Quiz 1, STATS 401 W18

In lab on 10/5

This document produces different random quizzes each time the source code generating it is run. The actual quiz will be a realization generated by this random process, or something similar.

This version lists all the questions currently in the quiz generator. The quiz will have one question drawn at random from each of the five categories. No new questions will be added after Wednesday 10/3. Small changes may be made.

Instructions. You have a time allowance of 40 minutes, though the quiz may take you less time and you can leave lab once you are done. The quiz is closed book, and you are not allowed access to any notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor.

Formulas

The following formulas will be provided. To use these formulas properly, you need to make appropriate definitions of the necessary quantities.

(1) $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$

(2) $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}[X^2] - \left(\mathrm{E}[X]\right)^2$

(3) The probability density function of the standard normal distribution is $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$

(4) Syntax from `?pnorm`:

```
pnorm(q, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1)
q: vector of quantiles.
p: vector of probabilities.
```

Q1. Matrix exercises

Q1-1.

(a). Evaluate $\mathbb{AB}$ when

$$\mathbb{A} = \begin{bmatrix} 2 & 3 \\ 1 & 3 \\ -1 & -2 \end{bmatrix}, \quad \mathbb{B} = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}$$

Solution:

$$\mathbb{AB} = \begin{bmatrix} 6 & 5 \\ 3 & 4 \\ -3 & -3 \end{bmatrix}$$

(b). For $\mathbb{A}$ as above, write down $\mathbb{A}^{\mathrm{T}}$.

Solution:

$$\mathbb{A}^{\mathrm{T}} = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 3 & -2 \end{bmatrix}$$

(c). For $\mathbb{B}$ as above, find $\mathbb{B}^{-1}$ if it exists. If $\mathbb{B}^{-1}$ doesn't exist, explain how you know this.

Solution:

$$\mathbb{B}^{-1} = \frac{1}{3}\begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix}$$

Q1-2.

(a). Evaluate $\mathbb{A}\mathbb{B}$ when

$$\mathbb{A} = \begin{bmatrix} -1 & -1 & 3 \\ 2 & 0 & 3 \end{bmatrix}, \quad \mathbb{B} = \begin{bmatrix} -1 & 1 & -2 \\ 0 & 0 & 0 \\ -2 & 3 & 0 \end{bmatrix}$$

Solution:

$$\mathbb{A}\mathbb{B} = \begin{bmatrix} -5 & 8 & 2 \\ -8 & 11 & -4 \end{bmatrix}$$

(b). For $\mathbb{A}$ as above, write down $\mathbb{A}^{\mathrm{T}}$.

Solution:

$$\mathbb{A}^{\mathrm{T}} = \begin{bmatrix} -1 & 2 \\ -1 & 0 \\ 3 & 3 \end{bmatrix}$$

(c). For $\mathbb{A}$ as above, find $\mathbb{A}^{-1}$ if it exists. If $\mathbb{A}^{-1}$ doesn't exist, explain how you know this.

Solution:

Only square matrices can be invertible. $\mathbb{A}$ is 2×3 and so cannot have an inverse.

Q2. Summation exercises

Q2-1.

Calculate $\sum_{i=k}^{k+3}(i+3)$, where k is a whole number. Your answer should depend on k.

Solution:

$\sum_{i=k}^{k+3}(i+3) = k + (k+1) + (k+2) + (k+3) = 4k + 6.$

Q2-2.

Evaluate $\sum_{i=1}^{30} 10 - \sum_{i=10}^{20} 20$.

Solution:

$\sum_{i=1}^{30} 10 - \sum_{i=10}^{20} 20 = 30 \times 10 - 11 \times 20 = 300 - 220 = 80.$

Q2-3.

Calculate $\sum_{k=m}^{n} a$, where m and n are whole numbers and a is a real number.

Solution:

$\sum_{k=m}^{n} a = (n - m + 1)a$ since the sum has $(n - m + 1)$ terms each of which is a.

Q2-4.

Evaluate $3\sum_{k=1}^{5} 2 - 0.5\sum_{i=2}^{11} 6$.

Solution:

$3\sum_{k=1}^{5} 2 - 0.5\sum_{i=2}^{11} 6 = 3 \times 10 - 0.5 \times 60 = 0$

Q3. R exercises

Q3-1.

(a) Which of the following is the output of `matrix(c(rep(0,times=4),rep(1,times=4)),ncol=2)`

(i). $\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$ (ii). $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$ (iii). $\begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$ (iv). $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$

Solution:

(i), since R fills matrices by columns.

(b) Suppose we define an R vector by `y <- c(3,NA,-1,4,NA,-2)`. What will `y[y>0]` give you?

(i). A vector of the positive elements and NA values of `y`.

(ii). A vector of the negative elements of `y`.

(iii). A vector of all `NA`s.

(iv). A vector of `TRUE`s and `FALSE`s.

(v). A vector of `TRUE`s and `FALSE`s and `NA`s.

Solution:

(i). Indexing by `y>0` should pick out positive terms, but `NA` terms remain `NA` since R cannot tell if missing data are positive.

Q3-2.

(a) Which one of the following lines of code successfully constructs the matrix $\mathbb{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$

(i). `A <- matrix(c(1,1,2,2,3,3) ,nrow=3)`

(ii). `A <- cbind(c(1,1),c(2,2),c(3,3))`

(iii). `A <- t(matrix(c(1,1,2,2,3,3) ,nrow=2))`

(iv). `A <- c(c(1:3),c(1:3))`

Solution:

TBD

(b) Suppose `X` is a matrix in R. Which of the following is NOT equivalent to `X`?

(i). `t(t(X))`

(ii). `X %*% matrix(1,ncol(X)`

(iii). `X*1`

(iv). `X%*%diag(ncol(X))`

Solution:

(ii). for (i), the transpose of the transpose gets back the original matrix. (iii) is elementwise multiplication by 1. (iv) is matrix multiplication with the identity matrix.

Q3-3.

(a) Which of the following is the matrix $\mathbb{A}$ generated by

```
A <- t(matrix(c(rep(1,times=2),rep(3,times=2), 6, 4),ncol=3))
```

(i) $\mathbb{A} = \begin{bmatrix} 1 & 1 \\ 3 & 3 \\ 6 & 4 \end{bmatrix}$

(ii) $\mathbb{A} = \begin{bmatrix} 1 & 3 & 6 \\ 1 & 3 & 4 \end{bmatrix}$

(iii) $\mathbb{A} = \begin{bmatrix} 1 & 3 \\ 1 & 6 \\ 1 & 3 \end{bmatrix}$

(iv) $\mathbb{A} = \begin{bmatrix} 1 & 1 & 3 \\ 3 & 6 & 4 \end{bmatrix}$

Solution:

(i). The matrix generated by `matrix(c(rep(1,times=2),rep(3,times=2), 6, 4),ncol=3)` is answer (iv) and the code then transposes this.

(b) Which of the following successfully select the first five odd elements of the vector `x <- c(1,2,3,4,5,6,7,8,9,10,11)`? (List all that apply. Do not list commands that will give an error)

(i) `x[rep(c(TRUE,FALSE),each=5)]`

(ii) `x[rep(c(TRUE,FALSE),times=5)]`

(iii) `x[rep(c(TRUE,FALSE),length=9)]`

(iv) `x[rep(c(TRUE,FALSE)][1:5]`

(v) `x[rep(c("TRUE","FALSE"),5)]`

(vi) None of the above

(vii) All of the above

Solution:

Only (iv). Here's what they give:

```
x <- c(1,2,3,4,5,6,7,8,9,10,11)
x[rep(c(TRUE,FALSE),each=5)]
```

```
## [1]  1  2  3  4  5 11
```

```
x[rep(c(TRUE,FALSE),times=5)]
```

```
## [1]  1  3  5  7  9 11
```

```
x[rep(c(TRUE,FALSE),length=9)]
```

```
## [1]  1  3  5  7  9 10
```

```
x[rep(c(TRUE,FALSE))][1:5]
```

```
## [1] 1 3 5 7 9
```

```
x[rep(c("TRUE","FALSE"),5)]
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA
```

Q3-4.

(a) Define the matrix `A` as:

```
##      [,1] [,2]
## [1,]    0    3
## [2,]    1    3
## [3,]    1    2
```

What is the output of `apply(A,2,mean)`?

(i). A vector of length 3 corresponding to the average of each row of `A`.

(ii). A vector of length 2 corresponding to the average of each column of `A`.

(iii). The mean of all the values in `A`.

(iv). The mean of the second column of `A`.

(v). The mean of the second row of `A`.

Solution:

(ii). A vector of length 2 corresponding to the average of each column of `A`.

(b) For each of the lines of code below, say whether it will correctly make 50 draws from the normal$(100, 20)$ distribution. Among the correct answers, comment briefly on some strengths and weaknesses from the perspective of writing good R code. Which answer do you think is the best code, and why?

(i) `rnorm(50,20,100)`

(ii) `rnorm(100,20,50)`

(iii) `rnorm(100,20,n=50)`

(iv) `rnorm(mean=100,sd=20,n=50)`

(v) `rnorm(n=50,mean=100,sd=20)`

(vi) `replicate(rnorm(100,20),50)`

(vii) `replicate(rnorm(n=1,mean=100,sd=20),n=50)`

(viii) `rnorm(50)*20+100`

(ix) `100+sqrt(20)*rnorm(50)`

Solution:

(iii), (iv), (v), (vii), (viii) are all correct.

(i), (ii), (vi) make incorrect assumptions about how R matches arguments and how R chooses default arguments when they are not provided. The convenience of default arguments, and matching arguments by position or name, comes at a price. If in doubt, use named arguments.

(ix) has the wrong scaling. Recall $\text{SD}(aX) = a\text{SD}(X)$ and $\text{Var}(aX) = a^2\text{Var}(X)$.

The easiest to read is probably (v). Arguments are labeled but also appear in the order matching the default, for familiarity.

```
head( rnorm(50,20,100) )
```

```
## [1]  198.22290 -211.10691  107.86046   23.58067  121.28287   63.22652
```

```
head( rnorm(100,20,50) )
```

```
## [1]  45.31174 -20.99976 -79.94235  -3.96463  24.20900 -24.77433
```

```
head( rnorm(100,20,n=50) )
```

```
## [1]  73.74268  82.32061 141.54190  58.01549  75.22988 119.80866
```

```
head( rnorm(mean=100,sd=20,n=50) )
```

```
## [1] 102.77372 109.02927 124.29548  73.51038  77.20744 133.51297
```

```
head( rnorm(n=50,mean=100,sd=20) )
```

```
## [1]  87.05917  72.01126 138.62349 109.23145 110.54877  99.26617
```

```
dim( replicate(50,rnorm(100,20)) )
```

```
## [1] 100  50
```

```
head( replicate(rnorm(n=1,mean=100,sd=20),n=50) )
```

```
## [1] 109.87415  68.05709 105.91662 126.67787  77.28649 124.95498
```

```
head( rnorm(50)*20+100 )
```

```
## [1]  94.51544 102.01456 100.91902 117.78477 107.70187 116.39337
```

```
head( 100+sqrt(50)*rnorm(50) )
```

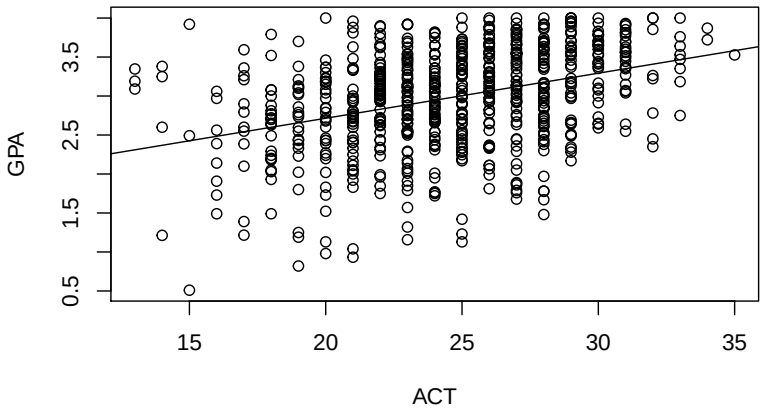
```
## [1]  95.53564  95.78399 101.77652  98.60334  96.12892 101.21262
```

Q4. Fitting a linear model by least squares

Q4-1.

The admissions officer at a large state university wants to assess how well academic success can be predicted based on information available at admission. She collects data on freshman GPA and highschool ACT exam scores for 705 students in an R dataframe called **gpa**. The plot below shows a line fitted to a scatterplot of the points in the dataset.

```
gpa_lm <- lm(GPA~ACT,data=gpa)
plot(GPA~ACT,data=gpa)
abline(coef(gpa_lm))
```



(a) Explain in words the criterion that is used to obtain the fitted line in the plot above.

Solution:

The line is fitted by least squares. This minimizes the sum of squared residuals, where the residual for each student is the difference between the value of GPA for that student and the value predicted by their ACT score.

(b) Defining appropriate notation, write an equation for the fitted model in subscript form. At this point, you don't have to explain how the coefficients are calculated.

Solution:

Let y_i be the freshman GPA for student i, $i = 1, \ldots, n$ with $n = 705$. Let x_i be the corresponding ACT score. The model in subscript form is

$$y_i = b_1 x_i + b_2 + e_i, i = 1, \ldots, n$$

where e_i is the residual for student i.

(c) Defining appropriate notation, write an equation for the fitted model in matrix form. You still don't have to explain how the coefficients are calculated.

Solution:

Define the column vector of coefficients as $\mathbf{b} = (b_1, b_2)$. Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_n)$ and let

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$$

Finally, let $\mathbf{e} = (e_1, \ldots, e_n)$ be a column vector of residuals. The matrix form of the linear model is

$$\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$$

(d) Now, explain using matrix notation how the model coefficients are calculated.

Solution:

The least squares choice of $\mathbf{b}$ is calculated using the equation

$$\mathbf{b} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$$

(e) Write an equation using subscript notation for the *fitted value* for the ith baseball player. Write a sentence to explain the interpretation of this fitted value.

Solution:

The fitted value for height x_i of player i is

$$\hat{y}_i = b_1 x_i + b_2$$

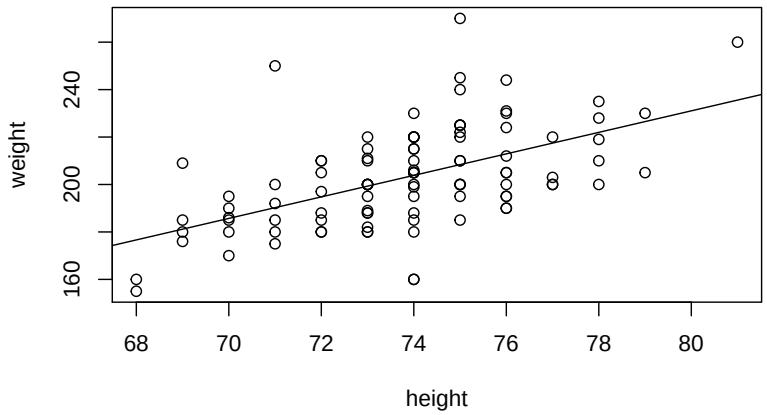
This is the predicted value of weight from the best fit line for weight of an individual with height x_i.

Q4-2.

A statistician employed by a major league baseball team is asked to assess the range of typical weights for major league baseball players of a given height. She obtains data from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights and reads them into R as a dataframe including variables 'Height' (in inches) and 'Weight' (in pounds) for each of 1035 Major League Baseball players. She starts by analyzing just the first 100 players.

She fits a linear model and plots the data and the resulting fitted line using the following R code:

```
weight_lm <- lm(weight ~ height)
plot(height,weight)
abline(coef(weight_lm))
```



(a) Write out the fitted linear model using subscript notation, including the following coefficients from `weight_lm`. This means you are asked to use actual numbers, rather than letters, for the model coefficients. Make sure to define any notation you introduce.

```
round(coef(weight_lm),3)
```

```
## (Intercept)      height
##    -131.652       4.534
```

Solution:

Let y_i be the weight of observation i, $i = 1, \ldots, 100$, and x_i be the corresponding height. The model in subscript form is

$$y_i = 4.53 \times x_i - 131.7 + e_i, \quad i = 1, \ldots, 100$$

where e_i is the residual for observation i.

(b) Use matrix notation to explain how these coefficients were calculated.

Solution:

Define the column vector of coefficients as $\mathbf{b} = (4.53, -131.7)$. Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_{100})$ and let

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_{100} & 1 \end{bmatrix}$$

We obtain $\mathbf{b}$ using the equation

$$\mathbf{b} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$$

(c) The tenth observation corresponds to Adam Stern, an outfielder for the Baltimore Orioles. His recorded height is 71 inches. Write out the formula for the fitted value for this observation. You do not need to simplify your calculation.

Solution:

We have the fitted value

$$\hat{y}_{10} = 4.53 \times 71 - 131.7$$

(d) Use matrix notation to write out an expression for the fitted values of the model. Make sure to define appropriate notation.

Solution:

Define $\mathbb{X}$ and $\mathbf{b}$ as above. Let $\hat{\mathbf{y}}$ be the column vector $(\hat{y}_1, \ldots \hat{y}_{100})$ where $\hat{y}_i$ is the fitted value corresponding to observation i. The fitted values are given by

$$\hat{\mathbf{y}} = \mathbb{X}\mathbf{b}$$

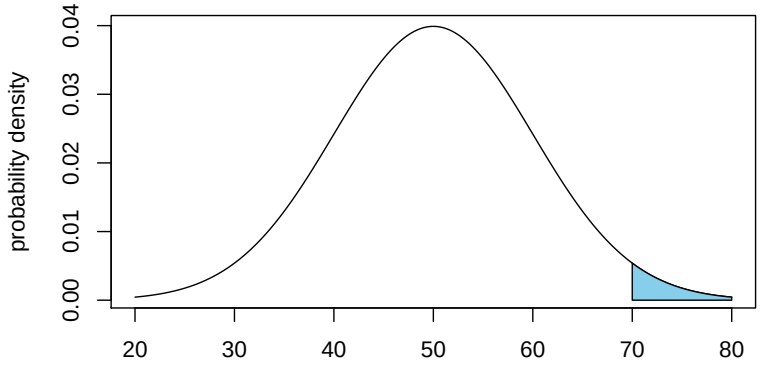
Since $\mathbf{b} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$ we can also write

$$\hat{\mathbf{y}} = \mathbb{X}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$$

Q5. Probability exercises

Q5-1.

The figure below shows the probability density function of a normal random variable X.



(a) By looking at the probability density function, estimate the mean and standard deviation of X. Use these estimates for the subsequent parts of this question.

Solution:

The center is at about 50. The points of inflection on the density are at about 40 and 60, which should be the mean plus/minus one standard deviation. It looks like about 95% of the area is between 30 and 70, which should be the mean plus/minus two standard deviations. These facts are consistent with a mean of 50 and an SD of 10.

(b) Write a probability statement about the random variable X that corresponding to the shaded area.

Solution:

The shaded area is $\mathrm{P}(X > 70)$.

(c) Write an integral corresponding to this shaded area.

Solution:

$$\int_{70}^{\infty} \frac{1}{\sqrt{2\pi 10^2}} \exp\left\{-\frac{(x-50)^2}{2\times 10^2}\right\} dx$$

(d) Write R code to evaluate this integral numerically.

Solution:

```
1-pnorm(70,mean=50,sd=10)
```

```
## [1] 0.02275013
```

It is acceptable not to label arguments, but then you have to get them in the right order!

Q5-2.

Let Y be a discrete random variable that takes values 0, 1, or 2 with probabilities 0.25, 0.5, and 0.25, respectively.

(a) What is the expected value of Y?

Solution:

The expected value is

$$0.25\times 0 + 0.5\times 1 + 0.25\times 2 = 1$$

(b) What is the variance of Y?

Solution:

The variance is

$$0.25\times(0-1)^2 + 0.5\times(1-1)^2 + 0.25\times(2-1)^2 = 0.5$$

Alternatively, we can calculate the expected value of X^2 first:

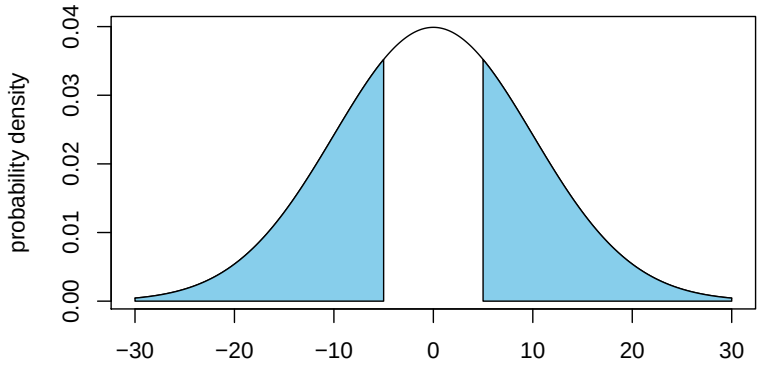
$$\mathrm{E}[X^2] = 0.25\times 0^2 + 0.5\times 1^2 + 0.25\times 2^2 = 0.5\times 1 + 0.25\times 4 = 1.5$$

and then

$$\mathrm{Var}(X) = 1.5 - 1^2 = 0.5$$

Q5-3.

The figure below shows the probability density function of a normal random variable X.



(a) By looking at the probability density function, estimate the mean and standard deviation of X. Use these estimates for the subsequent parts of this question.

Solution:

The center is at about 0. The points of inflection on the density are at about -10 and 10, which should be the mean plus/minus one standard deviation. It looks like about 95% of the area is between -20 and 20, which should be the mean plus/minus two standard deviations. These facts are consistent with a mean of 0 and an SD of 10.

(b) Write a probability statement about the random variable X that corresponding to the shaded area.

Solution:

The shaded area is $\mathrm{P}(-5 < X < 5)$.

(c) Write an integral corresponding to this shaded area.

Solution:

$$\int_{-5}^{5} \frac{1}{\sqrt{2\pi 10^2}} \exp\left\{-\frac{x^2}{2\times 10^2}\right\} dx$$

(d) Write R code to evaluate this integral numerically.

Solution:

The left tail is

```
pnorm(-5,mean=0,sd=10)
```

```
## [1] 0.3085375
```

The right tail is

```
1-pnorm(5,mean=0,sd=10)
```

```
## [1] 0.3085375
```

We need to add these up to get the total shaded area.

```
pnorm(-5,mean=0,sd=10)+1-pnorm(5,mean=0,sd=10)
```

```
## [1] 0.6170751
```

It is acceptable not to label arguments, but then you have to get them in the right order!
