STAT 210
Applied Statistics and Data Analysis
One Sample Problems I

Joaquín Ortega
KAUST

Fall 2020

Example: The Speed of Light

Example

In 1879 Albert Michelson carried out a series of experiments to measure the speed of light.

Some of the experiments (but not the one we are going to consider) were made together with Edward Morley, and they have become known as the Michelson-Morley experiments. Michelson went on to win the Nobel prize in Physics in 1907.

At the time, the 'accepted' value for the speed of light was 299990 km/s.

The file `michelson` in the `MASS` package has the data for this experiment. It is also stored as `morley` in the `base` package.

There were five experiments, each consisting of 20 different runs. The response is the speed of light measurement (with 299000 subtracted) in km/s.

Data

```
library(MASS)
data(michelson)
str(michelson, vec.len = 2)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Speed: int  850 740 900 1070 930 ...
##  $ Run  : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 5 ...
##  $ Expt : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 ...
```

To see if these experiments produced a different value for the speed of light, we need to compare the average of the speed from the data with the accepted value of 990 (recall we have subtracted 299000 from the results).

We will consider only the results for the 20 runs of the first experiment.

Data

We extract the values corresponding to the first experiment.

```r
(mich.exp1 <- michelson[michelson$Expt == 1,1])
```

```
##  [1]  850  740  900 1070  930  850  950  980  980
## [10]  880 1000  980  930  650  760  810 1000 1000
## [19]  960  960
```

The summary values for the data set are

```r
summary(mich.exp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     650     850     940     909     980    1070
```

Can the difference between accepted (990) and observed (909) be due to chance alone?

Data Analysis

Sample parameters

Suppose we have a sample $x_1, x_2, \ldots, x_n$. We assume that these values are independent (in the statistical sense) but, for the time being, make no assumptions about the distribution.

The mean of these values, denoted by $\bar{x}_n$ or $\hat{\mu}_n$ is given by

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

We can think of the values $x_1, x_2, \ldots, x_n$ as the *realization* of a collection of random variables $X_1, X_2, \ldots, X_n$ which represent the result of performing the i-th experiment.

We assume that these variables all have the same (unknown) distribution and are independent. We denote this by **iid** (independent and identically distributed).

Sample parameters

We also assume that the unknown common distribution for the X_is has a finite mean and variance, which we denote by

$$E(X) = \mu \qquad \text{and} \qquad Var(X) = \sigma^2.$$

If X has a continuous distribution with density $f(x)$ (as the normal distribution, for example) then

$$E(X) = \int x\, f(x)\, dx = \mu$$

and

$$Var(X) = E((X-\mu)^2) = E(X^2) - \mu^2 = \int x^2 f(x)\, dx - \mu^2.$$

Sample parameters

On the other hand, if X has a discrete distribution with values $a_1, a_2, \ldots$ and probability function $p_1, p_2, \ldots$ then

$$E(X) = \sum_{i \geq 1} a_i p_i$$

and

$$Var(X) = E((X - \mu)^2) = E(X^2) - \mu^2 = \sum_{i \geq 1} a_i^2 p_i - \mu^2$$

These are location and shape parameters for (almost) any distribution.

Sample parameters

We now have two means, the (unknown) theoretical or population mean μ, and the empirical or sample mean $\hat{\mu}_n$ that we can obtain from the sample.

An important result in Probability Theory, known as the Law of Large Numbers, says that as the sample size grows, the empirical mean will converge to the population mean:

$$\hat{\mu}_n \to \mu \quad \text{as } n \to \infty.$$

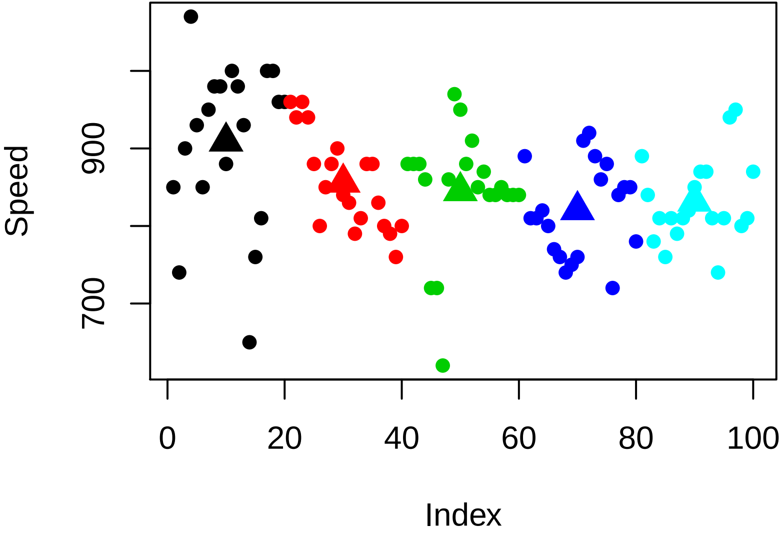
Therefore, it makes sense to use $\hat{\mu}_n$ to estimate μ.

Sample parameters

In our example, the ‘accepted’ value for the mean of the distribution was 990, and the sample mean obtained from Michelson’s experiment was 909. These are the quantities we want to compare.

Observe that $\hat{\mu}_n$ is also a random variable: it depends on the sample. If the experiments were to be repeated, we would obtain different values. Intuitively we would expect to see more variability in the sample than in their averages.

Sample parameters



Sample parameters

Since $\hat{\mu}_n$ is a random variable, we can consider its mean and variance:

$$E(\hat{\mu}_n) = E(\frac{1}{n}\sum_1^n X_i) = \frac{1}{n}\sum_1^n \mu = \mu.$$

So the mean value of $\hat{\mu}_n$ is μ. (Observe that here n is fixed).

On the other hand, the variance is (using independence)

$$Var(\hat{\mu}_n) = Var\Big(\frac{1}{n}\sum_1^n X_i\Big) = \frac{1}{n^2}\sum_1^n Var(X_i) = \frac{\sigma^2}{n}.$$

Therefore, the standard deviation of the sample mean decreases with the square root of the sample size.

Sample parameters

The standard deviation of the sample mean is known as the **standard error of the mean**.

$$se = \frac{\sigma}{\sqrt{n}}$$

Since this is a dispersion parameter, we would expect to have dispersion decrease as the sample size increases.

Sample parameters

Summing up, we have shown that for the sample mean

$$E(\hat{\mu}_n) = \mu$$

and

$$Var(\hat{\mu}_n) = \frac{\sigma^2}{n}$$

These results are always true, as long as the population distribution has finite mean and variance.

Sample parameters

How can we use this to compare Michelson's results with the accepted previous value?

Could the difference be attributed to randomness?

First Alternative

First Alternative

Let us assume that the speed of light measurements come from a Normal or Gaussian distribution.

Gaussian Distribution

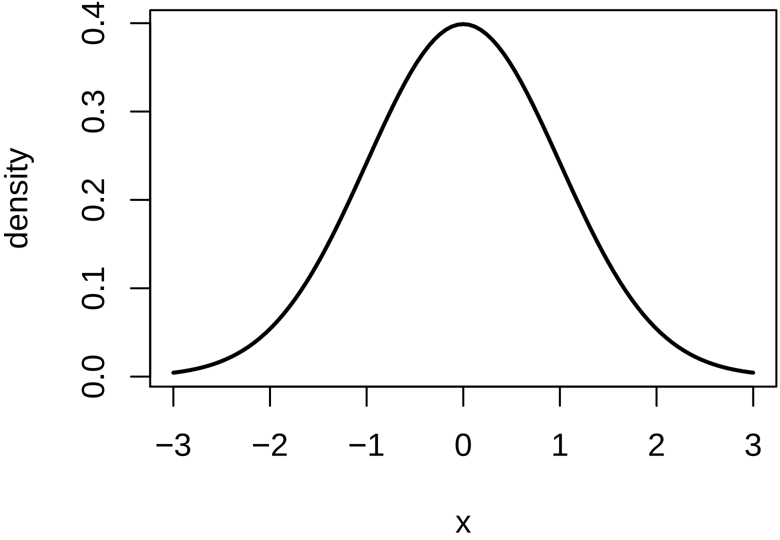
The Gaussian distribution plays a central role in Statistics. The standard Gaussian density is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It has mean 0 and variance 1.

Gaussian Distribution

```
curve(dnorm(x), -3,3,lwd=2, ylab = 'density')
```



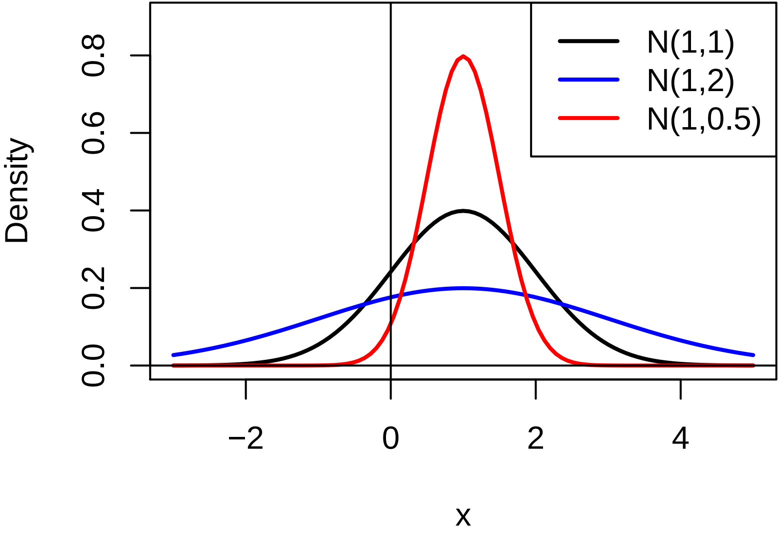
Gaussian Distribution

The Gaussian distribution with parameters μ and σ^2, denoted $N(\mu, \sigma^2)$ has density

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

The parameters are the mean μ and the variance σ^2

Gaussian Distribution



Gaussian Distribution

If $X \sim N(\mu, \sigma^2)$ then it is easy to see that

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

We say that

$$\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$$

is a parametric family of distributions with parameters

$$\mu \quad \text{and} \quad \sigma.$$

μ and σ are location and scale parameters.

Gaussian Distribution

If we assume that the measurements come from a normal distribution, then the average over a sample of size n has distribution

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

Now we can determine how likely it is for the difference between what Michelson observed and what was thought at the time to be the correct value to be due to chance.

Gaussian Distribution

Accepted value $\mu = 990$

Observed average $\hat{\mu}_n = 909$

Difference $= 81$

We want

$$P(|\hat{\mu}_n - \mu| \geq 81).$$

Gaussian Distribution

However, we are not ready yet to answer this question because the distribution of $\hat{\mu}_n$ depends on another (unknown) parameter: σ^2.

Solution: Estimate the variance from the sample and use it in the formula.

This is not a bad solution if n is large, again due to the Law of Large Numbers. In our case, $n = 20$, which is not too big.

The estimate for the variance is denoted s_n^2.

```r
var(mich.exp1); sd(mich.exp1)
```

```
## [1] 11009.47

## [1] 104.926
```

Gaussian Distribution

Using this value,

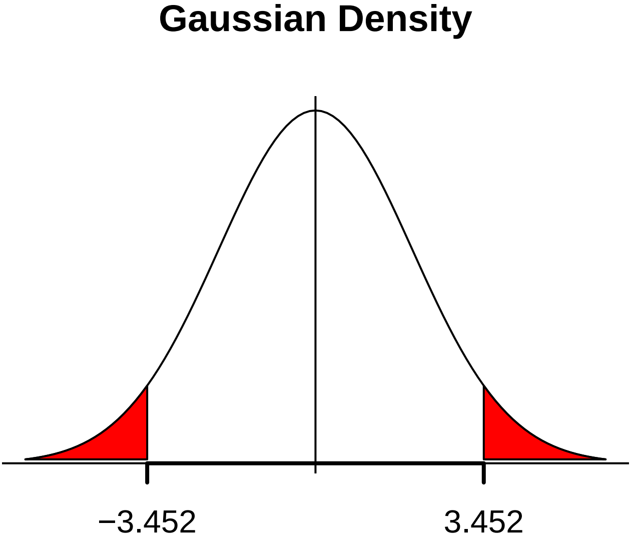
$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \approx N(0, 1).$$

We can calculate the probability we are interested in using `R`. Observe that

$$\begin{aligned} P(|\hat{\mu}_n - \mu| \geq 81) &= P(\frac{|\hat{\mu}_n - \mu|}{s_n/\sqrt{n}} \geq \frac{81}{s_n/\sqrt{n}}) \\ &= P(|Z| \geq 3.452) \\ &= 2P(Z \leq -3.452) \end{aligned}$$

where the last relation follows from the symmetry of the Gaussian distribution

Gaussian Distribution



Gaussian Distribution

```r
points.x <- seq(-3,3,length.out = 101)
points.y <- dnorm(points.x)
plot(points.x,points.y,type='l',xlab='',
     main='Gaussian Density')
xv <- points.x[points.x <= -1.74]
xv <-c(xv, -1.74, -3)
yv <- points.y[points.x <= -1.74]
yv <- c(yv,yv[1],yv[1])
polygon(xv,yv, col='red')
xw <- points.x[points.x >= 1.74]
xw <-c(xw, 1.74)
yw <- points.y[points.x >= 1.74]
yw <- c(yw,points.y[101])
polygon(xw,yw, col='red')
axis(1,at=c(-1.74,1.74),labels = c('-3.452','3.452'),
     line=0,pos=0,lwd=2)
```

```r
2*pnorm(-3.452)
```

```
## [1] 0.0005564477
```

Gaussian Distribution

Thus, the probability of observing a difference as large or larger than the one we found is about 0.00056.

The odds are about one in 1800.

This is a small number, so it would make us doubt the validity of the accepted value.

We have done a test of hypothesis. Assuming that the sample is Gaussian, we have tested the null hypothesis that the mean of the common distribution is 990 versus the alternative that it is not:

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

This test is frequently known as a z-test.

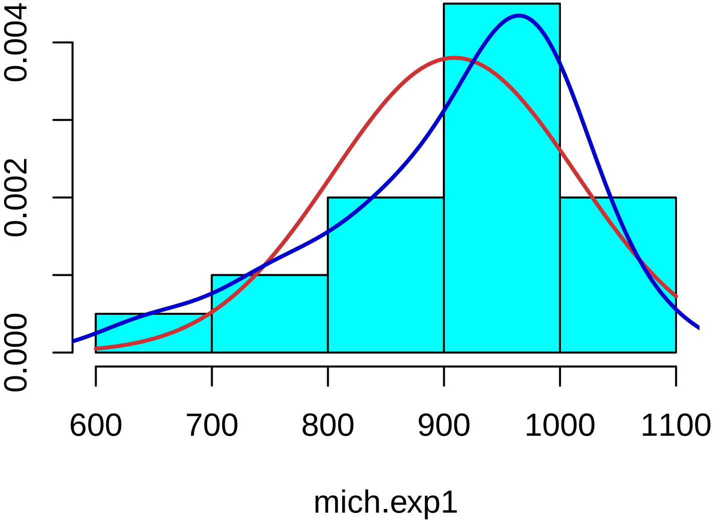
Gaussian Distribution

However, we made a fundamental assumption to carry out the previous calculations: We have assumed that our sample comes from a Gaussian distribution.

We need to ask ourselves whether this is a reasonable assumption.

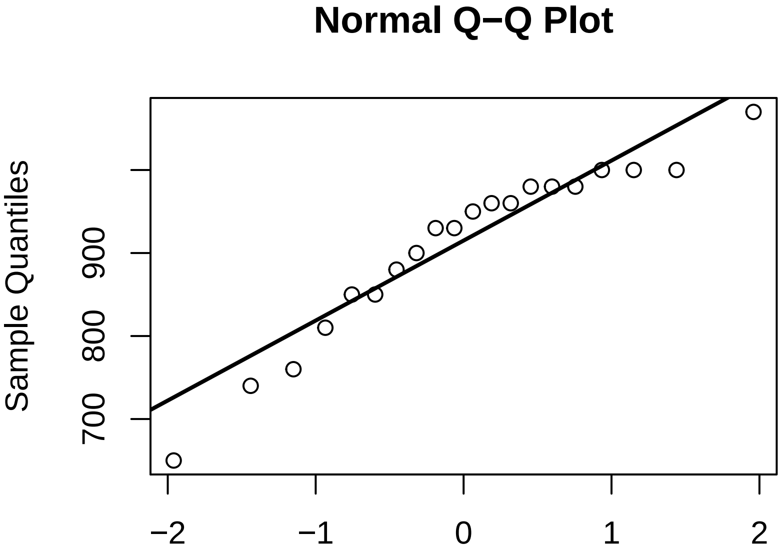
Gaussian Distribution

```
library(MASS); truehist(mich.exp1)
curve(dnorm(x,mean = mean(mich.exp1),sd = sd(mich.exp1)),
      600, 1100, add = TRUE, lwd=2, col='brown3')
lines(density(mich.exp1),col='blue3',lwd=2)
```



Gaussian Distribution

```
qqnorm(mich.exp1)
qqline(mich.exp1,lwd=2)
```



Gaussian Distribution

The qq plot does not look very good, but a test of normality does not reject the null hypothesis.

```r
shapiro.test(mich.exp1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mich.exp1
## W = 0.91992, p-value = 0.09876
```