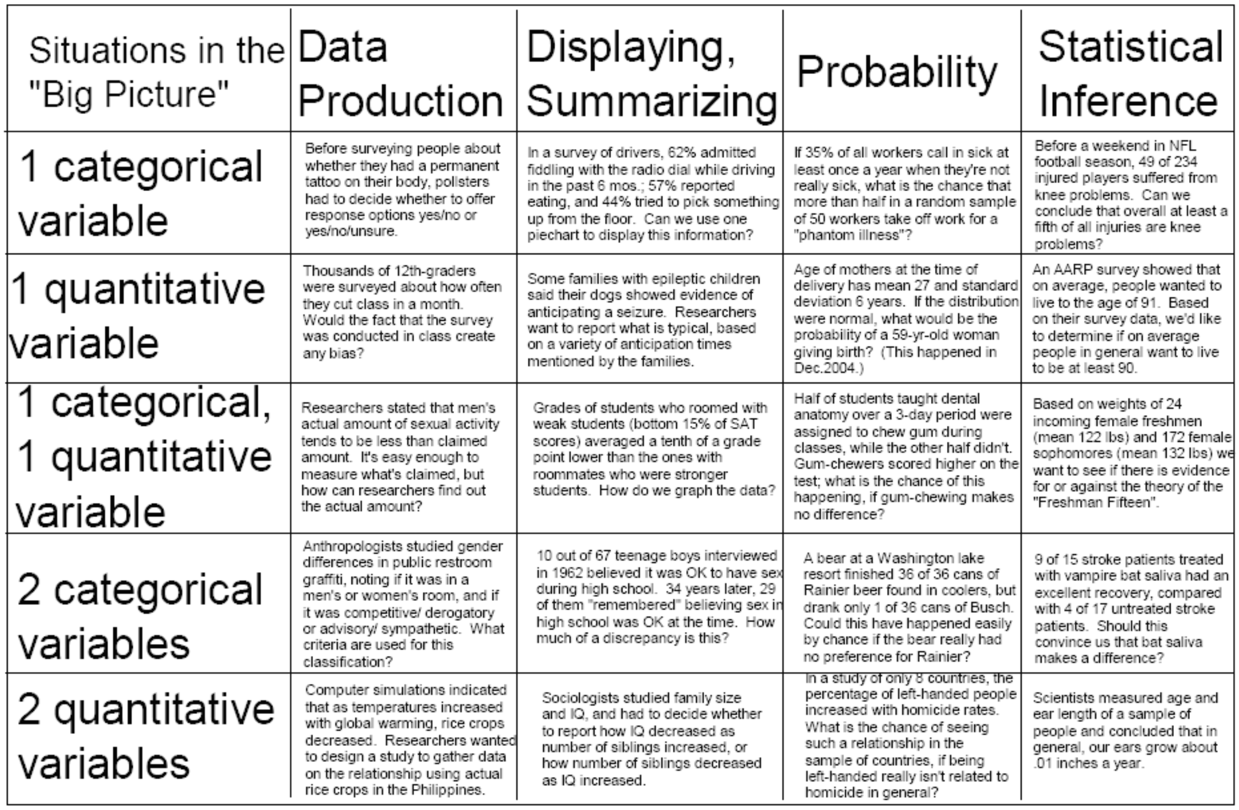
Looking at the Big Picture of Statistics

Nancy Pfenning University of Pittsburgh

Students may manage to solve statistics problems *in context*, but otherwise they often have trouble identifying the appropriate tools to even get started. To help students keep material in perspective throughout a statistics course, they can, from the very beginning, be taught to identify (A) which specific stage in the general four-stage process applies; and (B) which of five variable-type situations applies. Then, as the course progresses, they can be reminded of how the new material fits into the "big picture" of statistics.

This table presents examples within each of the four basic processes, classified according to five possible variable situations.



Situations in the "Big Picture"	Data Production	Displaying, Summarizing	Probability	Statistical Inference
1 categorical variable	Before surveying people about whether they had a permanent tattoo on their body, pollsters had to decide whether to offer response options yes/no or yes/no/unsure.	In a survey of drivers, 62% admitted fiddling with the radio dial while driving in the past 6 mos.; 57% reported eating, and 44% tried to pick something up from the floor. Can we use one piechart to display this information?	If 35% of all workers call in sick at least once a year when they're not really sick, what is the chance that more than half in a random sample of 50 workers take off work for a "phantom illness"?	Before a weekend in NFL football season, 49 of 234 injured players suffered from knee problems. Can we conclude that overall at least a fifth of all injuries are knee problems?
1 quantitative variable	Thousands of 12th-graders were surveyed about how often they cut class in a month. Would the fact that the survey was conducted in class create any bias?	Some families with epileptic children said their dogs showed evidence of anticipating a seizure. Researchers want to report what is typical, based on a variety of anticipation times mentioned by the families.	Age of mothers at the time of delivery has mean 27 and standard deviation 6 years. If the distribution were normal, what would be the probability of a 59-yr-old woman giving birth? (This happened in Dec.2004.)	An AARP survey showed that on average, people wanted to live to the age of 91. Based on their survey data, we'd like to determine if on average people in general want to live to be at least 90.
1 categorical, 1 quantitative variable	Researchers stated that men's actual amount of sexual activity tends to be less than claimed amount. It's easy enough to measure what's claimed, but how can researchers find out the actual amount?	Grades of students who roomed with weak students (bottom 15% of SAT scores) averaged a tenth of a grade point lower than the ones with roommates who were stronger students. How do we graph the data?	Half of students taught dental anatomy over a 3-day period were assigned to chew gum during classes, while the other half didn't. Gum-chewers scored higher on the test; what is the chance of this happening, if gum-chewing makes no difference?	Based on weights of 24 incoming female freshmen (mean 122 lbs) and 172 female sophomores (mean 132 lbs) we want to see if there is evidence for or against the theory of the "Freshman Fifteen".
2 categorical variables	Anthropologists studied gender differences in public restroom graffiti, noting if it was in a men's or women's room, and if it was competitive/ derogatory or advisory/ sympathetic. What criteria are used for this classification?	10 out of 67 teenage boys interviewed in 1962 believed it was OK to have sex during high school. 34 years later, 29 of them "remembered" believing sex in high school was OK at the time. How much of a discrepancy is this?	A bear at a Washington lake resort finished 36 of 36 cans of Rainier beer found in coolers, but drank only 1 of 36 cans of Busch. Could this have happened easily by chance if the bear really had no preference for Rainier?	9 of 15 stroke patients treated with vampire bat saliva had an excellent recovery, compared with 4 of 17 untreated stroke patients. Should this convince us that bat saliva makes a difference?
2 quantitative variables	Computer simulations indicated that as temperatures increased with global warming, rice crops decreased. Researchers wanted to design a study to gather data on the relationship using actual rice crops in the Philippines.	Sociologists studied family size and IQ, and had to decide whether to report how IQ decreased as number of siblings increased, or how number of siblings decreased as IQ increased.	In a study of only 8 countries, the percentage of left-handed people increased with homicide rates. What is the chance of seeing such a relationship in the sample of countries, if being left-handed really isn't related to homicide in general?	Scientists measured age and ear length of a sample of people and concluded that in general, our ears grow about .01 inches a year.

(A): Statistics as a Four-Stage Process

In general, statistics is used to take information from a sample and use it to draw conclusions about the larger population from which the sample came. To accomplish this goal, we

1. Learn about **Data Production**: how to take a sample that truly represents the larger population of interest, and how to gather information so that it accurately reflects what is true about the variables or relationships in that sample.

2. Learn how to **Display and Summarize** single quantitative or categorical variables of interest, or relationships between variables if there are two variables involved.

3. Learn about the science of **Probability**: assume we actually know what is true for the entire population; what is likely to be true for a sample drawn at random from that population?

4. Learn how to perform **Statistical Inference**: Use what we have discovered about variables of interest in a random sample, and draw conclusions about those variables for the larger population.

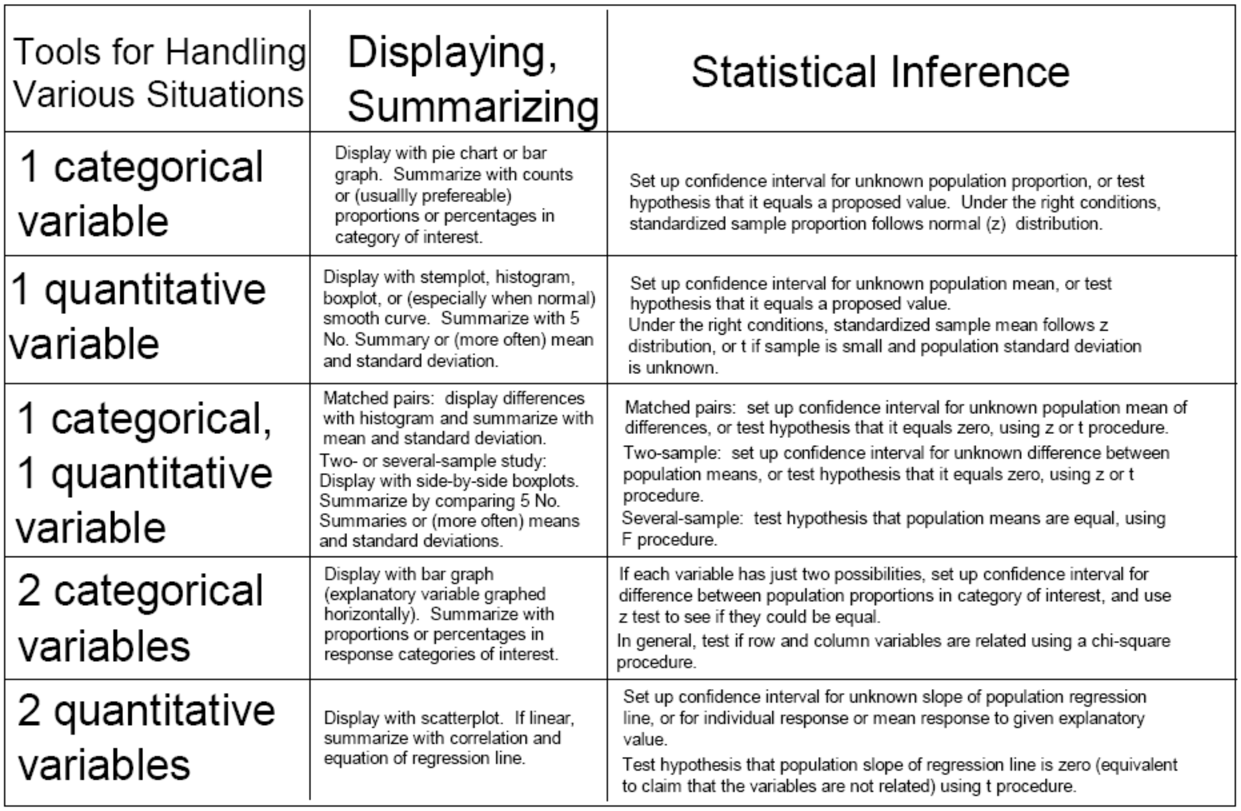
(B): Five Possible Variable Situations

Another crucial part of the statistical picture is what kind of variable or variables are involved. In introductory statistics, we are almost always concerned with one of just five possible situations: either a single categorical or quantitative variable, two categorical variables, two quantitative variables, or one of each.

"Where Am I?" "You Are Here..."

This game has players pick a scenario from a sample like those in the table above, and then identify what type of variables are involved, and what basic statistical process is being addressed. (At the lower level of play, the scenario can be chosen from a list. At the more challenging level, it is picked at random from a box.)

The following table reminds users of statistics about the various display, summary, and inference tools to be applied, depending on which situation is at hand. These are encountered gradually throughout an introductory statistics course, until a student has accumulated enough tools to handle any of the situations in the "Where Am I?" table.



Tools for Handling Various Situations	Displaying, Summarizing	Statistical Inference
1 categorical variable	Display with pie chart or bar graph. Summarize with counts or (usuallly prefereable) proportions or percentages in category of interest.	Set up confidence interval for unknown population proportion, or test hypothesis that it equals a proposed value. Under the right conditions, standardized sample proportion follows normal (z) distribution.
1 quantitative variable	Display with stemplot, histogram, boxplot, or (especially when normal) smooth curve. Summarize with 5 No. Summary or (more often) mean and standard deviation.	Set up confidence interval for unknown population mean, or test hypothesis that it equals a proposed value. Under the right conditions, standardized sample mean follows z distribution, or t if sample is small and population standard deviation is unknown.
1 categorical, 1 quantitative variable	Matched pairs: display differences with histogram and summarize with mean and standard deviation. Two- or several-sample study: Display with side-by-side boxplots. Summarize by comparing 5 No. Summaries or (more often) means and standard deviations.	Matched pairs: set up confidence interval for unknown population mean of differences, or test hypothesis that it equals zero, using z or t procedure. Two-sample: set up confidence interval for unknown difference between population means, or test hypothesis that it equals zero, using z or t procedure. Several-sample: test hypothesis that population means are equal, using F procedure.
2 categorical variables	Display with bar graph (explanatory variable graphed horizontally). Summarize with proportions or percentages in response categories of interest.	If each variable has just two possibilities, set up confidence interval for difference between population proportions in category of interest, and use z test to see if they could be equal. In general, test if row and column variables are related using a chi-square procedure.
2 quantitative variables	Display with scatterplot. If linear, summarize with correlation and equation of regression line.	Set up confidence interval for unknown slope of population regression line, or for individual response or mean response to given explanatory value. Test hypothesis that population slope of regression line is zero (equivalent to claim that the variables are not related) using t procedure.