R and Stats - PDCB topic
Hypothesis testing: parametric tests

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

March 25th, 2011

Practical approach

- We use them to compare a sample to an expected distribution
- To compare a sample to another sample
- To check if two samples are from the same distribution

What we can conclude

- By default, we have a null hypothesis that we accept and we'll test it against an alternative hypothesis.
- If the p-value is significant, we can reject the null hypothesis in favor of the alternative one. Yet, we are not giving definite proof that the alternative hypothesis is true!
- It is very important to take into account the assumptions of a given test!

T test

- It is the main parametric test used for hypothesis testing.
- What is the difference between parametric and non parametric tests?

One sample case

- We have $x_1, \ldots, x_n$ which are assumed to be independent realizations of random variables with a distribution $N(\mu, \sigma^2)$.
- Our null hypothesis is that $\mu = \mu_0$
- Do we know μ?

Estimates

- We estimate μ with the empirical mean $\overline{x}$
- Likewise, we estimate σ with the standard deviation s
- The *standard error of the mean* (SEM) describes the variation of the average of n random values with mean μ and variance σ^2. $SEM = \sigma/\sqrt{n}$

SEM

- The SEM will tell us how far or close we were from estimating the real mean μ
- Basically, if you repeat an experiment, the means from the experiments should have a tight distribution around the true mean.
- Yet, one sample is enough to get SEM.
- The t test will check if μ_0 is within $2xSEM$ of μ within an acceptance region at a given significance level.

$$t = \frac{\overline{x} - \mu_0}{SEM} \tag{1}$$

Degrees of freedom

- Small samples have *heavier* tails than N(0,1) simply because *SEM* might be too small.
- Therefore, we correct t distribution with $f = n - 1$ degrees of freedom

Is the result significantly different?

- If it falls outside the acceptance region, it is.
- More exactly, we calculate the p-value.
- If the p-value is smaller than the significance level we reject the . . . hypothesis.

Why do we use the one side test?

- Simply if you have other information that points you to the direction of the effect.
- In such cases you only test against one of the tails of the t distribution.
- Note that doing so changes the acceptance region and the p-value.
- If your result is not significant, then it isn't! Don't change to a two ways test just to get a significant result!

Quick exercise

- Below we have the daily energy intake in kJ for 11 women. Is it different from the recommended value of 7725 kJ?

```
> daily <- c(5260, 5470, 5640, 6180,
+       6390, 6515, 6805, 7515, 7515,
+       8320, 8770)
```

- What is our null hypothesis? Our alternative one?
- Which function do we use to do the t test?
- What is our conclusion at a 5% significance level?

Quick exercise

- *t* test:

```
> t.test(daily, mu = 7725)

One Sample t-test

data:  daily
t = -2.7682, df = 10, p-value =
0.01985
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
 5986.539 7537.098
sample estimates:
mean of x
 6761.818
```

Quick exercise

- Note that the output shows information on:
  1. the data that we are testing
  2. the degrees of freedom
  3. the p-value
  4. the alternative hypothesis
  5. the 95% confidence interval, what is it for?
  6. the sample mean of x

Quick exercise

- What did we do wrong?

Quick exercise

- If our H_0 is $\mu = 7225$ and our H_1 is $\mu < 7725$ and we are using a significance level of 5%, what do we conclude?

More info on CIs

- It's calculated with:

$$\overline{x} - t_{0.975}(f) * SEM < \mu < \overline{x} + t_{0.975}(f) * SEM$$

- It is the interval where you expect the true mean to lie on. It's basically the range of μ_0 values that cause t to lie within its acceptance region.
- With a larger sample, the interval should be smaller given the same variation.
- If you decrease the confidence, then the interval is larger for the same data set.

Theory

- We used the one way t test to check if the true mean is significantly different from a given value.
- Two-sample t tests are used to test the hypothesis that two samples come from distributions with the same mean.
- It's nearly the same, just that we wave two independent groups.
- SEDM is the *standard error of difference of means* and the t test is:

$$t = \frac{\overline{x}_2 - \overline{x}_1}{SEDM} \tag{2}$$

Same variance?

- That's the question you need to ask before doing the t test with two samples.
- The underlying statistical methods vary quite a bit depending on the answer to this question.
- Which functions can we use to answer this question visually?

Practice

- We'll use a data set from the ISwR package.
- You can install it quicklly with:

```
> install.packages("ISwR")
```

- Lets check the data first:

```
> library(ISwR)
> attach(energy)
> head(energy)
```

Practice

```
  expend stature
1   9.21   obese
2   7.53    lean
3   7.48    lean
4   8.08    lean
5   8.09    lean
6  10.15    lean
> class(energy)
[1] "data.frame"
> dim(energy)
[1] 22  2
```

Practice

- We want to test whether both samples come from the same distribution.
- We can do so by specifying x and y:

```
> t.test(energy$expend[energy$stature ==
+      "lean"], energy$expend[energy$stature ==
+      "obese"])
```

```
Welch Two Sample t-test

data:  energy$expend[energy$stature == "lean"] and ener
t = -3.8555, df = 15.919, p-value =
0.001411
alternative hypothesis: true difference in means is not
95 percent confidence interval:
```

Practice

```
 -3.459167 -1.004081
sample estimates:
mean of x mean of y
 8.066154 10.297778
```

- Or we can take advantage of the formula notation:

  > t.test(expend ~ stature)

```
Welch Two Sample t-test

data:  expend by stature
t = -3.8555, df = 15.919, p-value =
0.001411
alternative hypothesis: true difference in means is not
95 percent confidence interval:
```

Practice

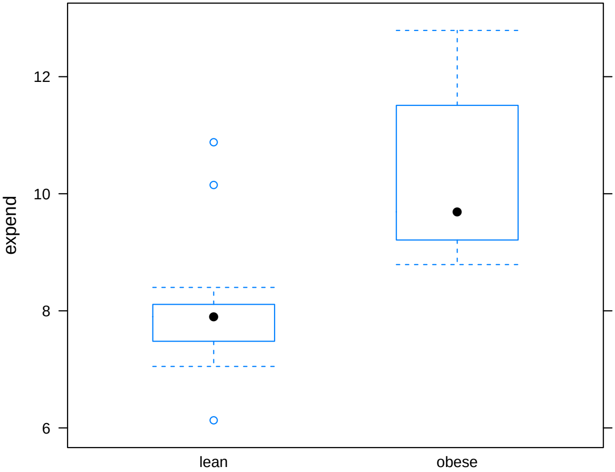
```
 -3.459167 -1.004081
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

Practice

- However, we missed the important step of checking whether the variance is the same for the two groups.
- We can do so easily with boxplots

```
> library(lattice)
> print(bwplot(expend ~ stature,
+     data = energy))
```

Practice



Practice

- So, what is our conclusion in this case at a 5% significance level?

Practice

```
> t.test(expend ~ stature, var.equal = TRUE)

Two Sample t-test


data:  expend by stature
t = -3.9456, df = 20, p-value =
0.000799
alternative hypothesis: true difference in means is not
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778

> t.test(expend ~ stature, var.equal = FALSE)
```

Practice

```
Welch Two Sample t-test

data:  expend by stature
t = -3.8555, df = 15.919, p-value =
0.001411
alternative hypothesis: true difference in means is not
95 percent confidence interval:
 -3.459167 -1.004081
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

Testing equality of variance

- To properly test whether the variance of the two group is equal, we use the function var.test:

  > var.test(expend ~ stature)

```
F test to compare two variances

data:  expend by stature
F = 0.7844, num df = 12, denom df =
8, p-value = 0.6797
alternative hypothesis: true ratio of variances is not
95 percent confidence interval:
 0.1867876 2.7547991
sample estimates:
```

Testing equality of variance

```
ratio of variances
          0.784446
```

- It's actually a *F* (Fisher) test
- In this case, the samples are small so it's also important to guide our decision by the CI.

Basic idea

- This case of the t test is useful when you take measurements on the same group two times. Meaning that there is no independence between the two groups.

Lets jump right into it

- With the *intake* data set, how can you observe visually the relationship between the two measurements?

```
> library(ISwR)
> attach(intake)
> intake
```

```
   pre post
1  5260 3910
2  5470 4220
3  5640 3885
4  6180 5160
5  6390 5645
6  6515 4680
7  6805 5265
```

Lets jump right into it

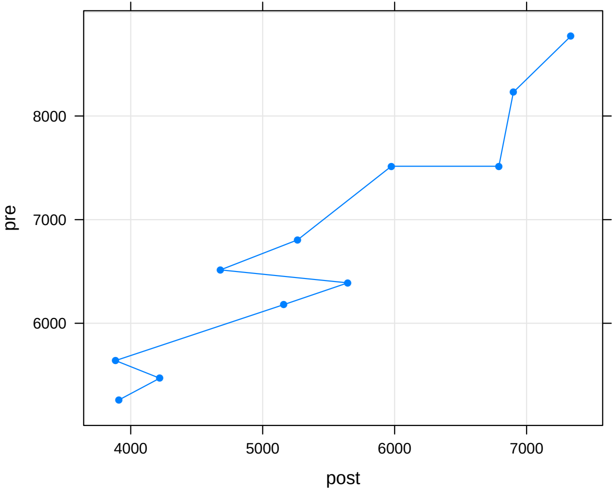
```
8   7515 5975
9   7515 6790
10  8230 6900
11  8770 7335
```

- It's data from the same 11 women that are measured twice for their daily intake.

A scatterplot works just fine

```
> print(xyplot(pre ~ post, data = intake,
+     type = c("o", "g"), pch = 16))
```

A scatterplot works just fine



Paired t test

- So, what do we conclude with a significance level of 5%?

Paired t test

- So, what do we conclude with a significance level of 5%?

  > t.test(pre, post)

```
Welch Two Sample t-test

data:  pre and post
t = 2.6242, df = 19.92, p-value =
0.01629
alternative hypothesis: true difference in means is not
95 percent confidence interval:
  270.5633 2370.3458
sample estimates:
mean of x mean of y
 6753.636  5433.182
```

Paired t test

```
> t.test(pre, post, paired = TRUE)

Paired t-test


data:  pre and post
t = 11.9414, df = 10, p-value =
3.059e-07
alternative hypothesis: true difference in means is not
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
               1320.455
```

htest object

- Note that we can save the result in an object and extract the information later on:

```
> res <- t.test(pre, post, paired = TRUE)
> class(res)

[1] "htest"

> names(res)

[1] "statistic"   "parameter"
[3] "p.value"     "conf.int"
[5] "estimate"    "null.value"
[7] "alternative" "method"
[9] "data.name"

> res$p.value
```

htest object

```
[1] 3.059021e-07
```

- This will be true for all hypothesis testing functions.

So. . .

- How do you find more functions for doing hypothesis testing?

So...

- Simply use apropos!!

  > *apropos("test")*

```
 [1] ".valueClassTest"
 [2] "ansari.test"
 [3] "bartlett.test"
 [4] "binom.test"
 [5] "Box.test"
 [6] "chisq.test"
 [7] "cor.test"
 [8] "file_test"
 [9] "fisher.test"
[10] "fligner.test"
[11] "friedman.test"
```

So...

```
[12] "kruskal.test"
[13] "ks.test"
[14] "mantelhaen.test"
[15] "mauchley.test"
[16] "mauchly.test"
[17] "mcnemar.test"
[18] "mood.test"
[19] "oneway.test"
[20] "pairwise.prop.test"
[21] "pairwise.t.test"
[22] "pairwise.wilcox.test"
[23] "poisson.test"
[24] "power.anova.test"
[25] "power.prop.test"
```

So. . .

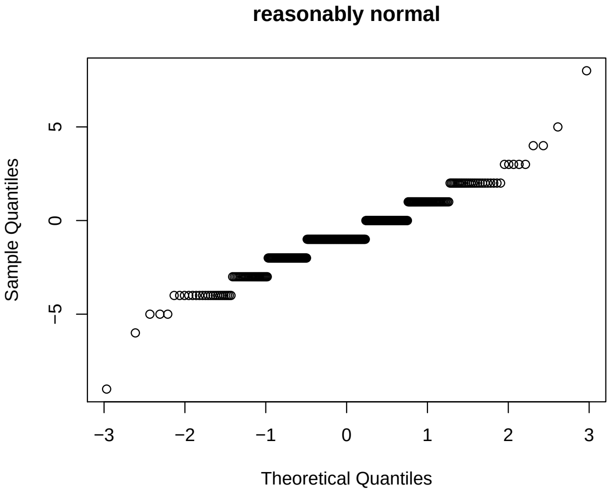
```
[26] "power.t.test"
[27] "PP.test"
[28] "prop.test"
[29] "prop.trend.test"
[30] "quade.test"
[31] "shapiro.test"
[32] "t.test"
[33] "testInheritedMethods"
[34] "testPlatformEquivalence"
[35] "testVirtual"
[36] "var.test"
[37] "wilcox.test"
```

Practice I

Do the values from the `react` data set look reasonably normally distributed? Does the mean differ significantly from zero according to a t test?

```
> qqnorm(react, main = "reasonably normal")
```

Practice I


reasonably normal
Sample Quantiles
Theoretical Quantiles
5
0
−5
−3
−2
−1
0
1
2
3

Practice I

```
> t.test(react)

One Sample t-test

data:  react
t = -7.7512, df = 333, p-value =
1.115e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.9985214 -0.5942930
sample estimates:
 mean of x
-0.7964072
```

Practice I

```
> t.test(react)$p.value < 0.05
[1] TRUE
```

Practice II

In the data set `vitcap`, use a *t* test to compare the vital capacity for the two groups. Calculate a 99% CI for the difference. The result of this comparison may be misleading. Why?

```
> var.test(vital.capacity ~ group,
+     data = vitcap)

F test to compare two variances

data:  vital.capacity by group
F = 2.3105, num df = 11, denom df =
11, p-value = 0.1806
alternative hypothesis: true ratio of variances is not equa
95 percent confidence interval:
 0.6651437 8.0260128
```

Practice II

```
sample estimates:
ratio of variances
          2.310509

> t.test(vital.capacity ~ group,
+     conf = 0.99, data = vitcap)

Welch Two Sample t-test


data:  vital.capacity by group
t = -2.9228, df = 19.019, p-value =
0.008724
alternative hypothesis: true difference in means is not equ
99 percent confidence interval:
 -2.06447665 -0.02219002
```

Practice II

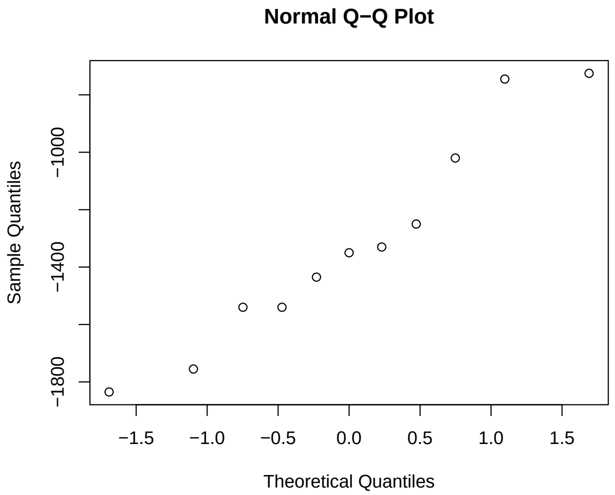
```
sample estimates:
mean in group 1 mean in group 3
       3.949167        4.992500
```

Practice III

Perfom graphical checks on the assumptions for a paired t test in the `intake` data set.

```
> qqnorm(intake$post - intake$pre)
```

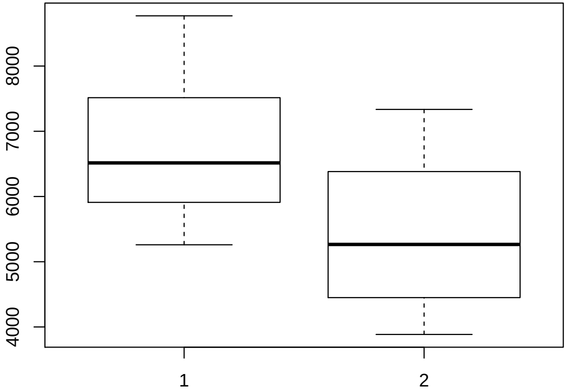
Practice III



Practice III

```
> boxplot(intake$pre, intake$post)
```

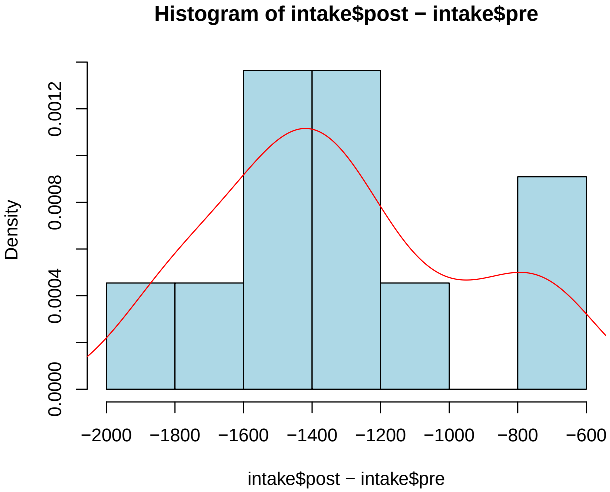
Practice III



Practice III

```
> hist(intake$post - intake$pre,
+      prob = TRUE, col = "light blue")
> lines(density(intake$post - intake$pre),
+      col = "red")
```

Practice III



Practice IV

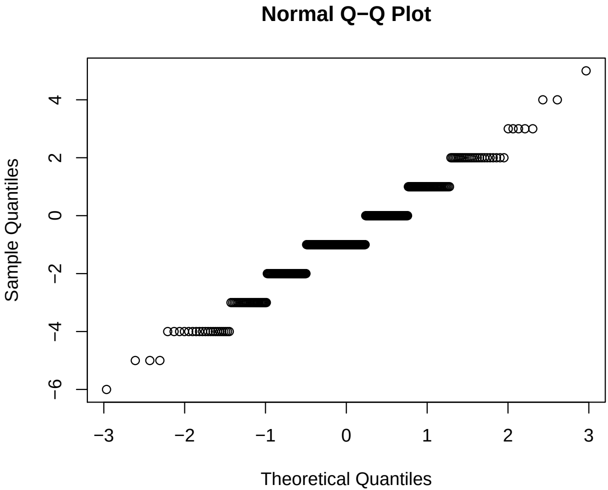
The function `shapiro.test` computes a test of normality based on the degree of linearity of the Q-Q plot. Apply it to the `react` data. Does it help to remove outliers?

```
> shapiro.test(react)

Shapiro-Wilk normality test


data:  react
W = 0.957, p-value = 2.512e-08

> shapiro.test(react[-c(1, 334)])

Shapiro-Wilk normality test


data:  react[-c(1, 334)]
W = 0.9687, p-value = 1.376e-06
```

Practice IV

```
> qqnorm(react[-c(1, 334)])
```

Practice IV


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3

Practice V

The crossover trial in `ashina` can be analysed for a drug effect in a simple way (how?) if you ignore a potential period effect. However, you can do better. Hint: Consider the intra-individual differences; if there were *only* a period effect present, how should the difference behave in the two groups? Compare the results of the simple method and the improved method.

```
> attach(ashina)
> t.test(vas.active, vas.plac, paired = TRUE)
```

Practice V

```
Paired t-test


data:  vas.active and vas.plac
t = -3.2269, df = 15, p-value =
0.005644
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of the differences
                -42.875

> t.test((vas.active - vas.plac)[grp ==
+     1], (vas.plac - vas.active)[grp ==
+     2])
```

Practice V

```
Welch Two Sample t-test

data:  (vas.active - vas.plac)[grp == 1] and (vas.plac - va
t = -3.2517, df = 13.97, p-value =
0.005807
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -130.56481  -26.76853
sample estimates:
mean of x mean of y
-53.50000  25.16667
```

Practice VI

Perform 10 one-sample *t* tests on simulated normally distributed data sets of 25 observations each. Repeat the experiment, but instead simulate samples from a different distribution; try the *t* distribution with 2 degrees of freedom and the exponential distribution (in the latter case, test for the mean being equal to 1). Can you find a way to automate this so that you can have a larger number (say 10k) of replications?

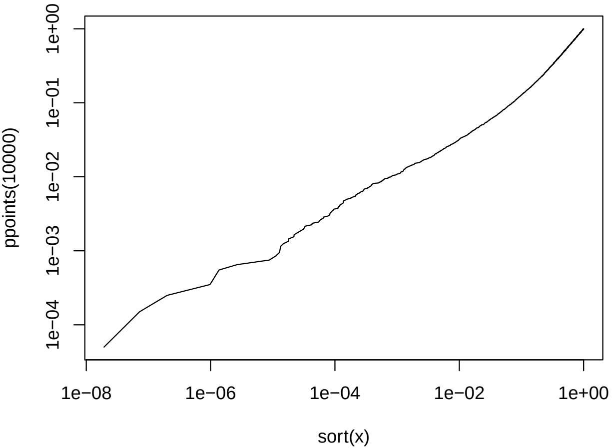
```
> t.test(rnorm(25))$p.value

[1] 0.6118598

> t.test(rt(25, df = 2))$p.value

[1] 0.7829499

> t.test(rexp(25), mu = 1)$p.value
```

Practice VI

```
[1] 0.5847691
```

```
> x <- replicate(10000, t.test(rexp(25),
+     mu = 1)$p.value)
> qqplot(sort(x), ppoints(10000),
+     type = "l", log = "xy")
```

Practice VI


1e+00
1e−01
1e−02
1e−03
1e−04
ppoints(10000)
1e−08
1e−06
1e−04
1e−02
1e+00
sort(x)

Practice VII

Calculate manually the equivalent to the one sample t test for the `daily` vector:

```
> t.test(daily, mu = 7725)$p.value

[1] 0.01984965

> tvalue <- tvalue <- (mean(daily) -
+     7725)/(sd(daily)/sqrt(length(daily)))
> pt(tvalue, df = length(daily) -
+     1) * 2

[1] 0.01984965
```

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

other attached packages:
[1] lattice_0.19-13 ISwR_2.0-5
```

Session Information

```
loaded via a namespace (and not attached):
[1] grid_2.12.0
```