Big Data for Cities

Week 2

Curt Savoie
Connor McKay

Agenda

- Recap on last week
- What does the Government Analytics Space look like?
- Stats Review
- More R Demo

Recap

- How did things go with R installation?
- Dataverse sign up?
- Other Data Sources?

Government Analytics Overview

- Analytics have mostly been in Cities (as opposed to states)
- They come from different areas and that is reflected in the work the put out
  - Performance Management
  - Urban Planning
  - IT
  - Administration and Finance

Government Analytics Overview

Types of Services include:

- Open Data
- Analytics
- Data Policy and Governance
- Business Intelligence and Data Warehousing
- Performance Management

Government Analytics Overview

Common operational sources of data:

- 311
- Building Permits
- Zoning and economic regulations
- 911
- Transportation
- Environment & Energy
- Public Health
- GIS including tax assessing

Government Analytics Overview

Common Issues:

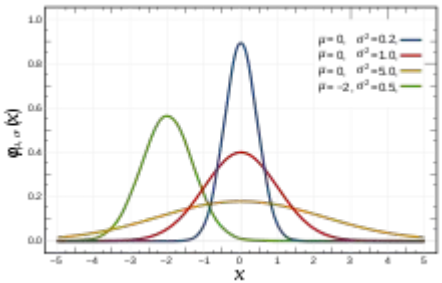
- Messy or nonexistent data with little documentation
- Lack of technical resources
- Outdated technology
- Lack of policy to allow for data work
- Lack of executive buy-in

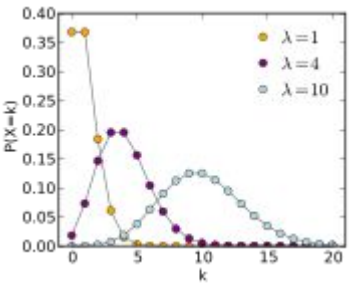
Government Analytics Overview

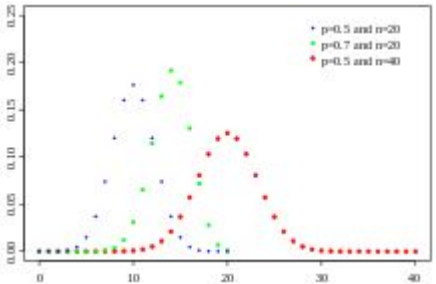
- New York City (http://www1.nyc.gov/site/analytics/index.page)
- Chicago (https://www.cityofchicago.org/city/en/depts/doit/provdrs/data_sciences.html)
- Boston (https://www.boston.gov/departments/analytics-team)
- Metropolitan Area Planning Council (https://www.mapc.org/)
- Pittsburgh / Allegheny County (http://pittsburghpa.gov/innovation-performance/ip-data-analytics.html)

Stats Overview

- Distributions
  - Gaussian (“Normal”), Poisson, Binomial
- Important information about distributions:
  - Mean, Median, Variance (many, many more)







Mean

- Also known as the “expected value” of a distribution denoted by a big E.
- Simple means are calculated with:

$$E(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- In R: mean(x)

Median

- Measures central value of distribution, regardless of variance.
- “Robust” statistic, resilient to outliers
- Means are suspect without medians!

Variance

- A measure of how “spread out” your data is
- Data with values ranging from -1,000 to 1,000 will have higher variance than data taking on values between 1 and 10
- Variance is calculated with:

$$Var(X) = E[(X - \mu)^2]$$

- In R: var(x)

For Next Week

- **Reading on theory and practice**
  - http://labs.openviewpartners.com/curt-savoie-future-of-open-data/#.WcHFJdOGO2w
  - http://datasmart.ash.harvard.edu/news/article/planning-the-data-driven-city-1003
  - http://www.govtech.com/data/Cities-Are-Having-a-Data-and-Analytics-Driven-Moment-and-Its-Likely-to-Stay.html
- In R
  - R4DS reading in syllabus
  - Homework 1