Handy Dandy Stats Checklist: *All you need to know from Stats for Metrics*

Random Variables and Distributions

1. What are random variables (rv's)? distributions? probabilities? probability density functions?
2. … characteristics/properties: Means, variances, covariances, correlations, independence, …

Estimation: Gather data and estimate true underlying parameter values

3. Estimators v. estimates: Start with estimators (rules; random variables) … gather data … and generate estimates (what you get when you apply those estimator rules to Your data)
4. Example: Sample Mean estimator; random sampling from X (unknown: mean μ and variance σ^2)
   a. ex ante: Independent random variables X_i's *iid X* (You do remember *iid*? Yes?)
   b. ex ante: Sample Mean Estimator (it's a random variable… until you have data!): $\overline{X}_n = \frac{1}{n}\sum X_i$
   c. ex post (w/ data!): Estimated sample mean (remember all those sample statistics?): $\overline{x}$
5. ***LUE***s: Linear Unbiased Estimators … linear and unbiased
6. ***BLUE***: Best Linear Unbiased Estimator … minimum variance in the class of LUEs
7. Return to example: The Sample Mean estimator is a LUE, and also BLUE! So far, we've assumed random sampling from the same distribution, and finite mean and variance… that's it!
8. So far: No distributional assumptions! Nothing about whether X has a Normal distribution or …

Inference: How did we do? How close is the estimate to the actual true parameter value?

9. Stay with the example: Now!... we need to make distributional assumptions! Assume $X \sim N(\mu, \sigma^2)$ … given the other assumptions, we know the distribution of the Sample Mean estimator:
   $\overline{X}_n \sim N(\mu, \sigma^2 / n)$ … and the standardized estimator: $\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
10. But Wait! We don't know the variance σ^2!
11. No problemo! Just estimate it with the Sample Variance, S_x^2, which is an unbiased estimator of σ^2. And then we can generate an estimate of the *standard deviation* of the Estimator, which we confusingly call the *standard error* $se = S_x / \sqrt{n}$ … *Why is stats so confusing?*
12. Substitute your estimate, *se*, into the formula… and you have the ***t statistic***, which has a t distribution with n-1 dofs: $\frac{\overline{X}_n - \mu}{se} \sim t_{n-q}$ … ***The Foundational Cornerstone of Inference***:
13. I say again: ***The t statistic is The Cornerstone of Inference***
14. ***Confidence Intervals:*** Interval Estimators…. bounds are Sample Means +/- a certain number of standard errors (*se*'s)… How many se's? Several. More precisely: Choose a Confidence Level and work with *The Cornerstone of Inference*, the t statistic and the t distribution with n-1 dofs.
    a. Interpretation of CIs: If confidence level is C%, then C% of the random intervals generated in this fashion (with random sampling of the data), contain the true unknown parameter value, μ.
15. ***Hypothesis Testing:*** In metrics we are almost always testing Null: $H_0: \mu = 0$. Use a two-tailed test. Reject if the sample mean is large in magnitude, far far away from 0. How far away? c *se*'s (standard errors); so reject if $|\overline{x} / se| > c$. Where does c come from? Choose a *significance level* α, and use *The Cornerstone of Inference*. What's a good α? 10%? 5%? 1%? You decide! … but make it small!
16. Or better yet, compute the p value and reject if $p < \alpha$. Where do *p* values come from? Use The Cornerstone of Inference! $p = prob\left(|t_{n-1}| > |\overline{x} / se|\right)$, assuming $H_0: \mu = 0$
17. *Statistical significance*: If we reject the Null Hypothesis that the true parameter value is 0 (at some significance level α), we say that our estimate is *statistically significant (at significance level α)*.

Done! I hope that wasn't too painful! What did I miss?