Memory Aids–Cumulative Final

$$\text{percentage}=(\text{width})\,(\text{height})$$

$$\text{z} = \frac{\text{x} - \text{avg}}{SD}$$

$$\text{x} = \text{avg} + \text{z}\,(SD)$$

$$\text{rms error } = \left(\sqrt{1 - r^2}\right) SD_Y$$

$$\text{slope} = \frac{r\,(SD_Y)}{SD_X}$$

$$\text{intercept} = \text{avg}_Y - (\text{slope})(\text{avg}_X)$$

$$SD_{box} = \left(\begin{smallmatrix} big \\ number \end{smallmatrix} - \begin{smallmatrix} small \\ number \end{smallmatrix}\right) \times \sqrt{\begin{smallmatrix} fraction\ \ with \\ big\ \ number \end{smallmatrix} \times \begin{smallmatrix} fraction\ \ with \\ small\ \ number \end{smallmatrix}}$$

$$EV_{\text{sum}} = \text{number of draws} \times \text{avg}_{\text{box}}$$

$$SE_{\text{sum}} = \sqrt{\text{number of draws}} \times SD_{\text{box}}$$

$$EV_{\%} = \text{pop }\%$$

$$SE_{\%} = \sqrt{\frac{(\text{pop }\%)\,(100\% - \text{pop }\%)}{\text{sample size}}}$$

$$EV_{\text{avg}} = \text{pop avg}$$

$$SE_{\text{avg}} = \frac{\text{pop SD}}{\sqrt{\text{sample size}}}$$

$$z = \frac{obs - EV}{SE}$$

$$\text{estimate} \pm z^*\,(SE)$$

$$SD^+ = \sqrt{\frac{\text{sample size}}{\text{sample size-1}}} \times SD$$

$$\text{df} = \text{sample size} - 1$$

$$SE_{\text{diff}} = \sqrt{(SE_A)^2 + (SE_B)^2}$$

$$\text{expected count} = \frac{(\text{row total})\,(\text{column total})}{\text{grand total}}$$

$$\chi^2 = \text{sum of } \left[\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}\right]$$

0.1 Chapter 1–Controlled Experiments

Controlled Experiments: In a controlled experiment, the investigators (experimenters or researchers) determine who receives the drug and who does not.

Well Designed Experiments:

- the control and treatment groups should be very similar
- **randomization**–randomly assign patients to each group
  - this reduces human bias
- **blind**–subjects don't know which group they are in
- **double blind**–neither the subjects or doctors know which group the subjects are in.
- **historical vs contemporary controls**–it is always better to use contemporary (current) people as the controls than to use past historical data.

confounding factors: If the two groups differ with respect to something other than the treatment, we say there is a confounding factor. (i.e. age, income, initial health, etc.)

- We should control for confounding factors to avoid their ill-effects.
- We can't control for all possible confounding factors.
- But we control for as many confounding factors as we can.

placebo: A placebo is a "neutral" treatment such as a sugar pill or saline injection.

0.2 Chapter 2–Observational Studies

Controlled Experiments: investigator decides who gets treatment and who gets the placebo.

Observational Study: subjects assign themselves to groups and investigators watch.

A very good controlled experiment can try to show causation. That the treatment causes the response.

An observational study can only show association. It can NOT show causation.

Simpson's Paradox: An association or comparison that holds for a combined group, can reverse direction when the data is examined by subgroups.

- Overall averages or percentages can be misleading.
- Conclusions that seem obvious when we only look at the total combined data can become quite different when the data is examined in more detail (broken into subgroups).

0.3 Chapter 3–Variables & Histograms

0.3.1 Variables

- A variable is a characteristic of a person or thing which is of interest in a study. For example, in a health study we might be interested in people's blood pressure, weight, height, cholesterol level, etc.

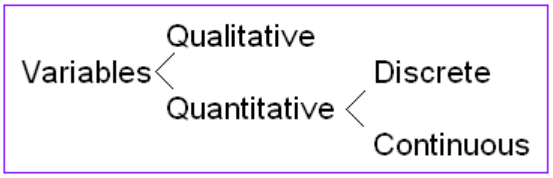
0.3.1.1 Types of Variables

- A quantitative variable is measured by a number (described with numbers)
  - age, height, weight, income, family size. etc.
- A qualitative variable is measured by a category (described with words)
  - occupation, race, religion, color of car

0.3.1.2 Even More Specific Types of Variables

Quantitative (number) variables can be *discrete* or *continuous*

- Discrete: values can only be specific values
  - Family size (0,1,2,3,4,5,6,7,8,...)
  - Number of classes you are taking (0,1,2,3,4,...)
  - Quiz scores where I grade in half point increments (0,.5, 1, 1.5, 2, 2.5, ...)
- Continuous: values can be any value in an interval
  - Age (any number from 0 to 120 including 3.2, 3.10384, 3.0293829, etc.)
  - Height
  - Weight
  - Income
- Continuous variables are always recorded to a certain number of decimal places, so they may look discrete (e.g. age in years).



0.3.2 Histograms

- the ***area*** of each block represents the percentage of data.
- We can find the height by
$$height = \frac{area}{width}$$
and we can rearrange the formula to find the area
$$area = height\,(width)$$
- The total area has to be 100%.
- Label the axes
  - x is height in inches $\Longrightarrow$ y label is "Percent per inch"
  - x is weight in kilograms $\Longrightarrow$ y label is "Percent per kilogram"
- The shape of the histogram is called the distribution of the data.

0.3.3 Percentiles

- 50th **percentile** means 50% of the data is *less* than the 50th percentile.
- 75th percentile means 75% of the data is *less* than the 75th percentile.

0.4 Chapter 4–Average & Standard Deviation

0.4.1 Measure Center & Spread

Average: a way to measure the center of the data; a "typical" value
$$\text{average} = \frac{\text{sum of data values}}{\text{number of data values}}$$
- the average is the "balance point" of the histogram.

Median: another way to measure the center of the data; a "typical" value
- Half the data values in a list are less than the median and half the data values are greater than the median.
- Half the area of a histogram is to the left of the median, and half the area is to the right.
- Find the Median:
  - Write the numbers from smallest to largest.
  - If the number of values is ODD, the median is the middle number.
  - If the number of values is EVEN, the median is the average of the two middle numbers.

Sensitive Measures of the Center:
- The average is sensitive. Outliers or data values in a long tail can pull the average towards the tail.
- The median is NOT sensitive. It is not as affected by long tails or outliers.
- For histograms that are not symmetric, statisticians sometimes prefer to use the median to measure the center of the data instead of the average.

Standard Deviation: a way to measure how spread out the data values are

- The more spread out the data is, the bigger the standard deviation.
- The standard deviation is the typical distance of the data from the average.
- The standard deviation is the **root mean square (r.m.s)** of the **deviations** of the data values from the average.

Facts about averages, medians, and standard deviations

- If we add the same number to each data value
  - the average goes up by that number
  - the median goes up by that number
  - the standard deviation does NOT CHANGE
- If we multiply each data value by the same number
  - the average is multiplied by that number
  - the median is multiplied by that number
  - the SD is multiplied by that number

0.4.2 Outliers

Outlier: An observation that doesn't fit with the rest of the data.

- Usually we call something an outlier if it is more than 3 SDs from the average.

0.4.3 Cross-Sectional versus Longitudinal Studies

0.4.3.1 Cross-Sectional versus Longitudinal Studies

- NHANES is a cross-sectional study because it consists of a cross-section of the U.S. population at some time point.
  (They looked at lots of different groups of people at the same time.)
- Cross-sectional studies give us a "snapshot" of a population at just one moment in time.
- A longitudinal or cohort study is one in which we follow a group of people over time.
- If you want to draw conclusions about what happens over time, you must have a longitudinal study!

0.5 Chapter 5–Normal Distribution

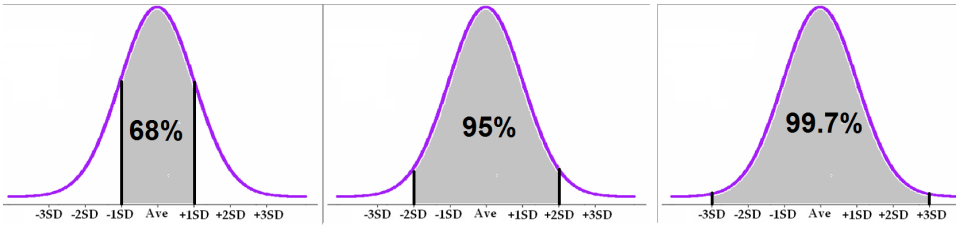
0.5.1 Normal Distribution

- The normal distribution is symmetric.
- The normal curve is bell shaped.
- The normal curve is centered at the average.
- There are many normal curves with different averages and standard deviations.
  - They all have a similar shape and share similar properties.
  - The shape is affected by the average and SD.
- The total area under a normal curve is 100%.

0.5.2 The 68-95-99.7 Rule of Thumb

In *any* normal distribution:

- Approximately 68% of the observations fall within 1 standard deviation of the average.
- Approximately 95% of the observations fall within 2 standard deviations of the average.
- Approximately 99.7% of the observations fall within 3 standard deviations of the average.



0.5.3 Normal Distribution–Standard Normal Distribution

- There is a special normal distribution.
- We call it the standard normal distribution.
- The average is 0.
- The standard deviation is 1.

0.5.4 Normal Distribution–Standardization

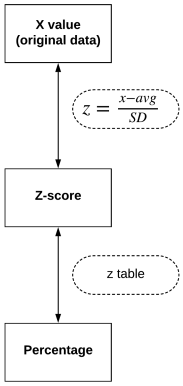
In general we can find the z-score for any observation and any normal distribution:

$$z = \frac{x - \text{average}}{\text{SD}}$$

z tells us the number of standard deviations from the average.

0.5.5 Normal Approximation

If the data looks like a normal curve we can use the normal curve to estimate percentages.



We can estimate the percentage of people in an interval by

1. use $z = \dfrac{x - avg}{SD}$ to get a z-score
2. use the standard normal table to find the percentage

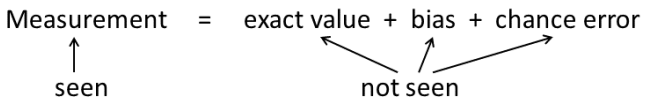
If we know the percentage and we want to find the x value

1. use the standard normal table to find the z score with that percentage
2. use $z = \dfrac{x - avg}{SD}$ and solve for x
   (You can also use $x = z \cdot SD + avg$ to make it easier.)

Remark. There are many data sets that are normally distributed. But there are so many more that are NOT normally distributed. The normal curve approximation doesn't work unless the data actually looks like the normal curve.

0.6 Measurement Error

0.6.1 Measurement Error



Bias: is something in the measurement process that **affects all measurements the same**, either by increasing all measurements, or decreasing all measurements.

Chance error: is variation in the measurements that are **simply due to chance**.

Repeated measurements: when we measure something over and over we get slightly different results each time.

- To estimate the exact value, we take the average of the measurements.
- The SD of a series of repeated measurements estimates the chance error in a single measurement.
- (The SD tells us how much we are likely to be off when we take a measurement.)
- If you can figure out the bias remove it.
- The more measurements you can do, the more accurate you will be.
- Repeated measurements usually follow a normal distribution.
  - So about 68% of the time you will be within

    $$\text{true value} \pm 1\,(SD)$$

    So we say you are likely to be off by $1 \cdot SD$.
  - And about 95% of the time, a new measurement will be between

    $$\text{true value} \pm 2\,(SD)$$

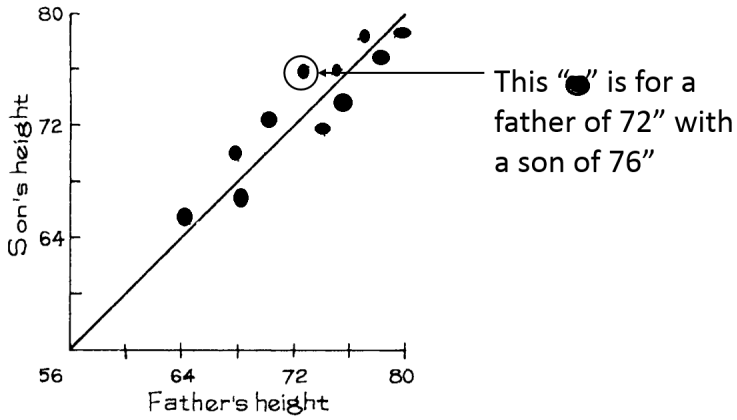
Remark. There is always chance error when you measure.

There may or many not be bias.

0.8 Correlation

A scatterplot shows the relationship between two variables.

Example 1. A scatterplot of the heights of father-son pairs.

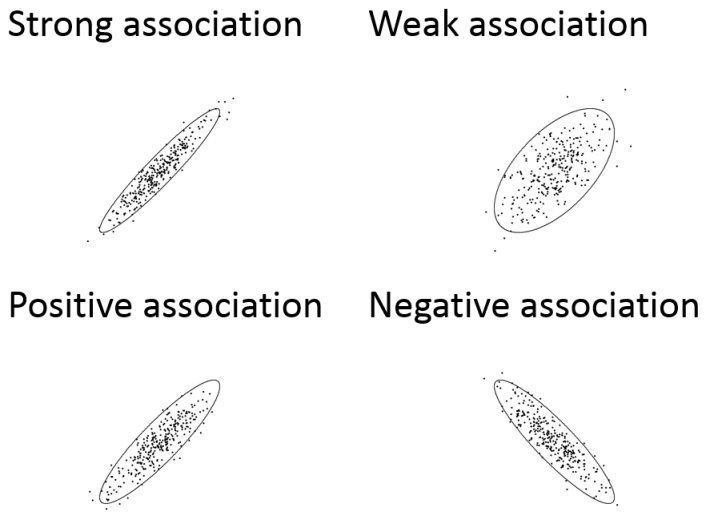


Each dot represents one father-son pair.

Scatterplot Axes

- The independent variable is plotted on the x-axis.
- The dependent variable is plotted on the y-axis.
- The independent variable (x-axis) is sometimes used to predict the dependent variable (y-axis).

0.8.1 Scatter Plots and Association



Strong Association: When the points of a scatter diagram are tightly clustered around a line, there is a strong linear association between the variables.

Information about one variable helps in predicting the other.

Weak Association: When there is a weak linear association, the points are scattered loosely about the line.

Knowing about one variable does not help much in predicting the other.

Positive Association: As the values of one variable increase, so do the values of the other.

(If x goes up, y tends goes up as well.)

Negative Association: As the values of one variable increase, the values of the other variable decrease.

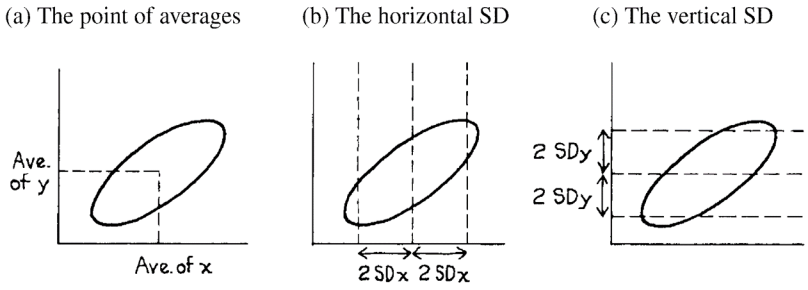
(If x goes up, y tends to goes down.)

Remark. **Association does not imply causation!!!**

0.8.2 Measuring the Relationship

The *linear* relationship between two variables can be described with the following 5 number summary:

ave_x, ave_y, SD_x, SD_y, r



Ave_x**:** average of the x values, measures the center of cloud

Ave_y**:** average of the y values, measures the center of cloud

SD_x**:** standard deviation of the x values, measures the spread of the cloud from side to side

SD_y**:** standard deviation of the y values, measures the spread of the cloud from top to bottom.

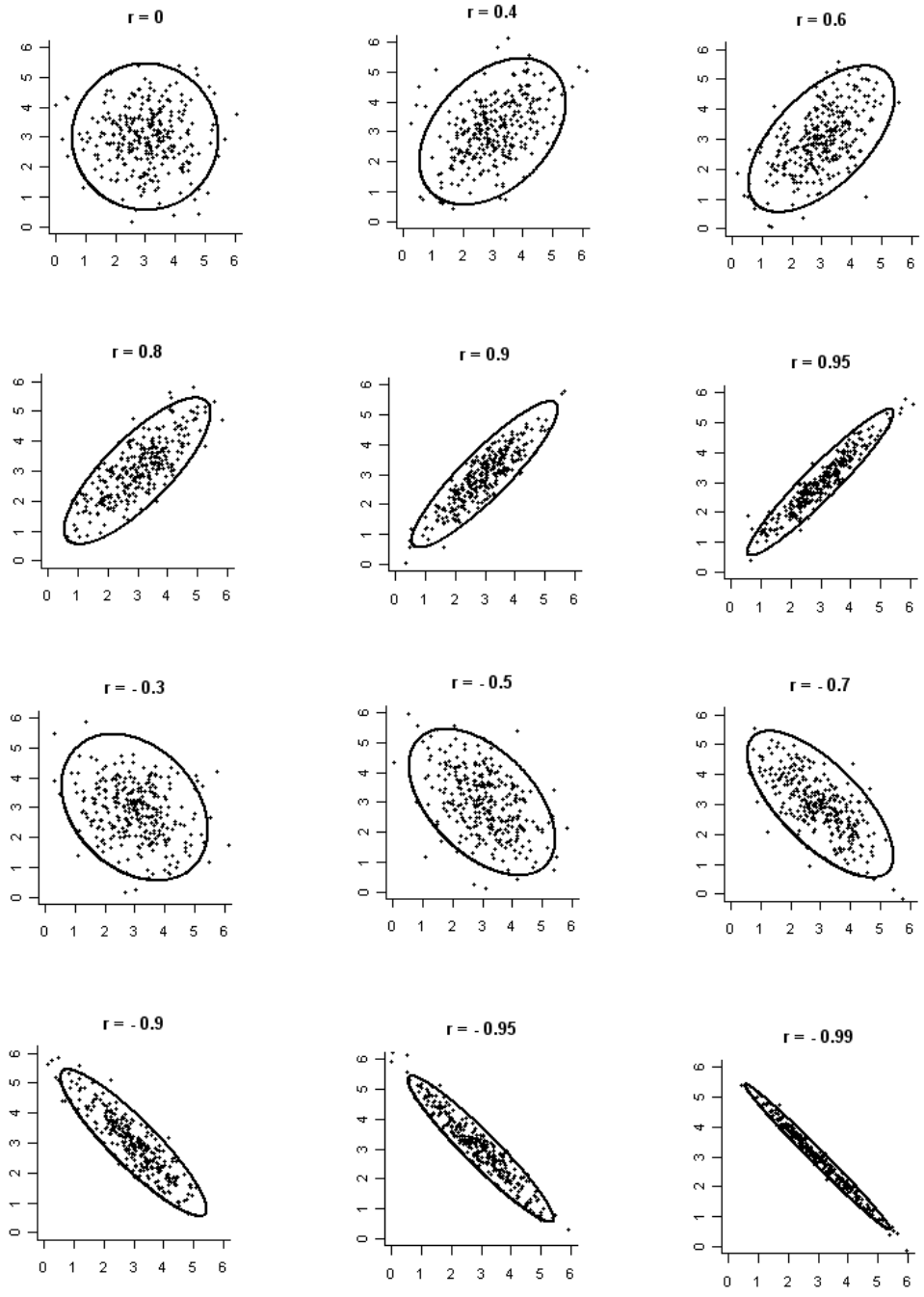
r**:** the correlation, measures how close the points are to a line

The point of averages is the point found by the average of x and the average of y.

Most of the data falls within 2 SDs of the point of averages.

0.8.3 Correlation

The correlation coefficient, r, measures the strength of the *linear* association between two variables.



- r is always between -1 and 1
- r is -1 if the points are on a line with negative slope
- r is 1 if the points are on the line with positive slope
- r is 0 if there is no linear relationship (points seem to be randomly scattered)

0.9 Chapter 9–More about Correlation

Some facts about correlation:

- r is a pure number (it has no units)
- r does not change if you
  - switch x and y
  - add the same number to each x value
  - add the same number to each y value
  - multiply each x value by a *positive* number
  - multiply each y value by a *positive* number

0.9.1 Interpreting the correlation

Be careful how you interpret r:

- $r = 0.6$ does NOT mean "twice as much association" as $r = 0.3$.
- $r = 0.6$ does NOT mean that 60% of the points are clustered tightly around the line.
- r can be misleading if there are outliers
- r measures linear association
  - r is misleading if there is a nonlinear association

ALWAYS DRAW THE SCATTER DIAGRAM!

Remark. Correlation is very sensitive to outliers.

Remark. The correlation from a plot of averages or rates is called an ECOLOGICAL CORRELATION.

Ecological Correlations are artificially strong!

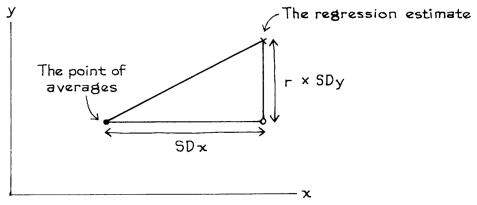
0.10 Chapter 10–Regression

The regression line is used to predict the y variable when we know the x variable

The regression line:

- goes through the point of averages (ave_x, ave_y)
- has slope $\dfrac{r \cdot SD_y}{SD_x}$

Figure 2. Regression method. When x goes up by one SD, the average value of y only goes up by r SDs.



0.10.1 The Regression Method for Predictions

To predict or estimate the value of y when you know the value of x

1. Find out how many SDs your x value is above or below the average in the x variable.

$$z = \frac{x - avg_x}{SD_x}$$

2. Multiply the answer to step 1 by r.
3. The answer to step 2 tells you how many SDs the new y variable is above or below the average in the y variable.

0.10.2 Notes

- If you switch the x and y variables the correlation stays the same but the regression line changes.
- Don't use a regression line if the data doesn't look linear.

0.10.2.1 Regression Effect

Have you ever heard of "regression toward the mean"?

You might have heard it used to describe the fact that extremely tall fathers tend to have sons that are still taller than average, but not quite as tall as their dads. On average, when the fathers are short, their sons will be slightly taller than their dads.

Test-Retest Situations: In test-retest situations, people with low scores tend to improve and people with high scores tend to do worse.

Why? Chance error.

$$\text{observed value} = \text{true value} + \text{chance error}$$

The chance error can be positive or negative.

Example 2. Imagine taking an IQ test twice. If you got lucky the first time, you might get a higher score than you deserved. The next time you took the test, your score would probably be lower. If you scored very low, you were probably unlucky to some extent and your score will probably be higher the next time.

0.10.2.2 The Regression Fallacy

Regression fallacy is attributing the regression effect to something other than chance error.

Example 3. A group of people get their blood pressure measured.

Only those that have high blood pressure return and have their blood pressure measured again.

We expect their second measurements to have a *lower* average than their first measurements, JUST due to the regression effect.

Attributing this apparent reduction of blood pressure to a change in behavior or your medicine or them exercising more is an example of the regression fallacy.

0.11 Chapter 11–r.m.s. error for regression

The distance of a point above or below the regression line is the *error* or *residual*.

$$\text{error} = \text{actual } y \text{ value} - \text{predicted } y \text{ value}$$

The *r.m.s. error for regression* says how far typical points are above or below the regression line.

The r.m.s. error measures how good a prediction is. It says how large the errors are likely to be.

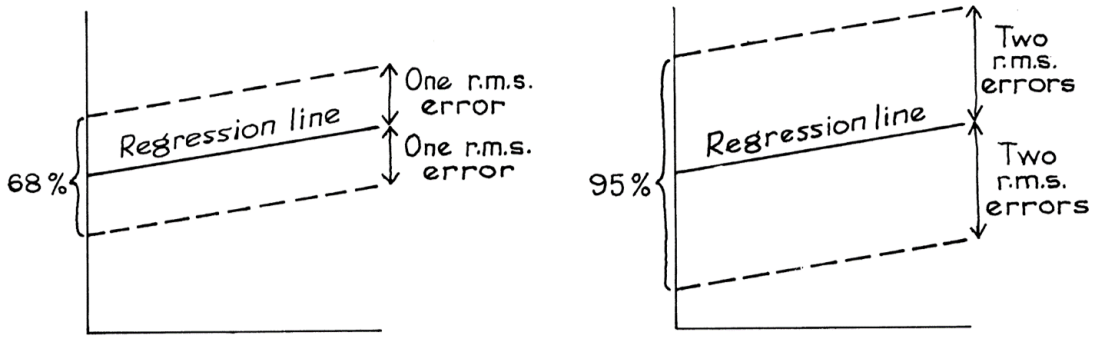
The r.m.s. error is the root mean square (r.m.s.) size of the errors.

$$\text{r.m.s. error} = \sqrt{\frac{(error_1)^2 + (error_2)^2 + (error_3)^2 + \cdots + (error_n)^2}{n}}$$

To calculate the r.m.s. error use the following shortcut:

$$\text{r.m.s error} = \sqrt{1 - r^2}\,(SD_y)$$

If the scatter diagram is football-shaped, the r.m.s. error is like an SD for the regression line.



About 68% of the points in a scatter diagram are within 1 r.m.s. error of the regression line.

About 95% of the points are within 2 r.m.s. errors of the regression line.

What we like to say:

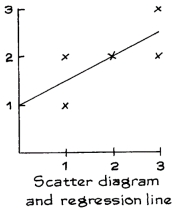
- you are likely off by 1 r.m.s. error
- we aren't surprised unless we are off by more than 2 r.m.s. errors

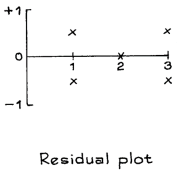
0.11.1 Plotting the Residuals

*Remember, *residual* is another term for *error*.

The residual says how far the point is above or below the line.

In some situations, regression is not appropriate. Plotting the residuals can help to determine whether or not this is the case.

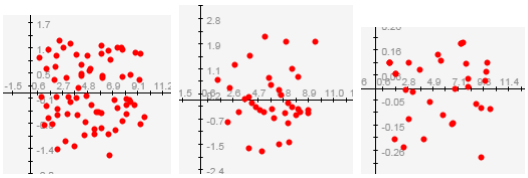




If regression is appropriate (scatterplot is football shaped):

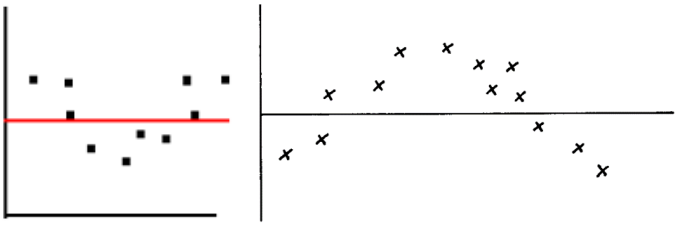
- The residuals look like they are randomly scattered around the horizontal line at 0.
  - The residuals average out to zero.
- The vertical spread of the residuals seems to be fairly constant for all the x values.

Example 4. Here are some residual plots that indicate regression is appropriate.

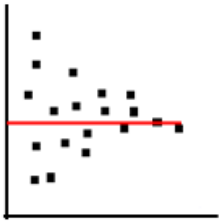


Cautions: Don't use regression if:

- If there is a strong pattern in the residuals, it is a mistake to fit the regression line.



- If the vertical spread isn't similar for all the x values, don't use regression.



0.11.1.1 Homoscedastic versus Heteroscedastic

- If the vertical spread is the same for all the x values, we call the data "homoscedastic".
- If the vertical spread isn't the same for all the x values, we call the data "heteroscedastic".

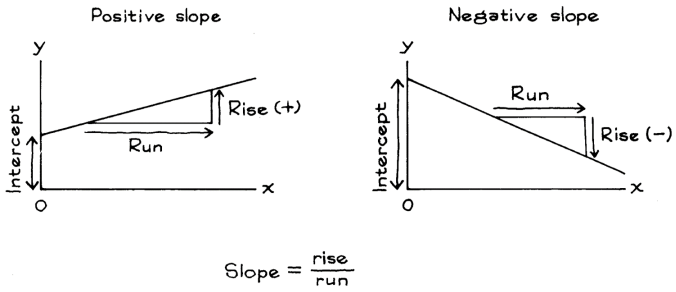
The r.m.s error is only appropriate for homoscedastic scatter diagrams.

0.11.1.2 Caution

If you haven't seen the diagram, you have to assume that it is football-shaped to do regression. If this assumption is incorrect, your answers may not be accurate.

0.12 Chapter 12–The Regression Line

Any line can be described in terms of its slope and intercept. $\boxed{y = \text{slope}\,(x) + \text{intercept}}$



The regression line has

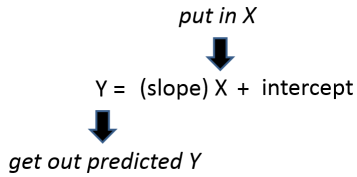
$$\text{slope} = \frac{r \cdot SD_Y}{SD_X}$$

$$\text{intercept} = (\text{average}_Y) - (\text{slope})\,(\text{average}_X)$$

0.12.1 Using the equation of the regression line

The equation can be used to get a prediction by putting in the value for X and getting out the predicted value for Y.

The regression equation:



0.12.2 Association is not causation!

The regression line describes the data that you see, but it can NOT be relied on for telling us how Y will respond if the investigator changes X unless it is a controlled experiment. In an observational study there are too many confounding factors.

0.12.3 Intercept might not be logical

Sometimes the y-intercept doesn't have a practical or logical interpretation.

That doesn't mean that the y-intercept is wrong, it just means that we picked the y-intercept that makes the line the best fit for the data.

0.12.4 Extrapolation

Never use regression to predict outside the range of your data!

Extrapolation is when we use a regression line for predictions, but the x values are far outside the range of the x values used to obtain the line.

We do extrapolate sometimes if it is the only information we have, but these predictions are often inaccurate.

We can't guarantee that the relationship between x and y remains the same for different x values.

0.12.5 More cautions

A regression line can be found for any two continuous variables. But it may be misleading if:

- There is a non-linear association between the variables.
- The regression is silly.

0.12.6 Regression Method or Regression Line

You can do the regression method from chapter 10 or the regression line from chapter 12 to find predictions. They both give the same results. On the exam you will probably be asked to use both methods to show that you can.

0.1 Chance

The **chance** of something occurring is the percentage of times it is expected to happen when the basic process is repeated independently, many times.

The Rules of Chance:

- Chances are between 0% and 100 %.
  - (or between 0 and 1 for fractions)
  - chances are never negative!
  - chances are never more than 100%
- The chance of something happening is 100% minus the chance of it not happening
  - e.g. if the chance you win is 45%, then the chance you don't win is 100% - 45% = 55%
  - e.g. if the chance of getting no prizes is .25, then the chance of getting at least one prize is $1 - .25 = .75$.
  - e.g. the chance something happens at least once is $1 - P(\text{no times})$
- When drawing at random, all possible outcomes have the same chance of being picked.

0.1.1 Drawing from a box

- with replacement (we look at the first ticket and then we put it back in the box and mix up the tickets before choosing the second ticket)
- without replacement (we do not put the first ticket back)

0.1.2 Conditional Probability

A conditional probability is when an event is affected by something.

0.1.3 The Multiplication Rule

The Multiplication Rule: The chance that two things will both happen equals the chance that the first will happen multiplied by the chance that the second will happen given that the first has happened.

0.1.4 Independence

Independent: Two things are independent if they don't affect each other.

Examples:

- Coin tosses are independent of each other
- Rolls of a die are independent of each other
- Draws from a box are independent if we are drawing with replacement
- Draws from a box are NOT independent if we are drawing without replacement

Two things are independent if the chances for the second (given the first) are the same no matter how the first turns out. Otherwise, the two events are dependent.

0.1.5 Mutually Exclusive

Mutually Exclusive: Two outcomes are mutually exclusive if the occurrence of one prevents the occurrence of the other (if one happens, the other cannot).

Addition Rule: If two events are mutually exclusive, the chance one happens *or* the other happens is equal to the sum of the two chances. If they are not mutually exclusive, this sum will be too big.

0.1.5.1 What's the difference between *mutually exclusive* and *independent*?

"Mutually exclusive" and "Independence" are two completely different ideas.

- Two events are *mutually exclusive* when one event can't happen if the other event already happened.
- Two events are *independent* when knowing that one event already happened, does NOT change the chance of the other event occurring.

0.1.6 When do I add? When do I multiply?

- You add the chances when you want the probability that A **or** B happens.
  - This only works if A and B are mutually exclusive!
- You multiply the chance when you want the probability that A **and** B **both** happen.
  - Don't forget to use conditional probabilities, unless you know A and B are independent.

0.2 Chance Error and Law of Averages

0.2.1 Chance Error

If we toss a coin many times,

$$\text{number of heads} = \text{half the number of tosses} + \text{chance error}$$

The "law of averages" says that for a large number of tosses, the **chance error** is likely to be

- LARGE in absolute terms (see pg 89 in course reader)
- SMALL compared to the number of tosses

0.2.2 Calculate chance error

$$\text{chance error} = \text{actual results} - \text{expected results}$$

Remark 1. **The chance error** is the difference between what we expect and what we actually get.

0.2.3 The Law of Averages

In terms of percentages, the "law of averages" says that as the number of tosses increases, the percentage of heads is likely to get closer and closer to 50%.

Remark 2. The Law of averages does not say "you are due for a win". The law of averages does not change or affect the chance for an individual trial.

0.3 Box Models

0.3.1 Questions to ask to build box model

1. What is the quantity of interest? Are we interested in
   - the sum of the draws
   - the average of the draws
   - the percentage of 1's in the draws
2. How many draws?
3. How many tickets go in the box?
4. What numbers go on the tickets?
5. Are the draws made with or without replacement? **(for this class we will do with replacement for all box models)**

"The total is like the sum of _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ draws from the box."

0.3.2 Classifying and Counting

If you want to count how many times something happens, redo your box model.

Use 1's for what you want to count.

Use 0's for everything else.

Then the number of times something happens is like "the sum of _ _ _ _ _ _ _ draws from the box".

0.3.3 Find SD of Box

You can find the average of the box quickly enough. But I will find the SD of the box for you, unless there are just 2 numbers in the box.

0.3.3.1 Find the SD of a box with only two numbers

If your box only has two different numbers, a big one and a small one, there is an easy shortcut to find the SD_{box}. (The two different numbers can be repeated.)

$$SD_{box} = \left(\begin{smallmatrix} big \\ number \end{smallmatrix} - \begin{smallmatrix} small \\ number \end{smallmatrix}\right) \times \sqrt{\begin{smallmatrix} fraction\ with \\ big\ number \end{smallmatrix} \times \begin{smallmatrix} fraction\ with \\ small\ number \end{smallmatrix}}$$

0.4 Expected Value & Standard Error

The *observed value* is the number of heads that you actually get.

The observed value differs from the expected value by some amount of chance error.

The standard error (SE) gives us the likely size of the chance error.

0.4.1 Different possible sample sums

Chance Variability: For each sample we take, we expect to get different results.

Remark 3. If we drew 25 tickets from a box and looked at the sum and recorded it, and then we drew a different 25 tickets and recorded the sum, then we kept doing that, we would get a whole list of the sums we got.

Expected Value: EV_{sum} tells us the average of all those sums we got

Standard Error: SE_{sum} tells us the standard deviation of all those sums we got

Remark 4. So when we draw from a box and look at the sum...

- EV_{sum} tells us what we expect the sum to be.
- SE_{sum} tells us how much we are likely off by.
  - SE_{sum} tells us how big the chance error will probably be.
  - 1 SE_{sum} is the "give or take"

- We say we are likely to be within 1 SE of the *expected value EV*.
- We say we won't be surprised unless we are more than 2 SEs from the *expected value EV*.

Everything we said about Sums applies to Percentages and Averages as well.

$$\begin{aligned} EV_{\text{sum}} &= \text{number of draws} \times \text{avg}_{\text{box}} \\ SE_{\text{sum}} &= \sqrt{\text{number of draws}} \times SD_{\text{box}} \end{aligned}$$

$$\begin{aligned} EV_{\%} &= \text{population \%} \\ SE_{\%} &= \sqrt{\frac{(\text{population \%})\,(100\% - \text{population \%})}{\text{sample size}}} \end{aligned}$$

$$\begin{aligned} EV_{avg} &= \text{population average} \\ SE_{avg} &= \frac{\text{population SD}}{\sqrt{\text{sample size}}} \end{aligned}$$

Remark 5. In general we use "standard deviation" when we find the SD of a list of numbers.
We use the words "standard error" when we talk about our possible sums (or percentages or averages).

0.5 Normal Curve and Sums, Averages, Percentages

0.5.1 Normal Curves and Sums

For a *large number of draws*, the sum of the draws will follow the normal curve with average EV_{sum} and standard deviation SE_{sum}.

68% of the time the sum of the draws will be within $EV_{sum} \pm SE_{sum}$

95% of the time the sum of the draws will be within $EV_{sum} \pm 2SE_{sum}$

We can use the normal curve to find chances for sums.

We use EV_{sum} and SE_{sum} to get standard units (the z-score):

$$z = \frac{sum - EV_{sum}}{SE_{sum}}$$

Notes for using the normal curve for sums:

- It does not matter what is in the box.
- The histogram for the tickets in the box does not have to follow the normal curve.
- The sum of the draws will follow the normal curve, even if the tickets in the box are 0's and 1's!
- In fact, we don't even need to know what is in the box, we just need to know the average and the SD of the box.
- **But we do need to have a LARGE NUMBER OF DRAWS!**

0.5.2 Percentages

Everything in about sums in 0.5.1 is true for Percentages and Averages as well.

68% of the time the sample percentage will be between $EV_{\%} \pm SE_{\%}$.

95% of the time the sample percentage will be between $EV_{\%} \pm 2\,(SE_{\%})$.

$$z = \frac{\% - EV_{\%}}{SE_{\%}}$$

0.5.3 Averages

Everything in about sums in 0.5.1 is true for Percentages and Averages as well.

68% of the time the sample averages will be within $EV_{avg} \pm SE_{avg}$

95% of the time the sample average will be within $EV_{avg} \pm 2\,(SE_{avg})$

$$z = \frac{avg - EV_{avg}}{SE_{avg}}$$

Remark 6. For averages, we need at least 30 draws to use the normal curve (or the data needs to look normal).

0.5.3.1 The SD and the SE

The SD says how far data values are from average – for typical *individual* data points.

The SE for the average says how far sample averages are from the population average – for typical *samples*.

0.6 Normal Approximation for Histograms

0.6.1 Empirical and Probability Histograms

Empirical Histograms represent the actual data

Probability Histograms represent the theoretical chances

For both empirical and probability histograms, the areas represent the percentages.

The more repetitions we do, the closer the empirical histogram will get to the probability histogram.

Remark 7. The probability histograms can be any shape, they don't have to be normal, even for large draws unless it is **sums, percentages, or averages.**

Remark 8. If you are looking at **sums, percentages, or averages**, the histogram will look like the normal curve for a large number of draws.

0.7 Populations and Samples

Population: The set of all people, items, events, objects, etc. that are of interest.

- everyone in the US
- every Honda car built in 2010
- every English major at 4 year universities

Sample: A smaller part of the population on which we actually collect data.

- every thousandth person in the US
- every 500th car to come out of the Honda Factory
- 300 English majors selected from 4 year universities

It is usually too difficult, time consuming, or expensive to collect data on the entire population. If our sample is representative of the population, we can infer information about the population from our sample data.

Parameter: numerical facts about the population that we are interested in

Parameters are usually unknown.

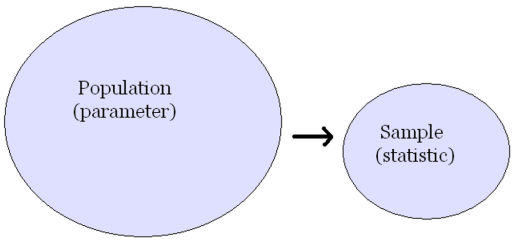
A goal of statistics is to estimate parameters.

Statistic: a value calculated from the sample

We use statistics to estimate parameters.

So we use the sample statistics to draw conclusions about our population parameters.

The science of deducing properties of a population from a sample is known as statistical inference.



Example 9. An investigator wishes to know what percentage of voters in the US will vote 'Republican' in the next election. We take a sample of 100 voters and find 37% plan to vote republican.

- *Population*: Voters in the US.
- *Parameter*: percentage of US voters who will vote Republican in the next election.
- *Sample:* the 100 voters in sample
- *Statistic:* the 37% in our sample who plan to vote Republican

0.7.1 Representative Samples

If we want to be able to generalize our sample results to a population, we need our samples to be representative of the population. This means that the characteristics of the sample should be similar to the population.

0.7.2 Bias

Bias is any systematic error in an estimate.

If our sampling method will systematically favor an outcome, we say we have "sampling bias".

So if for some reason our sample is more likely to have republicans (such as we asked outside a republican caucus), we say we have "sampling bias".

0.7.2.1 Selection Bias

Selection Bias: A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample.

This often occurs if the sample is self-selected or if there is any human judgment in picking a sample.This happens if you select people from telephone books, or you only select students from the honors class, you only select students inside the library, etc.

0.7.2.2 Non-response bias

Non-response Bias: is bias introduced by important differences between those who respond and those who do not.

In general:

- The non-respondents can be very different from the people who did respond.
- Lower-income and upper-income people tend not to respond to questionnaires. So the middle-class is over-represented.
- If you interview people at home:
  - women are more likely to answer the phone

  - people who are not at home when the interviewer calls may be different from those who are at home
    * working hours, social background, etc.

- If you let people choose to participate in a survey (click-in polls or mail in surveys, etc.) people who take the time to respond often care much more than the general population.
  - They often feel differently about the issue than the general population as well.
- People with strong feelings (positive or negative) are more likely to respond.
- Dissatisfied people are the most likely to respond.

0.7.3 Convenience Sampling (not a good choice)

Subjects are chosen because of their convenient accessibility to the researcher.

This also includes when you let willing people self-select to be in the study. (This would include all mail-in surveys and click-in online polls.)

The subjects are selected just because they are easiest to recruit for the study.

Convenience samples are almost NEVER representative of the general population.

0.7.4 Probability Methods for Sampling

To ensure that the sample is representative of the population, it should be chosen using *probability methods.*

The elements in the sample should be chosen randomly from the population.

When probability-based sampling methods are used, the interviewers don't have any discretion about whom they interview.

0.7.4.1 Simple Random Sampling (SRS)

A simple random sample (SRS) is selected by drawing at random without replacement from the population.

By design, each member of the population has an equal chance of being selected.

Simple random sampling is great when it is feasible.

0.7.5 More problems with sampling

Probability methods attempt to minimize bias, but there are still problems due to:

- badly asked questions
- interviewer control (persuading people to give certain answers whether consciously or subconsciously)
- talk is cheap (people lie, have good intentions but don't follow through, etc)

0.7.6 Are bigger samples better?

Bigger samples are always better IF they are good samples.

Bad samples will always be bad no matter how big the sample size.

0.7.7 Accuracy of a Sample (Sample Size versus Population Size)

Example 10. Brigham City has a population of about 49,000 while about 8.2 million people live in New York City. If we draw a sample of 2,000 from each of these populations and compute the expected percentage of senior citizens, which would we expect to be more accurate?

As long as the population is much bigger than the sample, the population size doesn't affect our estimates. Only the actual sample size affects the estimates.

The **absolute size** of the sample (not the size of the sample relative to the size of the population) determines accuracy.

Analogy: Suppose we take a drop of liquid from a bottle for chemical analysis. If the liquid is well mixed, it doesn't matter how large the bottle is. (It doesn't matter if we take a drop from a cup or a drop from a gallon.)

0.7.8 Sampling WITHOUT Replacement

Sampling without replacement from a LARGE population is just like sampling with replacement.

If our original population is much bigger than our sample, then it doesn't matter if we sample without replacement.

Technically, our probabilities change with each person, but the changes will be so minuscule that they won't really affect our answer.

So to make the problem doable, we pretend that we are sampling with replacement for the calculations.

0.7.9 Summary

The method of choosing the sample matters a lot.

Important features of all probability methods **(random samples)** for sampling:

- Interviewers have no discretion as to whom they interview.
- There is a definite procedure for selecting the sample.
- Selecting the sample involves the planned use of chance.

Even when probability methods are employed to select the sample, the estimate of the parameter will differ from the true parameter value due to bias and chance error.

If you use good sampling techniques, a bigger samples are more accurate.

0.8 Confidence Intervals

In general, the formula for a confidence interval for the population percentage is

$$\text{estimate} \pm z^* (SE)$$

For percentages we get

$$\text{sample } \% \pm (z^*) (SE_{\%})$$

For averages we get

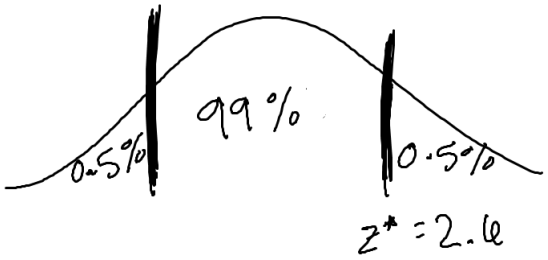
$$\text{sample average} \pm z^* (SE_{avg})$$

Margin of Error: The margin of error is the part that we add/subtract for our confidence intervals. So the $z^* (SE)$ is the margin of error.

0.8.1 Find z^*

Draw a picture and use the normal table.

For 95% we have memorized $z^* = 2$.



0.8.2 Interpret confidence interval

We are [90%, 95%, 99%] confident that the [average / percentage] of the [population / all] is between ___ & ___ .

0.8.3 Interpret confidence level

What does 95% confidence mean?

It means that there are millions of possible samples we could pick. Each sample would give us a different average (or percentage). So each sample gives a different confidence interval. Of all those, 95% would contain the population average (or percentage).

We hope that our confidence interval is one of the 95% that contain the population average, but there is no way to know.

0.8.4 Some Notes on Interpretations

- the confidence interval does NOT contain 95% of the individual data (in the population or sample)
- there is NOT a 95% chance that the *population* average is in the confidence interval
- there is NOT a 95% chance that the *sample* average is in the confidence interval

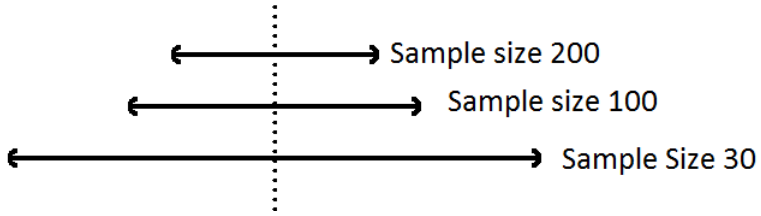
Example 11. Education level is measured for a SRS of 400 people and the average is found to be 11.6 years with an SD of 4.1 years.

The 95% confidence interval for the average education level of all people is 11.2 to 12.

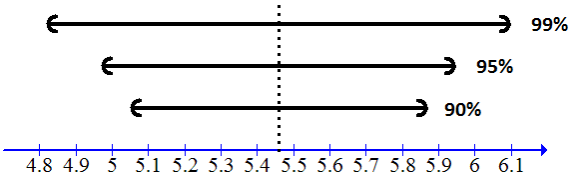
- **Correct Interpretations:**
  - We are 95% confident that the AVERAGE for ALL the people is between 11.2 and 12 years.
  - Of all the samples we could have chosen, 95% of the resulting confidence intervals would cover the true average for all the people.
  - (5% of the confidence intervals would NOT cover the true average for all the people.)
- **Incorrect Interpretations:**
  - There is a 95% chance that the average for all the people is between 11.2 and 12.
  - There is a 95% chance that the average for the people in this sample is between 11.2 and 12.
  - 95% of all the possible sample averages will be between 11.2 and 12.
  - 95% of all the people will have an education level between 11.2 and 12 years.

0.8.5 More Notes

- Increasing the sample size makes the confidence interval shorter.



- To increase the confidence level (to go from 95% to 99%) you have to make the confidence interval wider.



0.1 Ch 26 –Tests of Significance

Estimation (Confidence Intervals): We want to use sample data to estimate a population parameter

Tests of Significance: We want to know if a proposed population value is plausible or not.
We want to know if a difference is just due to chance.

Remark 1. The key idea is that if the observed value is too many SEs away from its expected value, that is hard to explain by chance.

0.1.1 Statistical Terminology

Null Hypothesis: A statement that *the difference between what we expect to observe and what we actually observe is due to chance.*

- The null hypothesis is the statement being tested.
- We either reject or fail to reject the null.

Alternative Hypothesis: A statement that the observed difference is "real" or not due to chance.

- The alternative is usually the statement of what we suspect may be true or *what we're trying to show.*

*If we reject the null hypothesis, we conclude that the alternative hypothesis is probably true.

Remark. We always write our conclusion as whether or not we have evidence for the *alternative hypothesis.*

Test Statistic: a number that we use to measure the difference between the data we observed and what we expected

- When we calculate the test statistic, we assume the null hypothesis is true.
- The further our actual observed data is from what we expected, the more evidence we have against the null hypothesis.
- There are many different test statistics for different situations. We will only learn a few in this class.
- For the Z test, we used a z-value for our test statistic. Z test statistics use the formula

$$z = \frac{\text{observed-expected}}{\text{SE}}$$

- The z-value tells us how many SEs our observed value is from our expected value.

P-value: The P-value is the chance of getting a sample value or test statistic at least as weird (extreme) as the one we got, if the null hypothesis were true.

- Extreme means far from what we would expect if the null hypothesis were true.
- The P-value is called the "observed significance level".
- We find the P-value as the *area under the curve.*

- The smaller the P-value, the less likely we are to believe the null hypothesis is plausible.
  - So the smaller the P-value, the more evidence we have against the null hypothesis.
  - And the smaller the P-value, the more evidence we have for the alternative hypothesis.
- The P-value is NOT the chance that the null hypothesis is true.
  - The P-value is the chance of us seeing DATA as far away as what we saw, if the null hypothesis were true.
  - So if the P-value is small, we tend to believe the null is false.

Remark. Smaller p-values are always STRONGER evidence for the alternative.

Conclusion: Reject the null hypothesis if the p-value is small.

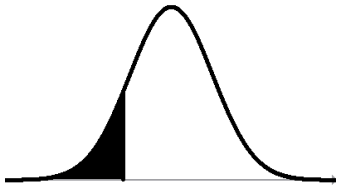
- How small?
  - Less than 5% is "statistically significant"
  - Less than 1% is "highly statistically significant"
- Then write your conclusion in the real-world context of the problem.
  - "We reject the null hypothesis and conclude that ...." or
  - "We fail to reject the null hypothesis and conclude that ...."

Remark 2. A hypothesis test tells us *how sure we are that there is a mathematical difference* (statistically significant). It doesn't tell us if that difference is actually big enough that anyone will care in the real world.
So **"Statistical Significance does not always mean Practical Significance".**

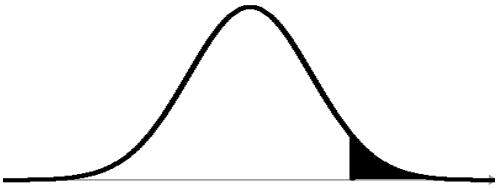
0.1.2 Two Tailed Tests

If we have a suspicion, before we do the experiment or take the sample, that the alternative hypothesis will be only in one direction, then we do a 1-tailed test.

Left Tail Example:
Null: the average of the population is 25.
Alternative: the average is LESS than 25.

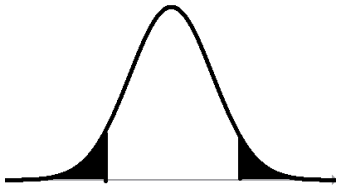


Right Tail Example:
Null: the average of the population is 25
Alternative: the average is MORE than 25.



If we don't know in which direction it will go, then we should do a 2-tailed test.

Two Tail Example:
Null: the average of the population is 25.
Alternative: the average is NOT 25.



Caution: It might be tempting to look at the data and then write the alternative hypothesis. DON'T! Pick the hypotheses before computing anything for the hypothesis test.

Remark 3. If you do a two tailed test all you can say is that the average is NOT 25. You can't say whether it is bigger or smaller (even if you can guess based on the data).

0.1.3 Rejecting or Failing to Reject the Null Hypothesis

- A hypothesis test rejects a null hypothesis only if there is strong statistical evidence against it.
  - This is similar to the legal system. We declare someone guilty, or reject their innocence, only if the evidence of guilt is beyond reasonable doubt.
- Due to convention, if our evidence isn't strong enough to convince us to reject the null hypothesis, we do **not** say, "We accept the null hypothesis".
  - Instead we say, **"We fail to reject the null hypothesis"** or "we do not have sufficient evidence to reject the null hypothesis".
  - This is like the legal system. If we don't have enough evidence to prove that a person is guilty beyond a reasonable doubt, we pronounce him **not guilty**. We do not say that we have proved his innocence.

Remark 4. If we fail to reject the null, we do NOT say we have evidence for the null.
Failing to find evidence against the null hypothesis only means that the data is consistent with the null hypothesis. It does *not* mean that we have clear evidence that the null hypothesis is true.

0.1.4 Never say "proved"

When we reject the null hypothesis, we have evidence that the alternative hypothesis is true. But we haven't proven it for sure. Because we are dealing with chance, you can't know which hypothesis is correct for sure. Even if you get a p-value like 0.000000002%.

That means you have a 0.000000002% of getting your actual sample data IF THE NULL WAS TRUE.

That might be a small chance, but it isn't impossible.

So even though the null seems very, very, very implausible, it MIGHT still be true. So we never say "proved".

0.1.5 Ch 26–One Sample Z Test for the Average

We use this test if:

- We are interested in the population average.
- We know the *population SD*.
- The data is normally distributed *or* the sample size is large (at least 30).

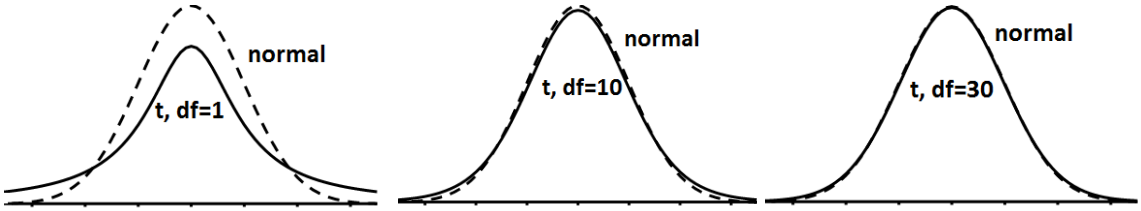
The test statistic is

$$z = \frac{\text{sample average} - EV_{avg}}{SE_{avg}}$$

Remark 5. EV_{avg} = population average (in null hypothesis) and $SE_{avg} = \frac{\text{population SD}}{\sqrt{\text{sample size}}}$.

The T Curve

The T curve is similar to the standard normal curve. But it has one extra property: *degrees of freedom.*



The normal curve is the dashed curve in each figure. As the degrees of freedom get bigger, the T curve gets closer and closer to the normal curve.

The T curve is symmetric. It is always centered at 0.

0.1.6 Ch 26–One Sample T Test for the Average

We use this test if:

- We are interested in the population average.
- We do NOT know the *population SD.* Instead, we have to estimate it with the *sample standard deviation.*
- The data is normally distributed *or* the sample size is at least 30.

The test statistic is

$$t = \frac{\text{sample average} - EV_{avg}}{SE_{avg}}$$

The degrees of freedom are

$$df = \text{sample size} - 1$$

We have to use the T table.

Remark 6. $EV_{avg} = $ population average (in null hypothesis) and $SE_{avg} = \dfrac{\text{population SD}}{\sqrt{\text{sample size}}} \approx \dfrac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$.

We don't know for sure what the population average is, but for our calculations, we assume that the null hypothesis is true.

We also don't know what the population SD is, but we estimate it with the sample standard deviation.

0.1.7 Two Sample Z Test for Averages

We use this test if:

- We want to compare two *population* averages
  - or we want to compare two randomized groups to see if a treatment is effective.
- We have independent samples.
- For **both** groups, the data is normally distributed, or the sample size is at least 30.

The test statistic is

$$z = \frac{\text{obs}_{\text{diff}} - \text{EV}_{\text{diff}}}{SE_{\text{diff}}}$$

where

$$SE_{\text{diff}} = \sqrt{(SE_A)^2 + (SE_B)^2}$$

Remark. If you think Population average A is **LESS** than Population average B, find the area to the **LEFT**.
If you think Population average A is **GREATER** than Population average B, find the area to the **RIGHT.**
This only works if you always write everything in the same order with A first.

Remark 7. We won't worry about a small sample size T test for two samples in this class.

Remark 8. But even if the *population* averages are the same, we would expect to see some difference in the *sample* averages just due to chance error (which sample we picked). The question is:

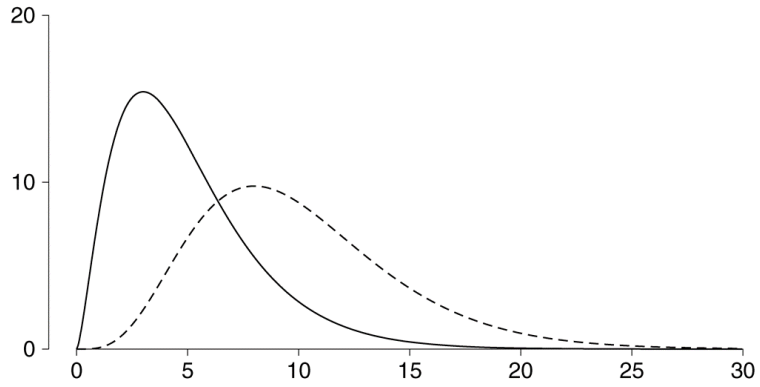
Is the difference in the *sample* averages so big that we believe it couldn't be just chance error and instead the difference must be due to a real difference in the *population* averages?"

Remark 9. Remember that unfortunately, if we do a two-tail test, all we can say is whether the population averages are the same or different. The test itself doesn't actually say which population average is higher.

Chi-Square Distribution

Here are some χ^2-curves. The curves have long right-hand tails. As the degrees of freedom go up, the curves flatten out and move off to the right.

The solid curve is for 5 degrees of freedom and the dashed is for 10 degrees of freedom.



0.1.8 The Chi-Square Test for Goodness of Fit

We use this test if:

- We want to know if a proposed distribution is a good fit for our data.
- The expected frequency for each category is at least 5.

Null Hypothesis: The proposed distribution is a good fit for our data. (The theoretical percentages seem to fit the data.)

Alternative Hypothesis: The proposed distribution is not a good fit for our data. (The theoretical percentages do NOT seem to fit the data.)

The test statistic is

$$\chi^2 = \text{sum of } \left[\frac{(\text{observed frequency-expected frequency})^2}{\text{expected frequency}}\right]$$

Degrees of freedom are

$$df = \text{number of categories} - 1$$

The P-value is the area in the right tail.

Remark 10. The observed frequency is the actual data of how many times it actually happened.

The expected frequency is found by

$$\text{observed frequency} = (\text{theoretical percentage}) \cdot (\text{sample size})$$

0.1.9 The Chi-Square Test for Independence

We use this test if:

- We want to know if two variables are independent or if they affect each other.
- Each of the expected frequencies is at least 5.

Null Hypothesis: The two variables are independent (do NOT affect each other).

Alternative Hypothesis: The two variables are not independent (do affect each other).

The test statistic is

$$\chi^2 = \text{sum of } \left[\frac{(\text{observed value-expected value})^2}{\text{expected value}}\right]$$

Degrees of freedom are

$$df = (\text{number of rows-1})\,(\text{number of columns} - 1)$$

To find the expected frequency for a table cell:

$$\text{expected frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

The P-value is the area in the right tail.

Remark 11. Unfortunately, all the χ^2-Test of Independence tells us is whether or not the two variables affect each other. They don't tell us which variable affects the other or if they both do. It doesn't tell you which category is more likely to do something compared to another category.

0.2 Chapter 29–A Closer Look at Tests of Significance

0.2.1 How small of a p-value do you need to reject the null?

The p-value tells us the strength of the evidence against the null hypothesis.

If the p-value is:

- less than 5% – significant
- less than 1% – highly significant

The 1% and 5% cut-off levels are guidelines only – a P-value of 4.9% is not very different from one of 5.1% and in a perfect world should not be treated differently but we just made a rule to simplify things.

0.2.2 Interpreting P-values

The P-value tells you how likely the result is, under the null hypothesis.

The P-value only tells you the strength of the evidence against the null hypothesis. It does NOT tell you about the:

- importance of the result
- strength of a relationship
- reliability of the design of an experiment

To decide how important a result is, or how strong a relationship is, think of it applying to the whole population and think about the real-world consequences. e.g. if increasing exposure to a toxin 100-fold increases cancer rates by only 1%, it is not as important as a treatment that can cut fatalities by 30%.

To decide how good an experiment is, ask questions about how the study was conducted.

0.2.3 Errors in Hypothesis Testing

Hypothesis testing deals with chances. So we can never be sure if our conclusions are correct.

- Sometimes we accidentally reject a true null hypothesis.
- Sometimes we fail to reject a null hypothesis that is false.

The point of testing is to help distinguish between real differences and chance variation.

But big differences due to chance variation can happen occasionally. So when we find a statistically significant result (small p-value), it might have just been a really big chance error.

If we have a large sample, we are less likely to make errors in our hypothesis testing.

0.2.4 Data Snooping

0.2.4.1 Doing Lots of Tests

If we do a LOT of statistical tests, we expect to find some "statistically significant" results due to chance error.

e.g. testing at the 5% level, we expect 5% of our tests to show up as "statistically significant" just due to chance, even if ALL null hypotheses are true!

- if we do 100 tests, we expect 5 "false positives"
- if we do 1000 tests, we expect 50 "false positives"

Always report how many tests you do, not just the ones that are "statistically significant". There are ways to adjust the p-values to take into account how many tests you have done (beyond the scope of this class).

0.2.4.2 1 Tail vs 2 Tail

Researchers like 1-tailed tests because they are more likely to get "statistically significant" results. (p-values are smaller if you only look at the area in one tail)

HOWEVER, the decision should be made BEFORE looking at the data and if there is any doubt about which direction to go, it should be a 2-tailed test.

In principle, it doesn't matter whether investigators make a one-sided or a two-sided test as long as they report what they did.

Data snooping (aka deciding what hypothesis to test AFTER the results have come in) is NO GOOD. Investigators need to decide what they want to test, then gather the data.

If they realize after looking at the data that they should have picked another alternative hypothesis (different tail), they need to run the analysis again on a *new, independent,* set of data.

0.2.5 Was the result important?

0.2.5.1 Statistical versus Practical Significance

Significance is not the same as importance.

Statistical significance: we are sure statistically speaking that there is a real difference and not just chance variation.

Practical significance: the difference would actually be important, practically speaking, in the real world (would it make a difference to you?)

0.2.5.2 Sample Size

If we have a LARGE SAMPLE, even tiny differences can show up as "statistically significant" – they might not be important.

If we have a SMALL SAMPLE, even an important difference might not show up as "statistically significant" (we say the test lacks "power").

0.2.6 Design of the Experiment

A test of significance doesn't check the design of the study.

A test of significance can tell us whether a difference is real, but not what caused the difference.

It is up to you as the investigator to use logic and design good experiments if you want to show what causes differences.

0.2.6.1 Population and Samples

Don't do tests of significance on the whole population!

- if we know the population data, many times there is no reason to do a hypothesis test.
  - e.g. we can just look at the population to see what the population average is, we don't need a Z or T test for average
- the size and types of errors would be different in a population than sample

Don't do tests of significance on samples of convenience.

- Your results won't be accurate.
- Instead, make sure you use probability samples.

0.2.7 Replication

If we conduct a test, and think we have found something important, we can replicate the study to convincingly show the result.

Also, since we know errors happen in hypothesis testing, replication lets us "double check" our results.

Important studies that cannot be designed properly (e.g. smoking studies) become convincing when they are replicated and show consistent effects, the effects respond appropriately to dose (e.g. higher doses show higher rates of disease), and are confirmed with lab experiments.

0.2.8 Conclusion

Tests of significance answer one and only one question:

How easy is it to explain the difference between the data and what is expected for the null hypothesis, on the basis of chance variation alone? (If the null hypothesis is true, how likely is it to get such a big difference between the observed data and the expected value, *just due to chance error*?)

The hypothesis test:

- does not check the design of an experiment
- does not tell you if a significant difference is important
- does not tell you what causes the difference

Hypothesis tests are useful, but they aren't appropriate for all situations.