PHYS 100 EXPERIMENT 3 (week 4)
Quantifying Distributions

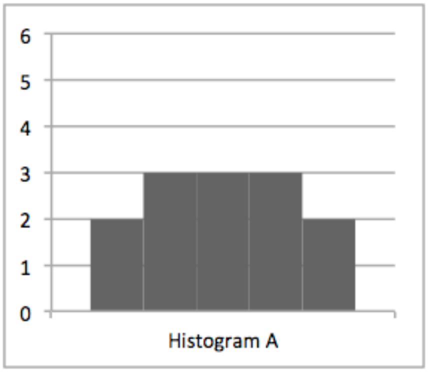
Name: ______________________ Student #: ____________ Section: ____ Date: __________

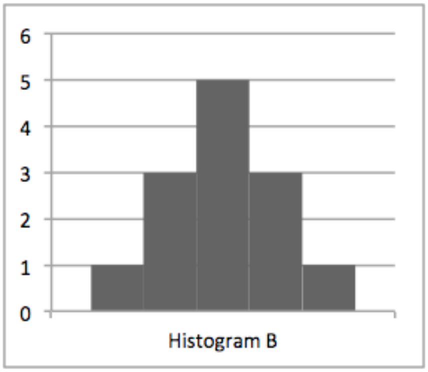
Outline:
- Introduction (5m)
- Comparisons (10m)
- Solution examples (25m)
- Common solution (20m)
- Practice (15m)

Topic

Measuring distributions of data

Introduction (5 min, entire class)
One important factor that affects uncertainty is *distribution* of data. For example, Histogram A has a wider distribution than Histogram B (and Histogram B has a narrower distribution).





So far we learned two ways to analyze distributions:

1. Eyeballing the measurements.
2. Comparing histograms.

Today we will add another tool to our toolbox, as we will learn about mathematical ways to quantify distributions, that is, the spread (or scatter) of the data

☞**Clicker 1:** Jack and Jill measured their walking speed multiple times going uphill. Jack got 1.3 m/s and Jill got 1.4 m/s on average. However, they did not report their distributions. If they were to compete again, who would be faster? Would you bet $5 on one of them?

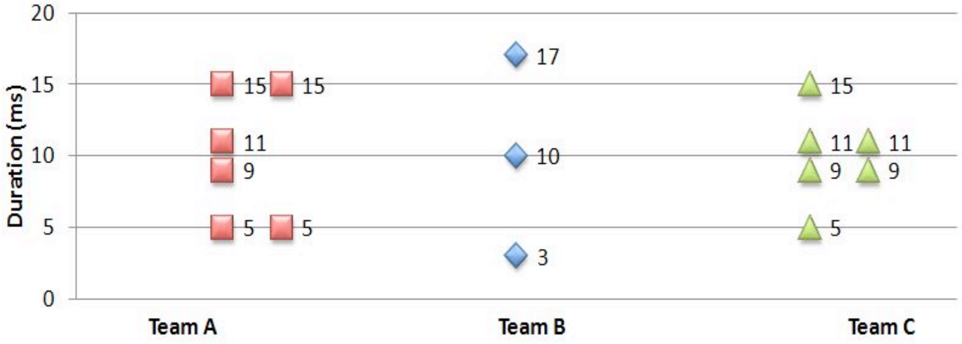
A. Jill is faster, as her average is lower.
B. They were equally fast, as their averages are rather similar.
C. Impossible to determine without the distributions of their results.

Tasks

Task 1: Comparing the distributions of data (10 min, individual)

Three teams of scientists measured the time it takes a certain beam of particles to travel a given distance. Here are their results, reported in milliseconds (ms) in random order:

Team A:	15	9	5	11	5	15
Team B:	3	17	10			
Team C:	5	11	9	11	15	9



For each of the following pairs, determine qualitatively (without calculations) which dataset has a narrower distribution. Explain your choice.

- Team A vs. Team B:

  Why?

- Team B vs. Team C:

  Why?

- Team A vs. Team C:

  Why?

Once you are done, compare your answers *and your explanations* to those of your neighbour. If there are disagreements, resolve them.

Task 2: Examples for Methods to Quantify Distribution (25 min, pairs)

Below you can see how a group of students from last year developed a method to quantify the distribution. The students tried several different solution approaches shown on the following pages. Their solutions reflect good ideas, however they are incomplete.

Your task is to explain how each solution works and where it may fail. Use your comparisons from task 1 to help you.

Solution idea 1:

Mari: *Essentially, distribution is a function of the distances between the data points. I think that narrow distribution means that the data is less spread out. For each Team let's calculate the distance between the maximum and the minimum:*

	A	B	C
1) Calculate the distance from the maximum to the minimum	15-5 = 10	17-3 = 14	15-5 = 10

Jeff: *Okay… the maximum and the minimum are equal for Teams A and C. This means that Teams A and C have similar distributions. For Team B the result is bigger, so its distribution is wider.*

Does the method work? If not, please explain where this method fails.

Solution Idea 2:

Mari: *This doesn't make sense. We can see that the distribution of Team C is narrower than Team A, but our method gave the same value to both. How can we improve it?*

Jeff: *A narrow distribution means that more numbers are at the mean or close to the mean. Let's calculate deviations from the mean.*

Mari: *Why the mean?*

Jeff: *Well, we need to account for all the distances between all the points, but that is too much to calculate. Instead, we can calculate the distances to a number that represents all the points, such as the mean. Using the mean allows us to determine the distance to the central tendency of the data.*

	A	B	C
1) Calculate the mean	(15+9+5+11+5+15)/6 = 10	(3+17+10)/3 = 10	(5+11+9+11+15+9)/6 = 10
2) Subtract the mean from each value	15-10 = 5 9-10 = -1 5-10 = -5 11-10 = 1 5-10 = -5 15-10 = 5	3-10 = -7 17-10 = 7 10-10 = 0	5-10 = -5 11-10 = 1 9-10 = -1 11-10 = 1 15-10 = 5 9-10 = -1
3) Add up	5-1-5+1-5+5 = 0	-7+7+0 = 0	-5+1-1+1+5-1 = 0

Mari: *Meh, this gives us zeros for all teams. Thus, according to this method Team A, B, and C all have the same distribution – and it equals zero…*

Does the method work? If not, please explain where this method fails.

Solution 3:

Jeff: *Ouch, the distances cancel each other out. I think that distributions depend on the distances between the points, regardless of whether the data points are above or below the mean. We somehow need to use only positive values. How about using squared values?*

	A	B	C
1) Calculate the mean	(15+9+5+11+5+15) /6 = 10	(3+17+10)/3 = 10	(5+11+9+11+15+9) /6 = 10
2) Subtract the mean from each value	15-10 = 5 9-10 = -1 5-10 = -5 11-10 = 1 5-10 = -5 15-10 = 5	3-10 = -7 17-10 = 7 10-10 = 0	5-10 = -5 11-10 = 1 9-10 = -1 11-10 = 1 15-10 = 5 9-10 = -1
3) Square the distances	$5^2 = 25$ $(-1)^2 = 1$ $(-5)^2 = 25$ $1^2 = 1$ $(-5)^2 = 25$ $5^2 = 25$	$(-7)^2 = 49$ $7^2 = 49$ $0^2 = 0$	$(-5)^2 = 25$ $1^2 = 1$ $(-1)^2 = 1$ $1^2 = 1$ $5^2 = 25$ $(-1)^2 = 1$
4) Add up	25+1+25+1+25+25 = 102	49+49+0 = 98	25+1+1+1+25+1 = 54

Mari: *Team A has the largest value, so their measurements have the widest distribution. Team C has the smallest value, so their measurements have the narrowest distribution.*

Does the method work? If not, please explain where this method fails.

Solution 4:

Mari: *Better, but still not perfect. It is good that C is the narrowest – but what about A and B? I thought that B was wider than A!*

Jeff: *Also, the more I think about it, the more I realize that units are wrong. Our units now are time^2 because we squared the distances*

Mari: *Perhaps we can take the square root of the expression we calculated earlier. This will bring us back to units of time. Let's add to the previous attempt:*

	A	B	C
5) Take the square root	*√102 = 10.1*	*√98 = 9.9*	*√54 = 7.3*

Mari: *Well, even if the units are correct now, we still get the same ranking for the three teams.*

Does the method work? If not, please explain where this method fails.

Task 3: A common approach to quantifying distribution (20 min, pairs)

One method that experts use to evaluate distribution is Standard Deviation, or SD. This is how standard deviation is calculated:

$SD = \sqrt{\frac{\sum(x_i - \overline{x})^2}{N}}$	x_i – each of the data points $\overline{x}$ – the mean of all the x_i's N – the number of data points.

In words, SD is $\sqrt{\frac{sum\ of\ all\ (distances\ from\ the\ mean)^2}{number\ of\ points}}$

3.1 Read the rationale below and explain it to each other. Make sure you understand the formula.

Rationale for the formula:

Essentially, distribution is a function of the distances between the data points. Thus, a narrower distribution corresponds to shorter distances between data points.

Why the mean?
To account for the distances between all the points, we can calculate the distances to a number that represents all the points. The **mean** does that, as it depends on all the points. It allows us to determine the distance to the central tendency of the data.

Why squaring?
The distribution depends on the distances between the points, regardless of whether the data points are above or below the mean. Therefore we need to use only **positive values**. One method to obtain positive values is taking squared values.

Why sum of all?
We sum all squared distances to take **all data points** into account.

Why N?
We divide by N to compensate for differences in **sample size**.

Why square-root?
Remember that we squared the distances. Thus, the units now are time^2. **To fix the units** we take a square root of the entire expression. Notice that N has no units.

Below is an example for applying SD, $\sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$

Let us break it down into steps, using the data set you have evaluated in the beginning:

3.1 Explain to each other *in your own words* how each step is useful. You can use the given rationale to help you. Then, write your explanations below.

Team A: 15 9 5 11 5 15

1) Calculate the mean
(15+9+5+11+5+15)/6 = 10
Why is this step useful?

2) Calculate the distance from each data point to the mean
15-10 = 5
9-10 = -1
5-10 = -5
11-10 = 1
5-10 = -5
15-10 = 5
Why is this step useful?

3) Square the results
$5^2 = 25$
$(-1)^2 = 1$
$(-5)^2 = 25$
$1^2 = 1$
$(-5)^2 = 25$
$5^2 = 25$
Why is this step useful?

4) Add up
25+1+25+1+25+25 = 102
Why is this step useful?

5) Divide the sum by the number of data points
102/6 = 17
Why is this step useful?

6) Take the square root (and add units)
$\sqrt{17} = 4.12$ms
Why is this step useful?

Now let's look at the data of the other teams. Below you can see how SD is applied to their data. However, several steps in the solutions are missing. Please fill in the blanks.

Team B: 3 17 10

1) Calculate the mean
   (3+17+10)/3 = 10

2) Calculate the distance from each data point to the mean
   3-10 = -7
   17-10 = 7
   10-10 = 0

3) Square the results
   $(-7)^2 = 49$
   $7^2 = 49$
   10-10 = 0

Fill in the missing steps:

4) Add up

5) Divide the sum by the number of data points

6) Take the square root (and add units)

Team C: 5 11 9 11 15 9

1) Calculate the mean
(5+11+9+11+15+9)/6 = 10

Fill in the missing steps:

2) Calculate the distance from each data point to the mean

3) Square the results

4) Add up

5) Divide the sum by the number of data points

6) Take the square root (and add units)

Now you have the SD for all three teams. Please copy the results for the three teams here:

Team A: Team B: Team C:

In each of the following pairs, which dataset has a narrower distribution according to the calculated values?

- Team A vs. Team B:

- Team B vs. Team C:

- Team A vs. Team C:

Do this method and your initial observations agree? If not, why not?

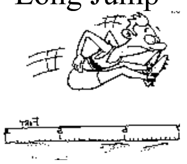
Task 3: Practice (10 min, pairs)

1. What is the SD of the following data sets? You can leave the answer without calculating the square root, but be sure to indicate the values that should be in the root (√*x).*

   Data set A: 2m 2m 5m 7m

   Data set B: 5sec 4sec 6sec 5sec

2. (bonus question) An equal number of students competed in high jump, long jump, and pole vault. The achievements of the champions in the three competitions are shown below, as well as the mean and SD of all finalists .

	High jump	Long Jump	Pole Vault
			
Champion:	130cm	440cm	308cm
Mean of the finalists:	120cm	410cm	305cm
SD of the Finalists:	5cm	20cm	4cm

Relative to the other finalists, who is the best athlete among the three champions? Who is the worst? Invent a method that can be used to compare the champions between the three competitions. The method should assign a value to each champion for how good his or her performance was relative to the other finalists in his or her competition. Explain your calculations.

PHYS 100 Homework 3 (for week 4)
Pendulum (can do in pairs, but hand in individually)

Name: __________________ Student #: _____________ Section: ___

In this homework you will use the methods you have learned so far to answer the following scientific question:
Do either mass or length affect the oscillation time of a pendulum?

Design and execute an experiment to test **one** variable (length or mass). Test only two set-ups of a pendulum (that is, test only two lengths or masses). Obviously, take multiple measurements in each set-up

Submit using both sides of this page **only**:

1. A description and a diagram of the apparatus you used. What was your pendulum? How did you measure oscillation time?
2. What did you vary between measurements. How did you test the effect of the factor you focused on?
3. Your raw data. Keep a copy of your results and bring it with you to the next lab.
4. A histogram of the data. You can either use one histogram with multiple data sets, or multiple histograms.
5. Calculate the mean and SD of each experimental set-up. Report your values and, on the histogram, mark three values: mean; mean+SD; mean-SD. If you only have these three numbers, what can you infer about the histogram?
6. Did your factor affect the oscillation time?

Rubric

	Sufficient (✓)	Lacking (✓-)	Insufficient (x)
Experimental design is valid			
Sufficient high-quality data was collected			
Histograms are correct and support comparison between set-ups			
SD is calculated correctly, and its meaning is well understood			

Clicker Resource Page

☞**Clicker 1:** Jack and Jill measured their walking speed multiple times going uphill. Jack got 1.3 m/s and Jill got 1.4 m/s on average. However, they did not report their distributions. If they were to compete again, who would be faster? Would you bet $5 on one of them?

A. Jill is faster, as her average is lower.
B. They were equally fast, as their averages are rather similar.
Impossible to determine without the distributions of their results.