The Gaussian distribution and the Central Limit Theorem

Peter Ralph

1 October 2020 – Advanced Biological Statistics

Stochastic minute: the Central Limit Theorem and the Normal distribution

The CLT

The Central Limit Theorem says, roughly, that net effect of the *sum* of a bunch of small, *independent* random things can be well-approximated by a Gaussian distribution, almost regardless of the details.

The CLT

The Central Limit Theorem says, roughly, that net effect of the *sum* of a bunch of small, *independent* random things can be well-approximated by a Gaussian distribution, almost regardless of the details.

For instance: say $X_1, X_2, \ldots, X_n$ are independent, random draws with mean μ and standard deviation σ.

Then, the difference between the "true" mean, μ, and the sample mean is Gaussian,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mathrm{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The Gaussian distribution

Also called the *Normal distribution*: see previous slide.

The Gaussian distribution

Also called the *Normal distribution*: see previous slide.

Saying that a random number Z "is Normal":

$$Z \sim \mathrm{Normal}(\mu, \sigma)$$

means that

$$\mathbb{P}\left\{Z \geq \frac{x-\mu}{\sigma}\right\} = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

The Gaussian distribution

Also called the *Normal distribution*: see previous slide.

Saying that a random number Z "is Normal":

$$Z \sim \mathrm{Normal}(\mu, \sigma)$$

means that

$$\mathbb{P}\left\{Z \geq \frac{x-\mu}{\sigma}\right\} = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

What to remember:

1. Z is probably no more than a few times σ away from μ
2. Using R,

```
rnorm(10, mean=3, sd=2)    # random simulations
pnorm(5, mean=3, sd=2)     # probabilities
qnorm(0.975, mean=3, sd=2) # quantiles
```


A demonstration

Let's check this, by doing:

> *find the sample mean of 100 random draws from some distribution*

lots of times, and looking at the distribution of those sample means.

A demonstration

Let's check this, by doing:

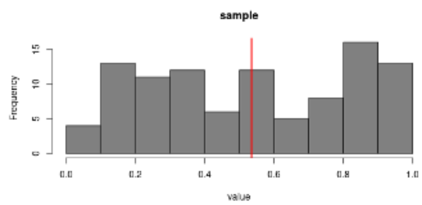
> *find the sample mean of 100 random draws from some distribution*

lots of times, and looking at the distribution of those sample means.

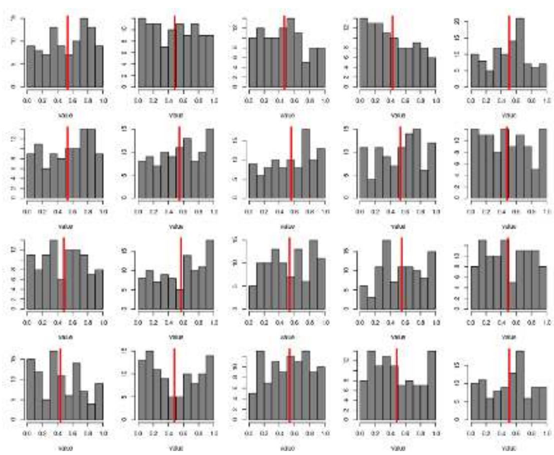
Claim: no matter the distribution we sample from, it should look close to Normal.

One sample

```r
n <- 100
x <- runif(n)
hist(x, xlab='value', main='sample', col=grey(0.5))
abline(v=mean(x), col='red', lwd=2)
```

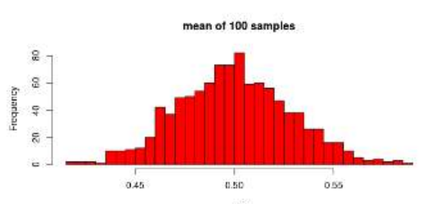


More samples



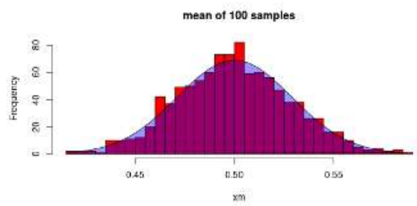
Distribution of 1,000 sample means

```r
xm <- replicate(1000, mean(runif(n)))
xh <- hist(xm, breaks=40, main=sprintf('mean of %d samples', n), col='red')
```



Distribution of 1,000 sample means

```r
plot(xh, main=sprintf('mean of %d samples', n), col='red')
xx <- xh$breaks
polygon(c(xx[-1] - diff(xx)/2, xx[1]),
        c(length(xm)* diff(pnorm(xx, mean=0.5, sd=1/sqrt(n*12))), 0),
        col=adjustcolor("blue", 0.4))
```



Relationship to the t distribution

If Y and $Z_1, \ldots, Z_n$ are independent $\text{Normal}(0, \sigma)$, and

$$X = \frac{Y}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} Z_j^2}}$$

then

$$X \sim \text{StudentsT}(n).$$

Relationship to the t distribution

If Y and $Z_1, \ldots, Z_n$ are independent $\text{Normal}(0, \sigma)$, and

$$X = \frac{Y}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} Z_j^2}}$$

then

$$X \sim \text{StudentsT}(n).$$

More usefully, *a sample mean divided by its standard error is** *t distributed.*

Relationship to the t distribution

If Y and $Z_1, \ldots, Z_n$ are independent $\text{Normal}(0, \sigma)$, and

$$X = \frac{Y}{\sqrt{\frac{1}{n}\sum_{j=1}^n Z_j^2}}$$

then

$$X \sim \text{StudentsT}(n).$$

More useful ly, *a sample mean divided by its standard error is* t distributed.*

This is thanks to the Central Limit Theorem. (* usually, approximately)

// reveal.js plugins

/