データサイエンス特論
Data Science

確率の基礎

1．　確率とは
2．　確率分布の見方と計算

授業で必要な最小限の簡単な説明に留めます。
予習，復習をしっかり行い、理解しておいてください。
統計のみならずリテラシーとして必須です。

(C) 創造技術コース　橋本洋志／大久保友幸
{hashimoto, ohkubo-tomoyuki}@aiit.ac.jp
http://hhlab.org/

確率とは

1. 身の回りの例
2. 確率密度関数と確率分布関数
3. 例題

身の回りの例

下記の事例，いずれも，10秒以内で直感で答えてほしい

□ 居酒屋の例

- サイコロ1個，客が振るとする。偶数が出ればある飲み物の料金が半分，奇数が出れば2倍の料金を支払うものとする。あなたは，このゲームに乗るか？

□ 誕生日の例

- 1クラス40人生徒がいるとする。誕生日が少なくとも二人以上の確率は何%か？

人間の直感は、時として当たらない。

なぜならば、直感は　その人の知識と経験に基づくから。

人の知識と経験は有限である。

そのため、森羅万象の現実世界を全て見渡すことは不可能である。

では、その現実世界から真実を見抜くにはどうすればよいか？

真実を見抜こうとするため、人間は科学を立脚した。科学は普遍性のあるモデルを打ち立てることを最大の使命としている。このモデルの意味は、現実世界の現象を推定、または、予測するものである。

確率と確率変数

確率（probability）とは，次式で表される比のことである。

$$\frac{\text{期待するある事象（event）の起こる場合の数}}{\text{起こり得る全ての事象の場合の数}}$$

◻ 離散型確率変数

- サイコロの目，コインの表裏，勝敗など

表 4.1　二つのサイコロの目の和の確率分布，一つこぶ型分布の様子を示す

出た目の和 $X=\{x_i\}$	2	3	4	5	6	7	8	9	10	11	12
確率 $P(X=x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

◻ 連続型確率変数

- 離散型確率変数と同じ考え方はできない
- 待ち時間など

連続型確率変数

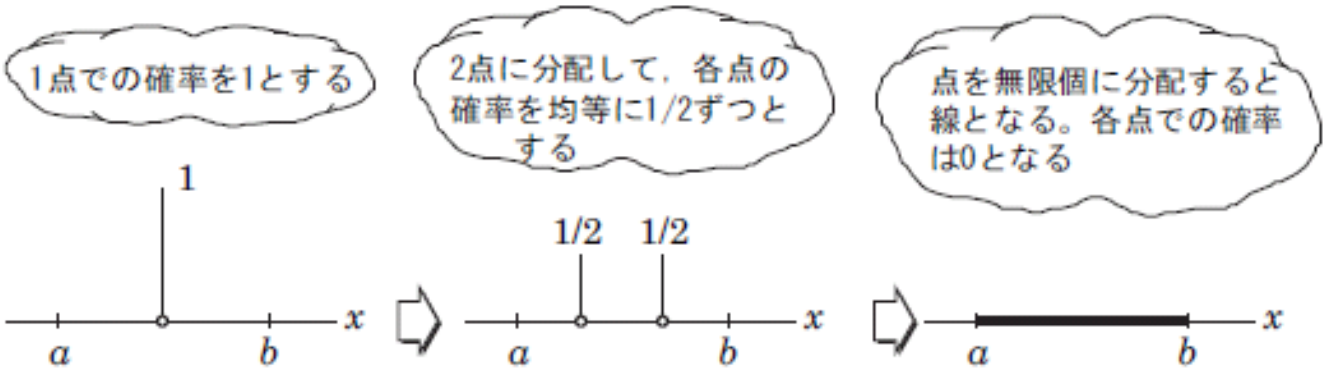


図 4.1　連続型確率変数の図的説明

表 4.2　ATM の利用時間の分布例

利用時間	1 分以内	1~2 分	2~3 分	3~4 分	4~5 分
確率	1/2	1/4	1/8	1/16	1/16

利用時間は1分15秒125などは，離散型確率変数では表現できません！

確率密度関数と確率分布関数

連続型の確率変数

$$F(x) = \int_{-\infty}^{x} f(\tau)d\tau$$

確率密度関数

確率分布関数
（累積分布関数を表すとする）

- $F(-\infty) = 0,\ \ F(+\infty) = 1$
- $x_1 \leq x_2$ ならば $F(x_1) \leq F(x_2)$　　（単調非減少）

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(\tau)d\tau = 1$
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$

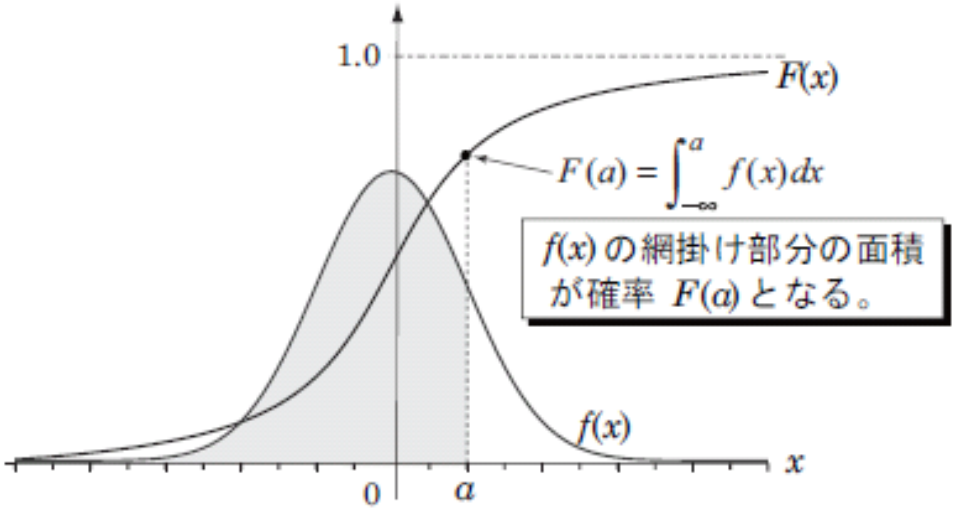


図 4.2　確率密度関数と確率分布関数の関係

平均，分散，分散（標準偏差）

$$E[X] \triangleq \mu = \sum_{i=1}^{N} x_i p_i$$

$$V[X] \triangleq \sigma^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 p_i$$

E[X]： ある数値の集合の期待値（Expected value）
V[X]； ある数値の集合の分散（Variance）

$$X = \{x_1, x_2, \cdots, x_n\}$$

$$P(X = x_i) = p_i$$

x_iが生じる確率がp_iとも読める

$$E[X] \triangleq \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

$$V[X] \triangleq \sigma^2 = E\left[(X - E[X])^2\right] = \int_{-\infty}^{+\infty} (x - \bar{x})^2 f(x) dx$$

- 注意：
  - 関数E と Vだけは強い理由は無いが□を付ける。これは，離散値と連続値とで演算内容が異なるためである。他の関数を表す場合，f()と丸かっこを付ける。
  - ここでのNは，母集団の要素の総数であり，サンプル数のNとは異なる。
  - 平均値は期待値の一部であり，Eと見たら期待値，平均値と称するのは，そのときの使用条件による。
  - 小文字のxは確率変数である。これに対する演算E□,V□は全要素に対しての操作ゆえ，得られる結果は確定値である。
  - 一方，サンプル値（集合の一部）に対する操作の結果は，確率変数である。

期待値の例：宝くじ

期待値＝（確率変数×その値をとる確率）の総和

＝（当せん金(a)×当せん確率(b)）の総和

等級	当せん金 (a)	当せん本数	当せん確率 (b)	a×b
1等	200,000,000 円	2本	0.00002%	40 円
前後賞	50,000,000 円	4本	0.00004%	20 円
組違い賞	100,000 円	198 本	0.00198%	2 円
2等	10,000,000 円	3本	0.00003%	3 円
3等	1,000,000 円	40 本	0.0004%	4 円
4等	100,000 円	100 本	0.001%	1 円
5等	3,000 円	100,000 本	1%	30 円
6等	300 円	1,000,000 本	10%	30 円
夏祭り賞	50,000 円	3,000 本	0.03%	15 円
			期待値→	145 円

この宝くじの場合，300円買って，平均的に145円戻ってくる。

48%の還元率，残りの52%は国庫金に入る。

この計算は，母集団全体に対して行っているので，得られた数値は全て確定値であり，確率変数ではない。

期待値の例：平均点

問題

あるクラスのテストの点数： N = 10人、10点満点

4, 7, 5, 9, 7, 8, 10, 5, 8, 5 ⇒ 合計 68点

先の離散型確率変数の期待値を求める式において
確率 p = 1/N = 1/10 は、どの点数においても同じ

期待値 ＝ $\sum x_i \frac{1}{N} = \frac{1}{N} \sum x_i$

良く知られている、平均値の計算に一致する。
よって、確率が同じ場合には、期待値は平均値となる、

各種分布

1. **連続型確率変数**
   - **正規分布**
   - **t分布**
   - **χ2乗分布**
   - **一様分布**
2. **離散型確率分布**
   - **ベルヌーイ分布**
   - **二項分布**

この二つのみを説明します。
後の分布は、授業では用いません。
興味ある学生が自学習してください。

略語(**abbreviation**)
pdf (Probability density function) 確率密度関数
cdf (Cumulative density function) 累積分布関数
ppf (Percent point function) パーセント点関数
sf (Survival function) 生存関数
isf (Inverse survival function) 生存関数の逆関数

正規分布

正規分布（normal distribution）[*8]は，統計学上最も応用の多い有用な確率密度関数であり，次式で表される。

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{4.35}$$

ここに，平均値が μ，分散が σ^2，π は円周率，e（式中では exp と表現している）は自然対数の底である。また，確率変数 X が正規分布に従う変数であることを示すために

$$X \sim N(\mu, \sigma^2) \tag{4.36}$$

- ガウス分布ともいう
- 分散の平方根である標準偏差σで表現されることが多い。これは，物理量を扱うとき，計測データが長さ[m]であるとき，そのばらつきを分散で見ると[m]×[m]=[m^2]が面積となり，比較が難しくなる。この点，標準偏差ならば単位を[m]で見るのでばらつきの度合いを理解しやすくなるためである。
- ギリシャ文字μは、英語のmに相当するため、平均値（mean value)の記号として良く用いられる。ただし、英語は比較的新しい言語のため、ギリシャ文字とは順番が異なることに注意。

正規分布の見方

平均値 μ を中心とした裾拡がりの単峰形状

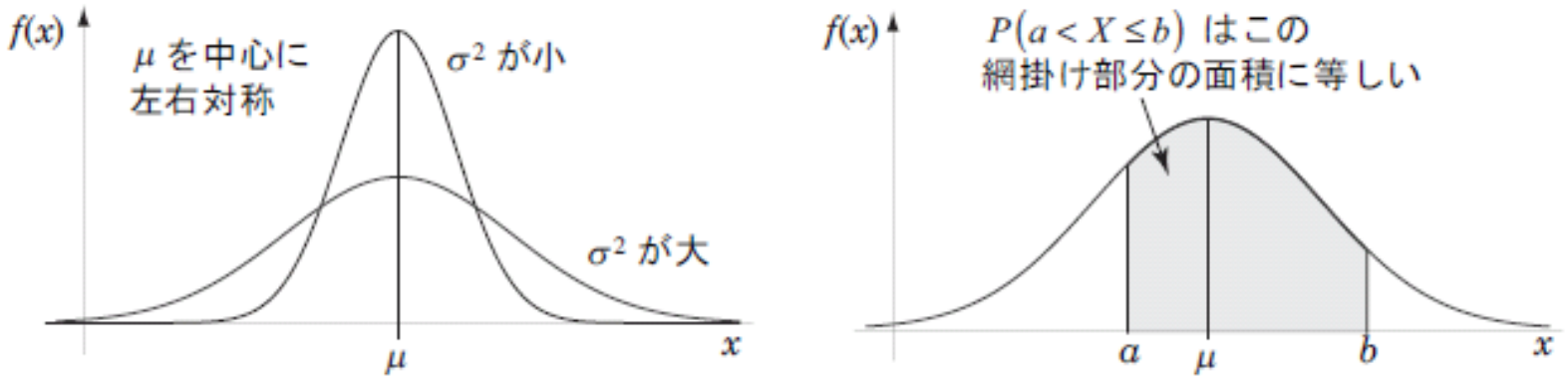


図 4.7　正規分布

P(a < X < b)の読み方は，ある確率変数の値がaより大きくてbより小さい値を示す確率，である。

標準正規分布

- 次のような変換を**標準化**という

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

区間 $[a, b]$ は次のように変換される

$$[a, b] \rightarrow \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]$$

- この確率は，P(a<X<b)に等しい。
- この標準化により，平均値は0，標準偏差は1となる。

- 標準正規分布におけるパーセント点

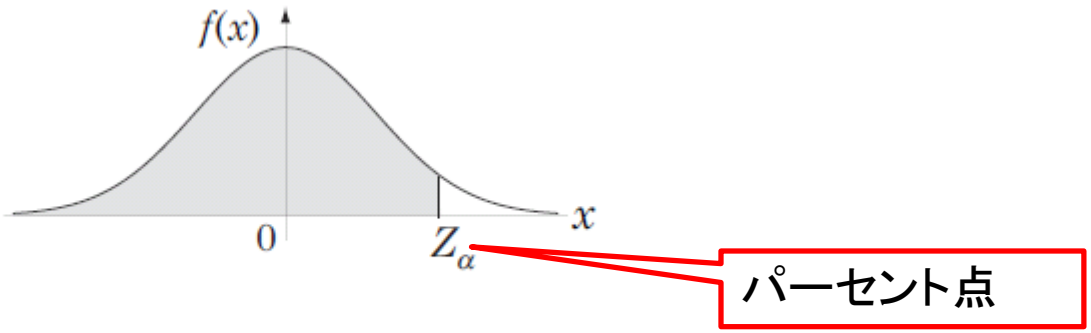


図 4.8　パーセント点

中心極限定理

- この定理があるおかげで正規分布が重要となる。

【定理】 中心極限定理（central limit theorem）

標本 $x_1, x_2, \cdots, x_N$ が独立で，期待値が μ，標準偏差 σ のある同じ分布に従うとする。標本平均を

$$\hat{\mu} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{4.40}$$

とする（$\hat{\mu}$ は，新たな確率変数となることに留意されたい）。標本の大きさ N が限りなく大きくなると，この $\hat{\mu}$ は，平均 μ，標準偏差 $\sigma/\sqrt{N}$ の正規分布に限りなく近づく。別表現で述べると，N が限りなく大きくなると

$$z = \frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{N}} \tag{4.41}$$

は標準正規分布に限りなく近づく。

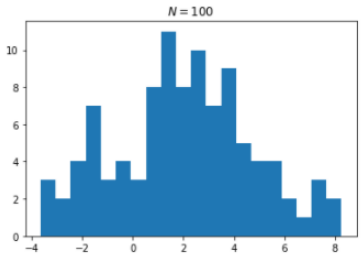
- 標本は，正規分布でないことに注意。シミュレーション例では，一様分布を発生させている。

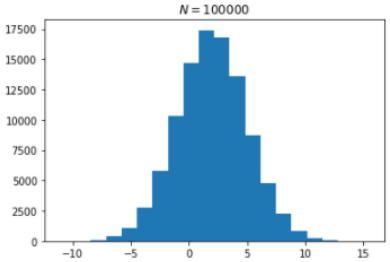
正規分布の計算

PRB_NormalDistribution

正規分布に従う確率変数の発生

- scipy.stats.norm.rvs(loc=mean, scale=std, size=N)

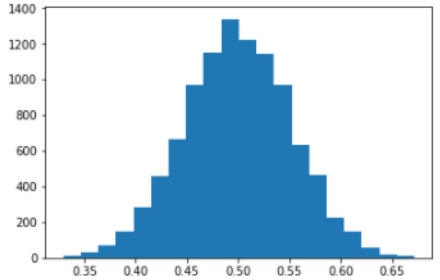




中心極限定理

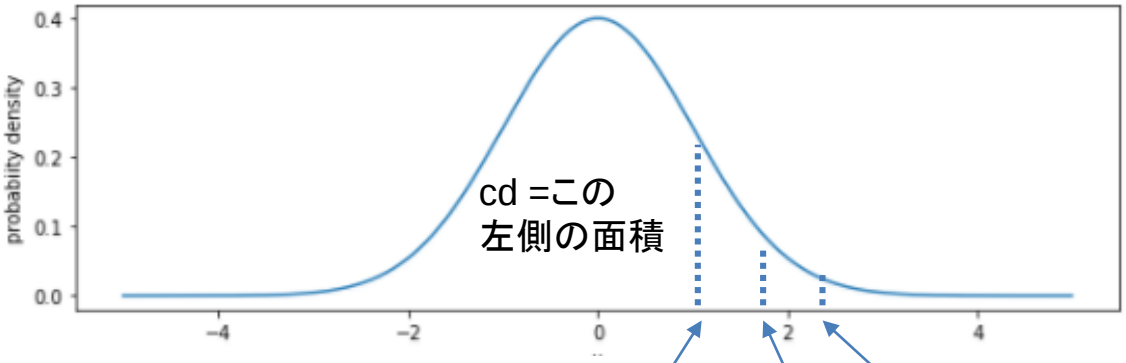
- 一様乱数　20個の平均，10000個の分布

数を変えたら分布の変化は？



パーセント点と確率

- scipy.stats.norm.cdf(x_pp, loc=m, scale=std)
- scipy.stats.norm.ppf(cd, loc=m, scale=std)



x= 1.65,　1.88,　 2.33

パーセント点から確率を求める

例えば，区間 $[-1.96, 1.96]$ の確率を求めたいとする。この場合，**図 4.9** に示すように，区間を $x \leq 1.96$ と $x < -1.96$ に分けて考える。

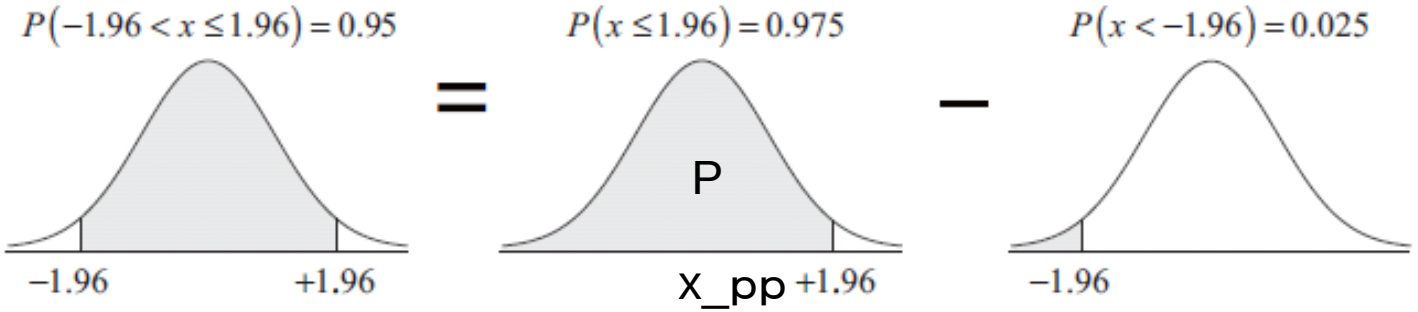


図 4.9　正規分布の確率の求め方

区間 $x \leq 1.96$ の確率は，表 4.4 の $x = 1.96$ を見て，これに対応する $F(x) = P(x \leq 1.96) = 0.975$ を得る。区間 $x \leq -1.96$ の確率は $1 - P(x \leq 1.96)$ より求まり，この値は 0.025 である。したがって，図 4.9 に基づき，求めたい確率は $0.975 - 0.025 = 0.95$ を得る。

または，次の計算でも求まる

```
• x_pp = 1.96
• prob_0 = scipy.stats.norm.cdf(x_pp, loc=m, scale=std)
• prob = 1-2*(1-prob_0)
• print('both side probability = ',prob) # 0.950004209704
```

確率分布関数の各種値

- 接頭語の“norm”を変えれば，他の確率分布関数に適用できる。
- ドット以下の関数名（.ppfなど）の意味が読めるようにしておく
- α（アルファ）は，有意水準または危険率を意味する。

scipy.stats.norm は，正規分布に関する幾つかの計算を行う。

norm.ppf (percent point function，パーセント点関数) α を与えて，確率 $(1-\alpha)$ となるパーセント点（片側）を求める。

norm.isf (inverse survival function，逆生存関数) norm.ppf の $(1-\alpha)$ を計算することなく，直接 α からパーセント点を求める。

norm.interval 区間 $[z_a, z_b], (|z_a| = z_b)$ の確率 $(1-\alpha)$ を与えて，パーセント点 z_a, z_b を求める。

norm.cdf (cumulative density function，累積分布関数) パーセント点を与えて，この確率を求める。

norm.pdf (probability density function，確率関数関数) 確率変数 x を与えて確率密度関数 $f(x)$ の値を求める。

norm.rvs (random variates) 平均値 loc, 標準偏差 scale, サイズ size のランダム変数を発生する。

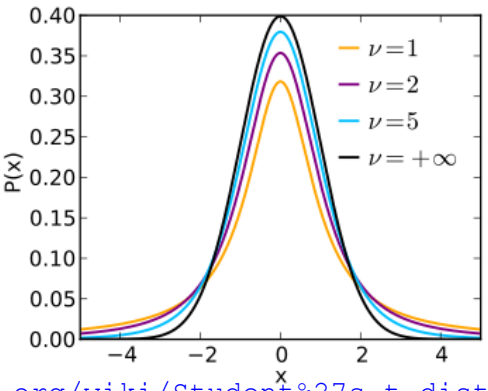
t 分布

標準正規分布N(0,1)は，母分散が既知であった

この母分散がわからない場合，標本分散で代用するという次の考え方

スチューデントの t統計量（Student’s t-statistic）

$$t = \frac{\hat{\mu} - \mu}{\sqrt{\hat{\sigma}^2 / N}}$$



性質

- t は 自由度df = N-1のt分布に従う。 https://en.wikipedia.org/wiki/Student%27s_t-distribution
- t分布は左右対称である。正規分布と同じような裾拡がりの形状であるが，Nの値により，分布のピーク値や拡がり度合いは変わる。
- N -> ∞ のとき，t は標準正規分布に一致する
- 母平均の推定，異なる集団の平均値の差の検定などに用いられる（t検定と言われる）
- t分布の具体的関数を知ることは必要なく，興味があれば他書を参照
- 計算の仕方だけを知ればよい
- ちなみに，平均値と分散は

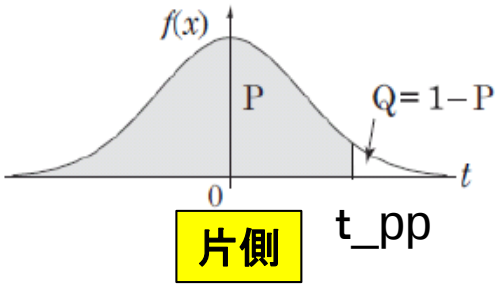
$$E[t] = 0$$

$$V[t] = \frac{df}{df - 2}$$

- 余話：N>30程度のとき，t分布が正規分布に近くなることから，正規分布関数を用いて計算していた。これは，昔は正規分布関数の方が良く計算されていたためである。現在は，ライブラリが計算してくれるから，分散が未知ならばNに関わらずt分布関数を用いて計算してもよい。

t分布のパラメータの計算

- 片側(右図)を求めたい
- t ＝ Tの値を求める
  - Pの値は与えられることが多い。(Q=1%、5%とすることが多いため)
  - scipy.stats.t.ppf(P, df)
  - http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html
  - 引数の意味, P:図の網掛け部分の確率, df ＝自由度(＝N－1)
  - 例、自由度＝5の場合




```
• t_pp = scipy.stats.t.ppf(0.95, 5)   #= 2.01504837267
• prob = scipy.stats.t.cdf(t_pp, 5)   #= 0.949999999958
```

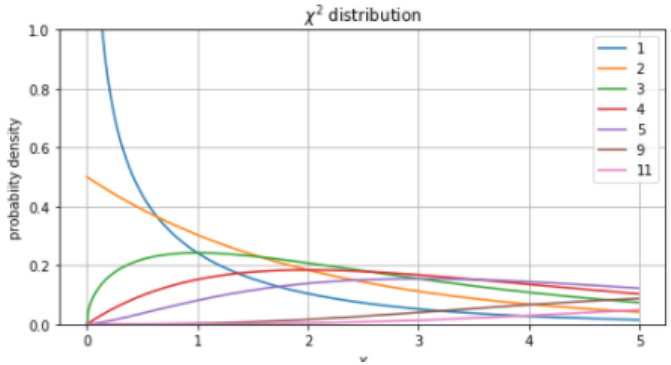
χ^2乗分布（カイ二乗分布）

式

- 標準正規分布に従う確率変数 z_i ～ N(0,1), (i = 1, N)

$$X = z_1^2 + z_2^2 + \cdots + z_N^2$$

- この確率変数Xは，自由度Nのχ2分布に従う。



性質

- 密度関数の分布は，右図のように，N（図ではk）による
- 左右対称ではない。
- Γ関数で表される
- 平均と分散は次で与えられる

$$E[X] = N$$
$$V[X] = 2N$$

用途

- 独立性検定，適合度検定，尤度比検定などに用いられる。
- 後のクロス集計の独立性検定に現れる。

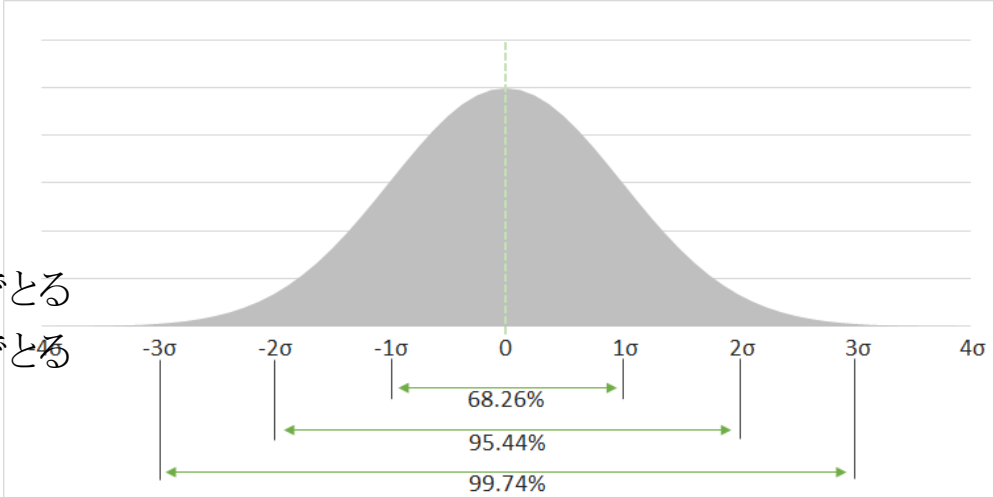
```
•  X_pp = scipy.stats.chi2.ppf(0.95, 6)   #= 12.5915872437
•  prob = scipy.stats.chi2.cdf(X_pp, 6)   #= 0.95
```

χ^2乗分布（カイ二乗分布）

式の解釈

□ 標準正規分布zを考える

- 右図でσ＝1，このとき，zの値は
- ±1の範囲の値を，約68.3%の確率でとる
- ±2の範囲の値を，約95.4%の確率でとる



引用　http://toukeigaku-jouhou.info/2015/08/25/426/

□ z^2を考えているから

- N=1
  - 68.3%の確率で，0～1の値をとる
  - 95.4%の確率で，0～4の値をとる
- N=2
  - z_1^2 + z_2^2 ，この二つともが0～1の範囲になる確率は 68%×68% ＝ 約46% で，0～2の範囲をとる。

□ 見方を変えると

- χ2乗分布は，次のような食い違い度を測っていると見ることができる

$$\text{食い違いの測度} = \sum_{i=1}^{N} \frac{\left(\text{実現値}_i - \text{期待値}_i\right)^2}{\text{期待値}_i}$$

この後の説明は行いません。
また、授業でも用いません。
興味ある学生が自ら学習してください。

ベルヌーイ分布（Bernoulli distribution）

- 0か1をとる離散型の確率変数
- Wikipedia「ベルヌーイ分布」引用

Xをベルヌーイ分布に従う確率変数とすれば、

$$P(X=1)=p, \qquad P(X=0)=q=1-p$$

である。確率変数 X の平均は p, 分散は $pq = p(1-p)$ である。

ベルヌーイ分布の(離散)確率分布は次のように表される。

$$f(k;p)=p^k(1-p)^{1-k} \quad \text{for } k\in\{0,1\}.$$

上式が確率分布であることは、変数kが0, 1の時の分布の値の和をとることで確かめられる。$k=1$ のとき $f(1;p)=p$, $k=0$ のとき $f(0;p)=1-p$ なので、和は 1 である。従って、上式は確率分布の定義を満足する。

ベルヌーイ分布は、離散型確率分布なので「確率密度関数」ではなく「確率質量関数」という

- 平均値と分散

$$E[X]=p$$

$$V[X]=p(1-p)=pq$$

- 二項分布との関係
  - 例えば，ある特定のコインの表の出る確率をp_0として，これを1回投げる。これはベルヌーイ分布である。この「同じコイン」をN回投げたとき，表，裏の出る確率分布を表したのが二項分布である。
  - この定義に従えば，同じNであっても，次にN枚のコインをそれぞれ独立に1回だけ投げて，そのN枚の表裏の分布はそれぞれのベルヌーイ分布に従う。ここで，N枚のコインの表の出る確率p_i，これら全てと，先のp_0が同じであれば，この試行も二項分布とみなせないことはない（理論的に苦しいが，実際において認めたいことが多々ある）。p_0, p_i，一つでも異なるものがあれば言えない。

二項分布（離散型）

1. 各試行において，その事象が発生するか否かのみを問題にする。
2. 各試行は独立である。
3. 事象が発生する確率は一定とする。

1 回の試行において，事象 A が発生する確率を p とする。n 回の試行において，事象 A が起こる回数を表す確率変数を X とすると，k 回事象 A が発生する確率は次式で表される。

$$P(X = k) = {}_nC_k p^k (1-p)^{n-k} \tag{4.19}$$

ここに，${}_nC_k$ は**組合せ**（combination）であり，次で計算される。

$${}_nC_k = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \times \cdots \times 1} \tag{4.20}$$

二項分布（binomial distribution）といい，$\boldsymbol{X \sim B(n,p)}$

$$E[X] = \mu = np$$
$$V[X] = \sigma^2 = np(1-p)$$

二項分布（例１）

【例題1】
- 試行回数がn=2回，ある事象が生じる確率が0.05，このとき，事象が生じない（k=0）の場合の確立を求めよ。
- 表の出る確率が p = 0.5 の硬貨を3回投げたときの確率分布を求めよ。

$$P(X = k) = {}_nC_k p^k (1-p)^{n-k} \qquad (4.19)$$

```
print scipy.stats.binom.pmf(0, 2, 0.05)
n, p  = 3, 0.5
tries = range(n+1)
rv = scipy.stats.binom.pmf(tries, n, p)
print rv
```

実行結果
```
0.9025
[ 0.125  0.375  0.375  0.125]
```

2番目の結果の解釈
- 「k回事象Aが発生する確率」を表しているのだから，
- 事象Aは，「表が出る」だった，
- したがって，実行結果で，例えば k= 0 のとき確率が0.125 とは，「3回投げて，表が0回出る確率は　0.125」と読む。

二項分布（例２）

【例題 2】　サイコロを 8 回振るとき，2 以下の目の出た回数を X とするとき，$P(X \geq 3)$ の値を求めよう。

考え方：補集合で考える方が楽

[解]　目が 2 以下となる確率は $2/6 = 1/3$ であるから，$X \sim B(8, 1/3)$ を考えればよい。これより，

$$P(X \geq 3) = 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$
$$= 1 - \frac{2^8 + 8 \times 2^7 + 28 \times 2^6}{3^8} \approx 0.53$$

```
n = 8
p = 1.0/3.0
tries = range(n+1)
a = scipy.stats.binom.pmf(tries, n, p)
sol = 1.0 - (a[0]+a[1]+a2])
print sol
```

実行結果
0.531778692273

上記の　a(0)はP(X=0), a(1)はP(X=1), a(2)はP(X=2)に対応する。

二項分布（例３）

【例題 3】　75 万人の読者を持つ新聞に広告を出すと，1/3 の読者が注目することがわかっている。この新聞に 6 回広告すると，何万人の読者が注目するかを求めよう。ただし，注目する人数は 2 項分布に従うものとする。

［解］　$B(n,p) = B(6,1/3)$ を用いて，広告回数に対応する注目人数の分布を**表 4.3** に示す。

表 4.3 において，$P(X = x_i)$ は **プログラム**　　より求めた。また，注目する人数は $z_i = P(X = x_i) \times 750$〔千人〕で求めた。

表 4.3　広告回数 $n = 6$ のときの注目する人数の二項分布

注意する回数 $X = x_i$	0	1	2	3	4	5	6
$P(X = x_i)$	0.0878	0.263	0.329	0.219	0.0823	0.0165	0.00137
注意する人数 z_i〔千人〕	66	198	247	165	62	12	1

実行結果は次となる。
0.0877915　0.2633745　0.3292181　0.2194787　0.0823045　0.0164609　0.0013717

二項分布（例３）

表の結果の解釈：

- 6回　広告を出したとき，2回注目する確率（0.329）が最も高く，0回注目する場合もあることがわかる。
- 「何万人の読者が注目するか」 ⇒　表の「注意する人数」を単純に足すことはできない。 なぜなら，ある人Aさんは，注意深くて，必ず　広告を読むならば，全ての　xi　に含まれるためである。
- 結果から言えることは，2回は注目する人は　24.7万人いること。
- １人当たりが注目する平均の回数は，右の計算により2回となった。

$$\frac{\sum_{k=0}^{6} x_i z_i}{750} = 2.0$$

では，１回の広告費用が300万円かかるとしよう。
何回広告を出すのが良いか，シミュレーションで，調べてみよ。

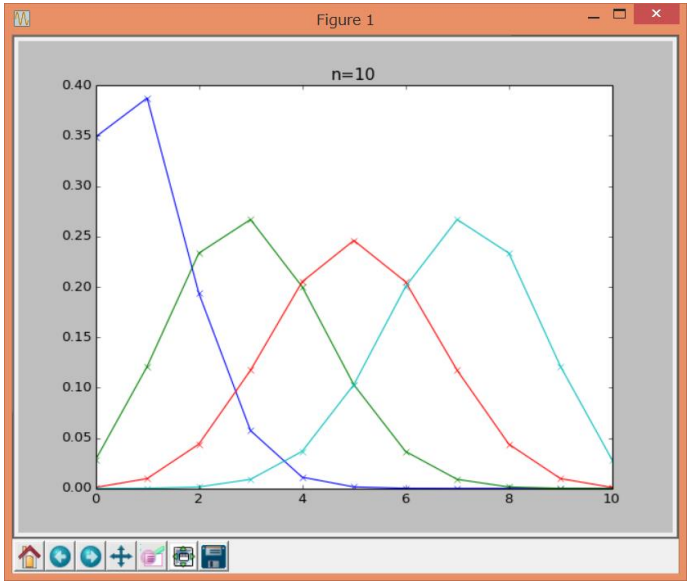
二項分布のプロット

右のグラフは、

n = 10　とおいて

　p = 0.1, 0.3, 0.5, 0.7

の場合の二項分布のプロットです。



ポアソン分布

ポアソン分布（poisson distribution）は，離散型確率変数の一種で，時間間隔 $(0, t]$ の中で平均 λ 回[*6]発生する確率事象が k 回 $(k = 0, 1, 2, \cdots)$ 発生する確率を表現するのに用いられ，次式で示される。

$$P(X = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \tag{4.24}$$

ここに，k は自然数 $(0, 1, 2, \cdots)$，e は自然対数の底（$= 2.71828\cdots$，ネイピア数ともいう）

ポアソン分布に従う過程は，次の性質がある。

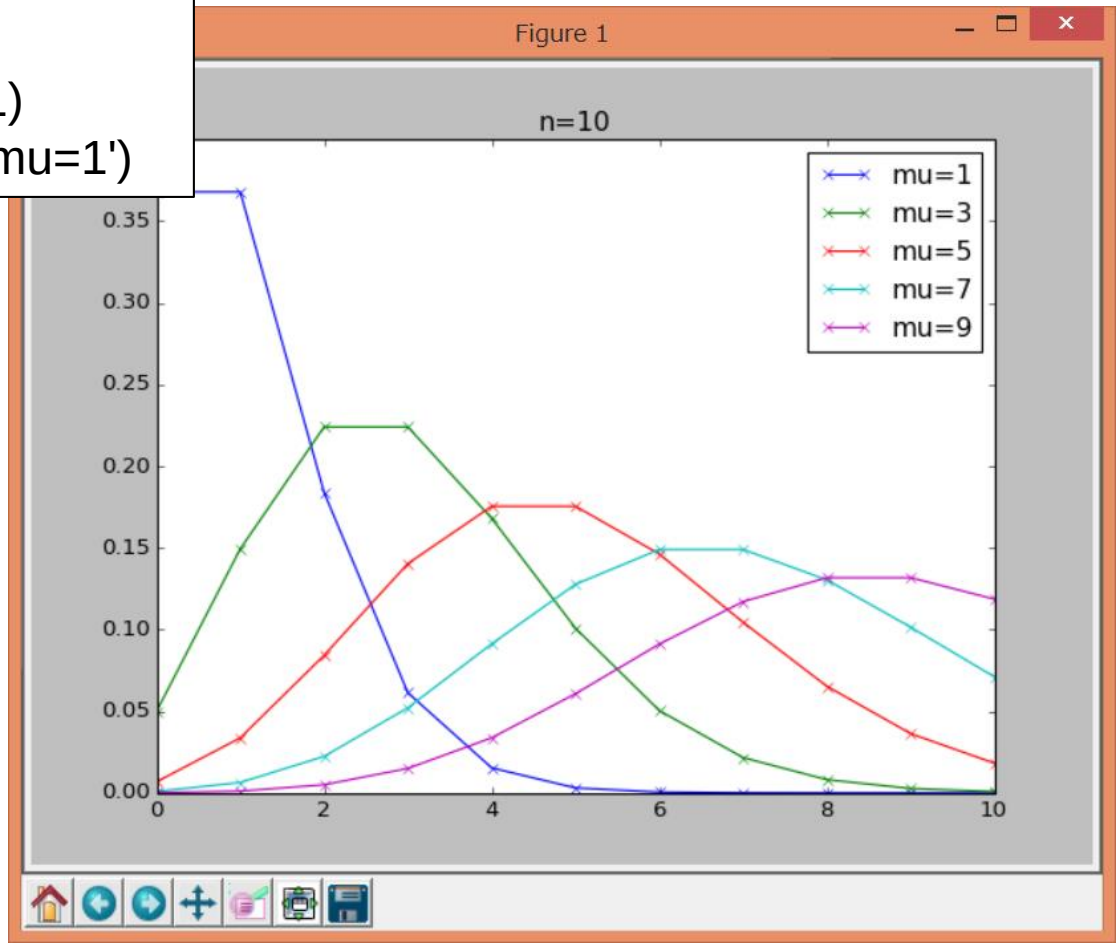
1. （独立性）事象が起きるのは互いに独立である。
2. （定常性）事象が起きる確率はどの時間帯でも同じである。
3. （希少性）微小時間 Δt の間にその事象が 2 回以上起きる確率は無視できるくらい小さいとする。

ポアソン分布が適応されている例として，交通事故発生確率，1 日に受け取る電子メールの件数，単位時間当たりの Web サーバへのアクセス数，単位時間当たりに店や ATM などに訪れる客の数，などがある。

ポアソン分布のグラフ

- λを変数としたとき，P(X=k)，k＝0,1,・・・,10のグラフ

```
n=10
x = scipy.linspace(0,n,n+1)
pmf1 = scipy.stats.poisson.pmf(x, 1)
plt.plot(x, pmf1, marker='x', label='mu=1')
```



ポアソン分布（例１）

【例題 4】　ある都市の交通事故は 1 日平均 2.4 件ある。1 日に起こる交通事故の件数がポアソン分布に従うと仮定したとき，1 日の交通事故が 2 件以下になる確率を求めよう。

[解]　求める確率は $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ である。この右辺の計算は，例えば，平均値と事象の数をそれぞれ変数 m,k に代入して次のようにすれば $P(X = 0)$ を求めることができる。

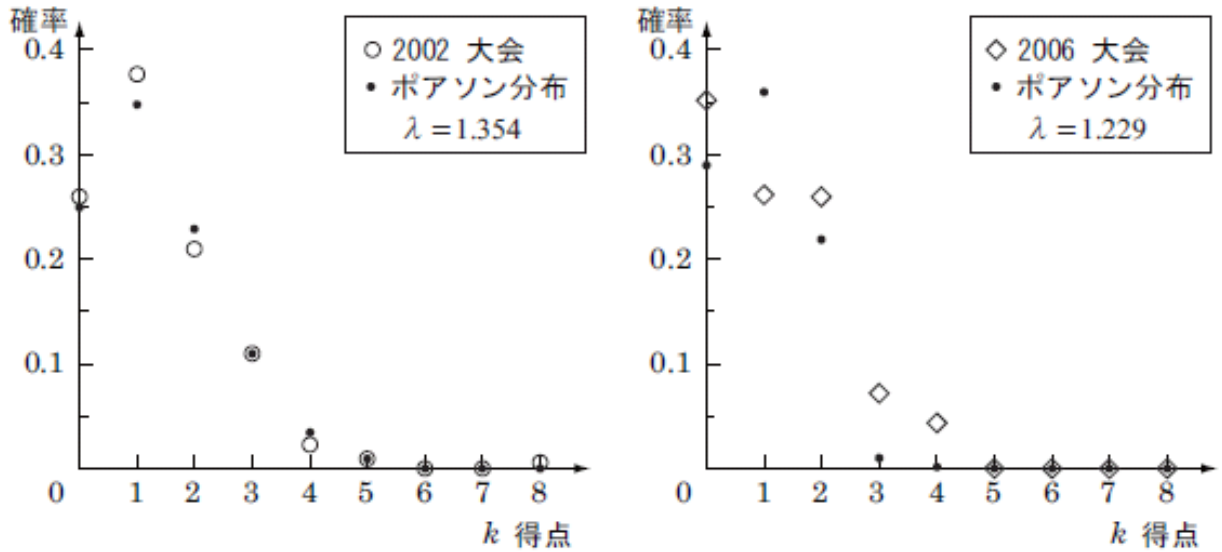
```
m = 2.4
sum=0
for k in [0, 1, 2]:
        sum += scipy.stats.poisson.pmf(k,m)
print sum
```

$P(X = 1)$, $P(X = 2)$ も同様にして求めると，$P(X \leq 2) = 0.56971$ を得る。すなわち，約 57 % の確率で 2 件以下の事故が発生する。

ポアソン分布（例２）

【例題 5】　FIFA ワールドカップ 2002 年と 2006 年大会における 1 次リーグ全 48 試合の得点を調べ，試合で対戦する 2 チームの得点両方を合計した値を 1 試合の得点として，この頻度を求める。頻度を全試合数で割った値を縦軸にとり，横軸に得点分布をとり，これを**図 4.4** に示す。

ここに，図中に示す λ は 1 試合当たりの平均得点である。



(a) 2002年大会　　(b) 2006年大会

図 4.4　試合の得点分布

［解］　図 4.4(a),(b) において，それぞれの λ を用いてポアソン分布 (4.24) 式を計算した結果を図中 • で示す。図 4.4(a) は実際の得点分布がポアソン分布に近いことが認められる。一方，図 4.4(b) は両者が似かよった分布とは言い難い。これは，無得点試合が増えたためで近代サッカー戦法の影響であろうか？

ここで，少し乱暴な議論ではあるが，サッカーの得点は図 4.4(a) よりポアソン分布で表されると仮定しよう。この仮定のもとで，サッカー試合で，弱いチーム (A) と強いチーム (B) が対戦したとき，弱いチームの勝つ確率を考えてみる。

A チームの 1 試合当たり平均得点が $\lambda_A = 1$ 点，B チームのそれを $\lambda_B = 2$ 点とする。例えば，A チームが 1 点，B チームが 2 点得点したことを (A,B)=(1,2) と表記するとき，A チームが勝つ場合は，(A, B)=(1,0), (A, B)=(2,1) と (2,0), (A,B)=(3,2) と (3,1) と (3,0), (A,B)=(4,3) と $\cdots$，$\cdots$ である。各場合の確率を求めると，例えば，(A, B)=(2,1) となる確率は

$$\frac{\lambda_A^{k_A}}{k_A!}e^{-\lambda_A} \cdot \frac{\lambda_B^{k_B}}{k_B!}e^{-\lambda_B} = \frac{1^2}{2!}e^{-1} \cdot \frac{2^1}{1!}e^{-2} = 0.0498 \qquad (4.25)$$

となる。このように，A チームが勝つ場合の確率の和がある程度収束するまで計算を続けると，A チームが勝つ確率は 18.2 % となる。ちなみに，引き分けの確率は同様にして求めると 21.2 % である。これより，A チームが引き分け以上となる確率は 39.4 % となり，10 回対戦中 4 回程度は勝ち点を得られることになる。

ところが，得点力に 2 倍の差があって，野球のように A が平均 3 点，B が平均 6 点の場合に A チームが勝つ確率は 9.5 %，引き分け確率は 8.0 % とぐっと低くなる。

このように確率論から考えると，攻撃力がサッカーのように少ない得点力でかつ 1 点差程度ならば番狂わせが高い頻度で生じることが指摘できる。

ポアソン分布を用いた演習

演習：2億円の宝くじのあたる確率を1000万分の1とする。このとき，

① 番号不揃いの宝くじを2000万枚買った場合、2億円のあたりくじの枚数の期待値は何枚か？

② 2000万枚買って、2億円の宝くじが１枚もあたらない確率はいくらか？　また、１枚だけ当たる確率はいくらか？　ポアソン分布を用いて求めよ。ただし、e^{-2} ≒ 0.135,　　0! = 1　を用いて計算すること。

演習：確率1/400のスロットルマシーンを400回　回したとき，少なくとも1回当たる確率を求めよ。ただし，ポアソン分布に従うものとする。

解答（PPT）

□ 時間があれば，もっと説明（PoissonDistrib.pdf）

一様分布

ある実数区間 $[a, b]$ において，全ての値を同等に取る分布を**一様分布**(uniform distribution) といい，確率密度関数は次式で表される。

$$f(x) = \frac{1}{b-a} \quad (a \leq x \leq b) \tag{4.26}$$

確率変数 X が一様分布に従うことの表記として，

$$\boldsymbol{X \sim U(a, b)} \tag{4.27}$$

を用いる。なお，後に用いるため，一様分布に従う確率変数 X の平均と分散を示す。平均は，(4.11) 式より

$$E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{b+a}{2} \tag{4.28}$$

分散は，(4.12) 式より次式で計算される。

$$V[X] = \int_a^b \left(x - \frac{b+a}{2} \right)^2 \cdot \frac{1}{b-a} dx = \frac{(b-a)^2}{12} \tag{4.29}$$

プログラミング言語が提供する一様乱数（一様分布に従う乱数）は $a = 0, b = 1$ であることが多い。この場合，平均は $1/2$，分散は $1/12$ となる。

一様分布（例）

[解]　(i) 電車を待つ時間を X 分とすると，$X \sim U(0, 10)$ であるから $f(x) = 1/10$，これより求める確率は次で計算される（4.1.2 節参照）。

$$P(X \geq 5) = \int_5^{10} f(x)\,dx = \left[\frac{x}{10}\right]_5^{10} = 0.5 \tag{4.30}$$

(ii) i 日目に電車を待つ時間を X_i 分とすると，1 週間の中で電車を最も待つ時間は

$$X_{\max} = \max(X_1, X_2, X_3, X_4, X_5, X_6, X_7) \tag{4.31}$$

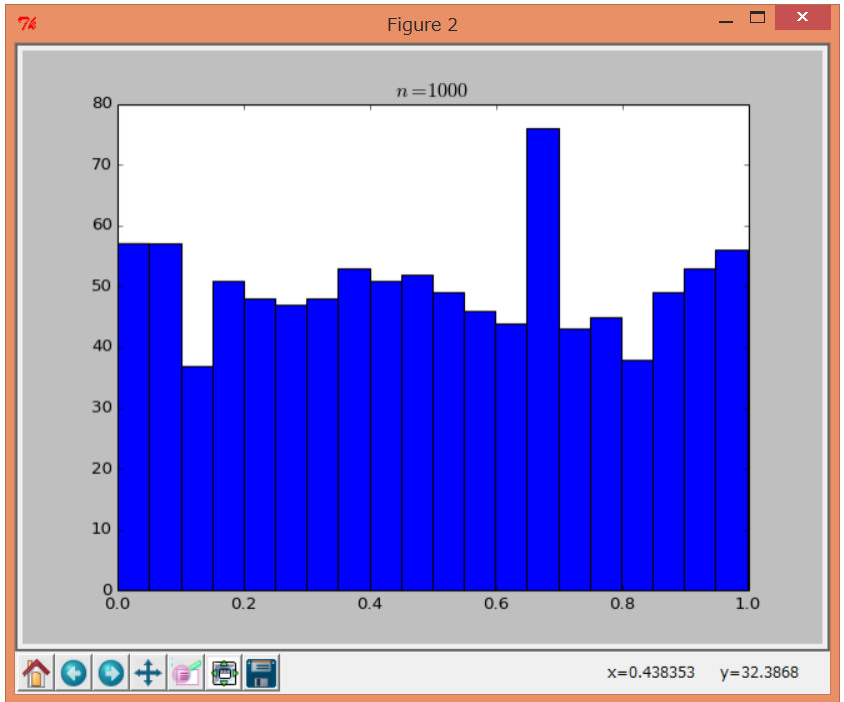
と表せる。これより

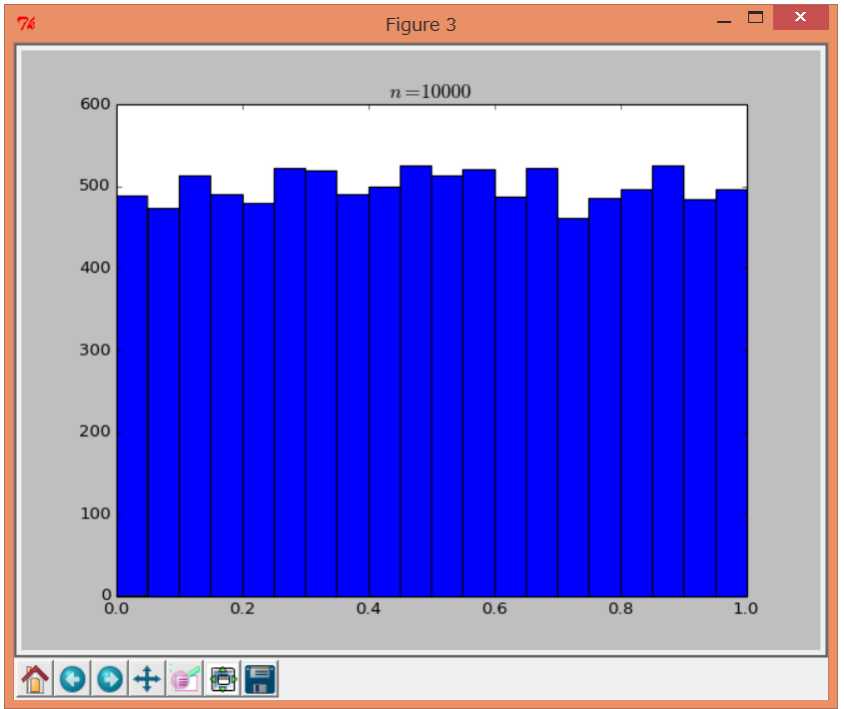
$$\begin{aligned} P(X_{\max} \leq 8) &= P(\max(X_1, X_2, X_3, X_4, X_5, X_6, X_7)) \\ &= P(X_1 \leq 8,\ X_2 \leq 8,\ X_3 \leq 8,\ X_4 \leq 8,\ X_5 \leq 8,\ X_6 \leq 8,\ X_7 \leq 8) \\ &= \prod_{i=1}^{7} P(X_i \leq 8) \\ &= \left(\int_0^8 f(x)\,dx\right)^7 = 0.8^7 = 0.2097 \end{aligned} \tag{4.32}$$

一様分布

```
for n in [100, 1000, 10000]:
        x=scipy.stats.uniform.rvs(size=n)
#       print x.min(), x.max(), len(x)
        plt.figure()
        plt.hist(x, bins=20)
        plt.title('$n=%i$' % (n) )
```

左から n=1000,10000
この程度だと、一様性が認められない。

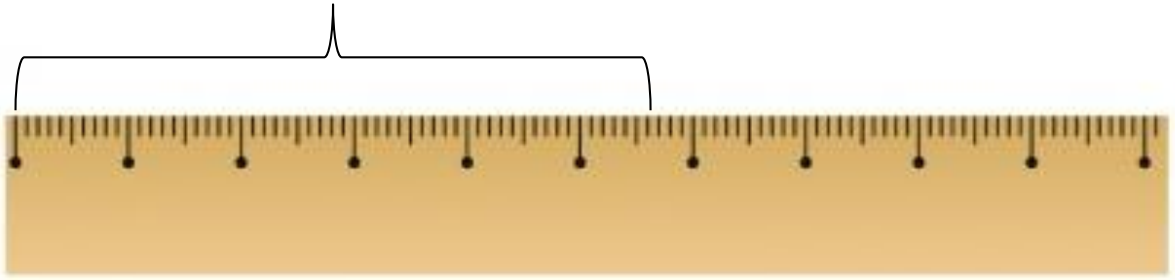




余話：AD変換誤差

- コンピュータを用いたセンシング技術では
  - 外界や物理量はほとんどアナログ量である
  - これをディジタルに変換しなければならない
  - これを Analog to Digital よりAD変換という。
  - 先ほどの例より，10進数を2進数に変換して，有限桁長で打ち切るから，誤差が生じる ⇒ 量子化誤差
  - これは，一様分布に従うノイズが信号に重畳している

アナログで見ると、無限の点が存在する ⇒ 1.234567・・・・ という点は 1.2 となる。
ディジタルで見ると、有限の点である。



指数分布

指数分布（exponential distribution）は，正のパラメータ λ を用いて，確率密度関数が次式で表されるものをいう。

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & if \ \ x \geq 0 \\ 0 & if \ \ x < 0 \end{cases} \tag{4.33}$$

銀行の窓口に客が到着する時間間隔や，発作を起こしてから死亡するまでの時間間隔などに用いられる。

上の定義から，次を得る

$$P(X \leq x) = \int_{-\infty}^{x} f(\tau)\, d\tau = \int_{0}^{x} f(\tau)\, d\tau = \left[-e^{-\lambda\tau} \right]_{0}^{x} = 1 - e^{-\lambda x}$$

$$P(X \geq x) = \int_{x}^{\infty} f(\tau)\, d\tau = e^{-\lambda x}$$

指数分布の例

【例題 7】 ある銀行の ATM の使用時間は平均 2 分間の指数分布に従うとする。このとき，3 分以上使用する確率を求めよう。

［解］ ATM 使用時間を X とすると，求める確率は次で計算される（4.1.2 節参照）。

$$P(X \geq 3) = \int_3^\infty \frac{1}{2} e^{-\frac{1}{2}x} dx = \left[-e^{-\frac{1}{2}x} \right]_3^\infty = 0.2231 \qquad (4.34)$$

最後の数字は次のように計算したものである。

コマンド窓から
> Python
>>> import numpy
>>> numpy.exp(-3.0/2.0)
0.22313016014842982
>>>

指数分布とポアソン分布（例）

ホエール・ウォッチングツアーで，

- ✓ 平均して1時間に1．5頭の鯨が見られる。
- ✓ 鯨の出現回数はポアソン分布に従うものとする。

(a) このツアーで，1頭の鯨を見ることができたとき，それから15分以内にもう1頭見られる確率を求めよ。

(b) このツアーが2時間の場合，ツアー中に鯨が1頭も見られない確率を求めよ。

指数分布とポアソン分布（例）

- 単位時間は1時間，　15分＝1/4 時間 を用いて

$$P\left(X \le x\right) = 1 - e^{-\lambda x} = 1 - e^{-1.5 \times 1/4} = 0.312710\cdots$$

すなわち，15分以内に鯨をもう1頭見られる確率は約32%となる。

- 事象（鯨の出現回数）Xが　ポアソン分布に従う。　単位時間を2時間とする。
- 2時間での鯨の平均出現回数は　1.5 × 2 = 3 となるから，λ＝3とおく。
  3の0乗＝1，　0！＝1　に注意して

$$P(X = 0) = \frac{3^0}{0!} e^{-3} = 0.0497\cdots$$

すなわち，ツアー中に一頭も見ることのできない確率は　約5%である。

後に，品質管理で危険率がでてくる。この危険率は，えいやっー！と5%に決めているテキストが多いが，できる限り，上記のように，できない確率を求めて，それにたいして，費用対効果を計算するのが望ましい。