Understanding p-values and assumptions behind statistical tests

Cyril Matthey-Doret

December 8, 2016

Introduction

The goal of this report is to assess the sensitivity of statistical tests to violations of their assumptions.

Example 1: Dependent samples and two-sample t-test

The two-sample t-test allows to test if the means of two normally distributed samples are significantly different, assuming they are independent. Here, I investigate the sensitivity of the two-sample t-test to violation of the assumption of independence between two datasets (Figure 1) using simulated data. For each simulation, I generate 2 normally distributed datasets with the same mean and standard deviation. When the two datasets are independent, we obtain uniformely distributed p-values between 0 and 1, as should be the case when the distributions are similar. I will look at how dependence between similar samples affects the distribution of p-values.

To observe the effect of dependence between datasets (or paired data), I generate pairs of datasets with different degrees of correlation using Cholesky decomposition[1]. For every simulation, a new pair of datasets is generated and a two-sample t-test is performed. The p-values are stored and the proportion of p-values below 0.05 (false positives) is measured at different correlation coefficients.

```r
library(ggplot2); library(gridExtra);library(parallel)
n.cores <- detectCores() # Probing number of cores available

n <- 100000  # Number of simulations
dep_sim <- function(r.coef){
  # Function performing simulation
  X <- rnorm(mean=0, 100)  # First dataset (X)
  Y <- rnorm(mean=0, 100)  # Second dataset (Y)
  Z <- r.coef*X+sqrt(1-r.coef^2)*Y  # Making Y correlated to X
  return(c(t.test(X,Z,paired=F)$p.val,t.test(X,Z,paired=T)$p.val))  # Extracting p-values
}
rep_sim <- function(r){
  # replicates simulation and extracts proportion of p<0.05
  Rep <- replicate(n,dep_sim(r))  # Replicating simulation n times
  out <- list()
  out[["NP"]] <- sum(Rep[1,]<0.05)/length(Rep[1,])
  # p<0.05 with two-samples t-test
  out[["P"]] <- sum(Rep[2,]<0.05)/length(Rep[2,])
  # p<0.05 with paired two-samples t-test
  return(out)
}
```

[1] Dependency of variable Y on variable X is induced by applying $Y_d = r * X + \sqrt{1 - r^2} * Y$ where r is the desired correlation coefficient and Y_d is the dependent Y variable.

```r
tmp <- mclapply(round(seq(-1,1,0.05),2), rep_sim, mc.cores = n.cores, mc.cleanup = TRUE)
# parallelizing simulation
merged <- unlist(tmp)
prop_p <- unname(merged[names(merged)=="P"])
# Proportions of false positives for paired t-test
prop_np <- unname(merged[names(merged)=="NP"])
# Proportions of false positives for non-paired t-test
```

When using the two-sample t-test, the proportion of false positive is dependent on the coefficient of correlation (Figure 2A) between the 2 samples. Negative correlation increases the proportion of false positives while positive correlation decreases it. Therefore, if the paired datasets are positively correlated, the test will be biased towards conservative results, whereas if they are negatively correlated, it will be biased in a liberal way. In the simulations I ran here, we can see false positive rates up to 16.661% when samples are negatively correlated, and down to 0% when samples are positively correlated.

This strong bias induced by correlated datasets highlights the importance of knowing if the data is paired (e.g. comparing measures taken twice on the same individuals at different times, or pairs of individuals from the same families) and using the appropriate test if this is the case. Here, we can see that the paired-two sample t-test works well, no matter the strength of correlation between samples (Figure 2B) and should be used whenever datasets are not independent.

Example 2: Different variances across groups in one-way ANOVA

One-way ANOVA is used to compare the means of three or more samples. It tests the null hypothesis that all samples distributions have the same means. One of the assumptions of ANOVA is that the different samples have equal variance. Here, I test the sensitivity of ANOVA to unequal variance across groups by comparing 3 normally distributed samples with equal means but different variances. If the test is not sensitive to violations of this assumption, the p-values should remain uniformely distributed between 0 and 1. Therefore, under an α threshold of 0.05, there should be 5% of false positives.

I test this sensitivity by varying the ratio of standard deviations between groups (Figure 3) and observing the impact on the proportion of false positives. At each ratio of variances, I run 10^{5} simulations and compute the proportion of false positives.

```r
simanov <- function(rt){  # Function performing the simulation
  df_anova <- data.frame(fact=c(rep("A",100),rep("B",100),rep("C",100)),
                         val=c(rnorm(n=100, mean=0,sd=1),
                               rnorm(n=100, mean=0,sd=1*rt),
                               rnorm(n=100, mean=0,sd=1*rt^2)))
  # Generating data with 3 groups and a continuous numeric response variable
    panov <- summary(aov(df_anova$val ~ df_anova$fact))[[1]][1,5]
    # Extracting p-values from the summary
  return(panov)
}
repanov <- function(rt){
  # replicates simulation and extracts proportion of p<0.05
  Rep <- replicate(n,simanov(rt))  # Replicating simulation n times
  out <- sum(Rep<0.05)/length(Rep)  # Proportion of false positives
  return(out)
}
tmp <- mclapply(1:10, repanov, mc.cores = n.cores, mc.cleanup = TRUE)
# parallelizing the simulation
propanov <- unlist(tmp)
```

Unequal variances between groups increase the false positive rate of the one-way ANOVA to up to 8.519% in the range of tested values. It appears that this difference of 3.465% is quite weak, considering the standard deviations differed up to a ratio of 100 between groups, which is rather extreme compared to what be expected in most real-world situations. In summary, the one-way ANOVA seems rather robust to violations of variance homogeneity across groups and as long as the difference is not dramatic, the test could still be applied without risking major bias in the results.

Figures:

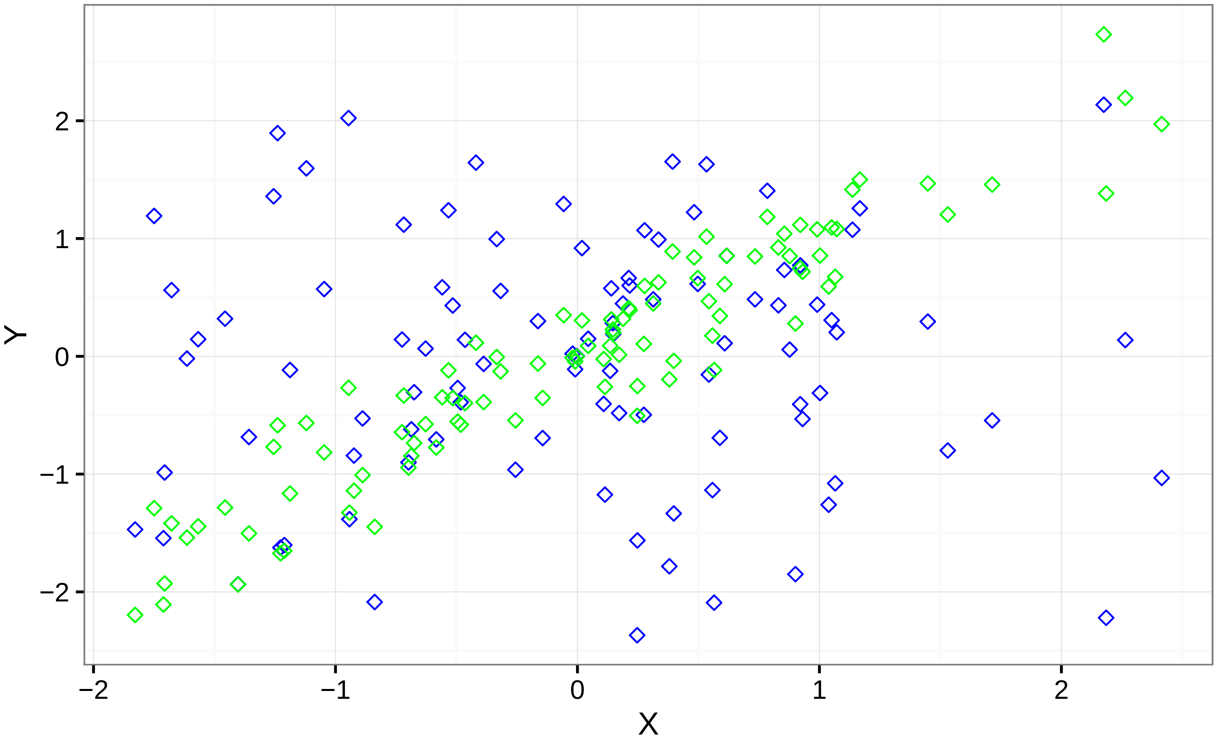
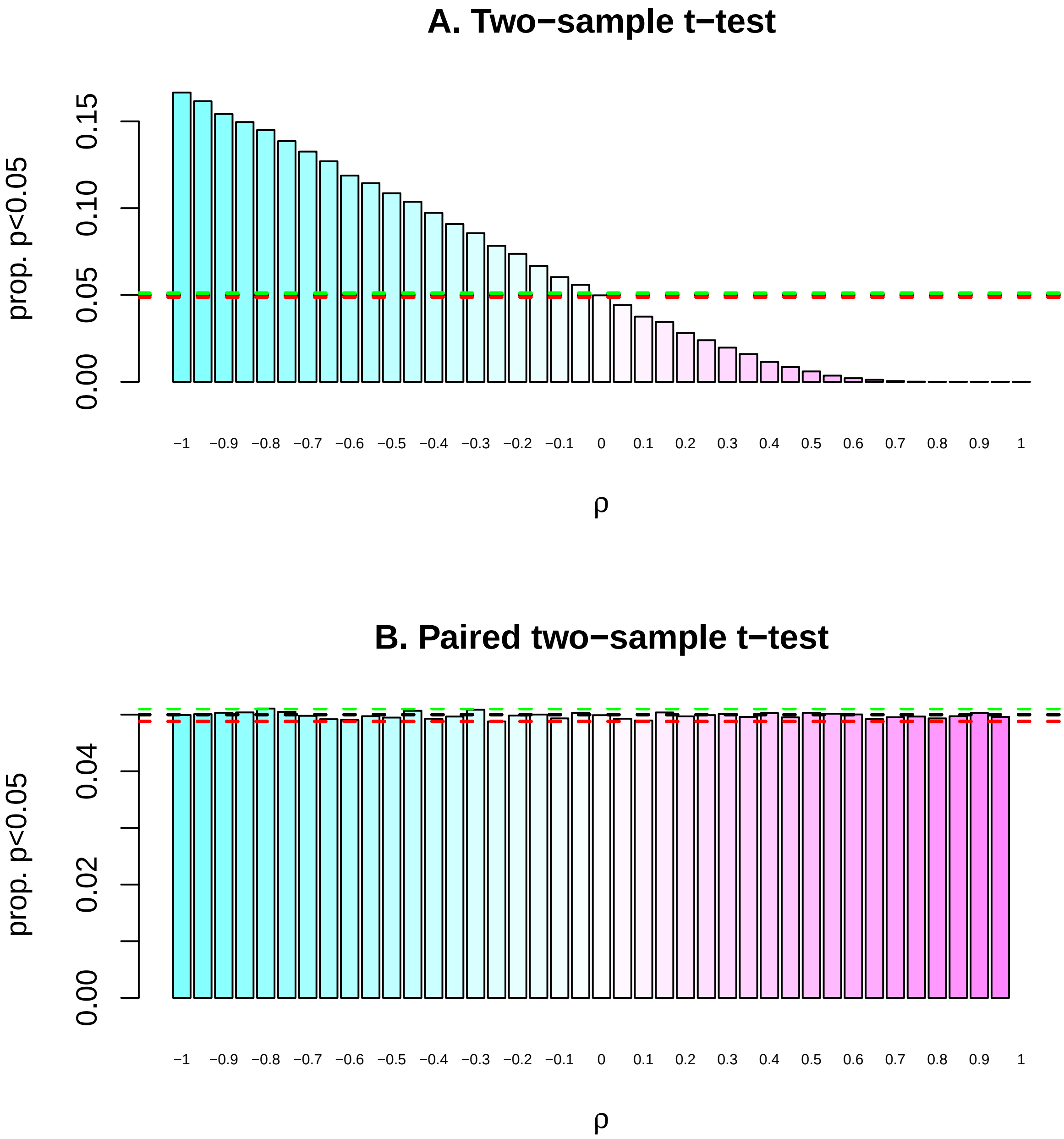


Figure 1: Illustration of two simulated datasets X and Y with or without dependence of Y on X. Blue: X and Y are independent with a correlation coefficient of ρ=0. Green: X and Y are highly dependent with ρ=0.95. In both cases, X and Y have equal means and variances.


A. Two−sample t−test

B. Paired two−sample t−test

Figure 2: Proportion of false positives across 10,000 simulations at different correlation coefficients between datasets. Proportion of p-values below 0.05 (False positives) when comparing two samples with equal means and standard deviations at different degrees of correlation using A. two-sample t-tests and B. paired two-sample t-tests. Green and red dashed horizontal lines represent the maximum and minimum proportions of false positives obtained with the paired two-sample t-test to better appreciate the scale. Black dashed horizontal line represent the 0.05 α threshold.

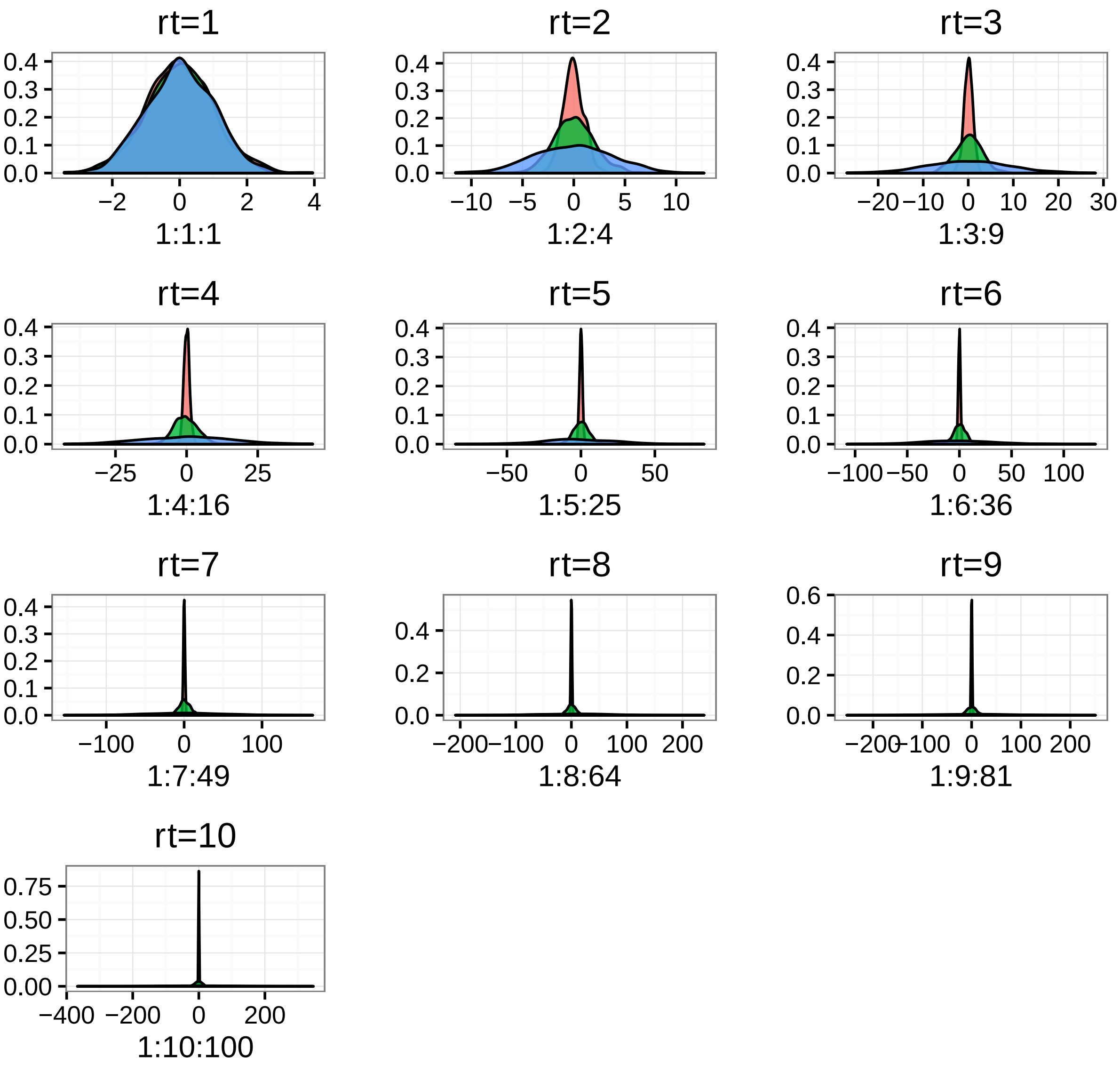


Figure 3: Illustration of the method used to generate different variances between groups. The factor rt is set to 10 different levels and dictates the standard deviation of each group. Standard deviations of groups A (blue), B (green) and C (red) are set respectively to $\sigma_A = 1$, $\sigma_B = 1 * rt$ and $\sigma_C = 1 * rt^2$. For each value of rt, a plot is generated and the corresponding standard deviations of each group are written below the plots in the format $\sigma_A : \sigma_B : \sigma_C$

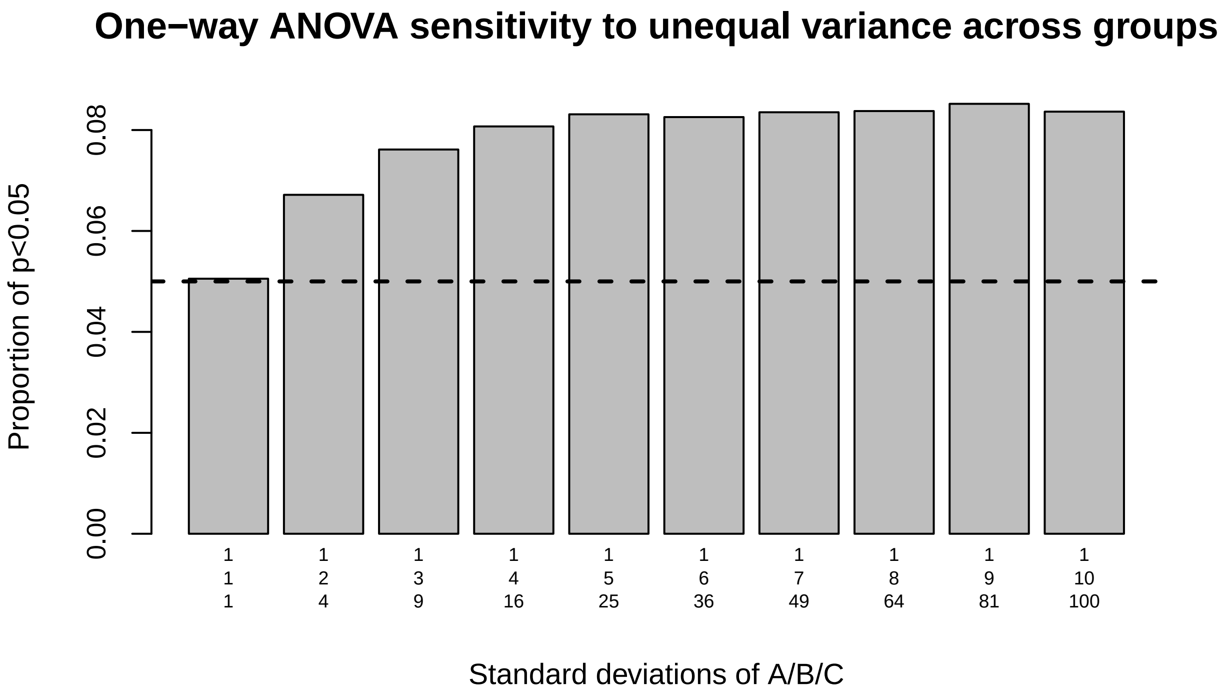


Figure 4: False positive rates (p<0.05) of one-way ANOVA when comparing 3 normally distributed groups of equal means with increasingly different standard deviations. Standard deviations of the three groups are displayed on the X axis. The black horizontal dashed line represents the 0.05 threshold.