# STAT 311: Describing Data

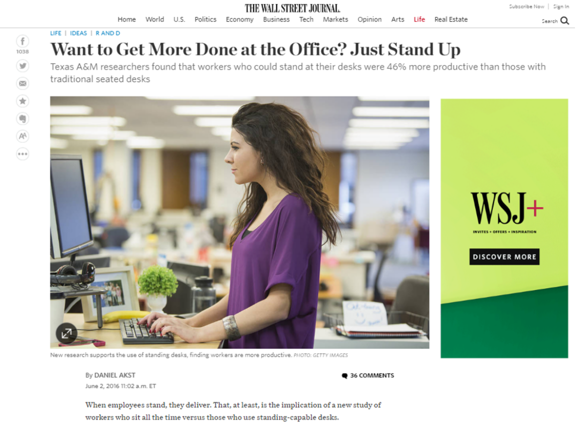
Y. Samuel Wang

Summer 2016

# Gathering Data

Fill out the survey that I just emailed out

# Standing Desk Study



THE WALL STREET JOURNAL.

Subscribe Now | Sign In
Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate
Search
1038

LIFE | IDEAS | R AND D

## Want to Get More Done at the Office? Just Stand Up

Texas A&M researchers found that workers who could stand at their desks were 46% more productive than those with traditional seated desks

New research supports the use of standing desks, finding workers are more productive. *PHOTO: GETTY IMAGES*

By DANIEL AKST

June 2, 2016 11:02 a.m. ET

36 COMMENTS

When employees stand, they deliver. That, at least, is the implication of a new study of workers who sit all the time versus those who use standing-capable desks.

Figure: Researchers found that individuals at a call center were 46% more productive when they used a standing desk compared to individuals who did not have a standing desk

# Basic Terminology

- **Observational Unit**: Individual in our study
- **Variable**: Characteristic that differs from unit to unit
- **Population**: The set of all units we are interested in
- **Parameter**: Some description of the population
- **Sample**: The set of all units on which we have data
- **Statistic**: Some description of our sample

# Types of Variables

How variables can be recorded

- Numeric: Variables that take the form of quantitative measurements
  - Discrete: Variables which can only take on certain values. Typically a count
  - Continuous: Variables which can be measured to arbitrary precision
- Categorical: Variables that take labels or categories
  - Ordinal: Categorical variable which has logical ordering

# Types of Variables

How variables can be recorded

- Numeric: Variables that take the form of quantitative measurements
    - Discrete: Variables which can only take on certain values. Typically a count
    - Continuous: Variables which can be measured to arbitrary precision
- Categorical: Variables that take labels or categories
    - Ordinal: Categorical variable which has logical ordering

How we think about variables

- Response: Typically the variable of interest. The variable which we want to measure change in. Can be explained partially by explanatory variable
- Explanatory: Typically what we want to measure the effect of

# Standing Desk Study

Methods from **(author?)** [1]

- 167 Employees at a health call center (118 females, 49 males)
- Examined 2 groups: Those with standing desks, those with sit only desks
- Measured productivity in the form of “Encounters per hour”
- Gathered self reported comfort/discomfort ratings

# Visualizing Numeric Data

- 167 Employees at a health call center (118 females, 49 males)
- Examined 2 groups: Those with standing desks, those with sit only desks
- Measured productivity in the form of “Encounters per hour”
- Gathered self reported comfort/discomfort ratings

# Describing Distributions: Centrality

- **Mean**: $\frac{1}{N}\sum_i X_i$; Often denoted by $\bar{x}$
  - $3, 5, 2, 1, 3, 7 \Rightarrow (3 + 5 + 2 + 1 + 3 + 7)/6 = 21/6$
- **Median**: "Middle observation"
  - Sort the elements
  - Select the element in the middle
  - Ex: $3, 5, 2, 1, 7 \Rightarrow 1, 2, \mathbf{3}, 5, 7$
  - If there is an odd number, take the average of the "middle two" elements
  - Ex: $3, 5, 2, 1 \Rightarrow 1, \mathbf{2}, \mathbf{3}, 5$, so the median $= 2.5$
- **Mode**: Most common observation
  - $3, 5, 2, 1, 3, 7 \rightarrow \mathbf{3}, 5, 2, 1, \mathbf{3}, 2$, so the mode is 3

# Summarizing Numeric Data

Five Number Summary

- Min: Smallest value
- First Quartile (Q1): Median of all values **below** the median
- Median: Middle Value
- Third Quartile (Q3): Median of all values **above** the median
- Max: Largest value

# Describing Distributions: Spread

- Standard Deviation: $\sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}$
- Interquartile Range (IQR): Q3 - Q1
- Range: Min - Max

# What should I be using?

With so many different measures of the same idea, what should I be using?

# What should I be using?

With so many different measures of the same idea, what should I be using?

Depends what you care about

# What should I be using?

Robust to outliers

- Median
- IQR

Not robust to outliers

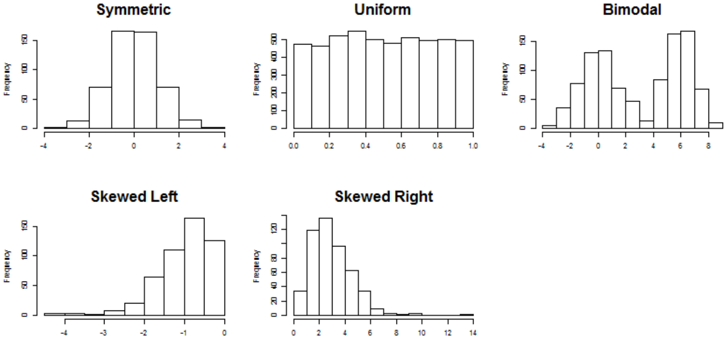
- Mean
- SD
- Range

**Outlier**: Observation that is not consistent with the bulk of the data

# Data Analysis

Let's take a look at the average number of hours spent thinking about statistics

# Describing Distributions: Shape



What shapes do each of the gathered variables have?

# Visualizing Numeric Data

Boxplot: Shows the five number summary in visual form

# Visualizing Numeric Data

Histogram: Bar plot which shows the number of occurrences for each value

# Visualizing Numeric Data

Stem and Leaf: Similar to histogram, but displays actual values

# Summary

- Basic terminology
- Types of variables
- How to summarize sets of numbers

[1] Gregory Garrett, Mark Benden, Ranjana Mehta, Adam Pickens, Camille Peres, and Hongwei Zhao. Call center productivity over 6 months following a standing desk intervention. *IIE Transactions on Occupational Ergonomics and Human Factors*, 0(ja):00–00, 0.