Sample Surveys Question 3

Chesia Anyika

2024-03-13

Libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.5.0     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors
```

Question

In a study to estimate the total sugar content of a truckload of oranges, a random sample of oranges was juiced and weighed. The total weight of all the oranges, obtained by first weighing the truck loaded and unloaded, was found to be 1800 pounds.

Orange	Sugar Content	Weight of orange
1	0.021	0.40
2	0.030	0.48
3	0.025	0.43
4	0.022	0.42
5	0.033	0.50
6	0.027	0.46
7	0.019	0.39
8	0.021	0.41
9	0.023	0.42
10	0.025	0.44

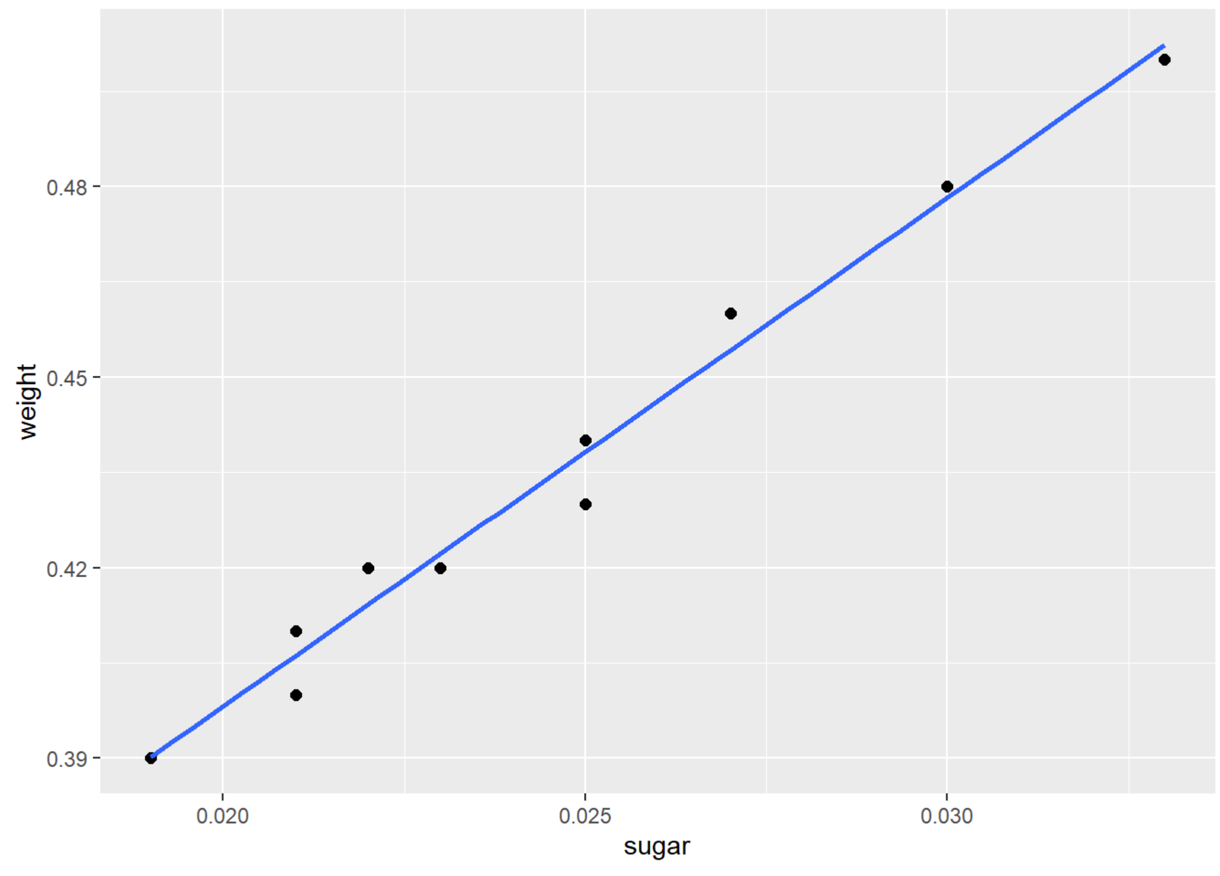
a) Make a Scatter Plot of the ten sample values. Do you think there is an association between sugar content and weight?

```
#input values into vectors
sugar <- c(0.021,0.030,0.025,0.022,0.033,0.027,0.019,0.021,0.023,0.025)
weight <- c(0.40,0.48,0.43,0.42,0.50,0.46,0.39,0.41,0.42,0.44)

#create a dataframe using the vectors
oranges <- data.frame(sugar, weight)

#plot the scatterplot with a regression line
ggplot(oranges, aes(x=sugar, y=weight)) +
  geom_point(size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



There is a strong positive association between sugar content and weight of an orange, as evidenced by the regression plot above. This means that the higher the sugar content of the orange, the higher the weight.

We can find the specific regression equation $\hat{y} = \beta_0 + \beta_1 x$ using the `lm()` function in R, as follows:

```
# Fit linear regression model
model <- lm(weight ~ sugar, data = oranges)

# Print the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ sugar, data = oranges)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0082019 -0.0022274  0.0008005  0.0033121  0.0058121
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.238086   0.009585   24.84 7.38e-09 ***
## sugar       8.004640   0.384212   20.83 2.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005045 on 8 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9796
## F-statistic: 434.1 on 1 and 8 DF,  p-value: 2.955e-08
```

From the regression model the equation is

$$\hat{y} = 0.238 + 8.005x$$

Where:

- $\hat{y}$ represents the predicted value of y.
- x is the predictor variable.
- β_1 is 8.005
- β_0 is 0.238

b) Estimate τ_y, the total Sugar content for the oranges.

We can use the above regression equation to compute this, by inputting our known total weight of oranges as the x parameter, and thus estimating the total sugar content in the oranges.

First, I defined my known parameters

```
#define the parameters
N = 1

beta0 <- 0.238
beta1 <- 8.005
```

I then used these to solve for $\hat{y}$ in my regression equation.

```
#solve for regression equation
yhat <- beta0 + (beta1*N)

#print results
cat('Estimator of total sugar content is', yhat)
```

```
## Estimator of total sugar content is 8.243
```

Thus the total sugar content in the oranges has been estimated at 8.243 units of concentration.

c) Place a 90% CI on the Estimation

As per the formula

$$CI = ME \pm Z * SE$$

Where:

- CI is the Confidence Interval
- ME is the Margin of Error
- Z is the Critical value obtained from the Normal distribution
- SE is the Standard Error

First I accessed the Standard Error value from the regression model created as follows:

```
# Get the standard error of the estimate from the regression output
SE <- summary(model)$sigma

#print the values
cat('SE is', SE)
```

```
## SE is 0.005044753
```

I then calculated the Z score, aka the Critical Value for a 90% confidence interval

```
#dedfine confidence level of 90%
CL <- 0.90

#define alpha values
alpha <- 1 - CL
alphahalf <- alpha/2

#define cumulative probability
p <- 1-alphahalf

# Calculate the critical value for a 90% confidence interval
CV <- qnorm(p)

#print results
cat('Z Score is', CV)
```

```
## Z Score is 1.644854
```

I then computed the margin of error by multiplying the Critical Value and the Standard Error Obtained.

```
# Calculate the margin of error
ME <- CV * SE

cat('Margin of Error is', ME)
```

```
## Margin of Error is 0.008297881
```

Using the previously computed estimator of total sugar content, I then calculated the upper and lower bounds as follows

```
# Calculate the lower and upper bounds of the confidence interval
lower <- yhat - ME
upper <- yhat + ME

# Print the confidence interval
cat('90% Confidence Interval for Total Sugar Content: [', lower, ',', upper, ']' )
```

```
## 90% Confidence Interval for Total Sugar Content: [ 8.234702 , 8.251298 ]
```

The confidence Interval Obtained is $CI = [8.234702, 8.251298]$. This is a narrow confidence interval suggesting high accuracy of the regression estimator.