Sampling

Sampling

- How do we draw samples in such a way that polls accurately reflect what is going to happen?
- A sample is a small share of the population in that we are interested in
- How to construct accurate samples?

Sampling

- Example: We want to know the voting intentions of American voters

- We can hardly ask all eligible voters about their intention

- So we take a sample from the population and ask those in the sample

Literary Digest Poll was the largest poll ever undertaken: 2.4 million individuals

PREDICTION: LANDON WINS LANDSLIDE in 1936

RESULT: ROOSEVELT WON LANDSLIDE

Chicago Daily Tribune
Home
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in City
REPUBLICAN TICKET AHEAD
RECORD CITY VOTE SEEN IN LATE TALLIES

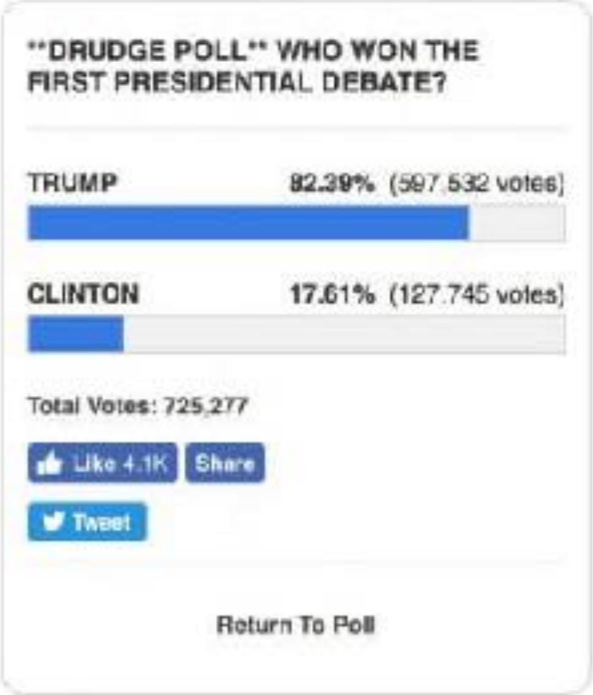
Dewey predicted as winner by large margin in 1948

Truman wins by 5 percentage points

Why were these initials polls so bad?

Even with more than 2 million respondents, how where these polls so far off?

It doesn’t matter how large your sample is, if it is not representative of the overall population!



DRUDGE POLL WHO WON THE FIRST PRESIDENTIAL DEBATE?

TRUMP 82.39% (597,532 votes)

CLINTON 17.61% (127,745 votes)

Total Votes: 725,277

Literary Digest Sample

Mail questionnaire to 10 million people

Addresses came from phone books and club memberships

Much less likely to reach poor voters who often didn’t have phone or were rarely members of clubs

Sample was **biased**

When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.

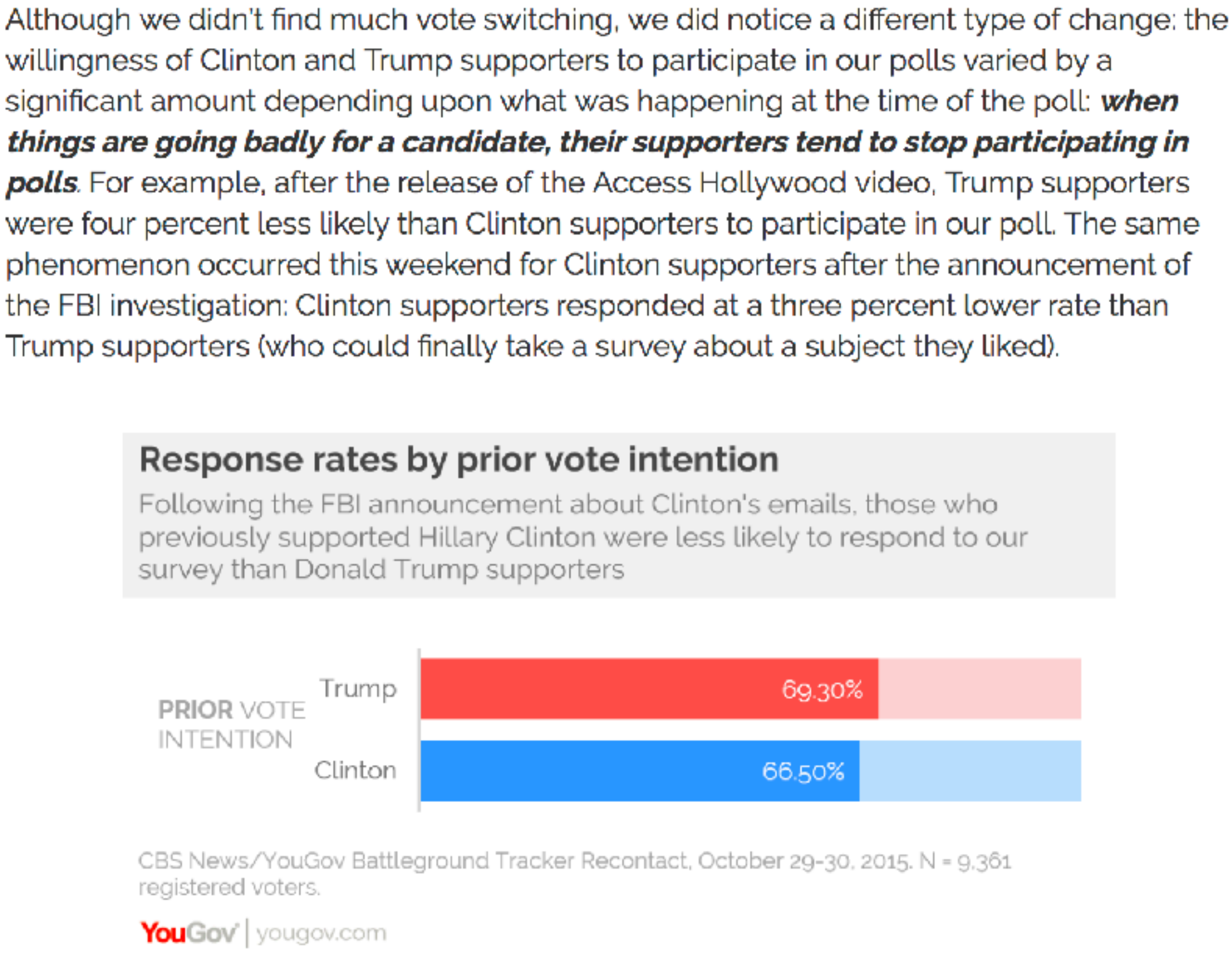
Literary Digest

In addition to a biased sample, the digest got thrown off by something else

Of their 10 million requests only 2.4 million individuals replied

If **non-respondent** and **respondents** differ from each other, the poll will be off even further

Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.



Although we didn't find much vote switching, we did notice a different type of change: the willingness of Clinton and Trump supporters to participate in our polls varied by a significant amount depending upon what was happening at the time of the poll: ***when things are going badly for a candidate, their supporters tend to stop participating in polls***. For example, after the release of the Access Hollywood video, Trump supporters were four percent less likely than Clinton supporters to participate in our poll. The same phenomenon occurred this weekend for Clinton supporters after the announcement of the FBI investigation: Clinton supporters responded at a three percent lower rate than Trump supporters (who could finally take a survey about a subject they liked).

Response rates by prior vote intention

Following the FBI announcement about Clinton's emails, those who previously supported Hillary Clinton were less likely to respond to our survey than Donald Trump supporters

CBS News/YouGov Battleground Tracker Recontact, October 29-30, 2015. N = 9,361 registered voters.

YouGov | yougov.com

What about quota sampling?

What is it?

Why does it not work?

What is quota sampling?

Quota sampling is an attempt to survey members of important groups according to their share in the population

Why does it not work?

Interviewers often don’t randomly choose within groups

Variables that are not part of the quota may affect the outcome

In quota sampling, the sample is hand-picked to resemble the population with respect to some key characteristics. The method seems reasonable but does not work very well. The reason is unintentional bias on the part of the interviewers.

So how should we sample?

Randomized Sampling

Think of all voters sitting in a box and a survey firm randomly drawing voters and asking them about vote intentions

Random draws without replacement would give us an unbiased estimate of the population

Everybody has the same chance of being in the sample

Simple random sampling means drawing at random without replacement

Simple Random Sampling is really hard

So how is it done?

- No discretion for the interviewer on who to survey

- Preselected potential respondents based on chance alone

- Multistage cluster sampling or today random digit dialing

Multistage cluster sampling

1. Randomly selected towns in regions
2. Within towns, randomly select wards
3. Within wards, randomly select precincts
4. Within precincts, randomly select households
5. Even in households, detailed instructions on who to interview

Random Digit Dialing

In the 90s survey firms realized that now almost everybody has a landline phone

- Computer randomly calls numbers
- **Not from phonebook,** but by randomly combining digits
- Why might phonebooks be problematic?
- What has happened in the last ~15years or so that might make things problematic?

- What is weighting and why would we do it?
- How is it different from quota sampling?

Sampling at random: percentages

- For a random sample, the expected value of the sample percentage is equal to the population percentage

- SE for percentages = (SE for number/size of sample)*100

- When we increase the sample size by a factor of X, the standard error gets smaller (divide by square root of X)

Sampling at random: percentages

A university has 25,000 students, of whom 10,000 students are older than 25. The registrar draws a simple random sample of 400 students.

A. Find the expected value and SE for the number of students in the sample who are older than 25.

B. Find the expected value and SE for the percentage of students in the sample who are older than 25.

C. The percentage of students in the sample who are older than 25 will be _____, give or take _____ or so.

Normal Approximation

- Again, as before we can use the normal approximation for percentages
- Just use expected value and standard error and transform to standard units
- Everything else is the same
- Remember to translate the questions to box problems

But when we sample it is without replacement!

- What do we do?

But when we sample it is without replacement!

- What do we do?

- It doesn’t really matter, as long as the population is large enough such that the composition doesn’t really change throughout out our draws

Guess what?

College Basketball Season Starts Today

But when we sample it is without replacement!

- But we can correct for it anyways:
  sqrt((population size - sample size)/(population size -1))

Example

During the 2004 presidential election the two major parties want to know how they are doing in Texas and New Mexico. There are about 1.5 million eligible voters in NM and 15 million in Texas. Firms in both states take samples of 2500 voters. Assume for now that both parties have 50% support in each state. What are the expected values and SEs in percent?

Example

One public opinion poll uses a simple random sample of size 1500 draws from a town with a population of 25,000. Another poll uses a simple random sample of size 1,500 from a town with population 250,000. The polls are trying to estimate the percentage of voters who favor single-payer health insurance. Other things being equal:

A. The first poll is likely to be quite a bit more accurate

B. The second poll is likely to be quite a bit more accurate

C. There is likely not to be much difference in accuracy between the two

The accuracy of percentage

- Or: how do we make inference about the contents of the 📦 — the population?

- We use our sample to make inference about contents of the population

   1. We assume that the random sample is pretty good at representing the population

The accuracy of percentage

- Or: how do we make inference about the contents of the 📦 — the population?

  1. Take the expected value from sample and apply it to population

  2. How do we calculate the standard errors?

The accuracy of percentage

- Or: how do we make inference about the contents of the 📦 — the population?

    1. Take the expected value from sample and apply it to population

    2. How do we calculate the standard errors?

        - Need the standard deviation of the population!

        - Assume fractions 1/0 of sample apply to population

The bootstrap: When sampling from a 0-1 box whose composition is unknown, the SD of the box can be estimated by substituting the fractions of 0's and 1's in the sample for the unknown fractions in the box. The estimate is good when the sample is reasonably large.

Example

Sample 400 students from a population of 25,000 are asked whether they lived at home or not. We want to find out how many live at home.

317 say they live at home, 83 do not.

Expected value: 79%

What is the standard error?

Example

Sample 400 students from a population of 25,000 are asked whether they lived at home or not. We want to find out how many live at home.

317 say they live at home, 83 do not.

Expected value: 79%

> What is the standard error? Apply the proportions to the population:
> SD = sqrt(0.79 * 0.21) = 0.41
> SE = sqrt(400) * 0.41 = 8

So we might be 8 students or so off, but let's convert that to %

(8/400)*100% = 2% so our survey result is 79% +/- 2%

Confidence Intervals

- Now that we make inference from the sample to the population we want to quantify our confidence with respect to the results

- The SE helps us quantify the uncertainty, but we can use the normal approximation to create confidence intervals

- Confidence Intervals give us an idea of how confident we should be that our parameter of interest is within the interval

Confidence Intervals

- Expected value +/- 1 SE is the 68% CI
- Expected value +/- 2 SE is the 95% CI
- Expected value +/- 3 SE is the 99.7% CI

Example

Random sample of 1,600 people for a town of 25,000. We want to figure out the share of Democrats in this town. In the sample we found 917 Democrats. What is the expected value for Democrats in the population and the 95% confidence interval?

Confidence Intervals

Remember: the confidence interval is based on the normal approximation & using the sample to infer population shares. If either is not accurate, the SE and CI will be wrong!

- Need large and accurate sample!

- Box not to be too lopsided!

Something weird about confidence intervals

- The probabilities we are talking about are really about our samples and not the population…
- Even though we say the we are 95% confident that there are between 52% and 57% Democrats, this is either true or not
- Confidence intervals are about the frequency (share) with which our samples correctly include the parameter of interest

A confidence interval is used when estimating an unknown parameter from sampled data. The interval gives a range for the parameter, and a confidence level that the range covers the true value.

Reminder about multiple regression

- Regression with more than 1 independent variable

- Intercept is the expected value of Y when both X1 and X2 are zero

- Coefficient on X1 is the change in Y associated with an increase of X1 by 1, holding X2 constant.

- Coefficient on X2 is the change in Y associated with an increase of X2 by 1, holding X1 constant.

Reminder about multiple regression

- RMSE is still the same

Accuracy of Averages
— everything stays the same—

Assume we are interested in the average value of a population

- Maybe average income, average partisanship, average education

- We use our sample to draw inference on the population

- The average is just the sum of draws divided by the number of samples —- SIMPLE!

Assume we are interested in the average value of a population

- The SE of the average is just the SE of the sums divided by the number of draws —- SIMPLE!

Example

- 25 draws w. replacement from 1, 2, 3, 4, 5, 6, 7
- Expected value of average will be?
- SE of average will be?

Example

- 25 draws w. replacement from 1, 2, 3, 4, 5, 6, 7
- Expected value of average will be? 4
- SE of average will be? 0.4
- What if we have 100 draws?

Averages

- EV of average: average of box
- SE of average: SE of sum divided by number of draws

SE of Sum vs SE of Average

- As the number of draws increase, the absolute SE (of the sum) increases
- What happens to the SE of the average?

Confidence interval and chances

- Normal curve applies again!

- Turns out averages are distributed normally, even if the contents of the box are not!

- So to estimate confidence intervals or other probability intervals we can use the normal approximation

Example

- EV= 4, SE = 0.2
- What are the chances that the average draws will be more than 4.2?

Exercise

A box contains 10,000 tickets. The numbers on these tickets average out to 50 and the SD is 20.

A) One hundred tickets are drawn at random with replacement. The average of these draws will be ___, give or take ____.

B) What if 100 draws are made without replacement?

C) What if 100 draws are made without replacement and there are only 100 tickets in the box?

Exercise

Two hundred draws are made at random with replacement from the box 1, 2, 2, 3. True or False?

What is the expected value for the average?

What is the SE for the average?

What is the chance the average is above 2.5?

Sample Average

- When making inference over population we again use the sample to do so
- We use the sample average to approximate the population average
- Use the sample SD to approximate population SD

Example

- We want to know the average income in a town of 25,000
- Survey organization asks 1000 families about their income
- Total income of 1000 families is $62,396,714
- SD of sample is 53,000
- What is the expected value of average income?
- What is the SE of the average income?

Example

- We want to know the average income in a town of 25,000
- Survey organization asks 1000 families about their income
- Total income of 1000 families is $62,396,714
- SD of sample is 53,000
- EV of Average 62,400 with SE of 1,700
- NOTE that the SE of the Average says nothing about the distribution of income in the town!

So many Standard Errors!

- SE for sum = sqrt(Draws)* SD of box
- SE for average = SE for sum/Draws
- SE for count = SE for sum, from 0-1 box
- SE for percent = (SE for count*100)/draws

Accuracy of Averages

- Works because as the number of samples increase, their average will be approximately normally distributed
- But, requires random sampling!

Exercise

A utility company serves 50,000 households. As part of a survey of costumer attitudes, they take a simple random sample of 750 of these households. The average number of television sets in the sample turns out to be 1.86 and the SD is 0.8. Find a 95% confidence interval for the average number of television sets in the 50,000 households.

Exercise

As part of the survey, all persons age 16 or over in the 750 sample households are interviewed. This makes 1,528 people. On average these people watched 5.2 hours of television on Sunday before the survey and the SD was 4.5 hours. Can we estimate the 95% confidence interval? If not, why not?

Exercise

A box contains 250 tickets. Two people want to estimate the average of the numbers on the tickets in the box. They agree to take a sample of 100 tickets, and use the sample average as their estimate. Person A wants to draw with replacement, person B without. Which would give the more accurate estimate? Does it make a difference?

Exercises

A town has 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters. A box model is used to work out the expected value and SE for the percentage of Democrats in the sample. Match each phrase on list A with a phrase or number on list B. (Items on list B may be used more than once or not at all)

A: population, population percentage, sample, sample size, sample number, sample percentage, denominator for sample percentage

B: number of 1's among draws, percentage of 1's among draws, 40%, box, draws, 1,000, 12,000

Exercises

You are drawing at random from a large box of red and blue marbles. Fill in the blanks:

A) the expected value for the percentage of reds in the _____ equals the percentage of reds in the ______.

B) As the number of draws goes up, the SE for the ____ of reds in the sample goes up but the SE for the _____ of reds goes down.

Exercises

According to the Census, a certain town has a population of 100,000 people age 18 and over. Of them, 60% are married, 10% have incomes over $75,000 a year, and 20% have college degrees. As part of a pre-election survey, a random sample of 1,600 people will be drawn from the population.

A) to find the chance that 58% or less of the people in the sample are married, a box model is needed. Should the number of tickets in the box be 1,600 or 100,000? Explain. Then find the chance.

B) To find the chance that 11% or more of the people in the sample have incomes over $75,000 a year, a box model is needed. Should each ticket in the box show the person's income? Explain. Then find the chance.

C) Find the chance that between 19% and 21% of the people in the sample have a college degree.

Exercises

You have hired a polling organization to take a simple random sample (without replacement) from a box of 100,000 tickets and estimate the percentage of 1’s in the box. Unknown to them the box contains 50% 0’s and 50% 1’s. How far off should you expect them to be:

A) if they draw 2,500 tickets

B) if they draw 25,000 tickets

C) if they draw 100,000 tickets.

Exercises

A box contains 2 red marbles and 8 blue ones. Four marbles are drawn at random. Find the SE for the percentage of red marbles drawn, when the draws are made

A) with replacement

B) without replacement

Exercises

A coin is tossed 10,000 times. Estimate the chance of getting

A) 4,900 to 5,050 tails

B) 4900 or fewer tails

C) 5,050 or more tails

Exercises

A biased coin has one chance in ten of landing heads. It is tossed 400 times. Estimate the chance of getting exactly 40 heads.

Exercises

The registrar keeps an alphabetical list of all undergraduates, with their current addresses. Suppose there are 10,000 undergraduates in the current term. Someone proposes to choose a number of random from 1 to 100, count that far down the list, taking that name and every 100th after it to draw a sample.

A) Is this a probability method?

B) Is is the same as random sampling?

C) Is there selection bias in this method?

Exercises

In about 1930, a survey was conducted in New York on the attitude of former slaves towards their owners and conditions of servitude. Some of the interviewers were black, some white. Would you expect the two groups of interviewers to get similar results? Give your reasons.

Exercises

In a certain city, there are 100,000 persons age 18 to 24. A simple random sample of 500 such persons is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in that city who are currently enrolled in college. Put a give or take number on the estimate.

A) the first step in solving the problem is:

1. Finding the SD of the box
2. Finding the average of the box
3. Writing down the box model

B) Now solve the problem