AP Statistics / Mr. Hansen
8/6/2015 [slightly rev. 9/2/2015]

Name: _____

Reserved Words: Be Careful!

> *It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so.*
>
> — Artemus Ward

Definition You May Have Heard	**Statistics Definition**
average = mean, mode, or median, depending on context	**average** = mean
bar graph = any graph that uses bars	**bar graph** = a graph that depicts quantities associated with different values of a categorical variable, with bars having lengths proportional to those values *Note:* If a graph shows the distribution of a quantitative variable, we call it a *histogram*, not a bar graph.
bias = ~~error~~ *Note:* This is a bad definition, and it is crossed out because it is simply wrong. Bias does not refer to an error; it refers to a *systematic* error. In fact, it's possible to have a large error with little or no bias. People sometimes mistakenly say "bias" when what they mean is, "The s.d. (variability) of your data set is larger than I would like." Maybe you have heard a student say, "So-and-so is a biased teacher, because his grades are unpredictable, and his averages are all over the place and usually wrong." The student is incorrect to say that the teacher is biased. The teacher's grades probably have a high s.d., and the grades may even be wrong, but that doesn't mean that the teacher is biased.	**bias** = a systematic tendency to compute statistics that are either larger (on average) than the true parameter value or smaller (on average) than the true parameter value *Note:* Bias can be either positive (on the high side of the parameter) or negative (on the low side of the parameter). Bias refers to the overall net tendency of the statistical error and must therefore be positive or negative, not both. Positive bias means that the error is positive on average, but you might still obtain a negative error by chance. Similarly, negative bias means that the error is negative on average, but you might still obtain a positive error by chance. **If the mean error is 0, then there is no bias, regardless of how large some of the errors are.**

Definition You May Have Heard	**Statistics Definition**
cause and effect = something about which the world of mathematics has nothing to say	**cause and effect** = something that can be proved, in the statistical sense, but only by means of a controlled experiment
confidence = bluster, swagger, or a feeling of knowledge or power	**confidence** = the probability (usually expressed as a percentage) with which intervals generated by a certain process* would correctly include the parameter value, in the long run *Note:* It is incorrect to apply the word "probability" to the likelihood that a single interval correctly includes the parameter value. The reason is that with a single interval, there is no long run. Your confidence applies to the *process*, not to the specific interval you generate. This is a very difficult concept for students to master.
correlation = any pattern at all	**correlation** = Pearson's *r* value, i.e., the linear correlation coefficient *Note:* On those occasions when we need a less precise word to refer to patterns in general, we will say "association" instead of "correlation."
distribution = allocation, allotment	**distribution** = a set of possible quantitative outcomes, coupled with the counts (or relative frequencies) associated with each value *Note:* A distribution is almost always depicted by a histogram, a relative frequency histogram, a density plot, a boxplot, a stemplot, or a dotplot. These visual aids are so common that we often blur the distinction between a distribution (which is an abstract concept) and its visual representation. For example, we will often say, "Look at this distribution," as we point to a histogram or other visual aid, even though it would be more correct to say, "Look at this depiction of the distribution." This normally causes no confusion. (Similarly, nobody is ever confused in math class by a teacher who says, "Look at this function," even though it would be more correct to say, "Look at this *graph* of our function.")

Definition You May Have Heard	Statistics Definition
error = mistake	**error** = the difference between an observed value and the predicted or expected value *Synonym:* residual. *Note:* In statistics, error is almost never a mistake! It's simply a recognition that in the real world, variability is everywhere.
expected value = mode, maybe?	**expected value** = mean
experiment = a study undertaken for some purpose	**experiment** = a study in which the subjects are not merely observed ("observational study") but actually subjected to a *treatment* *Note:* A control group is not always used, nor is a control group even possible in some cases. (If there is a treatment but no control group, we call such a study an "uncontrolled experiment.") However, if we seek to prove cause and effect, we must have a treatment group and a control group.
explanatory = serving to make clear	**explanatory** = x *Note:* That's all it means, just x. By the way, that's regardless of whether or not the x variable has any causal link with y, and regardless of whether or not the values of x "explain" the values of y.
hypothesis = one of the following, depending on context: 1. In science: a conjecture to explain an observation. 2. In mathematics: the "if" part of a theorem. 3. In common speech: a conjecture, or an imaginary situation ("hypothetical").	**hypothesis** = one of two types of imaginary descriptions of parameters: 1. The *null hypothesis* asserts that the parameter values are such that essentially nothing interesting is happening; i.e., no effect exists. 2. The *alternative hypothesis* is what we seek to prove, and it asserts that the parameter values are different in some way from what the null hypothesis claims.
normal = customary	**normal** = Gaussian, i.e., following a specific bell-shaped distribution curve

Definition You May Have Heard	Statistics Definition
odds = likelihood (imprecise usage)	**odds** = ratio of favorable to unfavorable outcomes ("odds in favor"), or the ratio of unfavorable to favorable outcomes ("odds against") *Note:* "Odds" and "probability" are not synonyms.
outlier = someone or something extraordinary	**outlier** = a numeric data point that is more than 1.5 IQR above the third quartile or more than 1.5 IQR below the first quartile *Note:* In a regression setting, an outlier is simply any data point that has a large |residual|, and we use our judgment to decide what constitutes a "large" absolute value.
parameter = a boundary, or an adjustable constant that defines the boundary conditions of a more general problem	**parameter** = a number that describes a population *Note:* Population parameters are almost always unknown. The one-sentence summary of our course is, **"We use statistics to estimate parameters."**
probability = likelihood of occurrence	**probability** = long-run relative frequency of occurrence
prove = to establish as true *Note:* In math class this means using axioms, definitions, theorems, and accepted means of rigorous proof in order to establish the absolute truth of a proposition. At the end of a rigorous mathematical proof, it is customary to write "Q.E.D." or the Halmos sign, □. Although almost nobody outside high school ever uses a two-column proof, all rigorous mathematical proofs should be capable of being presented in a two-column format (statements in the left column, ironclad reasons in the right column).	**prove** = to establish that an observed difference is too great to be plausibly attributed to chance alone, i.e., that the null hypothesis is likely to be false *Note:* It is impossible to achieve 100% confidence in a statistical proof, and a two-column format is never used. The standards are not at all similar to those used in mathematics. In statistics, we consider a proposition (i.e., alternative hypothesis) proved if *fresh data uncontaminated by "data dredging"* reveal that the null hypothesis should be rejected at a certain significance level called "alpha" (often 0.05), and this occurs whenever the *P*-value associated with a test or experiment is less than that alpha value. Also, note that only *differences* from the null hypothesis can be proved in the statistical sense—there is no way to prove the null hypothesis itself. **

Definition You May Have Heard	**Statistics Definition**
random = arbitrary or haphazard	**random** = exhibiting long-run behavior that follows a certain distribution, even though the short-run behavior is unpredictable
range = set of all possible *y* values for a relation	**range** = max. minus min. (just a number, not a set at all!)
regression = reverse progression, decline	**regression** = the process of finding a function (usually linear) to predict *y* values when the *x* values are known
replication = duplication, or (in science) the repeating of an experiment for the purpose of confirming a previous result	**replication** = using a sufficiently large sample size (*n*) so that the results, if any, are less likely to be dismissed as a fluke
residual = something left over	**residual** = synonym for *error* (see "error" above)
response = action resulting from a stimulus	**response** = *y* *Note:* That's all it means, just *y*. By the way, that's regardless of whether or not the *x* variable has any causal link with *y*, and regardless of whether or not the values of *x* "explain" the values of *y*, and regardless of whether or not the values of *y* "respond" in any way to the values of *x*.

Definition You May Have Heard	**Statistics Definition**
significance = magnitude, size	**significance** = a situation in which a difference is too large to be plausibly explained by chance alone *Note:* The word "plausibly" is essential. After all, any difference can be explained by chance alone, at least with some extremely small probability. The question is not, "Can this difference be explained by chance alone, under any possible imaginary combination of unlikely circumstances?" but rather, "Can this difference be *plausibly* explained by chance alone?" The difference we are talking about, by the way, is typically the observed difference between a statistic and its expected value, or between a statistic and some claimed benchmark value. Example: My friend Bill claims that Obama's job performance rating is 44%. I poll a random sample of Americans and find that 46% of the people in my poll approve of Obama's job performance. Is that difference of 2 percentage points statistically significant? *It depends on how large my sample size was. For a sample of 100, no: The difference could plausibly be explained by chance alone. For a sample of 10,000, yes: The difference is too large to be plausibly explained by chance alone, and we would conclude that Bill's claim is faulty.*
skew (adj.) = tilted to one side	**skew** (adj.) = tilted to one side (said of a distribution), but note: "skew right" means that the distribution dribbles out to the right, and "skew left" means that the distribution dribbles out to the left In other words, "right" and "left" do not refer to the side that has the big lump; they refer to the side that has the long tail. Business and economic distributions are almost always skew right.
skew (v.) = to tamper with; to doctor	**skew** (v.) = to make asymmetric

Definition You May Have Heard	Statistics Definition
slope = $\frac{\Delta y}{\Delta x}$	**slope** = the predicted average change in *y* for each additional unit increase of the *x* variable Memorize this! It's on every AP Statistics exam, and it's on other standardized tests as well, beginning in 2015 (PSAT) and 2016 (SAT). *Note:* For full credit, you must replace the letters *x* and *y* with the "real-world context" variables, in words. For example, if the *x* (explanatory) variable is hours of TV viewing per night, and if the *y* (response) variable is GPA, you could explain a slope of –0.15 as follows: "The model predicts that for each additional hour of TV viewing per night, the average reduction in GPA is 0.15 points."
statistic = a fact (especially a numeric fact); sometimes figuratively used to refer to a person, e.g., an accident victim	**statistic** = a number computed from data *Note:* Isn't it amazing how people often use the word "statistic" without knowing what it means?
time series = a series of times, maybe? *Note:* This is just plain wrong.	**time series** = a collection of quantitative data (*y* values) associated with dates (*x* values), usually displayed graphically *Note:* If you go into investment banking, real estate, asset management, or finance, you'll spend a large chunk of your life looking at time series.
variable = a quantity that can change, denoted by a lowercase or uppercase letter	**variable** = a named column of data, or in the case of a random variable, a numeric outcome from a distribution described by a random process *Note:* We use uppercase letters for random variables.

Footnotes

* We learn how to do this process in the second semester.

** What if the *P*-value associated with a test or experiment is greater than or equal to alpha? In that case, all we can say is that there is insufficient evidence to reject the null hypothesis. In other words, there is no statistical significance, and we haven't proved anything. This situation often occurs in scientific research. For example, do a Google search on "pwoton decay expewiment" (with the Barry Kripke–style spelling) and see what you get! (It helps if you are a fan of *The Big Bang Theory*.)