** Final Exams**

Posted on: Sunday, August 19, 2012

Both exams will be comprehensive in spirit, covering the full course's timeframe.  But certainly there will be some topics from the course that are not addressed on the exams.  The take-home part will be more calculational in focus, but that doesn't mean you won't be asked to interpret things.  It will be comparable to the midterm and less challenging homework problems in terms of difficulty level.  The proctored part is online, 90 minutes total, and is really more essay-based and qualitative in nature.  Both exams are open books, notes, etc.  For the take-home you can use SPSS and Excel as well as calculator, but for the proctored part, NO SPSS AND EXCEL. CALCULATORS ONLY - BUT NO CELL PHONES.  So make sure you have a stand-alone calculator.  In principle you may not even need a calculator for the proctored part, but if it helps you in any way you're welcome to use it.    
  
***\*\*\*The two parts are independent of each other and can be dealt with in whichever order you prefer.\*\*\**** You have the full one-week window to deal with both parts during Session 10.  They will be open by Monday August 20th at noon Central Time (if not before), and the deadline for submission of both is next Sunday August 26th at 11:55pm Central Time.  If at all possible, please try to avoid submitting both exams at the last possible minute.  SCS is only giving us 2 days to grade this term, as opposed to the usual week or so, so I will be grateful to anyone who's able to submit exams earlier in the week rather than later so that I can get a jump on the grading.  But this is a request, not a requirement.  I plan to hold a Sync Session next Saturday, August 25th at 1pm Central Time to answer questions, clarify any points of lingering confusion on the exam, etc.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Follow-up: Correlation, Regression & Causation**

Posted on: Saturday, August 18, 2012

Lots of good meaty discussion in Session 7 on these topics.  Many of you mentioned the maxim that 'correlation does not imply causation', and we had some good discussion of this regarding the correlation topic on sales and calls.  But some of you wondered whether a high *r-*value might somehow imply causation in a situation like this. No, it doesn't.  It simply implies that a straight line will be highly effective at describing the relationship in the data.  There is no magic value or range of values that allows you to conclude that one variable officially causes the other.    
  
Again, the correlation value - whether you consider *r* or *r2 -* indicates how effective a straight line will be at describing the relationship between two quantities.  But it is inaccurate to say things like, '*r=0.5* means the relationship only correlated half the time', or 'half the time the data moves in the same direction, the other half of the time it moves in different directions.'  Correlation isn't a binary property of individual data points that turns on & off as you move through the data. It's a global property of all the data taken together. You can't say based on a single correlation value that some data points are correlated and others aren't. It's a high-level summary of the relationship between two quantities that can't be applied at the finer scale of individual points.   It's an 'all in' property - either the data points taken in totality are correlated to some quantifiable degree, or they're not.  And you don't get to choose which points enter the computation, unless you subdivide your data and speak of correlations for different subsets.  But if you want to speak of the whole dataset, then you have to include the whole dataset.  
  
The same is true for the coefficient of determination.  ***You can't have multiple values for correlation or determination for a single dataset.  Nor can you apply one value to different subsets or interpret it differently for subsets of the original data***.  On the other hand, you could subdivide your data into smaller sets and compute separate correlations & determinations for those subsets.  It might be the case that half the data, say for smaller x-values, is barely correlated at all, while the other half for higher x-values shows strong correlation.  In that way you could speak of a range of correlations across your original.  That might be a valid follow-up technique if the initial correlation isn't very strong.  It could indicate a nonlinear relationship in your data, if there is a range of moderate values.  And a good correlation could be viewed as empirical evidence that strongly suggests a causal relationship between the variables, especially in hard science fields.  But only over the dataset that it applies to!  Be careful when extrapolating to larger a dataset the correlation value of a smaller subset contained within it, because that can get you into trouble!  And it probably shouldn't be the only piece of evidence for concluding causation, however strong it is.  
  
Again, *r=0.5* doesn't mean that half the time the data move together and half the time they move in different directions.  By that logic you could just as well argue that the correlation is -0.5 because the same statement would hold. Again, it's not a point-level metric that quantifies how often the points move together or differently within the original dataset.  What you can say is that, since *r2* = 0.25, 25% of the overall variation in y-values is accounted for by variation over all x-values. But again, that's taken as a whole, not on a point-by-point basis.  
  
Nor can you say '...variables are not correlated at .5 and that a value closer to 1 would be a stronger indication that there is a correlation."  If the correlation is nonzero, then the variables are correlated to some degree, and the value tells you to what degree. Correlation isn't something that 'stays off' until the value approaches 1; rather it's something that turns on if the value differs from 0, and it becomes more strong as the value reaches 1.  
  
I'm not trying to pick on anyone or embarrass anybody with these comments.  I'm just trying to point out how easy it is to make vague, squishy statements about these concepts.  We might think we understand them but they are quite subtle and we really have to think carefully about them to avoid misstatements.  In any case, here's a fun paper that talks about different ways to think about the interpretation of the correlation coefficient.  Optional and strictly for fun!  If you go through it you'll learn some history as well as trigonometry....  <http://data.psych.udel.edu/laurenceau/PSYC861Regression%20Spring%202012/READINGS/rodgers-nicewander-1988-r-13-ways.pdf>.    
  
Lastly, a single *r-*value cannot tell you the nature of the best functional relationship to fit the data, or whether the data is curvilinear as opposed to linear.  A horizontal line and the parabola *y=x2* both can have zero correlation, but one is curvilinear and the other isn't.  The *r-*value can only tell you whether a straight line is likely to be effective at describing the relationship between the variables.  But even that conclusion can be problematic - remember Anscombe's quartet!  You can have points all lying exactly on a curve that nevertheless exhibit a high degree of linear correlation between the two variables.  This is why monkeys will never replace us as analysts.  Whew!  
  
And now, a new corollary to the correlation maxim...  Repeat after me everyone:  'Regression does not explain causation, regression does not explain causation...'  Some of you went on to say that regression somehow adds a causal explanation to the analysis, or implicitly assumes a causal direction or relationship between the variables.  Some of you even quoted other websites making such statements.  ***This is not correct***, and the authors of such websites should know better.  
  
A regression model ***DOES NOT*** introduce, assume, or explain a causal relationship between the two variables.  It is simply providing a plausible functional relationship between the variables, offering more detail than the correlation does about this relationship.  The correlation is a single value that provides a holistic summary of the scatterplot.  Like most overviews, it is broad and lacking in specific detail.  It is *symmetric* in *x* and *y,* meaning if you swap *x* and *y* and recompute the correlation, it will be the same.  Regression goes beyond this by providing an equation that relates specific *x-*values to specific *y-*values.  It describes the data in a more detailed way than the correlation does, and gives you predictive capability that the correlation does not.  It zooms in on the data and makes more precise what the correlation only alludes to.  And the regression slope is related to the correlation coefficient via a simple multiplicative ratio.  So the two concepts are directly connected.  But regression itself has no power to bestow causality upon the variables beyond what has already been contemplated through the correlation results, as discussed above.  
  
You can regress *y* on *x,* then turn the graph around 90 degrees and regress *x* on *y,* as discussed in the GPA & SAT example in the Announcement.The correlation remains exactly the same.  But if regression introduced a causal relationship, then the first line would indicate how *x* causes *y,* while the second line would indicate how *y* causes *x.* Which is it??  If regression introduces causality, and we can regress in either direction, then how could the causality flow in either direction?  Regression can go both ways but one-way causality can't.  Merely drawing a line over a set of points cannot tell you any information that isn't already contained in those points.  The line is determined by the points.  It doesn't determine anything new about the points.  
  
As someone said in a recent post, 'the data alone do not show causation'.  They can suggest it, but they can't prove it.  Causation is imputed by human reason and judgment after making observations and analyzing the data.  It is a logical inference that we make based on our interpretation of the data.  As such, it is the result of things beyond the data being brought to bear on the data.  This is the essence of the scientific method - interpreting the data in light of additional knowledge and experience that lie outside the data.  In a sense this is what that sales manager failed to appreciate in the discussion topic...  
  
Furthermore, even if it is clear that two variables are not directly causally related, if they are correlated, you can fit a regression line and predict the value of one given the value of the other.  For example, ice cream sales might correlate strongly in the summer with shark attacks, homicides, bikini sales etc. even though there's no direct causal link.  But you can still do regression and gain predictive power, in the sense that given a value for the variable 'ice cream sales', you can predict what the value for 'shark attacks' will be.  That doesn't mean you're saying that ice cream sales cause shark attacks.  Causality is not a prerequisite for regression.  Correlation is the necessary and sufficient prerequisite for regression.  One is nonzero if and only if the other is nonzero.  
  
I think part of the confusion stems from the wording regarding *r2,* the coefficient of determination, which indicates the proportion of variation in *y* that can be explained by variation in *x.* As one of you asked, is there a distinction between 'caused by' and 'explained by'?  Absolutely!  In everyday language these phrases can be synonymous, especially since the words 'causal explanation' so often go together.  But here in the statistical setting, 'explained by' might better be said as 'described by'.  *r2* is not saying how much variation in *y* is *caused* by, or *results from*, variation in *x;* it is only saying how much of that variation in *y* can be accounted for quantitatively, graphically, numerically, by variation in *x.* The predictive capability of the regression line should also not be confused with causal explanation.  'Prediction' is another word whose meaning can be conflated with 'causation' in our imprecise everyday speech, but not in statistics.  Predicting something - for example, the outcome of a sporting event - is not the same as offering an explanation of its cause.  Language language language!  In statistics we need to say what we mean and mean what we say.  
  
At an even more basic level, we have potential confusion from the terms *dependent* and *independent variable*, which are universal terms used to describe any functional relationship.  But they do not imply causality.  '*Depends on'* doesn't mean '*caused by'.* These terms are saying which variable is the function's input and which is the function's output.  Saying that '*y depends on x'* simply means that *x* is the input and *y* is the output.  Again, you could turn it around and create a different functional relationship where *x* depends on *y.* But still, no causality is expressed or implied by these statements, and any similarity to actual causal relationships, living or dead, is unintended and purely coincidental...  
  
To sum up:  'Correlation alone does not imply causation, and regression alone does not explain causation.'

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Tests for Various Situations**

Posted on: Wednesday, August 15, 2012

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Explanatory Variable |  |
|  |  | Categorical | Quantitative |
| Response Variable | Categorical | Z-test or T-test for proportions Chi-Square test Fisher's exact test | Logistic Regression |
|  | Quantitative | Z-test or T-test for means ANOVA | Regression |

 People sometimes ask me about a reference table summarizing various statistical tests, when to apply which one, etc.  There are lots of these out there but here's a nice simple one I came across.  It's not comprehensive, and it includes some topics we're not covering in this course.  But it might be a handy reference or starting point for you moving forward.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Session 9 Under Way**

Posted on: Wednesday, August 15, 2012

This is our last week of covering new concepts and material.  Next week is for catching up and taking the final exam.  This week we are going back in Sirkin to pick up Chapters 6, 11, and 12 on contingency tables, chi-squared testing, and so forth.  In a way we are studying the discrete or categorical counterpart, you could say, of the correlation analysis that we learned to apply to continuous data in Chapters 13 and 14.  
  
As always, your work on the discussion boards is impressive.  Well done folks!  This will be your last week with a full set of 2 topics for discussion.  If we have a topic next week, it will likely be more of a summary/reflection looking back on the whole course.  I'll say more about it as the time approaches.  
  
Two final comments re: this week's homework.  On p3 of the solutions, step 2 should read "2. Multiply the actual number of \*\*Good\*\* (40) by that 40%. So, 40 x 40% = 16." i.e. replace 'Democrats' with 'Good'.  The point is that if there is no relationship between affiliation and health, and things are truly random, you would populate the table according to the relative percentages of each category.  Since Democrats comprise 40% of the total, that applies to any sub-total, such as the number of 'Good's, in the absence of any relationship.  (Also, in computing the chi-squared statistic, the first sigma symbol did not render in my page-view, although the second one did.  You should have two sigma symbols in the calculation walk-through.)  
  
***Also the solutions compute d.f. by adding rather than multiplying.  This is a typo.  It's always (r-1) TIMES (c-1).  Probably the author (not me) overlooked it because of the happy coincidence that 2+2 = 2x2!***

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Session 8: More on Expected Value, Average, & Law of Large Numbers**

Posted on: Sunday, August 12, 2012

Let's consider the formulas covered in your handout for *expected value* and *variance* of a random variable:

E(X) = μ = Σ (xi \* p(xi))  
  
Var(X) = σ2 = Σ ((xi -μ)2 \* p(xi))

where the summations run from *i = 1* to *n,* to cover all possible outcomes.  The expected value is also called *expectation value*, *average*, or *mean of the distribution*.  These formulas should look familiar because they are the frequency-level formulas that we've discussed previously.   
  
One reason I like to avoid using the term "expected value" is because it might lead you to think that you should "expect" that value to come up for the variable somehow.  But this is not true.  Don’t let the term mislead you.  It is just an average, whatever you call it.  The expected value of a variable, distribution, etc. doesn’t have to actually be one of the possible values.  It may be, but it also may not be.  It is merely an average, and hopefully you already appreciate the fact that the average value of a quantity may or may not be one of the actual values of that quantity.  We saw this in the roulette example from the previous Announcement.  On any given instance of that wager, you will either win $35 or lose $1.  Those are the only two possibilities.  The expected value is -$0.053, but you will never lose exactly 5.3 cents on your wager.  Rather, if you play many many times, determine your final net profit or loss, and divide by the number of plays, that average will come closer and closer to -0.053, the longer you play.   
  
By the way, the technical definition of 'gamble' is 'a wager with negative expectation value'.  No surprise there.  But did you know that the technical definition of 'investment' is 'a wager with positive expectation value'?  It's true.  In light of this, as well as the financial meltdown and behavior of toxic 'investments' in recent years, we could argue that sub-prime mortgages and the like should really be called toxic 'gambles', and many so-called 'investment managers' should really be called 'gamble managers'!  Why pay someone to help you lose your money??  
  
Anyway, back to our main topic.  In a previous Announcement I refer to the formulas at top as *frequency-level formulas.* How are these related to the *observation-level formulas* such as μ = Σx / n and σ =√[Σ(x-μ)2 / n]?  Why do we have two different-looking creatures that are both called "average" and represented by the same symbol μ?  Here's the connection:  if you listed all the values of *x,* allowing for repeats, you could then construct a frequency distribution for those values.  If you chose a value at random from the list, the probability *p(xi)* of choosing a particular value *xi* would simply be the relative frequency of *xi*in the distribution.  So the frequency-level formula is just using the relative frequencies to add up all the values and compute a *weighted* average, where the weight of each distinct value is its relative frequency in the list.  You're computing the same average, but this is a slicker way to do it using the frequency distribution, as opposed to adding each individual value to a running total and dividing by *n* at the end.  
  
Now let's shift gears for a minute and imagine that we want to determine the theoretical probability of some gambling event, like 'ball lands on red number' or 'first 2 cards give blackjack'.  In principle we may be able to do this using our methods, but what if the event is too complicated to analyze from first principles, like many situations in poker?  Or even just take the standard tossing of a fair coin, if you prefer, where we know 'coin lands heads' has probability 1/2.  In short, how could we empirically confirm a known theoretical probability, or even determine an unknown theoretical probability?  We do the thing itself - or simulate it on a computer via Monte Carlo, etc - many many times, and keep track of how often our desired event occurs.  In this way we determine the empirical frequency of our event.  When we divide by the total number of trials, we get a relative frequency of occurrence for our event.  
  
Now, this relative frequency is an empirical quantity determined by observing many many trials.  The theoretical probability is an abstract quantity that applies to one single trial, like the examples above.  Are we justified in making any connection between our empirical data and abstract theory?  How do we know that there should even be a connection?  The answer is provided by the Law of Large Numbers.  We encountered this before in discussing the Central Limit Theorem, Normality Assumption, etc.  There we used LLN informally and didn't say much more about it.  But to be precise, what LLN truly says is that *in the long run, as the number of trials increases without bound, the empirical relative frequency equals the theoretical probability.* In other words, LLN says that the more we simulate our event, the closer our empirical result has to be to the actual abstract probability of that event.  
  
This is kind of amazing if you think about it.  It says that random events in the real physical world around us, with all its messiness and imperfection, must somehow conform to the abstractions of conceptual thought that we use to analyze those random events!  How does Nature know that it is supposed to behave in this way?  What's to stop Nature from saying, "Screw your theoretical abstractions!  In the long run I can do whatever I want!"  These are deeper philosophical questions that I'll leave you to ponder, but I will say that LLN is actually pretty easy to prove mathematically, so even if She wanted to capriciously disobey, Mother Nature is prevented from doing so by a straightforward mathematical law!  Where random events are concerned, empirical probability must approach theoretical probability in the long run.  If you know the theoretical probability already, you will see it confirmed.  If you don't know it, the long-run empirical probability will give you a damn good estimate.  You have to love mathematics!  
  
The phrase 'Law of Averages' is a popular colloquialism for the Law of Large Numbers.  It is LLN that justifies our intuitive notion of probability as, 'the ***long-run*** chances of a random event occurring'.  But Mother Nature can do whatever she wants to you in the short term!  Some people mistakenly speak of the 'Law of Averages' in the short run, without allowing sufficient time for the long run to kick in.  For example, when the roulette wheel stops on red several times in a row, someone may think, 'The Law of Averages requires balance, so it's time for black to come up.  I will now bet my whole life savings on black.'  Then red comes up again and they wonder what went wrong.  The answer is: nothing.  'Law of Averages' only applies in the long-run, over many many repeated trials.  In the long run, balance will *always* be maintained, although you may have to stick around a long time to see it in any particular situation.  But in the short run, anything can happen.  That's how people get lucky and actually win at gambling, or the lottery!  But if they keep playing, they will watch their winnings slowly dwindle...

** Session 8: More on Binomial Probabilities & Problem Set 7**

Posted on: Sunday, August 12, 2012

I want to share some comments about the binomial probability distribution and Problem Set 7, in hopes of explaining to you in more detail what's going on, and helping to understand more deeply the fundamental formula that is presented in the handouts for solving these problems.  
  
Consider #2b in Problem Set 7.  Now, you don't HAVE to use the formula to compute the answer shown in the solutions; the formula is merely a convenient shortcut alternative to doing the calculations by hand.  Let's consider the longhand approach for a moment.  Each pair of teachers leads to the expression (0.3)2(0.7)6 for the probability that they will both leave (note that the solution mistakenly reads 32 instead of (0.3)2 ).  So you need some way to determine how many different pairs of teachers there are out of the total.  Without a formula, you need to count them.  Go ahead and try to list them out one by one, using letters for each teacher.  How many different pairs do you get?  If you list them all systematically and thoroughly, you should get 28 - AB, AC, AD, and so on, up through GH.  Each one of these contributes a probability of (0.7)6(0.3)2, so the final answer would be 28 \* (0.7)6(0.3)2, as indicated by the formula.  Note that via cancellation, 8!/(6!2!) = 8\*7 / 2 = 28.  When simplifying binomial coefficients, it's much better to cancel common factors first, rather than multiplying each factorial all the way out.  Otherwise you might get some very big numbers to deal with!  This is the issue some of you encountered with the Birthday Problems, when you tried first to compute 243! or whatever instead of canceling out common factors in the factorial ratios, and Excel couldn't do it.  243! is an incomprehensibly large number, maybe on the order of 10500.  Most machines can't deal with numbers that large.  On a side note, one thing we can learn from this is how quickly factorials grow.  10! is already greater than 3.2 million, and each higher number adds an extra digit to the final answer - so factorial values increase astronomically as *n* gets larger.  
  
Anyway, in the case of this homework question, you could do it by hand and list all the pairs if you had to.  But what if you had 50 teachers instead of 8?  Are you going to systematically and exhaustively count every possible pair from the total of 50??  If you do it correctly without missing a single one - and with 50 teachers that's a big IF - you'll get the correct number.  Go ahead and try it for a few minutes...  ready for a shortcut?  The shortcut is provided by the binomial coefficient, the first factor in the formula.  The point is that given *n* teachers, if you want to consider *r* of them leaving together, you're talking about the ***combination of n things taken r at a time***.  This is written a few different ways, and you can read Sirkin's discussion of this material on p262.  However you write it, you compute it using the "factorial fraction" given in both handouts as well as Sirkin.  For the given problem, we're talking 8!/(6!2!) = 28 pairs, so you could enumerate these by hand without much difficulty.  But for 50 teachers, we're talking 50!/(48!2!) = 1225 pairs.  Good luck listing/counting all of those correctly by hand.  Now you can appreciate the shortcut offered by the binomial coefficient!  It's not magic, but rather a concise way to count the total number of possibilities without listing them individually by hand.  Some people describe combinatorics, the branch of mathematics that deals with such things, as "counting without really counting".  
  
Now, as to the tables mentioned in the solutions... the handouts don't mention them, and you're not supposed to know them already.  You could easily Google 'binomial distribution tables' and find reliable ones in 0.2 nanoseconds - probably Wikipedia has some decent ones.  Once you know the binomial expression you need to compute, you can do it by hand or look it up in a table of values for the binomial distribution corresponding to your specific parameters *n, p,* and *r*.  This saves you the trouble of computing by hand.  Again, for the 8-teacher case, it's not hard to compute by hand, but for the 50-teacher case, you have 1225\*(0.7)48(0.3)2 to compute, and that's a lot more challenging by hand, unless you have a reliable computer package to do it for you.  Maybe Excel could do it accurately, I'm not sure...  So it's another shortcut to spare you tedious calculation.  
  
That's just for computing one specific value in the distribution.  When questions ask for P(2 or fewer leave) or P(4 or more leave), then you have to compute the specific values for each possibility and add them all up - e.g. P(4) + P(5) + ... + P(8) - to get the answer.  You could look each of these up in the distribution table and add them all up, OR you could use a CUMULATIVE distribution table, which gives you the sums of the individual terms for your parameters.  For example, the solutions indicate that you could look up P(>= 4 successes) from a table corresponding to 8 trials with a success probability of *p=0.3* (note that here I define 'success' to be 'teacher leaving' - you could also define it as 'teacher staying', in which case *p=0.7* and then you'd have to consult that table).  The point is that you have to essentially add up areas under the binomial distribution curve, and again for 8 teachers that's not hard to do by hand, but for 50 teachers, if you want P( >= 17 teachers leave), good luck with that one by hand!  Most people would use a computer or table to give them the answer.  In this respect the cumulative binomial table functions just like the normal distribution table in Appendix 1 of Sirkin, which gives areas under the curve for various intervals along the x-axis.  These values are values of the CUMULATIVE normal distribution.  In fact, the reading discusses conditions under which the binomial distribution can actually be approximated closely by the normal distribution, and that's a topic that we've already mentioned.  In such cases you could use the cumulative normal table to compute cumulative binomial probabilities.  But you cannot really use the normal approximation to compute a single binomial probability such as P(n=4); it only works for inequalities like P(n >= 4) or P(2 <= n <=5).   
  
So note that you have 2 different binomial tables to keep straight if you want to use them.  To compute P(exactly 4 successes), you would use the non-cumulative table.  To compute P( >= 4 successes), you would use the cumulative table to get the answer in one shot (as opposed to using the non-cumulative table to add up the individual values, which quickly becomes unwieldy for any more than a few terms to add together).  Since the HW problems only use 8 teachers, I suggest you try to work all the answers out by hand to be sure you understand the methodology, then use lookup tables to confirm your calculations.  That would be the best way to make sure you understand both approaches to these types of problems.  
  
One final comment just on the wording for #1.  The problem is discussing funding of grant proposals, a subject near and dear to the hearts of faculty throughout academia.  When the problem compares the two scenarios by describing how much time they take, it seems like it's talking about time needed to prepare the proposals, but it really means research time that would be funded by the grant if the proposal were accepted.  So the wording should really read something like, "If accepted, each proposal would fund 1/3 of your work time", or "If accepted, this proposal would fund 100% of your work time".  Then the question in 1b) about average funded time makes more sense.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Session 8, or How to Go to Vegas, Break the Casinos, and Come Home a Multi-Gazillionaire**

Posted on: Wednesday, August 8, 2012

OK, not really.  But you will learn about the mathematics underlying many games of chance this week, and in doing so you will learn why casino owners are multi-gazillionaires, even if you're not.  Welcome to the week where we get to study one of my favorite corners of statistics!

If any of you have ever stepped foot into a casino, played poker in someone’s basement or garage, or even just watched poker on TV, then you have encountered the concepts that we will cover in this session. Even if you are not a gambler and just like to play the occasional board game, you have – whether you know it or not – used binomial probabilities.

The idea of expectation value (sometimes called *expected value)* often comes into play when money enters the situation. In layman’s terms, the expectation value is the long-run average amount of money you can expect to win or lose per bet, if you keep playing over and over.  To make this more concrete, think of one of the most lucrative games for any casino: roulette. Why is this game lucrative? Let’s find out.

Roulette is a game where you spin a ball around a wheel and it lands in a numbered slot. In the American version of it, there are 38 possible slots: 00, 0 and 1 – 36. (Note that European roulette wheels only have 0.  We Americans are never content with just a little, and have to have more.  Why have a little edge for the house when you could have a little more edge?)  Imagine you bet $1 that the ball lands on one specific number.  The probability of hitting that number is of course 1/38.  Now, here is where the ‘house’ makes its money. The payout odds (what the casino pays the winner) are 35-to-1, meaning you will win $35 additional.  This means you have a 1/38 chance of winning, but get paid as if it were 1/36. The payoff is not quite as much as it should be, based on the rarity of the outcome and the true likelihood of winning.  That might not seem like a huge difference, but it is why they have all of those big, shiny buildings and you do not.  
  
To make it more quantitative, consider the expectation value of a $1 bet on your number. This is calculated according to the formulas that you will encounter in this week's handout and homework (and which you should recognize from a previous Announcement and homework):

E(x) = sum(outcome\*probability of outcome)

Here we have two outcomes: you win $35 or you lose $1.  Plug them in and watch what happens:

E(1) = $35(1/38) + (-$1)(37/38) = $-0.053

In other words, in the long run, you should expect to lose 5.3 cents per every dollar that you bet.  Of course in the short run you could win big, but if you stay and keep playing more and more, eventually the long-run average of -$0.053 takes over.  Does this sound like a lot?  Probably not.  But consider that every time someone places this bet on the roulette wheel, you could say they lose on average 5.3% of their money.  The casinos are basically taxing you for the privilege of playing.  That negative expectation value is from your perspective as the player.  But your loss is the house's gain.  The casino's expectation value is, naturally, +$0.053.  And how many roulette wheels are being played by how many people at how many casinos on how many nights of the year?  You can see why it's good to be in the casino business.  
  
That negative expectation value is what people mean by 'the house edge', and I guarantee you that it is negative, even if only slightly negative, for every game on the casino floor.  If casinos ever had a game with positive edge, they would lose money.  You can perform this same little analysis for any casino bet you like, crunch the numbers, and compute the expectation value.  If it's not negative, you've likely made an error - although, on occasion, casinos will tweak a certain bet in a specific game like keno to give players the advantage, as part of a promotion or to generate interest "for a limited time only!"  But it will never last forever.  Eventually the promotion will end.  And casinos are actually some of the nicest money-stealers, relatively speaking!  Go do the same calculations for carnival games or sideshow bets at the circus, or the video poker machine at your favorite tavern, and the edge will likely be even worse for the player.  
  
You should be able to confirm that if the casino paid you for winning your roulette bet according to the ***true*** chances of winning (i.e. $37 instead of $35), the expectation value would be zero. This is called "breaking even", and neither party has the edge now.  For fun, you could go further and figure out the expected value of hitting the same number twice in a row, or two times out of three spins. This brings the previously encountered topic of combinations back into the conversation, and we will see in our week's material how combinations help us to analyze situations involving so-called ***binomial probabilities,*** such as we have here in our example.  
  
By the way, you should be able to understand at this point why expectation value is often called 'average value' or 'mean'.  It's because expectation values ARE averages!  Look again at the formula above:  the outcomes are the possible values in the distribution of winnings or losses, and the probabilities are... probabilities, which are also relative frequencies.  Right?  So adding up those products gives you the overall average win or loss for the distribution, just like any situation where you compute the average via the frequency-level formula.  It's as though you had a dataset consisting of the winning values occurring at their respective relative frequencies, and the losing values occurring at their respective relative frequencies.  Doesn't make it any less painful when you lose, but at least you understand what the numbers are doing!  
  
In closing, you might say that casinos are an anomaly, only frequented by people with, um, gambling problems. Have you ever played the lottery? Just for fun, go to the Powerball web site, check out the odds of winning, and do your own calculations of the expected value. Then you will fully understand why (and this quote is attributed to numerous people), “The lottery is a voluntary tax on people who cannot do math."  Some people correlate lottery ticket purchases with demographics like annual income, and say "The lottery is a voluntary tax on the poor", etc etc.  (There are several different statistical analyses that could be carried out here for further study...)  At any rate, since you're here, you can do math, and so if you play the lottery, or gamble, or whatever, at least make sure that it's an informed decision with awareness of the calculations involved.  Enjoy yourself, but remember that knowledge is power!   And don't expect to win in the long run!

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Limitations of Correlation & Linear Regression**

Posted on: Monday, August 6, 2012

We have been reading a lot about how the strength of linear relationship is reflected by the correlation coefficient, the coefficient of determination (the square of the former) and the fitted regression line.  These analytic tools can be very effective ***when they apply***.  But they have limitations.  Be wary of applying them too quickly to everything you encounter.  Remember the old adage: when all you have is a hammer, everything looks like a nail!

Blindly jumping into computation without stepping back to consider the big picture can lead an analyst astray.  On p469, Sirkin has a nice set of scatterplots to illustrate the main concepts at play here, as well as the limitations of correlation & linear regression.  They show that correlation doesn't automatically capture every possible form of association between two variables.  But this is not the most dramatic way to drive home the importance of looking at the data before you dive into the analysis.

To see how important it is to look at the data, check out the famous example from Anscombe (1973).  We see here scatterplots of 4 datasets whose *(x,y)* values are quite different, yet they have the **very same** parameter values, such as mean, standard deviation etc. - and the regression lines are the same as well.  Moral of the story?  Do not "shoot first and ask questions later"!  
  
Here is a quick look via Wikipedia:   
[http://en.wikipedia.org/wiki/Anscombe's\_quartet](http://en.wikipedia.org/wiki/Anscombe%27s_quartet)  
  
Remember that correlation is strictly linear for our purposes in this class.  We assume that the data fall along a straight line, with random deviations (errors) that pull the observed values off the line.  This is a big assumption.  Other functional forms for the data are possible that linear correlation can't capture.  Maybe a power curve or some other curve fits the data better than a line.  Nonlinear regression and rank-order regression can handle more general relationships between *x* and *y.*  And certainly there can be multiple input variables *x1, x2, x3...* rather than one*.* But that's all beyond the scope of this course so don't worry about knowing any of it for now.  As you progress through the MSPA program you might encounter some of these more sophisticated variations.  
  
In everyday language, when we say things are correlated we often just mean "associated in some way", maybe more qualitatively than quantitatively, and not necessarily linearly.  But in the statistical context, *correlation measures* ***linear dependence*** *by default.*

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Applet for Linear Regression**

Posted on: Monday, August 6, 2012

From this link, you can check out the Applets for '11. Linear Models' for a nice hands-on demo illustrating the concepts of coefficient of determination & linear regression.  (Remember that page numbers, exercises etc. refer to a different textbook that we don't use.)  In the first applet you can drag lines around to see how errors, r2, and so forth change as the regression line changes.  Here r2 refers to the coefficient of determination offered by various candidate lines.  In general, when a straight line offers some reasonable fit to the data, the horizontal line through (x-bar, y-bar) is the reference point; for this line, r2=0 and the sum of squared errors (SSE) is rather large.  Then as you add some slope to the line in the right direction, the fit improves, r2 increases, and SSE decreases. When you hit the line of best fit, r2 reaches maximum value and now equals the r2 correlation coefficient for the data. At the same time SSE reaches minimum value. As you overshoot the best-fit line, the quantities revert to sub-optimal values.    
  
The point is that correlation is an intrinsic property of the data and doesn't depend on the proposed line, while coefficient of determination varies strongly in dependence upon the line being fit to the data.  The two quantities are equal only for the line of best fit, determined via OLS.  
  
Note the last example on 'Curvilinear Relationship' in the same applet. Here the opposite situation occurs. A straight line is not appropriate here, so the horizontal mean line is the best *linear* fit, and r2=0 is the best you can do.  SSE is minimal for the mean line and only increases as you add slope to the line in either direction.  So you're better off with no line at all!  In this case the nonlinear curve *y=x2* would fit the data perfectly.  
  
The second applet is a great exercise in understanding the impact of outliers on linear regression.  We already know that outliers can affect parameters like the mean, so it's no surprise that they can affect regression parameters as well - sometimes significantly.  Play around with the examples and see the outlier effect 'come alive'.  
  
<http://www.wadsworth.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9780495110811&token=>

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Comments on Multimedia: Ordinary Least Squares**

Posted on: Monday, August 6, 2012

In the Session 7 Multimedia presentation, we have an opportunity to ponder some issues surrounding linear regression.

Screen 8 asks us to solve for the SAT of a student who has a 4.0 GPA using the formula:

GPA = 0.0005(SAT) + 2.4779.  The calculations look like this:

0.0005(SAT) + 2.4779 = GPA

0.0005(SAT) + 2.4779 = 4.0

0.0005(SAT) = 4.0 - 2.4779

0.0005(SAT) = 1.5221

SAT = 1.5221 / 0.0005

SAT = 3044.2

So technically the correct answer is 3044.2.  But, of course, an inquisitive person might wonder how any high school student could get an SAT score higher than 2400.

This problem points to some of the limitations of linear regression.  Using regression to predict across the entire range of the independent variable (600 to 2400 in this case) may lead to poor predictions.  Also, extrapolating outside this range can lead to problems as well.  These caveats should be kept in mind when using regression models.

Furthermore, note that in this instance, instead of predicting a value of GPA from SAT, you are being asked to use the regression of GPA (dependent variable) on SAT (independent variable) as a basis for finding a value of SAT from GPA.  In other words, you are going in the opposite direction using the same formula.  But the regression formula for predicting SAT from GPA is different from the regression formula for predicting GPA from SAT.   If we were to use the data in the multimedia presentation to fit the regression of SAT on GPA, we would get a very different prediction for what SAT score corresponds to a GPA of 4.0.

The regression of SAT on GPA looks like this:

SAT = 341(GPA) + 804

Then for a GPA of 4, we would have SAT = 341(4) + 804 = 2168

Note that doing the problem this way yields a realistic answer for the SAT score, given that SAT's have a maximum of 2400.  Generally it's better to determine a separate regression line for going in the opposite direction, rather than using the original line "in reverse".

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Follow-up: Type I & II Errors, Jury Trials, and Drug Testing**

Posted on: Thursday, August 2, 2012

Hi all, we had a great discussion on the question of Type I & II errors in a jury trial setting, and I wanted to pass along some very interesting links that discuss the parallels between jury trials and hypothesis testing.  Related to this is the concept of Type I & II errors as they pertain to drug testing, quality control, medical diagnostics, and so forth.  We have discussed these topics at some length already.     
  
At the outset, let's agree that the null-vs-research hypothesis distinction is ***not*** whether the accused is presumed innocent until proven guilty.  The presumption of innocence is actually a founding principle of our judiciary, and it is not the question that the jury is being asked to decide.  Rather, the jury is being asked to decide whether to reject the presumption of innocence for this particular defendant.  It is thequestion of whether the accused party actually ***is*** innocent, in accordance with the presumption.    
  
Unfortunately, our incessant 24-hour news cycle and immediate media and information availability often lead to the conviction of an accused party in the court of public opinion before they appear in the court of law.  'Innocent until proven guilty' doesn't always apply in our snap judgments based on limited information.  But sometimes the court of public opinion is probably right.  Does anyone really believe that Casey Anthony had nothing to do with her poor daughter's demise?  Or that OJ Simpson really 'didn't do it'?  Still, the judicial process must be allowed to play out fairly with the presumption of innocence even for these people, lest we return to the benighted era of lynch mobs and vigilante justice.  
  
I think that the parallel between trials and statistical reasoning could be made even more apt if our verdicts were termed "Innocence Rejected" and "Innocence Not Rejected" instead of "Guilty" and "Not Guilty". That would truly reflect the spirit of hypothesis testing that the jury is asked to undertake.  The burden of proof rests with the state to present sufficient evidence to convince the jury to reject beyond a reasonable doubt the null hypothesis - namely, that the defendant is innocent as presumed.  "Beyond a reasonable doubt" is analogous to the significance level, say p=0.05 for a hypothesis test. 

** Follow-up: Misinterpretation & Diagnostic Testing**

Posted on: Sunday, August 5, 2012

This Discussion topic shows yet another aspect of the tricky relationship between probability and intuition.  Another cognitive slip that we often make with our intuition, in situations involving conditional probabilities, is to ***reverse the direction of the conditional.*** The problem posed in the forum can be analyzed via conditional probabilities.  For example, what is the probability of a positive test, given that someone has cancer?  This would be P(A|B) = P(positive test | cancer).  This information is given as 95%.  The problem is basically asking you to compute the reverse conditional P(B|A) = P(cancer | positive test), the probability of having cancer given that someone tests positive, and to determine in fact whether or not those two conditionals are equal.  (They are not.)  
  
Many people make that cognitive reversal without even consciously realizing it; some people make that switch consciously, by assuming that P(A | B) is the same as P(B | A).  'After all,' goes the thought process, 'it's just word order and syntax, right?  How could switching the word order materially affect the probability?  My intuition tells me that there couldn't possibly be a difference.'  Without engaging in the detailed computation to check that assumption, we fall prey to confusion and invalid conclusions.  Again I say:  syntax and wording may not make much difference in everyday non-technical speech, but they can make a huge difference in a technical setting.  
  
The proper way to go about computing the reverse of a given conditional probability is using something called Bayes' Theorem.  (This is the same Bayes whose name is attached to Bayesian inference - see the recent Announcement on Confidence Intervals.)  Several of you mentioned this or used it in your forum posts.  But others of you are likely unfamiliar with it.  This theorem lies at the heart of Bayesian statistics, for it shows us how to update our beliefs about a probability or parameter value (e.g. mean), given some empirical data.  You don't need to know all this to carry out the calculations to answer the question in the forum, but whether you know it or not, this is basically what you're doing.  
  
By now we all know that P(A and B) = P(A|B)\*P(B) by definition.  But we can also say P(A and B)=P(B and A)=P(B|A)\*P(A).  This is because order is very important for ***conditional*** probabilities - P(B|A) is generally ***not*** equal to P(A|B) - but order is immaterial for ***joint*** probabilities - P(A and B) always equals P(B and A).  If you're describing the probability of A and B occurring simultaneously, it doesn't matter whether you say "A and B" or "B and A" - it's the same outcome.  So we can say P(A|B)P(B) = P(B|A)P(A).  Divide both sides by P(A) and we get

P(B|A) = P(A|B) \* P(B) / P(A)

This is Bayes' Theorem, also called Bayes' Rule.  Given the values of P(A), P(B), and the forward conditional P(A|B), this equation tells you how to compute the reverse conditional P(B|A) in straightforward fashion.  
  
In terms of the Discussion problem, if A={positive test} and B={person has cancer}, we are given P(B) = 0.005 and P(A|B) = 0.95.  What about P(A)?  We have to work a little harder to get this, but we can do it on the basis of the given information.  The basic idea is  
  
P(A) = P(A and B) + P(A and -B)  
  
because for any two events A & B, either B occurs or does not occur in conjunction with A.  These two outcomes are exhaustive and mutually exclusive so we just add their probabilities to get the total probability P(A).  Now apply the 'AND' rule two more times to get  
  
P(A) = P(A|B)P(B) + P(A|-B)P(-B)  
  
Now, P(-B) = 1 - P(B) = 1-0.005 = 0.995, and P(A|-B) = 0.01 (this is the given 'false positive' rate - probability of a positive result when there is no cancer).  Putting this all together, we have  
  
P(B|A) = P(A|B) \* P(B) / P(A) = P(A|B)P(B) / [P(A|B)P(B) + P(A|-B)P(-B)]  
  
Notice that P(A|B)P(B) - the probability of a *correct* positive test - appears in both numerator and denominator.  This is because we are basically computing the proportion of positive tests ***that actually are correct***out of all positive tests, correct or incorrect.  Please be aware that this formula is exhibited and discussed in the Cartoon Guide.    
  
Plugging in the numbers gives  
P(B|A) = 0.95\*0.005 / [0.95\*0.005 + 0.01\*0.995] = 0.3231... < 1/3  
  
So less than 1/3 of positive test results actually correspond to cancer!  This is because the bulk of positive test results come from the 'false positive' category.  False positives may be rare, but the cancer rate is even *more* rare, so if you test enough people you will get an appreciable number of false positives, and these will outnumber the 'true positives'.  In this case the false positives numbered almost twice the true positives, simply due to the numbers involved.  Suprising results like this are collectively known as the ***false-positive paradox***, which is discussed very nicely, along with Bayes' Theorem, in the Cartoon Guide.  If you're more comfortable with the suggestion given in the problem, you can take the alternative approach of assuming a concrete total number of people to be tested - say, 10,000 or 100,000, which some of you did - and multiply all the probabilities by this total to get concrete people counts for each category.  But you still have to play the same game - out of all the people with positive tests, you have to compute the proportion of them that actually have cancer.  You will get the same answer.  
  
Results like this make a strong case against mandatory disease screening, mandatory drug testing, and so forth.  The problem is that the condition being tested for - cancer, drug use, whatever - can occur so rarely in the general population that positive results are wrong more often than they're right.  How many doctors or companies do you think understand this?  How many medical malpractice lawyers understand this?

Note how easy it is to confuse ***joint probabilities*** with ***conditional probabilities*** when we're thinking through these things.  Also note again how language can play a role in our misinterpretation.  The word 'accuracy' is given a precise definition in this context, but without that definition, if we just hear that a test is '95% accurate', we don't really know what that means.  It's a slippery statement that can easily be turned around in our mind when we're interpreting the results, so the layman might think that accuracy goes both ways, meaning symbolically that  P(A|B) = P(B|A), which clearly is not the case here.  To say that the test is 95% accurate, without specifying exactly what means, is to invite misinterpretation.    
  
In other words, imprecise language, or fuzzy understanding of the terminology, can lead us to reverse the direction of the conditional. We are given P(A|B) but we turn this around and interpret it as P(B|A), unconsciously equating the two.  But these two conditional probabilities are generally far from equal in a given situation, and that is precisely the point that this problem is making.  On the contrary, given P(A|B), we have to go to a fair amount of trouble to compute the value of P(B|A).  Given one conditional, we are computing the reverse conditional through this process.  How many people even understand how to do that?  No wonder confusion can occur so easily.    
  
This happens a lot in our everyday lives, and it's related to the similar cognitive misstep of ***reversing the direction of implication (fallacy of affirming the consequent),*** mentioned by Sirkin, whereby a person hears that A implies B.  Then they know B to be true, so they turn it around and conclude that A is true since B implies A!  But they were not told that B implies A.  E.g. "All hedge-fund managers are arrogant jerks; this guy at the gym is an arrogant jerk; he must be a hedge-fund manager!"  (Apologies to all hedge-fund managers out there...)    
  
We see how easy it is for our brains to reverse the orders of conditional probabilities and logical implications, and once again our language plays a role in that.  These are subtle issues so we're often not even aware that we're doing it.  I've no doubt that some marketing is based on manipulating people's tendencies to turn these things around.  If we understand these things clearly, we are less likely to be fooled.  Knowledge is power!

** Session 7 Under Way**

Posted on: Wednesday, August 1, 2012

Session 7 is under way, with both Discussion topics posted and open for your consideration. Overall I continue to be impressed with the quality and caliber of the many stimulating conversations, links, references, ideas etc. that you all are posting. I am definitely learning too from all of you!  Keep up the good work.  
  
For this week's material we jump ahead to Chapters 13 & (the first part of) 14 on Linear Regression and Correlation. These are two very important topics in statistical analysis that I'm sure a number of you are familiar with. Whole textbooks are devoted to the theory of regression.  We're just scratching the surface here with simple linear regression.  Do your best to understand the big-picture flow of the arguments without getting bogged down in every last detail of the math, if you're not keen on the details or it doesn't all click for you. Try to understand where the formulas come from and what they represent.   
  
And try to have a feel in words for what these concepts mean. You should be able to give a qualitative verbal description of these key concepts to your non-analytical poet and artist friends who ask you about them, even if you don't understand every detail of the calculations. Certainly SPSS & Excel are equipped to handle the calculations for you in the homework, examples, etc. but still it behooves you to appreciate the reasoning behind the formulas you will encounter.  The Discussion topics should help to clarify the key concepts as well.  
  
Minor correction on Solutions to the Problem Set:  at the end of #2, on page 8, the sentence reading "In other words, every additional year of experience yields..." should read "In other words, every additional year of \*\*education\*\* yields..."  I think everything else in there is OK.

** Errata: Problem Set 5 & Sirkin Tables**

Posted on: Wednesday, July 25, 2012

A few minor errors from Problem Set #5 and a typo from Chapter 10 that most of you have probably already caught.  
  
On #1, the problem comes from p311 of Sirkin (not p273 as stated).  Also, the square-root symbol in the denominator of the *t-*score expression seems to have rendering issues - at least for me - and may not show up on your copy of the answers.  So don't forget to take that square root when computing your *t-*score!  Also note that the expression for pooled variance has a few extraneous symbols in it.  The numerator of the first term should read *n1s12 + n2s22.*  
  
On #2, you are supposed to use the data set 'Employee Data' from the Course Content of Session One.  Also, at the end, I don't agree with the statement that the critical value for F is in the low 1's.  When I use an F-table with α=0.05, df-B (sometimes called dfnumerator)=1, and df-W (sometimes called dfdenominator)=infinity (since 472 >> 120), I get 3.84.  The conclusion doesn't change at all, though, so see if you agree with me on this detail.  
  
Finally, on p330 when he discusses the F-tables, Sirkin says that df-B is on the left column and df-W is on the top row, but this is backwards.  Really df-W is on the left column and df-B is on the top row.  Although he misstates it verbally, he does use the values properly in all his Chapter 10 examples, from what I can tell.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Session 6 Under Way**

Posted on: Wednesday, July 25, 2012

This week begins with an extension of what we covered in Sessions 4 & 5, when we learned about z-tests and t-tests and their application to hypothesis testing.  The idea behind both of these was that we were comparing the mean from a population with the mean from a sample in order to, in essence, determine whether the sample comes from the given population.  We spent a fair amount of time and effort working out the details of this.

What we are doing now is asking similar questions, but this time they are about what happens when we compare the means from two (or more) samples. In other words, we are no longer looking directly at populations.  Rather, we are now testing whether the different samples could all come from the same population, or whether they come from populations with genuinely different means.  If we are comparing two sample means, we will use what is called a ‘two-sample t test.’  If there are more than two sample means to compare, then the approach we take is called 'one-way analysis of variance,’ also known as ANOVA.  ANOVA is used in applications such as:

--When six different kinds of fertilizer are used for the same plants, we can test whether there is a difference in their growth.

--When three different machines are used to produce car doors, we can test to see if there is a difference in their output.   
  
There are dozens of other scenarios we could describe here.  One of this week's discussion topics asks you to come up with and discuss more examples like this.  
  
It might seem a little strange that we use analysis of *variance* to make decisions about population *means* as opposed to the *variances*.  However, remember that we calculate the variance (and standard deviation) by first determining the mean, and then subtracting the mean from each value, squaring, etc.  Thus, the variance is really dependent on the mean.    
  
Another way to think about it is that *variance* is a synonym for *variation,* and ANOVA helps us determine to what extent the observed variation in the dependent variable can be accounted for by the different categories of the independent variable.  We break down the total variance into two components:  a *between-groups variance* - the variation due to genuine differences between the sampled populations - and a *within-groups variance* - the variation due to random sampling error within each individual population.  We compare these two components of variation to each other, in order to test our null hypothesis that all samples come from the same population, i.e. the different *treatments* have no real effect.  If the between-groups variance is large enough relative to the within-groups variance, we can conclude that the samples likely do come from genuinely different populations, which means that the treatments do have a real effect.  
  
Another question you might have, is why bother with a new test when a t-test seems like it would suffice? If we had three samples, why not just compare the means for two of them at a time and skip Chapter 10 entirely? This approach requires six different hypothesis tests - if we want to test directionally - so if we set alpha at 0.05, the degree of confidence in our combined tests would be lowered to 0.956 (or 0.735). Thus the risk of a Type I error (rejecting the null when it is true) is unacceptably high at 26.5%.  ANOVA lets us avoid this pitfall by using one test - meaning only *one alpha* - for equality of several means.  I hope that is a convincing enough argument for you!    
  
As with last week, please do not get too wrapped up in the somewhat complicated mathematics - unless you actually enjoy such details like I do! :-)  Try to read through it, let the computer do the calculations and focus your own energy on understanding what they mean, interpreting the output, and drawing the proper conclusions.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Follow-up: Meaning of Confidence Intervals**

Posted on: Wednesday, July 25, 2012

Well, in reviewing all your discussion posts, I've been delighted to see all the different attempts you've made to parse the meaning of the 2 statements on confidence intervals, and to understand the proper interpretation of a confidence interval.  One thing this question shows is how much lies beneath the surface of such an innocuous phrase as "confidence interval".  We see once again how the language of statistics takes words from our everyday speech but uses them in a precise technical way.  'Confidence' in this setting doesn't mean exactly the same thing as 'confidence' in the ways we normally use it.  That language overlap can make it confusing to understand exactly what is being said by the term, which is a point I've tried to make consistently in my Announcements.  This is part of the curse of mathematics, that its concepts must be expressed in terms that often have different connotations or inflections in our everyday speech.  No wonder people get confused!  
  
Loosely speaking, we might think 'confidence' is synonymous with 'probability', because in our daily lives we hear and say statements like, "I'm 95% confident that this will happen", or "There's a 70% chance of rain today".  But it's a little more subtle than that.  As a few people have mentioned, this topic speaks to the distinction between 'frequentist' statistics and 'Bayesian' statistics.  This is a deep topic that is hard to summarize, and it involves genuine philosophical differences between those camps about how data should be analyzed and what we're actually doing when we make predictions based on observations.  I encourage you all to read about it in Wikipedia or other sources.    
  
The frequentist approach assumes that a parameter like population mean has a definite value, even if it is unknown.  So whenever we construct a confidence interval with specific endpoints, that unknown mean either lies in that interval with 100% probability or lies outside it with 100% probability.  It's all or nothing, and to a frequentist it would make no sense to say Statement 1, a la 'There's a 90% chance that the mean lies in this interval'.  If you think of a specific number that you keep to yourself, and ask me to give you a range of numbers within which I think it falls, without knowing your number I can say that it is either inside or outside my interval.  Although I don't know its value, I know it is a single definite value, which is either in or out of my proposed range.    
  
So the frequentist interpretation - ***which we use in this course*** - is basically Statement 2.  This is kind of technical and hard to unwrap, so that's why people including Sirkin tend to gloss over the point when they teach it.  But strictly speaking, this is our interpretation here:  If we were to take many repeated samples from the population and play the 90% "confidence game" - ha ha, I mean, the confidence *interval* game; no wonder statistics gets a bad rap! - then the unknown but single-valued population mean would lie within 90% of those intervals.  And this ***particular*** interval, constructed according to the procedure we've been using, ranges from 5 to 15.  The next one might range from 4.7 to 14.7, and so on.  (Note that these endpoints can and most likely will be different each time, ***so we are NOT saying that 90% of the time, the interval will range from 5 to 15.  Nor are we saying that 90% of the time, the mean will fall between 5 and 15.***)  Anyway, the kicker, of course, is that we have no way of knowing WHICH 90% of our intervals actually contain that unknown mean!  We only know that it will lie within 90% of them constructed that way in the long run.  See the blue figure about halfway down on the Wikipedia page for a visual illustration of this (<http://en.wikipedia.org/wiki/Confidence_interval>).  Note also on that same page, the section entitled '**Meaning of the term 'confidence' '.**  
  
The Bayesian approach says, well, not only do we not know the population mean, but it has a distribution of possible values before we start sampling.  Through the ongoing process of observation and estimation we refine our assessment of the situation. On the basis of observed data, we modify the distribution and update our original belief about where that mean value will end up lying.  This is the crux of Bayesian methods - to continually update our belief about some value based on empirical data.  So the Bayesian interpretation is basically Statement 1.  Because we consider the mean to have a distribution of possible values, we can say "we are 90% confident that the mean lies within this interval", and for a Bayesian that is essentially equivalent to "There's a 90% probability that the mean is in this range."  The next confidence interval would be used to update our belief about where that mean lies, and so on.  Bayesians actually use the term 'credible interval' to refer to this interpretation of confidence interval.  
  
Here's a simple analogy to drive it home:  take 20 blank index cards, number them from 1 to 20, and mark an 'X' on the back of 18 of them.  Put them all in a hat and draw one at random.  Then you have a 18/20=90% chance of drawing one with an 'X' on it.  Similarly, if you construct 10,000 different "90% confidence intervals" from 10,000 different random samples of the same population, and then ***choose one of those intervals at random,*** you have a 90% chance of choosing an interval that contains the population mean (which is constant over all intervals), even if we don't know and can't even figure out what that mean value is.  So in a sense 'confidence' does refer to 'probability', but not in the simplest sense of 'probability that this specific interval contains the mean'; rather it is 'probability that an interval constructed according to the procedure will contain the mean'.

** Midterm Now Available**

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Scores, Confidence Intervals & Hypothesis Testing - Comments & Clarifications**

Posted on: Sunday, July 22, 2012

* Regarding the use of z-score vs. t-score, sample size greater than 30 vs. less than 30, etc,  I think this quote from Sirkin (p249) is a nice summary:  "Thus, if we know σ and can therefore do a one-sample z-test, we always do the z-test, not the t-test.  We do the one-sample t-test only if σ is unknown and we must estimate it with σ-hat [or s].  In fact, when n gets large (say 30 or more), many statisticians advocate the use of the z-test, with S substituting for σ in the formula, rather than the use of the t-test.  But with a smaller n where σ is unknown, we must always do the t-test and retain the normality assumption for the population."  So if n is large enough, we can cheat a little by 'combining' the z- and t-test in this informal way.  But I'm not saying that you *should* do this or that you *have to* do this.  You can play around with the various approaches as you work through examples and see for yourself how the different calculations approximate each other.  Note carefully that the converse statement is NOT being made.  That is, for n large, you can use z instead of t with s substituting for σ, but that does not imply that for n small, you can use t instead of z with σ substituting for s.  It's not a two-way street.  Small sample size alone does not imply the use of t over z.
* Smaller sample means a greater chance that the difference between sample mean and assumed population mean is due to chance, i.e. sampling error.  That's why the critical values of z or t - whichever one we use - increase as n decreases, making it harder to reject the null hypothesis.  This is evident when you consider the formula for z- or t-score:  as n decreases, σ/√n increases in the denominator, so the score decreases as well.  The curve is broader and less steep, and our sample mean has to be farther away from the population mean in order to pass the requisite threshold (95% or whatever).
* There is a trade-off between the confidence level (e.g. 95%) and the size of the confidence interval.  If we want to increase our level of confidence, say from 95% to 99%, this comes at the cost of lengthening the interval, in order to capture more area under the curve.  So our range of estimates is greater as a consequence of increased confidence.  The only way to have full 100% confidence in an interval is to take the entire range of possible values as the interval!  "I am 100% confident that the true mean lies somewhere within the range of all possible values."  The *Cartoon Guide* has Sherlock Holmes making this very statement to his senator-employer.  Gee, thanks for the insight!  This is obviously not very helpful or informative.  At the other extreme, we could tighten our estimate by shortening our interval more and more, but then our confidence would be approaching 0%!  This also has limited utility.  If you want a shorter interval without sacrificing confidence, the only option is to increase your sample size.  In polling language, smaller margin of error at equal confidence implies larger sample, and vice-versa.
* Forget about hypothesis testing for a second, and just consider random sampling of a given population.  If n < N (sample size < population size), then there is always some finite p > 0 chance that our sample mean will lie quite far from the population mean, due to sampling error.  As n increases, we can be more confident (in the everyday, non-technical sense) that our sample accurately reflects the underlying population.  The only way to remove ALL uncertainty around our sample is to sample the entire population!  But if we could do that, we wouldn't even be having this conversation.  ***The whole point of inferential statistics is to determine to what degree we can make meaningful statements about a population - that is to say, infer characteristics of a population - when we can only access a sample of that population.***
* Hypothesis testing can be thought of as determining the likelihood that the population we have sampled and the population to which we are comparing that sample are the same - or at least have identical characteristics, such as the mean.  The p-value represents the probability of attaining a score at least as big as our computed score, *under the assumption that H0 is true*.  It is thus the probability that the observed value of the test statistic is in fact due merely to chance (*sampling error*).  This is equivalent to the observed sample actually coming from the same population, or at least a population with the same parameters (e.g. mean) as the original population.  That's why if p is small enough, we deem the evidence sufficient to reject H0, because it is so unlikely that our data resulted from simple sampling error.  BUT, no matter how large our z- or t-score may be, if n < N then there is always some finite chance (given by the p-value) that our conclusion to reject H0 is wrong, meaning the sample happens to be an outlier that really does come from the original population via sampling error.  So the p-value can be thought of as the probability of committing a Type I error with the specific sample that we have drawn - again, assuming *H0* is true.  We can decrease this chance of Type I error if we are willing to pay the price of increasing our sample size.  No free lunches in statistics!

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Sample Size vs. Population Size**

Posted on: Sunday, July 22, 2012

Some of you - especially if you are contemplating this material for the first time - might feel strongly that the sample size should be related in some way to the population size, in order to be able to draw valid inferences about the population from that sample, for a given confidence level.  This is an understandable feeling when we just think abstractly about the situation, but the magic of the sampling distribution is that this is not correct.  There is NO need to base our sample size n IN ANY WAY on the size of the population in order to draw valid statistical inferences about the population from that sample.  Intuitive though it might seem, we do NOT need to take a certain percentage of the population size as our sample size, or anything else along those lines, in order to play the game effectively.  Whether our population has size 10,000, 1 million, or 1 billion, if we want to construct a 95% confidence interval around our sample mean, then basic statistical theory says that we use the same fixed sample size in all situations.  
  
Of course, if we want that CI to be smaller and more tightly centered around our sample mean, we do need to increase the sample size.  But we still do not bring the population size into consideration.  In polling and survey work, which we'll consider a little later in the course, the sample size to be used depends on the desired confidence level and the desired margin of error.  That's all.  Nothing more.  Of course we do sometimes need to bring the population mean and/or standard deviation into the calculation, but never the population size.  The mean and standard deviation summarize all the important information that we need about the population.  In a sense, those parameters account for the population size already - think about the formulas we use to compute those parameters, wherein we divide by N and so forth - and remove the influence of population size from any further aspects of the analysis.  Populations of different size with the same values of mean and std dev will give rise to the same z-score or t-score, all else being equal.  
  
This kind of gets back to the point I made in my Announcement about CLT and so forth.  Once we start sampling, the population size is immaterial.  The sample says, "Give me the high-level parameters summarizing the population. I don't care about street-level details like population size, possible values outside the sample, etc."  I encourage you to go back and re-read that post, because this stuff is subtle.  It does require some clear, deep thinking to really get comfortable with it.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Errata: Problem Set 4 Solutions**

Posted on: Friday, August 3, 2012

First of all, the writer of the problem set (not yours truly) has throughout committed the same error that I have already seen from several of you - not putting parentheses in the numerator of a fractional expression.  a - b/c is not the same as (a-b)/c.  Order of operations matters.  That's why it's one of the first things that we learn in math.  
  
Here's a text rendering of the expression for t-score:  t = (x-bar - mu) // (s / sqrt(n))  
  
The point is the parentheses around the numerator (and denominator too, for that matter).  It could be argued that the formatting & large-bar style used for this expression in the solutions render the parentheses unnecessary, but you should never take these chances on mathematical expressions that leave them subject to misinterpretation by your colleagues or managers.  Order of operations means you divide first, then subtract; if you want the reader to subtract first, you put parentheses around the subtraction.    
  
Moving on...  assume normal distributions throughout all problems.  
  
Page 1, #1:  the 4th bold line should have "...t0.05,99..." since df=99 here.  
  
Page 4, #3a: the 4th bold line should have "...t0.01,99..." since df=99 here.  
  
            #3b:  the parenthetical clause should read "(t at 0.01 with \*\*24\*\* degrees of freedom)".  I think they're still stuck on #2 where n=50.  
  
Page 5, #4c & d: Instead of s/sqrt(n), they should use s/sqrt(n-1). Dividing by sqrt(99)=9.95 is of course not much different from dividing by sqrt(100)=10, but it's the principle of the thing.  Also in d), note that they have technically listed the z-value instead of the t-value, but for such a high df-value, the two are virtually identical so it's immaterial.  
  
Hope that helps --Matthew

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Follow-up: Birthday Problem, Probability & Intuition**

P(A) = 1- P(A')  
  
It is straightforward to compute P(A') through repeated multiplication for independent events, and by subtracting from one you have your desired probability.  Set this equal to 0.5 and determine the value of *n*, and you have solved the problem indirectly by considering the negative possibilities instead.    
  
A number of posts on the discussion thread provide details for the calculations.  In brief, for #1 with *n* people,   
  
P(A') = (364/365)\*(363/365)\*(362/365)\*...\*[365-(n-1)]/365 = 364\*363\*362\*...\*[365-(n-1)]/365n  
  
because each new person in the room must have a different birthday from all the others.  For #2 with *n* people,   
  
P(A') = (364/365)\*(364/365)\*...\*(364/365) = (364/365)n   
  
because each new person in the room need only have a different birthday from the instructor.  Personally I find #2 simpler and more intuitive to solve.  And it's worth noting that when we first hear #1, many of us fix a specific birthday in mind, perhaps in order to get a feel for the abstract situation via a concrete instance.  This is actually a great problem-solving strategy!  And it's probably something that our brains have evolved to do over time.  But if we forget that that fixed date can vary, we then end up solving #2 instead of #1!  Also, #2 is more amenable to direct solution via logarithms, if you know how to use those - whereas #1 really requires numerical iteration via a spreadsheet to arrive at the answer.    
  
1 - (364/365)n = 0.5 --> (364/365)n = 0.5 --> n = log(0.5)/log(364/365) = 253  
  
Take a look at the Wikipedia page for a great selection of graphs, tables, etc. that show all kinds of results related to the solutions.  There is also some discussion of the fact that in the real world, birthdays do not exhibit the uniform distribution throughout the year that we have to assume in order to answer these questions.  It's simplistic and not very 'real-world', but it makes the problems much more doable.  This is a common approach in mathematics - when faced with a complex problem, make simplifying assumptions until you can get a handle on it.  Then once you understand that, you can start to relax some of your assumptions and investigate more general and more complicated set-ups.  Here you could assume some particular form for the distribution of birthday probabilities for each calendar day - some research out there has tried to examine this - and then play the game based on that.  But note that your probabilities of matching/not matching would be conditioned on which specific day you're looking at, and this could make things quite a bit more complicated.  I encourage you to play around with this variation if you're interested, and you'll see how unwieldy the calculations can become.  
  
The sections titled 'Calculating the Probability' and 'Same Birthday as You' discuss the solutions for #1 and #2:  
  
<http://en.wikipedia.org/wiki/Birthday_problem>  
  
Mathworld is also a great website for diving into an overview of a topic or problem of interest.  The Birthday problem page is also interesting here:  
  
<http://mathworld.wolfram.com/BirthdayProblem.html>  
  
Another way to think about it is that the two problems are the same in terms of testing all possible pairs of birthdays for a match.  What distinguishes them is the criterion defining a pair to be tested.  For #1, the pairings are between all the students, which makes the problem harder to grasp conceptually because it tangles the birthdays all up with each other and makes things more complicated.  For #2, the pairings are all directly with the instructor, so the independence is fairly obvious and the analysis is straightforward.  This reveals a subtle connection between the problems.  To a close approximation we can view each paired comparison in #1 as independent of all the others.  Each comparison between 2 people has a 364/365 chance of failing.  So we have 1 - (364/365)n(n-1)/2 = 0.5 --> (364/365)n(n-1)/2 = 0.5 --> n(n-1)/2 = log(0.5)/log(364/365) = 253  
  
Notice how this is the same as #2 except the exponent is now C(n,2) - the number of combinations of *n* people taken 2 at a time.  But in both situations, the exponent is the number of comparisons.  It's just that in #2, the number of comparisons equals the number of students, since they're all comparing with the instructor, not each other.  So we have a little quadratic equation to solve for *n*; the use of factoring, quadratic formula, or trial-and-error yields... *n*=23!  In other words, since there are 253 2-person combinations in a group of 23 people (i.e. C(23,2)=253), the answer to #1 is 23.  
  
In closing, this elegant trick of 'turning the problem around' and approaching indirectly is a very helpful problem-solving technique.  One thing math teaches us is the value of holding a problem up to the light like the sparkly little jewel that it is, and turning it around to consider it from different angles.  In the Birthday case the Complement Rule is about the only reasonable way to attack the problems.  In other cases it's not necessary but it can be used as a shorter alternative.  In Problem Set 2, question 2e, the probability of absence on one or both days can be determined directly with a little effort, but a slicker approach is to compute it indirectly via the complement.  P(absence on one or both days) = 1 - P(absence on neither day) = 1 - 0.9\*0.9 (independent events) = 1 - 0.81 = 0.19.  Over and out.

** It's kind of subtle... CLT, NA, and LLN**

Posted on: Wednesday, July 18, 2012

Folks, these are subtle topics that run quite deep mathematically.  There is much that lies beneath the surface of these concepts and Sirkin keeps things very broad and straightforward but still useful, to his credit.  However, he doesn't state things exactly right, and he kind of mashes up all 3 of these concepts together.  So let me clarify some things.  
  
The Central Limit Theorem (CLT) is actually one of the deepest results in all of statistics, although if we had some more advanced material under our belts we would see it's not that hard to prove.   Sirkin misstates part of the CLT.  Remember that CLT is a statement about the sample, and not the original population.  More precisely, CLT is a statement about MULTIPLE REPEATED samples of the same size from the same population.  In fact, if you think about it, the mean values of all equal-sized samples form a population in their own right, and it is THAT population whose distribution is described by CLT.  When *n* is small, it's hard to say much about that distribution; as *n* increases, CLT says that the distribution of sample means will approach a bell curve, with mean and standard deviation related to the original population as specified by the theorem.  CLT doesn't actually say that if you start with a normal population, your sample means will also follow a normal distribution with the indicated mean and std dev.  That statement is correct, but CLT doesn't say it; it is actually just a straightforward result that can be shown with basic statistical theory.   
  
What CLT does say relates to the situation where the population distribution is arbitrary, maybe even unknown.  In that case, CLT says that as sample size *n* increases, then the sample means will follow more and more closely a normal distribution with the indicated mean and std dev, *regardless of the original population distribution****.*** Sirkin calls this statement the Law of Large Numbers, but that's not quite right either.  That's a different theorem that we'll discuss later in the course.  
  
On the other hand, the Normality Assumption (NA) is a statement about the population being sampled, rather than the sample itself.  It could be true that even if you have *n >> 30* (>> means 'much greater than'), the population being sampled still might not follow a normal distribution.  And it could also be true that a particular sample might follow a normal distribution when the population does not, or vice-versa.  The point here is that roughly for *n > 30*, Sirkin's Law of Large Numbers (LLN) - which is really the CLT - says that you do not NEED the Normality Assumption about the population in order to conclude that the sample mean will follow a normal distribution with mean and std dev as indicated - even if the population being sampled is non-normal.  For *n < 30*, you DO need the NA about the population in order to draw the same conclusion about the sample mean.   
  
These are high-level statements about the distribution of mean values from repeated samples.  They do not tell us anything about a particular sample or its relationship to the larger population.  CLT does NOT tell us that if *n > 30*, then any sample will follow a normal distribution; it also does NOT tell us that if *n > 30*, then the population will follow a normal distribution.  How could the nature of your sample affect the larger population that existed before you decided to sample it?  In short, the distribution of any particular *sample* is always irrelevant in determining the distribution of *sample means*; CLT says that for *n* sufficiently large, the distribution of the *population* is also irrelevant!  
  
Please understand that no matter what you do with "bootstrapping", variable transformations, or other exotic techniques that some of you might be familiar with, there is no way to magically force a non-normal population to suddenly follow a normal distribution.  You might be able to approach it from different perspectives, and you might be able to represent the data in different ways that shed light on the situation.  But ultimately the population distribution is what it is, and you may or may not even know what it is.  The point is that, whatever the population distribution might be, whether you even know what it is or not, under the right circumstances you can come to powerful conclusions about the distribution of the sample means. 

** Frequency-Level Calculations**

Posted on: Wednesday, July 11, 2012

Another topic I want to clarify relates to some of your homework problems, wherein you are given a frequency distribution of observed values in the form of a percentage table, and asked to compute mean, standard deviation, t-scores or z-scores, etc.  Later in the course you will have a handout that presents some of this material to you, but since it shows up now I want to explain it to anyone who might be unfamiliar with it.    
  
We have discussed the formulas for mean, standard deviation, and so forth.  We might call these "observation-level" formulas (Sirkin calls them 'individual data' formulas in Chapter 5), because you diligently process each observation in computing the mean and standard deviation via

μ = Σx / n and σ =√[Σ(x-μ)2 / n]

You add each observation to get the mean, or subtract each observation from the mean in computing standard deviation, etc etc.  In this way you compute the mean and standard deviation of the given sample - or maybe it's the whole population.  Doesn't matter.  But when you're given a frequency distribution, you don't have individual observations.  Rather, you have the relative frequencies or percentages of occurrence for each different value observed.  You could convert this information into observation-level information by repeating each value a certain number of times, in accord with its percentage, and then you could play the game with the observation-level formulas.  But there is a slicker way, using what we could call "frequency-level" formulas.  
  
Given the relative frequency data, we can use the expression

μ = Σx\*p(x)

to compute the mean.  This says, take the product of each value observed for x with its probability of occurrence p(x), and add up these products to get the mean.  What is the probability of occurrence p(x)?  The relative frequency for that value of x!  ***Relative frequencies ARE probabilities of occurrence.*** If a particular value in the distribution has a relative frequency  of 0.4, that's the p(x) value for that observation. 40% of the observations have that value.  Sometimes you see this with subscripts, which makes it a little more concrete: μ = Σxi\*p(xi), where each xi is an observed value.  But both formulas mean the exact same thing.  This certainly is more expedient than listing 40 occurrences of that value on a 100-slot list and doing the same for all other observed values in order to play the observation-level game.  
  
Once you've done this, you can move on to compute standard deviation via

σ2 = Σ(x-μ)2 \* p(x)

Note that this gives you the *variance*, which is σ2.  To get σ you have to take a final square-root.  In any case, this formula says, square the difference of each observed value from the mean you have just computed, just as you would do in the observation-level case, then multiply by the probability of occurrence for that value.  Finally you add up all these products to get σ2. A final square-root gives you σ.    
  
These are the frequency-level formulas that are used in the homework problems involving frequency distributions, e.g. the maternal health clinic problem (this week) and the SAT-gain problem (next week).  Note that they're quite similar to the observation-level formulas.  You still have to add, subtract, and/or square the same things.  The difference is multiplying by p(x), which saves you the trouble of having to list out and process each observation individually.  You can just multiply each factor by its frequency of occurrence instead.  But both approaches give you the same answers.  If you took the SAT-gain data, for example, and wrote down 10 instances of -100, 5 instances of -50, etc. and played the usual game to compute the parameters, you would get the same answers.  If you don't, you've made an error.  But frequency-level formulas save you some tedious bookkeeping work where silly errors love to lurk and trip you up.  If you had a frequency distribution for 35 different values over 10,000 observations, good luck converting that to an observation-level table and computing the parameters without error!  
  
One final note about the formulas...  These formulas as given are used to calculate ***population*** parameters:  μ, the population mean, and σ, the population standard deviation.  We can also use them to compute ***sample*** parameters, with a little more care.  As far as the mean goes, if we want to compute a sample mean *x-*bar, we change the symbol from μ to *x-*bar but it's still the same formula.  Now, if you are computing the standard deviation for an entire population of data, you just use the given formulas without further complication.  If you are computing the standard deviation for a ***sample*** of data from a larger population, then as the formula stands you're computing Sirkin's *s,* not σ. So again you change the symbols and use the same formula, with *x-*bar in place of μ and *s* in place of σ.  ***However***, you need to keep in mind that the frequency-level formula implicitly uses a denominator of *n*.  This denominator is explicitly clear in the case of the observation-level formula when you choose to divide by *n* and write it down that way, but it is hidden in the frequency-level formula because you never have to write down a denominator.  Thus if you want to compute a *t-*score on the basis of this sample standard deviation, computed via the frequency-level formula, you then need to use *n-1* in the subsequent square-root term in the formula for *t.*  If you convert the data to observation level in order to use the observation-level formulas, then you can choose whether to divide by *n* or *n-1* when computing *s.*  But if you go with the frequency-level formulas, that *n* value is 'baked into' the denominator, and the choice is made for you.  Remember, your calculation of the *t-*score must involve one factor of *n* and one factor of *n-1* in order to be correct.

1)  Sirkin p253:  First sentence at top of page should read, "Suppose you did not have access to \*\*σ-hat\*\* but did know .... "  The point he's making here is what to do if you know *S* but not σ-hat (although they're related through simple algebra).  
  
2)  In the Multimedia on Confidence Intervals, slide 7 uses a splashy graphic to say that S is an "unbiased estimate" of σ.  This is a technical statement meaning that if you take many samples of size *n* and compute S2 for each one, then the average value of S2 will approach σ2 in the long run.  (Technically, owing to the square-root involved, S2 is an unbiased estimate of σ2, but S itself is not exactly an unbiased estimate of σ.  You actually have to massage S further to make it an unbiased estimate of σ.) The problem is that this is all true only if you use *(n-1)* in the denominator of S2, rather than *n.* They don't show their formula anywhere in the slides, so we have to take it on faith that they're doing that (you could check the computations in the interactive slide at the end to be sure).  But most people use *(n-1)* by default when computing S or S2, so it's a safe assumption.  Also note that Sirkin's formula for S uses *n,* whereas he uses σ-hat to refer to the *(n-1)* version.  Whichever version you use, the discrepancy between them is only substantial if *n* is small; this is what led to the development of the unbiased estimator in the first place, to deal properly with small samples.  That discrepancy gets smaller and smaller as *n* gets larger and larger, rendering the choice less and less important for large samples.  
  
This is the point that Sirkin is making in the sidebar box on p138.  Most people - including the Multimedia - use *(n-1)* in their formula for S and call it the '*sample standard deviation*'.  Sirkin sticks with *n* and calls his S 'sample standard deviation' as well, using 'σ-hat' for the *(n-1)* version*.* To make matters worse, people sometimes refer to the *(n)* version as 'standard deviation of the sample', to distinguish it from the '*sample standard deviation*' with *(n-1).* (See why statistics terminology can drive you crazy if you're not careful??)  So multimedia S2 = Sirkin σ-hat2 = unbiased estimate of σ2; Sirkin's version of S2 is a *biased estimate* of σ2, due to his use of *(n)* rather than *(n-1)* in the formula.  
  
The point is that you can use whatever symbol and name for it you want, and you can even start out by computing Sirkin's S first if you want, but if you're ultimately calculating a t-score, you have to have that *(n-1)* factor somewhere in the calculation to get everything to work out exactly right.  This brings us back to the text that I'm referring to in point (1) above.  Sirkin is trying to explain that when you compute a t-score, you can use his S, then divide by sqrt(n-1), or you can use his σ-hat (which is everyone else's S) and then divide by sqrt(n).  But it's the principle of the thing.  I want people to understand the different formulas and not be confused by the fact that the book uses one thing and the slide show uses another, unfortunately without showing their formula to minimize confusion.  This is sound advice for analytic communication:  if there's any possible room for ambiguity of interpretation, it's always good practice to explicitly state the formulas used in your calculations.  
  
Besides that one complaint, the Multimedia slideshow is overall quite a good supplement on Confidence Intervals, and if you go through it with that caveat in mind I think you'll find it pretty helpful.  
  
3)  In the solutions for this week's Problem Set 3, #5b) should have "S/√(n-1)" rather than "S/√n" in the expression for confidence intervals based on the *t-*score.  I have just explained the reason for this in 2).  The calculation of S in the previous step basically used 'n' by virtue of using the frequency-level formula, so now you need to use 'n-1' in the square root.  The end result will change very little, but I want to use the correct expression on principle.   
  
4)  Other Problem Set 3 clarifications:  Assume normal distributions throughout all problems, if not explicitly stated; for #4, you're working with confidence intervals for the *mean* score of all enrollees in the program, and the standard deviation of 10 refers to the difference between enrollee scores and the Harvard average.

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

** Answer Key Corrections**

answer key for Chapter 7.  I'll post it on the Course Content page for this week.  
  
Exercise 7.2  
1. One-sample z  
2. ANOVA  
3. Two-sample t  
4. Chi-square  
5. Two-sample t  
6. Chi-square  
7. ANOVA  
8. One-sample t  
  
Exercise 7.5  
1. p<0.001  
2. p<0.05  
3. p<0.01  
4. Not significant at 0.05 level  
5. p<0.01 (p=0.01 if you round off)  
6. Not significant at 0.05 level  
7. p<0.01  
8. Not significant at 0.05 level  
9. p<0.05 (p=0.02 if you round off)  
10. p<0.02  
  
Exercise 8.1  
1. 0.0495  
2. 0.4505  
3. 0.9505  
4. 0.4483  
5. 0.0517  
6. 0.5022  
7. 0.4978  
8. 0.123  
9. 0.3238  
10. 0.701

Posted by: Matthew McCallum

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

Posted to: 2012Summer PREDICT\_401-DL\_SEC56 Introduction to Statistical Analysis 2012SU\_PREDICT\_401-DL\_SEC56

Here's an interesting question to keep in mind throughout the course: what should we do when someone makes a statement in a Discussion Board thread that is obviously incorrect?  Suppose, for example that a student were to say: "Whenever a distribution has high-end outliers, the median will be higher than the mean."  By now I hope you all realize why this statement is problematic.  
  
  
Another thing we can do is refer to the textbook or another reference source: "My reading of the course material for this week (Sirkin 2006, 111-112) suggests that just the opposite would be true:  When high-end outliers are present, as we would expect for a positively skewed distribution, the mean is likely to be higher than the median."

** Mean, Median, Mode and All That Jazz**

Posted on: Saturday, June 30, 2012

**In our topics this week we're grappling with *Measures of Central Tendency* (mean, median, mode) and *Measures of Dispersion* (range, variance, standard deviation).  These comprise the essential aspects of *Descriptive Statistics,* where we focus on how to best summarize and describe a collection of data points.  Soon we'll start to discuss the other fundamental component of stats, *Inferential Statistics*, where we seek to draw inferences about a larger group - typically called the 'population' - based on data from a sample of that group.  Anyway, I've been quite happy with the discussion so far, but I often see some subtle misstatements when I teach this material.  So I want to clarify some things and shore up our collective understanding.**

**Just to set the stage a little... if you have a collection of data that you have to describe to someone - whether a frequency histogram, a distribution curve, or simply just a list of raw numbers - you should ask yourself, "Self, what is the most effective way for me to summarize these data?  What key parameters could I give that would best describe these values?"  Upon reflection, you could decide to convey two main features of the data:  the 'middle' of the data, whatever that means, and the spread of the data around that middle.  Measures of central tendency represent various ways of addressing the first feature - where is the center, the 'middle' of my data?  Measures of dispersion represent various ways of addressing the second feature - what is the spread of my data around that middle?**

**Keep in mind that all metrics have their own pros and cons; no statistic is perfect in every situation.  Part of our job when we work with data and communicate about data is to be aware of the various pros and cons of different statistics, and to make informed decisions about which metrics to use in a given situation.  One of the main purposes of this course, then, is to make sure that you all have a thorough grounding in such thought processes, if you don't already, so that you don't just blindly or reflexively regurgitate numbers without really understanding what you're doing.**

**Part of the confusion we sometimes run into when discussing or thinking about these concepts has to do with language itself.  The very words we use have precise meaning in a technical context, but since they also exist in our everyday speech, they can have broader or looser connotations that we don't realize are influencing our interpretation and use of them.  When we're stumped on a problem or confused by something, the first thing we should do is make sure we know exactly what's being said by the words used.  We may think something is being said that actually isn't upon closer inspection.  We will then start to realize what's not being said, and sometimes that distinction is the difference between confusion and clarity.  Sirkin's definitions are a little wordy for my taste, but he is trying to reach a non-technical audience, which more people should be doing, so I applaud him for that.**

**The flip side of this is that language that can be misinterpreted is language that can be misused, sometimes deliberately.  The bottom line I think is to be a little wary of the language used in *any* data analysis, and to be aware of the subtle ways that it could be misused or misconstrued.  After all, even if we're on our guard about it, the sources of our information, analysis, etc. might not be so concerned with precision and exactly-right-ness themselves....  In any case, we will see this theme of slippery language and precise wording recur throughout the course.**

**So, on to the first group, *measures of central tendency.* We'll start with mode.  We could define *MODE* succinctly as the value in the distribution (i.e. the histogram) having the greatest frequency.  Sirkin generalizes this to, values in the distribution having greater frequency than their nearest neighbors.  If you know calculus, this is nothing more than the distinction between global and relative maxima on a curve.  In this language, a mode is a distribution value where the frequency curve attains a relative maximum.  \*\*\*NOTE THAT THIS IS DIFFERENT FROM THE MAXIMUM OF THE DISTRIBUTION ITSELF.  THE MAXIMUM OF THE FREQUENCY IS NOT THE SAME CONCEPT AS THE MAXIMUM OF THE VARIABLE  WHOSE FREQUENCY DISTRIBUTION IS GRAPHED.\*\*\*  The mode of a distribution can easily occur at a value of the distribution that is close to or at the minimum of *x*.  Graphically, the mode is the *x-*value with maximum *y-*value, NOT the maximum *x-*value.  Although in theory those two values could coincide, they certainly don't have to.**

**So, even though it's lumped in with other measures of central tendency, note that the mode of the distribution can occur anywhere, far from the mean or median.  In this way the mode is really a measure of clustering of the data around certain values, rather than center-ness of the data.**

**The men and median are the two most common values used to describe the 'middle' of the data.  *MEAN* is used interchangeably with *average*, which is OK, but in everyday discourse we often think of *average* as "typical" or "normal".  (Arrgh, yet another word - *normal* - with differing technical and non-technical definitions!  We will discuss the famous 'normal distribution' a little later in the course.)  And as we're learning, a distribution's mean can be far from a "typical" or common *x-*value in the distribution, because of the influence of outliers.  *MEDIAN* is another name for *50th percentile -* i.e. the *x-*value below which and above which 50% of the distribution lies, after accounting for the frequency of the values.  Graphically, it is the *x-*value that divides the area under the frequency curve into two equal halves.  And note that while the mode must be an actual *x-*value in the distribution, the mean may or not be an actual value of the distribution.  The median also may or may not be an actual *x-*value, but in general we can say that for a discrete distribution (that is, one having a histogram barchart for its distribution, rather than a continuous curve)  with an odd number of data values, the median will necessarily be one of those values.  For a discrete distribution with an even number of values, the median will most likely not be an actual value in the distribution, since it is found by averaging the 2 values next to the middle position.**

**For those of you who know some physics, there's a nice way to think about mean vs. median.  If you had a very thin metal plate in the shape of your frequency curve, histogram, etc., then the mean would be the center of mass of that plate, the point where the plate would exactly balance if placed on a fulcrum.  This need not be at all the same as the median, which would simply be the value that divides the plate into 2 halves of equal mass.  The median doesn't care about the distribution of mass on either side of it, as long as both sides remain at 50% of the total.  The center of mass - the mean - on the other hand, cares very much about the distribution of mass on either side of it!  The distance of the mass from the center is an important component of the mean, but the median only cares about the total amount of mass on either side.**

**A see-saw can balance with different weights on each end, but only if they are placed at different distances from the fulcrum.  The balance point is the center of mass.  The mean cares about the distribution of the weight on each side, but the median only cares whether the weight on each side is half of the total.  You should be able to convince yourself with some drawings that a frequency curve can have a fixed median while the mean slides all over the place, if the curve grows a long tail on one side (skewness).  This is the outlier effect, plain and simple.  As the weights slide around and move along the see-saw, the balance point must move as well in order to maintain the balance.  If you care to go farther with this, the analogy can be extended by interpreting the variance of the distribution as the plate's moment of inertia about its center of mass (the mean).  Ahhh, rotational mechanics.... but that's another discussion!**

**This brings us to *measures of dispersion.* The *RANGE* is simply the difference between maximum and minimum *x-*values in your dataset.  While this is helpful for getting an initial feel for the distribution, kicking the tires on your data to make sure there are no red flags, etc., it doesn't tell you anything about how the individual data points are distributed within that range.  The intermediate data points could be anywhere in that range, with any frequency.  So we need to move to more sophisticated measures in order to understand the distribution more deeply.  There are various ways of doing this, but the most common, and the ones we'll focus on in this class, are *variance* and *standard deviation.* The *VARIANCE* is the average squared distance of each data value from the mean, which can only be computed once the mean is known.  The *STANDARD DEVIATION* is simply the square root of the variance, making it the root-mean-square distance of the data from the mean.  Those of you with an engineering background will already be familiar with things like root-mean-square voltage in other settings.**

**Finally, *SKEWNESS* is technically a property of an entire distribution, rather than of a single value like the mean.  We can talk about skewness of the mean, for example, which is kind of loose usage.  But again, this is forgivable because in our everyday language we talk about "skewed perceptions" or "skewed perspective on the matter" resulting from a specific influence on a person's viewpoint.  Even in statistics we speak of a single value "skewing the distribution", meaning there is a long tail on one side or the other of the curve, due to an extreme outlier.  So in our thinking, we conflate the distributional property with the effect of one or a few values on the distribution.  Again, this is OK as long as we know what we really mean.  But if we say that 'the distribution skews the mean', or 'the mean skews the distribution', or 'the mean is skewed', what is that relative to?  Skewness of a single value has no intrinsic meaning without a reference point.  Also, the distribution determines the mean, not vice-versa.  Many distributions can have the same mean, so the mean doesn't by itself determine skewness.  It's the spread of data around the mean, the location of the rest of the data in relation to the mean, that does so.  In this way, skewness is also a measure of dispersion.**

**It turns out that skewness is computed like the variance, only the third power of distance from mean is used instead of the second.  And if you move up to the fourth power, you get something called *kurtosis,* which some of you may be familiar with.  But that one is harder to grasp intuitively, and we won't discuss it in this class...**

**As a final thought, I think what we often mean when we say e.g. 'the mean is skewed' is that the distribution is noticeably asymmetric.  The degree of difference between mean and median indicates the degree of symmetry in your distribution curve/histogram.  The more mean and median are different from each other, the more skewness you see in the data. The closer they are, the more symmetric the distribution.  Some skewness = asymmetric = mean and median unequal.  Little or no skewness =  highly symmetric = mean and median close to each other, or equal.**

**So you see folks, in this harsh world of ours, it really is "skew or be skewed!"**