

## 应用统计学的基本概念（三）

在这部分，我们继续学习针对性地处理具体数据。

有三种数据类型：

第一类是分类数据。分类数据就是一些彼此之间没有数学关系的数据。既无分级也无顺序关系。例如：性别（男性或女性），种族，血型。一个人血型可以是 A 型或 B 型或 O 型或 AB 型，但不可能既是 A 又是 B 型。

等级数据就是有序或分级数据。尽管是有序资料但是组距无明确规定，并且不一定相等。例如，社会经济等级，从最易到最难分组。生活质量的评估依据或将人群任意分组，如三分组，五分组，十分组，其中的关系或顺序已知，但是这些组的组距无明确规定。例如分五组时，第一组的组距可能是从 1 到 3, 而最后一组的组距可能是从 75 到 100。组距大小可以不同但顺序是固定的。

最后还有区间和比率数据。这些是连续性数据。这些数据的共性是有等级或有顺序，并且各个数值之间具有已知的相等的间距。这类资料可以进行算术运算如加法减法。 我

们已经对正态分布的概念非常熟悉。这张幻灯片是正态分布人群的图形显示。在这里，我们看到一条线，代表人群的正态分布，X 轴代表人群百分比，Y 轴代表感兴趣的指标数值。人群 50%点值是人群均数。1 个标准差内包含 67%的人群，2 个标准差内包含 95%的人群。这就意味着，根据定义，即使是正态分布人群，也有 2.5%的人在 2 个标准差之下，2.5%的人在 2 个标准差之上。这就是说，在正态分布人群中，占很小百分比的个体，即 5%的人，其指标是在所谓的正常范围之外。

不同的检验方法用来分析不同类型的资料。分类变量通常可以使用  $\chi^2$  检验。列  $2 \times 2$  表进行  $\chi^2$  检验。 $\chi^2$  检验代表  $(\text{实际值} - \text{预期值})^2 / \text{预期值}$ 。其它类型的分析也可以用于分类变量。包括 Fisher 精确检验和 logistic 回归。当整体样本量均较小或任意一格预期值较小时，Fisher 精确检验尤为有用。logistic 回归是一种模型，评价二分变量结果(是或否)发生的概率。

这张幻灯片显示的是数据分布数值。在这种情况下，可以确定这张幻灯片左上角的具体数据是代表实际实验的真实值还是误差值。这些数据来自于我本人的临床试验。这张幻灯片显示血清胰岛素水平(Y 轴)和体重指数(X 轴)之间的

关系。体重指数介于 15-60：从体瘦到非常肥胖，可以看到体重指数的增加和空腹血清胰岛素水平增加之间的关系。

现在让我们看一下左上角的数据。这是体重指数为 22 的受试者，其预期的空腹血清胰岛素水平较低，但是实际空腹血清胰岛素水平接近 200。这是误差(实际上值是 17，小数点位置标错)还是此值真是介于 170-180 之间？这个特殊的例子确实是一个真实的实验值。这个个体代表的是胰岛素受体出现基因突变由此引起严重的胰岛素抵抗的人。这个数据针对相对体重指数与胰岛素水平的关系，将个体与正常人群之间的差异以图形形式表示。

下一张幻灯片显示图表数据的其它方式。这些是箱式图，可以提供大量信息。它们提供了中位数，数值范围，数值的变异范围，可以用于特定临床试验的不同结果的表示。在这里，具体数据是不重要的，图上的数据所代表的含义是重要的。

对连续性数据可以使用不同的统计学分析方法。我们最常使用的方法之一是所谓的 Student t 检验。比较两样本结果的均数，以确定它们相同或是不同。要使用和解释 t 检验需要知道样本量，两样本均数间差异的大小，每一样本中数

据的变异性。应注意这些条件与计算临床试验样本量所需要的条件相似。

应当记住当进行多重  $t$  检验评价同一干预的多个结果时，必须应用校正因子。也就是说，如果同一干预有 40 个不同的结果时，使用 Student  $t$  检验进行时，应该给出校正因子。否则，就有可能这些结果中的 1 或 2 个单独是由于机会的原因(偶然性)具有统计学显著性，有  $1/20$  的机会， $P$  值等于 0.05，其结论将是不正确的。

也可以使用方差分析对连续资料进行分析。方差分析适用于连续变量的多组比较，它不能区分不同的活性治疗组间效果的差异，仅能评价这些组与对照组之间的差异。另一方面，就连续性变量而言，给定  $X$  值线性回归就可以算出  $Y$  估计值。它的优点在于可以直观地观察数据，帮助识别异常数值。

这张幻灯片显示连续性数据的评价。在这里，我们看到用简易智力状态检查 (MMSE) 所测定的智力状态的变化，有记忆力障碍的个体被划分入安慰剂组或两种不同剂量的雌激素组观察：正方形表示低剂量组，三角形表示较高剂量组。注意，从这个相对较容易进行的临床试验中可以看到，与安

慰剂组相比，雌激素干预可以有效地降低记忆力的恶化率。现在常对不同类型的资料进行生存分析。我们已经逐渐认识到生存数据不仅仅是指观察的结果是死亡，最初我们是从癌症试验开始生存分析，但有时结果可以是特定的终点事件，例如，疾病复发、青春期出现、需行冠状动脉旁路移植术、或者心肌梗死。做生存分析有两种不同的方法。

一种是 Kaplan Myer 分析，它是围绕单变量对两组进行比较。另一种是 Cox 比例风险模型，它是比较几种自变量对生存时间的共同影响。这张幻灯片使用 Kaplan Myer 生存曲线显示了，外科手术干预(上图)或药物干预(下面这条线)对患有颈动脉狭窄的个体继续发展为同侧卒中的影响。从此图中你可以了解到，外科手术与药物相比(下面这条曲线)在一段时间内(X 轴代表年)可以降低发展为卒中的危险。

该图还提供了另外一些信息(图的下部)，接受外科手术或药物干预后每年分析的研究对象的数量。在一个图表中以提供大量关于临床试验的信息，包含的研究对象的数量，不同干预的结果。

我们已经学习了数据，如何把数据制成图表，如何分析数据，在临床试验设计的范围内如何解释数据。在最后一部

分我们将谈谈在临床试验设计过程中， 未对受试者的选择和分组进行严格控制而造成数据的混杂。

