Introduction to Probability and Statistics

1 Introduction to Probability and Statistics

1.1 Why we need to study Math & Statistics?

The following tracks are main points in the area of Data Scientist. All tracks are related with Mathematics and Statistics to variable degree. If you want to be a professional Data Scientist, you need have a good command of math & statistics knowledge. Following are the responsibilities a data scientist may have in the daily work:

- The definition of data scientist
- Analytics or Core Data Science
- Maintain Dashboard or Modeling
- PM track or Machine Learning Engineer track
- Analyze product or Develop algorithms

1.2 Outline

- Introduction and Basic Probability
- Joint Probability, Independence and Conditional Probability
- Bayes Rule
- Important Distributions
- Hypothesis Testing
- T-test
- Confusion Matrix and Power Analysis

2 Introduction and Basic Probability

2.1 Introduction of Probability

Define Uncertainty: There are too many uncertain events in our life, so we need probability to define those uncertainties. For example, we watch a football game, the winner team is an uncertainty, and we can use probability to guest which team is winner.

Understand Probability: Probability is a measure of the likelihood. The best example for understanding probability is flipping a coin: There are two possible outcomes—heads or tails.

2.2 Introduction of Statistics

Define Statistics: The science of data. Statistics is a discipline which is concerned with: designing experiments and other data collection, summarizing information to aid understanding, drawing conclusions from data, and estimating the present or predicting the future.

Purpose of Statistics: In general, the purpose of statistical tests is to determine whether some hypothesis is extremely unlikely given the observed data. Then we can make some predictions.

2.3 Random Variable

Random Variable: To describe a random event and its possible outcomes.

Examples: Coin toss, Lottery, The spend of next week.

3. More Events' Relation

3.1 Joint and Conditional Probability

Joint Probability: P(A,B),

The probability of A/B happens together.

Conditional Probability: $P(A|B) = P(A,B)/P(B)$,

Given that B happens, the probability of A happens.

Example:

1. One coin flip: A: head, B: tail. $P(A|B)= 0$
2. Two coin flips: A: 1^{st} flip head, B: 2^{nd} flip tail. $P(A|B)= \frac{1}{4}$

3.2 Independence

Definition: $P(A,B) = P(A)P(B)$ or $P(A|B) = P(A)$

Extension: $P(A,B,C) = P(A)P(B)P(C)$

Question: Does pairwise independence mean independence?

Answer: Pairwise independence does not imply mutual independence. For example, we throw two dice. Let A be the event "the sum of the points is 7", B the event "die #1 came up 3",

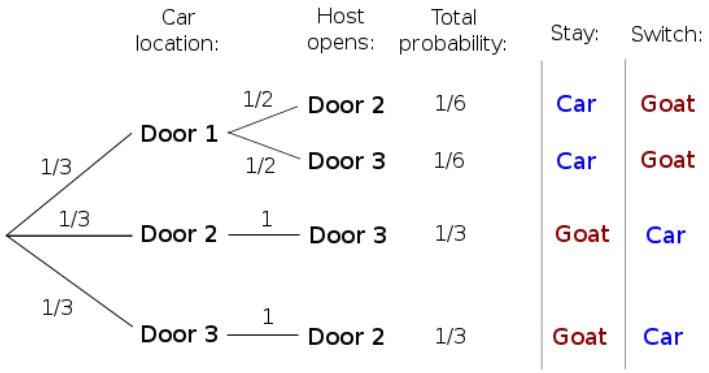
and C the event "die #2 came up 4". So P(A)=P(B)=P(C)=$\frac{1}{6}$, P(A,B)=P(B,C)=P(A,C)=P(A,B,C)=$\frac{1}{36}$, but P(A)P(B)P(C)=$\frac{1}{216}$.

4.Bayes Rule

Define: Bayes' rule describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

For proposition A and evidence B, P (A): the prior, is the initial degree of belief in A. P (A | B): the "posterior," is the degree of belief having accounted for B. B: Support (evidence)

Example: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



The Bayes Rule:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Right now we know No.3 has no Prize. So the probability of No.1 has a prize is:

$$P(Prize\ in\ No.1|No\ Prize\ in\ No.3)$$
$$= \frac{P(No\ Prize\ in\ No.3|Prize\ in\ No.1) \cdot P(Prize\ in\ No.1)}{P(No\ Prize\ in\ No.3|Prize\ in\ No.1)P(Prize\ in\ No.1) + P(No\ Prize\ in\ No.3|Prize\ in\ No.2)P(Prize\ in\ No.2)}$$
$$= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}$$

Which means that if we decide not to switch, our winning probability is only $\frac{1}{3}$, so we decide to switch.

5. Distributions

5.1 Probability Distributions

- Probability Mass Function (PMF): For discrete distributions

$$f_X(x) = P(X = x), \qquad \sum_{x \in A} f_X(x) = 1$$

- Probability Density Function (PDF): For continuous distributions

$$P(a \le X \le b) = \int_a^b f_X(x)dx$$

- Cumulative Distribution Function(CDF): Any distributions

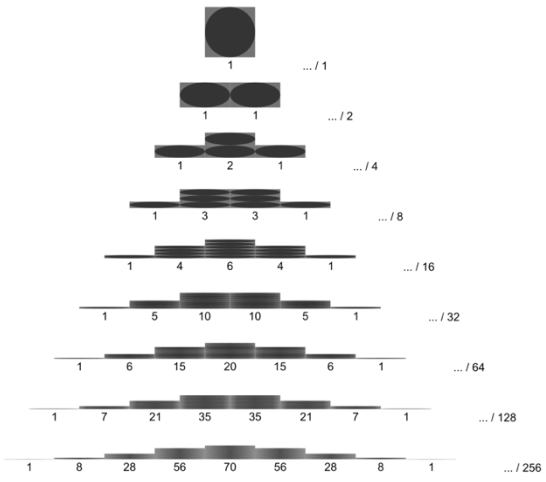
$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

5.2 Important Discrete Distributions

5.2.1 Binomial Distribution

A binomial random variable is the number of successes x in n repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution.

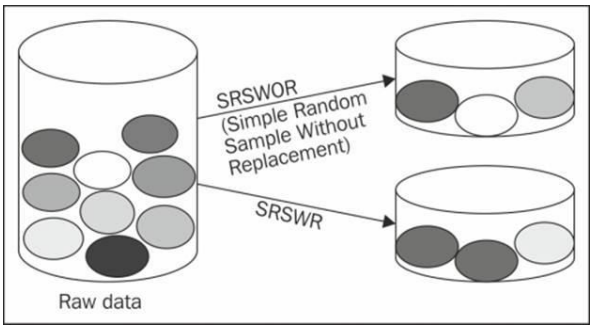
Notation	$B(n, p)$
Parameters	$n \in \mathbb{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$



5.2.2 Hypergeometric Distribution

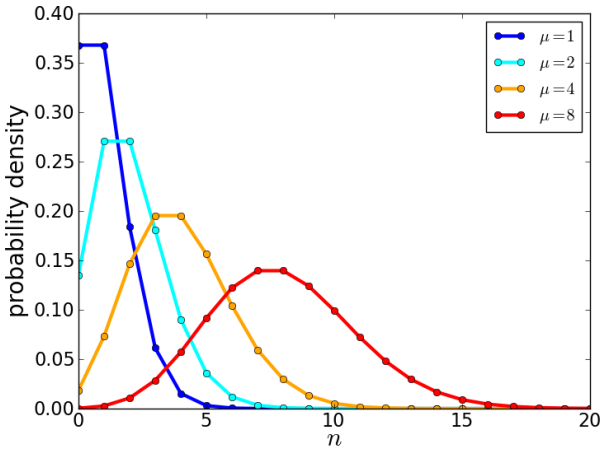
The **hypergeometric distribution** is a discrete probability distribution that describes the probability of *k* successes (random draws for which the object drawn has a specified feature) in *n* draws, without replacement, from a finite population of size *N* that contains exactly *K* objects with that feature, wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of *k* successes in *n* draws with replacement.

Parameters	$N \in \{0, 1, 2, \ldots\}$ $K \in \{0, 1, 2, \ldots, N\}$ $n \in \{0, 1, 2, \ldots, N\}$
Support	$k \in \{\max(0, n+K-N), \ldots, \min(n, K)\}$
pmf	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
CDF	$1 - \frac{\binom{n}{k+1}\binom{N-n}{K-k-1}}{\binom{N}{K}} {}_3F_2\left[\begin{matrix}1,\ k+1-K,\ k+1-n \\ k+2,\ N+k+2-K-n\end{matrix}; 1\right]$, where ${}_pF_q$ is the generalized hypergeometric function
Mean	$n\frac{K}{N}$
Mode	$\left\lfloor \frac{(n+1)(K+1)}{N+2} \right\rfloor$
Variance	$n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$



5.2.3 Poisson Distribution

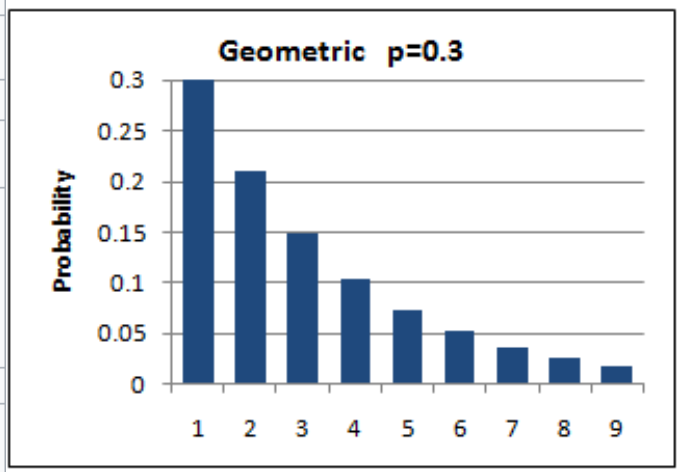
Parameters	$\lambda > 0$ (real)
Support	$k \in \mathbb{N} \cup 0$;
pmf	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda}\sum_{i=0}^{\lfloor k \rfloor}\frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ



The horizontal axis is the index k, the number of occurrences. λ is the expected number of occurrences. The vertical axis is the probability of k occurrences given λ. The function is defined only at integer values of k. The connecting lines are only guides for the eye.

5.2.4 Geometric Distribution

Parameters	$0 < p < 1$ success probability (real)
Support	k trials where $k \in \{1, 2, 3, \ldots\}$
Probability mass function (pmf)	$(1-p)^{k-1}\,p$
CDF	$1-(1-p)^k$
Mean	$\frac{1}{p}$
Median	$\left\lceil \frac{-1}{\log_2(1-p)} \right\rceil$ (not unique if $-1/\log_2(1-p)$ is an integer)
Mode	1
Variance	$\frac{1-p}{p^2}$

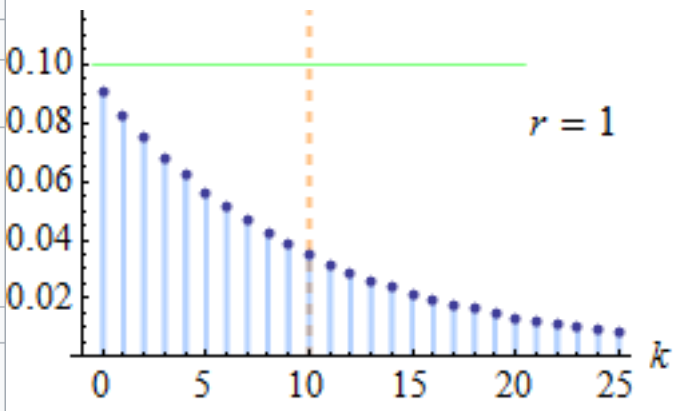


The geometric distribution models the number of failures before one success in a series of independent trials, where each trial results in either success or failure, and the probability of success in any individual trial is constant. For example, if you toss a coin, the geometric distribution models the number of tails observed before getting a head. The geometric distribution is discrete, existing only on the nonnegative integers.

5.2.5 Negative Binomial Distribution

In probability theory and statistics, the negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures occurs.

Parameters	$0 < p < 1$ success probability (real)
Support	k trials where $k \in \{1, 2, 3, \ldots\}$
Probability mass function (pmf)	$(1-p)^{k-1}\,p$
CDF	$1-(1-p)^k$
Mean	$\frac{1}{p}$
Median	$\left\lceil \frac{-1}{\log_2(1-p)} \right\rceil$ (not unique if $-1/\log_2(1-p)$ is an integer)
Mode	1
Variance	$\frac{1-p}{p^2}$

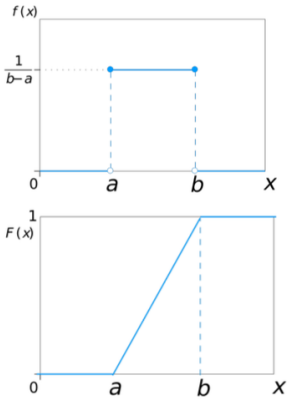


5.3 Important Continuous Distribution

5.3.1 Uniform Distribution

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability. The probability density function and cumulative distribution function for a continuous uniform distribution on the interval are.

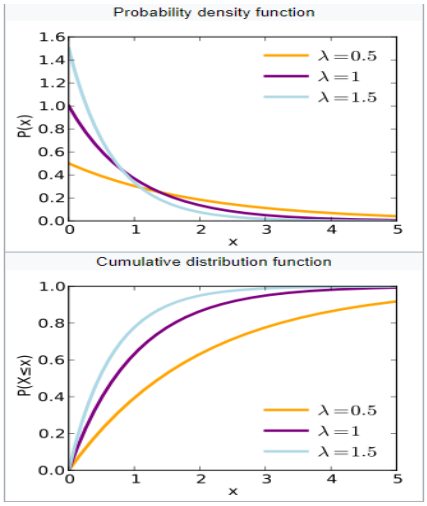
Notation	$\mathcal{U}(a,b)$ or $\operatorname{unif}(a,b)$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a,b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$
Mean	$\frac{1}{2}(a+b)$
Median	$\frac{1}{2}(a+b)$
Mode	any value in (a,b)



5.3.2 Exponential Distribution

In probability theory and statistics, the exponential distribution (also known as negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless. In addition to being used for the analysis of Poisson processes, it is found in various other contexts.

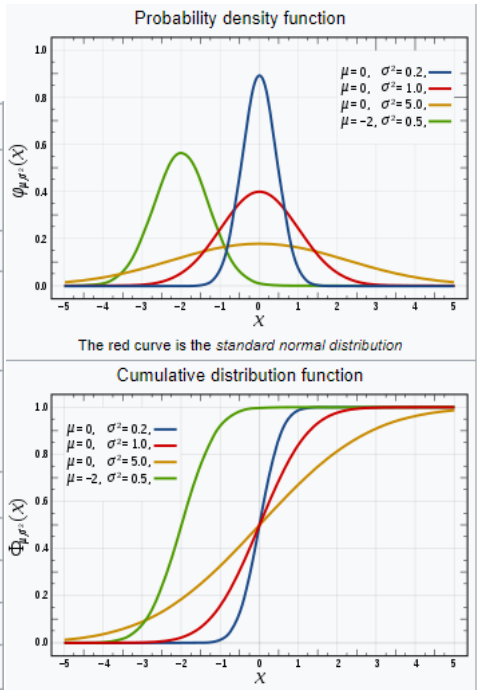
Parameters	$\lambda > 0$ rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$
Quantile	$-\ln(1-F)/\lambda$
Mean	λ^{-1} $(= \beta)$
Median	$\lambda^{-1} \ln(2)$
Mode	0
Variance	λ^{-2} $(= \beta^2)$



5.3.3 Normal Distribution

In probability theory, the normal (or Gaussian) distribution is a very common continuous probability distribution. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.

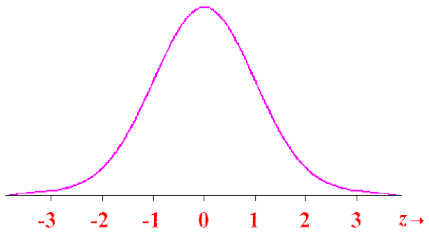
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2



5.3.4 Standard Normal Distribution

This is the "bell-shaped" curve of the Standard Normal Distribution. It is a Normal Distribution with mean 0 and standard deviation 1. It shows you the percent of population: between 0 and Z (option "0 to Z"). Every normal distribution can be transformed to a standard normal distribution by change the value to: $\frac{X-\mu}{\sigma}$

Then we can find the CDF (probability) from Z table.



5.4 Exercise

Question: $X_i \sim iid\ U(o, 1)$, If $X_{i+1} > X_i$, then stop; otherwise continuous drawing. We will get a sequence of $\{X_i\}$, I = 1, 2, …, N. What is E(N)? Tip: iid: independent and identically distributed, E(N): Expect value = $\sum_{N \epsilon \emptyset} N * P(N)$. E(X)= $\int_{-\infty}^{\infty} X f_X(x) dx$

Answer: min is P(N=2) = $\frac{1}{2}$, P(N=3) = $\frac{1}{2} * \frac{1}{3}$, P(N=4)= $\frac{1}{4!}$. So E(N) = $_{N=2}^{\infty} 2N * \frac{1}{N!}$ = $\sum_{N=1}^{\infty} \frac{1}{N!}$

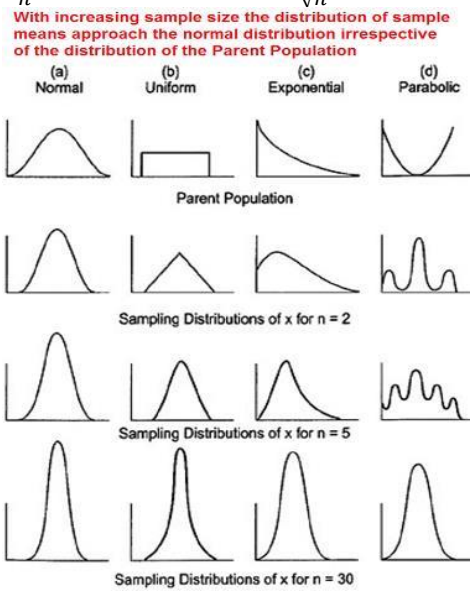
6. Hypothesis Test

6.1 Hypothesis Testing Process

Step 1: Make a hypothesis
Step 2: Calculate the test statistic
Step 3: Find the critical value
Step 4: Confirm the statement
Step 5: Interpret your result

6.2 Central Limit Theorem

Suppose $\{X_1, X_2, ...\}$ is a sequence of i.i.d. random variable with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as *n* approaches infinity, $(\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu \xrightarrow{n \to \infty} N(0, \frac{\sigma^2}{\sqrt{n}})$



6.3 P-value

Define: The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis(H_0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested.

Exercise: $H_0: \mu = 20$, $H_a: \mu \neq 20$. P-value=0.1. If we change $H_a: \mu > 20$, what' s the P-value?

Answer: P-value = 0.05

7. T-test

7.1 T-test Assumption

Assumption:

1. Independent
2. n>30
3. If n<30, then need normal assumption

7.2 One Sample T-test

Hypothesis:

$$H_0: \mu_0 = \mu, H_a: \mu_0 \neq \mu$$

Test Statistics: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

7.3 Proportion T-test

Hypothesis:

$$H_0: p = p_0, H_a: p \neq p_0$$

Test Statistics: $t = \frac{\bar{p}-p_0}{\sqrt{\bar{p}(1-\bar{p})/n}}$

7.4 Paired T-test

Paired t-tests are conducted for a sample of matched pairs of similar units or one group of units has been tested twice.

Hypothesis:

$$H_0: \mu_1 = \mu_2, H_a: \mu_1 \neq \mu_2$$

Test Statistics: $t = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{n}}}$

7.5 Two Sample T-test

1. Control group: we cannot have only one experiment group or compare to historical result

2. Randomization: the object will be randomized selected into control or treatment group

3. Keep control and treatment almost equal size

Hypothesis:

$$H_0: \mu_1 = \mu_2, H_a: \mu_1 \neq \mu_2$$

Test Statistics: $t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$

7.6 Paired vs Two Sample

Whether the test subject can be matched!

Example:

In clinical trial, patient want test whether a new drug has effects: we should use paired t-test to get the value about before& after the new drug. Then we compare with two group with two sample t-test to get the result.

7.7 T-test Example: AB Test

A/B testing (sometimes called split testing) is comparing two versions of a web page to see which one performs better. You compare two web pages by showing the two variants (let's call them A and B) to similar visitors at the same time.

Original

Test

8. Power Analysis

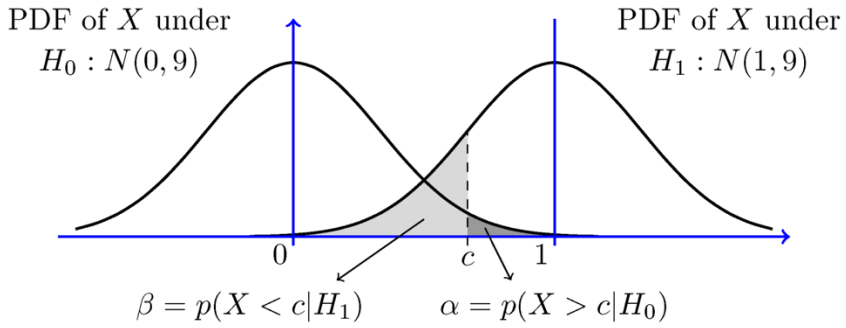
8.1 Confusion Matrix

Confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

- true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- true negatives (TN): We predicted no, and they don't have the disease.
- false positives (FP): We predicted yes, but they don't actually have the disease.
- false negatives (FN): We predicted no, but they actually do have the disease.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

8.2 Power Analysis



In summary, we have determined that we have a H_1 chance of rejecting the null hypothesis H_0:μ=0 in favor of the alternative hypothesis H_a:μ>1 if the true unknown population mean is in reality μ=1.

8.3 Why We Need Power Analysis?

- Sample could be expensive
  - Expensive experiment (Rocket)
  - A rare used feature
  - New secret product/feature

- Whether the difference makes sense

- 0.01% gain worth it?
- Effect size

Data Scientist Materials

Interview Questions ROI
http://www.1point3acres.com/bbs/thread-292951-1-1.html

Product Sense:
https://www.amazon.com/Cracking-PM-Interview-Product-Technology/dp/0984782818/ref=sr_1_1?ie=UTF8&qid=1506459497&sr=8-1&keywords=cracking+pm+interview

DS Probability Exercise
https://www.amazon.com/How-Became-Quant-Insights-2007-07-09/dp/B01NBOUEEU/ref=sr_1_4?s=books&ie=UTF8&qid=1506459582&sr=1-4&keywords=how+I+became+a+quant