Lecture 17

Ciprian M Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

November 14, 2013

p-value : include extreme ($X >= x_0$) / ($X <= x_0$)

Table of contents

Outline

1. Paired difference hypothesis tests
2. Independent group differences hypothesis tests

Hypothesis testing for comparing two means

- When comparing groups, first determine whether observations are paired or not
- When dealing a single set of paired data, one strategy is to take the difference between the paired observation and do a one-sample t test of $H_0 : \mu_d = 0$ versus $H_a : \mu_d \neq 0$ (or one of the other two alternatives)
- Test statistic is

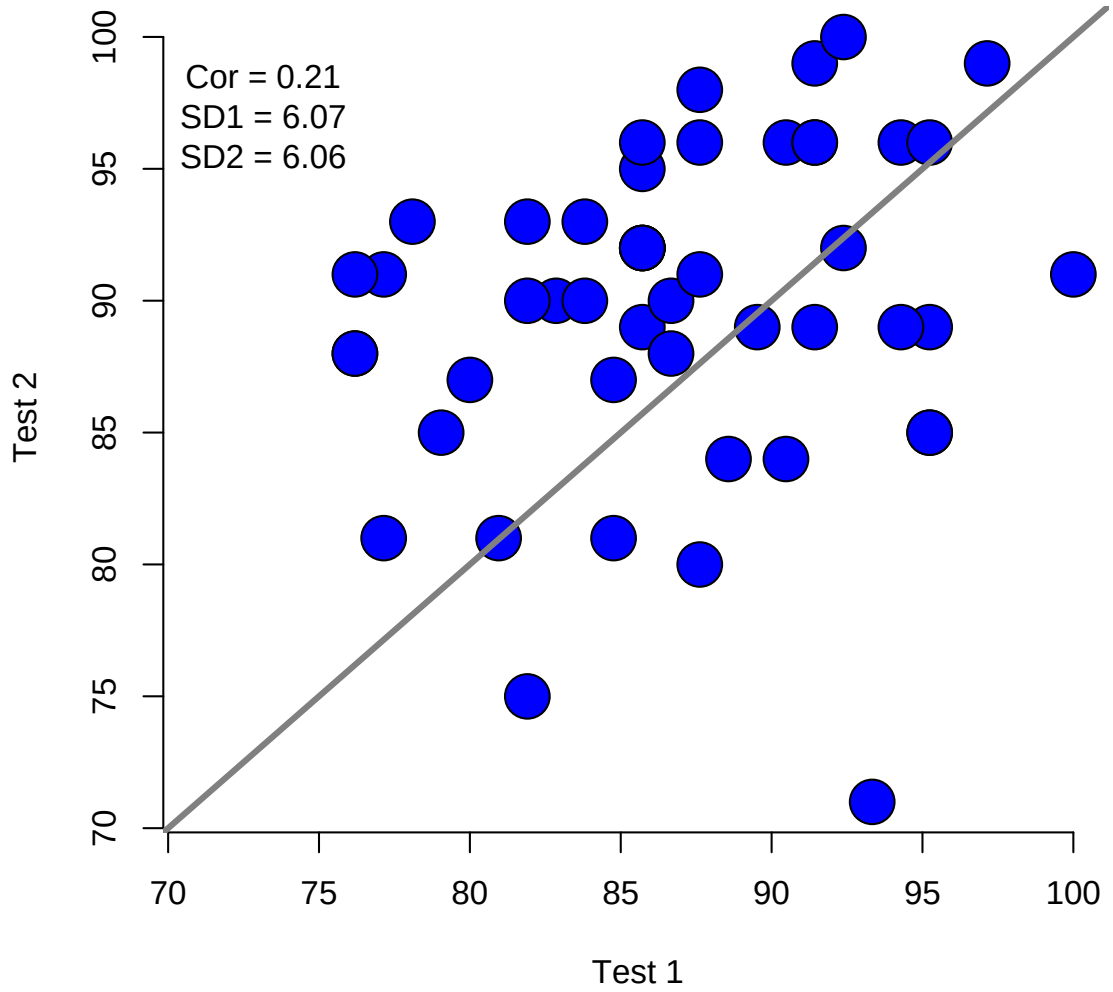
$$\frac{\bar{X}_d - \mu_{d0}}{S_d/\sqrt{n_d}}$$

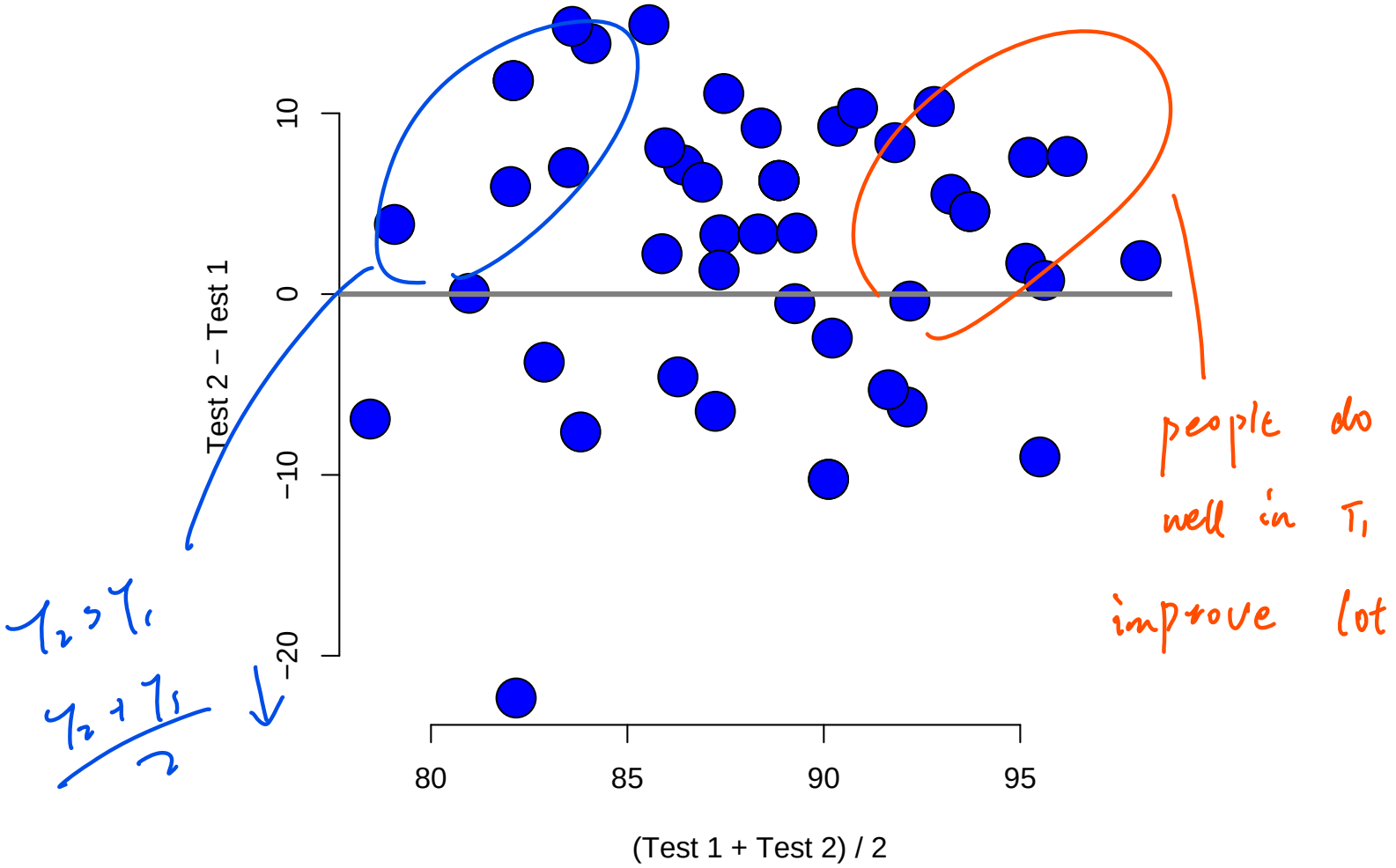
where μ_{d0} is the value under the null hypothesis (typically 0); compare this statistic to a t_{n_d-1} or z statistic

Example

- Consider Exam 1 and Exam 2 grades from a previous class
- Is there any evidence that the second exam was easier or harder than the first?
- The same students took both exams with none dropping out in-between
- Summaries for the two exams

```
> summary(test1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  76.19   82.86   86.67   86.94   91.43  100.00
> summary(test2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   87.00   90.00   89.82   93.00  100.00
```


Cor = 0.21
SD1 = 6.07
SD2 = 6.06
Test 2
Test 1
70
75
80
85
90
95
100


Test 2 – Test 1
10
0
−10
−20
80
85
90
95
(Test 1 + Test 2) / 2
people do well in T1
improve lot
T2 > T1
(T2 + T1) / 2

R Code

```
diff <- test2 - test1
n <- sum(!is.na(diff)) #49
mean(diff) #2.88
sd(diff) #7.61
testStat <- sqrt(n) * mean(diff) / sd(diff) #2.65
# below works out to be 0.01
2 * pt(abs(testStat), n -1, lower.tail = FALSE)
##uses the R function
t.test(diff)
```

Discussion of matched data

- Also to consider, "are ratios more relevant than pair-wise differences?"; if so, try doing the test on the log-observations
- When considering matched pairs data, you often want to plot the first observation by the second
- A more efficient plot displays the average of the observations by the difference or doing this on the log scale
- The previous plot is called a "mean/difference" plot, invented by Tukey (sometimes it is called a "Bland/Altman" plot after researchers who effectively described and used the method for considering measurement error)

Regression to mediocrity

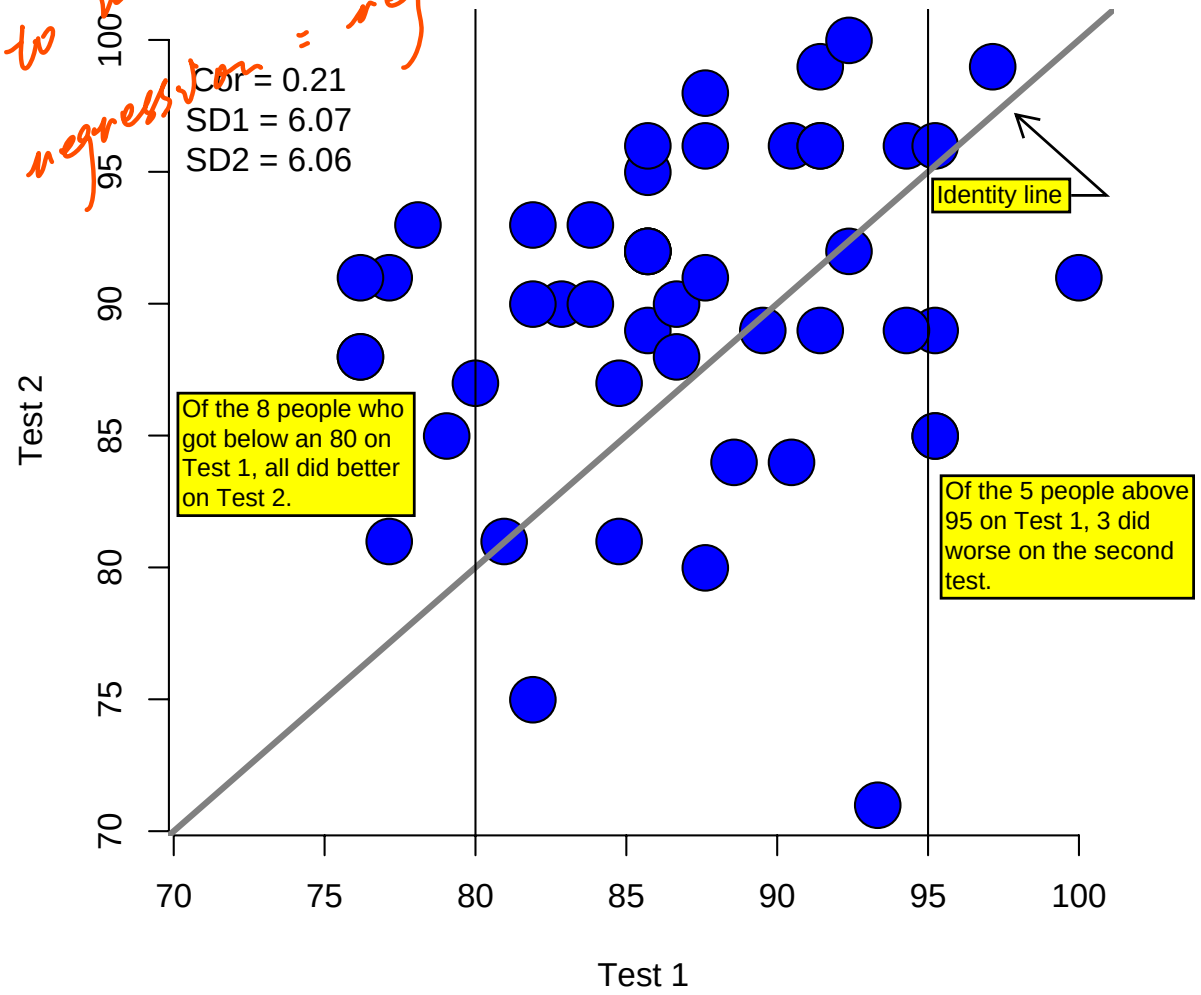
if 1st extreme 2nd → mean

if 2nd extreme 1st → mean.

- Francis Galton was the first to recognize that for matched data, high initial observations tended to be associated with lower second observations and low initial observations tended to be associated with higher second observations
- Example: Sons of very tall fathers tend to be a little shorter (also fathers of very tall sons tend to be shorter)
- Example: Second exams for those who scored very high on a first exam tend to be a little lower

regression to mean = regression.

mean of regression = regression line



RTM continued

- To investigate more, we normalize both scales (so that their means are both 0 and standard deviations 1)
- If there was no regression to the mean, the data would scatter about an identity line
- The best fitting line goes through the average and has slope

$$\text{Cor}(\textit{Test}1, \textit{Test}2)\frac{\text{SD}(\textit{Test}2)}{\text{SD}(\textit{Test}1)}$$

and passes through the point

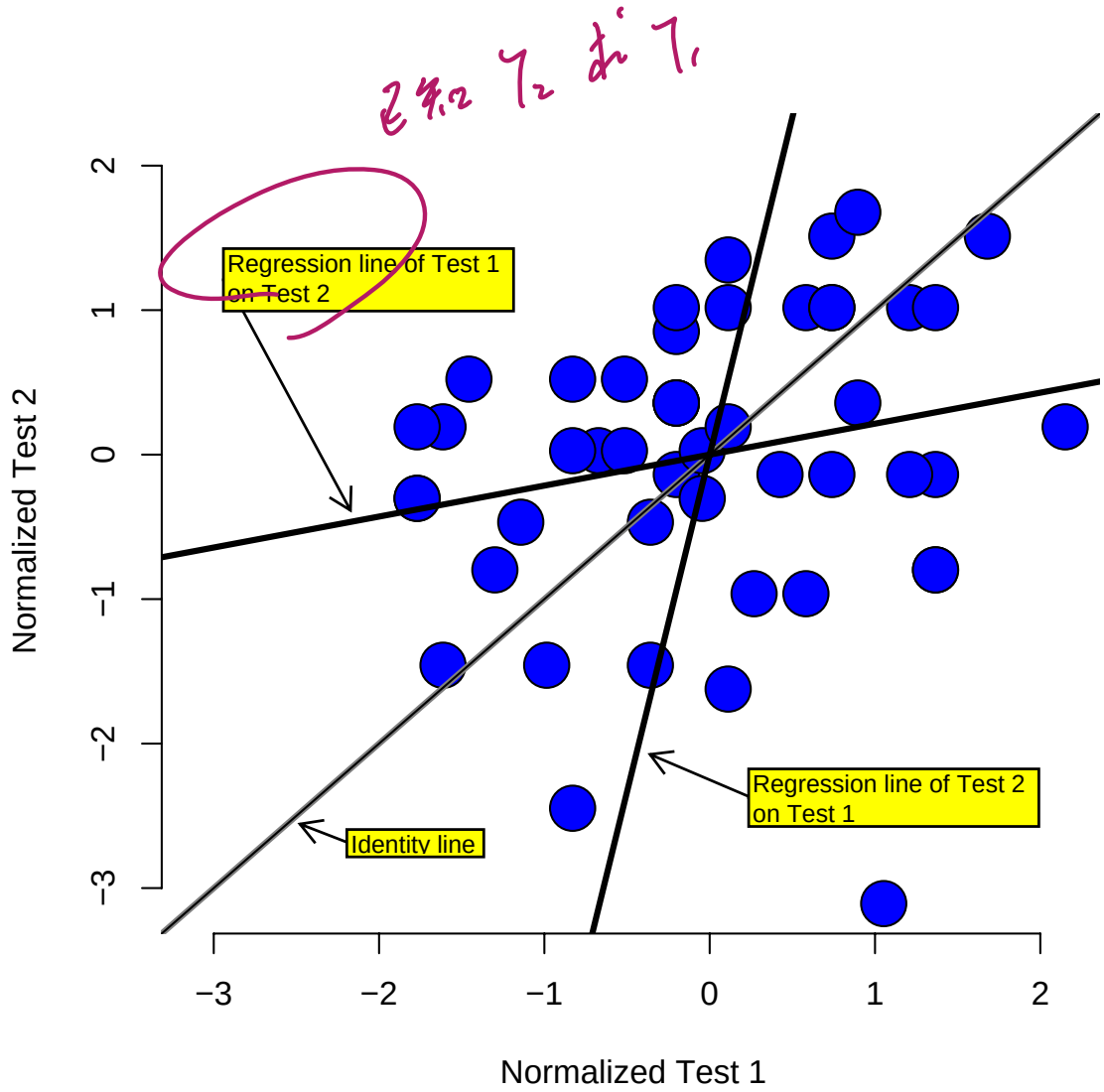
$$\{\text{mean}(\textit{Test}1), \text{mean}(\textit{Test}2)\}$$

.

- Because we normalized the data, our line passes through $(0, 0)$ and has slope $\text{Cor}(\textit{Test}1, \textit{Test}2)$ (normalizing doesn't impact the correlation)

RTM continued

- The best fitting "regression line" has slope $\mathrm{Cor}(\mathit{Test}1, \mathit{Test}2) < 1$
- This will be shrunk toward a horizontal line, telling us our expected normalized test score for Test 2 will be $\mathrm{Cor}(\mathit{Test}1, \mathit{Test}2)$ times the normalized Test 1 score
- This line appropriately adjusts for regression to the mean for Test 2 conditioning on Test 1; we could similarly do the same for Test 1 conditioning on Test 2; this line will have slope $\mathrm{Cor}(\mathit{Test}1, \mathit{Test}2)^{-1}$ if plot with Test 1 on the horizontal axis
- The latter line will be shrunk toward a vertical line; the identity line will fall between the two


Regression line of Test 1 on Test 2
Regression line of Test 2 on Test 1
Identity line
Normalized Test 2
Normalized Test 1
−3
−2
−1
0
1
2

Final comments

- An ideal examiner would have little difference between the identity line and the fitted regression line
- The more unrelated the two exam scores are, the more pronounced the regression to the mean
- If you watch sports you have to wonder how much discussion is over regression to the mean
- Athletes who perform the best will often perform worse the next year; is this regression to the mean or an actual decline in the athlete's ability?

Two independent groups

- The extension to two independent groups should come as no surprise
- $H_0 : \mu_1 = \mu_2$, versus $H_a : \mu_1 \neq \mu_2$ (or one of the other two alternatives)
- Assuming a comon error variance we have

$$\frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

which will follow a $t_{n_x+n_y-2}$ distribution under the null hypothesis and the usual assumptions

Continued

- If the assuming a common error variance is questionable

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

follows a standard normal distribution for large n_x and n_y. It follows an approximate Students T distribution if the X_i and Y_i are normally distributed

- The approximate degrees of freedom are

$$\frac{(S_x^2/n_x + S_y^2/n_y)^2}{(S_x^2/n_x)^2/(n_x - 1) + (S_y^2/n_y)^2/(n_y - 1)}$$

- Note the connection between hypothesis testing and confidence intervals still holds; for example, if zero is in your independent group T interval, you will fail to reject the independent group T test for equal means $\neq$ Accept
- Don't test for equality of means by comparing separately constructed intervals for the two groups and rejecting the null hypothesis if they do not overlap
- This procedure has lower power than just doing the right test
- Also, it leads to the potential abuse of comparing intervals for paired data

Example

- Suppose that instead of having repeated data on two consecutive exams, students were randomized to two teaching modules and took the same exam
- We might obtain data like the following

Group	N	Mean Exam	SD Exam
Module 1	50	86.9	6.07
Module 2	50	89.8	6.06

- Pooled standard deviation 6.065[1]
- Test stat

$$\frac{89.8 - 86.9}{6.065\sqrt{\frac{1}{50} + \frac{1}{50}}}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

(you do the rest)

[1] note this is not obtained by averaging the two standard deviations, it's obtained by averaging the variances then square rooting

- Look over the review notes on formal tests for equality of variances between two groups
- These tests, employing the F distribution, rely heavily on normality assumptions
- Alternatively, for moderate sample sizes, consider creating a bootstrap confidence interval for the ratio of the two variances
- For smaller sample sizes, consider basing decisions on exploratory data analysis

Final comments

- Suppose you have equal numbers of observations for two groups (say X and Y)
- If the data are truly matched, then the standard error of the difference is estimating

$$\sqrt{\frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n} - 2\frac{\mathrm{Cov}(X, Y)}{n}}$$

- If you ignore the matching, then the standard error of the difference is estimating

$$\sqrt{\frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n}}$$

- Since, generally, matched data is positively correlated, by ignoring the matching you are likely (violating assumptions and) inflating your standard error unnecessarily