# Statistics for Data Science

## 1 Exploratory Data Analysis

### Histograms

- The histogram is one of the most basic tools for visualizing metric variables.
- To make a histogram,
  - X-Axis: First have to partition the range of a variable into intervals.
    - This is called "binning" the variable.
  - Y-axis: Then we draw a bar above each bin
    - The area of the bar is proportional to the number of observations that fall in that interval.
    - The Y axis could be a proportion or the number of observations.
      - If it represents the fraction of cases in each bin, the total area is 1.
      - If we plot the number of cases on the y-axis instead, it just scales the y-axis, but the plot looks the same.

| Class | 1 – <3 | 3 – <5 | 5 – <7 | 7 – <9 | 9 – <11 | 11 – <13 | 13 – <15 | 15 – <17 | 17 – <19 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 1 | 11 | 21 | 25 | 17 | 9 | 4 | 1 |
| Relative frequency | .011 | .011 | .122 | .233 | .278 | .189 | .100 | .044 | .011 |

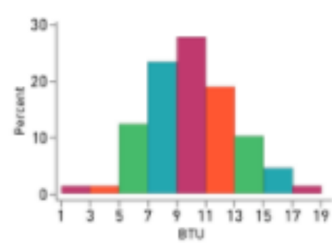
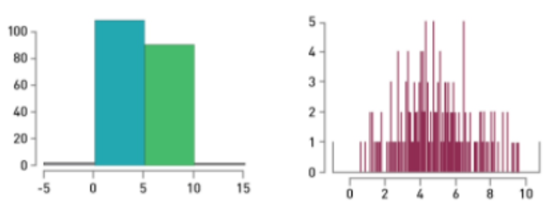
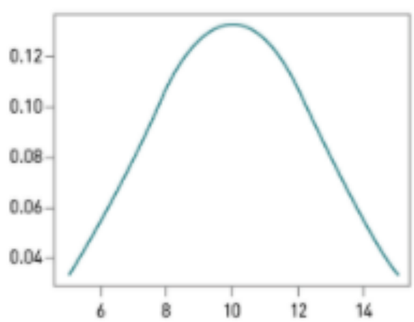



Figure 1.8 Histogram of the energy consumption data from Example 1.10

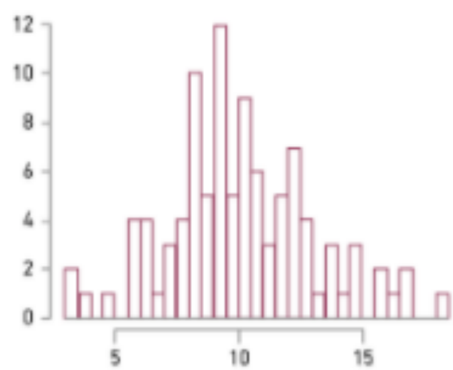
- The choice of bin width is very important.
  - If you choose one that's too wide, you don't get a lot of information.
  - If you choose one that's too narrow, the plot is very noisy.



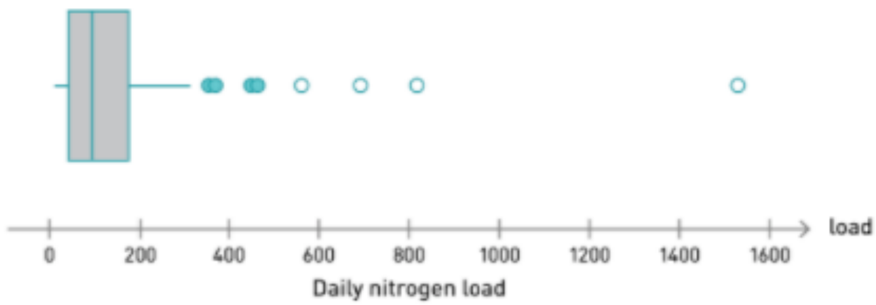
- A histogram shows us a specific sample.
- It doesn't directly show us the population the sample came from.
- When we do inference, we're going to assume that our sample comes from a population distribution.
- The histogram will give us an idea of what that distribution looks like.
- The below plot represents a basic model for a population.
  - This is a normal curve.




- The histogram from our data approximates the shape of the model.
  - It is not exact.
  - It gives clues about what the population model that created it could be.



# Boxplots




- A boxplot is another tool that's useful for visualizing numeric variables.
- The focus is on quantiles of the sample.
- First, there's a center line that represents the median.
  - On either side of this line, we draw a box that extends from the first quartile to the third quartile.
  - The box encloses the center 50% of the data.
  - The length of the box is called the interquartile range.
  - Length = Third quartile − first quartile
- On either end of the box, we draw whiskers.
  - Each whisker extends to the most extreme value, unless that value is more than **1.5 times the interquartile range** away from the box.
  - In that case, the whisker ends at 1.5 times the interquartile range, and the extreme outliers past that point are represented by dots.
- The boxplot helps us see quartiles, ranges, and outliers in a way that a histogram cannot.

# Graphs and plots

- Graphs capture much more detail than numerical summaries.
- Graphs are very useful for learning and communicating the features of a set of data.
- At the same time, graphical interpretation is not standard in the way that numerical summaries are.
- Be aware that our eyes can fool us.

- **General rules**
    - Show the data.
    - Induce the reader to think about the data being presented.
    - Avoid distorting the data.
    - Present many numbers with minimum ink.
    - Try to make large data sets coherent.
    - Encourage the reader to compare different pieces of data.
    - "Reveal" data at different levels of detail.
    - Serve a clear purpose with the visualization of the graph.
- **Things to Avoid**
    - Avoid distractions, chart junk, and distorting data.
    - Refine focus and purpose in order to create a clear representation of the data.
    - Pie Charts Are Usually Terrible
        - Eyes struggle to compare a given area within a pie chart.
        - Pie charts do not effectively convey information clearly.

## Guidelines for Statistical Reporting

- **Guideline One**
  - A statistical analysis is a written argument.
  - A good writing style is key.
  - This is technical writing.
  - Aim for clarity and exposition.
  - All the rules of good writing apply.
    - Organize your argument clearly.
    - Guide the reader through the evidence in the data.
    - Proofread.
- **Guideline Two**
  - If you don't have something nice to say (about your output), don't display it at all.
  - There should be no output dumps.
  - Every graph should be mentioned in your writing.
    - It should have some purpose.
  - Explain what the graphs and numbers mean.

- **Guideline Three**
  - You should document decisions.
  - If you decide that observations should be removed, state which ones.
  - If values are suspicious, but you leave them in, state that too.
  - If you transform a variable, for example, by taking the logarithm, state that.
  - Your justification can often be very brief (just a sentence), but make sure the reader can follow your logic.
- **Guideline Four**
  - Identify features that should be reflected in statistical models.
    - This will make more sense once you have experience building models.
  - Keep in mind the purpose of the analysis.
    - E.g., if you're interested in explaining the price of a house, look to see what kind of relationship that variable has with the explanatory variables.
    - Is it linear?
    - Is it exponential?
    - Are there values that don't seem to fit with the overall trend?
- **Guideline Five**
  - Remember the difference between sample and population.

- At this point, we don't know how to model a population.
  - This means that you must confine your conclusions to the sample.
  - You can talk about sample means, sample covariances.
- You can't say anything about the population that generated your sample.
- Be wary of technical words—in particular, the word significant.
  - People might casually say one value is significantly bigger than another.
  - But this has a technical meaning, and it implies that we've built a model and performed a statistical test.

- **Guideline Six**
  - In most professional situations, you have to think about different levels of detail for different audiences.
  - It's usually a good idea to provide an executive summary.
    - Not everyone can read 50 pages of output.
    - Often, you'll want to move details like your script to an appendix.

**NOTE on deleting data**

- Do not ever delete a data point just because it's an outlier.
- An outlier is a point that doesn't fit some statistical model we want to apply.
  - For novice statisticians (and many professionals who should know better), the obvious response is to delete outliers so the data and the model match.
- Unfortunately, this is distorting data to fit a preconceived idea of what it should look like.
  - Moreover, the outlying data points are usually the most interesting part of the story.
- The real test of whether to leave an outlier in your data is whether it's meaningful.
  - Does it represent a real feature of the process that generated the data?
  - If you suspect that the datapoint is erroneous in some way, then it's necessary to remove it.
  - Otherwise, it must be kept.
- Goal is always to choose a model that's flexible enough to describe the data we have.

# 2 Research design and Probability

## Correlational Research (observational study)

- Correlational study: We observe variables as they naturally occur.
- Variables aren't manipulated (we just measure them).
- Variation in the variables comes from nature, i.e., variation isn't introduced by the researcher.
- Correlational research also known as "observational research"
- We function as observers, separate from the process we are studying.
- **Why Correlational Research?**
  - When we can't manipulate a variable directly, correlational research may be our only option.
    - E.g., it's unethical to assign one group of people to be smokers and another to be nonsmokers in order to measure the effect of smoking.
    - Instead, we could ask people whether they are smokers or nonsmokers, and compare the health outcomes of the two groups.
- **Weakness of Correlational Research**
  - The interest is in whether smoking causes poor health or not.
  - The research reveals how variables are distributed in a population but doesn't tell us what causal pathways link variables.
  - We may find that smokers have worse health than nonsmokers do.
    - Is that because smoking degrades health?
    - Is it because people in poor health turn to smoking for relief?
    - Or is it because some other factor causes both smoking and poor health?
  - We know population averages, but the group of smokers is different in many ways from the group of nonsmokers.
  - Correlational research doesn't tell us about causal effects.
- **Important Exceptions**
  - Identification strategies:
    - instrumental variables,
    - regression discontinuity,
    - difference in difference
  - Assumptions needed for these to work are usually quite strict
  - Can only justify them in special circumstances

## Experimental Research

- It allows us to directly investigate causal claims:
  - Does variable X actually cause a change in variable Y?
- We divide our individuals or (units of analysis) into different treatments or conditions.
- A key feature of a true experiment is that the treatments are randomly assigned.
  - Each individual must have an equal chance of getting any particular treatment.
- Because of random assignment, we know that individuals in treatment condition are the same (on average) as those in control condition.
  - Groups should have similar characteristics.
  - Differences in behavior can be interpreted as being caused by the treatment.

- If you want to understand causal relationships, you can't just look at the statistics.
  - Stats that we'll learn can be applied to correlational or experimental data.
  - Data won't inform you where variation comes from.
- Statistics can always help us measure a relationship between two variables.
- To know if that relationship is a causal one, we have to look at
  - Design of the study
  - Process that leads different individuals to have different levels of a variable

# Types of datasets

- **Cross-Sectional Data Sets**
  - A cross section is a sample of individuals (or cities, countries, or other units of observation) at a given point in time.
    - No time dimension.
  - Ideally, each observation is independent.
    - E.g., we have a population, and we draw one person at a time with an equal chance of picking each person (random sampling).
  - Sometimes, we can't get perfect random sampling.
    - E.g., some people may not respond to surveys or units may be clustered.
  - Special techniques are used to respond to these situations.
  - Example:

Table 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

| obsno | wage | educ | exper | female | married |
|---|---|---|---|---|---|
| 1 | 3.10 | 11 | 2 | 1 | 0 |
| 2 | 3.24 | 12 | 22 | 1 | 1 |
| 3 | 3.00 | 11 | 2 | 0 | 0 |
| 4 | 6.00 | 8 | 44 | 0 | 1 |
| 5 | 5.30 | 12 | 7 | 0 | 1 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| 525 | 11.56 | 16 | 5 | 0 | 1 |
| 526 | 3.50 | 14 | 5 | 1 | 0 |

Indicator variables (1 = yes, 0 = no)

Observation number

Hourly wage

    - Each row represents measurements on one individual.
    - Each column records the value of a different variable.
    - Note that each individual is only measured once.
    - No extra structure that relates different rows

- **Time Series Data**
  - Time series: set of observations of a variable over time
    - Could also be several variables
  - Popular time series: stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, automobile sales, etc.
  - Different frequencies: daily, weekly, monthly, quarterly, annually, etc.
  - The ordering of observations conveys important information.
  - Observations cannot be assumed to be independent.
    - Temporal structure: past values of a variable contain info about future values
  - Time series looks at one individual (or one country, one stock market, etc.) at many times.

- **Pooled Cross Sections**

    - We may have two or more cross sections from the same population, taken at different times.
        - Can place these in one data set, called pooled cross sections
    - Each cross section is drawn independently, so the individuals in one are different than the individuals in the other.
        - Can't track changes in individuals from one period to another
    - Pooled cross sections are often used to evaluate policy changes.
    - Example:

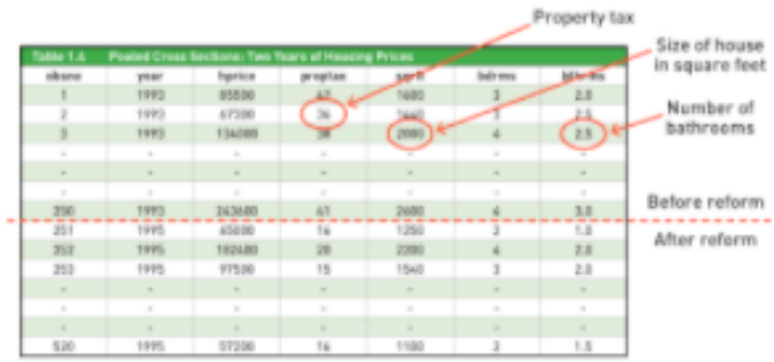


Table 1.4 Pooled Cross Sections: Two Years of Housing Prices

| obsno | year | hprice | proptax | sqrft | bdrms | bthrms |
|---|---|---|---|---|---|---|
| 1 | 1993 | 85500 | 42 | 1600 | 3 | 2.0 |
| 2 | 1993 | 67300 | 36 | 1440 | 3 | 2.5 |
| 3 | 1993 | 134000 | 38 | 2000 | 4 | 2.5 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 250 | 1993 | 243600 | 41 | 2600 | 4 | 3.0 |
| 251 | 1995 | 65000 | 16 | 1250 | 2 | 1.0 |
| 252 | 1995 | 182400 | 20 | 2200 | 4 | 2.0 |
| 253 | 1995 | 97500 | 15 | 1540 | 2 | 2.0 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 520 | 1995 | 57200 | 16 | 1100 | 2 | 1.5 |



        - Data can be used to evaluate the effect of change in property taxes on house prices.
        - Random sample of house prices for the year 1993
        - New random sample of house prices for the year 1995
        - Comparison of before/after (1993: before reform, 1995: after reform)

- **Panel or Longitudinal Data**
  - A panel contains data on multiple individuals, taken at multiple points of time.
  - Unlike pooled cross sections, the same units are measures in each time period.
    - Can follow the change to each individual over time
  - Panel data have a cross-sectional and a time series dimension.
  - A model for panel data has to account for both variation across individuals and variation across time.
    - Variety of advanced techniques to do this
  - Example:

Table 1.5 A Two-Year Panel Data Set on City Crime Statistics

| obsno | city | year | murders | population | unem | police |
|---|---|---|---|---|---|---|
| 1 | 1 | 1986 | 5 | 350000 | 8.7 | 440 |
| 2 | 1 | 1990 | 8 | 359200 | 7.2 | 471 |
| 3 | 2 | 1986 | 2 | 64300 | 5.4 | 75 |
| 4 | 2 | 1990 | 1 | 65100 | 5.5 | 75 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 297 | 149 | 1986 | 10 | 260700 | 9.6 | 286 |
| 298 | 149 | 1990 | 6 | 245000 | 9.8 | 334 |
| 299 | 150 | 1986 | 25 | 543000 | 4.3 | 520 |
| 300 | 150 | 1990 | 32 | 546200 | 5.2 | 493 |

Each city has two time series observations

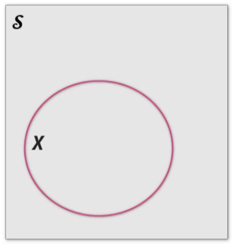
Number of police in 1986

Number of police in 1990

    - Each city may have different characteristics, some of which we may assume are constant through time.
    - Effect of police on crime rates may exhibit time lag.

# Probability Theory

- Probabilities follow same mathematical structure as areas.
- The probability of event X is the area of X over the area of ***S***.




- We roll a die with six sides.
- Rolling a 1 would be an event, rolling a 2 would be another, etc.
- These are **elementary events**, the list of all possible outcomes.


- **Composite events** are formed by combining several elementary events (e.g., rolling an even number)


- **Mutually exclusive** events are events that don't overlap at all.
  - They have no area in common.
  - No more than one can occur at a time.
- **Exhaustive events** are events that cover the entire event space, ***S***.
  - Every point in ***S*** falls in at least one of the events.
  - At least one of the events occurs.
- Events could be both mutually exclusive and exhaustive.
  - Exactly one of them occurs but never more than one.

- **Addition Rule (Mutually exclusive)**
  - Events X and Y are mutually exclusive.
  - We want to know the probability that X or Y occurs, or $X \cup Y$.
  - The addition rule for mutually exhaustive events says: $P(X \cup Y) = P(X) + P(Y)$

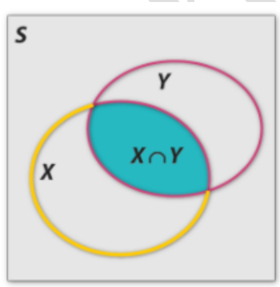
- **Addition Rule (general)**
  - Events X and Y are not mutually exclusive.
  - $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$

- **Axioms of probability**
  - A set $\boldsymbol{S}$, called the sample space.
    - Each element of $\boldsymbol{S}$ is called an outcome.
  - A set of events, $\mathscr{F}$.
    - Each event is a subset of $\boldsymbol{S}$.
  - Define $\mathscr{F}$ to be the set of all subsets of $\boldsymbol{S}$ (power set of $\boldsymbol{S}$).
  - A function, P, from the set of events to the real numbers
  - We assume that this function follows a set of properties (axioms of probability).
    - **Axiom 1**: $P(A) \geq \emptyset$ for any event A in $\mathscr{F}$
    - **Axiom 2**: $P(\boldsymbol{S}) = 1$
    - **Axiom 3**: For any countably infinite set of disjoint events {A1, A2, A3,…}

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

  - Putting these elements together, a Probability Space is the triple ($\boldsymbol{S}$, $\mathscr{F}$, P).

- **Conditional Probability**
  - Probability that event X occurs given that event Y occurs, $P(X|Y)$
  - Ratio of areas: the area of the shaded intersection over the area of Y




  - Same as probability of $X \cap Y$ over the probability of Y
  - Written as $P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$

- **Multiplication Rule**
  - From the above, we get

    $P(X \cap Y) = P(X|Y). P(Y)$

- **Bayes Rule**

$$P(X|Y) = \frac{P(Y|X).P(X)}{P(Y)}$$

  - Where,
    - P(X|Y) → Posterior
    - P(X) → Prior
    - P(Y|X) → Likelihood
    - P(Y) → Normalizer (using total probability)

- **Law of total probability**

$$P(Y) = P(Y|X).P(X) + P(Y|\overline{X}).P(\overline{X})$$

- **Independent Events**
  - Two events are independent if P(X|Y) = P(X)
  - This gives us: $P(X \cap Y) = P(X).P(Y)$
  - If 2 events are mutually exclusive, they cannot be independent

# 3 Discrete Random Variables

## Binomial Probability Distribution

- **Binomial Experiment**
    - An experiment that satisfies the below conditions:
        - Experiment consists of a sequence of n trials where n is predetermined.
        - **Dichotomous**: Each trial can result in one of 2 possible outcomes
        - **Independent**: Each trial is independent and doesn't affect other trials.
        - **Homogenous**: The probability of success is constant from trial to trial.
    - **Independence Assumption**
        - If sampling (i.e trials) is done ***without*** replacement, the experiment will not yield independent outcomes.
        - However, if the number of trials (n) is at most 5% of the population size(N), we can assume independence even without replacement.
        - i.e. without replacement if $n/N < 0.05$ ➔ trials assumed independent

- **Binomial Random Variable**
    - A binomial random variable X is defined as,
        - X = Number of successes among n trials
        - Denoted as: $X \sim Bin(n, p)$
            - Where,
                - n: number of trials
                - p: probability of success
    - The **pmf of X** is denoted as: b(x;n,p)

$$P(X = x) = b(x; n, p) = \begin{cases} \binom{n}{x} . p^x (1-p)^{n-x} & , x = 0,1,2 \ldots n \\ 0 & , otherwise \end{cases}$$

- The **CDF of X** is denoted as: B(x;n,p)

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p) \ , x = 0,1,2 \ldots n$$

- If $X \sim Bin(n, p)$, let q = (1-p)
    - E(X) = np
    - var(X) = npq
    - $Std\ deviation, \sigma_x = \sqrt{npq}$

## Hypergeometric Distribution

- The binomial distribution is the approximate probability model for sampling without replacement from a finite dichotomous population (of size N), provided the sample size n << N. (less than 5%)
- The hypergeometric distribution is the exact probability model for the number of success's in a sample.
- The assumptions leading to a hypergeometric distribution are as follows:
    - Population consists of **N** individuals
    - Each individual can be characterized either a success or failure.
        - There are **M** successes in the population.
    - A sample of **n** individuals is selected without replacement such that each subset of size n is equally likely.
- A hypergeometric random variable X is defined as,
    - X = Number of successes in a sample

- The **pmf of X** is denoted as: h (x; n, M, N)

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, where\ max(0, n - N + M) \leq x \leq min(n, M)$$

- The **CDF of X** is given by:

$$P(X \leq x) = \sum_{y=0}^{x} h(y; n, M, N)\ , x = 0,1,2 \ldots n$$

- For a hypergeometric random variable, $X$, let p = $\frac{M}{N}$ i.e proportion of successes in the population. We get:
    - E(X) = $np$
    - var(X) = $\frac{N-n}{N-1}. n. p. (1-p)$
    - $Std\ deviation, \sigma_x = \sqrt{var(X)}$

Negative Binomial Distribution

Poisson Probability Distribution

# END OF DOCUMENT