P4 - Central Limit Theorem

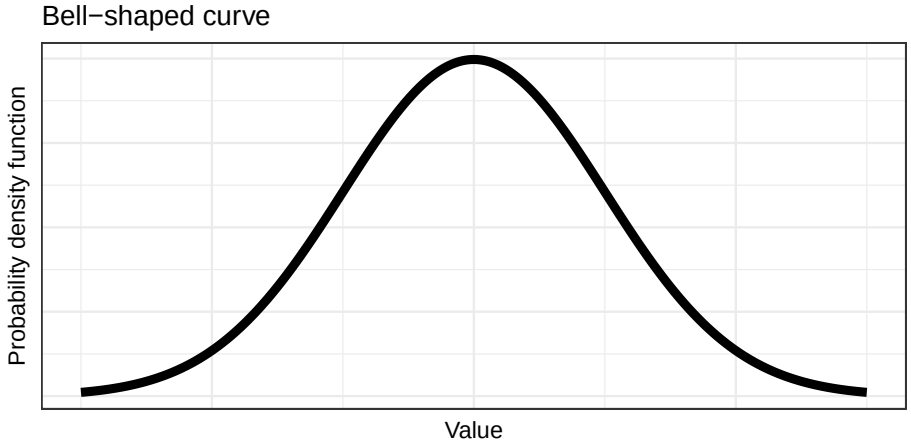
STAT 5870 (Engineering)
Iowa State University

November 22, 2024

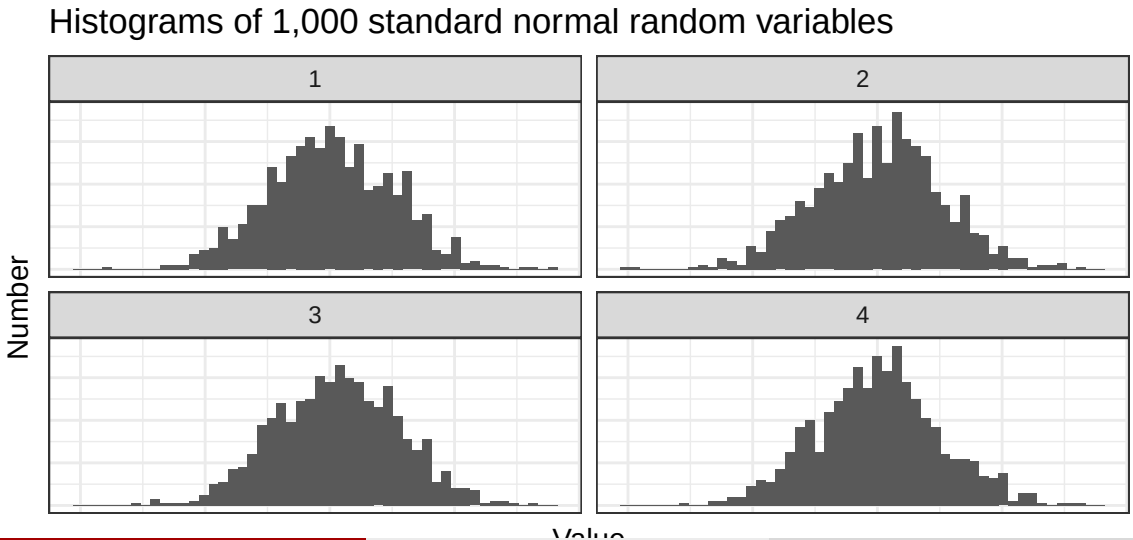
Main Idea: Sums and averages of iid random variables from any distribution have approximate normal distributions for sufficiently large sample sizes.

Bell-shaped curve

The term bell-shaped curve typically refers to the probability density function for a normal random variable:



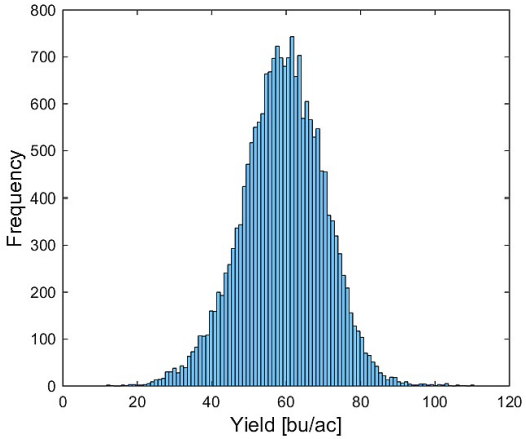
Histograms of samples from bell-shaped curves



Yield

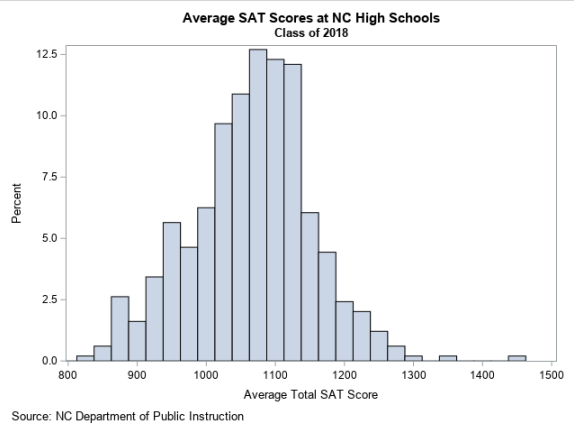
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184198

S1 Fig. Histogram of yield.

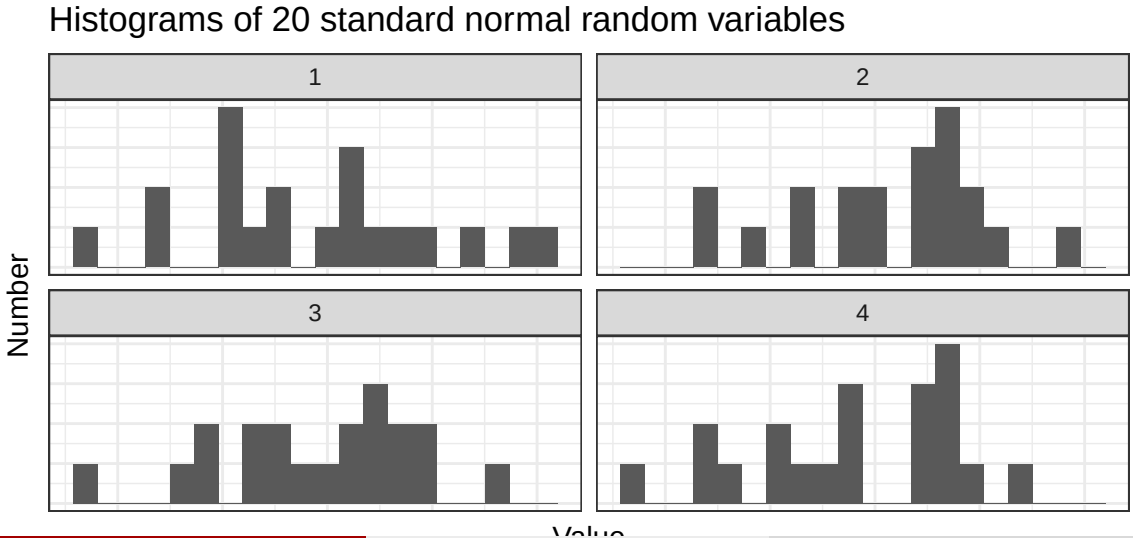


SAT scores

https://blogs.sas.com/content/iml/2019/03/04/visualize-sat-scores-nc.html

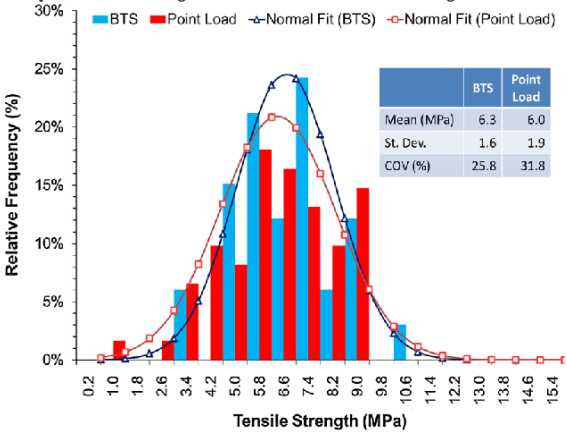


Histograms of samples from bell-shaped curves



Tensile strength

https://www.researchgate.net/figure/Comparison-of-histograms-for-BTS-and-tensile-strength-estimated-from-point-load_fig5_260617256



Sums and averages of iid random variables

Suppose $X_1, X_2, \ldots$ are iid random variables with

$$E[X_i] = \mu \quad Var[X_i] = \sigma^2.$$

Define

$$\begin{aligned} \text{Sample Sum: } S_n &= X_1 + X_2 + \cdots + X_n \\ \text{Sample Average: } \overline{X}_n &= S_n/n. \end{aligned}$$

Using properties of expectations and variances, we can find

- for S_n

$$E[S_n] = n\mu, \quad Var[S_n] = n\sigma^2, \quad \text{and} \quad SD[S_n] = \sqrt{n}\sigma$$

- for $\overline{X}_n$

$$E[\overline{X}_n] = \mu, \quad Var[\overline{X}_n] = \sigma^2/n, \quad \text{and} \quad SD[\overline{X}_n] = \sigma/\sqrt{n}.$$

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots$ are iid random variables with

$$E[X_i] = \mu \quad Var[X_i] = \sigma^2.$$

Define

$$\begin{aligned} \text{Sample Sum: } S_n &= X_1 + X_2 + \cdots + X_n \\ \text{Sample Average: } \overline{X}_n &= S_n/n. \end{aligned}$$

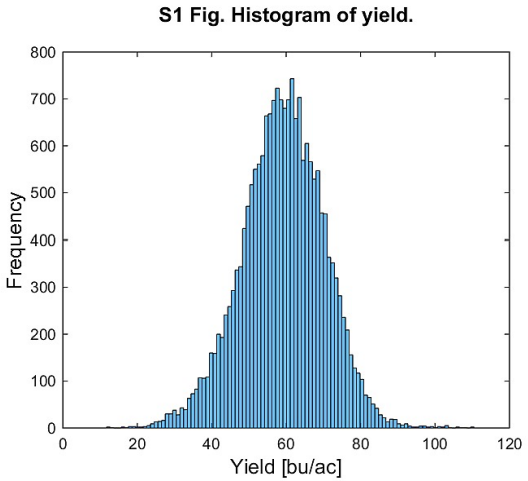
Then the Central Limit Theorem says

$$\lim_{n\to\infty} \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{and} \quad \lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1).$$

Main Idea: Sums and averages of iid random variables from any distribution have approximate normal distributions for sufficiently large sample sizes.

Yield

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184198



Approximating distributions

Rather than considering the limit, I typically think of the following approximations as n gets large.

For the sample average,

$$\overline{X}_n \dot{\sim} N(\mu, \sigma^2/n).$$

where $\dot{\sim}$ indicates *approximately distributed*. Recall

$$E\left[\overline{X}_n\right] = \mu \qquad \text{and} \qquad Var\left[\overline{X}_n\right] = \sigma^2/n.$$

For the sample sum,

$$S_n \dot{\sim} N(n\mu, n\sigma^2).$$

Recall

$$\begin{aligned} E[S_n] &= n\mu \\ Var[S_n] &= n\sigma^2. \end{aligned}$$

Averages and sums of uniforms

Let $X_i \overset{ind}{\sim} Unif(0,1)$. Then

$$\mu = E[X_i] = \frac{1}{2} \quad \text{and} \quad \sigma^2 = Var[X_i] = \frac{1}{12}.$$

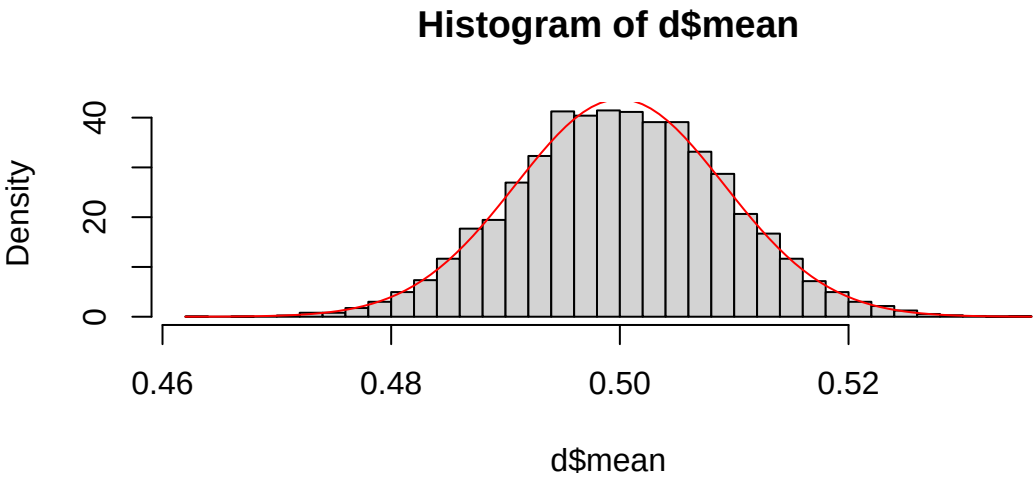
Thus

$$\overline{X}_n \dot{\sim} N\left(\frac{1}{2}, \frac{1}{12n}\right)$$

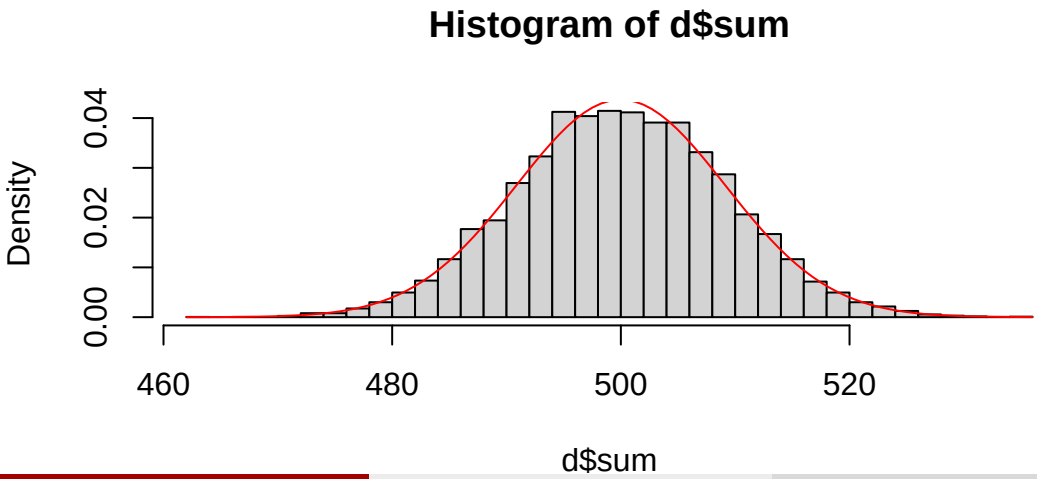
and

$$S_n \dot{\sim} N\left(\frac{n}{2}, \frac{n}{12}\right).$$

Averages of uniforms



Sums of uniforms



Normal approximation to a binomial

Recall if $Y_n = \sum_{i=1}^{n} X_i$ where $X_i \overset{ind}{\sim} Ber(p)$, then

$$Y_n \sim Bin(n, p).$$

For a binomial random variable, we have

$$E[Y_n] = np \qquad \text{and} \qquad Var[Y_n] = np(1-p).$$

By the CLT,

$$\lim_{n\to\infty} \frac{Y_n - np}{\sqrt{np(1-p)}} \to N(0,1).$$

If n is large,

$$Y_n \dot{\sim} N(np, np[1-p]).$$

Roulette example

A European roulette wheel has 39 slots: one green, 19 black, and 19 red. If I play black every time, what is the probability that I will have won more than I lost after 99 spins of the wheel?

https://isorepublic.com/photo/roulette-wheel/

Roulette example

A European roulette wheel has 39 slots: one green, 19 black, and 19 red. If I play black every time, what is the probability that I will have won more than I lost after 99 spins of the wheel?

Let Y indicate the total number of wins and assume $Y \sim Bin(n, p)$ with $n = 99$ and $p = 19/39$. The desired probability is $P(Y \geq 50)$. Then

$$P(Y \geq 50) = 1 - P(Y < 50) = 1 - P(Y \leq 49)$$

```
n <- 99
p <- 19/39
1 - pbinom(49, n, p)

[1] 0.399048
```

Roulette example

A European roulette wheel has 39 slots: one green, 19 black, and 19 red. If I play black every time, what is the probability that I will have won more than I lost after 99 spins of the wheel?

Let Y indicate the total number of wins. We can approximate Y using $X \sim N(np, np(1-p))$.

$$P(Y \geq 50) \approx 1 - P(X < 50)$$

```
1 - pnorm(50, n * p, sqrt( n * p * (1 - p) ) )
```

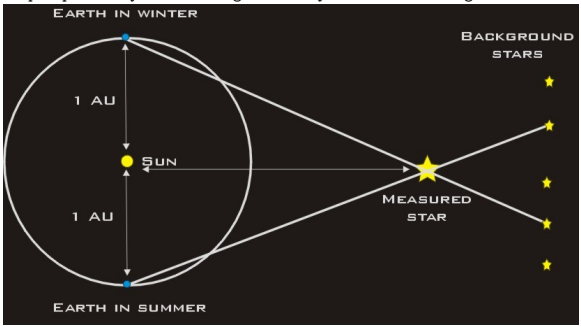
```
[1] 0.3610155
```

A better approximation can be found using a continuity correction.

Astronomy example

An astronomer wants to measure the distance, d, from Earth to a star. Suppose the procedure has a known standard deviation of 2 parsecs. The astronomer takes 30 iid measurements and finds the average of these measurements to be 29.4 parsecs. What is the probability the average is within 0.5 parsecs?

http://planetary-science.org/astronomy/distance-and-magnitudes/



Astronomy example

Let X_i be the i^{th} measurement. The astronomer assumes that $X_1, X_2, \ldots X_n$ are iid with $E[X_i] = d$ and $Var[X_i] = \sigma^2 = 2^2$. The estimate of d is

$$\overline{X}_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n} = 29.4.$$

and, by the Central Limit Theorem, $\overline{X}_n \dot{\sim} N(d, \sigma^2/n)$ where $n = 30$. We want to find

$$\begin{aligned} P\left(|\overline{X}_n - d| < 0.5\right) &= P\left(-0.5 < \overline{X}_n - d < 0.5\right) \\ &= P\left(\frac{-0.5}{2/\sqrt{30}} < \frac{\overline{X}_n - d}{\sigma/\sqrt{n}} < \frac{0.5}{2/\sqrt{30}}\right) \\ &\approx P(-1.37 < Z < 1.37) \end{aligned}$$

```
diff(pnorm( c(-1.37, 1.37) ))
```

```
[1] 0.8293131
```

Astronomy example - sample size

Suppose the astronomer wants to be within 0.5 parsecs with at least 95% probability. How many more samples would she need to take?

We solve

$$
\begin{aligned}
0.95 \leq P\left(\left|\overline{X}_n - d\right| < .5\right) &= P\left(-0.5 < \overline{X}_n - d < 0.5\right) \\
&= P\left(\tfrac{-0.5}{2/\sqrt{n}} < \tfrac{\overline{X}_n - d}{\sigma/\sqrt{n}} < \tfrac{0.5}{2/\sqrt{n}}\right) \\
&= P(-z < Z < z) \qquad\qquad z = 0.5/(2/\sqrt{n}) \\
&= 1 - [P(Z < -z) + P(Z > z)] \\
&= 1 - 2P(Z < -z)
\end{aligned}
$$

where $z = 1.96$ since

```
1 - 2 * pnorm(-1.96)
```

```
[1] 0.9500042
```

and thus $n = 61.47$ which we round up to $n = 62$ to ensure the probability is *at least* 0.95.

Summary

- Central Limit Theorem
  - Sums
  - Averages
- Examples
  - Uniforms
  - Binomial
    - Roulette
- Sample size
  - Astronomy