Machine Learning for Business

Module 4: Data exploration and preparation

Day 2, 13.00 – 16.00

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

BIG DATA EXPERIENCE CENTER

Module 4 Overview

- Sampling
- Distribution
- Abnormalities and outliers

BIG DATA EXPERIENCE

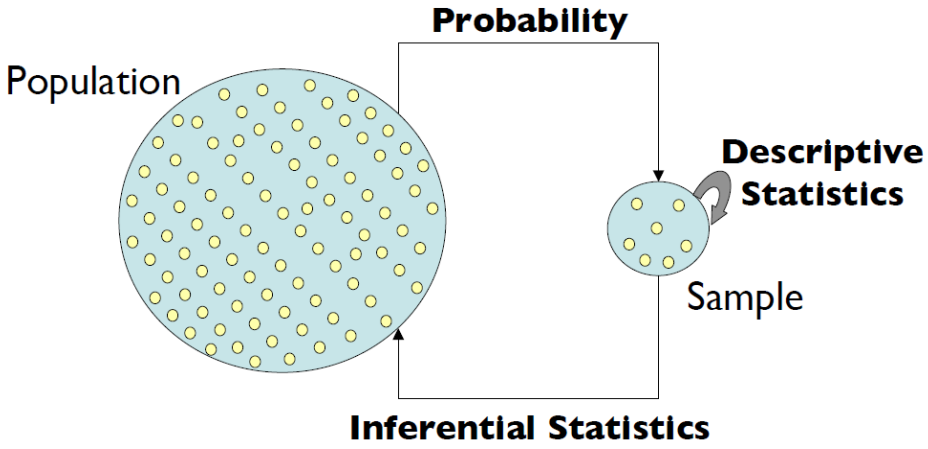
Sampling

Part 1

g·able

BIG DATA EXPERIENCE

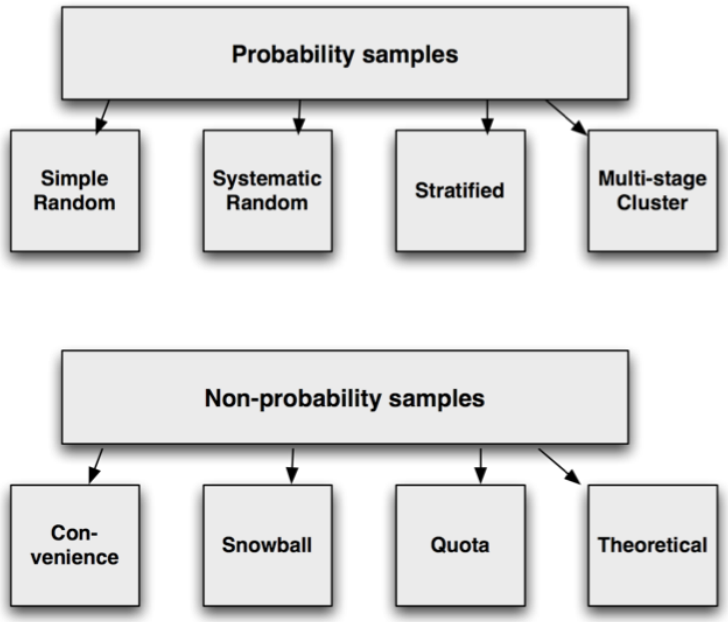
Central Dogma in Statistics



g·able

BIG DATA EXPERIENCE

Sampling



g·able

BIG DATA EXPERIENCE

Sampling in R

- Sampling the indices
  - Generate sample indices with “sample”
  - Use indices to subset the data frame
  - (optional) Obtain the remaining data frame
- Sampling using dplyr functions
  - Obtain the random sample of the data frame

g·able

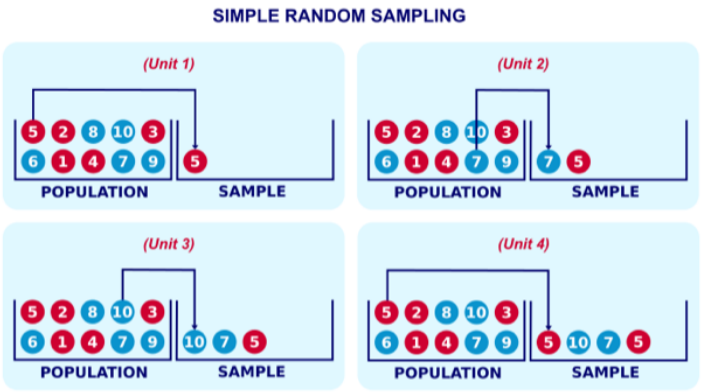
R: Load data
bank-data.csv

```
bankData <- read.csv("bank-data.csv",sep=";")
summary(bankData)
```

```
      age                job          marital
 Min.   :18.00   blue-collar:9732   divorced: 5207
 1st Qu.:33.00   management :9458   married :27214
 Median :39.00   technician :7597   single  :12790
 Mean   :40.94   admin.     :5171
 3rd Qu.:48.00   services   :4154
 Max.   :95.00   retired    :2264
                 (Other)    :6835
```

Simple Random Sampling

- Randomly select individual from populations
- All individuals in the population are assumed to have equal chance



Sampling error

- The difference between a sample statistic (e.g. $\bar{x}$) and its corresponding population parameter (e.g. μ).
- This error is inherent in
  - the sampling process (since sample is only part of the population)
  - The choice of statistics (since a statistics is computed based on the sample).

R: Simple random sampling With specified number (dplyr)

```
sample_n(bankData,5)
```

	age	job	marital	education	default	balance	housing
32898	34	blue-collar	married	primary	no	3986	yes
43860	49	admin.	single	tertiary	no	228	yes
31307	28	student	single	tertiary	no	111	no
39726	39	technician	married	tertiary	no	437	no
22003	30	technician	single	secondary	no	790	no

	loan	contact	day	month	duration	campaign	pdays	previous
32898	no	cellular	17	apr	345	2	-1	0
43860	no	cellular	2	jun	407	1	92	1
31307	no	cellular	10	mar	70	2	-1	0
39726	no	telephone	27	may	186	2	-1	0
22003	no	cellular	20	aug	696	3	-1	0

R: Simple random sampling
With specified fraction (dplyr)

```
dim(sample_frac(bankData, 0.1))
```

```
[1] 4521   17
```

R: Sampling error

```
bankData %>% group_by(education) %>%
    summarise(age = mean(age))
bankData %>% sample_n(100) %>%
    group_by(education) %>%
    summarise(sampled_age = mean(age))
```

Full

```
# A tibble: 4 x 2
  education      age
      <fctr>    <dbl>
1   primary 45.86557
2 secondary 39.96427
3  tertiary 39.59364
4   unknown 44.51050
```

100

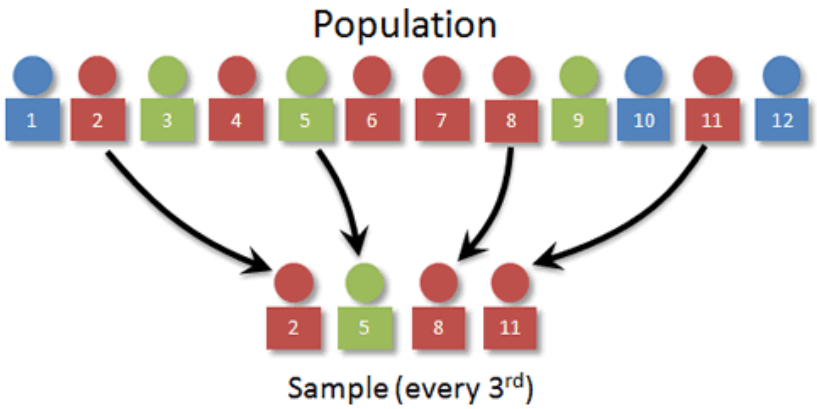
```
# A tibble: 4 x 2
  education sampled_age
      <fctr>       <dbl>
1   primary    44.27778
2 secondary    41.14583
3  tertiary    38.45161
4   unknown    37.00000
```

10000

```
# A tibble: 4 x 2
  education sampled_age
      <fctr>       <dbl>
1   primary    45.61365
2 secondary    39.99942
3  tertiary    39.47412
4   unknown    44.45592
```

Systematic Random Sampling

• Have a fixed rule (interval) to select the sample



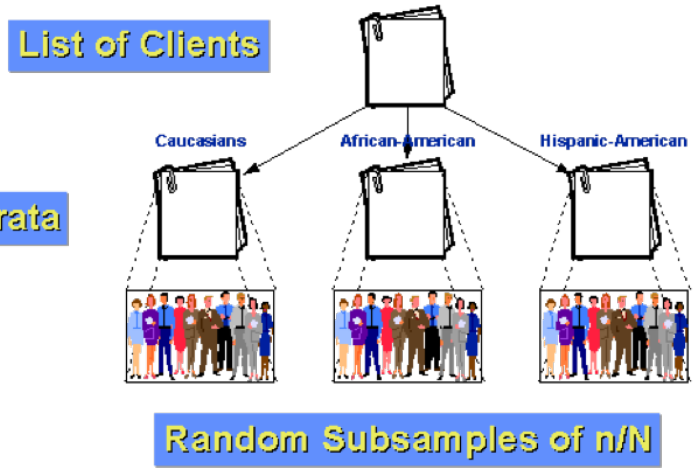
R: Systematic random sampling

```
dim(bankData[seq(1,45211,2),])
```

```
[1] 22606    17
```

Stratified Sampling

- With stratified sampling, the investigator divides the population into separate groups, called strata.
- Then, a probability sample (often a simple random sample) is drawn from each group.

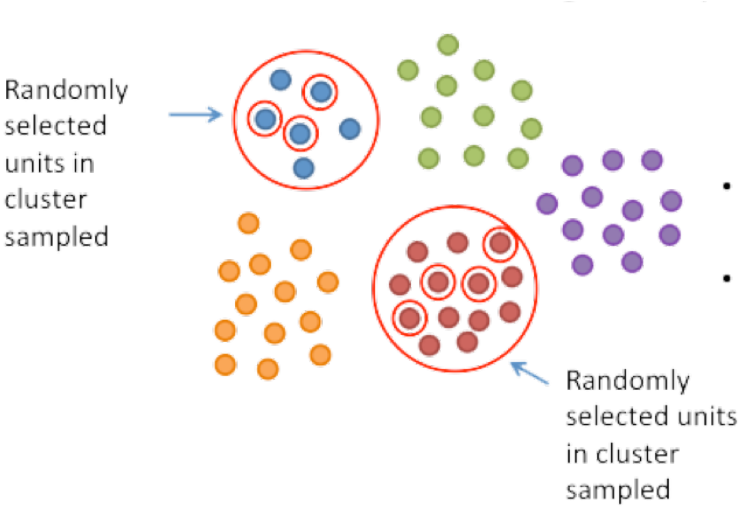


R: Stratified random sampling

```
bankData %>% group_by(y) %>% sample_n(3)
```

```
# A tibble: 6 x 17
# Groups:   y [2]
    age          job marital education default balance housing
  <int>       <fctr>  <fctr>    <fctr>  <fctr>   <int>  <fctr>
1    38   management married  tertiary      no    1021      no
2    58   management married  tertiary      no    2764      no
3    42 entrepreneur married  tertiary      no     -14     yes
4    56  blue-collar married   primary      no    8163      no
5    35   management married  tertiary      no      43     yes
6    31       admin. married secondary      no      35      no
# ... with 10 more variables: loan <fctr>, contact <fctr>,
#   day <int>, month <fctr>, duration <int>, campaign <int>,
#   pdays <int>, previous <int>, poutcome <fctr>, y <fctr>
```

Multistage Sampling



- Natural grouping within population
- Units within groups sampled

R: multistage sampling

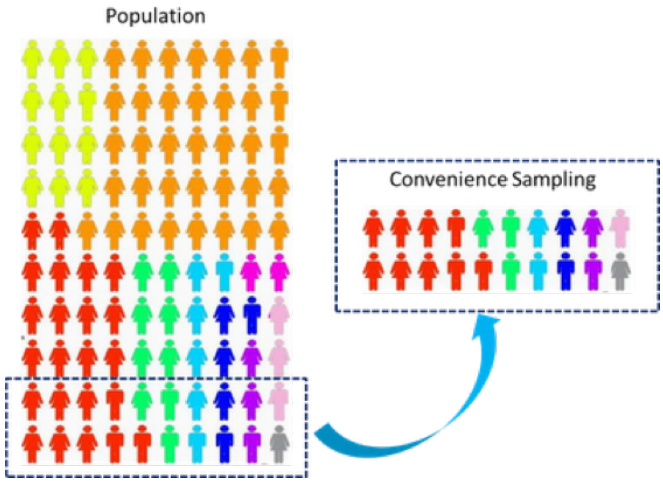
```
res <- kmeans(mtcars,5)
data1 <- mtcars
data1$cluster <- res$cluster
data1 %>% group_by(cyl) %>% sample_n(2)
```

```
# A tibble: 6 x 12
# Groups:   cyl [3]
    mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  33.9     4  71.1    65  4.22 1.835 19.90     1     1     4
2  22.8     4 108.0    93  3.85 2.320 18.61     1     1     4
3  17.8     6 167.6   123  3.92 3.440 18.90     1     0     4
4  21.0     6 160.0   110  3.90 2.875 17.02     0     1     4
5  15.5     8 318.0   150  2.76 3.520 16.87     0     0     3
6  14.3     8 360.0   245  3.21 3.570 15.84     0     0     3
# ... with 2 more variables: carb <dbl>, cluster <int>
```

BIG DATA EXPERIENCE

Convenience Sampling

- Members of the population are chosen based on their relative ease of access.
- To sample friends, co-workers, or shoppers at a single mall, are all examples of convenience sampling.

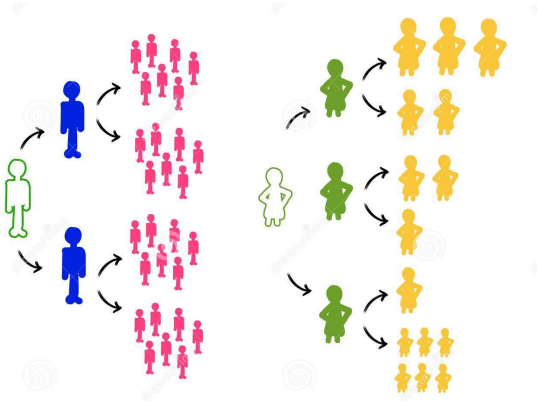


g·able

BIG DATA EXPERIENCE

Snowball Sampling

- The first respondent refers an acquaintance. The friend also refers a friend, and so on.
- Such samples are biased because they give people with more social connections an unknown but higher chance of selection.



g·able

Quota Sampling

- A quota is established (e.g. 65% women) and investigators are free to choose any respondent they wish as long as the quota is met

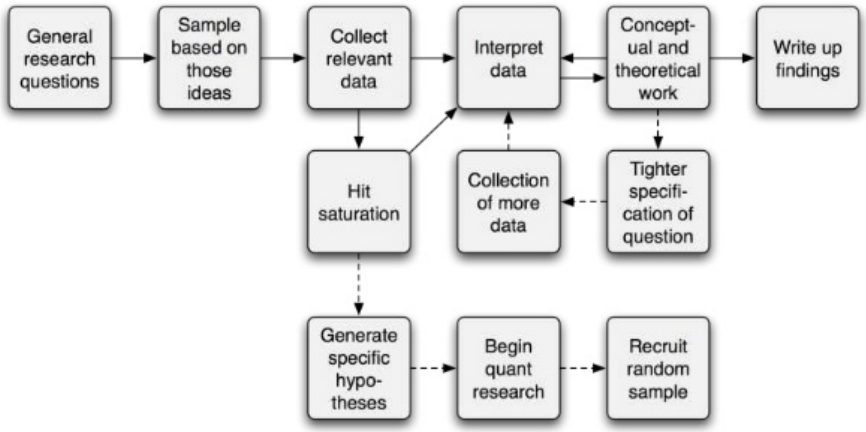
Symbol	Age Group	No.
	11-21 Years	11
	22-31 Years	16
	32-41 Years	15
	42-51 Years	18
Total	11-51 Years	60

Quota vs Stratified

- In Stratified Sampling, selection of subject is random. Call-backs are used to get that particular subject.
- Stratified sampling without call-backs may not, in practice, be much different from quota sampling.
- In Quota Sampling, interviewer selects first available subject who meets criteria: is a convenience sample.
- Highly controlled quota sampling uses probability sampling down to the last block or telephone exchange

Theoretical Sampling

- Theoretical sampling is a process of data collection for generating theory whereby the analyst jointly collects codes and analyses data and decides what data to collect next and where to find them, in order to develop a theory as it emerges.



Sample size

- Heterogeneity: need larger sample to study more diverse population
- Desired precision: need larger sample to get smaller error
- Sampling design: smaller if stratified, larger if cluster
- Nature of analysis: complex multivariate statistics need larger samples
- Accuracy of sample depends upon sample size, not ratio of sample to population

BIG DATA EXPERIENCE

g·able

Sampling in practice

- Often a non-random selection of basic sampling frame (city, organization etc.)
- Fit between sampling frame and goals must be evaluated
- Sampling frame as a concept is relevant to all kinds of studies (including nonprobability)
- Nonprobability sampling means you cannot generalize beyond the sample
- Probability sampling means you can generalize to the population defined by the sampling frame

BIG DATA EXPERIENCE

g·able

Distribution

Part 2

BIG DATA EXPERIENCE

g·able

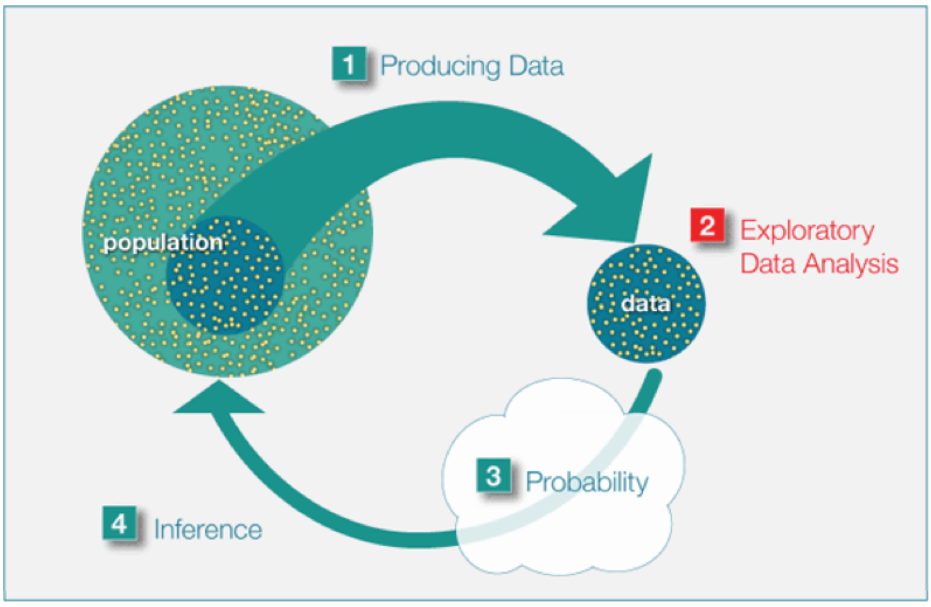
Exploratory Data Analysis

- Exploratory data analysis or “EDA” is a critical step in analyzing the data
- The main reasons are
  - detection of mistakes, outliers or abnormalities
  - checking of assumptions
  - preliminary selection of appropriate models
  - determining relationships among the explanatory variables
  - assessing the direction and rough size of relationships between explanatory and outcome variable

BIG DATA EXPERIENCE

g·able

Exploratory Data Analysis



Data format

- Data from either experiments or operations are generally collected in databases (e.g. spreadsheet)
- One row per record and one column for each identifiers, outcome variables, and explanatory variables
- Each column contains the numerical value of a particular quantitative variable (aka measure) or the levels for a categorical variable (aka dimension)

Why EDA?

- People are not very good at looking at column of numbers or a whole spreadsheet
- Why? Surely you know…tedious, boring, and overwhelming
- EDA techniques have been devised as an aid in this situation
- Most of these techniques work by reducing dimension of data to make some aspects clearer

Types of EDA

- Graphical or Non-graphical
  - Non-graphical methods usually involve with calculation of summary statistics
  - Graphical methods obviously summarize the data in a diagrammatic or pictorial way
- Univariate or Multivariate
  - Univariate methods look at one variable (column) at a time while multivariate methods look at two or more variables at a time
  - Usually it is a good idea to perform univariate EDA first for each of a component of multivariate EDA

Four basic types of EDA

This lecture

- Univariate non-graphical
- Univariate graphical

Next lecture

- Multivariate non-graphical
- Multivariate graphical

BIG DATA EXPERIENCE

g·able

Univariate non-graphical EDA

- This is to measure certain characteristics (e.g. age, gender, speed at a task, or response to a stimulus) of data of all subjects/records
- We should think of measurements as representations of a "sample distribution", which in turn more or less representing the "population distribution"
- The goal is to better understand the "sample distribution" and make some conclusion about the "population distribution"

BIG DATA EXPERIENCE

g·able

Univariate analysis Categorical (factors/strings)

- Use frequency table
- Inspect the prior probabilities, aka proportions
- Identify types of categories: nominal vs ordinal
- Identify the ID columns. These columns will need to be removed before analysis. Why?
- Take note at minority categories

R: Use summary to look at frequency table

```
bankData <- read.csv("bank-data.csv", sep = ";")
summary(bankData)
```

```
#       age                  job           marital
# Min.   :18.00   blue-collar:9732   divorced: 5207
# 1st Qu.:33.00   management :9458   married :27214
# Median :39.00   technician :7597   single  :12790
# Mean   :40.94   admin.     :5171
# 3rd Qu.:48.00   services   :4154
# Max.   :95.00   retired    :2264
#                 (Other)    :6835
# ........
```

R: Create frequency table with dplyr

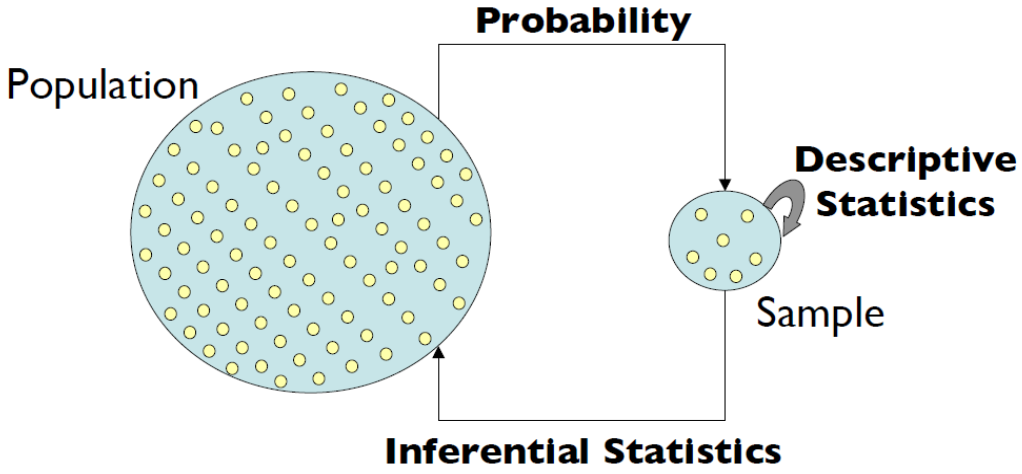
```
library(dplyr)
bankData %>%
   group_by(y) %>%
   summarise(n = n())
## A tibble: 2 x 2
#         y     n
#    <fctr> <int>
# 1      no 39922
# 2     yes  5289
```

Univariate non-graphical EDA

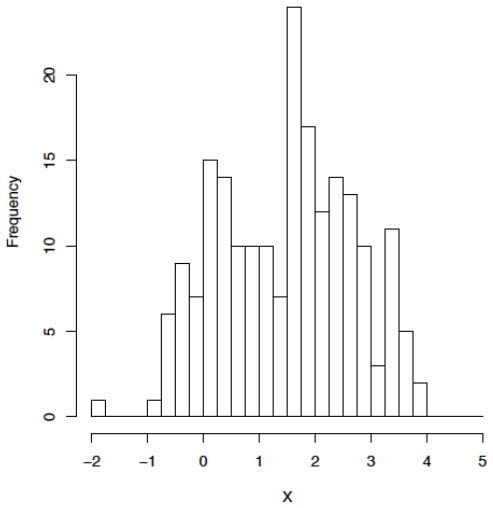
Quantitative data

- Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.
- The characteristics of the population distribution of a quantitative variable are its center, spread, modality (number of peaks in the pdf), shape (including “heaviness of the tails"), and outliers.

Sample statistics



Histogram



Central tendency
Spread
Skewness
Etc.

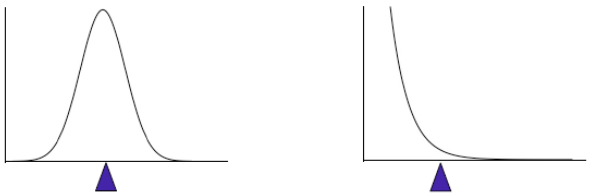
What can you see from this histogram?

Central tendency
Mean

1. The Mean

To calculate the average $\overline{x}$ of a set of observations, add their value and divide by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Central tendency
Median

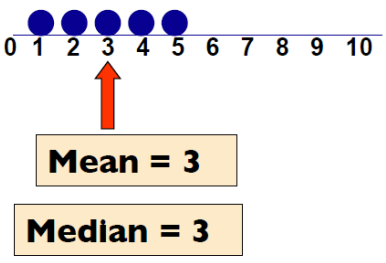
- **Median** – the exact middle value
- **Calculation**:
  - If there are an odd number of observations, find the middle value
  - If there are an even number of observations, find the middle two values and average them
- **Example**

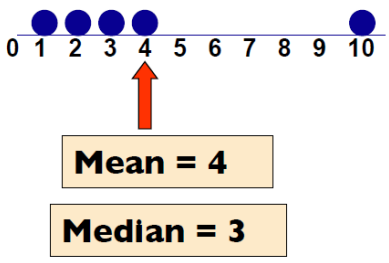
  Some data:
  Age of participants: 17 19 21 22 23 23 23 38

  Median = (22+23)/2 = 22.5

Central tendency
Which location measure is the best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers





BIG DATA EXPERIENCE

Scale: Variance

- Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}$$

g·able

BIG DATA EXPERIENCE

Why squared deviations?

- Adding deviations will yield a sum of ?
- Absolute values do not have nice mathematical properties (non-linear)
- Squares eliminate the negatives

- Results:
  - Increasing contribution to the variance as you go farther from the mean

g·able

Scale: Variance

- Variance is somewhat arbitrary
- What does it mean to have a variance of 8.9? Or 1.5? Or 1245.34? Or 0.00001?
- Nothing. But if you could "standardize" that value, you could talk or compare about any variance (i.e. deviation) in equivalent terms
- Standard deviations are simply the square root of the variance

Scale: Standard deviation

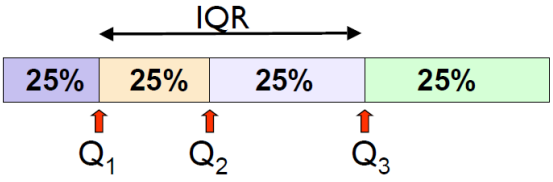
$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

1. Score (in the units that are meaningful)
2. Mean
3. Each score's deviation from the mean
4. Square that deviation
5. Sum all the squared deviations (Sum of Squares)
6. Divide by n-1
7. Square root – now the value is in the units we started with!!!

R: mean, median and sd

```
bankData %>%
   group_by(education) %>%
   summarise(n = n(),
             mean = mean(balance),
             median = median(balance),
             sd = sd(balance))
## A tibble: 4 x 5
#   education     n     mean median       sd
#       <fctr> <int>    <dbl>  <dbl>    <dbl>
# 1   primary  6851 1250.950    403 2690.744
# 2 secondary 23202 1154.881    392 2558.257
# 3  tertiary 13301 1758.416    577 3839.088
# 4   unknown  1857 1526.754    568 3152.228
```

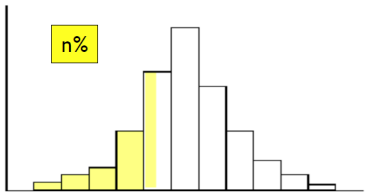
Scale: Quartiles and IQR



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Percentiles (aka Quantiles)

In general the **n^{th} percentile** is a value such that n% of the observations fall at or below or it



Q_1 = 25th percentile

Median = 50th percentile

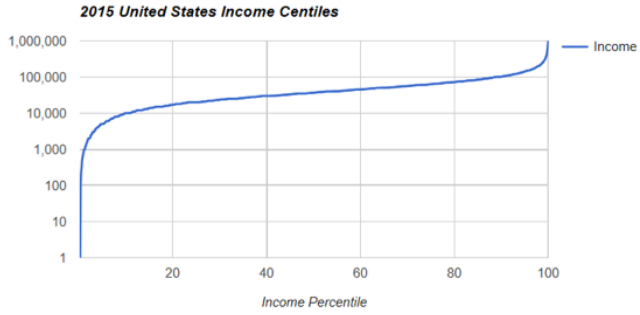
Q_2 = 75th percentile

R: Quartile & IQR

```
bankData %>%
   group_by(education) %>%
   summarise(Q1 = quantile(balance,0.25),
             mean = mean(balance),
             median = median(balance),
             Q3 = quantile(balance,0.75),
             IQR = IQR(balance))
## A tibble: 4 x 6
#   education    Q1     mean median    Q3   IQR
#      <fctr> <dbl>    <dbl>  <dbl> <dbl> <dbl>
# 1   primary    61 1250.950    403  1390  1329
# 2 secondary    55 1154.881    392  1234  1179
# 3  tertiary   104 1758.416    577  1804  1700
# 4   unknown   106 1526.754    568  1699  1593
```

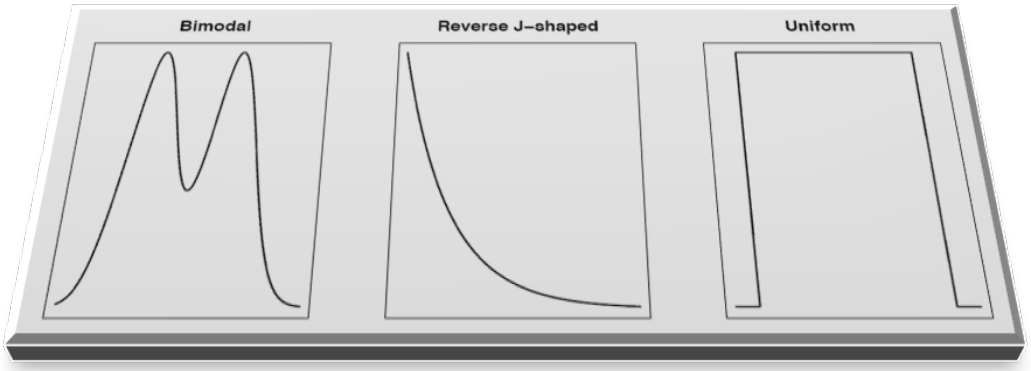
Outliers

- Extreme observations in the data
- Example
  - US Median Household Income: $57,616
  - US Mean Household Income: $72,641
  - But
    - Q1 = $20,000
    - Q3 = $64,200
- What happen?



Common distribution shapes

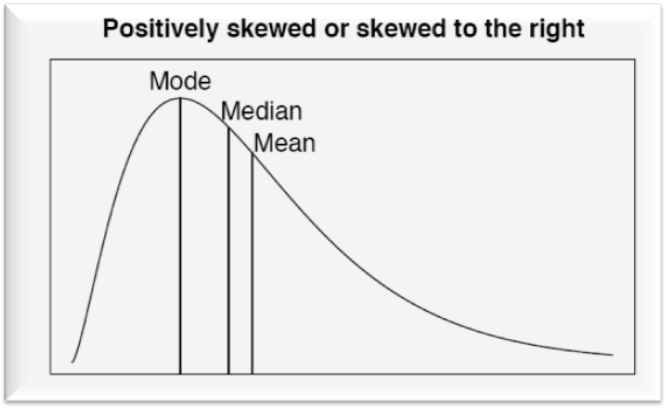
Other distribution shapes



Skewness I

Positively skewed

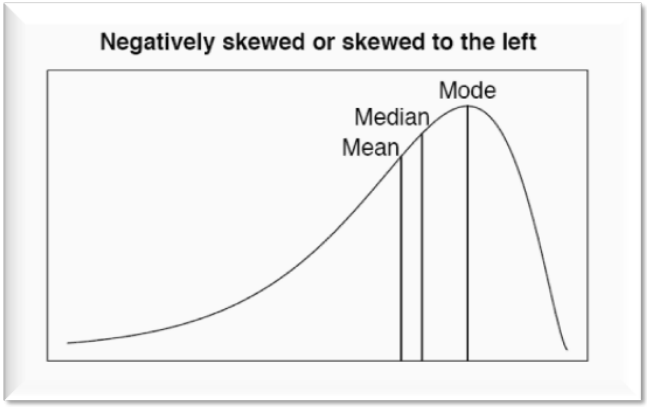
- Longer tail in the high value
- Mean > Median > Mode



Skewness II

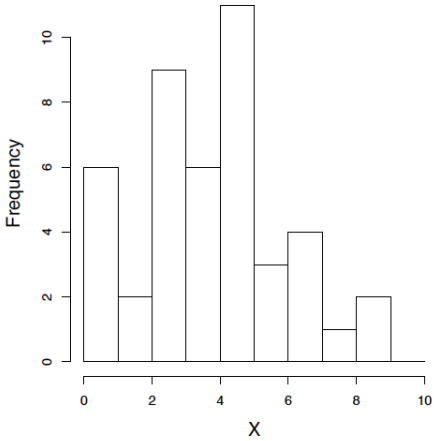
Negatively skewed

- Longer tail in the low value
- Mode > Median > Mean



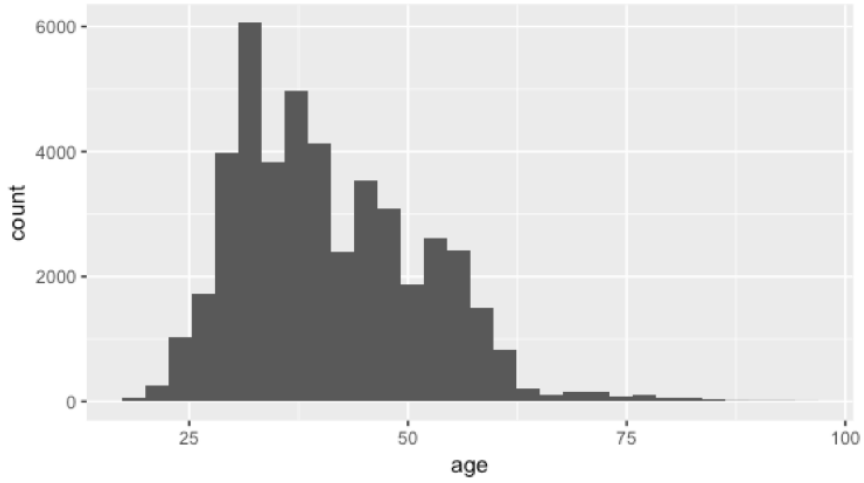
Univariate Graphical EDA
Histogram

- Histogram is a graphical representation of the distribution of numerical data
- It provides a view of data density and the shape of data distribution
- To construct a histogram, the first step is to
  - bin the range of values
  - count how many values fall into each interval
- The bins are usually specified as consecutive, non-overlapping intervals of a variable.
- The bins (intervals) must be adjacent, and are usually equal size.



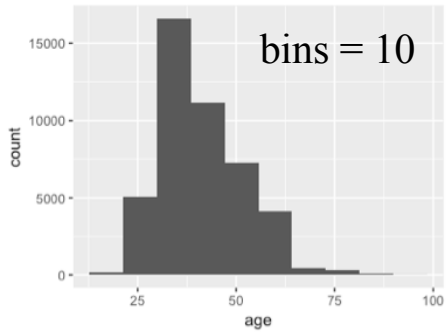
R: Histogram

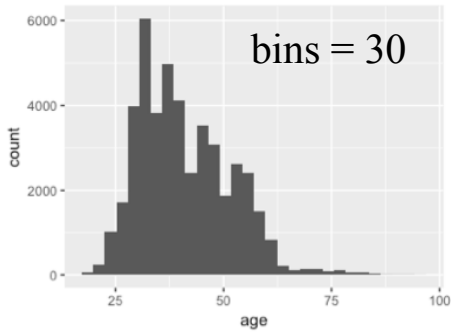
```
bankData %>%
    ggplot(aes(x = age)) +
    geom_histogram()
```

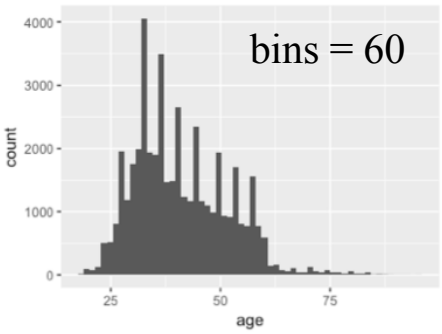


Univariate Graphical EDA
Effects of Histogram Bin

```
bankData %>%
   ggplot(aes(x = age)) +
   geom_histogram(bins = 10)
```

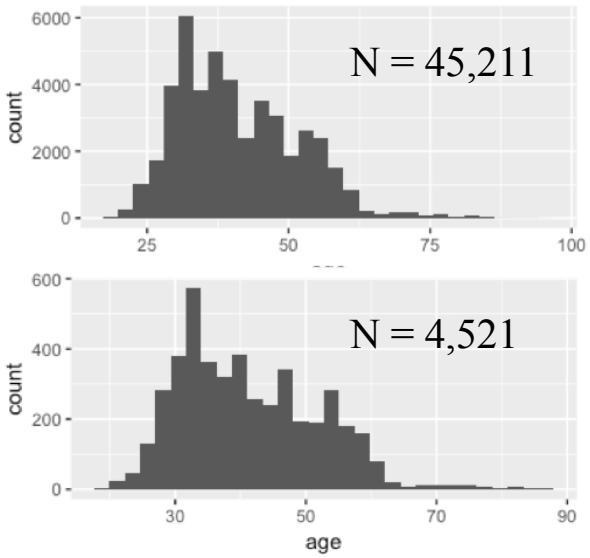




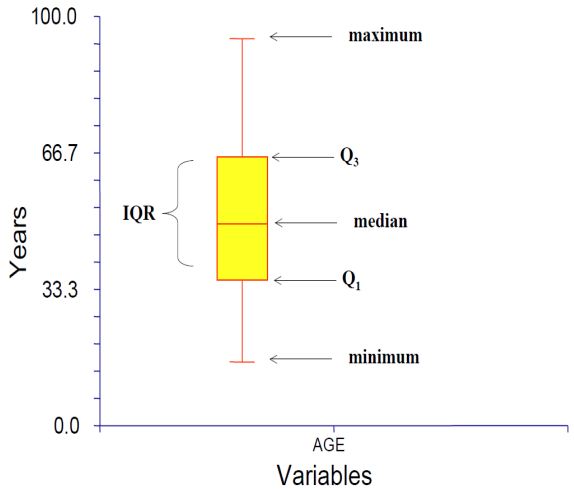


Univariate Graphical EDA
Effects of Number of Samples

```
bankData %>%
    sample_frac(0.1) %>%
    ggplot(aes(x = age)) +
    geom_histogram(bins = 30)
```



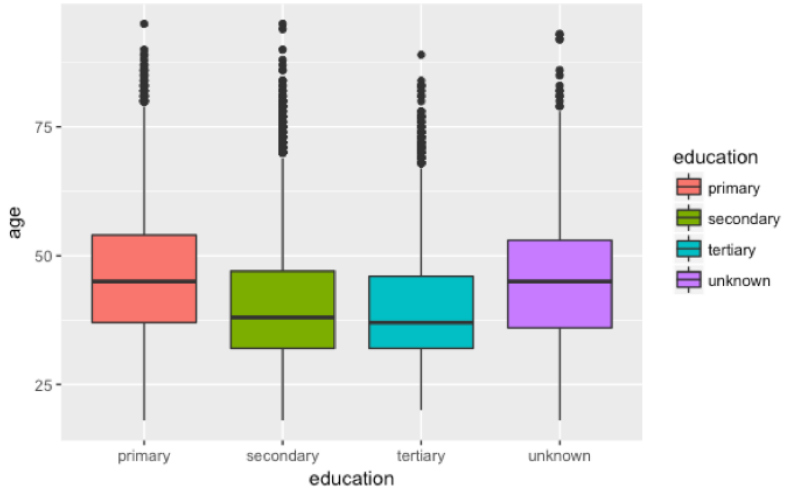
Boxplot



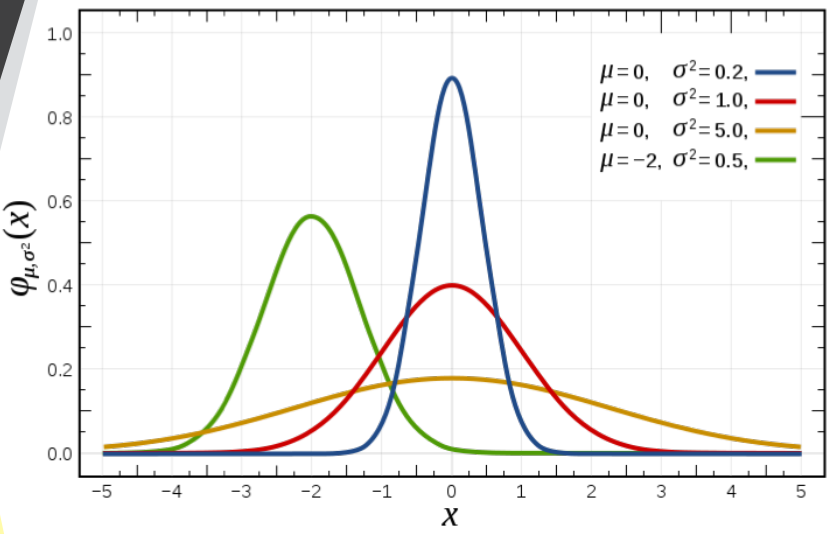
- The box in boxplot represents the middle 50% of the data
- The middle line indicates median
- Whiskers can be designated as either
  - Max/Min
  - Outlier boundaries
    - Upper = Q3 + 1.5*IQR
    - Lower = Q1 – 1.5*IQR

R: Boxplot

```
bankData %>%
    ggplot(aes(x = education,
               y = age,
               fill = education)) +
  geom_boxplot()
```



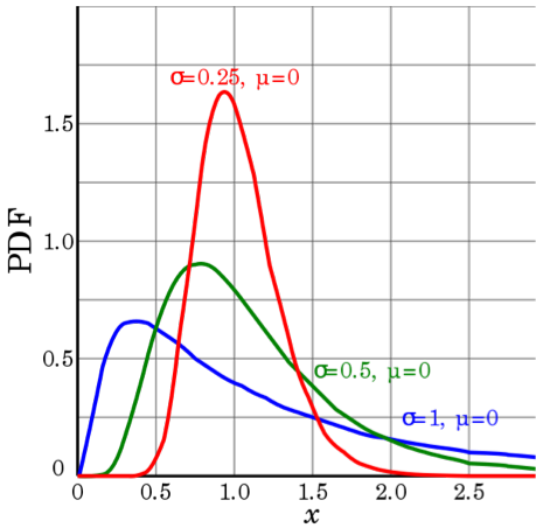
Common probability distribution
Normal distribution



The normal distribution is useful because of the central limit theorem.

It states that averages of samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal.

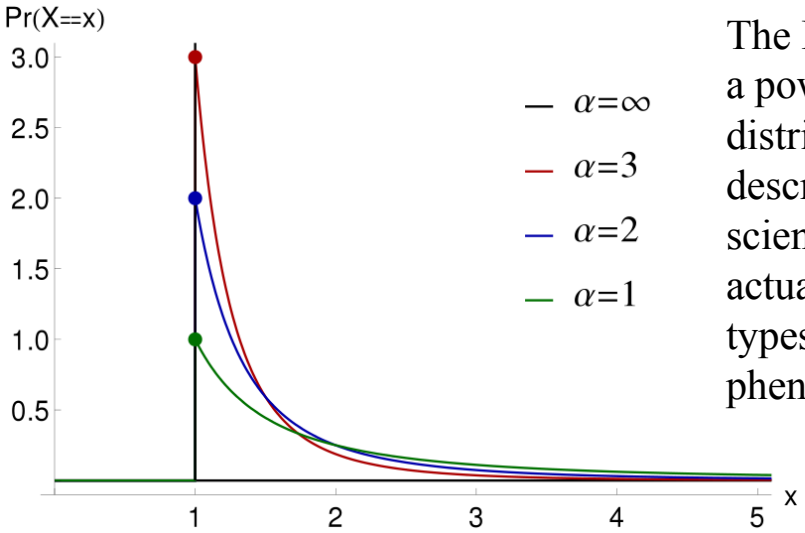
Common probability distribution
Log-normal distribution



Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution. Likewise, if Y has a normal distribution, then the exponential function of Y, X = exp(Y), has a log-normal distribution.

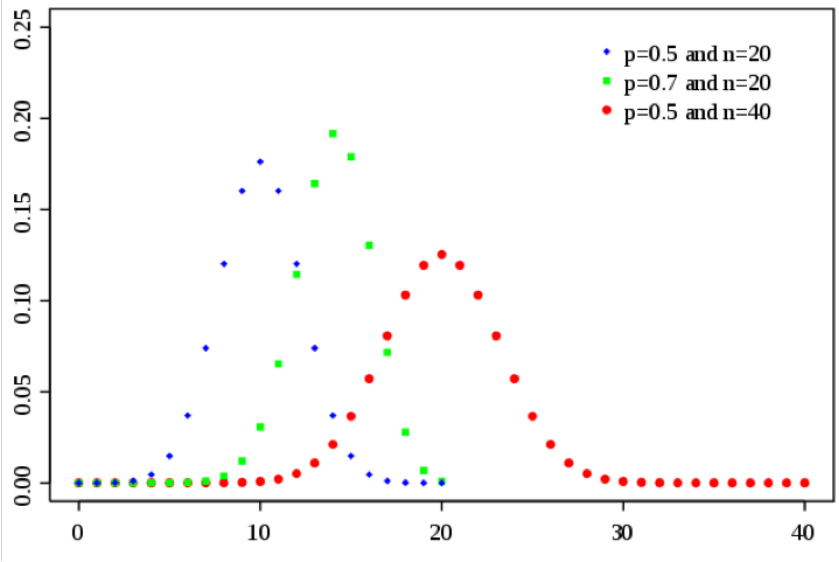
A log-normal process is the statistical realization of the multiplicative product of many independent random variables, each of which is positive.

Common probability distribution
Pareto distribution



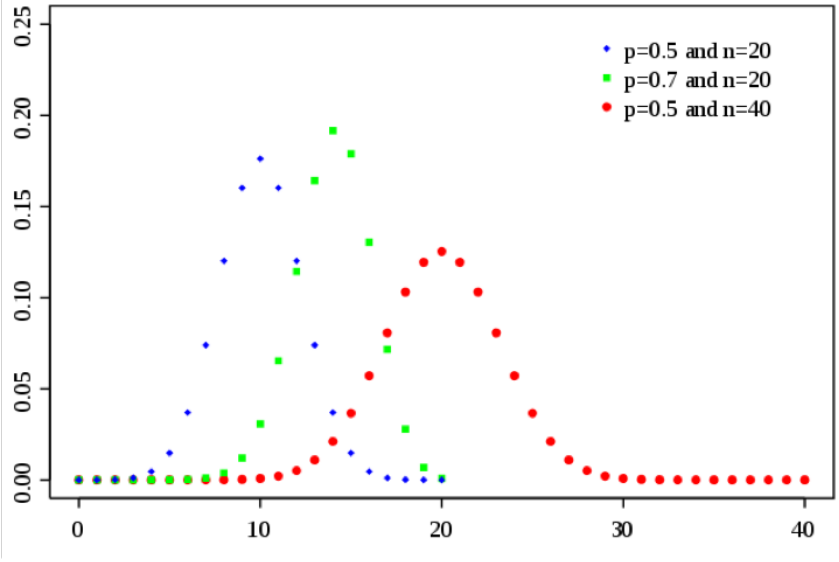
The Pareto distribution is a power law probability distribution that is used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena.

Common probability distribution
Binomial distribution



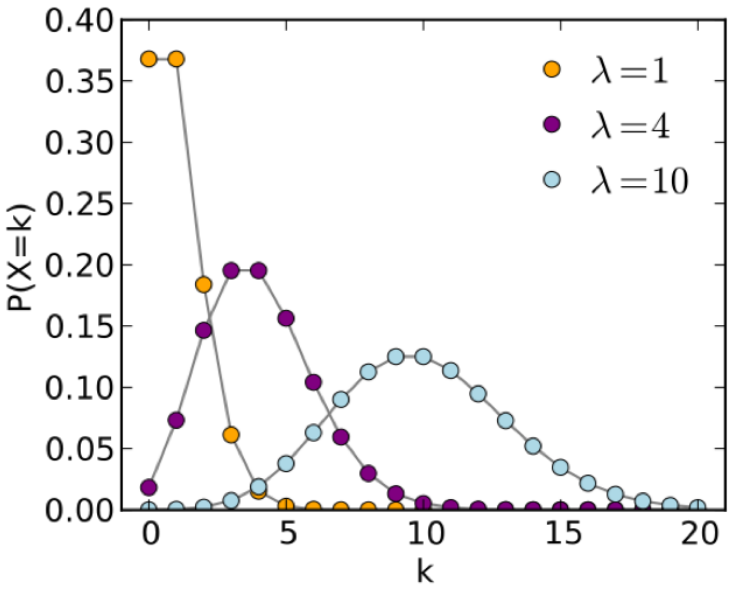
for the number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences

Common probability distribution
Binomial distribution



for the number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences

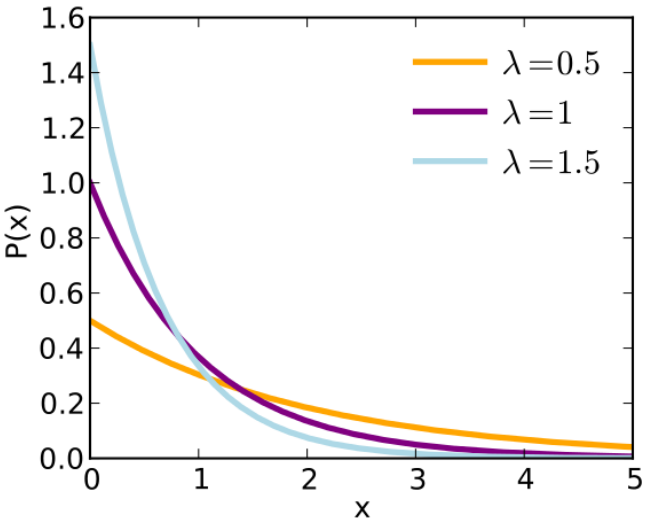
Common probability distribution
Poisson distribution



Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time.

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Common probability distribution
Exponential distribution



Exponential distribution is the probability distribution that describes the time between events in a Poisson point process, i.e. a process in which events occur continuously and independently at a constant average rate.

BIG DATA EXPERIENCE

g·able

Distribution and models

- Understand distributions allow us to understand the data and process better
- It also allows us to detect outliers
- Different distribution may indicates
  - Types of models to be used, e.g. Poisson regression, survival analysis
  - Data transformation, e.g. log transform for log-normal distribution

BIG DATA EXPERIENCE

g·able

Thank you

Question?