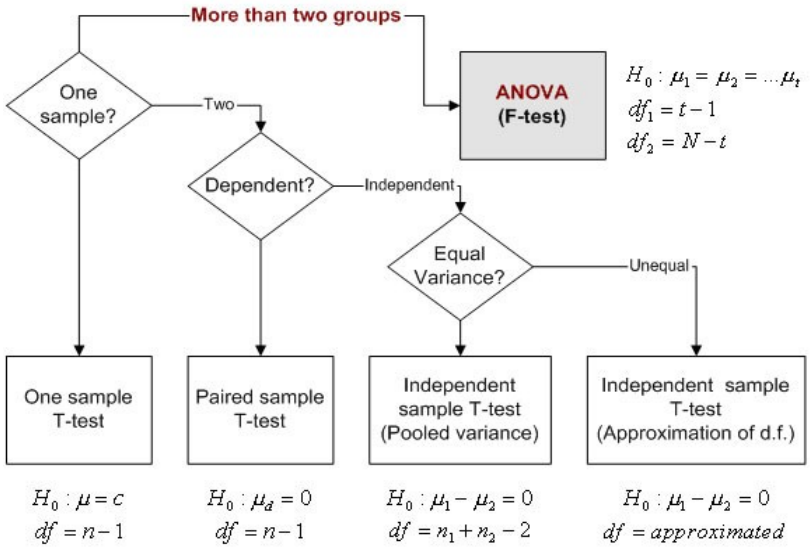
6. 집단간 평균비교

집단간 평균을 비교하는 것은 기본 방법이다. 따라서 비교할 변수는 평균을 계산할 수 있어야 하고, 의미있게 해석할 수 있어야 한다. 두 집단을 비교하는 것은 T-test 로, 두 집단 이상이라면 ANOVA 를 사용한다. 그림 6.1 은 T-test 각 유형과 ANOVA 를 비교하고 있다.

그림 6.1



6.1 T-test

T-test 는 두 집단의 평균이 같은지를 검정하는 방법이다. 따라서 의미있는 평균을 계산할 수 없는 분류변수를 비교할 수는 없다. T-test 는 비교되는 변수가 연속이어야 한다는 것 외에 다음 세가지 가정을 가지고 있다.

6.1.1 T-test 가 가정하는 것

The t-test assumes that samples are randomly drawn from normally distributed populations with unknown population variances. If such assumption cannot be made, you may try nonparametric methods. The variables of interest should be random variables, whose values change randomly. A constant such as the number of parents of a person is not a random variable. In addition, the occurrence of one measurement in a variable should be independent of the occurrence of others. In other word, the occurrence of an event does not change the probability that other events occur. This property is called statistical independence. Time series data are likely to be statistically dependent because they are often autocorrelated.

T-tests assume random sampling without any selection bias. If a researcher intentionally selects some samples with properties that he prefers and then compares them with other samples, his inferences based

on this non-random sampling are neither reliable nor generalized. In an experiment, a subject should be randomly assigned to either the control or treated group so that two groups do not have any systematic difference except for the treatment applied. When subjects can decide whether or not to participate (non-random assignment), however, the independent sample t-test may under- or over-estimate the difference between the control and treated groups. In this case of self-selection, the propensity score matching and treatment effect model may produce robust and reliable estimates of mean differences.

Another, yet closely related to random sampling, key component is population normality. If this assumption is violated, a sample mean is no longer the best measure (unbiased estimator) of central tendency and t-test will not be valid. Figure 1 illustrates the standard normal probability distribution on the left and a bimodal distribution on the right. Even if the two distributions have the same mean and variance, we cannot say much about their mean difference.

T-test 는 비교양태에 따라 크게 네가지 경우로 나눈다. 먼저 자료판을 가져온다.

```
> cancer<-read.dta('/users/kucc625/smoking.dta')
> attach(cancer)
> satisfy<-read.table('/users/kucc625/satisfy.txt', header=T)
> attach(satisfy)
```

6.1.2 One Sample T-test

한 변수의 평균이 특정값과 같은지를 알아보기 위한 방법으로 가장 간단한 T-test 이다. mu 는 0 가 아닌 특정값을 지정할 수 있다. 유의수준과 검증형식은 conf.level 과 alternative 로 지정할 수 있다. 전자는 신뢰수준을 계산할 때, 후자는 p-value 를 계산할 때 사용한다. 각각 .95 와 two.sided 을 기본값으로 갖는다. 따라서 아래 두 명령어는 같은 결과를 보여준다.

```
> t.test(lung)
> t.test(lung, mu=0, conf.level=.95, alternative='two.sided')
```

폐암발생률 평균이 20 이라는 가설을 검증해보자. 신뢰수준은 .99 (유의수준 .01)로 한다.

```
> t.test(lung, mu=20, conf.level=.99)

        One Sample t-test

data:  lung
t = -0.5441, df = 43, p-value = 0.5892
alternative hypothesis: true mean is not equal to 20
99 percent confidence interval:
 17.93529 21.37108
sample estimates:
mean of x
 19.65318
```

암발생률 평균은 19.6532 이고, T 검정치는 -.5441 이고, 자유도는 43 (N=44)이다. 위험치는 .5892 으로 커서 기준가설(H_a: μ=20)을 버리기 어렵다. 또한 표본평균이 신뢰구간 안에 있다. 따라서 폐암발생률 평균은 20 이라 결론지을 수 있다. t.test()는 분산이나 표준편차를 보여주지 않는다. 계산된 검정치를 확인하기 위해 var()나 sd()로 분산이나 표준편차를 계산한다.

```
> var(lung)
[1] 17.87701
> sd(lung)
[1] 4.228122
```

검정치 -.5441 과 신뢰구간[17.9353, 21.3711]은 다음과 같이 계산된다. `4.228122/sqrt(44)` 는 평균오차(standard error)이며, 2.695 는 유의수준 .01 에서 t 기준치(critical value)이다.

```
> (19.65318-20)/(4.228122/sqrt(44))
[1] -0.5441053

> 19.65318-2.695*(4.228122/sqrt(44))
[1] 17.93535

> 19.65318+2.695*(4.228122/sqrt(44))
[1] 21.37101
```

폐암발생률 평균이 20 보다 작다는 가설은 alternative 에서 한쪽검증(less or greater)을 지정하면 된다. 다음에서 유의수준은 1 할임에 유의하라.

```
> t.test(lung, mu=20, conf.level=.90, alternative=c('less'))
```

6.1.3 Paired Sample T-test

짝지어진 변수값 차이의 평균이 특정값과 같은지 검증한다. 방법론은 7.1.2 과 같다. 2008 년과 2009 년 IT 서비스 만족도를 비교한다고 해보자. 변화가 없다면 각 항목 차이를 계산한 결과 그 평균은 0 이 될 것이다.

```
> t.test(iub2008, iub2009, mu=0, paired=T)

        Paired t-test

data:  iub2008 and iub2009
t = 3.7607, df = 10, p-value = 0.003718
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8632042 3.3731595
sample estimates:
mean of the differences
               2.118182
```

표본평균 2.1182 는 0 과 많이 떨어져 있고 검정치 3.7607 는 매우 크다. 그러므로 실제 평균이 0 이라는 기준가설을 기각해도 틀릴 확률(위험치)이 겨우 .0037 이다. 따라서 실제 평균은 0 이 아니라 결론지을 수 있다. 다음 iub2008_2009 는 iub2008 에서 iub2009 를 뺀 값이다. 예컨대, 2.1=99.3-97.2 이다.

```
   iub2008 iub2009 iub2008_2009
1     99.3    97.2          2.1
2     97.4    94.0          3.4
3     97.9    93.0          4.9
4     99.8    94.6          5.2
5     96.4    94.6          1.8
6     98.1    95.4          2.7
7     97.6    98.6         -1.0
8     93.3    92.8          0.5
9     96.2    95.4          0.8
10    97.8    95.8          2.0
11    98.5    97.6          0.9
```

위에서 행한 paired t-test 는 다음과 같은 one sample test 와 같다.

```
> t.test(iub2008_2009)

        One Sample t-test

data:  iub2008_2009
t = 3.7607, df = 10, p-value = 0.003718
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.8632042 3.3731595
sample estimates:
mean of x
 2.118182
```

다음 예는 2008 년과 2009 년의 만족도 차이가 2 보다 크다는 가설을 검정한다.

```
> t.test(iub2008, iub2009, paired=T, mu=2, alternative=c('greater'))

        Paired t-test

data:  iub2008 and iub2009
t = 0.2098, df = 10, p-value = 0.419
alternative hypothesis: true difference in means is greater than 2
95 percent confidence interval:
 1.097331      Inf
sample estimates:
mean of the differences
               2.118182
```

표본평균이 2 와 차이가 없고, 신뢰구간$[1.0973, +\infty]$ 안에 위치해 있다. 검정치가 매우 작은 반면 위험치가 크다. 즉, 만족도 차이가 2 보다 크다는 기준가설을 기각한다면 10 번에 4 번정도는 틀릴 가능성이 있다. 따라서 기준가설을 기각하지 않는 것이 안전하다. 다음 표준편차를 이용하여 검정치를 확인해 보라.

```
> sd(iub2008_2009)
[1] 1.868057
```

검정력을 계산하기 위해서는 power.t.test()를 사용하고 t-test 의 종류를 지정한다. 검정력은 .9224 인데, 두 연도간 차이가 정말 있다면 이 검정은 그 효과를 9 할 이상 찾아낼 수 있다는 뜻이다(반대로 100 번 검정을 하면 7-8 번은 효과가 있어도 찾아내지 못한다).

```
> power.t.test(n=11, delta=2.1182, sd=1.8681, sig.level=.05, type=c("one.sample"))

     One-sample t test power calculation

              n = 11
          delta = 2.1182
             sd = 1.8681
      sig.level = 0.05
          power = 0.9223524
    alternative = two.sided
```

6.1.4 Independent Sample T-test with Equal Variance

독립된 변수의 평균을 비교하는 방법으로 두 변수의 분산이 같다는 것을 가정한다. 보통 한 변수의 분산이

다른 변수 분산의 3 배 이상 크면 분산이 다르다고 볼 수 있다. 두 변수의 분산이 같은지는 var.test()로 점검한다. 다음(두 명령어 중 하나)은 흡연자와 비흡연자의 폐암발생률 분산을 비교한 것이다.

```
> var.test(cancer[smoke==1,][,4],cancer[smoke==0,][,4])
> var.test(heavy, light)

        F test to compare two variances

data:  heavy and light
F = 1.1666, num df = 21, denom df = 21, p-value = 0.7273
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4843455 2.8098353
sample estimates:
ratio of variances
          1.166590
```

작은 분산에 대한 큰 분산의 비가 F 값이다. 즉, 1.1666=11.6838/10.0153. 분자와 분모의 자유도는 각 표본수에서 1 을 뺀 값이다. 여기서 표본수는 각각 22 이기 때문에 자유도는 21 이다. F 값이 작고 위험도가 크기 때문에 두 변수의 분산이 같다는 기준가설을 기각할 수 없다. 두 변수의 분산행렬을 이용하여 F 값을 확인해 보라.

```
          heavy     light
heavy 11.683757  2.865445
light  2.865445 10.015311
```

자료판이 넓은 모양새로 되어 있다면 다음 명령어를 사용해보자. var.equal=T 는 두 변수의 분산이 같다는 것을, paired=F 는 두 변수가 짝지어 있지 않고 독립적라는 것을 의미한다. mu=0 는 두 변수의 평균차가 0 임을 검정한다. 여기서 paired=F 와 mu=0 는 기본값으로 군더더기이다.

```
> t.test(light, heavy, var.equal=T, paired=F mu=0)

        Two Sample t-test

data:  light and heavy
t = -5.3714, df = 42, p-value = 3.164e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.338777 -3.330314
sample estimates:
mean of x mean of y
 16.98591  22.32045
```

긴 모양새로 된 자료판이라면 집단을 구분하는 두쪽변수를 가지고 있어야 한다. 이때 두쪽변수를 tilde ~ 다음에 적어놓는다. mu 를 이용하여 두 평균의 차이를 지정할 수도 있다.

```
> t.test(lung~smoke, var.equal=T, conf.level=.95)

        Two Sample t-test

data:  lung by smoke
t = -5.3714, df = 42, p-value = 3.164e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.338777 -3.330314
sample estimates:
mean in group 0 mean in group 1
       16.98591        22.32045
```

pairwise.t.test()는 집단간 t-test 를 수행하여 위험치를 행렬형태로 보여준다. 아래 명령어에서 집단을 구분하는 변수는 비교할 변수 뒤에 따라오며, pooled.sd=T 와 paired=F 는 기본값이다.

```
> pairwise.t.test(lung, smoke, pooled.sd=T, paired=F)

        Pairwise comparisons using t tests with pooled SD

data:  lung and smoke

  0
1 3.2e-06

P value adjustment method: holm
```

다음 예는 leukemia 와 신장암 발생률의 평균을 비교한다. 두 변수의 분산은 비슷하다(F=1.5119). 표본평균이 큰 차이가 있으며, t 검정치도 크며, 위험치는 미미하기 때문에 두 암발생률의 평균은 서로 다르다고 결론내린다.

```
> var.test(leukemia, kidney)

        F test to compare two variances

data:  leukemia and kidney
F = 1.5119, num df = 43, denom df = 43, p-value = 0.1794
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8249702 2.7708448
sample estimates:
ratio of variances
          1.511907

> t.test(leukemia, kidney, var.equal=T)

        Two Sample t-test

data:  leukemia and kidney
t = 32.5356, df = 86, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.788673 4.281781
sample estimates:
mean of x mean of y
 6.829773  2.794545
```

6.1.5 Independent Sample T-test with Unequal Variance

분산이 다른 독립된 변수의 평균을 비교한다. 분산이 다르기 때문에 자유도를 수정해서(approximation of degrees of freedom) 사용해야 한다. R 은 Welch 수정치를 사용한다.

다음 var.test()를 사용하여 동부와 서부의 신장암발생률 분산이 같은지 살펴본다. F 검정치 3.9749 (=.3580/.0901)가 크고 위험치가 작아 두 변수의 분산이 다르다고 결론내린다.

```
> var(east)
[1] 0.0900779
> var(west)
[1] 0.3580493
```

```
> var.test(west, east)

        F test to compare two variances

data:  west and east
F = 3.9749, num df = 23, denom df = 19, p-value = 0.003404
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.612668 9.438230
sample estimates:
ratio of variances
          3.974885
```

두 변수의 분산이 다르기 때문에 var.equal=F 를 지정하여 t-test()를 수행한다. Welch 조정치를 사용하기 때문에 자유도가 정수가 아닌 실수임에 유의하라. T 검정치가 크고 위험치가 작아서 동부와 서부 신장암 발생률은 평균이 서로 다르다고 결론내린다(p<.0086).

```
> t.test(kidney~west, var.equal=F)

        Welch Two Sample t-test

data:  kidney by smoke$west
t = 2.7817, df = 35.11, p-value = 0.008641
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1047722 0.6705611
sample estimates:
mean in group 0 mean in group 1
       3.006000        2.618333
```

위에서 계산한 분산과 표본평균을 이용하여 T 검정치를 계산해 본다. 2.7817=(3.006-2.6183)/ sqrt(.0901/20+.3580/24).

자료판이 긴모양새로 되어 있다면 다음과 같은 명령어를 실행한다. 먼저 분산이 같은지를 살펴본다. F 값이 크고 위험치가 작아서 서로 다른 분산을 가졌다고 결론내린다.

```
> var.test(bladder, kidney)

        F test to compare two variances

data:  bladder and kidney
F = 3.4556, num df = 43, denom df = 43, p-value = 8.733e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.885522 6.332942
sample estimates:
ratio of variances
          3.455561
```

T 검정값이 크고 위험치가 작으므로 두 암발생률은 평균이 다르다고 결론내린다. 방광암이 신장암보다 발생률이 높다고 할 수 있다. 자유도가 정수가 아니라 실수임에 유의하라. Welch 조정치를 사용했기 때문이다.

```
> t.test(bladder, kidney)

        Welch Two Sample t-test
```

```
data:  bladder and kidney
t = 8.0312, df = 65.964, p-value = 2.337e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9967938 1.6563880
sample estimates:
mean of x mean of y
 4.121136  2.794545
```

pairwise.t.test()는 집단을 구분하는 변수가 두 개 이상 값을 가질 때 유용하다. 각 집단끼리 t-test 를 수행하고 위험치를 행렬형태로 보여준다. 예컨대, 집단 1 과 3 의 방광암 발생률 평균은 서로 다르다 (p<.016).

```
> pairwise.t.test(bladder, area, pooled.sd=F, paired=F)

        Pairwise comparisons using t tests with pooled SD

data:  bladder and area

  1     2     3
2 0.081 -     -
3 0.016 1.000 -
4 0.050 1.000 1.000

P value adjustment method: holm
```

6.2 비율 비교

두쪽 변수를 비교할 때는 binomial 분포를 이용한다. R 에서는 prop.test(x, n, p)를 사용한다. 30 번 시행 중에 20 번 성공한 경우 성공확률이 .5 인지를 확인해보자. correct=F 는 Yates's continuity correct 을 끈다. X-squared 는 z 값을 제곱한 것이다. 따라서 실제 z 값은 1.8257=sqrt(3.3333) 이다.

```
> prop.test(20, 30, p=.5, alternative=c('two.sided'), correct=F)

        1-sample proportions test without continuity correction

data:  20 out of 30, null probability 0.5
X-squared = 3.3333, df = 1, p-value = 0.06789
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4878005 0.8076950
sample estimates:
        p
0.6666667
```

두쪽변수 둘을 비교해 보자. 30 번 중에 20 번 성공한 경우와 10 번 성공한 경우를 따져보다. 검정치가 크고 위험치가 작기 때문에 두 변수의 성공비율은 다르다고 말할 수 있다.

```
> prop.test(c(20, 10), c(30, 30), correct=F)

        2-sample test for equality of proportions without continuity correction

data:  c(20, 10) out of c(30, 30)
X-squared = 6.6667, df = 1, p-value = 0.009823
alternative hypothesis: two.sided
95 percent confidence interval:
 0.09477411 0.57189255
sample estimates:
```

```
   prop 1    prop 2
0.6666667 0.3333333
```

z 값은 다음과 같이 계산한다.

```
> sqrt(6.6667)
[1] 2.581995
```

6.3 One-Way ANOVA

ANOVA 는 aov( )을 사용한다. 다음 예에서 F 검정치 28.852 는 7.1.4 의 t 검정치 -5.3714 의 제곱이다.

```
> summary(aov(lung~smoke))
            Df Sum Sq Mean Sq F value    Pr(>F)
smoke        1 313.03  313.03  28.852 3.164e-06 ***
Residuals   42 455.68   10.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

선형회귀모형도 같은 분석을 해준다. 다만 보여주는 통계량이 좀 다를 뿐이다. 7.1.4 에서 얻은 t 통계량은 회귀계수의 t 값에서 보여준다. F 값은 one-way ANOVA 와 같다.

```
> summary(lm(lung~smoke))

Call:
lm(formula = lung ~ smoke)

Residuals:
      Min        1Q    Median        3Q       Max
-10.21045  -1.40341  -0.04091   2.28818   8.46409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9859     0.7023  24.188  < 2e-16 ***
smoke         5.3345     0.9931   5.371 3.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.294 on 42 degrees of freedom
Multiple R-squared: 0.4072,       Adjusted R-squared: 0.3931
F-statistic: 28.85 on 1 and 42 DF,  p-value: 3.164e-06
```