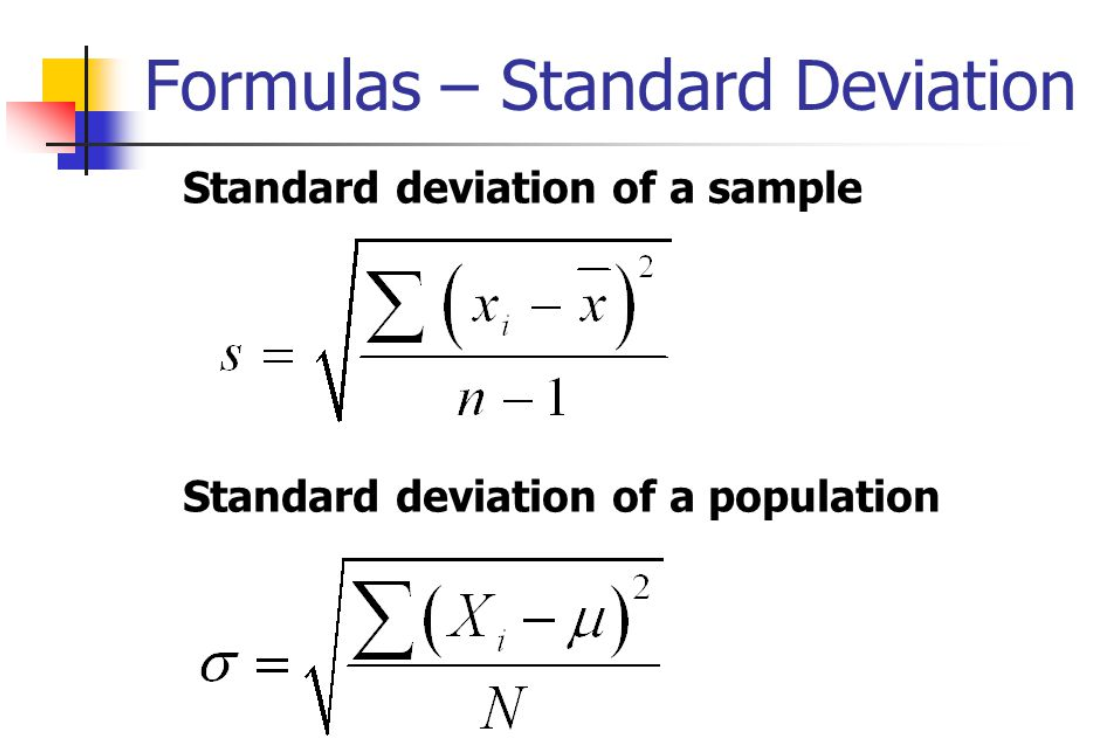
Stats Review Notes

**PART 1:**

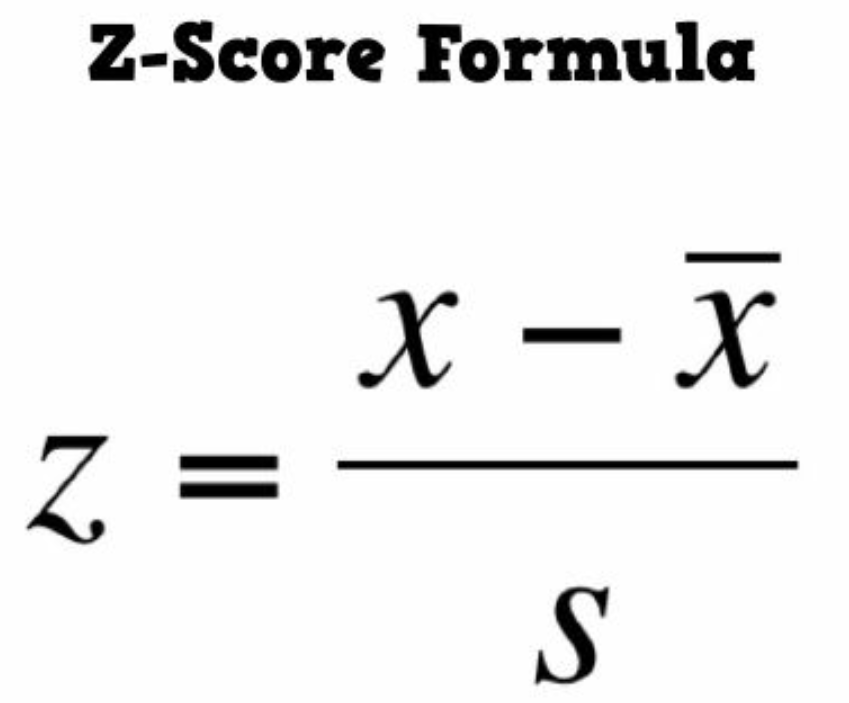
* **Standard Deviation** – average squared distance from the mean
  + % of data that usually falls within number of deviations is as follows:   
    -68% between -1 and 1  
    -95% between -2 and 2  
    -99.7% between -3 and 3
  + Use standard deviation to add or subtract from mean to get range   
    -ex. S.D. = 4.32 and mean = 8 therefore, 1 and -1 standard deviation are 3.68

(8 - 4.32) and 12.32 (8 + 4.32)

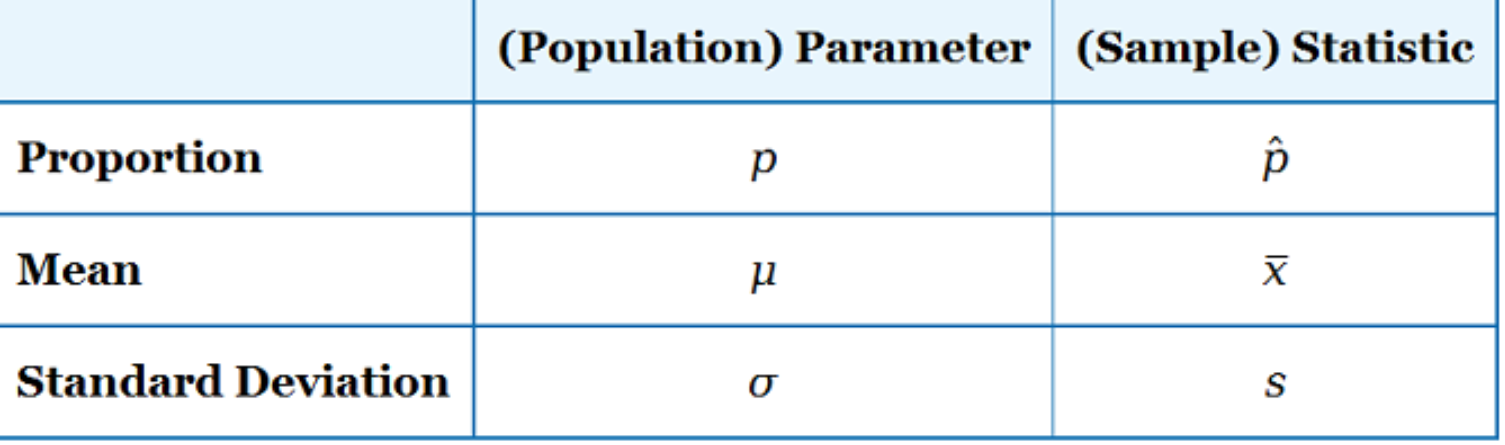
Formula:



x-bar = mean  
n = sample size  
Xi = individual point

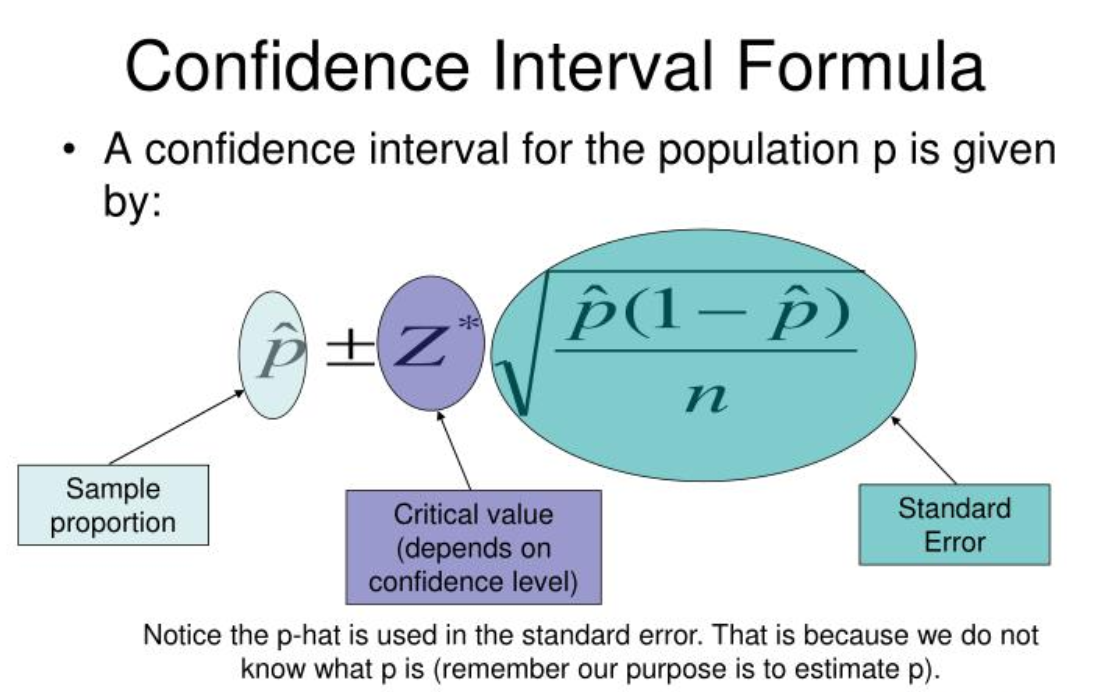
* **Z-Score** – how many standard deviations from the mean a certain data point is  
  

X = data point  
x-bar = mean  
s = standard deviation

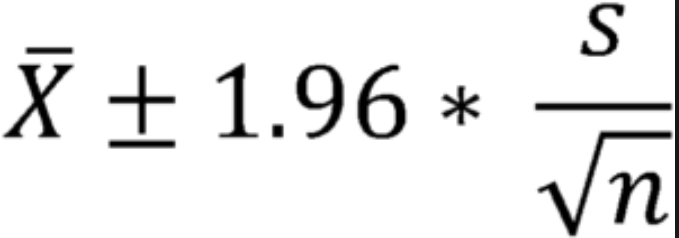


**PART 2:**

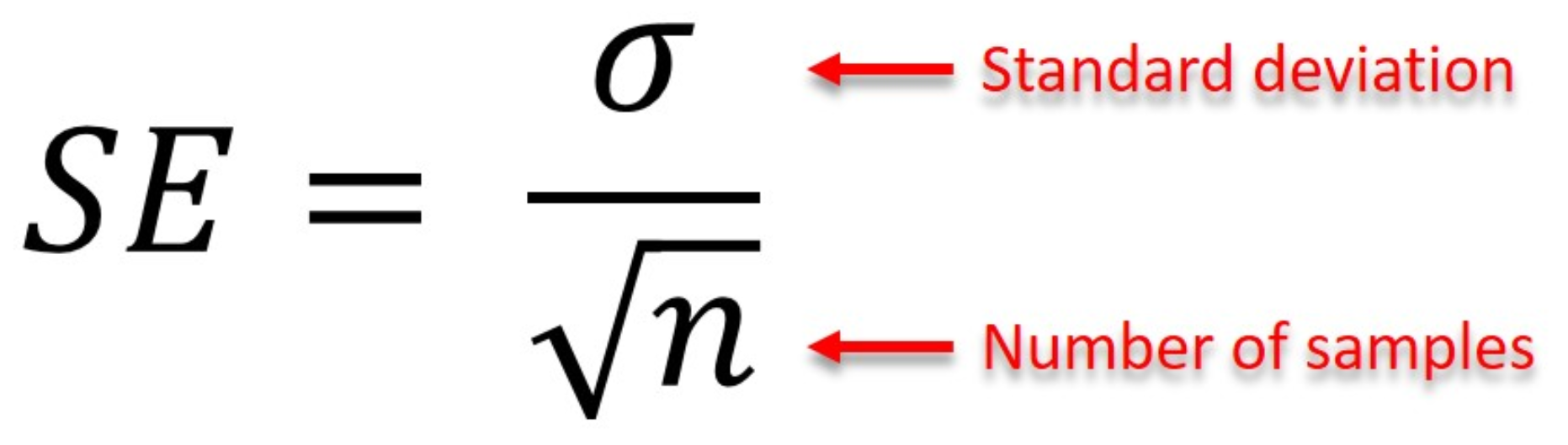
* **Confidence Interval** – creating an interval (range) in which we are x% confident that the true population proportion lies in
  + Ex. Upper limit = 69.6%, lower limit is 50.4% and p-hat (sample mean) = 60%  
    - this means we are 95% (or whatever limit is chosen) confident (based on the sample collected) that our true result (population proportion) will fall within this range of 50.4% to 69.9%



\*\*for population proportion use above formula



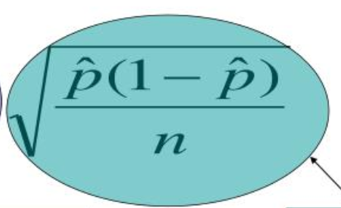
\*\*for sample means use above formula with z-score multiplied by standard error (below)



-In summary, for population given, use population prop. +/- z-score\*(standard error of population AKA standard deviation) formula 1  
-for sample mean given, use sample mean +/- z-score\*(sampling error) formula 2  
-sample error is formula 3 which is standard deviation divided by # of samples

**More on confidence interval formula:**

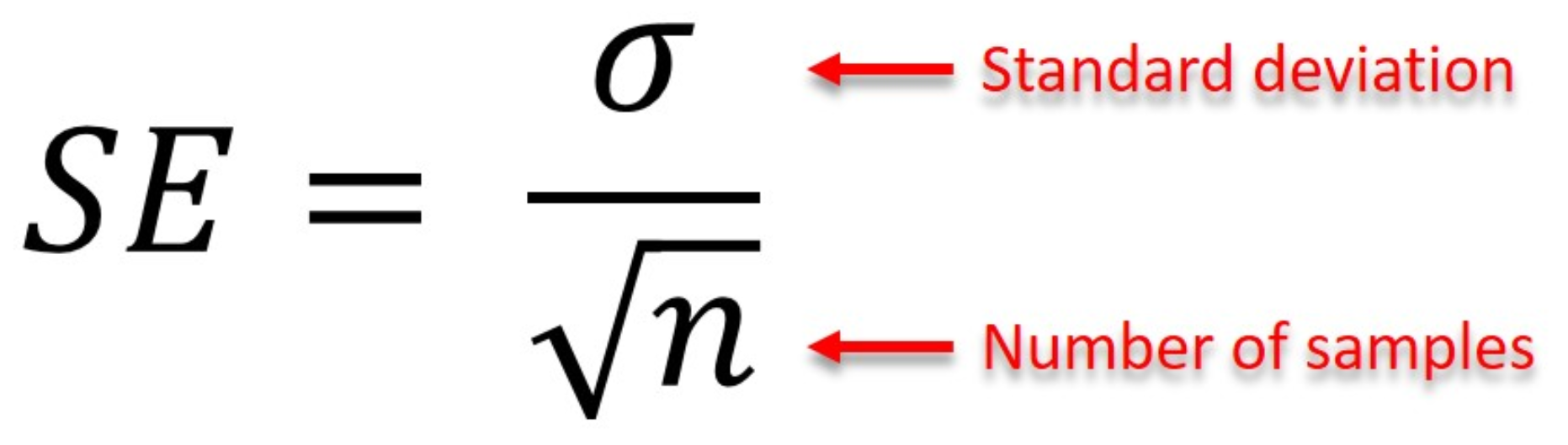
-**Standard error (proportions)** = standard deviation of proportion distribution (sample)  
-ex. We are comparing our sample results to our given population proportion such as 60% of our nation is satisfied with nation being the population proportion   
-our sample can’t possibly represent everyone in our population therefore, each sample will carry some level of standard error: standard error =



P = pop. Proportion  
n= sample size  
p-hat = sample proportion

-this allows us to set up a range around the population proportion that extends the equivalent of one standard deviation

-**Standard error (mean)** – standard deviation of sample means



-standard error = standard deviation of population / square root of sample

**Important z-scores with their confidence level**

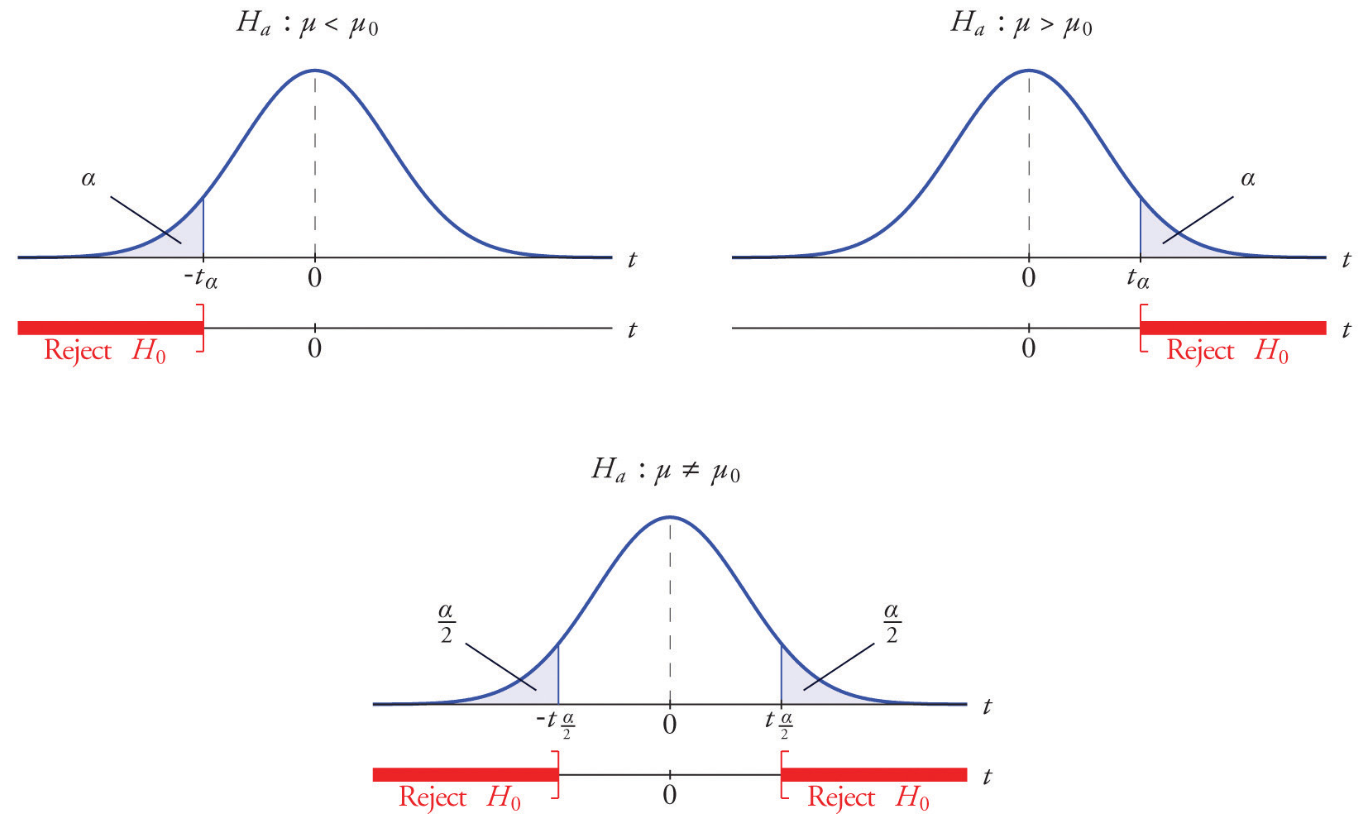
**1.28 z-score = 80% confidence  
1.645 🡪 90%  
1.96 🡪 95%  
2.33 🡪 98%  
2.58 🡪 99%**

**\*side note – sample size goal should be no less than 30 and at 100 samples the distribution starts looking normal and 500 sample size should be goal**

**PART 3: Vocabulary**

* **Covariance** – measures how changes in one variable are able to predict changes in another
  + tests HOW variables interact AKA the level of association between 2 variables (not normalized)
  + positive result means variables move same direction
  + negative result means they move opposite to each other
* **Correlation** – measure of the strength of the linear association or relation between 2 variables
  + Ranges between -1 and 1 with symbol “r”
  + Measure of how well our linear model line fits is also between -1 and 1 symbol “R^2”
* **Test statistic** – our calculated # that we get when applying our particular test for hypothesis testing
* **Critical value** – the # we get from the chart of our particular hypothesis test to determine whether to accept (fail to reject) or reject hypothesis
* Ex. Calculated test statistic is .004 compared to 95% p-value .05 (critical value) therefore we reject the null hypothesis
* **P-value** is rejection region
* **Type I error** – false positive ex. test positive for cancer when you don’t actually have it
* **Type II error** – false negative ex. test negative for cancer when you actually do have it

**PART 4: Hypothesis testing**



1. (Top Right)

**Left tailed:**Ho µ ≥ 12  
Ha µ < 12

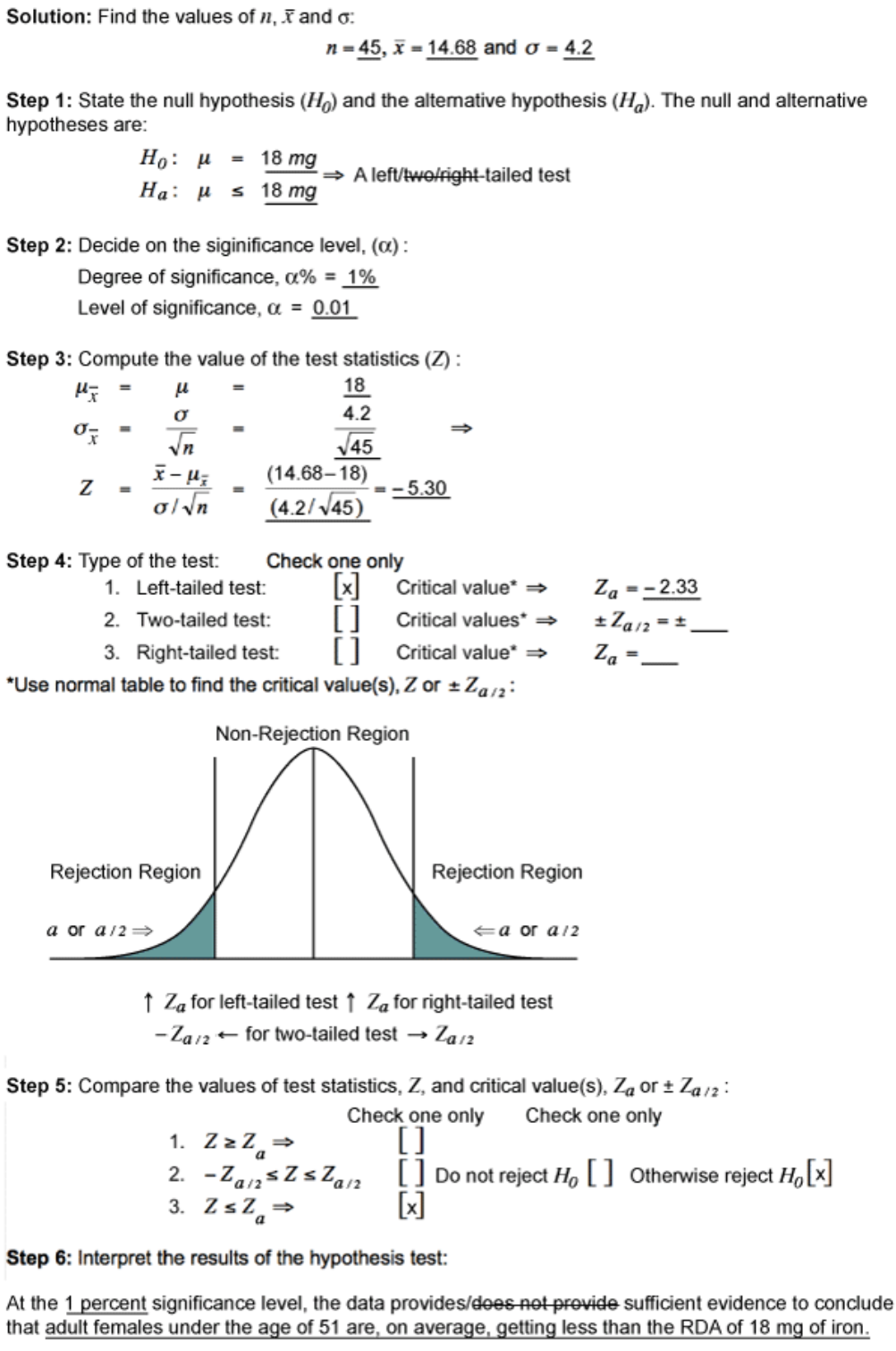
1. (Top Left) **Right Tailed:**

Ho µ ≤ 12  
Ha µ > 12

1. (Bottom Middle) **Two-Tailed:**

Ho µ = 12  
Ha µ ≠ 12

**Steps to Hypothesis Testing:**



**PART 5: Popular Tests**

* **T-test** – compares the means of 2 samples for significance
  + samples DO NOT need to be from the same population
  + use if n < 100 and population variance or standard deviation is unknown
  + 2 sample t test – samples came from same population means
* **Z-test** – compares larger sample mean to population mean and the sample is assumed to come from the population
  + use if n > 100 and population variance or standard deviation is known
* **Chi Square** – Goodness of Fit test, used to compare categorical data
  + Observed vs Expected (usually categorical data)
* **Chi Square** – test for independence, used to determine if 2 sets of categorical variables are associated or independent
  + Ho = no association, Ha = there is an association
* **Chi Square** – test for single variance, determines if true population variance is different than hypothesized population variance by using a sample from the population
  + Ex. comparing long time measured population to current population after an adjustment was make by taking a sample from new population
* **F-test** – for 2 variances, tests difference in variances between 2 independent populations
  + Ex. NYSE variance vs. NASDAQ variance
* **ANOVA** – compares means (using variance) of 2 or more categorical samples
  + used to determine if the variation between reported outputs is the result of a particular factor or randomness
  + two-way ANOVA would be examining multiple factors of each sample such as age, groups, names
  + Ex. mean sales of grocery store vs. drug store vs. convenience store
* **Regression Analysis** – helps us investigate the relationship between 2 variables such as education and lifetime earnings
  + Expressed graphically with a scatterplot and best fit line
  + Finds the formula that will create the best line to fit the data to make future predictions
* **Linear Regression** – finds line which best describes relationship between 2 variables
  + Also used to predict a dependent variable using an independent variable
  + X axis = independent variable
  + Y axis = dependent variable
* **R^2 –** coefficient of determination, score between 0 and 1 that tells us if our regression line is a good or bad fit for our plotted data
* **r** – correlation coefficient, score between -1 and 1 that shows how one variable affects the other, r = square root of R^2