STA2002 Assignment 3

Chen Ang (118010009)

1

(a)

$$\bar{x} = 26, s^2 = 27.14$$

(b)

$$H_0 : \mu = \mu_0 = 26$$
$$H_1 : \mu > \mu_0 = 26$$
$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \alpha = 0.05$$

(c)

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{26 - 26}{4/\sqrt{15}} = 0$$

(d)

Under null hypothesis,

$$T \sim N(0, 1)$$

And so the p-value

$$p = P(T \geq t = 0) = 0.5 > 0.05$$

So Charles does not reject H_0 at 0.05 significance level.

(e)

We need

$$p = P(T \geq t) \leq 0.05$$

which means

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{0.05}$$

Equivalently

$$\begin{aligned}\bar{x} &\geq \mu_0 + z_{0.05}\sigma/\sqrt{n} \\ &= 26 + 1.645 \cdot 4/\sqrt{15} \\ &= 27.70\end{aligned}$$

So $\bar{x}$ should at least be 27.70 for Charles to reject H_0.

(f)

The test statistics becomes

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

with realization

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 0$$

Under H_0,

$$T \sim t(n-1)$$

The p-value is given by

$$p = P(T \geq t = 0) = 0.5 > 0.05$$

Charles does not reject H_0 at 0.05 significance level.

2

Denote the outcome of each coin flip by $X_1, \cdots X_{n=50} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. The hypotheses are:

$$H_0 : p = p_0 = 0.5$$
$$H_1 : p \neq p_0 = 0.5$$

(a)

Since n is relatively large, by CLT

$$\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

We construct the statistic

$$T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

which has realization

$$t = \frac{30/50 - 0.5}{\sqrt{0.5(1-0.5)/50}} = 1.414$$

Under H_0,

$$T \overset{\text{approx}}{\sim} N(0,1)$$

We compute the p-value

$$p = P(|T| \geq 1.414) = 2 \cdot P(Z \geq 1.414) = 0.157 > 0.05$$

Hence Michael fails to reject H_0 at $\alpha = 0.05$.

(b)

Keeping $n = 50$ and $\alpha = 0.05$, we want the p-value

$$p = P(|T| \geq t) = 2 \cdot P(Z \geq t) \leq \alpha$$

which means

$$P(Z \geq t) \leq \frac{\alpha}{2} \iff t = \frac{h/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z_{\alpha/2}$$

Equivalently

$$h \geq n \cdot \left(p_0 + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}\right) = 31.9$$

So h at least needs to be $\lceil 31.9 \rceil = 32$ for Michael to reject H_0 at $\alpha = 0.05$.

3

(a)

Given $\mu = \mu'$,

$$\bar{X} \sim N\left(\mu', \frac{\sigma^2}{n}\right)$$

We falsely accept H_0 when the statistic $\bar{x} \notin C$, i.e.

$$\bar{x} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

with probability

$$\begin{aligned} P\left(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) &= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} \leq z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \leq z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) \\ &= \beta(\mu') \end{aligned}$$

which is to be shown.

(b)

We want

$$\beta(\mu') = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

Taking the inverse of Φ on both sides,

$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = \Phi^{-1}(\beta) = z_{1-\beta} = -z_\beta$$

which yields

$$n = \left(\frac{\sigma\,(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right)^2 \tag{1}$$

as desired.

(c)

Plugging $\alpha = 0.025, \beta = 0.05, \mu' = 25000, \mu_0 = 24000, \sigma = 1300$ into (1), we need

$$n = \left(\frac{1300\left(z_{0.025} + z_{0.05}\right)}{24000 - 25000} \right)^2 = 21.96 \approx 22$$

samples.

4

(a)

The common CDF of the Beta RVs is

$$F_Y(y) = \int_0^y 2t\, dt = y^2, \quad y \in (0, 1)$$

Hence the CDF of $Y_{(10)}$ is given by

$$\begin{aligned} F_{Y_{(10)}}(y) &= P\left(Y_{(10)} \le y\right) \\ &= P(Y_1, \cdots, Y_{10} \le y) \\ &= \prod_{i=1}^{10} P(Y_i \le y) \\ &= \prod_{i=1}^{10} y^2 = y^{20}, \quad y \in (0, 1) \end{aligned}$$

Differentiation w.r.t. y gives the PDF

$$f_{Y_{(10)}} = 20y^{19}$$

with support $(0, 1)$.

(b)

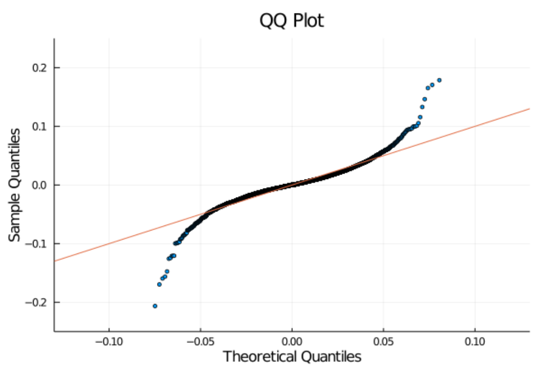
$$P\left(Y_{(10)} > 0.9\right) = 1 - F_{Y_{(10)}}(0.9) = 0.878$$

5

(a)

$$\begin{aligned} \bar{r} &= 8.87 \times 10^{-4} \\ s_r^2 &= 4.68 \times 10^{-4} \end{aligned}$$

(b)



(c)

The QQ plot fails to fit the identity line very well. In particular, the sample quantiles deviate from the theoretical quantiles significantly in both tails. This suggests that the actual distribution of R_i might be more heavy-tailed than what the model predicted.

6

(a)

$$H_0 : p_A - p_B = 0$$
$$H_1 : p_A - p_B \neq 0$$

(b)

$$\hat{p}_A = 0.1205$$
$$\hat{p}_B = 0.1188$$

(c)

The two-sided test statistic is given by

$$t = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n + 1/m)}} = 1.3626$$

Here $\hat{p}$ denotes the value of the pooled estimator of p.

d)

The p-value is given by

$$p = P(|T| \geq 1.3636) = 2 \cdot P(T \geq 1.3626) = 0.173 > 0.05$$

Hence we fail to reject H_0 at 5%.

7

(a)

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y > 0$$

(b)

Since $\sigma_X^2 = \sigma_Y^2$, we select the pooled t-test. The test statistic T is realized at

$$t = \frac{\bar{x} - \bar{y}}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} = -8.3286$$

where S_p is the pooled standard deviation. Under H_0,

$$T \sim N(0, 1)$$

We see the p-value is well above 0.05 as

$$p = P(T \geq t) > P(T \geq 0) = 0.5 > 0.05$$

Thus we fail to reject H_0 at 5%.

(c)

Since $\sigma_X^2 \neq \sigma_Y^2$, we select Welch's t-test. The Welch's statistic T is realized at

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} = -8.3369$$

Under H_0, since n, m are both large,

$$T \overset{\text{approx}}{\sim} N(0,1)$$

and the p-value is well above 0.05 for the same reason as in **(b)**. So we fail to reject H_0 at 5%.

(d)

The F statistic T is realized at

$$t = \frac{S_X^2}{S_Y^2} = 1.0049$$

Under H_0,

$$T \sim F(n-1, m-1) = F(13390, 5671)$$

We adopt the critical region approach and see

$$F_{0.025}(13390, 5671) = 0.9572 < t < 1.0451 = F_{0.975}(13390, 5671)$$

Hence we fail to reject H_0 at 5%.

Julia Code

5

```
using CSV
using Distributions
using Plots
```

```
# (a)
```

```
s = CSV.read("MSFT_stock.csv")[2]
r = log.(s[i] / s[i-1] for i in 2:size(s)[1])
```

```
r̄ = mean(r)
```

```
0.0008868566305898707
```

```
sᵣ² = var(r)
```

```
0.0004675170547290336
```

```
# (b)
```

```
n = size(r)[1]
theo_Q = quantile.(Normal(r̄, sqrt(sᵣ²)), [i / (n + 1) for i in 1:n])
samp_Q = sort(r)

plot(theo_Q, samp_Q, seriestype = :scatter, title = "QQ Plot",
     markersize = 3, legend = false, xlims = (-0.13, 0.13), ylims = (-0.25, 0.25),
     xlabel = "Theoretical Quantiles", ylabel = "Sample Quantiles")
plot!(x -> x)
```

6

```julia
# (b)
```

```julia
df = CSV.read("AB_2020.csv")
A = df[df[:landing_page] .== "A", end]
B = df[df[:landing_page] .== "B", end]
n, m = size(A)[1], size(B)[1]
```

```julia
p̂ᴬ = mean(A)
```

```
0.1204794729871982
```

```julia
p̂ᴮ = mean(B)
```

```
0.1188495364868077
```

```julia
# (c)
```

```julia
p̂ = mean([A; B])
```

```julia
t = (p̂ᴬ - p̂ᴮ) / sqrt(p̂ * (1 - p̂) * (1/n + 1/m))
```

```
1.362602512273659
```

7

```julia
df = CSV.read("traffic.csv")
X = df[df[:weather_main] .== "Clear", end]
Y = df[df[:weather_main] .== "Rain", end]
n, m = size(X)[1], size(Y)[1]
Sₚ = (n - 1) * var(X) + (m - 1) * var(Y)
Sₚ /= n + m - 2
Sₚ = sqrt(Sₚ)
```

```julia
t = mean(X) - mean(Y)
t /= Sₚ * sqrt(1/n + 1/m)
```

```
-8.328589157910267
```

```julia
# (c)
```

```julia
t = mean(X) - mean(Y)
t /= sqrt(var(X)/n + var(Y)/m)
```

```
-8.336873582326676
```

```julia
# (d)
```

```julia
t = var(X) / var(Y)
```

```
1.0049224336665912
```

```julia
quantile.(FDist(13390, 5671), [0.025,0.975])
```

```
2-element Array{Float64,1}:
 0.9572211255062762
 1.0451043846031538
```