Univariate probability

Adam Boult (www.bou.lt)

February 10, 2023

Contents

Preface

This is a live document, and is full of gaps, mistakes, typos etc.

Part I

Probability

Chapter 1

Events, the probability function and the Kolgomorov axioms

1.1 Events

1.1.1 Elementary events

We have a sample space, Ω consisting of elementary events.

All elementary events are disjoint sets.

1.1.2 Non-elementary events

We have a σ-algebra over Ω called F. A σ-algebra takes a set a provides another set containing subsets closed under complement. The power set is an example.

All events E are subsets of Ω

$\forall E \in F E \subseteq \Omega$

1.1.3 Mutually exclusive events

Events are mutually exclusive if they are disjoint sets.

1.1.4 Complements

For each event E, there is a complementary event E^C such that:

$E \vee E^C = \Omega$

$E \wedge E^C = \varnothing$

This exists by construction in the measure space.

1.1.5 Union and intersection

As events are sets, we can define algebra on sets. For example for two events E_i and E_j we can define:

- $E_i \wedge E_j$
- $E_i \vee E_j$

1.2 Kolmogorov axioms

1.2.1 The probability function

For all events E in F, the probability function P is defined.

1.2.2 Measure space

This gives us the following measure space:

(Ω, F, P)

1.2.3 First Kolmogorov axiom

First axiom

The probability of all events is a non-negative real number.

$\forall E \in F[(P(E) \geq 0) \wedge (P(E) \in \mathbb{R})]$

1.2.4 Second Kolmogorov axiom

The probability of one of the elementary events occuring is 1.

The probability of the outcome set is 1.

$P(\Omega) = 1$

1.2.5 Third Kolmogorov axiom

The probability of union for mutually exclusive events is:

$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

1.3 Basic results

1.3.1 Probability of null

$P(\Omega) = 1$

$P(\Omega \vee \varnothing) = 1$

$P(\Omega) + P(\varnothing) = 1$

$P(\varnothing) = 0$

1.3.2 Monotonicity

Consider $E_i \subseteq E_j$:

$E_j = E_i \vee E_k$

$P(E_j) = P(E_i \vee E_k)$

Disjoint so:

$P(E_j) = P(E_i) + P(E_k)$

We know that $P(E_k) \geq 0$ from axiom 1 so:

$P(E_j) \geq P(E_i)$

1.3.3 Bounds of probabilities

As all events are subsets of the sample space:

$P(\Omega) \geq P(E)$

$1 \geq P(E)$

From axiom 1 then know:

$\forall E \in F[0 \leq P(E) \leq 1]$

1.3.4 Union and intersection for null and universal

$P(E \wedge \varnothing) = P(\varnothing) = 0$

$P(E \vee \Omega) = P(\Omega) = 1$

$P(E \vee \varnothing) = P(E)$

$P(E \wedge \Omega) = P(E)$

1.3.5 Separation rule

Firstly:

$P(E_i) = P(E_i \wedge \Omega)$

$P(E_i) = P(E_i \wedge (E_j \vee E_j^C))$

$P(E_i) = P((E_i \wedge E_j) \vee (E_i \wedge E_j^C))$

As the latter are disjoint:

$P(E_i) = P((E_i \wedge E_j) + (E_i \wedge E_j^C))$

1.3.6 Addition rule

We know that:

$P(E_i \vee E_j) = P((E_i \vee E_j) \wedge (E_j \vee E_j^C))$

By the distributive law of sets:

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee E_j)$

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee (E_j \wedge (E_i \vee E_i^C))$

By the distributive law of sets:

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee (E_j \wedge E_i) \vee (E_j \wedge E_i^C))$

As these are disjoint:

$P(E_i \vee E_j) = P(E_i \wedge E_j^C) + P(E_j \wedge E_i) + P(E_j \wedge E_i^C)$

From the separation rule:

$P(E_i \vee E_j) = P(E_i) - P(E_i \wedge E_j) + P(E_j \wedge E_i) + P(E_j) - P(E_j \wedge E_i)$

$P(E_i \vee E_j) = P(E_i) + P(E_j) - P(E_i \wedge E_j)$

1.3.7 Probability of complements

From the addition rule:

$P(E_i \vee E_j) = P(E_i) + P(E_j) - P(E_i \wedge E_j)$

Consider E and E^C:

$P(E \vee E^C) = P(E) + P(E^C) - P(E \wedge E^C)$

We know that E and E^C are disjoint, that is:

$E \wedge E^C = \varnothing$

Similarly by construction:

$E \vee E^C = \Omega$

So:

$P(\Omega) = P(E) + P(E^C) - P(\varnothing)$

$1 = P(E) + P(E^C)$

1.4 Other

1.4.1 Odds

Given a set of outcomes for a variable, the odds of the outcome are defined as:

$o_f = \frac{P(E)}{P(E^C)}$

For example, the odds of rolling a 6 are $\frac{1}{5}$.

1.4.2 Discrete and continous probability

We know that:

$\sum_y P(X \wedge Y) = P(X)$

So for the continuous case

$P(X) = \int_{-\infty}^{\infty} P(X \wedge Y)dy$

This behaves like the probability for a single event, or multiple events with one fewer event if there were more than 2 events to start with.

1.4.3 Marginalisation

Chapter 2

Conditional probability and Bayes' theorem

2.1 Introduction

2.1.1 Conditional probability

We define conditional probability

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_j)}$$

We can show this is between 0 and 1.

$$P(E_j) = P(E_i \wedge E_j) + P(\bar{E}_i \wedge E_j)$$

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_i \wedge E_j) + P(\bar{E}_i \wedge E_j)}$$

We know:

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_j)}$$

$$P(E_j|E_i) := \frac{P(E_i \wedge E_j)}{P(E_i)}$$

So:

$$P(E_i|E_j)P(E_j) = P(E_j|E_i)P(E_i)$$

$$P(E_i|E_j) = \frac{P(E_j|E_i)P(E_i)}{P(E_j)}$$

Note that this is undefined when $P(E_j) = 0$

Note that for the same event,

$$P(E_i|E_j) = \frac{P(E_i \wedge E_j)}{P(E_j)}$$

$$P(E_i|E_j) = 0$$

For the same outcome:

$$P(E_i|E_i) = \frac{P(E_i \wedge E_i)}{P(E_i)}$$

$$P(E_i|E_i) = \frac{P(E_i)}{P(E_i)}$$

$$P(E_i|E_i) = 1$$

2.1.2 Bayes' theorem

From the definition of conditional probability we know that:

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_j)}$$

$$P(E_j|E_i) := \frac{P(E_i \wedge E_j)}{P(E_i)}$$

So:

$$P(E_i \wedge E_j) = P(E_i|E_j)P(E_j)$$

$$P(E_i \wedge E_j) = P(E_j|E_i)P(E_i)$$

So:

$$P(E_i|E_j)P(E_j) = P(E_j|E_i)P(E_i)$$

2.1.3 Independent events

Events are independent if:

$$P(E_i|E_j) = P(E_i)$$

Note that:

$$P(E_i \wedge E_j) = P(E_i|E_j)P(E_j)$$

And so for independent events:

$$P(E_i \wedge E_j) = P(E_i)P(E_j)$$

Chapter 3

Entropy

3.1 Entropy

3.1.1 Information

Criteria

Self information measures surprise of outcome. also called a surprisal.

When we observe an outcome we get information. We can develop a measure for how much information is associated with a specific measurement.

Rule 1: Information is always positive

Rule 2: If $P(x) = 1$, the the information for $I(P(x)) = 0$.

Rule 3: If two events are independent, then their information is additive.

- $P(C) = P(A)P(B)$
- $I(P(C)) = I(P(A)P(B))$
- $I(P(A)) + I(P(B)) = I(P(A)P(B))$

Choice of function

A function which satisifes this is $I(P(A)) = -\log(P(A))$

Any base can be used. 2 is most common, information is in units of bit then.

3.1.2 Entropy

Introduction

Entropy measures the expected amount of information produced by a source.

$H(P(x)) = E(I(P(x))$

Entropy is similar to variance, is the sense that both measure uncertainty.

Entropy, however, has no references to specific values of x. If all values were multiplied by 100, or if parts of the distribution were cut up and swapped, entropy would be unaffected.

For a probability function $p(z)$, its entropy is :

$H(p) = -\int p(z) \ln p(z) dz$.

This is a measure of the spread of a distribution.

Negative infinity means no uncertainty

For a multivariate gaussian $H = d/2ln(2\pi e|\Sigma)$.

Part II

Variables

Chapter 4

Variables

4.1 Variables

4.1.1 Random variables

Defining variables

We have a sample space, Ω. A random variable X is a mapping from the sample space to the real numbers:

$X : \Omega \to \mathbb{R}$

We can then define the set of elements in Ω. As an example, take a coin toss and a die roll. The sample space is:

$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$

A random variable could give us just the die value, such that:

$X(H1) = X(T1) = 1$

We can define this more precisely using set-builder notation, by saying the following is defined for all $c \in \mathbb{R}$:

$\{\omega | X(\omega) \leq c\}$

That is, for any number random variable map X, there is a corresponding subset of Ω containing the ωs in Ω which map to less than c.

Multiple variables

Multiple variables can be defined on the sample space. If we rolled a die we could define variables for

- Whether it was odd/even

- Number on the die
- Whether it was less than 3

With more die we could add even more variables

Derivative variables

If we define a variable X, we can also define another variable $Y = X^2$.

4.1.2 Probability mass functions

$P(X = x) = P(\omega | X(\omega) = x)$

For discrete probability, this is a helpful number. For example for rolling a die.

This is not helpful for continuous probability, where the chance of any specific outcome is 0.

4.1.3 Cumulative distribution functions

Definition

Random variables all valued as real numbers, and so we can write:

$P(X \leq x) = P(\omega | X(\omega) \leq x)$

Or:

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

$F_X(x) = \sum_{x_i \leq x} P(X = x_i)$

Partitions

$P(X \leq x) + P(X \geq x) - P(X = x) = 1$

Interval

$P(a < X \leq b) = F_X(b) - F_X(a)$

4.1.4 Probability density functions

Definition

If continuous, probability at any point is 0. We instead look at probability density.

Derived from cumulative distribution function:

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

The density function is $f_X(x)$.

Conditional probability distributions

For probability mass functions:

$$P(Y = y|X = x) = \frac{P(Y = y \wedge X = x)}{P(X = x)}$$

For probability density functions:

$$f_Y(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

4.2 Multiple variables

4.2.1 Joint and marginal probability

Joint probability

$P(X = x \wedge Y = y)$

Marginal probability

$P(X = x) = \sum_y P(X = x \wedge Y = y)$

$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$

4.2.2 Independence and conditional independence

Independence

x is independent of y if:

$\forall x_i \in x, \forall y_j \in y(P(x_i|y_j) = P(x_i)$

If $P(x_i|y_j) = P(x_i)$ then:

$P(x_i \wedge y_j) = P(x_i).P(y_j)$

This logic extends beyond just two events. If the events are independent then:

$P(x_i \wedge y_j \wedge z_j) = P(x_i).P(y_j \wedge z_k) = P(x_i).P(y_j).P(z_k)$

Note that because:

$$P(x_i|y_j) = \frac{P(x_i \wedge y_j)}{P(y_j)}$$

If two variables are independent

$$P(x_i|y_j) = \frac{P(x_i)P(y_j)}{P(y_j)}$$

$P(x_i|y_j) = P(x_i)$

Conditional independence

$P(A \wedge B|X) = P(A|X)P(B|X)$

This is the same as:

$P(A|B \wedge X) = P(A|X)$

Chapter 5

Expected value, conditional expectation and Jensen's inequality

5.1 Moments

5.1.1 Functionals of probabilities

$\phi(P) \in \mathbb{R}$ is a functional on $P(X)$.

Examples include the expectation and variance.

We can define derivatives on these functionals.

$\phi(P) \approx \phi(P^0) + D_\phi(P - P^0)$

Where D_ϕ is linear.

5.1.2 Expected value

Definition

For a random variable (or vector of random variables), x, we define the expected value of $f(x)$ as :

$E[f(x)] := \sum f(x_i)P(x_i)$

The expected value of random variable x is therefore this where $f(x) = x$.

$E(x) = \sum_i x_i P(x_i)$

Linearity of expectation

We can show that $E(x+y) = E(x) + E(y)$:

$E[x+y] = \sum_i \sum_j (x_i + y_j) P(x_i \wedge y_j)$

$E[x+y] = \sum_i \sum_j x_i [P(x_i \wedge y_j)] + \sum_i \sum_j [y_j P(x_i \wedge y_j)]$

$E[x+y] = \sum_i x_i \sum_j [P(x_i \wedge y_j)] + \sum_j y_j \sum_i [P(x_i \wedge y_j)]$

$E[x+y] = \sum_i x_i P(x_i) + \sum_j y_j P(y_j)$

$E[x+y] = E[x] + E[y]$

Expectations of multiples

Expectations

$E(cx) = \sum_i cx P(x_i)$

$E(cx) = c \sum_i x P(x_i)$

$E(cx) = cE(x)$

Expectations of constants

$E(c) = \sum_i c_i P(c_i)$

$E(c) = cP(c)$

$E(c) = c$

Conditional expectation

If Y is a variable we are interested in understanding, and X is a vector of other variables, we can create a model for Y given X.

This is the conditional expectation.

$E[Y|X]$

$E[P(Y|X)Y]$

In the continuous case this is

$E(Y|X) = \int_{-\infty}^{\infty} yP(y|X)dy$

We can then identify an error vector.

$\epsilon := Y - E(Y|X)$

So:

$Y = E(Y|X) + \epsilon$

Here Y is called the dependent variable, and X is called the dependent variable.

Iterated expectation

$E[E[Y]] = E[Y]$

$E[E[Y|X] = E[Y]$

5.1.3 Jensen's inequality

If ϕ is convex then:

$\phi(E[X]) \geq E[\phi(X)])$

Chapter 6

Variance and covariance

6.1 Introduction

6.1.1 Variance

Definition

The variance of a random variable is given by:

$Var(x) = E((x - E(x))^2)$

$Var(x) = E(x^2 + E(x)^2 - 2xE(x))$

$Var(x) = E(x^2) + E(E(x)^2) - E(2xE(x))$

$Var(x) = E(x^2) + E(x)^2 - 2E(x)^2$

$Var(x) = E(x^2) - E(x)^2$

Variance of a constant

$Var(c) = E(c^2) - E(c)^2$

$Var(c) = c^2 - c^2$

$Var(c) = 0$

Variance of multiple

$Var(cx) = E((cx)^2) - E(cx)^2$

$Var(cx) = E(c^2x^2) - [\sum_i cxP(x_i)]^2$

$Var(cx) = c^2E(x^2) - c^2[\sum_i xP(x_i)]^2$

$Var(cx) = c^2[E(x^2) - E(x)^2]$

$Var(cx) = c^2 Var(x)$

Link between variance of expectation

$E(x)^2 + Var(x) = E(x)^2 + E((x - E(x))^2)$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2 + E(x)^2 - 2xE(x))$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2) + E(E(x)^2) - E(2xE(x))$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2) + E(x)^2 - 2E(x)E(x))$

$E(x)^2 + Var(x) = E(x^2)$

Covariance

$Var(x + y) = E((x + y)^2) - E(x + y)^2$

$Var(x + y) = E(x^2 + y^2 + 2xy) - E(x + y)^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - E(x + y)^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - [E(x) + E(y)]^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - E(x)^2 - E(y)^2 - 2E(x)E(y)]$

$Var(x + y) = [E(x^2) - E(x)^2] + [E(y^2) - E(y)^2] + E(2xy) - 2E(x)E(y)$

$Var(x + y) = Var(x) + Var(y) + 2[E(xy) - E(x)E(y)]$

We then define:

$Cov(x, y) := E(xy) - E(x)E(y)$

Noting that:

$Cov(x, x) = E(xx) - E(x)E(x)$

$Cov(x, x) = Var(x)$

So:

$Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$

$Var(x + y) = Cov(x, x) + Cov(x, y) + Cov(y, x) + Cov(y, y)$

$Cov(x, c) = E(xc) - E(x)E(c)$

$Cov(x, c) = cE(x) - cE(x)$

$Cov(x, c) = 0$

6.1.2 Covariance matrix

With multiple events, covariance can be defined between each pair of events, including the event with itself.

The covariance between 2 variables is:

$Cov(x_i, x_j) := E(x_i x_j) - E(x_i)E(x_j)$

Which is equal to:

$Cov(x_i, x_j) = E[x_i - E(x_i)][x_j - E(x_j)]$

We can therefore generate a covariance matrix through:

$\sum = E[(X - E[X])(X - E[X])^T]$

Chapter 7

Higher moments

7.1 Introduction

7.1.1 Moments

Moments

The nth moment of variable X is defined as:

$E[X^n] = \sum_i x_i^n P(x_i)$

The mean is the first moment.

Central moments

The nth central moment of variable X is defined as:

$\mu_n = E[(X - E[X])^n] = \sum_i (x_i - E[X])^n P(x_i)$

The variance is the second central moment.

Standardised moments

The nth standardised moment of variable X is defined as:

$$\frac{E[(X - E[X])^n]}{(E[(X - E[X])^2]^{\frac{n}{2}}} = \frac{\mu_n}{\sigma^n}$$

Kertosis

Kertosis is the third standardised moment.

Skew

Skew is the fourth standardised moment.

Chapter 8

Markov's inequality and Chebyshev's inequality

8.1 Other

8.1.1 Markov's inequality and Chebyshev's inequality

Lemma 1

$E[I_{X \geq a}] = P(X \geq a)$

Consider the indicator function.

$I_{X \geq a}$

This is equal to 0 if X is below a and 1 otherwise.

We can take expectations of this.

$E[I_{X \geq a}] = P(X \geq a).1 + P(X < a).0 = P(X \geq a)$

$E[I_{X \geq a}] = P(X \geq a)$

Lemma 2

$aI_{X \geq a} \leq X$

While X is below a the left side is equal to 0, which holds.

While X is equal to a the left side is equal to X, which holds.

While X is above a the left side is equal to a, which holds.

Markov's inequality

$P(X \geq a) \leq \frac{\mu}{a}$

From above:

$aI_{X \geq a} \leq X$

We can take expectations of both sides:

$E[aI_{X \geq a}] \leq E[X]$

$aP(X \geq a) \leq E[X]$

$P(X \geq a) \leq \frac{\mu}{a}$

Chebyshev's inequality

We know from Markov's inequality that:

$P(X \geq a) \leq \frac{\mu}{a}$

Let's take the variable X to be $(X - \mu)^2$

$P((X - \mu)^2 \geq a) \leq \frac{E[(X - \mu)^2]}{a}$

$P((X - \mu)^2 \geq a) \leq \frac{\sigma^2}{a}$

$P(|X - \mu| \geq \sqrt{a}) \leq \frac{\sigma^2}{a}$

Take a to be a multiple k^2 of the variance σ^2.

$a = k^2\sigma^2$

$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2}$

$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Chapter 9

Characteristic functions

9.1 Characteristic functions

9.1.1 Characteristic functions

Transformations

Summary

Cumulative probability function

$F = \int_{-\infty}^{\infty} xP(x)$

Moment generating function

$F = \int_{-\infty}^{\infty} e^{tx}P(x)$

Characteristic function

$F = \int_{-\infty}^{\infty} e^{itx}P(x)$

Moment generating function

Take random variable X. This has moments we wish to calculate.

We can transform our function in other forms which maintain all of the required information. For example we could also use the cumulative probability function to calculate moments. We now look for an alternative form of the probability density function which allows us to easily calculate moments.

One method is to use the probability density function and the definitions of moments, but there are other options. For example, consider the function:

$E[e^{tX}]$

Which expands to:

$E[e^{tX}] = \sum_{j=1}^{\infty} \frac{t^j E[X^j]}{j!}$

By taking the mth derivative of this, we get

$E[X^m] + \sum_{j=m+1}^{\infty} \frac{t^j E[X^j]}{j!}$

We can then set $t = 0$ to get

$E[X^m]$

Alternatively, see that differentiating m times gets us

$E[X^m e^{tX}]$

If we can get this function, we can then easily generate moments.

The function we need to get is:

$E[e^{tX}]$

In the discrete case this is:

$E[e^{tX}] = \sum_{i=1} e^{tx_i} p_i$

In the continuous case:

$E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} P(x) dx$

Characteristic function

It may not be possible to calculate the integral for the moment generating function. We now look for an alternative formula with which we can generate the same moments.

Consider

$E[e^{itX}]$

As this can be broken down into sinusoidal functions it can more readily be integrated.

This expands to

$E[e^{itX}] = \sum_{j=1}^{\infty} \frac{i^j t^j E[X^j]}{j!}$

By taking the mth derivative we get.

$E[X^m] i^m + \sum_{j=m+1}^{\infty} \frac{t^j E[X^j]}{j!}$

By setting $t = 0$ we then get:

$E[X^m] i^m$

Alternatively see that differentiating m times gets us

$E[(iX)^m e^{itX}]$

So we can get the moment by differentiating m times, and multiplying by i^{-m}.

Inverses of these functions

Moment generating function

Characteristic function

Moments of constants added to variables

$\phi_{X+c}(t) = E[e^{it(X+c)}]$

$\phi_{X+c}(t) = E[e^{itX} e^{i} tc]$

$\phi_{X+c}(t) = e^{itc} E[e^{itX}]$

$\phi_{X+c}(t) = e^{itc} \phi_X(t)$

$\phi_X(t) = e^{-itc} \phi_{X+c}(t)$

Moments of constants multiplied by events

$\phi_{cX}(t) = E[e^{itcX}]$

$\phi_{cX}(t) = \phi_X(ct)$

Taylor series of a characteristic function

$\phi_X(t) = E[e^{itX}]$

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{\phi_X^j(a)(t-a)}{j!}$

Around $a = 0$

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{\phi_X^j(0)(t)}{j!}$

The characteristic function is now given in terms of its moments.

We know:

$\phi_X^j(0) = E[X^j] i^j$

So:

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{E[X^j] i^j (t)^j}{j!}$

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{E[X^j] (it)^j}{j!}$

We know:

$$\frac{E[X^0](it)^0}{0!} = E[1] = 1$$
$$\frac{E[X^1](it)^1}{1!} = E[X](it) = it\mu_X$$
$$\frac{E[X^2](it)^2}{2!} = \frac{-E[X^2]t^2}{2} = \frac{-(\mu_X + \sigma_X^2)t^2}{2}$$

So:

$$\phi_X(t) = 1 + it\mu_X - \frac{(\mu_X + \sigma_X^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](it)^j}{j!}$$

Part III

Single observation probability distributions

Chapter 10

Degenerate, Bernoulli and categorical distributions

10.1 Degenerate distribution

10.1.1 Degenerate distribution

10.1.2 Dirac delta distribution

10.2 Bernoulli distribution

10.2.1 Introduction

The outcome of a Bernoulli trial is either 0 or 1. We can describe it as:

$P(1) = p$

$P(0) = 1 - p$

With a single parameter p.

10.2.2 Moments of the Bernoulli distribution

The mean of a Bernoulli trial is $E[X] = (1-p)(0) + (p)(1) = p$.

The variance of a Bernoulli trial is $E[(X-\mu)^2] = (1-p)(0-\mu)^2 + (p)(1-\mu)^2 = (1-p)p^2 + p(1-p)^2] = p(1-p)$.

10.3 Categorical distribution

10.3.1 The categorical distribution

Bernoulli with three or more discrete possible outcomes.

Chapter 11

Simple continuous distributions

11.1 Continous distributions

11.1.1 Uniform distribution

There is a set s such that:

$P(x \in s) = p$

$P(x \notin s) = 0$

Moments of the uniform distribution

The mean is the mean of the set s.

If the set is all numbers of the real line between two values, a and b, then:

The mean is $\frac{1}{2}(a + b)$.

The variance is $\frac{(b-a)^2}{12}$ in the continuous case.

The variance is $\frac{(b-a+1)^2-1}{12}$ in the discrete case.

11.2 Other

11.2.1 Weibull distribution

11.2.2 Power law

$P(X) = \frac{\alpha - 1}{a}(\frac{x}{a})^{-\alpha}$

Where a is the lower bound.

$P(X) = 0$ for $X < a$.

Moments of the power law

$E[X^m] = \frac{\alpha - 1}{\alpha - 1 - m}a$

If $m \geq \alpha - 1$ then this is not well defined.

Higher order moments, such that the variance, cannot be identified.

11.2.3 Logistic distribution

The logistic distribution has the cumulative distribution function:

$$F(x) = \frac{1}{1 + e^{-\frac{x - \mu}{s}}}$$

11.2.4 Laplace distribution

11.2.5 Lévy distribution

Definition

The Lévy distribution is a continuous probability distribution.

The marginal probability is:

$$P(X) = \sqrt{\frac{c}{2\pi}}\frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{\frac{3}{2}}}$$

11.2.6 Split-normal distribution

Part IV

The central limit theorem

Chapter 12

Independent and identically distributed variables

12.1 Identically Independently Distributed variables (IID)

12.1.1 IID

Identically distributed

x is identically distributed to y if:

$\forall i(\exists x_i \rightarrow P(x_i) = P(y_i))$

Covariance matrix of IID variables

For IID varaibles, the covariance matrix is:

$\Sigma = \sigma^2 I$

Chapter 13

The weak law of large numbers

13.1 Weak law of large numbers

13.1.1 Weak law of large numbers

The sample mean is:

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

The variance of this is:

$Var[\bar{X}_n] = Var[\frac{1}{n}\sum_{i=1}^{n} X_i]$

$Var[\bar{X}_n] = \frac{1}{n^2} nVar[X]$

$Var[\bar{X}_n] = \frac{\sigma^2}{n}$

We know from Chebyshev's inequality:

$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Use $\bar{X}_n$ as X:

$P(|\bar{X}_n - \mu| \geq \frac{k\sigma}{\sqrt{n}}) \leq \frac{1}{k^2}$

Update k so $k := \frac{k\sqrt{n}}{\sigma}$

$$P(|\bar{X}_n - \mu| \geq k) \leq \frac{\sigma^2}{nk^2}$$

As n increases, the chance that the sample mean lies outside a given distance from the population mean approaches 0.

Chapter 14

Levy's continuity theorem

14.1 Lévy's continuity theorem

14.1.1 Lévy's continuity theorem

Chapter 15

The central limit theorem and the gaussian/normal distribution

15.1 Central limit theorem

15.1.1 Central limit theorem

Generalise weak law of large numbers

Characteristic function of summed IID events

$Z = \sum_{i=1}^{n} Y_i$

$\phi_Z(t) = E[e^{itZ}]$

$\phi_Z(t) = E[e^{it\sum_{i=1}^{n} Y_i}]$

$\phi_Z(t) = E[e^{itY}]^n$

$\phi_Z(t) = \phi_Y(t)^n$

Taylor series: first moments dominate with means

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X}{n}$

$\phi_Z(t) = \phi_Y(t)^n$

$\phi_Z(t) = \phi_{\frac{X}{n}}(t)^n$

$$\phi_Z(t) = \phi_X(\frac{t}{n})^n$$

$$\phi_X(t) = 1 + it\mu_X - \frac{(\mu_X + \sigma_X^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](it)^j}{j!}$$

$$\phi_X(\frac{t}{n}) = 1 + i\frac{t\mu_X}{n} - \frac{(\mu_X + \sigma_X^2)(\frac{t}{n})^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](i\frac{t}{n})^j}{j!}$$

$$\phi_X(\frac{t}{n}) = 1 + i\frac{t\mu_X}{n} - \frac{(\mu_X + \sigma_X^2)t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[X^j](i\frac{t}{n})^j}{j!}$$

Eliminating the imaginary term

We want μ to be 0.

$$Z = \sum_{i=1}^{n} Y_i$$

$$Y = \frac{X - \mu_X}{n}$$

$$\phi_Y(t) = 1 + it\mu_Y - \frac{(\mu_Y + \sigma_Y^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\mu_Y = E[\frac{X - \mu_X}{n}] = \mu_X - \mu_X n = 0$$

$$\phi_Y(t) = 1 - \frac{\sigma_Y^2 t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\sigma_Y^2 = E[(\frac{X - \mu_X}{n})^2]$$

$$\sigma_Y^2 = E[\frac{X^2 + \mu_X^2 - 2X\mu_X}{n^2}]$$

$$\sigma_Y^2 = \frac{E[X^2] + E[\mu_X^2] - E[2X\mu_X]}{n^2}] \; \sigma_Y^2 = \frac{E[X^2] - \mu_X^2}{n^2}]$$

$$\sigma_Y^2 = \frac{\sigma_X^2}{n^2}$$

$$\phi_Y(t) = 1 - \frac{\sigma_X^2 t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X - \mu}{n})^j](it)^j}{j!}$$

$$\phi_Z(t) = \phi_Y(t)^n$$

$$\phi_Z(t) = [1 - \frac{\sigma_X^2 t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X - \mu}{n})^j](it)^j}{j!}]^n$$

$$\phi_Z(t) = [1 - \frac{\sigma_X^2 t^2}{2n^2}]^n$$

Eliminating σ^2

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X - \mu_X}{\sigma n}$

$\phi_Y(t) = 1 + it\mu_Y - \frac{(\mu_Y + \sigma_Y^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$

$\mu_Y = E[\frac{X - \mu_X}{\sigma_X n}] = \mu_X - \mu_X \sigma_X n = 0$

$\phi_Y(t) = 1 - \frac{\sigma_Y^2 t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$

$\sigma_Y^2 = E[(\frac{X - \mu_X}{\sigma n})^2]$

$\sigma_Y^2 = E[\frac{X^2 + \mu_X^2 - 2X\mu_X}{\sigma^2 n^2}]$

$\sigma_Y^2 = \frac{E[X^2] + \mu_X^2 - 2E[X]\mu_X}{\sigma^2 n^2}$

$\sigma_Y^2 = \frac{E[X^2] - \mu_X^2}{\sigma^2 n^2}$

$\sigma_Y^2 = \frac{\sigma_X^2}{\sigma^2 n^2}$

$\sigma_Y^2 = \frac{1}{n^2}$

$\phi_Y(t) = 1 - \frac{t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X - \mu}{\sigma n})^j](it)^j}{j!}$

$\phi_Z(t) = \phi_Y(t)^n$

$\phi_Z(t) = [1 - \frac{t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X - \mu}{\sigma n})^j](it)^j}{j!}]^n$

$\phi_Z(t) = [1 - \frac{t^2}{2n^2}]^n$

Preparing for exponential expansion

We know that

$[1 + \frac{x}{n}]^n = e^x$

As $n \to \infty$.

With:

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X - \mu_X}{\sigma n}$

We have:

$\phi_Z(t) = [1 - \frac{t^2}{2n^2}]^n$

With:

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X - \mu_X}{\sigma \sqrt{n}}$

We have:

$\phi_Z(t) = [1 - \frac{t^2}{2n}]^n$

Which tends towards

$\phi_Z(t) = e^{-\frac{1}{2}t^2}$

Rescaling

The average of random variables, less their mean, and divided by their standard deviation multiplied by the square root of the sample size, follows a normal distribution as n increases.

What does this say about the actual distribution of sample averages?

$Z = \sum_{i=1}^{n} Y_i$

$Y_i = \frac{X_i - \mu_X}{\sigma_X \sqrt{n}}$

$\sum_{i=1}^{n} Y_i$

$Y = \frac{X}{n}$

Let's create Q.

$Q = \frac{Z\sigma_X}{\sqrt{n}} + \mu_X$

$Q = \frac{(\sum_{i=1}^{n} Y_i)\sigma_X}{\sqrt{n}} + \mu_X$

$Q = \frac{(\sum_{i=1}^{n} (\frac{X_i - \mu_X}{\sigma_X \sqrt{n}}))\sigma_X}{\sqrt{n}} + \mu_X$

$$Q = \sum_{i=1}^{n} (\frac{X_i - \mu_X}{n}) + \mu_X$$

$$Q = \sum_{i=1}^{n} (\frac{X_i - \mu_X}{n} + \frac{\mu_X}{n})$$

$$Q = \sum_{i=1}^{n} (\frac{X_i}{n})$$

This is the sample average.

$$\phi_Q(t) = \phi_{\frac{Z\sigma_X}{\sqrt{n}} + \mu_X}(t)$$

$$\phi_Q(t) = \phi_Z(\frac{t\sigma_X}{\sqrt{n}}) e^{it\mu_X}$$

$$\phi_Z(\frac{t\sigma_X}{\sqrt{n}}) = e^{-\frac{1}{2}(\frac{t\sigma_X}{\sqrt{n}})^2}$$

$$\phi_Z(\frac{t\sigma_X}{\sqrt{n}}) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}}$$

$$\phi_Q(t) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}} e^{it\mu_X}$$

Normal distribution

We name the normal distribution this function when $n = 1$

$$N(\mu_X, \sigma_X^2) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}} e^{it\mu_X}$$

$$N(\mu_X, \sigma_X^2) = e^{-\frac{1}{2}t^2\sigma_X^2} e^{it\mu_X}$$

Getting the probability distribution function

$$\phi_X(t) = e^{-\frac{1}{2}t^2\sigma_X^2} e^{it\mu_X}$$

$$\phi_X(t) = e^{-\frac{1}{2}t^2\sigma_X^2} [\cos(t\mu_X) + i\sin(t\mu_X)]$$

15.2 Convergence

15.2.1 Convergence in distribution (converge weakly)

15.2.2 Convergence in probability and o-notation

Introduction

Converges in probability

$P(distance(X_n, X) > \epsilon) \to 0$

For all ϵ.

$X_n \to^P X$

Little o notation

Little o notation is used to describe convergence in probability.

$X_n = o_p(a_n)$

mean that

$\frac{X_n}{a_n}$

Converges to 0 and n approaches something

Can be wrtiten:

$\frac{X_n}{a_n} = o_p(1)$

Big O notation

Big O notation is used to describe boundedness.

$X_n = O_p(a_n)$

means that:

If something is little o, it is big O.

15.2.3 Almost sure convergence

X_n converges almost surely to X if:

$d(X_n, X) \to 0$

Where $d(X_n, X)$ is a distance metric.

$X_n \to^{as} X$

15.3 Gaussian distributions

15.3.1 Gaussian

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

15.3.2 The error function and the complementary error function

15.3.3 Multivariable Gaussian distribution

Definition

For univariate:

$x \sim N(\mu, \sigma^2)$

We define the multivariate gaussian distribution as the distribution where any linear combination of components are gaussian.

For multivariate:

$X \sim N(\mu, \Sigma)$

Where μ is now a vector, and Σ is the covariance matrix.

Density function is :

$$f_x = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

For normal gaussian it is:

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

This is the same wher $n = 1$.

Singular Gaussians

Need det $|\Sigma|$ and Σ^{-1}. These rely on the covariance matrix not being degenerate.

If the covariance matrix is degenerate we can instead use the pseudo inverse, and the pseudo determinant.

Part V

More probability distributions from IID

Chapter 16

Statistics

16.1 Creating statistics

16.1.1 Creating statistics

We take a sample from the distribution.

$x = (x_1, x_2, ..., x_n)$

A statistic is a function on this sample.

$S = S(x_1, x_2, ..., x_n)$.

16.2 Moments of statistics

16.2.1 Bias from single and joint estimation

Bias from single estimation

$\mathbf{x}_i$ and $\mathbf{z}_i$ are not independent, so we cannot estimate just $y_i = \mathbf{x}_i\theta$.

Bias from joint estimation

We could estimate our equation with a single ML algorithm.

$y_i = f(\mathbf{x}_i, \theta) + g(\mathbf{z}_i) + \epsilon_i$

For example, using LASSO.

However this would introduce bias into our estimates for θ.

Bias from iterative estimation

We could iteratively estimate both θ and $g(\mathbf{z}_i)$.

For example iteratvely doing OLS for θ and random forests for z_i.

This would also introduce bias into θ.

16.3 Asymptotic properties of statistics

16.3.1 Asymptotic distributions

$f(\hat{\theta}) \rightarrow^d G$

Where G is some distribution.

16.3.2 Asymptotic mean and variance

16.3.3 Asymptotic normality

Many statistics are asymptotically normally distribution.

This is a result of the central limit theorem.

For example:

$\sqrt{n}S \rightarrow^d N(s, \sigma^2)$

Confidence intervals for asymptotically normal statistics

We have the mean and variance, and know the distribution. This allows us to calculare confidence intervals.

Chapter 17

Order statistics

17.1 Order statistics

17.1.1 Order statistics

Defining order statistics

The kth order statistic is the kth smallest value in a sample.

$x_{(1)}$ is the smallest value in a sample, the minimum.

$x_{(n)}$ is the largest value in a sample, the maximum.

Probability distributions of order statistics

The probability distribution of order statistics depends on the underlying probability distribution.

Probability distribution of sample maximum

If we have:

$Y = \max \mathbf{X}$

The probability distribution is:

$P(Y \leq y) = P(X_1 \leq y, X_2 \leq y, ..., X_n \leq y)$

If these are iid we have:

$P(Y \leq y) = \prod_i P(X_i \leq y)$

$F_y(y) = F_X(y)^n$

The density function is:

$f_y(y) = nF_X(y)^{n-1} f_x(y)$

Probability distribution of the sample minimum

If we have:

$Y = \min \mathbf{X}$

The probability distribution is:

$P(Y \leq y) = P(X_1 \geq y, X_2 \geq y, ..., X_n \geq y)$

If these are iid we have:

$P(Y \leq y) = \prod_i P(X_i \geq y)$

$F_y(y) = [1 - F_X(y)]^n$

The density function is:

$f_y(y) = -n[1 - F_X(y)]^{n-1} f_x(y)$

Chapter 18

Totals of independent draws: Binominal and Poisson distributions

18.1 Binomial

18.1.1 Binomial distribution

If we repeat a Bernoulli trials with the same parameter and sum the results, we have the binomial distribution.

We therefore have two parameters, p and n.

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

18.1.2 Moments of the binomial distribution

The mean is np, which can be seen as the trials are independent.

Similarly, the variances can be addeded together giving $np(1-p)$.

18.1.3 Multinomial distribution

The mass function for the binomial case is:

$f(x) = \frac{n!}{x!(n-x)!} p^k (1-p)^{n-k}$

18.1.4 The multinomial distribution

This generalises the binomial distribution where there are more than 2 outcomes.

$f(x_1, ..., x_n) = \frac{n!}{\prod_i x_i!} \prod_i p_i^{x_i}$

18.2 Poisson

18.2.1 Poisson distribution

18.2.2 Definition

We can use the Poisson distribution to model the number of indepedent events that occur in an a time period.

For a very short time period the chance of us observing an event is a Bernoulli trial.

$P(1) = p$

$P(0) = 1 - p$

18.2.3 Chance of no observations

Let's consider the chance of repeatedly getting 0: $P(0; t)$.

We can see that: $P(0; t + \delta t) = P(0; t)(1 - p)$.

And therefore:

$P(0; t + \delta t) - P(0; t) = -pP(0; t))$

By setting $p = \lambda \delta t$:

$\frac{P(0; t + \delta t) - P(0; t)}{\delta t} = -\lambda P(0; t))$

$\frac{\delta P(0; t)}{\delta t} = -\lambda P(0; t)$

$P(0; t) = Ce^{-\lambda t}$

If $t = 0$ then $P(0; t) = 0$ and so $C = 1$.

$P(0; t) = e^{-\lambda t}$

18.2.4 Deriving the Poisson distribution

Chapter 19

Time between draws: geometric and exponential distributions

19.1 Geometric distribution

19.1.1 Geometric distribution

19.2 Exponential distribution

19.2.1 Exponential distribution

Chapter 20

Extreme value distributions

20.1 Extreme value distributions

20.1.1 Type-I - Gumbel distribution

The probability function is:

$$f(x) = \frac{1}{\beta} e^{-(\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}})}$$

We can use:

$$z = \frac{x-\mu}{\beta}$$

To get:

$$f(x) = \frac{1}{\beta} e^{-(z+e^{-z})}$$

Link to the logistic function

The difference between two draws from a Gumbel distribution is drawn from the logistic function.

20.1.2 Type-II - Frechet distribution

20.1.3 Type-III - Reversed Weibull distribution

Chapter 21

The geometric distribution

21.1 Introduction

21.1.1 Introduction

Part VI

Distributions with multiple variables

Chapter 22

Mixture distributions

22.1 Mixture models

22.1.1 Gaussian Mixture Models

Mixture models

We have a latent variable which is part of the process

The variable is distributed according to parametric distribution, but parameters are different for differnet latent classes.

There are K latent classes, and so K sets of parameters.

The population is weighted into the K classes.

We have a distribution, but we have different parameters for the distribution for different populations.

For example we could observe the height of men and women, where both are normally distributed but with different parameters.

Where there is a normal distribution, this is a Gaussian mixture model.

If there is more than one variable to observe, this is a multivariate Gaussian mixture model.

Gaussian Mixture Models (GMM)

In a Gaussian Mixture Model each non latent variable has a normal distriubtion with a mean and variance. For multiple variables there is a covariance matrix.

Chapter 23

Latent class analysis and the expectation-maximisation algorithm

23.1 Latent variable models

23.1.1 Latent class analysis

23.2 The Expectation-Maximisation (EM) algorithm

23.2.1 The Expectation-Maximisation algorithm

Expectation-Maximisation algorithm

This is used to learn the parameters for a Gaussian Mixture Model

We cannot simply maximise the likelihood function, because this cannot be specified for a latent model.

The log likelihood function normally is:

$L(\theta; X) = p(X|\theta)$

With hidden variables it is:

$L(\theta; X, Z) = p(X|\theta) = \int p(X, Z|\theta) dZ$

1: Expectation step

We consider the expected log likelihood. We call this

$E[\log L(\theta; X, Z)]$

2: Maximisation step

23.3 Stochastic Expectation-Maximisation

Part VII

The empirical distribution

Chapter 24

The empirical distribution

Part VIII

Exploratory data analysis

Chapter 25

Data cleaning

Chapter 26

Summary statistics and visualisation for one variable

26.1 Basis statistics for a single variable

26.1.1 N

The is the size of the sample.

26.1.2 Sample range

Minimum

This is the smallest value in the sample.

Maximum

This is the largest value in the sample.

Range

This is the difference between the maximum and minimum.

Median

This is the value whereby 50% of the sample can be found below the value.

Percentiles

The xth percentile is the value by which $x\%$ of the values can be found below it.

Interquartile range

This is the differnence between the 25th percentile and the 75th percentile.

26.1.3 Sample mode

The is the most common value in the sample.

26.2 Sample moments

26.2.1 Sample mean

We previously defined the population mean is defined as $\mu = E[X]$.

The sample mean is defined as $\bar{x} = \frac{1}{n}\sum_i x_i$.

Centred mean

We can subtract the mean from each entry in the sample. This will leave a new mean of 0. This is convenient for many calculations.

26.2.2 Sample variance

We previously defined the population variance as $\sigma^2 = E[(X-\mu)^2]$.

We define the sample variance as $\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$.

We can calculate this using matrices:

$M = X - \bar{x}$

$\sigma^2 = \frac{1}{n} M^T M.$

Centred variance

If $\bar{x} = 0$ then:

$\sigma^2 = \frac{1}{n} X^T X.$

26.3 Other

26.3.1 Standard error

26.3.2 Standard deviation

26.3.3 Sample size

26.4 Updating statistics

26.4.1 Updating the mean

$$\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{n+1}$$

26.4.2 Updating the variance

If it is centred:

$$\sigma_n^2 = \frac{1}{n} X_n^T X_n$$

So:

$$\sigma_{n+1}^2 = \frac{n\sigma_n^2 + x_{n+1}^t x_{n+1}}{n+1}$$

26.5 Visualising a single continous variable

26.5.1 Box and whisker plots

26.5.2 Density plot

Chapter 27

Testing population means with Z-tests and T-tests

27.1 Z-test

27.1.1 Z-test for variable significance

The standard score

We may want to see how different a mean statistic is from a specific value.

The standard score allows us to measure this, by taking this distance and standardising by the standard deviation.

$$z = \frac{\bar{x} - x_0}{\sigma}$$

This requires us to know the standard deviation, which is in general not known.

If the sample size is large, we know this converges to the normal distribution through the central limit theorem.

The Z-test

We can see how likely our statistic was to be produced if it was drawn from a normal distribution with mean x_0 and standard deviation s_0.

P-values

This is the chance of the statistic being produced by chance.

27.2 t-test

27.2.1 T-test for variable significance

T-statistic

In practice we don't know the population standard deviation and so must estimate it instead.

We use the standard deviation on the sample.

$$t = \frac{\bar{x} - x_0}{s_0}$$

Student's t-distribution

As we have used the sample standard deviation we have lost a degree of freedom, and can no longer model the variable as a normal distribution, as we did for the z-statistic.

We now have a distribution with an addition parameter, the number of degrees of freedom.

The number of degrees of freedom is $n - 1$.

As the sample size tends towards infinity, the distribution tends towards the normal distribution.

Student's t-test

Confidence interval

27.2.2 Welch's t-test

Alternative to student.

Chapter 28

Pivotal quantities

28.1 Pivotal quantity

28.1.1 Introduction

A pivotal quantity is a statistic whose distribution does not depend on the parameters of the underlying distribution.

For example, the z statistic if the underyling distribution is a normal distribution.

Chapter 29

Jackknifing

29.1 Jackknifing

29.1.1 The jackknife

We have a statistic:

$S(x_1, x_2, ..., x_n)$

We may want to estimate moments for this statistic, but are unable to do so.

The jackknife estimator

The jackknife is an approach for getting moments for statistics.

We start by creating n statistics each leaving out one observation.

$\bar{S}_i(x_1, x_2, ...x_{i-1}, x_{i+1}, ..., x_n)$

We define:

$\bar{S} = \frac{1}{n} \sum_i \bar{S}_i$

Moments of the jackknife estimator

We want to know the variance.

$Var\bar{S} = \frac{n-1}{n} \sum_i (\bar{S}_i - \bar{S})^2.$

29.1.2 The infintesimal jackknife

The jackknife as a weighting

In the jackknife we calculate the statistic leaving one observation out.

This is the same as weighting observations and giving one a weighting of 0 and the others 1.

The infintesimal jackknife

For the infintesimal jackknife we reduce the weight not to 0, but by an infintesimal amount.

29.1.3 Variance of jackknife estimators

Chapter 30

Bootstrapping

30.1 Bootstrapping

30.1.1 Bootstrapping

If we have a sample of n, we can create bootstrap samples by drawing with replacement for other sets with n members.

30.1.2 Variance of bootstrap estimators

Part IX

Estimating generative probability distributions

Chapter 31

Non-parametric estimation of probability distributions

31.1 Histograms

31.1.1 Histograms

31.2 Kernels

31.2.1 Kernel density estimation

31.2.2 Smoothing kernel estimation

Smoothed kernels

We have $K(x - x_i)$

We can smooth this to:

$K_h(x - x_i) = \frac{1}{h} K(\frac{x - x_i}{h})$

Where $h > 0$ is the smoothing bandwidth.

$f(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$

Chapter 32

Bayesian parameter estimation

32.1 Bayesian parameter estimation

32.1.1 Bayesian parameter estimation

Bayes rule

We want to generate the probability distribution of θ given the evidence X.

We can transform this using Bayes rule.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Here we have:

- Our prior - $P(\theta)$
- Our likelihood function - $P(X|\theta)$
- Our posterior - $P(\theta|X)$

Normal priors and posteriors

If our prior is a normal distribution then:

$$P(\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_0|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma_0^{-1}(x-\mu)}$$

Similarly, if our likelihood function $P(X|\theta)$ is a normal distriubtion then:

$$P(X|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We can now plug these into Bayes rule:

$$P(\theta|X) = \frac{1}{P(X)} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(\theta|X) \propto e^{-\frac{1}{2}[\frac{(\theta-\mu_0)^2}{\sigma_0^2} + \frac{(x-\mu)^2}{\sigma^2}]}$$

We can then set this an a new Gaussian:

$$P(\theta|X) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|^{\frac{1}{2}}}} e^{-\frac{1}{2}[\frac{(\theta-\mu_0)^2}{\sigma_0^2} + \frac{(x-\mu)^2}{\sigma^2}]}$$

32.1.2 Empirical Bayes

Bayes rule

We can calculate the posterior probability for θ, but we need a prior $P(\theta)$.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Empirical Bayes

With empirical Bayes we get our prior from the data.

We have $P(X|\theta)$

And $P(\theta|\rho)$

We observe X and want to estimate θ.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)}{P(X)} \int P(\theta|\rho)P(\rho)d\rho$$

32.1.3 Prior and posterior predictive distributions

Prior predictive distribution

Our prior predictive distribution for X depends on our prior for θ.

$P(\mathbf{x}) = \int_\Theta P(\mathbf{x}|\theta)P(\theta)d\theta$

Posterior predictive distribution

Once we have calculated $P(\theta|X)$, we can calculate a posterior probability distribution for X.

$P(\mathbf{x}|\mathbf{X}) = \int_\Theta P(\mathbf{x}|\theta)P(\theta|\mathbf{X})d\theta$

32.1.4 Bayesian risk

Risk and Bayes risk.

Chapter 33

Point estimates of probability distributions

33.1 Point estimates for parameters

33.1.1 Estimators

When we take statistics we are often concerned with inferring properties of the underlying probability function.

As the properties of the probability distribution function affect the chance of observing the sample, we can analyse samples to infer properties of the underlying distribution.

There are many properties would could be interested in. This includes moments and parameters of a specific probability distribution function.

An estimator is a statistic which is our estimate of one of these values.

Emphasise that statistics and estimators are different things. A statistic may be terrible estimator, but be useful for other purposes.

33.1.2 Sufficient statistics

We can make estimates of a population parameter using statistics from the same.

A statistic is sufficient if it contains all the information needed to estimate the parameter.

We can describe the role of a parameter as:

$P(x|\theta, t)$

t is a sufficient statistic for θ if:

$P(x|t) = P(x|\theta, t)$

33.2 Properties of point estimators

33.2.1 Estimator error and bias

Error of an estimator

The error of an estimator is the difference between it and the actual parameter.

$Error_\theta[\hat{\theta}] = \hat{\theta} - \theta$

Bias of an estimator

The bias of an estimator is the expected error.

$Bias_\theta[\hat{\theta}] := E_\theta[\hat{\theta} - \theta]$

$Bias_\theta[\hat{\theta}] := E_\theta[\hat{\theta}] - \theta$

33.2.2 Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of an estimator

Mean squared error

Mean squared error

$MSE = E[(\hat{\theta} - \theta)^2] = E[((\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta))^2]$

$MSE = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}])^2 + (E[\hat{\theta}] - \theta)^2 + 2(E[\hat{\theta}] - \theta)(\hat{\theta} - E[\hat{\theta}])]$

$MSE = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E[\hat{\theta}])^2] + E[(E[\hat{\theta}] - \theta)^2] + E[2(E[\hat{\theta}] - \theta)(\hat{\theta} - E[\hat{\theta}])]$

$MSE = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2 + 2(E[\hat{\theta}] - \theta)E[\hat{\theta} - E[\hat{\theta}]]$

$MSE = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$

Root Mean Square Error (RMSE)

This is the square root of the MSE.

It is also called the Root Mean Square Deviation (RMSD)

33.2.3 Asymptotic properties of estimators

33.2.4 Consistency and efficiency of estimators

Consistency

A statistic $\hat{\theta}$ is a consistent estimator for θ if its error tends to 0.

That is:

$\hat{\theta} \rightarrow^p \theta$

We can show that an estimator is consistent if we can write:

$\hat{\theta} - \theta$ as a function of n, causing it to tend to 0.

Efficiency

Efficiency measures the speed at which a consistent estimator tends towards the true value.

The speed of this convergence is the efficiency. could be fairly efficient plus biased too p Measured as:

$$e(\hat{\theta}) = \frac{\frac{1}{I(\theta)}}{Var(\hat{\theta})}$$

If an estimator as an efficiency of 1 and is unbiased, it is efficient.

Relative efficiency

We can measure the relative efficiency of two consistent estimators:

The relative efficiency is the variance of the first estimator, divided by the variance of the second.

Root-n estimators

An estimator is root-n consistent if it is consistent and its variance is:

$O(\frac{1}{n})$

n^δ-convergent

A consistent estimator is n^δ-consistent if its variance is:

$O(\frac{1}{n^{2\delta}})$

33.2.5 Cramér-Rao lower bound

For an unbiased estimator, the variance cannot be below the Cramer-Rao lower bound.

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Where $I(\theta)$ is the Fisher information.

We can prove this.

We have the score:

$V = \frac{\delta}{\delta\theta} \ln f(X, \theta)$

$V = \frac{1}{f(X,\theta)} \frac{\delta}{\delta\theta} f(X, \theta)$

The expectation of the score is 0:

$E[V] = E[\frac{1}{f(X,\theta)} \frac{\delta}{\delta\theta} f(X, \theta)]$

$E[V] = \int \frac{1}{f(X,\theta)} \frac{\delta}{\delta\theta} f(X, \theta) dx$

33.2.6 Bias-Variance trade-off

Bias-variance trade-off. if we care about $E[(y - xt)^2]$ then we may not want an unbiased estimator. by adding some bias we could reduce the variance a lot.

33.3 Sort

33.3.1 Testing estimators

Assessing estimators of parametric models: do monte carlo simulations

33.3.2 Loss

loss functions for point estimates. point estimate confidence interval h3

33.3.3 Estimator properties

best asymptotically normal (BAN) estimators AKA consistently asymptotically normal efficience (CANE)

these are root n consistent!

33.3.4 Feasible and infeasible estimators

Feasible uses known terms. Infeasible uses those that aren't

Eg Ω is infeasible, unless we assume its form, making it feasible.

33.3.5 Bias etc

pages: + Cramer rao + Minimum-Variance Unbiased Estimators (MVUE)

Unbiased estimators for some kernel value. Can use used to estimate population moments.

33.3.6 Rao-Blackwell theorem

33.3.7 One step and k-step estimators

in cramer rao stuff?

33.3.8 Delta method

in bias section?

We can consider X_n to be a sequence. We are interest in asymptotic properties of this sequence.

33.3.9 Fat tails

section on fat tails + can't estimate pop mean from sample mean + method of moments requires non-fat tails + correlation/covariance with fat tails.

Chapter 34

Likelihood functions

34.1 Likelihood functions

34.1.1 Likelihood function

We want to estimate parameters. One way of looking into this is to look at the likelihood function:

$L(\theta; X) = P(X|\theta)$

The likelihood function shows the chance of the observed data being generated, given specific parameters.

If this has high peaks then it provides information that θ is located in this region.

34.1.2 IID

For multiple events, the likelihood function is:

$L(\theta; X) = P(X|\theta)$

$L(\theta; X) = P(A_1 \wedge B_2 \wedge C_3 \wedge D_4 \ldots |\theta)$

If the events are independent, that is the chance of a flip doesn't depend on any other outcomes, then:

$L(\theta; X) = P(A_1|\theta).P(B_2|\theta).P(C_3|\theta).P(D_4|\theta)...$

If the events are identically distributed, the chance of flipping a head doesn't change across flips (for example the heads side doesn't get heavier over time) then:

$L(\theta; X) = P(A|\theta).P(B|\theta).P(C|\theta).P(D|\theta)...$

$L(\theta; X) = \prod_{i=1}^{n} P(X_i|\theta)$

Chapter 35

The score, Fisher information and orthogonality

35.1 Score functions

35.1.1 The score

The score is defined as the differential of the log-likelihood function with respect to θ.

$V(\theta, X) = \frac{\delta}{\delta\theta} l(\theta; X)$

$V(\theta, X) = \frac{1}{\prod_{i=1}^{n} P(X_i|\theta)} \frac{\delta}{\delta\theta} L(\theta; X)$

35.1.2 Expectation of the score

The expectation of the score, given the true value of θ is:

$E[V(X|\theta)] = \int V(X|\theta) dX$

$E[V(X|\theta)] = E[\frac{1}{\prod_{i=1}^{n} P(X_i|\theta)} \frac{\delta}{\delta\theta} L(\theta; X)]$

$E[V(X|\theta)] = \int \frac{1}{\prod_{i=1}^{n} P(X_i|\theta)} \frac{\delta}{\delta\theta} L(\theta; X)$

$E[\frac{1}{\prod_{i=1}^{n} P(X_i|\theta)}]$

$\int \frac{1}{\prod_{i=1}^{n} P(X_i|\theta)} P(\theta) d\theta$

We can show that the expected value of this is 0.

35.1.3 Variance of the score

The variance of the score is:

$var[\frac{\delta}{\delta\theta} l(\theta; X)]$

$var[\frac{1}{\prod_{i=1}^{n} P(X_i|\theta)}]$

35.2 Fisher information

35.2.1 Fisher information

The Fisher information is the variance:

$E[(\frac{\delta}{\delta\theta} \log f(X, \theta))^2 | \theta]$

$E[\frac{\delta^2}{\delta\theta^2} \log f(X, \theta) | \theta]$

Same as expectation of score squared, because centred around 0.

35.2.2 Fisher information matrix

We have k parameters.

$I(\theta)_{ij} = E[(\frac{\delta}{\delta\theta_i} \log f(X, \theta))(\frac{\delta}{\delta\theta_j} \log f(X, \theta)) | \theta]$

35.2.3 Observed Fisher information matrix

The Fisher information matrix contains informatio about the population

The observed Fisher infoirmation is the negative of the Hessian of the log likelihood.

We have:

- $l(\theta|\mathbf{X}) = \sum_i \ln P(\mathbf{x}_i|\theta)$
- $J(\theta^*) = -\nabla\nabla^T l(\theta|mathbfX)|_{\theta=\theta^*}$

The Fisher information is the expected value of this.

$I(\theta) = E[J(\theta)]$

35.3 Orthogonality

35.3.1 Orthogonality

Two variables are called orthogonal if their entry in fisher info matrix is 0

This means that the parameters can be calculated separately. MLE estimates are separate

This can be written as a moment condition

δ

Chapter 36

Quasi-likelihood functions

36.1 Quasi-likelihood function

36.1.1 Quasi-likelihood function

Chapter 37

Maximum Likelihood Estimation (MLE)

37.1 Maximising the likelihood function

37.1.1 Maximising the likelihood function

We have a likelihood function of the data.

$L(\theta; X) = P(X|\theta)$

We choose values for θ which maximise the likelihood function.

$argmax_\theta P(X|\theta)$

That is, for which values of θ was the observation we saw most likely?

This is a mode estimate.

37.1.2 IID

$L(\theta; X) = \prod_i P(x_i|\theta)$

37.1.3 Logarithms

We can take logarithms, which preserve stationary points. As logarithms are defined on all values above 0, and all probabilities are also above zero (or zero), this preserves solutions.

The non-zero stationary points of:

$\ln L(\theta; X) = \ln \prod_i P(x_i|\theta)$

$\ln L(\theta; X) = \sum_i \ln P(x_i|\theta)$

37.1.4 Example: Coin flip

Let's take our simple example about coins. Heads and tails are the only options, so $P(H) + P(T) = 1$.

$P(H|\theta) = \theta$

$P(T|\theta) = 1 - \theta$

$\ln L(\theta; X) = \sum_i \ln P(x_i|\theta)$

If we had 5 heads and 5 tails we would have:

$\ln L(\theta; X) = 5 \ln(\theta) + 5 \ln(1 - \theta)$

So $P(H) = \frac{1}{2}$ is the value which makes our observation most likely.

37.2 Properties of the MLE estimator

37.2.1 Asymptotic normality of the MLE

37.3 Results for specific distributions

37.3.1 MLE of the Gaussian distribution

The parameters are the population means and covariance matrix.

The MLE estimator for the mean is the sample mean.

The MLE estimator for the covariance matrix is the unadjusted sample covariance.

37.3.2 MLE of the Poisson distribution

37.3.3 MLE of the Bernoulli and binomial distributions

37.4 Other

37.4.1 Restricted Maximum Likelihood

We can partition out Likelihood functions, and include a part only with variance.

37.4.2 Targeted Maximum Likelihood Estimation

37.4.3 Scores

Existing score: rename Maximum Likelihood score

MLE bad if true theta not at where score is 0

Eg if one sided tails, true theta is not at MLE condition.

Can we find other scores?

37.4.4 Orthogonality

Score of one parameter depends on other parameters

If we misestimate one, then estimate another, will be bad answer

We want the score not to change around bad estimates

We want nuisance parameter bias not to affect score

separate page for orthogonality for sets of parameters. eg nuisance; of interest

Chapter 38

Maximum A-Priori (MAP) estimation

38.1 Maximum A-Priori Estimation

38.1.1 Maximum A-Priori (MAP) estimation

Mode estimate

$Argmax_\theta p(\theta|X)$

Using Bayes theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

So:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$Argmax_\theta p(\theta|X) = Argmax_\theta \frac{p(X|\theta)P(\theta)}{P(X)}$$

The denominator isn't affected so:

$Argmax_\theta p(\theta|X) = Argmax_\theta p(X|\theta)P(\theta)$

If $P(\theta)$ is a constant then this is the same as the MLE estimator.

Other

$Argmax_\theta p(\theta|X)$

Mode estimate

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

$$Argmax_\theta \frac{p(X|\theta)p(\theta)}{p(X)}$$

θ doesn't change denominator so can instead use:

$Argmax_\theta p(X|\theta)p(\theta)$

It is the same as maximum likelihood estimator if $p(\theta)$ is a constant.

38.1.2 MAP of the Gaussian distribution

Chapter 39

The Method Of Moments (MOM)

39.1 Method of Moments

39.1.1 Method of moments

Introduction

If we have k parameters to estimate, we can solve this if we have k equations.

We generate these

First, we link each first k moments to functions of the parameters.

Then we replace the momenets with sample estimates.

Estimation

The moments of this population distribution are:

$\mu_i = E[X^i] = g_i(\theta_1, ..., \theta_k)$

We have a sample.

$X = [X_1, ..., X_n]$

We now define the method of moments estimator

$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$

Chapter 40

Testing generative parameter estimates with Z-tests and T-tests

Chapter 41

Choosing parametric probability distributions

41.1 AIC

41.1.1 Introduction

41.2 AICc

41.2.1 Introduction

41.3 Bayes factor

41.3.1 Introduction

41.4 BIC

41.4.1 Introduction

41.5 Kullback-Leibler divergence

41.5.1 Kullback-Leibler divergence

Bayesian inference means we have full distribution of $p(w)$, not just moments of a specific point estimate

41.5.2 Cross entropy:

$H(P, Q) = E_P(I(Q))$

So for a discrete distribution this is:

$H(P,Q) = -\sum_x P(x) \log Q(x)$

Q is prior

P is posterior

41.5.3 Kullback-Leibler divergence

When we move from a prior to a posterior distribution, the entropy of the probability distribution changes.

$D_{KL}(P||Q) = H(P,Q) - H(P)$

KL divergence is also called the information gain.

41.5.4 Gibb's inequality

$D_{KL}(P||Q) \geq 0$

41.6 Bayesian model selection

41.6.1 Introduction

41.7 Cross entropy

41.7.1 Introduction

Chapter 42

Estimating population moments

42.1 Plug-in estimators

42.1.1 Estimating the population mean

42.1.2 Estimating the population variance

42.1.3 Estimating the population standard deviation

Part X

Stochastic methods

Chapter 43

Creating pseudo-random numbers

43.1 Pseudo random numbers

43.1.1 Seeds

43.1.2 Period

Chapter 44

Stochastic methods for integration

44.1 Introduction

Chapter 45

Stochastic optimisation

45.1 Random search

45.1.1 Random search

We start with a random set of parameters, x.

We then loop through the following:

- We define a search space local to our current selection.
- We randomly select a point from this space.
- We compare the new point to our current point. If the new point is better we move to that.

45.1.2 Random optimisation

This is similar to random search, however we use a multivariate Gaussian distribution around our current point rather than a hypersphere.

45.1.3 Simulated annealing

Introduction

We can use a version of Metropolis-Hastings to find the global maximum of a function $f(x)$.

We start with an arbitrary point x_0.

We move randomly from this to identiy a candidate point x_c.

We accept this with probability depending on the the relationship between x_0 and x_c.

This process will converge on the global maximum.

Hyperparameter

There is a hyperparameter for selection. At the extreme this becomes a greedy function.

45.2 Bayesian optimisation

45.2.1 Bayesian optimisation

Introduction

If we have sampled from the hyperparameter space we know something about the shape.

Can we use this to inform where we should next look?

The shape of the function is $y = f(\mathbf{x})$

We have observations $\mathbf{X}$ and $\mathbf{y}$.

So what's our posterior, $P(y|\mathbf{X}, \mathbf{y})$?

Exploration and exploitation

The can be a tradeoff between:

- Exploring - which gives us a better shape for $y = f(x)$; and
- Exploiting - which gives us a better estimate for the global optimum.

The surrogate function

We do not know $y = f(x)$, but we model it as:

$z(x) = y(x) + \epsilon$

We can then maximise z

Proposing new candidates

We want an algorithm which maps from our history of observations to a new candidate.

There are different approaches:

- Probability of improvement - Choosing one with the highest chance of a more optimal value
- Expected improvement - Choosing one with the biggest expected increase in the optimal value

- Entropy search - choosing one which reduces uncertainty about the global maximum.

45.3 Evolutionary algorithms

45.3.1 Evolutionary algorithms

Initialisation

We generate a set of candidate parameter values, x.

Evaluate using the fitness function

We evaluate each of these against a fitness function (the function we are optimising).

We assign fitness values to each individual.

Crossover and mutation

We generate a second generation. We select "parents" randomly using the fitness values as weightings.

The values of the new individual are a function of the values of the parents, and noise (mutation).

We do this for each member in the next generation.

We iterate this process across successive generations.

45.4 Differential evolution

45.4.1 Differential evolution

45.5 Particle swarms

45.5.1 Particle swarms

Chapter 46

Calculus of stochastic processes

46.1 Introduction

46.1.1 Ito integrals

46.1.2 Stochastic differential equations

Chapter 47

Lossy compression

47.1 Lossy compression

Chapter 48

Non-cryptographic hashes

48.1 Data integrity checks

48.1.1 Hash functions

Hash functions (take input and return fixed length output) (h=hash(m))

Data integrity checks

Needs to be very different for small changes. so typo has different hash for example. corrput data needs to be noticed.

Checksums

if two files are the same then hashes the same

Introduction

Want following properties for a hash function

Deterministic, so the same hash is always created.

Quick to compute hash

Cannot generate input from hash, except for brute forcing inputs

Small changes to document should cause large charges to hash, such that the two hashes appear uncorrelated

Can't find multiple documents with the same hash, practically.

Can be used to verify files, check passwords.

So possible vulernabilities are:

Given hash, find message (Pre-image resistence)

Given input, find another input with the same hash (second pre-image resistance)

Collission resistance (find two inputs with same hash)

We want to prevent accidental changes to file, and deliberate changes to file. Vulerabilites are more importnat for latter.

48.2 Example of non-cryptogrphic hash functions

48.2.1 Introduction

Part XI

Sampling

Chapter 49

Rejection sampling

49.1 Direct sampling

49.1.1 Density estimation through direct sampling

I THINK THE STUFF HERE IS LIMITATIONS TO REJECTION SAMPLING??

DIRECT SAMPLING IS DOING PHYSICAL SAMPLES, MANUALLY PICKING BALLS FROM URL ETC?

There is distribution $P(x)$ which we want to know more about.

If the function was closed, we could estimate it by using values of x.

49.1.2 Limitations of direct sampling

However if the function does not have such a form, we cannot do that.

We can't plug in values, because the function is complex.

Sometimes we may know a function of the form:

$f(x) = cP(x)$

That is, a multiple of the function.

This can happen from Bayes' theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

We may be able to estimate $P(x|y)$ and $P(y)$, but not $P(x)$

This means be have

$P(y|x) = cP(x|y)P(y)$

49.2 Acceptance-rejection sampling

49.2.1 Introduction

Used to sample from propability distribution function.

Useful when can't use direct sampling, because no closed form.

MORE GENERALLY FRAME THESE FIRST AS SAMPLING FROM PROBABILITY FUNCTION.

Generate pairs of (x, y). If $y < P(x)$ then keep x.

Metropolis-Hastings and Gibb's sampling are extensions of this.