

医学统计学

第一章 绪论

- 1、统计步骤：研究设计、收集资料、整理资料、分析资料
- 2、研究思路：研究总体随机抽样得到样本，对样本进行统计描述（统计表、统计图、统计指标），目的是描述样本特征，从而根据样本对总体进行统计推断（参数估计、假设检验）
- 3、数据类型：**资料类型是基本功**

数据类型	特点	举例	怎么描述?	其他
定量数据 (计量资料)	有大小, 有单位	身高、体重、血压、 温度、脉搏、WBC 数	频数表、频数 图、算指标	
定性数据 (计数资料)	有类别, 没差别	性别、阴阳性、血型	率、构成比、 相对比	二分类的都是 计数资料
有序数据 (等级资料)	有类别, 有差别	治疗效果（显效、有 效、好转、无效）		多分类并且有 等级差别的

- 4、总体：是指**研究对象的全体**，通常由所有的同质观察单位或个体组成
- 5、样本：是指从总体中选取的**有代表性的一部分**观察单位或个体，通常使用随机选取（随机抽样）方法获得。（目的：保证样本代表性）
- 6、同质：指观察单位或研究个体间具有相同或相近的性质，通常要求主要研究指标的影响因素相同或基本相同。
- 7、变异：同一种测量在总体中不同观察单位或个体之间的差别。
- 8、误差

系统误差 由一些固定因素产生，如仪器未进行归零校正、研究对象选择不合适

随机误差	随机测量误差	可以通过多次测量对真实值进行比较准确的估计。
	★抽样误差	由于抽样而引起的样本统计量与总体参数间的差异 解决办法：增加样本量 n。

9、参数与统计量

		均数	标准差	总体率	总体相关	总体回归
参数	描述总体的	μ	σ	π	ρ	β
统计量	描述样本的	\bar{X}	S	P	r	b

- 10、概率：描述随机事件出现可能性大小的定量度量。
- 11、小概率事件：发生的可能性很小，认为在一次实验中不太可能发生；如果发生，我们就有理由怀疑前提假设不成立。（在假设检验中用到）

第二章 定量资料的统计描述

- 1、描述方法：图、表、指标
- 2、编制频数表的步骤：
 - a) 求极差
 - b) 确定组段
 - c) 确定组距
 - d) 归组计数
 - e) 整理成表
- 3、绘制频数图：连续型：直方图；离散型：直条图
- 4、频数表和频数图的作用：
 - a) 便于观察数据的分布类型：正态？偏态？
 - b) 大致看出观察值的形态和特征：高度看出平均水平，两侧是离散趋势

5、描述集中趋势的统计学指标

算数均数	\bar{X}	适用于正态分布 或 近似正态分布的资料	$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X}{n}$	几组数的均值
几何均数	G	①呈倍数关系的等比资料 ②对数正态分布	$G = \sqrt[n]{X_1 X_2 \cdots X_n}$ 次方。为了计算方便,常改用对数的形式计算,即 $G = \lg^{-1} \left(\frac{\lg X_1 + \lg X_2 + \cdots + \lg X_n}{n} \right) = \lg^{-1} \left(\frac{\sum \lg X}{n} \right)$	抗体滴度细菌计数
中位数	M	①偏态分布 ②一端或两端没有确定数据的资料	$M = L + \frac{i_M}{f_M} (n \times 50\% - f_L)$	先排序,再找中间数
百分位数	P _x	③分布不明的资料	$P_x = L + \frac{i_x}{f_x} (nx\% - f_L)$ <i>注意: n 为频数, nx% 为百分位数, f_L 为小于 P_x 的频数之和</i>	通过频数表计算

6、描述离散趋势（变异程度）的统计学指标

极差	R	适用于任何资料	$R = X_{\max} - X_{\min}$	较片面缺点较多
四分位数间距	Q	偏态分布	$Q = P_{75} - P_{25}$	较片面
方差	S ²	正态分布、对称分布	$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$	
标准差	S	正态分布、对称分布	$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$	越大越偏离均数
变异系数	CV	①观察单位指标不同,如身高体重 ②同单位资料,但均数相差悬殊	$CV = \frac{S}{\bar{X}} \times 100\%$	描述相对离散程度

7、正态分布：均数、标准差

8、偏态分布：中位数、四分位数间距

第三章 正态分布

1、图形特征：以 μ 为中心，左右对称，对称区域的面积相等，曲线下总面积为 1。

2、两个参数： μ （位置参数）：增大向右移，减小向左移
 σ （形状参数）：越大越矮胖，越小越瘦高

3、标准正态分布

做 z 变换， $z = \frac{X - \mu}{\sigma}$ ， $Z \sim (0, 1)$

90%	$\mu \pm 1.64\sigma$
95%	$\mu \pm 1.96\sigma$
99%	$\mu \pm 2.58\sigma$

4、医学参考值范围：绝大多数正常人某个指标的波动范围

概率 (%)	正态分布法			百分位数法		
	双侧	单侧		双侧	单侧	
		下限	上限		下限	上限
90	$\bar{x} \pm 1.64S$	$\bar{x} - 1.28S$	$\bar{x} + 1.28S$	$P_5 \sim P_{95}$	P_{10}	P_{90}

95	$\bar{x} \pm 1.96S$	$\bar{x} - 1.64S$	$\bar{x} + 1.64S$	$P_{2.5} \sim P_{97.5}$	P_5	P_{95}
99	$\bar{x} \pm 2.58S$	$\bar{x} - 2.33S$	$\bar{x} + 2.33S$	$P_{0.5} \sim P_{99.5}$	P_1	P_{99}

5、过低异常，算单侧下限；过高异常，算单侧上限。

第四章 定性数据的统计描述

1、描述方法：率、构成比、相对比

2、率：表示频率或强度的指标 = $\frac{\text{某事物或现象发生的实际数}}{\text{可能发生该事物或现象的总例数}} \times \text{比例基数}$

3、构成比：事物内部各组成部分在整体中所占的比重 = $\frac{\text{某一组成部分的观察单位数}}{\text{所有观察单位之和}} \times 100\%$ （和为 100%）

4、相对比：A 和 B 两个有关联指标值之比。相对比 = $\frac{A}{B}$

5、使用相对数要注意的问题：

- 样本数不能太少，分母不能太小，否则误差较大，波动较大。
- 不能把构成和率相混淆。
- 注意资料的可比性。（用标准化消除可比性差）
- 正确计算合计率
- 样本率或构成比存在抽样误差。

第五章 统计表和统计图

1、表格的构成：

- 标题：上方居中
- 标目：横标目、纵标目
- 线条：只有横线，不宜使用竖线或斜线
- 数字：用阿拉伯数字表示，小数位数一致，无数字用“—”，缺数字用“…”，0 就用 0 表示。
- 备注：需要说明时可用“*”标出，将说明文字写在表格的下面。

2、统计图的构成：

- 标题：下方居中
- 图域
- 标目
- 图例
- 刻度

纵横比例 5：7

3、图形的分类

1	直方图	用于表示连续变量频数分布情况	纵轴必须从 0 开始
2	直条图	用于相互独立的资料。单式、复式条图。 考的多一点	
3	圆图	用于描述构成比	—
4	百分条图		
5	散点图	描述双变量的关系	
6	线图	用于描述随时间变换的资料。 普通线图（变化的幅度）、半对数线图（变化的速度）两种。	

第六章 参数估计与假设检验

1、标准误：描述抽样误差大小的指标， $S_{\bar{x}} = \frac{S}{\sqrt{n}}$ ，增大样本量，减小抽样误差

2、标准误越大，样本均数 \bar{X} 估计总体均数的可靠性越差。

3、标准差越大，均数的代表性越差。

t 分布

1、t 分布，对样本均数 \bar{x} 做 t 变换 $(\bar{X}-\mu) / (S/\sqrt{n})$

2、t 分布图形特征：以 0 为中心，左右对称；只有一个特征参数 $v=n-1$ ， v 增大，峰越高。当 v 为 ∞ 时，为标准正态分布。

3、了解 t 分布特征，可以查 t 界值表。

4、区间估计：按照预先指定的概率（=可信度 $1-\alpha$ ）估计总体均数所在的范围，范围就是 $1-\alpha$ 的可信区间。

t 分布法	样本量较小， $n<50$	$\bar{X} \pm t_{\alpha/2, v} S_{\bar{X}}$
正态近似法	样本量较大， $n>50$	$\bar{X} \pm 1.96 S_{\bar{X}}; \bar{X} \pm 2.58 S_{\bar{X}}$

5、可信区间的含义：区间包含总体参数的概率就是 $1-\alpha$ ，

6、准确度和精密度评价可信区间

准确度	$1-\alpha$ 的大小，越大准确度越高。	95 的准确度低一点，精密度高
精密度	指区间的宽度，越窄精度越好。	99 的准确度高一点，精密度低

假设检验

差异性	t 检验、 χ^2 检验、秩和检验
关联性	相关、回归

1、假设检验的步骤

建立假设，确定单双侧检验	$H_0: \mu = \mu_0; \pi_1 = \pi_2; M_d=0$ $H_1: \mu \neq \mu_0; \pi_1 \neq \pi_2; M_d \neq 0$
确定检验水准	$\alpha=0.05$
选定检验方法，计算检验统计量	资料类型：计量资料、计数资料、等级资料 设计类型：单样本设计、配对设计、成组设计（完全随机设计）
确定 P 值	查附表可得
作出推断结论	$P \leq \alpha$ ，拒绝 H_0 ，接受 H_1 ，差别有统计学意义，可以认为 xxx 和 xxx 是有差别的 $P > \alpha$ ，不拒绝 H_0 ，差别无统计学意义，还不能认为 xxx 和 xxx 是不同的

2、两类错误

I	H_0 成立，拒绝 H_0	假阳性	α	样本量一定的情况下， α 越大， β 越小
II	H_0 不成立，不拒绝 H_0	假阴性	β	

3、检验效能：按规定的检验水准 α 能发现其差别的概率，其值为 $1-\beta$ 。

t 检验

1、用于计量资料，目的：比较均数的差别

2、设计：单样本设计、配对设计、成组设计（完全随机设计）

单样本设计	是否满足独立性、正态性，不满足就用单样本的秩和检验	$H_0: \mu = \mu_0;$ μ_0 是已知的总体均数 $H_1: \mu \neq \mu_0;$	$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, v = n-1$	本例自由度 $v=n-1$ ，查表得 $t_{0.05/2, v} = ***$ 。因为 $t_{0.05/2, v} > t$ ，故 $P > 0.05$ ，差异无
-------	---------------------------	---	---	---

配对设计	计算差值是否服从正态分布，不服从就用配对秩和检验	$H_0: \mu_d=0;$ $H_1: \mu_d \neq 0;$	$t = \frac{\bar{d} - \mu_d}{S_d} = \frac{\bar{d} - 0}{S_d} = \frac{\bar{d}}{S_d / \sqrt{n}}$	统计学意义，按 $\alpha = 0.05$ 水准不拒绝 H_0 ，还不能认为
成组设计	<pre> graph LR A[先作正态性、方差齐性检验] --> B[正态性 方差齐性] A --> C[非正态性 方差不齐] B --> D[t 检验] C --> E[t' 检验] C --> F[变量变换] C --> G[秩和检验] </pre>	$H_0: \mu_1 = \mu_2;$ $H_1: \mu_1 \neq \mu_2;$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}, \nu = n_1 + n_2 - 2$	xxx 与 xxx 存在差异。(算 < 查, $P >$, 算 > 查, $P <$)

x² 检验

- 1、用于计数资料，目的：比较率的差别，构成比的差异。
- 2、X² 值反映了实际频数与理论频数的吻合程度。

成组设计和配对设计，两个四格表的检验思路和检验步骤

配对设计	$b+c \geq 40$	不校正	$X^2 = \frac{(b-c)^2}{b+c}, \nu=1$
	$b+c < 40$	校正	$Xc^2 = \frac{(b-c -1)^2}{b+c}, \nu=1$

- 3、R×C 表中格子的理论频数不应小于 1，并且 $1 \leq T < 5$ 的格子数不宜超过格子总数的 1/5。
- 4、 $V = (\text{行}-1) \times (\text{列}-1)$
- 5、理论频数 $P = (\text{行合计} \times \text{列合计}) / \text{总合计}$
- 6、二分类比较率，多分类比较构成比

秩和检验

- 1、参数检验和非参数检验的适用条件、特点、优点、缺点。
- 2、知道资料类型、设计类型、大概的检验步骤， $H_0 H_1$ 怎么建立假设，目的比较总体的分布位置，配对和单样本比较对差值的总体中位数 M_d ，对成组设计比较总体的分布。
- 3、检验思路，是否满足正态性，方差齐性，满足参数检验的条件就用 t 检验，不满足的成组的还有 t' 检验、变量变换、秩和检验。

相关

- 1、相关反映互依关系
- 2、线性相关：双变量正态分布的随机变量。
- 3、秩相关：非正态分布、等级资料、有不确定数值的资料
- 4、特点：-1~1 之间，绝对值越大，相关性越强，没有单位
- 5、步骤：绘制散点图、计算相关系数、作假设检验、解释相关系数

回归

- 1、回归反映依存关系
- 2、公式： $\hat{y}_i = a + bx$
- 3、b 的统计学意义：x 每增加（减）一个单位，y 平均改变 b 个单位。
- 4、回归分析的步骤：
 - a) 散点图有无线性趋势
 - b) 用最小二乘法求解 a 和 b，列回归方程
 - c) 作假设检验，推断总体回归关系是否成立， β 不为 0
- 5、假设检验的两种方法：方差分析和 t 检验
- 6、线性回归要求：线性、独立性、正态性、等方差
- 7、评价拟合效果的指标：剩余标准差和拟合系数