Stat 425 HW6 Solution

Fritz Scholz

1. Write a function

```
Estimate.compare = function(m = 10, n = 10, Nsim = 1000, Delta = 2,
                            dist = rnorm, parm = c(0,1)){...}
```

that generates independent samples $X_, \ldots, X_m \sim F(x)$ and $Y_1, \ldots, Y_n \sim F(x-\Delta)$ and computes the estimates $\bar{\Delta} = \bar{Y} - \bar{X}$ and $\hat{\Delta} = \text{med}_{i,j}(Y_j - X_i)$ and repeats this process `Nsim` times and collects the estimates in vectors `Delta.bar` and `Delta.hat`, respectively. Here the input argument `dist` indicates which F is to be used. This builds on what you learned from the previous assignment. Inside the function body of `Estimate.compare` you may want to use the function `Using.dist` given in the solution to HW5. Of course, you have to make sure that `Using.dist` exists inside your R work space.

The output of `Estimate.compare` should be the mean, median and variance of the two generated vectors of length `Nsim`, i.e., the estimated mean, median and variance of the respective estimator distributions. We know that the variance of the $\bar{\Delta}$ sampling distribution is

$$\text{var}(\bar{\Delta}) = \text{var}(\bar{Y} - \bar{X}) = \text{var}(\bar{Y}) + \text{var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \, \frac{m+n}{mn}$$

where σ^2 is the variance of $F(x)$ and thus also of $F(x-\Delta)$. We see that $1/\text{var}(\bar{\Delta})$ is proportional to $mn/(m+n)$. Illustrate this fact by writing a function

```
recip.var.plot = function(Nsim = 1000, mvec = c(5,10,20,30,50,100),
                          nvec = c(5,10,20,30,50,100), dist = rnorm,
                          parm = c(0,1), dist.name = "normal"){...}
```

that calls `Estimate.compare` for several different m, n, i.e., $m =$ `mvec[i]` and $n =$ `nvec[i]`, `i` $= 1, 2, \ldots$, and which plots the reciprocal of the estimated variances, i.e., the estimated values of $1/\text{var}(\bar{\Delta})$, against the respective values of $mn/(m+n)$. The point pattern should look roughly linear. Fit a line through this point pattern such that it goes through the origin (i.e., with intercept zero) using `out.ls1=lsfit(...,intercept=F)`. The component `out.ls1$coef` will give you the slope of that line. Does the slope of this linear pattern make sense? Once you have accomplished this, add the corresponding points for the estimates of $1/\text{var}(\hat{\Delta})$ in relation to $mn/(m+n)$ and fit a similar line to them. Does this suggest that $\text{var}(\hat{\Delta})$ is proportional to $(m+n)/mn$? Make sure that all 12 points show in the plot. Annotate your plot such that the sampled distribution is shown (that's is where you use the argument value for `dist.name`). In the annotation also show the ratio of slopes (slope of the $1/\text{var}(\hat{\Delta})$ pattern over the slope of the $1/\text{var}(\bar{\Delta})$ pattern) and also show those slopes separately. Further, add a legend explaining which points/lines belong to which type of estimate. Using an appropriate choice for `M` you may want to experiment with:

```
text(0,.9*M,substitute("slope for "~1/var(hat(Delta))==xx,
     list(xx=format(signif(out.ls2$coef,3)))),adj=0)
text(0,.85*M,substitute("slope for "~1/var(bar(Delta))==xx,
     list(xx=format(signif(out.ls1$coef,3)))),adj=0)
legend(0,.8*M,c(expression(1/var(bar(Delta))),
     expression(1/var(hat(Delta)))),lty=1:2,
     col=c("black","blue"),pch=c(1,16),bty="n")
```

We have shown above why to expect proportionality with respect to $mn/(m+n)$ in the case of $1/\text{var}(\bar{\Delta})$ and based on the simulation evidence we will accept the same type of proportionality for $1/\text{var}(\hat{\Delta})$.
If you were to match $\text{var}(\hat{\Delta}_{m.n}) = \text{var}(\bar{\Delta}_{m',n'})$ by proper choice of sample sizes $m = n$ and $m' = n' = \rho \times m$, what concept would ρ resonate with and how does it relate to the ratio of slopes that annotates your plot. We have seen two instances of this concept already.
By changing `dist`, `parm`, `dist.name` make such plots for `dist = rnorm`, `dist = runif`, `dist = rlogis`, and `dist = rexp`.

What is expected:

1. The code for `Estimate.compare`.
2. The code for `recip.var.plot`.
3. The 4 plots produced by `recip.var.plot` for the 4 indicated distributions.
4. Discussion of the resonance of ρ and its relation to the slope ratio.
5. Does the slope for $1/\text{var}(\bar{\Delta})$ in the normal distribution plot make sense? Explain. Try to extend that insight to the other distributions. Take as given: The variance of the logistic distribution with location 0 and scale 1 is $\pi^2/3$, the variance of the $U(0,1)$ distribution is $1/12$, and the variance of the exponential distribution with mean 1 is 1.

1. The code for `Estimate.compare`

```
Estimate.compare = function (m = 10, n = 10,
           Nsim = 100, Delta = 2, dist = rnorm,parm=c(0,1)){
Delta.bar=rep(0,Nsim); Delta.hat=rep(0,Nsim)
for(i in 1:Nsim){
  y=Using.dist(N=n,dist=dist,parm=parm)+Delta
  x=Using.dist(N=m,dist=dist,parm=parm)
  Delta.bar[i]=mean(y)-mean(x)
  Delta.hat[i]=median(outer(y,x,"-"))
}
```

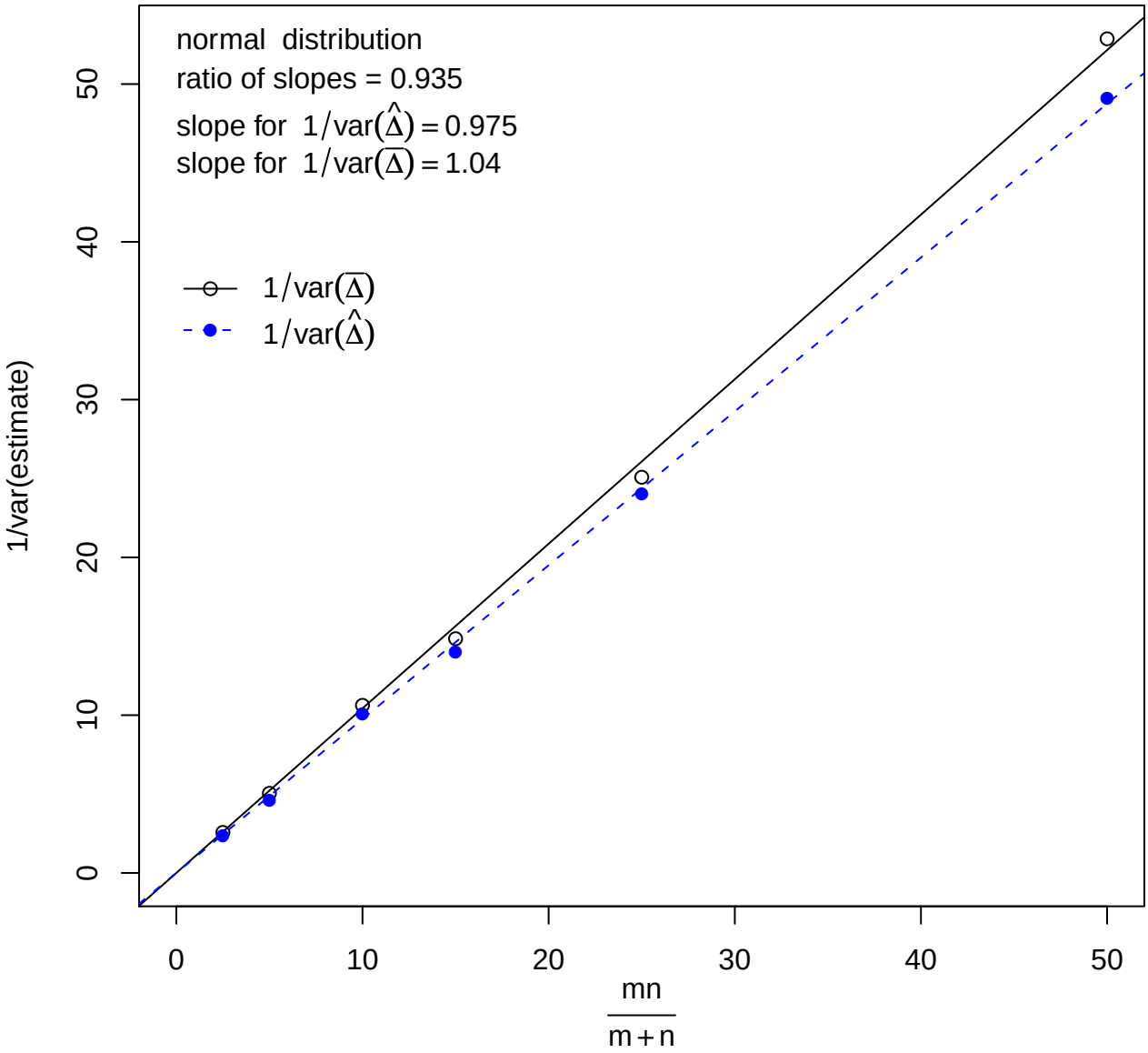
```
mean.hat=mean(Delta.hat)
med.hat=median(Delta.hat)
var.hat=var(Delta.hat)
mean.bar=mean(Delta.bar)
med.bar=median(Delta.bar)
var.bar=var(Delta.bar)
out=c(mean.hat,med.hat,var.hat,mean.bar,med.bar,var.bar)
names(out)=c("mean.hat","med.hat","var.hat",
              "mean.bar","med.bar","var.bar")
out}
```

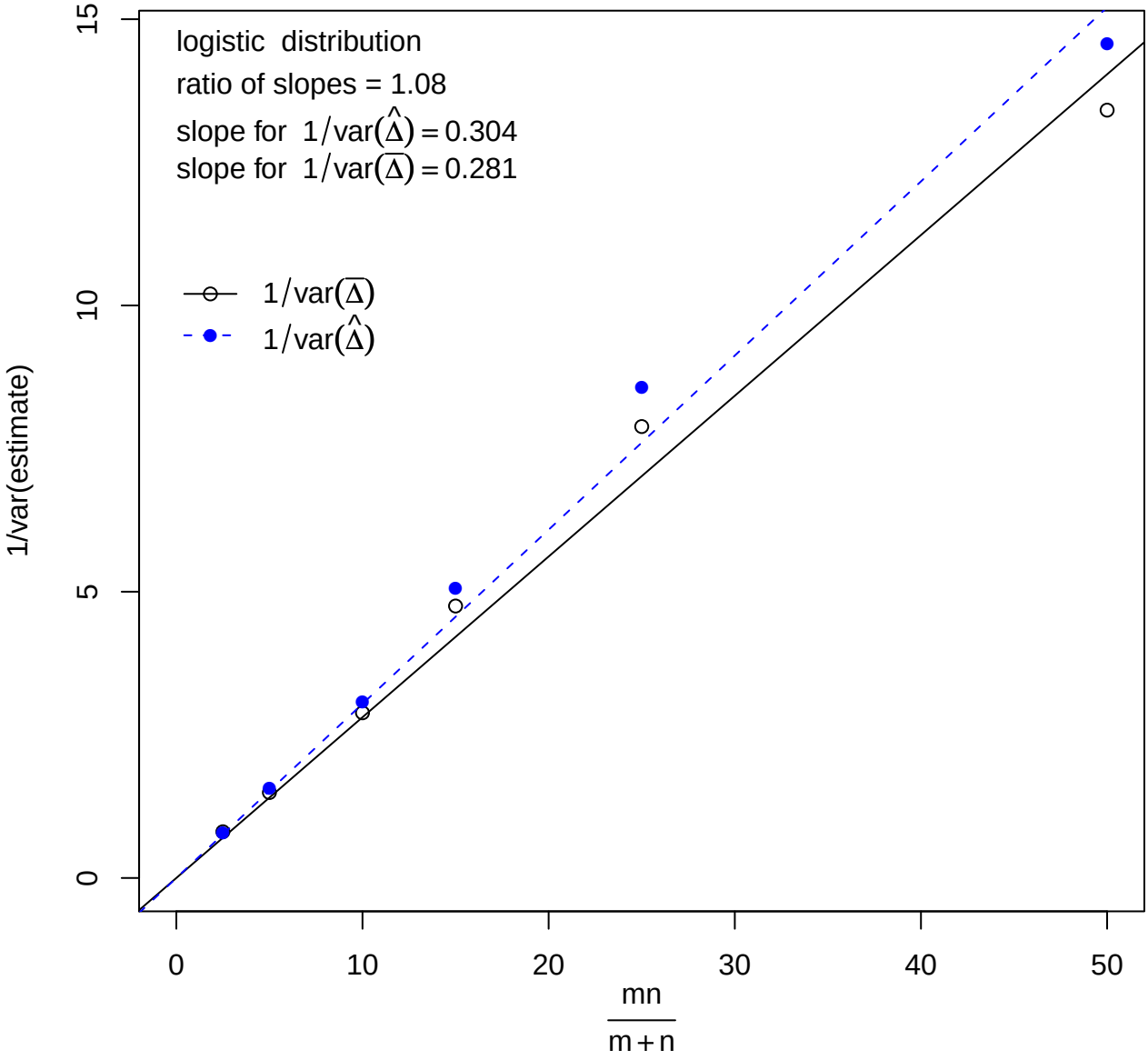
2. The code for `recip.var.plot`

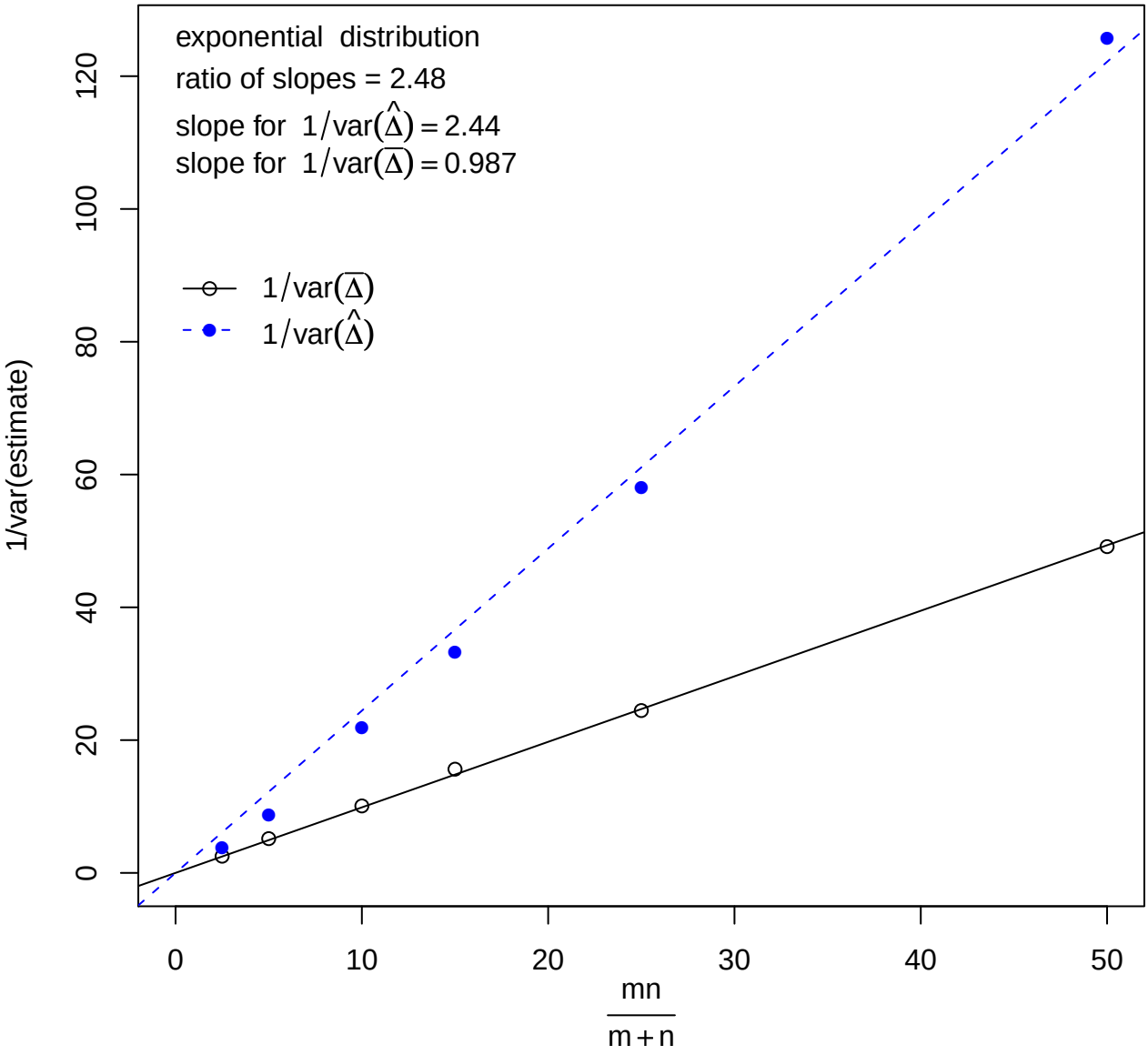
```
recip.var.plot=function(Nsim=1000,mvec=c(5,10,20,30,50,100),
       nvec=c(5,10,20,30,50,100),dist=rnorm,
       parm=c(0,1),dist.name="normal"){
k=length(mvec)
var.bar=rep(0,k)
var.hat=rep(0,k)
for(i in 1:k){
out=Estimate.compare(m=mvec[i],n=nvec[i],Nsim=Nsim,
      Delta=2,dist=dist,parm=parm)
var.bar[i]=out[6]
var.hat[i]=out[3]
}
M=max(c(1/var.bar,1/var.hat))
plot(mvec*nvec/(mvec+nvec),1/var.bar,ylim=c(0,M),
  xlim=c(0,max(mvec*nvec/(mvec+nvec))),xlab=expression(over(m * n,m+n)),
ylab="1/var(estimate)")
out.ls1=lsfit(mvec*nvec/(mvec+nvec),1/var.bar,intercept=F)
abline(0,out.ls1$coef)
points(mvec*nvec/(mvec+nvec),1/var.hat,pch=16,col="blue")
out.ls2=lsfit(mvec*nvec/(mvec+nvec),1/var.hat,intercept=F)
abline(0,out.ls2$coef,lty=2,col="blue")
text(0,M,paste(dist.name," distribution"),adj=0)
var.ratio=out.ls2$coef/out.ls1$coef
text(0,.95*M,paste("ratio of slopes =",
format(signif(var.ratio,3))),adj=0)
text(0,.9*M,substitute("slope for "~1/var(hat(Delta))==xx,
   list(xx=format(signif(out.ls2$coef,3)))),adj=0)
text(0,.85*M,substitute("slope for "~1/var(bar(Delta))==xx,
   list(xx=format(signif(out.ls1$coef,3)))),adj=0)
legend(0,.75*M,c(expression(1/var(bar(Delta))),
```

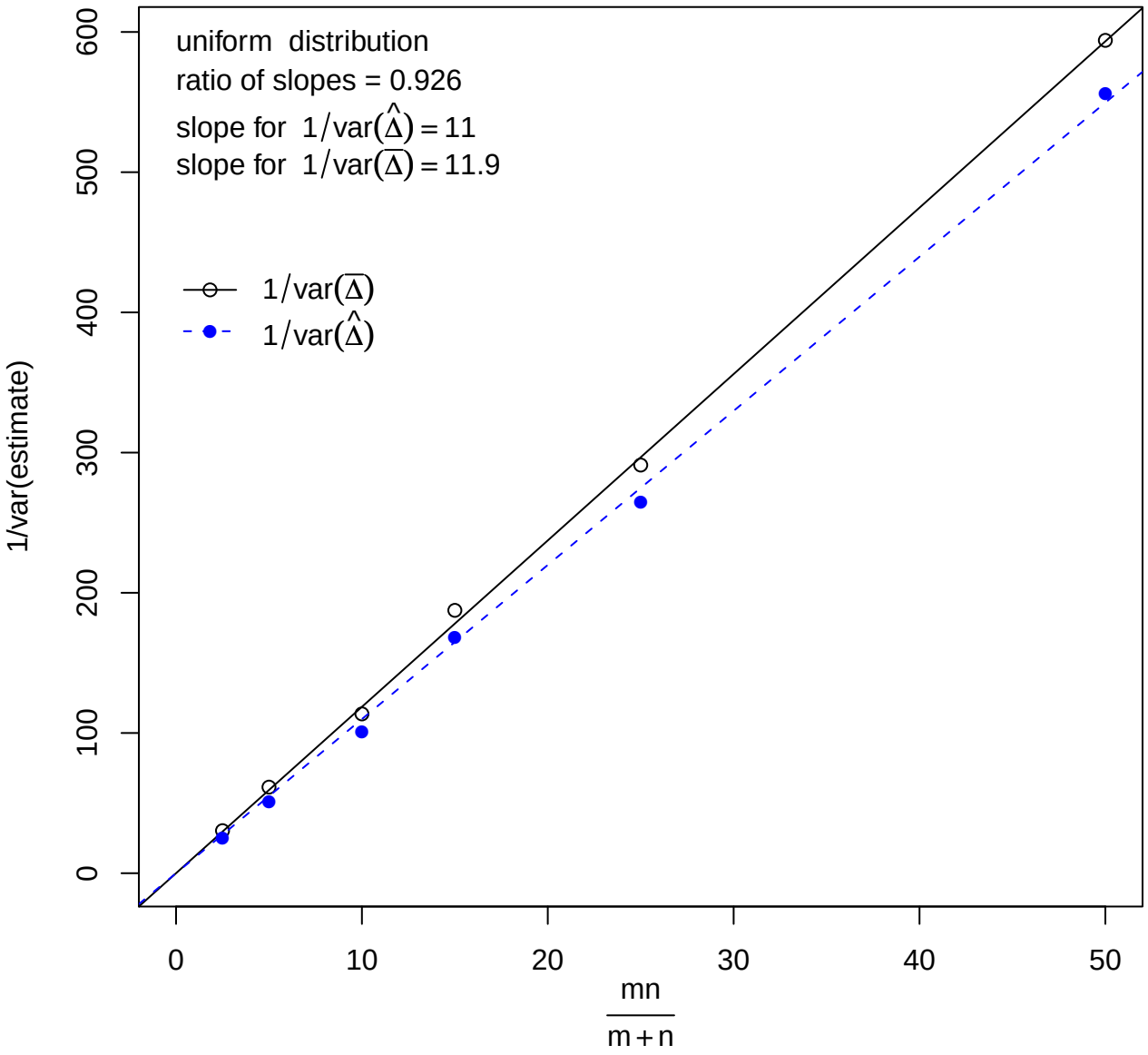
```
   expression(1/var(hat(Delta)))),lty=1:2,
col=c("black","blue"),pch=c(1,16),bty="n")}
```

3. The 4 plots:




logistic distribution
ratio of slopes = 1.08
slope for $1/\mathrm{var}(\hat{\Delta}) = 0.304$
slope for $1/\mathrm{var}(\overline{\Delta}) = 0.281$
$1/\mathrm{var}(\overline{\Delta})$
$1/\mathrm{var}(\hat{\Delta})$
1/var(estimate)
0
5
10
15
0
10
20
30
40
50
$\frac{mn}{m+n}$


exponential distribution
ratio of slopes = 2.48
slope for $1/\mathrm{var}(\hat{\Delta}) = 2.44$
slope for $1/\mathrm{var}(\overline{\Delta}) = 0.987$
$1/\mathrm{var}(\overline{\Delta})$
$1/\mathrm{var}(\hat{\Delta})$
1/var(estimate)
0
20
40
60
80
100
120
0
10
20
30
40
50
$\frac{mn}{m+n}$



4. Discussion of resonance. Matching variances of the two estimators with different sample sizes $m = n$ and $m' = n' = \rho m$ has very much the flavor of efficiency where we matched the powers for the Wilcoxon test with that of the t-test, or we matched the dispersion probabilities of the corresponding estimators $\hat{\Delta}$ and $\bar{\Delta}$ by appropriate choice of sample sizes.
This suggests that we take $m'/m = \rho$ as a measure of efficiency of $\hat{\Delta}$ relative to $\bar{\Delta}$ with respect to estimator variances. If $\rho = .5$ then $\bar{\Delta}$ would require only half as many observations as $\hat{\Delta}$ would require to have the same variance.

Assuming the following proportionalities

$$\frac{1}{\text{var}(\bar{\Delta})} = b \times \frac{m'n'}{m'+n'} = b \times \frac{m'}{2} = b \times \frac{\rho m}{2}$$

$$\frac{1}{\text{var}(\hat{\Delta})} = a \times \frac{mn}{m+n} = a \times \frac{m}{2}$$

and having chosen m' and m such that the variances are matched we would get

$$a \times \frac{m}{2} = b \times \frac{m'}{2} \qquad \text{or} \qquad \frac{m'}{m} = \rho = \frac{a}{b}$$

the ratio of the calculated slopes. The ratios in the plots are approximately the same as the efficiencies given in the corresponding comparison of the Wilcoxon test relative to the t-test. The only exception is in the case of the exponential distribution where we had an ARE of 3 and the ratio in the plot only gives 2.48. It may be that the large sample behavior takes some time in that case. While the linear pattern is quite strong for $1/\text{var}(\bar{\Delta})$ (as it should be by theory, previously explained) the pattern for $1/\text{var}(\hat{\Delta})$ shows strong initial curvature which may starighten out only for large $m = n$.