Homework questions, week 5

Econ 103

1 Daily Homework questions

The questions in bold font are due on **Thursday 23rd June**. You do not need to hand in the questions that are not in bold, though these will be useful to complete for your own understanding.

Lecture 16: Confidence Intervals III

Textbook questions:
Chapter 8: **17, 19**, 21

Additional questions:

1. **Suppose you want to construct a 99% confidence interval for the average height of US males above the age of 20. Based on past studies you think the standard deviation of heights for this population is around 6 inches. How large a sample should you gather to ensure that your confidence interval has a width no greater than 1 inch?**

Solution: Assuming the population is normal and σ is known, our confidence interval takes the form:

$$\bar{X}_n \pm \texttt{qnorm}(1-\alpha/2) \times \sigma/\sqrt{n}$$

Thus, the width equals $2(1-\alpha/2) \times \sigma/\sqrt{n}$. From the problem statement $\sigma = 6$. For a 99% confidence interval we set $\alpha = 0.005$. Plugging this into R, we find $\texttt{qnorm}(1 - 0.01/2) \approx 2.58$. Thus, in terms of n, the width of our interval is approximately

$$2 \times 2.58 \times 6/\sqrt{n} \approx 31/\sqrt{n}$$

Solving $1 = 31/\sqrt{n}$ for n gives $n = 961$. Double checking this in R:

```
n <- 950:961
width <- 2 * qnorm(1 - 0.01/2) * 6/sqrt(n)
cbind(n, width)

##         n  width
##  [1,] 950 1.0029
##  [2,] 951 1.0023
##  [3,] 952 1.0018
##  [4,] 953 1.0013
##  [5,] 954 1.0007
##  [6,] 955 1.0002
##  [7,] 956 0.9997
##  [8,] 957 0.9992
##  [9,] 958 0.9987
## [10,] 959 0.9981
## [11,] 960 0.9976
## [12,] 961 0.9971
```

so the exact answer is 956, which is pretty close to what we got using a rounded value for $2 \times$ `qnorm(1 - 0.01/2)` $\times 6$ as we did above.

2. A well-known weekly news magazine once wrote that the width of a confidence interval is inversely related to sample size: for example, if a sample size of 500 gives a confidence interval of plus or minus 5, then a sample of 2500 would give a confidence interval of plus or minus 1. Explain the error in this argument.

Solution: This question involves symmetric confidence intervals, i.e. intervals of the form $\widehat{\theta} \pm ME$. As we have seen in class, the width of such intervals, whether based on the normal or t distributions, depends on $\sqrt{n}$ rather than n:

$$\sigma \text{ Known:} \quad \bar{X}_n \quad \pm \quad \texttt{qnorm}(1-\alpha/2) \times \frac{\sigma}{\sqrt{n}}$$

$$\sigma \text{ Unknown:} \quad \bar{X}_n \quad \pm \quad \texttt{qt}(1-\alpha/2, df = n-1) \times \frac{S}{\sqrt{n}}$$

Thus, all other things equal, we would have to quadruple the sample size to cut the width of the interval in half. (There is a slight complication that arises from the fact that the quantile of the t interval also involves n, but as discussed in class, this only makes a practical difference in the confidence interval when n is very small.)

Lecture 17: Confidence Intervals IV

Textbook questions:
Chapter 8; 25

Additional questions:

Table 1: Fictional summary statistics for midterms last semester

	Midterm 1	**Midterm 2**	**Difference**
Sample size	100	100	100
Sample mean	110	100	10
Sample standard deviation	20	25	
Sample variance	400	625	
Sample correlation	0.5		

1. **Calculate the variance of differences between midterm 1 and midterm 2 scores**

> **Solution:** $var(d) = var(m1) + var(m2) - 2sd(m1)sd(m2)corr(m1, m2) = 400 + 625 - 2 * 0.5 * 20 * 25 = 525$

2. **Suppose we want to know whether midterm 2 was 'harder' than midterm 1. Using your answer to the previous question, and the numbers in the table, write down an expression for the 95% confidence interval for the difference between scores on the midterm.**

> **Solution:** $10 \pm 2\sqrt{525/100} \approx 10 \pm 2\sqrt{5.25} \Rightarrow (5.41, 14.58)$

Lecture 18: Hypothesis Testing I

Textbook questions:
Chapter 9: **7, 9**

Past exam question: Final 2012, Q10

Lecture 19: Hypothesis Testing II

Textbook questions:
Chapter 9: 19 (a,b only), **27, 29**

Additional questions:

1. In April of 2013, Public Policy Polling carried out a survey of 1247 registered voters to determine whether Republicans and Democrats differ in their beliefs about various conspiracy theories. To answer this question, you'll need to download the full results of their survey which are on Prof DiTraglia's website:
   `http://www.ditraglia.com/econ103/conspiracy.pdf`
   In this question you'll use the data to carry out hypothesis tests. Throughout you may assume that the sample size is large enough for the approximation based on the central limit theorem to be valid.

   (a) Suppose we wanted to test the null hypothesis that 20% of registered voters believe that a UFO crashed at Roswell, New Mexico in 1947 and the US Government covered it up. There are two possible test statistics we could use. Calculate them both and explain the difference. Which is preferable?

   Solution: Overall percentages appear on page 2 of the report, and this question refers to Q3. The sample size is 1247 and $\widehat{p} = 0.21$

```r
p.hat <- 0.21
n <- 1247
```

   We calculate the numerator of the test statistic as follows

```r
p.null <- 0.20
numerator <- p.hat - p.null
```

   For the denominator we need the standard error of $\widehat{p}$. There are two possibilities. The first is to use the estimated standard error

```r
n <- 1247
SE.est <- sqrt(p.hat * (1 - p.hat)/n)
```

   The second option is to use the exact standard error *under the null hypothesis*

```r
SE.0 <- sqrt(p.null * (1 - p.null)/n)
```

   The two test statistics are as follows:

```r
test.stat <- numerator / SE.est
test.stat.refined <- numerator / SE.0
test.stat
## [1] 0.8669819
test.stat.refined
## [1] 0.8828222
```

The refined test statistic is preferable since it *fully imposes* the null hypothesis. This is the test statistic that we will use below.

(b) Suppose that we wanted to test the null hypothesis from the preceding part against the one-sided alternative that more than 20% of registered voters believe in the UFO conspiracy. Calculate the p-value for this test.

Solution:

```r
1 - pnorm(test.stat.refined)
## [1] 0.1886662
```

(c) Repeat the preceding part for the *two-sided* alternative.

Solution:

```r
2 * (1 - pnorm(test.stat.refined))
## [1] 0.3773324
```

(d) Calculate the p-value for a test of the null hypothesis that equal proportions of Romney and Obama voters believe in the UFO conspiracy against the two-sided alternative. There are two test statistics you could use. Calculate the p-value using each and explain the difference. Which should we prefer?

Solution: Percentages broken down by 2012 vote appear in page 5 of the survey results. Overall percentages of Romney and Obama voters in the sample appear on page 3. Of the 1247 registered voters in the sample, 50% voted for Obama and 44% voted for Romney. We'll call this $n_O = 623$ and $n_R = 547$. The sample proportions are $\widehat{p}_O = 0.16$ for Obama voters versus $\widehat{p}_R = 0.27$ for Romney voters:

```
n.R <- 547
p.R <- 0.27
n.O <- 623
p.O <- 0.16
diff <- p.R - p.O
```

The two statistics correspond to different ways of calculating the standard error of the difference of sample means. The first possibility is to use the estimated standard errors for each population and combine them using the independence of the samples, as we did when constructing confidence intervals:

```
SE.R <- sqrt(p.R * (1 - p.R)/n.R)
SE.O <- sqrt(p.O * (1 - p.O)/n.O)
SE <- sqrt(SE.R^2 + SE.O^2)
```

The second possibility is to construct a *pooled* estimator of the standard error based on a *pooled* sample proportion. This is preferable because it fully imposes the null hypothesis:

```
n.total <- n.O + n.R
p.pooled <- ((n.O * p.O) + (n.R * p.R)) / n.total
SE.pooled <- sqrt(p.pooled * (1 - p.pooled) * (1/n.O + 1/n.R))
```

The resulting test statistics are as follows:

```
test.stat <- diff / SE
test.stat.refined <- diff / SE.pooled
test.stat
## [1] 4.583097
test.stat.refined
## [1] 4.597651
```

The two test statistics are quite similar in this particular example and both p-values are essentially zero:

```
2 * (1 - pnorm(test.stat))
## [1] 4.581394e-06
2 * (1 - pnorm(test.stat.refined))
## [1] 4.272809e-06
```

Using either test statistic, we would find extremely strong evidence against the null hypothesis.

Lecture 20: Hypothesis Testing III

Textbook questions:
Chapter 9: **23, 25**

Additional questions:

1. **Professor Neil is interested in determining whether viewing different colors affects subjects' mental states in a way that alters their athletic ability. As a part of her research she carries out the following experiment. Each subject is randomly assigned to wait in one of two rooms: a room in which all of the walls have been painted pink or another in which all of the walls have been painted red. After waiting for five minutes, each subject is taken to a track and asked to run a 5K as fast as possible. Using the data from this experiment, Professor Neil carries out a statistical test of the null hypothesis that the population mean 5K time is equal across groups (those who waited in the pink room versus those who waited in the red room). Testing at the 1% level, she finds a statistically significant difference. For each of the following, answer True or False. If false, explain.**

   (a) The p-value for the null hypothesis that population means are equal across groups is greater than 0.01.

   > **Solution:** False: the p-value is *less than or equal to* 0.01.

   (b) Professor Neil would also have found a statistically significant difference had she carried out her test at the 5% level.

   > **Solution:** True.

   (c) If there were really no difference in population means across the two groups, the chance of observing a test statistic at least as extreme as that observed by Professor Neil would be 0.01 or less.

   > **Solution:** True.

   (d) Professor Neil's findings have important practical implications for sports regulatory organizations such as the International Olympic Committee: all locker rooms should be painted exactly the same color to keep from throwing off the outcomes of sporting events.

Solution: False. Professor Neil has found strong evidence of a difference in population means across the two groups. However, none of the information given above provides any indication of whether this difference is large enough to have any practical importance. In the words of the textbook "statistical significance and practical significance are two entirely different matters." For example, the difference of population means could be one second. This is far too small to be likely to change the outcome of a 5K race, but with a large enough sample size we would still be able to detect it. Unlike a confidence interval, which gives us a range of plausible values for the difference in population means, a hypothesis test merely tells us whether we have strong evidence that a difference exists.

Lecture 21: Hypothesis Testing IV

Textbook questions: None

Past exam question: Final, Spring 2013, Q7

2 R Tutorials

You should complete R Tutorial #5 by **Thursday 23rd June**.

R tutorials will be posted in Piazza, with solution code.