STAT 521: Assignment 6

Make sure to show your computation and/or attach appropriate output.

Problem 1

Conduct the most appropriate test for Example 8 (page 16) in Lecture 8 note. Make sure to state null and alternative hypotheses. Is the test one-tailed or two-tailed? What is your conclusion?

Problem 2 (Computer exercise)

Researchers investigated spinal canal dimensions in 30 subjects symptomatic with disc herniation selected for a discectomy (Group 1) and 46 asymptomatic individuals (Group 2). One of the areas of interest was determining if there is a difference between the two groups in the spinal canal cross-sectional area (cm^3) between vertebrae L5/S1. Do data provide evidence to conclude that a difference in the spinal canal cross-sectional area exists between a population of subjects with disc herniations and a population of those who do not have disc herniations?

a) State the null and alternative hypotheses.

b) Compute the mean and standard deviation of the spinal canal cross-sectional area in symptomatic and asymptomatic subjects.

c) Produce box plots to compare spinal canal cross-sectional area between the two groups.

d) Produce a normal probability plot for each of the groups. What do you find?

e) Conduct an appropriate test at $\alpha = 0.05$. What do you conclude? Is the assumption of equal variances met?

f) Produce a 95% confidence interval for the difference of two means. Does the interval contain zero? Would you have expected that it would?

Problem 3

A randomized controlled, double-blind trial was conducted to see if sustained-release bupropion (a pharmaceutical normally used to treat depression) provided benefit over use of the nicotine patch alone in helping people stop smoking. The control group (nicotine patch alone) included 244 smokers who wanted to stop smoking. The treatment group consisting of 245 individuals received sustained-release bupropion in combination with a nicotine patch. After 1 year, 40 individuals in the control group remained smoke-free. In contrast, 87 in the treatment group had done the same.

a) Calculate the incidence proportions of cessation in the groups.

b) Test the difference in proportions for significance at $\alpha = 0.05$. State the null and alternative hypotheses. What is your conclusion?

Problem 4 (Computer exercise)

Beney et al. (2002) evaluated the effect of telephone follow-up on the physical well-being dimension of health-related quality of life in patients with cancer. One of the main outcome variables was measured by the physical well-being subscale of the Functional Assessment of Cancer Therapy Scale-General (FACT-G). A higher score indicates higher physical well-being. The data show the baseline FACT-G score and the follow-up score to evaluate the physical well-being during the 7 days after discharge from hospital for 66 patients who received a phone call 48-72 hours after discharge that gave patients the opportunity to discuss medications, problems, and advice. Is there sufficient evidence to indicate that quality of physical well-being significantly changes among patients who receive a phone call?

a) State the null and alternative hypotheses. What test should be used?

b) Use SAS, R or Python to conduct the test to answer the research question. Use $\alpha = 0.05$. Attach the output. What is your conclusion? Does the follow-up phone call make difference in physical well-being?

c) Report a 95% confidence interval of changes in FACT-G score.

Problem 5

Refer to Example 1 in Lecture 8 note. Suppose the mean difference in the trace element between male and female donors is expected to be 6.0 ppm (two-tailed).

a) If you wish to collect equal numbers of male and female donors, how many subjects will be needed? Use $\alpha = 0.05$ and power of 80%. Do this by hand and compare the result using SAS, R, Python or G*Power.

b) Suppose that female donors are harder to find. If you want twice as many males as females in your study, how many male and female donors are needed? Use $\alpha = 0.05$ and power of 80%. Do this by hand and compare the result using SAS, R, Python or G*Power.

Problem 6 (Biostats students only)

Monte Carlo expectation in R:

In Assignment 4, we did a simple example of Monte Carlo simulation. We generated a large number of random values from the standard normal distribution $Z \sim N(0,1)$, and then calculated the fraction of those taking $Z \leq z$ to approximate a normal CDF $F_Z(Z) = P(Z \leq z)$. We can use the same approach to *estimate the expected value* of any function of a random variable X, as long as its PDF $f_X(x)$ is known.

Suppose we wish to know the expected value of some function of a standard normal random variable Z, i.e., $E[g(Z)]$ where $Z \sim N(0,1)$. In order to approximate this by Monte Carlo, we can:

1. Generate a large number of random values from $Z \sim N(0,1)$
2. For each random value of z, calculate $g(z)$
3. Approximate $E[g(Z)]$ by taking the average of all $g(z)$, i.e., $E[g(Z)] \approx \frac{1}{n}\sum_{i=1}^{n} g(z_i)$ where n is the number of random draws.

Example: Suppose we wish to know the expected value of Z^2 where $Z \sim N(0,1)$, i.e., $E[g(Z)] = E(Z^2)$. From the definition of the expected value, you may be tempted to solve $E(Z^2) = \int_{-\infty}^{\infty} z^2 f_Z(z)\, dz$ where $f_Z(z)$ is the PDF of Z, but this is not easy to do. So let's approximate $E(Z^2)$ using R:

```r
# Generate one million random values from Z ~ N(0, 1)
z <- rnorm(10 ^ 6)

# Take a look at the first 6 random values
head(z)
```

```
## [1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774  1.71506499
```

In the R code above, first I generated a vector called `z` that has 1 million of random values from $Z \sim N(0, 1)$. The second command, `head()`, displays the first 6 values of the vector. Now let's square each of the values in the vector.

```r
# Square each value of the vector
z.sq <- z ^ 2
head(z.sq)
```

```
## [1] 0.314132950 0.052981677 2.429571609 0.004971433 0.016715318 2.941447909
```

Notice the first line above where I take the square (`^2`). A nice thing about R is that if you square the vector, each value of the vector gets squared (actually, any arithmetic operators work the same way). The squared vector is saved into the object which I named `z.sq`. Take a look at the first value of `z.sq`: The first value is 0.314133, which is $(-0.5604756)^2$

Thus, now I have one million random values of Z^2. If I want to estimate $E(Z^2)$, all I have to do is take the mean of `z.sq`.

```r
# Take the average
mean(z.sq)
```

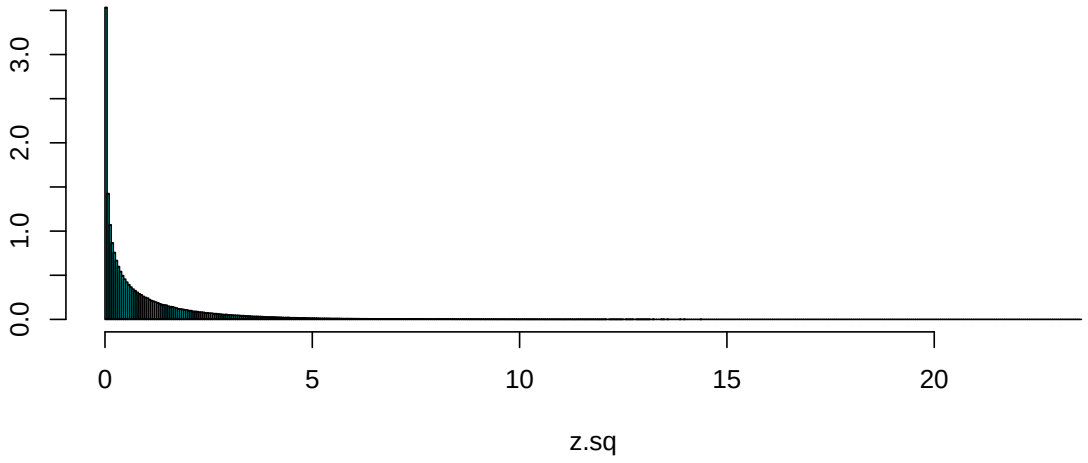
```
## [1] 0.9998534
```

In this case, I got $E(Z^2) \approx 0.9998534$. This is very close to the true value of $E(Z^2) = 1$.

By the way, how do I know $E(Z^2) = 1$? It's actually very simple. You already know for a standard normal random variable Z, its expected value is $E(Z) = 0$ and variance is $var(Z) = 1$ and that is why we write as $Z \sim N(0, 1)$. You also know, from Assignment #2, $var(Z) = E(Z^2) - [E(Z)]^2$. Solving for $E(Z^2)$, $E(Z^2) = var(Z) + [E(Z)]^2 = 1$.

Now take a look at a histogram of Z^2 using `truehist()` command from `MASS` library.

```r
# See a histogram of Z^2
library(MASS)
truehist(z.sq)
```



You will see a highly right-skewed distribution like above. As it turns out, Z^2 follows the chi-square χ^2 distribution of degree of freedom of 1: $Z^2 \sim \chi^2$. The χ^2 distribution has an expected value equal to its *df*, and that's why $E(Z^2) = 1$.

Then how about the variance of Z^2? Again, we can approximate the variance using `z.sq`. To get the variance, use `var()` function:

```
# Variance of Z^2
var(z.sq)
```

```
## [1] 2.000237
```

In this case, I got $var(Z^2) \approx 2.0002369$. This is very close to the true variance of $var(Z^2) = 2$. The χ^2 distribution has a variance equal to $2 \times df$.

Here's some exercise for you: We want to approximate $E(W)$ where $W = Z_1^2 + Z_2^2$ and (Z_1, Z_2) are independent standard normal random variables. Please submit your R code as well as answers:

1. Generate 1 million random values from $Z \sim N(0, 1)$. Save the vector to an object named `z1`. Take the square of `z1`.
2. Generate another 1 million random values from $Z \sim N(0, 1)$. Save the vector to an object named `z2`. Take the square of `z2`.
3. Add squares of `z1` and `z2`, element by element. In other words, for each $i, i = 1, 2, \cdots, 10^6$, calculate $Z_1^2 + Z_2^2$. To do this, you simply need to add two squared vectors using `+` operator. Save the added vector into an object named `w`.
4. Approximate $E(Z_1^2 + Z_2^2)$ and $var(Z_1^2 + Z_2^2)$. Also approximate the distribution of $Z_1^2 + Z_2^2$ by creating its histogram.
5. A random variable $W = Z_1^2 + Z_2^2$ follows the χ^2 distribution with $df = 2$. What are the true expected value and variance of W? Are they very close to your approximation?