The statistical idea of causation is truly described as follows: X causes Y only if, every time X occurs, Y occurs.

TRUE

**FALSE**

The statistical idea of causation is truly as described as follows: if X causes Y, then Y only occurs if X occurs.

TRUE

**FALSE**

The statistical idea of causation entails which of the following claims:

**We can infer causation with real confidence only by intervening and performing experiments.** 

We can infer causation with confidence by making at least 1000 observations in good perceptual conditions.

We can infer causation from correlation if we think about it in the abstract, and it seems to make sense.

We can infer causation with confidence if we observe the causal power of the cause with an electron microscope.

Why do proper medical trials require controls?

**There is no other way to justify a causal claim. A control group is a group not given the treatment, and without it, our claim that the observed change was due to that treatment is unjustified.**

Without controls, we won’t get funding for the study.

There is no other way to justify a causal claim. Without a control group, we have no control over who is given the treatment and who isn’t.

There is no other way to justify a causal claim. A control group is just a randomly selected group, and without random selection, we can’t reliably infer causation.

What is the correct specification of the main purpose of randomized controlled trials (RCTs)?

To make the individuals in each group are treated equally.

So that the individuals in each group don’t know which group they’re in.

To ensure that the evidence of the resultant multiple studies is reviewed systematically.

**To account for any confounding variables of which we’re ignorant.**

What is adjustment and what is it for?

**The inclusion into a regression model of known confounders which are not of direct interest, but are intended to allow a more balanced comparison between groups. The hope is that estimated effects associated with explanatory variables of interest should then be closer to causal effects.**

Adjusting conditions to make those engaged in a clinical trial not know what group (treatment or control) a given individual falls into. Its purpose is to avoid bias in outcome assessments.

A way of assessing the quality of an algorithm for prediction or classification by systematically adjusting some cases to act as a test set.

Adjusting a trial to make sure participants are analyzed according to whatever intervention they were supposed to get, not what they actually received.

Which situation exemplifies Simpson’s Paradox?

A situation in which a variable x appears to be correlated with a variable y, but y appears negatively correlated with x.

A situation in which a variable x appears correlated with one variable y, but also appears correlated with a variable z that appears negatively correlated with y.

**A situation where the apparent direction of an association is reversed by adjusting for a confounding factor, requiring a complete change in the apparent lesson from the data.**

A situation in which an association that appears causal is shown to be a negative correlation.

What are the signs of ‘Direct evidence,’ in the sense specified by Jeremy Horwick and colleagues?

a. I. The size of the effect is so large that it can’t be explained by plausible confounding II. There is a plausible mechanism of action. III. The effect fits with what’s already known.

b. I. The effect is found when the study is already replicated. II. There is a plausible mechanism of action. III. The effect is found in similar but not identical studies.

**c. I. The size of the effect is so large that it can’t be explained by plausible confounding II. There is appropriate temporal and/or spatial proximity, in that cause precedes effect and effect occurs after a plausible interval III. Dose responsiveness and reversibility: the effect increases as the exposure increases.**

d. I. The effect fits with what’s known already. II. The effect is found when the study is replicated. III. The effect is found in similar, but not identical studies.

The purpose of forensic epidemiology is just and only to use evidence derived from populations to draw conclusions about correlations between individual events

TRUE

**FALSE**

What is the name for a systematic review of multiple studies?

a. A multiple analysis

**b. A meta-analysis**

c. A super-analysis

d. A final analysis

**Models & Algorithms**

In this subunit, you'll learn about the bread and butter of data science: models and algorithms. If you had to boil data science down to its essence, you'd end up with proper statical models implemented with algorithms. The material in this subunit will give you the conceptual knowledge you'll need to navigate the data science libraries in Python successfully.

Chapter Five covers the fundamental information you need to know about modeling and regression.

While working through Chapter Six, you'll learn about the distinctions between predictive analytics, supervised and unsupervised learning, and feature engineering.

### Chapter 5: Modelling Relationships Using Regression

Save



1 - 2 Hours

27 Points

Please read Chapter Five of *The Art of Statistics* (pp. 121-141). This reading will teach you about some of the most important concepts in data science: statistical modeling with linear and logistic regression. You'll delve into concepts like the least-squares fitted line, regression coefficients, and signal versus noise ratios.

Regression modeling is the staple technique of predictive statistics, and will be your go-to when you want a mathematical representation of the relationship between a set of explanatory variables and a single, quantitative response variable.

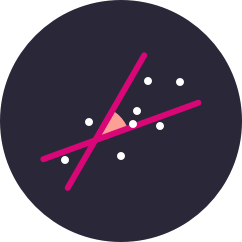
#### Chapter 6

# Regression Analysis

Linear regression is an approach for modeling the linear relationship between two variables.

A picture containing chart

Description automatically generatedOrdinary Least Squares

Correlation

A picture containing schematic

Description automatically generatedAnalysis of Variance

**Ordinary Least Squares**

The ordinary least squares (OLS) approach to regression allows us to estimate the parameters of a linear model. The goal of this method is to determine the linear model that minimizes the sum of the squared errors between the observations in a dataset and those predicted by the model. Explore the OLS method through the four infamous datasets contained in [Anscombe's Quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet).

Choose one of the quartets to investigate.

Top of Form

Bottom of Form

Drag and drop data points to explore how this affects the OLS line.

Click on a column of the regression table to learn more about this parameter.

|  |  |
| --- | --- |
|  | *n* |
|  | *x*¯ | |

|  |  |
| --- | --- |
|  | *y*¯ |

|  |  |
| --- | --- |
|  | *B*0^ |

|  |  |
| --- | --- |
|  | *B*1^ |

|  |  |
| --- | --- |
|  | *SSE* |

|  |
| --- |
|  |
| Model |  |  |  |  |  |  |

**Correlation**

Correlation is a measure of the linear relationship between two variables. It is defined for a sample as the following and takes value between +1 and −1 inclusive:

*r*=*sxysxx*−−−√*syy*−−−√

*sxy*,*sxx*,*syy*

are defined as:

*sxysxxsyy*=∑*i*=1*n*(*xi*−*x*¯)(*yi*−*y*¯)=∑*i*=1*n*(*xi*−*x*¯)2=∑*i*=1*n*(*yi*−*y*¯)2

It can also be understood as the cosine of the angle formed by the ordinary least square line determined in both variable dimensions. Explore this concept through Edgar Anderson's famous [Iris flower dataset](https://en.wikipedia.org/wiki/Iris_flower_data_set).

Check which species to investigate.

Top of Form

Bottom of Form

Click on a cell of the correlation matrix to visualize the relationship between these traits.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sepal Length | Sepal Width | Petal Length | Petal Width |
| Sepal Length |  |  |  |  |
| Sepal Width |  |  |  |  |
| Petal Length |  |  |  |  |
| Petal Width |  |  |  |  |

**Analysis of Variance**

Analysis of Variance (ANOVA) is a statistical method for testing whether groups of data have the same mean. ANOVA generalizes the t-test to two or more groups by comparing the sum of square error within and between groups.

Choose one of the following datasets to investigate.

Drag and drop data points to explore how this affects the result of the ANOVA test.

Click on a column of the ANOVA table to learn more about this parameter.

|  |  |  |
| --- | --- | --- |
|  | *SSE* | |
|  | *df* |

|  |  |
| --- | --- |
|  | *MS* |

|  |  |
| --- | --- |
|  | *F* |

|  |  |
| --- | --- |
|  | *p* |

|  |
| --- |
|  |
| Treatment |  |  |  |  |  |
| Error |  |  |  |  |  |
| Total |  |  |  |  |  |

What’s the name for the form of regression developed for proportions?

a. Multiple linear regression

b. Linear regression

c. Regression to the mean

**d. Logistic regression**

When doing linear regression, we want a line that makes the residuals:

a. Large

b. The same

**c. Small**

d. Powers of 2

Regression to the mean occurs when:

a. A non-extreme observation is followed by a high or low one, through the process of chance

**b. A high or low observation is followed by one that is less extreme, through the process of natural variation**

c. A regression model draws a straight line right through the mean of a dataset

d. A non-extreme observation is followed by one that is even less extreme

The regression coefficient expresses how weak the relationship is between an explanatory variable and an outcome in multiple regression analysis

TRUE

**FALSE**

What should we always do when we observe some residual error (that is, some data that can’t be explained by our model, and is thereby said to be due to chance variation)?

a. Iterate on the model - perhaps our next model will be better.

b. Try a different statistical model, like a classification algorithm

**c. Acknowledge that some residual error is the inevitable inability of a model to exactly represent what we observe**

d. Change the parameters to our model slightly and see what’s output the next time

When is multiple regression often used?

a. When researchers are interested in multiple explanatory variables, and there are no other variables that need to be adjusted for.

b. When researchers are interested in multiple explanatory variables, but there are other variables that need to be adjusted for to allow for imbalances.

c. When researchers are interested in one particular explanatory variable, and no other variables need to be adjusted for.

**d. When researchers are interested in one particular explanatory variable, and other variables need to be adjusted for to allow for imbalances.**

What was British statistician George Box’s famous aphorism?

a. ‘All models are useful, only some are right.’

b. ‘Only some models are right, and only some of those are useful.’

**c. ‘All models are wrong, some are useful**.’ 

D. ‘Some models are useful. Only some of those are right.’

Which of the following are all developments to regression helped by modern computing?

a. I. Having fewer explanatory variables II. Having response variables that are continuous variables III. Having relationships that are not straight lines and adapt flexibly to the pattern of data

**b. I. Having relationships that are not straight lines and adapt flexibly to the pattern of data II. Having many explanatory variables III. Having response variables that are not continuous variables**

c. I. Having explanatory variables that are numbers, rather than categories II. Having many explanatory variables III. Having relationships that are not straight lines and adapt flexibly to the pattern of data

d. I. Having relationships that are categories, rather than numbers II. Having relationships that are not straight lines and adapt flexibly to the pattern of data III. Having response variables that are continuous variables

When we do statistical modelling, what equation do we assume?

a. Observation = indeterministic model + chance

b. Observation = deterministic model + chance

c. Observation = indeterministic model + natural laws

**d. Observation = deterministic model + residual error**

What does it mean to say that a machine learning algorithm is a ‘black box?’

a. They make adequate predictions, and their internal structure is always clear to us.

**b. They make good predictions, and their internal structure is not always clear to us.**

c. They make initially poor predictions, but they learn how to make great ones on their own.

d. They make perfect predictions, and their internal structure is clear to us.

Data Science Career Track  
The Art of Statistics, Chapter 6: Algorithms, analytics, and  
prediction  
Take-Away Notes  
This chapter gives us the lay of the land when it comes to using algorithms to do statistical  
modeling.  
- Supervised learning = using an algorithm to make predictions when you know in advance what  
you’re looking for. To do it properly, we need to strictly train the algorithm on a subset of the data  
(the training set, and test it on another (the test set).  
- Unsupervised learning = using an algorithm to find patterns in data, when you don’t know in  
advance what you’re looking for.  
- Sometimes, dimensionality reduction (or reduction in the number of columns) is a necessary  
prerequisite to using a classification algorithm; this is known as feature engineering.  
A binary classification tree asks a sequence of yes/no questions, examining features in  
sequence until a classification is made. Such trees are evaluated with error matrices which  
represent accuracy (% correctly classified) sensitivity (% of those with the feature of interest  
correctly classified) and specificity (% of those without that feature correctly classified) done on  
both the training and test set.  
- Algorithms that output a probability (or a number) rather than a simple yes/no binary  
classification are often compared using Receiver Operating Characteristic (or ROC) curves ,  
which pick a threshold value above which a data-point is classified as a ‘yes’ and below which as  
a ‘no’. Different choices of value here change the sensitivity and specificity values of the  
algorithm.  
This document is authorized for use only by Charito Mascio (charrie@k5ryu.com). Copying or posting is an infringement of  
copyright.

An algorithm also needs calibration: a check that the observed frequencies of events match  
those expected by the probabilistic predictions of the outcome. E.g: if some event is given the  
probability of 0.85, we need to check that that event actually occurs roughly 85% of the time.  
The mean-squared error measures the performance of our algorithm, and is the average of the  
squares of the errors; the average mean-squared error is known as the Brier score.  
Excessively adapting a classification tree to the training data makes its predictive power decline,  
and is known as over-fitting. We overfit when we take into account all the available information  
(reducing bias) but thereby increasing variation in our estimates. Since too much bias is bad, we  
need to make a judgment on the bias/variance trade-off. We can use repetitions of  
cross-validation (removing a percentage of the training data, developing the algorithm on the  
remaining data, and then testing it on the removed data) to help counter this. We typically build  
a classification tree by deliberately overfitting, and then pruning the tree back to something  
simpler and more robust.  
This document is authorized for use only by Charito Mascio (charrie@k5ryu.com). Copying or posting is an infringement of  
copyright.

What is supervised learning?

a. Supervised learning involves the construction of algorithms that spot patterns in data when we don’t know in advance what we’re looking for.

b. Supervised learning is making algorithms that can think for themselves, provided we supervise them occasionally.

**c. Supervised learning involves the construction of classification algorithms based on cases with confirmed membership of classes.**

d. Supervised learning involves building a classificatory neural net that learns from failure.

What is unsupervised learning?

a. Unsupervised learning involves the construction of clustering algorithms when you know in advance what you’re looking for.

**b. Unsupervised learning involves the construction of algorithms that spot patterns in data when we don’t know in advance what we’re looking for.**

c. Unsupervised learning involves the use of game-theoretic concepts to build algorithms that ‘learn’ from their own errors.

d. Unsupervised learning involves building a clustering neural net that learns from failure.

It can be good practice to do feature engineering with, for example, a clustering algorithm prior to using those fields in a classification algorithm

**TRUE**

FALSE

For any problem that can be solved to some degree of accuracy with a basic classification model, there is some neural network we could construct that could solve it with a higher degree of accuracy

TRUE

**FALSE**

Suppose we have a classification tree that tries to predict a success rate for a given job interview. In this context, what is Accuracy?

**a. % correctly classified**

b. % of successful applicants correctly classified

c. % of unsuccessful applicants correctly classified

d. % incorrectly classified

With such a classification tree, what does the area under the ROC curve measure? What does it not measure?

a. It measures how good the probabilities are, but it doesn’t measure how well the algorithm splits the successful applicants from the unsuccessful ones.

b. It measures how well the algorithm is calibrated.

c. It measures the efficiency of the algorithm.

**d. It measures how well the algorithm splits the successful applicants from the unsuccessful ones, but it doesn’t measure how good the probabilities are.**

7 / 10

Under what conditions does over-fitting afflict a statistical model?

a. When the model over-adapts to the test data, rendering its training less useful, and its predictive power reduced.

b. When the model over-fits to the training data, increasing its predictive power.

**c. When it is over-adapted to the training data so that its predictive ability starts to decline.**

d. When the model is cross-validated.

What is the bias-variance trade-off?

**a. Over-fitting leads to less bias but at a cost of more uncertainty or variation in the estimates.**

b. Avoiding over-fitting leads to less bias, but at a cost of more uncertainty or variation in the estimates.

c. Cross-validation leads to less bias but at a cost of more uncertainty or variation in the estimates.

d. Avoiding cross-validation leads to less bias but at a cost of more uncertainty or variation in the estimates.

What is the phenomenon of interactions, and when does it occur?

a. Interactions occur when single explanatory variables produce the same effect when combined that they produce on their own.

**b. Interactions occur when multiple explanatory variables combine to produce an effect different from that expected from their individual contributions.**

c. Interactions occur when two or more of the explanatory variables correlate positively with one another.

d. Interactions occur when there is a negative correlation between two or more of the explanatory variables.

What makes an algorithm contain problematic implicit bias?

**a. When the algorithm is based on associations that we normally think are irrelevant to the task at hand, and that may unjustly disadvantage people as a result.**

b. When an algorithm is opaque due to its sheer complexity.

c. When an algorithm is based on limited data.

d. When the algorithm is based on real associations, that disadvantage some justifiably.

Get two extra months in the course

when you achieve 50% progress!

Your mission

1. Set a realistic weekly point goal. Strive for consistency week over week.

2. Work with your mentor to hit your targets.

3. Reach 50% progress to unlock an extra two free months in the course!

#### Chapter 4

# Frequentist Inference

Frequentist inference is the process of determining properties of an underlying distribution via the observation of data.

### Point Estimation

One of the main goals of statistics is to estimate unknown parameters. To approximate these parameters, we choose an estimator, which is simply any function of randomly sampled observations.

To illustrate this idea, we will estimate the value of *π*

by uniformly dropping samples on a square containing an inscribed circle. Notice that the value of *π* can be expressed as a ratio of areas.

*Scircle*=*πr*2*Ssquare*=4*r*2⟹*π*=4*ScircleSsquare*

We can estimate this ratio with our samples. Let *m* be the number of samples within our circle and *n* the total number of samples dropped. We define our estimator *π*^ as:

*π*^=4*mn*

It can be shown that this estimator has the desirable properties of being unbiased and consistent.

|  |
| --- |
| *m*= |

932  
*n*=

|  |  |
| --- | --- |
| 1200 | *π*^= |

|  |
| --- |
| 3.1067 |

Drop 100 Samples

Drop 1000 Samples

### Confidence Interval

In contrast to point estimators, confidence intervals estimate a parameter by specifying a range of possible values. Such an interval is associated with a confidence level, which is the probability that the procedure used to generate the interval will produce an interval containing the true parameter.

Choose a probability distribution to sample from.

Choose a sample size (*n*) and confidence level (1−*α*)

.

*n*

= 5   
1−*α* = 0.90

Start sampling to generate confidence intervals.

Contains μExcludes μ0.00.51.0

Start Sampling

This visualization was adapted from Kristoffer Magnusson's fantastic visualization of [confidence intervals](http://rpsychologist.com/d3/CI/).

### The Bootstrap

Much of frequentist inference centers on the use of "good" estimators. The precise distributions of these estimators, however, can often be difficult to derive analytically. The computational technique known as the Bootstrap provides a convenient way to estimate properties of an estimator via resampling. In this example, we resample with replacement from the empirical distribution function (which is itelf generated by sampling once from the population) in order to estimate the standard error of the sample mean.

Choose a probability distribution from which we will sample once to generate the empirical distribution function.

Choose a sample (and resampling) size (*n*)

and sample from your chosen distribution.

*n*

= 10

Sample

Resample to get an idea of the spread of the sample mean's distribution.

Resample

Resample 100 times

#### Chapter 4

# Frequentist Inference

Frequentist inference is the process of determining properties of an underlying distribution via the observation of data.

### Point Estimation

One of the main goals of statistics is to estimate unknown parameters. To approximate these parameters, we choose an estimator, which is simply any function of randomly sampled observations.

To illustrate this idea, we will estimate the value of *π*

by uniformly dropping samples on a square containing an inscribed circle. Notice that the value of *π* can be expressed as a ratio of areas.

*Scircle*=*πr*2*Ssquare*=4*r*2⟹*π*=4*ScircleSsquare*

We can estimate this ratio with our samples. Let *m* be the number of samples within our circle and *n* the total number of samples dropped. We define our estimator *π*^ as:

*π*^=4*mn*

It can be shown that this estimator has the desirable properties of being unbiased and consistent.

|  |
| --- |
| *m*= |

932  
*n*=

|  |  |
| --- | --- |
| 1200 | *π*^= |

|  |
| --- |
| 3.1067 |

Drop 100 Samples

Drop 1000 Samples

### Confidence Interval

In contrast to point estimators, confidence intervals estimate a parameter by specifying a range of possible values. Such an interval is associated with a confidence level, which is the probability that the procedure used to generate the interval will produce an interval containing the true parameter.

Choose a probability distribution to sample from.

Choose a sample size (*n*) and confidence level (1−*α*)

.

*n*

= 5   
1−*α* = 0.90

Start sampling to generate confidence intervals.

Contains μExcludes μ0.00.51.0

Start Sampling

This visualization was adapted from Kristoffer Magnusson's fantastic visualization of [confidence intervals](http://rpsychologist.com/d3/CI/).

### The Bootstrap

Much of frequentist inference centers on the use of "good" estimators. The precise distributions of these estimators, however, can often be difficult to derive analytically. The computational technique known as the Bootstrap provides a convenient way to estimate properties of an estimator via resampling. In this example, we resample with replacement from the empirical distribution function (which is itelf generated by sampling once from the population) in order to estimate the standard error of the sample mean.

Choose a probability distribution from which we will sample once to generate the empirical distribution function.

Choose a sample (and resampling) size (*n*)

and sample from your chosen distribution.

*n*

= 10

Sample

Resample to get an idea of the spread of the sample mean's distribution.

Resample

Resample 100 times

* ☰
* Chapter 1: Basic Probability

#### Chapter 1

# Basic Probability

This chapter is an introduction to the basic concepts of probability theory.

### Chance Events

Randomness is all around us. Probability theory is the mathematical framework that allows us to analyze chance events in a logically sound manner. The probability of an event is a number indicating how likely that event will occur. This number is always between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

A classic example of a probabilistic experiment is a fair coin toss, in which the two possible outcomes are heads or tails. In this case, the probability of flipping a head or a tail is 1/2. In an actual series of coin tosses, we may get more or less than exactly 50% heads. But as the number of flips increases, the long-run frequency of heads is bound to get closer and closer to 50%.

Flip the Coin

Flip 100 times

For an unfair or weighted coin, the two outcomes are not equally likely. You can change the weight or distribution of the coin by dragging the true probability bars (on the right in blue) up or down. If we assign numbers to the outcomes — say, 1 for heads, 0 for tails — then we have created the mathematical object known as a [random variable](https://seeing-theory.brown.edu/probability-distributions/index.html#section1).

### Expectation

The expectation of a random variable is a number that attempts to capture the center of that random variable's distribution. It can be interpreted as the long-run average of many independent samples from the given distribution. More precisely, it is defined as the probability-weighted sum of all possible values in the random variable's support,

E[*X*]=∑*x*∈X*xP*(*x*)

Consider the probabilistic experiment of rolling a fair die and watch as the running sample mean converges to the expectation of 3.5.

Roll the Die

Roll 100 times

Change the distribution of the different faces of the die (thus making the die biased or "unfair") by adjusting the blue bars below and observe how this changes the expectation.

### Variance

Whereas expectation provides a measure of centrality, the variance of a random variable quantifies the spread of that random variable's distribution. The variance is the average value of the squared difference between the random variable and its expectation,

Var(*X*)=E[(*X*−E[*X*])2]

Draw cards randomly from a deck of ten cards. As you continue drawing cards, observe that the running average of squared differences (in green) begins to resemble the true variance (in blue).

Draw a Card

Draw 100 times

Toggle which cards you want to include in the deck by clicking on them below.

* [Download](https://seeing-theory.brown.edu/doc/basic-probability.pdf)
* Share

#### [Next](https://seeing-theory.brown.edu/compound-probability/index.html)

[Compound Probability →](https://seeing-theory.brown.edu/compound-probability/index.html)

Logo, icon

Description automatically generatedChance Events

Icon

Description automatically generatedExpectation

Icon

Description automatically generatedVariance

Observed outcomesTrue probabilities

* ☰
* Chapter 3: Probability Distributions

#### Chapter 3

# Probability Distributions

A probability distribution specifies the relative likelihoods of all possible outcomes.

### Random Variables

Formally, a random variable is a function that assigns a real number to each outcome in the probability space. Define your own discrete random variable for the uniform probability space on the right and sample to find the empirical distribution.

Click and drag to select sections of the probability space, choose a real number value, then press "Submit."

Top of Form

Bottom of Form

|  |  |
| --- | --- |
| Color | Value |
|  | 0 |

Sample from probability space to generate the empirical distribution of your random variable.

0.00.51.0

Sample Distribution

Reset

### Discrete and Continuous

There are two major classes of probability distributions.

A discrete random variable has a finite or countable number of possible values.

If *X*

is a discrete random variable, then there exists unique nonnegative functions, *f*(*x*) and *F*(*x*)

, such that the following are true:

*P*(*X*=*x*)*P*(*X*<*x*)=*f*(*x*)=*F*(*x*)

Choose one of the following major discrete distributions to visualize. The probability mass function *f*(*x*)

is shown in yellow and the cumulative distribution function *F*(*x*)

in orange (controlled by the slider).

### Central Limit Theorem

The Central Limit Theorem (CLT) states that the sample mean of a sufficiently large number of i.i.d. random variables is approximately normally distributed. The larger the sample, the better the approximation.

Change the parameters *α*

and *β*

to change the distribution from which to sample.

*α*

= 1.00   
*β* = 1.00

Choose the sample size and how many sample means should be computed (draw number), then press "Sample." Check the box to display the true distribution of the sample mean.

Sample size = 1

Draws = 1

Sample

This visualization was adapted from Philipp Plewa's fantastic visualization of the [central limit theorem](https://bl.ocks.org/pmplewa/4120c2929ede7e336d9b55b760e496f6).

* [Download](https://seeing-theory.brown.edu/doc/probability-distributions.pdf)
* Share

#### [Next](https://seeing-theory.brown.edu/frequentist-inference/index.html)

[Frequentist Inference →](https://seeing-theory.brown.edu/frequentist-inference/index.html)

A picture containing shape

Description automatically generatedRandom Variable

Logo, icon

Description automatically generatedDiscrete and Continuous

Icon

Description automatically generatedCentral Limit Theorem

What is the main purpose of the Null Hypothesis being the default hypothesis?

**a. To protect us against false discoveries**

b. To speed up discoveries

c. To make our discoveries more certain

d. To make our discoveries more important

A permutation test is a form of hypothesis test in which the distribution of the test statistic under the null hypothesis is obtained by permuting the labels of the data, rather than through a detailed statistical model for the random variables

**TRUE**

FALSE

What is the p-value of an observation?

a. The probability of the Null Hypothesis

b. The probability of getting a result less extreme than that observation, if the Alternative hypothesis is true

**c. The probability of getting a result at least as extreme as that observation, if the Null hypothesis is true**

d. The probability of the observation, if the Null Hypothesis is true

What’s the difference between a one-sided and two-sided test?

a. A one-sided hypothesis test would be appropriate for a Null hypothesis that a treatment effect, say, is exactly 0, and so both positive and negative estimates would lead to the Null being rejected. A two-sided test is used when a Null hypothesis specifies that, say, the effect of a medical treatment is negative

**b. A one-sided hypothesis test is used when a Null hypothesis specifies that, say, the effect of a medical treatment is negative. A two-sided test would be appropriate for a Null hypothesis that a treatment effect, say, is exactly 0, and so both positive and negative estimates would lead to the Null being rejected.**

Why is the use of the idea of statistical significance debated?

a. Because saying that an observation is ‘statistically significant’ if it has a p-value of 0.05000000001 (assuming a significance level of 0.05) is altogether sensible

b. Because saying that an observation is ‘not statistically significant’ if it has a p-value of 0.05000000001 (assuming a significance level of 0.05) is not hugely misleading

**c. Because saying that an observation is ‘not statistically significant’ if it has a p-value of 0.05000000001 (assuming a significance level of 0.05) is hugely misleading**

d. Because saying that an observation is ‘statistically significant’ if it has a p-value of 0.05000000001 (assuming a significance level of 0.05) is hugely misleading

What is the purpose of the t-statistic?

**a. It serves to tell us whether the association between an explanatory variable and a response is statistically significant.**

b. It serves to test whether our data is normally distributed.

c. It serves to tell us which statistical test to employ.

d. It serves to tell us whether the Alternative Hypothesis is true.

What’s the difference between Type 1 and Type 2 Errors?

a. Type 1 Error = not rejecting a false Alternative hypothesis; Type 2 Error = rejecting a true Alternative hypothesis

**b. Type 1 Error = rejecting a true Null hypothesis; Type 2 Error = not rejecting a false Null hypothesis**

c. Type 1 Error = not rejecting a false Null hypothesis; Type 2 Error = rejecting a true Null hypothesis

d. Type 1 Error = rejecting a true Alternative hypothesis; Type 2 Error = not rejecting a false Alternative hypothesis

What is the particular danger always incurred by carrying out many significance tests?

a. It is hugely expensive on resources

b. It is boring for the testers

**c. As we do more tests, the chance of getting a misleading ‘significant’ result increases**

d. As we do more tests, the chance of misleadingly not getting a ‘significant’ result increases

What is ‘the size of a test’ vs ‘the power of a test’ distinction?

a. The size of a test = The Type 2 error rate of a statistical test, and the power of a test = the probability of correctly rejecting the Null hypothesis

b. The size of a test = The Type 1 error rate of a statistical test, and the power of a test = the probability of incorrectly rejecting the Null hypothesis

c. The size of a test = the probability of correctly rejecting the Null hypothesis, and the power of a test = The Type 1 error rate of a statistical test

**d. The size of a test = The Type 1 error rate of a statistical test, and the power of a test = the probability of correctly rejecting the Null hypothesis**

Which claim is true?

a. All statistically insignificant results are also practically insignificant.

b. All practically significant results are statistically significant.

c**. Not all statistically significant results are practically significant.**

d. Not all statistically significant results are statistically significant.

* ☰
* [Seeing Theory](https://seeing-theory.brown.edu/index.html)

#### Chapter 5

# Bayesian Inference

Bayesian inference techniques specify how one should update one’s beliefs upon observing data.

### Bayes' Theorem

Suppose that on your most recent visit to the doctor's office, you decide to get tested for a rare disease. If you are unlucky enough to receive a positive result, the logical next question is, "Given the test result, what is the probability that I actually have this disease?" (Medical tests are, after all, not perfectly accurate.) Bayes' Theorem tells us exactly how to compute this probability:

*P*(Disease|+)=*P*(+|Disease)*P*(Disease)*P*(+)

As the equation indicates, the posterior probability of having the disease given that the test was positive depends on the prior probability of the disease *P*(Disease)

. Think of this as the incidence of the disease in the general population. Set this probability by dragging the bars below.

P(Healthy)P(Disease)

The posterior probability also depends on the test accuracy: How often does the test correctly report a negative result for a healthy patient, and how often does it report a positive result for someone with the disease? Determine these two distributions below.

P(-|H)P(+|H)P(-|D)P(+|D)

Finally, we need to know the overall probability of a positive result. Use the buttons below to simulate running the test on a representative sample from the population.

Test one patient

Test Remaining

|  |  |
| --- | --- |
| **Negative** | **Positive** |
| 0.70 | 0.30 |

We now have everything we need to determine the posterior probability that you have the disease. The table below gives this probability among others using Bayes' Theorem.

|  |  |  |
| --- | --- | --- |
|  | **Negative** | **Positive** |
| **Healthy** | 0.96 | 0.75 |
| **Disease** | 0.04 | 0.25 |

Sort

Reset

### Likelihood Function

In statistics, the likelihood function has a very precise definition:

*L*(*θ*|*x*)=*P*(*x*|*θ*)

The concept of likelihood plays a fundamental role in both Bayesian and frequentist statistics.

Choose a sample size *n*

and sample once from your chosen distribution.

*n*

= 1

Sample

Use the purple slider on the right to visualize the likelihood function.

### Prior to Posterior

At the core of Bayesian statistics is the idea that prior beliefs should be updated as new data is acquired. Consider a possibly biased coin that comes up heads with probability *p*

. This purple slider determines the value of *p*

(which would be unknown in practice).

*p*

= 0.5

The pink sliders control the shape of the initial Beta(*α*,*β*)

prior distribution, the density function of which is also plotted in pink.

*α*

= 1

*β*

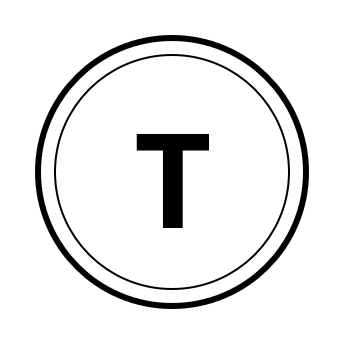
= 1

As we acquire data in the form of coin tosses, we update the posterior distribution on *p*

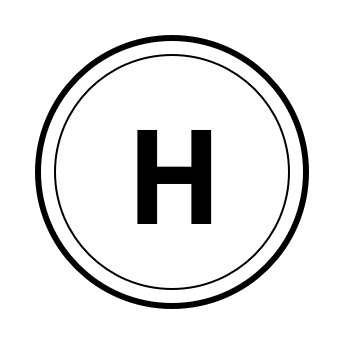
, which represents our best guess about the likely values for the bias of the coin. This updated distribution then serves as the prior for future coin tosses.

Flip the Coin

Flip 10 times



= 0



= 0

* [Download](https://seeing-theory.brown.edu/doc/bayesian-inference.pdf)
* Share

#### [Next](https://seeing-theory.brown.edu/regression-analysis/index.html)

[Regression Analysis →](https://seeing-theory.brown.edu/regression-analysis/index.html)

A picture containing text

Description automatically generatedBayes' Theorem

Icon

Description automatically generatedLikelihood Function

A picture containing text, clock

Description automatically generatedPrior to Posterior

* ☰
* Chapter 3: Probability Distributions

#### Chapter 3

# Probability Distributions

A probability distribution specifies the relative likelihoods of all possible outcomes.

### Random Variables

Formally, a random variable is a function that assigns a real number to each outcome in the probability space. Define your own discrete random variable for the uniform probability space on the right and sample to find the empirical distribution.

Click and drag to select sections of the probability space, choose a real number value, then press "Submit."

Top of Form

Bottom of Form

|  |  |
| --- | --- |
| Color | Value |
|  | 0 |

Sample from probability space to generate the empirical distribution of your random variable.

0.00.51.0

Sample Distribution

Reset

### Discrete and Continuous

There are two major classes of probability distributions.

A discrete random variable has a finite or countable number of possible values.

If *X*

is a discrete random variable, then there exists unique nonnegative functions, *f*(*x*) and *F*(*x*)

, such that the following are true:

*P*(*X*=*x*)*P*(*X*<*x*)=*f*(*x*)=*F*(*x*)

Choose one of the following major discrete distributions to visualize. The probability mass function *f*(*x*)

is shown in yellow and the cumulative distribution function *F*(*x*)

in orange (controlled by the slider).

### Central Limit Theorem

The Central Limit Theorem (CLT) states that the sample mean of a sufficiently large number of i.i.d. random variables is approximately normally distributed. The larger the sample, the better the approximation.

Change the parameters *α*

and *β*

to change the distribution from which to sample.

*α*

= 1.00   
*β* = 1.00

Choose the sample size and how many sample means should be computed (draw number), then press "Sample." Check the box to display the true distribution of the sample mean.

Sample size = 1

Draws = 1

Sample

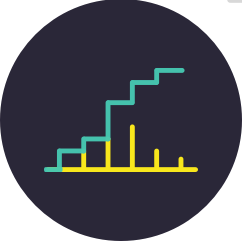
This visualization was adapted from Philipp Plewa's fantastic visualization of the [central limit theorem](https://bl.ocks.org/pmplewa/4120c2929ede7e336d9b55b760e496f6).

* [Download](https://seeing-theory.brown.edu/doc/probability-distributions.pdf)
* Share

#### [Next](https://seeing-theory.brown.edu/frequentist-inference/index.html)

[Frequentist Inference →](https://seeing-theory.brown.edu/frequentist-inference/index.html)

Random Variable

Discrete and Continuous

Central Limit Theorem