Probability & Statistics Exam 3 Note Sheets

Chapter 6: Point Estimation

Empirical Rule. Approximately 68%, 95%, and 99.7% live within 1, 2, and 3 standard deviations of the mean if the data is roughly bell shaped.

Order Statistics: Suppose X is a continuous random variable with CDF $F(x)$ and pdf $f(x)$ on the interval $x \in (a,b)$ and we select a random sample of size n. Then, the random variables $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics; that is, Y_1 is the smallest of X_i, Y_2 is the second smallest, Y_n is the largest.

The CDF of Y_r is: $G_r(y) = P(Y_r \le y) = \sum_{k=r}^{n} \binom{n}{k}[F(y)^k][1-F(y)]^{n-k}$.
The pdf of Y_r is: $g_r(y) = P(Y_r = y) = \frac{n!}{(r-1)!(n-r)!}[F(y)^{r-1}][1-F(y)]^{n-r} f(y)$.

Maximum Likelihood: Find $L(\theta) = \sum_{i=1}^{n} f(x_i,\theta)$ where $f(x_i,\theta)$ is from the distribution that you are looking for. Then, find $\frac{d}{d\theta}L'(\theta) = 0$.

Sample Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. **S. Variance:** $s^2 = \frac{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)}{n-1}$.

Examples:

1. Suppose that X has CDF $F(x) = x^2$ for $0 < x < 1$. Let $n = 6$ and Y_r be the 6 order statistics. Find $P(Y_4 < 0.7)$. (We can either use CDF as is, or use the pdf and integrate from 0 to 1. If asked for $E(Y)$, use pdf and add a y.)
   (a) *CDF Version:* $\displaystyle\sum_{k=4}^{6} {}_6C_k(.7^2)^k(1-.7^2)^{6-k}$.
   (b) *pdf Version:* $\displaystyle\int_0^{.7} \frac{6!}{3!2!}(y^2)^3(1-y^2)^2(2y)\,dy$.
2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta) = (1/\theta^2)xe^{-x\theta}$, for $0 < x < \infty$ and $0 < \theta < \infty$. Find $\hat{\theta}$.
   Solution. $L(\theta) = \displaystyle\prod_{i=1}^{n}(1/\theta^2)xe^{-x/\theta} = \left(\prod_{i=1}^{n} 1/\theta^2\right)\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} e^{-x_i/\theta}\right) \Rightarrow$
   $\ln(L(\theta)) = -2n\ln(\theta) + \sum \ln(x_i) - \frac{1}{\theta}\sum x_i \Rightarrow \ln(L'(\theta)) = \frac{-2n}{\theta} + \frac{1}{\theta^2}\sum x_i$
   $\Rightarrow 0 = \ln(L'(\theta)) \Rightarrow \frac{2n}{\theta} = \frac{1}{\theta^2}\sum x_i \Rightarrow \theta = \left(\sum x_i\right)/2n \Rightarrow \theta = \overline{x}/2$.
3. Suppose $X \sim N(\mu, 1)$. Find $\hat{\mu}$.
   Solution. $L(\mu) = \displaystyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right)$
   $\Rightarrow \ln(L(\mu)) = \sum \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right)\right) = \sum\left(\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2}\right)$
   $\Rightarrow \frac{d}{d\mu}\ln(L(\mu)) = \sum(x_i - \mu) \Rightarrow 0 = n\overline{x} - n\mu \Rightarrow \mu = \overline{x}$.
4. Consider the data given by 6, 10, 14, 16, 18, 18, 18, 20. (Store in L_1 and run `1-Var Stats`)
   (a) Find the sample mean, $\overline{x}$: *Solution.* $1/8(6 + 10 + \cdots + 20) = 15$.
   (b) Find the sample standard deviation, s:
   Solution: $s^2 = \frac{1}{(8-1)}(6^2 + \cdots + 20^2 - 8(15)^2) = \frac{170}{7} \implies s = 4.78$

Chapter 8: Tests of Statistical Hypotheses

For the confidence interval, if one tailed, use $CI = 1 - 2\alpha$.

Examples:

1. You flip a coin 300 times and get 128 heads. Use the calculator to build a 99% confidence interval for the proportion of heads this coin produces. What can you say about the fairness of the coin? ***Solution.*** Using `1-PropZInt` with $x = 128$ and $n = 300$, we get $(0.353, 0.500)$ as our 99% confidence interval. Since the value 0.5 is included within the 99% CI, we cannot conclude that the coin is unfair at $\alpha = 0.1$.
2. Week-old ducklings have a mean weight of 423 g. You are concerned about a new industrial plant upstream from your favorite pond, and so weigh a random sample of $n = 43$ week-old ducklings. You find $\overline{x} = 413.7$ g and $s = 33.9$. At $\alpha = 0.05$, is there evidence that the plant is causing the ducks to be underweight? Use the NHST steps. ***Solution.*** **Population:** Week-old ducklings in the pond. **Tailed:** one. **Hypotheses:** $H_1 : \mu \ge 423$ – They weight more than 423; H_0 : They weigh less. **Test:** 1-sample t-test: $t_{\text{samp}} = -1.799$, $p = 0.04$. **Decision:** Since $p < \alpha$, we reject the null hypothesis. **CI:** 90% : $(405, 422.4)$. **Effect Size:** $d = \frac{\overline{x}-\mu_0}{s} = -0.247$, small effect. **Conclusion:** These data lead us to conclude that the industrial plant upstream has a small negative effect upon week-old ducklings in the pond.
3. Prof. Seme gives the same 5 NHST questions to his MATH 215 and MATH 310 students, and assigns each student a grade. We will treat each as a random sample of all Intro Stats and Math-Major-Adjacent Stats students. Using the data in the table below, is there evidence, at $\alpha = 0.01$, that there is a difference in performance between the two groups? Use the NHST steps.

	MATH 215	MATH 310
Mean Score	74.7	81.2
Sample St. Dev.	12.3	9.1
Count	35	23

   Solution. **Population:** All students in MATH 215 and all students in MATH 310. **Tailed:** two. **Hypotheses:** $H_0 : \mu_{215} = \mu_{310}$ – no difference; $H_1 : \mu_{215} \ne \mu_{310}$, difference. **Test:** 2-sample t-test with not equal variance: $t_{\text{samp}} = -2.309$, $p = 0.02$. **Decision:** fail to reject the null hypothesis. **CI:** 99% : $(-14, 1)$, which we note includes 0. **Effect Size:** (Find s_p, and then use that to find d). **Conclusion:** While the sample data shows a medium effect size, we do not have enough evidence at $\alpha = 0.01$ to support that there is a difference in the populations.
4. Determine if there is a difference in the variance between the two groups using the data from the previous problem. ***Solution.*** **Population:** same as before. **Tailed:** two (keyword *difference*). **Hypotheses:** $H_0 : \sigma_{215}^2 = \sigma_{310}^2$ – The two variances are the same; $H_1 : \sigma_{215}^2 \ne \sigma_{310}^2$. **Test:** We will use a 2-sample F-test: $F_{\text{samp}} = 1.827$, $p = 0.14$. **Decision:** fail to reject. **Conclusion:** We do not have evidence to suggest that the variances are different.

Chapter 8 (cont.)

Examples:

1. A professor wonders if students are more likely to be absent on Friday in the 3-day-a-week class than Monday or Wednesday (i.e. not exactly equal across all days). The professor uses a single course one semester as though it is a random sample, and counts that of 100 total absences, 41 were on a Friday. Is there evidence, at $\alpha = 0.05$ to support the professor's suspicion about Fridays? Use the NHST steps. ***Solution.*** **Population:** All absences in the semester. **Tailed:** one. **Hypotheses:** $H_0 : p = 1/3$ – Friday absences occur at the expected rate; H_1 : They don't. **Test:** 1-proportion z-test: $z_{\text{samp}} = 1.63$, $p = 0.052$ **Decision:** fail to reject. **CI:** 90% : $(0.329, 0.491)$, which contains 1/3. **ES:** $d = 2\arcsin\left(\sqrt{\hat{p}}\right) - 2\arcsin\left(\sqrt{p_0}\right) = 0.159$, small. **Conclusion:** We conclude that there is not sufficient evidence at $\alpha = 0.05$ to support the claim that absences are more likely on Fridays, though the result is borderline. Additionally, even if they were significant, our result would would have a small effect.
2. A random sample of Hendrix and Ozarks students shows that 25 of 48 Hendrix students have taken a statistics class in college and 17 of 50 Ozarks students have. Is there evidence, at $\alpha = 0.1$ that there is a difference in the proportion of students who take stats at the two schools? Use the NHST steps. ***Solution.*** **Population:** All Hendrix students and all Ozarks students. **Tailed:** 2. **Hypotheses:** – There is no difference between the proportion of Ozarks students that take statistics and that of Hendrix students; H_1 : is a difference. **Test:** two sample z-test: $z_{\text{samp}} = 1.81$, $p = 0.07$. **Decision:** Reject. **CI:** 90% : $(0.18, 0.34)$. **ES:** $d = 2\arcsin\left(\sqrt{\hat{p}_1}\right) - 2\arcsin\left(\sqrt{\hat{p}_2}\right) = .37$; small to medium. **Conclusion:** We have evidence to support the claim that there is a difference between the proportion of students who take stats at Ozarks and Hendrix. Additionally, our effect size tell us that our result is small to medium.

Chapter 9: Additional Tests

GOF: Store observed data in `L1`, and expected in `L2`. *df* is found by subtracting 1 from the number of categories. The hypotheses are H_0 : the sample is drawn from a population with the claimed distribution, or H_1 : from a different distribution.
Test for Independence: Store the contingency table in matrix `[A]`, then use `X2-Test`. The hypotheses are H_0 : the two categories are independent, H_1 : they are dependent (some relationship exists between them).

Examples:

1. You are interested in whether there is a relationship between a student's major area, and the amount of sleep they get. You survey students and find the data in the table below. What can you say, at $\alpha = 0.05$? Use the NHST.

Sleep	Human.	Nat. Sci.	Soc. Sci.	Inter.
8+ Hours	12	34	36	9
6.5 - 7.9 Hours	17	31	42	11
6.4- Hours	13	29	43	12

Chapter 9: (cont.)

Examples:

Solution. (for 1) **Population:** All college students. **Hypotheses:** H_0 : Sleep amount and major are independent; H_1 : they are dependent. **Test:** χ^2-test of independence: $\chi^2 = ...$ $p = 0.91$. **Decision:** fail to reject. **Conclusion:** We lack evidence to support the claim that there is a relationship between the amount of sleep a student gets and their major.

2. It is known that, over the past 10 years, 50% of Hendrix students come from AR, 29% from TX, 10% from OK, 5% from MO, 3% from TN, and 3% elsewhere. A survey of 120 students who have recently taken Calculus I shows that 54 are from AR, 23 from TX, 17 from OK, 8 from MO, 8 from TN, and 10 elsewhere. Is there evidence, at $\alpha = 0.05$, that the distribution of students who take Calculus I is different from the overall Hendrix population? *Solution:* Population is all Hendrix Students. We will use the χ^2-GOF-Test. From the given data, we have the following table:

	AR	**TX**	**OK**	**MO**	**TN**	**Else.**
Obs.	54	23	17	8	8	10
Exp.	60	34.8	12	6	3.6	3.6

I got $\chi^2_{\text{samp}} = 24.107$, $p = 0$. Since $p < \alpha$, we **reject** H_0 and conclude that there is a difference between the students who take Calculus I and the overall Hendrix population.

3. You want to know if Hendrix students have differences in their average number of M&Ms eaten per week, depending on their home state. You run an ANOVA, at $\alpha = 0.05$ and find the following table:

ANOVA

Source of Variation	SS	df	MS	F	P-value
Between Groups	20975.92	4	5243.98	23.31	1.2879E-14
Within Groups	29921.96	133	224.98		
Total	50897.89	137			

Solution. **Population:** All Hendrix students. **Hypotheses:** H_0: $\mu_{AR} = \mu_{LA} = \mu_{MO} = \mu_{OK} = \mu_{TX}$ – The mean number of M&Ms eaten are the same among all groups; H_1 : There is at least one group that has a different weekly mean M&Ms eaten. We see that $F_{\text{samp}} = 23.31$, $p = 1.29\text{E}\text{ -}14$. **Decision:** reject H_0. **Effect Size:** $\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} = \frac{20975.92}{50897.89} = 0.412$; we have a very large effect size ($\eta^2 = 0.01, 0.06, 0.14$: small, medium, large, respectively). **Conclusion:** Thus, we have found a strong relationship between home state and M&M consumption. It appears that the home state and candy eating habits are strongly related.