Post2

Sangwook Kim

2017 11 23

Probablity Distributions

Introduction

The purpose of this post is to explore the various ways of plotting thress basic probability distributions, Normal Distribution, Binormal Distribution, and Poisson Distribution by using the package, ggplot2.

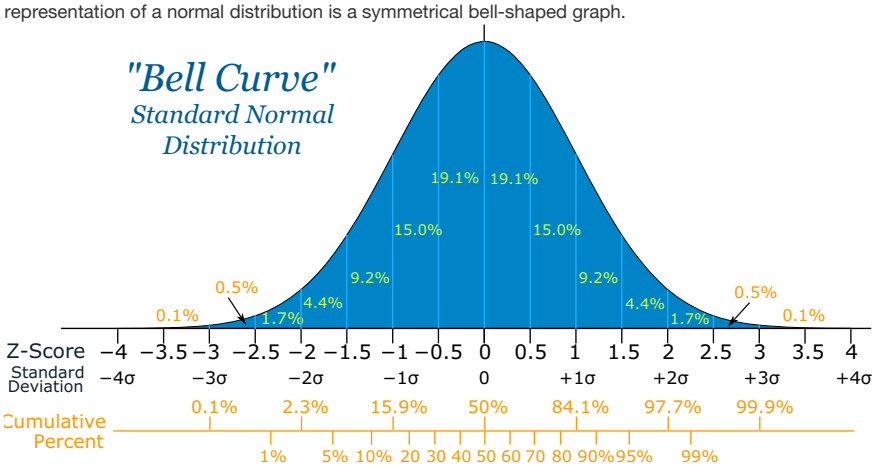
```
library(ggplot2)
```

The data we use for plotting distributions is cleansocres.csv, which is from the last homework.

```
dat <- read.csv("../hw04/data/cleandata/cleanscores.csv")
```

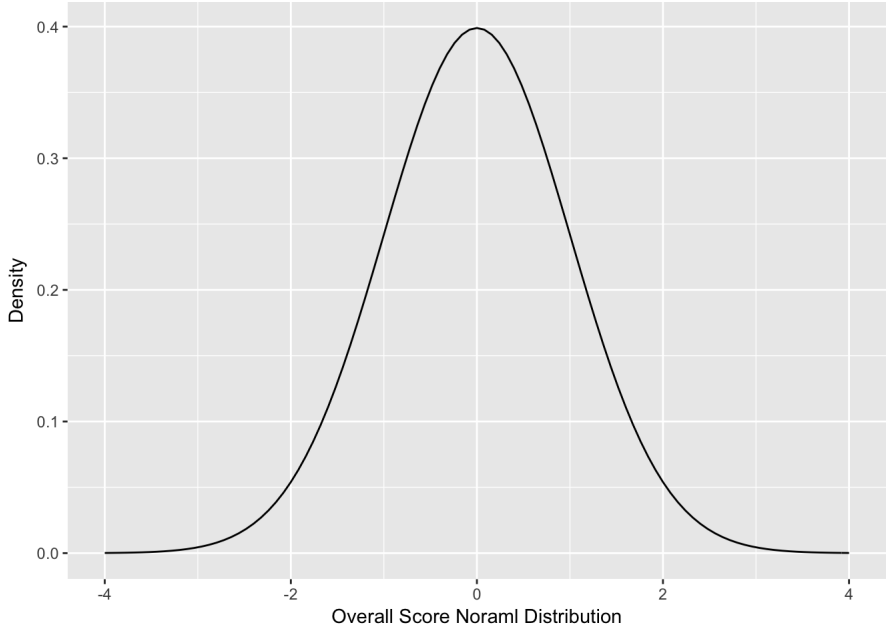
Normal Distribution

A normal distributon is simply the distribution of a data set in which most values are gathered in the middle of the range. So, a graphical representation of a normal distribution is a symmetrical bell-shaped graph.



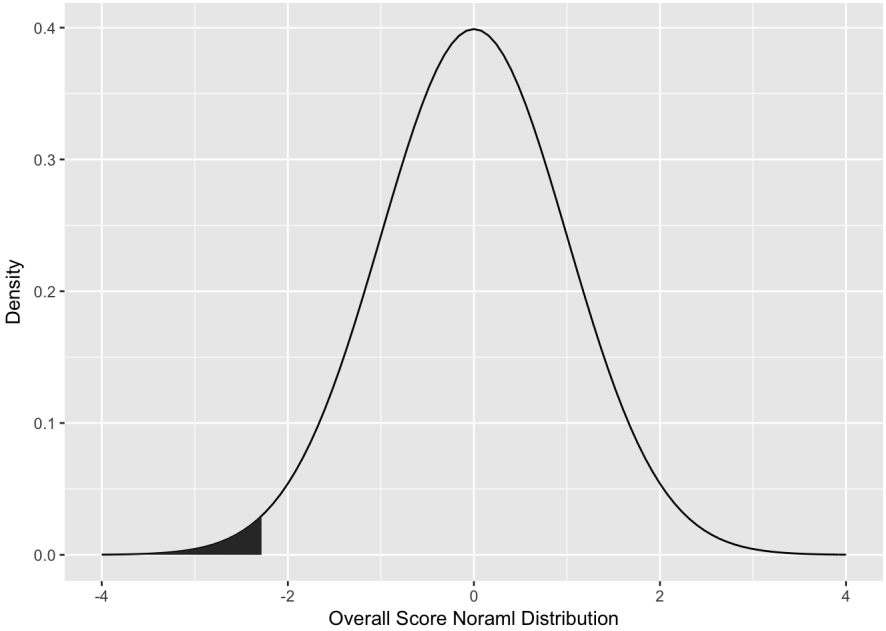
Let's plot a normal distributon of the overall scores in the data by using ggplot2.

```
# This is a standard normal distribution, which its mean is 0 and sd is 1.
p <- ggplot(dat, aes(Overall)) + stat_function(fun=dnorm) + xlim(c(-4,4)) + labs(x="Overall Score Noraml Distribut
ion", y="Density")
p
```



Let's use this graph to see probability of students whose overall score is less than 50.

```
less50 <- dat$Overall[dat$Overall < 50]
# Normalize
normLess50 <- (less50 - mean(dat$Overall))/sd(dat$Overall)
normLess50 <- seq(min(normLess50), max(normLess50), by=.1)

less50 <- data.frame(x=normLess50, y=dnorm(normLess50))
less50 <- rbind(c(min(dat$Overall), 0),
                less50,
                c(max(normLess50), 0)
                )
p + geom_polygon(data = less50, aes(x=x, y=y))
```



Let's find the probability of students whose overall score is less than 50.

```
# We can use the function pnorm to calculate the cumulative Distribution.
pnorm(max(normLess50), 0, 1)
```

```
## [1] 0.01122337
```

We've got 0.01122337, which means about 1% of students' overall grades is less than 50.

You might notice that this result is not precisely showing that less than 50 because we use **max(normLess50)**. This means that the probability of students less than 50 in the data. So, if you want to calculate the accurate probability of less than the overall grade of 50, we need to use a different value.

```
# Normalizing at 50
norm50 <- (50 - mean(dat$Overall)) / sd(dat$Overall)
pnorm(norm50, 0, 1)
```

```
## [1] 0.01257885
```

Binomial Distribution

A binomail distribution is a distribution of the possible numebr of success in a given number of trials and each trials have the same probability of sucess.

Based on data, we can find the probability of weather a student can get a grade above A-.

```
# The total number of students above 'A-'
total_of_a <- length(dat$Grade[dat$Grade == "A+" | dat$Grade == "A" | dat$Grade == "A-"])

# The chance of getting above 'A-'
p <- total_of_a / nrow(dat)

p
```

```
## [1] 0.3023952
```

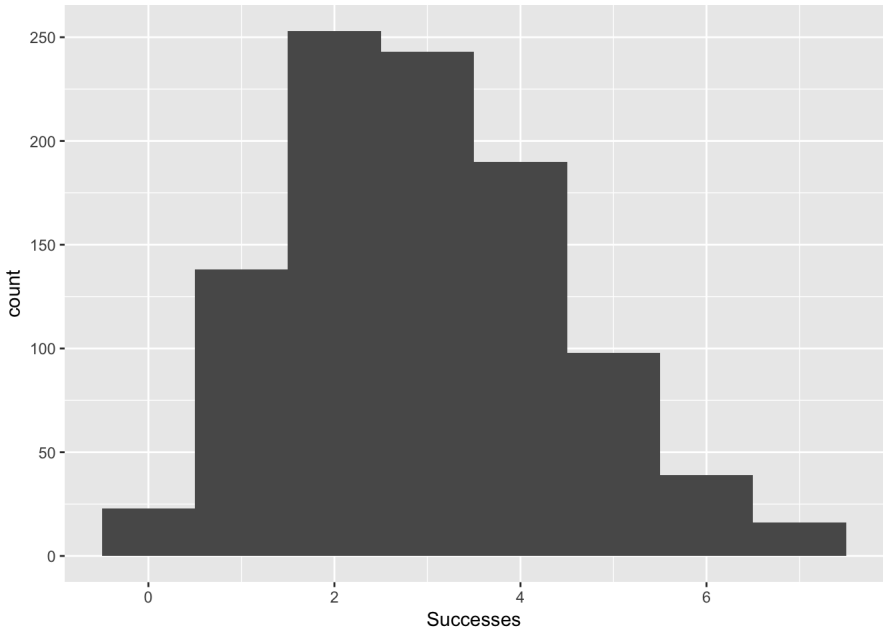
About 30% of students get above 'A-'. Let's use this probability to see how a binomial distribution works by using the function **rbinom**

```
# In the parameter, n is the number of observation, size is the number of trials.

rbinom(n=10, size = 1, prob = p)
```

```
##  [1] 1 0 1 1 1 0 0 0 0 0
```

Now, let's plot the binomial distribution by ggplot2.

```
# Let's use the same probability with n=1000, size=10

binom = data.frame(Successes = rbinom(1000, 10, p))
ggplot(binom, aes(x = Successes)) + geom_histogram(binwidth = 1)
```



We can calculate the probability of a specific value:

```
# The probability of a student got above 'A-' exactly 5 time in trials of 10.
dbinom(x=5, 10, p)
```

```
## [1] 0.1052743
```

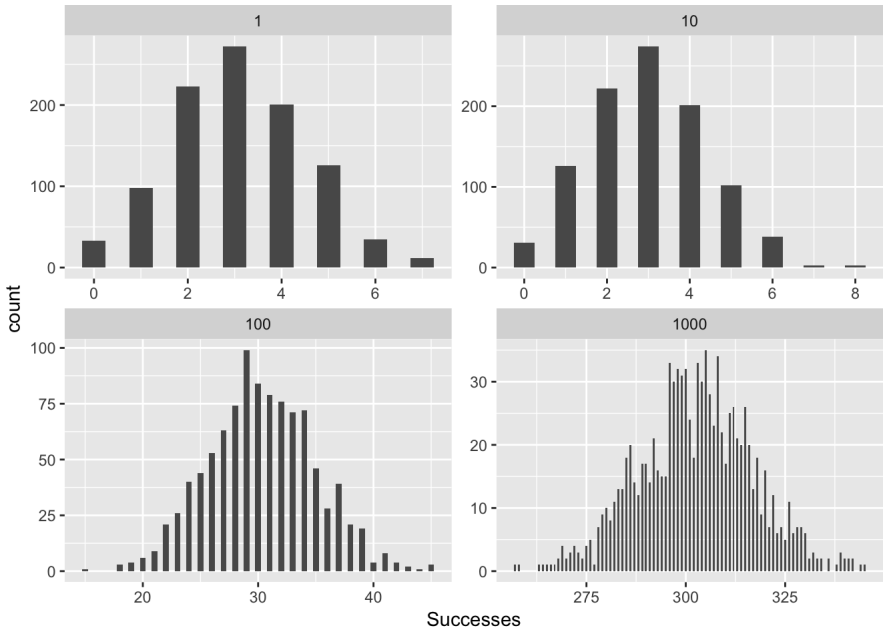
For the cumulative probability:

```
# The probability of a student got above 'A-' less equal to 5 time in trials of 10.
pbinom(q=5, 10, p)
```

```
## [1] 0.950867
```

There is a interesting thing about a binormal distribution. As the number of size increase, the binomial distribution is similar to the normal distribution.
Let's see the difference by faceting each different binomial distributions by its size(trial).

```
# Create each different size of a binomial data frame.
binom0 = data.frame(Successes = rbinom(1000, 10, p), size = 1)
binom1 = data.frame(Successes = rbinom(1000, 10, p), size = 10)
binom2 = data.frame(Successes = rbinom(1000, 100, p), size = 100)
binom3 = data.frame(Successes = rbinom(1000, 1000, p), size = 1000)

# To compare each distribution, let's use facet_wrap by size.
ggplot(rbind(binom0, binom1, binom2, binom3), aes(x=Successes)) + geom_histogram(binwidth = 0.5) + facet_wrap(~ si
ze, scales = "free")
```



Poisson Distribution

A poisson distribution is a distribution of the probability of a number of **independent** events occuring in a fixed time. Unlike a normal distribution, it is not graphically symmetrical but skewed to the left or right of the median.
The following formula is to calulate a probability of a poisson distribution.

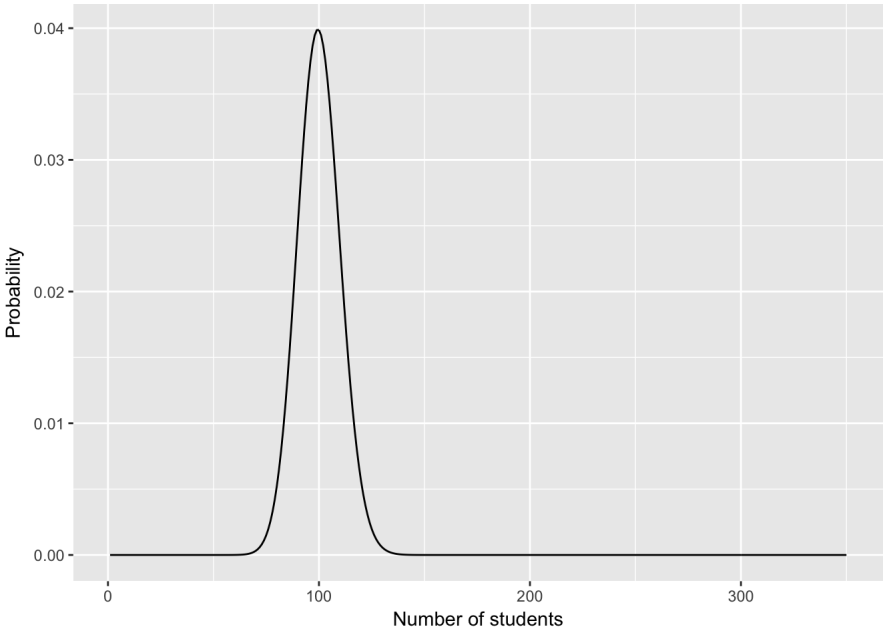
$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$

λ is the mean occurence per a given interval.
x is the number of occurences within a given interval

Then, let's assume that the average number of students get a grade above 'A-' is 100 in STAT133 per semester. Each semester STAT133 contains 350 students in the class.

Let's use this information to plot a possion distribution.

```
p <- ggplot(data.frame(x=c(1:350)), aes(x))

# The function dpois to calculate the probablity of each number of student.
# Use args=list(100) for the lambda value, 100.

p + stat_function(geom="line", n=350, fun=dpois, args=list(100)) + labs(x="Number of students", y="Probability")
```



What is the probability of exactly 110 students will get above 'A-' in next semester?

```
# lambda is 100.
# x is 110

dpois(110, 100)
```

```
## [1] 0.02342255
```

The chance of exactly 110 students get above is about 2.3%.

What is the probability of more than equal to 110 students will get above 'A-' in next semester?

```
# We need to calculate the cumulative probablity.

ppois(110, 100)
```

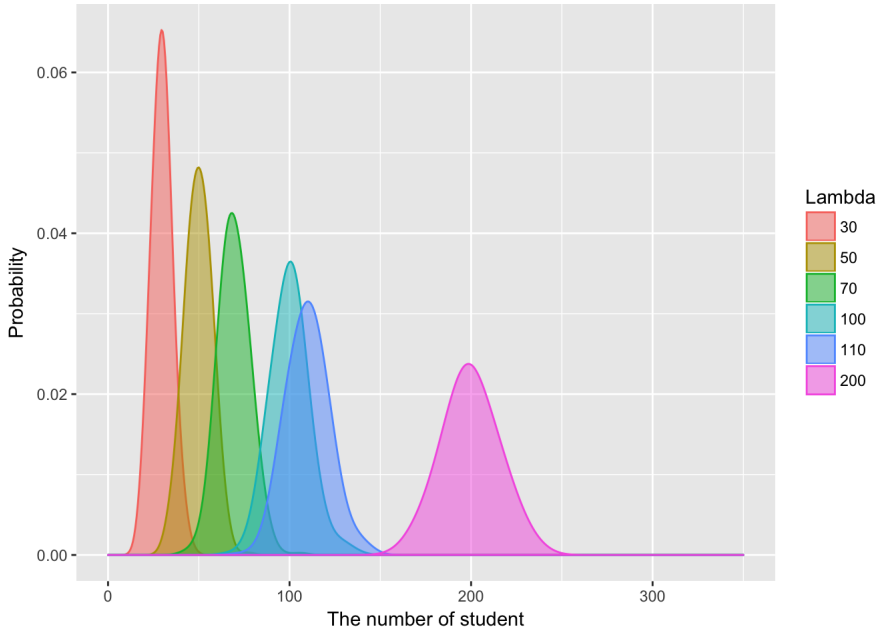
```
## [1] 0.8528627
```

The chance is 85%

As you can imagine, when lambda changes it gives a different shape of poisson distribution.

Let's see the difference by using ggplot2:

```
pois1 = rpois(n=350, lambda = 30)
pois2 = rpois(n=350, lambda = 50)
pois3 = rpois(n=350, lambda = 70)
pois4 = rpois(n=350, lambda = 100)
pois5 = rpois(n=350, lambda = 110)
pois6 = rpois(n=350, lambda = 200)

# Create a dataframe that contains each differenct lambda and its values.
pois = data.frame(Lambda=c(rep(30, 350), rep(50, 350), rep(70, 350), rep(100, 350), rep(110, 350), rep(200, 350)),
x = c(pois1, pois2, pois3, pois4, pois5, pois6))

pois$Lambda <- as.factor(pois$Lambda)

ggplot(pois, aes(x)) +
  geom_density(aes(group = Lambda, color=Lambda, fill=Lambda), adjust = 2, alpha = 0.5) + labs(x = "The number of
student", y = "Probability") + xlim(c(0,350))
```



As you can see, when the value of lambda increases, the poisson distribution tend to be skewed to the right.

Conclusion

We have explored the basic three probability distribution by using the data, **cleanscores.csv** and **ggplot2**. Each probability distritions has own characteristics. Thanks to the powerful graphical package, ggplot2, now we know the differences between them and how to plot each of them.