Problem Set 3

LING280 Winter 2013

Distributed 23 January, due 1 February via email to mwagers@ucsc.edu by 5pm.

1. Warm-up: central limit theorem.

In this problem, you will demonstrate via simulation that the average wait time observed in a sample of wait times approaches in distribution the normal distribution.

Consider, first, the *exponential distribution*. The exponential distribution is a continuous probability distribution defined as

> $f(x) = \lambda \cdot e^{-\lambda \cdot x}$ for $x \geq 0$; and 0 otherwise

This distribution is commonly used to model interarrival times for events that occur at a constant rate per unit time (i.e., wait times for such processes as spikes from a firing neuron, failure of a mechanical part, etc. - technically, these are events whose counts follow a *Poisson* distribution)[1].

The exponential distribution has one parameter, λ, which describes the rate -- for example, if you receive emails at a constant rate of 10 per hour during the work day, then λ = 10 emails/hour.

1. Using the `rexp` function in R, figure out what the relationship is between rate (λ), mean (μ) and standard deviation (σ)? Create 1 × 2 plot showing the PDF and CDF. Mark the position of μ in both. What is $P(X \leq \mu)$?

   > Hint: investigate the function `density`. Remember that `lines()` will overplot.

2. Simulate a very large population of wait times for some λ.
3. Demonstrate that mean wait times of sample from this population approaches a normal distribution as *n*, the sample size, increases.
4. Use the following code snippet to help make your argument. `n.sim` is the number of sample draws you simulated. `my.means` is a vector of sample means. In your response, embed the snippet in an appropriately commented function. Give the `plot` command additional arguments so that any resulting visualization is helpfully labeled.

   ```
   > ps <- ((1:n.sim) - 0.5) / n.sim
   > qs <- qnorm(ps)
   ```

```
> plot(qs, sort(my.means))
```

2. Comparing two samples

Download the data file from http://people.ucsc.edu/~mwagers/ling280/data/ps3.Rdata. Use the `ls` function to inspect the objects in the workspace.

1. Consider the vector **condition.a**: suppose these simulated data represent reaction times in a priming task from 64 individuals. Perform a *t*-test on this vector that answers the question: does the sample in **condition.a** come from a population with a mean value of 370? Because this is a single-sample test, do this part without using an in-built R function.
2. What is the 95% confidence interval from your test in (1)? Show via simulation how the width of this interval changes as sample size increases from 10 to 64. Assuming there were a true difference between **condition.a** and a population of mean 370, below what number of experimental participants would you have mistakenly concluded that there probably was no difference?
3. Use the `t.test` function of R to perform a two-sample t-test, comparing **condition.a** to **condition.b** (corresponding to reaction times from the same 64 individuals, in another condition of the priming task). Report and interpret the results of this test. Are **condition.a** and **condition.b** significantly different?
4. The following commands will give you different answers. Why? What implications, if any, can you draw about how to design experiments?

    ```
    > t.test(condition.a, condition.b)
    > t.test(condition.a - condition.b)
    ```

> Some or all of the following calculations will help you make your argument.

```
> mean(condition.a) - mean(condition.b)
> mean(condition.a) - mean(sample(condition.b, replace=FALSE))
> sqrt((var(condition.a)+var(condition.b))/64) -> x
> sqrt(var(condition.a - condition.b)/64) -> y
> t.test(condition.a - sample(condition.b, replace=FALSE)) -> my.test
> my.test$statistic * x
```

3. Materials design

Kayne (1983)[2] is about the theory of empty categories. It annotates the following pair of sentences with a ? and a * respectively (*pg*: parasitic gap; *t*: trace).

(a) ? [a person]$_i$ that people who read descriptions of *pg*$_i$ usually end up fascinated with *t*$_i$

(b) * [a person]$_i$ that people to whom descriptions of *pg*$_i$ are read usually end up fascinated with *t*$_i$

This contrast is used to motivate a distinction in English between empty categories on left and right branches. In this problem, you will propose an experimental design that would allow this claim to be tested against a larger sample of English-speakers, and, more importantly, to understand the extent to which the contrast between (a)/(b) is due to the position of *pg*.

1. What are the ways in which (a) and (b) are distinct from one another *other* than the configuration of the parasitic gap?
2. Considering the two positions in (a)/(b) currently occupied by the empty categories with subscript *i*, enumerate four versions each of (a) and (b) that vary whether or not an empty category is present or a pronoun is there instead.
3. Enumerate four versions each of (a) and (b) that vary whether or not the two empty category positions in (a)/(b) even exist. (What do you have to change to vary this?)
4. Would patterns of acceptability across any of the alternative sentences you constructed in 2. or 3. allow you to determine whether or not subject-embedded parasitic gaps on right branches are better than subject-embedded parasitic gaps on left branches? Why or why not? Which versions are crucial?
5. Based on your answer to 4., select either the sentence set from 2., the sentence set from 3., or some alternative version, and create four more just like it. To begin this process, consider the lexical item choices Kayne (1983) made: are any particularly clever, or are any suboptimal? can you do any better? (Hint: start by thinking about *read*).
6. What design do your sentence sets instantiate? I.e., what are the factors and what are their levels? Describe a pattern of pairwise comparisons across individual factors or levels that would confirm Kayne's claim. Describe a pattern of comparisons that would undercut it.

Footnotes

1. Distributions of interarrival times are useful in studying language as well -- for example, we might care to ask questions like how much time/how many 'units' elapse between certain linguistic features/morphemes/words re-occur? The exponential distribution is really a poor approximation for questions about spacing of linguistic features because a given linguistic feature can occur in bunches and introduce dependencies (for reasons you should be familiar with). See, e.g., Altmann EG, Pierrehumbert JB, Motter AE (2009) Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLoS ONE* 4(11): e7678. doi:10.1371/journal.pone.0007678.
2. Kayne, R.S. (1983). Connectedness. *Linguistic Inquiry, 14*, 223-248.