# Basic notation and properties

Notation:

|  |  |
| --- | --- |
| X | A random variable |
| µ or EX | The mean of X (average value) |
| σ or σ(X) | The standard deviation = square-root of mean squared error = sqrt(µ((X - µ(X))2)) |
| σ2 or V(X) | The variance of X = standard deviation squared |
| C(X, Y) | Covariance of X and Y = 1/2(µ(X - µ(X) µ(Y - µ(Y)) |
| Pr(expression) | Probability that expression is true |
|  |  |

Simple properties that follow from the definition:

1. E(aX + bY) = aE(X) + bE(Y)
2. C(X,Y) = E[(X−EX)(Y−EY)]=E[XY]−(EX)(EY)
3. C(X,X) = V(X)
4. C(X,Y) = 0 if X and Y are independent
5. V(X+Y) = V(X) + V(Y) + 2C(X, Y)
6. V(X+Y) = V(X) + V(Y) if X and Y are independent
7. V(X) = E(X2) – E(X)2
8. V(aX) = a2V(X)

# Basic theorems

## [Markov's inequality](https://en.wikipedia.org/wiki/Markov%27s_inequality)

**Pr(X ≥ a) < E(X)/a** if X has only positive values

Proof: E(X) = Pr(X < a)\*E(X|X < a) + Pr(X ≥ a) \* E(X|X ≥ a) ≥ Pr(X ≥ a) \* E(X|X ≥ a) ≥ Pr(X ≥ a) \* a

## [Chebyshev's inequality](https://en.wikipedia.org/wiki/Chebyshev%27s_inequality)

**Pr(|X – µ| > kσ) ≤ 1/k2**

Informally: The probability that a sample from X is within k standard deviations is < 1/k2.

Proof (using Markov’s inequality):

Pr(|X – µ| > kσ) = Pr((X - µ)2 > k2) ≤ E((X - µ)2)/(k2σ2) = σ2/(k2σ2) = 1/k2

This gives the following upper bounds on % of samples more than k SD’s from the mean:

**k Chebyshev’s bound Standard normal distribution (for comparison)**

1 0% 66%

2 75% 95%

3 88.8889% 99%

4 93.75%

5 96%

## [Law of large numbers](https://en.wikipedia.org/wiki/Law_of_large_numbers) (aka law of averages)

**σ(Xn) = σ(X)/√n** If X has finite variance; where Xn = (X + X + ... X)/n

Which means limn->ꝏ Xn = EX. Informally: Taking many samples of X and averaging them approaches the mean.

Proof: V(Xn) = V(ΣX/n) = 1/n2V(ΣX) = 1/n2(n\*V(X)) = V(X)/n. Per Chebyshev’s inequality, the percentage of values greater than any distance d will approach zero.

### Example: Compute n so that Pr(Xn ≥ 2σ(N)) < .05

Suppose we have an unknown distribution. How many samples do we need to average so that the probability of the sample average being more than 2 standard deviations is < 5% (the same range as standard normal with 1 sample)?

Per Chebyshev, Pr(|Xn – µ| > kσ(Xn)) ≤ 1/k2

For k=5, Pr(|Xn – µ| > 5σ(Xn)) ≤ 1/25 < .04 < .05.

So, we need to find n so that 5σ(Xn) ≤ 2σ(X). Or 25V(Xn) ≤ 4V(X)

Per law of large numbers, that’s the same as 25V(X)/n <= 4V(X)

Or 25/4 <= n. So, n = 6 would do.

## Law of sums

**σ(Xn) = √n\*σ(X)** If X has finite variance and where Xn = (X + X + ... X)

Instead of averages, if you do sums, then you multiply by sqrt(n).

Proof: Use law of large numbers + the fact that σ(aX) = aσ(X).

## [Central limit theorem](https://en.wikipedia.org/wiki/Central_limit_theorem)

Informally: Taking many samples of X and averaging them approaches a standard distribution. The tails of the distribution converge more slowly.

Statement: limn->ꝏ (√n/σ \* (Xn - µ)) approaches the [standard normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) if X has finite variance and Xn = ΣX/n (average of n samples of X)

## Binomial basics

Let B = binomial random variable with probability p of outcome 1, and (1-p) of outcome 0. Then:

* E(B) = p
* V(B) = p(1-p)
* σ(B) = √p(1-p)

### Example: If you flip a coin 100 times, you would that 66% of the time the number of heads would be in what range?

Let X = heads after 1 flip and X100 = total heads after 100 flips. Then E(X100)=50. σ(X100) = sqrt(100)\*σ(X) = 10\*1/2 = 5.

About 66% of outcomes are within one standard deviation, per central limit theorem below. So that gives a range of 45-55.

## Example: Binomial distribution

For a [binomial distribution](https://en.wikipedia.org/wiki/Binomial_distribution), X, which averages n samples from [0, 1] with probability p of a 1, what is the probability of being within k standard deviations for k=1, 2, and 3?

Looks non-trivial, per <https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval>

## Linear Regression

**Let X={x1, .., xn} and Y={y1, ..., yn} be paired n-dimensional data points. Then the equation of the line, y=a + bx, with least squared error, has:**

**b = COV(X, Y)/VAR(X)**

**E(Y) = a + bE(X)**

Proof: We need to minimize L = Σ(yi – (a + bxi))2

Take the derivative with respect to a and b and set to zero and solve

# Statistical testing

As a concrete example to ground things, suppose we have some treatment (e.g., changing how ads are displayed), some variable we care about (e.g., click through rate), and we want to perform a statistical test to see if the treatment results in improvements to the variable. How do we go about this?

## Frequentist basics

Traditional statistics would set things up like so:

* Define the ***null hypothesis***, H0. For example, that the click through rate will not change from its non-treatment value. Calculate the ***mean*** and ***standard deviation*** for this distribution.
* Define an ***alternative hypothesis***, H1. For example, that the click through rate will improve with the treatment.
* Define ***alpha***; the acceptable rate of false positives, aka type 1 errors, typically 5%
* Define ***beta***; the acceptable rate of false negatives, aka type 2 errors, typically 20%), which also defines the tests ***statistical power*** (1 – beta, typically 80%) to find true positives. You must also define the following inter-related statistics: the **effect size** you need to detect and the ***number of samples*** you will collect; increasing the effect size or number of samples will increase the statistical power. Beta can be calculated as a function of
* Run your experiment and use some statistical measure such as the ***student t-test*** calculate the ***p-value*** of the results (which measures the probability of observing results as far away from the null-hypothesis mean if the null-hypothesis was true). If the p-value is less than alpha, the result is considered significant.
* Beware – if you make multiple t-tests in an experiment (e.g., early stopping, multiple hypothesis, etc) you will have a higher percent chance of finding something with significant p-value. Solve this by pre-registering the experiments and lowering the alpha level of each experiment to get a cumulative alpha level of .05. E.g, [Pocock boundary](https://en.wikipedia.org/wiki/Pocock_boundary) is an example of a way to adjust the alpha at each look for early stopping in order to get the desired alpha level for the overall experiement.

See basics.py in this repo for some example code.

For further reading:

* <https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/>
* <https://en.wikipedia.org/wiki/Effect_size>
* <https://en.wikipedia.org/wiki/Student%27s_t-test>
* <https://en.wikipedia.org/wiki/Analysis_of_variance>

## Bayesian basics

The frequentist approach has several drawbacks. The biggest criticisms being:

1. It’s hard to understand because it’s now how people think. People want to know “what is your best guess as to how much better the treatment really is” and “what is your confidence level in the results” – rather than “what is the percent of time we’d have seen such extreme results if the null hypothesis is true” (try explaining that to an executive in a meeting).
2. It doesn’t account for prior beliefs. For example, if I flip heads three times in a row, I should have a statistical method that allows me to infer that the coin is more likely fair than always heads (unless I get a lot more evidence to the contrary).

Enter [Bayesian statistics](https://en.wikipedia.org/wiki/Bayesian_statistics).

It is based on Bayes’ theorem: P(A|B) = P(B|A) P(A) / P(B). While the proof is trivial (cross-multiply by P(B), and then both sides give P(A and B)), it’s the interpretation of this formula for Bayesian statistics is the important part:

P(A|B) = P(B|A) × P(A) ÷ P(B)

Posterior = Likelihood × Prior ÷ Evidence

Here’s what each part means and how it is calculated:

* P(A) represents your prior beliefs (e.g., that it is likely a fair coin) quantified as a distribution (how much more do you think it is fair vs mostly heads vs always heads). Using conjugate priors (below) can simplify how this is specified.
* B represents the new evidence you observed (e.g., 8 heads in 10 flips).
* P(B|A) represents the probability of observing B given your prior probability.
* P(B) represents the probability of observing B. This is in general really tricky to calculate as you have to integrate/average the probability of B over all possible distributions. Fortunately, there is no need to calculate this when using conjugate priors.
* P(A|B) represents your posterior belief given your prior belief and the new evidence.

### Using a conjugate prior

For the case of binomial trials, the [Beta function](https://en.wikipedia.org/wiki/Beta_function) is an effective way quantify prior beliefs because it’s simple, intuitive, and makes the math very efficient. The beta function takes two parameters, alpha and beta, which has the following characteristics:

1. B(1, 1), it means you have no prior beliefs and whatever you observe becomes what you believe.
2. The posterior belief will also be a beta distribution (see [conjugate prior](https://en.wikipedia.org/wiki/Conjugate_prior)).
3. The math is simple: If your prior was B(a, b) and you observe x positives and y negatives, the after applying Baye’s rule you’re your posterior belief will be B(x+a, y+b).
4. The intuition is simple: B(x, y) is equivalent to having previously observed (x-1) heads, (y-1) tails. So (1, 1) means no prior belief, (2, 2) means low conviction it’s a fair coin (equivalent to adding one head and one tail to any observation), and (10, 10) would be a stronger prior belief that it’s a fair coin.

### Likelihood ratios

Suppose one wants to just compute the relative likelihood of two distributions, H0 and H1, given some observed data, D. There are two approaches one can take:

1. Frequentist. Here we calculate the ratio Pr(D|H0)/Pr(D|H1). For a binomial distribution, after observing D = x heads in n trials, the scipy code would be binom.pmf(x, n, H0)/binom.pmf(x, n, H0), where H0 and H1 are different success probabilities.
2. Bayesian. Here we can include prior beliefs, using the formula:

P(H1|D)/P(H0|D) = P(D|H1)/P(D|H0) \* P(H1)/P(H0)

Note the difference in approaches; the first ratio is the likelihood of the observed data for the two hypotheses. The latter ratio is the likelihood of the hypothesis given the observed data and prior beliefs for the two hypotheses.

In terms of interpreting the ratio, a rule of thumb is: adjust numerator/denominator so the ratio that is >=1. For that one, a likelihood ratio of < 3 is considered inconsequential. Between 3 and 10 is moderately evidence the numerator is more likely. Over 10 is compelling evidence.

### Credible interval

Bayesian inference provides a probability distribution that reflects our current beliefs. The mean is our current belief. The 95% confidence interval is where the cumulative density function is between .025 and .975. The scipy code for these two points for a beta distribution would be beta(a, b).ppf(.025) and beta(a, b).ppf(.975).

See likelihood.py in this repo for some example code on the above.