STAT 311: Sampling Distributions

Y. Samuel Wang

Summer 2016

Logistics

- Midterm passed back at end of class
- Homework posted later today, due next Friday

Long run behavior

- Each individual outcome of a random variable is unpredictable
- Probability models define the long run behavior
- How can we apply these ideas to real world data which we gather

Long run behavior

- Each individual outcome of a random variable is unpredictable
- Probability models define the long run behavior
- How can we apply these ideas to real world data which we gather

Theorem

Law of Large Numbers: *Let $n \geq 1$, and let $X_1, X_2 \ldots X_n$ be a sequence of mutually independent random variables with finite mean μ and finite variance σ^2. Define*

$$\bar{x}_n = \frac{X_1 + X_2 + \ldots X_n}{n}$$

then for any $\epsilon, \delta > 0$ there exists some n such that

$$P(|\bar{X}_n - \mu| < \epsilon) = 1 - \delta$$

Implications of the LLN

- Given a large enough sample, we can achieve any level of precision (determined by ϵ and δ)
- Exactly how large n has to be depends on σ^2, which in practice we don't always know
- Typically, taking a larger sample will give us a more precise estimate of the true parameters
- Sample statistic will "converge" to the truth
- LLN cannot make up for bad study design

Central Limit Theorem

Let $n \geq 1$, and let $X_1, X_2 \ldots X_n$ be a sequence of mutually independent random variables such that $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$. Define $S_n = X_1 + X_2 \ldots X_n$

- Note that because we are summing random variables, $E(S_n) = \mu_1 + \mu_2 \ldots \mu_n$
- Since the random variables are independent $Var(S_n) = \sigma_1^2 + \sigma_2^2 + \ldots \sigma_n^2$
- If $Y = S_n - E(S_n)$, then $E(Y) = 0$ and $Var(Y) = Var(S_n)$
- If $Z = Y/\sqrt{Var(S_n)}$, then $E(Z) = 0$ and $Var(Z) = 1$

Central Limit Theorem

Let $n \geq 1$, and let $X_1, X_2 \ldots X_n$ be a sequence of mutually independent random variables such that $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$. Define $S_n = X_1 + X_2 \ldots X_n$

- Note that because we are summing random variables, $E(S_n) = \mu_1 + \mu_2 \ldots \mu_n$
- Since the random variables are independent $Var(S_n) = \sigma_1^2 + \sigma_2^2 + \ldots \sigma_n^2$
- If $Y = S_n - E(S_n)$, then $E(Y) = 0$ and $Var(Y) = Var(S_n)$
- If $Z = Y/\sqrt{Var(S_n)}$, then $E(Z) = 0$ and $Var(Z) = 1$

Can we say more?

Central Limit Theorem

Theorem

Central Limit Theorem: *Let $n \geq 1$, and let $X_1, X_2 \ldots X_n$ be a sequence of mutually independent random variables such that $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$. Define $S_n = X_1 + X_2 \ldots X_n$*

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \rightarrow_d N(0, 1)$$

The $\rightarrow_d$ symbol means converges in distribution. This means that the distribution of my random variable Z_n, which is simply formed from the X_i's, begins to look more and more like a standard normal distribution as n increases.

Central Limit Theorem

What if all my random variables are identical?

- $E(S_n) = n\mu$
- $Var(S_n) = n\sigma^2$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\frac{1}{n}S_n - \mu)}{\sqrt{n}\sqrt{\sigma^2}} = \sqrt{n}\frac{(\bar{x}_n - \mu)}{\sigma}$$

Central Limit Theorem

What if all my random variables are identical?

- $E(S_n) = n\mu$
- $Var(S_n) = n\sigma^2$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\frac{1}{n}S_n - \mu)}{\sqrt{n}\sqrt{\sigma^2}} = \sqrt{n}\frac{(\bar{x}_n - \mu)}{\sigma}$$

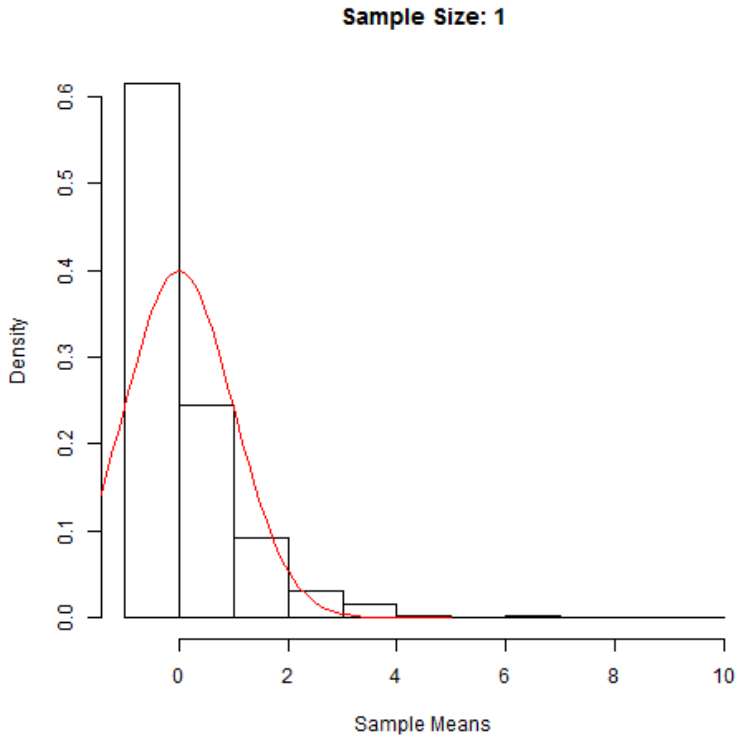
The distribution of (almost) any sample mean goes to a standard normal distribution when we account for the mean and standard deviation of the population.

Central Limit Theorem Caveats

- Note that this does say anything about individual random variables
- This only makes a statement about the $\bar{x}_n$ as n grows large
- Imagine taking many samples of size $n = 100$ and plot the histogram, that will look roughly normal; then take many samples of $n = 1000$, that will look even more like a normal distribution, etc.
- How large n has to be for a "roughly normal" histogram of by $\bar{x}_n$ depends on many things, but as long as n keeps growing, it will eventually get there

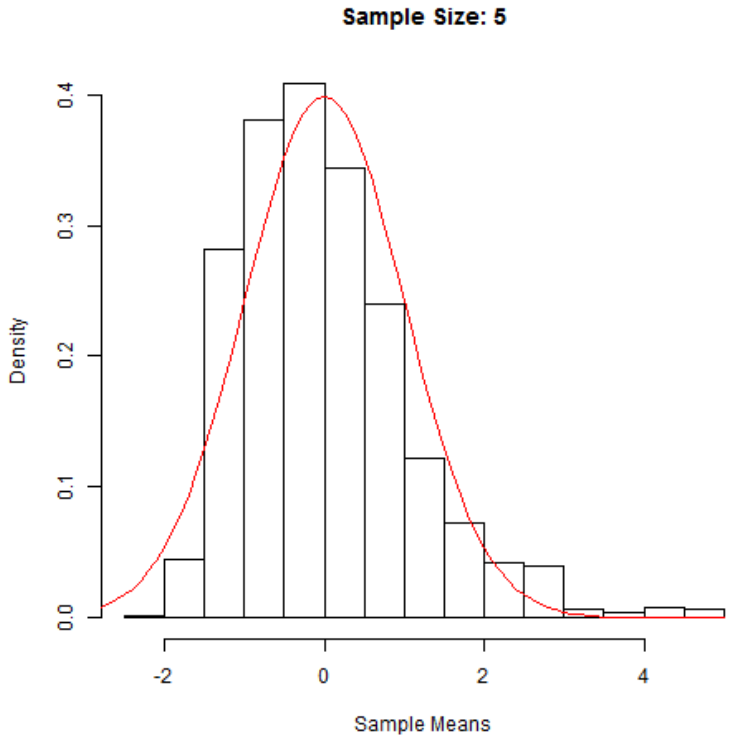
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$. Then repeat with a new sample 1000 times.



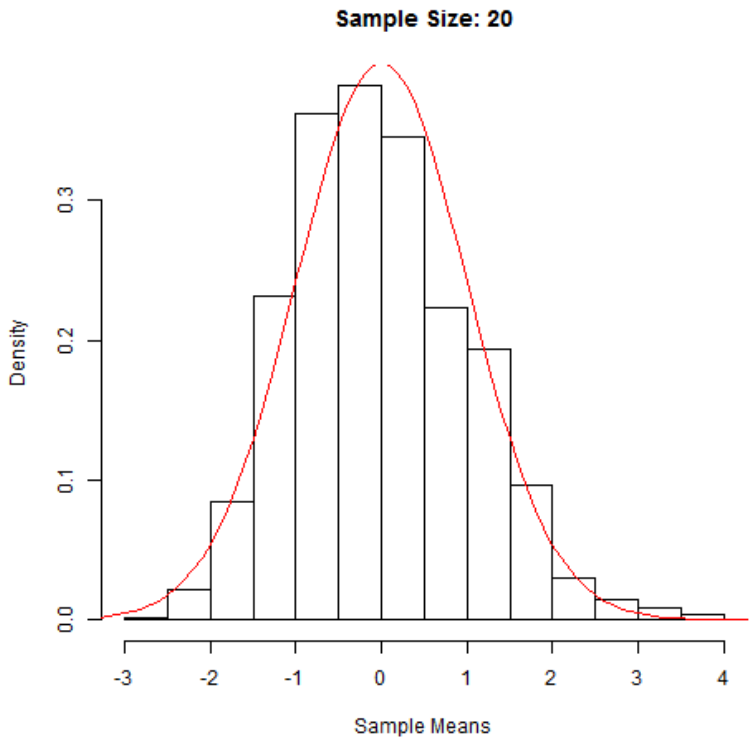
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$. Then repeat with a new sample 1000 times.



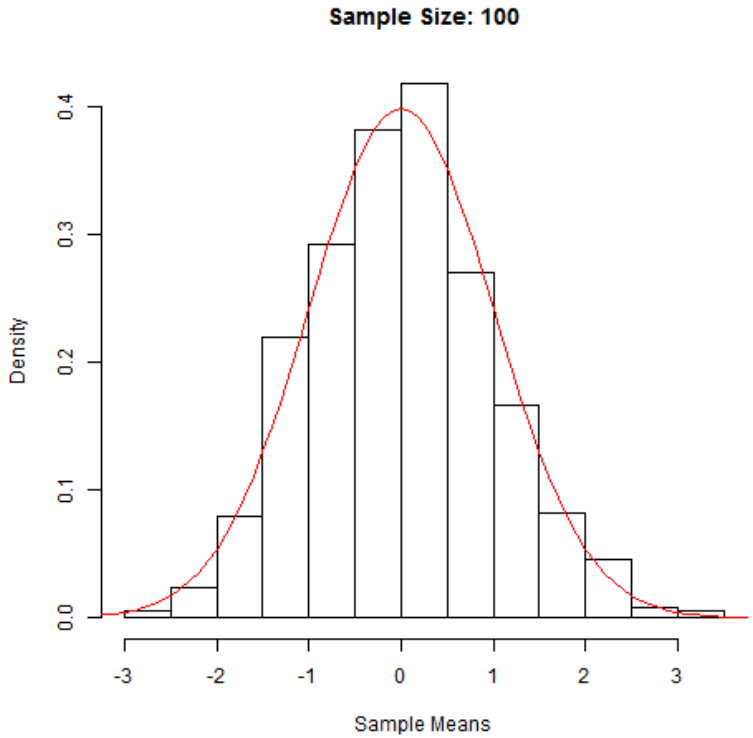
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$. Then repeat with a new sample 1000 times.



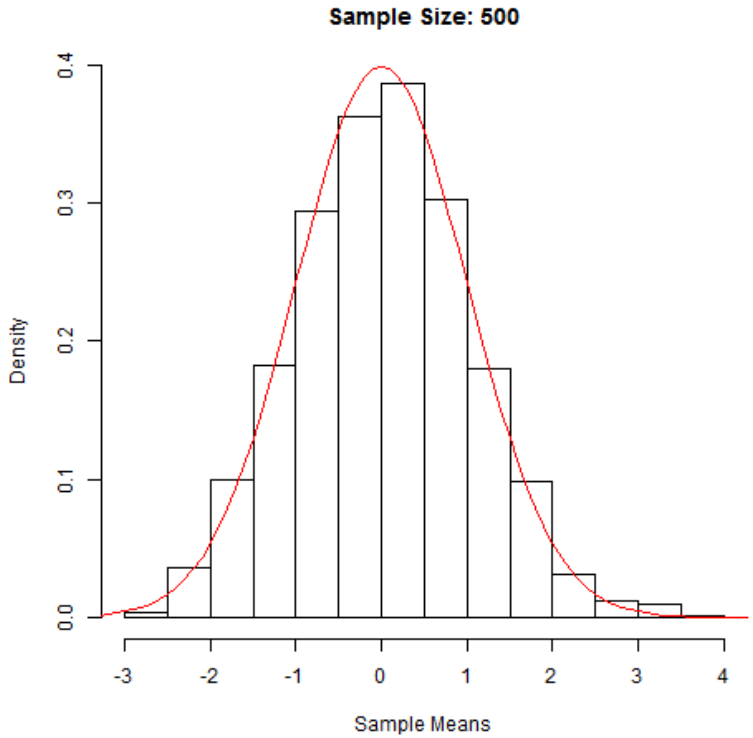
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$ and standardize the sample average. Then repeat with a new sample 1000 times.



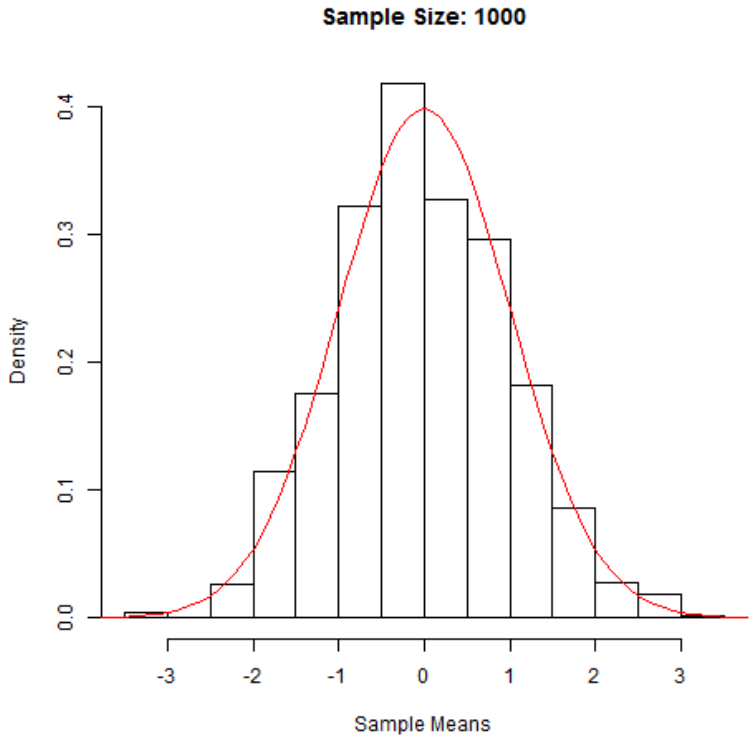
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$ and standardize the sample average. Then repeat with a new sample 1000 times.



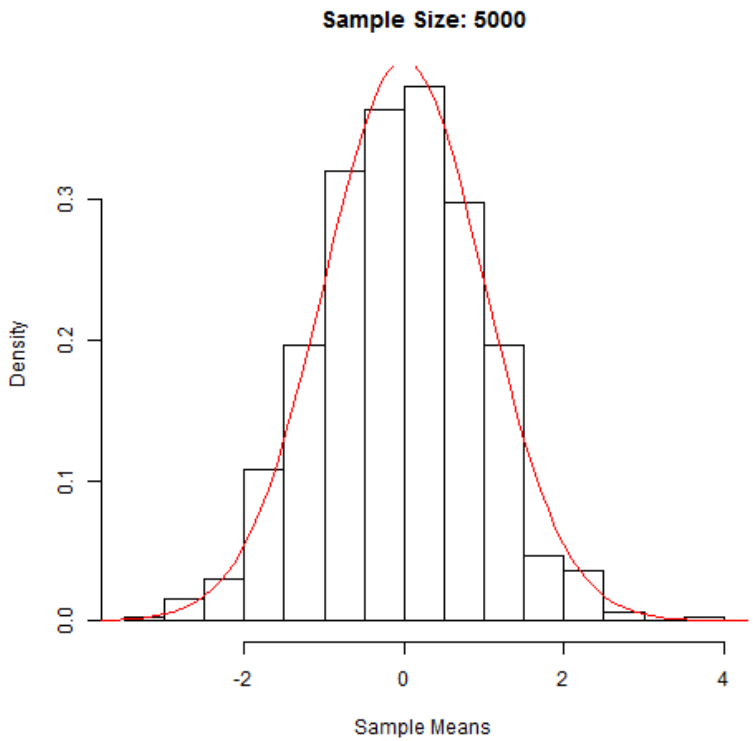
Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$ and standardize the sample average. Then repeat with a new sample 1000 times.



Central Limit Theorem Example

For a given sample size n, take the sample average of n exponential random variables with rate $= 1$ and standardize the sample average. Then repeat with a new sample 1000 times.



Normal Distributions in Nature

Why is height (approximately) normally distributed?

- How tall are you today?
- That is the sum of how tall you grew each day for the X number of days you have been alive so far.
- So the sum of those lengths is a sample of size X
- Each person is a separate sample

How else can we use the CLT?

We know a hypothetical distribution of a sample statistic $\bar{x}$

- In practice we don't do many "resamples"
- We can still estimate the hypothetical distribution if we did resample many times, if we know σ^2
- If we don't know σ^2, we can estimate it, but this changes things slightly

Sampling distributions

If the assumptions are satisfied,

$$\bar{x} \sim N(0, \sigma^2/n)$$

Sampling distributions

If the assumptions are satisfied,

$$\bar{x} \sim N(0, \sigma^2/n)$$

$$Var(\bar{x}) = Var(\frac{1}{n}\sum_i x_i) = \frac{1}{n^2} Var(\sum_i x_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$$