Math 472: Homework 03

Due Monday Feb 9

Problem 1 (Binomial distribution, revisited)**.** Let $X \sim \text{Bin(n,p)}$. (That is, X is a binomial random variable with number of trials n and success probability p). One way to interpret the random variable X is as follows

$$X = (\text{the number of heads in } n \text{ coin flips}),$$

where p is the probability that the coin lands on heads. Or, more formally,

$$X = \xi_1 + \xi_2 + \ldots + \xi_n$$

where the random variables $\xi_1, \ldots, \xi_n$ are independent with

$$\xi_i = \begin{cases} 1 & : \text{ with probability } p \\ 0 & : \text{ with probability } 1-p \end{cases}$$

for all $i = 1, \ldots, n$.

(a) Use this characterization to compute the expected value and variance of X.

(b) Use `R` to draw a random sample of size 50 from the distribution of X (pick any value of $p \in (0,1)$ that you want) and compute the sample mean and sample variance. Compare with the values you got in part (a).

Problem 2 (`R` problem)**.**

(a) Let $Z_1, \ldots, Z_6$ be a sample of standard normal random variables. Use `R` to find $\mathbb{P}\left[\sum_{i=1}^{6} Z_i^2 \leq 6\right]$.

(b) Let $Y_1, \ldots, Y_{10}$ be a sample of normal random variables with mean μ and variance $\sigma^2 = 1$. Let S^2 be the sample variance. Use `R` to find $\mathbb{P}\left[S^2 \geq 3\right]$.

Problem 3. Refer to example 7.2 in the textbook (i.e., examples 57 in the typed lecture notes). The amount of fill dispensed by a bottling machine is normally distributed with $\sigma = 1$. If $n = 9$ bottles are randomly selected from the output of the machine, we found that the probability that the sample mean will be within .3 ounce of the true mean is .6318. Suppose that Y is to be computed using a sample of size n.

(a) If $n = 16$, what is $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$?

(b) When $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ when $\overline{Y}$ is computed using samples of sizes $n = 25, 36, 49$ and 64.

(c) What patterns do you observe among the values for $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ for the various values of n?

(d) Do the results you obtained in part (b) appear to be consistent with the result obtained in Example 7.3 in the textbook?

Problem 4. Refer to example 7.2 in the textbook. Assume now that the amount of fill dispensed by the bottling machine is normally distributed with $\sigma = 2$ ounces.

(a) If $n = 9$ bottles are randomly selected, what is $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$? Compare this with the answer found in Example 7.2

(b) When $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ when $\overline{Y}$ is computed using samples of sizes $n = 25, 36, 49$ and 64.

(c) What patterns do you observe among the values for $\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq .3\right]$ for the various values of n?

(d) How do the respective probabilities obtained in this problem (where $\sigma = 2$) compare to those obtained in Problem 3 (where $\sigma = 1$)?

Problem 5. A forester studying the effects of fertilization on certain pine forests in the Southeast is interested in estimating the average basal area of pine trees. In studying basal areas of similar trees for many years, he has discovered that these measurements (in square inches) are normally distributed with standard deviation approximately $\sigma = 4$ square inches. If the forester samples $n = 9$ trees, find the probability that the sample mean will be within 2 square inches of the population mean.

Problem 6. Suppose the forester in Problem 5 would like the sample mean to be within 1 square inch of the population mean, with probability .90. How many trees must he measure in order to ensure this degree of accuracy?

Problem 7 (Another `R` problem)**.**

(a) Use `R` to plot histograms comparing the standard normal distribution with t-distribution, for various degrees of freedom (e.g., 5, 10, etc). [The point of this problem is for you to visually observe how the t-distribution converges to the standard normal as the degrees of freedom increase.] The following code may be helpful:

```
## Plotting a histogram of a t-distribution
d = 10 # degrees of freedom for the t-distribution (play around with this)

# generate a ton of samples from t and z distributions
t_samples = rt(n=100000, df=d)
z_samples = rnorm(n=100000,mean=0,sd=1)

# truncate the datasets to only samples within the range (-6,6). This is
# needed to make the plot pretty
z_truncated<- z_samples[z_samples >= -6 & z_samples <= 6]
t_truncated<- t_samples[t_samples >= -6 & t_samples <= 6]

# set histogram box widths to 0.1, and have them range from -6 to 6.
my_breaks<-seq(-6,6, by=.1)

# plot the histograms
hist(z_truncated, probability=TRUE, col=rgb(1,0,0,0.4), breaks=my_breaks,
     border="red", main="t-distribution vs standard normal distribution")

hist(t_truncated, probability=TRUE, col=rgb(0,0,1,0.4),
     breaks=my_breaks, border="blue",add=TRUE)

legend("topright",
       legend = c("approx. distribution of Z",
                  paste("approx. t-distribution with df=",degrees_of_freedom)),
       fill=c(rgb(1,0,0,0.4), rgb(0,0,1,0.4)))
```

(b) Suppose T is a random variable with t-distribution and 5 degrees of freedom. The `R` code `pt(t,df=5)` returns $\mathbb{P}[T \leq t]$. Use this to find the probability that T is greater than 2.

(c) Use `R` to find the probability that T is less than -2.

(d) Find the probability that T is between -2 and 2.

(e) Your answer to part (d) is considerably less than $.9544 = \mathbb{P}[-2 \leq Z \leq 2]$. Provide 1-2 sentence explanation of why.

The remaining problems involve the idea of *quantiles*, defined here:

Definition (p^{th}-quantile)**.** Let Y be a continuous random variable, and fix a real number $p \in (0,1)$. Then the p^{th} ***quantile*** of Y, denoted ϕ_p, is the smallest value such that

$$\mathbb{P}[Y \leq \phi_p] = p,$$

For example, when $p = 1/2$, then $\phi_{\frac{1}{2}}$ is the median. Quantiles are also referred to using the "percentile" terminology, e.g., if $p = .98$, then $\phi_{.98}$ is the 98^{th} percentile (i.e., the cutoff below which lies 98% of the population). When reasoning about quantiles, I find it very helpful to draw pictures of bell curves with shaded areas.

Problem 8. Suppose T is a random variable with t-distribution and 5 degrees of freedom.

(a) For any $p \in (0,1)$, the `R` code `qt(p,df=5)` gives ϕ_p (the p^{th} quantile of T). Let $t_{.10}$ be the number such that $\mathbb{P}[T > t_{.10}] = .10$. Find the value of $t_{.10}$.

(b) What quantile does $t_{.10}$ correspond to? What percentile?

(c) Find the value of $t_{.10}$ for t-distributions with $df = 30, 60$ and 120.

(d) When Z has a standard normal distribution, $\mathbb{P}[Z > 1.282] = .10$, and $z_{.10} = 1.282$. What property of the t distribution (when compared to the standard normal dsitribution) explains the fact that all the values obtained in (c) are larger than $z_{.10} = 1.282$?

(e) Guess what $t_{.10}$ converges to as the number of degrees of freedom gets large.

Problem 9 (Exercise 4.12 in textbook)**.** The length of time to failure (in hundreds of hours) for a transistor is a random variable Y with cumulative distribution function given by

$$F(y) = \begin{cases} 1 - e^{-y^2} & : \quad y \geq 0 \\ 0 & : \quad y < 0 \end{cases}$$

(a) Show that F has the properties of a distribution function (i.e., that $F(-\infty) = 0$, $F(\infty) = 1$, and F is nondecreasing).

(b) Find the .3-quantile, $\phi_{.3}$ of Y.

(c) Find the probability density function of Y.

(d) Find the probability that the transistor operates for at least 200 hours (remember the units are in hundreds of hours. You don't need to evaluate an integral).

(e) Find $\mathbb{P}[Y > 100 \mid Y \leq 200]$. (Again, no need to evaluate an integral :)).

Problem 10 (Refer to Problem 5)**.** Suppose that in the forest fertilization problem the population, standard deviation of basal areas is not known and must be estimated from the sample. If a random sample of $n = 9$ basal areas is to be measured, find a statistic A such that

$$\mathbb{P}\left[\left|\overline{Y} - \mu\right| \leq A\right] = .90$$

Assume that neither the mean μ nor the variance σ^2 of the population are known.