LS 6 Learning From Data

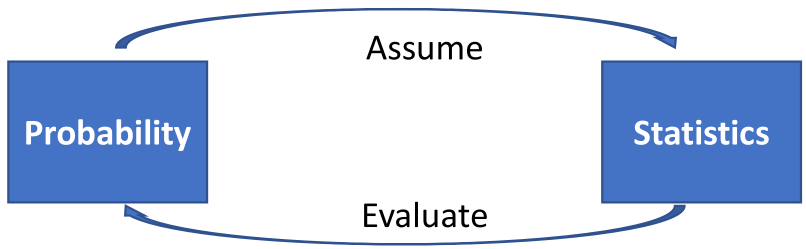
Data Scales (from https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/)

Nominal variables -"*name*," or label"
Ordinal scales provide measures about the *order* or preference of choices,
Interval scales give us the preference order + the ability to quantify *the difference*.
Ratio scales give us the ultimate–order, quantified distance plus a true zero

			Metric	
Provides:	**Nominal**	**Ordinal**	**Interval**	**Ratio**
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Where We Are

- We are at the point where are starting to learn from data



- How are we doing this?
  - Collect Random Samples (iid samples of random variables)
  - Compute statistics, estimate parameters
  - Make Assertions
  - Evaluate Assertions (hypothesis testing)

Assertion Evaluation And Supporting Results

- Sampling dist'ns are the basis evaluating assertions
  - Sampling dist'n is the prob dist'n of the statistic of interest
  - A function of the underlying RV, the statistic , estimator and the popl'n (probability law)
- Law of Large Numbers: If $E[|X_i|]<\infty$, statistic=$n^{-1}*\sum X_i$, $\{X_i\}$ is i.i.d. then $| n^{-1}*\sum X_i - E[X]| \xrightarrow{p} 0$ as n↑
  - LLN connects random samples, sample size and estimation to expected values
- Evaluating and Comparing Estimators
  - Bias: $E[\theta-\theta] \equiv$ accuracy
  - Variance: $Var(\theta) \equiv$ precision
  - Unbiased and lower variance estimates are generally preferred
- We can in theory evaluate our assertion under the sampling distribution of the statistic whatever it may be, this may be difficult, so ….
- The Central Limit Theorem ("The Statistician's Shim")
  - Number of definitions, we will use the following: Distribution
  - If $-\infty<E[X]==\mu<\infty$ and $Var(X)==\sigma^2<\infty$ and the sample $\{X_i\}$ are iid random variables e.g. independent from the same distributions
  - Then FOR ANY DISTRIBUTION the sums and averages of sample realizations of size n converge in [redacted] to the normal distribution with mean = μ, and variance = σ^2/n, **convergence speed and asymptotics in part dependent upon dist'n skewness and size of n**

Hypothesis Testing And Confidence Intervals – On being A Kenny Rodgers - We are going to examine "Know when to hold them and know when to fold them"

1. Let {X(i)} be iid, σ^2 known, i=1,…,n > 30
   - Xbar =n^-1*∑ X(i) estim by xbar = n^-1*∑ x(i) (must be metric)
   - Z* = (Xbar-μ)/(σ^2/n)^0.5 ~ approx N(0,1)
2. Same as 1., but σ^2 unknown
   - t* = (Xbar-μ)/(s^2/n)^0.5 ~ approx t(n-1)

- If dist'n X unknown, n>30 and skewness "not bad" relay on CLT, otherwise use non-parametrics
- Do this to make inferences about popl'n from sample*, the assertion.*
- Go on to evaluate an inference in form of test with two alternatives
  - Null hypothesis; H(0): θ=a vs Alternative one of H(a): θ≠a or H(a): θ<a or H(a): θ>a   must specific H(a) a priori

Make decision

- Under unknowable that H(0) correct, using transform above and the realizations of Z* or t* and a predetermined "rule" we find H(0) likely and cannot we reject
  - Otherwise reject H(0)

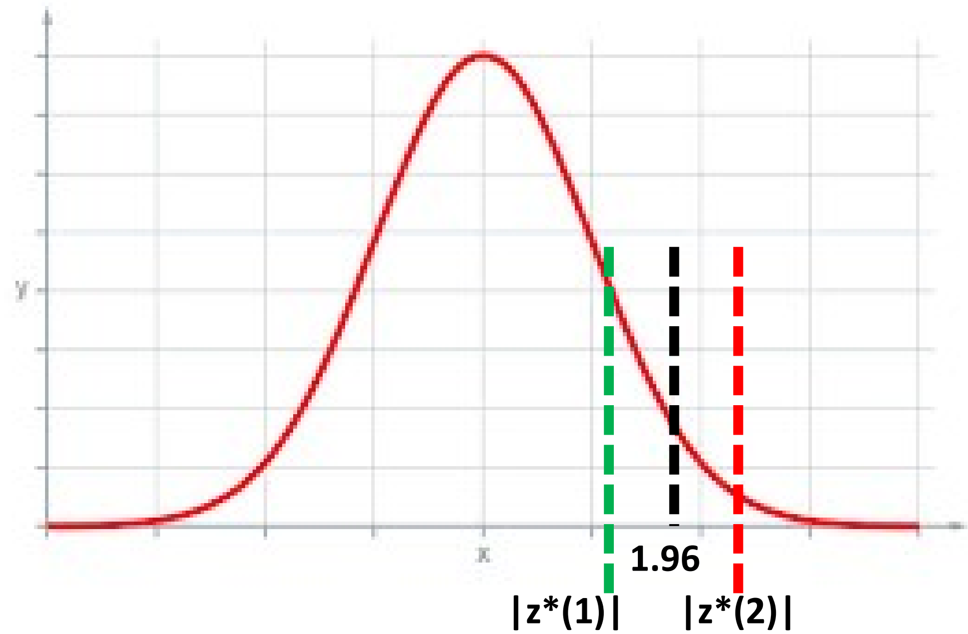
But Problem!

- Under the assertion H(0) true what can we say about the values of the RVs Z* or t*
  - They can take on a support from - ∞ to ∞.
- So statisticians must be gamblers and any time you are a gambler you can make an error
  - How do we gamble: select level of confidence, say CL=95% ➔a level of risk taking, α = 1-CL =0.05 or 5%
- Let's evaluate the transform Z* = (Xbar-μ)/(σ^2/n)^0.5 ~ approx N(0,1) to see how these rules lay out

Rules (Using long run thinking)
Three mathematically equivalent rules
1. Use Confidential Interval (CI)
2. Use upper and lower critical quantile
3 Use P-value

CI & Critical Values (Z Score)

- P(-1.96<=Z*<=1.96) = 0.95
- Where Z is defined as above:
- P(-1.96<= (Xbar-mu)/(sigma/n^0.5) <= 1.96) =0.95
- = P(se*-1.96<= Xbar-mu <= se * 1.96)
- = P(-Xbar – se*1.96<= -mu <= -Xbar + se*1.96)
- = P((Xbar + se*1.96 => mu => Xbar - se*1.96)
- Under frequentist statistics meaning 95 out of 100 such CI's so construct will include μ, NOT there is a 95% chance that CI will include μ

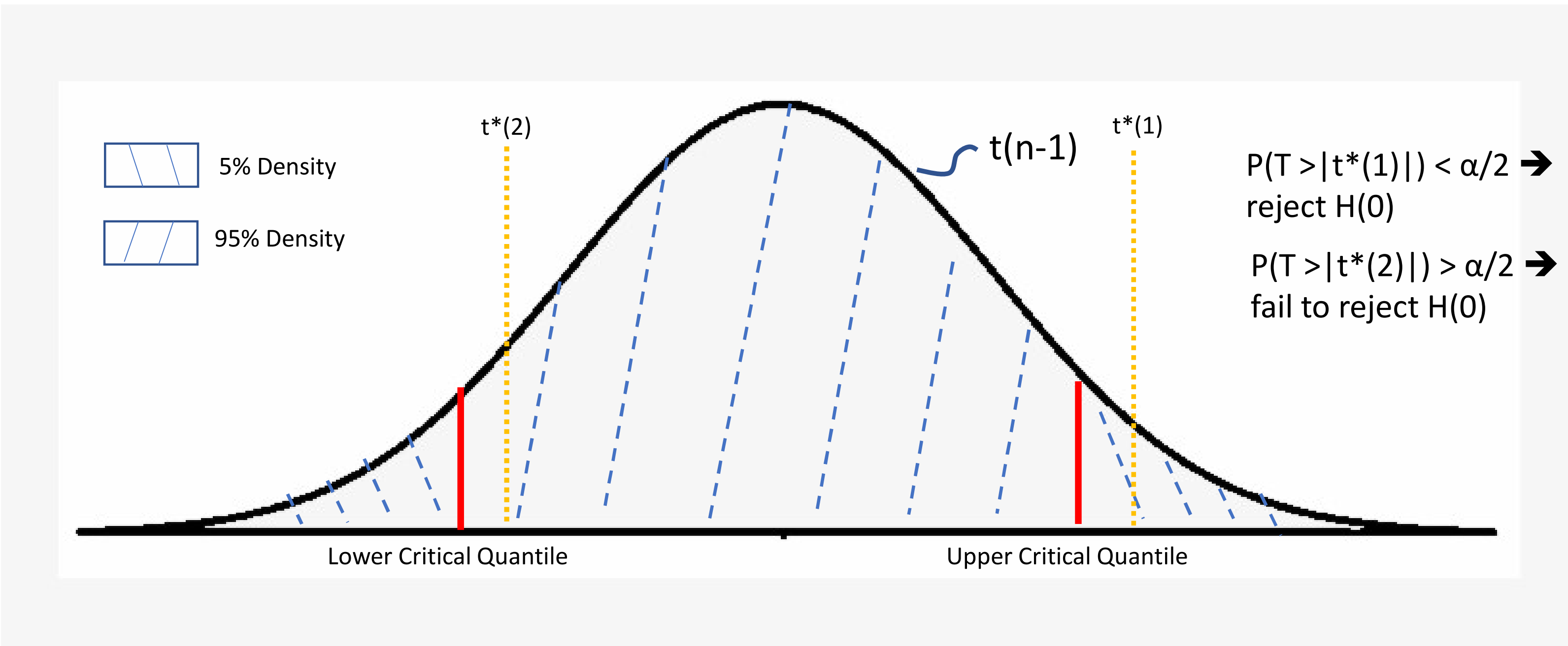


If α is set at 0.05 ➔ α/2 =0.025
P(Z>1.96) = 0.025
P(Z>|z*(1)|) > P(Z>1.96) ➔ Fail to reject
P(Z>|z*(2)|) < P(Z>1.96) ➔ Reject|

Hypothesis Testing Relationships

We concrete-ize all of this “likely” discussion and relate CL, α, P_values, t*’s, n, H(0) and H(A) as follows:

Let T be a RV ~ t(n-1)



Numbers to Know

- $P(-a <= Z <= a) = p$, where $Z \sim N(0,1)$

p	a
0.90, 90%	+/- 1.64
0.95, 95%	+/- 1.96
0.99, 99%	+/- 2.576