第三章 正态分布与医学参考值范围

第一节 正态分布

一、正态曲线

> 正态曲线本质上是：当样本足够大、分组足够细时，频率分布对真实概率结构的连续逼近。

1️⃣ 医学数据中常见的分布形态

- 在医学研究中，**许多定量变量**的频数分布：
  - 以**均数（mean）为中心**
  - **靠近均数的观测值多**
  - **远离均数的观测值少**
  - **左右两侧基本对称**

👉 这种分布反映了“多数人接近平均水平，极端值较少”的客观现实。

2️⃣ 从频数表 → 频率分布 → 平滑曲线

以 **140 名正常成年男性红细胞计数**为例：

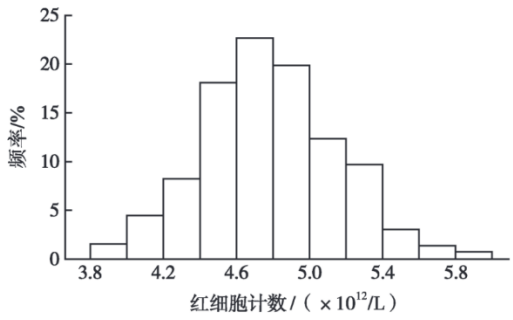

图 3-1　红细胞计数的频率分布图

(李康,贺佳主编, 2024, p. 22) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- **频数分布图**：纵轴是“频数”

- **频率分布图**：纵轴改为“频率”
- 👉 **纵轴含义不同，但图形形状完全一致**

随着：

- 样本量不断增大（n ↑）
- 分组越来越细（组距 ΔXi ↓）

📉 直方图的矩形条会：

- 越来越窄
- 顶端中点连线
- 最终逼近一条**连续、光滑的曲线**

3️⃣ 概率密度函数的意义（Probability Density Function）

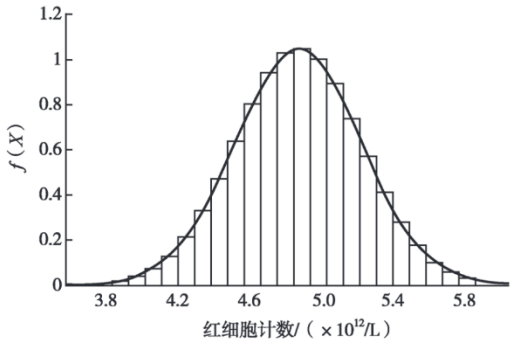

图 3-2　红细胞计数的概率密度曲线

(李康,贺佳主编, 2024, p. 22) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

这条光滑曲线对应的是：

概率密度函数 $f(X)$

其定义来自分组数据：

$$f(X) = \frac{(f_i/n)}{\Delta X_i} \qquad (3\text{-}1)$$

- f_i：第 i 组频数
- ΔX_i：第 i 组组距

- n：总例数

关键理解点：

- **每个直条的面积**

$$\Delta X_i \cdot f(X) = \frac{f_i}{n}$$

  👉 等于该组的**频率**

- 当例数足够大时：
  - **频率 ≈ 概率**
  - **曲线下总面积 = 1**

4️⃣ 正态曲线与正态分布

当概率密度曲线满足：

- 中间高、两边低
- 左右对称
- 钟形外观

➡️ 该分布**近似于数学上的正态曲线（normal curve）**
➡️ 在医学统计处理中，**可视为服从正态分布（normal distribution）**

二、正态分布的特征

> **正态分布是一类由“中心位置 + 离散程度”完全决定，其曲线下面积可直接解释为概率的连续型分布模型。**

0️⃣ 定义一句话

若连续型随机变量 X 的**概率密度函数**为

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{X-\mu}{\sigma})^2},$$

则称 X **服从正态分布**，记为

$$X \sim N(\mu, \sigma^2)$$

- μ：总体均数（mean）
- σ：总体标准差（standard deviation）
- σ^2：总体方差（variance）

---

1 形态特征（Shape）

- **单峰分布（unimodal）**
- 以 $X = \mu$ 为中心
- **左右完全对称**
- 正态曲线与 x 轴**渐近**，但永不相交
- $X \in (-\infty, +\infty)$，但极端值概率趋近于 0

👉 生物学直觉：极端个体存在，但极少。

---

2 高度与拐点（Height & Inflection）

- 曲线在 $X = \mu$ 处取最大值

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

- 距离 μ 越远，$f(X)$ 越小
- 在

$$X = \mu \pm \sigma$$

处出现**拐点（inflection points）**

👉 决定“钟形”的关键数学特征。

---

3 参数决定一切（Parameters）

正态分布**完全由 μ 与 σ 决定**:

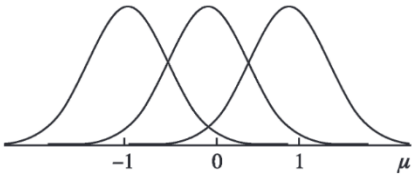

图 3-3　正态分布位置参数变化示意图（$\sigma=1$）

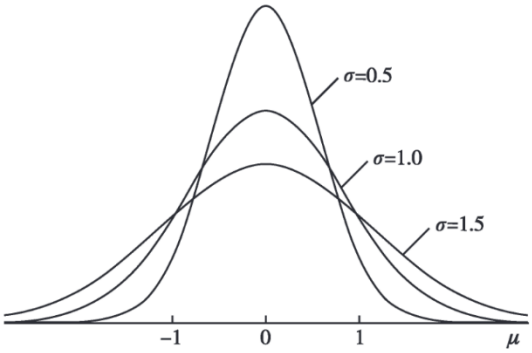

图 3-4　正态分布形状参数变化示意图（$\mu=0$）

(李康,贺佳主编, 2024, p. 23) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

（1）μ：位置参数（location parameter）

- 决定曲线在 x 轴上的**左右位置**
- 改变 μ，曲线**整体平移**
- **形状不变**

（2）σ：形状参数（scale / shape parameter）

- 决定数据的**离散程度**
- σ 大 → 曲线**矮胖** → 变异大
- σ 小 → 曲线**瘦高** → 变异小

👉 不同 μ、σ ⇒ 不同正态分布

4️⃣ 面积 = 概率（Area = Probability）

(1) 基本原则

- **曲线下面积 = 概率**
- 区间 $[a, b]$ 下的面积
  $= P(a < X < b)$

（2）整体性质

- 总面积 = 1（100%）
- 以 μ 为中心：
  - 左侧 50%
  - 右侧 50%
- 越靠近 μ，单位区间内概率越大

5️⃣ 经典面积规律

对**任意正态分布**都成立（与 μ, σ 无关）：

- $\mu \pm \sigma$：≈ **68.27%**
- $\mu \pm 1.96\sigma$：≈ **95.00%**
- $\mu \pm 2.58\sigma$：≈ **99.00%**

👉 这是：

- 置信区间
- 正态近似
- 假设检验

的**统计学核心支点**

三、标准正态分布

> **标准正态分布的意义不在于“多一种分布”，而在于：它把所有正态分布的概率问题，统一成一次查表问题。**

1️⃣ 为什么一定要“标准化”？

- 一般正态分布：

$$X \sim N(\mu, \sigma^2)$$

> 👉 **每个 μ, σ 都不一样**

- 查表、算概率如果每种都单独来 —— **不可操作**

✅ 解决方案：
把任意正态分布统一拉回到同一个“公共标尺”

2️⃣ 标准化变换（Z transformation）

定义

$$z = \frac{X - \mu}{\sigma} \qquad (3\text{-}3)$$

- 称为：**随机变量的标准化变换**（standardized transformation）
- 变换后：

$$z \sim N(0,1)$$

👉 含义非常直观：

> **z = 该观测值距离均数“差了几个标准差”**

3️⃣ 标准正态分布本身

概率密度函数

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

分布函数

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2}\, dz$$

- 中心：0
- 左右完全对称
- 曲线下面积 = 概率

4️⃣ 查表的本质逻辑（非常重要）

（1）表里给的是什么?

- **$\Phi(z)$：z 左侧的累计概率**
- 只列 **负值部分**

（2）为什么只列负数?

- 利用对称性：

$$\Phi(z) = 1 - \Phi(-z)$$

👉 正数不用单独存，**算出来即可**

5️⃣ 区间概率的通用公式（必背）

$$P(z_1 < z < z_2) = \Phi(z_2) - \Phi(z_1)$$

任何正态区间概率题，最终都会走到这一步。

6️⃣ μ、σ 未知时怎么办？（现实中最常见）

用样本估计：

$$z = \frac{X - \bar{X}}{S}$$

- $\bar{X}$：样本均数
- S：样本标准差

👉 医学统计中 **99% 的应用场景**

7️⃣ 两个“必须形成肌肉记忆”的 z 值

区间	概率	含义
$\mu \pm 1.96\sigma$	95%	95% 置信区间核心
$\mu \pm 2.58\sigma$	99%	极端值判断

👉 这两个数会在：

- 正态近似
- 置信区间
- 异常值判断
  里反复出现

8️⃣ 解题统一流程（模板）

Step 1｜写一句话“转化目标”

求某区间内的人数比例
= 求该区间下**正态曲线面积**

Step 2｜标准化（核心操作）

$$z = \frac{X - \bar{X}}{S}$$

- 下限 → z_1
- 上限 → z_2

Step 3｜转成 z 区间概率

- 单侧：

$$P(X < a) = \Phi(z)$$

- 双侧：

$$P(a < X < b) = \Phi(z_2) - \Phi(z_1)$$

Step 4 | 查表 + 对称性

- 记住：

$$\Phi(z) = 1 - \Phi(-z)$$

【例 3-1】若 $X \sim N(\mu, \sigma^2)$，试计算 X 取值在区间 $\mu \pm 1.96\sigma$ 上的概率。

先做标准化变换，求 X 所对应的 z 值，根据式（3-3）计算

$$z_1 = \frac{X_1 - \mu}{\sigma} = \frac{(\mu - 1.96\sigma) - \mu}{\sigma} = -1.96$$

$$z_2 = \frac{X_2 - \mu}{\sigma} = \frac{(\mu + 1.96\sigma) - \mu}{\sigma} = 1.96$$

通过查附表 1，由式（3-6）和式（3-7）可得

$$P(-1.96 < z < 1.96) = \Phi(1.96) - \Phi(-1.96) = (1 - \Phi(-1.96)) - \Phi(-1.96)$$
$$= 1 - 2\Phi(-1.96) = 1 - 2 \times 0.025 = 0.95$$

即 X 取值在区间 $\mu \pm 1.96\sigma$ 上的概率为 95%。同理，$P(-2.58 < z < 2.58) = 0.99$，即 X 取值在区间 $\mu \pm 2.58\sigma$ 上的概率为 99%。对于正态分布而言，1.96 和 2.58 这 2 个数值经常会被用到。

(李康,贺佳主编, 2024, p. 25) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

【例 3-2】已知某地 140 名正常成年男性红细胞计数近似服从正态分布，$\bar{X}=4.78\times10^{12}$/L，$S=0.38\times10^{12}$/L，试估计：①该地正常成年男性红细胞计数在 4.0×10^{12}/L 以下者占该地正常成年男性总数的百分比；②红细胞计数在（4.0~5.5）$\times10^{12}$/L 者占该地正常成年男性总数的百分比。

估计红细胞计数在某个范围内的人数占总人数的比例，可以转化为求此区间内正态曲线下的面积问题。

（1）将 $X=4.0$ 代入式（3-8）得

$$z=\frac{X-\bar{X}}{S}=\frac{4.0-4.78}{0.38}=-2.05$$

于是问题转化成了求标准正态分布 z 值小于−2.05 的概率，查附表 1 得 $\Phi(-2.05)=0.020\,2$，表明该地成年男性红细胞计数低于 4×10^{12}/L 者约占该地正常成年男性总数的 2.02%。

（2）$X_2=5.5$ 所对应的 z 值

$$P(4.00<X<5.50)=P\left(\frac{4.00-4.78}{0.38}<\frac{X-\mu}{\sigma}<\frac{5.50-4.78}{0.38}\right)=P(-2.05<z<1.89)$$

$$=(1-\Phi(-1.89))-\Phi(-2.05)=(1-0.029\,4)-0.020\,2=0.950\,4$$

表明红细胞计数在（4.0~5.5）$\times10^{12}$/L 者约占该地正常成年男性总数的 95.04%。

(李康,贺佳主编, 2024, p. 25) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

四、对数正态分布

对数正态分布不是一种“新分布”，而是一种“把偏态世界拉回正态秩序”的工具。

1 一句话定义

如果一个随机变量本身不服从正态分布，但它的对数值服从正态分布，那么这个随机变量服从对数正态分布。

形式化说法是：

- 若

$$Y=\ln X\sim N(\mu,\sigma^2)$$

- 则

$$X\sim \text{Log-normal}$$

👉 **正态的是“对数后的值”，不是原始数据。**

2️⃣ 典型分布形态（你一眼要能识别）

- 原始数据 X：
  - **明显右偏（正偏态）**
  - 小值多，大值少
  - 有“长右尾”
  - **下界通常是 0（或正数）**
- 对数变换后 $\log X$：
  - 分布趋于对称
  - 可能近似正态

📌 这是判断是否“值得做对数变换”的**第一视觉信号。**

3️⃣ 为什么医学和生物里特别常见?

因为这些变量的生成机制本身是“乘法型”的：

- 环境监测：
  - 有害物质浓度
  - 污染物暴露水平
- 食品安全：
  - 农药残留量
- 临床与流行病学：
  - 某些检验指标
  - 疾病潜伏期
  - 住院天数
- 血清学 / 微生物学：
  - 抗体滴度
  - 血清凝集效价
  - **倍数变化数据**

👉 **乘法累积 → 对数后变成加法 → 更接近正态**

4️⃣ 对数变换在统计上的真实目的

不是“为了好看”，而是为了三件事：

1️⃣ **减少偏态（skewness）**
2️⃣ **稳定方差（variance stabilization）**
3️⃣ **让数据满足正态分布假定**

一旦对数后的数据近似正态，就可以：

- 用均数 ± 标准差
- 用 t 检验 / 方差分析
- 用正态分布理论做区间估计

5️⃣ 对数变换的底数问题（别被绕晕）

- 常见底数：
  - $\ln$（以 e 为底）
  - $\log_{10}$
  - $\log_{2}$

📌 统计推断的**结论不取决于底数**
📌 只影响数值尺度，不影响分布形态

6️⃣ 和正态分布的本质区别（非常关键）

正态分布	对数正态分布
对称	右偏
可取负值	只能取正值
加法波动	乘法波动
均数有直观意义	**几何均数更有意义**

👉 对数正态数据中：

- **直接用原始均数往往会被极端大值“拉歪”**
- 实际解释更常用：
    - 几何均数（geometric mean）
    - 中位数（median）

五、质量控制图

1️⃣ 理论前提（这一步必须成立）

质量控制图**不是万能的**，它成立的前提是：

- 影响指标的**随机因素很多**
- 每个因素影响都不大
- **不存在明显系统误差**
  （如设备故障、操作失误、环境突变）

👉 在这种情况下：

- 指标的波动主要来自**随机误差**
- 数据 **近似服从正态分布**
- 可以使用**正态分布曲线下面积规律**

📌 这一步没想清楚，后面的“判异常”都是空谈。

2️⃣ 质量控制图的结构

图 3-8　质量控制图示意图

(李康,贺佳主编, 2024, p. 26) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

核心构成：7 条平行线

以中心线为基准：

- **中心线（CL）**：μ
- **±1σ 线**
- **±2σ 线** → ⚠️ 警戒限（warning limits）
- **±3σ 线** → ❌ 控制限（control limits）

  若 μ, σ 未知：
  👉 用样本均数 $\bar{X}$ 和标准差 S 代替

3️⃣ 质量控制图的核心判断逻辑

一句话版本

> **随机误差 ⇒ 点在 ±3σ 内随机波动**
> **系统误差 ⇒ 出现“不太可能由随机性造成的模式”**

4️⃣ 最关键的：8 种“系统误差预警情形”

这些不是背诵条款，而是**“概率异常模式”**。

---

① 1 个点越过控制限

- 距中心线 > 3σ
  👉 随机出现概率极低
  👉 **高度怀疑系统误差**

---

② 连续 3 点中有 2 点越过警戒限

- 距中心线 > 2σ
  👉 单次可能偶然
  👉 连续出现就不正常

---

③ 连续 5 点中有 4 点超过 1σ

👉 波动开始“偏向一侧”
👉 不再是纯随机

---

④ 连续 6 点单调上升或下降

👉 明显趋势性变化
👉 常见于：

- 设备老化
- 试剂变质
- 操作习惯改变

---

⑤ 连续 14 点上下交替

👉 “太有规律”
👉 反而不随机
👉 常提示人为干预或测量方式问题

---

⑥ 连续 9 点落在中心线同一侧

👉 均值发生偏移
👉 系统误差典型信号

---

⑦ 连续 15 点都在 ±1σ 内

👉 波动异常小
👉 可能：

- 数据被“修饰”
- 分辨率不足
- 人为筛选数据

---

⑧ 连续 8 点都在 1σ 以外

👉 波动异常大
👉 控制系统可能失效

5️⃣ 这些规则的统计本质（别死记）

所有规则背后只有一句话：

> **“在正态分布假设下，这种模式出现的概率太低了。”**

所以：

- 不是“违规”，而是**“概率上说不通”**

6️⃣ 质量控制图在医学中的真实角色

- 不是为了“证明数据好看”
- 而是为了**第一时间发现异常来源**

应用场景包括：

- 临床检验质量控制
- 实验室长期监测
- 卫生管理指标追踪

👉 **它是“统计 + 现场管理”的交汇点**

第二节 医学参考值范围

一、医学参考值范围的概念

> **医学参考值范围不是“正常值”，而是：**
> **在生物变异存在的前提下，用统计学为“正常”划出的概率区间。**

1️⃣ 一句话定义（先立住概念）

> **医学参考值范围，是指从“正常人群”中获得的指标观测值，在统计意义上覆盖大多数个体的波动区间。**

关键点有三层含义（缺一不可）：

- 对象是：**正常人（reference population）**
- 内容是：**个体值的波动范围**
- 方法是：**统计学建立的区间，而不是某个固定值**

2️⃣ 为什么不能用“一个数”当标准?

因为**生物变异不可避免**:

- 个体间差异:
  - 遗传
  - 性别
  - 年龄
- 个体内变异:
  - 时间
  - 环境
  - 生理状态

👉 即便都是“正常人”:

- 测量值也不可能完全相同
  👉 即便是同一个人:
- 不同时点的结果也会波动

📌 所以:
“正常”不是一个点，而是一个区间。

3️⃣ 医学参考值范围的严格统计含义

不是经验拍脑袋，而是:

> **从选定的参照总体中，收集个体观测值，用统计方法建立百分位数界限，由此得到个体值的波动区间。**

🔑 常用规则

- **95% 参考值范围**（最常见）
  - 覆盖“正常人群中 95% 的个体”
  - 剩余 5%:

- 2.5% 在下限之外
- 2.5% 在上限之外

⚠️ 重要提醒（考试和临床都爱考）：

落在参考值范围外 ≠ 一定是病人
落在范围内 ≠ 一定没问题

这是概率意义，不是诊断结论。

4️⃣ 两个核心应用方向（必须分清）

（1）临床医学：**个体判定**

用途：

- 判断某个个体指标
- 是否“可能异常”

例如：

- 成年人白细胞计数

$$(3.5 \sim 9.5) \times 10^9/\mathrm{L}$$

👉 本质是：

- **分类 / 划界**
- 为临床决策提供参考，而不是下诊断

（2）预防医学 / 公共卫生：**人群评价**

用途：

- 评价群体发育或健康水平
- 制定等级或分层标准

例如：

- 不同年龄、性别儿童的：
  - 身高
  - 体重
  - 生化指标

👉 本质是：

- **人群比较**
- **健康监测与干预依据**

5️⃣ 和前面正态分布知识的关系

- 正态分布
- 百分位数
- 95% 区间
- 对数正态（处理偏态指标）

👉 **医学参考值范围 = 统计分布 + 百分位界限 + 实际解释**

二、制订医学参考值范围的注意事项

> **医学参考值范围不是算出来的，是在正确人群、可靠数据、清楚目的前提下“被允许算出来的”。**

1️⃣ 参照总体必须“同质”（这是第一性原则）

❌ 常见误解

“正常人 = 什么病都没有的人”

✅ 正确理解

> **“正常人”是：对该研究指标而言，没有关键干扰因素的人群**

- 绝对健康不存在

- 每个人都可能有轻微异常
- 关键在于：**是否影响所研究的指标**

📌 例：制订 **SGPT（ALT）参考值范围**
必须排除：

- 肝、肾、心、脑、肌肉疾病
- 近期使用肝毒性药物
- 检测前剧烈运动

同时必须考虑并**分层或分组**：

- 性别
- 年龄
- 地区（如高原 vs 平原）
- 民族
- 妊娠状态
- 时间因素（季节等）

👉 **参照总体不“干净”，后面所有统计都是假的。**

2️⃣ 参照样本量要“够大”（否则区间不稳）

- 没有统一的“标准例数”
- 但遵循一个逻辑：

数据情况　　样本量需求

近似正态、变异小 相对可少

明显偏态、离散大 必须更多

📌 原则一句话：

样本量不足 → 参考值范围不可靠

3️⃣ 必须严控检测误差（否则“假异常”泛滥）

控制目标

- 随机误差：尽量小
- 系统误差：必须避免
- 过失误差：必须杜绝

实操层面包括：

- 方法标准化
- 仪器灵敏度一致
- 试剂质量控制
- 操作人员培训
- 环境条件统一（温度、季节、运动、饮食、妊娠等）
- 必要时重复测定

👉 **检测误差不控制，参考值范围没有任何临床价值。**

4️⃣ 单侧 vs 双侧界值：不是数学问题，是医学问题

判断原则：

> **异常是“只往一边”还是“两边都不行”**

- **双侧界值**（上下限都有）
  - 如：白细胞计数（高、低都异常）
- **单侧界值**（只设一端）
  - 如：
    - 肺活量（只低异常）
    - 血铅（只高异常）

📌 这是**专业判断先行，统计随后**的典型场景。

---

5 百分数范围的选择 = 假阳性 vs 假阴性的权衡

常用范围

- 80%、90%、95%、97%、99%
- **95% 最常用，但不是“唯一正确”**

核心权衡逻辑

研究目的	百分数范围	结果倾向
减少假阳性（确诊、科研）	大（95–99%）	少误诊
减少假阴性（初筛）	小（80–90%）	少漏诊

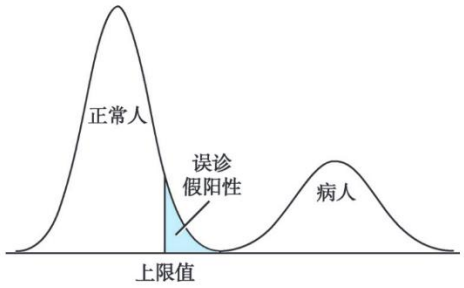

图 3-9　两组人群的数据分布无重叠示意图

(李康,贺佳主编, 2024, p. 27) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

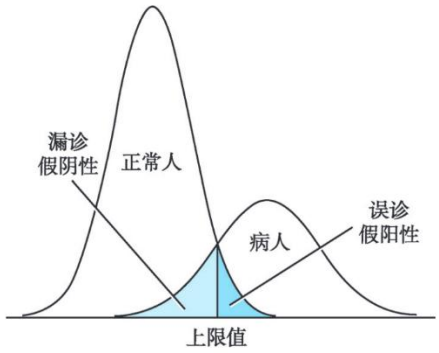

图 3-10　两组人群的数据分布有重叠示意图

(李康,贺佳主编, 2024, p. 27) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

📌 当“正常人”和“病人”分布**明显重叠**时：

- 可能需要设立：
  - **可疑值范围**
- 范围不宜过大
- 只覆盖主要重叠区域

👉 临床诊断中常结合 **ROC 曲线**综合判断。

6️⃣ 计算方法必须“服从数据”，不能反过来

方法选择原则

数据特征	推荐方法
近似正态	正态近似法
对数后正态	对数正态法
明显非正态	百分位数法
复杂分布	曲线拟合法

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields