Topic 2
Describing Distributions with Numbers
Measuring Center

Outline

Medians

The median take the middle value for $x_1, x_2, \ldots, x_n$ after the data has been sorted from smallest to largest,

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

($x_{(k)}$ is called the k-th order statistic. Sorting can be accomplished in R by using the `sort` command.) If n is odd, then this is just the value of the middle observation $x_{((n+1)/2)}$. If n is even, then the two values closest to the center are averaged.

$$\frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}).$$

For a vector `x`, we can write `median(x)` to compute the median.

Medians

Example. For the $n = 10$ observations,

$$1 \quad 3 \quad 3 \quad 3 \quad 4 \quad 2 \quad 4 \quad 2 \quad 1 \quad 3$$

We can compute the median by sorting

$$1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4.$$

and then averaging the 5th and 6th order statistic, $x_{(5)} = 3$,and $x_{(6)} = 3$. Thus, the median,

$$\frac{1}{2}(x_{(5)} + x_{(6)}) = \frac{1}{2}(3 + 3) = 3.$$

Means

For a collection of numeric data, If $x_1, x_2, \ldots, x_n$, the sample mean is the numerical average

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Alternatively, if the value If x occurs $n(x)$ times in the data, then use the distributive property to see that

$$\bar{x} = \frac{1}{n}\sum_{x} xn(x) = \sum_{x} xp(x), \quad \text{where} \quad p(x) = \frac{n(x)}{n}.$$

So the mean $\bar{x}$ depends only on the proportion of observations $p(x)$ for each value of x.

Means

Example. For the $n = 10$ observations,

$$1 \quad 3 \quad 3 \quad 3 \quad 4 \quad 2 \quad 4 \quad 2 \quad 1 \quad 3$$

We can compute the mean by adding

$$\bar{x} = \frac{1}{10}(1 + 3 + 3 + 3 + 4 + 2 + 4 + 2 + 1 + 3) = \frac{26}{10} = 2.6,$$

or by considering the number of occurrences of each value

$$\bar{x} = \frac{1}{10}(1n(1) + 2n(2) + 3n(3) + 4n(4)) = \frac{1}{10}(1 \cdot 2 + 2 \cdot 2 + 3 \cdot 4 + 4 \cdot 2) = \frac{26}{10} = 2.6.$$

Means

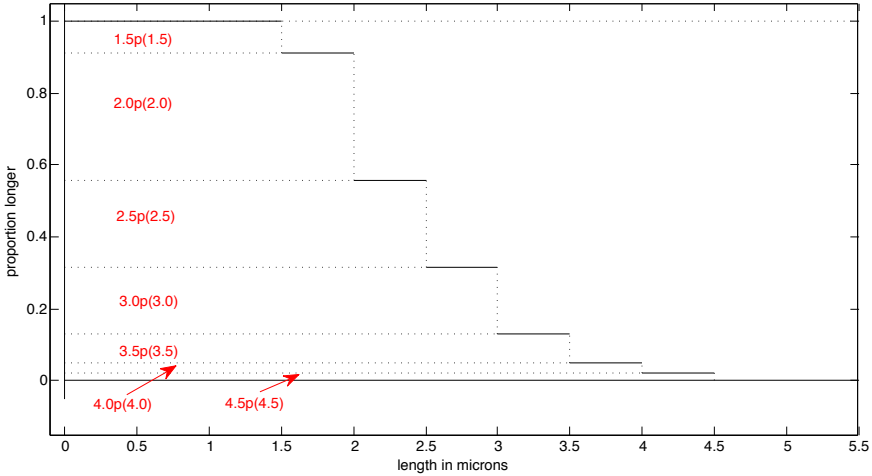
Example. For the data on the length in microns of wild type *Bacillus subtilis* data, we have

length x	frequency $n(x)$	proportion $p(x)$	product $xp(x)$
1.5	18	0.090	0.135
2.0	71	0.355	0.710
2.5	48	0.240	0.600
3.0	37	0.185	0.555
3.5	16	0.080	0.280
4.0	6	0.030	0.120
4.5	4	0.020	0.090
sum	200	1	2.490

Thus, the mean $\bar{x} = 2.490$ microns.

Connection to the Empirical Survival Function

For the survival function of the bacteria length, we compute the area of each rectangle.



Connection to the Empirical Survival Function

Notice that for each length x,

- the height of each of the rectangles is x,
- the width of each of the rectangles is $p(x)$.
- So, the area is the product $xp(x)$.

The sum of these areas are presented is the sample mean. In this way, we see that the area under the empirical survival function is the sample mean.

Connection to the Empirical Survival Function

Exercise. For the $n = 10$ observations,

$$1 \quad 3 \quad 3 \quad 3 \quad 4 \quad 2 \quad 4 \quad 2 \quad 1 \quad 3,$$

- draw the empirical survival function and
- show that the area under the function is the sample mean $\bar{x}$.

Weighted Mean

Many times, we do not want to give the same weight to each observation.

For example, in computing a student's grade point average,

- we set values x_i corresponding to grades

$$\mathrm{A} \mapsto 4 \quad \mathrm{B} \mapsto 3 \quad \mathrm{C} \mapsto 2 \quad \mathrm{D} \mapsto 1 \quad \mathrm{F} \mapsto 0$$

- give weights $w_1, w_2, \ldots, w_n$ equal to the number of units in a course. We then compute the grade point average as a weighted mean

.

Weighted Mean

- Multiply the value of each course by its weight $x_i w_i$, called the quality points in computing the grade point average.
- Add up the quality points:

$$x_1 w_1 + x_2 w_2 + \ldots + x_n w_n = \sum_{i=1}^{n} x_i w_i.$$

- Add up the weights, i e., the number of units attempted:

$$w_1 + w_2 + \ldots + w_n = \sum_{i=1}^{n} w_i.$$

- Divide the total quality points by the number of units attempted:

$$\frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}.$$

For weights `w`, then we can compute the weighted mean using `weighted.mean(x,w)`.