Quantitative Methods for Cognitive Scientists

Probability Theory and Maximum Likelihood Parameter Estimation

Emre Neftci

Department of Cognitive Sciences, UC Irvine,

May 9, 2019

Emre Neftci

eneftci@uci.edu

https://canvas.eee.uci.edu/courses/16991

Basic methods seen last week are effective for fitting the parameters of a model, but lack a statistical interpretation.

Example

In the class example $RMSE(d_1, d_2) = 1.613$ (see lecture02.pdf, slide entitled ‘Example: calculating an RMSE”).

- What does 1.613 mean? An average distance between the predicted points and the data points
- Does $RMSE = 1.613$ mean that our model is a good or bad fit? Generally, we can’t tell because RMSE does not have a statistical interpretation.

Plan

In this module, we will see another technique called **maximum likelihood estimation** that is deeply rooted in probability and statistics.

1. Review/Study Probability Theory (today and next week)
2. Understand the concept of likelihood in probability & statistics
3. Optimization techniques applied to likelihood (= Maximum Likelihood Estimation)

For further information read Farrell and Lewandowsky, 2018 chapter 4.

Samples, Outcomes and Sample Space

A strict definition of probability relies on the notion of samples, events, and outcomes

Example

Each time the croupier spins the wheel, and the ball is thrown in and settles in a slot, we obtain a new **sample**. *The* **outcome** *for a spin corresponds to the slot in which the ball came to rest, which is one possible outcome from the* **sample space** *of all possible slots.*

Events

An **event** is a sub-set of the sample space. In the roulette example an event could be:

- A single number
- Even number
- A number between 1-18

A **probability** P can be assigned to an **event**. It is a numerical value reflecting our expectation of the event. Probability follow these fundamental assumptions:

- Probabilities of events must lie between 0 and 1 (inclusive);

Note that P is a **function** that associates a numerical value between 0 and 1 to each possible outcome or event.

A **probability** P can be assigned to an **event.** It is a numerical value reflecting our expectation of the event. Probability follow these fundamental assumptions:

- Probabilities of events must lie between 0 and 1 (inclusive);
- The probabilities of all possible outcomes must sum exactly to 1;

Note that P is a **function** that associates a numerical value between 0 and 1 to each possible outcome or event.

A **probability** P can be assigned to an **event**. It is a numerical value reflecting our expectation of the event. Probability follow these fundamental assumptions:

- Probabilities of events must lie between 0 and 1 (inclusive);
- The probabilities of all possible outcomes must sum exactly to 1;
- In the case of mutually exclusive events (that is, two events that cannot both occur simultaneously, such as the ball in roulette settling on both an odd and an even number), the probability of any of the events occurring is equal to the sum of their individual probabilities.

Note that P is a **function** that associates a numerical value between 0 and 1 to each possible outcome or event.

Joint Probability and Mutual Exclusion

Joint probability gives the probability that multiple events occur simultaneously.

- For two events a and b, joint probability is denoted $P(a, b)$

Joint probability gives the probability that multiple events occur simultaneously.

- For two events a and b, joint probability is denoted $P(a, b)$
- $P(a, b)$ gives the probability that both a and b occur.

Joint probability gives the probability that multiple events occur simultaneously.

- For two events a and b, joint probability is denoted $P(a, b)$
- $P(a, b)$ gives the probability that both a and b occur.
- Joint probabilities allow us to formally define the concept of mutual exclusivity, introduced in the third property above: two events are mutually exclusive if $P(a, b) = 0$.

Joint probability gives the probability that multiple events occur simultaneously.

- For two events a and b, joint probability is denoted $P(a, b)$
- $P(a, b)$ gives the probability that both a and b occur.
- Joint probabilities allow us to formally define the concept of mutual exclusivity, introduced in the third property above: two events are mutually exclusive if $P(a, b) = 0$.

List three different events in the roulette example.

- What are the probabilities of each event?
- Are they mutually exclusive?

Sample spaces are denoted Ω, events are denoted ω.

- Roulette: $\Omega =$ all numbers between 0 and 36. If the roulette is unbiased, the probability of each outcome is $\frac{1}{37}$.

More examples 1/3

Sample spaces are denoted Ω, events are denoted ω.

- Roulette: $\Omega =$ all numbers between 0 and 36. If the roulette is unbiased, the probability of each outcome is $\frac{1}{37}$.
- Die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the die is unbiased, the probabilities for each outcome is $\frac{1}{6}$.

Sample spaces are denoted Ω, events are denoted ω.

- Roulette: $\Omega =$ all numbers between 0 and 36. If the roulette is unbiased, the probability of each outcome is $\frac{1}{37}$.
- Die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the die is unbiased, the probabilities for each outcome is $\frac{1}{6}$.
- One coin toss: $\Omega = \{H, T\}$. For an unbiased coin $P(H) = P(T) = .5$

Sample spaces are denoted Ω, events are denoted ω.

- Roulette: $\Omega =$ all numbers between 0 and 36. If the roulette is unbiased, the probability of each outcome is $\frac{1}{37}$.
- Die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the die is unbiased, the probabilities for each outcome is $\frac{1}{6}$.
- One coin toss: $\Omega = \{H, T\}$. For an unbiased coin $P(H) = P(T) = .5$
- Two coin tosses: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. For unbiased coin, each event has probability $\frac{1}{4}$

Sample spaces are denoted Ω, events are denoted ω.

- Roulette: $\Omega =$ all numbers between 0 and 36. If the roulette is unbiased, the probability of each outcome is $\frac{1}{37}$.
- Die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the die is unbiased, the probabilities for each outcome is $\frac{1}{6}$.
- One coin toss: $\Omega = \{H, T\}$. For an unbiased coin $P(H) = P(T) = .5$
- Two coin tosses: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. For unbiased coin, each event has probability $\frac{1}{4}$
- Reaction times in the sequential sampling model: $\Omega = \mathbb{R}^+$ (= all positive real numbers) .

More examples 3/3

- What is the sample space of randomly choosing a letter out of the alphabet?
- What is the sample space of two dice rolls?

The **conditional probability** of event a given event b, denoted P(a|b), is the probability of observing event a given that we have observed event b.
The joint probability is given by:

$$P(a,b) = P(a|b) \times P(b) \text{ or } P(a|b) = \frac{P(a,b)}{P(b)}$$

If a and b are independent, then $P(a|b) = P(a)$

Example

Probability that it will rain today, given that yesterday was rainy

Example

Probability that roulette outcome is number 35 given that the color is black is $\frac{1}{18}$

If two events a,b are independent if and only if:

$$P(a, b) = P(a)P(b)$$

Or equivalently

$$P(a|b) = P(a)$$

Rule of Addition

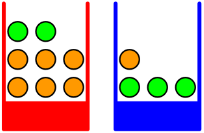
We have two events a and b, we want to know the probability that either event occurs:

$$P(a \text{ or } b) = P(a) + P(b) - P(a, b)$$

- In the roulette example: what is the probability that P(outcome is 11 or 12)?
- In the roulette example: what is the probability that P(outcome is 11 or smaller than 18)?
- In the roulette example: what is the probability that P(outcome is 11 or 12 | color is black)?

Example

Balls and Boxes

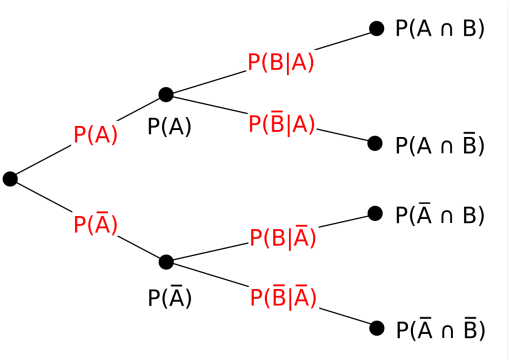


Assume we pick a ball from the red box with probability 40% and from the blue box with probability 60%.
Two example queries:

- What is the probability that a green ball is picked?
- Given that we've picked a green ball, what is the probability that we picked from the blue box (we'll solve this type of problem later)

Summary So far

- **Sample space:** the space of possible outcomes
- **Event:** A collection of outcomes
- **Probability:** a numerical value between 0 and 1, reflecting how probable an event is
- **Joint probability:** $P(a, b, c)$ is the probability that that multiple events a,b,c occur simultaneously
- **Rule of addition:** $P(a \text{ or } b) = P(a) + P(b) - P(a, b)$
- **Conditional probability:** $P(a, b) = P(a|b)P(b)$
- **Independence:** $P(a, b) = P(a)P(b)$ or $P(a|b) = P(a)$



A is an event and $\bar{A}$ is its complement (all outcomes except those in event A)

In probability theory, we mainly use **Random Variables** instead of outcomes and events. A random variable (rv) is a mapping from an outcome ω to a space, such as integers or reals.

In probability theory, we mainly use **Random Variables** instead of outcomes and events. A random variable (rv) is a mapping from an outcome ω to a space, such as integers or reals.

Example

Coin Toss

$$\omega \in \Omega = \{\text{heads}, \text{tails}\}$$

For ease of notation, we match heads and tails to the number 1 and 0, respectively. An rv matches ω to the numerical values 0, 1

$$X = \begin{cases} 1, \text{ if } \omega = heads \\ 0, \text{ if } \omega = tails \end{cases}$$

Random Variables are like Measurements

Random variables are useful tools to relate random outcomes to measurements.

Random variables are useful tools to relate random outcomes to measurements.

Example

Suppose a book is picked at the library. The sample space consists of all the books. An outcome is a book. We can define several rvs. One could be the number of pages in that book. Another might be the year of print. Yet another might be its condition (poor, acceptable, good, like new). For each rv each outcome (book) is mapped to some numerical value or category.
With rvs, One can perform queries of interest like: how many pages do books have on average? or what is the average condition for all books before printed before 1980?

Example

Other examples of rv:

- Die: X is equal to 1 if the die rolls an odd number.
- WBCD: X are features. Another example is the diagnostic Y.
- Balls and Boxes: $B = 1$ if a green ball was drawn, 0 otherwise.
- Test questions: D is the number of correct answers.

Random variables are useful tools to relate random outcomes to measurements.

Example

Roulette X is the croupier's *payout* per unit for a given event

Event	Payout (X)
Any single number	35
Corner	8
1st column	2
1 to 12	2
Red or Black	1

The **probability** that a random variable X takes value x is written $P(X = x)$. In discrete space:

$$P(X = x) = \sum_{\omega \text{ with } X(\omega)=x} p_\omega.$$

Example

Probability of an odd dice roll Assuming a fair die, $p_\omega = \frac{1}{6}$ for all $\omega \in \{1, 2, 3, 4, 5, 6\}$. Then $P(X = \text{odd}) = p_1 + p_3 + p_5 = \frac{1}{2}$ Note that in this example, X is equal to the event $\{1, 3, 5\}$.

Probabilities of an rv can also be estimated from measurements: $P(X = x)$ is equal to the fraction of trials such that X is equal to x.

Example

Coin Tosses Assume the outcome of N coin tosses is n_h heads and n_t tails.

$$P(X = 1) = \frac{n_h}{N}$$

$$P(X = 0) = \frac{n_c}{N}$$

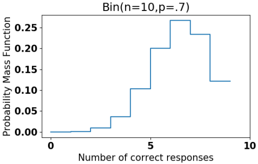
So far, we determined probabilities for each outcome/event explicitly. In many cases, we can make use of a probability function.
The function we choose depends on whether the sample spaces is **discrete** or **continuous**.

In the **discrete case**, the outcomes are one out of a number of elements. Probabilities are defined by associating a probability with each outcome. The probability function is called the **probability mass function.**

Examples with discrete sample spaces:

- Roulette
- Coin toss
- Recall a correct answer to a question

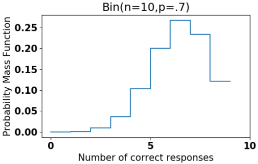
Example: the probability of n correct responses out of 10 questions.



- Note that this task is similar to the memory recall example

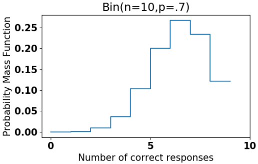
The Binomial Distribution

Example: the probability of n correct responses out of 10 questions.



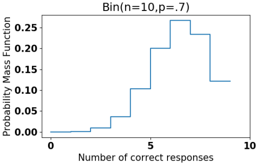
- Note that this task is similar to the memory recall example
- This function is given by the **Binomial probability function**.

Example: the probability of n correct responses out of 10 questions.



- Note that this task is similar to the memory recall example
- This function is given by the **Binomial probability function**.
- The Binomial distribution describes the number of successes in a sequence of s biased coin toss trials with probability p.

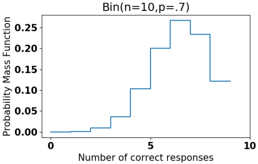
Example: the probability of n correct responses out of 10 questions.



- Note that this task is similar to the memory recall example
- This function is given by the **Binomial probability function.**
- The Binomial distribution describes the number of successes in a sequence of s biased coin toss trials with probability p.
- The probability of k correct responses out of n, $Bin(n,p) = \binom{n}{k} p^k (1-p)^{n-k}$

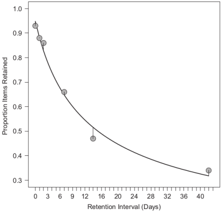
The Binomial Distribution

Example: the probability of n correct responses out of 10 questions.



- Note that this task is similar to the memory recall example
- This function is given by the **Binomial probability function**.
- The Binomial distribution describes the number of successes in a sequence of s biased coin toss trials with probability p.
- The probability of k correct responses out of n, $Bin(n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
- Parameter p is the expected proportion of positive outcomes.

Example Model: Memory recall example

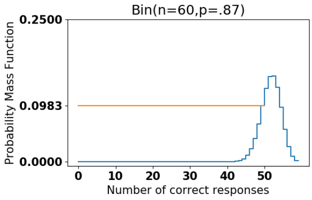


Predicted probability of recall, as a function of time:

$$r = a(bt + 1)^{-c}$$

- Measured response is a two category reponse (True or False)
- r is the proportion of correct responses.

Example Model: Memory recall example

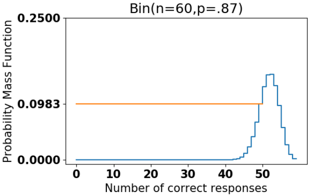


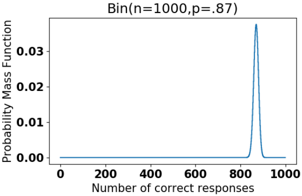
Predicted probability of recall, as a function of time:

$$r = a(bt + 1)^{-c}$$

- The number of correct reponses is distributed as $Bin(n, r)$
- $r = a(bt + 1)^{-c}$ is the parameter of the Binomial distribution.
- For 60 questions, the probability of obtaining 50 correct responses is $\binom{60}{50} r^{50}(1-r)^{10}$
- For example $r(1\text{ day}) = .87$. The probability of observing 40 correct responses is $.0983$

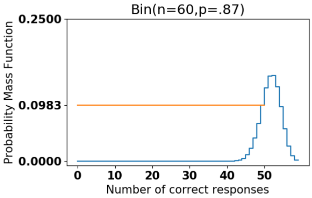
Example Model: Memory recall example

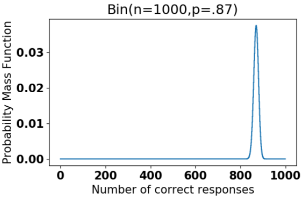




- Note that the larger n is, the more concentrated the distribution is around r

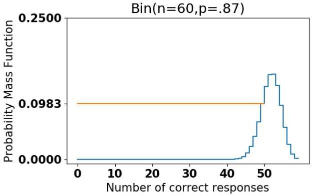
Example Model: Memory recall example

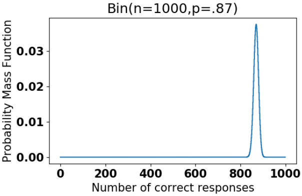




- Note that the larger n is, the more concentrated the distribution is around r
- Suppose that we don't know $r(1\text{ day})$, and our goal is to estimate it. Between the following two experiments, which one will give a more *precise* estimate of $r(1\text{ day})$?
  1. A subject answers to 60 questions
  2. A subject answers to 1000 questions

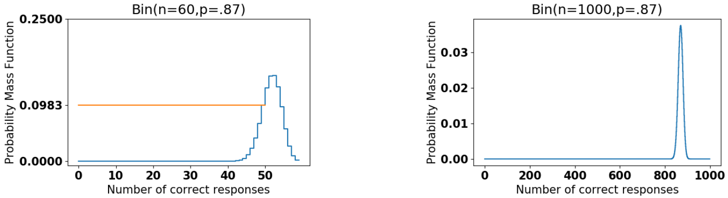
Example Model: Memory recall example





- Note that the larger n is, the more concentrated the distribution is around r
- Suppose that we don't know $r(1\text{ day})$, and our goal is to estimate it. Between the following two experiments, which one will give a more *precise* estimate of $r(1\text{ day})$?
  1. A subject answers to 60 questions
  2. A subject answers to 1000 questions

Example Model: Memory recall example



- Note that the larger n is, the more concentrated the distribution is around r
- Suppose that we don't know $r(1\text{ day})$, and our goal is to estimate it. Between the following two experiments, which one will give a more *precise* estimate of $r(1\text{ day})$?
  1. A subject answers to 60 questions
  2. A subject answers to 1000 questions
- Suppose that we observe 52 correct responses out of 60, what is the most *likely* value of r?

In the **continuous case**, outcomes are real numbers. In this case it is not possible to enumerate all probabilities. Instead we define probabilities over intervals, *e.g.* the probability that RT is between a and b. **Probability density functions** generalize this idea:

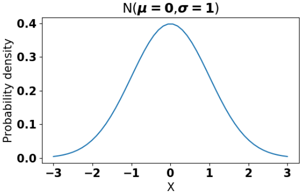
$$P(a < X < b)$$

Note: The probability of observing exactly one value in the continuous case is zero.

Examples:

- Nudges in the sequential sampling model
- Reaction times (sequential sampling, skill acquistion)
- BOLD signals in fMRI images

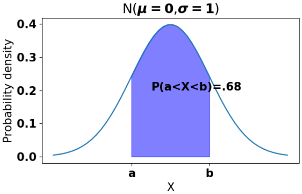
The Gaussian Distribution



- The **Gaussian** or **Normal** distribution is important due to the fact that the distribution of sums of independent variables tend to a Gaussian distribution (= Central Limit Theorem).
- The probability density function (pdf) of the Gaussian distribution is:

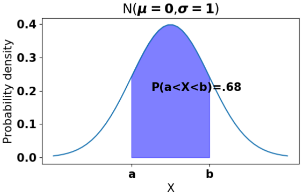
$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cumulative Density Function



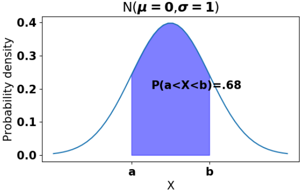
- The probability of observing X between two values is the area under the pdf between these two values.

Cumulative Density Function



- The probability of observing X between two values is the area under the pdf between these two values.
- $P(a < X < b)$ is equal to the *area* under the probability density function between a and b

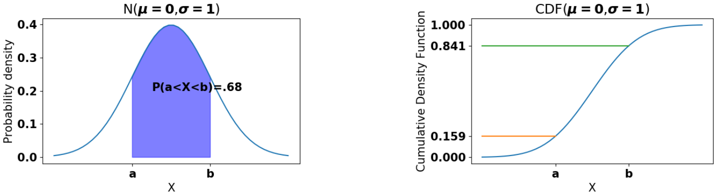
Cumulative Density Function



- The probability of observing X between two values is the area under the pdf between these two values.
- $P(a < X < b)$ is equal to the *area* under the probability density function between a and b
- This area is equal to the *integral* of the function

$$P(a < x < b) = \int_a^b P(x)\mathrm{d}x$$

Cumulative Density Function



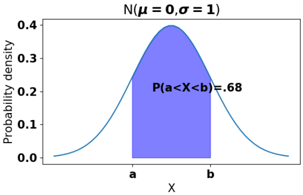
- It is more common to use the cumulative distribution function (CDF) defined as:

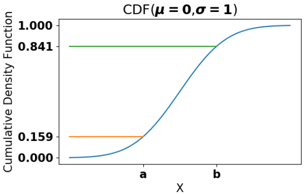
$$F(x < b) = \int_{-\infty}^{b} f(x)\mathrm{d}x$$

- Following the properties of the integral, probabilities of x can then be calculated as

$$P(a < x < b) = F(x < b) - F(x < a)$$

Cumulative Density Function

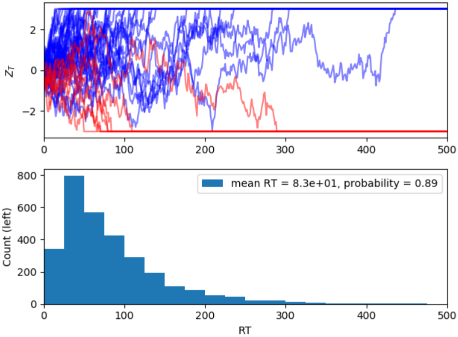




- Take home message: probabilities over continuous rvs are calculated using a difference of CDFs. CDFs for common distributions are available in most programming environments.

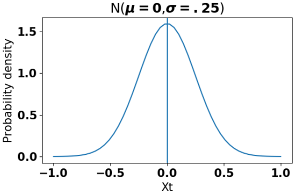
$$P(a < x < b) = F(x < b) - F(x < a)$$

Recall from the first week: 1000 trials of random walk model with $X_t \sim N(\mu_D = 0, \sigma_D = 0.25)$, starting at zero ($X_0 = 0$)



- Assume top boundary = left, bottom boundary = right
- Bottom: Recorded RTs

Example: Nudges in the Sequential Sampling Model

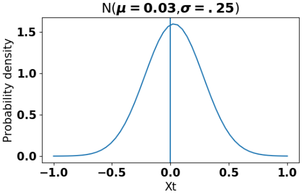


- Recall that the "nudges" in the sequential sampling model were sampled from a Gaussian distribution:

$$X_t \sim N(\mu_D = 0, \sigma = 0.25)$$

- $\sim$ means "is distributed as"
- $P(-\sigma < X_t < \sigma) = 0.68$ means that 68% of the nudges are going to be between $-.25$ and $.25$.

Example: Nudges in the Sequential Sampling Model



- When a majority of dots moved to the right, the "nudges" were sampled from a Gaussian distributions with $\mu_D > 0$:

$$X_t \sim N(\mu_D = 0.03, \sigma = 0.25)$$

- $P(-\sigma + \mu_D < X_t < \sigma + \mu_D) = 0.68$ means that 68% of the nudges are going to be between $-.22$ and $.28$.

Summary

- Random variables (rv) are functions (mappings) from sample space to numbers $P(X = x) = \sum_{\omega \text{ with } X(\omega)=x} p_\omega$
- Rvs can be discrete or continuous. The distribution is called the probability mass function in the discrete case, and probability density function in the continuous case.
- Discrete rv example: Binomial distribution $Bin(n,p) = \binom{n}{k} p^k (1-p)^{n-k}$
- Continuous rv example: Gaussian distribution $N(\mu, \sigma)$
- In the continuous case, probabilities are non-zero for intervals only, for example $P(a < X < b)$ and is equal to the difference of cumulative density function.

Expectation and Sample Mean (Discrete)

Distribution functions can be characterized with expectations and varances.
The **Expectation** of a discrete random variable X characterizes its “average value”. Mathematically it is:

$$\mathbb{E}(X) = \sum_x x P(X = x)$$

The sum runs over all possible values that X can take. The expectation can be *estimated* them from samples: **Sample mean**

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Where N is the total number of samples and x_i are samples. The larger the number of samples, the closer the sample mean μ gets to the expectation $\mathbb{E}(X)$ (= the law of large numbers).

Expectation Properties

The expectation has the properties:

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, where X and Y are two random variables.
- $\mathbb{E}(aX) = a\mathbb{E}(X)$, where a is a scalar

Expectation Properties

The expectation has the properties:

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, where X and Y are two random variables.
- $\mathbb{E}(aX) = a\mathbb{E}(X)$, where a is a scalar

Assume $X_t \sim N(0.02, .3)$

- What is the expectation of $X_t + 3$
- What is the expectation of $a * X_t$

The expectation has the properties:

- $\mathbb{E}(X + Y)$ = $\mathbb{E}(X) + \mathbb{E}(Y)$, where X and Y are two random variables.
- $\mathbb{E}(aX) = a\mathbb{E}(X)$, where a is a scalar

Assume $X_t \sim N(0.02, .3)$

- What is the expectation of $X_t + 3$
- What is the expectation of $a * X_t$
- In the sequential samples, recall that $Z_T = \sum_{t=0}^{T} X_t$.
- What is $\mathbb{E}(Z_T)$?

Expectation Properties

The expectation has the properties:

- $\mathbb{E}(X+Y) = \mathbb{E}(X)+\mathbb{E}(Y)$, where X and Y are two random variables.
- $\mathbb{E}(aX)=a\mathbb{E}(X)$, where a is a scalar

Assume $X_t \sim N(0.02, .3)$

- What is the expectation of X_t+3
- What is the expectation of $a*X_t$
- In the sequential samples, recall that $Z_T=\sum_{t=0}^{T} X_t$.
- What is $\mathbb{E}(Z_T)$?
- What is the expectation of Z_T+3

Example 1, binomially distributed random variable

- What is the expectation of a random variable distributed as $Bin(60, p = .87)$?
- What is the sample mean of $55, 59, 52, 51, 48, 53, 56, 50, 50, 49$

Example 1, binomially distributed random variable

- What is the expectation of a random variable distributed as $Bin(60, p = .87)$?
- What is the sample mean of $55, 59, 52, 51, 48, 53, 56, 50, 50, 49$

- What is the expectation of a random variable distributed as $Bin(60, p = .87)$?
- What is the sample mean of $55, 59, 52, 51, 48, 53, 56, 50, 50, 49$
- Note that samples are not required for computing the expectation. It is a property of the distribution. Because of this, expectations are looked up in a table/reference (=wikipedia).

Example 1, binomially distributed random variable

- What is the expectation of a random variable distributed as $Bin(60, p = .87)$?
- What is the sample mean of $55, 59, 52, 51, 48, 53, 56, 50, 50, 49$
- Note that samples are not required for computing the expectation. It is a property of the distribution. Because of this, expectations are looked up in a table/reference (=wikipedia).
- Note that the distribution function is not required for computing the sample mean.

Example 2, Gaussian distributed random variable

The Expectation of continuous random variables is computed in a similar fashion but using integrals (harder).

- What is the expectation of a random variables X_t distributed as $N(.3, 2)$?

Example 2, Gaussian distributed random variable

The Expectation of continuous random variables is computed in a similar fashion but using integrals (harder).

- What is the expectation of a random variables X_t distributed as $N(.3, 2)$?
- The Gaussian is a special case where its parameter μ is equal to its expectation.

Example 2, Gaussian distributed random variable

The Expectation of continuous random variables is computed in a similar fashion but using integrals (harder).

- What is the expectation of a random variables X_t distributed as $N(.3, 2)$?
- The Gaussian is a special case where its parameter μ is equal to its expectation.
- What is the sample mean of $X_0 = 0.0$, $X_1 = 1.1$, $X_2 = -0.8$, $X_3 = 0.5$, $X_4 = 1.9$, $X_5 = -1.4$, $X_6 = -0.8$, $X_7 = -1.7$, $X_8 = 0.3$, $X_9 = -0.3$.
- The calculation of the sample mean is the same as with discrete random variables

The **Variance** of a discrete random variable X characterizes its "spread" around its Expectation:

$$\mathrm{Var}(X) = \sum_x (x - \mathbb{E}(X))^2 P(X = x)$$

Similarly to sample mean, we can *estimate* the variance:
Sample variance

$$\mathrm{Var}(X) \cong s = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

Examples: Calculating Variances and Sample Variances

- Find the variance of $Bin(n, p)$.
- What is the variance of $Bin(1, .5)$ (=Coin Toss)
- What is the sample variance of (1, 0, 1 , 1 ,1, 0, 1)
- Find the variance of $N(\mu, \sigma)$.
- What is the sample variance of $X_0 = 0.0$, $X_1 = 0.3$, $X_2 = 0.2$, $X_3 = 0.3$, $X_4 = 0.4$, $X_5 = 0.2$, $X_6 = 0.2$, $X_7 = 0.2$, $X_8 = 0.3$, $X_9 = 0.3$?

Gaussian distribution: Some properties

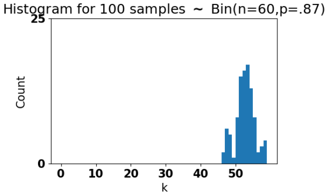
$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

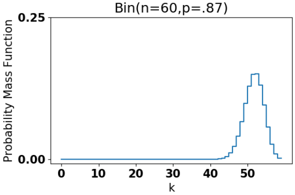
- If $X \sim N(\mu, \sigma)$, then $X + a \sim N(\mu + a, \sigma)$
- If $aX \sim N(\mu, \sigma)$, then $X \sim N(a\mu, a\sigma)$
- The probability of the pdf $N(\mu, \sigma)$ is highest at $x = \mu$
- **Central Limit Theorem:** Suppose X_1, X_2, ... , X_N are N **independent, identically distributed (iid)** random variables with mean μ and variance σ. Then the random variable

  $$Z_t = \frac{1}{N} \sum_{i=0}^{N} X_t$$

  tends towards $N(\mu, \sigma)$. Note that this is true even when X_t are not themselves Gaussian distributed!

- A **histogram** is a representation of the distribution of a continuous or discrete random variable. A histogram counts the number of observations that fall into n bins. The following is the histogram for 100 samples distributed as Bin(60, .87).





- As the number of samples increase, the shape of a histogram will increasingly resemble the probablity distribution.

Likelihood of data

- Assuming $X_1 \sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$?

Likelihood of data

- Assuming $X_1 \sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$?
- Assuming X_2 is also $\sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$ and $X_2 = 56$?

Likelihood of data

- Assuming $X_1 \sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$?
- Assuming X_2 is also $\sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$ and $X_2 = 56$?
- We observed $X_1 = 52$ and $X_2 = 56$, how likely is it that p is .87?

Likelihood of data

- Assuming $X_1 \sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$?
- Assuming X_2 is also $\sim Bin(60, .87)$, how probable is it to observe $X_1 = 52$ and $X_2 = 56$?
- We observed $X_1 = 52$ and $X_2 = 56$, how likely is it that p is .87?
- What is the most likely value of p, given that we observed $X_1 = 52$ and $X_2 = 56$

The last step is an example of a Maximum Likelihood Estimate.

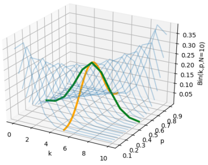
We are now ready to define the maximum Likelihood estimate (MLE). We need two elements:

- The **Joint probability** of two or more random variables is $P(X_1, X_2, ...)$.
- $P(X_1, X_2, ...|p)$ is the **Likelihood function**, describing how likely it is to observe the data $X_1, X_2, ...$ given the parameter is equal to p.

Maximum likelihood is the value of p that maximizes $P(X_1, X_2, ...)$

$$MLE(p) = \max_p P(X_1, X_2, ...|p).$$

The Likelihood function takes the same form as the probability function, but it is used for the purposes of estimating parameters.



Probability Estimation:

- The data determines the value on the p axis
- The probability estimation consists in finding the probability function (example, green curve)

Likelihood:

- The data determines the value on the k axis
- The likelihood consists in evaluating the parameter p (orange curve)

Recall that if two random variables are independent, then:

$$P(X_1, X_2) = P(X_1)P(X_2)$$

So for for N variables:

$$P(X_1, X_2, ..., X_N) = P(X_1)P(X_2), ..., P(X_N)$$

And:

$$MLE(p) = \max_p P(X_1|p)P(X_2|p)...P(X_N|p).$$

Finally, it is common to take the $\log$ of the likelihood function. This is because the maximum value does not change, and because computers can better handle the resulting numbers

$$MLE(p) = \max_p \log\left(P(X_1|p)P(X_2|p)...P(X_N|p)\right)$$
$$MLE(p) = \max_p \underbrace{\sum_{i=1}^{N} \log P(X_i|p)}_{Log-Likelihood} \quad (1)$$

Log-Likelihood as a Cost function

$$MLE(p) = \max_p \underbrace{\sum_{i=1}^{N} \log P(X_i|p)}_{Log-Likelihood} \quad (2)$$

- Log-Likelihood can be viewed as a loss function. Therefore all the basic parameter estimation techniques we've seen can be used to maximize the log-likelihood with respect to its parameters
- Many probabilistic estimation and machine learning algorithms use Maximum Likelihood estimation.
- The G^2 loss function is derived from a ratio of log-likelihoods (data likelihood divided by model likelihood).
- In some simple cases, the maximum likelihood estimates can be computed mathematically (no fitting algorithm needed)

Log Likelihood of the Binomial Distribution*

Taking the logarithm of the joint distribution becomes

$$\log P(X_1, X_2) = \log P(X_1)P(X_2) = \log P(X_1) + \log P(X_2)$$

So for for N variables:

$$\log P(X_1, X_2, ..., X_N) = \sum_{i=1}^{N} \log P(X_i)$$

For $P(X_i) \sim Bin(1, p)$ we get :

$$\begin{aligned}\log P(X_i = k) &= \log(\binom{N}{k}) + \log(p^k(1-p)^{1-k}) \\ &= \log(\binom{N}{k}) + k\log(p) + (1-k)\log((1-p))\end{aligned}$$