

深圳北理莫斯科大學

УНИВЕРСИТЕТ МГУ-ППИ В ШЭНЬЧЖЭНЕ

SHENZHEN MSU-BIT UNIVERSITY

Математическое моделирование и исследование моделей с помощью математических программ

数学建模及数学软件的使用

Лекция № 10 (统计)

张晔

ye.zhang@smbu.edu.cn

Статистика/ Statistics/ 统计学

数理统计学派 vs 社会统计学派

Теория вероятностей изучает математические модели случайных явлений. Математическая статистика решает обратные задачи: разрабатывает различные методы, которые позволяют по статистическим данным, которые носят случайный характер, подобрать подходящую теоретико – вероятностную модель.

主要术语

- **统计学(statistics): 收集、处理、分析、解释数据并从数据中得出结论的科学。**
- **描述统计(descriptive statistics): 研究数据收集、处理和描述的统计学方法。**
- **推断统计（inferential statistics）：研究如何利用样本数据来推断总体特征的统计学方法。**
- **均值（mean）：均值也就是平均数，有时特指算术平均数，这是相对其他方式计算的均值，求法是先将所有数字加起来，然后除以数字的个数，这是测量集中趋势，或者说平均数的一种方法。**
- **中位数（median）：也就是选取中间的数，要找中位数，首先需要从小到大排序，排序后，再看中间的数字是什么。**
- **众数（mode）：众数也就是数据集中出现频率最多的数字。**

总体与个体

定义 在数理统计中，将所研究对象的全体称为**总体**(母体), 其中每个对象称为**个体**。

- 通常关注的是研究对象的某些个数量指标, 因此也称这些数量指标取值的全体为总体, 其中每个元素称为个体.

- 例如, 检验灯泡厂生产的灯泡寿命:受检的全体灯泡就是总体, 每个灯泡就是个体。也可理解：全体灯泡寿命数值构成总体，每个灯泡的寿命数值为一个体。

- 又如, 调查 深北莫 男生身高情况：深北莫全体男生就是总体，每个深北莫男生就是一个个体。
- 也可理解：全体深北莫男生身高数值构成总体，每个深北莫男生身高数值就是一个个体。

- 灯泡的寿命检验是一个**破坏性试验**, 即当得知一个灯泡寿命时, 该灯泡的使用价值也就消失了. 因此, **不可能抽检每个灯泡!**

- 可以逐一测量每个工大男生的身高, 但**工作量大**. 而我们仅需对深北莫男生身高情况有个大致了解, 因此, **不必要抽测每个深北莫男生!**

做法　从总体中随机地抽取若干个体(灯泡、深北莫男生)，测试其所需数据(寿命、身高)，最后对所得数据通过整理加工和分析来推断总体(这批灯泡寿命、深北莫男生身高)的分布情况，从而了解整体情况.

一般，我们所研究的总体的某项数量指标 X 是一个随机变量，其取值在客观上有一定的分布. 因此，对总体的研究，就是对相应的随机变量 X 的研究。

今后，我们称 X 的分布函数和数字特征分别为总体的分布函数和数字特征，并不再区分总体与相应的随机变量X. 对总体的称呼：总体，总体X与总体F.

总体：Population/ Генеральная совокупность

样本与样本值

数理统计的基本任务就是通过对从总体中抽取的一部分个体(称为总体的样本)进行观察,根据所记录的数据(样本值)经整理与加工,以推断总体的某些性质.

“从总体中抽取一个个体”就是对总体进行一次观察(试验),并记录其数据结果.

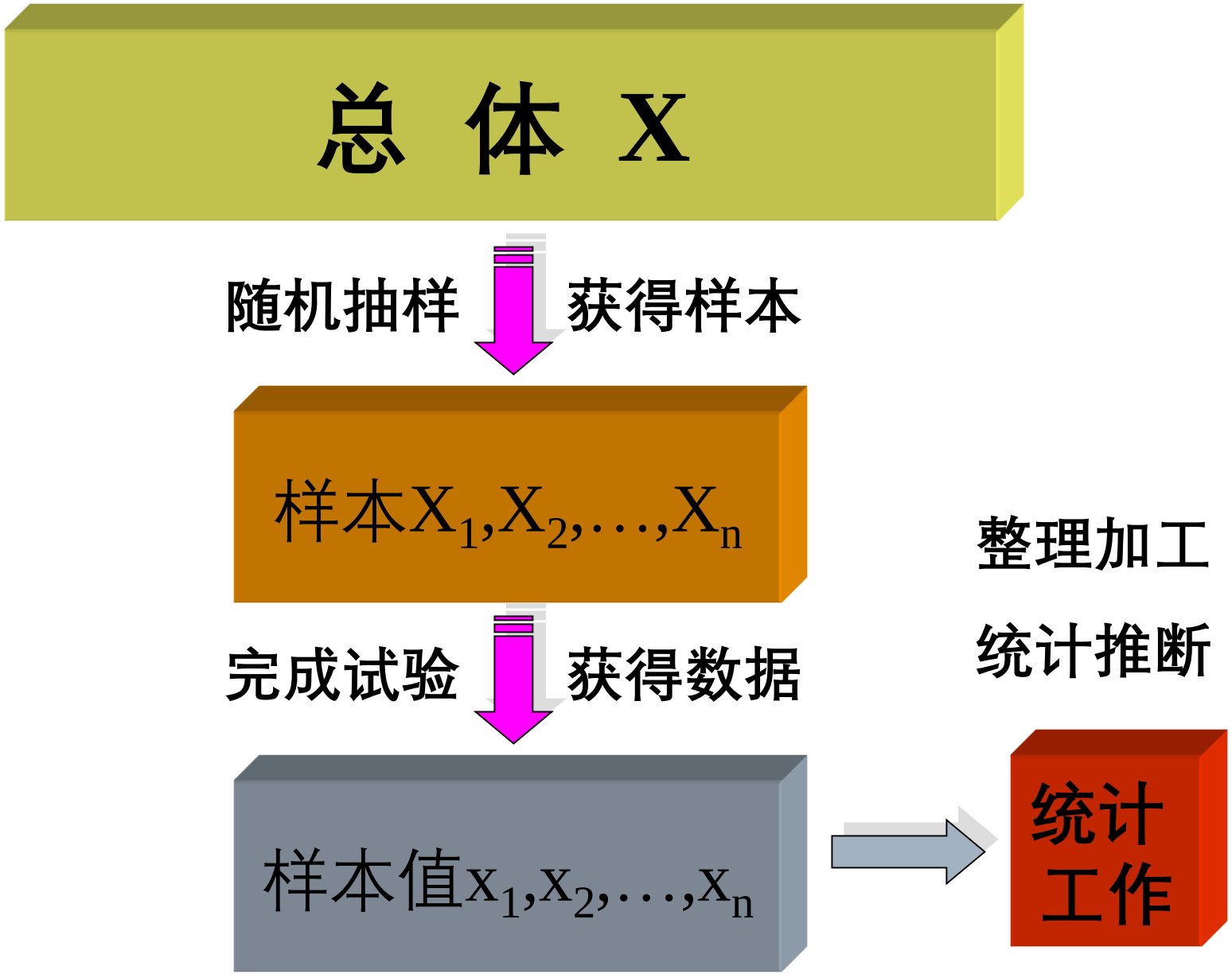
在相同条件下对总体X进行n次独立、重复的观察,将n次试验结果依次记为 $X_1, X_2, \cdots, X_n$,则称之为来自总体X的容量为n的一个简单随机样本;n次试验完成后所得样本的一组观察值 $x_1, x_2, \cdots, x_n$ 称为样本值.

样本：Sample/ Выборка

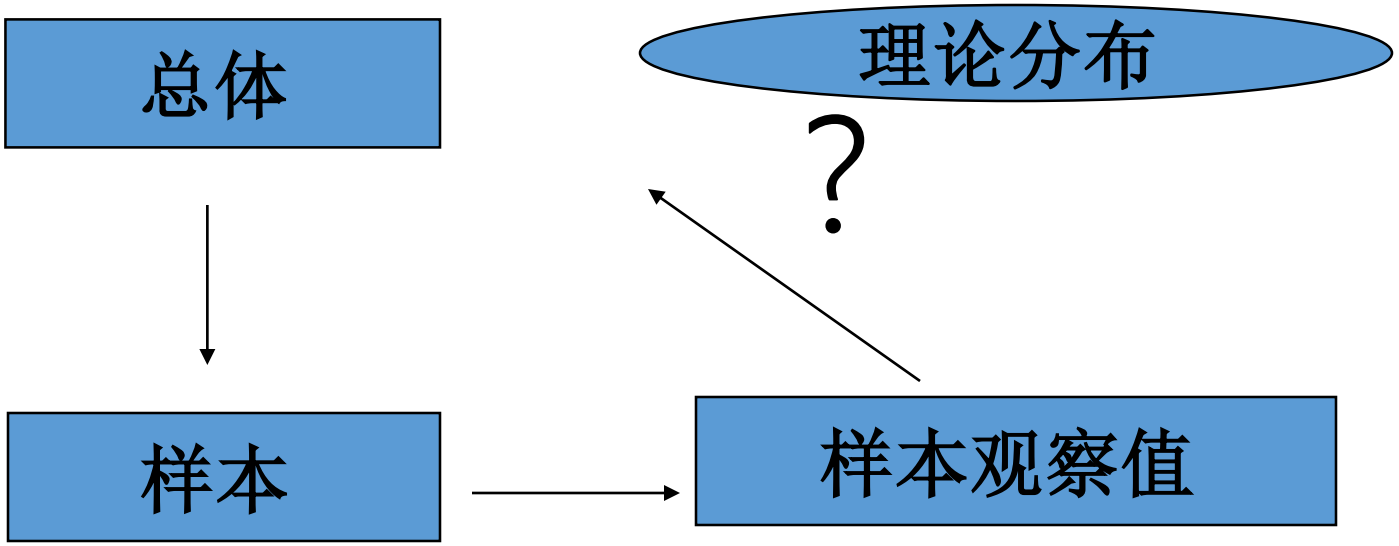
数理统计的基本任务

- 样本来自总体, 必然携带有反映总体性质的各种信息。

➢ 数理统计的基本任务就是通过对样本的研究来对总体的未知参数或分布类型作出估计，对有关总体的假设作出推断。


总 体 X
随机抽样
获得样本
样本$X_1,X_2,\ldots,X_n$
整理加工
统计推断
完成试验
获得数据
样本值$x_1,x_2,\ldots,x_n$
统计
工作

总体、样本、样本观察值的关系



统计是从手中已有的资料——样本观察值，去推断总体的情况——总体分布。　样本是联系两者的桥梁。

总体分布决定了样本取值的概率规律，也就是样本取到样本观察值的规律，因而可以用样本观察值去推断总体

统计量

样本是进行统计推断的依据。但在实际应用时，一般不是直接使用样本本身，而是对样本进行整理和加工，即**针对具体问题构造适当的函数 — 统计量**，利用这些函数来进行统计推断，揭示总体的统计特性.

定义 设X_1，X_2，…，X_n是来自总体X的样本，x_1，x_2，…，x_n为其样本值，则称**不含任何总体分布中未知参数的连续函数** $g(X_1, X_2, \cdots, X_n)$ **为统计量**，相应实数 $g(x_1, x_2, \cdots, x_n)$ 称为其观察值。

常用统计量有：

样本均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

(修正)样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

(修正)样本标准差 $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

样本k阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ (k = 1,2,\cdots)$

样本k阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \ (k = 1,2,\cdots)$

- 上述各统计量的观察值为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \,(k = 1,2,\cdots)$$

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \,(k = 1,2,\cdots)$$

- 重要结论：样本矩(的连续函数)依概率收敛于总体矩(的连续函数)[矩估计的理论基础]。

抽样分布

- 完全由样本确定的函数就是统计量。

统计量是随机变量，它的分布称为**抽样分布**。

下面, 介绍来自**正态总体**的几个重要统计量的分布.

1、χ^2-分布(卡方分布)

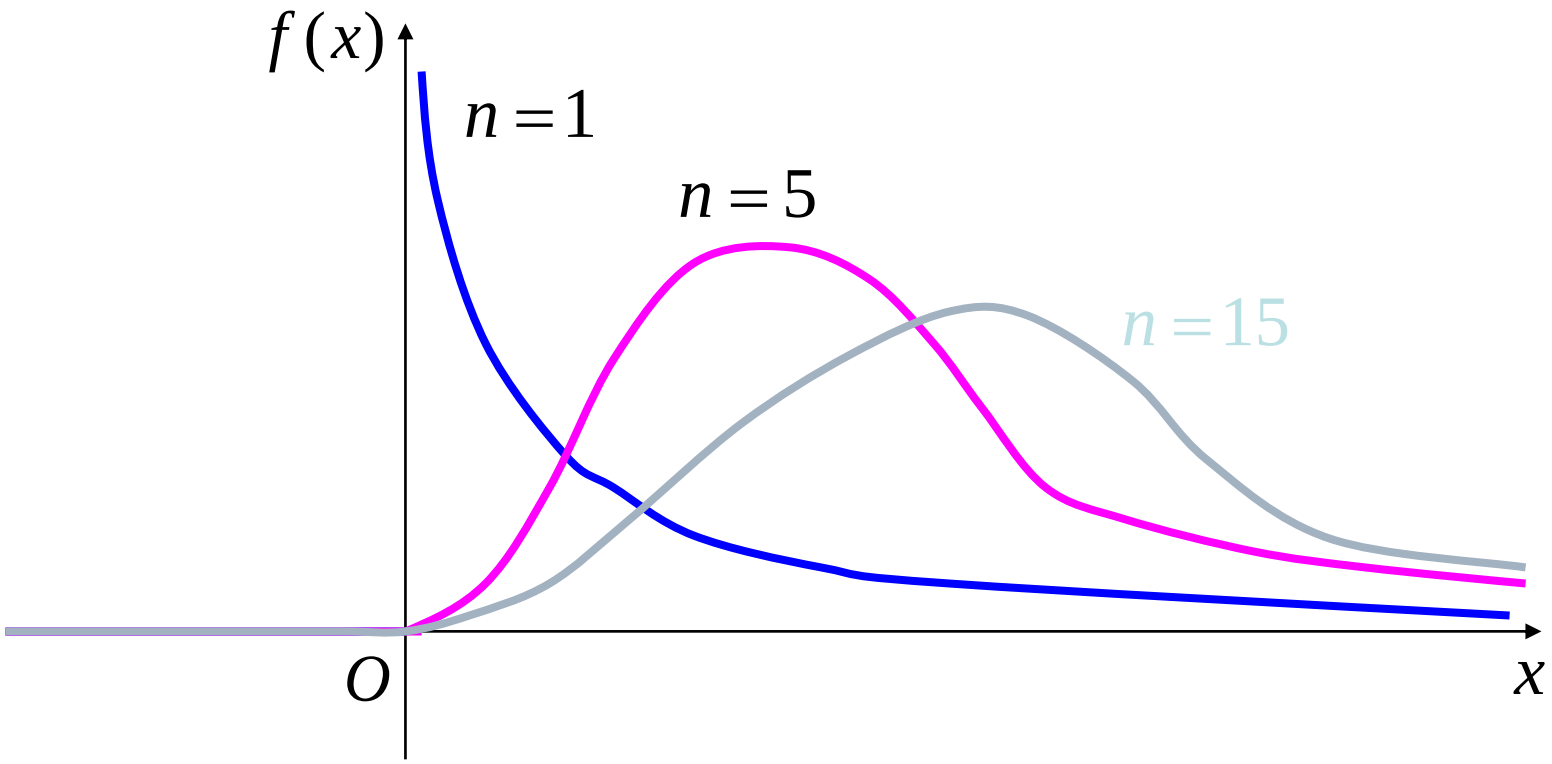
❶ **定义** 设X_1，X_2，…，X_n是来自标准正态总体N(0, 1)的样本, 称统计量

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为n的 x ^{2}–分布, 记为 $\chi^2 \sim \chi^2(n)$.

❷ $\chi^2(n)$-分布的**概率密度**为

$$f(x)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0,\\ 0, & \text{其它.}\end{cases}$$



❸ $\chi^2(n)$-分布的性质与数字特征

$\chi^2(n)$ -分布的可加性:

$$X \sim \chi^2(n_1), Y \sim \chi^2(n_2), \text{且} X, Y \text{独立} \Rightarrow X+Y \sim \chi^2(n_1+n_2)$$

$\chi^2(n)$ -分布的期望与方差为:

$$E(\chi^2)=n, D(\chi^2)=2n.$$

❹ 上α分位点(双侧α/2分位点)

定义 点 $\chi^2_\alpha(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点

$$\Leftrightarrow P\{\chi^2 > \chi^2_\alpha(n)\}=\alpha(0<\alpha<1).$$

分位数/Quantile/ Квантиль

- 亦称分位点，是指用分割点（cut point）将一个随机变量的概率分布范围分为几个具有相同概率的连续区间。分割点的数量比划分出的区间少1，例如3个分割点能分出4个区间。
- 常用的有中位数（即二分位数）、四分位数（quartile）、十分位数（decile ）、百分位数等。q-quantile是指将有限值集分为q个接近相同尺寸的子集。
- 分位数指的就是连续分布函数中的一个点，这个点对应概率p。

α-квантилем (или квантилем уровня α) распределения $\mathbb{P}^X$ называется число $x_\alpha \in \mathbb{R}$, такое что

$$\mathbb{P}(X \leqslant x_\alpha) \geqslant \alpha\,,$$
$$\mathbb{P}(X \geqslant x_\alpha) \geqslant 1-\alpha.$$

2、t-分布

❶ **定义** 设 $X \sim N(0,1), Y \sim \chi^2(n)$,且X与Y独立，则称随机变量

$$t = \frac{X}{\sqrt{Y/n}}$$

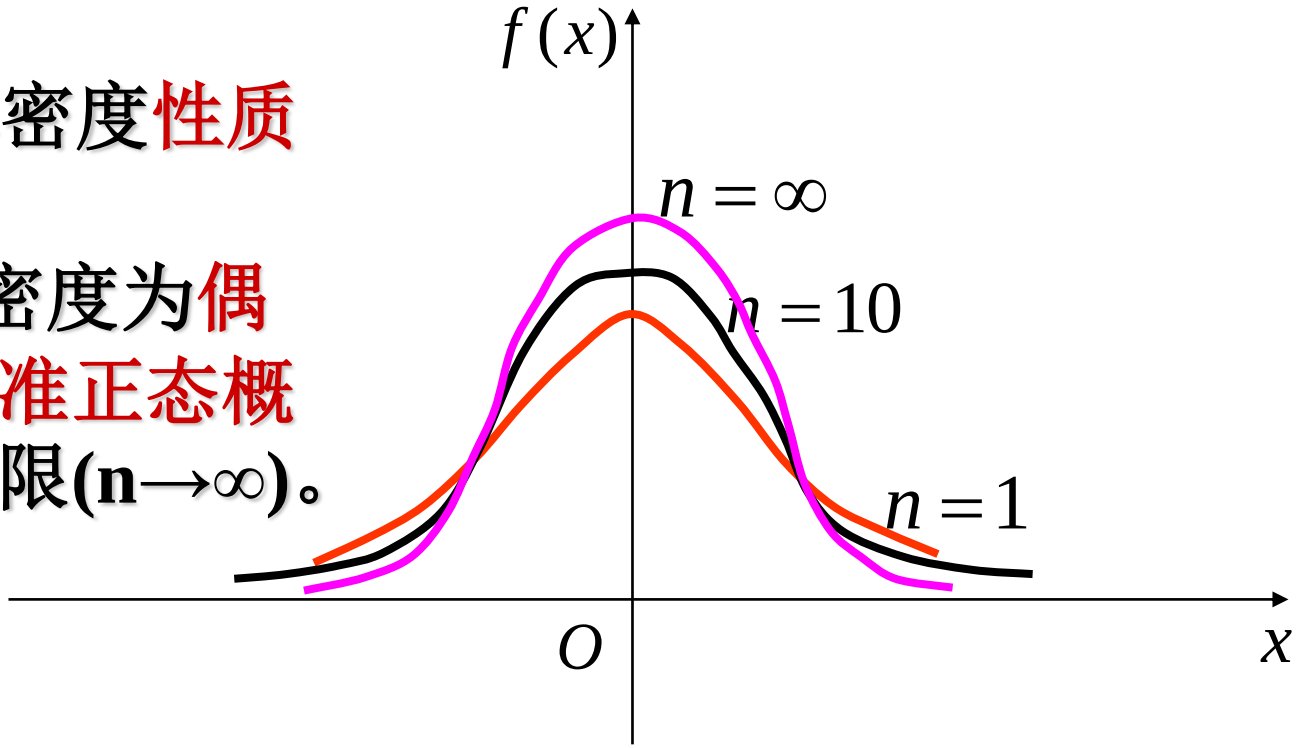
服从自由度为n的t-分布，记为 $t \sim t(n)$.

❷ t-分布的概率密度为

$$f(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty < x < +\infty)$$

❸ t-分布的概率密度性质

- t-分布的概率密度为偶函数，且以标准正态概率密度为其极限(n→∞)。

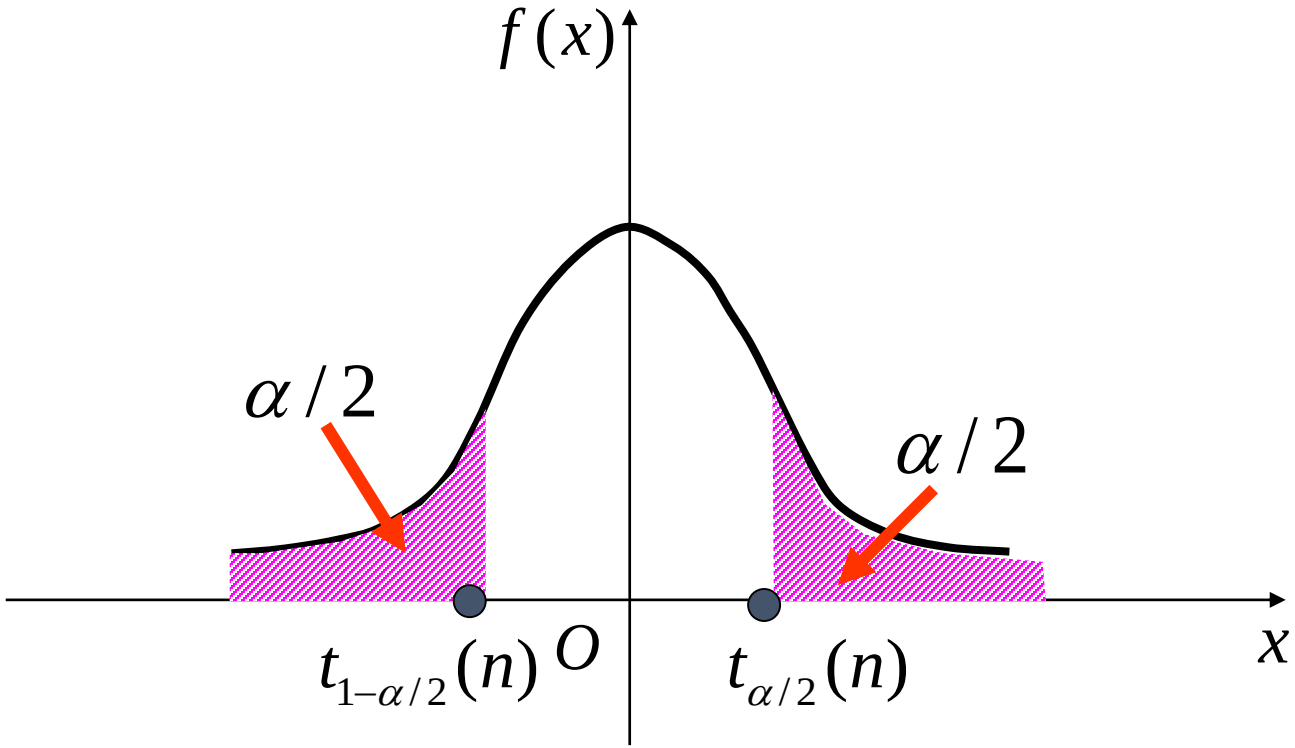


❹ 上α分位点(双侧α/2分位点)

定义 点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位点

$$\Leftrightarrow P\{t > t_\alpha(n)\} = \alpha (0 < \alpha < 1).$$

双侧α/2分位点： $t_{1-\alpha/2}(n), t_{\alpha/2}(n)$



显然，

$$t_{1-\alpha/2}(n) = -t_{\alpha/2}(n)$$

3、F-分布

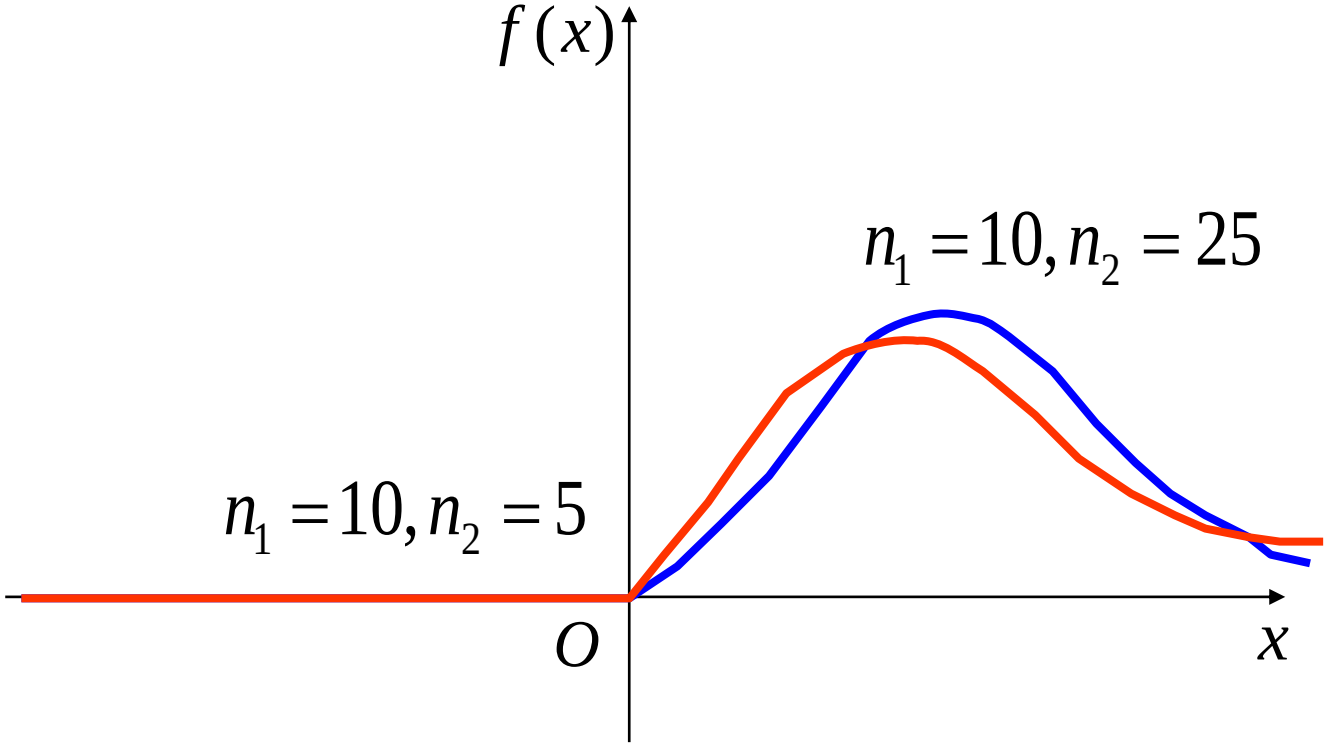
❶ **定义** 设 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$, 且X与Y独立，则称随机变量

$$F = \frac{X / n_1}{Y / n_2}$$

服从自由度为(n_1, n_2)的F-分布, 记为 $F \sim F(n_1, n_2)$.

❷ F-分布的概率密度为

$$f(x) = \begin{cases} \dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1x/n_2)]^{\frac{n_1+n_2}{2}}}, & x > 0, \\ 0, & \text{其它.} \end{cases}$$



❸ **F-分布的性质**

由F分布定义可得：

$$F \sim F(n_1, n_2) \Rightarrow \frac{1}{F} \sim F(n_2, n_1)$$

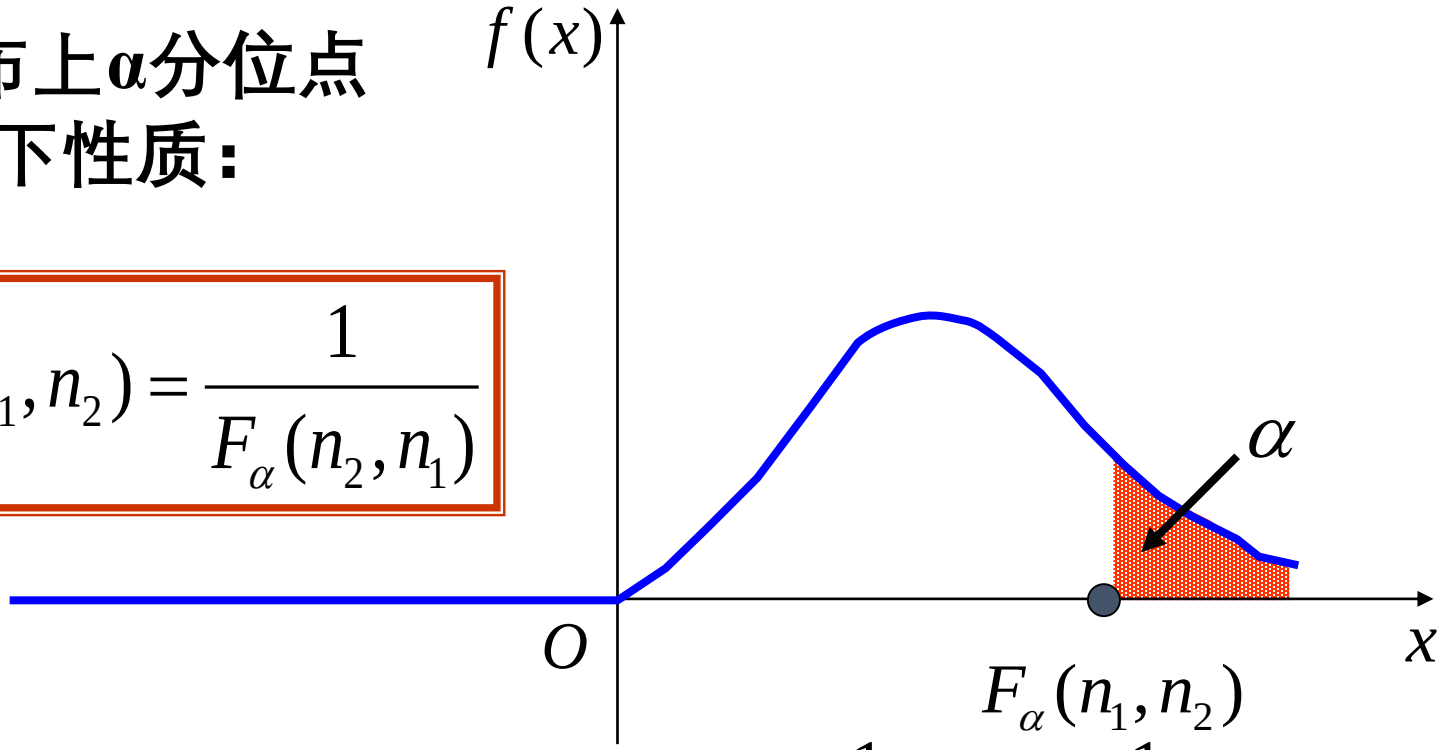
❹ 上α分位点(双侧α/2分位点)

定义 点$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$分布的上 α 分位点

$$\Leftrightarrow P\{F > F_\alpha(n_1,n_2)\} = \alpha(0<\alpha<1).$$

F分布上α分位点有如下性质：

$$F_{1-\alpha}(n_1,n_2) = \frac{1}{F_\alpha(n_2,n_1)}$$



查**附表5[P.447]**：$F_{0.95}(12,9) = \dfrac{1}{F_{0.05}(9,12)} = \dfrac{1}{2.80} = 0.357$

样本均值与样本方差的分布

设总体X有均值与方差：

$$E(X)=\mu, D(X)=\sigma^2,$$

$X_1, X_2, \cdots, X_n$ 是来自X（无论X服从何种分布!）的一个样本，则总有：

$$E(\overline{X})=\mu, D(\overline{X})=\frac{\sigma^2}{n}.$$

特别的，当 $X \sim N(\mu, \sigma^2)$ 时，样本均值

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

对于**单正态总体**$N(\mu,\sigma^2)$的**均值与方差**有:

定理1 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体$N(\mu,\sigma^2)$的样本, 则

①、$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$

②、$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1);$

③、$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1);$

④、$\overline{X}, S^2$ **独立.**

注意: $\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \sim \chi^2(n-1),$ 即2

$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n).$ 卡方分布定义

上面介绍的3个重要分布与4个重要公式在数理统计的区间估计与假设检验中有着重要应用，必须牢记!

χ^2-分布　　t-分布　　F-分布

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1);$$

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1);$$

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_w \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

参数估计

- 点估计
- 估计量的评选标准
- 区间估计

点估计

一、参数估计的概念

定义 设X_1，…，X_n是总体X的一个样本，其分布函数为$F(x;\theta)$, $\theta\in\Theta$。其中θ为未知参数, Θ为参数空间, 若统计量$g(X_1, \cdots, X_n)$可作为θ的一个估计,则称其为θ的一个估计量，记为 $\hat{\theta}$,

即$\hat{\theta} = g(X_1, \cdots, X_n)$.

注：$F(x;\theta)$也可用 分布律 或 密度函数 代替.

若x_1，…，x_n是样本的一个观测值。

$\hat{\theta} = g(x_1, \cdots, x_n)$称为$\theta$的估计值，

由于$g(x_1$，…，$x_n)$是实数域上的一个点，现用它来估计θ，故称这种估计为**点估计**。

点估计的经典方法

矩估计法 与 **极大似然估计法**。

矩估计法/ method of moments / Ме́тод моме́нтов

关键点：1.用样本矩作为总体同阶矩的估计，即

$$E(\hat{X}^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

2.约定：若 $\hat{\theta}$是未知参数θ的矩估计，则g(θ)的矩估计为g($\hat{\theta}$),

极大似然估计法/
maximum likelihood estimation, MLE/
Ме́тод максима́льного правдоподо́бия

1、极大似然思想

有两个射手，一人的命中率为0.9, 另一人的命中率为0.1, 现在他们中的一个向目标射击了一发，结果命中了，估计是谁射击的?

一般说，事件A发生的概率与参数$\theta\in\Theta$有关，θ取值不同，则P(A)也不同。因而应记事件A发生的概率为$P(A|\theta)$. 若A发生了，则认为此时的θ值应是在Θ中使$P(A|\theta)$ 达到最大的那一个。这就是**极大似然思想**

2.设总体X为连续型随机变量，概率密度f(x;θ)

现有样本观察值$x_1,x_2,\ldots x_n$,

问：根据极大似然思想，如何用$x_1,x_2,\ldots x_n$估计θ?

设$X_1,\cdots,X_n \overset{iid}{\sim} f(x;\theta),\quad \theta\in\Theta,$ 则称

$$L(\theta)=L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$$

为该总体的似然函数。

定义：若有 $\hat{\theta} \in \Theta$, 使得

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta) \overset{\text{或}}{=} \sup_{\theta \in \Theta} L(\theta),$$

则称 $\hat{\theta}$ 为θ的极大似然估计.记为$\hat{\theta}_{MLE}$.

$$\frac{d[\ln L(\theta)]}{d\theta} = 0$$

为什么取 log ?

注：极大似然估计具有下述性质：

若$\hat{\theta}$是未知参数θ的极大似然估计， g(θ)是θ的严格单调函数，则g(θ)的矩极大似然估计为g($\hat{\theta}$),

估计量的评选标准

一、一致性/ Consistency/ Состоя́тельная оце́нка

设 $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$是 θ 的估计量，若 $\hat{\theta} \xrightarrow{P} \theta$，则称 $\hat{\theta}$ 是 θ 的一致性估计量。

二、无偏性/ unbiased / Несмещённая оце́нка

设$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$为$\theta$的估计量, 若$E\hat{\theta} = \theta$

则称$\hat{\theta}$是θ的无偏估计量.

三、有效性

设$\hat{\theta}_i = \hat{\theta}_i(X_1, \cdots, X_n), i = 1, 2$分别是参数$\theta$的两个无偏估计, 若$D\hat{\theta}_1 < D\hat{\theta}_2$, 则称$\hat{\theta}_1$比$\hat{\theta}_2$有效.

区间估计/ interval estimation /

Интервальная оценка

定义： 设总体X的分布函数F(x；θ)含有未知参数θ，对于给定值α（0< α<1),若由样本$X_1, \cdots, X_n$确定的两个统计量$\underline{\theta}, \overline{\theta}$使

$$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1 - \alpha \qquad *$$

则称随机区间 $(\underline{\theta}, \overline{\theta})$为θ的置信度为**1–α的**置信区间

$\underline{\theta}$,和$\overline{\theta}$分别称为置信度为$1-\alpha$的置信上限和置信下限 。

注：F(x;θ)也可换成概率密度或分布律。

Доверительный интервал/ Confidence interval/置信区间

- 数值为95%置信区间意味着如果在同样情况下重复样本分析（这回生成不同的数据集），95%的区间会得出符合（总体）情况的实际结果。

Пример

设$X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\ \sigma^2)$,给定$\alpha$，由观测值$x_1,\cdots,x_n$，求出$\mu$的$1-\alpha$置信区间.

1、σ^2已知

$$令 p\{\overline{X}-a<\mu<\overline{X}+b\}=1-\alpha$$

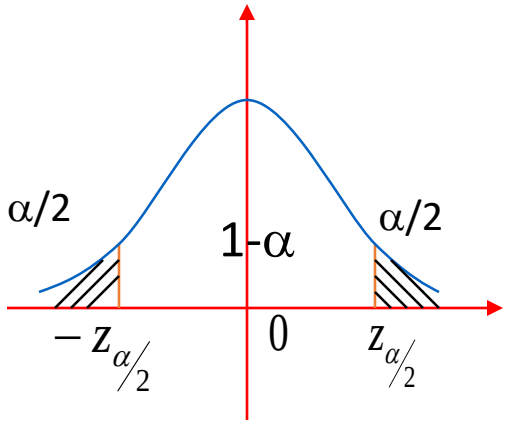
$$\Leftrightarrow\ p\{-b<\overline{X}-\mu<a\}=1-\alpha$$

$$\because\ \mathrm{U}=\frac{\overline{\mathrm{X}}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

$$\Leftrightarrow p\{-b\sqrt{n}/\sigma<U<a\sqrt{n}/\sigma\}=1-\alpha$$

可取

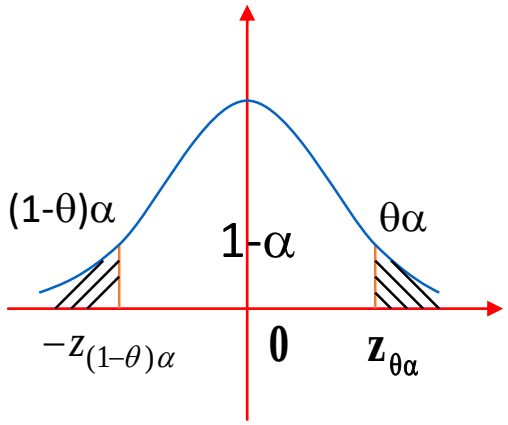
$$-b\sqrt{n}/\sigma=-z_{\alpha/2}\Rightarrow b=\frac{\sigma}{\sqrt{n}}z_{\alpha/2}$$



$$a\sqrt{n}/\sigma=z_{\alpha/2}\Rightarrow a=\frac{\sigma}{\sqrt{n}}z_{\alpha/2}$$

μ的置信度为**1–α**的置信区间为

$$(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\quad \overline{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2})。$$



注： μ的1–α置性区间不唯一。

$$\forall\theta,(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{(1-\theta)\alpha},\quad \overline{X}+\frac{\sigma}{\sqrt{n}}z_{\theta\alpha})。$$

都是μ的1–α置性区间.但θ=1/2时区间长最短.

求正态总体参数置信区间的解题步骤：

(1)根据实际问题构造样本的函数，要求仅含待估参数且分布已知；

(2)令该函数落在由分位点确定的区间里的概率为给定的置信度$1-\alpha$，要求区间按几何对称或概率对称；

(3)解不等式得随机的置信区间；

(4)由观测值及α值查表计算得所求置信区间。

假设检验/ Statistical hypothesis testing/ Проверка статистических гипотез

- 是用来判断样本与样本、样本与总体的差异是由抽样误差引起还是本质差别造成的统计推断方法。

- **显著性检验**是假设检验中最常用的一种方法，也是一种最基本的统计推断形式，其基本原理是先对总体的特征做出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受做出推断。

- 常用的假设检验方法有Z检验、t检验、卡方检验、F检验等

假设检验的基本思想:“小概率事件”原理

- 其统计推断方法是带有某种概率性质的反证法。
- 小概率思想是指小概率事件在一次试验中基本上不会发生。
- 反证法思想是先提出检验假设，再用适当的统计方法，利用小概率原理，确定假设是否成立。
- 即为了检验一个假设H0是否正确，首先假定该假设H0正确，然后根据样本对假设H0做出接受或拒绝的决策。如果样本观察值导致了“小概率事件”发生，就应拒绝假设H0，否则应接受假设H0

假设检验的基本思想:“小概率事件”原理

- 假设检验中所谓“小概率事件”，并非逻辑中的绝对矛盾，而是基于人们在实践中广泛采用的原则，即小概率事件在一次试验中是几乎不发生的，但概率小到什么程度才能算作“小概率事件”，显然，“小概率事件”的概率越小，否定原假设H0就越有说服力，常记这个概率值为α(0<α<1)，称为检验的显著性水平。

- 对于不同问题，检验的显著性水平α不一定相同，一般认为，事件发生的概率小于0.1、0.05或0.01等，即“小概率事件”

Нулевая гипотеза/ null hypothesis/ 零假设

假设检验的例子（法庭窘境/Courtroom trial）：

- **零假设**：H0-认为被告是清白的。
- **备择假设 (Alternative hypothesis)**：H1-则认为被告有罪。

□ 起诉是因为怀疑被告有罪。H0（现状）与H1对立并且被认可，除非H1被“超过合理质疑”的证据证伪。然而，“无法排除H0”并不能代表被告清白，只是说证据无法将其定罪。所以，陪审团没有必要在H0“无法推翻”的情况下将其“接受”。

- 第一型错误中零假设被错误地证伪，得出测试结果为“假阳性”。
- 第二型错误中零假设没有被及时排除，总体中的实际差异被错误判断为“假阴性”。

	H_0 is true Truly not guilty	H_1 is true Truly guilty
Accept null hypothesis Acquittal	Right decision	Wrong decision Type II Error
Reject null hypothesis Conviction	Wrong decision Type I Error	Right decision

I 型错误的发生概率为显著性水平α

步骤

1. 最初研究假设为真相不明。
2. 第一步是**提出相关的零假设和对立假设**。这是很重要的，因为错误陈述假设会导致后面的过程变得混乱。
3. 第二步是考虑检验中对样本做出的**统计假设**；例如，关于独立性的假设或关于观测数据的分布的形式的假设。这个步骤也同样重要，因为无效的假设将意味着试验的结果是无效的。
4. 决定哪个检测是合适的，并确定相关检验统计量 *T*。
5. **在零假设下推导检验统计量的分布**。在标准情况下应该会得出一个熟知的结果。比如检验统计量可能会符合学生t-分布或正态分布。
6. 选择一个显著性水平 (α)，若低于这个概率阈值，就会拒绝零假设。最常用的是 5% 和 1%。
7. 根据在零假设成立时的检验统计量 *T* 分布，找到数值最接近备择假设，且几率为显著性水平 (α)的区域，此区域称为“拒绝域”，意思是在零假设成立的前提下，落在拒绝域的几率只有α。
8. 针对检验统计量 *T*，根据样本计算其估计值 t_{obs}。
9. 若估计值 t_{obs}未落在“拒绝域”，接受零假设。若估计值 t_{obs}落在“拒绝域”，拒绝零假设，接受对立假设。

Пример

- 假设甲乙两人共比赛100次，其中**甲61胜39败**。
  - 下面是他们对该结果的争论
- **甲:** 我比你更强
- **乙:** 不，这纯属偶然
- **甲:** 偶然？那也差距太大了吧！这明显是我们的实力差距
- **乙:** 也不一定吧，就算实力差不多，偶尔出现这样的结果也不奇怪吧?
- **甲:** 没那种事
- **乙:** 真的吗？你具体计算一下试试?
- **甲:** 好，我现在就算。如果出现现在这个结果(甲61胜39败)的概率小于5%，你就承认这是实力的差距吧

- 通过一次比赛结果来判断总体的实力水平，就可以通过假设检验来进行推断。甲认为自己的比赛水平高于乙，故可以提出如下的备择假设:
- H1: **甲获胜的概率 > 1/2**
- 我们无法直接对我们期望证实的H1假设来进行判定，所以，我们需要提出一个虚无假设来进行判定推断
- H0: **甲获胜的概率 = 1/2**
- 上述场景中的5%，即为我们根据实际需要来设定的显著性水平 α
- 根据虚无假设H0的条件，计算出现上述比赛结果的概率，易知100次比赛中甲获胜X次数的概率符合二项分布: $X \backsim B(100, 1/2)$

- 这里利用Matlab计算可得:

$$P(X \geq 61)=0.0176 < \alpha=5\%,$$

故我们知道在H0虚无假设下，出现当前比赛结果的概率远低于我们之前设定的显著性水平 $\alpha=5\%$ ，即在H0假设下，出现当前比赛结果是一个小概率事件，而实际样本数据即是这样的，故发生了不合理现象，所以，我们拒绝了虚无假设H0，接受了备择假设H1，即出现这样的比分结果，是由于甲的实力水平高于乙，而不是偶尔恰好发生的

命令行窗口

```
>>
>> clear;
startX = 61;
endX = 100;
P = 0; % P(X>=61)
for count = startX:1:endX
    temp = binopdf(count, 100, 1/2);
    P = P + temp;
end
display(P);

P =

    0.0176
```

Линейная регрессия/ Linear regression/ 线性回归

例子：牙膏的销售量

问题

建立牙膏销售量与价格、广告投入之间的模型

预测在不同价格和广告费用下的牙膏销售量

收集了30个销售周期本公司牙膏销售量、价格、广告费用，及同期其它厂家同类牙膏的平均售价

销售周期	本公司价格(元)	其它厂家价格(元)	广告费用(百万元)	价格差（元）	销售量(百万支)
1	3.85	3.80	5.50	-0.05	7.38
2	3.75	4.00	6.75	0.25	8.51
…	…	…	…	…	…
29	3.80	3.85	5.80	0.05	7.93
30	3.70	4.25	6.80	0.55	9.26

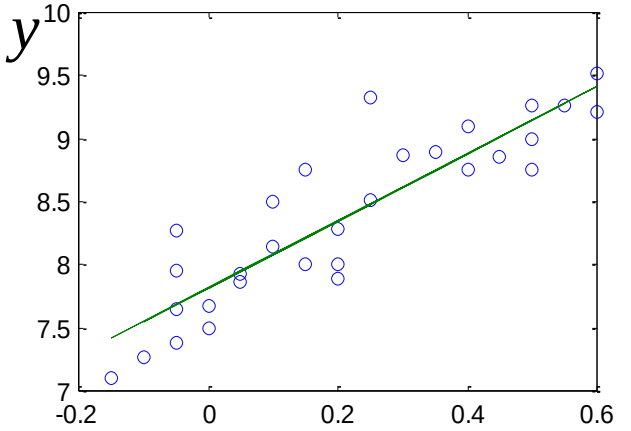
基本模型

y ~公司牙膏销售量

x_1~其它厂家与本公司价格差

x_2~公司广告费用

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$



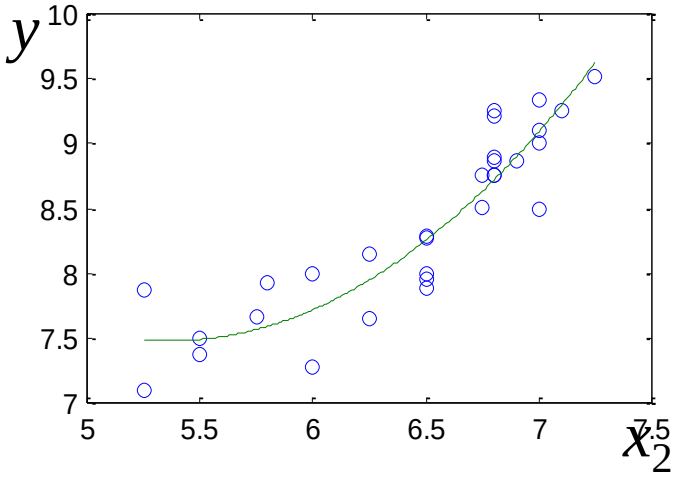
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y~被解释变量（因变量）

x_1, x_2~解释变量(回归变量, 自变量)

$\beta_0, \beta_1, \beta_2, \beta_3$ ~回归系数

ε~随机误差（均值为零的正态分布随机变量）



$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon$$

模型求解

MATLAB 统计工具箱

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$　由数据 y, x_1, x_2 估计 β

```
[b,bint,r,rint,stats]=regress(y,x,alpha)
```

输入

y~*n*维数据向量

x= $[1 \quad x_1 \quad x_2 \quad x_2^2]$ **~*n*×4数据矩阵, 第1列为全1向量**

alpha(置信水平,0.05)

输出

b~β的估计值

bint~b的置信区间

r ~残差向量y-xb

rint~r的置信区间

参数	参数估计值	置信区间
β_0	17.3244	[5.7282　28.9206]
β_1	1.3070	[0.6829　1.9311]
β_2	-3.6956	[-7.4989　0.1077]
β_3	0.3486	[0.0379　0.6594]
R^2=0.9054　F=82.9409　p=0.0000		

Stats~

检验统计量

R^2, F, p

p值/ *p*-value/ P-значение

- **定义：**为拒绝零假设 H0 的最低显著性水平。
- p值是当原假设为真时所得到的样本观察结果或更极端结果出现的概率。如果P值很小，说明原假设情况的发生的概率很小，而如果出现了，根据小概率原理，我们就有理由拒绝原假设，P值越小，我们拒绝原假设的理由越充分。
- **p-value的作用：**p-value就是用来判断H0假设是否成立的依据。因为期望值是基于H0假设得出的，如果观测值与期望值越一致，则说明检验现象与零假设越接近，则越没有理由拒绝零假设。如果观测值与期望值越偏离，说明零假设越站不住脚，则越有理由拒绝零假设，从而推出对立假设的成立。
- **p-value计算：**计算chi-suqare，计算自由度，查卡方分布表。

p值/ *p*-value/ P-значение

- **这个p-value到底是个什么鬼？** p值可通过计算chi-square后查询卡方分布表得出，用于判断H0假设是否成立的依据。
- **为什么小于0.05就很重要？** 大部分时候，假设错误拒绝H0的概率为0.05，所以如果p值<0.05，说明错误拒绝H0的概率很低，则我们有理由相信H0本身就是错误的，而非检验错误导致。
- 大部分时候p-value用于检验独立变量与输入变量的关系，H0假设通常为假设两者没有关系，所以若p值小于0.05，则可以推翻H0（两者没有关系），推出H1（两者有关系）。
- **很重要是什么意思？** 当p值小于0.05时，我们就说这个独立变量重要（significant），因为这个独立变量与输出结果有关系。

结果分析

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]
β_2	-3.6956	[-7.4989 0.1077]
β_3	0.3486	[0.0379 0.6594]
R^2=0.9054	F=82.9409	p=0.0000

y的90.54%可由模型确定

F远超过F检验的临界值

p远小于α=0.05

模型从整体上看成立

β_2的置信区间包含零点
(右端点距零点很近)

x_2对因变量y 的
影响不太显著

x_2^2项显著

可将x_2保留在模型中

销售量预测

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2$$

价格差x_1=其它厂家价格x_3-本公司价格x_4

估计x_3 调整x_4 ⇨ 控制x_1 ⇨ 通过x_1, x_2预测y

控制价格差x_1=0.2元，投入广告费x_2=650万元

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 = 8.2933 \text{ (百万支)}$$

销售量预测区间为 [7.8230，8.7636]（置信度95%）

上限用作库存管理的目标值　下限用来把握公司的现金流

若估计x_3=3.9，设定x_4=3.7，则可以95%的把握知道销售额在 7.8320×3.7≈ 29（百万元）以上

模型改进

x_1和x_2对y的影响独立

x_1和x_2对y的影响有交互作用

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]
β_2	-3.6956	[-7.4989 0.1077]
β_3	0.3486	[0.0379 0.6594]
R^2=0.9054	F=82.9409	p=0.0000

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	29.1133	[13.7013 44.5252]
β_1	11.1342	[1.9778 20.2906]
β_2	-7.6080	[-12.6932 -2.5228]
β_3	0.6712	[0.2538 1.0887]
β_4	-1.4777	[-2.8518 -0.1037]
R^2=0.9209	F=72.7771	p=0.0000

两模型销售量预测比较

控制价格差x_1=0.2元，投入广告费x_2=6.5百万元

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

$\hat{y}=8.2933$ **(百万支)**

区间 [7.8230，8.7636]

$$\hat{y}=\beta_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2+\hat{\beta}_4x_1x_2$$

$\hat{y}=8.3272$ **(百万支)**

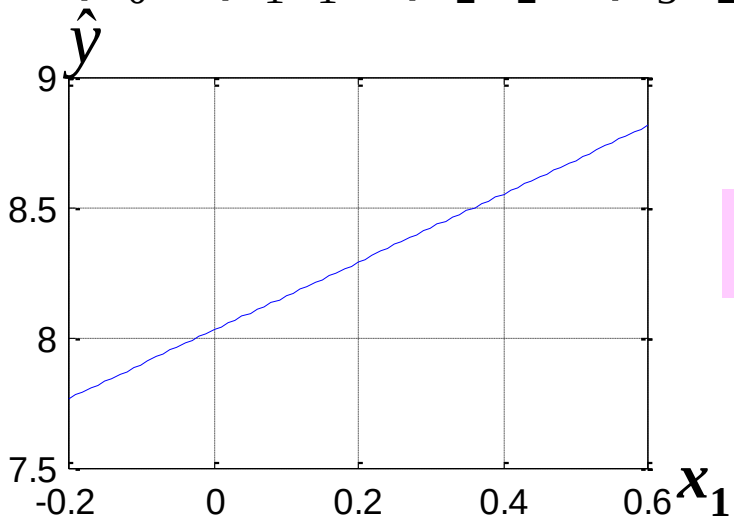
区间 [7.8953，8.7592]

$\hat{y}$ **略有增加**

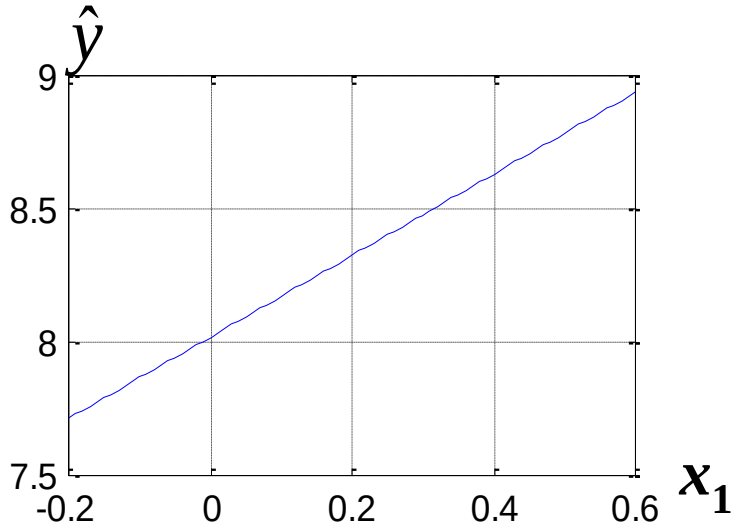
预测区间长度更短

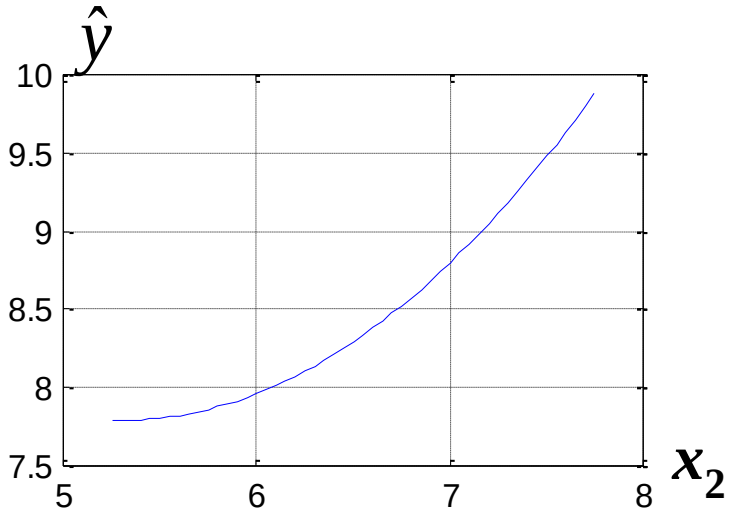
两模型 $\hat{y}$ 与x_1, x_2关系的比较

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 \qquad \hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$

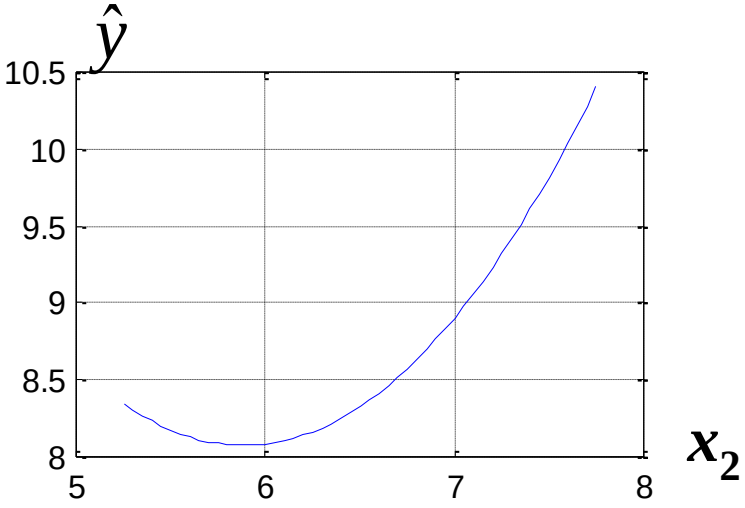


x_2=6.5





x_1=0.2



交互作用影响的讨论

$$\hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$

价格差 x_1=0.1

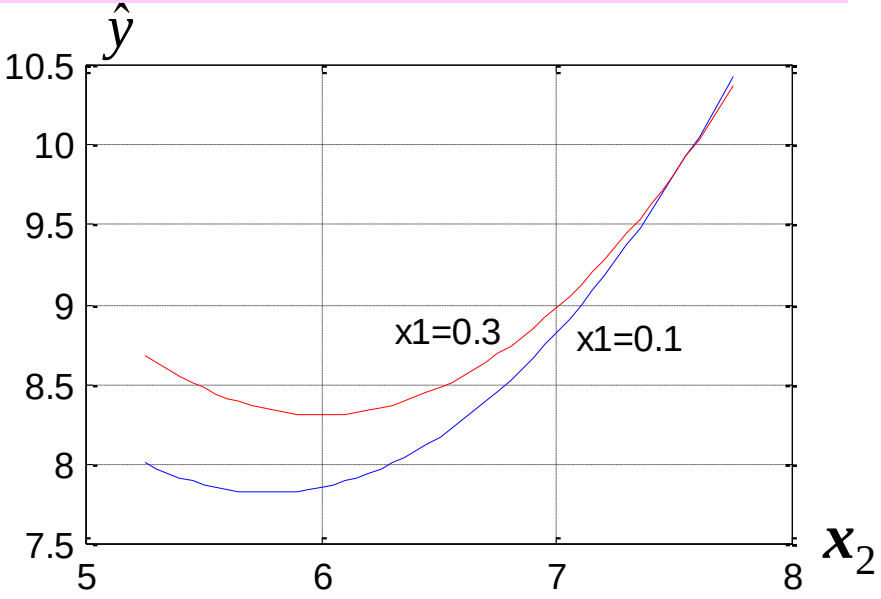
$$\hat{y}\big|_{x_1=0.1} = 30.2267 - 7.7558\,x_2 + 0.6712\,x_2^2$$

价格差 x_1=0.3

$$\hat{y}\big|_{x_1=0.3} = 32.4535 - 8.0513\,x_2 + 0.6712\,x_2^2$$

$$x_2 < 7.5357 \Rightarrow \hat{y}\big|_{x_1=0.3} > \hat{y}\big|_{x_1=0.1}$$

⇨ **价格优势会使销售量增加**



加大广告投入使销售量增加

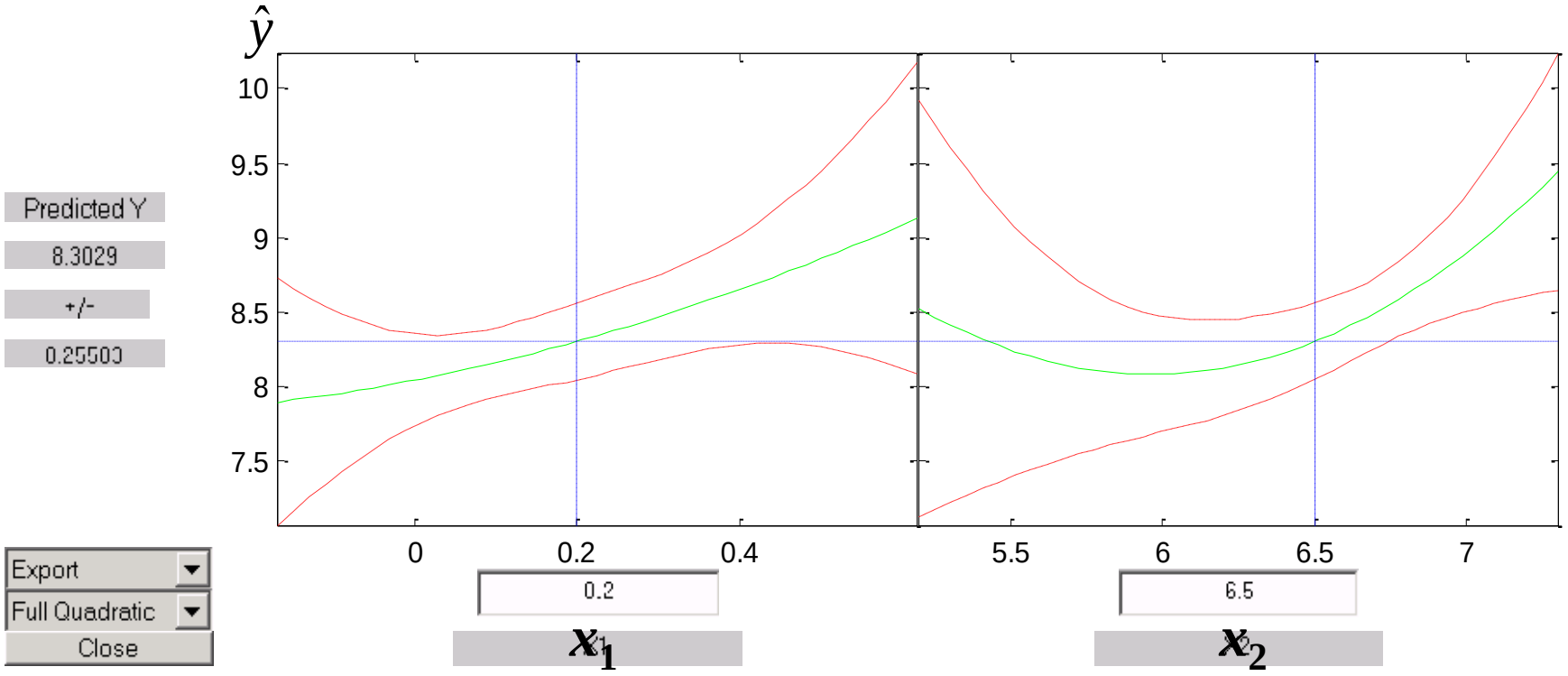
（x_2大于6百万元）

价格差较小时增加的速率更大 ⇨ **价格差较小时更需要靠广告来吸引顾客的眼球**

完全二次多项式模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

MATLAB中有命令rstool直接求解



从输出 Export 可得 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$

Метод главных компонент/ Principal Component Analysis/ 主成分分析

- Снижение размерности/ dimensionality reduction/降维。

- 原理：在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

Машинное обучение/ Machine learning/机器学习

- **Обучение с учителем/ Supervised learning/监督式学习**：从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。常见的监督学习算法包括回归分析和统计分类。
- **Обучение без учителя/ Unsupervised learning/ 无监督学习**：输入无监督算法的数据都没有标签，也就是只为算法提供了输入变量(X)而没有对应的输出变量。在无监督学习中，算法需要自行寻找数据中的有趣结构。无监督学习的主要运用包含：聚类分析（cluster analysis）、关系规则（association rule）、维度缩减（dimensionality reduce）。
- **Обучение признакам / Feature learning/ 表征学习**：将原始数据转换成为能够被机器学习来有效开发的一种形式。学习如何学习。分为：监督的和无监督的。

Глубокое обучение/ Deep learning/深度学习

- 是一种以人工神经网络为架构，对数据进行表征学习的算。
- 用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。
- 卷积深度神经网络（Convolutional Neural Networks, CNN）在计算机视觉领域得到了成功的应。