Unit 9, Chapter 20: Chance Errors in Sampling

Marius Ionescu

11/1/2011

Average of the draws

Example

Suppose that a city has an average income of $24,000 with SD of $10,000. A survey collects data at random from 400 households. What do we expected for the average of the survey?

Fact

The average of draws is

$$
\begin{aligned}
EV_{avg} &= \frac{EV_{sum}}{n} = Avg_{box} = \$24,000 \\
SE_{avg} &= \frac{SE_{sum}}{n} = \frac{SD_{box}}{\sqrt{n}} = 500
\end{aligned}
$$

Example

Example

Suppose that we want to estimate the percentage of students who live on the hill. We take a survey of 100 students. Assume that 1400 students live on the hill and 1200 do not, and assume that there is no bias.

- Estimate the count of students living on the hill from the sample.
- Estimate the percentage of students living on the hill from the sample and the chance error.
- What if we survey 1000 students?

Note

Fact

- *To estimate a percentage, first calculate the corresponding number, then convert to a percent relative to the size of the sample.*
- *Increasing the size of the survey by a factor of 10 decreases the SE by* $\sqrt{10}$ *(the EV is the same).*
- *This does not depend on the size of the population only on the size of the sample.*
- *That is why Gallup uses around 3000 people to estimated the outcome of the election.*

Correction factor

Fact

- *When drawing without replacement, the box does get a bit smaller, reducing the variability slightly.*
- *The for drawing without replacement is a little less than the SE for drawing with replacement:*

$$SE_{WITHOUTH\ REPL} = correction\ factor * SE_{WITH\ REPL}.$$

- *The correction factor is*

$$\sqrt{\frac{\#in\ box - \#\ of\ draws}{\#in\ box - 1}}.$$

- *When the number of tickets in the box is large relative to the number of draws, the correction factor is near 1 and can be ignored.*

Colgate Students living on the hill

Example

Recall that we had 1400 students living on the hill and 1200 not living on the hill.

- Do we need the correction factor if the sample has a size of 100?
- What if the sample size is 1000?

Using the normal approximation

Example

A town has 100,000 people age 18 or over. 10% of them have incomes over \$50,000 and 20% of the people have college degree. A sample of 1600 people is taken.

- Find the chance that 11% or more of the people in the sample have income above \$50,000.
- Find the chance that between 19% and 21% of the people in the sample have a college degree.

Chapter 21: Inference from the sample to the population

Fact

- *Inference=going from the sample to the population*

The bootstrap

Fact (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 reasonable large.*

Example

Example (Colgate students living on the hill)

Suppose that we extract a sample of 200 students at random from the 2600 students at Colgate. 108 of the students in the sample live on the hill. Estimate the percentage of students living on the hill. How accurate is our estimate?

Confidence intervals

Fact

- *The interval "sample percentage* ±1 *SE" is a 68%-confidence interval for the population percentage*
- *The interval "sample percentage* ±2 *SE" is a 95%-confidence interval for the population percentage*
- *The interval "sample percentage* ±3 *SE" is a 99.7%-confidence interval for the population percentage*

Example

Example (Democrats vs Republicans in Hamilton)

Suppose that a random sample of 400 people is taken to estimate the percentage of Democrats among the 4000 eligible voters in Hamilton. It turns out that 240 in the sample are Democrats. Find a 95%-confidence interval for the percentage of Democrats among all 4000 eligible voters.

Interpreting a confidence interval

Fact

- *The chances are in the sampling procedure, not in the parameter.*
- *The confidence interval gives a range for the parameter and a confidence lever that the range covers the true value.*
- *A sample percentage will be off the population percentage, due to chance error. The SE tells you the likely size of the amount off.*

Example

Example (Internet access in Hamilton)

Assume that a random sample of 200 households is taken to estimate the percentage of households with broadband internet among the 2000 households in Hamilton. It turns out that 140 of them have internet access. Find a 68% and a 95% confidence interval for the percentage of households with fast internet access in Hamilton.