# Basic notation and properties

Basic notation and definitions:

|  |  |
| --- | --- |
| 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)

Heuristic: if n>=20, then approximating Xn with a normal distribution works well. When n<20, one can’t assume it is normal and so must use other techniques for statistical analysis (e.g., using the student t-test for significance testing, as described later).

## 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 correlation

## Person Correlation coefficient

Def: ***ρxy*​ = COV(X, Y)/(σ(X)σ(Y))**

Properties:

* -1<=ρxy<=1
* If ρxy = 1, then X and Y are perfectly linearly correlated with positive slope
* If ρxy = -1, then X and Y are perfectly linearly correlated with negative slope
* If ρxy = 0, there is no linear correlation what-so-ever

## 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) = ρxy \* σ(Y)/σ(X)**
* **E(Y) = a + bE(X)**

Proof: We need to minimize L = Σ(yi – (a + bxi))2. Set the derivative with respect to a and b to zero and solve the two equations for a and b.

## Residuals

Residual = different between predicted vs actual value (- yi)

Sum of squared residuals, SSres = Σ(- yi)2

Total sum of squares, SStot = Σ(- E(y))2

**R2 = 1 – SSres/SStot** = Percent of variation in the y-values explained by the regression line (so 1 - R2 is the percent of variation due to the residuals). 1 = perfect fit. 0 = residuals account for all the error.

# Error estimation

Given the average of n observations of a random variable X, we use the sample mean as an estimate of the population mean. But how do we quantify how accurate that estimate is? It turns out when creating estimates for a large population (e.g., average height of someone in India, or average voting preference of someone in the US), we can get a pretty good estimate with a moderate number of samples such as 100. In subsequent sections we’ll make this more precise using the following terminology:

1. X = random variable (often with unknown distribution)
2. E = population mean (often unknown)
3. σ = population standard deviation (often unknown)
4. Xn = n samples of X computed during an experiment
5. E(Xn) = sample mean
6. σ(Xn) = sample standard deviation
7. SE = σ/√n = standard error = the standard deviation of all possible n-sample averages of X (often unknown)

## SE estimate (aka bootstrap principle)

For large values of n (>20), we can estimate σ as σ(Xn). So:

**SE ~ σ(Xn)/√n**

For smaller n, this will typically underestimate the true SE.

## Confidence interval

Motivation: After conducting a poll, people want to know not just the average results, but the error range estimate in our results. Intuitively, one knows that the error estimate would be lower if more people are polled. But how do we quantify this?

***A 95% confidence interval*** (often abbreviated as just a confidence interval) is an interval computed from the experiment samples in such a manner that 95% of experiments will contain the true value of the statistic. For a normal distribution X or for large values of n, this would be:

**E(Xn) +- 2σ(Xn)/√n**

Similarly, a 99% confidence interval would be E(Xn) +- 3σ(Xn). The scaler is often called the ***z-value***, which is desired standard deviations for a normalized version of the metric (Xn – E(Xn))/σ(Xn)). The z-value has the same meaning for any normal distribution.

Proof: If X is normal or n is large (and the central limit theorem applies), then 95% of the time:

E – 2\*SE <= E(Xn) <= E + 2\*SE

Per the bootstrap principle, this can be approximated as:

E – 2\* σ(Xn)/√n <= E(Xn) <= E + 2\* σ(Xn)/√n

Solving for E gives the desired result.

***Example***: 20% of 1000 likely voters approve of the way the president is doing his job. What is the confidence interval for presidential approval?

The standard error for the sample is σ(X)/√1000 ~ √ ((.2)(.8))/ √1000 (by the bootstrap principal) = .4/31.6 = .01265

Using normal approximation per central limit theorem, we get a 95% confidence interval of .2 +- 2\*.01265 = .2 +-.025 = [.175, .225] = [17.5%, 22.5%]

## Significance testing

Motivation: Suppose we wish to understand if a coin is fair or not, given the number of heads flipped. Or if a new type of ad results in a higher click-through rate. In general, there is always a chance the observation was due to luck, rather than an actual effect. Significance testing is a mechanism to quantify the results. For the next few sections, we’ll use the following terminology:

1. H0 = null hypothesis = default assumption that there is no effect (e.g., coin is fair, ad CTR doesn’t change).
2. H1 = alternative hypothesis = there is an effect (e.g., coin flips more heads, ad CTR is higher for the new type of ad).
3. E(H0) = population mean if H0 is true
4. σ(H0) = standard deviation if H0 is true

## z-test, p-value, and z-value

The ***z-value*** (aka z-score aka z-statistic) measures the standard deviations between the sample mean and the expected value under the null hypothesis. In particular:

**z = (E(Xn) – E(H0))/σ(H0)**

If σ(H0) is unknown, then for large n or if H0 is normal we can substitute σ(Xn) per SE estimation.

The ***p-value*** is the probability that the z statistic would be as large as it is, given the null hypothesis.

For ***z-test*** checks if the p-value is below a given ***alpha*** threshold, typically 5%.

* If it is, the result is said to be ***statistically significant***. That means we claim it as evidence that H0 is incorrect, and we state that such a claim would be correct at least 95% of the time (the rest of the time it would be a false positive, aka type 1 error)
* If it is not, the result is said to be ***not statistically significant***. That means we do not claim it as evidence that H0 is incorrect, and we live with the fact that this might be a type 2 error (false negative)

z-tests can be one-sided or two sided, depending on if we need to check for H1 being on one side or both sides of H1. For example, checking if an ad treatment results in better CRT would be a 1-sided test (as we only want to check if things are better), so we’d need a z-value slightly higher than 2. While checking if a coin is not fair would be a 2-sided test (so need a z-value between -1.96 and 1.96).

### t-test

As mentioned previously, if n < 20, then the approximation of SE by using the sample mean tends to give too small of an error estimate. This is overcome by using the t-test, which is the exact same test as above, except instead of approximating σ(H0) with σ(Xn), we approximate it with:

σ(Xn) \* sqrt(n)/sqrt(n-1)

In other words, when computing the variance, we sum the squares and divide by n-1 instead of n.

## Comparing 2 population samples: 2-sample z-test (or two sample t-test)

In the previous sections, we sampled from one population. However, we sometimes wish to sample multiple populations which may or may not be different. For example, we want to check if a drug treatment or placebo treatment results in better outcomes. If the two populations are represented by random variables X and Y, then we use the z-test on a new random variable V = X – Y. In this case:

**z = (observed difference – expected difference)/SE for the differences**

Where:

Observed difference = E(X) – E(Y)

Expected difference under the null hypothesis (typically zero)

SE for the differences = sqrt (Var(X) + Var(Y))

The two-sample t-test is the exact same thing, but using the t version of the variances (divide by n-1 instead of n), and is better for small n/non-normal distributions.

## Comparing n population samples: ANOVA

ANOVA stands for analysis of variance. It is a direct generalization of the two-sample t-test to n samples (the produce the same value if n=2), although it’s quite a bit more complicated to understand.

Suppose we have k populations we have sampled from; p1, .., pk. Suppose we have a total of n samples, of which si samples were taken from population i. Let E be the overall sample average, and E(pi) be the average of the samples within pi.

**MSS =** **mean sum of squares** = 1/(n-1) \* Σy(y – E)2. It defines the total average variance. Note it has n-1 degrees of freedom.

**MSW =** **mean sum of squares within** each population = 1/(n-k) Σy (y-Ei), where Ei is the mean of y’s group. It defines the part of the variance within each population. It has n-k degrees of freedom.

**MSB = mean sum of squares between** populations = 1/(k-1)\* Σy (E-Ei). It defines the part of the variance between populations. k-1 degrees of freedom.

Note: MSS = MSW + MSB. The degrees of freedom also sum.

**F score = MSB/MSW.**

If F >> 1, then the variation is mostly due to the differences between means, which suggests rejecting the null hypothesis. One can compute the r value by plugging in the F score against the F distribution that has the numerator and denominators degrees of freedom.

Assumptions for the test to be valid: the populations must be approximately normally distributed, have the same variance, and the samples must be independent; the same assumptions as required for a 2-sample t-test. A box plot of each group is a good way to visualize if this assumption is approximately true.

Note: ANOVA gives the same result as a 2-sample t-test if k = 2.

## Comparing paired population samples: paired difference test

Sometimes we wish to test for differences across paired values. For example, are husbands typically older than their wives? Or are children typically tall if their parents are tall? In this case, averaging over the populations will miss out on the correlation between specific paired sets in the two populations.

In this case, we create a new random variable V made up of specific matched pairs from X and Y, and do the z-test on those paired differences. So:

**z = (observed difference – expected difference)/SE for the differences**

Where:

Observed difference = E(differences of paired values)

Expected difference under the null hypothesis (typically zero)

SE for the differences, approximated from the sample differences

## Categorical fit: Chi-Square test

What if the random variable X produces categorical (rather than a numeric) value? E.g., a person’s race, or religion, or the color of an M&M? There are a number of tests, called chi-squared tests, that allow us to perform statistical inference over such categorical data.

First a definition: **the chi-square distribution with n degrees of freedom is the sum of squares of n standard normal distributions**. As n becomes large, this approaches a normal distribution. Here’s what it looks like for smaller values of n:

A picture containing text, screenshot, line, plot

Description automatically generated

Python and other software packages allow easy probability lookup of these distributions.

The three main chi-square tests, goodness of fit, homogeneity, and independence, are covered next.

### Goodness of fit

Suppose we want to check how well categorical variable X fits the expected value for number of items in each category. For example, if the percentage of various demographics (race, religion, voting preference, etc) in a new poll match a previous value.

We first compute the chi-square value as the sum of squares of these variables:

**chi-square =** **Σcat(observed – expected)2/expected**

In the above, expected = expected count for each category under the null hypothesis, observed = observed count. Note: this method is only valid if all the expected counts are reasonably large (>= 5 for all categories). If your desired metric is percentages, just convert to counts based on the number of observations.

Next, we compute the **p value as the one-sided (right tail) probability of the computed chi-square value in the chi-square distribution with #cat-1 degrees of freedom**. The “-1” is because the last category is not independent – as each observation must belong to at least one category.

***XXX – why does this work?*** The chi-square formula is not a sum of standard normal distributions because the denominator is not the variance of each variable, yet we compute the p value by looking up against a chi-square distribution that is a sum of standard normal distributions.

### Homogeneity

This is basically the same test, but across populations. For example, comparing voting preferences across states. Or other demographics across counties. Or percent of M&M colors across different types of M&Ms (peanut, chocolate, etc). We perform a similar calculation:

**chi-square = Σpop + cat(observed – expected)2/expected**

The main difference being we sum across a matrix of population and category counts, and the degrees of freedom are (# populations - 1)(# categories – 1).

### Independence

In this case we want to test if two variables are independent. For example, gender vs political party. In his case we create an nxm table of observation counts (e.g, 3x5 for 3 genders and 5 political parties) and perform the same test we did for homogeneity.

## Multiple comparisons

Running many tests will produce many false positives (type 1 errors) – according to the alpha level selected (e.g., using alpha = 5% will lead to 5% false positives). How can this be controlled? There are three common techniques:

* **Bonferroni correction**: If there are m tests, multiply the p value by m. The downside is it introduces more type 2 errors (false negatives), so only works well if m is small.
* **Benjamini-Hochberg procedure:** Control the **False discovery proportion (FDP =** # false positives/#positives) by:
  1. Defining the alpha value for false positives, say alpha = 5%
  2. Sorting the p values of all m experiments from smallest to largest
  3. Find the smallest k such that pk < alpha \* k/m
  4. Declare discoveries for just the first k tests
* **Validation set**: Hold out a validation set and run the test on that if the test on the non-validation data shows significance. Experiments are significant if both tests have p < alpha.

# Bayesian 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 recap

Traditional statistics would set things up like so:

* Define the ***null hypothesis***, H0. For example, 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.