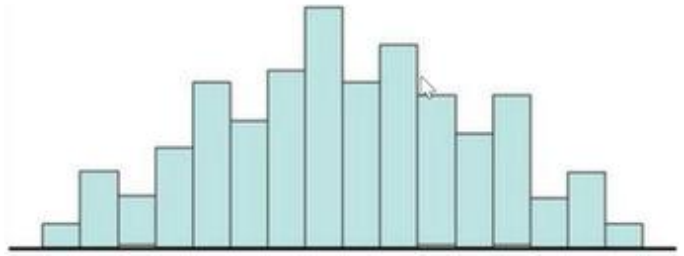
(1)总体分布已知;
(2)总体区间划分;
(3)每个区间对应的概率可以算出来

对样本

$$\sum_{i=1}^{m}(n_i - np_{i0})^2$$

理论与观察之间的偏离

设$(x_1, x_2, \cdots, x_n)$为来自总体$F(x)$的容量为n的一组样本观测值, n_i为观测值落入A_i的频数, $\sum_{i=1}^{m} n_i = n$. 若H_0成立, 则实际频数n_i与理论频数np_{i0}比较接近, 因此分布的拟合优度检验可转化为分类数据的实际频数与理论频数的一致性检验.



分布检验-卡方检验

理论与观察之间的偏离

- 构造统计量，判断分布

$$\sum_{i=1}^{m}(n_i - np_{i0})^2$$

平方和

Karl Pearson
1857.3～1936.4

英国数学家
生物统计学家

定理 7.2.1 (Pearson定理)

1) 若$F_0(x)$完全已知(不带有未知参数)，则当H_0成立时，统计量

$$\chi^2 = \sum_{i=1}^{m}\frac{(n_i - nP_{i0})^2}{nP_{i0}} \sim \chi^2(m-1).$$

实际频数

理论频率

R实现?

chisq.test( )的调用格式

```
chisq.test(x, y = NULL, correct =TRUE,p=rep(1/length(x),length(x)),
           rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

- 卡方检验根本思想是比较理论频数和实际频数的吻合程度或拟合优度问题。

分布检验-卡方检验

离散数据的卡方检验

每个点数出现的概率均为$\frac{1}{6}$（25次）

如果掷一骰子150次，并得到以下分布，则此骰子是均匀的吗？

点数	1	2	3	4	5	6	合计
出现次数	22	21	22	27	22	36	150

25

问题：是否符合均匀分布，关键是看观测值与期望值离得有多远

统计量：$X^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} \sim X^2(n-1)$

```
> freq=c(22,21,22,27,22,36)
> probs=c(1,1,1,1,1,1)/6
> chisq.test(freq,p=probs)

        Chi-squared test for given probabilities

data:  freq
X-squared = 6.72, df = 5, p-value = 0.2423
```

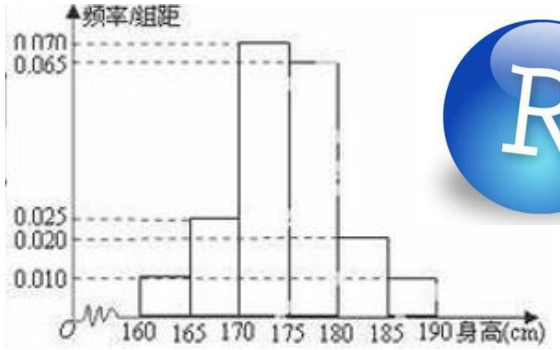
建立假设：

原假设：观测值=理论值

备择原假设：观测值与理论值不完全相等

结论：该骰子是均匀的

卡方检验—连续数据实例



- 假如要检验成人男性是否服从$N(170, 8^2)$分布，

  调查了20名男性的身高如下：

159.8	178.5	168.9	183.2	174.0	160.9	180.0	171.7	152.4	174.3
170.2	185.3	169.6	160.1	158.9	164.6	172.2	168.0	182.1	171.1

H_0: 成年男性身高服从$N(170, 8^2)$的正态分布

H_1: 成年男性身高不服从该正态分布

该检验依赖于分组，当不能拒绝原假设时，不能武断的承认原假设；离散化的过程会导致信息丢失。

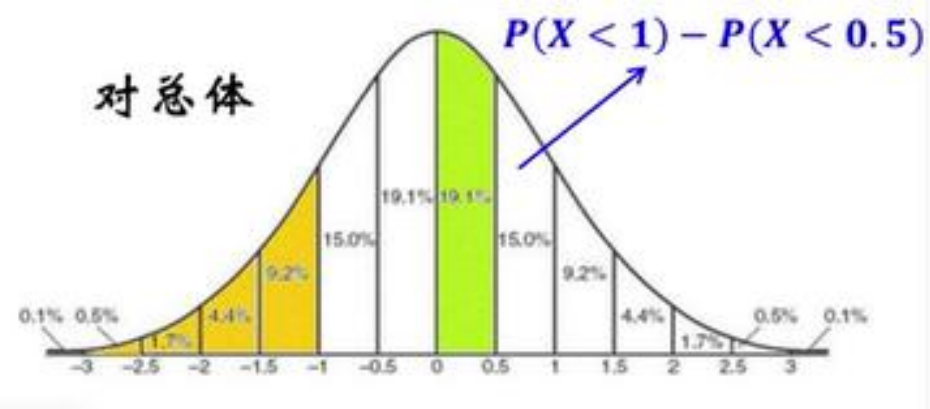
```
> x<-c(159.8,178.5,168.9,183.2,174.0,160.9,180.0,171.7,
+      152.4,174.3,170.2,185.3,169.6,160.1,158.9,164.6,
+      172.2,168.0,182.1,171.1)
> fn<-table(cut(x,breaks=c(min(x),160,170,180,190,max(x))))
> F<-pnorm(c(min(x),160,170,180,190,max(x)),170,8)
> P<-c(F[1],F[2]-F[1],F[3]-F[2],F[4]-F[3],1-F[4])
> chisq.test(fn,p=P)

        Chi-squared test for given probabilities

data:  fn
X-squared = 26.537, df = 4, p-value = 2.466e-05
```

对样本数据分组—离散化—实际频数

每个分组的理论概率（总体）

结论：拒绝原假设
身高不服从$N(170, 8^2)$的正态分布

卡方检验



2. 理论分布依赖于若干个未知参数的情况

$H: X$ 的分布属于 $\{F(x, \theta_1, \theta_2, \cdots, \theta_r)\}$

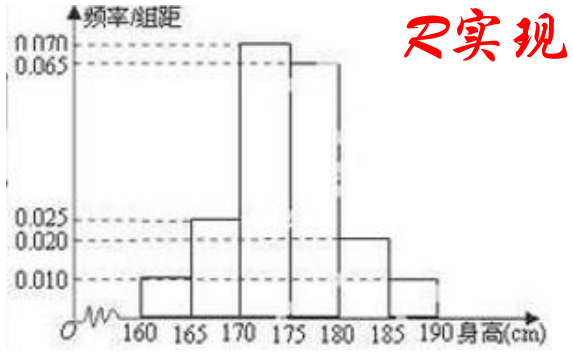
解决这个问题的步骤是，先通过样本作出 $(\theta_1, \theta_2, \cdots, \theta_r)$ 的极大似然估计 $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r)$ 再检验假设

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n\hat{P}_{i0})^2}{n\hat{P}_{i0}} \sim \chi^2(m - r - 1),$$

其中 r 表示未知参数的个数，m表示划分的区间（种类）数。

先用参数估计估计出分布的未知参数 → 再做卡方检验

卡方检验—实例



R实现

- 理论分布依赖于未知参数

20名男性身高如下，问该身高是否服从正态分布？

159.8	178.5	168.9	183.2	174.0	160.9	180.0	171.7	152.4	174.3
170.2	185.3	169.6	160.1	158.9	164.6	172.2	168.0	182.1	171.1

H_0: 成年男性身高服从正态分布
H_1: 成年男性身高不服从正态分布

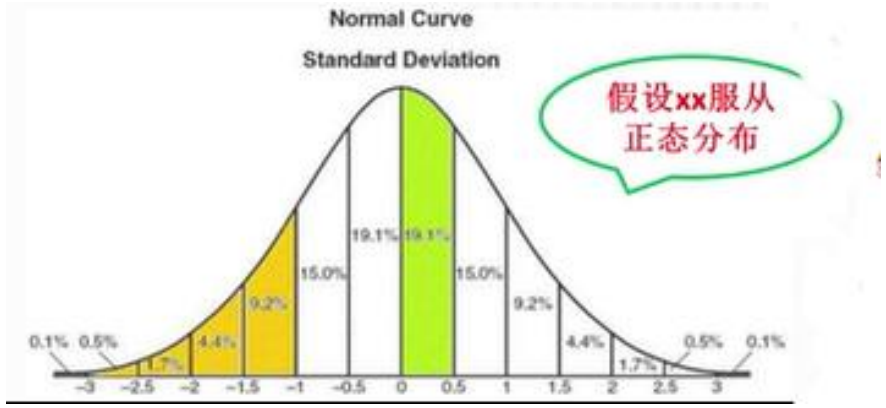
```
> mean.pop=mean(x)
> sd.pop=sqrt((length(x)-1)/length(x))*sd(x)
> fn<-table(cut(x,breaks=c(min(x),160,170,180,190,max(x))))
> F<-pnorm(c(min(x),160,170,180,190,max(x)),mean.pop,sd.pop)
> P<-c(F[1],F[2]-F[1],F[3]-F[2],F[4]-F[3],1-F[4])
> chisq.test(fn,p=P)

        Chi-squared test for given probabilities

data:  fn
X-squared = 21.028, df = 4, p-value = 0.0003126
```

点估计求总体均值和方差

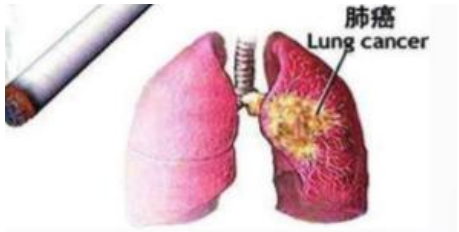
结论：拒绝原假设，身高不服从正态分布

5.3 非参数假设检验

- 分布类型检验
  - 多样本独立性检验



假设检验—卡方独立性检验



吸烟与肺癌有关吗？

⇕

吸烟与肺癌是否独立

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

若随机变量 X, Y 的分布函数分别为 $F_1(x)$和$F_2(y)$, 且联合分布为$F(x,y)$, 则X与Y的独立性归结为假设检验问题：

$$H_0 : F(x,y) = F_1(x)F_2(y) \longleftrightarrow H_1 : F(x,y) \neq F_1(x)F_2(y).$$

联合分布为$F(x,y)$是否等于两个独立分布$F_1(x)$与$F_2(x)$的乘积

假设检验—卡方独立性检验

(1) 构建列联表

		患癌 Y_1	不患癌 Y_2	$\cdots$	Y_s	总和
吸烟	X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
不吸烟	X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
总和		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	n

(2) 极大似然估计

吸烟概率 $\widehat{p_{i.}} = n_{i.}/n$　　$\widehat{p_{.j}} = n_{.j}/n$　　患癌概率

既吸烟又患癌概率 $\widehat{p_{i.}} \cdot \widehat{p_{.j}} = n_{i.} \cdot n_{.j}/n^2$

既吸烟又患癌的实际频率 $p_{ij} = n_{ij}/n$

令 $p_{ij} = P(X = X_i, Y = Y_j)$, $p_{i\cdot} = P(X = X_i)$, $p_{\cdot j} = P(Y = Y_j)$, $i, 1,2, \ldots, r$, $j = 1,2, \ldots, s$, 则 X 与 Y 的独立性检验就等价于下述检验：

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}, \forall 1 \le i \le r, 1 \le j \le s \leftrightarrow H_1: \exists (i,j), p_{ij} \ne p_i p_j$$

统计量：$x^2 = \sum_{i=1}^{r}\sum_{k=1}^{s}\left[n_{ij} - \frac{n_i n_j}{n}\right]^2 \Big/ \frac{n_i n_j}{n}$，$x^2$近似服从$x^2((r-1)(s-1))$

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - nP_{i0})^2}{nP_{i0}} \sim \chi^2(m-1).$$

chisq.test( )的调用格式

```
chisq.test(x, y = NULL, correct =TRUE,p=rep(1/length(x),length(x)),
           rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

假设检验—卡方独立性检验

为了研究吸烟是否与患肺癌有关，对63位肺癌患者及43名非肺癌患者（对照组）调查了其中的吸烟人数，得到2X2列联表

	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

H_0: 肺癌与吸烟相互独立；H_1: 肺癌与吸烟相关

```
> x<-c(60,3,32,11)
> dim(x)<-c(2,2)
> x
     [,1] [,2]
[1,]   60   32
[2,]    3   11
> chisq.test(x)

        Pearson's Chi-squared test with Yates' continuity correction

data:  x
X-squared = 7.9327, df = 1, p-value = 0.004855
```

结论：拒绝原假设，即认为肺癌与吸烟相关。

练习

练习1：调查某美发店上半年各月顾客数量，如下表所示，问该店每月的顾客数量是否服从均匀分布

月份	1	2	3	4	5	6	合计
顾客人数（百人）	27	18	15	24	36	30	120

练习2：从某地区高中二年级学生中随机抽取45位学生测得他们的体重如下表所示，问该地区同学的体重是否服从正态分布？

36	36	37	38	40	42	43	43	44	45	48	48	50	50	51
52	53	54	54	56	57	57	57	58	58	58	58	58	59	60
61	61	61	62	62	63	63	65	66	68	68	70	73	73	75