$P(\text{A} \mid \text{B}) = P(\text{A \& B}) / P(\text{B})$

$$z = \frac{x - \mu}{\sigma} = \frac{x - \text{Mean}}{\text{Standard Deviation}}$$

$$\mu = \frac{\sum x}{N} \qquad \sigma^2 = \frac{\sum(x^2) - \frac{(\sum x)^2}{N}}{N} = \frac{\sum(x-\mu)^2}{N} \qquad \sigma = \sqrt{\sigma^2} \qquad \bar{x} = \frac{\sum x}{n}$$

N stands for population size n stands for sample size x stands for data

Central Limit Theorem (and related topics):

1. $\mu_{\bar{x}} = \mu$
2. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3. If x is normal, then so is $\bar{x}$
4. Even if x isn't Normal, it's pretty safe to assume $\bar{x}$ is Normal if $30 \leq n \leq 50$.

$\mu_{X+c} = \mu_X + c$ $\qquad (\sigma_{X+c})^2 = (\sigma_X)^2$

$\mu_{cX} = c\mu_X$ $\qquad (\sigma_{cX})^2 = c^2(\sigma_X)^2$

$\mu_{X+Y} = \mu_X + \mu_Y$ $\qquad (\sigma_{X+Y})^2 = (\sigma_X)^2 + (\sigma_Y)^2$

$\mu_{X-Y} = \mu_X - \mu_Y$ $\qquad (\sigma_{X-Y})^2 = (\sigma_X)^2 + (\sigma_Y)^2$

For a multi-step process to find out how many ways it can be done, multiply how many choices at each step.

$n! = n(n-1)(n-2)\cdots 1$ gives the number of ways of arranging n objects $\qquad 0! = 1$

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} = \text{How many ways to pick } x \text{ things from } n \text{ things if order doesn't matter}$$

Binomial Probability: $P(\text{Exactly } x \text{ successes in } n \text{ trials}) = \binom{n}{x}(p^x)(q^{n-x})$

$\mu = np$ $\sigma = \sqrt{npq}$ Normal is a good approximation if np and nq exceed 10

Correlation & Regression:

First calculate these:

$$\sum x \quad \sum y \quad \sum (x^2) \quad \sum (y^2) \quad \sum (xy)$$

Then calculate the following three numbers:

$$\sum (x^2) - \frac{(\sum x)^2}{n} \qquad \sum (y^2) - \frac{(\sum y)^2}{n} \qquad \sum (xy) - \frac{(\sum x)(\sum y)}{n}$$

Now you can find the following:

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

$$y = mx + b \qquad m = \frac{\sum (xy) - \frac{(\sum x)(\sum y)}{n}}{\sum (x^2) - \frac{(\sum x)^2}{n}} \qquad b = \frac{\sum y - m(\sum x)}{n}$$

r^2 tell us what percent of differences in y can be attributed to the regression line on x

$$\bar{x} = \frac{\sum x}{n} \qquad s = \sqrt{\frac{\sum (x^2) - \frac{(\sum x)^2}{n}}{n-1}} \qquad E = \frac{(z_{\alpha/2})(\sigma)}{\sqrt{n}} \qquad n = \left[\frac{(z_{\alpha/2})(\sigma)}{E}\right]^2$$

To calculate percentiles:

1. Rank data from smallest to biggest.

2. Let n be the number of pieces of data for your data set, r represents the rank of each piece of data (1^{st}, 2^{nd}, etc.)

3. Calculate the percentiles, p, of your data pieces with the formula $p = \frac{100}{n}\left(r - \frac{1}{2}\right)$

4. To find percentile, p, find what percent of the way p is between two consecutive p's from your data set and go the percent of the way between the two corresponding pieces of data. If the percentile you are looking for actually corresponds to the percentiles of one of your pieces of data, then that percentile is just that piece of data.

1.5 IQR rule for outliers:

1. Find $Q_3 - Q_1$

2. Multiply by 1.5

3. Outliers are anything above $Q_3 + 1.5(Q_3 - Q_1)$ or below $Q_1 - 1.5(Q_3 - Q_1)$

Hypothesis Tests and Confidence Intervals:

Basic Formula: z (or t) = (non z (or t) of interest – mean) / SD

Comment: For a hypothesis test, if given a level of significance, first set up the rejection area(s) and find the edge(s) of it by using the z (or t) table.

Comment: For hypothesis tests, the z (or t) from the data is (Column 3 – Column 4) / Column 5

Comment: The confidence interval for each case below is Column 3 $\pm$ (Column 2 $\times$ Column 5)

Comment: In any of the cases below, if the population standard deviation(s) are known, then use them and z.

Comment: The sample size needed for CI's (for means) is $n = \left(\frac{(z_{\alpha/2})(\sigma)}{E}\right)^2$, and the sample size for CI's (for proportions) is $n = \frac{(z_{\alpha/2})^2(p)(q)}{E^2}$ (use $p = q = 0.50$ to guarantee the sample size is large enough, use p' in place of p to get a reasonable estimate)

Please see the next page…

Column 1	Column 2	Column 3	Column 4	Column 5
Situation:	z/t:	Non z (or t) of Interest:	Mean:	Standard Deviation or Standard Error:
1 Sample Mean	t, df $= n - 1$	Sample Mean	Population Mean (from H_0)	$\sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$
Difference of Means from 2 Dependent Samples (a.k.a. Matched Pairs)	t, df $= n - 1$	Sample Mean of the differences in appropriate order	Population Difference Mean (from H_0, often 0)	$\sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$, where s is the SD of the differences
Difference of Means from 2 Independent Samples	t, df = Minimum of the Sample Sizes – 1	The Difference of the two Sample Means subtracted in the appropriate order	Difference of the Population Means (from H_0, often 0)	$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$
Difference of Means from 2 Independent Samples with the assumption that the Population Standard Deviations are Equal	t, df $= n_1 + n_2 - 2$	The Difference of the two Sample Means subtracted in the appropriate order	Difference of the Population Means (from H_0, often 0)	$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$, where **Next Page, Note 1**
1 Sample Proportion	z	Sample Proportion, p' = Number of Successes / n	Population Proportion, p (from H_0)	HT: $\sqrt{\frac{pq}{n}}$ CI: $\sqrt{\frac{(p')(q')}{n}}$
Difference of Proportions (percentages, or probabilities of success) from 2 Samples	z	The Difference of the two Sample Proportions subtracted in the appropriate order	Difference in Population Proportions (from H_0, often 0) 0 and non-zero differences have different standard deviations see ⟶	**Next Page, Note 2**

Note 1:

$$s^2 = \frac{(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2}{n_1 + n_2 - 2}$$

Note 2:

HT (0 case): $\sqrt{\frac{(p_{\text{Pool}})'(q_{\text{Pool}})'}{n_1} + \frac{(p_{\text{Pool}})'(q_{\text{Pool}})'}{n_2}}$ where $(p_{\text{Pool}})' = \frac{x_1 + x_2}{n_1 + n_2}$

CI & HT (non-zero case): $\sqrt{\frac{(p_1)'(q_1)'}{n_1} + \frac{(p_2)'(q_2)'}{n_2}}$

Confidence Interval for Population Variance:

Using $\chi^2 = \frac{(\text{df})(s^2)}{\sigma^2}$, get two values from the χ^2 table and solve for σ^2 twice. (Take the square root if you want to find σ).

Hypothesis Test for Population Variance:

Get critical value(s) from the χ^2 table, and use $(\chi^2)_{\text{Data}} = \frac{(\text{df})(s^2)}{\sigma^2}$, where σ^2 is from H_0.

Hypothesis Test for the Ratio of Two Variances:

Use $F_{\text{Data}} = \frac{(s_1)^2}{(s_2)^2}$, making sure the top is bigger than the bottom. Keep track of the two different df's. Use the F-table for the critical value(s). Note that making F(Data) > 1 even if you have two critical values as in a two-tail test, only the right-hand one matters.

***O* and *E* Stuff:**

Right tails only.

Use $(\chi^2)_{\text{Data}} = \sum\left[\frac{(O - E)^2}{E}\right]$

Type I: H_0 : Data distributed in a certain way H_a : It's not distributed in that way

df = Number of Categories – 1

E's are found using *np* where p is the probability of being in a category in H_0

Type II: H_0 : Two characteristics are independent H_a : They are related

df = $(r-1)(c-1)$

E's are found by [(Row Total)(Column Total)] / (Grand Total)

Analysis of Variance (ANOVA)

Collect data from SRS's from *S* different sources.

Assume the populations are normal, the variances are equal, and the data are collected independently.

The ANOVA test is always a one-tailed test to the right.

H_0 : All population means are equal.

H_a : There is some difference in the population means.

The *F*-statistic is found using the following:

$$n = \sum(n_i) \qquad \bar{x} = \frac{\sum(n_i \bar{x}_i)}{n} = \frac{\sum(x_i)}{n} \qquad \mathrm{df}_{\mathrm{Factor}} = S - 1 \qquad \mathrm{df}_{\mathrm{Error}} = \sum(n_i) - S$$

$$(s^2)_{\mathrm{Factor}} = \frac{\sum[n_i(\bar{x}_i - \bar{x})^2]}{\mathrm{df}_{\mathrm{Factor}}} \qquad (s^2)_{\mathrm{Error}} = \frac{\sum[(n_i - 1)(s_i)^2]}{\mathrm{df}_{\mathrm{Error}}} \qquad F_{\mathrm{Data}} = \frac{(s^2)_{\mathrm{Factor}}}{(s^2)_{\mathrm{Error}}}$$

$\bar{x}$ is the overall mean

n_i is how many pieces of data from source i

$\bar{x}_i$ is the sample mean of source i

$(s_i)^2$ is the sample variance of source i

Correlation & Regression HT & CI

Test Statistic: $t_{\mathrm{Data}} = \sqrt{\dfrac{n-2}{1-r^2}}$ $\qquad \mathrm{df} = n - 2$

CI for the average of all y's with $x = x_0$:

$$y_0 \pm (t_{\alpha/2}) \sqrt{\frac{\sum(y^2) - b\sum y - m\sum(xy)}{n-2}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x^2) - \frac{(\sum x)^2}{n}}}$$

CI for a single value of y with $x = x_0$:

$$y_0 \pm (t_{\alpha/2}) \sqrt{\frac{\sum(y^2) - b\sum y - m\sum(xy)}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x^2) - \frac{(\sum x)^2}{n}}}$$

y_0 is obtained by the regression line $y = mx + b$ with x_0 in place of x.

$t_{\alpha/2}$ is the critical value off the t-table with $n - 2$ degrees of freedom.

Hypothesis Tests Outline:

1. Read the question to decide what it's asking about.

2. Set up H_0 and H_a (H_a is what you hope to prove and H_0 is the opposite.)

3. You assume that H_0 is true.

4. You often have some evidence against H_0, a HT is seeing if the evidence is statistically significant, meaning that it is unlikely that you would have such good evidence against a true H_0 by just luck.

5. Determine the rejection region(s) and find the edge(s) of them from the appropriate table. The edge(s) are the critical value(s). The total area of the rejection region is α.

6. Collect the data.

7. Check to see if the data were gathered in an appropriate way and the conditions are OK to proceed.

8. Take the number from the data that corresponds to what you are testing in H_0 and H_a and convert it to the appropriate test statistic. We call this t_{Data} or z_{Data} or F_{Data} or $(\chi^2)_{\text{Data}}$.

9. Check to see if the test statistic (the data number) beats the critical value(s) (the table number(s)). If so, then you can reject H_0.

10. If H_0 is true, then the chance you will reject it by mistake is the significance level (or α).

11. If H_0 is false, then you do not know the chance you will mistakenly fail to reject it.

Finding & interpreting the *p*-value:

1. If the test is right-tailed, then the *p*-value is the area to the right of your test statistic (the data number).

2. If the test is left-tailed, then the *p*-value is the area to the left of your test statistic (the data number).

3. If the test is two-tailed, then the *p*-value is the smaller of the area to the left and right of your test statistic (the data number) multiplied by 2.

4. If you were able to reject H_0, then the *p*-value will be smaller than α.

5. If you were unable to reject H_0, then the *p*-value will be larger than α.

6. To say what the *p*-value means in non-technical terms, a nice format is the following: In the case that **H_0 is true**, the chance we would find evidence as strong or stronger than what we got in favor of **H_a** is ***p*-value**. This is assuming that all the conditions are met and there were no problems with the methodology of obtaining the data. (You should fill in the bold face parts so they are relative to your problem).

7. In practice, some problems may not have an α, in which case only the *p*-value is given for each person to decide for themselves if the evidence against H_0 is strong enough.