Sample Answer Lab01: Data Transformations, Bivariate Regression Analysis, Numerical Integration & Distributions

Handed out: Monday, February 1, 2021

Task 1. Univariate Variable Exploration and Transformations [2 points]

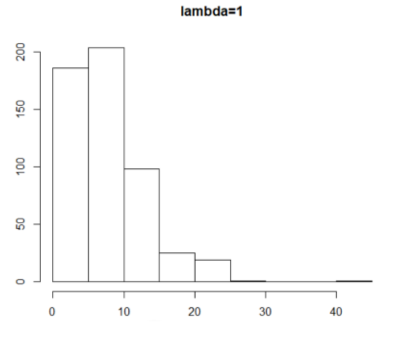
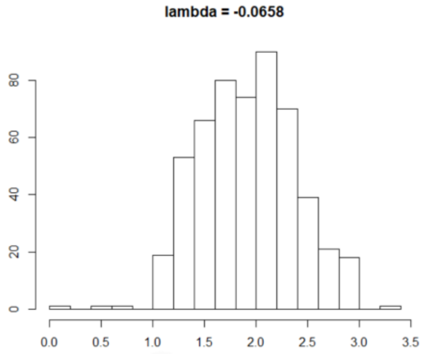
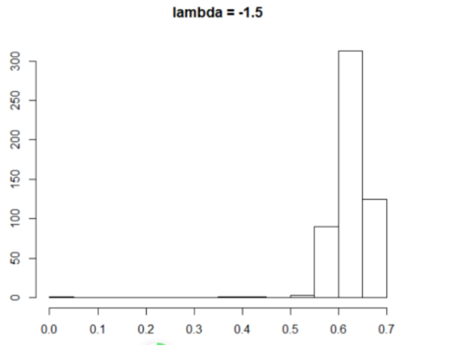
Use the **CPS1985** dataset[1] in the library **AER** to explore the distribution of the respondents' **wage**.

Task 1.1: Find the best λ^{best} -value for the Box-Cox transformation (see **summary(car::powerTransform(*varName*))**). Could the *log*-transformation (i.e., $\lambda = 0.0$) instead of λ^{best} be used? Justify your answer. [0.5 points]

```
library(AER);library(car);library(e1071)
data(CPS1985)
summary(powerTransform(CPS1985$wage))
bcPower Transformation to Normality
               Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
CPS1985$wage   -0.0658           0         -0.1997        0.068
Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                            LRT df pval
LR test, lambda = (0) 0.9245 1 0.33628
Likelihood ratio test that no transformation is needed
                            LRT   df    pval
LR test, lambda = (1) 232.5055 1 < 2.22e-16
```

Comment: The estimated transformation power (λ= −0.0658), which is very close to 0. The p-value of the likelihood ratio test (H_0: $\lambda = 0$) is larger than 0.05, so we fail to reject the null hypothesis. Instead of using (λ= −0.0658), the log-transformation should be used in this case.

Task 1.2: For the untransformed ($\lambda = 1$), optimal ($\lambda = \lambda^{best}$) and over-adjusted ($\lambda = -1.0$) Box-Cox transformed **wage** variable repeat the following tasks and ***comparatively interpret*** the results:

	$\lambda = 1$	$\lambda^{best} = -0.0658$	$\lambda = -1$
Histograms			

[1] See pages 1-3 in Kleiber & Zeileis, 2008. Applied Econometrics with R. Springer Verlag. Available as *e-book* in UTD's library. To import the data-frame use the statement **data(CPS1985, package="AER")** .

Skewness	1.687762	0.001027028	-2.79321396

Shapiro-test	`> shapiro.test(wage)`
	`Shapiro-Wilk normality test`
	`data: wage`
	`W = 0.8673, p-value < 2.2e-16`
		`> shapiro.test(wage.bc)`
		`Shapiro-Wilk normality test`
		`data: wage.bc`
		`W = 0.98923, p-value = 0.000586`
			`> shapiro.test(wage.bc.over)`
			`Shapiro-Wilk normality test`
			`data: wage.bc.over`
			`W = 0.83125, p-value < 2.2e-16`

[a] Draw properly constructed histograms of *all three distributions* and discuss their properties,

[b] evaluate their sknewness (see **`e1071::skewness( )`**), and

[c] test whether these variables are approximately normal distributed (see **`ks-test( )`** or **`shapiro.test( )`**).

Address also the questions: Which transformed variable *comes the closest* to the normal distribution? Is the transformation with $\lambda = -1.0$ over-compensating the inherent positive skewness in the **`wage`** variable? [1.5 points]

Comments: The untransformed **wage** variable has positive skewness with an outlier $44500. The optimal transformation makes the transformed distribution almost symmetric with a tiny skewness value. However, the positive skewness is over-compensated when using ($\lambda = -1.5$). This leads to substantial negative skewness.

The *p*-values of Shapiro-Wilk normality tests with $H_0: X \sim N\left(\hat{u}, \widehat{\sigma^2}\right)$ for the properly Box-Cox transformed data is much smaller than 0.05, therefore transformed **wage** still deviates from the normal distribution. However, this *p*-value is the largest one among all three scenarios. Therefore, we can conclude the optimal transformed variable becomes closest to the normal distribution.

Task 2: Explore the function `powerTransform` to achieve a multivariate normal distribution [1 point]

Read up in R’s online help about the function **`car::powerTransfrom( )`** in the **`car`** library. Use the variables **`water81`**, **`water80`**, and **`water79`** from the **`Concord1.sav`** data file.

Task 2.1: ***Simultaneously*** estimate the ***optimal set*** of Box-Cox transformation parameters for all variables so that the set of transformed variables becomes approximately multivariate normal distributed. Report your code to do the estimation. [0.5 points]

```
> Concord <- foreign::read.spss("Concord1.sav",to.data.frame=TRUE)
> summary(powerTransform(cbind(water79, water80, water81) ~ 1, Concord))
bcPower Transformations to Multinormality
        Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
water79    0.1651        0.17       0.0824       0.2478
water80    0.1782        0.18       0.1005       0.2558
water81    0.2717        0.33       0.2022       0.3412
```

```
Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                               LRT df       pval
LR test, lambda = (0 0 0) 77.71615  3 < 2.22e-16

Likelihood ratio test that no transformations are needed
                               LRT df       pval
LR test, lambda = (1 1 1) 716.6848  3 < 2.22e-16
```

Task 2.2: Show the output and interpret the results. [0.5 points]

Comments: The *p*-values for transformation vectors (λ= 0,0,0), i.e., perform a log-transformation on all variables, and (λ= 1,1,1), i.e., leave all variables untransformed, are much smaller than α=0.05 so we can reject the two null hypotheses assuming these sets of transformation parameters lead to a multivariate normal distribution for the set of variables. The estimated transformation vector (λ= 0.17,0.18,0.33) comes the closest to the multivariate normal distribution.

Task 3: Confidence Intervals [2 points]

Continue with the **CPS1985** dataset for this task. To simplify things do not perform variable transformations.

Task 3.1: Run a bivariate regression of **wage** (dependent variable) on **education** (independent variable) and interpret the model estimates. [0.5 points]

```
> reg <- lm(wage~education, data=CPS1985)
> summary(reg)

Call:
lm(formula = wage ~ education, data = CPS1985)

Residuals:
   Min     1Q Median     3Q    Max 
-7.911 -3.260 -0.760  2.240 34.740 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.74598    1.04545  -0.714    0.476    
education    0.75046    0.07873   9.532   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.754 on 532 degrees of freedom
Multiple R-squared:  0.1459,  Adjusted R-squared:  0.1443 
F-statistic: 90.85 on 1 and 532 DF,  p-value: < 2.2e-16
```

Comment: The R^2 is only 0.1459 which means only about 14.6% of the variation in the dependent variable **wage** can be explained by the independent variable **education**. The statistically significant slope ($H_0\colon \beta_1 = 0$) is positive meaning with each additional year of education the income will increase the hourly wages by $0.75. While the intercept is not statistically different from zero, it should be keep in the model because [a] there is no logical reason why a person without education may have zero wages, and [b] because otherwise some statistics of the OLS model, such as the R^2 lose their properties.

Task 3.2: Calculate the 90 % confidence intervals around the estimated regression parameters. Can you draw the same conclusion as you did using the *t*-test in the **`summary`** output of task 3.1? [0.5 points]

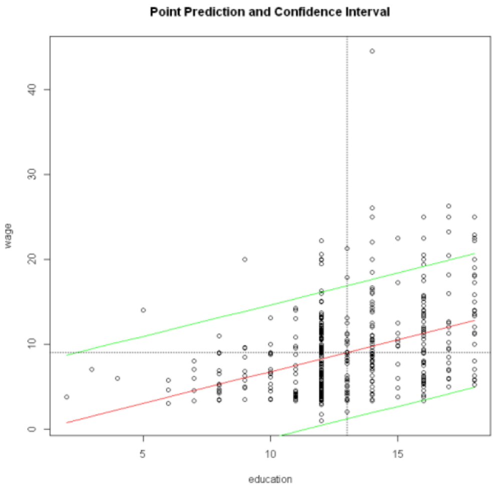
```
> cbind("Coef"=coef(reg), confint(reg, level=0.9))
                  Coef        5 %      95 %
(Intercept) -0.7459797 -2.4685982 0.9766389
education    0.7504608  0.6207294 0.8801921
```

Comments: The *t*-test investigates the null hypotheses that the estimated regression parameters are zero. That is, $H_0: \beta_0 = 0$ for the intercept and $H_0: \beta_1 = 0$ for the slope. As long as the $1-\alpha$ confidence intervals cover the values under the null hypothesis, that is $\beta_0=0$ and $\beta_1=0$, the null hypothesis cannot be rejected with an error probability of α.
Intercept: 0 is inside the confidence interval so we fail to reject the null hypothesis. *P* value for the *t*-test in task 3.1 is larger than $\alpha=0.01$ so we fail to reject the null hypothesis. Both methods lead to the identical conclusions: intercept is not different from 0.
Slope: 0 is outside the confidence so we can reject the null hypothesis. *P* value for the *t*-test in task 3.1 is smaller than $\alpha=0.01$ so we can reject the null hypothesis. Both methods lead to identical conclusions that income is significantly influenced by education.

Task 3.3: Scatterplot both variables and add the predicted regression line as well as the lower and upper 90% confidence interval lines around the ***point predictions***.(i.e., prediction interval in Hamilton and **`interval="prediction"`** in the **`predict`** function).

You may also want to enhance your plot by adding lines for the means of education and income as well as adding a title. [1 point]

```
> regPred <- predict(reg, interval="prediction", level = 0.90)
Warning message:
In predict.lm(reg, interval = "prediction", level = 0.9) :
      predictions on current data refer to _future_ responses
> CPSPred <- data.frame(CPS1985,regPred)
> CPSPred <- CPSPred[order(CPSPred$education),]
> plot(wage~education,data=CPSPred, main = "Point Prediction and Confidence
Interval")
> lines(CPSPred$education,CPSPred$fit,col="red")
> lines(CPSPred$education,CPSPred$lwr,col="green")
> lines(CPSPred$education,CPSPred$upr,col="green")
> abline(h=mean(CPSPred$wage), lty=3)
> abline(v=mean(CPSPred$education), lty=3)
```



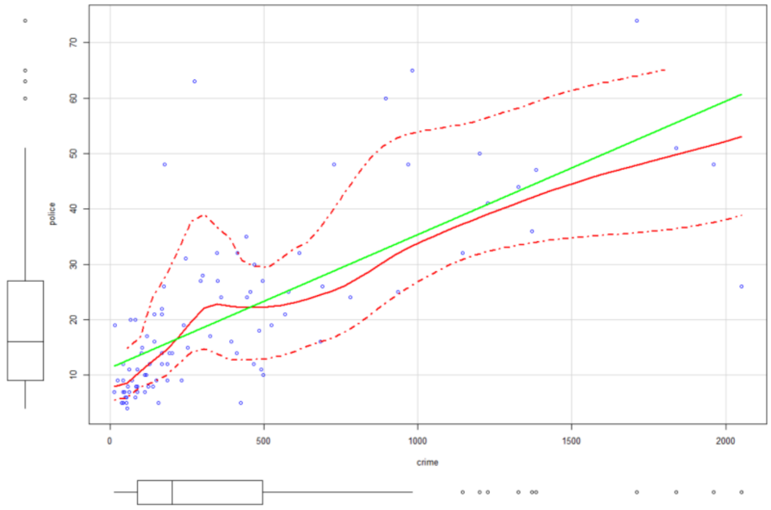
Task 4: Calibration and Prediction of a Bivariate Regression Model with Skewed Variables [4 points]

The **DBASE** file **`CampusCrime.dbf`** has among other variables the count variables **`crime`** (number of crimes committed on university campuses) and **`police`** (size of the campuses police forces).

Note: To import **DBASE** files use the R function **`foreign::read.dbf( )`**.

Task 4.1: Scatterplot **`police`** in dependence of **`crime`** including their box-plots along the margins. Is a data transformation advisable?

```
Crime <- foreign::read.dbf("CampusCrime.dbf" )
scatterplot(police ~ crime , data = crime)
```



Comments: The dependent variable **police** and the independent variable **crime** should be transformed because both are positively skewed. Furthermore, the residuals of the linear regression model are also slightly positively skewed. To make sure the residuals satisfy the assumption of ordinary least squares, it is advisable that both variables are transformed.

Task 4.2: Find a proper transformation of both variables in a way that the independent variable `crime` is approximately symmetrically distributed and that the transformation of the dependent variable `police` leads to approximately symmetrically distributed regression residuals.

```
> ## Transformation of independent variable
> summary(powerTransform(lm(crime~1, data=crime)))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.0258           0      -0.1444        0.196
Likelihood ratio test
                             LRT df    pval
LR test, lambda = (0) 0.08843754  1 0.76617
LR test, lambda = (1) 107.9414  1 < 2.22e-16

> ## Transformation of dependent variable so residuals become approx. symmetric.
> summary(powerTransform((lm(police~log(crime), data=crime))))

bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0051           0      -0.2474       0.2372

Likelihood ratio test
                              LRT df    pval
LR test, lambda = (0) 0.001726968  1 0.96685
LR test, lambda = (1) 63.30872  1 1.7764e-15
```

The suggested lambda parameters are λ=0.0258 for the independent variable and λ=−.0051 for the regression model **lm(police~log(crime), data=crime)** so that the residuals are approximately normal or at least symmetrically distributed.

Task 4.3: Test whether a ***log***-transformation (i.e., $\lambda = 0$) is appropriate for both, the dependent and the independent, variables.

Important: If the ***log***-transformation is appropriate ($H_0: \lambda = 0$ is insignificant) then ***use it*** because their relationship can now be interpreted *in terms of elasticities*.

Comments: The likelihood ratio tests in task 4.2 suggests that the estimated lambda coefficients are not significantly different from zero. Therefore, both crime and police can be log-transformed.

Task 4.4: Estimate the model in the transformed system and interpret the estimates. Also ***test*** if the elasticity (i.e., slope parameter) differs significantly from the neutral elasticity of 1, i.e., $H_0: \beta_1 = 1$. This could be done manually by using β_1's standard error from the regression output or by using the function `car::linearHypothesis`.

Tips:
1. Interpret the estimates as elasticity.
2. Use *t*-test to test if the slope parameter differs significantly from 1, $H_0: \beta_1 = 1$.
3. This is a two-sides *t*-test, you should double your p-value obtained from the cumulative *t*-distribution.

$$H_0: \beta_1 = 1$$

$$H_1: \beta_1 \neq 1$$

```
> ## Estimate the elasticity model
> elast.lm <- lm(log(police)~log(crime), data=crime)
> summary(elast.lm)
Call:
lm(formula = log(police) ~ log(crime), data = Crime)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4893 -0.2910 -0.0575  0.3301  1.3459 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.27907    0.23950   1.165    0.247    
log(crime)   0.46573    0.04326  10.766   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4793 on 91 degrees of freedom
Multiple R-squared:  0.5602,  Adjusted R-squared:  0.5554 
F-statistic: 115.9 on 1 and 91 DF,  p-value: < 2.2e-16
```

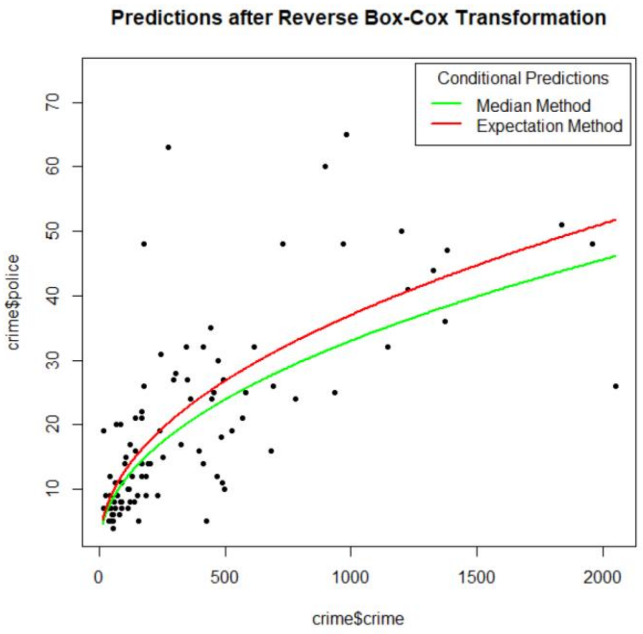
Comments: Since the bivariate regression model is specified in the log-log form, the results can be interpreted in terms of elasticity. One percent change in the number of crimes committed on university campuses will lead to 0.47 percent change in the size of the campuses police forces. The meaningful null hypothesis for elasticity is that 1% in the independent variable will lead to 1% change in the dependent variable.

```
> ## Test for H0: beta_log(crime)=1
> slope <- coef(elast.lm)[2]
> se <- summary(elast.lm)$coefficients[2, 2]
> df <- nrow(crime) - 2
> (t.value <- (slope-1)/se) # Note E(slope)=1 under H0
log(crime) 
 -12.35116 
> 2*pt(-abs(t.value),df = df) # one-sided significance using cumulative 
distribution
  log(crime) 
3.686307e-21 

> ## Alternative approach. Note: F == t.value^2
> linearHypothesis(elast.lm, c("log(crime) = 1"))
Linear hypothesis test

Hypothesis:
log(crime) = 1

Model 1: restricted model
Model 2: log(police) ~ log(crime)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     92 55.962                                  
2     91 20.910  1    35.052 152.55 < 2.2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: The *p* value is virtually zero for this test, thus we can reject this null hypothesis $H_0: \beta = 1$. The elasticity 0.46573 of the regression model is significantly different from unity, which means the relative change of police is much less elastic than the relative change of crime. Ultimately it means that the relationship exhibits decreasing rates of law enforcement allocation.
Notice: This is a two-sided test. Thus the *p*-value needs to be doubled when you look up the table for only one tail. The *F*-statistic of the **linearHypothesis( )** function is equal to the squared *t*-value.

Task 4.5: Perform a prediction in the original data units and plot the ***median*** and ***expectation*** curves. Interpret the plot both in terms of the median and expectation curves.



Comments: For the predictions being mapped back into the original scale, the expected predicted value is larger than the median predicted value because mean is larger than median in the positively skewed distribution. This applies over the full data range of the independent variable.
Notice the nonlinear relationship: at lower crime values proportionally a higher number of additional police officers is needed than at higher crime values. Ultimately a larger police force gives the police department more flexibility to react to crimes with less down-time of individual police officers.

Task 4.6: Provide a ***clean*** R script that documents your analysis. Do ***not show*** the provided functions in the sample script.

See above code for the individual tasks.

Task 5: Numerical Integration [2 points]

Evaluate the three distance decay functions along a line of around the central reference point zero.

$$f_1(x) = \exp(-|x|^1)$$

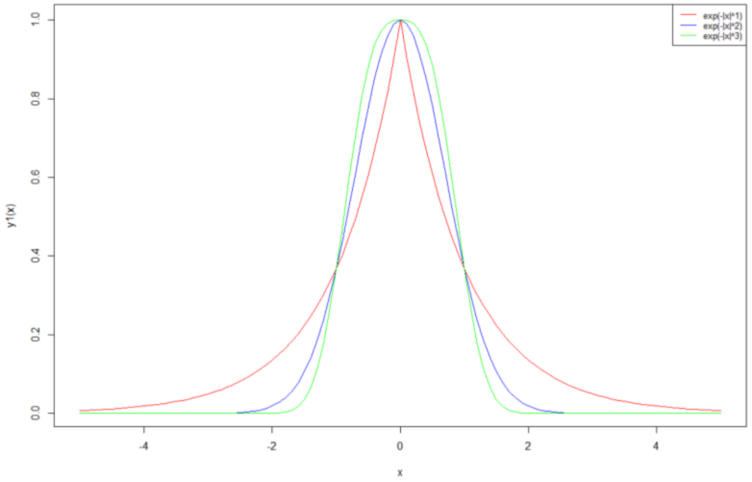
$$f_2(x) = \exp(-|x|^2)$$

$$f_3(x) = exp(-|x|^3)$$

over their full distance range $-\infty \leq x \leq \infty$. Show the ***clean*** code for the tasks below.

Task 5.1: Plot these three functions within a reasonable value range of the distance x. How does the shape of the curves change with increasing power?

```
> x <- seq(from = -5,to = 5,by = 0.1)

> y1 <- function(x){return(exp(-(abs(x)^1)))}

> y2 <- function(x){return(exp(-(abs(x)^2)))}

> y3 <- function(x){return(exp(-(abs(x)^3)))}

> plot(x,y1(x),col = 'red',type = 'l')

> lines(x,y2(x),col = 'blue',type = 'l')

> lines(x,y3(x),col = 'green',type = 'l')

> legend("topright", legend=c("exp(-x^1)", "exp(-x^2)","exp(-x^3)"), col=c("red", "blue",'green'), lty=1, cex=0.8)
```



Task 5.2: Evaluate the areas A_1, A_2 and A_3 underneath the three curves over their full support $-\infty \leq x \leq \infty$ with R's **`integrate( )`**-function.

Task 5.3: Calculate the expectation of the distances $E_i(x) = \int_{-\infty}^{\infty} x \cdot f_i(x)/A_i \cdot dx$ for all three distance decay functions $i \in \{1,2,3\}$. Could you predict the Expectation by just looking at the plot of the three curves?

Task 5.4: Calculate the variances of the distances $Var_i(x) = \int_{-\infty}^{\infty}(x - E_i(x))^2 \cdot f_i(x)/A_i \cdot dx$ for all three distance decay functions $i \in \{1,2,3\}$.

Hints: [a] The distance decay function $f_i(x)$ must the rescaled by its area A_i to make it a proper statistical density function, which integrates to one. [b] In R the professional integration function is **`integrate( )`**. Study the online help for this R functions. [c] To learn how to properly setup the functions to be integrated (aka, integrand), see for example the **RIEMANNSUM.R** or the online help of the **`integrate( )`**-function. [d] The **`integrate( )`**-function is more robust and may accept as upper and lower bounds the parameter **`Inf`** and **`-Inf`** rather than an arbitrary truncation point. [e] Additional parameters, other than **`x`**, to the integrand function, such as $A_i$or $E_i(x)$, can be passed in the **`integrate( )`**-function using its placeholder argument "…". See the online help for the function **`integrate( )`**.

```
>f <- function(lambda){
  A <- integrate(function(x) {exp(-abs(x)^lambda)}, -Inf, Inf)
  fnE <- function(x) {x*exp(-abs(x)^lambda)/A$value}
  E <- integrate(fnE, -Inf, Inf)
  fnV <- function(x) {((x-E$value)^2)*exp(-abs(x)^lambda)/A$value}
  Var <- integrate(fnV, -Inf, Inf)
  return(cbind(Lambda=lambda, Area=A$value, Expectation=E$value,
                Variance=Var$value)) }
> f(1)
     Lambda Area Expectation Variance
[1,]      1    2           0        2
> f(2)
     Lambda     Area Expectation Variance
[1,]      2 1.772454           0      0.5
> f(3)
     Lambda     Area Expectation  Variance
[1,]      3 1.785959           0 0.3732822
```

Task 6: Hamilton Appendix 1 [2 points]

Task 6.1: Why is a t-distributed random variable T with df_T degrees of freedom when it is squared T^2 identically to the F-distributed random variable with one degree of freedom in the numerator and df_T degrees of freedom for the denominator? Hint: use the definition of both random variables. [0.8 points]

The elementary variables $z \sim N(0,1)$ are independently standard normal distributed. The sums for squared variables z_i and z_j , that is, $s_1 = \sum_{i=1}^{n} z_i^2$ and $s_2 = \sum_{j=1}^{m} z_j^2$, are χ^2-distributed with n and m degrees of freedom, respectively. A variable $z^2 \sim \chi^2_{df=1}$. Because

$$F_m^n = \frac{s_1/n}{s_2/m} \; and \; t_m = \frac{z}{\sqrt{s_2/m}}$$

the squared *t*-statistic $t_m^2 = \frac{z^2}{s_2/m}$ becomes equivalent with the *F*-statistic if the numerator consists of just a *sum-of-one* standard normal distributed variable, that is, F_m^1 with the numerator having just one degree of freedom.

Task 6.2: Answer the questions **4 a**, **b**, **c**, and **d** in Hamilton's exercises on page 300. [1.2 points]

Use Word's equation editor to typeset your answers.

4a. since $\tilde{X} = X + \varepsilon$, so the expectation of $\tilde{X}$ equal the expectation of X plus the error term, ε. That is, $E(\tilde{X})$=$E(X)$+$\mathrm{E}(\varepsilon)$. An unbiased estimator of X is the expected value of $\tilde{X}$ minus the expectation of ε.

4b.

$$var(\tilde{X}) = var(X + \varepsilon)$$
$$= var(X) + var(\varepsilon) + cov(X, \varepsilon)$$

4c. if $cov(X, \varepsilon) = 0$, then $var(\tilde{X}) = var(X) + var(\varepsilon)$

4d. if $cov(X, \varepsilon) \neq 0$, the conclusion in part c does no hold and the covariance needs to be explicitly accounted for.