LECTURE 3: UNUSUAL DESCRIPTIONS AND PERCENTILES

AP STATS

Recall that whenever you are asked to describe the shape of a distribution that you need to use SOCS:

- **S**hape: is it symmetric? how many modes? is it skewed?
- **O**utliers: are there any outliers or *unusual features*?
  - unusual features include *gaps* (intervals of no data points) and outliers
- **C**enter: where is the *median*? what is the *mean* (average)?
- **S**pread: what is the *range* $(\max - \min)$ and what is the *standard deviation*?

1. Percentiles

Given a quantitative variable (numerical data set) with $X = \{13, 7, 1, 4, 19, 16\}$, we can create rankings, **but the points in the data must be ordered correctly**.

Rearranging X gives us that $X = \{1, 4, 7, 10, 13, 16, 19\}$ and so we can determine the individual *percentile rankings* of each point in the data.

Definition. Given a data set X, the **pth percentile of** X, for $0 < p \leq 100$, is the smallest value $k \in X$ such that *no more than* p percent of the data is strictly less than k **and** at least p percent of the data $\leq k$.

A few properties:

- The 100th percentile is defined to be the largest value in the ordered list.
- A percentile calculated should refer to a member on the original ordered list

Given an ordered list, there's 2 different methods we can consider to talk about the data:

$$\boxed{Ordinals} \longleftrightarrow \boxed{Percentiles}$$

Ordinals are enumerations- whenever we say the first, second, third, etc...these are ordinals.

To go from Orindals to Percentiles, follow the algorithm:

(1) Order the list. If k is the object we are interested in, mark the position of k as n (this is the ordinal)

(2) k is in the pth percentile where

$$\boxed{p = \frac{100}{N} \times n, \quad N = \text{size of } X}$$

So in our working example, where $X = \{1, 4, 7, \ldots, 16, 19\}$, the if we are interested in the 13, we compute the percentile that 13 lives in by:

$$\begin{aligned} p &= \frac{100\%}{(N = 7)} \times (n = 5) && \text{since 13 is 5th} \\ &\approx 71.428\% \end{aligned}$$

so the 13 is in the 71.428th percentile.

To go from Percentiles to Ordinals, we follow the algorithm (e.g. given p, find n in the formula above)

(1) Order the list.
(2) If k is in the nth position of the list, then

$$\boxed{n = \lceil \frac{p}{100} \times N \rceil, \quad N = \text{size of } X}$$

(3) Your answer is k.

Question 1. Given that $X = \{13, 7, 1, 4, 19, 16\}$ what number would mark the 20th percentile?

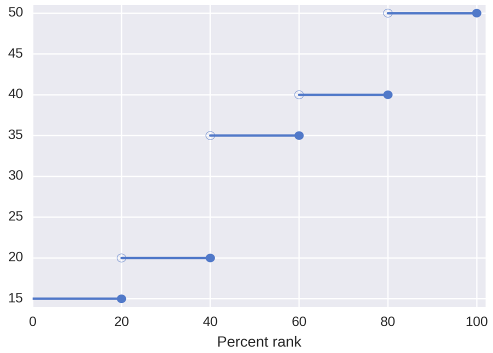
Solution. Following the algorithm,

(1) Reordering X, we get $X = \{1, 4, 7, 10, 13, 16, 19\}$
(2) Compute that

$$\begin{aligned} n &= \lceil \frac{(p = 20)}{100} \times (N = 7) \rceil \\ &= \lceil \frac{7}{5} \rceil && \text{simplifying} \\ &= 2 && \text{ceiling function rounding} \end{aligned}$$

So our answer is the 2nd number: the 20th percentile would be the 4.

The percentile →ordinal is a ceiling function- it always rounds up no matter what. Graphically, if $X = \{15, 20, 35, 40, 50\}$ then we can map out the percentiles accordingly:



the x-axis is the percentile ranking (the p) while the y-axis marks the actual values of $x \in X$.

2. Probability Measures

Percentiles, which rank the data, leads us into the theory of probability, the mathematical language of uncertainty. In probability, the objects we are interested in are random variables and the way we study random variables is by associating a *probability measure*, which *measures the probability of obtaining a certain data point* in the set.

Definition. [1] Let X be a quantitative (random) variable. A **probability measure of X** (oftentimes called the **probability mass function** (PMF) or **probability density function** (PDF)) is any function $f : \Omega \to \mathbb{R}$ with the following 3 properties:

(1) Non-negativity $f(A) \geq 0$ for all $A \subseteq X$ $\leftrightsquigarrow$ "you can never have negative probabilities"
(2) Normalization $f(X) = 1$ $\leftrightsquigarrow$ "the probability of obtaining the everything possible is 1"
(3) Additivity If $A, B \subseteq X$ and $A \cap B = \emptyset$, then $f(A \cup B) = f(A) + f(B)$ $\leftrightsquigarrow$ "prob(this or that) = prob(this) + prob(that)...as long as this and that can't both happen at the same time"

Some properties about probability measures: If we let f be a probability measure on a random variable X, then

(1) $f(\emptyset) = 0$
(2) If $A \subseteq B$ then $f(A) \leq f(B)$
(3) $f(A^c) = 1 - f(A)$

Question 2. Suppose we have $X = \{1, 3, 5\}$ and suppose $p(x) = kx^2$ for some constant k. What is the value of k if we want p to be a probability measure?

Solution. Rewriting X as $X = \{1\} \cup \{3\} \cup \{5\}$, from Additivity we get that

$$\begin{aligned} p(X) = p(\{1,3,5\}) = p(\{1\} \cup \{3\} \cup \{5\}) & \\ &= p(\{1\}) + p(\{3\}) + p(\{5\}) && \text{using additivity} \\ &= k(1)^2 + k(3)^2 + k(5)^2 && \text{plugging } x \text{ into } p(x) \\ &= k + 9k + 25k = 35k \end{aligned}$$

Using the fact that $p(X) = 1$ from Property (2) Normalization, we conclude that $p(X) = 35k = 1 \Longrightarrow k = \dfrac{1}{35}$.

[1] The definition provided is that of a probability measure, which is slightly more general than PDF or PMF's, which have very specific *measures* defined on them...we won't define what a measure is...