Name: ________________________
ID: __________________________

8 + 28

Stat 421, Test 1, Fall, Oct. 19, 2015; Marzban

ONLY a half-size "cheat sheet" is allowed

Multiple choice: Circle all the correct answers; there is wrong-answer penalty

For rest, SHOW answer & work; NO CREDIT for correct answer without explanation

Points

Lect 2

∼ 1 **1.** Comparative boxplots are useful for dealing with
a) 1 continuous variable and 1 categorical variable.
b) 1 continuous variable and multiple categorical variables.
c) multiple/different continuous variables (e.g., height and weight).
d) None of the above.

Lect 2

∼ 1 **2.** When comparing two boxplots, it is reasonable to conclude that the true medians of the two populations are different if/when
a) the median of one is larger than the median of the other.
b) the 5 numbers in one boxplot (min, 1st quartile, median, 3rd quartile, max) are **respectively** larger than the 5 numbers in the other.
c) If the spread of one boxplot (e.g., 3rd quartile - 1st quartile) is larger than that of the other.
d) None of the above.

Lect 2,3

2 **3.** Circle all of the quantities for which one can build a sampling distribution.
a) sample median b) sample standard deviation c) sample proportion d) sample quantile.

4. Circle the correct answer. Suppose $X \sim$ uniform over [2,4]. Let M denote the sample maximum.
a) For a sample of size $n = 2$, do you expect the mean of the sampling distribution of M to be
∼ 1 closer to 2 or to 4?
b) Which do you expect to be larger, the mean of the sampling distribution of M for samples of
∼ 1 size $n = 2$ or for samples of size $n = 100$?
c) Will the variance of the sampling distribution of M for samples of size $n = 2$ be the same or
∼ 1 different from the variance of the sampling distribution of M for samples of size $n = 100$?

Lect 7

5. Which of the following is/are NOT appropriate for a 1-way ANOVA? We want to know
a) about the mean tomato growth under three different lighting conditions.
1 b) if there is a difference in the time it takes for 5 different computers to execute a given code.
c) if 3 different drugs have an effect on the proportion of men with hair-loss problems.
d) about the number of return visits into a website with three different advertisements.

hw-M

∼ 2 **6.** What is the value of the y-intercept in the power curve, if one is testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$? Just answer. No explanations. α

∼ 2 **7.** If data on x and y are paired, and we are trying to see if there is a difference between the true mean of x and the true mean of y, does it make sense to compare the two boxplots? Yes/No? Why?

No, because each obs of x may be larger than each obs of y, which means that $\mu_x > \mu_y$. But there may still be a lot of overlap between the boxplots, in which case we will conclude we don't know.

Moral: So, boxplots should not be used on paired data. Instead, one should look at the boxplot of differences.

~ 3 hw-L **8.** Jane and Alice are planning on making measurements of the mass of the electron. Jane aims to use 121 measurements. Assuming that both Jane and Alice are using equipment which are known to have the same precision, what sample size should Alice use if she wants the variance of her measurements to be less than half the variance of Jane's measurements, with probability 10%? You can make all the usual normality assumptions.

$0.1 = \text{prob}\left(s_A^2 < \frac{1}{2} s_J^2\right) = \text{prob}\left(\left(\frac{s_J}{s_A}\right)^2 > 2\right) \sim F_{n_J-1,\, n_A-1}$ if $\sigma_J = \sigma_A$

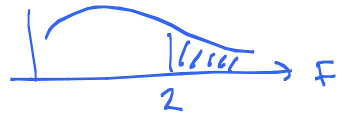
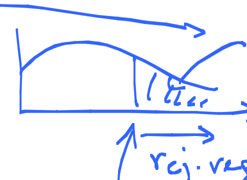


Table IV (p.617) for $\alpha = 0.1$

with $df_J = 120 \Rightarrow df_A = 11 \Rightarrow \boxed{n_A = 12}$

~ ~~4~~ 3 hw-J,M **9.** Suppose we are testing $H_0 : \sigma^2 \leq \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$, at level $\alpha = 0.05$, where $\sigma_0^2 = 15$. What is the power of the test if the sample size is $n = 16$, and the true variance is in fact 25? You may use the Tables to get an approximate answer.



First, find critical region:

$0.05 = \text{prob}(s^2 > s_c^2 \mid H_0 = T) \leftarrow \sigma^2 = \sigma_0^2$

$\therefore\ 0.05 = \text{prob}\left(\frac{(n-1)s^2}{\sigma_0^2} > \frac{(n-1)s_c^2}{\sigma_0^2} \mid H_0 = T\right)$

$= \text{prob}\left(\chi^2 > \underbrace{\frac{\cancel{15}\, s_c^2}{\cancel{15}}}_{25}\right) \Rightarrow \chi_c^2 = 25 \Rightarrow s_c^2 = 25$ $\leftarrow \chi^2$ table

$\frac{(n-1)s^2}{\sigma^2}$ X s^2

$\frac{(n-1)s_c^2}{\sigma^2}$ X s_c^2

Correction

Then, find power:

$\text{power} = pr(s^2 > s_c^2 \mid H_1 = T) = pr\left(\frac{(n-1)s^2}{\sigma^2} > \frac{15(\cancel{25})}{\cancel{25}}\right) = pr(\chi^2 > 15) \simeq 0.5$

$\sigma^2 = 25$ standardize

Table: $pr(\chi^2 > 14.34) = 0.5$

~ 3 Lect 3 **10.** Suppose X has a pdf given by $f_X(t) = \lambda e^{-\lambda t}$, $t > 0$. It can be shown that the pdf for $Y = X^2$ is given by $2\lambda t e^{-\lambda t^2}$, $t > 0$. **Starting from the integral defn.** of $E[.]$, show that $E_X[\sqrt{x}] = E_Y[Y]$.

$E_X[\sqrt{x}] = \int_0^\infty \sqrt{t}\, f_X(t)\, dt = \int_0^\infty \sqrt{t}\, \lambda e^{-\lambda t}\, dt$ $\leftarrow =$

$E_Y[Y] = \int_0^\infty t\, f_Y(t)\, dt = \int_0^\infty t\, 2\lambda t\, e^{-\lambda t^2}\, dt = \int_0^\infty 2\lambda u\, e^{-\lambda u} \frac{du}{2\sqrt{u}} = \int_0^\infty \sqrt{u}\, \lambda e^{-\lambda u}\, du$

$u = t^2,\ du = 2t\,dt$

~ 3 hw-C **11.** Suppose we know that $E_{X,Y}[X+Y] = E_X[X] + E_Y[Y]$. **Starting from the defining integral** of $Cov[.,.]$, show that $Cov[X, X+Y] = V_X[X] + Cov[X,Y]$. You may NOT assume that X and Y are independent.

$Cov[X, X+Y] = \int (t_1 - E_X[X])\left(t_1 + t_2 - \underbrace{E_{X,Y}[X+Y]}_{E_X[X] + E_Y[Y]}\right) f_{X,Y}(t_1,t_2)\, dt_1\, dt_2$

$= \iint (t_1 - E_X[X])^2 f_{X,Y}(t_1,t_2)\, dt_1\, dt_2 + \int (t_1 - E_X[X])(t_2 - E_Y[Y]) f_{X,Y}(t_1,t_2)\, dt_1\, dt_2$

$= \int (t_1 - E_X[X])^2 f_X(t_1)\, dt_1 + Cov[X,Y]$

$= V_X[X] + Cov[X,Y]$

~ 3 Lect 6 **12.** It is known that the sample standard deviation has an approximately normal sampling distribution with mean and variance paramaters given by σ_y and $\frac{\sigma_y^2}{2n}$, where σ_y is the population standard deviation, and n is the sample size. Starting from this fact, build a 2-sided CI for σ_y.

$S \sim N(\sigma_y, \sigma_y^2/2n) \Rightarrow Z = \frac{S-\sigma_y}{\sigma_y/\sqrt{2n}} \sim N(0,1) \rightarrow pr\left(1 + \frac{Z_{-\alpha/2}}{\sqrt{2n}} < \frac{S}{\sigma_y} < 1 + \frac{Z_{\alpha/2}}{\sqrt{2n}}\right)$

$prob(Z_{-\frac{\alpha}{2}} < Z < Z_{\alpha/2}) = 1-\alpha$

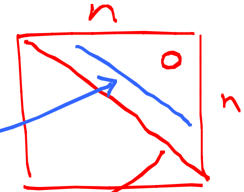
$pr\left(Z_{-\alpha/2} < \frac{S-\sigma_y}{\sigma_y/\sqrt{2n}} < Z_{\alpha/2}\right) = 1-\alpha$

$pr\left(Z_{-\frac{\alpha}{2}}/\sqrt{2n} < \frac{S}{\sigma_y} - 1 < Z_{\alpha/2}/\sqrt{2n}\right) = 1-\alpha$

C.I. σ_y: $\frac{S}{1+\frac{Z_{\pm\alpha/2}}{\sqrt{2n}}}$

~ 2 Sample Test **13.** Consider a sample of size n. We have shown that if the individual elements of the sample are drawn independently, then the variance of the sample mean is σ^2/n, where σ is the standard deviation of the population. Now, suppose the individual samples are not independent; they are drawn sequentially (one at a time) and every element has covariance K with the element "prior" to it. Derive the expression for the variance of the sample mean.

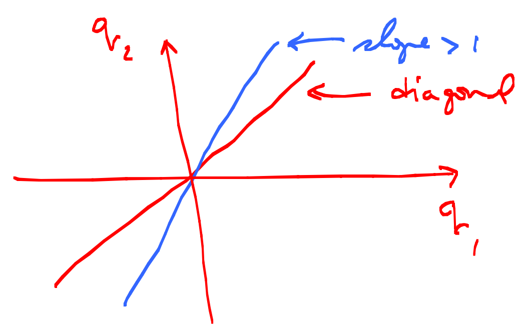
$V[\bar{Y}] = V\left[\frac{1}{n}\sum_i Y_i\right] = \frac{1}{n^2}\left(\underbrace{\sum_i V[Y_i]}_{n\sigma^2 (iid)} + 2\underbrace{\sum_{\substack{i,j\\ i>j}} Cov[Y_i, Y_j]}_{\text{only } \frac{n(n-1)}{2} \text{ are non-zero}} \right)$

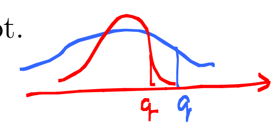


$= \frac{1}{n}\left(\sigma^2 + K - \frac{1}{n}K\right)$

Moral: comparable!

~ ~~2~~ 3 Sample Test **14.** The traditional notion of the qq-plot places quantiles of a theoretical distribution on the x-axis, and the corresponding quantiles of the observed sample on the y-axis. But it is also possible to place quantiles of a different **sample** on the x-axis. Suppose we put quantiles of x_1 on the x-axis, and quantiles of x_2 on the y-axis. Also suppose the distribution of x_2 is wider than that of x_1, but they are both normal. Draw the general shape of the resulting 2-sample qq-plot. No explanations.





~ 4 **15.** Consider the distribution of apples that have fallen onto the ground from an apple tree. Suppose the x and y coordinates of the apples have a normal distribution with zero mean (i.e., centered about the tree), and a standard deviation of σ. What is the probability that a random apply will fall outside a circle of radius $\sqrt{6}\sigma$? Hint: The sum of the square of n standard normal variables has a chi-squared dist. with df = n. Show work. r = radius

$pr(r > \sqrt{6}\sigma) = pr(r^2 > 6\sigma^2) = pr(X^2 + Y^2 > 6\sigma^2)$

standardize

$= pr\left(\left(\frac{X-0}{\sigma}\right)^2 + \left(\frac{Y-0}{\sigma}\right)^2 > 6\right) = pr(Z_x^2 + Z_y^2 > 6)$

$= pr(\chi^2_{df=2} > 6) = \underline{0.05}$

χ^2 Table.