A (sort of) e-Text for Core Statistical Procedures

The goal of any statistical analysis is to be wrong in the best possible way.
Tyler Johnson, April 2021.

Good statistical practices rests on three foundations:

(1) Statistical concepts
(2) Statistical tools
(3) Subject matter expertise

I presume you have (3) pretty much covered; my intention is to help you with (1) and (2).

Special sections:

	**Core Concepts:** Some topics are core material for *any* statistical analysis.
	Some notes herein are for the **statistical aficionado**, and can be skipped by others.
	**Nerd alert…** Some topics are particularly nerdy, and can usually be skipped. The photo to the left will be your warning.

Stats FAQs is a file I hope you find useful for repeated exploration and reading, so I let it sit here on its own..

	**Stats FAQs.** An interactive exploration of questions frequently asked by my students. A useful "overview"-type study guide. I encourage a tour.

"Our brains are statistical organs that are built simply to predict what will happen next," said Karl Friston, a professor of neuroscience at University College London.

Chapter Listings

Each chapter has the relevant documents (Word or pdf files) listed in a table, with annotation so you know what you are getting into when you open any of them. For your convenience, I created hyperlinks on the document titles.

Statistical Foundations	
This section features notes on background concepts that you need to understand for *any* statistical analysis. In particular, the Central Limit Theorem, and properties and procedures for confidence intervals and doing hypothesis tests are essentials.	
	**An English-Statistics Translation Dictionary.** A number of statistical terms use words borrowed from everyday English, but then impose a new definition for their technical use. If one is not careful, miscommunication can ensue because of this cross-cultural dissonance.
	**Central Limit Theorem.** A central idea indeed. It is the statistical law of gravity that shows that certain statistics (means, differences in means, regression coefficients, and more) will at least approximately have a Normal distribution, given sufficient sample size. This in turn justifies the use of the t-distribution for many applications.
	**SE is synonymous with SD.** It is quite useful to understand that the SE of a statistic is in fact an estimate of the SD in the distribution of the statistic. That, coupled with the Normality that the CLT brings us, forms the basis of many statistical methods.
	**Intro to Hypothesis Testing.** Such a culturally engrained tool, you simply *must* learn about it, although I would be glad to toss it into the ditch and drive on using properly done estimates. Sigh.
	**Intro to Confidence Intervals.** Any time you do a statistical analysis, you are either taking aim at estimating something or testing something. A proper estimate requires some statement of precision; a C.I. is a formal way of doing that.

Advanced Notes on Core Topics

	**Construction of Standard Confidence Intervals.** In many cases, a stat package will do the work for you, but not always; for instance CIs for regression coefficients aren't usually given. Understanding CI mechanics can you are more effective data analyst
	**Connections Between Intervals and Tests.** Why do we choose confidence levels to be complementary to alpha levels? How might one rethink confidence intervals when doing a one-tailed (a.k.a. one-sided) test? Finally, the aficionado section, which argues that with well-done estimation, hypothesis tests are redundant.

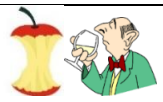
Excel Tools for Core Concepts

	**Central Limit Theorem Demo.** One of the most profound facts that enables simple statistical analyses is that the sample mean[1] will tend to have a Normal distribution, so long as sample size is large enough. This tool demonstrates that fact (known as the Central Limit Theorem)
	**Samdist Binomial:** The CLT applies to Binomial proportions as well as measured values.
	**SD.** A standard deviation calculation demonstrator for those who want to know…
	**Confidence Interval Properties.** Understanding the properties of confidence intervals will make you a more confident user of them. This tool demonstrates the effect of confidence levels and sample size, both of which are, in principle at least, chosen by you.
	**CI Calc** demonstrates the steps in a simple confidence interval calculation
	**Connect CI to p.** A wee demonstration of the relationship between confidence intervals and *p*-values from tests.
	**Random.** A demonstration of the behavior of truly random numbers, compared to "random" values chosen by humans. Also shows that predictability in random phenomena comes with increased sample size.
	***Z* table** These days, you hardly ever need to look up values in a *z*-table; statistics packages do that work for you automagically. This particular table is quite easy to use, although you might well go through your entire career and not need it.
	***T*-tables.** These days, you hardly ever need to look up values in a *t*-table; statistics packages do that work for you automagically. Even though I think this particular table is quite easy to use, you might well go through your entire career and not need it.

[1] And its cousins: the difference in two means, coefficients from regression models, sample proportions (and differences in proportions)…

Summarizing Data

	**Graphical Summary.** A general introduction to commonly used statistical terms and concepts, in the service of summarizing a single data set.
	**Types of data.** A brief and possibly useful graphical summary of data types and tools pertaining thereto.
	The data are skewed: **Means or medians?** There is an urban legend in the science community that says, "If the data are skewed, you should use medians rather than means, since means are sensitive to outliers." I strongly disagree with this advice.
	**Boxplots.** Interpretation and uses

Methods for One, Two, and Paired Samples Some scientific studies are composed of comparisons between two groups, in which case the tools described here are pertinent. That said, such tools are quite frequently used in management-oriented work for, usually for government agencies.	
	**Methods for Single Samples.** One does not often do much with a single sample, aside from describe it (mean, SE, CI, perhaps a histogram or boxplot), so this section is included mostly for completeness.
	**Methods for two Independent Samples.** In most cases (when interest is in the difference in two means, and sample sizes are large enough), the tool of choice for testing or estimation is the two-sample *t*. Necessary assumptions are pretty simple, but there is a twist: they differ when doing a test from when making an estimate.
	**Degrees of Freedom.** For those interested, a brief foray into the genesis of the idea, and their implementation…
	**Variance Assumptions for the two-sample *t*.** The first page of this note is core material, explaining my argument for when and when not to assume equal variances. The second dives into some of the math related to the assumptions.
	**Behind the Scenes.** This note examines in detail the math that is used when you do or do not assume equal variances for the two-sample t.
	**Methods for paired data.** A paired data[2] design, when feasible, is usually a more powerful[1] design than using two independent samples.
	**Binomial Proportions from a Single sample.** Details for estimating (with confidence intervals) and testing a single proportion. The Agresti-Coull so-called "+4" confidence interval is introduced.
	**Two proportions.** Protocols and procedures for testing and estimating (with confidence intervals) the difference in two proportions.
	**The Agresti-Coull.** CI procedure**.** Technical notes on how and why the Agresti-Coull method came to be. Short story: the classical interval method sucks; and the A-C is a clever approximation to a sophisticated improvement (which works if you use 95% as your chosen confidence level.

[2] Paired data can arise from a "before/after" study on the same elements, or from (say) having identical twins, or sibs in a study. The idea is that you might expect the data *within* pairs to be at least somewhat similar due to the nature of the pairing.

Simple and multiple linear regression.

Since it is usual and useful to study simple linear regression first, I have ordered the contents accordingly.

Simple Linear Regression In many instances, multiple regression (as such or by any other name) is a go-to tool for analyzing scientific data. The notes in this section present essential background material that will make your foray into multiple regression much easier.	
	**Simple Linear Regression.** SLR is itself not often-used (because only rarely will you have only a single numerical predictor), but the concepts you learn by studying it carry over seamlessly to multiple regression.
	**Goodness of Fit.** It's not what you might think (i.e. a small *p*-value for the slope and a high R^2). It is a bit more subtle, and something you check *before* you look at the numerical output (including the *p*-value for the slope and R^2).
	**Skewed Y and X.** A demonstration that the distribution of Y and/or X do not determine whether assumptions are met.
	**Creating Categories From a Numerical Predictor.** I see this done from time to time (Low, Medium, and High, for example, from elevation data). It generally leads to a loss of statistical power and an increase in complexity of telling the underlying story. As a general rule, if a variable is innately numerical, you will likely come out ahead by treating it as such.
	**The meaning of "Linear Model".** It's not what you might think (i.e. the name does not imply a straight line model only). If you are curious, click away…
	**Technical notes**. Short notes on a variety of background items (formulae for coefficients, parsing R^2 and adjusted R^2).

Excel tools for studying SLR.

	**Influence in Regression**. An interactive simulator to learn under what conditions a datum is at risk for being influential in a regression.
	**Correlation** An interactive tool that can give you a feel for what various values of correlation might look like in a scatterplot between Y and X.

On logarithmic Transformations… the whys and wherefores.

	**Intro to Log Transformations.** Relativity in statistical relationships is ubiquitous, and they can be handled easily by logarithmic transformations of the response variable, one or more predictors, or perhaps all variables. It is essential to have at least a superficial facility with this topic.
	**Examples of log-transformations** Several visual examples where you can see the consequences of log-transforming Y, X, both, or none.
	**Characteristics of Log-transformations.** This note shows the math behind logarithmic transformations in regression models. It is not necessary to understand things at this level (you don't have to look under the hood to know how to drive a car), but some will find it interesting.

Excel tool for log-transformations

	**Inference from log-transformations.** In the event you have log-transformed Y, a predictor (or both), and if you need to interpret the slope coefficient, this tool will help get there.

Multiple Regression	
Multiple regression is a central tool for scientific research; many questions are addressable by properly created regression models.	
	**A Brief Intro to MLR.** Multiple regression is a widely used tool, and this note is an attempt to introduce some of the essentials; I introduce a metaphor (people working in groups on a project) that I think has some analogies that will aid learning about multiple regression.
	**Model Creation.** This note suggests a flow to the process of multiple regression. It is worth re-reading as you learn more about the elements in multiple regression.
	**Multicollinearity.** Sometimes the presence of multicollinearity (referring to the fact that predictors are often correlated with one another) can be a nuisance; sometimes it can be ignored. Learn how and where to measure it (and how *not* to measure it).
	**Model selection.** I will likely sometime merge this note and the preceding one, but just now need to give it some space before I can make a good decision of which bits to keep and which not. So two related notes, but not completely overlapping.
	**Creating and Interpreting Interactions.** I think interactions are a pretty important and useful concept, but I know that much science gets done without giving them a second thought.
	**Poking at Interactions.** In a spirit similar to the foregoing on model selection/creation, I will likely someday merge these two, attempting then to keep the best of both and losing the rest. Until then… ☺.

Analysis of Variance	
	**A Brief Introduction to ANOVA.** This introduction includes a few bits of philosophical advice arguing against jumping full on into doing an ANOVA (in brief, it doesn't always answer the question(s) of greatest interest, so why bother?)
	**The Bonferroni correction.** When one has a multiplicity of comparisons to make the risk of at least one false significance grows exponentially as the number of comparisons goes up. I show you the classical correction to it, and I also present some caveats and concerns I have about its use.

Excel Tools for ANOVA

	**Contrasts** The arithmetic for making comparisons other than simple two-treatment comparisons can be a bit tedious. This tool will help with that.