draft October 25, 2021

Notes on the Mean, the Standard Deviation, and the Standard Error.
Practical Applied Statistics for Sociologists.

An introductory word on philosophy of the class:

My goal in this class is to give you an intuitive understanding of some basic statistical ideas, and to give you practical experience with using basic statistics. Toward these pedagogical ends, I dispense with as much statistical formality as I possibly can. We will talk a little bit about linear algebra and about calculus (two bedrocks of statistical theory) in this class, but only as asides.

Consider a variable X. E(X) is the expected value of X or the mean of X.

The formal definition of E(X) is

$$E(X)=\sum_i x_i p(x_i)$$

if X is a discrete function, meaning you sum up the different outcomes weighted by how likely each different outcome is. If X is a continuous function, the expectation is defined this way:

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$

where f(x) is the probability density function. Expectation is an important idea, but it is somewhat abstract. If X is Normally distributed (an assumption that is actually quite reasonable in the kinds of data and questions we will be looking at), and we have a bunch of x's to observe, the sample mean of our x's is actually equal to the expectation of X. Since the sample mean is very concrete and tangible and already familiar to you, I am going to talk a lot about the sample mean and not so much about E(X).

These are notes on the Sample mean, the Variance, the Standard Deviation, and so on. In this discussion you will have to know a few basic things about summary notation:

$$\sum_{i=1}^{n} X_i=(X_1+X_2+...+X_n)$$

$$\sum_{i=1}^{n} aX_i=a\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n}(aX_i+bY_i)=a\sum_{i=1}^{n} X_i+b\sum_{i=1}^{n} Y_i$$

In words, summary notation is just a sum of things. No big deal.
When you multiply each value by a constant, it is the same as multiplying the sum by the same constant. If the sight of summary notation scares you, don't worry. Summary notation is just shorthand for a simple idea.

1) The Sample mean, or the average. If we have n observations, X_1, X_2,....X_n, the average of these is simply

$$Avg(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

In other words, you take sum of your observations, and divide by the number of observations. We will generally write this more simply as

$$Avg(X_i) = \frac{1}{n}\sum X_i$$

This is a formula you are all familiar with. The simple formula has some interesting implications.

2) How the Average changes when we add or multiply the X_i's by constant values. In all the below equations, *a* and *b* are constants.

$$Avg(aX_i) = \frac{1}{n}\sum aX_i = \frac{a}{n}\sum X_i = a(Avg(X_i))$$

When we take a variable and double it, the average also doubles. That should be no surprise.

To be slightly more general:

$$Avg(a + bX_i) = a + b(Avg(X_i))$$

(this is easy enough to show. See Homework 2)

3) Also, the Average of a sum of two variables is the sum of the Averages. More formally:

$$Avg(X_i + Y_i) = \frac{1}{n}\sum (X_i + Y_i) = \frac{1}{n}\sum (X_i) + \frac{1}{n}\sum (Y_i) = Avg(X_i) + Avg(Y_i)$$

If X is January income, and Y is February income, then the Average of January plus February income is the same as the Average for January plus the Average for February. No surprise there.

4) The sample variance, defined:

$$Var(X_i) = \frac{1}{n}\sum(X_i - Avg(X_i))^2$$

The Variance is basically the average squared distance between X_i and Avg(X_i). Variance can't be negative, because every element has to be positive or zero. If all of the observations X_i are the same, then each X_i= Avg(X_i) and Variance=0. Variance has some down sides. For one thing, the units of Variance are squared units. If X is measured in dollars, then Var(X) is measured in dollars squared. That can be awkward. That's one reason we more usually use the standard deviation rather than the variance is that the standard deviation (just the square root of the variance) puts the units back to the units of X. Sometimes the sample variance is calculated with 1/(n-1) rather than 1/n. With large enough samples, the difference is small. For simplicity's sake, we will stick with the 1/n.

5) How does variance respond to changes in scale?

$$Var(a + bX_i) = b^2 Var(X_i)$$

If you move the bell curve over (displacement by a), the variance does not change. If you increase the X_i by a factor of b, the variance increases by b^2.

6) How about the Variance of the combination of two variables?

$$Var(X_i + Y_i) = Var(X_i) + Var(Y_i) + 2Cov(X_i, Y_i)$$
$$Var(X_i - Y_i) = Var(X_i) + Var(Y_i) - 2Cov(X_i, Y_i)$$

If X and Y are independent, then covariance(X,Y)=0, and life becomes simple and sweet. Variance of (X+Y) is simply Var(X)+ Var(Y). Also note that Var(X-Y)= Var(X)+Var(Y), because you could think of -Y as (-1)Y. If you take the distribution and move it to the negative numbers, the variance is still the same. Of course we could just calculate the covariance (it's not that hard). But most of the time it is simpler and cleaner to make the assumption of independence (and sometimes, it is even true!)

7) Standard Deviation

$$StDev(X_i) = \sqrt{Var(X_i)}$$

Standard Deviation is simply the square root of the variance. Standard deviation of X has the same units as X, whereas variance has squared units. When you want to know the standard deviation of the combination of two variables, the easiest thing to do is first calculate the variances, and then take the square root last.

8) Standard Error of the Mean

Usually in social statistics we are interested not only distribution of a population (let's say, the income of Nurses), but also in the mean and in the comparison of means (do nurses earn more than sociologists? How sure are we?)

So let's look at the variance and standard error of the mean. How sure are we about the mean earnings of nurses?

$$Var(Avg(X_i)) = Var(\frac{1}{n}\sum X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i)$$

Because $Var(bX_i)=b^2Var(X_i)$. Now we take advantage of the fact that the X's are independent, and identically distributed, so that the covariance between them is zero:

$$\frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(X_1 + X_2 + ...X_n) = (\frac{1}{n^2})nVar(X_i) = (\frac{1}{n})Var(X_i)$$

On the importance of sample size. Standard Deviation of the mean is usually called the Standard Error:

$$\text{Standard Error}=Stdev(Avg(X_i)) = \frac{\sqrt{Var(X_i)}}{\sqrt{n}}$$

What is new here is the factor of square root of n in the denominator. What this means is the larger the sample size, the smaller the standard error of the mean. High standard error means we are less sure of what we are trying to measure (in this case the average of X). Small standard error implies that we are more sure. Sample size is crucial in social statistics, but if you want a standard error half as large, you need a sample size 4 times as big (because of the square root). If you increase the sample size, the population variance of nurse's income stays the same, but the standard error of the mean of nurse's income decreases. It is important to keep in mind the difference between standard deviation of the population and the standard error of the mean.

9a) Now let's say we want to compare two means, X_i and Y_i, say they are the incomes of lawyers and nurses. The difference between the means is easy to calculate- it's just average income of the lawyers minus average income of the nurses. But what about the standard error of the difference? You take the standard errors of the individual means of X and Y, and you square them, to get the variance of the mean. Then you add them together to get the variance of the difference (because Var(X-Y)= Var(X)+ Var(Y) under independence), then you take the square root to get the standard error of the difference.

$$StdError(Avg(X_i)-Avg(Y_j))=\sqrt{(StdError(Avg(X_i)))^2+(StdError(Avg(Y_j)))}^2$$

$$\frac{Avg(X_i)-Avg(Y_j)}{StdError(Avg(X_i)-Avg(Y_j))}=\frac{\text{the difference}}{\text{standard error of that difference}}=T-Statistic$$

which we will compare to the Normal distribution or sometimes to the T distribution (a close and slightly fatter relative of the Normal distribution). The T-statistic is unit free, because it has units of X in the numerator and denominator, which cancel. The T-statistic is also immune to changes in scale. I know all the algebra looks a little daunting, but the idea is simple, and it is the basis for a lot of hypothesis testing. See my Excel sheet for an example.

Why does the average divided by its standard error take a Normal (or close to Normal) distribution? There is a famous theorem in statistics called the Central Limit Theorem which explains why. This Theorem requires a lot of advanced mathematics to prove, but the basic point is this: No matter what shape the distribution of the X's take- it could be flat, it could have three modes, etcetera, the mean of the X's approaches a Normal distribution as n grows large.

10a) The T-Statistic I describe above is the T-statistic which acknowleges that the variance and standard deviations of sample X and sample Y may be different. This is called the T-Test with unequal variances, and can be written this way (where n_x is the sample size of X, and n_y is the sample size of Y). Note that in the denominator we just have the square root of the sum of the variance of the mean of X and the variance of the mean of Y :

$$T-Statistic=\frac{Avg(X)-Avg(Y)}{\sqrt{\left(Var(X)/n_x\right)+\left(Var(Y)/n_Y\right)}}$$

10b) If we are comparing our mean to a constant, note that constants have variance of zero, so

$$T-Statistic=\frac{Avg(X)-const}{\sqrt{\left(Var(X)/n_x\right)}}=\frac{Avg(X)-const}{\text{Std Error(Avg(X))}}=\sqrt{n_x}\,\frac{Avg(X)-const}{\sqrt{(Var(X))}}$$

Our basic T-statistic is proportional to the square root of n, and takes n-1 degrees of freedom.

11) Although the T-statistic with unequal variance is the most intuitive, the more common T-statistic, which is also the T-statistic we will encounter in Ordinary Least Squares (OLS) regression is the T-Statistic which assumes equal variances in X and Y. The assumption of equal variance is called homoskedasticity. In fact, real data very frequently have heteroskedasticity, or unequal variances in different subsamples. The equal variance or homoskedastic T-Statistic is:

$$T-Statistic=\frac{Avg(X)-Avg(Y)}{\sqrt{\frac{(n_x-1)Var(X)+(n_Y-1)Var(Y)}{n_x+n_Y-2}}\sqrt{\left(\frac{1}{n_x}+\frac{1}{n_Y}\right)}}$$

You can show that these two formulas (T-statistic for unequal variance and T-statistic for equal variance) are the same when Var(X)=Var(Y). And note that in the equal variance T statistic, the denominator is the square root of the weighted sum of the variances of mean of X and mean of Y.

12) When looking up the T-statistic on a table or in Stata, you need to know not only the T-statistic but also the degrees of freedom, or the n of the test. For the equal variance T-statistic, the df of the test is n_x+n_y-2. The degrees of freedom for the unequal variance T-statistic is given by Satterthwaite's formula, and it is a little more complicated (you can look the formula up in the STATA documentation or online, but you don't need to know it…. For Satterthwaite's formula, that is for the df of the unequal variance T-test, if $(Var(X)/n_x)\approx(Var(Y)/n_y)$, meaning that the standard errors of our two means are similar, then for the unequal variance T-test df $\approx n_x+n_y$, which means the df is similar to the df we would get with the equal variance test (which makes sense since the standard errors are nearly equal, so the equal variance assumption is valid).

If $(Var(X)/n_x)>>(Var(Y)/n_y)$, then for the unequal variance T-test the df$\approx n_x$, because the $(Var(X)/n_x)$ will dominate the combined variance, and that means that the Xs are determining the combined variance, then the sample size of the Xs should determine our degrees of freedom for the T-statistic.

But don't worry too much about the degrees of freedom of the T-test! Big changes in the df of the T-test may not change the substantive outcome of the test- the T distribution changes with changes in df, but for df>10 the changes are fairly subtle. Even a change from 10 to 1000 df might not result in a different substantive answer. A T-statistic of 2.25 will correspond to a one-tail probability of 2.4% with 10 df (2 tail probability of 4.8%), while the same statistic of 2.25 would result in a one-tail probability of 1.2% on 1,000 df (2 tail probability of 2.4%). So for 10 df or 1,000 df or any higher number of df, a T-statistic of 2.25 yields a 2-tail probability of less than 5%, meaning we would reject the null hypothesis.

12.1) The Finite Population Correction and the Sampling Fraction

Above in Section 8, I defined the Standard Error of the mean this way:

Standard Error=$Stdev(Avg(X_i)) = \frac{\sqrt{Var(X_i)}}{\sqrt{n}}$

In fact, this definition leaves something out: the Finite Population Correction. A more accurate formula for the Standard Error of the mean is:

Standard Error=$Stdev(Avg(X_i)) = \frac{\sqrt{Var(X_i)}}{\sqrt{n}} \sqrt{\left[1 - \frac{n-1}{N-1}\right]}$

where n is the sample size of our sample (133,710 in the CPS), and N is the sample size of the universe that our sample is drawn from (274 million people in the US), n/N is the sampling fraction (about 1/2000 in the CPS), and $\left[1 - \frac{n-1}{N-1}\right]$ is the Finite Population Correction, or FPC. Note the following:

When n<<N, which is the situation we usually face, FPC≈1 which is why we usually ignore the FPC. Also note that when n<<N, and FPC≈1, it is only the small n (sample size of our sample) and not large N (size of the universe our sample was drawn from) that matters in the standard error formula. What this means, in practice, is that a 500 person opinion survey is just as accurate an instrument to test opinions in a small state like Maine as in a large state like California. As long as n<<N, we don't really care how big N is. When n is 500 and N is 100,000, FPC is 0.995. When n is 500 and N is 1,000,000, FPC is 0.9995. When n is 500 and N is 35,000,000, FPC is 0.999986. Generally we treat these FPC all as 1, and we ignore it.

When n=N, sampling fraction =1, and FPC=0. When n=N, the standard error of the mean is zero, which makes sense because if we have the entire sample universe measured in our hand, there is no statistical uncertainty left, and our measured mean is the true mean. When you have the whole sample universe in your data, for instance the votes of all 100 senators or the data from all 50 states, you can still run models, but you cannot treat the standard errors and probabilities that STATA reports as real probabilities describing uncertainty in what we know about the sample universe, since there is no uncertainty. We know what we know. When .01<sampling fraction<.9, then we have to think seriously about the FPC, or let STATA think seriously about it.

13) **Ordinary Least Squares regression, or OLS**.

OLS is the original kind of regression and the kind we will be dealing with in this class, is a method of comparing means.

$$Y = Const + B_1 X_1 + B_2 X_2 + ... + B_n X_n + residual$$

or

$$Y^{predicted} = Const + B_1 X_1 + B_2 X_2 + ... + B_n X_n$$

Where Y is the dependent variable, i.e. the thing we are trying to predict. It is important to note that the predicted values of the model will not in general equal the real values Y. The X's are the independent, or predictor variables. The B's are the coefficients for each variable which are produced by the regression. Each B will have a standard error and a resulting T-statistic to tell us whether the B is significantly different from zero or not. The residuals have a mean of zero. In theory, the residuals are supposed to be Normally distributed and well behaved, like pure noise. In reality, the residuals are only Normally distributed pure noise if the model fits the data very well. If the model fits poorly, the residuals can have all sorts of odd patterns.

The constant equals the average of $Y^{predicted}$ when all the X's are zero. Sometimes zero values for the X's make sense (i.e. correspond to real subgroups in the population). Sometimes zero values of the X's don't correspond to real subgroups in the population, and in that case the constant does not have a useful or substantive interpretation.

13.1) Correlation, r, and R-squared.

$$Cov(X,Y) = \frac{1}{n}\sum (X_i - Avg(X))(Y_i - Avg(Y))$$

and note that Cov(X,X)=Var(X).

r, the correlation coefficient, also known as Pearson's correlation, is defined this way:

$$r_{x,y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

Pearson's correlation r ranges from -1 (a perfect negative correlation) to 1 (a perfect positive correlation). When r=0, there is no linear relationship between X and Y.

With a single predictor X_1, and an OLS equation $Y=a+bX_1$, the regression line slope will be: $b = \frac{Cov(X_1,Y)}{Var(X)}$, so the slope of the regression line must be in same direction (positive or negative) as $Corr(X_1,Y)$, but the slope b is not bounded the way the correlation r is. Furthermore, the slope b can only be zero if r is zero. Once you know b, then $a = Avg(Y) - b(Avg(X_1))$

Now consider the following sums of squares from a generalized OLS regression model (which can have many predictor variables):

$$SS_{tot} = \sum (Y_i - Avg(Y))^2 = nVar(Y)$$
$$SS_{reg} = \sum (Y_i^{predicted} - Avg(Y))^2$$
$$SS_{res} = \sum residual_i^2$$

In OLS regression, $SS_{tot}=SS_{reg}+SS_{res}$

The R-square, or R^2, also known as the coefficient of determination, is defined:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}} = \frac{SS_{reg}/n}{SS_{tot}/n} = \frac{SS_{reg}/n}{Var(Y)}$$

R-square can be interpreted as the proportion of Var(Y) (the variance of the outcome variable) that is explained by the model. R-square varies between 0 and 1; R-square is 0 if the model explains none of Var(Y), and the value of R-square is 1 if the model fits the data perfectly, so that all residuals are zero. In practice, with models predicting social outcomes, R-square of 0.2 or 0.3 is often as good as you can expect to get with a reasonable model. It is also worth noting that a model with n predictor variables, that is one term for every observation in the dataset, could be designed to fit the data exactly, so that residuals would be zero and R-square would be 1. But if n is large (in our CPS dataset, n is typically in the tens of thousands), a model with n predictors would be useless to us because it would not have simplified the data at all.

In an OLS regression with one predictor variable, X_1, the R-square of the model $Y=a+bX_1$ is the square of the $Corr(X_1,Y)$, so that $R^2=(r)^2$. That is how R-square got its name. One thing to remember is that Cov(X,Y)=Cov(Y,X), and that Corr(X,Y)=Corr(Y,X), which means the R-square for the OLS regression $Y=aX_1+b$ will be the same as the R-square for the OLS regression $X_1=cY+d$. With regressions, however, usually only one of the models makes sense, we usually want the predictor variables to be logically or temporally prior to the dependent variable.

Units: Var(X) is in units of X^2. Std(X) is in units of X. Cov(X,Y) is in units of XY. Corr(X,Y) is unit-free. The regression line slope b is in units of Y/X. R-square is unit-free.

The more independent variables or predictors you put into the model, the closer your $Y^{predicted}$ should be to Y, and the lower SS_{reg} should be. New predictors cannot make the model fit worse; the new predictor would be exactly zero (and not change the predicted values at all) if the new predictor had no value at predicting Y at all (net of the other predictors already in the model). R-square increases monotonically as you add terms to nested models, so we need some other way to compare goodness of fit between models.

Note: Two models are nested if they contain the same set of observations, the same functional form, the same dependent variable, and one model's predictor variables are a subset of the second model's predictor variables. When comparing model fits in this class, we will want the models to be nested (though there are some who argue that Bayesian ways of comparing model fits, such as the BIC, don't require the models to be nested).

First approach: adjust the R-square for the number of terms in the model. The definition of the adjusted R-square is:

$$\text{Adjusted R-Square=}R^2-(1-R^2)\frac{k}{n-k-1}$$

where R^2 is the regular R-square statistic, k is the number of terms in the model (not counting the constant term), and n is the sample size of your dataset. The Adjusted R-square is not bounded by [0,1] the way R-square is. Adjusted R-square is adjusted because the adjusted R-square penalizes you for adding more terms to the model; the larger k is, the larger the penalty is. Higher values of adjusted R-square indicate better fit, but as far as I know there is no statistical test for the comparison of adjusted R-squares. You can simply examine the adjusted R-squares of nested models, and whichever model has the higher adjusted R-square is the better fitting model (by the metric of adjusted R-square). For regular (unadjusted) R-square, you would not want to simply compare the R-squares of nested models, because the R-square of the model with more terms would have to be higher. The regular unadjusted R-square statistics, when derived from OLS regression, on the other hand, can be compared with a statistical test, specifically the F-test:

$$F(m,n-k-1)=\frac{\left(R_B^2-R_A^2\right)/m}{\left(1-R_B^2\right)/(\text{n}-k-1)}$$

Where R_B^2 is the R-squared (regular R-squared rather than adjusted) from a larger model B, R_A^2 is the R-squared from a smaller model A nested within B, m, is the number of degrees of freedom difference between A and B, and k is the number of predictor variables in B (besides the constant), and n is the sample of subjects in both models. Because of some nice properties of OLS regression, this goodness of fit test is the same test as the F-test that the m new terms in model B are all jointly nonzero, and the test is easy to get Stata to perform after you run model B.

A very brief introduction to Logistic Regression

14) Ordinary Least Squares regression, OLS regression, has all sorts of nice features and is the original kind of regression, sort of like the Rose Bowl is the original college football bowl game. But there are some kinds of data that don't lend themselves easily to OLS regression. Take, for instance the case of dependent variable Y which is a Yes/No kind of variable, i.e. can be coded 0 or 1. There are lots of variables like this. If we try to run dependent variable Y through our OLS regressions, we might get predicted values of Y that were greater than 1 or less than 0, because Y is assumed to be Normally distributed, and could in theory take on any values. But greater than 1 or less than zero might be out of the acceptable range, so we need to impose a transformation on Y to keep all predicted values in the (0,1) range. The most common and popular transformation is the logistic transformation. The logit transformation looks like this:

$$Logit(Y^{\text{Pred}}) = Ln(\frac{Y^{\text{Pred}}}{1-Y^{\text{Pred}}})$$

The logit transformation covers the entire real numbers when predicted values of Y are between 0 and 1, which is what we want. Here Ln is the natural logarithm, that is the logarithm with base e (where e≈2.7183)

Logistic regression has a similar form on the right hand side:

$$\textit{(14.1)}\ \ Ln(\frac{Y^{\text{Pred}}}{1-Y^{\text{Pred}}}) = Const + B_1X_1 + B_2X_2 + ... + B_nX_n$$

It won't be visible to you, but logistic regression gets estimated recursively, whereas OLS regression gets solved directly.

Because we have transformed the left hand side of our regression equation, the coefficients on the right, the betas, need to be interpreted differently... If we exponentiate both sides

$$(\frac{Y^{\text{Pred}}}{1-Y^{\text{Pred}}}) = e^{(Const+B_1X_1+B_2X_2+...+B_nX_n)}$$

and then, because of a property of exponents that $e^{(a+b)}=e^ae^b$

$$(14.2)\ (\frac{Y^{\text{Pred}}}{1-Y^{\text{Pred}}}) = e^{Const}e^{B_1X_1}e^{B_2X_2}...e^{B_nX_n}$$

The left hand side is the odds of Y, So you can think of the exponentiated betas as factors in the odds. If we increase X_1 by 1, we increase the odds by a factor of e^{B_1}
Which means

$$(14.3) \qquad (\frac{Y^{\text{Pred}(X_1=\tilde{X}+1)}}{1-Y^{\text{Pred}(X_1=\tilde{X}+1)}}) = e^{Const}e^{B_1(X_1+1)}e^{B_2X_2} = e^{Const}e^{B_1X_1}e^{B_1}e^{B_2X_2}$$

The value of the predicted odds before incrementing X is:

$$(14.4) \qquad (\frac{Y^{\text{Pred}(X_1=\tilde{X})}}{1-Y^{\text{Pred}(X_1=\tilde{X})}}) = e^{Const}e^{B_1X_1}e^{B_2X_2}$$

So if we take the ratio of equations 14.3/14.4, we get that the exponentiated coefficient is actually an odds ratio, the ratio of the predicted odds with X_1=X+1 to the odds when X_1=X. If you think of X_1 as a categorical variable, for married (compared to unmarried) or for black (compared to white), then e^{B_1} is just the ratio of the predicted odds of the outcome variable for married (compared to unmarried) or for black (compared to white).

$$e^{B_1} = \frac{(\frac{Y^{\text{Pred}(X_1=\tilde{X}+1)}}{1-Y^{\text{Pred}(X_1=\tilde{X}+1)}})}{(\frac{Y^{\text{Pred}(X_1=\tilde{X})}}{1-Y^{\text{Pred}(X_1=\tilde{X})}})}$$

Which means e^{B_1} is an odds ratio, or the ratio of two odds.

In practice, we don't have to go through this algebra very much. Also, the logistic regression will produce coefficients with familiar Normal distributions and Z-statistics. In the coefficient (the un-exponentiated) version of logistic regression output, equation 14.1 above, the coefficients are Normally distributed with 95% confidence interval (coef-1.96SE, coef + 1.96 SE). As in OLS regression, the null hypothesis for coefficients is that the coefficient is zero. In the exponentiated version, equation 14.2 above, the null hypothesis is that the exponentiated coefficient=1, because e^0=1, and because 1 is the multiplicative identity. Adding zero doesn't change the predicted values in equation 14.1, and multiplying by 1 does not change the predicted values in equation 14.2.

Note that while the unexponentiated coefficients are nice and Normally distributed, with a nice symmetric 95% confidence interval with the coefficient in the middle, the exponentiated coefficient is not Normally distributed and its confidence interval is not symmetric any more. The exponentiated version of the confidence interval around the coefficient is
($e^{\text{coef-1.96SE}}$, $e^{\text{coef + 1.96 SE}}$), and since the exponentiation function increases large numbers more than it increases small numbers, the confidence interval is no longer symmetric around the coefficient. If the coefficient is 5 and its standard error is 1, the 95% CI for the coefficient would be approximately (3, 7). The coefficient would be significant because it is 5 standard errors greater than zero, so the Z-score would be 5. In exponentiated terms, e^5=148, and the 95% confidence interval would be (e^3, e^7), or (20,1097), which as you can see is not symmetric around 148.

It is important to keep in mind that an odds ratio of 2, for instance, does *NOT* mean that the predicted probabilities will be doubled. The whole point of the logit transformation is that the odds ratio can be as high as it wants without ever allowing the predicted probability of the outcome to be as high as 1 (or as low as zero). If we start with an odds ratio as the ratio of two predicted odds of a positive outcome,

$$\frac{\left(\frac{P_2}{1-P_2}\right)}{\left(\frac{P_1}{1-P_1}\right)} = e^{\beta}$$

And we solve this for P_2, (because we want to know how great the probability of success will be given starting probability P_1), we get:

$$P_2 = \frac{\left(\frac{P_1}{1-P_1}\right)e^{\beta}}{1+\left(\left(\frac{P_1}{1-P_1}\right)e^{\beta}\right)}$$

Let's say we go back to our equation 14.1, the (unexponentiated) coefficient version of loglistic regression:

$$Ln(\frac{Y^{\text{Pred}}}{1-Y^{\text{Pred}}}) = Const + B_1X_1 + B_2X_2 + ... + B_nX_n$$

The right side of the equation is very familiar, it is just a sum of coefficients multiplied by X values, with the constant added on. So how would we get to actual predicted values, that is Y^{Pred}? This is simply a matter of solving the above equation for Y^{Pred}. Let's start by saying that the total on the right side of the equation is W, and W can be any value, positive or negative. Then solving for Y^{Pred} we have:

$$Ln(\frac{Y^{\mathrm{Pred}}}{1-Y^{\mathrm{Pred}}}) = W$$

then exponentiate both sides,

$$(\frac{Y^{\mathrm{Pred}}}{1-Y^{\mathrm{Pred}}}) = e^{W}$$

and

$$Y^{\mathrm{Pred}} = e^{W}(1-Y^{\mathrm{Pred}})$$

and

$$Y^{\mathrm{Pred}} = e^{W} - e^{W}(Y^{\mathrm{Pred}})$$

and

$$Y^{\mathrm{Pred}} + e^{W}(Y^{\mathrm{Pred}}) = e^{W}$$

Factoring,

$$Y^{\mathrm{Pred}}(1+e^{W}) = e^{W}$$

and

$$Y^{\mathrm{Pred}} = \frac{e^{W}}{(1+e^{W})}$$

If you look at the right side of that equation, regardless of W’s value, e^{W} must be positive. And since 1+ e^{W} must be greater than e^{W}, we know that $0< Y^{\mathrm{Pred}}< 1$, which is what we want.

Some Comments on the Chisquare Distribution:

with n (integer) degrees of freedom,

$$f(x) = \frac{(1/2)^{n/2} x^{(n/2)-1} e^{-x/2}}{\Gamma(n/2)}$$

for $x \geq 0$

Mean=n
Variance=2n
$\text{Standard Deviation} = \sqrt{2\text{n}}$

a) The chisquare distribution has a domain of all positive real numbers, meaning $x \geq 0$. The Greek letter Γ in the denominator is just Gamma, indicating the Gamma function. The Gamma function is nothing more than a fancy way of extending the factorial function to all the real numbers. Γ(x)= (x-1)!, when x is an integer.

b) $\chi^2(1)$, or the Chisquare distribution with one degree of freedom is defined as the square of a Standard Normal variable. In other words, if z has the familiar Normal(0,1) distribution whose cumulative distribution is the source of tables in the back of every statistics text book (i.e. Normal with mean of zero and variance of 1), and if $y=z^2$, then y has a $\chi^2(1)$ distribution. This also means that if you have a statistic expressed a value from a $\chi^2(1)$ distribution, you can take the square root and you will have the familiar z-score. When switching back and forth from $\chi^2(1)$ and Normal(0,1) you do have to keep in mind that the Normal distribution has two tails (in the positive and negative directions), whereas as the Chisquare distribution only has the one tail.

c) Under independence, $\chi^2(a)+ \chi^2(b)= \chi^2(a+b)$. Another way to look at this is that $\chi^2(a)= \chi^2(1)+ \chi^2(1)+ \chi^2(1)+$.... a times (with each component being independent). Given what we know about the Central Limit Theorem, you would expect that $\chi^2(n)$ would look more and more like the Normal distribution, the larger n gets (since $\chi^2(n)$ is just n combinations of independent $\chi^2(1)$ variables). The examples of the chisquare distribution will verify that that $\chi^2(16)$ looks quite Normal (and in this case it approximates Normal(16,32)).

d) One property of the Chisquare distribution that is relevant to evaluating the relative fit of different models that are fit by likelihood maximization (such as logistic regression), is that if Model 1 has -2logLikelihood, or -2LL=V, and Model 2 adds m additional terms and has -2LL=U, then the comparison of Model 1 and Model 2 is $\chi^2(m)$=V-U. This comparison only works if Model 1 is nested within Model 2 (that is, if Model 2 contains Model 1). This is the Likelihood Ratio Test, or LRT, which I describe in a little more detail below.

e) The F-distribution is defined as a ratio of two chisquare distributions. If P is distributed as $\chi^2(m)$ and Q is independent from P and is distributed as $\chi^2(n)$, and $W = \frac{P/m}{Q/n}$, then W is has the $F_{m,n}$ distribution, that is the F distribution with m and n degrees of freedom.

Likelihood and the Likelihood Ratio Test:

Let's say that θ is the vector of all the parameters in our model, the several right-sided predictor variables we are trying to find the best values for.

The likelihood of a model is the joint point probability that every data point in our sample would have their exact values of Y_i

$$\text{Likelihood}(\theta) = f(y_1, y_2, y_{3,}...., y_n \mid \theta)$$

and

$$\text{Likelihood}(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

There are a few things to know about the likelihood. First of all, the point probability p of any single exact outcome value, even given the best fitting model, will almost always be $0<p<1$ (although probability density functions can yield point values of equal to or greater than 1 under some unusual circumstances, think of the uniform distribution, or the Chisquare(1) distribution for small values of x). If n is at all substantial in size (as it almost always is in the datasets we use like the CPS), then the product of the n (already very small) likelihoods is going to be phenomenally small- infinitesimal. We don't really care about the value of the likelihood of the model, except that Stata will choose θ to maximize the likelihood. We don't care about the value of the likelihood itself, but we do care about the comparison of the likelihoods of two nested models.

First things first: If we want to find the value (or values) of θ that will maximize the joint likelihood, we don't want to deal with an enormous product of probabilities, because products of functions are difficult to differentiate. What we do instead is we take the natural log of the likelihood, the log likelihood:

$$\text{log likelihood } l(\theta) = \sum_{i=1}^{n} \ln(f(y_i \mid \theta))$$

Because the logarithm is a function that is monotonic, that is it has no maxima or minima of its own, the θ that maximizes the log likelihood is the same θ that maximizes the likelihood, and the log likelihood almost always is easier to work with.

Likelihoods are a kind of probability, so in general $0<\text{likelihood}(\theta)<1$. Log likelihood, or $l(\theta)$ is going to be negative, because it is a sum of negative elements. Every time we add new variables to our model (keeping the same set of data points), the likelihood will be larger (or the same if the added terms are all zero), and the log likelihood will be greater (meaning less negative, i.e. closer to zero). The -2LL, or minus two times the log likelihood will be positive, and will get smaller and smaller with each new term that improves the fit of the model. The likelihood ratio test is the a test of the ratio of the likelihoods of two nested models, which we convert into the difference of the -2LL of the two nested models (because it is always easier to work with the log likelihood than to work with the likelihood itself), and the difference of the -2LLs is chisquare with m degrees of freedom, where the larger model has m additional terms beyond what the smaller model has.

The Normal and T distributions:

The probability density function of the Normal distribution is defined this way:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

for $-\infty<x<\infty$

Expected value (or mean), E (X)=μ

Variance (X)=σ^2

You can see from the distribution that it is symmetrical around its mean.

If Z is distributed as Normal(0,1)- the standard Normal, and U is distributed as $\chi^2(n)$, and U and Z are independent, and if $T=\frac{Z}{\sqrt{U/n}}$, then T is distributed as the t-distribution, with n degrees of freedom.

OLS as an example of Maximum Likelihood:

Recall the functional form of our Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The Maximum Likelihood estimates for our linear regression will be OLS, that is Ordinary Least Squares estimates if the errors are Normally distributed, i.e. if the errors are Gaussian (Gauss being the mathematician credited with describing the Normal distribution). To get Stata's glm function to assume that the errors are Normally distributed, you specify the option family(gaussian).

Now, if we assume that the errors are Normally distributed, what would the Likelihood function for the B_0 and B_1 look like, in our regression line of $Y_i = B_0 + B_1 X_i + residual$

$$L(Y_i \mid B_0, B_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(Y_i - B_0 - B_1 X_i)^2}{2\sigma^2}}$$

Where B_0 is simply our constant term, and $(Y_i - B_0 - B_1 X_i)^2$ is the square of each residual. Note that the errors of the true regression model and the residuals of the actual regression model we use are not the same, but the residuals are what we have to estimate the errors with.

We don't much care here about how σ is going to be estimated, because σ will be estimated by the square root of the sample variance of our data (with a minor correction that need not concern us here). Here we are interested in the estimates of B_0 and B_1 that will maximize the Likelihood, so let us press on a bit further. Remember that $e^{(a+b+c)} = e^a e^b e^c$. We can get rid of the Product notation, and convert the exponential function to an exponential of a sum.

$$L(Y_i \mid B_0, B_1, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{\left(\frac{-1}{2\sigma^2} \right) \sum_{i=1}^{n} (Y_i - B_0 - B_1 X_i)^2}$$

I am not going to put you through the calculus of finding the maximum here, but the story should be clear from looking at the likelihood function. The sum in the above equation is the sum of squared residuals. In order to maximize the likelihood, we need to maximize the exponential function. In order to maximize the exponential function, which is of the form $e^{-f(x)}$ where f(x)≥0, we need to minimize f(x). In other words, the Maximum Likelihood solution for B_0 and B_1 is the solution which minimizes the sum of squared residuals, which is the Least Squares solution. OLS is the MLE solution if we assume that the errors, and therefore the residuals are Normally distributed. The next question you should ask is: Is it reasonable to assume that the residuals will be Normally distributed? If the model is the true model, so that the data are perfectly predicted by the model except for noise, then of course the residuals will be Normal. If the model is not the true model (and with real life data, there are no "true" models), then the Normality assumption might not be valid. Plotting the residuals is a useful exercise, and one often will find that the residuals are very non-Normal. On the other hand, there is statistical theory showing that OLS yields the Best Linear Unbiased Estimates (BLUE) for the coefficients, even if the errors are not Normally distributed.

On Independence:

One way you will sometimes see the Chisquare distribution invoked is for tests of independence between two variables. What does Independence mean in this context?

If Variables X and Y are Independent, then P(X∩Y)= P(X)P(Y).

The ∩ just means intersection. You can read the above statement as "The probability of X and Y both happening is the product of the probability of X multiplied by the probability of Y." Let's say X is the probability of husbands being black, and Y is the probability of wives being black. Let's further say that P(X)=P(Y)=0.1, because blacks are about 10% of the US population. So does that mean that P(X∩Y)= P(X)P(Y)=0.1(0.1)=0.01? NO. In fact in this case P(X∩Y)=0.09≠0.01 because husband's race and wife's race are not independent from each other: there is a strong pattern of selective mating by race.

If we compare crosstabulated data from two models, we can generate the chisquare statistic two ways. Here n_{ij} are the cell counts from one model (or the actual data), and u_{ij} are the cell counts from the other model, or in our case the cell counts from predicted values :

Pearson Chisquare or $X^2 = \sum \frac{(n_{ij} - u_{ij})^2}{u_{ij}}$

Likelihood Ratio Chisquare or $G^2 = 2\sum n_{ij} Ln(\frac{n_{ij}}{u_{ij}})$

In Chisquare tests for independence, the data itself has RC df, and the independence model has R+C-1 df, so the Chisquare test for independence is on RC-(R+C-1) df.