Problems class 7-Feb-2013

(This handout provides detailed solutions using R of problems mostly done in class. Problem 2 is slightly different (part (a) we didn't do in class, and does not strictly belong here).)

Solving problems with R

1. Assume you have a regular coin ($P(\text{heads}) = P(\text{tails}) = 0.5$).

   (a) What is the probability that you get 6 or more heads in 10 throws?

   (b) What is the probability that you get 60 or more heads in 100 throws?

   (c) What is the probability that you get 600 or more heads in 1000 throws?

   Answer:

   Let's compute $P(X \geq 6) = 1 - F_X(5)$ for a binomial $b(p = 0.5, n = 10)$ distribution:

```
> 1 - pbinom(5, 10, 0.5)

[1] 0.3769531
```

   Similarly, for the second and third question:

```
> 1 - pbinom(59, 100, 0.5)

[1] 0.02844397

> 1 - pbinom(599, 1000, 0.5)

[1] 1.364232e-10
```

   Six heads or more in ten throws is quite likely. Sixty or more out of $n = 100$ is fairly rare. Six hundred or more out of $n = 1000$ has such a small probability that we would hardly be willing to believe the coin is regular!

   Let's compute now the normal approximations: we will work through the first case, and then show the results for the remaining two.

$$\begin{aligned} P(X \geq 6) &\approx P\left(\frac{X - np}{\sqrt{npq}} \geq \frac{6 - np}{\sqrt{npq}}\right) \\ &= 1 - \Phi\left(\frac{6 - np}{\sqrt{npq}}\right) = 1 - \Phi\left(\frac{6 - 10 \times 0.5}{\sqrt{10 \times 0.25}}\right) \\ &\approx 1 - \Phi(0.6325) \end{aligned}$$

If we compute this last value,

```
> 1 - pnorm(0.6325)

[1] 0.2635301
```

we get a rather poor approximation to the exact value computed before. We can try a continuity correction to improve the approximation,

$$P(X \geq 6) \approx 1 - \Phi\left(\frac{6 - \frac{1}{2} - 5}{\sqrt{2.5}}\right) = 1 - \Phi(0.3162278)$$

which can be readily computed as:

```
> 1 - pnorm(0.3162278)

[1] 0.3759148
```

For small n, the continuity correction really matters. Let's see what the respective values are in the remaining cases:

$$P(X \geq 60) \approx 1 - \Phi\left(\frac{60 - 50}{\sqrt{100 \times 0.25}}\right)$$

```
> 1 - pnorm( 10   / sqrt(100*0.25) )

[1] 0.02275013

> 1 - pnorm( 9.5 / sqrt(100*0.25) )

[1] 0.02871656
```

$$P(X \geq 600) \approx 1 - \Phi\left(\frac{600 - 500}{\sqrt{1000 \times 0.25}}\right)$$

```
> 1 - pnorm( 100   / sqrt(1000*0.25) )

[1] 1.269814e-10

> 1 - pnorm(  99.5 / sqrt(1000*0.25) )

[1] 1.557618e-10
```

2. There are two candidates A and B running for office in an upcoming election. You want to estimate the proportion of people who will vote for A; you interview a sample of 1000 randomly chosen individuals, out of which 550 declare they will vote A.

(a) Give a 95% confidence interval for the true proportion p of people willing to vote A.

(b) What is the probability of getting 550 or more A-voters out of 1000 if, in fact, only 45% of the people are willing to vote for A?

Answer:

We know that the number X of A-voters is distributed as $b(p, n = 1000) \approx N(np, \sigma^2 = npq)$. Therefore, X/n (the *binomial frequency*) is distributed approximately as $N(p, \sigma^2 = pq/n)$. Hence,

$$P\left(z_{\alpha/2} \le \frac{X/n - p}{\sqrt{pq/n}} \le z_{\alpha/2}\right) \approx 1 - \alpha$$

which gives the following interval for p:

$$P\left(\frac{X}{n} - z_{\alpha/2}\sqrt{pq/n} \le p \le \frac{X}{n} + z_{\alpha/2}\sqrt{pq/n}\right) \approx 1 - \alpha$$

```
> Xn    <- 550/1000
> zalfa <- qnorm(0.975)
> zalfa

[1] 1.959964

> pmerror <- zalfa * sqrt(0.25/1000)
> c(Xn - pmerror, Xn + pmerror)

[1] 0.5190102 0.5809898
```

If people willing to vote for A is 45%, the number of A-voters in a random sample of size $n = 1000$ is distributed as $X \sim N(m = 450, \sigma^2 = 1000 \times 0.45 \times 0.55)$, so:

$$P(X \ge 550) \approx 1 - \Phi\left(\frac{550 - \frac{1}{2} - 450}{\sqrt{1000 \times 0.45 \times 0.55}}\right)$$

readily computed as:

```
> 1 - pnorm( 99.5 / sqrt(1000*0.45*0.55) )

[1] 1.269158e-10

> 1 - pnorm( 100 / sqrt(1000*0.45*0.55) )

[1] 1.032568e-10
```

If indeed $p = 0.45$, what we have observed is extremely unlikely. The evidence does not quite agree with our assumption, which would cast some doubt on whether that assumption is tenable.

3. You are selling airplane tickets for a plane with 340 seats. You know from experience that 15% of the people who buy a ticket never board the plane, because of last minute problems, late connections, etc.

   (a) If you sell 355 tickets, what is the expected number of people showing up at the boarding gate? What is the probability that you will not have enough room?

   (b) How many tickets can you sell if you want that the probability of not being able to accommodate all passengers be less than 0.01?

   Answer:

   For (a), note $E[X] = 355 \times 0.85 = 301.75$, and $X \sim N(301.74, \sigma^2 = 355 \times 0.85 \times 0.15)$. Therefore,

$$P(X > 340) = 1 - P(X \leq 340) \approx 1 - \Phi\left(\frac{340 + \frac{1}{2} - 301.75}{\sqrt{355 \times 0.85 \times 0.15}}\right)$$

```
> 1 - pnorm( (340 + .5 - 301.75) / sqrt(355 * 0.85 * 0.15) )

[1] 4.212326e-09
```

   How far can you push the overbooking if insufficient seats with probability $p = 0.01$ seems acceptable to you? Now you want n such that

$$P(X > 340) \approx 1 - \Phi\left(\frac{340 + \frac{1}{2} - n \times 0.85}{\sqrt{n \times 0.15 \times 0.85}}\right) < 0.01$$

   or:

$$\begin{aligned} \Phi\left(\frac{340 + \frac{1}{2} - n \times 0.85}{\sqrt{n \times 0.15 \times 0.85}}\right) &> 0.99 \qquad (1) \\ \left(\frac{340 + \frac{1}{2} - n \times 0.85}{\sqrt{n \times 0.15 \times 0.85}}\right) &> \Phi^{-1}(0.99) \qquad (2) \end{aligned}$$

   You can solve by hand (needed in an exam) or just by trial and error with R:

```
> qnorm(0.99)

[1] 2.326348

> ratio <- function(n) { (340.5 - 0.85*n) / sqrt(0.15*0.86*n) }
> ratio(390)
```

```
[1] 1.268865

> ratio(385)

[1] 1.880142

> ratio(382)

[1] 2.250767

> ratio(381)

[1] 2.374963
```

Let's check:

$$P(X > 340) \approx 1 - \Phi\left(\frac{340 + \frac{1}{2} - 381 \times 0.85}{\sqrt{381 \times 0.15 \times 0.85}}\right)$$

```
> 1 - pnorm( (340.5 - 381*0.85) / sqrt(381*0.15*0.85) )

[1] 0.008449617
```

A more elaborate approach would be to note that we are looking for the largest integer value n such that (1) holds. If we solve

$$\begin{align} \Phi\left(\frac{340 + \frac{1}{2} - n \times 0.85}{\sqrt{n \times 0.15 \times 0.85}}\right) = 0.99 \tag{3} \\ \tag{4} \end{align}$$

or equivalently

$$\begin{align} \Phi\left(\frac{340 + \frac{1}{2} - n \times 0.85}{\sqrt{n \times 0.15 \times 0.85}}\right) - 0.99 = 0 \tag{5} \\ \tag{6} \end{align}$$

and then round to one unit less. To do that, we may define a function,

```
> f <- function(n) {
+         pnorm((340.5 - n*0.85)/sqrt(n*0.15*0.85)) - 0.99
+       }
```

and solve as follows:

```
> uniroot(f,lower=340,upper=500)
```

```
$root
[1] 381.5003

$f.root
[1] -4.625811e-11

$iter
[1] 9

$estim.prec
[1] 6.103516e-05
```

`uniroot` is a function that searchs for roots of arbitrary functions in a given interval. For `uniroot` to find a root, function `f` passed as argument has to take values of opposite signs at the ends of the searching interval.

4. An insurance company specializes in fire risks. They charge a premium of 500€ per year per house. The probability that in a year a house catches fire, is 0.002, in which case the indemnity the insurance company has to pay is 200000€. They have insured 10000 houses.

   (a) What is the expected gross profit (excess of premiums over the cost of claims) per year? What is the probability that they incur a loss?

   Answer:

   The company pockets 500€ per house; that makes 5.000.000€ per 10.000 houses insured, no matter what. However, some among the 10.000 houses will catch fire: call that number Z. The gross profit is then

   $$GP = 5000000 - 200000Z$$

   in euros. Z is random, can be seen as the sum of 10.000 independent binaries, therefore $b(p = 0.002, n = 10000)$; hence the gross profit is also random. The expected gross profit (or mean value of the gross profit) is:

   $$E[GP] = 5000000 - 200000 \times (10000 \times 0.002) = 1000000€.$$

   How could the company slip into a loss? If the total cost of claims, $200000Z$ exceeds 5000000, or equivalently if

   $$Z > 5000000/200000 = 25$$

   So the question boils down to: "What is the probability that a random variable distributed as $b(p = 0.002, n = 10000)$ exceeds 25?".

```
> 1 - pbinom(25,size=10000,prob=0.002)
[1] 0.1119618
> 1 - pnorm( (25 + 0.5 - 20) / sqrt(10000*0.002*0.998) )
[1] 0.1091485
```

(b) Assume that the company enters a reinsurance agreement with another similar company (same number of houses insured, same premium, same indemnity in case of fire). They agree to share all premiums and claims 50% each. What is now the expected gross profit and probability of loss for the first company?

Answer:

The company has twice as many houses insured, but the expected profit from each one is only 50% as big. So we would have

$$GP = 20000 \times 250 - 100000Z$$

where Z is now $b(p = 0.002, n = 20000)$, and it is easily checked that $E[GP]$ is the same as before (1000000€). Just as before, the probability of a loss is as before the probability that $Z > 5000000/100000$, i.e. the probability that a $b(p = 0.002, n = 20000)$-distributed random variable exceeds the value of 50:

```
> 1 - pbinom(50,size=20000,prob=0.002)
[1] 0.05245092
```

or if we use a normal approximation, $N(np = 40, \sigma^2 = npq = 39.92)$,

```
> 1 - pnorm( (50 + 0.5 - 40) / sqrt(39.92) )
[1] 0.04827058
```

So the reinsurance aggreement does have an effect! It cuts roughly by half the probability of incurring a loss.

Things to investigate on your own.

1. What exactly is a confidence interval? In problem 2, do we mean that the true p (which we don't know) lies in $(0.5190, 0.5809)$ with probability 0.95? (Hints: Can we, at least in principle, compute the true proportion p of A-voters? What meaning has the statement "With probability 0.95, $p = 0.734$"?)[1].

2. In several instances we have found that, assuming p to have some given value, the probability of an event (like "finding 550 or more A-voters in a sample of 1000") has an extremely small probability. What would you suspect if one of those extremely unlikely events crops up in your experimentation[2]?

3. Everything in this handout you can solve by hand, only it takes longer. Look at equation (2) and satisfy yourself that you would be able to deal with it in an exam.

4. In the last problem, we had a fairly good idea of what values of n to try. Sometimes we don't. Write in R a loop such as:

   ```
   > for(n in 350:390) { print(ratio(n)) }
   ```

   and make sure you understand what it does. (`ratio` must be a function defined as in Problem 3 above.)

5. You have a ton of free documentation about R, that you can obtain from any mirror of the central CRAN archive (you have one mirror right here, at Sarriko, check `http://www.et.bs.ehu.es/cran`.) You can also look at books such as Ugarte et al. (2008) or Dalgaard (2002), available in the library.

References

P. Dalgaard. *Introductory Statistics with R.* Statistics and Computing. Springer-Verlag, 2002. Signatura: 519.682 DAL.

M. Ugarte, A. Militino, and A. Arnholt. *Probability and Statistics with R.* CRC Press, 2008.

[1] We will discuss this in class in an upcoming seminar.

[2] We will return to this when we discuss hypothesis testing, much later in the course. Just ask yourself this question and keep it in the back of your mind.