Assignment 3

This assignment gives experience with interpretation & computation of discrete data statistics, sensitivity and specificity and Gaussian and probability calculations.

1. Read the article by Julian et. al. on the EMIAT trial.

1a. What is the primary outcome?

1b. Did the authors correctly report statistics that took into account time to event outcomes? (note: The Cox model produces hazard rates.)

1c. What do the "risk" ratios in Table 2 imply (ie RR < 1 vs RR=1 vs RR > 1) in terms of mortality in this study? We can assume that the RRs reported are hazard rate ratios. Be sure to interpret this in the context of THIS study, not just what this means in general.

1d. Why did the authors present both intent-to-treat and on-treatment analysis? Is one better than the other? Briefly explain.

1e. Figure 4 shows the percent of patients continuing study medication over time. Does this strengthen or weaken the author's conclusions or have no effect? Is there a possible bias in the results?

2. A new antibiotic is compared to a standard antibiotic for treating bacterial urinary tract infections (UTIs) in a randomized trial. Below are the percentages of patients who are cured after one week.

Treatment Group	sample size	pct cured	pct disease (not cured)
Standard	25	60%	
New	25	72%	

Based on the above, compute

2a. Percent (pct) with disease in each group

2b. Number needed to treat - NNT

2c. Risk ratio – RR of disease in new/standard

2d. Odds ratio – OR of disease in new/standard

2e. Relative risk reduction (RRR)

3. The table below shows the results from a new "electronic chip" test for detecting bacterial urinary tract infections. The data is taken from a prospective study.

	true-no infection	true-has infection
test-negative	950	10
test-positive	50	40
total	1000	50

Based on the above, compute the following statistics

3a. Sensitivity

3b. Specificity

3c. Unweighted accuracy

3d. Positive predictive value (PPV)

3e. Negative predictive value (NPV)

4 The **log** (base 10) of PSA (prostate specific antigen) is known to have bell curve (Gaussian) distribution. One cutoff (threshold) used to diagnose prostate disease (hypertrophy, cancer) on the original scale is a PSA higher than 4 ng/nl (positive if PSA > 4 ng/ml). On the log 10 scale, this cutoff corresponds to $\log_{10}(4)=0.6$ log ng/ml.

Below are population $\log_{10}$PSA means and standard deviations (SD) in healthy and diseased individuals

$\log_{10}$ PSA statistics in log ng/ml

Group	mean=μ	SD=σ
Healthy	-0.092	0.477
Disease	1.009	0.382

4a. Compute the $\log_{10}$ scale 95% prediction intervals in log ng/ml for each group using Gaussian theory and the information above.

4b. Compute the original scale means and 95% prediction intervals for each group in ng/ml. Are these geometric means or arithmetic means?

4c. If the 4 ng/ml cutoff is used, compute the Gaussian theory sensitivity, specificity and accuracy.

4d. Without further calculation, if the threshold cutoff is decreased, would the sensitivity and/or specificity increase, decrease or stay the same?

5. About 72% of the US population is age 18 or older (Census, 2005). Also, about 3% of the US population has been exposed to asbestos. Assuming **independence** between age and asbestos exposure, what is the probability that someone in the US is both under 18 and has been exposed to asbestos? Is this the same as the probability of a person under 18 being exposed to asbestos?

6. Breast tumors that are removed from breast cancer patients can be tested for overexpression of the HER2/new gene. Therefore, we know that about 15% of Caucasian females age 50-59 with breast cancer have HER2/neu gene overexpression (positive for HER2/neu). Also, a representative sample of all Caucasian females 50-59 females shows that 5% have HER2/neu overexpression. And 2% of all Caucasian females age 50-59 have breast cancer. Based on this, **compute the chance that a Caucasian female age 50-59 with HER2/neu overexpression will have breast cancer or explain why this cannot be done**.

7. The number of births per woman during the 10 years of fertility are known to follow a Poisson distribution. Compute the change in the percent of women with 3 or more births if the mean birth rate per woman decreases from 2.5 children to 1.5 children. Can use the formula in the notes or the EXCEL Poisson function - POISSON(X, mean, cum). If cum=0, this function gives the probability of X for the specified mean. If cum=1, this function gives the cumulative probability of X **or less** for the specified mean. For example, POISSON(4, 1.9, 0) gives the probability of X=4 if the mean is 1.9. POISSON(4, 1.9, 1) gives the cumulative probability of 4,3,2,1 and 0 if the mean is 1.9.

8. Three students miss their final exam and ask the instructor for a makeup exam. When asked by their instructor for the reason, they said they were all in the same car and the car had a flat tire. On the makeup exam, in addition to the other questions, the instructor asks, which tire was flat?

Compute the probability that all three give the same answer. Hint, if there are four tires (call them A, B, C and D), how many possible answers could three students give? How many are the same for all three?

JMP assignment

The "PSA" dataset contains prostate specific antigen (PSA) values on normals and those with enlarged prostate disease ("disease").

1. Determine if the data in each group follows a Gaussian distribution better on the original scale or a log scale.

2. Using the best scale, report summary statistics including the mean, SD and range for each group. Is there overlap between the two distributions?

3. Consider a threshold that is halfway between the two means. Using this threshold, compute sensitivity, specificity and accuracy. Hint: Create a variable that is equal to '1" if an individual is above the threshold and is equal to "0" otherwise. Then make a cross tabulation.

4. Compute the Gaussian theoretical sensitivity and specificity using the threshold and the means and SDs reported above. Compare to the results from #3 above. Hint: Compute the Z scores and use the JMP function "NormalDistribution(z)".