Introduction to Statistical Inference

Edwin Leuven

Introduction

- Define key terms that are associated with inferential statistics.
- Revise concepts related to random variables, the sampling distribution and the Central Limit Theorem.

Introduction

Until now we've mostly dealt with descriptive statistics and with probability.

In descriptive statistics one investigates the characteristics of the data

- using graphical tools and numerical summaries

The frame of reference is the observed data

In probability, the frame of reference is all data sets that could have potentially emerged from a population

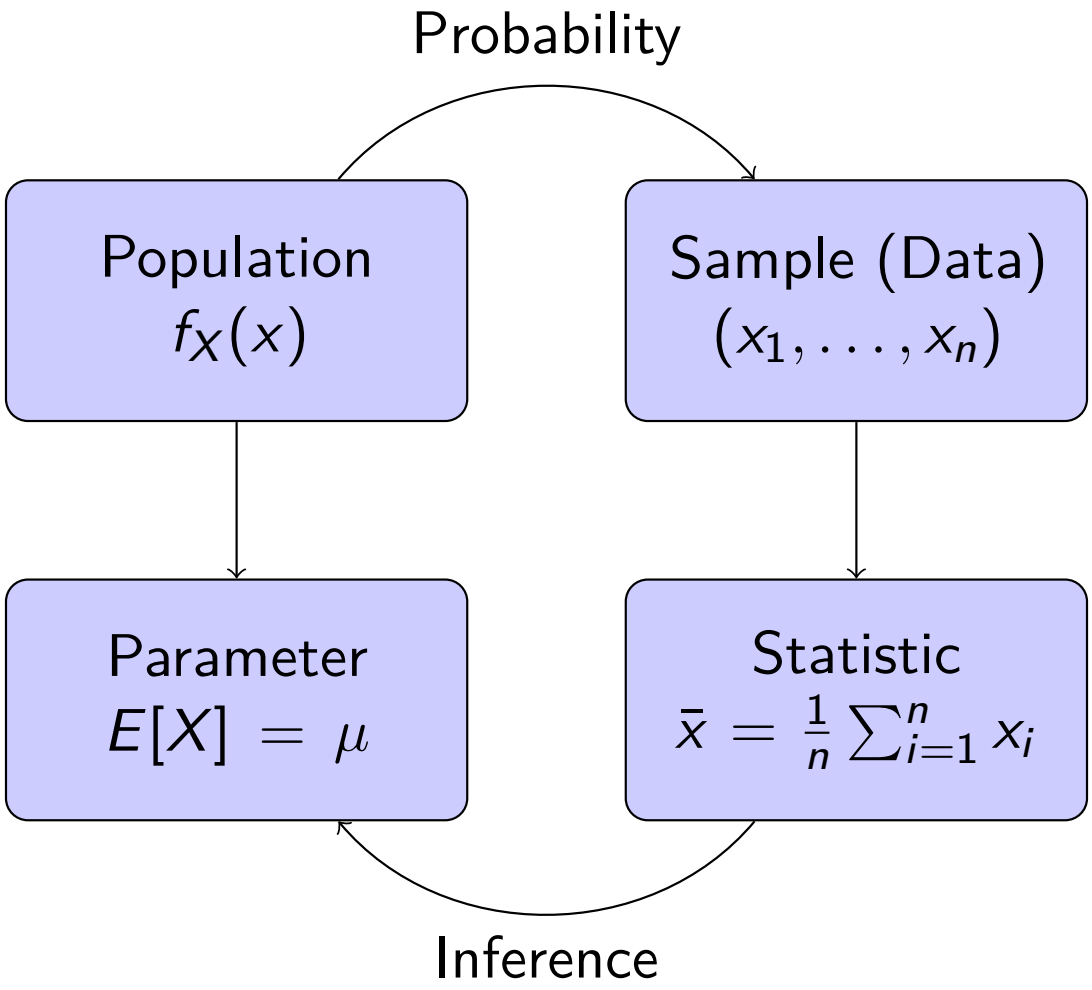
Introduction

The aim of *statistical inference* is to learn about the population using the observed data

This involves:

- computing something with the data
  - a statistic: function of data
- interpret the result
  - in probabilistic terms: sampling distribution of statistic

Introduction



Point estimation

We want to estimate a population parameter using the observed data.

- f.e. some measure of variation, an average, min, max, quantile, etc.

Point estimation attempts to obtain a best guess for the value of that parameter.

An *estimator* is a statistic (function of data) that produces such a guess.

We usually mean by "best" an estimator whose sampling distribution is more concentrated about the population parameter value compared to other estimators.

Hence, the choice of a specific statistic as an estimator depends on the probabilistic characteristics of this statistic in the context of the sampling distribution.

Confidence Interval

We can also quantify the uncertainty (sampling distribution) of our point estimate.

One way of doing this is by constructing an interval that is likely to contain the population parameter.

One such an interval, which is computed on the basis of the data, is called a *confidence interval*.

The sampling probability that the confidence interval will indeed contain the parameter value is called the *confidence level*.

We construct confidence intervals for a given confidence level.

Hypothesis Testing

The scientific paradigm involves the proposal of new theories that presumably provide a better description of the laws of Nature.

If the empirical evidence is inconsistent with the predictions of the old theory but not with those of the new theory

- then the old theory is rejected in favor of the new one.
- otherwise, the old theory maintains its status.

Statistical hypothesis testing is a formal method for determining which of the two hypothesis should prevail that uses this paradigm.

Statistical hypothesis testing

Each of the two hypothesis, the old and the new, predicts a different distribution for the empirical measurements.

In order to decide which of the distributions is more in tune with the data a statistic is computed.

This statistic t is called the *test statistic*.

A threshold c is set and the old theory is reject if $t > c$

Hypothesis testing consists in asking a binary question about the sampling distribution of t

Statistical hypothesis testing

This decision rule is not error proof, since the test statistic may fall by chance on the wrong side of the threshold.

Suppose we know the sampling distribution of the test statistic t

We can then set the probability of making an error to a given level by setting c

The probability of erroneously rejecting the currently accepted theory (the old one) is called the *significance level* of the test.

The threshold is selected in order to assure a small enough significance level.

Multiple measurements

The method of testing hypothesis is also applied in other practical settings where it is required to make decisions.

Consider a random trial of a new treatment to a medical condition where the

- treated get the new treatment
- controls get the old treatment

and measure their response

We now have 2 measurements that we can compare.

We will use statistical inference to make a decision about whether the new treatment is better.

Statistics

Statistical inferences, be it point estimation, confidence intervals, or hypothesis tests, are based on statistics computed from the data.

A *statistic* is a formula which is applied to the data

and we think of it as a statistical summary of the data

Examples of statistics are

- the sample average and
- the sample standard deviation

For a given dataset a statistic has a single numerical value.

it will be different for a different random sample!

The statistic is therefore a random variable

Statistics

It is important to distinguish between

1. the statistic (a random variable)
2. the realisation of the statistic for a given sample (a number)

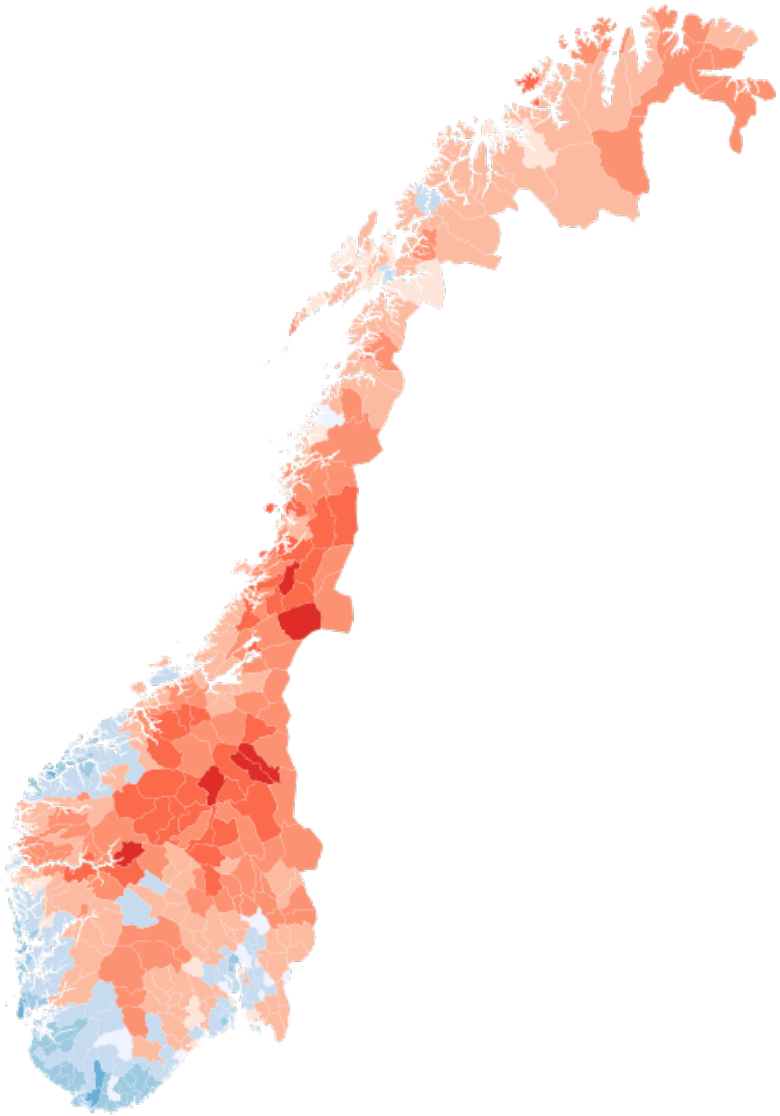
we therefore denote the statistic with capitals, f.e. the sample mean:

- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

and the realisation of the statistic with small letters:

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Example: Polling



Example: Polling

Imagine we want to predict whether the left block or the right block will get a majority in parliament

Key quantities:

- $N = 4{,}166{,}612$ - Population
- $p = (\#$ people who support the right) / N
- $1 - p = (\#$ people who support the left) / N

We can ask the following questions:

1. What is p?
2. Is $p > 0.5$?
3. We estimate p but are we sure?

Example: Polling

We poll a random sample of n = 1,000 people from the population *without replacement*:

- choose person 1 at random from N, choose person 2 at random from N-1 remaining, etc.
- or, choose a random set of n people from all $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible sets

Let

$$X_i = \begin{cases} 1 & \text{if person } i \text{ support the right} \\ 0 & \text{if person } i \text{ support the left} \end{cases}$$

and denote our data by $x_1, \ldots, x_n$

Then we can estimate p by

$$\hat{p} = (x_1 + \ldots + x_n)/n$$

Example: Polling

To construct the poll we randomly sampled the population

With a random sample each of the n people is equally likely to be the ith person, therefore

$$E[X_i] = 1 \cdot p + 0 \cdot (1 - p) = p$$

and therefore

$$\begin{aligned} E[\hat{p}] &= E[(X_1 + \ldots + X_n)/n] \\ &= (E[X_1] + \ldots + E[X_n])/n = p \end{aligned}$$

The "average" value of $\hat{p}$ is p, and we say that $\hat{p}$ is *unbiased*

Unbiasedness refers to the average error over repeated sampling, and not the error for the observed data!

Example: Polling

Say 540 support the right, so $\hat{p} = 0.54$

Does this mean that in the population:

- $p = 0.54$?
- $p > 0.5$?

The data are a realization of a random sample and $\hat{p}$ is therefore a random variable!

For a given sample we will therefore have estimation error

$$\text{estimation error} = \hat{p} - p \neq 0$$

which comes from the difference between our sample and the population

Example: Polling

When sampling with replacement the X_i are independent, and

- $Var[\hat{p}] = \frac{p(1-p)}{n}$

When sampling without replacement the X_i are *not* independent

$$\frac{N_1 - 1}{N - 1} = \Pr(X_i = 1 | X_j = 1) \neq Pr(X_i = 1 | X_j = 0) = \frac{N_1}{N - 1}$$

and we can show that

- $Var[\hat{p}] = \frac{p(1-p)}{n}\left(1 - \frac{n-1}{N-1}\right)$

For $N = 4,166,612$, $n = 1,000$, and $p = 0.54$, the standard deviation of $\hat{p} \approx 0.016$.

But what is the distribution of $\hat{p}$?

The Sampling Distribution

Statistics vary from sample to sample

The sampling distribution of a statistic

- is the nature of this variability
- can sometimes be determined and often approximated

The distribution of the values we get when computing a statistic in (infinitely) many random samples is called the *sample distribution* of that statistic

The Sampling Distribution

We can sample from

- population
  - eligible voters in norway today
- model (theoretical population)
  - Pr(vote right block) $= p$

The sampling distribution of a statistic depends on the population distribution of values of the variables that are used to compute the statistic.

Sampling Distribution of Statistics

Theoretical models describe the distribution of a measurement as a function of one or more parameters.

For example,

- in n trials with succes probability p, the total number of successes follows a Binomial distribution with parameters n and p
- if an event happens at rate λ per unit time then the probability that k events occur in a time interval with length t follows a Poission distribution with parameters λt and k

Sampling Distribution of Statistics

More generally the sampling distribution of a statistic depends on

- the sample size
- the sampling distribution of the data used to construct the statistic

can be complicated!

We can sometimes learn about the sampling distribution of a statistic by

- Deriving the finite sample distribution
- Approximation with a Normal distribution in large samples
- Approximation through numerical simulation

Finite sample distributions

Sometimes we can derive the finite sample distribution of a statistic

Let the fraction of people voting right in the population be p

Because we know the distribution of the data (up to the unknown parameter p) we can derive the sampling distribution

In a random sample of size n the probability of observing k people voting on the right can be derived and follows a binomial distribution

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This depends on p which is unknown.

This approach is however often not feasible

The statistic may be complicated, depend on different variables, the population distribution of these variables is unknown

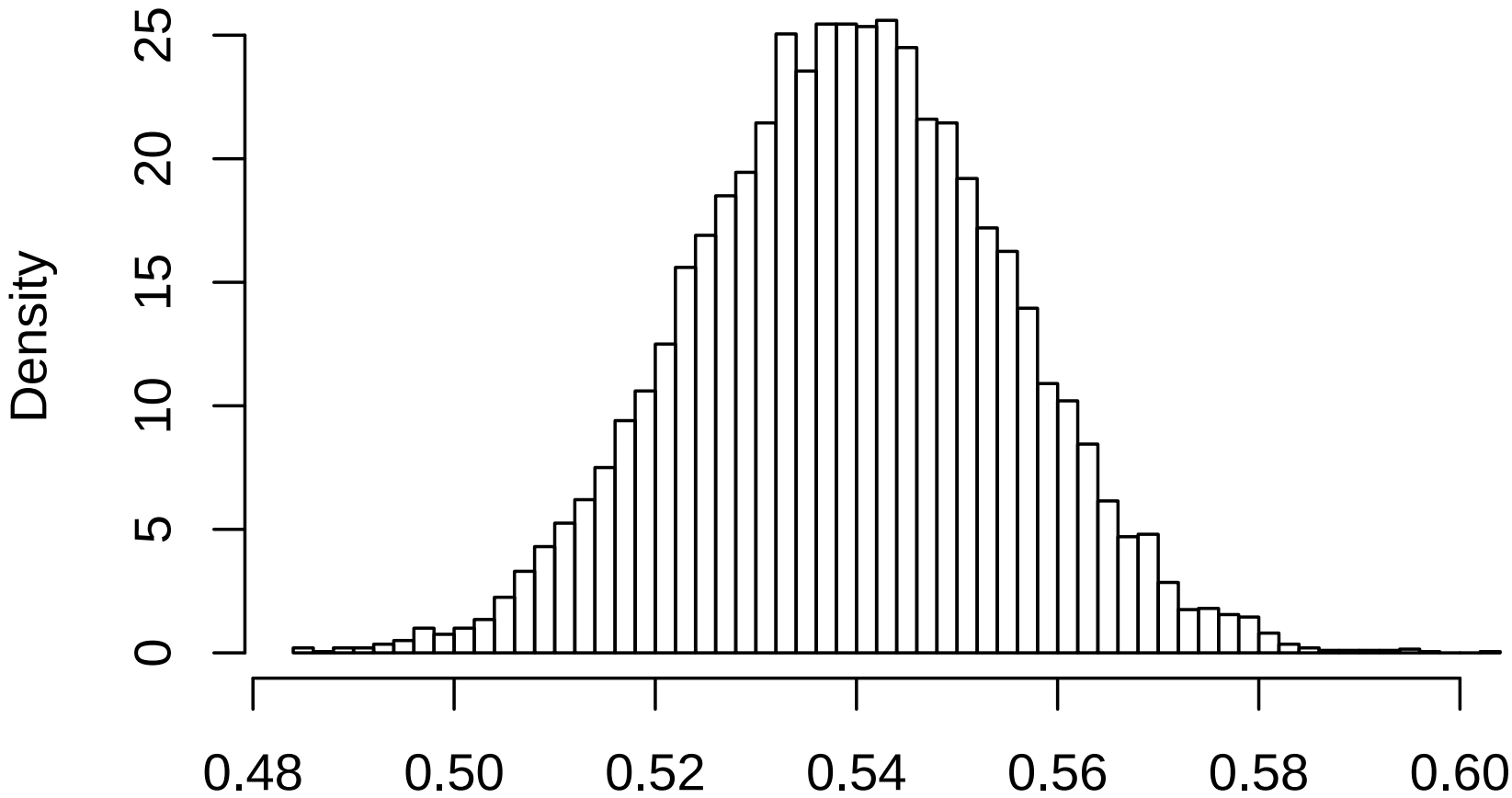
Theoretical Distributions of Observations (Models)

Distribution	Sample Space	$f(x)$
Binomial	$1, \ldots, n$	$\binom{n}{k} p^k (1-p)^{n-k}$
Poisson	$1, 2, \ldots$	$\lambda^k \exp(-\lambda)/k!$
Uniform	$[a, b]$	$1/(b-a)$
Exponential	$[0, \infty)$	$\lambda \exp(-\lambda x)$
Normal	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$

Distribution	$E[X]$	$Var(X)$	R
Binomial	np	$np(1-p)$	d,p,q,rbinom
Poisson	λ	λ	d,p,q,rpoisson
Uniform	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	d,p,q,runif
Exponential	λ^{-1}	λ^{-2}	d,p,q,rexp
Normal	μ	σ^2	d,p,q,rnorm

Example: Polling

```
hist(
  replicate(
    10000,mean(rbinom(1000, 1, .54)))
  , main="", xlab="p_hat",prob=TRUE,breaks=50)
```



The Normal Approximation

In general, the sampling distribution of a statistic is not the same as the sampling distribution of the measurements from which it is computed.

If the statistic is

1. (a function of) a sample average and
2. the sample is large

then we can often approximate the sampling distribution with a Normal distribution

Example: Polling

In the graph $\hat{p}$ looked like it had a Normal distribution with mean 0.54 and s.d. 0.16

If $N \gg n$ then X_i are approximately independent, and if n is large then

$$\sqrt{n}(\hat{p} - p) \sim N(0, p(1-p))$$

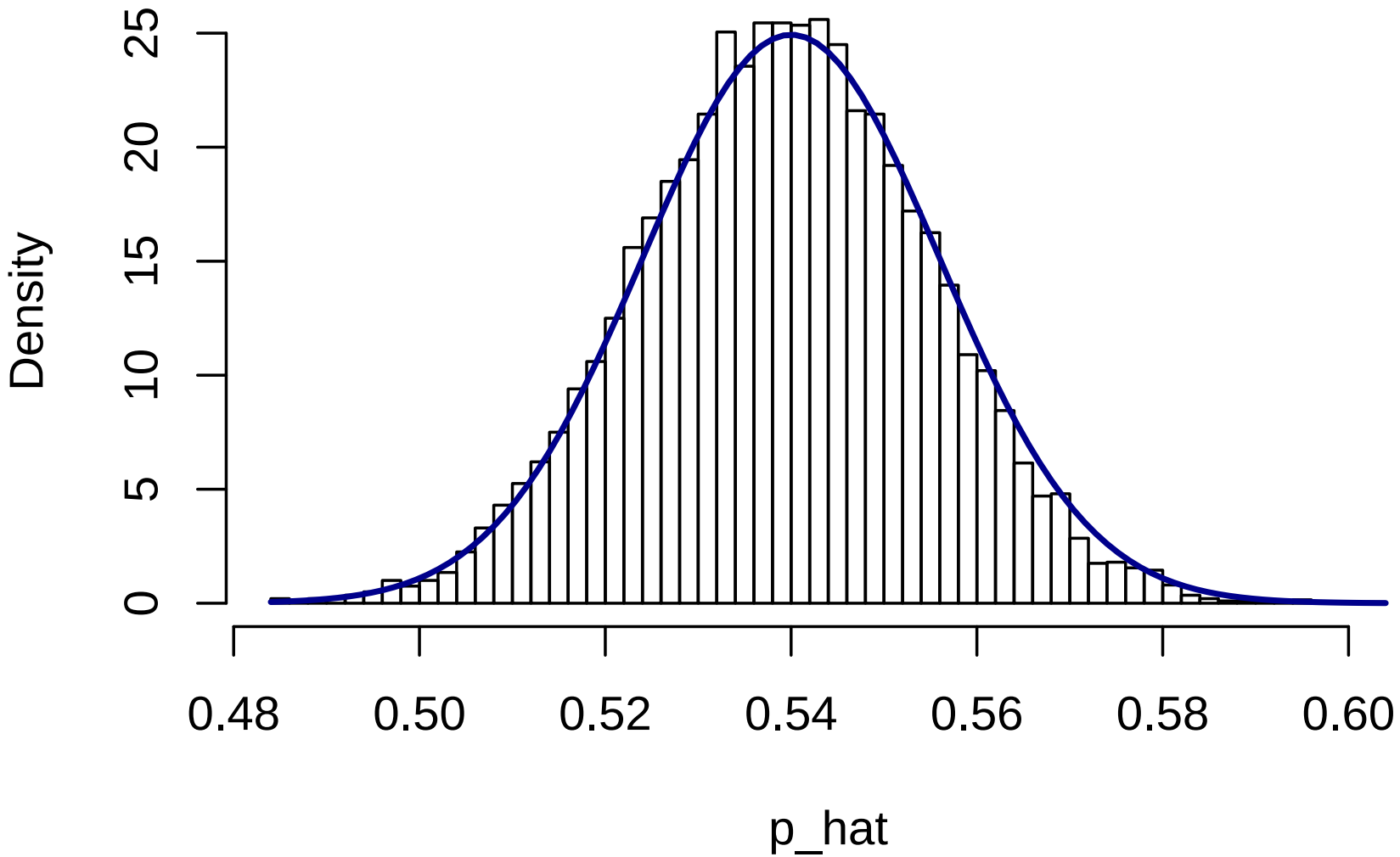
or equivalently

$$\hat{p} \sim N(p, \frac{p(1-p)}{n})$$

by the Central Limit Theorem

Example: Polling

```
curve(dnorm(x, mean=.54, sd=0.016),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



Approximation through numerical simulation

Computerized simulations can be carried out to approximate sampling distributions.

With a model we can draw many random samples, compute the statistic, and characterize it's sampling distribution.

Assume $price \sim Expontential(\lambda)$

Consider samples of size $n = 201$

$E[price] = \lambda^{-1}$ and $Var[price] = \lambda^{-2}$

and therefore

$$Var(\overline{price}) = \sqrt{(1/\lambda^2)/201} \approx 0.0705/\lambda$$

Approximation through numerical simulation

Remember that 95% of the probability density of a normal distribution is within 1.96 s.d. of its mean.

The Normal approximation for the sampling distribution of the average price suggests

$$1/\lambda \pm 1.96 \cdot 1/(\lambda\sqrt{n})$$

should contain 95% of the distribution.

Approximation through numerical simulation

We may use simulations in order to validate this approximation

Assume $\lambda = 1/12,000$

```
X.bar = replicate(10^5,mean(rexp(201,1/12000)))
mean(abs(X.bar-12000) <= 1.96*0.0705*12000)
```

```
## [1] 0.95173
```

Which shows that the Normal approximation is adequate in this example

How about other values of n or λ?

Approximation through numerical simulation

Simulations may also be used in order to compute probabilities in cases where the Normal approximation does not hold.

Consider the following statistic

$$(\min(x_i) + \max(x_i))/2$$

where $X_i \sim \mathrm{Uniform}(3, 7)$ and $n = 100$

What interval contains 95% of the observations?

Approximation through numerical simulation

Let us carry out the simulation that produces an approximation of the central region that contains 95% of the sampling distribution of the mid-range statistic for the Uniform distribution:

```
mid.range <- rep(0,10^5)
for(i in 1:10^5) {
  X <- runif(100,3,7)
  mid.range[i] <- (max(X)+min(X))/2
}
quantile(mid.range,c(0.025,0.975))
```

```
##      2.5%     97.5%
## 4.9409107 5.0591218
```

Observe that (approximately) 95% of the sampling distribution of the statistic are in the range $[4.941680, 5.059004]$.

Approximation through numerical simulation

Simulations can be used in order to compute any numerical summary of the sampling distribution of a statistic

To obtain the expectation and the standard deviation of the mid-range statistic of a sample of 100 observations from the $\mathrm{Uniform}(3, 7)$ distribution:

```
mean(mid.range)
```

```
## [1] 4.9998949
```

```
sd(mid.range)
```

```
## [1] 0.027876151
```

Approximation through numerical simulation

Computerized simulations can be carried out to approximate sampling distributions.

1. draw a random sample of size n with replacement from our data
2. compute our statistic
3. do 1. & 2. many times

The resulting distribution of statistics across our resamples is an approximation of the sampling distribution of our statistic

The idea is that a random sample of a random sample from the population, is again a random sample of the population

This is called *the bootstrap* and computes the sampling distribution without a model!

Approximation through numerical simulation

```
n = 1000
data = rbinom(n, 1, .54) # true distr, usually unknown
estimates = rep(0,999)
for(i in 1:999) {
  id = sample(1:n, n, replace=T)
  estimates[i] = mean(data[id])
}
sd(estimates)
```

```
## [1] 0.015946413
```

```
sqrt(.54*(1-.54)/1000)  # true value, usually unknown
```

```
## [1] 0.015760711
```

Summary

Today we looked at the elements of statistical inference

- Estimation:
  - Determining the distribution, or some characteristic of it. (What is our best guess for p?)
- Confidence intervals:
  - Quantifying the uncertainty of our estimate. (What is a range of values to which we're reasonably sure p belongs?)
- Hypothesis testing:
  - Asking a binary question about the distribution. (Is $p > 0.5$?)

Summary

In statistical inference we think of data as a realization of a random process

There are many reasons why we think of our data as (ex-ante) random:

1. We introduced randomness in our data collection (random sampling, or random assigning treatment)
2. We are actually studying a random phenomenon (coin tosses or dice rolls)
3. We treat as random the part of our data that we don't understand (errors in measurements)

The coming weeks we will take a closer look at how this randomness affects what we can learn about the population from the data