Formative Assessment 1

PATAYON, SPIKE LEE-ROY V

2025-01-27

## Instruction

Submit exactly two files: (i) a PDF/HTML file with your write-up and graphs and (ii) a .r/.txt/.Rmd AND/OR Github link to files with code to reproduce your graphs.

1. Download the CyTOF data (see our GC).
2. Choose one of the columns of the dataset and make at least two plots showing its distribution (choose from ECDF, quantile, histogram, density estimate). What does the plot tell you about the distribution of the values in that column?
3. Choose any two of the columns of the dataset and make a Q-Q plot comparing the distributions. What does the Q-Q plot tell you about similarities or differences between the distributions of the values in the two columns? Note: You may use Lecture 2 to answer the items above.

import necessary libraries

library(lattice)  
 library(ggplot2)  
 library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

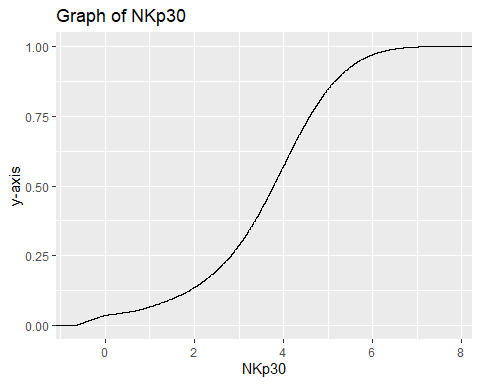
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data <- read.csv("C:\\Users\\spike\\Downloads\\cytof\_one\_experiment.csv")  
View(data)

head(data)

## NKp30 KIR3DL1 NKp44 KIR2DL1 GranzymeB CXCR6 CD161  
## 1 0.1875955 3.6156932 -0.5605694 -0.2936654 2.477893 -0.14470053 -0.3152872  
## 2 1.0348518 1.7001820 -0.2889611 -0.4798280 3.261016 -0.03392447 -0.4112129  
## 3 2.9996398 6.1411419 1.9032606 0.4823102 4.277562 1.94654156 -0.5022347  
## 4 4.2998594 -0.2211586 0.2425707 -0.4831267 3.351808 0.92622195 3.8772370  
## 5 -0.4386448 -0.5035892 -0.1526320 0.7506128 3.194145 -0.05893640 1.0907379  
## 6 2.0883050 -0.3992646 3.4550676 -0.5200856 4.345102 -0.36434277 -0.5705891  
## KIR2DS4 NKp46 NKG2D NKG2C X2B4 CD69 KIR3DL1.S1  
## 1 1.94497046 4.0818316 2.6200784 -0.3573817 -0.2711557 3.849965 -0.2554637  
## 2 3.80251714 3.7339299 -0.4832788 -0.4675984 -0.5594752 2.910197 -0.2909482  
## 3 -0.32010171 4.5594631 -0.5069090 2.6193782 -0.4554785 3.113454 3.6613886  
## 4 -0.16969487 4.4831486 1.9272290 -0.3110146 1.6350771 3.045998 0.2871241  
## 5 -0.05033025 0.8379358 -0.4581674 0.9216947 1.2419054 2.644422 0.4218294  
## 6 -0.45033591 4.0550848 3.4283565 0.6272837 -0.4157104 3.958158 0.7993406  
## CD2 KIR2DL5 DNAM.1 CD4 CD8 CD57 TRAIL  
## 1 5.3529769 -0.5092906 0.8811347 -0.32347280 -0.2822405 3.3254704 -0.6084228  
## 2 4.3132510 3.7774776 1.5406568 -0.13208167 0.9161920 2.4946442 -0.5034739  
## 3 5.5969513 0.8128166 1.0005903 -0.59933641 1.8382744 3.9897914 -0.2749380  
## 4 -0.5002885 0.3612212 1.2663267 -0.12568567 0.7667204 1.9950916 -0.5130930  
## 5 -0.5479527 1.0638327 0.8722272 -0.07107408 -0.1059012 3.4291302 -0.1433044  
## 6 5.1028564 3.0918867 0.8717267 -0.47986180 -0.2577198 -0.5784575 -0.5731323  
## KIR3DL2 MIP1b CD107a GM.CSF CD16 TNFa  
## 1 -0.30668543 1.2497120 -0.1295305 -0.43074102 3.9951417 0.90143498  
## 2 -0.54320954 2.8693060 -0.1887180 -0.16283845 4.4082309 1.93590153  
## 3 2.06488239 4.0955112 -0.1998480 3.18853825 6.0023244 -0.02336999  
## 4 2.11247859 3.3726018 -0.5720339 0.91310694 5.8238698 -0.60793749  
## 5 -0.02505141 -0.3099826 -0.1068511 -0.60370379 4.0122501 -0.61989100  
## 6 -0.28337673 -0.4108283 -0.1797545 -0.06372458 -0.5832926 0.14311030  
## ILT2 Perforin KIR2DL2.L3.S2 KIR2DL3 NKG2A NTB.A CD56  
## 1 -0.386027758 6.431983 1.22710292 2.660657999 -0.5220613 4.348923 2.897523  
## 2 2.983874845 6.814827 -0.04141081 3.841304627 4.6771149 3.474335 3.782870  
## 3 -0.521099944 5.099562 -0.16705075 -0.009694396 -0.4730573 5.634341 5.701186  
## 4 -0.043783559 5.841797 -0.51753289 -0.592990887 -0.4059049 4.598021 6.065672  
## 5 1.182703288 4.888777 -0.36251589 -0.398123704 -0.5440881 3.606101 1.966169  
## 6 -0.003258955 3.952542 -0.20194392 -0.202592720 3.8882776 2.346275 6.473243  
## INFg  
## 1 -0.3841108  
## 2 2.7186296  
## 3 2.5321763  
## 4 2.4564582  
## 5 3.1470092  
## 6 2.8282987

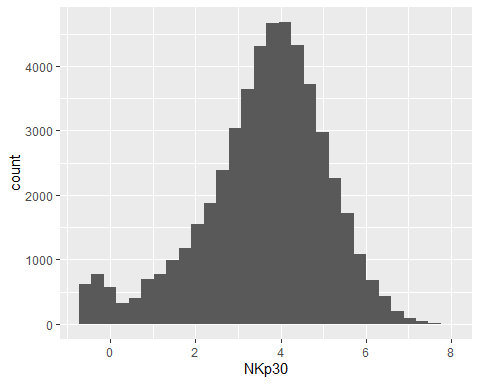
ggplot(data, aes(x = NKp30)) +   
 stat\_ecdf() +   
 xlab("NKp30") +   
 ylab("y-axis") +   
 ggtitle("Graph of NKp30")



Insights: The NKp30 values fall within the range shown by the x-axis, which extends from little below 0 to around 8. The curve begins at zero and rises steadily until rising sharply between two and six, then dropping down near one. This indicates that there are fewer values at the extremities of the NKp30 range, with the majority of values falling between 2 and 6. The ECDF’s S-shape indicates a sigmoidal or normal-like distribution with possible skewness, which means that most of the data points are not evenly distributed but rather grouped around a certain range.

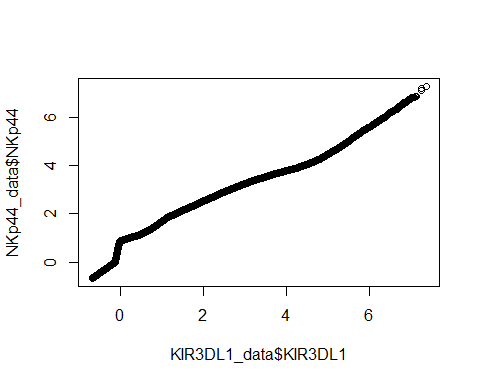
ggplot(data, aes(x = NKp30)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Insights: The x-axis displays values that are consistent with the ECDF plot, ranging from about 0 to just over 8. The bell-shaped, symmetrical histogram resembles a normal distribution. The majority of values are located in the middle, approximately between 3 and 4. The majority of NKp30 values are centered close to 4, as indicated by the highest bar, which is around 4. There is a small extension on the right tail, indicating a slight positive skew, even if the distribution is largely symmetrical. Although the right tail extends slightly, there isn’t any severe skewness. Standard statistical techniques like mean and standard deviation would be helpful descriptors because the distribution appears to be well-behaved.

KIR3DL1\_data <- data %>% subset(select = KIR3DL1)  
NKp44\_data <- data %>% subset(select = NKp44)  
  
qqplot(KIR3DL1\_data$KIR3DL1, NKp44\_data$NKp44)



Insights: The distributions of KIR3DL1 and NKp44 appear to have similar shapes, as most of the points fall on a straight diagonal line. There is a modest difference between the distributions’ tails, as seen by the deviations at both the lower and upper ends. While the modest deviation at the top-right hints that NKp44 may have more extreme high values (possible outliers), the small curvature at the bottom end shows that NKp44 has a heavier lower tail (more small values). The variances at the tails may indicate that scaling or modification (such as log transformation) is required to properly align these two variables, which should have the exact same distribution. The low and high ends show differences, indicating that NKp44 may have somewhat more extreme values.