Medical Statistic Report

Introduction

Drugs company Develop has produced a new drug Vunder to reduce recovery time after surgery. Recovery time is defined by length of time (in hours) that a patient remains in hospital after surgery. This report covers 3 parts (4 main tasks) and aims to do some statistical analysis.

Part 1

In market, there is another drug named Xanadu where the recovery time after surgery is known to follow a Normal distribution with mean 148 hours and standard deviation 41 hours. And early trials by Develop suggest that the mean recovery time using Vunder is 142 hours which means that Vunder is more effective than Xanadu in general.

Now we set a null hypothesis H0: μ_0 =148, and the alternative hypothesis H1:μ_1=142. **Task 1** is aimed to calculate sample size to reject null hypothesis with significance level 10% ($\alpha = 10\%$) and power 90% ($1 - \beta = 90\%$).

To calculate sample size we have already know α=0.1 and β=0.9, we can use this formula to calculate: $n = (\frac{\sigma}{\mu}(Z_{1-\alpha} - Z_{\beta}))^2$, μ=(μ_1-μ_0), σ denotes the standard deviation of recovery time.

Using R to calculate and the final result is about 307. In conclusion we need more than 307 samples to reject null hypothesis.

Part 2

In the hospital Atesta a trail is carried out with 80 patients. Now we have the recovery time for each patient.

Task 2 aimed to construct a 99% confidence interval for μ. To construct a 99% confidence interval, we can use this formula to calculate: $\bar{x} \pm t_{n-1,0.995}\frac{s}{\sqrt{n}}$. We can also use t-test to construct confidence interval for μ. The results of these two methods are the same. Results show below.

As I set μ_0=148 and 99% confident level for t-test, and I use two t-tests to calculate.

In the first t-test, null hypothesis is true mean less than 148.we get a final p-value which is 0.4679. That means we can accept null hypothesis. In the second t-test, null hypothesis is true mean greater than 148. And p-value is 0.5321. That means we can accept null hypothesis. We can see these t-tests below. In conclusion, the mean value of recovery time has two possibilities:

1. True mean of recovery time is less than 148.
2. True mean of recovery time is greater than 148.

Meantime we can see the confidence interval is between 137.3616 and 159.3884. However, we cannot support that Vunder is an improvement on Xanadu. It is because there are two possibilities for the recovery time. We cannot determine whether it must be less than 148 or greater than 148.

```
##
##  One Sample t-test
##
## data:  data$Recovery.time
## t = 0.08085, df = 79, p-value = 0.4679
## alternative hypothesis: true mean is greater than 148
## 99 percent confidence interval:
##  137.3616      Inf
## sample estimates:
## mean of x
##   148.375
```

```
##
##  One Sample t-test
##
## data:  data$Recovery.time
## t = 0.08085, df = 79, p-value = 0.5321
## alternative hypothesis: true mean is less than 148
## 99 percent confidence interval:
##      -Inf 159.3884
## sample estimates:
## mean of x
##   148.375
```

Part 3

Now consider the performance if Vunder could depend upon a wide range of covariates. We select data (including age, sex, weight, height, smoker, hospital) from m hospitals M patients.

We need to create a new feature named bmi. The bmi is given by (weight in kg)/(height in m)^2. And there are four main categories of bmi underweight (<18.5),

healthy (18.5-25), overweight (25-30) and obese (>30). Task 3 is aimed to judge if the bmi category of a patient receiving Vunder independent of hospital. Create a table to observe.

	1	2	3	4
1	8	12	23	28
2	13	37	32	33
3	12	49	17	24

To identify independent, we can use chi-test to do that. In R we use chisq.test to do chi-test. The result shows below. The p-value of chi-test is about 0.002. We can find the p-value is small so we must reject independence. The bmi category does depend on hospital.

```
        Pearson's Chi-squared test

data:  table(data$hospital, data$bmi_category)
X-squared = 20.52, df = 6, p-value = 0.002237
```

Finally in Task 4 we need to find the most appropriate linear regression for predicting the recovery time. Meantime we can use the most appropriate model to predict the recovery time of a 59-year-old male who smokes in hospital 3 given that their height is 177cm and weight is 99.1kg.

We can use lm() function to do linear regression, and we can also use step() function to find the best model. The result shows below. The best model has three covariates (smoker, hospital and bmi). We can predict y by the formula: $y = 19.14 * smoker + 5.22 * hospital + 0.63 * bmi + 114.08$. Finally we use predict function to predict y and we get the value is 168.80.

```
Step:  AIC=1464.16
Recovery.time ~ smoker + hospital + bmi

           Df Sum of Sq   RSS    AIC
<none>                  45210 1464.2
- bmi       1    3839.8 49050 1485.6
- hospital  1    4455.0 49665 1489.2
- smoker    1   13119.6 58329 1535.5

Call:
lm(formula = Recovery.time ~ smoker + hospital + bmi, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-31.768  -8.149   0.019  10.449  33.968

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 114.0829     4.3299  26.347 < 0.0000000000000002 ***
smoker       19.1362     2.1079   9.078 < 0.0000000000000002 ***
hospital      5.2241     0.9875   5.290          0.000000245 ***
bmi           0.6293     0.1281   4.911          0.000001530 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.62 on 284 degrees of freedom
Multiple R-squared:  0.315,     Adjusted R-squared:  0.3077
F-statistic: 43.53 on 3 and 284 DF,  p-value: < 0.00000000000000022
```

Conclusion

In Task 1 we calculate the sample size that we need more than 307 samples to reject null hypothesis.

In Task 2 we construct 99% confident interval and conclude that the t-test does not support that Vunder is an improvement on Xanadu.

In Task 3 we create a new variable bmi and in chi-test we found that bmi category depends on the hospital.

In Task 4 we built a linear regression model by using step() function and predict the recovery time of a 59-year-old male who smokes in hospital 3 given that their height is 177cm and weight is 99.1kg and the result was 168.80.