STA 113 Spring 2005
Final Exam

NAME

Calculator and one 8.5x11 sheet of notes allowed.

Useful Numbers: $\Phi^{-1}(.99) = z_{.01} = 2.33, z_{.025} = 1.96, z_{.05} = 1.65, z_{.10} = 1.28$

Circle True or False.

1. Assume the null and alternative hypotheses for a statistical test are $H_O : \mu = 15$ mg *versus* $H_a : \mu < 15$ mg. We decide to do the usual one-sided t-test with test statistic is $\frac{\bar{x}-15}{s/\sqrt{10}}$ at level $\alpha = .01$. We have $n = 10$ observations.

   (a) F If the p-value is greater than .01, then we reject H_0.

   (b) T The test statistic has the t-distribution with 9 degrees of freedom, if $\mu = 15$.

   (c) T The p-value is the area under the $t(9)$ probability density to the left of the observed value of the test statistic.

   (d) F The two-sided confidence interval will contain 15 if and only if the one-sided t-test accepts H_0.

   (e) F The critical value $t_{.005,9}$ is smaller than the value $z_{.005}$.

   (f) T The probability of the Type I error of the test is .01.

   (g) F The probability of the Type II error of the test is .01.

Give complete solutions and show your work for problems 2.

2. Suppose a point (X, Y) (in rectangular coordinates) is uniformly distributed throughout the interior of the unit disk (radius 1).

   (a) Compute the marginal density $f_X(x)$ for each $x \in [-1, 1]$.

   (b) To compute $E(|X|^{2/3})$, we are going to use a Monte Carlo method. We generate n points in the disk uniformly, then compute $|x_1|^{2/3}, \ldots, |x_n|^{2/3}$ for each, and then use the sample mean as the value. If we want the sample mean to be within 0.01 of the true value with confidence 99%, what sample size n is necessary? A good approximation is sufficient.

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} 1/\pi \; dy = \frac{2\sqrt{1-x^2}}{\pi}, \;\; x \in [-1, 1].$$

We would get the desired accuracy by the Central Limit Theorem if $z_{.005}\sigma/\sqrt{n} \leq .01$, where σ is the standard deviation of $|X|^{2/3}$. Since $|X|^{2/3} \in [0, 1]$, $\sigma \in (0, 1)$ so an upper bound of 1 can be used to give a sufficient size of $n \geq 66,564$.

To get σ, we need $\texttt{Var}(|X|^{2/3}) = E(|X|^{4/3}) - E(|X|^{2/3})^2$, two difficult integrals. One finds that σ is less than .25, which would give a sufficient sample size of around 4,150.

For problems 3-16, choose the best answer.

3. Suppose the lifetime T of a hard disk drive has probability density $f(x)$ proportional to $x^3 e^{-x}$ for $x > 0$, and $f(x) = 0$ when $x \leq 0$. Then the sample mean $\bar{T}$ of a random sample of 100 such lifetimes has standard deviation $\sigma_{\bar{T}}$ equal to

   (a) 0.30

   (b) 300.0

   (c) 0.20 X

   (d) 2.0

   (e) none of the above

4. Assume $X_1, X_2, X_3, X_4, X_5, X_6$ is a random sample from some unknown distribution with mean $\mu = 10$ months, $\sigma = 2$ months. If each represents the lifetime of a single battery, then $T = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$ is the total lifetime when they are used in sequence. Find approximately the 5^{th} percentile of the distribution of T, the time above which T falls with probability .95.

   (a) 40.3 months

   (b) 51.9 months X

   (c) 54.3 months

   (d) 65.7 months

   (e) none of the above

5. Suppose two random variables X, Y have means $\mu_X = 0.01$ and $\mu_Y = 0.02$ (mm), standard deviations $\sigma_X = 0.01, \sigma_Y = 0.03$ (mm), and correlation $\rho_{X,Y} = .40$. Find the standard deviation of $\frac{1}{3}X + \frac{2}{3}Y$.

   (a) 0.020
   (b) 0.021
   (c) 0.022 X
   (d) 0.023
   (e) none of the above

6. A random sample of $n = 9$ fiber test specimens gave data $x_1, \ldots, x_9$ with sample mean 30.2 kg and sample standard deviation of 5.5. Assuming the data is a random sample from some Normal distribution, compute a 95% confidence interval for the mean μ.

   (a) (26.0, 34.4) X
   (b) (26.8, 33.6)
   (c) (26.6, 33.8)
   (d) (27.2, 33.2)
   (e) none of the above

7. Two independent samples $x_1, \ldots, x_{10}$ and $y_1, \ldots, y_{20}$ from Normal distributions gave $\bar{x} = 1.1, s_x = .52, \bar{y} = 2.3, s_y = 1.5$ in units of seconds. The degrees of freedom for comparing the means μ_x and μ_y in a two sample test is:

   (a) 30

   (b) 28

   (c) 26 X

   (d) 20

   (e) 15

8. Two independent samples $x_1, \ldots, x_{10}$ and $y_1, \ldots, y_{10}$ from Normal distributions gave $\bar{x} = 1.0, s_x = 1.0, \bar{y} = 2.0, s_y = 1.0$ in units of seconds. The p-value for testing $H_0 : \mu_x = \mu_y$ *versus* $H_a : \mu_x \neq \mu_y$ is:

   (a) The area to the left of -2.24 under the $t(18)$ density.

   (b) The combined area to the left of -2.24 and to the right of 2.24 under the $t(18)$ density. X

   (c) The area to the left of -2.24 under the $t(20)$ density.

   (d) The combined area to the left of -2.24 and to the right of 2.24 under the Standard Normal density.

   (e) None of the above.

9. Three 1 Gbyte hard drives store 2 Gb of data as follows: the first Gb is on disk 1, the second Gb is on disk 2, and on disk 3 the xor of the bits on the first two drives is stored (the xor is 1 if the other two are both the same, 0 otherwise, so if you know two of the three values you can get the third). If disks fail and lose data independently with probability p=.001, find the probability that the total RAID configuration loses data.

   (a) 3×10^{-6} X
   (b) 3×10^{-3}
   (c) 1×10^{-3}
   (d) 1×10^{-6}
   (e) none of the above

10. A teacher wants to estimate the proportion p of students at Duke who would answer "yes" to the question:
    Q_1 : `Should there be a limit on the proportion of computers that run Microsoft Windows on campus?`
    To get around possible intimidation with a direct response in class, a second question is formulated:
    Q_2 : `Are you a senior at this time in the University?`
    The instructions were: "Each of you flip a coin, and if it's heads answer Q_1 , if it's tails answer Q_2 but don't tell me which you answer." Assuming students now answer truthfully because it is not clear to others which question they are answering, estimate the proportion p if 45 out of 100 students turn in "yes" responses. Also, the teacher knows that there are 60 seniors among the 100 students in the class.

    (a) $\hat{p} = 0.50$
    (b) $\hat{p} = 0.45$
    (c) $\hat{p} = 0.30$ X
    (d) $\hat{p} = 0.15$
    (e) none of the above

11. Let X be the number of times I get a pair in the pocket (first two cards) in 1000 games of well-shuffled Texas Hold'em. Approximate $P(X \geq 60)$.

    (a) 0.40
    (b) 0.46 X
    (c) 0.52
    (d) 0.58
    (e) none of the above

12. Suppose 10 newly manufactured items show a number of defects described in the following table:

Number of defects per item	0	1	2
Observed frequency	6	3	1

If the number of defects on an item is modeled with the Poisson(λ) distribution, the numerical value of the maximum likelihood estimator for λ is:

    (a) $\hat{\lambda} = 0.5$ X
    (b) $\hat{\lambda} = 5.0$
    (c) $\hat{\lambda} = 1.0$
    (d) $\hat{\lambda} = 10.0$
    (e) none of the above

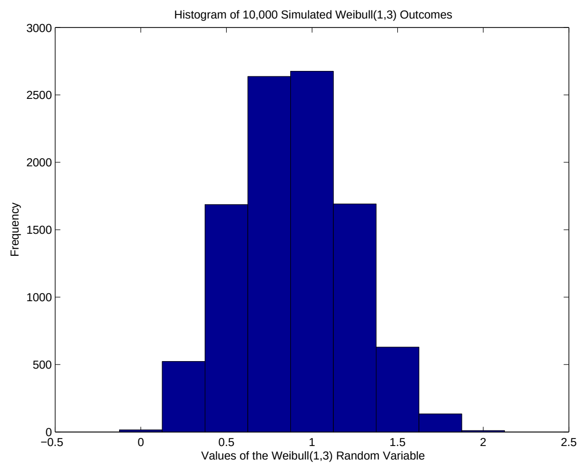
13. The Matlab command to compute the .05 critical value for the t-distribution with 10 degrees of freedom is

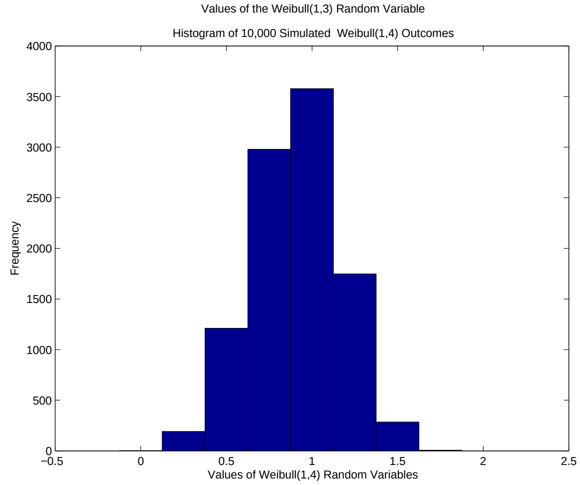
    (a) `tinv(.05,10)`
    (b) `tinv(.95,9)`
    (c) `tcdf(.025,10)`
    (d) `tcdf(.05,10)`
    (e) `tinv(.95,10)` X

14. The Matlab command to generate a column of 10000 pseudo-random numbers from the exponential distribution with rate $\lambda = 2$ is

    (a) `randexp(2,10000)`
    (b) `rndexp(2,10000,1)`
    (c) `exprnd(2,10000,1)`
    (d) `exprnd(.5,10000,1)` X
    (e) `exprnd(.5,10000)`

15. I cannot understand the documentation for the Weibull distribution in Matlab, so I am trying to figure out the meaning of the parameters. I simulate 10,000 Weibull (1,3) outcomes and 10,000 Weibull (1,4) outcomes, whatever the parameters mean, and get the following histograms using exactly the same bins:





Which of the distributions seems to have the smaller standard deviation?

(a) Weibull(1,4) X

(b) Weibull(1,3)

(c) They both have the same standard deviation.

(d) This question cannot be answered from histograms.

16. It is known that 40% of incoming mail is spam. A central mail server can look at basic features of the message to try to classify it as spam or legitimate mail. If a piece of mail is spam, there is a 30% chance the subject line is in all capital letters, and there is a 30% chance 100 or more people on campus received exactly the same message body. On the other hand, if the mail is legitimate, there is a 5% chance the subject line is in all capital letters, and there is a 20% chance 100 or more people on campus received exactly the same message body.

    Suppose a particular piece of mail just received by the mail server 1) has all capital letters in its subject line and 2) fewer than 100 received it on campus. Compute the conditional probability the mail is spam, given these features. (Assume the two features are independent of one another for simplicity.)

    (a) none of the below
    (b) 0.84
    (c) 0.78 X
    (d) 0.21
    (e) 0.08