《生物实验设计》
第三章　概率和概率分布

王超

广东药科大学

Email: wangchao@gdpu.edu.cn

2022-09-07

广東藥科大學 GUANGDONG PHARMACEUTICAL UNIVERSITY

第四章　统计推断

内涵和作用

统计推断主要包括

- 假设检验
- 参数估计

统计推断的任务

- 分析误差产生的原因
- 确定差异的性质
- 排除误差的干扰
- 对总体的特征做出正确的判断

- 试验数据往往存在一定的差异，这种差异可能
  - 由于随机误差产生
  - 由于试验处理所引起
- 试验处理的效应往往和随机误差混淆，不容易分开
- 通过概率的计算和假设检验作出正确判断

假设检验

- 根据总体的理论分布和小概率原理，对未知或不完全知道的总体提出两种彼此对立的假设，然后由样本的实际结果，经过一定的计算，作出在一定概率意义上应该接受的那种假设的推断

如果：

- 抽样结果使小概率事件发生
  - 则拒绝假设
- 抽样结果没有使小概率事件发生
  - 则接受假设

小概率事件：概率 ≤ 0.05 或 ≤ 0.01 的事件为小概率事件

1. 提出假设
2. 确定显著水平
3. 计算统计数与相应的概率
4. 推断是否接受假设

（一）提出假设

- 对总体提出假设，一般是两个彼此对立的假设
  - 无效假设或零假设 H_0：
    - 处理的效应跟总体参数之间没有真实的差异，试验结果中的差异是误差所致，即处理“无效”
  - 备择假设 H_A：
    - 处理结果中的差异是由于总体参数不同所引起的，即处理“有效”
  - 无效假设与备择假设是对立事件：接受 H_0 则否定 H_A，接受 H_A 则否定 H_0
- H_0 随研究内容的不同而不同：
  - H_0 必须有意义
  - 根据 H_0 可以算出因抽样误差而获得样本结果的概率

(一)提出假设

以样本平均数的假设为例:

- 对一个样本平均数的假设(样本与总体)
  - 假设平均数为 $\bar{x}$ 的样本来自于一组具有 μ 的总体,提出:
    - $H_0 : \mu = \mu_0$
    - $H_A : \mu \neq \mu_0$
- 对两个样本平均数相比较的假设(样本与样本)
  - 假设两个样本平均数 $\bar{x}_1$ 和 $\bar{x}_2$ 分别来自具有平均数 μ_1 和 μ_2 的两个总体,提出:
    - $H_0 : \mu_1 = \mu_2$
    - $H_A : \mu_1 \neq \mu_2$
- 可从假设的总体中推论其平均数的随机抽样分布,从而可以算出其一个样本平均数指定值出现的概率,这样就可以根据样本与总体的关系,作为假设检验的理论依据

（一）提出假设

- 克矽平能否治疗矽肺病？
  矽肺病患者血红蛋白含量的平均数 $\mu_0 = 126(mg/L)$，$\sigma^2 = 240(mg/L)^2$ 的正态分布 $N(126, 240)$
  克矽平对 6 名患者进行治疗，治疗后测得平均血红蛋白含量 $\bar{x} = 136(mg/L)$
- $\bar{x}$ 和 μ_0 之间的差值是由抽样误差还是药物治疗造成的？

（二）确定显著水平

- 确定一个否定 H_0 的概率标准，显著水平 α
- 人为规定的小概率界限
- 常用 $\alpha = 0.05$ 和 $\alpha = 0.01$
- 根据研究需要调整

（三）计算统计数与相应的概率

（四）推断是否接受假设

（二）频率

- 在 n 次试验中，事件 A 出现的次数 m 称为事件 A 出现的频数，比值 $\frac{m}{n}$ 称为事件 A 出现的频率

$$W(A)=\frac{m}{n}, 0 \leq W(A) \leq 1$$

为测定某批玉米种子的发芽率，分别取 10，20，50，100，200，500，1000 粒种子。在相同条件下进行发芽试验：

Table 1: 某批种子的发芽试验结果

种子总数	发芽种子总数	种子发芽率
10	9	0.900
20	19	0.950
50	47	0.940
100	91	0.910
200	186	0.930
500	459	0.918
1000	920	0.920

（三）概率

- 假设在相同的条件下，进行大量重复试验，若事件 A 的频率稳定地在某一确定值 p 的附近摆动，则称 p 为事件 A 出现的概率

$$P(A) = p = \lim_{x \to \infty} \frac{m}{n}$$

不可能完全准确得到 p，在 n 充分大时，频率 $W(A)$ 作为 $P(A)$ 的近似值。

- 概率的基本性质：
  1. 任何事件的概率都在 0 和 1 之间 $0 \leqslant P(A) \leqslant 1$;
  2. 必然事件的概率等于 1 $P(U) = 1$
  3. 不可能事件的概率等于 0 $P(V) = 0$

（一）事件的相互关系

- 和事件
  - 事件 A 和事件 B 至少有一件发生而构成的新事件，$A+B$
- 积事件
  - 事件 A 和事件 B 同时发生而构成的新事件, $A\cdot B$
- 互斥事件
  - 事件 A 和事件 B 不能同时发生，$A\cdot B=V$
- 对立事件
  - 事件 A 和事件 B 必有一个事件发生，但二者不能同时发生，$A\cdot B=V, A+B=U, \bar{A}=B, \bar{B}=A$
  - 新生儿要么为男孩，要么为女孩

（一）事件的相互关系

- 独立事件
  - 事件 A 的发生与事件 B 的发生毫无关系
  - 独立事件群：多个事件 $A_1, A_2, A_3, \cdots, A_n$ 彼此独立
- 完全事件系
  - 多个事件 $A_1, A_2, A_3, \cdots, A_n$ 两两相斥，且每次试验结果必然发生其一

（二）概率计算法则

- 加法定理

  互斥事件 A 和 B 的和事件的概率等于事件 A 和事件 B 的概率之和，称为**加法定理**。

$$P(A+B)=P(A)+P(B)$$

推理 1：如果 $A_1, A_2, A_3, \cdots, A_n$ 为 n 个互斥事件，则其和事件的概率为

$P(A_1+A_2+A_3+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$

推理 2：对立事件 $\bar{A}$ 的概率为 $P(\bar{A})=1-P(A)$

推理 3：完全事件系和事件的概率等于 1

（二）概率计算法则

- **乘法定理**
  如果事件 A 和事件 B 为独立事件，则事件 A 与事件 B 同时发生的概率等于事件 A 和事件 B 各自概率的乘积，称为**乘法定理**。

$$P(A \cdot B) = P(A) \cdot P(B)$$

推理：如果 $A_1, A_2, A_3, \cdots, A_n$ 彼此独立，则
$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$

- 研究随机变量主要是研究变量的取值范围，也就是取值的概率
- 随机变量的概率分布：随机变量的取值与取这些值的概率之间的对应关系
- 随机变量的概率分布可以用分布函数表述
- 离散型变量的概率分布
    - 二项分布
    - 泊松分布
- 连续型变量的概率分布
    - 正态分布

（一）离散型随机变量的概率分布

【复习】离散型变量/非连续变量：在变量数列中仅能取得固定数值，并且通常是整数

- 离散型随机变量 x 所有可能的取值为 $x = x_i(i = 1, 2, \cdots, n)$
- 对于任意一个 x_i，都有一个相应的概率为 $p_i(i = 1, 2, \cdots, n)$

可以用下式表示为，

$$P(x = x_i) = p_i \quad (i = 1, 2, \cdots, n)$$

- x_i 与 p_i 为数值，表示事件“变量 x 取值为 x_i 时”的概率等于 p_i

并且，

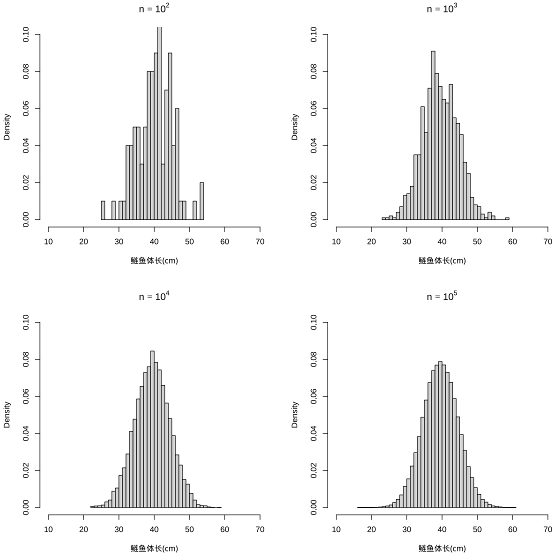
$$\sum_{i=1}^{n} p_i = 1$$

（二）连续型随机变量的概率分布

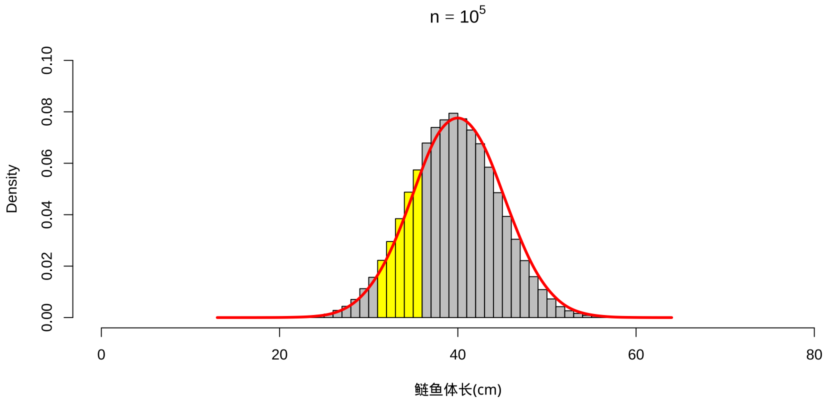
【复习】连续变量：在变量范围内可抽出某一范围内的所有值，变量之间是连续的、无限的

- 对于连续型随机变量，可以通过分组整理成次数分布表
- 如果从总体中抽取样本的容量 n 相当大，则频率分布就趋于稳定，近似地看成总体的概率分布
- 对连续型随机变量的次数分布表作直方图，直方图中同一间距内的频率密度是相等的
- 当 n 无限大，频率转化为概率，频率密度转化为概率密度，直方图逼近光滑连续曲线
    - 概率密度曲线（曲线下的总面积为 1）
    - 概率密度函数 $f(x)$

（二）连续型随机变量的概率分布



（二）连续型随机变量的概率分布



对于一个连续型变量 x，取值于区间内的概率即黄色阴影部分的面积，也就是概率密度函数 $f(x)$ 的积分，即

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} f(x)dx$$

- 事件 A 发生的频率 $W(A)$ 和概率 $P(A)$ 之间的关系，实际上就是样本统计数和总体参数的关系。
- 当 n 足够大的时候，为什么可以用样本中的 $W(A)$ 代替?
- 大数定律是阐述大量随机现象平均结果稳定性的一系列定律的总称。

伯努利大数定律：m 是 n 次独立试验中事件 A 出现的次数，p 是事件 A 在每次试验中出现的概率，对于任意小的正数 ϵ，有：

$$\lim_{n\to\infty} P\left\{\left|\frac{m}{n} - p\right| < \epsilon\right\} = 1$$

以上，P 为实现 $\left|\frac{m}{n} - p\right| < \epsilon$ 这一事件的概率，$P = 1$ 是必然事件。

- 设一个随机变量 x_i，是由一个总体平均数 μ 和随机误差 ϵ_i 构成 $x_i = \mu + \epsilon_i$
- 从总体中抽取 n 个随机变量构成一组样本，样本的平均数是

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(\mu + \epsilon_i) = \mu + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$$

- 当样本容量 n 越来越大，$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i$ 就越小，使得 $\bar{x}$ 逼近 μ
- 样本容量越大，样本统计数与总体参数之差越小

随机变量的概率分布可以用分布函数来表述。

- 离散型变量的概率分布
  - 二项分布
  - 泊松分布
- 连续型变量的概率分布
  - 正态分布

（一）二项分布的概率函数

- 二项分布是一种离散型随机变量的分布。
  - 每次试验只有两个对立结果，A 和 $\bar{A}$，出现的概率分别记为 p 和 q $(q=1-p)$。
  - 试验具有重复性和独立性。
    - 重复性：每次试验条件不变，在每次试验中事件 A 出现的概率都是 p
    - 独立性：任何一次试验中事件 A 的出现与其余各次试验中出现的任何结果无关。

$P(x)$ 为随机变量 x 的二项分布，记为 $B(n,p)$，概率分布函数为：

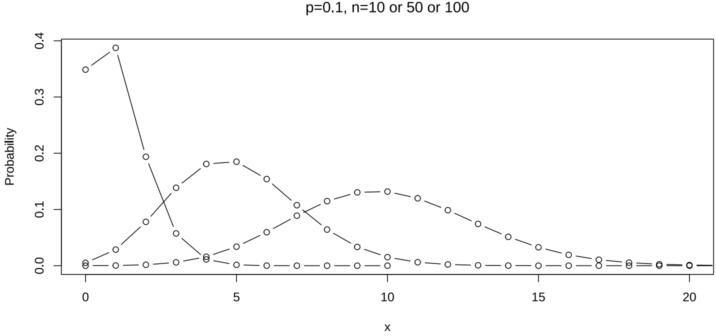
$$P(x)=\mathrm{C}_n^x p^x q^{(n-x)}$$

其中，$\mathrm{C}_n^x=\frac{n!}{x!(n-x)!}$，$q=1-p$。

重复实验 N 次，每次在 n 个试验中出现事件 A 为 x 的理论次数等于 $N\cdot P(x)$

（二）二项分布的的性状和参数

- 二项分布的形状由 n 和 p 两个参数决定
  - n 值不同的情况下，p 值较小的时候，分布是偏倚的。



  - p 值趋于 0.5 的时候，分布趋于对称（如何图形化？）

（二）二项分布的的性状和参数

- 二项分布的参数
  - 二项分布的平均数

$$\mu = np$$

  - 二项分布的总体标准差

$$\sigma = \sqrt{npq}$$

- 很多事件的发生概率很小，但是样本容量很大，即 n 很大和 p 很小
- 这是二项分布的特殊情况，即泊松分布
- 泊松分布的概率函数由二项分布推导得到：

$$P(x) = \mathrm{C}_n^x p^x (1-p)^{(n-x)}$$

由于是二项分布，所以 $P(x) = np = \mu$，即 $p = \frac{\mu}{n}$：

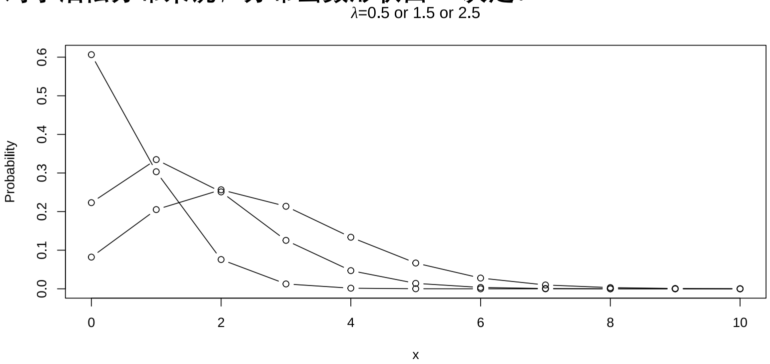
$$P(x) = \frac{n!}{(n-k)! \cdot k!} (\frac{\mu}{n})^x (1 - \frac{\mu}{n})^{n-x}$$

考虑到 n 无限大，μ 和 x 相对较小，可以近似后得到：

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

其中 λ 是参数，$\lambda = np$；e 为自然对数，x 为正整数。

- 泊松分布的平均数、方差和标准差为
  - $\mu = \lambda$
  - $\sigma^2 = \lambda$
  - $\sigma = \sqrt{\lambda}$
- 对于泊松分布来说，分布函数形状由 λ 决定。



- 正态分布是一种连续型随机变量的概率分布
- 多数变量都围绕在平均值左右，由平均值到分布的两侧，变量数逐渐减少
- 在统计理论和应用上最重要的分布
  - 试验误差的分布一般都服从于这种分布
  - 许多生物现象的计量资料也服从于这种分布
  - 正态分布还可作为离散型随机变量或其他变量的近似分布（中心极限定理）

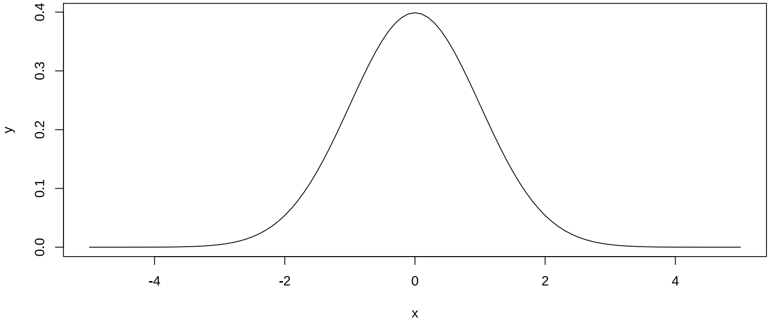
(一) 正态分布概率密度函数

- 正态分布的概率密度函数根据二项分布的函数在 $n \to \infty$ 时推导出

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

其中 μ 为总体平均数，π 为圆周率，e 为自然对数底

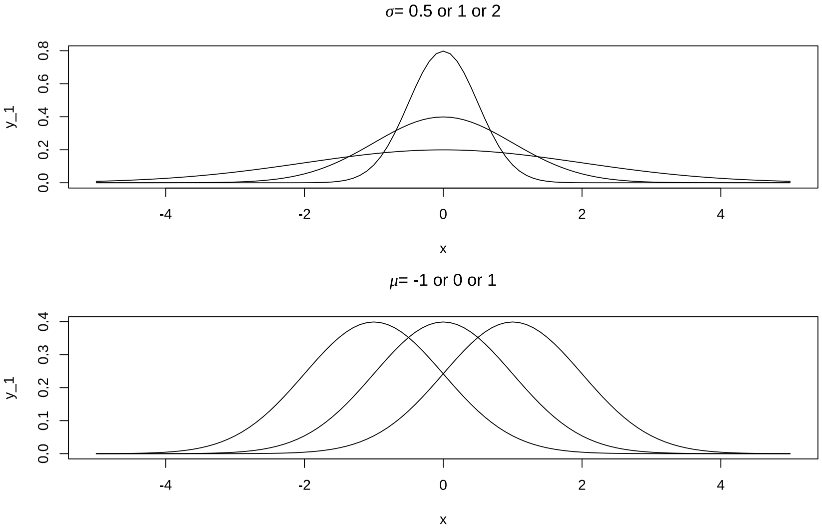
- 正态分布记为 $N(\mu, \sigma^2)$，表示平均数为 μ，方差为 σ^2 的正态分布



(二)正态分布特征

- 当 $x=\mu$ 时，$f(x)$ 有最大值为 $\frac{1}{\sigma\sqrt{2\pi}}$
- 当 $x-\mu$ 的绝对值相等，$f(x)$ 值也相等
- $\frac{x-\mu}{\sigma}$ 的绝对值越大，$f(x)$ 的值越小，逼近但不等于 0
- 正态分布曲线完全由参数 μ 和 σ 决定
- 正态分布在 $x=\mu\pm\sigma$ 处各有一个拐点
- 正态分布曲线在 $x\in(-\infty,\infty)$ 皆能取值（x 取值的完全事件系）

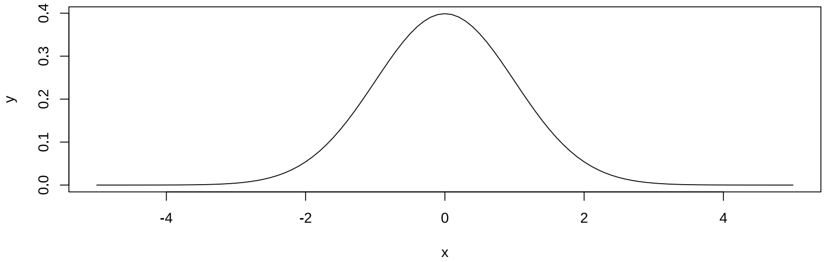
（二）正态分布特征


σ= 0.5 or 1 or 2
y_1
0.0
0.2
0.4
0.6
0.8
-4
-2
0
2
4
x
μ= -1 or 0 or 1
y_1
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

（三）标准正态分布

- μ 确定了分布曲线的中心位置
- σ 确定了分布曲线的变异度
- 对于 $N(\mu, \sigma^2)$ 来说，是一条曲线系
- 为了便于一般化应用，令 $\mu = 0, \sigma = 1$，则

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$



（三）标准正态分布

- 对于任何一个服从 $N(\mu, \sigma^2)$ 的随机变量，都可以通过 u 进行标准化变换

$$u = \frac{x - \mu}{\sigma}$$

- u 为标准正态离差，表示离开平均数 μ 几个标准差 σ