Statistical methods for linguistic research: Foundational Ideas

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

August 3, 2015

What this course is about

1. In this course, I aim to provide a first encounter with the foundational ideas in frequentist statistical theory.
2. We will cover what I consider the absolute minimum you need to know before you do any data analysis.
3. We will mainly use simulation to understand the key concepts, and this requires some knowledge of the language R.
4. So download and install R from: http://cran.r-project.org/
5. Why R? Why not Excel, or SPSS? R is free, and is more powerful than any of these alternatives.

Why I prepared this course

1. A poor understanding of statistical theory has been causing a lot of confusion in psychology, linguistics, and other areas.
2. My hope is that, with a little study, we can stop making the same mistakes over and over again.

Why I prepared this course

1. The meaning of the p-value is widely misunderstood, and the p-value is grossly abused, leading to invalid claims.
2. Even experienced scientists can't give a correct definition of some basic concepts in statistics.
3. There is a very strong tendency to use statistical models as a black box for delivering p-values to make decisions about "significance".
4. Beginning users of statistics look for "guidelines" from experts; but guidelines usually distort the details.
5. There is no substitute for judgement tempered with understanding; no one-size-fits-all solution.

Why I prepared this course

At a minimum, I hope that you will realize that the following binary decision strategy is not the purpose of statistical data analysis:

1. Get $p < 0.05 \rightarrow$ decide that your favored hypothesis is true
2. Get $p > 0.05 \rightarrow$ decide that the null hypothesis is true

Prerequisites for this course

1. For this course, I only assume that you are willing to put in some work on your own.
2. This means reading the lecture notes, and doing the homework.
3. A certain amount of fearlessness is also assumed.
4. Only pre-calculus math and high school algebra is assumed, but if you know some calculus (perhaps a passive knowledge?), you will get more out of this course.

Course lecture notes (optional during ESSLLI)

I have two sets of notes. Choose the one you like more.

1. A less technical, more intuitive presentation: https://github.com/vasishth/Statistics-lecture-notes-Potsdam/tree/master/IntroductoryStatistics
2. A more technical presentation assuming basic calculus and linear algebra: https://github.com/vasishth/LM
3. During ESSLLI, it is enough to just follow my slides. Read the notes later, after you go home!

I suggest reading the first set of notes and then the second set.

Self-paced exercises

You can do short auto-graded exercises each day on the datacamp course page:
https://www.datacamp.com/courses/statistical-methods-for-linguistic-research-foundational-ideas

What this course is about

We will cover the following topics:

1. Basic R constructs needed for this course. (But please do one of the datacamp intro courses or one of the online R introductions.)
2. Basic probability theory, random variables, including jointly distributed RVs, univariate probability distributions, Maximum Likelihood Estimation.
3. The sampling distribution of the mean, null hypothesis, t-tests, confidence intervals.
4. Type I error, Type II error, power, Type M and Type S errors.
5. An introduction to linear modeling and generalized linear models.
6. An introduction to linear mixed models.

for-loops

One construct we will use often is calculating some (varying) quantity repeatedly, and then storing the result of that calculation in a vector.
An example:

```
## number of iterations:
nsim<-10
## vector for storing results:
results<-rep(NA,10)
for(i in 1:nsim){
  results[i]<-1+2*i
}
results

##  [1]  3  5  7  9 11 13 15 17 19 21
```

The definition of a random variable

A random variable X is a function $X : S \to \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$.
S_X is all the x's (all the possible values of X, the support of X).
I.e., $x \in S_X$.
Discrete example: number of coin tosses till H

- $X : \omega \to x$
- ω: H, TH, TTH,... (infinite)
- $x = 0, 1, 2, \ldots; x \in S_X$

We will write $X(\omega) = x$:
$H \to 1$
$TH \to 2$
$\vdots$

Probability mass/distribution function

Every discrete random variable X has associated with it a **probability mass function (PMF)**. Continuous RVs have **probability distribution functions** (PDFs). We will call both PDFs (for simplicity).

$$p_X : S_X \to [0, 1] \tag{1}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{2}$$

This pmf tells us the probability of having getting a heads on 1, 2, . . . tosses.

The cumulative distribution function

The **cumulative distribution function** in the discrete case is

$$F(a) = \sum_{\text{all } x \leq a} p(x) \tag{3}$$

The cdf tells us the *cumulative* probability of getting a heads in 1 or less tosses; 2 or less tosses,. . . .
It will soon become clear why we need this.

Discrete example: The binomial random variable

Suppose that we toss a coin $n = 10$ times. There are two possible outcomes, success and failure, each with probability θ and $(1 - \theta)$ respectively.
Then, the probability of x successes out of n is defined by the pmf:

$$p_X(x) = P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \tag{4}$$

[assuming a binomial distribution]

Discrete example: The binomial random variable

Example: $n = 10$ coin tosses. Let the probability of success be $\theta = 0.5$.

We start by asking the question:

What's the probability of x or fewer successes, where x is some number between 0 and 10?

Let's compute this. We use the built-in CDF function `pbinom`.

Discrete example: The binomial random variable

```
## sample size
n<-10
## prob of success
p<-0.5
probs<-rep(NA,11)
for(x in 0:10){
  ## Cumulative Distribution Function:
probs[x+1]<-round(pbinom(x,size=n,prob=p),digits=2)
}
```

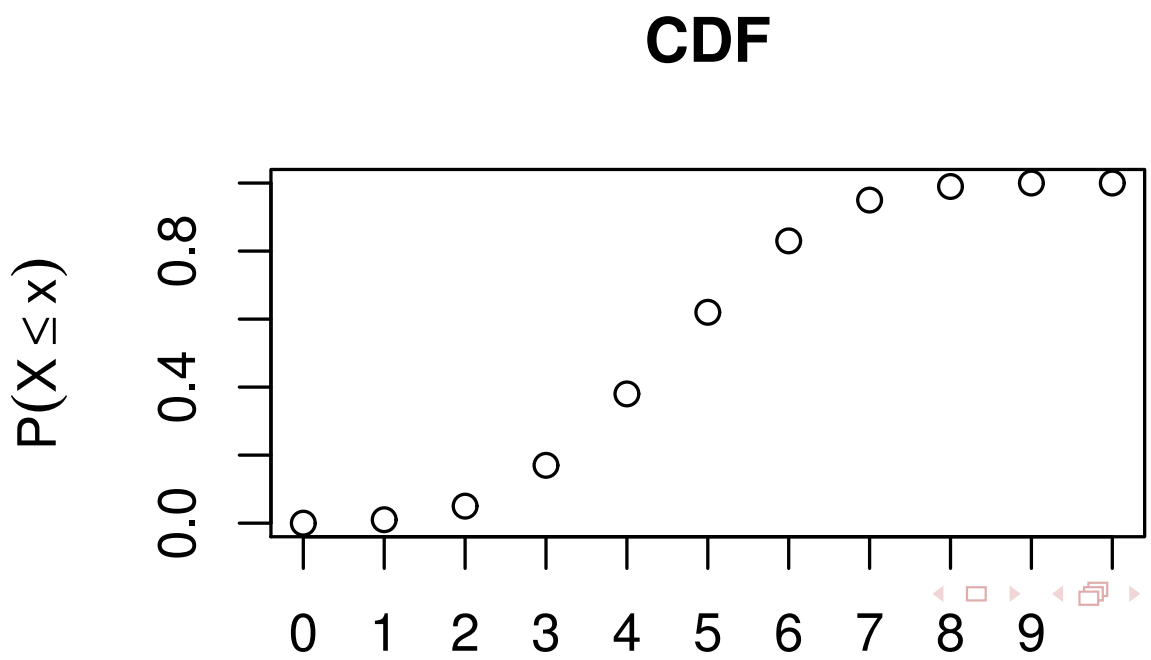
We have just computed the cdf of this random variable.

Discrete example: The binomial random variable

	$P(X \leq x)$	cumulative probability
1	0	0.00
2	1	0.01
3	2	0.05
4	3	0.17
5	4	0.38
6	5	0.62
7	6	0.83
8	7	0.95
9	8	0.99
10	9	1.00
11	10	1.00

Discrete example: The binomial random variable

```
## Plot the CDF:
plot(1:11,probs,xaxt="n",xlab="x",
     ylab=expression(P(X<=x)),main="CDF")
axis(1,at=1:11,labels=0:10)
```



Discrete example: The binomial random variable

Another question we can ask involves the pmf: What is the probability of getting exactly x successes? For example, if x=1, we want P(X=1).
We can get the answer from (a) the cdf, or (b) the pmf:

```
## using cdf:
pbinom(1,size=10,prob=0.5)-pbinom(0,size=10,prob=0.5)

## [1] 0.009765625

## using pmf:
choose(10,1) * 0.5 * (1-0.5)^9

## [1] 0.009765625
```

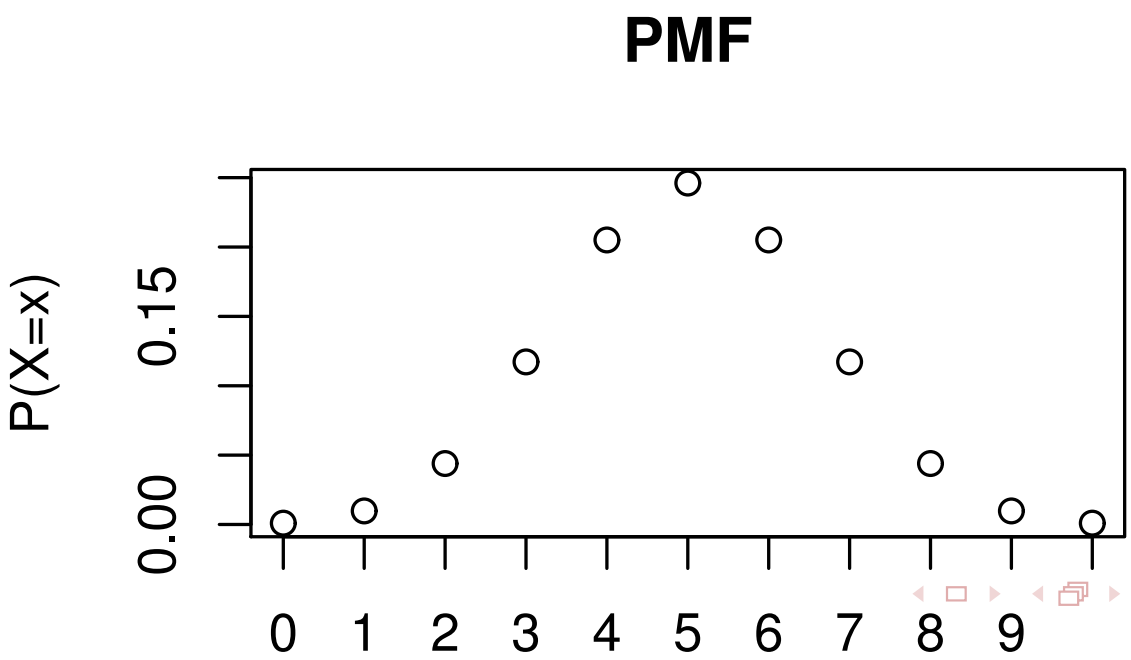
Discrete example: The binomial random variable

The built-in function in R for the pmf is `dbinom`:

```
## P(X=1)
choose(10,1) * 0.5 * (1-0.5)^9

## [1] 0.009765625

## using the built-in function:
dbinom(1,size=10,prob=0.5)

## [1] 0.009765625
```

Discrete example: The binomial random variable

```
## Plot the pmf:
plot(1:11,dbinom(0:10,size=10,prob=0.5),main="PMF",
     xaxt="n",ylab="P(X=x)",xlab="x")
axis(1,at=1:11,labels=0:10)
```



Summary: Random variables

To summarize, a discrete random variable X will be defined by

1. the function $X : S \to \mathbb{R}$, where S is the set of outcomes (i.e., outcomes are $\omega \in S$).
2. $X(\omega) = x$, and S_X is the **support** of X (i.e., $x \in S_X$).
3. A PMF is defined for X:

$$p_X : S_X \to [0, 1]$$

$$p_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{5}$$

4. A CDF is defined for X:

$$F(a) = \sum_{\text{all } x \leq a} p(x)$$

Continuous example: The normal random variable

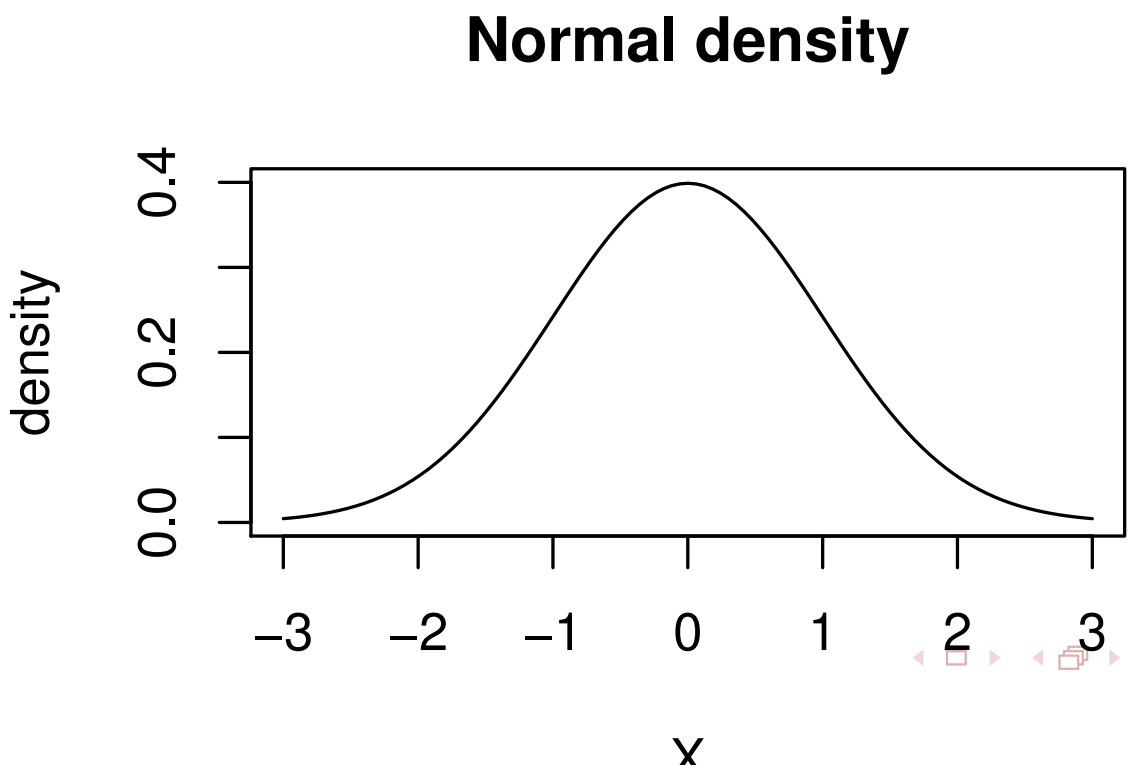
The pdf of the normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty \tag{6}$$

We write $X \sim \texttt{norm}(\texttt{mean} = \mu, \texttt{sd} = \sigma)$.
The associated `R` function for the pdf is `dnorm(x, mean = 0, sd = 1)`, and the one for cdf is `pnorm`.
Note the default values for μ and σ are 0 and 1 respectively. Note also that R defines the PDF in terms of μ and σ, not μ and σ^2 (σ^2 is the norm in statistics textbooks).

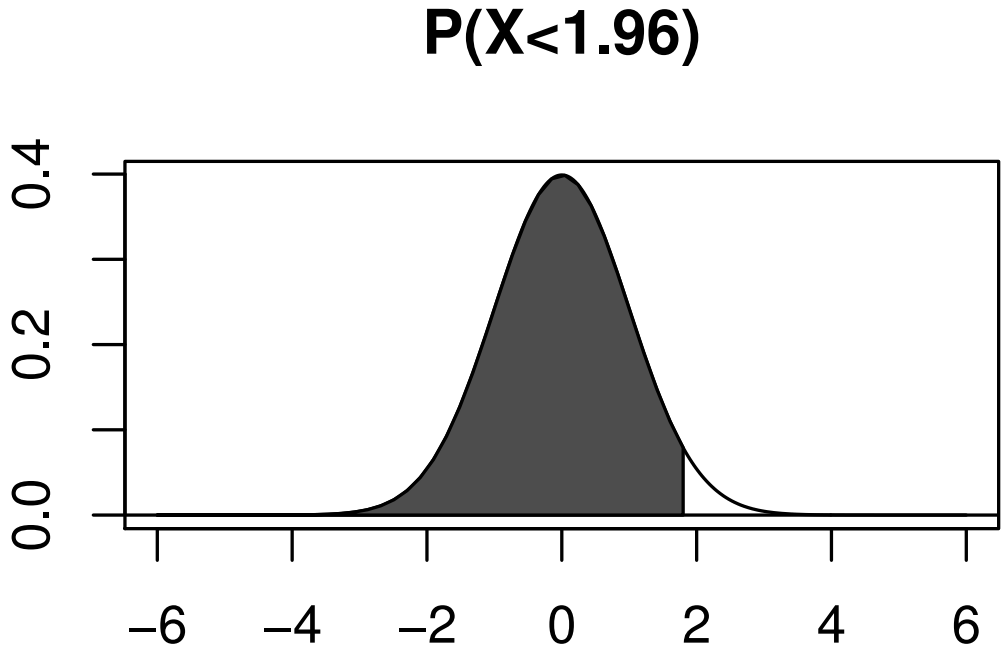
Continuous example: The normal RV

```
plot(function(x) dnorm(x), -3, 3,
      main = "Normal density",ylim=c(0,.4),
                ylab="density",xlab="X")
```



Probability: The area under the curve

Also see shiny app



Continuous example: The normal RV

Computing probabilities using the CDF:

```
## The area under curve between +infty and -infty:
pnorm(Inf)-pnorm(-Inf)

## [1] 1

## The area under curve between 2 and -2:
pnorm(2)-pnorm(-2)

## [1] 0.9544997

## The area under curve between 1 and -1:
pnorm(1)-pnorm(-1)

## [1] 0.6826895
```

Finding the quantile given the probability

We can also go in the other direction: given a probability p, we can find the quantile x of a $Normal(\mu, \sigma)$ such that $P(X < x) = p$.
For example:
The quantile x given $X \sim N(\mu = 500, \sigma = 100)$ such that $P(X < x) = 0.975$ is

```
qnorm(0.975,mean=500,sd=100)

## [1] 695.9964
```

This will turn out to be very useful in statistical inference.

Standard or unit normal random variable

If X is normally distributed with parameters μ and σ, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters $\mu = 0, \sigma = 1$.
We conventionally write $\Phi(x)$ for the CDF of N(0,1):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} \, dy \quad \text{where } y = (x - \mu)/\sigma \tag{7}$$

Standard or unit normal random variable

For example: $\Phi(2)$:

```
pnorm(2)

## [1] 0.9772499
```

For negative x we write:

$$\Phi(-x) = 1 - \Phi(x), \quad -\infty < x < \infty \tag{8}$$

See the shiny app for a visualization.

Standard or unit normal random variable

In R:

```
1-pnorm(2)

## [1] 0.02275013

## alternatively:
pnorm(2,lower.tail=F)

## [1] 0.02275013
```

Standard or unit normal random variable

If Z is a standard normal random variable (SNRV) then

$$p\{Z \leq -x\} = P\{Z > x\}, \quad -\infty < x < \infty \tag{9}$$

Since $Z = ((X - \mu)/\sigma)$ is an SNRV whenever X is normally distributed with parameters μ and σ, then the CDF of X can be expressed as:

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right) \tag{10}$$

The standardized version of a normal random variable X is used to compute specific probabilities relating to X.
We will soon see the relevance of the SNRV in hypothesis testing.

dnorm, pnorm, qnorm

1. For the normal distribution we have built in functions:
   1. dnorm: the pdf
   2. pnorm: the cdf
   3. qnorm: the inverse of the cdf
2. Other distributions also have analogous functions:
   1. Binomial: dbinom, pbinom, qbinom
   2. t-distribution: dt, pt, qt

We will be using the t-distribution's dt, pt, and qt functions a lot in statistical inference.

Maximum Likelihood Estimation

We now turn to an important topic: maximum likelihood estimation.

MLE: The binomial distribution

Suppose we toss a fair coin 10 times, and count the number of heads each time; we repeat this experiment 5 times in all. The observed sample values are $x_1, x_2, \ldots, x_5$.

```
(x<-rbinom(5,size=10,prob=0.5))

## [1] 7 4 3 4 2
```

The joint probability of getting all these values (assuming independence) depends on the parameter we set for the probability θ:

$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
$= f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$

MLE: The binomial distribution

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$
$$= f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$$

So, the above probability is a function of θ. When this quantity is expressed as a function of θ, we call it the **likelihood function**.

MLE: The binomial distribution

The value of θ for which this function has the maximum value is the **maximum likelihood estimate**.

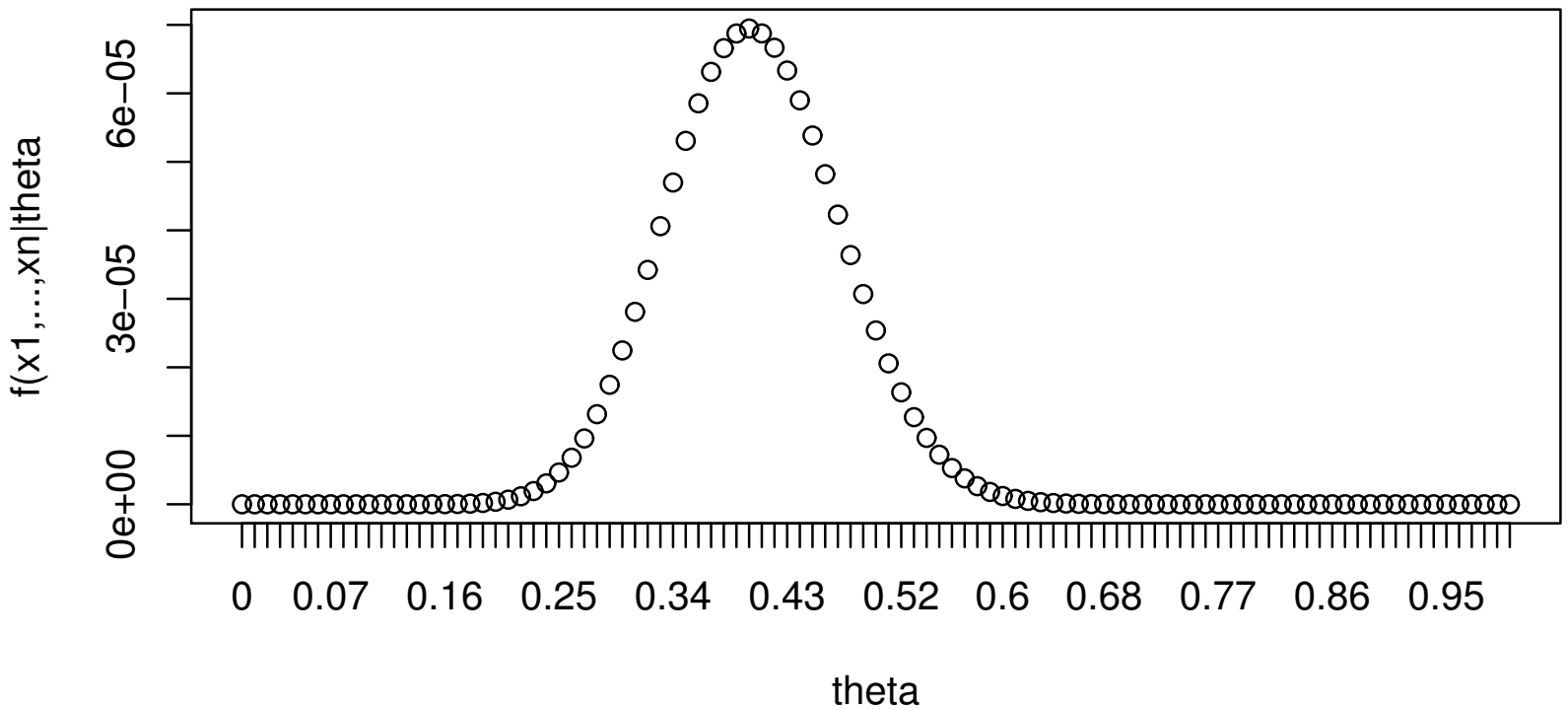
```
## probability parameter fixed at 0.5
theta<-0.5
prod(dbinom(x,size=10,prob=theta))

## [1] 2.53813e-05

## probability parameter fixed at 0.1
theta<-0.1
prod(dbinom(x,size=10,prob=theta))

## [1] 1.211404e-11
```

MLE: The binomial distribution

Let's compute the product for a range of probabilities:

```
theta<-seq(0,1,by=0.01)
store<-rep(NA,length(theta))
for(i in 1:length(theta)){
store[i]<-prod(dbinom(x,size=10,prob=theta[i]))
}
```

MLE: The binomial distribution



MLE: The binomial distribution

Detailed derivations: see lecture notes

We can obtain this estimate of θ that maximizes likelihood by computing:

$$\hat{\theta} = \frac{x}{n} \tag{11}$$

where n is sample size, and x is the number of successes.

For the analytical derivation, see the Linear Modeling lecture notes: https://github.com/vasishth/LM

MLE: The normal distribution

Detailed derivations: see lecture notes

For the normal distribution, where $X \sim N(\mu, \sigma)$, we can get MLEs of μ and σ by computing:

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x} \tag{12}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \tag{13}$$

you will sometimes see the "unbiased" estimate (and this is what R computes) but for large sample sizes the difference is not important:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \tag{14}$$

The significance of the MLE

The significance of these MLEs is that, having assumed a particular underlying pdf, we can estimate the (unknown) parameters (the mean and variance) of the distribution that generated our particular data.
This leads us to the distributional properties of the mean **under repeated sampling**.

The sampling distribution of the mean

When we have a **single sample**, we know how to compute MLEs of the sample mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$.
Suppose now that you had many repeated samples; from each sample, you can compute the mean each time. We can simulate this situation:

```
x<-rnorm(100,mean=500,sd=50)
mean(x)

## [1] 501.3919

x<-rnorm(100,mean=500,sd=50)
mean(x)

## [1] 497.2572
```

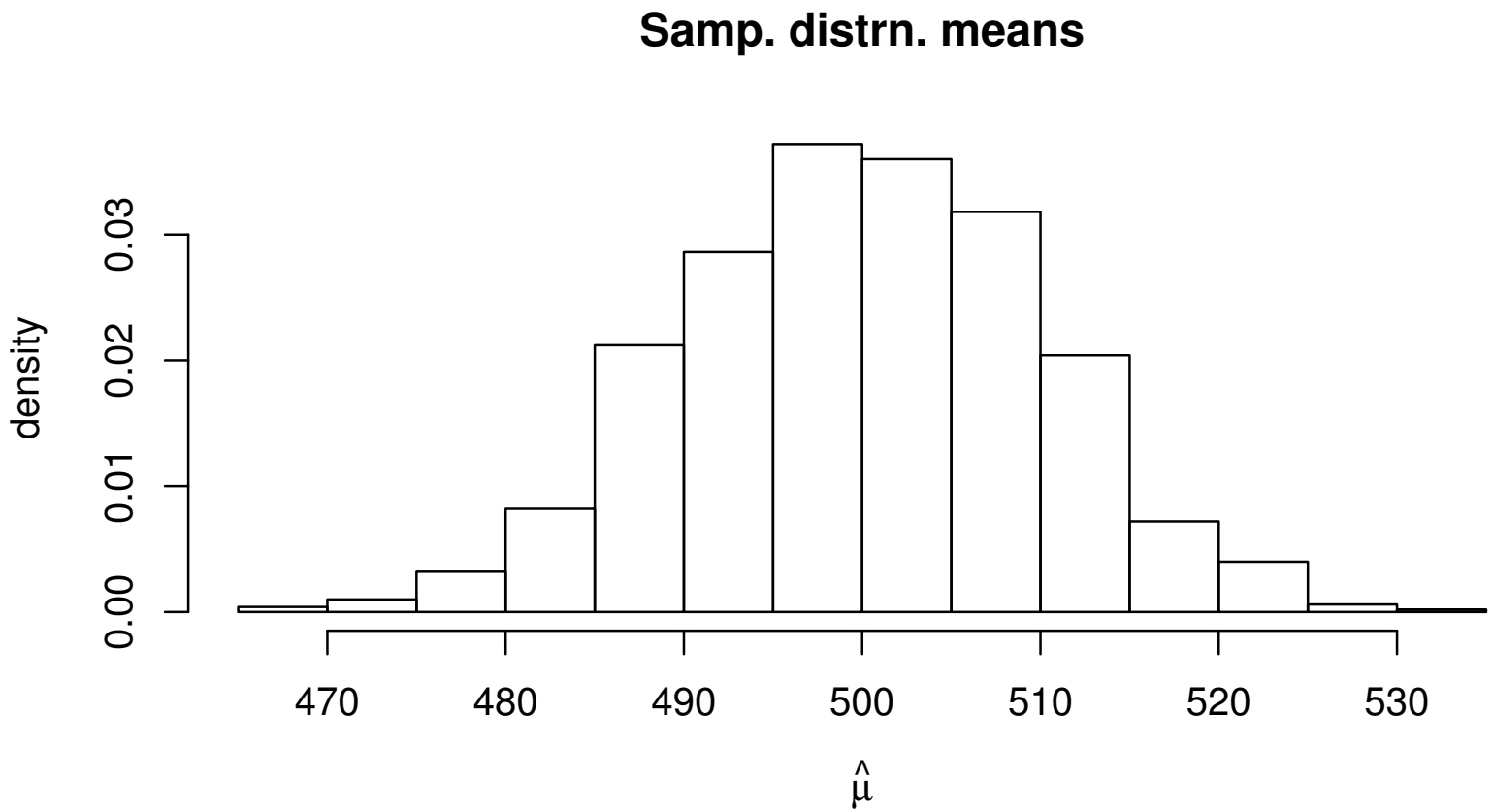
The sampling distribution of the mean

Let's repeatedly simulate sampling 1000 times:

```
nsim<-1000
n<-100
mu<-500
sigma<-100
samp_distrn_means<-rep(NA,nsim)
samp_distrn_sd<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n,mean=mu,sd=sigma)
  samp_distrn_means[i]<-mean(x)
  samp_distrn_sd[i]<-sd(x)
}
```

The sampling distribution of the mean

Plot the distribution of the means under repeated sampling:

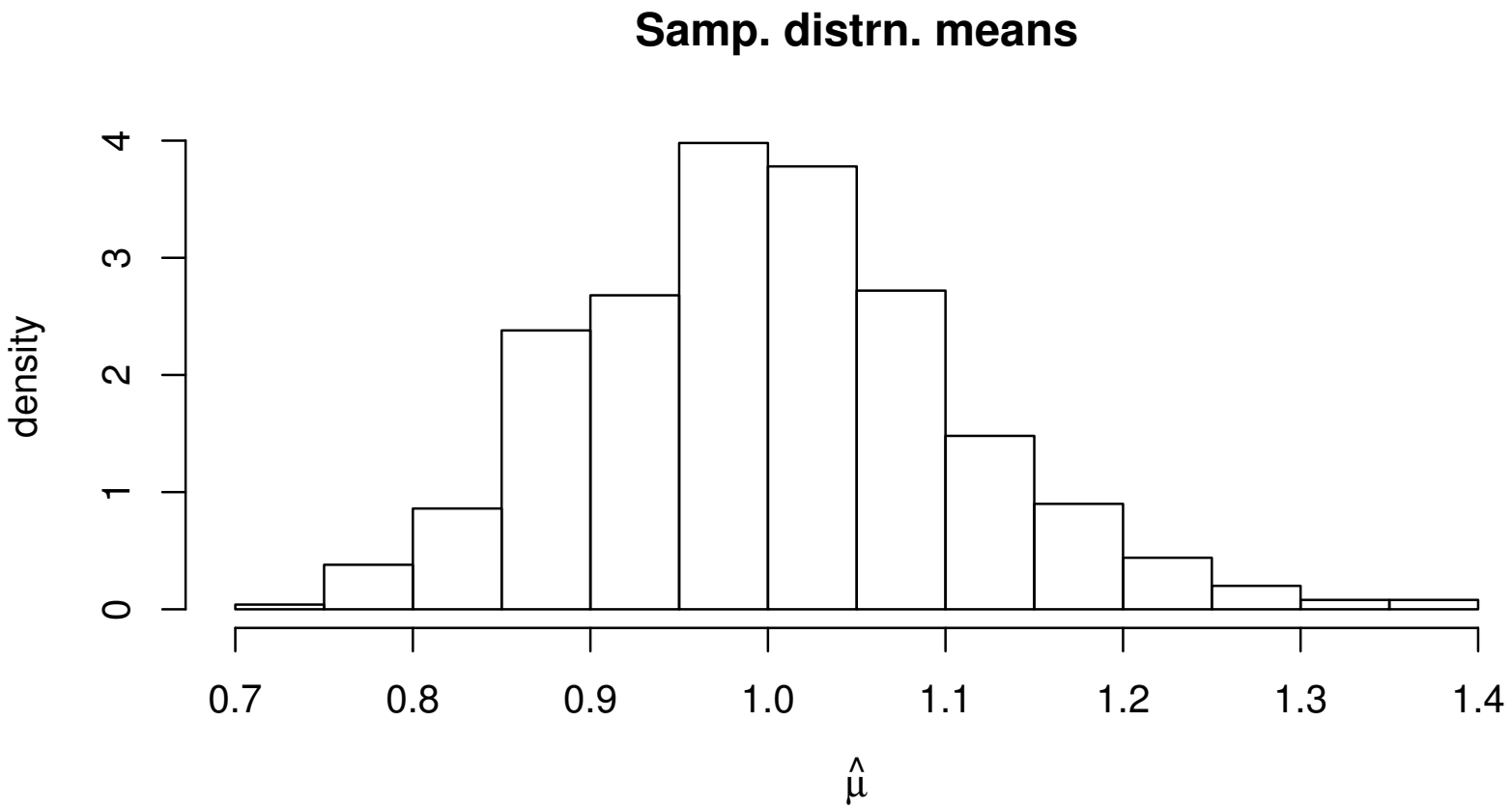


The sampling distribution of the mean

Interestingly, it is not necessary that the distribution that we are sampling from be the normal distribution.

```
for(i in 1:nsim){
  x<-rexp(n)
  samp_distrn_means[i]<-mean(x)
  samp_distrn_sd[i]<-sd(x)
}
```

The sampling distribution of the mean



The central limit theorem

1. For large enough sample sizes, the sampling distribution of the means will be approximately normal, regardless of the underlying distribution (as long as this distribution has a mean and variance defined for it).
2. This will be the basis for statistical inference.

The sampling distribution of the mean

We can compute the standard deviation of the sampling distribution of means:

```
## estimate from simulation:
sd(samp_distrn_means)

## [1] 0.1059066
```

The sampling distribution of the mean

A further interesting fact is that we can compute this standard deviation of the sampling distribution **from a single sample** of size n:

$$\frac{\hat{\sigma}}{\sqrt{n}}$$

```
x<-rnorm(100,mean=500,sd=100)
hat_sigma<-sd(x)
hat_sigma/sqrt(n)

## [1] 10.1317
```

See linear modeling notes on github for an analytical proof.

The sampling distribution of the mean

1. So, from a sample of size n, and sd σ or an MLE $\hat{\sigma}$, we can compute
   the standard deviation of the sampling distribution of the means.
2. We will call this standard deviation the estimated **standard error**.
   $SE = \frac{\hat{\sigma}}{\sqrt{n}}$
   I say *estimated* because we are estimating SE using an an estimate of σ.

Confidence intervals

The standard error allows us to define a so-called **95% confidence interval**:

$$\hat{\mu} \pm 2SE \tag{15}$$

So, for the mean, we define a 95% confidence interval as follows:

$$\hat{\mu} \pm 2\frac{\hat{\sigma}}{\sqrt{n}} \tag{16}$$

Confidence intervals

In our example:

```
## lower bound:
mu-(2*hat_sigma/sqrt(n))

## [1] 479.7366

## upper bound:
mu+(2*hat_sigma/sqrt(n))

## [1] 520.2634
```

The meaning of the 95% CI

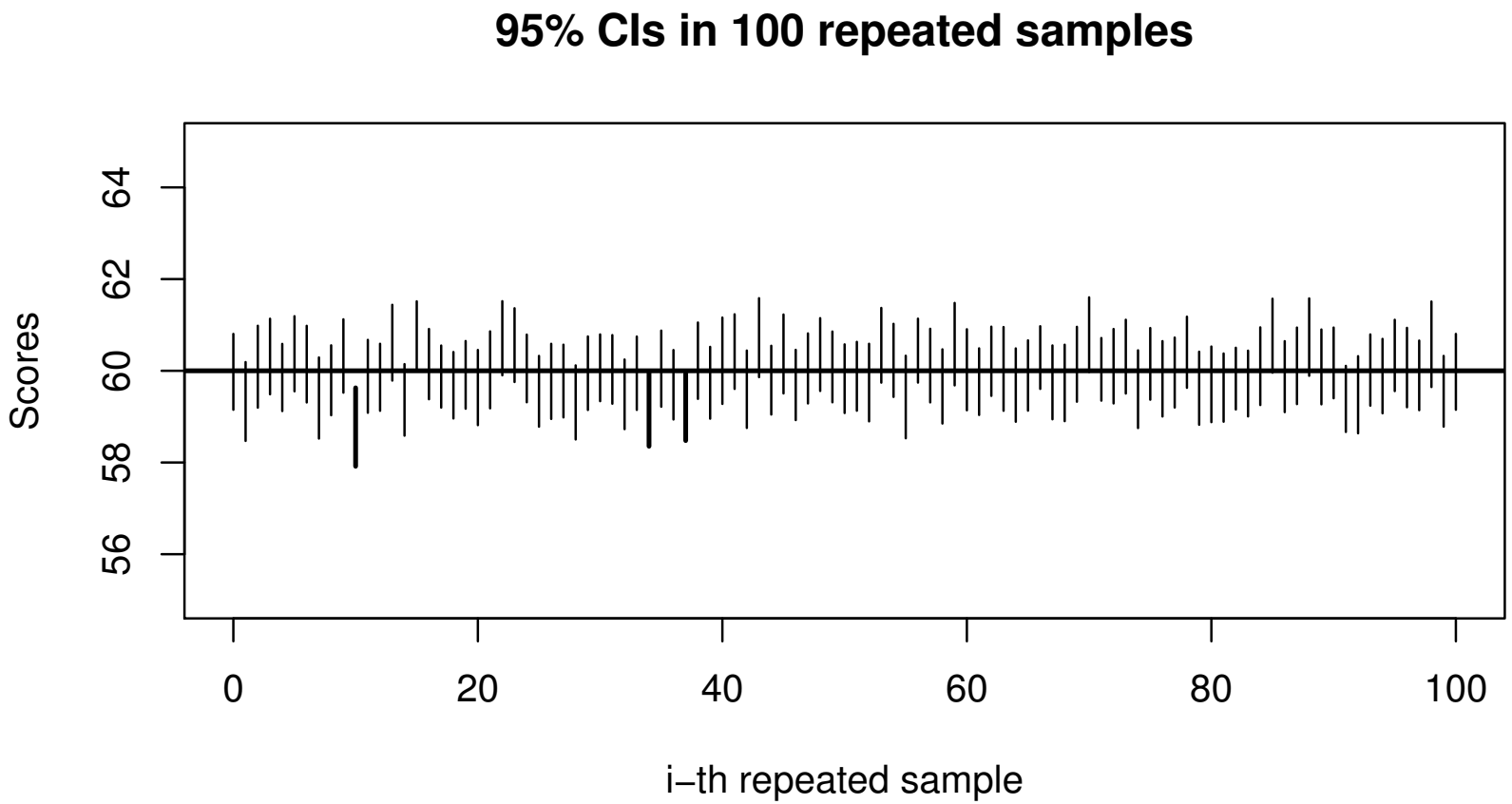
If you take repeated samples and compute the CI each time, 95% of those CIs will contain the true population mean.

```
lower<-rep(NA,nsim)
upper<-rep(NA,nsim)
for(i in 1:nsim){
  x<-rnorm(n,mean=mu,sd=sigma)
  lower[i]<-mean(x) - 2 * sd(x)/sqrt(n)
  upper[i]<-mean(x) + 2 * sd(x)/sqrt(n)
}
```

The meaning of the 95% CI

```
## check how many CIs contain mu:
CIs<-ifelse(lower<mu & upper>mu,1,0)
table(CIs)

## CIs
##   0   1
##  45 955

## approx. 95% of the CIs contain true mean:
table(CIs)[2]/sum(table(CIs))

##     1
## 0.955
```

The meaning of the 95% CI



The meaning of the 95% CI

1. The 95% CI from a particular sample does **not** mean that the probability that the true value of the mean lies inside that particular CI.
2. Thus, the CI has a very confusing and (not very useful!) interpretation.
3. In week two we will use the credible interval, which has a much more sensible interpretation.

However, for large sample sizes, the credible and confidence intervals tend to be essentially identical.
For this reason, the CI is often treated (this is technically incorrect!) as a way to characterize uncertainty about our estimate of the mean.

Main points from this lecture

1. We compute maximum likelihood estimates of the mean $\bar{x} = \hat{\mu}$ and standard deviation $\hat{\sigma}$ to get estimates of the true but unknown parameters.
   $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
2. For a given sample, having estimated $\hat{\sigma}$, we estimate the standard error:
   $SE = \hat{\sigma}/\sqrt{n}$
3. This allows us to define a 95% CI about the estimated mean:
   $\hat{\mu} \pm 2 \times SE$

From here, we move on to statistical inference and null hypothesis significance testing (NHST).