Corpus Statistics
2008/11/06
Baayen (2008: pp. 53-58)

Toru Hitomi

3.2 Discrete distributions (continued)

これまでは、2 項分布[1]（binomial distribution）を用いて、ある 100 万語コーパスに発現しうる the の頻度確率を算出してきた。
→視点を変えてみる：the の頻度を発現率と捉える。

⇩

重要な離散型確率分布 **Poisson distribution[2]（ポワソン分布）** の 1 つのパラメーター（λ）
試行回数は多いが、発現確率が小さい場合、2 項分布は、ポワソン分布と近似する。
→ポワソン分布は、語彙頻度分布をモデル化するのに便利！（Baayen, 2001）

- R でのポワソン分布を用いる関数←2 項分布の関数と同じつくり
  1. dpois()：頻度分布
  2. rpois()：乱数
  3. qpois()：分位
  4. ppois()：累積分布

Figure 3.4 : 様々なλ値でのポワソン分布　λ値の増加にしたがって連続正規分布に近似する。

- データセット（havelaar）の het の観察データの分布を出したい！
  1000 語あたりの het の token 平均値は 0.0134→1000 語につき、λ = 13.4
  （2 項分布では、1000 回の独立試行で het が発現する確率（p）＝0.0134）
  1. het が発現した頻度ごとに text fragments の頻度をリスト化
     havelaar.tab = xtabs(~havelaar$Frequency)
     havelaar.tab

het の頻度	1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	32	36
text fragments の頻度	1	1	2	2	5	5	8	7	12	8	10	3	4	10	1	5	2	3	1	2	2	1	2	1	1

  2. それぞれの text fragments の数を総数で割る。
     havelaar.probs = xtabs(~havelaar$Frequency) / nrow(havelaar)
     round(havelaar.probs, 3)
  3. plot()を用いてプロットする。
     plot(as.numeric(names(havelaar.probs)), havelaar.probs, xlim=c(0, 40),
     + type="h", xlab="counts", ylab="relative frequency")
  4. mtext()でグラフにラベル付け
     mtext("observed", 3, 1)

---

[1] 背反する事象を複数回繰り返すことにより生起する確率分布
${}_nC_r\, p^r q^{n-r}$ の分布の状態

[2] 発生する事例の確率は低いが発生回数の多い交通事故や原子分裂のような連続して発生する個々の事例の結果に用いられる確率分布；

Corpus Statistics
2008/11/06
Baayen (2008: pp. 53-58)

Toru Hitomi

- het の発現確率（2 項分布）を出したい！
  1. パラメーターをセットする。

     > n = 1000
     > p = mean(havelaar$Frequency/n)
     > p
     [1] 0.01338384
  2. プロットする。

     > counts = 0:40
     > plot(counts, dbinom(counts, n, p), type="h", xlab="counts",
     + ylab="probability")
     + mtext("binomial (1000, 0.013)", 3, 1)
- het の発現確率（ポワソン分布）を出したい！
  1. ラムダ値を定義する。

     > lambda = n * p
  2. プロットする。

     > plot(counts, dpois(counts, lambda), type="h", xlab="counts",
     + ylab="probability")
     + mtext("Poisson (13.4)", 3, 1)

Figure 3.5 からわかること：

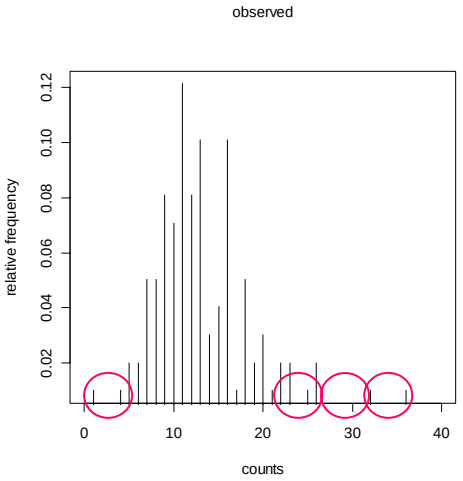
① 実際に観察した頻度分布と、確率分布（2 項分布・ポワソン分布）とに大きな違いがある。
←特定のサンプル（Dekker の小説）だから。

② 2 項分布とポワソン分布が近似している。
←n が大きく、p が小さいから。

③ 実際に観察した頻度分布内に「隙間」がある。
＋対称性が低い。

⇩

実際の観察データと確率分布からの期待データの間の違いが偶然によるものか、これらの確率モデルが不適切であるからなのかどうかをどのように検定するかという問題が生じてくる！！



- ある 100 万語コーパスで 100 tokens という頻度で現れる語が、別の 100 万語コーパスの中で、80 tokens で起きる確率を算出したい。

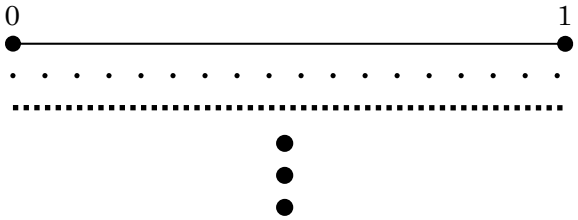
  Pattern 1. > sum(dpois(0:80, 100)) [1] 0.02264918
  Pattern 2. > ppois(80, 100) [1] 0.02264918

Corpus Statistics
2008/11/06
Baayen (2008: pp. 53-58)

Toru Hitomi

3.3 Continuous distributions

ここから、連続確率変数（continuous random variables）での重要な確率分布に入る。
基本的な概念は、離散型の確率変数のときと同じ。
異なる部分は、扱う変数が実数であり、特有の数学的特徴を有している点：
e.g. 0 と 1 の間で、等しい確率で起こるある値（**uniform random variable**）を想定したい。



しかし！無限に想定できるので、それぞれの値での確率を想定できない。
→値と値のある間隔に、ある値が発現する確率を想定する。
e.g. ±1SD の間に入る確率が 68%
→確率密度関数の描き方：連続曲線（Figure 3.6）