UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: Ivo Dinov,

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Jacquelina Dacosta & Chris Barr

University of California, Los Angeles, Fall 2006

http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 9

Paired Data

Comparison of Paired Samples

- In chapter 7 we discussed how to compare two independent samples
- In chapter 9 we discuss how to compare two samples that are paired
  - In other words the two samples are not independent, Y_1 and Y_2 are linked in some way, usually by a direct relationship
  - For example, measure the weight of subjects before and after a six month diet

Paired data

- To study paired data we would like to examine the differences between each pair
  - $d = Y_1 - Y_2$
  - each Y_1, Y_2 pair will have a difference calculated
- With the paired t test we would like to concentrate our efforts on this difference data
  - we will be calculating the mean of the differences and the standard error of the differences

Paired data

- The mean of the differences is calculated just like the one sample mean we calculated in chapter 2

$$\bar{d} = \frac{\sum d}{n_d} = \bar{y}_1 - \bar{y}_2$$

  - it also happens to be equal to the difference in the sample means – this is similar to the t test
- This sample mean differences is an estimate of the population mean difference $\mu_d = \mu_1 - \mu_2$

Paired data

- Because we are focusing on the differences, we can use the same reasoning as we did for a single sample in chapter 6 to calculate the standard error
  - aka. the standard deviation of the sampling distribution of $\bar{d}$
- Recall: $SE = \frac{s}{\sqrt{n}}$
- Using similar logic: $SE_{\bar{d}} = \frac{s_d}{\sqrt{n_d}}$
  - where s_d is the standard deviation of the differences and n_d is the sample size of the differences

Paired data

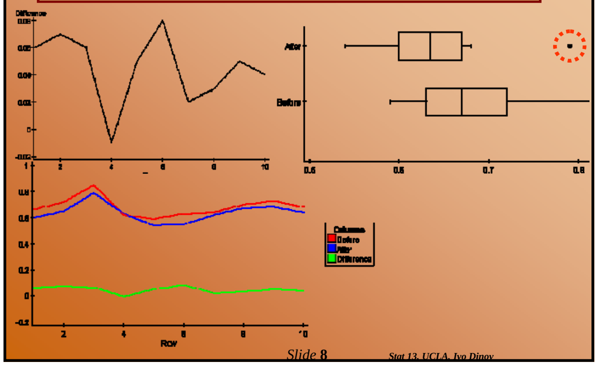
Example: Suppose we measure the thickness of plaque (mm) in the carotid artery of 10 randomly selected patients with mild atherosclerotic disease. Two measurements are taken, thickness before treatment with Vitamin E (baseline) and after two years of taking Vitamin E daily.

What makes this paired data rather than independent data?

Why would we want to use pairing in this example?

Subject	Before	After	Difference
1	0.66	0.60	0.06
2	0.72	0.65	0.07
3	0.85	0.79	0.06
4	0.62	0.63	-0.01
5	0.59	0.54	0.05
6	0.63	0.55	0.08
7	0.64	0.62	0.02
8	0.70	0.67	0.03
9	0.73	0.68	0.05
10	0.68	0.64	0.04
mean	0.682	0.637	0.045
sd	0.0742	0.0709	0.0264

Paired data



Paired data

Calculate the mean of the differences and the standard error for that estimate

$$\bar{d} = 0.045$$

$$s_d = 0.0264$$

$$SE_{\bar{d}} = \frac{s_d}{\sqrt{n_d}} = \frac{0.0264}{\sqrt{10}} = 0.00833$$

Paired CI for μ_d

- A 100(1 - α)% confidence interval for μ_d

$$\bar{d} \pm t(df)_{\alpha/2}(SE_{\bar{d}})$$

where df = n_d - 1

- Very similar to the one sample confidence interval we learned in section 6.3, but this time we are concentrating on a difference column rather than a single sample

Paired CI for μ_d

Example: Vitamin E (cont')

Calculate a 90% confidence interval for the true mean difference in plaque thickness before and after treatment with Vitamin E

$$\bar{d} \pm t(df)_{\alpha/2}(SE_{\bar{d}})$$
$$= 0.045 \pm t(9)_{0.05}(0.00833)$$
$$= 0.045 \pm (1.833)(0.00833)$$
$$= (0.0297, 0.0603)$$

Paired CI for μ_d

CONCLUSION: We are highly confident, at the 0.10 level, that the true mean difference in plaque thickness before and after treatment with Vitamin E is between 0.03 mm and 0.06 mm.

- Great, what does this really mean?
- Does the zero rule work on this one?

Paired t test

- Of course there is also a hypothesis test for paired data
- #1 Hypotheses:
  - Ho: $\mu_d = 0$
  - Ha: μ_d != 0 or Ha: $\mu_d < 0$ or Ha: $\mu_d > 0$
- #2 test statistic
  - Where df = $n_d - 1$
- #3 p-value and #4 conclusion similar idea to that of the independent t test

$$t_s = \frac{\bar{d} - 0}{SE_{\bar{d}}}$$

Paired t test

Example: Vitamin E (cont')

Do the data provide enough evidence to indicate that there is a difference in plaque before and after treatment with vitamin E for two years? Test using $\alpha = 0.10$

H_o: $\mu_d = 0$ (thickness in plaque is the same before and after treatment with Vitamin E)

H_a: μ_d != 0 (thickness in plaque after treatment is different than before treatment with Vitamin E)

df = 10 – 1 = 9

p < 2(0.0005) = 0.001, so we reject H_o.

$$t_s = \frac{0.045 - 0}{0.00833} = 5.402$$

Paired t test

CONCLUSION: These data show that the true mean thickness of plaque after two years of treatment with Vitamin E is statistically significantly different than before the treatment (p < 0.001).

In other words, vitamin E appears to be a effective in changing carotid artery plaque after treatment

- May have been better to conduct this as an upper-tailed test because we would hope that vitamin E will reduce clogging
  - however, researchers need to make this decision before analyzing data

Paired t test

Paired T-Test and CI: Before, After

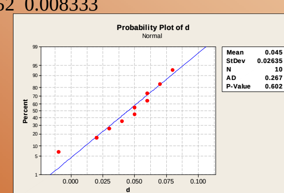
Paired T for Before - After

	N	Mean	StDev	SE Mean
Before	10	0.682000	0.074207	0.023466
After	10	0.637000	0.070875	0.022413
Difference	10	0.045000	0.026352	0.008333

90% CI for mean difference: (0.029724, 0.060276)

T-Test of mean difference = 0 (vs not = 0):

T-Value = 5.40 P-Value = 0.000



Results of Ignoring Pairing

- Suppose we accidentally analyzed the groups independently (like an independent t-test) rather than a paired test?
  - keep in mind this would be an incorrect way of analyzing the data
- How would this change our results?

Results of Ignoring Pairing

Example Vitamin E (con't)

Calculate the test statistic and p-value as if this were an independent t test

$$SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.0742^2}{10} + \frac{0.0709^2}{10}} = 0.0325$$

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{0.682 - 0.637}{0.0325} = 1.38$$

df = 17

2(0.05) < p < 2(0.1)

0.10 < p < 0.2

Fail To Reject H_o!

Results of Ignoring Pairing

- What happens to a CI?

  Calculate a 90% confidence interval for μ_1 - μ_2

$$\bar{y}_1 - \bar{y}_2 \pm t(df)_{\alpha/2}(SE_{\bar{y}_1-\bar{y}_2})$$
$$= (0.682 - 0.637) \pm t(17)_{0.05}(0.0325)$$
$$= 0.045 \pm (1.740)(0.0325)$$
$$= (-0.0116, 0.1016)$$

How does the significance of this interval compare to the paired 90% CI (0.03 mm and 0.06 mm)?

Why is this happening?

Is there anything better about the independent CI? Is it worth it in this situation?

Paired T-Test and CI: Before, After

```
       Paired T for Before - After
                N      Mean     StDev   SE Mean
Before         10  0.682000  0.074207  0.023466
After          10  0.637000  0.070875  0.022413
Difference     10  0.045000  0.026352  0.008333
90% CI for mean difference: (0.029724, 0.060276)
T-Test of mean difference = 0 (vs not = 0):
T-Value = 5.40  P-Value = 0.000
```

```
   Two Two-Sample T-Test and CI: Before, After
 Two-sample T for Before vs After
          N    Mean   StDev  SE Mean
 Before  10  0.6820  0.0742    0.023
 After   10  0.6370  0.0709    0.022
 Difference = mu (Before) - mu (After)
 Estimate for difference:  0.045000
 90% CI for difference:  (-0.011450, 0.101450)
 T-Test of difference = 0 (vs not =):
 T-Value = 1.39  P-Value = 0.183  DF = 17
```


Results of Ignoring Pairing

- Why would the SE be smaller for *correctly* paired data?
  - If we look at the within each sample at the data we notice variation from one subject to the next
  - This information gets incorporated into the SE for the independent t-test via s_1 and s_2
  - The original reason we paired was to try to control for some of this inter-subject variation
  - This inter-subject variation has no influence on the SE for the paired test because only the differences were used in the calculation.
- The price of pairing is smaller df.
  - However, this can be compensated with a smaller SE if we had paired correctly.

Conditions for the validity of the paired t test

- Conditions we must meet for the paired t test to be valid:
  - It must be reasonable to regard the differences as a random sample from some large population
  - The population distribution of the differences must be normally distributed.
    - The methods are approximately valid if the population is approximately normal or the sample size n_d is large.
  - These conditions are the same as the conditions we discussed in chapter 6.

Conditions for the validity of the paired t test

- How can we check:
  - check the study design to assure that the differences are independent (ie no hierarchical structure within the d's)
  - create normal probability plots to check normality of the differences
  - NOTE: p.355 summary of formulas

The Paired Design

- Ideally in the paired design the members of a pair are relatively similar to each other
- Common Paired Designs
  - Randomized block experiments with two units per block
  - Observational studies with individually matched controls
  - Repeated measurements on the same individual
  - Blocking by time – formed implicitly when replicate measurements are made at different times.
- IDEA of pairing: members of a pair are similar to each other with respect to extraneous variables

The Paired Design

Example: Vitamin E (cont')

- Same individual measurements made at different times before and after treatment (controls for within patient variation).

Example: Growing two types of bacteria cells in a *petri dish* replicated on 20 different days.

- These are measurements on 2 different bacteria at the same time (controls for time variation).

Purpose of Pairing

- Pairing is used to reduce bias and increase precision
  - By matching/blocking we can control variation due to extraneous variables.
- For example, if two groups are matched on age, then a comparison between the groups is free of any bias due to a difference in age distribution
- Pairing is a strategy of design, not analysis
  - Pairing needs to be carried out before the data are observed
  - It is not correct to use the observations to make pairs after the data has been collected

Paired vs. Unpaired

- If the observed variable Y is not related to factors used in pairing, the paired analysis may not be effective
  - For example, suppose we wanted to match subjects on race/ethnicity and then we compare how much ice cream men vs. women can consume in an hour
- The choice of pairing depends on practical considerations (feasibility, cost, etc…) and on precision considerations
  - If the variability between subjects is large, then pairing is preferable
  - If the experimental units are homogenous then use the independent t

The Sign Test

- The **sign test is a non-parametric version of the paired t test**
- We use the sign test when pairing is appropriate, but we can't meet the normality assumption required for the t test
- The sign test is not very sophisticated and therefore quite easy to understand
- Sign test is also based on differences

  $d = Y_1 - Y_2$

  The information used by the sign test from this difference is the sign of d (+ or -)

The Sign Test

- #1 Hypotheses:

  H_o: the distributions of the two groups is the same

  H_a: the distributions of the two groups is different

  or H_a: the distribution of group 1 is less than group 2

  or H_a: the distribution of group 1 is greater than group 2
- #2 Test Statistic B_s

The Sign Test - Method

- #2 Test Statistic B_s:
  1. Find the sign of the differences
  2. Calculate N_+ and N_-
  3. If H_a is non-directional, B_s is the larger of N_+ and N_-

     If H_a is directional, B_s is the N that jives with the direction of Ha:

     if H_a: $Y_1<Y_2$ then we expect a larger N_-,

     if H_a: $Y_1>Y_2$ then we expect a larger N_+.

NOTE: If we have a difference of zero it is not included in N_+ or N_-, therefore n_d needs to be adjusted

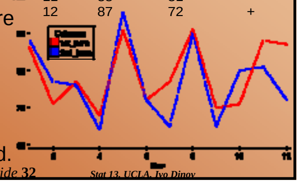
The Sign Test

- #3 p-value:
  - Table 7 p.684
  - Similar to the WMW
  - Use the number of pairs with “quality information”
  - http://www.socr.ucla.edu/htmls/SOCR_Analyses.html
- #4 Conclusion:
  - Similar to the Wilcoxon-Mann-Whitney Test
  - Do NOT mention any parameters!

The Sign Test

Example: 12 sets of *identical twins* are given psychological tests to determine whether the first born of the set tends to be more aggressive than the second born. Each twin is scored according to aggressiveness, a higher score indicates greater aggressiveness.

Set	1st born	2nd born	Sign of d
1	86	88	-
2	71	77	-
3	77	76	+
4	68	64	+
5	91	96	-
6	72	72	Drop
7	77	65	+
8	91	90	+
9	70	65	+
10	71	80	-
11	88	81	+
12	87	72	+



- Because of the natural pairing in a set of twins these data can be considered paired.

The Sign Test (cont’)

Do the data provide sufficient evidence to indicate that the first born of a set of twins is more aggressive than the second? Test using $\alpha = 0.05$.

H_o: The aggressiveness is the same for 1st born and 2nd born twins

H_a: The aggressiveness of the 1st born twin tends to be more than 2nd born.

NOTE: Directional Ha (we’re expecting higher scores for the 1st born twin), this means we predict that most of the differences will be positive

N_+ = number of positive = 7

N_- = number of negative = 4

n_d = number of pairs with useful info = 11

The Sign Test

$B_s = N_+ = 7$ (because of directional alternative)

$P > 0.10$, Fail to reject H_o

CONCLUSION: These data show that the aggressiveness of 1st born twins is not significantly greater than the 2nd born twins ($P > 0.10$).

$X \sim B(11, 0.5)$

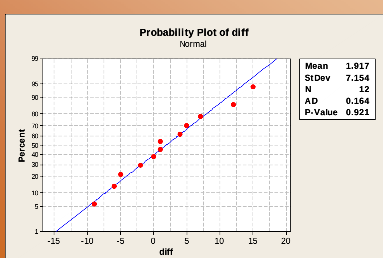
$P(X>=7)=0.2744140625$

http://socr.stat.ucla.edu/htmls/SOCR_Distributions.html (Binomial Distribution)

http://socr.stat.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm

The Sign Test

- Hold on did we actually need to carry out a sign test? What should we have checked first?



Practice

- Suppose H_a: one-tailed, $n_d = 11$
- And $B_s = 10$
- Find the appropriate p-value
  - $0.005 < p < 0.01$
  - Pick the smallest p-value for $B_s = 10$ and bracket
    - NOTE: Distribution for the sign test is discrete, so probabilities are somewhat smaller (similar to Wilcoxon-Mann-Whitney)

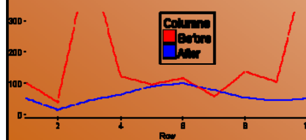
Applicability of the Sign Test

- Valid in any situation where d's are independent of each other
- Distribution-free, doesn't depend on population distribution of the d's
  - although if d's are normal the t-test is more powerful
- Can be used quickly and can be applied on data that do not permit a t-test

Applicability of the Sign Test

Example: 10 randomly selected rats were chosen to see if they could be trained to escape a maze. The rats were released and timed (sec.) before and after 2 weeks of training. Do the data provide evidence to suggest that the escape time of rats is different after 2 weeks of training?

Test using $\alpha = 0.05$.



Rat	Before	After	Sign of d
1	100	50	+
2	38	12	+
3	N	45	+
4	122	62	+
5	95	90	+
6	116	100	+
7	56	75	-
8	135	52	+
9	104	44	+
10	N	50	+

N denotes a rat that could not escape the maze.

Applicability of the Sign Test

- H_o: The escape times (sec.) of rats are the same before and after training.
- H_a: The escape times (sec.) of rats are different before and after training.

$N_+ = 9$; $N_- = 1$; $n_d = 10$

B_s = larger of N_+ or $N_- = 9$

$0.01 < p < 0.05$, reject H_o

X~Bin(10, 0.5)
P(X>=9)=0.0107421875
http://socr.stat.ucla.edu/Applets.dir/Normal_T_Chi2_F_Tables.htm

- CONCLUSION: These data show that the escape times (sec.) of rats before training are different from the escape times after training ($0.01 < p < 0.05$).

Further Considerations in Paired Experiments

- Many studies compare measurements before and after treatment
  - There can be difficulty because the effect of treatment could be confounded with other changes over time or outside variability
    - for example suppose we want to study a cholesterol lowering medication. Some patients may have a response because they are under study, not because of the medication.
  - We can protect against this by using randomized concurrent controls

Further Considerations in Paired Experiments

Example: A researcher conducts a study to examine the effect of a new **anti-smoking pill on smoking behavior**. Suppose he has collected data on 25 randomly selected smokers, 12 will receive treatment (a treatment pill once a day for three months) and 13 will receive a placebo (a mock pill once a day for three months). The researcher measures the number of cigarettes smoked per week before and after treatment, regardless of treatment group. Assume normality. The summary statistics are:

# cigs / week	n	$\bar{y}_{before}$	$\bar{y}_{after}$	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	163.92	152.50	11.42	1.10
Placebo	13	163.08	160.23	2.85	1.29

Further Considerations in Paired Experiments

Test to see if there is a difference in number of cigs smoked per week ***before and after the new treatment***, using $\alpha = 0.05$

H_o: $\mu_d = 0$

H_a: $\mu_d \,!= 0$

# cigs / week	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

$$t_s = \frac{11.42}{1.10} = 10.38$$

$df = n_d - 1 = 12 - 1 = 11$

$p < (0.0005)2 = 0.001$, reject H_o.

These data show that there is a statistically significant difference in the true mean number of cigs/week before and after treatment with the new drug

Further Considerations in Paired Experiments

- This result does not necessarily demonstrate the effectiveness of the new medication
  - Smoking less per week could be due to the fact that patients know they are being studied (i.e., difference statistically significantly different from zero)
  - All we can say is that he new medication appears to have a significant effect on smoking behavior

Further Considerations in Paired Experiments

Test to see if there is a difference in number of cigs smoked per week ***before and after in the placebo group***, using α = 0.05

H_o: $\mu_d = 0$

H_a: μ_d != 0

$$t_s = \frac{2.85}{1.29} = 2.21$$

# cigs / week	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

df = n_d – 1 = 13 – 1 = 12

(0.02)2 < p < (0.025)2

0.04 < p < 0.05 , reject H_o

These data show that there is a statistically significant difference in the true mean number of cigs/week before and after treatment with the a placebo

Further Considerations in Paired Experiments

- Patients who did not receive the new drug also experienced a statistically significant drop in the number of cigs smoked per week
  - This doesn't necessarily mean that the treatment was a failure because both groups had a significant decrease
  - We need to isolate the effect of therapy on the treatment group
  - Now the question becomes: was the drop in # of cigs/week significantly different between the medication and placebo groups?
  - How can we verify this?

Further Considerations in Paired Experiments

Test to see if there is the *difference in number of cigs smoked per week before and after treatment* was significant ***between the treatment and placebo groups***, using α = 0.05

$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{(1.10)^2 + (1.29)^2} = 1.695$$

Ho: $\mu_d = 0$

H_a: μ_d != 0

$$t_s = \frac{11.42 - 2.85}{1.695} = 5.06$$

# cigs / week	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

df = 22

p < (0.0005)2 = 0.001, reject H_o

These hypothesis tests provide strong evidence that the new anti-smoking medication is effective. If the experimental design had not included the placebo group, the last comparison could not have been made and we could not support the efficacy of the drug.

Reporting of Paired Data

- Common in publications to report the mean and standard deviation of the two groups being compared
  - In a paired situation it is important to report the mean of the differences as well as the standard deviation of the differences
  - Why?

Limitations of $\bar{d}$

- There are two major limitations of $\bar{d}$
  1. we are restricted to questions concerning $\bar{d}$
     - When some of the differences are positive and some are negative, the magnitude of does not reflect the “typical” magnitude of the differences.
     - Suppose we had the following differences: +40, -35, +20, -42, +61, -31.

Descriptive Statistics: data

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Max
data	6	0	2.17	17.9	43.9	-42.0	-36.8	-5.50	45.3	61.0

- What is the problem with this? Small average, but differences are large.
- What other statistic would help the reader recognize this issue?

Limitations of $\bar{d}$

2. limited to questions about aggregate differences
   - If treatment A is given to one group of subjects and treatment B is given to a second group of subjects, it is impossible to know how a person in group A would have responded to treatment B.

- Need to beware of these viewpoints and take time to look at the data, not just the summaries
- To verify accuracy we need to look at the individual measurements.
  - Accuracy implies that the d's are small

Inference for Proportions

- We have discussed two major methods of data analysis:
  - Confidence intervals: quantitative and categorical data
  - Hypothesis Testing: quantitative data
- In chapter 10, we will be discussing hypothesis tests for categorical variables
- RECALL: Categorical data
  - Gender (M or F)
  - Type of car (compact, mid-size, luxury, SUV, Truck)
- We typically summarize this type of data with proportions, or probabilities of the various categories