·讲座·

医学统计学的基本方法(二)

潘子昂 王石麟 李占魁

二 数据特性及其分布

1 数据特性

在实际工作中,通过对样本进行测量,可取得大量数据。如在身体检查中,获得100名某一年龄男性的身高数据,他们分别为:

141 137 140 132 142 147 139 141 136 140
134 142 142 145 135 142 139 144 142 139
142 142 130 134 142 137 136 137 134 137
137 144 145 132 148 140 145 139 146 139
153 136 148 140 139 138 140 136 145 150
143 138 143 141 148 139 145 137 137 139
145 131 141 144 144 142 147 135 136 139
140 138 135 142 143 142 142 142 140 141
137 146 136 137 127 137 138 142 134 143
142 141 141 144 148 155 137 136 149 149

厘米。这些数据粗看起来参差不齐、杂乱无章,但就这些数据来说,具有下述二个特性:①波动性。哪怕同一地区同日出生相同性别也不可能有相同身高。实验工作也一样,即使严格控制了实验条件,各次测量所得结果还是不尽相同,即数据具有一定的分散性。②规律性。它们在一定范围内变化,而且有向某一中心值集中的趋势。数理统计就是从波动的数据中找出其规律性。

2 频数分布

要找出规律性,需要对数据进行整理,找出频数分布。以上述数据为例,介绍数据整理的步骤。

1) 找出最大值与最小值。最大值为155,最小值为127。

2) 决定组距和组数。先决定组距,然后定组数。组距决定于极差。组数的决定依据样本容量:容量较大时,可分为10~20组,少于50时,分成5~6组。在上述例子中,极差R=155-127=28,所以组距定为3,可分为10组。

3) 决定各组的边界值和组中值。为了避免既可分在这一组又可分在那一组这种骑墙状态,通常使边界值数字比原测定值的有效数字位数多一位,分组情况见表1第1列。

4) 频数分布表和直方图。数出样本值落在每个组的数目,即频数,并把它列成频数分布表,见表1。其中相对频数为频数与样本容量之比。

表1 频数分布表

分 组	组中值	频数	累积频数	相对频数	累积相对频数
126.5~129.5	128	1	1	0.01	0.01
129.5~132.5	131	4	5	0.04	0.05
132.5~135.5	134	7	12	0.07	0.12
135.5~138.5	137	22	34	0.22	0.34
138.5~141.5	140	24	58	0.24	0.58
141.5~144.5	143	24	82	0.24	0.82
144.5~147.5	146	10	92	0.10	0.92
147.5~150.5	149	6	98	0.06	0.98
150.5~153.5	152	1	99	0.01	0.99
153.5~156.5	155	1	100	0.01	1.00
Σ		n=100		1.00	

为了更为直观,可在横坐标标出分组的点(边界值),纵坐标为对应的频数,以组距为底边,划出高度为频数的矩形,便得图1。在统计上称为直方图。如果直方图的纵坐标标的是相对频数,就得到相对频数分布的直方图。见图2。

如果我们能够取得更多的样本值,而且把组分得更细,那么各组的相对频数将趋于一个

作者单位:100029,中日友好医院临床医学研究所

稳定的值，直方图的形状逐渐趋于一条曲线，这条曲线可反映数据分布的规律，叫做频数分布曲线。如图2中的平滑曲线。数据波动规律不同，分布曲线的形状也就不一样。实际工作中，形状如图2中的曲线较为常见，称为正态分布曲线。

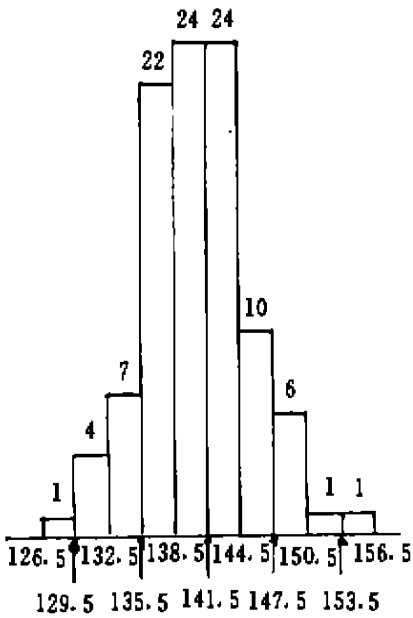


图1 频数分布直方图

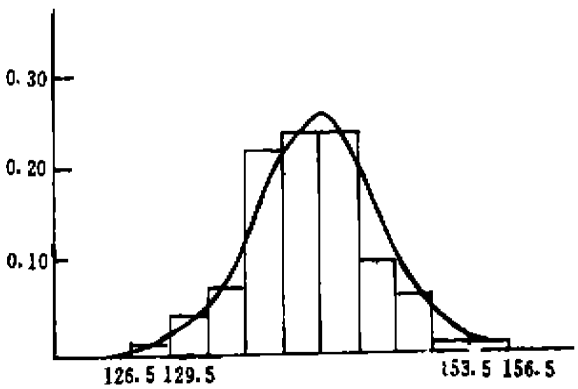


图2 相对频数分布直方图

3 正态分布

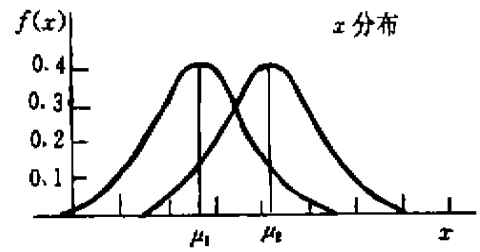
3.1 x 分布——测量值的分布

正态分布也称高斯分布，其分布曲线由正态概率密度函数

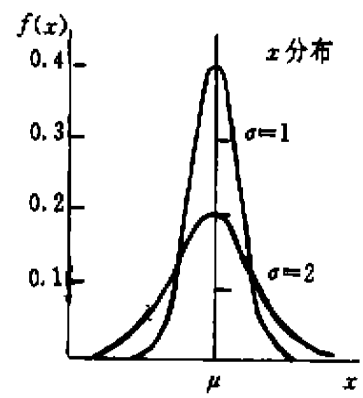
$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2],$$

$$(-\infty<x<\infty)$$

给出，式中 x 为随机变量 x 的各个取值，对应于实验观测的随机样本值，μ 为随机事件的期望值，对应于实验观测中的总体平均值，为曲线最高点的横坐标，曲线对 μ 对称；σ 为总体标准差。exp 为 $e\approx2.718$ 的指数式。μ 反映样本值的集中趋势，而 σ 可表征样本值的离散特性，σ 越大，数据越分散，分布曲线越宽，σ 越小，数据越集中，分布曲线越窄。见图3。



(a)μ 不同，σ 相同



(b)μ 相同，σ 不同

图3 测量值的正态分布

可见，μ 和 σ 是两个重要的参数，只要有了这两个参数，正态分布曲线完全确定下来。我们用 $N(\mu,\sigma^2)$ 表示均值为 μ，标准差为 σ 的正态分布。如果某事物的表现受大量微弱因素的影响，其总效果使实验数据分布形状接近于正态分布。正态分布是某些实验数据频数分布的极限状态，例如二项分布、泊松分布 χ^2 分布等。

3.2 u 分布——标准正态分布

对于遵从正态分布 $N(\mu,\sigma^2)$ 的样本值 x，经过如下的变换：

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

这时 u 的分布就成为标准正态分布。相当于 $\mu=0,\sigma=1$ 时的正态分布，可用 $N(0,1)$ 来表示。其概率密度函数为

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

3.3 正态分布表的使用

已知一个总体遵从正态分布 $N(\mu,\sigma^2)$，其平均值为 μ，标准差为 σ。这个总体的随机样本值落在某个区间(a,b)中的概率由下式给出：

$$P(a\leqslant x\leqslant b)=\frac{1}{\sigma\sqrt{2\pi}}\int_a^b exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx。$$

由于这个积分的计算比较麻烦，故需要经过上述变换，先使正态分布标准化为标准正态分布。

若正态分布 $N(\mu,\sigma^2)$的累积分布函数为

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]dx,$$

标准正态分布 $N(0,1)$的累积分布函数为

$$\phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} exp(-\frac{1}{2}u^2)du。$$

其函数值都列成表格，可从统计学书的附录中查到。

对于任何遵从正态分布的随机变量，样本值落在区间(a,b)的概率 $P(a\leqslant x\leqslant b)$，可由下列标准正态分布的累积分布函数给出：

$$\begin{aligned}P(a\leqslant x\leqslant b)&=P\left(\frac{a-\mu}{\sigma}\leqslant u\leqslant\frac{b-\mu}{\sigma}\right)\\&=\frac{1}{\sqrt{2\pi}}\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}}exp\left(\frac{-u^2}{2}\right)du\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{b-\mu}{\sigma}}exp\left(\frac{-u^2}{2}\right)du\\&-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{a-\mu}{\sigma}}exp\left(\frac{-u^2}{2}\right)du\\&=\phi\left(\frac{b-\mu}{\sigma}\right)-\phi\left(\frac{a-\mu}{\sigma}\right),\end{aligned}$$

即 $P(a\leqslant x\leqslant b)$可从$\frac{b-\mu}{\sigma},\frac{a-\mu}{\sigma}$这两个数的标准正态分布的累积分布函数值相减得到。

例：已知 x 遵从正态分布 $N[66.62,(0.21)^2]$，求 x 落在(66.15～67.04)中的概率。

解：已知 $\mu=66.62,\sigma=0.21$

当 $x=67.04$ 时，$u=\frac{x-\mu}{\sigma}=\frac{67.04-66.62}{0.21}=2.0$

$x=66.15$ 时，$u=\frac{x-\mu}{\sigma}=\frac{66.15-566.62}{0.21}=-2.24$

所以 $P(66.15\leqslant x\leqslant 67.04)=\phi(2.0)-\phi(-2.24)$，

又所查表中 u 都是正值，根据标准正态分布是对 $u=0$ 轴对称，对于$-u$，可用 $\phi(-u)=1-\phi(u)$求出。

查表得到 $\phi(2.0)=0.9772,\phi(2.24)=0.9875$

所以 $P(66.15\leqslant x\leqslant 67.04)=0.9772-(1-0.9875)=0.9617=96.47\%$

样本值落在(66.15,67.04)以外的概率为

$1-P=1-0.9647=3.53\%$

4 与正态分布有关的特殊分布

4.1 t 分布

在大样本测定中，可用大样本的方差代替总体方差 σ^2。但在一些医学研究中，只进行次数不多的测试，在这类小样本测试中，除非总体为正态分布，否则不能把样本平均值 $\bar{x}$ 视为正态分布变量。一般情况下，总体方差是不知道的，用小样本的方差 s^2 代替总体方差 σ^2 是不准确的，由此计算得到的与 u 相类似的变量，也不能视为标准正态分布的变量。因此，基于正态分布的经典误差理论不能直接用于这类小样本测定的数据处理。

在小样本测定中，用样本标准差代替总体标准差，得到的统计量 t

$$t=\frac{\bar{x}-\mu}{s_{\bar{x}}}=\frac{\bar{x}-\mu}{s}\sqrt{n}$$

遵从 t 分布(又称司都顿分布)。式中 $s_{\bar{x}}$是小样本平均值的标准差。t 分布的概率密度由司都顿分布概率密度函数

$$p(t,f)=\frac{1}{\sqrt{\pi f}}\frac{\Gamma(\frac{f+1}{2})}{\Gamma(\frac{f}{2})}(1+\frac{t^2}{f})^{-\frac{f+1}{2}}$$

$(-\infty<t<\infty)$

(下转第 75 页)

长无明显差异。12岁至16岁之间桡、尺骨面密度迅速增长。16岁以后男性骨密度值明显高于女性，并保持终生。39岁以后男女两性骨密度有显著性差异($P<0.01$)；女性桡、尺骨面密度逐渐减少起于40－49岁组，明显减少则始于50岁组。男性桡、尺骨面密度减少的情况约较女性晚10年。

本调查结果表明，正常人性别不同，桡、尺骨密度亦不同。吉林(松花江上游)地区男女两性骨密度均于35～39岁达到一生中高峰，男性桡骨面密度峰值为0.658±0.072，尺骨为0.708±0.111；女性桡骨面密度峰值为0.654±0.059，尺骨为0.681±0.061。以后随年龄的增长，骨密度逐渐减少。男性骨密度减少较女性晚10年，且下降的速度较慢。

本次调查首次获得吉林(松花江上游)地区居民桡、尺骨密度正常值，对骨质疏松症诊断和治疗方法的疗效评价，提供了客观的定量的和有价值的依据。

参考文献

1 刘忠厚主编，骨质疏松症．第1版．北京：化学工业出版社，1992：1

2 刘忠厚主编，骨质疏松研究与防治．第一版．北京：化学工业出版社，1994.10，1.

(上接第92页)

给出。式中$f=n-1$，是计算s的自由度。由上式看出，t分布只取决于自由度f。当f较小时，t分布曲线与正态分布曲线差别较大，随f逐渐变大，其越接近正态分布，当$f\to\infty$时，t分布曲线与正态分布曲线严格一致，这时t分布变成u分布。

实用上都将t分布列成表供查阅。

4.2 χ^2分布

若x是正态分布$N(\mu,\sigma^2)$的随机变量，x_1、x_2、…、x_n是相互独立的n个样本值，样本方差为s^2，统计量

$$\chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2=\frac{(n-1)}{\sigma^2}s^2$$

遵从自由度$f=n-1$的χ^2分布，χ^2分布的概率密度函数为

$$p(\chi^2,f)=\frac{1}{2^{f/2}\Gamma(\frac{f}{2})}(\chi^2)^{\frac{f-2}{2}}exp(-\frac{\chi^2}{2})$$

$$(0\leqslant x^2\leqslant\infty)$$

χ^2分布的概率密度曲线是不对称的，随着样本容量的增加，即自由度f增加，不对称性减小，当$f\to\infty$时，χ^2分布逐渐接近正态分布。

为应用方便，统计学书中都附有制成的χ^2分布表。

4.3 F分布

方差或标准差是反映测试精度的重要标志，由两个样本方差构成统计量

$$F(f_1,f_2)=\frac{s_1^2}{s_2^2},$$

其可作为检验统计量。式中s_1^2和s_2^2分别是两个遵从χ^2分布的变量，数值较大的为s_1^2，较小的为s_2^2，f_1和f_2分别是其自由度。F分布的概率密度是由菲歇尔分布函数

$$p(F,f_1,f_2)=\frac{\Gamma(\frac{f_1+f_2}{2})}{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}f_1^{\frac{f_1}{2}}f_2^{\frac{f_2}{2}}$$

$$\frac{F^{(f_1-2)/2}}{(f_2+f_1F)^{(f_1+f_2)/2}}\quad(0\leqslant F<\infty)$$

给出。它只取决于计算方差的自由度f_1和f_2。

同样，一般统计学书中也附有制成的F分布表可利用。