Estimators

Key distribution used:

$$\overline{\mathbb{P}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_i \quad \text{, where } \mathbb{P}_1 = \mathbb{P}_2 \dots = \mathbb{P}_n$$

(NOTE: $\mathbb{P}_n = \sum_{i=1}^{n} \mathbb{P}_i$)

↳ This is the sample mean distribution. It models the distribution ~~from~~ of sample means from n independent samples drawn from a common distribution $\mathbb{P}$.

Properties of sample mean distr.:

1) $\mu(\overline{\mathbb{P}_n}) = \mu(\mathbb{P})$

2) $\text{Var}(\overline{\mathbb{P}_n}) = \frac{\text{Var}(\mathbb{P})}{n}$

The purpose of sample mean is to estimate the actual distribution mean.
i.e. sample mean is an estimator of the distribution's mean.

↳ To generalise...

An estimator ~~is~~ is a collection (i.e. a map for each n value) of maps

$$T^n : X^n \to \mathbb{R}$$

implementing some operation tuples (size n) of samples

The purpose of an estimator is to estimate some real value associated with probability measures on X (ex. mean, variance, etc.)

In statistics, we study the distr. of estimators applied to ~~products~~ the product measure

$$\underbrace{\mathbb{P} \otimes \mathbb{P} \otimes \dots \mathbb{P}}_{n \text{ times}} \quad \text{on} \quad X^n$$

↳ (i.e. n independently & identically distr.ed samples)

2 properties of estimators:

1) Unbiasedness: on average, they estimate correctly, i.e. if an estimator T^n is meant to estimate parameter $\phi(\mathbb{P})$ of a ~~distn.~~ prob. measure $\mathbb{P}$, then T^n is unbiased if

$$\mu(T^n_*(\underbrace{\mathbb{P}\otimes\dots\mathbb{P}}_{n\text{ times}})) = \phi(\mathbb{P})$$

$\Rightarrow$ distn. mean

2) Consistency: as no. of samples grow, estimator converges to true value of the parameter it is meant to estimate, i.e. T^n is consistent if

$$\lim_{n\to\infty} T^n_*(\underbrace{\mathbb{P}\otimes\dots\mathbb{P}}_{n\text{ times}})\left([\phi(\mathbb{P})-\varepsilon, \phi(\mathbb{P})+\varepsilon]\right) = 1$$

where the expression within the limit is the prob. mass of the interval $[\phi(\mathbb{P})-\varepsilon, \phi(\mathbb{P})+\varepsilon]$ w.r.t. the estimator's distn. & where ε is an arbitrarily small +ve value.

Some results

1) Sample mean is unbiased

2) Biased sample variance i.e.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \text{ for sample } (x_1, x_2 \dots x_n)$$

is biased

Limit Laws

1) Law of Large numbers

1.1) Weak law

If $\mathbb{P}$ is a prob. measure with mean $\mu(\mathbb{P})$, then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \overline{\mathbb{P}}_n\left(\left[\mu(\mathbb{P}) - \varepsilon, \mu(\mathbb{P}) + \varepsilon\right]\right) = 1$$

i.e. sample mean is a consistent estimator of true (distr.) mean.

> NOTE: WLLN presupposes that $\mu(\mathbb{P})$ actually exists, it cannot be applied else.

1.2) Strong law

If $\mathbb{P}$ is a distr. with mean $\mu(\mathbb{P})$, then

$$\mathbb{P}^\infty\left(\left\{(x_1, x_2 \ldots) \,\middle|\, \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \mu(\mathbb{P})\right\}\right) = 1$$

where

$\mathbb{P}^\infty = \mathbb{P} \otimes \mathbb{P} \ldots \infty$

$(x_1, x_2 \ldots x_n)$ = samples, where $n \to \infty$

i.e. the probability mass of an infinite set of samples whose finite means converge to the true mean is 1 (provided the samples are independently & identically distr.ed)

↳ All ~~??~~ IID ~~sa~~ infinite samples have this property.

> NOTE: SLLN also presuppose that $\mu(\mathbb{P})$ actually exists.

---

2) Central limit theorem

If $\mathbb{P}$ is a prob. measure with mean $\mu(\mathbb{P})$ & variance $\mathrm{Var}(\mathbb{P})$, then

$$\lim_{n \to \infty} \sqrt{n}\left(\overline{\mathbb{P}}_n - \mu(\mathbb{P})\right)((-\infty, t]) = N(0, \mathrm{Var}(\mathbb{P}))((-\infty, t])$$

In other terms

$$\sqrt{n}\,(\overline{IP}_n - \mu(IP)) \approx N(0, Var(IP))$$

as $n \to \infty$

Represents the prob. distr. of the deviation of sample means from true mean, multiplied by $\sqrt{n}$.

We can rearrange the terms to get:

a) $$\frac{\sqrt{n}\,(\overline{IP}_n - \mu(IP))}{\sqrt{Var(IP)}} \approx N(0, 1) \text{ as } n \to \infty$$

b) $$\overline{IP}_n \approx N\left(\mu(IP), \frac{Var(IP)}{n}\right) \text{ as } n \to \infty$$

Hypothesis testing

Basics

For given data X, we define an estimator to serve a required purpose (ex. if we want to test for distn. mean, the estimator is sample mean). This estimator or some function of it is used as the test statistic

test statistic

Then, we either identify or approximate the ~~estimator's~~ probability distn. Using this, we can see how likely or unlikely is the particular value we obtain for the estimator for the given data X.

We formulate such a test by formulating the appropriate hypotheses...

Null hypothesis (H_0): The assumption we want to test. Called "null" because it is the assumption of no difference btw. assumed & true value. (of a parameter) ↳ (or estimated)

↳ We reject H_0 if our observation is deemed sufficiently unlikely.

Defining "sufficiently unlikely"

We define a parameter α, the "confidence ~~interval~~ level". This is ~~a proportion~~ the probability mass of some proportion (indicated by α itself, naturally) of the most likely / high-probability / high-density values of the distn. of the test statistic.

ex. $\alpha = 90\% \Rightarrow$ the prob. mass of 90% of the most likely values of the distn.

↳ ~~$\alpha = 90\%$~~ $\Rightarrow$ if an observed value lies outside this range of values, we will consider it sufficiently unlikely

Usually, α is chosen as 95%, 99%, 99.9% etc etc.

Critical region

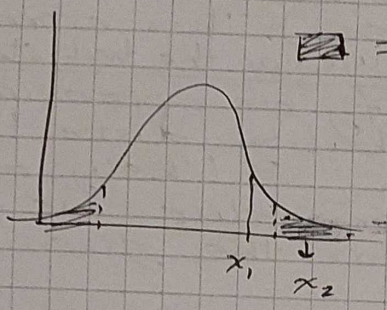
The least likely values of the distn. that together have a mass of $1-\alpha$ (called the significance level) form the critical region.

i.e. they are the $1-\alpha$ ~~least likely~~ proportion

of least likely values in the distr.

If an observation lies in the critical region, it is considered sufficiently unlikely.

Visual example:



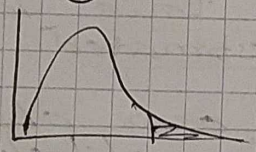
▨ ⇒ critical region

$x_2 \in$ Crit. region
⇒ reject H_0

$x_1 \notin$ Crit-region
⇒ cannot reject H_0.

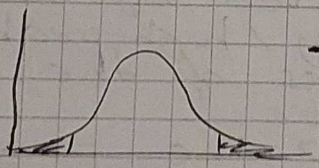
NOTE: Depending on the type of distribution the $1-\alpha$ proportion of least likely (i.e. most extreme) values may either:

i) lie only on one side of the distr.

ex.  → requires one-sided test

↳ most extreme values (at the moment proportion) lie only to the right
↳ (at a higher proportion, we may begin to include from left also).

ii) lie on either side of the distr.

ex.  → requires two-sided test

↳ is the case for any symmetric distr.

---

p-value (to critical region) + logically

An alternative ↑ (& practically equivalent) method to test extremeness of an observation.

p-value = probability mass ↑ (under the assumption that H_0 is true)
↳ (i.e. assuming a distr. for test statistic)

of all values at least as extreme (i.e. unlikely) as the observed value.

NOTE:

For ~~one~~ two-sided test,
if p-value $< \frac{1-\alpha}{2}$, reject H_0

For ~~two~~ one-sided test,
if p-value $< 1-\alpha$, reject H_0.

↳ This is if we calculate p-value as $\min(CDF(t), 1-CDF(t))$ (where t is observed value)

↳ If we instead calculate p-value as $2\min(CDF(t), 1-CDF(t))$ for two-sided test only, then

if p-value $< 1-\alpha$, reject H_0.

WHY? Consider: what does p-value mean? What set of values would be consider "at least as extreme" as the observed value in two-sided test? Think about it...

Testing errors

Type 1: Rejecting H_0 even if it holds.

By definition, the probability of type 1 error is $1-\alpha$ (i.e. the significance level)

Type 2: Failing to reject H_0 even if it does not hold

Prob. of committing type 2 is called β. It is harder to derive.

NOTE: We can never validly say we "accept" H_0, only that we "fail to reject" it.

WHY? Consider: what kind of evidence does statistical testing give? Can it give positive ~~ev~~ evidence for something, or only a lack of negative evidence against something?

NOTE: H_0 either explicitly or implicitly hypothesis that a test statistic follows a certain distr. IP.

Exact test: if exact distr. of test statistic is known

Approximate test: if we can only approx. the distr. of the test statistic.

Approximate test

Preliminary concepts: useful distributions (helpful in approximating test distr.)

1) $N(0,1)$ → standard normal distr.

2) $\chi^2(k)$ → Chi squared distr. with k degrees of freedom (df)

NOTE: (extra info)

$$\sum_{i=1}^{n} \left(N(0,1) - \overline{N(0,1)}_n \right)^2 = \chi^2(n-1)$$

3) $t(k)$ → Student's t-distr. with k df

NOTE: (useful later)

$$t(k) = \frac{N(0,1)}{\sqrt{\frac{\chi^2(k)}{k}}}$$

Preliminary concepts: key facts

1) By the CLT (central limit theorem):

$$\frac{\sqrt{n}\left(\overline{IP}_n - \mu(IP)\right)}{\sqrt{Var(IP)}} \approx N(0,1) \quad \text{as } n \to \infty$$

2) $$\frac{SVar_n(IP)}{Var(IP)} \approx \frac{\chi^2(k)}{k}$$

where $k = \frac{2n}{c - \frac{n-3}{n-1}}$ (c is some constant)

$SVar_n IP$ = distr. of sample variance of a sample of size n drawn from IP

where c = Kurtosis (IP)

NOTE: $SVar_n(IP)$ is a distr., but $Var(IP)$ is a single value!

To present more clearly: Here, $k = \frac{2n}{c - \frac{n-3}{n-1}}$, where c = Kurtosis (IP)

3) (1) & (2) $\Rightarrow$

$$\frac{\frac{\sqrt{n}\left(\bar{P}_n - \mu(P)\right)}{\sqrt{Var(P)}}}{\sqrt{\frac{SVar_n(P)}{Var(P)}}} \approx \frac{N(0,1)}{\sqrt{\frac{\chi^2(k)}{k}}}$$

IMPORTANT

where

$$k = \frac{2n}{c - \frac{n-3}{n-1}}$$

(where c is a constant, n is the sample size)

where $c = \text{Kurtosis}(P)$

$$\approx t(k)$$

$$\Rightarrow \frac{\sqrt{n}\left(\bar{P}_n - \mu(P)\right)}{\sqrt{SVar_n(P)}} \approx t(k)$$

↳ Notice that $Var(P)$ is cancelled out; this is useful if $Var(P)$ is unknown but we want to test for mean.

NOTE: Estimating the above df i.e. k.

In general, we may estimate

$df = n$ (sample size)

with reasonable accuracy, if Kurtosis (P) is unknown.

Uses of the above:

- z-test: using test statistic in (1)
  → testing mean with known variance
- χ^2-test: using test statistic in (2)
  → testing ~~sample~~ variance
- t-test: using test statistic in (3)
  → testing mean with unknown variance.

Goodness of fit

(comes under hypothesis testing)

Preliminary concept: Distance btw. distrs

Formally, a distance function (defined on $X \times Y$ where X & Y are 2 sets) should ~~fulfil~~ fulfil the following:

- ~~$d(x,x)$~~ dist.$(x,x) = 0$
- dist(x,y) = dist(y,x)
- dist$(x,z) \leq$ dist(x,y) + dist(y,z)

Distance (as general concept) serves as a measure similarity or dissimilarity:
higher dist. ⇒ lower similarity
lower " ⇒ higher "

OUR AIM: Measure i.e. quantify the similarity (or dissimilarity) btw. 2 distrs.

GENERAL IDEA TO MEASURE DISTANCE BTW DISTRS:

Two distributions are close if the compute similar integrals.

(NOTE: One of the main things we do with prob. distrs is integrate them i.e. obtain expectations.)

Generalising with "test function":

We can compute the integrals along with different functions that would focus on different aspects ~~or~~ of the distr. → (unclear, need to clarify!)

↳ There are "test functions"

Suppose distrs P_1 & P_2 have densities f_1 & f_2, then P_1 & P_2 are close if

$\int g(x) f_1(x)\,dx$ ~~dist~~ $\overset{\text{is}}{\text{close}}$ to $\int g(x) f_2(x)\,dx$

for all test functions g.

Measuring distances btw. distrs using integrals

Suppose distrs $\mathbb{P}_1$ & $\mathbb{P}_2$ have densities f_1 & f_2. Then, distance may be calculated in one of the following ways:

1) Max. value of $\left| \int g(x) f_1(x) dx - \int g(x) f_2(x) dx \right|$ over all test functions, i.e.

$$\text{dist}(\mathbb{P}_1, \mathbb{P}_2) = \sup_g \left| \int g(x) f_1(x) dx - \int g(x) f_2(x) dx \right|$$

"Integral probability pseudo-metric"

2) Weighted avg. of $\left[\int g(x) f_1(x) dx - \int g(x) f_2(x) dx \right]^2$ over all test functions.

"Quadratic EDF statistics"

Preliminary concept: Empirical PDF, CDF & probability measure

Empirical PDF:

Suppose we have

- n samples $x_1, x_2 \ldots x_n$ from distr. $\mathbb{P}$

then → Any density-histogram we build from these is the empirical density from distr.

↳ For this, choose a sequence of k bins $y_1 < y_2 < \ldots y_k$
(a good choice could be $k = \sqrt{n}$)

↳ Then, we define the empirical probability density function as

$$f_{emp}(x) = \sum_{i=1}^{k-1} 1_{[y_i, y_{i+1}]}(x) \cdot \left(\frac{\sum_{j=1}^{n} 1_{[y_i, y_{i+1}]}(x_j)}{n} \right)$$

NOTE: $1_{[y_i, y_{i+1}]}(x) = 1$ if $x \in [y_i, y_{i+1}]$, 0 if $x \notin [y_i, y_{i+1}]$

Also, $\sum_{j=1}^{n} 1_{[y_i, y_{i+1}]}(x_j)$ = (Count) of all samples belonging to bin $[y_i, y_{i+1}]$

(Proportion)

+NOTE: As $n \to \infty$, we expect ~~empirical~~ empirical PDF to converge to actual PDF of $\mathbb{P}$.

CONSIDER: Empirical PDF's define prob. measures. Why?

Empirical CDF

As before, suppose we have

- n samples $x_1, x_2 \ldots x_n$ from distr. $\mathbb{P}$

↳ Then, the unique empirical CDF ~~for~~ w.r.t. these samples is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[x_i, \infty)}(x)$$

CONSIDER: what does empirical ~~probability~~ cumulative prob. mean? It means you add up the empirical ~~probab~~ proportional frequencies i.e. the proportion of the

Essentially equals the proportion of ~~times~~ values the no. of times ~~value~~ samples ~~at least~~ less than or equal to x appeared in the sample data set.

+NOTE: As $n \to \infty$, we expect F_n to converge to the actual CDF of $\mathbb{P}$.

Empirical measure

Again, suppose we have

- n samples $x_1, x_2 \ldots x_n$ from distr. $\mathbb{P}$.

Also, define ~~δ_{x_i}~~ $\delta_x(A) = \begin{cases} 1, & \text{if } x \in A \\ 0 & \text{else} \end{cases}$

For a single value, $\delta_x(a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{else} \end{cases}$

∴ Then, the unique ~~empirica~~ empirical prob. measure corresponding to the samples is

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

Hence, $\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(A)$

Clearly, we see that

$$\mathbb{P}_n(A) = \frac{\text{No. of samples from } x_1, x_2 \ldots x_n \text{ that are present in } A}{n}$$

+NOTE: As $n \to \infty$, $\mathbb{P}_n \to \mathbb{P}$

+NOTE: It is easy to show that empirical ~~CD~~ CDF can be derived from $\mathbb{P}_n$.

Main topic: Goodness of fit tests

Discrete case

- ~~Test~~ To evaluate distance between ~~true~~ theoretical & empirical distr. :
- Test function used: $1_t(x) = \begin{cases} 1 & \text{if } x = t \\ 0 & \text{else} \end{cases}$
- Distance method used:
  Weighted avg. of

$$\left(\int g(x) f_1(x)\,dx - \int g(x) f_2(x)\,dx\right)^2$$

NOTE: In discrete case, the "integral" is essentially summation.

- Consider $\{z_1, z_2 \ldots z_k\}$ as the ~~exhaustive~~ support of the theoretical distr, $\mathbb{P}$.
- Consider ~~a sample x~~ n samples $x_1, x_2 \ldots x_n$ drawn from $\mathbb{P}$

- ~~Let~~ f Let f be the PMF of $\mathbb{P}$. Then, we can obtain the probabilities of each point of the support as

$$\sum_{i=1}^{k} 1_{z_k}(z_i) f(z_i) = \mathbb{P}(z_k) \quad \ldots (1)$$

- From $x_1, x_2 \ldots x_n$, we build the empirical measure

(turn to next page)

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$ → also the ~~PMF~~ empirical PMF

Hence, we have defined

- The theoretical distr.
- The empirical distr.
- The test function

"Integral" of test function w.r.t. $\mathbb{P}$ is in fact $\sum_{i=1}^{k} 1_{z_k}(z_i) f(z_i) = \mathbb{P}(z_k)$

Further, "integral" of test function w.r.t. $\mathbb{P}_n$ is in fact

$$\sum_{i=1}^{k} 1_{z_k}(z_i) \mathbb{P}_n(z_i) = \mathbb{P}_n(z_i)$$

Hence we have the "integrals" needed to calculate distance btw. $\mathbb{P}_n$ & $\mathbb{P}$.

But weighting scheme?

↳ we choose a scheme s.t. the distr. of the distance can be approximated using central limit theorem (CLT)

Hence, we define

$$\text{dist}(\mathbb{P}, \mathbb{P}_n) = \sum_{i=1}^{k} \frac{1}{\mathbb{P}(z_i)} \left(\mathbb{P}(z_i) - \mathbb{P}_n(z_i)\right)^2$$

NOTE: This is for discrete case

$\frac{1}{\mathbb{P}(z_i)}$ → weighting scheme

$\left(\mathbb{P}(z_i) - \mathbb{P}_n(z_i)\right)^2$ → difference of "integrals"

Results on the above computed distance measure for discrete case:

1) As $n \to \infty$, $\text{dist}(\mathbb{P}, \mathbb{P}_n) \to 0$

(is derived from the next theorem)

2) Theorem (Pearson):

$n \cdot \text{dist}(\mathbb{P}, \mathbb{P}_n) \to \chi^2(k-1)$ as $n \to \infty$

NOTE: k = size of support of $\mathbb{P}$

↳ Hence, we can approximate ~~this~~ sample distances' distribution!

The goodness of fit test tests how well the distr. $\mathbb{P}$ agrees with the empirical distr. of the drawn samples i.e. $\mathbb{P}_n$.

This Thanks to the previous results, we can approximate how the distance btw. theoretical & empirical distr.s are distributed in general & thus, we can evaluate the plausibility of observed cases (i.e. observed empirical distr.s).

H_0: The samples $x_1, x_2 \dots x_n$ are drawn from $\mathbb{P}$

Under H_0, $n \cdot \text{dist}(\mathbb{P}, \mathbb{P}_n) \approx \chi^2_{k-1}$

H_0 is rejected if $n \cdot \text{dist}(\mathbb{P}, \mathbb{P}_n)$ is too large i.e. too unlikely.

Since this test is based on Pearson's theorem, it is called

Pearson's χ^2-test

NOTE: χ^2-test $\Rightarrow$ one-sided test (right-tailed) $\Rightarrow$ p-value $= \chi^2([n\,\text{dist}(\mathbb{P}, \mathbb{P}_n), \infty))$

(Goodness of fit, next topic...)

Continuous case

To evaluate distance btw. theoretical & empirical distr.:

- Test function used: ~~$1_{(-\infty, t]}$~~ (rather, the class of test functions used)

$$g_t(x) = 1_{(-\infty, t]}(x)$$

- Distance method used: Since we need to deal with a class of test functions, we use max. value of

$$\left| \int g_t(x) f_1(x) - \int g_t(x) f_2(x)\, dx \right|$$

(maximised using t) dx

Now, note that

$$\int g_t(x) f(x)\,dx = \int_{-\infty}^{t} f(x)\,dx = F(t)$$

(f: density func. of $\mathbb{P}$ (theoretical distn.); $F(t)$: theoretical CDF)

Similarly, note that

$$\int g_t(x) f_{emp}(x)\,dx = \int_{-\infty}^{t} f_{emp}(x)\,dx = F_{emp}(t)$$

(f_{emp}: empirical PDF)

Hence, we get the distance btw. theoretical & empirical distn.s as

$$dist(\mathbb{P}, \mathbb{P}_n) = \sup_t |F(t) - F_{emp}(t)|$$

[Oops, F_{emp} is notated as F_n !]

$$dist(\mathbb{P}, \mathbb{P}_n) = \sup_t |F(t) - F_n(t)|$$

Results for distances btw. distn.s in continuous case:

1) Glivenko-Cantelli theorem:

$$dist(\mathbb{P}, \mathbb{P}_n) \to 0 \text{ as } n \to \infty$$

2) Kolmogorov theorem:

$$\sqrt{n}\, dist(\mathbb{P}, \mathbb{P}_n) \to K \text{ as } n \to \infty$$

(K here is the Kolmogorov distn.)

← Just like for discrete case, now we can do hypothesis testing (goodness of fit testing) for continuous distn. too.

This test is the Kolmogorov-Smirnov (KS) test, wherein ~~we~~

KS-statistic $= \sqrt{n}\, dist(\mathbb{P}, \mathbb{P}_n)$

p-value $= K([\sqrt{n}\, dist(\mathbb{P}, \mathbb{P}_n), \infty))$

↳ also a one-sided right tailed test.

NOTE: H_0 is the same as before (except for distribution used)

H_0: Samples $x_1, x_2, \ldots x_n$ were drawn from $\mathbb{P}$ (under H_0, $\sqrt{n}\, dist(\mathbb{P}, \mathbb{P}_n) \approx K$)

Independence, dependence & linear dependence

Preliminary concept: Joint distributions (or distr.s over product spaces)

A product space is the ~~cross~~ product of 2 or more sets, i.e. $X_1 \times X_2 \cdots \times X_n$. (Cartesian)

A distribution on a product space ~~of~~ is often called a joint distribution.

NOTE: We have seen product measures, ex. $\mathbb{P}_1 \otimes \mathbb{P}_2 (X_1 \times X_2)$, which are a kind of joint distr..

~~Basic~~ Independence

Product measure captures the idea of independence. For example

$\mathbb{P}_1 \otimes \mathbb{P}_2 (A \times B) = \mathbb{P}_1(A)\,\mathbb{P}_2(B)$

= Prob. that the random experiment modelled by $\mathbb{P}_1$ returns outcome A & Prob. that the random experiment modelled by $\mathbb{P}_2$ returns outcome B

One outcome does not affect the other, & order of outcomes / of observing the outcomes does not change the joint probability above.

i.e. no information flows btw. the experiments.

KEY POINT TO REITERATE: All product measures are joint distr.s, but not all joint distr.s are product measures.

Hence, $\mathbb{P}(A \times B) \neq \mathbb{P}_1(A)\,\mathbb{P}_2(B)$ for any $\mathbb{P}_1$ & $\mathbb{P}_2$.

In general, we cannot assume independence in a joint distr.

Perfect dependence

Consider the "copy" function

$$\text{copy}(x) = (x, x), \quad \forall x \in X$$

i.e. $\text{copy} : X \to X \times X, \quad x \mapsto (x, x)$

We can use this to create a perfectly dependent joint distr.:

$$\mathbb{P}(\{(x,x) \mid x \in X\})$$

> SIDE NOTE: Alternatively, the above joint distr. is representable as
>
> $$\text{copy}_* \mathbb{P}(A) = \mathbb{P}(\{x \mid (x,x) \in A\})$$

Marginals (key concept)

Any product space $X_1 \times X_2 \ldots X_n$ comes with n special maps called projections:

~~$\pi_1 : X_1 \times X_2 \ldots X_n \to X_1, (x_1, x_2 \ldots x_n) \to x_1$~~

~~$\pi_2 : X_2 \times X_2 \ldots X_n \to X_2, (x_1, x_2 \ldots x_n) \to x_2$~~

$\pi_1 : X_1 \times X_2 \ldots X_n \to X_1, (x_1, x_2 \ldots x_n) \to x_1$

$\pi_2 : X_2 \times X_2 \ldots X_n \to X_2, (x_1, x_2 \ldots x_n) \to x_2$

etc. upto π_n

Basically, projection i i.e. π_i maps any given tuple from the product space to the ith coordinate of the tuple i.e. $\pi_i((x_1, x_2 \ldots x_n)) = x_i$

(It's basically a way to pick a particular coordinate from a tuple of the product space.)

The marginals of a joint distr. $\mathbb{P}$ on $X_1 \times X_2 \ldots \times X_n$ are the pushforward of the projections

~~i.e.~~ (turn to next page)

i.e. marginals are measures on particular axes of the product space, obtained via the ~~joint distro's mea~~ measure defined for the product space as a whole.

Hence, if $\mathbb{P}$ is a measure on the whole product space $X_1 \times X_2 \dots X_n$, then,

$\pi_1 * \mathbb{P}$ is a measure on X_1

$\pi_2 * \mathbb{P}$ is a measure on X_2

etc. upto $\pi_n * \mathbb{P}$

Hence, if $A \subseteq X_i$, then

$\pi_i * \mathbb{P}(A) = \mathbb{P}(X_1 \times X_2 \dots X_{i-1} \times A \times X_{i+1} \times \dots X_n)$

Basically, it's the joint distro. where you keep ~~all but~~ one coordinate as constant. ~~all others~~

Illustrative example:

Consider product space ~~$\{0,1\} \times \{0,1\}$, $A \times A$, where A~~ $= X \times X$, where $X = \{0,1\}$. Consider the following joint distr.:

$\mathbb{P}(0,0) = \frac{1}{4}$, $\mathbb{P}(0,1) = \frac{1}{4}$, $\mathbb{P}(1,0) = \frac{1}{2}$ and $\mathbb{P}(1,1) = 0$

Then, $\pi_1 * \mathbb{P}$ is a measure defined on $\{0,1\}$, s.t.

$$\begin{aligned} \pi_1 * \mathbb{P}(0) &= \mathbb{P}(0 \times X) \\ &= \mathbb{P}(0 \times \{0,1\}) \\ &= \mathbb{P}(\{(0,0),(0,1)\}) \\ &= \mathbb{P}(0,0) + \mathbb{P}(0,1) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

(Essentially, we have defined the prob. of observing 0 in the 1st coordinate)

Similarly, $\pi_1 * \mathbb{P}(1) = \mathbb{P}(1,0) + \mathbb{P}(1,1)$
$= 1/2 + 0 = 1/2$

— Similarly we can define $\pi_2 * \mathbb{P}$ —

Continuous example...

Let $\mathbb{P}$ be a ~~product~~ joint measure defined on $\mathbb{R} \times \mathbb{R}$, with its density being $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Now, consider $\pi_1 * \mathbb{P}$ defined for a specific interval $[a, b] \subseteq \mathbb{R}$

$$\pi_1 * \mathbb{P}([a,b]) = \int_{-\infty}^{\infty} \int_{a}^{b} f(x,y)\, dx\, dy$$

i.e. x is integrated over range $[a,b]$ (1st coordinate) while y is integrated over range $(-\infty, \infty)$ (no bounds).

Some results on marginals:

1) Theorem: Let $\mathbb{P}$ on $\mathbb{R} \times \mathbb{R}$ have density $f(x,y)$. Then,

$$\mu_x = \mu(\pi_1 * \mathbb{P}) = \int\int x f(x,y)\, dx\, dy$$

↳ mean of 1st marginal's distr. *

2) Similarly, we have

$$\mu_y = \mu(\pi_2 * \mathbb{P}) = \int\int y f(x,y)\, dx\, dy$$

Results: Marginals for independent & perfectly dependent joint distrs

For independent measure $\mathbb{P}_1 \otimes \mathbb{P}_2$ on $X \times Y$,

$$\pi_1 * (\mathbb{P}_1 \otimes \mathbb{P}_2)(A) = \mathbb{P}_1 \otimes \mathbb{P}_2 (A \times Y)$$
$$= \mathbb{P}_1(A)\, \mathbb{P}_2(Y)$$
$$= \mathbb{P}_1(A)$$

i.e. total prob. of A.

Similarly,

$\pi_2 * (\mathbb{P}_1 \otimes \mathbb{P}_2)(B) = \mathbb{P}_2(B)$

Hence, marginals of a product measure are just its components.

For perfectly dependent distrs, $\text{copy}_* \mathbb{P}$ defined on $X * X$,

$\pi_1 * (\text{copy}_* \mathbb{P})(A) = \mathbb{P}(A)$

$\pi_2 * (\text{copy}_* \mathbb{P})(A) = \mathbb{P}(A)$

(the 1st & 2nd coordinates are perfectly dependent, so the total prob. for each is equal)

Linear dependence

Preliminary: Covariance, correlation

Theoretical ~~covariance~~ covariance:

Discrete case:

$$\text{Cov}(\mathbb{P}) = \sum_x \sum_y (x - \mu_x)(y - \mu_y) f(x, y)$$

NOTE: $\mathbb{P}$ is the joint distr., μ_x & μ_y are the means of $\pi_1 * \mathbb{P}$ & $\pi_2 * \mathbb{P}$ respectively (i.e. 1st & 2nd marginals) & f is the density of $\mathbb{P}$.

Continuous case

$$\text{Cov}(\mathbb{P}) = \iint (x - \mu_x)(y - \mu_y) f(x, y) \, dx \, dy$$

+ Special cases:

~~Cov. for independent joint distr.:~~

$\text{Cov}(\mathbb{P}) =$

Key result:

$$\text{Cov}(\mathbb{P}) = \iint xy f(x, y) \, dx \, dy - \mu_x \mu_y$$

+ Corollary: $\text{Cov}(\mathbb{P}_1 \otimes \mathbb{P}) = 0$ ↳ independent joint distr.

+ ~~Special~~
+ Corollary: $Cov(copy_* \mathbb{P}) = Var(\mathbb{P})$
↓ perfectly dependent joint distr.

Sample covariance

$$SCov = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where we have n samples $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$
& SCov is the sample cov. for them.

NOTE: For $(x_1, x_1), (x_2, x_2) \ldots (x_n, x_n)$,
clearly, SCov = Unbiased sample variance.

+ POINTS:

$SCov_* (\mathbb{P} \underbrace{\otimes \ldots \otimes}_{n \text{ times}} \mathbb{P})$ (where $\mathbb{P}$ is a joint distr., & $\mathbb{P} \otimes \ldots \otimes \mathbb{P}$ is the distr. of n-sized tuples of samples each drawn from $\mathbb{P}$)
gives the distr. of sample cov.s when samples are drawn from $\mathbb{P}$.

+ NOTE: Sample cov. is unbiased estimator of cov.

Correlation

It is essentially "standardised" covariance...

$$Corr(\mathbb{P}) = \frac{Cov(\mathbb{P})}{\sigma_x \sigma_y}, \text{ where}$$

$\sigma_x^2 = Var(\pi_1 * \mathbb{P})$ (variance of 1st marginal)
$\sigma_y^2 = Var(\pi_2 * \mathbb{P})$ (" " 2nd ")

NOTE:

$$\sigma_x^2 = \sum\sum (x-\mu_x)^2 f(x,y) \cdot \Big\} \text{ discrete case}$$
$$\sigma_y = \sum\sum (y-\mu_y)^2 f(x,y) \Big\}$$

$$\sigma_x^2 = \iint (x-\mu_x)^2 f(x,y)\,dx\,dy \Big\} \text{ continuous case}$$
$$\sigma_y^2 = \iint (y-\mu_y)^2 f(x,y)\,dx\,dy \Big\}$$

Results on correlation:

1) $-1 \leq Corr(\mathbb{P}) \leq 1$

2) $Corr(\mathbb{P}_1 \otimes \mathbb{P}_2) = 0$

3) $Corr(copy_* \mathbb{P}) = 1$

NOTE: No dependence $\Rightarrow$ no correlation, but no correlation does not imply no dependence (ex. dependence may be non-linear)

Sample correlation

$$SCorr = \frac{SCov}{\sqrt{SVar_x \cdot SVar_y}}$$

$SVar_x$: sample variance of 1st coordinate (x's); $SVar_y$: sample variance of 2nd coordinate (y's)

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

+ POINTS: $SCorr_*(\mathbb{P} \otimes \overset{n \text{ times}}{\cdots} \otimes \mathbb{P})$ (where $\mathbb{P}$ is joint distr.) is distr. of sample correlation for n-sized samples drawn from $\mathbb{P}$.

Linear dependence

Consider the function

$$lin_{a,b}(x) = (x, ax+b)$$

(it is a generalisation of the previous copy function; copy is basically $lin_{0,1}$)

Now, consider a distribution $\mathbb{P}$ with support X.

Then, $lin_{a,b} * \mathbb{P}$ defines a joint distr. with support

$$\{ (x, ax+b) \mid x \in X \}$$

~~$lin_{a,b}\mathbb{P}$~~ $lin_{a,b} * \mathbb{P}$ is a joint distr. displaying linear dependence, wherein one coordinate in a ~~tuple~~ pair is a linear function of the other.

Results on linear dependence:

1) Theorem: $Cov(lin_{a,b} * \mathbb{P}) = a Var(\mathbb{P})$

2) Theorem: $Corr(lin_{a,b} * \mathbb{P})$

$$= \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$