

고고학 자료 통계분석

Week 6 : 검정과 상관분석

숭실대학교 사학과 석사과정 1학기
주 찬 혁

계획

주차	제목	내용
1	Intro	소개, R 설치
2	기초 통계(1)	모집단과 표본, 기술통계량
3	기초 통계(2)	변수의 종류, 기설과 검정, 오류, 분석절차
4	전처리	데이터 전처리
5	시각화	다양한 종류의 그래프
6	검정과 상관분석	검정, 상관분석
7	회귀분석	선형회귀, 다중선형회귀, 로지스틱회귀
8	군집분석	K-means,
9	판별분석	DA, MDA
10	주성분분석	PCA

복습

- 시각화의 목적과 유용성에 대해 안다.
- 다양한 시각화 방법의 장단점에 안다.
- R을 사용하여 데이터를 시각화 할 수 있다.

검정

- 표본의 정보를 통해 **가설의 합당성을 판정**하는 과정
- 귀무가설을 수립하고 이를 기각하여 대립가설을 채택할 수 있는지 판정
- 항상 [기각 0] / [기각 X] 로 구분되지만, 이를 100%로 신뢰할 수는 없으므로 일정한 **유의수준**을 통해 범위 내에서만 설명

검정

				종속변수			
				연속 변수		범주 변수	
				정규분포	비정규분포	순위 변수	명목 변수
독립 변수	명목 변수	두군	독립된 자료	independent samples t-test	Mann-Whitney U test		chi-squared test / Fisher's exact test
			짝지어진 자료	paired samples t-test	wilcoxon signed rank test		McNemar test
		세군 이상	독립된 자료	one-way ANOVA	Kruskal-Wallis test		chi-squared test
			짝지어진 자료	repeated Measure ANOVA	Friedman test		cochran Q test

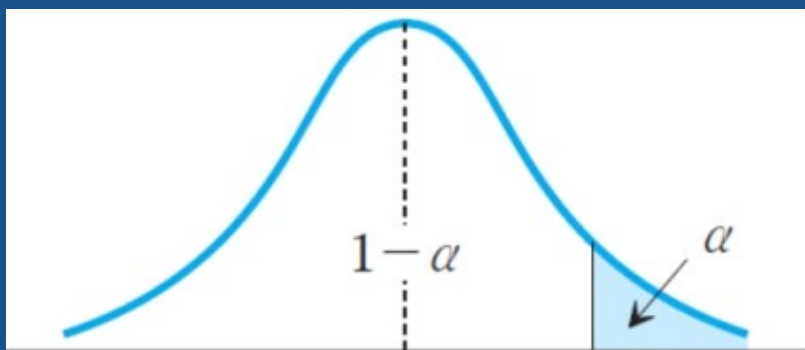
***오늘은 t-test와 Chi-squared test만 다루도록 하겠습니다.**
+ Median Polish

T-검정

- ~~One sample t-test~~ : 모수와 표본이 같은가 → 고고학자료는 모수 모름
- Two sample t-test : 독립적인 두 표본이 실제로 다른지, 혹은 단순 표본상의 차이인지 확인
- **평균**을 사용하기 때문에 저항성이 작음 → 이상치에 민감
- 전제
 1. 연속형 변수
 2. 양자 모두 정규분포를 이루어야함
 3. 두 표본 사이에 분산차가 없어야함(완화 가능)

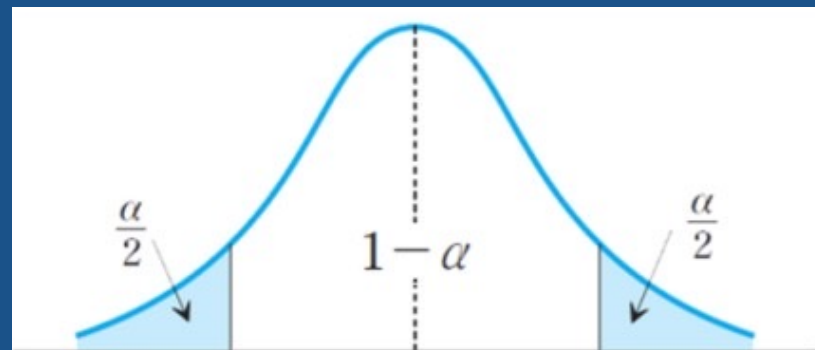
T-검정

단측검정(one-tailed)



귀무가설 : $A = B$
대립가설 : $A < B$,
 $A > B$

양측검정(two-tailed)



귀무가설 : $A = B$
대립가설 : $A \neq B$

T-검정

A지구



B지구



***상황가정**

A지구 주거지들의 면적이 B지구 주거지들의 면적보다 크지 확인하고 싶음

귀무가설 : A지구 주거지와 B지구 주거지 면적은 같다

-> 기각시킬 수 있다면 두 지구의 주거지들의 면적은 서로 다른 것

*고전적인 방법

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2} (x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

cum. prob one-tail two-tails	t _{.50}	t _{.75}	t _{.80}	t _{.85}	t _{.90}	t _{.95}	t _{.975}	t _{.99}	t _{.995}	t _{.999}	t _{.9995}
	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

T-검정

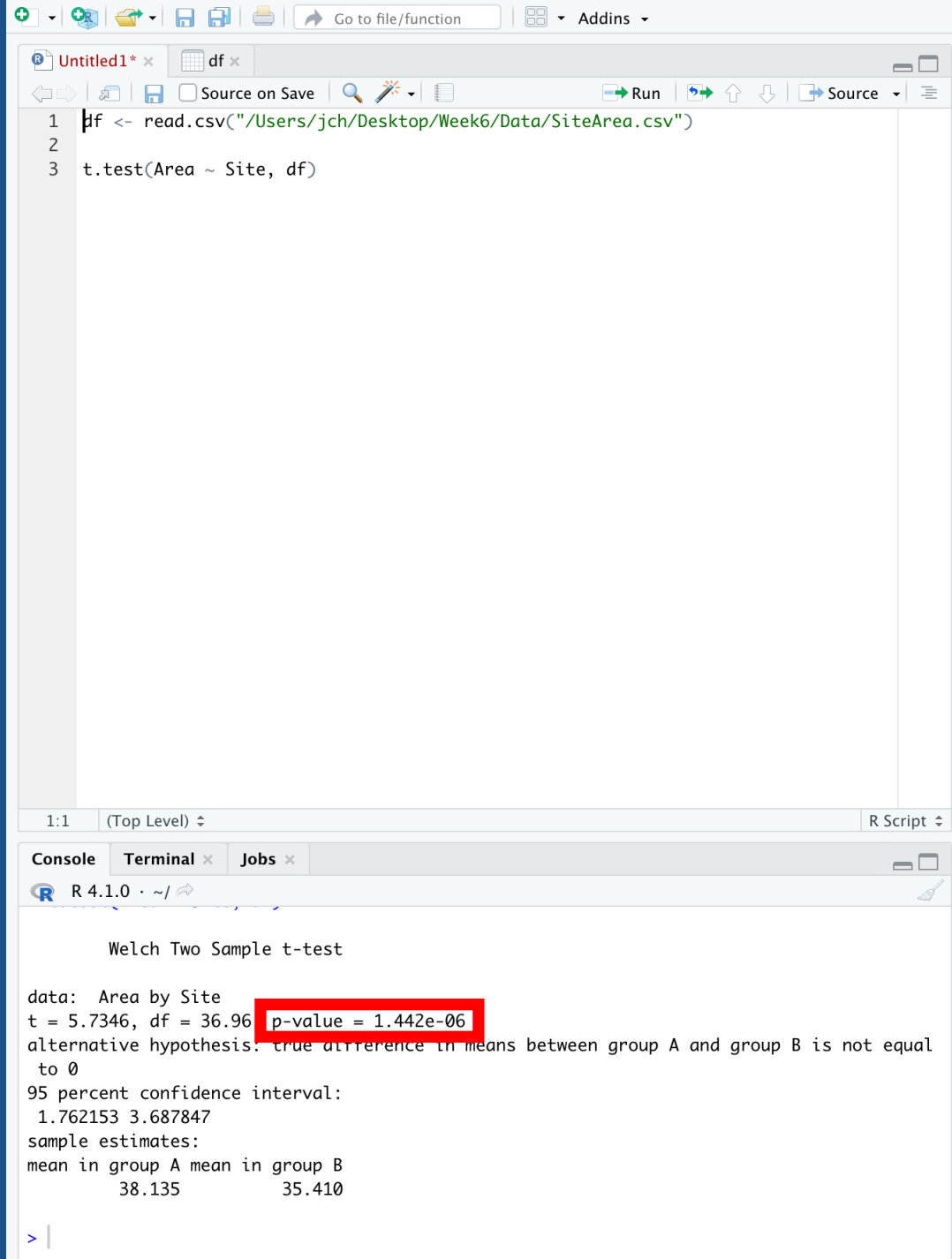
*R과 함께하는 아주 편한 방법 (^_^)

```
> t.test(변수값 ~ 변수구분, 데이터)
```

p-value가 0.05보다 작으면 귀무가설 기각

*p-value : 유의확률

*인문사회과학에서는 관습적으로 0.05(5%)를
유의수준으로 사용



The screenshot shows the RStudio environment. The script editor at the top contains the following R code:

```
1 df <- read.csv("/Users/jch/Desktop/Week6/Data/SiteArea.csv")
2
3 t.test(Area ~ Site, df)
```

The console at the bottom displays the output of the t-test:

```
Welch Two Sample t-test

data: Area by Site
t = 5.7346, df = 36.96, p-value = 1.442e-06
alternative hypothesis: true difference in means between group A and group B is not equal to 0
95 percent confidence interval:
 1.762153 3.687847
sample estimates:
mean in group A mean in group B
      38.135      35.410
```

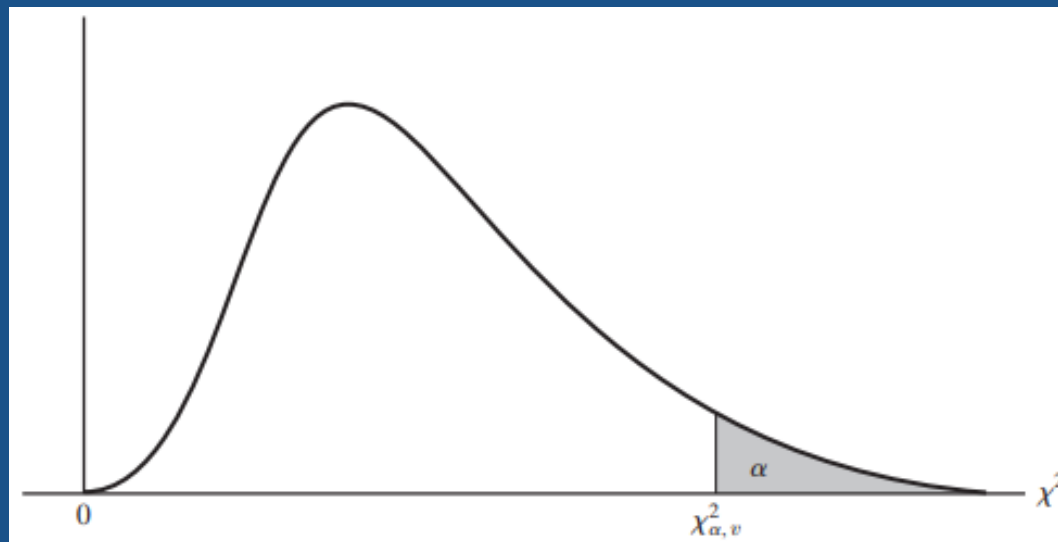
The p-value, 1.442e-06, is highlighted with a red box in the original image.

Chi-squared 검정

- 두 명목형 혹은 순서형 변수의 결합이 유의미한지 판정
- 기대치와 관찰치 사이의 차이를 검증
- 표본의 수가 많을 수록 귀무가설 기각이 쉬워짐
 - > 통계학적으로 유의미 but 실제 의미를 갖는지는 검토 필요
- 전제
 1. 교차분할표 상의 각 셀의 값이 5 이상일 것을 권장
 2. 교차분할표 상의 셀 값 중 0이 없을 것을 권장

Chi-squared 검정

Chi-squared 확률밀도함수



Chi-squared 검정

무덤



***상황가정**

성별에 따라 매장위치가 다른지
확인하고 싶음

귀무가설 : 성별과 매장위치는
상관이 없다.

대립가설 : 성별과 매장위치는
상관이 있다.

Chi-squared 검정

*고전적인 방법

(1)교차분할표 작성 -> (2)x-squared 계산
-> (3) Chi-Squared test table의 값에
알맞는 부분을 찾아 기각 확인

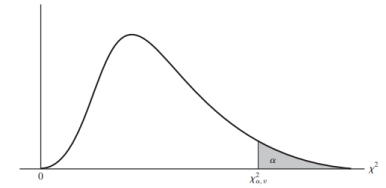
	<i>M</i>	<i>F</i>	
RHS	29	14	43
LHS	11	33	44
	40	47	87

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed frequency counts in each category

E_i = expected frequency counts in each category

k = number of categories



ν	$\alpha = 0.250$	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77529	4.60517	5.99146	7.37776	9.21034	10.5966	13.816
3	4.10834	6.25139	7.81473	9.34840	11.3449	12.8382	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8603	18.467
5	6.62568	9.23636	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2189	13.3616	15.5073	17.5345	20.0902	21.9550	26.125
9	11.3888	14.6837	16.9190	19.0228	21.6660	23.5894	27.877
10	12.5489	15.9872	18.3070	20.4832	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7568	31.264
12	14.8454	18.5493	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3620	24.7356	27.6882	29.8195	34.528
14	17.1169	21.0641	23.6848	26.1189	29.1412	31.3194	36.123
15	18.2451	22.3071	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3689	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1565	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1909	38.5823	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6706	35.4789	38.9322	41.4011	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7957	48.268
23	27.1413	32.0069	35.1725	38.0756	41.6384	44.1813	49.728
24	28.2412	33.1962	36.4150	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9279	52.618
26	30.4346	35.5632	38.8851	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1945	46.9629	49.6449	55.476
28	32.6205	37.9159	41.3371	44.4608	48.2782	50.9934	56.892
29	33.7109	39.0875	42.5570	45.7223	49.5879	52.3356	58.301
30	34.7997	40.2560	43.7730	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8051	55.7585	59.3417	63.6907	66.7660	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9815	74.3970	79.0819	83.2977	88.3794	91.9517	99.607
70	77.5767	85.5270	90.5312	95.0232	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	128.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449

Chi-squared 검정

*R과 함께하는 아주 편한 방법 (^_^)

> table(변수1, 변수2)

- 교차분할표 생성

> chisq.test(변수1, 변수2)

- Chi-squared test 실행

p-value가 0.05보다 작으면 귀무가설
기각

*p-value : 유의확률

*인문사회과학에서는 관습적으로
0.05(5%)를 유의수준으로 사용

The screenshot shows the RStudio interface. The script editor on the left contains the following R code:

```
1 #data
2 df <- read.csv("/Users/jch/Desktop/Week6/Data/SiteArea")
3 chidf <- read.csv("/Users/jch/Desktop/Week6/Data/Chi_squar
4
5 #data 분할
6 df_A <- subset.data.frame(df, df$Site=="A")
7 df_B <- subset.data.frame(df, df$Site=="B")
8 summary(df_A)
9 summary(df_B)
10
11 #t-test
12 t.test(Area ~ Site, df)#두 표본의 분산이 다를 때
13 t.test(Area ~ Site, var.equal=TRUE,df)#두 표본의 분산이
14
15 #chi-squared test
16 table(chidf$Sex, chidf$RorL)
17 chisq.test(chidf$Sex, chidf$RorL)
18
```

The Environment pane on the right shows the following objects:

Object	Description
chidf	87 obs. of 2 variables
df	40 obs. of 3 variables
df_A	20 obs. of 3 variables
df_B	20 obs. of 3 variables
med.d	List of 5

The Console pane at the bottom shows the output of the R script:

```
> chidf <- read.csv("/Users/jch/Desktop/Week6/Data/Chi_squar
ed_Test.csv")
> #chi-squared test
> table(chidf$Sex, chidf$RorL)

  LHS RHS
F  33  14
M  11  29
> chisq.test(chidf$Sex, chidf$RorL)

Pearson's Chi-squared test with Yates' continuity co
rrection

data: chidf$Sex and chidf$RorL
X-squared = 14.109, df = 1, p-value = 0.0001725
> |
```

The p-value of 0.0001725 is highlighted with a red box in the original image.

Median Polish

- 교차분할표를 작성한 뒤 각 셀의 값들을 중앙값(Median)으로 빼는 행위를 반복하여 자료의 패턴을 찾는 방법
- 탐색적 데이터 분석(EDA, Exploratory Data Analysis) 중 하나
- 단순히 변수 사이의 관계여부를 떠나 **관계의 강도를 파악**할 수 있음
- Chi-squared test의 대안으로 활용될 수 있음

Median Polish

*R과 함께하는 아주 편한 방법 (^_^)

```
> 변수명 <- medpolish(table(변수1,  
변수2))  
> print(변수명)
```

- 변수명에 Median polish 결과를 저장하고 이를 출력
- Overall은 각 셀의 값을 중앙값으로 뺀 횟수
- Residuals 부분에서 양수일 경우 양의 관계, 0일 경우 관계 없음, 음수일 경우 음의 관계로 해석

The screenshot shows the RStudio interface with a script editor on the left and a console on the bottom. The script editor contains R code for reading data, performing a t-test, a chi-squared test, and a median polish. The console shows the output of the median polish, including the overall value, row and column effects, and residuals. The residuals section is highlighted with a red box.

```
#data  
df <- read.csv("/Users/jch/Desktop/Week6/Data/SiteArea")  
chidf <- read.csv("/Users/jch/Desktop/Week6/Data/Chi_s")  
#data 분할  
df_A <- subset.data.frame(df, df$Site=="A")  
df_B <- subset.data.frame(df, df$Site=="B")  
summary(df_A)  
summary(df_B)  
#t-test  
t.test(Area ~ Site, df)#두 표본의 분산이 다를 때  
t.test(Area ~ Site, var.equal=TRUE,df)#두 표본의 분산이 같을 때  
#chi-squared test  
table(chidf$Sex, chidf$RorL)  
chisq.test(chidf$Sex, chidf$RorL)  
#median polish  
med.d <- medpolish(table(chidf$Sex, chidf$RorL))  
print(med.d)
```

Console Output:

```
R 4.1.0 ~/  
> print(med.d)  
Median Polish Results (Dataset: "table(chidf$Sex, chidf$RorL)")  
Overall: 21.75  
Row Effects:  
  F    M  
1.75 -1.75  
Column Effects:  
 LHS  RHS  
0.25 -0.25  
Residuals:  
      LHS  RHS  
F  9.25 -9.25  
M -9.25  9.25  
>
```

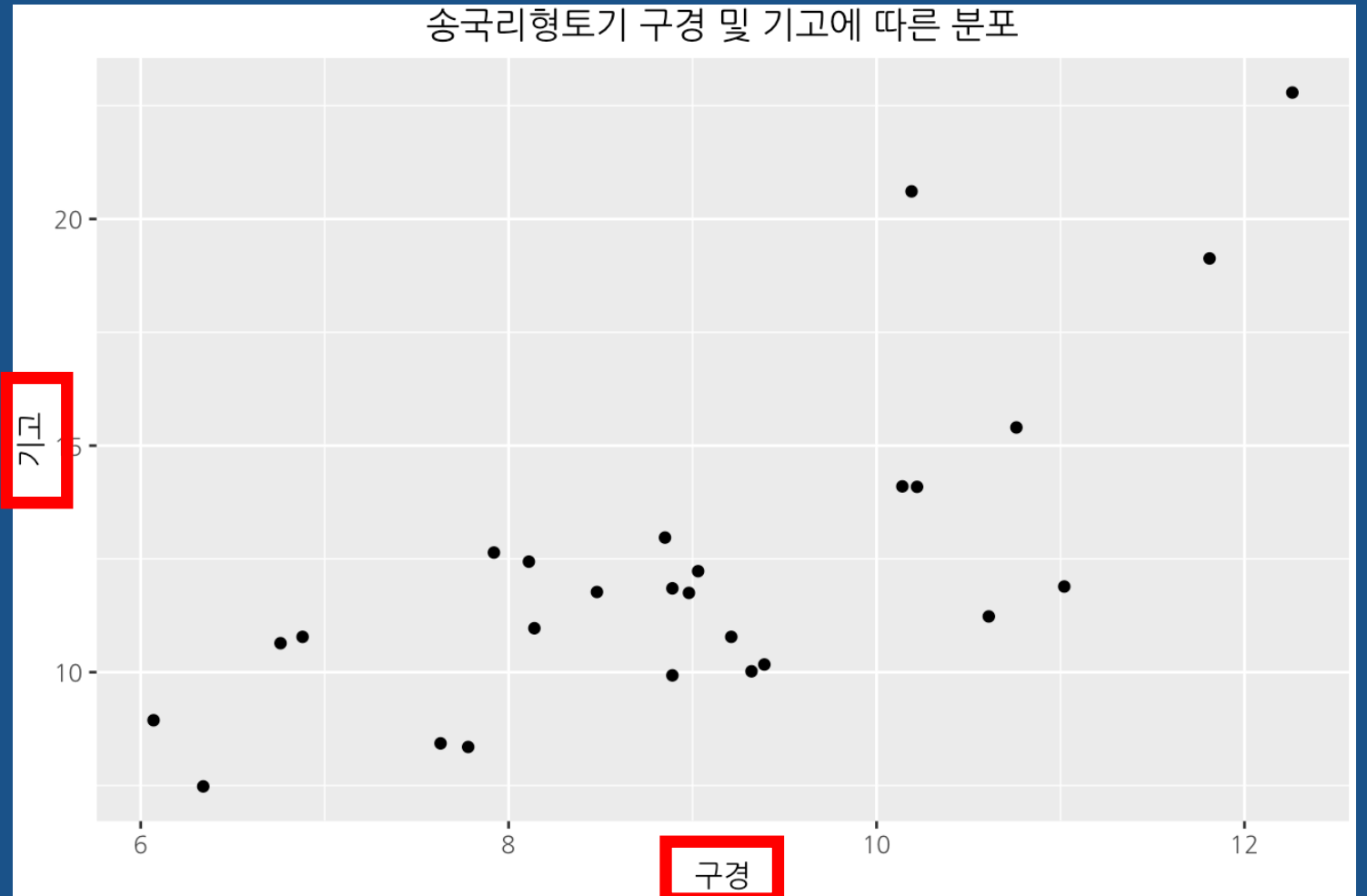
상관분석

- 두 변수 사이의 어떠한 **선형관계**가 존재하는지 분석하는 방법
- 단, 상관분석은 단순히 관계의 정도만을 나타낼 뿐이며 이를 통해 **인과관계에 대해서는 알 수 없다.**
- 상관관계 : 두 변수 사이의 관계의 강도
- 상관계수 : 두 변수 사이의 관계를 수치로 표현
 - Pearson
 - Spearman
 - Kendall

상관관계

- 두 변수 간의 관계(선형)를
상관관계라고 하며, 양의
상관, 0, 음의 상관이 있다.

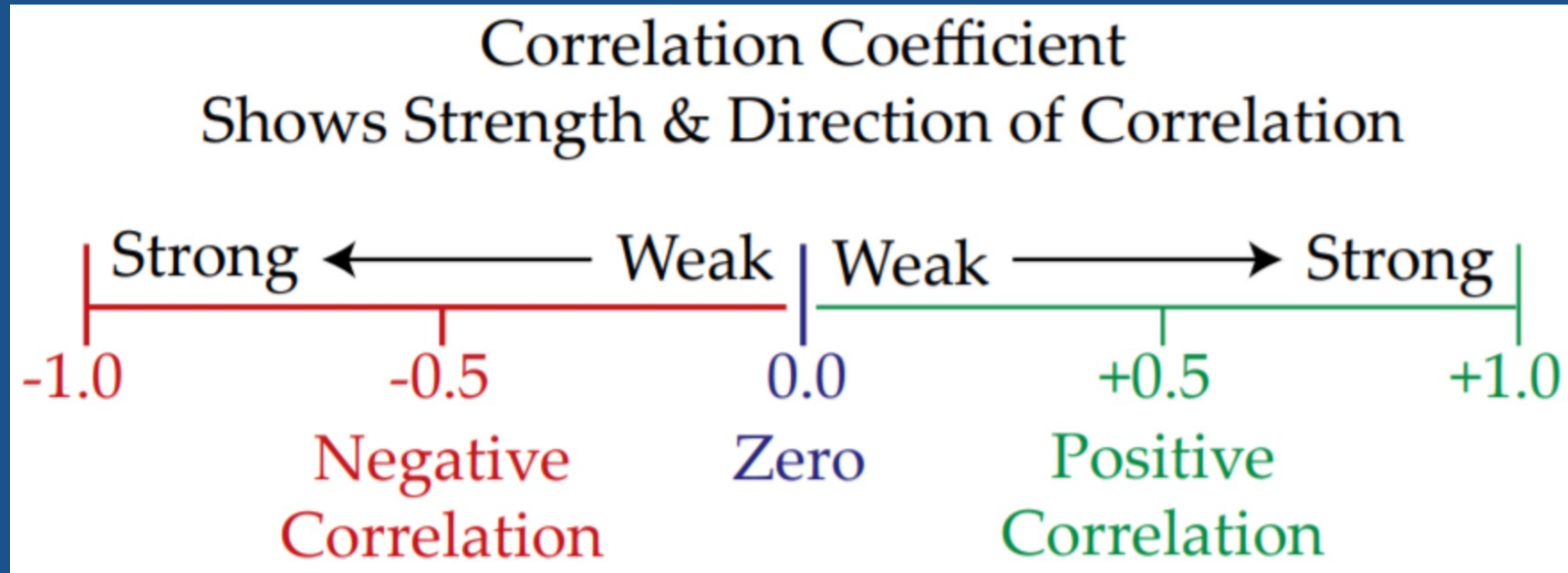
- 양의 상관 : 비례
- 0 : 무상관
- 음의 상관 : 반비례



상관계수

- 상관계수 : 두 변수 사이의 관계를 수치로 표현
 - **Pearson** : 가장 기본적인 상관계수
 - **연속형 변수** 사이의 상관관계 측정
 - Spearman
 - 변수를 순서형 변수로 변환하여 상관관계 측정
 - 저항성이 낮음
 - Kendall
 - 변수를 순서형 변수로 변환하여 상관관계 측정
 - 표본수가 적을 때 용이

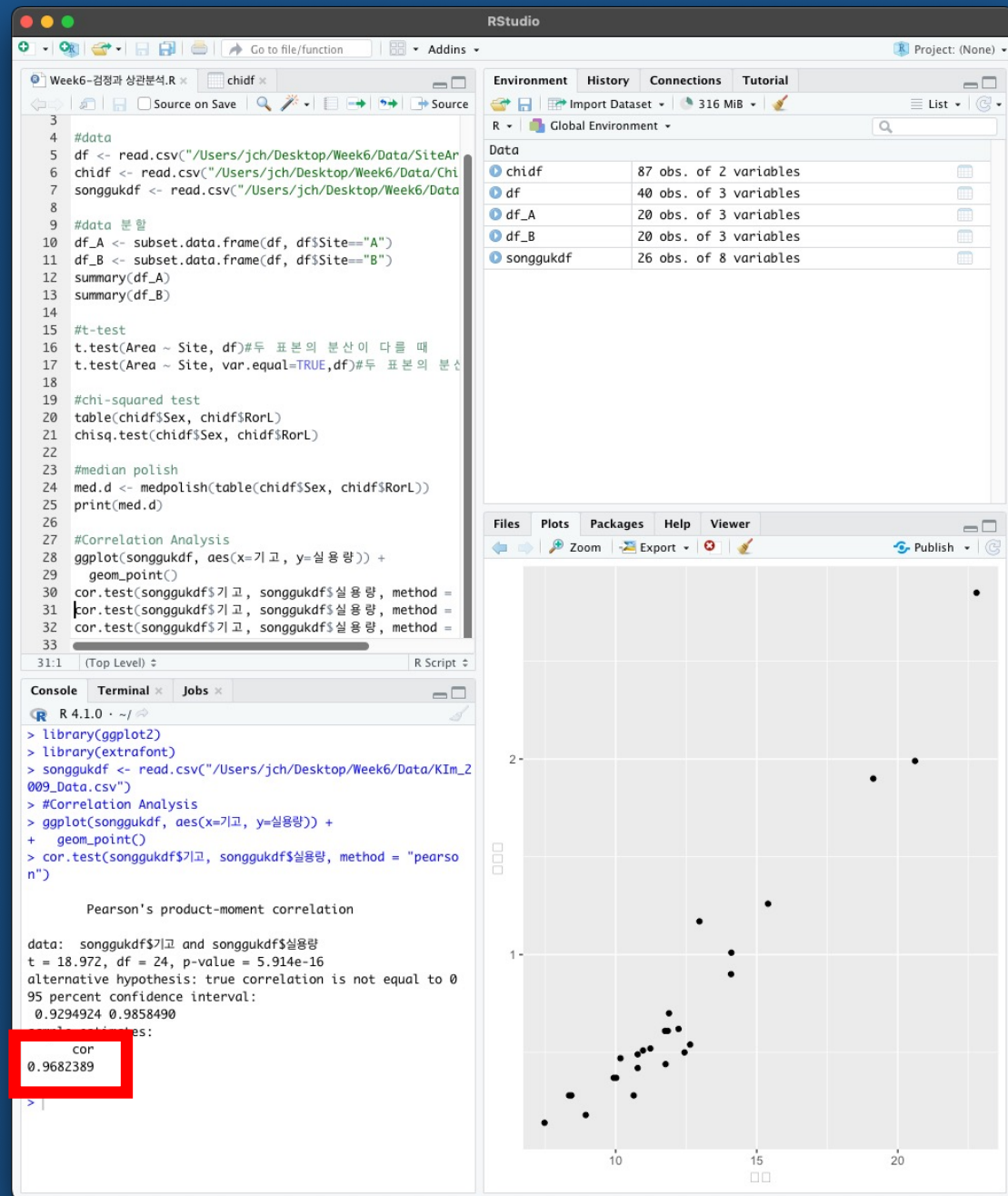
상관계수



상관분석

> cor.test(변수1, 변수2, method = "상관계수")

- 변수를 지정
- 사용을 원하는 상관계수 설정
- 상관계수 값을 확인하여 두 변수 사이의 관계를 확인



상관분석

- 단, 상관분석은 단순히 관계의 정도만을 나타낼 뿐이며 이를 통해 **인과관계에 대해서는 알 수 없다.**
- 인과관계를 파악하기 위해서는 다음 주의 주제인 **회귀분석**을 시행해야함