1 Finding confidence intervals

We discussed in class confidence intervals of the form

$$\hat{p} \pm z^* \mathsf{SE}(\hat{p}), \quad \bar{X} \pm z^* \mathsf{SE}(\bar{X}),$$

and the following type

$$[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}]$$

follows in a similar manner.

For each of these, we need to be able to find the quantities $\hat{p}, \bar{X}, S$ and z^*, a, b and t^*. To find the first 3 we use `mean` or `sd` as appropriate. The z^* and t^* and a and b require us to learn about quantiles.

1.1 Finding quantiles

For example, z^*, solves the following problem

$$P(-z^* < Z < z^*) = 1 - \alpha, \text{ or } P(Z < z^*) = 1 - \frac{\alpha}{2}$$

The latter by symmetry. Thinking about quantiles, we note that for a *continuous* distribution, the q quantile, x, solves $F(x) = q$, where F is the cumulative distribution function. Funny, that is exactly what we have above, where $F(z) = P(Z < z)$ for the normal.

Thus to generate z^*'s for values of α we have (note the funny way I chose to define α.)

```
 > alpha = 1-.68;qnorm(1-alpha/2)
[1] 0.9944579                    # 68% within 1 SD
> alpha = 1-.95;qnorm(1-alpha/2)
[1] 1.959964                     # 95% within 2 SD
> alpha = 1-c(.68,.95,.998);qnorm(1-alpha/2)
[1] 0.9944579 1.9599640 3.0902323
```

For t^* and a and b it is not different conceptually, you just need to know that the distributions (the t and χ^2) need a degrees of freedom parameter. Here we find the 90% values for t and for χ^2 with 10 degrees of freedom:

```
> alpha = 1-.9;qt(1-alpha/2,df=10)
[1] 1.812461
> a = qchisq(.05,df=10);b=qchisq(.95,df=10);c(a,b)
[1]  3.940299 18.307038
```

(Why did I use .05 and .95 in the latter, and what are they in terms of α?)

1.2 Putting it together

Now to make a confidence interval is easy. Suppose we have the following data.

- ($\hat{p}$, α, n) We can construct as follows

```
> phat = .42;n=500;alpha = 1 - .90
> SE = sqrt(phat*(1-phat)/n)
> zstar = qnorm(1-alpha/2)
> c(phat - zstar*SE, phat + zstar*SE)
[1] 0.3836938 0.4563062
```

- ($\bar{X}$, S,α,n) We have using the t distribution (n is small)

```
> Xbar = 7.2;S=1;alpha = 1-.9;n = 25
> SE = S/sqrt(n)
> zstar = qt(1-alpha/2, df=n-1)
> c(Xbar - zstar*SE, Xbar + zstar*SE)
[1] 6.857824 7.542176
```

- (S,α,n) For estimating σ we have

```
> S = 5;alpha = 1 - .9; n=15;
> ab = qchisq(c(alpha/2,1-alpha/2),df=n-1)
> ab
[1]  6.570631 23.684791
> (n-1)*S^2 / rev(ab)           # rev to switch
[1] 14.77742 53.26733
```

1.3 Built in functions

If we have the data – not the summarized values – we can use the following built-in commands to do the work: `prop.test`,`t.test`. (There is not a `z.test` or a test for the variance for one-sample). The `prop.test` is a little different, but does the same thing as we want.

All are easy to use. Let's do so with random data that we generate to see if we get the right answers:

First for proportions. The values are 0 or 1:

```
n = 1000;p = .4;
> x  = rbinom(n,1,p)
> prop.test(sum(x),n)

    1-sample proportions test with continuity correction

data:  sum(x) out of n, null probability 0.5
X-squared = 42.025, df = 1, p-value = 9.011e-11
```

```
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3666330 0.4281683
sample estimates:
    p
0.397
```

Notice, this confidence interval contains the true value of $p = .4$. This only happens 95% of the time for the default confidence level. To change the confidence level, you would add the switch `conf.level` as follows:

```
> prop.test(sum(x),n,conf.level=.9)
```

Next, for the t-test. Notice we "know" $\mu = 5$.

```
> x = rnorm(15,mean=5,2)
> t.test(x)

    One Sample t-test

data:  x
t = 8.9948, df = 14, p-value = 3.412e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.456718 5.621360
sample estimates:
mean of x
 4.539039
```

2 A problem

You are working on a computer program and part of it involves a *hard* mathematical computation (such as factoring into prime factors). You are clever and write an algorithm using random numbers that can factor quickly, but may not always get the correct answer. (This is not an uncommon practice for computer scientists.)

Suppose you have the following knowledge.

1. You chose 100 numbers at random from the interval 1 to a billion and found that 93% of the time the factorization was correct. Based on this, what is 95% confidence interval for the true proportion of correct times?

2. The "run" time seems to vary randomly with the input number. You ran the program 10 times on randomly chosen numbers to investigate this and found the following run times

   ```
   90 110 106 123 111 100 113 103  92 111
   ```

(a) Check graphically to see if these look like they are from a normal sample.

(b) Find a 90% confidence interval for the mean run time. What test did you use and why?

3. Now find a 80% confidence interval for the standard deviation σ based on your sample. What assumptions about the data are you making?

4. You show you friend your algorithm, and she modifies it a bit, and finds that she gets run times of

```
117  86  93  98  96  79 101  85  81  94
```

(a) Again check for normality

(b) Find a confidence interval for her mean

(c) Find a 90% confidence interval for the mean difference of the two algorithms. Does it include 0? (Read the help pages to see the easy way to do this.)

5. Your teacher notes that you have an algorithm that actually grows like $n\log(n)$. To check if this is correct, you take 10 randomly chosen values of n, compute the time it takes and then subtract $n\log(n)$. If your teacher is correct, your times should have mean 0. Suppose your data is

```
 13 -8 -4 17  7 -2  4  0  3  8
```

Find a 95% confidence interval and check if it includes 0.