Data Analysis

1. Comp. Programming
2. Statistics
3. Applications

R package Slidify – http://rawnathw.github.com/slidify

Slides – github.com / jtleek / dataanalysis

↑ all classes are available there as well as slides

to compile them:

library (slidify)
slidify ("index.Rmd")

Week 1

Data – values of ~~quant~~ qualitative or quantitative variables, belonging to a set of items

Set of items – subjects
variables – measurements

Data

kind of observation

Raw data

hard to use
complex format

Processed

ready for analysis
each variable forms a column
each observation forms a row
each file stores data about one kind of observation

Big data – data that is too huge to manipulate on a single computer

How to represent data?

H for height, W for weight } informative

Randomness

- represents incompletely ~~or~~ measured variable
- a random mechanism

Distributions

X – random value

P – probability $\quad 0 \le x \le 1$

$p_1 + \dots + p_n = 1.$

↙ continuous

↘ discrete

Parameters:

$N(\mu, \sigma)$ $\quad$ Poisson(λ)

$X \sim N(\mu, \sigma)$ means X has the $N(\mu, \sigma)$ distribution

The most important parameters

$E[X]$ – expected value

$\mathrm{Var}[X]$ – how "spread out" a distribution is measured in (unit of X)^2

$SD[X] = \sqrt{\mathrm{Var}[X]}$ how spread out, but in the same unit

Conditioning

to indicate that smth is fixed

$X|\mu$ X rand variable with fixed μ

$Y|X=2$ when X is fixed at 2

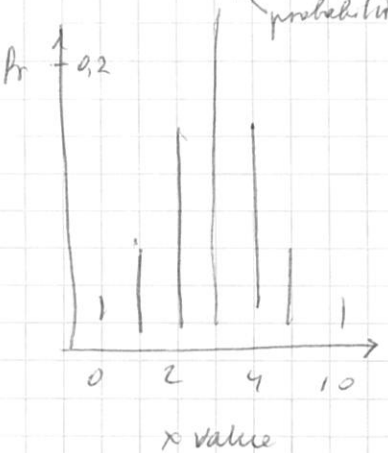
Distributions

- Binomial

number of coins

$X \sim Bin(10, 0.5)$

probability of Heads



dist. that describes a sum of coin flips

$E[X] = n \cdot p$

$Var[X] = n \cdot p \cdot (1-p)$

you flop 10 coins and count when they come up Heads

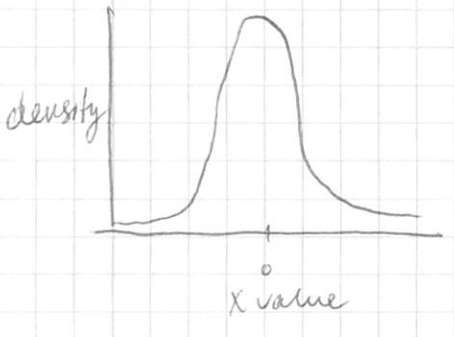
- Normal

mean

$N(\mu, \sigma)$

std

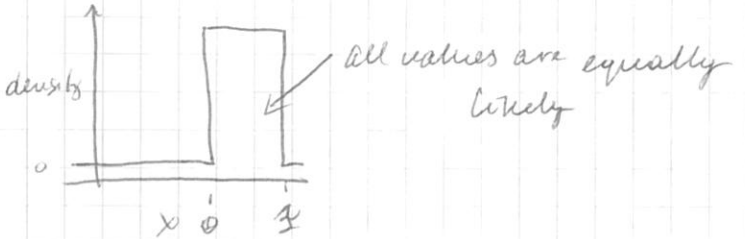
$X - N(0, 1)$



density – explains smth about probability for ranges of values of X.

– Uniform $U(\alpha, \beta)$

$X \sim U(0, 1)$



Conditioning

Law of total variance
Law of total expectation
=) Wiki

---

openintro.org/stat/textbook.php - free Statistics course

Representing data in R

myDataFrame[firstName == "jeff",]

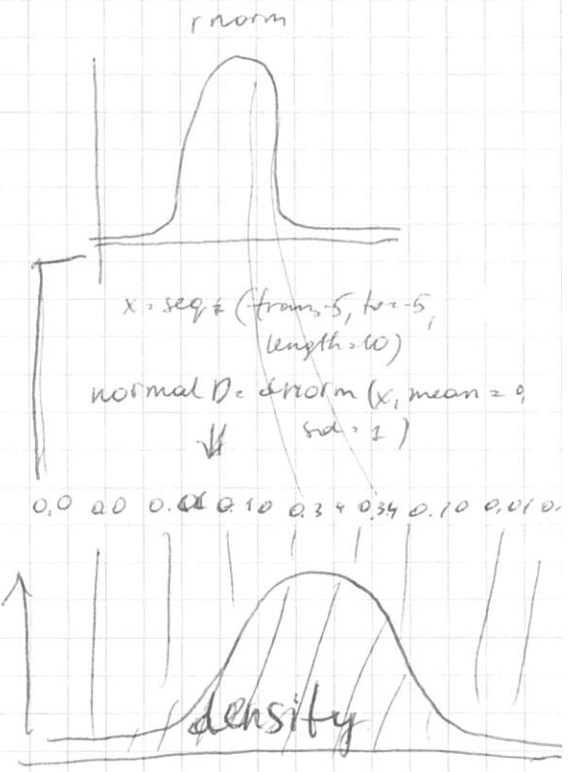
consider reading on style guides in R

Simulation Basics

Distributions	Densities
rbeta	dbeta
rbinom	dbinom
rgamma	dgamma
...	...
runif	dunif

Sampling

sample =



Sample draws a random sample

args(sample)

x size ← Number of samples replace = F, prob = NULL

h = rnorm(10, mean = 188, sd = 3)

s = sample(heights, size = 10, replace = T)

↓

the same value is used several times

Sample with probability

s = c(0.4, 0.3, 0.2, 0.1, 0 ... 0)

↓ ↓ ↓ ↓ ↓ ↓

1 2 3 4 5 ... 10 el of vec X

sample(heights, size = 10, replace = T, prob = s)

set.seed(12345)

↑

for reproducing

Types of Data Analysis

- Descriptive

  goal: to describe a set of data

  commonly applied to census data
  (for generalizing?) (not for explaining!)

- Explanatory

  goal: find relationships

  ideas for following studies

- Inferential analysis

  goal: use relatively small sample of data
  to say smth about a bigger population

  most common goal of statistics models
  depends on sample you got

- Predictive

  goal: use data on some object to predict
  values for another object

---

census – перепись, сбор сведений

- Causal

goal: to find out what happens to one variable when you make another variable change

explores an average effect
randomized, to see if X causes Y

- Mechanistic

goal: understand the exact changes in variables that lead to changes in other variables for individual objects

(in one object we change X → Y in the same object changes in predictible way)

for finding formula, etc by empirical analysis

Sources of Data Sets

censuses
observational study
randomized trials

etc

sample(1:8, size=4, replace=F)

census - all
observation study (sample with in replacements)
convenience (sample with probabilities) ∃ no replacement
randomized → get two groups and apply different

prediction: get a subset, do analysis,
get a subset of remaining
and check if it works

Study over time - cross-sectional
sub set only at one diary

longitudinal
the same individuals over time

retrospective
sampling at the end
(relationship between outcome and exposure)