Steve Sawin
Statistics Math 3352

Final (Solutions)

Sign your name below to indicate that the only resources you used on the computer was Excel and the data and template files on my website, the only paper resources you used were your four pages of notes and scratch paper, and the only person you communicated with during the exam was Professor Sawin.

---

1. According to the General Sociological Survey, the number of sex partners American adults had in the last 12 months had a mean of 1.0 and a standard deviation of 1.0.

   (a) [7pt] Identify the **population** and **variable** under consideration. Population is American adults, variable is how many sex partners each American had in the last year.

   (b) [5pt] Does this variable have a normal distribution? Explain. No, a normal dist with a mean of 1 and a standard deviation of 1 would be negative about 17% of the time, while this variable is never negative. Also, it is discrete with only a few likely values.

   (c) [7pt] For a random sample of 100 adults, describe the **shape** of the sampling distribution of $\overline{X}$, and give its **mean** and **standard error**. Be sure to say how your answer to (b) relates to the shape of the sampling distribution. Since $n > 40$, the distribution of $\overline{X}$ is approximately normal even though we said X was not, by the Central Limit Theorem.
   Mean: $\mu_{\overline{X}} = \mu_X = 1$.
   St. Error: $\sigma_{\overline{X}} = \sigma_X/\sqrt{n} = 1/10 = .1$.

   (d) [6pt] What is the chance that sample will have an average numbers of sex partners bigger than 1.5?

$$P(\overline{X} > 1.5) = 1 - \text{Normdist}(1.5, 1, .1, 1) = 2.8 \times 10^{-7}.$$

2. [12pt] The length of the wings of a random African swallow is a continuous random variable with density function $f(x) = 3x^2$ for $0 < x < 1$ (OK, I made that up). Find the density function of the area of the wing, which is given by $Y = X^2/2$.

$$P(X < x) = F(x) = \int_0^x 3x^2 dx = x^3 \qquad 0 < x < 1.$$

$$F(y) = P(Y < y) = P\left(X^2/2 < y\right) = P\left(X < \sqrt{2y}\right) = F\left(\sqrt{2y}\right) = 2^{3/2}y^{3/2}.$$

$$f(y) = F'(y) = \frac{3}{2}2^{3/2}y^{1/2} = 2^{1/2}y^{1/2} \qquad 0 < y < 1/2$$

because when $x = 1$ then $y = 1/2$.

3. (a) [10pt] Find the moment generating function for the sum Y of n independent **Poisson** variables, each with mean λ. The moment generating function for a Poission variable with mean λ is

$$e^{\lambda(e^t-1)}.$$

When you add independent variables you multiply the moment generating functions so

$$M_Y(t) = e^{\lambda(e^t-1)} \times e^{\lambda(e^t-1)} \times \cdots \times e^{\lambda(e^t-1)} = e^{n\lambda(e^t-1)}.$$

(b) [12pt] Identify the distribution of Y. We recognize the moment generating function of Y as the moment generating function of a Poisson distribution with mean $n\lambda$. Since a destribution is uniquely determined by its MGF, Y is a Poisson variable with mean $n\lambda$.

4. (a) [6pt] Find the probability density function $L(x_1, x_2, \ldots, x_n|n, \theta)$ of the n independent Bernoulli variables $X_1, X_2, \ldots, X_n$ that are each 1 with probability θ and 0 with probability $1-\theta$.

$$L(x_1, x_2, \ldots, x_n|n, \theta) = \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\cdots\theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

(b) [7pt] For L as in (a) show that

$$\frac{\partial}{\partial\theta}\ln\left[L(x_1, x_2, \ldots, x_n|n, \theta)\right] = \frac{\sum_{i=1}^n x_i - n\theta}{\theta(1-\theta)}$$

$$\ln\left[L(x_1, x_2, \ldots, x_n|n, \theta)\right] = \sum_{i=1}^n x_i \ln\theta + \left(n - \sum_{i=1}^n x_i\right)\ln(1-\theta)$$

$$\frac{\partial}{\partial\theta}\ln\left[L(x_1, x_2, \ldots, x_n|n, \theta)\right] = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{\sum_{i=1}^n x_i - n}{1-\theta} = \frac{\sum_{i=1}^n x_i(1-\theta) - (n-\sum_{i=1}^n x_i)\,\theta}{\theta(1-\theta)} = \frac{\sum_{i=1}^n x_i - n\theta}{\theta(1-\theta)}.$$

(c) [9pt] Use (b) (even if you didn't complete it) to find the maximum likelihood estimate for the θ of a Bernoulli random variable from a sample of size n. The maximum likelihood estimate of θ is the solution to

$$0 = \frac{\partial}{\partial\theta}\ln\left[L(x_1, x_2, \ldots, x_n|n, \theta)\right] = \frac{\sum_{i=1}^n x_i - n\theta}{\theta(1-\theta)} \qquad \text{so}$$

$$n\theta = \sum_{i=1}^n x_i$$

$$\theta = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

which is to say that the maximum liklihood estimate for the population proportion θ is the sample proportion $\sum_{i=1}^{n} x_i/n$, which makes sense.

(d) [8pt] Show your estimate in part (c) (previous page) is an unbiased estimator for θ.

$$E\left[\overline{x}\right] = E\left[x_i\right] = \theta$$

using the mean of the Bernoulli distribution (actually $\sum_{i=1}^{n} x_i$ has a binomial distribution, which makes this even easier).

(e) [5pt] Find $E\left[(X-\theta)^2\right]$ for X a Bernoulli variable with parameter θ. If it helps you can always think of the Bernoulli variable as a binomial variable with $n = 1$. If it doesn't help, don't.

$$E\left[(X-\theta)^2\right] = E\left[(X-\mu_X)^2\right] = \mathrm{Var}\left[X\right] = \theta(1-\theta).$$

(f) [11pt] Show your answer in (c) has minimum variance among all unbiased estimators. You will end up using (d). Notice that the relevant $f(x)$ for Cramer-Rao is the L of parts (a) and (b) with $n = 1$. Since $f(x) = \theta^x(1-\theta)^{1-x}$ we can steal from (b) or reproduce its reasoning to get that $\partial/\partial\theta f = \frac{x-\theta}{\theta(1-\theta)}$. We have to check that $\mathrm{Var}\left[\overline{X}\right]$ equals

$$\begin{aligned}
\frac{1}{nE\left[(\partial/\partial\theta \ln f)^2\right]} &= \frac{1}{nE\left[\frac{(X-\theta)^2}{\theta^2(1-\theta)^2}\right]} \\
&= \frac{1}{\frac{n}{\theta^2(1-\theta)^2}E\left[(X-\theta)^2\right]} \\
&= \frac{1}{\frac{n}{\theta^2(1-\theta)^2}\theta(1-\theta)} \\
&= \frac{1}{\frac{n}{\theta(1-\theta)}} \\
&= \frac{\theta(1-\theta)}{n}.
\end{aligned}$$

On the other hand $\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[X_1\right]/n = \theta(1-\theta)/n$.

5. [11pt] I have a bivariant distribution (X, Y) with

$$\begin{aligned}
\mu_X &= 10, & \sigma_X &= 2 \\
\mu_Y &= 4, & \sigma_Y &= 1 \\
\mathrm{COV}(X,Y) &= -4.
\end{aligned}$$

Find the coefficient of correlation ρ and the Linear Regression relationship between these two variables.

$$\begin{aligned}
\rho &= \mathrm{COV}(X,Y)/(\sigma_X\sigma_Y) = -4/(1*2) = -2 \\
\beta &= \mathrm{COV}(X,Y)/\sigma_X^2 = -4/2^2 = -1 \\
\alpha &= \mu_Y - \beta\mu_X = 4 - 1*10 = -6.
\end{aligned}$$

$$m(x) = -6 - x.$$

6. [10pt] Find the equation of the least squares line for the data set $(-1,0),(0,1),(1,5)$.

$$\overline{x} = (-1+0+1)/3 = 0$$
$$\sum_i (x_i - \overline{x})^2 = (-1)^2 + 0^2 + 1^2 = 2$$
$$\overline{y} = (0+1+5)/3 = 2$$
$$\sum_i (y_i - \overline{y})^2 = (-2)^2 + (-1)^2 + (3)^2 = 14$$
$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = (-1)(-2) + 0(-1) + 1(3) = 5b \qquad = \frac{5}{2}$$
$$a = overline y - b\overline{x} = 2 - \frac{5}{2} * 0 = 2$$
$$y = 2 + \frac{5}{2}x$$

7. [20pt] 16 randomly selected runners are asked to run a one kilometer race on each of two consecutive weeks. In one race the runners wear one brand of shoe and in the other a second brand. Which brand each wears in which race is determined at random. All runners are timed and asked to run their best in each race. The results in minutes are in the "runners" tab of the file alt_data.xls on the Data file. You want to determine if there is evidence at the 5% significance level that one brand is better than the other for running races. Decide **what test** to use, identify **each assumption** of the test and say briefly **why it is or is not met,** the **alternate hypothesis,** the **p-value** and a **clear English sentence** giving the conclusion.(Hint: The only relevant thing for each runner is how much *longer* they took with Brand X than with Brand Y, which is the difference in the last column.)

For each individual the relevant variable is the difference between their running times, which is numerical, so we use the one-sample t-procedure.

Since the question did not specify one brand, we ask the alternate hypothesis $\mu_1 \neq \mu_2$.

It says the runners were chosen at random. Since there are 16 runners we check the shape of the distribution. This is a bit skewed, but is probably OK. The population of runners needs to be at least 160.

The p-value is 92.8%. This is not significant evidence that there is a difference in running times between the shoes.

8. [14pt] For the following situation identify **what samples** you would take (e.g. "a sample of American men"), **what variables** you would measure or questions you would ask (e.g. "I.Q. and number of cavities "), **what test** you would use (e.g. "linear regression"), the **alternate hypothesis** (e.g. "$\beta > 0$") and what your **conclusion would be in an English sentence** if the data turned out to be significant (e.g. "This data is significant evidence that there is a positive correlation between number of cavities and IQ in American men"). In some cases more than one study design and statistical test may be appropriate.

You want to know whether people generally score higher on a final if they take vitamins than if they don't.

Take a sample of people who have not taken vitamins and give half of them (randomly chosen) vitamins before their test, see how well each do on the test, then use the two sample t-test with the alternate hypothesis $\mu_{\text{vit}} > \mu_{\text{no-vit}}$. There is significant evidence that people who take vitamins do better on tests.

OR, take a sample of people, give them all vitamin before one test (randomly selected) and require them not to take vitamins for another. Record both scores. Use matched pairs (one sample t on the differences) with the alternate hypothesis $\mu_{\text{vit}} > \mu_{\text{no-vit}}$ (which would be $\mu_{\text{vit}-\text{no vit}} > 0$). Same conclusion.

9. You survey 100 adult American smokers and ask each their age when they first started to smoke. the 95% confidence interval was 13.2 ± 2.7.

   (a) [5pt] What is the quantity this confidence interval is estimating (describe it in words, not give a number). the parameter it is estimating is the average age at which an adult American smoker started smoking.

   (b) [5pt] If instead you found a 90% confidence interval would the width of the confidence interval be bigger or smaller? Why? The width of the confidence interval would be smaller for the 90% interval because if we are required to be less sure we can make our guess more precise. Also, the t^* value is smaller for smaller confidence levels.

   (c) [5pt] If you took a different sample of size 500 and computed the 95% confidence interval, would you expect the width of the interval to be larger or smaller? Why? the interval would be smaller because more data allows you to be more precise. Also, we divide by the square root of n.

   (d) [5pt] If we took another sample of size 100 with the same sample mean but the sample standard deviation was bigger, how would the 95% confidence interval be different? It would have the same middle, but the margin of error/width would be bigger because there would be more variation in the data. Also, the margin of error is proportional to s.

10. [20pt] Archeologists can tell if archeological sites represent different cultures or time periods by looking at the percentage of various styles of manufactured goods present. Excavation of the Cliff Talus site revealed 81 Mesa Verde type pot shards, 70 McElmo type pot shards, and 62 Mancos type pot shards. The Canyon Bench site contained 92 Mesa Verde shards, 68 McElmo shards, and 66 Mancos shards. Does this data represent evidence that the Cliff Talus and Canyon Bench were inhabited by people of different cultures? Give a null and alternate hypothesis in words, test your claim at the 5% significance level, report a p-value and your conclusion in a simple English sentence, and check all assumption We use χ^2 to test the claim that the variables location and shard type are independent (that is, where it is found affects its chances of being of a given type), against a null hypothesis that the distribution of shard types is independent of site. The p-value is 79%, which is bigger than the significance level and therefore this data is not signficant evidence that the sites represent different cultures. We do not know if this is a simple random sample. We would need to know that there are at least 2130 shards in the Cliff sit and 2260 shards in the Canyon site, which is not at all clear. There are at least five in *all* the expected cells, so that assumption is met.

Out of 200 points