# Assignment 8: Statistics
# Due November 21, 2023

John McCuan

Recall the measurements in inches for the length of Milkyway minis from Assignment 7:

```
0.88 1.04 0.77 1.11 0.88 1.04 0.89 0.90 0.97 0.90
1.14 0.96 1.10 1.02 0.94 0.75 0.82 0.83 0.94 1.02
0.90 0.62 0.89 0.88 0.81 1.01 0.82 0.67 0.96 0.89
0.90 0.92 1.02 0.79 0.96 0.80 0.91 0.99 0.72 1.02
0.83 0.90 0.95 1.03 0.78 0.79 0.94 0.85 0.86 0.98
```

**Problem 1** (sample mean and sample variance) The **sample mean** of a collection of continuous type data $\ell_1, \ell_2, \ldots, \ell_n$ like the Milkyway minis lengths, is just the avarage of the values:

$$\bar{\ell} = \frac{1}{n}\sum_{j=1}^{n} \ell_j.$$

The **sample variance** is defined to be the number

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (\ell_j - \bar{\ell})^2.$$

The **sample standard deviation** is defined to be $s = \sqrt{s^2}$.

**(a)** Find the sample mean of the Milkyway mini data.

**(b)** Find the sample variance of the Milkyway mini data.

**(c)** Find the sample standard deviation of the Milkyway mini data.

**Problem 2** (sample variance versus emperical variance) The number

$$\nu^2 = \frac{1}{n}\sum_{j=1}^{n}(\ell_j - \bar{\ell})^2$$

is called the **emperical** (distribution) **variance** associated with the continuous type sample data $\ell_1, \ldots, \ell_n$.

**(a)** Simulate Milkyway minis lengths using the command

`rnorm(50, mean =` $\bar{\ell}$`, sd = s)`

in R.

**(b)** Simulate Milkyway minis lengths using the command

`rnorm(50, mean =` $\bar{\ell}$`, sd =` $\nu$`)`

where $\nu = \sqrt{\nu^2}$ in R.

**(c)** Does one value give a better simulation than the other? What is the basic difference between the **sample standard deviation** $s$ and the **emperical deviation** $\nu$? Why do you think one might be used instead of the other?

**Problem 3** (sample variance) Both the sample variance and the emperical variance involve the quantity[1]

$$\sum_{j=1}^{n}(\ell_j - \bar{\ell})^2.$$

**(a)** Show

$$\sum_{j=1}^{n}(\ell_j - \bar{\ell})^2 = \sum_{j=1}^{n}\ell_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n}\ell_j\right)^2.$$

**(b)** Show

$$\nu^2 = -\bar{\ell}^2 + \frac{1}{n}\sum_{j=1}^{n}\ell_j^2.$$

[1]The alternative expressions in this problem can be easier to calcualate (by hand) and also can incur less round off error when calculated using a computer.

**(c)** Show

$$s^2 = \frac{1}{n-1}\left[-\frac{1}{n}\left(\sum_{j=1}^{n}\ell_j\right)^2 + \sum_{j=1}^{n}\ell_j^2\right] = \frac{1}{n-1}\left[-\frac{1}{n}\left(\sum_{j=1}^{n}\ell_j\right)^2 + n(\nu^2+\overline{\ell}^2)\right].$$

**Problem 4** (histogram) A data histogram is not always constructed using a stem-and-leaf display as presented in Assignment 7. More generally, one can specify values $c_0 < c_1 < \cdots < c_k$ determining **class intervals**

$$(c_0, c_1], (c_1, c_2], \ldots, (c_{k-1}, c_k]$$

whose union contains all the data. The numbers $c_0, \ldots, c_k$ are called **class boundaries** and the midpoints

$$a_j = \frac{c_j + c_{j-1}}{2}, \qquad j = 1, 2, \ldots, k$$

are called **class marks**. With each class is associated a frequency $f_j$, $j = 1, \ldots, k$ which may be plotted above the class mark or as the height in a bar graph over the class itself. The result is a generalization of the histogram(s) consideed in Assignment 7.

Consider the following[2] ordered data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 5 |
| 5 | 5 | 5 | 6 | 6 | 6 | 6 | 8 | 8 | 9 |
| 15 | 17 | 22 | 23 | 24 | 24 | 25 | 27 | 32 | 43 |

**(a)** Create/draw a histogram with class boundaries $1.5, 2.5, 6.5, 29.5, 49.5$.

**(b)** Is the length of each frequency proportional to the height of the histogram class in the sense that for some fixed $\alpha \in \mathbb{R}$

$$f_j = \alpha(c_j - c_{j-1}), \qquad j = 1, 2, \ldots, k \ ?$$

**(c)** Create/draw a histogram for the relative frequencies.

[2]This "colorful" data comes from the text *Probability and Statistical Inference* by Hogg, Tanis, and Zimmerman. It is supposed to represent the costs of property losses in millions of dollars.

**(d)** Find a formula for the relative frequency $r_j$, $j = 1, 2, \ldots, k$ associated with each class $(c_{j-1}, c_j]$ in terms of the frequency $f_j$, the class boundaries and the number $n$ of sample data points.

**Problem 5** (modal class) The **modal class** associated with a general frequency histogram is the class with the largest frequency. The **mode** of the data is the class mark of the modal class.

**(a)** Find the modal class for the frequency histogram you found in part **(a)** of Problem 4 above.

**(b)** Find the mode of the data in Problem 4 above.

**(c)** Show the mode depends not only on the data but also on the choice of class boundaries.

**Problem 6** (sample PMF) The relative frequency histogram values $r_1, r_2, \ldots, r_k$ associated with the class marks $a_1, a_2, \ldots, a_k$ are sometimes said to constitute the **sample PMF**. Under what conditions can one expect the sample PMF to be proportional to the frequency histogram (function)?

**Problem 7** (binomial experiments and frequency) In the framework/terminology of Problems 8 and 9 of Assignment 6 consider the following two data descriptions:

**(i)** One hundred experiments consisting of one Bernoulli trial each.

**(ii)** One experiment consisting of one hundred Bernoulli trials.

**(a)** Which description describes the output `tottrials` of the `R` command

```
tottrials <- rbinom(100,1,rbinom(100,10,0.6)/10)
```

from Problem 10 of Assignment 6?

**(b)** Compose an `R` command producing data `totexp` corresponding to the description from among **(i)** and **(ii)** above which is not the answer to part **(a)**.

**(c)** Assuming you used (only) the base probability `prob = 0.6` in the function `rbinom` in composing your answer to part **(b)**, can you tell the difference between `tottrials` and `totexp`? If so, how?

**Problem 8** (simulated probabilities) Here is a data set, call it data set $A$, containing 90 samples:

```
 [1] 0.57 0.59 0.68 0.60 0.61 0.69 0.62 0.54 0.57 0.58 0.66 0.62 0.62 0.61 0.57
[16] 0.69 0.62 0.50 0.64 0.58 0.55 0.59 0.55 0.56 0.57 0.52 0.64 0.61 0.54 0.66
[31] 0.62 0.59 0.64 0.64 0.64 0.63 0.63 0.60 0.58 0.58 0.57 0.59 0.54 0.71 0.66
[46] 0.54 0.58 0.58 0.64 0.60 0.61 0.60 0.60 0.67 0.59 0.68 0.52 0.63 0.61 0.61
[61] 0.62 0.57 0.58 0.55 0.55 0.62 0.62 0.60 0.65 0.70 0.58 0.48 0.65 0.56 0.57
[76] 0.65 0.59 0.54 0.61 0.59 0.60 0.62 0.58 0.63 0.59 0.62 0.65 0.62 0.58 0.66
```

A second data set, call it data set $B$, is the following:

```
 [1] 0.52 0.58 0.55 0.66 0.54 0.59 0.50 0.62 0.61 0.56 0.72 0.71 0.64 0.57 0.63
[16] 0.61 0.60 0.59 0.59 0.58 0.56 0.66 0.62 0.67 0.52 0.64 0.60 0.70 0.61 0.57
[31] 0.64 0.60 0.58 0.59 0.61 0.46 0.61 0.62 0.59 0.55 0.59 0.59 0.57 0.57 0.46
[46] 0.63 0.62 0.57 0.62 0.66 0.61 0.65 0.61 0.67 0.61 0.60 0.55 0.63 0.58 0.61
[61] 0.53 0.62 0.70 0.55 0.67 0.56 0.60 0.52 0.60 0.60 0.60 0.71 0.61 0.61 0.58
[76] 0.56 0.69 0.57 0.64 0.72 0.67 0.60 0.61 0.62 0.57 0.58 0.57 0.65 0.52 0.57
```

**(a)** One of the data sets above was generated using the `R` command

```
round(rnorm(90,0.6,0.05),digits = 2)
```

The other was generated using

```
rbinom(90,100,0.6)/100
```

Can you tell which is which? If so, how? If not, why?

**(b)** Compute the mean and variance of the discrete measure associated with

`rbinom(90,100,0.6)/100`.

The steps on the next page are intended to walk you through this calculation if you need hints.

**(i)** Let $\beta : \mathcal{P}(\{0, 1, 2, \ldots, 100\}) \to [0, 1]$ denote the binomial induced measure. Define a new measure $\pi : \mathcal{P}(\{0, 0.1, 0.2, \ldots, 1\}) \to [0, 1]$ corresponding to the sampling values given by `rbinom(90,100,0.6)/100`. Hint: Ask yourself: What is the probability associated with the sample value $k/100$?

**(ii)** Remember[3] that the mean of the binomial induced measure on $n$ Bernoulli trials with base probability $p$ is $\mu = np$. Use this fact and the formula for the mean applied to the measure $\pi$ to compute the desired mean $\mu_\pi$.

**(iii)** Look up the formula for the variance of the binomial induced measure on $n$ Bernoulli trials with base probability $p$, and use this fact and the formula for the variance applied to the measure $\pi$ to compute the desired variance $\sigma_\pi^2$.

**Problem 9** (scaling and translating the range) Generalize the calculations from part **(b)** of Problem 8 above to calculate the mean and variance of the measure $\nu : \wp(\{a\omega + b : \omega \in S\}) \to [0, 1]$ by

$$\nu(\{a\omega + b\}) = \pi(\{\omega\})$$

where $a, b \in \mathbb{R}$ with $a > 0$ and $\pi : \wp(S) \to [0, 1]$ is a probability measure on a measure space $S \subset \mathbb{R}$ with $\#S < \infty$. What happens if $a < 0$?

**Problem 10** (variation of David Chan's probability problem) What is the probability that a best of seven game series goes to seven games (given the series is between teams A and B, and the probability team A beats team B in any particular game is $p$)?

[3] You can also look this up on Wikipedia or elsewhere.