Statistical Functions in Excel

- There are many statistical functions in Excel. Moreover, there are other functions that are not specified as statistical functions that are helpful in some statistical analyses. The rest of these notes provides several lists of functions with short descriptions of how to use them.

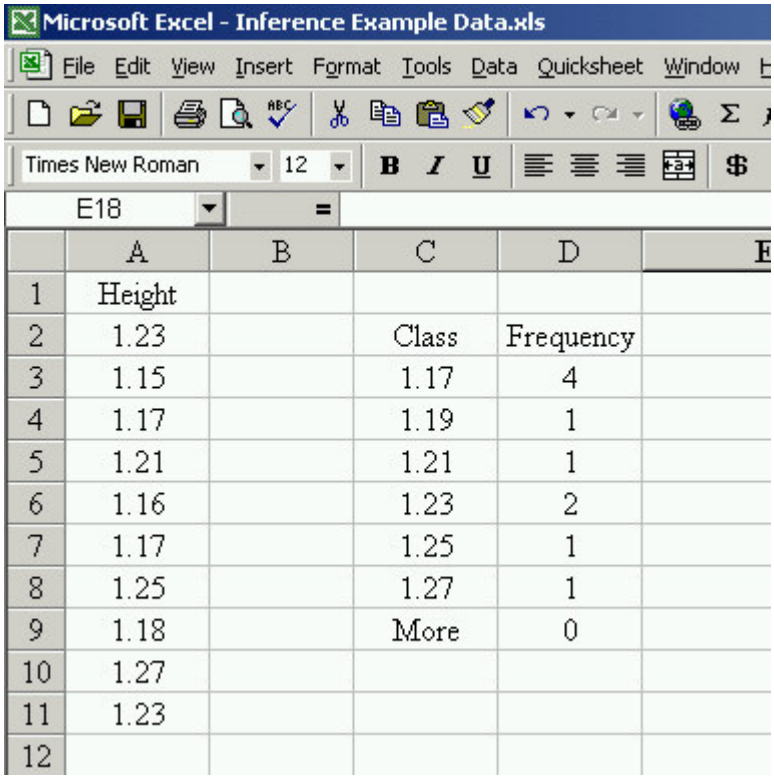
Summarizing Data Functions

- The table below shows the spreadsheet functions that can be used to find the most common measures of central location of a data set. The entry data_range means that you would enter the range of cells containing the data you want to analyze. For example, if you had data in cells A2:A51, data_range would be A2:A51. If you prefer, you can give names to cell ranges, and then use the name in place of the cell range. Finally, you can enter the numbers directly but I don't recommend it. If you choose to do so, the numbers should be separated by commas. For example, I could use =average(10,11,14,18,22).

Measure	Function
Mean	=average(data_range)
Median	=median(data_range)
Mode	=mode(data_range)
p^{th} Percentile	=percentile(data_range, p/100)
1^{st} Quartile	=quartile(data_range, 1)
3^{rd} Quartile	=quartile(data_range, 3)
Range	=max(data_range) - min(data_range)
Interquartile Range	=quartile(data_range,3) - quartile(data_range,1)
Population Variance	=varp(data_range)
Sample Variance	=var(data_range)
Population Std. Dev.	=stdevp(data_range)

Measure	Function
Sample Std. Dev.	=stdev(data_range)
Population Covariance	=covar(data_range1, data_range 2)
Sample Correlation	=correl(data_range1, data_range 2).
Frequency Distribution	=frequency(data_range,bin_range)

- The FREQUENCY function requires a little more explanation. First, it is an array function so you must use the rules that apply to array functions. Next, the bin_range argument is a range of cells containing the class interval boundaries that you want to use. In the example below, I highlighted the range D3:D9, typed in the function =FREQUENCY(A2:A11,C3:C8), and then used CNTRL-SHIFT-ENTER. The results show that there are 4 numbers in column A that are in (-∞,1.17], 1 that is in (1.17,1.19], etc. The 0 at the end indicates that none are bigger than 1.27. Note that I typed in "More" at the end of the bin values, but I did not use its cell reference in the second argument of the function.



	A	B	C	D	E
1	Height				
2	1.23		Class	Frequency	
3	1.15		1.17	4	
4	1.17		1.19	1	
5	1.21		1.21	1	
6	1.16		1.23	2	
7	1.17		1.25	1	
8	1.25		1.27	1	
9	1.18		More	0	
10	1.27				
11	1.23				
12					

Probability Distribution Functions

- There are several functions for common probability distributions. I have listed below those that are encountered in most introductory statistics classes.

  Discrete Distributions:

  Arbitrary: If we have a given discrete random variable X with probability mass function f(x), we can find the probability P($a \le X \le b$) using

  =PROB(x_range, f(x)_range, *a*, *b*)

  where x_range is the cell range containing the possible values of X, f(x)_range is the range of cells containing f(x).

- When finding the expected value and variance of discrete random variables, the computation requires the sum of products. A useful function is =SUMPRODUCT.

- As an example, suppose that we had the following discrete distribution (including labels) entered in cells A1:B6:

x	f(x)
0	.1
1	.2
2	.2
3	.3
4	.2

- To find the P($1 \le X \le 3$) we would use =PROB(A2:A6,B2:B6,1,3). The result would be 7.

- To find the expected value of X we would use =SUMPRODUCT(A2:A6,B2:B6). The result is 2.3.

- To find the variance we would need to insert another column which would be $(x-\mu)^2$ (so in cell C2 we would enter =(A2-2.3)^2 and we would copy it down to cell C6). Then we would enter =SUMPRODUCT(C2:C6,B2:B6) to get the variance. The result is 1.61.

  Binomial: Concerns the number of successes *x* in *n* trials.

  =BINOMDIST(*x*,*n*,*p*,I)

  Finds P(X=*x*) when I=0 P(X≤*x*) when I=1. The other parameters are *n* the sample size and *p* the probability of success in one trial.

  Poisson: Concerns the number of occurrences *x* in a given time or space.

  =POISSON(*x*, μ, I)

  Finds P(X=*x*) when I=0 P(X≤*x*) when I=1. The other parameter is μ the average *rate* of occurrences in the given time or space.

  Hypergeometric: Concerns the number of successes *x* in a sample of size *n* drawn from a population of size *N*, and in the population there are *r* items that we would consider successes.

  =HYPGEOMDIST(*x*, *n*, *r*, *N*)

Related Functions:

Counting Functions

- Here are some functions related to finding probabilities when counting.

Counting Method	Symbol	Function
Factorial	$n!$	=fact(n)
Combination	${}_NC_n = \binom{N}{n}$	=combin(N,n)
Permutations	${}_NP_n$	=permut(N,n)

Continuous Distributions:

- Excel has two types of functions for continuous distributions. The first type generally ends with "dist" and finds tail areas for a given value of the random variable. The second type generally ends with "inv" and finds the inverse function, i.e., the value of the random variable for a given tail area.

Standard Normal: (assumes a mean of 0 and standard deviation of 1)

=NORMSDIST(z)

Finds the area to the *left* of a standardized value z.

=NORMSDIST(A)

Finds the value of z that gives a *left* tail area A.

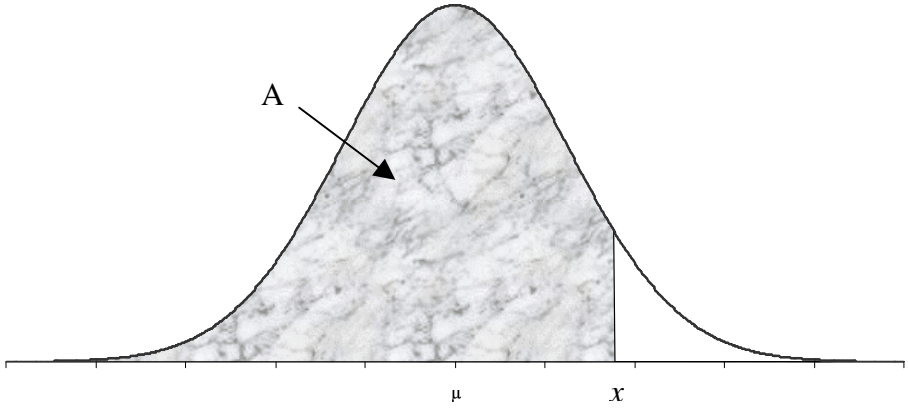
Normal:

=NORMDIST(x, μ, σ, 1)

Finds the area to the *left* of x for a mean of μ and a standard deviation σ. If the 1 in the last argument is changed to 0, this function finds the height of the normal bell curve (pdf).

=NORMSDIST(A, μ, σ)

Finds the value of *x* that gives a *left* tail area A.



Exponential: There is no inverse function in Excel for the exponential distribution.

=EXPONDIST(*x*,μ,1).

Finds the area to the *left* of *x* for a mean *rate* of μ (in other words, the units on μ are the reciprocal of the units on *x*).

Student t: Unlike previous functions, those for the t-distribution utilize the right rather than the left tail area.

=TDIST(*T*, *df*, *tails*)

Finds the area to the *right* of *T* for for *df* degrees of freedom. If *tails* is 1, then it finds one tail area. If *tails* is 2, it finds the two-tailed area (which is the one-tailed area multiplied by 2). In this function, *T* must be positive.

=TINV(A, *df*)

This function always assumes two tails! It finds what value of *T* would give an area to the right of *T* equal to A/2 when there are *df* degrees of freedom. This is a useful feature when we want to find confidence intervals and do two-tailed hypothesis tests. For example, for a (1-α)100% confidence interval on the mean, we use $\bar{X} + t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$. We can use the TINV function to obtain $t_{\alpha/2,\, n-1}$. The function would be =TINV(α,*n-1*).

F: These functions also utilize the right tail area.

=FDIST(*F*, *ndf*, *ddf*)

Finds the area to the *right* of *F* for for *ndf* numerator degrees of freedom and *ddf* denominator degrees of freedom.

=FINV(A, *ndf*, *ddf*)

Finds the value of *F* that would give an area to the right of *F* equal to A when there are *ndf* numerator degrees of freedom and *ddf* denominator degrees of freedom.

Chi-squared (χ^2): These functions also utilize the right tail area.

=CHIDIST(χ^2, *df*)

Finds the area to the *right* of χ^2 for *df* degrees of freedom.

= CHIINV(A, *df*)

Finds the value of χ^2 that would give an area to the right of χ^2 equal to A when there are *df* degrees of freedom.

Others: There are also functions for the Beta, Gamma and Lognormal distributions that are very similar to the functions above.

Random Samples

- To make a random selection, we can number each member of the population to be sampled from 1 to *N* (the size of the population). Then to determine which will be sampled, we can use the function RANDBETWEEN(1,*N*) and then copy it down to reach the desired sample size.

- This function is interesting, because if you do anything to the worksheet (e.g., when you copy the formula down), the value recalculates. That is a useful property for doing simulation, but it is annoying when trying to do other things. Once you have copied the function down to the desired sample size, I suggest "freezing" the numbers by copying them, selecting Edit/Paste Special, and choosing values. That will remove the function and just leave the values.

- The RANDBETWEEN function does sampling with replacement and hence duplicates are possible. The best way to check for duplicates is to sort the data. I suggest using either the RANK function or the SMALL function to do so. I will describe the SMALL function, which sorts. Suppose I had my random numbers in cells C1:C25. To dynamically sort them, I would enter the numbers 1 to 25 in cells D1:D25. Then in cell E1 I would type

  =SMALL(C1:C25,D1)

  and copy it down through E25.

  The function says to look for the D1st smallest number in cells C1:C25. Since D1 =1, it finds the very smallest number in the range. When the function is copied down, it looks for the 2nd smallest and puts it in E2, etc.

- Once the data are sorted, you can look at neighbors for duplicates. I suggest using the =IF function to do so. In the example above I would use

  =IF(E1=E2,1,0)

  in cell F1. Then I would copy it down through F24 and look at the sum of the results. If the sum is 0, there are no duplicates.

Inference Functions

- Most of the functions needed to do statistical inference have already been introduced. For example, to form a confidence interval on a population proportion, we generally use $z_{\alpha/2}$. To get its value in Excel, we would use =normsinv(1-α/2) or =-normsinv(α/2). (Remember that the area used as an input in the NORMSINV function is the left area, but for $z_{\alpha/2}$ we need a right area.) Below I have put some of these functions in the context of inference, and have also introduced a few other functions unique to inference.

Confidence Intervals:

- For the case of estimating a population mean when we know the population standard deviation σ, we can use the function

  =CONFIDENCE(α, σ, *n*).

  The result of the function is the margin of error.

Hypothesis Tests on One Parameter:

Statistic	Tails	Function for CR	Function for p-value
$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$	1	=normsinv(1-α)	=1-normsdist(abs(Z))
	2	=normsinv(1-α/2)	=2*(1-normsdist(abs(Z))
$T = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$	1	=tinv(2α, n-1)	=tdist(abs(T), n-1,1)
	2	=tinv(α, n-1)	=tdist(abs(T), n-1, 2)
$Z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	1	=normsinv(1-α)	=1-normsdist(abs(Z))
	2	=normsinv(1-α/2)	=2*(1-normsdist(abs(Z))

In the equations above the values with subscript 0 are the hypothesized values. For these tests, compare the *absolute value* of the computed test statistic to the CR value, or compare the p-value to α.

- If we have the data given for the first two cases above, there are built-in functions in Excel to give us the p-value. For the case when we are testing the hypothesis on one mean and we have the population standard deviation given, we can use

  =ztest(data_range, μ_0, σ)

  The result of this function is the one-sided p-value. *NOTE* that this is different than what the Excel documentation says. The documentation mistakenly says that it is the two-sided p-value.

- If the population standard deviation were not known, the function would be

  =TTEST(data_range, μ_0_range, *tails*,1)

  Here the μ_0_range is a column of values all equal to the hypothesized mean with the same size as the column of data (i.e., the same length as data_range) and *tails* is 1 if it is a 1-sided test and 2 if it is a 2-sided test. The last argument tails Excel what type of t-test to do (see below).

  Testing Two Means:

- In the case of comparing two means, we can also use the TTEST function. There are only a few changes. The function is

  =TTEST(data_range1, data_range2, *tails*, *type*).

  Here the data_range values are the cell references for the data from the two samples (one from each population). The *tails* argument is the same as above (1 if it is a 1-sided test and 2 if it is a 2-sided test). The *type* argument tells Excel what kind of t-test is being done.

  type=1: Paired (or matched) sample t-test
  type=2: Unmatched 2-sample t-test assuming equal variances
  type=3: Unmatched 2-sample t-test assuming unequal variances

  Testing Two Variances:

  =FTEST(data_range1, data_range2)

  The result of this function is a 2-sided p-value. If you want a one-sided test, you need to divide the result by 2.

Chi-squared(categorical) tests:

=CHITEST(observed_range, expected_range)

where observed_range is the range of cells containing the observed number in each category, and expected_range is the range of cells containing the expected number in each category. The result will be the p-value.

- Be sure that all categories have at least 5 expected observations before you use this function.

- Note that the chi-squared test for independence is equivalent to testing the equality of two population proportions.

ANOVA

There are no built in functions in Excel to do ANOVA. However, the multiple regression function described below can be used to obtain the results. You can also use the Analysis Tools to do it. If you want to use the function, you should create one column containing the observed values from all groups. Call it y (it is the dependent variable). If there are a total of *p* groups, then you would need to create *p*-1 independent variables. If the value of y comes from group 1, then x_1 would be 1 otherwise it would be 0. The other independent variables would have the same coding. If an observation came from group *p*, all of the x values would be equal to 0. Then follow the instructions for the LINEST function described below, being sure that the *stats* argument is equal to 1. The resulting F statistic would be used to test the hypothesis that all means are equal.

Regression:

- Simple Regression:

Statistic	Function
Intercept	=intercept(y_range, x_range)
Slope	=slope(y_range, x_range)
R^2	=rsq(y_range, x_range)
s_{ylx}	=steyx(y_range, x_range)

In all cases y_range is the range of cells containing the dependent variable and x_range is the range of cells containing the independent variable.

The last statistic, s_{ylx}, is the standard error of the y given x. In other words, it is the estimate of the standard deviation in the y-direction about the line. It is also equal to $\sqrt{MSE}$.

- Multiple Regression:

=linest(y_range,x_range,*intercept*,*stats*)

- This function calculates several statistics associated with a multiple regression, depending on the value of the argument *stats*. Here y_range is the range of cells containing the dependent variable and x_range is the range of cells containing *all* of the independent variable. Therefore, all independent variables must be in a contiguous set of cells.

- The argument *intercept* is usually omitted. If you make it equal to 0 or FALSE, the regression will force the intercept to be 0. The argument *stats* is equal to 0 or FALSE if you only want the estimates of the slopes and intercept; it is 1 or TRUE if you want see several other statistics.

- The function =LINEST is an array function. The number of cells that you highlight depends on whether *stats* is 0 or 1. If it is 0, you need to highlight one row whose length is equal to the number of parameters being estimated (the number of independent variables+1). If *stats* is 1, you need to highlight 5 rows and the number of columns will again be the number of independent variables+1.

- The coefficients are listed with the last slope first, then the second-to-last slope, etc., until reaching the first slope followed by the intercept.

- When *stats* is 1, the values in the resulting matrix are as follows:

b_k	b_{k-1}	…	b_1	b_0
s_{b_k}	$s_{b_{k-1}}$	…	s_{b_1}	s_{b_0}
R^2	$s_{y\|x}$			
F	n-k-1			
SSR	SSE			

In the above, I am assuming that there are k independent variables. The b values are the coefficients (b_0 is the intercept), the s_b values are the standard errors of the coefficients, F is the F-statistic calculated in the ANOVA section of the regression analysis (and used to test if the overall relationship is significant), SSR is the regression (or explained) sums of squares, and SSE is the error (or unexplained) sums of squares. From the above numbers most of the other basic numbers from a regression output can be calculated. For example, to get MSE I would divide SSE by n-k-1. To get the t-statistic for slope k, I would use b_k / s_{b_k} .

Matrix Functions

- Many statistical computations require matrix manipulations. Hence matrix functions are quite useful. Below I list the matrix functions in Excel. The arguments are arrays of cells in Excel. Except for MDETERM, these are all array functions.

Function	Purpose
=TRANSPOSE(matrix)	Transposes the matrix
=MINVERSE(matrix)	Inverts a square matrix
=MDETERM(matrix)	Finds the determinant of a square matrix
=MMULT(matrix1,matrix2)	Multiplies the two matrices

Other Useful Functions

- There are several other functions that I use often. I will just list them and say what they do, but I will not describe them in detail here. Instead I will let you learn about them on your own.

Function	Purpose
=SQRT	Finds the square root of the argument
=STANDARDIZE	Standardizes a value by subtracting the mean and dividing by the standard deviation
=COUNT	Counts the number of non-blank and non-text cell entries
=IF	Does if/then computations
=COUNTIF	Counts values based on a particular criterion
=SUMIF	Sums values based on a particular criterion