# Inductive Logic: Basic Concepts of Statistics

1. What is **statistics?**
2. What is the relationship between a **population** and a **sample?**
3. What is the difference between the **median, mean,** and **mode?**
4. What is a **histogram?** A **normal distribution?**

**Statistics** is the branch of mathematics and inductive logic that deals with the relationship between populations and samples. While it can get pretty complex, it is based on a number of fairly simple concepts, which we’ll be talking about today:

* A **population** is a set of things. Some populations: people in the United States, squirrels currently living on campus, things under my bed, blood cells in my body, and so on.
* A **sample** is a subset of the population. We get samples by selecting some, but not all, members of the population.
* A **random** sample is a sample in which each member of the population has an equal chance of being picked for the sample. Random samples are important because they are thought to be **representative** samples—i.e., the properties of the sample resemble the properties of the population at large. Very few samples are *perfectly* random or representative.
* A **biased** sample is an unrepresentative sample. That is, the sample differs in some significant way from the population. For example, using a sample made up only of college students to draw conclusions about the whole U.S. population would be biased. When dealing with human experiments, potential sources of bias are numerous:
  + The sample should resemble the population in gender, race, age, income level, and other basic demographic categories.
  + For medical experiments, we need to make sure the sample matches the population in health-related issues: the presence or absence of disease, medications taken, amount of physical activity, diet, etc.
  + Many studies that involve human subjects have samples are, at least to some, **self-selected** (where people *choose* to participate). This almost always introduces *some* level of bias, but this need to be minimized as much as possible.

**Three Meanings of “Average.”** Suppose we have the following set of numbers: **1, 2, 2, 7, 10, 11, 25.** If someone asks “What is the average value?” we might give one of three answers.

* The **mean** is obtained by (1) adding up all the numbers and (2) dividing by how many numbers there are. So, (1+2+2+7+10+11+25)/7 = 58/7 = 8.3
* The **median** is the “middle” number when the numbers are arranged in order (as they are above). In the example above, this is 7.
* The **mode** is the most frequently appearing number. In the example above, this is 2.

## Measuring Dispersion

Along with figuring out the “average” of the data items, we are often concerned with the **dispersion**, or how close the data are to one another. The **range** is probably the simplest measure of dispersion, and is obtained by subtracting the *smallest data item* from the *largest data item*. For example, if our data set is {1, 2, 3}, the range is 3 – 1 = 2. The **variance ()** is calculated by determining the *mean squared distance from the mean of the data set*. In other words, it provides us with a measure of how different (on average) each member of the data is from the mean of the data set. While this may sound confusing, it can be obtained by following a step-by-step process (don’t be afraid to use a calculator!)

1. Calculate the mean by summing all of the data, and dividing by the number of items.
2. Make a *new* set of data by subtracting the mean from each of the original items. In some cases, this will be a negative number. Then, square each of these numbers.

|  |  |  |
| --- | --- | --- |
| **2a. Write Down Data** | **2b. Subtract Mean from Data** | **2c. Square the result from 2b** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. Add up the results from step 2c, and divide by the number of data.
   1. (rounded to two decimals).

The **standard deviation ()** is simply the *square root of the variance*. For example, the standard deviation for 1, 2, 3 is . (Remembering that is OK to use a calculator for this!) It’s important to remember that the dispersion of two data sets can be *very, very* different, even if they have the same means. For example, another data set with a mean of 2 is -200, -100, 306. The range for this is 506, the variance is 35,906, and the standard deviation is 189.5.

## Understanding Dispersion: Histograms and the Standard Distribution

Statistical arguments (of both the deductive and inductive variety) frequently involve claims about how the data is distributed. The success (or failure) of arguments often hinges on claims about these distributions.

|  |  |
| --- | --- |
| **Histograms.** A histogram plots the frequency of each data item. That is, it tells you *how frequently* each data item occurred. So, for example, in the following graph the horizontal For example this graph (from the U.K.’s National Health Service) shows that almost 4,500 women over the age of 75 were admitted to the hospital because of unintentional injury. By contrast, less than 500 women between the ages of 15 and 25 were admitted for this reason. The data displayed here can be used in a variety of inductive arguments. For example, we might use it to make a *prediction* about what sorts of patients will be most likely to suffer unintentional injuries in the future. | http://www.qihub.scot.nhs.uk/media/261618/histogram.png |
| **The Normal Distribution.** In many cases, the data we are interested in are distributed *normally,* with around 68% of the data lying with one standard deviation of the mean, and 95% within two standard deviations.Normal distributions with *small* variance will have data clustered very close to the mean, while those with *large* variance will have data that are much more spread out. Again, we can use these data in arguments: in normally distributed data sets with small variance, we can give a strong argument *predicting* that any individual datum will be close to the mean. This same argument would be much weaker for data with larger variance. |  |

## Problems With Percentages

Statistical arguments frequently involve the use of percentages, which can sometimes mislead us into accepting unwarranted conclusions. A few important points about working with percentages include:

**Consider the baseline.** Whenever a percentage change is reported (“Profits are up 25%!”), it’s important to ask: “Relative to what?”. If the answer is “relative to last year’s already impressive profits,” 25% is impressive. If it is “relative to last year’s squeaking by with $1 profit,” 25% amounts to a pretty large failure.

**Remembering baselines, part 2.** When dealing with *multiple* changes reported in percentages, its important to remember that each change is relative to the new baseline. So, for example, suppose that Jeanie takes a paycut of 50% one year, but then *increases* her pay by 50% the next year. It’s tempting to think “since 50% plus 50% is 100%, she’s back where she started!” But this is wrong: in fact, she is only making 75% of what she was originally making. Why? Suppose she made $100,000. After taking a 50% paycut, she was making $50,000 (the new baseline). A 50% raise to *this* baselineput her at $75,000.

**Wholes and parts: On not adding percentages.** Suppose that you are trying to save money. You decide to cut your spending on food by 10%, entertainment by 30%, and everything else by 20%.. Does this mean you’ve cut your budget by 60%? No! Since the budget is made up the components, each cut to a part ends up cutting a (smaller) part of the whole.

## Arguments using Statistics: How to Avoid Mistakes

**Review: Inductive and Deductive Arguments.** In order to assess the success of any argument (including those involving statistical data), it’s important to determine what the argument is trying to *do.* If the argument claims that some statistical data *guarantees* the truth of the conclusion, it is **deductive**. By contrast, if the argument merely claims the that truth of the statistical data makes the conclusion *likely* to be true, it is **inductive.** In many real-world cases, we’ll need to make use of both deductive and inductive reasoning to figure out whether a given statistical argument really “works.”

**Deductive Arguments Using Statistics.** An example of a deductive argument would be “Since the three childrens’ ages are 3, 4, and 5, their mean age must be 4.” Since this particular argument is **valid,** we know that *if* the premises are true, *then* we can be absolutely certain that the conclusion is true as well. This might be categorized as an **argument from mathematics** or an **argument from definition** (we are arguing from the definition of the *mean*). In deductive arguments, you can check the validity simply by taking the time to review the reasoning carefully and correctly (which can sometimes be tricky, especially when working with percentages and statistics). It’s also important to remember that the statistical data we start from might be mistaken, and that the argument may be valid but **unsound** (the children might have lied about their ages!).

**Inductive Arguments Using Statistics.** Once we have used the statistical data to make various deductive inferences, we often want to draw conclusions about things we *haven’t* already measured. So, for example, we might predict that “Since the mean age of the first three birthday party attendees is 4, the next child to arrive will probably be between 3 and 5.” This involves inductive reasoning, which can be either *weak* or *strong*: Inductive arguments use premises regarding statistical data to make **predictions, generalizations, causal inferences,** and **arguments to the best explanation.** Inductive arguments are also involved whenever we attempt to interpret charts or graphs (**arguments from signs**) or determine whether we can *really* trust the accuracy of the statistical data we are presented with (**arguments from authority**). Finally, unlike deductive arguments, determining the strength and cogency of an inductive argument always requires that we consider the possibility of **suppressed evidence,** and of relevant data that may have been “left out.”

## Inductive Generalizations

|  |  |  |  |
| --- | --- | --- | --- |
| Sample size | Confidence Level | | |
| **99.0%** | **95.0%** | **90.0%** |
| 10 | 40.5% | 31.0% | 25.9% |
| 50 | 18.1% | 13.9% | 11.6% |
| 100 | 12.8% | 9.8% | 8.2% |
| 250 | 8.1% | 6.2% | 5.2% |
| 500 | 5.7% | 4.4% | 3.7% |
| 1000 | 4.1% | 3.1% | 2.6% |
| 2000 | 2.9% | 2.2% | 1.8% |
| 3000 | 2.3% | 1.8% | 1.5% |

Many arguments that use statistical data take the form of **inductive generalizations**, which use premises concerning a sample to draw conclusions about a population. One common form is as follows:

* Premise 1: S is a sample of population P. “3,000 randomly chosen likely voters are a sample of all voters in the next election.”
* Premise 2: In sample S, the value of parameter A is X. “60% of those polled favored candidate A”
* Conclusion: So, the mean value of A in P is X (+/- **margin of error**%). This is with **confidence level** L. “With a confidence level of 95%, we can predict that between 58.2% and 61.8% of all voters favor candidate A.” (This is 60% +/- 1.8%, which is the margin of error for a confidence level of 95% and sample size of 3,000)

**What Does All this Mean?** The **margin of error,** or **confidence interval (CI)**, gives a range of possible “difference” between the sample (in our premises) and the population (in the conclusion). So, X +/-1.8% means that the true population value might differ by as much as 1.8% from X. The **confidence level** Lis a claim about how accurate the sampling/polling procedure is. 95% confidence claims that, using this sampling procedure, the true population mean will fall within the margin of error 95% of the time, at least under ideal conditions. However, it’s important to remember that conditions are never actually ideal—so, the fact that a study or poll reports 95% confidence never *literally* means “this study deductively proves that there is exactly a 95% chance the true result falls within the margin of error.”

**Why Should I Care?** A good inductive generalization should make it easy for the audience to see the degree to which the premises actually provide support for the conclusion. Because of this, you should be very wary of “polls” or “studies” that don’t report things such as the sample size, confidence level, or margin of error. In general, it is important to remember that we can draw conclusions with a higher confidence level by (1) increasing our sample size (**larger sample sizes *strengthen* inductive generalizations)** and (2) allowing for a *greater* margin of error **(highly specific claims with low margins of error *weaken* inductive generalizations)**.

## Evaluating Inductive Generalizations

When evaluating the strength or weakness of inductive generalizations (and other arguments using statistical premises), it’s important to keep in mind a few things

**How Representative was the Sample?** The confidence level and margin of error only tell you how good the argument *would be* if the sampling procedure were *perfect* and the population is *normally distributed.* This is never exactly true. Look closely, and see if there are any reasons for thinking that the sample chosen by the researchers might not have been representative of the population as a whole. In many cases, researchers try to ensure representativeness by choosing **randomly**, but this is often impossible. For example, many political polling agencies have (in the past) relied on contacting people who owned landline phones. This may no longer be representative of the population in general, however, since many younger people (who might vote differently) rely only on mobile phones. Remember to avoid **Hasty Generalizations** from small or biased samples.

**How Does this Fit With Other Evidence?**  Whenever you see the result of a “surprising” new poll or scientific study in the media, it’s important to remember that assessing the strength or weakness of an inductive argument requires that we consider ALL of the relevant evidence, and not just these particular results. For many of us, it’s all too easy to focus on those statistical results that seem to *support* our preexisting beliefs, and to (perhaps unconsciously) ignore those that challenge them. In presidential election years, for instance, there are almost always some people who argue that “The polls must be wrong! Except the ones that favor my candidate!” Something similar holds true of research regarding diet and nutrition (with various people focusing on the studies whose conclusions they like), and on many other issues. It’s important not to fall into this trap. While polls and studies can be wrong, we shouldn’t just believe the ones we *want* to believe. Instead we should take a “holistic” view at the evidence. (For nutrition-related issues, the U.S. government actually does this when it issues dietary guidelines). Relevant fallacies here include **Suppressed Evidence** and **Argument from Ignorance** (“Even though 99 studies say I’m wrong, this one says I’m right! So, who’s to say what the right answer is?”)

**Causation, Correlation, and Alternative Explanations.** We often use statistical data to make inductive arguments about *causes.* For example, statistical data about the strong correlation between *smoking cigarettes* and *contracting lung cancer* provides evidence for thinking that smoking *causes* lung cancer. However, it’s important to remember that correlation between X and Y doesn’t mean that X causes Y. For example, *spending time in the hospital* might be correlated with *chance of death.* But this doesn’t mean that spending time in the hospital *caused* death!This is closely related to the **False Cause** fallacy discussed in previous classes.

## Don’t Forget about Distributions!

Statistical arguments don’t always involve drawing conclusions about populations from samples. In other cases, we use statistical data about *populations* to draw conclusion about a particular *individuals* or *groups.* When doing this, determining whether a given (inductive) argument is strong or weak will often involve using information concerning about the distribution of the population, and the dispersion of this distribution. For example, consider the following two arguments, both of which assume a normal distribution. In this sort of distribution, around 68% of the population is within one standard deviation of the mean, and 95% is within two standard deviations of the mean:

* The mean age of a high school senior is 18 years old, with a standard deviation of .5 years. So, if I called up a random high school senior, it is probably true that she or he would be between 17 and 19 years old.
* The mean wage of an employee at Firm X is $18/hr, with a standard deviation of $8. So, if I called up a random employee of Firm X, it is probably true that she or he would earn between $17 and $19.

Even though the mean is the same in both cases (18), these arguments are very different. In particular, the first argument is inductively *strong,* while the second one is inductively *weak.* This is because the ages of high school seniors are, on average, much *closer* to the mean than are the average wages of the employees of firm X.

**Not All Data Are Distributed Normally!** Perhaps because normal distributions are so common in the natural world, we often tend to think and act as if *all* data are normally distributed, even when they are not. In particular, we tend to think that knowing the “mean” gives us information about the “typical” (median) or “most common” (mode) case. But this isn’t always true. For example, *wealth* is not normally distributed. Instead, most of the wealth is clustered near the “top” of the distribution (the top 10%, 1%, etc.). So, for example, suppose we learn that the mean wealth of an American citizen is $250,000. This does NOT mean that there are roughly equal numbers of people who are worth more or less than this, as would be the case if wealth were distributed normally. Instead, we should expect that the majority of people are worth less than this, with a few being worth much more. Other examples of non-normal distribution include *annual health care spending* and *alcoholic beverages consumed per week.* So, when you hear that the “average” American spends $9,000 on health care per year, or drinks 11 drinks/week, just keep in mind that *most* people are below these numbers, and that the “top” end of the distribution is much higher (they may spend $100,000/year, or drink 60 drinks/week). This makes a real difference when we consider inductive arguments about how to solve problems related to these issues.

## Review Questions

1. Calculate the *mean, median, mode, variance,* and *standard deviation* of the following data set: {1, 1, 3, 4, 6}. It’s OK to use a calculator (or the calculator app on your phone.
2. What is the normal distribution? What difference does the variance (and SD) make to the “shape” of this distribution?
3. On the website for the local tennis club, you can “vote” for whether or not you approve of the new mayor’s performance, given her recent decision to demolish six tennis courts in a local park, and to replace them with playground equipment. You decide to vote, and it shows you the results so far. It says that 100 people have voted so far, and that 70% disapprove. You like the mayor, so you decide to click “back” and vote again. This works fine.
   1. Name at least THREE things that are wrong with using the results of this poll as evidence for the conclusion “the mayor is very likely to lose the election coming up next month.”
   2. Suppose that you were in charge of designing a poll that would more accurately predict the results of the upcoming mayoral election. How might you go about doing this? How would your poll differ from this one?
4. Identify the following argument as deductive-valid, deductive-invalid, inductive-strong, or inductive-weak. Briefly explain your reasoning.
   1. The final grade in Professor’s X class is determined entirely by the mean percentage on three exams. Bonnie received scores of 100%, 90%, and 50%. So, her final grade is 80%.
   2. 60% of the students in Professor X’s class on strength training class are capable of bench-pressing at least 200 lbs. So, it is likely that between 58% and 62% of all students at the college can bench at least 200 lbs.
   3. In a survey of 1,000 randomly selected students at the local college, 55% were women. From this we, can conclude that between 50% and 60% of the college’s students are women.
   4. I got a 60% on exam 1. My score on exam 2 was 50% higher than this, and my score on exam 3 was 50% lower than my score on exam 2. I can conclude that my score on exam 3 was 60%, just like exam 1.
   5. In his career, Michael has taken almost 10,000 free throws, and has made 70% of them. Given this, we can predict he will probably make *exactly* seven of the next ten free throws he shoots.
   6. Publicly available statistics show that Harvard admits only 5% of applicants. We can conclude from this that, if Bill Gates (the founder of Microsoft) were to decide to apply to Harvard to finish his degree, he would have only a 1 in 20 chance of being admitted.