Review Solutions, Exam 3, Math 338

1. Define a **random sample:** A random sample is a collection of random variables, typically indexed as $X_1, \ldots, X_n$.

2. Why is the $t-$distribution used in association with sampling? Why is the χ^2 distribution used? (In particular, pay attention to any conditions on their use)

   The $t-$distribution is typically used to model the standardized rv:

   $$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

   especially when n is small and the variance σ is unknown. In that case, we estimate σ using the sample mean s. If the random sample comes from a normal distribution, this distribution is exact; otherwise, we use this as an approximate distribution.

   As a comparison, if the random sample comes from a normal distribution and σ is known, then

   $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

   follows a standard normal distribution.

   The χ^2 distribution is used to model the random variable S^2 from a random sample. In particular, if the random sample of size n comes from a normal distribution, then

   $$\frac{(n-1)S^2}{\sigma^2}$$

   follows a χ^2 distribution with $n-1$ degrees of freedom. If the random sample does not come from a normal distribution, we may be able to use χ^2 as an estimate, which gets better as n increases.

3. The four principal game fish in Clear Lake are bluegills, crappie, small-mouth bass and large-mouth bass. The weights of these, amazingly, has a Chi-square distribution with the following parameters:

Name	deg. of freedom
Bluegills	8 oz
Crappie	13 oz
Small MB	1 lb 3 oz
Large MB	3 lb 2 oz

   where one pound is 16 oz. Assuming that the numbers of these fishes are independent, if Fisherperson Terry catches 6 bluegills, 5 crappie, 3 small mouth bass, 3 large mouth bass, what is the probability that the total weight of these 17 fish is over 19 pounds?

The fish are particular values of random variables from different distributions. Call the random variables B, C, S, L where, for example, $B_1, B_2, \ldots, B_6$ is a random sample of Bluegills.

The random variable we're interested in is the total weight. Symbolically, we let X be that rv:

$$X = \sum_{k=1}^{6} B_k + \sum_{k=1}^{5} C_k + \sum_{k=1}^{3} S_k + \sum_{k=1}^{3} L_k$$

The sum of χ^2 rvs is a χ^2 rv, and the degrees of freedom are summable. Therefore X is a random variable that is χ^2, and its degrees of freedom is:

$$6 \cdot 8 + 5 \cdot 13 + 3 \cdot 20 + 3 \cdot 50 = 323$$

With this many degrees of freedom, we'll use the Central Limit Theorem to justify a normal approximation of

$$\frac{\frac{1}{n}X - \mu}{\sigma/\sqrt{n}}$$

We know the theoretical distribution of X, therefore we know the mean and standard deviation of $\bar{X}$ (see Appendix or look it up):

$$\mu_{\bar{x}} = \frac{1}{17}E(X) = \frac{\nu}{17} = \frac{323}{17}$$

$$\sigma^2_{\bar{x}} = \frac{1}{n^2} 2\nu = \frac{2 \cdot 323}{17^2}$$

Therefore, the following rv is approximately standard normal:

$$\frac{\frac{1}{17}X - \frac{323}{17}}{\frac{\sqrt{646}}{17}} = \frac{X - 323}{\sqrt{646}}$$

Now we compute the probability that the combined weight of the fish is over 19 pounds (or 304 ounces):

$$P(X > 304) = P\left(\frac{X - 323}{\sqrt{646}} > \frac{304 - 323}{\sqrt{646}}\right) = P\left(Z > -0.75\right) =$$

$$\frac{1}{2} + P(0 < Z < 0.75) = 0.5 + 0.2734 = 0.8734$$

4. If X_1, X_2, X_3, X_4, X_5 are iid with standard normal distributions, find c so that the random variable:

$$\frac{c(X_1 + X_2)}{\sqrt{X_3^2 + X_4^2 + X_5^2}}$$

has a t distribution.

To build a $t-$distribution, we need a standard normal distribution, Z, and a χ^2 distribution, Y (see Theorem 8.12). We know that if X_1, X_2 are standard normal, then

$$Z = X_1 + X_2$$

is also standard normal with standard deviation $\sigma = \sqrt{2}$ (the MGF technique). To make Z have unit standard deviation, we should multiply Z by $\frac{1}{\sqrt{2}}$.

We know that $X_1^2 + X_2^2 + X_3^2$ has a χ^2 distribution with 3 degrees of freedom. Therefore, the following has a $t-$distribution:

$$\frac{\frac{1}{\sqrt{2}}(X_1 + X_2)}{\sqrt{\frac{X_1^2+X_2^2+X_3^2}{3}}}$$

so $c = \sqrt{\frac{3}{2}}$.

5. Given a random sample of size n from a Gamma distribution with known parameter $\alpha = 2$, find the MLE of the parameter β.

(See Exercise 10.61)

With $\alpha = 2$, the pdf is: $f(x) = \frac{1}{\beta^2} x \mathrm{e}^{-x/\beta}$. Therefore the likelihood function is:

$$L(\beta) = \frac{1}{\beta^{2n}} \sum_{i=1}^{n} x_i \mathrm{e}^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i}$$

Before differentiating to find the max, take the log of both sides:

$$\ln(L(\beta)) = -2n\ln(\beta) + \ln(\sum x_i) - \frac{\sum x_i}{\beta}$$

Differentiate and set to zero to find the critical points:

$$\frac{d\ln(L(\beta))}{d\beta} = -\frac{2n}{\beta} + \frac{\sum x_i}{\beta^2} = 0 \quad \Rightarrow \quad \beta = \frac{\overline{x}}{2}$$

To show that this is indeed a maximum, consider that the derivative can be written as:

$$\frac{d\ln(L(\beta))}{d\beta} = \frac{2n}{\beta^2}\left(-\beta + \frac{\bar{x}}{2}\right)$$

which changes from positive to negative as β changes from smaller than $\bar{x}/2$ to larger.

6. The claim that the variance of a normal population is 4 will be rejected if the variance of a random sample of size 9 exceeds 7.7535. What is the probability that the claim will be rejected (even though the actual parameter is $\sigma^2 = 4$)?

   We're looking for the probability that the variance S^2 will be greater than 7.7535. Recall that the rv $(n-1)S^2/\sigma^2$ is χ^2, so we look for:

$$P\left(\frac{8S^2}{4} > 15.507\right)$$

   Where this rv is χ^2 with 8 dof. If you look at the χ^2 table in the row with 8 dof, you will see 15.507 for $\alpha = 0.05$.

   This means that the probability is 5% of rejecting the claim.

7. A random sample of size 100 is taken from an infinite population with the mean $\mu = 75$ and $\sigma^2 = 256$. Compare the probabilities we get that $\bar{X}$ will fall between 67 and 83 using (i) Chebyshev's Inequality, and (ii) CLT.

   We compute the mean and variance of $\bar{x}$: $\mu_{\bar{x}} = \mu = 75$ and $\sigma_{\bar{x}} = \frac{16}{\sqrt{100}} = 1.6$. We'll also standardize the rv:

$$67 < \bar{X} < 83 \quad \Rightarrow \quad -8 < \bar{x} - \mu < 8 \quad \Rightarrow \quad \frac{|\bar{X} - \mu|}{\sigma} = 5$$

   For Chebyshev, $k = 5$ and the probability:

$$P\left(|\bar{X} - 75| < 5 \cdot 1.6\right) \geq 1 - \frac{1}{5^2} = \frac{24}{25} = 0.96$$

   For the normal approximation,

$$P(67 < \bar{X} < 83) = P(|Z| < 5) =$$

$$2P(0 < Z < 5) = 2 \cdot 0.4999997 = 0.9999994 \approx 1$$

8. Given an iid random sample $X_1, X_2, \ldots, X_n$, we said that the random variable $\bar{X}$ has its own distribution- What is its mean and variance? Is there a relationship between that and the sample mean and sample variance (biased or unbiased estimators perhaps?)

   Similar to our last problem, if μ is the expected value from the identical distribution, and σ is its standard deviation, then

$$\mu_{\bar{x}} = E(\bar{X}) = \mu \qquad \sigma_{\bar{x}}^2 = \frac{1}{n^2}(\sigma^2 + \ldots + \sigma^2) = \frac{1}{n}\sigma^2$$

   The sample mean and variance, as random variables, are:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X - \overline{X})^2$$

   We saw that $\overline{X}$ and S^2 are unbiased estimators of μ, σ (given infinite populations).

9. The width of a fence board that is marked 6 inches will actually have a mean width measurement of $\mu = 5.5$ inches (that is true) with a standard deviation of 0.24 inches (that is made up). What is the probability (using the CLT) that the total length of 100 boards placed side by side (with the gap between negligible) will be between 546 and 554 inches?

   To use the CLT, we need to form: $(\bar{X} - \mu)/(\sigma/\sqrt{n})$. Let X_i be the width of the i^{th} board. Then:

$$546 < \sum_{i=1}^{100} X_i < 554 \quad \Rightarrow \quad 5.46 < \frac{1}{100}\sum X_i < 5.54 \quad \Rightarrow$$

$$|\bar{X} - \mu| < 0.04 \quad \Rightarrow \quad \left|\frac{\bar{X} - \mu}{0.0204}\right| < 1.96$$

   or, what is $P(|Z| < 1.96)$? From the table,

$$P(|Z| < 1.96) = 2P(0 < Z < 0.98) = 2 \cdot 0.3365 \approx 67.3\%$$

10. In a study of television viewing habits, it is desired to estimate the number of hours that teenagers spending watching pe week. If it is reasonable to assume that $\sigma = 3.2$ hours, how large a sample is needed so that it will be possible to assert with 95% confidence that the mean is off by less than 20 minutes?

   (See Exercise 11.29, p 371)

   We take $\sigma = 3.2$, how large should n be so that it will be possible to assert with 95% confidence that the sample mean is off by less than 1/3 (where units are in hours).

$$|\bar{X} - \mu| < 1.96 \cdot \frac{3.2}{\sqrt{n}} = \frac{1}{3}$$

   Solving for n,

$$18.816 = \sqrt{n} \quad \Rightarrow \quad n \geq 354$$

11. Let Y be a random variable with pdf $f(y)$. Let $Z = G(Y)$. What should G be in order for Z to be uniform on $(0, 1)$?

   See the notes from Section 7.1 (in the text, see Example 7.8). In class, we said that G should be F, which is the CDF of Y.

   The importance of this:

   > This is how we use computers to simulate general distributions given only the uniform distribution. That is, $F^{-1}(Z) = Y$ means that if I take a random sample from Z (uniform), and take $y_i = F^{-1}(x_i)$, then y_i comes from the pdf defined as f.

Example from class: How would we find n random numbers coming from an exponential distribution (using a uniform distribution)?

12. Given a random sample of size n from a geometric population, find formulas for estimating its parameter θ by (a) Method of moments, (b) MLE (Maximum Likelihood Estimation)

    (a) Method of moments:

    $$m_1' = \frac{1}{\theta} \qquad \text{Easy!}$$

    (b) MLE:

    $$L(\theta) = \theta(1-\theta)^{(x_1-1)} \cdot \theta(1-\theta)^{(x_2-1)} \cdots \theta(1-\theta)^{(x_n-1)} = \theta^n (1-\theta)^{\sum x_i - n}$$

    Take the logarithm and differentiate, then find the critical points:

    $$\ln(L(\theta)) = n\ln(\theta) + \left(\sum_{i=1}^{n} x_i - n\right)\ln(1-\theta)$$

    $$\frac{d\ln(L(\theta))}{d\theta} = \frac{n}{\theta} - \frac{\sum_{i=1}^{n} x_i - n}{1-\theta} = 0$$

    $$\frac{\sum x_i - n}{1-\theta} = \frac{n}{\theta} \quad \Rightarrow \quad \theta\sum_{i=1}^{n} x_i - n\theta = n - n\theta \quad \Rightarrow \quad \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

    which is the same answer as in part (a).

13. If X and Y are two independent rvs having identical gamma distributions, find the joint pdf of the random variables $U = \frac{X}{X+Y}$ and $V = X + Y$.

    We will use the transformation method to get the new pdf. First we invert the given functions, then compute the Jacobian. To do the inversion, you might use substitution- For example, substitute $y = v - x$ into the equation for u:

    $$u = \frac{x}{x+y} \quad \Rightarrow \quad u = \frac{x}{x+v-x} \quad \Rightarrow \quad x = uv$$

    and solve for y in the same way:

    $$y = v - x = v - uv = v(1-u)$$

    We now have x and y in terms of u, v. Compute the Jacobian:

    $$\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \begin{bmatrix} v & u \\ -v & 1-u \end{bmatrix}$$

whose determinant is v (or the absolute value of v).

Now, the original pdf was a joint Gamma pdf:

$$f(x,y) = \frac{1}{\left(\beta^\alpha \Gamma(\alpha)\right)^2} x^{\alpha-1} y^{\alpha-1} \mathrm{e}^{-(1/\beta)(x+y)}, \quad x > 0, y > 0$$

Now convert this into a joint pdf in u, v:

$$g(u,v) = \frac{1}{\left(\beta^\alpha \Gamma(\alpha)\right)^2} (u(1-u))^{\alpha-1} v^{2\alpha-1} \mathrm{e}^{-(1/\beta)v}$$

And the restrictions were originally $x > 0, y > 0$:

$$u = \frac{x}{x+y} \qquad \Rightarrow \quad 0 < u < 1 \qquad \text{and} \qquad 0 < v < \infty$$

The inequality for u was obtained by inspection of the numerator and denominator- The denominator is always greater than the numerator.

14. The standard error is either $\sigma/\sqrt{n}$ or $s/\sqrt{n}$. What is the length of a confidence interval (using σ, then using s)?

    The CI is $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$. Respectively, the lengths are:

    $$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad 2t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

    How would n have to change in order to halve a confidence interval?

    The size n would have to be increased fourfold, since it is proportional to $1/\sqrt{n}$.

15. Find $z_{\alpha/2}$ if we are to construct an 85% confidence interval (two sided as is our usual practice):

    Using 85, we have $1 - \alpha = 85$, or $\alpha/2 = 0.075$. The area is then $0.5 - 0.075 = 0.425$. Look that up in the table to find that z is 1.44.

16. Given 8 data values, and

    $$\sum x_i = 108 \qquad \sum x_i^2 = 1486$$

    Compute the sample mean and sample variance.

    The sample mean is simple- 108/8=13.5. For the sample variance, (could you give this general formula?)

    $$s^2 = \frac{8 \cdot 1486 - (8 \cdot 13.5)^2}{8 \cdot 7} = \frac{11888 - 11664}{8 \cdot 7} = 4$$

17. Prove the following, using the MGF technique: If X_1, X_2 are independent rvs, and X_1 has a χ^2 distribution with dof ν_1, and $X_1 + X_2$ has χ^2 with dof $\nu > \nu_1$, then X_2 is χ^2 with dof $\nu - \nu_1$.

    (See Theorem 8.11) Here's a nice proof using MGFs:

$$M_{X_1}(t)M_{X_2}(t) = M_{X_1+X_2}(t) \qquad \Rightarrow \qquad (1-2t)^{-\frac{1}{2}\nu_1}M_{X_2}(t) = (1-2t)^{-\frac{1}{2}\nu} \quad \Rightarrow$$

$$M_{X_2}(t) = (1-2t)^{-\frac{1}{2}(\nu-\nu_1)}$$

    Therefore, X_2 is χ^2 with $\nu - \nu_1$ degrees of freedom.

18. Consider two random variables X and Y whose joint pdf is given by: $f(x,y) = \frac{1}{2}$ if $x > 0, y > 0$ and $x + y < 2$ (zero elsewhere). Find the pdf of $U = Y - X$.

    (See Exercise 7.34)

    We set up the solution to this one in class. The hard part is in figuring out what region in the plane the new PDF is over.

    The pdf was uniform, $f(x,y) = 1/2$ for $x > 0$, $y > 0$ and $x + y < 2$. This is a triangle in Quadrant I. With the change of variables,

$$\begin{array}{ll} u_1 & = x \\ u_2 & = y - x \end{array} \qquad \Rightarrow \qquad \begin{array}{ll} x & = u_1 \\ y & = u_1 + u_2 \end{array} \qquad \Rightarrow \qquad |J| = 1$$

    Therefore, the new joint pdf is:

$$f(u_1, u_2) = \frac{1}{2}$$

    You should plot the region before integrating- We're finding the marginal pdf for y, therefore we need to break the integration into two parts:

    - If $0 < u_2 < 2$, then:

$$g(u_2) = \frac{1}{2}\int_0^{1-\frac{1}{2}u_2} du_1 = \frac{1}{4}(2-u)$$

    - If $-2 < u_2 < 0$, then:

$$g(u_2) = \frac{1}{2}\int_{-u}^{1-\frac{1}{2}u_2} du_1 = \frac{1}{4}(2+u)$$

19. Let X and Y be iid exponential with $\theta = 1$. Let $Z = \frac{1}{2}(X+Y)$. Find the pdf of Z by the CDF technique.

To find the CDF, we compute the following- which we interpret as an integral of the joint pdf.

$$F(Z) = P(Z \leq z) = P(X + Y \leq z)$$

To graph the area in the $x - y$ plane, think of z as a fixed parameter, and we plot $y = -x + z$, which in the first quadrant has both $x-$ and $y-$intercepts equal to z.

Therefore, the probability is found by integrating the joint PDF over the appropriate area:

$$P(X + Y \leq z) = \int_0^z \int_0^{-x+z} \mathrm{e}^{-(y+x)}\,dy\,dx = \int_0^z -\mathrm{e}^{-z} + \mathrm{e}^{-x}\,dx =$$

so that

$$F(z) = -\mathrm{e}^{-z}z - \mathrm{e}^{-z} + 1 \text{ and therefore } f(z) = z\mathrm{e}^{-z} \quad z > 0$$

20. Show that the following is an unbiased estimator of σ^2 (which is the shared variance of X_1, X_2):

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

(See Exercise 11.9)

Plug-n-Chug.. Note that $E(S_1^2) = E(S_2^2) = \sigma^2$, since the variance of X_1, X_2 was said to be shared:

$$E(S_p^2) = \frac{n-1}{(n+m-2)}E(S_1^2) + \frac{m-1}{(n+m-2)}E(S_2^2) = \frac{n+m-2}{n+m-2}\sigma^2 = \sigma^2$$