

Instructions: This is a three hour exam. It is "open book" meaning any notes or texts are allowed. The marks assigned total 100. Attempt all questions – allocate your time accordingly.

1. (30 marks = 5+5+5+5+5+5) Write an explanatory paragraph **in your own words** on each of the following topics:

a) A small P-value associated with a statistical test of hypothesis suggests that the hypothesis is not credible.

a) The P-value is computed on the assumption that the hypothesis is true, and it is the probability of getting a statistic value as far from its expected value (or further) than was observed. When this probability is small, it indicates that a rare event has occurred, and since the data is what was actually observed, this suggests that the calculation is wrong, and the only thing that can be wrong is the hypothesis assumed. So this suggests the hypothesis is not credible.

b) A discrete time process that mimics a Poisson process can be easily generated using a simulation of a Bernoulli random variable.

b) If we use discrete time intervals to represent a small duration, and set the probability of an event in that small duration to be a small probability (Bernoulli RV), then the time between events will have a geometric duration and it will closely mimic the exponential distribution between events in the continuous time Poisson Process. Clearly, the binomial probability of the number of events occurring in the discrete time approximation to $(0,t)$ will be very close to the Poisson probability of the number of events occurring in $(0,t)$.

c) A 95% Confidence Interval for a population mean is usually wider than a 90% Confidence Interval.

c) A 95% Confidence interval would contain the population mean with probability .95. A 90% Confidence interval would contain the population mean with probability .90. The .05 probability difference must be those instances where the population mean is outside the 90% CI but inside the 95% CI. So the 95% CI is wider than the 90% CI.

d) An experiment producing equally likely events has probabilities associated with it that are often computed using combinatorial coefficients.

d) If there are N events in the sample space, then each event has probability $1/N$. The probability of any event E is then (the number of sample space points in E)/ N , and the numerator, a count of the "number of ways", is often found using combinatorial coefficients.

e) The hypergeometric distribution is useful for describing certain outcomes when sampling without replacement. (No formula required).

e) In SWOR and a sample size n , for a population of two kinds of things ("0" or "1" for example), the probability of m things of one kind ("1" say), is given by the hypergeometric distribution.

f) A regression line has the property that it is the "best" line for prediction of y from x .

f) The regression line is chosen to minimize the sum of squared prediction errors in the data set. The regression line is used to translate a value x into a prediction \hat{y} . The prediction errors are $(y - \hat{y})$ and the "best" criterion in this situation is defined as the line that minimizes the sum of squared errors.

2. (15 marks=5+5+5) Describe the role of the sampling distribution of the sampling mean in the application simulations described below:

a) We showed that a portfolio of companies could provide a good profit even if every company in the portfolio had only a small chance of being profitable.

a) If the outcomes of the companies are independent, the average return of the portfolio will be an average just like a sample mean, and its variability will be much less than the variability of the return from a single company, according to the square root law. As long as the "expected" return is positive, a large number of companies in the portfolio will almost guarantee a positive return for the portfolio.

b) We showed that a large auto insurance company could have a better chance of being profitable than a smaller auto insurance company, even when the policies held were individually equally profitable between the two companies.

b) Each policy is like the companies in part a). As long as the outcome from the policies are independent, and the "expected" return from each policy is positive, the average outcome will be positive if the company, since the variability of the outcome can be made as precise as necessary by increasing the number of policies. In other words, the mean less 2 SEs will be > 0 .

c) We showed that a student receiving a B on every component of a course could receive a A on the course as a whole, in a setting where the letter grades are assigned on the basis of percentiles.

c) Imagine all the components are marked out of 100, and a typical B score is in the 65th percentile and a mark of 70, while a typical A score would be in the top 20% of the component scores. The average for the student described would be 70 but relative to the average over all students, the mark of 70 would be a higher percentile than the 65th

percentile. The reason for this is that the variability of the average grade (across students) has a smaller variability than the individual components, and so if 70 is 1 SD above the mean on a component, it might be 2 SDs above the mean on the distribution of average scores. This effect is strongest when there are a large number of components (like the n – sample size) and the performances on them are independent (i.e. when a student does well on one component, it does not mean they will do well on other components.)

3. (5 marks) A gamma distribution with a shape parameter that is an integer k can be thought of as a model for the waiting time T in a Poisson process until the k th event occurs. What can you say about the shape of the distribution of T for large k ?

3. It should be approximately bell-shaped (and symmetrical) like the normal distribution. (This is a result of the CLT).

4. (10 marks) An automobile company wishes to poll its customers to determine their level of satisfaction with the vehicles it produces. There are 12,132 vehicles purchased in the last year and it is decided to contact a sample of these customers and ask them "Are you satisfied with your vehicle?" The answers are to be coded as "Yes", "No", or "Other". The company is particularly interested in the proportion of the 12,132 customers that would provide "No" answers. It is proposed to estimate this proportion based on a sample of either 50 or 100 customers. Advise the company on the precision of estimation they might expect with these sample sizes. (You may assume that each customer bought one vehicle.)

4. The proportion in the population potentially giving "No" answers is unknown. However, the worst it can be for sample size considerations is $p=1/2$. The SD of the sample proportion in this case is $\sqrt{((1/2)*(1-1/2)/n)} = 0.5/\sqrt{n}$ which is about .07 for $n=50$ and .05 for $n=100$. So the company could be advised that the sample estimate of the population proportion they want to know would be accurate to within about ± 0.14 for $n=50$ and ± 0.10 for $n=100$. (In other words, ± 14 percentage points for $n=50$ and ± 10 percentage points for $n=100$.) The company could be told that the precision might be a little better if the proportion was not close to 0.5.

5. (5 marks) The chance that a particular student of STAT 100 be involved in an automobile accident was estimated to be 1% per month. Find the probability that this student has an accident in the next twelve months.

5. The average number of accidents in 12 months is 0.12, and the actual number could be modeled as Poisson RV N . $1 - P(N=0) = 1 - \exp(-0.12) = 0.113$. The possible answer 0.12 is not quite right since the 12 months events are not mutually exclusive.

6. (10 marks=5+5) An IQ test is calibrated so that, in the national population tested, the mean is 100 and the standard deviation is 15. A small one-room country school uses this IQ test and on all of its nineteen students: the mean is 107 and the standard deviation is 10. The IQ distribution in the national population is designed to be normal.

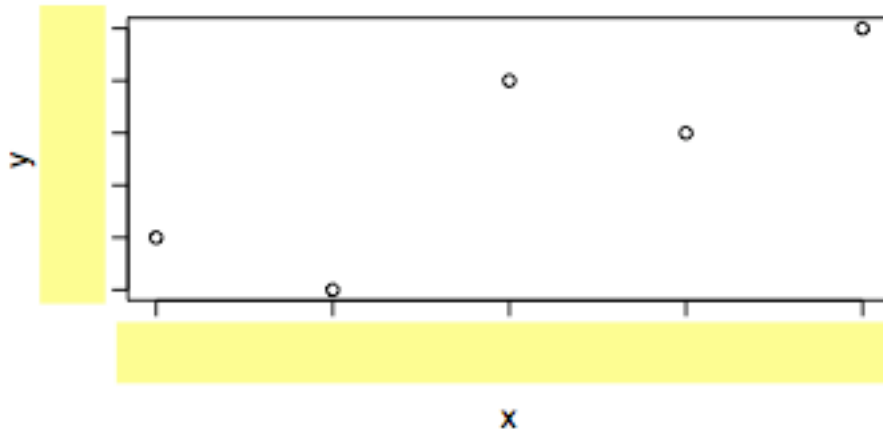
a) Use a test of hypothesis of the mean to assess whether the school results are typical of a random sample of 19 students from the national population. Use the P-value method.

b) Find a confidence interval for the standard deviation of the national population based on the 19 student's values, and assuming these nineteen to be a random sample from the national population (but without using the assumed value of the SD=15).

6. a) $z = (107 - 100) / (15 / \sqrt{19}) = 2.03$ If H_0 : mean=100 and H_a : mean \neq 100 then the P-value is $.0212 + (1 - .9788) = .04$ approx. This small P-value suggests H_0 is false so population mean \neq 100. So these students are not typical of the national population.

b) A 95% CI for the pop. mean is $107 \pm 2.101 * 10 / \sqrt{19} = 107 \pm 4.8$

7. (5 marks) Compute, to two decimal places, the correlation coefficient between x and y for the data shown:



7. Assign numbers to the points $x=0,1,2,3,4$ $y=1,0,4,3,5$.
 $cor(x,y) = cov(x,y) / (sd(x) * sd(y)) = 0.84$. (If someone gets 0.73, just subtract 1 mark)

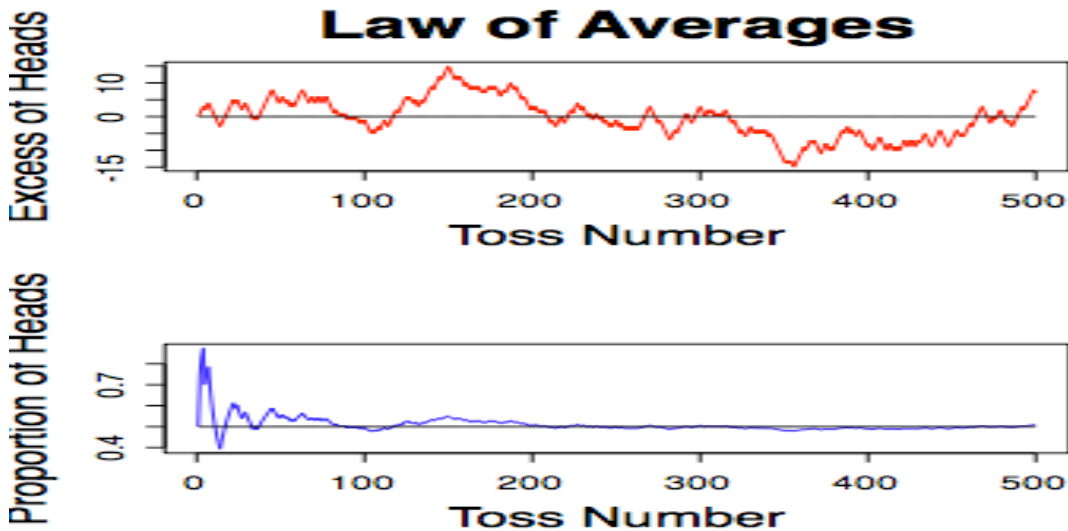
8. (5 marks) In describing a distribution numerically, when would the **median and interquartile range** be preferred to **mean and standard deviation**? Consider the shape of the distribution, and the audience of the description.

8. use median & IQR when
shape: skewed or multi modal, or when
audience: non-technical

otherwise use mean and SD.

9. (10 marks=5+5) A simulation called "The Law of Averages" was demonstrated in class, in which a fair coin was repeatedly tossed and at each toss the accumulation of EXCESS=(the number of heads minus the number of tails) was recorded. Also recorded at each toss was the accumulated proportion of heads, PROP. We were studying the

tendency of DIFF and PROP as the number of tosses increased to a large number, like 500. To speed up the demonstration, we used the computer for the simulation and the graphical summary, like this:



a) Compute the standard deviation of each of DIFF and PROP, for 500 throws, for this experiment. (Hint: $DIFF = 2 * (\text{number of heads}) - \text{number of Tosses}$)

b) Comment on the general graphical behavior of these experimental outcomes if the coin is biased so that $P(\text{Head}) = .55$. (no calculations required for this part).

9. a) $SD(DIFF) = 2 * \sqrt{500 * .5 * .5} = 22.4$ (Binomial SD)

$SD(PROP) = \sqrt{.5 * .5 / 500} = 0.0224$

b) red curve trends upward linearly (with similar variability to that shown), blue curve converges to 0.55.

10. (5 marks) Write down the aspect of this course content that was of most interest to you, and explain why it was of interest to you. ("most interest" can be replaced by "least irrelevant" if necessary).

10. Almost anything is acceptable as long as both parts are answered.