The Central limit theorem, Section 3.6

Grethe Hystad

October 2, 2012

The sample mean, Normal distribution

- Recall that if $X_1, X_2, ..., X_n$ are n independent, identically distributed **normal** random variables with mean μ and standard deviation σ, then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is $N(\mu, \frac{\sigma^2}{n})$.

The sample mean, general

- In general, suppose that $X_1, X_2, ..., X_n$ are n independent, identically distributed random variables from some distribution with mean μ and standard deviation σ.
- Define the sum of the random variables,

$$S_n = X_1 + X_2 + ... + X_n.$$

Define the standardize random variable,

- $$Z_n = \frac{\overline{X} - E(\overline{X})}{\sqrt{\text{Var}(\overline{X})}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

- We have $E(Z_n) = 0$ and $\text{Var}(Z_n) = 1$.

The sample mean, general

- By the law of larges number, $\overline{X} - \mu$ approaches zero as $n \to \infty$, while the factor $\frac{\sqrt{n}}{\sigma}$ in $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} = Z_n$ spreads out the probability enough to prevent Z_n from going to zero. Thus, we see fluctations about zero.
- In the limit as $n \to \infty$, the distribution of Z_n is approximately $N(0, 1)$.

The Central limit theorem

Theorem

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variable with mean μ and variance σ^2. Define

$$S_n = \sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then the distribution of

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

is $N(0,1)$ in the limit as $n \to \infty$.

The Central limit theorem

- The Central limit theorem says that

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

converges in distribution to a standard normal random variable, Z.

- That is

$$\lim_{n\to\infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt = \Phi(z).$$

- In practice, we have

$$P(a < Z_n \leq b) \approx P(a < Z \leq b) = \Phi(b) - \Phi(a).$$

R-simulation

Use R to simulate the sample mean, $\overline{x}$, of $30, 100$, and then 1000 uniformly distributed random variable, $U(0, 1)$, 1000 times. Create a histogram for these simulations. Determine the mean and variance for these simulations. How does these values compare to the distributional values?
The following code shows the simulation for a sample of $n = 30$ Uniformly distributed random variables.

```
> xbar=numeric(1000)
> for (i in 1:1000){x=runif(30);xbar[i]=mean(x)}
> hist(xbar,probability=TRUE, col="orange")
> mean(xbar)
[1] 0.5000515
> var(xbar)
[1] 0.002712586
> 
```

R-simulation

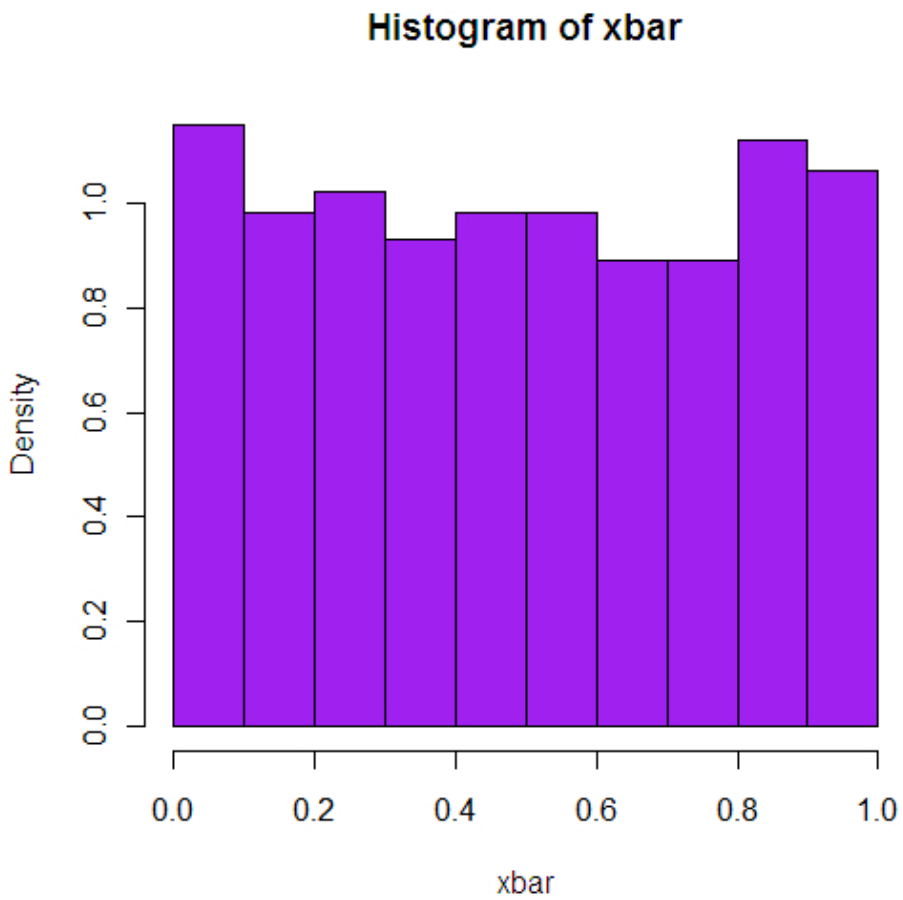

Figure: n=1

R-simulation

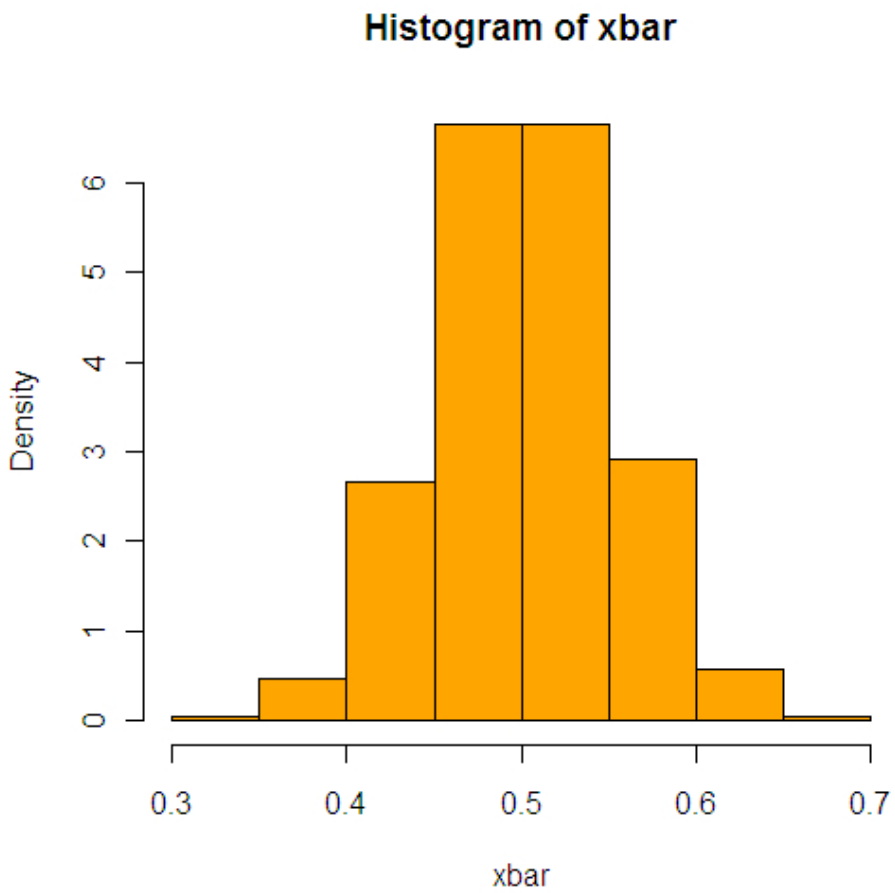

Figure: n=30

R-simulation

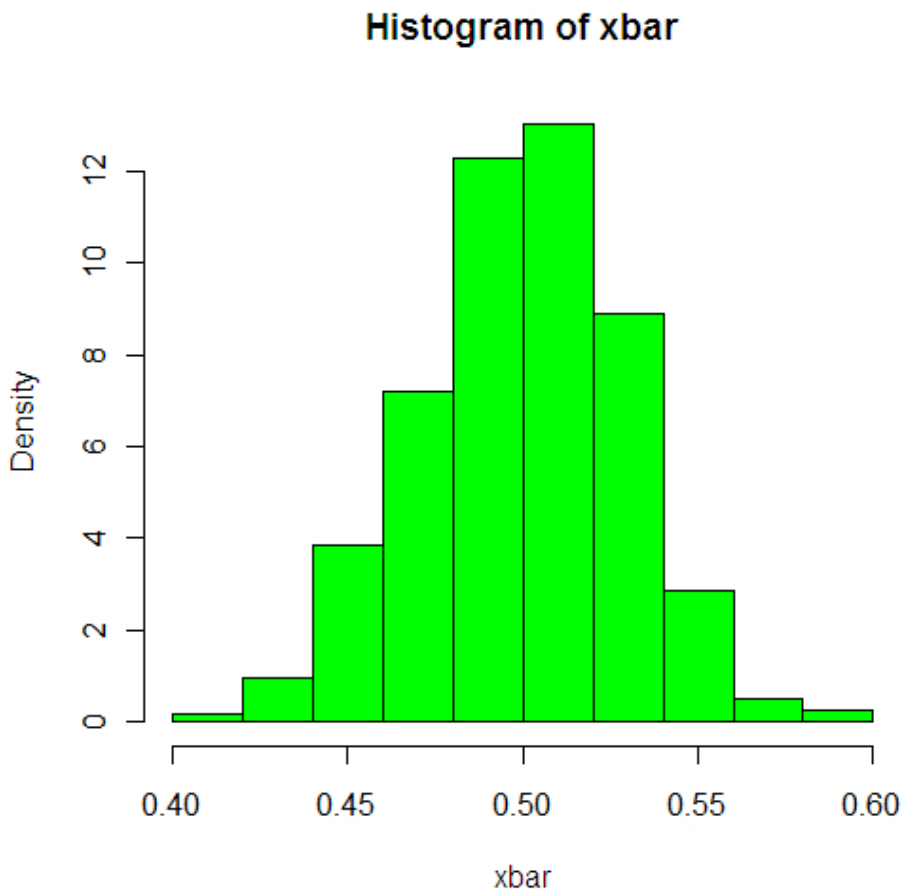

Figure: n=100, $mean(\overline{x}) = 0.4994481$, $var(\overline{x}) = 0.0008427071$

R-simulation

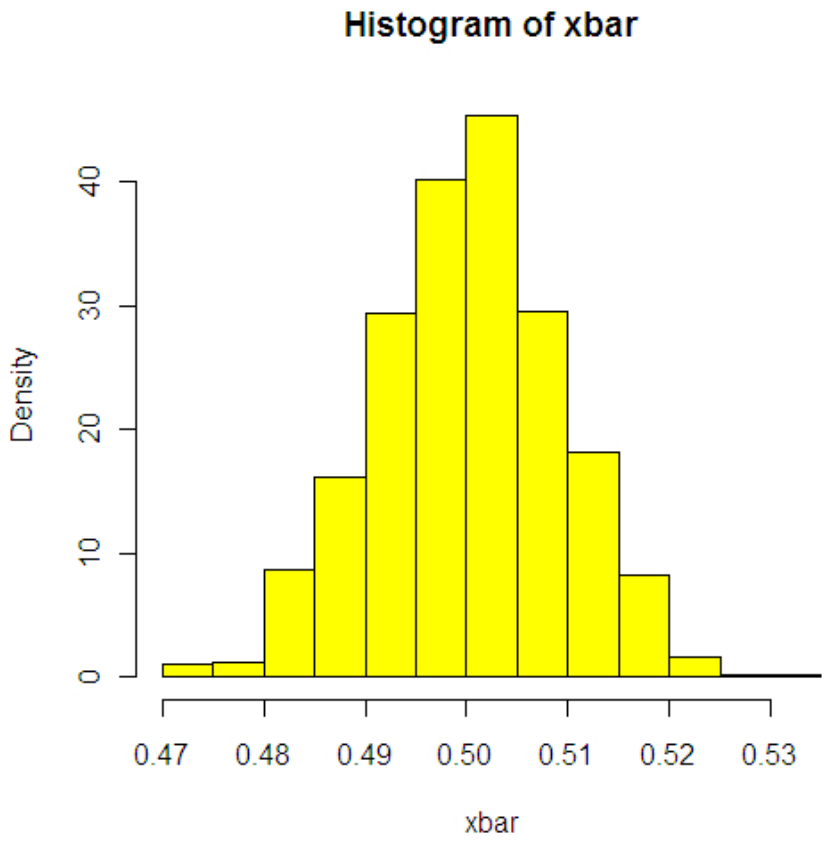

Figure: n=1000, $mean(\overline{x}) = 0.5001582$, $var(\overline{x}) = 8.306938e^{-05}$

R-simulation

- Recall that the mean of X in $U(0,1)$ is $E(X) = 0.5$ and the variance is $\text{Var}(X) = \frac{1}{12}$.
- The mean $\overline{X}$ of 100 random variable is $E(\overline{X}) = 0.5$ and the variance is $\text{Var}(\overline{X}) = \frac{1}{12*100} = 0.0008333333$
- These value are close to the values we obtained by simulation.

- Generally if n is greater than 25 or 30, the Normal approximation is a good approximation of the mean.
- If the underlying distribution is symmetric, unimodal, and of continuous type, the Normal approximation can be a good approximation of the mean for n as small as 4 or 5.
- If the underlying distribution is approximately normal, the mean is approximately normal when n is 2 or 3.

Example

Service times at a checkout counter in a retail store are exponentially distributed with a mean of 3 minutes. Suppose that 100 customers come to the checkout counter in a day. Find the approximate probability that the mean service time for those 100 customers are between 3 and 4 minutes.

Solution

- *Let* X_1, X_2,..., X_{100} *denote the service times to serve the* 100 *customers.*
- *We have* $E(\overline{X}) = 3$ *and* $\text{Var}(\overline{X}) = \frac{\text{Var}(X_i)}{n} = \frac{3^2}{100} = 0.09$.
- *Define*

$$Z_{100} = \frac{\overline{X} - 3}{\sqrt{0.09}} = \frac{\overline{X} - 3}{0.3}.$$

- *By the central limit theorem, the probability that* $\overline{X}$ *is between* 3.5 *and* 4 *minutes is:*

$$\begin{aligned} P(3.5 < \overline{X} < 4) &= P\left(\frac{3.5 - 3}{0.3} < \frac{\overline{X} - 3}{0.3} < \frac{4 - 3}{0.3}\right) \\ &= P(1.67 < Z_{100} < 3.33) \\ &\approx \phi(3.33) - \phi(1.67) = 0.047 \end{aligned}$$

Problem (A) is from Introduction to Probability and its application by R. Scheaffer:

Problem

The strength of a thread is a random variable with mean 0.5 *lb and standard deviation* 0.2 *lb. Assume that the strength of a rope is the sum of the strengths of the threads in the rope.*
(A) Find the probability that a rope consisting of 100 *threads will hold* 45 *lb.*
(B) How many threads are needed for a rope to provide 99% *assurance that the average thread will hold* 0.45 *lb.*

Solution given in class.