LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: Dr **James Connolly**

Assessment Title: ____________________

Work to be submitted to: Blackboard

Date for submission of work: ____________________

Place and time for submitting work: ____________________

To be completed by the Student

Student's Name: **Danfeng Wang**

Class: ____________________

Subject/Module: Data Science

Word Count (where applicable): ____________________

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: ____________________ Date: __________

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

Abstract

Smoking contributes to man diseases, we hypothesized that ciggrette smoking could result in an observable effect on stroke age. There are ___ patients in 5 years, ___ of them have had stroke. ___ stroke patients has mean age of __ who had never smoked, while the average age of smoking subgroup is __ younger than it. The difference in mean age for smoker, or non-smokers is statistically significantly (p-value <...). There is no statistically significant association between obesity and stroke at patients who has had stroke subgroup. Smoking was associated with a younger age at .

xxxx is performed because of the distribution of xx is not normally distributed.

We calculated the stroke patients by gender, and compare the outcome by using a t test for continuous variable and chi-square test for categorical variables.

All hypothesis tests in this reports are two-sides test by default, and the significant level is 0.05.

The contribution to the stroke of bmi, hypertension, gender is described in this report. These are all risk factors to stroke. Our goal is to find the patterns of health habit and the risk of getting stroke. We test the correlations between age, obesity, smoking and stroke separately with chi-square test of independence. The dependent variable (DV) is stroke which is categorical variable.

Data is collected from from 2015 to 2021.

This data set consists of xx rows and xx columns. For patients records from 2015 to 2021, the following variables are reported: gender (Male, Female), …..It contains categorical variables and continuous variables.

We evaluated the 5000 patients, 50% is woman, 50% is men. Men are younger than woman (数字范围均值+最大值) , 3% of the men has smoke history, 50% of the woman has smoke history. Men had more stroke patients than woman (p-value = 0.4,)

Overall, there is a significant age difference between stroke and non-stroke patient groups. The mean age of stroke group patients is xx which is statistically higher than that of the non-stroke group. So that age is a risk factor contributes to stroke.

Research questions

1. Is there relationship between age and stroke?
2. Is the age in stroke patient group and non-stroke patient group significantly different? Or does the stroke group has higher mean age than that of the non-stroke group?
3. Does smoking have relationship with stroke?
4. Is smoking history has relationship with age of stroke patients? The average age reduced as the increased smoking level?

5. Is there relationship between average blood glucose level and BMI?

Data preparation

Dataset description

The dataset being used in this report is stroke patients' information from 2015 to 2021 (See below table). From this table, we can see the data describes the patient information from aspects of basic information, health status, life styles. The dataset is stored as csv format, before using it, we are going to clean and transform it to make it prepared for the next analysis.

Table 1 The raw dataset and description

No.	Name	Description
1	Id	Unique id
2	Gender	Male / Female / Other
3	Age	Age of the patient
4	Hypertension	Has hypertension or not (1/0)
5	Heart disease	Has heart disease or not (1/0)
6	Ever married	Yes / No
7	Work type	Children / Gov_jov / never_worked / Private / Sel-smployed
8	Residence type	Rural / Urban
9	Avg glucose level	Average blood glucose level
10	BMI	Body mass index
11	Smoking status	Formerly smoked / never smoked / smokes / Unknown
12	Stroke	The patient has had a stroke or not (1/0)
13	Date	Record date

Data clean and transformation

The purpose of this stage is to make sure the dataset for analysis is tidy and structured, and prepared for investigate our research questions. The relevant steps will be involved to prepare the dataset for this study.

Step1 – load the dataset from csv, and store each variable into suitable types. Below table is the type conversion plan. The hypertension, heart disease variables' values are 0 or 1. 0 means patient has no such disease, and 1 has the opposite meaning. In order to know clearly what the data represents, we convert them to *No* and *Yes* accordingly when converting them to factor type.

Table 2 Type conversion

No.	Name	Target type and format
1	Gender	Factor
2	Hypertension	Factor with Yes/No labels
3	Heart_disease	Factor with Yes/No labels

4	Ever_married	Factor
5	Work_type	Factor
6	Residence_type	Factor
7	Smoking_status	Factor
8	Stroke	Factor with Yes/No labels
9	Date	Date with format yyyy-mm-dd

The other four columns, which has not been listed in the above table includes id, age, avg_glucose_level, bmi, have been loaded as numeric.

After conversion, we check the structure and sample data to make sure all of the variables have been loaded into the corresponding types and the total amount of the data is 5110 rows with 13 columns.

Step2 – process missing value. Before dealing with the missing data, we calculate and plot the missing values of the dataset first. It can be seen that only bmi variable has a small proportion of missing values (201 rows). We drop these part of data directly.

Step3 – process outliers. Looking from the summary of the dataset, there are rows with *Other* gender, and *unknown* smoking status. After filtering out these outliers by subset function, re-check the data summary to make sure they have been remove properly. Moreover, some factor values are no longer existed because of the filter operation, we refresh the factor level by reconverting them to char and to factor type again.

Step4 – Based on research questions, smoking_level, diabetes variables are generated to support the next analysis. Final variables are listed in below table.

Table 3 final variables of the dataset

No.	**Name**	**Target type and format**
1	Age	Numeric
2	Bmi	Numeric
3	Obesity	Factor (Yes/No)
4	Smoking_level	Ordinal Factor (never smoked/ formerly smoked / smoke)
5	Stroke	Factor (Yes/No)

Descriptive statistics

After the dataset has been prepared, descriptive analysis will be applied to give an initial view about the variables correlations.

We plot the distribution and the correlation between the variables.

Assumption of age dependent variable

We assume age independent variable is normally distributed, and the age of all patients on both stroke group and non-stroke group is normally distributed. In addition, we assume stroke patients are older than that of the non-stroke patients.

Table 4 age distribution by stroke

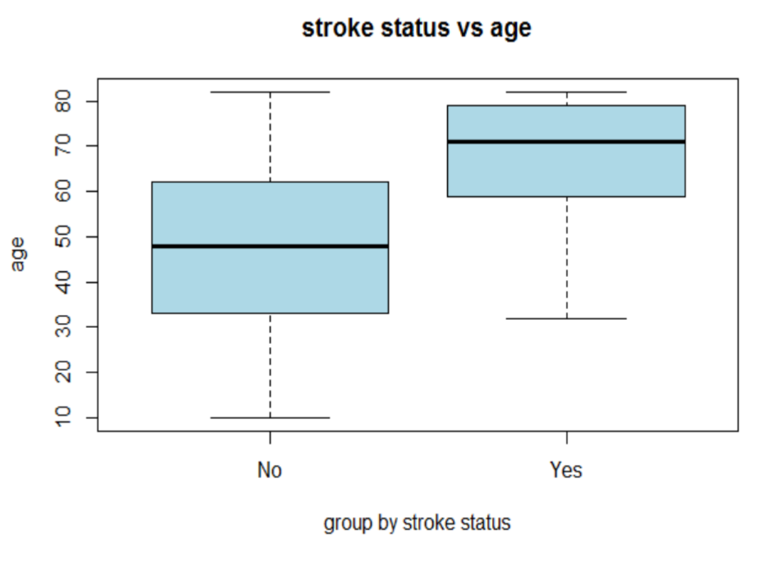
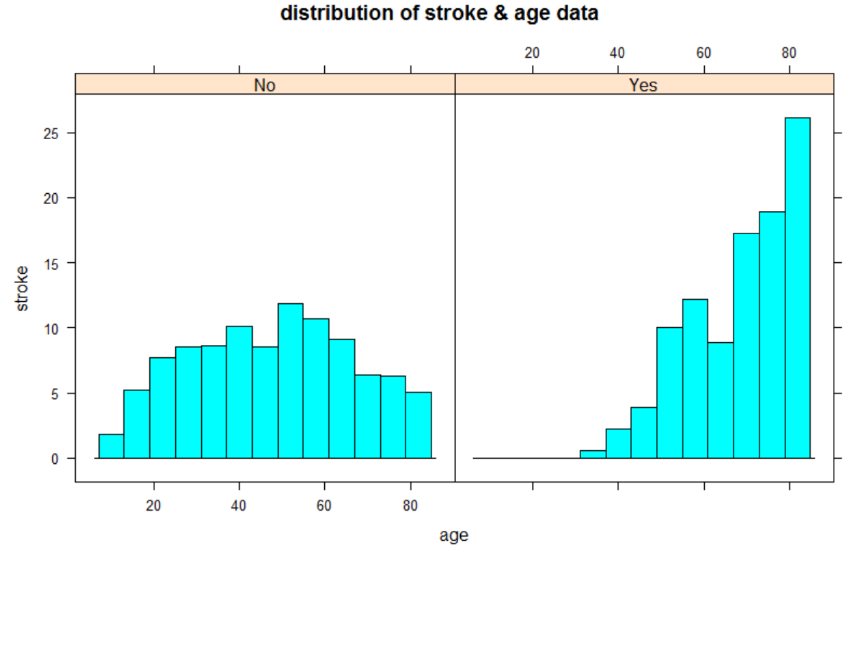
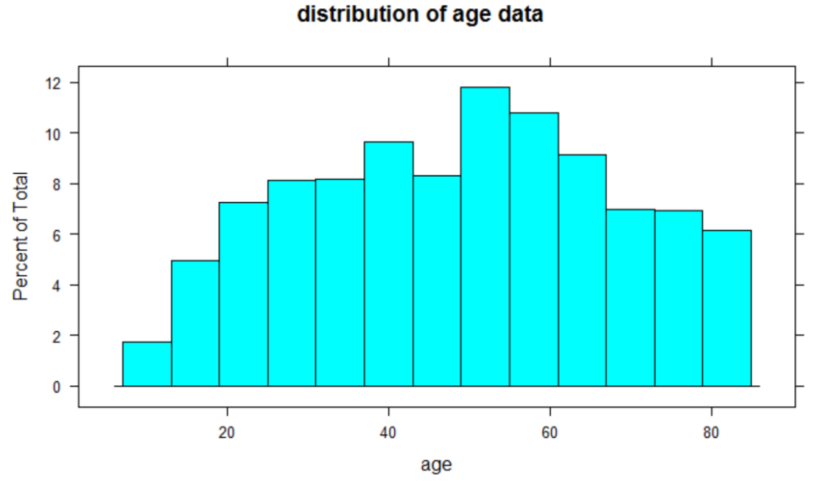
	Check the distribution of patients' age. The median age of stoke group are higher than that of the non-stroke group. We calculate the accurate mean value of both group, the result shows that non-stroke patients is approximately at 47.57 of mean age, while the mean age of stroke group is 68.05. It looks a bell-shaped but not really symmetrical. Both sides are cut off at specific values. It seems to be normally distributed.
Age – continuous	
Stroke – categorical dichotomous	

Table 5 age normality check by stroke

	Check the distribution of patients' age in stroke and non-stroke group. Because our investigation revolves around stroke, so we also split the population by stroke status to check normally distribution for further analysis. The non-stroke group looks a bell-shaped but not really symmetrical, it seems to be normally distributed. The stroke patients group is unlikely to be normally distributed.
Age – continuous	
Stroke – categorical dichotomous	



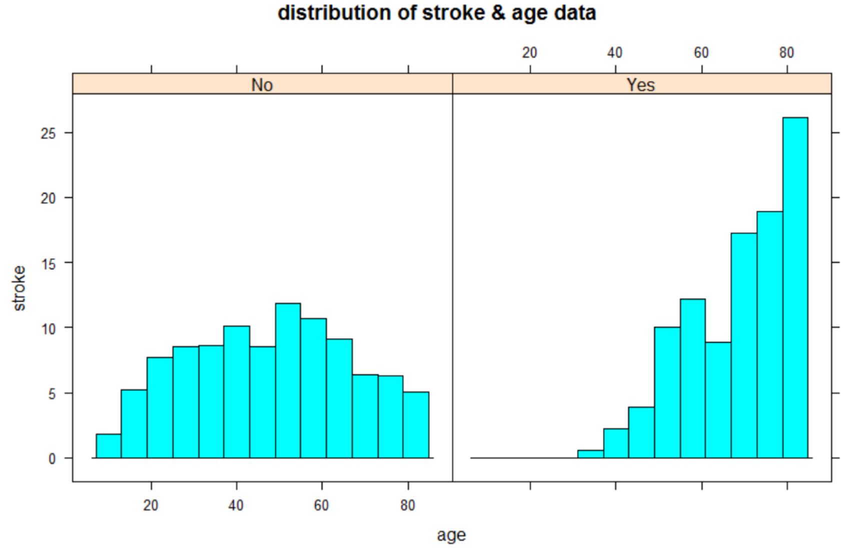
Check the distribution of patients' age.

It looks a bell-shaped but not really symmetrical. Both sides are cut off at specific values.

It seems to be normally distributed.

Age – continuous

Stroke – categorical dichotomous



Check the distribution of patients' age in stroke and non-stroke group.

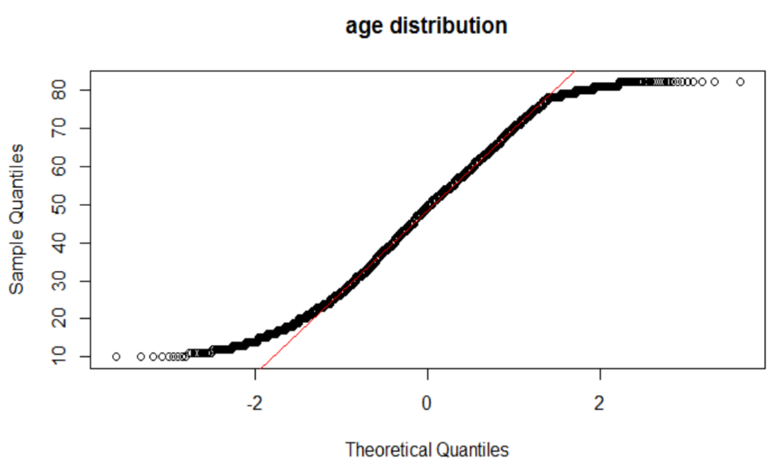
Because our investigation revolves around stroke, so we also split the population by stroke status to check normally distribution for further analysis.

The non-stroke group looks a bell-shaped but not really symmetrical, it seems to be normally distributed.

The stroke patients group is unlikely to be normally distributed.

Age – continuous

Stroke – categorical dichotomous

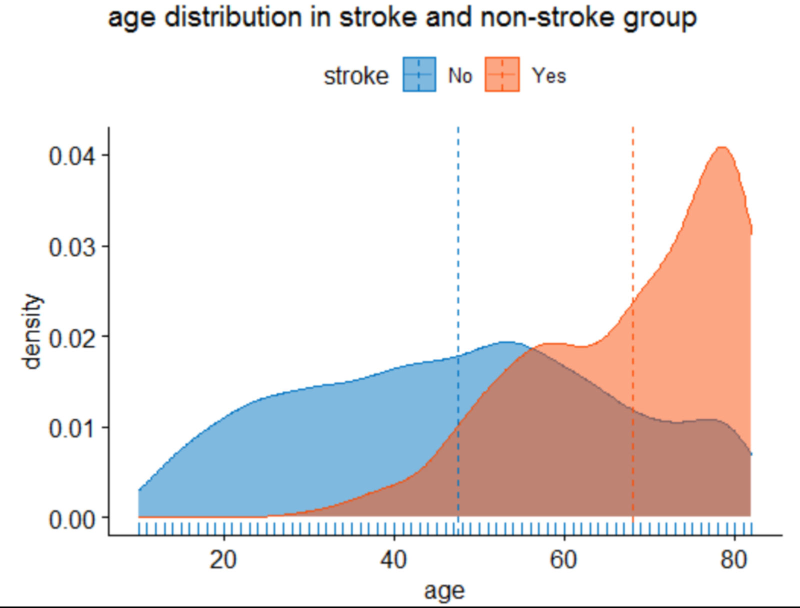


Check the distribution of patients' age with Q-Q norm with Q-Q line.

Though the age from 20 ~ 75 data points distributed along the normal line, the other data points didn't fall on the normal straight line. The data points of age under 20 skewed left and the age above 75 skewed right

It's unlikely to be normally distributed, but still we need air-tight proof.

Age – continuous
Stroke – categorical dichotomous

	Visualize the underlying distribution of patient age by density group. The marks along the x axis indicates the data location. This graph gives us an evidence of the median value of non-stroke group is obviously different with that of the stroke patient group.
Age – continuous	
Stroke – categorical dichotomous	

We did Shapiro-Wilk test in order to get a precise answer of whether the age is normally distributed in the entire population. The null hypothesis of Shapiro-Wilk test is that *a variable is normally distributed in some population*. If p-value < 0.05, the null hypothesis is rejected.

According to the test result, the p-value of non-stroke group, stroke group and entire population is 6.236922e-23, 4.133363e-09, 8.500619e-25 respectively, which are all far smaller than 0.05. In this case, we conclude that age is not normally distributed in our dataset.

Assumption of average glucose variable

We assume that *average glucose level* variable is normally distributed. and the *average glucose level* of all patients on both stroke group and non-stroke group is normally distributed.

We use boxplots of ggplot2 to see the outliers for stroke patient group and non-stroke-patient group. The middle line of the boxplots indicates both stroke and non-stroke patients group have similar median values (90-110). Non-stroke patients group has many outliers, while the stroke patients group has no outliers.

Side-by-side boxplots are provided by ggplot2. The boxplots below seem to indicate one outlier for treatment group C and D. Furthermore, both the mean (circle with +)

and median (middle line) values are at the 75th percentile. This indicates that the data is highly skewed by the effects of the outlier(s).

用 descriptive statistics techniques 分析。描述过程，用数据和图表 discuss。直到 dataset prepared

Hypothesis testing and statistical methods

Question 1

Question - Is there relationship between age and stroke?		
Age	Independent continuous variable	Numeric
Stroke	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
H0	age has no correlation with stroke	
H1	age has a correlation with stroke	

Many diseases have significant different in age, gender, and other group identity. We use this hypothesis to test whether there is correlation between patients age and stroke. The dependent variable of this test is categorical dichotomous which should apply Chi-squared test regardless the independent variable is normally distributed or not.

Question 2

Question - Is the age in stroke patient group and non-stroke patient group significantly different? Or does the stroke group has higher mean age than that of the non-stroke group?		
Age	Independent continuous variable	Numeric
Stroke	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
H0	patient age in stroke and non-stroke patients group are similar	
H1	patient age in stroke and non-stroke patients group are different	

First we visualize the data by boxplot, density plot to see if the age distribution in both group, and whether it’s normally distributed. Then we prove the normality by Shapiro-Wilk test. According to the test result, age is not normally distributed variable in both group. So, we use *Wilcoxon* test to confirm whether the age distribution in both groups are significantly different.

Furthermore, according to the boxplot, it seems the mean age of patients in stroke group is higher than that of the non-stroke group. A summary by tapply function will be used to see the mean age in both groups. Then we use Wilcoxon with different *alternative* parameter to do one-tailed test to check if it's true.

From the boxplot and density plot of the above analysis, we see the distribution of patients age in stroke and non-stroke group look different, and the stroke group has higher mean value than that of the non-stroke patient group. We make this hypothesis to test whether there is a difference influence between population's age and stroke

The histograms show that both distribution do not seem to follow a normal distribution, and the p-values of the Shapiro-Wilk test confirm it (reject the null hypothesis of normality for both distributions at the 5% significance level)

Both stroke patients and non-stroke patients are independent, which are the patients who has stroke or not respectively, and do not affect each other.

As the age of each group is not normally distributed, we use non-parametric test - *Wilcoxon* Man-Whitney test- to test whether the age distribution in both groups are significantly different.

Question 3

Question - Does smoking have relationship with stroke?		
Stroke	Dependent categorical variable	Yes/No
Smoking level	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
H0	smoking has no relationship with stroke	
H1	smoking has relationship with stroke	

Question 4

Question - Is smoking history has relationship with age of stroke patients? The average age reduced as the increased smoking level?		
Age	Independent continuous variable	Numeric
Smoking level	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
H0	smoking level has correlation with age in stroke patient group	
H1	smoking level has no correlation with age in stroke patient group	

From the boxplot of stroke patient group, we can see the age distribution by smoking level (never smoked / formerly smoked / smokes). Though there are outliers in *never smoked* and *formerly smoked* group, the mean age seems to have negative correlation with the smoking level, which means the fewer smoking history the older of the mean age. With this hypothesis test, we can know if the smoking history

Question 3

Question - Is there relationship between average blood glucose level and BMI?		
Age	Independent continuous variable	Numeric
Smoking level	Independent ordinal variable	Has been converted to ordinal by the smoking degree, which is *never smoker, formerly smoked, smoke*
H0	average glucose level has no correlation with BMI	
H1	average glucose level has a correlation with BMI	

Average glucose level is the key indicator of diabetes. Diabetes used to be believed a risk factor of other diseases such as heart diseases[1]. We use the above hypothesis to prove this relationship.

Describe the questions in hypothesis test.

Describe the relationship between variables.

Explain how each of the hypothesis test will be valuable when answering each question. 参照问题定义

Statistical methods

Using the hypothesis test to describe examine each test.

Describe in detail each statistical test by examining the following characteristics

- The structure of the data variables I will examine
- How selected variables enable me to answer the question
- Any assumptions you are making about your data variables such as **normality**

There exists gender bias with some diseases. Before making hypothesis test of the correlation of gender and stroke status, we see the data summation first from these two dimensions. Below table shows the number of patients by gender by stroke. Percentage of female patients had stroke in this dataset, is approximately 4.14% (120/2897), and the according percentage of man is about 4.25% (209/4908).

Number of patients by gender by stroke			
Gender	**Non-stroke**	**Stroke**	**Total**
Female	2777	120	2897
Male	1922	89	2011

[1] Whiteley, L., Padmanabhan, S., Hole, D. and Isles, C., 2005. Should diabetes be considered a coronary heart disease risk equivalent?: results from 25 years of follow-up in the Renfrew and Paisley survey. *Diabetes Care*, *28*(7), pp.1588-1593.

Total	4699	209	4908

Judging from the percentage difference between male who got stroke and that of the female group, it appears that there is a slight difference between gender and stroke. However, in order to prove it, we need to do hypothesis test.

We are going to test whether there is correlation between gender and the stroke based on this dataset. Gender is a typical categorical variable, the value of gender consists of Male and Female. Because we found there are outliers in the gender variable in data preparation stage, we have removed the row with gender values are neither Male nor Female, and then convert gender from char to factor. Stroke in this dataset means whether the patient ever got stroke or not, we convert the variable into factor and label the original 0 and 1 value with No and Ye respectively.

Both gender and stroke are categorical variables. Gender is independent and stroke we take it as dependent variable. When both the dependent variable and independent variable are categorical, we use Chi-square test to do hypothesis testing.

The null hypothesis in chi-square is that there is no relationship between the independent variable and the dependent variable. So our test is defined as the above Table shows.

We choose Alpha = 0.05 in this test which means when these two variables have a 0.05 or less probability, there is no relationship between the two variables.

p-value of this test is 0.6805 which indicates weak evidence against the null hypothesis, so that there is no correlation between gender and stroke based on this patients dataset.

Test 2

Number of patients by smoking status by stroke			
Smoking	**Non-stroke**	**Stroke**	**Total**
No	1768	84	1852
Yes	1477	96	1573
Total	3245	180	3425

Test 3

The distribution of stroke is not normally distributed. The dependent variable is not normally distributed in this dataset, so that we use Kruskai-Wallis test of Non-parametric test.

These data samples are independent, and the samples do not affect each other. Using this test, we can decide if the population distribution is identical without assuming them to follow the normal distribution.

The p-value of the test is 2.2e-16 which is significantly less than the significant level 0.05, we can conclude that there are significant age differences between the patients of stroke and non-stroke groups.

Question 4

In stroke patient group, is *Smoker* patients has lower average age than that of *Male* patients? Assume the sample data (Male group & Female group) is independent.

Assumption - If the sample data is not normally distributed. If the mean age of Male and Female group are different.

We plot the stroke patient age by gender, and it seems there is a significant of mean value in both group. After summarize the data by gender,

Result

用 R 对 variables 应用统计学方法回答问题。解释每个测试的输出。解释如何决定每个假设的

From the summary of the prepared dataset, the patient's information gathered from 1st Jan 2015 to 1st Jan 2021 shows,

XX percent of the sample are woman, the mean age of the sample is xxx years, and woman is sigficantly older than men (55 years' vs 50 years). The two groups did not differ in health.

Recognition of stroke

Blood pressure, smoking, and obesity are risk factors of stroke. BMI in a good level (according to WHO) has lower risk in getting stroke.

The definition of obesity (BMI>30 kg/m2) [2]is a risk factor of diabetes. It has been wildly used in white population. In this dataset, the personal information does not include race, we cannot have a accurate level to define the obesity status which might lead bias of our investigation.

The data summary is shown in Table n. ___ patient has had stroke, ___ patients have not. The average age of stroke patients is ___, the average of non-stroke patients is ___. The stroke is associated with age (p-value < ___). The older patients are more like to have had stroke (p-value < ___) which is significantly different with the patients who has never had stroke.

Smoking is associated with younger age in patient subgroup. In these stroke patients, ____ has never smoked, ___ has smoked formerly, and ___ is smoker. The smoking patients shows youngest mean age of __, the formerly smoked stroke patients have ___ of mean age, and the mean age of patients who never smoke is the oldest, which is about __.

The other differences in risk profiles in these two subgroups were small and not statistically significant.

[2] Chiu, M., Austin, P.C., Manuel, D.G., Shah, B.R. and Tu, J.V., 2011. Deriving ethnic-specific BMI cutoff points for assessing diabetes risk. *Diabetes care*, *34*(8), pp.1741-1748.

Men and women with diabetes only and with CHD only formed an intermediate risk group. Men with diabetes only had **marginally higher** mortality than men with CHD only (54.0 vs. 50.5 deaths per 1,000 person-years), whereas women with diabetes only appeared to have a considerably higher risk of vascular death than women with CHD only (46.7 vs. 29.2 deaths per 1,000 person-years) (Fig. 1). Similar trends were observed for each group of causes of death. Specifically, men with diabetes only had marginally higher CHD and other vascular mortality than men with CHD only, whereas women with diabetes only had higher CHD and other vascular mortality than women with CHD only (Fig. 1).

Overall survival over the course of 25 years is shown in Fig. 2. This confirms the similarity in outcome between men with diabetes only and men with CHD only (log-rank $\chi^2 = 0.19$, $P = 0.664$) as well as the difference in outcome between women with diabetes only and women with CHD only (log-rank $\chi^2 = 8.54$, $P = 0.004$). Survival was least in men and women with both diabetes and CHD and greatest in those with neither (Fig. 2).

The similarity in men and the difference in women persisted after adjustments for age, smoking, hypertension, serum cholesterol, BMI, and social class (Table 2). Adjusted hazard ratios (HRs) for CHD and all-cause mortality in men with diabetes only compared with men with CHD only were 1.17 (95% CI 0.78–1.74; $P = 0.450$) and 1.20 (0.92–1.56; $P = 0.172$), respectively. Corresponding HRs for women were 1.97 (1.27–3.08; $P = 0.003$) for CHD mortality and 1.80 (1.37–2.35; $P < 0.001$) for all-cause mortality. Table 2 also shows HRs for the other covariates in the Cox model. Increasing age, cigarette smoking, hypertension, and hyperlipidemia were all associated with CHD mortality. There were trends toward increased CHD mortality with increasing BMI, but these did not achieve statistical significance. Low social class predicted CHD mortality in women but not men.

Increasing age, smoking, and obesity are each associated with stroke.
, whereas BMI ≥35.0 kg/m^2 did (Table 2).

Conclusion

The p-value of the chi-squared tests is less than 0.05, we infer dependence. If it not less than 0.05, we failed to prove the dependence. The p-value of test 1 is roughly 0.00005, which is quite tiny. In other words, we have evidence that xxxxx and stroke are not independent. The proportion of stroke patients in this dataset who is obesity is roughly half that of the non-stroke patient group.

The main finding of our study is that, age, smoking, obesity is the risk factor of stroke.

However, one should note that in the dataset of our investigation, there are only xxx stroke patients in the entire population. and were therefore unable to examine outcome in men and women separately.

The limitation of the smoking factor is that the smoking status is only described with never smoke, formerly smoked, smoke, and we don’t know the specific smoking years. That means, our hypothesis test is based on the cognition of formerly smoked group has less degree than the smoker, and greater degree than non-smoker.

Although all recognize that obesity is associated with higher stroke probability, the belief that obesity is a risk factor of stroke is not supported as strongly as we expected.
The study focus on age, bmi and the relationship with stroke. Due to space limitation, we have ignored some other important factors of stroke, such as hypothesis, heart disease, and etc.
Studies such as ours have strengths and limitations. We do not know if there were differences between respondents and nonrespondents because we did not have permission to track the nonrespondents, although we believe that a 79% response rate means that subjects in the Renfrew and Paisley Survey were likely to have been representative of the general population from which they were drawn. The inclusion of both sexes, the long duration of follow-up, and the large number of deaths are also strengths, as is the adjustment for the effects on outcome of six possible confounding variables including age, smoking habit, blood pressure, cholesterol, BMI, and social class.

讨论每个问题的分析，统计结果表明。对结果明细进行批判。提供深度讨论这个结果表明。