I hear, I forget. I do, I understand: a modified Moore-method mathematical statistics course

Nicholas J. Horton*

Department of Mathematics

Amherst College, Amherst, MA

September 24, 2013

*Address for correspondence: Dept of Mathematics, Seeley Mudd, Amherst College, Amherst, MA 01002-5000. Phone: 413-542-5655, email: nhorton@amherst.edu

Online Appendix: Additional Example Problems

I hear, I forget. I do, I understand:
a modified Moore-method mathematical statistics course

The following material is proposed as an online appendix.

5.4 Estimating σ using IQR

Assume that we observe n iid observations from a normal distribution. Questions:

i. Use the IQR of the list to estimate σ.

ii. Use simulation to assess the variability of this estimator for samples of $n = 100$ and 400.

iii. How does the variability of this estimator compare to $\hat{s}$ (usual estimator)?

```
numsim=1000; mu=42; n1=100; n2=400
runsim = function(numsim, n, mu, sigma) {
  res1 = numeric(numsim); res2 = res1
  for (i in 1:numsim) {
    mynorms = rnorm(n, 0, sigma)
    vals = quantile(mynorms)
    res1[i] = (vals[4]-vals[2])/1.34898
    res2[i] = sd(mynorms)
  }
  return(data.frame(IQR=res1, S=res2))
}
res100 = runsim(numsim, n1, mu, pi)
res400 = runsim(numsim, n2, mu, pi)
boxplot(res100$IQR, res100$S, res400$IQR, res400$S,
  names=c("n=100 (IQR)","n=100 (S)",
  "n=400 (IQR)", "n=400 (S)"),
  ylab="distribution of sigma-hat")
text(3.5, 4.0, "True sigma is 3.14159")
abline(h=pi); abline(v=2.5)
```

Figure 6: R code to carry out simulation study (estimation of σ)

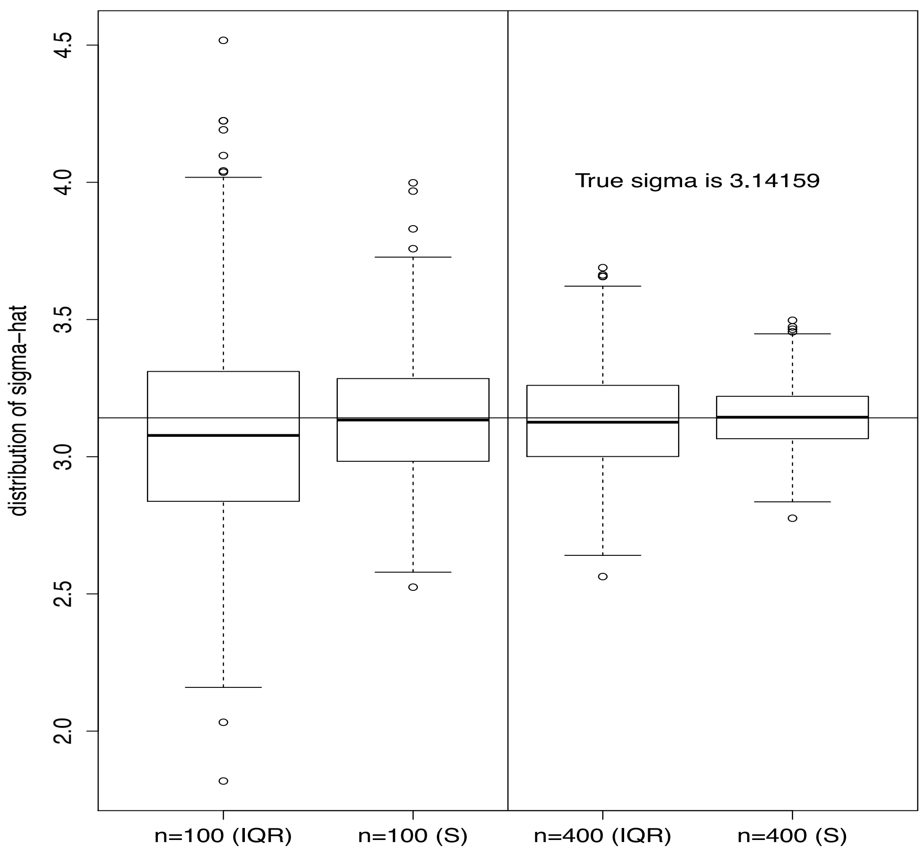

Figure 7: Distribution of sample estimates by estimator and sample size

5.4.1 Solution

i. We know that for a standard normal random variable $P(Z > 0.675) = 0.25$. So we would expect that the IQR (interquartile range) would extend to $2 * .6745 = 1.35$ standard units. We use this expectation to determine the estimator: $\tilde{s} = IQR/1.35$.

ii. We carried out a simple simulation study with a fixed mean and standard deviation (set to π). A thousand simulations of samples were taken using $\tilde{s}$ and $\hat{s}$ (MLE). The results are displayed in Figure 7. We note that both estimators are less variable when $n = 400$ than for $n = 100$ and conclude that the variability of the estimators goes down as a function of $\sqrt{n}$.

iii. The IQR for $\tilde{s}$ is 0.50 for n=100 and 0.25 for n=400, while the IQR for $\hat{s}$ is 0.30 for n=100 and 0.16 for n=400. We conclude that the MLE is more efficient than our ad-hoc estimator.

5.4.2 Commentary

This exercise was included with a problem set mid-way through the class as the nature and properties of estimators were explored. This problem introduced the idea of a simulation study to investigate the behavior of a new estimator. While the analytic solution was straightforward, it required the students to think about estimation in a different way, and tap properties of the normal distribution. The empirical solution provided a glimpse into the additional variability of the IQR estimator compared to the standard estimator of standard deviation. A full analytic solution for this problem was beyond the scope of the course, but can be undertaken for specific values of n.

5.5 Assessing robustness of chi-square statistic to small cell counts

Perform a simulation study on the sensitivity of the χ^2 test for the uniform distribution to expected cell counts below 5. Simulate the distribution of the test statistic for 16 and 64 observations from a uniform distribution using 8 equal-length bins (from Nolan & Speed (2000)).

5.5.1 Solution

We know that the chi-square test is recommended only in situations where the expected cell count is 5 or more in each cell. In this simulation study, we generate repeated samples from the null distribution and compare these to the large-sample distribution of the chi-square (χ^2) statistic (see Figure 8). We know that in this setting, the appropriate degrees of freedom are equal to the number of bins minus 1. The main work is done using the `simchisq()` function, which generates data from a continuous uniform variable, then constructs the observed and expected cell counts and the chi-square statistic. This is repeated for the two scenarios and displayed in Figure 9. We see that the observed distribution under

```
simchisq = function(n, bins) {
  vals = cut(runif(n, 0, bins), breaks=0:bins)
  obs = c(table(vals))
  exp = c(rep(n/bins, bins))
  return(sum(((obs - exp)^2)/exp))
}
library(mosaic); par(mfrow=c(1, 2))
bins = 8; n = 16    # Expected count per cell equal to 2
res = do(10000)*simchisq(n, bins)
plot(density(res$result), main="", lwd=2,
     xlab=paste("N=",n,", ",bins," bins", sep=""), xlim=c(0, 20))
curve(dchisq(x, bins-1), 0, max(res$result), add=TRUE, lwd=2, lty=2)

n = 64              # Expected count per cell equal to 8
res = do(10000)*simchisq(n, bins)
plot(density(res$result), main="", lwd=2,
  xlab=paste("N=",n,", ",bins," bins", sep=""), xlim=c(0, 20))
curve(dchisq(x, bins-1), 0, max(res$result), add=TRUE, lwd=2, lty=2)
```

Figure 8: R code to carry out simulation study (chi-square problem)

the null is somewhat jumpy (due to the discreteness of the possible values) when the expected cell counts are low (left figure), and that the observed curve is quite similar to the chi-square distribution when the expected cell count is 8 (right figure).

5.5.2 Commentary

This problem was intended to provide more practice in the construction of simulation studies as well as introduce new idioms related to looping and writing of functions. It also serves to highlight the importance of assumptions and the idea of sampling under the null distribution (as a precursor to resampling based inference). This was included with a group of problems mid-way through the class as the nature and properties of tests of hypotheses along with sampling distributions under the null were explored.

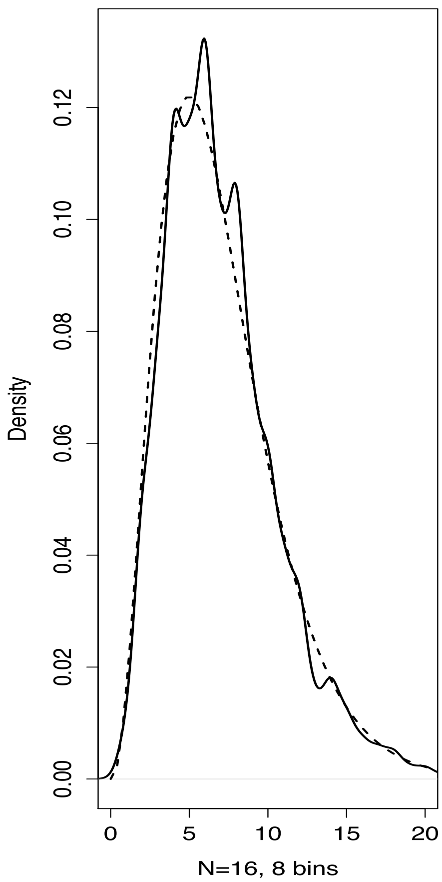

Figure 9: Observed and expected distribution for chi-square statistic