R and Stats - PDCB topic
Determining pvalues

LCG Leonardo Collado Torres
lcollado@wintergenomics.com – lcollado@ibt.unam.mx

March 11th, 2011

Apropos

- Just as a reminder, use apropos when you know that a function should include a given word:

  > apropos("history")

  ```
  [1] "history"     "loadhistory"
  [3] "savehistory"
  ```

  > apropos("help")

  ```
  [1] "help"           "help.request"
  [3] "help.search"    "help.start"
  [5] "link.html.help"
  ```

- Once you are familiar with a given function, args can be useful:

  > args(history)

  ```
  function (max.show = 25, reverse = FALSE, pattern, ...)
  NULL
  ```

Quick practice

- What function can we use to create a sequence of numbers?
- How can we repeat a vector?

Quick answers :P

- What function can we use to create a sequence of numbers? seq or : in some cases
- How can we repeat a vector? rep

Numbers from a population

- One way to get random numbers is to define a population and extract X elements from it.
- Which function can we use to do this step?

Sample

- We use sample:

```
> x <- 1:10
> args(sample)
function (x, size, replace = FALSE, prob = NULL)
NULL
> sample(x, 9)
[1] 10  2  4  1  8  6  3  7  5
> sample(x, 11, TRUE)
 [1]  2  3  4  9 10  1  1  7  8  7  2
```

Random from a distribution

- Using sample works for some cases. But what if we want random numbers from a given distribution?
- R has functions for all the important distributions which makes thing very easy for us :)
- Which one do we use to generate random numbers from the normal distribution?

rnorm

- We simply use rnorm:

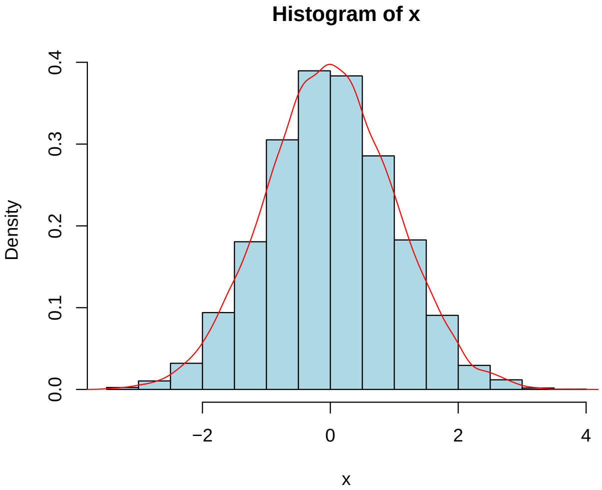
```
> args(rnorm)

function (n, mean = 0, sd = 1)
NULL

> set.seed(123)
> rnorm(10)

 [1] -0.56047565 -0.23017749  1.55870831
 [4]  0.07050839  0.12928774  1.71506499
 [7]  0.46091621 -1.26506123 -0.68685285
[10] -0.44566197
```

- Quick exercise: prove visually that rnorm gets random numbers from the normal distribution.

A histogram could work

```
> x <- rnorm(10000)
> hist(x, prob = TRUE, col = "light blue")
> lines(density(x), col = "red")
```

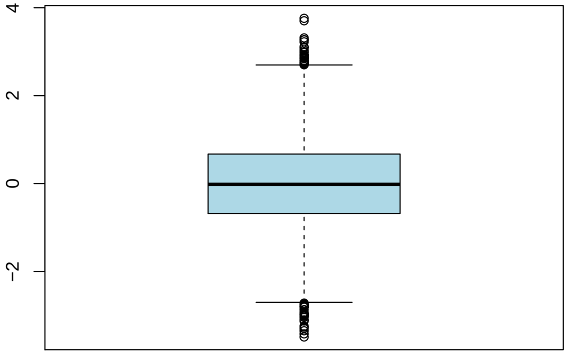
A histogram could work



Boxplot seems better

```
> boxplot(x, col = "light blue")
```

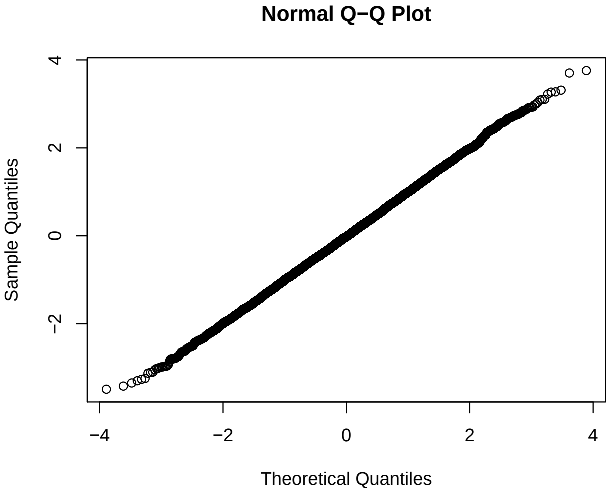
Boxplot seems better



QQnorm is the way to go

```
> qqnorm(x)
```

QQnorm is the way to go



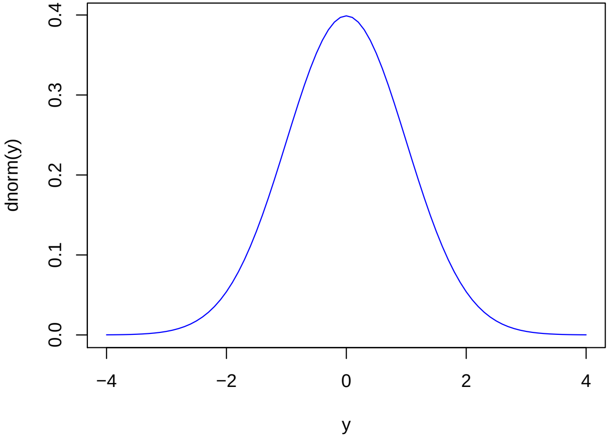
Density

- Now that we know how to get random numbers from a given distribution, the next step is to understand the distribution.
- One way of doing so is by plotting the density of the distribution. It's like plotting the probability of getting X, Y and Z values.
- Which function would help us to get the density values of the normal distribution?
- As a short exercise, plot the density from -4 to 4 of a normal dist with mean 0 and standard dev. 1
- Does it seems likely that a random number with this dist. would have a value of 0? -2? 2? 4?

dnorm

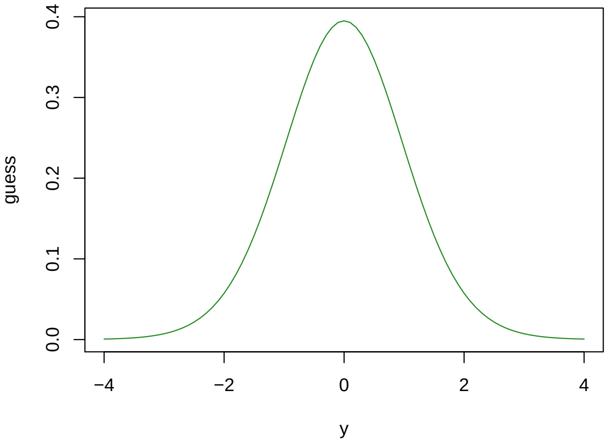
```
> y <- seq(-4, 4, 0.1)
> plot(y, dnorm(y), type = "l", col = "blue")
```

dnorm



What is the following distribution?

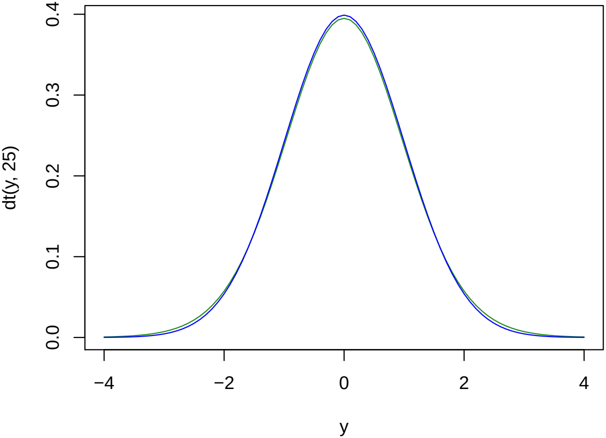
What is the following distribution?



It's a t-student!

```
> plot(y, dt(y, 25), type = "l",
+      col = "forest green")
> lines(y, dnorm(y), type = "l",
+      col = "blue")
```

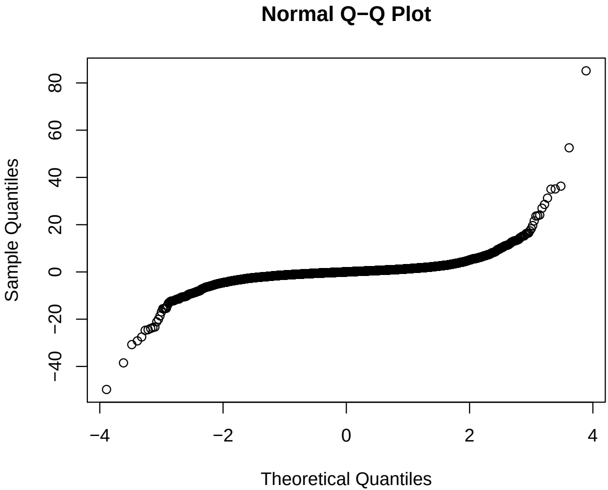
It's a t-student!



QQnorm of a t-student

```
> tt <- rt(10000, 2)
> qqnorm(tt)
```

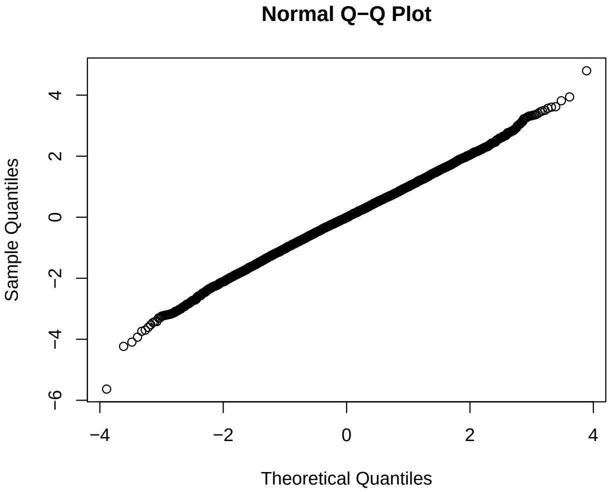
QQnorm of a t-student



QQnorm of a t-student 2

```
> tt2 <- rt(10000, 25)
> qqnorm(tt2)
```

QQnorm of a t-student 2



Central limit theorem

- Basically, if you have a large number of variables with finite mean and variance, the distribution will end up looking as the normal distribution.
- That's why the t-student with 25 degrees of freedom is nearly like the normal dist. with mean 0 and variance 1.
- For the curious ones, `http://en.wikipedia.org/wiki/Central_limit_theorem` might be a good place to start.

Distributions

We have used the normal dist. so far, but others exist:

- Lognormal: plnorm(x, mean, sd)
- Student's *t*: pt(x, df)
- *F* dist.: pf(x, n1, n2)
- Chi-square: pchisq(x, df)
- Binomial: pbinom(x, n, p)
- Poisson: ppois(x, lambda)
- Uniform: punif
- Exponential: pexp(x, rate)
- Gamma: pgamma(x, shape, scale)
- Beta: pbeta(x, a, b)

Finding probabilities

- Earlier, we saw that 4 (and -4) seemed pretty unlikely to pop up in random numbers following a normal dist. with mean 0 and variance of 1.
- We can empirically determine the probability of finding a value equal to or larger than 4 or equal to or smaller than -4:

```
> set.seed(123)
> lots <- rnorm(1e+07)
> length(which(lots >= 4 | lots <=
+     -4))/length(lots)

[1] 6.39e-05
```

Finding probabilities II

- While the above seems to work, we can find the actual probabilities using the p family of functions such as pnorm:

  ```
  > args(pnorm)

  function (q, mean = 0, sd = 1, lower.tail = TRUE, log.p
  NULL

  > pnorm(4)

  [1] 0.9999683
  ```

- What did we find with pnorm(4)?

Finding probabilities III

- pnorm(4) gave us the probability of finding a value smaller than 4.
- Below we determine the probability of finding a number equal to or larger than 4:

```
> 1 - pnorm(4)

[1] 3.167124e-05

> pnorm(4, lower.tail = FALSE)

[1] 3.167124e-05
```

- Determine the probability of finding a value greater than 4 or smaller than -4.

Finding probabilities IV

- We have two simple ways to do so:

```
> 2 * pnorm(4, lower.tail = FALSE)
[1] 6.334248e-05
> pnorm(4, lower.tail = FALSE) +
+     pnorm(-4)
[1] 6.334248e-05
```

- Remember that the normal dist. is symmetrical :)
- Our empirical value was close enough:

```
> length(which(lots >= 4 | lots <=
+     -4))/length(lots)
[1] 6.39e-05
```

- What we just did was find the p-value :)

Other options

- Surely, if you want you can use printed tables such as `http://www.statsoft.com/textbook/distribution-tables/`[1]
- You can also use more rigid calculators such as `http://www.graphpad.com/quickcalcs/pvalue1.cfm` Yet they do not give you detailed values for extreme cases.
- Try finding the p-value for a Z value of 4.

[1] They are common in books.

Practice

- With the previous calculator, find the p-value for a Z value of 2.
- Can you reproduce it in R?

Practice II

- The calculator gives the following answer: *Z=2 The two-tailed P value equals 0.0455 By conventional criteria, this difference is considered to be statistically significant.*
- Doing the following gives us a different answer:

```
> pnorm(2, lower.tail = FALSE)
[1] 0.02275013
> 1 - pnorm(2)
[1] 0.02275013
```

- Where is the incongruency?

Practice III

- We only calculated the p-value for a one-tail distribution. Meaning, that we were only interested on values larger than 2 (or smaller than -2) and we completely ignored extreme values smaller than 2 (or larger than 2). We simply need to multiply our p-value by 2:

  ```
  > 2 * pnorm(2, lower.tail = FALSE)
  ```

  ```
  [1] 0.04550026
  ```

- Our answer is more exact than the one given by the calculator :)

Exercises

Taken from the *Introductory Statistics with R* book.

1. Calculate the probability for each of the following events:
   - A standard normally distributed variable is larger than 3.

     ```
     > 1 - pnorm(3)
     [1] 0.001349898
     > pnorm(3, lower.tail = FALSE)
     [1] 0.001349898
     ```

   - A normally distributed variable with mean 35 and standard deviation 6 is larger than 42.

     ```
     > 1 - pnorm(42, mean = 35, sd = 6)
     [1] 0.1216725
     > pnorm(42, mean = 35, sd = 6, lower.tail = FALSE)
     ```

Exercises

```
[1] 0.1216725
```

- Getting 10 out of 10 successes in a binomial distribution with probability 0.8

  ```
  > dbinom(10, size = 10, prob = 0.8)
  [1] 0.1073742
  ```

- $X < 0.9$ when X has the standard uniform distribution.

  ```
  > punif(0.9)
  [1] 0.9
  ```

- $X > 6.5$ in a chi-squared distribution with 2 degrees of freedom.

  ```
  > 1 - pchisq(6.5, df = 2)
  [1] 0.03877421
  > pchisq(6.5, df = 2, lower.tail = FALSE)
  ```

Exercises

```
[1] 0.03877421
```

2. A rule of thumb is that 5% of the normal distribution lies outside an interval approx +-2s about the mean. To what extent is this true? Where are the limits corresponding to 1%, 0.5% and 0.1%? What is the position of the quartiles measured in standard deviation units?

```
> pnorm(-2) * 2
[1] 0.04550026
> qnorm(1 - 0.1/2)
[1] 1.644854
> qnorm(0.1/2, lower.tail = FALSE)
[1] 1.644854
```

Exercises

```
> qnorm(0.005/2, lower.tail = FALSE)
[1] 2.807034
> qnorm(0.001/2, lower.tail = FALSE)
[1] 3.290527
> qnorm(0.25)
[1] -0.6744898
> qnorm(0.75)
[1] 0.6744898
```

Exercises

3. For a disease known to have a postoperative complication frequency of 20%, a surgeon suggests a new procedure. He tests it on 10 patients and there are no complications. What is the probability of operating on 10 patients successfully with the traditional method?

   ```
   > dbinom(0, size = 10, prob = 0.2)
   ```

   ```
   [1] 0.1073742
   ```

4. Simulated coin-tossing can be done using `rbinom` instead of `sample`. How exactly would you do that?

   ```
   > rbinom(10, 1, 0.5)
   ```

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

Exercises

```
> ifelse(rbinom(10, 1, 0.5) == 1,
+      "Head", "Tail")

 [1] "Head" "Tail" "Head" "Head" "Tail"
 [6] "Tail" "Tail" "Head" "Head" "Head"

> c("Aguila", "Sol")[1 + rbinom(10,
+      1, 0.5)]

 [1] "Aguila" "Aguila" "Aguila" "Aguila"
 [5] "Aguila" "Sol"    "Sol"    "Aguila"
 [9] "Aguila" "Sol"
```

Session Information

```
> sessionInfo()

R version 2.12.0 (2010-10-15)
Platform: i386-pc-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices
[4] utils     datasets  methods
[7] base
```