Statistical Methods - 732A93

Teachers:

Hector Rodriguez-Deniz

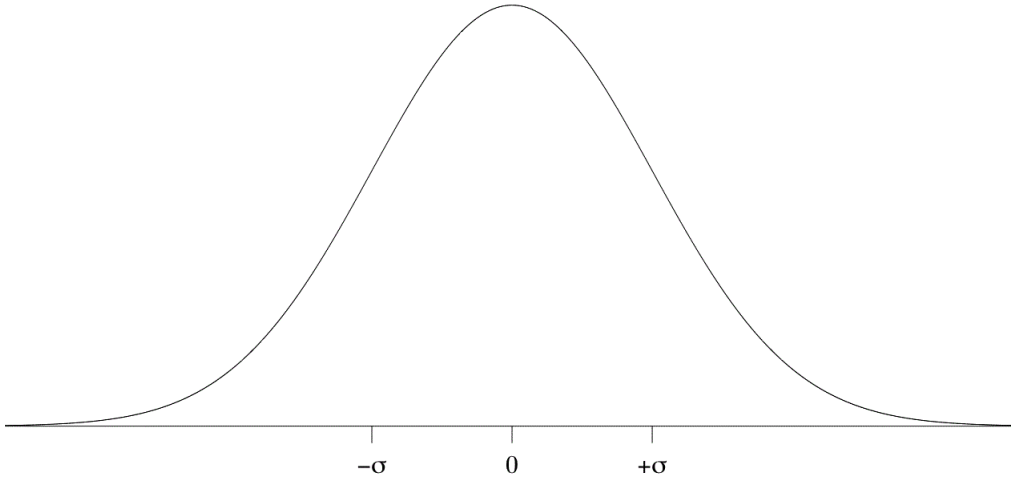
Ann-Charlotte Hallberg

Quick facts about the course

- Information will be posted in LISAM
  - Registration problems? → Contact Annelie Almquist in E-Building (Room E-3G:488)
- Course starts 27th August and ends 15th October 2019 (8 weeks)
- 12 seminars - theory + exercises
- Examination: Written exam and an assignment. Written exam at 2019-10-23
- Re-exam the 2019-11-27, second re-exam in Spring 2020
- Literature: Wackerly et al. (2008)
- Self-study course (**LOT** of theory) → No time to go through all theory!
- We will go through most theory by solving exercises (i.e. learning by doing)

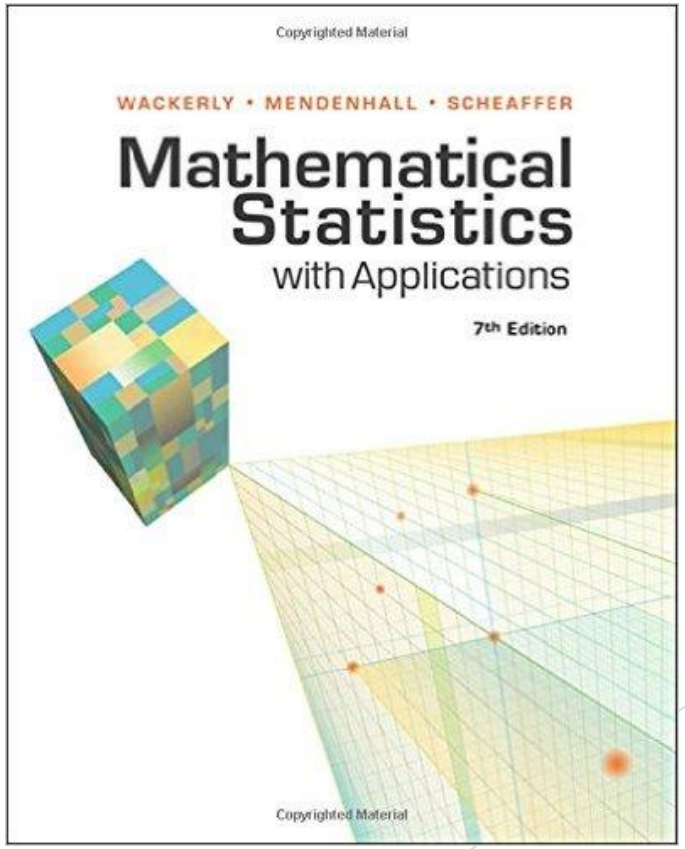
Aim of the course

- Basic grounds in probability and statistics
- Concept of probability. Random variable, common statistical distributions and their properties. Point and interval estimation. Hypothesis testing. Linear regression. Bayesian estimation.
- Supportive course for those without a degree in Statistics
- If you already know Statistics → good refreshing of basic concepts
- However, we assume you know at least some basics…
  - Recognize this distribution?



Course Literature

- Mathematical Statistics with Applications, 7th edition
- Wackerley, Mendenhall and Scheaffer, 2008
- Odd numbered exercises have answers
- There is also a solution manual
- We will solve relevant exercises in class
- Exercises with an * are more difficult



Examination

- Examination is written exam + computer assignment (3+3 credits)
- Written exam the 23th October → Don't forget to register in Studentportalen!!
  - 4 hours written exam
  - Grades from A to F → A is top, F is fail.
  - Aids: Pocket calculator + One handwritten A4 paper (both sides) with your notes
  - Probability distribution formulas + tables will be given with the exam
  - Some past exams with solutions will be available in advance
  - Do a lot of exercises to prepare the exam!
  - Re-exam the 27th November 2019
  - Re-exam in Spring 2020
- Computer assignment on multiple imputation techniques
  - Details will be given later on the course.

Course Syllabus

- From Wackerly's book:
  - Chapter 1: What is Statistics?
  - Chapter 2: Probability
  - Chapter 3: Discrete random variables and their probability distributions
  - Chapter 4: Continuous random variables and their probability distributions
  - Chapter 5: Multivariate probability distributions
  - Chapter 7: Sampling distributions and the central limit theorem
  - Chapter 8: Estimation
  - Chapter 9: Properties of point estimators and methods of estimation
  - Chapter 10: Hypothesis testing
  - Chapter 11: Linear models and estimation by least squares
  - Chapter 16: Introduction to Bayesian methods for inference
- Additional:
  - Multiple Imputation (Computer assignment)

Recommended exercises

- From Wackerly's book:
  - Chapter 2: 2.163
  - Chapter 3: 3.12, 3.24, 3.50, 3.66, 3.70, 3.122
  - Chapter 4: 4.8, 4.12, 4.30ab, 4.48, 4.74, 4.94, 4.96, 4.126
  - Chapter 5: 5.4, 5.6, 5.26, 5.34, 5.56, 5.82
  - Chapter 7: 7.20, 7.26, 7.38, 7.46
  - Chapter 8: 8.10, 8.44, 8.48, 8.64
  - Chapter 9: 9.74, 9.80, 9.90
  - Chapter 10: 10.6, 10.18
  - Chapter 11: 11.4, 11.26, 11.40, 11.47, 11.68, 11.70, 11.76
  - Chapter 16: 16.6, 16.10, 16.18, 16.24

About the teachers

- Hector Rodriguez-Deniz (PhD Student)
  - Course responsible
  - Office at B building, room B 3D:451 → Same aisle as Von Neumann's room
  - hector.rodriguez@liu.se
- Ann-Charlotte Hallberg (Lecturer)
  - Course examiner
  - Office at B building, room B 3E:489 → Same aisle as Oleg's office
  - ann-charlotte.hallberg@liu.se

Questions?

- Any question/suggestion??

Basic probability definitions from Ch. 2

Definitions 2.1,2.3, 2.5

An experiment is the process by which an observation is made. The outcome of an experiment is one or more events (usually represented by capital letters)

Example of experiment: throwing a dice

The sample space S associated to an experiment is the set consisting of all possible outcomes (simple events)

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

An event (in general) in a discrete sample space is a collection of single events, i.e. any subset of S. For example, $A = \{E_1, E_3, E_5\}$ is the event of the odd results in the dice experiment.

Basic probability definitions from Ch. 2

Definition 2.6: Classic definition of probability

Suppose S the sample space associated to an experiment. To every event A in S we assign a number P(A) called the probability of A, so that the following three axioms hold:

Kolgomorov Axioms:

I. $P(A) \geq 0$

II. $P(S) = 1$

III. $If\ A_1, A_2, A_3 \ldots$ is a sequence of events and $A_i \cap A_j = empty\ set$ then $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum P(A_i)$

In words,

I. *The probability of an event is a real number greater or equal to zero*

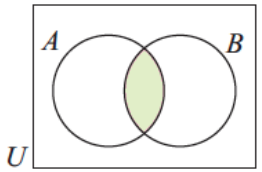
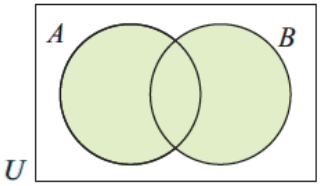
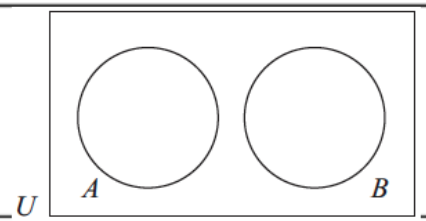
II. *The probability of the entire sample space is 1, and there are no events outside the sample space*

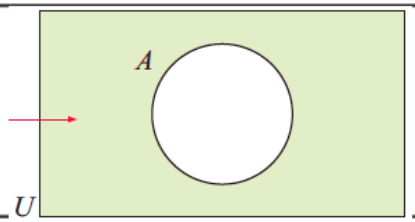
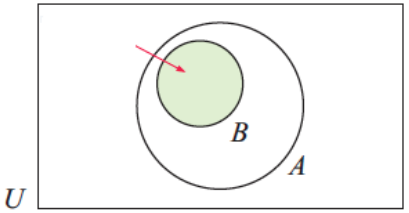
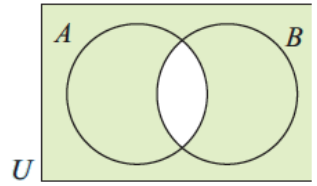
III. *The probability of a set of disjoint events is equal to the sum of the probabilities of every event in the set*

Basic probability definitions from Ch. 2

Theorems 2.5-2.7: Intersection, union and complement

$P(A \cap B) = P(A)P(B|A)$, *if A,B are independent then* $P(A \cap B) = P(A)P(B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(\bar{A}) = 1 - P(A)$

Set Notation	Visual representation
$A \cap B$	
$A \cup B$	
$A \cap B = \varnothing$	

Set Notation	Visual representation
A'	
$B \subseteq A$	
$(A \cap B)'$	

Basic probability definitions from Ch. 2

Definition 2.9 Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ where } P(B)>0$$

$P(A|B)$ is read "the probability of A given B".

Definition 2.10 Independence

$$P(A \cap B) = \mathrm{P(A)} \cdot P(B)$$

$P(A \cap B)$ is also called the **joint probability** of A and B, also represented as P(A,B)

Theorem 2.8 Law of total probability

Let $B_1, B_2, \ldots, B_k$ be a partition of S, $B_i \cap B_j = empty\ set$

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Theorem 2.9 Bayes theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

$P(B_i)$ is our prior belief (i.e. uncertainty) in outcome B_i → Prior probability

After seeing some data A we can compute conditionals $P(A|B_i)$

Using the formula above we get the "updated" belief $P(B_i|A)$ in B_i → Posterior probability

Example:

B_1: The patient is smoker

B_2: The patient is non-smoker

A: The patient has lung cancer (we have DATA on this)

What is the probability that a randomly chosen patient with cancer is smoker?, i.e. $P(B_1|A)$?

A priori, we believe that $P(B_1) = 0.45$ and $P(B_2) = 0.55$

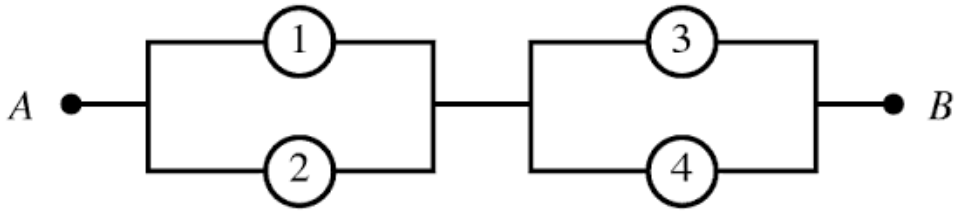
Now from the data that we have on our patients, we can calculate $P(A|B_1) = 0.9$ and $P(A|B_2) = 0.05$

Using the above theorem:

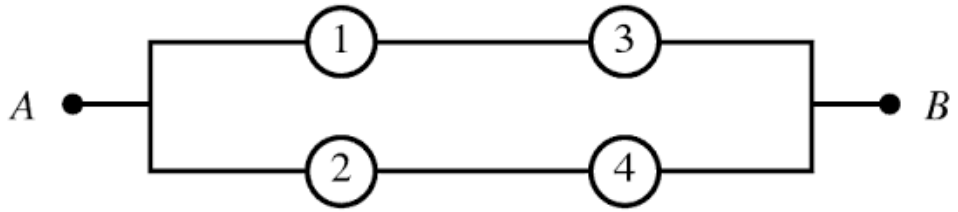
$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{2} P(A|B_i)P(B_i)} = \frac{0.9\ x\ 0.45}{(0.9\ x\ 0.45) + (0.05\ x\ 0.55)} = 0.9364$$

Solve exercise 2.163 from the book.

2.163 Relays used in the construction of electric circuits function properly with probability .9. Assuming that the circuits operate independently, which of the following circuit designs yields the higher probability that current will flow when the relays are activated?



A



B

Chapter 3: Discrete Random Variables and Their Probability Distributions

Definition 3.1:

A random variable Y is said to be discrete if it can assume only a finite or a countably infinite number of distinct values.

Def 3.2 and 3.3

The probability of Y taken on the value y is $P(Y = y)$

The probability distribution of Y is denoted $\mathrm{p}(y) = P(Y = y)$ for all values y

That is, if you know all the probabilities then you know the distribution!

Definition 3.4 Expected value

$$\mu = E[Y] = \sum_{y} y \cdot p(y)$$

Theorem 3.2

$$E[g(Y)] = \sum_{y} g(y) \cdot p(y)$$

Definition 3.5 Variance

$$\sigma^2 = V[Y] = Var[Y] = \mathrm{E}[(Y - \mu)^2]$$

Theorem 3.6

$$Var[Y] = E[Y^2] - E^2[Y]$$

Theorem 3.4 *if* c is a constant

$$E[cY] = cE[Y]$$

Theorem 3.5 The expected value of a sum of random variables is the sum of the expected values

$$E\left[\sum Y\right] = \sum E[Y]$$

Proof Theorem **3.6** (page 96)

Solve exercise **3.12** (page 97)

Theorem 3.6 $Var[Y] = E[Y^2] - E^2[Y]$

3.12 Let Y be a random variable with $p(y)$ given in the accompanying table. Find $E(Y)$, $E(1/Y)$, $E(Y^2 - 1)$, and $V(Y)$.

y	1	2	3	4
$p(y)$	.4	.3	.2	.1

Binomial distribution

Def. 3.6 (read), Def. 3.7, Theo. 3.7

Y is a random variable

Y = number of successful trials out of *n* independent trials

p = P(successful trial)

Y has probability distribution:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0,1,\dots,n$$

$$Y \sim Bin(n,p)$$

$$E[Y] = \mathrm{np} \qquad\qquad Var[Y] = \mathrm{np}(1-\mathrm{p})$$

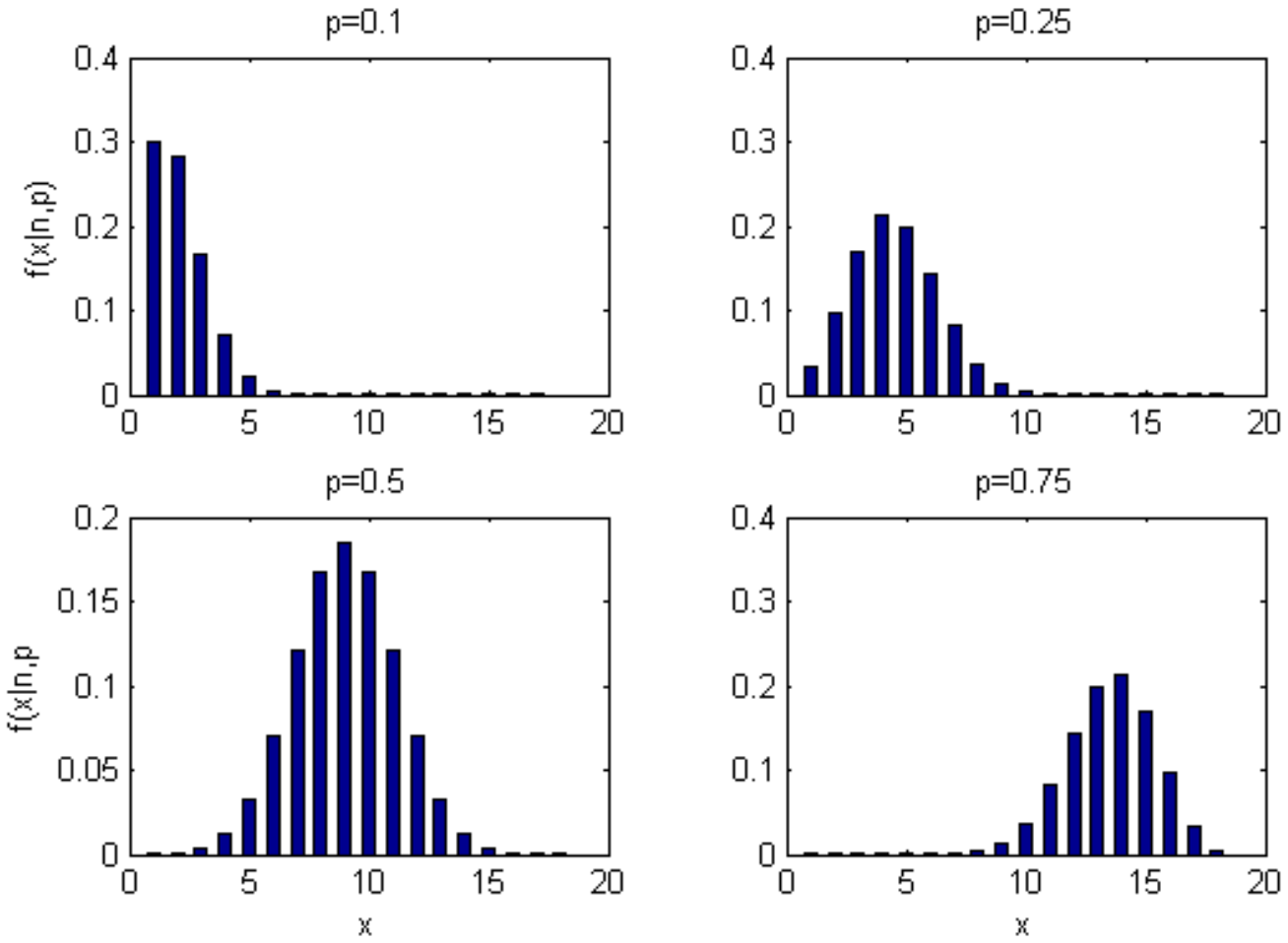
When *n=1* (a single trial) we have a Bernoulli distribution, i.e.

$Y \sim Bin(1,p) \equiv Y \sim Ber(p)$, with probability $p(y) = p^y(1-p)^{1-y}, \quad y = \{0,1\}$

$$E[Y] = \mathrm{p} \qquad\qquad Var[Y] = \mathrm{p}(1-\mathrm{p})$$

Binomial distribution

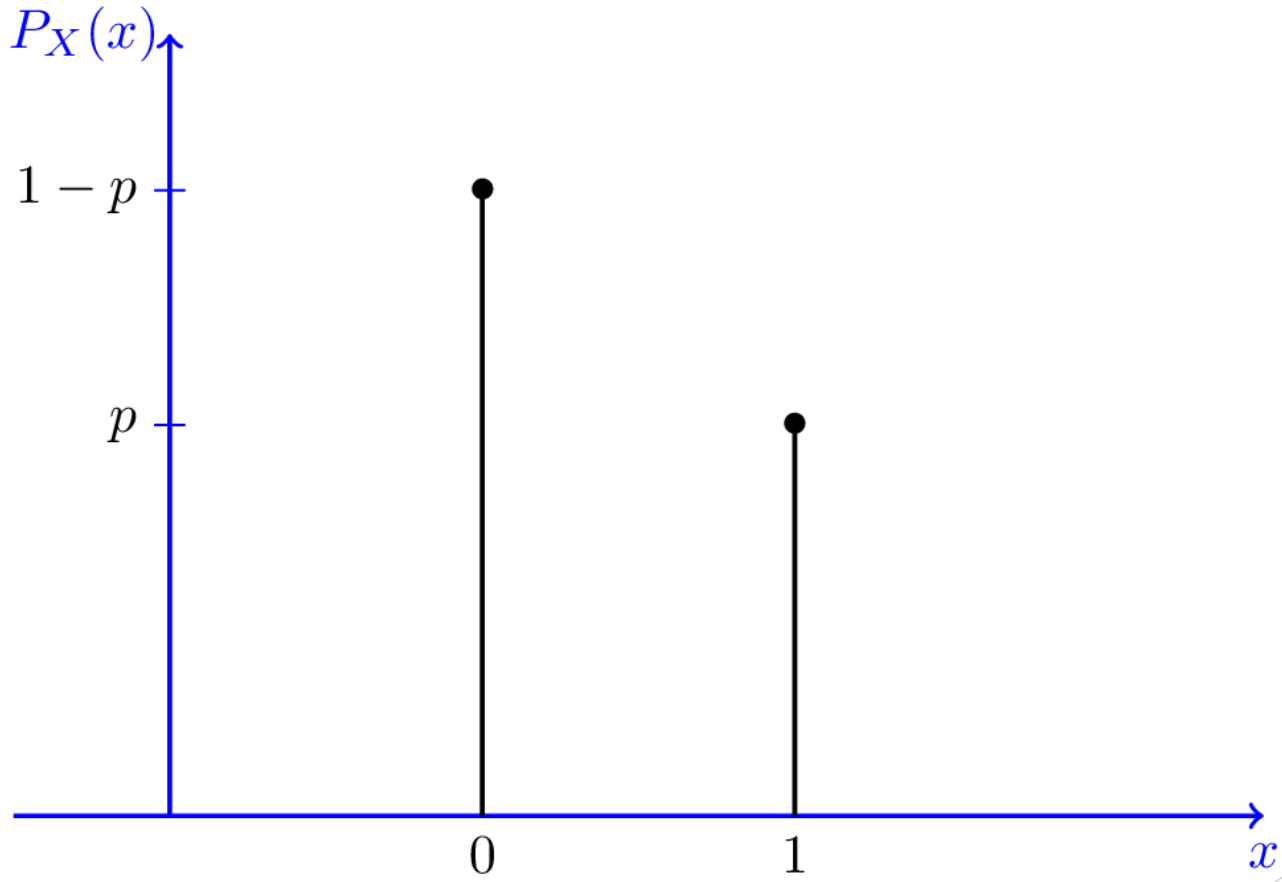
Some Binomial distributions for n = 20



Bernoulli distribution

P is probability of success, (1-P) probability of failure

$$X \sim Bernoulli(p)$$



Solve exercises **3.24 (p. 99)**, **3.50 (p. 112)**

3.24 Approximately 10% of the glass bottles coming off a production line have serious flaws in the glass. If two bottles are randomly selected, find the mean and variance of the number of bottles that have serious flaws.

3.50 A missile protection system consists of n radar sets operating independently, each with a probability of .9 of detecting a missile entering a zone that is covered by all of the units.

a If $n = 5$ and a missile enters the zone, what is the probability that exactly four sets detect the missile? At least one set?

b How large must n be if we require that the probability of detecting a missile that enters the zone be .999?

Geometric distribution

Def. 3.8, Theo. 3.8

Y is a random variable

Y = the number of trials until (first) success occur

Trials are independent.

p = P(successful trial)

Y has probability distribution:

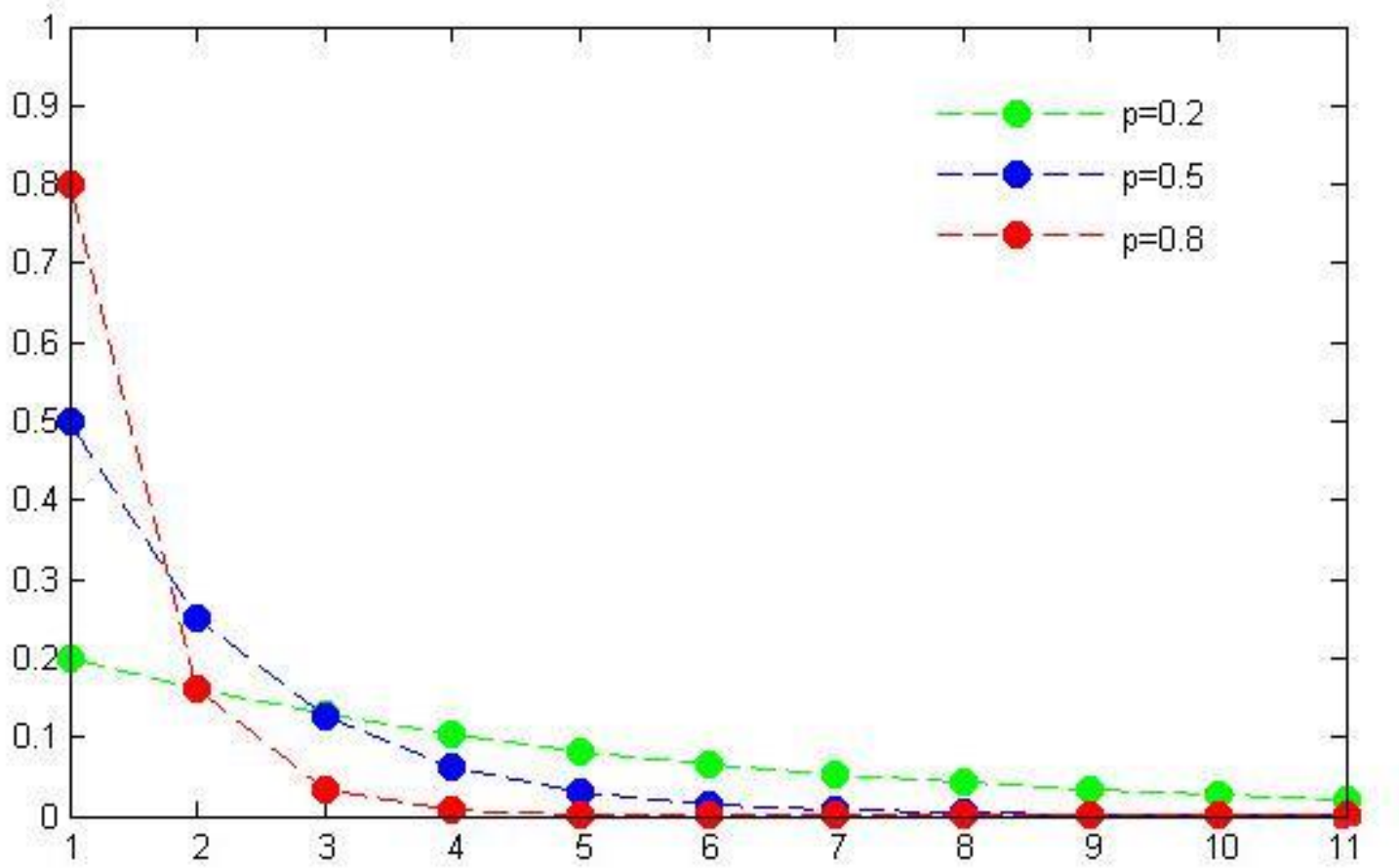
$$p(y) = (1-p)^{y-1}p, \qquad \text{y} = 1,2,\ldots$$

$$Y \sim Geo(p)$$

$$E[Y] = \frac{1}{p} \qquad Var[Y] = \frac{1-p}{p^2}$$

Geometric distribution

Some Geometric distributions for different values of p



Solve exercise **3.70** (p. 119)

3.70 An oil prospector will drill a succession of holes in a given area to find a productive well. The probability that he is successful on a given trial is .2.

a What is the probability that the third hole drilled is the first to yield a productive well?

b If the prospector can afford to drill at most ten wells, what is the probability that he will fail to find a productive well?

Poisson distribution

Def. 3.11, Theo. 3.11

Y is a random variable

Y = number of events during an interval (often time interval)

Example events:

accidents, telephone calls to a switchboard, number of cars to a gas station

Y has probability distribution:

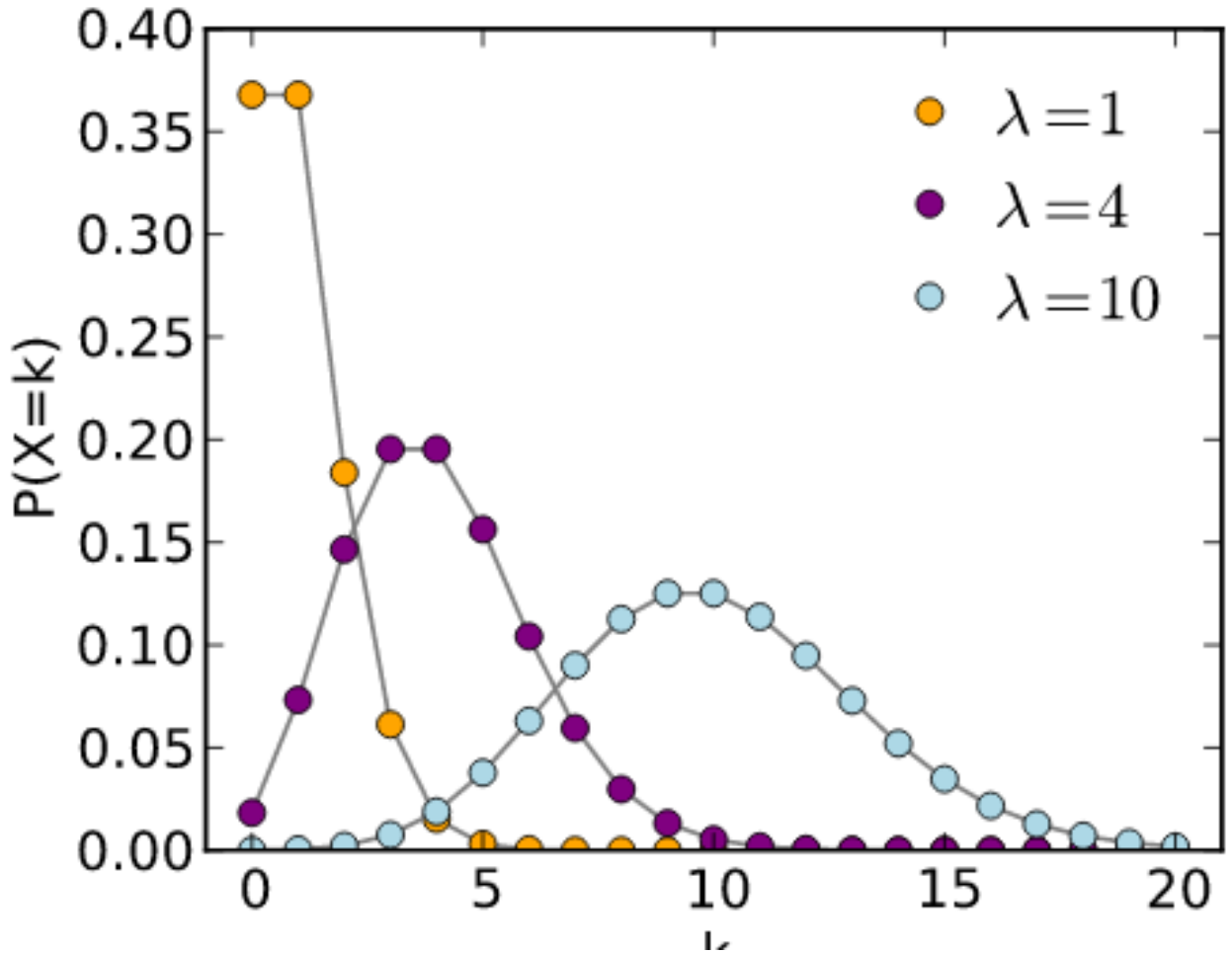
$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \qquad y = 0,1,2, \ldots \qquad \lambda > 0$$

$$Y \sim Poi(\lambda)$$

$$E[Y] = \lambda \qquad\qquad Var[Y] = \lambda$$

Poisson distribution

Some Poisson distributions for different values of λ



Solve exercise **3.122** (p.136)

3.122 Customers arrive at a checkout counter in a department store according to a Poisson distribution at an average of seven per hour. During a given hour, what are the probabilities that

a no more than three customers arrive?

b at least two customers arrive?

c exactly five customers arrive?