

What is Statistics?

The subject of Statistics deals with the process of finding out more about a scientific question in a discipline by collecting information (data) and then try to make some sense out of that information.

Broadly speaking ‘**Statistics**’ is the science of

- 1) Planning studies and experiments to collect data.
- 2) Obtaining and summarizing data.
- 3) Analyzing, interpretation and drawing conclusions based on data.

Data are usually collected on “Variables”. A variable is a characteristic that ‘varies’ across subjects, objects or items.

Types of Variables:

- (A) **Qualitative or Categorical:** The variable can take a small set of prescribed values, or a set of categories.

Examples:

- (1) Political ideology: liberal, moderate, conservative
- (2) Accommodation: house, condo, apartment
- (3) Opinion: disagree, neutral, agree
- (4) Preference: Brand A, Brand B, Brand C
- (5) Binary: yes, no; accept, reject; correct, incorrect.

Further subclassifications: (a) nominal (b) ordinal.

Categorical variables having unordered scales are called “nominal” variables.

Examples:

- (a) Religious affiliation
Catholic, Jewish, Hindu, Muslim
- (b) Primary mode of transportation to work
Walk, Bicycle, Bus, Subway, Car

-the order of listing the categories is irrelevant.

Categorical variables having ordered scales are called “ordinal” variables.

Examples:

(a) Response to medical treatment

Categories: Excellent, Good, Fair, Poor

(b) Stockpile of PPE at a hospital

Categories: Too low, About right, Too much

Statistical Methods designed for ordinal variables may not be used with nominal variables since nominal variables do not have ordered categories.

However, methods designed for nominal variables **can** be used with ordinal variables but not recommended.

(B) Quantitative Data: Data that consists of measurements on a scale of real numbers.

Examples: Weights, Ages, Body temperatures, Incomes of a group of people.

Two subclassifications: (1) Discrete (2) Continuous

When the variable is a count, the data are known as discrete.

Example: Number of children in a hospital

When the variable is a measurement, the data are known as continuous data.

Example: Body temperatures, Weights, Heights etc.

Some more examples:

The region of the US in which an individual lives: East, West, North, South

The time it takes for a student to complete an exam.




The maximum temperatures in U.S. cities

Response (Y) and Explanatory (X) variables:

We use the term response, outcome or dependent variable for measurements that are free to vary in response to other variables called explanatory variables or predictor variables or independent variables.

Response variables (y) are regarded as random variables. Explanatory variables (x) are usually treated as though they are non-random or known variables, for example, they may be fixed by the experimental design or survey.

Summary of Statistical Methods:

Explanatory Variable (x) 	Response (y)	
	Categorical	Quantitative
Categorical	Contingency Tables GLM's	T-test ANOVA
Quantitative	GLM's	Linear Regression
Probability models or distributions	 Binomial, Poisson Multinomial	 Normal Multivariate normal

Visualizing Data

The very simplest way to present or visualize a dataset is to produce a table.

Tables can be helpful but aren't much use for large datasets. We normally make bar charts and pie charts.

Example: Psychologists collected data from students in grades 4 – 6 in three school districts to understand what factors students thought made other students popular.

This data was presented by Chase and Dunner in a paper published in Research Quarterly for Exercise and Sport in 1992.

Among other things, they asked each student whether their goal was to make good grades (Grades), to be popular (Popular), or to be good at sports (Sports).

The sample size was 478. A table would be hard to read and it's even harder to draw any serious conclusions from the data. We need a more effective tool than eyeballing the table.

Table 1.2 Chase and Dunner, in a study described in the text, collected data on what students thought made other students popular

Gender	Goal	Gender	Goal
Boy	Sports	Girl	Sports
Boy	Popular	Girl	Grades
Girl	Popular	Boy	Popular
Girl	Popular	Boy	Popular
Girl	Popular	Boy	Popular
Girl	Popular	Girl	Grades
Girl	Popular	Girl	Sports
Girl	Grades	Girl	Popular
Girl	Sports	Girl	Grades
Girl	Sports	Girl	Sports

As part of this effort, they collected information on (a) the gender and (b) the goal of students. This table gives the gender (“boy” or “girl”) and the goal (to make good grades—“Grades”; to be popular—“Popular”; or to be good at sports—“Sports”). The table gives this information for the first 20 of 478 students; the rest can be found at <http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html>. This data is clearly categorical, and not ordinal

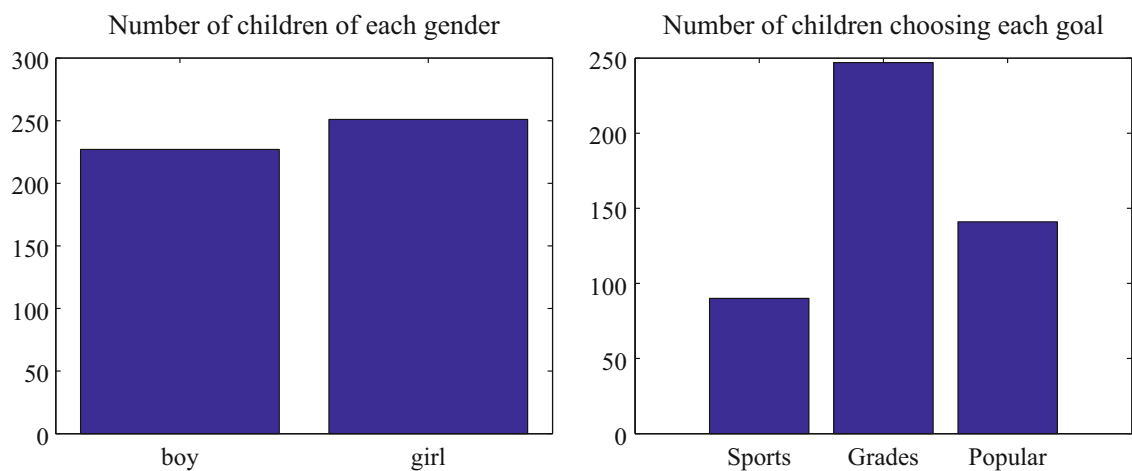


Fig. 1.1 On the *left*, a bar chart of the number of children of each gender in the Chase and Dunner study. Notice that there are about the same number of boys and girls (the bars are about the same height). On the *right*, a bar chart of the number of children selecting each of three goals. You can tell, at a glance, that different goals are more or less popular by looking at the height of the bars

be very hard to read. Table 1.2 shows the gender and the goal for the first 20 students in this group. It’s rather harder to draw any serious conclusion from this data, because the full table would be so big. We need a more effective tool than eyeballing the table.

1.2.1 Bar Charts

A **bar chart** is a set of bars, one per category, where the height of each bar is proportional to the number of items in that category. A glance at a bar chart often exposes important structure in data, for example, which categories are common, and which are rare. Bar charts are particularly useful for categorical data. Figure 1.1 shows such bar charts for the genders and the goals in the student dataset of Chase and Dunner. You can see at a glance that there are about as many boys as girls, and that there are more students who think grades are important than students who think sports or popularity is important. You couldn’t draw either conclusion from Table 1.2, because I showed only the first 20 items; but a 478 item table is very difficult to read.

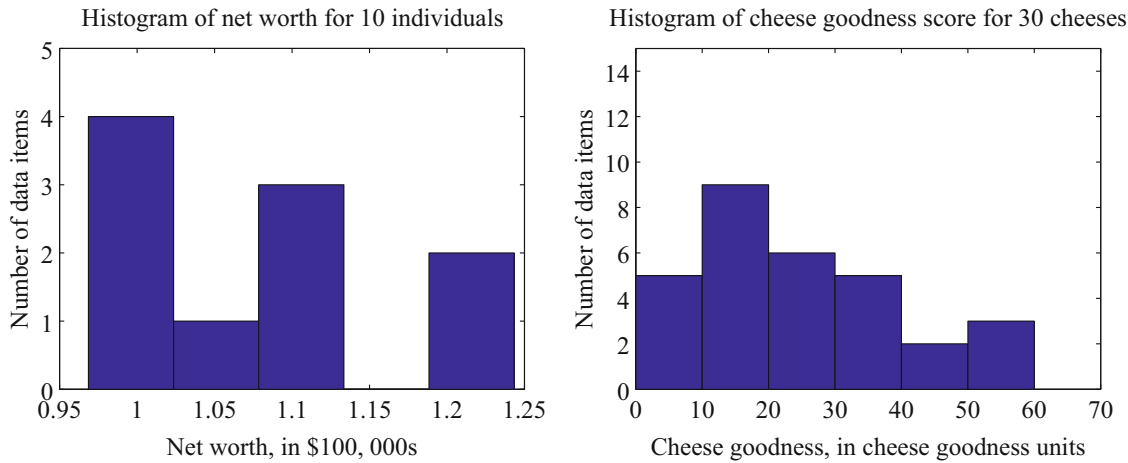


Fig. 1.2 On the *left*, a histogram of net worths from the dataset described in the text and shown in Table 1.1. On the *right*, a histogram of cheese goodness scores from the dataset described in the text and shown in Table 1.1

1.2.2 Histograms

Data is continuous when a data item could take any value in some range or set of ranges. In turn, this means that we can reasonably expect a continuous dataset contains few or no pairs of items that have *exactly* the same value. Drawing a bar chart in the obvious way—one bar per value—produces a mess of unit height bars, and seldom leads to a good plot. Instead, we would like to have fewer bars, each representing more data items. We need a procedure to decide which data items count in which bar.

A simple generalization of a bar chart is a **histogram**. We divide the range of the data into intervals, which do not need to be equal in length. We think of each interval as having an associated pigeonhole, and choose one pigeonhole for each data item. We then build a set of boxes, one per interval. Each box sits on its interval on the horizontal axis, and its height is determined by the number of data items in the corresponding pigeonhole. In the simplest histogram, the intervals that form the bases of the boxes are equally sized. In this case, the height of the box is given by the number of data items in the box.

Figure 1.2 shows a histogram of the data in Table 1.1. There are five bars—by my choice; I could have plotted ten bars—and the height of each bar gives the number of data items that fall into its interval. For example, there is one net worth in the range between \$102,500 and \$107,500. Notice that one bar is invisible, because there is no data in that range. This picture suggests conclusions consistent with the ones we had from eyeballing the table—the net worths tend to be quite similar, and around \$100,000.

Figure 1.2 also shows a histogram of the data in Table 1.1. There are six bars (0–10, 10–20, and so on), and the height of each bar gives the number of data items that fall into its interval—so that, for example, there are 9 cheeses in this dataset whose score is greater than or equal to 10 and less than 20. You can also use the bars to estimate other properties. So, for example, there are 14 cheeses whose score is less than 20, and 3 cheeses with a score of 50 or greater. This picture is much more helpful than the table; you can see at a glance that quite a lot of cheeses have relatively low scores, and few have high scores.

1.2.3 How to Make Histograms

Usually, one makes a histogram by finding the appropriate command or routine in your programming environment. I use Matlab and R, depending on what I feel like. It is useful to understand the procedures used to make and plot histograms.

Histograms with Even Intervals: The easiest histogram to build uses equally sized intervals. Write x_i for the i 'th number in the dataset, x_{\min} for the smallest value, and x_{\max} for the largest value. We divide the range between the smallest and largest values into n intervals of even width $(x_{\max} - x_{\min})/n$. In this case, the height of each box is given by the number of items in that interval. We could represent the histogram with an n -dimensional vector of counts. Each entry represents the count of the number of data items that lie in that interval. Notice we need to be careful to ensure that each point in the range

Steps for making a histogram

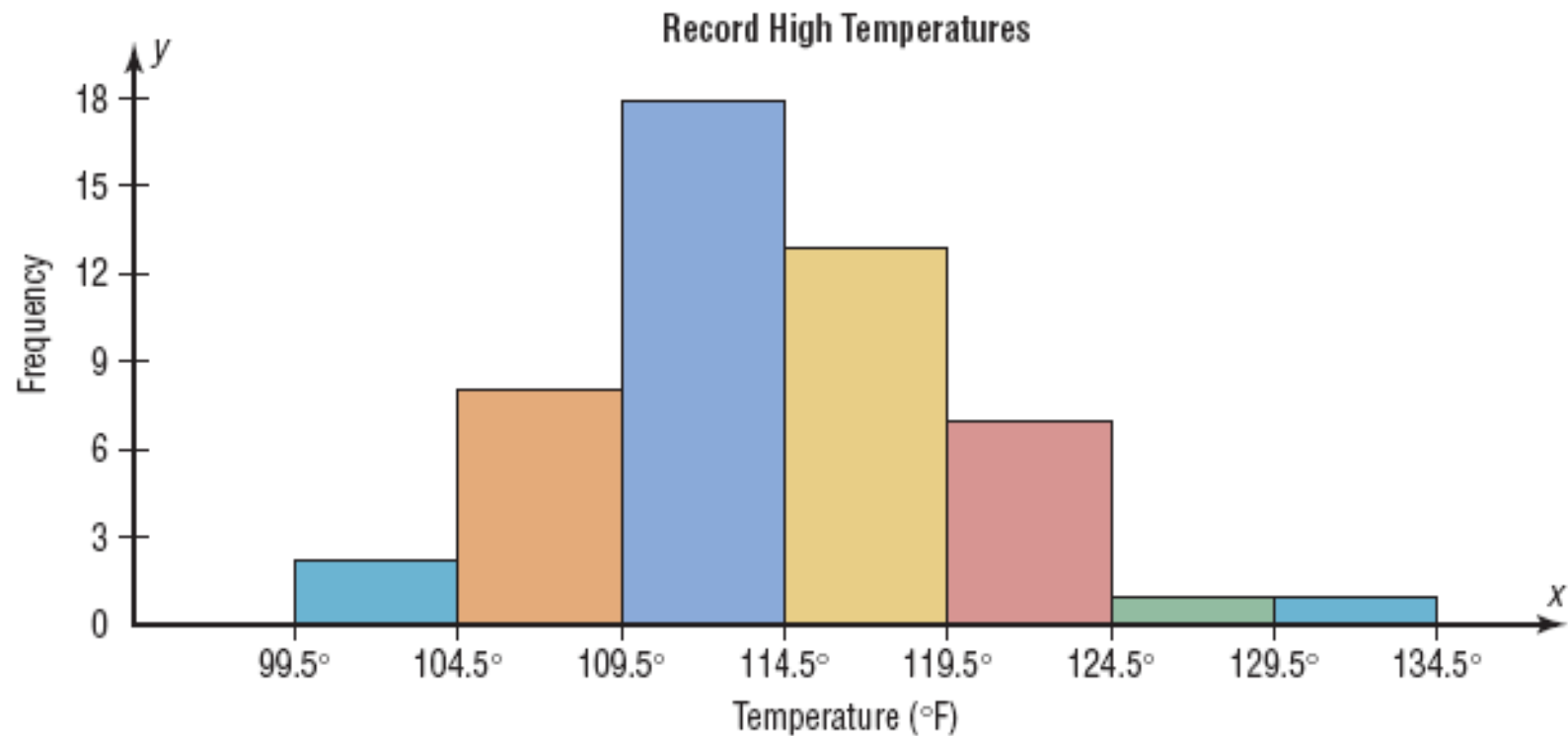
1. Choose the number of classes (usually between 5 – 10).
2. Find the highest and lowest values in the sample.
3. Calculate

$$\text{Class width} = \frac{\text{highest value} - \text{lowest value}}{\text{\# of classes}}$$

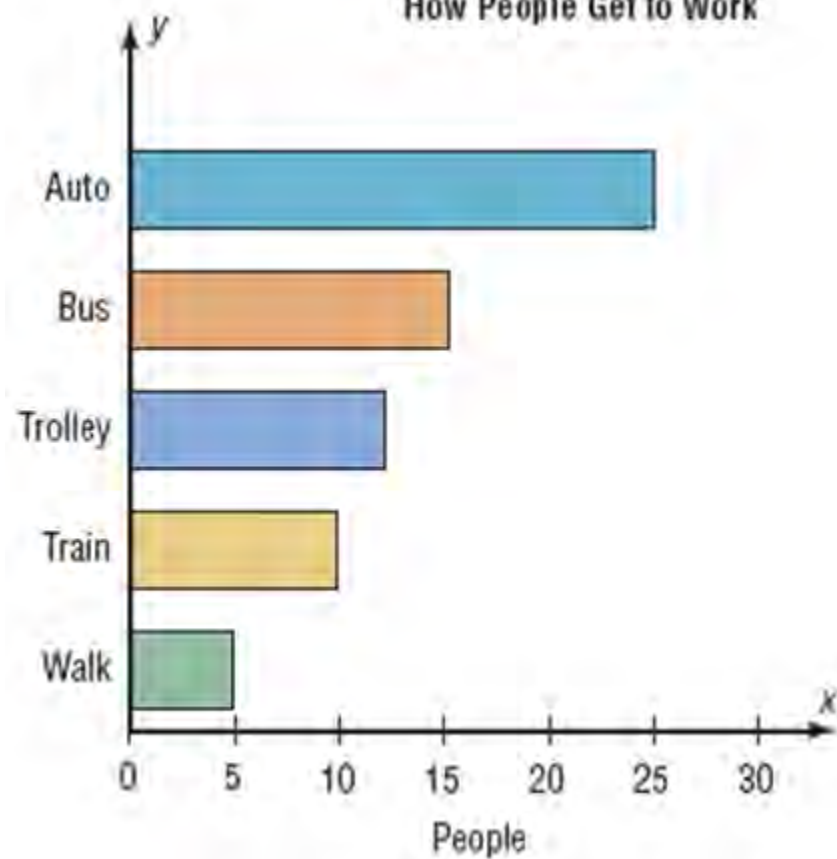
4. Determine class boundaries
5. Count the number of data values that fall into each of the classes and write the frequency distribution as a table.

You can create a histogram in excel, python etc.

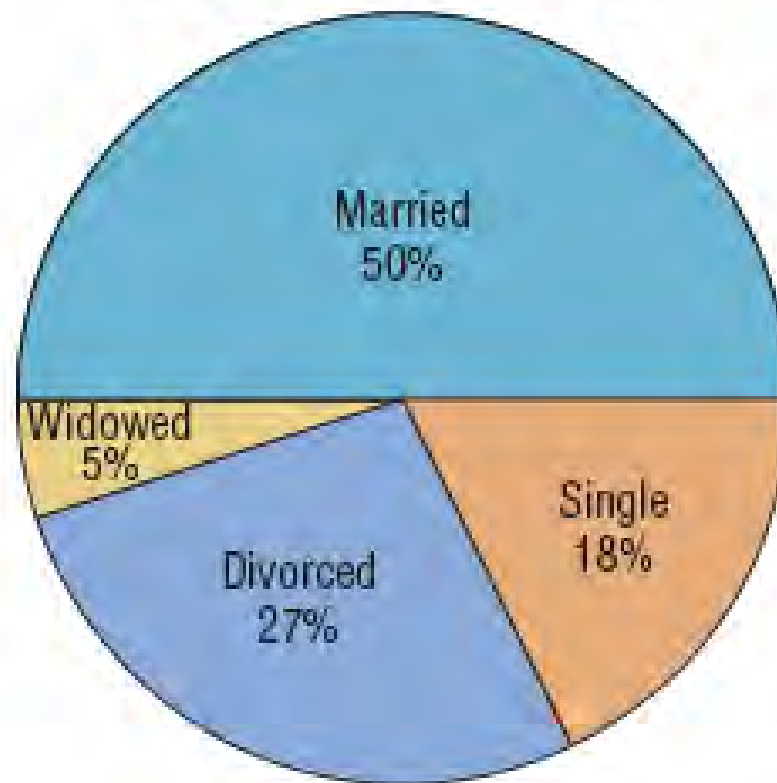
Histograms



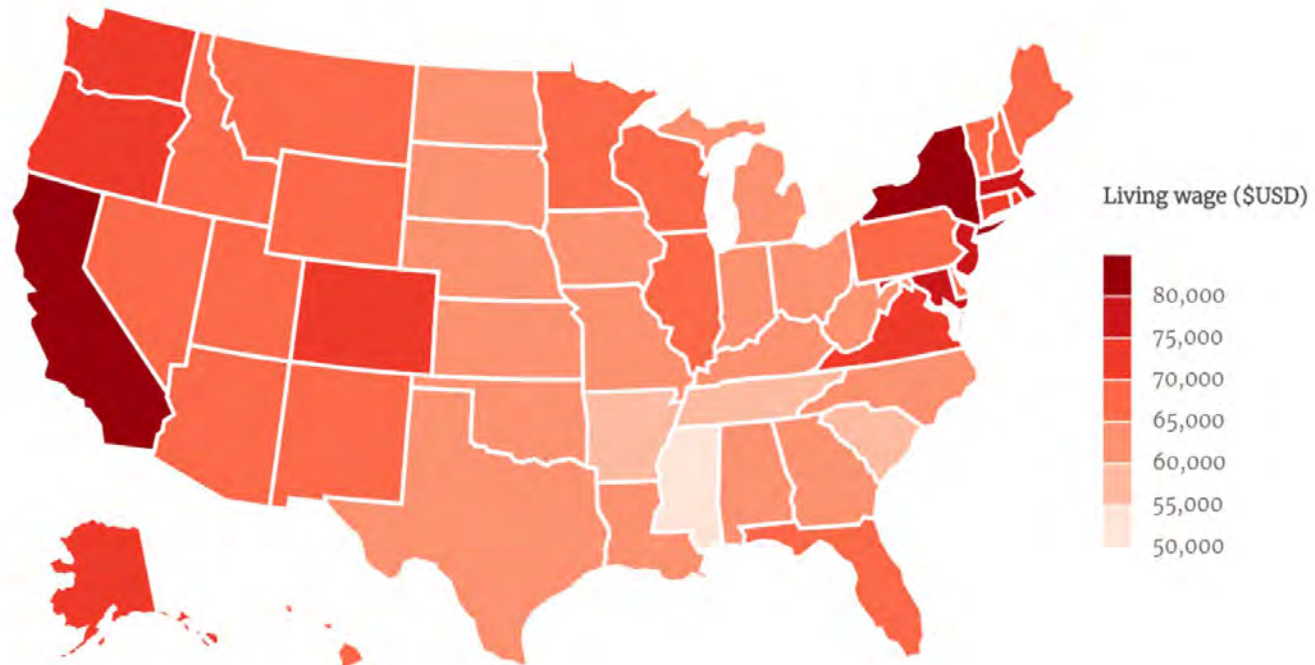
How People Get to Work



**Marital Status of Employees
at Brown's Department Store**



Minimum income a family of four needs to get by in each state



3D Map Shows U.S. Economic Contribution by City

Two Centuries of US Immigration