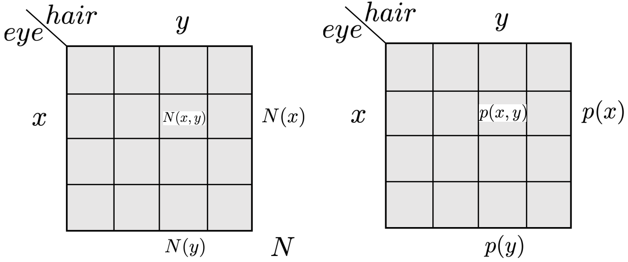
STATS 100A: Two or More Random Variables

Ying Nian Wu

Department of Statistics
University of California, Los Angeles

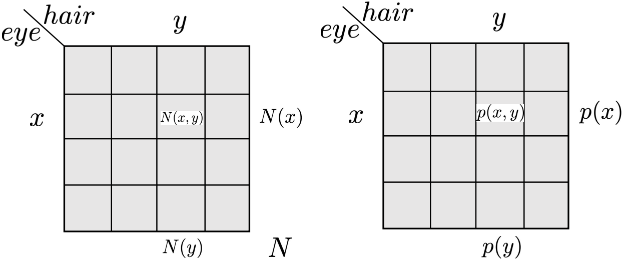
Some pictures are taken from the internet.
Credits belong to original authors.

Discrete distribution



N: number of people in population.
$N(x, y)$: number of people with eye color x and hair color y.
$N(x) = \sum_y N(x, y)$: number of people with eye color x.
$N(y) = \sum_x N(x, y)$: number of people with hair color y.

Joint and marginal



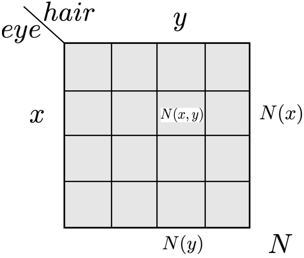
$$p(x,y) = \frac{N(x,y)}{N}.$$

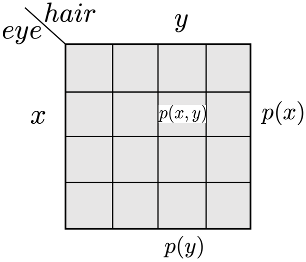
$$p(x) = \frac{N(x)}{N} = \frac{\sum_y N(x,y)}{N} = \sum_y p(x,y).$$

$$p(y) = \frac{N(y)}{N} = \frac{\sum_x N(x,y)}{N} = \sum_x p(x,y).$$

Prob = population proportion ≈ sample proportion / frequency

Conditional



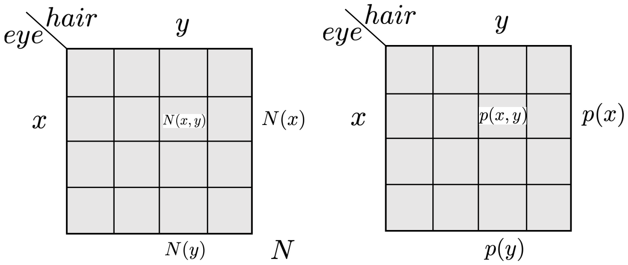


$$p(x|y) = \frac{N(x,y)}{N(y)} = \frac{N(x,y)/N}{N(y)/N} = \frac{p(x,y)}{p(y)}.$$

$$p(y|x) = \frac{N(x,y)}{N(x)} = \frac{N(x,y)/N}{N(x)/N} = \frac{p(x,y)}{p(x)}.$$

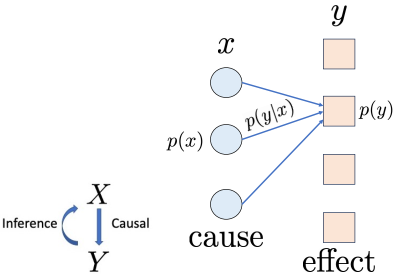
Prob = population proportion ≈ sample proportion / frequency

Rules



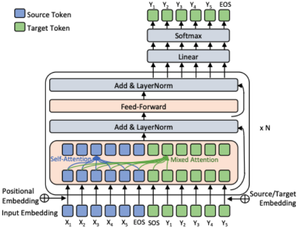
Marginalization: $p(y) = \sum_x p(x, y)$.
Conditioning: $p(x|y) = p(x, y)/p(y)$.
Chain rule: $p(x, y) = p(x)p(y|x)$.

Random variables, probability mass functions



Marginal: prior $p(x) = P(X = x)$, marginal $p(y) = P(Y = y)$.
Conditional: forward generation $p(y|x) = P(Y = y|X = x)$
backward inference $p(x|y) = P(X = x|Y = y)$.
Chain rule: joint $p(x, y) = p(x)p(y|x)$.
Rule of total probability: marginal
$p(y) = \sum_x p(x, y) = \sum_x p(x)p(y|x)$.

Generative Pre-trained Transformer (GPT)

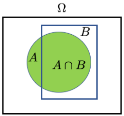


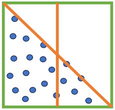
$x = (x_1, ..., x_{T_x})$ (e.g., "Can you write a poem?")
$y = (y_1, ..., y_{T_y})$ (e.g., "Certainly. Below is the poem...")
$p(y|x) = \prod_{t=1}^{T_y} p(y_t|y_{<t}, x)$.
Learn from training data $(x^{(i)}, y^{(i)}, i = 1, ..., n)$ by maximizing

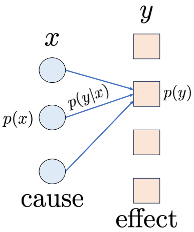
$$\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y^{(i)}|x^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t} \log p_\theta(y_t^{(i)}|y_{<t}^{(i)}, x^{(i)}).$$

memorize and generalize (interpolation).

Bayes rule







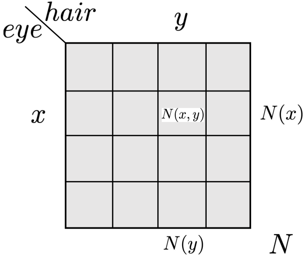
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

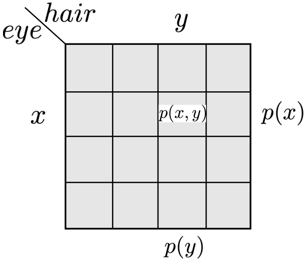
Bayes rule: backward inference, back tracing, posterior

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

$$= \frac{p(x, y)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}.$$

Expectation





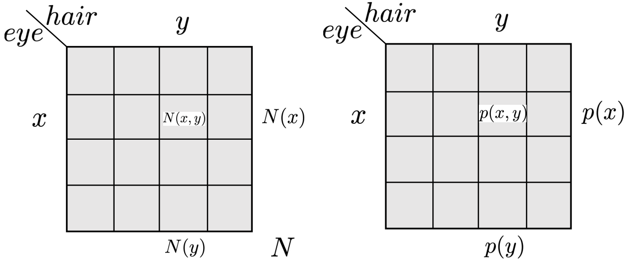
$$\mathbb{E}[h(X,Y)] = \sum_{x,y} h(x,y)p(x,y).$$

Population average or long run average.

$$\frac{1}{N}\sum_{x,y} h(x,y)N(x,y) = \sum_{x,y} h(x,y)\frac{N(x,y)}{N}$$

$$= \sum_{x,y} h(x,y)p(x,y) = \mathbb{E}[h(X,Y)].$$

Expectation



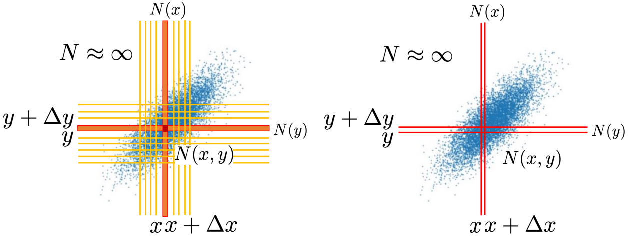
$$\mathbb{E}(X) = \sum_{x,y} xp(x,y) = \sum_x x \sum_y p(x,y) = \sum_x xp(x).$$

same for $\mathbb{E}[h(X)]$.

$$\mathrm{Var}(h(X,Y)) = \mathbb{E}[(h(X,Y) - \mathbb{E}[h(X,Y)])^2].$$

Two continuous random variables

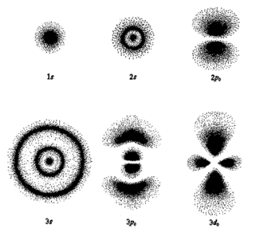
X = height, Y = weight.

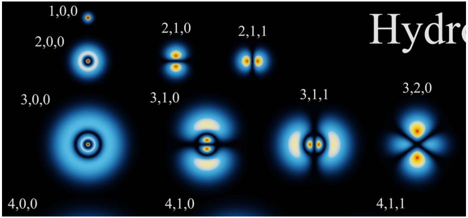


$$f(x, y) = \frac{P(X \in (x, x + \Delta x), Y \in (y, y + \Delta y))}{\Delta x \Delta y} = \frac{N(x, y)/N}{\Delta x \Delta y}.$$

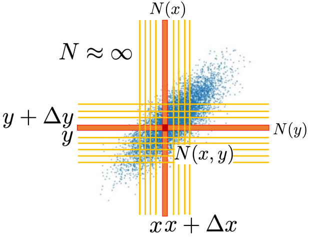
density = probability / size

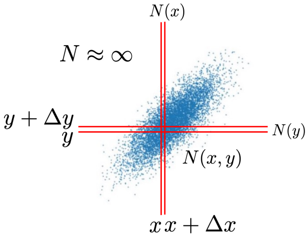
Probability density function





Marginal

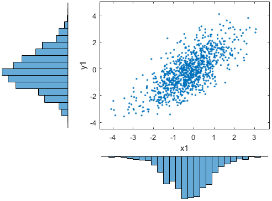


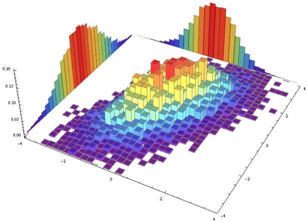


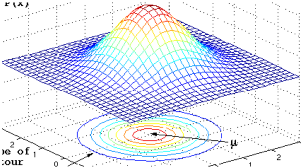
density = prob / size

$$
\begin{aligned}
f(x) &= \frac{P(X \in (x, x + \Delta x))}{\Delta x} = \frac{N(x)/N}{\Delta x} \\
&= \frac{\sum_y N(x, y)/N}{\Delta x} = \frac{\sum_y f(x, y)\Delta x \Delta y}{\Delta x} = \int f(x, y)dy. \\
f(y) &= \int f(x, y)dx.
\end{aligned}
$$

Joint and marginal densities

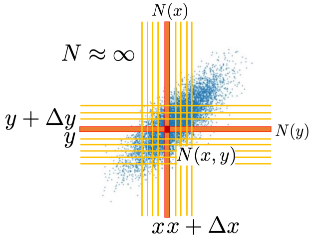


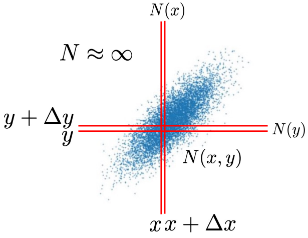




Sample points under the surface, collapse on the plane.

Conditional density



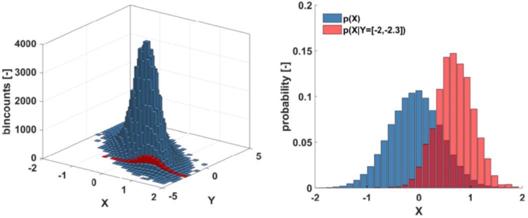


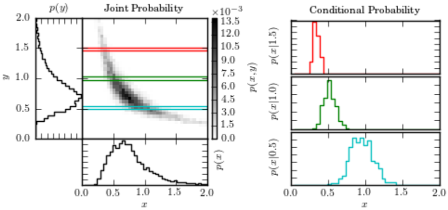
density = prob / size

$$
\begin{aligned}
f(y|x) &= \frac{P(Y \in (y, y+\Delta y) \mid X \in (x, x+\Delta x))}{\Delta y} \\
&= \frac{N(x,y)/N(x)}{\Delta y} = \frac{N(x,y)/N}{(N(x)/N)\Delta y} \\
&= \frac{f(x,y)\Delta x \Delta y}{f(x)\Delta x \Delta y} = \frac{f(x,y)}{f(x)}.
\end{aligned}
$$

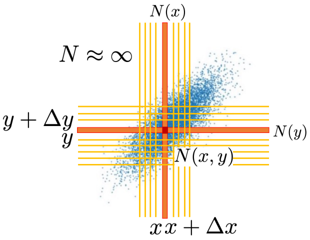
$f(x|y) = f(x,y)/f(y)$.

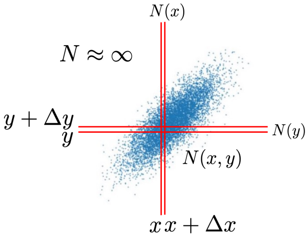
Conditional density

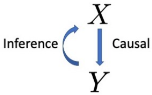




Rules

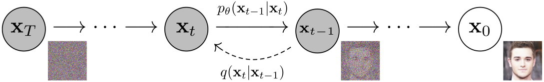






Marginalization: $f(y) = \int f(x, y)dx$.
Normalization (conditioning): $f(x|y) = f(x, y)/f(y)$.
Factorization (chain rule): $f(x, y) = f(x)f(y|x)$.
$f(y|x)$: prediction. $f(x|y)$: inference.

Denoising Diffusion Probability Model



x_0: clean image.

$x_t = x_{t-1} + e_t$, e_t: small noise

Forward noising $q(x_t|x_{t-1})$, $t = 1, ..., T$. x_T: big noise.

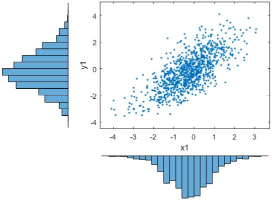
Backward denoising $p(x_{t-1}|x_t)$.

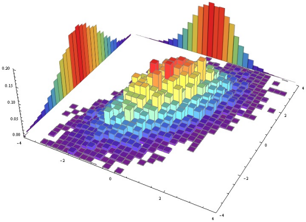
Learn from training data $(x_0^{(i)}, i = 1, ..., n)$ by maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=T}^{1}\log p_\theta(x_{t-1}^{(i)}|x_t^{(i)}).$$

memorize and generalize (interpolation).

Expectation





If $(X, Y) \sim p(x, y)$, then

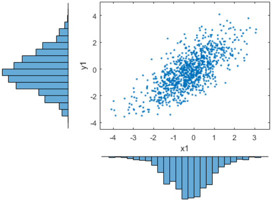
$$\mathbb{E}(h(X, Y)) = \sum_x \sum_y h(x, y)p(x, y).$$

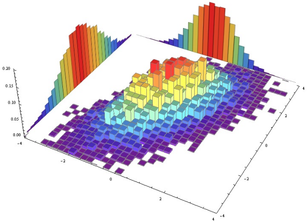
If $(X, Y) \sim f(x, y)$, then

$$\mathbb{E}(h(X, Y)) = \int\int h(x, y)f(x, y)dxdy.$$

$$\mathrm{Var}(h(X, Y)) = \mathbb{E}[(h(X, Y) - \mathbb{E}[h(X, Y)])^2].$$

Expectation

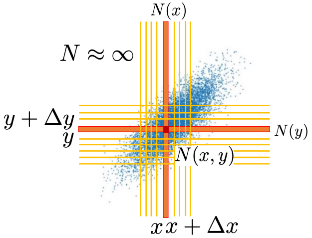


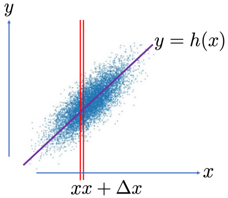


Population average or long run average of $h(X, Y)$.

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i) &= \frac{1}{n}\sum_{\text{cells}} h(x, y) n f(x, y)\Delta x \Delta y \\
&\to \int\int h(x, y) f(x, y) dx dy.
\end{aligned}
$$

Conditional expectation and variance





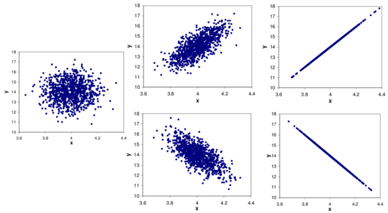
Recall $\mathbb{E}(Y) = \int y f(y) dy$.

$$h(x) = \mathbb{E}[Y|X = x] = \int y f(y|x) dy.$$

Regression, prediction.

$$\text{Var}(Y|X = x) = \mathbb{E}[(Y - h(X))^2 | X = x] = \int (y - h(x))^2 f(y|x) dy.$$

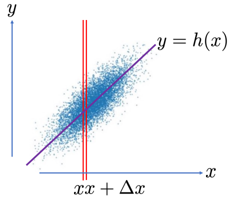
Bivariate Normal



$$X \sim \mathrm{N}(0,1),$$
$$Y = \rho X + \epsilon;\ \epsilon \sim \mathrm{N}(0, 1-\rho^2),\ (|\rho| \leq 1).$$

ϵ is independent of X. Given $X = x$, $Y = \rho x + \epsilon$.

Bivariate Normal



The distribution of points within a vertical slice at x.

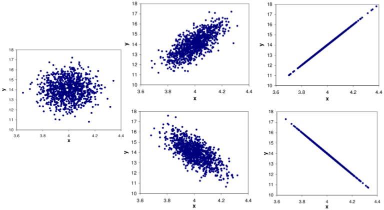
$$\mathbb{E}(Y|X = x) = \mathbb{E}(\rho x + \epsilon) = \rho x.$$

Regression towards the mean ($\rho < 1$), e.g., son's height given father's height.

$$\mathrm{Var}(Y|X = x) = \mathrm{Var}(\rho x + \epsilon) = \mathrm{Var}(\epsilon) = 1 - \rho^2.$$

$$[Y|X = x] \sim \mathrm{N}(\rho x, 1 - \rho^2).$$

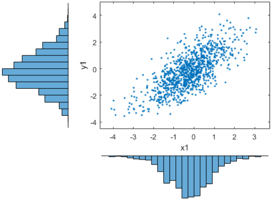
Bivariate Normal

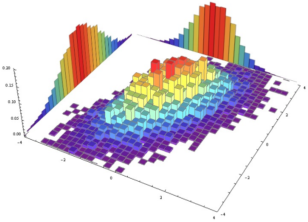


$$
\begin{aligned}
f(x,y) &= f(x)f(y|x) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right].
\end{aligned}
$$

symmetric in (x,y)

Covariance



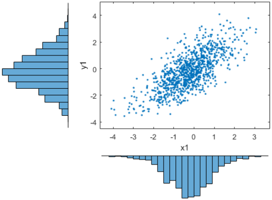


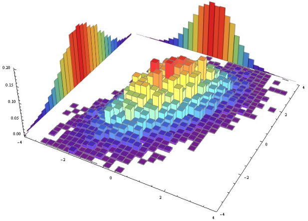
Let $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$, we define the covariance

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

It is defined for both discrete and continuous random variables.

Covariance



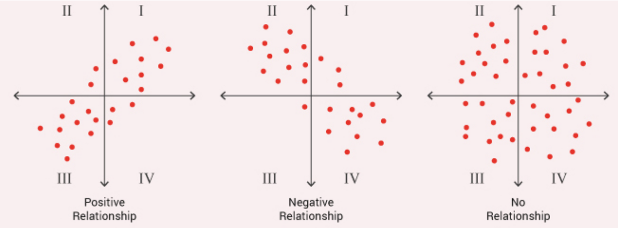


$$(X_i, Y_i) \sim f(x, y),\ i = 1, ..., n.$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

$$\text{Cov}(X, Y) \doteq \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Covariance



$$\text{Cov}(X, Y) \doteq \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

I, III: $(X_i - \bar{X})(Y_i - \bar{Y}) > 0.$
II, IV: $(X_i - \bar{X})(Y_i - \bar{Y}) < 0.$

Covariance

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathbb{E}[XY - \mu_X Y - X\mu_Y + \mu_X \mu_Y] \\
&= \mathbb{E}(XY) - \mu_X \mathbb{E}(Y) - \mu_Y \mathbb{E}(X) + \mu_X \mu_Y \\
&= \mathbb{E}(XY) - \mu_X \mu_Y \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}
$$

Clearly, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ and $\mathrm{Cov}(Y, Y) = \mathrm{Var}(Y)$.

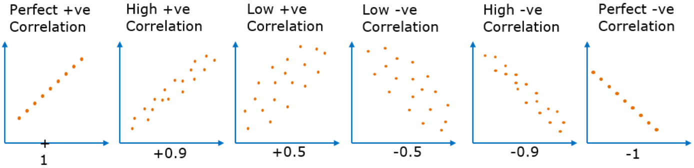
Linearity

$$
\begin{aligned}
& \mathrm{Cov}(aX+b, cY+d) \\
= \; & \mathbb{E}[(aX+b-\mathbb{E}(aX+b))(cY+d-\mathbb{E}(cY+d))] \\
= \; & \mathbb{E}[a(X-\mathbb{E}(X))c(Y-\mathbb{E}(Y))] = ac\mathrm{Cov}(X,Y).
\end{aligned}
$$

Covariance depends on units (meter/foot, kilogram/pound).

$$
\begin{aligned}
& \mathrm{Cov}(X+Y, Z) = \mathbb{E}[(X+Y-\mathbb{E}(X+Y))(Z-\mathbb{E}(Z))] \\
= \; & \mathbb{E}[(X-\mathbb{E}(X)+Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\
= \; & \mathbb{E}[(X-\mathbb{E}(X))(Z-\mathbb{E}(Z))] + \mathbb{E}[(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))] \\
= \; & \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z).
\end{aligned}
$$

Correlation

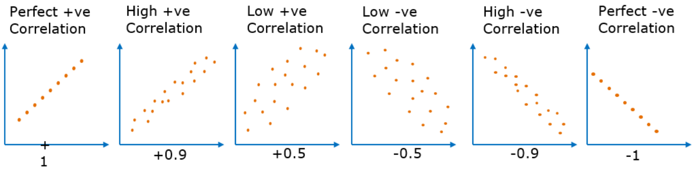


Standardize: $X \to (X - \mu_X)/\sigma_X$, $Y \to (Y - \mu_Y)/\sigma_Y$.

$$\mathbb{E}\left[\frac{X-\mu_X}{\sigma_X}\right] = \frac{\mathbb{E}(X)-\mu_X}{\sigma_X} = 0; \ \mathrm{Var}\left[\frac{X-\mu_X}{\sigma_X}\right] = \frac{\mathrm{Var}(X)}{\sigma_X^2} = 1.$$

$$\mathrm{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{Corr}(X,Y).$$

Correlation



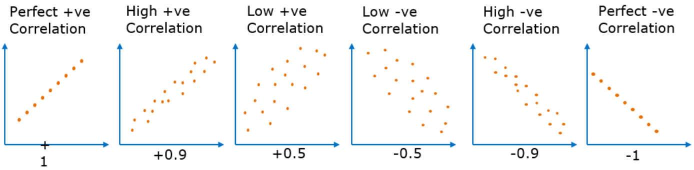
$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

$$\mathrm{Cov}(X,Y) \doteq \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

$$\mathrm{Var}(X) \doteq \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2; \;\; \mathrm{Var}(Y) \doteq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

$$\mathrm{Corr}(X,Y) \doteq \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Correlation



Centralize: $\tilde{X}_i = X_i - \bar{X}$; $\tilde{Y}_i = Y_i - \bar{Y}$.

$$\text{Corr}(X,Y) \doteq \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i}{\sqrt{\sum_{i=1}^n \tilde{X}_i^2}\sqrt{\sum_{i=1}^n \tilde{Y}_i^2}}.$$

Correlation

$$\begin{aligned}\mathrm{Corr}(X,Y) &= \frac{\sum_{i=1}^n \tilde{X}_i\tilde{Y}_i}{\sqrt{\sum_{i=1}^n \tilde{X}_i^2}\sqrt{\sum_{i=1}^n \tilde{Y}_i^2}}\\ &= \frac{\langle \mathbf{X}, \mathbf{Y}\rangle}{\|\mathbf{X}\|\|\mathbf{Y}\|} = \cos\theta.\end{aligned}$$

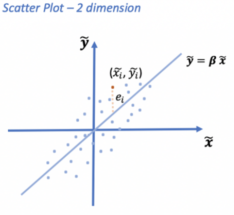
$$\frac{1}{n}\langle \mathbf{X}, \mathbf{Y}\rangle = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{Y}_i = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \doteq \mathrm{Cov}(X,Y).$$

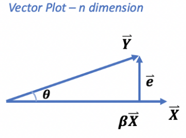
$$\frac{1}{n}\|\mathbf{X}\|^2 = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \doteq \mathrm{Var}(X).$$

$$\frac{1}{n}\|\mathbf{Y}\|^2 = \frac{1}{n}\sum_{i=1}^n \tilde{Y}_i^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 \doteq \mathrm{Var}(Y).$$

Correlation and regression

	$\vec{X}$	$\vec{Y}$	$\vec{e}$
1	$\tilde{x}_1$	$\tilde{y}_1$	e_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	$\tilde{x}_i$	$\tilde{y}_i$	e_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\tilde{x}_n$	$\tilde{y}_n$	e_n





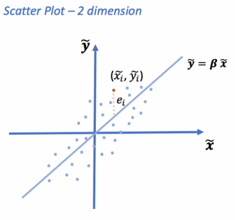
Strength of linear relationship:

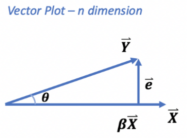
$$\frac{\|\mathbf{e}\|^2}{\|\mathbf{Y}\|^2} = \frac{\sum_i e_i^2}{\sum_i (Y_i - \bar{Y})^2} = \sin^2\theta = 1 - \cos^2\theta = 1 - \rho^2.$$

$$\frac{\|\beta\mathbf{X}\|}{\|\mathbf{Y}\|} = \cos\theta = \rho;\ \beta = \rho\frac{\|\mathbf{Y}\|}{\|\mathbf{X}\|} = \rho\frac{\sigma_Y}{\sigma_X}.$$

Bivariate normal

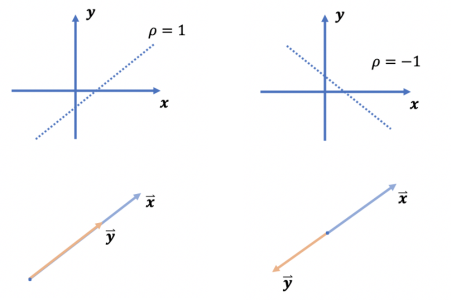
	$\vec{X}$	$\vec{Y}$	$\vec{e}$
1	$\tilde{x}_1$	$\tilde{y}_1$	e_1
⋮	⋮	⋮	⋮
2	$\tilde{x}_i$	$\tilde{y}_i$	e_i
⋮	⋮	⋮	⋮
n	$\tilde{x}_n$	$\tilde{y}_n$	e_n



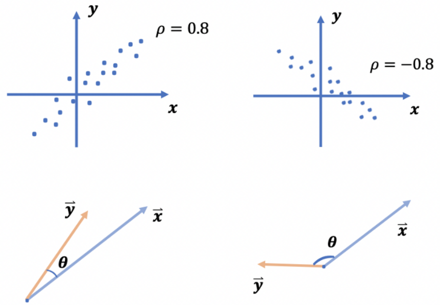


$$X_i \sim \mathrm{N}(0,1),$$
$$Y_i = \rho X_i + \epsilon_i;\ \epsilon_i \sim \mathrm{N}(0, 1-\rho^2),\ i = 1, ..., n.$$
$$\mu_X = \mu_Y = 0,\ \sigma_X = \sigma_Y = 1.$$
$$\frac{\|\mathbf{e}\|^2}{\|\mathbf{Y}\|^2} = 1 - \rho^2.$$
$$\beta = \rho\frac{\|\mathbf{Y}\|}{\|\mathbf{X}\|} = \rho\frac{\sigma_Y}{\sigma_X} = \rho.$$

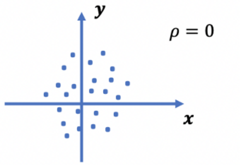
Correlation and regression

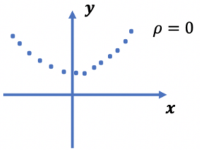


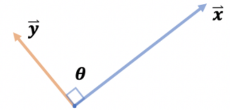
Correlation and regression

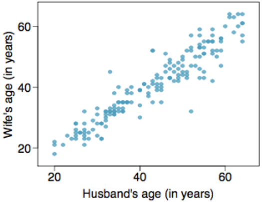


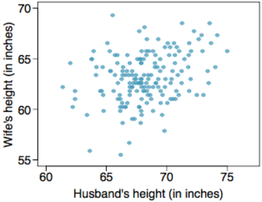
Correlation and regression

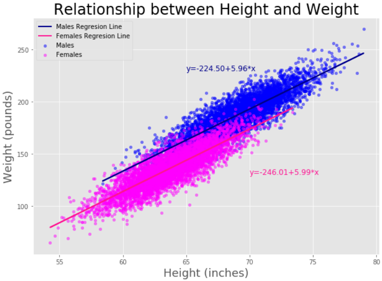




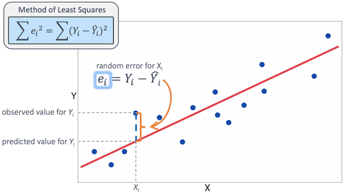


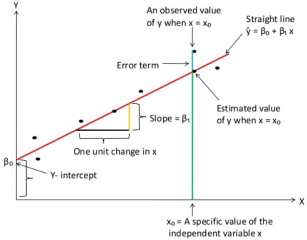

Wife's age (in years)
60
40
20
20
40
60
Husband's age (in years)


Wife's height (in inches)
70
65
60
55
60
65
70
75
Husband's height (in inches)


Relationship between Height and Weight
Males Regresion Line
Females Regresion Line
Males
Females
y=-224.50+5.96*x
y=-246.01+5.99*x
Weight (pounds)
250
200
150
100
55
60
65
70
75
80
Height (inches)

Correlation and regression





Regression line:

$$\hat{Y} - \bar{Y} = \beta_1 (X - \bar{X}).$$

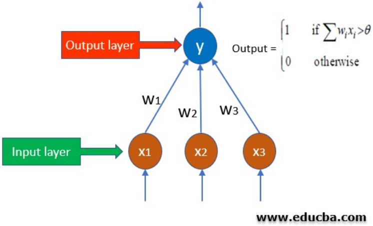
$$\hat{Y} = \beta_1 X + (\bar{Y} - \beta_1 \bar{X}) = \beta_1 X + \beta_0.$$

Multiple regression:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Deep learning: non-linear regression

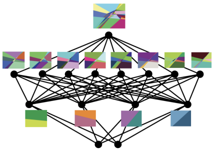
Perceptron

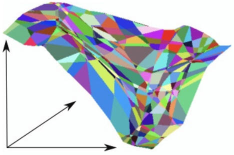


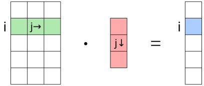
Rectified Linear Unit (ReLU$(a) = \max(0, a)$):

$$y = \max\left(0, \sum_i w_i x_i + b\right).$$

Deep learning: multi-layer perceptron





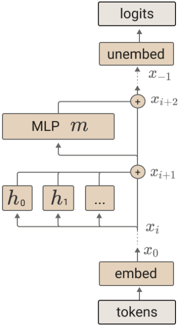


Each node = Linear combination of nodes at layer below $\sum_i w_i x_i$, and then ReLU $\max(0, \sum_i w_i x_i - \theta)$.

$$h_l = \max\left(0, W_l h_{l-1} + b_l\right).$$

h_l: embedding, encoding, representation, thought vector.
W_l: weight matrix. b_l: bias vector.
Piecewise linear mapping from input to output
Weights can be learned from training data .
Learned weights can be used for testing

Deep learning: GPT



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

Token embedding.

$$x_0 = W_E t$$

Embed: word $\rightarrow$ vector

Compute: vectors operated by learned matrices

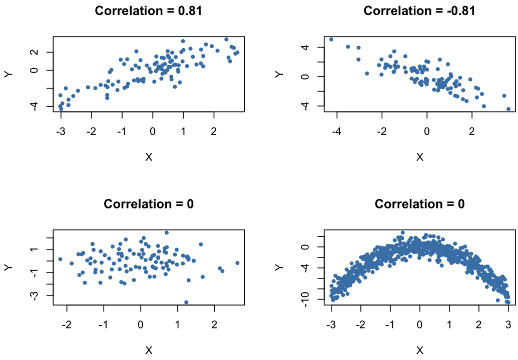
Unembed: vector $\rightarrow$ probabilities for next word

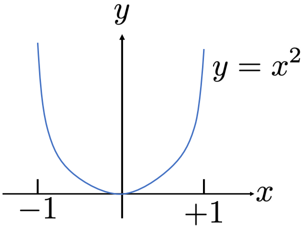
Independence

$P(A \cap B) = P(A)P(B)$.
$p(x, y) = p_X(x)p_Y(y)$; $p(y|x) = p_Y(y)$.
$f(x, y) = f_X(x)f_Y(y)$; $f(y|x) = f_Y(y)$.

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) \\
&= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p_X(x)p_Y(y) \\
&= \sum_x (x - \mu_X)p_X(x) \sum_y (y - \mu_Y)p_Y(y) \\
&= \left(\sum_x x p_X(x) - \mu_X\right)\left(\sum_y y p_Y(y) - \mu_Y\right) = 0.
\end{aligned}
$$

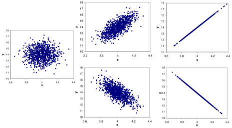
Correlation

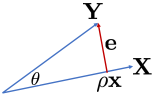




Let X be a uniform distribution over $[-1, 1]$. Let $Y = X^2$.
Then X and Y are not independent.
However, $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$, and $\mathbb{E}(X) = 0$. Thus $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

Bivariate normal





$$X \sim \mathrm{N}(0, 1),$$
$$Y = \rho X + \epsilon;\ \epsilon \sim \mathrm{N}(0, 1 - \rho^2),$$

$$\mathbb{E}(Y) = \mathbb{E}(\rho X + \epsilon) = 0.$$

ϵ and X are independent.

$$\mathrm{Var}(Y) = \mathrm{Var}(\rho X + \epsilon) = \rho^2 \mathrm{Var}(X) + \mathrm{Var}(\epsilon) = 1.$$

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \rho X + \epsilon) = \rho \mathrm{Cov}(X, X) + \mathrm{Cov}(X, \epsilon) = \rho.$$

Variance of sum

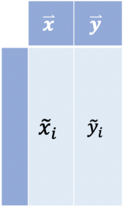
$\mathbb{E}(X+Y) = \sum_x \sum_y (x+y)p(x,y) =$
$\sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[((X+Y) - \mu_{X+Y})^2] \\
&= \mathbb{E}[((X-\mu_X) + (Y-\mu_Y))^2] \\
&= \mathbb{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2] + 2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).
\end{aligned}
$$

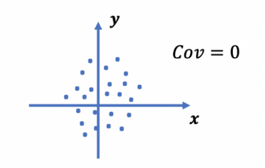
If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$, and

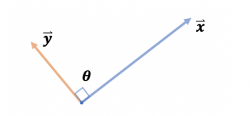
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Variance of sum

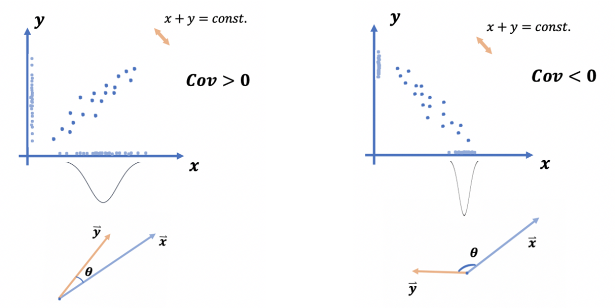


$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i{}^2 = Var(X) = \frac{1}{n}|\vec{x}|^2$$



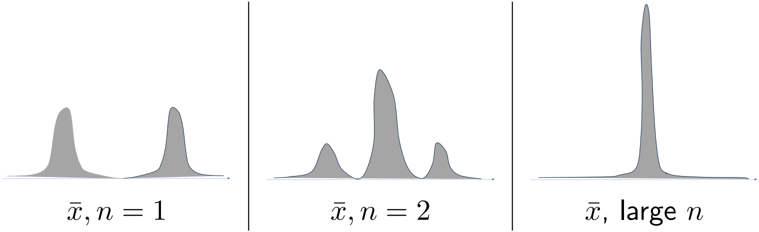


Variance of sum



Average of iid

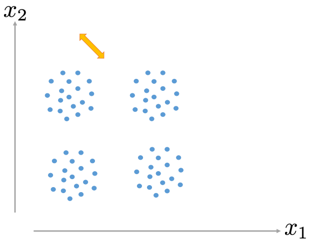
$X_1, X_2, ..., X_n \sim f(x)$ independently.
independent and identically distributed, iid



$x_1 \backslash x_2$	small	large
small	small	medium
large	medium	large

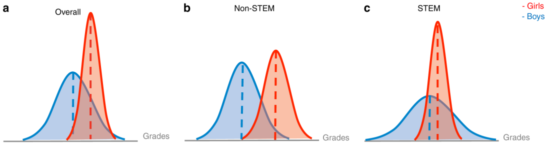
Variance becomes smaller, distribution becomes smoother.

Average of iid



$x_1 \backslash x_2$	small	large
small	small	medium
large	medium	large

Average of iid



- On average girls grades better,
- Girls grades more consistent,
- Fewer top scoring girls

- On average girls grades much better,
- Girls grades similarly variable,
- More top scoring girls

- On average girls grades slightly better,
- Girls grades much more consistent,
- Many fewer top scoring girls

Sum and average of iid

$X_i \sim f(x)$, $i = 1, ..., n$, iid: independent and identically distributed.

$$S = \sum_{i=1}^{n} X_i. \; \bar{X} = \frac{S}{n}.$$

$$\mathbb{E}(X_i) = \mu; \; \mathrm{Var}(X_i) = \sigma^2, \; i = 1, ..., n.$$

$$\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mu.$$

$$\mathrm{Var}(S) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2.$$

$$\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S)}{n} = \mu.$$

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Monte Carlo method

Want to compute $I = \int a(x)dx$. Key equation:

$$I = \int a(x)dx = \int \frac{a(x)}{p(x)}p(x)dx = \mathbb{E}_p\left[\frac{a(X)}{p(X)}\right] = \mathbb{E}_p[h(X)],$$

where $p(x)$ is probability density function, and $h(x) = a(x)/p(x)$.
Sample $X_i \sim p(x)$, $i = 1, ..., n$, iid. Approximate I by

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

$$\mathbb{E}[\hat{I}] = \mathbb{E}[h(X)] = I.$$

$$\mathrm{Var}[\hat{I}] = \mathrm{Var}(h(X))/n.$$

Law of large number

$$\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S)}{n} = \mu.$$

$$\text{Var}(\bar{X}) = \frac{\text{Var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0.$$

$$\bar{X} \to \mu, \text{ in probability.}$$

$$P(|\bar{X} - \mu| < \epsilon) \to 1, \; \forall \epsilon > 0.$$

Average $\to$ expectation.

Law of large number

Special case:

$$X = \sum_{i=1}^{n} Z_i, \; Z_i \sim \text{Bernoulli}(p) \text{ iid.}$$

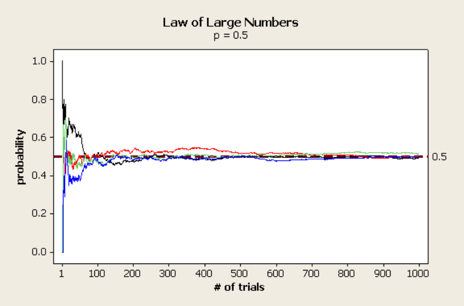
$$\mathbb{E}(X) = np; \; \text{Var}(X) = np(1-p).$$

$$\mathbb{E}(X/n) = p; \; \text{Var}(X/n) = p(1-p)/n \to 0.$$

$$X/n \to p, \text{ in probability.}$$

Frequency $\to$ probability.
X/n is average of Z_i. Probability is expectation of Z_i.

Special case:

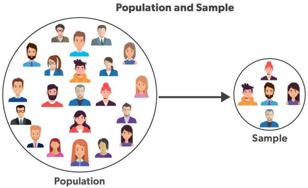


Keep flipping a fair coin, frequency $\to 1/2$.

Intuition: most of 2^n sequences have frequencies close to 1/2.

Survey sampling: N^n reasoning

Ω_1: Population of N people.
Each person $a \in \Omega_1$, $X(a)$ = height.
$\mu = \mathbb{E}(X)$ = population average height.
Repeat random sampling n times independently



$\rightarrow N^n$ equally likely sequences: Ω_n.
For a sequence $\omega \in \Omega_n$, $\bar{X}(\omega)$ = sequence average.
$A = \{\omega : |\bar{X}(\omega) - \mu| \leq .01\}$: representative sequences.
$P(A) = \frac{|A|}{|\Omega_n|} \rightarrow 1$ as $n \rightarrow \infty$.

Cube

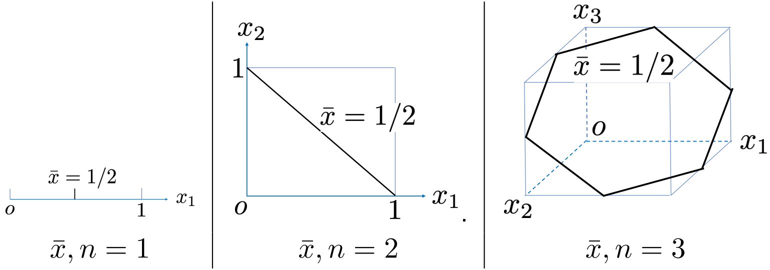
Special case: $X_i \sim \text{Uniform}[0, 1] = \Omega_1$, iid, $i = 1, ..., n$.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \to \mathbb{E}(X_i) = 1/2.$$

$$P(|\bar{X} - 1/2| < .01) \to 1.$$

Intuition: a sequence $(X_1, ..., X_i, ..., X_n)$ is a random point in $\Omega_n = [0, 1]^n$, n-dimensional unit cube.
$A = \{(x_1, ..., x_i, ..., x_n) : |\bar{x} - 1/2| < .01\}$ is the central diagonal piece.
$P(A)$ is the volume of A. $P(A) \to 1$.
The volume of the central diagonal piece is almost the same as the volume of the whole n-dimensional unit cube Ω.
Most of the points in Ω belong to A. Concentration of measure (volume).

Cube


$\bar{x} = 1/2$
o
1
x_1
$\bar{x}, n = 1$
x_2
1
$\bar{x} = 1/2$
o
1
x_1
$\bar{x}, n = 2$
x_3
$\bar{x} = 1/2$
o
x_1
x_2
$\bar{x}, n = 3$

Statistical physics

Most of the points in Ω belong to A. Concentration of measure.
Suppose $(x_1, ..., x_i, ..., x_n)$ describes a physical system, e.g., $n = 10^{23}$ molecules.
It evolves **deterministically** over time, by traversing with Ω.
Ergodic: it traverses every point in Ω with equal number of visits in the long run.
At any **random moment**, $(x_i, ..., x_i, ..., x_n) \sim \text{Unif}(\Omega)$.
Then most likely it will be in A, with fixed statistical properties (e.g., temperature, pressure, magnetism).

Central limit theorem

$X_i \sim f(x)$, $i = 1, ..., n$, iid. $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$.

$$\mathbb{E}(\bar{X} - \mu) = 0;\ \mathrm{Var}(\bar{X} - \mu) = \frac{\sigma^2}{n}.$$

$\bar{X} - \mu \to 0$ in probability.
Magnify: $Y_n = \sqrt{n}(\bar{X} - \mu)$.

$$\mathbb{E}(Y_n) = \mathbb{E}[\sqrt{n}(\bar{X} - \mu)] = 0.$$

$$\mathrm{Var}(Y_n) = \mathrm{Var}[\sqrt{n}(\bar{X} - \mu)] = (\sqrt{n})^2 \frac{\sigma^2}{n} = \sigma^2.$$

Central limit theorem: $Y_n = \sqrt{n}(\bar{X} - \mu) \to \mathrm{N}(0, \sigma^2)$ in distribution.

$$P(Y_n = \sqrt{n}(\bar{X} - \mu) \in [a, b]) \to \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy,$$

regardless of the original distribution of X_i, or whether X_i is discrete or continuous.

Central limit theorem

$X_i \sim f(x)$, $i = 1, ..., n$, iid. $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$.

$$S = \sum_{i=1}^{n} X_i; \; \bar{X} = \frac{S}{n}.$$

$$\mathbb{E}(S) = n\mu, \mathrm{Var}(S) = n\sigma^2; \; \mathbb{E}(\bar{X}) = \mu, \mathrm{Var}(\bar{X}) = \sigma^2/n.$$

Normalization = (random variable - mean)/standard deviation.

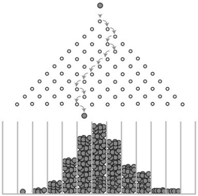
$$Z_n = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{S - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central limit theorem: $Z_n \to \mathrm{N}(0, 1)$ in distribution.

$$P(Z_n \in [a, b]) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz,$$

regardless of the original distribution of X_i, or whether X_i is discrete or continuous.

Coin flipping, random walk, diffusion



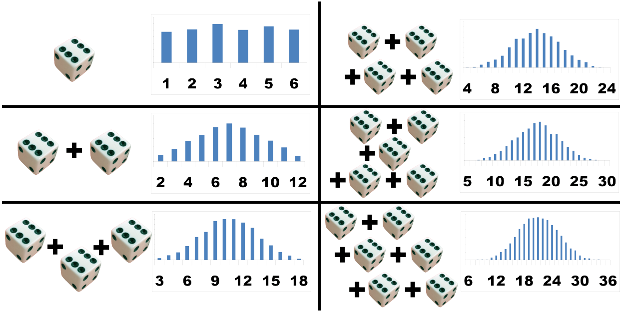
$$X = \sum_{i=1}^{n} \epsilon_i, \;\; \epsilon_i \sim \text{Bernoulli}(1/2) \text{ iid.}$$

$X \sim \text{Binomial}(n, 1/2)$. $\mu = \mathbb{E}(X) = n/2$; $\sigma^2 = \text{Var}(X) = n/4$.

$$P\left(Z = \frac{X - n/2}{\sqrt{n}/2} = z\right) \doteq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \frac{2}{\sqrt{n}} = f(z)\Delta z.$$

In general, ϵ_i can be any discrete or continuous random variable with $\mathbb{E}(\epsilon_i) = 0$.

Die rolling



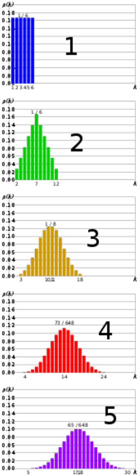
Repeat and plot histogram

$$S = \sum_{i=1}^{n} X_i.$$

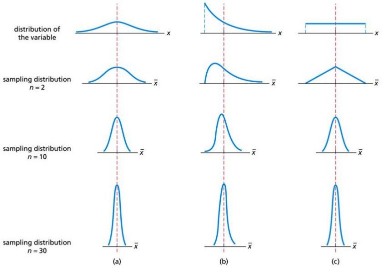
$$\mathbb{E}(X_i) = \mu;\ \mathrm{Var}(X_i) = \sigma^2,\ i = 1, ..., n.$$
$$S \sim \mathrm{N}(n\mu, n\sigma^2).$$

Population of sequences

6^n equally likely sequences $\to$ 6^n equally likely sums $\to$ histogram.



Central limit theorem

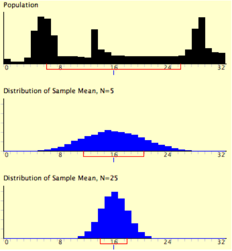


$$S = \sum_{i=1}^{n} X_i.\ \bar{X} = S/n.$$

$$\mathbb{E}(X_i) = \mu;\ \mathrm{Var}(X_i) = \sigma^2,\ i = 1, ..., n.$$

$$S \sim \mathrm{N}(n\mu, n\sigma^2).\ \bar{X} \sim \mathrm{N}(\mu, \sigma^2/n).$$

Central limit theorem



Universal, regardless of the distribution of each X_i.

$$S \sim \mathrm{N}(n\mu, n\sigma^2).\ \bar{X} \sim \mathrm{N}(\mu, \sigma^2/n).$$

$$Z = \frac{S - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

Take home message

As long as you can count (and average)
(1) **Population of equally likely possibilities**
Probability = population proportion
(2) **Large sample of repetitions**
Frequency (fluctuating) $\approx$ probability (fixed)
(3) N^n **reasoning:** hyper-population of sequences (1) $\rightarrow$ (2).
(a) **Probability**: population proportion, long run frequency
(b) **Expectation**: population average, long run average
(c) **Conditional**: sub-population, when something happens
Forward conditional: cause $\rightarrow$ effect
Backward conditional: effect $\rightarrow$ cause
Population migration: cause state $\rightarrow$ effect state
Continuous: discretize, infinitesimal analysis