Introduction to statistics: The sampling distribution, t-test

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

June 16, 2019

The sampling distribution of the mean

When we have a **single sample**, we know how to compute MLEs of the sample mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$.
Suppose now that you had many repeated samples; from each sample, you can compute the mean each time. We can simulate this situation:

```
x<-rnorm(100,mean=500,sd=50)
mean(x)

## [1] 503.62

x<-rnorm(100,mean=500,sd=50)
mean(x)

## [1] 495.19
```

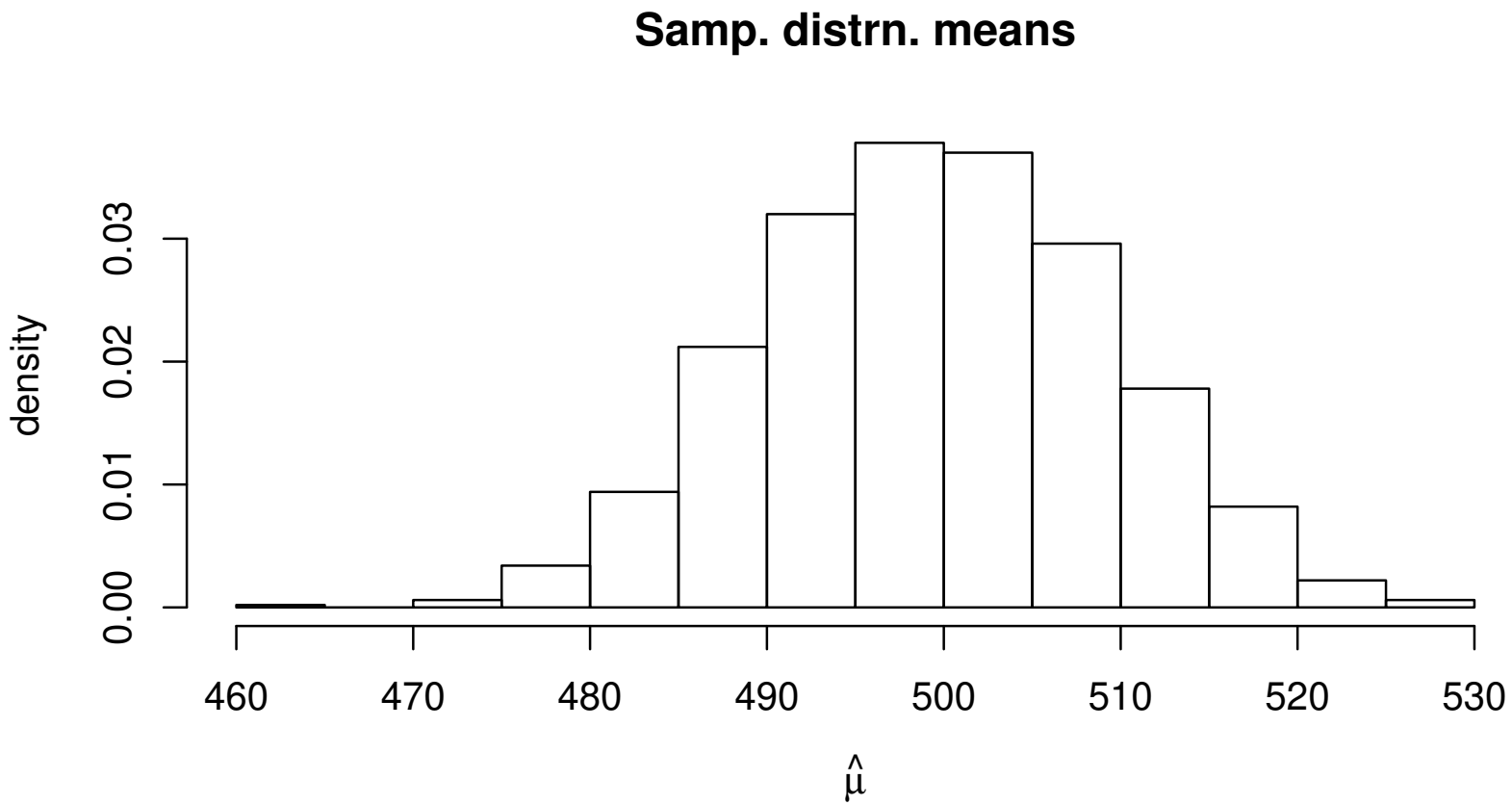
The sampling distribution of the mean

Let's repeatedly simulate sampling 1000 times:

```
nsim<-1000
n<-100
mu<-500
sigma<-100
samp_distrn_means<-rep(NA,nsim)
samp_distrn_sd<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n,mean=mu,sd=sigma)
  samp_distrn_means[i]<-mean(x)
  samp_distrn_sd[i]<-sd(x)
}
```

The sampling distribution of the mean

Plot the distribution of the means under repeated sampling:

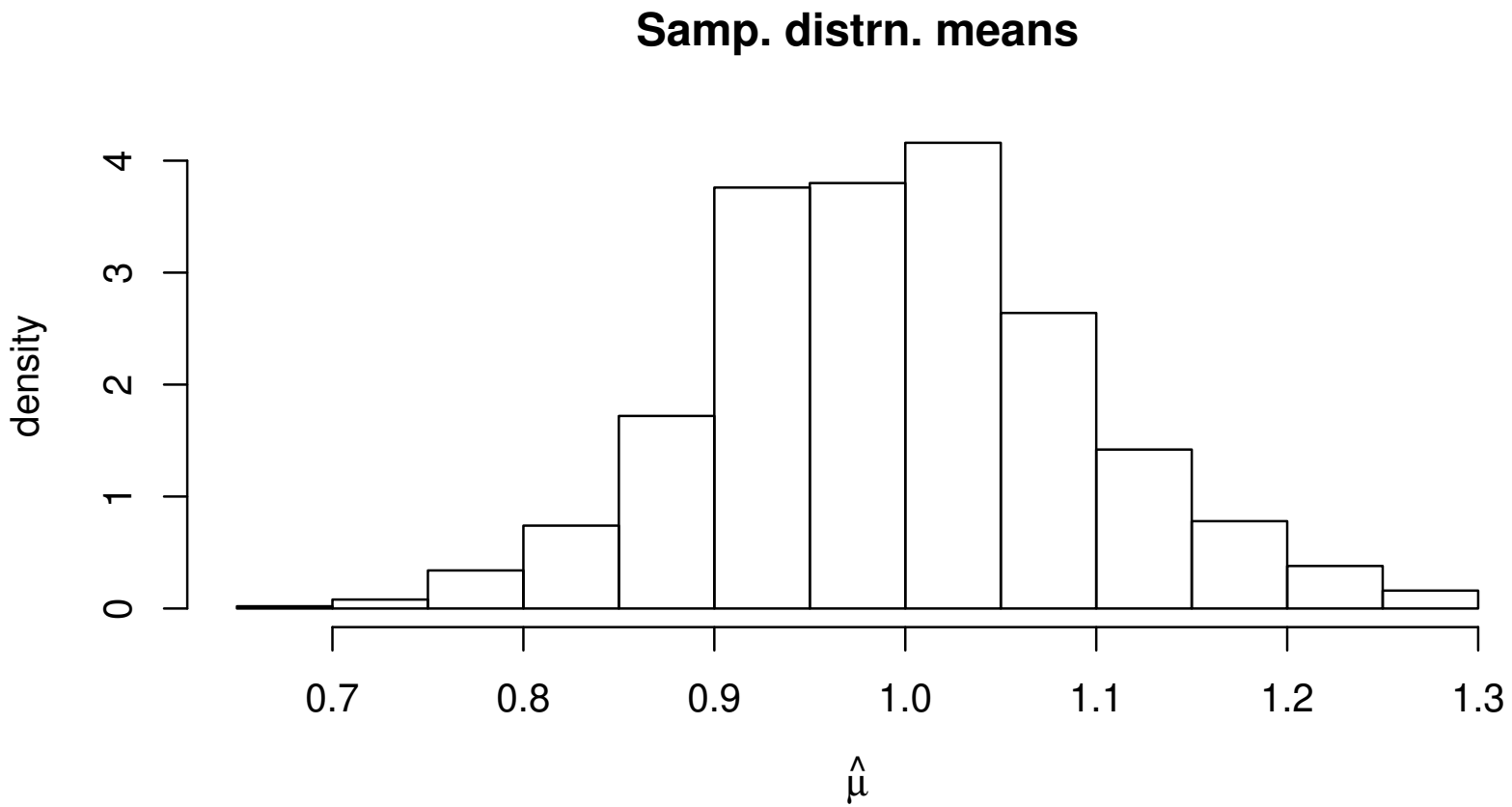


The sampling distribution of the mean

Interestingly, it is not necessary that the distribution that we are sampling from be the normal distribution.

```
for(i in 1:nsim){
  x<-rexp(n)
  samp_distrn_means[i]<-mean(x)
  samp_distrn_sd[i]<-sd(x)
}
```

The sampling distribution of the mean



The central limit theorem

1. For large enough sample sizes, the sampling distribution of the means will be approximately normal, regardless of the underlying distribution (as long as this distribution has a mean and variance defined for it).
2. This will be the basis for statistical inference.

The sampling distribution of the mean

We can compute the standard deviation of the sampling distribution of means:

```
## estimate from simulation:
sd(samp_distrn_means)

## [1] 0.097254
```

The sampling distribution of the mean

A further interesting fact is that we can compute this standard deviation of the sampling distribution **from a single sample** of size n:

$$\frac{\hat{\sigma}}{\sqrt{n}}$$

```
x<-rnorm(100,mean=500,sd=100)
hat_sigma<-sd(x)
hat_sigma/sqrt(n)

## [1] 9.8761
```

The sampling distribution of the mean

1. So, from a sample of size n, and sd σ or an MLE $\hat{\sigma}$, we can compute
   the standard deviation of the sampling distribution of the means.
2. We will call this standard deviation the estimated **standard error**.
   $SE = \frac{\hat{\sigma}}{\sqrt{n}}$
   I say *estimated* because we are estimating SE using an an estimate of σ.

Confidence intervals

The standard error allows us to define a so-called **95% confidence interval**:

$$\hat{\mu} \pm 2SE \tag{1}$$

So, for the mean, we define a 95% confidence interval as follows:

$$\hat{\mu} \pm 2\frac{\hat{\sigma}}{\sqrt{n}} \tag{2}$$

Confidence intervals

In our example:

```
## lower bound:
mu-(2*hat_sigma/sqrt(n))

## [1] 480.25

## upper bound:
mu+(2*hat_sigma/sqrt(n))

## [1] 519.75
```

The meaning of the 95% CI

If you take repeated samples and compute the CI each time, 95% of those CIs will contain the true population mean.

```
lower<-rep(NA,nsim)
upper<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n,mean=mu,sd=sigma)
  lower[i]<-mean(x) - 2 * sd(x)/sqrt(n)
  upper[i]<-mean(x) + 2 * sd(x)/sqrt(n)
}
```

The meaning of the 95% CI

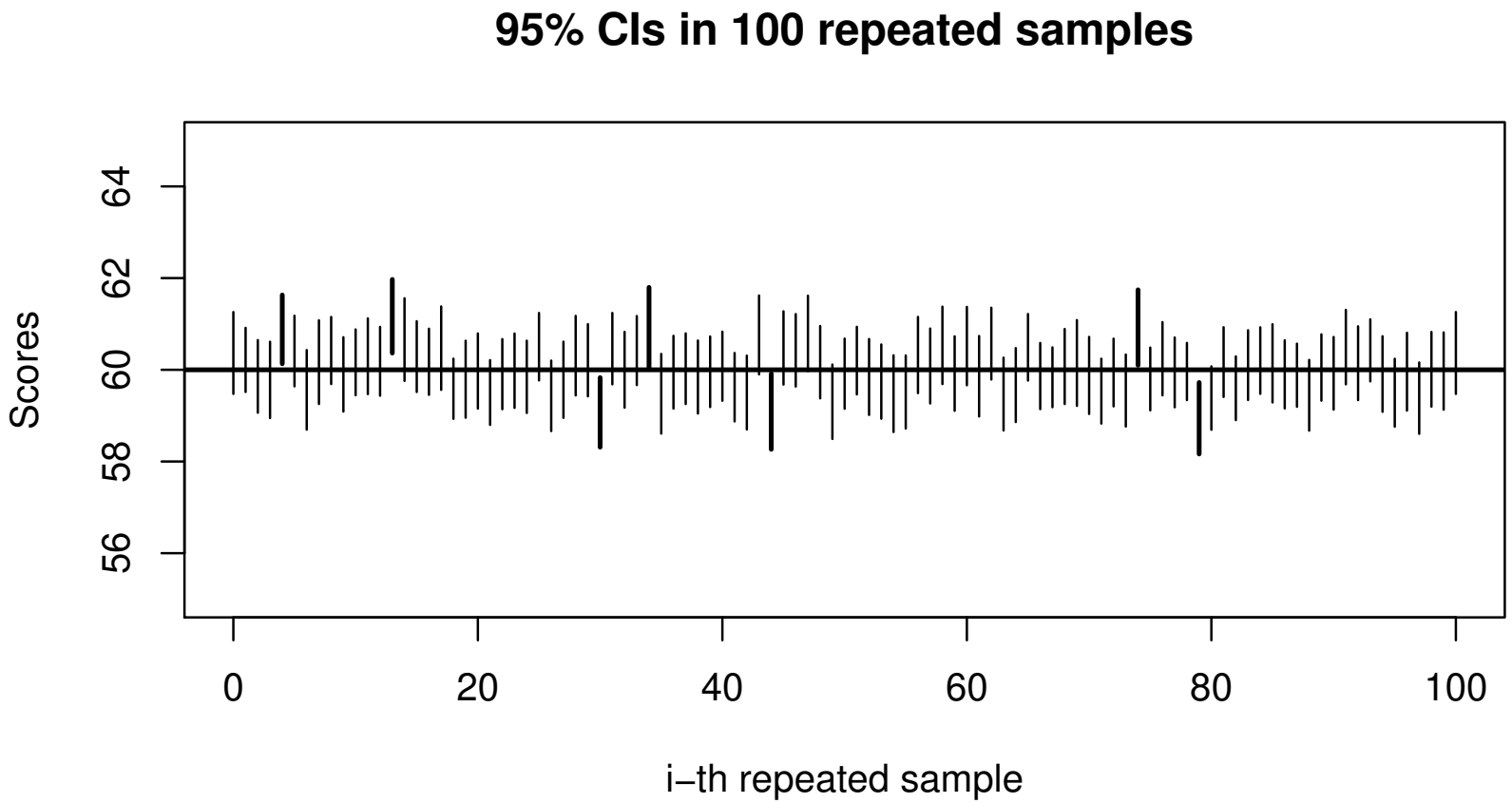
```
## check how many CIs contain mu:
CIs<-ifelse(lower<mu & upper>mu,1,0)
table(CIs)

## CIs
##   0   1
##  42 958

## approx. 95% of the CIs contain true mean:
table(CIs)[2]/sum(table(CIs))

##     1
## 0.958
```

The meaning of the 95% CI



The meaning of the 95% CI

1. The 95% CI from a particular sample does **not** mean that the probability that the true value of the mean lies inside that particular CI.
2. Thus, the CI has a very confusing and (not very useful!) interpretation.
3. In Bayesian statistics we use the credible interval, which has a much more sensible interpretation.

However, for large sample sizes, the credible and confidence intervals tend to be essentially identical.
For this reason, the CI is often treated (this is technically incorrect!) as a way to characterize uncertainty about our estimate of the mean.

Main points from this lecture

1. We compute maximum likelihood estimates of the mean $\bar{x} = \hat{\mu}$ and standard deviation $\hat{\sigma}$ to get estimates of the true but unknown parameters.
   $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
2. For a given sample, having estimated $\hat{\sigma}$, we estimate the standard error:
   $SE = \hat{\sigma}/\sqrt{n}$
3. This allows us to define a 95% CI about the estimated mean:
   $\hat{\mu} \pm 2 \times SE$

From here, we move on to statistical inference and null hypothesis significance testing (NHST).

1. We defined random variables.
2. We learnt about pdfs and cdfs, and learnt how to compute $P(X < x)$.
3. We learnt about Maximum Likelihood Estimation.
4. We learnt about the sampling distribution of the sample means.

This prepares the way for null hypothesis significance testing (NHST).

Hypothesis testing

Suppose we have a random sample of size n, and the data come from a $N(\mu, \sigma)$ distribution.
We can estimate sample mean $\bar{x} = \hat{\mu}$ and $\hat{\sigma}$, which in turn allows us to estimate the sampling distribution of the mean under (hypothetical) repeated sampling:

$$N(\bar{x}, \frac{\hat{\sigma}}{\sqrt{n}}) \tag{3}$$

The one-sample hypothesis test

Imagine taking an **independent** random sample from a random variable X that is normally distributed, with mean 12 and standard deviation 10, sample size 11. We estimate the mean and SE:

```
sample <- rnorm(11,mean=12,sd=10)
(x_bar<-mean(sample))

## [1] 8.1447

(SE<-sd(sample)/sqrt(11))

## [1] 3.0064
```

The one-sample test

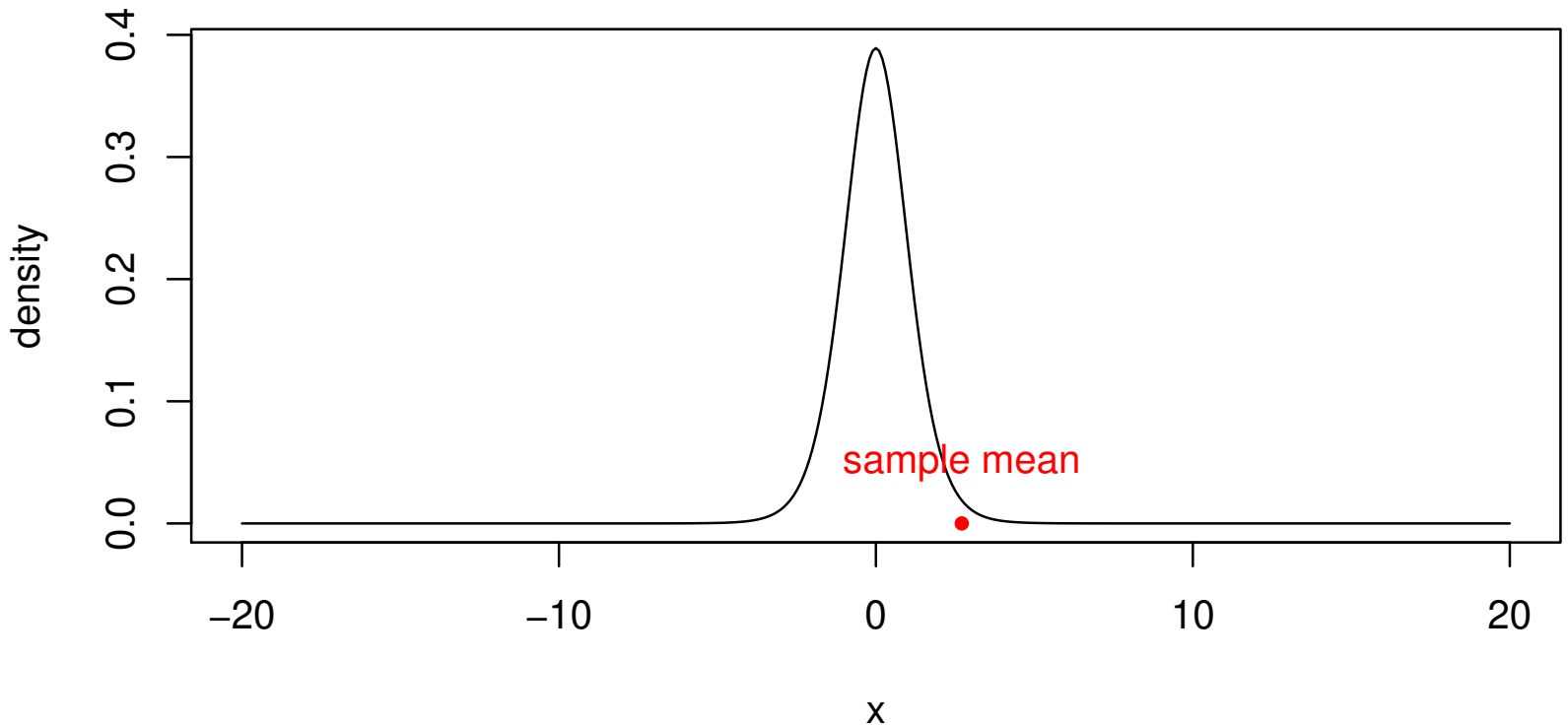
The NHST approach is to set up a null hypothesis that μ has some fixed value. For example:

$$H_0 : \mu = 0 \tag{4}$$

This amounts to assuming that the true distribution of sample means is (approximately*) normally distributed and centered around 0, *with the standard error estimated from the data*.

* I will make this more precise in a minute.

Null hypothesis distribution



NHST

The intuitive idea is that

1. if the sample mean $\bar{x}$ is near the hypothesized μ (here, 0), the data are (possibly) "consistent with" the null hypothesis distribution.
2. if the sample mean $\bar{x}$ is far from the hypothesized μ, the data are inconsistent with the null hypothesis distribution.

We formalize "near" and "far" by determining how many standard errors the sample mean is from the hypothesized mean:

$$t \times SE = \bar{x} - \mu \tag{5}$$

This quantifies the distance of sample mean from μ in SE units.

NHST

So, given a sample and null hypothesis mean μ, we can compute the quantity:

$$t = \frac{\bar{x} - \mu}{SE} \tag{6}$$

Call this the **t-value**. Its relevance will just become clear.

NHST

The quantity

$$T = \frac{\bar{X} - \mu}{SE} \tag{7}$$

has a t-distribution, which is defined in terms of the sample size n.
We will express this as: $T \sim t(n-1)$
Note also that, for large n, $T \sim N(0,1)$.

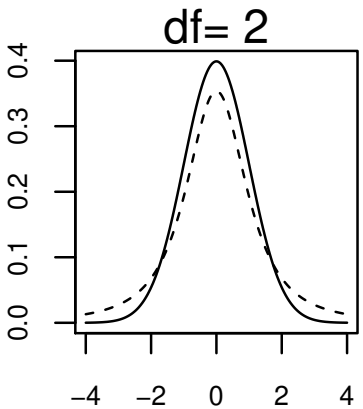
NHST

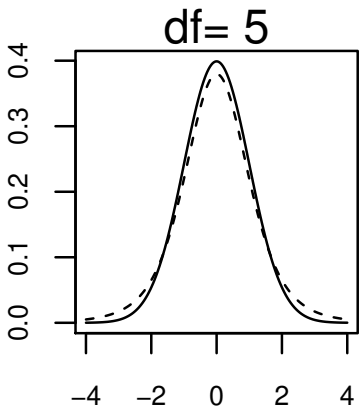
Thus, given a sample size n, and given our null hypothesis, we can draw t-distribution corresponding to the null hypothesis distribution.
For large n, we could even use N(0,1), although it is traditional in psychology and linguistics to always use the t-distribution no matter how large n is.

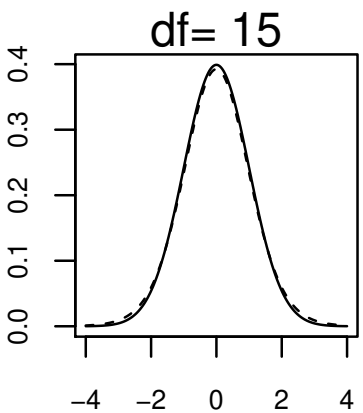
The t-distribution vs the normal

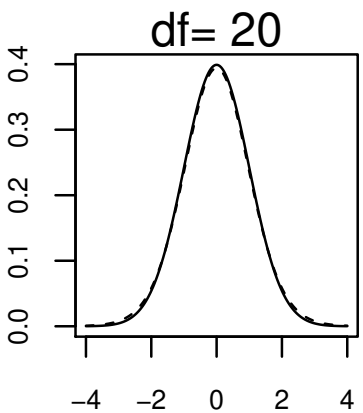
1. The t-distribution takes as parameter the degrees of freedom $n-1$, where n is the sample size (cf. the normal, which takes the mean and variance/standard deviation).
2. The t-distribution has fatter tails than the normal for small n, say $n < 20$, but for large n, the t-distribution and the normal are essentially identical.

The t-distribution vs the normal









t-test: Rejection region

So, the null hypothesis testing procedure is:

1. Define the null hypothesis: for example, $H_0 : \mu = 0$.
2. Given data of size n, estimate $\bar{x}$, standard deviation s, standard error $s/\sqrt{n}$.
3. Compute the t-value:

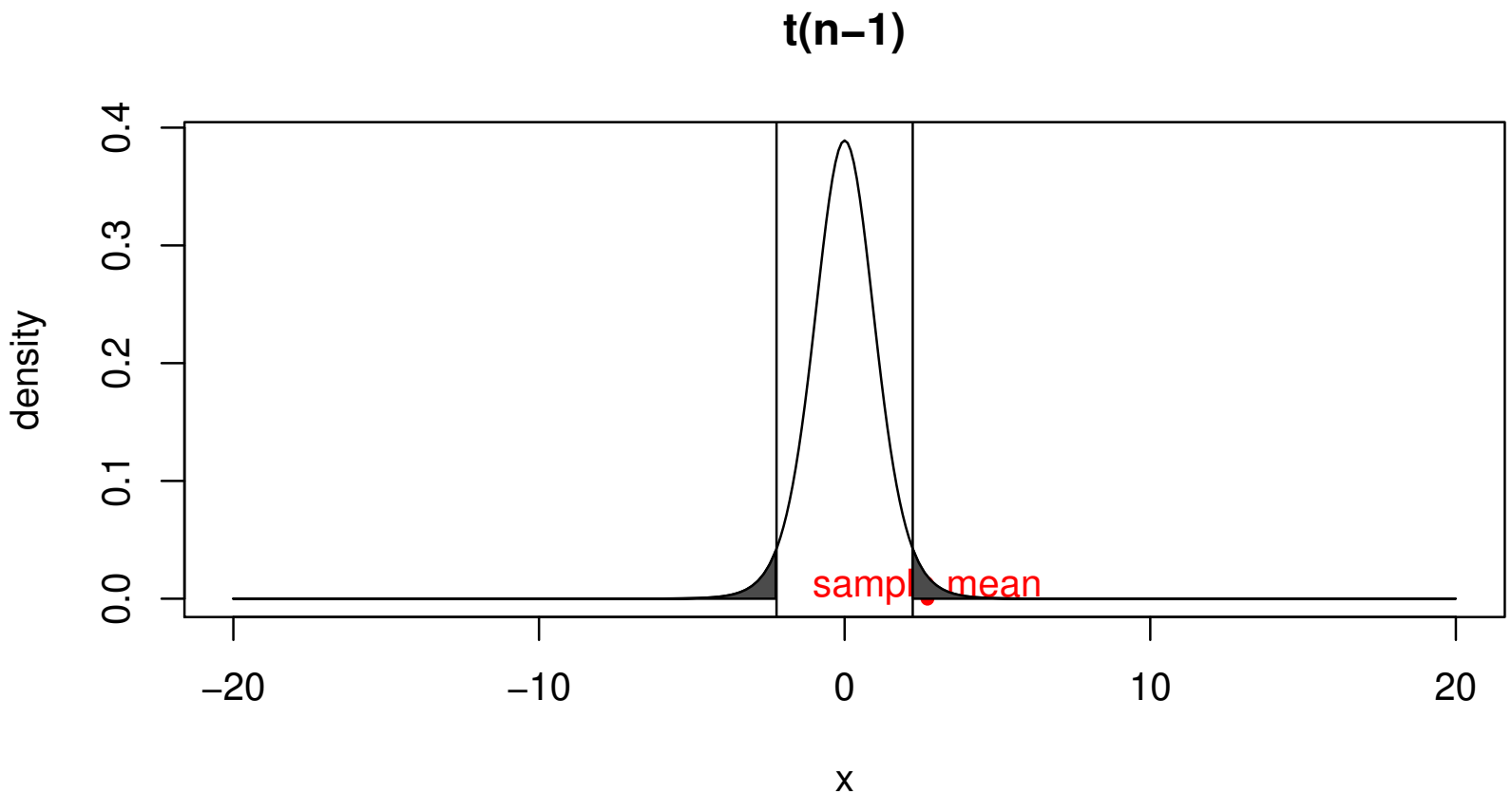
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{8}$$

4. Reject null hypothesis if t-value is large (to be made more precise next).

t-test

How to decide when to reject the null hypothesis? Intuitively, when the t-value from the sample is so large that we end up far in *either* tail of the distribution.

t-test



Rejection region

1. For a given sample size n, we can identify the "rejection region" by using the qt function (see lecture 1).
2. Because the shape of the t-distribution depends on the degrees of freedom (n-1), the **critical t-value** beyond which we reject the null hypothesis will change depending on sample size.
3. For large sample sizes, say $n > 50$, the rejection point is about 2.

```
abs(qt(0.025,df=15))

## [1] 2.1314

abs(qt(0.025,df=50))

## [1] 2.0086
```

t-test: Rejection region

Consider the t-value from our sample in our running example:

```
## null hypothesis mean:
mu<-0
(t_value<-(x_bar-mu)/SE)

## [1] 2.7091
```

Recall that the t-value from the sample is simply telling you the distance of the sample mean from the null hypothesis mean μ in standard error units.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \text{ or } t\frac{s}{\sqrt{n}} = \bar{x} - \mu \tag{9}$$

t-test: Rejection region

So, for large sample sizes, if $\mid t \mid > 2$ (approximately), we can reject the null hypothesis.
For a smaller sample size n, you can compute the exact critical t-value:

```
qt(0.025,df=n-1)
```

This is the critical t-value on the **left**-hand side of the t-distribution. The corresponding value on the right-hand side is:

```
qt(0.975,df=n-1)
```

Their absolute values are of course identical (the distribution is symmetric).

The t-distribution vs the normal

Given the relevant degrees of freedom, one can compute the area under the curve as for the Normal distribution:

```
pt(-2,df=10)

## [1] 0.036694

pt(-2,df=20)

## [1] 0.029633

pt(-2,df=50)

## [1] 0.025474
```

Notice that with large degrees of freedom, the area under the curve to the left of -2 is approximately 0.025

The t.test function

The t.test function in R delivers the t-value:

```
## from t-test function:
## t-value
t.test(sample)$statistic

##      t
## 2.7091
```