WHAT STATISTICIANS DO

THIS IS AN ATTEMPT TO EXPLAIN WHAT STATISTICS IS BY LOOKING AT WHAT STATISTICIANS DO, AS OPPOSED TO WHAT YOU OR THEY SHOULD OR SHOULDN'T DO.

1. IDEALLY, STATISTICIANS SHOULD START WITH A DESIGN PHASE: HOW LARGE A SAMPLE IS NEEDED, WHERE SHOULD OBSERVATIONS BE TAKEN, WILL WE HAVE THE RIGHT DATA FOR THE ANSWERS WE NEED. THIS IS A PRACTICAL RARITY, THOUGH EXPERIMENTAL DESIGN IS A BOOMING BUSINESS, OFTEN MORE IN THEORY THAN PRACTICE. PARENTHETICALLY, MUCH PROGRESS IN DESIGN CENTERS AROUND THE FOLLOWING PROBLEM: WE HAVE A SPACE WITH A DISTANCE, WE WANT TO PUT n POINTS DOWN SO THAT THE MAXIMUM OF THE DISTANCE TO ANY OTHER POINT IS MINIMIZED.

2. THE BULK OF STATISTICAL WORK IS ANALYSIS, AND THIS IS CRUCIAL TO ANSWERING THE DESIGN QUESTION. WHEN STATISTICIANS COMMUNICATE TO EACH OTHER, THE FOLLOWING TREE-LIKE DESCRIPTION IS OFTEN IMPLICIT.

3. DATA TYPE 1 IS THE DATA A SAMPLE FROM A WELL DEFINED LARGER POPULATION, IF SO WHAT KIND OF SAMPLE (SIMPLE, PROPORTIONAL (STRATIFIED) TO A COVARIATE)... . IF NOT, WE CALL THE DATA SET A SAMPLE OF CONVENIENCE OR A BATCH AND WORRY ABOUT THE VALUE OF OUR ANALYSIS WHEN APPLIED TO ANOTHER POPULATION. TODAY, ~~THE~~ DATA ~~ARE~~ OFTEN ARISES AS THE OUTPUT OF A COMPLEX COMPUTER SIMULATION; AT ANOTHER EXTREME, THE DATA MAY BE A COMPLETE ENUMERATION OF A FINITE POPULATION
4. DATA TYPE 2 THE DATA MAY BE UNIVARIATE LENGTHS OF A SAMPLE OF FISH, BIVARIATE (X_i, Y_i) HEIGHT AND WEIGHT FOR EXAMPLE, OR MULTIVARIATE. IT MAY BE A TIME SERIES (STOCK PRICES) OR DATA ALONG THE

GENOME, IT MAY BE FROM A SPATIAL PROCESS. (LOCATIONS OF INCIDENCE OF CANCER ON A MAP).

4. DATA TYPE 3 WE OFTEN HAVE DIFFERENT ANALYTICAL TOOLS FOR CONTINUOUS DATA VERSUS DISCRETE DATA (CONTINGENCY TABLES). IF EACH IS PRESENT WE TALK ABOUT MIXED DATA. NOWADAYS WE ALSO SEE DATA IN THE FORM OF CURVES (eg. GAIT ANALYSIS, GROWTH CURVES OF CHILDREN) AND DATA IN THE FORM OF RANKINGS OR PHYLOGENETIC TREES!

CLASSICAL.

5. STATISTICAL TASKS THERE ARE THREE CLASSICAL STATISTICAL TASKS: ESTIMATION (WHERE IS THIS BUTCH OF DATA CENTERED) TESTING (DOES THE BELL-CURVE MODEL FIT THE DATA) AND REGRESSION GIVEN SOME TRAINING DATA $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$ FIND A FUNCTION $f(X)$ WHICH ~~DOES~~ WILL MAKE $f(X_i)$ CLOSE TO Y_i. THIS FUNCTION MAY BE USED FOR PREDICTION OR FOR UNDERSTANDING (~~DOES~~ WILL INCREASING MONEY TO SCHOOLS AFFECT STUDENT PERFORMANCE).

6. MODERN STATISTICAL TASKS THE PURPOSE OF INVESTIGATION MAY BE JUST TO GET A HUMANELY INTERPRETABLE DESCRIPTION OF THE DATA AND ITS RELATIONS. THIS IS OFTEN CALLED EXPLORATORY DATA ANALYSIS (EDA). AND IS OFTEN CENTERED AROUND GRAPHICAL DISPLAYS, WAYS OF VISUALIZING IN HIGH DIMENSIONS. A RECENT HYBRID WHICH IS WIDELY PRACTICED IS MODEL BUILDING; THIS IS THE ITERATIVE CYCLE OF FITTING A PRELIMINARY MODEL (BY ESTIMATING A CURVE OR SOME PARAMETERS) CHECKING ITS GOODNESS OF FIT (BY TESTING OR GRAPHICAL ANALYSIS OF RESIDUALS) CHANGING THE MODEL TO INCORPORATE HIGHER ORDER APPROXIMATIONS, AND SO ON.

7. TYPES OF MODELS THESE MAY BE PARAMETRIC (eg. FIT A GAMMA DISTRIBUTION $\Gamma(a)\sigma^{a-1} \cdot x^{a-1} e^{-\sigma x}$ ON $[0, \infty)$) NOW PARAMETRIC

(FIND AN ESTIMATE FOR THE BEST FITTING CURVE WITH $\int (f'')^2 \leq 10$) OR SEMI-PARAMETRIC (ESTIMATE THE BEST MONOTONE FUNCTION OR THE BEST SYMMETRIC DENSITY (THE CENTER OF SYMMETRY BEING THE PARAMETER).

8. INFERENTIAL MODE. FREQUENTIST ANALYSIS TRYS TO CHOOSE PROCEDURES WHICH WILL DO WELL IF MANY INSTANCES OF THE PROBLEM UNDER INDEPENDENT REPETITIONS. FOR EXAMPLE, THE AVERAGE VALUE MINIMIZES THE AVERAGE SQUARED ERROR ~~OVER~~ ABOUT THE TRUE MEAN (FOR UNIVARIATE DATA). BAYESIAN ANALYSIS INVOLVES PUTTING A PRIOR DISTRIBUTION OVER ANY UNKNOWN PARAMETERS AND CHOOSES PARAMETER ESTIMATES TO MINIMIZE THE BAYES RISK. THUS, IN A BAYESIAN SETUP, ONE HAS A FAMILY OF PROBABILITIES P_θ AND A PRIOR DISTRIBUTION $\pi(\theta)$; TO ESTIMATE θ $\hat{\theta}^{BAYES}(x)$ IT IS USUAL TO SPECIFY A LOSS FUNCTION $L(\theta, \hat{\theta}(x))$. WE CHOOSE AN ESTIMATE TO MINIMIZE $\int L(\theta, \hat{\theta}(x)) P_\theta(x) \pi(\theta) d\theta$. THIS IS OFTEN THE MEAN OR MODE OF THE POSTERIOR DISTRIBUTION $\pi(\theta|x)$ GIVEN BY BAYES THEOREM.

I WANT TO MENTION TWO OTHER INFERENTIAL MODES: PERTURBATIONAL -- WILL A SMALL CHANGE IN THE DATA EFFECT MY CONCLUSIONS SUSAN HOLMES THINKS THE BOOTSTRAP ANALYSIS IS MOSTLY USED FROM THIS POINT OF VIEW. FINALLY, THERE IS THE ROBUSTIST VIEW: THEORY TELLS US THAT A MEDIUM TO LARGE AMOUNT OF DATA, THE CONCLUSIONS DRAWN FROM BAYESIAN OR FREQUENTIST ANALYSIS WILL AGREE TO GOOD APPROXIMATION. ONE MAY USE A BAYES PROCEDURE TO GET SOMETHING GOOD WITHOUT TAKING THE PRIOR AS A SERIOUS QUANTIFICATION OF APRIORI KNOWLEDGE. ALTERNATIVELY, MANY BAYESIANS CHOOSE THE PARAMETERS IN THEIR PRIORS FROM THE AVAILABLE DATA (EMPIRICAL BAYES).

9. THE ~~ABOVE~~ CATEGORIES ABOVE GIVE A TREE LIKE DESCRIPTION TO STATISTICS; FOLLOWING A PATH GIVES VARIOUS AREAS WITHIN THE FIELD BAYESIAN NON-PARAMETRIC ESTIMATION OF UNIVARIATE TIME SERIES DATA,

10. THERE ARE ~~MANY~~ SOME CONSIDERATIONS LEFT OUT. ONE OF THE BIGGEST AREAS OF STATISTICAL DEVELOPMENT TWENTY YEARS AGO WAS ROBUST ANALYSIS DEVELOPING AUTOMATED WAYS TO DEAL WITH OUTLIERS. AN IMPORTANT SUB-AREA IS MISSING DATA AND SURVIVAL ANALYSIS WHAT TO DO IF SOME DATA ARE MISSING. MATHEMATICAL STATISTICS TRIES TO MAKE MATHEMATICALLY PRECISE CHOICES OF GOOD PROCEDURES ON A GIVEN BRANCH OF THE TREE. IT SUPPOSES GIVEN A FAMILY $P_\theta(x)$ AND A LOSS FUNCTION LC