Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

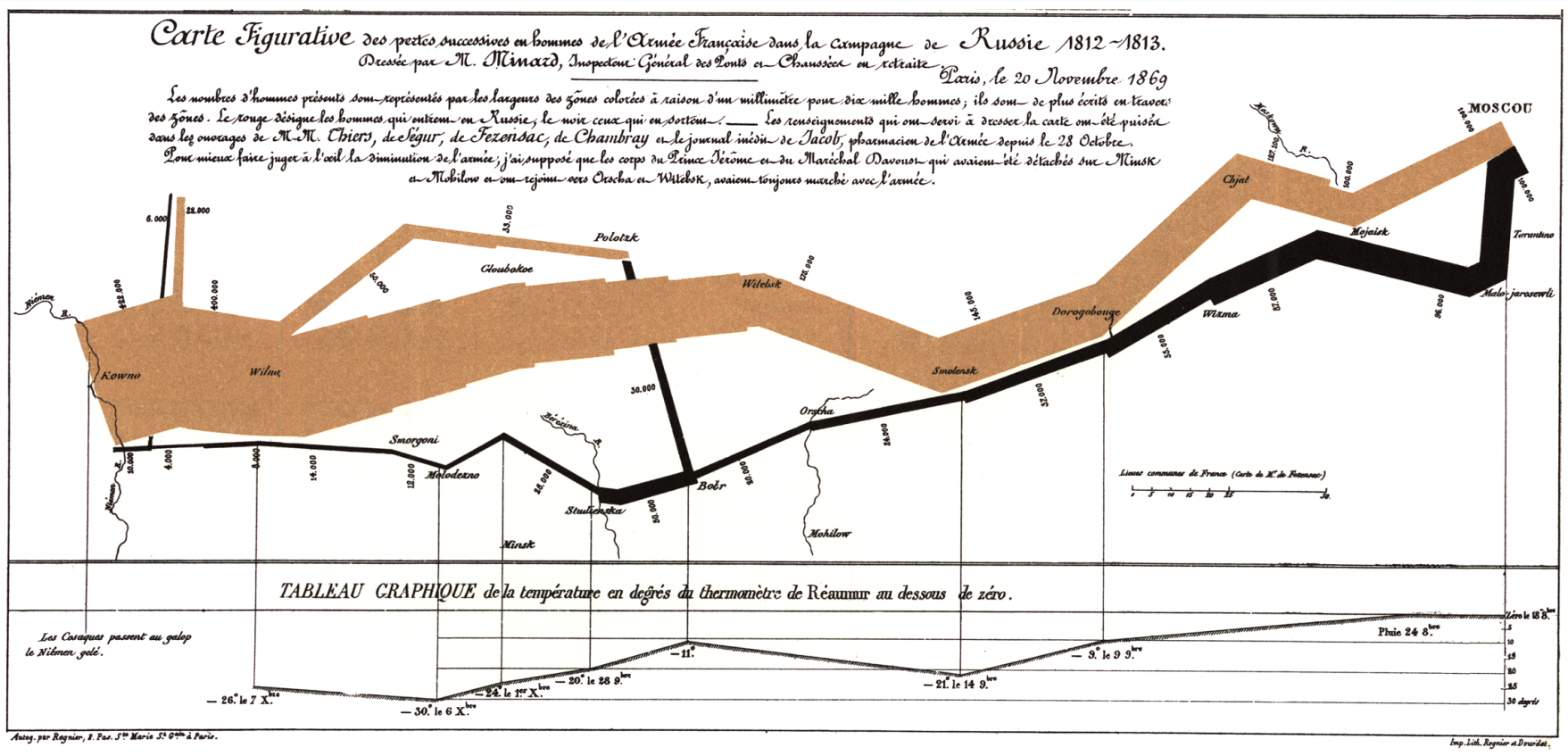
January 14, 2016

The Voinovich School of Leadership and Public Affairs

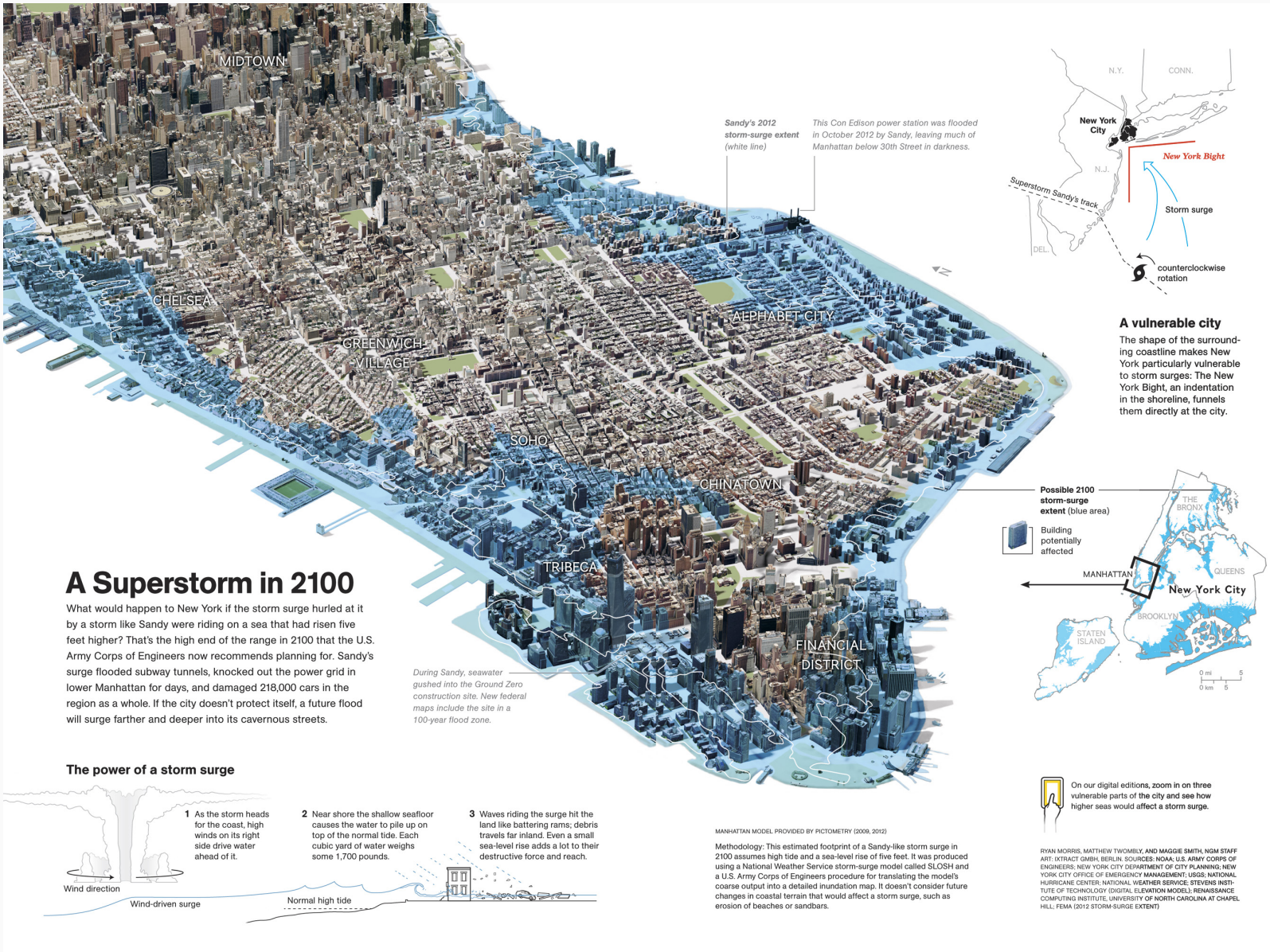
Table of Contents

Visualizing Data

Minard's Map


Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite
Paris, le 20 Novembre 1869
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Kowno
Wilna
Smorgoni
Minsk
Molodexno
Studienska
Polotzk
Bobr
Houbokoe
Witebsk
Orscha
Mohilow
Dorogobouge
Smolensk
Wixma
Chjat
Mojaisk
MOSCOU
Malo-jaroswetz
Tarantino
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU CRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26° le 7 Xbre
— 30° le 6 Xbre
— 24. le 1er Xbre
— 20° le 28 9bre
— 11°
— 21° le 14 9bre
— 9° le 9 9bre
Pluie 24 8bre
Zéro le 18 8bre
Anteg. par Regnier, 8. Pas. St Marie St Gbin à Paris.
Imp. Lith. Regnier et Dourdet.

Super Storm & NYC?



Displaying Frequency Distributions

Frequency Tables: Categorical Data

Table 1: Frequencies

Cause	No.deaths
Accidents	6688
Homicide	2093
Suicide	1615
Malignant tumor	745
Heart disease	463
Congenital abnormalities	222
Chronic respiratory disease	107
Influenza and pneumonia	73
Cerebrovascular diseases	67
Other tumor	52
All other causes	1653

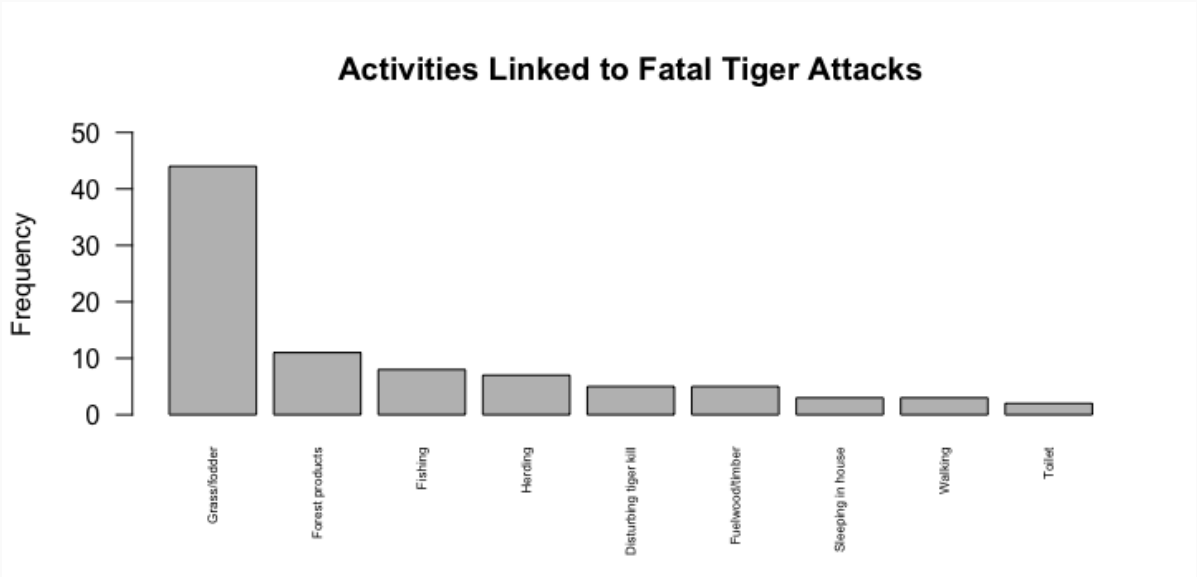
Table 2: Relative Frequencies

	No.deaths
Accidents	0.49
Homicide	0.15
Suicide	0.12
Malignant tumor	0.05
Heart disease	0.03
Congenital abnormalities	0.02
Chronic respiratory disease	0.01
Influenza and pneumonia	0.01
Cerebrovascular diseases	0.00
Other tumor	0.00
All other causes	0.12

Bar Graphs: Categorical Data

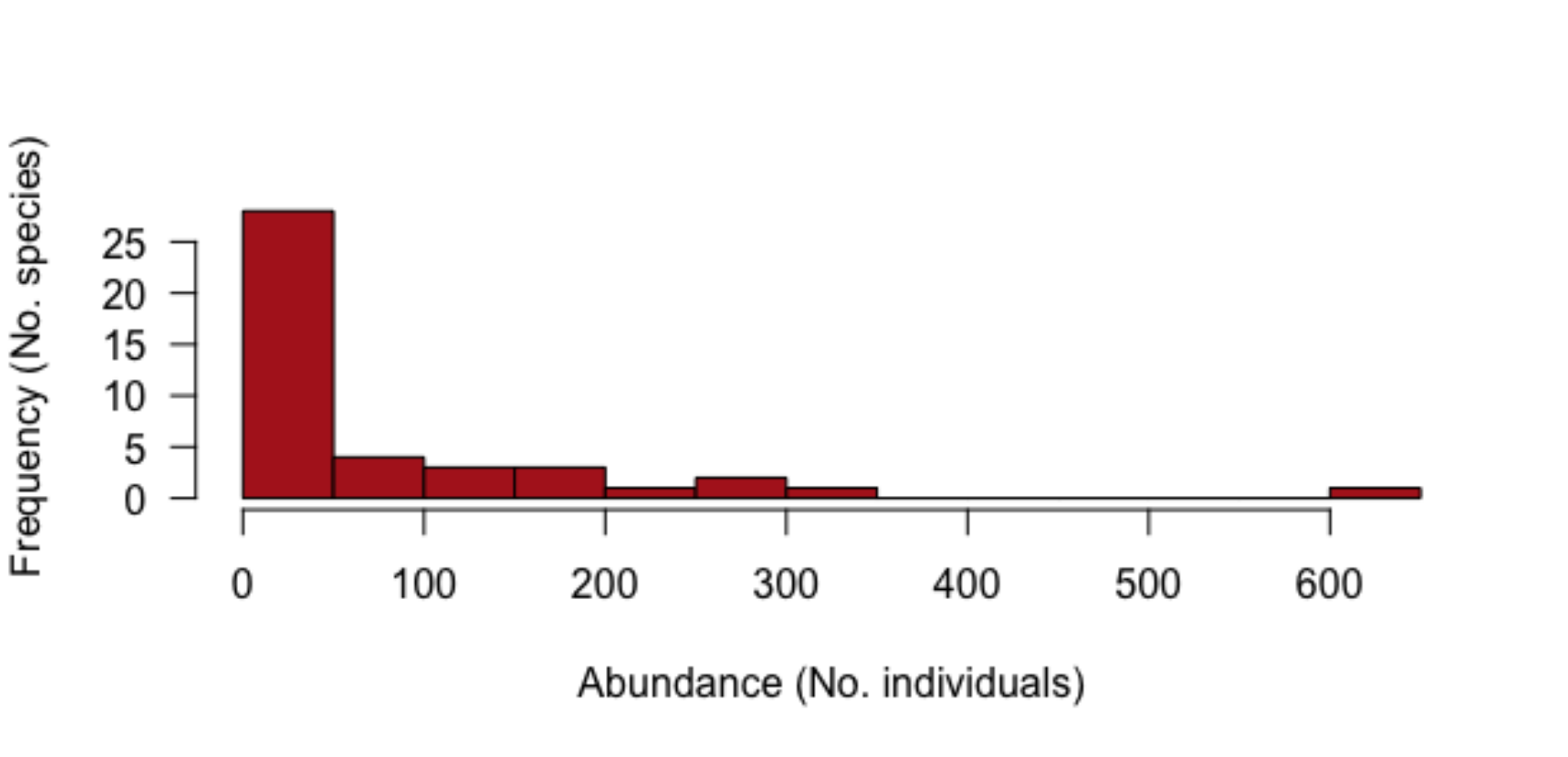
Data on humans killed by tigers while victims engaged in specific activities by tigers near Chitwan National Park (Nepal)

	Frequency
Grass/fodder	44
Forest products	11
Fishing	8
Herding	7
Disturbing tiger kill	5
Fuelwood/timber	5
Sleeping in house	3
Walking	3
Toilet	2
Sum	88



Histograms: Numerical Data

Data from survey of breeding birds of Organ Pipe Cactus National Monument in southern Arizona

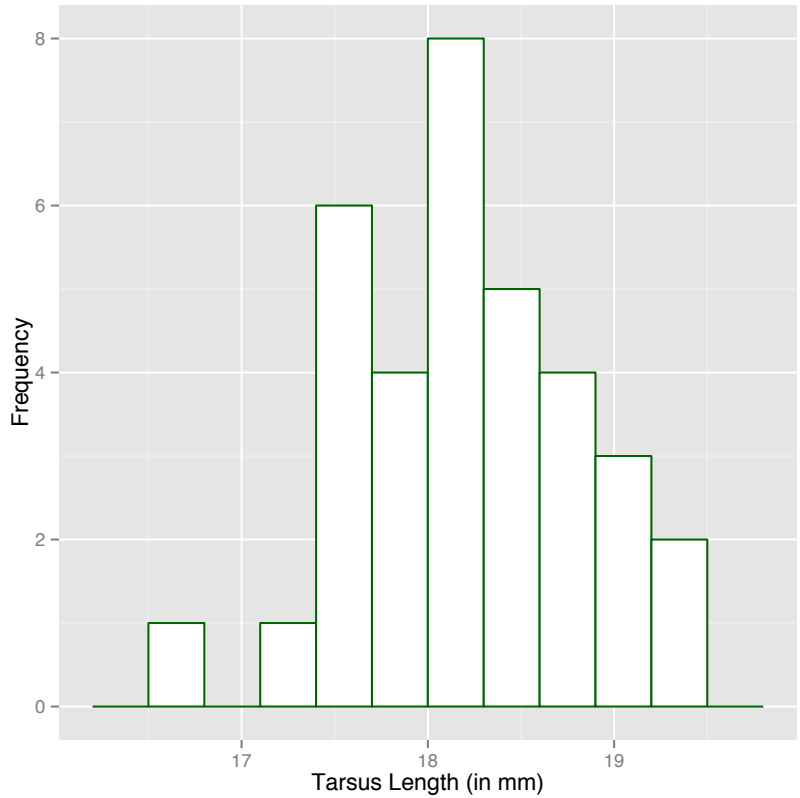


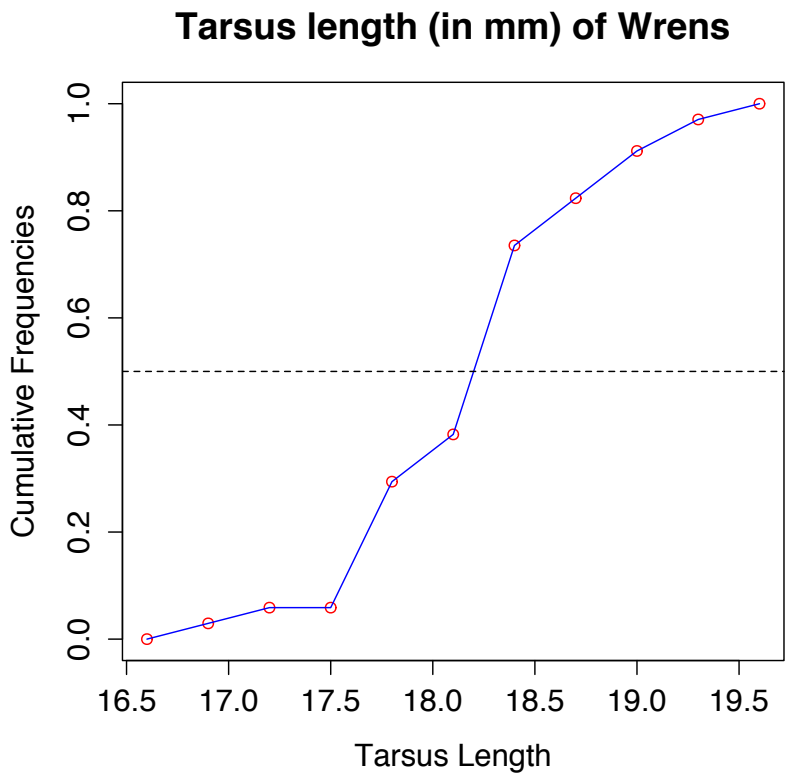
Tarsus lengths (in mm) of Wrens

Group	Freq	Relative Freq	Cumulative Freq	Relative Cumulative Freq
[16.5,17)	1.00	0.03	1.00	0.03
[17,17.5)	1.00	0.03	2.00	0.06
[17.5,18)	10.00	0.29	12.00	0.35
[18,18.5)	13.00	0.38	25.00	0.74
[18.5,19)	6.00	0.18	31.00	0.91
[19,19.5)	3.00	0.09	34.00	1.00

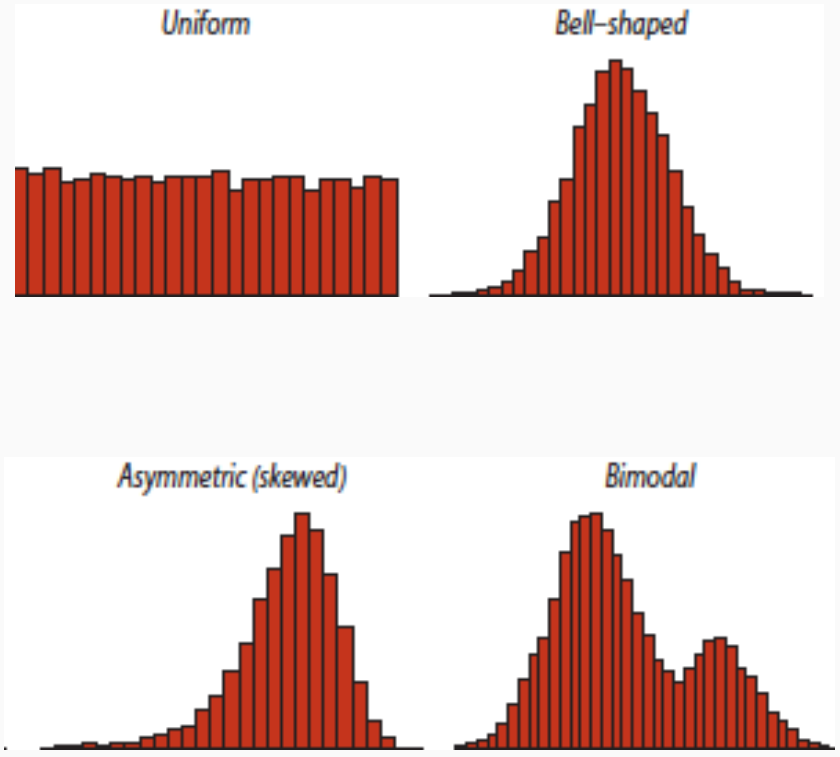
- [and) indicate "left-closed" and "right-open", i.e.,
  - include 16.5 and everything up to but not including 17 in the first group
  - Include 17 in the second group and then everything above 17 up to but not including 17.5
  - ... and so on
- Relative Freq is Frequency divided by the total number of units in the sample. For e.g., $\frac{1}{34} = 0.03; \frac{10}{34} = 0.29$
- Cumulative Freq for a group is the group's frequency $+$ all preceding frequencies

Plotting Tarsus Lengths





Describing the Shape of a Histogram



Key Points About Histograms

- Histograms vary ...
  1. Symmetric (cells split symmetrically)
  2. Skewed Left (easy exam so most score high, only a few low scores)
  3. Skewed Right (tough exam so most score low, only a few high scores)
  4. Uniform (Penguins)
  5. Bimodal (interval between geyser eruptions, drug inactivity in humans)
- Watch your bin width ... alters the shape of the Histogram
  1. Some pre-set rules:
     - Sturges: $h = 1 + \frac{ln(n)}{ln(2)}$; then round up to nearest integer
     - Freedman–Diaconis: $h = 2\left[\frac{IQR}{n^{\frac{1}{3}}}\right]$
     - Scott: $h = \frac{3.5\sigma}{\sqrt[3]{n}}$

Associations Between Categorical Variables

Associations Between Categorical Variables

- Many ways to evaluate how two or more categorical variables are related
- Easiest method is a contingency table
- Note: Columns = Explanatory variable; Rows = Outcome of interest (i.e., Response variable)
- Does reproduction make the wild great tit (Parus major) more susceptible to malaria? $\cdots$ see below

	Experimental Treatment Group		
	Control	Egg-Removal	Row Total
Malaria	7	15	22
No Malaria	28	15	43
Column Total	35	30	65

Grouped Bar Graphs

Definition

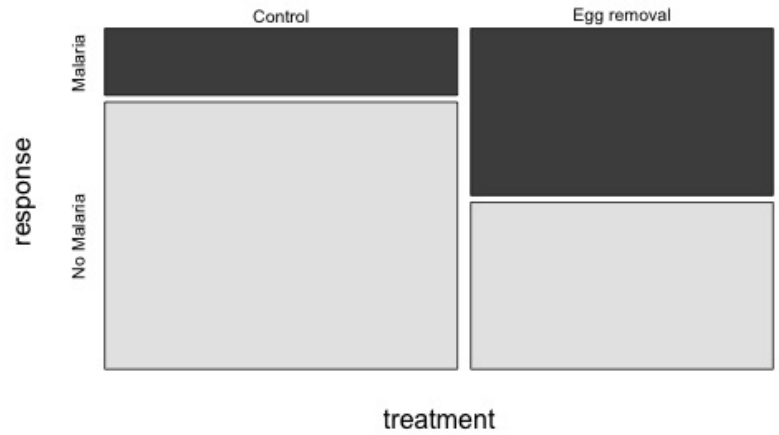
Grouped bar graphs show the frequency of all combinations of two or more categorical variables

Mosaic Plot

Definition

Mosaic plots use the area of rectangles to display the relative frequency of occurrence of all combinations of two or more categorical variables

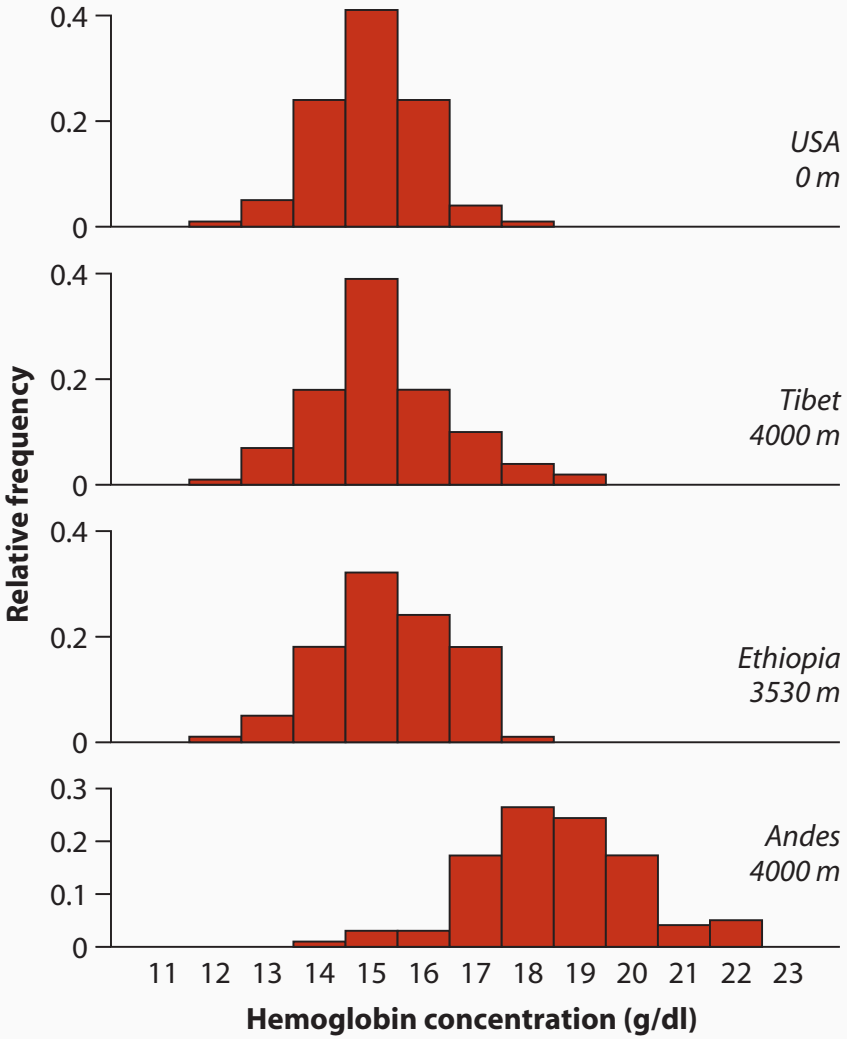
	Experimental Treatment Group		
	Control	Egg-Removal	Row Total
Malaria	7	15	22
No Malaria	28	15	43
Column Total	35	30	65



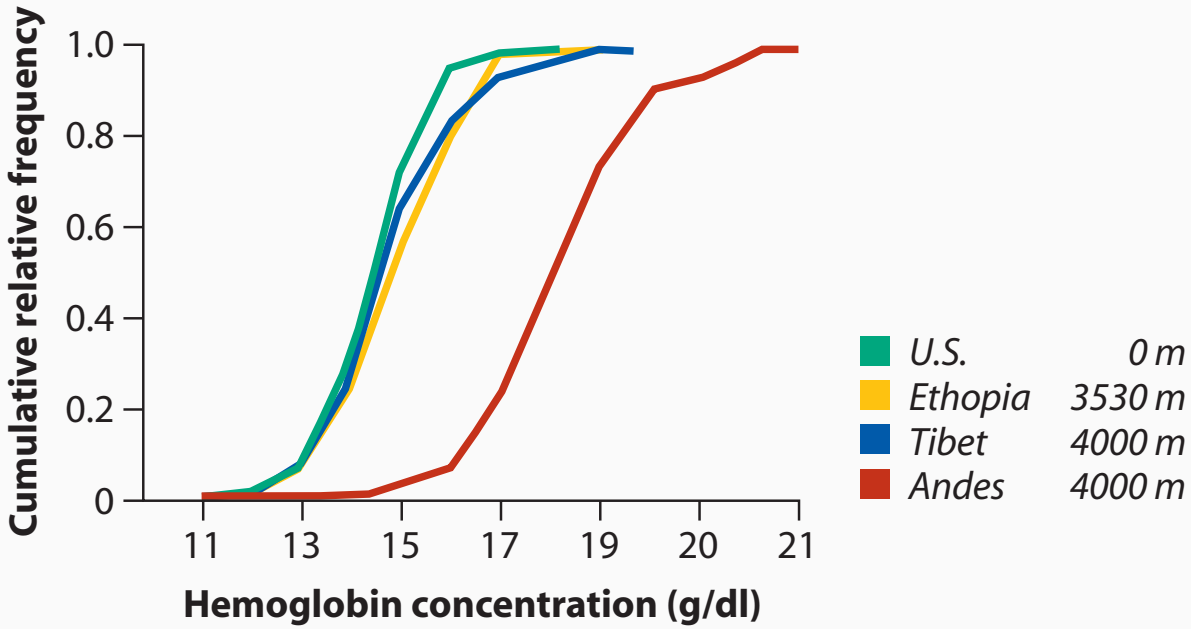
Comparing Numerical Variables

Comparing Histograms across Groups

- Do indigenous peoples who live at high altitudes have physiological attributes that compensate for oxygen deprivation?
- Beall et al. (2002) shed some light; USA (sea-level) versus three high-altitude populations
- Andean males have higher concentrations of hemoglobin but not so Tibetan and Ethiopian males (compared to American males)

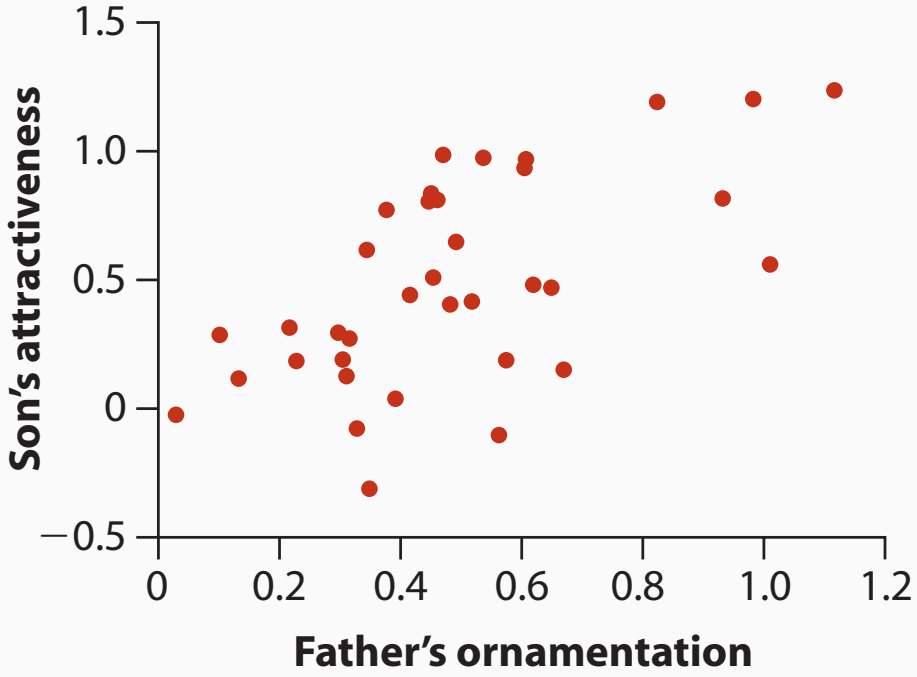


Comparing Cumulative Frequencies across Groups



Displaying Relationships between Numerical Pairs

- What explains bright colors and elaborate courtship displays of the males of many species?
- Brooks (2000) gives us some clues
- Explored how fathers' ornamentation (a composite index of color & brightness) is related to sons' attractiveness (rate of female visits to corralled males, relative to a standard)
- Presumably females are attracted to more ornamented males

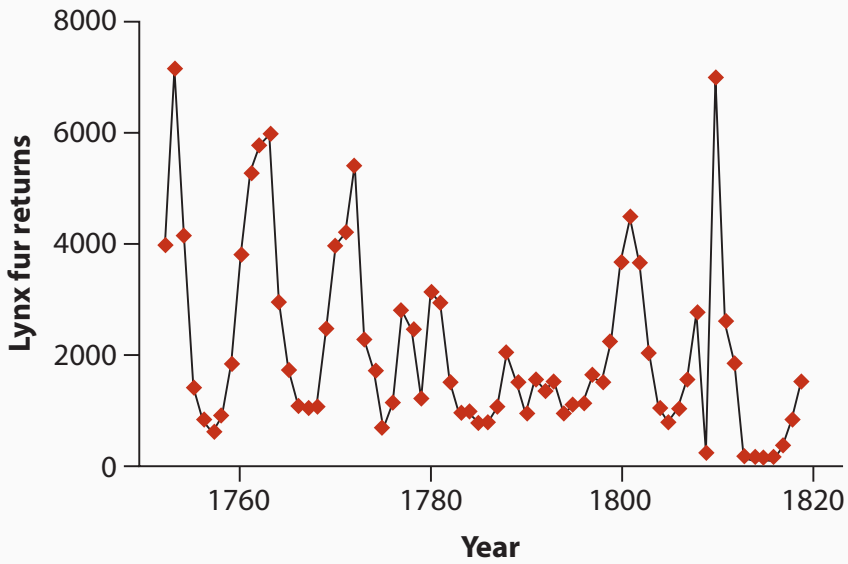


Line Graphs

Definition

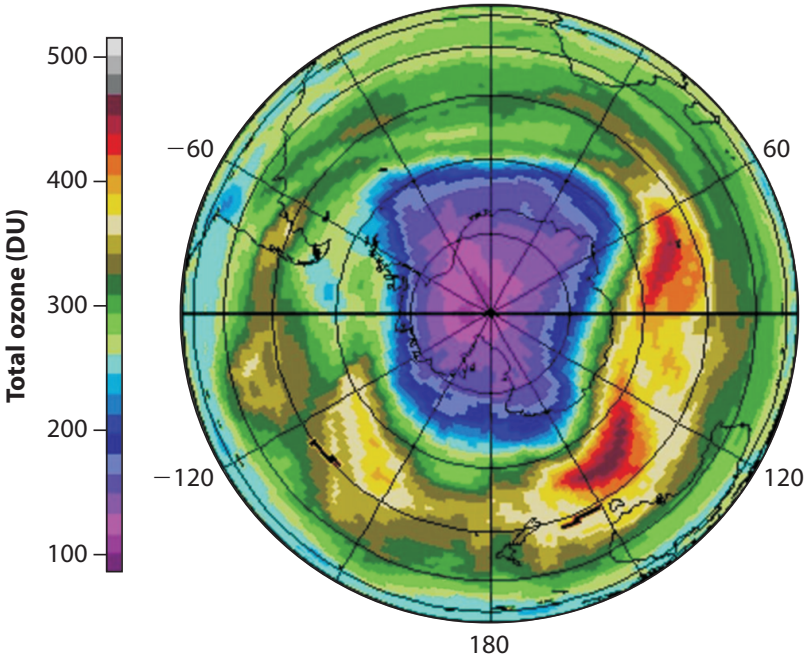
Line graphs connect observations ordered over time (or some other ordered dimension)

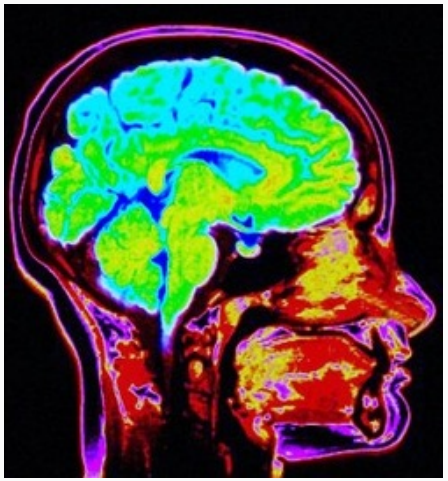
- Lynx pelts turned in at fur trading posts in Canada (1752–1819)
- Line graph shows patterns over time
- Note a cyclical pattern of peaks and troughs
- Note also the steep slopes
- Useful for multiple time series so long as it isn't too cluttered



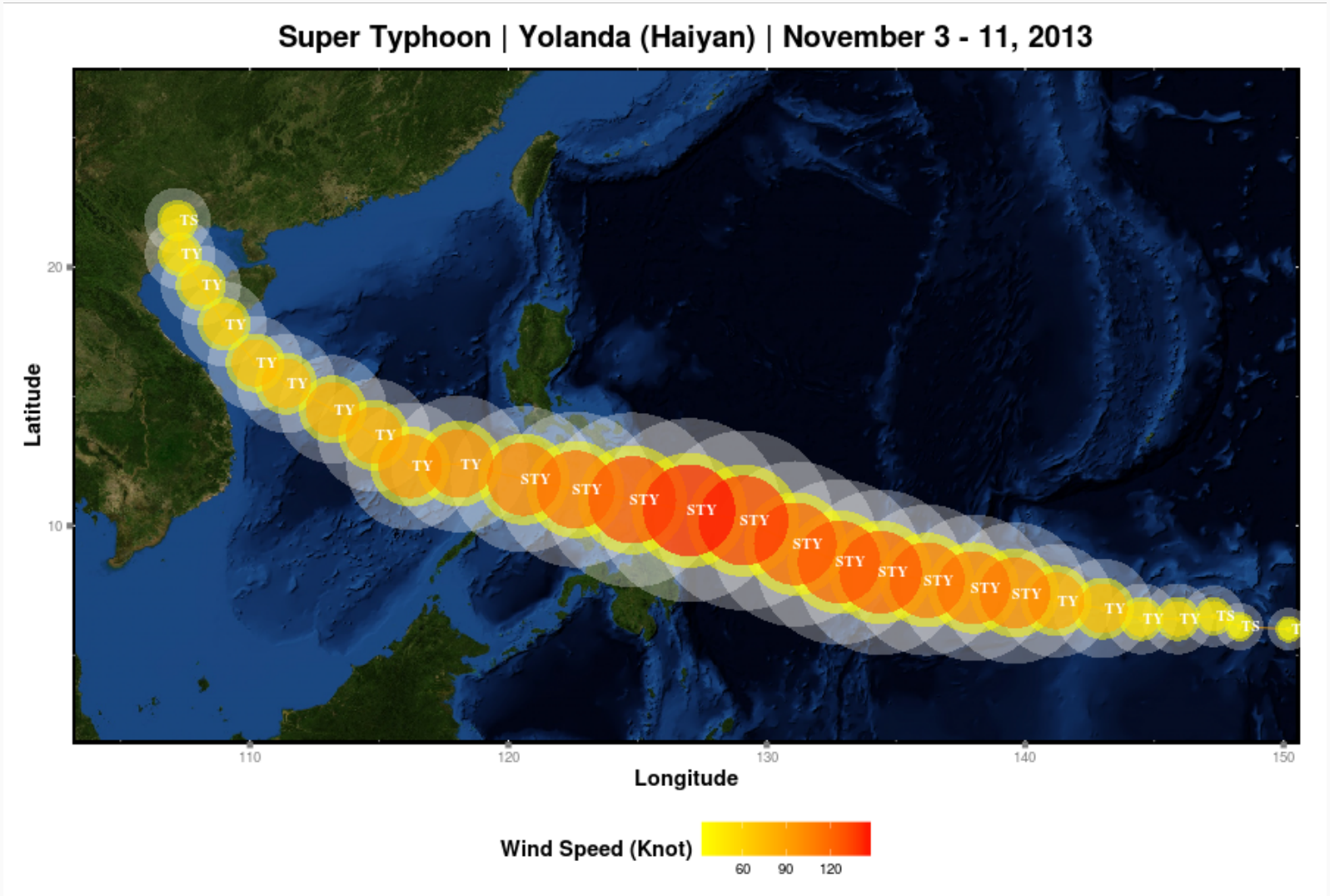
Maps

- Ozone concentrations on October 6, 1987 over the Southern Hemisphere
- Center is the South Pole, outer edge is 15 degrees south of the equator
- Heat Map shows varying levels of Ozone concentrations (note the “hole” above the South Pole)
- Note: Maps can also be a graphic with a heatmap; see here for Brain mapping project



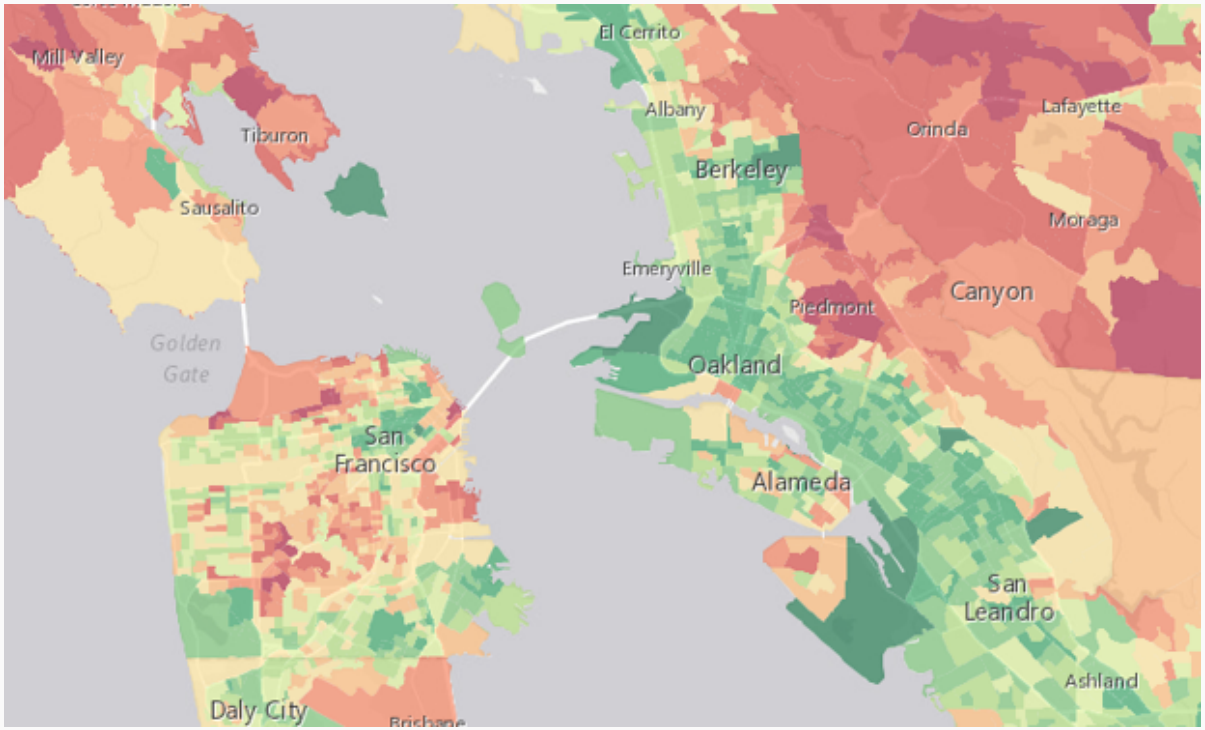


Mapping the Path of Super Typhoon Yolanda (Haiyan)



Source: Analysis with Programming

This map shows carbon emissions from the consumption of goods, with red marking high rates of emissions and green marking low.

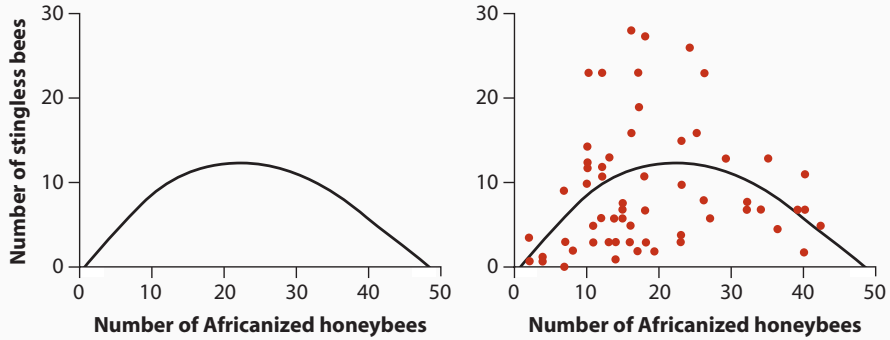


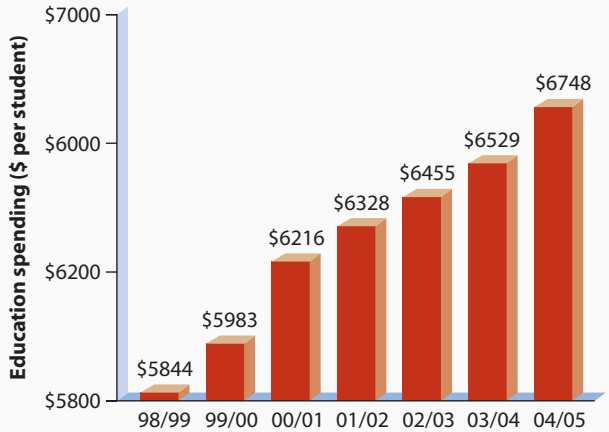
Source: City Carbon Footprint

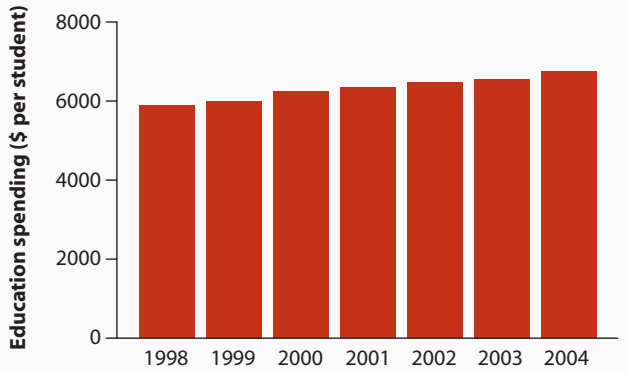
Principles of Effective Displays

Making Effective Displays

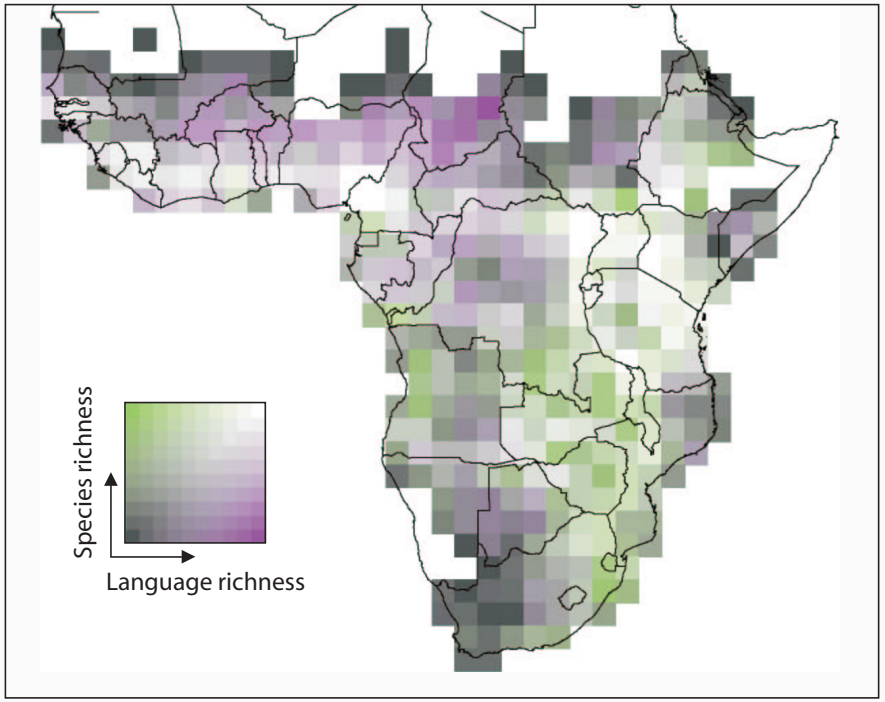
- Show as much data as you can
  - Plot 1 (left) hints at a curvilinear link between Africanized honeybees and stingless bees
  - Adding the actual data points shows more details
- Do not distort magnitudes. y-axis must start at 0
- Minimize chartjunk ... for e.g., three-dimensional bars, shadow effects, etc.
- Avoid jargon for non-technical audience
- Data graphic $\neq$ work of art; must be informative







- Too much data or too complex a plot can defeat the purpose of visualizations
- See this map – simultaneously plots linguistic richness and diversity of bird species
- You could improve this display with a better color scheme, maybe some labeling of Low-Low, Low-High, High-Low, and High-High blocks
- Avoid red-green colors; one-fifth of males cannot distinguish between shades of these colors
- Better yet, if you can avoid colors altogether, do so



WTF Data Visualizations

Selecting your graphic

1. Nominal or Ordinal variable(s)
    - Frequency Table
    - Bar-chart
    - Mosaic plot
2. Continuous or Discrete variable(s)
    - Grouped Frequency table
    - Grouped Histogram
    - Line graph
    - Scatter plot
    - Box-plot (coming soon)
    - Ogive curves (coming soon)
    - Strip charts (coming soon)
    - Violin plots (coming soon)