Statistics Formula Glossary

Please keep in mind that this document changes as we add more to the UB online math resources; the copy you download today may be different from the copy available tomorrow. This is by no means a comprehensive list of all possible statistical formulas, but instead a list which may be representative of your needs as a student taking a statistics course at UB.

Binomial Distribution

The binomial distribution is a type of mathematical model. Mathematical models allow us to easily calculate the probability of occurrence of any specific value of the variable of interest. The binomial distribution is used in situations where the discrete variable is the number of occurrences in a sample of n observations.

$$P\ (X = x \mid n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

And

P(X = x | n,p) = probability that X = x events of interest, where n and p are as follows:

n = number of observations

p = probability of an event of interest (prob.of success)

I - p = q = probability of not having an event of interest (prob. of failure)

x = number of events of interest (no. of successes) in the sample (X = 0,1,2, …, n)

Standard Deviation for BD: $\sqrt{Var(X)} = \sqrt{np(1-p}$

Chi-Square Statistic

This kind of test is used to determine whether there is a significant difference between expected frequencies (in one or more categories) and the actual observed frequencies. The following formula poses Fo as the observed value and Fe as the expected value.

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Cohen's d

Effect size is used to quantify the differences between two groups, rather than just stating the probability that the differences between two groups is just a fluke. Cohen's d is a measure of effect size used when there are two means to compare, like with a t-test. When working with a single sample t test, for the numerator, you need to find the difference between the mean of the sample and the mean of the population. For independent measures t-tests, you subtract the two sample means from one another in the numerator, and you divide by the square root of the pooled variance (the equation for which is in this glossary) If you already know the t-value, there's another equation. If you're working with a repeated-measures t-test, simply divide the mean difference of the two groups by the standard deviation.

Cohen's d in theory: $d = \frac{mean\ difference}{standard\ deviation}$

Cohen's d for Single Sample t test: $d = \frac{M-\mu}{s}$

Estimated Cohen's d for Independent Measures t-test: $d = \frac{M_1 - M_2}{\sqrt{s_p^2}}$ or $d = t\frac{\sqrt{N_1+N_2}}{N_1 N_2}$

Estimated Cohens' d for Repeated-Measure t-test: $d = \frac{M_D}{s}$

Coefficient of Determination (r^2)

This is pretty straightforward to calculate. Once you've calculated *r* or your Pearson Correlation number, you only need to square it.

Combinations

The number of combinations of selecting x objects out of n objects.

$$ {}_nC_x = \frac{n!}{x!\,(n-x)!} \; where\; n! = (n)(n-1)\dots(1)\; and\; 0! = 1 $$

Degrees of Freedom

Single Samples t-test: $n - 1$

Independent Samples t-test: $df = df_1 + df_2 = (n_1 - 1) + (n_2 - 1)$

Estimation (Confidence Intervals)

Single sample t-test: $\bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right)$

Single sample t-test with n>30: $\bar{X} \pm z\left(\frac{s}{\sqrt{n}}\right)$

Independent samples t-test: $(\bar{x}_1 - \bar{x}_2) \pm tS_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Repeated Measures t-test: $\mu_D = M_D \pm ts_{M_D}$

Estimated Standard Error

This formula is used as a second step when needing to find the t-statistic for an individual samples t-test. The first step would be to calculate the pooled variance.

$$ S_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n^2}} $$

Independent-Measures ANOVA

Solving an ANOVA by hand can be tedious and there are many steps, so it's important to do everything in the correct order. First and foremost, you need to find **n, M, T,** and **SS** for every level of your independent variable (aka group). It's best to write these out in a table of sorts. For n, you want to find the number of participants in each group (so if you have 3 groups, you need 3 n's). M is the mean of each group, T is the total score of each group, and SS is the sum of squares of each group (remember that that's the mean subtracted from each score and then squared all added up). Next, you need to find **k, ΣX^2, G,** and **N** and write them in the margins of your table. k is the number of levels in your independent variable. For ΣX^2, you need to square each individual score in the data set and then add them all together. G stands for Grand Total, which means to get G, you just need to add all the T's together. Finally, N is the total number of scores, so all the n's added together. The second part of all this is finding SS_{total}, SS_{within}, and $SS_{Between}$. The formulas for each are as follows:

$SS_{within} = \Sigma SS_{inside\ each\ treatment}$

$SS_{Between} = SS_{total} - SS_{within}$

$SS_{total} = \Sigma X^2 - \frac{G^2}{N}$

The third part is to find df_{total}, df_{within}, and $df_{Between}$. The formulas for each are as follows:

$$df_{total} = N - 1$$
$$df_{within} = N - k$$
$$df_{between} = k - 1$$

Finding df_{total} is a way to make sure your math is correct for the other df's, since adding the other two should equal the total. The fourth part is to find the mean of the squared deviations, or MS for short. For this we need to find $MS_{between}$ and MS_{Within}.

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$
$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

The final step is to solve for the F-value.

$$F = \frac{MS_{between}}{MS_{within}}$$

If it turns out that F falls within the critical region, then it's time to run a post-hoc.

Mean

This measure of central tendency is often referred to as the "average." It's found by adding up all the scores of one set and dividing that number by the number of scores.

Population: $\mu = \frac{\sum X}{N}$

Sample: $M = \frac{\sum X}{n}$

Binomial Distribution: $\mu = E(X) = n\pi$

Median

When working with raw scores, the easiest way to find the median is to make a cumulative frequency table, divide the sum of all the frequencies by 2, and then the number which has a cumulative frequency (meaning the number's frequencies plus all of the frequencies before it) of that number or just above that number is the median. When working with grouped or continuous scores, you make a cumulative frequency table, divide the total frequencies by 2, and use the equation below, where **L** is the lower class boundary of the group containing the median, **n** is the total number of scores, **B** is the cumulative frequency of the groups before the median group, **G** is the frequency of the median group, and **w** is the group width. Remember the order of operations!

Estimated Median of Grouped/Continuous Frequencies: $L + \frac{(n-2)-B}{G} * w$

Partial Eta Squared (η^2)

Independent One-Way ANOVA: $\eta^2 = \frac{SS_{between\ treatments}}{SS_{total}}$

Repeated Measures One-Way ANOVA: $\eta^2 = \frac{SS_{between\ treatments}}{SS_{total} - SS_{between\ subjects}} = \frac{SS_{between\ treatments}}{SS_{error}}$

Two-Way ANOVA:

$$\eta_A^2 = \frac{SS_A}{SS_{total} - SS_B - SS_{AxB}} = \frac{SS_A}{SS_A + SS_{within\ treatments}}$$

$$\eta_B^2 = \frac{SS_B}{SS_{total} - SS_A - SS_{AxB}} = \frac{SS_B}{SS_B + SS_{within\ treatments}}$$

$$\eta_{AxB}^2 = \frac{SS_{AxB}}{SS_{total} - SS_A - SS_B} = \frac{SS_{AxB}}{SS_{AxB} + SS_{within\ treatments}}$$

Pearson Correlation (r)

Although it's easy to calculate a correlation on SPSS, it's a little more involved when calculating it by hand. It's not *difficult*, per say, but very similarly to calculating standard deviation by hand, one needs to use tables. Follow this link to watch someone do this step by step: https://www.youtube.com/watch?v=2SCg8Kuh0tE. The formula itself is as follows:

$$r = \frac{\sum\left((x - \bar{x})(y - \bar{y})\right)}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Poisson Distribution

The Poisson distribution is a type of mathematical model. The Poisson distribution applies when we want to determine the number of occurrences of a particular event in some fixed interval of time and space. This fixed interval of time and space is often called an area of opportunity. Within the area of opportunity, there can be multiple occurrences of an event.

$$P\ (X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x!}{x!}$$

where

P (X = x | λ) = probability that X = x events in an area of opportunity given λ

λ = expected number of events per unit

e = mathematical constant approximated by 2.71828

x = number of events (x = 0,1,2, …, n)

Pooled Variance

Pooled variance is a variable one has to find in order to be able to solve for an independent variables t-statistic. Before you can do this step, be sure to calculate the sum of squares (SS) for each sample.

$$s_p^2 = \frac{(SS_1 + SS_2)}{df_1 + df_2}$$

Proportion

$$p = \frac{X}{n} = \frac{Number\ of\ items\ having\ the\ characteristics\ of\ interest}{Sample\ Size}$$

Regression

The point of doing a regression is to come up with a prediction line for the data you have: Y=bX+a. To calculate b, you need to know the standard deviation of Y and of X. To calculate a, you need to know the mean of Y and the mean of X.

The other formulas in this section are about testing the significance of a regression. For all variability formulas related to regression, look under standard error of estimate.

$$\text{Slope: } b = r\left(\frac{S_y}{S_x}\right)$$

$$\text{Y-intercept: } a = \bar{y} - b\bar{x}$$

$$\text{MS Regression: } \frac{SS_{regression}}{df_{regression}}$$

$$\text{MS Residual: } \frac{SS_{residual}}{df_{residual}}$$

$$\text{F= } MS_{regression}/MS_{residual}$$

Repeated-Measures ANOVA

Please refer to the independent one-way ANOVA section for more detailed instructions for the first portions of an ANOVA. Original table designed by Dr. Rebecca Thompson.

Sums of Squares	DF	Variances	F-ratio
1. *SS* total	6. df_{total} $N-1$	11. $MS_{\text{between tmt}}$	13. *F*
2. *SS* Within $\sum SS$	7. df_{within} $\Sigma df_{\textit{inside each treatment}}$		
3. *SS* Btwn Treat	8. df_{between} k - 1		
4. SS Btwn Subjects	9. $df_{\text{btw subjects}}$ $n-1$	12. MS_{error}	
5. *SS* Error $SS_{\text{within}} - SS_{\text{btw subjects}}$	10. df_{error} $df_{\text{within}} - df_{\text{btw subjects}}$		

Scheffe Test

- Step 1: Calculate the absolute values of the differences between each sample mean (A vs B, A vs C, B vs C, for example). It's best to use a table.
- Step 2: Use this formula to find a set of Scheffe values you will use in the next step. $\sqrt{(k-1) * F * MSE * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$. Calculate this value for each pair. Keep in mind that k is condition number, F is the F value found from your ANOVA calculation, and MSE is the mean square error from your ANOVA calculation.

- Step 3: Compare the 2 numbers you got from both steps in each pair. If the absolute difference between the two means is greater than the Scheffe value you found for it, then that means there is a significant difference between those two groups.

Standard Deviation

The standard deviation describes how far the scores of a data set deviate from the mean and is very similar to variance. In fact, in order to find the standard deviation, you first need to calculate the variance (the equation for which is further along in this glossary).

$$\text{Population: } \sigma = \sqrt{\frac{SS}{N}}$$

$$\text{Sample} s = \sqrt{\frac{SS}{n-1}}$$

$$\text{Binomial Distribution} = \sqrt{\sigma^2} = \sqrt{n\pi(1-\pi)}$$

Standard Error

$$(\sigma_M) : \frac{\sigma}{\sqrt{n}} \text{ or } \sqrt{\frac{o^2}{n}}$$

Standard Error of Estimate

The precision of the estimate made by a regression line is measured by the standard error of the estimate, as not every data point will fall perfectly on the prediction line.

$$SS_{regression} = r^2 SS_Y$$

$$SS_{residual} = (1 - r^2) SS_Y$$

$$\sqrt{\frac{SS_{residual}}{df}} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}}$$

Standard Error of the Proportion

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Sum of Squares (SS)

The sum of squares is a variable which represents scores' deviations from the mean. Unless your data set consists of a repetition of one number over and over again, there are going to be scores which are different from the mean and therefore deviate from the mean. Sum of squares is important for formulas regarding *variance, standard deviation, pooled variance,* and *regression.* To find the Sum of Squares, use the formula below. Alternatively, one can subtract each score from the mean, square each deviation, and then add the squared deviations together, but this is a more tedious way of finding the SS.

$$\text{Computational (recommended): } SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

Definitional: $SS = \sum(x - \mu)^2$

Tukey's Range Test

- Step 1: Find the difference between the means of all conditions (A vs B, B vs C, A vs C, etc.)
- Step 2: Calculate Honest Significant Difference using the following formula. You should have everything except q after running an ANOVA. Q can be found in the appropriate table using the degrees of freedom found during the ANOVA calculation process.

$$HSD = q\sqrt{\frac{MS_{within}}{n}}$$

- Step 3: Compare the values. Whatever you got for HSD should be compared to all the mean differences calculated in step 1. If HSD is smaller than the difference, then it's significantly different, otherwise, the differences are not significant.

t Statistic for Independent measures

After finding the estimated standard error, you can finally calculate t. The following equation is the whole theoretical equation, but you can ignore the first parentheses when actually calculating, as both population means should be 0.

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

t Statistic for Related Samples

Remember that there are three kinds of t statistics. These equations are for repeated measures or paired t tests, meaning that the individuals each condition are related in some sort of way, either being the same participants or participants matched up by relationship or participant characteristics. First, calculate the variance (s^2) or standard deviation (s) (the equations for which can be found further along in this glossary), then calculate the estimated standard error (s_{M_D}), and then you should have all you need to find t.

Mean difference: $M_D = \frac{\Sigma D}{n}$

Sum of Squares for difference: $SS = \Sigma D^2 - \frac{(\Sigma D)^2}{N}$

Estimated Standard Error: $s_{M_D} = \sqrt{\frac{s^2}{n}}$ or $S_{M_D} = \frac{s}{\sqrt{n}}$

t Statistic: $t = \frac{M_D - \mu_D}{S_{M_D}}$

Two-Factor ANOVA

Please refer to the section on independent one-way ANOVAs to know what smaller things you need to calculate before beginning this process.

Stage 1:

- $SS_{total} = \Sigma X^2 - \frac{G^2}{N}$

- $SS_{between-treatments} = \Sigma \frac{T^2}{n} - \frac{G^2}{N}$
- $SS_{within-treatments} = \Sigma\, SS_{inside\ each\ treatment}$

Stage 2:

- $SS_A = \Sigma \left(\frac{T_{row}^2}{n_{row}}\right) - \frac{G^2}{N}$
- $SS_B = \Sigma \left(\frac{T_{col}^2}{n_{col}}\right) - \frac{G^2}{N}$
- $SS_{AxB} = SS_{between-treatments} - SS_A - SS_B$

Stage 3:

- $df_{total} = N - 1$
- $df_{within-treatments} = \Sigma df_{inside\ each\ treatment}$
- $df_{between-treatments} = k - 1$
- $df_A = number\ of\ rows - 1$
- $df_B = number\ of\ columns - 1$
- $df_{error} = df_{within-treatments} - df_{between-subjects}$

Stage 4:

- $MS_A = \frac{SS_A}{df_A}$
- $MS_B = \frac{SS_B}{df_B}$
- $MS_{AxB} = \frac{SS_{AxB}}{df_{AxB}}$
- $MS_{within-treatments} = \frac{SS_{within-treatments}}{df_{within-treatments}}$

Stage 5:

- $F_A = \frac{MS_A}{MS_{within}}$
- $F_B = \frac{MS_B}{MS_{within}}$
- $F_{AxB} = \frac{MS_{AxB}}{MS_{within}}$

Variance

Simply put, the variance is a number which represents how much the scores in a data set deviate from the mean. When finding the standard deviation, you must find the variance first. Standard deviation is basically the same concept of variance just restored to the same unit of measurement as the original data. To find the variance, you must first find the sum of squares (the equation for which can be found in this glossary).

Population: $\sigma^2 = \frac{SS}{N}$

Sample: $s^2 = \frac{SS}{n-1}$

---

z-score (for locating an X value)

$$z = \frac{X - \mu}{\sigma}$$

---

z-score (for locating a sample mean)

$$z = \frac{M - \mu}{\sigma_M}$$

---

z-score (for the sampling distribution of the proportion)

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$