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.

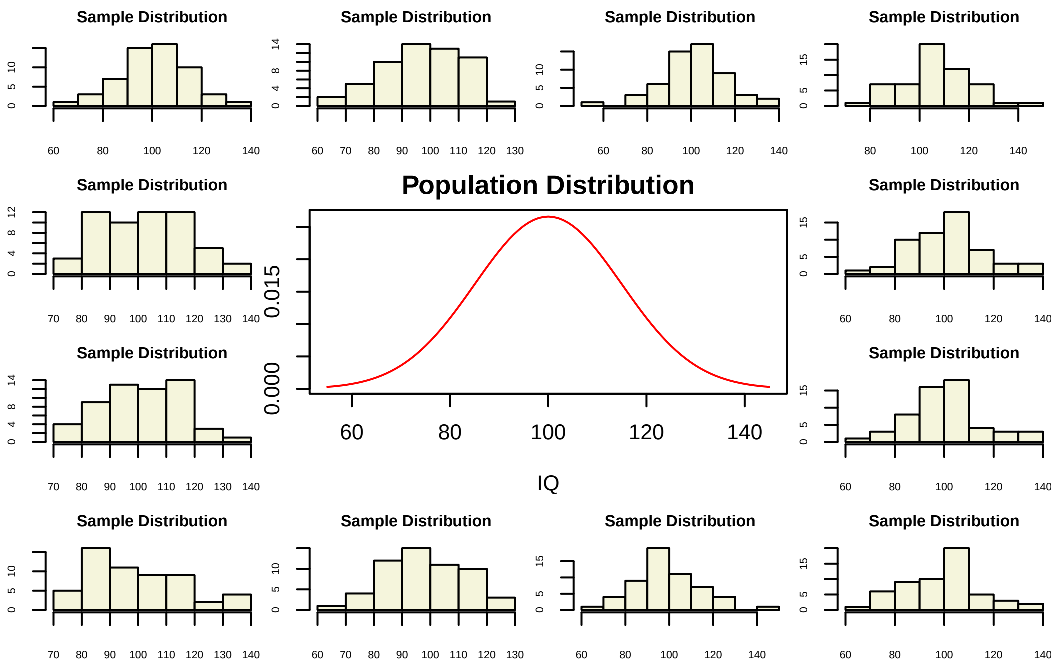
In the English-language literature it takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples, for example the chemical properties of barley where sample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

Source: 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    = 100
sigma = 15

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

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

n.samples = 12

for(i in 1:n.samples) {

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

}
```

Figure 1: William Sealy Gosset (aka Student) in 1908 (age 32)



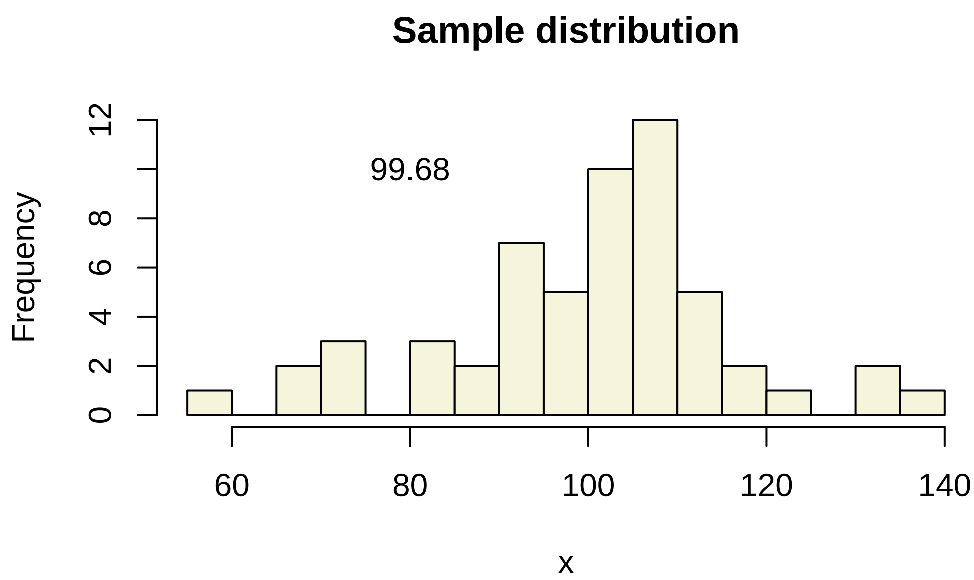
A sample

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

```
x <- rnorm(n, mu, sigma); x
```

```
 [1]  82.15232 101.63595 115.28672 109.87158  70.51442 105.32608  89.19363
 [8] 103.90399  92.40117 106.72208 106.52567 108.13867  66.11985  92.78027
[15]  97.88056 110.92711  72.73882 110.32494 102.18272  95.70328  99.89787
[22]  71.42069 112.92535 132.90416 106.35493 108.01691 104.90114  95.11672
[29]  97.48502 130.41333 106.06373 111.18564 101.57385  92.09770 102.34523
[36] 104.50496 100.92133  91.13382  57.70358 112.77222  92.65271  86.17285
[43] 107.64381 105.62078  84.70532 101.87855  91.65065 137.00274  82.47433
[50] 100.88022 115.84041 108.67502 120.72857 106.48097  93.89546  65.90284
```

```
hist(x, main = "Sample distribution", col = "beige", breaks = 15)
text(80, 10, round(mean(x),2))
```



More samples

let's take more samples.

```
n.samples     <- 1000
mean.x.values <- vector()
se.x.values   <- vector()

for(i in 1:n.samples) {
  x <- rnorm(n, mu, sigma)
  mean.x.values[i] <- mean(x)
  se.x.values[i]   <- (sd(x) / sqrt(n))
}
```

Mean and SE for all samples

```
head(cbind(mean.x.values, se.x.values))
```

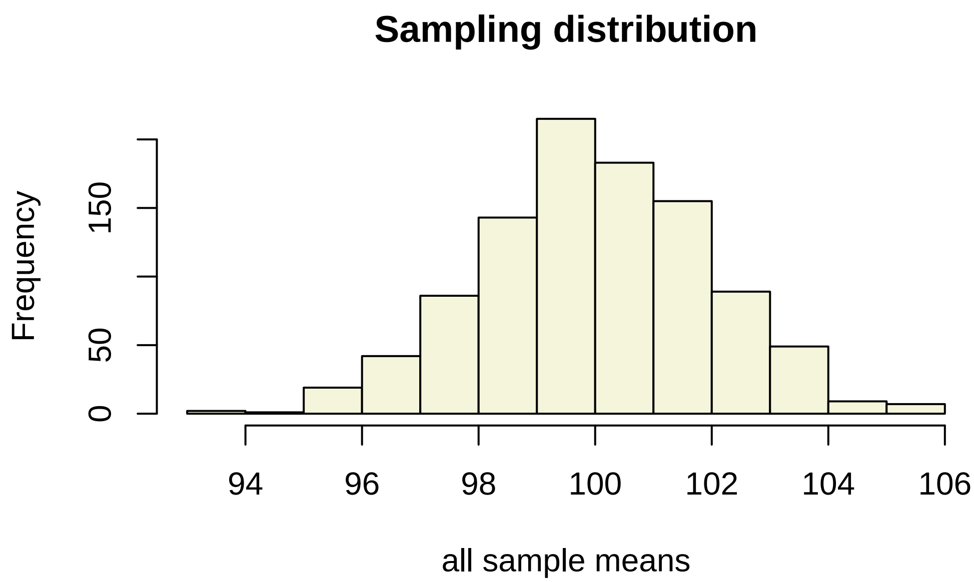
```
     mean.x.values se.x.values
[1,]      98.03299    1.742415
[2,]      98.63867    2.095539
[3,]     100.94955    1.908056
[4,]      99.43912    2.028078
```

```
[5,]       99.66210     2.283267
[6,]      102.42883     2.073420
```

Sampling distribution

Of the mean

```r
hist(mean.x.values,
     col  = "beige",
     main = "Sampling distribution",
     xlab = "all sample means")
```



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 $df = n - 1$

t-value

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

```
tStat <- (mean(x) - mu) / (sd(x) / sqrt(n))
tStat
```

```
[1] -0.7775873
```

Calculate t-values

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

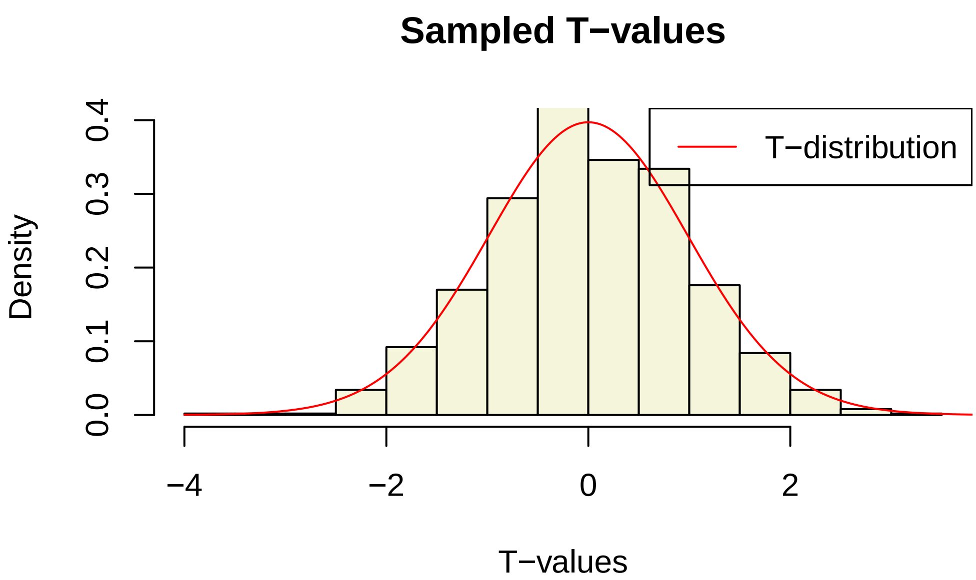
```
t.values <- (mean.x.values - mu) / se.x.values

tail(cbind(mean.x.values, mu, se.x.values, t.values))
```

```
       mean.x.values  mu se.x.values   t.values
 [995,]      96.34620 100    1.953362 -1.8705191
 [996,]      99.13851 100    2.085015 -0.4131808
 [997,]      99.37245 100    2.155226 -0.2911773
 [998,]      98.93876 100    2.050231 -0.5176203
 [999,]      98.28284 100    1.925408 -0.8918409
[1000,]      98.40228 100    2.054709 -0.7775873
```

Sampled t-values

What is the distribution of all these t-values?

```
hist(t.values,
     freq = FALSE,
     main = "Sampled T-values",
     xlab = "T-values",
     col  = "beige",
     ylim = c(0, .4))
myTs = seq(-4, 4, .01)
lines(myTs, dt(myTs,df), col = "red")
legend("topright", lty = 1, col="red", legend = "T-distribution")
```



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 ($df = n - 1$).

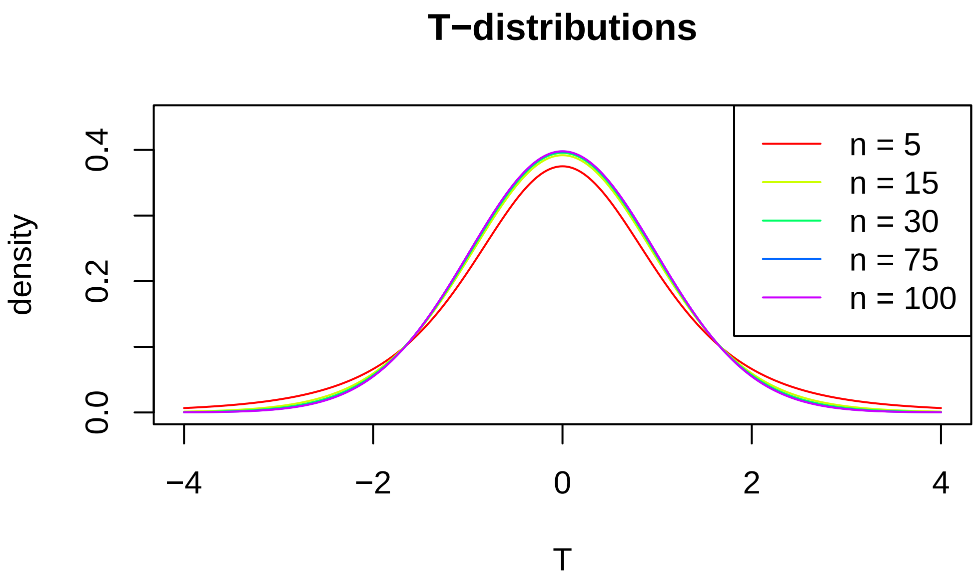
The t-distibution therefore is different for different sample sizes and converges to a standard normal distribution if sample size is large enough.

The t-distribution is defined by:

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where ν is the number of degrees of freedom and Γ is the gamma function.

Source: wikipedia



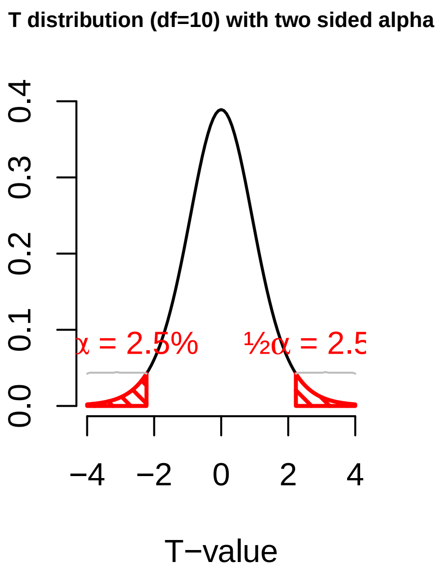
One or two sided

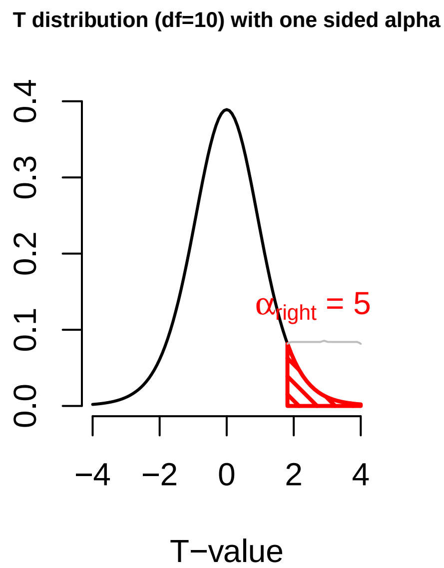
Two sided

- $H_A : \bar{x} \neq \mu$

One sided

- $H_A : \bar{x} > \mu$
- $H_A : \bar{x} < \mu$





Effect-size

The effect-size is the standardized difference between the mean and the expected μ. In the t-test effect-size is expressed as r.

$$r = \sqrt{\frac{t^2}{t^2 + \text{df}}}$$

Tukey (1969): >being so disinterested in our variables that we do not care about their units can hardly be desirable.

```
r <- sqrt(tStat^2 / (tStat^2 + df))

r
```

```
[1] 0.2603778
```

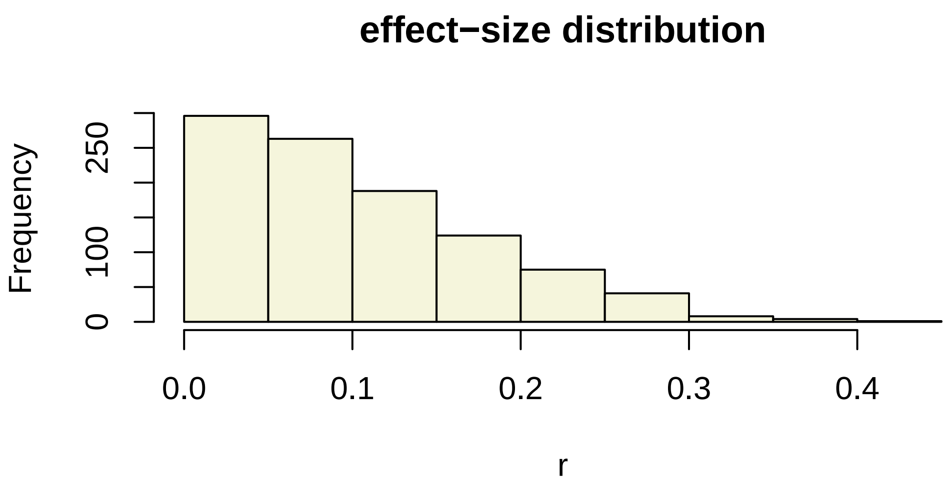
Effect-sizes

We can also calculate effect-sizes for all our calculated t-values. Under the assumption of H_0 the effect-size distribution looks like this.

```r
r <- sqrt(t.values^2/(t.values^2 + df))

tail(cbind(mean.x.values, mu, se.x.values, t.values, r))
```

```
       mean.x.values  mu se.x.values   t.values          r
 [995,]     96.34620 100    1.953362 -1.8705191 0.24456174
 [996,]     99.13851 100    2.085015 -0.4131808 0.05562702
 [997,]     99.37245 100    2.155226 -0.2911773 0.03923211
 [998,]     98.93876 100    2.050231 -0.5176203 0.06962652
 [999,]     98.28284 100    1.925408 -0.8918409 0.11939559
[1000,]     98.40228 100    2.054709 -0.7775873 0.10427823
```

Effect-size distribution



Cohen (1988)

- Small: $0 \leq .1$
- Medium: $.1 \leq .3$
- Large: $.3 \leq .5$

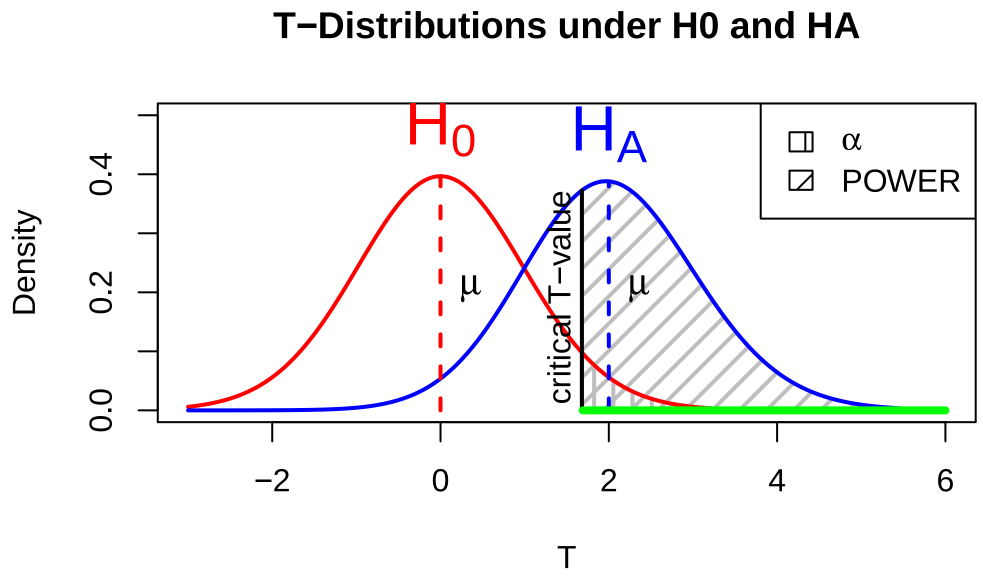
Power

- Strive for 80%
- Based on know effect size
- Calculate number of subjects needed
- Use G*Power, JASP, or SPSS to calculate

Alpha Power

```
tValues <- seq(-3,6,.01)
N <- 45
E <- 2

# Set plot
plot(0,0,
     type = "n",
     ylab = "Density",
     xlab = "T",
     ylim = c(0,.5),
     xlim = c(-3,6),
     main = "T-Distributions under H0 and HA")

critical_t = qt(.05,N-1,lower.tail=FALSE)

# Alpha
range_x = seq(critical_t,6,.01)
polygon(c(range_x,rev(range_x)),
        c(range_x*0,rev(dt(range_x,N-1,ncp=0))),
        col      = "grey",
        density = 10,
        angle    = 90,
        lwd      = 2)
```

```
# Power
range_x = seq(critical_t,6,.01)
polygon(c(range_x,rev(range_x)),
        c(range_x*0,rev(dt(range_x,N-1,ncp=E))),
        col     = "grey",
        density = 10,
        angle   = 45,
        lwd     = 2)

lines(tValues,dt(tValues,N-1,ncp=0),col="red", lwd=2) # H0 line
lines(tValues,dt(tValues,N-1,ncp=E),col="blue",lwd=2) # HA line

# Critical value
lines(rep(critical_t,2),c(0,dt(critical_t,N-1,ncp=E)),lwd=2,col="black")
text(critical_t,dt(critical_t,N-1,ncp=E),"critical T-value",pos=2, srt = 90)

# H0 and HA
text(0,dt(0,N-1,ncp=0),expression(H[0]),pos=3,col="red", cex=2)
text(E,dt(E,N-1,ncp=E),expression(H[A]),pos=3,col="blue",cex=2)

# Mu H0 line
lines(c(0,0),c(0,dt(0,N-1)), col="red",  lwd=2,lty=2)
text(0,dt(0,N-1,ncp=0)/2,expression(mu),pos=4,cex=1.2)
# Mu HA line
lines(c(E,E),c(0,dt(E,N-1,ncp=E)),col="blue",lwd=2,lty=2)
text(E,dt(0,N-1,ncp=0)/2,expression(paste(mu)),pos=4,cex=1.2)

# t-value
lines( c(critical_t+.01,6),c(0,0),col="green",lwd=4)

# Legend
legend("topright", c(expression(alpha),'POWER'),density=c(10,10),angle=c(90,45))
```



R-Psychologist