Power Tools for Epidemiologists

Charles DiMaggio

Departments of Anesthesiology and Epidemiology
College of Physicians and Surgeons
Columbia University
New York, NY 10032

cjd11@columbia.edu

P9489
applications of epidemiologic methods II
Spring 2014

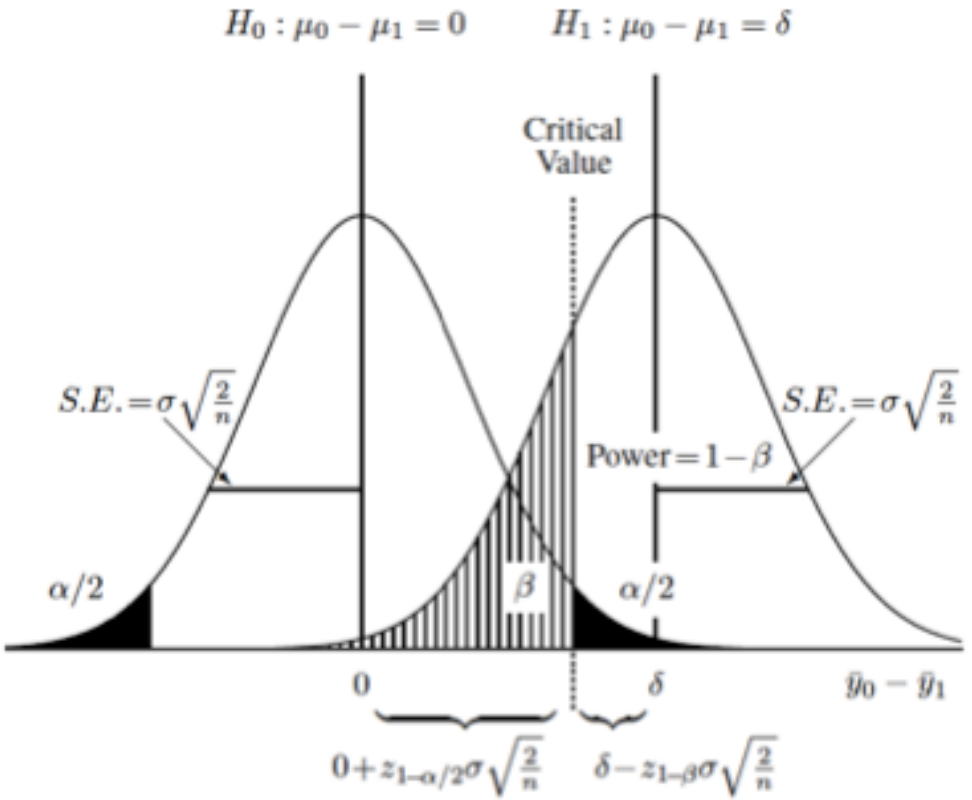
Outline

1. about power
2. Lehr's equation
3. poisson distributed or count data
   - relative risks and odds ratios
4. binomial data or proportions
5. Bayesian approaches to power calculations
   - "adaptive" calculations
6. Bayesian Sample Sizes

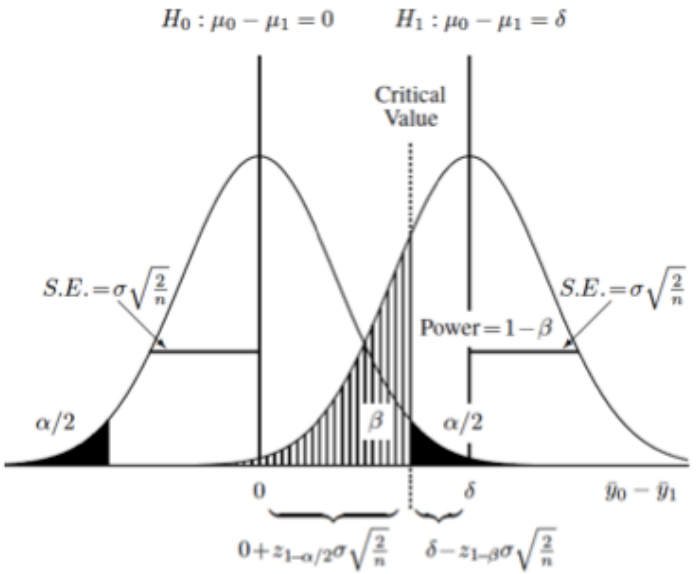
Acknowledgements

Gerald van Belle "Statistical Rules of Thumb"
Andrew Gelman and Jennifer Hill's "Data Analysis Using Regression and Multilevel/Hierarchical Models" (Chapter 20)
Brad Carlin
JoAnn Alvarez

difference between two groups on continuous measurement



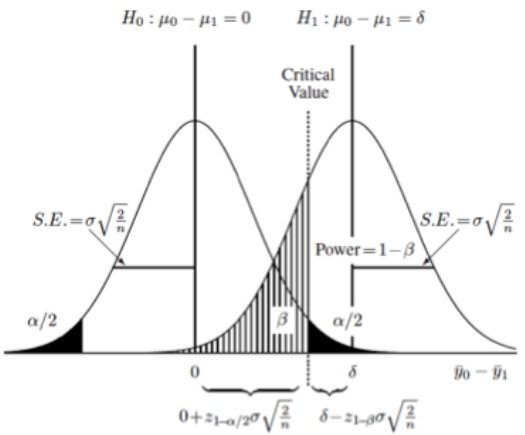
The Recipe



- $\Delta = y_0 - y_1$ - Difference
- $\mu_\Delta = 0$ - No mean difference (Curve on the left)
- $\mu_\Delta = \delta$ - Appreciable mean difference (Curve on the right)
- σ^2 - Variances equal
- $se_\Delta = \sigma * \sqrt{\frac{2}{n}}$ - *relationship between s.e. of differences in means and s.d. allows us to set up calculations for sample size, n.*

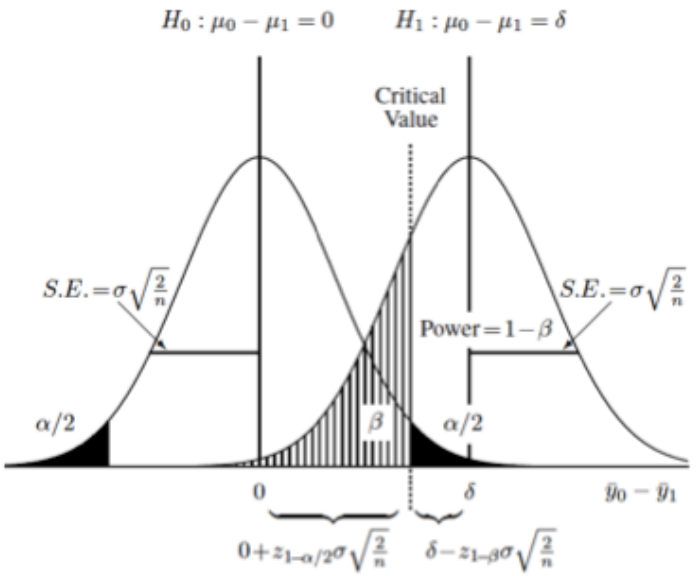
"Classic" Formulation

conditional probabilities



- type I error or $\alpha = \Pr[+ \text{ result} \mid \text{- difference}]$, ie. false positive
  - the black-shaded 0.05/2 tails on the null hypothesis
- type II error or $\beta = \Pr[\text{- result} \mid + \text{ difference}]$, i.e. false negative
  - the hatched 0.2 left-handed tail on the alternative hypothesis
- power or $1 - \beta = \Pr[+ \text{ result} \mid + \text{ difference}]$, i.e. true positive
  - the area to the right of β on the alternative hypothesis

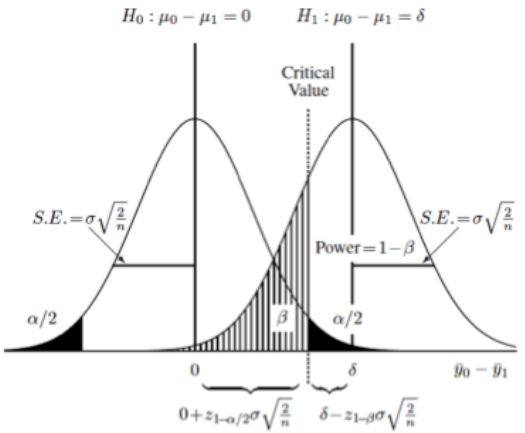
Two Key Ingredients



- difference between the two groups
- variability of the measurements

more power

(without a larger sample...)



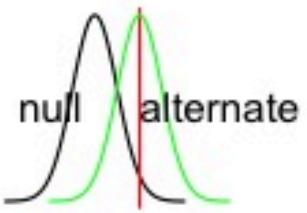
- Increase AUC to the right of the critical value
- Measure your outcome more precisely
  - "shrinks" curves towards means
- study a greater effect size
  - increases difference between $\mu_\Delta = 0$ and $\mu_\Delta = \delta$, separates the curves more

relationship of delta to power

critical value

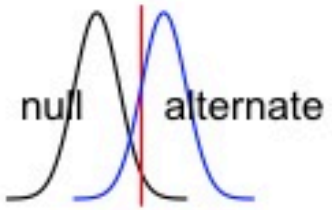
delta

50% Power



delta

80% Power



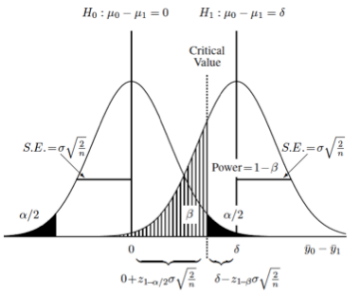
delta

99% Power

delta

The Critical Value

the secret ingredient for the sample size calculation



- where the upper value of α on the null curve meets the value for β on the alternative curve
- point at which you accept or reject your null hypothesis

$$0 + z_{1-\alpha/2} * \sigma\sqrt{\frac{2}{n}} = \delta - z_{1-\beta} * \sigma\sqrt{\frac{2}{n}}$$

- Solve for n

$$n_{group} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\frac{\mu_1 - \mu_2}{\sigma})^2}$$

Lehr's Equation

Difference Between Two Means

- Substitute the usual values of 1.96 for α and 0.84 for β
- denominator is the square of the standardized difference (Δ^2)

$$n_{group} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\frac{\mu_1 - \mu_2}{\sigma})^2}$$

$$n_{group} = \frac{2(1.96 + 0.84)^2}{(\Delta)^2}$$

$$n_{group} = \frac{16}{(\Delta)^2}$$

10-point difference in IQ between two groups

μ=100 vs. 90, σ^2=20

- $$n_{group} = \frac{16}{((100-90)/20)^2}$$
$$n_{group} = \frac{16}{(.5)^2}$$
- total sample size of 128.
- function in R:

```
nMeans<-function(x0, x1, sd){
    d<-(x1-x0)/sd
    n<-16/d^2
    n
    }
nMeans(100,90,20)
```

Evaluate range of values

- μ between 2 and 10 IQ points

```
unexp<-100
exp<-seq(88,98,2)
nMeans(unexp, exp, 20)
```

- σ^2 between 10 and 20 IQ points

```
dev<-seq(10,20,2)
nMeans(100, 90, dev)
```

- Combination of μ and σ^2

```
for(i in seq(10,20,2)){
    size<-nMeans(unexp,exp, i)
    print(size)
}
```

Detectible Difference

- Rearrange Lehr's equation to solve for the detectible *standardized* difference between two populations:

$$n_{group} = \frac{16}{(\Delta)^2}$$

$$(\Delta)^2 = \frac{16}{n_{group}}$$

$$\Delta = \frac{4}{\sqrt{n_{group}}}$$

- *unstandardized* difference

$$\delta = \frac{4\sigma}{\sqrt{n_{group}}}$$

R functions for detectible difference

```
nDD<-function(sd, n){
    dd<-4*sd/(sqrt(n))
    dd
    }
```

- Detectible difference in mean IQ with a sample size of 50

```
nDD(20,50)
```

- about 12 IQ points.

Percentage Change in Means

- $$n_{group} = \frac{16(c.v.)^2}{(ln(\mu_0) - ln(\mu_1))^2} = \frac{16(c.v.)^2}{(ln(r.m.))^2}$$
- where,
  - $c.v. = \sigma/\mu$
  - r.m - percentage change translated into ratio of means

$$= \frac{\mu_0 - \mu_1}{\mu_0} = 1 - \frac{\mu_1}{\mu_0}$$

- e.g., 20% difference between two groups in data with about 30% variability.
  - 20% change translates to a ratio of means of $1 - .20 = .80$.

$$n_{group} = \frac{16(.3)^2}{(ln(.8))^2} = 29$$

R function for percent change in means

```
nPC<-function(cv, pc){
    x<-16*(cv)^2/((log((1-pc)))^2)
    print(x)
}
```

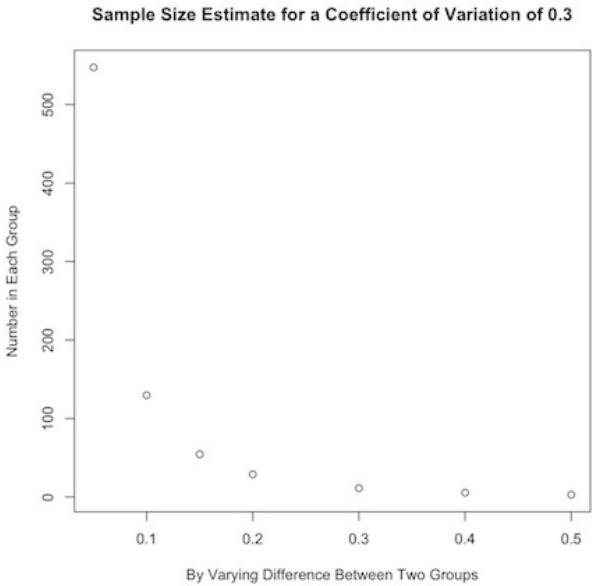
- 15% change from one group to another, but uncertain about how the data varies

```
a<-c(.05,.10,.15,.20,.30,.40,.50,.75,1)
nPC(a,.15)

plot(a,nPC(a,.15),  ylab="Number in Each Group",
xlab="By Varying Coefficent of Variation",
main="Sample Size Estimate for a 15% Difference")
```

Or look at how desired effect size changes sample calculations

```
plot(b, nPC(.3,b), ylab="Number in Each Group",
xlab="By Varying Difference Between Two Groups",
main="Sample for a Coefficient of Variation of 0.3")
```



The square root of poisson-distributed data is approximately normally distributed

- $$y_i \sim Pois(\lambda)$$
$$\sqrt{y_i} \sim Nl(\sqrt{\lambda}, 0.25)$$
- apply Lehr's equation by taking the square root

$$n_{group} = \frac{4}{(\sqrt{\lambda_1} - \sqrt{\lambda_2})^2}$$

```
nPois<-function(x0,x1){
    n<-4/(sqrt(x0)-sqrt(x1))^2
    n
    }
nPois(30,36)
```

demonstrate change in mean count per unit of observation

- mean count of health outcome = 36 per unit of observation (e.g. person-year)
- want to demonstrate decrease to 30 per unit of observation in a comparison population
- $n_{group} = \frac{4}{(\sqrt{30}-\sqrt{36})^2} = 15$ units of observations in each group.
- mean counts per unit time (or volume)
  - mean is $T * \lambda$
  - $n_{group} = \frac{4}{T*(\sqrt{\lambda_1}-\sqrt{\lambda_2})^2}$

high background rates of disease, $\lambda*$

add background rate to the individual means

$$n_{group} = \frac{4}{(\sqrt{\lambda_1 + \lambda*} - \sqrt{\lambda_2 + \lambda*})^2}$$

- e.g. two Poisson sets data with means $\mu_1 = 1$ and $\mu_2 = 2$
- $n_{group} = \frac{4}{(\sqrt{2}-\sqrt{1})^2} = 25$ observations in each group.
- background population rate of 1.5,
- $n_{group} = \frac{4}{(\sqrt{3.5}-\sqrt{2.5})^2} = 50$

```
nbrPois<-function(x,x0,x1){
    n<-4/(sqrt(x+x0)-sqrt(x+x1))^2
    n
    }
nbrPois(1.5,1,2)
```

rate above background for significant result (Δ_λ)

need 4 times the square root of the background rate more outcomes

- set λ_1 to 0 and λ_2 to Δ_λ
- $\Delta_\lambda = 4\sqrt{\lambda *}$
- e.g. usually 50,000 deaths in a population
- need $4\sqrt{50000} = 895$ more deaths to demonstrate the effect of some event

```
nbbrPois<-function(x){
    n<-4*sqrt(x)
    n
    }
nbbrPois(50000)
```

Upper Bound for Count Data when No Events Have Occurred (Rule Of 3)

- given no outcomes, 95% upper bound rate occurrance $= \frac{3}{n}$
- e.g. 20 surgeries, no bad outcomes, $3/20 = 0.15$ or 15% chance of an adverse event possible
- probability at least one is complement of probability of none

$$
\begin{aligned}
Pr[some_1] &= Pr[1 - \forall_0] = 0.05 \\
0.05 &= Pr[\Sigma y_1 = 0] = e^{n\lambda} \\
-ln(0.05) &= n\lambda \\
2.99/n &= \lambda \\
\lambda &\approx 3/n
\end{aligned}
$$

sample to achieve relative risk

- (Again) based on Poisson distribution
- $n_{group} = \frac{4}{\frac{c}{c+d}(\sqrt{RR}-1)^2}$
- reasonably accurate for prevalences under 20%
- E.g. to detect RR at least 3 in study of disease with prevalence of 1% in unexposed
- $n_{group} = \frac{4}{0.01(\sqrt{3}-1)^2} = 746.4$
- increasing prevalence, decreases required sample size
  - can increase period of observation

number of outcomes needed to detect a relative risk

- $$n_{outcomes_{unexp}} = \frac{4}{(\sqrt{RR}-1)^2}$$
$$n_{outcomes_{exp}} = RR * n_{outcomes_{unexp}}$$
- E.g. for RR at least 3
- $4/(\sqrt{3}-1)^2 \approx 8$ outcomes in unexposed
- $3 * 8 = 24$ outcomes in the exposed
- from rearranging the previous formula
  - i.e., rates of $1/100$ in exposed and $3/100$ in unexposed require 747 people in each group
- number of *outcomes* important, not necessarily number of exposed and unexposed.

log-based formula

- $n_{group} = \frac{8(RR+1)/RR}{\frac{c}{c+d}(ln(RR))^2}$
- E.g. unexposed prevalence 1%, to achieve RR 3
  - $n_{group} = \frac{8(4)/3}{0.01(ln(3))^2} \approx 884$
- log-based formula more conservative than Poisson-based
- ... and allows easier estimation for odds ratio

odds ratios

log based

- $n_{group} = \frac{8\sigma^2_{ln(OR)}}{(ln(OR))^2}$
  where,
- $*\sigma^2_{ln(OR)} = \frac{1}{p_0} + \frac{1}{1-p_0} + \frac{1}{p_1} + \frac{1}{1-p_1}$
  - for prevalence p_0 unexposed and p_1 exposed

  and,
- $\ln(OR) = ln(\frac{(p_1)(1-p_0)}{(1-p_1)(p_0)})$
- E.g. unexposed prevalence 1%, for OR 3
  - $n_{group} = 8 * (\frac{1}{.01} + \frac{1}{.99} + \frac{1}{.03} + \frac{1}{.97}) / ln(\frac{(.03)(.99)}{.97)(.01)})^2 \approx 865$
- (see website for R code...)

The Rule of 50

to demonstrate halving of risk

- For rare, discrete outcomes, at least 50 events in control group, and *an equal number* in a treatment group, to detect a halving of risk.
- E.g. risk of MI 8% over 5 years
- want to demonstrate drug reduces risk to 4%
- need enough patients to have *50 MI's* in each arm of the study
  - 50 / 0.08 = *625 patients in each group*, or 1,250 subjects in total
  - conservative ballpark, actual calculations return value of 553 each group.
- (see web for derivation)

Lehr's equation for binomial data or proportions

- average the two proportions to adapt Lehr's equation
  - $\bar{p} = (p_0 + p_1)/2$
  - $n_{group} = \frac{16\bar{p}\bar{q}}{(p_0 - p_1)^2}$
- E.g. outcome 30% in untreated population
- want to demonstrate treatment decreases to 10%.
  - $\bar{p} = (.3 + .1)/2 = .2$ and $\bar{q} = 1 - .2 = .8$, and $n_{group} = \frac{16 * .2 * .8}{(.3 - .1)^2} = 64$

conservative estimate for proportions

- maximize variance
  - $p_0 = .5$
- $\bar{p} = (.5 + .5)/2 = .5$
  - numerator of sample size calculation $16 * .5 * .5 = 4$:
  - $n_{group} = \frac{4}{(p_0 - p_1)^2}$
- E.g. $n_{group} = \frac{4}{(.3-.1)^2} = 100$
- conservative approximation only valid where sample size each group between 10 and 100

some more rules of thumb from Gerald Van Belle

- A quick estimate of the standard deviation is $s = \frac{Range}{\sqrt{n}}$
- Confidence intervals can overlap as much as 29% and still be statistically significantly different
- In a logistic regression, aim for about *10 outcomes for every parameter* in the model
  - some say 15...
- Little to be gained, beyond 4 or 5 controls per case
- Surprisingly little optimization in overall costs with unequal sample sizes even the costs for one population are appreciably greater than the other,

power calculations based on many assumptions

- power calculation are a guide, not a requirement
- should be run over range of possible effect sizes and standard deviations
- logical step to run over a *distribution*
- incorporating uncertainty in parameters inherently Bayesian

Bayesian twist to Lehr

- E.g. 10 point difference in IQ two groups, sd=20
- estimate from Lehr's formula 64 each group
- say believe estimate for sd has its own distribution
  - $\mu = 20$ and $\sigma^2 = 1$
- also believe estimate for difference has its own distribution
  - $\mu = 10$ and $\sigma^2 = 1$
- redefine basic Lehr's formula R function in terms of difference and standard deviation.
- sample from a truncated distribution (negative sd makes no sense)

R function for difference in means

```
nMeans.vary<-function(diff, sd){
    d<-diff/sd
    n<-16/d^2
    n
    }

install.packages("truncnorm")
library(truncnorm)

sd.r<-rtruncnorm(n=1000, a=0, mean=20, sd=1)
diff.r<-rtruncnorm(n=1000, a=0, mean=10, sd=1)

n1<-nMeans.vary(diff.r,sd.r)
summary(n1)
plot(density(n1))
```

allowing for uncertainty

- can achieve same power under the same conditions with a few as 30 or as many as 160 observations in each group.
- alternatively, look at how power varies for the point estimate of 64 observations in each group
  - solve for power given a sample size n, a difference between two groups of θ and a standard deviation of σ:
  - Power $= \phi(\sqrt{\frac{n\theta^2}{2\sigma^2}} - 1.96)$

```
p1<-sqrt(64/2) * diff.r/sd.r -1.96

summary(pnorm(p1))
plot(density(pnorm(p1)))
```

- power of study with 64 observations in each group can be as low as 42% or has high as 98%.

drug testing (binomial outcome)

traditional approach

- disease 25% prevalence
- want to demonstrate decrease to 10%
  - efficacy 40% (.1/.25)
- use Lehr's equation
- $n_{group} = \frac{16\bar{p}\bar{q}}{(p_0 - p_1)^2}$

```
nBin<-function(p0,p1){
    p<-(p0+p1)/2
    n<-16*p*(1-p)/(p0-p1)^2
    n
    }
nBin(.25,.1)
```

limited resources

say only enough money to enroll 20 patients?

- reframe the question: should you do the trial at all?
- perhaps only if likely demonstrate 15 of 20 pts will respond?
- add uncertainty to efficacy: ranges between 20% to 60% with sd of 10%
- use to define a Beta distribution to use as a prior in a binomial simulation
  - define parameters for Beta prior from mean and variance
  - use the Beta distribution to define a binomial simulation
  - power is proportion of simulations that meet or exceed the criteria greater than or equal to 15

```
estBetaParams <- function(mu, var) {
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta <- alpha * (1 / mu - 1)
  return(params = list(alpha = alpha, beta = beta))
}

estBetaParams(.4,(.1)^2)

N=1000
theta<-rbeta(N,9.2,13.8)
x<-rbinom(N,20, theta)
y<-0
accept<-ifelse(x>14.5, y+1, y+0)
prob<-sum(accept)/N
prob
library(coda)
densityplot(accept)
```

- about 2% prob study of 20 pts will show efficacy in 15 people.

learn from previous studies

- find study that demonstrates efficacy in 15 of 20 patients
- incorporate information by updating the Beta prior
- use "conjugacy"
  - *Beta* priors and *Binomial* likelihoods
  - if prior is
    $$\theta \sim Beta(a, b)$$
  - and the likelihood is
    $$X|\theta \sim Bin(N, \theta)$$
  - then the posterior is
    $$\theta|X \sim Beta(X + a, N - X + b)$$
- new prior for θ in planned study is $Beta(15 + 9.2, 20 - 15 + 13.8)$
  - combines Binomial $\binom{20}{15} p^{15} q^{20-15}$ likelihood of external data with the $Beta(9.2, 13.8)$

```
N=1000
theta<-rbeta(N,24.2,18.8)
x<-rbinom(N,20, theta)
y<-0
accept<-ifelse(x>14.5, y+1, y+0)
prob<-sum(accept)/N
prob
```

- probability of demonstrating target efficacy has gone up
  - but perhaps not as much as we want

from updating to adapting

- do the study and find 14 of 20 patients respond
- make a "decision rule": if see 25 responses in next 40 patients will continue drug development
- how likely is that to be?
- again a matter of updating the prior and running the simulation
- updated prior for θ is $Beta(14 + 24.2, 20 - 14 + 18.8)$
- probability of 25 or more successes in an additional 40 trials is about 38%
    - judgement on whether that is a good "rule"

```
N=1000
theta<-rbeta(N,38.2,24.8)
x<-rbinom(N,40, theta)
y<-0
accept<-ifelse(x>25, y+1, y+0)
(prob<-sum(accept)/N)
```

about adapative designs

- learn from and adapt to accumulating evidence
- decisions based on predictive probabilities
  - early stopping, re-estimation of sample sizes, weighting the randomization process
- best suited to settings where patients enrolled quickly, responses ascertained quickly, costs high, uncertainty about effect
- requires additional, sometimes complex initial planning, considerable flexibility, regulatory approvals, concerns *may increase type I error* because multiple analyses, concerns *may introduce bias* because know results or through patient selection

Bayesian sample size calculations

- rather than run calculations for a set sample size (as above) replace with a vector of values

```
Nrep<-1000
s<-40:100
theta<-rbeta(Nrep,38.2,24.8)
y<-0
results<-matrix(0, nrow=length(s), ncol=1)
for(i in 1:length(s)){
x<-rbinom(Nrep,s[i], theta)
accept<-ifelse(x>25, y+1, y+0)
prob[i]<-sum(accept)/Nrep
results[i,]<-prob[i]
}
(tab<-cbind(s, results))
```

- hit the usual 80% threshold with a sample of about 49.

weighting evidence

- till now have applied equal weighting to prior study or evidence
- can *downweight* previous study data with weight $0 \leq w \leq 1$l
  - likelihood, $Beta(a * w, b * w)$
  - each previous patient info worth fraction w of new patient info
- E.g. not very confident of the evidence from the literature
- downweight it by 50%
- take $Beta(24.2, 18.8)$ prior and multiply through by .5
  - $Beta(24.2 * .5, 18.8 * .5)$
- then update with the evidence from the trial
  - $Beta(14 + 12.2, 20 - 14 + 14.4)$

down-weight the prior

```
Nrep<-1000
s<-40:100
theta<-rbeta(Nrep,26.2,20.4)
y<-0
results<-matrix(0, nrow=length(s), ncol=1)
for(i in 1:length(s)){
x<-rbinom(Nrep,s[i], theta)
accept<-ifelse(x>25, y+1, y+0)
prob[i]<-sum(accept)/Nrep
results[i,]<-prob[i]
}

(tab<-cbind(s, results))
```

- now don't hit 80% power until a sample size of about 54.

shift the prior

- can *shift* or manipulate the prior
  - keep overall sample size same
  - increase or decrease the number of successes in beta-binomial
  - e.g. shift 7 successes of $Beta(14 + 24.2, 20 - 14 + 18.8)$ prior to failures as $Beta(7 + 24.2, 20 - 7 + 18.8)$

```
Nrep<-1000
s<-40:100
theta<-rbeta(Nrep,31.2,31.8)
y<-0
results<-matrix(0, nrow=length(s), ncol=1)
for(i in 1:length(s)){
x<-rbinom(Nrep,s[i], theta)
accept<-ifelse(x>25, y+1, y+0)
prob[i]<-sum(accept)/Nrep
results[i,]<-prob[i] }
(tab<-cbind(s, results))
```