AP Stats Notes

Franklin Chen

11 November 2024 - idk some time in may

using *The Practice of Statistics for the AP Exam: 6th Edition* by Starnes and Tabor

1 Data Analysis

1.1 What is Statistics?

Definition 1 (Statistics)**.** The science of collecting, analyzing, and drawing conclusions from data.

Data is collected from *individuals* about certain *variables*.

Definition 2 (Individual, Variable)**. Individuals** are objects described in a dataset. Typically people, but not always.

Variables are attributes that can take different values for different individuals.

For example, *individuals* may be households, and *variables* may be region, nnumber of people, household income, etc. It's important to distinguish between *categorial* and *quantative* variables:

Definition 3 (Categorical and Quantative Variables)**. Categorical Variables** are variables whose values can be placed into distinct categories.

Quantative Variables are variables whose values are quantities, typically counts or measurements.

For example, region would be categorical, while household income would be quantative. *Not all numbers are quantative*; eg. zip code.

1.2 Analyzing Categorical Data

1.2.1 One-Variable Categorical Data

Definition 4 (Frequency and Relative Frequency Tables)**. Frequency Tables**shows the number of individuals that have values of a certain category. **Relative Frequency Tables** shows the proportion or percent of individuals in each category.

Note (relative) frequencies are not data; they summarize *data.* Bar graphs and Pie Graphs summarize relative frequency tables.

Beware of misleading graphs; we mainly react to the area of each bar, not the actual height.

	Like Skateboards	Do Not Like Skateboards	Totals
Like Snowmobiles	80	25	105
Do not like Snowmobiles	45	10	55
Totals	125	35	160

MathBits.com

Figure 1: An example two-way table with additional summary information.

1.2.2 Two-Variable Categorical Data

Use a two-way table to summarize data about two categorical variables. These tables can be used to answer questions about *marginal, joint, and conditional relative frequencies.*

Margial relative frequencies give the percent or proportion of individuals that have a given value for one categorical variable. For example, the marginal relative frequency of liking skateboards is $\frac{125}{160} \approx 78.125\%$.

Joint relative frequencies give the percent or proportion of individuals that have a specific value for both categorical variables. For example, the joint frequency of liking both skateboards and snowmobiles is $\frac{80}{160} = 50\%$.

Conditional relative frequencies give the percent or proportion of individuals that have a specific value for one categorical variable relative to other individuals with the same other categorical variable. For example, the conditional relative frequency of those who like snowmobiles out of all individuals that like skateboards is $\frac{80}{125} = 64\%$.

These frequencies can be summarized in *side by side bar graphs, segemented bar graphs, or mosaic plots.*

Graphs and these tables can be used to show **assocation** between two variables. There is association between two variables if knowing the value of one helps to predict the other. For example, knowing that an individual likes skateboards helps predict whether they like snowmobiles ($\frac{80}{125} = 64\%$ vs $\frac{25}{35} \approx 71.4\%$). **ASSOCIATION DOES NOT IMPLY CAUSATION!**

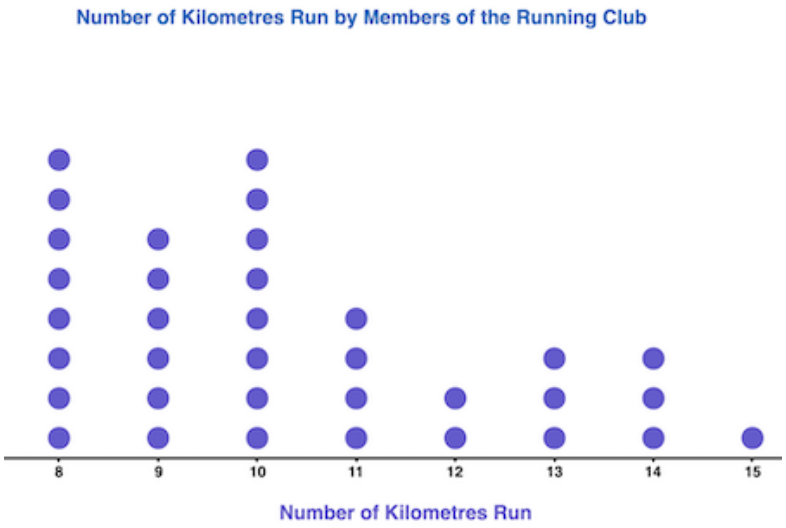

Figure 2: A dotplot showing the distribution of kilometers run by members of the running club.

1.3 Analyzing Quantative Data with Graphs

Dotplots (as shown above) show each individual as a dot above their quantative data value.

When describing the shape of a dotplot (or other quantative graphs), *focus on main features*: major peaks, clusters, or gaps. Especially note whether the distribution is roughly symmetric or skewed:

Definition 5 (Symmetric, Skewed)**.** A distribution is rougly **symmetric** if the right side of the graph has rougly the same shape as the left side.

A distribution is **skewed to the right** if the right 'tail' has less values than the left; typically, the left has a peak whereas the right does not. **Left-skewed** definition are defined similarly to right-skewed distributions.

For example, the distribution of the number of kilometers run is right-skewed because the right 'tail' has less values.

Graphs with a single peak are considered *unimodal*, like the dotplot. Distributions with two peaks are considered *bimodal*, and beyond that is considered *multimodal.*

When describing a distribution of quantative data, use the acronym ROCS: **R**ange (max - min), **O**utliers (clear departures from the data), **C**enter (mean or median), and **S**hape (symmetry, skew, gaps, peaks).

Leaf plots exist. Stem represents first few digits, leaf represents final digit.

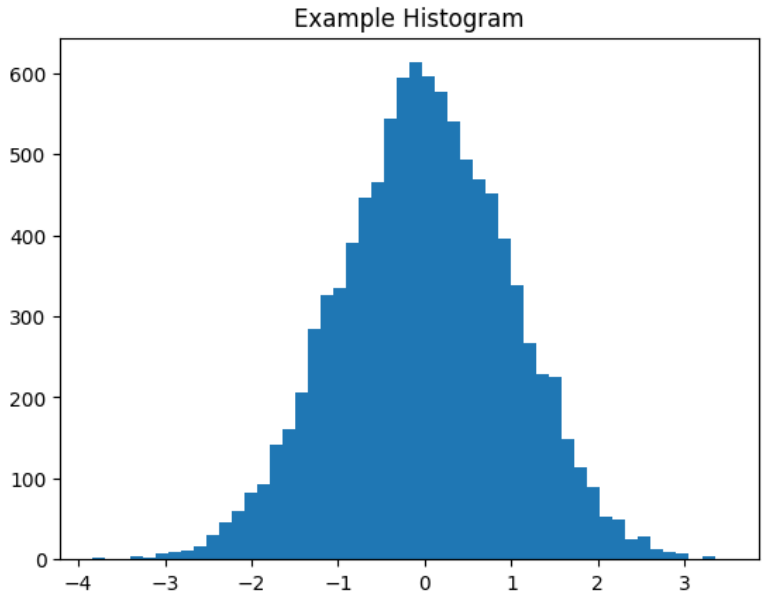

Figure 3: An example histogram with a normal distribution.

Histograms are a notable way of displaying quantative data, as they avoid showing individual data points. Histograms divide the variable into many 'bins' (bars), with the height representing the frequency. Smaller bins show more detail at the cost of a less clear pattern.

Don't confuse histograms and bar graphs. Histograms are used for quantative data, while bar graphs are used for qualatative data.

Use percentages when comapring to distributions in order to remove the effect of a larger sample.

1.4 Describing Quantative Data with Numbers