AP Statistics

Arnav Patri

March 12, 2022

Contents

Part I

One-Variable Data

Chapter 1

One-Variable Quantitative Data

Statistics is the science of collecting and analyzing data.
A data set is a collection of data on several **individuals**. These individuals can be anything.
Data provides values for **variables**, which describe some characteristic of an individual. A variable's **distribution** describes the frequency with which a variable takes on its possible values.

1.1 Graphically Displaying Distributions

Quantitative variables can either be **discrete**, having some countable set of possible values, or **continuous**, having an uncountably infinite set of possible values.

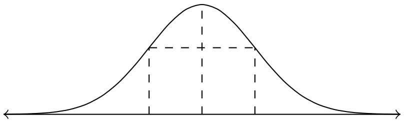
> The quantitative variable being described must always be defined, typically as a capital letter. An arbitrary particular value is denoted with a lowercase letter and a superscript is added to denote a defined value.

Dot plots assign the horizontal axis to the variable and show each value's frequency with a number of dots above, each corresponding to an individual.
Stem (and leaf) plots can only display quantities. The stem (vertical axis) corresponds to the first digit while the leaf (horizontal axis) corresponds to the remaining digits. The stem is always on the left, and the leaves should always be ordered from least to greatest. A key is required to denote the magnitudes displayed.
Histograms show bars that are assigned equal intervals. To find the length of each interval, the range can be divided by the number of classes and the result rounded up. Each bar's height shows the number of individuals within the class.
Time plots measure a variable over time.
A **Cumulative frequency curve** (OGIVE) is a line graph that shows the cumulative relative frequency, the sum of all lower classes (inclusive). Rather than each class being represented as a range, they are labeled by their medians.

Graphically Describing Distributions

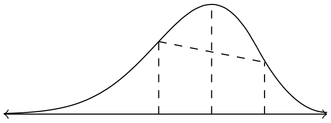
A distribution can be described by its shape, outliers, center, and spread (SOCS).

Shape In order for a distribution to be symmetric, values that are equidistant from the mode must have the same frequencies.

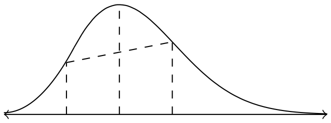


Symmetric

Skew is dependent on whether the tail is on the direction of the tail.



Skew Left



Skew Right

Outliers An **outlier** is a point that does not fit the general trend of the data. It is marked as an asterisk on a modified box plot.

Center The mean ($\bar{x}$), median (Q_2), and mode are measures of center. The relationship between these quantities is related to the skew.

Skew	Left	Center	Right
Relationship	$\bar{x} < Q_2 <$ mode	$\bar{x} = Q_2 =$ mode	$\bar{x} > Q_2 >$ mode

Spread The standard deviation, range, and interquartile range are measures of spread, The greater they are, the more variability there is in the data.

1.2 Numerically Describing Distributions

The **mean** is the sum of each value (not necessarily unique) in the distribution divided by the total number of values. It is also referred to as the average or expected value of its variable. For a sample, it is denoted using a bar over the lowercase form of its variable (X becoming $\bar{x}$).

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

The first, second, and third **quartiles**, denoted Q_1, Q_2, and Q_3, are the values with 25%, 50%, and 75% of the data below them respectively.[1] The second quartile is the **median**, and is, along with the mean, a measure of center.

The **range** is the difference between the highest and lowest values of a data set.

The **interquartile range**, denoted IQR, is the difference between the values of the third and first quartiles.[2] It therefore shows the "middle half" of the distribution.

$$IQR = Q_3 - Q_1$$

The **standard deviation** is the average distance from the mean. It is denoted by s with the subscript of its variable's lowercase form (for a sample). Along with the range and interquartile range, it is a measure of spread/variability.

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

[1] The probability of X falling below each quartile is equal to its number multiplied by 25%.

$$P(X < Q_1) = 0.25 \qquad P(X < Q_2) = 0.5 \qquad P(X < Q_3) = 0.75$$

[2] A box plot shows the interquartile range as the box and the median as the line within it.

It is crucial to note that the mean and standard deviation are only to be used when there are no outliers[a] due to their sensitivity.

[a] An **outlier** is any value that varies from the middle 50% by over 1.5 multiplied by the interquartile range.

$$(|Q_1 - x_i| \vee |x_i - Q_3|) > 1.5IQR \implies x_i \text{ is an outlier}$$

The **variance** is the square of the standard deviation and is accordingly denoted by s^2 with the appropriate subscript. [3]

A **linear transformation** shifts all values by the same amount b and/or proportionally to their values (with constant of proportionality a), resulting in a new variable.

$$y_i = ax_i + b$$

It should be noted that the measures of center are changed by both a and b, but only the former alters the measures of spread.

1.3 Density Curves

A **density curve**, typically denoted f, displays the relative frequencies of every possible value of its variable, making it effectively continuous. As such, the total area under the curve must be exactly 1 and its values must always be positive.

Because a density curve is effectively an idealized distribution curve, representing a population rather than a sample, different variables are used.

Value	Sample Data	Population Value
Mean	$\bar{x}$	μ
Standard Deviation	s	σ

The area under the density curve from its leftmost value to any given value is that value's **percentile**. It is the percentage of the data that that particular value exceeds, which means that it is also equal to the probability of the value being below it, and can be found using a cumulative distribution function, generally notated F.[4]

$$\text{Percentile} = P(X < x)$$

As a density curve is categorically continuous, the inclusivity of the binary relation of the probability statement is irrelevant to the value of the percentile.[a]

[a] The lack of regard for inclusivity with percentile, as well as other probability statements made regarding continuous distributions, is due to the area being $\int_{x_{\min}}^{x} f(t)\,dt$. As dt is a differential, it is infinitesimal, so adding or subtracting $f(x)\,dt$ to or from the integral has no tangible impact on its value.

The probability of X falling within two values is equal to the difference of their percentiles and the chance of X falling above a value is 1 minus its percentile.

$$(x_1 < X < x_2) = P(X < x_2) - P(X < x_1) \qquad P(X > x) = 1 - P(X < x)$$

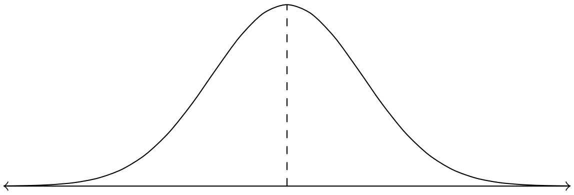
[3] The values of n, $\sum x_i$, $\bar{x}$, Q_1, Q_2, Q_3, range, IQR, s_x, and s_x^2 are all calculable automatically given a data set. The data can be entered into `L`n (`Stat/EDIT/1`), and `1-Var Stats` (`Stat/CALC/1`) can be performed with `L`n (`ALPHA/`n) as its parameter.

[4] The cumulative distribution function is the area under f from its minimum x value to the desired x value. It is thusly defined as the following integral:

$$F(x) = \int_{x_{\min}}^{x} f(t)\,dt$$

The median of a density curve is the point that splits the curve into two regions of equal area[5] while its mean is the point at which the entire curve would be balanced[6].

1.4 The Normal Distribution



The Normal Distribution

The **normal** distribution (typically denoted φ) is a density curve that is bell-shaped and symmetrical with inflection points one standard deviation from the mean.[7] Its mean and standard deviation are 0 and 1 respectively.

For a curve to be normal, it must be some transformation[8] of the normal distribution curve. A normal curve with parameters μ and σ is denoted $\mathcal{N}(\mu, \sigma)$, and its application to a variable X is denoted $X \sim \mathcal{N}(\mu, \sigma)$.[9]

To justify the normality of a data set, a modified box plot can be created and symmetry and a lack of variance shown.

Percentile can be found using the **cumulative normal distribution function** Φ.[10]

$$\Phi(x) = P(X < x)$$

[5] $P(X < x) = P(X > x) = 0.5$, as the areas are equal and the total area must be equal to 1. This is also consistent with the its prior definition.

[6] The mean of the curve is the x-value of the centroid of $f(x)$ over its entire domain.

$$\mu = \frac{\int_{x_{\min}}^{x_{\max}} x f(x)\,dx}{\int_{x_{\min}}^{x_{\max}} f(x)\,dx}$$

[7] The normal distribution is defined as such:

$$\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

[8] Changing the mean shifts the graph of the normal distribution horizontally while changing the standard deviation stretches or shrinks it vertically (the larger, the flatter, and the more data in the "tails").

[9] A normal distribution of X with parameters μ and σcan also be denoted as such:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

[10] The cumulative normal distribution function is defined as such, where erf is the error function:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,dt = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

Percentile can be evaluated on a calculator (for a standard or nonstandard normal distribution) using the `normalcdf` (`2nd/distr/2`) function:

$$P(X < x)\,\texttt{normalcdf}\,(\text{lower} : -\infty, \text{upper} : x, \mu : \mu, \sigma : \sigma)$$

The probability of X falling within a specific range can be found with this function as well:

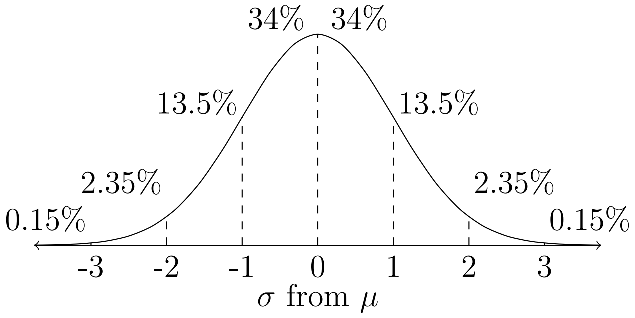
$$P(x_1 < X < x_2) = \texttt{normalcdf}\,(\text{lower} : x_1, \text{upper} : x_2, \mu : \mu, \sigma : \sigma)$$

The **inverse normal function** Φ^{-1} takes a percentile as an input and returns a value.[11]

$$\Phi^{-1}(P(X < x)) = x$$

The normal distribution follows the **empirical rule**, which states the following:

Range	$\mu \pm \sigma$	$\mu \pm 2\sigma$	$\mu \pm 3\sigma$	$(\mu, \mu+\sigma)$	$(\mu+\sigma, \mu+2\sigma)$	$(\mu+2\sigma, \mu+3\sigma)$	$(\mu+3\sigma, \infty)$
Proportion of Data	60%	95%	99.7%	34%	13.5%	2.35%	0.15%



To simulate a situation that shows normality, random numbers can be generated following a normal distribution.[12].

Standardization and z-Scores

The ***z*-score** is the number of standard deviations away from the mean.[13]

$$z = \frac{x - \mu}{\sigma}$$

z-scores allow the standard normal distribution to always be used[14], as it turns x into z, μ into 0 (as the numerator becomes zero), and σ into 1 (as z is measured in units of σ).

$$f(x) = \varphi(z) \qquad\qquad F(x) = \Phi(z)$$

z-scores can always be calculated, but normality must be confirmed in order for percentile to be calculable.

When answering a question involving calculating the value with a corresponding percentile, the following should be done:

1. Define the variable its distribution, using the given parameters (and justifying normality).

[11] A percentile's corresponding x value can be found using the `invNorm` (`2nd/distr/3`)function:

$$x = \Phi(P(X < x) = \texttt{invNorm}\,(\text{area} : P(X < x), \mu : \mu, \sigma : \sigma, \text{Tail} : \text{LEFT})$$

[12] To generate n random numbers following a normal distribution, the `randomnorm` (`math/PROB/6`) function can be used as such:

$$\texttt{randomnorm}\,(\mu : \mu, \sigma : \sigma, n : n) = \{x_1, x_2, \ldots, x_n\}$$

This list can then be stored using `sto` → $\texttt{L}_n$

[13] z-scores follow the same conventions as variables, Z being used for any value, z for any particular value, and z_i for a defined particular value. The subscript may also denote the variable that it is the z-score of, though, in which case the latter would be uppercase, as in Z_X, z_X, and $z_{x,i}$

[14] z should be used in place of x when working with φ and Φ.

(a) To show normality, the situation can be quoted, a modified box plot can be used, or a **normal probability plot**, which shows x vs z, can be created, and its straightness verified.

(b) If a graph is not given, draw the distribution, labeling the mean, given/desired particular value(s), and given/desired percentiles.

2. Calculate/denote the z-scores of any particular values.

   (a) Use the definition of z-score as the number of standard deviations from the mean if given particular values.

   (b) Use the inverse normal function if given percentiles.

3. Use the formula for z-score to calculate the desired variable.

$$x = \mu + z\sigma \qquad \mu = x - z\sigma \qquad \sigma = \frac{x - \mu}{z}$$

Chapter 2

One-Variable Categorical Data

Categorical variables assign labels that place each individual into one of several groups, while **quantitative variables** provide values that describe or measure some characteristic.
A **frequency table** shows the number of individuals that have a certain value while a **relative frequency table** shows the percentage of all individuals in the data set that have that particular value.
Bar graphs show each category as a bar, the height of which corresponds to its frequency.
Pie charts show each category as some fraction of a circle that is bounded by two radii. The areas of each slice is proportional to the frequency.

Part II

Two-Variable Data

Chapter 3

Two-Variable Quantitative Data

A **response** (dependent) variable measures the outcome while a **explanatory** (independent) variable explains the data.
The term **bivariate data** refers to data with two variables that are recorded for the same set of individuals.
A **scatterplot** is, as the name suggests, a set of points, each representing an individual, scattered about a grid with x and y axes that correspond to the response and explanatory variables respectively.[1] They are the most effective way for the relationship between two quantitative variables measured on the same individuals.

When storing data within lists, it is crucial that each element x_n and y_n correspond to the same individual.

A scatterplot can be described by its form, direction, and strength.
Form describes the nature (shape) of the variable's relationships (linear, polynomial, root, exponential, logarithmic, sinusoidal, etc.). More generally, from can be described as linear or nonlinear. **Strength** describes how strongly correlated the variables are.
Association direction describes how the explanatory

3.1 Correlation

The **correlation coefficient *r*** is equal to the sum of the products of the z-scores of each variable for each individual divided by one less than the number of individuals.

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum z_{xi} z_{yi}}{n-1}$$

The magnitude of r corresponds to strength while its sign indicates the **association direction**.

[1] To plot the data of two lists as a scatterplot, `stat plot` (`2nd/stat plot`)) can be pressed, and then a plot selected and toggled `On`, its first option selected for type, and the `X` and `Ylist` set to their corresponding lists. To cycle through individuals in X index order, it `trace` can be selected.

The following should be noted regarding r:

- It is the same for x against y and y agains x (commutative).
- It requires that both variables be quantitative.
- It is independent of the units of the variables.
- Its magnitude cannot exceed 1.
- It only measures the strength of linear relationships.

The correlation may be 0 even when there is a strong nonlinear pattern.

- It is not resistant, as its formula includes non-resistant means and standard deviations, meaning that it is prone to being affected by outliers.
- It does not completely describe bivariate data.

A strong correlation alone is not enough to ensure linearity.

- It is nor affected by linear transformations of either variable, even if they are not transformed correspondingly.

3.2 Regression and Residuals

The **line of best fit** of a scatterplot is the line that best predicts the data. A method for finding a linear line of best fit is **least-squares regression**.[2] The line generated by this method is referred to as the **least-squares regression line** or **LSRL**.

The LSRL enables predictions to be made regarding how the response variable when the explanatory variable is changed. It is defined as the line the minimizes the sum of the squares of the **residuals** (denoted e), the differences between the predicted and actual values of the response variable.[3] Due to the fact that it uses a sum of every point's values, the LSRL is not resistant to **influential observations**.[4]

The predicted value for the explanatory variable is denoted by putting a "hat" over the variable. The LSRL can be described as a linear transformation that maps x onto y:

$$\hat{y} = a + bx$$

Outliers are points with large residuals.

Predictions made within the data are **interpolations** while those made outside the bounds of the data are **extrapolations**.

Extrapolation should be avoided, as it is more likely not to be accurate than interpolation.

[2] On a calculator, a linear regression can be carried out by entering the data into two corresponding lists. From there, `LinReg(ax+b)` (`stat/CALC/4`) with the appropriate `X` and `Ylists` entered. This also calculates the r value. To store the equation, `Store RegEQ` can be set to $\mathtt{Y}_n$. Before a regression is run, though, `DiagnosticOn` (`2nd/catalog/DiagnosticOn`) should be selected.

[3] A list of residuals can be created by setting $\mathtt{L}_n$ to `RESID` (`2nd/list/7`), though the regression should first be run.

[4] Influential observations are often outliers in the y direction, especially in smaller data sets. To identify whether a point is an outlier, a modified box plot can be created. A point is influential if removing it markedly changes r

a and b can be derived as such: [5]

$$a = \bar{y} - bx \qquad\qquad b = r\frac{s_y}{s_x}$$

The definition of a makes it clear that the LSRL always passes through $(\bar{x}, \bar{y})$.
The **coefficient of determination r^2** is the percentage of the change/variation in the response variable that can be explained by the LSRL relating the response variable to the explanatory variable. It is maximized by the LSRL.
A **residual plot** shows the residuals of each data point. A random pattern suggests linearity.[6]
The regression line of a residual plot is always $\hat{y} = 0$, as the sum of all residuals is equal to 0. This also means that the mean of the residuals is equal to 0.
Clusters are groups of points that are similar.
To add a categorical variable, the data can be displayed using different symbols for each group.

3.3 Transformations to Achieve Linearity

The form of a relationships is not always linear, but the LSRL calculates a linear line of best fit. To resolve this, x and/or y can be transformed using powers, roots, or logarithms. This process is known as **linearization**.
The axes of the scatterplot are changed in accordance with the linearization.
An **exponential model** takes the form of an exponential function.

$$\hat{y} = ab^x$$

A **logarithmic model** takes the form of a natural logarithmic function.

$$\hat{y} = a + b\ln x$$

$$\boxed{\text{Exponentiating one variable and taking the logarithm of the other are equivalent operations.}}$$

[5] a and b can be derived by differentiating the sum of the squares of the residuals.

$$E = \sum (y_i - \hat{y})^2 = \sum (y - a - bx_i)$$

Because the error is being minimized, its derivatives are equal to 0.

$$\begin{aligned}
\frac{\partial E}{\partial a} &= \frac{\partial}{\partial a}\sum (y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i)^2 \left(\frac{\partial}{\partial a}[y_i - a - bx_i]\right) \\
&= \sum (y_i - a - bx_i)(-1) \\
&= \sum y_i - \sum a - \sum bx_i = \sum y_i - an - b\sum x_i \\
a &= \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial E}{\partial b} &= \frac{\partial}{\partial b}\sum (y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i) \left(\frac{\partial}{\partial b}[y_i - a - bx_i]\right) \\
&= \sum (y_i - a - bx_i)(-x_i) = \sum (x_i y_i - x_i\bar{y} + b\bar{x}x_i - bx_i^2) \\
&= \sum (x_i y_i - x_i\bar{y}) - b\sum (x_i^2 - \bar{x}x_i) \\
b &= \frac{\sum (x_i y_i - \bar{y}x_i)}{\sum (x_i^2 - \bar{x}x_i)} = \sum \left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right) = \sum \left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}\right) \\
&= \frac{1}{s_x(n-1)}\sum \left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x}\right) \\
&= \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \times \frac{s_y}{s_x} = r\frac{s_y}{s_x}
\end{aligned}$$

[6] This can be found on a calculator by setting `Ylist` to `RESID` (`2nd/list/7`) when creating a scatterplot. This should be done after running a regression.

A **power model** takes the form of a polynomial with direct variation.

$$\hat{y} = ax^b$$

Taking the logarithm of both sides simply results in a power model after it is re-expressed.

$$\begin{aligned} \ln \hat{y} &= a + b \ln x \\ e^{\ln \hat{y}} &= e^{a + b \ln x} \\ \hat{y} &= e^a e^{\ln x^b} \\ &= cx^b \end{aligned}$$

The model that bests fits the data is the one that minimizes s, the standard deviation of the residuals.

Chapter 4

Two-Variable Categorical Data

Segmented bar graphs have one bar per variable. Each bar is divided into segments according to their relative frequency. The height of each bar is equal to 1.

Side-by-side bar graphs show the distribution of one categorical variable for each value of another. The grouping of the bars is based on one of the categorical variables while the bars themselves show the frequencies of the values of the other.

Mosaic plots are variants of segmented bar graphs that display the number of individuals in a category by making each bar's width proportional to it.

A **two-way table** shows data on the relationship between two categorical variables for some group of individuals.

A **marginal relative frequency** is the proportion of individuals with a specific value for one categorical variable. A **joint relative frequency** is the proportion of individuals with a specific value for one categorical variable as well as a specific value for another categorical variable.

A **conditional relative frequency** is the proportion of individuals that have a specific value for one variable that also have a specific value for another variable.

If knowing the value of one variable helps in predicting that of another, there is an **association** between them. To test for association, the marginal and conditional relative frequencies compared. If they vary significantly, there is an association.

Part III

Collecting Data

Part IV

Probability, Random Variables, and Probability Distributions

Part V

Sampling Distributions

Part VI

Inference

Chapter 5

Inference for Categorical Data

5.1 Proportions