T-distribution

Gosset

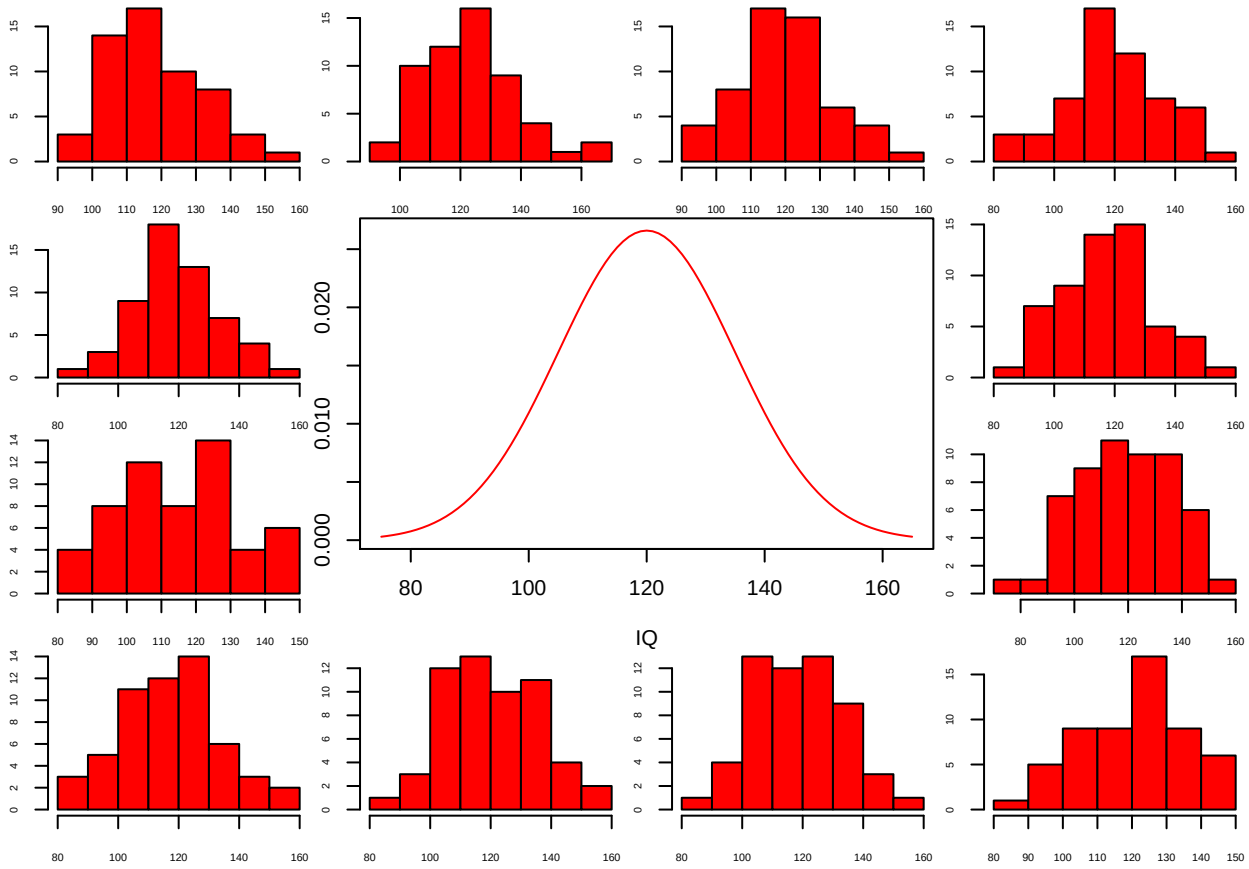
In probability and statistics, Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

In the English-language literature it takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples, for example the chemical properties of barley where sample sizes might be as low as 3.

Wikipedia

Population distribution

```
layout(matrix(c(2:6,1,1,7:8,1,1,9:13), 4, 4))

n  = 56    # Sample size
df = n - 1 # Degrees of freedom

mu    = 120
sigma = 15

IQ = seq(mu-45, mu+45, 1)

par(mar=c(4,2,0,0))
plot(IQ, dnorm(IQ, mean = mu, sd = sigma), type='l', col="red")

n.samples = 12

for(i in 1:n.samples) {

  par(mar=c(2,2,0,0))
  hist(rnorm(n, mu, sigma), main="", cex.axis=.5, col="red")

}
```



T-statistic

$$T_{n-1} = \frac{\bar{x} - \mu}{SE_x} = \frac{\bar{x} - \mu}{s_x/\sqrt{n}}$$

So the t-statistic represents the deviation of the sample mean $\bar{x}$ from the population mean μ, considering the sample size, expressed as the degrees of freedom $df = n - 1$

A samples

Let's take one sample from from our normal populatiion and calculate the t-value.

```
x = rnorm(n, mu, sigma); x
```

```
##  [1] 116.00911 123.42038  96.06892 125.05637 119.46173 114.07599 121.80648
##  [8] 142.37255 120.91049 120.99107 104.15695 116.33302 118.17362 134.16723
## [15] 131.63021 102.85196 114.88232 125.76760 133.67369  87.12215 122.86204
## [22] 124.46962 122.01964 118.75073 108.12744 131.85769 115.61038 108.04812
## [29] 122.60557 129.72557 124.60866 108.57841 136.72071 155.29043 113.69737
## [36] 111.06194 114.26217  97.11223  89.24808 106.94682 132.21406 137.47853
## [43] 103.72935 122.60753 129.34830 115.23041 152.74112 120.70792 138.77568
## [50] 132.55697 117.97826 118.16951 106.89011 129.49820 115.77496 108.48918
```

```r
mean(x)
```

```
## [1] 119.8701
```

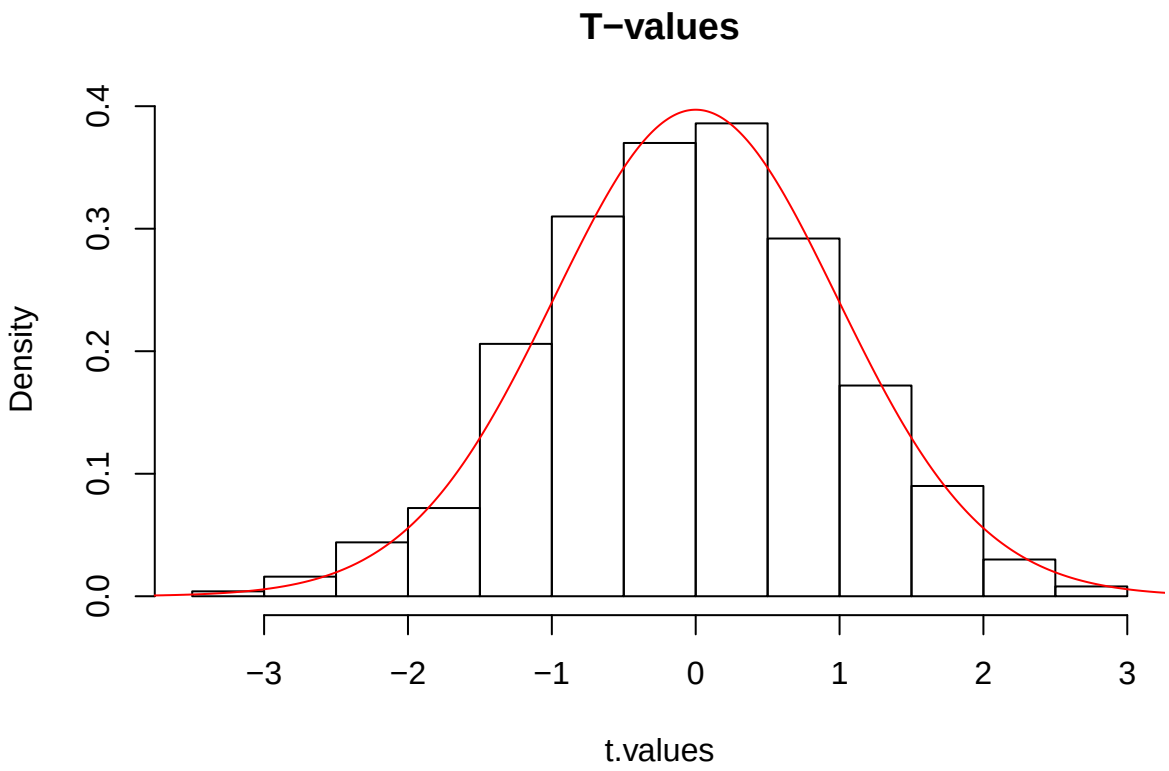
```r
t = (mean(x) - mu) / (sd(x) / sqrt(n)); t
```

```
## [1] -0.07111018
```

More samples

let's take more samples and calculate the t-value every time.

```r
n.samples = 1000

t.values = vector()

for(i in 1:n.samples) {

  x = rnorm(n, mu, sigma)

  t.values[i] = (mean(x) - mu) / (sd(x) / sqrt(n))

}

hist(t.values, freq = F, main="T-values")
T = seq(-4, 4, .01)
lines(T, dt(T,df), col = "red")
```



T-distribution

So if the population is normaly distributed (assumption of normality) the t-distribution represents the deviation of sample means from the population mean (μ), given a certain sample size ($df = n - 1$).

The t-distibution therefore is different for different sample sizes and converges to a standard normal distribution if sample size is large enough.

```
multiple.n  = c(5, 15, 30, 75, 100)
multiple.df = multiple.n - 1
col         = rainbow(length(multiple.df))

plot(T,  dt(T, multiple.df[1]), type = "l",
     xlim = c(-4,4), ylim = c(0,.45),
     xlab = "T", ylab="density",
     col = col[1], main="T-distributions" )

for(i in 2:length(multiple.df)) {

  lines(T, dt(T, multiple.df[i]), type="l", col=col[i])
  }
legend("topright", legend = paste("n =",multiple.n), lty=1, col = col)
```