AP Stats Notes

Franklin Chen

11 November 2024 - idk some time in may

Contents

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!**

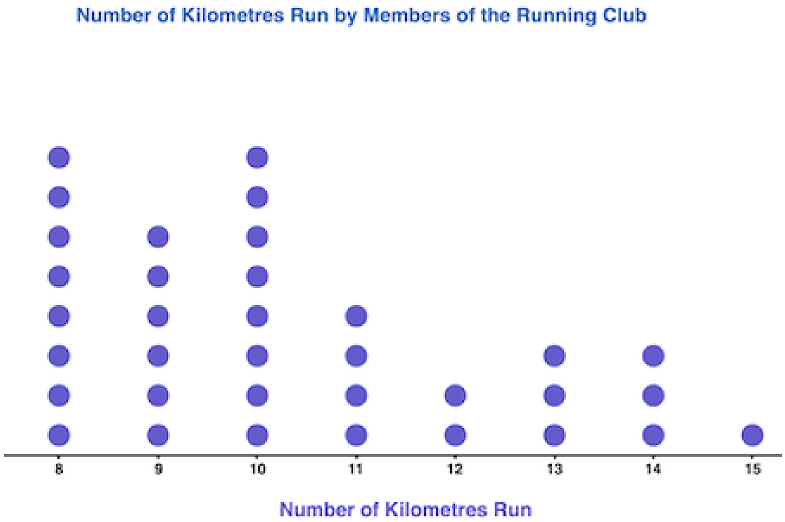

Figure 2: A dotplot showing the distribution of kilometers run by members of the running club.

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.

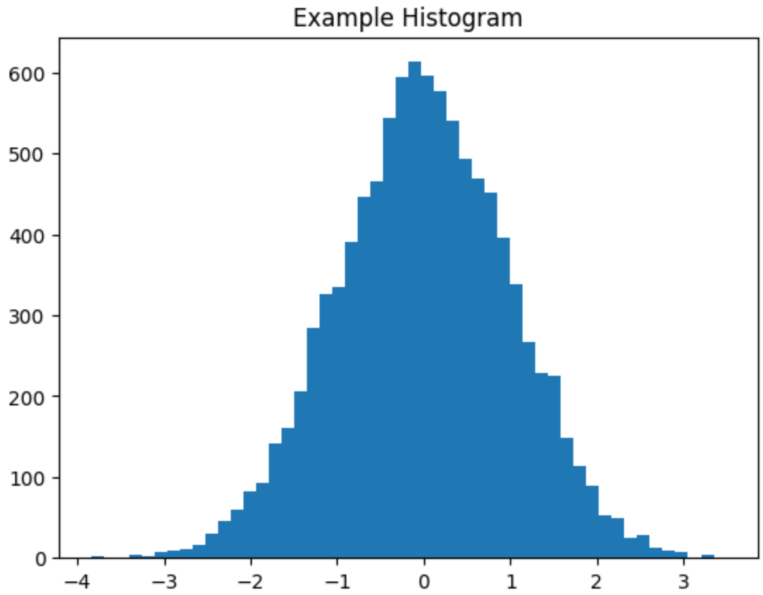

Figure 3: An example histogram with a normal distribution.

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

Definition 6 (Mean: $\bar{x}, \mu$)**.** The average of all individual data values. If there are n observations $x_1, x_2, ..., x_n$, the sample mean is calculated by

$$\bar{x} = \frac{\sum x_i}{n}$$

The mean of a **sample** is referred to using $\bar{x}$, while the mean of a **population** is referred to using μ.

Statistics come from **samples** (small subset of population) and **parameters** come from **populations** (all possible samples of what's being tested).

The mean is not **resistant** as it is sensitive to strong outliers in a distribution. The median *is* a resistant measure of the center of the distribution.

Definition 7 (Median)**.** The 'midpoint' of a distribution. Either the middle element (n is odd) or the average of the two middle elements (n is even) in a **SORTED** distribution.

Using both the mean and the median, one can predict the skew of the data. If a distribution is rougly symmetric without outliers, **the mean and median will be similar**. If a distribution is strongly skewed, **the mean will be pulled in the direction of the skew**. (Mean < Median for left-skewed, Mean > Median for right-skewed)

The **range** (max - min) is one way to show the variability of a distribution. Note that the range is *not* resistant.

Definition 8 (Standard Deviation)**.** The **standard deviation** (s_x, σ) measures the 'average' distance of the values in a distribution from the mean. Standard deviation is calculated by

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The squared stdev is known as **variance** (s_x^2, σ^2). Remember, s_x refers to a sample while σ refers to a population. Larger stdev indicates greater variation, but is not a resistant measure of variability. **Stdev measures variance around the mean; if the mean is skewed, so will stdev!**

The **Interquartile Range (IQR)** is another way to measure variance, using $IQR = Q_3 - Q_1$ where Q represents the quartiles. IQR can be thought of as the range of the 'middle half' of the distribution. *IQR is a resistant measure.*

$$\text{Lower Outliers} < Q_1 - 1.5 \times IQR \text{ or High Outliers} > Q_3 + 1.5 \times IQR$$

boxplots and the five-number summary = min, Q1, median, Q3, and max exist. **Boxplots don't show gaps, clusters, or multiple peaks.**

be careful with language- 'skews' is a shape, IQR and range are single numbers (no 'in the middle of the IQR')

2 Modelling Distributions

3 Two-Variable Data

4 Collecting Data

5 Probability

6 Random Variables and Distributions

7 Sampling Distributions

8 Confidence Intervals

9 Significance Tests

Significance Tests are like the opposite of confidence intervals. Instead of using a statistic to find a parameter, significance tests use statistics to test claims about a parameter.

Definition 9 (Significance Test)**.** A formal procedure for using observed statistics in order to decide between to competing claims (*hypotheses*) about parameters.

9.1 Basics of Significance Tests

9.1.1 Hypotheses

Definition 10 (Null Hypothesis, H_0)**.** A claim about a parameter that we weigh evidence **against** in a significance test. Usually a statement of 'no difference' (as claimed).

Definition 11 (Alternative Hypothesis, H_a)**.** The claim that we are trying to find evidence for. Directly contradicts the null hypothesis.

This "Null Hypothesis" and "Alternative Hypothesis" can be thought as trying to prove someone "guilty" or "not guilty."

For example, if a player claims they're a 80% free throw player, the null hypothesis would be $H_0 : p = 0.80$, and the alternative hypothesis would be $H_a : p < 0.80$.

The alternative hypothesis is **one-sided** because we suspect that the player makes less than 80% of his free throws. If we believe that it's equally plausible that they make more than 80% of their free throws, then we would use a **two-sided hypothesis**- $H_a : p \neq 0.80$.

Hypotheses express our beliefs before looking at the data. Molding a hypothesis around data shows nothing.

9.1.2 P-values

Definition 12 (P-value)**.** The probability of getting the values observed in the data under the assumption that the null hypothesis H_0 is true.

Small P-values provide convincing evidence for the alternative hypothesis because small values suggest the observed result is unlikely to happen when the null hypothesis is true. Similarly, large P-values provide convincing evidence for the null hypothesis because large values suggest the observed result is likely to happen due to chance if the null hypothesis is true.

In terms of probability notation: P-value = P(observed data | null hypothesis is true).

For two-sided tests, we look at the distance between the null hypothesis and the observed data. For example, if $H_0 : p = 0.5$, $H_a : p \neq 0.5$ and $\hat{p} = 0.65$ (observed proportion), then the P-value = P($\hat{p} \leq 0.35$ or $\hat{p} \geq 0.6$ | $p = 0.5$). We look at $\hat{p} \leq 0.35$ because $|p - 0.35| = |p - 0.65|$.

Based on the P-value, we make a conclusion about data:

- If the P-value is small (unlikely to happen by chance), then we "reject H_0" and conclude that there is convincing evidence for H_a (in context).
- If the P-value is large (likely to happen by chance), then we "fail to reject H_0" and conclude that there is not convincing evidence for H_a (in context).

How small does a P-value have to be in order to reject H_0? We use a given **significance value** for this boundary.

Definition 13 (Significance Level, α)**.** The value that we use as a boundary for deciding whether a P-value is significant enough to disqualify the null hypothesis. *α should be stated before data is produced (cherrypicking).*

$$\text{P-value} < \alpha \Rightarrow \text{reject } H_0 \ \Rightarrow \text{convincing evidence for } H_a \text{ in context}$$

$$\text{P-value} > \alpha \Rightarrow \text{fail to reject } H_0 \ \Rightarrow \text{not convincing evidence for } H_a \text{ in context}$$

If P is less than the significance level, we say that the result is "statistically significant at the $\alpha =$ ____ level." Alternatively, "the results were significant ($P = 0.03 < \alpha = 0.05$)." *Keep in mind a P-value is more informative than a statement of significance!*

NEVER "accept H_0" or conclude that H_0 is true! Always use 'reject' or 'fail to reject!'

9.1.3 Type I and Type II Errors

When drawing a conclusion from a significance test, our conflusion may be wrong. There are two types we can make with the conclusion process, helpfully named "Type I" and "Type II" errors. Only one type of error is possible at once.

Definition 14 (Type I and II errors)**.** A **Type I error** occurs when we reject H_0 when H_0 is true; the data gives convincing evidence that H_a is true despite being false.

A **Type II error** occurs when we fail to reject H_0 when H_a is true; the data fails to give convincing evidence that H_a is true despite it being true.

Note $P(\text{Type I error}) = \alpha$. However, as the significance decreases, $P(\text{Type II erorr})$ increases.

9.2 Tests About a Population Proportion

9.3 Tests About a Difference in Proportions

10 Estimating Means

11 Confidence with Means

12 Chi-Square Tests

13 Inference for Slopes and Tables