Name: OUTLINE SOLUTIONS

University of Chicago
Graduate School of Business

Business 41000: Business Statistics

Special Notes:

1. This is a closed-book exam. You may use an 8×11 piece of paper for the formulas.
2. Throughout this paper, $N(\mu, \sigma^2)$ will denote a normal distribution with mean μ and variance σ^2.
3. This is a 2 hr exam.

Honor Code: By signing my name below, I pledge my honor that I have not violated the Booth Honor Code during this examination.

Signature:

Problem A. True or False: Please Explain your answers in detail. Partial credit will be given (50 points)

1. The sample variance is unaffected by outlying observations.

   False. We have the following formula for the sample variance:

   $$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

   If x_i is large relative to $\bar{x}$ then it has an undue influence.

2. If $P(A|B) = 0.5$ and $P(B) = 0.5$, then the events A and B are necessarily independent.

False. This is not necessarily true. We need more information about each event to definitively say so.

3. Consider the standard normal random variable $Z \sim N(0,1)$. Then the random variable $-Z$ is also standard normal.

True. If $X \sim N(0,1)$, then $-X \sim N\Big(-1(0), (-1)^2 \times 1\Big) = N(0,1)$. Moreover, we have that the pdf of Z and $-Z$ is the same:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}z^2 = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(-z)^2 = f(-z)$$

Lastly, we can observe that all the even moments are the same due to symmetry and all the odd moments are zero. Thus, they are the same distribution.

4. The Central Limit Theorem states that the distribution of the sample mean $\bar{X}$ is Normally distributed for large samples.

True. If X_i is a random sample (or iid) with mean μ, then $\sqrt{n}(\bar{X} - \mu) \longrightarrow^d N(0, \sigma^2)$

5. Suppose that you toss a biased coin with probability 0.25 of getting a head. The probability of getting five heads out of ten tosses is less than thirty percent.

True This follows from a binomial distribution with $n = 10$.
$Prob(5H) = \begin{pmatrix} 10 \\ 5 \end{pmatrix} (0.25)^5(0.75)^5 = 0.0583$

6. There is much discussion of the effects of second-hand smoke. In a survey of 500 children who live in families where someone smokes, it was found that 10 children were in poor health. A 95% confidence interval for the probability of a child living in a smoking family being in poor health is then 2% to 4%.

False The 95% confidence interval should be:

$$\left[.02 - 1.96\sqrt{\frac{.02(1-.02)}{500}}, .02 + 1.96\sqrt{\frac{.02(1-.02)}{500}}\right]$$

So, the lower bound on this interval will definitely be less than 2 percent.

7. Arsenal are playing Liverpool at home in an EPL game this weekend. You think that the number of goals to be scored by both teams follow a Poisson distribution with rates 2.2 and 1.6 respectively. Given this, the odds of a scoreless $0-0$ draw are $45-1$.

False Assuming each scoring's ability is independent, we have that:

$$\begin{aligned} Pr(Arsenal = 0) &= \frac{(2.2)^0 e^{-2.2}}{0!} = e^{-2.2} \\ Pr(Liverpool = 0) &= \frac{(1.6)^0 e^{-1.6}}{0!} = e^{-1.6} \\ \longrightarrow \quad Pr(Arsenal = 0 \; and \; Liverpool = 0) &= e^{-2.2} \times e^{-1.6} = e^{-3.8} \end{aligned}$$

The odds are: $O = (1 - e^{-3.8})/e^{-3.8}$ or 43.7 to 1.

8. A local bank experiences a 2% default rate on residential loans made in a certain city. Suppose that the bank makes 2000 loans. Then the probability of more than 50 defaults is 25 percent.

False The probability of having more than 50 defaults can be calculated using a Normal approximation to the $Bin(2000, 0.02) \approx N(2000 \times 0.02, 2000 \times 0.02 \times 0.98)$. So, the mean is 40 and the standard deviation is 6.261. Therefore,

$$Pr(default > 50) = Pr\Big(\frac{default - 40}{6.261} - \frac{50 - 40}{6.261}\Big) = Pr(z > 1.5972) \approx .055$$

9. The average movie is Netflix's database has an average customer rating of 3.1 with a standard deviation of 1. The last episode of *Breaking Bad* had a rating of 4.7 with a standard deviation of 0.5. The $p - value$ for testing whether Breaking Bad's rating is statistical different from the average is a lot less than 1%.

   False The formula for a T-ratio is

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}}$$

   However, since we do not know n_1 and n_2 we cannot get the T-ratio and thus p-value.

10. You are finding a confidence interval for a population mean. Holding everything else constant, an interval based on an unknown standard deviation will be wider than one based on a known standard deviation no matter what the sample size is.

    True Because you need to use the same data to estimate the standard deviation, and thus contains more noise or error. In other words, we are comparing between z distribution and t distribution.

Problem B. (20 points)
Several spam filters use Bayes rule.
Suppose that you empirically find the following probability table for classifying emails with the phrase "buy now" in their title as either "spam" or "not spam".

	Spam	Not Spam
"buy now"	0.02	0.08
not "buy now"	0.18	0.72

1. What is the probability that you will receive an email with spam?

   Answer: The probability that you will receive an email with spam is:

   $$Pr(Spam) = .02 + .18 = .2$$

2. Suppose that you are given a new email with the phrase "buy now" in its title

   What is the probability that this new email is spam?

   Answer: The posterior probability is

   $$Pr(Spam|buy\ now) = \frac{Pr(Spam\ and\ buy\ now)}{Pr(buy\ now)} = \frac{.02}{.02 + .08} = .2$$

3. Explain clearly any rules of probability that you use.

   Answer: The marginal probability is given by the law of total probability: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$. We also use the definition of conditional probability (not Bayes' rule) $Pr(A|B) = \frac{Pr(A\ and\ B)}{Pr(B)}$

Problem C. (20 points)
You are given the following `R` output in three variables. You know that the data generated can from three distributions: a Normal, Binomial and Poisson.

Figure 1: Probability Distributions

```
> summary(z1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    3.00    5.00    5.04    6.25   14.00
> summary(z2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   5.000   4.987   6.000  14.000
> summary(z3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -3.141   3.261   4.889   4.997   6.693  13.910
```

- Identify which variable can from which distribution. Explain your reasoning.

  Answer: The variables `z1, z2`, and `z3` are Binomial, Poisson, and Normal respectively.

  The only distribution whose realization can be negative is the normal. The Poisson distribution has more mass at zero. The first is Binomial which has an upper bound.

- Why do these three distributions have similar looking histograms? Explain the theoretical relationships between the three distributions.

  Answer: For large sample sizes, the Binomial is approximation $N(np, np(1-p))$ and if $\lambda = np$ where p is small the Poisson will look like a Binomial.

Problem D. (20 points)
Google is test marketing a new website design to see if it increases the number of click-throughs on banner ads. In a small study of a million page views they find the following table of responses

	Total Viewers	Click-Throughs
new design	700,000	10,000
old design	300,000	2,000

1. Find a 99% confidence interval for the increase in the proportion of people who click-through on banner ads using the new web design.

   Answer: We are interested in the distribution of $p_1 - p_2$. We have that:

$$\hat{p}_1 = \frac{10000}{700000} = \frac{1}{70}$$
$$\hat{p}_1 = \frac{2000}{300000} = \frac{1}{150}$$

   Following the hint, we have that the 99% confidence interval is:

$$\begin{aligned} p_1 - p_2 &\in \left[\hat{p}_1 - \hat{p}_2 - 2.58\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + 2.58\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right] \\ &= [0.00762 - 0.00053, 0.00762 + 0.00053] \\ &= [0.00709, 0.00815] \end{aligned}$$

[Hint: a 99% confidence interval for a difference in proportions $p_1 - p_2$ is given by $(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$]