**Hypothesis Tests for Comparisons**

**Problem** A company is considering two procedures for measuring the mass of gizmos produced on an assembly line. It would like to know whether, on average, the two procedures produce the same results. It gathers data by randomly selecting 200 gizmos off the assembly line, measuring each of their masses using both procedures. How do we determine whether this data provides strong evidence that, on average, the two procedures produce different results?

**The parameter**  = average (procedure1 – procedure2) for all gizmos

**The hypotheses** H0 :  = 0

Ha :  ≠ 0

**The data** 200 differences in measurements d = x1 – x2

**The sample statistics** = average of the 200 measurement differences

sd = standard deviation of the 200 measurement differences

**The sampling distribution** If the distribution of all possible differences is normal and the null hypothesis is true, then t = has a t-distribution with 199 degrees of freedom. If the distribution is not perfectly normal, then t = has approximately a t-distribution.

**The p-value** 2Prob(T199 ≥ |t|)

The method illustrated above is called the **paired t-test**, since the data is paired.

Suppose = 0.24 gm and sd = 0.12 gm. Then t =  = 28.28.

> 1-pt(28.28,199)

[1] 0

**General Format for a paired t-test**

**Hypotheses**

H0 : = 

Ha : > or < or ≠

**Data** d1, d2, …, dn paired differences

**Test statistic** t =  has (approximately) a t-distribution with n-1 degrees

of freedom

**p-value** Prob(Tn-1 ≥ t) or Prob(Tn-1 ≤ t) or 2 Prob(Tn-1 ≥| t|) depending on

the alternative hypothesis.

**Example** On average, is a student’s MathSAT score higher than the student’s VerbalSAT score?

If = average difference MathSAT – VerbalSAT

H0 :

Ha :

The data-frame Students contains values for 200 students

> dbar<-mean(~(MathSAT-VerbalSAT),data = Students)

> sdd<-sd(~(MathSAT-VerbalSAT),data = Students)

> dbar

[1] 20.195

> sdd

[1] 72.19437

> t<-(dbar-0)/(sdd/sqrt(200))

> t

[1] 3.955993

> 1-pt(t,199)

[1] 5.298035e-05

**Letting R to do all of the work**

Suppose we want to do a paired t-test where the variables are in the column COL1 and COL2 of the data-frame FRAME. The difference of interest is COL1 – COL2.

t.test(COL1 , COL2 , data=FRAME, alternative = “ “ , paired = TRUE)

**Apply to the example above**:

> t.test(MathSAT,VerbalSAT, data = Students, alternative = "greater", paired = TRUE)

Paired t-test

data: MathSAT and VerbalSAT

t = 3.956, df = 199, p-value = 5.298e-05

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

11.75889 Inf

sample estimates:

mean of the differences

20.195

**A New Problem**

Suppose the company has two different assembly lines that produce gizmos. It would like to know whether, on average, the masses of the gizmos produced by the two lines are the same.

**The Parameters**

 = average mass for assembly line 1

 = average mass for assembly line 2

**The Hypotheses**

H0 :  = 

Ha :  ≠ 

**The Data**

Two independent samples of gizmos are selected, a sample of size n1 from line1 and a sample of size n2 from line2. The masses for line1 sample are and the masses for line2 are.

**The sample statistics**

, , , 

**The sampling distribution**

t =  has (approximately) a t-distribution with



Rule of thumb df = smaller of n1-1 and n2 -1.

**P-value**

Prob(T\* ≥ t) or Prob(T\* ≤ t) or 2 Prob(T\* ≥| t|) depending on

the alternative hypothesis.

Example

**Using R**

If the two samples are contained in the columns COL1 and COL2 of the data-frame FRAME

**t.test(COL1,COL2,data=FRAME,alternative =** )

**Example**  The data-frame Salaries contains annual salaries (in $1,000) for 50 individuals with gender 0 and 50 individuals with gender1. The data are contained in the columns Salary0 and Salary1. Use this data to test whether there is, on average, a difference between annual salaries of gender0 and gender1.

The parameters

The Hypotheses

**Letting R do all the work**

> t.test(Salary0,Salary1,data=Salaries,alternative = "two.sided")

Welch Two Sample t-test

data: Salary0 and Salary1

t = -2.658, df = 77.064, p-value = 0.009555

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-38.108788 -5.465212

sample estimates:

mean of x mean of y

41.631 63.418

**In general, the samples for a two-sample t-test do not contain the same number of values, so they cannot be used to create a data.frame. However, if the two samples are in the vectors S1 and S2, the R command t.test(S1,S2, alternative = “ “) will do the two sample t-test**.

**Example** The dataframe **Students** contains a column **VerbalSAT** that contains student VerbalSAT scores and the dataframe **StudentSurvey** contains a column **MathSAT** that contains student MathSAT scores. These two sets of data are independent of each other, so we must apply the two-sample t-test to test whether, on average, MathSAT scores are higher than VerbalSAT scores.

> verbal<-Students$VerbalSAT

> math<-StudentSurvey$MathSAT

> t.test(math,verbal,alternative="greater")

Welch Two Sample t-test

data: math and verbal

t = 3.3991, df = 412.14, p-value = 0.0003709

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

10.53267 Inf

sample estimates:

mean of x mean of y

609.4365 588.9850

> verbal<-StudentSurvey$VerbalSAT

> math<-Students$MathSAT

> t.test(math,verbal,alternative="greater")

Welch Two Sample t-test

data: math and verbal

t = 2.4076, df = 438.06, p-value = 0.008235

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

4.726989 Inf

sample estimates:

mean of x mean of y

609.1800 594.1906

**Exercises 20**

1. Is there a gender difference, on average, in the number of hours spent sleeping. The Excel Workbook (that I will send by email) contains data from 254 college students. For one gender, the values are in the column Sleep0, and for the other gender they are in the column Sleep1.
2. In R create the columns Sleep0 and Sleep1.
3. Define the parameter(s) and give the hypotheses.
4. Is this data paired or two (independent) sample?
5. Let R to do all of the work in computing the p-value. Include the R command and the output as your “solution”. Based on the p-value, what is your conclusion?
6. On average, do students do better on the final exam than they do on tests? The Excel Workbook Statsgrades (which I will send you by email) contains data from 50 students in a statistics class (not at Calvin.) The columns are Test, which contains the average test score, and Final, which contains the final exam score. All values are percents.
7. Create the data-frame Statsgrades in R whose columns are Test and Final. Use the R command head(Statsgrades) to produce the first 5 rows of the data-frame. Include the R command and the output as your “solution”.
8. Define the parameter(s) and state the hypotheses.
9. Is this data paired or two-sample?
10. Use R to do all of the work in computing the p-value. Include the R command and the output as your “solution”. Based on the p-value, what is your conclusion?