Inference Of Population Means

Mallikarjuna

2023-09-26

# Load necessary packages  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

# Set ggplot theme for visualizations  
theme\_set(ggthemes::theme\_few())  
  
# Set options for flextables  
set\_flextable\_defaults(na\_str = "NA")  
  
# Load function for printing tables nicely  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

# Get a sample  
  
sample(1:100, size=10)

## [1] 3 81 92 32 11 56 49 21 60 52

### Let’s import data from the website GitHub to use for this activity.

# Load Palmer penguins data  
penguins <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/penguins.csv", show\_col\_types = FALSE)

### *As a researcher interested in penguins, we may be interested in the average flipper length (mm) of penguins from Palmer Archipelago, Antarctica. From our previous experience, we believe that penguins typically have an average flipper length of 200mm, but is this true for these penguins from Antarctica as well? We can explore this using a one-sample t-test.*

**Formally state the hypotheses for our question of interest.**

We are interested in testing whether the average flipper length of all penguins in Palmer Archipelago, Antarctica is equal to 200mm or not.

Formally, the hypotheses for our corresponding one-sample -test are below.

H0: vs. Ha:

where is average flipper length of all penguins in Palmer Archipelago, Antarctica.

**What is the null value, 0, for this test?**

# Calculating descriptive statistics  
quant1Stats <- penguins %>%   
 dplyr::summarize(  
 Minimum = min(flipper\_length\_mm, na.rm = TRUE),  
 Q1 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.25),  
 M = median(flipper\_length\_mm, na.rm = TRUE),  
 Q3 = quantile(flipper\_length\_mm, na.rm = TRUE, probs = 0.75),  
 Maximum = max(flipper\_length\_mm, na.rm = TRUE),  
 Mean = mean(flipper\_length\_mm, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(flipper\_length\_mm, na.rm = TRUE),  
 n = n()  
)  
  
# Printing table of statistics  
quant1Stats %>%   
 make\_flex(caption = "Quantitative summary statistics for penguin flipper lengths (mm).")

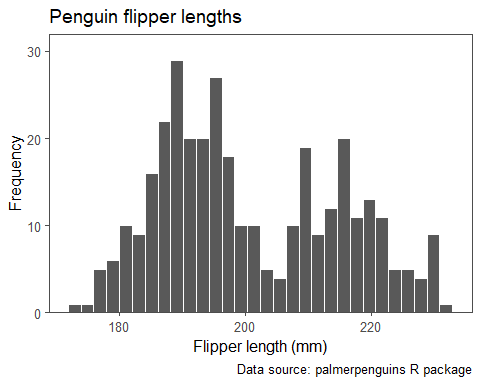
Table 1: Quantitative summary statistics for penguin flipper lengths (mm).

| Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 172.00 | 190.00 | 197.00 | 213.00 | 231.00 | 200.92 | 59.00 | 14.06 | 344 |

# Creating a histogram  
penguins %>%   
 ggplot(aes(x = flipper\_length\_mm)) +   
 geom\_histogram(color = "white") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Penguin flipper lengths",  
 x = "Flipper length (mm)",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing non-finite values (`stat\_bin()`).



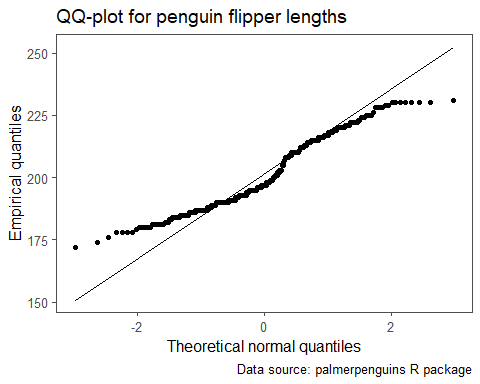
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Quantile-Quantile plot

# Creating a Quantile-Quantile (QQ) plot  
penguins %>%   
 ggplot(aes(sample = flipper\_length\_mm)) +   
 stat\_qq\_line() +  
 stat\_qq() +  
 labs(title = "QQ-plot for penguin flipper lengths",  
 x = "Theoretical normal quantiles",  
 y = "Empirical quantiles",  
 caption = "Data source: palmerpenguins R package")

## Warning: Removed 2 rows containing non-finite values (`stat\_qq\_line()`).

## Warning: Removed 2 rows containing non-finite values (`stat\_qq()`).



##Checking the model assumptions

**State each assumption for the one-sample t-test, and indicate whether or not each assumption is met, citing specific evidence from the output obtained.**

1. **Independent Observations**: These measurements are for separate penguins and we assume they were randomly selected for the sake of the problem.
2. **Normality Assumption**: Since n = 344, which is larger than 100, the normality assumption is met as per Central Limit theorem.

## Next, we implement the one-sample t-test.

# Implementing a one-sample t-test using R  
ttestRes <- t.test(penguins$flipper\_length\_mm,  
 mu = 200, conf.level = 0.95)  
  
# Printing model output  
ttestRes %>%   
 broom::tidy() %>%   
 make\_flex()

| estimate | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 200.92 | 1.20 | 0.23 | 341.00 | 199.42 | 202.41 | One Sample t-test | two.sided |

**State our p-value, decision, and conclusion in the context of the problem testing at the 5% significance level, citing specific evidence from the obtained output.**

***p-value***: 0.23

***decision***: Since the p-value of 0.23 is grater than 0.05, we fail to reject H0.

***conclusion***: We have insufficient evidence that the mean flipper lengths of penguins in Palmer Archipelago, Antarctica differs from 200mm.

**Provide and interpret a 95% confidence interval for the average flipper length of penguins from Palmer Archipelago, Antarctica.**

CI limits for : (199.42,202.41)

We are 95% confident that the mean flipper length of penguins in Palmer Archipelago, Antarctica is between 199.42mm and 202.41mm.

**Do the results of the t-test and 95% confidence interval align? Why or why not?**

## Two-sample t-test

A formal statement of our hypotheses for testing if penguins from the Dream island weigh the same as penguins from the Biscoe island, on average, is below:

H0: vs Ha:

where is the average body mass of penguins from Dream island and is the average body mass of penguins from Biscoe island.

**What would the statement of hypotheses be for testing if Biscoe penguins have a greater average body mass in grams than penguins from the Dream island?**

In that case, our hypotheses would be .

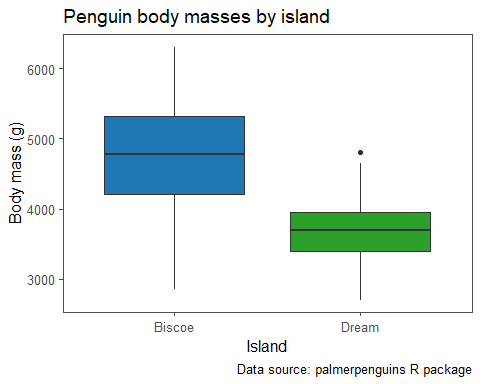
# Calculating descriptive statistics  
quant2Stats <- penguins %>%   
 dplyr::filter(island %in% c("Biscoe", "Dream")) %>%   
 group\_by(island) %>%   
 summarize(  
 Minimum = min(body\_mass\_g, na.rm = TRUE),  
 Q1 = quantile(body\_mass\_g, na.rm = TRUE, probs = 0.25),  
 M = median(body\_mass\_g, na.rm = TRUE),  
 Q3 = quantile(body\_mass\_g, na.rm = TRUE, probs = 0.75),  
 Maximum = max(body\_mass\_g, na.rm = TRUE),  
 Mean = mean(body\_mass\_g, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(body\_mass\_g, na.rm = TRUE),  
 n = n()  
)  
  
# Printing table of statistics  
quant2Stats %>%   
 make\_flex(caption = "Summary statistics for penguin body masses (g) by species.")

Table 3: Summary statistics for penguin body masses (g) by species.

| island | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Biscoe | 2,850.00 | 4,200.00 | 4,775.00 | 5,325.00 | 6,300.00 | 4,716.02 | 3,450.00 | 782.86 | 168 |
| Dream | 2,700.00 | 3,400.00 | 3,687.50 | 3,956.25 | 4,800.00 | 3,712.90 | 2,100.00 | 416.64 | 124 |

# Creating side-by-side box plots  
penguins %>%   
 dplyr::filter(island %in% c("Biscoe", "Dream")) %>%   
 ggplot(aes(x = island, y = body\_mass\_g, fill = island)) +   
 geom\_boxplot() +   
 scale\_fill\_manual(values = c("#1F77B4", "#2CA02C")) +  
 labs(title = "Penguin body masses by island",  
 x = "Island",  
 y = "Body mass (g)",  
 caption = "Data source: palmerpenguins R package") +  
 theme(legend.position = "none")

## Warning: Removed 1 rows containing non-finite values (`stat\_boxplot()`).



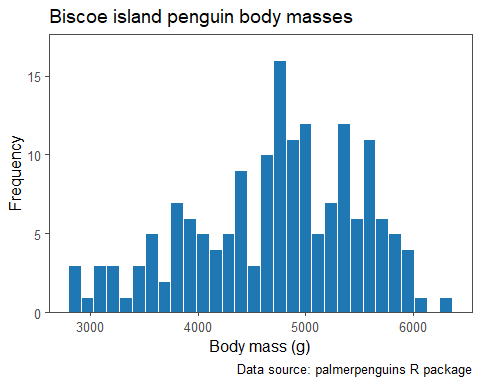
## Warning: Removed 1 rows containing non-finite values (`stat\_boxplot()`).

**Describe what the box plots show us in terms of the center and symmetry of the body masses for both the Biscoe and Dream island penguins.**

# Creating a histogram  
penguins %>%   
 dplyr::filter(island %in% c("Biscoe")) %>%   
 ggplot(aes(x = body\_mass\_g)) +   
 geom\_histogram(color = "white", fill = "#1F77B4") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Biscoe island penguin body masses",  
 x = "Body mass (g)",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

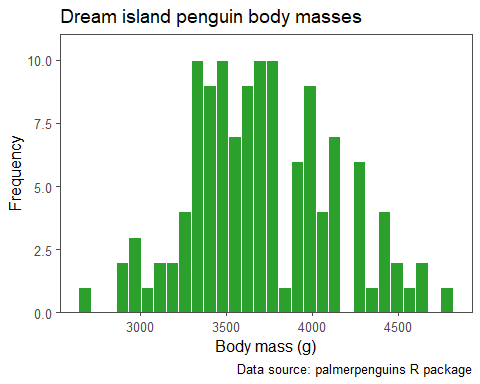
## Warning: Removed 1 rows containing non-finite values (`stat\_bin()`).



## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

# Creating a histogram  
penguins %>%   
 dplyr::filter(island %in% c("Dream")) %>%   
 ggplot(aes(x = body\_mass\_g)) +   
 geom\_histogram(color = "white", fill = "#2CA02C") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Dream island penguin body masses",  
 x = "Body mass (g)",  
 y = "Frequency",  
 caption = "Data source: palmerpenguins R package")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

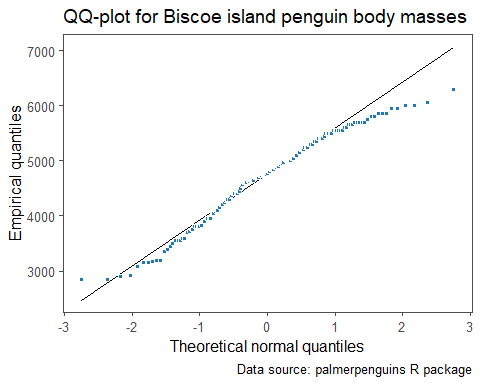
**Describe what the histograms show us in terms of the body masses for the penguins.**

Fairly Symmetric and unimodal.

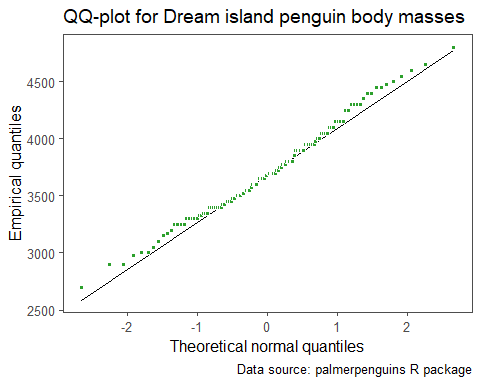
# Creating a Quantile-Quantile (QQ) plot  
penguins %>%   
 dplyr::filter(island %in% c("Biscoe")) %>%   
 ggplot(aes(sample = body\_mass\_g)) +   
 stat\_qq\_line() +  
 stat\_qq(color = "white", fill = "#1F77B4", pch = 21) +  
 labs(title = "QQ-plot for Biscoe island penguin body masses",  
 x = "Theoretical normal quantiles",  
 y = "Empirical quantiles",  
 caption = "Data source: palmerpenguins R package")

## Warning: Removed 1 rows containing non-finite values (`stat\_qq\_line()`).

## Warning: Removed 1 rows containing non-finite values (`stat\_qq()`).



# Creating a Quantile-Quantile (QQ) plot  
penguins %>%   
 dplyr::filter(island %in% c("Dream")) %>%   
 ggplot(aes(sample = body\_mass\_g)) +   
 stat\_qq\_line() +  
 stat\_qq(color = "white", fill = "#2CA02C", pch = 21) +  
 labs(title = "QQ-plot for Dream island penguin body masses",  
 x = "Theoretical normal quantiles",  
 y = "Empirical quantiles",  
 caption = "Data source: palmerpenguins R package")



### Provide our decision for the normality assumption for this data set, citing specific evidence from output obtained.

## Two-sample t-test

**Let’s conduct the unpooled independent samples t-test:**

# Creating vectors of values  
dreamMasses <- dplyr::filter(penguins, island == "Dream") %>%   
 dplyr::pull(body\_mass\_g)  
  
biscoeMasses <- dplyr::filter(penguins, island == "Biscoe") %>%   
 dplyr::pull(body\_mass\_g)  
  
# Implementing the two-sample t-test  
ttestRes2 <- t.test(x = dreamMasses,  
 y = biscoeMasses,  
 mu = 0,  
 conf.level = 0.95)  
  
# Printing model output  
ttestRes2 %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1)

| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -1,003.1 | 3,712.9 | 4,716.0 | -14.1 | <2e-16 | 264.8 | -1,143.3 | -862.9 | Welch Two Sample t-test | two.sided |

### Implementing the test

### Provide our test statistic, p-value, and decision at the 5% significance level citing evidence from the output.

**test statistic:** -14.1

**p-value:** <2e-16

**decision:** Since the p-value is <0.05, we reject the H0.

**Provide our interpretation of the result of the two-sample t-test based on our decision.**

Since the p-value is <0.05, we have sufficient evidence that the average body mass in grams for penguins from the Dream island is less that of penguins from Biscoe island at the 5% significance level.

*Any time we have a two sample t-test that is statistically significant (p-value < 0.05), we should indicate which group was higher or lower, on average in our interpretation.*

**Provide and interpret the confidence interval for the difference in the means between the two groups.**

**confidence interval:** (-1,143.3, -862.9) -> (862.9, 1143.3)

We are 95% confident that the mean body mass in grams for penguins from the Dream island is between 862.9 and 1143.3 grams less than that of penguins from the Biscoe island.

*We never include negative numbers / bounds in confidence interval interpretations.*

**Which is more useful for describing the difference, if any, between Dream island and Biscoe island penguins in terms of their average body masses - the two-sample t-test or the confidence interval? Why?**

The confidence interval was more useful since it tells us whether or not there is a statistically significant difference between the two groups, and it quantifies the magnitude of the difference between the two groups.