Are Male Gentoo Penguins Significantly Heavier Than Females?

CID: 02385539

2026-02-09

Why This Question

I stumbled upon the Palmer Penguins dataset[1] a few months ago while reading a book on Machine Learning. While exploring it alongside the author, I found that not only is it one of the cutest datasets, but also one of the most beginner friendly for data cleaning, exploration, and interpretation (and I happen to be a beginner). It contains the species, sex, body mass, bill length, and more on 344 penguins (333 after cleaning), which is big enough to satisfy the Central Limit Theorem but small enough that I can easily calculate estimators manually or visualise every datapoint without overplotting.

Moreover, when exploring the data, I discovered a distinct bimodality in the body mass distribution of Gentoo penguins (one of the three species in the dataset). Rather than being a bell curve, as expected, the distribution has a double-hump. After colouring the Gentoo penguin data according to sex, it became clear that this was really the amalgamation of two normal distributions, a phenomenon I found fascinating (see Figure 1). I wondered then: within a penguin species, is the size difference between the two sexes statistically significant enough to be modeled parametrically?[2]

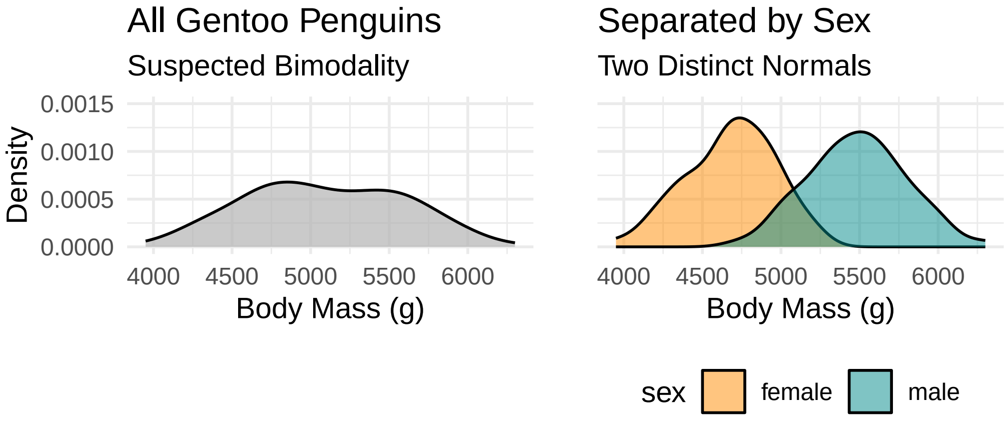

Figure 1: Left: The initial bimodal distribution. Right: The same data separated by sex. (Note: Y-axes fixed to same scale for comparison)

[1] Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

[2] The full reproducible code repository is available at: github.com/recursivechockler/MATH50011-Coursework-1.

Theoretical Framework

Derivation of MLE

Another convenience of this dataset is the ease with which one can assume i.i.d. observations (with the concession that family-related penguins would likely share similar characteristics). Hence, I chose to treat the body mass measurements $x_1, \dots, x_n$ of Gentoo penguins as independent and normally distributed:

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n$$

where the parameters are $\theta = (\mu, \sigma^2)^T \in \mathbb{R} \times (0, \infty)$.

To work out the MLE: First, note that the probability density function (PDF) of a single observation X_i is:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Therefore, the Likelihood function $L(\theta)$ is the product of the individual PDFs:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Simplifying this expression:

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

Taking logs of both sides gives the Log-Likelihood function, which we denote as $\ell(\mu, \sigma^2)$:

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

To find the Maximum Likelihood Estimators, we differentiate with respect to μ and σ^2, equate to 0, and solve:

1. **For** μ:
$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \implies \sum x_i - n\hat{\mu} = 0 \implies \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

2. **For** σ^2:
$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Now, while the MLE for $\hat{\mu}$ is unbiased, the MLE for $\widehat{\sigma^2}$ is biased (as proven in lectures).

To determine if our estimator $\hat{\mu} = \bar{X}$ is efficient, we compare its variance to the Cramér-Rao Lower Bound.

Efficiency of Statistics

1. Fisher Information ($I_n(\mu)$): The log-likelihood function derived previously is:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

We take the first and second partial derivatives with respect to μ:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1) = -\frac{n}{\sigma^2}$$

The Fisher Information is defined as the negative expectation of the second derivative:

$$I_n(\mu) = -E\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] = -E\left[-\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2}$$

2. The Cramér-Rao Lower Bound: For an unbiased estimator, the lower bound for variance is the reciprocal of the Fisher Information:

$$\text{CRLB} = \frac{1}{I_n(\mu)} = \frac{\sigma^2}{n}$$

3. Variance of our Estimator: We calculate the variance of our MLE $\hat{\mu} = \bar{X}$:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2}\sum Var(x_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Conclusion: Since $Var(\hat{\mu}) = \text{CRLB}$, the estimator $\hat{\mu}_{MLE}$ is **efficient**.

4. Efficiency Check for Variance ($\hat{\sigma}^2_{MLE}$):

We repeat the process for the variance parameter σ^2.

A. Fisher Information ($I_n(\sigma^2)$)

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\implies \frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

Taking the negative expectation (noting that $E[\sum(x_i - \mu)^2] = n\sigma^2$):

$$I_n(\sigma^2) = -E\left[\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(n\sigma^2)\right] = -\left(\frac{n}{2\sigma^4} - \frac{n}{\sigma^4}\right) = \frac{n}{2\sigma^4}$$

B. The Cramér-Rao Lower Bound:

$$\text{CRLB}_{\sigma^2} = \frac{1}{I_n(\sigma^2)} = \frac{2\sigma^4}{n}$$

C. Variance of the Estimator: We know that for Normally distributed data, the sum of squares follows a Chi-Square distribution:

$$\sum(x_i - \bar{x})^2 \sim \sigma^2\chi^2_{n-1}$$

Therefore, our estimator $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(x_i - \bar{x})^2$ has variance:

$$Var(\hat{\sigma}^2_{MLE}) = \frac{1}{n^2}Var(\sigma^2\chi^2_{n-1}) = \frac{\sigma^4}{n^2}(2(n-1)) = \frac{2(n-1)\sigma^4}{n^2}$$

Conclusion: Comparing the actual variance to the CRLB:

$$\frac{2(n-1)\sigma^4}{n^2} \neq \frac{2\sigma^4}{n}$$

The MLE for variance does **not** attain the CRLB for finite n. However, it is **asymptotically efficient**, as the ratio approaches 1 when $n \to \infty$:

$$\lim_{n\to\infty} \frac{2(n-1)\sigma^4/n^2}{2\sigma^4/n} = 1$$

Application & Results

These estimators for $\hat{\mu}$ and $\widehat{\sigma^2}$, as well as the true σ^2, are as follows[3]:

For Gentoo Males ($n = 61$):

$$\hat{\mu}_m = 5484.84 \text{ g}, \quad \hat{\sigma}^2_m = 9.646063 \times 10^4 \text{ g}, \quad \sigma^2_m = 9.806831 \times 10^4$$

For Gentoo Females ($n = 58$):

$$\hat{\mu}_f = 4679.74 \text{ g}, \quad \hat{\sigma}^2_f = 7.791933 \times 10^4 \text{ g}, \quad \sigma^2_f = 7.928634 \times 10^4$$

In the following sections, I will investigate whether the difference in mean body mass between male and female Gentoo penguins is statistically significant. In these sections, I will use the sample mean to form my conclusions. As shown above, no other estimator could possibly be more precise, as the sample mean is both the maximum likelihood estimator of μ and hits the Cramér-Rao Lower Bound.

[3] R's `mean()` function was used to calculate the above mean MLEs, as the MLE for the mean is the sample average. However, R's `var()` function divides by $n-1$ rather than by n, making for an unbiased estimator. Hence, a helper function was written to calculate the MLE for the variance of the data through multiplying the result of this function by $n-1$ then dividing it by n.

Hypothesis Testing

To determine if the difference in body mass is statistically significant, we test the null hypothesis that there is no difference in mean body mass between the sexes against the alternative that a difference exists:

$$H_0 : \hat{\mu}_{male} - \hat{\mu}_{female} = 0$$

$$H_1 : \hat{\mu}_{male} - \hat{\mu}_{female} \neq 0$$

As seen above, this dataset is big enough that using the MLE for variance versus the unbiased variance will make little difference. As it is the standard for t-tests, I will be using the unbiased estimator. Also, it will guarentee that we do not underestimate the variance of the data, keeping the probability of a Type I error (a false positive) to a minimum.

As $n \to \infty$, the estimator $\hat{\mu}$ is asymptotically Normal. Since our samples are independent, the difference between the sample means is also Normally distributed:

$$(\hat{\mu}_m - \hat{\mu}_f) \sim \mathcal{N}\left(\delta, \frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}\right)$$

Where $\delta = \hat{\mu}_{male} - \hat{\mu}_{female}$.

Confidence Interval for the Difference in Mass

We aim to derive a $(1 - \alpha)$ Confidence Interval for the true difference in means δ. To do this, we use the pivotal quantity based on the difference of the sample means. Since the sample sizes are large, we can apply the Central Limit Theorem to approximate the distribution as Normal:

$$1 - \alpha = P\left(-c_{\alpha/2} < \frac{(\hat{\mu}_m - \hat{\mu}_f) - \delta}{\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}} < c_{\alpha/2}\right)$$

Rearranging this inequality to isolate δ, we get the formula for the confidence interval:

$$(\hat{\mu}_m - \hat{\mu}_f) \pm c_{\alpha/2}\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}$$

Plugging in our known values for $\hat{\mu}_m, \hat{\mu}_f, \sigma_m^2, \sigma_f^2, n_m$, and n_f with $\alpha = 0.05$, we calculate the lower and upper bounds. The observed difference in means is 805.09g. With a critical value of $c_{0.025} \approx 1.96$, the 95% Confidence Interval is:

$$[698.2, \quad 911.99]$$

Since this interval does not contain 0, we conclude that the difference in body mass is statistically significant. We can then repeat this analysis across the other species of penguin:

Table 1: 95% Confidence Intervals for Difference in Mean Body Mass (Male - Female) by Species

Species	Lower Bound (g)	Upper Bound (g)
Adelie	573.92	775.39
Chinstrap	256.79	566.73
Gentoo	698.20	911.99

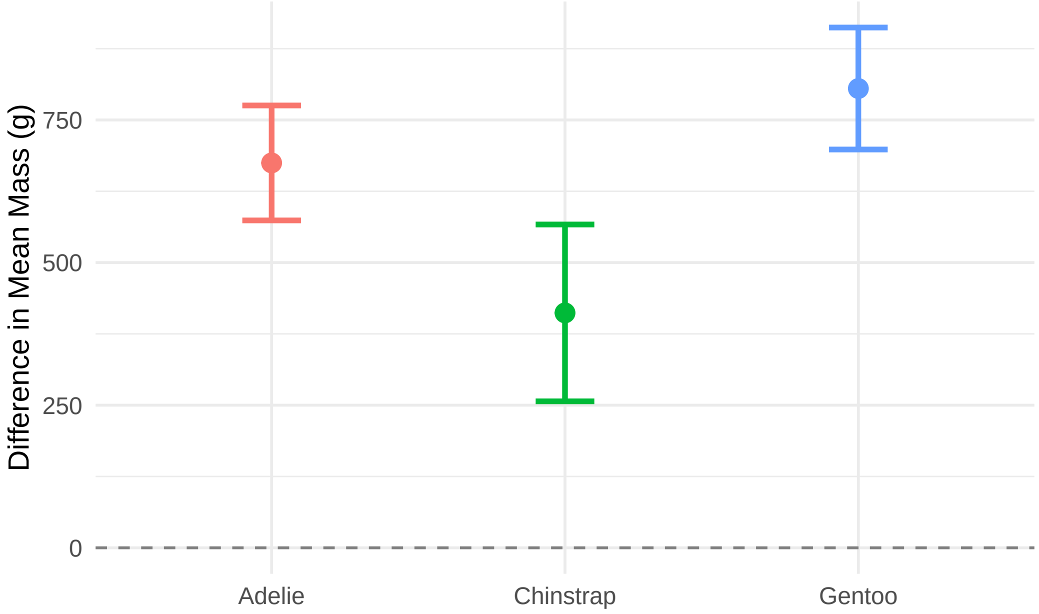

Figure 2: 95% Confidence Intervals for the difference in mean body mass (Male - Female). The dashed line at 0 represents the null hypothesis (no difference).

P-Value for Hypothesis Testing

To formally quantify the strength of evidence against the null hypothesis $H_0 : \mu_m - \mu_f = 0$, we calculate the p-value. The p-value represents the probability of observing a difference in means as extreme as, or more extreme than, the one observed in our sample, assuming the null hypothesis is true.The test statistic Z is calculated as:

$$Z = \frac{(\hat{\mu}_m - \hat{\mu}_f) - 0}{\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}}$$

Plugging in our data for the Gentoo penguins:

$$Z_{obs} = \frac{805.09}{54.54} \approx 14.76$$

This yields a p-value of:

$$p = P(|Z| > 14.76) < 0.001$$

Since $p < 0.05$, we reject the null hypothesis. In fact, the p-value is so small ($\approx 1.33 * 10^{-49}$), that we almost certainly did not incorrectly reject the null hypothesis. The probability of observing such a large difference in body mass by random chance alone is effectively zero.

Power of the Test

The power function of a hypothesis test, denoted as $\beta(\delta)$, is defined as the probability of rejecting the null hypothesis given a specific value of the true difference δ. High power is desirable when the true parameter lies in the alternative hypothesis region (i.e., $\delta \neq 0$).We calculate the power of our test to detect the observed difference ($\delta = 805.09$ g) at a significance level of $\alpha = 0.05$. The power is calculated as the probability that the test statistic falls into the rejection region determined by the null hypothesis:

$$\beta(\delta) = P\Big(Z > c_{\alpha/2} - \frac{\delta}{SE}\Big)$$

where $SE = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}$ is the standard error of the difference in means.

Substituting our values:

$$\beta(805.09) = P\Big(Z > 1.96 - \frac{805.09}{54.54}\Big) \approx 1$$

The power is approximately 1.0 (or 100%). This confirms that our sample size and the magnitude of sexual dimorphism were more than sufficient to detect the difference. We were virtually guaranteed to reject the null hypothesis given the true biological difference between male and female Gentoo penguins.

Discussion

It is clear that the conclusion we set out to prove is not only almost certainly true, but also not exclusive to the Gentoo species. Rather, it seems that all three species of penguin observed in the Palmer Archipelago exhibit what zoologists call “sexual dimorphism”, which is a significant difference in size between the sexes of an animal. While this is a universal trait across these penguins, the magnitude varies drastically. Gentoo penguins exhibit the most dimorphism, with males being approximately 800g heavier than females on average. Adelie penguins sit in the middle, with an average difference of about 650g. Chinstrap penguins show the smallest difference: males are only about 400g heavier than females. Importantly, the confidence interval for the mass difference in Chinstrap penguins is relatively lower and has no overlap with the interval for Gentoo penguins nor the interval for Adelie penguins.

This has important consequences. If a significant distinction between the species, these confidence intervals could suggest that biological drivers for size difference may be weaker in Chinstrap penguins compared to penguins of different descent. Moreover, the overlap between the confidence intervals for the Adelie and Gentoo intervals could suggest that, while Gentoo males are much bigger, the magnitude of the mass difference between the sexes is generally closer between these two species.

Our analysis could suggest many evolutionary and environmental hypotheses for the three species of Pygoscelis penguin. Perhaps there is a significant mating advantage in having a large difference in weight for Gentoo penguins – more so than for Chinstrap penguins but similarly as for Adelie penguins. This could be an indicator to similarity, or lack of similarity, in the origins of each the three species. Perhaps the mass difference is universally significant due to heavier males being able to better protect themselves from predators, so are able to produce more offspring in their lifetime, passing on their heaviness characteristic. Or perhaps there is a different reason for this completely; we know only the correlation, not the causation.

The application of robust statistical frameworks—specifically Maximum Likelihood Estimation (MLE), the Cramér-Rao Lower Bound, and high-power hypothesis testing—ensures our conclusions are grounded in rigorous evidence. While the dataset may harbor deeper correlations that warrant future investigation, the statistical proofs provided here establish a reliable confirmation of the hypothesis we set out to prove.

Critical Use of AI

Google Gemini was used to assist with the following tasks:

Installation and Initialisation of RStudio and Quarto: As I had never used these tools before, I used AI to guide the setup process and ensure the environment was configured correctly.

Writing Mathematics using LaTeX: I have rarely used LaTeX previously, so I used AI to assist with the syntax for the mathematical derivations and formulae.

Explanation of R Syntax: while my primary resource for learning R was the official documentation, I used AI to help explain specific syntax and translate certain concepts from Python into R.

Refining Written Communication: I used AI to polish the clarity and flow of a judicious few of my written explanations, ensuring that my arguments were presented as concisely and professionally as possible. The core analysis and conclusions remain entirely my own.

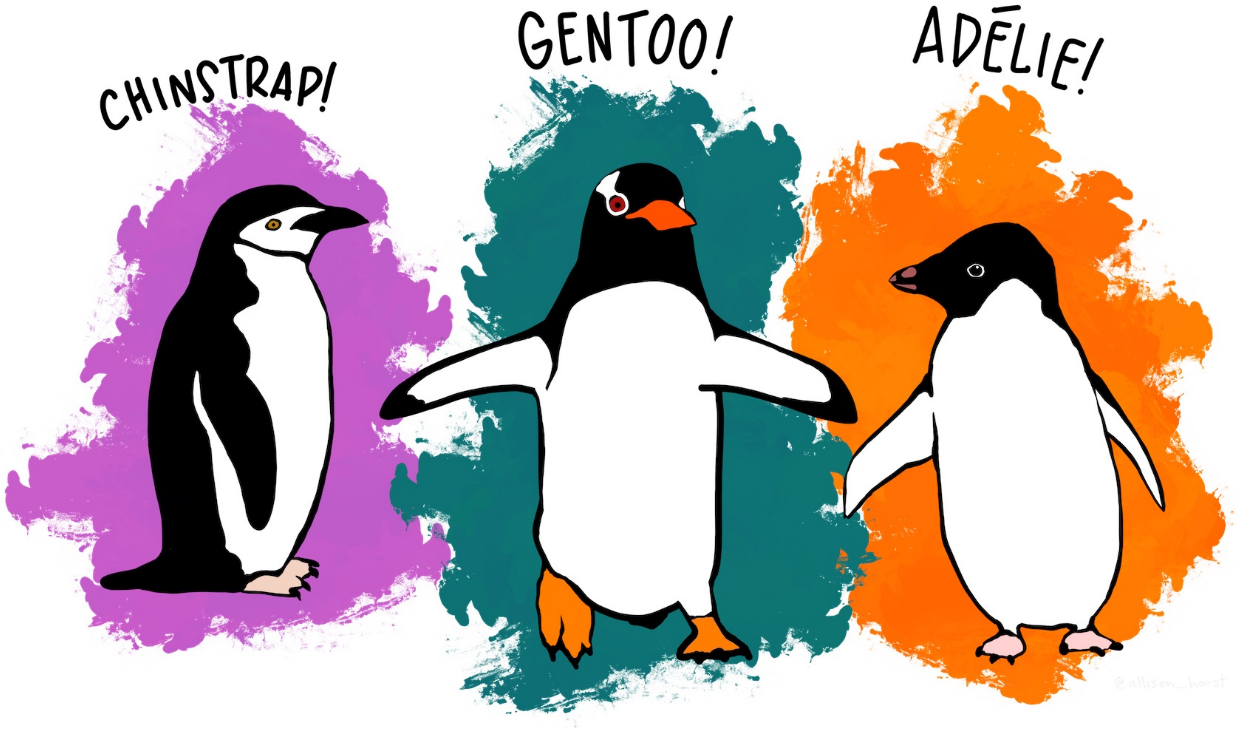

Figure 3: The Palmer Archipelago penguins. Artwork by @allison_horst.