

# Can Body Measurements Reliably Distinguish Penguin Species?

CID: 02385539

2026-02-09

## Why This Question

I stumbled upon the Palmer Penguins dataset (1, reference dataset as instructed to on their website) a few months ago while reading a book on Machine Learning (2, reference interpretable ML book). 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 plenty for calculating rough estimators of the different (continuous) features. Consequently, I selected it to be the focal point of this work.

When exploring the data, I discovered that a plot of the body mass distribution of Gentoo penguins (one of the three species in the dataset) was not a bell curve as expected, but rather a (bimodal) double-humped curve. 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?

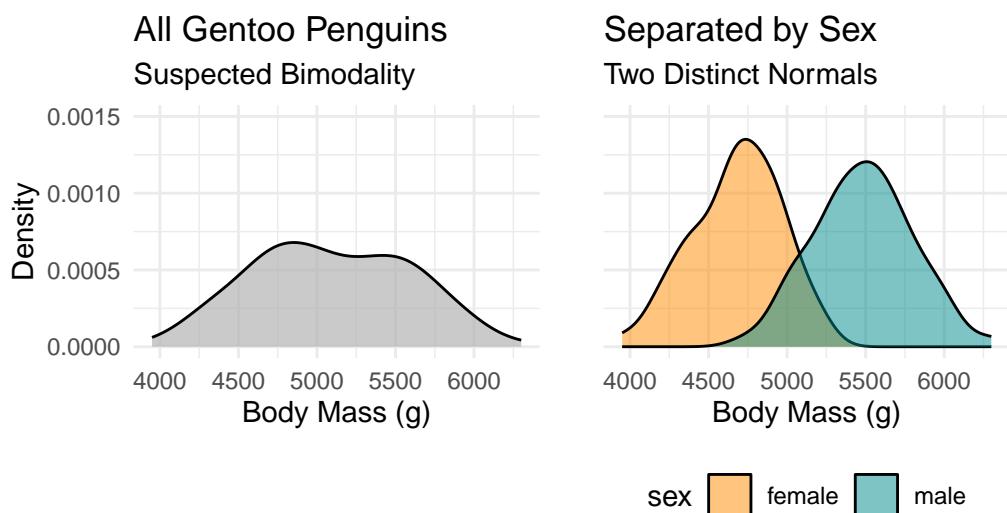


Figure 1: Left: The initial bimodal distribution. Right: The same data separated by sex. (Note: Y-axes fixed to same scale for comparison)

# Theoretical Framework

## Derivation of MLE

We choose to treat the body mass measurements  $x_1, \dots, x_n$  of a specific sex of Gentoo penguins as independent observations. We model these as:

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

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  $\mu$  and  $\sigma^2$ , equate to 0, and solve:

1. **For  $\mu$ :**

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

2. **For  $\sigma^2$ :**

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

Now, while the MLE for  $\hat{\mu}$  is unbiased, the MLE for  $\hat{\sigma}^2$  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 (CRLB).

**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  $\mu$ :

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

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

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  $\hat{\mu} = \bar{X}$ :

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

**Conclusion:** Since  $\text{Var}(\hat{\mu}) = \text{CRLB}$ , the estimator  $\hat{\mu}_{MLE}$  is **efficient** and attains the Cramér-Rao Lower Bound.

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

We repeat the process for the variance parameter  $\sigma^2$ .

**A. Fisher Information ( $I_n(\sigma^2)$ ):** First, differentiate the log-likelihood with respect to variance ( $\sigma^2$ ):

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

Second derivative:

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

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:

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

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

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

**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  $n \rightarrow \infty$ :

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

## Application & Results

These estimators for  $\hat{\mu}$  and  $\hat{\sigma}^2$  are as follows<sup>1</sup>:

**For Gentoo Males ( $n = 61$ ):**

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

**For Gentoo Females ( $n = 58$ ):**

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

**Male:**  $9.806831 \times 10^4$ , **Female:**  $7.928634 \times 10^4$

These unbiased variances are larger than the MLEs, demonstrating the existence of bias in the latter.

- **Hypothesis Test (Syllabus §7):** Perform a Two-Sample t-test (or Likelihood Ratio Test) to check if the difference in mean Culmen Depth between species is significant.
- **Confidence Region (Syllabus §6):** Plot the 95% Confidence Interval for your estimator.

---

<sup>1</sup>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  $n$ .

- Visuals:

## Critical Use of AI (0.5 Pages)

## Discussion (0.5)

## References



Figure 2: A solitary penguin heading towards the mountains (Source: Werner Herzog's 'Encounters at the End of the World')

## Data Analysis

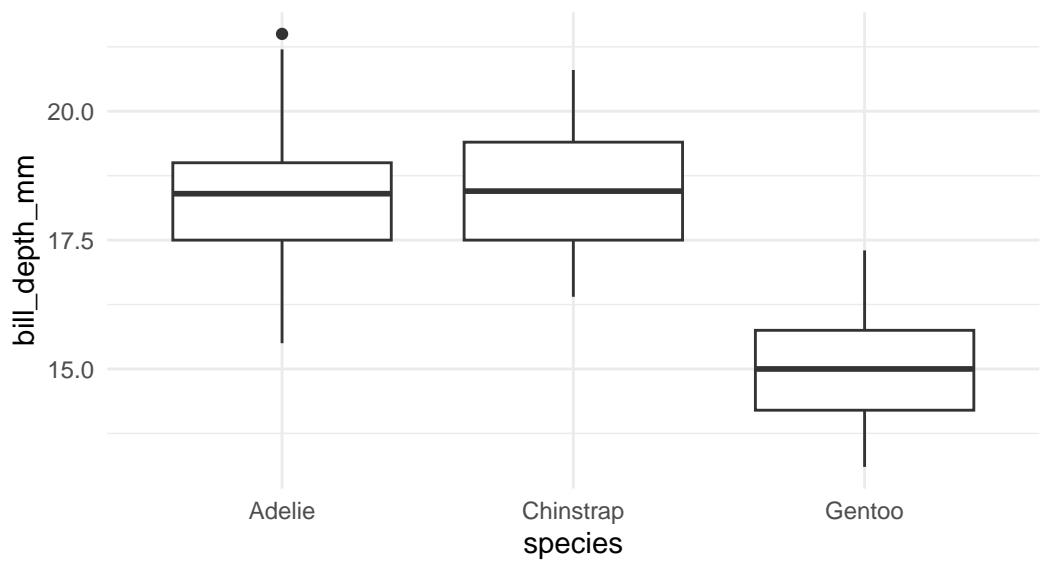


Figure 3: Distribution of Culmen Depth by Species