Descriptive analysis of categorical variables

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn

- Categorical data
- Probability
- Statistical description of
  - Prevalence
  - Incidence
  - Rate

Measurement and comparison

To find out whether a community is healthy or unhealthy:

- first measure one or more indicators of health (deaths, new cases of disease, etc)
- compare the results with another community or group.

Measures of Disease Occurrence

- Incidence proportion (risk)
- Incidence rate (density)
- Prevalence

All three are *loosely* called "**rates**" (but only the second is a *true* rate)

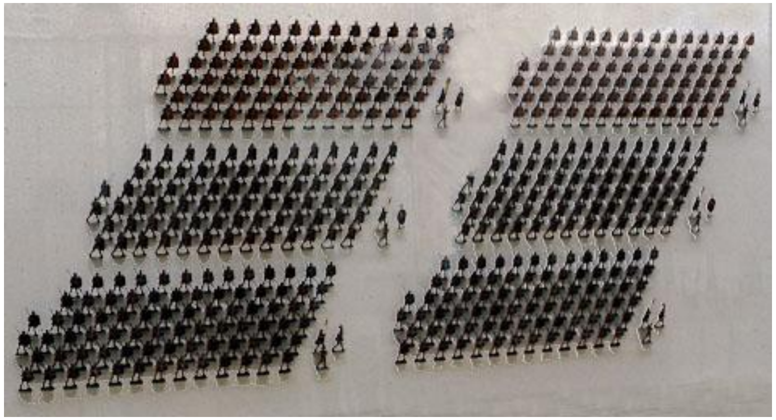
Types of populations

We measure disease occurrence in two types of populations:

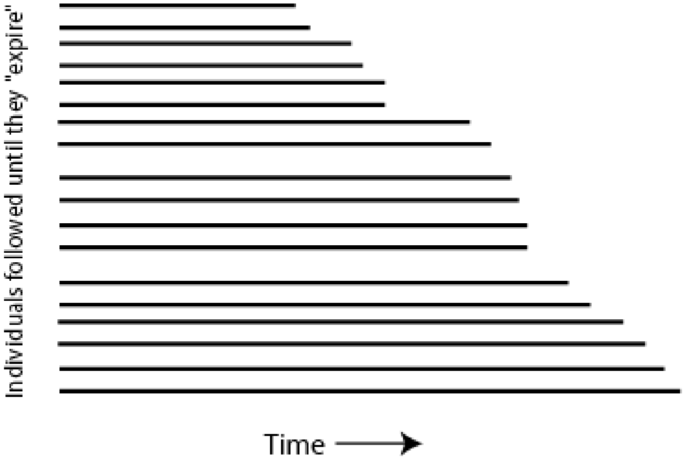
- **Closed populations $\Rightarrow$ "cohorts"**
- **Open populations**

Closed population = cohort

Cohort word origin (Latin *cohors*) basic tactical unit of a Roman legion

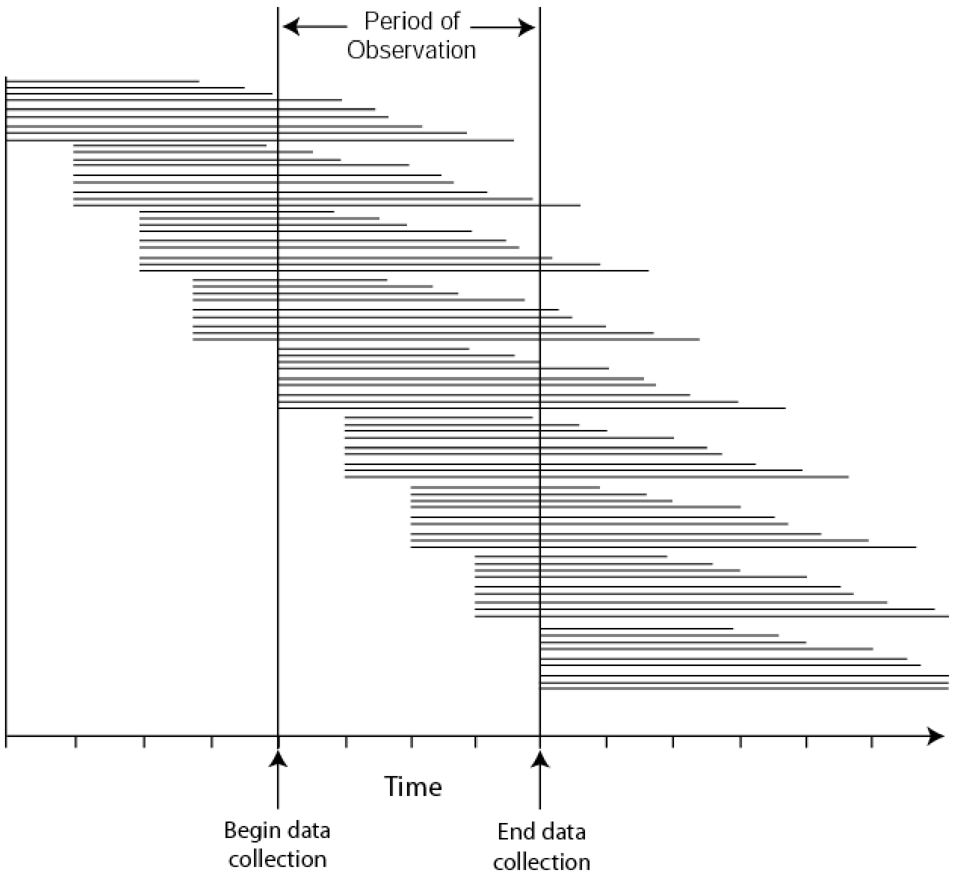


Epi cohort ≡ a group of **individuals** **followed over time**



Open population

- Inflow (immigration, births)
- Outflow (emigration, death)
- An open population in “steady state” (constant size) is said to be **stationary**

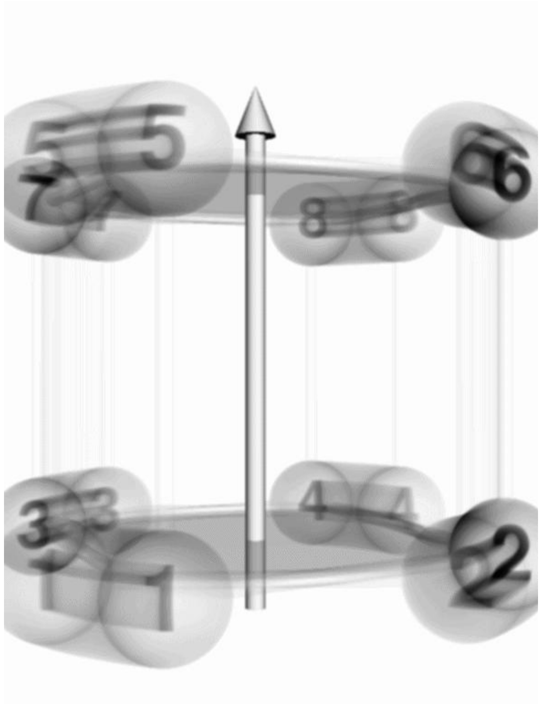


Numerators and denominators

- "Rates" are composed of numerators and denominators
- **Numerator** ⇒ case count

  Incidence count ⇒ onsets

  Prevalence count ⇒ old + new cases
- ***Denominators*** ⇒ reflection of population size



Denominators

Denominators: reflection of population size

Incidence proportion

Can be calculated *only* in cohorts

$$\mathrm{IP} = \frac{\text{no. of onsets over time}}{\text{no. @ risk at beginning of study}}$$

- **Synonyms: risk, cumulative incidence, attack rate**
- **Interpretation: average risk**

Example of IP

- Objective: estimate risk of uterine cancer
- Recruit cohort of 1000 women
- 100 had hysterectomies, leaving 900 *at risk*
- Follow at risk individuals for 10 years
- Observe 10 onsets of uterine cancer

$$\text{IP} = \frac{\text{no. of onsets}}{\text{no. @ risk}} = \frac{10 \cancel{\text{ women}}}{900 \cancel{\text{ women}}} = 0.0111$$

10-year average risk is .011 or 1.1%.

Incidence rate

$$\mathrm{IR} = \frac{\text{no. onsets}}{\text{Sum of person-time @ risk}}$$

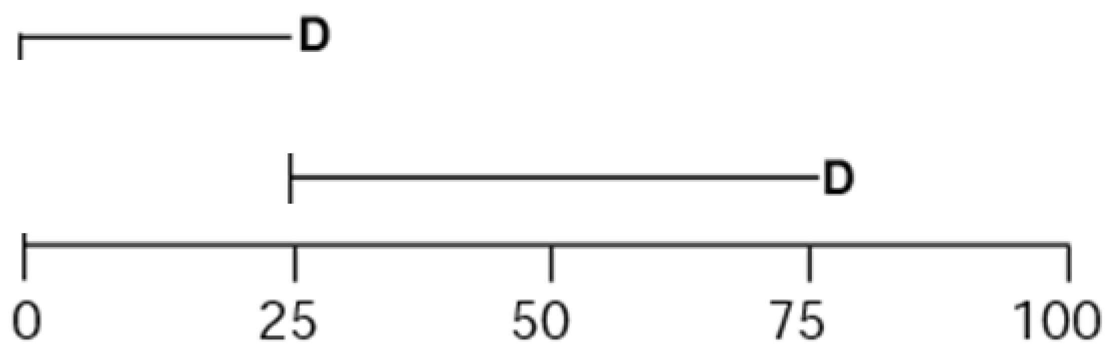
- Synonyms: incidence density, person-time rate
- Interpretation A: “Speed” at which events occur
- Interpretation B: When disease is rare:
  rate per person-year ≈ one-year risk
- Calculated differently in closed and open populations

- Objective: estimate rate of uterine cancer
- Recruit cohort of 1000 women
- 100 had hysterectomies, leaving 900 *at risk*
- Follow at risk individuals for 10 years
- Observe 10 onsets of uterine cancer

$$IR = \frac{\text{no. of onsets}}{\text{person-time}} = \frac{10 \cancel{\text{ women}}}{900 \cancel{\text{ women}} \times 10 \text{ years}} = \frac{10}{9000 \text{ years}} = \frac{.00111}{\text{year}}$$

Rate is .00111 per year or 11.1 per 10,000 years

Individual follow-up over time



$$\mathrm{IR} = \frac{\text{onsets}}{\sum \text{person-time}} = \frac{2\ \text{onsets}}{25\ \text{years} + 50\ \text{years}} = \frac{2\ \text{onsets}}{75\ \text{years}}$$

$$= 0.0267\ \text{per person-years} = 2.67\ \text{per}\ 100\ \text{person-years}$$

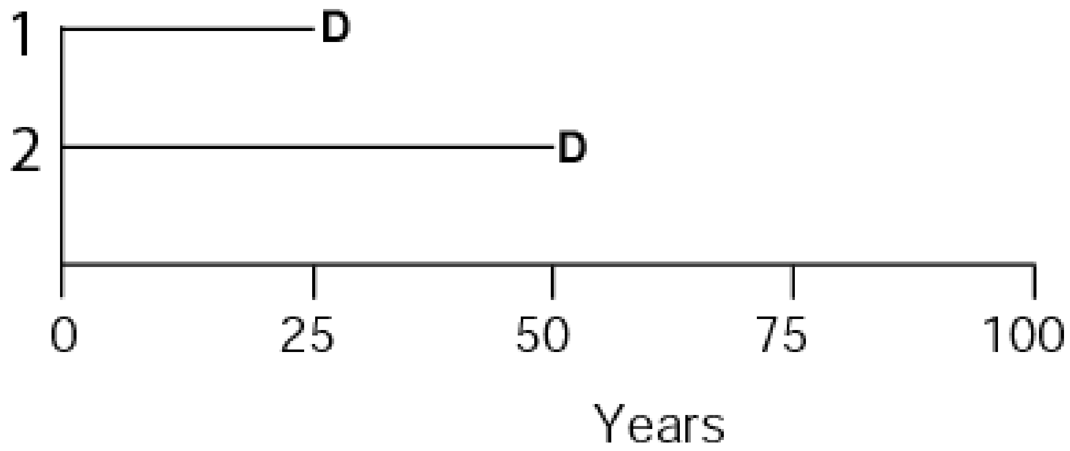
Mortality and life expectancy

In stationary populations, and in cohorts with complete follow-up, the mortality rate is the reciprocal of life expectancy (and vice versa).

$$\text{Life expectacy} = \frac{1}{\text{Mortality Rate}}$$

Example: for a mortality rate of .0267 per year

$$\text{Life expectacy} = \frac{1}{.0267/\text{year}} = 37.5 \text{ years}$$



$$\text{This cohort has a mortality rate of } \frac{2 \text{ deaths}}{(25+50) \text{ years}} = 0.0267 \text{ year}^{-1}$$

$$\text{This cohort has life expectancy } \frac{(25+50) \text{ years}}{2} = 37.5 \text{ years}$$

Incidence rate in open population

$$IR = \frac{\text{onsets}}{\text{Avg population size} \times \text{duration of observation}}$$

Example: 2,391,630 deaths in 1999 (one year)
Population size = 272,705,815

$$IR = \frac{2{,}391{,}630 \text{ deaths}}{272{,}705{,}815 \text{ persons} \times 1 \text{ year}} = 0.008770 \text{ deaths year}^{-1}$$

$$= 877 \text{ per } 100{,}000 \text{ person-years}$$

Prevalence

$$\text{Prevalence} = \frac{\text{no. old and new cases}}{\text{no. of people}}$$

- Point prevalence ≡ prevalence at a particular point in time
- Period prevalence ≡ prevalence over a period of time
- Interpretation A: proportion with condition
- Interpretation B: probability a person selected at random will have the condition

Example of prevalence

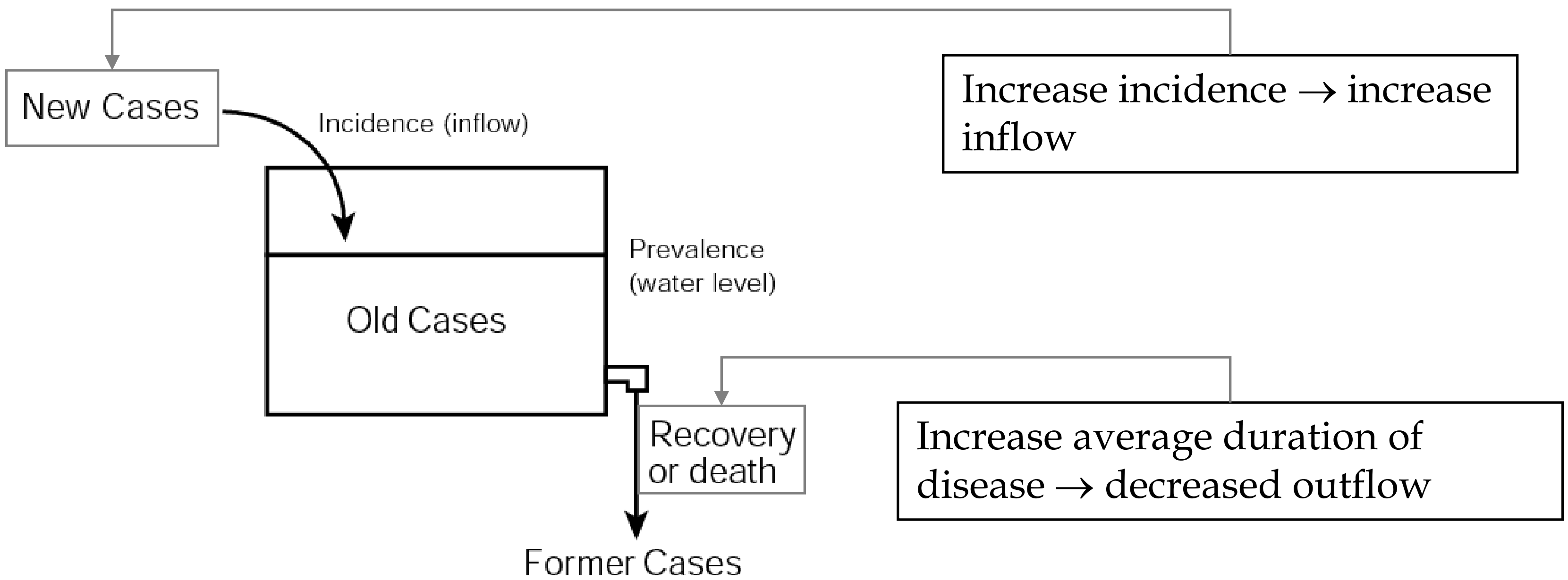
- Recruit 1000 women
- Ascertain: 100 with hysterectomies

$$\text{Prevalence} = \frac{\text{no. cases}}{\text{no. of people}} = \frac{100\ \cancel{\text{people}}}{1000\ \cancel{\text{people}}} = 0.10$$

Prevalence in sample is 10%

Dynamic prevalence

Ways to increase prevalence



Prevalence and incidence

When disease rare & population stationary

$$\text{prevalence} \approx (\text{incidence rate}) \times (\text{average duration})$$

Example:

- Incidence rate = 0.01 / year
- Average duration of the illness = 2 years.
- Prevalence ≈ 0.01 / year × 2 years = 0.02

Estimation of 95% confidence interval

Proportions

- **Proportion** of event in the sample, denoted "p hat":

$$\boxed{\hat{p} = \frac{x}{n}}$$

where *x* = no. of events and *n* = sample size

Proportion, cont

Two of 10 individuals in the sample have a risk factor for disease X

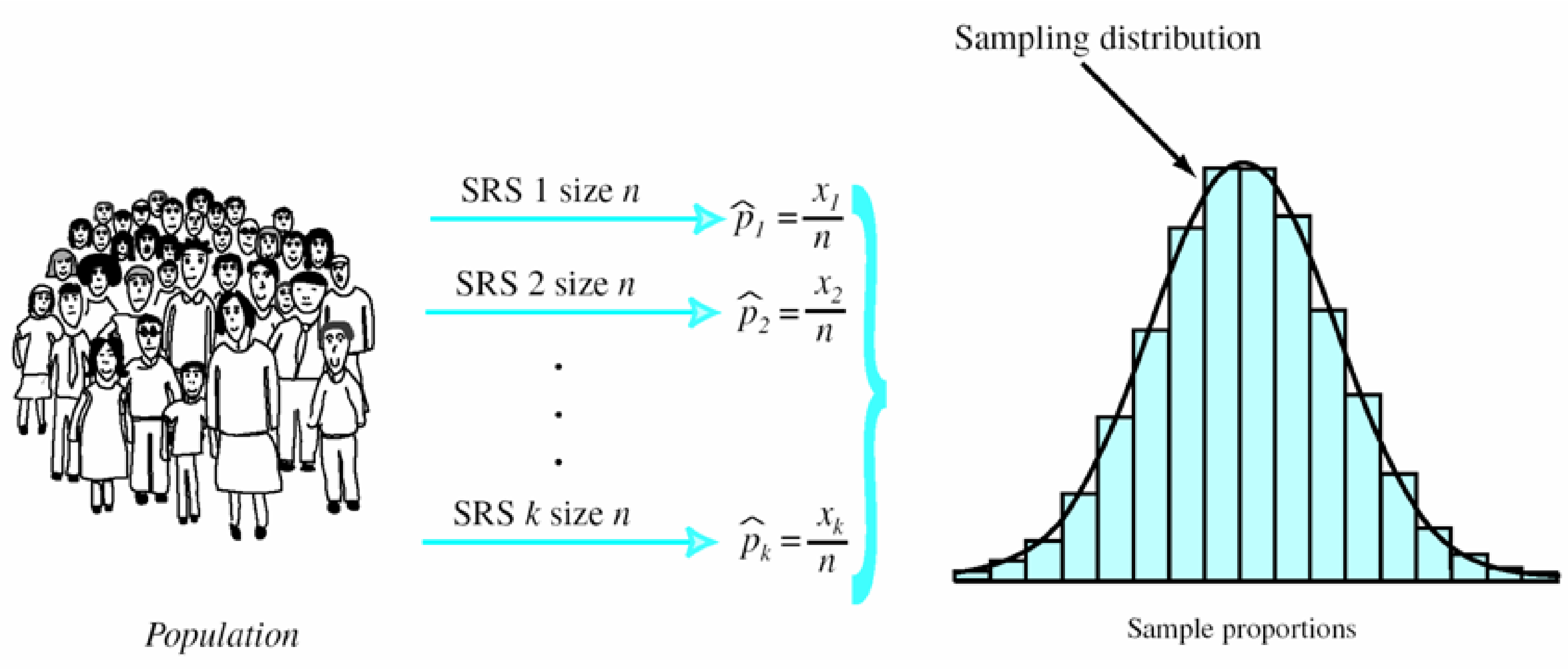
The prevalence of this risk factor in the sample is:

$$\boxed{\hat{p} = \frac{x}{n} = \frac{2}{10} = 0.1(\text{or}10\%)}$$

Inference about a Proportion

How good is sample proportion at estimating population proportion *p*?

Consider what would happen if we took repeated samples, each of size *n*, from the population? How would sample proportions be distributed?



Normal Approximation for Proportions

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

$$\text{where } q = 1 - p$$

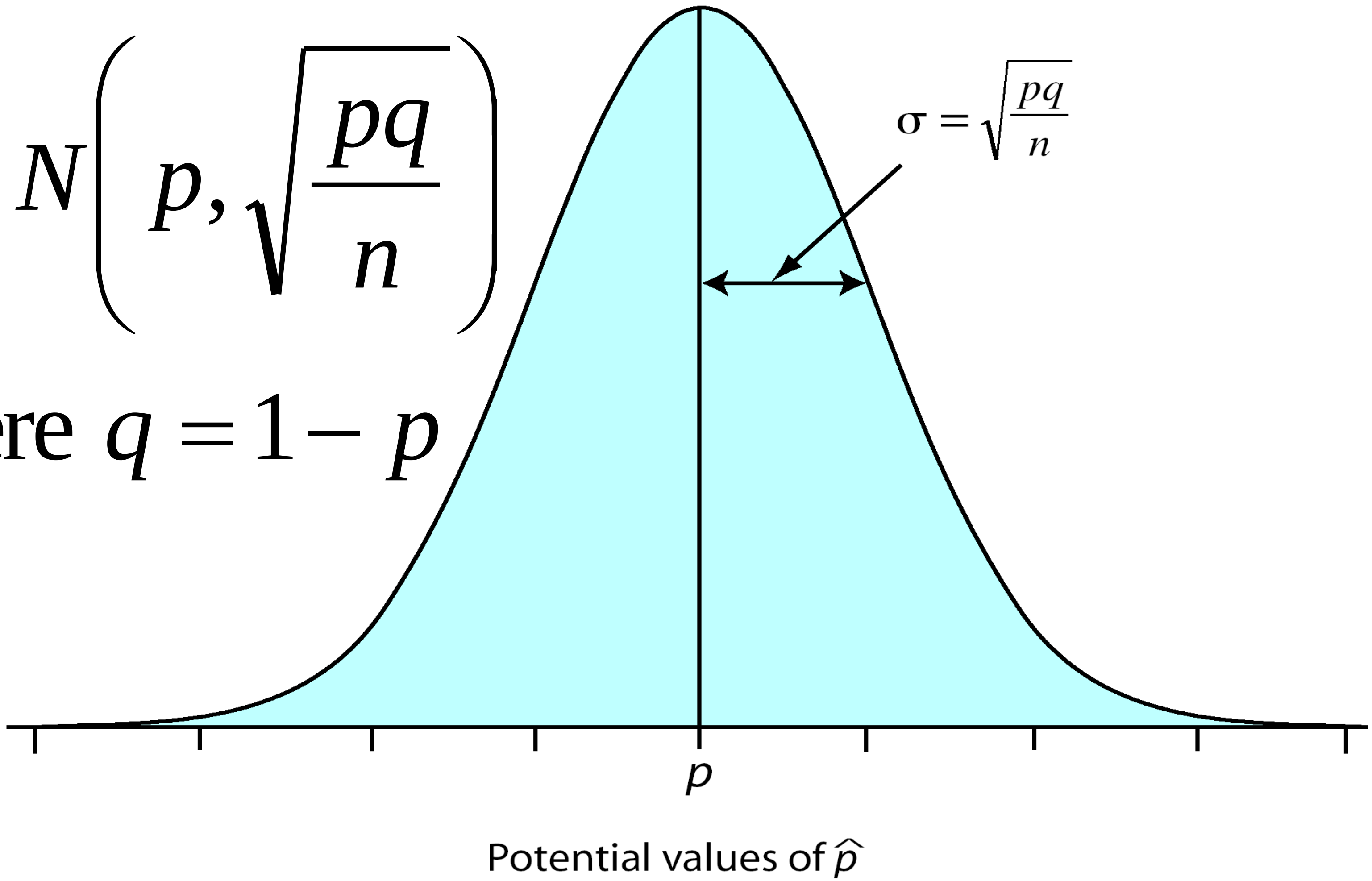


FIGURE 16.4 Sampling distribution of a proportion, Normal approximation.

Normal approximation

H_0: $p = p_0$ vs. H_a: $p \neq p_0$ where p_0 represents the proportion specified by the null hypothesis

Test statistic

$$\boxed{z_{\text{stat}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}}$$

Example

n = 57 finds 17 smokers (p-hat = 17 / 57 = 0.2982).

The national average for smoking prevalence is 0.25. Is the proportion in the sample significantly different than the national average?

H_0: p = 0.25 vs. H_a: $p \neq 0.25$

$$z_{\text{stat}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{.2982 - .25}{\sqrt{.25 \cdot .75/57}} = 0.84$$

The sample proportion is *not* significantly different than the national average.

Confidence Interval for Proportion

This method is called the "plus four method" because it adds four imaginary points during calculations. It is much more accurate than the traditional Normal method.

A 1−α(100%) confidence interval for *p* is:

$$\tilde{p} \pm z_{1-\frac{a}{2}} \times \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} \quad \text{where}$$

$$\tilde{x} = \tilde{x} + 2, \quad \tilde{n} = n + 4, \quad \tilde{p} = \frac{\tilde{x}}{\tilde{n}}, \quad \text{and } \tilde{q} = 1 - \tilde{p}$$

Confidence Interval, example

Based on n = 57 and x = 17, the 95% CI for the prevalence of smoking in the population is:

$$\tilde{x} = x + 2 = 17 + 2 = 19; \ \tilde{n} = n + 4 = 57 + 4 = 61$$

$$\tilde{p} = \frac{19}{61} = .3115; \ \tilde{q} = 1 - .3115 = .6885$$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} = \sqrt{\frac{(.3115)(.6885)}{61}} = .0593$$

$$z = 1.96 \text{ for } 95\% \text{ confidence}$$

$$95\% \text{ CI for } p = \tilde{p} \pm z \cdot SE_{\tilde{p}} = .3115 \pm (1.96)(.0593)$$

$$= .3115 \pm .1162 = (.1953, .4277)$$

Sample Size and Power

Three approaches:

- *n* needed to estimate *p* with margin of error *m* (for confidence interval)
- *n* needed to test H_0 at given α level and power
- The power of testing H_0 under stated conditions

n need to achieve margin of error *m*

$$\boxed{n = \frac{z_{1-\frac{\alpha}{2}}^{2}\, p^{*} q^{*}}{m^{2}}}$$

- where p^* represent an educated guess for population proportion p (when no educated guess for p^* is available, let $p^* = .5$)
- Round *up* to next integer to ensure stated precision

n need to achieve *m*, example

Suppose our educated guess for the proportion is p^* = 0.30

For margin of error of .05, use:

$$n = \frac{(1.96^2)(.30)(.70)}{.05^2} = 322.7 \Rightarrow 323$$

For margin of error of .03, use:

$$n = \frac{(1.96^2)(.30)(.70)}{.03^2} = 896.4 \Rightarrow 897$$

n to test H_0: $p = p_0$

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \sqrt{p_0 q_0} + z_{1-\beta} \sqrt{p_1 q_1}}{p_1 - p_0} \right)^2$$

where

- α ≡ alpha level of the test (two-sided)
- $1 - \beta$ ≡ power of the test
- p_0 ≡ proportion under the null hypothesis
- p_1 ≡ proportion under the alternative hypothesis

n to test H_0: $p = p_0$, example

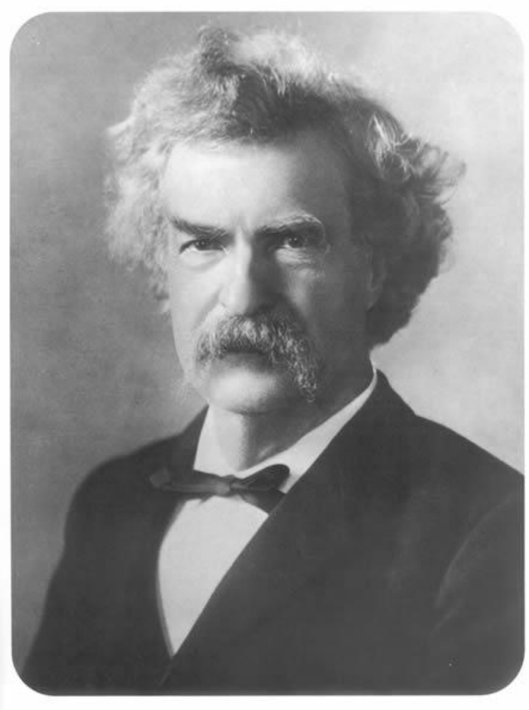
How large a sample is needed to test H_0: $p = 0.21$ against H_a: $p = 0.31$ at $\alpha = 0.05$ (two-sided) with 90% power?

$$\boxed{\begin{aligned} n &= \left(\frac{1.96\sqrt{(0.21)(0.79)} + 1.28\sqrt{(0.31)(0.69)}}{0.31 - 0.21} \right)^2 \\ &= 193.3 \Rightarrow 194 \end{aligned}}$$

$\Rightarrow$ means round up to ensure stated power

Conditions for Inference

- Sampling independence
- Valid information
- The plus-four confidence interval requires at least 10 observations
- The z test of H_0: $p = p_0$ requires $np_0q_0 \geq 5$



I'd rather have a sound judgment than a talent.
Mark Twain

Bayesian analysis of proportion

Review

- When X $\sim$ Binomial(n, π) we know that
- p = X/n is the MLE for π
- Var(p) = p(1 − p)/n

- Wald interval for π

$$\boxed{p \pm Z_{1-a/2}\sqrt{p\left(1-p\right)}}$$

Problems of Wald CI

- 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 adjustment

- A simple fix for the problem is to add 2 successes and 2 failures
- That is let p = (X + 2) / (n + 4)
- Lead to the Agresti-Coull interval

$$\boxed{p \pm Z_{1-a/2}\sqrt{p\left(1-\ p\right)}}$$

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

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

- The likelihood is the factor by which our prior beliefs are updated to produce conclusions in the light of the data

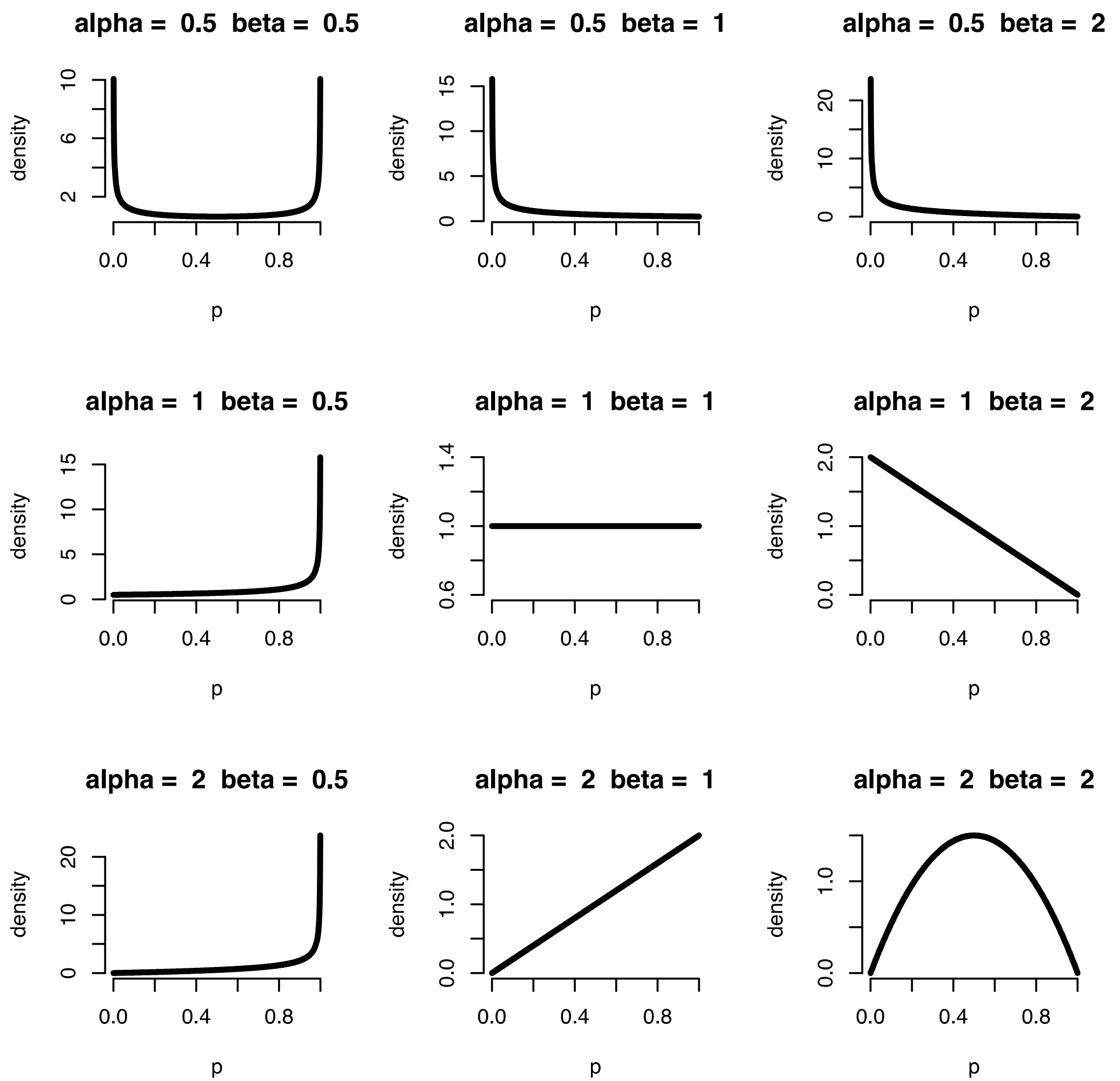
Beta priors

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

Some beta distributions



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**

Posterior ∝ Likelihood × Prior

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

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

Posterior mean

- **This density is just another beta density with parameters α* =x+α and β =n−x+β**

$$E[p \mid X] = \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}}$$

$$= \frac{x+\alpha}{x+\alpha+n-x+\beta}$$

$$= \frac{x+\alpha}{n+\alpha+\beta}$$

$$= \frac{x}{n} \times \frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+\beta}{n+\alpha+\beta}$$

$$= \text{MLE} \times \pi + \text{Prior Mean} \times (1-\pi)$$

Posterior variance

- **Posterior variance is**

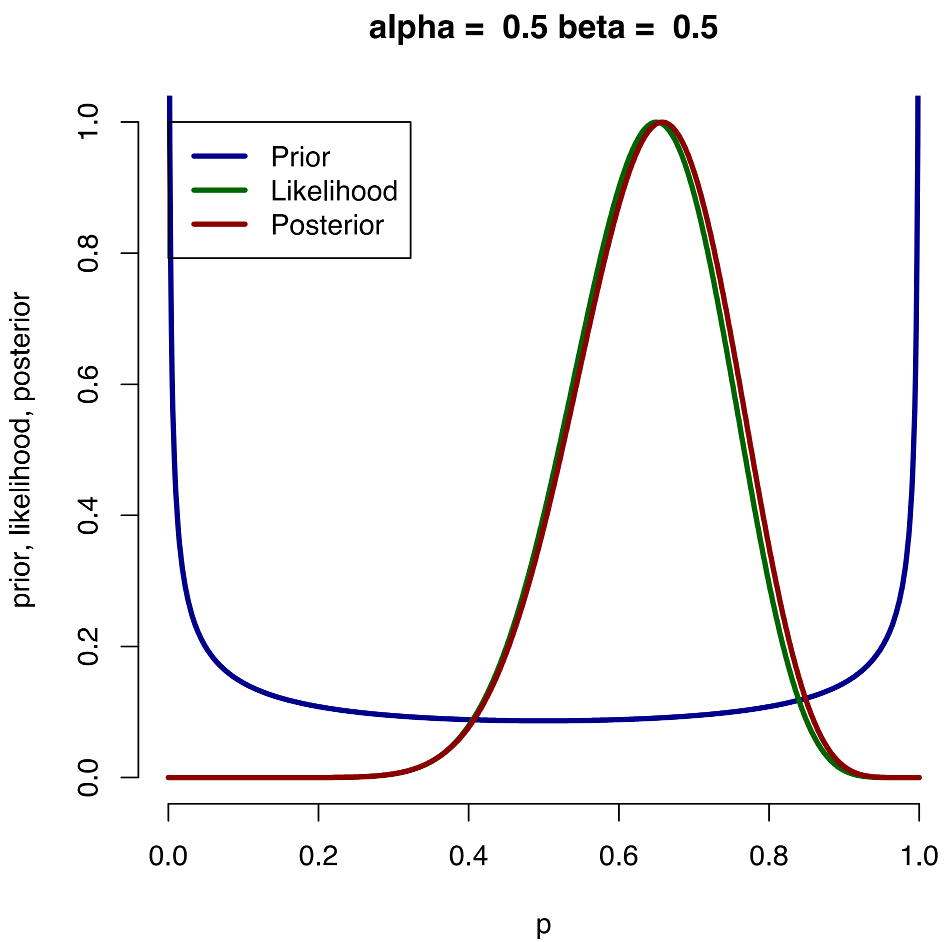
$$\mathrm{Var}(p \mid x) = \frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta})^2(\tilde{\alpha}+\tilde{\beta}+1)} = \frac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)}$$

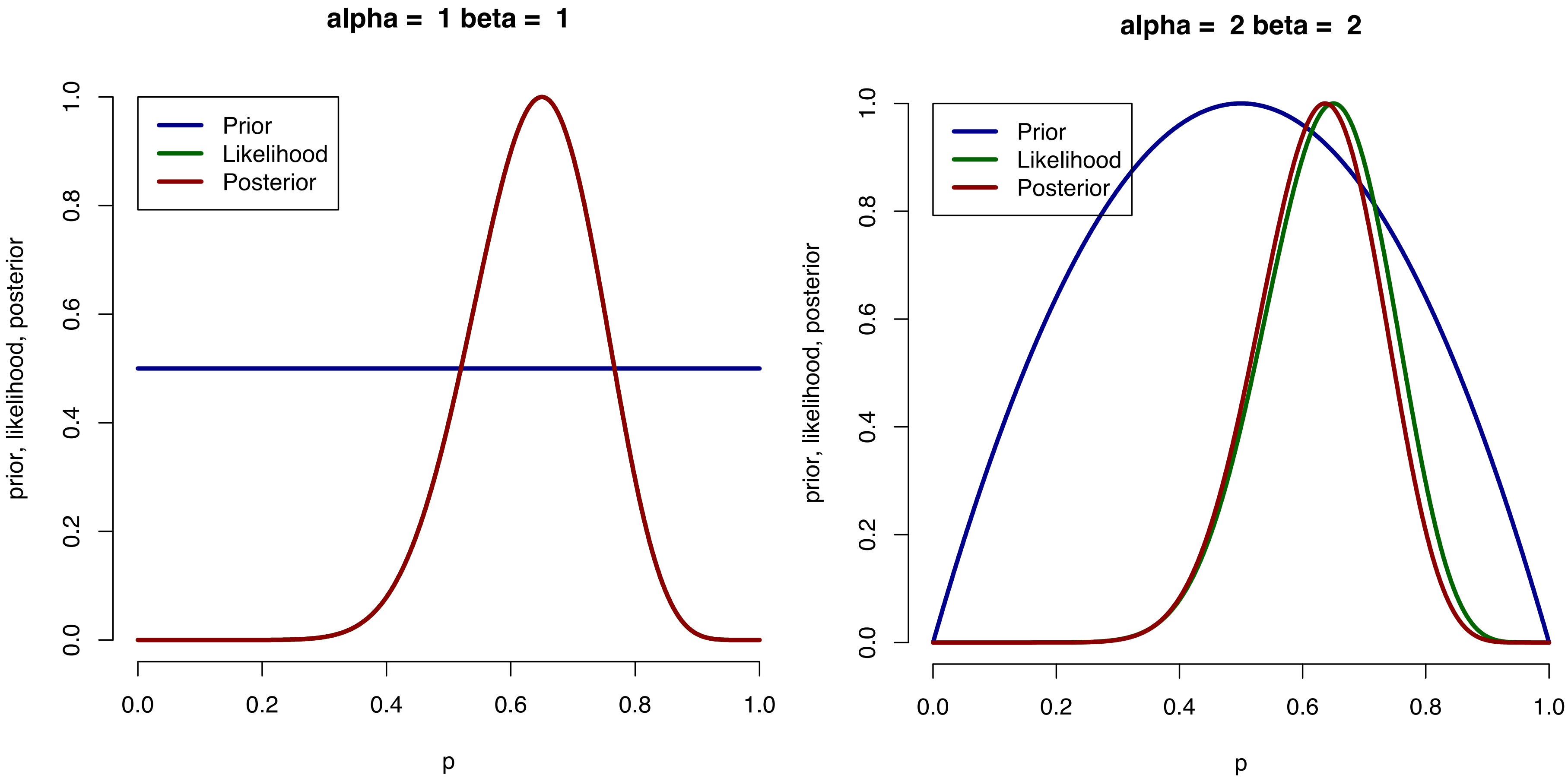
- **Let $p^* = (x + \alpha)/(n + \alpha + \beta)$ and $n^* = n + \alpha + \beta$ then we have**

$$\mathrm{Var}(p \mid x) = p^*(1 - p^*) / (n^* + 1)$$

Jeffreys prior

- **The "Jeffrey's prior" has some theoretical benefits puts $\alpha = \beta = 0.5$**





alpha = 1 beta = 1
Prior
Likelihood
Posterior
prior, likelihood, posterior
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
p
alpha = 2 beta = 2
Prior
Likelihood
Posterior
prior, likelihood, posterior
1.0
0.8
0.6
0.4
0.2
0.0
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.