深入浅出统计学(1) 信息图形化

1. 基本概念

1.1. 数据与信息

- 数据指的是手机的原始事实和数字。
- 信息指的是加入了某种意义的数据。

1.2. 频数

- 表示在一个特定组、或在一个特定区间内的统计对象的数目。类似于数数。
- 在设计以百分数为表现内容的图形时：设法指出频数——或是将频数标在图形中间，或标在图形旁边。
- 若没有频数，只有百分比，要小心，有可能隐藏真实数据。

1.3. 类别数据与数值型数据

- 类别数据：用以描述某类的性质或特征，也称为定性数据。
- 数据型数据：涉及数字，数值具有数字的意义，还设计计量和计数，由于数值型描述的是数量，也称为定量数据。

1.4. 累计频数

- 向原来的综合中添加新值得到的总和。

2. 图形化

2.1. 饼状图

方便对比不同类别各自所占的比例。

2.2. 条形图：

- 能对频数的相对大小进行比较。
- 相对于饼状图，优点在于更精确。

- 条形图可以是水平（体现百分数）或垂直（体现频数）条形图。
- 特殊条形图：
    - 堆积条形图（用不同长方形代表不同指标的频数）。
    - 分段条形图，可同事体现频数和百分比（即同一长方形中显示不同类别数据）。

2.3. 直方图

- 直方图用长方形表示得分范围，而不是表示一项（条形图）。
- 与条形图外观区别：
    - 每个长方形的面积和频数成正比。
    - 图上的长方形之间没有间隔。
- 求直方图过程：
    - 长方形面积=每组频数
    - 频数 = 长方形宽度 * 长方形高度
    - 长方形高度（即频数密度） = 频数/长方形宽度
- 为什么用频数代表面积：保证每个组的相对大小和与数据成正比，且不失真实。

2.4. 折线图

- 很好的体现数据的趋势。
- 与条形图比较：
    - 很容易添加新的数据，不会让图形面目不清。
    - 更好体现数据趋势。
    - 虽然能够显示数值，但不如条形图清晰。
- 只应用于展示数值型数据，不应用于类别数据。

深入浅出统计学(2) 集中趋势的度量

1. 均值

- 为什么不叫平均数：因为平均数不止一种。
- 如何计算：这……不用说了吧。
- 符号：μ。
- 异常值的影响：
  - 异常值会导致数据倾斜。
  - 向左倾斜/向右倾斜/对称数据。
- 不能应用于类别数据。
- 在数据非常对称，且仅显示出一种趋势时使用。

2. 中位数

- 中位数：另一种平均数。
- 用于解决异常值问题。
- 求解：
  - 排序。
  - 如果有奇数个数值，则以中间位置的数据为中位数。
  - 如果有偶数个数值，以中间两个数的均值作为中位数。
- 异常情况
  - 如 `1 1 1 1 1 1 1 1 1 1 11 11 11 11 11 11 11 11 11 11`，平均数和中位数都是 6，但不能表达数据的情况。
- 不能应用于类别数据。
- 数据由于异常值而发生偏斜时使用

3. 众数

- 平均数的一种。
- 一批数字中最常见的数，即频数最大的数值。
- 如果一批数据中有两个众数，成为双峰数据。
- 一批数据中的所有众数，组成众数组。
- 应用场景：众数数目较少，或数据为类别数据而不是数据型数据。

深入浅出统计学(3) 强大的"距"

1. 全距

- 定义：全距，又名极差，计算方法为`上界（最大值） - 下界（最小值）`。
- 作用：用于度量数据集分散程度的方法。
- 缺陷：
    - 全距仅仅描述了数据的宽度，没有描述数据的分布形式。
    - 容易受到异常值的影响。

2. 迷你距

- 定义：不在度量数据集全距，而是找出全距的一部分（不包含异常值的部分）。
- 作用：忽略异常值。
- 分类：
    - 四分位距（IQR）：
        - 定义：迷你距的一种，计算方法为`上四分位数 - 下四分位数`。
        - 优点：与全距相比，较少收到异常值影响。
        - 通过三个数字将数据集分为四部分，三个数字称为：下四分位数（第一四分位数）、中位数、上四分位数（第三四分位数）。
    - 百分位数：不常用，对于划分名词、排名很有用。
    - 十分位数
- 箱线图（或箱型图）
    - 箱的左右两边分别代表下四分位数和上四分位数。
    - 箱中话一条线，标出中位数。
    - 箱两边画先，表示全距的上界和下界。

3. 方差与标准差

- 作用：度量数据分散情况。
- 方差：数值与均值距离的平方的平均数。
    - 计算公式：

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

$$\sigma^2 = \frac{\sum x^2}{n} - \mu^2$$

- 标准差：方差开根号。
  - 意义：根据与均值的距离指出分散性。
  - 符号：σ
  - 有时候，会用`距离均值若干个标准差`来表示特定数值的相对位置。

4. 标准分

- 定义（标准分以字母 `z` 表示）：

$$z = \frac{x-\mu}{\sigma}$$

- 作用：比较不同数据集中的数值。使用后，可以将所有数值视为来自同一数据集和数据分布，从而进行比较。
- 其他：
  - 理论上新分布的均值为 0，标准差为 1。
  - 标准分 = 距离均值的标准差个数。

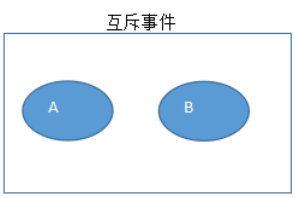
深入浅出统计学(4) 概率计算

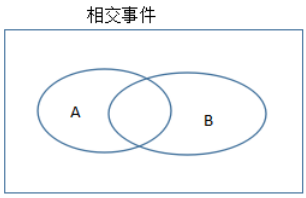
1. 概率基本概念

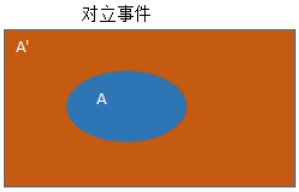
- 概率：度量某事发生几率的数量指标。
  - 进一步理解：概率只是对事件发生可能性的一种表达，概率并非担保。
- 事件：有概率可言的一个结果或一件事。
- 计算公式：$P(A)=\frac{n(A)}{n(S)}$ ，其中 S 称为概率空间，或样本空间。
- 概率的直观表现形式：
  - 维恩图
  - 概率树
- 对立事件: “A 不发生”事件可以用 A'表示。A'被称为 A 的对立事件。A'包含事件 A 所不包含的任何事件。 P(A')=1-P(A)
- 互斥事件
- 相交事件
- 独立事件：几个事件互相不影响。P(A|B)=P(A). 如果两个事件相互独立，则 P(A∩B)=P(A|B)P(B)=P(A)P(B)
- 穷举事件：表示两个事件的并为全集。
- 全概率公式：根据条件概率计算一个特定事件的全概率。P(B)=P(A∩B)+P(A'∩B)=P(A)* P(B|A)+P(A')* P(B|A')
- 贝叶斯定理：提供了一种计算逆条件概率的方法，再无法预知每种概率的情况下，非常有用。贝叶斯定理：已知 P(A),P(B|A),P(B|A');求 P(A|B).

  P(A|B) = P(A∩B) / P(B) = P(A)* P(B|A) / P(A)* P(B|A)+P(A')* P(B|A')

- 公式：P(A∪B)=P(A)+P(B)-P(A∩B)
- 条件概率：P(A|B)=P(A∩B)/P(B)
- 相关事件： 如果 P(A|B)不等于 P(A)，就说事件 A 与事件 B 的概率相互影响。







深入浅出统计学(5) 离散概率分布的应用

1. 期望和方差的引入

- 概率的缺陷：无法指出发生这些事情的整体影响。
- 期望的作用：利用概率预测长期结果。
- 方差的作用：度量预测结果的确定性。

2. 期望和方差的计算方法

- 计算都需要使用 概率分布
- $E(X)=\sum xP(X=x)$
- $Var(X)=E(X-\mu)^2$
- 标准差：$\sigma=\sqrt{Var}$

3. 线性变换公式

- $E(aX+b)=aE(x)+b$
- $Var(aX+b)=a^2Var(X)$

4. 相互独立的两个随机变量

- $E(X)+E(Y)=E(X+Y)$
- $E(X)-E(Y)=E(X-Y)$
- $Var(X+Y)=Var(X)+Var(Y)$
- $Var(X-Y)=Var(X)+Var(Y)$
- $E(aX+bY)=aE(X)+bE(Y)$
- $E(aX-bY)=aE(X)-bE(Y)$
- $Var(aX+bY)=a^2Var(X)+b^2Var(Y)$
- $Var(aX-bY)=a^2Var(X)+b^2Var(Y)$

深入浅出统计学(6) 排序、排位、排

1. 排位

- 普通排位：有 n 个对象进行排位，方式一共有 $n!$。
- 圆形排位：有 n 个对象进行圆形排位，方式一共有$(n-1)!$。
  - 圆形排位定义：n 个对象围成一个圈。
  - 如果圆形排位，且要考虑对象的绝对位置，则排位方式一共有 $n!$。
- 按类型排位：
  - 对 n 个对象排位，按类型进行排位。其中包括第一类对象 k 个，第二类对象 j 个，第三类对象 m 个……则排位方式一共有

$$\frac{n!}{j!k!m!\dots}$$

2. 排列与组合

- 排列：
  - 定义：一个较大（n 个）对象群体中取出一定的数目(r 个）对象进行排序。
  - 计算方式：$\frac{n!}{(n-r)!}$.
  - 特点：与顺序有关。
- 组合：
  - 从 n 个对象中选取 r 个对象，不必进行排序。
  - 计算方式：$\frac{n!}{r!(n-r)!}$。
  - 特点：与顺序无关。

深入浅出统计学(7) 几何分布、二项分布及泊松分布

1. 几何分布

- 条件：
  - 进行一系列相互独立试验。
  - 每一次试验都存在成功和失败的可能，且每次可能性都相同。
  - 想得到的结果是，为了取得第一次成功所需要进行多少次试验。
- 表示：

$$X\sim Geo(p)$$

第 r 次试验取得成功的概率：$P(X=r)=pq^{r-1}$

- 需要 r 次以上才能获得第一个成功的概率：$P(X>r)=q^r$
- 需要试验 r 次或不到 r 次即可取得第一次成功的概率：$P(X<=r)=1-q^r$
- 期望：$E(X)=1/p$
- 方差：$Var(X)=q/p^2$

2. 二项式分布

- 条件：
  - 进行一系列独立试验。
  - 每一次试验都存在成功和失败的可能，且每次成功概率相同。
  - 试验次数有限。
- 与几何分布的不同之处：
  - 几何分布感兴趣的是取得第一次成功所需要进行多少次试验。
  - 二项式分布感兴趣的是获得成功的次数。
- 表示：

$$X\sim B(n,p)$$

- 在 n 次试验中，取得 r 次成功的概率为：$P(X=r)=C_r^n p^r q^{n-r}$
- 期望：$E(X)=np$
- 方差：$Var(X)=npq$

3. 泊松分布

- 条件：
  - 单独时间在给定区间内随机、独立地发生，给定区间可以是时间或空间。

    - 一直该区间内的时间平均发生的次数（或者叫做发生率），且为有限数值。该时间平均发生次数通常用希腊字母 λ 表示。
- 表示：

$$X\sim Po(\lambda)$$

- 给定区间内发生 r 次时间的概率是：$P(X=r)=\frac{e^{-\lambda}\lambda^{r}}{r!}$
- 期望：$E(X)=\lambda$
- 方差：$Var(X)=\lambda$

深入浅出统计学(8) 正态分布的运用

1. 连续随机变量

- 用概率密度函数来描述连续随机变量的概率分布。
- 连续随机变量的概率通过面积表示。
- 处理连续数据时，所计算的是一个数值范围的概率。

2. 正态分布

- 如果一个连续随机变量 X 符合均值为 μ、标准差为 σ 的正泰分布，则通常写作

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

- 无论图形多大，概率密度永远不会等于 0。
- 计算方式：
  - 确定分布的范围
  - 标准化
  - 查找概率
- $P(Z>z)=1-P(Z<z)$
- $P(a<Z<b)=P(Z<b)-P(Z<a)$

深入浅出统计学(9) 再谈正态分布的运用

线性变换与独立观察

- 线性变换：
  - 概念：描述了数据的基本变化。
  - $aX+b \sim N(a\mu+b, a^2\sigma^2)$
- 独立观察：
  - 概念：描述了有多少数值。
  - $X_1+X_2,+\ldots+X_n \sim N(n\mu, n\sigma^2)$
- 两者区别：
  - 线性变换影响概率分布中的基本数值。
  - 独立观察影响所处理的事件的数量

几个常用计算公式：

- 如果 X 和 Y 为独立变量，则
  - $X+Y \sim N(\mu_x+\mu_y, \sigma^2_x+\sigma^2_y)$
  - $X-Y \sim N(\mu_x-\mu_y, \sigma^2_x+\sigma^2_y)$
- 如果 X 满足正态分布，且 a 和 b 都是数字，则：
  - $aX+b \sim N(a\mu+b, a^2\sigma^2)$
- 如果 $X_1X_2..X_nX1X2..Xn$ 是独立的观察结果，则
  - $X_1+X_2+\ldots+X_n \sim N(n\mu, n\sigma^2)$

正态分布近似代替其他分布

- 替代二项分布：
  - 如果 $X \sim B(n,p)$，且 np > 5, nq > 5，则可以使用 $X \sim N(np, npq)$近似替代二项分布。
  - 连续性修正：
    - 对于 <= ，要增加一个额外的 0.5。
    - 对于 >= ，要减去一个额外的 0.5。
    - 对于 a <= X <= b，是上述两种的合并，要计算 a - 0.5 <= X <= b + 0.5 。
    - 对于 < ，要减去一个额外的 0.5。
    - 对于 > ，要增加一个额外的 0.5。
- 近似泊松分布：
  - 如果 $X \sim Po(\lambda)$且 $\lambda>15$，则可用 $X \sim N(\lambda, \lambda)$进行近似

深入浅出统计学(10) 统计抽样的运用

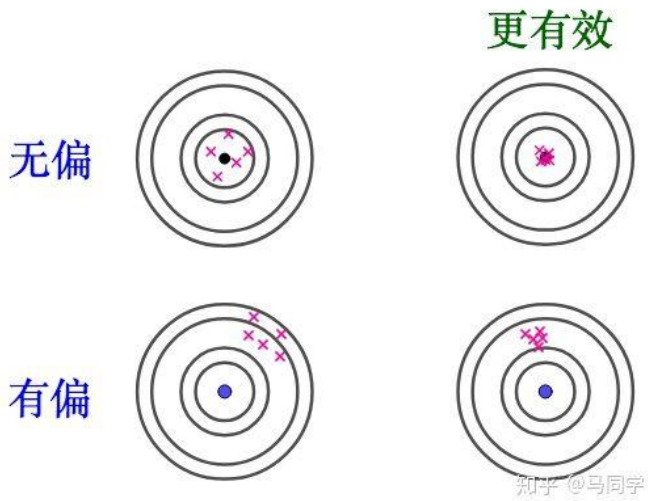
- 总目标：成功收集数据。

样本与总体

- 总体：统计学上的总体指的是对其进行测量、研究或分析的整个群体。
  - 普查指的是对总体进行研究或调查。
- 样本：一个统计样本就是从总体中选取的一部分对象。
- 样本调查：仅对总体的一个样本进行的调查或研究称为样本调查。
  - 大多数情况下，对样本调查比进行普查更切实可行，通常样本调查所费的时间和费用都比较低，且不用考虑总体。
- 如何设计样本：
  - 确定总体目标。总体目标指的是正在研究、并打算为其采集结果的群体。
  - 确定抽样单位。
  - 确定抽样空间。需要列一张表，对目标总体范围内的所有抽样单位，最好给每个取个名或编个号。这张表就叫抽样空间。

偏倚

- 无偏样本：可以代表目标总体，即该样本与总体样本具有相似特性，可以利用这些特性对总体本身做出判断。
- 偏倚样本：无法代表目标总体，由于样本与总体的特性的不相似，无法根据样本对总体作出判断。
- 偏倚来源：
  - 抽样空间中条目不齐全，因此未包含目标总体中的所有对象。
  - 抽样单位不正确。
  - 为样本选取的一个个抽样单位未出现在实际样本中。
  - 调查问卷的问题设计不当。
  - 样本缺乏随机性。
- 偏倚来源广泛，大部分归咎于样本选取方法。



样本选择

- 简单随机抽样：通过随机过程选取一个大小为 m 的样本，所有大小为 n 的可能样本被选中的可能性都相同。
- 简单随机抽样的具体方法：
  - 重复抽样：在选取一个样本单位并记录下这个抽样单位的相关信息之后，再将这个单位放回总体。
  - 不重复抽样：不再将抽样单位放回总体。
- 简单随机抽样的主要方法：
  - 抽签
  - 随即编号
- 分层抽样：将总体分割为几个相似的组，每个组具有类似的特性。
- 整体抽样：对群进行简单随机抽样，然后对每一个群的各种特性进行调查。
- 系统抽样：按照某种顺序列出总体名单，然后每 k 个单位进行一次调查，其中 k 为一个特定数字。

深入浅出统计学(11) 总体和样本的估计

1. 均值

- 样本均值被称为总体均值的点估计量，即，作为一个基于样本数据的计算结果，它给出了总体均值的良好估计。
- 点估计量是由样本数据得出，是对总体参数的估计。
- 样本均值：$\bar{x} = \frac{\sum x}{n}$
- 总体均值：μ
- 总体均值点估计量：$\hat{\mu}$

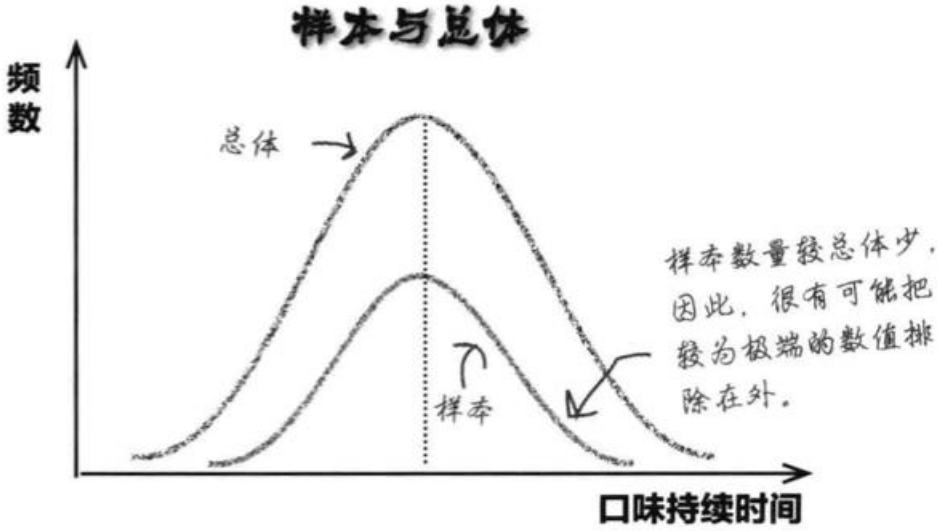
2. 方差：

- 样本数据的方差可能不是总体方差的最好估计办法。
- 总体方差计算：$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$
- 总体方差估计量：$\hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
- 使用 n-1 的原因：结果与总体方差的数值更接近，大部分情况下，样本估计的总体方差比真实方差略小一些。

3. 比例与概率

- 计算成功概率的方法与计算成功比例的方法完全一样。
- p = probability(概率) = proportion(比例)
- 样本比例可以作为总体比例的点估计。

4. 为样本计算概率



- 总体过程：
  - 查看与特定样本大小相同的所有样本。
  - 观察所有样本形成的分布，然后求出比例的期望和方差。
  - 得出上述比例后，利用该分布求出概率。
- 抽样分布：从一个总体中用相同的办法抽取许多大小相同，但存在差异的样本，然后用每个样本的某个属性形成一个分布，则所得的结果为抽样分布。
  - 由此可见，用每个样本的比例形成的抽样分布就是比例的抽样分布。通过 `P_s` 来代表样本比例随机变量。
  - 期望：$E(P_s)=p$
  - 方差：$Var(P_s)=pq/n$
  - 比例标准误差：$\sqrt{Var(P_s)}$
  - 如果 n > 30，则 P_sPs 符合正态分布，即 $P_s \sim N(p,pq/n)$，需要进行连续性修正。

5. 中心极限定理

- 计算样本均值的概率：
  - 查看我们所研究的样本大小相同的所有可能性样本。
  - 查看所有样本形成的分布，求出样本均值的期望和方差。
  - 得知样本均值的分布后，用该分布求出概率。
- 均值的抽样分布：为我们提供了一种计算样本均值的概率的方法。
- 均值的分布：
  - 期望：$E(\bar{X}) = \mu$
  - 方差：$Var(\bar{X}) = \frac{\sigma^2}{n}$
  - 均值标准误差：$\frac{\sigma}{\sqrt{n}}$

- 中心极限定理：
  - 定义：如果从一个非正态总体中去除一个样本，且样本很大，则 $\bar{X}$ 的分布近似为正态分布。
  - n 很大指的是 n > 30。
  - $\bar{X} \sim N(\mu, \sigma^2/n)$
- 点估计量与抽样分布的关系：
  - 总体均值的点估计量、均值的抽样分布的期望、样本均值的期望，三者相等。

深入浅出统计学(12) 置信区间的构建

通过样本估计总体：点估计与区间估计

点估计：估计一个精确值（点）

点估计问题：依赖唯一样本进行精确估计。虽然能够确保估计无偏，但无偏是在平均抽样 n 的的结果上，参数估计平均下来无偏。具体到每一个样本，往往是有细微偏差的，具体偏差多少，无法确定。

区间估计：在点估计的基础上，提供一个误差界限，形成一个取值范围，叫做置信区间。” 总体参数落在区间[a,b]内“这一结果具有特定概率，这个概率，叫做置信程度(多高概率，即多大把握估计正确)
目的：降低准确度（从点到区间），提高命中率。

置信区间

- 另一种估计总体统计量的方法，考虑了不确定性。
- 点估计量的缺陷：存在小小的误差。
- 求解过程：选择总体统计量，求出其抽样分布，决定置信水平，求出置信上下限。
- 置信水平：表明对于“置信区间包含总体统计量”这一说法有多大把握。
    - 置信水平越高，区间越宽，包含总体统计量几率越大。
    - 把置信区间弄得太宽会导致其市区意义。

置信区间的构造流程

置信区间，目的是根据样本构造一个区间，然后希望这个区间可以把真值包含进去，但是并不知道这个真值是多少。

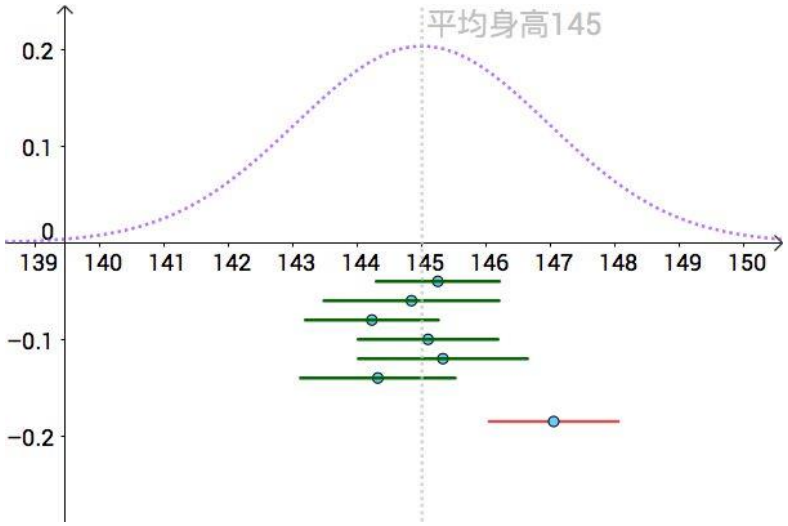
求解步骤

1. 选择总体统计量
2. 求出其抽样概率分布（总体参数的期望与方差）
    - 若总体参数已知，则直接代入计算
    - 若总体参数未知，使用样本点估计近似
3. 确定置信水平
    - 常见置信水平为 95%。
4. 求出置信上下限
    - 通过置信水平，概率表查并求出总体统计量的取值范围。

置信水平

置信水平即我们构造的置信区间，成功将总体参数包括在内的概率。
若置信水平为 95%的含义：若重复抽样 n 次，每次抽样构造一次置信区间，我们一共构造了 n 个置信区间。

例：如下图，蓝点为每次样本的点估计，横线为每次样本的区间估计。我们一共重复了 7 次抽样，其中 6 次横线都将真实值 mu=145 包含在内，除了红色的那根。则置信水平为 6/7 = 0.86



PS: 显著水平 P-value = 1-置信水平。若置信水平为 95%，则显著水平为 0.05。显著水平 P-value 的详细介绍，见”假设检验“相关章节。

置信区间的便捷公式

总体统计量是否为正态分布	总体方差是否已知	n 是否足够大	样本抽样分布	置信区间

t 分布

- 正态分布的缺陷：并非任何情况都能进行良好近似。
  - 原因：可能不知道总体方差的确切值，因此必须利用样本书估计方差；样本太小时，估计值可能出现较大误差。
- 定义：$T\sim t(v)$，其中 v 是自由度，且 $T=\frac{\bar{X}-\mu}{s/\sqrt{n}}$
- 在用小样本估计总体方差时，t 分布更精确。

深入浅出统计学(13) 假设检验的运用

假设检验

- 目标：判断一个假设是否可信。
- 假设检验：做出假设或断言，对照证据进行检验。
- 步骤：
    - 确定要进行检验的假设。
    - 选择检验统计量。
    - 确定用于做出决策的拒绝域。
    - 求出检验统计量的 p 值。
    - 查看样本结果是否位于拒绝域内。
    - 做出决策。

详细过程

- 确定假设
    - 所需要检验的断言被成为原假设。
    - 与原假设对立的被成为备择假设。
    - 原假设与备择假设不用覆盖所有可能。
- 选择检验统计量
    - 检验统计量：用于对假设进行检验的统计量，是与该检验关系最为密切的统计量。
- 确定拒绝域
    - 拒绝域：一组数值，给出反驳元假设的最极端证据。
    - 为求拒绝域，先定显著性水平，即所度量的一种愿望，希望在样本结果不可能程度达到多大时，就拒绝原假设，一般选择 5%或 1%。
    - 检验分类：
        - 单尾检验：检验的拒绝域在可能的数据集的一侧。
        - 双尾检验：拒绝域一分为二位于数据集的两侧。
- 求出 p 值
    - 定义：某个小于或等于拒绝域方向上的一个样本数值的概率。
    - 为取得样本中的各种结果或取得拒绝域方向上的某些更为极端的结果的概率。
- 样本结果位于拒绝域中吗。
- 做出决策。

第一类错误与第二类错误

- 即使证据很有力，也无法确定断言是错误的。
- 假设检验可能出现的错误有两种：

- 第一类错误：错误地拒绝真是假设。
    - P(第一类错误)=α，其中 α 为假设的显著性水平。
- 第二类错误：错误地接受假的原假设。
    - P(第二类错误)=β
    - 计算过程：检查是否拥有 H_1 的特定数值，求检验拒绝域以外的数值范围，假定 H_1 为真，求得到这些数值的概率。

	接受原假设H_0	拒绝原假设H_0
原假设H_0为真	✔	第一类假设
原假设H_0为假	第二类假设	✔

- 功效：在 H_0 为假的情况下，拒绝 H_0 的概率。
    - 功效=$1-\beta$

错误概率的计算

第一类错误概率：显著水平

发生了第一类错误，则我们拒绝了原假设，即原假设的发生概率落于拒绝域内。
故而发生第一类错误的概率，等于原假设落于拒绝域内的概率，等于显著水平 alpha。
P(第一类错误) = alpha

第二类错误概率

第二类错误概率，则原假设错误情况下，接受原假设的概率。
即备择假设正确的情况下，接受原假设的条件概率。
PS: 要计算第二类错误概率，必须拥有备择假设 H_1 的具体数值，否则无法计算。因为需要根据备择假设构建新的置信区间。

计算步骤如下

1. 我们接受了原假设：根据原假设的置信区间，确定接受原假设时，检验统计量 X 所需要的取值范围
2. 备择假设正确：根据备择假设，得到新的概率分布
3. 根据概率分布，得到 X 取值范围的发生概率，即为第二类错误概率。

深入浅出统计学(14) χ^2 分布

- 目标：利用 χ^2 分布，判断期望与事实之间存在的差别。
- 举例：老虎机赢钱概率较高，要进行某种假设检验，检查观察频数与期望频数之前的差别。

χ^2 分布

- 检验统计量：

$$X^2 = \sum \frac{(O-E)^2}{E}$$

  - 提供了一种对观察频数与期望频数之间的差异进行量度的办法
  - 值越小，总差异值越小。
- 主要用途：
  - 检测拟合优度，即检验一组给定的数据与指定分布的吻合程度。
  - 检验两个变量的独立性，即检查变量之间是否存在某种关联。
- ν：自由度数目，用于计算检验统计量的独立变量的数目，也可以说是独立信息段的数目。
  - 计算方式：ν=(组数)−(限制数)
- α：显著性。
- 单尾检验，右尾作为拒绝域。

χ^2 假设检验

- 步骤：
  - 确定要进行检验的假设及其备择假设。
  - 求出期望频数和自由度。
  - 确定用于做决策的拒绝域。
  - 计算检验统计量 X^2。
  - 查验检验统计量是否位于拒绝域内。
  - 做出决策。
- χ^2 检验是假设检验的特殊形式，总使用右尾。

深入浅出统计学(15) 相关与回归

引入

- 相关与回归的目标：说明变量之间的关系，发现事物关系的秘诀
- 单变量数据：考虑的是一个单一变量的频数或概率。
- 二变量数据。
  - 其中一个变量以某种方式受到控制，或者被用来解释另一个变量，则成为自变量，另一个变量称为因变量。
  - 可视化：散点图，用于显示数据之间的相关性。
- 线性：
  - 如果散点图上的点几乎呈线性分布，则相关性为线性。
  - 正线性相关、负线性相关、不相关。

深入

- 两个变量之间存在相关关系，并不意味着一个变量会影响另一个变量，也不意味着两者存在实际关系。
  - 相关关系意味着存在数学关系，但并不一定是实际关系。
- 最小二乘回归：
  - 引入：最好地接近所有数据点的线被称为最佳拟合线。
  - 实际点与拟合直线的距离平方之和被称为误差平方和，即$SS-E=\sum(y-\hat{y})^2$。
  - 计算直线 y=ax+b，其中

$$b=\frac{\sum((x-\bar{x})(y-\bar{y}))}{\sum(x-\bar{x})^2}$$

  - 再通过平均数求 a

$$a=\bar{y}-b\bar{x}$$

  - 预测一个特定 x 值对应的 y 值时，应避免对已知数据点范围以外的值进行预测。
- 相关系数
  - 作用：判断拟合直线的准确性。

$r=\frac{bs_x}{s_y}$，其中$s_x=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$，$s_y=\sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$

  -