Central Limit Theorem for proportions & means

It's freaking MAGIC people!

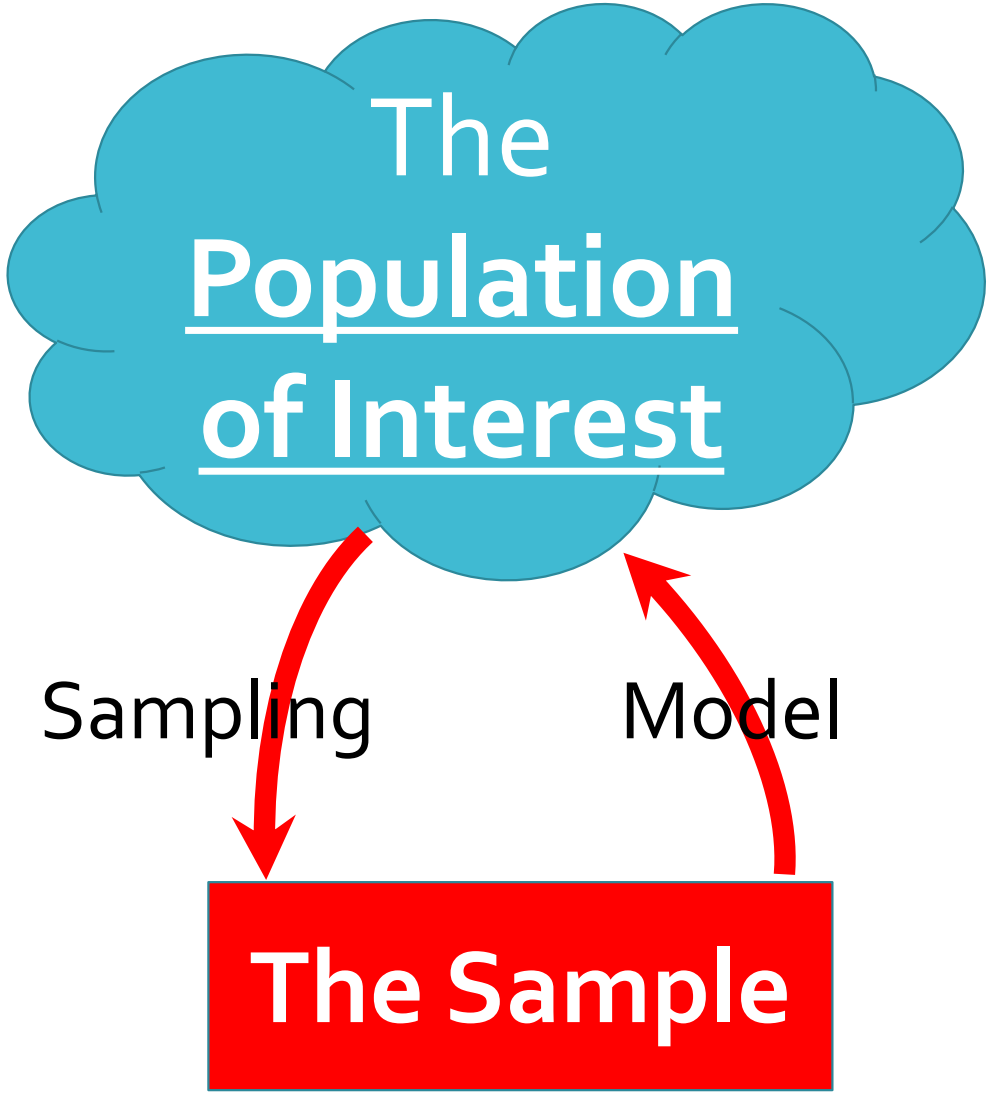
All models are wrong, but some are useful.
– G. Box.

- Sampling distribution models are important because
  - they act as a bridge from the real world of data to the imaginary world of the statistic and
  - enable us to say something about the population when all we have is data from the real world.

Freaking magic

- We would expect the histogram of the sample proportions to center at the true proportion, *p*, in the population.
- It turns out that the histogram is **unimodal**, **symmetric**, and **centered** at *p*.
- More specifically, it's ~~an amazing and fortunate~~ a FREAKING MAGICAL fact that a Normal model is just the right one for the histogram of sample proportions.

Remember this?



- We should not be surprised if 95% of various polls gave results that were near the mean but varied above and below that by no more than two standard deviations.
- This is what we mean by **sampling error**. It's not really an error at all, but just variability you'd expect to see from one sample to another. A better term would be **sampling variability**.

Applet to explore

- http://onlinestatbook.com/stat_sim/sampling_dist/

2 kinds of data

- Categorical
- Quantitative

Categorical

- First up, categorical data.

Assumptions

- Assumptions
  - Independence – Do we have good justification to believe samples are independent? What is it?

Conditions (or the Goldilocks principle)

- Conditions
  - Randomization – do we have random samples?
  - n<10% of population (is our sample too big?)
  - Success/Failure (is our sample too small?)
    - np≥10 AND n(1-p)≥10

Mean

- To use a Normal model, we need to specify its mean and standard deviation. We'll put μ, the mean of the Normal, at p.
- $\mu(\hat{p}) = E(\hat{p}) = p$
- [why is it p-hat?] <- big deal!

Standard Deviation

- $S.D.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

Normal model is:

- $\mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Quantitative

- Next up, quantitative data

The Fundamental Theorem of Statistics

- The Fundamental Theorem of Statistics is called the Central Limit Theorem (CLT).

The CLT

The Central Limit Theorem (CLT)

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.

Assumptions

- Assumptions
  - Independence – Do we have good justification to believe samples are independent? What is it?

Conditions (or the Goldilocks principle)

- Conditions
  - Randomization – do we have random samples?
  - n<10% of population when we draw w/o replacement (is our sample too big?)
  - Is our sample Large Enough (is our sample too small?)

Mean

- $\mu(\bar{x}) = E(\bar{x}) = \bar{x}$

Standard Deviation

- $SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Normal Model

- $N\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right)$

All models are wrong, some are useful.
-G. Box

- Be careful! Now we have *two* distributions to deal with.
- The first is the real world distribution of the sample, which we might display with a histogram.
- The second is the math world *sampling distribution* of the statistic, which we model with a Normal model based on the Central Limit Theorem.