讲座

简明生物统计方法

王 广 仪

(吉 林 医 科 大 学)

第二讲 计数资料的统计推断方法(上)

一、用样本百分率估计总体百分率

用样本百分率($\hat{p}$)估计总体百分率(p),一般常用区间估计法,即根据$\hat{p}$求出p的95%(或99%)的置信区间。

(一)样本比较大($n>50$),且p不太靠近0或1(可由$\hat{p}$窥测)时,样本百分率趋近于正态分布,可用下列公式求置信区间的上、下限(这是最常见的方法,但样本较小或者p较偏向0或1时准确性差):

$$p_L=\hat{p}-t_{0.05(或0.01)}S_{\hat{p}} \tag{2.1}$$

$$p_u=\hat{p}+t_{0.05(或0.01)}S_{\hat{p}} \tag{2.2}$$

式中,p_L——置信区间的下限;

p_u——置信区间的上限;

$t_{0.05(或0.01)}$——自由度$df=n-1$,显著标准$\alpha=0.05$(或0.01)的t值*;

$S_{\hat{p}}$——$\hat{p}$的标准误差;其式为

$$S_{\hat{p}}=\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例 2-1 为了解麻疹减毒活疫苗接种后血凝抑制抗体产生的最早时间,在接种后15天随机抽测79名被接种者,其中有60名测出抗体,占75.9%(称抗体阳转率);问接种该疫苗15天后的总体阳转率至少能达到百分之几?

为此,应求出总体阳转率(p)的95%置信区间下限。将

$$\hat{p}=75.9\%$$

$$S_{\hat{p}}=\sqrt{\frac{75.9(100-75.9)}{79}}\%=4.81\%$$

代入式(2.1)得:

$$p_L=75.9\%-1.99^{*}\times4.81\%=66\%$$

结论:该麻疹减毒活疫苗接种15天后的阳转率至少可达66%。

(二)当样本比较小($n<50$)时,用下列公式计算置信区间较为准确:

$$p_L=\frac{n_2}{n_1F_{n_2}^{n_1}\left(\frac{\alpha}{2}\right)+n_2} \tag{2.3}$$

式中,$n_1=2(n-x+1)$,$n_2=2x$;

$$p_u=\frac{m_1F_{m_2}^{m_1}\left(\frac{\alpha}{2}\right)}{m_1F_{m_2}^{m_1}\left(\frac{\alpha}{2}\right)+m_2} \tag{2.4}$$

式中,$m_1=2(x+1)$,$m_2=2(n-x)$;$F_{n_2}^{n_1}\left(\frac{\alpha}{2}\right)$与$F_{m_2}^{m_1}\left(\frac{\alpha}{2}\right)$可查$F$值表[3]。如查95%的置信区间,即查$P=0.025$的$F_{n_2}^{n_1}$与$F_{m_2}^{m_1}$值。

例 2-2 曾用某药治疗38例慢性气管炎患者,31例有效,有效率为81.6%;问某药对慢性气管炎的普遍有效率可达多少?

为此应求出该药的总体有效率(p)的95%置信区间。因本例$n<50$,故宜用式(2.3)、式(2.4)计算。将$n=38$,$x=31$代入式(2.3)和式(2.4)得:

$$n_1=2(38-31+1)=16,$$
$$n_2=2\times31=62,$$
$$F_{62}^{16}\left(\frac{0.05}{2}\right)\approx2.04^{**},$$
$$p_L=\frac{62}{16\times2.04+62}\approx66\%;$$
$$m_1=2(31+1)=64,$$
$$m_2=2(38-31)=14,$$
$$F_{14}^{64}\left(\frac{0.05}{2}\right)\approx2.60^{**},$$
$$p_u=\frac{64\times2.60}{64\times2.60+14}\approx92\%$$

结论:该药的普遍有效率在66%—92%之间。

本例如用式(2.1)、(2.2)计算,其95%的置信区间为69%—94%,不如66%—92%较为准确。而例2-1如用式(2.3)、(2.4)计算其95%的置信区间,则结果为65%—85%,便与用式(2.1)、(2.2)计算的结果相差无几。

(三)在实际应用中,可直接由置信区间便查图(表)查出比较准确的结果。现将置信区间便查图(图

* 可由t分布表查出,见参考资料[2]第217页

** 见参考资料[3]第441、第437页

2-1，图 2-2）的用法简介如下：

1．本图系根据式（2.3）、（2.4）计算结果绘出，可直接查出样本含量为 50—1,000 的 99% 置信区间（如 $n<50$ 时，可直接查“置信区间便查表”[1]）；如用样本含量的 1.65 倍（即 $1.65n$）查本图时，即得 95% 的置信区间。

2．如果样本百分率（$\hat{p}$）大于 50%，则取从 100% 减去 $\hat{p}$ 所得之差查图，把查得的结果再从 100% 减去即得所求的置信区间；但此时从下限图查得的结果为上限，反之亦然。

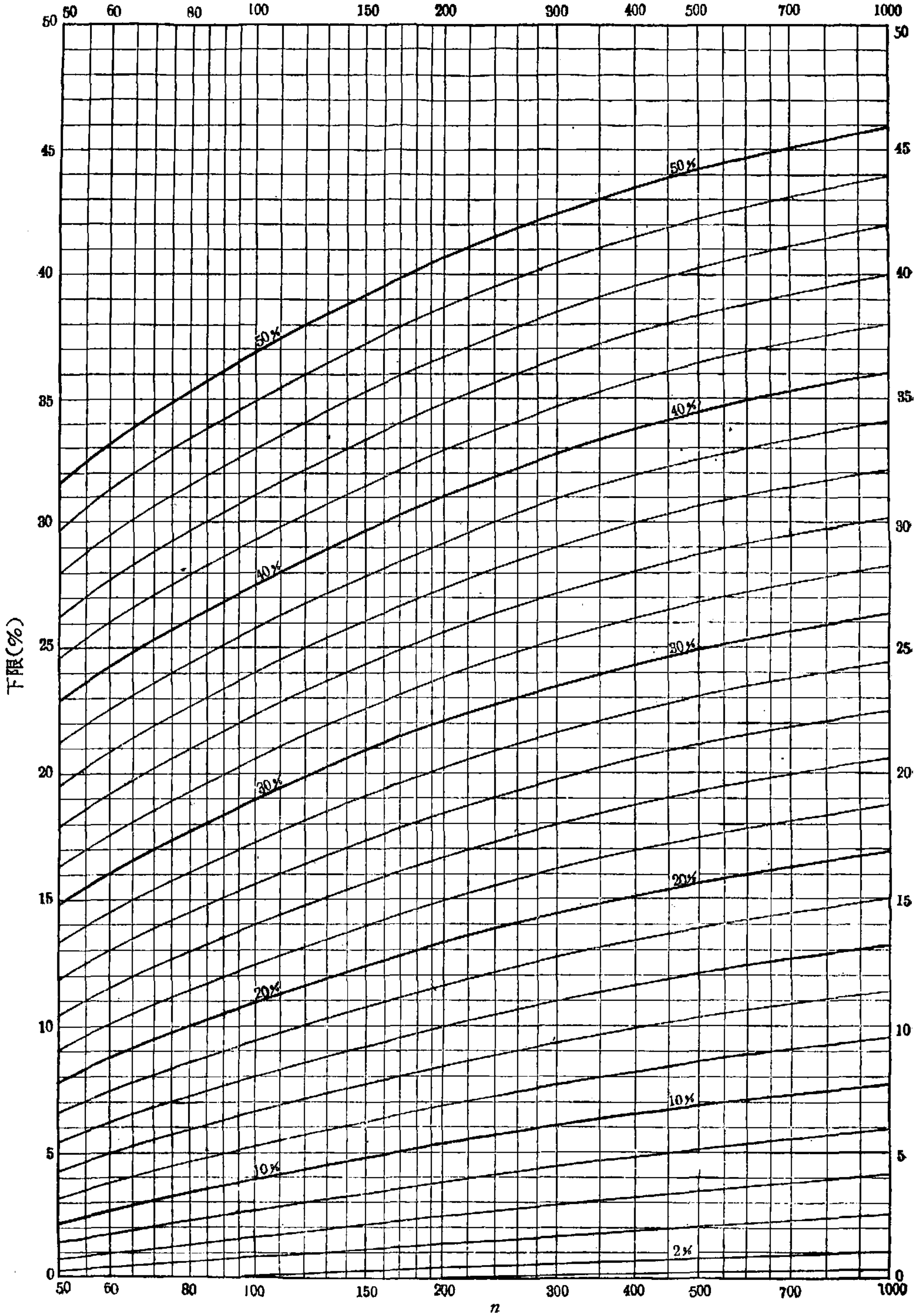

图 2-1 总体百分率（p）的 99% 置信区间下限便查图

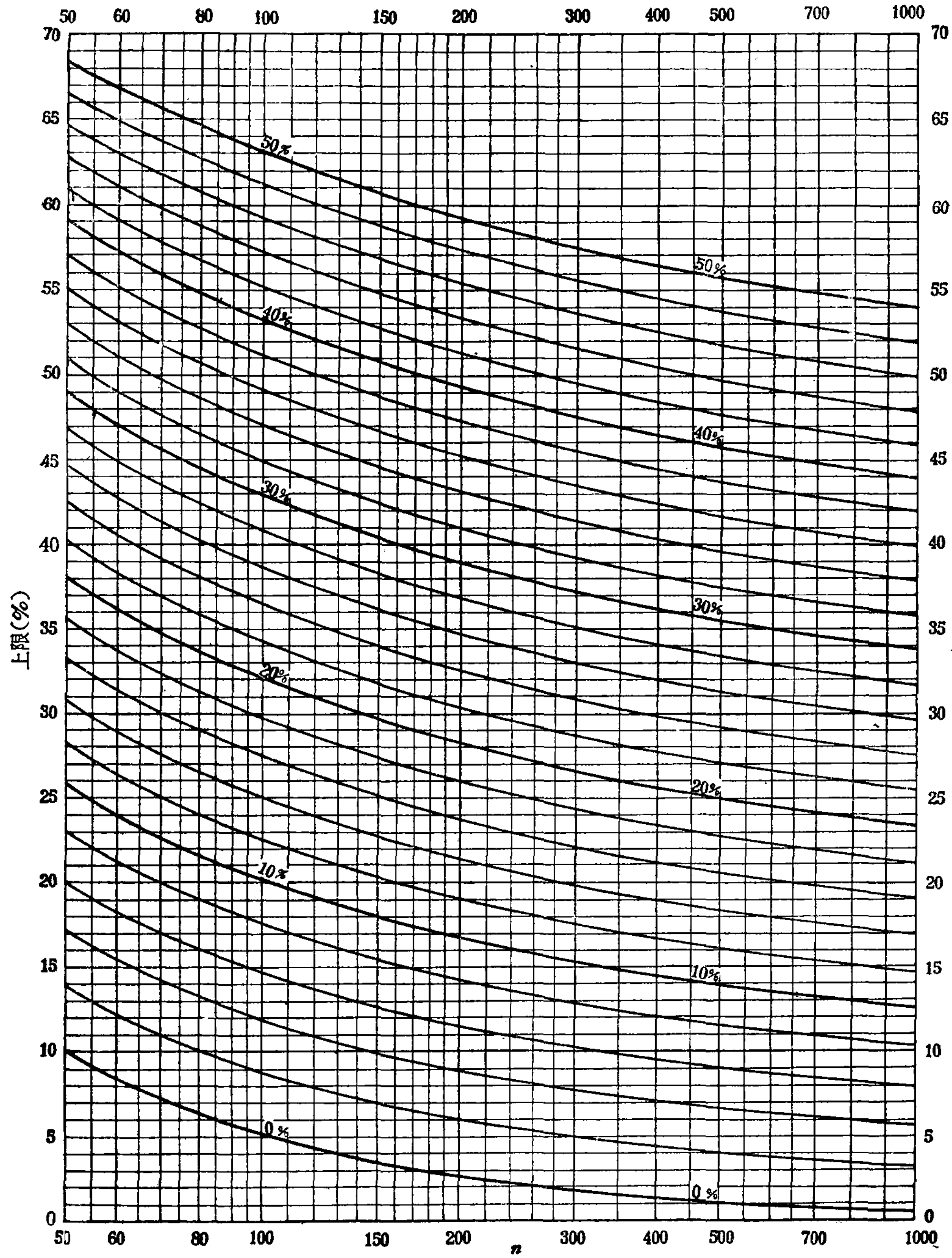

图 2-2　总体百分率 (p) 的 99% 置信区间上限便查图

3．图的横座标 (n) 为样本含量，纵座标为置信区间，图内曲线为样本百分率 ($\hat{p}$)。查图时先在横座标找到 n，如例 2-1，查 95% 的置信区间时，$n = 1.65\times79 \approx 130$；然后在图内找到相当于24.1%(100%−75.9%＝24.1%)的曲线(在曲线 24% 与 26% 之间，可由目测决定)；通过 $n = 130$ 的垂直线与 $\hat{p} = 24.1\%$ 的曲线的交点所对应之纵座标接近 15%；再从 100% 减去 15%，得 85%，即所求的 95% 置信区间上限。同理，可查得 95% 置信区间的下限为 65%。如果要查 99%的置信区间，则可直接按 $n = 79$ 查图，手续同上；可查得 99% 的置信区间为 62%—87%。

二、样本百分率的差异显著性测验方法

根据调查或实验所获数据的实际情况，一般可分为两个(或多个)样本百分率之间和两列(或多列)样本百分率之间的差异显著性测验两种。

(一)两个样本百分率的差异显著性测验方法

两个样本百分率的比较，是生物学与医学研究中最常应用的、最基本的样本百分率差异显著性测验方法。这是因为即或同时比较多个百分率时，最后也往往需要进行两两比较。因此这类测验方法也比较多，诸如：u 检验法，t 检验法，2×2表 χ^2 检验法，以及精确概率计算法（亦称直接概率计算法）等等；此外还有各种用于进行测验的工具图（表），如"统计分析纸"[2]（二项概率纸），"百分比比较简查表"[4]等。这些方法，

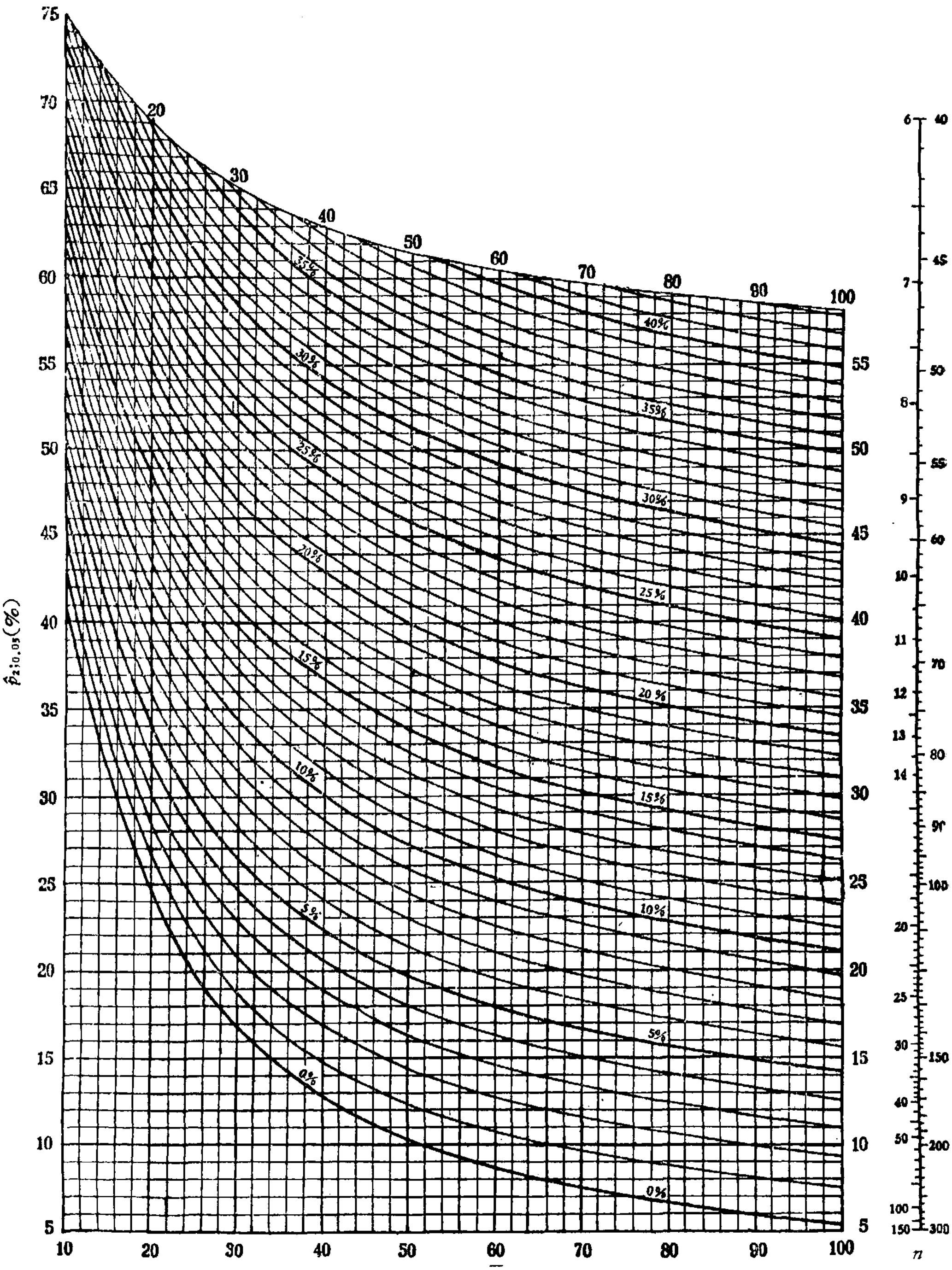

图 2-3　两个样本百分率的差异显著性测验便查图之一

$P=0.05$（双侧）

繁简不同，精度亦异，各有其优缺点。本着方法既应简便易行、结果又要精确可靠的要求，我们绘制出两个样本百分率的差异显著性测验便查图（图 2-3，图 2-4）。这是根据精确概率计算结果绘制的，其精确程度基本上相当于用四格表精确概率计算公式计算的结果。凡样本含量平均在 10—1,000 之间的任意两个样本百分率差异显著性测验，皆可由“便查图”查出。以下主要介绍“便查图”的用法，并扼要介绍其它各种测验方法及其结果，以供参考。

1. “便查图”（图 2-3，图 2-4）的用法

例 2-3 某医生用美蓝复红染色法检查流感活疫苗接种者 75 例，其中发现鼻粘膜柱状上皮细胞浆内有病毒包涵体者 35 例，阳性率为 46.7%；同时检查流行期间流感住院者 20 例，阳性者 16 例，阳性率为 80%；

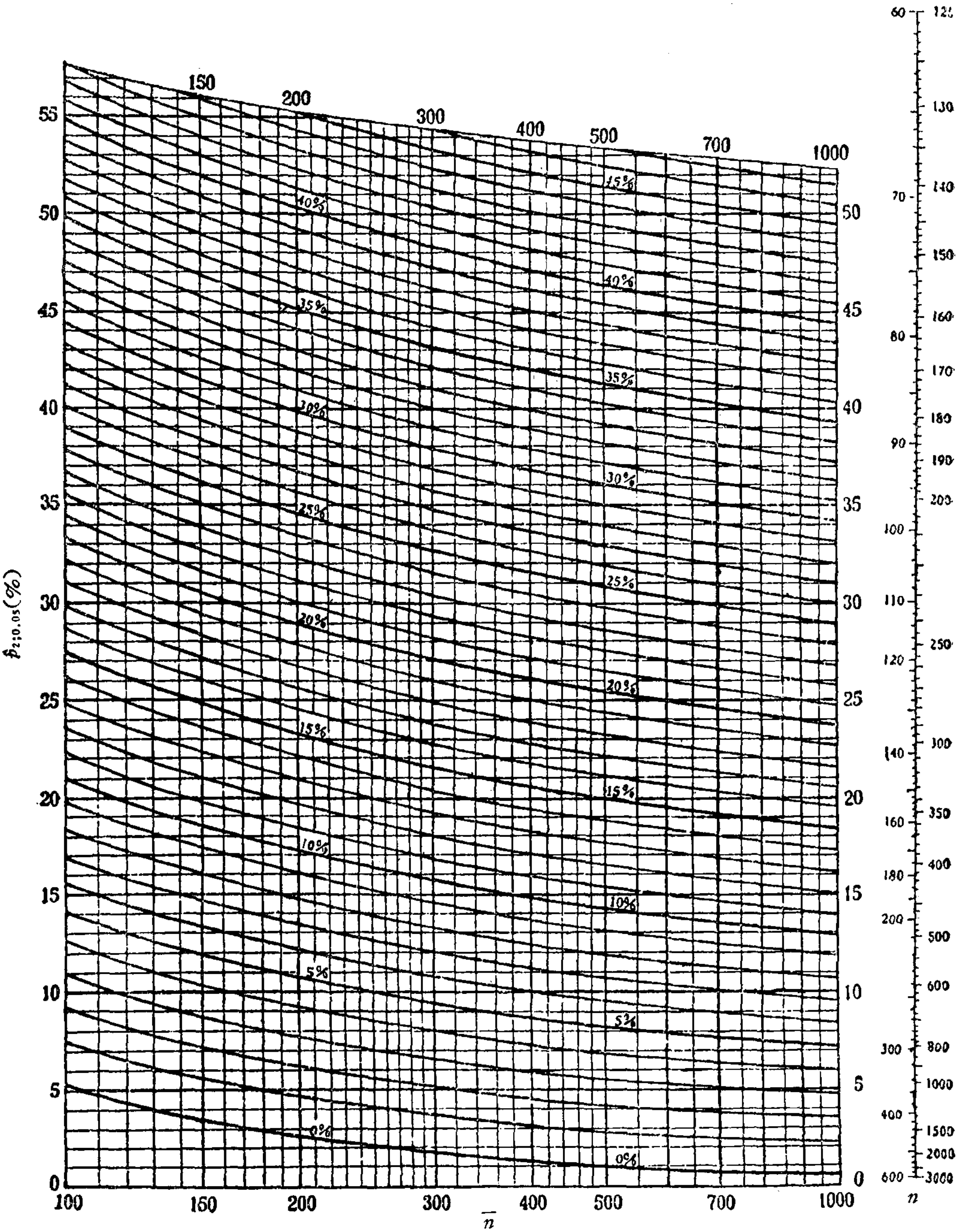

图 2-4　两个样本百分率的差异显著性测验便查图之二

$P=0.05$（双侧）

问患者的阳性率与活疫苗接种者有无显著性差异？

（1）定 $\hat{p}_1$ 与 n_1　如果两个（或有一个）样本百分率未超过50%时，即定样本百分率小者为 $\hat{p}_1$，其样本含量为 n_1。如果两个样本百分率都超过50%，或虽有一个未超过但在图内查不到相应的曲线时，则都取其从100%减去（有效率，或阳性率）后的结果（即无效率，或阴性率）进行测验，而定 $\hat{p}_1$ 与 n_1 的原则不变。据此，本例取其阴性率查图，即：

$$\hat{p}_1 = 20\%, \qquad n_1 = 20;$$
$$\hat{p}_2 = 53.3\%, \qquad n_2 = 75$$

（2）求样本含量平均数（$\bar{n}$）　若 $n_1 = n_2$，则 $\bar{n} = n_1 = n_2$；若 $n_1 \fallingdotseq n_2$，则

$$\bar{n} = \frac{2n_1n_2}{n_1 + n_2} + \frac{n_1 - n_2}{100\hat{p}_1 + a} \qquad (2.5)$$

式中，

$$a = \begin{cases} 5, 当\ n_1 > n_2; \\ 10, 当 n_1 < n_2 \end{cases}$$

上式中 $\frac{2n_1n_2}{n_1 + n_2}$ 是 n_1 及 n_2 的调和平均数，可从“便查图”旁所附之计算尺直接查出。如本例，先在尺上查到20与75，其中点（可用纸条对折法）即是其调和平均数，约为31.5。式中 $\frac{n_1 - n_2}{100\hat{p}_1 + a}$ 是校正数，如本例为 $\frac{20 - 75}{100 \times 0.20 + 10} \approx -1.8$，于是得 $\bar{n} = 31.5 - 1.8 \approx 30$。

（3）查图　图内曲线为 $\hat{p}_1$，本例 $\hat{p}_1 = 20\%$，恰好能找到一条相应的曲线（如果 $\hat{p}_1$ 有小数时，可由目测决定该曲线位置）。图的纵座标为 $\hat{p}_{2;0.05}$（即相当于 $P = 0.05$ 的 $\hat{p}_2$ 的界限值）。图的横座标为 $\bar{n}$。通过 $\bar{n} = 30$ 的垂直线与曲线 $\hat{p}_1 = 20\%$ 的交点所对应的纵座标47%，即为 $\hat{p}_{2;0.05}$。

（4）判定显著性测验结果　如果 $\hat{p}_2 \geqslant \hat{p}_{2;0.05}$，则 $P < 0.05$，差异显著；如果 $\hat{p}_2 < \hat{p}_{2;0.05}$，则 $P > 0.05$，差异不显著。本例 $\hat{p}_2 = 53.3\%$ 大于 $\hat{p}_{2;0.05} = 47\%$，$P < 0.05$，故差异显著。结论：流感活疫苗接种者的阳性率与流感住院患者的阳性率有显著性差异，前者低于后者。

如果需用 $P = 1\%$ 的标准进行显著性测验时，则只需以系数0.65乘 $\bar{n}$，即按 $0.65\bar{n}$ 查图，所对应的纵座标即为 $\hat{p}_{2;0.01}$。本例 $0.65\bar{n} = 19.5$，查得 $\hat{p}_{2;0.01} \approx 55.5\%$，它大于 $\hat{p}_2 = 53.3\%$，故此时 $P > 0.01$。

以上所述是“便查图”的一般（正规）查法。此图的特点是：例数越少，对应的 $\hat{p}_{2;0.05}$（或 $\hat{p}_{2;0.01}$）越大，因而对于 $\hat{p}_2$ 来说，如果已经超过 $n_1(n_1 < \bar{n})$ 所对应的 $\hat{p}_{2;0.05}$ 时，当然更能超过 $\bar{n}$ 所对应的 $\hat{p}_{2;0.05}$ 了；反之，如果 $\hat{p}_2$ 尚未超过 $n_2(n_2 > \bar{n})$ 所对应的 $\hat{p}_{2;0.05}$ 时，当然更不能超过 $\bar{n}$ 所对应的 $\hat{p}_{2;0.05}$ 了。因此在实际应用中，当 $n_1 \fallingdotseq n_2$，如 $n_1 < n_2$ 时，可先用 n_1 代替 $\bar{n}$ 查图，如已经“差异显著”（$\hat{p}_2 > \hat{p}_{2;0.05}$），便可下结论；如“差异不显著”（$\hat{p}_2 < \hat{p}_{2;0.05}$）时，再改用 n_2 代替 $\bar{n}$ 查图，如果差异仍不显著，亦可下结论。只有如本例情况，当用 $n_1 = 20$ 查图得“差异不显著”，再用 $n_2 = 75$ 查图又“差异显著”了，这时才需按式（2.5）求出 $\bar{n}$ 而后查图。

2. 2 × 2 表（四格表）χ^2 测验法

关于 χ^2 测验的详述，可参阅一般的统计方法书[3]。这里只是仍用例2-3，给出 2 × 2 表 χ^2 测验的计算结果，供对照比较。

先列出例2-3的 2 × 2 表，见表2-1。

表 2-1　例 2-3 的 2 × 2 表

项　目	阳性人数	阴性人数	合　计
流感活疫苗接种者	35 (a)	40 (b)	75 ($a+b$)
流感住院患者	16 (c)	4 (d)	20 ($c+d$)
合　计	51 ($a+c$)	44 ($b+d$)	95 ($a+b+c+d$)

然后把表内数据代入下列公式

$$\chi^2 = \frac{[\,|ad-bc| - (a+b+c+d)/2\,]^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} \qquad (2.6)$$

得：

$$\chi^2 = \frac{[\,|35\times4 - 40\times16| - 95/2\,]^2\times95}{75\times20\times51\times40} = 5.78$$

所得 χ^2 值大于 $\chi^2_{0.05(df=1)} = 3.841$*，$P < 0.05$，差异显著，结论同上。

严格说来，在 2 × 2 表中如有一项数据小于5（如本例 $d = 4$），则用精确概率计算法较为准确。

3. 精确概率计算法

仍用例2-3说明。此法也需先列出 2× 2表，参见表2-1，因35，40，16，4中最小者是4，故使周围合计项不变，列出由4到0的所有2×2表如下：

a_i	b_i	75
c_i	d_i	20
51	44	95

上表中，　$i = 1, 2, 3, 4, 5$

$a_{i=1-5} = 35, 34, 33, 32, 31$

$b_{i=1-5} = 40, 41, 42, 43, 44$

$c_{i=1-5} = 16, 17, 18, 19, 20$

$d_{i=1-5} = 4, 3, 2, 1, 0$

上表内的数据皆属 $\hat{p}_1 - \hat{p}_2 \geqslant 80\% - 46.7\% = 33\%$ 的。

---

* 见参考资料[3]第435页

把上表的五组数据代入下式：

$$P=\frac{(a_1+b_1)!(c_1+d_1)!(a_1+c_1)!(b_1+d_1)!}{(a_1+b_1+c_1+d_1)!}\times\sum_i\frac{1}{a_i!b_i!c_i!d_i!} \qquad (2.7)$$

则有

$$P=\frac{75!20!51!44!}{95!}\left(\frac{1}{35!40!16!4!}+\frac{1}{34!41!17!3!}+\frac{1}{33!42!18!2!}+\frac{1}{32!43!19!1!}+\frac{1}{31!44!20!0!}\right)=0.007$$

因式(2.7)的 P 是单侧概率，故用 0.025 作为双侧测验 $P=0.05$ 的标准，0.005 作为 $P=0.01$ 的标准。对于例 2-3，

$$0.005<P=0.007<0.025$$

结论同前，差异显著。

例 2-4 某生物制品所从先后制备的两批某种活菌苗里各随机抽取 200 个某种杆菌进行培养观察，结果分别有77个和88个活菌，样本活菌率分别为 38.5% 和 44.0%；问两批活菌苗的活菌率有无显著差异？

本例用"便查图"进行测验时，则

$$\hat{p}_1=38.5\%,\quad n_1=200;$$
$$\hat{p}_2=44.0\%,\quad n_2=200;$$
$$\bar{n}=n_1=n_2=200$$

查得 $\hat{p}_{2;0.05}\approx 48.6\%$，因 $\hat{p}_2<\hat{p}_{2;0.05}$，$P>0.05$，故结论：尚不能认为两批活菌苗的活菌率有明显差异。

4. u 测验法及 t 测验法

当两个样本含量 n_1 和 n_2 都比较大，而且两个样本百分率 $\hat{p}_1$ 和 $\hat{p}_2$ 都不太靠近 0% 或 100% 时，还可利用服从正态分布的 u 测验法或服从 t 分布的 t 测验法进行测验。

例 2-4 满足 u，t 测验法的条件。将有关数据代入

$$u=\frac{|\hat{p}_1-\hat{p}_2|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \qquad (2.8)$$

得：

$$u=\frac{44.0-38.5}{\sqrt{\frac{38.5(100-38.5)}{200}+\frac{44.0(100-44.0)}{200}}}=1.12$$

$u=1.12$ 小于 $u_{0.05}=1.96^*$，$P>0.05$，差异不显著。

若将例 2-4 数据代入：

$$t=\frac{|\hat{p}_1-\hat{p}_2|}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad (2.9)$$

式中，$\hat{p}_1=\frac{x_1}{n_1}$，$\hat{p}_2=\frac{x_2}{n_2}$，$\hat{p}_c=\frac{x_1+x_2}{n_1+n_2}$，自由度 $df=n_1+n_2-2$；则得

$$t=\frac{44.0-38.5}{\sqrt{41.25(100-41.25)\left(\frac{1}{200}+\frac{1}{200}\right)}}=1.12$$

再查 $df=200+200-2=398$ 的 t 值表 ($df\to\infty$)，得 $t_{0.05(df\to\infty)}=1.96^*$ 大于 $t=1.12$，结论同前，差异不显著。

参 考 资 料

[1] 吉林医科大学：医学统计方法讲义，1966。

[2] 中国科学院数学研究所统计组：常用数理统计方法，科学出版社，1973。

[3] Steel, R. G. D., Torrie, J. H.: Principles and procedures of statistics, 1960.

[4] А. Я. Бодрский: Статистические методы в экспериментальных медицинских насследованиях, 1955.

（待 续）

* 查 t 分布表 $df\to\infty$ 项，见参考资料[2]第 217 页

简 讯

提高二苯胺反应灵敏度的简易方法

在生物学研究中，脱氧戊糖-二苯胺反应是常用于定量测定 DNA 含量的方法。但是如果溶液中 DNA 浓度低于每毫升 5 微克时，这一反应往往不够灵敏。据报道，如果用乙酸戊酯把反应中形成的蓝色脱氧戊糖-二苯胺络合物从溶液中抽提出来，灵敏度就可提高 4 倍，能精确地测定溶液中浓度低至每毫升 1.5 微克的 DNA。由于乙酸戊酯能有效的把溶液中二苯胺反应特异的蓝色产物抽提到有机溶剂层，因此当样品溶液混浊时也可以直接在 595 毫微米读数，无需再扣除 700 毫微米处为校正混浊的光密度校正读数。

这种方法简单易行，只要按照常规方法，把 2 毫升含 DNA 的 10% 过氯酸溶液与 2 毫升二苯胺试剂混合，加入 0.1 毫升乙醛水溶液(浓度为 1.6 毫克/毫升)，在 56℃ 温浴一小时，待反应完全以后，在反应混合液中加入乙酸戊酯，充分摇匀，在室温下以每分钟 1,600 转的速度离心一分钟，使混合液中有机溶剂和水相分层，就可把反应产物抽提并浓缩于乙酸戊酯层中。测定有机溶剂层在 595 毫微米的光密度，即可得知 DNA 含量。

标准溶液的分析结果表明，乙酸戊酯的用量以 1 毫升为宜。这样，经过抽提的样品，光密度可提高大约 4 倍，而且可以获得光密度值与溶液浓度变化之间的良好正比关系；适用于溶液中浓度在每毫升 1.5—12.5 微克范围内的 DNA 含量测定。

[陶毓顺摘译自 *Analytical Biochemistry*, **49**, 2, 547]