Why is the normal distribution so normal?

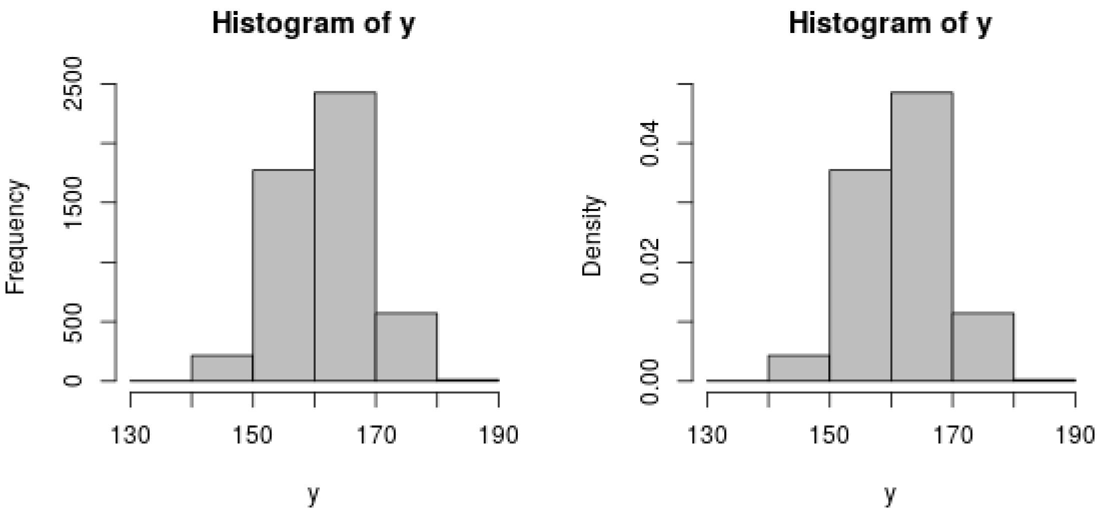
- and is it really so?

Søren Højsgaard

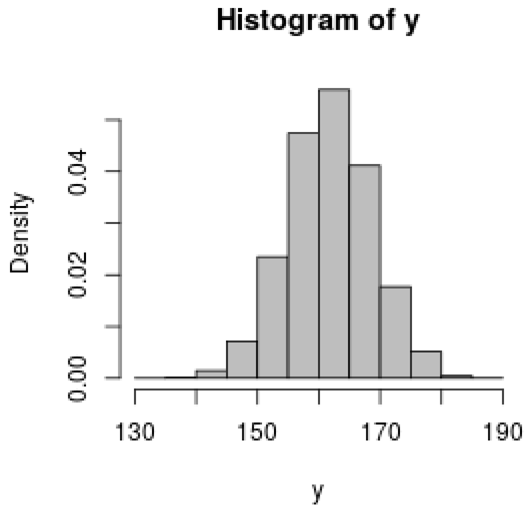
Department of Mathematical Sciences, Aalborg University

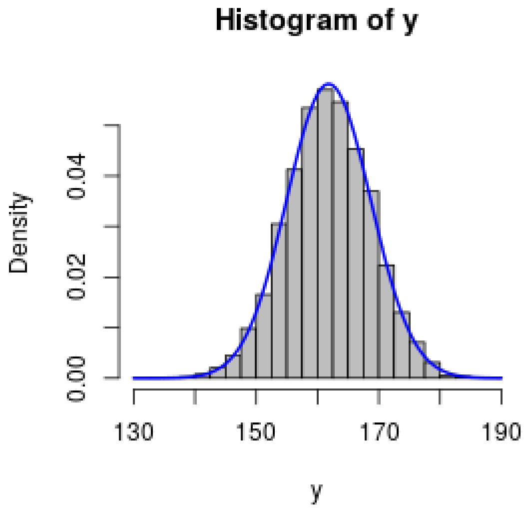
(updated:2019-04-19)

Height of women



If you divide into smaller groups, you can imagine that the histogram becomes a "smooth bell".





Described with the normal distribution density

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

But why is the normal distribution so normal?

So: Why can so many phenomena be described with a normal distribution?

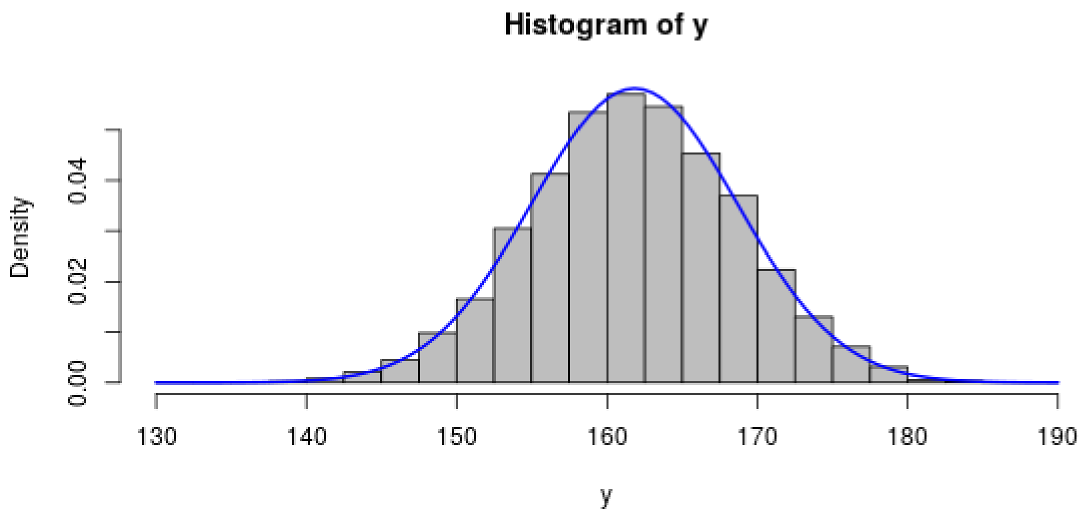
- The answer lies in the central limit value theory - "the central limit theorem" or CLT.
- Resolved: The sum of many independent contributions is approximate normally distributed; and the approximation gets better the more contributions there is.

What determines a person's height?

- Genetics; a pile of small contributions
- Food
- Living conditions
- …

A total of small (independent) contributions

Therefore, the height (with approximation) is normally distributed.



The central limit value setting

There are many different CLTs; the simplest is:

Let $X_1, X_2, \ldots, X_n$ be independent random variables, each with the same mean $ E(X_i) = \mu$ and same variance $V(Y_i) = \sigma^2$.

Let

$$S_n = \sum_i X_i, \quad Z_n = \frac{1}{n} S_n$$

Easy to prove that

- $E(S_n) = n\mu, V(S_n) = n\sigma^2$.
- $E(Z_n) = \mu, V(Z_n) = \sigma^2/n$.

CLT tells more: When $n \to \infty$ then the distribution of S_n and Z_n **is approximately normal** (written $S_n \sim_A N(,)$)

$$S_n \sim_A N(n\mu, n\sigma^2), \quad Z_n \sim_A N(\mu, \sigma^2/n)$$

Example:

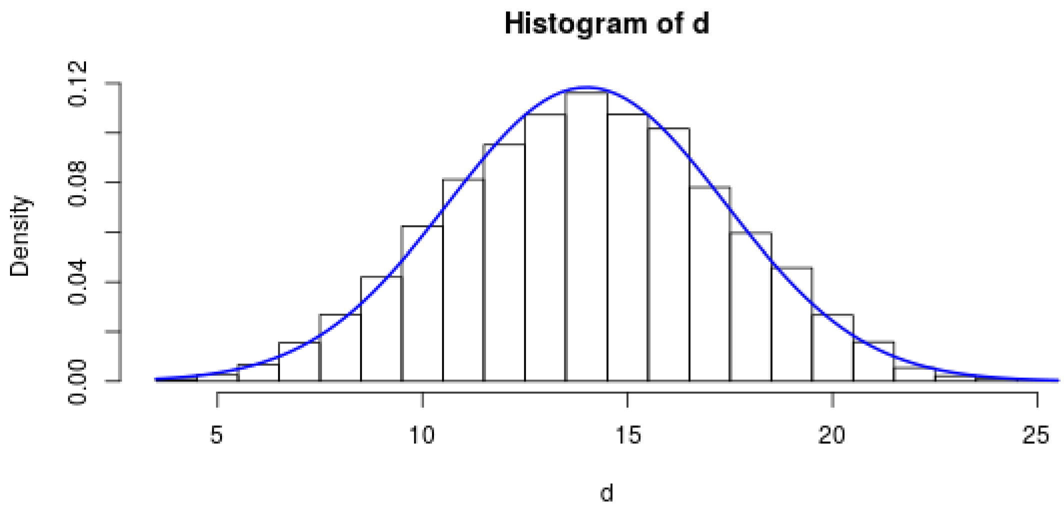
Experiment: Throw four dice and note the total number of eyes. Repeat experiment 10 times:

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    5    6    2    5    3    5    2    1    3     4
## [2,]    2    4    2    3    6    5    6    2    2     4
## [3,]    3    2    3    5    6    1    2    3    5     1
## [4,]    3    5    3    5    4    4    4    6    4     6
```

```
##  [1] 13 17 10 18 19 15 14 12 14 15
```

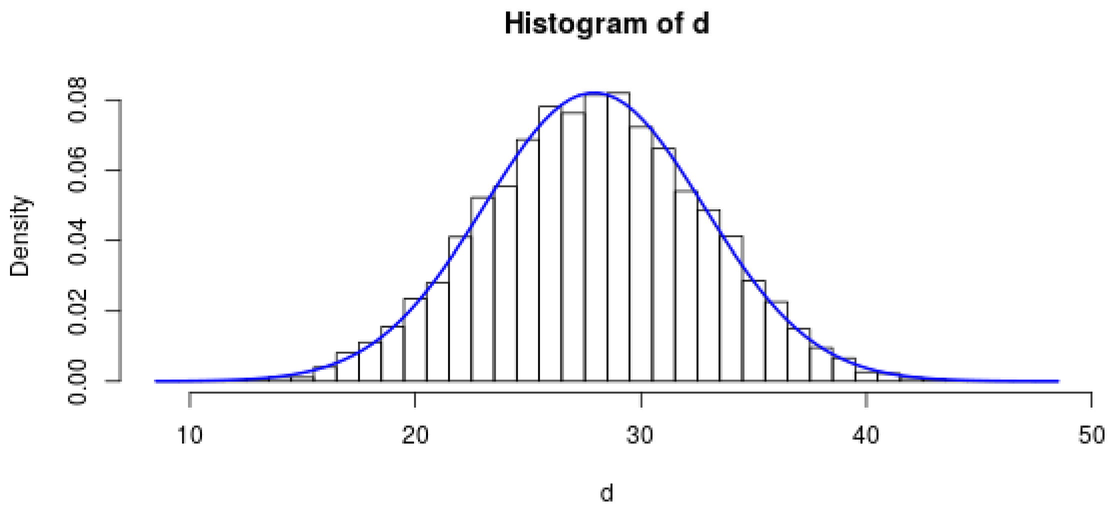
Instead, repeat the experiment 10,000 times. How do the number of eyes distribute?

```
##   [1] 12  8 12 12 13 18 19 11 19 14  9 17 20 14 19 12 11 10
##  [19] 16 13  9 16 15 14 16  6 11 22 21 11 12 17 18 17  8 15
##  [37] 18 10 16  6 18 20 14 12 13  7 18 15 17 10 17 14 14  8
##  [55]  7 15 12 17 14 15 19 15 13 16 10 13 20  9 19  9 12 10
##  [73] 13 10 14 11 17 15 11 12 12 14 13  9 12 15  8 10 16 17
##  [91] 12 18 12 21 18 12 10 17 17 15
```



Instead, let the experiment consist of throwing 8 dice and noting the total number of eyes. Repeat this experiment 10,000 times:

```
##   [1] 34 23 20 37 26 33 30 31 33 24 28 26 27 32 32 23 31 35
##  [19] 31 31 25 32 36 35 29 24 25 35 27 21 25 26 36 30 33 25
##  [37] 33 27 28 18 26 25 27 31 23 29 35 27 22 33 29 23 23 26
##  [55] 35 31 23 35 25 27 25 34 33 29 35 28 23 33 30 28 24 24
##  [73] 36 27 22 29 24 30 32 29 27 27 33 24 37 29 23 28 21 23
##  [91] 31 26 33 34 30 26 37 25 30 32
```



Simulate data from a $N(\mu, \sigma^2)$ distribution.

fact:

1. A standard normal distribution is the normal distribution with mean 0 and variance 1, written $N(0, 1)$.

2. All normal distributions are similar: Let $Y \sim N(\mu, \sigma^2)$ and $X = a + bY$. Then $X \sim (a + b\mu, b^2\sigma^2)$.

Quiz:

1. If U is standard normally distributed, $U \sim N(0, 1)$, then what is the distribution of $Y = \mu + \sigma U$?

2. If Y is $N(\mu, \sigma^2)$, then what is the distribution of $U = (Y - \mu)/\sigma$?

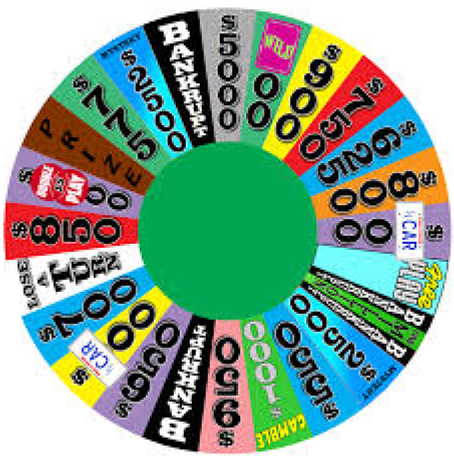
Hence, all normal distributions "look like" a $N(0,1)$ distribution.

Therefore: simulating from a $N(\mu, \sigma^2)$ distribution can be managed by simulating from $N(0,1)$ distribution.

Uniform distribution

Random variable Z has uniform distribution on the interval $[a, b]$, written $Z \sim unif(a, b)$ if all values in the interval are equally probable and values outside the interval cannot occur.

Z values can be simulated with a wheel of fortune :)



In Excel, Z can be simulated with RAND () (it is called SLUMP () in the Danish version of Excel).

fact

1. For a random variable $Z \sim unif(0,1), E(Z) = 1/2, V(Z) = 1/12.$

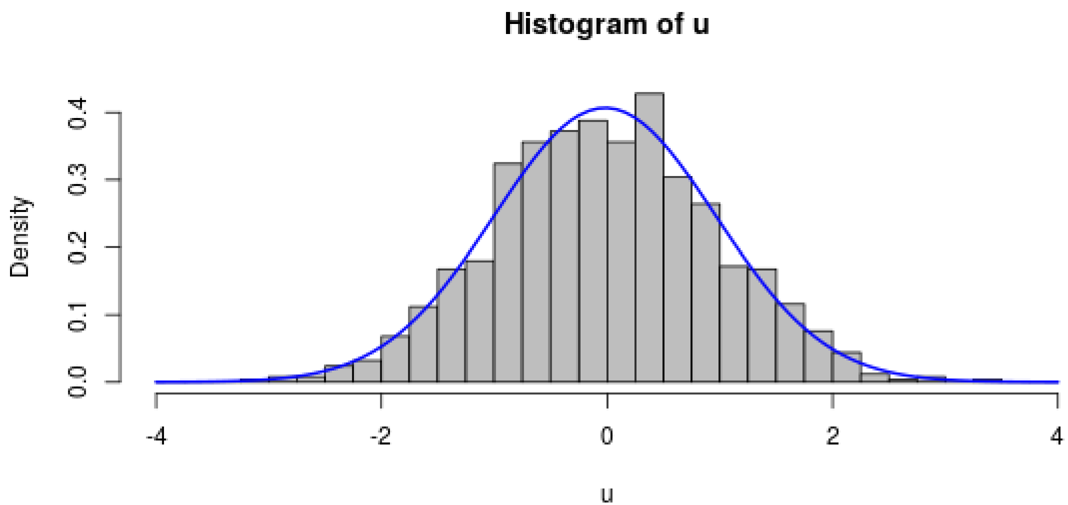
Quiz:

1. Prove the above.

Let $Z_1, \ldots Z_{12}$ be independent and $unif(0,1)$ - distributed and let
$U = \sum_{j=1}^{12} Z_j - 6$.

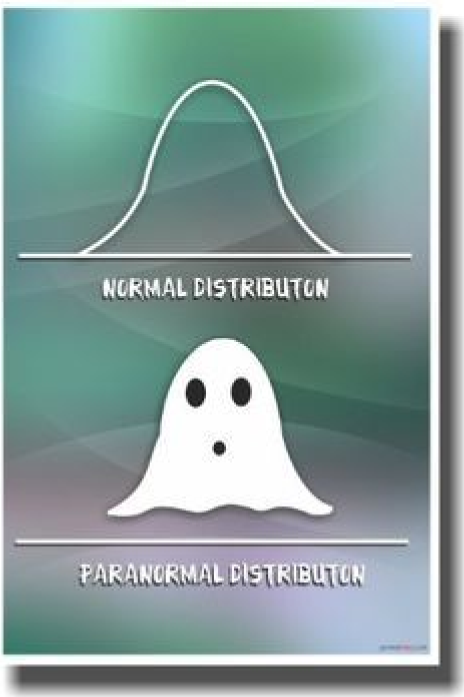
Then CLT gives $U \sim_A N(0,1)$

"Proof:" Simulate U many times and draw histogram; Then one should see the bell shape and get the right mean and variance



Mean: -0.0188. Variance: 0.9617

And maybe the normal distribution is not so normal anyway ...



Exercise 1 (computer)

Let $Z_1, \ldots, Z_n$ be independent or uniformly distributed $unif(0.1)$.

1. The sum of 2 independent contributions? Let $U = \sum_{j=1}^{2} Z_j$ Simulate many times and draw histogram. Does it look like a normal distribution?
2. The sum of 4 independent contributions? Let $U = \sum_{j=1}^{4} Z_j$ Simulate many times and draw histogram. Does it look like a normal distribution?

Exercise 2 (paper and pencil)

fact:

1. Let X be a binomial distributed random variable, $X \sim bin(N, \theta)$. Then $E(X) = N\theta$ and $V(X) = N\theta(1-\theta)$.
2. Let $X_1 \sim bin(N_1, \theta)$ and $X_2 \sim bin(N_2, \theta)$ and let X_1 and X_2 be independent. So is $Y = X_1 + X_2 \sim bin(N_1 + N_2, \theta)$

Quiz:

1. Argue that X approximate normal distribution for large N
2. What is the approximate distribution of $\frac{X}{N}$?
3. What is the approximate distribution of $\frac{\frac{X}{N}-\theta}{\sqrt{\theta(1-\theta)/N}}$?