Statistics 5, Section D

Jizhou Kang jkang37@ucsc.edu

Ph.D. Student, Department of Statistics, UCSC

O.H.: Monday 2pm - 3pm Tuesday: 5pm-6pm

Zoom link is on Canvas.

Section: Section D, Tuesday 3:20 pm - 4:55 pm.

What's in the discuss sections?

Examples, Q&A.

Introduce yourself (grade, major, why you take this course, goals).

Histogram:

0, 1, 4, 4, 5, 2, 3, 11, 15, 6,
8, 18, 20, 17, 8, 12, 16, 16, 14, 9. } 20 data.

Step 1: List all data in increasing order.

0, 1, 2, 3, 4, 4, 5, 6, 8, 8,
9, 11, 12, 14, 15, 16, 16, 17, 18, 20.

Step 2: Decide on class interval size ("class interval" is the technical word for each of the "bins" that define the bars). Every bin should have the same size.

Data range from 0 to 20. let's have 4 bins.
[0, 5] (5, 10], (10, 15], (15, 20].

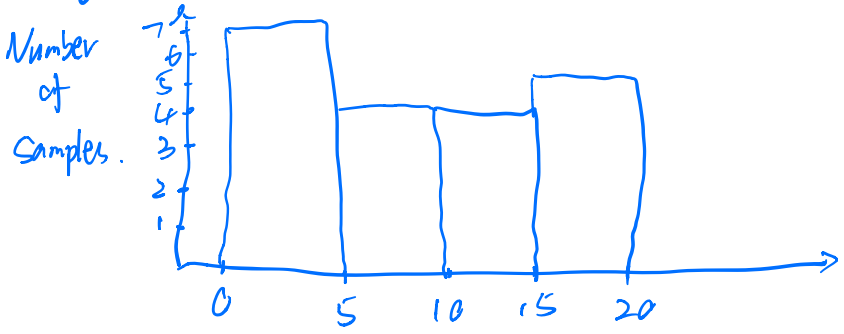
Step 3: Make a table. Count how many samples you have in each interval.

[0, 5]: 7.
(5, 10]: 4
(10, 15]: 4
(15, 20]: 5.

Step 4. Decide the type of the histogram, then decide the height of the bar.

type 1.
(absolute) frequency histogram. height of a bar is the count.

height of bars: 7, 4, 4, 5.

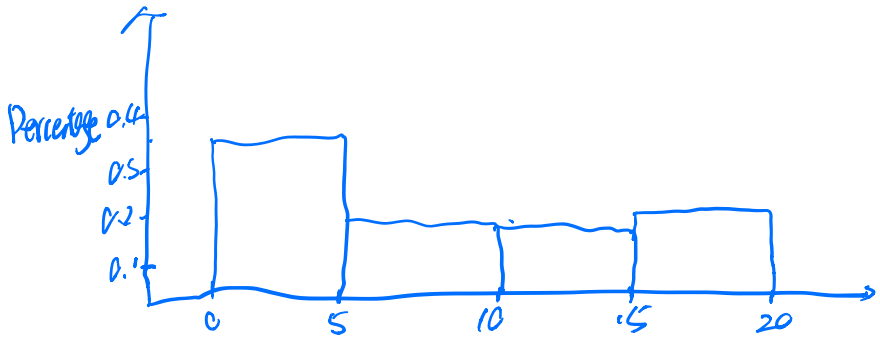


type 2: relative frequency histogram

height of bar is how many counts you observed in that interval divided by total number of observations.

Total number of observations: 20.

height of bar: $\frac{7}{20}=0.35$, $\frac{4}{20}=0.2$, $\frac{4}{20}=0.2$, $\frac{5}{20}=0.25$.



type 3:

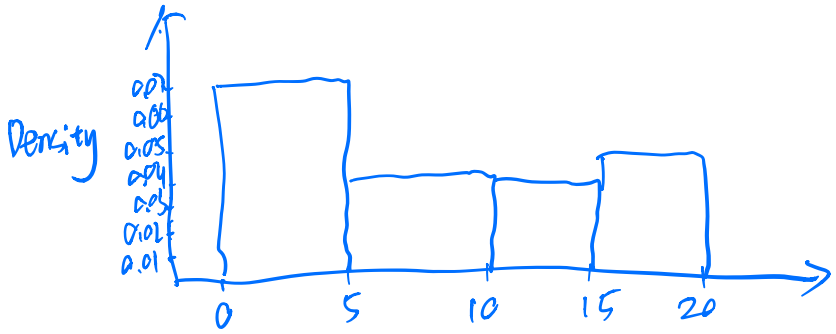
density histogram: the height of a bar is how many count you observed in that interval devided by total number of observations and then divide this number by the size of the interval.

height of bars: $\frac{7}{20} = 0.35$ $\frac{0.35}{(5-0)} = 0.07$

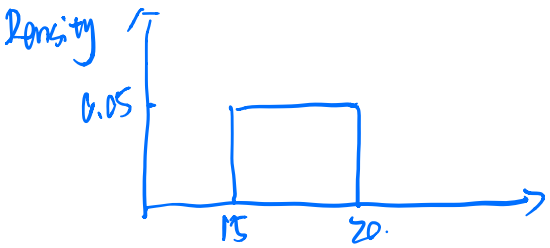
$\frac{4}{20} = 0.2$ $\frac{0.2}{(10-5)} = 0.04$

$\frac{4}{20} = 0.2$ $\frac{0.2}{(15-0)} = 0.04$

$\frac{5}{20} = 0.25$ $\frac{0.25}{(20-15)} = 0.05$.



Questions may also look like:



What's the percentage of samples belonging to (15, 20] ?

The height of density histogram is the height of percentage histogram divided by length of the interval.

So. percentage = density × length of the interval.

$= 0.05 \times 5 = 0.25 = 25\%$.

(This is also the area of the bin.

The area of the bars in density histogram is the percentage of the total population).

Average / Median / Mode / Standard deviation.

How to describe your data? Center, spread.

Average: $\frac{\text{sum of your data}}{\text{total counts}} = \frac{\sum_{i=1}^{n} x_i}{n}$.

0, 1, 2, 3, 4, 4, 5, 6, 8, 8,

9, 11, 12, 14, 15, 16, 16, 17, 18, 20

$$\frac{0+1+2+3+4+4+5+6+8+8+9+11+12+14+15+16+16+17+18+20}{20}$$

$= \frac{10+31+61+87}{20} = \frac{189}{20} = 9.45.$

Median: rank your data from the smallest to largest.

number in the middle of your data.

0, 1, 2, 3, 4, 4, 5, 6, 8, 8, (10th)

(11th) 9, 11, 12, 14, 15, 16, 16, 17, 18, 20

the total number is even: average of the middle. $\frac{8+9}{2} = 8.5$

the total number is odd:

0, 1, 2, 3, 4, 4, 5, 6, 8, (8,) median is 8.

9, 11, 12, 14, 15, 16, 16, 17, 18

Median can be some number that does not include in the original data!

Mode: count the times each number appears.

0, 1, 2, 3, 4, 4, 5, 6, 8, 8,

9, 11, 12, 14, 15, 16, 16, 17, 18, 20

4: 2 times, 8: 2 times, 16: 2 times. all the other: 1 time.

Mode : 4, 8, 16.

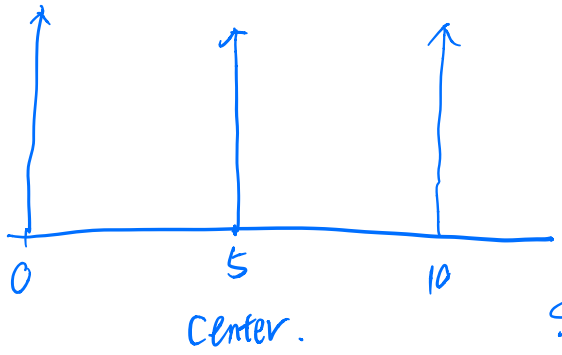
Mode does not need to be unique!

Standard deviation:

SD = root mean square differences from the mean.

$$= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \text{mean}(x_i))^2}{n}}$$

Why? SD measures the spread of data.



how far is each sample to the center.

0-5=-5, 10-5=5.

Sum up got 0. does this make sense?

negative value and positive values cancels, so we need to make negative values positive, but keep the scale. how?

absolute value of square.

absolute value: mean absolute difference: $\frac{\sum_{i=1}^{n} |x_i - \text{mean}(x_i)|}{n}$.

square: mean square difference: $\frac{\sum_{i=1}^{n} (x_i - \text{mean}(x_i))^2}{n}$.

why take square root? keep the unit the same. → SD.

0 1 2 3 4. $\frac{0+1+2+3+4}{5} = \frac{3+7}{5} = \frac{10}{5} = 2$.

$$\sqrt{\frac{(0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2}{5}} = \sqrt{\frac{2^2 + 1^2 + 0^2 + 1^2 + 2^2}{5}}$$

$$= \sqrt{\frac{4+1+0+1+4}{5}} = \sqrt{\frac{10}{5}} = \sqrt{2}.$$