American University of Beirut
Stat 230 Summer 2013

Midterm Exam

Name: _____ ID Number: _____

Rules

1. No questions.
2. No talking.
3. No looking at someone else's exam or any other action which could be construed as cheating.
4. No mobile phones.
5. No calculators.
6. There is no need to simplify solutions for which a calculator is needed (such as $\binom{n}{k}$ or $\sum$). However, you are expected to integrate and differentiate where needed.

Failure to comply with Rules 1-5 will result in a score of zero for the exam and your immediate dismissal from the room. There will be no warnings.

Some Possibly Useful Formulae

Geometric(p): $f(i) = (1-p)^{i-1}p, \quad i = 1, 2, \ldots$

Binomial(n, p): $f(i) = \binom{n}{i}p^i(1-p)^{n-i}, \quad i = 0, 1, \ldots, n$

Poisson(λ): $f(i) = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \ldots$

NegBin(r, p): $f(n) = \binom{n-1}{r-1}p^r(1-p)^{n-r}, \quad n = r, r+1, \ldots$

HypGeo(N, n, k): $f(i) = \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}}, \quad \max\{0, n-(N-k)\} \le i \le \min\{n, k\}$

Uniform: $f(x; a, b) = \frac{1}{b-a}, \quad a < x < b$

Normal: $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad -\infty < x < \infty$

Gamma: $f(x; \alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}, \quad x, \alpha, \beta > 0$

Beta: $f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < x < 1, \ \alpha, \beta > 0$

Lognormal: $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\Big(-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2\Big), \quad x \ge 0$

Weibull: $f(x; \alpha, \beta) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}, \quad x, \alpha, \beta > 0$

Section One (*12 points*)

1. (12 points) For the following questions, circle **TRUE** or **FALSE**. No work need be shown. Each correct answer will be awarded +2 points and each incorrect answer will be penalized −2 points. Blank answers will be marked as zero.

T **F** It is possible that $P(A) = P(B) = 0.6$ and that A and B are mutually exclusive.

Solution: False. If they were mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) = 1.2$$

T **F** For 3 events A, B and C, $P(A \cup (B' \cup C')') = P(A \cup (B \cap C))$

Solution: True (by DeMorgan's Laws).

T **F** Suppose there is a pile of 157 socks on your bedroom floor; 19 of them are clean and the remainder are dirty. Your curious mother bravely reaches into the pile and grabs 22 socks. Let X be a random variable denoting the number of clean socks she is holding; then X has a hypergeometric distribution.

Solution: False, $22 > 19$.

T **F** For a normal random variable X with mean 6 and variance 9, $P(3 < X < 9) > 0.7$.

Solution: False , it is $P(-1 < Z < 1) = .8413 - .1587 = .6826$.

T **F** Mutually exclusive events can be independent, but independent events cannot be mutually exclusive.

Solution: False, they are just different things...

T **F** The Central Limit Theorem states that as the sample size tends to infinity, then regardless of the distribution in the population, the distribution of the random sample will be approximately standard normal.

Solution: False, the CLT is a theorem about the distribution of the sample mean, not the distribution of the sample.

Section Two (*48 points*)

Answer the questions in the spaces provided. If you run out of room for an answer, continue on the back of the page. **Show clearly all work and make sure your final answer is clearly indicated.**

2. (6 points) Consider a typical 'Pick-Six' Lottery: a person purchases a ticket and can choose 6 distinct (no repeated) numbers in the set $\{1, 2, 3, \ldots, 50\}$. Later a Lottery Machine picks 6 distinct numbers *at random* from the set $\{1, 2, 3, \ldots, 50\}$. A winning ticket is one which matches all six numbers chosen by the Machine (in any order–the order does not matter). What is the probability of winning if you purchase only one ticket?

Solution: The sample space are all the subsets of $\{1, 2, 3, \ldots, 50\}$ with six elements. There are $\binom{50}{6}$ possible outcomes. All outcomes are equally likely, so the probability of winning is

$$\frac{\binom{6}{6}\binom{44}{0}}{\binom{50}{6}} = \frac{1}{15890700}$$

where $\binom{6}{6}\binom{44}{0}$ is the number of ways to choose six winning numbers from six and zero losing numbers from the remaining 44.

3. (7 points) Show that if the joint p.d.f. of X, Y is $f(x, y) = ke^{x+y}$ for $0 < x < 1$ and $0 < y < 1$, then X and Y are independent. [*Note: First you need to find the value of any unknown constants in the density.*]

Solution: First, solve for k,

$$1 = \int_0^1 \int_0^1 f_{X,Y}(x, y)\, dxdy = \int_0^1 \int_0^1 ke^{x+y}\, dxdy = k(e-1)^2$$

Thus $k = (e-1)^{-2}$, and

$$f_{X,Y}(x, y) = (e-1)^{-2}e^{x+y}, \quad x, y \in [0, 1]$$

Furthermore, we can see that

$$f_X(x) = \int_0^1 (e-1)^{-2}e^{x+y}\, dy = e^x(e-1)^{-1}, \;\; x \in [0, 1]$$

Similarly, $f_Y(y) = e^y(e-1)^{-1}$. Thus $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ and therefore X and Y are independent.

4. (7 points) Suppose X has a continuous uniform distribution on the interval $[-1, 1]$. Let $Y = 3X + 4$. Derive the correlation between X and Y using the definition (i.e. do not use any properties without showing them).

Solution: Most people used properties of covariance, variance and the uniform distribution without showing them as they were required to do. You don't actually need to find $E(X)$ and $\text{var}(X)$ to answer this question. However, if you do, then you need to find them using the uniform density. We have $E(X) = \int_{-1}^{1} \frac{1}{2} dx = \frac{x}{2}\Big|_{-1}^{1} = 0$ and thus $\text{var}(X) = E(X^2) = \int_{-1}^{1} \frac{x^2}{2} dx = \frac{x^3}{6}\Big|_{-1}^{1} = \frac{1}{3}$. Then $E(Y) = 4$ and $\text{var}(Y) = 3$. Also,

$$\text{cov}(X, Y) = E([X - E(X)][3X + 4 - E(3X + 4)]) = E[3[X - E(X)]^2] = 3\text{var}(X) = 1.$$

Thus,

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = 1$$

5. (7 points) The brothers Assi and Mansour Rahbani decided that all of the songs they wrote for Fairuz should be credited as being written by 'The Rahbani Brothers' without specifying which one had actually written the song, even though every song was written by either Assi or Mansour, but never together. In fact, Assi and Mansour each wrote half of the songs. Any Fairuz song written by Assi would become a number one hit with probability 0.6 and a song written by Mansour would become a number one hit with probability 0.3. Given that a particular Fairuz song credited to the Rahbani Brothers was a number one hit, what is the probability that it was written by Mansour?

Solution: Bayes' Theorem. Let H be the event that a song is a number one hit, A means Assi and M means Mansour. We have $P(H|A) = 0.6$ and $P(H|M) = 0.3$. Also $P(A) = P(M) = 0.5$. Then

$$P(M|H) = \frac{P(H|M)P(M)}{P(H|M)P(M) + P(H|A)P(A)} = \frac{(0.3)(0.5)}{(0.3)(0.5) + (0.6)(0.5)} = \frac{.15}{.45} = \frac{1}{3}$$

6. (7 points) On any given day, independently of what happens on any other day, a bird poops on my car with fixed probability 0.9. Consider the probability that in the next 1000 days, I will be able to enjoy at least 50 but no more than 150 days **without** a bird pooping on my car. Give 3 different answers to this question: (a) the exact answer, (b) an approximation using a discrete distribution and (c) an approximation using a continuous distribution. Make sure to clearly define the distributions and parameters you are using.

Solution: This is a binomial problem with $n = 1000$ and $p = 0.1$. We have

$$P(50 \leq X_{bin} \leq 150) = \sum_{k=50}^{150} \binom{1000}{k} (0.1)^k (0.9)^{1000-k}$$

For the Poisson approximation, let $\lambda = np = 100$. Then

$$P(50 \leq X_{bin} \leq 150) \approx \sum_{k=50}^{150} \frac{100^k e^{-100}}{k!}$$

and the normal approximation uses the continuity correction, so

$$P(50 \leq X_{bin} \leq 150) \approx P(\frac{49.5 - 100}{\sqrt{90}} < Z < \frac{150.5 - 100}{\sqrt{90}})$$

7. (7 points) Suppose that X is a continuous random variable with cumulative distribution function given by $F(x) = x$, $0 \leq x \leq 1$. Let $Y = 3X + 4$. Find $E(Y)$ and $\text{var}(Y)$.

Solution: Clearly $f(x) = 1$ for $0 \leq x \leq 1$. We could first find $E(X) = \int_0^1 x dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$ and $E(X^2) = \int_0^1 x^2 dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$. Therefore

$$\text{var}(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Then $E(Y) = 3E(X) + 4 = \frac{11}{2}$ and $\text{var}(Y) = 3^2\text{var}(X) = \frac{3}{4}$. Alternatively, one could simply calculate

$$E(Y) = \int_0^1 (3x + 4)dx = \frac{3x^2}{2} + 4x\Big|_0^1 = \frac{11}{2}$$

and

$$E(Y^2) = \int_0^1 (9x^2 + 24x + 16)dx = 3x^3 + 12x^2 + 16x\Big|_0^1 = 31$$

which yields $\text{var}(Y) = E(Y^2) - [E(Y)]^2 = 31 - \frac{121}{4} = \frac{3}{4}$.

8. (7 points) State the Central Limit Theorem as precisely as you can (i.e. in mathematical terms, not with words). Make sure to state all of the assumptions and define every quantity.

Solution: See slides. A shocking number of people clearly do not understand the CLT. A lot of people wrote things like, 'when the random samples are large, we take the random variable X and approximate it to $Z = \frac{X - E(X)}{\sqrt{\text{var}(X)}}$,' or something like this, which makes absolutely no sense. A complete answer would include all of the following elements:

- given a random sample of size n, which is a set $\{X_1, \ldots, X_n\}$ of independent and identically distributed random variables
- the population has finite variance σ^2
- as the sample size n tends to infinity, the distribution of the random variable defined by
$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{var}(\bar{X})}}$$
is approximately standard normal (or some equivalent mathematical statement about this main result)
- make it clear that you understand this theorem is a theorem about the distribution of $\bar{X}$, not the distribution of the sample or anything else
- the form of the population distribution does not matter
- $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \sigma^2/n$ (these are properties of $\bar{X}$, not part of the theorem)

There are many equivalent ways to express the CLT, as given in the slides and book. It makes no sense to talk about 'the number of samples'. We are talking about a *random sample of size* n, not n samples. Also it makes no sense to mention $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ without making it clear that this is a random variable and the CLT gives an approximation to its asymptotic distribution. Moreover, Z is not synonymous with standard normal; you can't just say Z without telling me which distribution it has.

Bonus Question (worth 4 points, but not counted at all unless all other questions in Section 2 are answered)

9. Explain why the normal distribution with mean μ and variance σ^2, which has positive probability over all intervals from $(-\infty, \infty)$, may still be appropriate to model the distribution of a random variable which, in reality, can take only positive values.

Solution: No partial credit. A graph would be helpful in answering this question. If $\mu >> 0$ and σ^2 is small relative to the mean, then almost all of the area under the normal density will be over positive values. We know that more than 99% of the values of a normal random variable are within 3 standard deviations of its mean. Thus, while a normal random variable will always take negative values with positive probability, this probability can be negligibly small for large positive values of μ and small relative values of σ^2. Thus the normal distribution could in principle still be a good approximation to the distribution of a random variable which only takes positive values.

Bonus Question (worth 4 points, but not counted at all unless all other questions in Section 2 are answered)

10. Let X represent a random variable having the same distribution as the population and suppose that the distribution is normal with mean 2 and unknown variance σ^2. Further suppose we have a very small random sample from this population. Determine which is greater, $P(1 < X < 3)$ or $P(1 < \bar{X} < 3)$.

Solution: No partial credit. Here we know that $\bar{X}$ is normal without the CLT because it is a sum of iid normal random variables. In fact we are told that we have a 'very small' random sample, so the CLT would not apply here. Thus $P(1 < X < 3) = P(\frac{1-2}{\sigma} < Z < \frac{3-2}{\sigma})$. Also $P(1 < \bar{X} < 3) = P(\frac{1-2}{\sigma/\sqrt{n}} < Z < \frac{3-2}{\sigma/\sqrt{n}})$. If $n > 1$, the interval for $\bar{X}$ contains a higher concentration of the density, because $\text{var}(\bar{X}) < \text{var}(X)$ and therefore $P(1 < \bar{X} < 3) > P(1 < X < 3)$. If $n = 1$, they are the same.