中国海洋大学

数学模型

专题：统计模型-回归模型

统计学的基本定义

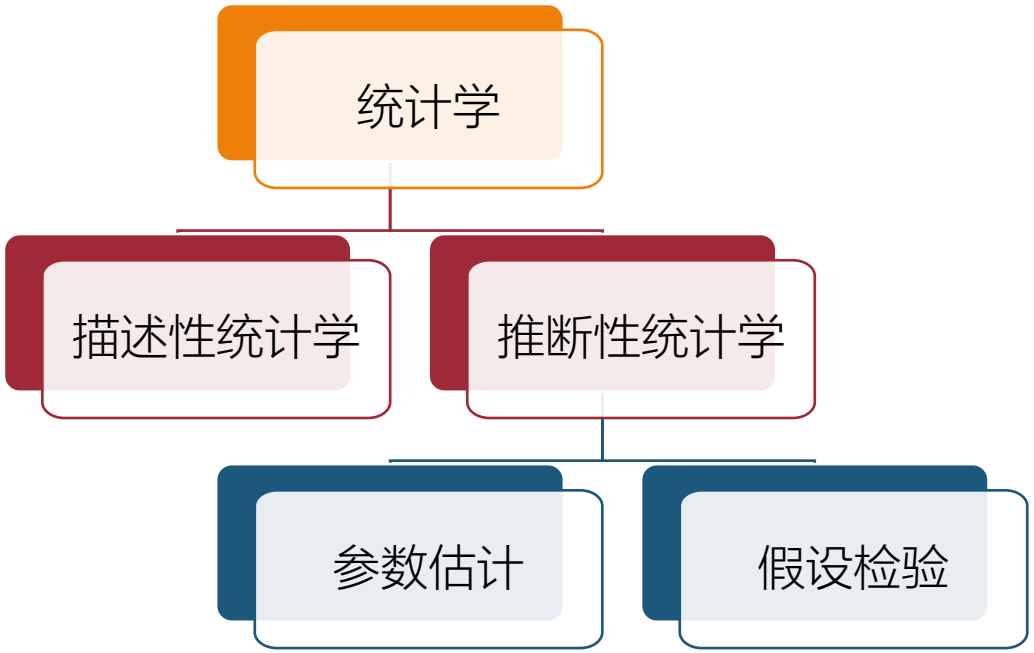
一、统计学的定义（Definition of Statistics）

定义：*Statistics is the art of learning from data.*

Data(数据) : information coming from observations, counts, measurements, or responses.

Statistics(统计学) : the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

二、统计学的分支（Branches of Statistics）



回归模型

- 描述性统计学
- 统计推断
- 回归模型
- 例：孕妇吸烟与胎儿健康
- 课堂练习

1、描述性统计学

要学会阅读数据

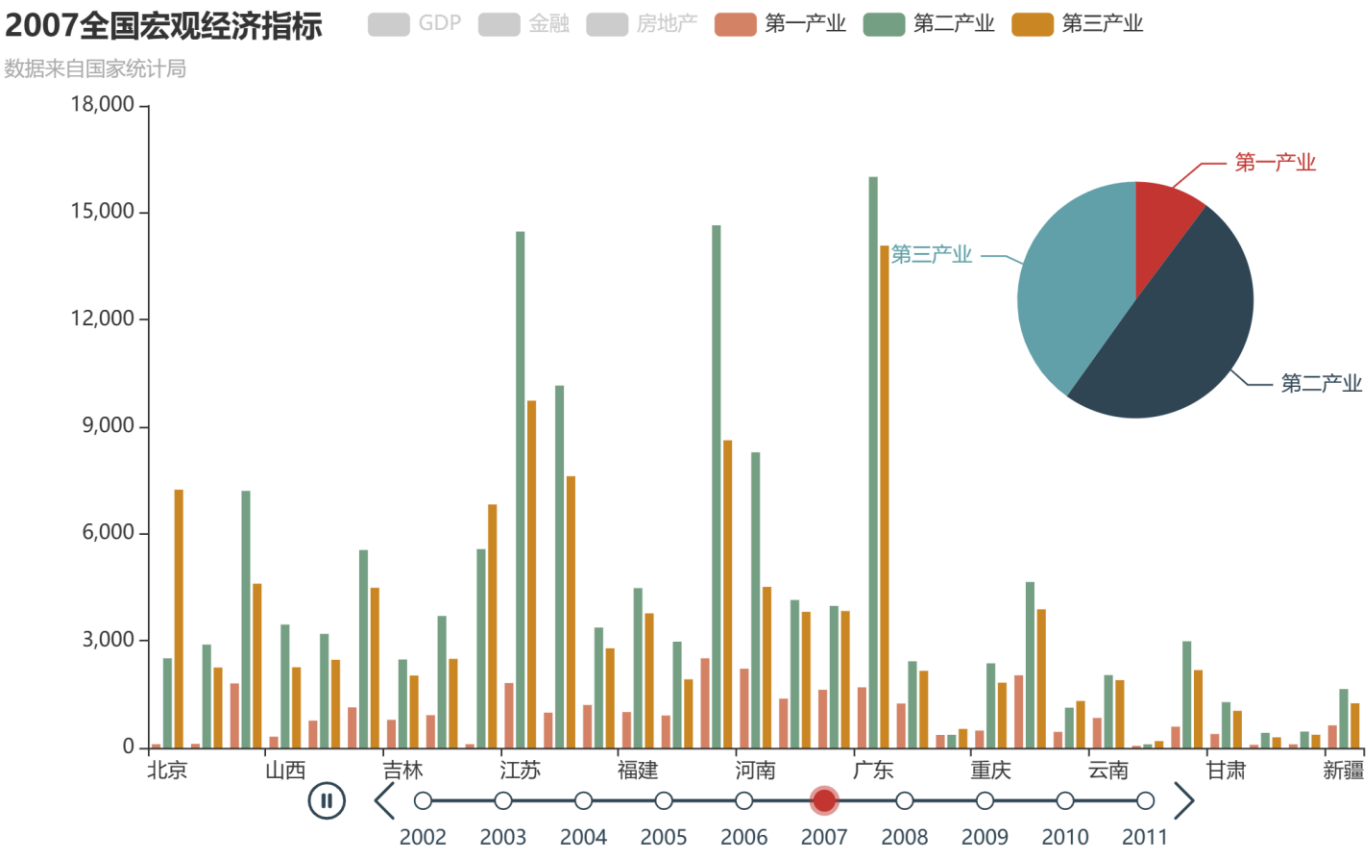
1、基本统计量：均值、方差、协方差、相关系数、分位数等

2、数据的图形表示

基本图表		直角坐标系图表	
Calendar:	日历图	Bar:	柱状图/条形图
Funnel:	漏斗图	Histogram:	直方图
Gauge:	仪表盘	Polygon:	折线图
Graph:	关系图	Boxplot:	箱形图
Liquid:	水球图	EffectScatter:	涟漪特效散点图
Parallel:	平行坐标系	HeatMap:	热力图
Pie:	饼图	Kline/Candlestick:	K线图
Polar:	极坐标系	Line:	折线/面积图
Radar:	雷达图	PictorialBar:	象形柱状图
Sankey:	桑基图	Scatter:	散点图
Sunburst:	旭日图	Overlap:	层叠多图

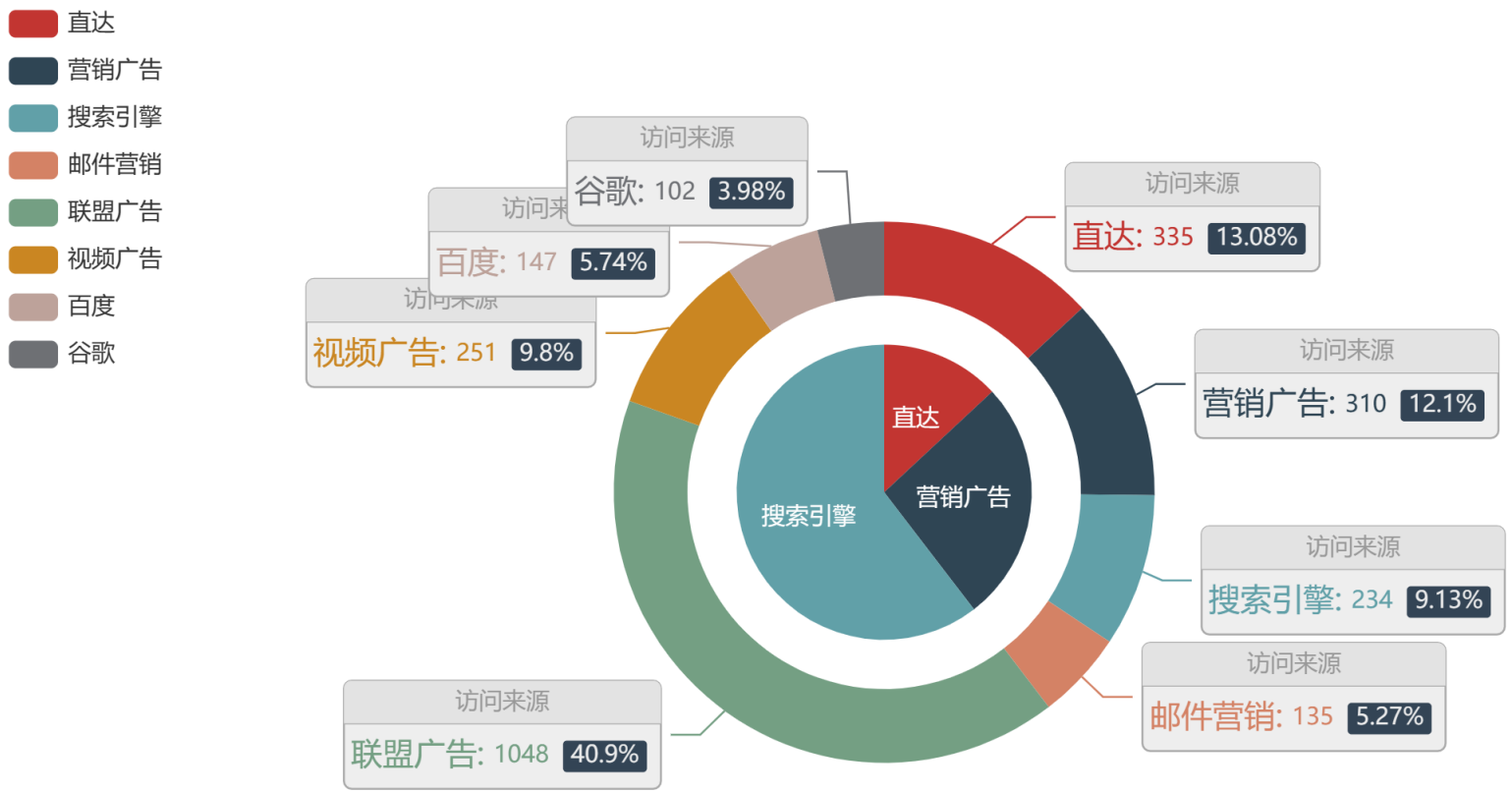
1、描述性统计学

Bar chart (条形图)



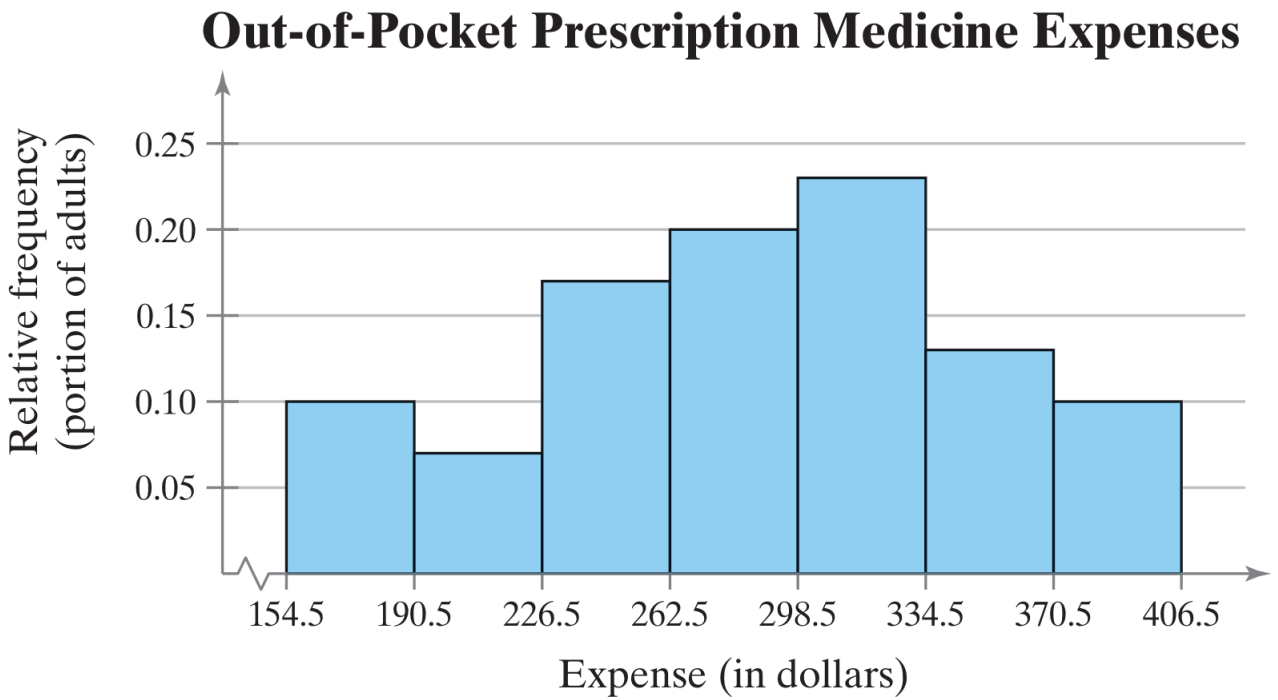
1、描述性统计学

Pie chart (饼图)



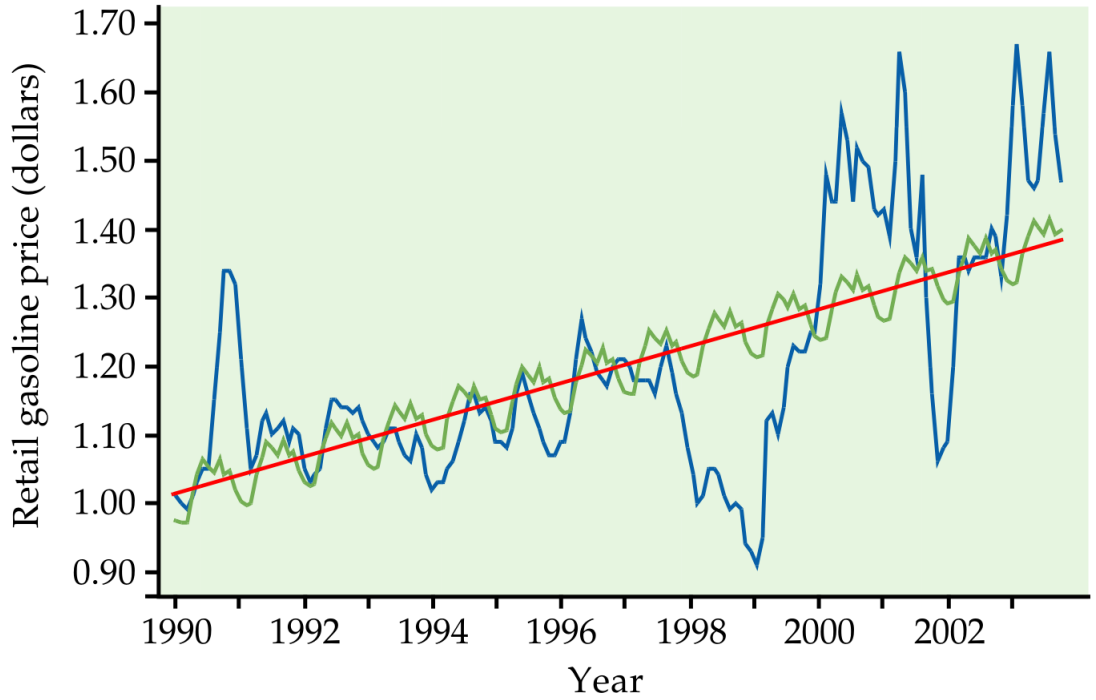
1、描述性统计学

Histogram(直方图)



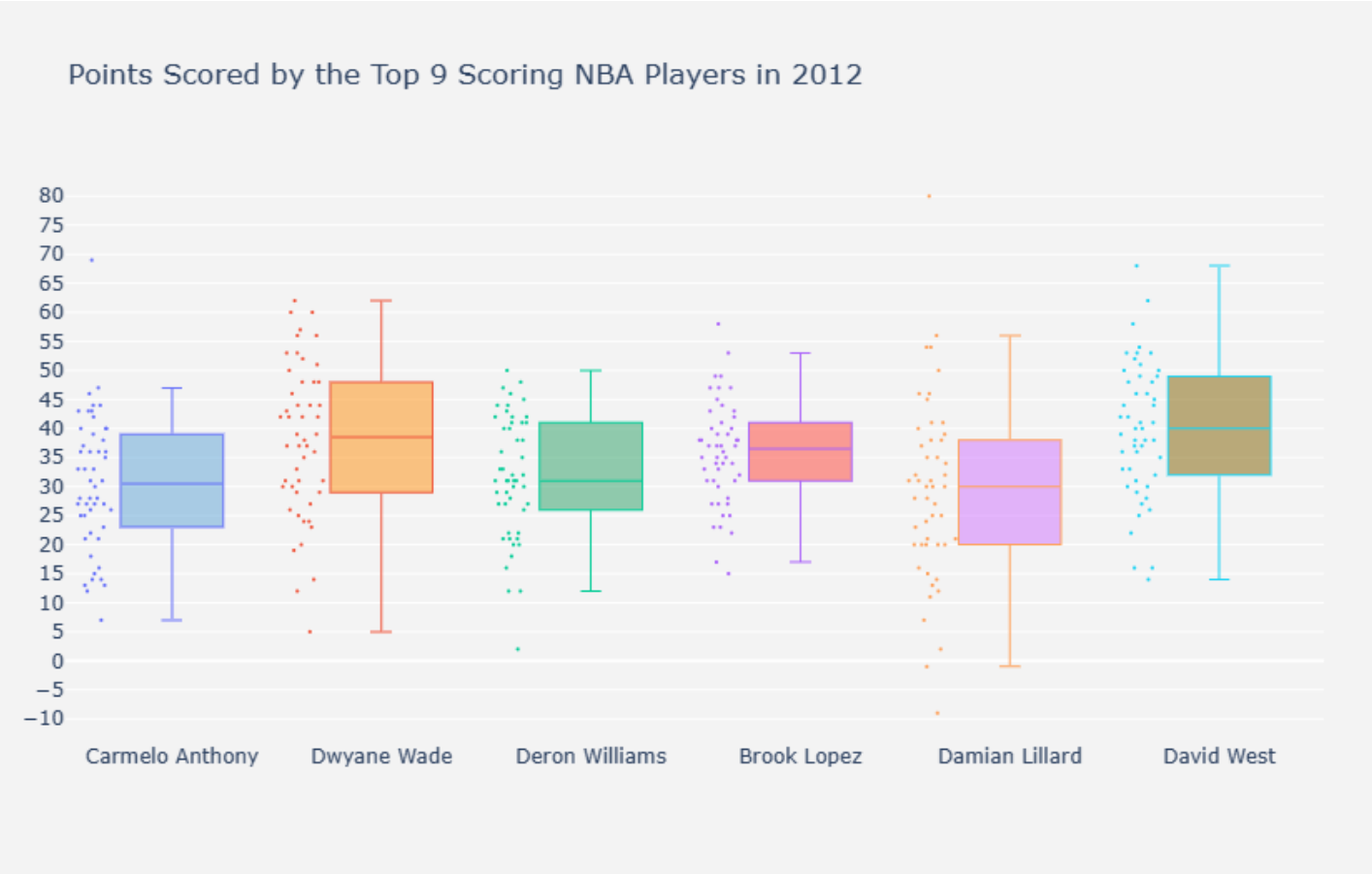
1、描述性统计学

Polygon (折线图)



1、描述性统计学

Box plot (箱线图)

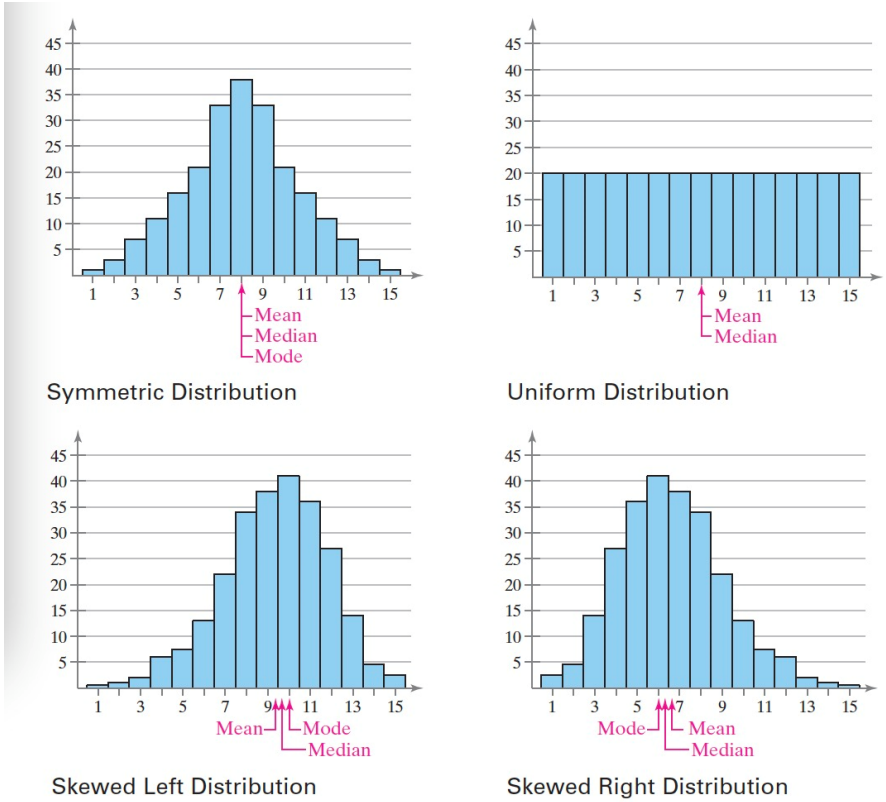


1、描述性统计学

分布的图形（Shape of Distributions）

- 对称（Symmetric）
- 均匀（Uniform）
- 偏态（Skewed： left / right）

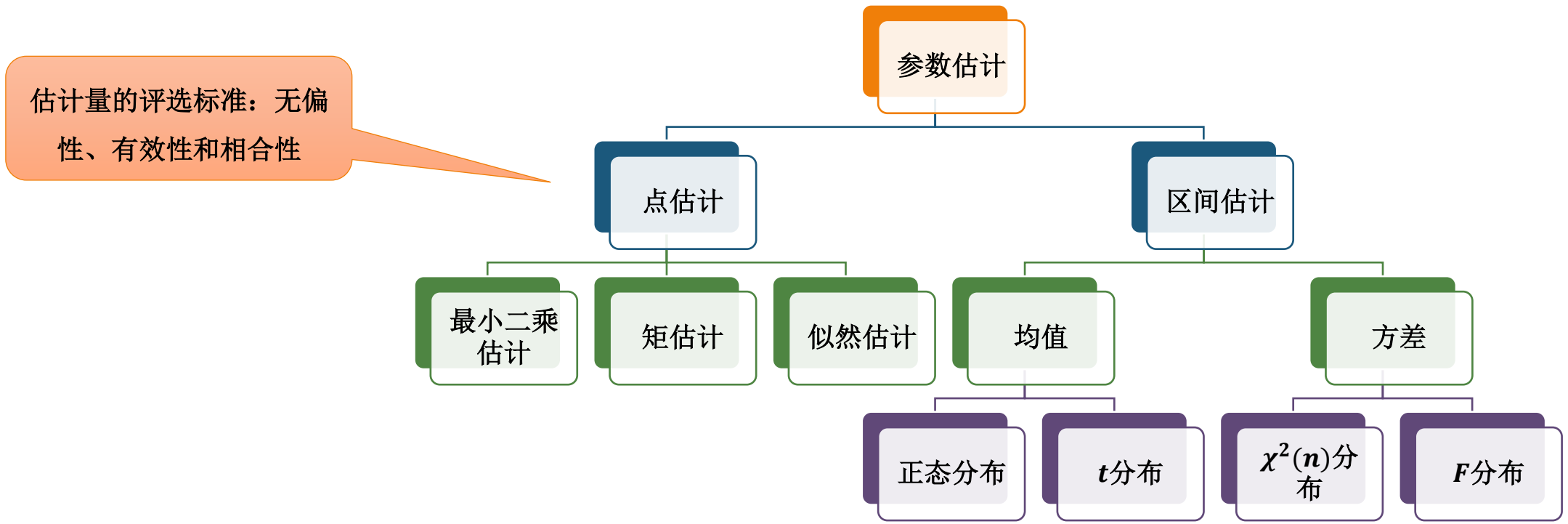
均值始终落在分布偏斜的方向上。



2、统计推断

参数估计

所谓参数估计，就是利用样本信息对总体数字特征作出推断和估计，即用样本估计量推断总体参数的具体数值或者一定概率保证下总体参数所属区间。



2、统计推断

假设检验

假设检验问题是统计推断的另一类重要问题.

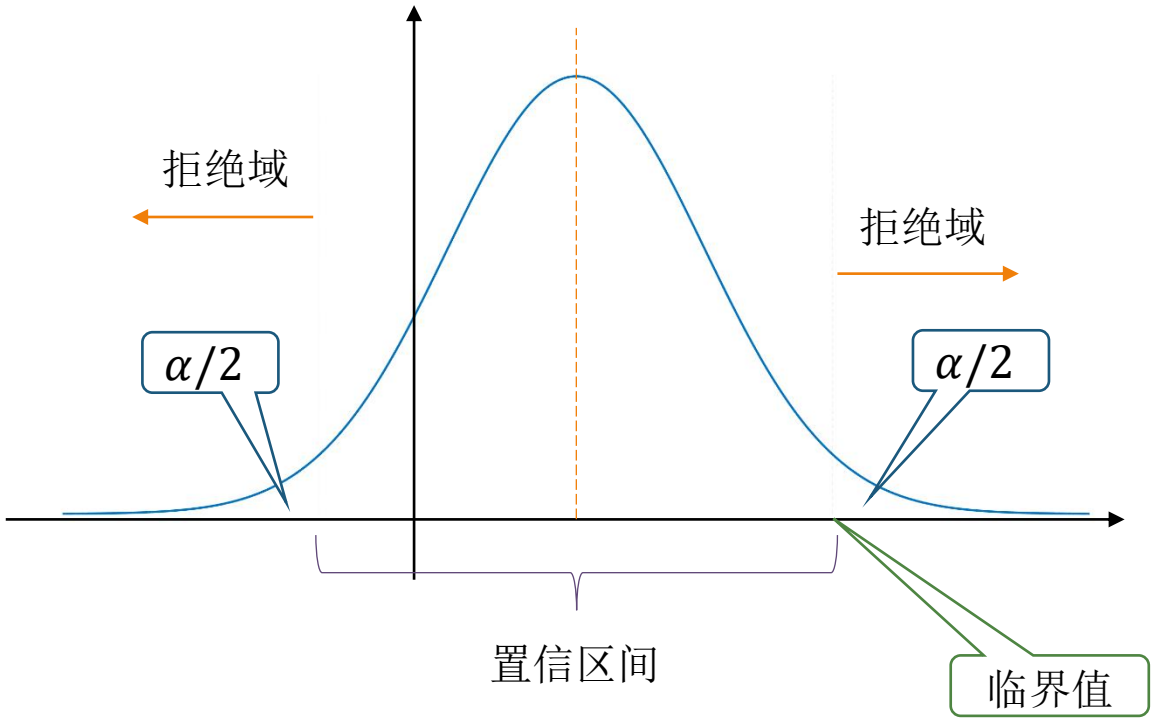
如何利用样本值对一个具体的假设进行检验?

通常借助于直观分析和理论分析相结合的做法，其基本原理就是

人们在实际问题中经常采用的所谓实际推断原理

“一个小概率事件在一次试验中几乎是不可能发生的”

2、统计推断



2、统计推断

方差已知情况下，单个正态总体的均值μ的假设检验（Z检验法）

函数　$ztest$

格式

$h = ztest(x, \mu_0, \sigma)$　　% x为正态总体的样本，μ_0为均值，σ为标准差，显著性水平为0.05(默认值)

$h = ztest(x, \mu_0, \sigma, \alpha)$　　%显著性水平为α

$[h, p, ci, zval] = ztest(x, \mu_0, \sigma, \alpha, tail)$ %p为观察值的概率，当p为小概率时则对原假设提出质疑，ci为真正均值μ的$1-\alpha$置信区间，$zval$为统计量的值。

说明　若$h=0$，表示在显著性水平α下，不能拒绝原假设；

若$h=1$，表示在显著性水平α下，可以拒绝原假设。

原假设：$\mu = \mu_0$，

tail=‘both’，表示备择假设：$\mu \neq \mu_0$（默认，双边检验）；

tail=‘right’，表示备择假设：$\mu > \mu_0$（右边检验）；

tail=‘left’，表示备择假设：$\mu < \mu_0$（左边检验）。

2、统计推断

例：某车间用一台包装机包装葡萄糖，包得的袋装糖重是一个随机变量，它服从正态分布。当机器正常时，其均值为0.5公斤，标准差为0.015。某日开工后检验包装机是否正常，随机地抽取所包装的糖9袋，称得净重为(公斤)

0.497, 0.506, 0.518, 0.524, 0.498, 0.511, 0.52, 0.515, 0.512

问机器是否正常？

2、统计推断

方差未知情况下，单个正态总体的均值μ的假设检验（t检验法）

函数　$ttest$

格式

$h = ttest(x, \mu_0)$　　% x为正态总体的样本，μ_0为均值，显著性水平为0.05(默认值)

$h = ttest(x, \mu_0, \alpha)$　　%显著性水平为α

$[h, p, ci, zval] = ttest(x, \mu_0, \alpha, tail)$　%p为观察值的概率，当p为小概率时则对原假设提出质疑，ci为真正均值μ的$1-\alpha$置信区间，$zval$为统计量的值。

说明　若$h=0$，表示在显著性水平α下，不能拒绝原假设；

若$h=1$，表示在显著性水平α下，可以拒绝原假设。

原假设：$\mu = \mu_0$，

tail=‘both’，表示备择假设：$\mu \neq \mu_0$（默认，双边检验）；

tail=‘right’，表示备择假设：$\mu > \mu_0$（右边检验）；

tail=‘left’，表示备择假设：$\mu < \mu_0$（左边检验）。

2、统计推断

方差未知情况下，单个正态总体的均值μ的假设检验（t检验法）

函数　$ttest$

格式

$h = ttest(x, \mu_0)$　　% x为正态总体的样本，μ_0为均值，显著性水平为0.05(默认值)

$h = ttest(x, \mu_0, \alpha)$　　%显著性水平为α

$[h, p, ci, zval] = ttest(x, \mu_0, \alpha, tail)$　%p为观察值的概率，当p为小概率时则对原假设提出质疑，ci为真正均值μ的$1-\alpha$置信区间，$zval$为统计量的值。

说明　若$h=0$，表示在显著性水平α下，不能拒绝原假设；

若$h=1$，表示在显著性水平α下，可以拒绝原假设。

原假设：$\mu = \mu_0$，

tail=‘both’，表示备择假设：$\mu \neq \mu_0$（默认，双边检验）；

tail=‘right’，表示备择假设：$\mu > \mu_0$（右边检验）；

tail=‘left’，表示备择假设：$\mu < \mu_0$（左边检验）。

2、统计推断

课堂练习：在平炉上进行一项试验以确定改变操作方法的建议是否会增加钢的产率，试验是在同一只平炉上进行的。每炼一炉钢时除操作方法外，其他条件都尽可能做到相同。先用标准方法炼一炉，然后用建议的新方法炼一炉，以后交替进行，各炼10炉，其产率分别为

（1）标准方法：78.1　72.4　76.2　74.3　77.4　78.4　76.0　75.5　76.7　77.3

（2）新方法：　79.1　81.0　77.3　79.1　80.0　79.1　79.1　77.3　80.2　82.1

设这两个样本相互独立，且分别来自正态总体和，均值和方差均未知。问建议的新操作方法能否提高产率？（取$\alpha = 0.05$）

3、回归模型

回归是研究因变量对自变量的依赖关系的一种统计分析方法，目的是通过自变量的给定值来估计或预测因变量的均值。它可用于预测、时间序列建模以及发现各种变量之间的因果关系。

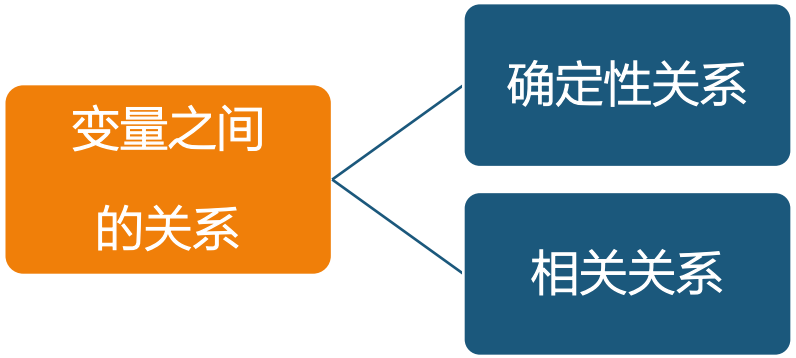
使用回归分析的益处良多，具体如下：

1）指示自变量和因变量之间的显著关系；

2）指示多个自变量对一个因变量的影响强度。

回归分析还可以用于比较那些通过不同计量测得的变量之间的相互影响，如价格变动与促销活动数量之间的联系。这有利于市场研究人员，数据分析人员以及数据科学家排除和衡量出一组最佳的变量，用以构建预测模型。

3、回归模型



$S = \pi r^2$：　　确定性关系

身高和体重：　　相关关系

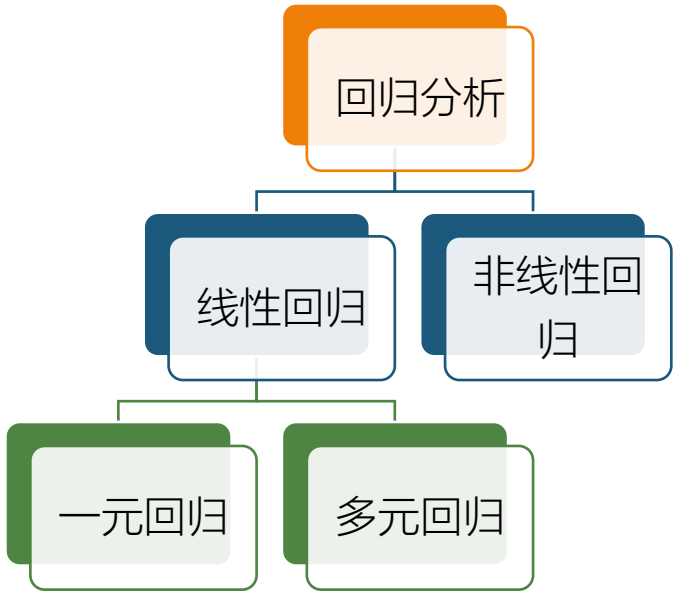
相关关系的特征是：变量之间的关系很难用一种精确的方法表示出来.

3、回归模型

确定性关系和相关关系的联系：

由于存在测量误差等原因，确定性关系在实际问题中往往通过相关关系表示出来；另一方面，当对事物内部规律了解得更加深刻时，相关关系也有可能转化为确定性关系.

回归分析——处理变量之间的相关关系的一种数学方法，它是最常用的数理统计方法.



3、回归模型

求解步骤

1. 推测回归函数的形式

- 根据专业知识或者经验公式确定;
- 作散点图观察.

2. 建立回归模型:$\hat{y} = f(X, \theta)$

3. 估计未知参数:$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y})^2$

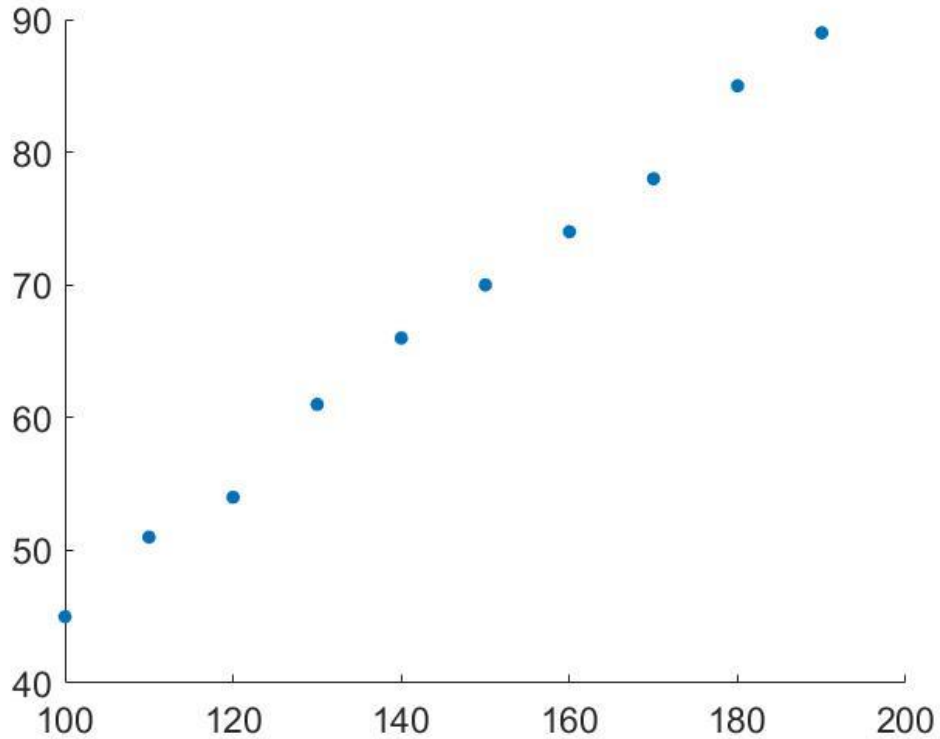
4. 模型诊断：残差的独立性、正态性检验(QQ图，JB检验，Lilliefors检验(KS检验的修正，适合大样本)等)。

3、回归模型

例:为研究某一化学反应过程中,温度x对产品得率Y(%)的影响, 测得数据如下:

温度x	100	110	120	130	140	150	160	170	180	190
得率Y	45	51	54	61	66	70	74	78	85	89

试确定温度和得率之间关系?



计算样本相关系数：0.9981

通过观察散点图，确立函数关系为

一元线性回归模型：

$$Y = ax + b + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

利用matlab回归函数fitlm(x,y)得：

$$Y = 0.48303x - 2.7394$$

4、孕妇吸烟与胎儿健康

吸烟有害健康！孕妇吸烟是否会伤害到腹中的胎儿？

- 对于新生儿体重，吸烟比妇女怀孕前身高、体重、受孕历史等因素的影响更为显著——美国公共卫生总署警告

美国儿童保健和发展项目(CHDS)提供的数据(1236个出生后至少存活28天男性单胞胎新生儿体重及其母亲的资料)

数据说明：						
2.新生儿体重(oz)	120	113	128	123	108	…
3.孕妇怀孕期(天)	284	282	279	999	282	…
4.新生儿胎次(1:第1胎,0:非第1胎)	1	0	1	0	1	…
5.孕妇怀孕时年龄	27	33	28	36	23	…
6.孕妇怀孕前身高(in)	62	64	64	69	67	…
7. 孕妇怀孕前体重(lb)	100	135	115	190	125	
8.孕妇吸烟状况(1:吸烟,0:不吸烟)	0	0	1	1	1	

研究目的

利用CHDS的数据建立新生儿体重与孕妇怀孕期、吸烟状况等因素的数学模型，定量地讨论：

- 对于新生儿体重来说，孕妇吸烟是否是比孕妇年龄、身高、体重等更为显著的决定因素；
- 孕妇吸烟是否会使早产率增加，怀孕期长短对新生儿体重有影响吗；
- 对每个年龄段来说，孕妇吸烟对新生儿体重和早产率的影响是怎样的。

问题背景及分析

美国公共卫生总署的警告容易受到人们的质疑：按照是否吸烟划分人群所做的研究，只能依赖于观测数据，而无法做人为的实验，很难确定新生儿体重的差别是因为吸烟，还是其它因素(如怀孕期长短、吸烟孕妇多是体重较轻的年青人等).

“孕妇吸烟可能导致胎儿受损、早产及新生儿低体重”的警告不如“吸烟导致肺癌”来得强，是由于对孕妇吸烟与胎儿健康间的生理学关系研究得不够.

参数估计

参数估计	不吸烟孕妇（ $n = 742$ ）	吸烟孕妇（$n = 484$）
新生儿体重均值的点估计	$\mu_{y0} = 123.0472$	$\mu_{y1} = 114.1095$
新生儿体重均值的区间估计	[121.7932,124.3011]	[112.4930,115.7260]
新生儿体重低比例的点估计	$r_0 = 0.0310$	$r_1 = 0.0826$
怀孕期均值的点估计	$\mu_{x0} = 280.1869(n = 733)$	$\mu_{x1} = 277.9792$
怀孕期均值的区间估计	$[278.9812, 281.3926]$	$[276.6273, 279.3311]$
早产率(≤ 37周)的点估计	$q_0 = 0.0764$	$q_1 = 0.0854$

- 吸烟比不吸烟孕妇新生儿体重平均低9 oz （250g ），新生儿体重低(≤ 2500)的比例明显高.
- 吸烟比不吸烟孕妇怀孕期平均短2天, 早产率(≤ 37周)差不多.
- 新生儿体重和怀孕期的差别在统计学上是否显著？

假设检验

假设检验	假设	检验结果（α =0. 05）
新生儿体重均值	$H_0: \mu_{y0} \leq \mu_{y1}, H_1: \mu_{y0} > \mu_{y1}$	拒绝H_0, 接受H_1
新生儿体重低比例	$H_0: r_0 \geq r_1, H_1: r_0 < r_1$	拒绝H_0, 接受$H_1(t = 4.0304)$
怀孕期均值	$H_0: \mu_{x0} \leq \mu_{x1}, H_1: \mu_{x0} > \mu_{x1}$	拒绝H_0, 接受H_1
早产率	$H_0: q_0 = q_1, \quad H_1: q_0 \neq q_1$	接受H_0, 拒绝$H_1(t = 0.5663)$

- 吸烟孕妇的新生儿体重比不吸烟孕妇的低、且新生儿体重低（≤ 2500）的比例高，在统计学上有显著意义.

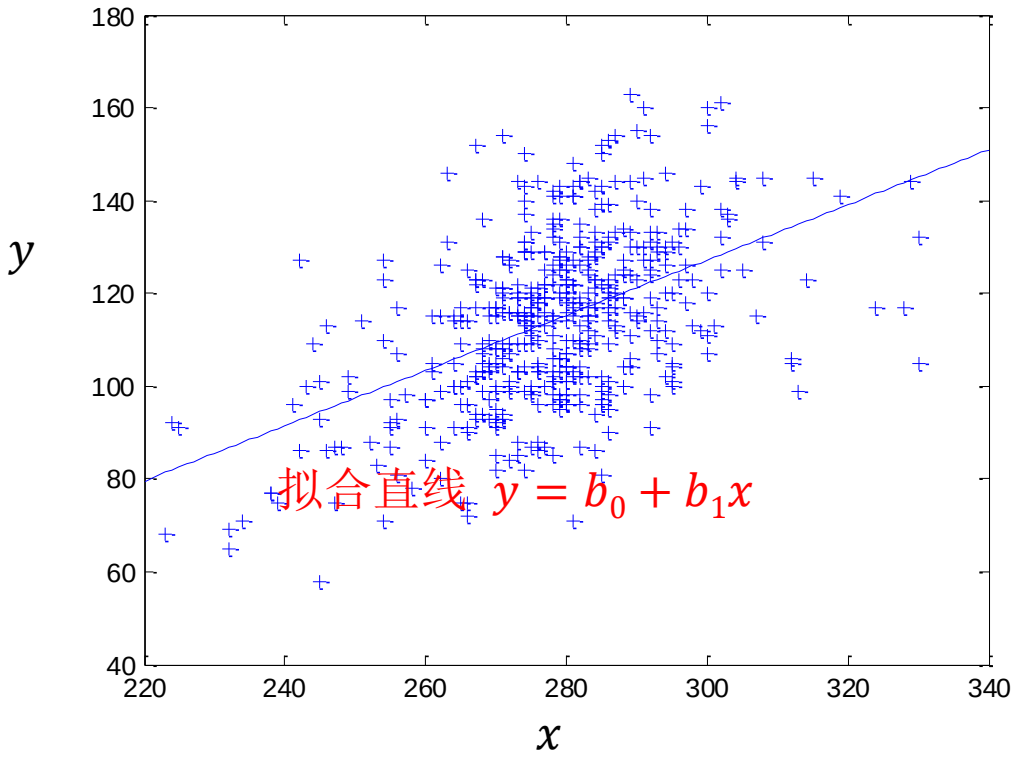
吸烟与不吸烟孕妇孕期和早产率（≤ 37周）的差别难以肯定是显著的（若$\alpha = 0.01$将接受怀孕期均值相等的假设）

一元线性回归分析

假设检验结果：孕妇吸烟状况对新生儿体重大小有显著影响，但是对怀孕期长短的影响难以确定。

- 新生儿体重与怀孕期的关系如何？

480位吸烟孕妇的怀孕期x和新生儿体重y



直线$y = b_0 + b_1 x$描述了数据的变化趋势，但是拟合得不好.

- 怎样衡量由拟合得到的模型的有效性？
- 模型系数精确度和模型预测的数值范围多大？

模型求解, 模型检验

一元线性回归模型 $y = b_0 + b_1 x + \varepsilon$，（怀孕期$x$，新生儿体重$y$）

随机变量ε是除x外，影响y的随机因素的总和，对于不同的x，ε相互独立且服从$N(0, \sigma^2)$分布.

480位吸烟孕妇数据 x, y ⇨

系数	系数估计值	系数置信区间
b_0	-51.2983	$[-77.5110 \quad -25.0856]$
b_1	0.5949	$[0.5008 \quad 0.6891]$
$R^2 = 0.2438, F = 154, p < 0.0001, s^2 = 249$		

- b_1置信区间不含零点，$F = 154 >> F(1, n-2) = 3.8610\ (\alpha = 0.05)$，应拒绝$H_0$: $b_1 = 0$的假设，模型有效。
- b_1置信区间较长，决定系数R^2较小(y的24.38%由x决定)，剩余方差s^2较大，模型的精度不高.

模型解释、模型预测

一元线性回归模型：$y = b_0 + b_1 x + \varepsilon$(怀孕期$x$，新生儿体重$y$)

- $\hat{b}_1 = 0.5949$：吸烟孕妇怀孕期增加一天，新生儿体重平均增加约0.6 oz.
- $\hat{b}_0 = -51.2983$：不是$x = 0$时y的估计，只能在数据范围内($x = 220 \sim 340$天）估计.

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = -51.2983 + 0.5949x$$

若怀孕期$x = 280$天，新生儿体重$\hat{y} = 114.5937\ oz$, 预测区间为

$$[88.0949, 141.0925]$$

- 模型精度不高导致预测区间如此之大！

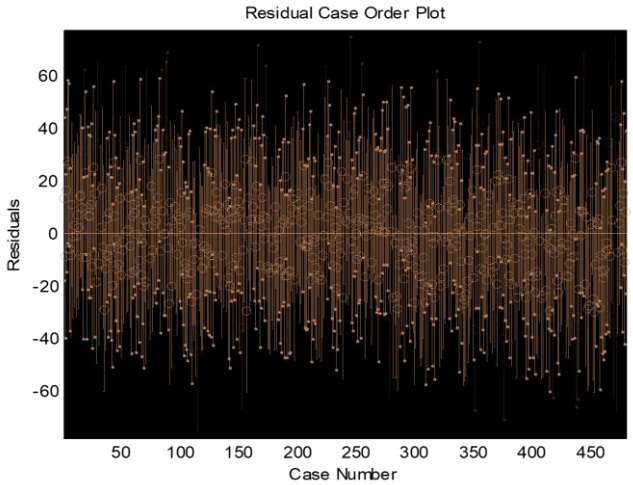
残差分析

一元线性回归模型 $y = b_0 + b_1 x + \varepsilon$(怀孕期$x$，新生儿体重$y$)

模型残差$e = y - \hat{y}$：误差ε的估计值(均值为0的正态分布)

若数据残差的置信区间不含零点，称为异常点(偏离整体数据的变化趋势)，应剔除，剔除后重新估计参数

系数	系数估计值	系数置信区间
b_0	-53.6126	$[-77.0606 \quad -30.1645]$
b_1	0.6007	[0.5164 0.6850]
$R^2 = 0.3040 \quad F = 196 \quad p < 0.0001 \quad s^2 = 182$		



虽然b_0和b_1的估计值变化不大，但置信区间变短，且R^2和F变大，s^2减小，说明模型精度得到提高.

一元线性回归模型　$y = b_0 + b_1 x + \varepsilon$

690位不吸烟孕妇数据x, y（剔除异常点后）⇨

系数	系数估计值	系数置信区间
b_0	33.5330	[14.9989　52.0671]
b_1	0.3201	[0.2541　0.3860]
$R^2 = 0.1165$　$F = 90$　$p < 0.0001$　$s^2 = 181$		

- $\hat{b}_1 = 0.3201$不吸烟孕妇怀孕期增加一天，新生儿体重平均只增加$0.32oz$.
- 对吸烟孕妇是增加约$0.6oz$，二者相差很大！

将吸烟状况作为另一自变量，建立新生儿体重与2个自变量的回归模型，利用全体孕妇数据进行分析.

多元线性回归分析

模型 $y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$

y：新生儿体重，

x_1：孕妇怀孕期，

$x_2 = 0,1$：不吸烟，吸烟.

1145位全部孕妇数据（剔除异常点后）

$$\hat{y} = 0.7698 + 0.4365 x_1 - 8.7610\, x_2$$

- $\hat{b}_1 = 0.4365$：对于吸烟状况x_2相同的孕妇，x_1增加一天y平均增加$0.44oz$. 在吸烟孕妇的0.6与不吸烟孕妇的$0.32oz$之间.
- $\hat{b}_2 = -8.7610$：x_1相同时，吸烟比不吸烟孕妇的新生儿体重平均约低$8.8oz$. 与参数估计的数值相同，但增加了x_1相同的条件.

多元线性回归分析

模型

$$y = b_0 + b_1x_1 + b_2x_2 + \varepsilon$$

$$\longrightarrow y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + \varepsilon$$

增加乘积项x_1x_2：x_1和x_2对y的综合影响

系数	系数估计值	系数置信区间
b_0	34.0925	[15.4605,52.7244]
b_1	0.3181	[0.2517,0.3844]
b_2	−87.0738	[−116.9656, −57.1820]
b_3	0.2804	[0.1734,0.3875]
$R^2 = 0.2766\,, F = 145, p < 0.0001, s^2 = 183$		

模型有效，但是R^2较小，s^2较大，仍有改进余地.

$$\hat{y} = 34.0925 + 0.3181x_1 - 87.0738\,x_2 + 0.2804x_1\,x_2$$

$x_2 = 0 \Rightarrow \hat{y} = 34.0925 + 0.3181x_1$:不吸烟孕妇的一元模型

$x_2 = 1 \Rightarrow \hat{y} = -52.9813 + 0.5985x_1$:吸烟孕妇的一元模型

变量选择与逐步回归

- CHDS提供的数据中除孕妇怀孕期和吸烟状况外,还有孕妇怀孕时的年龄、体重、身高和胎次状况.

新生儿体重模型中是否应该加入其他的自变量?

变量选择 ： 从应用的角度希望将所有影响显著的自变量都纳入模型，又希望最终的模型尽量简单.

逐步回归 ： 迭代式的变量选择方法.

- 利用CHDS数据提供的全部信息，通过逐步回归方法选择变量，建立新生儿体重的线性回归模型.

逐步回归

用逐步回归方法建立新生儿体重y的线性回归模型

x_1（孕妇怀孕期），x_2（胎次状况），x_3（年龄），x_4（身高），x_5（体重），x_6（吸烟状况）组成候选变量集合S.

- 选取x_1, x_6为初始子集S_0
- 从S_0外的S中引入一个对y影响最大的x，$S_0 \to S_1$.
- 对S_1中的x进行检验，移出一个影响最小的，$S_1 \to S_2$.
- 继续进行，直到不能引入和移出为止 .
- 引入和移出都以给定的显著性水平为标准.
- 显著性水平取缺省值（引入$\alpha = 0.05$，移出$\alpha = 0.10$）

逐步回归

用逐步回归方法建立新生儿体重y的线性回归模型

$$\hat{y} = -80.7132 + 0.4441x_1 - 3.2876x_2 + 1.1550x_4 + 0.0498x_5 - 8.3939x_6$$

x_1（怀孕期），x_2（胎次状况），x_4（身高），x_5（体重），x_6（吸烟状况）.

$$\hat{b}_6 = -8.3939, \hat{b}_1, \hat{b}_4, \hat{b}_5 > 0, \hat{b}_2 = -3.2876$$

- x_1, x_2, x_4, x_5相同时，吸烟比不吸烟孕妇的新生儿体重平均低$8.4\ oz$.
- 孕妇的怀孕期、身高、体重对新生儿体重的影响是正面的.
- 第1胎新生儿体重比非第1胎平均约低$3.3\ oz$（第1胎$x_2 = 1$）.

相关分析

y和各自变量的相关系数矩阵							
	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1.0000	0.4075	−0.0439	0.0270	0.2037	0.1559	−0.2468
x_1		1.0000	0.0809	−0.0534	0.0705	0.0237	−0.0603
x_2			1.0000	−0.3510	0.0435	−0.0964	−0.0096
x_3				1.0000	−0.0065	0.1473	−0.0678
x_4					1.0000	0.4353	0.0175
x_5						1.0000	−0.0603
x_6							1.0000

- 与y相关性较强的是怀孕期x_1，吸烟状况x_6，身高x_4.
- 自变量间相关性较强的有：孕妇体重x_5与身高x_4的正相关；年龄x_3与胎次状况x_2的负相关(年龄越大第1胎$x_2 = 1$越少).

当几个自变量间有较强相关性时，删除多余的只保留一个不会对模型有效性和精确度有多大影响.

相关分析

不同年龄段孕妇吸烟对新生儿体重的影响

孕妇按年龄分组建立y与x_1, x_2, x_4, x_5, x_6的回归模型

	小于25岁	25~30岁	30~35岁	大于35岁
b_0	−66.3893	−39.1296	−157.1307	−130.1740
b_1(怀孕期)	0.3972	0.3521	0.5951	0.6728
b_2	−0.9978	−7.4124	−0.0932	−4.1835
b_4	1.2144	0.8409	1.6828	0.8747
b_5	−0.0021	0.0959	0.0557	0.0732
b_6(吸烟状况)	−8.4119	−8.2656	−10.5411	−6.4008
R^2	0.2549	0.2330	0.3394	0.3136
s^2	211.6359	239.7201	272.6021	304.7208
n	444	362	211	157

对于x_1和x_6两个影响y的主要因素，30岁以下两组结果差别不大，而与30岁以上两组则有一定差异.

课堂练习

风暴数据包含龙卷风、雷暴、洪水、闪电、极端温度和其他天气现象。下表总结了 1953 年至 2005 年间美国每年的龙卷风次数。

(a) 按年份绘制龙卷风总数趋势图。 试判断线性趋势是否合理?

(b) 是否存在异常值或异常模式? 解释你的答案。

(c) 运行简单线性回归并总结结果

(d) 计算残差并出其随年份的趋势图

(e) 如何说明残差是否正常?

(f) 如果用二次回归结果如何?

年份	次数	年份	次数	年份	次数	年份	次数
1953	421	1967	926	1981	783	1995	1235
1954	550	1968	660	1982	1046	1996	1173
1955	593	1969	608	1983	931	1997	1148
1956	504	1970	653	1984	907	1998	1449
1957	856	1971	888	1985	684	1999	1340
1958	564	1972	741	1986	764	2000	1076
1959	604	1973	1102	1987	656	2001	1213
1960	616	1974	947	1988	702	2002	934
1961	697	1975	920	1989	856	2003	1372
1962	657	1976	835	1990	1133	2004	1819
1963	464	1977	852	1991	1132	2005	1194
1964	704	1978	788	1992	1298		
1965	906	1979	852	1993	1176		
1966	585	1980	866	1994	1082		