Statistical Methods

Data sets

Data set is typically a **Sample, S**

Normally a subset of a bigger **population, P**

Depend on some underlying **distribution, p**

Individual events are **Independent and Identically distributed, (i.i.d)**

P follows some "true" distribution

"Infer" or "estimate" props of P from Sample S.

Events S and p are drawn $P(\vec{x})$ with no dependence → **i.i.d**

Histograms:

Collection of i.i.d events $X_i = \{x_1, x_2, \dots, x_n\}$

→ vis w/ hist predict bin edges

Cont events in each bin : $l_b \leq x < h_b$

Bin Contents:

$$N_b = \sum_i^N \begin{cases} 1 & \text{if } l_b \leq x < h_b \\ 0 & \text{otherwise} \end{cases}$$

Error: $\sigma_b = \sqrt{N_b}$ Poisson distribution

Can Scale Counts with a weight

$$N_b = \sum_i^N \begin{cases} w_i & \text{if } l_b \leq x \leq h_b \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_b = \sqrt{\sum_{i \in b} w_i^2}$$

Look into Seaborn.

Histograms as probability distributions

Scaled to give prob density $\sum_b w_b p_b = 1$

Quantifying Data

Data Sample, S with N i.i.d events $\{x_1, x_2, x_3 \dots x_N\}$

Mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ **Function Mean:** $\bar{f} = \frac{1}{N} \sum_{i=1}^{n} f(x_i)$

If binned then average bin counts:

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{B} n_j x_j \quad \bar{f} = \frac{1}{N} \sum_{j=1}^{B} n_j f(x_j)$$

(x_j: central value; n_j: counts in bin; B: total no. bins)

$\bar{x}$ is the "Sample mean" - estimation of 'population mean' - depends on 'true mean, N'

↳ $N = \langle x \rangle$

Sample mean: $\bar{x}$ 'An estimate'

True mean: N

Measuring Spread

Variance: $V(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$ (Squared deviation from mean)

Sample Variance $= \overline{x^2} - \bar{x}^2$

Proof:

$$V(x) = \frac{1}{N} \sum_i^N (x_i - \bar{x}^2)$$

$$= \frac{1}{N} \sum_i^N (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$$

$$= \frac{1}{N} \left(\sum_i^N x_i^2 - 2\bar{x} \sum_i^N x_i + \sum_i^N \bar{x}^2 \right)$$

$$= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2$$

$$= \overline{x^2} - \bar{x}^2$$

Function Variance $V(y) = \frac{1}{N}\sum_{i=1}^{N}(y(x_i) - \bar{y})^2 = \overline{y^2} - \bar{y}^2$

Standard Deviation:

$\sigma = \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$ Sample Stand deviation $= \sqrt{\frac{1}{N}\sum (x_i - \bar{x})^2}$ → Biased Estimator as used Sample data twice → For mean and the Std → non independent of each other

these are Sample estimates

True variance / Std Deviation

$V(x) = \langle x^2 \rangle - \langle x \rangle^2$

$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$

$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

Corrected Standard-Deviation - Bessel Correction

$$S = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}$$

Higher Moments

n^{th} algebraic moment : $a^n = \frac{1}{N}\sum_i x_i^n$

n^{th} Central Moment : $C^n = \frac{1}{N}\sum_i (x_i - \bar{x})^n$

1st Algebraic: Mean

2nd Central: Variance

'Skew' (measures asymmetry)

Related to 3rd Central: $\gamma = \frac{1}{N\sigma^3}\sum (x_i - \bar{x})^3$ → ensures Skew is dimensionless.

Curtosis (measures tails)

Related to 4th Central moment $k = \frac{1}{N\sigma^4}\left[\sum (x_i - \bar{x})^4\right] - 3$ → ensures 0 for gaussian distribution; → ensures dimensionality.

Covariance + Correlation

For datasets containing multiple variables $(x_i, y_i) = \{(x_1, y_1) \dots (x_N, y_N)\}$

Covariance: $Cov(x,y) = \frac{1}{N}\sum (x_i - \bar{x})(y_i - \bar{y})$

$= \frac{1}{N}\sum_i (x_i y_i - x_i\bar{y} - \bar{x}y_i - \bar{x}\bar{y})$

$= \overline{xy} - \bar{x}\,\bar{y}$

For a data set of M variables

M×M Covariance (diagonal) matrix

$V_{xy} = Cov(x,y) = \overline{xy} - \bar{x}\bar{y}$

Covariance in words

+ve Covariance: If values of $x > \bar{x}$ then $y > \bar{y}$

−ve Covariance: Opposite.

Correlation (Dimensionless $-1 \leq \rho \leq 1$)

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

For multi-dimensional:

$$V_{xy} = \rho_{xy} \sigma_x \sigma_y$$

ML algorithms

'attempt to fit some empirical function based on a set of features in data'

Sample data $\vec{x}_i = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_N\}$ → typically X

$\vec{x}$ is multidimensional - no. features

Overall (no. events, no. features)

Attempt to learn:

$$y_i = f(\vec{x}_i)$$

ML evaluation metric

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

All positive / Signal event: $P = TP + FN$

All negative / background events: $N = TN + FP$

All classified positive: $C_P = TP + FP$

All classified negative: $C_N = TN + FN$

True positive rate: $TPR = \frac{TP}{TP+FN}$ } True positive / All positive

True Negative rate: $TNR = \frac{TN}{TN+FP}$

False Positive rate: $FPR = \frac{FP}{FP+TN}$

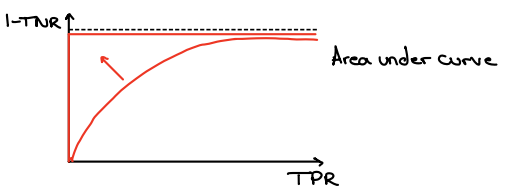
False Negative rate: $FNR = \frac{FN}{FN+TP}$

$$TPR + FNR = \frac{TP+FN}{TP+FN} = 1$$

Type-I Error – High False Positive Rate

Type II Error – High False Negative Rate

Reciever Operating Characteristic (ROC) Curve



Precision:

Fraction of all positively classified events that are correct

$$\rho = \frac{TP}{FP+TP}$$

Accuracy

↳ Fraction correctly classified

$$\alpha = \frac{TP+TN}{P+N}$$

Error Rate

↳ Fraction incorrectly classified

$$\varepsilon = \frac{FP+FN}{P+N}$$

Purity:

$$\rho_P = \frac{TP}{TP+FP} \qquad \rho_N = \frac{TN}{TN+FN}$$

Significance

$$\sigma = \frac{TP}{\sqrt{TP+FP}}$$

Signal to Noise (SNR)

$$SNR = \frac{TP}{FP}$$

F-Score

$$F = \frac{2TPR}{2TPR+FPR+FNR}$$

Avoiding Overfitting and Hyper parameter Tuning

↳ training Sample - fit the model

Validation Sample - optimize the hyper parameters

test Sample. - evaluate the performance

Cross - Validation

k fold – Split data into k (=4) folds : A, B, C, D

	Train	Test
Fold 1	BCD	A
Fold 2	ACD	B
Fold 3	ABD	C
Fold 4	ABC	D

Probabilistic Theory

Mathematical Probability

Kolmogorov Axioms (for X_i exclusive outcomes in Ω)

↳
1) $P(X_i) \geq 0 \quad \forall i$ (probability must be zero or positive)
2) $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$ (exclusive probabilities sum)
3) $\sum_{\Omega} P(X_i) = 1$ (all probabilities must sum to unity)

Frequentist Probability

given by relative frequency of particular outcome.

$$P(X) = \lim_{N \to \infty} \frac{n}{N}$$

N - total no.

n - observed outcome of X

True Probability - fixed probability but never attainable

Only applicable to repeatable experiments - keep relevant conditions stable

does not require any prior belief (purely experimental)

Frequentist vs Bayesian

↳ Both equally valid but cannot be mixed

Often depends on 'field'.

Should converge to the same inference in the 'asymptotic Nirvana'. → Sample size → ∞

Bayesian Probability

given by a degree of belief

$P(X)$ = degree of belief that X happens

Based on - Finelli's Coherent bet:

$P(X) =$ largest amount you would bet / amount you would gain if win → must still obey Kolmogorov Axioms

Subjective as property of observer – 'amount willing to bet'.

depends on observers knowledge so will change

Properties of Probability

Addition - non exclusive sets A and B in χ_i

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

↳ Stop overcounting

Union and Intersection

Union $A \cup B$ events in A or B or Both

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$$

Intersection $A \cap B$ events in both A and B

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$

Conditional Probability

$P(A|B)$ - probability of 'A given B'

probability event known to be in B also in A

$$P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

Note $P(A|B) \neq P(B|A)$

↳ Leads to Bayes's Theorem

Independence:

A and B are independent if:

$P(A|B) = P(A)$ (occurance of B has no input on A)

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

If both A and B are independent:

$P(A \text{ and } B) = P(A) \cdot P(B)$ ↳ ie. Probability factorises

↳ Can factorise probability into two independent

Formal Proof of Probability Theory

Prove if $p(A) = p_A$ then $p(\text{not } A) = 1 - p_A$

Declare all other outcomes in $\Omega - B$ $\quad P(B) = p_B$

Using Kol III: $\sum_{\Omega} P(\chi_i) = 1$; $p_A + p_B = 1$

$$p_B = 1 - p_A$$

Prove Probability in A or B but not both

Call $p(\text{not } A) = p(A^c)$ then $p(A \cup A^c) = 1$

$$p(A \cup B) = p(A \cup B)\underbrace{p(A \cup A^c)}_{1}$$

$$p(A \cap (B \cap A^c)) = p(A) + p(B \cap A^c)$$

$$p(B) = p(B \cap A) + p(B \cap A^c)$$

$$p(B \cap A^c) = p(B) - p(B \cap A)$$

$$p(A \cup B) = p(A) + p(B) - p(B \cap A)$$

Bayes Theorem (discrete)

$$P(A \text{ and } B) = P(A|B)\,p(B) = p(B|A)\,p(A)$$

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

for Bayes' theorem often dont know $p(B)$

Law of Total Probability

Exclusive Sets $\{A_1, \ldots, A_N\}$ and Seperate event B

$$P(B) = P(\{B \text{ and } A_1\} \text{ or } \{B \text{ and } A_2\} \text{ or } \ldots \text{ or } \{B \text{ and } A_N\})$$

$$= \sum_i^N P(B \text{ and } A_i)$$

$$= \sum_i^N P(B|A_i)\,P(A_i)$$

Law of Total Probability and Bayes Theorem

$$p(A|B) = \frac{P(B|A)\,P(A)}{\sum_i^N P(B|A_i)\,P(A_i)}$$

Monty Hall Problem - Bayes Theorem

Stick:

$$P(\text{car in A} \mid \text{host opens B}) = \frac{p(\text{open B} \mid \text{car in A})\,p(\text{car in A})}{p(\text{opens B})}$$

$p(\text{open B}) = \sum_i p(B|A_i)\,P(A_i)$

$= p(\text{open B} \mid \text{car in A})\,p(\text{car in A})$

$+ P(\text{open B} \mid \text{car in C})\,p(\text{car in C})$

$= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$

$$P(\text{car in A} \mid \text{opens B}) = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

Switch:

$$P(\text{car in C} \mid \text{open B}) = \frac{p(\text{open B} \mid \text{car in C})\,p(\text{car in C})}{p(\text{open B})} = \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Random Variables

Example: rolling a dice

random event: roll

random variable: $x = \{1, 2, 3, 4, 5, 6\}$

Probabilities: $P(x) = \{p(x_1), p(x_2), \ldots, p(x_6)\}$

Probability Mass Functions

↳ Discrete outcomes

Sum of probability over all possible values

$$\sum_i p(x_i) = 1$$ - Kolmogorov axioms

Probability Density Functions

↳ Continuous outcomes - can take any real value

Example: heights of people

$$p(x) = \lim_{\Delta x \to 0} \frac{P(x - \Delta x/2 < x < x + \Delta x/2)}{\Delta x}$$

→ Dimensions of density $[x^{-1}]$

Prob $x \to$ in a small interval Δx

Always normalised: 3rd Kolmogorov axiom

$$\int_\Omega p(x)\,dx = 1$$

For all x $(\forall x)$

$$p(x) \geq 0$$

Probability that $a < x < b$:

$$p(a < x < b) = \int_a^b p(x)\,dx$$

Change of variables

If $x \sim f(x)$ (x is distributed as $f(x)$) and $Y = h(x)$

then what is $g(y)$? → Prob distribution of y

$Y \sim g(Y)$

Must conserve probability:

$$g(Y)\,dY = f(x)\,dx$$

If h is 'monotonic' and 'invertible'

Use $x = h^{-1}(Y)$

$$g(Y)\,dY = f(h^{-1}(Y)) \frac{dx}{dY}\,dY$$

$$g(Y) = f(h^{-1}(Y)) \left|\frac{dx}{dY}\right| = f(x) \left|\frac{dx}{dY}\right| = \frac{f(x)}{|h'(x)|}$$

↳ Jacobian of transform

'modulus of the transformation differential'

For disjoint regimes ($h(x)$ is not a 1-1 mapping)

$$g(Y) = \sum_i \frac{f_i(x)}{|h'(x)|}$$

Extended to two Indep. Variables X, Y

Distributed by joint density $f(x,y)$ – transformed by $U = U(x,y)$ $V = V(x,y)$

Joint density (U,V):

$$G(U,V) = F(x,y) \left| J\left(\frac{x,y}{U,V}\right) \right| = F(x,y) \begin{vmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{vmatrix}$$

Extended to multi-Variable

For $\vec{X} = (X_1, X_2, \ldots, X_n)$ Variables With transformation functions $\vec{Y} = (Y_1, Y_2, \ldots, Y_m)$

$$g(\vec{Y}) = \left| J\left(\frac{\vec{x}}{\vec{y}}\right) \right| f(\vec{x})$$

Cumulative Distribution

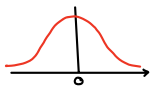
$$F(x) = \int_{-\infty}^{x} f(x)\, dx$$

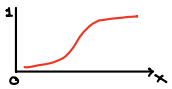
Note: $F(x_{min}) = 0$

$F(x_{max}) = 1$

pdf: $f(x)$ – Lower case

cdf: $F(x)$ – Upper case

pdf 

cdf 

$$P(x < x') = \int_{x_{min}}^{x'} f(x)\, dx = F(x')$$

$$P(x' < x < x'') = \int_{x'}^{x''} f(x)\, dx = F(x'') - F(x')$$

Percentage Point Functions F^{-1}

'The inverse of the c.d.f'

Maps us from random variable, X $[0,1]$ ⟶ random variable X

Useful to generate samples from a distribution.

$X = F^{-1}(p)$ ⟶ Uses F^{-1} takes IP as an argument ⟶ returns x

Truncated Distribution

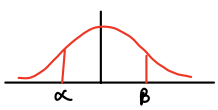
Standard distributions often $x \in [-\infty, \infty]$ or $x \in [0, \infty]$

Often require: Restricted range $x \in [\alpha, \beta]$ ← Requires Renormalisation

Truncated pdf

↳ uses truncated c.d.f to normalise

$$f(x) \longrightarrow f'(x) = \frac{f(x)}{F(\beta) - F(\alpha)}$$



Truncated c.d.f

$$F(x) \longrightarrow F'(x) = \frac{F(x) - F(\alpha)}{F(\beta) - F(\alpha)}$$

Truncated p.p.f

Def $G(x) = \frac{F(x) - \delta}{N} = p$ $\quad \delta = F(\alpha)$

$N = F(\beta) - F(\alpha)$

$F(x) = Np + \delta$

$F^{-1}(F(x)) = x = F^{-1}(Np + \delta)$

$F^{-1}(p) = F^{-1}((F(\beta) - F(\alpha))p + F(\alpha))$

Normal Distribution: notation

p.d.f ~ $N(x; \mu, \sigma)$

c.d.f ~ $\Phi(x; \mu, \sigma)$

p.p.f ~ $\Phi^{-1}(p; \mu, \sigma)$ ⟶ 'erf'

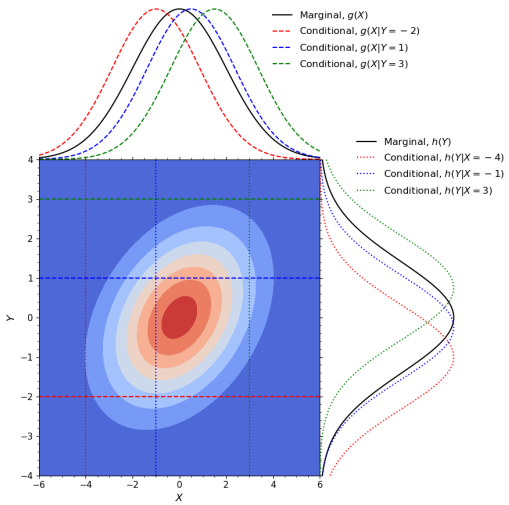
Joint p.d.f

Prob dist of two random variables

Consider p.d.f $X, Y \sim f(x, y)$

If X, Y are independent: $\boxed{f(x,y) = g(x)h(y)}$

↳ $g(x), h(y)$ - marginal distributions.



Working backwards:

Can 'integrate out' other variables

$g(x) = \int f(x,y)\,dy$ "marginal in x"

$h(y) = \int f(x,y)\,dx$ "marginal in y"

Conditional p.d.f

↳ a 'slice through the joint density' 3D-Slice through condition in 1 dimension

i) Prob of x given y: $\boxed{g(x|y) = \frac{f(x,y)}{h(y)}} = \frac{f(x,y)}{\int f(x,y)dx}$

ii) Prob of y given x: $\boxed{h(y|x) = \frac{f(x,y)}{g(x)}} = \frac{f(x,y)}{\int f(x,y)dy}$

An example:

$$\boxed{f(x,y) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \cdot e^{-y}} \qquad \begin{cases} x \in \{-\infty, \infty\} \\ y \in \{0, \infty\} \end{cases}$$

Unit Gaussian · Decaying Exponential

Marginal in x

$$g(x) = \int_0^\infty f(x,y)\,dy = \left[-\frac{e^{-x^2/2}}{\sqrt{2\pi}} e^{-y} \right]_0^\infty = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \left[-e^{-y} \right]_0^\infty = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Marginal in y

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,dy = \frac{e^{-y}}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-x^2/2}\,dx}_{\sqrt{2\pi}\text{ - see proof}} = e^{-y}$$

$f(x,y) = g(x)h(y)$ ⟶ $g(x|y) = \frac{f(x,y)}{h(y)} = g(x)$ $\quad h(x|y) = \frac{f(x,y)}{g(x)} = h(y)$

↳ Independent. ↳ In independent - no effect 'given'

Integral Proof: e^{-ax^2}

$$I = \int_{-\infty}^{\infty} e^{-ax^2}\,dx \longrightarrow I^2 = \int_{-\infty}^{\infty} e^{-ax^2}dx \int_{-\infty}^{\infty} e^{-ay^2}dy = \int_{-\infty}^{\infty} e^{-a(x^2+y^2)}\,dx\,dy$$

Convert to polar: $x(r,\theta) = r\cos\theta \quad y(r,\theta) = r\sin\theta \quad r^2 = x^2 + y^2$

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-a\cdot r^2}\,dx\,dy$$

Jacobian: $(x,y) \longrightarrow (r,\theta) \qquad J = \begin{pmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{pmatrix}$

$$dx\,dy = |J|\,dr\,d\theta$$

$$= \begin{vmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{vmatrix} dr\,d\theta = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} dr\,d\theta$$

$$= r(\cos^2\theta + \sin^2\theta)\,dr\,d\theta$$

$$= r\,dr\,d\theta$$

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-ar^2} r\,dr\,d\theta \xrightarrow{u = r^2,\ du/dr = 2r} I^2 = \int_0^{2\pi} d\theta \int_0^{\infty} \frac{e^{-au}}{2}\,du = 2\pi\left[-\frac{e^{-au}}{2}\right]_0^{\infty} = \frac{\pi}{a}$$

$$\boxed{I = \sqrt{\frac{\pi}{a}}}$$

Bayes Theorem - Continuous

$$f(x|y) = g(x|y)h(y) = h(y|x)\,g(x)$$

$$\boxed{g(x|y) = \frac{h(y|x)\,g(x)}{h(y)}}$$ holds for any r.v.s X, Y

↳ for bayesian all parameters are random variables

The likelihood function

For a continuous random variable X and some data $\{x_1, x_2, \ldots, x_n\}$

And a continuous range of hypothesis, θ:

Assume definition $x \sim f(x)$ then

Likelihood is:

$$\boxed{L(\theta) = p(x|\theta) = \prod_{i=1}^{N} f(x_i|\theta)}$$

↳ Product of p.d.f's of assumed distribution over all observations

Not strictly a p.d.f → p.d.f is a function of (x) → depends on data set.

A function of the parameters θ (and the data) not the random variable x

"The likelihood that we observe the data/measurement x_i given particular values of parameters θ" → Likelihood: $L(\theta)$ Function of θ

Represents: agreement between our model and observed data.

Bayesian Interference

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)}$$

$p(\theta|x)$ → Posterior distribution, probability distribution for parameter θ given the data we observed x

$p(\theta|x) \propto p(x|\theta)\,p(\theta)$

Posterior ∝ likelihood × prior

$p(x|\theta)$ → Likelihood Function, likelihood we observe data x given particular value, θ

$p(\theta)$ → The Prior, prior beliefs of θ (bayesian - prior belief required)

$p(x)$ → The Evidence, normalisation factor - ensures posterior is a p.d.f - just a number

Statistical Inference

↳ deducing properties of a population or distribution from observed data

From posterior distribution

↳ Can determine: Inferred value (point estimate)

↳ Normally use 'maximum a posteriori' probability (MAP) i.e. mode

Uncertainty:

↳ Often High Density Interval (HDI)

(narrowest interval containing $(1-\alpha)\%$

Properties of Distribution

Expectation

Random variable, x with p.d.f $f(x)$ $(x \sim f(x))$

Function $g(x)$ operating on x

Expectation of $g(x)$ (continuous distribution)

$$E[g(x)] = \langle g(x) \rangle = \int g(x) f(x)\,dx$$

↳ Integral over all x

Expectation of $g(x)$ (discrete distribution)

$$E[g(x)] = \langle g(x) \rangle = \sum_x g(x) P(x)$$

↳ Upper case (discrete)

Acts as a linear operator:

$$E[ag(x) + bh(x)] = aE[g(x)] + bE[h(x)]$$

Mean Density, μ (of probability function)

↳ expectation value of x itself ($g(x) = x$)

$$E[x] = \langle x \rangle = \mu = \int x f(x)\,dx$$

Variance, $V(x), \sigma^2$

Expectation value of $(x-\mu)^2$

$$V(x) = \sigma^2 = E[(x-\mu)^2]$$

σ: Std deviation

$$= E[x^2 - 2\mu x - \mu^2]$$

$$= E[x^2] - 2\mu \underbrace{E[x]}_{\mu} + E[\mu^2]$$

↳ μ^2 - ⟨constant⟩ = constant

$$= E[x^2] - 2\mu^2 + \mu^2 = E[x^2] - \mu^2$$

$$V(x) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$$

Moments

$\mu_n = E[x^n]$ → n^{th} algebraic moment

$\alpha_n = E[(x - \mu_1)^n]$ → n^{th} Central moment.

Mean:	μ_1	Skew	$\gamma_1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$
Variance:	α_2	Kurtosis	$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$

Covariance and Correlation

Between two random variables $X, Y \sim f(x, y)$

Covariance: $V(x,y) = E[(X - \mu_x)(Y - \mu_y)]$

$$= \iint (x - \mu_x)(y - \mu_y) f(x,y)\, dx\, dy$$

$$= \iint (xy - x\mu_y - \mu_x y - \mu_x \mu_y) f(x,y)\, dx\, dy$$

$$= \iint xy\, f(x,y)\, dx\, dy - \mu_y \underbrace{\int x f(x,y)\, dx\, dy}_{\mu_x} - \mu_x \underbrace{\int y f(x,y)\, dx\, dy}_{\mu_y} + \mu_x \mu_y$$

$$= \iint xy\, f(x,y)\, dx\, dy - \mu_x \mu_y$$

$$V(X,Y) = E[XY] - \mu_x \mu_y = E[XY] - E[X]E[Y]$$

Correlation $\rho(x, y) = \frac{V(X,Y)}{\sigma_x \sigma_y} = \frac{E[XY]}{\sigma_x \sigma_y} - \frac{\mu_x \mu_y}{\sigma_x \sigma_y}$

An example: $X \sim f(x) = Ne^{-x^2}$

1) Find N: $\int_{-\infty}^{\infty} Ne^{-x^2} dx = 1$ (Normalised)

Using $\int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi}$ → $N = 1$

$X \sim f(x) = \frac{e^{-x^2}}{\sqrt{\pi}}$

2) Find mean

$\mu = \langle x \rangle = E[x] = \int_{-\infty}^{\infty} x N e^{-x^2} dx$

Method 1: Odd function → 0

Method 2: $u = x^2$ $\frac{du}{dx} = 2x$ $du = 2x\, dx$

$$\int x e^{-x^2} dx \rightarrow \int_{-\infty}^{\infty} \frac{e^{-u}}{2} du = \left[-\frac{e^{-u}}{2} \right]_{-\infty}^{\infty} = 0 \qquad \mu = 0$$

3) Variance $E[(x - \mu)^2] \rightarrow E[x^2]$ as $\mu = 0$

$$E[x^2] = \int_{-\infty}^{\infty} \frac{x^2 e^{-x^2}}{\sqrt{\pi}} dx$$

$\int uv' = uv - \int vu'$ $u = x$ $u' = 1$ $V(x) = \frac{1}{2}$

$v' = x e^{-ax^2}$ $v = \frac{-e^{-x^2}}{2a}$

Proof: $\int_{-\infty}^{\infty} x^2 e^{-ax^2}$ $\quad u = x \longrightarrow u' = 1$

$v = \frac{e^{-ax^2}}{2a} \longleftarrow v' = xe^{-ax^2}$

$$= \int x^2 e^{-ax^2} dx = \left[\frac{-x e^{-ax^2}}{2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{e^{-ax^2}}{2a} dx = \frac{1}{2a}\sqrt{\frac{\pi}{a}}$$

Standard Normal $\quad f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

- Shift in x by μ $\quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ $\qquad x \longrightarrow z = \frac{x-\mu}{\sigma}$
- Scale by $1/\sigma$

Characteristic Function

Fourier Transform of p.d.f, $f(x)$

For random variable $X \sim f(x)$ with C.d.f $F(x)$

$$\varphi(t) = E[e^{itx}] = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

→ Characteristic function

Completely defines $f(x)$

Can freely switch between $f(x) \longleftrightarrow \varphi(t)$ → 'Characteristic function can completely define probably distribution'

Inverse Fourier transform

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(t) e^{-ixt}\, dt$$

Algebraic Moments (from Characterstic function)

- μ_n - n^{th} moment around mean

Previously $\mu_n = E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$ → μ_n can be obtained from n^{th} differential of $\varphi(t)$ at $t=0$

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

$\frac{d^n}{dt^n}$

n^{th} differential $\varphi^n(t) = \frac{d^n \varphi}{dt^n} = i^n \int_{-\infty}^{\infty} x^n e^{itx} f(x)\, dx$ → 'Can generate moments by differentiating the Characteristic function'

$t=0$

$$\varphi_n(0) = i^n \int_{-\infty}^{\infty} x^n f(x)\, dx = i^n \mu_n = i^n E[X^n]$$

$$\mu_n = E[X^n] = i^{-n}\left[\frac{d^n \varphi(t)}{dt^n}\right]_{t=0}$$

→ Advantage

Often much more useful to differentiate using Characteristic function than integrate p.d.f

Central Moment / Moment about mean

(Sub $Y = X - \mu$)

Example : Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\varphi(t) = e^{\mu t} e^{-1/2\sigma^2 t^2}$ → proved below

$$\mu = E[x] = i^{-1}\left[\frac{d\varphi(t)}{dt}\right]_{t=0}$$

$$\frac{d\varphi(t)}{dt} = i\mu e^{i\mu t} e^{-\frac{1}{2}\sigma^2 t^2} - e^{i\mu t}\frac{1}{2}\sigma^2 2t e^{-\frac{1}{2}\sigma^2 t^2}$$

$$= (i\mu - \sigma^2 t)\varphi(t)$$

$$\mu = i^{-1}(i\mu - \sigma^2 t)\Big|_{t=0} = \mu \quad \text{Q.E.D}$$

$$V(x) = E[(x-\mu)^2] = E[x^2] - \mu$$

Where $E[x^2] = i^{-2}\left[\frac{d^2\varphi(t)}{dt^2}\right]_{t=0}$

$$\frac{d^2\varphi(t)}{dt^2} = -\sigma^2\varphi(t) + (i\mu - \sigma^2 t)\frac{d\varphi(t)}{dt}$$

$$= -\sigma^2\varphi(t) + (i\mu - \sigma^2 t)^2\varphi(t)$$

Thus $E[x^2] = -1\left(-\sigma^2 - \mu^2 - 2i\mu\sigma^2 t + \sigma^4 t^2\right)\varphi(t)\Big|_{t=0}$

$$E[x^2] = \sigma^2 - \mu^2$$

$V(x) = E[x^2] - \mu^2 = \sigma^2$ → Characteristic function often easier to compute moments → Used to show μ, σ can be undefined.

Relation to moments

Expand exponential

$$\varphi(t) = E[e^{itx}] = E\left[\sum_{n=0}^{\infty}\frac{(itx)^n}{n!}\right] = \sum_{n=0}^{\infty}\frac{(it)^n}{n!}E[x^n] \qquad \text{Using } e^{itx} = \sum_{n=0}^{\infty}\frac{(itx)^n}{n!}$$

$$= \sum_{n=0}^{\infty}\frac{(it)^n}{n!}\mu_n \rightarrow$$

$\varphi(t)$, Characteristic function can be expanded as the sum of the moments which appear as coefficients of $\frac{(it)^n}{n!}$

Moment Generating Function (MGF) $E[e^{tx}]$

Characteristic Function → Always exists

Disadvantages: Only exists if finite for all t (real)

Properties of Characteristic Function

$$\varphi_x(t) = E[e^{itx}]$$

1) Sum of many independent random variables

$\varphi_z(t)$ where: $z = \sum_i^N x_i$, $\varphi_z(t) = E[e^{itz}] = E[e^{it\Sigma x_i}]$

$$\varphi_z = \int e^{itz} f(z)\,dz = \int e^{it\Sigma x_i} \cdot f(x_1)f(x_2)\cdots f(x_n) \cdot d(x_1)d(x_2)\cdots d(x_n)$$

$f(z)$ → Independent r.v's

This Seperable p.d.f $f(x_1,\ldots x_n) = f(x_1)\cdot\ldots\cdot f(x_n)$

$$= \int\int\cdots\int_{x_1 x_2 \ldots x_n} e^{itx_1}f(x_1)dx_1\; e^{itx_2}f(x_2)dx_2 \cdots e^{itx_n}f(x_n)dx_n$$

$$= \prod_{i=1}^{N}\int e^{itx_i}f(x_i)dx_i = \boxed{\prod_{i=1}^{N}\varphi_{x_i}(t)} \rightarrow$$

Charac func for Sum of random variables is product of individual Charac functions

2) Random Variable multiplied by a constant

$$\varphi_{ax}(t) = E[e^{t a x}] = E[e^{i a t x}] = \varphi_x(at)$$

Proof of $\varphi(t)$ for Normal

Starting With: Std Normal: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ → even function

$$\varphi(t) = E[e^{itx}] = E[\cos(tx) + i\sin(tx)] = E[\cos(tx)] + iE[\sin(tx)]$$ → useful expansion

$$= \int_{-\infty}^{\infty} \cos(tx) f(x) + i \int_{-\infty}^{\infty} \sin(tx) f(x)$$

$\int_{-\infty}^{\infty} \text{odd} = 0$

→ Difficult to Solve: trick, Investigate first derivative of $\varphi(t)$

Investigating trick

$$\frac{d}{dt}\varphi(t) = \frac{d}{dt} E[e^{itx}] = E\left[\frac{d}{dt}(e^{itx})\right]$$

$$= E[ix e^{itx}]$$

$$= iE[x\cos(xt)] - E[x\sin(xt)] \quad (i^2 \sin x)$$

$$= i\int_{-\infty}^{\infty} x\cos(tx) f(x)\,dx - \int_{-\infty}^{\infty} x\sin(tx) f(x)\,dx$$

$i\int_{-\infty}^{\infty} x\cos(tx) f(x)\,dx$: O (odd)

$x f(x) = x\left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right) = -\frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right)$

$= -\frac{d}{dx} f(x)$

$$= \int \sin(tx) \frac{df(x)}{dx}\,dx$$ → By parts

$$= [\sin(tx) f(x)]_{-\infty}^{\infty} - \int t\cos(tx) f(x)\,dx$$

$$= -t\int_{-\infty}^{\infty} \cos(tx) f(x)\,dx$$

Using $\varphi(t) = \int_{-\infty}^{\infty} \cos(tx) f(x)\,dx$

$\frac{d}{dt}\varphi(t) = -t\,\varphi(t)$ → Condition $\varphi(0) = E[e^0] = 1$

General sol: $\varphi(t) = e^{-t^2/2}$

For non-standard normal

$$X \longrightarrow Z = \frac{X-\mu}{\sigma}$$

$$\varphi(t) = E[e^{itx}] = E[e^{it(\mu+\sigma z)}] = e^{i\mu t} E[e^{i\sigma t z}]$$

$$= e^{i\mu t} \varphi_z(\sigma t) = \underline{e^{i\mu t} e^{-\sigma^2 t^2/2}}$$

Common Probability Densities

Probability distribution depends on r.v.s, $\vec{x}$ and parameters, $\vec{\theta}$

$\boxed{p(\vec{x};\theta)}$ eg normal $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Binomial Distribution (fixed no. outcomes, two outcomes)

n trials, $p(\text{Success}) = p$, $p(\text{fail}) = 1-p$

p that first k are Success: $p = p^k(1-p)^{n-k}$ × Combinations that k Success in n trials: $C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial p.d.f $\boxed{p(k;p,n) = \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}}$

Mean: $\boxed{\mu = E[k] = np}$

$$\mu = E[k] = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}$$

$$= \sum_{k=0}^{n} \frac{n(n-1)!}{(k-1)!(n-k)!} p\, p^{k-1} (1-p)^{(n-1)-(k-1)} \quad \text{Sub } j = k-1,\ m = n-1$$

$$= np \underbrace{\sum_{j=0}^{m} \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}}_{1} = np$$

Variance: $\boxed{V(k) = np(1-p)}$

$$V(k) = E[(k-\mu)^2] = E[k^2] - E[k]^2 = E[k^2] - (np)^2$$

$$E[k^2] = \sum_{k=0}^{n} k^2 \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}$$

$$= \sum_{k=1}^{n} kn \frac{(n-1)!}{(k-1)!(n-k)!} p\, p^{k-1} (1-p)^{(n-1)-(k-1)} \quad \text{Sub } j = k-1,\ m = n-1$$

$$= np \underbrace{\sum_{j=0}^{m} (j+1) \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}}_{B} = np\left[\underbrace{\sum_{j=0}^{m} jB}_{mp=(n-1)p} + 1\right]$$

$$= np((n-1)p+1) = np(np-p+1) = (np)^2 - np^2 + np$$

$$V(k) = E[k^2] - E[k]^2 = (np)^2 - np(p-1) - (np)^2 =$$

Standard Deviation: $\boxed{\sigma = \sqrt{np(1-p)}}$

Poisson Distribution

Binomial where $n \to \infty$ but $np = \lambda$ is fixed

Binomial becomes:

$$P = \left(k; \frac{\lambda}{n}, n\right) = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

As $n \to \infty$: $\frac{n!}{(n-k)!} = n(n-1)(n-2)\dots(n+k+1) \underset{n\to\infty}{=} n^k$

$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \longrightarrow \left(1 - \frac{\lambda}{n}\right)^{n} \longrightarrow e^{-\lambda}$$

$$p(k;\lambda) = \frac{n^k \lambda^k}{k!\, n^k} e^{-\lambda} = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\boxed{p(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}}$$

Mean: $E[k] = \sum_{k=0}^{\infty} \frac{k\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$

$$= \lambda e^{-\lambda} \underbrace{\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}}_{e^{\lambda}} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$\boxed{E[k] = \lambda}$$

Variance: $E(k^2) - E(k)^2$

$$E[k^2] = \sum_{k=0} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^{k-1}}{(k-1)!}$$

$$= \lambda e^{-\lambda}\left[\underbrace{\sum_{k=1}^{\infty} (k-1)\frac{\lambda^{k-1}}{(k-1)!}}_{\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}\,/\,e^{\lambda}} + \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}}\right]$$

$$= \lambda e^{-\lambda}(\lambda e^{\lambda} + e^{\lambda}) = \lambda(\lambda+1) = \lambda^2 + \lambda$$

$$\mathrm{Var}(k) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$\boxed{\mathrm{Var}(k) = \lambda}$$

$$\boxed{\sigma = \sqrt{\lambda}}$$

Note: Error for weights per event

If each entry has its own weight, w_i

Count in bins $\quad \boxed{n_B = \sum_{i \in B} w_i}$

Compute Variance in each bin $\quad \boxed{V(n_B) = \sum_{i \in B} w_i^2}$

Uncertainty $\quad \boxed{\sigma(n_B) = \sqrt{\sum_{i \in B} w_i^2}}$

Normal/Gaussian Distribution

Standard normal $\quad \mu = 0, \sigma = 1 \sim N(0,1)$

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}}$$

Properties of Normal Distribution

1) Any linear combination of norm dist r.v.s is also normal dist.
2) Sample mean and Variance are independent
3) If x_i are Standard normally distributed:
   → p.d.f is constant on the hypersphere $\left(\sum_i^N x_i^2 = \text{constant}\right)$ → Unique for Normal dist

Non Standard Normal

↳ Shift by μ; Scale by $1/\sigma$ $\quad z = \frac{x-\mu}{\sigma}$

$$\boxed{p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}$$

Characteristic Function

$$\boxed{\phi_x(t) = e^{i\mu t} e^{-\frac{1}{2}\sigma^2 t^2}}$$

Useful Integral for Normal

$$\boxed{\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}}$$

Cummalative Distribution Function (c.d.f) for Normal/Gaussian

$$\boxed{\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]} \qquad \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

Inverse c.d.f Φ^{-1}

↳ Z-score

or $\quad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$

Mean

$$E[x] = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \left(\text{Sub } z = \frac{x-\mu}{\sigma}\right)$$

$$= \int_{-\infty}^{\infty} \frac{z}{\sigma\sqrt{2\pi}} e^{-z^2/2} dz + \int \frac{\mu}{\sqrt{2\pi}} e^{-z^2/2} dz = \mu \qquad E[x] = \mu$$

$$\mu \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = \mu$$

Variance

$$E[x^2] = \int_{-\infty}^{\infty} \frac{x^2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \left(\text{Sub } z = \frac{x-\mu}{\sigma} \longrightarrow x^2 = \sigma^2 z^2 + 2\mu\sigma z + \mu^2\right)$$

$$= \underbrace{\int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz}_{\sigma^2} + \int_0^{\infty} \frac{2\mu z}{\sqrt{2\pi}} e^{-z^2/2} + \underbrace{\int \frac{\mu^2}{\sqrt{2\pi}} e^{-z^2/2} dz}_{\mu^2}$$

$$= \sigma^2 + \mu^2$$

$$V(x) = E[x^2] - \mu^2$$

$$\boxed{V(x) = \sigma^2}$$

Multi-Variate Normal

For independant r.v.s $\vec{x}$ then $\longrightarrow$ $\vec{x}$ random variable with means $\vec{\mu}$

$$p(\vec{x}; \vec{\mu}, \vec{\sigma}) = \prod_i p(x_i; \mu_i, \sigma_i)$$

$$= \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

We can include dependent r.v.s with a correlation term:

term. $\vec{\sigma}^2 = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_n^2 \end{pmatrix}$ $\quad \underline{\underline{\sigma}}^2 = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ \rho_{1n}\sigma_1\sigma_n & & \cdots & \sigma_n^2 \end{pmatrix}$

Term in exponential becomes:

$$(\vec{x} - \vec{\mu})^T \underline{\underline{V}}^{-1} (\vec{x} - \vec{\mu}) \quad \text{where } \underline{\underline{V}} = \underline{\underline{\sigma}}^2$$

Multivariate Normal P.d.f

$$\boxed{p(\vec{x}; \vec{\mu}, \underline{\underline{V}}) = \frac{1}{(2\pi)^{n/2}\sqrt{|V|}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^t \underline{\underline{V}}^{-1} (\vec{x} - \vec{\mu})\right]}$$

Properties of Multi-Variate Normal

1) Contours of constant probability where:

$\longrightarrow (\vec{x} - \vec{\mu})^T \underline{\underline{V}}^{-1} (\vec{x} - \vec{\mu}) = \text{Const.}$

2) Any Slice through a m.v.n distribution is also a m.v.n (Proof below)

$\longrightarrow$ If $p(\vec{x})$ is a m.v.n at dim n

$p(\vec{x}_{n-i} | \vec{x}_i)$ is also a m.v.n in dim n-1

With Covariance matrice V_{n-1} obtained by removing i^{th} row and column from V^{-1} and inverting

3) Any projection to a lower dimension is also a m.v.n

get $\underline{\underline{V}}_{n-1}$ by removing relevant row + column from $\underline{\underline{V}}$

Mathematical Proof 2)

Partion $\vec{x} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix}$ with dimensions $\begin{bmatrix} q \times 1 \\ (N-q)\times 1 \end{bmatrix}$

Thus $\vec{\mu} = \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix}$ w/ $\begin{bmatrix} q\times 1 \\ (N-q)\times 1 \end{bmatrix}$ and $\underline{\underline{V}} = \begin{pmatrix} \underline{\underline{V}}_{11} & \underline{\underline{V}}_{12} \\ \underline{\underline{V}}_{21} & \underline{\underline{V}}_{22} \end{pmatrix}$ w $\begin{bmatrix} q\times q & q\times(N-q) \\ (N-q)\times q & (N-q)\times(N-q) \end{bmatrix}$

the $p(\vec{x}_1 \mid \vec{x}_2) \sim$ m.v.n $N(\vec{\mu}', \underline{\underline{V}}')$ with:

$$\vec{\mu}' = \vec{\mu}_1 + \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}(\vec{x}_1 - \vec{\mu}_2)$$
$$\underline{\underline{V}}' = \underline{\underline{V}}_{11} - \underline{\underline{V}}_{12}\underline{\underline{V}}_{22}^{-1}\underline{\underline{V}}_{21}$$

Example in 2D

r.v.s X, Y

$$\vec{\mu} = (\mu_x, \mu_y) \qquad \underline{\underline{V}} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

$$|V| = \sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2 = \sigma_x^2\sigma_y^2(1-\rho^2)$$

$$\sqrt{|V|} = \sigma_x\sigma_y\sqrt{1-\rho^2}$$

$$V^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

$$V^{-1}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} \sigma_y^2(x-\mu_x) & -\rho\sigma_x\sigma_y(x-\mu_y) \\ -\rho\sigma_x\sigma_y(x-\mu_x) & \sigma_x^2(y-\mu_y) \end{pmatrix}$$

$$(x-\mu_x \mid V^{-1}\begin{pmatrix} v-\mu_x \\ y-\mu_y \end{pmatrix} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\left[\sigma_y^2(x-\mu_x)^2 - \rho\sigma_x\sigma_y(y-\mu_y)(x-\mu_x) - \rho\sigma_x\sigma_y(x-\mu_x)(y-\mu_y) + \sigma_x^2(y-\mu_y)^2\right]$$

$$= \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]$$

Thus: $$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right\}\right]$$

Exponential Decay

$p(x) = Ne^{-\lambda x}$

Normalise: $\int_0^\infty Ne^{-\lambda x}\,dx = 1$

$$\frac{1}{N} = \left[\frac{-e^{-\lambda x}}{\lambda}\right]_0^\infty \qquad \frac{1}{N} = \frac{1}{\lambda} \qquad N = \lambda$$

p.d.f $\boxed{p(x) = \lambda e^{-\lambda x}}$

c.d.f $F(x) = \lambda \int_0^x e^{-\lambda x'}\,dx' = \lambda\left[-\frac{1}{\lambda}e^{-\lambda x'}\right]_0^x$

$$\boxed{F(x) = 1 - e^{-\lambda x}}$$

Mean: $\mu = \mathbb{E}[x] = \int_0^\infty x\lambda e^{-\lambda x}\,dx$

$u = \lambda x \qquad v' = e^{-\lambda x}$

$u' = \lambda \qquad v = \frac{e^{-\lambda x}}{\lambda}$

$$= \left[-\lambda e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = 0 + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda} = \tau$$

$$\boxed{\mu = \mathbb{E}[x] = \frac{1}{\lambda} = \tau}$$

p.d.f alternative $\boxed{p(x) = \frac{e^{-x/\tau}}{\tau}}$

Variance $\mathbb{E}[x^2] - E[x]^2$

$u = \lambda x^2 \qquad v' = e^{-\lambda x}$

$E[x^2] = \int x^2 \lambda e^{-\lambda x}\,dx \qquad u' = 2\lambda x \qquad v = -\frac{e^{-\lambda x}}{\lambda}$

$$= \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2x e^{-\lambda x}\,dx$$

$$= 0 + \frac{2}{\lambda}E[x] = \frac{2}{\lambda^2}$$

$$V(x) = \mathbb{E}[x^2] - E[x]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \qquad \boxed{V(x) = \frac{1}{\lambda^2}} \qquad \boxed{\sigma = \frac{1}{\lambda}}$$

Relationship to Poisson:

Poisson – Counting for n occurences (in same interval)

Exp – length of time between occurences.

Take Poisson w/ λ in time t

Expectation is λt with Poisson distribution

$\boxed{p(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}}$ → expectation λt

Probability of 0 occurences in time t

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

Probability that event occours within time t

$P(T \le t) = \underbrace{1 - e^{-\lambda t}}_{\text{c.d.f of exp.}}$

Polynomial P.d.f's

Advantages: very flexible Disadvantage: Go negetive

↳ Kol axiom: $p(x) \ge 0$ for $\forall x$

Bernstein Basis (Basis of Polynomials)

↳ positive definite everywhere

defined for $0 \le x \le 1$ ⟶ Can be easily transformed to any range $[a,b]$

$x \longrightarrow \frac{x-a}{x-b}$

Bernstein Basis Polynomials

↳ $b_{i,n} = \binom{n}{i} x^i (1-x)^{n-i}$ for $i = 0, \dots, n$

↳ degree n ⟶ degree $n+1$ polynomial.

$b_{0,0} = 1$

$b_{0,1} = 1-x \qquad b_{1,1} = x$

$b_{0,2} = (1-x)^2 \qquad b_{1,2} = 2x(1-x) \qquad b_{2,2} = x^2$

} Change given order n

Bernstein in generic range $[a,b]$:

$$b_{i,n}(x) = \frac{1}{(b-a)^n} \frac{n!}{i!(n-i)!} (x-a)^i (b-x)^{n-i}$$

Derivative: $\frac{d\, b_{i,n}(x)}{dx} = \frac{n}{(b-a)} [b_{i-1,n-1}(x) - b_{i,n-1}(x)]$

Integral: $\int b_{i,n}(x)\, dx = \frac{(b-a)}{n+1} \sum_{j=i+1}^{n+1} B_{j,n+1}(x)$

Def Integral: All basis funcs of same order have same def integral over $[a,b]$

$$\int_a^b b_{i,n}(x)\, dx = \frac{b-a}{n+1}$$

The Bernstein Polynomial: $B_n = \sum_{i=0}^{n} C_i b_{i,n}$

$p(x) = N \sum_{i=0}^{n} C_i\, b_{i,n}$ $\int p(x)\, dx = 1$

$\frac{1}{N} = \int \sum_{i=0}^{n} C_i b_{i,n}\, dx = \sum C_i \int b_{i,n}(x)\, dx$

$= \sum_{i=0}^{n} C_i \frac{1}{n+1} = \frac{\sum C_i}{(n+1)}$

$$p(x) = \frac{n+1}{\sum C_i} \sum C_i b_{i,n}(x)$$

χ^2 distribution

K Degrees of Freedom

↳ Distribution of Sum of Squares of k-independent normal variables

If x_i are K independent Std. normal

$$Z = \sum_{i=1}^{k} x_i^2 \quad \text{(Sum of Squares)}$$

↳ Distributed according to $\chi^2(k)$

χ^2 P.d.f

$$p(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Mean: $E[x] = k$

$$E(x) = \int_0^\infty x \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} dx$$

$$= \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty x^{k/2} e^{-x/2} dx$$

$u = \frac{x}{2}$ $\quad dx = 2du$

$$= \frac{2^{k/2+1}}{2^{k/2}\Gamma(k/2)} \int_0^\infty u^{k/2} e^{-u} du$$

gamma function: $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ $\quad a = k/2+1$

$$= \frac{2}{\Gamma(k/2)} \cdot \Gamma\left(\frac{k}{2}+1\right)$$

gamma identity: $\Gamma(a+1) = a\Gamma(a)$

$$= \frac{2}{\Gamma(k/2)} \frac{k}{2} \Gamma(k/2)$$

$$= k$$

Variance $V(x) = 2k$ $\quad \sigma = \sqrt{2k}$

$$E[x^2] - E[x]^2 = E[x^2] - \mu^2$$

$$E[x^2] = \int_0^\infty x^2 \left(\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}\right) dx$$

$$= \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty x^{k/2+1} e^{-x/2} dx$$

$u = \frac{x}{2}$ $\quad dx = 2du$

$$= \frac{2}{2^{k/2}\Gamma(k/2)} \int_0^\infty (2u)^{k/2+1} e^{-u} du$$

$$= \frac{4}{\Gamma(k/2)} \int_0^\infty u^{k/2+1} e^{-u} du$$

gamma function: $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ $\quad a = k/2+2$

$$= \frac{4}{\Gamma(k/2)} \Gamma(k/2+2)$$

gamma identity: $\Gamma(a+1) = a\Gamma(a)$ $\quad \Gamma(k/2+2) = (k/2+1)\Gamma(k/2+1)$

$$= (k/2+1)(k/2)\Gamma(k/2)$$

$$= \frac{4}{\Gamma(k/2)} (k/2+1)(k/2)\Gamma(k/2)$$

$$= 4(k/2+1)(k/2) = k^2 + 2k$$

$$V(x) = (k^2+2k) - \mu^2 = k^2 + 2k - k^2 = 2k$$

C.d.f: allows us to read off fraction contained within given Standard deviation of multivariate normal distribution → Note: Fraction contained within given Z score for different degrees of freedom (k)

DoF (k)	Within 1σ
k = 1	0.683
k = 2	0.393
k = 3	0.199

Convolutions

Imagine 2 independent r.v.s X, Y then $Z = X + Y$

$Z = X + Y$

$f(x,y) = g(x)h(y)$

C.d.f $F(z)$

$$F(z) = P(x+y<z) = \int_{-\infty}^{\infty} g(x)\,dx \int_{-\infty}^{z-x} h(y)\,dy$$

$$= \int_{-\infty}^{\infty} h(y)\,dy \int_{-\infty}^{z-y} g(x)\,dx$$

P.d.f $f(z)$

$$f(z) = \frac{dF(z)}{dz} = \int_{-\infty}^{\infty} g(x)h(z-x)\,dx$$

$$= \int_{-\infty}^{\infty} h(y)\,g(z-y)\,dy$$

} known as Convolution: $f(z) = g(x) \otimes h(y)$

If $Z = XY$ (mellin Convolution)

$$f(z) = \int_{-\infty}^{\infty} \frac{g(x)h(z/x)}{|x|}\,dx$$

$$= \int_{-\infty}^{\infty} \frac{g(z/y)\,h(y)}{|y|}\,dy$$

Generating Samples from distributions and inverse c.d.f

Simulation knowing Inverse c.d.f

C.d.f: Random Variable → Cummulative probability $[0,1]$ $F(x) = \int_{-\infty}^{x} f(x)dx$

Inverse c.d.f: $[0,1]$ → Random variable x $\boxed{F^{-1}(p) = x}$

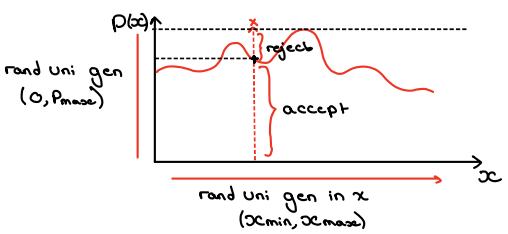
↳ Commonly 'Percentage point function' (p.p.f)

↳ $F^{-1}(p)$ or $\Phi^{-1}(p)$

1) Generate Uniform random numbers
2) Sub into ppf to transform into relevent dist.

Accept Reject Method (Brute force)

1) Compute max value p.d.f (f_{max})
   → Not necessarily normalised
2) Generate random Uniform normal in x
   ↳ along random variable x
3) Generate random Uniform normal in y
   ↳ between
4) If $y < p(x)$ then accept
5) If $y > p(x)$ then reject.



Convergence

Convergence in Distribution

For $N \to \infty$, distribution of sample tends to $p(x)$
- N: Sample size
- $p(x)$: probability density (true)

For $(x_1, x_2, \ldots, x_n)$ with empirical c.d.fs $(F_1, F_2, \ldots, F_n)$

$$\lim_{n \to \infty} F_n(x) = F(x)$$ → Requirement for 'convergence in distribution'

Convergence in Probability

Implies Convergence in distribution (Stronger requirement)
- Places a requirement on distance x_n from x (bounding limit)

'Probability for 'outliers' becomes smaller as Sample Size increases

For $(x_1, x_2, \ldots, x_n)$ converges in probability to the random variable x
- If for all $\varepsilon > 0$:

$$\lim_{n \to \infty} P(|x_n - x| > \varepsilon) = 0$$

Law of Large Numbers

Often distribution not a 'priori' known — obtained using experiment

Frequentist Probability

Expectation of frequency tends to probability

h – Sample frequency/mean

$$h = \frac{1}{n} \sum x_i$$
- n: no trials
- x_i: 'outcome of i^{th} trial' → 1 if 'match'

$\sum x_i$ – no. of times event occoured in n Samples

$\frac{1}{n} \sum x_i$ – frequency/proportion of trials in which event occured

Although $E[x] = np$ → whole distribution

$E[x_i] = p$ → outcome of i^{th} trial

$$E[h] = \hat{h} = \frac{1}{n} \sum E[x_i] = \frac{1}{n} np = p$$

Law of Large no.: As $n \to \infty$ then $h \to p$

$E[h] = p$ (consistent).

Central Limit Theorem

For any x_i generated from any distribution (mean, N_i and variance, σ_i^2)

Sum of random variables:

$S = \sum_i x_i$ has mean: $N = \sum_i N_i$

variance $\sigma^2 = \sum_i \sigma_i^2$

Central Limit Theorem: as $N \to \infty$, distribution of $S \longrightarrow$ Normal/Gaussian $G(N, \sigma)$

Any distribution with defined mean and variance with large $N \longrightarrow$ tends to Normal Dist

Proof of CLT

Assume N independent random variables x_i

from any distribution with mean N_i and variance σ_i^2

Standardise them:

$$Z_i = \frac{x_i - N_i}{\sigma_i}$$

Wish to find probability distribution: $Y = \frac{\sum_i Z_i}{\sqrt{N}}$ as $N \to \infty$

Moments

$$E[z_i] = \frac{\overbrace{E[x_i]}^{N_i} - \overbrace{E[N_i]}^{N_i}}{\sigma_i} = 0$$

$$E[z_i^2] = E\left[\frac{(x_i - N_i)^2}{\sigma_i^2}\right] = \frac{1}{\sigma_i^2} E[x_i^2 - 2x_i N_i + N_i^2]$$

$$= \frac{1}{\sigma_i^2}\left\{E[x_i^2] - 2N_i \underbrace{E[x_i]}_{N_i} + N_i^2\right\}$$

$$= \frac{1}{\sigma_i^2}\left\{\underbrace{E[x_i^2]}_{\sigma_i^2 + N_i^2} - N_i^2\right\} = \frac{1}{\sigma_i^2}\left\{\sigma_i^2 + N_i^2 - N_i^2\right\}$$

$$= 1$$

Characteristic Functions

$$\varphi_y = \prod_i \varphi_{z_i}\left(t/\sqrt{N}\right)$$

Studying $\varphi_{z_i} = E[e^{itz_i}] = E\left[\sum_{n=0}^{\infty} \frac{(it)^n}{n!} z_i^n\right]$ → Expansion around moments

$$= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} E[z_i^n] = 1 + it \underbrace{E[z]}_{0} + \frac{(it)^2}{2} \underbrace{E[z_i^2]}_{1} + \frac{(it)^3}{3!} E[z_i^3] + \dots$$

$$= 1 - \frac{t^2}{2} + \frac{(it)^3}{3!} E[z_i^3] + \dots$$

$$\varphi_y(t) = \prod_{i=0}^{N} \varphi_{z_i}\left(\frac{t}{\sqrt{N}}\right) = \left(1 - \frac{t^2}{2N} + \frac{(it)^3}{3! N^{3/2}} E[z_i^3] + \dots\right)^N$$

↳ changes: $\prod_{i=0}^{N}$ and $\frac{1}{\sqrt{N}}$

$$\ln \varphi_y(t) = N \ln\left(1 - \frac{t^2}{2N} + \frac{(it)^3}{3! N^{3/2}} E[z_i^3] + \dots\right)$$ → Vanishes as $N \to \infty$

$$= N \ln\left(1 - \frac{t^2}{2N}\right) = N\left(-\frac{t^2}{2N} - \frac{t^4}{4N^2} + \dots\right) \xrightarrow{0;\, N \to \infty} -\frac{t^2}{2}$$

Using $\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3}$

Recalling Properties of Characteristic Function

$\varphi_x(t) = E[e^{itx}]$

$$\varphi_y(t) = \varphi_{\frac{\sum z_i}{\sqrt{N}}}(t) = E\left[e^{it\left[\sum z_i/\sqrt{N}\right]}\right]$$

$$= \varphi_{\sum z_i}\left(t/\sqrt{N}\right)$$

$$= \prod_i \varphi_{z_i}\left(t/\sqrt{N}\right)$$

↳ Char func of Sum of r.v's = Product of indiv Characteristic functions

$\ln \varphi_Y(t) = -\frac{t^2}{2}$ $\quad \boxed{\varphi_Y(t) = e^{-t^2/2}}$ → c.f for std normal distribution

Propagation of Errors

Often give Central Values and Uncertainties
↳ Convention: Uncertainty at 1 std

$x \pm \sigma_x$ → Require σ_y on $y(x)$

Linear Function

Say $Z = ax + b$

$$\begin{aligned} V(Z) &= E[Z^2] - E[Z]^2 \\ &= E[(ax+b)^2] - E[ax+b]^2 \\ &= E[a^2x^2 + abx + b^2] - (aE[x] + b)^2 \\ &= a^2E[x^2] + 2ab\,E[x] + b^2 - (a^2E[x]^2 + 2ab\,E[x] + b^2) \\ &= a^2\{E[x^2] - E[x]^2\} = a^2V(x) \end{aligned}$$

$\boxed{V(Z) = a^2 V(x)}$

$\boxed{\sigma_Z = a\,\sigma_x}$ → Shift of b: does nothing to Spread
Scale of a: increases Spread

Generic Functions $y(x)$ - Using Taylor Expansion

Taylor Expansion:

$$y(x) = y(x_0) + (x - x_0)\left.\frac{dy}{dx}\right|_{x=x_0} + \mathcal{O}(x^2) + \dots$$

$$\begin{aligned} V(y(x)) &= E[y(x)^2] - E[y(x)]^2 \\ &= E\Big[\underbrace{y(x_0)^2}_{y(x_0)^2} + (x-x_0)^2\left.\frac{dy}{dx}\right|_{x=x_0}^2 + 2\,y(x_0)(x-x_0)\left.\frac{dy}{dx}\right|_{x=x_0}\Big] \\ &\quad - \underbrace{E[y(x_0)]^2}_{y(x_0)^2} - E\Big[(x-x_0)\left.\frac{dy}{dx}\right|_{x=x_0}\Big]^2 \\ &= \cancel{y(x_0)^2} + \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right)^2 E[x^2 - 2xx_0 + x_0^2] - \cancel{y(x_0)^2} - \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right)^2 \left(E[x]^2 + E[x_0]^2 - 2E[x_0]E[x]\right) \\ &= \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right)\left\{E[x^2] - E[2xx_0] + E[x_0^2] - E[x^2] - E[x_0]^2 - 2E[x_0]E[x]\right\} \\ &= \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right)\left\{E[x^2] - \cancel{2x_0E[x]} + \cancel{x_0^2} - E[x]^2 - \cancel{x_0^2} - \cancel{2x_0E[x]}\right\} \\ &= \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right)\left\{E[x^2] - E[x]^2\right\} = \left(\left.\frac{dy}{dx}\right|_{x=x_0}\right) V(x) \end{aligned}$$

$\boxed{\sigma_y \approx \left|\frac{dy}{dx}\right| \sigma_x}$ → Only valid for small errors

Error Propagation — Two variables $y(x,y)$

$\boxed{V(y) = \left(\frac{dy}{dx}\right)^2 V(x) + 2\left(\frac{dy}{dx}\right)\left(\frac{dy}{dy}\right) \mathrm{Cov}(x,y) + \left(\frac{dy}{dy}\right)^2 V(y)}$ $\qquad \sigma_y^2 = \left(\frac{dy}{dx}\right)^2\sigma_x^2 + \left(\frac{dy}{dy}\right)^2\sigma_y^2 + 2\left(\frac{dy}{dx}\right)\left(\frac{dy}{dy}\right)\rho\,\sigma_x\sigma_y$

Error propogation - Many variables

$$\mathrm{Cov}(y_k, y_l) = \sum_i \sum_j \left(\frac{\partial y_k}{\partial x_i}\right)\left(\frac{\partial y_l}{\partial x_j}\right) \mathrm{Cov}(x_i, x_j)$$

In terms of Jacobian Matrix → Jacobian Matrix $\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \\ \frac{\partial f_m}{\partial x_1} & & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$

$f: \mathbb{R}^n \to \mathbb{R}^m$ where $\frac{\partial f_i}{\partial x_j}$ - partial derivative of i^{th} output function wrt j^{th} input variable

$$V_y = \underline{\underline{J}}\, \underline{\underline{V}}\, \underline{\underline{J}}^T$$

Averaging and Combining Measurements

$$\mu = \frac{1}{N} \sum_i \mu_i$$

$$V(\mu) = \frac{1}{N^2} \sum V_i = \frac{\sigma^2}{N}$$ → Standard deviation falls like: $\sigma \to \frac{\sigma}{\sqrt{N}}$

Combining measurements with different uncertainties

'Weighted' average according to precision

$$\mu = \frac{\sum_i (\mu_i/\sigma_i^2)}{\sum 1/\sigma_i^2} \qquad V(\mu) = \frac{1}{\sum_i (1/\sigma_i^2)}$$

Estimates

Measurement (or estimate) of a parameter based on a limited no. observations

Produced by 'estimator'

$\hat{\theta} = f(x_i)$ — estimate ← $\hat{\theta}$; f → estimator; x_i → observables in data

Note: $\hat{\theta} \neq \theta_0$ — $\hat{\theta}$ → Estimator; θ_0 → True value

Consistency

↳ Estimator consistent - estimate produced tends to true value as data size increases

Difference should get smaller as $N \to \infty$

$\lim_{N\to\infty} \hat{\theta} = \theta_0$ → Variance ↓ as N ↑

Bias

↳ Deviation of the expectation of the estimate, $E[\hat{\theta}]$ from true value θ

$$b_N(\hat{\theta}) = E[\hat{\theta}] - \theta_0 = E[\hat{\theta} - \theta_0]$$

Estimator unbiased: for any Sample Size, N. $b_N(\hat{y}) = 0$

Efficency

estimator produces narrow spread of possible values, minimal variance

Bias - Variance Trade Off

Often trade off between Bias and Efficiency
↳ Mathematical description of accuracy vs precision

Estimates of mean, variance and Standard deviation

Mean:

Arithmetic mean of Sample → Consistent and Unbiased as an estimate → True mean of Population

$$\hat{\mu} = E[x] = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Central Limit Theorem tells us variance of the estimate :
(ie Variance of the estimate – Not the estimate of Variance)

$$\hat{V}(\bar{x}) = \frac{\sigma^2}{N}$$

σ^2 → Variance of parent distribution
N → Size of Sample

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Variance:

1) If we know the true mean, μ:

$$\hat{V}(x) = \frac{1}{N}\sum_i (x_i - \mu)^2$$

$$E[V(x)] = \frac{N E[(x-\mu)^2]}{N} = V(x)$$

↳ unbiased and consistent estimate

2) Usually do not know μ_{true}, thus estimate $\hat{\mu}$:

$$\hat{V} = \frac{1}{N}\Sigma(x_i - \bar{x})^2 = \frac{1}{N}\Sigma(x_i^2 - \bar{x}^2)$$

Take expectation:

$$E[\hat{V}] = \frac{N E[x^2 - \bar{x}^2]}{N} = E[x^2] - E[\bar{x}^2]$$

$$= \{\underbrace{E[x^2] - E[x]^2}_{V(x)}\} - \{\underbrace{E[\bar{x}^2] - E[\bar{x}]^2}_{V(\bar{x}) = \frac{V(x)}{N}}\}$$

↳ variance on mean est.

$$= V(x) - \frac{V(x)}{N}$$

$$E[\hat{V}] = \frac{N-1}{N} V(x) \neq V(x)$$

↳ Thus biased As $N \to \infty$, $\frac{N-1}{N} \to 0$

Shows origin of Bessel Function

$$\hat{V} = S^2 = \frac{1}{N-1}\Sigma(x_i - \bar{x})^2$$

→ Exactly Compensates for bias
Consistent and unbiased

Standard error on Variance

$$V(\hat{V}) = \frac{2\sigma^4}{N-1}$$

Standard Deviation on Variance

$$\hat{\sigma} = \sqrt{\hat{v}} = S$$
$$= \frac{\sigma}{\sqrt{2(N-1)}}$$

Summary of Sample estimates

Given i.i.d Sample $x_i = \{x_1, x_2, \dots, x_N\}$

Mean: $\hat{\mu} = \frac{1}{N} \sum x_i$

Std dev: $\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_i (x_i - \hat{\mu})^2}$

Error on mean $\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{N}}$

Error on Standard $\hat{\sigma}_{\hat{\sigma}} = \frac{\hat{\sigma}}{\sqrt{2(N-1)}}$

The likelihood function

↳ likelihood of observing the data set we have given a particular set of parameters

For N independent observations of random variable X $\{x_1, \dots, x_N\}$

p.d.f: $p(X|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$

Replace variable x with particular set of observations x_0

$$L(\theta) = L(x_0, \theta) \equiv p(x_0|\theta)$$

Maximum Likelihood Method

Estimate of θ obtained by finding value $\hat{\theta}$ which minimises $L(\theta)$

$$\left.\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

Log-Likelihood

$$\text{Log}(L(\theta)) = \sum_{i=1}^{N} \ln p(x_i|\theta)$$

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = \frac{\partial}{\partial \theta} \sum_{i=1}^{N} \ln p(x_i|\theta) = 0$$

→ often requires numerical minimisation.

Example: Normal Dist

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\ln p(x; \mu, \sigma) = -\frac{(x-\mu)^2}{2\sigma^2} + C$$

The Score

$S(\theta) = \frac{\partial \ln L(\theta)}{\partial(\theta)}$ partial derivative of likelihood function wrt θ

Maximum Likelihood Estimation: Where $S(\theta) = 0$

Expectation Value of Score

Liklihood function gives weight: expected Score function over all possible observations x given θ.

$$E[S(\theta)] = \int \frac{\partial \ln L(\theta)}{\partial \theta} L(\theta)\, dx = 0$$

→ total p.d.f is normalised $\int L(\theta) d\,dx = \int \prod_i p(x_i;\theta)\, dx = 1$

$\int L(\theta)\,dx$

$\hookrightarrow \frac{\partial}{\partial\theta}\underbrace{\int L(\theta)\,dx}_{1} = \int \frac{\partial L(\theta)}{\partial\theta}\,dx = \int \frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx$

$= \frac{\partial}{\partial\theta}\cdot 1 = 0$

Overall: $\int \frac{\partial \ln L(\theta)}{\partial\theta} L(\theta) = 0 \quad \therefore E[S(\theta)] = 0$ QED.

Information (variance of score)

$I(\theta) = V(S(\theta)) = E[S(\theta)^2] - \cancel{E[S(\theta)]^2}^{0:\text{above}} = E[S(\theta)^2]$

$= E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right]$

Using knowledge

$E[S(\theta)] = E\left[\frac{\partial \ln L(\theta)}{\partial\theta}\right] = \int \frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx = 0$

$\frac{\partial}{\partial\theta} E[S(\theta)] = \frac{\partial}{\partial\theta}\int \frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx$ $\quad \left[\frac{\partial \ln L(\theta)}{\partial\theta}\cdot L(\theta)\right]$

$= \int \frac{\partial^2 \ln L(\theta)}{\partial\theta^2} L(\theta)\,dx + \int \frac{\partial \ln L(\theta)}{\partial\theta}\frac{\partial L(\theta)}{\partial\theta}\,dx$

$= \int \frac{\partial^2 \ln L(\theta)}{\partial\theta^2} L(\theta)\,dx + \int \left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2 L(\theta)\,dx = 0$

$\int \frac{\partial^2 \ln L(\theta)}{\partial\theta^2} L(\theta)\,dx = -\int \left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2 L(\theta)\,dx$

$= E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta^2}\right] = -E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right] = -I(\theta)$

$$\boxed{I(\theta) = E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta^2}\right]}$$

The Minimum Variance Bound

Def: Unbiased $\hat{\theta}$

$E[\hat{\theta}] = \theta = \int \hat{\theta} L(\theta)\,dx = \theta$

Differentiate w.r.t θ (provided integral limits do not depend on θ)

$\int \hat{\theta}\frac{\partial L(\theta)}{\partial\theta}\,dx = 1$ $\quad \left[\frac{\partial L(\theta)}{\partial\theta} = \frac{\partial \ln L(\theta)}{\partial\theta}\cdot L(\theta)\right]$

$\int \hat{\theta}\frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx = 1$

Substract $\theta\cdot E[S(\theta)] = \int \theta\frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx = 0$

$\int (\hat{\theta}-\theta)\frac{\partial \ln L(\theta)}{\partial\theta} L(\theta)\,dx = 1$ Exploit Cauchy Schwartz Inequality: $\int u^2 dx \int v^2 dx \geq \left(\int uv\,dx\right)^2$

With $u = (\hat{\theta}-\theta)\sqrt{L(\theta)}$ $\quad v = \frac{\partial \ln L(\theta)}{\partial\theta}\sqrt{L(\theta)}$

$$\left(\int uv\,dx\right)^2 = \left(\int (\hat{\theta}-\theta)\frac{\partial \ln(L(\theta))}{\partial\theta} L(\theta)\,dx\right)^2 = 1 \leq \int (\hat{\theta}-\theta)^2 L(\theta)\,dx \int \left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2 L(\theta)\,dx$$

$$1 \leq \underbrace{\int (\hat{\theta}-\theta)^2 L(\theta)\,dx}_{V(\hat{\theta})} \underbrace{\int \left(\frac{\partial \ln L}{\partial \theta}\right)^2 L(\theta)\,dx}_{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]}$$

$$V(\hat{\theta}) \geq \left(E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]\right)^{-1}$$

$$\geq -\left(E\left[\frac{\partial^2 \ln(L)}{\partial \theta^2}\right]\right)^{-1}$$

→ Shows minimum variance bound from likelihood

Allows calc of variance

Can show as $n \to \infty$:

Maximum Likelihood Estimate = Minimum Variance bound

i.e. ML efficient as $n \to \infty$

Many parameter likelihood

Covariance of the estimates:

$$\mathrm{Cov}(\theta_i, \theta_j) = -E\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right] = -\left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\Big|_{\vec{\theta}=\hat{\vec{\theta}}}\right)^{-1}$$

For $N \to \infty$:

ML estimates: Normally distributed estimate for $\hat{\theta}$

Unbiased.

distributed about the true value

Variance ≈ Minimum variance bound

Useful Properties of MLE

$$E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \frac{\partial^2 \ln(L)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}$$

At minimum variance bound $V(\hat{\theta}) = -E\left[\frac{\partial^2 \ln(L)}{\partial \theta^2}\right]^{-1}$

Large Sample → ML is best estimator (Unbiased, Consistent, efficient)

Small N → Variable

Profiled Log-Likelihood

↳ Scan log likelihood in dimension of interest.

At each point; fix parameter to that value

Minimize $-\ln L$ with respect to all other parameters

Uncertainties profiled log-likelihoods

nσ error $-\Delta \ln L = \frac{n^2}{2}$

Extended Maximum Likelihood

Previously: assumed total probability density is normalised

↳ i.e: for N i.i.d observations x_i with $x \sim f(x, \vec{\theta})$

$$\int f(x;\theta)\,dx = 1 \qquad L(\theta) = \prod_i^N f(x_i;\vec{\theta})$$

Often: Total no. events in dataset is also a random variable

Wish to determine overall scale/yield as well as shape parameters

N — Poisson distributed random variable with expectation ν

$N \sim \mathrm{Poiss}(N;\nu)$

Likelihood

↳ $$L(\nu,\vec{\theta}) = \frac{\nu^N e^{-\nu}}{N!} \prod_{i=1}^{N} f(x_i;\theta)$$

$$\ln L(\nu,\vec{\theta}) = N\ln\nu - \nu - \ln(N!) + \sum_{i=1}^{N} \ln f(x_i;\vec{\theta})$$

Binned Maximum Likelihood

Used for very large datasets → more computationally efficient
More numerically stable

Dataset split into N bins

Relevent p.d.f is now poisson with expectation ν_b for each bin.

Observed data in each bin n_b:

$$p(x_b) = \frac{\nu_b^{n_b} e^{-\nu_b}}{n_b!}$$

n_b – no. observed events

ν – Poisson expectation

$$\nu_b = n_b \int_{b_l}^{b_u} f(x;\theta)\,dx$$

If c.d.f known: $f_b(\vec{\theta}) = n_b [F(b_u;\vec{\theta}) - F(b_l;\vec{\theta})]$

b_u – upper edge bin boundary

b_l – lower edge bin boundary

Likelihood:

$$L = \prod_{b=1}^{N} \frac{\nu_b^{n_b} e^{-\nu_b}}{n_b}$$

$$-\ln L = -\ln \prod_{b=1}^{N} \frac{\nu_b^{n_b} e^{-\nu_b}}{n_b!}$$

$$= -\sum n_b \ln \nu_b - \nu_b - \ln(n_b!)$$

Stirling approx: $\ln(n!) = n\ln n - n$

$$-\sum [n_b \ln \nu_b - \nu_b - n_b \ln(n_b) + n_b]$$

$$-\ln L(\theta) = -\sum_{b=1}^{N} [n_b \ln(F(h_u,\theta) - F(h_l,\theta)) - (F(h_u,\theta) - F(h_l,\theta))] + C$$

Least Squares Method

$$\chi^2 = \sum_{i=1}^{N} (y_i^{obs} - y_i^{pred})^2 = \sum_{i=1}^{N} [y_i - f(x_i,\theta)]^2$$

→ Assumes all y_i are meased to same precision

If y_i have associated errors σ_i

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i;\theta))^2}{\sigma_i^2}$$

Number of Degrees of Freedom

$$n_{d.o.f} = n_{bins} - n_{params}$$

Binned Chi Squares Fits

For Sufficiently large N in each bin:

→ y_i - Sum of events in bin

y_{error} - $\sqrt{y_i}$

→ As poisson distributed

Relationship to log likelihood (if Gaussian Distributed)

→ x_i , $y_i \pm \sigma_i$ (gaussian distributed)

$$L(\vec{\theta}) = \prod_{i=1} \text{Gauss}(y_i - f(x_i,\sigma_i))$$

$$= \prod_{i=1} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{(y_i - f(x_i,\vec{\theta}))^2}{2\sigma_i^2}\right]$$

$$\ln L(\theta) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) + \sum_{i=1}^{N} -\frac{1}{2}\frac{(y_i - f(x_i,\vec{\theta}))^2}{\sigma_i^2} = \sum_{i=1}^{N} -\frac{1}{2}\frac{(y_i - f(x_i,\vec{\theta}))^2}{\sigma_i^2} + C$$

Notice $\boxed{\chi^2 = -2\ln L + C}$

$\boxed{\Delta\chi^2 = -2\,\Delta\ln L}$

↳ Significant: If uncertainties are normal

$-2\Delta\ln L$ is distributed as χ^2 with $n_{dof} = n_{pars}$

Which under CLT true for $N\to\infty$ ⟶ Thus true for all $N\to\infty$

Wilkes Theorem

As $N\to\infty$:

$\boxed{-2\Delta\ln(\theta) \longrightarrow \chi^2}$

↳ m parameters ↳ m degrees of freedom

Allows drawing contours on profile likelihood plots

↳ Z Score - fraction contained within χ^2 distribution

Log-Likelihood ratio (LLR) $\boxed{\Delta\ln L = \ln L_1 - \ln L_0 = \ln\left(\frac{L_1(\hat{\theta})}{L_0(\hat{\theta})}\right)}$ ⟶ Test Statistic

↳ a function of estimated parameters

$\boxed{T = f(\hat{\theta})}$

Example: LLR

Note drawing contours

Z Score	$2\Delta\ln L$ 1D	$2\Delta\ln L$ 2D	Frac
1	1	2.29575	0.6827
2	4	6.18007	0.9545
3	9	11.8292	0.9973

Method of Moments

↳ Often difficult to implement Maximum Likelihood and Least Squares Methods

Recall Law of Large Numbers

↳ $\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \longrightarrow \mu \text{ as } N\to\infty}$ 'Sample estimates ⟶ true moments for large N'

↳ For α^{th} moment about zero:

$\boxed{\hat{\mu}_\alpha = \frac{1}{N}\sum_{i=1}^{N} x_i^\alpha \longrightarrow \mu_\alpha \text{ as } N\to\infty}$

For α^{th} moment about the mean:

$\boxed{\mu_\alpha' = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^\alpha \longrightarrow \mu_\alpha \text{ as } N\to\infty}$

Data: random variable $x_i = \{x_1, x_2, \dots, x_N\}$

Model: $f(x;\theta)$

True mean: $\mu = g(\theta) = \int_{-\infty}^{\infty} x f(x;\theta)\,dx$ ⟶ If θ is the mean; $g(\theta) \longrightarrow \mu = \theta$

Estimate $\hat{\mu}$: using law of large numbers

$\bar{x} = \hat{\mu} = \frac{1}{N}\Sigma x_i \longrightarrow \mu$ as $N\to\infty$

$\longrightarrow g(\theta)$ as $N\to\infty$

Estimate $\hat{\theta}$: Invert $g(\theta)$ - Solve for θ

$\boxed{\hat{\theta} = g^{-1}(\hat{\mu})}$ (Method of Moments estimate of θ)

For m params $\vec{\theta} = \{\theta_1, \dots, \theta_m\}$

↳ m unknowns

Cannot Solve $\nu = g(\vec{\theta})$

Exploiting Higher Powers

m^{th} moment: $\hat{\nu}_\alpha = \frac{1}{N} \sum_i x_i^\alpha \longrightarrow \nu_\alpha$ as $N \longrightarrow \infty$

have m Simultaneous equations:

$\nu_1 = g_1(\theta_1, \theta_2, \dots, \theta_m)$
$\nu_2 = g_2(\vec{\theta})$
$\vdots$
$\nu_m = g_m(\vec{\theta})$

Gives m functions $g(\vec{\theta})$

θ is expressed as a function $\vec{h}(\vec{\nu})$ of moments ie $\theta = h_1(\vec{\nu})$

Plug in Sample estimates of moments, $\vec{\nu}$ ⟶ gives estimates $\vec{\theta}$

Uncertainties from MoM estimates

Estimate of Covarience of moment - with Bessel Correction

$$\widehat{Cov}(\hat{\nu}_\alpha, \hat{\nu}_\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i^\alpha - \bar{x}_i)(x_i^\beta - \bar{x}_i)$$

Propogate to Covariance on parameters (θ)

↳ Using functions of θ in terms of ν

$$Cov(\theta_i, \theta_j) = \sum_{k,l} \frac{\partial \theta_i}{\partial \nu_k} \frac{\partial \theta_j}{\partial \nu_l} Cov(\nu_k, \nu_l)$$

1) Example with Normal Distribution $\sim N(\nu, \sigma^2)$

1st moment $\nu_1 = \nu$ (mean of normal dist)

2nd moment: $\nu_2 = \nu^2 + \sigma^2$

↳ (St. dev of normal distribution)

Solve for $\vec{\theta} = (\nu, \sigma)$

$\nu = \nu_1$

$\sigma^2 = \nu_2 - \nu_1^2$

Plugging in Sample Moments

$\hat{\nu} = \hat{\nu}_1 = \frac{1}{N} \sum_{i=1}^{N} x_i$

$$\hat{\sigma} = \sqrt{\hat{\nu}_2 - \hat{\nu}_1^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} x_i\right)^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

2) Particle physics example

$x \sim f(x; \alpha) = N(1 + \alpha \cos x)$ ⟶ Normalise by finding N

$$\frac{1}{N} = \int_{-\pi}^{\pi} 1 + \alpha \cos x \, dx = 2\pi \longrightarrow f(x; \alpha) = \frac{1 + \alpha \cos x}{2\pi}$$

1st Moment: $\nu_1 = \int_{-\pi}^{\pi} \frac{x(1 + \alpha \cos x)}{2\pi} dx = 0$ ⟶ ∴ (the $\alpha \cos x$ term integrates to 0)

2nd Moment: $\nu_2 = \int_{-\pi}^{\pi} \frac{x^2(1 + \alpha \cos x)}{2\pi} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} x^2 dx + \frac{1}{2\pi} \int_{-\pi}^{\pi} \alpha x^2 \cos x \, dx$

$$= \frac{1}{2\pi}\left[\frac{x^3}{3}\right]_{-\pi}^{\pi} + \frac{\alpha}{2\pi}\left\{\left[x^2\sin(x)\right]_{-\pi}^{\pi} - \int_{-\pi}^{\pi} 2x\sin x\,dx\right\}$$

$$= \frac{1}{2\pi}\left[\frac{x^3}{3}\right]_{-\pi}^{\pi} + \frac{\alpha}{2\pi}\left\{\left[x^2\sin(x)\right]_{-\pi}^{\pi} + \left[2x\cos x\right]_{-\pi}^{\pi} - \left[2\sin x\right]_{-\pi}^{\pi}\right\}$$

$$= \frac{1}{\pi}\left[\frac{\pi^3}{3}\right] + \frac{\alpha}{2\pi}\left[4\pi\right] = \frac{\pi^2}{3} - 2\alpha$$

Solve for α in terms of N_2

$$\hat{\alpha} = \frac{1}{2}\left(\frac{\pi^2}{3} - \hat{N}_i\right) = \frac{1}{2}\left(\frac{\pi^2}{3} - \bar{x}^2\right)$$

Goodness-of-Fit Tests

- determining how well parameters values and uncertainties match dataset

Test Statistic for Goodness of Fit

- quantify agreement between data and model:

'How well does the data, x_0 agree with the null hypothesis, H_0'

- Hypothesis - model with fitted parameters

For some Test Statistic ~T distributed according to P(T)

- Probability I got this fit or worse:

$$p_{gof} = \int_{T_0}^{\infty} P(T|H_0)$$ — T_0 - T evaluated at data given

Chi-Squared, χ^2 Test Statistic

χ^2 value - great test statistic

- χ^2 with k degrees of freedom:
  - $E[\chi^2] = k$ $\therefore$ $\frac{\chi^2}{d.o.f} \approx 1$

For a χ^2 test: $n_{d.o.f} = n_{obs} - n_{pars}$

- k = no. observations (or bins) − no. free parameters

The Chi Squared Value

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

O_i - Observed value for i^{th} data point

E_i - Expected value for i^{th} data point

Chi Squared Distribution

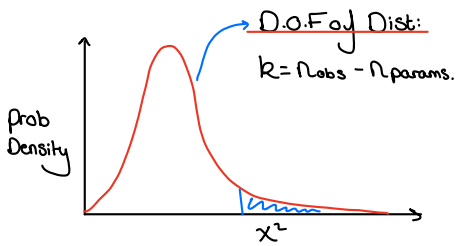
Expected outcome of the sum of Squared 'independant Standard normal variable' under null hypothesis

As DoF increases: Distribution resembles normal

χ^2 Test/Probability: $1 - F(\chi^2)$

where $F(\chi^2)$ - C.d.f of χ^2 dist with k degrees of freedom

- Probability that a function that does define the data has as large or larger χ^2 value than the one you find. — Percentage change of better distribution



Small p-value: model doesn't describe data well (large discrepencies - large χ^2 value)

larger than expected by random chance: Small chance of obtaining such a poor fit (or worse)

if model was true but data independant Standard normally distributed

Large p-value: model describes data too well - overfitting - too many free parameters (very small discrepency)

Very high chance χ^2 value for true model would be worse (larger)

ie overfitting

Step 1: Define Hypothesis, H_0 (model with parameters)

Step 2: Calculate χ^2 value given x_i and model: χ^2

Step 3: Calculate c.d.f of χ^2 distribution with $k = n_{obs} - n_{params}$ D.O.F
With χ^2-value, $F(\chi^2)$

Step 4: Percentage change of better distribution: $1 - F(\chi^2)$

Note: Not necessarily very powerful
Only asseses compatability between data and hypothesis
Not good at comparing two.

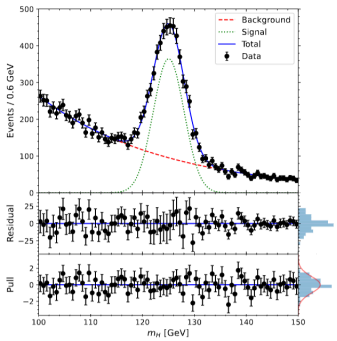
Residuals and Pull Distribution

↳ 'pull': residual normalised by the Error

$$\text{pull} = \frac{n_{obs} - n_{pred}}{\sigma_{obs}}$$

↳ Error on pull - unity

For a good fit: Expect pulls to be 'Standard normally distributed'



Kolmogorov - Smirnoff, KS Test

Two Sample KS Test

Compares two empirical c.d.f's of two datasets to assess whether they are likely drawn from the same underlying distribution

KS score: Maximium deviation between two distributions
c.d.f (times $\sqrt{\text{Sample size}}$)

$$p_{KS} = \sqrt{N} \cdot |\max_x(F(x_1, x) - F(x_2, x))|$$

often Empirical c.d.f:

$$F(x) = \frac{\text{No. elements in sample} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty, x]}(x_i)$$

One Sample KS Test

Compares dataset to specified theoretical distribution
Difference between empirical c.d.f and cdf of hypothesized theoretical distribution

$$p_{KS} = \sqrt{N} \cdot |\max_x (F_n(x_1, x) - F(x))|$$

↳ Hypothesised CDF
↳ Empirical CDF

Confidence Intervals

Classical frequentist Confidence Interval

We believe: "True value of the parameter θ_0 lies within some interval, $\theta_a < \theta_0 < \theta_b$ with some probability, β"

↳ True value of parameter, θ_0 is considered fixed

Confidence Intervals: true value will fall within interval β% of time

Typically quote: 1σ $\beta = 68.3\%$

2σ $\beta = 95.4\%$

Z Score: (Std: 1σ ⟶ Z Score: 1)

Any normal distribution can be translated to the Standard normal

$$Z = \frac{x - \mu}{\sigma}$$

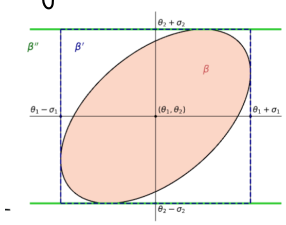
Under Covage: too small a interval

Over Confident.

Over Coverage: too large an interval

'Conservative'

Confidence Levels 2D



	$Z=1$	$Z=2$	$Z=3$
β	0.393	0.865	0.989
β' for $\rho = 0.00$	0.466	0.911	0.995
β' for $\rho = 0.20$	0.471	0.912	0.995
β' for $\rho = 0.50$	0.498	0.917	0.995
β' for $\rho = 0.80$	0.561	0.929	0.995
β' for $\rho = 0.90$	0.596	0.936	0.996
β' for $\rho = 0.95$	0.622	0.941	0.996
β' for $\rho = 1.00$	0.683	0.954	0.997
β''	0.683	0.954	0.997

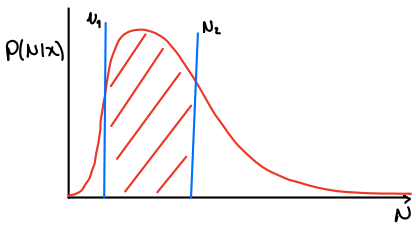
Bayesian Credible Interval

defined on Posterior p.d.f

$$\beta = \int_{\theta_1}^{\theta_2} p(\theta | x)\, d\theta$$

⟶ Infinitely many by shifting θ_2, θ_1 around

Convention: quote narrowest (central) interval



To quote upper limit: Set θ_1 to θ_{min}

Neyman - Pearson (Classical) Intervals

Frequentist Parameter, θ - fixed value

Confidence Interval

$$p(\mu \in [\mu_1, \mu_2]) = \beta$$

Interval is built from the p.d.f of the observation, x - will vary over experiments

Difference from Bayesian

Classical $p(x|\mu)$ - p.d.f of observation given fixed parameter μ

Bayesian $p(\mu|x)$ - fixed observation and varying parameter, μ

For different values, μ

Create Confidence belt: (for central intervals - equal probability either side)

Compute values x_1, x_2 for each μ:

$$p(x < x_1 | \mu) = p(x > x_2 | \mu) = \frac{1-\beta}{2}$$

Constraints in fits

When fitting to a combination of p.d.f's

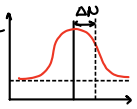
$f_1 P_1 + f_2 P_2 + f_3 P_3$

Constraints: $f_1, f_2, f_3 \geq 0$

: $f_1 + f_2 + f_3 = 1$

Step1: $f_1 P_1 + f_2 P_2 + (1 - f_1 - f_2) P_3$

Step2 (External Constraint)

Example 

$\Delta N = \sigma_{\Delta N}$ - Calibration Shift.

$N' \rightarrow N + \Delta N$

Likelihood Model: $\mathcal{L} = \frac{N}{i} P(x_i; N = \Delta N, \sigma) \times \underbrace{p(N + \Delta N; \Delta N \sigma_{\Delta N})}_{\text{Constraint term}}$

p.d.f of $N + \Delta N$

Estimates and Intervals Near Physical Boundaries

Bayesian Approach Advantages:

Allows removal of physical boundaries using prior.

ie $N \geq 0$: prior $\rightarrow p(N) = 0 \ \forall N > 0$

Frequentist approach:

Frequentist typically allows parameter value and interval which goes into unphysical region

↳ possible to measure value with empty Confidence Interval ⟶ Undercoverage

Suppose physical restriction $N > 0$ is required

Confidence belt cuts off at $N < 0$

Likelihood abruptly stops at physical boundary

↳ Cannot calculate: gradient and double differential near minimum.

Values 'stack up' at boundary

Flip-Flopping (Incorrect)

1. Quote central value and uncertainty above threshold
2. Quote upper limits below some threshold up to the boundary
3. Quote upper limit at boundary for any value beyond boundary

} Leads to discontinuities in the Confidence belt

Feldman-Cousins Method:

↳ Purely frequentist approach:

Creates interval directly from Coverage - determined by iterations

↳ Producing Simulation Samples and Compute were β% lie.

Set Confidence level on one axis ⟶ Corresponds to confidence level on other axis

Procedure: For each X, start with Maximum Likelihood Estimate of N, $\hat{N}$

With physical bound applied

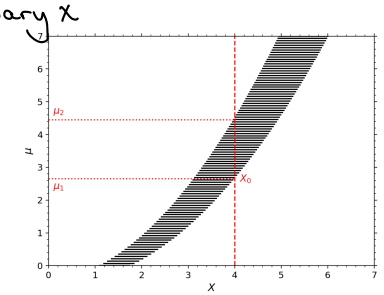
For some fixed (different) N:

Compute Likelihood Ratio: $R = \frac{p(x|N)}{p(x|\hat{N})}$ → Compares likelihood of observing X under two different assumptions

Arrange values of X from higher to lower R

Add values to interval under desired probability content reached

} Can do these in 1 to stop over computing R for unnecessary X



Feldman Cousins - Monte Carlo Simulation

1. Initial Fit and Maximum Likelihood

Run a maximum likelihood fit for a dataset to find best estimate $\hat{N}$ of parameter N

Calculate likelihood value, to serve as reference ($-2\ln L(\hat{N})$)

2. Choosing Test values of N

To find Confidence Interval - Choose some 'test' values of N around $\hat{N}$

For each test value N_0

↳ Wish to determine confidence level that corresponds to it

Iterate over values until you find a range that meets desired Confidence level

3. Define Test Statistic

$T = -2\Delta\ln L(N_0)$ → $T = -2\ln(R) = -2(\ln p(x|N) - \ln p(x|\hat{N})) = -2(\ln L(N) - \ln L(\hat{N})) = -2\Delta\ln(N)$

↳ Compares likelihood at N_0 against maximum

4. Generate Pseudo Experiments

Simulate datasets generated under assumption $N = N_0$

For each toy dataset:

↳ repeat max likelihood fitting process to find best fit N_i for that toy

Calculate likelihood at both $\hat{N}_i$ and test value N_0

Use these to calculate Test Statistic: $T_i = -2\Delta\ln(N_0)_i$ for each toy

5. Determine Confidence Level (p-value)

CL at N_0 → fraction of toy experiments where $T_i >$ Observed T

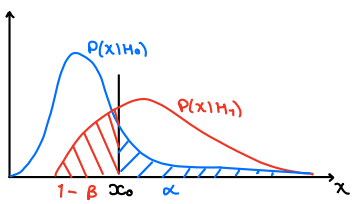
↳ Indicates how often simulated data (under N_0) produces worse results than actual data.

Hypothesis Testing

Goodness of fit: test a hypothesis H_0 against all other hypotheses

Wish to: Test between Null hypothesis, H_0 and alternate hypothesis, H_1

Use PDF of H_0 and H_1 to ascertain how much of each distribution falls above/bellow particular value.



$\alpha = \int_{x_0}^{\infty} p(x|H_0)$ $\beta = \int_{x_0}^{\infty} p(x|H_1)$

2 Hypothesis: H_0, H_1

4 Outcomes:

Accept	True	
H_0	H_0	(Correct)
H_1	H_0	(Incorrect - Type I error)
H_0	H_1	(Correct)
H_1	H_1	(Incorrect - Type II error)

Two types of error:

Type I: "Loss" reject H_0 when true (probility: α)

Type II: "Contamination" accept H_0 when false (probility: $1-\beta$)

α: Probability of incorrectly rejecting H_0

$1-\beta$: Probability of incorrectly accepting H_0

$\alpha = p(x > x_0|H_0)$ $\beta = p(x > x_0|H_1)$

↳ Typically low α (low Type I error rate) favoured over larger $1-\beta$ (high type II error rate).

Power of Test: β

Probability of Correctly rejecting null Hypothesis H_0 (when it is false)
↳ Higher power - test more trustworthy rejection

β: Probability Correctly rejecting H_0

Assume hypothesis depends on some parameter, θ
↳ either θ - continuous
or θ_0, θ_1 - distinct two values.

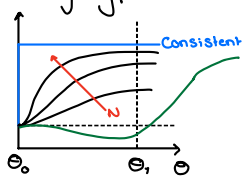
Power ρ

$\rho(\theta_1) = \beta$ in general $\rho(\theta) = \beta(\theta)$

$\rho(\theta_0) = \beta(\theta_0) = \alpha$

Power Curve

↳ function of hypothesis:



Want test to be:

i) Consistent: $\lim_{N\to\infty} P(x > x_0)$ (Sample Size N increaseses, Prob correctly rejecting $H_0 = 1$)

ii) unbiased: $p(\theta_1|H_0) > p(\theta_0|H_0)$ (Higher probability rejecting H_0 under θ_1 than θ_0

The Neyman-Pearson Lemma

↳ Most powerful test: maximises power

Critical Region, W_α:

region in x which is rejected $W_\alpha \in [x_0, \infty]$

w.r.t rv $\vec{x} = (x_1, x_2, \ldots, x_N)$ w p.d.f $f(x|\theta)$

Type 1 error: $\alpha = \int_{x_0}^{\infty} f(\vec{x}|\theta_0)\, dx$

Type II error: $1 - \beta = 1 - \int_{x_0}^{\infty} f(\vec{x}|\theta_1)\, dx$

find W_α (given fixed α) that maximises β (ie Maximal power by choosing optimal w)

$$\beta = \int_{x_0}^{\infty} \frac{f(\vec{x}|\theta_1)}{f(\vec{x}|\theta_0)} f(\vec{x}|\theta_0)\, dx = E_{W_\alpha}\left[\left. \frac{f(\vec{x}|\theta_1)}{f(\vec{x}|\theta_0)} \right|_{\theta=\theta_0} \right]$$

↳ maximal where: $\frac{f(\vec{x}|\theta_1)}{f(\vec{x}|\theta_0)}$ likelihood ratio maximal.
With constraint α.

Best most powerful test:

$$T = -2\ln\left(\frac{L(x|H_0)}{L(x|H_1)}\right)$$

↳ T will be χ^2 distributed ($k = 1$)

$$f(\vec{x}|\theta_1) = f(x_1|\theta_1) f(x_2|\theta_1) \ldots f(x_N|\theta_1)$$
$$= \Pi_i f(x_i|\theta_1) = L(\theta_1$$

Limit Setting

Useful when insufficient precision to make a measurement.
↳ ie: Given didnt see evidence for H_1, what is the confidence level of rejecting H_1
↳ Ie For a given confidence level, which H_1's can I reject

Signal Searches

Consider model of Hypotheses:

$$p(\vec{x};\vec{\theta}) = \mu \cdot s(\vec{x};\vec{\theta}_s) + (1-\mu)\cdot b(\vec{x};\vec{\theta}_b)$$

↳ Set of H dependant on μ - each of which dependent on $\vec{\theta}_b$

H_0: null hypothesis (background only) Subset $p(x;\vec{\theta})$ where $\mu = 0$.

H_1: Continuous set of models with different values μ.

↳ $\hat{\mu}$ - best fit value of μ for the given data under alternative hypothesis, H_1

Maximises likelihood function for observed data

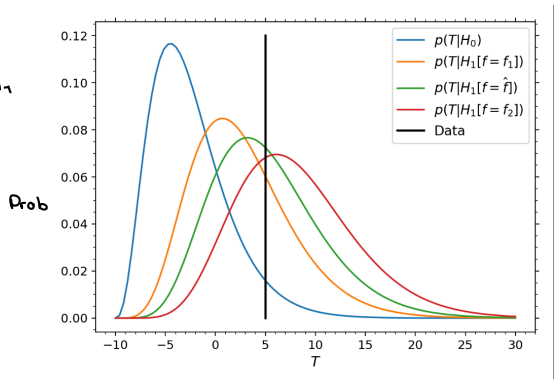
'value μ maximises favourability of rejecting H_0 given data.

Changing into 'T Space' $x \to T$

Once determined $\hat{\mu}$ by maximising likelihood of H_1

Calculate T (test statistic):

$$T = -2\ln\left(\frac{L(\mu=\hat{\mu})}{L(\mu=0)}\right)$$



Also Convert $x \to T$

$p(x|H) \to p(x|T)$

In T space: $\alpha = \int_{T_0}^{\infty} p(T|H_0)\,dT$ $\quad 1-\beta = \int_{-\infty}^{T_0} p(T|H_1)\,dT$

Observed Limit

For given observed data:

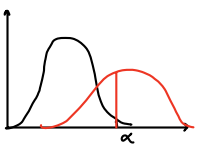
With confidence level β can limit values μ

ie: probability of incorrectly accepting H_0:

$$1-\beta = \int_{-\infty}^{T_0} p(T|H_1(\mu_{lim}))\,dT$$

'Values of $\mu > \mu_{lim}$ excluded at β% Confidence level'

Note: One Sided vs two Sided tests

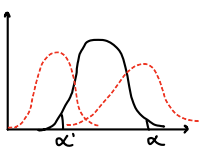


One Sided:

eg null hypothesis → no signal

Alternate → presence Signal

$p_{val} = \alpha$



Two-Sided test

$p_{val} = \alpha + \alpha'$

Convert Z-Score accordingly

Expected Limits/Bands

Show the expected limits on μ if H_0 is actually true.

Simulate many pseudo-datasets under H_0

↳ Calculating μ_{lim} under pseudo-data set

Determine μ_{lim} median - Expected Limit

Determine μ_{lim} Spread quantile to some confidence level - Expected Band

Example:

For given observed data

↳ Calculate $\hat{\mu}$

Simulate set pseudo experiments H_1 ($\mu = \hat{\mu}$)

↳ Calculate T for each pseudo-experiment.

↳ Distribution T under H_1 ($\mu = \mu_0$) → Gives $T_{exp}(H_1, \mu = \hat{\mu})$

Simulate set pseudo experiments H_1 ($\mu = 0$)

↳ Repeat. → Gives $T_{exp}(H_0, \mu = 0)$

Compare T_0 with $T_{exp}(H_0)$ and $T_{exp}(H_1)$

↳ T_0 closer to $T_{exp}(H_0)$ – data more consistent with H_0

T_0 closer to $T_{exp}(H_1)$ – data more consistent with H_1

CL_s Method

Issues arise when: data inconsistent with both H_0 and H_1

or poor sensitivity to seperation

CL_s Method: Down weighting 'p-value' value by exclusion

p_i - probability of attaining a value as extreme or more extreme

p_b - Background only

$$p_b = \int_{T_0}^{\infty} p(T|H_0)\,dT \quad \text{or} \quad \int_{-\infty}^{T_0} p(T|H_0)\,dT$$

p_{sb} - Signal and Background

$$p_{sb} = \int_{T_0}^{\infty} p(T|H_1)\,dT \quad \text{or} \quad \int_{-\infty}^{T_0} p(T|H_1)\,dT$$

Adjusted p-value:

$$\boxed{p_s = \frac{p_{sb}}{1 - p_b}}$$

Sometimes written $p_b \sim CL_b$ $p_{sb} \sim CL_{sb}$

Reduces risk of false rejecting signal hypothesis due to an overlap between H_0 and H_1 distributions. ⟶ Intentionally conservative

Resampling Methods:

↳ Excellent for estimation methods and Cross checking

Jackknife Resampling

$X_i = \{x_1, x_2, \ldots, x_N\}$

$\hat{\theta}_N$ - Biased estimate of θ (on N observations)

Can write out bias as a Taylor expansion in N:

$$E[\hat{\theta}_N - \theta] = \underbrace{\frac{a_1}{N}}_{\text{bias at } O(N)} + \frac{a_2}{N^2} + \frac{a_3}{N^3}$$

Defn: $\boxed{\theta' = N\hat{\theta}_N - (N-1)\hat{\theta}_{N-1}}$

$$E[\theta'] = N\left(\theta + \frac{a_1}{N} + \frac{a_2}{N^2} + \ldots\right) - (N-1)\left(\theta + \frac{a_1}{N-1} + \frac{a_2}{(N-1)^2} + \frac{a_3}{(N-1)^3} + \ldots\right)$$

$$= N\theta + a_1 + \frac{a_2}{N} + \frac{a_3}{N^2} - (N-1)\theta - a_1 - \frac{a_2}{N-1} - \frac{a_3}{(N-1)^2} + \ldots$$

$$= \theta + \frac{a_2}{N(N-1)} + \ldots = \theta + O\left(\frac{1}{N^2}\right)$$

↳ θ' is biased to $O\left(\frac{1}{N^2}\right)$

Produce: N jackknife samples omitted one point at a time

$x_{(1)} = \{x_2, x_3, \ldots, x_N\}$ → $\hat{\theta}_{(1)}$

$x_{(2)} = \{x_1, x_3, \ldots, x_N\}$ → $\hat{\theta}_{(2)}$

$x_{(N)} = \{x_1, x_2, \ldots, x_{N-1}\}$ → $\hat{\theta}_{(N)}$

Create jackknife estimates using: $\hat{\theta} = f(x)$

Note: $\hat{\theta}_{(i)} = f(x_1, \ldots, x_{i-1}, x_{i+1}, x_N)$

Average of jackknife: $\hat{\theta}_{(\cdot)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{(i)}$

From which estimate bias: $\hat{b}(\hat{\theta}) = (N-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$ → **Bias Corrected estimate**

$\hat{\theta}_c = \hat{\theta} - \hat{b}(\hat{\theta}) = \hat{\theta} - (N-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}$

Often See Jackknife pseudo values:

$\tilde{\theta}_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(i)}$

↳ average of these gives bias corrected est

$\hat{\theta}_c = \frac{1}{N}\tilde{\theta}_{(i)} = \frac{1}{N}\Sigma\,(N\hat{\theta} - (N-1)\hat{\theta}_{(i)})$

$= N\hat{\theta} - \frac{(N-1)}{N}\Sigma\hat{\theta}_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}$

$= N\hat{\theta} - \frac{(N-1)}{N}\Sigma\hat{\theta}_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}$

Jackknife variance:

↳ "pseudo values" - $\tilde{\theta}_{(i)}$ are approx i.i.d

Sample variance, $\hat{S}^2$, givenest of variance on the estimate

Using CLT (std error on mean)

$\hat{V}(\hat{\theta}_c) = \frac{\hat{S}^2}{N} = \frac{1}{N}\frac{1}{N-1}\Sigma(\tilde{\theta}_{(i)} - \hat{\theta}_c)^2$

$= \frac{1}{N}\cdot\frac{1}{N-1}\Sigma\,[N\hat{\theta} - (N-1)\hat{\theta}_{(i)} - N(\hat{\theta}_{(i)}) - N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)}]^2$

$= \frac{N-1}{N}\sum_{i=1}^{N}\left[\hat{\theta}_{(i)} - \frac{1}{N}\Sigma\hat{\theta}_{(j)}\right]^2 = \boxed{\frac{N-1}{N}\sum_{i=1}^{N}\left[\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right]^2}$

↳ variance on jk average w/ $\frac{N-1}{N}$ Correctio

Boot Strap (non-parametric) resampling

Sample with replacement:

↳ Same Sample size as original

Randomly Choose N points - with replacement

Commonly end up with duplicate/missing events in Sample.

Bootstraping Intervals

↳ Can produce Confidence intervals from quantiles of boot strap distribution

Large Samples needed for reasonable accuracy at high/low CL

↳ Typically accurate to $1/N$

Empirical CDF: $\tilde{F}$ **with Inverse:** $\Phi = \tilde{F}^{-1}$

$[\theta_l, \theta_h] = [\Phi(\alpha/2), \Phi(1-\alpha/2)] = [\Phi(z_\alpha), \Phi(z_{1-\alpha})]$

α - Confidence level of interval

Bias-Corrected-accelerated (BCa) Intervals

Corrects for bias and skewness in Bootstrap distribution.

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}} = \underbrace{N(0,1)}_{\text{Standard Normal}}$$

Bias Correction:

$$[\theta_L, \theta_U] = [\tilde{\Phi}(\hat{z}_0 + z_\alpha), \tilde{\Phi}(\hat{z}_0 + z_{1-\alpha})]$$

↳ $\tilde{z}_0$ is a bias parameter - estimated from bootstraped sample

→ $$\frac{\hat{\theta}-\theta}{\hat{\sigma}} = N(-z_0, 1)$$

Additional acceleration

$$[\theta_L, \theta_U] = \left[\tilde{\Phi}\left(\tilde{z}_0 + \frac{\tilde{z}_0 + z_\alpha}{1-\hat{a}(\hat{z}_0+z_\alpha)}\right), \tilde{\Phi}\left(\tilde{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha}}{1-\hat{a}(\hat{z}_0+z_{1-\alpha})}\right)\right]$$

↳ $\hat{a}$ is accelerated using a jackknife of sample

→ $$\frac{\hat{\theta}-\theta}{\hat{\sigma}} = N\left(-z_0(1+a\theta), (1+a\theta)^2\right)$$

Parametric Bootstrap

Sampling from the p.d.f of the model directly (rather than via replacement)

'Bootstraping based on model itself'.

Used for estimation - and check estimation procedures.

'Pull' of parameters

↳ Should find over an ensemble:

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}} \sim N(0,1)$$ → often plot this distribution. Called 'pull' distribution

Aidan Plot (condensed version)

If instead find:

$N(z_0, a)$ → biased by z_0

Under/overcover by a or $1/a$

Computing Systematic

Bootstrapping procedure:

Change an assumption in procedure and/or modelling

Run parametric bootstrap:

↳ generate under change of assumption

fit back with nomial procedure

Obtain covariance matrix using covariance estimate for BS sample.

Nuisance Parameters

Statistics in Machine Learning

Measurement error (observation error)

↳ difference between measured (observed) value and true value

$$\hat{\theta} - \theta$$

Different from bias:

$$b(\hat{\theta}) = E[\hat{\theta} - \theta]$$

Two potential Sources:

↳ Statistical Fluctuations (random) - not problematic

Systematic Biases - problematic

Forward Modelling

↳ used in supervised machine learning algorithms.

Predictions made based on input features

↳ wish to minimise difference between prediction and some reference target.

Training Sample (size N):

$$\vec{x_i} = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_N}\}$$

Target

$$Y_i = \{Y_1, Y_2, \ldots, Y_N\}$$

Aim: build function:

$$y_i = f(\vec{x_i}, \vec{\theta})$$ $\vec{\theta}$: weights, w_j and biases, b_j

Done by minimising:

$$\chi^2 = \sum_i^N (y_i - Y_i)^2 = \sum_i^N (f(x_i) - Y_i)^2$$

overtraining avoided by:

Statistically independent adjoint Samples

Optimisation:

given a function: $f(\vec{x})$

↳ find values $\vec{x_0}$ which either minimise / maximise $f(\vec{x})$

Optimisation Algorithms

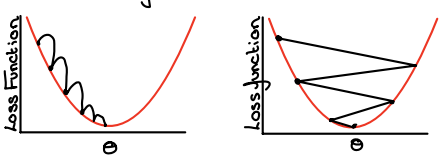
↳ Scipy.optimize

Iminuit - migrad algorithm

Gradient Descent

free parameter: learning rate

↳ Step size of descent.



Local minima/saddle points

↳ Training gets stuck over epochs

Parameters in Optimisations

Scaling parameters

↳ Optimisers are designed for / like parameters $\mathcal{O}(1)$

Scale parameters accordingly

Parameter limits

↳ help limit region in which optimiser is to search.

Regularisation

↳ terms added to cost function in order to steer or change the optimisation outcome.

Help learn trends not statistical fluctuations

Loss function becomes:

$$L = \sum_i^N V(f(x_i), y_i) + \lambda R(f)$$

V - empirical loss function

R - regularisation term

λ - input hyper parameter

- tune importance of regularisation term

Non-Parametric Density Estimation

Histograms as densities

↳ Approximate density based on independent 'control' sample

↳ Created by: Adjoint dataset

MC simulation

For a binned fit: provides a template

Unbinned fit: approximation

↳ Any x value in bin will evaluate to same fit.

Note: Require large independant sample.

↳ avoid/smooth statistical fluctuations.

Kernel Density Estimation

↳ provides smooth continuous distribution (histogram step func ⟶ Smoothing Kernel.

For dataset: $X = \{x_1, x_2, \dots, x_N\}$

Approximate the Shape: $\hat{f}(x) = \frac{1}{N} \sum_i^N K\left(\frac{x - x_i}{h}\right)$

Choice of bandwidth parameter:

Larger Bandwidth

↳ Smoother distribution

More bias

Smaller Bandwidth

↳ More fluctuating distribution

Greater variance.

Decision:

Low Dimensions: by eye

Higher Dimensions: Scott's, Silverman Algorithms

Types of Kernel Shape

↳ Less important than band width.

Common options: Gaussian Exponential

Top hat Linear.

KDEs at boundaries

All kernels typically fall off at end.

Often deal with truncated distributions

↳ Mirror distribution to other side.

Expectation Maximisation

→ Set of observed random variables: $\vec{x}$

Set of hidden random variables: $\vec{y}$

Normal Likelihood:

$$L(\vec{\theta}|x,y) = \prod_i p(x_i, y_i|\vec{\theta})$$

Issue: don't observe $\vec{y}$, so unsure of $P(x,y)$

Question: Is it possible to estimate, $\hat{\theta}$ if we dont observe $\vec{y}$?

Marginalise out y:

$$L(\vec{\theta}|x) = \prod_i p(x_i|\vec{\theta}) = \prod_i \int p_i(x_i, y|\vec{\theta})\, dy$$

$$= \prod_i \int p(x_i|y,\theta)\, p(y|\theta)\, dy$$

→ Maximise to give θ estimate

↳ Still unknown

Using Bayes Theorem:

$$p(y|x,\theta) = \frac{p(x|y,\theta)\,p(y|\theta)}{p(x|\theta)}$$

$$\rightarrow P(x|\theta) = \frac{P(x|y,\theta)\,P(y|\theta)}{P(y|x,\theta)} = \frac{P(x|y,\theta)\,P(y,\theta)}{q(y)} \cdot \frac{q(y)}{P(y|x,\theta)}$$

$$\rightarrow \ln(P(x|\theta)) = \ln\left(\frac{P(x|y,\theta)\,P(y,\theta)}{q(y)}\right) + \ln\left(\frac{q(y)}{P(y|x,\theta)}\right)$$

Expectation over $q(y)$:

$$E[\ln p(x|\theta)] = \underbrace{\int q(y) \ln\left(\frac{P(x|y,\theta)\,P(y,\theta)}{q(y)}\right) dy}_{G(\theta)\text{- Lower Bound Functional}} + \underbrace{\int q(y) \ln\left(\frac{q(y)}{P(y|x,\theta)}\right) dy}_{\text{Kullback-Liebler Divergence}}$$