Once again: the Central Limit Theorem and hypothesis testing

INFO 1301

Profs. Michael Paul and William Aspray

11/28/16

Central Limit Theorem

- Discussed but probably understated its importance earlier this semester
- CLT: the sample means $\bar{x}$ for a population will be distributed roughly in a normal distribution around the population mean μ
  - True even if the population does not form a normal distribution.
  - The sample size must be large enough (typically n>30) to smooth out the sample's distribution.
  - The sample must be randomly chosen.
- This means that roughly 68% of the sample means will lay within one standard error of the population and roughly 95% will lay within two standard errors.

Implications of CLT

1. If we have detailed info about the population, we can draw inferences about a properly drawn sample.
2. If we have detailed info (mean, standard deviation) about a properly drawn sample, we can draw inferences about the population.
3. If we have data describing a particular sample and data describing the population, we can infer whether that sample was drawn from that population (in a proper way).
4. If we have data from two samples, we can tell whether they were drawn from the same population (in a proper way).

Yet More About CLT

- The Standard Error is a measure of the dispersion of the sample means (all samples of size n)
- SE = standard deviation of the sample means
  - But it is also connected to the dispersion of the population (σ)
  - More specifically, SE = σ/ √n
- Implications
  - If SE large, samples are spread out widely around the population mean.
  - By choosing a larger n, the samples are clustered more closely to the population mean.
  - If the population has large dispersion, then the samples will tend to be more widely dispersed.

An Example relating to CLT

- US household income is heavily right-skewed (thanks to Bill Gates and Warren Buffett, and others like them)
- Can tell this from the large difference between median = 51.9 (K$) and the mean = 71.9 (K$)
- Even though the population is right skewed, when you take 1000 samples of size n=100, they will form a normal distribution around 71.9
- If you took similarly many samples of size n=25, you would have a normal distribution with twice the standard error.
- If you changed your population to single-earner households where the employed person was a public school teacher, the population would be less dispersed and thus the sample means would also be less dispersed (μ=56.3 in 2014, NCES)
- If you surveyed a sample of 100 people and the sample mean of their salary was 82.1, you would know either that this sample was not drawn from public school teachers or that it was not a random sample of teachers because 82.1 is far removed from 56.3) (In 2014 the average salary for top 10% of public school teachers ranked by compensation, the average is 88.9.)

Type 1 and Type 2 Error Tradeoff in Hypothesis Testing

- Which would be more socially problematic, a Type 1 error, Type 2 error, or both?
- Remember:
  - Type 1 Error: incorrectly rejecting a true null hypothesis
  - Type 2 Error: incorrectly retaining a false null hypothesis
- Spam filters. $H_{0:}$ a particular email is not spam.
- Screening for cancer. H_0: No cancer is present.
- Capturing terrorists: An individual is not a terrorist