HW1 Report: Statistical Analysis

1 Basic assumptions and approaches

a) The assumptions that one-way ANOVA are based on

在使用 one-way ANOVA 进行方差分析时需要基于以下几条假设：

- 数据由随机采样得来，数据间是相互独立的
- 需要假定各组数据的方差是“相同”的，可以解释为以下两种情况：
  - 各组数据分布的方差是相等的，称为同方差
  - 最大标准差与最小标准差的组之间的比值小于 2，标准差为方差的平方根
- 数据的偏差 (residuals) 服从高斯分布

b) One-way ANOVA with non-normal data

对于非正态分布数据上的单变量 ANOVA，可以使用以下几种更具鲁棒性的模型进行方差分析：

- Welch ANOVA
- Brown-Forsythe ANOVA

这些 ANOVA 方差分析的模型在方差相等的假设不成立时，优于传统基于 F 统计量的方差分析。也可以使用非参数检验的 **Kruskal–Wallis test**[1]进行 ANOVA。Kruskal–Wallis test 是一种与单变量 ANOVA 相对应的非参数估计方法，用于检验样本是否来自于同一分布。Kruskal–Wallis 检验基于变量的 rank 进行，首先需要对所有的样本从小到大从 $1-N$ 进行标号，对于值相同的样本，将他们在值不同情况下得到的标号求均值赋给他们，这样操作后所有的样本都会得到一个 rank 值 r_{ij}。根据样本的 rank 构造统计量为：

$$H = (N-1)\frac{\sum_{i=1}^{k} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

当所有样本中没有相等的数据时统计量可以简化为：

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i \bar{r}_i^2 - 3(N+1)$$

[1] https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

其中 N 为样本总数，k 为样本分组数量，$\bar{i}_i$ 为组内秩的平均值，$\bar{r}$ 为全体样本秩的平均值。可以通过 χ^2 分布进行检验，给定显著性水平 α 和自由度 $g-1$ 后，可以通过查表得到检验的临界值 $H_c = \chi^2_{\alpha:g-1}$ ，与统计量 H 进行比较，若 $H \geq H_c$ 则零假设会被拒绝，否则接受零假设。

2 Null(H0) and the alternative(H1) hypotheses

不同分类的群组中平均年龄的差异性可以通过均值进行衡量，按群组类别对数据进行分组，假设年龄可以近似服从同方差的分布，通过比较均值比较不同分类的差异性，可以得到待检验的假设如下：

零假设 (H_0): 所有类别的群组中平均年龄的均值都相等$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

备择假设 (H_1): 至少存在一个类别的群组均值与其他分类中的平均年龄均值不相同

3 ANOVA experiments

在本次作业中使用统计分析工具 R 语言进行，使用`Rscript 2022210870.r`运行实验代码。

a) Empirical probability density funtion and Normality test

选择平均年龄一列的数据作为源数据，调用 R 语言中绘制经验概率密度分布的函数`epdfPlot()`得到平均年龄的经验概率密度分布：

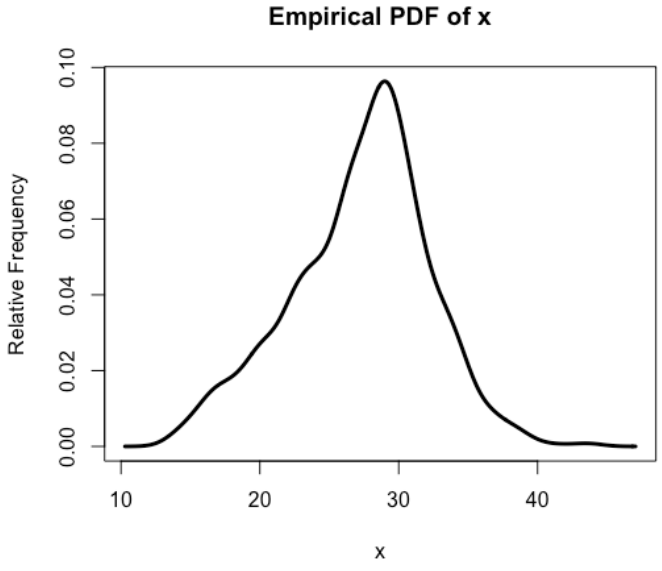


Figure 1: Empirical probability density function

观察 epdf 图像有尖峰和偏度，下面通过正态性检验测试这一列数据是否服从正态分布。使用假设检验的方法，选取显著性水平 $\alpha = 0.05$，建立假设：

$$H_0: \text{总体 X 的分布为正态分布} N(\mu, \sigma^2)$$

$$H_1: \text{总体 X 的分布不服从正态分布} N(\mu, \sigma^2)$$

使用 Lilliefors 检验（Kolmogorov-Smirnor(KS) 检验的修正）[2] 方法对这组数据的正态性进行检验，这种方法基于经验分布函数，用大样本近似，检验标化后的数据是否服从理论分布。

使用 R 中的`lillie.test(x)`函数进行检验，得到检验的结果:

$$\text{检验统计量：} D = 0.055093, \text{p 值：} p = 8.586e-16$$

由于 $p < \alpha$，所以拒绝零假设，即这一列数据不服从正态分布。

b) Empirical probability density funtion and Normality test within category

根据群类别将所有的数据划分成五组，对五组数据使用 Lilliefors 方法进行正态性检验，得到结果如下:

Table 1: Normality test for different categorys

Category	D	p
1 Online Game	0.045094	0.01994
2 Stock Market	0.044512	**0.1574**
3 House & Living	0.049218	**0.2933**
4 School Alumni	0.073593	9.325e-06
5 Organization & Industry	0.081527	8.414e-11

根据检验的结果可以得出，第 2 类和第 3 类这两个类别的群聊中平均年龄近似服从正态分布，而其他类别的群聊平均年龄不具有这一性质。

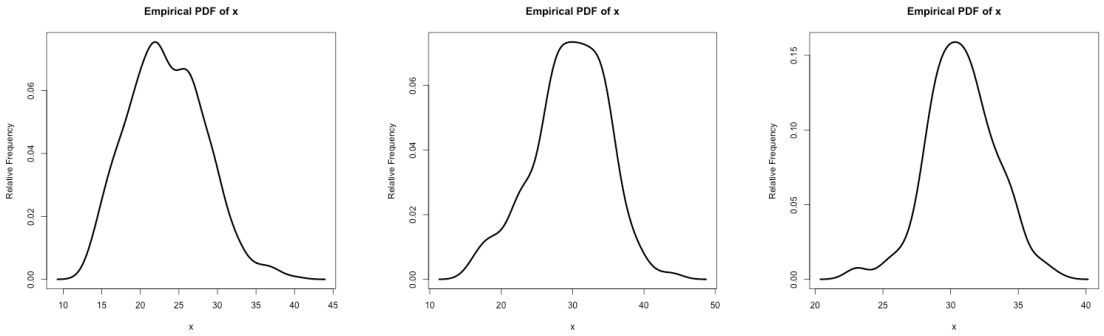
接下来对不同类别群组平均年龄的分布进行方差检验，测试不同类别的分布方差是否相同，检验中的零假设为 H_0：**各组分布的方差相同**，备择假设为 H_1：**至少有一组的方差与其他组不同**。由于这几个类别平均年龄的分布不完全服从正态分布，所以使用鲁棒性更高的方差检验方法 Fligner-Killeen's test。

使用 R 中的`fligner.test()`进行检验，由于共五个类别，所以自由度为 $df = 4$，得到的检验统计量值和 p 值如下:

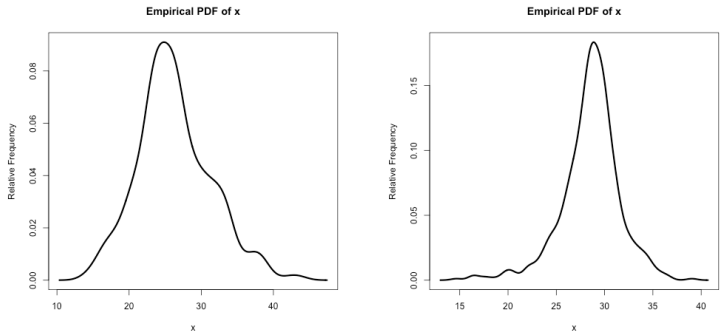
$$\chi^2 = 231.33 \quad p < 2.2e-16$$

由于 $p < \alpha$，所以拒绝零假设，不同群组间的平均年龄分布方差不一致。从各类别群组平均年龄的经验分布函数2也可以直观地看出上述结论。

[2] https://en.wikipedia.org/wiki/Lilliefors_test



(a) Online Game (b) Stock Market (c) House & Living



(d) School Alumni (e) Organization & Industry

Figure 2: Epdf for different categories

c) One-way ANOVA test

首先对不同类别群聊的平均年龄分布进行可视化，绘制得到箱形图3:

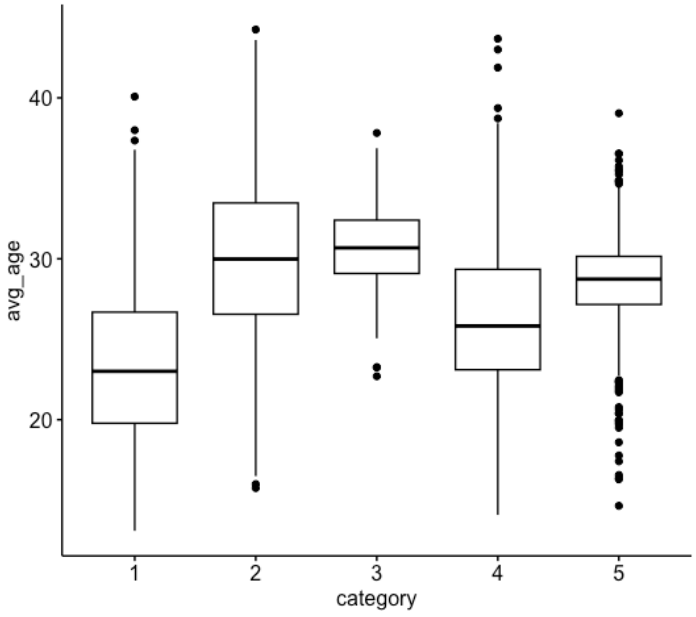


Figure 3: Data visualization

对不同类别的群聊平均年龄的分布进行单变量的 ANOVA 测试，待检验的为问题 2 中做出的假设1，使用 R 中的方差分析函数`aov()`进行检验，得到结果：

Source	SS	df	MS	F	p
Between	12783	4	3196	**171.5**	**<2e-16**
Within	37919	2035	19		

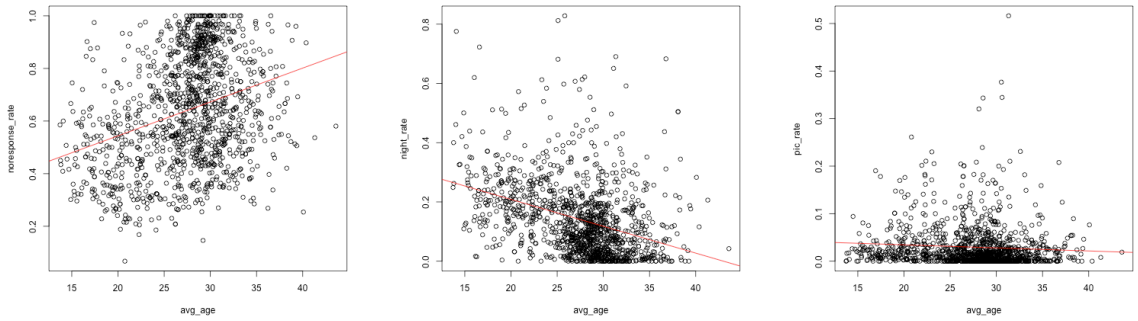
Table 2: ANOVA result table

从单变量 ANOVA 的结果可以得出结论：由于 $p < \alpha$，拒绝零假设，说明不同类别的群聊中平均年龄分布的均值不完全相同。使用非参数的 Kruskal-Wallis 检验可以得到一样的结论，检验统计量 $\chi^2 = 544.1$,$p < 2.2e-16$。不同群聊类型的平均年龄分布均值差异较大，从箱形图中可以直观地看出这一结论。

4 Regression problems

a) Chatting behaviors

筛选聊天会话数大于等于 20 的所有聊天群组，将平均年龄作为自变量 x，y_1: 无回应比例、y_2：夜聊比例，y_3：图片比例作为因变量，根据一元线性回归模型 $y_i = a_i x + b_i$ 建立三个线性模型进行预测，对应 R 中的线性模型`lm()`，回归得到的结果如下，其中红色直线为线性回归得到的模型：



(a) 无回应比例
$y_1 = 0.286293 + 0.012875x$

(b) 夜聊比例
$y_2 = 0.3899289 - 0.0090721x$

(c) 图片比例
$y_3 = 0.0473841 - 0.0006501x$

Figure 4: Result of Linear Regression

使用标准差 $\sqrt{\frac{\sum_k (y_k - \hat{y_k})^2}{df}}$ 和 R^2 对回归的结果进行评估，R^2 越接近 1 证明直线的拟合程度越好，标准差越低说明样本点偏离回归直线的程度越小：

	Std. error	R^2
无回应比例	0.1998	0.09641
夜聊比例	0.1182	0.1316
图片比例	0.04608	0.005089

从两个评估指标可以看出线性回归的效果并不是很好，说明这三组变量的关系不能用线性模型进行很好的描述，从图像中数据点的分布也可以比较明显地看出，所以只能通过拟合结果观测三个待预测的变量与群聊平均年龄间的大致关系。从回归的结果可以总结出平均年龄与上面几个因素的相关性：随着群聊平均年龄的增大，群聊的无回应比例升高，夜聊比例和图片比例下降，与我们的通常认知相符合，年龄越大的聊天者回应频率低，通常熬夜较少所以夜聊频率低，且在聊天过程中更倾向于使用正式的文字交流，导致图片占比较低；更年轻的聊天者聊天的回应频率和深夜聊天的比例都较高，且喜欢在聊天过程中使用图片进行交流。

b) Weighted multivariate linear regression

使用群聊的会话数 n_i 作为样本数据的权重，进行加权回归。权重在回归的误差目标函数中发挥作用，用于最小化 $\sum n_i(y_i - wx_i)^2$，以会话数作为权重可以使得聊天数更多的群聊数据点在回归过程中产生的影响更大，让回归的结果更贴近高频样本点，符合实际。

将`Col[3~10]`的特征组成的向量 $\boldsymbol{x}$ **(群人数，消息数，稠密度，性别比，平均年龄，年龄差，地域集中度，手机比例)** 作为自变量，y_1: 无回应比例、y_2：夜聊比例，y_3：图片比例作为因变量，建立多元线性回归模型：

$$y_i = b_i + \boldsymbol{w_i}^\top \boldsymbol{x}$$

使用 R 中的线性模型进行回归后得到结果如表格3，其中每个参数对应的 p 值越小，说明这个参数对预测的结果更重要；p 值越大，说明这个参数与预测值的相关性可能更小。由于三个待预测指标都为比例值，所以回归结果中的参数数量级都较小。

$$\boldsymbol{w_1} = 10^{-4} \times (-5.384 \times 10^{-1}, -4.681 \times 10^{-2}, -412.2, -213.6, 79.49, -24.92, -424.5, -1859)^\top$$
$$p_1 = (0.00550, 2 \times 10^{-16}, 0.50204, 0.30893, 1.74 \times 10^{-14}, 0.34943, 0.00114, 2 \times 10^{-16})$$
$$\boldsymbol{w_2} = 10^{-4} \times (-4.685 \times 10^{-1}, 3.541 \times 10^{-3}, 453.3, 392.1, -83.81, 105.1, -425.6, 1301)^\top$$
$$p_2 = (7.59 \times 10^{-6}, 0.000503, 0.170733, 0.000536, 2 \times 10^{-16}, 3.45 \times 10^{-13}, 1.59 \times 10^{-9}, 2 \times 10^{-16})$$
$$\boldsymbol{w_3} = 10^{-4} \times (-2.706 \times 10^{-1}, 6.035 \times 10^{-3}, 2.527, 236.3, -7.818, 4.882, -117.8, -85.21)^\top$$
$$p_3 = (1.03 \times 10^{-6}, 2 \times 10^{-16}, 0.98848, 8.02 \times 10^{-5}, 0.00773, 0.52013, 0.00153, 0.09080)$$

	w	b	p-value	Std. error	R^2
无回应比例	w_1	0.48	p_1	1.683	0.3859
夜聊比例	w_2	0.2805	p_2	0.9069	0.2979
图片比例	w_3	4.385×10^{-2}	p_3	0.4796	0.1019

Table 3: Weighted multivariate linear regression result

从回归的结果可以观察出，与单一自变量的线性回归相比，加权多元线性回归的 R^2 值更优，得到的模型能够更好地接近原始数据的相关性和变化趋势，其中对于每一个自变量，若 $\boldsymbol{w_i}$ 中对应的参数分量为正，则说明这个自变量与待预测的标签值正相关，反之则为负相关。

c) Logistical Regression

选择群组分类为 1 和 4 的数据作为数据集，其他维度的数据作为特征进行逻辑回归。以 0.8 为比例对数据集进行划分，80% 作为训练集，其余 20% 作为测试集。使用 R 中的包`tidymodels`实现逻辑回归[3]，

记群聊类别为 1 的概率为 p，群聊类别为 4 的概率为 $1-p$，则类别为 1 的几率 (odds) 为 $odds = \frac{p}{1-p}$。对 odds 取对数作为待预测的因变量，其他群聊相关特征组成的特征向量 $\boldsymbol{x}$ 作为自变量，建立逻辑回归的模型：

$$logit(p) = w_0 + \boldsymbol{w_1}^\top \boldsymbol{x}$$

R 中的一般化线性模型`glm()`提供了一般化的线性模型，选择分类模式 ("classification") 即为逻辑回归。对 1 和 4 两种类别的群聊进行逻辑回归后得到的模型参数如下，第一列为每个特征对应的系数，第二列为该特征对应的 p 值：

	Coefficients	p-value
w_0	-5.983910	8.054363×10^{-11}
群人数	2.825725×10^{-3}	4.409503×10^{-2}
消息数	-1.343529×10^{-5}	3.106366×10^{-1}
稠密度	7.787515	2.431379×10^{-10}
性别比	-1.546691	7.735975×10^{-3}
平均年龄	2.231625×10^{-1}	6.301176×10^{-17}
年龄差	-2.709520×10^{-1}	1.266673×10^{-6}
地域集中度	2.058343	7.040901×10^{-7}
手机比例	1.241220	2.193484×10^{-3}
会话数	-2.855675×10^{-3}	1.770263×10^{-2}
无回应比例	-3.426696×10^{-1}	4.979310×10^{-1}
夜聊比例	7.823027×10^{-1}	1.315708×10^{-1}
图片比例	-2.788621×10^{-1}	8.281111×10^{-1}

Table 4: Logistical Regression result

在测试集上对逻辑回归的模型进行预测，将预测得到的标签 $\hat{y_i}$ 与真实群类别 y_i 进行比较，得到预测的准确率为 88.5%。

5 ANOVA test for sampling data

选择简单随机采样和分层随机采样两种采样方式进行对比。

a) Simple Random Sampling

从原始数据中使用简单随机采样的方式采样得到 200 组数据，原始数据中每个样本被采样的概率相等，第 i 次采样得到的数据集记为 d_i，重复采样十次，对每次采样得到的子数据集的群聊

[3] https://www.datacamp.com/tutorial/logistic-regression-R

类别～平均年龄进行如 3 c）中的方差分析，对 10 次方差分析中的 F_i 统计量计算均值和标准差：

$$\bar{F} = \frac{\sum F_i}{10}$$
$$stddev = \frac{\sum (F_i - \bar{F})^2}{9}$$

R 中的简单随机采样通过`sample()`实现，进行十次采样得到的 F 统计量的均值和标准差为：

F-value	26.14731	11.80287	15.68029	10.72240	21.02451
	26.06658	16.05215	19.93289	18.60128	15.23844
Mean	18.12687		**Std dev**	5.298867	

Table 5: ANOVA for simple random sampling

b) Stratified Random Sampling

选择群聊的类别作为层次 (strata) 对数据进行分层，在每种类别的群聊中进行简单随机采样，根据不同群聊类别占总样本数的比例，五种群类别中得到的样本数量为 $48, 30, 19, 42, 63$，共 202 个样本。连续进行 10 次采样得到的 F 统计量的结果为：

F-value	20.91813	21.92411	10.72849	13.72324	17.00150
	16.77531	18.92694	16.66461	15.16108	14.88814
Mean	16.67115		**Std dev**	3.338393	

Table 6: ANOVA for stratified random sampling

对比简单随机采样和分层随机采样两种方法 ANOVA 测试的结果，可以观察出分层随机采样方法下的 F 检验统计量的标准差更小，说明在这种采样方式下每次得到的不同类别群聊的数据分布是相似的，更接近原始数据中的分布，所以 ANOVA 中 F 统计量的波动小；而简单随机采样因为不考虑采样结果中各类别的比例，可能导致采样得到的结果偏向某单一聊天类别，所以每次采样得到的数据较原始的真实数据分布差异较大，导致每次采样的 F 统计量波动较大。可以得出结论，**分层随机采样比简单随机采样更稳定**。

与在不经过采样的原始数据集上进行 ANOVA 的结果 $F = 171.5$ 进行对比，采样后的方差分析得到的 F 统计量值更小，因为样本数为 200，进行采样后可能导致每个类别中群聊的样本数较少，具有**样本选择性偏差**，组内数据分布的方差变小，不能很好地反映出原始不同类别群聊平均年龄的分布，体现在 ANOVA 中 SS_w 这一项相比不进行采样的情况下更小。导致对采样后各组的小样本数进行方差分析得到更小的 F 值，即各组间的均值相似，随着样本量的增大，不同类别的样本更接近于原始数据的真实分布，各组间分布的均值差异变得明显，所以 F 统计量的值增大，以分层随机采样为例，变化采样的总数目，进行 ANOVA 得到的 F 统计量值随总数目的增大而增大，可以支持这个结论：

Samples num	202	407	678	1019	2040
F-value	16.67115	35.12119	59.30885	86.47738	171.507

Table 7: F-value for different sampling numbers