Bill the Statypus says: Buckle up! This worksheet is 3 blocks or 75 minutes long!

The Universal Rulers: The Architecture of Distance

If you take a traditional math class, you expect exact answers: if you solve an equation, x equals exactly 5. In manufacturing, we try to force this exactness onto the physical world. A machine is built to cut exactly 10.00mm bolts. Any variation between two bolts is considered an "error" or a "defect."

But statistics is the mathematics of the real world, and in biology, exactness is an illusion. Variability is not a mistake; it is the engine of survival. If every single platypus on a given river weighed exactly 1.4 kg, we wouldn't need biologists—we would need an investigator to figure out who was cloning them!

Sally the Statypus says: Every animal and every measurement is unique. The "noise" in the data is actually the most important part of the story. If we only calculate the center, we are flying blind. We must learn to measure the spread and that **understanding variation is the heart of statistics.**

1. The Raw Data & The Global Span

Below is a raw, unsorted field log containing the weights (kg) of 16 platypuses captured along the Snowy River.

`1.5, 0.9, 2.1, 1.1, 1.4, 1.4, 0.8, 2.4, 1.0, 1.2, 1.0, 1.7, 1.2, 2.3, 1.3, 1.1`

Task A: Order the Data
Statistics begins with order. Sort the 16 weights from lightest to heaviest in the boxes below.

Task B: Range (Textbook Section 4.2.1)
Now that the data is sorted, we can immediately identify the boundaries of the population. Note that we use the word "Range" to describe two different concepts. Circle the smallest value and label it "min" and then put a triangle around the biggest value and label it "Max."

1. **Range Interval (The Territory):** The span between the min and Max. Record the interval:

2. **Range Length (The Distance):**

Sally the Statypus says: We often blur the line and just say "Range" to mean either the length or the interval allowing context to make things clear. But look closely at your sorted data. If that massive 2.4kg "Megastatypus" at the end hadn't been caught, how much would the Range Length shrink? Is the Range a sturdy ruler, or is it too fragile for serious science?

2. The Anchor: Construction of the Mean

Go to the **Intro to Section 4.1**. The textbook explains that the Mean is the only number that balances the data perfectly. It is constructed so that the sum of the deviations $(x - \bar{x})$ is **exactly zero**.

Task C: Find the Center
Calculate the exact Mean $(\bar{x})$ of your 16 platypuses. You may add them by hand and divide by 16, or type the values into an R console.

The Sample Mean $(\bar{x})$ is: _____

Task D: The Zero-Sum Proof
We are going to prove the balance point. Using your sorted list from Page 1, calculate the deviation $(x - \bar{x})$ for the first **15** animals and add them all together. (Hint: Keep track of your negatives!)

1. What is the **sum of the deviations** for the first 15 animals?

   _____

2. Look at your 16th animal (the 2.4kg Max). What is its single deviation from the mean?

   _____

Bill the Statypus says: Look at those two numbers! Once the first 15 animals were weighed, the final animal was a hostage of the math. It HAD to perfectly balance out the rest of the group to ensure the sum reached zero.
Because the last animal has no freedom to be anything else, we say a sample of size n only has $n-1$ **Degrees of Freedom**. This is exactly why we divide by $n - 1$ when calculating our "Standard" deviation—we only have $n - 1$ pieces of truly free information!

3. Team Mean's Ruler: Standard Deviation

Because deviations always sum to zero, we cannot simply average them. Look at **Example 4.7** in the textbook to see the pseudo-manual calculation to fix this.

We square the deviations to make them positive, average them (dividing by $n-1$), and get the **Variance**. However, Variance has a major problem: **Units**.

Statypus Insight: Escaping Squared Space

If we measure weights in kilograms (kg), the Variance is measured in kg^2. What is a "squared kilogram" of platypus? It makes no physical sense! We take the square root of the Variance to return to our original units. This final number is the **Standard Deviation**.

Coding Corner: The Standard Function

We rarely calculate this by hand. In R, the function to find the Standard Deviation is `sd()`.

Bill the Statypus says: Let's stop playing around. I put this data into the Chapter 4 .RData file. Head to r.statypus.org to load that and then run `sd( PlatyPlayData )` and just record your value in the margins. I don't feel like making a silly little box to put your answer in.

Task E: Standard Units

If the standard deviation is our ruler for the data, we can use it to measure how extreme the largest platypus in our data. How many "SD Rulers" away is our 2.4kg Megastatypus?

4. Notation Audit

Biologists must distinguish between a "Sample Estimate" and the "Population Truth." There is a second version of the standard deviation. See Section 4.2.5 at r.statypus.org for more info.

Concept	**Sample Estimate (Statistic)**	**Population Truth (Parameter)**
Notation	s	σ (Sigma)
Denominator	$n-1$ (Degrees of Freedom)	N (Population Size)

Sally the Statypus says: If you are curious why σ is found by dividing by N while s uses $n-1$, it is because the population contains every animal in existence. It has no missing values—it is the whole story. The use of $n-1$ appeared as a consequence of the sum of deviations, but it's effect is to account for the fact that a sample isn't the whole story.

5. Quartiles: The Cheese Cleaver

Turn to Section 4.2.2 and then read**Definition 4.6** and the **Remark** that follows.

Bill the Statypus says: Let's not overcomplicate this! The Median cuts the data into 2 equal pieces. Quartiles cut data into 4 equal pieces.

Statypus Insight: The Cleaver not the Cheese

Midnight is like a cleaver slicing a block of cheese. It isn't part of the cheese, but it cuts it into two pieces. If we are looking at individual data values and not cheese, the cleaver (quartile) may hit a value exactly. In that case, that value is the quartile. If not, the space between values that the cleaver finds is where our quartile lives, and there is some ambiguity in how we label it.

Task F: Partitioning the Catchment ($n = 16$)
Below is your sorted log of 16 weights. Use your pencil to draw three physical "Cleaver Slices" that divide the animals into **four equal groups** of 4 animals each.

`0.8 0.9 1.0 1.0 1.1 1.1 1.2 1.2 1.3 1.4 1.4 1.5 1.7 2.1 2.3 2.4`

Because we have exactly 16 animals, your cleaver slices fell into the "ambiguous space" between values every single time. In this course, we simply average the two values on either side of the slice.

1. **Cut 1 (Q_2 / Median):** Falls between _____ and _____. Result: _____

2. **Cut 2 (Q_1):** Falls between _____ and _____. Result: _____

3. **Cut 3 (Q_3):** Falls between _____ and _____. Result: _____

Bill the Statypus says: Slow down there, bucko! In *real* statistics, we don't just *average* the values to find the quartiles. There are many different conventions for this. All reasonable statisticians agree that that is ok for the median, but we argue about how to handle Q_1 and Q_3 in some cases.

Sally the Statypus says: (SIGH) Bill is right... in fact, R has numerous ways of handing the quartiles and will sometimes give slightly different values based of the function you use to find it. You can ask AI to explain it further if you want, my head hurts... I think I need some of Bill's tea.

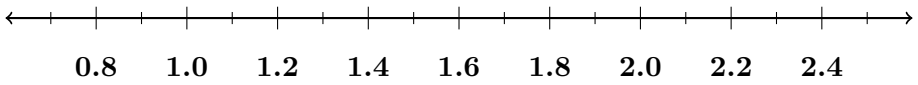
Statypus Insight: Reality Check

Let's face it, learning can be difficult at times. Even our lead statypi need to take a break now and then. Never forget that being confused is part of the process of learning and not a sign that you are failing.

6. Visualizing the Fences

Task G: The Manual Dot Plot

Plot your 16 platypus weights on the number line below. Place a large dot for each animal. **If a weight appears more than once, stack the dots vertically.**



7. Team Median's Ruler: The IQR

Task H: On the number line above: Do the following:

- Label the minimum, Maximum, median, mean, Q_1, and Q_3.
- Draw a line indicating the **Range** below the number line.
- Draw a line below the range indicating the **Interquartile Range (IQR)** by connecting Q_1 to Q_3.
- Draw a line below the IQR going from one standard deviation to the left and mean to one standard deviation to the right of the mean.

Sally the Statypus says: Just like Range, we use the term IQR to describe an Interval or a Length depending on the context.

1. **IQR Interval (The Middle 50% Territory):** $[Q_1, Q_3]$. Record it:

2. **IQR Length (The Sturdy Distance):** $Q_3 - Q_1$. Calculate it:

Reflection: The Ruler of Steel

Look at your dot plot. If the 2.4kg animal was actually a 5.0kg mutant, the Range bracket would stretch completely off the page. Would your IQR bracket move *at all*? Why is the IQR considered the best ruler for skewed data?

8. Scaling Up with R: Scavenger Hunt

In statistics, we use summary "packages." You must pair the center with the correct ruler.

- **The Mathematicals (Symmetric Data):** Mean ($\bar{x}$) & Standard Deviation (s).
- **The Resistants (Skewed Data):** Median (Q_2) & Interquartile Range (IQR).

Coding Corner: The Efficiency Switch

Most R functions like `sd()`, `IQR()`, and `range()` will fail if there is even one missing value. Look back at Section 4.1. What is the argument (the "switch") that tells R to ignore NAs?

Bill the Statypus says: Want to see a magic trick? Run `summary( PlatypusData2$WeightF )` in your console. Notice that you don't need the switch! The summary function is built to handle NAs automatically and report them as a separate count, while simultaneously giving you the Min, Max, Mean, Median, Q_1, and Q_3.

9. Final Replication Report

Reflection: Final Regional Audit

Use R to find the final truth for the **Snowy River Catchment** ($N = 259$). Record the results for the `WeightF` variable:

- **Regional Range Length:**
- **Regional IQR Length:**
- **Regional Sample SD (s):**

Final Conclusion

If you are writing a report for a biologist and the data is **heavily skewed** with several massive outliers, which team (The Mathematicals or The Resistants) provides the most honest summary of the population? Why?