The Urn Model, Simulation, and Resampling methods

Deborah Nolan
University of California, Berkeley

Overview

- Introductory example
- Basics elements of the Urn model
- Simulation studies
- Hypothesis testing
- Confidence Interval and Survey Sampling

Salk Field Trials

Polio left many children crippled and some who could survive only with respirators

Field Trial

- Jonas Salk had developed a vaccine that showed promising results in the laboratory
- Ready to test on a larger scale
- 1954 - a year-long field trial conducted
- 1.8M children from 217 areas in US, Canada, and Finland
- Cost $17.5M

Random assignment

- A carefully applied *chance process* that gives each volunteer an equal probability of getting vaccine or salt solution.
- The randomization turns potential biases into chance error, i.e. the two groups will be similar with respect to factors that might bias the result, even if these factors are unobserved.

Placebo controls

- Children were inoculated with simple salt solution.
- **Placebo effect** – bias that comes from reassurance of taking an otherwise worthless drug substitute.

Double-blind evaluation

- Neither children nor physicians who evaluated their subsequent health status know who had been given the vaccine/salt.
- Behavior of parents and children wouldn't be affected by knowledge that they received the vaccine.
- Physician wouldn't be biased when facing a borderline case that was difficult to diagnose.

Randomized placebo control method

- Randomly assign treatment to children.
- Maximum efforts to eliminate "observer bias" by use of placebo (injection with salt solution) and "blinding".

Outcome

- 56 of the children receiving the vaccine contracted polio
- 142 of those who did not receive the vaccine contracted the disease.
- It seems like the vaccine works, but could this result have easily happened by chance?
- The randomized design let's us answer that question

Ineffective Vaccine Scenario

- The 198 children who contracted polio would have become sick whether or not they received the vaccine
- It was only the randomization that led to 142 of the 198 sick children being assigned to the salt solution. Nothing else was going on

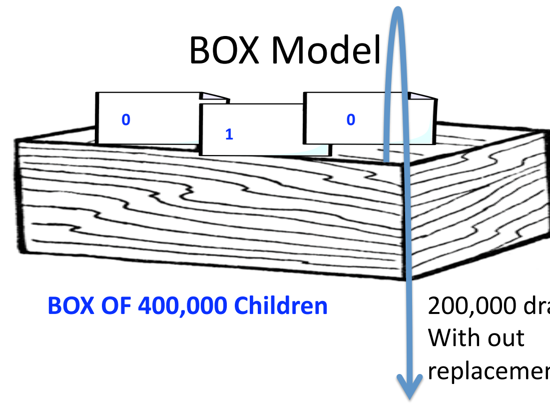
How Likely Is That?

- 400,000 children: 198 are sick and 399,802 are healthy.
- 200,000 of the 400,000 children are picked at random to receive the vaccine.
- What's the chance that as few as 56 or fewer of the sick children are given the vaccine?

The Urn Model

- Marbles: 400,000 – one for each child
  - 0 for healthy and 1 for contract polio
  - 399,802 0s and 198 1s
- Draws:
  - 200,000 draws with out replacement from the urn (the treated children)
- Summary:
  - Sum the values drawn

BOX Model

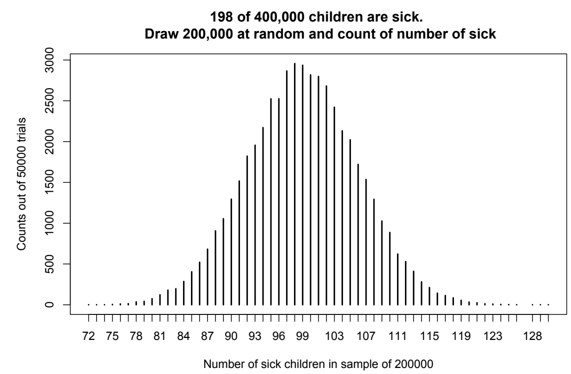


BOX OF 400,000 Children

200,000 draws
With out
replacement

Sum of draws = number of polio cases in treatment group

```
> vals = c(0, 1)
> counts = c(399802, 198)
> urn = rep(vals, times = counts)
> results = replicate(50000,
                 sum(sample(urn, size = 200000, replace = FALSE)))
> mean(results)
[1] 98.97672
> sd(results)
[1] 7.029245
> sum(results <= 56 )
[1] 0
> hist(results, breaks = 50)  OR plot(table….)
```



- In 50,000 trials it never happened – we did not get a single sample of 200,000 with 56 or fewer cases
- The Salk field trial was a turning point in medical research

Another example

Calcium and blood pressure

- Experiment: Study the effect of calcium supplements on blood pressure for male subjects
- Randomized into 2 groups –
  - One received calcium supplements for 12 wks
  - One received a placebo for 12 wks
- Double-blind experiment
- Response: reduction in blood pressure:

  (initial bp – bp after 12 wks)

Outcome

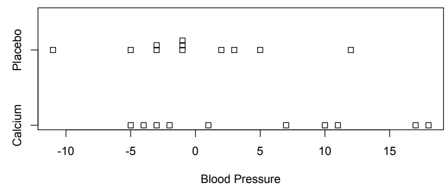
- Treatment group:

  7, -4, 18, 17, -3, -5, 1, 10, 11, -2
- So the first persons blood pressure decreased by 7 from the start to the end of the program.
- The second person's increased by 4, the third decreased by 18, and so on.
- For the group that received the placebo:

  -1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3

Informal Analysis

Those in the calcium group reduced their blood pressure by 5 on average, but..



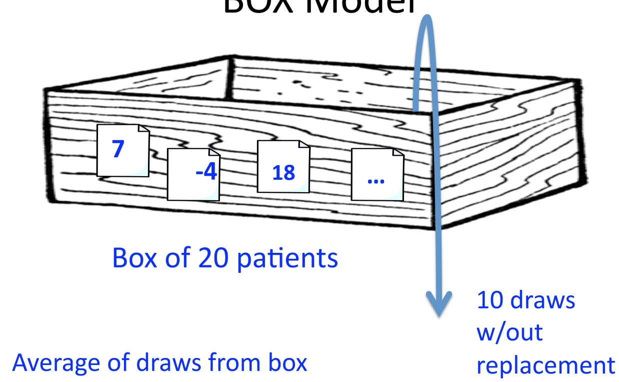
Informal Analysis

- The distributions have roughly the same spread
- The mean of the calcium group looks to be about 5 mm higher
- The mean of the control group looks to be about 0 mm

Scenario

- What if the calcium makes no difference, and it's just by chance that in the random assignment the calcium group got more people who had a lower blood pressure after 12 weeks.
- We can use an urn model to examine this chance process.

BOX Model



The Urn Model

- Marbles: 20
  - Values: blood pressure change 7, -4, 18, 17, -3, -5, 1, 10, 11, -2, -1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3
  - One for each subject
- Draws:
  - 10 draws from the urn (the calcium treatment)
- Summary:
  - Average change in blood pressure

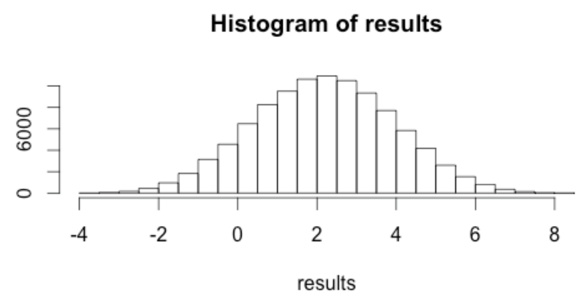
Using the urn in R

If the calcium supplement has no effect then the all subjects would have the same response whether they received the supplement or not

```
> avgs = replicate(100000,
         mean(sample(bp, 10,
                  replace = FALSE)))
> plot(table(avgs))
```

Let's find the approximate distribution of the sample average under the Urn Model.

```
> sum(results >=5)/100000
[1] 0.0616
```



The Basic Urn Model

Information about urn:

- Marbles:
  - What values do we write on the marbles?
  - How many marbles of each value do we have?
- Draws:
  - How many draws do we take from the urn?
  - We we replace marble between draws?
- Summary:
  - How do we summarize the values drawn?

Using the Urn in R: Simulation

- Set up an urn with marbles
- Sample from the urn (specify number of draws and with or without replacement)
- Do something with the results, e.g. take sum
- Repeat many, many, many times
- Use empirical distribution to approximate true distribution

Setting up the urn

```
> vals = c(…)    # values on marbles

> counts = c(…)  #counts for each val

# Create the urn
> urn = rep(vals, counts)
```

Sampling from the urn in R

Draw from the urn *n* times with replacement

```
> sample(urn, n, replace = TRUE)

# sum the draws from the urn
> sum(sample(urn, n, replace = TRUE))

# repeat this process m times
> replicate(m, sum(sample(urn, n,
                              replace = TRUE)))
```

Use urn to answer questions

- Roll a die 9 times and take the sum of the rolls.
- What's the chance the sum is 25 or less?

Simulation

- Draw 9 marbles with replacement from the urn, record the sum.
- Repeat 10,000 times to get 10,000 sums
- Find the proportion of those 10,000 sums that are 25 or less
- This **empirical proportion** should be close to the **chance** that the sum will be 25 or less

Steps in R

- Set up the urn with the "marbles"

```
urn = 1:6
```

- Sample from the urn

```
sample(urn, 9, replace = TRUE)
```

- Do something with the results, e.g. take sum

```
sum(sample(urn, 9, replace = TRUE) )
```

Steps in R

- Repeat many, many, many times

```
observations = replicate(10000,
          sum(sample(urn, 9, replace = TRUE) )
```

- Make a table of the proportions of times observed each value

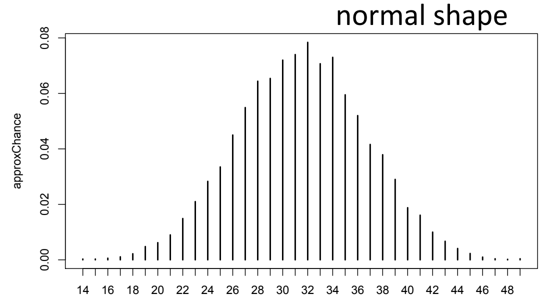
```
approxChance = table(observations)/10000
```

cumsum(approxChance)["25"]

25

0.1222

The PMF has a normal shape



Connection to Probability

- The Urn model is a simple chance process
- Consider the first draw:
  - Value drawn is random
- Avg of urn = Expected value for one draw
- SD of urn = SD for one draw

Connection to Probability

- Consider *sum* of *n* draws:
- Expected Value of sum = Avg of urn
- Standard Error for sum = root(n)SD of urn
- When draw without replacement
  - Expected value remains same
  - Correction factor for SE

Our simple example

- We can find the expected value for the sum:

$$n \times Urn\,AVG = 9 * 3.5 = 31.5$$

- We can determine the standard deviation for the sum:

$$\sqrt{9}\sqrt{\frac{6^2-1}{12}} = 5.12$$

- But how do we compute the chance the sum is 25 or less?

Why use the Urn Model?

- Connects statistical techniques,
  - Hypothesis Tests
  - Confidence Interval

  to underlying chance process
- If clear on the chance process then clear on the statistical method to apply
- Simulation study powerful tool for data analysis

Why use the Urn Model?

- Simulation study powerful tool for understanding a chance process
- Discover important properties – law of large numbers, CLT, etc.
- Offer alternative approach to learning concepts
- Help student discover features of a chance process, e.g. root(n) behavior

Urn model for distributions

- Discrete Uniform(1, 2, …, m)
  - *m* marbles
  - each marble has a unique value: 1 to *m*
  - Draw a marble from the urn, note the value, replace it
- Binomial(n,p)
  - *m* marbles in the urn
  - marbles are marked with a 1 (for success) or a 0 (for failure)
  - proportion *p* of the marbles are marked 1
  - Draw *n* marbles with replacement and sum the values
- Hypergeometric(m, n_1, n_2)
  - $n_1 + n_2$ marbles in the urn
  - n_1 marbles are marked with a 1 (for success)
  - n_2 marbles are marked 0 (for failure)
  - Draw *m* marbles **without** replacement and sum the values

Hypothesis Testing

Testing

- Permutation tests
- Fisher's exact test
- Chi-square tests
- Can also use for *z* and *t* tests, where place assumptions on the contents of the urn (and use large sample theory)

Chi-square test

- Grade expect in class and Gender

```
> table(grades, sex)
        sex
grades   F  M
     A   9 22
     B  21 31
     C   8  0
```

Is gender independent of expected grade?

Urn

- Values on marbles:

```
vals = c("A","B", "C")
```

- Counts of each type:

```
counts = c(31, 52, 8)
```

- Urn:

```
urn = rep(vals, counts)
```

Simulate

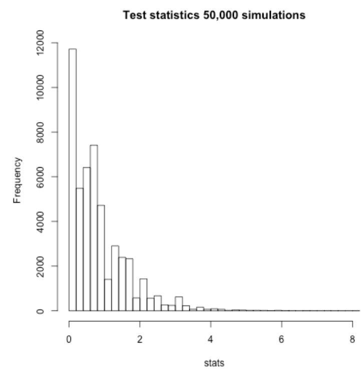
- Sample from urn:

```
table(sample(urn, 53, replace = FALSE))
A  B  C
15 33  5
```

- Repeat many times and calculate the test statistic for each table

```
results = replicate(50000,
  sum((table(sample(urn, 53)) –
                   expect)^2/expect)))
```

```
> sum(stats >= 5.538192)
[1] 45
```



Survey Sample

Simple Random Sampling

- The random selection process gives us samples that are tend to be **representative of the population**
- The random selection process means we can **use chance** to determine the chance someone is included in the sample AND the **probability distribution of the sample average**

Urn Model for Simple Random Sampling

- One marble for each unit in the population
- Value on the marble is the response to the question (We consider only one question)
- Draw *n* marbles with out replacement to get a sample
- Use a summary of the values drawn to make inference about the population

3 Views – A Triptych

- Population – Describes the Urn
- Your Sample – What you found from the execution of the chance process, what you drew from the urn
- Sampling Distribution – Understand the behavior of the chance process and use it along with Your Sample to make statements about the Population; We find this behavior by simulating with the urn or by applying large sample theory

Population View

Population = Urn

- N marbles -
- Unknown distribution of values
- Unknown Population Average
- Unknown Population SD

Sample View

Sample

- n draws without replacement
- Observe
  - Sample distribution
  - Sample mean
  - Sample SD

Simple Random Sample should resemble population – due to chance process

Sampling Distribution View

Probability Model for how the sample might turn out

- EV(Sample AVG) = Population AVG
- SE(Sample AVG) = $\frac{SD(Pop)}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

The Pop AVG and SD are unknown, and we don't know the sampling distribution of our statistic

Inference about the Population

- SRS tells us the sample should be representative of the population
- Create a **bootstrap population** from the sample
- Find the the approximate **sampling distribution** of the statistic for the bootstrap population and use it to make inferences about the true population

Cluster Sampling

Urn Model for Cluster Sampling

- Urn : packets of marbles
- N = number of packets in the urn
- n = number of draws / packets from the urn
- Draw packets without replacement
- Open up the packet and take all of the marbles from the packet

Stratified Sampling

Urn Model for Stratified Sampling

- M urns: one for each stratum
- Urn *i* has N_i marbles in it
- Draw n_i marbles without replacement from Urn *i*

Terence's Stuff: Simulation IMS Bulletin May 2011

It now seems to me that we are heading into an era when *all* statistical analysis can be done by simulation.

We don't need likelihood functions; we just need to know how to simulate from the models we entertain for our data.

That's been a *sine qua non* of statistical analysis for some time now. …

We don't need theory to tell us our method works; we just need to simulate and see.

Resources

- Box model in *Statistics* by Freedman, Pisani, Purves
- R Video #13: http://www.stat.berkeley.edu/share/rvideos/R_Videos/R_Videos.html