数据分析入门项目一

1. 自变量是什么？因变量是什么？

1.1 自变量是表格类型，文字与其颜色的一致性。
1.2 应变量是参与者分别在不同表格下说出油墨颜色的时间。

2. 此任务的适当假设集是什么？你想执行什么类型的统计测试？为你的选择提供正当理由。

2.1 此任务的适当假设集：

- μ_c 表示文字和颜色一致的条件下读出颜色时间的总体均值；
- μ_i 表示文字和颜色不一致的条件下读出颜色时间的总体均值；
- 零假设：使用文字和颜色一致的列表和文字和颜色不一致的列表，说出对应颜色的时间相同。即：$H_0 : \mu_c = \mu_i$
- 对立假设：使用文字和颜色一致的列表和文字和颜色不一致的列表，说出对应颜色的时间不相同。即：$H_A : \mu_c \neq \mu_i$

2.2 该实验是在同一总体下面采用不同影响条件得到了两组样本值，总体的均值和标准差都是未知的，样本小于 30。所以采用**相依样本配对 t 检验**(dependent t test for paired samples)。对立假设是不同条件下均值不等，所以选择 **α = 0.05 的双尾检验**。

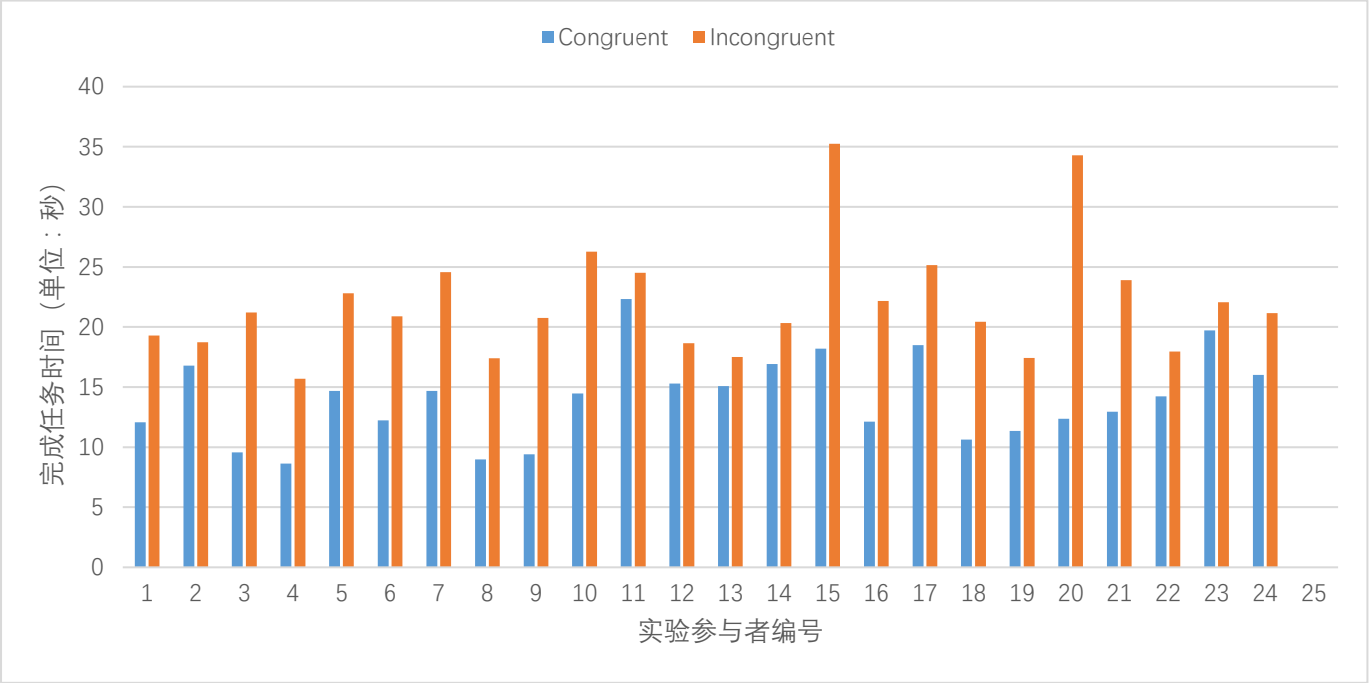
3. 报告关于此数据集的一些描述性统计。包含至少一个集中趋势测量和至少一个变异测量。

3.1 集中趋势
均值：
文字颜色一致条件下均值 $\bar{x}_c$ = 14.05113；
文字颜色不一致条件下均值 $\bar{x}_i$ = 22.01592；
中位数：
文字颜色一致条件下中位数 14.3565；
文字颜色不一致条件下中位数 21.075；

3.2 差异性
样本标准偏差：
文字颜色一致条件下 s_c = 3.559357958
文字颜色不一致条件下 s_i = 4.797057122

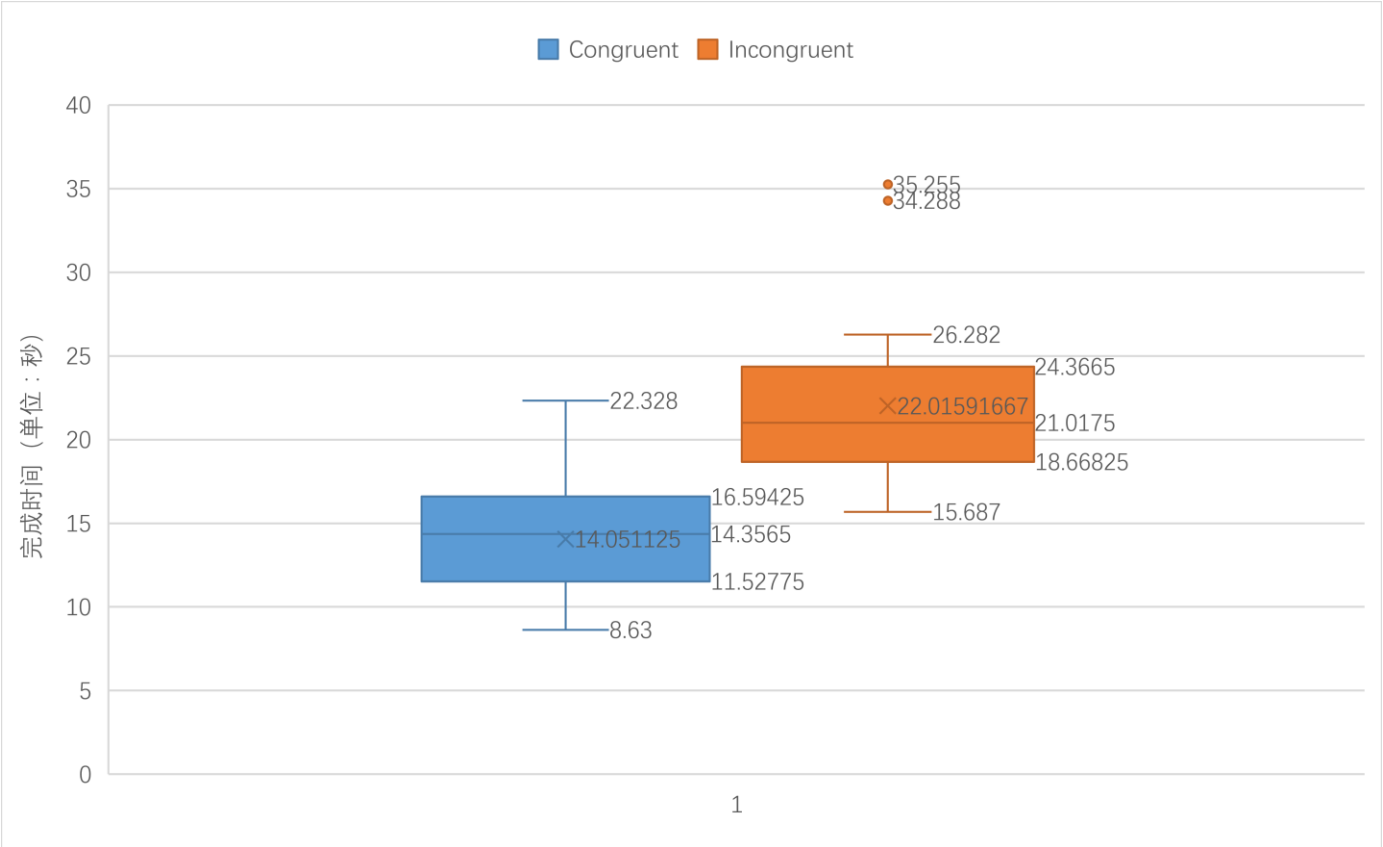
4. 提供显示样本数据分布的一个或两个可视化。用一两句话说明你从图中观察到的结果。

4.1 柱形图：



从上方的柱形图中可以看出每一个实验对象在文字和颜色不一致下的时间都明显高于在文字和颜色一致下的时间。

4.2 箱形图



分别对两个不同条件下的样本数据绘制了箱形图，从图中可以看出基本上中位数平分了 IQR ，可以大概判断样本是正态分布的。其中文字和颜色不一致条件下，出现了两个异常值。

5. 现在，执行统计测试并报告你的结果。你的置信水平和关键统计值是多少？你是否成功拒绝零假设？对试验任务得出一个结论。结果是否与你的期望一致？

这里统一用后一次实验结果减去前一次结果的值来表示前后数据的差异。
已知样本大小为 24，自由度为 23，α 为 0.05，双尾检验
5.1 通过 excel 计算得出：

差异的均值为

$$\bar{d} = 7.964792$$

差异的标准偏差

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})}{n-1}} = 4.864827$$

t 统计量

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = 8.020706$$

查 t 表知道，在自由度为 23，α为 0.05 的双尾检验下，t 的临界值为正负 2.069，显然 t 远大于临界值，所以 t 统计量位于临界区内，P 值远小于 0.025。综上我们拒绝零假设，也就是说当文字与颜色不一致时显著增加了实验参与者完成任务的时间。

参考资料

假设检验相关资料：为什么要使用配对 t 检验？ - Minitab
在线公式编辑器：https://www.codecogs.com/latex/eqneditor.php?lang=zh-cn