The t distribution

Peter Ralph

1 October 2020 – Advanced Biological Statistics

Stochastic minute: the t distribution

The t statistic

The t statistic computed from a collection of n numbers is the sample mean divided by the estimated standard error of the mean, which is the sample SD divided by $\sqrt{n}$.

The t statistic

The t statistic computed from a collection of n numbers is the sample mean divided by the estimated standard error of the mean, which is the sample SD divided by $\sqrt{n}$.

If $x_1, \ldots, x_n$ are numbers, then

$$\begin{aligned} \text{(sample mean)} \qquad \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \text{(sample SD)} \qquad s &= \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

so

$$t(x) = \frac{\bar{x}}{s/\sqrt{n}}.$$

Consistency check

```
n <- 20
x <- rnorm(n)
c(t.test(x)$statistic,
  mean(x) / (sd(x) / sqrt(n)))
```

```
##        t
## 1.318919 1.318919
```


The t approximation

Fact: If $X_1, \ldots, X_n$ are independent random samples from a distribution with mean μ, then

$$t(X - \mu) = \frac{\bar{x} - \mu}{s/\sqrt{n}} \approx \text{StudentsT}(n-2),$$

as long as n is not too small and the distribution isn't too wierd.

A demonstration

Let's check this, by doing:

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

lots of times, and looking at the distribution of those t scores.

A demonstration

Let's check this, by doing:

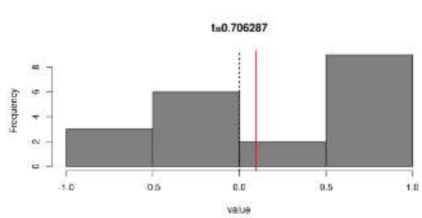
> *find the sample t score of 100 random draws from some distribution*

lots of times, and looking at the distribution of those t scores.

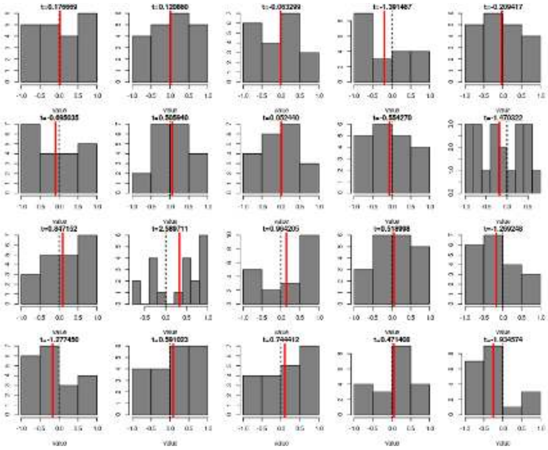
Claim: no matter* the distribution we sample from, the *sampling distribution* of the t statistics should look close to the t distribution.

One sample

```
n <- 20
x <- 2 * runif(n) - 1
hist(x, xlab='value', col=grey(0.5),
     main=sprintf("t=%f", mean(x)*sqrt(n)/sd(x)))
abline(v=0, lwd=2, lty=3)
abline(v=mean(x), col='red', lwd=2)
```

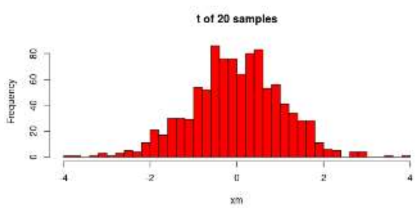


More samples



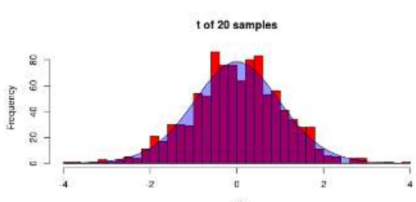
Distribution of 1,000 sample t scores

```
xm <- replicate(1000, {
          x <- 2 * runif(n) - 1;
          mean(x) * sqrt(n) / sd(x) })
xh <- hist(xm, breaks=40, main=sprintf('t of %d samples', n), col='red')
```



Distribution of 1,000 sample t scores

```
plot(xh, main=sprintf('t of %d samples', n), col='red')
xx <- xh$breaks
polygon(c(xx[-1] - diff(xx)/2, xx[1]),
        c(length(xm)* diff(pt(xx, df=(n-1))), 0),
        col=adjustcolor("blue", 0.4))
```



Exercise:

Do this again (use my code) except using

```
x <- rexp(n) - 1
```

instead of `2 * runif(n) - 1`.

// reveal.js plugins