Test for Binomial Proportions

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Outline

- Tests for a binomial proportion
- Score test versus Wald
- Exact binomial test
- Tests for differences in binomial proportions
- Intervals for differences in binomial proportions

Motivation

- Consider a randomized trial where 40 subjects were randomized (20 each) to two drugs with the same active ingredients but different excipients
- Consider counting the number of participants with side effects for each drug

	Side Effects	None	Total
Drug A	11	9	20
Drug B	5	15	20
Total	16	24	40

Hypothese tests for binomial proportions

- Consider testing $H_0 : p = p_0$ for a binomial proportion
- The **score** test statistic

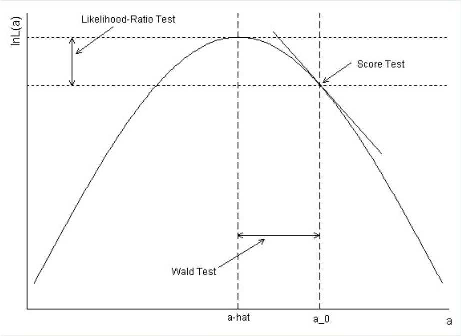
$$\frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

  follows a Z distribution for large n
- This test performs better than the Wald test

$$\frac{\widehat{p} - p_0}{\sqrt{\widehat{p}(1 - \widehat{p})/n}}$$

- The Wald interval for p

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\hat{p}(1 - \hat{p})/n}$$



- Likelihood ration test statistics: $-2[\ell(\tilde{\beta}) - \ell(\widehat{\beta})]$
- Wald test statistic: $(\mathbf{L}'\widehat{\beta} - \xi)'[\mathbf{L}'I(\widehat{\beta})^{-1}\mathbf{L}]^{-1}(\mathbf{L}'\widehat{\beta} - \xi)$
- Score test statistic: $s'(\tilde{\beta})I(\tilde{\beta})^{-1}s(\tilde{\beta})$
- $s(\beta)$ be the vector of score with j^{th} component, $s_j(\beta) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}$

Some discussion

- The Wald interval performs terribly
- Coverage probability varies wildly, sometimes being quite low for certain values of n even when p is not near the boundaries
  - Example, when $p = .5$ and $n = 40$ the actual coverage of a 95% interval is only 92%
- When p is small or large, coverage can be quite poor even for extremely large values of n
  - Example, when $p = .005$ and $n = 1,876$ the actual coverage rate of a 95% interval is only 90%

Simple fix

- A simple fix for the problem is to add two successes and two failures
- That is let $\tilde{p} = (X + 2)/(n + 4)$
- The (Agresti- Coull) interval is

$$\tilde{p} \pm Z_{1-\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

- Motivation: when p is large or small, the distribution of $\hat{p}$ is skewed and it does not make sense to center the interval at the MLE; adding the pseudo observations pulls the center of the interval toward 0.5

Example

- In our previous example consider testing whether or not Drug A's percentage of subjects with side effects is greater than 10%
- $H_0 : p_A = 0.1$ versus $H_A : p_A > 0.1$
- $\widehat{p} = \frac{11}{20} = 0.55$
- Score test statistic

$$\frac{0.55 - 0.1}{\sqrt{0.1 \times 0.9/20}} = 6.7$$

- Reject, p-value$=P(Z > 6.7) \approx 0$

Exact binomial tests

- Considering calculating an exact *p*-value
- What's the probability, under the null hypothesis, of getting evidence as extreme or more extreme than what we observed?

$$P(X_A > 11) = \sum_{x=11}^{20} \binom{20}{11} 0.1^x \times 0.9^{20-x} \approx 0$$

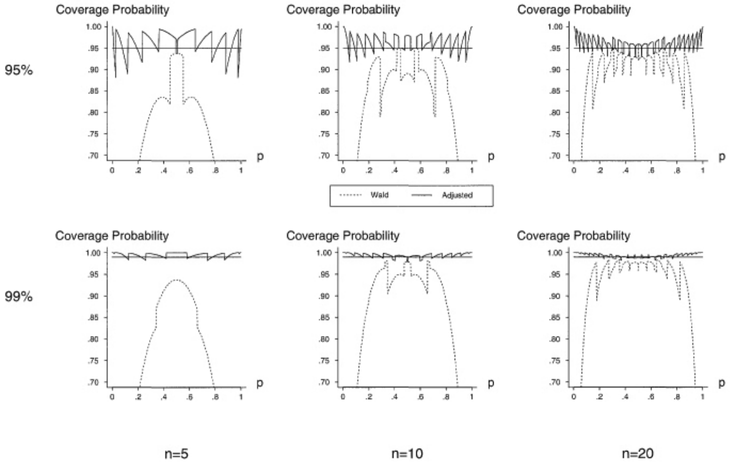
Exact binomial tests

- Considering calculating an exact *p*-value
- What's the probability, under the null hypothesis, of getting evidence as extreme or more extreme than what we observed?

$$P(X_A > 11) = \sum_{x=11}^{20} \binom{20}{11} 0.1^x \times 0.9^{20-x} \approx 0$$

- pbinom(10, 20, 0.1, lower.tail=FALSE)
- binom.test(11, 20, 0.1, alternative="greater")

Wald versus Agresti/Coull



Bayesian analysis

- Bayesian statistics posits a **prior** on the parameter of interest
- All inferences are then performed on the distribution of the parameter given the data, called the **posterior**
- In general,

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)},$$

hence

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- Therefore (as we saw in diagnostic testing) the likelihood is the factor by which our prior beliefs are updated to produce conclusions in the light of the data

Beta priors

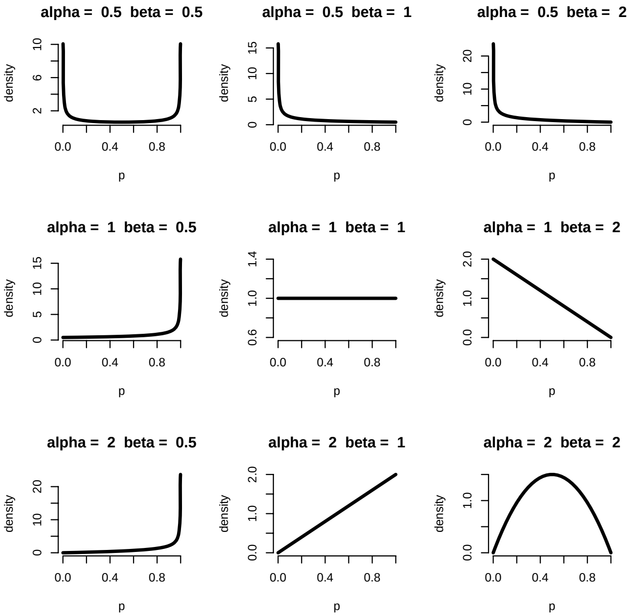
- The beta distribution is the default prior for parameters between 0 and 1.
- The beta density depends on two parameters α and β

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{for } \ 0 \leq p \leq 1$$

- The mean of the beta density is $\alpha/(\alpha+\beta)$
- The variance of the beta density is

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

- The uniform density is the special case where $\alpha = \beta = 1$


alpha = 0.5 beta = 0.5
alpha = 0.5 beta = 1
alpha = 0.5 beta = 2
alpha = 1 beta = 0.5
alpha = 1 beta = 1
alpha = 1 beta = 2
alpha = 2 beta = 0.5
alpha = 2 beta = 1
alpha = 2 beta = 2
density
p
0.0
0.4
0.8

Posterior

- Suppose that we chose values of α and β so that the beta prior is indicative of our degree of belief regarding p in the absence of data
- Then using the rule that

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

and throwing out anything that doesn't depend on p, we have that

$$\begin{aligned} \text{Posterior} &\propto p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \end{aligned}$$

- This density is just another beta density with parameters $\tilde{\alpha} = x + \alpha$ and $\tilde{\beta} = n - x + \beta$

Posterior mean

- Posterior mean

$$\begin{aligned}
E[p \mid X] &= \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} \\
&= \frac{x+\alpha}{x+\alpha+n-x+\beta} \\
&= \frac{x+\alpha}{n+\alpha+\beta} \\
&= \frac{x}{n} \times \frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+\beta}{n+\alpha+\beta} \\
&= \text{MLE} \times \pi + \text{Prior Mean} \times (1-\pi)
\end{aligned}$$

- The posterior mean is a mixture of the MLE ($\hat{p}$) and the prior mean
- π goes to 1 as n gets large; for large n the data swamps the prior
- For small n, the prior mean dominates
- Generalizes how science should ideally work; as data becomes increasingly available, prior beliefs should matter less and less
- With a prior that is degenerate at a value, no amount of data can overcome the prior

Posterior variance

- The posterior variance is

$$Var(p \mid x) = \frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta})^2(\tilde{\alpha}+\tilde{\beta}+1)} = \frac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)}$$

- Let $\tilde{p} = (x+\alpha)/(n+\alpha+\beta)$ and $\tilde{n} = n+\alpha+\beta$ then we have

$$Var(p \mid x) = \frac{\tilde{p}(1-\tilde{p})}{\tilde{n}+1}$$

Discussion

- If $\alpha = \beta = 2$ then the posterior mean is

$$\tilde{p} = (x + 2)/(n + 4)$$

and the posterior variance is

$$\tilde{p}(1 - \tilde{p})/(\tilde{n} + 1)$$

- This is almost exactly the mean and variance we used for the Agresti-Coull interval

Example

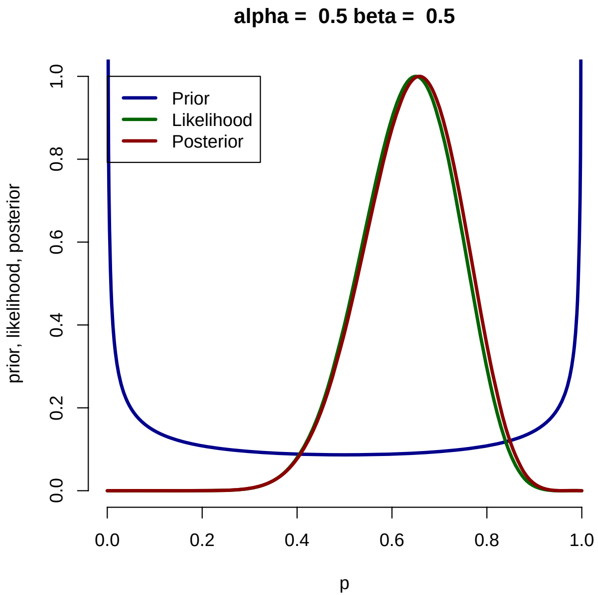
- Consider an example where $x = 13$ and $n = 20$
- Consider a uniform prior, $\alpha = \beta = 1$
- The posterior is proportional to (see formula above)

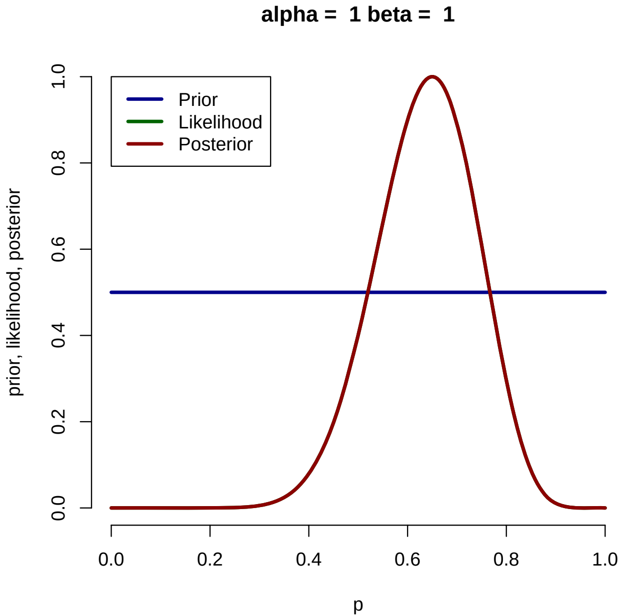
$$p^{x+\alpha-1}(1-p)^{n-x+\beta-1} = p^{x}(1-p)^{n-x}$$

  that is, for the uniform prior, the posterior is the likelihood
- Consider the instance where $\alpha = \beta = 2$ (recall this prior is humped around the point .5) the posterior is

$$p^{x+\alpha-1}(1-p)^{n-x+\beta-1} = p^{x+1}(1-p)^{n-x+1}$$

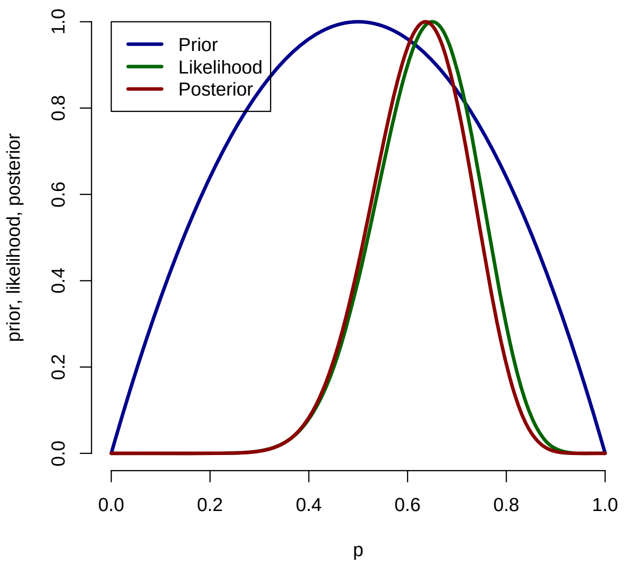
- The "Jeffrey's prior" which has some theoretical benefits puts $\alpha = \beta = .5$

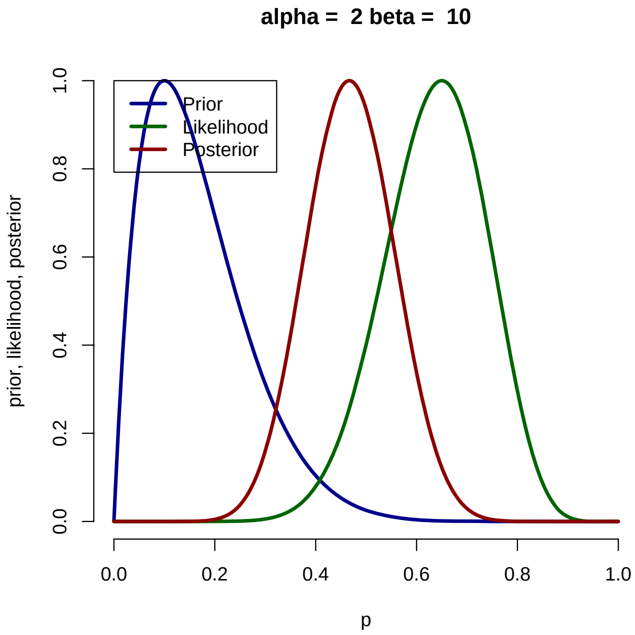

alpha = 0.5 beta = 0.5
Prior
Likelihood
Posterior
prior, likelihood, posterior
0.0
0.2
0.4
0.6
0.8
1.0
p

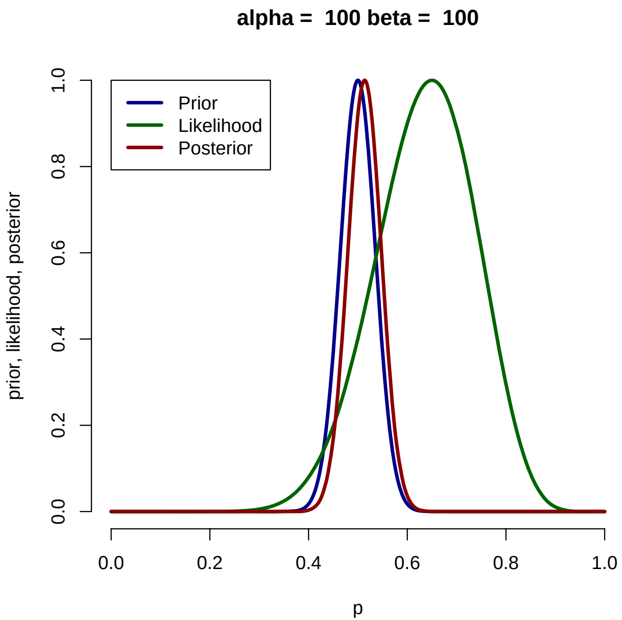


alpha = 1 beta = 1
Prior
Likelihood
Posterior
prior, likelihood, posterior
0.0
0.2
0.4
0.6
0.8
1.0
p

alpha = 2 beta = 2




alpha = 2 beta = 10
Prior
Likelihood
Posterior
prior, likelihood, posterior
0.0
0.2
0.4
0.6
0.8
1.0
p

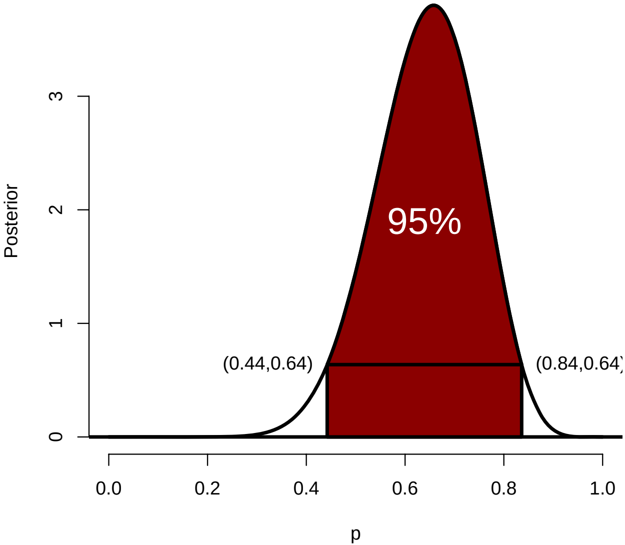

alpha = 100 beta = 100
Prior
Likelihood
Posterior
prior, likelihood, posterior
0.0
0.2
0.4
0.6
0.8
1.0
p

Bayesian credible intervals

- A *Bayesian credible interval* is the Bayesian analog of a confidence interval
- A 95% credible interval, $[a, b]$ would satisfy

$$P(p \in [a, b] \mid x) = .95$$

- The best credible intervals chop off the posterior with a horizontal line in the same way we did for likelihoods
- These are called highest posterior density (HPD) intervals


Posterior
3
2
1
0
95%
(0.44,0.64)
(0.84,0.64)
0.0
0.2
0.4
0.6
0.8
1.0
p

R code

Install the `binom` package, then the command

```
library(binom)
binom.bayes(13, 20, type = "highest")
```

gives the HPD interval. The default credible level is 95% and the default prior is the Jeffrey's prior.

Interpretation of confidence intervals

- Confidence interval: (Wald) [.44, .86]
- Actual interpretation:

  The interval .44 to .86 was constructed such that in repeated independent experiments, 95% of the intervals obtained would contain p.

Credible intervals

- Recall that Jeffrey's prior 95% credible interval was [.44, .84]
- Actual interpretation

The probability that p is between .44 and .84 is 95%.

Comparing two binomials

- Consider now testing whether the proportion of side effects is the same in the two grouops
- Let $X \sim Binomial(n_1, p_1)$ and $\widehat{p}_1 = \frac{X}{n_1}$
- Let $Y \sim Binomial(n_2, p_2)$ and $\widehat{p}_2 = \frac{Y}{n_2}$

$n_{11} = X$	$n_{12} = n_1 - X$	$n_1 = n_{1+}$
$n_{21} = Y$	$n_{22} = n_2 - Y$	$n_2 = n_{2+}$
n_{+1}	n_{+2}	

Comparing two proportions

- Consider testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$
- The score test statistic for this null hypothesis is

$$TS = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}(1 - \widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

  where $\widehat{p} = \frac{X+Y}{n_1+n_2}$ is the estimate of the common proportion under the null hypothesis
- This statisitc is normally distributed for large n_1 and n_2.

- The Wald test is

$$TS = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

- The resulting confidence interval is

$$\widehat{p}_1 - \widehat{p}_1 \pm Z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Posterior Inference

- As in the one sample case, the Wald interval and test performs poorly relative to the score interval
- For testing, always use the score test
- Another approach is to use the Agresti/Caffo interval which is calculated as $\tilde{p} = \frac{x+1}{n_1+2}, \tilde{n}_1 = n_1 + 2, \tilde{p} = \frac{y+1}{n_2+2}, \tilde{n}_1 = n_2 + 2.$
- Then use the Agresti/Caffo estimates to calculate the Wald interval.

Example

- Test whether or not the proportion of side effects is the same fo the two drugs
- $\widehat{p_A} = 0.55, \widehat{p_B} = 0.25, \widehat{p} = \frac{16}{40} = 0.4$
- Test statistic

$$\frac{0.55 - 0.25}{\sqrt{0.4 \times 0.6 \times (\frac{1}{20} + \frac{1}{20})}} = 1.61$$

- Fail to reject H_0 at $\alpha = 0.05$ level (compare with 1.96)
- p-value $P(|Z| > 1.61) = 0.11$

Wald versus Agresti/Caffo

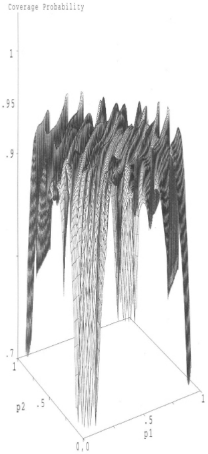


Figure 7. Coverage probabilities for 95% nominal Wald confidence interval as a function of p1 and p2, when n1 = n2 = 10.

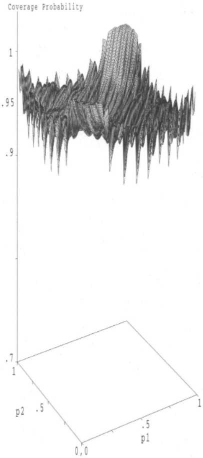


Figure 8. Coverage probabilities for 95% nominal adjusted confidence interval (adding t = 4 pseudo observations) as a function of p1 and p2, when n1 = n2 = 10.

Wald versus Agresti/Caffo

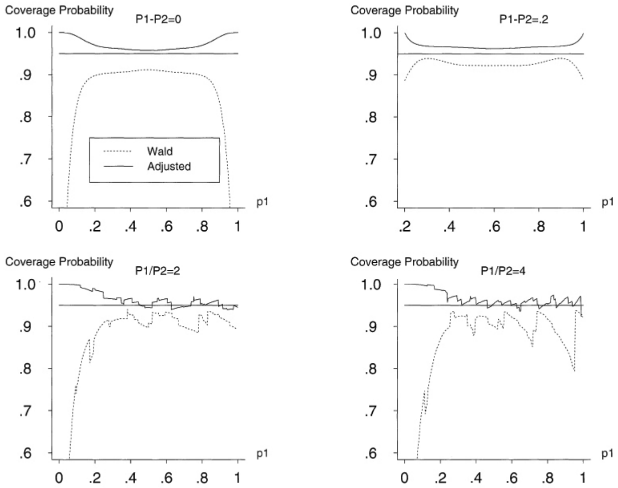


Figure 6. Coverage probabilities for nominal 95% Wald and adjusted confidence intervals (adding t = 4 pseudo observations) as a function of p1 when p1 − p2 = 0 or .2 and when p1/p2 = 2 or 4, for n1 = n2 = 10.

Bayesian inference for two binomial proportions

- Consider putting independent Beta(α_1, β_1) and Beta(α_2, β_2) priors on p_1 and p_2 respectively
- Then the posterior is

$$\pi(p_1, p_2) \propto p_1^{x+\alpha_1-1}(1-p_1)^{n_1-x+\beta_1-1} \times p_2^{y+\alpha_2-1}(1-p_2)^{n_2-y+\beta_2-1}$$

- Hence under this (potentially naive) prior, the posterior for p_1 and p_2 are independent betas
- The easiest way to explore this posterior is via Monte Carlo simulation

R Code

```
x <- 11; n1 <- 20; alpha1 <- 1; beta1 <- 1
y <- 5; n2 <- 20; alpha2 <- 1; beta2 <- 1
p1 <- rbeta(1000, x + alpha1, n1 - x + beta1)
p2 <- rbeta(1000, y + alpha2, n2 - y + beta2)
rd <- p2 - p1
plot(density(rd))
quantile(rd, c(.025, .975))
mean(rd)
median(rd)
```

Binomial posterior differences in proportions

- The function BinomPost on the course website automates a lot of this
- The output is

```
Post mn rd (mcse) = -0.274 (0.004)
Post mn rr (mcse) = 0.516 (0.007)
Post mn or (mcse) = 0.362 (0.008)

Post med rd        =  -0.281
Post med rr        =  0.483
Post med or        =  0.292

Post mod rd        =  -0.29
Post mod rr        =  0.437
Post mor or        =  0.203

Equi-tail rd       =  -0.529 0.011
Equi-tail rr       =  0.195 1.028
Equi-tail or       =  0.084 1.055
```

Bayesian Posterior Credible Interval