1 Probability Space

Set = A collection of unique objects

Example : a finite set : { sun, mon, tue, wed, thu, fri, sat }
Example : an infinite set of abstract things: $\{1, 2, 3, \cdots\}$
Order doesn't matter.
Important Sets: Natural numbers, integers, rational numbers, real numbers

Random Experiment = An experiment whose outcome is uncertain

Example : Coin toss
Example : Die roll
Example : Weather (temperature) 5 days after
Example : Lifetime of a laptop battery
Image Example: Intensity at a randomly chosen pixel in an image

Sample Space = A set Ω of all possible outcomes of a random experiment.

Discrete Examples:
- Coin toss: $\Omega = \{H, T\}$
- Die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Photographic B-W (grayscale) digital image pixel values: $\Omega = [0, M]$. For a 32-bit image, $M = 2^{32} - 1$.

Continuous Examples:
- Day temperature 5 days after: $\Omega = (0, 1.416785 \times 10^{32})$ Kelvin. Absolute zero to Plank temperature.
- Processed (e.g., contrast stretching) digital grayscale image pixel values.

Event = A subset of the sample space

Event A = The die roll produces an even number
Event A = The temperature is between 298K and 300K (25 and 30 degrees Celsius)
Event A = Intensity at chosen pixel falls between 10 and 20

Operations on Events. Given 2 events A and B of a sample space Ω:

Union = "or" = $A \cup B$
Intersection = "and" = $A \cap B$
Complement = "negation" = $\bar{A}$
Subtraction = A happens but B doesn't happen = $A - B$

Event Space = Space of all possible events

Given a sample space Ω, the space of all possible events F must satisfy several rules (basically, to deal with sample spaces that are infinite):

(1) Null Set $\phi \in F$
(2) Closed Under Countable Unions: If $A_1, A_2, A_3, \cdots \in F$, then $A_1 \cup A_2 \cup A_3 \cup \cdots \in F$.
Note: we can have a countably infinite number of sets here.
(3) Closed Under Complementation: If $A \in F$, then $\bar{A} \in F$

A set $F \subseteq 2^{\Omega}$ that satisfies the above rules is called a σ-algebra.

Note: 2^{Ω} is the power-set notation. Why choose this notation ? See Wikipedia. Motivation is: the number of all possible subsets of a finite set of size n is 2^n.

Note: The term "algebra" is used because the theory defines an algebra on sets (union, intersection, difference, ...) analogous to the algebra on numbers.

Example: $\Omega = \{a, b, c, d\}$. One possible sigma algebra on Ω is $F = \{\phi, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$

From Wikipedia: A more useful example is the set of subsets of the real line formed by starting with all open intervals and adding in all countable unions, countable intersections, and relative complements and continuing this process until the relevant closure properties are achieved (a construction known as the Borel σ-algebra [see Borel set].

Probability Measure : gives a precise notion of the "size" or "volume" of the sets.

A probability measure on a σ-algebra F is a function $P : F \mapsto [0, 1]$ such that
(1) $P(\phi) = 0$ and
(2) $P(\Omega) = 1$ (for a general measure, this needn't hold)
(3) For pairwise disjoint sets (events) $A_1, A_2, \cdots \in F$, we have $P(A_1, A_2, \cdots) = P(A_1) + P(A_2) + \cdots$.
We want the size of the union of disjoint sets to be the sum of their individual sizes, even for an infinite sequence of disjoint sets.

In simple words: the probability measure on a sample space Ω assigns every event $A \subseteq \Omega$, a number in $[0, 1]$, such that $P(\Omega) = 1$ and $P(A \cup B) = P(A) + P(B)$ for disjoint A, B.

In simple words: this number $\in [0, 1]$ assigned to an event is nothing but the probability/chance that the event occurs.

Properties of Probability Measures

Complement of an event A: $P(\bar{A}) = 1 - P(A)$.
Union of two overlapping events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Difference of events: $P(A - B) = P(A) - P(A \cap B)$

Probability Space = a triple (Ω, F, P) where :

(1) Ω is the sample space (= space of all possible outcomes)
(2) F is the event space (= space of all possible events; σ-algebra)

(3) P is the probability measure on (Ω, F) with $P(\Omega) = 1$

Example: Toss fair coin: What is Ω ? What is F ? What is P ?
Example: Roll fair die: $P(\{1\}) = 1/6$. All outcomes equally likely. $P(\{1, 2, 3\}) = 1/2$.
Example: Toss a drawing pin: $P(\{\text{pin pointing upwards}\}) \neq 1/2$. All outcomes not equally likely.

Probability Space for Multiple Random Experiments

(1) Repeated Experiments: If we do two runs of an experiment with sample space Ω, then we get a new experiment with sample space: $\Omega \times \Omega = \{(x, y) : x \in \Omega, y \in \Omega\}$
Note: (x, y) is an ordered pair; not a set. So, order matters.

Example: Toss two fair coins: What is Ω ? What is F ? What is P ?
If we do m runs of an experiment, then ...

(2) Two Different Experiments: If we do two experiment with sample spaces Ω_1 and Ω_2, then we get a new experiment with sample space: $\Omega_1 \times \Omega_2 = \{(x, y) : x \in \Omega_2, y \in \Omega_2\}$

Example: First toss a coin. Then roll a die. What is Ω ? What is F ? What is P ?

2 Conditional Probability, Total Probability, Independence

Conditional Probability

Definition: The conditional probability of event A given event B = $P(A|B) := P(A \cap B)/P(B)$

Dice example: $A = \{2\}$, $B = \{2, 4, 6\}$ (even number), $P(A) = 1/6$, but $P(A|B) = 1/3$

Image example: $P(\text{observed intensity})$ usually differs from $P(\text{observed intensity}|\text{true intensity})$! Former = image histogram. Latter = noise model.

Image example: $P(\text{intensity at a pixel})$ usually differs from $P(\text{intensity at a pixel}|\text{intensity at adjacent pixel})$! Former = image histogram. Latter = local dependence.

Partition

Definition: A set of events $B_1, \cdots, B_n$ is a **partition** of the sample space Ω if they are mutually exclusive (pairwise disjoint; $B_i \cap B_j = \phi$) and exhaustive ($\cup_i B_i = \Omega$)

Total Probability

Definition: Given a partition $B_1, \cdots, B_n$, the total probability $P(A) := \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$

Box Example: You have two boxes, one with 4 black balls and 1 white balls, the other with 1 black balls and 3 white balls. You pick one box at random and then select a ball from the box. What is the probability the ball is white ?
Solution: Tree diagram. $P(W) = P(W|Box1)P(Box1) + P(W|Box2)P(Box2)$

Independence

Definition: Two events A and B are independent iff $P(A|B) = P(A)$

Equivalent Definition: Two events A and B are independent iff $P(A \cap B) = P(A)P(B)$

Dice example: Roll two dice. Event A = see 1 on first die. Event B = see 1 on second die. $P(A \cap B) = 1/36$
Image example: $P(\text{“corruption in observed intensity at pixel a”} \cap \text{“corruption in observed intensity at pixel b”})$

Conditional Independence

Definition: Two events A and B are conditionally independent given C iff $P(A, B|C) = P(A|C)P(B|C)$

Equivalent Definition: Two events A and B are conditionally independent given C iff $P(A|B, C) = P(A|C)$

Interpretation: Given C, B doesn't add any information on the probability of A

Image Example: Events $A =$ intensity at pixel a in WM tissue and $B =$ intensity at pixel b in WM tissue are dependent, but they are conditionally independent given the event $C =$ the true intensity of WM tissue. Thus, $P(A, B|C) = P(A|C)P(B|C)$. Latter = noise model.

Image Example: Consider the sequence (Markov process) of values $x_0 =$ rand, $x_{i+1} = x_i + 10$. Then, $P(x_5|x_1, \cdots, x_4) = P(x_5|x_4)$. Thus, x_5 is conditionally independent of x_1, x_2, x_3, given x_4. But x_5 isn't independent of x_1, x_2, x_3 !

3 Random Variable (Discrete, Continuous), CDF, PMF/PDF, Transformation

Random Variable

– Definition: A random variable is a function defined on a probability space $\{\Omega, F, P\}$.
– $X : \Omega \mapsto V$ where V is the set of numerical values. e.g., set of integers or real numbers.
– Note: The name "random variable" is a misnomer.
– RVs are measurable functions.
– A RV is an abstraction.

Dice example: Roll one die and square the number. What is Ω ? What is X ?
Dice example: Roll two dice and sum the numbers. What is Ω ? What is X ?
Dice example: Roll one die and take the number. Here X is identity.

Image example: Pick a pixel location at random and take the intensity at that pixel. What is Ω, X ? The abstraction is helpful in applications e.g., histogram equalization, contrast stretching, where we don't care about where a certain intensity is found in the image. Other applications like denoising, segmentation, use more sophisticated models.

Discrete Random Variable

– Values taken by the RV form a discrete set.
– The cardinality of the set of possible values can be finite or (countable) infinite.
– Note: Sample space is usually a discrete (countable) set, but not necessarily.

Example 1: Roll die and take number on top face (finite sample space).
Example 2: Number of packets transmitted by a computer on a network.

Image Example: Number of photons hitting an X-ray detector.

Continuous Random Variable

– The RV takes a continuous range of values.
– The cardinality of the set of possible values is (uncountable) infinite.
– Note: Cardinality of the sample space must be uncountable infinite.

Example 1: Heights of children in school.
Example 2: Width of a leaf.
Image Example: Intensities in a magnitude-MR image.

Events via Random Variables

Examples:
$\{X = a\} = \{s \in \Omega : X(s) = a\}$

$\{X < a\} = \{s \in \Omega : X(s) < a\}$
$\{a < X < b\} = \{s \in \Omega : a < X(s) < b\}$
Notation: Upper case = random variable. Lower case = values taken by the random variable.

Image example: Consider an 8-bit grayscale image with intensities from 0 to 255. Event A = pixel intensity is 100. Event B = pixel intensity between 100 and 200.

Event Probabilities via Random Variables

Example: $P(\{a < X < b\}) = P(\{s \in \Omega : a < X(s) < b\})$

Image example: Consider an 8-bit grayscale image with intensities from 0 to 255. P (pixel intensity between 100 and 200). This takes us to the notion of a histogram (we will define it soon).

Cumulative Distribution Function (Discrete RV and Continuous RV)

Definition: For a real-valued random variable X, the CDF $f(x) := P(X \leq x)$
Note: $\forall x, f(x) \in [0, 1]$

Properties of CDFs:
(1) f is monotonically non-decreasing
(2) f is right continuous. $\lim_{\epsilon \to 0^+} f(x + \epsilon) = f(x), \forall x \in \mathbb{R}$
(3) $\lim_{x \to -\infty} f(x) = 0$
(4) $\lim_{x \to +\infty} f(x) = 1$

♣ See picture

Probability Mass Function (Discrete RV)

Definition: (lowercase) $p(x) = P(X = x)$
Note: $\forall x, p(x) \in [0, 1]$

The CDF can be defined in terms of the PMF: $f(x) = P(X \leq x) = \sum_{y \leq x} p(y)$

♣ See picture

Bernoulli Distribution (Discrete RV)

RV = success / failure at a task.
$p(x = 1) = \alpha$
$p(x = 0) = 1 - \alpha$
Example: (x = 1) = success in putting a basketball through the basket. (x = 0) = failure.

Binomial Distribution (Repeated Bernoulli Trials) (Discrete RV)

RV = number of successes observed in multiple tries
Probability of getting k successes from n trials = $p(k) = C_k^n \alpha^k (1-\alpha)^{1-k}$

Geometric Distribution (Repeated Bernoulli Trials) (Discrete RV)

Definition 1: RV = number of trials made to get the first success
Probability of making k trials to see a success = $p_1(k) = (1-\alpha)^{k-1}\alpha$, for $k = 1, 2, 3, \cdots$

Definition 2: RV = number of failures before the first success
Probability of k failures before we see a success = $p_2(k) = (1-\alpha)^k \alpha$, for $k = 0, 1, 2, 3, \cdots$

Note: $p_1(k=1) = p_2(k=0)$. In general: $p_1(k=n) = p_2(k=n-1)$
Note: $\sum_k p_1(k) = 1 = \sum_k p_2(k)$
Note: $p_1(k)$ is called the *shifted* geometric distribution

Probability Density Function (Continuous RV)

Definition: $p(x) = \frac{d}{dx} f(x)$, where the derivative exists (we consider cases when it always does).
Note: $\forall x, p(x) \geq 0$ (Because the CDF is non-decreasing)

The CDF can be defined in terms of the PDF: $f(x) = P(X \leq x) = \int_{-\infty}^{x} p(y)dy$.

Note: $\lim_{x\to\infty} f(x) = 1 \implies \int_{-\infty}^{\infty} p(y)dy = 1$

Note: PDF values aren't probabilities. Probabilities are obtained by integrating the PDF.
$P(a < X < b) = \int_a^b p(y)dy$

Uniform Distribution

X is uniformly distributed between a and b.
PDF:
$p(x) = 0, \forall x < a$ or $x > b$
$p(x) = 1/|b-a|, \forall a \leq x \leq b$
CDF:
$f(x) = 0, x < a$
$f(x) = 1, x > b$
$f(x) = (x-a)/(b-a), a \leq x \leq b$
Note: a and b are called parameters.

Exponential Distribution

PDF:
$p(x) = 0, \forall x < 0$
$p(x) = \lambda \exp(-\lambda x), \forall x \geq 0$
CDF:

$f(x) = 1 - \exp(-\lambda x), \forall x < 0$
$f(x) = 0, \forall x \geq 0$

Normal (Gaussian) Distribution

$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{(x-\mu)^2}{2\sigma^2})$, with parameters μ, σ

Definition: A *standard* normal distribution has $\mu = 0$ and $\sigma = 1$.

♣ See picture

Prove that the Gaussian integrates to 1.
1) Substitute $t = (x - \mu)/\sigma$
2) Then, Gaussian integral is $J/\sqrt{2\pi}$, where $J = \int_{-\infty}^{\infty} \exp(-t^2/2)dt$
3) Then $J^2 = (\int_{-\infty}^{\infty} \exp(-u^2/2)du)(\int_{-\infty}^{\infty} \exp(-v^2/2)dv) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-(u^2+v^2)/2)dudv$
4) Change to polar coordinates: $u^2 + v^2 = r^2; dudv = (rd\theta)(dr)$
5) Then $J^2 = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} \exp(-r^2/2)rdrd\theta = 2\pi \int_{r=0}^{\infty} \exp(-r^2/2)rdr$
6) Substitute $w = r^2/2; dw = rdr$
7) Then $J^2 = 2\pi \int_{r=0}^{\infty} \exp(-w)dw = 2\pi[-\exp(-w)]_0^{\infty} = 2\pi$
8) So $J = \sqrt{2\pi}$. QED.

Mixed Distribution: Discrete+Continuous RV

Example: Unusual cases: a (mixed) CDF can have both a continuous part and a discrete part (see Wikipedia).

Normal Distribution in Real Life

(From Wikipedia): Certain quantities in physics are distributed (*exactly / theoretically*) Normally, as was first demonstrated by James Clerk Maxwell.

(1) Velocities of the molecules in the *ideal gas*. An ideal gas is a theoretical gas composed of a set of randomly moving, non-interacting point particles. Generally, a gas behaves more like an ideal gas at higher temperature and lower pressure, as the work which is against intermolecular forces becomes less significant compared with the particles' kinetic energy, and the size of the molecules becomes less significant compared to the empty space between them.
More generally, velocities of the particles in any system in thermodynamic equilibrium will have normal distribution, due to the maximum-entropy principle.

(2) The position of a particle that experiences *diffusion*. If initially the particle is located at a specific point (that is its probability distribution is the Dirac-delta function), then after time t its location is described by a normal distribution with $\sigma^2 = t$, which satisfies the diffusion/heat equation: $\frac{\partial}{\partial t} f(x,t) = 0.5 \frac{\partial}{\partial x^2} f(x,t)$ with the initial condition $f(x, t = 0) = \delta(x)$.
If the initial location is given by a certain density function $g(X)$, then the density at time t is the convolution of g and the normal PDF.

♣ See dye-diffusion video at:
http://en.wikipedia.org/wiki/Diffusion
http://upload.wikimedia.org/wikipedia/commons/d/d5/Diffusion_v2_20101120.ogv

(3) Central Limit Theorem
The PDF of the average of a large number of RVs (irrespective of their individual PDFs) tends to a Gaussian.

♣ See pictures

Function of a RV

For a RV X and any continuous function $g : \mathbb{R} \to \mathbb{R}$,
$g(X)$ is also a RV that shares the same probability space as X.
Illustrate via diagram.

Transformation of a RV

Consider a RV X with PDF $p(X)$.
Consider a transformed variable $Y = g(X)$, where $g(\cdot)$ is an **increasing** function (we consider only the special case of monotonic functions).

– What is the PDF $p(Y)$?

– $P(a < X < b) := \int_a^b p(x)dx$
– Substitute $y = g(x)$ in the integral and write the integral in terms of y. Then, $dx = \frac{d}{dy}g^{-1}(y)dy$
– Then, $P(a < X < b) = \int_{g(a)}^{g(b)} p(g^{-1}(y))\frac{d}{dy}g^{-1}(y)dy$
– Consider $q(y)$ as the PDF of Y
– Now, $P(g(a) < Y < g(b)) = \int_{g(a)}^{g(b)} q(y)dy$
– Because we assumed increasing $g(\cdot)$, $P(g(a) < Y < g(b)) = P(a < X < b)$
– Thus, $q(y) = p(g^{-1}(y))\frac{d}{dy}g^{-1}(y)$

– If $g(\cdot)$ is increasing, then (i) $a < b \implies g(a) < g(b)$ and (ii) the derivative $\frac{d}{dy}g^{-1}(y)$ is non-negative. So, the above formula holds good.

– If $g(\cdot)$ is decreasing, then (i) $a < b \implies g(a) > g(b)$ and (ii) the derivative $\frac{d}{dy}g^{-1}(y)$ is negative. In this case, we can negate the derivative and switch the upper and lower limits to retain the same analysis.

– For convenience, $q(y) = p(g^{-1}(y)) \left|\frac{d}{dy}g^{-1}(y)\right|$

Classic Example 2 : Consider a RV $X \sim G(0, 1)$ (standard normal distribution). Consider $Y = aX$ with

$a > 0$. What is $q(Y)$?

$$y := ax \implies x = y/a \implies g^{-1}(y) = y/a$$
$$\left| \frac{d}{dy} g^{-1}(y) \right| = 1/a$$
$$q(y) := p(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = p\left(\frac{y}{a}\right) \frac{1}{a} = \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{y^2}{2a^2}\right)$$

– Thus, $p(Y)$ is also a Gaussian with σ^2 scaled by a factor of a^2

Classic Example 3 : Consider a RV $X \sim G(0, a^2)$. Consider $Y = b + X$. What is $q(Y)$?

$$y := b + x \implies x = y - b \implies g^{-1}(y) = y - b$$
$$\left| \frac{d}{dy} g^{-1}(y) \right| = 1$$
$$q(y) := p(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = p(y - b) \cdot 1 = \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{(y-b)^2}{2a^2}\right)$$

– Thus, $p(Y)$ is also a Gaussian with μ translated by b

4 Expectation

– Expectation = Expected Value = Mean
– Indicates the center of mass in the probability mass/density function.

Definition: Expectation of a Discrete RV: $E[X] := \sum_i x_i P(X = x_i)$

Dice Example: $E[X] = 3.5$

Note: "Expected value" isn't the value that is most likely to be observed in the random experiment.

Alternate Interpretation of Expectation: Illustrate with diagram:
$E[X] := \sum_i x_i p(X = x_i) = \sum_i x_i \left(\sum_{s \in \Omega : X(s) = x_i} P(s) \right) = \sum_{s \in \Omega} X(s) P(s)$.

Important Concept: "Expected value" is what one would "expect" to get if one could repeat the random experiment an infinite number of times and take the average (arithmetic mean) of the values obtained. Thus, $E[X] = \lim_{N \to \infty} (1/N) \sum_{n=1}^{N} x_n$, where $x_n \sim P(X)$.

Proof of the Important Concept:
In N random experiments, let the number of times that the outcome is s be $N_s := \sum_{n=1}^{N} I_s(n)$
Then, $\sum_s N_s = N$
By definition of (frequentist) probability, the probability of the outcome being s is $P(s) := \lim_{N \to \infty} N_s / N$
Let the value taken by the RV in the n-th random experiment be $x_n := \sum_{s \in \Omega} X(s) I_s(n)$
Substitute expression for $P(s)$ in $E[X] = \sum_{s \in \Omega} X(s) P(s) = \sum_{s \in \Omega} X(s) \left(\lim_{N \to \infty} (1/N) \sum_{n=1}^{N} I_s(n) \right) = \lim_{N \to \infty} (1/N) \sum_{n=1}^{N} \left(\sum_{s \in \Omega} X(s) I_s(n) \right) = \lim_{N \to \infty} (1/N) \sum_{n=1}^{N} x_n$

Definition: Expectation of a Continuous RV: $E[X] = \int_{-\infty}^{+\infty} x p(x) dx$

Example: Expectation of a Uniform Random Variable: $E[X] = \int_a^b x(1/(b-a))dx = [x^2/2]_a^b/(b-a) = (a+b)/2$
Example: Expectation of a Gaussian Random Variable: $E[X] = \mu$. Prove.
1) Substitute $t = (x - \mu)/\sigma$
2) Use property of odd functions. Part of integral is zero.
3) Other integral is easy when we use the fact that the Gaussian integrates to 1.

Linearity of Expectation

For RVs X and Y that share the same probability space (Ω, F, P), the following rules hold:
(1) $E[X + Y] = E[X] + E[Y]$. Prove.
(2) $E[X + c] = E[X] + c$. Prove.
(2) $E[aX] = aE[X]$. Prove.

Proof 1: $E[X + Y] = \sum_{s \in \Omega} (X(s) + Y(s)) P(s) = E[X] + E[Y]$

Proof 2 (c is like a RV that always takes the same value for any outcome of the random experiment): $E[X+c] = \sum_{s\in\Omega}(X(s)+c)P(s) = \sum_{s\in\Omega} X(s)P(s) + c\sum_{s\in\Omega} P(s) = E[X] + c$

Proof 3: $E[aX] = \sum_{s\in\Omega} aX(s)P(s) = aE[X]$

This extends to several RVs: $E[\sum_i a_i X_i] = \sum_i a_i E[X_i]$

Expectation of a Function of a RV

$E_{X\sim P(X)}[g(X)] = \sum_i g(x_i)p(x_i)$

$E_{X\sim P(X)}[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)dx$

5 Variance

– A measure of the spread of the mass, in the PMF or PDF, around the mean.

Definition: Var$(X) = E[(X - E[X])^2]$

Property: Variance is always non-negative.

Property: Var$(X) = E[X^2] - (E[X])^2$. Prove.

Definition: Standard deviation is the square root of the variance.

Units of variance = square of units of the random variable.
Units of standard deviation = units of the random variable.

Variance of a Uniform Random Variable

Evaluate.

Variance of a Gaussian Random Variable

– Variance of a Gaussian PDF is σ^2. Prove.
1) Substitute $t = (x - \mu)/\sigma$
2) Required integral = $(\sigma^2/\sqrt{2\pi}) \int_{-\infty}^{\infty} t^2 \exp(-t^2/2) dt$
3) Use following trick to evaluate $\int t^2 \exp(-t^2/2)$. Use integration by parts where first part is t and second part is $t \exp(-t^2/2)$
4) Finally, use the fact that the Gaussian integrates to 1.

6 Joint RV, Marginal Distribution, Conditional Distribution

Joint RV: Given two RVs X and Y sharing the same probability space (Ω, F, P), we can define a joint RV (X, Y) on the shared probability space.

Note: To join variables, we may have to construct a suitable probability space.

(Can Ignore) Example 1 (Independent RVs): Consider $X \equiv$ coin toss and $Y \equiv$ die roll. Then, the probability space for (X, Y) can be as follows:
– $\Omega := \Omega_1 \times \Omega_2 = \{(i, j) : i \in \{-1, 1\}, j \in \{1, 2, 3, 4, 5, 6\}\}$
– $F := F_1 \times F_2$
– $P(\{(X, Y) = (i, j)\}) := P(\{X = i\})P(\{Y = j\})$
– $X((i, j)) := f(i)$ (could be any function of i)
– $Y((i, j)) := g(j)$ (could be any function of j)
Property: $E[(X, Y)] = (E[X], E[Y])$. Prove.
Proof: $E[(X, Y)] = \sum_{s\in\Omega}(X(s), Y(s))P(s) = (\sum_{s\in\Omega} X(s)P(s), \sum_{s\in\Omega} Y(s)P(s))$
Here, $\sum_{s\in\Omega} X(s)P(s) = \sum_i \sum_j f(i)P((i, j)) = \sum_i f(i) \sum_j P((i, j)) = \sum_i f(i)P(i) = E[X]$

Example 2 (Dependent RVs): Consider a random experiment of a dice throw. For each throw, $X(i) = i$, $Y(i) = i^2$.
– $\Omega = \{1, 2, 3, 4, 5, 6\}$, F, $P(s \in \Omega) = 1/6$
Property: $E[(X, Y)] = (E[X], E[Y])$. Prove.
Proof: $E[(X, Y)] = \sum_{s\in\Omega}(X(s), Y(s))P(s) = (\sum_{s\in\Omega} X(s)P(s), \sum_{s\in\Omega} Y(s)P(s)) = (E[X], E[Y])$

Example 3 (Dependent RVs): Consider a 1D binary image of 10 pixels with intensities $[0, 0, 0, 1, 1, 1, 1, 1, 0, 0]$. Pick a pixel location s at random. $X(s) =$ intensity at location s. $Y(s) =$ intensity at location $s+1$ (wrap around boundary).
– $\Omega = \{1, 2, \cdots, 10\}$, F, $P(s \in \Omega) = 1/10$
– Joint PMF (table): $P(x = 0, y = 0) = 0.4$, $P(x = 1, y = 1) = 0.4$, $P(x = 1, y = 0) = 0.1$, $P(x = 0, y = 1) = 0.1$
– Marginal PMF: $P(x = 0) = 0.5$, $P(x = 1) = 0.5$, $P(y = 0) = 0.5$, $P(y = 1) = 0.5$

Joint CDF and Joint PMF/PDF

Joint CDF: $f(x, y) = P(X \leq x, Y \leq y) = P(\{X \leq x\} \cap \{Y \leq y\})$
Joint PMF (discrete): $p(x, y) = P(X = x, Y = y)$
Joint PDF (continuous): $p(x, y) = \frac{d^2}{dxdy} f(x, y)$

Marginal Distribution (similar to total probability)

Marginal PMF: $p(x) = \sum_j p(x, y_j)$. Because $P(\{X = x\}) = P(\{X = x\}, \{Y \in \{y_1, y_2, \cdots\}\}) = \sum_j P(\{X = x\}, \{Y = y_j\})$.
Marginal PDF: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$

Conditional Distribution (similar to conditional probability)

$p(x|y) = p(x,y)/p(y)$

Independent Random Variables

PDF: $p(x,y) = p(x)p(y)$
CDF: $f(x,y) = f(x)f(y)$

Conditional Expectation = expectation w.r.t. the conditional PDF

Discrete: $E[X|Y=y] := \sum_i x_i P(X = x_i|Y = y)dx$
Continuous: $E[X|Y=y] := \int_{-\infty}^{\infty} xp(x|y)dx$

7 Covariance, Correlation

Covariance

Assumption: RVs X and Y share the *same* probability space.

Covariance = A measure of how the values taken by the two RVs vary together ("co"-"vary").

Definition: Cov$(X, Y) := E[(X - E[X])(Y - E[Y])]$

What does this mean ?
– Example: Consider a probability space (Ω, F, P). Consider two RVs X and Y. Define another two RVs U and V, where $U(X) = X - E[X]$ and $V(Y) = Y - E[Y]$. Then, define a function $f(X, Y) := U(X)V(Y)$ and take its expectation over the PDF P.

Property: Covariance is symmetric : Cov$(X, Y) =$ Cov(Y, X).

Property: Cov$(X, Y) = E[XY] - E[X]E[Y]$. Prove (by expansion).

(From Wikipedia):
– If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive.
– In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative.
– The sign of the covariance therefore shows the tendency in the *linear* relationship between the variables.
– The magnitude of the covariance isn't easy to interpret.

Correlation

A scaled version of covariance ("co"-"relate").
Here, magnitude is easy to interpret. The magnitude shows the strength of the linear relation.

$\rho(X, Y) = E\left[\left(\frac{X - E[X]}{\text{SD}(X)}\right)\left(\frac{Y - E[Y]}{\text{SD}(Y)}\right)\right] = \text{Cov}(X, Y)/(\text{SD}(X)\text{SD}(Y))$

Property: $-1 \leq \rho(X, Y) \leq 1$. Prove.

First prove Cauchy-Schwartz inequality: $(E[XY])^2 \leq E[X^2]E[Y^2]$
Proof (of C-S inequality): $0 \leq E[(X + tY)^2] = E[X^2] + 2tE[XY] + t^2E[Y^2]$.
This is a quadratic in t which can't be negative. So it has either a single real root (discriminant = 0) or both imaginary roots (discriminant $<$ 0).
Thus, $(2E[XY])^2 - 4E[X^2]E[Y^2] \leq 0$. QED.

Now apply the C-S inequality to the definition of covariance to get:

$(\text{Cov}(X,Y))^2 \leq (E[X - E[X]])^2 (E[Y - E[Y]])^2 = \text{Var}(X)\text{Var}(Y)$
QED.

Independence Implies Uncorrelated

We know $\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$
But independence of X and Y implies $E[XY] = E[X]E[Y]$.
Thus, $\text{Cov}(X,Y) = 0 = \rho(X,Y)$.

Uncorrelated Doesn't Imply Independence

Correlation indicates the strength of a *linear* relationship between two variables, but its value generally doesn't completely characterize their relationship.

Example: Let X be *uniformly* distributed in $[-1,1]$ and let $Y = X^2$. Clearly, X and Y are dependent, but their covariance is 0.
$\text{Cov}(X, X^2) = E[XX^2] - E[X]E[X^2] = E[X^3] - 0 \cdot E[X^2] = 0$

♣ See pictures of examples from Wikipedia:

http://en.wikipedia.org/wiki/Correlation_and_dependence

8 Multivariate Gaussian

– Generalizes a univariate Gaussian.

– Consider a vector random variable $X = [X_1, X_2, \cdots, X_d]^T$. Nothing but a joint RV with d RVs. Represent as a $d \times 1$ vector.

Definition: The RV X has a multivariate (jointly) Gaussian PDF if $\exists$ a finite set of i.i.d. univariate standard-normal RVs $W_1, \cdots, W_n$ such that each X_d can be expressed as $X_d = \mu_d + \sum_n A_{dn} W_n$ (i.e., $X = AW + \mu$).

Example 1 (Zero Mean + Isotropic): The case of independent standard-normal RVs $W_1, \cdots, W_d$ with $A = I_{d \times d}$ and $\mu = 0$, i.e. $X = W$
Then, the Gaussian PDF is $p(w) = \Pi_d p(w_d) = \frac{1}{(2\pi)^{d/2}} \exp(-0.5 w^T w)$

Example 2 (Zero Mean + Anisotropic): What is the PDF $q(X)$ for arbitrary non-singular A and $\mu = 0$?
– Recall: Given PDF $p(w)$ and the transformation $X = g(W)$, the PDF $q(x) = p(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right|$
– In our case, $X = g(W) = AW$
– The inverse transformation $g^{-1}(X) = W = A^{-1}X$
– We want the *magnitude* of the derivative of the inverse transformation: $\left| \frac{d}{dx} A^{-1} X \right| = \left| A^{-1} \right| = \frac{1}{\det(A)}$
** Geometric intuition for $\left| A^{-1} \right| = \frac{1}{\det(A)}$
** Observe that the linear transformation A maps an infinitesimal hyper-cube $\delta \times \cdots \times \delta$ to an infinitesimal hyper-parallelepiped. If the axes of the hyper-cube were the cardinal axes, then the axes of the hyper-parallelepiped are the columns of A !!
** The volume of the hyper-parallelepiped is $\delta^d \det(A)$. In 3D, the volume can also be written as the scalar triple product $a_1 \cdot (a_2 \times a_3)$ where a_i is the i-th column of A
** Thus, $|dw| = \delta^d \implies |dx| = \delta^d \det(A)$
** Thus, $\frac{|dw|}{|dx|} = \frac{1}{\det(A)}$
– Finally, the transformation of variables gives :
$q(X) = p(A^{-1}X) \frac{1}{\det(A)} = \frac{1}{(2\pi)^{d/2} \det(A)} \exp(-0.5 X^T (A^{-1})^T A^{-1} X)$
– Simplify: Let $C := AA^T$. Then, $C^{-1} = A^{-T} A^{-1}$ and $\det(C) = \det(A) \det(A^T) = (\det(A))^2$
– So, the multivariate-normal PDF $q(X) = \frac{1}{(2\pi)^{d/2} |C|^{0.5}} \exp(-0.5 X^T C^{-1} X)$, where C is called the covariance matrix.

Property: The mean of $X = AW$ is zero. Prove.
Proof: $E[AX] = AE[X] = A \cdot 0 = 0$
Note: $E[X] = [E[X_1], E[X_2], ..., E[X_d]]^T$ (recall: all X_i share the same probability space).

Example 3 (Nonzero mean + Anisotropic): If X is multivariate normal, then $Y = X + \mu$ is multivariate normal with PDF $p(y) = \frac{1}{(2\pi)^{d/2} |C|^{0.5}} \exp(-0.5 (y - \mu)^T C^{-1} (y - \mu))$
Proof:
– Y is multivariate normal because each Y_d can be expressed as $Y_d = AW + \mu$ where W_n is i.i.d. standard normal.

– PDF $p(y) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5(y-\mu)^T C^{-1}(y-\mu))$ because of the transformation of the variables $Y = X + \mu$

Property: The *mean vector* of $X = AW + \mu$ is μ.
Proof: $E[AX + \mu] = AE[X] + \mu = \mu$

Covariance Matrix of $X = AW + \mu$ is $C = AA^T$ where $C_{ij} = \text{Cov}(X_i, X_j)$.
Analysis:
– Observe that $E[(X - E[X])(X - E[X])^T]$ equals such a matrix C where the outer-product structure implies that $C_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$ which equals $\text{Cov}(X_i, X_j)$.
– Covariance matrix $\text{Cov}(W) = E[WW^T] = I$ because (i) $\text{Cov}(W_i, W_i) = 1$ and (ii) $\text{Cov}(W_i, W_{j\neq i}) = 0$ because of independence of W_i and W_j
– Covariance matrix $\text{Cov}(X) = E[(X - E[X])(X - E[X])^T] = E[(AW)(AW)^T] = E[AWW^T A^T] = AE[WW^T]A^T = AA^T$

Property: If Y is multivariate normal, then $Z = BY + c$ is multivariate normal. Prove.
Proof: Because Y is multivariate normal, $Y = AW+\mu$. Thus, $Z = B(AW+\mu)+c = (BA)W+(B\mu+c)$

Property: Marginal PDFs of the multivariate Gaussian Z in n-dimensions, over any chosen subset of the variables, are (multivariate) Gaussian. Prove.
Proof: Choose the transformation B as the projection matrix of size $m \times n$ where $m < n$ with ones of diagonal and zeros elsewhere.

See 2D example. Explain what is a covariance matrix. See pictures from Wikipedia.

More properties of C:
1) $C = E[XX^T] - E[X](E[X])^T$. Proof: Expand the terms in the definition.
2) C is symmetric. Proof: $C_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = C_{ji}$
3) C is positive semi-definite. Prove.
Proof: For any $d \times 1$ vector z, $a^T C a = E[a^T(X - E[X])(X - E[X])^T a] = E[(f(X))^T f(X)] \geq 0$ that is the variance of a scalar RV $f(X) = (X - E[X])^T a$

Note: Every $N \times N$ *real symmetric* matrix A (like the covariance matrix C) has an eigen-decomposition $A = Q\Lambda Q^T$, where Q is an orthogonal matrix (i.e., $Q^T Q = QQ^T = I$).

Note: For a general $N \times N$ matrix A, the eigen-decomposition is $A = Q\Lambda Q^{-1}$, where the columns of Q are the eigenvectors. However, the eigenvectors needn't be orthogonal and Q^{-1} needn't equal Q^T.

9 Modeling, Model Fitting, Estimation, Noise Model

Statistical Model

Typically, a probabilistic description of real-world phenomena. Description involves a PDF that may involve some parameters.

Examples:
1) Distribution of numbers seen on top face of a die, after a die roll (for a *unfair* die; uniform; discrete; 5 parameters $p_1, \cdots, p_5, 1 - \sum_i p_i$).
2) Distribution of IQs of people (Gaussian; parametric functional form)
3) Distribution of heights (or weights) of a 1 year old (Gaussian)
4) Distribution of marks in a course (Gaussian)
5) Distribution of complex data in MRI (Gaussian)

Fitting a Model to Data: Parameter Estimation

(From Wikipedia): Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured/empirical data that has a random component. The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data.

Given: Data
Task: Find the model that generated the data.

Example:
1) Finding the distribution of heights (or weights) of a 1 year old:
– Assume a functional form of the PDF, e.g., Gaussian
– Then estimate the parameters μ, C
– What data do we need ?
2) Finding the distribution of heights (or weights) of a 1 year old (Gaussian)
– No assumption on the functional form of the PDF (typically used for a discrete RV)
– Estimate $p_1, \cdots, p_5$
– What data do we need ?

Estimator

(From Wikipedia):
In statistics, an estimator is a rule for calculating an estimate of a given quantity based on observed data: thus the rule and its result (the estimate) are distinguished.

Noise Model

Data can be corrupted.
(From Wikipedia): Noise is a general term for unwanted (and, in general, unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion

So, data is typically the signal corrupted by noise.

Example:
1) Human error in noting down the number of dots on the face of a die.
2) Human error and machine errors in height measurement of an infant.
3) MRI-machine error in measurement of a signal. Due to thermal noise.
4) Physics-based variation in signal received. e.g., Number photons arriving at a telescope (astronomy) or sensor (photography). Also called shot noise.

See pictures of noise:
1) In TV video: random dot pattern.
http://en.wikipedia.org/wiki/Noise_(video)
In TV/radio audio: hiss and hum.
2) In photographic images:
http://en.wikipedia.org/wiki/Image_noise
3) In medical images: X ray, MRI, DW MRI, fMRI,

Statistical model of noisy data:
(1) Per-pixel model: PDF $P(\text{Data}|\text{TrueSignal}, \text{PixelLocation})$
(2) Spatial model: spatial correlations (or their lack of).

Examples:
(1) (Weakly) White noise = A discrete signal whose samples are regarded as a sequence of serially *uncorrelated* RVs with *zero mean* and *finite variance*.
– In the image-processing context, "zero mean" means that $E[\text{Data}|\text{TrueSignal}] = \text{TrueSignal}$

(2) (Strongly) White noise = A discrete signal whose sample points are regarded as a sequence of serially *independent* RVs with *zero mean* and *finite variance*. Such a noise has a flat power spectrum.
– A special case of (strongly) white noise is where the noise at each pixel is (i) independent and (ii) identically distributed; called "i.i.d".

Notes: The power spectrum P of a random vector W can be defined as the expected value of the squared modulus of each coefficient of its Fourier transform W, that is, $P_I = E(|W_i|^2)$. Under that definition, a Gaussian white noise vector will have a perfectly flat power spectrum, with $Pi = \sigma^2$ for all i.

Notes: An example of a random *vector* that is "Gaussian white noise" in the weak but *not* in the strong sense is $X = [X_1, X_2]$ where X_1 is a normal random variable with zero mean, and X_2 is equal to $+X_1$ or $-X_1$, with equal probability. The joint PDF for this case is like a cross $\times$.

10 Noise Models in Medical Imaging

Gaussian PDF (i.i.d.)

– Most common type of noise. In the class of white-noise models.
– Thermal noise in all electronic systems. Unavoidable at non-zero temperature. Occurs in the (complex) data acquired by the MRI scanner.

(From Wikipedia): Electronic noise generated by the thermal agitation of the charge carriers (usually the electrons) inside an electrical conductor at equilibrium, which happens regardless of any applied voltage.

– Central Limit Theorem states that the sum of many RVs (irrespective of their PDFs) tends to a Gaussian.

– Signal-to-noise ratio = SNR$[X] = E[X]/$SD$(X) = \mu/\sigma$
– Noise level given by σ

Poisson PMF (i.i.d.)

– In the class of white-noise models.
– Situations where an image is created by the accumulation of photons over a detector, e.g., X ray, X-ray CT (monochromatic), CCD cameras, and infrared photometers.

(Wikipedia on Shot Noise): Shot noise exists because phenomena such as *light* and *electric current* consist of the movement of discrete (also called "quantized") 'packets' (*photon* or *electric charge*). Consider light—a stream of discrete photons—coming out of a laser pointer and hitting a wall to create a visible spot. The fundamental physical processes that govern light emission are such that these photons are emitted from the laser at random times; but the many *billions of photons* needed to create a spot are so many that the brightness, the number of photons per unit time, *varies only infinitesimally* with time. However, if the laser brightness is reduced until only a handful of photons hit the wall every second, the *relative fluctuations* in number of photons, i.e., brightness, will be significant, just as when tossing a coin a few times. These fluctuations are shot noise.
Basically, "billions of photons" $\implies$ sum *of a billion photon emitters* $\implies$ *SNR increased* $\implies$ *"relative fluctuations" reduced.*

– Noise level given by λ
– See picture from Wikipedia of simulated Poisson noise in CCD camera.

– Discrete RV
– Probability of a given number of events occurring in a fixed interval of time/space if these events occur (i) with a known average rate and (ii) independently of the time since the last event.
– Suppose someone gets 10 emails per day on average. The Poisson distribution specifies how likely it is that the count will be $n \geq 0$ emails for a single day.

– Given: average number of captured photons (parameter) $\lambda \in \mathbb{R}$ during a certain duration of imaging. This can be a non-integer.
– Probability of $k \in \mathbb{Z}^+$ events occurring is $P(X = k; \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$
– $\lambda = E[X] = \text{Var}(X)$
– Note: Mode of the Poisson PDF is either $\lfloor \lambda \rfloor$ or $\lceil \lambda \rceil - 1$. Note: floor$(x) = \lfloor x \rfloor$ is the largest integer not greater than x and ceiling$(x) = \lceil x \rceil$ is the smallest integer not less than x.
– Signal-to-noise ratio = SNR$[X] = E[X]/\text{SD}(X) = \sqrt{\lambda}$. SNR increases with increasing λ.

– In general, SNR increases by adding independent RVs: $E[X + X] = 2E[X]$. SD$[X + X] = \sqrt{\text{Var}[X + X]} = \sqrt{2}\text{Var [X]} = \sqrt{2}\text{SD}[X]$. SNR$[X + X] = \sqrt{2}\text{SNR}[X]$

Compound Poisson PMF (i.i.d.)

– Practical X-ray CT uses polychromatic X rays (broad-beam X ray)
– PMF of a sum of i.i.d. Poisson RVs: $Y = \sum_i X_i$ where $P(X_i) = \text{Poisson}(X_i; \lambda_i)$
– May be approximated by the Gaussian (due to the Central Limit Theorem)

Rician (Rice) PDF (i.i.d.)

– Magnitude images from MRI.
– $p(x|\nu, \sigma) = \frac{x}{\sigma^2} \exp(-\frac{x^2+\nu^2}{2\sigma^2}) I_0(\frac{x\nu}{\sigma^2})$, where $I_0(z)$ is the modified Bessel function of the first kind with order zero.
– ν, σ must be non-negative.
– $\nu \neq$ mean. $\nu =$ magnitude of the complex signal
– $\sigma \neq$ standard deviation. $\sigma =$ standard deviation of the zero-mean Gaussian noise in the real and imag parts of the complex signal
– $p(x|\nu, \sigma)$ is defined only for non-negative x
– Noise level given by σ
– NOT “zero mean”. PDF mass is shifted to the right of the signal ν.
– See example distributions.
– When $\nu = 0$, $p(x|\nu = 0, \sigma)$ is a Rayleigh distribution: $p(x; \sigma) = \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2})$. Rayleigh PDF has mean $\sigma\sqrt{\pi/2}$, variance $(2 - \pi/2)\sigma^2$, and mode σ.
– When $\nu \gg \sigma$, $p(x|\nu, \sigma)$ tends to a Gaussian distribution $G(x; \nu, \sigma^2)$.

Ultrasound

– Continues to be complicated to characterize because depends on (i) device type, e.g., coherent or incoherent waves (ii) scattering density of tissue.
– Variety of distributions have been used for magnitude images. (i) Rice, (ii) Rayleigh, (iii) K distributions.

– Characterization of Reperfused Infarcted Myocardium from High-Frequency Intracardiac Ultrasound Imaging Using Homodyned K Distribution. Hao, Bruce, Pislaru, Greenleaf. IEEE Trans ultrasonics, ferroelectrics, and frequency control 2002.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01049735

– One of the currently-used approximate models: $X = \mu + \sqrt{\mu}Y$, where $P(Y)$ is Gaussian with mean 0 and variance σ^2
– Approximate model because it incorrectly allows for negative magnitudes.

– Nonlocal Means-Based Speckle Filtering for Ultrasound Images. Pierrick Coupe, Pierre Hellier, Charles Kervrann, and Christian Barillot. IEEE TIP 2009.
URL: http://www.irisa.fr/vista/Papers/2009_IP_Coupe.pdf
– Based on Anisotropic Diffusion of Ultrasound Constrained by Speckly Noise Model. Karl Krissian, Kirby Vosburgh, Ron Kikinis, Carl-Fredrik Westin. Tech Report 2004.
URL: http://lmi.bwh.harvard.edu/papers/pdfs/2004/krissianLMI0004-04.pdf
– Also based on An Adaptive Speckle Suppression Filter for Medical Ultrasonic Imaging. Mustafa Karaman, M. Alper Kutay, Gozde Bozdagi. IEEE TMI 1995.
URL: http://yoksis.bilkent.edu.tr/doi_getpdf/articles/10.1109-42.387710.pdf

– So, $P(\sqrt{\mu}Y)$ is Gaussian with mean 0 and variance $\mu\sigma^2$
– So, noisy data X has mean μ and variance $\mu\sigma^2$
– Also, $P(X) = P(\mu + \sqrt{\mu}Y)$ is Gaussian with mean μ and variance $\mu\sigma^2$

– Coupe et al. replaced $\sqrt{\mu}$ by μ^γ, where γ is dependent on the ultrasound device.

11 Denoising Scanner Data

Denoising by Averaging Multiple Acquisitions

Assuming that the noise PDF is "centered" at the true signal value.
To be more precise: Assume that $E[\text{Data}|\text{TrueSignal}] = \text{TrueSignal}$
Examples: (i) Gaussian PDF, (ii) Poisson PDF, (iii) the specific ultrasound noise model $X = \mu + \sqrt{\mu}Y$

Theoretical Analysis

– Consider the noisy data to be represented by a RV X
– Consider at a certain pixel, the value of the noiseless signal value is fixed to be z

1) Gaussian PDF
– The signal z is the real (or imaginary) part of the complex scanner data
– Consider the noise level to be represented by some parameter σ
– $p(X = x|z, \sigma) := \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-z)^2}{2\sigma^2})$
– $E[X|z, \sigma] = z$

2) Poisson PDF
– The signal value z is the real and non-negative
– The noise level is related to the signal magnitude z
– $p(X = x|z) = \frac{z^x \exp(-z)}{x!}$
– $E[X|z] = z$

3) Ultrasound noise model
– $X := z + \sqrt{z}Y$, where $Y \sim G(0, \sigma^2)$
– $E[X|z, \sigma] = z$

What does the "expectation" mean ?
It means that if we performed a large number of acquisitions (i.e., random experiments) and take the average of the (noisy) data values, then the resulting value will be the (noiseless) signal value.

Practical Implications

In practice, we can't perform an infinite number experiments.
If we repeat the imaging process a finite N number of times, then how close is the average to the (noiseless) signal value ?

Sample Mean

Let $X_1, \cdots, X_N$ be N RVs that are i.i.d. with mean μ and variance σ^2.
Define the sample-mean RV, $X := (1/N) \sum_{n=1}^{N} X_n$.

We are interested in knowing the expectation and the variance of the sample mean.

Expectation of the sample mean:
$E[X] = (1/N) \sum_n E[X_n] = \mu$

Variance of the sample mean:
Var$(X) = E[X^2] - (E[X])^2 = E[X^2] - \mu^2$, where
$E[X^2] = (1/N^2) \left(\sum_n E[X_n^2] + \sum_n \sum_{m \neq n} E[X_n]E[X_m] \right) = (1/N^2) \left(N(\sigma^2 + \mu^2) + N(N-1)\mu^2 \right) = \sigma^2/N + \mu^2$
Thus, Var$(X) = \sigma^2/N$

So, consistent with the theoretical analysis, i.e., if we perform an infinite number of random experiments, then the expectation of the sample mean is μ and the variance of the sample mean is 0. This means that, as N increases, the sample-mean sequence *converges* to the (noiseless) signal value. This we already knew.

However, the sample-mean analysis tells us that the number of repeated experiments required so that the discrepancy between the average (denoised data) and the (noiseless) signal value isn't likely to exceed a specified amount.
The sample mean is an estimator of the (noiseless) signal value.

Sample Variance

Let $X_1, \cdots, X_N$ be N RVs that are i.i.d. with mean μ and variance σ^2.
The sample-mean is $\bar{x} = (1/N) \sum_{n=1}^{N} x_n$.
Define the sample variance as $\bar{\sigma}^2 := (1/N) \sum_n (x_n - \bar{x})^2$

Limitations of the Averaging Methods

– Works only when the noise PDF is centered at the true signal value
– Illustrate with diagram. What happens in case of Rician noise when the (noiseless) signal value is 0 ? In the presence of non-zero noise, the average will be greater than zero.
– More general methods are required.

Maximum-Likelihood (ML) Estimation

Likelihood function:
– A function of the parameters underlying the model, given the data
– Gives the likelihood that the data was generated from a specific model, i.e., specific parameter values
– Consider a noise model $P(X|\theta)$ with parameters θ
– Let data represented by x. The data is given / fixed
– $L(\theta|x) = P(x|\theta)$ = probability of observing the data
– Given the data, can we estimate the parameters θ ?
– The "most likely" parameters are those that maximize the likelihood function, i.e., maximize the probability of observing the data

Example: Gaussian

– Consider the denoising example
– Parameters: μ, σ
– Consider N repeated independent measurements (data) obtained from the MRI scanner: $x_1, \cdots, x_N$
– $L(\theta|x_1, \cdots, x_N) = P(x_1, \cdots, x_N|\mu, \sigma) = \Pi_n G(x_n; \mu, \sigma)$
– Optimize $\max_{\mu,\sigma} \Pi_n G(x_n; \mu, \sigma) = \max_{\mu,\sigma} \log \Pi_n G(x_n; \mu, \sigma)$. This is allowed because $\log()$ is a monotonic function.
– ML estimate for the mean μ equals the sample mean = $(1/N) \sum_n x_n$. Prove.
– ML estimate for the variance σ^2 equals $(1/N) \sum_n (x_n - \mu)^2$. Prove.

Example: Poisson
– Consider the denoising example
– Parameters: λ
– Consider N repeated independent measurements (data) obtained from the MRI scanner: $x_1, \cdots, x_N$
– Poisson PMF: $P(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$
– $L(\lambda|x_1, \cdots, x_N) = \Pi_n P(x; \lambda)$
– ML estimate for λ equals the sample mean = $(1/N) \sum_n x_n$. Prove.

Example: Rayleigh
– Consider the denoising example
– Parameters: σ
– Consider N repeated independent measurements (data) obtained from the MRI scanner: $x_1, \cdots, x_N$
– In case of MRI, we can consider the intensity in every "background" (air) pixel in the image as an independent experiment
– Rayleigh PDF: $p(x; \sigma) = \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2})$
– $L(\sigma|x_1, \cdots, x_N) = \Pi_n P(x; \sigma)$
– ML estimate for σ^2 equals $(1/2) \sum_n x_n^2/N$

Example: Rician
– Consider the denoising example
– Parameters: μ, σ^2
– Consider N repeated independent measurements (data) obtained from the MRI scanner: $x_1, \cdots, x_N$
– Rician PDF: $p(x|\nu, \sigma) = \frac{x}{\sigma^2} \exp(-\frac{x^2+\nu^2}{2\sigma^2}) I_0(\frac{x\nu}{\sigma^2})$
– $L(\nu, \sigma|x_1, \cdots, x_N) = \Pi_n P(x|\nu, \sigma)$
– Practical method: First estimate noise level σ using pixel intensities in the "background" where the signal is known to be zero. Use the ML estimate for σ for a Rayleigh PDF
– Use gradient descent algorithm on log likelihood to estimate ν. Use $\partial/\partial z I_0(z) = I_1(z)$
– Gradient for ν: $\frac{\partial}{\partial \nu} \log L(\nu, \sigma|x_1, \cdots, x_N) = -N\nu + \sum_n x_n \frac{I_1(x\nu/\sigma^2)}{I_0(x\nu/\sigma^2)}$

Limitations of ML Estimation

– What if we can't make repeated measurements ? May be costly ? Maybe time consuming ?
– We can compensate for lack of data by using "prior" information about what we want to estimate. For example, in our case, we want to estimate the noiseless image
– Now, if we have some "prior" information about the (noiseless) image, how can we use it in the denoising process ?

– More general methods are required