Simulations, the central limit theorem, and robustness

Patrick Breheny

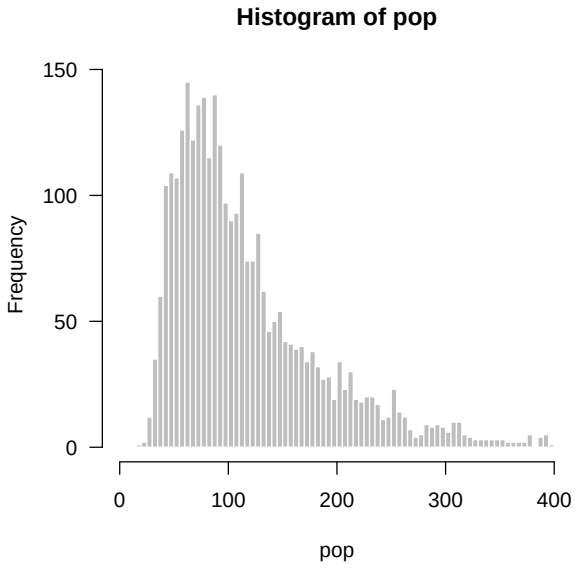
October 13, 2016

Last week we carried out our first simulations, investigating the actual coverage of various methods for constructing intervals for binomial proportions. We continue with simulations today, using them to investigate the central limit theorem as well as how the methods we derived in class (based on the t-distribution for the mean and the χ^2 distribution for the variance) perform when applied to non-normal data.

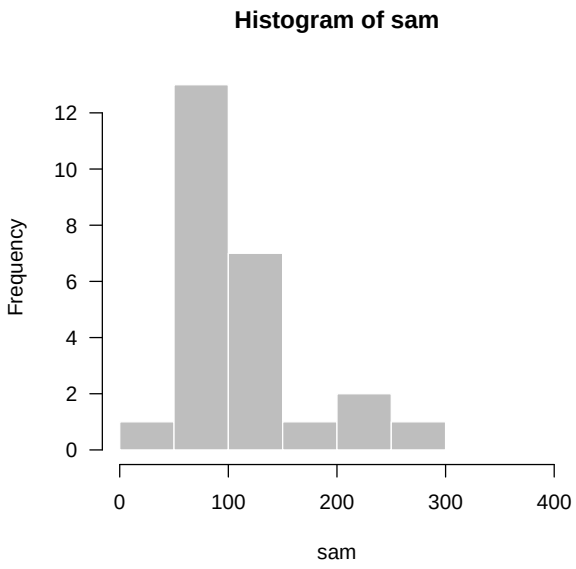
1 Central limit theorem

As part of the NHANES study, the triglyceride levels of 3,026 adult women were measured. Triglycerides, the main constituent of both vegetable oil and animal fat, have been linked to atherosclerosis, heart disease, and stroke. Let's consider this whole group of 3,026 women the "population" for the purposes of our simulation, and that we are going to conduct a study of this population by taking a small sample of, say, 25 women from it. So let's do this and take a look at the distribution of triglycerides in our population and in the sample:

```
> lipids <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lipids.txt")
> pop <- lipids$TRG
> sam <- sample(pop, 25)
> hist(pop, col="gray", border="white", breaks=seq(0, 400, 5), las=1)
```



```
> hist(sam, col="gray", border="white", breaks=seq(0, 400, 50), las=1)
```

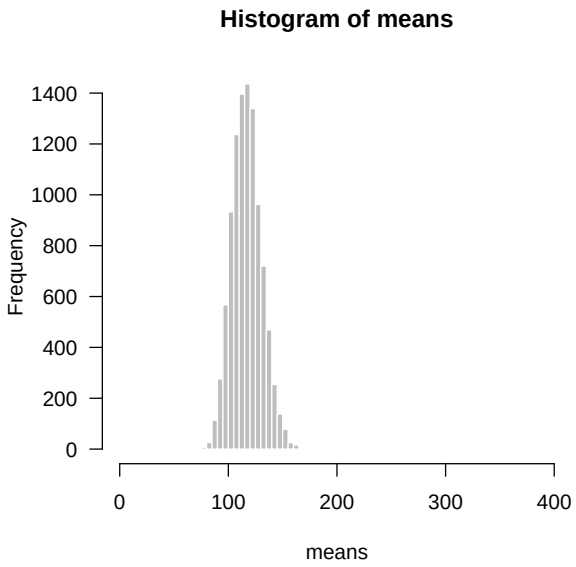



```
> mean(pop)

[1] 116.9451

> mean(sam)

[1] 108.28
```

It's worth noting that (a) the distribution of triglycerides in the population is clearly right-skewed, (b) the sample looks reasonably representative of the population (as it should, since it's a random sample), and (c) the sample mean and population mean are reasonably close, but the sample mean is clearly off by a bit in terms of estimating the population mean. Of course, this is just one sample; the means of other random samples might be much further away from 116.9, or much closer.

What the central limit theorem deals with is the distribution of the sample mean. To see that distribution, we'll have to repeat the above sampling process many times and obtain many sample means. This can be done in `R` using a `for` loop:

```
> N <- 10000          ## Number of simulations to run
> n <- 25             ## Sample size
> means <- numeric(N) ## Setting up an empty vector
> for (i in 1:N) {
+   sam <- sample(pop, n)
+   means[i] <- mean(sam)
+ }
> hist(means, col="gray", border="white", breaks=seq(0, 400, 5), las=1)
```

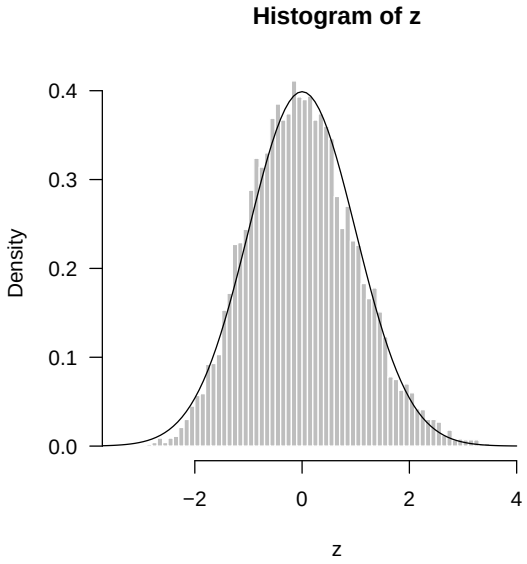


Note that, at least qualitatively, the central limit theorem seems to be holding up:

- The distribution of the means seems centered around the population mean of 117
- The spread of the distribution of the means is clearly much smaller than the spread in the original distribution of TRG values
- The shapes of the distributions are not the same; in particular, the distribution of means looks much less skewed and more normal-like

Let's put it through a more quantitative check, though, to see how exactly the CLT is working out:

```
> mean(means)

[1] 117.0655

> SD <- sd(pop)
> SE <- SD/sqrt(n)
> sd(means)

[1] 13.56125

> SE

[1] 13.58864

> z <- sqrt(n) * (means-mean(pop)) / sd(pop)
> hist(z, col="gray", border="white", freq=FALSE, breaks=99, las=1)
> zz <- seq(-4, 4, length=101)
> lines(zz, dnorm(zz))
```



So the approximation seems pretty good – the distribution isn't *exactly* normal, but it's pretty close, and the expectation and SD calculations match up with our in-class derivations. Let's look at some specific distributional predictions with respect to probability:

```
> ## Probability that a sample mean is less than 100 mg/dL
> mean(means <= 100)

[1] 0.0991

> pnorm(100, mean(pop), SE)  ## Using the location-scale normal

[1] 0.1061973

> pnorm((100-mean(pop))/SE) ## Using the standard normal

[1] 0.1061973

> ## 90th percentile
> quantile(means, .9)

   90%
134.92

> qnorm(.9, mean(pop), SE) ## Using the location-scale normal

[1] 134.3597

> qnorm(.9)*SE + mean(pop) ## Using the standard normal

[1] 134.3597
```

Questions for discussion:

- Try checking the accuracy with respect to: "What's the probability that the sample mean will be between 100 and 150 mg/dL?"

- Each of the above answers is an approximation. Why? Which should you trust? What does the accuracy of each approximation depend on?

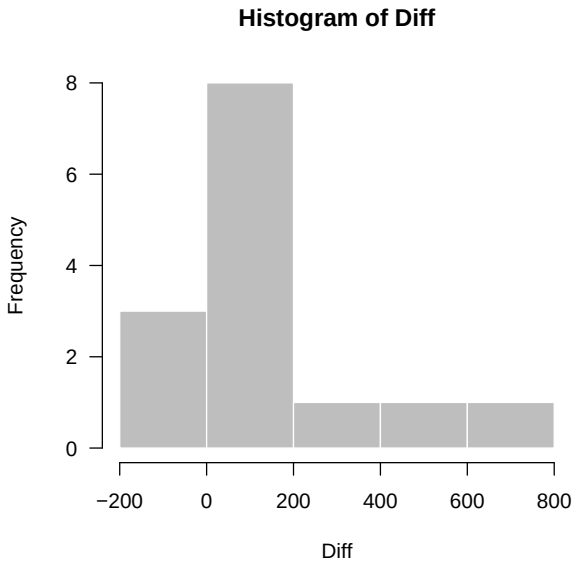
Additional exercises:

- Try re-running the above experiment(s) with different values for `N`, such as 100 and 1,000,000. What changes?
- Try re-running the above experiment(s) with different values for `n`, such as 5 and 1,000. What changes?

2 *t*-tests and confidence intervals

Carrying out one-sample t-tests and obtaining the corresponding confidence intervals is fairly straightforward in `R` using the function `t.test`. To illustrate with the cystic fibrosis study we discussed in class:

```
> cf <- read.delim("http://myweb.uiowa.edu/pbreheny/data/cystic-fibrosis.txt")
> Diff <- cf$Placebo - cf$Drug
> hist(Diff, col="gray", border="white", las=1)
```




```
> t.test(Diff)


One Sample t-test

data:  Diff
t = 2.2885, df = 13, p-value = 0.03949
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
   7.642646 265.357354
sample estimates:
mean of x
    136.5
```

As we saw in class, the p-value for testing the null hypothesis that amiloride has no effect on the average reduction in FVC is 0.04, and the confidence interval for the average difference is $[8, 265]$. As we see from the histogram, the distribution of differences doesn't look exactly normal, but with only 14 observations, it's difficult to tell.

3 Robustness

As we said in class, the derivation of the t distribution (and, therefore, its resulting confidence intervals) is based on an assumption of normality. How well do these intervals work when the data is not normally distributed? Or to put it a different way, how robust are these results to departures from normality? As we did earlier, let's investigate how well this works by resampling the NHANES triglyceride data. Recall that the underlying distribution here is somewhat skewed to the right. Does this cause problems for the t intervals?

```
> N <- 10000          ## Number of simulations to run
> n <- 25             ## Sample size
> covered <- numeric(N)
> for (i in 1:N) {
+   sam <- sample(pop, n)
+   covered[i] <- t.test(sam, mu=mean(pop))$p.value > 0.05
+ }
> mean(covered)

[1] 0.9338
```

The coverage is slightly under 95%, but the t intervals aren't doing too bad here. What if we lower the sample size to 15? To 5?

Now let's look at the variance. As we discussed in class, there is a nice pivotal quantity for the normal distribution that allows us to obtain confidence intervals for the variance. How about its coverage in this situation? Unfortunately there is no standard `R` function for constructing this interval, but it's pretty easy to code:

```
> N <- 10000          ## Number of simulations to run
> n <- 25             ## Sample size
> covered <- numeric(N)
> for (i in 1:N) {
+   sam <- sample(pop, n)
+   ci <- var(sam)*(n-1)/qchisq(c(0.975,0.025), n-1)
+   covered[i] <- (ci[1] < var(pop)) & (var(pop) < ci[2])
+ }
> mean(covered)

[1] 0.834
```

The coverage here is far worse than we had for the mean. As we said in class, the central limit theorem works for *any* distribution. For that reason, the t test is often a reasonable approximation even if the data departs somewhat from the normal distribution. There is no corresponding theorem for the variance: the pivotal CI works for the normal distribution and is not guaranteed to be accurate – even approximately – for other distributions.

This is a specific instance of an important general point in statistics: even though both of the above confidence interval methods are derived based on the assumption of normality, the two methods are not equally sensitive to that assumption. The t test is quite robust to departures from normality, while the

variance interval is very sensitive to such departures. This is one reason why simulations are so important and useful in statistics: to examine the performance of various methods when assumptions are not met (this is typically difficult to address analytically).