

数据分析-P1 项目作业

提交人：彭冰

- 1、 自变量：文字内容和打印颜色是否一致，取值为一致和不一致两种；

因变量：参与者说出列表墨色所消耗的时间

- 2、 建立假设：

在此任务中描述同一组样本量为 24 的参与者在一致文字条件和不一致文字条件下的测试情况，两个样本来自同一群人，所以它们是相关样本。又因为样本量为 $24 < 30$ 且需验证的内容为根据样本分析总体均值，故在此任务中，使用基于相关样本的 T 检验，检测方式为双尾检测 ($\alpha=0.05$)。

$H_0: \mu_1 - \mu_0 = 0$ 零假设，即两组测试中参与者完成时间保持一致

$H_1: \mu_1 - \mu_0 \neq 0$ 对立假设

- 3、 建立统计检验：

因为此任务中，对同样的 24 位参与者设定两组实验内容，实验目标为测试参与者在两组试验中的表现情况，即根据分析样本量为 24 的抽样分布推测对两次测试过程的总体均值的置信区间情况，故选择 T 检验，其中视两次试验是相关的。

使用工具：office excel；Tableau

- 4、 符号定义：

G_0 ：一致文字条件（以下简称第一组实验）；

G_1 ：不一致文字条件（以下简称第二组实验）；

μ_0 ：一致文字条件下的总体均值；

μ_1 ：不一致文字条件下的总体均值；

$H_0 : \mu_1 - \mu_0 = 0$ 零假设, 即两组测试中参与者完成时间保持一致

$H_1 : \mu_1 - \mu_0 \neq 0$ 或 $\mu_1 - \mu_0 < 0$ 或 $\mu_1 - \mu_0 > 0$ 对立假设

\bar{X}_0 : 第一组实验样本均值

\bar{X}_1 : 第二组实验样本均值

N_0 : $N_0 = 24$ 第一组实验样本量

N_1 : $N_1 = 24$ 第二组实验样本量

$\bar{X}_1 - \bar{X}_0$: 样本差异点估计

\bar{d} : 样本平均差异

S_d : 样本差异的标准偏差

$SE(\bar{d})$: 差异标准差

DF_0 : $DF_0 = 24 - 1 = 23$ 第一组实验样本自由度

DF_1 : $DF_1 = 24 - 1 = 23$ 第二组实验样本自由度

SD_0 : 第一组实验数据标准差

SD_1 : 第二组实验数据标准差

5、数据集的描述性统计 (集中趋势和变异测量)

$\bar{X}_0 = 14.05113$; 第一组测试中样本均值

$\bar{X}_1 = 22.01592$; 第二组测试中样本均值

$SD_0 : 3.559358$; 第一组测试中标准偏差 ;

$SD_1 : 4.797057$; 第二组测试中标准偏差 ;

6、 数据可视化

图 1 和图 2 分别为一致文字条件测试数据可视化和不一致文字测试条件可视化，从图中的数据分布可以看出两种情况下的抽样分布情况大致服从正态分布，且图 2 的数据分布更偏向横坐标 $+\infty$ 方向，假设期望的情况的是做不一致文字条件测试的平均使用时间要大于一致文字条件测试的平均使用时间。

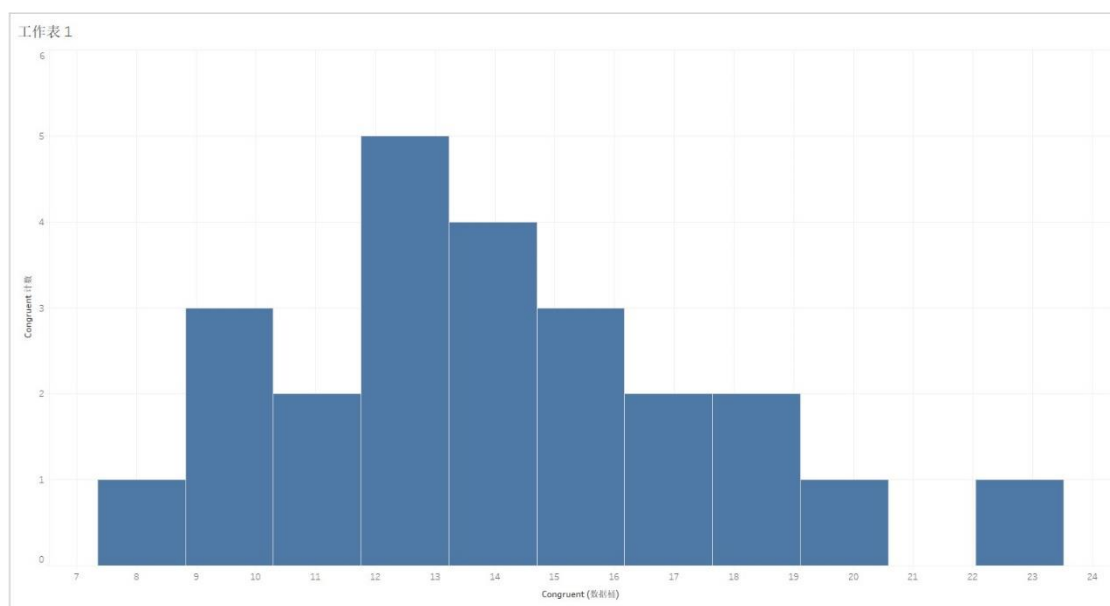


图 1 G_0 一致文字测试条件下的数据可视化

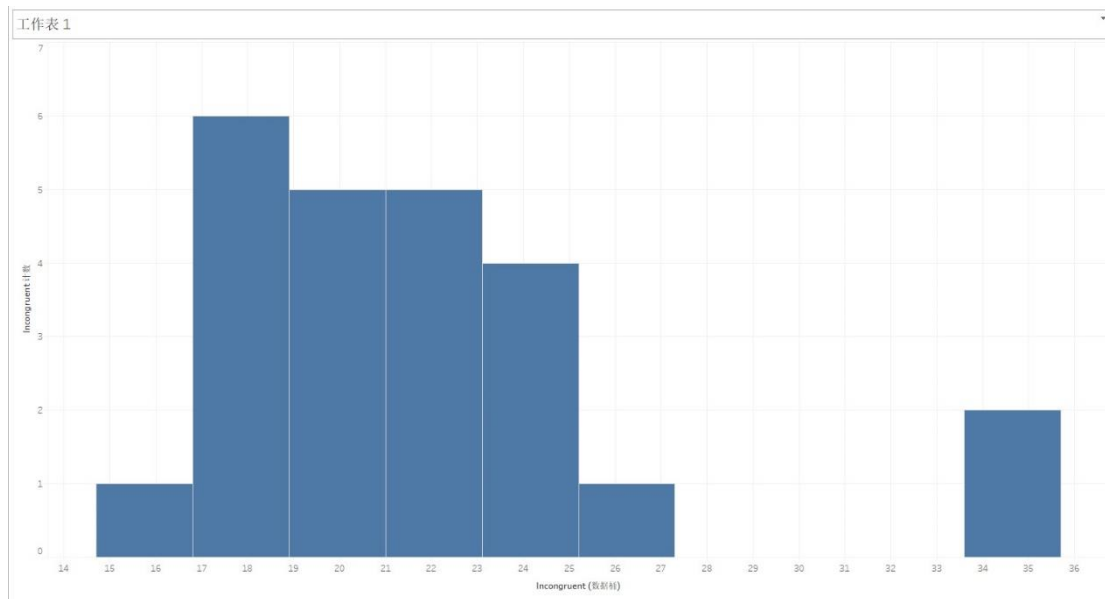


图 1 G_1 不一致文字测试条件下的数据可视化

7、 计算报告

∴ 采用相关样本 t 检验

∴ 根据两组 G_0 和 G_1 的实验数据计算样本差异点估计和样本平均差异

$$\bar{d} = \bar{X}_1 - \bar{X}_0 = 7.9648$$

计算得到样本差异的标准偏差 S_d

$$S_d = 4.8648$$

∴ 根据相关样本 t 检验得出 t 统计量

$$t\text{-statistic} = \frac{(\bar{X}_1 - \bar{X}_0) - (\mu_1 - \mu_0)}{\frac{S_d}{\sqrt{N}}} = 8.0209$$

又因为自由度为 $24 - 1 = 23$ 双尾检测 $\alpha = 0.05$ 的 t-score 为 ± 2.069

即 t-critical value ($\alpha = 0.05$) = 2.069

∴ t-statistic ($\alpha = 0.05$) > t-critical value ($\alpha = 0.05$)

证明 G_0 和 G_1 之间存在显著差异

∴ 拒绝零假设 H_0

∴ 对立假设 $H_1: \mu_1 - \mu_0 \neq 0$ 成立

$$\therefore 95\% \text{ CI} : (\bar{X}_1 - \bar{X}_0) \pm 2.069 * \frac{S_d}{\sqrt{N}}$$

∴ 95%置信区间为 (5.5103, 8.0293)

结论：参与者在 G_0 和 G_1 两组试验中的表现存在显著差异，此结论与提出的期望结果一致

8、 原因分析

我认为这种现象发生的原因是参与者心理因素导致的，类似物理学中的惯性原理，参与者习惯于正向思维逻辑，比如习惯了看到红色字体第一反应就是红色，而在看到红色却命令他说出其他颜色时大脑需要相应的反应时间。

类似任务模拟-可以创建两组测试

G_0 ：向参与者提供标有数字的提示板，数字范围(1-10)，要求参与者用右手做出相应的正确手势；

G_1 ：向参与者提供标有数字的提示板，数字范围(1-10)，要求参与者用右手做出相应的正确手势加 1 的值，比如看到数字 1 后做出 2 的手势；

（甚至可以要求参与者转换左手和右手来进行测试，因为大多数人都是右撇子）