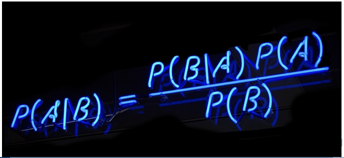
Statistics in Data Analysis

All you ever wanted to know about statistics but never dared to ask

part 2

Pawel Brückman de Renstrom
(pawel.bruckman@ifj.edu.pl)

March 13, 2024



Questions from the previous lecture

1. Using the Kolmogorov axioms, show that:

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

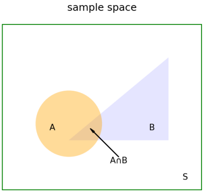
.

2. What is the standard deviation of the sample mean $\bar{x}$, i.e. calculate $Var\langle\bar{x}\rangle \equiv \langle(\bar{x} - \mu)^2\rangle$.
(Hint: On the way, you'll need to prove that $\langle x_i x_j \rangle_{i \neq j} = \mu^2$.)

Solutions to be sent to me before the next lecture

Solutions

Show that: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Using the Kolmogorov axioms, show that:

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

It is enough to note that:

$$A \cup B = A \cup (\bar{A} \cap B) \quad \text{and} \quad B = (A \cap B) \cup (\bar{A} \cap B)$$

and use the 2$^{\text{nd}}$ Kolmogorov's axiom about probability of disjoint subsets twice:

$$P(A \cup B) = P(A \cup (\bar{A} \cap B) = P(A) + P(\bar{A} \cap B)$$

$$P(B) = P((A \cap B) \cup (\bar{A} \cap B)) = P(A \cap B) + P(\bar{A} \cap B)$$

From where we get:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \therefore$$

Solutions

What is the standard deviation of the sample mean $\bar{x}$, i.e. calculate $Var(\bar{x}) \equiv \langle(\bar{x}-\mu)^2\rangle$.

$$
\begin{aligned}
Var(\bar{x}) \equiv \sigma_{\bar{x}}^2 &\equiv \langle(\bar{x}-\mu)^2\rangle \\
&= \langle(\frac{1}{n}\sum_i x_i - \mu)^2\rangle \\
&= \frac{1}{n^2}\sum_i \langle x_i^2\rangle + \frac{1}{n^2}\sum_{i\neq j}\langle x_i x_j\rangle - 2\mu\langle\bar{x}\rangle + \mu^2 \\
&= \frac{1}{n^2} n\langle x^2\rangle + \frac{n(n-1)}{n^2}\langle x_i x_j\rangle_{i\neq j} - 2\mu\langle\bar{x}\rangle + \mu^2 \\
&= \frac{\langle x^2\rangle}{n} + \frac{n-1}{n}\mu^2 - \mu^2 = \frac{\langle x^2\rangle - \mu^2}{n} = \frac{\sigma^2}{n} \qquad \therefore
\end{aligned}
$$

$$
\begin{aligned}
\langle x_i x_j\rangle_{i\neq j} &= \int\int x_i x_j g(x_i, x_j)\mathrm{d}x_i\mathrm{d}x_j = \int\int x_i x_j f(x_i) f(x_j)\mathrm{d}x_i\mathrm{d}x_j \\
&= \left(\int x f(x)\mathrm{d}x\right)^2 = \mu^2 \qquad \therefore \qquad \text{also: } \langle x^2\rangle = \sigma^2 + \mu^2
\end{aligned}
$$

Expectation value for the variance estimators s^2 and S^2

$$E[s^2] = \frac{1}{n-1}\sum_i E[(x_i-\bar{x})^2] = \frac{1}{n-1}\sum_i E[x_i^2 - 2x_i\bar{x} + \bar{x}^2] =$$

$$= \frac{1}{n-1}\sum_i \left(E[x_i^2] - \frac{2}{n} E\left[x_i \sum_j x_j\right] + \frac{1}{n^2} E\left[\sum_k x_k \sum_j x_j\right]\right) =$$

$$= \frac{1}{n-1}\sum_i \left(E[x_i^2] - \frac{2}{n}\sum_j E[x_i x_j] + \frac{1}{n^2}\sum_{k,j} E[x_k x_j]\right) =$$

$$=^* \frac{1}{n-1}\sum_i \left(\mu^2 + \sigma^2 - \frac{2}{n}(\mu^2 + \sigma^2 + (n-1)\mu^2) + \frac{1}{n^2}\left[(n^2-n)\mu^2 + n(\mu^2+\sigma^2)\right]\right) =$$

$$= \frac{1}{n-1}\sum_i \left(0 \times \mu^2 + \frac{n-1}{n}\sigma^2 \right) = \frac{1}{n-1} n \frac{n-1}{n}\sigma^2 = \sigma^2, \qquad \therefore \tag{1}$$

$$E[S^2] = \frac{1}{n}\sum_i E[(x_i-\mu)^2] = \frac{1}{n}\sum_i E[x_i^2 - 2x_i\mu + \mu^2] =^* \frac{1}{n}\sum_i \left(\mu^2+\sigma^2-2\mu^2+\mu^2\right) =$$

$$= \frac{1}{n} n\sigma^2 = \sigma^2, \qquad \therefore$$

* by virtue of identities used on the previous slide.

covariance & correlation

Let $a(\mathbf{x})$ and $b(\mathbf{x})$ be two functions of random variables $\mathbf{x} = (x_1, x_2, ..., x_n)$.

covariance matrix

$$\begin{aligned} V_{ab} = \mathrm{cov}[a,b] = E[(a-\mu_a)(b-\mu_b)] = \\ = E[ab] - E[a\mu_b] - E[\mu_a b] + E[\mu_a\mu_b] = \\ = E[ab] - \mu_a\mu_b - \mu_a\mu_b + \mu_a\mu_b = E[ab] - \mu_a\mu_b = \\ = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} a(\mathbf{x})b(\mathbf{x})f(\mathbf{x})dx_1...dx_n - \mu_a\mu_b \end{aligned} \tag{3}$$

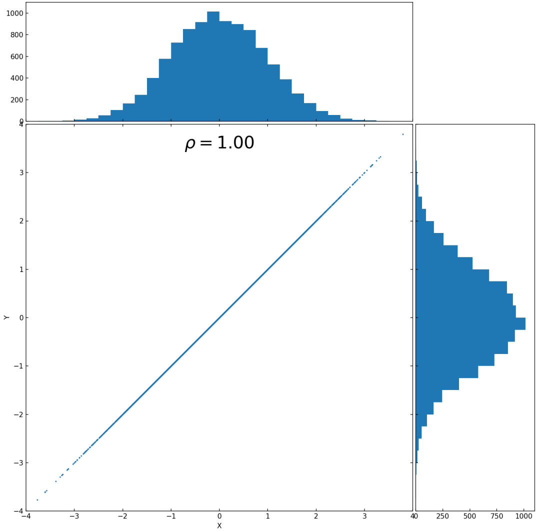
Note: $E\left[E[a(\mathbf{x})]\right] = E[a(\mathbf{x})]$ as $\int_S f(\mathbf{x})d\mathbf{x} \equiv 1$.

variance & correlation coefficient

$$V_{aa} = \mathrm{cov}[a,a] = \sigma_a^2 \qquad \rho_{a,b} = \frac{V_{ab}}{\sigma_a\sigma_b}. \tag{4}$$

Note that $-1 \leq \rho_{a,b} \leq 1$.

covariance & correlat



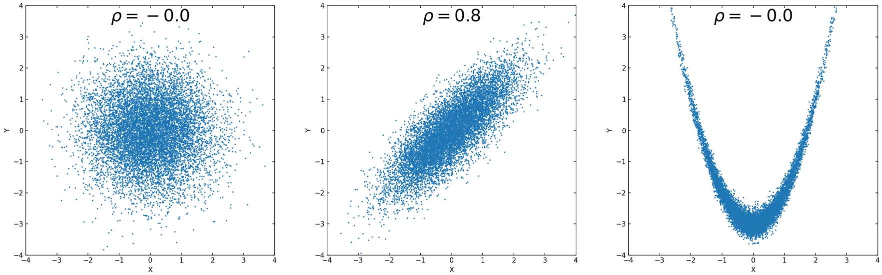
word of caution

For independent variables x and y the joint p.d.f. satisfies $f(x, y) = g(x)h(y)$ and hence:

$$E[xy] = E[x]E[y] = \mu_x \mu_y$$

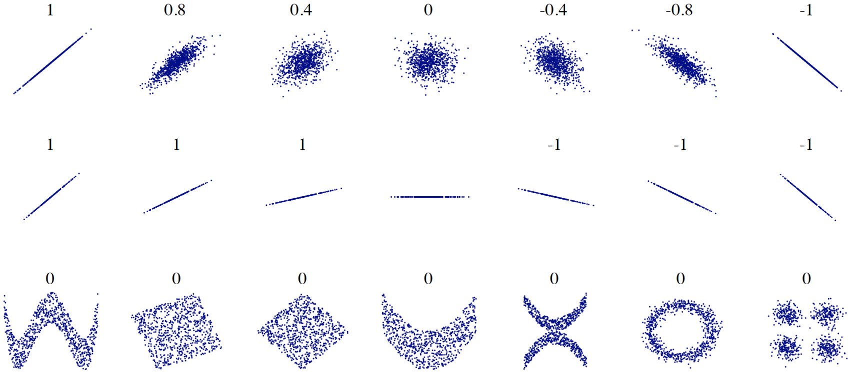
From the definition of covariance we get $V_{x,y} \equiv 0$.

The inverse cannot be inferred, though! I.e. $V_{x,y} = 0$ does not imply independence of the variables!



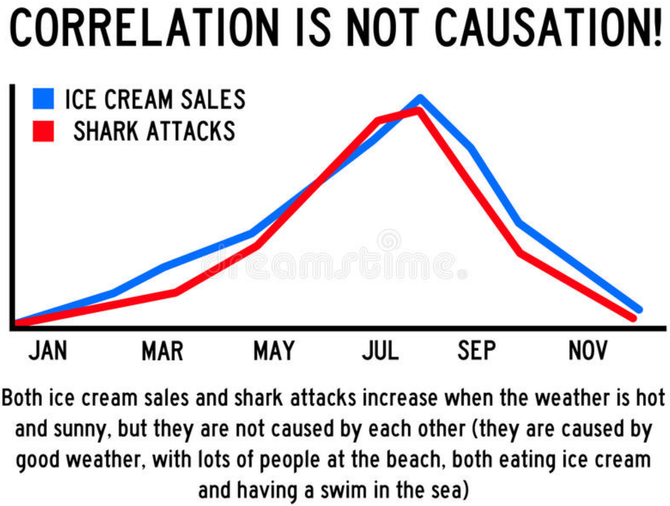
word of caution

correlated and uncorrelated variables (2D), examples



Correlation vs causation

The *hidden variable*



Derived random variable

mean of derived random variable

Let $y(\mathbf{x})$ be a function of n random variables $\mathbf{x} = (x_1, x_2, ..., x_n)$.
We know how to rigorously determine the p.d.f. of y. However, if the exact form of $f(\mathbf{x})$ is unknown and we only know the means and variances, we can approximate these properties for y:

$$y(\mathbf{x}) \approx y(\mu) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\mu} (x_i - \mu_i)$$

mean value

$$E[y(\mathbf{x})] \approx E[y(\mu)] = y(\mu), \tag{5}$$

as $E[x_i - \mu_i] \equiv 0$.

Error propagation

variance of derived random variable

$$E[y^2(\mathbf{x})] \approx y^2(\mu) + 2y(\mu) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\mathbf{x}=\mu} \overbrace{E[x_i - \mu_i]}^{0} + E\left[\left(\sum_{i=1}^{n} \frac{\partial y}{\partial x_i}_{\mathbf{x}=\mu} (x_i - \mu_i)\right)^2\right] =$$

$$= y^2(\mu) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\mathbf{x}=\mu} V_{ij}. \tag{6}$$

(co)variance

$$\sigma_y^2 = E[y^2] - (E[y])^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\mathbf{x}=\mu} V_{ij}. \tag{7}$$

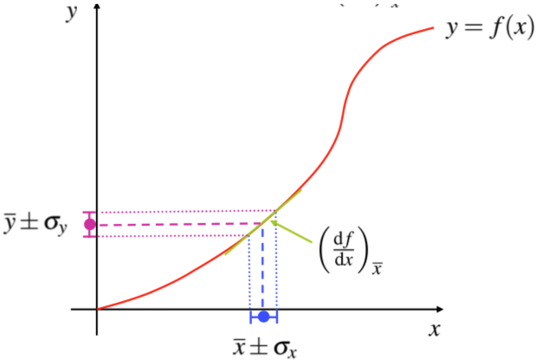
and analogously:

$$U_{kl} = \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right]_{\mathbf{x}=\mu} V_{ij}, \quad \text{in short} \quad U = AVA^T. \tag{8}$$

Error propagation

A simple 1D illustration

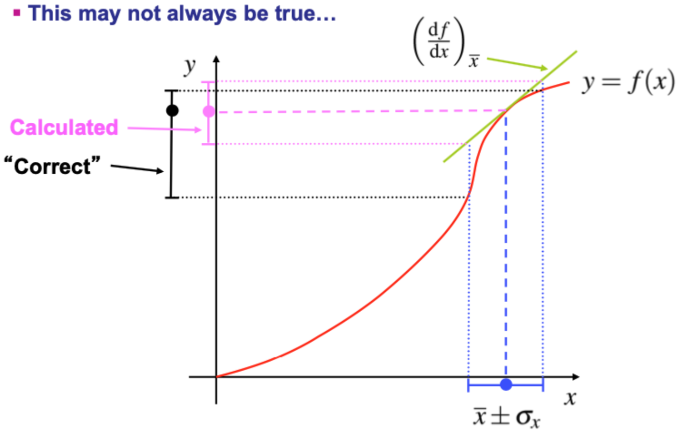
In the simplest case $y = f(x)$, it is easy to see the origin of $\sigma_y = \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)_{\bar{x}} \sigma_x$:



Error propagation

Watch out for traps...

In the previous example validity of the *linear expansion* was assumed, i.e. we considered higher order terms in the Taylor expansion to be negligible.



Error propagation

Watch out for traps...

- We measure the transverse momentum of a track (p_T) from the fitted track curvature which is inversly proportional to the radius of curvature (R) of the track in the solenoidal magnetic field:

$$R = 0.3B(\mathrm{T})p_T(\mathrm{GeV})$$

- We obtain a symmetric (Gaussian) uncertainty on $1/R$. Now we calculate the error on p_T. For simplicity, let us take $p_T = 1/x$, and we know σ_x:

$$\frac{\mathrm{d}p_T}{\mathrm{d}x} = -\frac{1}{x^2} = -p_T^2 \quad \mathrm{h}ence \quad \sigma_{p_T} = p_T^2 \sigma_x$$

- Take the measured x to be 0.01 ± 0.005 GeV^{-1}.
- We obtain: $p_T = 100 \pm 50$ GeV.
- The real variation corresponding to the uncertainty on x is: $p_T = 100 + 100 - 33$ GeV.
- The two results are very different (the latter being correct).

De-correlation

unitary rotation in the $\mathbf{x}_n$ space

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $V_{ij} = \mathrm{cov}[x_i, x_j]$ their (symmetric) covariance matrix. One can always find a linear transformation of $\mathbf{x}$ that diagonalizes the covariance:

$$y_i = \sum_{j=1}^{n} A_{ij} x_j, \qquad \mathrm{cov}[y_i, y_j] = \mathrm{cov}\left[\sum_{k=1}^{n} A_{ik} x_k \sum_{l=1}^{n} A_{jl} x_l\right] = AVA^T = U, \tag{9}$$

which is a special case of error propagation (exact, thanks to linear nature of the transformation!). The problem boils down to diagonalising the the matrix V, i.e. finding eigenvectors $\mathbf{r^i}$ and their corresponding eigenvalues λ_i satysfying the eigenequation:

$$V\mathbf{r^i} = \lambda_i \mathbf{r^i} \tag{10}$$

(note that: $\lambda_i {\mathbf{r^i}}^T \mathbf{r^j} = {\mathbf{r^i}}^T V \mathbf{r^j} = \lambda_j {\mathbf{r^i}}^T \mathbf{r^j} \overset{\lambda_i \neq \lambda_j}{\Longrightarrow} {\mathbf{r^i}}^T \mathbf{r^j} = \delta_{ij}$ if $\mathbf{r^i}$ normalised).

$$A \equiv \begin{pmatrix} \boxed{\quad \mathbf{r^1} \quad} \\ \boxed{\quad \mathbf{r^2} \quad} \\ \cdots\cdots \\ \cdots\cdots \\ \boxed{\quad \mathbf{r^n} \quad} \end{pmatrix} \qquad U = AVA^T = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & 0 & \\ & & \ldots & \\ 0 & & \ldots & \\ & & & \lambda_n \end{pmatrix} \tag{11}$$

De-correlation

example: rotation in the 2D space

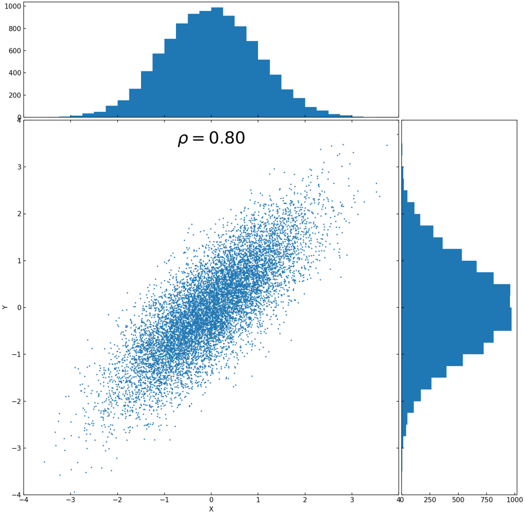
$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{12}$$

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{13}$$

$$\theta = \frac{1}{2}\arctan\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right) \tag{14}$$

Verify this result!

NOTE: Decorrelation will not necessarily make the variables independent!

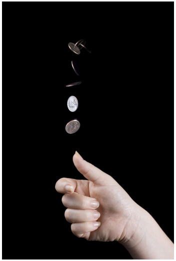


Tossing a coin

Binomial distribution

- Tossing a coin can yield two distinct results (usually with equal probability).
- What is the probability of scoring n heads in N trials?

$$P(n(N,p)) = \overbrace{\frac{N!}{n!(N-n)!}}^{\text{number of sequences}} \underbrace{p^n(1-p)^{N-n}}_{P \text{ of a particular sequence}} . \tag{15}$$



- The expectation value:

$$E[n(N,p)] = \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} = Np, \tag{16}$$

which agrees with our intuition, e.g. for a fair coin ($p = 0.5$) we expect heads and tail in 50/50 proportion.
Think of N independent trials, each with expectation value $E[1(1,p)] = p$.

Binomial distribution

Is it a proper p.d.f., i.e. normalised?

1. To start with, recall the binomial theorem:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k}^{*} a^k b^{n-k} \tag{17}$$

2. Now we can use the above to show that the binomial distribution is normalised:

$$\sum_{n=0}^{N} P(n(N,p)) = \sum_{n=0}^{N} \binom{N}{n} p^n (1-p)^{N-n} = (p+q)^N = 1^N = 1 \quad \therefore \tag{18}$$

$$* \qquad \frac{n!}{k!(n-k)!} \equiv \binom{n}{k}$$

Binomial distribution

$\langle n \rangle$ - rigorous calculation

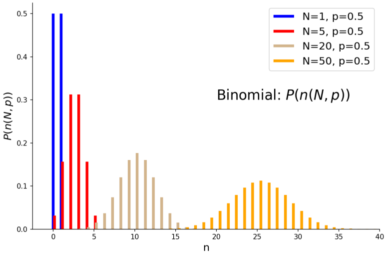
$$
\begin{aligned}
\langle n \rangle &= \sum_{n=0}^{N} nP(n(N,p)) = \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\
&= Np \sum_{n=1}^{N} \frac{(N-1)!}{(n-1)!(N-n)!} p^{n-1} (1-p)^{N-n} \\
&= Np \sum_{n=0}^{N-1} \frac{(N-1)!}{n!(N-1-n)!} p^n (1-p)^{N-1-n} \\
&= Np \underbrace{\sum_{n=0}^{N-1} P(n(N-1,p))}_{\text{normalised}} = Np \qquad \therefore
\end{aligned} \tag{19}
$$

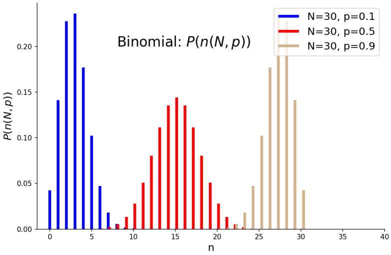
Binomial distribution

- The variance is:

$$Var[n(N,p)] = E[n^2(N,p)] - (E[n(N,p)])^2 = Np(1-p), \qquad (20)$$

which for a fair coin yields $1/4$ of the number of trials N.
This can be rigorously calculated, but can be thought of in terms of error propagation: $Var[1(1,p)] = p(1-p)$.





Binomial distribution

$Var[n(N,p)]$ - rigorous calculation

$$Var[n(N,p)] = \langle (n - \langle n \rangle)^2 \rangle = \langle n^2 \rangle - \langle n \rangle^2 \tag{21}$$

$$\begin{aligned}
\langle n^2 \rangle &= \sum_{n=0}^{N} n^2 P(n(N,p)) = \sum_{n=0}^{N} n^2 \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\
&= Np \sum_{n=1}^{N} n \frac{(N-1)!}{(n-1)!(N-n)!} p^{n-1} (1-p)^{N-n} \\
&= Np \sum_{n=0}^{N-1} (n+1) \frac{(N-1)!}{n!(N-1-n)!} p^n (1-p)^{N-1-n} \\
&= Np \left[\sum_{n=0}^{N-1} n \frac{(N-1)!}{n!(N-1-n)!} p^n (1-p)^{N-1-n} + \sum_{n=0}^{N-1} \frac{(N-1)!}{n!(N-1-n)!} p^n (1-p)^{N-1-n} \right] \\
&= Np \left[(N-1)p + 1 \right] = Np(Np - p + 1)
\end{aligned} \tag{22}$$

$$Var[n(N,p)] = \langle n^2 \rangle - \langle n \rangle^2 = Np(Np - p + 1) - (Np)^2 = Np(1-p) \quad \therefore \tag{23}$$

Binomial distribution

Simple example

Suppose you are assessing efficiency of a certain process (vaxine effectiveness, event selection, what have you...) and you observe n out of N passing the test. What is the efficiency and its uncertainty?
This is a binomial process (fixed number of trials).
The best estimate of the efficiency is:

$$\varepsilon = \frac{n}{N} \qquad (\langle\varepsilon\rangle = \frac{\langle n\rangle}{N} = \frac{Np}{N} = p)$$

How about the straightforward estimation of the variance:

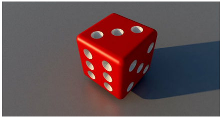
$$\sigma^2 = \frac{\varepsilon(1-\varepsilon)}{N}$$

$$\langle\sigma^2\rangle = \frac{\langle n\rangle}{N^2} - \frac{\langle n^2\rangle}{N^3} = \frac{Np - p(Np - p + 1)}{N^2} = \frac{N+1}{N^2}p(1-p) = \frac{N+1}{N}Var(\varepsilon)$$

Multinomial distribution

generalization of binomial

- Let us extend the process to $m > 2$ outcomes, e.g. rolling a dice.
- The only requirement is to have $\sum_{i=1}^{m} p_i = 1$.



- Probability distribution of a given sequence is given by:

$$P(n_1 ... n_m (N, p_1 ... p_m)) = \frac{N!}{n_1! ... n_m!} p_1^{n_1} ... p_m^{n_m}. \tag{24}$$

Can you derive the above?

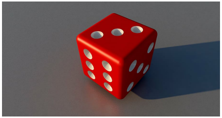
- One can calculate covariance from the joint probability distribution to get:

$$V_{ij} = E[(n_i - E[n_i])(n_j - E[n_j])] = -N p_i p_j \tag{25}$$

Note that for binomial $\rho_{1,2} = \frac{-Np(1-p)}{\sqrt{Np(1-p)}\sqrt{N(1-p)p}} = -1$ (dice: $\rho_{k,l} = ?$)

Multinomial distribution

generalization of binomial



- Let us extend the process to $m > 2$ outcomes, e.g. rolling a dice.
- The only requirement is to have $\sum_{i=1}^{m} p_i = 1$.
- Probability distribution of a given sequence is given by:

$$P(n_1 ... n_m(N, p_1 ... p_m)) = \frac{N!}{n_1! ... n_m!} p_1^{n_1} ... p_m^{n_m}. \tag{26}$$

Can you derive the above?

- One can calculate covariance from the joint probability distribution to get:

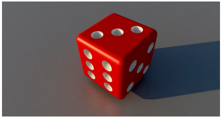
$$V_{ij} = E[(n_i - E[n_i])(n_j - E[n_j])] = -Np_ip_j \tag{27}$$

Note that for binomial $\rho_{1,2} = \frac{-Np(1-p)}{\sqrt{Np(1-p)}\sqrt{N(1-p)p}} = -1$ (dice: $\rho_{k,l} = -0.2$)

Multinomial distribution

generalization of binomial

- Let us extend the process to $m > 2$ outcomes, e.g. rolling a dice.
- The only requirement is to have $\sum_{i=1}^{m} p_i = 1$.



- Probability distribution of a single outcome is simply:

$$P(n_i(N, p_i)) = \frac{N!}{n_i!(N - n_i)!} p_i^{n_i} (1 - p_i)^{N - n_i}, \tag{28}$$

yielding $E[n_i] = Np_i$ and $V[n_i] = Np_i(1 - p_i)$.

Counting experiment

Do counting of a random process (e.g. number of cars passing by the IFJ main entrance in 10'). We want to know the probability distribution to find a certain number of occurences.

- A binomial limit when
  $N \to \infty, \quad p = \varepsilon \to 0, \quad N\varepsilon = \mu = \text{const.}$
- $$P(n) = \begin{pmatrix} N \\ n \end{pmatrix} \varepsilon^n (1-\varepsilon)^{N-n}$$

$$P(n;\mu) = \frac{N!}{(N-n)!n!}\left(\frac{\mu}{N}\right)^n \left(1-\frac{\mu}{N}\right)^{N-n} = \frac{\mu^n}{n!}\frac{N!}{(N-n)!}\frac{(N-\mu)^{N-n}}{N^N}$$

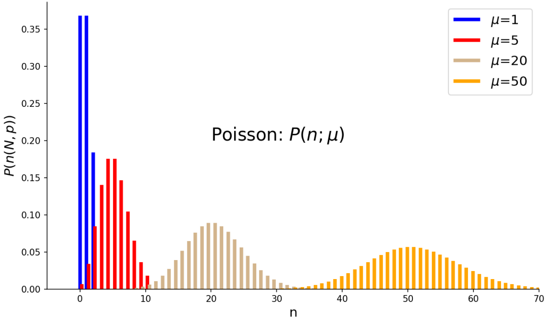
$$\stackrel{N\to\infty}{=} \frac{\mu^n}{n!}\left(\frac{N-\mu}{N}\right)^N = \frac{\mu^n}{n!}\left(1-\frac{\mu}{N}\right)^N \stackrel{*}{=} \frac{\mu^n}{n!}e^{-\mu} \qquad \therefore \tag{29}$$

- Counting random process is described by the **Poisson** distribution:

$$\mathbf{P(n;\mu) = \frac{\mu^n}{n!}e^{-\mu}} \tag{30}$$

* $\left|\ln\left[\left(1-\frac{\lambda}{x}\right)^x = e^{-\lambda}\right] = -x\ln\left(1-\frac{\lambda}{x}\right) \simeq \frac{\lambda}{1-\frac{\lambda}{x}} \stackrel{x\to\infty}{\to} \lambda \Rightarrow \lim_{x\to\infty}\left(1-\frac{\lambda}{x}\right)^x = e^{-\lambda}\right|$

Poisson distribution



For large values of μ Poisson distribution asymptotically tends to a Gaussian* $G(\mu, \sigma^2 = \mu)$

* See the Central Limit Theorem later in this lecture.

$$E[n] = \sum_{n=0}^{\infty} n \frac{\mu^n}{n!} e^{-\mu} = \sum_{n=1}^{\infty} n \frac{\mu^n}{n!} e^{-\mu} = \mu \sum_{n=1}^{\infty} \frac{\mu^{n-1}}{(n-1)!} e^{-\mu} = \mu \sum_{k=0}^{\infty} \frac{\mu^k}{(k)!} e^{-\mu} = \mu \quad (31)$$

$$V[n] = E[n^2] - (E[n])^2 = E[n(n-1) + n] - (E[n])^2 = \mu^2 + \mu - \mu^2 = \mu \quad (32)$$

Hence, the well known $\sigma(N) = \sqrt{N}$ for event counting.

Raindrops

Uniform distribution

- Some processes have uniform probability over a limited range of parameter (raindrops on a window sill). Usually these are selected fiducial region of a wider distributed random process.
- Characterised by a continuous uniform p.d.f.
- Must have finite range in order to allow normalisation.

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta-\alpha} \text{ for } \alpha < x < \beta \\ 0 \text{ otherwise} \end{cases} \tag{33}$$

- Mean and the variance are easily obtained:

$$E[x]=\int_\alpha^\beta \frac{x}{\beta-\alpha}dx=\frac{\mathbf{1}}{\mathbf{2}}(\alpha+\beta), \qquad V[x]=\int_\alpha^\beta [x-\frac{1}{2}(\alpha+\beta)]^2\frac{1}{\beta-\alpha}dx=\frac{\mathbf{1}}{\mathbf{12}}(\beta-\alpha)^{\mathbf{2}}. \tag{34}$$

- $f(x;0,1)$ (or simply $[0,1]$) is commonly used in statistics, notably for base random number generators.
- For any continuous p.d.f. $f(x)$, $y=F(x)$ is distributed according to $[0,1]$. (?) Hence, p.d.f. of $x=F^{-1}(y)$ will be $f(x)$ if y has a uniform distribution $[0,1]$.

Questions

1. Suppose two independent measurements of the same quantity gave the following results:

$$x_1 \pm \sigma_1 \quad \text{and} \quad x_2 \pm \sigma_2$$

Take the weighted mean to be $\bar{x} = wx_1 + (1-w)x_2$. Find the w which minimizes the error on the mean, hence provide expressions for the weighted mean $\bar{x}$ and its variance $\sigma_{\bar{x}}^2$.

Solutions to be sent to me before the next lecture

Thank you

Back-up