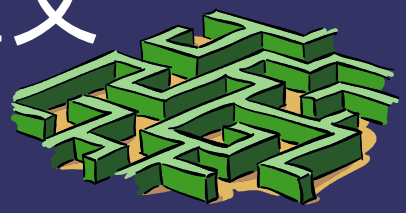
誤差と統計

画像処理を行う各段階や較正において、
真の物理量からのずれが混入する。
このズレをどう定式化し、
どう小さくしていくかについて、
具体例を交え解説する

・・・はずでしたが
誤差が大きくなりました。

八木雅文



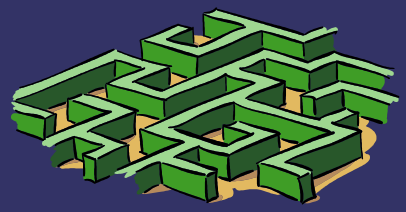
2007/10/09 微改訂版

誤差とは？

真の値からの測定値のズレ

「真の値」とは何かが問題。
測定値から、どのような値を求めたい
と考えているのか？

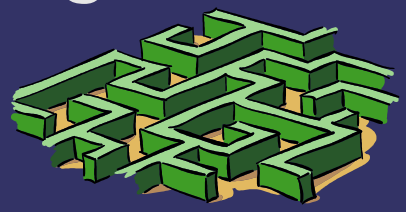
ちなみに測定値や推定値から
どの範囲に真の値があるかの
指標は「不確かさ」



系統誤差とランダム誤差

系統誤差:
　何回やっても同じ量ずれる誤差
ランダム誤差:
　毎回再現性のない誤差

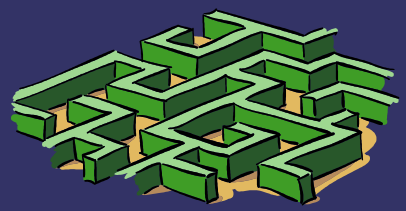
機器の目盛りが間違っていて
毎回同じ量ずれる⇒系統誤差
機器が不安定で結果に再現性がない
⇒ランダム誤差



問題設定

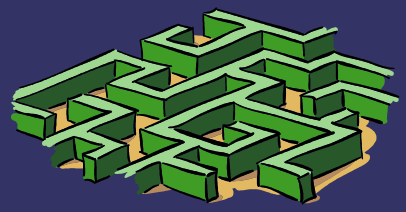
データがあった場合、
そこから欲しい情報を最も精度良く
取り出すにはどうすればよいか？

例えば光赤外の天文データの場合、
天体の光、バックグラウンドの光、
それに測定の際の雑音が乗って
観測される。ここから天体の光を
取り出すのが整約・解析。



基本方針

まず生の観測量にはどのような誤差が乗っているのかをモデル化する。
そのモデルから物理量を求める整約、解析の各段階で、誤差がどのように変化するかを追いかける。

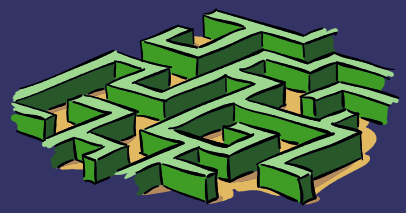


光の強さ

天球上のある範囲から、
ある波長範囲の光が
どれだけの強さで来ているか、つまり、

ある一定時間にあるエネルギー範囲
（＝波長範囲）の光子が観測者側の
単位面積をどれだけの数通過するか

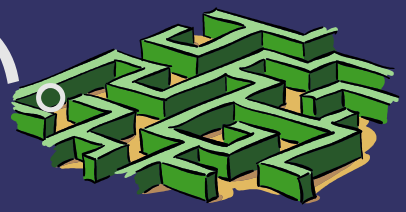
単位は光赤外分野では、
例えば[erg/sec/cm^2]
が良く使われる。



真の値？

時間変化しないものは、
真の値が何かを言うのは簡単。
例えば真空中の光速度。

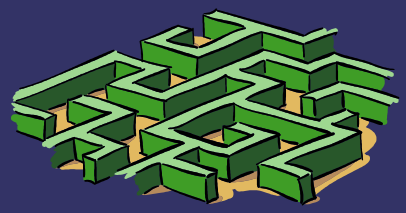
ところが、光はミクロな確率過程を経て
天体から出てくるため、
ある１秒間に天体から出てくる光の量は
別の１秒と同じではない。つまり
観測値は本質的に一定ではない。



光量の誤差とは？

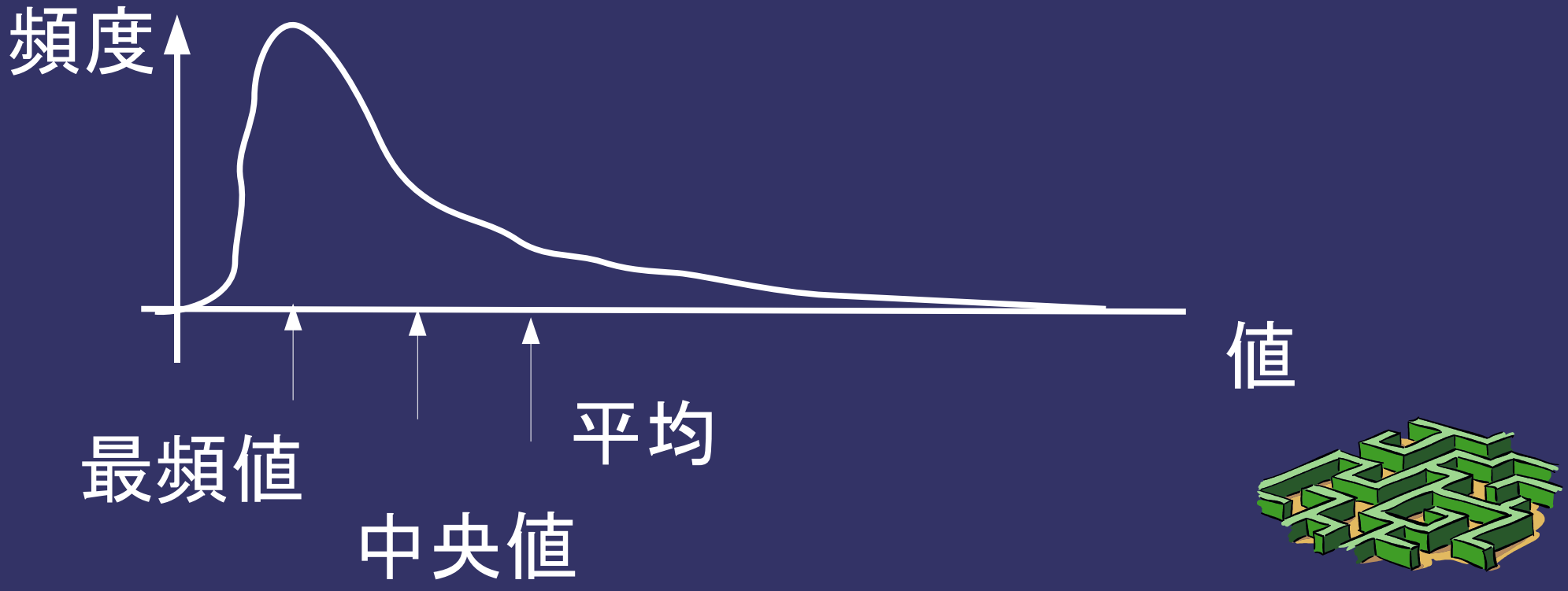
その上で、求めたい「真の値」とは、
「長時間平均」
これはある短時間に観測される光量の
「分布の代表値」

測光の場合の誤差とは、
実際に測定あるいは推定した値が
どれだけ精度良くその分布の代表値を
求められたかという指標。



分布と代表値

分布の代表値として良く使われるのは、
平均、中央値（メディアン）、
最頻値（モード）など。

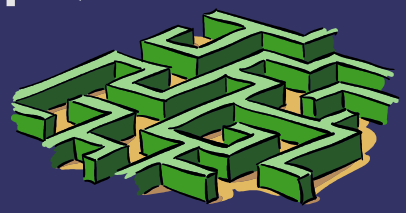


確率分布関数

観測量X：観測ごとに変わる量

があったとして、X＝X0である確率を
p(X0)と書ける場合、関数p(X)が
確率分布関数。

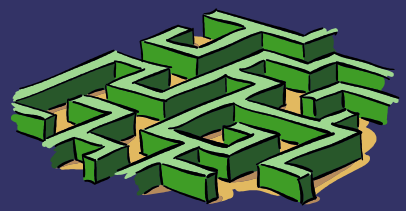
Xが連続量の時、
X0−dX/2<X<X0+dX/2 に入る確率が
p(X0)dX になるようなp(X)が
確率密度分布関数。



サイコロの確率分布

例えば6面サイコロ3個を振って
出た目の和を観測量Xとした時の
確率分布関数は以下。

X	p(X)	X	p(X)	X	p(X)
3	1/216	9	25/216	15	10/216
4	3/216	10	27/216	16	6/216
5	6/216	11	27/216	17	3/216
6	10/216	12	25/216	18	1/216
7	15/216	13	21/216		
8	21/216	14	15/216		



平均

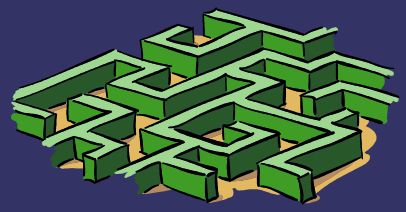
平均(mean)あるいは期待値は、
以下で定義される量。

$$E(X)=mean=\int_{-\infty}^{\infty} X\ p(X)dX$$

もっともタチのいい統計量。
ただし、外れ値に抜群に弱い。後述。

また平均は線形性を満たす

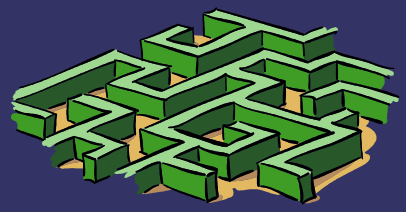
$$E(aX+bY+c)=aE(X)+bE(Y)+c$$



大数の法則

何回も観測して観測量Xの平均を
求めると、その値は確率分布
p(X)の平均に近づいていく。

※この性質はメディアンやモード
　にはない。

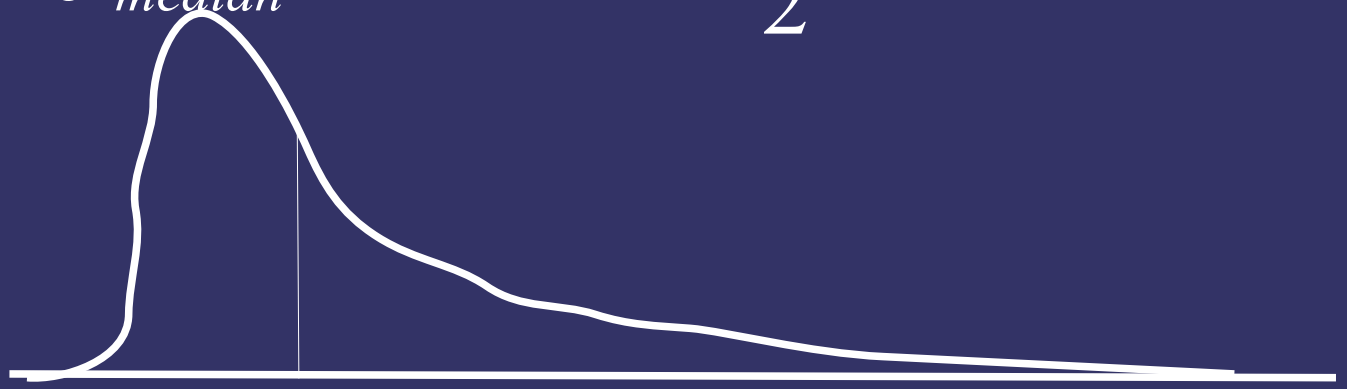


中央値（メディアン）

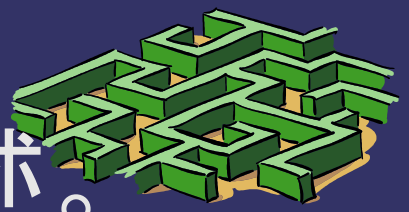
中央値(median)は、文字通り分布の
中央の値で、以下のように定義できる。

$$\int_{-\infty}^{median} p(X)\,dX \geq \frac{1}{2}$$

$$\int_{median}^{\infty} p(X)\,dX \geq \frac{1}{2}$$



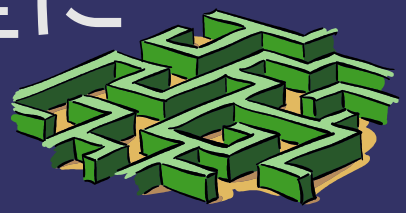
外れ値に強いが問題もある。後述。



誤差の評価

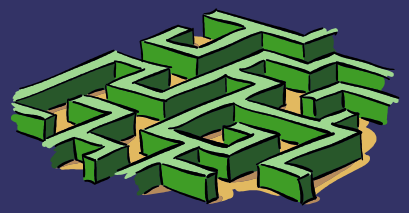
一回の測定での誤差は、
真の値がわからないと求まらない。
測定値の「分布」が何らかの方法で
わかっていれば「誤差の典型値」を
計算できる。

例えば良く使われるのが標準偏差
（誤差の二乗平均）σなどと書かれる。
※但しσは正規分布の標準偏差に
　相当する別の量の場合もある。



誤差の推定

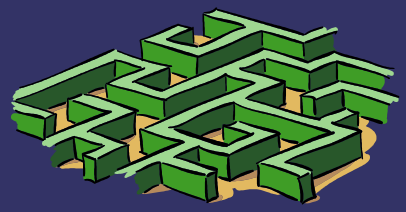
- 観測値の母分布がわかる場合
  ⇒　一回の観測値が典型値から
  どれだけずれるか（誤差）の
  分布も推定できて、典型値も
  求められる。
- 観測値の母分布の関数形はわかる
  あるいは仮定できるが、パラメタが
  分からない場合
  ⇒　母分布のパラメタを推定



誤差と不確かさ

なお、測定値が真の値からどれだけずれているか（誤差）と、測定値からどの範囲に真の値が収まっているか（不確かさ）は、別概念。

なのだが、以下ではこの辺りが曖昧に扱われている。



分散、標準偏差

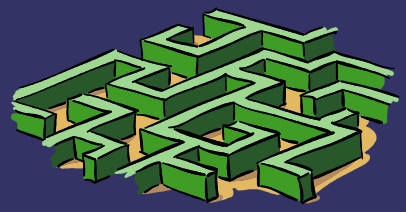
分散は期待値E(X)を使って

$$V(X)=\int_{-\infty}^{\infty}(X-E(X))^2\,p(X)\,dX$$

と書ける。
分散の平方根が標準偏差である。

分散には以下の性質がある。

$$V(aX+b)=a^2V(X)+b$$

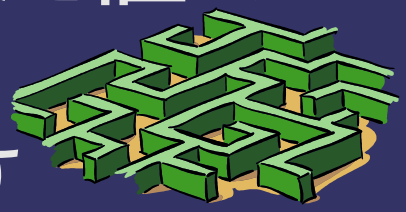


MAD

MAD(Median of Absolute Deviation)は、標準偏差に相当する統計量。
あまり広く使われてないが、
実は便利な量。

MAD=Median(|X−median|)

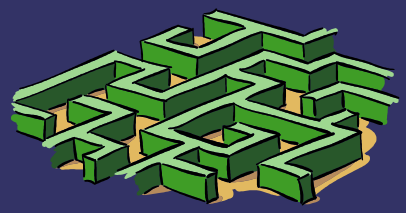
各データ点からメディアンを引いた値の絶対値のメディアン。
正規分布の場合 MAD=0.6745 σ



サイコロ3つの場合

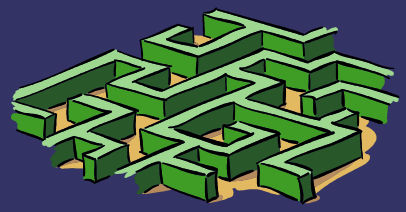
分布の典型値は10.5。
分散はゴリゴリと計算すると 8.75
標準偏差は 2.96 程度。
つまりσ=2.96。
これが理論から求まる誤差の典型値で
これを略して「誤差」と呼ぶことも多い。

この場合3σの範囲内と言えば、
10.5±8.88 である。



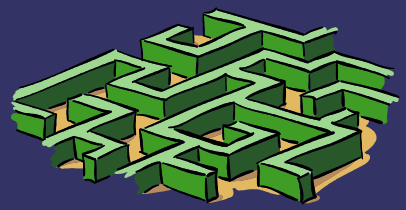
サイコロを振ってみた

さて、実際サイコロ3つを振ってみた。
　14,11,14,13,13,13
この6回の試行（観測）それぞれを
X1,X2...X6 とすると、
本来の平均10.5との差
 +3.5,+0.5,+3.5,+2.5,+2.5,+2.5
が、本来の意味での誤差。
14,11,14,13,13,13　の平均は13、
標準偏差は 1。これを標本平均、
標本標準偏差と呼ぶ。



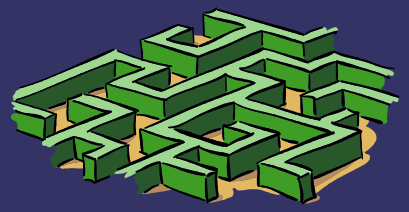
まとめ

- 誤差とは1回の観測量と「真の値」の差が本来の意味
- しかし通常使われるのは「誤差の典型値」の意味であり、この典型値として標準偏差が良く用いられる。
- 測光の場合「真の値」とは分布の典型値である。



誤差の推定

以下では
1）そもそも誤差はどうなるはずか
　から考え、誤差の伝播を使って
　観測量の誤差を推定する
　（限界等級の計算）
2）観測量の統計を考えた上で、
　解析での誤差の広がりを推定する
　（画像の並行移動）
を例として挙げる。



具体例：限界等級

撮像データの測光の限界等級とは、
天体の測光値（signal)が、測光誤差
(noise)の何倍かという比、S/N が
ある一定値、例えば、3以上の天体が
何等か、という値である。

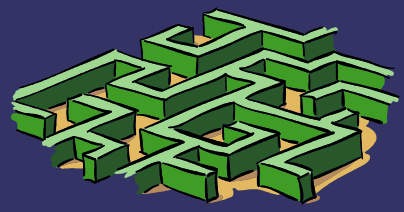
例えば28等の天体が S/N=5 の時、
S/N=5 での限界等級が28等、という。

等級誤差とS/N

等級のランダム誤差Δmと、
S/N との関係は、

$$\Delta m = \frac{1.086}{S/N}$$

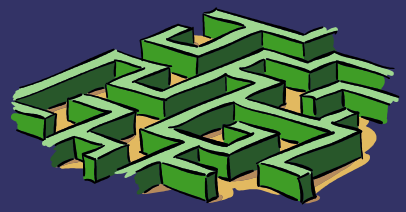
例えば 0.1 等以下の精度で
測光をしたいのであれば、
S/N>10.86 が必要である。



限界等級

この場合の限界等級は「測光精度」の意味での限界等級であって、検出限界の意味での限界等級ではない。

つまり測光精度の意味での限界等級が同じであっても、天体検出の割合が同じであるとは全く限らない。
全然別の話。

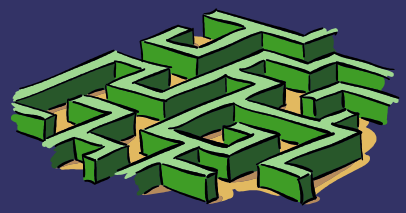


という話をしたら

検出の限界はどう推定するのか？
と質問されたのですが、これを理論から解析的に求めるのは極めて困難。

正方形がN個繋がった図形の数を数えるとかいう数学の未解決問題に関連してたりします。

■
■■■　　こんなペントミノとか
■

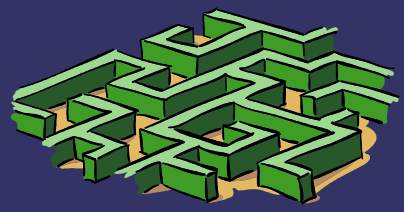


S／NのS

露出時間をt、
システムの透過率をf、
天体の光度をLとすると、
検出器で検出される信号は形式的に

$S=f \cdot t \cdot L$

（単位は光電荷数）と書ける。

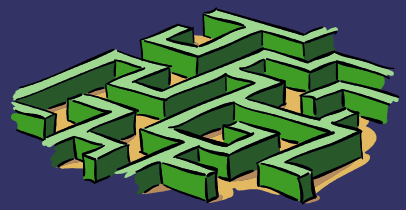


S／NのN

測光での大きな誤差源の一つは
機器のノイズ（読み出し雑音）。
読み出し雑音は露出時間によらず
一定であると考えてよい、
1ピクセルあたりのこの誤差を

$N1=r$

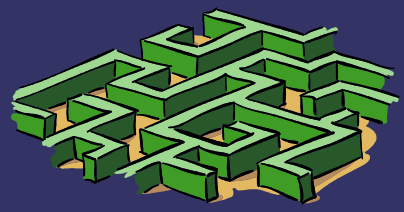
と書くことにする。
単位は電荷数に換算しておく。



さらっと書きましたが

N1=r
と書いたこの意味は、
ある画素での観測量をXとし、
読み出し雑音を δXとした場合に、
δXの標準偏差がrである
という意味。式で書くとこう。

$$E(\delta X)=0$$
$$V(\delta X)=E((\delta X)^2)=N1^2=r^2$$

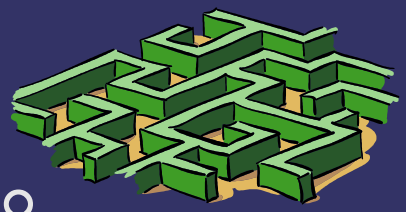


S／NのN

もう一つは
測定される電荷の分布の広がり
（ポアソン雑音）
素子に溜まった総電荷数を n、
ピクセル数をAとすると、

$$N2=\sqrt{n}=\sqrt{S+(f\cdot t\cdot A\cdot s_{SKY})+(t\cdot A\cdot s_{dark})}$$

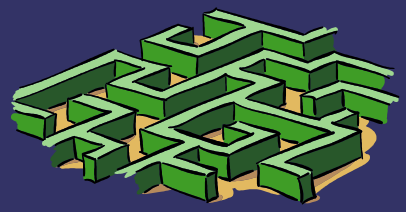
となる。
このN1,N2の合成が誤差Nとなる。



何故N2=√n

天体から出てくる光の量は、
「ポアソン分布」に従うと考えられる。

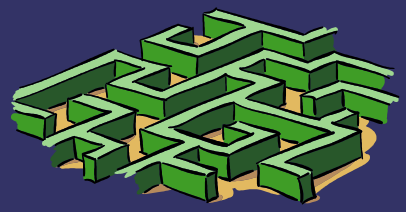
このポアソン分布とは
　非常に確率の低い事象について
　試行を非常に多数行った場合に
　起きた事象の数の分布
であり、二項分布の極限である。



二項分布

ある確率 p で事象が起きる試行を
N回行ったとき、X回事象が起きる
確率分布を二項分布と呼び、
B(N,p) と書く。
平均は Np、分散は Np(1−p)

例えば N=5, p=1/6 の時、
(6面サイコロ5個振った時に
1の目の出る数の確率分布)は
B(5,1/6) である。



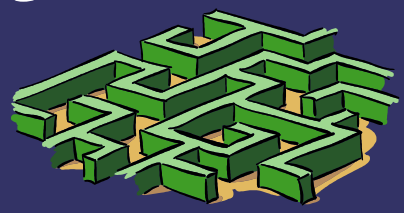
二項分布

$$B(N, p)(X) = {}_N C_X \, p^X (1-p)^{N-X}$$

$${}_N C_X = \frac{N!}{X!(N-X)!}$$

$$B\left(5, \frac{1}{6}\right)(0) = \left(1-\frac{1}{6}\right)^5 \sim 0.402$$

$$B\left(5, \frac{1}{6}\right)(2) = {}_5 C_2 \left(\frac{1}{6}\right)^2 \left(1-\frac{1}{6}\right)^{5-2} \sim 0.093$$



平均の計算(おまけ)

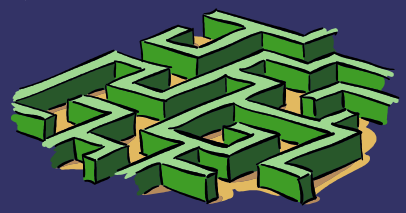
$$E(X)=\sum_{X=0}^{N} X\cdot {}_{N}C_{X}\, p^{X}(1-p)^{N-X}$$

ここでモーメント母関数と呼ばれる
以下の関数を導入する。

$$\phi(t)=E(e^{tX})=\sum_{X=0}^{N} e^{tX}\cdot {}_{N}C_{X}\, p^{X}(1-p)^{N-X}$$

$$=\sum_{X=0}^{N} {}_{N}C_{X}(pe^{t})^{X}(1-p)^{N-X}$$

$$=(pe^{t}+(1-p))^{N}$$



平均の計算（おまけ）

これを t で微分すると

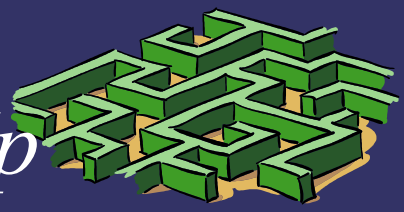
$$\phi'(t)=\sum_{X=0}^{N} X e^{tX}\cdot {}_N C_X\, p^X (1-p)^{N-X}$$

$$\phi'(0)=\sum_{X=0}^{N} X\cdot {}_N C_X\, p^X (1-p)^{N-X}$$

一方で、 $\phi'(t)=Np\,e^t(\,p\,e^t+(1-p))^{N-1}$

$$\phi'(0)=Np$$

$$E(X)=\sum_{X=0}^{N} X\cdot {}_N C_X\, p^X (1-p)^{N-X}=Np$$



分散の計算（おまけ）

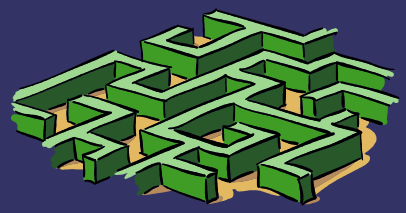
$$V(X)=\sum_{X=0}^{N}(X-Np)^2\cdot {}_NC_X\,p^X(1-p)^{N-X}$$

平均同様にモーメント母関数を使うと

$$\phi''(t)=\sum_{X=0}^{N}X^2e^{tX}\cdot {}_NC_X\,p^X(1-p)^{N-X}$$

$$\phi''(t)=Npe^t(pe^t+(1-p))^{N-1}$$
$$+N(N-1)p^2e^{2t}(pe^t+(1-p))^{N-2}$$

$$\phi''(0)=Np+N^2p^2-Np^2$$



分散の計算（おまけ）

$$V(X)=\sum_{X=0}^{N}(X^2-2NpX+(Np)^2)\cdot {}_NC_X\, p^X(1-p)^{N-X}$$

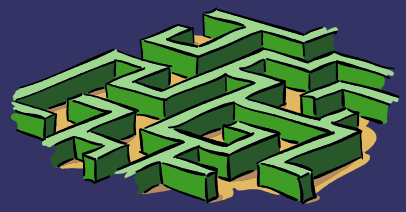
に代入して、

$$V(X)=(Np+N^2p^2-Np^2)-2(Np)(Np)+(Np)^2$$

$$=Np(1-p)$$

が導かれる。

※なお、(X−Np)周りの母関数を
考えたほうが計算は楽。



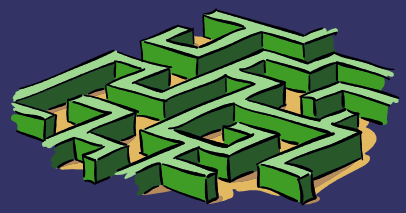
ポアソン分布

平均 E(X)=k
分散 V(X)=k

$$p(k)(x)=\frac{k^x e^{-k}}{x!}$$

二項分布は平均は Np、
分散は Np(1−p)だったのだが、
ポアソン分布は p<<1 の極限なので、
k=Np=Np となる。

従って標準偏差は√k

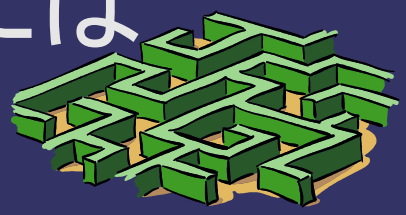


ポアソン分布の和

ポアソン分布に従うX、Yという2つの
変数があったとき、和　X＋Y　も
ポアソン分布に従う。

※二項分布の和は、pが等しければ
　二項分布になるが、その他は
　二項分布には必ずしもならない。

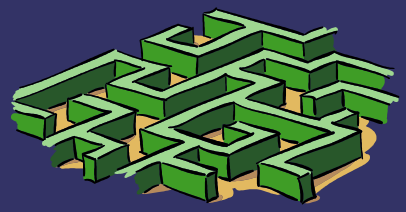
なお、差、X－Yはポアソン分布には
従わない。



余談

X，Yがポアソン分布なら和　X＋Y　も
ポアソン分布に従うのに、
何故X－Yはポアソン分布じゃないのか
気になりませんか？
直接の回答じゃないけど以下をどうぞ。

自然数A，Bの和A＋Bは自然数。
さてA－Bは自然数？？

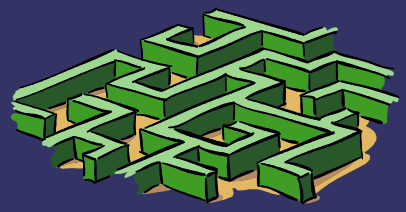


ポアソンの和

今回の場合、天体から出てくる光は
ポアソン分布に従い、空からの光も
ポアソンに従っていると考え、
またダーク電流もポアソン分布に
従っていると考える（ことにする）

この結果この独立な3つの量を足した
「電荷数」もポアソンに従うと
考えて良い。

※ここで中間赤外辺りだと、星から出る光は縮退の影響で
ポアソンに従わないという指摘を頂き調査中です。

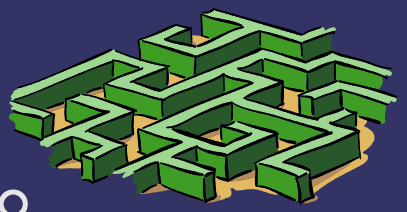


ポアソン分布の定数倍

ポアソン分布の定数倍は
ポアソン分布にならない。
これは要注意。

電荷数はポアソン分布に従うが、
これがFITSになって出てきた値は
ゲインで割り算されているので、
ポアソン分布には従わない。

これが式を電荷数で書いた理由。

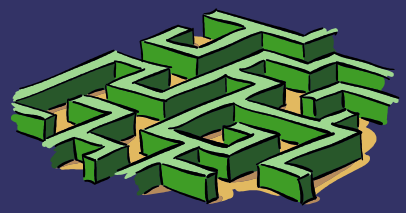


正規分布

別名ガウス分布。
誤差は良くこの分布だと仮定される。
平均m、標準偏差σの場合、

$$N(m,\sigma^2)(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

ポアソン分布はkが大きければ
正規分布で良く近似できる

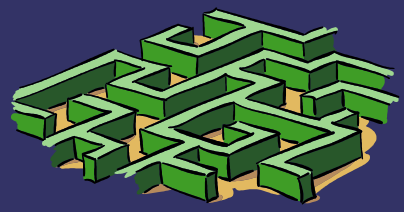


誤差の伝播

さて、このように2種の誤差源があった場合、その合成に用いられるのが

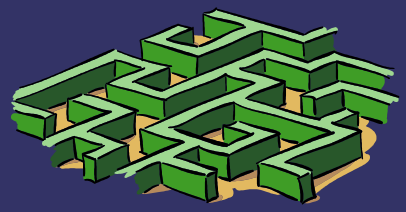
誤差の伝播の式

一般的には分散（標準偏差）が演算によってどのように変わるかをまとめたもの。



ここでの基本方針

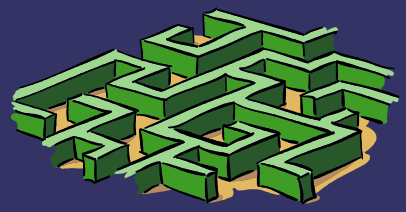
確率変数、X，Yと関数f(X,Y)が
あった場合、f(X,Y)の誤差は
１）確率変数Xと誤差δXを求める
　　Yも同様。
２）期待値E(f(X,Y))を求める
３）分散V(f(X,Y))=E((f(X,Y)−E(f(X,Y))^2)
　　つまり観測値から期待値を引いた
　　二乗平均を求める
という方針で求める。



誤差の伝播の例

X，Yの平均をμ_X, μ_Yと書き、
標準偏差を σ_X, σ_Y とする。
この時、E(X+Y)= $\mu_X + \mu_Y$

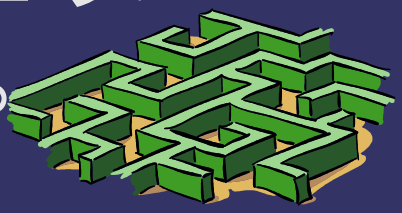
$$V(X+Y) = E(((X+Y)-(\mu_X+\mu_Y))^2)$$
$$= E(((X-\mu_X)+(Y-\mu_Y))^2)$$
$$= E((X-\mu_X)^2)+2E((X-\mu_X)(Y-\mu_Y)) + E((Y-\mu_Y)^2)$$
$$= \sigma_X^2+\sigma_Y^2+2E((X-\mu_X)(Y-\mu_Y))$$



共分散

ここで出てきた $E((X-\mu_X)(Y-\mu_Y))$
を共分散と呼んで $Cov(X,Y)$
と書く。
共分散が0の時、X、Yは独立という。

誤差の伝播で変数が独立かどうかは極めて重要。
しかし実際のデータでは独立かどうか判断するのは簡単とは限らない。



測光のS/N

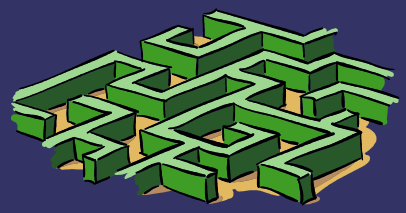
$$V(X+Y)=\sigma_X^2+\sigma_Y^2+2\mathrm{Cov}(X,Y)$$

測光誤差の場合、各ピクセルの読み出し雑音とポアソン雑音は全て独立と考えられるので、ピクセル数をAとおくと

$$N^2=A\,N_1^2+N_2^2$$

$$N=\sqrt{A r^2+S+(f\cdot t\cdot A\cdot s_{SKY})+(t\cdot A\cdot s_{dark})}$$

$$\frac{S}{N}=\frac{f\cdot t\cdot L}{\sqrt{A r^2+t\cdot(f\cdot(L+A\cdot s_{SKY})+A\cdot s_{dark})}}$$



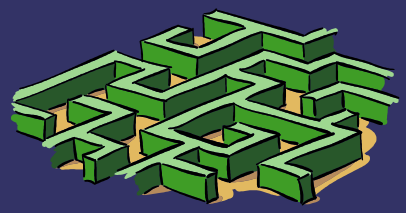
測光の S/N

$$\frac{S}{N}=\frac{f\cdot t\cdot L}{\sqrt{A r^2+t\cdot(f\cdot(L+A\cdot s_{SKY})+A\cdot s_{dark})}}$$

$$t \ll \frac{r^2}{f\cdot(L+A\cdot s_{SKY})+A\cdot s_{dark}} \qquad \frac{S}{N}\sim\frac{f\cdot L}{r}\cdot t$$

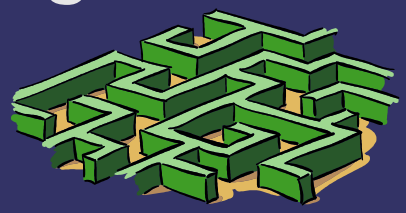
$$t \gg \frac{r^2}{f\cdot(L+A\cdot s_{SKY})+A\cdot s_{dark}}$$ の場合は

$$\frac{S}{N}\sim\frac{f\cdot L}{\sqrt{f\cdot(L+A\cdot s_{SKY})+A\cdot s_{dark}}}\cdot\sqrt{t}$$ ¿



しかし

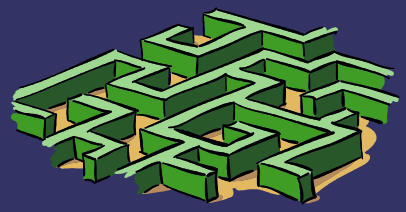
実際には露出時間には上限がある。
・サチり
・ガイドの影響
・宇宙線の増加
従って、露出をいくつかに分割して
撮ることになる。これを足し合わせた
結果は、必ずしも総露出時間の
平方根で S/N が上がるとは限らない



平均とメディアン

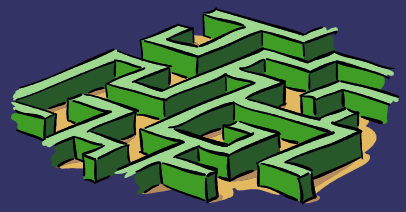
分散で N を評価した場合、S/N を
最も高くする足し合わせ方は単純平均。

しかし、単純平均は外れ値に弱い。
そこで、複数の露出の足し合わせには
sigma-clipped mean や、メディアンが
用いられる。



メディアン

例えば、平均100、分散20の
観測値のうち1つに非常に大きな
ノイズが乗った場合、
113,92,66,10000,104,68,90
これを単純平均を取ると ～1500。
一方、median は 92,
MAD でσを推定した上での、
3σ-clipped mean は、89
単純平均よりは遥かに良い推定

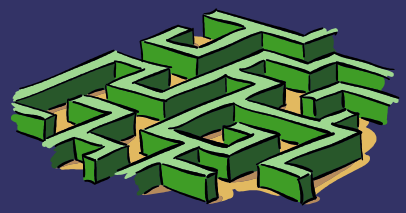


メディアンの誤差

しかし、メディアンは単純平均よりも
誤差が大きくなる。
分散1の正規分布の平均とメディアンを
比べると、Nが十分大きいところでは、

$$s(mean) = \frac{1}{\sqrt{N}}$$

$$s(median) = \frac{1.25}{\sqrt{N}}$$



メディアンの誤差

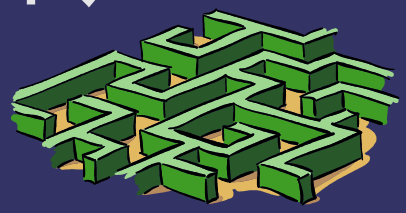
N	s(median)	s(1−clipped)
3	1.16	1.22
4	1.09	1.15
5	1.20	1.12
6	1.14	1.10
7	1.21	1.08

1つだけclip された場合とメディアンでは
N>4 では clipped の方が誤差が小さい

系統誤差の影響

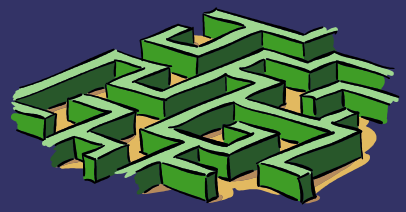
一方、1σ程度以下の系統誤差、
例えば画像の足し合わせでは
バックグラウンドの引き残りなど、
に対しては、
メディアンは引きずられにくいが、
clipped mean は引きずられる。

外れ値が常に大きい側に偏る場合、
メディアンは系統的にずれる



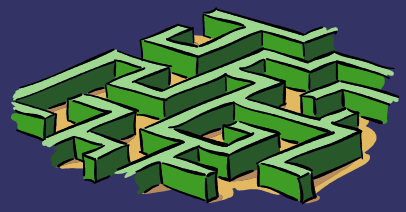
まとめ

- 天体からの光のS/Nのうち、ノイズはポアソン雑音と読み出し雑音の合成であると考えられている
- 露出時間が長くなると理想的な状態でも S/N は露出時間の平方根でしか上がらない
- メディアンよりも単純平均や clipped mean の方がランダム誤差は少ない



続く？

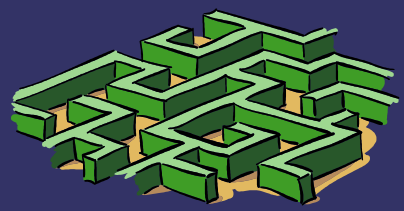
いや、この辺で時間切れ
じゃないかなと・・・



画像の並行移動

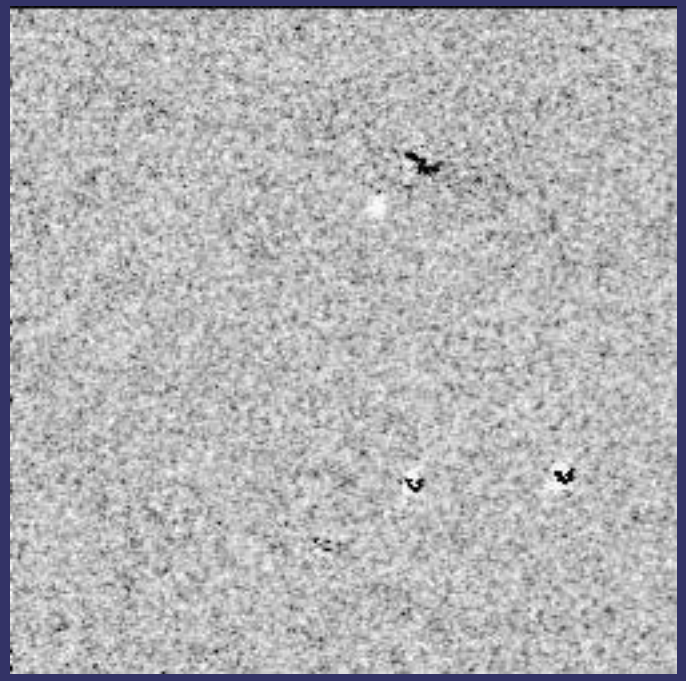
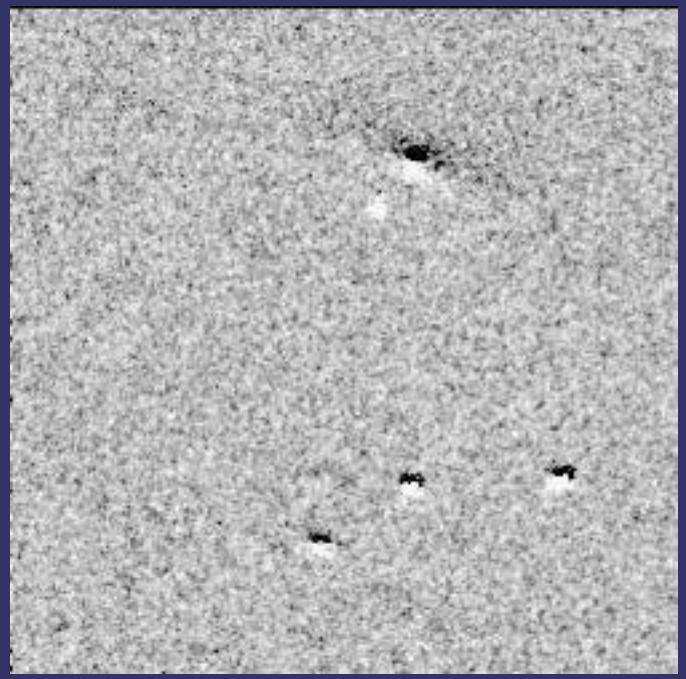
回転も類似の問題があるのだが、
ここでは並行移動だけ考える。

画像を非整数ピクセルだけ
並行移動させたい。
例えば位置を合わせて
足し合わせる場合など。



画像の差（余談）

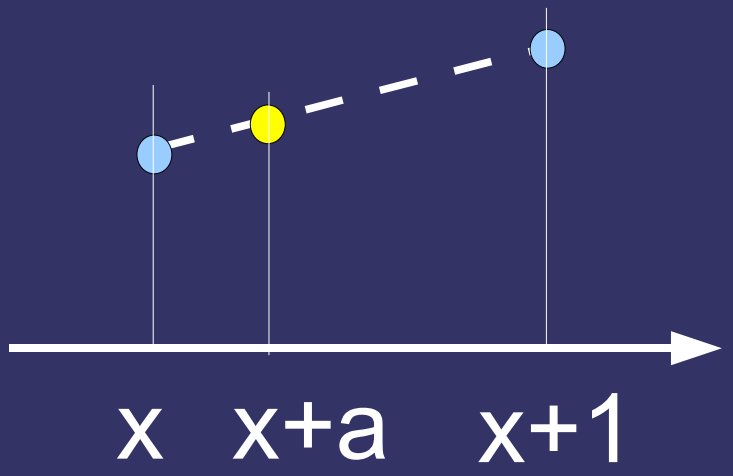
画像の差し引きの時は
特に精度が必要になる。
（右は0.2pixelずれた場合）

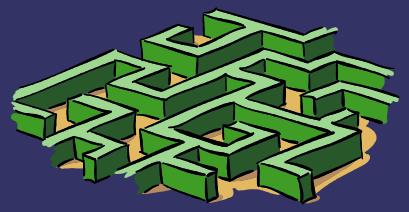




並行移動

並行移動の方法として、線形補間のリサンプリングを考える。
例えばx軸方向に −a (0<a<1) だけシフトさせるときは

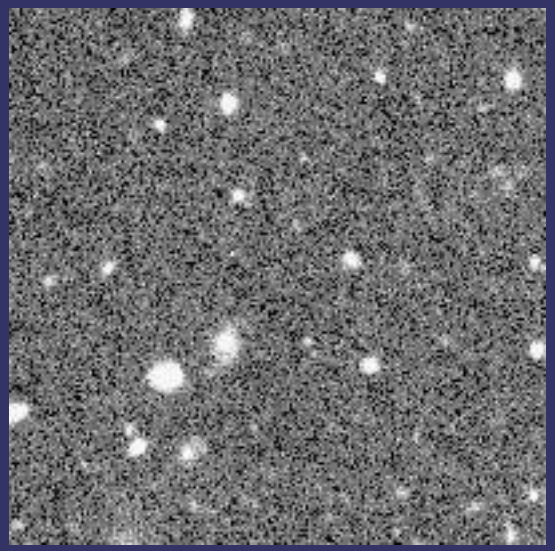
$$v(x')=(1-a)v(x)+a\,v(x+1)$$

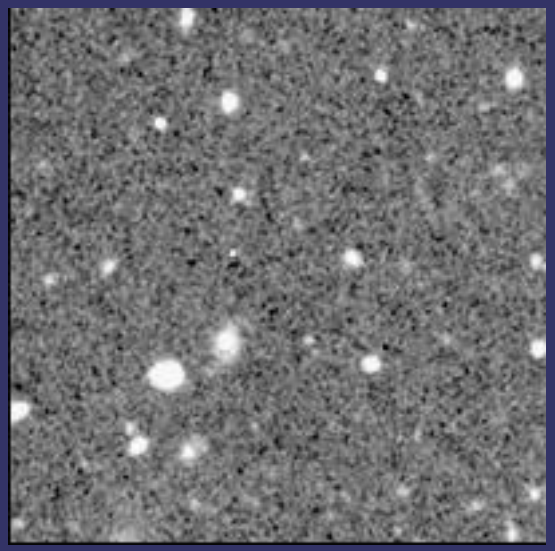




実際にシフトさせた

左がシフト前、右はx+0.5, y+0.5 シフト





ピクセルあたりのノイズは減っている
左67ADU, 右43ADU

誤差の伝播

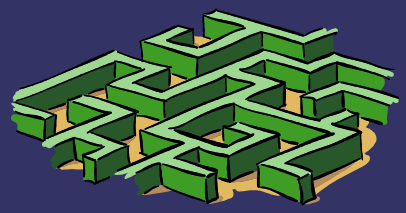
各点での誤差をσとする。X方向だけ考えて

$$v(x')=(1-a)v(x)+a\,v(x+1)$$

とする。各点の誤差が独立であれば、この分散は

$$V((1-a)v(x)+a\,v(x+1))=(1-a)^2\sigma^2+a^2\sigma^2$$

a=0.5 で $0.5\sigma^2$ と小さくなる。

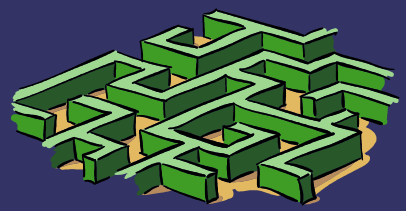


S/N は上がるのか？

一般的には上がらない。

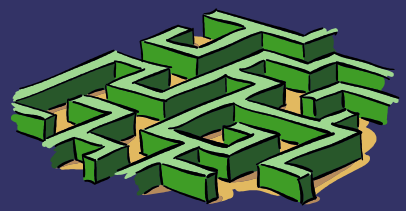
$$\frac{S}{N}=\frac{f\cdot t\cdot L}{\sqrt{A r^{2}+t\cdot\left(f\cdot\left(L+A\cdot s_{SKY}\right)+A\cdot s_{dark}\right)}}$$

まず、この式を求めた際に、
各点での誤差の独立を仮定していたが、
並行移動の結果、独立ではなくなる。



S/Nはあがらない

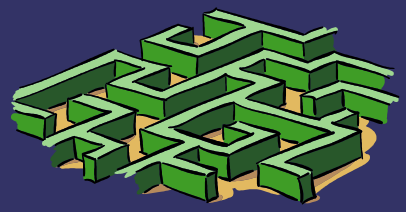
また、S/N は
　その天体の光量をどれだけ正しく
　見積もれたかの指標
だと思うと、並行移動によって
その天体周りの光量がほとんど
変化しない（開口の端での出入りだけ）
ので、S/N が変わるのはおかしい。



画像変形に伴うS/N

これは単純な線形補間による
並行移動の例だが、他の画像変形も
行うことによりピクセル当たりの
ノイズが減る場合が多い。

しかし、天体全体の光量が同じならば
S/N は結果的に変化していないと
考えるのが妥当。



まとめ

- 画像変形を行うとピクセルあたりのS/Nが見かけ上向上して見えることがある
- しかしこの場合、隣り合ったピクセルとの共分散も高くなり、一定以上の範囲でのS/Nは向上しない。
- この点は騙されやすいので要注意