베이즈데이터분석 / 이재용 교수

15강

베이즈 통계와 계층 모형

Figures don’t lie, but liars figure.

숫자는 거짓말을 하지 않는다. 하지만 거짓말쟁이는 궁리를 한다.

– *Carroll D. Wright*

한국방송통신대학교

목차

쥐의 종양 자료

- 종양(endometrial stromal polyps)을 가진 암컷쥐들의 수가 기록되었다.
- 현재의 자료에서는 14마리의 쥐 중에 4마리의 쥐가 종양을 보였다.
- 또한 비슷한 70개의 자료가 있다.
- 목표

  목표는 현재의 자료에서 쥐가 종양을 가질 확률 θ를 추정하는 것이다.

자료

```
> rats
    n  x
1  20  0
2  20  0
3  20  0
4  20  0
5  20  0
6  20  0
7  20  0
8  19  0
...
65 20  6
66 20  6
67 52 16
68 46 15
69 47 15
70 24  9
71 14  4
```

첫 번째 분석 : 독립모형

원하는 모수 θ에 관련된 자료는 오직 현재의 자료이므로,

과거의 모든 자료는 무시하고 현재의 자료만 이용해서 분석한다.

- (모형) $x|\theta \sim Bin(n = 14, \theta)$.
- (사전분포) $\theta \sim Beta(1,1) = \mathrm{U}(0,1)$.

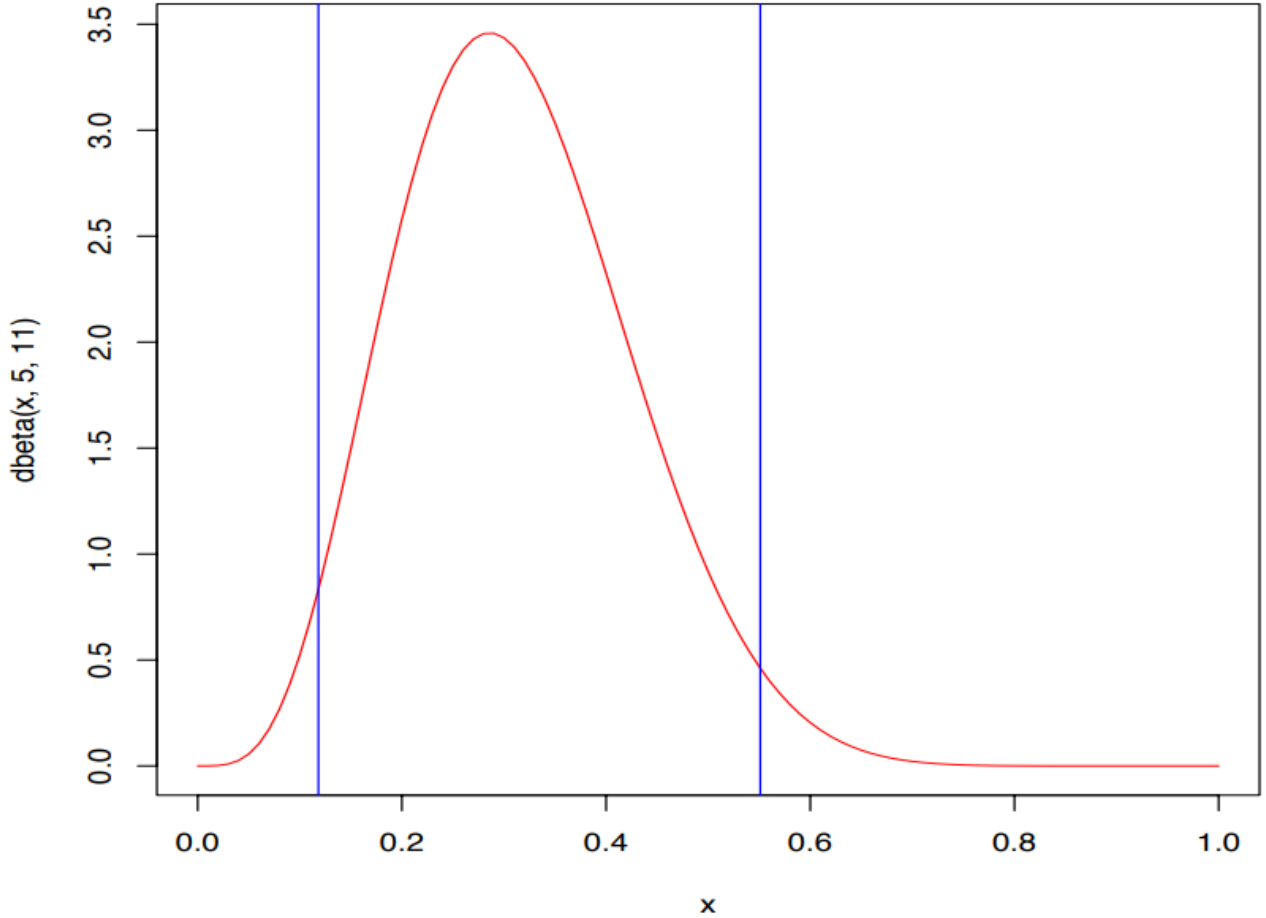
- (사후분포) $\theta|x \sim Beta(x+1, n-x+1) = Beta(5,11)$.

  사후분포의 평균: $\mathbb{E}(\theta|x) = \frac{5}{16} = 0.3125$

  사후분포의 표준편차: $sd(\theta|x) = \sqrt{\frac{5 * 11}{16^2 * 17}} = 0.1124183$

  95% 신용구간 : $(0.1182411, 0.5510032)$.

첫 번째 분석에서 얻은 사후분포의 밀도함수



두 번째 분석 : 통합모형

- 첫 번째 분석의 문제점은 과거의 70개의 자료를 분석에 사용하지 않아서 과거의 자료에 있는 정보를 사용하지 않는다는 것이다.

- 과거의 자료를 사용하기 위해 모든 과거의 자료와 현재의 자료가 동일한 조건에서 얻어진 실험 결과인 것을 고려해서, 모든 자료를 합쳐서 분석한다.

$$x = \sum x_i \, (= 267) \sim Bin(\sum n_i = 1739, \theta).$$

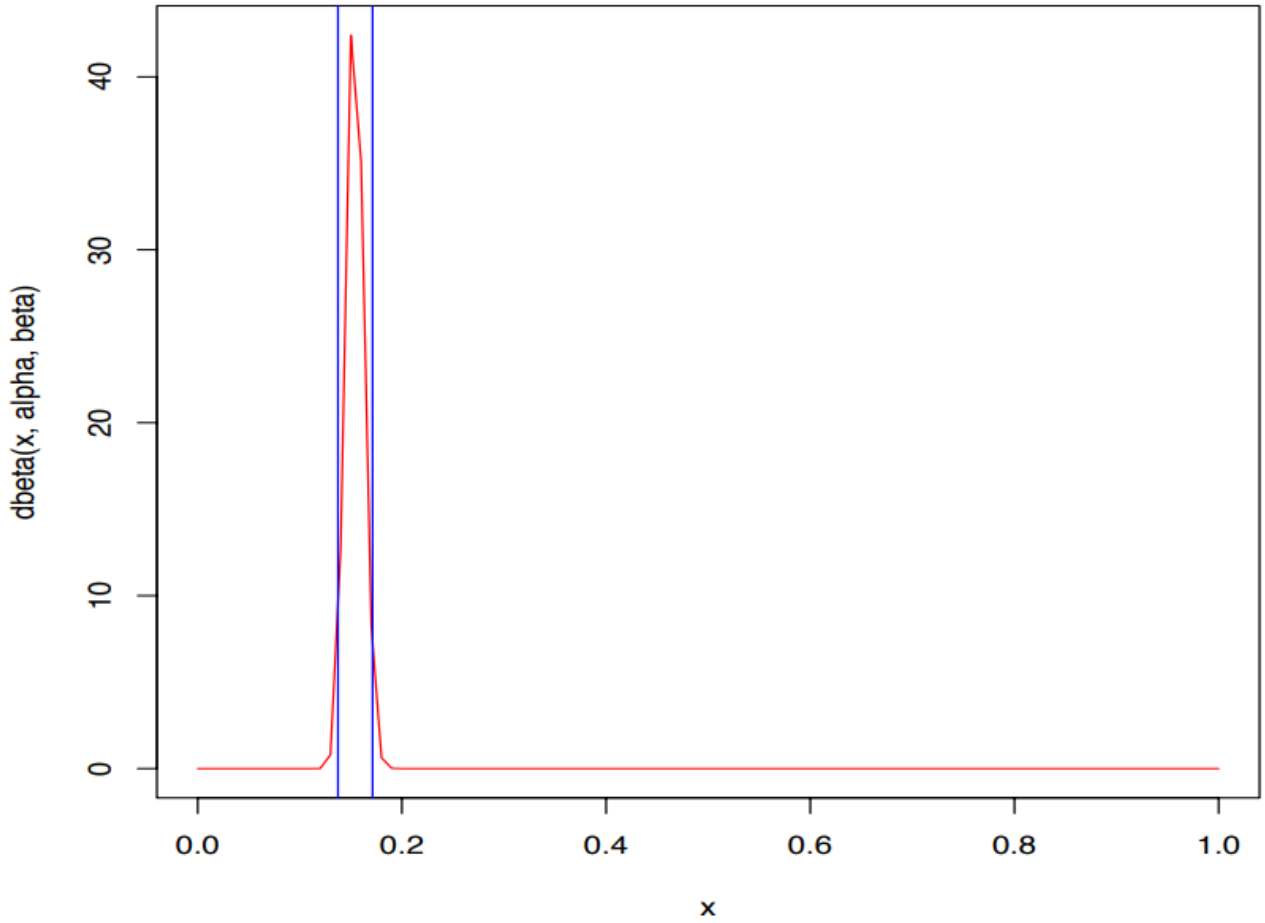
- (모형) $x|\theta \sim Bin(1739, \theta)$.
- (사전분포) $\theta \sim Beta(1,1) = \mathrm{U}(0,1)$.
- (사후분포) $\theta|x \sim Beta(x+1, n-x+1) = Beta(268, 1473)$.

  사후분포의 평균: $\mathbb{E}(\theta|x) = \frac{268}{1741} = 0.1539345$

  사후분포의 표준편차: $sd(\theta|x) = \sqrt{\frac{268 * 1473}{1741^2 * 1742}} = 7.476388e-05$

  95% 신용구간: $(0.1373692, 0.1712529)$.

두 번째 분석에서 얻은 사후분포의 밀도함수



$\theta_1 = \ldots = \theta_{71}$이라고 믿을 수 있는가?

- 두 번째 분석의 문제점은 모든 실험이 엄밀하게 동일한 조건에서 수행되었다고 믿을 수 없는데도 불구하고 71개의 이항실험의 모든 θ_i들이 같은 값을 갖는다고 가정했다는데 있다.

$$x_i \sim Bin(n_i, \theta_i), i = 1, \ldots, 71.$$

$\theta_1 = \ldots = \theta_{71}$이라고 믿을 수 있는가?

이를 알아보기 위해 다음을 정의하자.

$$\bar{\theta} = \frac{\sum x_i}{\sum n_i} = 0.1535365$$

$$\hat{\theta}_i = \frac{x_i + 0.5}{n_i + 1} \quad [1]$$

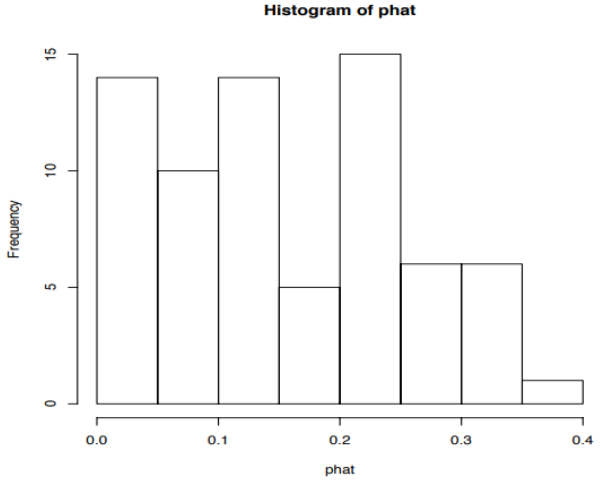
$$z_i = \frac{\hat{\theta}_i - \bar{\theta}}{\sqrt{\hat{\theta}_i (1 - \hat{\theta}_i)/n_i}}.$$

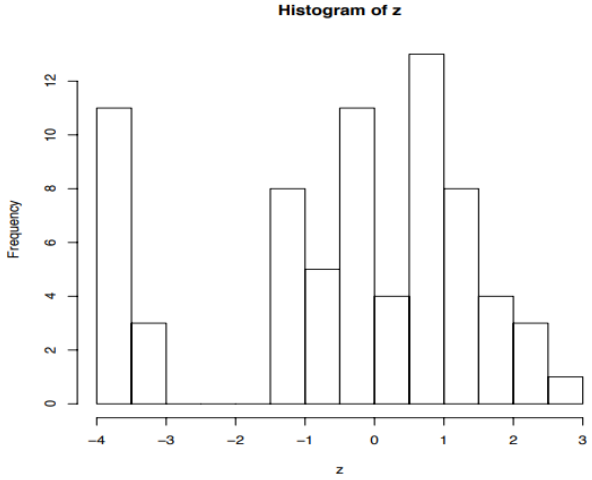
만약 모든 θ_i가 같다면 z_i는 근사적으로 $N(0, 1)$를 따를 것이다.

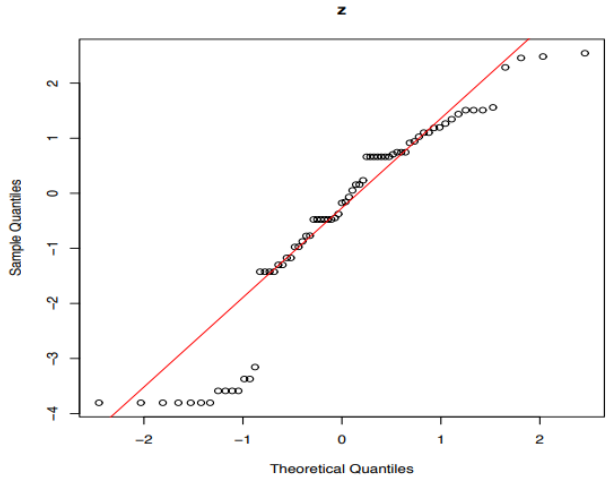
다음은 z_i들의 히스토그램과 정규 Q-Q 그림이다.

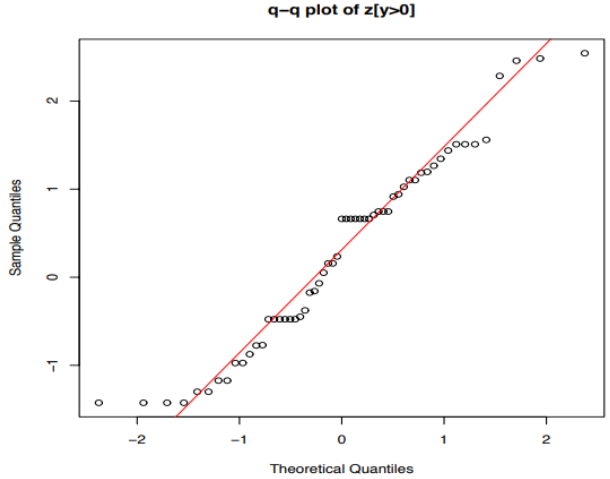
---

1 Instead of x_i/n_i, we use this estimator because there are many 0s.


Histogram of phat
Frequency
phat
0
5
10
15
0.0
0.1
0.2
0.3
0.4


Histogram of z
Frequency
z
0
2
4
6
8
10
12
−4
−3
−2
−1
0
1
2
3


z
Sample Quantiles
Theoretical Quantiles
−4
−3
−2
−1
0
1
2
−2
−1
0
1
2


q−q plot of z[y>0]
Sample Quantiles
Theoretical Quantiles
−1
0
1
2
−2
−1
0
1
2

- 위의 히스토그램과 정규 Q-Q 그림을 보면 z_i들은 $N(0,1)$에서 발생되었다고 보기 힘들다.

- 주요한 이유는 많은 0 때문이다. $x_i = 0$인 자료를 제외한 z_i들의 정규 Q-Q 그림은 직선에 가깝다. 하지만 많은 0을 무시할 수는 없다.

- 첫 번째와 두 번째 분석 모두 만족스럽지 못하다.

세 번째 분석 : 경험적 베이스

과거 70개의 자료를 이용하여 정보를 가진(informative) 사전분포를 구축한다. 70개의 θ_i가 정확히 동일한 값은 갖지 않지만 비슷하다는 가정은 매우 합리적이다. 다음을 가정한다.

$$\theta_i \sim Beta(\alpha, \beta), i = 1, \ldots, 71,$$

위의 가정하에서 α와 β를 추정한다.

과거 70개의 $\hat{\theta}_i$들의 표본평균과 표본표준편차는 각각 0.1524775, 0.00951101 8이다. [2] 아래의 식을 풀어서

$$\frac{\alpha}{\alpha + \beta} = 0.1524$$

$$\frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = 0.009511^2,$$

다음의 값들을 얻는다.

$$\alpha = 1.919$$

$$\beta = 10.66794.$$

---

[2] 이를 위해 추정량 $\hat{\theta}_i = (y_i + 0.5)/(n_i + 1)$이 사용되었다.

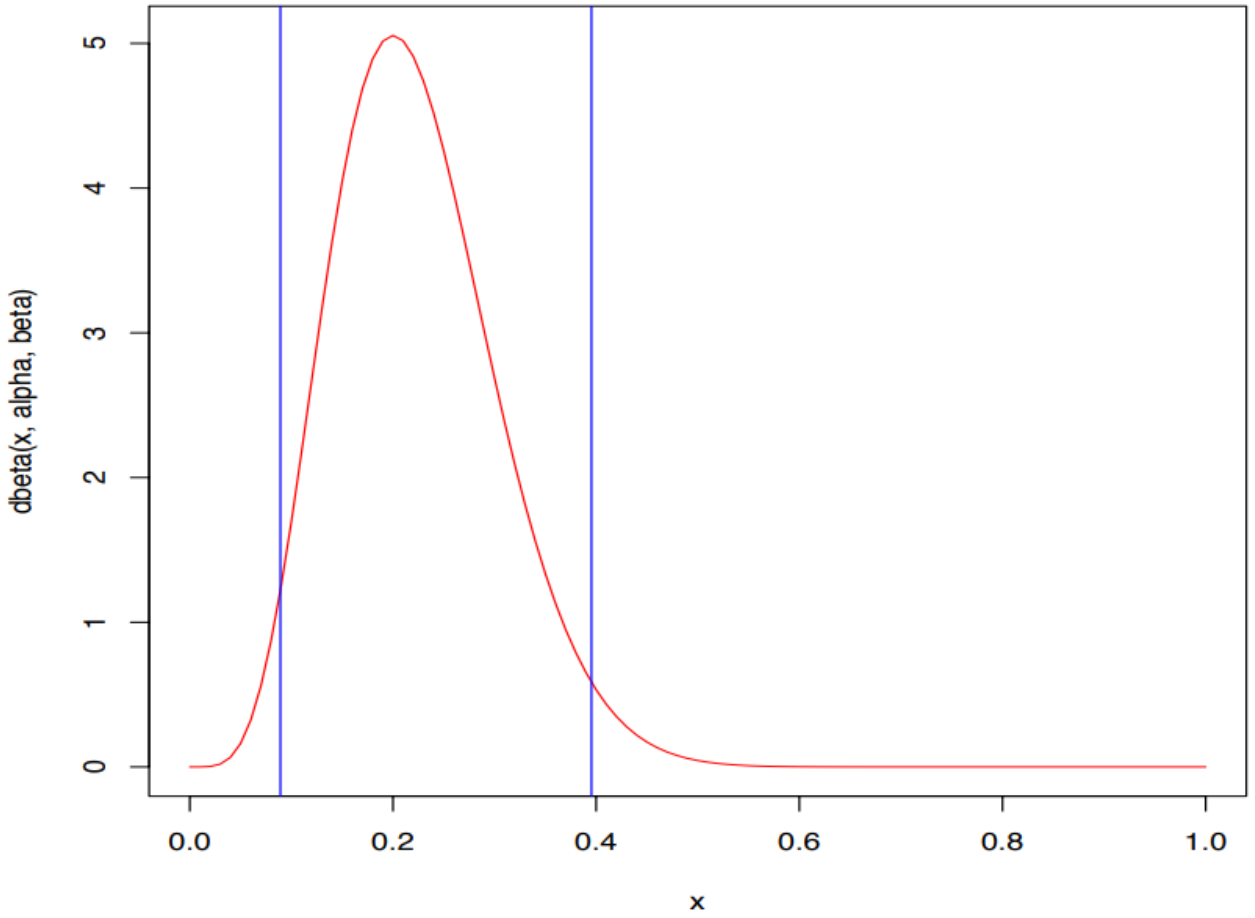
- (모형) $x|\theta \sim Bin(n = 14, \theta)$.
- (사전분포) $\theta \sim Beta(1.919, 10.66794)$.
- (사후분포) $\theta|x \sim Beta(5.919, 20.66794)$.

  사후분포의 평균: $\mathbb{E}(\theta|x) = 0.2226281$.

  사후분포의 표준편차: $sd(\theta|x) = 0.07920501$

  95% 신용구간 : $(0.0892141, 0.3953855)$.

세 번째 분석에서 얻은 사후분포의 밀도함수



세 번째 분석에 관한 고찰

- 명백하게 세 번째 분석은 첫 번째와 두 번째 분석보다 훨씬 합리적이다. 하지만 세 번째 분석도 몇 가지 문제점을 가지고 있다.
- 만약 우리가 $\theta_i, i = 1, 2, \ldots, 70$에 대해서도 동일한 분석을 원한다면, 각각의 θ_i에 대해 동일한 분석을 다시 해야 한다.
- 분석에서 초모수(Hyperparameter) α와 β를 추정하여 사용하였는데, 마치 이들이 고정된 값인 것처럼 사용하였다. 즉, 초모수의 추정에서 유래하는 불확실성을 무시하였다.
- 초모수 α와 β를 추정해야 하기 때문에, α와 β에 사전분포를 거는 것이 자연스럽다.

- 우리는 주어진 모든 자료를 사용하고 싶다. θ_i들이 서로 다른 값이기는 하지만 비슷한 값이라고 생각한다.

- 이를 위해서, θ_i들은 하나의 확률분포에서 발생하였고 이 분포는 비슷한 실험에서 발생하는 모든 가능한 θ_i들의 분포라고 가정한다.

$$\theta_i \overset{iid}{\sim} \pi(\cdot).$$

- θ_i들이 한 개의 분포에서 나왔다는 것은 이 값들이 서로 값은 다르지만 비슷하다는 것을 표현한다.

- 위의 가정은 교환가능성(Exchangeability)에 근거했다고 볼 수도 있다. 교환가능성은 자료를 보기 전에 θ_i들 간에 서로 구별하는 어떤 정보도 없을 때 타당한 가정이다.

계층모형(Hierarchical Model)

- (모형) x_i에 관한 모형은 전과 동일하다.

$$x_i|\theta_i \sim Bin(n_i, \theta_i), i = 1, \ldots, 71.$$

- (사전분포) θ_i는 한 개의 분포에서 발생하였고,
  이 분포는 모든 가능한 θ_i들의 분포를 나타낸다.

$$\theta_i \sim Beta(\alpha, \beta), i = 1, \ldots, 71.$$

- (초사전분포) 초모수 α와 β를 모르므로, 이들에 사전분포를 건다.

$$(\alpha, \beta) \sim \pi(\alpha, \beta).$$

  여기서는

$$\mu = \frac{\alpha}{\alpha + \beta} \sim U(0, 1)$$

$$\nu = \log(\alpha + \beta) \sim Logistic(0, 1)$$

  를 추천한다.

최종 모형 : 계층모형

- (모형) $x_i|\theta \sim Bin(n_i, \theta_i), i = 1, \ldots, 71.$
- (사전분포) $\theta_i \sim Beta(\alpha, \beta), i = 1, \ldots, 71.$
- (초사전분포)

$$\mu = \frac{\alpha}{\alpha + \beta} \sim U(0,1)$$

$$\nu = \log(\alpha + \beta) \sim Logistic(0,1).$$

쥐종양 자료: 계층모형

자료 준비

```
x = rats$x
n = rats$n
k = length(x)
data = list(x=x, n=n, k=k)
```

스탠 수행

```
fit3 = stan(model_code=rats3, data=data, seed=1234567,
        chains=2, iter=5000, thin=1)
```

사후분석

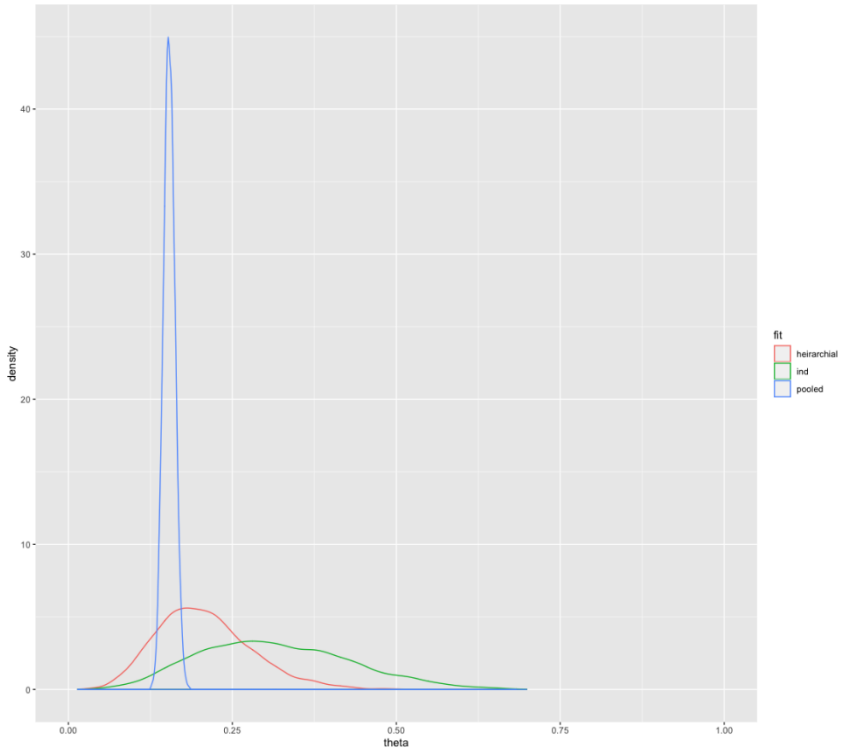
```
print(fit3)

plot(fit3, plotfun="plot", pars=c("theta"))
plot(fit3, plotfun="plot", pars=c("alpha", "beta"))
plot(fit3, plotfun="dens", pars=c("theta[71]", "alpha",
        "beta"))
plot(fit3, plotfun="hist", pars=c("theta[71]", "alpha",
        "beta"))
plot(fit3, plotfun="trace", pars=c("theta[71]", "alpha",
        "beta"))
plot(fit3, plotfun="ac", pars=c("theta[71]", "alpha",
        "beta"))
```

쥐종양 자료: 계층모형

스탠 코드

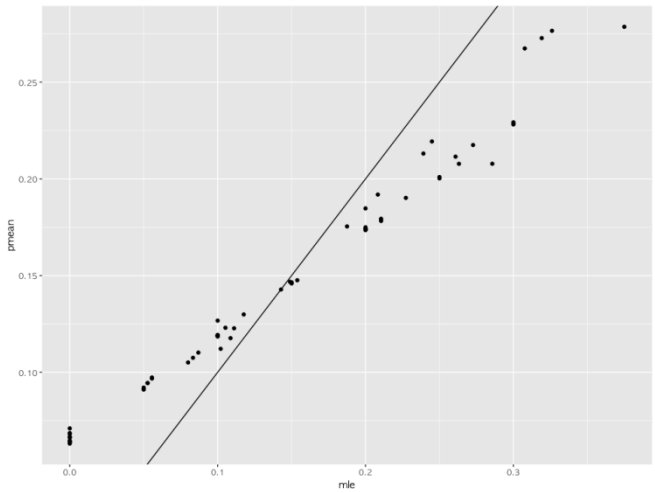
```
rats3 ="
data {
    // data
    int<lower=0> k;
    int<lower=0> x[k];
    int<lower=0> n[k];
}

parameters {
    real<lower=0, upper=1> theta[k];

    real<lower=0, upper=1> mu;
    real nu;
}

transformed parameters {
    real<lower=0> alpha;
    real<lower=0> beta;

    alpha = mu*exp(nu);
    beta = (1-mu)*exp(nu);
}

model {
    for(i in 1:k) {
        x[i] ~ binomial(n[i], theta[i]);
        theta[i] ~ beta(alpha, beta);
    }
    mu ~ uniform(0,1);
    nu ~ logistic(0, 1);
}
"
```

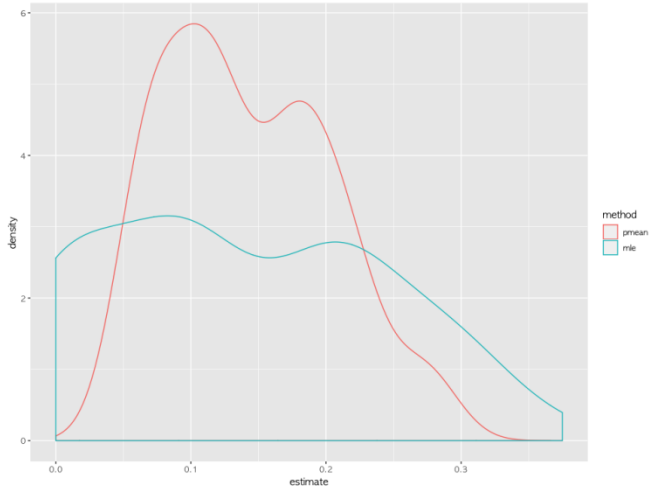
세 모형의 비교



모형	사후 평균	사후표준 평균	Q02.5	사후 중앙값	Q97.5
독립 모형	0.315	0.115	0.121	0.307	0.557
통합 모형	0.154	0.00862	0.137	0.154	0.171
계층 모형	0.213	0.0765	0.0861	0.207	0.382

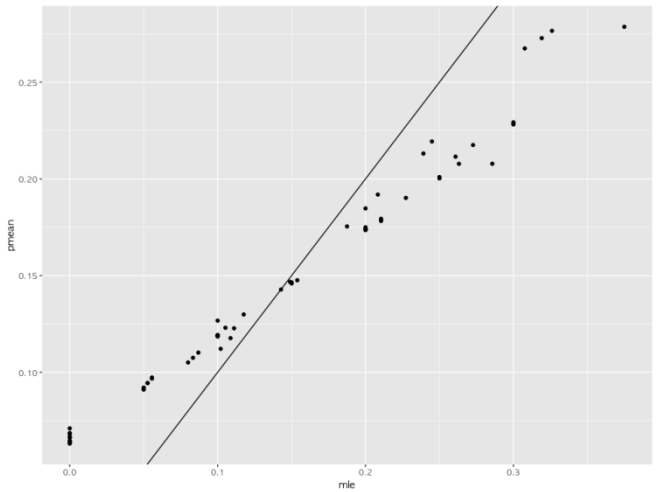
계층모형의 축소효과

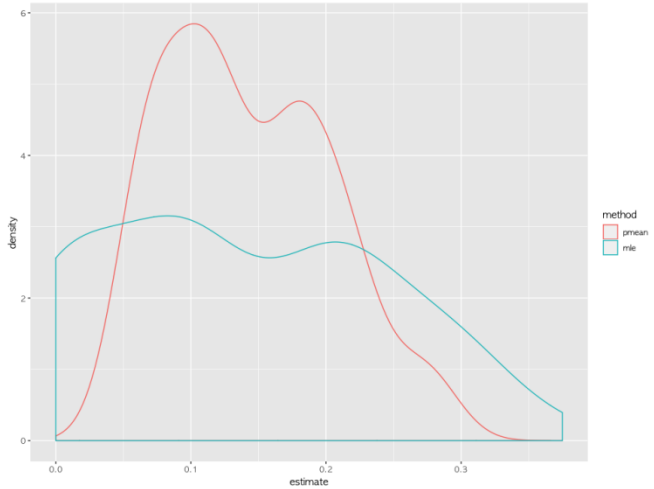




- 왼쪽 그림은 각 실험들의 최대가능도추정량 대 사후평균의 산점도이다. 직선은 $y = x$을 나타낸다. 만약 최대가능도추정량과 사후평균이 같다면 모든 점들이 이 직선 위에 있어야 한다. 최대가능도추정량이 큰 값들은 사후평균도 크지만 최대가능도추정량 보다는 작다. 최대가능도추정량이 작은 값들은 사후평균도 작지만 최대가능도추정량 보다는 크다.

계층모형의 축소효과





- 이를 축소효과(shrinkage effect)라 한다.
- 오른쪽 그림은 최대가능도 추정량들의 밀도함수와 사후평균의 밀도함수 그림이다. 사후평균의 밀도함수가 좀 더 중앙으로 집중되어 있다.

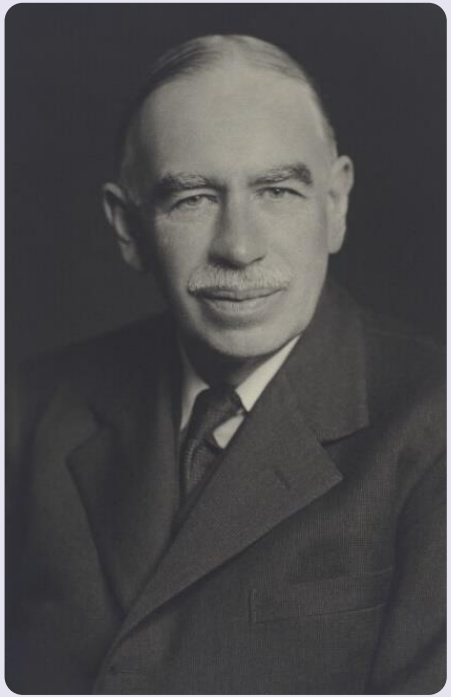
$$\hat{\theta}_i^{hier} \approx \alpha\hat{\theta}_i^{indep} + (1-\alpha)\bar{\theta}, \alpha \in (0,1).$$

계층모형의 장점

- 계층모형을 사용하지 않으면 쥐의 종양 자료 분석에서 봤듯이 71개의 데이터를 너무 적은 수의 모수로 표현하거나(모형2) 혹은 너무 많은 수의 모수로 표현하여(모형1) 예측력이 떨어지게 된다. 계층모형은 많은 수의 모수를 쓰지만 이 모수들이 하나의 분포를 따른다는 가정으로 모수의 자유도를 제한한다.
- 주변 그룹의 정보를 이용 (borrowing information from the neighbors).
  한 그룹의 자료만으로 그 그룹의 모수를 추정하는 것이 아니라
  비슷한 주변 그룹의 정보도 이용하여 그룹의 모수들을 추정한다.
- 평균으로의 축소효과(shrinkage effect)
  주변의 정보를 이용하는 형태는 너무 큰 추정치는
  약간 작게, 너무 작은 추정치는 약간 크게하는 경향이 있다.

계층모형의 일란성 쌍둥이들

- 경험적 베이즈 모형(empirical Bayes method)
- 다층 모형(multilevel model)
- 소지역 추정(small area estimation)



When the facts change,
I change my opinion.
What do you do, Sir?

사실이 바뀌면,
나는 의견을 바꿉니다.
당신은 어떤가요?

- John Maynard Keynes

감사합니다.