Statistics: Description and inference

Part II: the t-test

Statistical theory and process control

- Consider the problem facing Gossett or any quality control engineer. At a brewery (or any factory), beer (or widgets) are produced to meet certain specifications. There is a certain amount of variation from specifications that is acceptable, but you need to detect when something has gone wrong; i.e., when specifications are no longer being met. How can you tell if the product is being made up to specification?
- Two basic cases: Large samples and small samples

Data = Model + Error

- Almost all of statistics can be summarized as finding how well a model fits the data.
- We need to specify a model, observe the phenomenon, and see how far off the model is from the data.
- Always ask: What is the model? How well does it fit? What are the alternatives? How well do they fit?

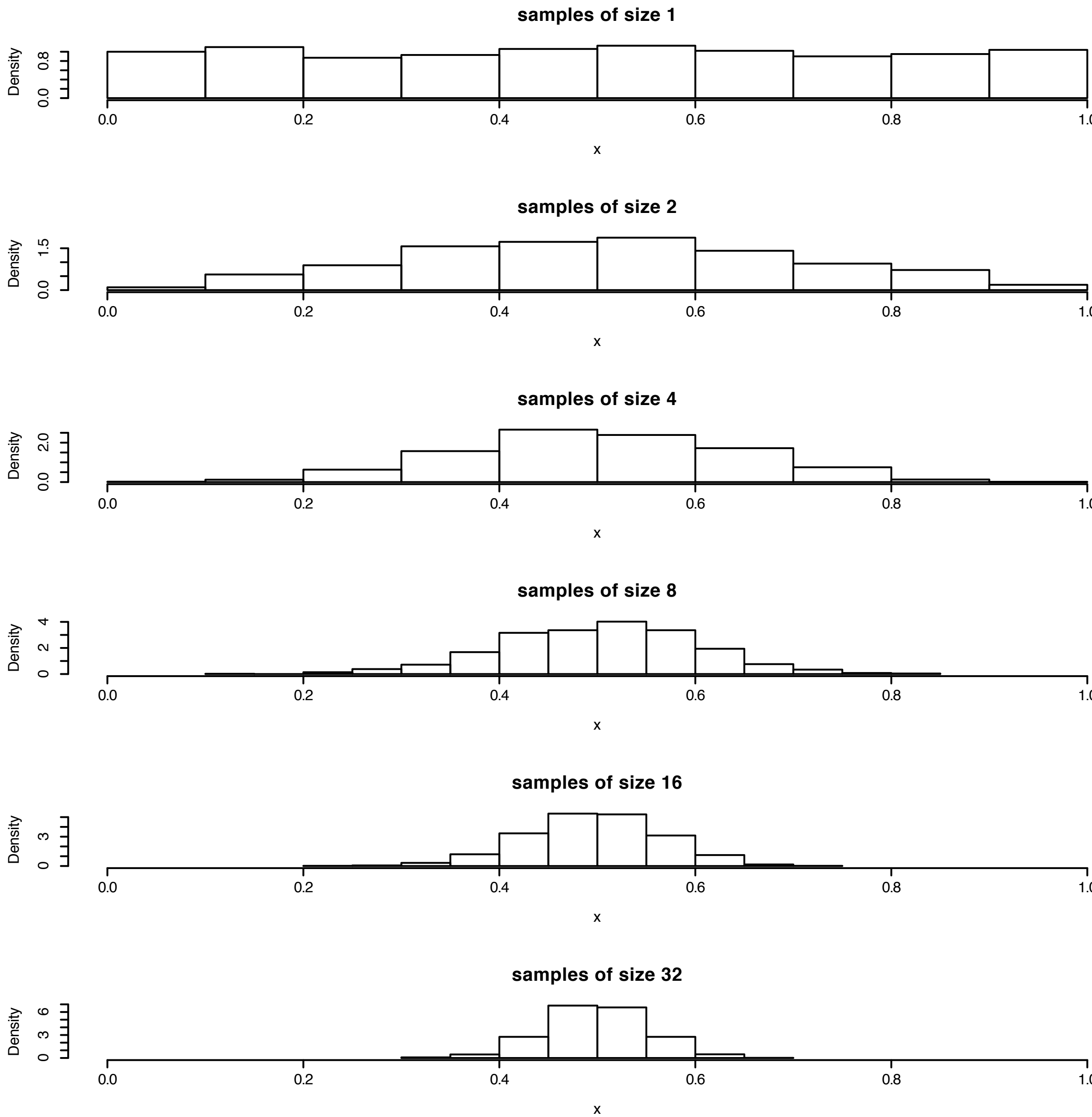
Normal distributions and the central limit theorem

- The distribution of sample means from a population with mean μ and variance σ^2 will tend towards a normal distribution with mean μ and variance σ^2/n

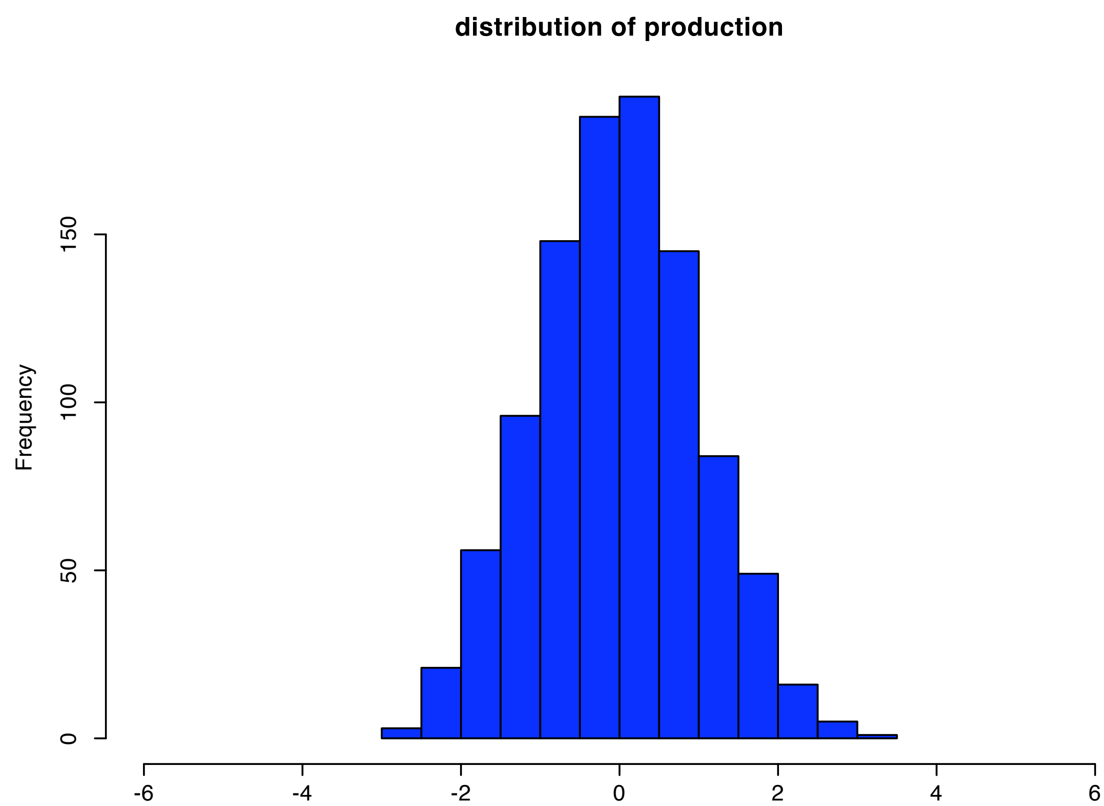
$$s^2 = \Sigma_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

$$s.e. = \sqrt{s^2/n} = s/\sqrt{n}$$

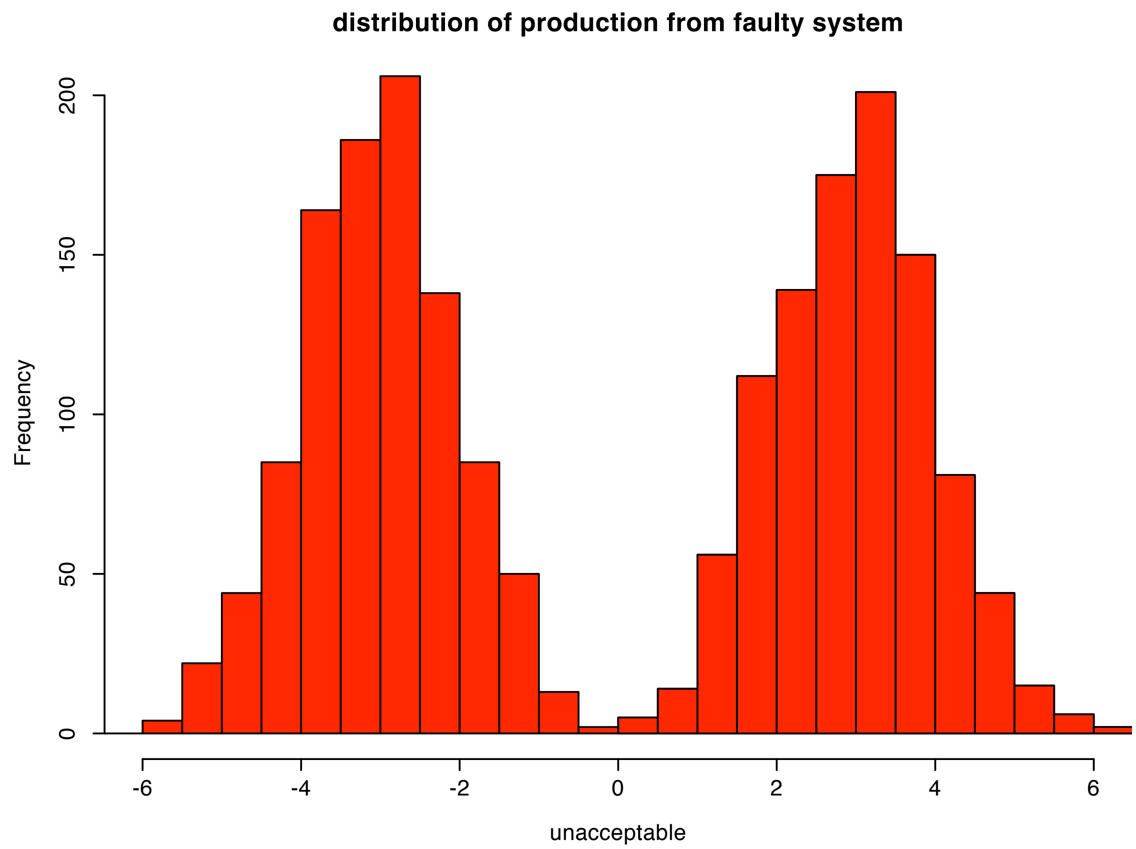
Distributions as f(sample size)



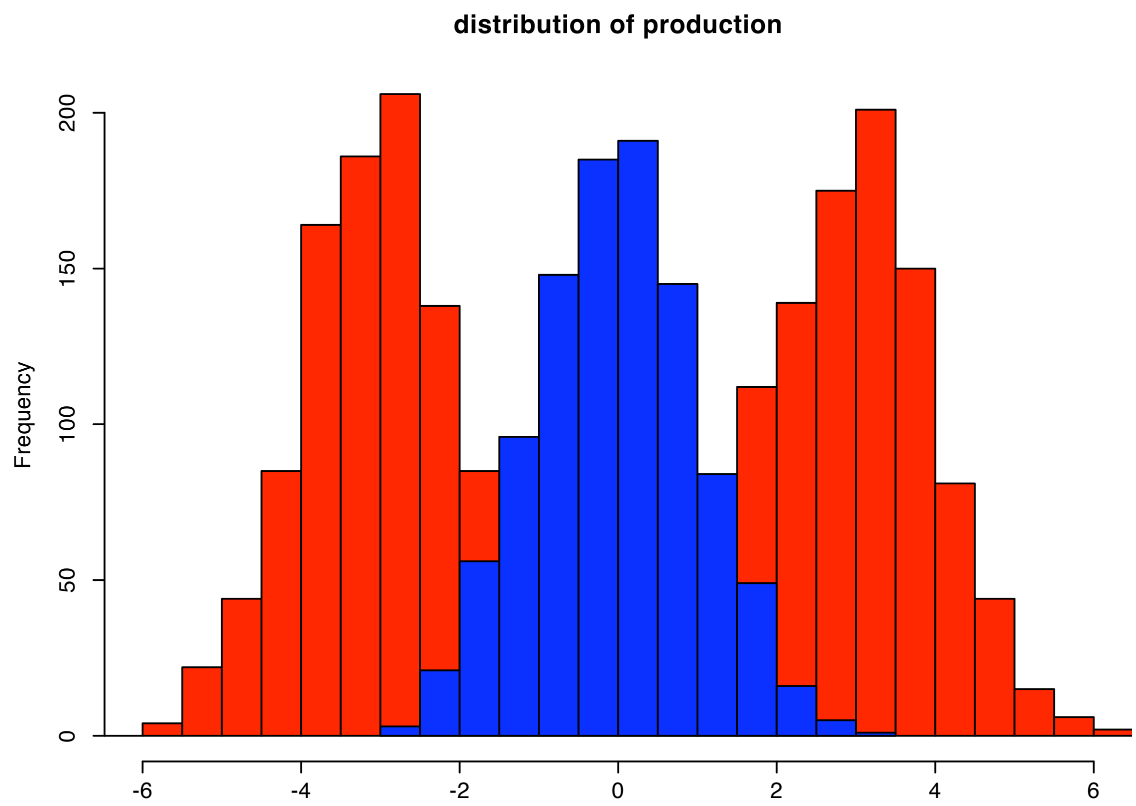
Basic Production process with mean = μ= 0 and σ = 1



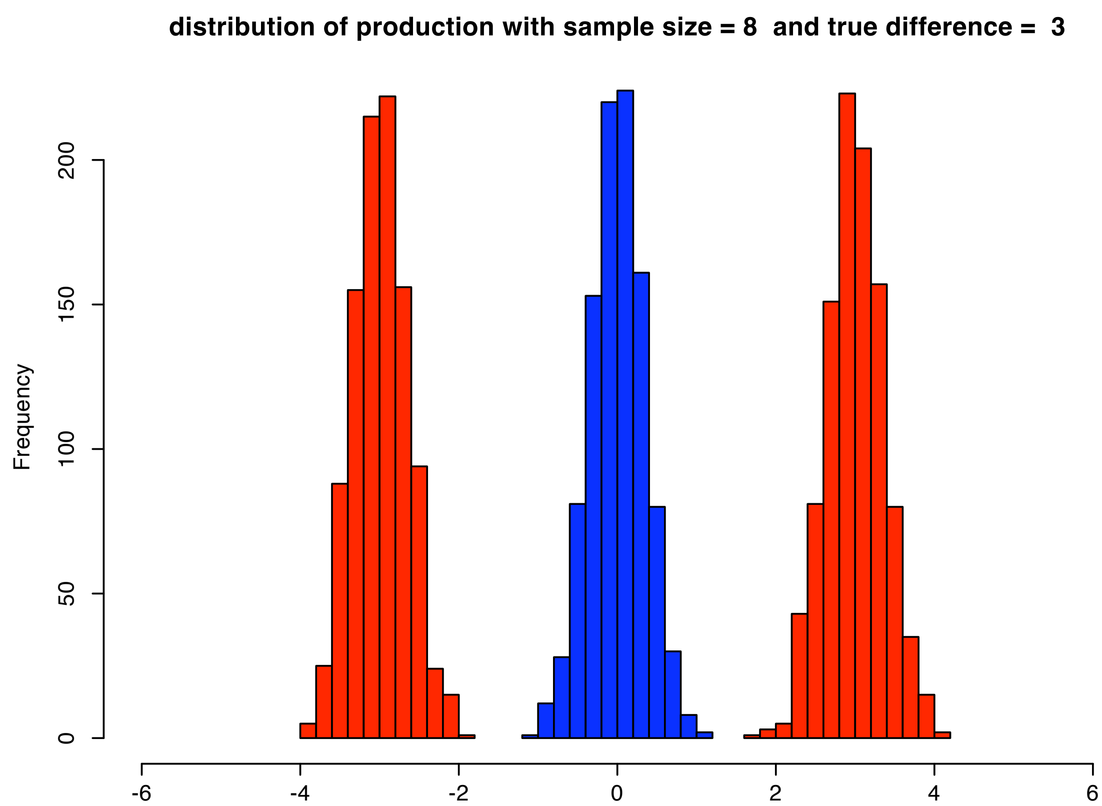
Production can go bad



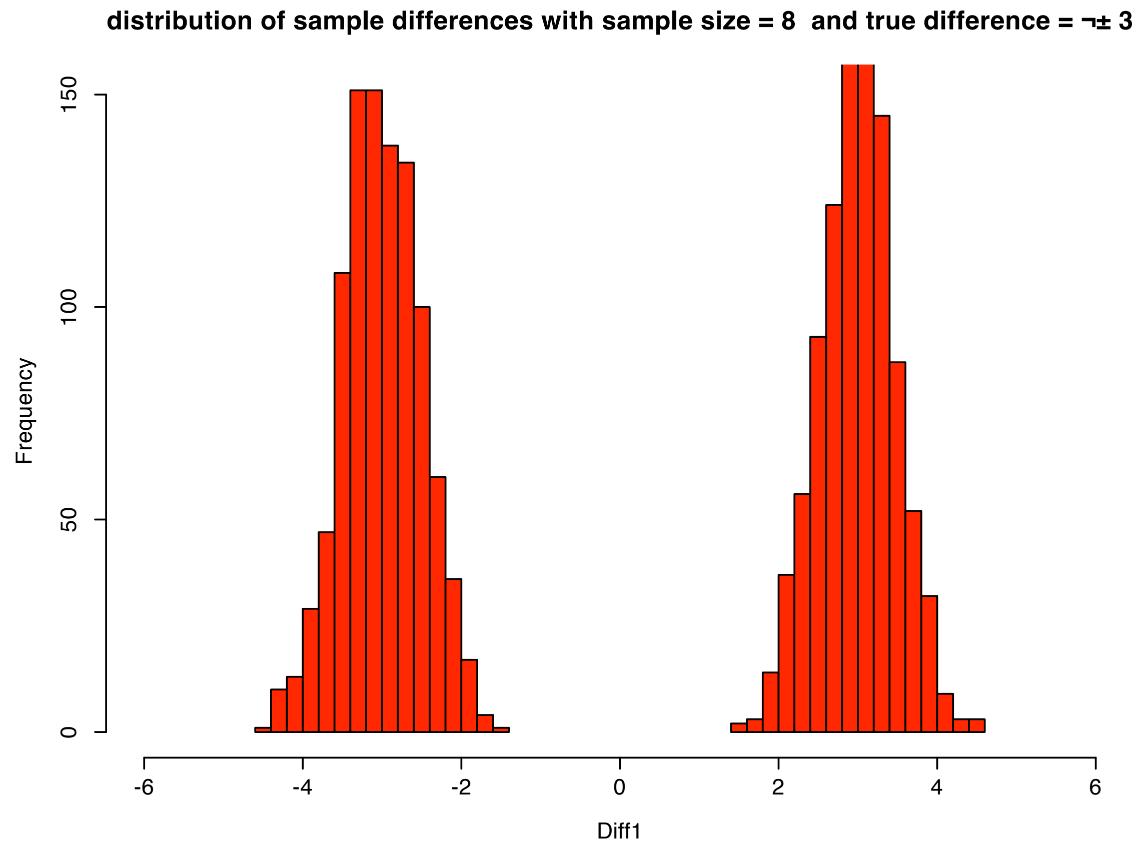
Problem of estimating which state the brewery is in



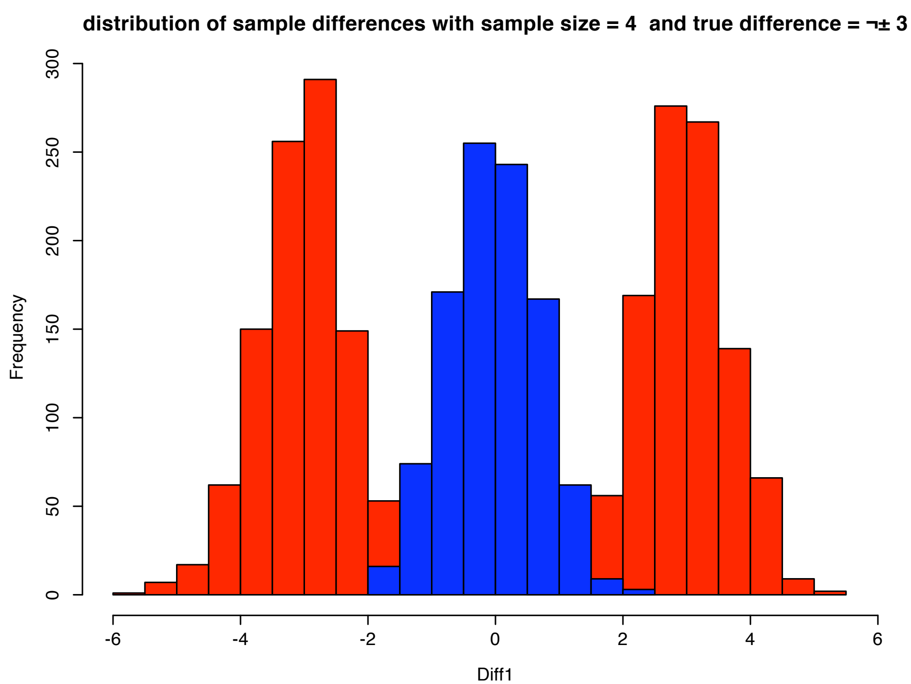
Consider samples of size n

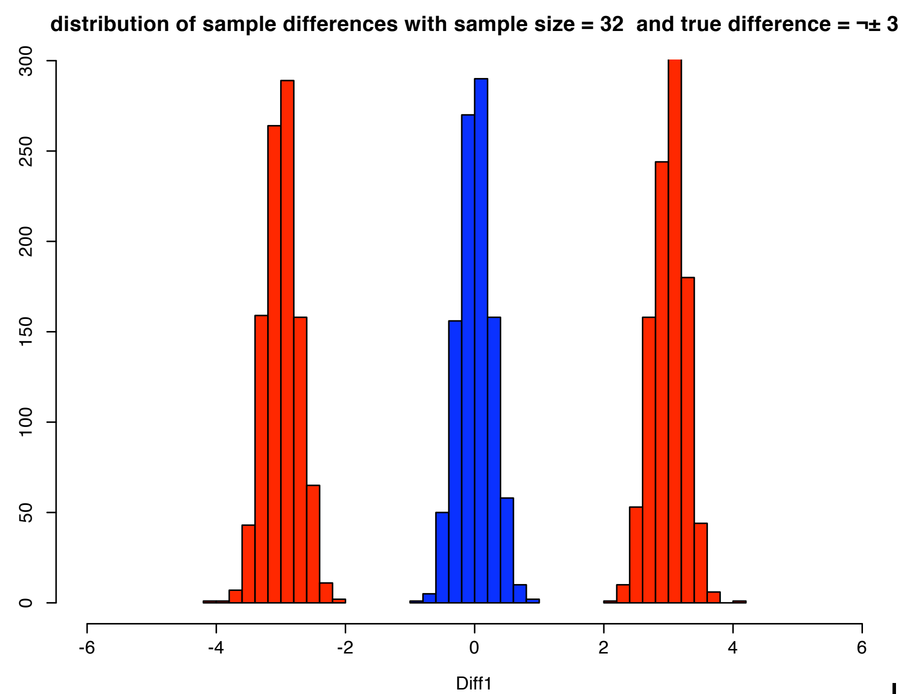


Consider distribution of sample differences

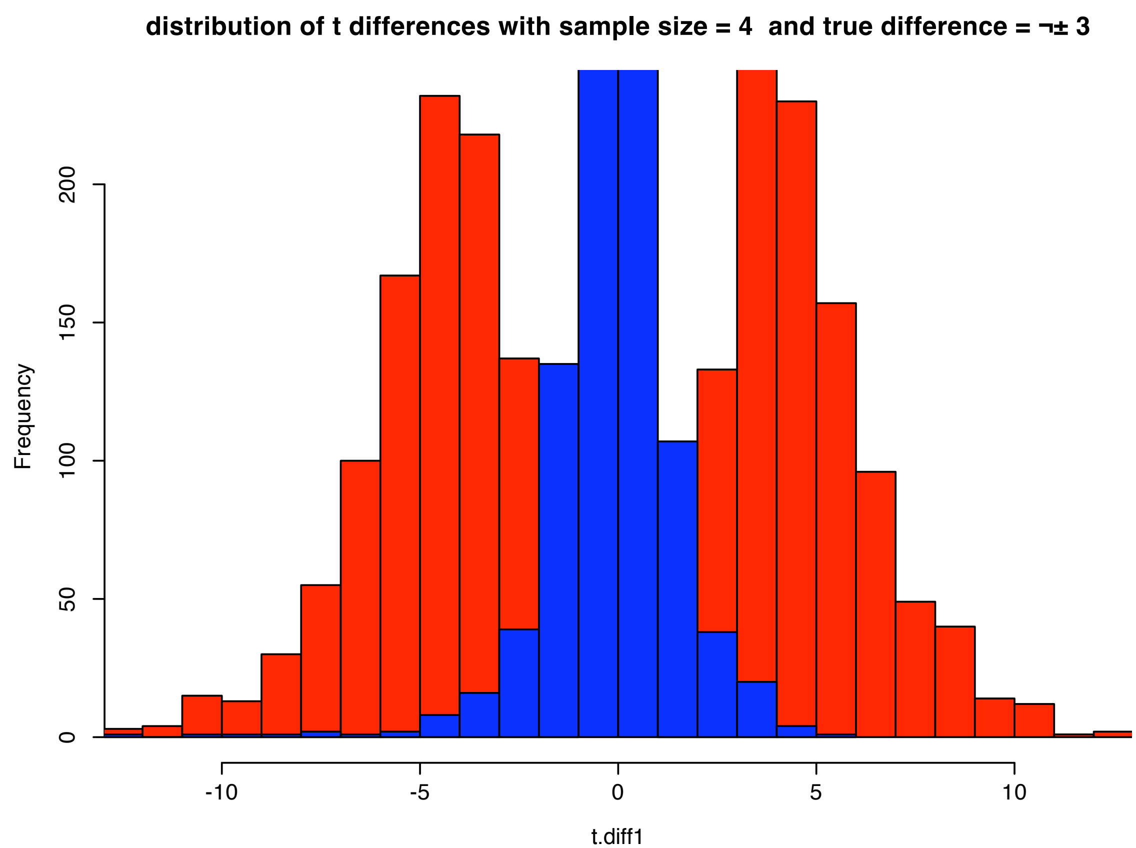


Variation of group differences depend upon sample size

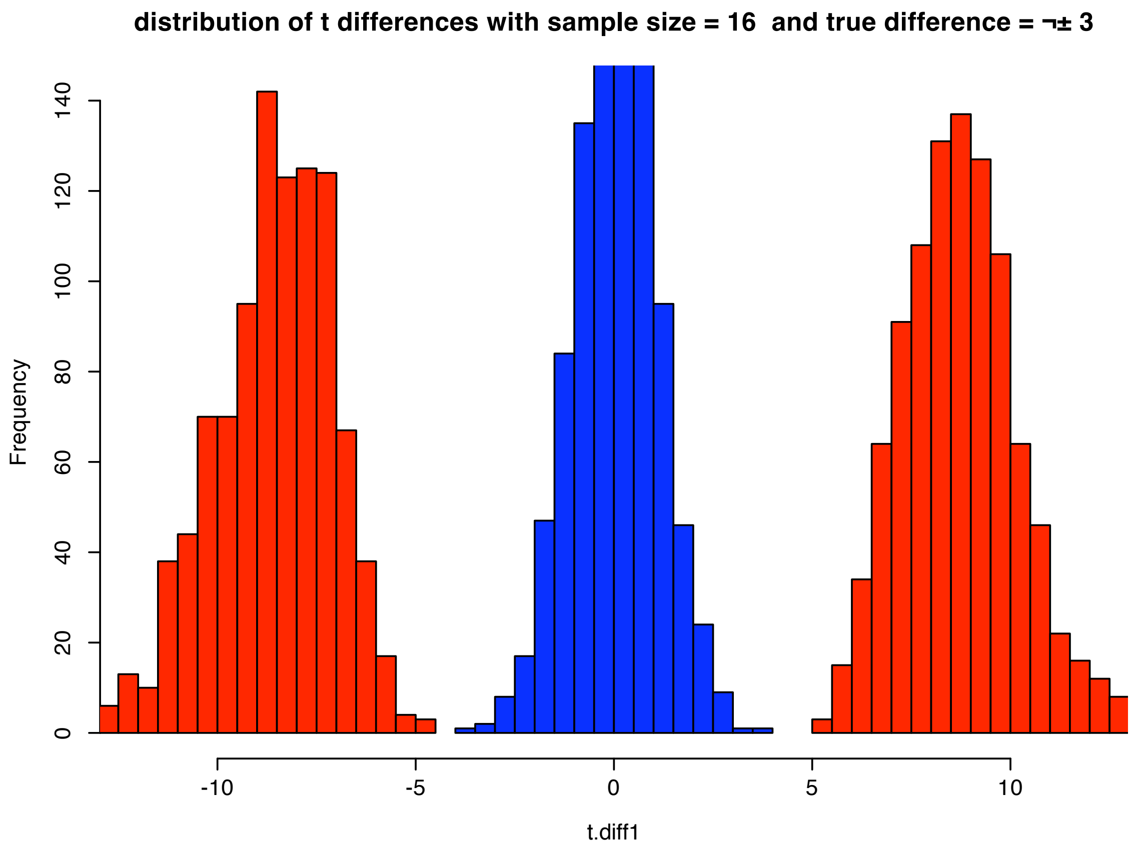




Gossett and the t-test
(compare differences of means to standard error of mean)

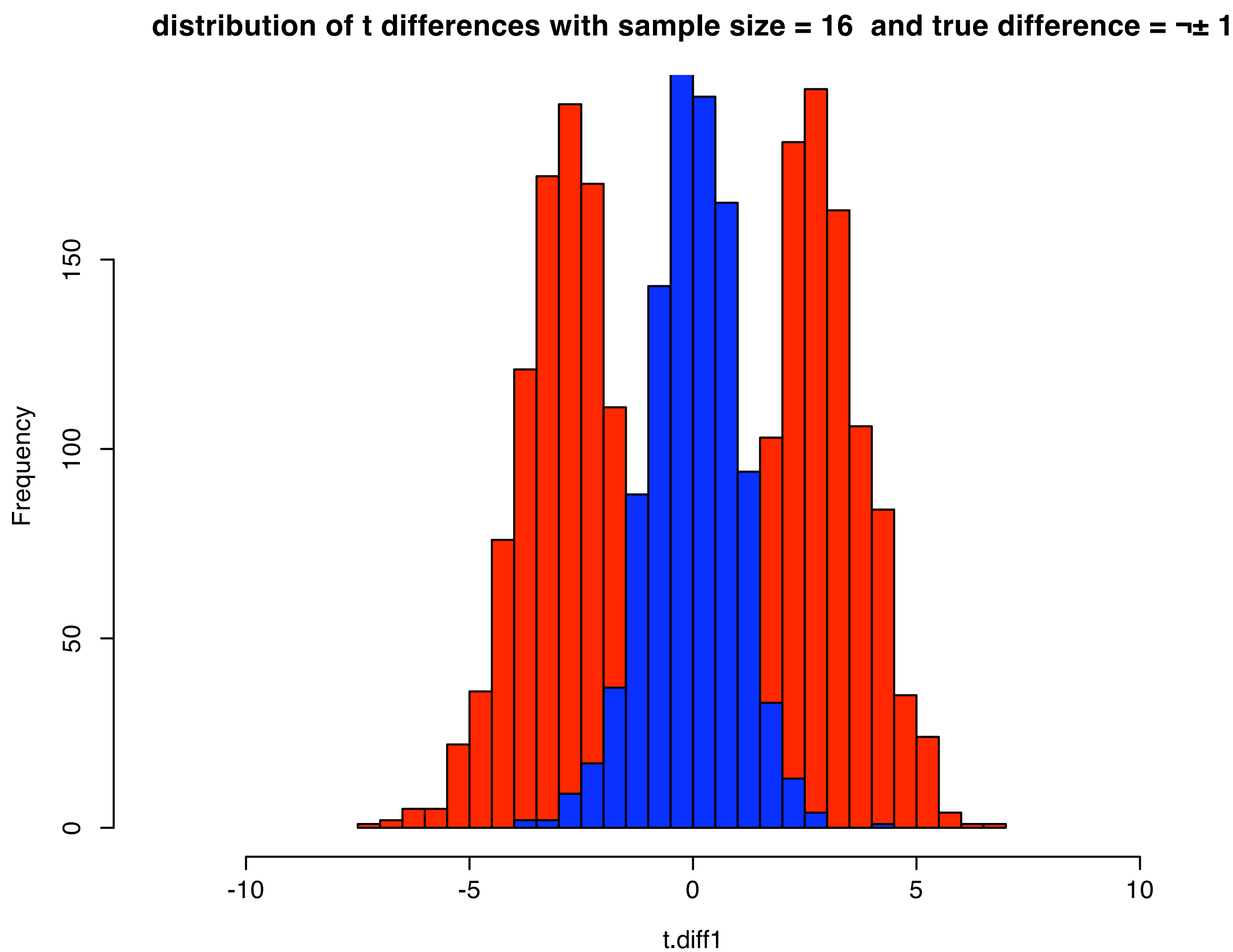


A sample of 16, diff = 3



Types of inferential errors

failure to detect, failure to reject



Descriptions and inference

- Classical "Null Hypothesis Inference Statistical Test" NHIST
- Descriptive statistics with confidence intervals
  - expressed in units of measurement
  - expressed in "effect sizes"

Null Hypothesis Testing

- The Null or Nill hypothesis of no difference
- Alternative hypothesis is that Null is wrong
- What is the likelihood of observing differences this big or bigger if Null is true
- If likelihood given Null is small, then reject Null
- Error of false rejection when Null is True (Type I)
- Error of failure to reject when Null is false (Type II)

Critique of NHIST

- Null is never true
- It is not that something has an effect, but we want to know how big the effect is.
- Hookes Law is not that if you pull on a wire it gets longer but rather that the amount it stretches is proportional to the force.
- We need to estimate quantities, not just see if they are $\neq 0$

Descriptive with confidence

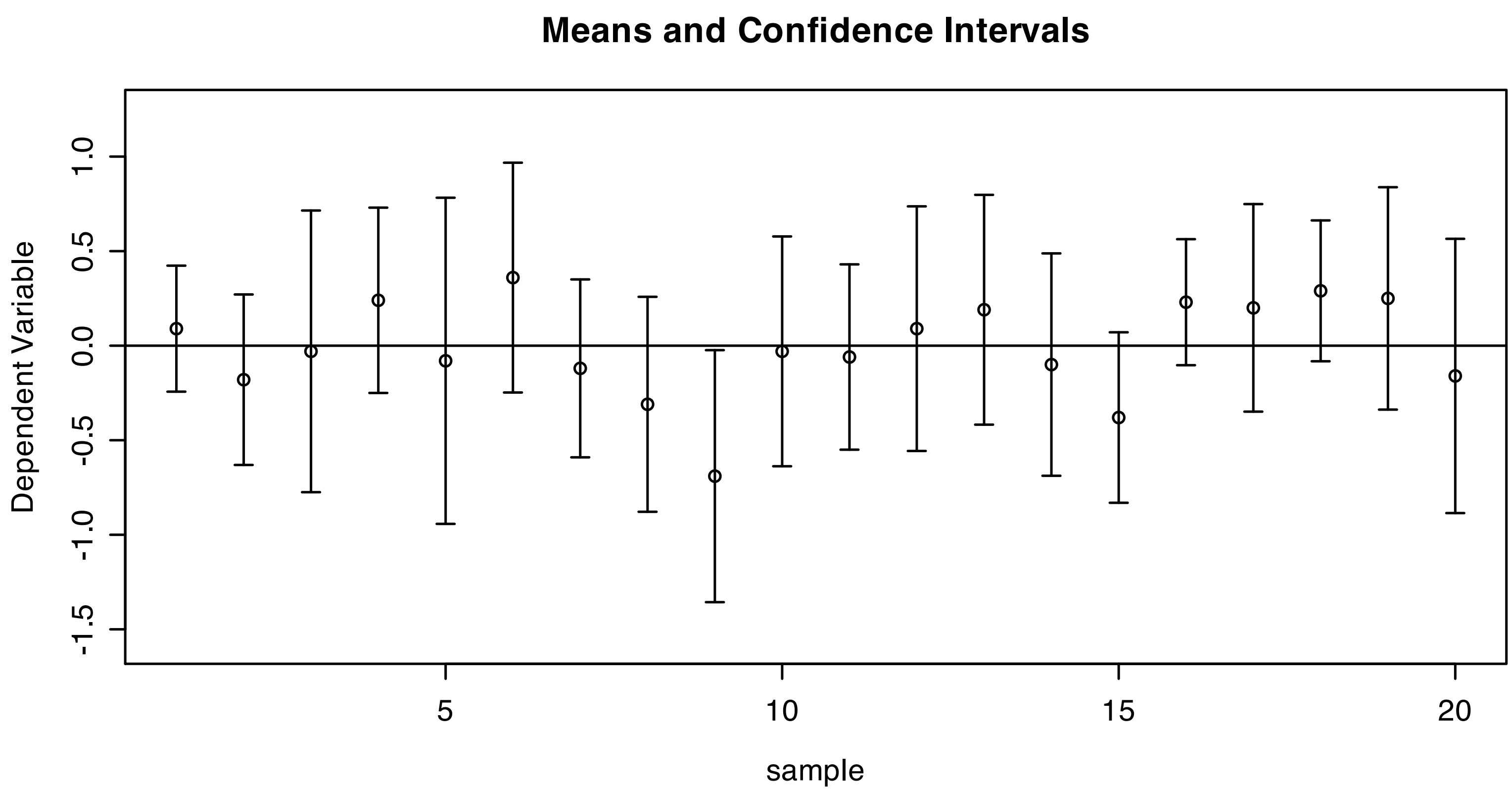
- Standard error = $s/\sqrt{N}$
  - observed standard deviation/sqrt(sample size)
- Report observed mean and the standard error of the mean. Allows us to estimate the precision of the estimate.
- If population mean is X, then 68% of observed means will be within 1 se of X, 95% within 2 se of X

Effect size comparisons

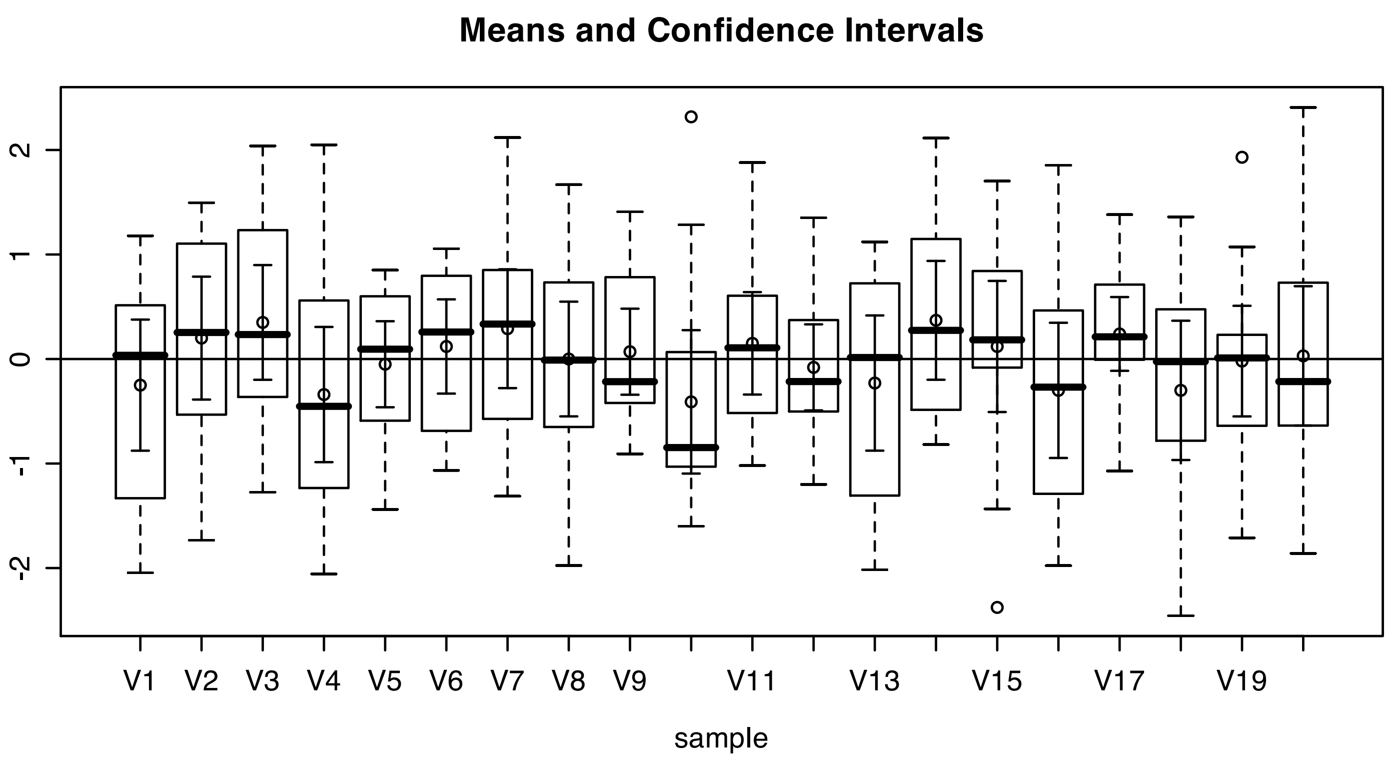
- Effect (e.g., difference of means) depends upon the scale we use (meters, feet, inches)
- Standardized effect = effect/within group standard deviation
- Note that while t = effect/sqrt(n) and thus varies as a function of sample size,standardized effect size does not depend upon sample size and thus allows one to compare effects across studies

Consider multiple samples

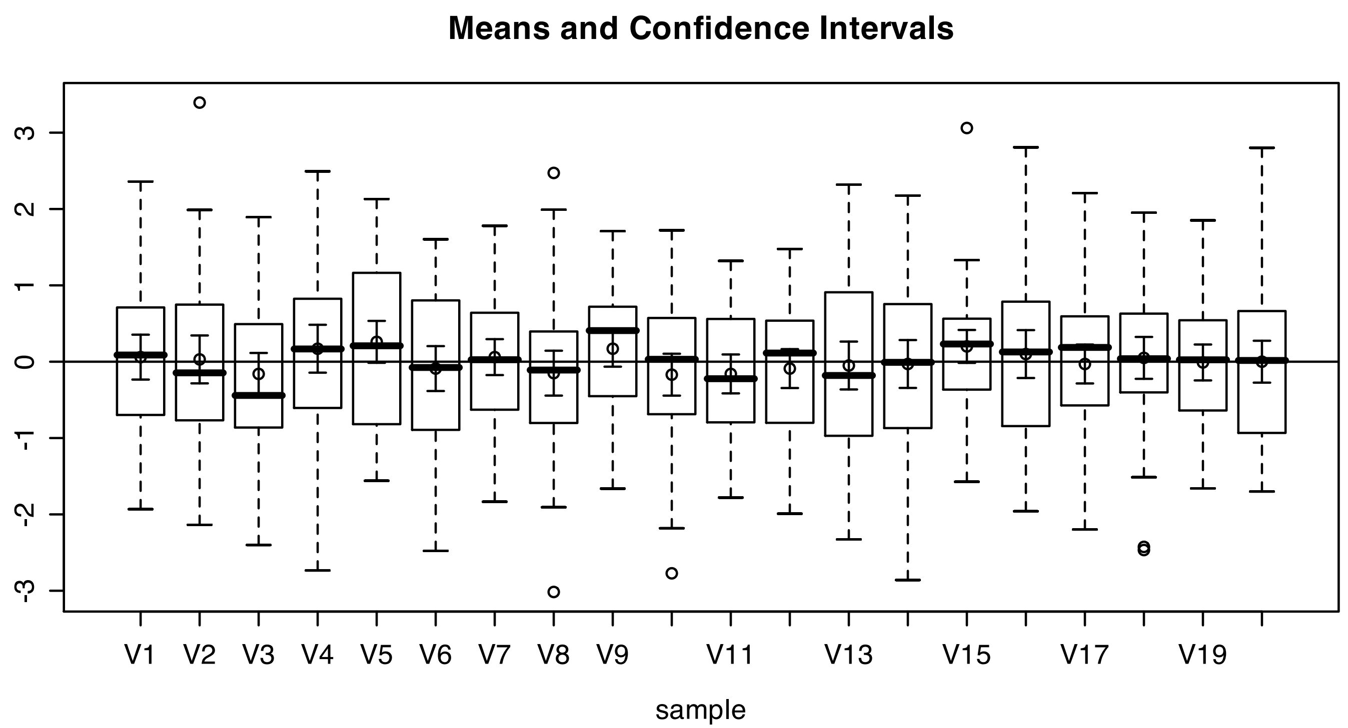
```
> x<-matrix(rnorm(240),ncol=20)
>error.bars(x,xlab="sample",main="Means and Confidence Intervals")
>abline(h=0)
```



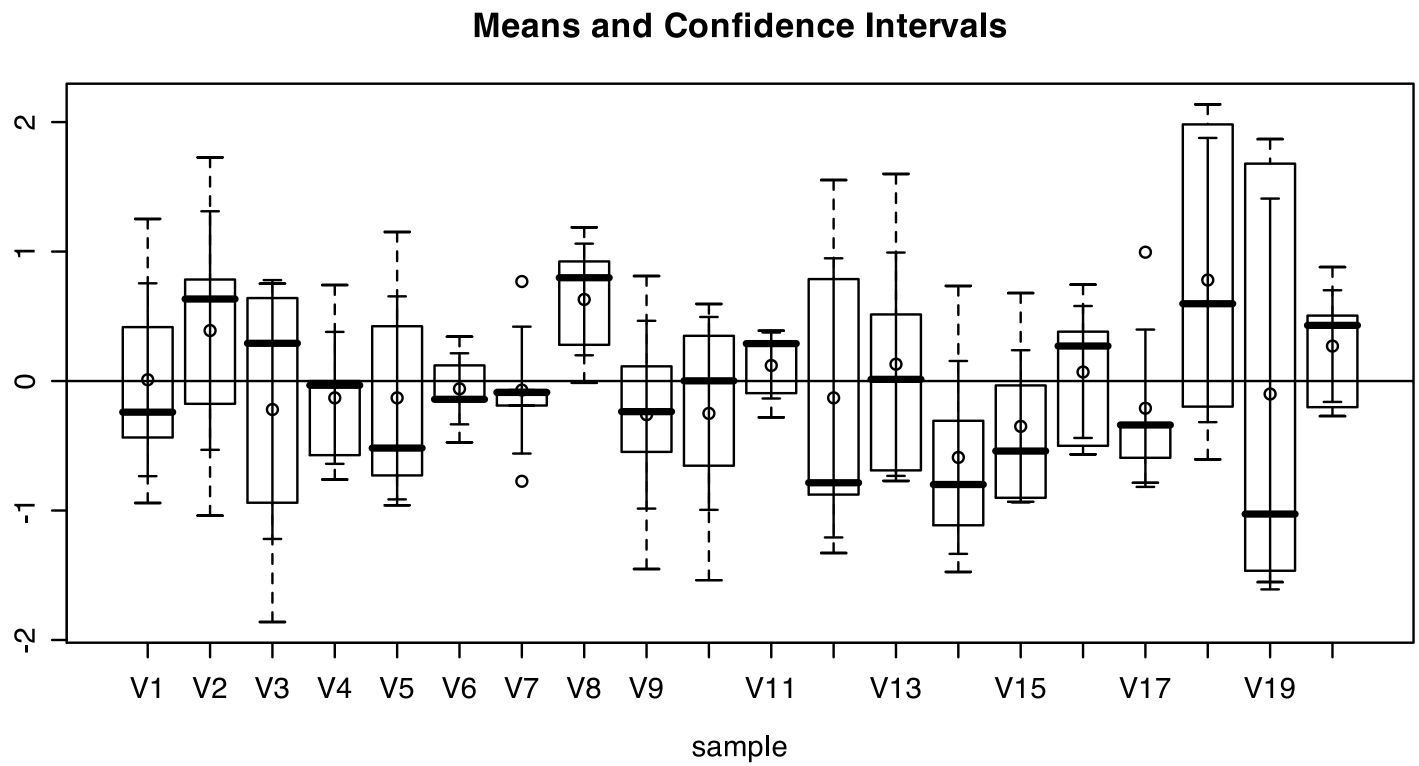
Another 20 samples of size 24



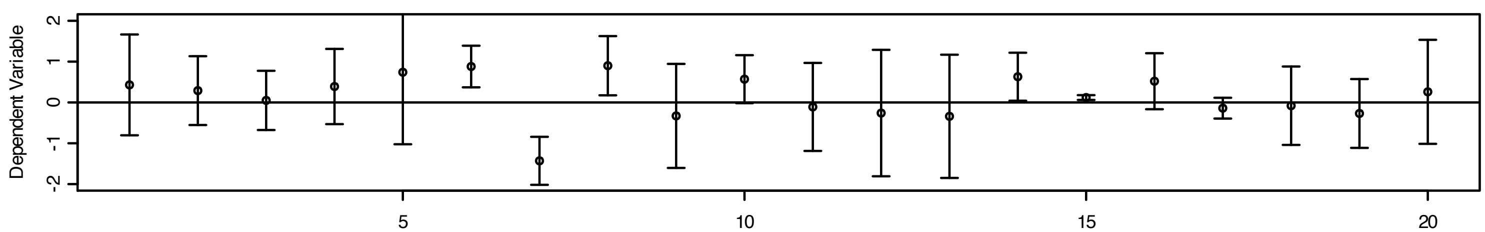
20 samples of size 50

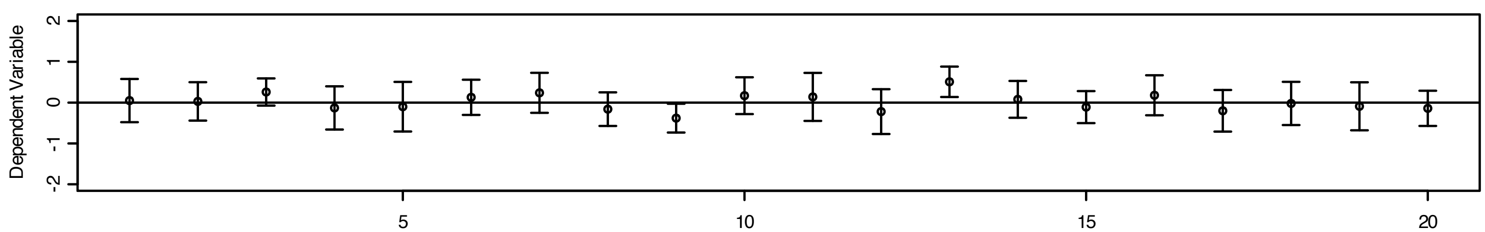


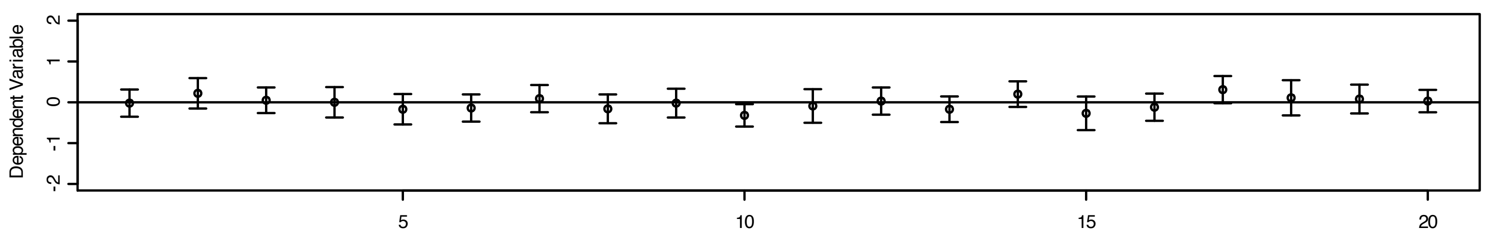
20 samples of size 5



20 samples of size 4, 16, 32







Confidence Intervals and precision

- As sample size increases, the confidence intervals get smaller
- But the probability of being included in a 95% confidence interval remains 95%!

Finding t

The data

placebo caffeine
24 24
25 29
27 26
26 23
26 25
22 28
21 27
22 24
23 27
25 28
25 27
25 26

Analysis

Get the data

```
spelling <- read.clipboard()
```

Describe the data

```
describe(spelling)
```

```
         var  n  mean   sd median  mad min max range  skew kurtosis   se
placebo    1 12 24.25 1.86   25.0 1.48  21  27     6 -0.33    -1.33 0.54
caffeine   2 12 26.17 1.85   26.5 2.22  23  29     6 -0.22    -1.33 0.53
```

t.test in R

```
> attach(spelling)
> t.test(placebo,caffeine,equal.var=TRUE)

   Welch Two Sample t-test

data:  placebo and caffeine
t = -2.5273, df = 21.999, p-value = 0.01918
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.4894368 -0.3438965
sample estimates:
mean of x mean of y
 24.25000  26.16667


         var  n  mean   sd median  mad min max range  skew kurtosis   se
placebo    1 12 24.25 1.86   25.0 1.48  21  27     6 -0.33    -1.33 0.54
caffeine   2 12 26.17 1.85   26.5 2.22  23  29     6 -0.22    -1.33 0.53
```

Reporting the t.test

- Formally (and formerly!):
  - The hypothesis of no difference between the groups was rejected with a probability of $p < .02$
- More typical:
  - Caffeine (26.17, sd. = 1.85) led to an increase in spelling performance when compared to placebo (24.25, sd. = 1.86), $t = 2.53$, $p < .02$.
  - Some also report the probability of replication:
    - p.rep = .95