Don't be mean, be above average: Understanding data distribution and descriptive statistics

K.W. Herbst ^{a,*}, G.M.A. Beckers ^b, L. Harper ^c, D.J. Bagli ^d, A.J. Nieuwhof-Leppink ^e, M. Kaefer ^f, M. Fossum ^{g,h}, N. Kalfa ⁱ, ESPU research committee

A Division of Urology, Department of Research, Connecticut Children's Medical Center, Hartford, CT, United States

B Department of Urology, Section of Pediatric Urology, Amsterdam UMC, Location VUmc, Amsterdam, the Netherlands

C Department of Pediatric Urology and Pediatric Surgery, Hopital Pellegrin-Enfants, CHU Bordeaux, France

D Division of Urology, Departments of Surgery and Physiology, University of Toronto, Developmental and Stem Cell Biology, The Hospital for Sick Children and Research Institute, Toronto, Ontario, Canada

E Department of Medical Psychology and Social Work, Urology, Wilhelmina Children's Hospital, University Medical Center Utrecht, PO Box 85090, 3508 AB, Utrecht, the Netherlands

F Riley Hospital for Children, Indiana University, Indianapolis, IN, United States

G Department of Pediatric Surgery, Surgical Clinic C, Rigshospitalet, Copenhagen University, Copenhagen, Denmark

H Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

I Service de chirurgie et urologie pediatrique, hopital Lapeyronie, CHU de Montpellier et Universite' de Montpellier, France

* Correspondence text: Katherine Herbst, 282 Washington Street, Hartford, CT, 06106, United States.

kherbst@connecticutchildrens.org (K.W. Herbst)

Understanding the characteristics of data is vital when conducting analysis and interpreting findings.

This article is intended to provide a base understanding of distribution and descriptive statistics.

This is the author's manuscript of the work published in final edited form as:

Herbst, K. W., Beckers, G. M. A., Harper, L., Bägli, D. J., Nieuwhof-Leppink, A. J., Kaefer, M., Fossum, M., Kalfa, N., & ESPU research committee. (2020). Don't be mean, be above average: Understanding data distribution and descriptive statistics. *Journal of Pediatric Urology*, 16(5), 712.
<https://doi.org/10.1016/j.jpurol.2020.09.007>

First, you must know some basic vocabulary. A variable is a characteristic, number, or classification that can be measured. A variable that can take on any value between two numbers is a continuous variable. A categorical variable can only take on one of a limited number of values. There are two types of categorical variables, nominal and ordinal. With nominal variables, such as gender, order does not matter. With ordinal variables, such as vesicoureteral reflux grade, order matters.

Mean, median, mode are used to describe a continuous variable's distribution. The mean is the average of all the values, the median the middle number, and the mode the most frequent value. Continuous variables are either normally distributed or have a skewed distribution. A normal distribution is bell shaped, symmetrical, and the mean, median, and mode are all the same number. A skewed distribution is where one of the distribution tails is longer than the other. A variable is right or positive skewed if the long tail is to the right of the graph, and left or negative skewed if the long tail is to the left.

When describing a normal distribution, mean and standard deviation (SD) are used. SD is a measure of how spread out, or dispersed, the values are from the mean. Sixty-eight percent of the values are within one SD of the mean, 95% within two SDs, and 99.7% within three SDs of the mean. This is known as the 68–95–99.7 rule or the 3-sigma rule. If a variable's distribution is normally distributed, there's a better chance that the true mean of the population was captured.

Variables often have a skewed distribution. The mean is now influenced by extreme values, and is no longer a representative measure of the data. For skewed variables, either the median or the mode are used, and range or interquartile range (IQR) used to describe the data's dispersion. Range is the lowest to the highest value. IQR is the 25th to the 75th percentile if the distribution is broken into four equal parts. While range allows a reader to know the lowest and highest values, it tells them nothing about how the data is dispersed between these values. IQR, on the other hand, represents 50% of the data so provides the reader with more information.

It's easy to understand why calculating a mean of a categorical variable doesn't make sense. Categorical variables are most often described as counts or percentages.

There are numerous reasons why it's important to understand and report distribution correctly. Most statistical tests have rules about the type of distribution that can be used [], and a variable's distribution can reveal bias in the data. Distribution also allows readers to determine if results can be applied to their own patient population. Thus, careful attention must be paid to data distribution and how it's reported.

Declaration of competing interest

All authors have no disclosures.

On behalf of the ESPU Research Committee.

Reference

- 1. Jaykaran C.: How to select appropriate statistical test?. J Pharm Negat Results 2010; 1: pp. 61-63. <http://www.pnrjournal.com/text.asp?2010/1/2/61/75708>