# Statistical Inference:
# Estimation and Confidence Intervals,
# Hypothesis testing

Brian D. Ripley

# Student's t distribution

Only rarely do we know $\sigma^2$. We do have an estimator of it, $s^2$. Can we just plug that in? Not quite!

We know

$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

if we replace $\sigma$ by $s$ we will increase the variability of the left-hand-side and so 'smear out' the distribution. However, the effect will be small if $s^2$ is an accurate estimator of $\sigma^2$.

The correct answer was guessed by W. S. Gossett (pseudonym 'Student') and proved by R. A. Fisher.

$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_\nu, \qquad \nu = n - 1$$

The parameter $\nu$ is known as the number of 'degrees of freedom'.

# Confidence Intervals – Example

We will illustrate this by a population of 334 measurements. Two nurses were asked to measure tuberculin reaction sizes (in mm), and the dataset is the set of 334 *differences*. So they are whole numbers, with summary

```
> print(summary(react), digits=4)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 -9.000 -2.000  -1.000 -0.796  0.000   8.000
```

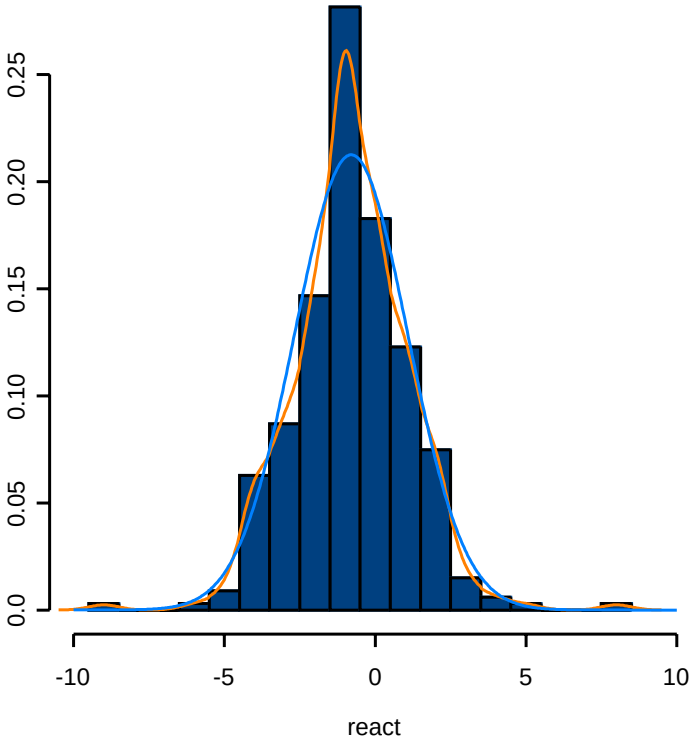
Viewed as a random sample from a large population, we can ask if the difference is significantly different from zero, and if so by how much:

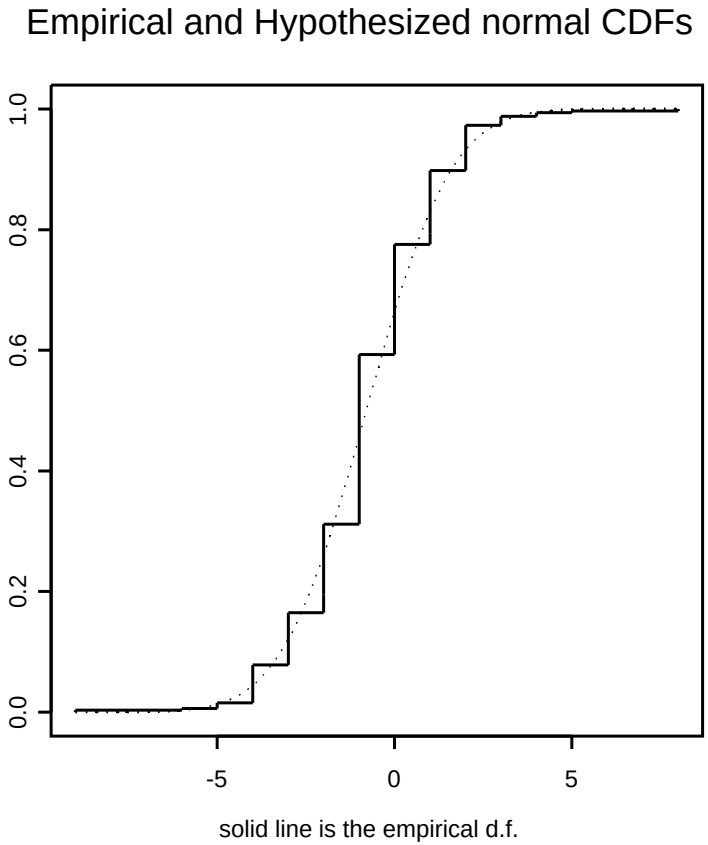
```
> t.test(react)
t = -7.75, df = 333, p-value = 0
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.999 -0.594
```

Alternatively, if we suspected in advance that the difference would be negative and did not care *at all* that it would be positive we could use

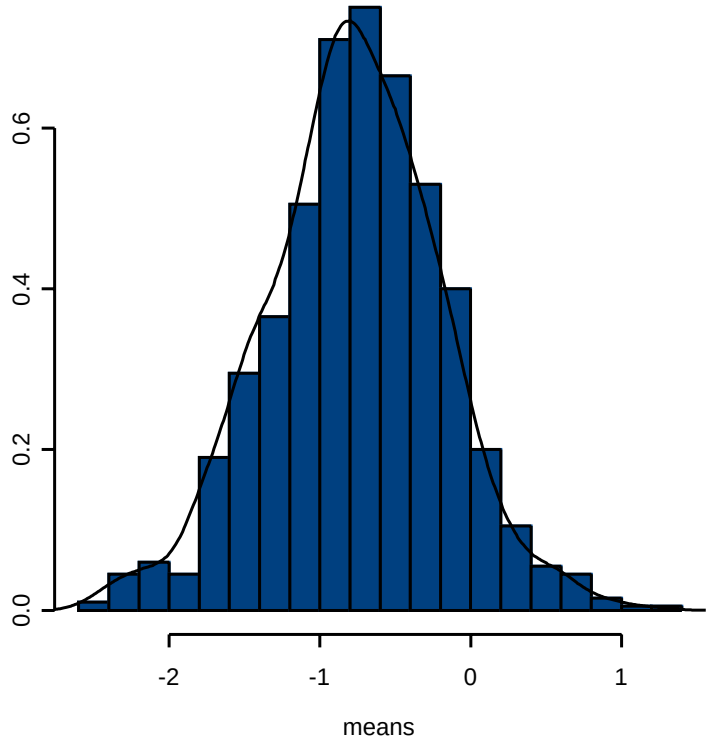
```
> t.test(react, alternative="less")
t = -7.75, df = 333, p-value = 0
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
     NA -0.627
```

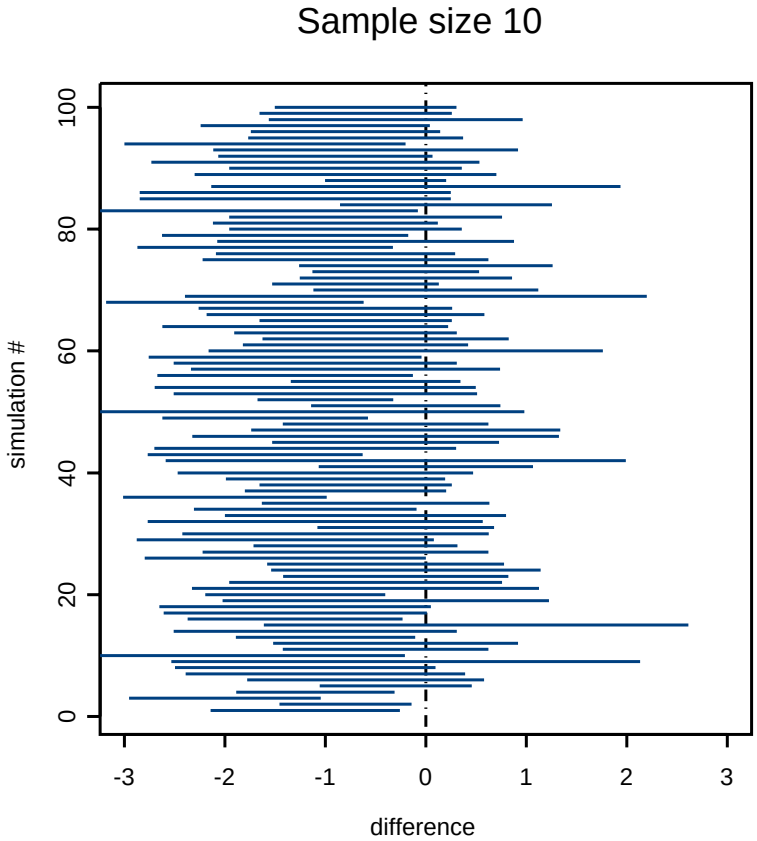
# Plots of the population



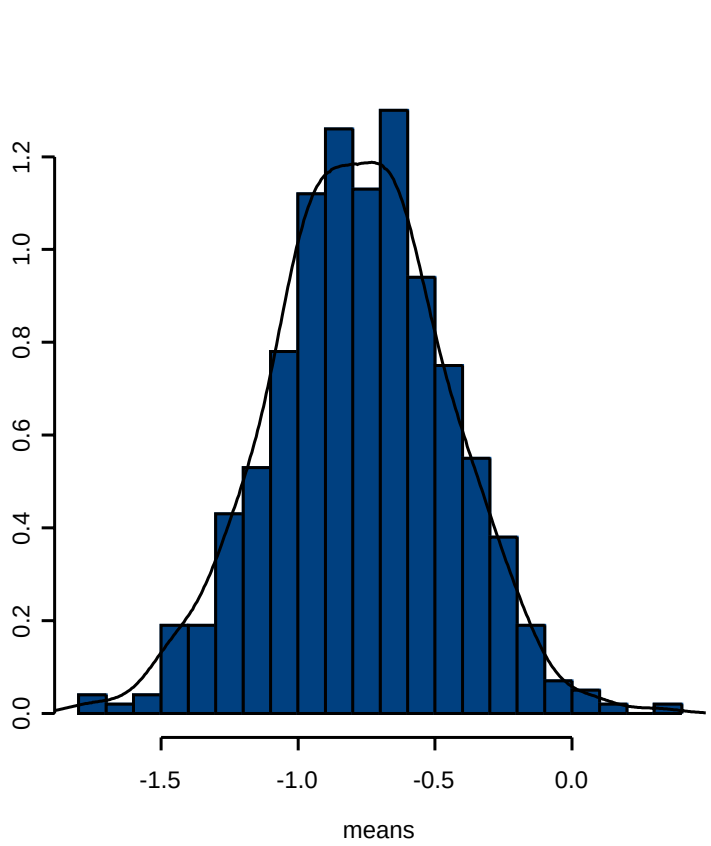



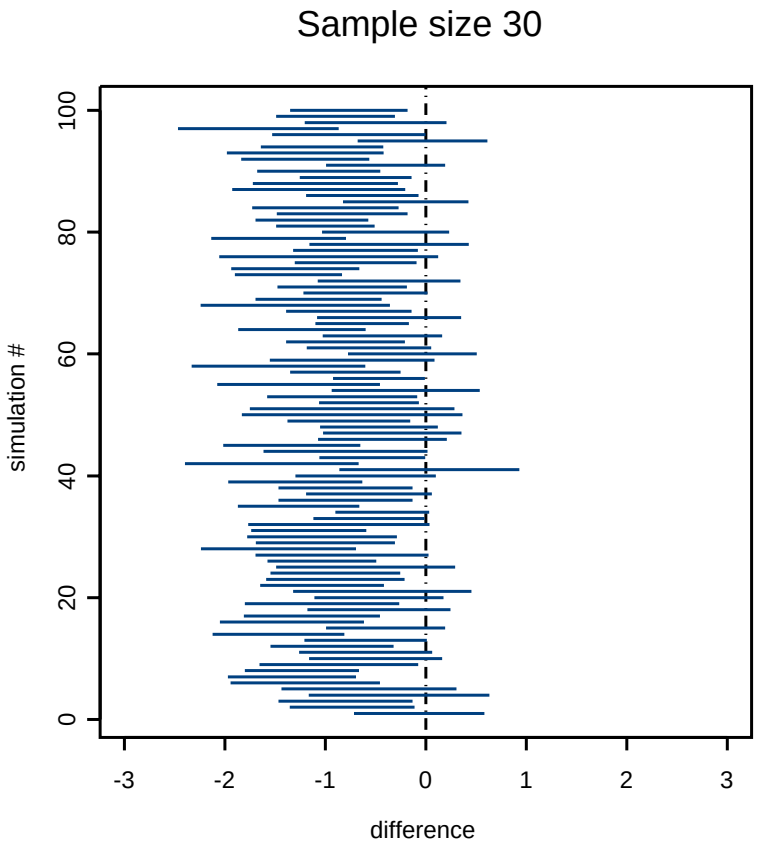
# Samples of size 10



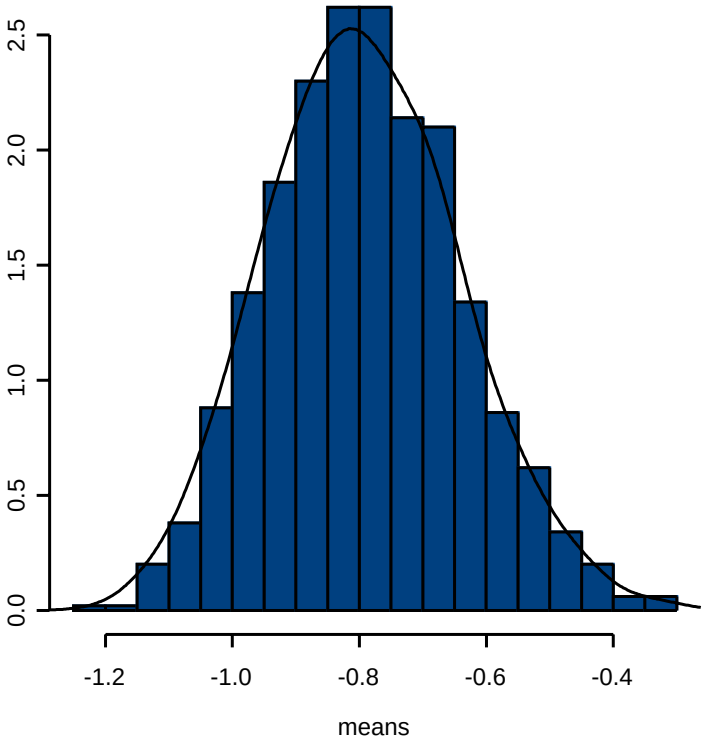



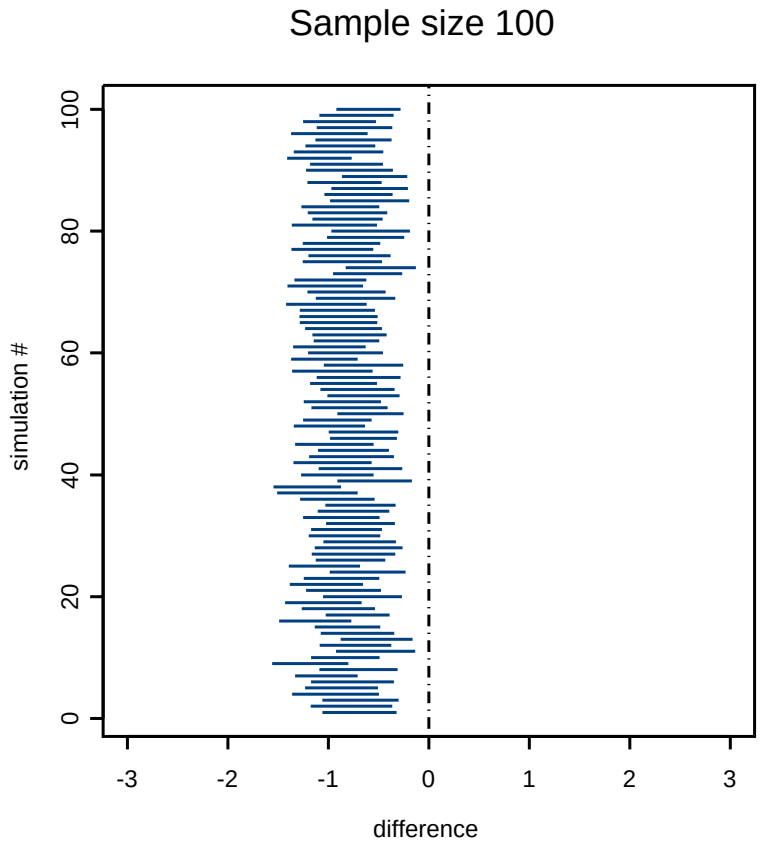
# Samples of size 30



# Samples of size 100



# Tests of location

Suppose we are not interested in the difference between the two nurses only if there is one. Then we want a hypothesis test of $H_0 : \mu = \mu_0 = 0$ against $H_1 : \mu \neq \mu_0$. To do so we choose a statistic $T$ that is larger (in probability) under $H_1$ than under $H_0$. The absolute value of the $t$ statistic

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

is an obvious candidate.

We can then proceed in either of two ways:

1. We compute a 'p-value'. Suppose we observed $T = t_0$. The p-value is the probability of observing a value of $T$ at least as extreme by chance under the null distribution. In our example that is $P_{\mu_0}(|T| > t_0) = 2P(Y > t_0)$ where $Y \sim t_{n-1}$. A small p-value is evidence against the null hypothesis.

2. Alternatively, we can set a *significance level* $\alpha$ and choose a *critical value* $c$ such that $P_{H_0}(T > c) \leqslant \alpha$. We then reject the null hypothesis if we observe a value greater than $c$.

   Note that we reject (or not) the null hypothesis: we never 'accept' hypotheses.

   Conventional significance levels are 5%, 1% and 0.1%, marked by one, two or three 'significance stars' respectively. Don't!

Note that we reject the null hypothesis at level $\alpha$ if and only if the p-value is less or equal to $\alpha$, and also if and only if a confidence interval of size $1 - \alpha$ does not cover $\mu_0$.

# Tests of spread

Sometimes we have two samples and we want to test if they differ in scale: taking logs usually works. What is we want to test if they differ in spread about possibly different locations. We can compute their variances. The salaries for male bank clerks had variance 477113, for females 291460? Are they really different?

Suppose we have sample variances $s_1^2$ and $s_2^2$ from samples of size $n_1$ and $n_2$. Their ratio looks like a good test statistic for $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$. Under the null-hypothesis

$$\frac{s_1^2}{s_2^2} \sim F_{n_1,n_2}$$

another famous distribution, Snecedor's F (for Fisher).

We reject the null hypothesis if the ratio is too far from one. In our example

```
> var.test(bankM, bankF)
F = 1.64, num df = 31, denom df = 60, p-value = 0.102
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.906 3.147
```

The confidence interval is for the ratio $\lambda = \sigma_1^2/\sigma_2^2$, and arises from considering $\log(s_1^2/s_2^2) - \log \lambda$.

# Two-sample inference

Now supposed we want to compare the locations of two samples, $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$.

The easier case is when we know (or have tested) that the variances are the same. Then we can suppose $X_i \sim N(\mu_1, \sigma^2)$, $Y_j \sim N(\mu_2, \sigma^2)$ and we are interested in $\delta = \mu_1 - \mu_2$. Then $\overline{x} - \overline{y}$ has mean $\delta$ and variance $\sigma^2/(1/n_1 + 1/n_2)$. This suggests working with the statistic

$$\frac{(\overline{x} - \overline{y}) - \delta_0}{s/\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where $s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ is the *pooled* variance estimate.

Otherwise we use

$$\frac{(\overline{x} - \overline{y}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_\nu$$

which is an approximation, with a formula for $\nu$ involving $n_i$ and $s_i^2$.

# Bank clerks' salaries

There were 32 male and 61 female bank clerks. Here we are only interested in discrimination against women (??)

```
> t.test(bankM, bankF, alternative="greater", var.equal = F)

        Welch Modified Two-Sample t-Test

t = 5.83, df = 51.3, p-value = 0
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 583  NA
t.test(bankM, bankF, alternative="greater", var.equal = F)$p.value
[1] 1.86e-007
```

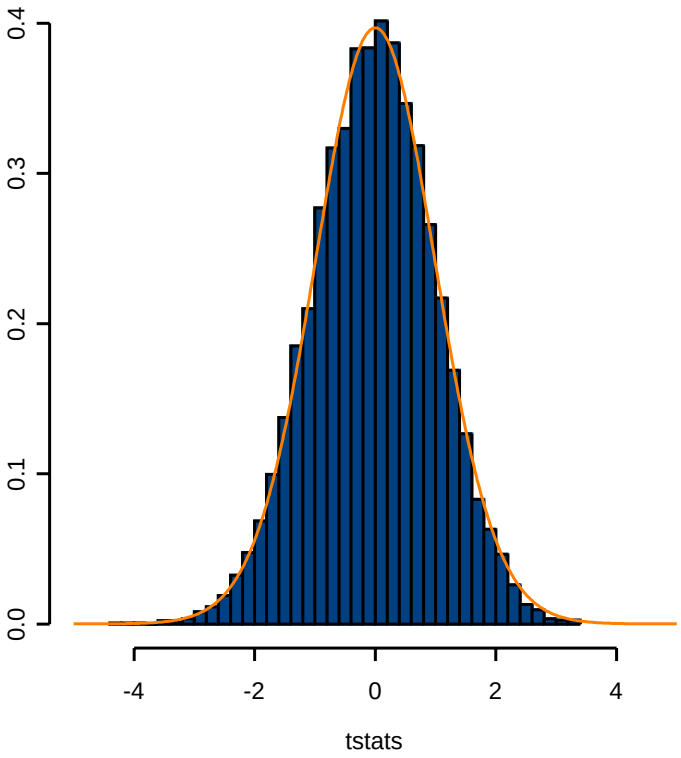
Another way to look at this is to ask if the difference would arise by chance if gender were irrelevant, and so we could randomly assign genders. Let's pick 32 of the 32 + 61 people randomly as 'male'. We did that 10000 times:

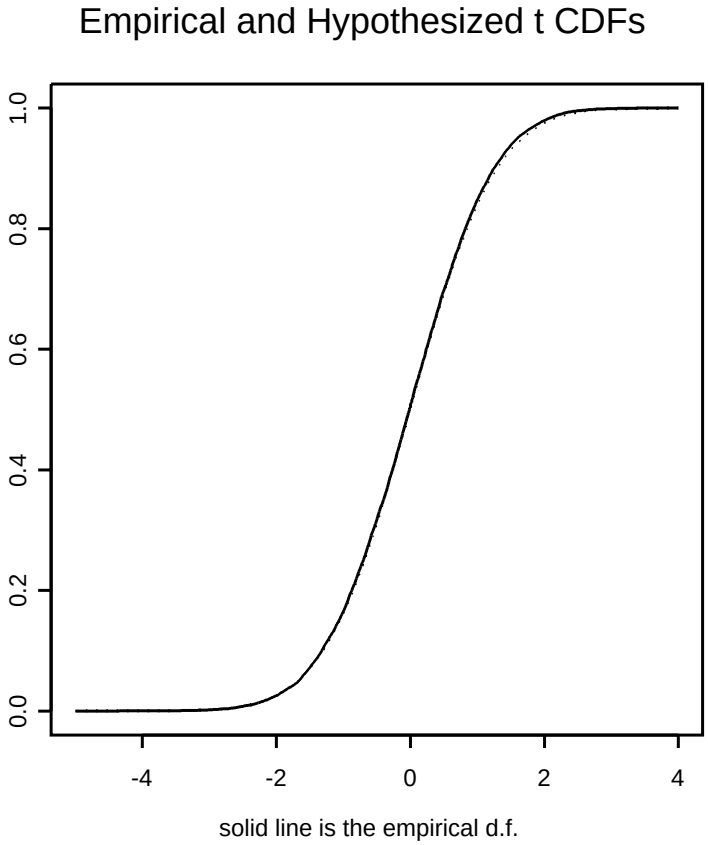
```
> summary(tstats)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 -4.34   -0.70  -0.02  -0.03   0.66    3.38
```

so the observed difference is very unlikely to have occurred by chance.

Note: this is not evidence of sex discrimination, only of association between gender and salary. The male clerks might have been older and more experienced, for example.

# 10000 randomizations of gender



# Robustness of the t statistics

We have demonstrated that the sample mean can be quite close to normality even if the original population is not too close to normal (and even if it is rather discrete). In general the 't tools' work well unless outliers upset the calculation of $s$.

Nevertheless, alternatives have been developed, for *continuous* distributions.

# Tests of location zero

These apply to data such as the differences in tuberculin reaction, and the null hypothesis is that the data have a symmetric continuous distribution about zero (and so zero will not occur).

## sign test

This just looks at the number of positive signs, which under the null hypothesis has a binomial distribution like coin-tossing.

Our example has 61 zeros.

# Wilcoxon signed-rank test

This uses the absolute values of the differences, and ranks them. The statistic is then the sum of the ranks for the differences for which the sign is positive.

For small samples without ties we use the permutation distribution, otherwise a large-sample normal approximation.

As a refinement, zero differences are dropped completely.

```
> wilcox.test(react, exact=F)

        Wilcoxon signed-rank test

signed-rank normal statistic with correction Z = -7.55, p-value = 0
alternative hypothesis: true mu is not equal to 0
> wilcox.test(react[react!=0], exact=F)

signed-rank normal statistic with correction Z = -7.34, p-value = 0
```

# Two-sample non-parametric tests

The Mann–Whitney or Wilcoxon rank-sum test is for location shifts between two samples.

First rank all the observations, assigning average ranks to the ties. Then sum the ranks for the observations in the first sample.

For small samples without ties we use the permutation distribution, otherwise a large-sample normal approximation.

```
> wilcox.test(bankM, bankF, exact=F, alternative="greater")

        Wilcoxon rank-sum test

rank-sum normal statistic with correction Z = 5.11, p-value = 0
alternative hypothesis: true mu is greater than 0
```

# Another two-sample example

Twenty-eight 14-year-old pupils were given a problem in coordinate geometry. They were divided into two groups of 14 and given different material to read, and timed (in seconds) as to how long it took them to solve the problem.

```
new   68  70  73  75  77  80  80 132 148 155 183 197 206 210
old  130 139 146 150 161 177 228 242 265 >300 >300 >300 >300 >300
```

Even though we don't know the exact times we can still do a Mann-Whitney test. If we jitter the data to break the ties we get

```
> wilcox.test(times[1:14], times[15:28], alternative = "less")

        Exact Wilcoxon rank-sum test

rank-sum statistic W = 137, n = 14, m = 14, p-value = 0.0009
```