Student Questions from January 7 and 8, 2015

Note: If the question doesn't interest you, skip reading it for now.
You can always come back later.
Everything here is potentially educational, but you're
much more likely to retain the answers that interest you.

(Think of these as a bunch of mini-TED talks on ted.com. Are you really going to surf over to the TED talks that don't interest you? Of course not. Just digest the interesting ones, and LEARN LEARN LEARN.)

Q. How can one determine significance just from *n* alone?

A. You can't. You need two numeric ingredients, namely population s.d. and the sample size(s).

In polls, where the parameter of interest is often a proportion, we can find an upper bound for the s.e. by using the "worst-case" result of $p = q = 0.5$. Therefore, it might appear that you are calculating the significance from *n* alone, but you really aren't. We talked a little bit about this in one of the sections, but since it is a topic for semester 2, we'll have to wait a little bit to discuss all the details.

For determining significance (in semester 2), there are also some non-numeric ingredients that we have to worry about. For example, we need to know whether the test is one-tailed or two-tailed, we need to know whether we are looking at means or proportions, and we need to know how many samples we have. All in due time . . .

Q. Error = difference?

A. Yes. The order is always "actual minus predicted."

Q. What is the s.e. of the LSRL?

A. You probably meant to ask what the s.e. of the LSRL *slope* is. The s.e. of the LSRL slope is given by the expression $\frac{b_1}{t}$ and equals the s.d. of all possible LSRL slopes that we might have observed under certain model assumptions that we will learn about in March or April.

Q. How is m.o.e. different from s.e.? How are each used?

A. The m.o.e. is always some multiple of the s.e., often roughly twice as big. That is why we say that a rough estimate of m.o.e. is $2 \cdot$ s.e. More accurate values for that multiple are found in the *t* table on the inner back cover of your textbook. We have not learned how to use the *t* table yet; that's a topic for semester 2.

Q. Why, in some cases, are we using s.e. in the place of s.d. if they are not the same thing? Like, with $\sigma_{\hat{p}}$.

A. Remember this: The s.e. is the s.d. of a statistic in a sampling distribution. In other words, the s.e. is a special kind of s.d. As for $\sigma_{\hat{p}}$, it's the s.e. of the sample proportion, i.e., the s.d. of

the sample proportion in the sampling distribution of all possible sample proportions. Since $\hat{p}$ equals the observed number of successes (X) divided by n, and since X (in a population that is large relative to n) is essentially binomial, we can adapt the binomial random variable formula $\sigma_X = \sqrt{npq}$ to get a formula for $\sigma_{\hat{p}}$. The derivation was given in class on Tuesday, 1/6/2015, and is as follows: $\sigma_{\hat{p}} = \sigma_{\frac{X}{n}} = \frac{1}{n}\sigma_X = \frac{1}{n}\sqrt{npq} = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$.

The third step in the derivation is true because the standard deviation of any constant multiple of a random variable is simply the constant times the standard deviation of the random variable. Is $\frac{1}{n}$ a constant? Yes.

If you like math, you can also do the derivation using variances as follows:

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \text{Var}\left(\frac{1}{n}X\right) = \left(\frac{1}{n}\right)^2 \text{Var}(X) = \left(\frac{1}{n}\right)^2 (npq) = \frac{npq}{n^2} = \frac{pq}{n}.$$

Here, we utilized the fact that Var(kX) = k^2Var(X) for any constant multiple k.

Q. I struggle with CLT and its uses.

A. We all do. Even Wall Street "geniuses" have struggled with this. Basically, the CLT is applicable when σ is finite and n is "large enough." The problem is that for highly skewed distributions, a "large enough" n may be completely impractical. Also note: A common mistake is to forget that in the real world, σ is a parameter and is, in general, unknown. An underestimation of σ (based on poor metaknowledge) has been a continuing theme throughout human history.

Q. How do we compute s.e.? How is s.e. different from s.d.? [Asked by multiple students.]

A. The s.e. is a special type of s.d., namely the s.d. of a statistic in a sampling distribution. That is why your AP formula sheet uses the labels "Standard Deviation of Statistic" in the second column of the third page; all those formulas are actually s.e. formulas. In each case, the ingredients of the s.e. formula include the population s.d. (in some form or other) and the sample size(s). Population s.d. always plays a role in the numerator, and the square root of sample size always plays a role in the denominator.

Q. I'm still a little fuzzy on what those possible differences could be.

A. Difference = effect size = E.S. = error = (observed – expected) = residual. These are all essentially synonyms. With statistical tests (semester 2), we usually compute E.S. as the difference of means or the difference of proportions. Exceptions are (1) the χ^2 tests, in which the χ^2 statistic is computed as a function involving observed *counts* minus expected *counts*, all of which are squared, and (2) the LSRL t-test, in which the t statistic is computed as the difference between the observed LSRL slope, i.e., b_1, and the expected LSRL slope if there

were no linear correlation between the variables, i.e., 0. However, other than those two exceptions, all our differences will be either differences of means or differences of proportions. In more advanced statistics classes, you can compute all sorts of other differences, each of which would have its own sampling distribution model.

Q. Binompdf vs. cdf and when to use.

A. Binompdf is for a single *k* value. Binomcdf is for left-tail probability *through* some value *k.* Before using either one, though, make sure that a binomial model is appropriate.

Q. Why is the s.e. of the LSRL slope equal to the LSRL slope divided by *t*?

A. This is a semester 2 issue. Learning why is required for MPQ question #80. (To find the MPQ, do a Google search on the words MUST PASS QUIZ in any order, and click "I'm Feeling Lucky.") Your question is answered in the MPQ answer key for #80.

Q. I still don't understand how to determine statistical significance. [Asked by multiple students.]

A. Good! You shouldn't understand the process, since we haven't learned how to do it yet. Statistical significance is quantified by something called the *P*-value of a test. *P*-value is a bit like a score in golf: Lower is better. Generally, $P < 0.05$ is considered significant, but in certain contexts, a much lower *P*-value would be needed. If you were publishing a scientific paper that purported to disprove a well-accepted theory, such as special relativity, you would be laughed at unless your *P*-value were extremely low, maybe 0.00001 or so.

Q. Is it always a must to use randomization?

A. No, since randomization is not always ethical. There are plenty of experiments that use what are called "case studies" instead. In other words, cases that appear to be closely matched in all important variables are monitored, and results are tabulated based on whether the experimental treatment was present or not. The problem, however, is that there could be subtle forms of selection bias (i.e., bias in who receives which treatment). The "gold standard" for determining the safety and efficacy of drugs remains the randomized double-blind trial.

A classic example of selection bias was in the case of Premarin, a drug that has been marketed to women since 1942 for the relief of hot flashes and other menopause-related symptoms. Doctors noticed over a period of decades that their patients on Premarin were healthier and had lower rates of heart attacks and strokes than women not on Premarin, and for years it was believed that long-term Premarin treatment was appropriate as an off-label way of reducing chronic disease, including heart disease and stroke risk. However, a large-scale randomized trial in 1991 through 2004, involving more than 10,000 women and a total cost of hundreds of millions of dollars, found that not only was there no evidence that Premarin was effective in reducing the risk of stroke, but there was some evidence that Premarin *caused* additional strokes.

What went wrong? Were the doctors' observations faulty? No, the doctors' anecdotal reports were true, and they were corroborated by scientific observational studies that proved that women allowed to choose their treatment did better with Premarin than without it. The

problem was that *women who chose treatment with Premarin were fundamentally different.* On average, they were better educated, wealthier, located in urban areas with better access to health care, and so on.

When the treatment and placebo groups are chosen in some way that biases the response variable, we call that *selection bias.* The best way to avoid that selection bias is to randomize the assignment of women to groups—so that the choice of who gets Premarin and who gets the placebo is not in anyone's hands. After the women volunteered for the study, they had no say in whether they got the real Premarin treatment or a placebo. The trials were also double-blind, meaning that neither the women nor the people who dispensed the pills knew which treatment was being administered.

Q. What constitutes an "adequate number" of observations for replication?

A. For opinion surveys, it is easy to compute the required sample size in advance, and we will learn how to do this in semester 2. (And we have to, since one of the AP requirements is to be able to compute *n* for a survey.) However, for experiments where the E.S. is unknown before the experiment is run, it is much harder to compute *n*. If you can run a pilot test in order to estimate the population s.d., then you can perform a nasty spreadsheet calculation called a "power analysis" in order to plan your sample size before running the full-scale experiment. The AP doesn't require this type of spreadsheet analysis, but as you might expect, we will do some of this in semester 2.

Q. What is an SRS of subjects?

A. An SRS is a sample in which each *subset* of the specified size is equally likely to be chosen. Because this is hard to accomplish in the real world, we usually have to settle for something considerably different from an SRS. Luckily, experiments usually do not require an SRS of subjects. The most important thing in experiments is to make sure that the subjects you have managed to recruit are randomly assigned to the treatment groups.

For surveys, though, you still need an SRS or something close to it. On our field trip to the Pew Research Center on Dec. 11, we saw some of the ways that the decidedly non-SRS samples that Pew uses for public opinion polling can be tweaked and weighted so that they can be made to resemble an SRS. The tricks of how to do that are covered in more advanced courses in statistics.

Q. [Paraphrased.] If there are so many blocks that the number of subjects within each block is tiny (e.g., one person per block), isn't [the result] going to be less significant because [there is] a higher variance?

A. Surprisingly, no. In fact, the ultimate type of blocking ("blocking to the max," so to speak) would be what we call *matched pairs*, in which each subject is contained within a "block of one" and serves as his or her own control. For example, if we wanted to compare the effectiveness of two sunscreens, we should block on individual subject, with one sunscreen applied to one arm and the other sunscreen applied to the other arm. (The choice of arm should be made randomly, so as to cancel out any possible lurking variable, e.g., the fact that people without automobile air conditioning would be more likely to put their left arms on the door sill of their cars.) If there are more than two treatments, we can adapt this idea (not for

people's arms, obviously, but maybe in other situations) to create what are called *matched triples, matched quadruples,* etc.

Q. What is an example of randomization?

A. As volunteers arrive to participate in your study, flip a coin to decide whether the volunteer receives the test treatment or the placebo treatment. This is certainly not the only way of randomly assigning subjects to treatment groups, but it works.

Q. I still want to know the logic/purpose of replication (what it is used to reduce).

A. "Replication" means that n is large enough so that people will not plausibly be able to say, "Oh, sure, you saw an effect, but it could easily have been a statistical fluke." If I flip a coin 8 times and receive 6 heads, I can truthfully say that the sample proportion of heads is 0.75. However, even though the sample proportion of heads is an unbiased estimator of the true probability, we certainly don't have anything interesting to report in seeing a 0.75 proportion of heads. *There is insufficient replication of trials.* Would 600 heads out of 800 flips be more convincing? Yes, definitely, since $n = 800$ provides adequate replication so that the E.S. of 0.25 is statistically significant. (We will learn how to calculate the P-value of the test in semester 2.)

Common sense tells you that not all sample proportions of 0.75 are created equal. The correct conclusion if I see 6 heads out of 8 flips? *No evidence of any unfairness in the coin.* The correct conclusion if I see 600 heads out of 800 flips? *Extremely strong evidence that the coin is unfair.*

To answer your question, it is the s.e. of the sampling distribution of differences that we are attempting to reduce by having adequate replication. The payoff with bigger sample sizes, unfortunately, is not as dramatic as we might hope. Multiplying the sample size by 100 does not reduce the s.e. by a factor of 100; the s.e. decreases only by a factor of 10. (The square root of n plays a role in the denominator of the s.e.)

Q. Do we ever use binompdf or normalpdf?

A. Yes and no.

We use binompdf quite often. The binompdf(n,p,k) function on your calculator is the exact same thing as the binomial $P(X = k)$ formula found on the second page of your AP formula sheet.

As for normalpdf, however, we never use normalpdf in AP statistics. Never. If you want to learn about the relationship between normalpdf and normalcdf, you have to have at least a rudimentary knowledge of calculus. More advanced statistics courses, especially those that are more theoretically oriented, would make the distinction between normalpdf and normalcdf crystal clear. Mr. Hansen is happy to teach you if you really want to know.

Q. Is [statistical] significance also known as confidence intervals?

A. Not exactly. Confidence intervals can be used to gauge what is called "two-tailed significance" (i.e., difference in both directions from a hypothesized value). However, in semester 2, we will also learn how to compute "one-tailed significance" (i.e., a difference in only one direction).

We will learn all about one-tailed and two-tailed tests in semester 2.

Q. Why is double-blind testing necessary? Why must the researchers be blinded as well?

A. Double-blind testing is desirable but is not always necessary or even feasible. For example, in a recent study to determine whether arthroscopic knee surgery is effective or not, some subjects, randomly chosen, were given a sham surgery. (Details are in *The New England Journal of Medicine,* 12/26/2013.) The "placebo treatment" patients were anesthetized, had their knees cut, and spent time in recovery, but the surgeons working on them did not actually do anything "arthroscopic" to treat them. Obviously, there is no way for a surgeon to be unaware of whether she is doing a real surgery or a sham surgery!

The logic behind blinding the test subjects is quite clear. If patients know they are receiving a real treatment, their minds may interact with their bodies in subtle ways that make the patients improve. (This is the principle behind faith healing, and the amazing thing is that it often actually works!) We want to make sure that the benefit of the treatment, if any, can be isolated to the treatment itself and not to any other mysterious lurking variables.

As for the logic behind blinding the researchers, we need to make sure that the people in contact with the patients and recording the health status of the patients (tabulating their moods and their reports of pain, freedom of movement, and so forth) are completely objective in their recordkeeping. If a researcher has, for example, a stock investment in the pharmaceutical company that is developing a new drug, the researcher may "tweak" the reports: "Um, so you're receiving the drug, and you say you feel so-so? I'll record that as 'no pain.' And this other patient, who is on the placebo, reports feeling so-so? I'll record that as 'still feeling pain.' After all, it's a judgment call, right?"

We call this type of bias in judgment a *conflict of interest.* You may be familiar with conflict-of-interest charges in connection with political figures and referees in sporting events. It is important to avoid *even the appearance* of conflict of interest, since an appearance of a conflict of interest can be devastating to community morale. A politician who appears to be biased toward certain special interests, even if no law has been broken, should face some serious questions, and a sports referee who appears to have a conflict of interest should be fired. Do you see why some leagues prohibit their players and referees from betting on sporting events, even unrelated sports?

Something for you to think about: STA currently prohibits teachers from accepting lavish gifts from students or parents. The definition of what constitutes a "moderate gift" is not precisely specified in the policy. Do you think this is adequate, or should STA take a stronger stand? Should STA prohibit all gifts except those that are designed (like bagels) to be shared with the entire class? Should teachers be required to refuse all other gifts? Can the objectivity

of STA teachers in their grading be relied upon, or does the acceptance of gifts, even small ones, create an appearance of conflict of interest?

Q. What creates the "sweet spot" of around 1000-1300 people in a sample for an experiment/poll? Why not fewer?

A. First, we need to clear up one thing. For experiments, a sample that large is almost never needed, unless the E.S. is so small that a large sample is required to be able to establish statistical significance. Most clinical trials use sample sizes of a few hundred at most. Social science research experiments usually use even smaller samples, often no more than 25 or 50.

The Premarin experiment (see above) was an exception, since it used a placebo group and a treatment group that each had more than 5000 women. That's why the study was so expensive, and even so, not all of the questions that the researchers hoped to answer were settled definitively.

For polls, where the parameter of interest is often a proportion (instead of a mean), the sweet spot for n is indeed in the 1000-1300 range, but that is because the m.o.e. of 3 percentage points and confidence level of 95% have become customary. In semester 2, we will learn how to calculate m.o.e. from n, or vice versa, using suitable estimates (either from a pilot study or from a worst-case analysis) for p and q.

Q. What does the sampling distribution of the E.S. [effect size] look like?

A. It depends on the model we are using for our data.

Case A: If we are looking at a single mean or a difference of means, then the t model is usually appropriate. (If there are three or more means involved, we need an F distribution model, which is beyond the scope of our course.) The t distribution looks quite a bit like a z (normal) curve, except with fatter tails.

Case B: If we are looking at a single proportion or a difference of proportions, then the z (normal) model is usually appropriate.

Case C: If we are looking at the differences among three or more proportions, then the χ^2 model is usually appropriate.

We will learn about all these cases in semester 2.

Q. How does the researcher distinguish groups if he or she does not know who receives the treatment in a double-blind experiment?

A. Imagine that the pills are in bottles labeled "A" and "B" but are identical in all other respects. Somebody has to know which bottle is the real treatment and which bottle is the placebo, but it can't be anyone who has contact with subjects. The person who administers the pills should know only the name (or better yet, only the subject ID number) of each subject and whether the subject is supposed to receive "A" pills or "B" pills.

The researchers and biostatisticians who compile the data have to know which treatment is which. Otherwise, there would be no way of computing the relevant statistics. However, the people behind the scenes must be scrupulous about never giving any clues to anyone who has contact with patients. Ideally, there should be no contact at all between the people who see patients and those who don't. There should definitely be no offhand remarks. ("Whoa, those patients on the 'A' pills are looking pretty good, aren't they?") Any information that passes in either direction could be a problem, since it may give a clue about which treatment is real and which treatment is the placebo. Even if the clue is dead wrong (i.e., going in the wrong direction), that could bias the recording of data.

Do you know how sometimes a knowing glance, or a nod, or a slight turn of the head can give away information? That's the problem. There should preferably be no information flow at all between the people with patient contact and those without, since any information could bias the data in subtle ways.

The rules concerning how experiments are administered are called the *protocols* of the experiment. Protocols take a lot of work to set up and administer carefully. Protocols are part of a broader topic known as *methodology*, which includes everything regarding the design and execution of an experiment.

Tests for ESP (extrasensory perception) conducted using methodology that includes protocols to minimize information flow have consistently failed to find any evidence for ESP. One of Mr. Hansen's favorite tricks is to do a phony ESP demonstration. Students usually can't figure out how information is being communicated, but under rigorous protocols, the "amazing" effect disappears.