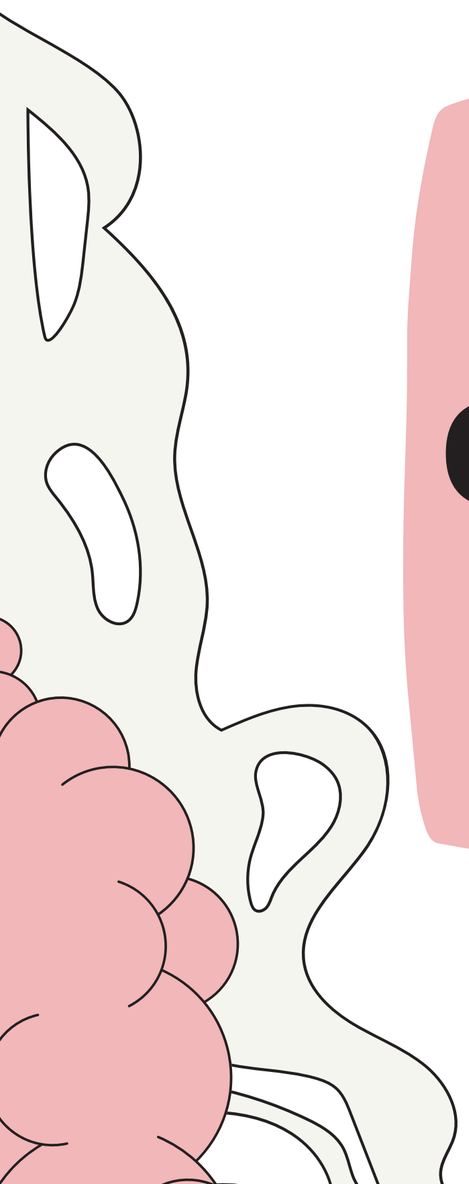
Project 2- Personal Key Indicators of Heart Disease

ANG YI QIN
FONG KHAH KHEH
SOO WAN YING
KEE SHIN PEARL



OBJECTIVE

- To prove whether the sleep time of females is more than the sleep time of males.
- To estimate the relationship between sleep time and BMI.
- To determine the risk of heart disease depending on the BMI values.
- To prove that there is a difference between the observed frequency and expected frequency of the people who have smoked.
- To test whether gender and heart disease has a relationship.

2 Sample Hypothesis Testing (Test on Mean, Variance Unknown)

To test whether the mean sleeping time of woman is larger than the mean sleeping time of man at 95% confidence level, assuming unequal variances.

$n_1 = 138$	$n_2 = 61$
$\bar{x}_1 = 7.297101$	$\bar{x}_2 = 7.327869$
$s_1 = 1.511076$	$s_2 = 1.795562$

calculated frequency (n), mean ($\bar{x}$) & standard deviation (s) for both groups

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

1. Hypothesis Statement

μ_1 = mean of sleep time of female, μ_2 = mean of sleep time of male

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$$

2. Test Statistic, t_0

test statistics, t_0 = -0.11679

group 1 = sleeptime of female
group 2 = sleeptime of male

$$v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

3. Degree of Freedom, v

degree of freedom, v = 99.19

4. Critical Value

Critical value, t =-1.66, p-value=0.9073

Conclusion

Using 95% confidence level, $\alpha = 0.05$, H_0 will be rejected if $t_{0.025,99.19} = -1.66$.
Critical value, $t_{0.025,99.19} = -1.66$, p-value = 0.9073

Since the test statistic, $t = -0.11679 > t_{0.025,99.19} = -1.66$. We reject the null hypothesis. There is sufficient evidence to prove that the sleep time of females is more than the sleep time of males.

```
> t.test(female_data$SleepTime, male_data$SleepTime)

        Welch Two Sample t-test

data:  female_data$SleepTime and male_data$SleepTime
t = -0.11679, df = 99.19, p-value = 0.9073
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5534717  0.4919369
sample estimates:
mean of x mean of y
 7.297101  7.327869
```

Correlation

CORRELATION

To test whether Linear relationship exists between the sleeping time and BMI using Pearson's Product Moment Correlation Coefficient, at 95% confidence level.

Step 1: Calculate the sample correlation coefficient using Pearson's method

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{\left[\left(\sum x^2\right) - \left(\sum x\right)^2/n\right]\left[\left(\sum y^2\right) - \left(\sum y\right)^2/n\right]}}$$

where:

- r = Sample correlation coefficient
- n = Sample size
- x = Value of the independent variable
- y = Value of the dependent variable

r = -0.090696, which indicates that there is a weak negative correlation between x and y.

Step 2: Significant Test for Correlation

Hypotheses:

H_0: $\rho = 0$	(no linear correlation)
H_A: $\rho \neq 0$	(linear correlation exists)

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Calculate test statistic

t= -1.2782

Find critical value:

From t-table, since this is a two-tailed test, there are two critical values:

Lower tail critical value -t/2=0.025,df=197=-1.9721

Upper tail critical value t/2=0.025,df=197=1.9721

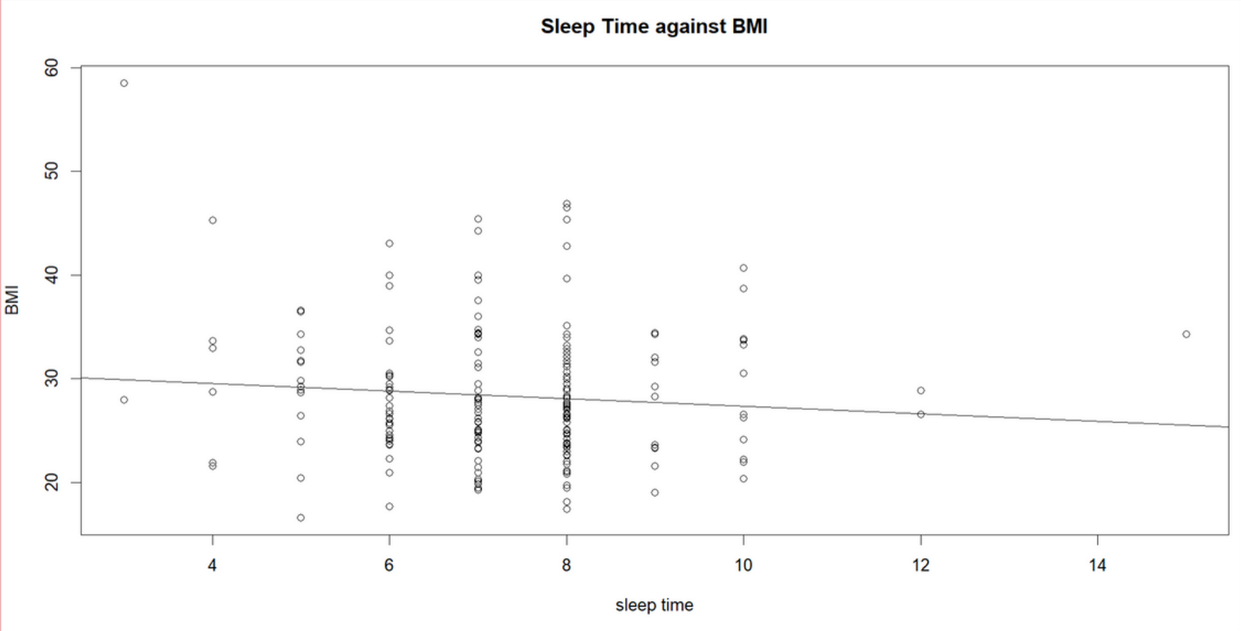
Conclusion

x is sleep time
y is BMI

Pearson's technique: cor(x, y, method = "person")

r = -0.090696 which is negative and falls between -0.5 and 0.5, hence it has a weak linear relationship between sleep time and BMI

Since the test statistic, t=-1.2782 > lower tail critical value t-/2=0.025,df=197 =-1.9721.It does not fall within the rejection region. Hence we failed to reject the null hypothesis. There is sufficient evidence to prove that there is no linear correlation between sleep time and BMI.



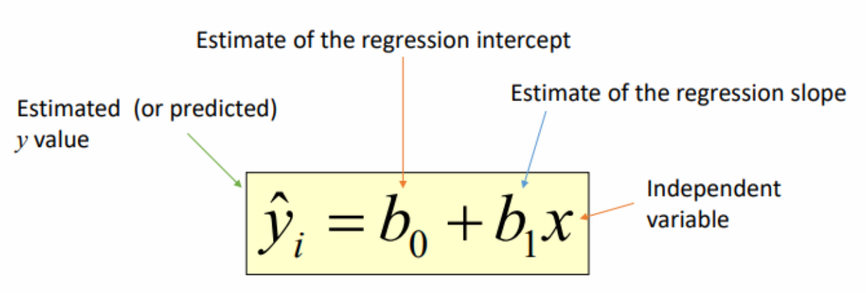

```
> # calculate correlation coefficient
> cor(x,y)
[1] -0.090696
> # calculate test statistic t
> n <- 200
> r <- cor(x,y)
> t <- r/(sqrt((1-(r^2))/(n-2)))
> # significance test for correlation
> # H0: no linear correlation
> # H1: linear correlation exists
> cor.test(Health$SleepTime,Health$BMI, method="pearson")

        Pearson's product-moment correlation

data:  Health$SleepTime and Health$BMI
t = -1.2782, df = 197, p-value = 0.2027
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.22692331  0.04901221
sample estimates:
       cor
-0.090696
```

Regression

REGRESSION

To test whether the rate of heart disease depends on BMI value, using the BMI values as the independent variable(x) and risk of facing physical health problems as the dependent variable(y).

Step 1: Estimated Regression model



a) Find least squares criterion:

- b_0 = Estimate of the regression intercept = -4.8074

  b_1 = Estimate of the regression slope =0.3546

Substitute the variable of b_0 and b_1 into the regression model equation:

$$\hat{y}_i = -4.8074 + 0.3546x$$

b) Explained and unexplained variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum (y - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

```
> SSR <- sum((yhat-mean(y))^2)
> SSR
[1] 1023.874
> SST <- sum((y-mean(y))^2)
> SST
[1] 16823.44
> SSE <- SST-SSR
> SSE
[1] 15799.56
> R2 <- SSR/SST
> R2
[1] 0.06085997
```

c) Find Standard Error of Estimate :

s=8.9555

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

d) Find Standard Deviation of Regression Slope

S =0.0992

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Step 2: Inference about the Slope: t-Test

Hypothesis statement

$H_0 : \beta_1 = 0$ (*no linear relationship*)

$H_1 : \beta_1 \neq 0$ (*linear relationship does exist*)

Find critical value

Lower tail critical value $- t_{\alpha/2=0.025, df=197} = -1.9721$

Upper tail critical value $t_{\alpha/2=0.025, df=197} = 1.9721$

Calculate test statistic

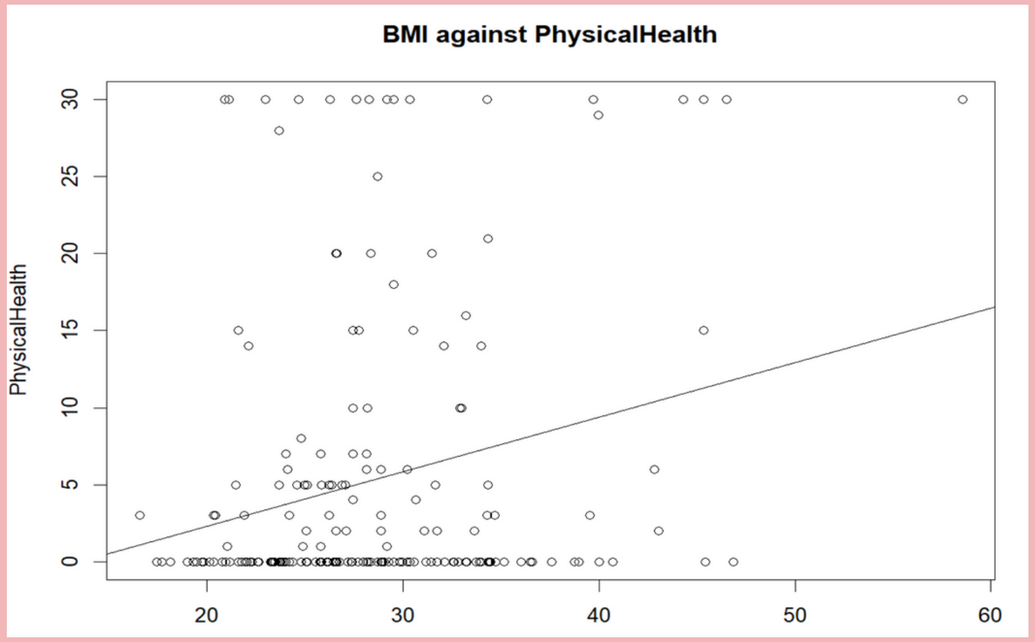
$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

t=3.5730

Conclusion

- x : BMI
- y : Physical Health

Since test statistic t=3.5730 >tα/2=0.025, df=197 = 1.9721 we reject the null hypothesis, H0. There is sufficient evidence that BMI affects physical health, at a=0.05.

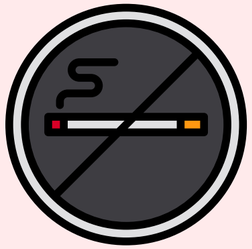



```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8135  -5.1201  -3.2318   0.8675  27.3817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.80735    2.88586  -1.666 0.097337 .
x            0.35462    0.09925   3.573 0.000444 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.955 on 197 degrees of freedom
Multiple R-squared:  0.06086,    Adjusted R-squared:  0.05609
F-statistic: 12.77 on 1 and 197 DF,  p-value: 0.0004437
```

Goodness of Fit Test

Goodness of Fit Test

To test whether there are differences between the observed frequency and expected frequency of a total number of people smoking, at a confidence level of 95%.



Step 1: Statement of test hypothesis:

H0: smoke=0.40, do not smoke=0.60

H1: At least one of the proportions is different from the claimed value

Step 2: Calculate the expected frequency:

	Smoke	Do not smoke	Total
Observed frequency, O	83	116	199
Expected Frequency, E	$np = 199 \times 0.4 = 79.6$	$np = 199 \times 0.6 = 119.4$	199

Step 3: Calculate the test statistic @ chi-square value:

- By using RStudio, test statistics value x2 = 0.6227.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Step 4: Find the critical value from the chi-square table/ RStudio:

- Degree of freedom=n-1=2-1=1, =0.05
- From the Chi-square distribution table, critical value, x2=3.84
- From RStudio, the critical value, x2=3.841459

```
#Finding critical value
x2.alpha <- qchisq(alpha,df=1, lower.tail = FALSE)
x2.alpha
```

```
[1] 3.841459
```

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09

Step 5: State the decision

Test statistic value (**x2=0.6227**) < critical value (**x2 k=1,=0.05= 3.84**), we **do not reject** H0 as it falls within the critical region. There is sufficient evidence to warrant rejection of the claim that the total number who smokes is distributed with the given percentages. Thus, it shows that there are differences between the observed frequency and expected frequency who smokes and who do not smoke.

R script:

```r
#Goodness of Fit Test
#H0: Psmoke = 0.4, Pdonotsmoke = 0.6
#H1: At least one of the proportion is different from the claimed value
smoking.freq <- table(Health[,3])
smoking <- c(83, 116)
prob <- c(0.4, 0.6)
alpha <- 0.05

#Finding critical value
x2.alpha <- qchisq(alpha,df=1, lower.tail = FALSE)
x2.alpha

#perform chi-square test on the data
#with their correspond probabilities
chisq.test(smoking, p = prob, correct = FALSE)
```

```
> #Goodness of Fit Test
> #H0: Psmoke = 0.4, Pdonotsmoke = 0.6
> #H1: At least one of the proportion is different from the claimed value
> smoking.freq <- table(Health[,3])
> smoking <- c(83, 116)
> prob <- c(0.4, 0.6)
> alpha <- 0.05
> #Finding critical value
> x2.alpha <- qchisq(alpha,df=1, lower.tail = FALSE)
> x2.alpha
[1] 3.841459
> #perform chi-square test on the data
> #with their correspond probabilities
> chisq.test(smoking, p = prob, correct = FALSE)

        Chi-squared test for given probabilities

data:  smoking
X-squared = 0.24204, df = 1, p-value = 0.6227
```

Chi-Square Test of Independence

Chi-Square Test of Independence

In this analysis, we are using variables sex and heart disease, where we will test whether the sex and heart disease are related using the Two Way Contingency Table, at a 95% confidence level.

Step 1: State the test hypothesis

H0 : There is no relationship between variables gender and heart disease

H1 :Variables gender and heart disease are related and dependent

Step 2: Find the critical value:

Critical value, x2=3.841459 (with df=(2-1)(2-1)=1,=0.05))

Step 3: Calculate the expected counts:

Sex	Heart Disease				Total
	No		Yes		
	Obs	Exp	Obs	Exp	
Female	124	$\frac{138\times177}{199} = 122.744$	14	$\frac{138\times22}{199} = 15.256$	138
Male	53	$\frac{61\times177}{199} = 54.256$	8	$\frac{61\times22}{199} = 6.744$	61
Total	177	177	22	22	199

Step 4: Calculate the test statistic value using RStudio:

```
> # perform chi-square test
> tbl = table(Health$Sex, Health$HeartDisease)
> chisq.test(tbl, correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 0.37943, df = 1, p-value = 0.5379
```

When we calculate the statistic using RStudio, we get x2=0.37943, with p-value=0.5379.

Step 5: State the decision:

∴ Since the **test statistic value (x2=0.37943) < critical value (x2k=1,=0.05=3.841)**, hence we **fail to reject H0**. There is sufficient evidence to conclude that there is no relationship between sex and heart disease, at =0.05.

“Happiness lies, first of all, in health.”

Thats's all
Thank you