STAT 210
Applied Statistics and Data Analysis
One Sample Problems II

Joaquín Ortega
KAUST

Fall 2020

Example: The Speed of Light

Recap

Recall that we are considering Michelson's experiment to measure the speed of light.

We want to determine whether Michelson's results are in agreement with the accepted value (at the time) for the speed of light.

We did a first test assuming that the sample came from a normal distribution, using the sample variance in place of the (unknown) population variance.

Second Alternative: Student's t distribution

Second Alternative: Student's t distribution

The t distribution arises as the sampling distribution of the (empirical) mean $\hat{\mu}_n$ when the data come from a normal distribution with unknown variance.

We saw that if $\hat{\mu}_n = \frac{1}{n}\sum_1^n X_i$ and the X_i are iid with $N(\mu, \sigma^2)$ distribution then

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

In practice, we hardly ever know the true value for the variance σ^2 and we must estimate it by the empirical variance s_n^2:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

Student's t distribution

It was shown by W.S. Gosset in a paper in Biometrika published in 1908 under the pseudonym *Student* that

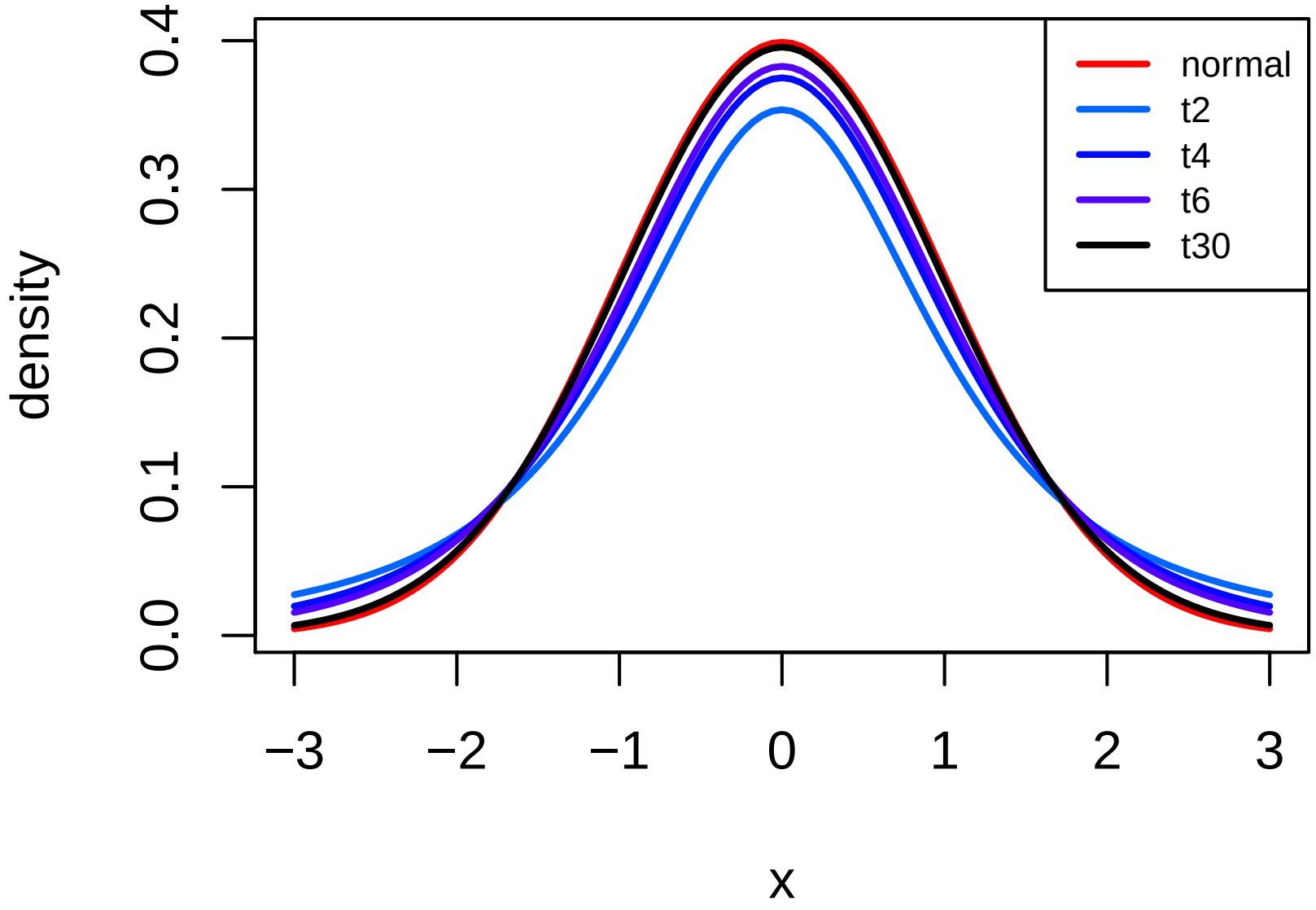
$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom.

For $n \geq 30$ the t distribution is very similar to the normal distribution but for n small there are important differences.

The t distribution has 'heavier' tails than the normal, which means that large values are more probable.

Student's t distribution



Student's t distribution

Now that we know the correct distribution for the sample mean with unknown variance (under the hypothesis that the data are normal), we can do a more precise test.

We want to test

$$H_0 : \mu = 990 \qquad vs. \qquad H_A : \mu \neq 990$$

and our *test statistic* is

$$\bar{x}_n = \hat{\mu}_n = \frac{1}{n} \sum_{1}^{n} x_i$$

where the x_i are the observed values.

Our null hypothesis specifies a single value for the mean of the distribution: $\mu = 990$.

Student's t distribution

Thus, *under the null hypothesis*, i.e., assuming the null hypothesis is true, we have that

$$\frac{\hat{\mu}_n - 990}{s_n/\sqrt{n}} \sim t_{n-1}.$$

Recall that we have a sample of size $n = 20$ and the standard deviation is

```
sd(mich.exp1)
```

```
## [1] 104.926
```

```
sd(mich.exp1)/sqrt(20)
```

```
## [1] 23.46218
```

Student's t distribution

Thus

$$\frac{\hat{\mu}_n - 990}{23.46} \sim t_{19}.$$

We observed $\hat{\mu}_n = 909$. How likely is this under this distribution?

We want $P(|\hat{\mu}_n - 990| \geq 81)$ but

$$\begin{aligned} P(|\hat{\mu}_n - 990| \geq 81) &= P(\frac{|\hat{\mu}_n - 990|}{23.46} \geq \frac{81}{23.46}) \\ &= P(|t_{19}| \geq 3.542) \\ &= 2P(t_{19} \leq -3.542) \end{aligned}$$

where the last equality follows from the symmetry of the t distribution.

Student's t distribution

To calculate this value, use the distribution function for the t distribution in R, given by pt:

```
2*pt(-3.452,df=19)
```

```
## [1] 0.002670794
```

which is bigger than what we obtained before assuming a normal distribution and using the estimated standard deviation (0.00056).

The odds now are about one in 375.

The reason for this is that, for small samples, you do not expect the estimated variance to be accurate, and you will have more variability in your sample. Hence you need a distribution that makes having larger values more likely, which is the t distribution in this case.

Student's t distribution

It is not necessary to do all the calculations every time we want to do a t test. The function `t.test` in `R` will do this.

```
t.test(mich.exp1, mu=990)
```

```
##
##  One Sample t-test
##
## data:  mich.exp1
## t = -3.4524, df = 19, p-value = 0.002669
## alternative hypothesis: true mean is not equal to 990
## 95 percent confidence interval:
##  859.8931 958.1069
## sample estimates:
## mean of x
##       909
```

Third Alternative: Non-parametric test

Third Alternative: Non-parametric test

So far, we have used the hypothesis of normality as the basis to build our tests.

What if we don't want to make this assumption?

Some tests are **distribution-free**, i.e., they do not make distributional assumptions.

They are known as **non-parametric tests** because they are not based on the assumption of a parametric family of distributions.

For the one-sample problem we are considering, the most popular choice is known as Wilcoxon's signed-ranks test.

Non-parametric test

Many non-parametric methods are based on **order statistics** and **ranks**.

Assume you have a sample $x_1, x_2, \ldots, x_n$ and that all values are different. The **order statistics** for this sample are the ordered values:

$$x_{(1)} < x_{(2)} < \cdots < x_{(n-1)} < x_{(n)}$$

The **rank** is the position that a particular value has in the ordered sample.

Non-parametric test

Example

Draw a sample of size 5 from the uniform distribution in (-1,1) and find the order statistics and the ranks:

```
(unif.spl <- runif(5,-1,1))
```

```
## [1] -0.8261034 -0.1073305 -0.8853697  0.9386212
## [5]  0.9507089
```

```
sort(unif.spl)
```

```
## [1] -0.8853697 -0.8261034 -0.1073305  0.9386212
## [5]  0.9507089
```

```
rank(unif.spl)
```

```
## [1] 2 3 1 4 5
```

Non-parametric test

Assume that your sample comes from a continuous distribution that is symmetric with respect to its average value μ.

We want to test the (null) hypothesis $H_0 : \mu = \mu_0$ versus the alternative that this is false.

By symmetry, it is equally likely that values will be above or below the mean.

Also, positive and negative differences of the same magnitude have the same probability of occurring.

Non-parametric test

If the sample values are $x_1, \ldots, x_n$ define

$$d_i = x_i - \mu_0$$

To compute the test statistic, follow these steps:

1. Take the absolute value of the d_i's.
2. Order these values and assign the ranks. If there are ties, use the midranks (average rank of the tied values).
3. Multiply the rank values obtained in step 2 by the original signs of the d_i's.
4. Sum the positive values in step 3 and denote the result by t^+.

Non-parametric test

t^+ is a sum of **ranks**, not of values, and is (the value of) the test statistic (we denote the corresponding r. v. by T^+).

T^- is defined similarly, but in fact, we only need one of them to carry out the test.

Non-parametric test

If there are n values in the sample, the possible values of T^+ are between 0 (if all the values in the sample are negative) and

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

if all values are positive.

Also,

$$T^+ + T^- = \frac{n(n+1)}{2}$$

and this is why we only need one of these random variables for the test.

Non-parametric test

If the hypothesis of symmetry is valid, we would not expect very small or very large values of T^+.

If we observe either of these situations, we will reject the null hypothesis.

The distribution of T^+ is challenging to calculate, particularly if there are ties in the sample.

There is a normal approximation to the distribution that is frequently helpful.

The test can be carried out in `R` with the command `wilcox.test`.

Non-parametric test

```
wilcox.test(mich.exp1,mu=990)
```

```
## Warning in wilcox.test.default(mich.exp1, mu =
## 990): cannot compute exact p-value with ties

##
##  Wilcoxon signed rank test with continuity
##  correction
##
## data:  mich.exp1
## V = 22.5, p-value = 0.00213
## alternative hypothesis: true location is not equal to 990
```

Non-parametric test

```
mich.dif <- mich.exp1-990
sort(mich.dif)
```

```
##  [1] -340 -250 -230 -180 -140 -140 -110  -90  -60
## [10]  -60  -40  -30  -30  -10  -10  -10   10   10
## [19]   10   80
```

```
length((mich.dif)[mich.dif<0])
```

```
## [1] 16
```

```
rank(abs(mich.dif))
```

```
##  [1] 15.5 19.0 13.0 12.0 10.5 15.5  9.0  3.5  3.5
## [10] 14.0  3.5  3.5 10.5 20.0 18.0 17.0  3.5  3.5
## [19]  7.5  7.5
```

Non-parametric test

```
mich.signrank <- rank(abs(mich.dif))*sign(mich.dif)
sum(mich.signrank[mich.signrank>0])
```

```
## [1] 22.5
```