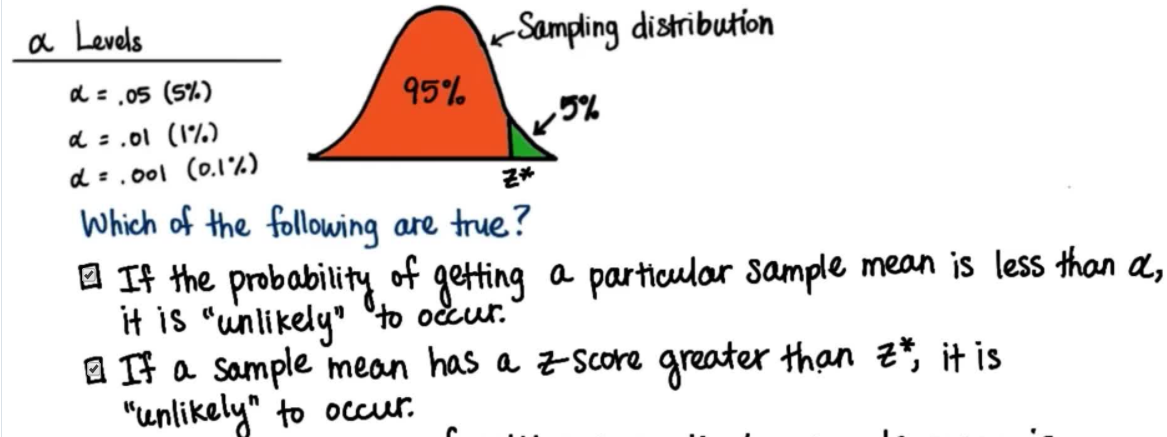
***Udacity Data Analyst Track***

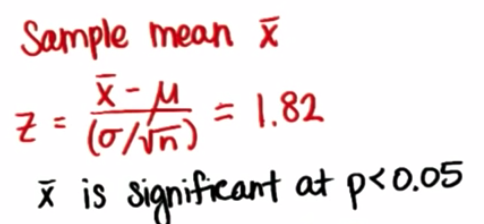
**I. Into to Descriptive Stats**

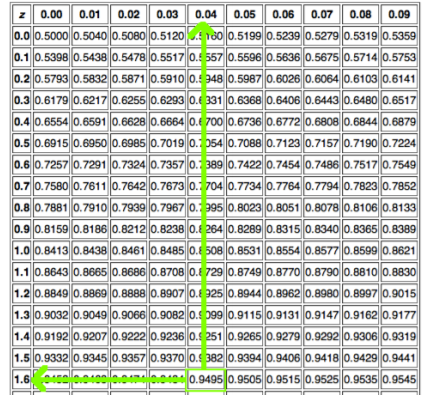
3. Hypothesis Testing

* **Levels of Likelihood/Alpha Levels**
* If probability of getting a sample mean is < 0.05, 0.01, 0.001 (5%, 1%, 0.1%) 🡺 UNLIKELY



* So, if probability of selecting a sample mean is < alpha ( p < a ), it falls in the **critical region**, which is cut off from the rest of the distribution by the z-score of the critical region = **z-critical value**
* If z-score of a sample mean > z-critical = evident the sample statistics are different from the regular/untreated population
* Alpha = 0.05 🡪 5% 🡪 z-crit = 1.65 0.01 🡪 1% 🡪 z-crit = 2.32 0.001 🡪 0.1% 🡪 3.08
* Those sample means that fall in any of the critical regions are those that are unlikely to have happened by chance
* sample statistics are very different to population parameters so there’s strong evidence of an effect from some intervention



* **A hypothesis test** = used to test a claim about how an observation may be different from the *known* population parameter.
* **Alpha level (a)** =helps us determine the **critical region** of a distribution.
* **Null Hypothesis** = always an equality = the claim we’re trying to provide evidence *against*.
* H0 : µ0 = µ
* H0 : µ0 >= µ
* H0 : µ0 <= µ
* **Alternative Hypothesis =** theresult we are checking *against the claim* 🡪 always some kind inequality.
* Ha : µa <> µ
* Ha : µa > µ
* Ha : µa < µ
* Ex: A towns census from 2001 reported average age of people living there = 32.3 w/ SD = 2.1 years.
* The town takes a sample of 25 people + finds average age = 38.4 years.
* Test the claim that the average age of people in the town has increased w/ a level = 0.05
* 1st define our hypotheses:
* H0 : µ0 = 32.3 years
* Ha : µ0 > 32.3 years
* ID the important info
* x¯ = 38.4 s = 2.1 n = 25 SE = 2.1/sqrt(25) = 0.42
* Last piece of important info we need is our **critical value**:
* Finding Z-critical value 🡪 look up as close as we can to 95% = **Z-crit =** 1.64
* 
* Once we have all our important info we can now find our **test statistic**:
* Z-score = obs – mean / SE 🡪 38.4 - 32.3 / 0.42 = **14.5238**
* Since our z-score *is much bigger than our z-crit* 🡪 REJECT the claim/null) that the average age of people was 32.3 years
* **Type I Error** 🡪 reject the null when null hypothesis is *actually true* (FP)
* Probability of committing a Type I error = alpha level (a)
* **Type II Error** 🡪 fail to reject the null when it is actually false (FN)
* Probability of committing a Type II error = alpha level (b)
* Ex: An insurance company is reviewing its current policy rates. When originally setting rates, they believed the average claim amount = $1,800. They’re concerned the true mean is actually higher than this, b/c they could potentially lose a lot of money. They randomly select 40 claims (n), + calculate a sample mean = $1,950. Assuming the SD of claims = $500, set a = 0.05 + test to see if the insurance company should be concerned.
* H0 : µ0 = $1,800
* Ha : µ0 > $1,800
* x¯ = $1,950 s = $500 n = 40 SE = 500/sqrt(40) = 79.0569415
* Z-crit = 1.64
* Z-score = obs – mean / SE 🡪 1,950 – 1800 / 79.0569415 = **1.897366596**
* Since our z-score is bigger than our z-crit 🡪 REJECT the claim/null)
* Ex: Explain a type I and type II error in context of the problem. Which is worse?