공분산분석

이용희

October 16, 2018

Contents

1 두 집단의 평균 비교

1.1 t-검정

통계학에서 나오는 가장 기본적이고 자주 쓰이는 가설검정 방법은 두 집단의 평균의 차이를 검정하는 t-검정(t-test)이다.

예를 들어 새로운 약을 개발하면 제일 먼저 해야할 일은 새로운 약이 병을 치료하는 유의한 효과가 있다는 사실을 통계적으로 보여야 한다. 즉, 위약을 사용한 집단과 새로운 약을 사용한 집단의 평균이 다르다는 가설을 검정해야 한다. 위약을 사용한 집단을 P, 새로운 약을 사용한 집단을 N 이라고 했을 때 각각의 집단의 평균을 μ_P, μ_N이라고 하자. 여기서 고려해야할 가설은 다음과 같다.

$$H_0 : \mu_P = \mu_N \quad \text{vesus} \quad H_1 : \mu_P \neq \mu_N \tag{1}$$

두 집단이 분산이 동일한 정규분포 $N(\mu_P, \sigma^2)$, $N(\mu_N, \sigma^2)$를 따른다고 가정하면 위의 가설을 다음과 같은 t-통계량을 이용하여 검정할 수 있다.

$$t = \frac{\bar{Y}_P - \bar{Y}_N}{s_p\sqrt{1/n_P + 1/n_N}} \tag{2}$$

여기서 $\bar{Y}_P$, $\bar{Y}_N$은 집단 P, N의 표본 평균을 나타내고 n_P, n_N 은 각 집단의 표본 개수를 나타낸다. 또한 s_p^2 은 두 집단의 공통분산 추정량(pooled variance estimator)이며 다음과 같이 계산한다.

$$s^2 = \frac{\sum_{j=1}^{n_P}(Y_{Pj} - \bar{Y}_P)^2 + \sum_{j=1}^{n_N}(Y_{Nj} - \bar{Y}_N)^2}{n_P + n_N - 2}$$

유의수준을 5%라고 했을 때 위에서 구한 t-통계량의 절대값이 크다면 귀무가설 H_0 에 반대되는 증거이고 따라서 t-통계량의 절대값이 자유도 $df = n_P + n_N - 2$ 를 가지는 t-분포의 상위 2.5% 백분위수보다 크면 귀무가설 H_0를 기각하고 대립가설 H_1을 채택한다.

$$\text{Reject } H_0 \quad \text{if} \quad |t| > t(0.975, n_P + n_N - 2)$$

이러한 가설 검정은 p-값(p-value)을 구하고 그 값이 5%보다 작으면 귀무가설을 기각하는 방법과 동일하다

$$\text{Reject } H_0 \quad \text{if p-value } = 2P[t(n_P + n_N - 2) > |t|] < 0.05$$

1.2 t-검정의 재구성

이제 두 집단에 대한 가설 검정을 세 개 이상인 여러 개의 집단에 대한 경우로 확장하는 것을 생각해보자. 여러 개의 집단에 대한 가설 검정을 고려하기 전에 두 집단에 대한 t-검정을 재구성하여 여러 개의 집단

에 대한 검정법의 일반적인 이론을 생각해보려고 한다. 이제 t-검정에서 검정 통계량의 분자와 분모를 따로 살펴보자

$$t = \frac{\bar{Y}_P - \bar{Y}_N}{s_p\sqrt{1/n_P + 1/n_N}}$$

t-검정 통계량의 분자는 집단 간의 평균의 차이를 나타낸다. 즉 $\bar{Y}_P - \bar{Y}_N$는 두 집단의 표본 평균의 차이를 추정하는 양이고 그 차이가 크면 클수록 두 집단의 모평균의 차이 $\mu_P - \mu_N$가 크다는 것을 의미한다. t-검정 통계량의 분모는 두 집단의 공통분산 추정량 s_p^2에 비례한다. 즉 집단 내의 변동 $\sum_{j=1}^{n_P}(Y_{Pj} - \bar{Y}_P)^2 + \sum_{j=1}^{n_N}(Y_{Nj} - \bar{Y}_N)^2$ 이 크면 클수록 t-검정 통계량은 그 크기가 작아져서 귀무가설의 기각을 어렵게 한다.

정리해보면 t-검정 통계량은 집단 간의 평균의 차이(between-group variation)를 집단 내의 변동 (within-group variation) 으로 나누어준 값이며 이를 t-검정 통계량을 제곱하여 다음과 같이 표현할 수 있다.

$$t^2 = \frac{(\bar{Y}_P - \bar{Y}_N)^2}{s_p^2(1/n_P + 1/n_N)} = \frac{\text{between-group variation}}{\text{within-group variation}}$$

두 집단의 비교뿐만이 아니라 3개 이상의 집단에 대해 평균이 같은지 검정하는 경우를 고려해보자. 이 경우 집단 간의 평균들의 차이(집단 간의 변동)와 집단 내의 변동을 추정할 수 있는 통계량을 찾아서 t-검정 통계량과 유사하게 사용할 수 있는 방법을 생각해 보려고 한다.

2 여러 집단의 평균의 비교

2.1 가설과 모형

앞 절의 예를 확장해서 이제 새로운 약을 사용한 집단, 기존의 약을 사용한 집단, 위약을 사용한 집단의 평균이 모두 같은지 또는 아닌지 가설을 검정해야 한다고 가정하자. 새로운 약을 사용한 집단을 N, 기존의 약을 사용한 집단을 O, 위약을 사용한 집단을 P 라고 했을 때 각각의 집단의 평균을 편의상 μ_1, μ_2, μ_3 이라고 하자(집단의 수가 많아질수록 집단을 나타내는 기호를 숫자로 바꾸는 것이 편리하다. 여기서는 $N \equiv 1$, $O \equiv 2$, $P \equiv 3$이라고 놓자) 여기서 고려해야할 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vesus} \quad H_1 : \text{ not } H_0$$

위의 가설에서 주의할 점은 귀무가설의 반대인 대립가설의 경우 평균들이 다른 여러 가지 가능성이 있다는 것이다 (예를 들어 $\mu_1 = \mu_2 < \mu_3$)

여기서 집단의 개수가 많아지면 각 집단의 평균을 직접 비교하는 것보다 다음과 같이 각 집단의 효과(처리 효과; treatment effect)를 고려하는 효과모형(effect model)이 편리하다.

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1,2,..,I \quad j = 1,2,3...,n \tag{3}$$

여기서 Y_{ij}를 i 번째 집단의 j번째 관측값이며 각 집단의 평균 μ_i을 전체집단의 평균 μ와 집단의 효과 α_i의 합으로 표현하였다($\mu_i = \mu + \alpha_i$). 예를 들어 각 집단에서 관측한 자료의 수를 각각 $n = 20$이라 하면 3개의 집단에서 총 60개의 관측치가 있고 이때 집단의 개수를 $I = 3$ 이다.

이러한 효과모형을 고려하면 위의 가설을 다음과 같이 바꿀수 있다. 집단에 대한 효과가 모두 0이 되면 모든 집단의 평균이 같아진다.

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_I = 0 \quad \text{vesus} \quad H_1 : \text{ not } H_0 \tag{4}$$

또한 위의 효과 모형 (3) 에서 e_{ij}는 관측에서 생기는 오차를 의미하며 모두 확률적으로 독립이고 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정한다.

2.2 분산분석과 F-검정

이제 앞 절에서 생각해본 t-검정의 재구성처럼 집단 간의 변동의 크기(각 집단의 평균의 차이가 얼마나 나는지에 대한 통계량)과 집단 내의 변동(각 집단내에서 관측값들의 퍼진 정도)를 측정하는 통계량을 찾아서 검정 통계량을 만들고자 한다.

집단 간의 변동의 크기는 다음과 같은 제곱합(summ of squares)으로 정의하고 이를 처리제곱합 (treatment SS; SS_t)이라고 하자.

$$SS_t = \sum_{i=1}^{I} \sum_{j=1}^{n} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{I} n(\bar{Y}_i - \bar{Y})^2$$

여기서 $\bar{Y}_i$는 i번째 집단의 평균이고 $\bar{Y}$는 모든 자료의 전체 평균을 의미한다. 즉

$$\bar{Y}_i = \frac{\sum_{j=1}^{n} Y_{ij}}{n}, \quad i = 1,2,..,I \qquad \bar{Y} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n} Y_{ij}}{In}$$

집단내의 변동의 크기는 다음과 같은 제곱합으로 정의하고 이를 오차제곱합(error SS; SSE)이라고 하자.

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$$

위의 식에서 볼 수 있듯이 집단 간의 변동의 크기를 나타내는 처리제곱합이 커질수록, 또는 집단내의 변동의 크기를 나타내는 오차제곱합이 작아질수록 귀무가설에 반대되는(즉, 집단 간의 평균이 유의한 차이가 난다는) 증거가 강해진다.

마지막으로 처리제곱합과 오차제곱합을 합한 통계량을 총제곱합(total SS; SST)라고 하고 다음 식이 성립한다.

$$SST = SS_t + SSE \tag{5}$$

위의 분해식에서 각 제곱합은 다음과 같이 정의 된다.

$$SS_t = \sum_{i=1}^{I} n(\bar{Y}_i - \bar{Y})^2$$
$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$
$$SST = \sum_{i=1}^{I}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2$$

이제 모든 집단의 평균이 같은지 아닌지 검정을 할 수 있는 검정 통계량은 다음과 같이 정의하며 이를 F-통계량이라고 한다.

$$F = \frac{SS_t/(I-1)}{SSE/I(n-1)} \tag{6}$$

귀무가설이 옳을 경우 위의 F-통계량은 자유도(Degree of Freedom; Df)가 $I-1$과 $I(n-1)$을 가지는 F-분포를 따른다. 이러한 분포이론을 이용하여 F-통계량으로 구한 p-값을 유의수준 5% 와 비교하여 5%보다 작으면 귀무가설을 기각한다.

$$p - value = P[F(0.95, I-1, I(n-1)) > F]$$

이렇게 전체의 변동을 집단 간의 변동과 집단 내의 변동으로 나누어 집단 간의 평균의 차이를 추론하는 방법을 분산분석(analsys of Variance, ANOVA)이라고 한다.

F-통계량을 정의할 때 편리하고 유용하게 사용되는 것이 다음과 같은 분산분석표(ANOVA table)이다.

Table 1: 분산분석표(ANOVA table)

source	SS	df	MS	F
Trt	SS_t	I-1	$MS_t = SS_t/(I-1)$	$F = MS_t/MSE$
Error	SSE	I(n-1)	$MSE = SSE/(In-I)$	
Total	SST	In-1		

2.3 모수의 추정

이제 모형 (3)에서 모수에 대한 추정을 생각해 보자. 추정해야할 모수는 전체 평균 μ와 각 그룹의 효과 α_i 그리고 분산 σ^2이다. 전체평균과 그룹의 효과는 다음의 오차제곱합(Sum of square error; SSE)을 최소로 하는 모수를 추정하는 최소제곱법(Least Square method; LSE)으로 구할 수 있다.

$$\min_{\mu,\alpha_1,\dots\alpha_I} SSE = \min_{\mu,\alpha_1,\dots\alpha_I} \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij}-\mu-\alpha_i)^2 \tag{7}$$

위의 오차제곱합이 모든 모수에 대하여 미분가능한 이차식으므로 최소제곱 추정량은 제곱합을 모수에 대하녀 미분하고 0으로 놓아 방정식을 풀어서 얻을 수 있다.

오차제곱합을 모수 μ와 α_i, $i=1,2,\dots,I$ 로 미분하여 0으로 놓은 방정식은 다음과 같다.

$$\frac{\partial}{\partial\mu}SSE = -2\sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij}-\mu-\alpha_i) = 0$$
$$\frac{\partial}{\partial\alpha_i}SSE = -2\sum_{j=1}^{n}(y_{ij}-\mu-\alpha_i) = 0, \quad i=1,2,\dots,I$$

위의 방정식을 정리하면

$$\mu + \frac{\sum_{i=1}^{I}\alpha_i}{I} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n} y_{ij}}{In} = \bar{y}$$
$$\mu + \alpha_i = \frac{\sum_{j=1}^{n} y_{ij}}{n} = \bar{y}_i, \quad i=1,2,\dots,I$$

위의 방정식에서 첫 번째 방정식은 다른 I개의 방정식을 모두 합하면 얻어진다. 따라서 모수는 $I+1$ 개이지만 실제 방정식의 개수는 I개이므로 유일한 해가 얻어지지 않는다. 따라서 유일한 해를 구하려면 하나의 부가조건이 필요하며 다음과 같은 두 개의 조건 중 하나를 사용한다.

1. $\alpha_I = 0$ (set-to-zero condition; 마지막 효과 α_I를 0으로 놓는 조건)

   set-to-zero 조건하에서는 다음과 같은 추정량이 얻어진다.

   $$\hat{\mu} = \bar{y}_I, \quad \hat{\alpha}_i = \bar{y}_i - \bar{y}_I, \;\; i=1,2,\dots,I-1, \;\; \hat{\alpha}_I = 0$$

2. $\sum_{i=1}^{I}\alpha_i = 0$ (sum-to-zero condition; 효과의 합은 0이다)

   sum-to-zero 조건에서는 계수의 추정치가 다음과 같이 주어진다.

   $$\hat{\mu} = \bar{y}, \quad \hat{\alpha}_i = \bar{y}_i - \hat{\mu}, \;\; i=1,2,\dots,I$$

이러한 현상은 여러 개의 그룹의 평균을 효과 모형으로 나타냈을 경우 일어난다. 만약에 여러 개의 그룹의 평균을 아래와 같은 평균모형으로 나타낼 경우에는 각 평균 μ_i 는 각 그룹의 표본 평균으로 추정된다.

$$y_{ij} = \mu_i + e_{ij}$$

위의 모형에서 각 그룹 평균에 대한 최소제곱 추정량은 $\hat{\mu}_i = \bar{y}_i$ 이다.

또한 모형에 관계없이 오차항의 분산 σ^2 에 대한 추정량은 다음과 같이 주어진다.

$$\hat{\sigma}^2 = \frac{\sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2}{I(n-1)} \tag{8}$$

3 공분산분석 (Analysis of Covariance)

4 공분산분석 개요

서로 다른 집단을 비교하는 실험이나 관측연구에서 관심이 있는 처리(treatment)나 요인(factor)뿐만 아니라 다른 예측변수들도 반응변수에 영향을 미친다. 이러한 예측변수들의 영향을 제거하기 위한 방법은 여러가지가 있지만 실험인 경우 임의화 방법(randomization)으로 그 영향을 상쇄시킬 수 도 있고 관측연구인 경우에는 사례-대조연구 방법을 이용하여 그 영향을 최소화하려고 노력을 한다. 하지만 많은 경우에 여러 가지 변수들이 반응변수에 영향을 미친다. 이러한 경우에 이러한 독립변수(또는 예측변수)를 모형에 포함시켜서 그 영향을 반영하고 동시에 자료의 변동을 부가적으로 설명해주는 방법이 공분산분석(analysis of covariance; ANCOVA)이다.

공분산 분석에서 고려되는 예측변수를 공변량(covarite)라고 부른다. 대부분의 실험연구에는 실험 전에 여러 가지 점수를 측정하는데 이 경우 이러한 점수를 공변량으로 모형에 포함시켜주는 것이 좋다 (예: 실험 전 상태에 대한 점수, 시험점수, IQ 점수). 또한 임상실험을 여러 개의 병원에서 진행하는 경우 병원 효과를 공변량으로 자주 사용한다.

여기서 주의해야 할 점은 공변량과 처리는 독립이 되야한다는 점이다. 만약 처리의 결과가 공변량에 영향을 미치게 되면 이러한 공변량은 모형에 포함시키는 것이 부적절하다. 예를 들어 자동차정비 교육을 위한 두 가지 학습법을 비교하는 실험을 생각해 보자. 학생들을 임의로 두 가지 학습법중 하나를 선택하여 3개월 동안 교육을 받게 하고 시험을 보아 평균 점수의 차이를 알아보았다. 이 때 공변량으로 총 학습시간을 고려하였는데 학습법의 선택이 총 학습시간에 영향을 줄 수 있다. 즉 고려된 학습법 중 하나는 컴퓨터를 이용한 학습법이며 이 학습법에 배정된 학생들은 컴퓨터 사용를 익히는 시간까지 학습시간에 포함되는 것이 나타났다. 이렇게 공변량이 처리에 의해 영향을 받는 경우(교호작용이 있는 경우)는 이를 모형에 포함시키는 것은 위험하다.

4.1 공분산분석의 모형

공분산분석의 모형은 분산분석 모형 (3)에 공변량 x의 효과를 다음과 같이 더해주는 것이다.

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \quad i = 1, 2, \ldots, I, \;\; j = 1, 2, \ldots, n \tag{9}$$

모형 (9)에서 x_{ij}는 관측값 y_{ij}의 공변량이며 $\bar{x}_{..} = \sum_{i=1}^{I} \sum_{j=1}^{n} x_{ij}$로 공변량의 전체 평균이다. 위의 효과모형은 다음과 같이 평균모형으로 나타낼 수 있다.

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta(x_{ij} - \bar{x}_i) + e_{ij} \\ &= \beta_0 + \alpha_i + \beta x_{ij} + e_{ij} \\ &= \beta_{0i} + \beta x_{ij} + e_{ij} \end{aligned}$$

4.2 모수의 추정

모형 (9)에서 각 모수의 추정은 ANOVA 모형에서와 같이 최소제곱법을 이용하여 추정하며 부가조건 $\sum_i \alpha_i = 0$ 을 이용하면 다음과 같은 추정량을 얻을 수 있다 (교과서 109 페이지 참조).

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\ \hat{\beta} &= \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2} \end{aligned}$$

5 가설검정

ANCOVA 모형에서는 다음과 같은 두 가지 가설을 검정할 수 있다.

ANOVA 모형에서와 같이 각 그룹의 평균에 대한 검정을 할 수 있고

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_I = 0 \quad \text{vesus} \quad H_1 : \text{ not } H_0$$

또한 공변량의 효과에 대한 검정도 할 수 있다.

$$H_0 : \beta = 0 \quad \text{vesus} \quad H_1 : \beta \neq 0 \tag{10}$$

가설검정을 위한 제곱합들을 다음과 같이 정의하자.

$$
\begin{aligned}
S_{xx(i)} &= \sum_{i=1}^{n}(x_{ij} - \bar{x}_{i.})^2 \\
S_{yy(i)} &= \sum_{i=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \\
S_{xy(i)} &= \sum_{i=1}^{n}(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}) \\
S_{xx} &= \sum_{i=1}^{I}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 \\
S_{yy} &= \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 \\
S_{xy} &= \sum_{i=1}^{I}\sum_{j=1}^{n}(x_{ij} - \bar{x})(y_{ij} - \bar{y}) \\
SST &= \sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 \\
SS_t &= \frac{(S_{xx})(S_{yy}) - (S_{xy})^2}{S_{xx}} - SSE \\
SSX &= \sum_{i=1}^{I} S_{yy(i)} - SSE \\
SSE &= \frac{(\sum_i S_{xx(i)})(\sum_i S_{yy(i)}) - (\sum_i S_{xy(i)})^2}{\sum_i S_{xx(i)}}
\end{aligned}
$$

이제 위의 두 검정은 다음과 같은 분산분석표를 이용한 F-검정법으로 수행할 수 있다. $N = In$으로 총 관측값의 개수이다.

Table 2: 공분산분석 모형의 분산분석표

source	SS	df	MS	F
X	SS_X	1	$MSX = SSX/1$	$F_1 = MSX/MSE$
Trt	SS_t	I-1	$MS_t = SS_t/(I-1)$	$F_2 = MS_t/MSE$
Error	SSE	N-I-1	$MSE = SSE/(N-I-1)$	
Total	SST	N-1		

위의 분산분석표에서 공변량 효과에 대한 가설 (10)은 다음과 같이 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(0.95, I-1, N-I-1) > F_1]$$

또한 그룹의 평균에 대한 검정은 ANOVA 검정과 유사하게 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(0.95, 1, N - I - 1) > F_2]$$

위의 두 F-검정에 쓰이는 F-분포의 두 번째 자유도가 ANOVA 검정에서 사용되는 자유도($N - I$)보다 하나가 작음($N - I - 1$)을 유의하자.

교과서 114-115 페이지에서 제 1종의 제곱합과 제 3 종의 제곱합에 대하여 구분하여 설명하였는데 모든 가설검정에서는 제 3 종의 제곱합을 이용하자.

5.1 최소제곱평균(Least Square Mean)과 각 평균의 비교

ANCOVA 모형 (9)에서 각 처리에 대한 평균을 구할 때 공변량의 값에 따라서 그 값이 변한다. 따라서 각 그룹의 평균을 비교하는 경우에는 모형의 공변량에 공변량의 전체 평균을 넣어 사용한다. 이러한 평균을 보정된 최소제곱평균이라고 한다 (교과서 117 페이지)

$$\bar{y}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}\bar{x}_{..}$$

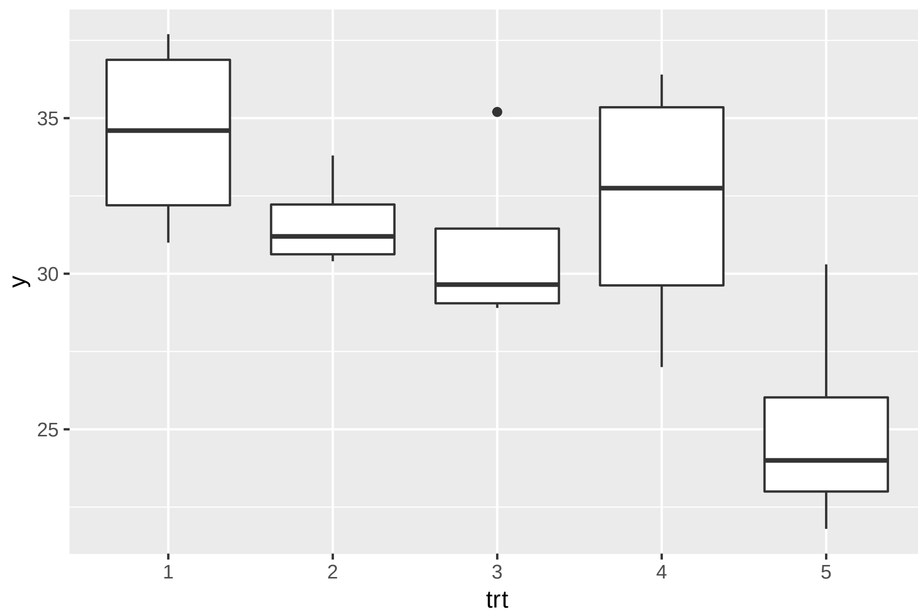
만약 그룹의 평균에 대한 검정 (4)가 기각된 경우 그룹의 평균을 각각 비교할 수 있다.

$$H_0 : \alpha_i = \alpha_j \text{ vs. } H_1 : \alpha_i \neq \alpha_j$$

만약 I 개의 그룹이 있는 경우 모든 쌍을 비교하면 $I(I - 1)/2$ 개의 검정이 필요하며 이러한 비교를 다중비교(multiple comparison)이라고 한다. (교과서 117-118 페이지)

5.2 예제 5.1 저혈당실험

다섯 개의 그룹에 대한 변수 y (1달 후 혈당량 수치)의 상자그림은 다음과 같다.



공변량 x (초기 혈당량 수치)와 y의 관계는 다음과 같은 산점도로 나타낼 수 있으며 강한 양의 선형 관계가 있다는 것을 알 수 있다. 즉 초기 형당량 수치가 크면 1달 후 혈당량 수치도 평균적으로 크다.