- we may've multiple params. that can be related → joint prob. model of these params.
- e.g: effectiveness of cardio. treatments
  - patients in hospital j have surviving prob. θ_j
  - est. of the θ_j's should be related
- y_{ij} can be used to est. aspects of the pop. dist. of the θ_j's
- these models have enough params. to fit the data well

Constructing a Parameterized Prior Dist.

- Case: - est. a parameter θ using data from small experiments.
  - a prior dist. constructed from historical experiments

- Risk of tumor in a group of rats

Previous experiments: 70 experiments

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Mean: 0.136
Std. Dev: 0.103

Current experiment: 71st
4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* $\frac{y_j}{n_j}$ *(number of rats with tumors)/(total number of rats).*

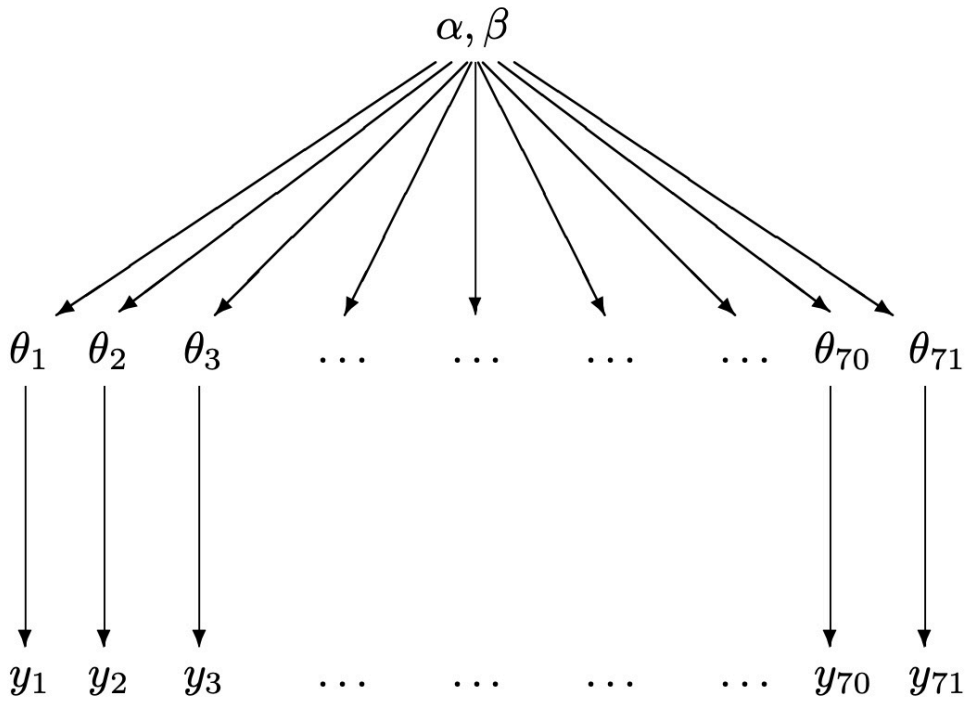
- goal: est. θ → prob. of tumor in a pop. of rats recieving a drug
  - $y \sim Bin(n, \theta)$
  - $\theta \sim Beta(\alpha, \beta)$

- Analysis with a fixed prior:
  - from historical data we know $\theta \sim Beta(\alpha, \beta)$, with known mean and std. dev.
  - $p(\theta|y) \propto \underbrace{p(y|\theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta|\alpha,\beta)}_{\text{prior}}$

$$\propto \underbrace{\left(\theta^4 (1-\theta)^{14-4}\right)}_{\text{current experiment}} \left(\theta^{\alpha-1} (1-\theta)^{\beta-1}\right)$$

$$\propto \left(\theta^{\alpha+4-1} (1-\theta)^{\beta+10-1}\right) \sim Beta(\alpha+4, \beta+10)$$

- Approx. Est of pop. dist. using historical data :
  - $y_j \sim Bin(n_j, \theta_j)$
  -



- $\alpha + \beta = \frac{E(\theta)(1-E(\theta))}{Var(\theta)} - 1$

  $= 10.0759$

- $\frac{\alpha}{10.0759} = 0.136 \Rightarrow \alpha = 1.37$, $\beta = 8.7$

- $(\alpha, \beta) = (1.4, 8.7)$
  - not a Bayesian calc.
  - not based on any full prob. model

  - using historical pop. dist. as a prior dist. for the current experiment yields a Beta(5.4, 18.7) posterior dist. for θ_{71}
    - mean: $0.224 < \left(\frac{4}{14} = 0.286\right)$
      - weight of experiance indicates # of tumors in current experiment is high

- it makes sense to est. pop. dist. from all of the data and help est. each θ_j than est. all 71 values θ_j seperately

---

Exchangeability and Setting Up Hierarchical Model

- set of experiments: $j = 1, \ldots, J$
- experiment j has data vector y_j and parameter vector θ_j
  - some params. in diff. experiments may overlap
  - to create a joint probability model for all params θ, we use exchangeability → (L1 : 25)

Exchangeability

- if we cannot distinguish θ_j's, the params. $(\theta_1, \ldots, \theta_J)$ are exchangeable
  - e.g. we roll a die and assign equal prob. to all outcomes; later on we may notice imperfections (eliminate symmetry)
  - $p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$ → ϕ is unknown

- $p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j | \phi) \right) p(\phi)\, d\phi$

- prob. of a die landing on each of its 6 faces:
  - $\theta_1, \ldots, \theta_6$ are exchangeable → $\sum_{j=1}^{6} \theta_j = 1$
  - cannot be modelled as iid

- select 8 states and record divorce rates/1000 people → $y_1, \ldots, y_8$
  - what can you say about y_8?
    - divorce rate in 7 states: 5.8, 6.6, 7.8, 5.6, 7.0, 7.1, 5.4
    - a reasonable posterior predictive dist. would be centered around 6.5 (mean)
    - y_j are exchangeable but not indep. b/c we assume 8_{th} state is similar to observed states

- the 8 states are Mountain States:
  - you've still not told which observed rate belongs to which state
  - prior dist. (before seeing data) changes; you assume:
    - Utah has a lower divorce rate
    - Nevada has a higher divorce rate
  - when you see the 7 observed vals. (they're close), you assume:
    - 8_{th} missing value is Utah or Nevada
  - y_8 is Nevada, so: $p(y_8 > \max(y_1, \ldots, y_7) \mid y_1, \ldots, y_7)$ is large

---

Exchangeability when additional info. is aval. on units

- if obs. can be grouped, we can make a hierarchical model
  - each group has its own submodel with properties unknown
  - if group properties are exchangeable → use a common prior dist.
- if y_i has additional info. x_i, (y_i, x_i) are exchangeable
  - we make a conditional model for $y_i | x_i$

- in rat tumors experiment, y_j is exchangeable → no additional knowledge of experiment conditions is given
  - if certain experiments happened in different labs → partial exchangeability

---

Full Bayesian Treatment of Hierarchical Models

- $p(\phi, \theta) = p(\theta|\phi) \cdot p(\phi)$
- joint posterior: $p(\phi, \theta|y) \propto p(y|\phi, \theta)\ p(\phi, \theta)$

  hyperparameter

  $\propto p(y|\theta)\ p(\phi, \theta)$ → ϕ only affects y through θ

Hyperprior Dist.

- start with a simple non-informative prior on ϕ
  - in rat tumor, hyperparameters (α, β) determine Beta dist. of θ

Fully Bayesian Analysis of Conjugate Hierarchical Models

1. write joint posterior density, $p(\phi, \theta|y)$, as a product of
   $\underbrace{p(\phi)}_{\text{hyperprior dist.}}$, $\underbrace{p(\theta|\phi)}_{\text{population dist.}}$, $\underbrace{p(y|\theta)}_{\text{likelihood dist}}$
2. find $p(\theta|\phi, y) = \prod_{j=1}^{J} p(\theta_j|\phi, y)$
3. find $p(\phi|y)$ using Bayesian paradigm → $\frac{p(\phi, \theta|y)}{p(\theta|\phi, y)}$ ① / ②

Drawing simulations from posterior dist ①

1. draw vector of hyperparams from $p(\phi|y)$
2. draw vector of params. from $p(\theta|\phi, y)$
3. ~~draw $\tilde{y}$ from $p(y|\theta)$~~
4. perform L times

- Rat tumors:
  - assign non-informative hyperprior dist

1. $p(\alpha, \beta, \theta|y) \propto p(\alpha, \beta) \cdot p(\theta|\alpha, \beta) \cdot p(y|\theta)$

   $\propto \underline{p(\alpha, \beta)} \cdot \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}} \cdot \prod_{j=1}^{J} \underbrace{\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}}$

   ↘ $\text{Beta}(\alpha+y_j, \beta+n_j-y_j)$ ↙
2. $p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$
3. $p(\alpha, \beta|y) \propto \underline{p(\alpha, \beta)} \cdot \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$

- we could use a reference prior on $\left(\frac{\alpha}{\alpha+\beta}, \alpha+\beta\right)$ which will be dominated by the likelihood
  - $\frac{\alpha}{\alpha+\beta}$: mean
  - $\alpha+\beta$: prior effective sample size
- we transform $p\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right) \propto 1$ to original scale
  - $u = \frac{\alpha}{\alpha+\beta}$ and $v = (\alpha+\beta)^{-1/2}$
  - b/c we've using a uniform prior on (u,v) so $p_{\alpha,\beta}\left(h_1(u,v), h_2(u,v)\right) = 1$
  - $J(\alpha,\beta) = \begin{bmatrix} \frac{\partial u}{\partial \alpha} & \frac{\partial u}{\partial \beta} \\ \frac{\partial v}{\partial \alpha} & \frac{\partial v}{\partial \beta} \end{bmatrix} \rightarrow |J| = (\alpha+\beta)^{5/2} \rightarrow |J|^{-1} = (\alpha+\beta)^{-5/2}$ (|J|: determinant)
  - $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$

Estimating exchangeable params. from a normal model

- observed data follows a normal dist.
  - diff. mean for each group
  - normal pop. dist. for group means

Data Structure

- we've J indep. experiments
  - experiment j est. param. θ_j from n_j iid data points y_{ij}, with known var. σ^2
  - $y_{ij} | \theta_j \sim N(\theta_j, \sigma^2)$ $\quad i = 1,\dots,n_j \quad j = 1,\dots,J$
    - $\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ - sample mean
    - $\sigma_j^2 = \sigma^2/n_j$ - sample var
  - we can write: $\bar{y}_{\cdot j} | \theta_j \sim N(\theta_j, \sigma_j^2)$

Constructing a prior dist.

- maybe est. θ_j by pooled est. $\bar{y}_{\cdot\cdot} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}\bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$ (or $\bar{y}_{\cdot j}$?)
  - ↳ not reliable if n_j is small.
  - which to use? ANOVA

	df	SS	MS	$E(MS\|\sigma^2,\tau)$
Between groups	$J-1$	$\sum_i\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	$SS/(J-1)$	$n\tau^2+\sigma^2$
Within groups	$J(n-1)$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot j})^2$	$SS/(J(n-1))$	σ^2
Total	$Jn-1$	$\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot\cdot})^2$	$SS/(Jn-1)$	

($\bar{y}_{\cdot\cdot}$ / $\bar{y}_{\cdot j}$: between groups / within groups)

- $n_j = n$
- $\sigma_j^2 = \sigma^2/n$

- τ^2 is the variance of $\theta_1, \ldots, \theta_J$ (group means)
- if $\frac{MS_1}{MS_2} > 1$ use $\hat{\theta}_j = \bar{y}_{\cdot j}$ → groups are meaningfully diff.

- alternatively, use a weighted combination:

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$$

- $\lambda_j \in [0,1]$

Hierarchical Model

- we assume $p(\theta_1, \ldots, \theta_J \mid \mu, \tau) = \prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2)$

$$p(\theta_1, \ldots, \theta_J) = \int \left[\prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2) \right] p(\mu, \tau)\, d\mu\, d\tau$$

$$p(\mu, \tau) = p(\mu \mid \tau)\, p(\tau) \propto p(\tau)$$

Joint Posterior Distribution

- $p(\theta, \mu, \tau \mid y) \propto p(\mu, \tau) \cdot p(\theta \mid \mu, \tau) \cdot p(y \mid \theta)$

$$\propto p(\mu, \tau) \cdot \prod_{j=1}^{J} N(\theta_j \mid \mu, \tau^2) \cdot \prod_{j=1}^{J} N(\bar{y}_{\cdot j} \mid \theta_j, \sigma_j^2)$$

Conditional Posterior Dist Given Hyperparams.

- $\theta_j \mid \mu, \tau, y \sim N(\hat{\theta}_j, V_j)$

- $\hat{\theta}_j = \dfrac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$ $\qquad V_j = \dfrac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$

Marginal Posterior Dist. of Hyperparams.

- $p(\mu, \tau \mid y) \propto p(\mu, \tau)\, p(y \mid \mu, \tau)$ → tough too derive
  - for a hierarchical normal model, we consider info. supplied by data about hyperparams.

$$\propto p(\mu, \tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} \mid \mu, \sigma_j^2 + \tau^2)$$

- $p(\mu, \tau | y) \propto p(\mu | \tau, y) \cdot p(\tau | y)$ (1) (2)

(1) $\mu | \tau, y \sim N(\hat{\mu}, V_\mu)$

- $\hat{\mu} = \dfrac{\sum \frac{1}{\sigma_j^2 + \tau^2} \bar{y}_{\cdot j}}{\sum \frac{1}{\sigma_j^2 + \tau^2}}$ $\qquad V_\mu^{-1} = \sum \frac{1}{\sigma_j^2 + \tau^2}$

(2) $p(\tau | y) = \dfrac{p(\mu, \tau | y)}{p(\mu | \tau, y)} \propto \dfrac{p(\tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu} | \hat{\mu}, V_m)}$

$\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(- \dfrac{(\bar{y}_{\cdot j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right)$

- $p(\tau) \propto 1$

Computation

- computing posterior dist of θ best done by simulation:

$$p(\theta, \mu, \tau | y) = p(\tau | y) \; p(\mu | \tau, y) \; p(\theta | \mu, \tau, y)$$ (2) (1) (3)

Ex. Parallel Experiments in 8 Schools

- analyze effects of coaching programs on test scores
  - seperate randomized experiments weve done to see effects of coaching on SAT-V scores
  - score range: [200, 800], mean: 500, SD: 100
  - all 8 schools assumed its coaching programs to be successful @ inc. scores
  - no prior belief one program is better, or some were similar in effect

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18
	$\bar{y}_{\cdot j}$	σ_j