Stats review

ADEC781001: Empirical Behavioral Economics

Lawrence De Geest (`lrdegeest.github.io`)

Boston College
Woods College of Advancing Studies

Random variables

- Random variable: any variable whose outcome is uncertain
  - Wind speed of hurricane when it hits land
  - Stock price of Apple in ten minutes
  - Number of Prussian soldiers kicked in the head by horses each year
- "Uncertain" does not imply "unpredictable"
- Most random variables have stable distributions
- Exploit these distributions to learn about random variable

Expected value

- All random variables have expected values $\mathbb{E}$
- Suppose we have random variable X
- Two kinds:
  1. Population mean: $\mathbb{E}[x_i] = \mu$
     - We cannot observe this
     - It's a parameter
  2. Sample mean: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
     - We observe this
     - It's a statistic
     - It's also an estimator for $\mathbb{E}[x_i]$
- We care about bias
  - A statistic is *unbiased* if $\mathbb{E}[\bar{x}] = \mathbb{E}[x_i]$

Variance

- All random variables have variance
- Two kinds:
  1. Population variance: $Var(x_i) = \mathbb{E}[(x_i - \mathbb{E}[x_i])^2] = \sigma_x^2$
     - We cannot observe this
     - It's a parameter
  2. Sampling variance: $Var(\bar{x}) = Var(\frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{\sigma_x^2}{N}$
     - We observe this
     - It's a statistic
     - It tells us how much the sample varies if we keep drawing samples
     - When $N \to \infty$ then $\frac{\sigma_x^2}{N} \to 0$ with implies $\mathbb{E}[\bar{x}] \to \mathbb{E}[x_i]$
     - This is the Law of Large Numbers

Hypothesis testing

- Suppose we have a sample X that is normally distributed with a mean of 10 and a standard deviation of 2
- Suppose we want to show $\mathbb{E}[x_i]$ is *significantly* different from zero
- Null hypothesis: $H_0 : \mu = 0$
- Alternative hypothesis: $H_0 : \mu \neq 0$
- First construct a t-statistic: $\frac{\bar{x}-0}{\sqrt{Var(\bar{x})}}$
- Significance: if H_0 were true, what is probability I would observe $\bar{x}$ in repeated sampling?
  - this probability is the p-value: area under the t-distribution
  - significance: check if $p < \alpha$ where α is a threshold set by the user (e.g. 0.05)
- Example: see `r-intro-stats-review.R`

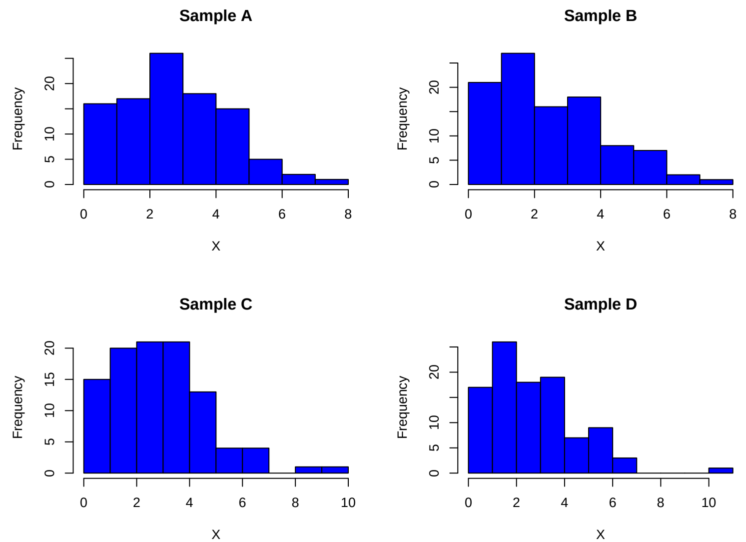
The Central Limit Theorem

- The t-test is valid for normally distributed variables
- Statistical inference is built around the assumption of normality
- But not all random variables out there are normally distributed
- Would be a lot cooler if they were
- Turns out there is way!

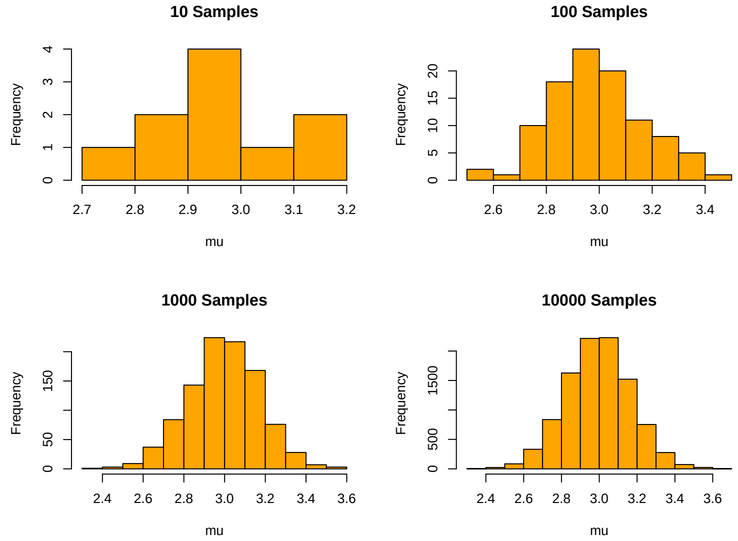
The Central Limit Theorem

- Suppose we drew samples from a population that followed a Poisson process
  - $X \in \{0, 1, 2, \ldots, \infty\}$
  - The population parameter is a constant μ but it is unknown
- For each sample we can calculate the expected value $\bar{X}$
- What will the distribution of $\bar{X}$ look like? Poisson, too?
- No! It turns out that if we keep sampling the sampling distribution for $\bar{X}$ will be a *Normal distribution*!
- This is the Central Limit Theorem

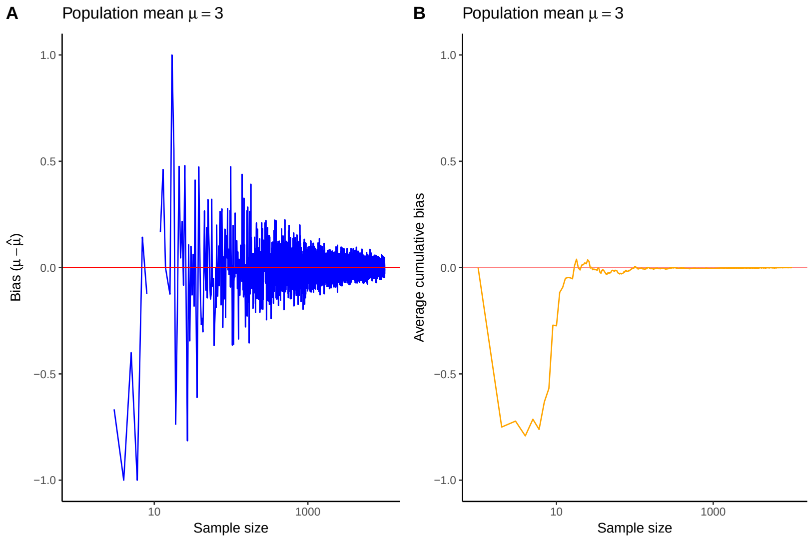
Example: sampling from a Poisson distribution with $\mu = 3$



Example: the sampling distribution



Example: the CLT and diminishing returns from sampling



Regression

- Example: see `r-intro-stats-review.R`
- Let's estimate the price of cars as a function of their weight
- Assume this function is linear: $Price_i = \beta_0 + \beta_1 weight_i + \varepsilon_i$
- The effect of car weight on car price is given by $\frac{\partial Price}{\partial weight} = \beta_1$
- Estimate β_1 using Ordinary Least Squares (OLS)
- The estimate is called "beta hat" ($\hat{\beta}_1$)

Regression: hypothesis testing

- Hypothesis test: weight has an effect on price
  - $H_0 : \beta_1 = 0$
  - $H_A : \beta_1 \neq 0$
- This is just a t-test
  - t-stat: $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$
    - $se(\hat{\beta}_1) = \frac{\sigma_e}{\sqrt{n}} \frac{1}{\sigma_x}$
    - Increasing in the variance of the residuals (σ_e).
    - Decreasing in sample size (n)
    - Decreasing in the variance of the regressor (σ_x).