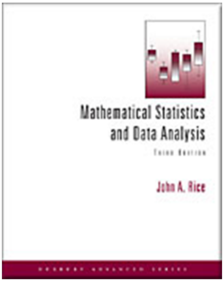
ST121:
Probability and Statistics I

Waleed A. Yousef, Ph.D.,

Human Computer Interaction Lab.,
Computer Science Department,
Faculty of Computers and Information,
Helwan University,
Egypt.

March 24, 2019

Lectures follow Rice, "*Mathematical Statistics and Data Analysis*", 3rd edition, Duxbury:



ISBN 0-534-39942-8

Course Objectives

- Developing rigorous treatment.
- Building intuition.
- Linking to real life problems.

Contents

Chapter 1

Probability

1.1 Introduction

"Probability" gives the meaning of chance or randomness; but how to formalize?
"Probability" is almost everywhere

- Genetics and Bioinformatics, e.g., mutation
- kinetic theory of gases.
- Queuing theory: tremendous applications
- Theory of finance
- $\vdots$

1.2 Sample Spaces

Definition 1 (Sample Space) *Ω is the set of all possible outcomes (we denote each outcome by ω).*

Definition 2 (Event) *is a subset of Ω.*

Example 3 *Passing by 3 traffic lights, at each either continue (c) or stop (s). Then*

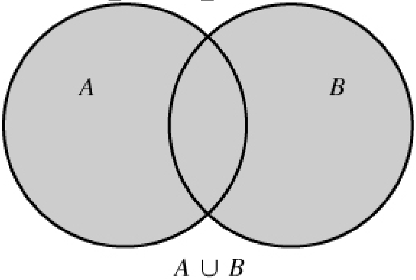
$$\Omega = \{ccc, ccs, csc, css, scc, scs, ssc, sss\},$$
$$A = \{sss, ssc, scc, scs\}$$

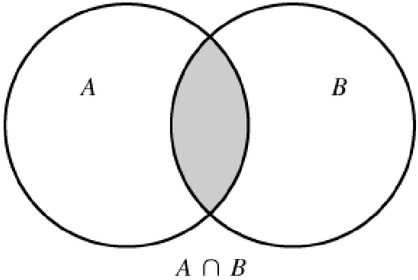
is the event of stopping at the first traffic light.

Example 4 *The length of time between two successive earthquakes in a particular region*

$$\Omega = \{t | t \geq 0\}, \text{(the set of nonnegative reals)}$$
$$A = \left\{t | t \geq 1 day\right\},$$

All set theory axioms and laws apply because "Sample Space" and "Events" are "Sets".





Intersection:

$$A \cap B = \{\omega | \omega \in A \wedge \omega \in B\}$$

$$A = \{sss, ssc, scs, scc\} \text{ (stopping at first)},$$

$$B = \{sss, scs, ccs, css\} \text{ (stopping at third)}$$

$$A \cap B = \{sss, scs\}$$

Union:

$$A \cup B = \{\omega | \omega \in A \vee \omega \in B\}$$

$$A = \{sss, ssc, scs, scc\},$$

$$B = \{sss, scs, ccs, css\}$$

$$A \cup B = \{sss, ssc, scs, scc, ccs, css\}$$

Complement:

$$A^c = \{\omega | \omega \notin A\}$$

$$A^c = \{ccc, ccs, \mathrm{csc}, css\}$$

1.3 Probability Measure

We define **rigorously** to meet what is in minds about probability:

- Probability as frequency and chance to happen
- Probability as a subjective belief

Definition 5 (Probability Measure) *is a function P from subsets of Ω to the real numbers that satisfies*

1. $P(\Omega) = 1$,
2. $\forall A \subset \Omega,\ P(A) \geq 0$,
3. *If A_i and A_j are disjoint $\forall i, j$ then*
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

No need for more axioms**; all come byproduct.**

Properties

$$
\begin{aligned}
P\left(A^c\right) &= 1 - P(A)\\
P(A) &\leq 1\\
P\left(\phi\right) &= 0\\
P(A) &\leq P(B),\ \forall A \subseteq B
\end{aligned}
$$

Proof.

$$
\begin{aligned}
P(\Omega) &= P\left(A \cup A^c\right)\\
&= P(A) + P\left(A^c\right) && \text{(Axiom 3)}\\
P\left(A^c\right) &= 1 - P(A)
\end{aligned}
$$

$$
\begin{aligned}
P(A) &= 1 - P\left(A^c\right),\\
P\left(A^c\right) &\geq 0 \text{ (Axiom 2)}\\
P(A) &\leq 1
\end{aligned}
$$

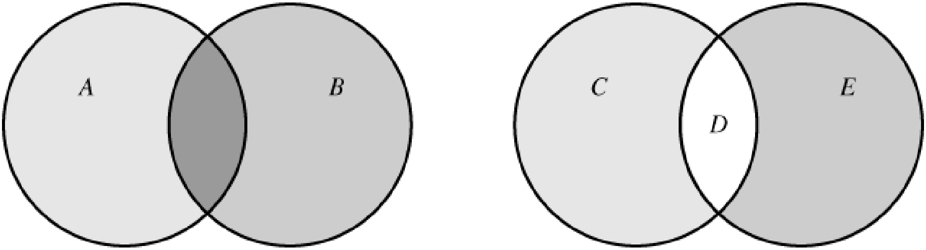
$$
\begin{aligned}
P(\Omega) &= P\left(\Omega \cup \phi\right)\\
&= P(\Omega) + P\left(\phi\right)\\
P\left(\phi\right) &= 0
\end{aligned}
$$

$$B = A \cup (B \setminus A) \;\; \forall A \subset B$$

$$P(B) = P(A) + P(B \setminus A)$$

$$P(A) = P(B) - P(B \setminus A)$$

$$P(A) \leq P(B)$$



$$
\begin{aligned}
A \cup B &= \underbrace{(A \setminus (A \cap B))}_{A \cap B^c} \cup (A \cap B) \\
&\cup \underbrace{(B \setminus (A \cap B))}_{B \cap A^c} \\
P(A \cup B) &= P(A \setminus (A \cap B)) + P(A \cap B) \\
&+ P(B \setminus (A \cap B)) \\
&= \underbrace{P(A \setminus (A \cap B)) + P(A \cap B)}_{P(A)} \\
&+ \underbrace{P(A \cap B) + P(B \setminus (A \cap B))}_{P(B)} \\
&- P(A \cap B)
\end{aligned}
$$

■

Example 6 *Tossing a coin:* $\Omega = \{hh, ht, th, tt\}$. *Assume equal probability of* $1/4$.

$$P\left(\textit{head on first tossing}\right)$$

$$\begin{aligned} &= P(\{ht, hh\}) \\ &= P(\{ht\} \cup \{hh\}) \\ &= P(\{ht\}) + P(\{hh\}) \\ &= 1/2 \end{aligned}$$

$$P\left(\textit{head on second tossing}\right)$$

$$\begin{aligned} &= P(\{th, hh\}) \\ &= 1/2 \end{aligned}$$

$$
\begin{aligned}
&P\left(\textit{head on first or second}\right)\\
&\quad = P\left(\underbrace{\textit{head on first}}_{A} \cup \underbrace{\textit{head on second}}_{B}\right)\\
&\quad = P\left(\{ht, hh\} \cup \{th, hh\}\right)\\
&\quad = P\left(\underbrace{\{ht, hh\}}_{A}\right) + P\left(\underbrace{\{th, hh\}}_{B}\right) - P\left(\underbrace{\{hh\}}_{A\cap B}\right)\\
&\quad = 1/2 + 1/2 - 1/4 = 3/4,
\end{aligned}
$$

or directly

$$
\begin{aligned}
&= P\left(\{hh, ht, th\}\right)\\
&= 1/4 + 1/4 + 1/4 = 3/4
\end{aligned}
$$

Example 7 *An extreme case of this rule: if Bassem cannot come except with Ahmed, then* $B \subset A$

$$
\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B)\\
&= P(A)
\end{aligned}
$$

1.4 Counting Methods

- Beneficial for finite sample space:

  $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$

- Denote $P(\omega_i)$ by p_i for simplicity.

- Special case: $p_i = p, |A| = n$, then $P(A) = n/N$

1.4.1 Multiplication Principle

Proposition 8 *Two experiments having m, n outcomes $\Longrightarrow N = mn$ pairs of outcomes.*

Proof. The outcomes N are the entries of $m \times n$ table: $\{a_1, \ldots, a_m\} \times \{b_1, \ldots, b_n\}$ ■

Proposition 9 *p experiments with n_i outcome in experiment $i \Longrightarrow N = \prod_{i=1}^{p} n_i$ total number of outcomes.*

Proof. By induction:
Base step: the statement is true for $p = 2$ (proven above).
Induction step: suppose it is true for $p = q$.Then total number of outcomes $N_q = \prod_{i=1}^{q} n_i$. For $p = q + 1$, we have two experiments one with $\prod_{i=1}^{q} n_i$ and the second with n_{q+1}

$$\begin{aligned} N_{q+1} &= n_{q+1} \times \prod_{i=1}^{q} n_i \\ &= \prod_{i=1}^{q+1} n_i, \end{aligned}$$

which completes the proof. ■

1.4.2 Permutations and Combinations

$C = \{c_1, c_{2,} \ldots, c_n\}$, how many ways to sample r elements:

Ordered sampling with replacement:
$n \times n \times \ldots n = n^r$

Ordered sampling without replacement:
$n \times (n-1) \times \ldots \times (n-r+1)$
special case: if $r = n \Longrightarrow n!$ ways.

Example 10 *If a plate has 3 letters and 3 numbers, we have* $26^3 \times 10^3 = 17,576,000$ *plates!*

Example 11 *If all letters and numbers have the same probability to occur, what is the probability that a plate has no duplicated letter or number?*

$$P(A) = \frac{26 \times 25 \times 24 \times 10 \times 9 \times 8}{17,576,000} = 0.64.$$

What is the probability that a car has, at least, two digits duplicated? Very easy!!!

(Finding P (A^c) is sometimes much easier than P $(A$
because of number of ways.)

Example 12 (Birthday Problem) *Given n persons is the probability that at least two of them have the same birthday? Assume all days have the same probability and each year has only 365 days.*

$$P\left(A^c\right) = \frac{365 \times 364 \times \ldots \times (365 - n + 1)}{365^n}$$

$$P(A) = 1 - \frac{365 \times 364 \times \ldots \times (365 - n + 1)}{365^n}.$$

n	$P(A)$
4	.016
16	.284
23	.507
32	.753
40	.891
56	.988

You may like to think as a "Frequentist"

Example 13 *How many persons must you ask to have .5 chance of finding someone who shares your birthday?*

$$
\begin{aligned}
P\left(A^c\right) &= \frac{364^n}{365^n} \\
P(A) &= 1 - \frac{364^n}{365^n} \\
n &= \frac{\log(1 - P(A))}{\log(364/365)},
\end{aligned}
$$

which gives $n = 253$. *Why it was easier in the above example?*

Unordered sampling without replacement:

$$\frac{n \times (n-1) \times \ldots \times (n-r+1)}{r!} = \frac{n!}{(n-r)!r!} = \binom{n}{r}.$$

It is used in

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

because $a^k b^{n-k}$ is obtained by summing $\binom{n}{k}$ different terms. Also,

$$2^n = \sum_{k=0}^{n} \binom{n}{k},$$

which is the number of all possible subsets.

Example 14 *A manufacture produced n items, t out of them are defective (we do not know them). We sample m items at random; what is the probability that r out of them may be defective. Why we do this?*

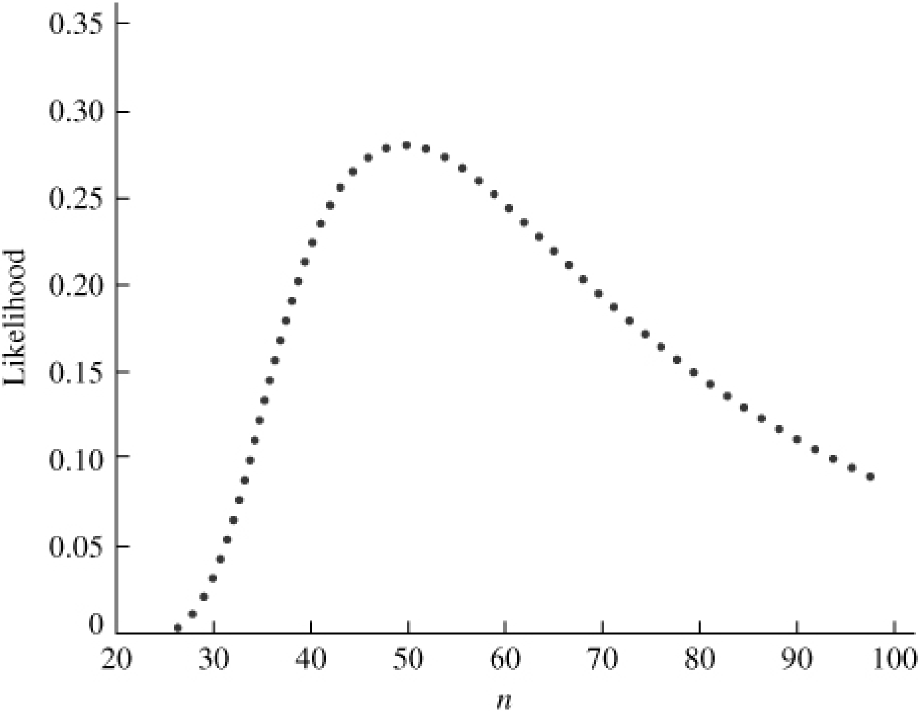
$$P\left(m \text{ defective items}\right) = \frac{\binom{t}{r}\binom{n-t}{m-r}}{\binom{n}{m}}.$$

to be used in estimating t. The following example is similar but it estimated n.

Example 15 (Capture/Recapture Method) *10 animals are captured, tagged and released. Later, 20 animals are captured, 4 of them are found to be tagged. What is the population n?*

$$\begin{aligned} P\left(r \text{ captured}\right) &= \frac{\binom{t}{r}\binom{n-t}{m-r}}{\binom{n}{m}} \\ &= \frac{\binom{10}{4}\binom{n-10}{20-4}}{\binom{n}{20}}. \end{aligned}$$

n must be the value that most probable to happen; then maximize P


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
Likelihood
20
30
40
50
60
70
80
90
100
n

$$
\begin{aligned}
L_n &= \frac{\binom{t}{r}\binom{n-t}{m-r}}{\binom{n}{m}} \\
&= \binom{t}{r}\frac{(n-t)!m!(n-m)!}{(m-r)!(n-t-m+r)!n!}
\end{aligned}
$$

$$
\begin{aligned}
\frac{L_n}{L_{n-1}} &= \frac{(n-t)!(n-m)!}{(n-t-m+r)!n!}\times \\
&\quad \frac{(n-1)!(n-t-m+r-1)!}{(n-t-1)!(n-m-1)!} \\
&= \frac{(n-t)(n-m)}{n(n-t-m+r)}.
\end{aligned}
$$

$\frac{L_n}{L_{n-1}} > 1\ if$

$$
\begin{aligned}
(n-t)(n-m) &> n(n-t-m+r) \\
n^2 - mn - tn + mt &> n^2 - tn - mn + rn \\
mt &> rn \\
n &< \frac{mt}{r}.
\end{aligned}
$$

Then,

$$\begin{aligned}\arg\max_{n}[L_n] &= \frac{mt}{r}\\ &= \frac{20\times 10}{4} = 50,\end{aligned}$$

it makes a lot of sense since 4/20 is related to 10/50.

Proposition 16 *The number of ways to partition n objects into r classes, each with n_i object (such that $\sum_{i=1}^{r} n_i = n$) is*

$$\begin{aligned}&\frac{n!}{n_1!(n-n_1)!}\cdot\frac{(n-n_1)!}{(n-n_1-n_2)!n_2!}\cdots\times\\ &\frac{(n-n_1-n_2-\cdots-n_{r-2})!}{(n-n_1-n_2-\cdots-n_{r-1})!n_{r-1}!}\times 1\\ &=\frac{n!}{n_1!n_2!\cdots n_r!}\end{aligned}$$

This is similar to

$$(x_1+x_2+\cdots+x_r)^n = \sum_{n_1,\cdots,n_r}\binom{n}{n_1\cdots n_2}x_1^{n_1}x_2^{n_2}\cdots x_r^{n_r}.$$

Very important: the proposition above assume **ordering between groups (inter)** and **non-orderi within a group (intra)**. If $\Omega = \{A, B, C, D\}$ $(n = 4)$, $n_1 = 2, n_2 = 2$,then selections are $\binom{4}{2}\binom{4-2}{2}$

$$\left\{ \begin{array}{c} \underline{(\{A, B\}, \{C, D\})}, \underline{\underline{(\{A, C\}, \{B, D\})}}, \\ (\{A, D\}, \{B, C\}), (\{B, C\}, \{A, D\}), \\ \underline{\underline{(\{B, D\}, \{A, C\})}}, \underline{(\{C, D\}, \{A, B\})} \end{array} \right\}$$

If we want non-ordering for both between-group and with-group

$$\left(\frac{n!}{n_1! n_2! \cdots n_r!} \right) / r!,$$

where r is the number of groups (2 above), and $r!$ is the number of ways to order r groups.

For more elaboration on counting methods, pleas refer to (Rosen, 2007, Ch. 5 and 7).

1.5 Conditional Probability

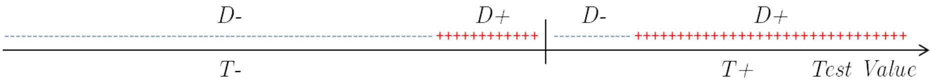
$T+$: high blood concentration (+ve test)

$T-$: low blood concentration (-ve test)

$D+$: toxicity (disease present)

$D-$: no toxicity (disease absent)

#	$D+$	$D-$	$Total$
$T+$	25	14	39
$T-$	18	78	96
$Total$	43	92	135



Conditioning on $T+$ changes Ω to $T+$.

$$P(D+|T+) = \frac{\#(D+\cap T+)}{\#(T+)} = \frac{25}{39}$$

$$
\begin{aligned}
P(D+|T+) &= \frac{\#(D+\cap T+)}{\#(T+)} = \frac{25}{39}\\
&= \frac{\#(D+\cap T+)/Total}{\#(T+)/Total} = \frac{25/135}{39/135}\\
&= \frac{P(D+\cap T+)}{P(T+)}.
\end{aligned}
$$

P	$D+$	$D-$	$Total$
$T+$	25/135	14/135	39/135
$T-$	18/135	78/135	96/135
$Total$	43/135	92/135	1

Definition 17 *A and B are events,* $P(B) \neq 0$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

- *B acts as the new* Ω.
- $P(A|B) =$ *normalized version of* $P(A \cap B)$.
- *This is a new probability measure; does it satisfy the axioms?*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

$$P(A|B) = \frac{+ve}{+ve} \geq 0.$$

$$P(B|B) = \frac{P(B \cap B)}{P(B)} = 1.$$

$$\begin{aligned}
P\left(\left(\bigcup_{i=1}^{\infty} A_i\right)|B\right) &= \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)} \\
&= \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)} \\
&= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} \\
&= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \\
&= \sum_{i=1}^{\infty} P(A_i|B).
\end{aligned}$$

Multiplication Law

$$P(A \cap B) = P(A|B)\,P(B)\,,\ P(B) \neq 0,$$
$$P(A \cap B) = P(B|A)\,P(A)\,,\ P(A) \neq 0.$$

What is the meaning of that?

Example 18 *An Urn contains 3 reds and one blue. What is the probability of selecting two red balls without replacement.*

Solution 1:

$$P(2\ reds) = \frac{\#\ ways\ of\ selecting\ 2\ reds}{\#\ ways\ of\ selecting\ 2\ balls} = \frac{\binom{3}{2}}{\binom{4}{2}} = \frac{1}{2}.$$

Solution 2:

$$P(2\ reds) = P(R_1)\,P(R_2|R_1) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

Lemma 19 (Law of Total Probability) *Let $B_i, i = 1, \ldots, n$ be such that $B_i \cap B_j = \phi\ \forall i \neq j$, $\bigcup_{i=1}^{n} B_i = \Omega$ (B_i **partition** Ω), and $P(B_i) \neq 0$. Then*

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

Proof.

$$\begin{aligned}
P(A) &= P(A \cap \Omega) \\
&= P\left(A \cap \bigcup_{i=1}^{n} B_i\right) \\
&= P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) \\
&= \sum_{i=1}^{n} P(A \cap B_i) \\
&= \sum_{i=1}^{n} P(A|B_i)\, P(B_i).
\end{aligned}$$

■

Example 20 *What is the probability that a red ball is selected on the second draw?*

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1)$$
$$= \frac{2}{3}\cdot\frac{3}{4} + \frac{3}{3}\cdot\frac{1}{4} = \frac{3}{4}.$$

Example 21 (Occupational Mobility) *This statistics is collected by Glass and Hall (1954).*

	U_2	M_2	L_2
U_1	.45	.48	.07
M_1	.05	.70	.25
L_1	.01	.50	.49

- U, M, L : *Upper, Middle, Lower levels.*
- $1, 2$*: Father, Son.*
- *ex:* $P(U_2|U_1) = .45 =$ *Probability a son occupies upper level after his father.*
- *Notice that* $P(M_2|L_1) > P(U_2|M_1)$.

- *What is $P(U_2)$, assuming that fathers occupations are 10%, 40%, 50% in U, M, L respectively.*

$$
\begin{aligned}
P(U_2) = {} & P(U_2|U_1)\,P(U_1) + P(U_2|M_1)\,P(M_1) \\
& + P(U_2|L_1)\,P(L_1) \\
& = .45 \times .1 + .05 \times .4 + .01 \times .5 \\
& = .07
\end{aligned}
$$

- *Notice that each column does not necessarily sum to one! Why? Where is the partition.*

- *Inverse problem: what is $P(U_1|U_2)$? This is Bayes's rule:*

Lemma 22 (Bayes' Rule) *Let A and $B_i, i = 1, \ldots, n$ be events, B_is are disjoint,* ***partition*** *Ω, and $P(B_i) \neq 0$. Then*

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

Proof.

$$\begin{aligned} P(B_i|A) &= \frac{P(A \cap B_i)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}. \end{aligned}$$

The proof is complete. ∎

Example 23 (polygraph test) *This is a lie-detector test:*

$+$: *polygraph +ve (+ve test)*

$-$: *polygraph -ve (-ve test)*

T : *Person telling truth*

L : *Person lying*

Based on Gastwirth (1987):

P	$\mid T$	$\mid L$
$+$	.14	.88
$-$	.86	.12
Sum	1	1

- *Why* $.14 + .88 \neq 1$ *and* $.86 + .12 \neq 1$ *while* $.14 + .86 = .88 + .12 = 1$?

- *What is the meaning of* $P(+|T) = .14$?

Now suppose that the majority of people are telling truth for a particular question, i.e., $P(T) = .99$. *What is the probability that a person is telling the* *truth even if the polygraph is +ve?*

$$P(T|+) = \frac{P(+|T)\,P(T)}{P(+|T)\,P(T) + P(+|L)\,P(L)}$$
$$= \frac{.14 \times .99}{.14 \times .99 + .88 \times .01} = .94$$

So, most of the innocent people in a screening setup will be placed under suspicion!!

The source of the problem is:

$$P(T|+) = \frac{1}{1 + \frac{P(+|L)P(L)}{P(+|T)P(T)}}$$
$$= \frac{1}{1 + \frac{.88\times.01}{.14\times.99}},$$

to have small $P(T|+)$: either $\frac{P(+|L)}{P(+|T)}$ or $\frac{P(L)}{P(T)}$ should be large.

When does a very naive test ($P(+|T) = P(+|L)$) give low $P(T|+)$?

Bayes' Rule in real life: more intuition

- $P(H), P\left(\overline{H}\right)$ $(= 1 - P(H))$: a prior probabilities (prior knowledge).
- H : Hypothesis.
- $E, P(E)$: Evidence and its probability.
- $P(E|H), P\left(E|\overline{H}\right)$: Likelihoods.
- $P(H|E)$: A Posterior Probability.

$$
\begin{aligned}
P(H|E) &= \frac{P(E|H)\,P(H)}{P(E)} \\
&= \frac{P(E|H)\,P(H)}{P(E|H)\,P(H) + P\left(E|\overline{H}\right)P\left(\overline{H}\right)} \\
&= \frac{1}{1 + \frac{P\left(E|\overline{H}\right)}{P(E|H)} \cdot \frac{1-P(H)}{P(H)}}
\end{aligned}
$$

Example from "Pattern Recognition"

- reading mammograms to find Breast Cancer. (H)
- probability of a woman having cancer = 1% ($P(H)$)
- if having a breast cancer 80% of radiologists observe it ($P(+|H)$)
- if having benign lesion 10% radiologists misinterpret it as a cancer
- how much, do you think (subjectively), is $P(H|+)$?

- however, when radiologists were asked abou $P(H|+)$ majority of them estimated it subjectively as 75%, although the Bayes' rule give:

$$P(H|+) = \frac{P(+|H)\,P(H)}{P(+|H)\,P(H) + P\left(+|\overline{H}\right)P\left(\overline{H}\right)} = \frac{.8 \times .01}{.8 \times .01 + .1 \times .99} = 7.5\% \quad !!!$$

1.6 Independence

Motivation towards independence:

$$\begin{aligned} P(A) &= P(A|B) \\ &= \frac{P(A\cap B)}{P(B)}, \text{ then} \\ P(A\cap B) &= P(A)P(B). \end{aligned}$$

Definition 24 *Events A and B are said to be independent if* $P(A\cap B) = P(A)P(B)$

Are disjoint events independent?

Definition 25 (Independence generalized) *Events* $A_i, i = 1\ldots, n$ *are said to be independent (or mutually independent) if for any subcollection* $A_{i_1}, \ldots,$

$$P\left(\bigcap_{j=1}^{m} A_{i_j}\right) = \prod_{j=1}^{m} P\left(A_{i_j}\right),$$

for which the definition above is a special case.

Example 26 *A backup system:*

- *has 3-mirror hard drives.*
- *the probability that one fails (F) is p.*
- *what is the probability that the system fail?*

$$\begin{aligned} P\left(\textit{system fail}\right) &= P\left(F_1 \cap F_2 \cap F_3\right) \\ &= P\left(F_1\right) P\left(F_2\right) P\left(F_3\right) = p^3. \end{aligned}$$

If $p = .001$ *then* $p^3 = 10^{-9}$.

Definition 27 (Weaker Independence) *Events A_i $1\ldots,n$ are said to be pair-wise independent if*

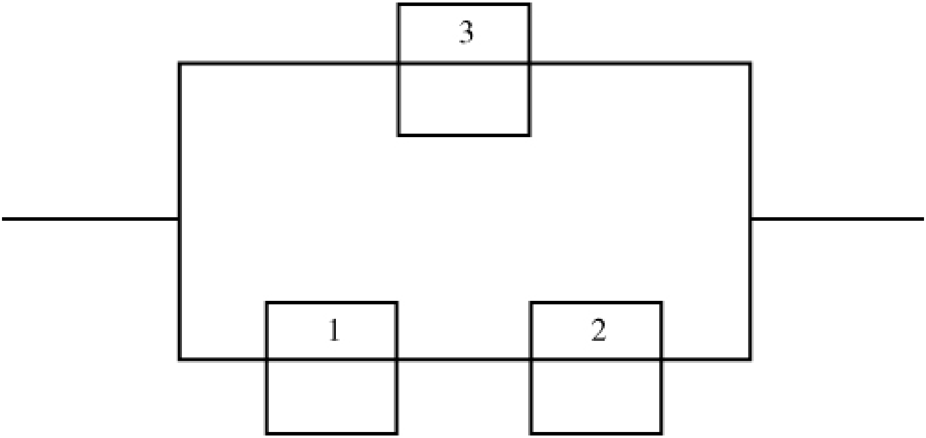
$$P\left(A_i \cap A_j\right) = P\left(A_i\right) P\left(A_j\right)\ \forall i \neq j.$$

Example 28 (counter example) *Tossing coin twice: A_1 = head on first, A_2 =head on second, A_3 = just one head in both. Clearly:*

$$\begin{aligned} P\left(A_i\right) &= \frac{1}{2}, \\ P\left(A_1 \cap A_2\right) &= P\left(\{hh\}\right) = \frac{1}{4} = P\left(A_1\right) P\left(A_2\right), \\ P\left(A_1 \cap A_3\right) &= P\left(\{ht\}\right) = \frac{1}{4} = P\left(A_1\right) P\left(A_3\right), \\ P\left(A_2 \cap A_3\right) &= P\left(\{th\}\right) = \frac{1}{4} = P\left(A_2\right) P\left(A_3\right); \end{aligned}$$

hence, they are pairwise independent. But,

$$P\left(\bigcap_{i=1}^{3} A_i\right) = P\left(\phi\right) = 0 \neq \prod_{i=1}^{3} P\left(A_i\right).$$



Example 29 *Consider the circuit in the figure. Ever* A_i *is "*i^{th} *relay works".* A_i *are independent with potability* p *each. Then, the event* C = *"passing current" has*

$$
\begin{aligned}
P(C) &= P(A_3 \cup (A_1 \cap A_2)) \\
&= P(A_3) + P(A_1 \cap A_2) - P(A_1 \cap A_2 \cap A_3) \\
&= p + p^2 - p^3.
\end{aligned}
$$

Chapter 2

Random Variables (r.v.)

A r.v. is essentially a number as:

Definition 30 *A r.v. X is a mapping from Ω to* $\mathbb{R} \cup \{\infty, -\infty\}$

$$X : \Omega \to [-\infty, \infty].$$

Example 31 *Consider*

$$\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$$

- *What is the number of heads?*
- *What is the number of tails?*

2.1 Discrete r.v.

Definition 32 *Consider the (infinite) partition* A_i, $1,\ldots,n$ *of* Ω. *We call* X *a discrete r.v. if it takes a value* $x_k \in [-\infty,\infty]$, $k = 1,\ldots,n$ *whenever* A_k *happens. More compactly*

$$X = \sum_{i=1}^{n} x_i I_{A_i},$$

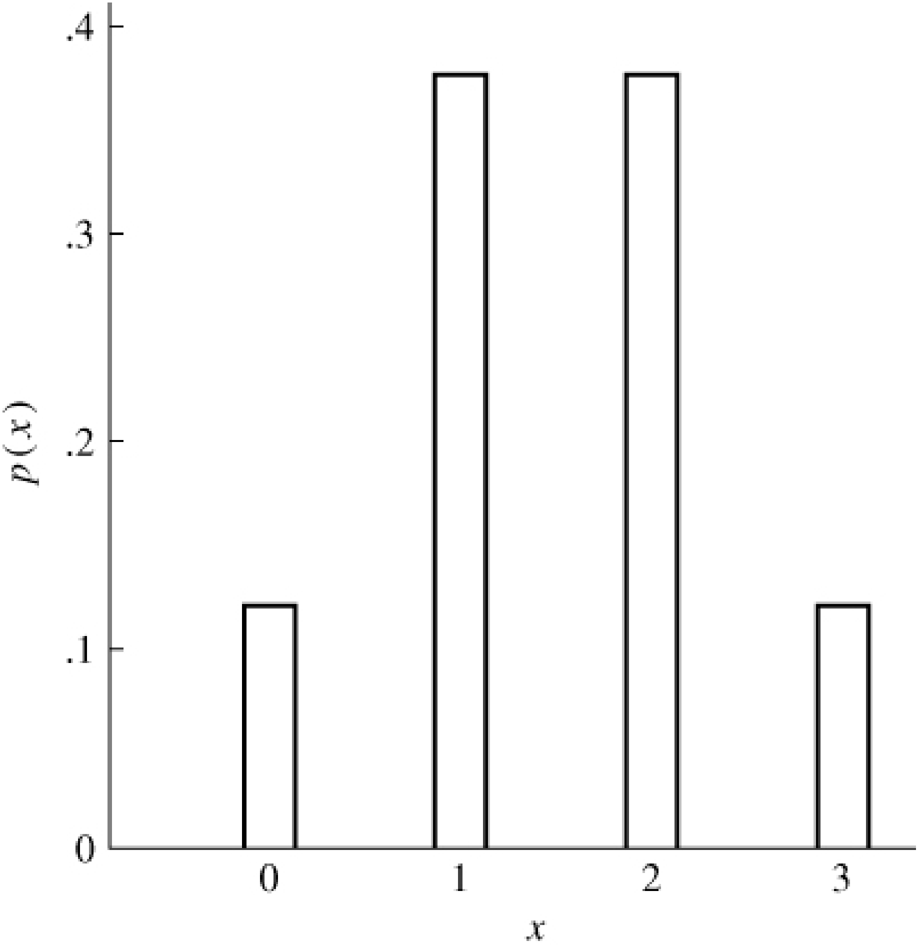
$$I_{A_i} = \begin{cases} 1, & if\, \omega \in A_i \\ 0, & if\, \omega \notin A_i \end{cases}$$

The probability of the discrete r.v. is called Probability Mass Function (pmf).

Example 33 *In the previous example, assume the coin is fair, and X is the number of heads. What is the probability mass function (pmf) of X?*

$$\begin{aligned}
0 \le X &\le 3,\\
P(X=0) = P(\{ttt\}) &= \frac{1}{8},\\
P(X=1) = P(\{htt, tht, tth\}) &= \frac{3}{8},\\
P(X=2) = P(\{thh, hth, hht\}) &= \frac{3}{8},\\
P(X=3) = P(\{hhh\}) &= \frac{1}{8}.
\end{aligned}$$

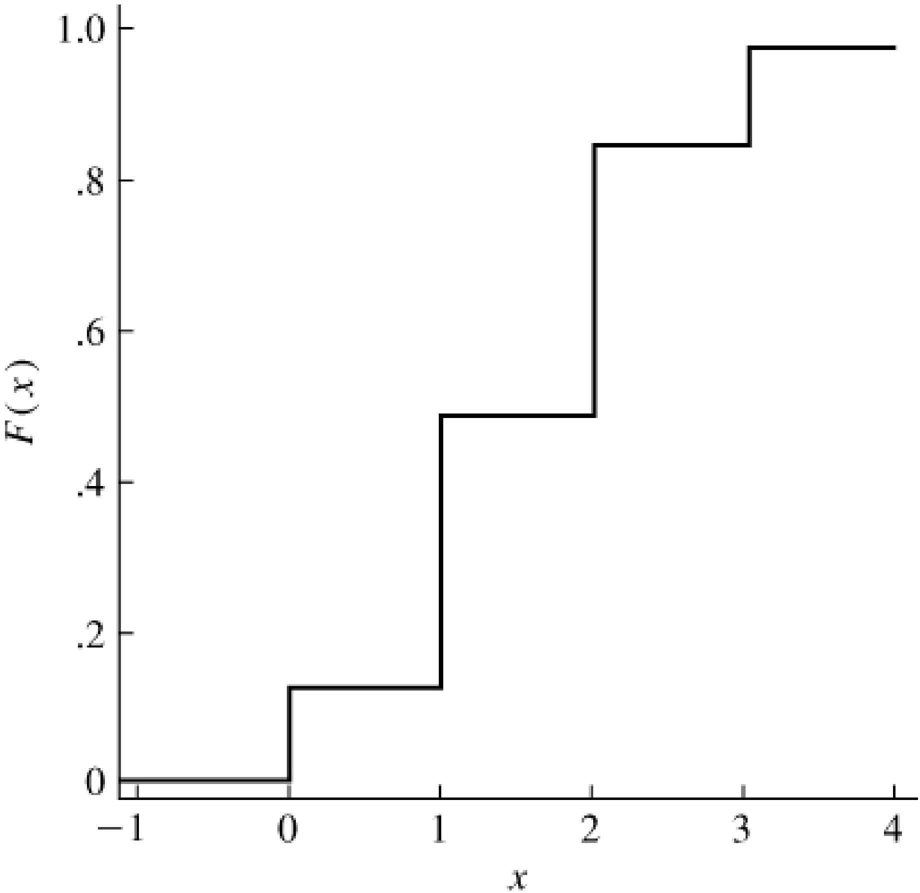
What is $P(X=4)$? Prove!


.4
.3
.2
.1
0
$p(x)$
0
1
2
3
x

Definition 34 *The cumulative density function (cdf) (sometimes called distribution function (df)) is defined as*

$$F(x) = P(X \leq x).$$

We may right F_X not to confuse with, e.g., F_Y.



Lemma 35 *Any cdf F has:*

- $F(x_1) \le F(x_2)\ \forall x_1 \le x_2$ *(monotonically non-decreasing),*
- $F(-\infty)\ (=\lim_{x\to-\infty} F(x)) = 0,$
- $F(\infty)\ (=\lim_{x\to\infty} F(x)) = 1,$
- $F(x) = F\left(x^{+}\right)$ *(continuous from the right)*

Proof. omitted ∎

2.1.1 $(Bernoulli(p))$

The r.v. X is Bernoulli if

$$P(X) = \begin{cases} p & X = 1 \\ 1-p & X = 0 \end{cases}.$$

Note that

$$P(1) + P(0) = 1.$$

If tossing a fair coin and the event $A = \{\text{head}\}$ then $I_A \sim Bernoulli(0.5)$

- Plot the pmf of $(Bernoulli(p))$
- Repeating an experiment many times under this pmf, how data looks like?

2.1.2 $(Binomial(n,p))$

If X is the number of successes of n trials, each with probability p, then

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ 0\leq k\leq n.$$

Note that

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}&=((p)+(1-p))^n\\&=1\end{aligned}$$

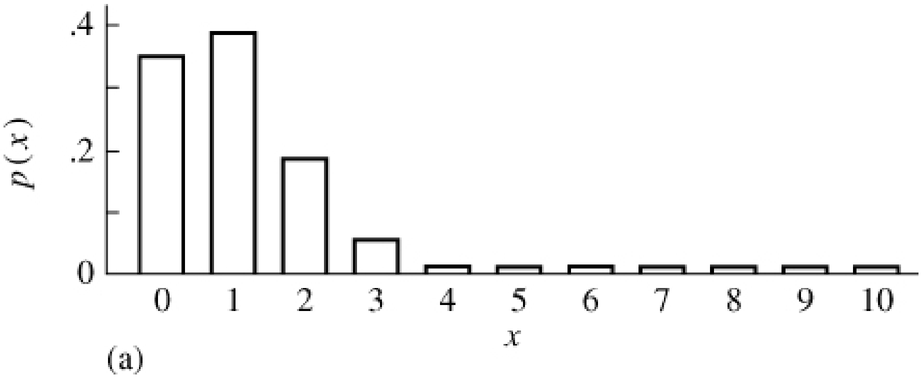
Also, observe that

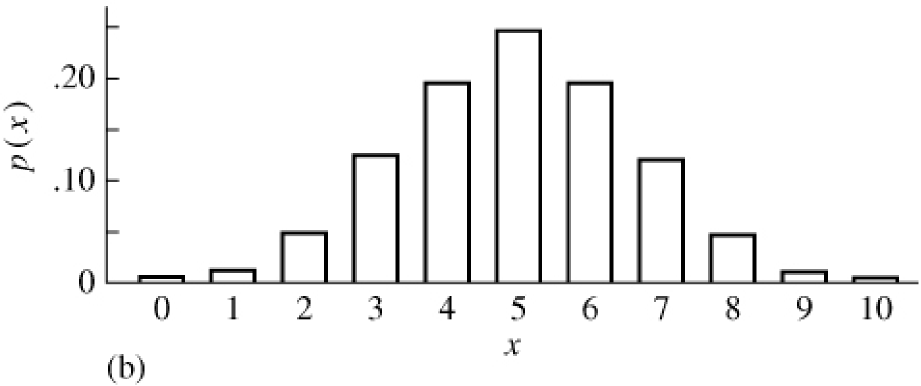
$$X=\sum_{i=1}^{n}I_i,$$

where $I_i\sim Ber(p)$, and I_is are independent (def. coming soon).

- **A plot for pmf functions of** $Bin(10,p)$, $p = .1, .5$ **respectively:**

- Repeating an experiment many times under one of those pmf's, how data looks like?





Example 36 (Tay-Sachs disease) *:*

- *common disease of Jewish or eastern European extraction.*
- *If a couple are both carriers, one of every four children of theirs is a carrier.*
- *If they have four children, then*

$$P(k) = \binom{4}{k} (.25)^k (1 - .25)^{4-k},\ 0 \le k \le 4$$

k	$P(k)$
0	.316
1	.422
2	.211
3	.047
4	.004

Example 37 (Error Correction) *:*

- *error in sending single bit is .1*
- *remedy: the system will send it 5 times*
- *the receiver takes a majority vote*
- *the probability of receiving one bit correctly is:*

$$P(\#errors \leq 2) = \sum_{k=0}^{2} \binom{n}{k} (p)^k (1-p)^{n-k}$$
$$= .9914$$

2.1.3 $\left(Geometric(p)\right)$

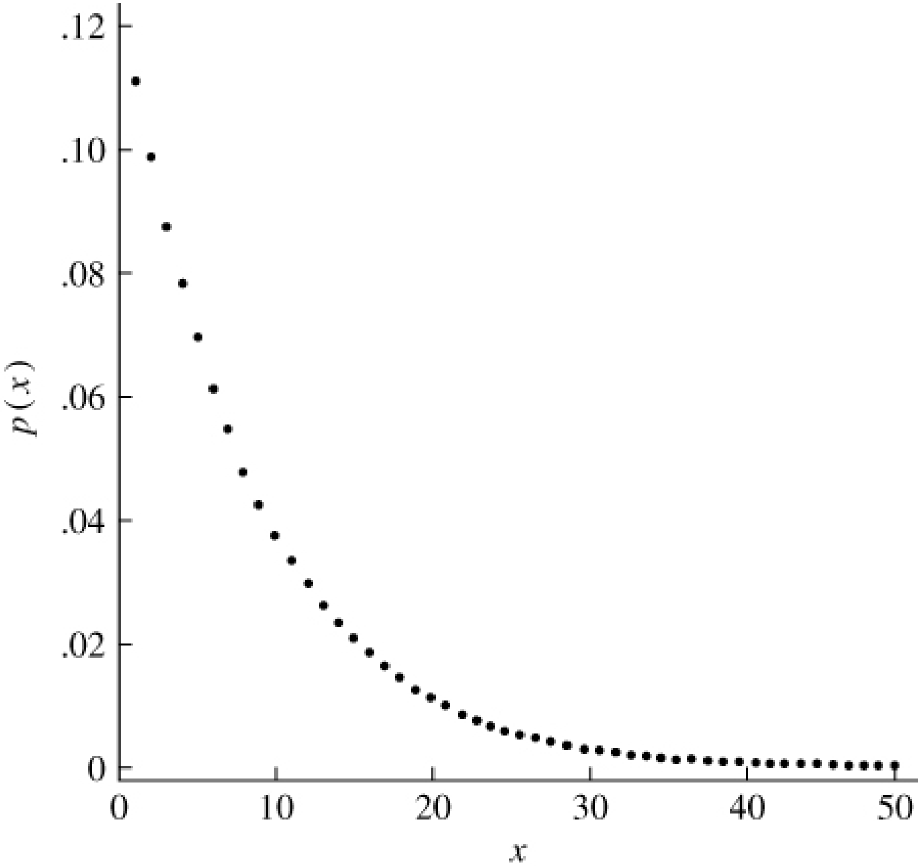
Constructed from infinite independent Bernoulli trials. $X = \#$ of trials to get a first success; hence:

$$P(X = k) = (1 - p)^{k-1} p,\ 1 \leq k$$

Note that

$$\begin{aligned}\sum_{k=1}^{\infty} p(1 - p)^{k-1} &= p \sum_{k=1}^{\infty} (1 - p)^{k-1} \\ &= p \frac{1}{p} \\ &= 1\end{aligned}$$

Example 38 *If the probability of wining for a successful draw is $p = 1/9$, then number of draws necessary for wining is $Geometric(1/9)$*



$(NBinomial(r, p))$

Generalization to Geometric distribution. $X = \#$ of trials necessary to get first r success. Then, last trial should be successful; and

$$P(X = k) = \underbrace{p}_{last\ trial} \times \underbrace{\binom{k-1}{r-1} p^{r-1} (1-p)^{(k-1)-(r-1)}}_{(r-1)\ success\ in\ (k-1)\ trial}$$

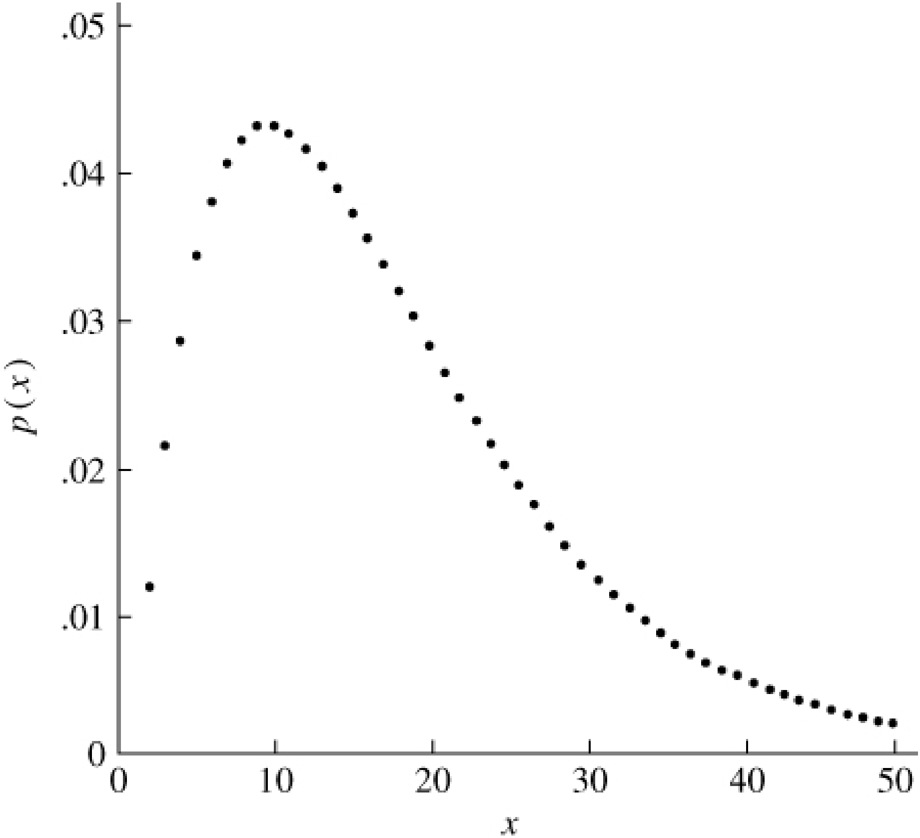
$$= \binom{k-1}{r-1} p^{r} (1-p)^{k-r}, \ r \le k.$$

Prove that

$$\sum_{k=r}^{\infty} \binom{k-1}{r-1} p^{r} (1-p)^{k-r} = 1.$$

Example 39 *Continuing the previous example, number of draws necessary to get the second wining draw is $NBin(2, 1/9)$. Then*

$$P(k) = (k-1)(1/9)^2 (8/9)^{k-2}$$



.05
.04
.03
.02
.01
0
$p(x)$
0
10
20
30
40
50
x

2.1.4 $Hypergeometric(n, r, m)$

Recall the capture/recapture method. Draw m objects from total of n with r labeled; then X is the number of selected labeled objects:

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}, \; 0 \le k \le m$$

Prove that (was a HW problem)

$$\sum_{k=0}^{m} \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}} = 1.$$

2.1.5 $Poisson(\lambda)$

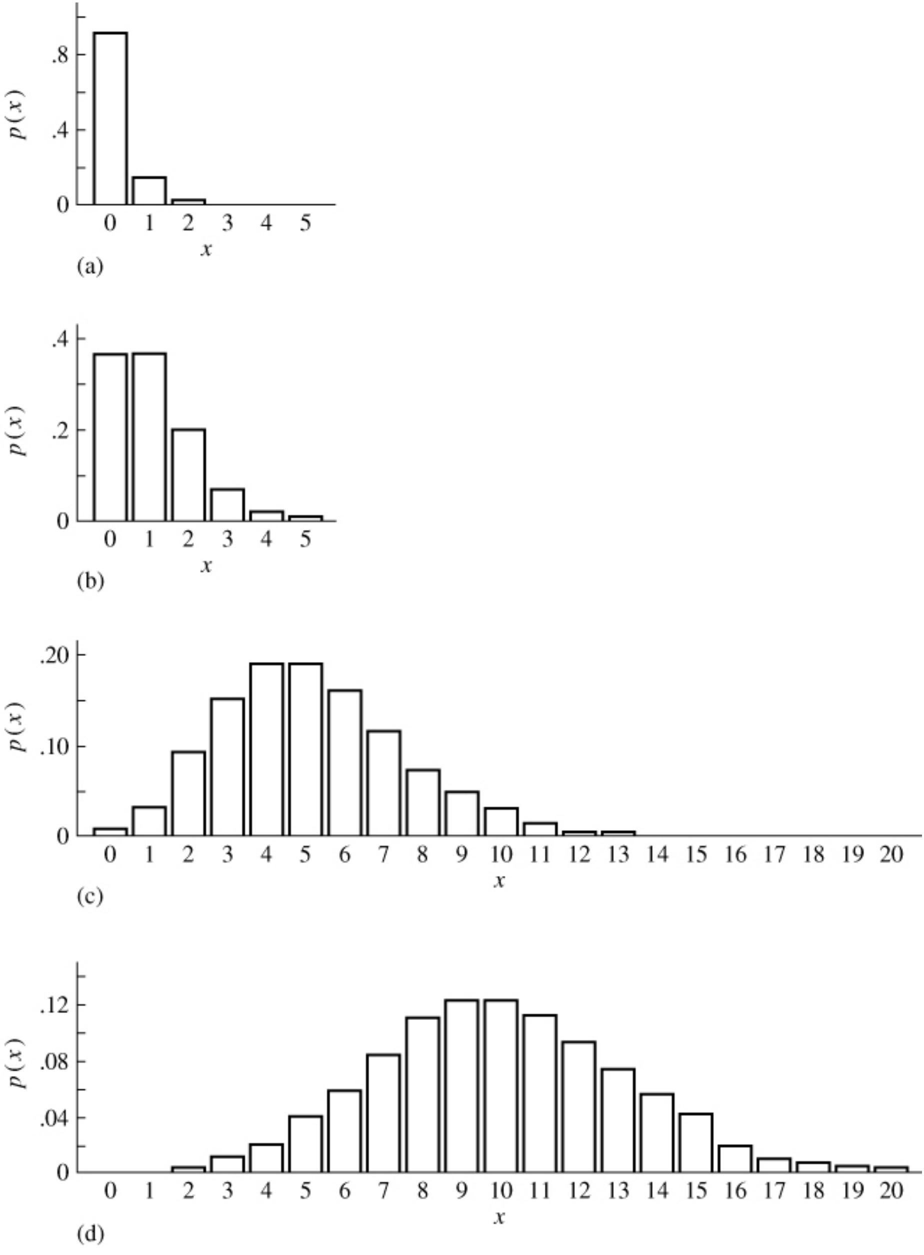
Poisson r.v. X is defined as

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \, 0 \le k.$$

Note that

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} &= e^{\lambda} e^{-\lambda} \\ &= 1.\end{aligned}$$

The following is a plot for the Poisson pmf at $\lambda = .1, \; 1, \; 5, \; 10$ respectively.


p(x)
.8
.4
0
0 1 2 3 4 5
x
(a)
p(x)
.4
.2
0
0 1 2 3 4 5
x
(b)
p(x)
.20
.10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x
(c)
p(x)
.12
.08
.04
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x
(d)

The Poisson r.v. can be seen as a limiting process for the Binomial, with $n \to \infty,\ p \to 0,\ np = \lambda$.

$$
\begin{aligned}
P(k) &= \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!\,(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \left(\frac{\lambda^k}{k!}\right) \underbrace{\left(\frac{n!}{(n-k)!\,n^k}\right)}_{\to 1} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1},
\end{aligned}
$$

$$
\lim_{n \to \infty} p(k) = \frac{\lambda^k}{k!} e^{-\lambda}
$$

Example 40 *In dice rolling, with $n = 100$ rolling, $P((6,6)) = 1/36$. If we consider Poisson approximation with $np = 100 \times \frac{1}{36} = 2.78$*

k	$Bionmail$ Pr	$Poisson$ Pr
0	*.0596*	*.0620*
1	*.1705*	*.1725*
2	*.2414*	*.2397*
3	*.2255*	*.2221*
4	*.1564*	*.1544*
5	*.0858*	*.0858*
6	*.0389*	*.0398*
7	*.0149*	*.0158*
8	*.0050*	*.0055*
9	*.0015*	*.0017*
10	*.0004*	*.0005*
11	*.0001*	*.0001*

Poisson r.v. and continuous time:

- Consider independent events in T time.
- subdivide T to many Δt.
- an event happens in Δt with small p.
- two events cannot happen in the same Δt

Applications:

- Modeling traffic in general:
  - incoming calls in telephone systems.
  - light traffic (but cars are not independent).
- modeling alpha particle emission.

Example 41 (Telephone calls) :

- *Modeling the calls as* $Poisson(\lambda)$.
- $\lambda = .5/\min$.

$$\begin{aligned} P(k) &= \frac{\lambda^k}{k!}e^{-\lambda} \\ P(no\ calls/\min) &= P(0) \\ &= e^{-.5} \\ &= 0.607, \\ P\left(just\ one\ call/\min\right) &= P(1) \\ &= .5e^{-.5} \\ &= 0.303 \end{aligned}$$

Notice that: if the period is, e.g., 2 min then

$$X = X_{1\min} + X_{1\min}$$

we will prove later that X *will be* $Poisson(\lambda_1 + \lambda_2)$.

2.2 Continuous r.v.

Instead of the pmf we define the Probability Density Function (pdf) so that:

$$P(a < X < b) = \int_a^b f(x)\ dx.$$

Notice that:

$$\int_{-\infty}^{\infty} f(x)\ dx = 1$$

$$\begin{aligned} P(X = c) &= \int_c^c f(x)\ dx \\ &= 0, \end{aligned}$$

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

More elaboration:

$$\begin{aligned} P\left(x - \frac{\delta}{2} \leq X \leq x + \frac{\delta}{2}\right) &= \int_{x-\delta/2}^{x+\delta/2} f(u)\ du \\ &\approx \delta f(x), \end{aligned}$$

Equivalently:

$$P(x \leq X \leq x + dx) = f(x)\ dx.$$

CDF:

$$F(x) = \int_{-\infty}^{x} f(u)\ du.$$

In very theoretical probability, what is defined first is F as before

$$F(x) = P(X \le x)$$

then if it is differentiable, the density f is defined as:

$$f(x) = F'(x).$$

It is clear that

$$\begin{aligned} P(a \le X \le b) &= F(b) - F(a) \\ &= \int_{a}^{b} f(x)\ dx. \end{aligned}$$

Inverse of CDF (F^{-1})

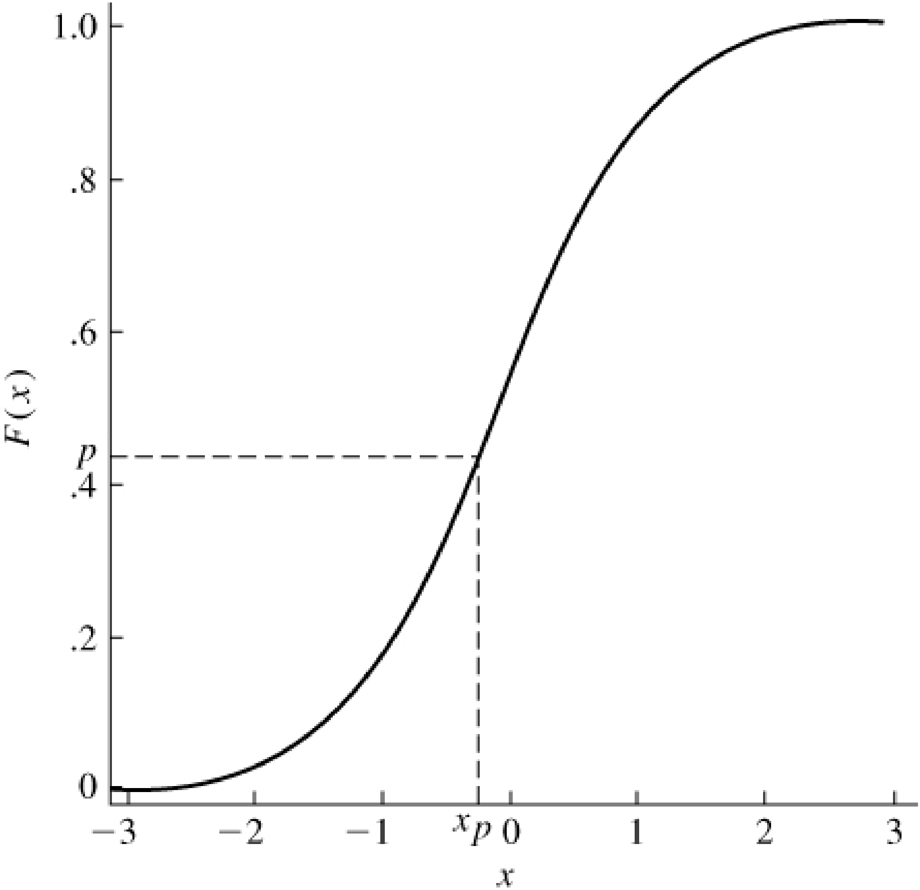
Definition 42 *The p^{th} quantile is defined as, the value x_p of the r.v. that satisfies $F\left(x_p\right) = p$.*

- If F is monotonically (strictly) increasing, the pth quantile is unique (see figure).
- $F^{-1}(.5)$ is the median.
- $F^{-1}(.25)$ and $F^{-1}(.75)$is the lower and upper quartile.

Example 43 *Suppose*

$$F(x) = x^2,\ 0 \le x \le 1,$$

$$\begin{aligned}
x_p^2 &= p, \\
x_p &= \sqrt{p}, \\
x_{.5} &= \sqrt{.5} = .707 \\
x_{.25} &= \sqrt{.25} = .5 \\
x_{.75} &= \sqrt{.75} = .866
\end{aligned}$$



1.0
.8
.6
$F(x)$
p
.4
.2
0
−3
−2
−1
x_p
0
1
2
3
x

$Uniform(a,b)$

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & otherwise \end{cases}$$

- Finite support ($\int_{-\infty}^{\infty} f(t)\ dt \stackrel{?}{=} 1$)
- Noninformative distribution.
- CDF:

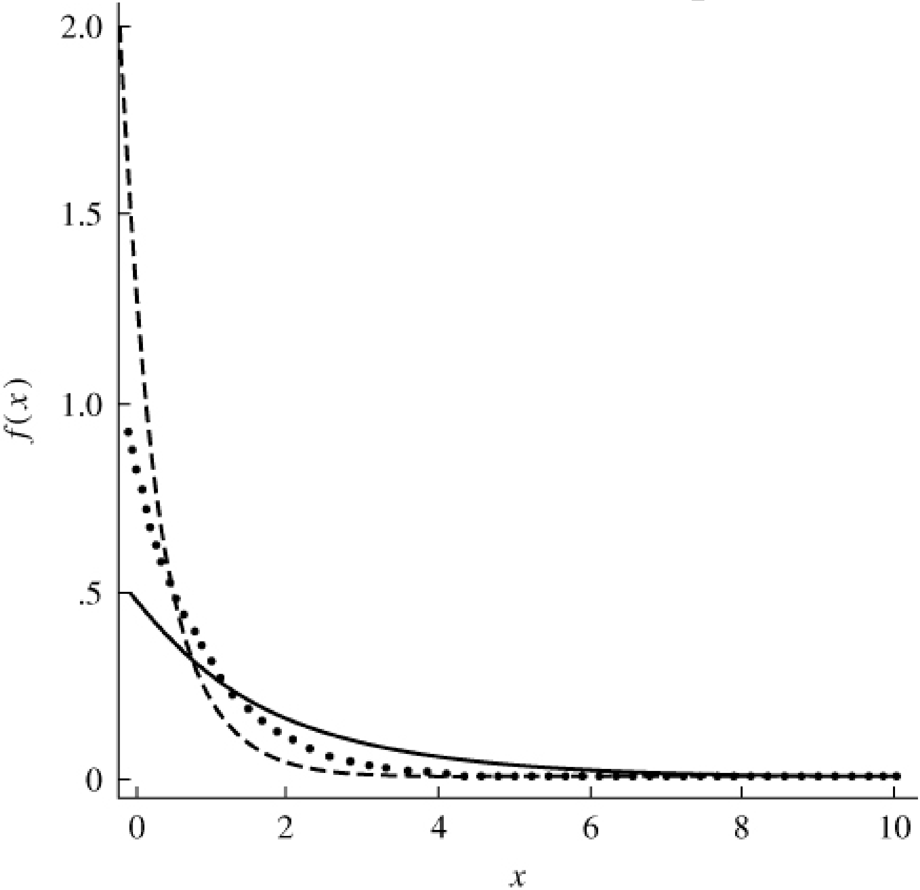
$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1. \end{cases}$$

- Plot the pmf of $\left(Uniform(0,1)\right)$
- Repeating an experiment many times under this pmf, how data looks like?
- HW: write and draw a discrete version of $Uniform(n)$, $X = 1, \ldots, n$. Write down 30 numbers drawn from $Uniform(10)$ using your mind first then using Matlab. Bring them next time ! :)

2.2.1 $Exponential(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$

Note the effect of λ (the rate) **on the pdf.**



$$\int_{-\infty}^{\infty} f(t)\ dt \stackrel{?}{=} 1$$

CDF:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)\ du \\ &= \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \end{aligned}$$

Median:

$$\begin{aligned} 1 - e^{-\lambda x_{.5}} &= \frac{1}{2}, \\ x_{.5} &= \frac{\log 2}{\lambda}. \end{aligned}$$

Application: modeling life time T

$$
\begin{aligned}
P(T > t+s | T > s) &= \frac{P(T > t+s \cap T > s)}{P(T > s)} \\
&= \frac{P(T > t+s)}{P(T > s)} \\
&= \frac{1-\left(1-e^{-\lambda(t+s)}\right)}{1-\left(1-e^{-\lambda s}\right)} \\
&= e^{-\lambda t},
\end{aligned}
$$

Notice that:

- $e^{-\lambda t} = P(T > t)$ (unconditional)

- $e^{-\lambda t}$ is not a function of s !!! This is called **memoryless property**

- This is not good modeling for human life time. Why?

- Probably good for electronic components.

Connection with $Poisson(\lambda)$:

- t_0 : time of an event.
- T : the time between two successive events
- λ: events per unit time ($Possion$ parameter).

$$
\begin{aligned}
P(T > t) &= P(no\ events\ in\ (t_0, t_o + t)) \\
&= \left.\frac{(\lambda t)^k e^{-(\lambda t)}}{k!}\right|_{k=0} \\
&= e^{-\lambda t}.
\end{aligned}
$$

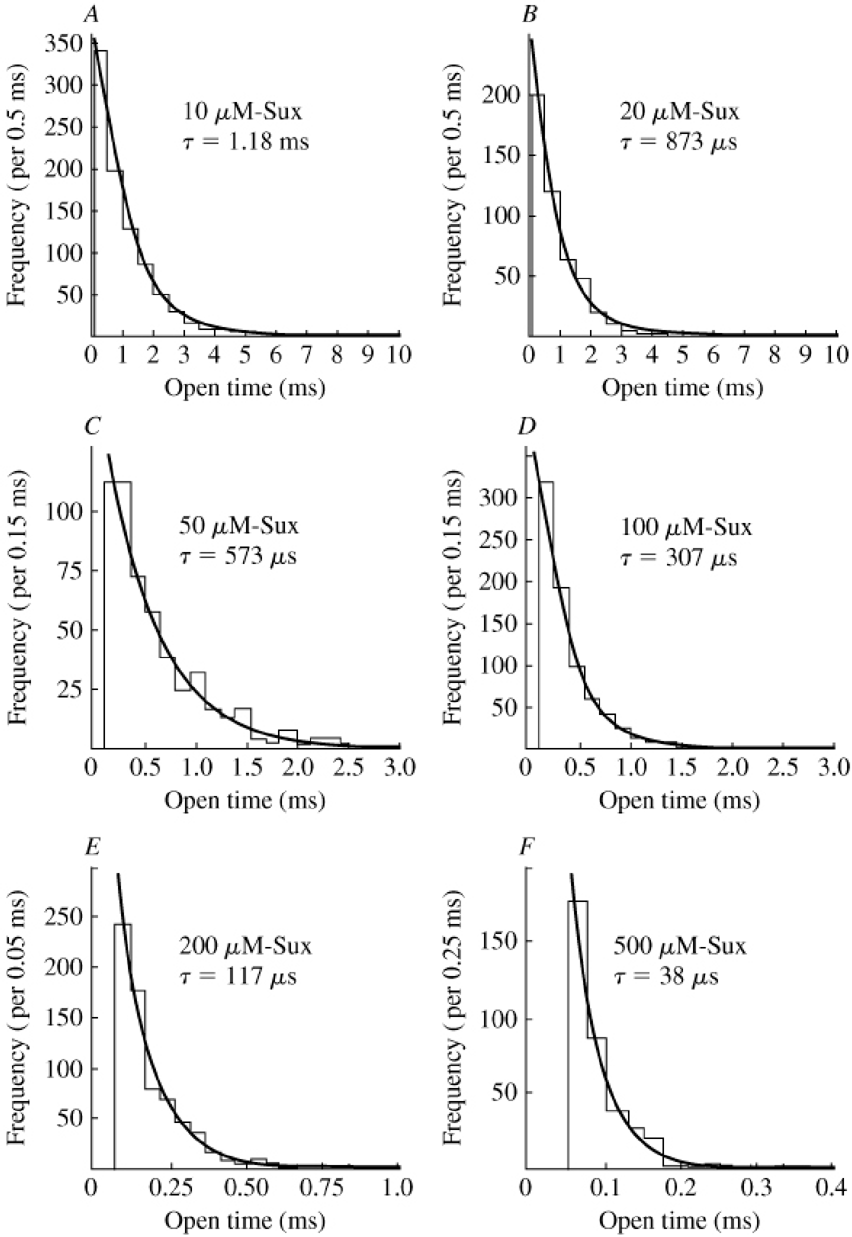
Example 44 (Marshall et al. (1990)) :

- *Blocking frogs nerve channel (passing current) using M-Sux drug.*
- *Drug closes the channel: T is the duration of opening.*
- *Effect of drug on opening time.*
- *Set* $\lambda = 1/\tau$

$$f(t) = \frac{1}{\tau} e^{-t/\tau}$$

Observations

- $T \sim Exponential(\lambda)$
- $\tau\ (= 1/\lambda)$ *increases with amount of M-Sux.*


A
10 μM-Sux
τ = 1.18 ms
Frequency (per 0.5 ms)
350
300
250
200
150
100
50
0 1 2 3 4 5 6 7 8 9 10
Open time (ms)
B
20 μM-Sux
τ = 873 μs
Frequency (per 0.5 ms)
200
150
100
50
0 1 2 3 4 5 6 7 8 9 10
Open time (ms)
C
50 μM-Sux
τ = 573 μs
Frequency (per 0.15 ms)
100
75
50
25
0 0.5 1.0 1.5 2.0 2.5 3.0
Open time (ms)
D
100 μM-Sux
τ = 307 μs
Frequency (per 0.15 ms)
300
250
200
150
100
50
0 0.5 1.0 1.5 2.0 2.5 3.0
Open time (ms)
E
200 μM-Sux
τ = 117 μs
Frequency (per 0.05 ms)
250
200
150
100
50
0 0.25 0.50 0.75 1.0
Open time (ms)
F
500 μM-Sux
τ = 38 μs
Frequency (per 0.25 ms)
150
100
50
0 0.1 0.2 0.3 0.4
Open time (ms)

$Exponential \Longleftrightarrow$ **Memorylessnes**

$$P(T > t + s | T > s) \overset{set}{=} P(T > t) \;\; \forall s \geq 0.$$

$$\frac{P(T > t + s \cap T > s)}{P(T > s)} = P(T > t)$$

$$\frac{P(T > t + s)}{P(T > s)} = P(T > t)$$

$$1 - F(t + s) = (1 - F(s))(1 - F(t))$$

$$F(s)(1 - F(t)) = F(t + s) - F(t)$$

$$\frac{F(s)}{s}(1 - F(t)) = \frac{F(t + s) - F(t)}{s}$$

$$(1 - F(t)) \lim_{s \to 0} \frac{F(s)}{s} = \lim_{s \to 0} \frac{F(t + s) - F(t)}{s}$$

$$\begin{aligned} \lim_{s \to 0} \frac{F(s)}{s} &= \left.\frac{F'(s)}{1}\right|_{s=0} \\ &= F'(0) \\ &= f(0) \;\; \text{(call it } \lambda\text{)} \end{aligned}$$

$$
\begin{aligned}
(1-F(t))\lim_{s\to 0}\frac{F(s)}{s} &= \lim_{s\to 0}\frac{F(t+s)-F(t)}{s}\\
(1-F(t))\,\lambda &= F'(t)\\
(1-F(t))\,\lambda &= -(1-F(t))'\\
-\lambda t + c &= \log(1-F(t))\\
F(t) &= 1-e^{-\lambda t + c}
\end{aligned}
$$

To find the constant c

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t + c}\\
\lambda &= \lambda e^{c}\\
c &= 0
\end{aligned}
$$

$$
\begin{aligned}
F(t) &= 1-e^{-\lambda t}\\
f(t) &= \lambda e^{-\lambda t}.
\end{aligned}
$$

2.2.2 $Gamma(\alpha, \lambda)$

$$f(t) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t},\ 0 \le t;\ \alpha, \lambda > 0,$$

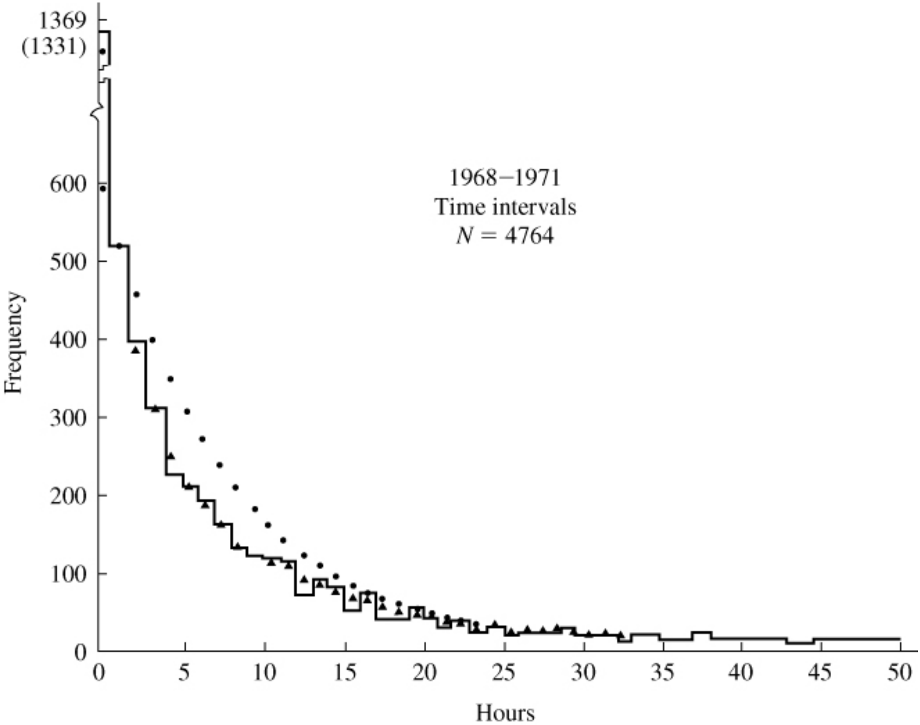
$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt.$$

$$\int_{-\infty}^{\infty} f(t)\, dt \stackrel{?}{=} 1$$

- α, λ are shape and scale parameters.
- See Mathematica Notebook.
- Prove $\alpha = 1$ gives $Exponential(\lambda)$.

Example 45 (Earthquake pattern) :

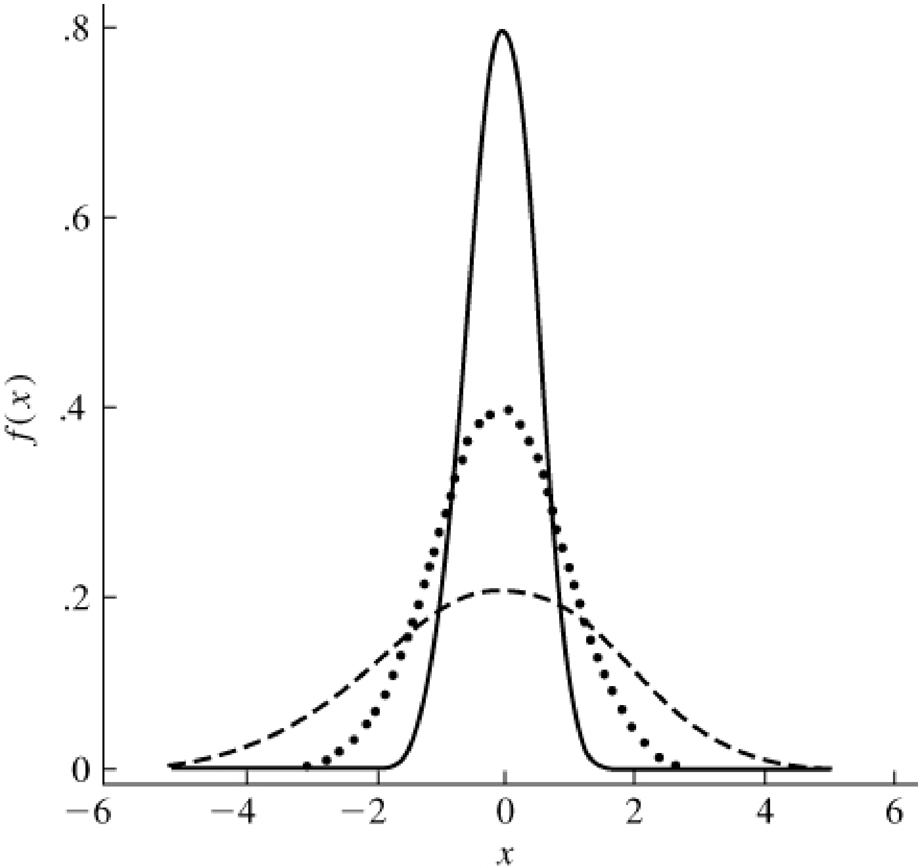
- *very erratic, and difficult to model*
- *Find the pdf of T, the time separating a sequence of small earthquakes.*
- *Exponential is not good because of memorylessnes property.*
- *Gamma (▲) looks fitting the data more than Exponential (•):* $\alpha = .509$, $\lambda = .0015$.
- *with these values, one can show that there is a large probability that the next earthquake immediately follows any given one; and this probability decreases with time (of course; why?).*


1369
(1331)
1968–1971
Time intervals
N = 4764
Frequency
Hours
0
100
200
300
400
500
600
0
5
10
15
20
25
30
35
40
45
50

2.2.3 $Normal\left(\mu,\sigma^2\right)$ (our ever friend)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2},\ \sigma > 0.$$

$$\int_{-\infty}^{\infty} f(t)\ dt \stackrel{?}{=} 1$$



- *Normal* because it is normal (statisticians)

- *Gaussian* after Carl Friedrich Gauss in measuring errors (applied scientists)

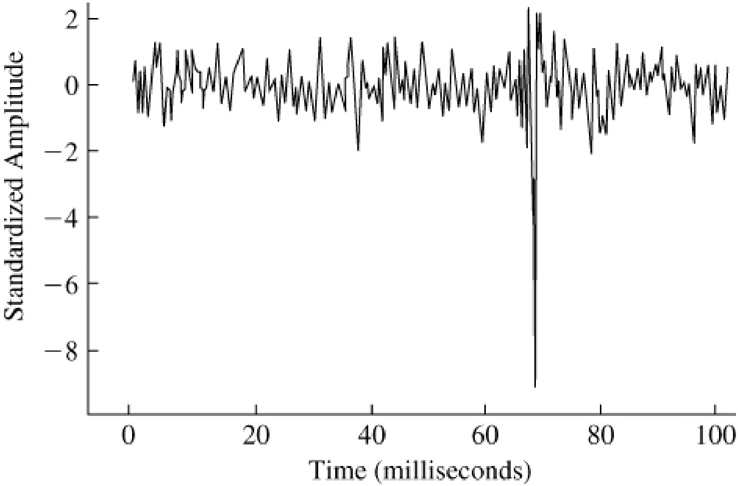
- *Bell* because it has a bell shape (some other parties)

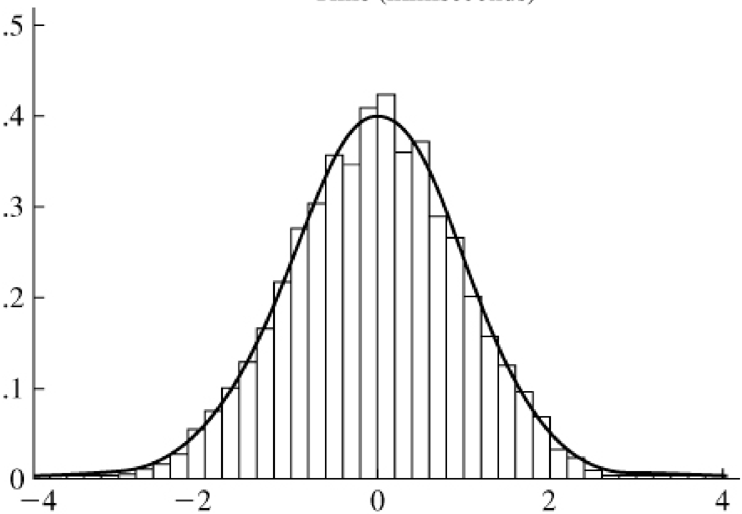
- symmetric around μ

- no closed form CDF; called Φ.

- Again: repeating an experiment many times under this pdf, how data looks like? How this apply to next example?

Example 46 (Veitch and Wilks (1985)) *: fitting amplitude of ice cracking noise in Arctic to Normal distribution.*





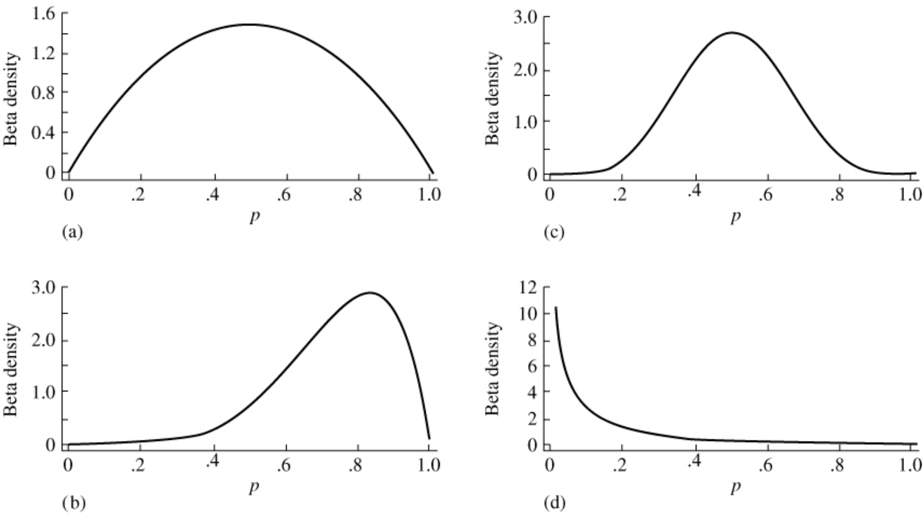
2.2.4 $Beta(a,b)$

$$f(x) = Beta(a,b)\, x^{a-1} (1-x)^{b-1},\ 0 \le x \le 1,$$

$$Beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$\int_{-\infty}^{\infty} f(t)\ dt \stackrel{?}{=} 1$$

- Prove that $a=1,\ b=1$ gives $Uniform(0,1)$
- Important in Bayesian approach (later).
- See Mathematica Notebook for shape parameters.



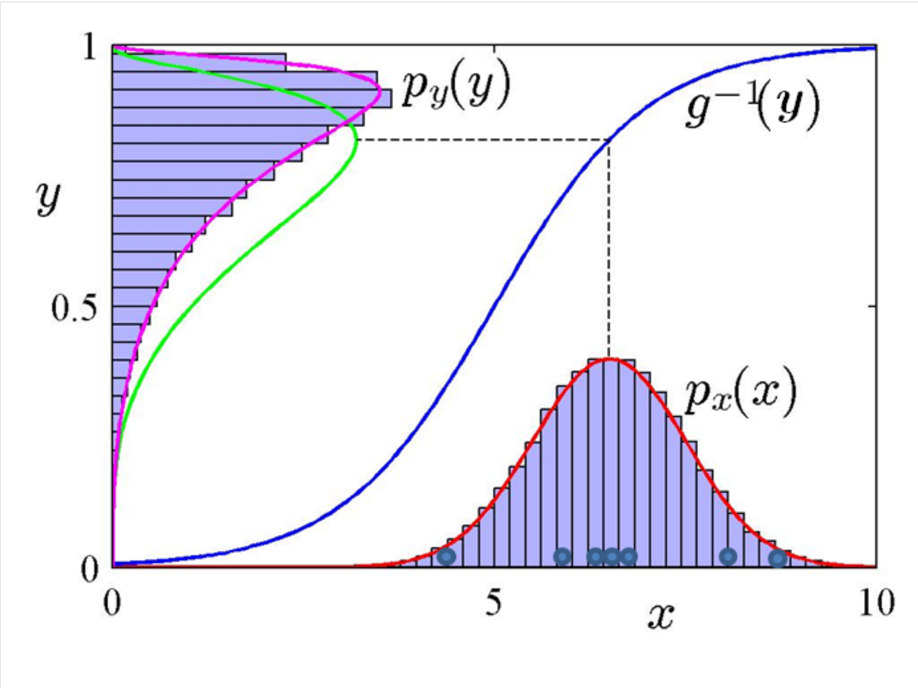
Example 47 (Mixtures) *We toss a coin, where*

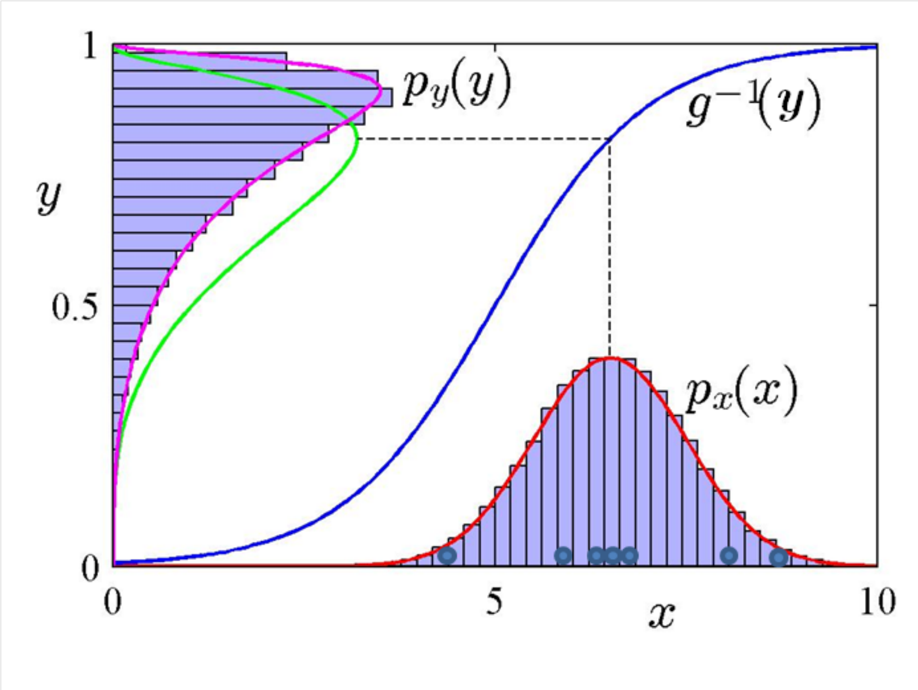
2.3 Functions of r.v.

Theorem 48 *If $Y = g(X)$, g is monotonically increasing (or decreasing) then g^{-1} exists and*

$$f_Y(y) = \left|\frac{1}{dy/dx}\right| f_X\left(g^{-1}(y)\right)$$

Meaning First:





"Not A Proof". At (x, y), where $y = g(x)$ (or $x = g^{-1}(y)$):

$$P(y < Y \leq y + dy) = P(x < X \leq x + dx)$$
$$f_Y(y)\,|dy| = f_X\left(g^{-1}(y)\right)|dx|$$

The "not a proof" is complete. ■

Proof.

$$
\begin{aligned}
P\left(Y \le y\right) &= P\left(X \le x\right), \\
y &= g\left(x\right) \\
F_Y\left(y\right) &= F_X\left(x\right) \\
\frac{d}{dy} F_Y\left(y\right) &= \frac{d}{dy} F_X\left(x\right) \\
f_Y\left(y\right) &= \frac{dx}{dy}\frac{d}{dx} F_X\left(x\right) \\
&= \frac{1}{dy/dx} f_X\left(x\right). \\
&= \frac{1}{dy/dx} f_X\left(g^{-1}\left(y\right)\right)
\end{aligned}
$$

Similarly, if g is monotonically decreasing, the only difference is

$$
P\left(Y > y\right) = P\left(X \le x\right)
$$

Then we will reach

$$
f_Y\left(y\right) = \frac{-1}{dy/dx} f_X\left(g^{-1}\left(y\right)\right).
$$

In general:

$$f_Y(y) = \left|\frac{1}{dy/dx}\right| f_X\left(g^{-1}(y)\right).$$

■

Take care if g^{-1} has two values; e.g., $X = Z^2$. You can do it from scratch also

$$
\begin{aligned}
P(X \le x) &= P\left(-\sqrt{x} \le Z \le \sqrt{x}\right) \\
&= F_Z\left(\sqrt{x}\right) - F_Z\left(-\sqrt{x}\right) \\
f_X(x) &= \left|\frac{d}{dx}\sqrt{x}\right| f_Z\left(\sqrt{x}\right) + \left|\frac{d}{dx}\sqrt{x}\right| f_Z\left(-\sqrt{x}\right) \\
&= \frac{1}{2}x^{-1/2} f_Z\left(\sqrt{x}\right) + \frac{1}{2}x^{-1/2} f_Z\left(-\sqrt{x}\right).
\end{aligned}
$$

Now, if $Z \sim N(0,1)$, then it is symmetric and

$$
\begin{aligned}
f_X(x) &= x^{-1/2} f_Z\left(\sqrt{x}\right) \\
&= x^{-1/2} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right]\Bigg|_{z=\sqrt{x}} \\
&= \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2},
\end{aligned}
$$

which is the pdf of $Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. Also, called chi-square (χ^2) distribution, with 1 degree of freedom.

Example 49 (solving by theorem) :

- $U \sim Uniform(0,1)$, $V = 1/U$.

- *First, pdf of U, its domain, V, its domain, and the function g.*

- *apply the theorem to get f_V, then draw it:*

$$
\begin{aligned}
f_V(v) &= f_U\left(\frac{1}{v}\right)\left|\frac{d}{dv}\left(\frac{1}{v}\right)\right| \\
&= 1 \times \frac{1}{v^2}
\end{aligned}
$$

Special case:

$$
\begin{aligned}
Y &= aX + b \\
F_Y\left(y\right) &= P\left(Y < y\right) \\
&= P\left(aX + b < y\right) \\
&= P\left(X < \frac{y-b}{a}\right), \; a > 0 \\
&= F_X\left(\frac{y-b}{a}\right), \\
\frac{d}{dy} F_Y\left(y\right) &= \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) \\
f_Y\left(y\right) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right).
\end{aligned}
$$

If $a < 0$; similarily:

$$\begin{aligned}
F_Y(y) &= P(aX + b < y)\\
&= P\left(X > \frac{y-b}{a}\right),\ a < 0\\
&= 1 - F_X\left(\frac{y-b}{a}\right),\\
\frac{d}{dy}F_Y(y) &= -\frac{d}{dy}F_X\left(\frac{y-b}{a}\right)\\
f_Y(y) &= -\frac{1}{a}f_X\left(\frac{y-b}{a}\right).
\end{aligned}$$

In general $\forall a \neq 0$ (why?):

$$f_Y(y) = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$$

We will see later (HW problem in Extra Materials) the condition for having maxima of f_Y coincide the maxima of f_X (the case of the green pdf above).

For case of $X \sim N\left(\mu, \sigma^2\right)$

$$
\begin{aligned}
f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \\
f_Y(y) &= \frac{1}{|a|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right]\Bigg|_{x=\frac{y-b}{a}} \\
&= \frac{1}{(|a|\sigma)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-(b+a\mu)}{a\sigma}\right)^2\right],
\end{aligned}
$$

which is

$$
N\left((b+a\mu), (a\sigma)^2\right).
$$

Interestingly: if

$$
\begin{aligned}
Z &= \frac{X-\mu}{\sigma} \\
&= \frac{1}{\sigma}X - \frac{\mu}{\sigma},
\end{aligned}
$$

Then, Z has the standard Normal density:

$$
Z \sim N(0,1)
$$

Application: Finding $P(x_0 < X < x_1)$ for any Normal r.v. requires knowing only the CDF of Z.

$$\begin{aligned}
P(x_0 < X < x_1) &= P\left(\frac{x_0-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{x_1-\mu}{\sigma}\right) \\
&= P\left(\frac{x_0-\mu}{\sigma} < Z < \frac{x_1-\mu}{\sigma}\right) \\
&= P\left(Z < \frac{x_1-\mu}{\sigma}\right) - P\left(Z < \frac{x_0-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{x_1-\mu}{\sigma}\right) - \Phi\left(\frac{x_0-\mu}{\sigma}\right).
\end{aligned}$$

Take care; e.g.,

$$x_0 < X \Leftrightarrow \frac{x_0-\mu}{\sigma} < \frac{X-\mu}{\sigma},$$

therefore

$$P(x_0 < X) = P\left(\frac{x_0-\mu}{\sigma} < \frac{X-\mu}{\sigma}\right).$$

But (show why by events and axioms?):

$$x_0 < X \Longrightarrow x_0^2 < X^2,$$

therefore

$$P(x_0 < X) \le P\left(x_0^2 < X^2\right).$$

Example 50 (IQ test Scores X) :

- *Found that* $X \sim N\left(100, 15^2\right)$.
- *What is the probability* Pr *that* $X \in [120, 130]$

$$\begin{aligned}
\Pr &= P(120 < X < 130) \\
&= P\left(\frac{120-100}{15} < \frac{X-100}{15} < \frac{130-100}{15}\right) \\
&= P(1.33 < Z < 2) \\
&= \Phi(2) - \Phi(1.33) \\
&= .9772 - .9082 \\
&= .069.
\end{aligned}$$

So, only 7% *of students takes grades in that range.*

Example 51 *[σ and μ]:*

$$
\begin{aligned}
P\left(\left|X-\mu\right|<\sigma\right) &= P\left(-\sigma<X-\mu<\sigma\right)\\
&= P\left(-1<\frac{X-\mu}{\sigma}<1\right)\\
&= P\left(-1<Z<1\right)\\
&= \Phi(1)-\Phi(-1)\\
&= .68,
\end{aligned}
$$

Similarly

$$
\begin{aligned}
P\left(\left|X-\mu\right|<2\sigma\right) &= \Phi(2)-\Phi(-2)\\
&= .9545,\\
P\left(\left|X-\mu\right|<3\sigma\right) &= \Phi(3)-\Phi(-3)\\
&= .9973
\end{aligned}
$$

So, all the probability almost is contained in $[-3\sigma, 3$
Also, $[-2\sigma, 2\sigma]$ *is a good approximation.*

Uniform Distribution & r.v. generators

Proposition 52 *Let $Z = F_X(X)$; then Z is distribut as $Uniform(0,1)$.*

Proof. First, draw the problem then notice that $0 \le Z \le 1$. Then,

$$\begin{aligned}
F_Z(z) &= P(Z \le z) \\
&= P(F_X(X) \le z) \\
&= P\left(X \le F_X^{-1}(z)\right) \\
&= F_X\left(F_X^{-1}(z)\right) \\
&= z,
\end{aligned}$$

which is a cdf of $Uniform(0,1)$.
If we want to prove it by using the theorem:

$$\begin{aligned}
f_Z(z) &= f_X\left(F_X^{-1}(z)\right)\left|\frac{d}{dz}F_X^{-1}(z)\right| \\
&= f_X\left(F_X^{-1}(z)\right) \times \frac{1}{f_X\left(F_X^{-1}(z)\right)} \\
&= 1
\end{aligned}$$

■

Proposition 53 *Let* $U \sim Uniform(0,1)$*, and let* $X = F^{-1}(U)$ *(some* F*). Then* $F_X = F$*.*

Proof. First, draw the problem.

$$
\begin{aligned}
F_X(x) &= P(X \le x) \\
&= P\left(F^{-1}(U) \le x\right) \\
&= P(U \le F(x)) \\
&= F(x),
\end{aligned}
$$

which is used to generate a r.v. if we know its distribution F_X and have and access to only uniform r.v. generator. ∎

Chapter 3

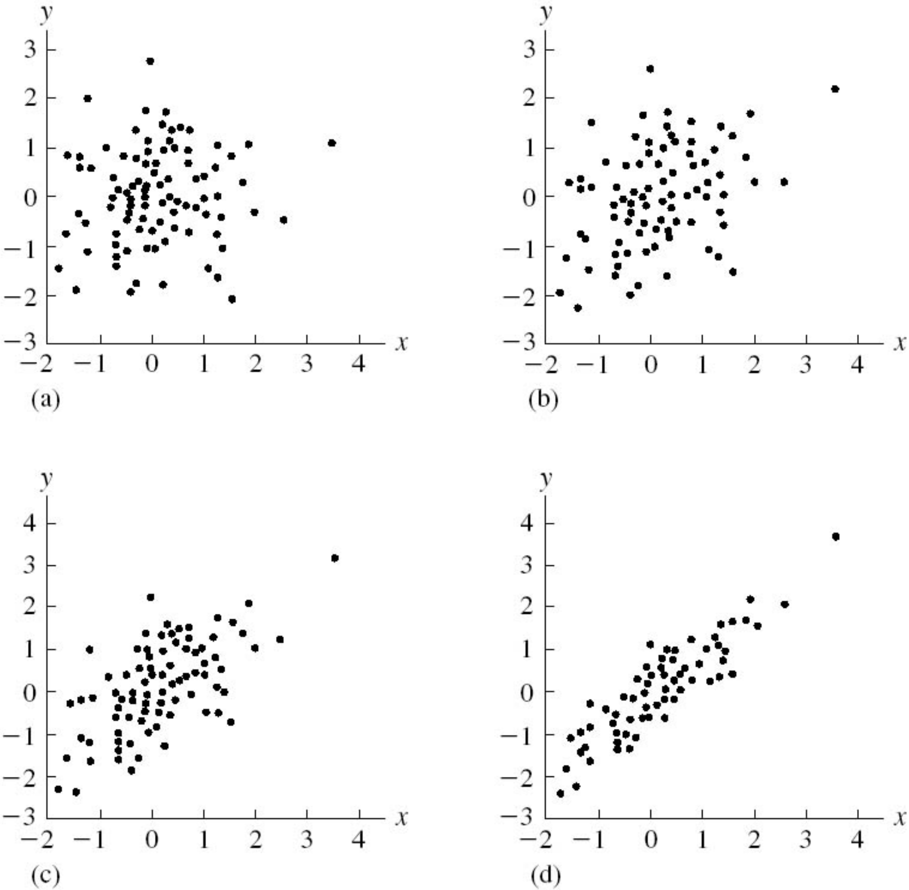
Joint Distributions

3.1 Introduction: examples for joint distributions

- Number of predators and Number of preys for a particular species in ecology.
- Hight and Weight of particular category of distribution of people.
- A model for joint distribution of Age and Length in a population of fish.

Motivation by very simple example: $\{(0,0),(1,1)\}$ has different joint distribution than $\{(1,0),(0,1)\}$. However each of X and Y have the same marginal distribution.

Another example: The Height and Weight of some species of fish are reported for 100 fishes. How they are related together?



Revision: relation between F, f, and p

For both discrete and continuous:

$$\begin{aligned} P(X \le x) &= F_X(x), \\ P(x_1 < X \le x_2) &= P(X \in (x_1, x_2]) \\ &= F_X(x_2) - F_X(x_1), \ x_1 < x_2. \end{aligned}$$

discrete only:

$$\begin{aligned} F_X(x_k) &= \sum_{i=-\infty}^{k} p(x_i), \\ p_X(x_k) &= F_X(x_k) - F_X(x_{k-1}), \end{aligned}$$

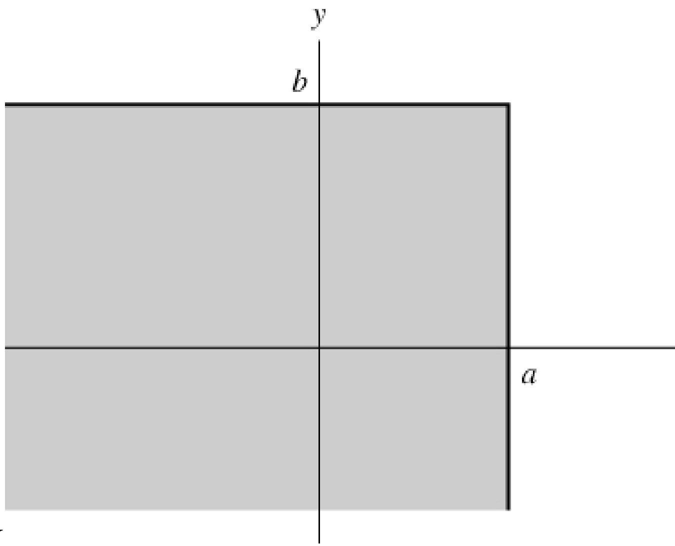
continuous only:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(u)\ du, \\ f_X(x) &= F'_X(x), \end{aligned}$$

Definition 54 *The joint distribution function of two r.v. is defined as*

$$F_{XY}\left(x, y\right) = P\left(X \leq x, Y \leq y\right).$$

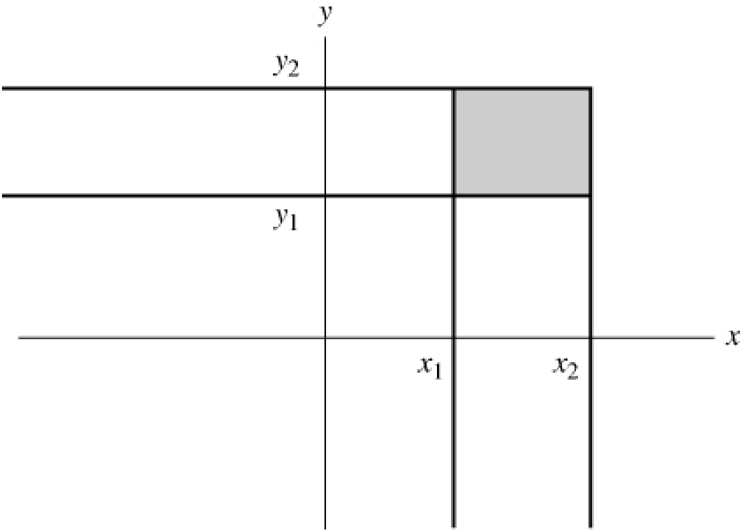
When there is no confusion we can write just F



instead of F_{XY}

Therefore, for any rectangle in the space

$$P\left(x_1 < X \leq x_2, y_1 < Y \leq y_2\right)$$
$$= F\left(x_2, y_2\right) - F\left(x_2, y_1\right) - F\left(x_1, y_2\right) + F\left(x_1, y_1\right)$$



It can be proven that: (proof is omitted) any more complex area than rectangles, e.g., circle, can be determined by limits of rectangles. Hence, F can determine the probability of any region in the spac

Definition 55 (Generalization) *The joint distribution function of several r.v., in p-dimensional subspace, is defined as:*

$$F_{X_1 \ldots X_p} = P\left(X_1 \le x_1, \ldots, X_p \le x_p\right)$$

3.2 Discrete r.v.

Definition 56 *The joint pmf of two r.v.:*

$$p\left(x_i, y_j\right) = P\left(X = x_i, Y = y_j\right)$$

Example 57 *coin tossing 3 times:*

$\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$

$X \backslash Y$	0	1	2	3
0	$\frac{1}{8}$ $\{ttt\}$	$\frac{2}{8}$ $\{tht, tth\}$	$\frac{1}{8}$ $\{thh\}$	0 $\{\}$
1	0 $\{\}$	$\frac{1}{8}$ $\{htt\}$	$\frac{2}{8}$ $\{hht, hth\}$	$\frac{1}{8}$ $\{hhh\}$

X, Y : # heads in 1st and 3 tosses respectively.

$$\begin{aligned}
p_Y(0) &= P(Y = 0) \\
&= P(\{(Y = 0, X = 0)\} \cup \{(Y = 0, X = 1)\}) \\
&= P(Y = 0, X = 0) + P(Y = 0, X = 1) \\
&= \frac{1}{8}, \\
p_Y(1) &= P(Y = 1) \\
&= P(Y = 1, X = 0) + P(Y = 1, X = 1) \\
&= \frac{3}{8}.
\end{aligned}$$

Definition 58 *The marginal pmf is defined as*

$$p_{X_1}(x_1) = \sum_{x_2} p(x_1, x_2),$$

more general

$$p_{X_1}(x_1) = \sum_{x_2,\ldots x_p} p\left(x_1, \ldots, x_p\right),$$

more more general

$$p_{X_1,\ldots,X_r}(x_1, \ldots, x_r) = \sum_{x_{r+1},\ldots x_p} p\left(x_1, \ldots, x_p\right).$$

Notice that:

$$\begin{aligned} F_{X_1,\ldots,X_p}\left(x_1, \ldots, x_p\right) &= P\left(X_1 \le x_1, \ldots, X_p \le x_p\right) \\ &= \sum_{a_1=-\infty}^{x_1} \cdots \sum_{a_p=-\infty}^{x_p} p\left(a_1, \ldots, a_p\right), \end{aligned}$$

Multinomial Distribution:

- Generalization to $Binomial(n, p)$
- n independent trials
- each can result in on of r types of outcomes (c.f. 2 types in $Binomial(n, p)$)
- each type has a probability p_r, $\sum_r p_r = 1$ (c.f. $p, 1-p$).
- outcome is: $N_1 = n_1, \ldots, N_r = n_r$, $\sum_i N_i = n$ (c.f. $N = k$)
- each with probability $p_1^{n_1} \ldots p_r^{n_r}$,
- For example: n students, each can get A, B, C, D, or F, with probabilities $p_1, \ldots, p_5$. What is the probability that n_1 get A, n_2 get B, …, n_5 fail?

- number of these choices (from Sec. 1.4.2) is $\frac{n!}{n_1!...n_r!}$. Therefore

$$
\begin{aligned}
p(n_1,\ldots,n_r) &= \frac{n!}{n_1!\ldots n_r!}p_1^{n_1}\ldots p_r^{n_r}\\
&= \binom{n}{n_1\ldots n_r}\prod_{i=1}^{r}p_i^{n_i}.
\end{aligned}
$$

Special case: $r=2$

$$
\begin{aligned}
p(n_1,n_2) &= \frac{n!}{n_1!n_2!}p_1^{n_1}p_2^{n_2}\\
&= \frac{n!}{n_1!(n-n_1)!}p_1^{n_1}(1-p_1)^{n-n_1}\\
&= \binom{n}{n_1}p_1^{n_1}(1-p_1)^{n-n_1}\\
&\equiv Binomial(n,p_1)
\end{aligned}
$$

Notice that: the marginal, e.g., $p_{N_1}(n_1)$ is very difficult to be obtained by

$$p_{N_1}(n_1) = \sum_{n_2,\ldots,n_r} p(n_1,\ldots,n_r).$$

Or in general

$$p_{N_i}(n_i) = \sum_{n_1,\ldots,n_{i-1},n_{i+1},\ldots,n_r} p(n_1,\ldots,n_r).$$

However, it can be obtained at once by noticing that:

$$N_i \sim Binomial(n, p_i),$$
$$p_{N_i}(n_i) = \binom{n}{n_i} p_i^{n_i} (1-p_i)^{n-n_i}.$$

Example 59 (histogram) :

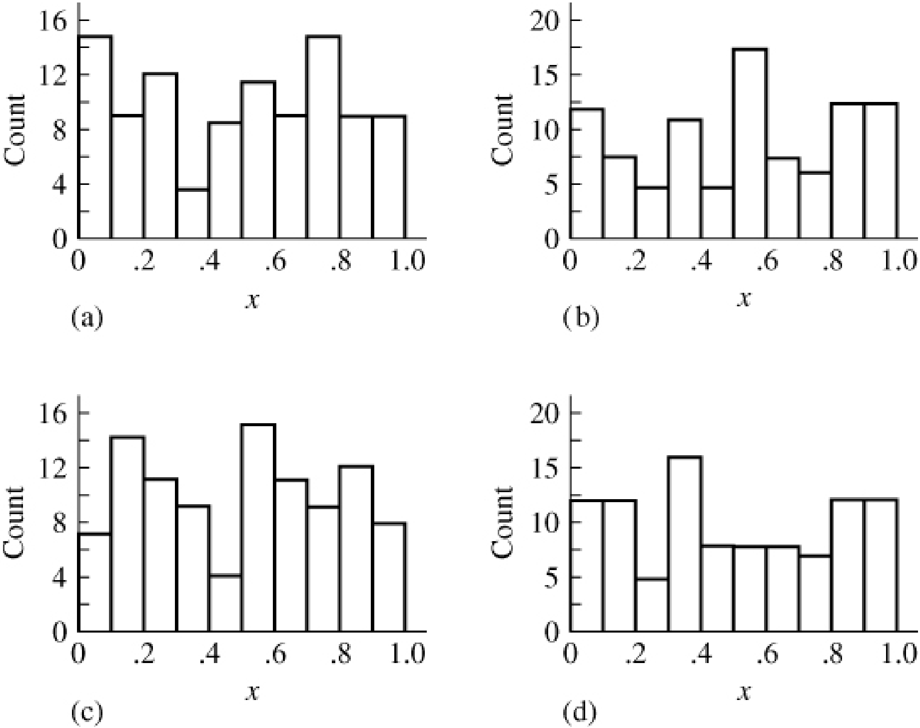
- *n independent observations, e.g., $n = 100$, from $Uniform(0,1)$*
- *Partition the interval $[0,1]$ to 10 equal bins*
- *n_i is the number of observations falling in the i^{th} bin; i.e,*

$$n_i = \sum_{j=1}^{n} I_{N_j \in i^{th} bin}.$$

- *Therefore, $N_1, \ldots, N_{10}$ are:*

  $Multinomial\left(100, p_{1,\ldots,}p_{10}\right)$, $p_i = 0.1$ (one tenth of the period $[0,1]$)

- *Notice the fluctuations in the following histograms; each corresponds to a new set of 100 observations.*



Later we will see, very interestingly, that with increasing** n**:

- *we can increase the number of bins; i.e., decrease the width of each bin.*
- *the histogram stabilizes ("converges")*
- *the histogram "converges" to the original pdf regardless whether it is uniform or not.*

3.3 Continuous r.v.

Definition 60 *If the joint cdf of two r.v. is differentiable, then the their joint pdf function is defined as*

$$f_{XY}\left(x,y\right) = \frac{\partial^2}{\partial x \partial y} F_{XY}\left(x,y\right);$$

and therefore

$$F_{XY}\left(x,y\right) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}\left(u,v\right)\, du\, dv.$$

and it can be also shown that (see extra material):

$$P\left(\left(X,Y\right) \in A\right) = \iint_A f_{XY}\left(u,v\right)\, du\, dv.$$

Again: for very small δ_x and δ_y,

$$\begin{aligned}
&P\left(x \le X \le x+\delta_x, y \le Y \le y+\delta_y\right)\\
&= \int_{x}^{x+\delta_x} \int_{y}^{y+\delta_y} f_{XY}\left(u,v\right)\, du\, dv.\\
&\approx f_{XY}\left(x,y\right)\delta_x \delta_y,
\end{aligned}$$

Marginal:

$$
\begin{aligned}
F_X(x) &= F_{XY}(x,\infty) \\
&= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(u,v)\ du\,dv.
\end{aligned}
$$

Generalization:

$$
F\left(x_1,\ldots,x_p\right) =
$$

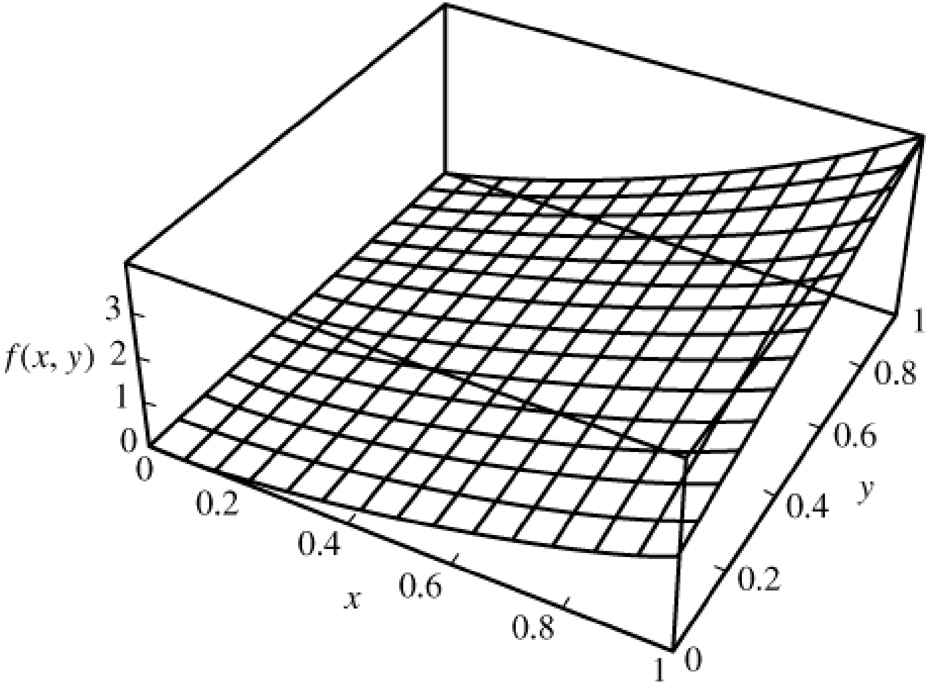
$$
\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_p} f\left(u_1,\ldots,u_p\right) du_1\ldots du_p,
$$

$$
P\left(\left(X_1,\ldots,X_p\right)\in A\right) =
$$

$$
\underset{A}{\int\cdots\int} f\left(u_1,\ldots,u_p\right)\ du_1\ldots du_p,
$$

$$
f_{X_1}(x_1) =
$$

$$
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f\left(x_1,x_2,\ldots,x_p\right) dx_2\ldots dx_p
$$



Example 61

$$f\left(x, y\right) = \frac{12}{7}\left(x^2 + xy\right),\ 0 \le x \le 1,\ 0 \le y \le 1.$$

First make sure that it integrates to one

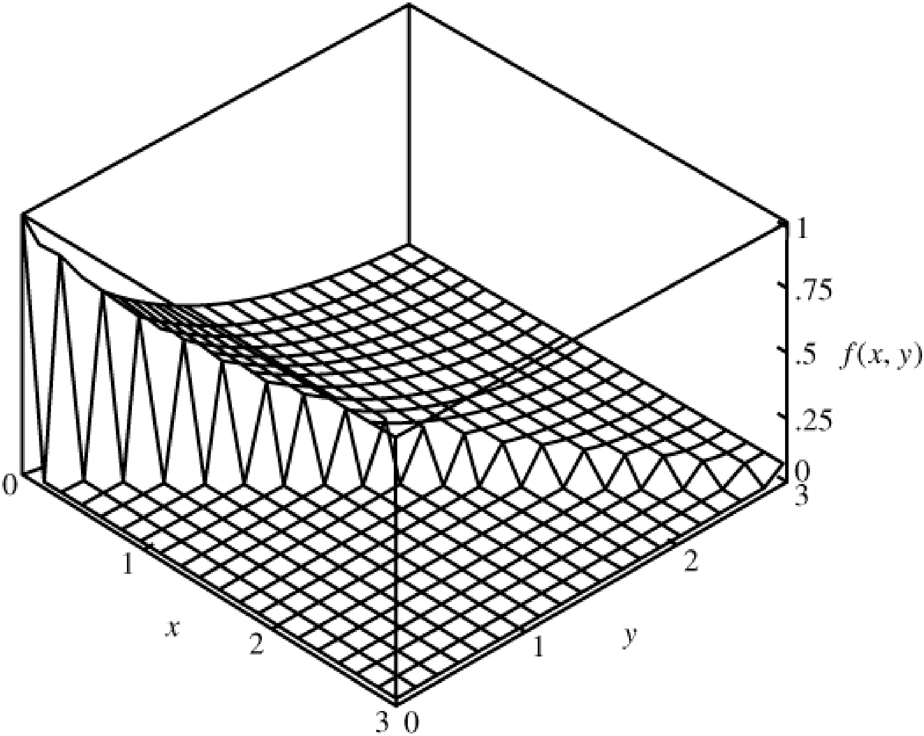
$$\int_{x=0}^{x=1} \int_{y=0}^{y=1} f\left(x, y\right)\ dy\ dx = 1$$

Next, find $P(X > Y)$. What does it mean when I pick up a pair (x, y) for the population?

$$P(X > Y) = \frac{12}{7}\int_0^1\int_0^x \left(x^2 + xy\right)\,dy\,dx$$
$$= \frac{9}{14}.$$

Next, find f_X.

$$f_X(x) = \frac{12}{7}\int_0^1 \left(x^2 + xy\right)\,dy$$
$$= \frac{12}{7}\left(x^2 + \frac{x}{2}\right).$$

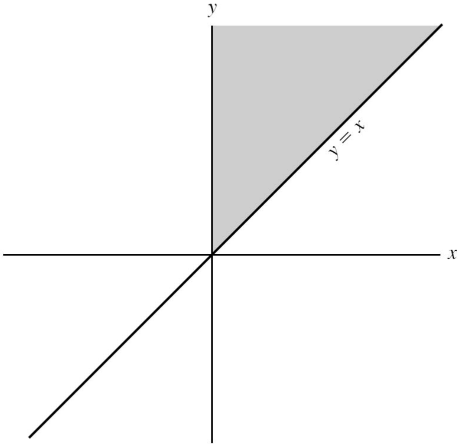


Example 62 *Consider the density*

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & 0 \le x \le y,\ \lambda > 0 \\ 0, & otherwise \end{cases}.$$

Take care, it can be re-written as

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} I_{0\le x} I_{x \le y}, & \lambda > 0 \\ 0, & otherwise \end{cases}.$$



$$f_X(x) = \int_x^\infty \lambda^2 e^{-\lambda y}\, dy$$
$$= \lambda e^{-\lambda x},\ x \geq 0,$$

$which\ is\ Exponential(\lambda)$

$$f_Y(y) = \int_0^y \lambda^2 e^{-\lambda y}\, dx$$
$$= \lambda^2 y e^{-\lambda y},\ 0 \leq y,$$

$which\ is\ Gamma(2,\lambda).$

Example 63 *A point is chosen randomly in a disk of radius 1. Then*

$$f_{XY}(x,y) = \begin{cases} \frac{1}{\pi} & x^2+y^2 \le 1 \\ 0 & otherwise. \end{cases}$$

What is $P(R \le r)$? In general, if $X = (X_1, \ldots, X_p)$ is uniformly distributed over an area A, then

$$\begin{aligned} f_X(x_1, \ldots, x_p) &= \frac{1}{|A|}, \\ P(X \in B) &= \int_B f_X \, dx \\ &= \frac{|B|}{|A|} \end{aligned}$$

$$\begin{aligned} P(R \le r) &= P\left(x^2+y^2 \le r\right) \\ &= \frac{\pi r^2}{\pi} \\ &= r^2. \end{aligned}$$

Later, we can do it differently: $r = x^2 + y^2$, which is a function of 2 r.v. (transformation).

What is f_X?

$$f_X(x) = \frac{1}{\pi}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy$$
$$= \frac{2}{\pi}\sqrt{1-x^2}, \; -1 \le x \le 1.$$

Curse of dimensionality:

For p-dimensional spheres, $P(R \le r) = r^p$, which means that data will be on the surface!!

3.4 Independent r.v.

Definition 64 *The r.v. $X_1, \ldots, X_n$ are said to be independent if*

$$F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n), \forall x_1, \ldots x_n.$$

For two r.v. (for intuition)

$$F_{X_1 X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2)$$
$$P(X_1 \le x_1, X_2 \le x_2) = P(X_1 \le x_1) P(X_2 \le x_2),$$

which is the independence of two events (from Ch. 1).

In particular for continuous r.v., if

$$F_{X_1 X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2)$$

then,

$$\begin{aligned} f_{X_1 X_2}(x_1, x_2) &= \frac{\partial^2 \left[F_{X_1}(x_1) F_{X_2}(x_2)\right]}{\partial X_1 \partial X_2} \\ &= f_{X_1}(x_1) f_{X_2}(x_2), \end{aligned}$$

Also, if

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2),$$

then

$$\begin{aligned}
F_{X_1X_2}(x_1, x_2) &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1X_2}(x_1, x_2)\; dx_1\, dx_2 \\
&= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1}(x_1)\, f_{X_2}(x_2)\; dx_1\, dx_2 \\
&= \int_{-\infty}^{x_1} f_{X_1}(x_1)\; dx_1 \int_{-\infty}^{x_2} f_{X_2}(x_2)\; dx_2 \\
&= F_{X_1}(x_1)\, F_{X_2}(x_2).
\end{aligned}$$

Also, it is easy to show that if X_1 and X_2 are independent, then

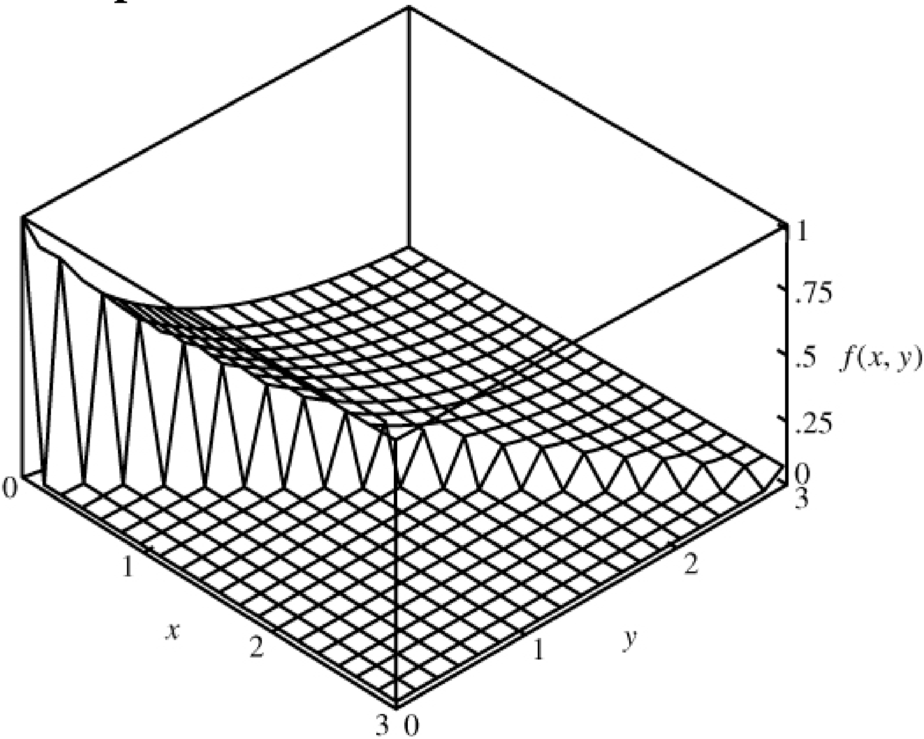
$$P(X_1 \in A, X_2 \in B) = P(X_1 \in A)\, P(X_2 \in B).$$

The proof can start easily from the pdf, or more generally from the cdf (see extra materials).

Lemma 65 *if X and Y are independent r.v. then the r.v. $G = g(X)$ and $H = h(Y)$ are independent as well.*

Proof. : is omitted ■

Example 66 (cont. Ex. 62) :



$$f_{XY}(x,y)=\begin{cases}\lambda^2 e^{-\lambda y}, & 0\le x\le y,\ \lambda>0\\ 0, & otherwise\end{cases}$$

Take care, it looks like it factors; however it is not since

$$f(x,y)=\begin{cases}\lambda^2 e^{-\lambda y} I_{0\le x} I_{x\le y}, & \lambda>0\\ 0, & otherwise\end{cases}.$$

The marginal was

$$f_X(x) = \lambda e^{-\lambda x},\ 0 \leq x,$$
$$f_Y(y) = \lambda^2 y e^{-\lambda y},\ 0 \leq y,$$

They are not independent, since

$$f_X f_Y \neq f_{XY}$$

Example 67 *Suppose that a point is selected uniformly from a unit square centered around 0. Then*

$$f_{XY}(x,y)=\begin{cases}1 & \frac{-1}{2}\le x\le\frac{1}{2},\ \frac{-1}{2}\le y\le\frac{1}{2}\\ 0 & otherwise\end{cases}$$
$$=I_{\frac{-1}{2}\le x\le\frac{1}{2}}I_{\frac{-1}{2}\le y\le\frac{1}{2}}.$$

then it factors and X and Y are independent (check by finding f_X and f_Y and prove that each is uniform and $f_{XY}=f_Xf_Y$).

However, if X and Y are uniformly distributed over the diamond area (rotating a square 90°) then let's draw it and formalize it:

$$f_{XY}(x,y)=\begin{cases}1 & \left|y\pm x\right|\le 1/\sqrt{2}\\ 0 & otherwise\end{cases}$$
$$=I_{\left|y\pm x\right|\le 1/\sqrt{2}},$$

which does not factor; hence they are not independent.

Check by finding f_X and f_Y

$$
\begin{aligned}
f_X(x) &= \begin{cases} \int_{-x-1/\sqrt{2}}^{x+1/\sqrt{2}} dy & -1/\sqrt{2} \le x \le 0 \\ \int_{x-1/\sqrt{2}}^{-x+1/\sqrt{2}} dy & 0 \le x \le 1/\sqrt{2} \end{cases} \\
&= \begin{cases} 2x + \sqrt{2} & -1/\sqrt{2} \le x \le 0 \\ -2x + \sqrt{2} & 0 \le x \le 1/\sqrt{2} \end{cases} \\
&= -2|x| + \sqrt{2},\ |x| \le 1/\sqrt{2},
\end{aligned}
$$

similarly

$$
f_Y(y) = -2\left|y\right| + \sqrt{2},\ \left|y\right| \le 1/\sqrt{2}.
$$

We see that

$$
f_{XY} \ne f_X f_Y
$$

3.5 Conditional Distributions

3.5.1 The Discrete Case

Definition 68 *The discrete conditional pmf is defined as:*

$$P_{Y|X}\left(Y = y_j | X = x_i\right) = \frac{P_{XY}\left(X = x_i, Y = y_j\right)}{P_X\left(X = x_i\right)},$$

where, $P\left(X = x_i\right) \neq 0$. *And hence,*

$$\begin{aligned} P_{XY}\left(x, y\right) &= P_{Y|X}\left(y|x\right) P_X\left(x\right), \\ P_Y\left(y\right) &= \sum_x P_{XY}\left(x, y\right) \\ &= \sum_x P_{Y|X}\left(y|x\right) P_X\left(x\right), \end{aligned}$$

which is nothing but the law of total probability.

Notice that: the conditional probability was proven to be a true probability measure; hence condition probability is a genuine pmf.

Example 69 (cont. Ex. 57) :

$X\backslash Y$	0	1	2	3
0	$\frac{1}{8}\ \{ttt\}$	$\frac{2}{8}\ \{tht, tth\}$	$\frac{1}{8}\ \{thh\}$	$0\ \{\}$
1	$0\ \{\}$	$\frac{1}{8}\ \{htt\}$	$\frac{2}{8}\ \{hht, hth\}$	$\frac{1}{8}\ \{hhh\}$

To find $P_{X|Y=1}$:

$$P_{X|Y=1}(0|1) = \frac{P_{X,Y}\left(x=0, y=1\right)}{P_Y\left(y=1\right)}$$
$$= \frac{P_{X,Y}\left(x=0, y=1\right)}{P_{XY}(0,1)+P_{XY}(1,1)}$$
$$= \frac{2/8}{2/8+1/8} = \frac{2}{3},$$
$$P_{X|Y=1}(1|1) = \frac{1/8}{2/8+1/8} = \frac{1}{3}.$$

Example 70 (Imperfect Counter) *:*

- *Particle counter, detects with probability* p.
- *Number of incoming particles/unit time* $\sim Poisson(\lambda)$.
- N *is the true number of particles,* X *is the counted; therefore*

$$X|N=n \sim Binomial\left(n,p\right),$$

$$P\left(X=k|N=n\right)=\binom{n}{k}p^{k}\left(1-p\right)^{n-k};$$

$$N \sim Poisson\left(\lambda\right)$$

$$P\left(N=n\right)=\frac{\lambda^{n}e^{-\lambda}}{n!}.$$

$$\begin{aligned}
P\left(X=k\right) &= \sum_{n=k}^{\infty} P\left(X=k|N=n\right)P\left(N=n\right) \\
&= \sum_{n=k}^{\infty} \frac{\lambda^{n}e^{-\lambda}}{n!}\binom{n}{k}p^{k}\left(1-p\right)^{n-k} \\
&= \frac{\left(\lambda p\right)^{k}}{k!}e^{-\lambda}\sum_{n=k}^{\infty}\lambda^{n-k}\frac{\left(1-p\right)^{n-k}}{(n-k)!} \\
&= \frac{\left(\lambda p\right)^{k}}{k!}e^{-\lambda}\sum_{j=0}^{\infty}\lambda^{j}\frac{\left(1-p\right)^{j}}{j!} \\
&= \frac{\left(\lambda p\right)^{k}}{k!}e^{-\lambda}e^{\lambda\left(1-p\right)} \\
&= \frac{\left(\lambda p\right)^{k}}{k!}e^{-\lambda p},
\end{aligned}$$

$$X \sim Poisson\left(\lambda p\right),$$

3.5.2 The Continuous Case

Definition 71 *The conditional density $f_{Y|X}(y|x)$ is defined as*

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

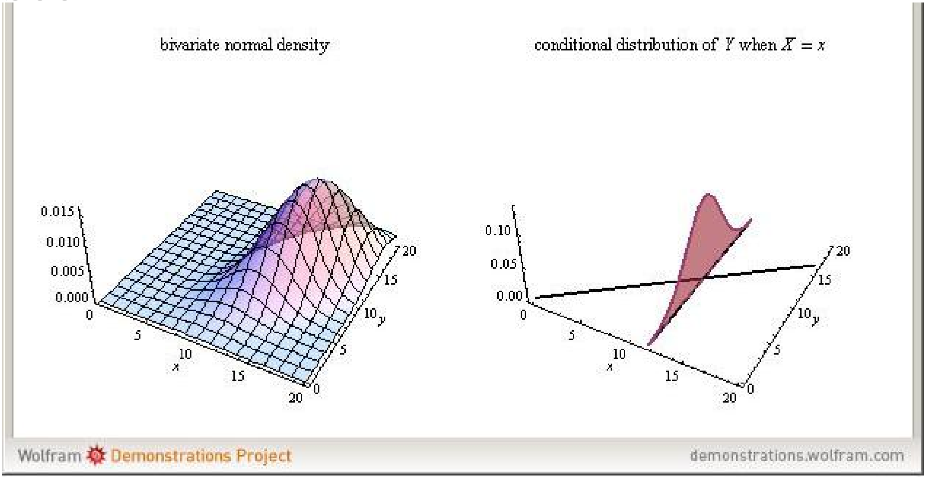
For intuition:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{\frac{P(x \le X \le x+dx, y \le Y \le y+dy)}{dx\,dy}}{\frac{P(x \le X \le x+dx)}{dx}}, \\ f_{Y|X}(y|x)\,dy &= \frac{P(x \le X \le x+dx, y \le Y \le y+dy)}{P(x \le X \le x+dx)} \\ &= P(y \le Y \le y+dy | x \le X \le x+dx). \end{aligned}$$

Alternatively:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy}, \end{aligned}$$

So the denominator is just a normalizing factor, and indeed $f_{Y|X}(y|x)$ is a pdf and integrates to one. **Let's try playing with Mathematica Notebook**



Analogously to discrete case:

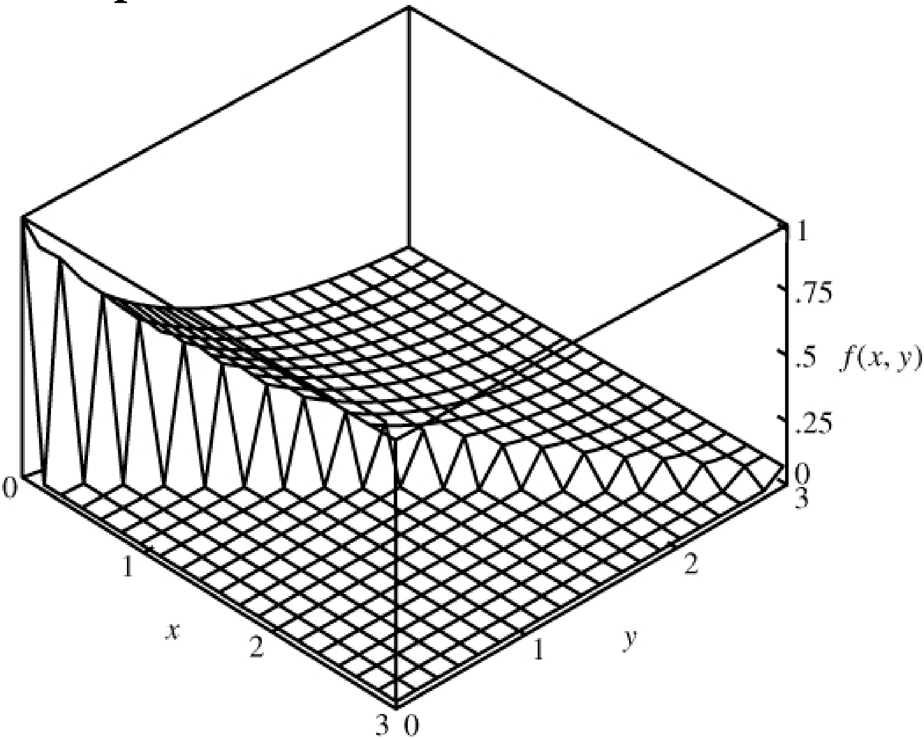
$$f_{XY}(x,y) = f_{Y|X}(y|x)\, f_X(x),$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\ dx$$

$$= \int_{-\infty}^{\infty} f_{Y|X}(y|x)\, f_X(x)\ dx,$$

which can be interpreted as a law of total probability for continuous case.

Example 72 (cont. Ex. 62) :



$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & 0 \le x \le y,\ \lambda > 0 \\ 0, & otherwise \end{cases},$$

$$f_X(x) = \lambda e^{-\lambda x},\ 0 \le x,$$

$$f_Y(y) = \lambda^2 y e^{-\lambda y},\ 0 \le y.$$

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{\lambda^2 e^{-\lambda y}}{\lambda e^{-\lambda x}} \\
&= \lambda e^{-\lambda(y-x)},\ 0 \le x \le y,\ \lambda > 0,
\end{aligned}
$$

$$
Y|X \sim Exponential(\lambda)
$$

$$
\begin{aligned}
f_{X|Y}(x|y) &= \frac{\lambda^2 e^{-\lambda y}}{\lambda^2 y e^{-\lambda y}} \\
&= \frac{1}{y},\ 0 \le x \le y,
\end{aligned}
$$

$$
X|Y \sim Uniform(0, 1/y).
$$

Notice that: *we can generate* (X, Y) *by generating* x*, followed by* $y|x$ *or by generating* y *followed* $x|y$

Bayesian Inference

- Coin tossing $\sim Bernoulli(\theta)$, θ is unknown
- Number of heads $x \sim Binomial(n, \theta)$.
- Given x, n what is θ?
- Conditional Probability of x is $f_{X|\Theta}(x|\theta)$.
- Prior Probability of Θ is $f_{\Theta}(\theta)$.
- Posterior Probability of Θ is $f_{\Theta|X}(\theta|x)$

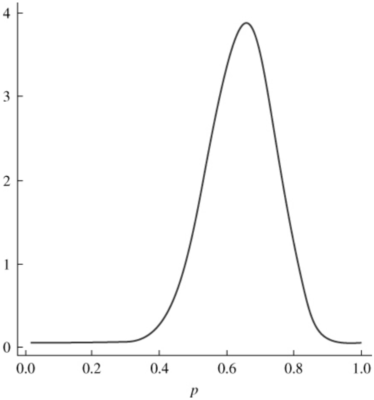
Therefore, we proceed very analogously to Bayes rule in Ch. 1

$$
\begin{aligned}
f_{\Theta|X}(\theta|x) &= \frac{f_{X,\Theta}}{f_X} \\
&= \frac{f_{X|\Theta} f_\Theta}{\int f_{X,\Theta}\, d\theta} \\
&= \frac{f_{X|\Theta} f_\Theta}{\int f_{X|\Theta} f_\Theta\, d\theta} \\
&= \underbrace{Const(x)}_{\int f_{X|\Theta} f_\Theta\, d\theta} f_{X|\Theta} f_\Theta \\
&= Const(x) \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot 1 \\
&= Const'(x)\ \theta^x (1-\theta)^{n-x},
\end{aligned}
$$

which is the density of $Beta(x+1, n-x+1)$,and $Const'(x)$ has to be equal to $Beta(x+1, n-x+1)$.

Anyway, this is a normalizing factor and has no significance on the shape of $f_{\Theta|X}(\theta|x)$.

Given: $n = 20,\ x = 13,\ f_{\Theta|X}(\theta|x)$ is drawn as the following and it is unlikely to be < 0.4.



Interpretation:

- Prior to observation:

  $\Theta \sim Uniform(0,1)$ (no-information case).

- Posterior to observations:

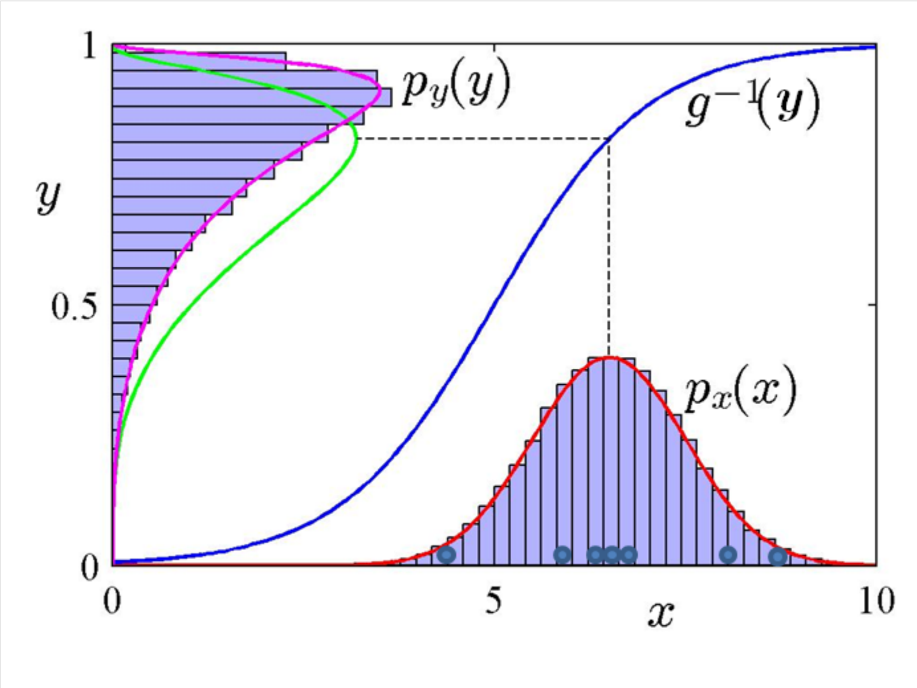
  $\Theta \sim Beta(x+1, n-x+1)$.

- A Posteriori represents my **belief** after observations based on my **subjective belief** prior to observations.

- We can estimate θ by the value that maximizes the posteriori as was done in Ch. 1

3.6 Functions of Jointly distributed r.v.s

Revision: Single function of single r.v.

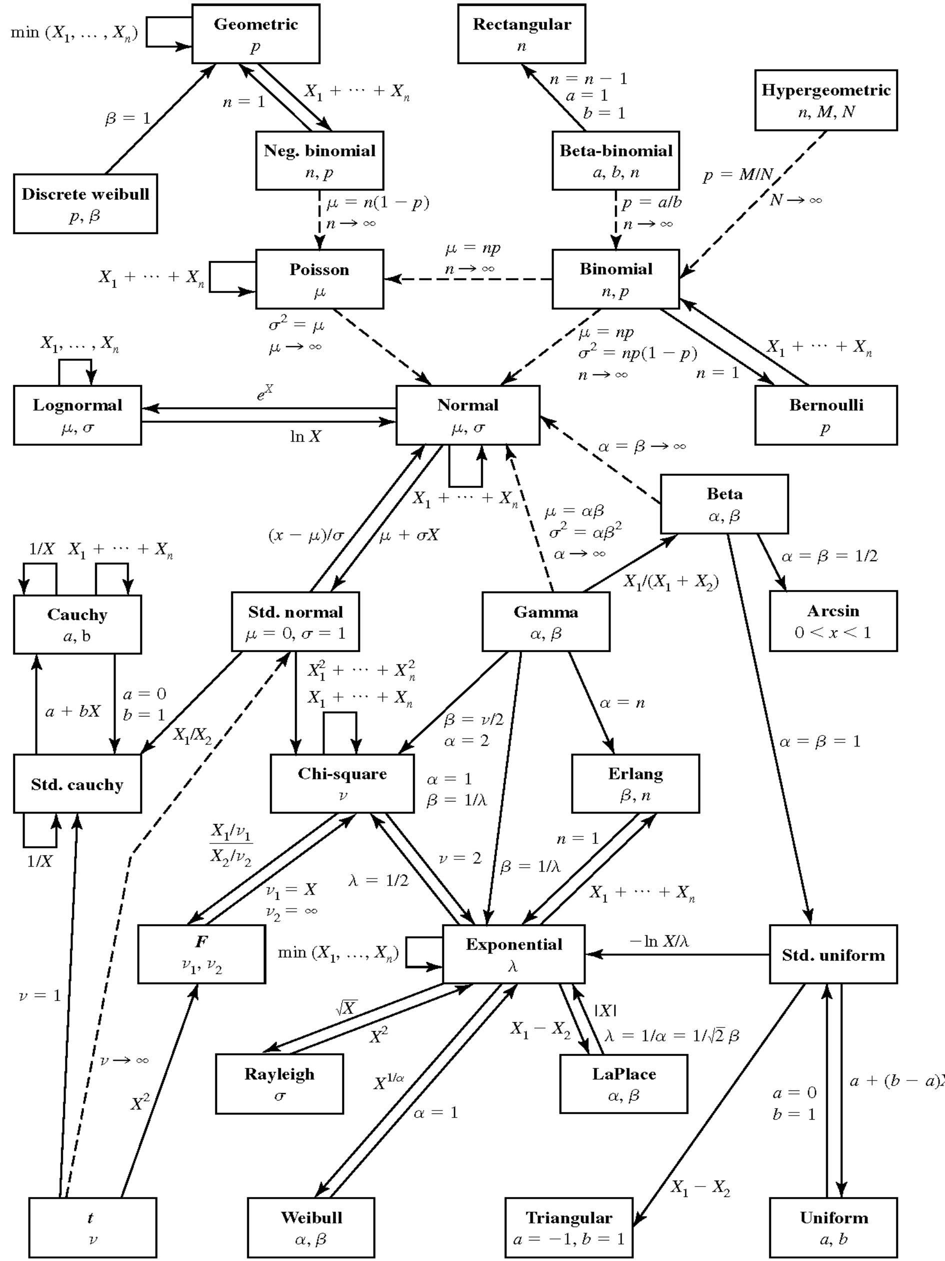


X is Discrete

$$p_Y(y) = p_X\left(g^{-1}(y)\right)$$

X is Continuous

$$f_Y(y)\left|dy\right| = f_X\left(g^{-1}(y)\right)\left|dx\right|$$



Adapted from Leemis, L. M. (1986), “Relationships among common univariate distributions,” *The American Statistician*, Vol. 40, No. 2, pp. 143–146. With permission.

3.6.1 Single Function of Jointly Distributed r.v.s

- Let's see examples below
- In each example we draw the 3D function:

$$Z = f(X, Y)$$

- We cut it at some level

$$Z = const,$$

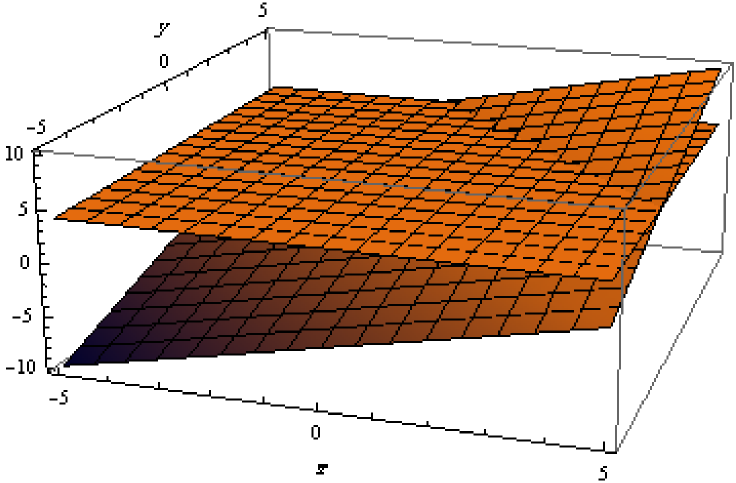
- We get the 3D line

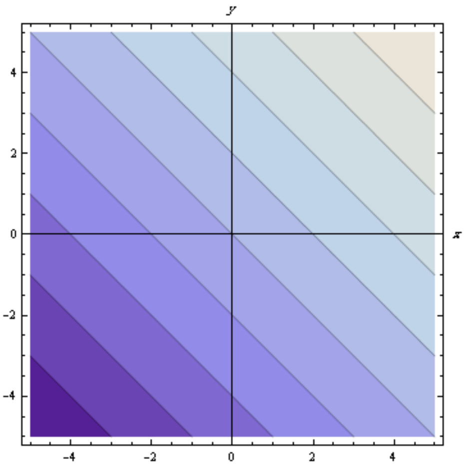
$$f(X, Y) = const,\ Z = const$$

- Project it down to the X-Y plan to get the contour plot

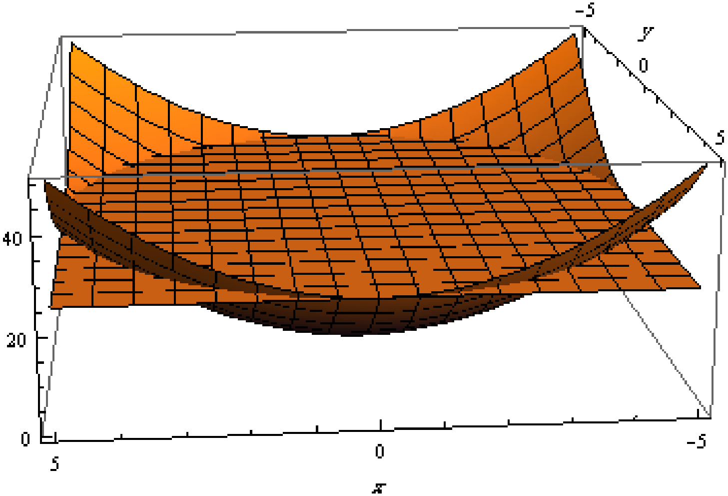
$$f(X, Y) = const,\ Z = 0.$$

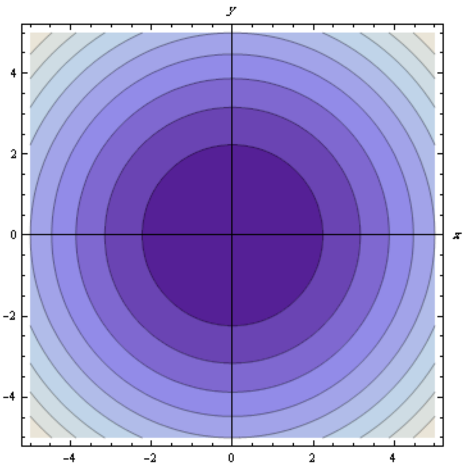
$$Z = X + Y, \; Z = 3$$



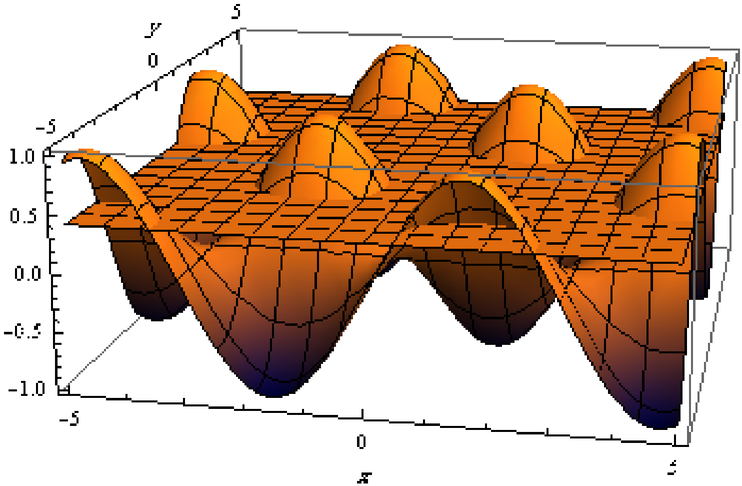


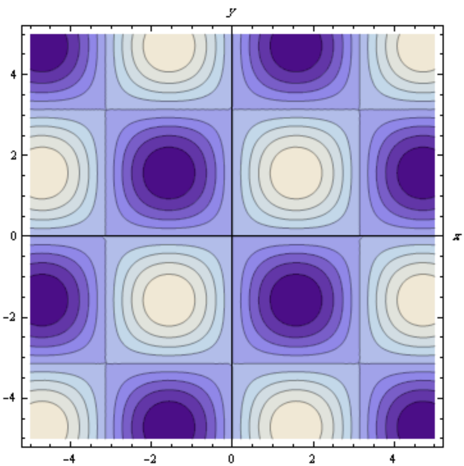
$$Z = X^2 + Y^2, \ Z = 25$$



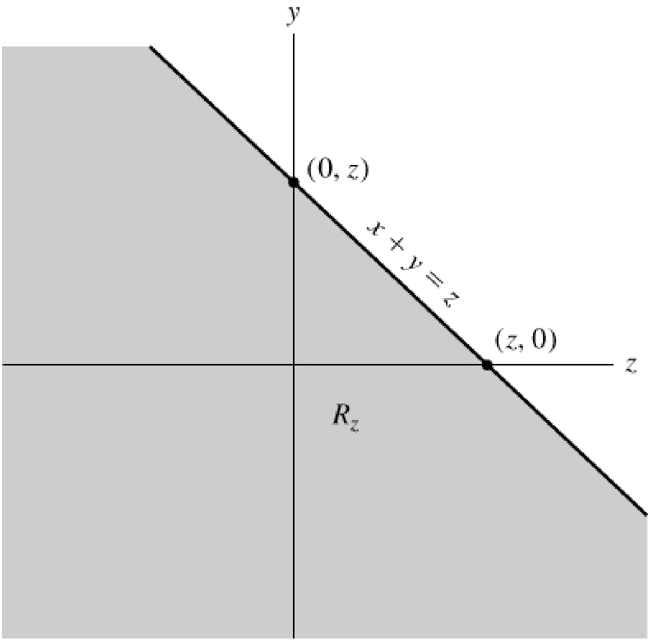


$$Z = \sin(X)\sin(Y),\ Z = 0.4$$





Example 73 (Sums) :



$$Z = X + Y$$

For X, Y *discrete*

$$P(Z = z) = P(X + Y = z)$$
$$= P\left(points\ on\ the\ line : X + Y = z\right)$$
$$p_Z(z) = \sum_{x=-\infty}^{\infty} p_{XY}(x, z - x)$$

E.g, integers: $1 \leq X \leq 3,\ 1 \leq Y \leq 3,\ p_{XY}\left(x, y\right) = \frac{1}{9}$*;Let's draw it:*

$$p_Z(2) = \sum_{x=1}^{1} p(x, 2-x) = \frac{1}{9}$$

$$p_Z(3) = \sum_{x=1}^{2} p(x, 3-x) = \frac{2}{9}$$

$$p_Z(4) = \sum_{x=1}^{3} p(x, 4-x) = \frac{3}{9}$$

$$p_Z(5) = \sum_{x=2}^{3} p(x, 5-x) = \frac{2}{9}$$

$$p_Z(6) = \sum_{x=3}^{3} p(x, 6-x) = \frac{1}{9}$$

What about: $Z = 2X + Y$ *on the same region (HW).*

Continuous case has to be done through CDF:

$$
\begin{aligned}
F_Z(z) = P(Z \leq z) &= P(R_z) \\
&= \iint_{R_z} f_{XY}(x, y)\; dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x, y)\; dy\, dx
\end{aligned}
$$

$$
f_Z(z) = \frac{d}{dz} \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x, y)\; dy\, dx
$$

Substitution

$$
\begin{aligned}
y &= v - x \\
dy &= dv \\
v &= y + x \implies \textit{(same limits)}
\end{aligned}
$$

$$
\begin{aligned}
f_Z(z) &= \frac{d}{dz} \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_{XY}(x, v - x)\; dv\, dx \\
&= \frac{d}{dz} \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{XY}(x, v - x)\; dx\, dv \\
&= \int_{-\infty}^{\infty} f_{XY}(x, z - x)\; dx
\end{aligned}
$$

If the region of f_{XY} is more complicated, integration limits will change and we have to do every thing from scratch. (See next example)

Application:

- *The life time of a system with two independent components is*

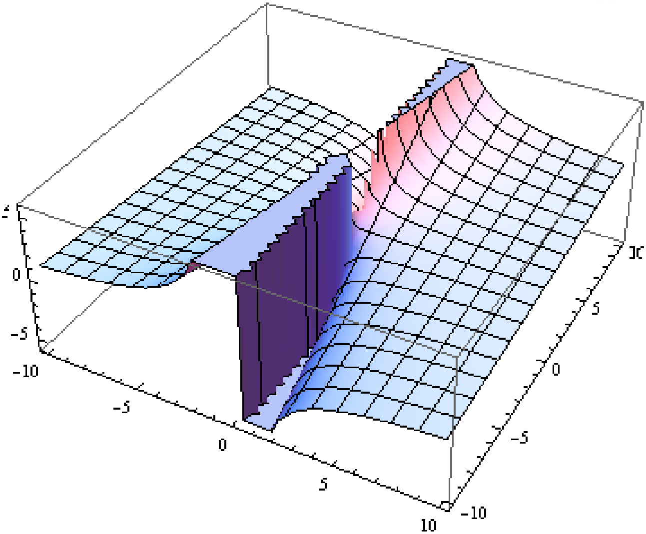
$$S = T_1 + T_2$$

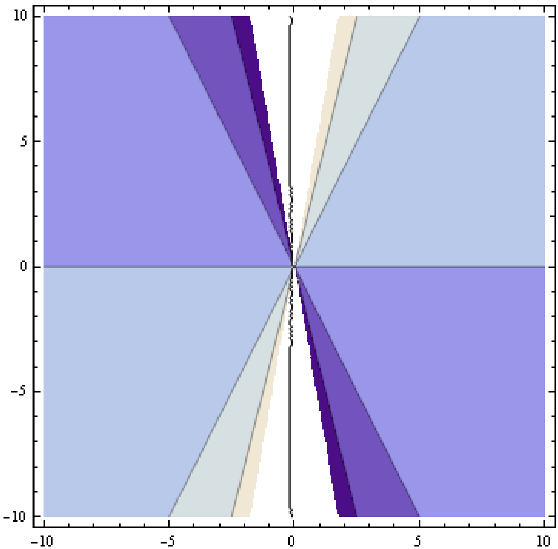
- *Both* T_1 *and* $T_2 \sim Exponential(\lambda)$.

- *Draw to find that* $f_{T_1 T_2}$ *is defined over the region* $0 \le T_1 \le S$ *and* $0 \le T_2 \le S - T_1$.

$$\begin{aligned}
f_S(s) &= \int_{-\infty}^{\infty} f_{T_1 T_2}(t, s-t)\ dt \\
&= \int_{-\infty}^{\infty} f_{T_1}(t)\, f_{T_2}(s-t)\ dt \\
&= \int_{-\infty}^{0} + \int_0^s \left(\lambda e^{-\lambda t}\right)\left(\lambda e^{-\lambda(s-t)}\right)\ dt + \int_s^{\infty} \\
&= \int_0^s \lambda^2 e^{-\lambda s} dt \\
&= \lambda^2 s e^{-\lambda s}.
\end{aligned}$$

Example 74 (Quotients) $: Z = \frac{Y}{X}$





No need to draw the 3D function every time. It suffices to draw the X-Y regions: then

$$
\begin{aligned}
Z &= \frac{Y}{X},\\
Z \leq z &\equiv \frac{Y}{X} \leq z\\
&\equiv \begin{cases} Y \geq zX & X < 0\\ Y \leq zX & X > 0 \end{cases}
\end{aligned}
$$

$$
F_Z(z) = \int_{-\infty}^{0}\int_{xz}^{\infty} f_{XY}(x,y)\,dy\,dx + \int_{0}^{\infty}\int_{-\infty}^{xz} f_{XY}(x,y)\,dy\,dx.
$$

Substitution:

$$
\begin{aligned}
y &= xv,\ v = y/x\\
dy &= x\,dv\\
xz &\Rightarrow z\\
\infty &\Rightarrow -\infty,\ x < 0\\
-\infty &\Rightarrow -\infty,\ x > 0
\end{aligned}
$$

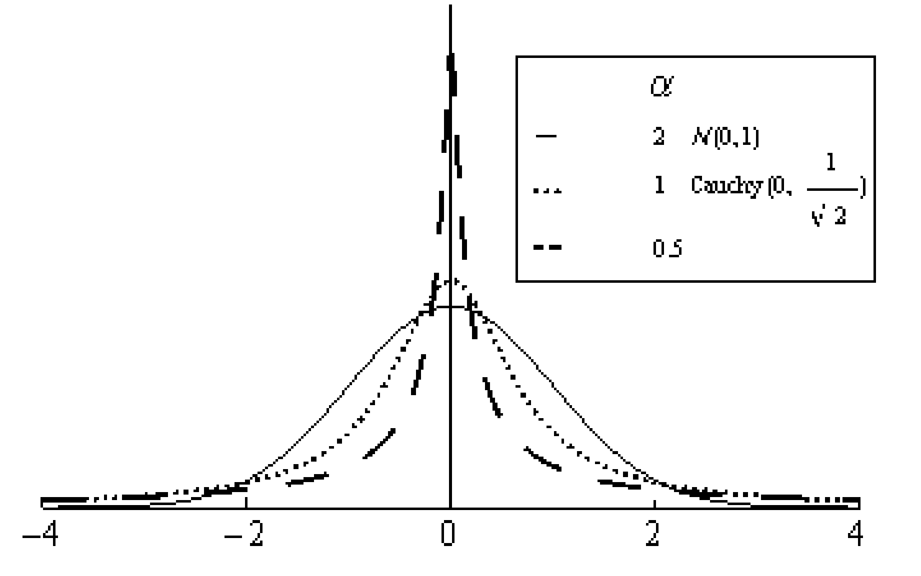
$$
\begin{aligned}
F_Z(z) &= \int_{-\infty}^{0}\int_{z}^{-\infty} x f_{XY}(x, xv)\ dv\, dx \\
&\quad + \int_{0}^{\infty}\int_{-\infty}^{z} x f_{XY}(x, xv)\ dy\, dx \\
&= \int_{-\infty}^{0}\int_{-\infty}^{z} -x f_{XY}(x, xv)\ dv\, dx \\
&\quad + \int_{0}^{\infty}\int_{-\infty}^{z} x f_{XY}(x, xv)\ dv\, dx \\
&= \int_{-\infty}^{z} dv\, [\int_{-\infty}^{0} -x f_{XY}(x, xv)\ dx \\
&\quad + \int_{0}^{\infty} x f_{XY}(x, xv)\ dx] \\
f_Z(z) &= \int_{-\infty}^{0} -x f_{XY}(x, xz)\ dx \\
&\quad + \int_{0}^{\infty} x f_{XY}(x, xz)\ dx
\end{aligned}
$$

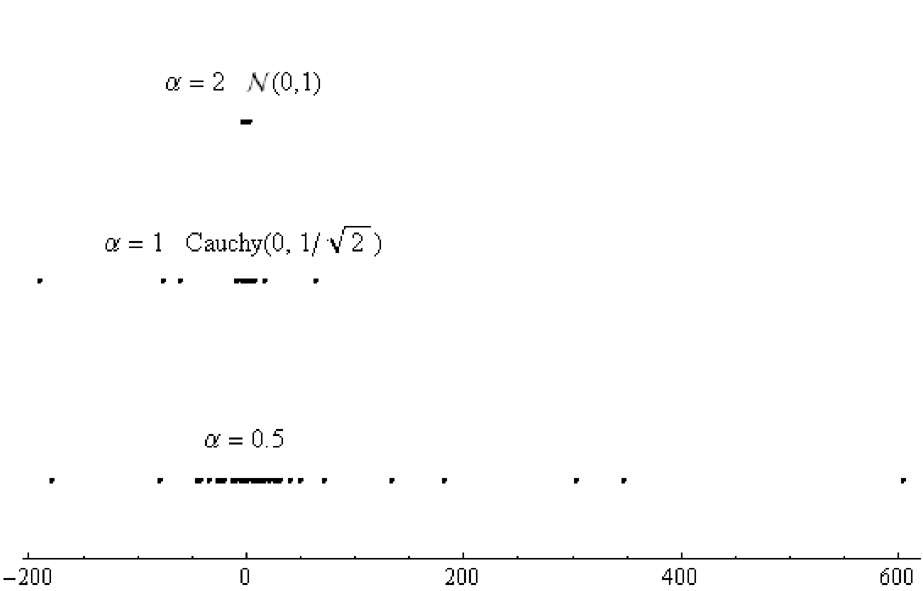
Application: $Z = \frac{Y}{X}$; *and* $X, Y \sim N(0,1)$.

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{0} -x f_{XY}(x, xz)\ dx \\
&\quad + \int_{0}^{\infty} x f_{XY}(x, xz)\ dx \\
&= 2\int_{0}^{\infty} x f_{XY}(x, xz)\ dx \\
&= 2\int_{0}^{\infty} x \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-(xz)^2/2}\right)\ dx \\
&= \frac{1}{\pi}\int_{0}^{\infty} x e^{-x^2\left(z^2+1\right)/2}\ dx \\
&= \frac{1}{\pi}\int_{0}^{\infty} x \frac{-2x\left(z^2+1\right)/2}{-2x\left(z^2+1\right)/2} e^{-x^2\left(z^2+1\right)/2}\ dx \\
&= \frac{1}{\pi\left(z^2+1\right)}\ e^{-x^2\left(z^2+1\right)/2}\Big|_{\infty}^{0} \\
&= \frac{1}{\pi\left(z^2+1\right)}
\end{aligned}
$$

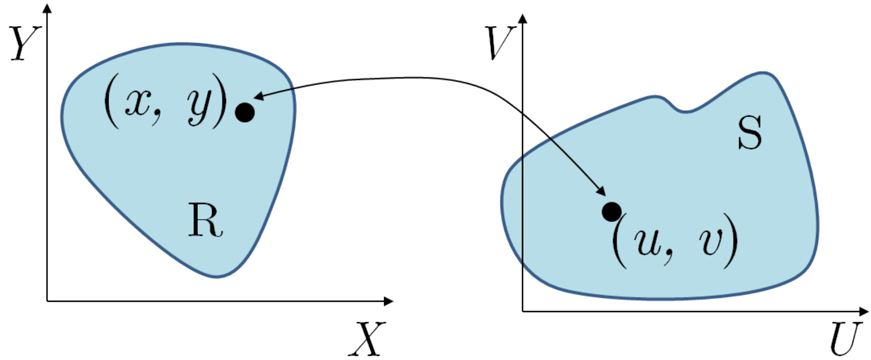
This is the density of Cauchy r.v.

Can be obtained also by univariate transformation!


α
2 N(0,1)
1 Cauchy(0, 1/√2)
0.5
−4
−2
0
2
4


α = 2 N(0,1)
α = 1 Cauchy(0, 1/√2)
α = 0.5
−200
0
200
400
600

3.6.2 p Functions of p r.v.s (Space Transformation):



- Transforming the whole space, not just a single variable $U = f(X, Y)$:

$$
\begin{aligned}
(X, Y) &\rightarrow (U, V), \\
U &= g_1(X, Y), \\
V &= g_2(X, Y), \\
X &= h_1(U, V), \\
Y &= h_2(U, V)
\end{aligned}
$$

- So, joint density f_{UV} is of interest.

Notice: If we want only f_U, $U = f(X, Y)$, then:

1. either as done before.
2. or, make up a variable $V = g_2(X, Y)$, find f_{UV}, then $f_U = \int f_{UV}\, dv$

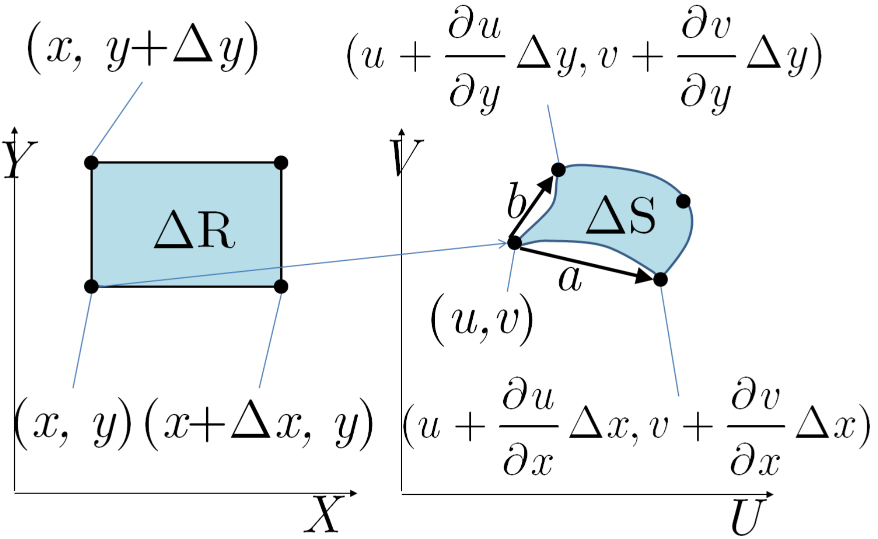
Theorem 75 *Suppose that X and Y are two jointly distributed r.v. with pdf* f_{UV}*, and mapped* ***onto*** *U and V by*

$$\begin{aligned}
u &= g_1(x, y) \\
v &= g_2(x, y), \\
x &= h_1(u, v) \\
y &= h_2(u, v)
\end{aligned}$$

and h_1, h_2, g_1, g_2 *are continuous and having first derivative. Then,*

$$\begin{aligned}
f_{UV}(u, v) &= f_{XY}(x, y)\, \mathbf{J}^{-1}(x, y) \\
\mathbf{J}(x, y) &= \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix}, \\
(x, y) &= (h_1(u, v), h_2(u, v))
\end{aligned}$$

Not a Proof. :



$$
\begin{aligned}
a &= \left(u + \frac{\partial u}{\partial x}\Delta x, v + \frac{\partial v}{\partial x}\Delta x\right) - (u, v) \\
&= \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\right)\Delta x, \\
b &= \left(u + \frac{\partial u}{\partial y}\Delta y, v + \frac{\partial v}{\partial y}\Delta y\right) - (u, v) \\
&= \left(\frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}\right)\Delta y.
\end{aligned}
$$

From elementary vector calculus, it is known that

$$\sin\theta = \frac{|b \times a|}{\|a\|\,\|b\|},$$
$$\cos\theta = \frac{b.a}{\|a\|\,\|b\|}.$$

Therefore, the area of the parallelogram

$$\begin{aligned} A &= \|a\|\,\|b\| \sin\theta \\ &= |a \times b| \\ &= \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} \Delta x\, \Delta y \\ &= \mathbf{J}\left(x, y\right) \Delta x\, \Delta y \end{aligned}$$

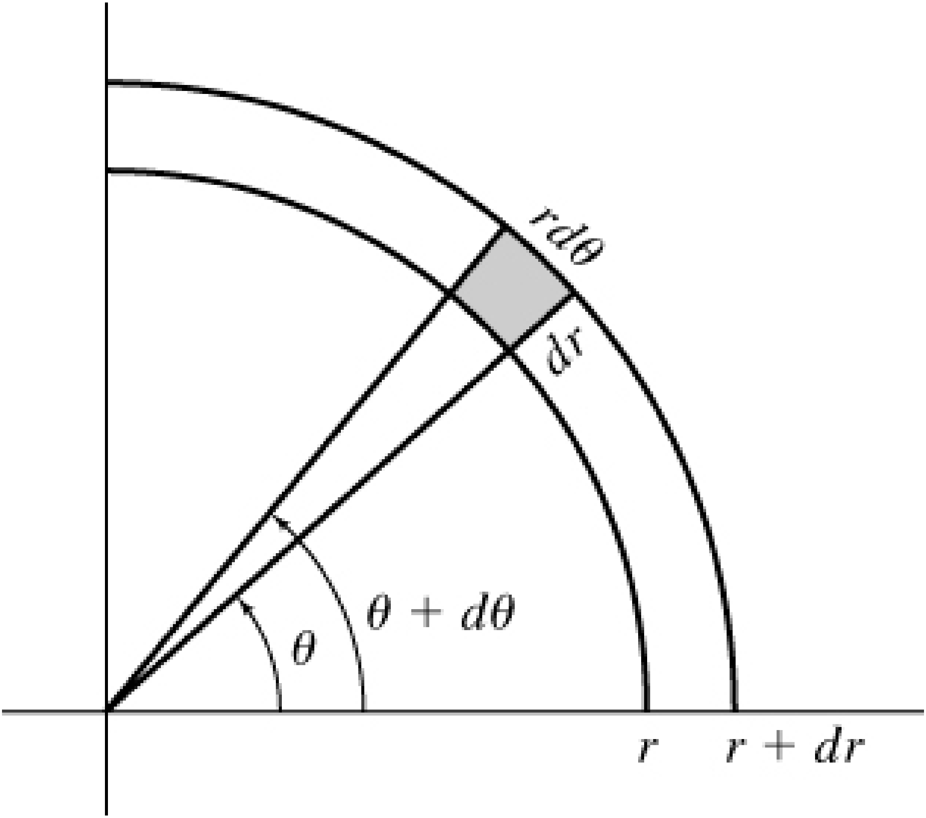
Hence,

$$P(S) = P(R),$$
$$f_{UV}(u, v)\,\mathbf{J}\left(x, y\right) \Delta x\, \Delta y = f_{XY}\left(x, y\right) \Delta x\, \Delta y,$$
$$f_{UV}(u, v) = f_{XY}\left(x, y\right) \mathbf{J}^{-1}\left(x, y\right),$$

where $\left(x, y\right) = (h_1(u, v), h_2(u, v))$. ■

Proof. is out of scope and omitted. ■

Example 76 (Polar System) *: If X and Y have* f_{XY} *what is the pdf of a point selected at radius* r *and angle* θ*?*



Let's draw, regions, transformed regions, and tran- formation functions h, g***:***

$$
\begin{aligned}
&(X,Y) \to (R,\Theta),\ 0 \le R,\ 0 \le \Theta \le 2\pi\\
&\quad R = \sqrt{X^2+Y^2},\\
&\quad \Theta = \begin{cases} \tan^{-1}\left(\frac{Y}{X}\right) & 0 < X\\ \tan^{-1}\left(\frac{Y}{X}\right)+\pi & X<0\\ \frac{\pi}{2}\operatorname{sign}(Y) & X=0, Y\neq 0\\ 0 & X=0, Y=0 \end{cases}\\
&\quad X = R\cos\Theta,\\
&\quad Y = R\sin\Theta,\\
&\quad \mathbf{J} = \begin{vmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y}\\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix}\\
&\quad\ \ = \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}}\\ \frac{1}{1+\left(y/x\right)^2}\frac{-y}{x^2} & \frac{1}{1+\left(y/x\right)^2}\frac{1}{x} \end{vmatrix}\\
&\quad\ \ = \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}}\\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix}\\
&\quad\ \ = 1/\sqrt{x^2+y^2} = \frac{1}{r}\\
&f_{R\Theta}(r,\theta) = r\, f_{XY}(r\cos\theta, r\sin\theta).
\end{aligned}
$$

Example 77 (Rayleigh density) *: Suppose that X and Y are independent standard normal; then*

$$
\begin{aligned}
f_{XY}\left(x,y\right) &= f_X(x)\, f_Y\left(y\right) \\
&= \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x^2+y^2\right)\right], \\
f_{R\Theta}(r,\theta) &= r\, f_{XY}(r\cos\theta, r\sin\theta) \\
&= r\frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x^2+y^2\right)\right] \\
&= \frac{1}{2\pi} r\exp\left[-\frac{1}{2}r^2\right], \\
f_R(r) &= \int_0^{2\pi}\frac{1}{2\pi} r\exp\left[-\frac{1}{2}r^2\right]\, d\theta \\
&= r\exp\left[-\frac{1}{2}r^2\right],\ 0\le r, \\
f_\Theta(\theta) &= \int_0^{\infty}\frac{1}{2\pi} r\exp\left[-\frac{1}{2}r^2\right]\, dr \\
&= \frac{1}{2\pi},\ 0\le\theta\le 2\pi.
\end{aligned}
$$

Notice that:

- *Independence of* R, Θ *is obvious from separability* $f_{R\Theta}(r,\theta)$ *or from:*

  $f_{R\Theta}(r,\theta) = f_R(r)\, f_\Theta(\theta)$.

- *It makes a lot of sense (from symmetry of the problem).*

- *Symmetry is consistent with that:*

  $\Theta \sim Uniform(0, 2\pi)$

- *R is called* $Rayleigh$ *r.v.*

3.7 Extrema and Order Statistics

Suppose that $X_1, \ldots, X_n$ are **continuous** i.i.d from CDF F. What is the pdf of:

$$U = \max(X_i), \; i = 1, \ldots, n$$
$$V = \min(X_i), \; i = 1, \ldots, n$$

What is the meaning of that for observations?

$$X \xrightarrow{Sample_1} x_1, x_2, \ldots, x_n$$
$$X \xrightarrow{Sample_2} x_1, x_2, \ldots, x_n$$
$$\vdots$$

Then, the r.v. $X_1, \ldots, X_n$, as well as of course, X, are i.i.d.

PDF derivation:

$$
\begin{aligned}
F_U(u) &= P(U \le u) \\
&= P(X_1 \le u, \ldots, X_n \le u) \\
&= P(X_1 \le u) P(X_2 \le u) \ldots P(X_n \le u) \\
&= [P(X_i \le u)]^n \\
&= [F(u)]^n, \\
f_U(u) &= nf(u)[F(u)]^{n-1}
\end{aligned}
$$

$$
\begin{aligned}
1 - F_V(v) &= P(V > v) \\
&= P(X_1 > v, \ldots, X_n > v) \\
&= [P(X > v)]^n \\
F_V(v) &= 1 - [1 - F(v)]^n \\
f_V(v) &= nf(v)[1 - F(v)]^{n-1}
\end{aligned}
$$

More intuition by studying $F_U(u)$ for $Uniform$

$$f_U(u) = n(1-u)^{n-1}.$$

Example 78 :

- *n independent components in series.*
- *Each has lifetime $T \sim Exponential(\lambda)$.*
- *The system fails when any component fails.*
- *V is the system lifetime; what is f_V?*

$$
\begin{aligned}
V &= \min(T_i), \; i = 1, \ldots, n \\
f_V(v) &= n f_T(v) \left[1 - F_T(v)\right]^{n-1} \\
&= n\left(\lambda e^{-\lambda v}\right)\left(1 - \left(1 - e^{-\lambda v}\right)\right)^{n-1} \\
&= n\lambda e^{-n\lambda v}.
\end{aligned}
$$

Therefore, $V \sim Exponential(n\lambda)$; $\rightarrow$ decays faster $\rightarrow$ concentrated at low values (ma a lot of sense).

If they are connected in parallel, then

$$
\begin{aligned}
U &= \max(T_i), \; i = 1, \ldots, n \\
f_U(u) &= n f_T(u) \left[F(u)\right]^{n-1} \\
&= n\lambda e^{-\lambda u}\left(1 - e^{-\lambda u}\right)^{n-1}
\end{aligned}
$$

Theorem 79 (Order Statistics) *:*
The density of the kth-order statistic $X_{(k)}$ is

$$f_{(k)}(x) = \frac{n!}{(k-1)!\,(n-k)!} f_X(x)\, F^{k-1}(x)\,[1-F(x)]^{n-k}$$

Proof.

$$P(X \le x) = F(x)$$

$$P\left(j \text{ observations } \le x\right) = Binomial(n, F(x)) = \binom{n}{j} F^j(x)\,(1-F(x))^{n-j}$$

$$F_{(k)}(x) = P(\textbf{at least } k \text{ observations } \le x) = \sum_{j=k}^{n} \binom{n}{j} F^j(x)\,[1-F(x)]^{n-j}$$

$$f_{(k)}(x) = \sum_{j=k}^{n} \binom{n}{j} j\, f(x)\, F^{j-1}(x)\,[1-F(x)]^{n-j} - \sum_{j=k}^{n} \binom{n}{j} F^j(x)\,(n-j)\, f(x)\,[1-F(x)]^{n-j-1}$$

$$
\begin{aligned}
f_{(k)}(x) = \binom{n}{k} k f(x) F^{k-1}(x) [1 - F(x)]^{n-k} \\
+ \sum_{j=k+1}^{n} \binom{n}{j} j f(x) F^{j-1}(x) [1 - F(x)]^{n-j} \\
- \sum_{j=k}^{n-1} \binom{n}{j} F^{j}(x) (n - j) f(x) [1 - F(x)]^{n-j-1}
\end{aligned}
$$

$$
\begin{aligned}
f_{(k)}(x) = \binom{n}{k} k f(x) F^{k-1}(x) [1 - F(x)]^{n-k} \\
+ \sum_{j=k}^{n-1} \underline{\binom{n}{j+1} (j+1)} f(x) F^{j}(x) [1 - F(x)]^{n-j-1} \\
- \sum_{j=k}^{n-1} \underline{\binom{n}{j} (n - j)} F^{j}(x) f(x) [1 - F(x)]^{n-j-1}
\end{aligned}
$$

$$
f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} f_X(x) F^{k-1}(x) [1 - F(x)]^{n-k}
$$

which completes the proof. ∎

Example 80 *Find $f_{(k)}$ when $X \sim Uniform(0,1)$.*

$$f(x) = 1,$$

$$F(x) = x,$$

$$\begin{aligned}
f_{(k)}(x) &= \frac{n!}{(k-1)!\,(n-k)!} f(x)\, F^{k-1}(x) \left[1 - F(x)\right]^{n-k} \\
&= \frac{n!}{(k-1)!\,(n-k)!} x^{k-1} (1-x)^{n-k},\ 0 \le x \le 1
\end{aligned}$$

Notice: $X \sim Beta(k, n-k+1)$

Chapter 4

Expected Values

4.1 The Expected Value of a r.v.

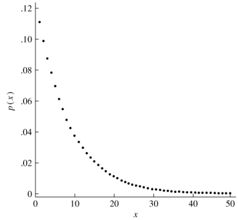
Definition 81 *If X is discrete r.v. with pmf $p(x)$, the expected value (mean) is defined as*

$$E(X) = \sum_i x_i p(x_i),$$

Provided that $\sum_i |x_i| p(x_i) < \infty$; otherwise, it is undefined.

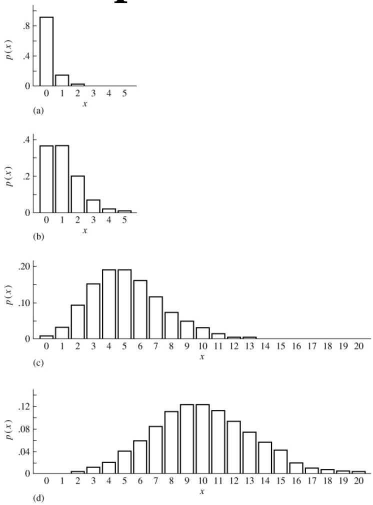
- $E(X)$ or μ_X
- It is a weighted sum
- It is also the center of mass.
- It is also the point of balance.
- All the above in the mathematical sense:

$$\sum_i x_i p(x_i)$$

Example 82 ($Geometric(p)$) : 

$$\begin{aligned} p_X(k) &= p\left(1-p\right)^{k-1},\ 1 \le k \\ E(X) &= \sum_{k=1}^{\infty} kpq^{k-1} \\ &= p\sum_{k=1}^{\infty} kq^{k-1} \\ &= p\sum_{k=1}^{\infty} \frac{d}{dq}q^k \\ &= p\frac{d}{dq}\sum_{k=1}^{\infty} q^k \\ &= p\frac{d}{dq}\frac{q}{1-q} \\ &= \frac{p}{\left(1-q\right)^2} \\ &= \frac{1}{p} \end{aligned}$$

Example 83 ($Poisson(\lambda)$) :



$$
\begin{aligned}
E(X) &= \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda
\end{aligned}
$$

Compare this to the peak that occurs at $k = \lfloor \lambda \rfloor$ *(coincidence).*

Example 84 (Decisions in life) :

- *Deciding A results in loss (-ve) or gain (+ve).*
- *with probabilities p and* $(1-p)$

$$L_A = \begin{cases} L_1 & p \\ L_2 & (1-p) \end{cases},$$

$$E(L_A) = L_1 p + L_2 (1-p)$$

- *Most times* $L_2 = 0$.
- *Travel (A) or not* (A')*?*
- *Calculate* $E(L_A)$ *and* $E(L_{A'})$.
- *the loss and gain are subjective.*
- *the self-consciousness of p is subjective too.*
- *the decision ultimately differs across people.*
- *"Acquire the most beneficial and prevent the most harmful"*

Definition 85 *If X is continuous r.v. then*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx,$$

If $\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$; otherwise, it is undefined.

Example 86 **($Gamma(\alpha,\lambda)$)** *:*

$$\begin{aligned}
E(X) &= \int_0^{\infty} x \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\
&= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \int_0^{\infty} \underbrace{\frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^{\alpha} e^{-\lambda x}}_{pdf\ of\ Gamma(\alpha+1,\lambda)}\ dx \\
&= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{1}{\lambda} \\
&= \frac{\alpha}{\lambda}.
\end{aligned}$$

Notice that: $Exponential(\lambda)$ is $Gamma(1,\lambda)$, hence its mean is $\frac{1}{\lambda}$.

Example 87 **($Normal(\mu, \sigma^2)$)** :

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right] dx$$

Substitute $z = x - \mu$

$$E(X); = \int_{-\infty}^{\infty} \underbrace{z \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z)^2}{2\sigma^2}\right]}_{symmetric} dz$$

$$+ \int_{-\infty}^{\infty} \mu \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z)^2}{2\sigma^2}\right]}_{pdf\ of\ Normal} dz$$

$$= \mu$$

Example 88 ($Cauchy$) :

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\pi} \frac{1}{1+x^2} dx$$
$$= \infty - \infty$$
$$undefined!!$$

Why

$$E(X) \neq 0\,?$$

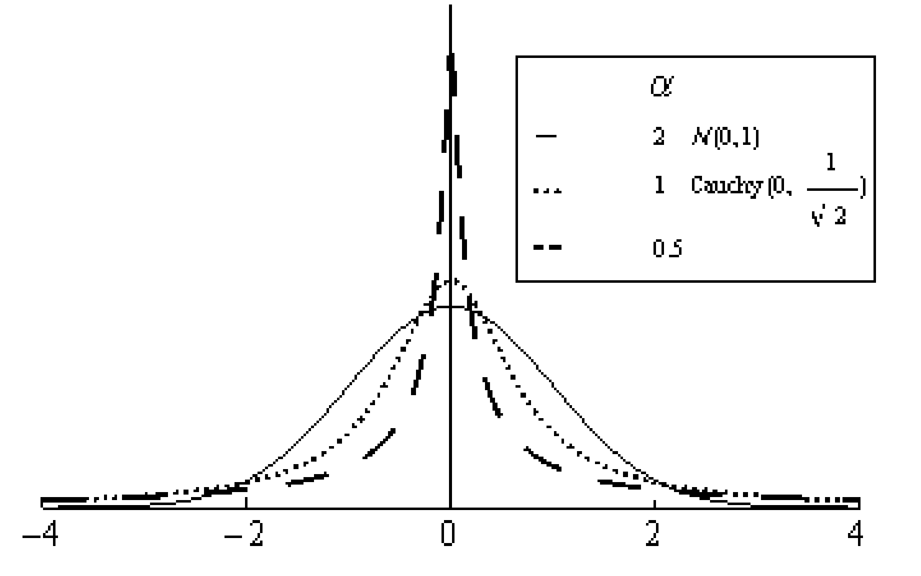
Because

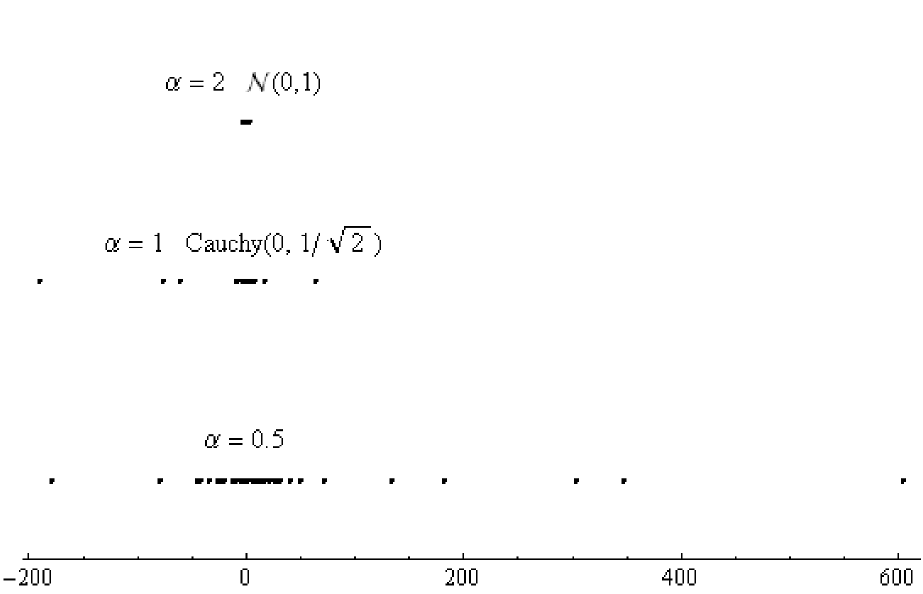
$$\int_{-\infty}^{\infty} |x| \frac{1}{\pi} \left(\frac{1}{1+x^2} \right) dx = \infty.$$

We will see later that this has a serious impact on the sample mean:

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

Please, recall the following figures:


α
2 N(0,1)
1 Cauchy(0, 1/√2)
0.5
−4
−2
0
2
4


α = 2 N(0,1)
α = 1 Cauchy(0, 1/√2)
α = 0.5
−200
0
200
400
600

Theorem 89 (Markov's Inequality) : *If X is a r.v. with $P(X \geq 0) = 1$, then*

$$P(X \geq t) \leq \frac{E(X)}{t} \; \forall t.$$

Proof.

$$\begin{aligned}
E(X) &= \int x f(x)\, dx \\
&= \underbrace{\int_{x<t} x f(x)\, dx}_{\geq 0} + \int_{x \geq t} x f(x)\, dx \\
&\geq \int_{x \geq t} x f(x)\, dx \\
&\geq t \int_{x \geq t} f(x)\, dx \\
&= t P(X \geq t).
\end{aligned}$$

Intuition:

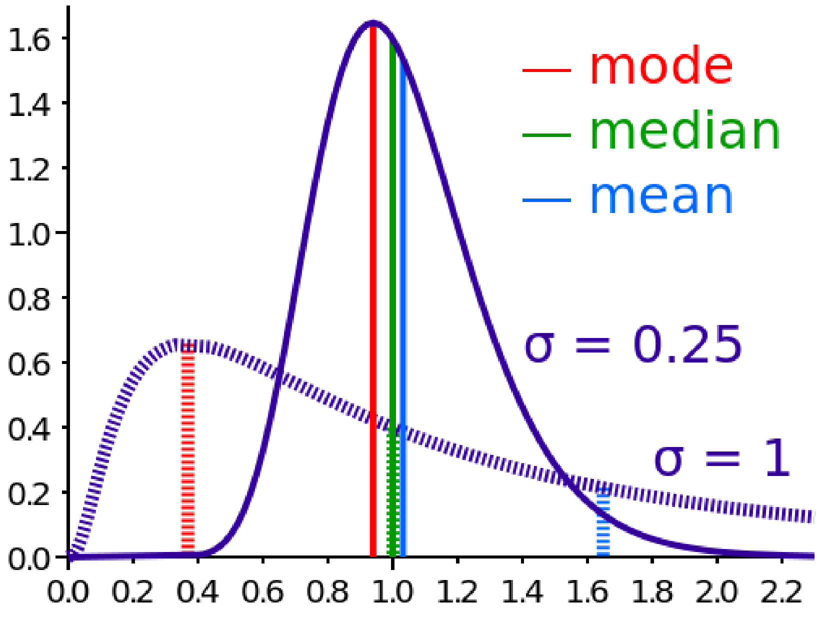
$$P(X \geq kE(X)) \leq 1/k.$$

■

Mean, Median, and Mode

- They do not have to coincide
- Any of them can be the left/right to another.
- More on that when discussing skewness.
- Example: Lognormal density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(-\mu + \log x)^2}{2\sigma^2}\right], \ 0 < x$$



4.1.1 Expectations of Functions of r.v.

Theorem 90 *Suppose that $Y = g(X)$, then for discrete X:*

$$E(Y) = \sum_x g(x)\, p(x),$$

if $\sum \left|g(x)\right| p(x) < \infty$*; and ior continuous X:*

$$E(Y) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx,$$

if $\int g(x)\, f(x)\, dx < \infty$.

This is much easier than:

$$f_Y(y) = f_X\left(g^{-1}(x)\right) \frac{1}{\left|dy/dx\right|}$$

then

$$E(Y) = \int y f_Y(y)\, dy.$$

Proof. For discrete case

$$p\left(Y=y_i\right)=p(\underbrace{\{x:g(x)=y_i\}}_{A_i})$$

$$\begin{aligned}
E(Y) &= \sum_i y_i p\left(y_i\right) \\
&= \sum_i y_i p(A_i) \\
&= \sum_i y_i \sum_{x\in A_i} p(x) \\
&= \sum_i \sum_{x\in A_i} y_i p(x) \\
&= \sum_x g(x)\, p(x)
\end{aligned}$$

■

Not a Proof. continuous case

$$\begin{aligned}
E(Y) &= \int y \underline{f_Y\left(y\right) dy} \\
&= \int g(x)\, \underline{f_X(x)\, dx}
\end{aligned}$$

■

Example 91 (Kinetic Energy) :

- *The velocity of a gas molecule is a r.v. with*

$$f_X(x) = \frac{\sqrt{2/\pi}}{\sigma^3} x^2 \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right]$$

- *The kinetic energy is*

$$Y = \frac{1}{2} m X^2.$$

- *What is the mean kinetic energy*

$$E(Y) = \int_0^\infty \frac{1}{2} m x^2 f_X(x)\, dx$$

$$= \frac{m}{\sqrt{2\pi}\sigma^3} \int_0^\infty x^4 \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right] dx$$

Change variables

$$u = x^2/2\sigma^2$$

$$x = \sqrt{2}\sigma u^{1/2}$$

$$dx = \frac{1}{\sqrt{2}}\sigma u^{-1/2}$$

$$\int_0^\infty \Longrightarrow \int_0^\infty$$

$$
\begin{aligned}
E(Y) &= \frac{2m\sigma^2}{\sqrt{\pi}} \int_0^\infty u^{3/2} \exp[-u]\, du \\
&= \frac{2m\sigma^2}{\sqrt{\pi}} \Gamma(3/2+1) \\
&= \frac{2m\sigma^2}{\sqrt{\pi}} 1.5 \times .5 \times \Gamma(.5) \\
&= \frac{3}{2} m\sigma^2.
\end{aligned}
$$

Notice we have used

$$
\Gamma(\alpha+1) = \alpha\Gamma(\alpha),
$$
$$
\Gamma(.5) = \sqrt{\pi}
$$

Theorem 92 (Generalization) *:Suppose $X_1, \dots X_n$ are jointly distributed r.v., and that $Y = g(X_1, \dots, X_n)$. Then*

1. *For discrete case:*

$$E(Y) = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n)\, p(x_1, \dots, x_n),$$

provided that

$$\sum_{x_1, \dots, x_n} \left| g(x_1, \dots, x_n) \right| p(x_1, \dots, x_n) < \infty$$

2. *For continuous case:*

$$E(Y) = \int \dots \int g(x_1, \dots, x_n) \times f(x_1, \dots, x_n)\, dx_1 \dots dx_n,$$

provided that

$$\int \dots \int \left| g(x_1, \dots, x_n) \right| f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

Corollary 93 *If X and Y are independent r.v. then*

$$E\left(g(X)\,h(Y)\right) = E\left(g(X)\right)E\left(h(Y)\right),$$

and in particular

$$E(XY) = E(X)\,E(Y).$$

Proof. Is trivial and left as an exercise (Problem 29). ■

4.1.2 Expectations of Linear Combinations of r.v.

Theorem 94 *Suppose $X_1, \ldots, X_n$ are jointly distribu* ***NOT NECESSARILY INDEPENDENT*** *and $E(X_i)$ ex- ists, and $Y = a + \sum_i b_i X_i$. Then*

$$E(Y) = a + \sum_i b_i E(X_i).$$

Proof. For continuous case:

$$\begin{aligned}
E(Y) &= \int \ldots \int \left(a + \sum_i b_i x_i\right) f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= \int \ldots \int a f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&\quad + \sum_i b_i \int \ldots \int x_i f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= a + \sum_i b_i \int x_i f(x_i)\, dx_i \\
&= a + \sum_i b_i E(X_i).
\end{aligned}$$

■

Example 95 (Mean of binomial) :

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

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

$$E(Y) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k},$$

which is not straight forward. However, thanks to indicator variables:

$$Binomial(n, p) \sim \sum_{i=1}^{n} Bernoulli(p),$$

$$Y = \sum_i I_i$$

$$\begin{aligned} E(Y) &= E\left(\sum_i I_i\right) \\ &= \sum_{i=1}^{n} E(I_i) \\ &= n\left(p \times 1 + (1-p) \times 0\right) \\ &= np. \end{aligned}$$

Example 96 (Coupon Collection) :

- *Collect n equal-likely different coupons*
- *# of rials required is r.v..*
- *Getting pmf may be hard; what about mean?*

$$X = \sum_{r=1}^{n} X_r,$$

$$X_r = \#\textit{trials to get } r^{th} \textit{ coupon}$$
$$\textit{after getting } r-1$$

$$X_r \sim Geometric\left(\frac{n-(r-1)}{n}\right),$$

$$E(X_r) = \frac{1}{p} = \frac{n}{n-r+1},$$

$$E(X) = \sum_{r=1}^{n} \frac{n}{n-r+1}$$
$$= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + \frac{n}{1}$$
$$= n\sum_{r=1}^{n} \frac{1}{r}.$$

E.g., for $n = 10$, $E(X) = 29.3$.

4.2 Variance and Standard Deviation

Definition 97 *If X is r.v with $E(X)$, then*

$$\sigma^2 \equiv \text{Var}(X) = E\left[(X - E(X))^2\right],$$
$$\sigma \equiv \text{SD}(X) = \sqrt{\text{Var}(X)},$$

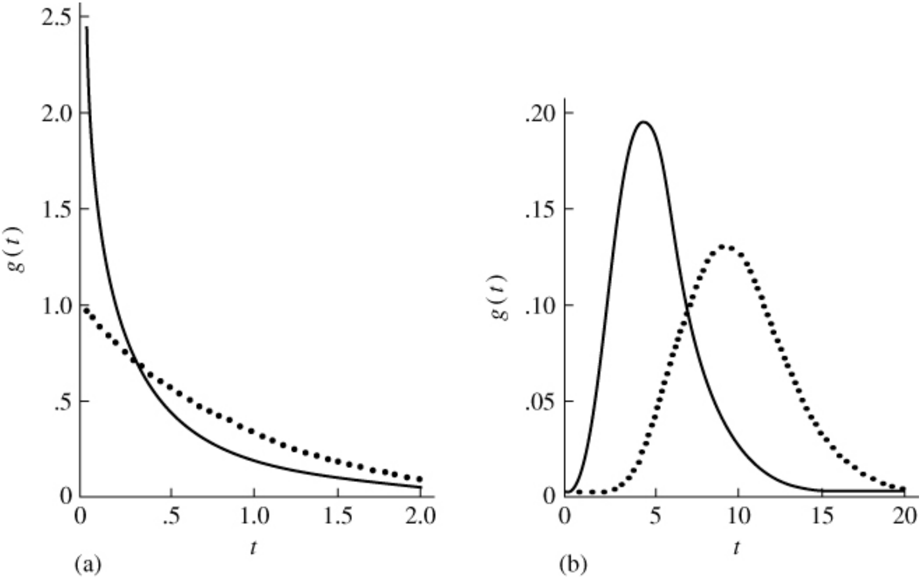
provided that $E\left[(X-E(X))^2\right] < \infty$.

Intuition:

- we need some measure for "dispersion".
- SD has same units, so more meaningful.
- What is the pdf of $Y = X - E(X)$
- We could have defined it as:

$$\underset{2}{\text{Var}}(X) = E\left(|X - E(X)|\right),$$

which is called absolute deviance.



2.5
2.0
1.5
1.0
.5
0
g(t)
0
.5
1.0
1.5
2.0
t
(a)
.20
.15
.10
.05
0
g(t)
0
5
10
15
20
t
(b)

Theorem 98 *If* Var(X) *exists, and* $Y = a+bX$ *then*

$$\mathrm{Var}(Y) = b^2 \mathrm{Var}(X).$$

Proof.

$$\begin{aligned}
\mathrm{Var}(Y) &= E(Y - E(Y))^2 \\
&= E(a + bX - [a + bE(X)])^2 \\
&= E(b(X - E(X)))^2 \\
&= b^2 E(X - E(X))^2 \\
&= b^2 \mathrm{Var}(X) \\
\mathrm{SD}(Y) &= |b| \mathrm{SD}(X),
\end{aligned}$$

■

which makes a lot of sense (Why?).

HW Problem: If $Z = \left(X - \mu_X\right)/\sigma_X$, then

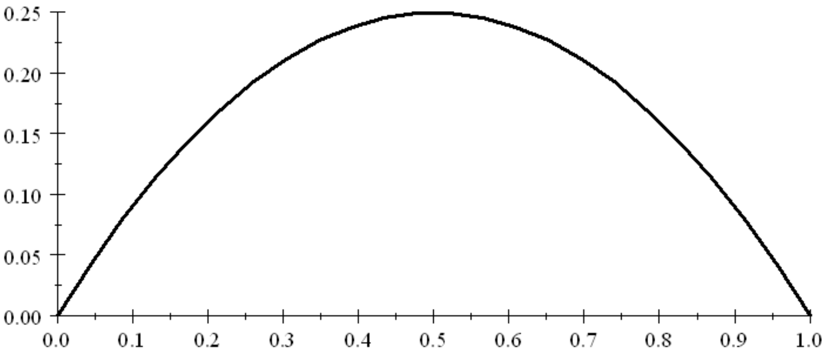
$$E(Z) = 0,$$

$$\mathrm{Var}(Z) = 1.$$

Example 99 (Bernoulli r.v.) :

$$\begin{aligned}
\operatorname{Var}(X) &= E\left(X - E(X)\right)^2 \\
&= \sum_x \left(x - E(X)\right)^2 p(x) \\
&= \sum_x \left(x - p\right)^2 p(x) \\
&= \left(1 - p\right)^2 p + \left(0 - p\right)^2 \left(1 - p\right) \\
&= \left(\left(1 - p\right) + \left(p\right)\right) p \left(1 - p\right) \\
&= p\left(1 - p\right).
\end{aligned}$$

Does it make sense?



Example 100 (Normal Distribution) :

$$\mathrm{Var}(X) = E(X-\mu)^2$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] dx,$$

Substitue (to transform to standard normal):

$$z = \frac{x-\mu}{\sigma}$$

$$dx = \sigma\, dz$$

$$\int_{-\infty}^{\infty} dx \Longrightarrow \int_{-\infty}^{\infty} dz$$

$$\mathrm{Var}(X) = \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 \exp\left[-z^2/2\right] dz$$

Substitute

$$u = z^2/2$$

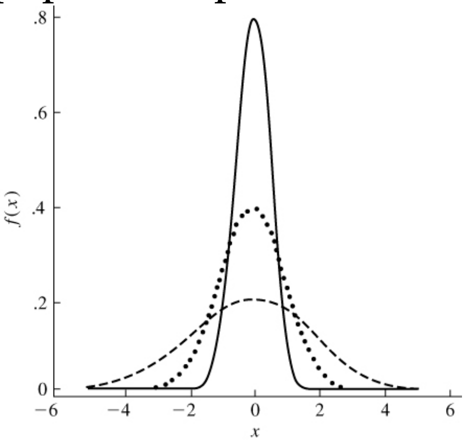
$$z = \sqrt{2}u^{1/2}$$

$$dz = \frac{\sqrt{2}}{2} u^{-1/2} du$$

$$\int_{-\infty}^{\infty} dz \Longrightarrow 2\int_{0}^{\infty} du$$

$$
\begin{aligned}
\operatorname{Var}(X) &= \sigma^2 2\int_0^\infty \frac{1}{\sqrt{2\pi}} 2u \exp[-u]\,\frac{\sqrt{2}}{2} u^{-1/2} du \\
&= \sigma^2 \frac{2}{\sqrt{\pi}} \int_0^\infty u^{1/2} e^{-u} du \\
&= \sigma^2 \frac{2}{\sqrt{\pi}} \Gamma(1.5) \\
&= \sigma^2 \frac{2}{\sqrt{\pi}} (.5)\,\Gamma(.5) \\
&\sigma^2
\end{aligned}
$$

Then, wonderful; the Normal pdf is expressed in terms of its population parameters



Corollary 101 *The variance, if exists, can be given by*

$$\operatorname{Var}(X) = E\left(X^2\right) - [E(X)]^2$$

Proof.

$$\begin{aligned}
\operatorname{Var}(X) &= E\left(X - \mu\right)^2 \\
&= E\left(X^2 - 2X\mu - \mu^2\right) \\
&= E\left(X^2\right) - 2\mu^2 + \mu^2 \\
&= E\left(X^2\right) - \mu^2.
\end{aligned}$$

■

Example 102 $(Uniform(a,b))$:

$$
\begin{aligned}
E(X) &= \int_a^b x f(x)\,dx \\
&= \int_a^b x \frac{1}{b-a} dx \\
&= \frac{1}{b-a}\,\frac{1}{2}x^2\Big|_a^b \\
&= \frac{b^2-a^2}{2(b-a)} \\
&= \frac{1}{2}(b+a) \\
E\left(X^2\right) &= \frac{1}{b-a}\int_a^b x^2 dx \\
&= \frac{b^3-a^3}{3(b-a)} \\
&= \frac{1}{3}\left(b^2+ab+b^2\right) \\
\mathrm{Var}(X) &= E\left(X^2\right) - (E(X))^2 \\
&= \frac{1}{3}\left(b^2+ab+b^2\right) - \frac{1}{4}(b+a)^2 \\
&= \frac{(b-a)^2}{12}
\end{aligned}
$$

Theorem 103 (Chebyshev's Inequality) *:Let X be a r.v. with μ and σ², then*

$$P\left(\left|X-\mu\right| \geq t\right) \leq \frac{\sigma^2}{t^2},\ \forall t > 0.$$

Proof. Set $Y = (X-\mu)^2$; then from Markov's Inequality

$$P\left(Y \geq t^2\right) \leq \frac{E(Y)}{t^2},\ \forall t > 0,\ Y \geq 0.$$

Then,

$$P\left((X-\mu)^2 \geq t^2\right) \leq \frac{E(X-\mu)^2}{t^2}$$

$$P\left(\left|X-\mu\right| \geq t\right) \leq \frac{\sigma^2}{t^2}.$$

■

Intuition: Probability of falling t-far from the mean $\nearrow \sigma^2$ and $\searrow t$

Corollary 104 *If* $\mathrm{Var}(X) = 0$, *then*

$$P\left(X = \mu\right) = 1.$$

Proof. Of course, it makes a lot of sense. But rigorously, from Chebyshev:

$$\begin{aligned} P\left(\left|X - \mu\right| \geq t\right) &\leq \frac{\sigma^2}{t^2} \; \forall t \neq 0, \\ &= 0. \end{aligned}$$

If $P\left(X = \mu\right) \neq 1$, then

$$\begin{aligned} P\left(X \neq \mu\right) &\neq 0, \\ P\left(\left|X - \mu\right| \geq t\right) &\neq 0, \text{ for some } t \neq 0, \end{aligned}$$

which contradicts with above. ∎

Example 105 *[Normal and Chebyshev]:For* $X \sim Normal(0,1)$*, we have learned before (Example 51):*

$$P\left(\left|X-\mu\right|>2\sigma\right)=0.0455$$

However, from Chebyshev

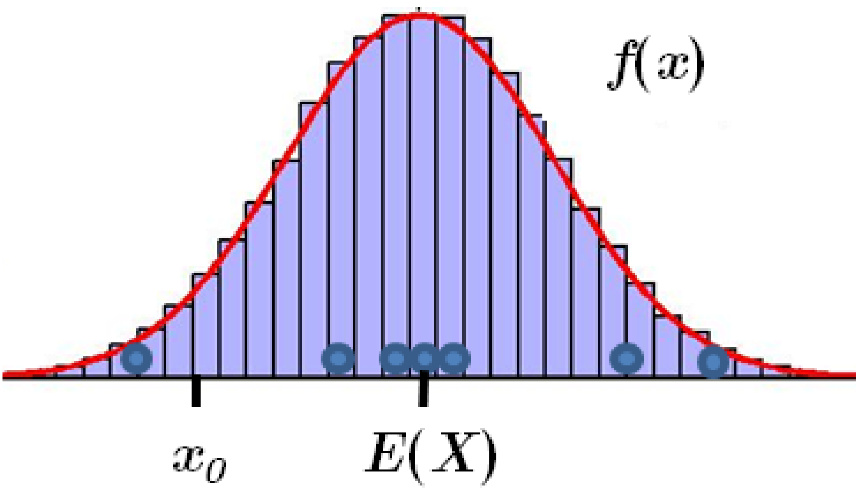
$$P\left(\left|X-\mu\right|\geq t\right)\leq\frac{\sigma^2}{t^2}$$
$$P\left(\left|X-\mu\right|\geq 2\sigma\right)\leq .25,$$

which is not tight as obtained from Normal density directly; because this is ***distribution-free*** *result*

4.2.1 A Model for Measurement Error

- Suppose that we measure a constant x_0.
- Measurements are r.v. X, with μ and σ
- The error $X - x_0$ is a r.v; analyze it!
- Mean Squared Error (MSE):

$$MSE = E(X - x_0)^2$$



Theorem 106 (Mean Squared Error (MSE)) :

$$\begin{aligned} MSE &= Variance + Bias^2 \\ &= \sigma^2 + \left(\mu - x_0\right)^2. \end{aligned}$$

Proof.

$$\begin{aligned} MSE &= E\left(X - x_0\right)^2 \\ &= \operatorname{Var}\left(X - x_0\right) + \left[E\left(X - x_0\right)\right]^2 \\ &= \operatorname{Var}(X) + \left(\mu - x_0\right)^2. \end{aligned}$$

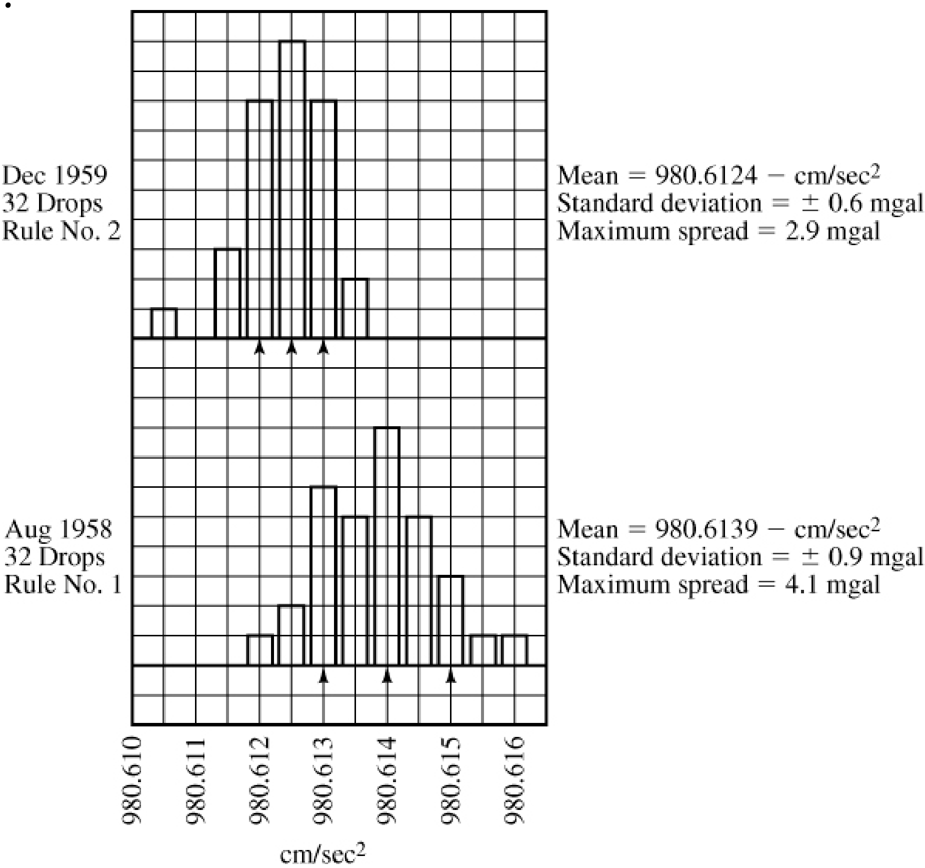
Another Proof (common trick):

$$\begin{aligned} MSE &= E\left(X - x_0\right)^2 \\ &= E\left(\left(X - \mu\right) + \left(\mu - x_0\right)\right)^2 \\ &= E\left(X - \mu\right)^2 + 2\left(\mu - x_0\right)E\left(X - \mu\right) + \left(\mu - x_0\right)^2 \\ &= \sigma^2 + \left(\mu - x_0\right)^2 \end{aligned}$$

■

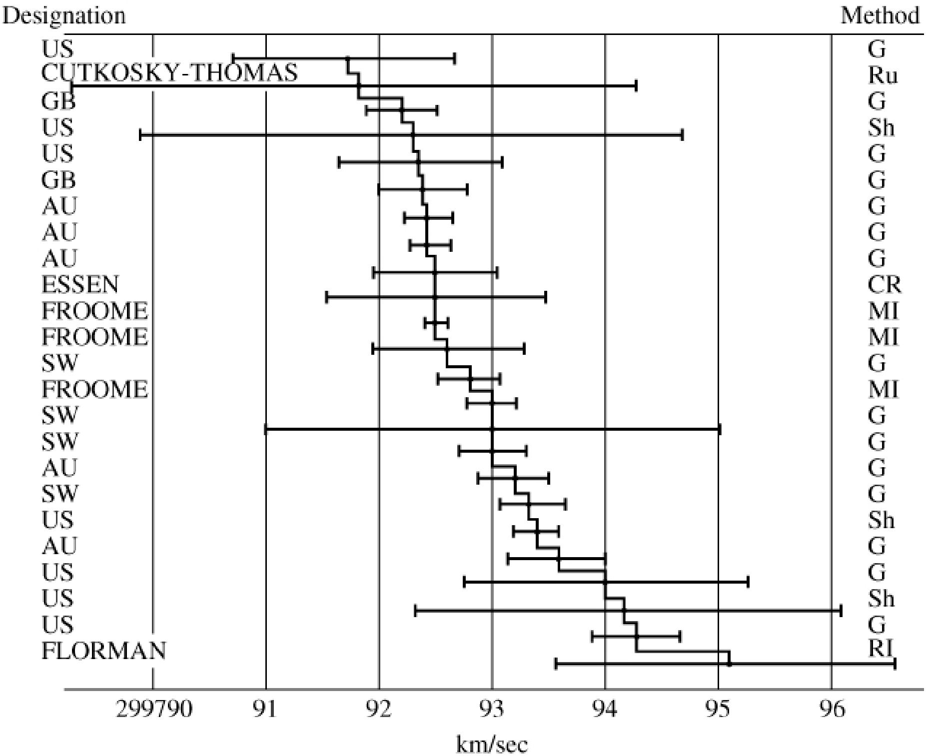
Example 107 (Measurement of the Gravity Const

:



Example 108 (Measurement of the Speed Light)

:



4.3 Covariance & Correlation

This section should have impact on your way of thinking and reading different situations in life

Definition 109 *If X and Y are two r.v. with μ_X and μ_Y*

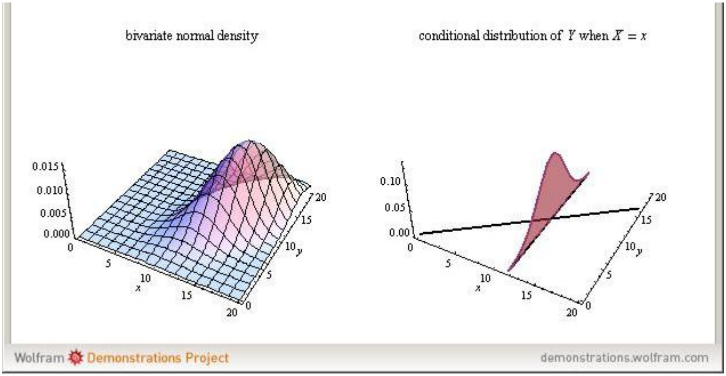
$$\operatorname{Cov}(X,Y) = E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right],$$

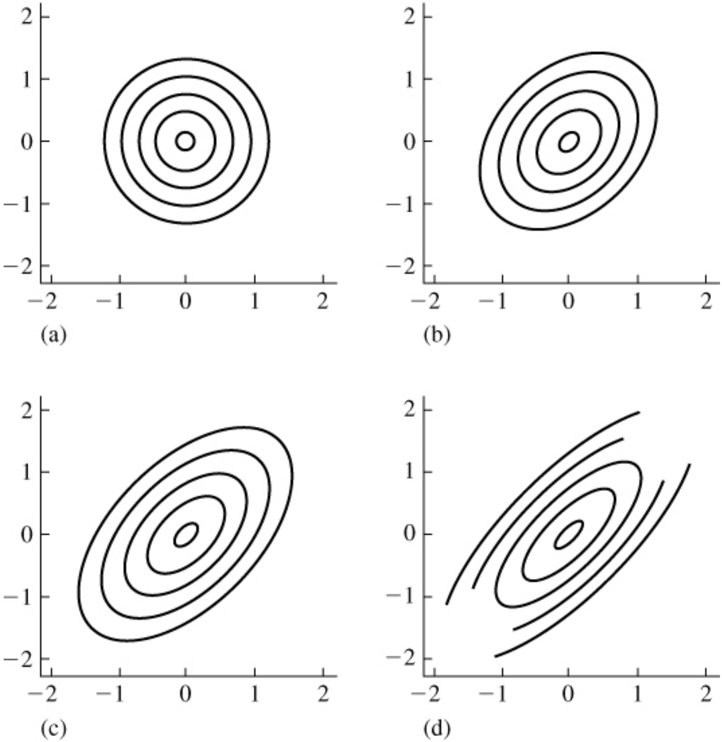
which can be rewritten as

$$\begin{aligned}
\operatorname{Cov}(X,Y) &= E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] \\
&= E\left[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y\right] \\
&= E[XY]-\mu_X\mu_Y-\mu_X\mu_Y+\mu_X\mu_Y \\
&= E[XY]-\mu_X\mu_Y.
\end{aligned}$$

Intuition:

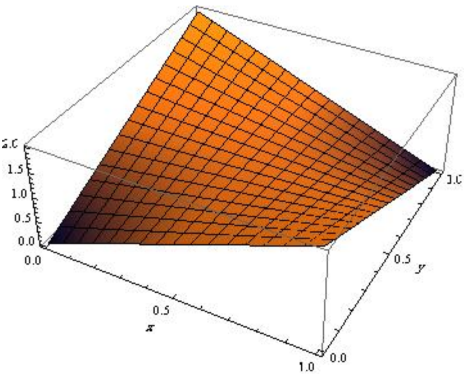
- we need to measure "Association".
- $\operatorname{Cov}(X,Y)$ has the units of XY.
- If X and Y are independent $\operatorname{Cov}(X,Y)=0$

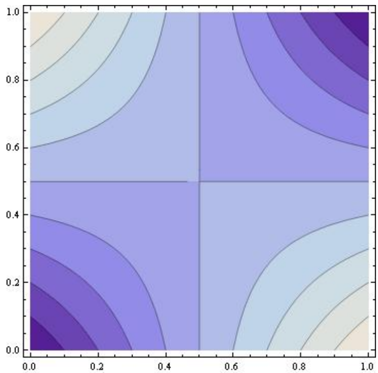

bivariate normal density
conditional distribution of Y when X = x
Wolfram Demonstrations Project
demonstrations.wolfram.com


(a)
(b)
(c)
(d)

Example 110

$$
\begin{aligned}
f(x,y) &= 2x+2y-4xy,\ 0\le x,y\le 1,\\
f_X(x) &= \int_0^1 (2x+2y-4xy)\,dy\\
&= 1,\ 0\le x\le 1\\
f_Y(y) &= 1,\ 0\le y\le 1\\
\mu_X &= \mu_Y = \frac{1}{2}\\
\operatorname{Cov}(X,Y) &= E(XY)-\mu_X\mu_Y\\
&= \int_0^1\int_0^1 xy(2x+2y-4xy)\,dx\,dy-\frac{1}{4}\\
&= \frac{2}{9}-\frac{1}{4}=\frac{-1}{36}
\end{aligned}
$$





Lemma 111 *Suppose that* $U = a_0 + a_1 X$, *and* $V = b_0 + b_1 Y$, *then*

$$\mathrm{Cov}(U, V) = a_1 b_1 \mathrm{Cov}(X, Y)$$

Proof.

$$\begin{aligned}
\mu_U &= a_0 + a_1 \mu_X \\
\mu_V &= b_0 + b_1 \mu_Y \\
\mathrm{Cov}(U, V) &= E\left[\left(U - \mu_U\right)\left(V - \mu_V\right)\right] \\
&= E\left[a_1\left(X - \mu_X\right) b_1 \left(Y - \mu_Y\right)\right] \\
&= a_1 b_1 E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] \\
&= a_1 b_1 \mathrm{Cov}(X, Y)
\end{aligned}$$

■

Intuition: Scaling is reflected, since covariance is unit dependent.

Theorem 112 *Suppose that:*
$U = a_0 + \sum_{i=1}^{n} a_i X_i$, *and* $V = b_0 + \sum_{j=1}^{m} b_j Y_j$, *then*

$$\mathrm{Cov}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{Cov}\left(X_i, Y_j\right)$$

Proof.

$$\begin{aligned}
\mu_U &= a_0 + \sum_{i=1}^{n} a_i \mu_{X_i} \\
\mu_V &= b_0 + \sum_{j=1}^{m} b_j \mu_{Y_j} \\
\mathrm{Cov}(U, V) &= E\left[\left(U - \mu_U\right)\left(V - \mu_V\right)\right] \\
&= E\left[\left(\sum_i a_i \left(X_i - \mu_{X_i}\right)\right)\left(\sum_j b_j \left(Y_j - \mu_{Y_j}\right)\right)\right] \\
&= E\left[\sum_i \sum_j a_i b_j \left(X_i - \mu_{X_i}\right)\left(Y_j - \mu_{Y_j}\right)\right] \\
&= \sum_i \sum_j a_i b_j E\left[\left(X_i - \mu_{X_i}\right)\left(Y_j - \mu_{Y_j}\right)\right] \\
&= \sum_i \sum_j a_i b_j \mathrm{Cov}\left(X_i, Y_j\right)
\end{aligned}$$

■

Corollary 113 *Consider* $U = a_0 + \sum_{i=1}^{n} a_i X_i$

1. *In general:*

$$\begin{aligned}
\mathrm{Var}(U) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}\left(X_i, X_j\right) \\
&= \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \,\mathrm{Cov}\left(X_i, X_j\right) \\
&= \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + 2\sum_{i > j} a_i a_j \,\mathrm{Cov}\left(X_i, X_j\right)
\end{aligned}$$

2. *If* X_i*s are uncorrelated (or independent):*

$$\mathrm{Var}(U) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i)$$

3. *If* X_i*s are i.i.d and* $a_i = 1$*:*

$$\mathrm{Var}(U) = n\sigma^2.$$

Proof. is immediate by noticing that $\mathrm{Cov}(U, U) = \mathrm{Var}(U)$. ■

Example 114 (Variance of Binomial) :

$$\begin{aligned}\operatorname{Var}(X) &= E\left[X^2\right] - (E[X])^2\\ &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - (np)^2.\end{aligned}$$

However; it is much easier to notice that

$$\begin{aligned}X &= \sum_{i=1}^{n} I_i,\\ I_i &\sim i.i.d\ Bernoulli(p)\\ \operatorname{Var}(X) &= n\operatorname{Var}(I_i)\\ &= np(1-p).\end{aligned}$$

Definition 115 (Correlation Coefficient) *: If X ar* *Y are jointly distributed r.v.s with existing means and variances*

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Intuition:

1. This is dimensionless, hoping that it has a meaningful figure instead of Covariance

2. invariant under linear transformation:

   $U = a_0 + a_1 X,\ V = b_0 + b_1 Y,\ a_1, b_1 > 0.$

$$\begin{aligned}
\rho_{UV} &= \frac{\mathrm{Cov}(a_0 + a_1 X, b_0 + b_1 Y)}{\sqrt{\mathrm{Var}(a_0 + a_1 X)\,\mathrm{Var}(b_0 + b_1 Y)}} \\
&= \frac{a_1 b_1 \mathrm{Cov}(X,Y)}{\sqrt{a_1^2 \mathrm{Var}(X)\, b_1^2 \mathrm{Var}(Y)}} \\
&= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \rho_{XY}
\end{aligned}$$

Theorem 116 $-1 \le \rho \le 1$. *Furthermore, $\rho = \pm 1$ **iff:** $P(Y = a + bX) = 1$ for some a, b.*

Proof.

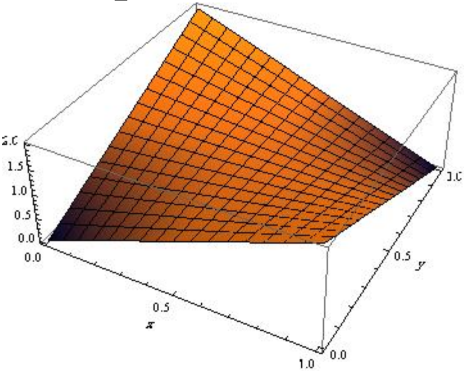
$$\begin{aligned}
0 &\le \operatorname{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) \\
&= \frac{\operatorname{Var}(X)}{\sigma_X^2} + \frac{\operatorname{Var}(Y)}{\sigma_Y^2} \pm 2\frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&= 2(1 \pm \rho) \\
-1 &\le \rho \le 1.
\end{aligned}$$

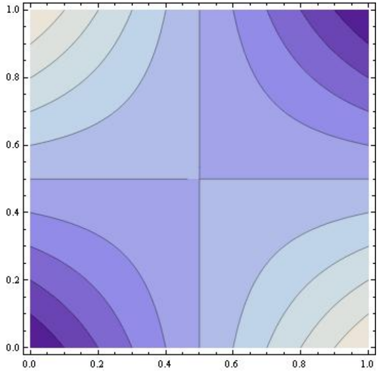
$$\operatorname{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 2(1+\rho)$$

$$\begin{aligned}
\rho = -1 &\iff \operatorname{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 0 \\
&\iff P\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c\right) = 1
\end{aligned}$$

First direction by Schwartz, second by Bernoulli variance. Similarly for $\rho = 1$. ■

Example 117 (Revisit Ex. 110) :





$$
\begin{aligned}
\rho &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \\
&= \frac{-1/36}{\sqrt{12}\sqrt{12}} \\
&= \frac{-1}{3}.
\end{aligned}
$$

X, Y	$\rho = 0$	$\rho \neq 0$
Dep.	T	T
Ind.	T	F

$$Independence \rightarrow \mathrm{Cov} = 0$$
$$\mathrm{Cov} \neq 0 \rightarrow Dependence.$$

Correlation is a measure of a linear relationship

Example 118 (Counter Example) $: X, Y$ *indep.,*

- $X \sim Uniform(-1,1)$; $f_X(x) = 1/2$.
- $Y \sim Uniform(0, 1/10)$; $f_Y(y) = 10$.
- $Z = X^2 + Y$: *what is* f_{XZ} *and* $\mathrm{Cov}(X, Z)$?

$$Z = X^2 + Y,$$

$$U = X,$$

$$\mathbf{J} = \begin{vmatrix} \frac{\partial Z}{\partial X} & \frac{\partial Z}{\partial Y} \\ \frac{\partial U}{\partial X} & \frac{\partial U}{\partial Y} \end{vmatrix} = \begin{vmatrix} 2X & 1 \\ 1 & 0 \end{vmatrix} = 1$$

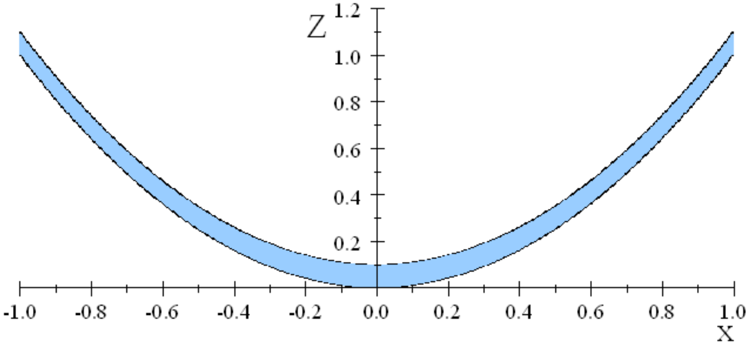
$$\begin{aligned} f_{UZ}(u,z) &= f_X(x)\, f_Y(y)\, \mathbf{J}^{-1}(x,y) \\ &= \left(\frac{1}{2}\right)(10)(1) \\ &= 5,\ -1 \le X \le 1,\ X^2 \le Z \le X^2 + .1 \end{aligned}$$

Much easier to say that:

$$f_{Z|X}(z|x) = 10,\ x^2 \le z \le x^2 + .1$$

$$f_{XZ}(x,z) = f_{Z|X}(z|x)\, f_X(x) = 10 \times \frac{1}{2} = 5.$$

$$f_{XZ}(x,z)=5,-1\le x\le 1,x^2\le z\le x^2+0.1$$



$$\begin{aligned}
\mathrm{Cov}(X,Z) &= E[XZ]-E[X]E[Z]\\
&= E\left[X\left(X^2+Y\right)\right]-0E[Z]\\
&= E\left[X^3\right]+E[XY]\\
&= 0+E[X]E[Y]\\
&= 0.
\end{aligned}$$

Intuition: There is dependency, yet **not linear**.

Observed Correlation Does Not Necessarily Imply Causation

May be one **or combination** of the following:

- **Example for** A **causes** B:

  Many

- **Example for** B **causes** A**:**

  Observation:

  (A): the more firemen fighting a fire

  (B): the bigger the fire is observed to be.

- **Example for** (C) **causes both:**

  Observation (Quinn et. al., 1999, Nature):

  (A): young children sleeping with the light

  (B): more likely to develop myopia

  Later study found that:

  infants sleeping with the light on caused the development of myopia!!

However, they found that:

parental myopia (C) is correlated with child myopia (B)

myopic parents (C) were more likely to leave a light on (A) in their children's bedroom.

4.4 Conditional Expectation and Prediction

4.4.1 Definitions and Examples

Definition 119 *The conditional expectation is defined as*

$$E(Y|X=x) = \sum_y y\, p_Y\left(y|x\right), \qquad \text{(Disc.)}$$

$$E(Y|X=x) = \int y\, f_{Y|X}\left(y|x\right) dy \qquad \text{(Cont.)}$$

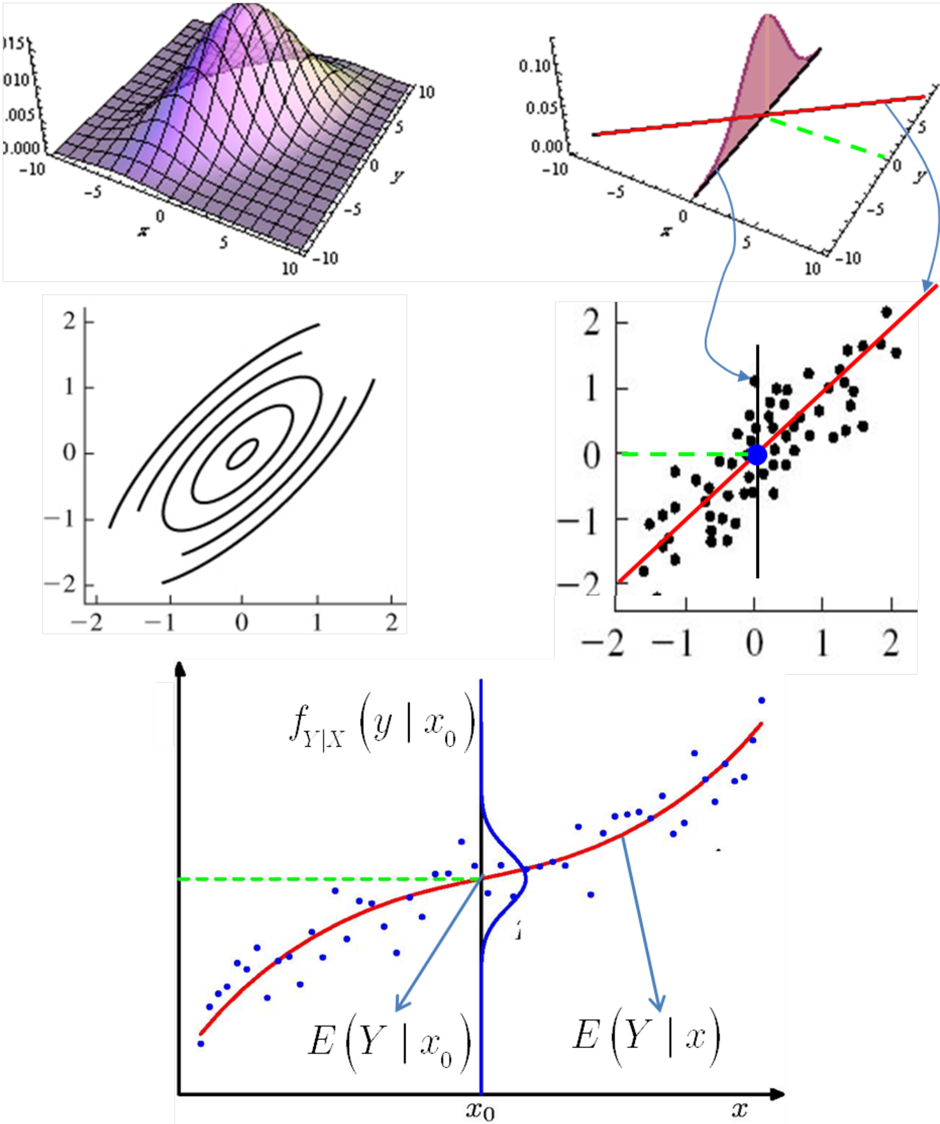
In general:

$$E(h(Y)|X=x) = \sum_y h\left(y\right) p_Y\left(y|x\right) \qquad \text{(Disc.)}$$

$$E(h(Y)|X=x) = \int h\left(y\right) f_{Y|X}\left(y|x\right) dy \qquad \text{(Cont.)}$$

Intuition:

For joint distribution $f_{XY}\left(x,y\right)$, at each x there is a conditional distribution $f_{Y|X}\left(y|x\right)$; e.g., **Age-Salary trend**.



This is called “Regression Function”

Example 120 $N \sim Poisson(\lambda)$

- N *is #obs. in* $[0,1]$
- X *is #obs. in* $\left[0,p\right]$
- *What is* $E\left[X|N=n\right]$

$$
\begin{aligned}
X &= \#obs.\ in\ \left[0,p\right]\ \sim Poisson\left(p\lambda\right)\\
Y &= \#obs.\ in\ \left[p,1\right]\ \sim Poisson\left(\left(1-p\right)\lambda\right)\\
N &= X+Y,\ X' = X\\
X &= X', \qquad Y = N-X.
\end{aligned}
$$

$$
\begin{aligned}
P_{X'N}(x,n) &= P_{XY}(x,n-x)\\
&= P_X(x)\,P_Y(n-x) \qquad\qquad \text{(no Jacobian)}\\
&= \frac{\left(p\lambda\right)^x e^{-p\lambda}}{x!}\frac{\left(\left(1-p\right)\lambda\right)^{n-x}e^{-(1-p)\lambda}}{(n-x)!}\\
&= \lambda^n e^{-\lambda}\frac{p^x\left(1-p\right)^{n-x}}{x!\,(n-x)!},
\end{aligned}
$$

which is nothing but $P_{X'|N}(x|n)$ scaled.

$$
\begin{aligned}
P_{X'|N}(x|n) &= \frac{P_{X'N}(x,n)}{P_N(n)} \\
&= \frac{\lambda^n e^{-\lambda} \frac{p^x (1-p)^{n-x}}{x!(n-x)!}}{\frac{\lambda^n e^{-\lambda}}{n!}} \\
&= \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!} \\
&= \binom{n}{x} p^x (1-p)^{n-x} \\
&\sim Binomial(n,p), \\
E[X|N=n] &= np,
\end{aligned}
$$

which is a discrete line at $n = 1, \ldots, n$. This makes sense because n is fixed.

Let's open a Mathematica notebook and understand it visually.

Theorem 121 (Law of Total Expectation:) $E(Y) = E[E(Y|X)]$. *We can also write it as:*

$$E(Y) = E_X\left[E_{Y|X}(Y|X)\right].$$

Disc. case:.

$$\begin{aligned}
E_{Y|X}(Y|X) &= \sum_y y\, p_{Y|X}\left(y|x\right)\\
E_X\left[E_{Y|X}(Y|X)\right] &= \sum_x E_{Y|X}(Y|X)\, p_X(x)\\
&= \sum_x \left(\sum_y y\, p_{Y|X}\left(y|x\right)\right) p_X(x)\\
&= \sum_y y \sum_x p_{Y|X}\left(y|x\right) p_X(x)\\
&= \sum_y y\, p_Y\left(y\right)\\
&= E(Y).
\end{aligned}$$

Proof for cont. case is very similar. ■

Intuition: (See Cond. Exp. figure)

Example 122 *Consider a unit and its backup*

- *Each has mean life time of μ*
- *a backup unit may not launch with probability p*
- *What is the mean life time of the system?*

Consider the launching Bernouli variable $I \sim Ber$

$$T = \begin{cases} T_1 & I = 1 \\ T_1 + T_2 & I = 0 \end{cases}$$

$$\begin{aligned} E(T|I=1) &= \mu, \\ E(T|I=0) &= 2\mu, \\ E(T) &= \mu p + 2\mu(1-p) \\ &= \mu(2-p). \end{aligned}$$

More generally

$$\begin{aligned} f_T(t) &= f_{T|I}(t|1)\, p + f_{T|I}(t|0)\,(1-p) \\ &= f_{T_1}(t)\, p + f_{T_1+T_2}(1-p). \end{aligned}$$

This is called mixture of r.v.

Theorem 123 (Variance Decomposition) :

$$\mathrm{Var}[Y] = \underset{X}{\mathrm{Var}}\left[E_{Y|X}[Y|X]\right] + E_X\left[\underset{Y|X}{\mathrm{Var}}[Y|X]\right].$$

Proof.

$$\begin{aligned}
&\underset{X}{\mathrm{Var}}\left[E_{Y|X}[Y|X]\right]\\
&= E_X\left[\left(E_{Y|X}[Y|X]\right)^2\right] - \left(E_X E_{Y|X}[Y|X]\right)^2\\
&= \underline{E_X\left[\left(E_{Y|X}[Y|X]\right)^2\right]} - \left(E[Y]\right)^2\\
&E_X\left[\underset{Y|X}{\mathrm{Var}}[Y|X]\right]\\
&= E_X\left[E_{Y|X}[Y^2|X] - \left(E_{Y|X}[Y|X]\right)^2\right]\\
&= E_X E_{Y|X}[Y^2|X] - E_X\left[\left(E_{Y|X}[Y|X]\right)^2\right]\\
&= E[Y^2] - \underline{E_X\left[\left(E_{Y|X}[Y|X]\right)^2\right]}.\\
sum &= E[Y^2] - \left(E[Y]\right)^2\\
&= \mathrm{Var}[Y].
\end{aligned}$$

■

Intuition: (See figure)

Example 124 (Random Sums) *: X_is are i.i.d.*

$$T = \sum_{i=1}^{N} X_i$$
$$T|(N = n) = \sum_{i=1}^{n} X_i$$
$$E[T|N = n] = nE[X]$$
$$E[T] = E_N[NE[X]]$$
$$= \mu_N \mu_X,$$

which makes sense.

$$\mathrm{Var}[T] = E_N\left[\underset{T|N}{\mathrm{Var}}[T|N]\right] + \underset{N}{\mathrm{Var}}\left[E_{T|N}[T|N]\right]$$
$$= E_N\left[n\sigma_X^2\right] + \underset{N}{\mathrm{Var}}\left[n\mu_X\right]$$
$$= \mu_N\sigma_X^2 + \mu_X^2\sigma_N^2.$$

A special case would be if N is constant n

$$\mathrm{Var}[T] = n\sigma_X^2,$$

as usual.

Typical Values: insurance company with number of claims $N \sim Poisson(\lambda = 900)$, claim value X_i

$$\begin{aligned}
\mu_N &= 900\\
\sigma_N &= 30\\
\mu_X &= 1000\$\\
\sigma_X &= 500\$\\
E[T] &= \mu_N \mu_X = 900,000\$\\
\text{Var}[T] &= \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2\\
&= 900 \times 500^2 + 1000^2 \times 30^2\\
&= 225M + 900M\\
&= 1125M\\
\sigma_T &= 33,541\$,
\end{aligned}$$

Therefore they should plan on

$$\mu_T = 900,000\$ \pm 33,541\$.$$

If N is fixed, then

$$\begin{aligned}
\sigma_T^2 &= 900 \times 500^2,\\
\sigma_T &= 15,000\$ << 33,541\$
\end{aligned}$$

4.4.2 Prediction

Let's predict a r.v. by some constant

$$
\begin{aligned}
MSE &= E(Y-c)^2 \\
&= E((Y-E(Y))+(E(Y)-c))^2 \\
&= E(Y-E(Y))^2+(E(Y)-c)^2 \\
&+2E[(Y-E(Y))(E(Y)-c)] \\
&= \underbrace{\operatorname{Var}[Y]}_{\text{irreducible error}} + (E(Y)-c)^2 \\
c_{\min} &= \arg\min_c [MSE] \\
&= E(Y).
\end{aligned}
$$

If we replace E by $E_{Y|X}$

$$c = h(X) = E_{Y|X}(Y|X),$$

which minimizes

$$E_{Y|X}\left[(Y-h(X))^2|X\right],$$

and therefore minimizes

$$
\begin{aligned}
MSE &= E(Y-h(X))^2 \\
&= E_X E_{Y|X}\left[(Y-h(X))^2|X\right],
\end{aligned}
$$

which is the regression function. **This is what is Machine Learning is about!**

4.5 The Moment-Generating Function

Definition 125 (Moments) :

$$r^{th} moment = E\left[X^r\right],$$

$$1^{st} moment = mean, \qquad \text{(Location)}$$

$$r^{th} central\ moment = E\left[(X - E[X])^r\right],$$

$$1^{st}\ central\ moment = 0,$$

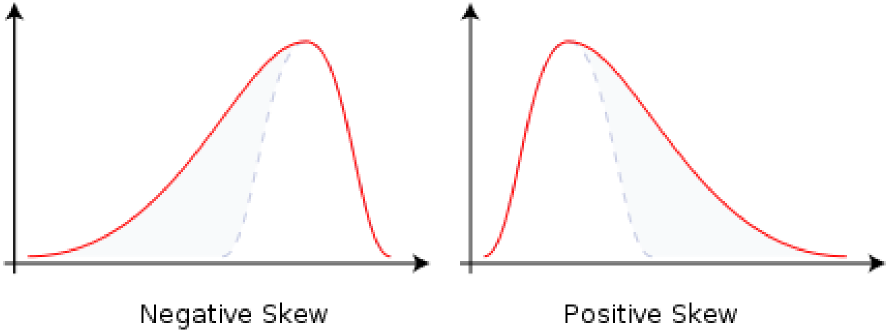
$$2^{nd} central\ moment = Variance, \qquad \text{(Dispersion)}$$

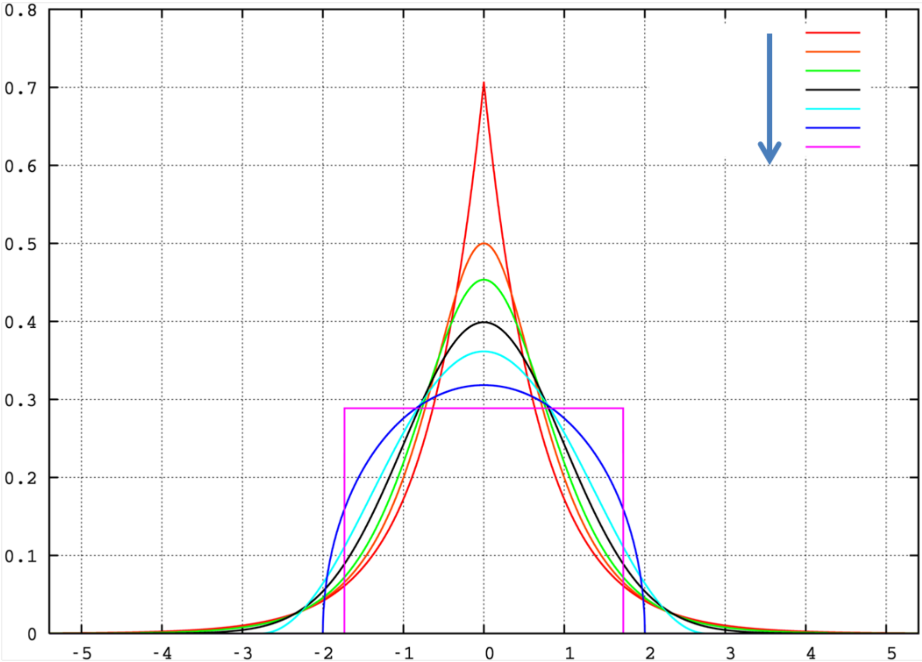
$$3^{rd} central\ moment = Skewness, \qquad \text{(Asymmetry)}$$

$$4^{th} central\ moment = Kurtosis \qquad \text{(Flatness)}$$

It is clear that if f_X is symmetric,

- its point of symmetry is $E[X]$
- $X - E[X]$ will be symmetric around 0.
- $E[(X - E[X])^r]$, for odd r, will be 0.
- r^{th} normalized moment $= \frac{E[(X-E[X])^r]}{\sigma^r}$


Negative Skew
Positive Skew


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Definition 126 *The Moment-Generating Function (mgf), $M_X(t)$, for a r.v. X is given by*

$$M_X(t) = E\left[e^{tX}\right]$$

$$\sum_x e^{tx} p_X(x), \qquad \text{(Disc.)}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \qquad \text{(Cont.)}$$

- *It may not exist, e.g., if $X \sim Cauchy$, since f_X fades out slowly.*
- *If exists it uniquely defines f_X*
- *The characteristic function always exists*

$$\begin{aligned}\phi_X(t) &= \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx \\ &= \mathcal{F}\left\{f_X\right\},\end{aligned}$$

  which is Fourier transform of f_X. This is because $\left|e^{itx}\right| \leq 1$.

- *Many nice properties for M and ϕ and connection to $\mathcal{L}$ and $\mathcal{F}$ Transforms in "Signals and Systems" course.*

Proposition 127 *If the mgf exists in an open interval containing zero then*

$$M^{(r)}(0) = E\left[X^r\right].$$

Proof.

$$\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \\
M'(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f(x)\,dx \\
&= \int_{-\infty}^{\infty} x e^{tx} f(x)\,dx \\
M''(t) &= \int_{-\infty}^{\infty} x^2 e^{tx} f(x)\,dx, \\
M^{(r)}(t) &= \int_{-\infty}^{\infty} x^r e^{tx} f(x)\,dx \\
M^{(r)}(0) &= \int_{-\infty}^{\infty} x^r f(x)\,dx \\
&= E\left[X^r\right].
\end{aligned}$$

■

Therefore:

$$\text{pdf} \Longleftrightarrow \text{mgf} \overset{\text{Taylor}}{\Longleftrightarrow} \text{mgf derv.} = \text{pdf mts.}$$

$$M(t) = M(0) + \sum_r \frac{1}{r!} t^r M^{(r)}(0) \qquad \text{(Taylor Series)}$$

Example 128 (Poisson) :

$$
\begin{aligned}
M(t) &= \sum_{0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{0}^{\infty} \frac{\left(\lambda e^t\right)^k}{k!} \\
&= e^{-\lambda} e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)}. \\
M'(t) &= \lambda e^t e^{\lambda\left(e^t - 1\right)} \\
E[X] &= M'(0) = \lambda \\
M''(t) &= \lambda e^t e^{\lambda\left(e^t - 1\right)} + \lambda^2 e^{2t} e^{\lambda\left(e^t - 1\right)} \\
E\left[X^2\right] &= M''(0) = \lambda + \lambda^2 \\
\mathrm{Var}[X] &= E\left[X^2\right] - (E[X])^2 = \lambda.
\end{aligned}
$$

Example 129 (Standard Normal) :

$$
\begin{aligned}
M(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t^2/2-(x-t)^2/2} dx \quad \text{(Comp. Sq.)} \\
&= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\
&= e^{t^2/2}, \\
M'(t) &= te^{t^2/2} \\
E[X] &= M'(0) = 0 \\
M''(t) &= e^{t^2/2} + t^2 e^{t^2/2} \\
E\left[X^2\right] &= M''(0) = 1, \\
\text{Var}[X] &= 1.
\end{aligned}
$$

Lemma 130 *If $Y = a + bX$ then*

$$M_Y(t) = e^{at} M_X(bt).$$

Proof.

$$\begin{aligned} M_Y(t) &= E\left[e^{tY}\right] \\ &= E\left[e^{at+btX}\right] \\ &= e^{at} E\left[e^{btX}\right] \\ &= e^{at} M_X(bt). \end{aligned}$$

■

Example 131 *If $Y \sim N\left(\mu, \sigma^2\right)$, then*

$$Y = \mu + \sigma Z$$

$$M_Y(t) = e^{\mu t} M_Z(\sigma t),$$

$$M_Z(t) = e^{t^2/2}$$

$$M_Y(t) = e^{\mu t} e^{\sigma^2 t^2/2}.$$

Lemma 132 *if X and Y are independent r.v. and $Z = X + Y$, then*

$$M_Z(t) = M_X(t)\, M_Y(t).$$

Proof.

$$\begin{aligned}
M_Z(t) &= E\left[e^{tZ}\right] \\
&= E\left[e^{tX+tY}\right] \\
&= E\left[e^{tX}e^{tY}\right] \\
&= E\left[e^{tX}\right] E\left[e^{tY}\right] \\
&= M_X(t)\, M_Y(t).
\end{aligned}$$

■

Example 133 (Sum of Poissons λ_1, λ_2)

$$\begin{aligned}
M_Z(t) &= M_X(t)\, M_Y(t) \\
&= e^{\lambda_1\left(e^t-1\right)} e^{\lambda_2\left(e^t-1\right)} \\
&= e^{(\lambda_1+\lambda_2)e^t-(\lambda_1+\lambda_2)} \\
&= e^{(\lambda_1+\lambda_2)\left(e^t-1\right)}
\end{aligned}$$

$$Z \sim Poisson(\lambda_1 + \lambda_2).$$

4.6 Approximate Methods

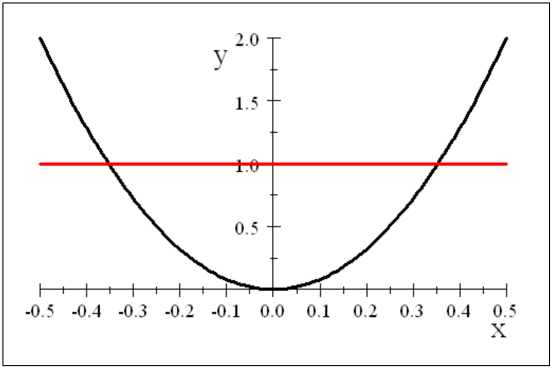
- X is r.v., and $Y = g(X)$
- We know μ_X and σ_X (or $\widehat{\mu}_X$ and $\widehat{\sigma}_X$).
- What is μ_Y and σ_Y? even approximately!

Delta Methods (Propagation Error):

$$
\begin{aligned}
Y &= g(X) \\
&= g(\mu_X) + (X - \mu_X) g'(\mu_X) + \\
&\frac{1}{2!}(X - \mu_X)^2 g''(\mu_X) + \cdots \qquad \text{(Taylor Series)} \\
&\approx g(\mu_X) + (X - \mu_X) g'(\mu_X) \qquad \text{(1st order aprox)} \\
\mu_Y &\approx g(\mu_X) \\
\sigma_Y^2 &\approx \sigma_X^2 \left(g'(\mu_X)\right)^2 \\
Y &\approx g(\mu_X) + (X - \mu_X) g'(\mu_X) \\
&+ \frac{1}{2!}(X - \mu_X)^2 g''(\mu_X) \qquad \text{(2nd order aprox)} \\
\mu_Y &= g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X).
\end{aligned}
$$

Example 134 (Very nonlinear) :

- $X \sim Uniform\left(-\frac{1}{2}, \frac{1}{2}\right)$.
- $Y = g(X) = 8X^2$



$$\mu_X = 0,$$

$$\sigma_X^2 = \frac{1}{12}.$$

$$\mu_Y = \int_{-1/2}^{1/2} 8x^2 dx = \frac{2}{3},$$

$$\sigma_Y^2 = E\left[Y^2\right] - \mu_Y^2$$

$$= \int_{-1/2}^{1/2} \left(8x^2\right)^2 dx - \left(\frac{2}{3}\right)^2 = \frac{16}{45}$$

However, by using 1st order approx.

$$\begin{aligned}\mu_Y &\approx g\left(\mu_X\right)\\ &= 8\mu_X^2 = 0,\\ \sigma_Y^2 &\approx \sigma_X^2\left(g'\left(\mu_X\right)\right)^2\\ &= \frac{1}{12}(0) = 0.\end{aligned}$$

By using 2nd order approx.

$$\begin{aligned}\mu_Y &\approx g\left(\mu_X\right) + \frac{1}{2}\sigma_X^2 g''\left(\mu_X\right) \qquad (4.1)\\ &= 0 + \frac{1}{2}\times\frac{1}{12}\times 16\\ &= \frac{2}{3}.\end{aligned}$$

This is exact because (4.1) is exact!

$$\begin{aligned}g(x) = g(x_0) &+ (x - x_0)\, g'(x_0) +\\ &\frac{1}{2!}(x - x_0)^2 g''(x_0) + \cdots\\ &= 8x_0^2 + (x - x_0)\, 16x_0 + \frac{1}{2}(x - x_0)^2\, 16\\ &= 8x^2\end{aligned}$$

$If\ X \sim Uniform\,(0.3, 0.5)\,, then$

$$\mu_X = .4,$$

$$\sigma_X^2 = \int_{.3}^{.5} x^2\,(5)\,dx - (.4)^2 = 3.3333 \times 10^{-3}$$

$$\mu_Y = \int_{.3}^{.5} \left(8x^2\right)(5)\,dx = 1.3067$$

$$\sigma_Y^2 = \int_{.3}^{.5} \left(8x^2\right)^2 (5)\,dx - (1.3067)^2 = 0.13702$$

By using 1st order approx.

$$\begin{aligned}\mu_Y &\approx g\left(\mu_X\right) \\ &= 8\,(.4)^2 = 1.28 \\ \sigma_Y^2 &\approx \sigma_X^2 \left(g'\left(\mu_X\right)\right)^2 \\ &= \frac{1}{300}\,(16 \times .4)^2 = 0.13653\end{aligned}$$

Chapter 5

Limit Theorems

We are getting into Statistics

Great concepts and intuition are here

5.1 The Law of Large Numbers

- It is always believed, **subjectively**, that tossing a fair coin will produce **ultimately** 0.5 heads proportion.

- Mathematician John Kerrich tried it in prison he got 5067 heads out of 10,000 tosses.

- LLN is a mathematical formulation of large sums.

- In particular, for coin tossing:

$$X_i \sim Bernoulli\,(0.5)\,, \qquad \text{(i.i.d)}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \text{ “approaches” } 0.5.$$

- What is “approaches” formally?

Definition 135 (Convergence in Probability) *: Le* $Z_1, \ldots, Z_n$ *be a sequence of r.v. (s.r.v.) We say that* Z_n *converges in probability to* α *if*

$$\lim_{n\to\infty} P\left(|Z_n - \alpha| > \varepsilon\right) = 0 \ \forall \varepsilon > 0.$$

This is written in different ways

$$Z_n \xrightarrow{p} \alpha,$$

$$\lim_{n\to\infty} P\left(|Z_n - \alpha| > \varepsilon\right) = 0 \ \forall \varepsilon > 0.$$

$$P\left(|Z_n - \alpha| > \varepsilon\right) \to 0 \ \forall \varepsilon > 0, \ \textit{as}\ n \to \infty.$$

Theorem 136 (Weak Law of Large Numbers) : *If $X_1, \ldots, X_n$ is a s.r.v., independent with* ***existing****, and common, μ and σ^2 (but not necessarily identical) then $\overline{X}_n (= S_n/n) \xrightarrow{p} \mu$.*

Proof.

$$\overline{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$E\left[\overline{X}_n\right] = \mu,$$

$$\mathrm{Var}\left[\overline{X}_n\right] = \sigma^2/n,$$

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \le \frac{\mathrm{Var}\left[\overline{X}_n\right]}{\varepsilon^2} \ \forall \varepsilon > 0$$

(Chebyshev's ineq.)

$$= \frac{\sigma^2}{n\varepsilon^2} \ \forall \varepsilon > 0,$$

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0 \ \forall \varepsilon > 0.$$

■

Example 137 (Repeated Measurements) *: A special case of the WLLN is when X_is are i.i.d.*

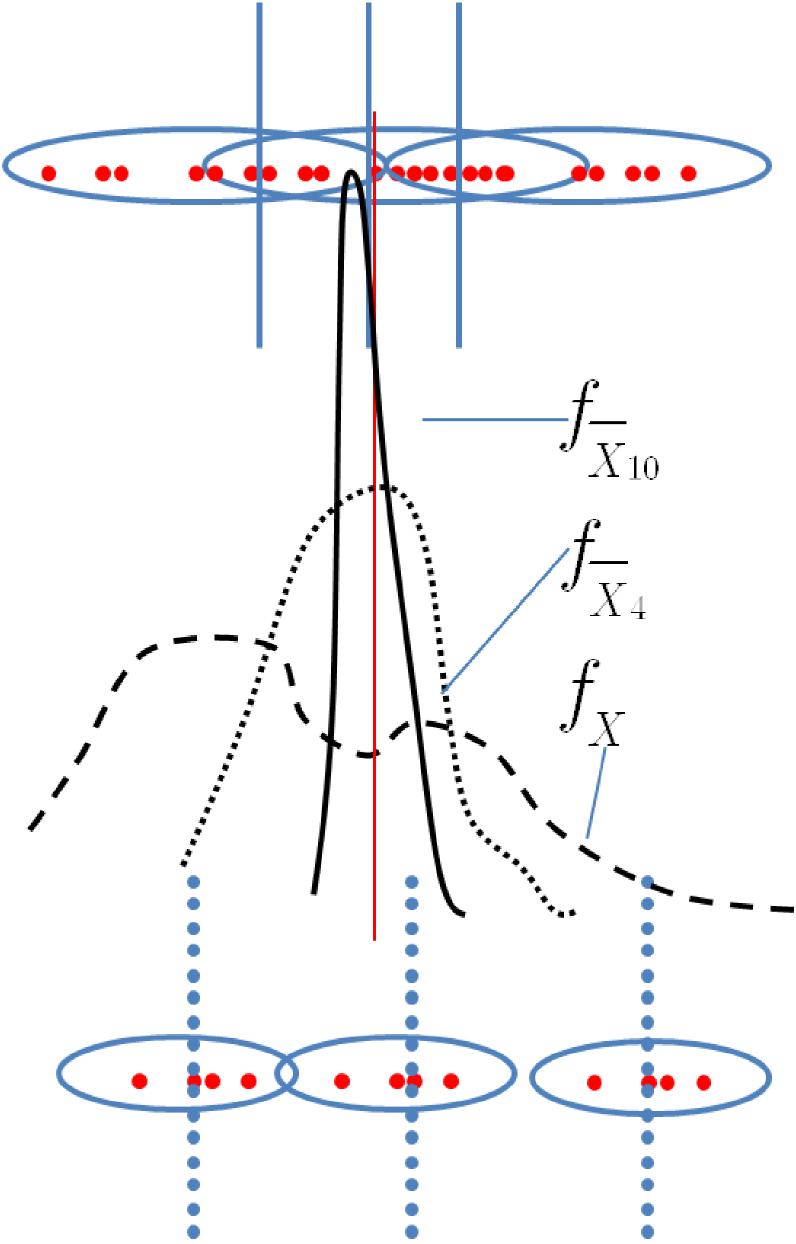
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For a particular sample $x_1, \ldots, x_n$, $\overline{X}_n$ becomes a number not a r.v.

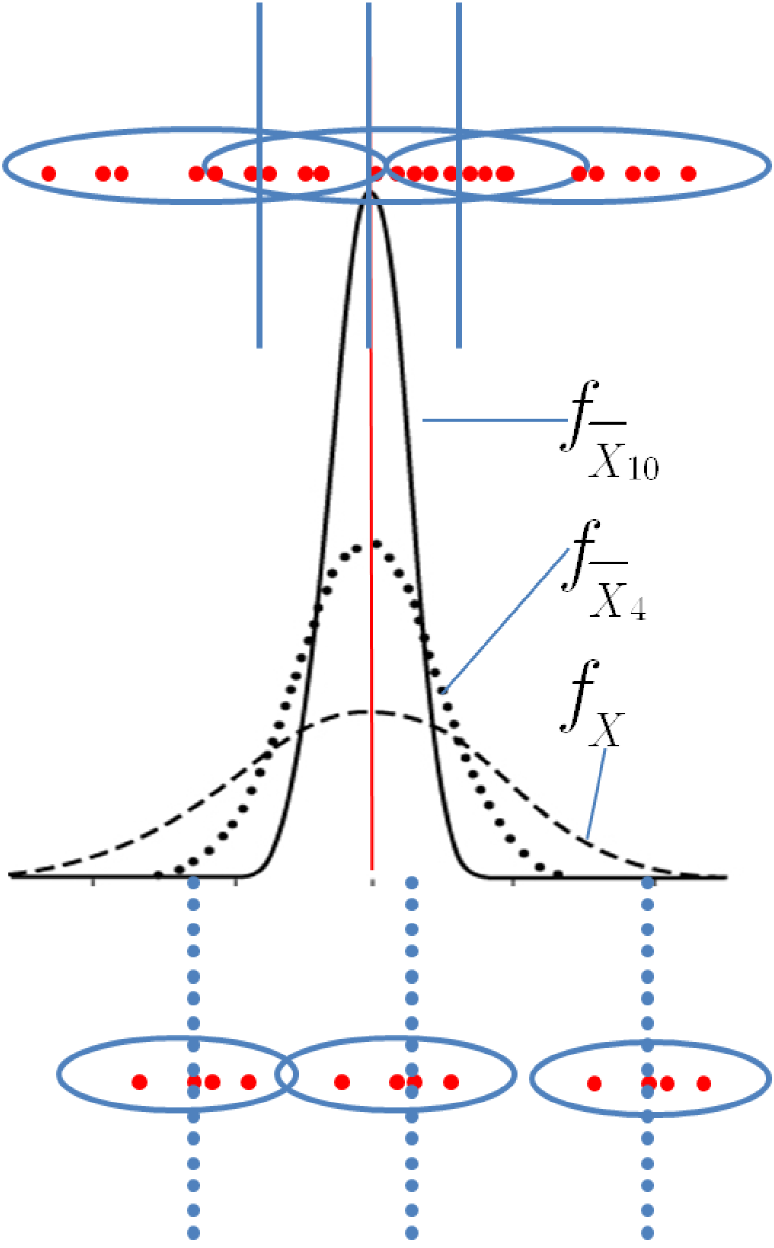
$$\overline{X}_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

$$\begin{aligned} & X_1, X_2, \ldots, X_n \\ X \xrightarrow{\underline{Sample_1}} & x_1, x_2, \ldots, x_n \\ X \xrightarrow{\underline{Sample_2}} & x_1, x_2, \ldots, x_n \\ & \vdots \end{aligned}$$

Let's see the meaning of the WLLN for $\overline{X}_n$:



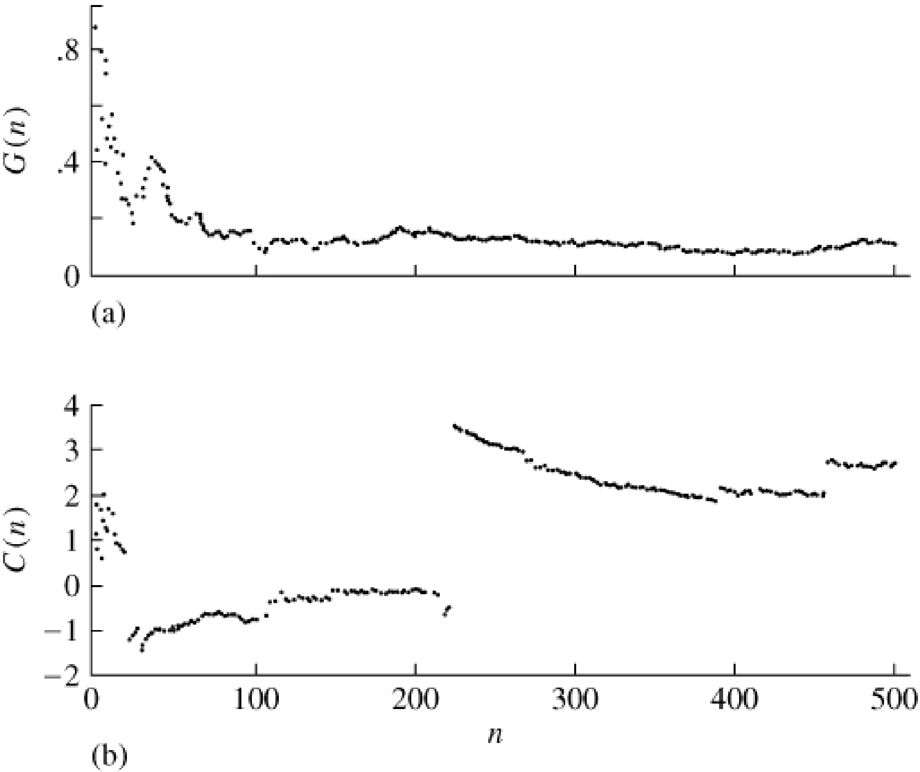
$f_{\overline{X}_{10}}$
$f_{\overline{X}_4}$
f_X



$f_{\overline{X}_{10}}$
$f_{\overline{X}_4}$
f_X

Example 138 (Normal vs. Cauchy) :

- $X \sim N(0,1)$, $Y \sim Cauchy$.
- *One sample (500 obs.) from each.*
- $G_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $C_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, $n = 1, \ldots, 500$.



Example 139 (Estimation of Moments) :

$$m_r = E\left[X^r\right].$$

Define

$$\widehat{m}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

Then

$$\begin{aligned} E\left[\widehat{m}_r\right] &= \frac{1}{n}\sum_{i=1}^{n} E\left[X^r\right] \\ &= m_r, \\ \widehat{m}_r &\xrightarrow{p} m_r. \end{aligned}$$

Example 140 (Monte Carlo Integration) *: How to calculate the integration*

$$I = \int_a^b g(x)\,dx.$$

Let $X \sim Uniform(a,b)$; *generate a sample* x_i, $i = 1,\ldots,n$. *Define*

$$\begin{aligned}
\widehat{I} &= \frac{1}{n}\sum_{i=1}^{n} g(X_i)\,,\\
E\left[\widehat{I}\right] &= E\left[g(X)\right]\\
&= \int_a^b g(x)\, f_X(x)\,dx\\
&= \int_a^b g(x)\,\frac{1}{b-a}dx\\
&= \frac{1}{b-a}I,\\
\widehat{I} &\xrightarrow{p} \frac{1}{b-a}I\\
I &\approx (b-a)\,\widehat{I}.
\end{aligned}$$

5.2 Convergence in Distribution and the Central Limit Theorem (CLT)

Definition 141 *Let $X_1, \ldots, X_n$ be s.r.v. with $F_1, \ldots, F$ and X is another r.v. with F. We say that X_n converges in distribution to X if*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

except at discontinuities. This can be written as

$$X_n \xrightarrow{d} X,$$

$$F_n(x) \to F(x), \text{ as } n \to \infty,$$

$$\lim_{n\to\infty} F_n(x) = F(x)$$

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x)$$

Theorem 142 *For the setting above, if $M_n(t) \to M(t)$ then $F_n(t) \to F(t)$ at all points of continuity.*

Proof. Omitted. ■

Theorem 143 (CLT) *: Let $X_1, \ldots, X_n$ be i.i.d, with common μ, σ^2, F, M. Then, the standardized version*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{S_n - n\mu}{\sigma \sqrt{n}} = \frac{\sum_i \left(X_i - \mu\right) / \sigma}{\sqrt{n}}$$

converges in distribution to a Standard Normal $N(0,1)$; i.e., $Z_n \xrightarrow{d} Z$.

Before rigorous proof, notice:

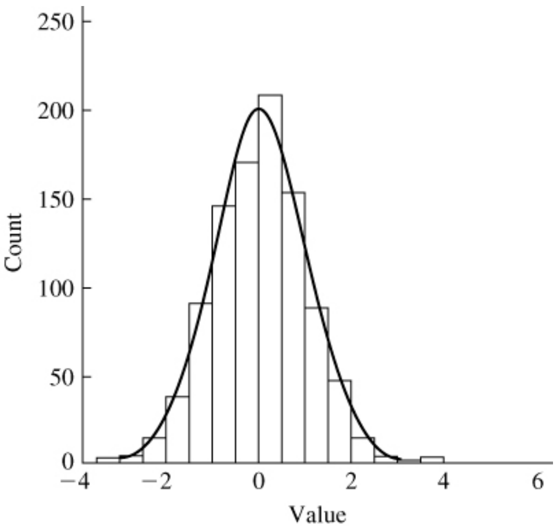
- This is regardeless to F !!!
- If WLLN shows that $\overline{X}_n$ **goes** to μ (in probability), CLT shows how it **fluctuates** around μ (i.e., distribution and rate)
- More precise than Chebyshev (Ex. 105, page 199).
- Several other versions of CLT

Example 144 $X \sim Uniform\left(-\sqrt{3}, \sqrt{3}\right)$; *then* $\mu = 0, \sigma = 1$.

$$\begin{aligned} Z_n &= \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \\ &= S_n/\sqrt{n}. \end{aligned}$$

Obtain 1000 samples, each with 12 obs. (i.e., $n = 12$*). We have 1000 values for* Z_{12}*. Notice that:*

$$Z_{n_{\max}} = \frac{nX_{\max}}{\sqrt{n}} = \sqrt{n}X_{\max} = \sqrt{12}\sqrt{3} = 6.$$



Example 145 (Measurement Error) :

$$\begin{aligned}
P\left(\left|\overline{X}_n-\mu\right|<c\right) &= P\left(-c<\overline{X}_n-\mu<c\right)\\
&= P\left(\frac{-c}{\sigma/\sqrt{n}}<\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}<\frac{c}{\sigma/\sqrt{n}}\right)\\
&\approx \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

Example 146 $X \sim Binomial\left(n,p\right)$, $n=100$, $p=0.5$. *E.g.,*

$$P\left(X \geq 60\right) = \sum_{k=60}^{100} \binom{n}{k} p^k \left(1-p\right)^{100-k}$$

is computationally expensive. However,

$$X = \sum_{i=1}^{n} I_i \qquad \left(I_i \sim Bernouli\left(p\right)\right)$$

$$\frac{X-np}{\sqrt{p\left(1-p\right)\sqrt{n}}} \xrightarrow{d} N\left(0,1\right),$$

Then,

$$\begin{aligned} P\left(X \geq 60\right) &= P\left(\frac{X-np}{\sqrt{p\left(1-p\right)\sqrt{n}}} \geq \frac{60-100/2}{\sqrt{\frac{1}{2}\frac{1}{2}\sqrt{100}}}\right) \\ &\approx P\left(Z \geq 2\right) \\ &= 1-\Phi\left(2\right) = 0.0228. \end{aligned}$$

Proof of CLT.

$$
\begin{aligned}
Z_n &= \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\\
&= \frac{\sum_i \left(X_i - \mu\right)/\sigma}{\sqrt{n}}\\
&= \frac{\sum_i Y_i}{\sqrt{n}}
\end{aligned}
$$

$$
\begin{aligned}
M_{\sum_i Y_i}(t) &= \left(M_Y(t)\right)^n\\
M_{Z_n}(t) &= M_{\sum_i Y_i}\left(t/\sqrt{n}\right)\\
&= \left(M_Y\left(t/\sqrt{n}\right)\right)^n
\end{aligned}
$$

$$
\begin{aligned}
M_Y\left(\frac{t}{\sqrt{n}}\right) &= \underbrace{M_Y(0)}_{E\left[Y^0\right]=1} + \left(\frac{t}{\sqrt{n}}\right)\underbrace{M_Y'(0)}_{E[Y]=0} +\\
&\quad \frac{1}{2!}\left(\frac{t}{\sqrt{n}}\right)^2 \underbrace{M_Y''(0)}_{E\left[Y^2\right]=1} + \sum_{k=3}^{\infty} \frac{M_Y^{(k)}(0)}{k!}\left(t/\sqrt{n}\right)^k\\
&= 1 + \left(\frac{t}{\sqrt{n}}\right)^2\left[\frac{1}{2} + \sum_{k=3}^{\infty}\frac{M_Y^{(k)}(0)}{k!}\left(\frac{t}{\sqrt{n}}\right)^{k-2}\right]\\
&= 1 + \frac{1}{n}t^2\left(\frac{1}{2} + \underbrace{r_n}_{\to 0 \text{ as } n\to\infty}\right).
\end{aligned}
$$

Therefore,

$$M_{Z_n}(t) = \left(1 + \frac{1}{n} t^2 \left(\frac{1}{2} + r_n\right)\right)^n,$$

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}.$$

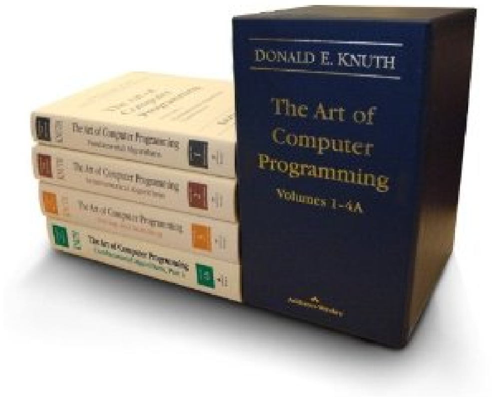
■

Appendix A

Simulation

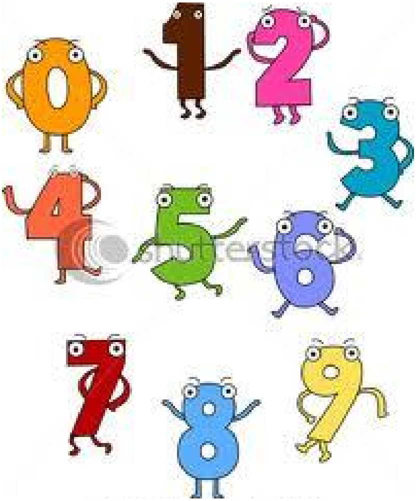
Very Nice Practical Chapter

This Chapter follows Knuth (1997, Vol. 2, Ch. 3)



and DeGroot and Schervish (2002, Ch. 11)

A.1 Generating r.v. by Simulation



Starting from 20's people thought of generating random numbers from ready made table.

John von Neumann in 1946 (although it is very deterministic but looks good scrambling that carries no physical significance):

1. start with a number, e.g., 5772156649
2. square it: 33317792380594909201
3. take the middle: 7923805949

Such methods are called *pseudorandom* or *quasirandom*.

Current methods are of two steps:

First: generate $Uniform(0,1)$ random number using Linear Congruential method

$$\begin{array}{lll} m, & \text{the modulus;} & 0 < m, \\ a, & \text{the multiplier;} & 0 \le a < m, \\ c, & \text{the increment;} & 0 \le c < m, \\ X_0 & \text{the seed;} & 0 \le X_0 < m. \end{array},$$

$$X_{n+1} = (aX_n + c) \bmod m, \; n \ge 0.$$

E.g.,

$$m = 10, X_0 = a = c = 7$$

$$X_n = 7,\ 6,\ 9,\ 0,\ 7,\ 6,\ 9,\ 0, \ldots$$

The choice of the parameters is a matter of research. Typical values are

$$a = 7^5,$$

$$c = 0,$$

$$m = 2^{31} - 1.$$

Second: use the generated uniform r.v. to convert it to any other r.v. by one of the following two methods

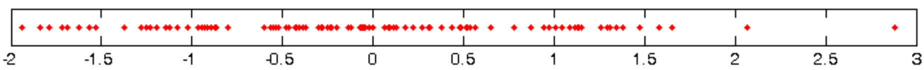
- Transformation methods (end of Ch. 2), if the cdf is known:

$$U \sim Uniform(0,1),$$
$$X = F^{-1}(U) \Longrightarrow F_X = F.$$

- Rejection method (Sec. 3.5), if the cdf cannot be found in closed form (only pdf is known)

Matlab Code A.1:

```
n=100; mu=0; sigma=1;
x = random('normal', mu, sigma, [n,1])
plot(x, zeros([length(x),1]), '.r');
```



A.2 Histograms

Let's define first the indicator function

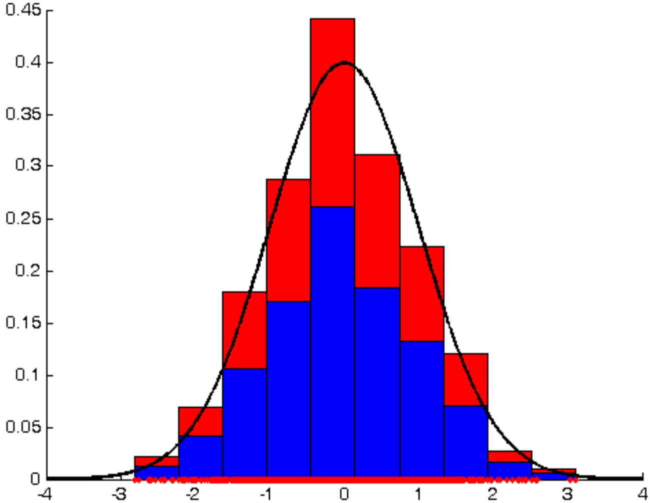
$$I_{(c)} = \begin{cases} 1 & \text{if } c \text{ is } T \\ 0 & \text{if } c \text{ is } F \end{cases},$$

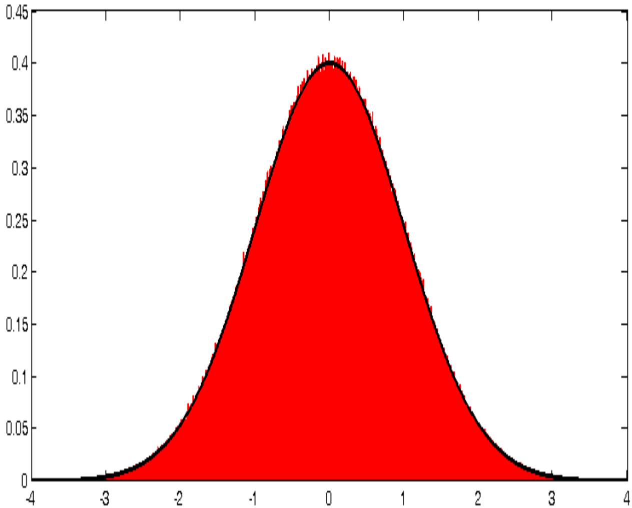
$$I_{(c)} \sim Bernoulli\left(\Pr(c)\right).$$

For data $x_1, \ldots, x_n$ divide the data range T to K equal regions of equal width Δ (so that $K = T/\Delta$)

$$\begin{aligned} T_k &= [t_0 + \Delta k, t_0 + \Delta(k+1)[\\ &= [t_k, t_{k+1}[, \; k = 0, \ldots, K-1, \end{aligned}$$

Notice: decreasing Δ increases K.


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4

We have three versions of histogram:

$$N_k = \sum_{i=1}^{n} I_{(X_i \in T_k)}, \qquad \text{(counts)}$$

$$R_k = \frac{N_k}{n}, \qquad \text{(relative counts)}$$

$$f_k = \frac{N_k}{\Delta n} \qquad \text{(normalized)}$$

$$\begin{aligned} Area(under\ N_k) &= \sum_k \Delta N_k \\ &= \Delta \sum_k N_k = \Delta n. \end{aligned}$$

$$\begin{aligned} R_k &\xrightarrow{p} \Pr(X \in T_k) \\ f_k &= \frac{\Pr(X \in T_k)}{\Delta} && \text{(for large } n\text{)} \\ &\approx \frac{f_X(t_k)\,\Delta}{\Delta} = f_X(t_k) && \text{(for small } \Delta\text{)} \end{aligned}$$

Matlab Code A.2:

```
n=1000; mu=0; sigma=1;
x = random('normal', mu, sigma, [n,1])

figure; hold on;

[N, xout]=hist(x);
bar(xout', N'/(n*(xout(2)-xout(1))), '
   barwidth', 1, 'facecolor', 'r');
bar(xout', N'/n, 'barwidth', 1, '
   facecolor', 'b');

z=-4:.01:4;
y=1/(sqrt(2*pi*sigma)) *exp(-(z-mu).^2
    / (2*(sigma^2)));
plot(z, y, 'k', 'LineWidth', 2);

plot(x, zeros([length(x),1]), '.r');
```

A.3 Population Parameters

$$\alpha = \int_R g(x) f_X(x)\, dx$$

$$\widehat{\alpha} = \frac{1}{M} \sum_{m=1}^{M} I_{(X_m \in R)} g(X_m)$$

$$I_{(X_m \in R)} g(X_m) = \begin{cases} g(X_m) & X_m \in R \\ 0 & X_m \notin R \end{cases},$$

$$\mathrm{E}\left[I_{(X_m \in R)} g(X_m)\right] = \int_R g(x) f_X(x)\, dx,$$

$$\widehat{\alpha} \xrightarrow{p} \alpha.$$

Matlab Code A.3: Monte Carlo (MC) Simulation

```
P={x1,...,xM}; %Data Generated
S=0;
for m=1:M
    if (R(xm))
        S=S+g(xm);
end;
ret=S/M;
```

General Case

$$\int_R g(x)\, f_X(x)\, dx \approx \frac{1}{M}\sum_{m=1}^{M} I_{(X_m \in R)}\, g(X_m)$$

Special Cases

$$\int g(x)\, f_X(x)\, dx \approx \frac{1}{M}\sum_{m=1}^{M} g(X_m) \qquad (\mu_g)$$

$$\int x f_X(x)\, dx \approx \frac{1}{M}\sum_{m=1}^{M} x_m \qquad (\mu_X)$$

$$\int_R f_X(x)\, dx \approx \frac{1}{M}\sum_{m=1}^{M} I_{(X_m \in R)} \qquad (\Pr(X \in R))$$

Other Cases

$$F^{-1}(0.5) \approx y^{(M/2)} \qquad \text{(Median)}$$

General rule: Generate a pseudo-population $P = \{x_1, \ldots, x_M\}$ for very large M, then treat the data as if it is the population to calculate your function.

Different Variances and MC size M

$$X, \mu_X, \sigma_X, \ldots \alpha_X$$
$$Y, \mu_Y, \sigma_Y, \ldots, \alpha_Y$$
$$P_k, \widehat{\mu}_k, \widehat{\sigma}_k, \ldots, \widehat{\alpha}_k,$$
$$\mathrm{SD}(\widehat{\alpha}), \widehat{\mathrm{SD}(\widehat{\alpha})}$$

In each MC repetition k, we get a pseudo-populati P_k, and hence different histogram and $\widehat{\alpha}_k$ (because of limited M):

$$\widehat{\mathrm{SD}(\widehat{\alpha})} = \sqrt{\frac{1}{K-1}\sum_k \left(\widehat{\alpha}_k - \overline{\widehat{\alpha}}\right)^2}.$$

In case of

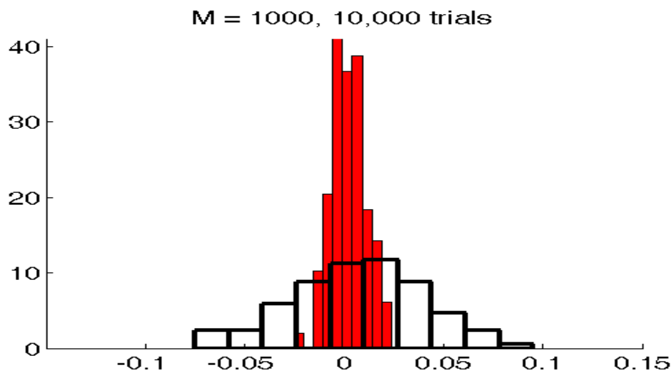
$$\widehat{\alpha} = \frac{1}{M}\sum_{m=1}^{M} y_m,$$

we can get $\mathrm{SD}(\widehat{\alpha})$ from a single MC repetition by

$$\mathrm{SD}(\widehat{\alpha}) = \frac{1}{\sqrt{M}}\mathrm{SD}\left(y\right),$$
$$\widehat{\mathrm{SD}(\widehat{\alpha})} = \frac{1}{\sqrt{M}}\sqrt{\frac{1}{M-1}\sum_m \left(y_m - \overline{y}\right)^2}$$

Matlab Code A.4: MC variation

```
M=1000; mu=0; sigma=1; K=100;
muhat=zeros([1,K]);
for k=1:K %repeating MC
    P=random('normal',mu,sigma,[M,1]);
    S=0;
    for m=1:M
        S=S+P(m);
    end;
    muhat(k)=S/M;
end;
%Shorter:
P=random('normal',mu,sigma,[M,K]);
muhat=mean(P,1);
```



A.4 Statistics

Example, order statistic

$$Y = X^{(n)}. \qquad (Y = Y(n))$$

In MC simulation, for each trial m we generate a dataset:

$$D_m = \{x_1, \ldots x_n\}, \; m = 1, \ldots, M,$$

from which we calculate a single value of our statistic

$$y_m = x^{(n)}$$

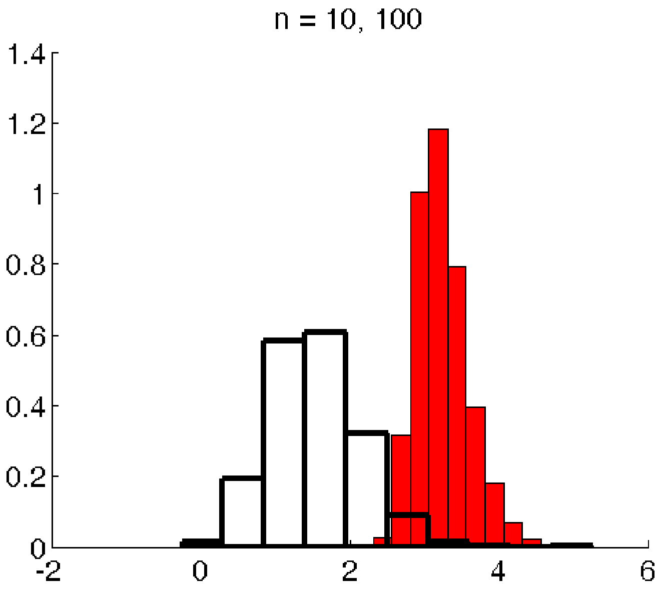
Then

$$\begin{aligned} &X, \mu_X, \sigma_X, \ldots \alpha_X, \\ &Y, \mu_Y, \sigma_Y, \ldots, \alpha_Y \\ &P_k, \widehat{\mu}_k, \widehat{\sigma}_k, \ldots, \widehat{\alpha}_k, \\ &\mathrm{SD}(\widehat{\alpha}), \widehat{\mathrm{SD}(\widehat{\alpha})} \end{aligned}$$

are for particular n.

Matlab Code A.5: Order Statistics MC

```
M=1000;  n=100;  mu=0;  sigma=1;
y=zeros([M,1]);
for m=1:M
    D=random('normal',mu,sigma,[n,1]);
    y(m)=max(D);
end;
```



A.5 Conditional Probability

We will condition on a non-zero probability event:

$$
\begin{aligned}
\alpha &= \int_R g(x)\, f_X(x|e) \\
&= \frac{\int_R g(x)\, f_{XE}(x,e)}{\Pr(e)}, \\
\widehat{\alpha} &= \frac{\frac{1}{M}\sum_{m=1}^{M} I_{(e_m \ \&\ X_m \in R)}\, g(X_m)}{\frac{1}{M}\sum_{m=1}^{M} I_{(e_m)}} \\
&= \frac{\sum_{m=1}^{M} I_{(e_m \ \&\ X_m \in R)}\, g(X_m)}{\sum_{m=1}^{M} I_{(e_m)}}.
\end{aligned}
$$

As if we generate a pseudo-population

$$
P = \{x_1, \ldots, x_M\}, \qquad (|P| = M)
$$

the make up the new dataset of size $\sum_{m=1}^{M} I_{(e_m)}$

$$
P' \{x | x \in P \ \&\ e(x) = T\}
$$

Matlab Code A.6: Algorithm: Cond. Prob. MC

```
P={x1,...,xM}; %Data Generated
S=0; Mprim=0;
for m=1:M
    if (e(xm))
        Mprim=Mprim+1;
        if (R(xm))
            S=S+g(xm);
        end;
    end;
end;
ret=S/Mprim;
```

Bibliography

DeGroot, M. H., Schervish, M. J., 2002. Probability and statistics, 3rd Edition. Addison-Wesley, Boston.

Gastwirth, J. L., 1987. The Statistical Precision of Medical Screening Procedures: Application To Polygraph and Aids Antibodies Test Data. Statistical Science 2 (3), 213–222.

Glass, D., Hall, J., 1954. Social Mobility in Britain.

Knuth, D. E., 1997. The art of computer programming, 3rd Edition. Addison-Wesley, Reading, Mass.

Rosen, K. H., 2007. Discrete mathematics and its applications, 6th Edition. McGraw-Hill Higher Education, Boston.