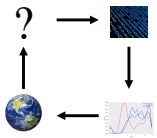
Data Science 100
Review: EDA & Viz & Probability & MLE & PCA

Slides by:
Deb Nolan
deborah_nolan@berkeley.edu



Philosophy of EDA

- Confirm understanding of the data
- Keep an open mind and be willing to find something surprising
- Iterate
  - Uncover new aspects of our data
  - Re-examine our understanding of the data
  - Continue exploration

Where does EDA help?

- Clean data
- Transform variables and derive new variables – put data in format suitable for analysis
- Better understand data/situation (no formal analysis pursued)
- Inform formal analysis – uncover important features that impact the analysis
- Examine formal analysis – explore output from an analysis, e.g., predictions, parameter estimates, model fit

How to Carry out EDA?

- Plot data in multiple ways to get different insights
- Transform variables to symmetrize distributions
- Transform variables to straighten relationships
- Derive new variables
- Consider effect of other variables on distributions & relationships

Connect what you find to the question and context

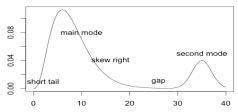
Histograms & Density Curves

- Histogram bar:
  Height * Width = Area = Proportion (or count)
- Similar property for area beneath a density curve

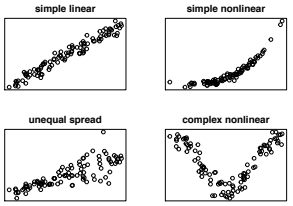
- No longer see individual values (as in rug plot)
- Smooth out values over a bin/region (histogram/density curve)
- Focus on the main features of the distribution
- Caution: Smoothing can reveal or disguise features (boxplot hides modes)

Features to Look for in a Distribution

- Mode(s) - values concentrate around particular points
- Symmetry – skew left, symmetric, skew right distribution of values about center
- Tails - long, short, normal (what expect for normal distribution)
- Gaps - regions where no values obs
- Outliers - unusually large/small value



Plotting Pairs of Variables



Match Data Type to Plot Type

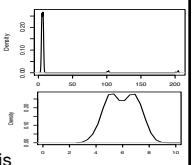
Type	Plot
Numeric –	Histogram or smooth Density curve Box plot or Violin plot Normal quantile plot Few Observations - Rug plot, Dot plot Caution: for discrete values, density curves and box plots may be misleading
Categorical – Counts of categories	Dot chart Bar chart Pie chart (avoid!) Caution for ordinal data, order the bars, dots, … to reflect category order; otherwise order by size

Bivariate Displays

	Numeric	Categorical
Numeric	Scatter plot Smooth scatter Smooth lines and curves Line plot for time	Multiple histograms next to each other, Multiple density curves super-posed, Avoid jiggling and stacking! Side-by-Side boxplots for each category
Categorical		Side-by-side bar plot Overlaid Lines plot Side-by-side dot chart Mosaic plot Avoid stacking!

Choosing the Scale

- Choose axis limit to fill the plotting region
- In necessary,
  - Zoom in to focus on region with bulk of data
  - Make multiple plots of different regions
  - Transform data to improve resolution (TBC)
- Don't change scale mid-axis
- Don't use two different scales for the same axis



Conditioning – Distributions & Relationships in subgroups

- Emphasize the important difference –
- Lines make it easier to see growth in gap
- Placement of one point above the other makes it easier to compare values for different sub-groups
- Plotting Techniques
  - Superpose density curves,
  - Superpose fitted curves and lines from different subgroups
  - Juxtapose scatter plots, histograms; keep same x (and y) scales
  - Use color and plotting symbols to represent additional variables

Perception: Color Guidelines

- 7-10% of males are red-green color blind
- Saturated/colorful colors are hard to look at for a long time. They tend to produce an after-image effect which can be distracting.
- Areas should be rendered with colors of similar luminance (brightness). Lighter colors tend to make areas look larger

Data Type and Color

- Qualitative – Choose a **qualitative** scheme that makes it easy to distinguish between categories
- Quantitative – Choose a color scheme that implies magnitude.
  - Does the data progress from low to high? Use a **sequential** scheme where light colors are for low values
  - Do both low and high value deserve equal emphasis? Use a **diverging** scheme where light colors represent middle values

Perception: Length
Bar plot, Pie chart, Dot chart

- Length is easier to compare than area or volume
- Lengths that fall on a line are easier to compare than lengths on parallel lines, i.e., judgments based on dot charts are easier to make than judgments based on bar plots
- Stacked bar plots and histograms are difficult to read because the base line moves from one bar to the next
- "Jiggled" Line plots where the area between successive lines represent the measurement are very difficult to read because the base line moves.

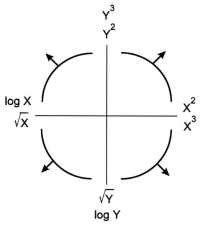
Why Transform Variables?

- Reveal distribution of most of the observations (otherwise much of the data is squashed in a small region)
- Numerical summaries of symmetric distribution are better summaries of data
- Choose a transformation that's simple and easily interpreted, e.g., a power of 2, 3, ½, 0 (log), -1
- Easier to uncover the form of the relationship if we can transform it to linear relationship
- Linear relationships are particularly simple to interpret & fit

Power Transformation

- Preserve order of values
- Effective when max / min > 5
- Sometimes add a shift before transform
- Ratio of hinges can help select a transformation

$$\frac{Upper\ Quartile - Median}{Median - Lower\ Quartile} \approx 1$$



Add Context

- Label axes, including units
- Add Reference lines and markers for important values
- Label points of unusual/interesting observations
- Include captions that describe data, how plotted, and describe important features

Large n (number of records)

- Use heat map or hexbin plot or transparency
- Add smooth curve that takes local averages to see the conditional center, i.e., average y in a neighborhood of x
- Large p (number of variables) – PCA to reduce dimensionality

Principle 1. Reveal the Data

- Choose scale appropriately
- Avoid having other graph elements interfere with data
- Use visually prominent symbols
- Eliminate superfluous material, aka chart junk
- Avoid over-plotting (jitter, shrink plotting symbol, transparency, smooth curve)

Principle 2. Facilitate Comparisons

- Put Juxtaposed plots on same scale
- Make it easy to distinguish elements of *superposed* plots, e.g. color, line type
- Avoid Stacking and Jiggling the baseline
- Avoid angles, extra dimensions (e.g., areas rather than lines)
- Maintain the visual metaphor, i.e., with rectangles, area should correspond to value
- Length easier to compare than angle, area, volume

Principle 3. Make a plot information rich

- Describe what you see in the Caption
- Add context with Reference Markers (lines and points) including text
- Add Legends and Labels
- Use color and plotting symbols to add more information
- Plot the same thing more than once in different ways/scales
- Reduce clutter

Probability, Random Variables, and Distributions

Formalize Rules of Probability

Ω = set of all possible outcomes from the chance process

A = a collection of outcomes, AKA an event

B = another collection of outcomes

- $P(\Omega) = 1$
- $0 \leq P(A) \leq 1$
- If A and B disjoint, then P(A or B) = P(A) + P(B)

From these 3 Rules

1. $P(\Omega) = 1$
2. $0 \leq P(A) \leq 1$
3. If A and B disjoint, then P(A or B) = P(A) + P(B)

- If B is contained in A, then $P(B) \leq P(A)$
- $P(A^c) = 1 - P(A)$
- P(A or B) = P(A) + P(B) – P(A and B)

Conditional Probability

- P(A|B) means the chance of A given B occurs
- P(A|B) = P(A and B)/P(B)

- Independence of A and B means P(A|B) = P(A)

- P(A and B and C) = P(A)P(B|A)P(C|A,B)
- If independent, then P(A and B and C) = P(A)P(B)P(C)
- Bayes RULE: P(A|B) = P(A)P(B|A)/ P(B)

Ways to solve probability problems

- Equally likely outcomes –
  P(A) = # outcomes in A / # outcomes in Ω
- Complement rule – $P(A) = 1 - P(A^c)$
- Symmetry – A and B are equivalent events so P(A) = P(B)
- Sequence of outcomes
  P(A and B and C) = P(A) P(B|A) P(C|A,B)

Probability Distribution

- Probability Distribution Table:
  - All Possible Values/Outcomes
  - Chance of each value/outcome

- Function or Rule for calculating the chance of each possible outcome

Random Variable

Numeric value for outcome

Use capital letter to denote the outcome, usually from the end of the alphabet

Example

U = number face up on roll of a fair die

X = 1 if 1st card drawn is diamond, 0 otherwise

Discrete Uniform Distribution

U ~ discrete uniform(k,m)

P(j) = 1/(m-k+1) for $k \leq j \leq m$
= 0 otherwise

Bernoulli(p) Distribution

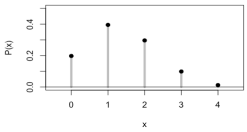
B = indicator for a success

P(B = 1) = p and P(B=0) = 1-p for $0 < p < 1$

Bernoulli(p) where *p* is the chance of success

Binomial(n,p) Distribution

- ***n*** trials
- *p* chance of success on a trial
- trials are independent
- Observe the number of successes in n trials



$$\mathbb{P}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$\binom{n}{k}$ is $n!/k!(n-k)!$

Poisson(λ)

- Count for rare events
- λ is the rate
- For $x = 0,1,2,\ldots$ $\quad P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$

Geometric(p)

- Count the number of failures until first success
- Trials are independent with the same probability of success
- P(k) = P(k failures followed by a success)
  = $p(1-p)^k$ for k = 0, 1, 2, …
- Alternative: count the number of trials until first success
- P(k) = P(k trials to first success)
  = $p(1-p)^{k-1}$ for k = 1, 2, …

Summarize a Distribution

- Recall, we summarize a data distribution with its average (center) and spread (SD)
- We can similarly summarize a probability distribution with its expected value and SD

$$E(X) = \sum_{i=1}^{m} x_i p_i$$

$$Var(X) = \sum_{i=1}^{m} (x_i - E(X))^2 p_i$$

$$SD(X) = \sqrt{Var(X)}$$

Bernoulli(p)

$$E(B) = 0(1-p) + 1p = p$$

$$Var(B) = (0-p)^2(1-p) + (1-p)^2 p$$

$$Var(B) = p(1-p)$$

Properties of Expected Value

$$E(aX+b) = aE(X) + b$$

$$E(aX+b) = \sum_{i=1}^{m} (ax_i + b) p_i$$

$$E(X+Y) = E(X) + E(Y)$$

Properties of Variance

$$Var(aX+b) = a^2 Var(X)$$

$$Var(aX+b) = \sum_{i=1}^{m} (ax_i + b - (aE(x)+b))^2 p_i$$

$$Var(X+Y) = Var(X) + Var(Y), \text{ if independent}$$

X ~Binomial(n,p) distribution

- X = $B_1+B_2+\ldots+B_n$ where B_i~ Bernoulli(p)
- $E(B_i) = p$
- $E(X) = E(B_1+\ldots+B_n) = np$

Maximum Likelihood

- Named probability distributions are defined in terms of parameters
- Given the data, we maximize the likelihood of the data over the possible parameter values
- In practice,
  - We might not be able to analytically solve for the parameters
  - We might not have the complete data
  - Computational considerations can be important for accuracy and efficiency

Maximum Likelihood for Binomial(200,p)

- The likelihood can be viewed as a function of *p* given the data.
- Suppose we saw 25 successes in 200 trials, the likelihood is
  $L(p) = C(200,25)\, p^{25}(1-p)^{200-25}$
- Find the *p* that maximizes the likelihood for our data.
- Often easier to maximize the log-likelihood:
  $\log(L(p)) = C(200,25) + 25\log(p) + (200-25)\log(1-p)$
- Differentiate: $25/p - 175/(1-p)$
- Set to 0 and solve for p: 25/200

Find MLE for p for a Geometric(p)

n observations: $k_1, \ldots, k_n$

$P(X_1=k_1, X_2=k_2, \ldots, X_n = k_n)$

$= P(X_1=k_1)P(X_2=k_2) \cdots P(X_n = k_n)$ independence

$= p(1-p)^{k1}\ p(1-p)^{k2} \ \ldots\ p(1-p)^{kn}$ geometric(p)

$= p^n(1-p)^{k1+k2+\ldots+kn}$ The likelihood function

log likelihood: $l(p) = \log(L(p)) = n\log(p) + (k_1+\ldots+k_n)\log(1-p)$

Differentiate wrt p $n/p - (k_1+\ldots+k_n)/(1-p)$

Set to 0 and solve: $\hat{p} = n/(n+ k_1+\ldots+k_n)$

Logistic Regression y_i ~ Bernoulli(π_i)

$$\pi_i = \frac{1}{1+\exp(-\mathbf{x}_i^t\boldsymbol{\beta})}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^t\boldsymbol{\beta}.$$

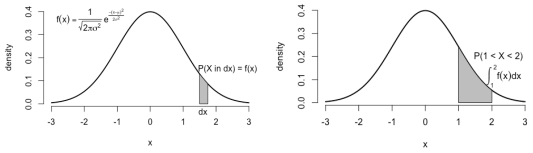
Link Function to connect explanatory variables to the mean

Maximum Likelihood

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, y_1, \ldots, y_n) &= \prod_{i=1}^{n} \pi_i (1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i) \\ &= \prod_{i=1}^{n} \left(\exp(\mathbf{x}_i^t \boldsymbol{\beta})\right)^{y_i} \frac{1}{1+\exp(\mathbf{x}_i^t \boldsymbol{\beta})}\end{aligned}$$

$$l(\boldsymbol{\beta}, y_1, \ldots, y_n) = \sum_{i=1}^{n} [y_i \mathbf{x}_i^t \boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^t \boldsymbol{\beta}))]$$

Normal Distribution:
P(X in dx) = f(x)dx here f is a density fcn
Find probabilities by integration



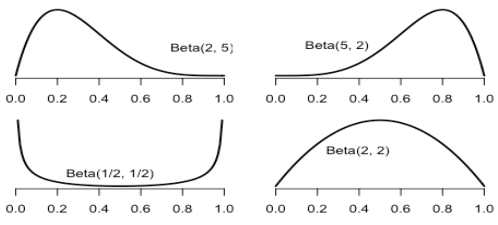
Continuous Uniform Distribution

- Uniform(a, b) distribution
- We say the random variable X has a Uniform(0, 1) distribution, if
  P(c < X < d) = (d-c)/(b-a) for any c & d in (a,b)
- f(x) = 1/(b-a) for x in (a,b)

Beta(α,β)

- The Beta distribution is for random variables in (0, 1)
- The Beta(1,1) is the Uniform(0,1) distribution
- The Beta includes symmetric, skewed, U-shaped distributions
- The probability density function is
  $f(x) = B(\alpha,\beta)\, x^{\alpha-1}(1-x)^{\beta-1}$ for x in (0, 1)
  where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$

The Family of Beta Distributions



MLE for α for sample from Beta(α,1)

- The likelihood
$$L(\alpha) = \prod_{i=1}^{n} \frac{1}{\alpha} x_i^{\alpha-1}(1-x_i)^0$$
$$l(\alpha) = -n\log(\alpha) + (\alpha-1)\sum \log(x_i)$$
- Differentiate
$$l'(\alpha) = -\frac{n}{\alpha} + \sum \log(x_i)$$
- Set to 0 and solve
$$\hat{\alpha} = \frac{1}{\frac{1}{n}\sum \log(x_i)}$$

MLE for μ for sample from Normal(μ,1)

- The likelihood $L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2}$

$$l(\mu) = -\frac{n}{2}\log(\pi) - \frac{1}{2}\sum (x_i - \mu)^2$$

- Differentiate $l'(\mu) = \sum (x_i - \mu)$
- Set to 0 and solve $\hat{\mu} = \bar{x}$

Principle Components

- When we have a large number of features, there is often multi-collinearity
  - Design doesn't fill the space
  - One (or more) variable(s) highly correlated with a combination of other variables
- Some times we have more features than observations
  - Design is over-determined
- Create a few new features that are combinations of original features and that preserve as much information as possible

Principal Component Idea

- Data X = [x_1, x_2, …, x_p]
  - x_j is a feature with n values (one for each record)
- Standardize data Z = [z_1, z_2, …, z_p]
  - z_j has mean 0 and variance 1
- Transform to new basis W = ZA = [w_1, …, w_p]
  - w_1 accounts for max collective variation in Z
  - w_2 orthogonal to w_1 and accounts for max remaining variation, etc.
- Reduce to the first few w_j

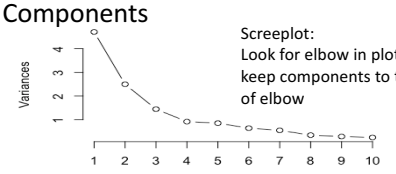
Properties of the new basis

- w_1 is the first principal component
- The w_j are orthogonal
- w_j has variance λ_j
- The eigenvalues are decreasing $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$
- The sum of the eigenvalues is p, which equals the sum of the variances of z_j

How to use this information

- Collinearity occurs when some of the basis vectors are very short (dimensions are nearly collapsed)
- The relative size of variances (the eigenvalues) serves as indicator for collinearity
- The loadings (the eigenvectors) can reveal important relationships among the features
- The new basis can reveal clusters in the data

Rules of Thumb for Selecting Components



Screeplot:
Look for elbow in plot and keep components to the left of elbow

Choose components such that at least 85% of variance captured

Ratio of largest to smallest variance > 10 indicates collinearity