第7讲 - 置信区间

本讲的目标是基于样本数据估计总体参数。

	样本统计量	**总体参数**
比例	$\hat{p}$	p
mean	$\bar{X}$	μ
标准差	s	σ^2
均值差异ff	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$
配对均值差ff	$\bar{d}$	μ_d

定义 一个 点估计量 $\hat{\theta}$ 是一个适用于一组样本数据以获得参数θ的点估计值的公式。所得值称为点估计。

在本课程中，重要的点估计量有：

点估计量	**总体参数**
$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$	$\mu =$ population mean
$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$	$\sigma^2 =$ population variance
$\hat{p} = \frac{k}{n}$ (where $k =$ number of “successes”)	$p =$ 总体比例
$\bar{X}_1 - \bar{X}_2$ (based on samples from two populations)	$\mu_1 - \mu_2 =$ difference of two population means
$\hat{p}_1 - \hat{p}_2$ (based on samples from two populations)	$p_1 - p_2 =$ difference of two population proportions

Example 假设你随机抽取了 $n = 10$ 名 BCIT 学生。对于每位学生，你记录变量 $X = \text{Age}$，得到的数据集为：

21	20	20	28	42
31	19	20	18	25

样本均值为： $\bar{X} = \frac{\Sigma X}{n} = \frac{24}{10} = 24.4$

我们的点估计 μ (总体均值) 为 24.4。

注意：点估计并不保证正确。事实上，我们预期任何点估计都会存在随机误差：

$$\theta = \hat{\theta} + \text{error}$$

幸运的是，中心极限定理为我们提供了一种估计误差大小 的方法。然后我们可以把对θ 的估计表述为一个区间 ，而不仅仅是一个点估计。

什么是置信区间?

置信区间是对总体参数θ的一个区间估计 $(\hat{\theta} - E, \hat{\theta} + E)$ 。

点估计 $\hat{\theta}$ 是区间的中心；误差幅度 E 延伸到置信区间的两端。我们认为 θ 位于该区间内，并以一定程度的置信度相信这一点。

μ 的置信区间

在深入复杂细节之前，我们先看看这在实践中是如何运作的。

示例 假设你随机抽取了 $n = 100$ 名 BCIT 学生。变量 $X =$ age 的样本数据如下所示。

26,	31,	26,	17,	26,	22,	22,	26,	31,	18,
30,	20,	30,	23,	28,	25,	26,	23,	24,	26,
33,	35,	27,	29,	25,	31,	25,	27,	22,	34,
17,	31,	22,	27,	25,	22,	17,	24,	27,	31,
32,	28,	27,	22,	26,	27,	27,	23,	21,	23,
31,	20,	25,	28,	28,	21,	23,	26,	25,	27,
18,	26,	27,	32,	26,	22,	23,	23,	26,	29,
28,	18,	17,	20,	29,	24,	28,	27,	26,	17,
25,	21,	26,	32,	32,	24,	28,	24,	25,	33,
31,	22,	25,	23,	26,	24,	25,	22,	30,	27

根据样本数据，你计算得到了样本统计量

$$n = 100$$

$$\bar{X} = 25.52$$

$$s^2 = 16.9996$$

由此我们计算出如下95%置信区间：

$$\text{lower limit} = \bar{X} - 1.984 \times \frac{s}{\sqrt{n}} =$$

$$\text{upper limit} = \bar{X} - 1.984 \times \frac{s}{\sqrt{n}} =$$

（数字1.9840是一个临界t-分数。）

结论：我们有 95% 的把握认为

使用 Z 表（已知 σ）

总体均值的置信区间 μ： X 变量来源于:

- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

- $\bar{X}$ 在以下任一情况下服从正态分布:
    - X 本身是正态变量，或
    - n 足够大（通常为$n \geq 30$）

这些观点表明，对于$\bar{X}$，Z分数服从标准正态分布

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

因此有95%的概率

$$-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

解出μ 告诉我们

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

注意： 这种形式的 95% 置信区间要求已知 σ（用于底层变量 X）。在不知道 μ 的情况下却设想我们会知道 σ，这是不现实的。

示例 设 $X =$ 为一个人在更换手机前使用手机的时长（以月为单位）。假设我们不知道 μ，但不知何故，我们知道 $\sigma = 13.0$ 月。

求 μ 的 95% 置信区间。

1. 为随机样本（样本量为 $n = 50$）收集 X 数据。

2. 计算样本统计量：

$$\bar{X} = 32.5$$
$$s = 13.2$$

对于此问题，我们假设 $\boldsymbol{\sigma = 13.0}$。

3. 计算误差幅度：

$$E =$$

4. 计算置信区间的上下限：

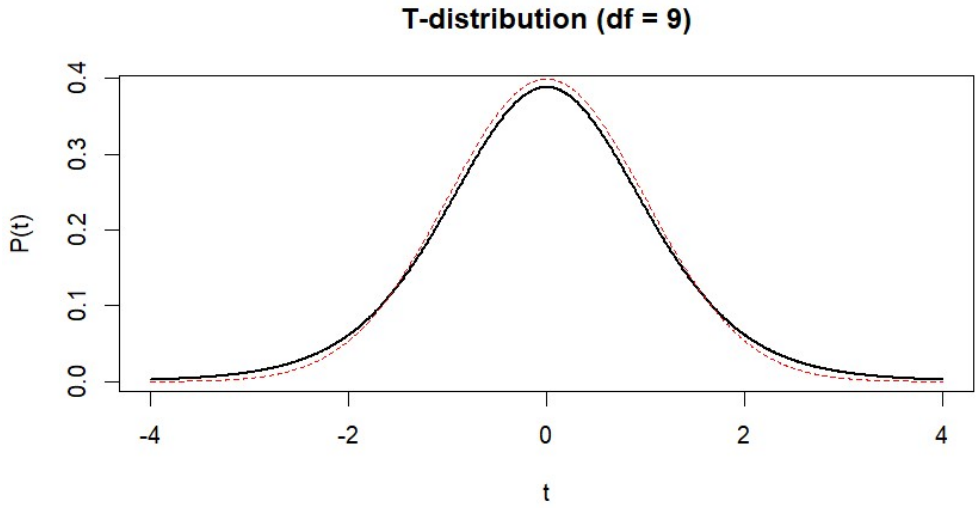
5. 结论：

使用 t 分布表（σ 未知）

在现实情况下，我们不知道 σ。于是我们使用 t 统计量：

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

变量 T 遵循 *Student's t-distribution* T_{n-1}，具有 $n-1$ “自由度”。该分布与标准正态接近，但在“尾部”包含更多概率。



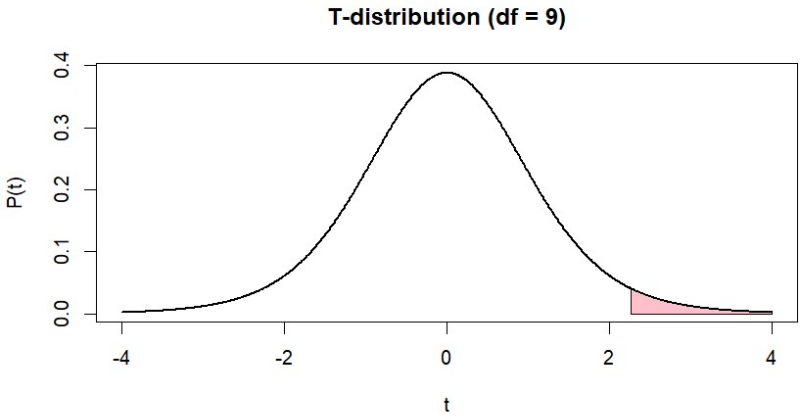
Critical t-Values

对于置信区间，我们常常需要 *critical t-value*

$t_{\alpha/2}$ = the value such that $P(T > t_{\alpha/2}) = \frac{\alpha}{2}$

我们使用以下公式计算临界 t 值：

- 学生t分布表，或
- the R function:
  `qt(1-alpha/2, df = n-1)`



Example Namzor 公司希望检验其硬盘抵御高温的能力。为控制成本，随机抽取了 10 个硬盘进行测试。故障温度似乎服从正态分布。根据样本，我们得到：

$$\bar{X} = 50.0\ ℃$$

$$s = 3.0\ ℃$$

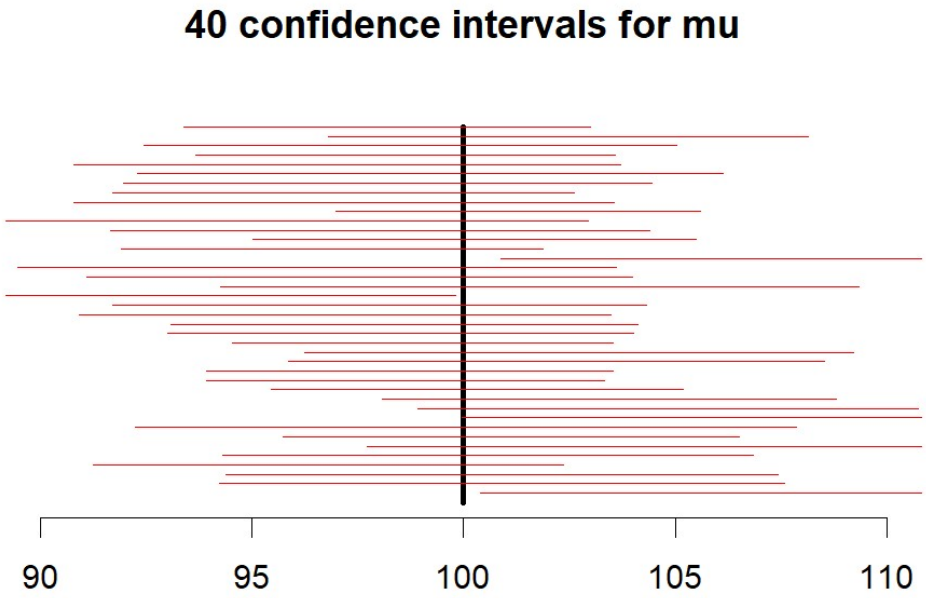
a. 构建μ的95%置信区间，即失效温度的总体均值。

b. 构造总体均值的 90% 置信区间。

我们说的 *95%* 置信度是什么意思?

我们知道，相对于 μ，有 95% 的概率 $\bar{X}$ 位于误差幅度 E 内。然而，说有 95% 的概率 μ 位于置信区间 $(\bar{X}-E, \bar{X}+E)$ 内是没有道理的，为什么？ *not*

是置信区间是随机的，而不是 μ!



我们可以说的是：*can*

如果我们从随机样本中收集数据，那么所得置信区间有 95% 的概率包含真实的 μ。

换句话说，如果我们能为许多 不同的随机样本构建置信区间，那么在 95% 的样本中，μ 会落在所得的置信区间内。

为简便起见，我们这样表述:

> “我们有 95% 的把握认为 μ 位于该区间内。”

示例 与访问内部存储器（RAM）相比，访问磁盘存储要慢得多。对于某一特定系统，针对访问磁盘存储所需时间进行了 35 次测量。测得平均值为 0.0293 秒，标准差为 0.0032。

a. 求真实平均磁盘访问时间 μ 的 95% 置信区间。

b. 如果我们错误地使用了临界Z 值（1.96）而不是T，95%置信区间会是多少?

确定样本量

我们已经基于已知数据构建了置信区间。假设我们尚未收集样本，那么我们会想确定为了在估计总体参数时达到某一误差幅度，所需的样本量 n 。

从下面开始

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

我们可以解出 n，得到：

$$n = \left(z_{\alpha/2}\frac{\sigma}{E}\right)^2$$

示例（确定样本量）求出在 BCIT 估计计算机平均使用年限所需的样本量。假设有 95% 的置信度，样本均值的误差不超过 0.25 年。在先前的研究中，BCIT 计算机年龄的标准差为 $\sigma = 0.5$ 年。