1 Evaluating estimators

Suppose you observe data $X_1, ..., X_n$ that are iid observations with distribution F_θ indexed by some parameter θ. When trying to estimate θ, one may be interested in determining the properties of some estimator $\hat{\theta}$ of θ. In particular, the *bias*

$$\text{Bias}(\hat{\theta}) = E\left(\hat{\theta} - \theta\right)$$

may be of interest. That is, the average difference between the estimator and the truth. Estimators with $\text{Bias}(\hat{\theta}) = 0$ are called *unbiased*.

Another (possibly more important) property of an estimator is how close it tends to be to the truth on average. The most common choice for evaluating estimator precision is the *mean squared error*,

$$\text{MSE}(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right).$$

When comparing a number of estimators, MSE is commonly used as a measure of quality. By directly using the identity that $\text{var}(Y) = E(Y^2) - E(Y)^2$, where the random variable $Y = \hat{\theta} - \theta$, the above equation becomes

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= E\left(\hat{\theta} - \theta\right)^2 + \text{var}(\hat{\theta} - \theta)\\ &= \text{Bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})\end{aligned}$$

where the last line follows from the definition of bias and the fact that $\text{var}(\hat{\theta}-\theta) = \text{var}(\hat{\theta})$, since θ is a constant. For example, if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$, then $\overline{X} \sim N(\mu, \sigma^2/n)$. So the bias of $\overline{X}$ as an estimator of μ is

$$\text{Bias}(\overline{X}) = E(\overline{X} - \mu) = \mu - \mu = 0$$

and the MSE is

$$\text{MSE}(\overline{X}) = 0^2 + \text{var}(\overline{X}) = \sigma^2/n$$

The above identity says that the precision of an estimator is a combination of the bias of that estimator and the variance. Therefore **it is possible for a biased estimator to be more precise than an unbiased estimator** if it is significantly less variable. This is known as the *bias-variance tradeoff*. We will see an example of this.

1.2 Using monte carlo to explore properties of estimators

In some cases it can be difficult to explicitly calculate the MSE for an estimator. When this happens monte carlo can be a useful alternative to a very cumbersome mathematical calculation. The example below is an instance of this.

Example: Suppose $X_1, ..., X_n$ are iid $N(\theta, \theta^2)$ and we are interested in estimation of θ. Two reasonable estimators of θ are the sample mean $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n X_i$ and the sample standard deviation $\hat{\theta}_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$. To compare these two estimators by monte carlo for a specific n and θ:

1. Generate $X_1, ..., X_n \sim N(\theta, \theta^2)$
2. Calculate $\hat{\theta}_1$ and $\hat{\theta}_2$
3. Save $(\hat{\theta}_1 - \theta)^2$ and $(\hat{\theta}_2 - \theta)^2$
4. Repeat step 1-3 k times
5. Then the means of the $(\hat{\theta}_1 - \theta)^2$'s and $(\hat{\theta}_2 - \theta)^2$'s, over the k replicates, are the monte carlo estimators of the MSEs of $\hat{\theta}_1$ and $\hat{\theta}_2$.

This basic approach can be used any time you are comparing estimators by monte carlo. The larger we choose k to be, the more accurate these estimates are. We implement this in R with the following code for $\theta = .5, .6, .7, ..., 10$, $n = 50$, and $k = 1000$.

```
k = 1000
n = 50

# Sequence of values of theta
THETA <- seq(.5, 10, by=.1)

# Storage for the MSEs of each estimator
MSE <- matrix(0, length(THETA), 2)

# Loop through the values in Theta
for(j in 1:length(THETA))
{

   # Generate the k datasets of size n
   D <- matrix(rnorm(k*n, mean=THETA[j], sd=THETA[j]), k, n)

   # Calculate theta_hat1 (sample mean) for each data set
   ThetaHat_1 <- apply(D, 1, mean)

   # Calculate theta_hat2 (sample sd) for each data set
   ThetaHat_2 <- apply(D, 1, sd)

   # Save the MSEs
   MSE[j,1] <- mean( (ThetaHat_1 - THETA[j])^2 )
   MSE[j,2] <- mean( (ThetaHat_2 - THETA[j])^2 )

}

# Plot the results on the same axes
plot(THETA, MSE[,1], xlab=quote(theta), ylab="MSE",
main=expression(paste("MSE for each value of ", theta)),
type="l", col=2, cex.lab=1.3, cex.main=1.5)
```

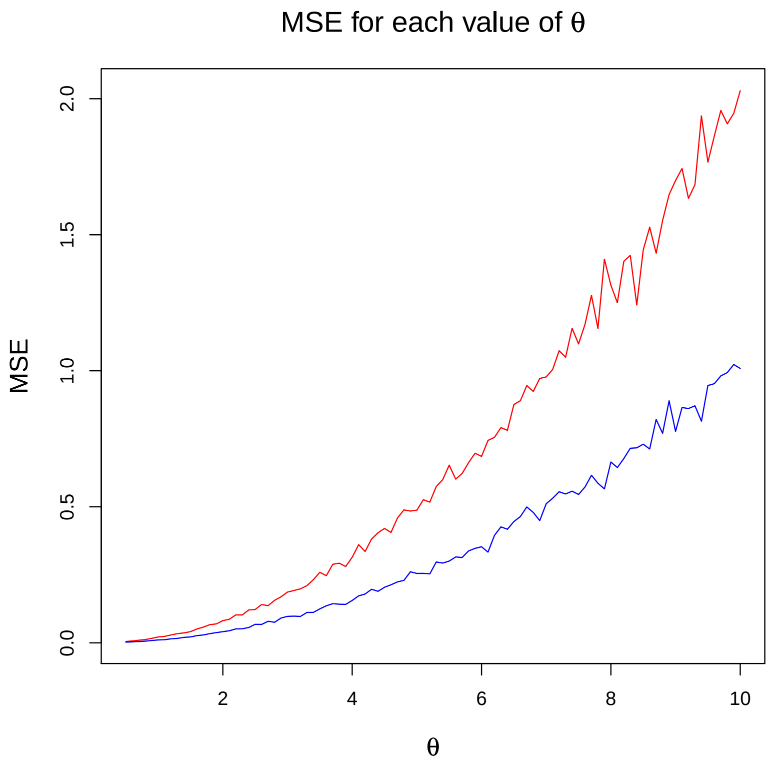

Figure 1: Simulated values for the MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$

```
lines(THETA, MSE[,2], col=4)
```

From the plot we can see that $\hat{\theta}_2$, the sample standard deviation, is a uniformly better estimator of θ than $\hat{\theta}_1$, the sample mean. We can verify this simulation mathematically. Clearly the sample mean's MSE is

$$\text{MSE}(\hat{\theta}_1) = \theta^2/n$$

The MSE for sample standard deviation is somewhat more difficult. It is well known that, in general, the sample variance from a normal population, V, is distributed so that

$$\frac{(n-1)V}{\sigma^2} \sim \chi^2_{n-1},$$

where σ^2 is the true variance. In this case $\hat{\theta}_2 = \sqrt{V}$. The χ^2 distribution with k degrees of freedom has density function

$$p(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

where Γ is the gamma function. Using this we can derive the expected value of $\sqrt{V}$:

$$
\begin{aligned}
E\left(\sqrt{V}\right) &= \sqrt{\frac{\sigma^2}{n-1}} E\left(\sqrt{\frac{(n-1)V}{\sigma^2}}\right) \\
&= \sqrt{\frac{\sigma^2}{n-1}} \int_0^\infty \sqrt{x} \frac{(1/2)^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})} x^{((n-1)/2)-1} e^{-x/2} dx
\end{aligned}
$$

which follows from the definition of expectation and the expression above for the χ^2 density. The trick now is to rearrange terms and factor out constants properly so that the integrand become another χ^2 density

$$
\begin{aligned}
E\left(\sqrt{V}\right) &= \sqrt{\frac{\sigma^2}{n-1}} \int_0^\infty \frac{(1/2)^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})} x^{(n/2)-1} e^{-x/2} dx \\
&= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \int_0^\infty \frac{(1/2)^{\frac{n-1}{2}}}{\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} dx \\
&= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \cdot \frac{(1/2)^{\frac{n-1}{2}}}{(1/2)^{n/2}} \underbrace{\int_0^\infty \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}\, dx}_{\chi^2_n \text{ density}}
\end{aligned}
$$

Now we know that the integral in the last line is 1, since it has the form of a χ^2 density with n degrees of freedom. The rest is just simplifying constants:

$$
\begin{aligned}
E\left(\sqrt{V}\right) &= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \cdot \frac{(1/2)^{\frac{n-1}{2}}}{(1/2)^{n/2}} \\
&= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \cdot \sqrt{2} \\
&= \frac{\sqrt{2} \cdot \Gamma(n/2)}{\sqrt{n-1} \cdot \Gamma(\frac{n-1}{2})} \sigma \\
&= \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \sigma
\end{aligned}
$$

Therefore $E(\hat{\theta}_2) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \theta$. So the bias is

$$
\text{Bias}(\hat{\theta}_2) = \theta - E(\hat{\theta}_2) = \theta \left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})}\right)
$$

To calculate the variance of $\hat{\theta}_2$ we also need $E(\hat{\theta}_2^2)$. $\hat{\theta}_2^2$ is the sample variance, which we know is an unbiased estimator of the variance, θ^2, so

$$
E(\hat{\theta}_2^2) = \theta^2
$$

so the variance of $\hat{\theta}_2$ is

$$\mathrm{var}(\hat{\theta}_2) = \theta^2 \left(1 - \frac{2}{n-1} \cdot \frac{\Gamma(n/2)^2}{\Gamma(\frac{n-1}{2})^2} \right)$$

Finally,

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_2) &= \theta^2 \left(\left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \right)^2 + \left(1 - \frac{2}{n-1} \cdot \frac{\Gamma(n/2)^2}{\Gamma(\frac{n-1}{2})^2} \right) \right) \\ &= 2\theta^2 \left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \right) \end{aligned}$$

It is a fact that

$$\left(1 - \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \right) < 1/2n$$

for any $n \geq 2$. This implies that $\mathrm{MSE}(\hat{\theta}_2) < \mathrm{MSE}(\hat{\theta}_1)$ for any n and any θ. We can check this derivation by plotting the MSEs and comparing with the simulation based MSEs:

```
# for each Q[1] is Theta, and Q[2] is n
# MSE of theta_hat1
MSE1 <- function(Q) (Q[1]^2)/Q[2]

# MSE theta_hat2
MSE2 <- function(Q)
{
   theta <- Q[1]; n <- Q[2];

   G <- gamma(n/2)/gamma( (n-1)/2 )
   bias <- theta * (1 - sqrt(2/(n-1)) * G )
   variance <- (theta^2) * (1 - (2/(n-1)) * G^2 )

    return(bias^2 + variance)
}

# Grid of values for Theta for n=50
THETA <- cbind(matrix( seq(.5, 10, length=100), 100, 1 ), rep(50,100))

# Storage for MSE of thetahat1 (column 1) and thetahat2 (column 2)
MSE <- matrix(0, 100, 2)

# MSE of theta_hat1 for each theta
MSE[,1] <- apply(THETA, 1, MSE1)
```

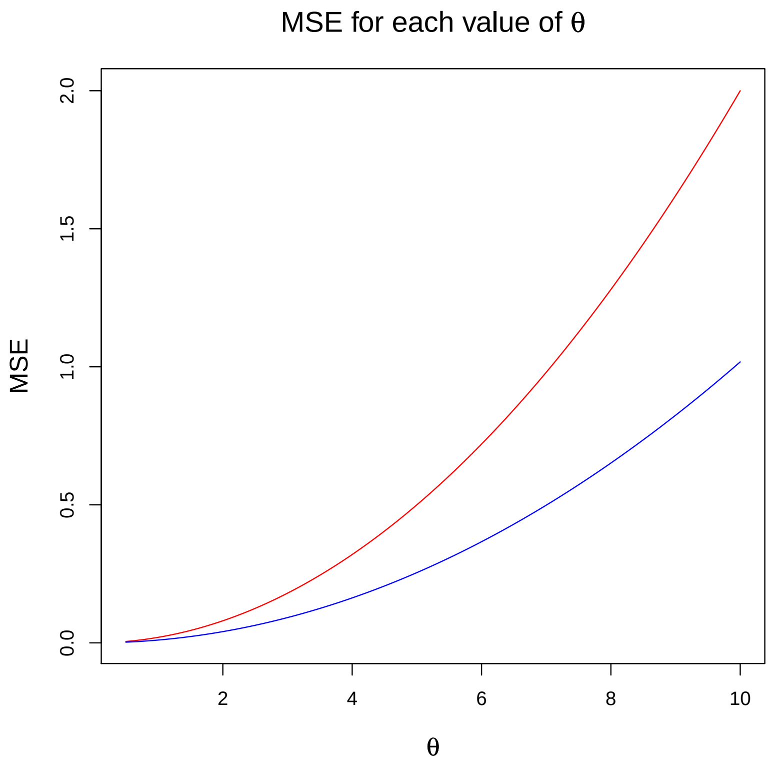

Figure 2: True values for the MSE of $\hat{\theta}_1$ and $\hat{\theta}_2$

```
# MSE of theta_hat2 for each theta
MSE[,2] <- apply(THETA, 1, MSE2)

plot(THETA[,1], MSE[,1], xlab=quote(theta), ylab="MSE",
main=expression(paste("MSE for each value of ", theta)),
type="l", col=2, cex.lab=1.3, cex.main=1.5)
lines(THETA[,1], MSE[,2], col=4)
```

Clearly the conclusion is the same as the simulated case– $\hat{\theta}_2$ has a lower MSE than $\hat{\theta}_1$ for any value of θ, but it was far less complicated to show this by simulation.

Exercise 1: Consider data $X_1, ..., X_n$ iid $N(\mu, \sigma^2)$ where we are interested in estimating σ^2 and μ is unknown. Two possible estimators are:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and the conventional unbiased sample variance:

$$\hat{\theta}_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Estimate the MSE for each of these estimators when $n = 15$ for $\sigma^2 = .5, .6, ..., 3$ and evaluate which estimate is closer to the truth on average for each value of σ^2.

2 Properties of hypothesis tests

Consider deciding between two competing statistical hypotheses H_0, the *null hypothesis*, and H_1, the *alternative hypothesis* based on data $X_1, ..., X_n$. A *test statistic* is a function of the data $T = T(X_1, ..., X_n)$ such that if $T \in R_\alpha$ then you reject H_0, otherwise you do not. The space R_α is called the *rejection region* and is chosen so that

$$P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(T \in R_\alpha | H_0 \text{ is true}) = \alpha$$

α is referred to as the *level* of the test, and is the probability of incorrectly rejecting H_0; α is typically chosen by the user; .05 is a common choice. For example, in a two-sided z-test of $H_0 : \mu = 0$, when σ^2 is known, the rejection region is

$$R_\alpha = (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$$

where z_a denotes the a'th quantile of a standard normal distribution. When $\alpha = .05$ this yields the familiar rejection region $(-\infty, -1.96) \cup (1.96, \infty)$.

A good hypothesis test is one that has for a small value of α, has a large *Power*, which is the probability of rejecting H_0 when H_0 is indeed false. When testing the hypothesis $H_0 : \theta = \theta_0$ for some specific null value θ_0, and $\theta_{\text{true}} \neq \theta_0$, the power is

$$\text{Power}(\theta_{\text{true}}) = P(T \in R_\alpha | \theta = \theta_{\text{true}})$$

Some primary determinants of the power of a test are:

- The sample size
- The difference between the null value and the true value (generally referred to as *effect size*
- The variance in the observed data

In many settings practioners are interested in either **a)** how far the true value of θ must be from θ_0 or **b)** for a fixed effect size, how large the sample size must be for the power to reach some nominal level, say 80%. Inquiries of this type are referred to as *power analysis*

Example 2: Power of the two-sample z-test

Suppose you observe $X_1, ..., X_n$ iid $N(\mu_X, \sigma^2)$ and $Y_1, ..., Y_m$ iid $N(\mu_Y, \sigma^2)$ where μ_X, μ_Y are unknown, and σ^2 **is known**. We are interested in a two-sided test of the hypothesis $H_0 : \mu_X - \mu_Y = 0$. A common statistic for testing such hypotheses is the z-statistic:

$$T = \frac{\sqrt{n}\left(\overline{X} - \overline{Y}\right)}{\sqrt{2\sigma^2}}$$

It is well known that, under H_0, T has a standard normal distribution. It can be shown that, for any value of $\mu_D = |\mu_X - \mu_Y|$, this test is the **most powerful level α-level test of H_0**. (Similarly, when the variances are unknown and the sample size/variances are potentially unequal, the students t-test is the most powerful α level tests of this null hypothesis).

μ_D is the measure of effect size in this test, and Power(μ_D) is a monotonically increasing function. For example, if μ_D is very small it intuitively that we would be less likely to reject H_0 than if μ_D was large.

We will investigate the power of the two-sample z-test for sample sizes $n = 10, 20, 30, 40, 50$ as a function of the true mean difference μ_D. The larger the true σ^2 the smaller the power will be (for a fixed n and μ_D), but we will not investigate this effect in this example. Each data set will be generated to have $\sigma^2 = 1$, the two samples will have equal sizes, and $\alpha = .05$. The basic algorithm is:

1. Generate datasets of the form $X_1, ..., X_n \sim N(0, 1)$, and $Y_1, ..., Y_n \sim N(\mu_D, 1)$.
2. Calculate T
3. Save $I = \mathcal{I}(|T| > z_{1-\alpha/2})$
4. Repeat k times
5. The mean of the k values of the I's is the monte carlo estimate of Power(μ_D).

```
#alpha level
alpha = .05

# number of simulation reps
k <- 1000

# sample sizes
n <- 10*c(1:5)

# the mu_D's
mu_D <- seq(0, 2, by=.1)

# storage for the estimated Powers
Power <- matrix(0, length(mu_D), 5)

for(i in 1:5)
{

   for(j in 1:length(mu_D))
   {

      # Generate k datasets of size n[i]
      X <- matrix( rnorm(n[i]*k), k, n[i])
      Y <- matrix( rnorm(n[i]*k, mean=mu_D[j]), k, n[i])

      # Get sample means for each of the k datasets
      Xmeans <- apply(X, 1, mean)
      Ymeans <- apply(Y, 1, mean)
```

```
        # Calculate the Z statistics
        T <- sqrt(n[i])*(Xmeans - Ymeans)/sqrt(2)

        # Indicators of the z-statistics being
        # in the rejectin region
        I <- (abs(T) > qnorm(1-(alpha/2)))

        # Save the estimated power
        Power[j,i] <- mean(I)

    }

}

plot(mu_D, Power[,1], xlab=quote(mu(D)), ylab=expression(
paste("Power(", mu(D), ")")), col=2, cex.lab=1.3,
cex.main=1.5, main=expression(paste("Power(", mu(D), ") vs.", mu(D))),
type="l" )
points(mu_D, Power[,1], col=2); points(mu_D, Power[,2], col=3)
points(mu_D, Power[,3], col=4); points(mu_D, Power[,4], col=5)
points(mu_D, Power[,5], col=6); lines(mu_D, Power[,2], col=3)
lines(mu_D, Power[,3], col=4); lines(mu_D, Power[,4], col=5)
lines(mu_D, Power[,5], col=6); abline(h=alpha)
legend(1.5, .3, c("n = 10", "n = 20", "n = 30", "n = 40", "n = 50"), pch=(1),
col=c(2:6), lty=1)
```

It is actually straightforward to calculate the power of the two-sample z-test. If μ_D is the true mean difference, then T is a standard normal random variable but shifted over by $\sqrt{n}\mu_D/\sqrt{2}$, since $E(\overline{X} - \overline{Y}) = \mu_D$. Letting Z denote a standard normal random variable, the power as a function of μ_D is:

$$
\begin{aligned}
\text{Power}(\mu_D) &= P\left(|T| > z_{1-\alpha/2}\right) \\
&= 1 - P\left(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}\right) \\
&= 1 - P\left(z_{\alpha/2} \leq Z + \sqrt{n}\mu_D/\sqrt{2} \leq z_{1-\alpha/2}\right) \\
&= 1 - P\left(z_{\alpha/2} - \sqrt{n}\mu_D/\sqrt{2} \leq Z \leq z_{1-\alpha/2} - \sqrt{n}\mu_D/\sqrt{2}\right) \\
&= 1 - \left(P\left(Z \leq z_{1-\alpha/2} - \sqrt{n}\mu_D/\sqrt{2}\right) - P\left(Z \leq z_{\alpha/2} - \sqrt{n}\mu_D/\sqrt{2}\right)\right) \\
&= 1 - \left(\Phi(z_{1-\alpha/2} - \sqrt{n}\mu_D/\sqrt{2}) - \Phi(z_{\alpha/2} - \sqrt{n}\mu_D/\sqrt{2})\right)
\end{aligned}
$$

where Φ denotes the standard normal CDF. Notice that as $n \to \infty$,

$$\lim_{n\to\infty} \Phi(z_{1-\alpha/2} - \sqrt{n}\mu_D/\sqrt{2}) = \Phi(-\infty) = 0$$

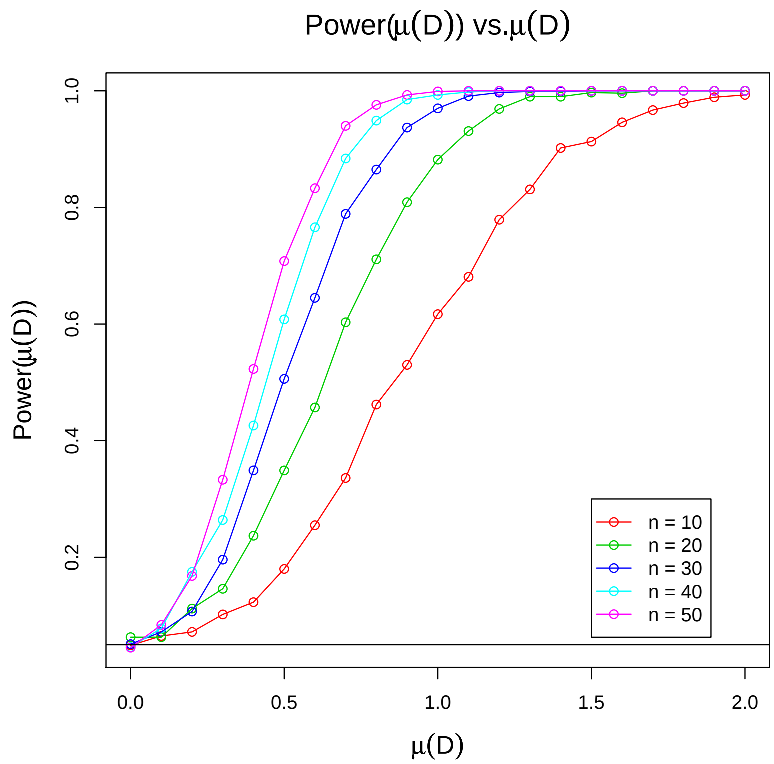

Figure 3: Simulated power of the two-sample z-test for sample sizes $n = 10, 20, 30, 40, 50$ and μ_D ranging from 0 up to 2.

and similarly for $\Phi(z_{\alpha/2} - \sqrt{n}\mu_D/\sqrt{2})$, therefore

$$\lim_{n\to\infty} \text{Power}(\mu_D) = 1$$

In other words, not matter how small $\mu_D > 0$ is, the power to detect it as significantly different from 0 goes to 1 as the sample size increases. To check this calculation we plot the theoretical power and compare it with the simulation:

```
# alpha level
alpha <- .05

# sample sizes
n <- 10*c(1:5)

# the mu_D's
mu_D <- seq(0, 2, by=.1)

# storage for the true Powers
Power <- matrix(0, length(mu_D), 5)

for(i in 1:5)
{
   for(j in 1:length(mu_D))
   {
```

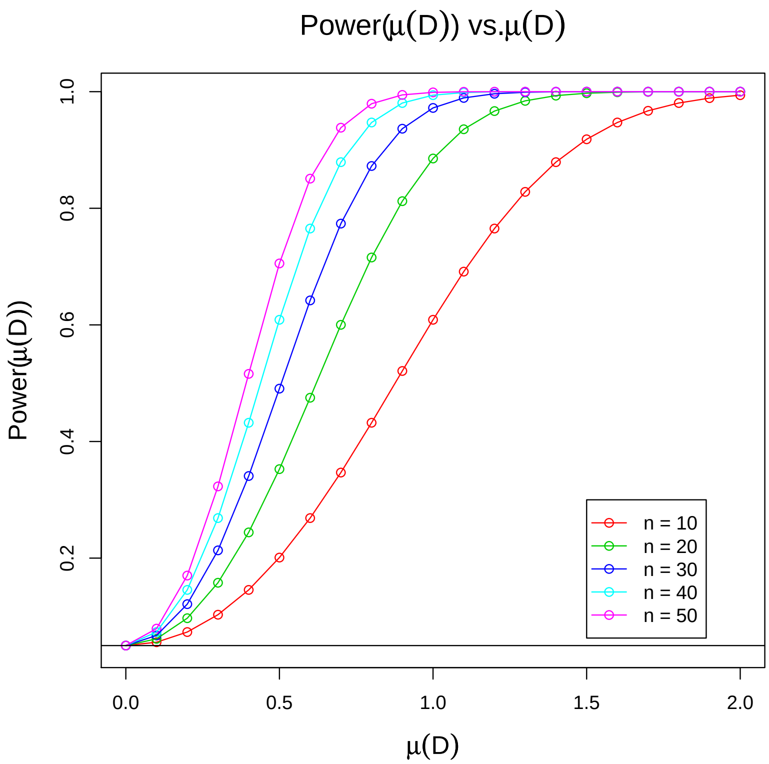

Figure 4: Theoretical power of the two-sample z-test for sample sizes $n = 10, 20, 30, 40, 50$ and μ_D ranging from 0 up to 2.

```
      Power[j,i] <- 1 - ( pnorm( qnorm(1-alpha/2) - sqrt(n[i])*mu_D[j]/sqrt(2) ) -
       pnorm( qnorm(alpha/2) - sqrt(n[i])*mu_D[j]/sqrt(2) ) )
   }
}

# plot the results
plot(mu_D, Power[,1], xlab=quote(mu(D)), ylab=expression(
paste("Power(", mu(D), ")")), col=2, cex.lab=1.3,
cex.main=1.5, main=expression(paste("Power(", mu(D), ") vs.", mu(D))),
type="l" )
points(mu_D, Power[,1], col=2); points(mu_D, Power[,2], col=3)
points(mu_D, Power[,3], col=4); points(mu_D, Power[,4], col=5)
points(mu_D, Power[,5], col=6); lines(mu_D, Power[,2], col=3)
lines(mu_D, Power[,3], col=4); lines(mu_D, Power[,4], col=5)
lines(mu_D, Power[,5], col=6); abline(h=alpha)
legend(1.5, .3, c("n = 10", "n = 20", "n = 30", "n = 40", "n = 50"),
pch=(1), col=c(2:6), lty=1)
```

We can see the theoretical calculation matches the simulation. In this case the power calculation is simple, but **for most hypothesis tests, power calculations are intractable, so simulation based power analysis is the only option.**

Exercise 2: Using a similar approach to the above, consider the same problem except $X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$ (both equal sample size) where σ_X^2, σ_Y^2 are not known and but are **assumed to be equal**. Use the statistic

$$T = \frac{\sqrt{n}\left(\overline{X} - \overline{Y}\right)}{\sqrt{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2}}$$

where $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ are the unbiased sample variances from exercise 1 calculated for each set of data. Under H_0, T has a t-distribution with $2n - 2$ degrees of freedom. Estimate the power of this test for sample sizes $n = 10, 20, 30, 40, 50$ and for the true $\mu_X - \mu_Y$ ranging from 0 up to 2.

In this case the theoretical power calculation, although possible, is significantly more difficult.