Introduction to statistics: Foundations

Shravan Vasishth

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

June 16, 2019

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)
}
```

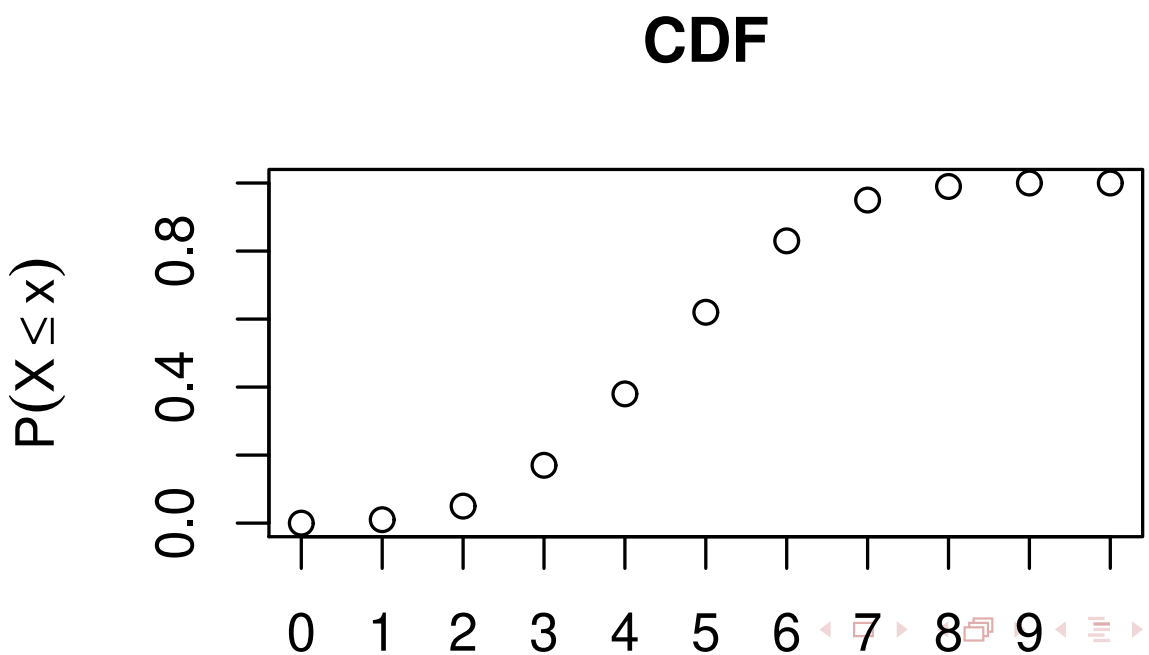
We have just computed the cdf of this random variable.

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] 0.0097656

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

## [1] 0.0097656
```

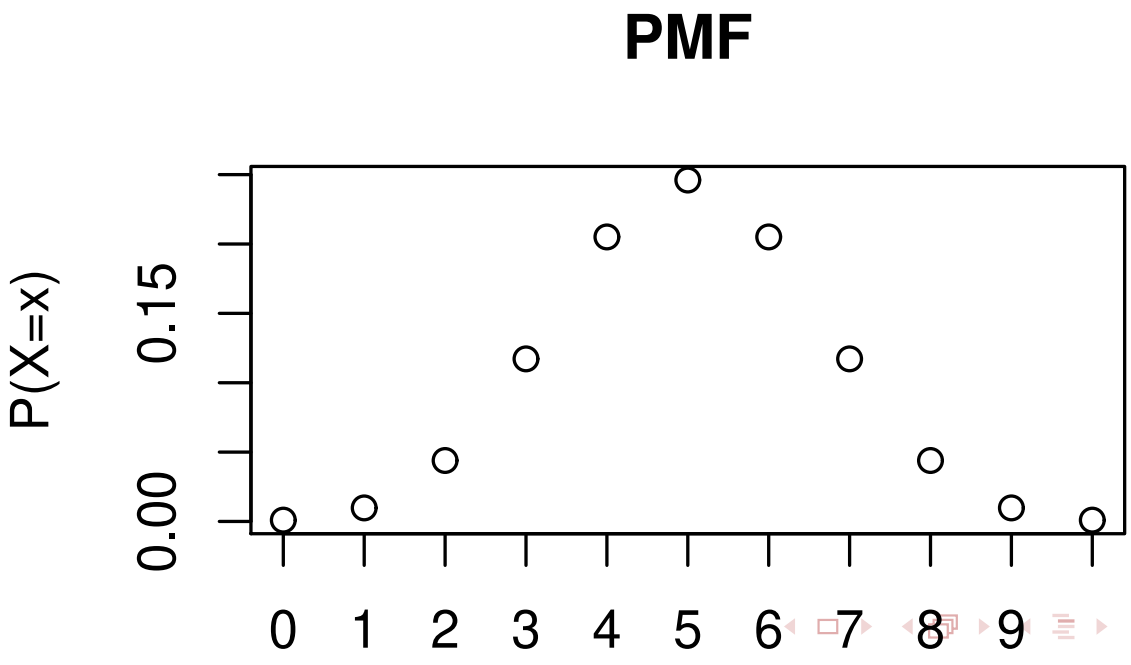
Discrete example: The binomial random variable

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

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

## [1] 0.0097656

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

## [1] 0.0097656
```

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, the discrete binomial 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:

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

Generating random binomial data

We can use the **rbinom** function to generate binomial data. So, 10 coin tosses can be simulated as follows:

```
rbinom(1,n=10,prob=0.5)

##  [1] 0 1 1 0 0 0 1 1 1 0
```

We switch now to class exercises.

Continuous example: The normal random variable

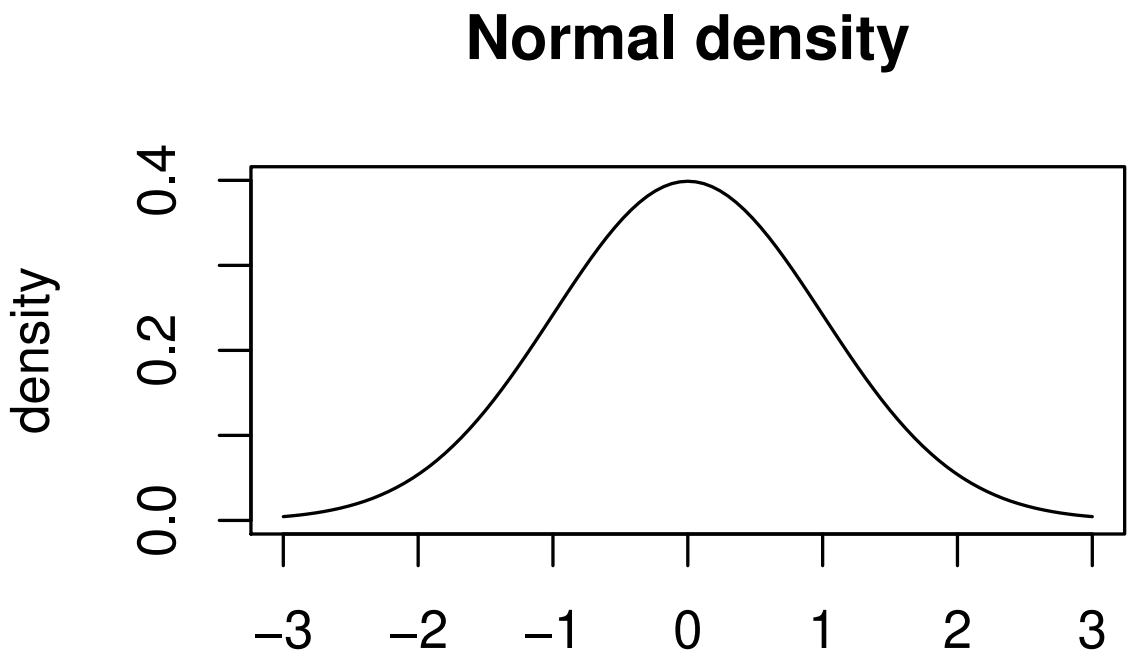
The pdf of the normal distribution is:

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

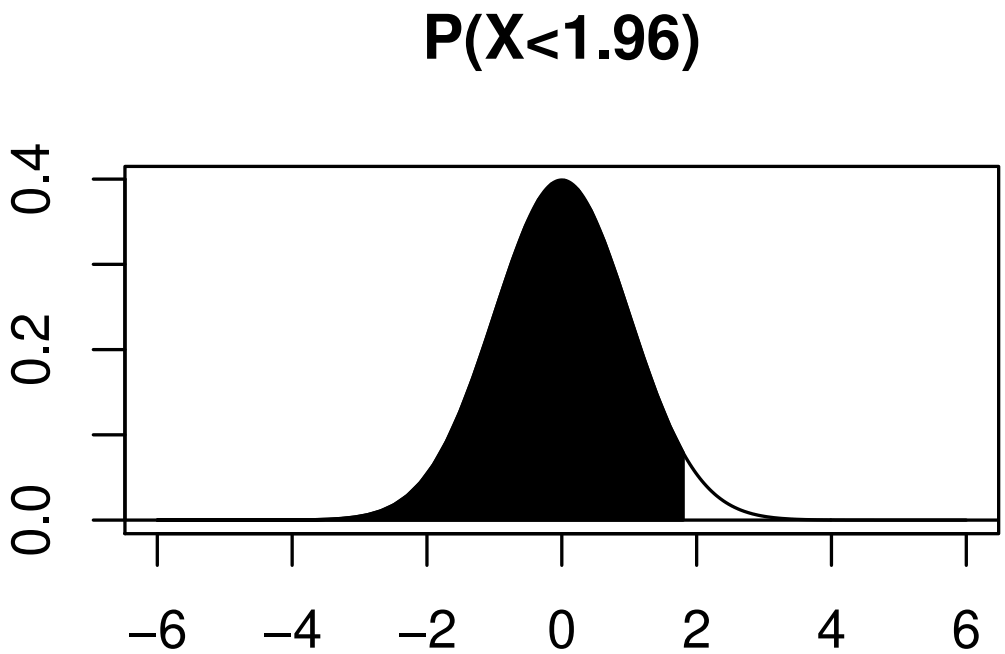
We write $X \sim \texttt{norm}(\texttt{mean} = \mu, \texttt{sd} = \sigma)$.
The associated `R` function for the pdf is `dnorm(x, mean = 0, sd = 1)`, and the one for cdf is `pnorm`.
Note the default values for μ and σ are 0 and 1 respectively. Note also that R defines the PDF in terms of μ and σ, not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



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] 0.9545

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

## [1] 0.68269
```

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] 696
```

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] 0.97725
```

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] 0.02275

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

## [1] 0.02275
```

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:
   - 1.1 dnorm: the pdf
   - 1.2 pnorm: the cdf
   - 1.3 qnorm: the inverse of the cdf
2. Other distributions also have analogous functions:
   - 2.1 Binomial: dbinom, pbinom, qbinom
   - 2.2 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] 5 4 3 5 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**.

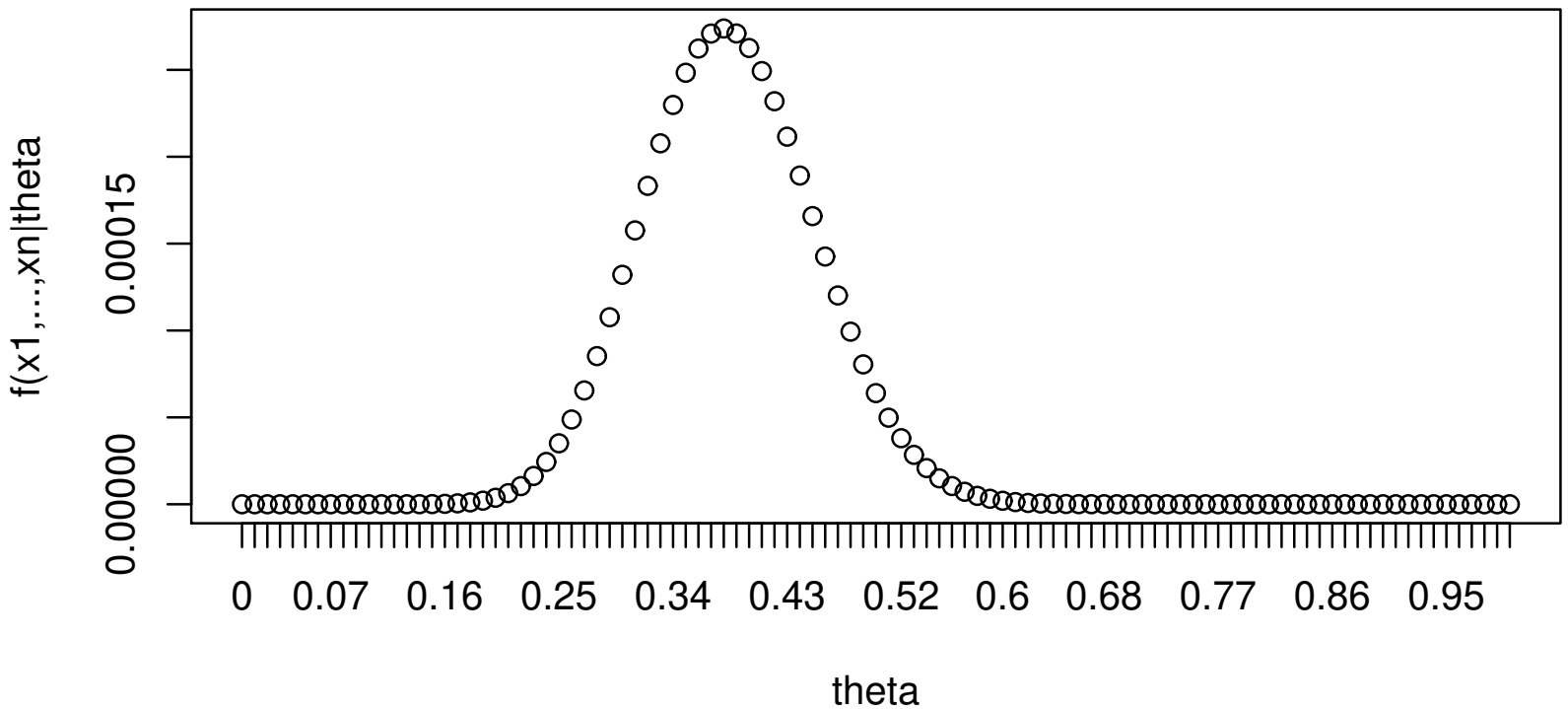
```
## probability parameter fixed at 0.5
theta<-0.5
prod(dbinom(x,size=10,prob=theta))

## [1] 6.3961e-05

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

## [1] 2.7475e-10
```

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**.