Tutorial 3
Distributions in R
Statistical Methods in Research
COSC 6323
Spring 2018

Ioannis Pavlidis
Dinesh Majeti
George Panagopoulos

Computational Physiology Lab

ipavlidis@uh.edu
dmajeti@uh.edu
gpanagopoulos@uh.edu

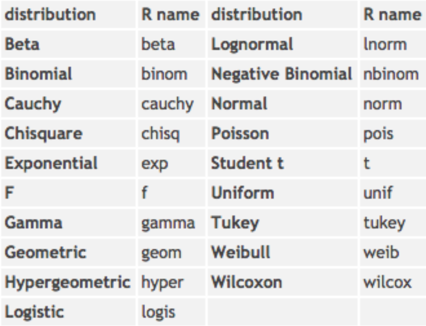
February 2, 2018

Distributions in R

Distributions in R

- Help about distributions

```
help.search("distribution")
```



distribution	R name	distribution	R name
Beta	beta	Lognormal	lnorm
Binomial	binom	Negative Binomial	nbinom
Cauchy	cauchy	Normal	norm
Chisquare	chisq	Poisson	pois
Exponential	exp	Student t	t
F	f	Uniform	unif
Gamma	gamma	Tukey	tukey
Geometric	geom	Weibull	weib
Hypergeometric	hyper	Wilcoxon	wilcox
Logistic	logis		

Figure 1: Common Priobability Distributions and their names in R

Distributions in R

- There are four commands for every distribution. A letter is used to indicate their functionality

  d gives the height of the probability density function

  p gives the cumulative density function

  q gives the inverse cumulative density function (quantiles)

  r gives randomly generated numbers

The Normal Distribution

- Functions associated with the normal distribution,

```
> ?Normal
```

- Probability value of the normal distribution at any point

```
> dnorm(0)
[1] 0.3989423

> dnorm(0,mean=3,sd=10)
[1] 0.03813878
```

The Normal Distribution - dnorm

- To plot the normal distribution

```
x = seq(-4, 4, length=100)
fx = dnorm(x)

plot(x, fx, type="l", lty=2, xlab="x value",
ylab="Density", main="Normal Distribution")

# ggplot
df = data.frame(x, y=fx)
ggplot(df, aes(x,y)) + geom_line(linetype="dashed")
+ xlab("x value") + ylab("Density")
+ theme(text= element_text(size=20))
```

The Normal Distribution

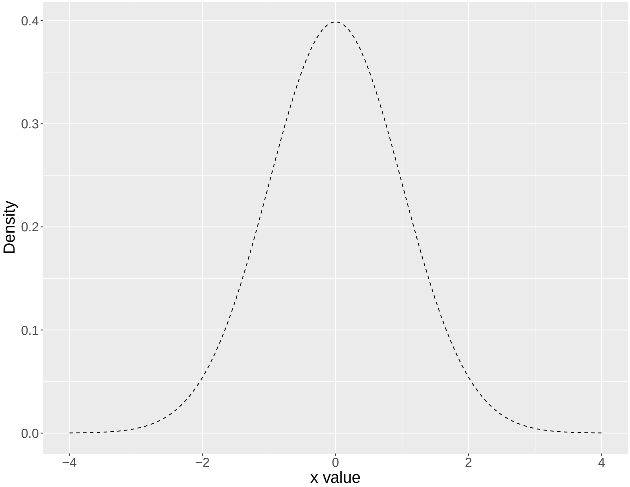


Figure 2: Normal Distribution with $\mu = 0, \sigma = 1$

The Normal Distribution Exercise

- Create a plot with normal distributions having the following parameters
  - $\mu = 0, \sigma = 0.5$
  - $\mu = 0, \sigma = 1$
  - $\mu = 0, \sigma = 2$
  - $\mu = -2, \sigma = 0.75$
- All the distribution should be in a single plot
- Each distribution should have a different color
- The plot should have a legend

The Normal Distribution Exercise

- To have multiple plots in a single plot, use

  ```
  lines()
  ```

- To have different colors and line types, use the following parameters

  ```
  col=..., lty=...
  ```

- To have a legend, use

  ```
  legend()
  ```

  with the colors and line styles as input.

The Normal Distribution Exercise

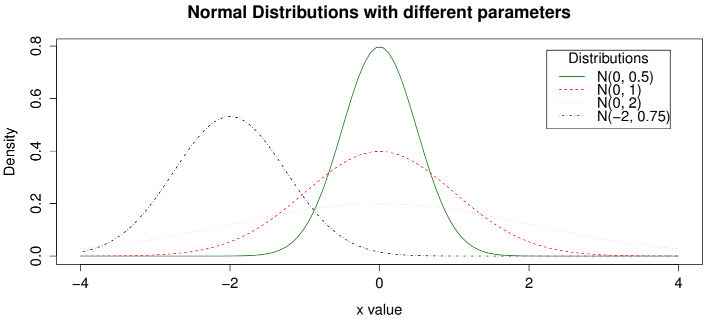


Figure 3: Normal distributions with different parameters

The Normal Distribution Exercise

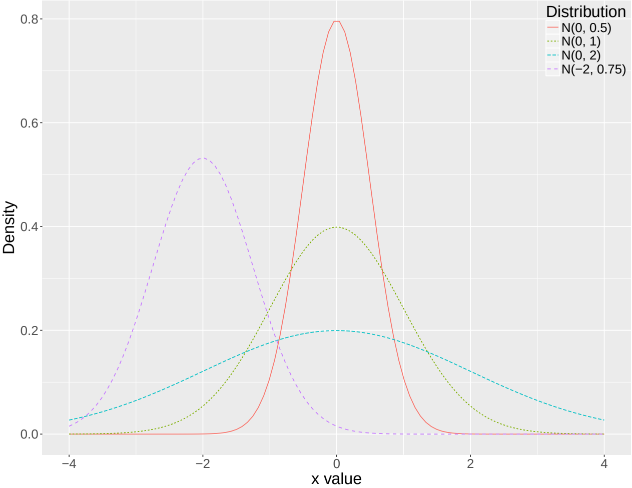


Figure 4: Normal distributions with different parameters - ggplot

The Normal Distribution Exercise

```
plot(x, dnorm(x,mean=0,sd=0.5), type="l", lty=1 ,
xlab="x value", ylab="Density",
main="Normal Distributions with different parameters",
col=’dark green’)

lines(x, dnorm(x), type="l", lty=2, col=’red’)
lines(x, dnorm(x,mean=0,sd=2), type="l", lty=3, col=’violet’)
lines(x, dnorm(x,mean=-2,sd=0.75), type="l", lty=4,
col=’black’)

labels <- c(’N(0, 0.5)’, ’N(0, 1)’, ’N(0, 2)’, ’N(-2, 0.75)’)
colors <- c(’dark green’,’red’,’violet’,’black’)
linetypes <- c(1, 2, 3, 4)
legend(’topright’,inset=.05, title="Distributions", labels,
lty = linetypes, col=colors)
```

The Normal Distribution Exercise

```
d1 = data.frame(D=’N(0, 0.5)’, x, y=dnorm(x,0,0.5))
d2 = data.frame(D=’N(0, 1)’, x, y=dnorm(x))
d3 = data.frame(D=’N(0, 2)’, x, y=dnorm(x,0,2))
d4 = data.frame(D=’N(-2, 0.75)’, x, y=dnorm(x,-2,0.75))

d = rbind(d1, d2, d3, d4)
ggplot(d, aes(x, y, col=D))
+ geom_line(aes(linetype=D))
+ xlab("x value") + ylab("Density")
+ theme(legend.justification=c(1,1),
      legend.position=c(1,1),
      legend.background = element_rect(fill="transparent"))
```

The Normal Distribution - pnorm

- *pnorm*(x) computes the probability that a normally distributed random number will be less than x
- It can be seen as the Cumulative Distribution Function

  ```
  > pnorm(0)
  [1] 0.5

  > pnorm(0,mean=2,sd=3)
  [1] 0.2524925

  > pnorm(1)
  [1] 0.8413447
  ```

- To find the probability that a number is larger than the given number, use the *lower.tail* option:

  ```
  > pnorm(1,lower.tail=FALSE)
  [1] 0.1586553
  ```

The Normal Distribution - qnorm

- *qnorm*() is the inverse of *pnorm*()
- Given a probability, *qnorm*() returns the number whose cumulative distribution matches the probability.

```
> qnorm(0.5)
[1] 0

> qnorm(0.5,mean=1,sd=2)
[1] 1

> qnorm(0.5,mean=2,sd=2)
[1] 2
```

The Normal Distribution - rnorm

- Generates random numbers whose distribution is normal

```
> rnorm(4)
[1]  1.2387271 -0.2323259 -1.2003081 -1.6718483

> rnorm(4,mean=3,sd=3)
[1] 4.580556 2.974903 4.756097 6.395894
```

The Normal Distribution - rnorm

```
> y <- rnorm(200)
> hist(y, main="Histogram of numbers generated by rnorm")

> qplot(y, binwidth=0.5)
```

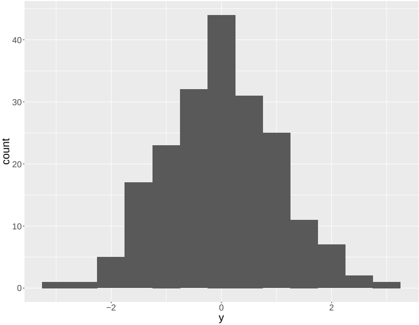


Figure 5: Distribution of numbers generated by *rnorm*

The Normal Distribution

- What is the probability that a randomly selected number from the standard normal distribution occurs within one standard deviation of the mean?

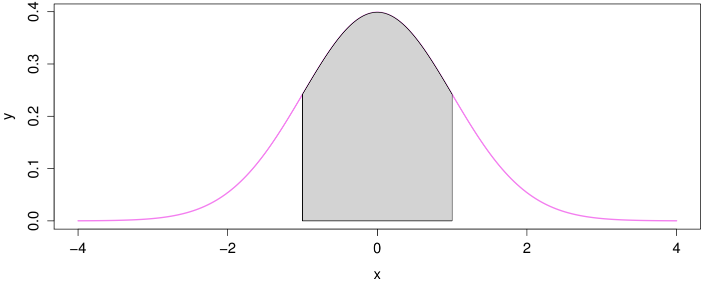


Figure 6: Normal distribution

The Normal Distribution

- Use *pnorm*

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
```

The Normal Distribution Exercise

- What is the probability that a randomly selected number from the standard normal distribution occurs within two standard deviations of the mean?

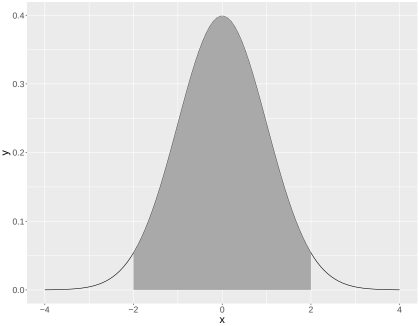


Figure 7: Normal distribution

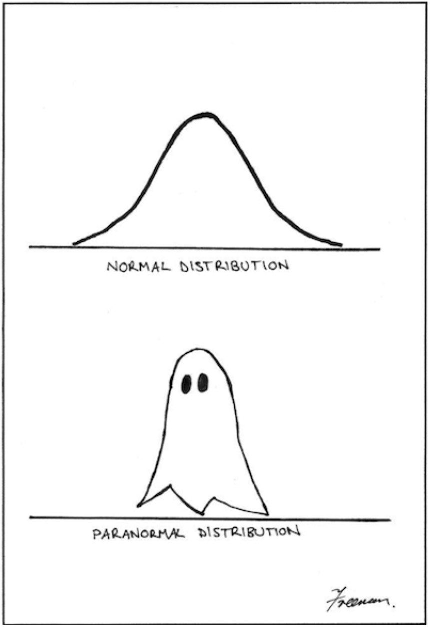
The Normal Distribution

- Use *pnorm*

```
> pnorm(2)-pnorm(-2)
[1] 0.9544997

x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="violet")
x1=seq(-2,2,length=200)
y1=dnorm(x1)
polygon(c(-2,x1,2),c(0,y1,0),col="light gray")

# ggplot
shade = data.frame(x=c(-2,x1,2),y=c(0,y1,0))
ggplot(df, aes(x,y)) + geom_line()
+ geom_polygon(data=shade, aes(x,y), fill="dark grey")
```


NORMAL DISTRIBUTION
PARANORMAL DISTRIBUTION
Freeman.

Exercise

- Suppose there are ten multiple choice questions in an English class quiz. Each question has 4 possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

The Binomial Distribution

- *binom* in R
- The probability of answering a question correctly by random is $1/4 = 0.25$.
- The probability of having exactly 4 correct answers by random attempts is

```
> dbinom(4, size=10, prob=0.25)
[1] 0.1459
```

The Binomial Distribution

- Using *dbinom* and *pbinom*

```
> dbinom(0, size=10, prob=0.25) +
+ dbinom(1, size=10, prob=0.25) +
+ dbinom(2, size=10, prob=0.25) +
+ dbinom(3, size=10, prob=0.25) +
+ dbinom(4, size=10, prob=0.25)
[1] 0.9218

> pbinom(4, size=12, prob=0.25)
[1] 0.9218
```

- Hence, the probability of correctly answering 4 or less questions in a multiple choice quiz is 92.2%

Exercise

- Operators of toll roads and bridges need information for staffing tollbooths so as to minimize queues (waiting lines) without using too many operators.
- Assume that in a specified time period the number of cars per minute approaching a toll booth has a mean of 10.
- Traffic engineers are interested in the probability that exactly 11 cars approach the tollbooth in a minute.

Exercise

- $\lambda = 10$

```
> dpois(11, lambda=10)
[1] 0.1137
```

- Therefore, there is approximately 11% chance that exactly 11 cars would approach the tollbooth in a minute.

More Exercises

- Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 79, and the standard deviation is 13.1. What is the percentage of students scoring 85 or more in the exam?

Solution

- Since we would like to find the percentages of students whose score is greater than or equal to 85, we are interested in the upper tail of the normal distribution

```
> pnorm(85, mean=79, sd=13.1, lower.tail=FALSE)
[1] 0.3234707
```

- Hence, the percentage of students scoring 85 or more in the exam is 32.3%

More Exercises

- Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

Solution

- The null hypothesis is that $\mu \geq 10000$.
- Compute the test statistic

```
> xbar = 9900            # sample mean
> mu0 = 10000            # hypothesized value
> sigma = 120            # population standard deviation
> n = 30                 # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                      # test statistic
[1] -4.5644
```

Solution

- Compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> -z.alpha                    # critical value
[1] -1.6449
```

- The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that mean lifetime of a light bulb is above 10,000 hours.

"I'm not an outlier; I just haven't found my distribution yet!"

References

- histogram - http://docs.ggplot2.org/0.9.3.1/geom_histogram.html
- polygon - http://docs.ggplot2.org/current/geom_polygon.html
- axis labels and legends - http://docs.ggplot2.org/0.9.2.1/labs.html