MATH-338 Midterm 1 Cheat Sheet

THEORY

Day 2: Independent events happen same time, not affecting one another (P(A ∩ B) = P(A)P(B)). Disjoint is the opposite (P(A ∩ B) = 0). Probability Mass Function (PMF) is a dictionary mapping of events to positive probabilities. Over an infinite amount of iterations, RVs converge to a number.

Day 3: Law of Large Numbers: more times means more precise result.

Day 4: Parameter: any numerical quantity that characterizes a given population. Population proportion: a percentage value associated with a population. Sample proportion: the proportion of individuals in a sample sharing a certain trait ($\hat{p}$). Sample Mean($\bar{X}$). Sampling distribution: probability distribution of statistic obtained through a large number of samples drawn (sample **must** be know).

Day 5: We want low bias and high variability. Bias bad. Variability ↓ as the sample size ↑. 'X' denotes the number of successes and 'n' is the number of elements in your sample. $\hat{P}$ does **NOT** have a binomial distribution.

Day 6: Interacting variables: one variable can affect the another variable (non-independent). Confounding variable: a factor that influences the results of an experiment. Block design: split sample initially based on traits (possibly confounding) then randomly assign in those groups. Matched Pairs Design: blocks sizes of two (only looking with two levels). Repeated Measures Design two similar subjects have the same tests and those results are compared. Hawthorne Effect: individuals know they are being experimented on.

Day 7: Sensitivity: proportion of actual positive. Specificity: proportion of actual negative. Positive Predictive Value: proportion of positive tests that were actually positive. Negative Predictive Value: same as above but for negative. Prevalence: base rate. In the tree diagram, sensitivity goes on top and the specificity goes on the bottom.

Day 8: Null Hypothesis: nothing unexpected (original hypothesis, H_0). Alternate Hypothesis: "something is happening and we should change our minds" (H_a). Critical region: range of values that corresponds to the rejection of H_0 at some chosen probability level. Type I Error: occurs when a significance test results in the rejection of a true null hypothesis. Type II Error: the data do not provide strong evidence that the null hypothesis is false. $\alpha < \beta$ and if not, switch hypothesis. $\beta \geq 0.8$. Compute CR: need α, H_0 (value of P under H_0) and sampling distribution of test statistic under H_0. Compute Power: need CR, H_1 (value of P under H_1) and sampling distribution of test statistic under H_1.

Day 9: We want low α and high power. Power analysis steps: define p (proportion in sample), let X be the number of successes, identify H_0 and H_1.

Day 10: One-tailed testing: The critical area of a distribution is either $<$ or $>$ a certain value but not both. Two-tailed the sample is greater than or less than a certain range of values. P-Value: a measure of "strength" of evidence against H_0 (always calculated after observation).

Day 11: Approximate the sampling distribution one of two ways: 1) Under H_0, χ^2 has approximately a χ^2 distribution with (number of categories - 1) ← degrees of freedom | 2) Simulate a lot of times assuming H_0 is true and compute their respective χ^2. When we expected ≤ 5 in each category in our sample, both approaches give similar results. Else, we use method 2. Find case(county), variable(leading digit in diabetes prevalence). You **need** prevalence to calculate the PPV and NPV.

Day 12: Examples will need us to find the probability within a sample population, then use that prevalence to make a more generalized claim for the larger population. P-Value is always above or equal to degrees of freedom.

No context information: A case is an entity of interest. Population is (∀) and sample is (⊂). Distribution of a variable is the relative number of each possible outcome will occur in 'N' trials. Explanatory variable → independent and variable → dependent. Levels of factor

---

FORMULAS

- Mean of Probability Dist. : $\mu_x = \Sigma x \times p(x)$
- Variance : ${\sigma^2}_x = \Sigma[x^2 \times P(x)] - {\mu^2}_x$
- Standard Deviation : $\sigma_x = \sqrt{\sigma_x}$ and $\sigma_{x+y} = \sqrt{\sigma_x + \sigma_y}$
- Number successes : $X \sim B(n, p)$
- Mean of binomial RV: nP
- Variance of Bernoulli RV: $P(1-P)$
- Variance of binomial RV: $nP(1-P)$
- Standard deviation of binomial RV: $\sqrt{nP(1-P)}$
- Bayes' Rule: $\frac{P(B|A)P(A)}{P(B)}$
- P(B|A) = $\frac{number\ of\ outcomes\ in\ A \cap B}{number\ of\ outcomes\ in\ A} = \frac{P(A \cap B)}{P(A)} > 0$
- Independent events: $P(A \cap B) = P(A) \times P(B)$
- Conditional probability: $P(A \cap B) = P(A) \times P(B|A)$ [Tree Mapping]

- Population proportion: $\hat{P} = \frac{X}{n}$
- Variance($\hat{P}$) = $\frac{P(1-P)}{n}$
- Standard Deviation($\hat{P}$) = $\sqrt{\frac{P(1-P)}{n}}$
- Sensitivity: $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$
- PPV: $\frac{TP}{TP+FP}$
- NPV: $\frac{TN}{TN+FN}$
- Prevalence: $\frac{Actual\ Positive}{Actual\ Positive + Actual\ Negative}$
- $\alpha = P(1)$−P(Concluded H_a | H_0 is true)
- Baseline $\alpha = 0.05$
- $\beta = P(2)$−P(Concluded H_0 | H_a)
- Power: $1 - \beta$
- Residual: $\frac{O-E}{\sqrt{E}}$ (O: Observed, E: Expected)
- $\chi^2 = \Sigma\ residual^2 = \frac{(O-E)^2}{E}$
- Degrees of freedom = $(r-1)(c-1)$. r = rows, c = columns.

TESTING FRAMEWORKS

Null Hypothesis Significance Testing: an experimental factor is tested against a hypothesis of no effect or no relationship based on a given observation. We start off assuming H_0 is true. Evidence is then collected and analyzed. An assessment is made upon those findings. If our significance level is breached, then we can reject H_0. Neyman-Pearson: will allow us to make preemptive decisions based on conditions presented before the study is conducted. These are the theoretical outcomes WITHOUT taking any sample data. Fisher Significance Hypothesis Testing: More concerned with model design rather than actual data collection/analysis. Interested in when/why the test failed to make a more efficient model. Goodness of Fit: how well did the data fit with the observations. Binomial Probability Distribution Conditions: Binary outcome (TF), Independent (previous outcomes do **not** affect next.), Number of outcomes, Success is equally likely. Test of Homogeneity: Is the variable's distribution the same in all populations (we initially assume it is and we consider the population to be the explanatory variable). Test of Independence: used to determine if there is a significant association between two variables.
