StatR 101: Fall 2012
Homework 9
Eli Gurarie
Due Thursday, November 29

Instructions: Please submit a single document with all the R code, short answers and figures. Upload the completed homework assignment into the course webpage drop-box. The homework follows Lab 9 closely. Don't hesitate to discuss the problems amongst yourselves on the forum.

The inference theory that we presented in class relies on the assumption that the distribution of the sample mean $\overline{X}$ is approximately normal. We will test this assumption wy working with the earthquake waiting data, which we know are highly non-normal. Recall that in order to load and convert these data you use the following code:

```
earthquakes <- read.csv("http://neic.usgs.gov/neis/gis/qed.asc")
DateTime <- strptime(paste(earthquakes$Date, earthquakes$Time),
                     format="%Y/%m/%d %H:%M:%S")
W <- difftime(DateTime[-length(DateTime)], DateTime[-1], units="mins")
```

Assume that the true standard deviation of earthquake waiting times is $\sigma = 38$ min, and the true mean $\mu = 29$ min.

1. **Point estimates and confidence Intervals:**

(a) Do Exercise 2 from the lab: i.e. Write a function that estimates a point estimate and confidence interval at a given confidence level from a vector X with known standard deviation. Note, that you can do this either "legitimately", by writing the function, or you may "cheat" by looking at the `Lab9.Rmd` file where I buried a function called `GetCI()`.

(b) Use this function to obtain a point estimate and 95% and 50% confidence intervals of three random samples from W of size 3, 10, and 30 respectively (i.e. `W.5 <- sample(W,5)`, `W.10 <- sample(W,10)`, and `W.30 <- sample(W,30)`).

(c) Test the validity of your confidence level by repeating this procedure 10,000 times, and count the proportion of times the 95% and 50% confidence intervals span the true mean. Comment on these results, with respect to the assumptions behind our construction of the confidence interval.

2. **Hypothesis Tests**: In Homework 8, Problem 3, we tested the hypothesis that the seismic activity of the world is increasing in anticipation of the apocalypse. We based on an observation that we observed mean waiting time between 10 earthquakes in a row was 15 minutes, compared to a known mean and standard deviation of 29 and 39 minutes respectively. Here, we decompose this hypothesis test into its constituent parts:

(a) State the null and alternative hypotheses, in words and symbolically.

(b) Last week we tested the sample mean $\overline{X}$ directly. This time, we would like to perform the test on the z_{test} statistic:

$$z_{test} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{1}$$

Compute this statistic.

(c) Name the approximate distribution of this statistic under the null hypothesis.

(d) Express the p-value of this test symbolically, and compute it. At a $\alpha = 5\%$ significance level, do you reject the null hypothesis?

3. **Bonus problem on t-tests**

(a) You can generate a sample of size 10 with mean 15 from an exponential distribution using, e.g., `W.obs <- rexp(10,rate=1/15)`[1]. Perform a t-test of `W.obs` for the hypothesis in problem 2(a) using the `t.test()` function, and report the result.

(b) Unlike our original test, most of the time the test above will reject the null hypothesis. This is somewhat surprising because the t-test is generally more conservative than the z-test. Think about why that may be. Recall that the t-statistic in this case is:

$$t_{obs} = \frac{\overline{X} - \mu_0}{s_x/\sqrt{n}} \tag{2}$$

where s_x is the sample standard deviation estimated from data X.

(c) Perform a t.test of the following observed data against the $\mu_0 = 29$ minutes:

```
      W.obs <- c(0.9,4.2,0.1,1.8,0.9,7.3,0.4,6.5,1.8,123)
```

Compare this p-value that the original z-test. Explain why it is larger, smaller, or equal.

4. **Final Project**: Write a brief proposal for the final project. Outline a larger research or synthesis question, the data you would like to analyze, what the source of the data will be. Optionally, also outline your proposed analysis methods. The more detailed the proposal, the more useful feedback I will be able to provide.

[1] Note: As a bonus-bonus problem, think of a way to produce an exponentially distributed sample whose mean is coerced to exactly 15