BIOL 220 Problem Set 03: Describing data
Answer Key

💡 Multiple choice questions

1. Which of the following accurately describes the numerator for the equation to calculate the variance? [1 point]

   a. The square root of the standard deviation
   b. The sum of the average deviations between the data values and the mean
   c. The sum of the squared deviations between the data values and the mean
   d. The average of all the data values

ℹ️ **Answer**

   c. The sum of the squared deviations between the data values and the mean

2. Using the data set shown below, what are the correct mean and median values for this data set [1 point]?

   Data set: 3, 4, 5, 5, 8, 9, 11, 11

   a. Mean = 6.5, median = 7
   b. Mean = 7.5, median = 6
   c. Mean = 6, median = 7.5
   d. Mean = 7, median = 6.5

ℹ️ **Answer**

   d. Mean = 7, median = 6.5

3. The standard deviation has the same units as the data values, whereas the variance does not. [1 point]

☐ TRUE
☐ FALSE

ℹ **Answer**

TRUE

Gliding snake undulation rate

We saw the flying snake undulation data in lecture, they come from our WS textbook. The data are reproduced below:

`0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0` Hz

These data are expressed in Hz which has units of 1/s (cycles per second). Frequency measurements are often expressed as angular velocity which is measured in rad/s (radians per second). Convert from Hz to angular velocity by multiplying by 2π, then use the *transformed* data to answer the following questions

💡 Use the transformed data to calculate the sample mean, standard deviation, and variance.

For your answers be sure:

- report values rounded to one significant digit greater than as reported in the data
- provide the **correct units**, using the ^ symbol for raising to a power. So if your answers were 2 square meters, you would write "2 m^2"

4. What is the transformed sample mean? [1 point]

> **ⓘ Answer**
>
> In R code:
>
> ```
> xHz <- c(0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0)
> xrps <- xHz * 2 * pi
>
> ybar <- mean(xrps)
> ybar
> ```
>
> [1] 8.63938
>
> You should report: 8.64 radians/second

5. What is the transformed sample variance? [1 point]

> **ⓘ Answer**
>
> In R code:
>
> ```
> s2 <- var(xrps)
> s2
> ```
>
> [1] 4.145234
>
> You should report 4.15 (radians/second)^2

6. What is the transformed sample standard deviation? [1 point]
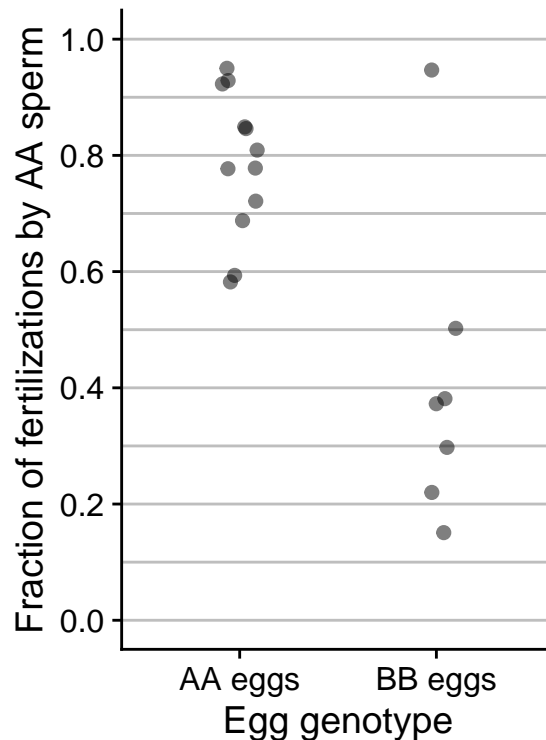
> **ⓘ Answer**
>
> In R code:
>
> ```
> s <- sd(xrps)
> s
> ```
>
> [1] 2.035985
>
> You should report 2.04 radians/second

Sea urchin reproduction

Reproduction in sea urchins involves the release of sperm and eggs in the open ocean. Fertilization begins when a sperm bumps into an egg and the sperm protein *bindin* attaches to the recognition sites on the egg surface. Gene sequences of *bindin* and egg-surface proteins vary greatly between closely related urchin species, thus reducing the chance of hybridization between species. In the burrowing sea urchin, *Echinometra mathaei*, the protein sequence for *bindin* varies even between populations within the same species. Do these differences affect fertilization? To test this, Palumbi (1999) carried out trials in which a mixture of sperm from AA and BB genotype males were added to dishes containing eggs of either the AA or the BB genotype. The results below indicate the proportion of successfully fertilized eggs (of both AA and BB genotypes) that were fertilized by AA genotype sperm. Note: we do not need to look at fertilization by BB sperm because by definition, the fraction will be 1 - *fraction fertilized by AA sperm.*

💡 Use the figure above to answer Questions 7–8.

7. Examine the graph and visually determine the approximate value of the median for each egg genotype. Round your answer to two decimal places. Your visual estimates should be within 0.05 of the true values. [1 point]

8. Inspect the plot. Based on how the data in each egg genotype group are distributed across the y-axis, answer the follow:

   a. Which method would be best to compare the spread of the fertilization data for each egg genotype group? Be sure to justify your answer [1 point]
   b. Use the raw data (shown below) to calculate the spread for each egg genotype group using the method you select in (8a.) [2 points]

You can copy the below code and run it in R to make two vectors: one with the data on fertilization proportions for eggs of AA genotype, the other for eggs of BB genotype:

```
eggsAA <- c(0.58, 0.59, 0.69, 0.72, 0.78, 0.78, 0.81,
            0.85, 0.85, 0.92, 0.93, 0.95)

eggsBB <- c(0.15, 0.22, 0.3, 0.37, 0.38, 0.5, 0.95)
```

> **i Answer**
>
> 8a. The best method would be the interquartile range because of the outlier for the BB genotype eggs. However, an argument could be made that one outlier in otherwise symmetric looking data is not a substantial concern. If no mention of the outlier is made, or if a different justification is used for picking the interquartile range method, full credit cannot be given

> **i Answer**
>
> 8b. See answers below:

```
eggsAA <- c(0.58, 0.59, 0.69, 0.72, 0.78, 0.78, 0.81,
            0.85, 0.85, 0.92, 0.93, 0.95)
```

```
eggsBB <- c(0.15, 0.22, 0.3, 0.37, 0.38, 0.5, 0.95)

# interquartile using the `quantile` function
iqrAA <- quantile(eggsAA, probs = c(0.25, 0.75), type = 2)
iqrBB <- quantile(eggsBB, probs = c(0.25, 0.75), type = 2)
```

Note: either the lower and upper bound can be given:

```
iqrAA
```

```
  25%    75%
0.705 0.885
```

```
iqrBB
```

```
 25%  75%
0.22 0.50
```

or the difference can be given:

```
diff(iqrAA)
```

```
 75%
0.18
```

```
diff(iqrBB)
```

```
 75%
0.28
```

There might be some confusion about how to use the quantile function. By default it will return a slightly different answer. We will accept that answer for full credit as well:

```
# interquartile using the `quantile` function default type
iqrAA <- quantile(eggsAA, probs = c(0.25, 0.75))
iqrBB <- quantile(eggsBB, probs = c(0.25, 0.75))
```

```
iqrAA
```

```
   25%    75%
0.7125 0.8675
```

```
iqrBB
```

```
 25%  75%
0.26 0.44
```

```
diff(iqrAA)
```

```
  75%
0.155
```

```
diff(iqrBB)
```

```
 75%
0.18
```

If the answer for (8a.) was "standard deviation" then those values should be:

```
sd(eggsAA)
```

```
[1] 0.1238859
```

```
sd(eggsBB)
```

```
[1] 0.2639444
```