第 9 回 続高橋セミナー
最尤法によるポアソン回帰分析入門
2020 年 4 月 19 日

第 1 章　ポアソン分布に従う各種のカウント・データ

ポアソン分布に従うカウント・データの特徴について概説する．まず，ポアソン分布の基本的な特徴である平均値と分散が等しいこと，平均値の増加に伴い分散も増加することを例示する．次に，ポアソン分布が稀に発生する事象の確率分布として導出されることを，二項分布の極限状態を仮定することにより導出されることを示す．さらに，ポアソン回帰に関連する各種の文献データから，ポアソン分布の特徴を浮き彫りにするための事例，各種のポアソン回帰の基本的な事例を取り上げる．また，ポアソン回帰に特徴的な，対数リンク，オフセットの使い分けについても概観する．さらに，過分散となり，通常のポアソン分布のあてはめができない場合の事例などを示す．

第 1 章　目 次

1. ポアソン分布に従う各種のカウント・データ ---------- 7
1.1. ポアソン分布の特徴 ---------- 7
1.2. 2 項分布からポアソン分布の導出 ---------- 10
1.3. 有害雑草の種子の数の分布（1 群） ---------- 13
1.4. 人工データ（恒等リンク, 3 水準, 回帰） ---------- 16
1.5. 冠動脈心疾患の死亡者数（対数リンク, 8 水準, オフセット, 回帰） ---------- 23
1.6. 満月と新月の日の犯罪件数に対する尤度比検定（2 群） ---------- 27
1.7. 細菌を用いた試験データ（2×2 要因配置） ---------- 32
1.8. 細菌を用いた用量反応試験（恒等リンク, 2 群, 8 水準, 効力比） ---------- 36
1.9. 植物の体サイズに関連した種子数（対数リンク, 2 群, 回帰） ---------- 40
1.10. 退役軍人における癌の発生（対数リンク, 2 群, 11 水準, オフセット） ---------- 46
1.11. 喫煙による冠動脈心疾患による死亡（対数リンク, 2 群, 5 水準, オフセット） -- 49
1.12. 医院への通院回数（過分散） ---------- 54
1.13. 雌のカブトガニに連結する雄の数（2 因子, 2 変量, 対数リンク, 過分散） ------ 56
文献索引, 索引, 解析用ファイル一覧 ---------- 63

連絡先:高橋 行雄, 〒105-0014 東京都 港区 芝 1―12－3－1005
takahashi.stat@nifty.com ， FAX : 03-342-8035

第 9 回 続高橋セミナー
最尤法によるポアソン回帰分析入門

第 9 回 続高橋セミナー「最尤法によるポアソン回帰分析入門」は，ページ数が多いので章ごとに公開する．全体の章立てを次に示す．

目 次

はじめに ---------- 1

1. ポアソン分布に従う各種のカウント・データ ----------7

2. ニュートン・ラフソン法によるポアソン回帰 ---------- 63

3 尤度比検定のためのデザイン行列---------- 95

4. デザイン行列を用いた回帰分析入門---------- 135

5. 反復重み付き最尤法によるポアソン回帰 ----------175

6. 過分散・ゼロ過剰への対応 ----------207

7. 過分散がある場合の探索的ポアソン回帰----------237

8. 2 本の回帰直線の比較----------269

9. 花数を共変量とした種子数の分析 ---------- 293

10. オフセットを含む探索的ポアソン回帰---------- 321

11. デビアンス・逸脱度・残差・テコ比---------- 357

12. パラメータの共分散行列の活用 ---------- 379

13. 最小 2 乗平均の謎を予測プロファイルで解く ---------- 417

文献，文献索引，索引，（解析用ファイル）一覧 ----------451

1. ポアソン分布に従う各種のカウント･データ

ポアソン分布に従うカウント・データの特徴について概説する．まず，ポアソン分布の基本的な特徴である平均値と分散が等しいこと，平均値の増加に伴い分散も比例して増加することを例示する．次に，ポアソン分布が稀に発生する事象の確率分布として，2 項分布の極限状態を仮定することにより導出されることを示す．引き続き，ポアソン回帰に関連する各種の文献データから，ポアソン分布の特徴を浮き彫りにするための事例，ポアソン回帰の基本的な事例を取り上げる．それらの事例を通じてポアソン回帰に特徴的な，対数リンク，オフセットの使い分けについても概観し，過分散となり，通常のポアソン分布のあてはめがためらわれるできない場合の事例なども例示する．

1.1. ポアソン分布の特徴

稀に起きるような現象についてある一定期間観察し，それが起きる事象の数が，ポアソン分布に従うことが経験的に知られている．実験研究においては，何らかの刺激 X を加えた場合には，稀に起きる現象 Y が X の強さに応じて多発するようになる．シャーレ上の数十万個の細菌の中で，何らかの異常を起こした細菌のコロニーをカウントするような場合，稀に起きる現象なので出現確率は求められないが，異常を起こしたコロニーの数は，ポアソン分布に従うことことが，経験的に知られている．

ポアソン分布は，分散が平均と同じなので，得られたカウント・データの平均値と分散の比を計算して，同程度であれば，ポアソン分布に従うと判断される．分散を平均で割った比が 1 を大きく超えるような場合は，“過分散”（Over Dispersion）が起きていると言い，実験条件が不均一となっていることが原因として疑われる．また，平均が異なるような複数の集団が存在するような場合にも過分散が起きやすい．

過分散が起きている場合には，平均に対し分散が大きいことを考慮するために，負の 2 項分布から導出されるガンマ・ポアソン分布を用いることもできる．事象が全く起きないゼロ・カウントの頻度が多い場合にも過分散が起きると判断される．この場合には，ゼロ・カウン

トの発生割合を新たなパラメータ ω として加味した，ゼロ過剰（Zero-Ifluraited）ポアソン分布，ゼロ過剰ガンマ・ポアソン分布を仮定することもできる．これらの詳細については，第 6 章で示す．

ポアソン分布は，位置パラメータとしての平均を μ とし，観測値を y としたときに，ポアソン分布の確率関数 $f(y)$ は，次式で与えられる．

$$f(y)=\frac{\mu^{y}e^{-\mu}}{y!},\quad y=1,2,\ldots \qquad (1.1)$$

ポアソン分布の形状の特徴を概観するために Excel でグラフ化を試みる．ポアソン分布の確率関数は，Excel の計算式および Poisson.dist() 関数で次のように計算することができる．

$$f(y)=(\mu\,\hat{}\,y)*\exp(-\mu)/\mathrm{Fact}(y)$$
$$=\mathrm{Poisson.dist}(y,\mu,\mathit{false})$$

3 番目の引数（*false*: 確率 or *true*: 下側確率）

Poisson.dist() 関数を用いて平均を $\mu=0.5,\ 0.8,\ 1,\ 2$ とし，y を 0，1，2,. . .，11 と変化させた場合のポアソン分布の確率を計算した結果を表 1.1 左に示す．さらに平均を $\mu=5,\ 10,\ 15,\ 20$ とし，y を 0，3，6,. . .，33 と 3 刻みで確率の計算をした結果を表 1.1 右に示す．平均が，

表 1.1　Excel の関数を用いたポアソン分布の確率計算

	平均 μ					平均 μ			
y	0.5	0.8	1	2	y	5	10	15	20
0	**0.6065**	0.4493	**0.3679**	0.1353	0	0.0067	0.0000	0.0000	0.0000
1	0.3033	0.3595	**0.3679**	0.2707	3	0.1404	0.0076	0.0002	0.0000
2	0.0758	0.1438	0.1839	0.2707	6	0.1462	0.0631	0.0048	0.0002
3	0.0126	0.0383	0.0613	0.1804	9	0.0363	0.1251	0.0324	0.0029
4	0.0016	0.0077	0.0153	0.0902	12	0.0034	0.0948	0.0829	0.0176
5	0.0002	0.0012	0.0031	0.0361	15	0.0002	0.0347	0.1024	0.0516
6	0.0000	0.0002	0.0005	0.0120	18	0.0000	0.0071	0.0706	0.0844
7	0.0000	0.0000	**0.0001**	0.0034	21	0.0000	0.0009	0.0299	0.0846
8	0.0000	0.0000	0.0000	0.0009	24	0.0000	0.0001	0.0083	0.0557
9	0.0000	0.0000	0.0000	0.0002	27	0.0000	0.0000	0.0016	0.0254
10	0.0000	0.0000	0.0000	0.0000	30	0.0000	0.0000	0.0002	0.0083
11	0.0000	0.0000	0.0000	0.0000	33	0.0000	0.0000	0.0000	0.0020

$y=0,\ \mu=0.5$ ： $\mathrm{Poisson.dist}(y,\mu,\mathit{false})=\mathrm{Poisson.dist}(0,0.5,\mathit{false})=0.6065$

Excel で表 1.1 のような関数計算を効率良く作成するためには，数式の中のパラメータのセルを設定する際に「相対参照」と「絶対参照」を組み合わせて設定するとよい．行方向の平均 μ 列方向の y を引用する際に =Poisson.dist (**$C6**, **D$5**, *false*) のようなセル参照とする．「$」が付いていると絶対参照となる．「$」の付与は，F4 キーを連続的に押すことで設定できる．このようなアドレスにすることにより，数式が縦横に自在にコピーしても表頭・表側のデータを自動的に変化させながら参照させることができる．

$\mu = 0.5$ の場合の $y = 0$ の確率は，Poisson.dist(0, 0.5, *false*)=0.6065 が示されている．平均が $\mu = 1$ の場合は，$y = 0$ の確率と $y = 1$ の確率は，0.3679 と同じになり，平均 $\mu = 7$ の場合には，0.0001 となり，それ以後は，0.0000 以下になる．

図 1.1 に表 1.1 で示したポアソン分布の確率を Excel の縦棒グラフで表示する．平均 μ が 1 よりも小さい場合には，指数分布的な片流れ的な形状であり，平均が 2 から 5 ぐらいまでは，右に裾を長く引くような分布であり，平均が 10 以上になるとやや右に裾を引くが，左右対称な正規分布に近づく．ただし，分散 σ^2 は，平均と同様に増大して行く．図中に，平均 μ，分散 σ^2，標準偏差 σ を上書きしてある．平均 μ 変化による標準偏差 σ の変化の程度を標準偏差を平均で除した変動係数 CV でみると，$\mu = 1$ の場合 $CV = (1/1) \times 100 = 100.0\%$，$\mu = 5$ の場合 $CV = (2.24/5) \times 100 = 44.8\%$，$\mu = 20$ の場合 $CV = (4.47/20) \times 100 = 22.4\%$ と減少している．

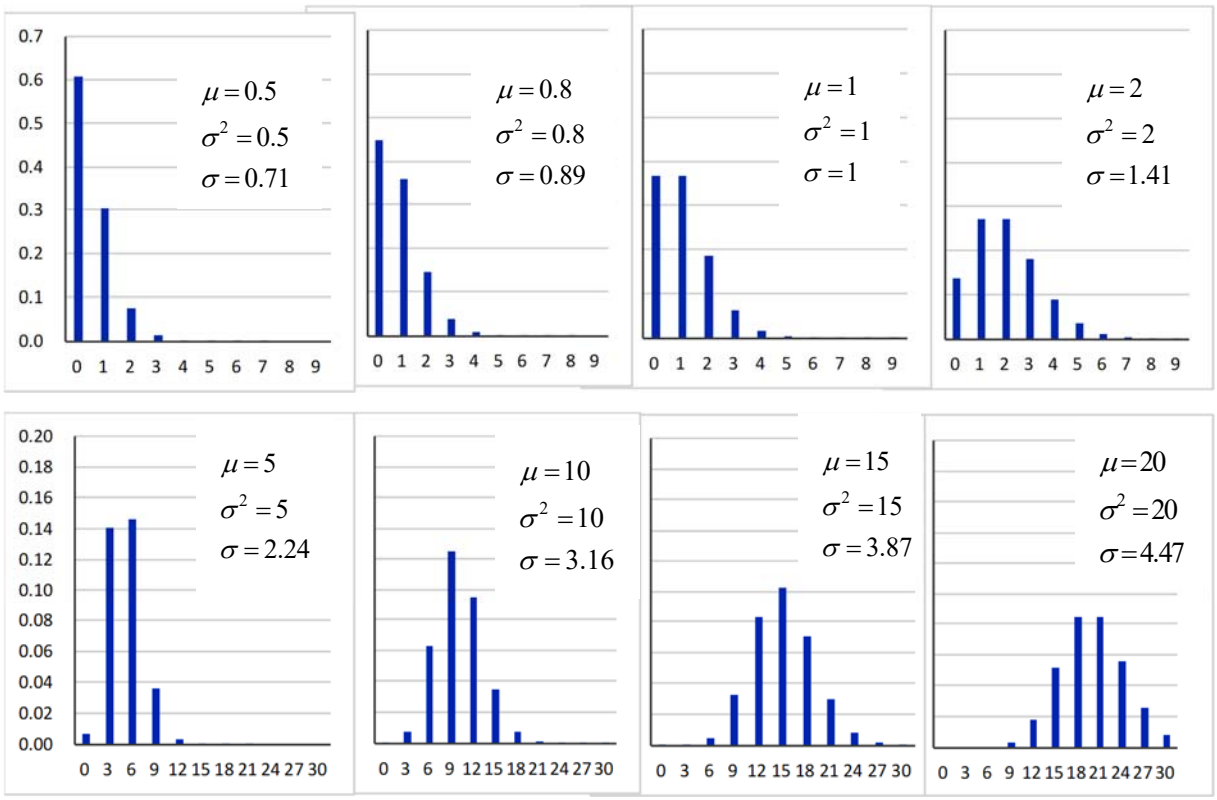


図 1.1　ポアソン分布の形状

1.2. 2 項分布からポアソン分布の導出

ある地域で一日に起きる交通事故を考える．一日を分単位で区切り $n=24\times60=1440$ 分とし，一分ごとに交通事故が起きれば 1，起きなければ 0 とする．一日あたりの事故が起きた件数を y とし，一日あたりの事故件数の平均を μ とする．一分ごとに事故が起きる確率を $\pi=\mu/n$，起きない確率を $1-\pi$ とする．一日あたりの事故件数を y とした場合の 2 項分布は，次式で与えられる．

$$\binom{n}{y}\pi^{y}(1-\pi)^{n-y}=\frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\mu}{n}\right)^{y}\left(1-\frac{\mu}{n}\right)^{n-y} \tag{1.2}$$

十分に大きい n に対して y はごく小さいので

$$n(n-1)\cdots(n-y+1)\approx n^{y} \tag{1.3}$$

に置き換えることができる．事故が起きない確率は，

$$\left(1-\frac{\mu}{n}\right)^{n-y} \tag{1.4}$$

であり，十分大きい n に対しては，

$$\left(1-\frac{\mu}{n}\right)^{n-y}\approx\left(1-\frac{\mu}{n}\right)^{n} \tag{1.5}$$

に置き換えることができる．数学の標準的な公式により，

$$\frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\mu}{n}\right)^{y}\left(1-\frac{\mu}{n}\right)^{n-y}=\frac{n^{y}}{y!}\left(\frac{\mu}{n}\right)^{y}e^{-\mu} \tag{1.6}$$

となる．これは，n が無限大に近付いたときに $(1-\mu/n)^{n}$ は，$e^{-\mu}$ に近付くためである．整理すると n 含まないポアソン分布 P_y となる．

$$\begin{aligned}P_y&=\frac{n^{y}}{y!}\left(\frac{\mu}{n}\right)^{y}e^{-\mu}\\&=\frac{\mu^{y}e^{-\mu}}{y!}\end{aligned} \tag{1.7}$$

ポアソン分布は，μ のみの関数であり，期待値と分散が共に μ となる．期待値は，

$$\begin{aligned}E(y)&=\sum_{y=0}^{\infty}yP_y\\&=\sum_{y=0}^{\infty}\frac{y\mu^{y}e^{-\mu}}{y!}\\&=\sum_{y=1}^{\infty}\frac{\mu^{y}e^{-\mu}}{(y-1)!}\end{aligned} \tag{1.8}$$

と変形でき，さらに，μ を Σ の外に出し，$i=y-1$ と置きなおすと，Σ の中は，i についてのポ

アソン分布となり 0 から∞までの和は，分布関数の性質から 1 となり，

$$
\begin{aligned}
E(y) &= \mu \sum_{y=1}^{\infty} \frac{\mu^{y-1} e^{-\mu}}{(y-1)!} \\
&= \mu \sum_{i=0}^{\infty} \frac{\mu^{i} e^{-\mu}}{i!}, \quad (i = y-1) \\
&= \mu
\end{aligned}
\tag{1.9}
$$

期待値が μ となることが確認される．簡便公式 $V(y) = E(y^2) - E(y)^2$ によって分散を求めるために，まず， $E[y(y-1)] = E(y^2) - E(y)$ を求める．

$$
\begin{aligned}
E[y(y-1)] &= \sum_{y=0}^{\infty} y(y-1) \frac{\mu^{y} e^{-\mu}}{y!} \\
&= \mu^2 \sum_{y=2}^{\infty} \frac{\mu^{y-2} e^{-\mu}}{(y-2)!} \\
&= \mu^2
\end{aligned}
\tag{1.10}
$$

これから，$E(y) = \mu$ なので，$E(y^2) - E(y) = \mu^2$ から，$E(y^2) = \mu^2 + \mu$ が得られる．これらから，分散 $V(y)$ は，

$$
\begin{aligned}
V(y) &= E(y^2) - E(y)^2 \\
&= \mu^2 + \mu - \mu^2 \\
&= \mu
\end{aligned}
\tag{1.11}
$$

のように期待値 μ と同じになる．つまり， μ に比例して分散が大きくなる．

どのくらいの大きさの n から，ポアソン分布は 2 項分布に近似できるのか検討してみよう．表 1.2 に示すように，事故件数 y の期待値を $\mu = 1.5$ と固定し，母数団の大きさを $n = 10, 100$，

表 1.2　ポアソン分布の 2 項分布に対する近似精度

		n	μ	π	n	μ	π	n	μ	π
		10	**1.5**	0.1500	100	**1.5**	0.0150	1440	**1.5**	0.0010
i	y	二項	ポアソン	差	二項	ポアソン	差	二項	ポアソン	差
1	0	**0.1969**	0.2231	**0.0263**	0.2206	0.2231	**0.0025**	0.2230	0.2231	**0.0002**
2	1	0.3474	0.3347	-0.0127	0.3360	0.3347	-0.0013	0.3348	0.3347	-0.0001
3	2	0.2759	0.2510	-0.0249	0.2532	0.2510	-0.0022	0.2512	0.2510	-0.0002
4	3	0.1298	0.1255	-0.0043	0.1260	0.1255	-0.0005	0.1255	0.1255	0.0000
5	4	0.0401	0.0471	0.0070	0.0465	0.0471	0.0005	0.0470	0.0471	0.0000
6	5	0.0085	0.0141	0.0056	0.0136	0.0141	0.0005	0.0141	0.0141	0.0000
7	6	0.0012	0.0035	0.0023	0.0033	0.0035	0.0003	0.0035	0.0035	0.0000
8	7	0.0001	0.0008	0.0006	0.0007	0.0008	0.0001	0.0008	0.0008	0.0000
9	8	0.0000	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0000
10	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

1440 と変化させ，ポアソン分布と 2 項分布の発生件数 y の確率を計算し，その差を求める．ただし，2 項分布の出現確率は $\pi = \mu / n$ で与えられる． ポアソン分布および 2 項分布の確率は，Excel の関数を用いて

ポアソン分布： $P_{y_i} = \text{Poisson.dist}(y_i, \mu, \mathit{false})$

2 項分布： $B_{y_i} = \text{Binom.dist}(y_i, n, (\mu / n), \mathit{false})$

のように計算する．ポアソン分布の計算には，n が含まれていないので，n を変えても同じ確率である．2 項分布の場合には，n が 10 の場合には $B(y_1 = 0) = 0.1969$ でポアソン分布の場合の $P(y_1 = 0) = 0.2231$ に比べて 0.0263 小さいが，n が 100 の場合は 0.0025 と差は減少し，n が 1440 の場合は 0.0002 と差は更に減少し，2 項分布はポアソン分布に漸近する．

一日に起きる事故件数の期待値を $\mu = 1.5$ としたときに，一年間 365 日の事故件数の分布を表 1.3 に示す．度数 n_i は，ポアソン分布の確率 P_i に 365 日を掛けて四捨五入して整数化し，平均を計算するために積和 $n_i P_i$ を計算し，その合計を 365 日で割って平均 1.5018 件が得られている．分散の計算のためには，偏差平方和 $\Sigma_{i=1}^{10} n_i (y_i - \mu)^2 = 550.47$ を求め，365 日で割って分散を計算し，1.5077 が得られ，分散と平均がほぼ等しいことが確認される．

表 1.3　ポアソン分布の平均と分散

	事故件数	ポアソン	度数	積和	偏差平方
i	y	P	n	$n\,y$	$n(y\text{-}\mu)^2$
1	0	0.2231	81	0	183.5850
2	1	0.3347	122	122	30.7679
3	2	0.2510	92	183	22.7372
4	3	0.1255	46	137	102.8055
5	4	0.0471	17	69	107.3469
6	5	0.0141	5	26	63.6352
7	6	0.0035	1	8	26.3042
8	7	0.0008	0	2	9.0691
9	8	0.0001	0	1	4.2227
10	9	0.0000	0	0	0.0000
		計	365	548	550.4738
			平均	**1.5018**	**1.5077**
				μ	**分散**

分散は，μ を既知としているので，偏差平方和を 365 で除している．

1.3. 有害雑草の種子の数の分布（1 群）

生物統計の名著として知られている スネデカー・コクラン著, 畑村・奥野・津村 訳 (1972) の「統計的方法，第 6 版」，の第 8.14 節に *Phleum praetense*（イチゴツナギ）の 98 副標本に含まれる有害雑草の種子の数が示されている．各副標本の重量は 1/4 オンスで，もちろん沢山の種子を含んでおり，その中のほんの少数が有害雑草のものであった．表 1.4 に観測度数，期待度数，その差，適合度のカイ 2 乗値を示す．

表 1.4 有害雑草の種子の数に対するポアソン分布のあてはめ

i	有害種子の数 y	観測度数 n	積和 ny	偏差平方 $n(y-\mu\hat{})^2$	ポアソン確率 P	期待度数 $n\hat{}=NP$	観測-期待 $n-n\hat{}$	適合度 $\frac{(n-n\hat{})^2}{n\hat{}}$	(-2)*対数尤度 $n[-2\ln(P)]$
1	0	3	0	27.3686	**0.0488**	4.7806	-1.7806	0.6632	18.1224
2	1	17	17	69.3948	0.1473	14.4393	2.5607	0.4541	65.1106
3	2	26	52	27.0721	0.2225	21.8062	4.1938	0.8065	78.1441
4	3	16	48	0.0067	0.2240	21.9546	-5.9546	1.6150	47.8717
5	4	18	72	17.2728	0.1692	16.5779	1.4221	0.1220	63.9682
6	5	9	45	35.2691	0.1022	10.0144	-1.0144	0.1028	41.0569
7	6	3	18	26.6339	0.0514	5.0413	-2.0413	0.8265	17.8038
8	7	5	35	79.1858	0.0222	2.1752	2.8248	3.6682	38.0783
9	8	0	0	0.0000	0.0084				0.0000
10	9	1	9	35.7555	0.0028	**1.2104**	-0.2104	0.0366	11.7474
11	10	0	0	0.0000	0.0008	(8以上)			0.0000
12	11	0	0	0.0000	0.0002				0.0000
	計	98	296	317.9592	=平方和	**分散/平均**	**カイ2乗=**	**8.2949**	**381.9035**
	平均 $\mu\hat{}$=	**3.0204**	**3.2779**	**=分散**		**1.0853**	***p*=**	**0.3073**	

結果が度数分布で与えられているので， N を観測度数 n_i の和とし， $n_i y_i$ の和を N で除して平均 $\hat{\mu}=3.0204$ が得られる．

$$N=\sum_i n_i=98,\quad i=1,\ 2,\ldots,12$$

$$\hat{\mu}=\frac{\sum_i n_i y_i}{N}$$

$$=\frac{296}{98}=3.0204$$

分散 $\hat{\sigma}^2=3.2779$ は，平方和を自由度で割り

$$\hat{\sigma}^2=\frac{\sum_i n_i(y_i-\hat{\mu})^2}{N-1}$$

$$=\frac{317.9592}{98-1}=3.2779$$

が得られる．分散/平均の比は， $3.2779/3.0204=1.0853$ であり，ほぼ 1 に近いと判断される．

ポアソン分布のあてはめが適切かを検討するために，y_i に対するポアソン分布の確率 P_i を求め，観測度数の和 N を掛けて期待度数 $\hat{n}_i = NP_i$ を計算し，観測度数 n_i と比較する．ポアソン分布の確率は，Excel の関数を用いて，

$$P_i = \text{Poisson.dist}\,(\,y_i,\ \hat{\mu}\,, \mathit{false})$$

で計算する．有害種種子の数 $y_1 = 0$ の場合は，

$$P_1 = \text{Poisson.dist}(\,0,\, 3.0204, \mathit{false}) = 0.0488$$

となる．期待度数 $\hat{n}_i$ は，$\hat{n}_i = NP_i$ で求められる．ただし，$\hat{n}_{10} = 1.2104$ の場合 Excel の関数で $y_{10} = 9$ 以上は，8 以下の下側確率から差し引いて N=98 を用いて

$$\hat{n}_{10} = 98 \times (1 - \text{Poisson.dist}(\,7,\, 3.0204,\, \mathit{true})) = 1.2104$$

のように上側確率を計算し観測度数 n_i の合計 $N = 98$ を掛けて期待度数を計算している．これは，期待度数の和が，観測度数の和 $N = 98$ と一致させるためである．観測度数と期待度数との差 $n_i - \hat{n}_i$ は，若干の凸凹があるものの大きな乖離はない．

ポアソン分布に従っているかの適合度の検定の χ^2 値は，

$$\chi^2{}_{9-1-1} = \sum_i \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 8.2949$$

であり，上側確率 p は，Excel の関数を用いて

$$p = 1 - \text{Chisq.dist}(8.2949,7, \text{true}) = 0.3073$$

となり，ポアソン分布のあてはめは棄却されない．

表の右端の「(–2) *対数尤度 n[-2ln(P)]」は，ポアソン分布の確率 P_i について対数を取り，$-2n_i$ を掛けた結果である．有害種子の数 $y_1 = 0$ のポアソン分布の確率 $P_1 = 0.0488$ なので，

$$\begin{aligned} n_1(-2)\ln L_1 &= -2n_1 \ln p_1 \\ &= (-2) \times 3 \times \ln(0.0488) \\ &= 18.1224 \end{aligned}$$

と計算されている．この欄の合計は，(–2) 倍の対数尤度 $\ln L$ は，

$$\begin{aligned} \ln L &= \sum_i n_i(-2)\ln L_i \\ &= 18.12 + 65.11 + \cdots + 11.75 \\ &= 381.9035 \end{aligned}$$

となっている．この「(–2) 倍の対数尤度」の和，または，単に「対数尤度」の和は，最尤法による各種の統計的方法の中心的な統計量で，最小 2 乗法での「偏差平方」の和と同等な役割を果す．表 1.5 に統計ソフト JMP の「一変量の分布」でポアソン分布をあてはめた結果の右の欄の「指標」の中の「(–2) *対数尤度」に 381.90 を見出すことができる．

表 1.5 には，JMP の「一変量の分布」での結果で，観測度数 n_i の縦棒グラフにポアソン分布の観測度数 n_i が上書きされている．「要約統計量の欄」に「平均＝3.0204」と「分散＝3.2779」が表示され，「Poissson 分布のあてはめ」欄に平均と同じ「尺度 λ=3.0204」が推定されている．なお，「尺度 λ」は，「位置 λ」と同義語的に使われている．

「適合度検定」の欄に，χ^2 :「X2＝105.2703」が計算され，自由度（98－1）のカイ 2 乗分布の上側確率が「Prob＝0.2659」と計算されている．Excel での適合度の計算は，出現度数に対する期待度数の関係から求めているのに対し，JMP の「一変量の分布」では，全 98 個の有害種子の数 Y に対して，Y の期待値である $\hat{\mu}$ との関係で，次のように計算されているからである．

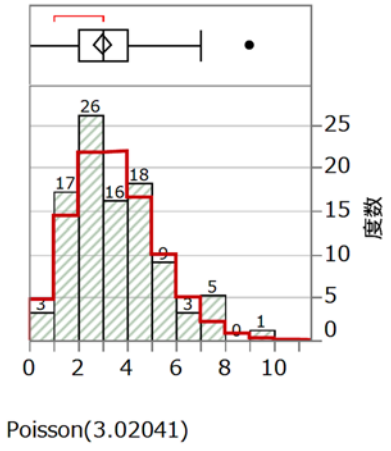
$$\text{Pearsonのカイ2乗} = \sum_i n_i \frac{(y_i - \mu)^2}{\mu}$$

$$= 3 \times \frac{(0 - 3.0204)^2}{3.0204} + 17 \times \frac{(1 - 3.0204)^2}{3.0204} + \cdots + 1 \times \frac{(9 - 3.0204)^2}{3.0204}$$

$$= 105.2703$$

$$p = 1 - \text{Chisq.dist}(105, 2703, 98\text{-}1, \text{true}) = 0.2659$$

表 1.5　JMP による有害雑草の種子の数に対するポアソン分布のあてはめ



要約統計量

平均	3.0204
標準偏差	1.8105
平均の標準誤差	0.1829
平均の上側95%	3.3834
平均の下側95%	2.6574
N	98.0000
分散	3.2779

Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値
尺度	λ	3.0204
指標		
(-2)*対数尤度		381.90
AICc		383.95
BIC		386.49

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
105.2703	0.2659

ここに示したように，統計ソフトの様々な出力結果を Excel で再現することは，様々な統計解析手法の計算理論について理解を深めることになる．その結果として，当該の統計的方法を適切に活用し，様々な応用ができるようになることが期待される．

1.4. 人工データ（恒等リンク，3 水準，回帰）

ドブソン著，田中・森川・山中・富田 訳（2008），「一般化線形モデル入門，原著 第 2 版」の第 4.4 節に「ポアソン反応変数に対する回帰分析の例」がある．この節では，一般化線形モデルの計算理論の理解を深めるために表 1.6 に示す人工データが例示され，反復重み付き回帰の行列計算のための式が示されている．さらに，反復過程での行列計算の結果の一部，回帰パラメータの収束過程が示されている．

表 1.6　ポアソン回帰の例（人工データ）

x	-1	-1	0	0	0	0	1	1	1
y	2	3	6	7	8	9	10	12	15

このデータを用いて，ドブソン(2008）で示されている反復重み付き回帰によるポアソン回帰を Excel の行列関数を用いて再現し，95%信頼区間の計算と結果の表示を追加する．

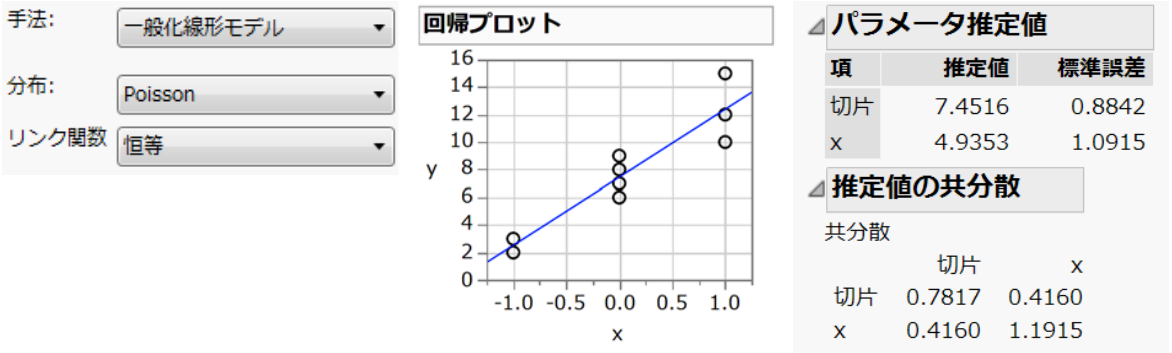
ポアソン回帰の適用

統計ソフト JMP の「モデルのあてはめ」から「一般化線形モデル」を選択し，「分布」で「Poisson」，「リンク関数」で「恒等」を選択し，実行した結果を表 1.7 に示す．回帰式は，

$$\hat{y}_i = 7.4513 + 4.9350 x_i$$

と推定されている．なお，「推定値の共分散」は，回帰直線の 95%信頼区間の計算のために必要となる．

表 1.7　JMP によるポアソン回帰の適用



注）　JMP では，ニュートン・ラフソン法を使用
（共分散行列の結果が反復重み付け回帰とは若干異なる）

ポアソン分布を仮定するカウント・データは，0 以上であるが，推定された直線を外挿すると推定値がマイナスになる場合が生ずる．これを防ぐために

$$y_i = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad \varepsilon_i \sim Poisson \text{ 分布} \tag{1.12}$$

のように，指数関数に線形式を含めてポアソン回帰を行い，推定値が 0 以下にならないようにすることが一般的であり，第 1.5 節で導入する．なお，一般化線形モデルでは，推定値に対して両辺に対数を取り，

$$\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (1.13)$$

$\ln \hat{y}_i$ についての線形式に対して反復重み付き回帰による最尤法が定式化されている．なお，JMP で「一般化線形モデル」を選び，「Poisson」を選ぶと「リンク関数」は自動的に「対数」となる．

ポアソン回帰の実際

統計ソフトでは，どのような計算が実際に行われているのだろうか．統計ソフトを使うユーザの立場であるならば，実際の計算がどのようなものか自ら計算を追試することは必要ないかもしれない．ポアソン回帰を説明する立場になった場合に，原理・原則を説明し，「実際の計算は統計ソフトを使えばいいのだ」と，言い切っていいのだろうか．通常の回帰分析の場合には，散布図上に回帰直線を描き，その 95%信頼区間を加え，さらに個別の 95%信頼区間（予測区間）を描くための数式は，ほとんどの教科書に示されている．統計ソフトでも標準的に出力されている．

ここに示した人工データで，ポアソン回帰直線の 95%信頼区間を描きたいのだけれども，どのようにしたらよいのだろうか．統計ソフトの外部出力の機能で 95%信頼区間の出力し，別途グラフ化することも可能ではある．Excel で奇麗な図を描きたいので，その計算式を知りたいが，どこに書いてあるのだろうか．ほとんどの統計の教科書の回帰分析の説明では，第 4.5 節で示すように偏差平方和を用いた計算式が示されているが，この方法ではポアソン回帰直線の 95%信頼区間の計算への応用はできない．ただし，推定された回帰パラメータの共分散行列を用いれば，通常の回帰分析でもポアソン回帰でも，本節の末尾の表 1.10 に示すように同じ考え方で求めることができる．

ポアソン回帰は，どのような計算方法で行なうのだろうか．ドブソン(2008）には，丁寧な記述がある．これに沿って，Excel のシート上で，行列関数を用いて再現してみよう．一般的な統計の教科書での回帰分析は，シグマを用いて記述されていて行列での記述を見いだすことはまれである．Excel の行列関数を用いることにより，煩雑なシグマを使った場合に比べ，すっきりとした回帰分析が可能となる．行列計算に不慣れであれば，この節は飛ばして，第 4 章の「デザイン行列を用いた回帰分析」，次いで第 5 章の「反復重み付き最尤法によるポアソン回帰」を先に読んでもらいたい．

反復重み付き回帰

ポアソン回帰のみならず一般化線形モデル全般の理解と応用には，対数尤度関数の知識が不可欠であり，さらにパラメータに関する偏微分を正確に求める計算力も必要である．最尤法は，推定したいパラメータについて対数尤度関数の 1 階の偏微分を行いベクトル化（スコアベクトル）し，さらに 2 階の偏微分を行い行列化（ヘッセ行列）し，逐次的な計算によってパラメータを推定（第 2 章で詳細を示す）する．反復重み付き回帰は，ヘッセ行列に変えて，反復重み付き回帰式

$$\hat{\boldsymbol{\beta}}^{(m)} = \left[(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{(m-1)} \right]^{-1} (\boldsymbol{X}^T \hat{\boldsymbol{W}} \hat{\boldsymbol{Z}})^{(m-1)} \tag{1.14}$$

の第 1 項を $\mathfrak{F}$

$$\mathfrak{F} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{(m-1)} \tag{1.15}$$

として，2 階の偏微分行列（ヘッセ行列）の代わりに使う．この方法は，ポアソン回帰を含む一般化線形モデルの各種のモデルに対する計算法として使われている．反復重み付き回帰を用いた最尤法についての詳細は，第 5 章で示すので，ここでは，Excel を使った反復計算の結果を表 1.8 に示す．

Excel による反復重み付き回帰の計算は，Excel の行列関数を活用している．通常の統計ソフトの回帰分析では，一般的に切片に対応する変数は省略している．ただし，行列計算では切片となる変数を明示する必要があり，それを 9×2 のデザイン行列 $\boldsymbol{X}$ とする．反応変数を 9×1 の $\boldsymbol{Y}$ ベクトルとし，回帰パラメータの初期値を 2×1 のベクトル $\hat{\boldsymbol{\beta}}^{(m-1)}$ として，推定値 $\hat{\boldsymbol{Y}}$ を Excel の行列の関数 Mmult() を使って

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m-1)} \ :=\text{Mmult}(\boldsymbol{X} \text{ の範囲}, \ \hat{\boldsymbol{\beta}}^{(m-1)} \text{ の範囲}) \tag{1.16}$$

のように計算する．なお，Excel の関数 Mmult() 内の引数 $\boldsymbol{X}$ は，（$\boldsymbol{X}$ の範囲）が存在するシートのセル範囲を選択して設定する．

デザイン行列の i 行目の行ベクトルを $\boldsymbol{x}_i = [x_{0,i}, x_{1,i}]$ とし，分布関数は，恒等リンク（何も変換しない）とする．一般化線形モデルの公式から

$$g(\mu_i) = \mu_i = \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^{(m-1)} = \eta_i \tag{1.17}$$

となる．重みは，恒等リンクなので，$\partial\mu_i / \partial\eta_i = 1$ となり，

$$\begin{aligned} w_i &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \frac{1}{\hat{\beta}_0 X_{0,i} + \hat{\beta}_1 X_{1,i}} = \frac{1}{\hat{Y}_i} \end{aligned} \tag{1.18}$$

分散は，ポアソン分布の分散は期待値に等しいので，$\hat{Y}_i$ となり，重みは分散の逆数となる．

リンク関数 $\eta_i^{(m)}$ は，一般化線形モデルの公式から，

$$\eta_i^{(m)} = \mu_i^{(m-1)} + (Y_i - \mu_i^{(m-1)})\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \qquad (1.19)$$
$$= Y_i$$

$\eta_i^{(m)} = Y_i$ となる．ここで，他のリンク関数への拡張も考慮して，行列計算の表記上

$$Z_i^{(m)} = \eta_i^{(m)} = Y_i$$

とする．

Excel による反復重み付き回帰

表 1.8 に示した反復重み付き回帰は，次式により段階的に計算する．なお，重みを分散の逆数とするのは，それぞれの X_i での Y_i の分散が明らかに異なる場合に対し分散を基準化するための基本的な方法である．なお，ここに示す行列計算は，第 5 章で詳しく説明する．

$(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(m-1)}$ ： = Mmult (Transpose ($\boldsymbol{X}$ の範囲 * $\hat{\boldsymbol{w}}$ の範囲), $\boldsymbol{X}$ の範囲)

$[(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(m-1)}]^{-1}$ ： = Minverse ($\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ の範囲)

表 1.8　Excel シート上での反復重み付き回帰による最尤法

	デザイン行列			回帰		重み付回帰		推定値
	$\boldsymbol{X}$		観測値	推定値$^{(m-1)}$	リンク関数	重み	推定値$^{(m)}$	差
i	x_0	x_1	y	y^	$z=\eta=y$	w^$=1/y$^	z^	y^ - z^
1	1	-1	2	2.0000	2.0	0.5000	2.5139	-0.5139
2	1	-1	3	2.0000	3.0	0.5000	2.5139	-0.5139
3	1	0	6	7.0000	6.0	0.1429	7.4514	-0.4514
4	1	0	7	7.0000	7.0	0.1429	7.4514	-0.4514
5	1	0	8	7.0000	8.0	0.1429	7.4514	-0.4514
6	1	0	9	7.0000	9.0	0.1429	7.4514	-0.4514
7	1	1	10	12.0000	10.0	0.0833	12.3889	-0.3889
8	1	1	12	12.0000	12.0	0.0833	12.3889	-0.3889
9	1	1	15	12.0000	15.0	0.0833	12.3889	-0.3889
			β_0^=	7.0000		β_0^=	**7.4514**	**1.4944E+06**
			β_1^=	5.0000		β_1^=	**4.9375**	平方和x10^6
				初期値 or $\boldsymbol{\beta}$^$^{(m-1)}$ 貼り付け			重み付き回帰係数 $\boldsymbol{\beta}$^$^{(m)}$	
		1.8214	-0.7500		0.7292	0.4375		9.8690
		-0.7500	1.2500		0.4375	1.0625		0.5833
		$\boldsymbol{X}^T\boldsymbol{W}$^$\boldsymbol{X}$	積和行列		$(\boldsymbol{X}^T\boldsymbol{W}$^$\boldsymbol{X})^{-1}$ 共分散行列			$\boldsymbol{X}^T\boldsymbol{W}$^$\boldsymbol{Z}$^

$(\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}})^{(m-1)}$ ： = Mmult (Transpose ($\boldsymbol{X}$ の範囲 * $\hat{\boldsymbol{w}}$ の範囲), $\hat{\boldsymbol{Z}}$ の範囲)

$\hat{\boldsymbol{\beta}}^{(m)} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}}$: = Mmult ($(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ の範囲, $\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}}$ の範囲)

$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)}$ ： = Mmult ($\boldsymbol{X}$ の範囲, $\hat{\boldsymbol{\beta}}^{(m)}$ の範囲)

ただし，ベクトル $\hat{\boldsymbol{w}}$ は， $\hat{\boldsymbol{W}}$ 行列の対角要素，* 演算子は，セル同士の積.

初期値としては，一般的に重みなしの回帰式

$$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 7.4545 \\ 4.9090 \end{bmatrix}$$

から得られる推定値を使うが，ドブソン（2008）に示されている

$$\hat{\boldsymbol{\beta}}^{(0)} = \begin{bmatrix} 7.0 \\ 5.0 \end{bmatrix}$$

を用いる．表 1.8 に示すように，初期値を入力すると，最初に $\hat{\boldsymbol{Y}} = (\boldsymbol{X}\,\hat{\boldsymbol{\beta}})^{(0)}$ が計算される．最初の $i=1$ の場合は， $\hat{y}_1 = 1\times 7.0 + (-1)\times 5.0 = 2.0$ となる．次に，重み $w_{ii} = 1/\hat{y}_i = 0.50$ が計算される．

第 1 回目の重み付き回帰のパラメータ $\hat{\boldsymbol{\beta}}^{(1)}$ は，次式のように推定されている．

$$\hat{\boldsymbol{\beta}}^{(1)} = \left[(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(0)}\right]^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}})^{(0)}$$

$$= \left\{\left[(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X}\right]^{(0)}\right\}^{-1}\left[(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\hat{\boldsymbol{Z}}\right]^{(0)}$$

$$= \begin{array}{|rr|} \hline 0.7292 & 0.4375 \\ 0.4375 & 1.0625 \\ \hline \end{array} \quad \begin{array}{|r|} \hline 9.8690 \\ 0.5833 \\ \hline \end{array}$$

$$= \begin{array}{|r|} \hline \mathbf{7.4514} \\ \mathbf{4.9375} \\ \hline \end{array}$$

ただし， $\boldsymbol{X} * \boldsymbol{w}$ は，行列 $\boldsymbol{X}$ とベクトル $\boldsymbol{w}$ のセル同士の掛け算である

この推定値 $\hat{\boldsymbol{\beta}}^{(1)}$ を用いて，推定値 $\hat{\boldsymbol{z}}^{(1)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(1)}$ が計算された結果が示されている．第 1 回目の推定値 $\hat{\boldsymbol{y}}^{(0)}$ との差 $\hat{\boldsymbol{y}}^{(0)} - \hat{\boldsymbol{z}}^{(1)}$ を計算し，平方和を次式で計算している．

$$S_e = \sum_{i=1}^{9}(\hat{y}_i - \hat{z}_i)^2 = 1.4944$$

なお，表には，収束の状況を把握しやすいように 10^6 倍で表示している．

1.4944E+06
平方和x10^6

反復計算

推定値 $\hat{y}_i$ と推定値 $\hat{z}_i$ の差の平方和を計算し，十分に小さくなるまで，推定値 $\boldsymbol{\beta}^{(m)}$ の結果をコピーし，“値のみ”を $\boldsymbol{\beta}^{(m-1)}$ にペーストする．これは，$\boldsymbol{\beta}^{(m)}$ には，計算式が埋め込まれているので，通常のペーストではセルの参照位置が異なりエラーとなるためである．表 1.9 に繰り返しコピー&ペーストした結果を示す．4 回目で平方和が 10^{-6} より小さくなったのでストップする．なお，手作業に換えソルバーで平方和を最小にするように $\boldsymbol{\beta}^{(m-1)}$ を変化させることにより解を得ることもできる．

表 1.9　反復重み付き回帰による逐次近似

反復 m		$(m\text{-}1)$	(m)	差	平方和x10^6	共分散行列	$(X^T WX)^{-1}$
1	β_0^=	7	7.451389	0.451389	1.494E+06	0.7292	0.4375
	β_1^=	5	4.937500	-0.062500		0.4375	1.0625
2	β_0^=	7.451389	7.451632	0.000243	1.581E+01		
	β_1^=	4.937500	4.935314	-0.002186		略	
3	β_0^=	7.451632	7.451633	0.000001	5.821E-04		
	β_1^=	4.935314	4.935300	-0.000013			
4	β_0^=	7.451633	7.451633	0.000000	2.142E-08	0.7817	0.4166
	β_1^=	4.935300	4.935300	0.000000		0.4166	1.1863

第 3 回目と第 4 回目の推定値の差は 10^{-6} 以下となり，差の平方和は，0.02142 と小さくなったので反復を中止し，解が求まったとする．さらに反復を行えば，差の平方和は，ゼロとなるが，ある一定の推定値の精度となれば，実用上は問題ないので，反復の中止基準をあらかじめ設定しておくのが，一般的である．

Excel の行列関数を用いた重み付き回帰分析を活用し，反復重み付き回帰による最尤法により解を求めたのであるが，基本は第 4 章に示す通常の回帰分析から初め，段階的に学習することが望ましい．第 5 章で，反復重み付き回帰の基礎について詳しく示す．

パラメータの共分散行列を用いた 95%信頼区間の計算

表 1.9 で計算されているパラメータの推定値および共分散行列を用いて回帰直線の 95%信頼区間を Excel で計算し図示する．ただし，反復重み付き回帰での結果であるため表 1.7 の JMP による共分散行列と若干異なる．

		パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})=(\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{X})^{-1}$					
	$\boldsymbol{\beta}\hat{}$		β_0^	β_1^			
β_0^=	7.4516	β_0^	0.7817	0.4166	=	$Var(\beta_0\hat{})$	$Cov(\beta_0\hat{}, \beta_1\hat{})$
β_1^=	4.9353	β_1^	0.4166	1.1863		$Cov(\beta_1\hat{}, \beta_0\hat{})$	$Var(\beta_1\hat{})$

推定値 $\hat{y}$ の分散 $Var(\hat{y})$ は，共分散行列の対角要素であり，回帰パラメータ $\hat{\beta}_0$ の分散 $Var(\hat{\beta}_0)=0.7817$，$\hat{\beta}_1$ の分散 $Var(\hat{\beta}_1)=1.1863$，$\hat{\beta}_0$ と $\hat{\beta}_1$ の共分散 $Cov(\hat{\beta}_0,\hat{\beta}_1)=0.4166$ を用いて，合成分散の一般式，

$$\begin{aligned} Var(\hat{y}) &= Var(\hat{\beta}_0+\hat{\beta}_1 x) \\ &= Var(\hat{\beta}_0)+2Cov(\hat{\beta}_0,\hat{\beta}_1)x+Var(\hat{\beta}_1)x^2 \\ \mathrm{L95\%} &= \hat{y}-1.96\sqrt{Var(\hat{y})} \\ \mathrm{U95\%} &= \hat{y}+1.96\sqrt{Var(\hat{y})} \end{aligned}$$

によって求められる．表 1.9 の $x_1=-2$ の場合については，

$$\hat{y}=7.4516+4.9353\times(-2)=-2.4190$$

$$Var(\hat{y})=0.7817+2\times0.4166\times(-2)+1.1863\times(-2)^2=3.8607$$

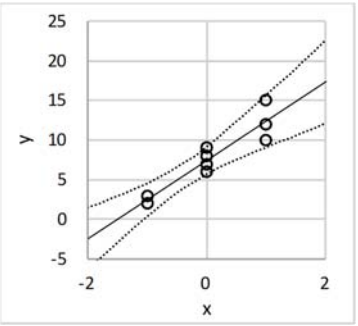
$$\mathrm{L95\%}=-2.4189-1.96\sqrt{3.8607}=-6.2702$$

$$\mathrm{U95\%}=-2.4189+1.96\sqrt{3.8607}=1.4322$$

であり，他の $x_i=(-1,0,1,2)$ についても同様に計算した結果を表 1.10 に示す．散布図上には，この結果に基づきポアソン回帰の回帰の 95%信頼区間を上書きしてある．通常の回帰分析とは異なり，x の大きい方が小さい方に比べて広がりが大きくなっている．

表 1.10　ポアソン回帰直線の 95%信頼区間

x_0	x_1	y^	$Ver(y$^$)$	SE	$L95\%$	$U95\%$
1	-2	-2.4190	3.8607	1.9649	-6.2701	1.4322
1	-1	2.5163	1.1349	1.0653	0.4283	4.6043
1	0	7.4516	0.7817	0.8841	5.7188	9.1845
1	1	12.3869	2.8011	1.6736	9.1066	15.6673
1	2	17.3222	7.1931	2.6820	12.0655	22.5790



実際の計算では，共分散行列が計算されているので，デザイン行列 $\boldsymbol{X}$ の $\boldsymbol{x}_i=(x_{0,i}\ x_{1,i})$ を行ベクトルとした場合に，次の 2 次形式で求められる．

$$\begin{aligned} Var(\hat{y}) &= \boldsymbol{x}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{x}^T \\ &= \mathrm{Mmult}\ (\ \mathrm{Mmult}\ (\ \boldsymbol{x}_i \text{の範囲},\ (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} \text{の範囲}, \mathrm{Transpose}(\boldsymbol{x}_i \text{の範囲})\ ) \end{aligned}$$

この，2 次形式での分散の計算方法の応用については，第 12 章に詳しく説明しているので参照されたい．ここで示した反復重み付き回帰の方法は，分布が 2 項分布である場合の「ロジット・リンク」の場合の「プロビット・リンク」の場合，補 2 重対数リンクの場合も同様であり，第 5.4 節で取り上げる．

1.5. 冠動脈心疾患の死亡者数（対数リンク，8 水準，オフセット，回帰）

ドブソン(2008）の第 3.5 節にオーストラリアのある地方の年齢 5 歳階級での冠動脈心疾患による死亡者数が示されている．これまでに示した事例は，事象の発現数のみが観測された場合であるが，稀な現象であっても表 1.11 に示すように対象とする部分母集団の大きさ n_i が人口統計学的に得られる場合もある．部分母集団の人数は十分大きく，極端な差がないので，死亡者数 y についてポアソン分布のあてはめが可能ではあるが，分母の大きさも考慮したい．

表 1.11　オーストラリアのある地方の冠動脈心疾患による死亡者数

	年齢層	死亡者数	母集団	死亡率	10万人比
No.	x	y	人数 n	%	人数
1	30	1	17,742	0.0056	5.6
2	35	5	16,554	0.0302	30.2
3	40	5	16,059	0.0311	31.1
4	45	12	13,083	0.0917	91.7
5	50	25	10,784	0.2318	231.8
6	55	38	9,645	0.3940	394.0
7	60	54	10,706	0.5044	504.4
8	65	65	9,933	0.6544	654.4
	全体	205	104,506	0.1962	196.2

年齢層は，30-34, 35-39 のように与えられている．

各年齢層の母集団の人数が分かっているので，死亡率を算出すると全体で 0.19%と小さな値となるので，1 万人あるいは 10 万人あたり（以下，10 万人比とする）の死亡者数にした方が理解しやすい．全体の人数が 104,506 人なので，10 万人比に換算すると 196.2 人となる．各年齢層では，50～54 歳が，231 人と平均的である．図 1.2 右に示すように年齢が高くなるにつれて指数関数的に増加する．図 1.2 左の直線のあてはめは，対数変換した 10 万人比に対して

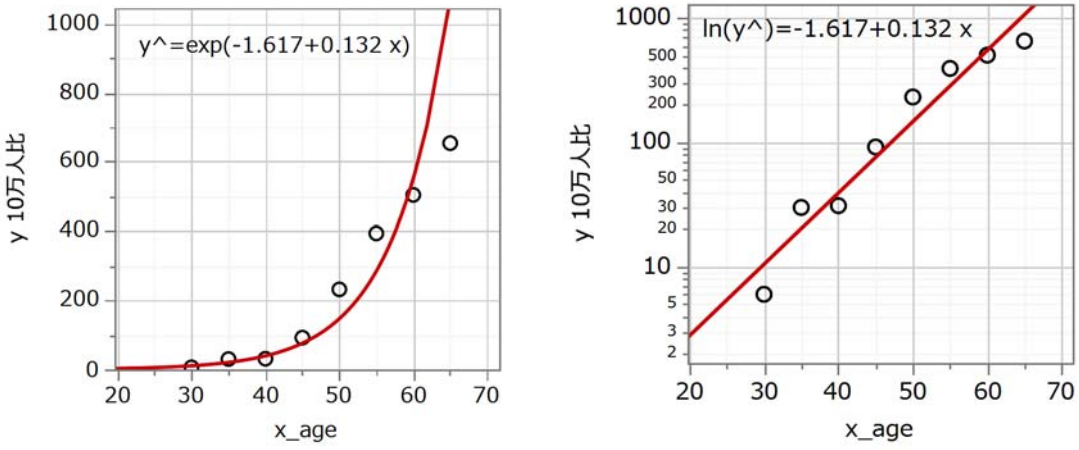


図 1.2　10 万人あたりの死亡者数の対数に対する便宜的な直線のあてはめ

通常の回帰分析をした結果であり，図 1.2 左は，単に元の 10 万人比に戻しただけで，ポアソン回帰の結果ではない．年齢が増加するにつれて頭打ちになる傾向があり，対数変換をして直線をあてはめることには無理そうであり，2 次式のあてはめも考慮する必要がある

ポアソン分布は，平均と分散が等しいので，説明変数 X の増加に伴い反応変数 Y の平均も直線的に増大するような場合に，通常の回帰分析で仮定する等分散性が成り立たない．さて，反応変数 y に対数を取った場合に，等分散性が成り立つのであろうか．説明変数 x が小さい場合に分散が相対的に大きくなってしましい，残念ながら等分散性が成り立たない．

死亡率が求められているので，ポアソン回帰ではなく 2 値データとしての一般化線形モデルの適用が可能である．誤差分布を 2 項分布，リンク関数に（プロビット or ロジット or 補 2 重対数）を選択することも可能である．リンク関数にロジットを選択した場合には，良く知られたロジスティック回帰と同じ結果となる．ただし，2 値データとしてロジスティック回帰を適用することには，疑問が残る．第 1 は，死亡の原因の一つである冠動脈心疾患の死亡率に対して上限の死亡率を 100%とするシグモイド曲線を仮定すること，第 2 に，1 パーセントに満たない死亡率について推定結果が得られても，そもそも解析する意義があるのか，結果の表示も極めて小さなパーセント表示では，妥当性に欠けると思われる．なお，第 5.4 節では，誤差分布を 2 項分布とした場合について示す．

分母 n_i が分かっている場合のポアソン回帰を次のように指数関数を使って定義し，

$$y_i = n_i \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad \varepsilon_i \sim \text{ポアソン分布}$$

推定値の形式にし，両辺に対数を取ると，

$$\ln(\hat{y}_i) = \ln n_i + \hat{\beta}_0 + \hat{\beta}_1 x_i$$

が得られる．この式で「$\ln n_i$」は，オフセットと言われている．この式で，未知パラメータは，切片 β_0 と傾き β_1 である．推定しようとしている切片 β_0 は，実際には，$(\ln n_i + \beta_0)$ なので β_0 から $\ln n_i$ だけ大きい方にずらした（オフセット）た点と解釈される．オフセットがゼロとなるのは，$n_i = 1$ の場合であり，切片 β_0 は，位置パラメータを $n_i = 1$ とした場合の切片である．

JMP でのオフセットの設定について表 1.12 に例示する．オフセットに設定するのは，元の分母 n_i ではなく，自然対数をあらかじめ計算して与える．

JMP を用いたオフセットを含んだ予測式は，JMP の計算式の出力で

$$\text{Exp}\left(-11.62782676 + 0.1044379989 \bullet \text{x\_age} + \text{ln\_n}\right)$$

表 1.12　オフセット値を用いたポアソン回帰でのパラメータ推定

となる．一般の式に変換すれば，

$$\hat{y}_i = \exp(-11.6278 \ + \ 0.1044x_i + \ln n_i)$$
$$= n_i \exp(-11.6278 \ + \ 0.1044x_i)$$

となる．推定されたパラメータを用い，オフセットを含んだ予測値を計算した結果を表 1.13 に示す．Excel による解析方法は，第 2.6 節，第 5.3 節で詳細に示す．第 5.4 節では，Excel の反復重み付き回帰による，2 値データの解析についても例示する．

表 1.13　オフセット含んだ予測式

				$\beta\hat{}_0$=	**-11.6279**	
				$\beta\hat{}_1$=	**0.1044**	
	年齢層	死亡者数	母集団	オフセット		推定値
No.	x	y	人数 n	$\ln n$	$\beta\hat{}_0$	$y\hat{}$
1	30	1	17,742	9.7837	-1.8442	3.6
2	35	5	16,554	9.7144	-1.9135	5.7
3	40	5	16,059	9.6840	-1.9438	9.3
4	45	12	13,083	9.4791	-2.1488	12.8
5	50	25	10,784	9.2858	-2.3420	17.8
6	55	38	9,645	9.1742	-2.4537	26.9
7	60	54	10,706	9.2786	-2.3493	50.2
8	65	65	9,933	9.2036	-2.4242	78.6
オフセット	65	65	**1**	0.0000	-11.6279	**0.00791**

オフセットを 1 にした 65 歳の場合の推定値は，

$$\hat{y}_i = 1 \times \exp(-11.6279 \ + \ 0.1044 \times 65) = 0.00791 \text{ 人}$$

であり，9,933 人当たりでは，78.6 人となる．このように，オフセットを考慮した推定値は，年齢階層別の母集団の人数を考慮した推定値となっている．

2 値データとして解析する際には，反応変数を死亡の場合を $y_i' = 0$ とし，その人数を $n_i' = y_i$ とする．生存を $y_i' = 1$ とし，その人数を $n_i' = n_i - y_i$ とする．JMP の一般化線形モデルで，「分布」を「二項」とし，「リンク関数」をロジットとし，人数 n' を「度数」とする．

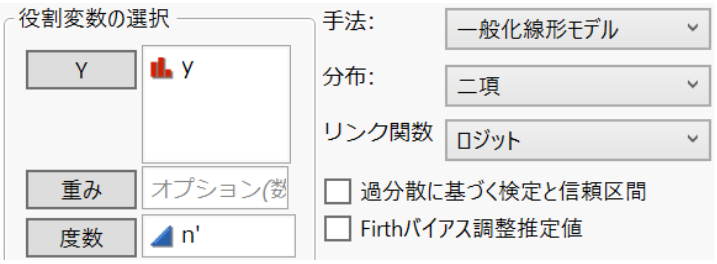


表 1.14 に示すような，推定結果を得る．ポアソン回帰での推定値とほぼ同様の結果が得られる．ただし，各種の出力結果は，死亡率のままなので，使い勝手はオフセットありのポアソン回帰に比べて良くない．

表 1.14　ロジスティック回帰を適用した場合の推定値

項	推定値	標準誤差
切片	-11.6395	0.4538
x	0.1047	0.0078

この式を用いた死亡率に対するロジスティック曲線は，図 1.3 に示すように 160 歳で 100%になるようなシグモイド曲線となる．

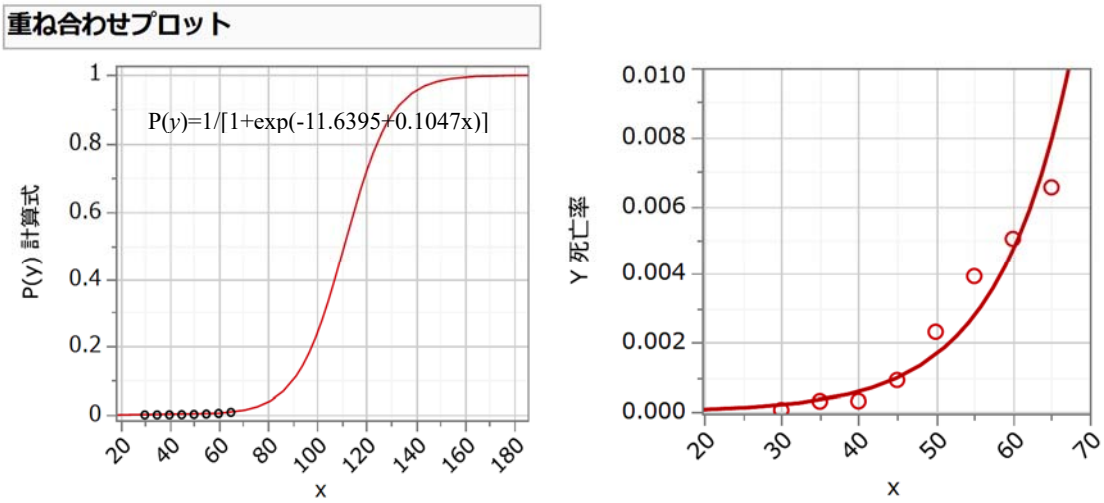


図 1.3 死亡率に対するロジスティック曲線のあてはめ

なお，第 2.6 節では，シグモイド曲線の上限もパラメータ化する方法を示す．

1.6. 満月と新月の日の犯罪件数に対する尤度比検定（2 群）

アルトマン著，木船・佐久間 訳（1999），「医学研究における実用統計学」の第 4.8 節に「満月と新月の日の犯罪件数」についての比較データが示されている．表 1.15 に示すように，このデータは，インドの 3 地域の 1978 年から 1982 年の間の満月と新月の日の一日当たりの犯罪件数である．満月と新月の日のどちら日に犯罪が多いのかを比較することが目的である．

表 1.15　インドの 3 地域の 1978 年から 1982 年の間の 1 日当たりの犯罪件数

	満月の日 x=0				新月の日 x=1			
犯罪件数	観測度数	ポアソン	期待値	残差	観測度数	ポアソン	期待値	残差
y	n	P	n^	n-n^	n	P	n^	n-n^
0	40	0.247	45.2	-5.2	114	0.603	112.2	1.8
1	64	0.345	63.2	0.8	56	0.305	56.7	-0.7
2	56	0.242	44.2	11.8	11	0.077	14.3	-3.3
3	19	0.113	20.6	-1.6	4	0.013	2.4	1.6
4	1	0.039	7.2	-6.2	1	0.002	0.3	0.7
5	2	0.011	2.0	0.0	0	0.000	0.0	0.0
6	0	0.003	0.5	-0.5	0	0.000	0.0	0.0
7	0	0.001	0.1	-0.1	0	0.000	0.0	0.0
8	0	0.000	0.0	0.0	0	0.000	0.0	0.0
9	1	0.000	0.0	1.0	0	0.000	0.0	0.0
合計	183		183.0		186		186.0	
平均	1.3989				0.5054			
分散	1.3620				0.5648			
分散/平均	0.9736				1.1177			

なお，平均と分散の比は，それぞれ 0.9736，1.1177 であり，ポアソン分布を仮定することが可能である．ポアソン分布の確率は，満月の日の $y=1$ の場合であれば，Excel の関数で

$$\text{Poisson.dist}(y,\ \text{平均}, \mathit{false}) = \text{Poisson.dist}(1,\ 1.3939, \mathit{false}) = 0.3453$$

で計算され，期待値 $\hat{n}_{y=1}$ は，出現総数は，$N=183$ なので，

$$\hat{n}_{y=1} = N\,p_{y=1} = 183 \times 0.3453 = 63.2$$

となる．出現数は $n_{y=1}=64$ で，その差は，0.8 件である．図 1.4 には，ボックス・プロットと

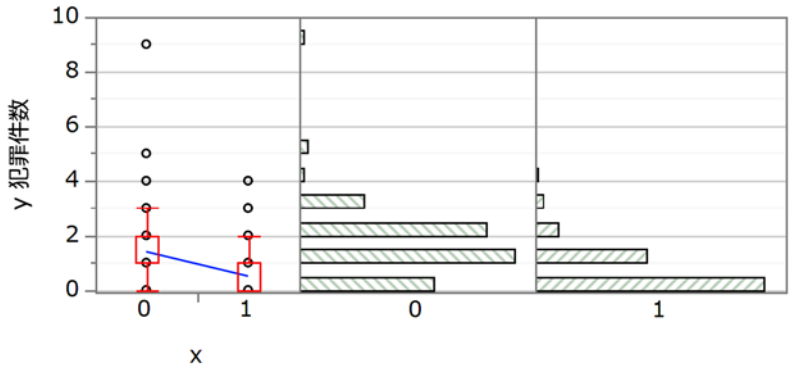


図 1.4　満月（x=0）と新月（x=1）の犯罪件数の比較

棒グラフの組み合わせた結果を示す．この図は，JMP の「二変量の関係」の「表示オプション」で「ヒストグラム」の選択で作成することができる．

順位和検定

満月の夜の平均犯罪件数は，1.3989 件，新月の夜は，0.5054 件となっていて，満月の夜は，2.8 倍の犯罪件数となっている．稀な現象の場合に 0 件の頻度が多くなるので，犯罪が起きたか否かで 2×2 の分割表にまとめて，カイ 2 乗検定を行えば簡単な検定で済ますこともできる．犯罪件数を順序尺度と見なして，表 1.16 に示すように順位和検定を便宜的に行うことも間違いではない．ただし，有意な差が有るか否かの単純な結果しか得られない．統計量としての「スコア平均」が，出力されているが，順位データの平均であり，これによる考察は無意味である．

表 1.16　JMP の「二変量あてはめ」による順位和検定の便宜的な適用

Wilcoxon/Kruskal-Wallisの検定(順位和)

水準	度数	スコア和	スコアの期待値	スコア平均
1:満月	183	42252.5	33855.0	230.888
2:新月	186	26012.5	34410.0	139.852

2標本検定(正規近似)

S	Z	p値(Prob>|Z|)
42252.5	8.70452	<.0001*

JMP のポアソン回帰による 2 群間比較

反応変数がポアソン分布に従うのならば，その特性を生かしたポアソン回帰を統計ソフトで行うことも容易である．しかし，適用事例が身近にある教科書になければ，どのように適用したらよいのか，結果の解釈はどうしたらよいのか迷いが生じ，使用経験がある統計手法で済ませたくなってしまう．なお，2 値反応にした場合の Pearson のカイ 2 乗検定，および，尤度比検定については，第 3.1 で詳細に示す．

反応変数がポアソン分布に従うことが確認できれば，満月を 0，新月を 1 とする目的変数を x として，ポアソン回帰により 2 群間の比較は容易にでき JMP および無償で使える OnDemand SAS を使った結果を表 1.20 に示す．

表 1.18 は JMP の一般化線形モデルのポアソン回帰を用いた結果である．切片は，0 を与えた満月の日の犯罪件数の推定値 1.3989 となり，新月の日に 1 を与えたので “x” の推定値が，満月と新月の平均値差（傾き）が -0.8935 なので，新月の日の犯罪の平均は，1.3989-0.8935＝0.5054 と推定される．この傾き−0.8935 についての検定は，尤度比カイ 2 乗検定を用いて p <0.0001 となり，95%信頼区間は，プロファイル尤度（正確な信頼区間の算出法）により（−1.0968～−0.6969）と計算されている．

表 1.17　JMP のポアソン回帰モデル実行のためのデータセット

	満月_新月	x	y 犯罪件数	n
1	1:満月	0	0	40
2	1:満月	0	1	64
3	1:満月	0	2	56
4	1:満月	0	3	19
5	1:満月	0	4	1
6	1:満月	0	5	2
7	1:満月	0	9	1
8	2:新月	1	0	114
9	2:新月	1	1	56
10	2:新月	1	2	11
11	2:新月	1	3	4
12	2:新月	1	4	1

表 1.18　JMP のポアソン回帰を用いた 2 群間の尤度比による比較

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)	下側信頼限界	上側信頼限界
切片	1.3989	0.0874	256.0000	<.0001*	1.2345	1.5773
x	-0.8935	0.1018	80.5774	<.0001*	-1.0968	-0.6969

説明変数が 2 水準なのでリンク関数は恒等でも対数でも同じ結果となる．

最尤法を用いた 2 群間の比較をポアソン回帰により実施し，表 1.18 に示したように，満月と新月の犯罪件数の検定統計量として尤度比カイ 2 乗値が 80.58 となっている．もちろん p<0.0001 と有意な差である．この尤度比検定は，満月と新月の犯罪件数の差についての対数用尤度のマイナス 2 倍の対数尤度が，自由度 1 のカイ 2 乗分布に従うとした場合である．

Excel による 2 群間の尤度比検定

表 1.19 に示すように Excel の計算シートで，実際に確認してみる．満月と新月の犯罪件数を加えた場合の平均は，$\mu_{0+1} = 0.9485$ となり，$y_1 = 0$ の場合のポアソン確率は，

$$P_{0+1,1} = \text{Poisson.dist}(0,\ 0.9485,\ false) = 0.3873$$

と計算されている．対数尤度は，

$$\begin{aligned} \ln L_{0+1,1} &= n_{0+1,1} \times \ln(0.3873) \\ &= 154 \times (-0.9485) \\ &= -146.0705 \end{aligned}$$

であり，全体の対数尤度は，

$$\begin{aligned}\ln L_{0+1} &= \sum_{i=1}^{10} n_{0+1,i}\,\mathrm{Poisson.dist}(y_i, \mu_{0+1}, \mathit{false}) \\ &= -146.0705 - 120.1648 - \cdots - 14.2261 \\ &= -484.8865\end{aligned}$$

となる．満月と新月それぞれ，

$$\ln L_0 = \sum_{i=1}^{10} n_{0,i}\,\mathrm{Poisson.dist}(y_i, \mu_0, \mathit{false}) = -268.4776$$

$$\ln L_1 = \sum_{i=1}^{10} n_{1,i}\,\mathrm{Poisson.dist}(y_i, \mu_1, \mathit{false}) = -176.1202$$

なので，

$$\begin{aligned}\chi^2 &= (-2\ln L_{0+1}) - [-2(\ln L_0 + \ln L_1)] \\ &= [-2\times(-484.8865)] - [2\times(-268.4776 - 176.1202)] \\ &= 80.5774\end{aligned}$$

と計算され，表 1.18 の JMP の尤度比カイ 2 乗に一致する．

表 1.19　Excel によるポアソン回帰を用いた尤度比検定

		満月の日 x=0			新月の日 x=1			満月＋新月		
i	y	n_0	P	$\ln L_0$	n_1	P_1	$\ln L_1$	n_{0+1}	P_{0+1}	$\ln L_{0+1}$
1	0	40	0.2469	-55.9563	114	0.6033	-57.6129	154	0.3873	-146.0705
2	1	64	0.3453	-68.0458	56	0.3049	-66.5184	120	0.3674	-120.1648
3	2	56	0.2416	-79.5576	11	0.0770	-28.1977	67	0.1742	-117.0747
4	3	19	0.1126	-41.4883	4	0.0130	-17.3780	23	0.0551	-66.6738
5	4	1	0.0394	-3.2342	1	0.0016	-6.4132	2	0.0131	-8.6760
6	5	2	0.0110	-9.0159	0	0.0002	0.0000	2	0.0025	-12.0006
7	6	0	0.0026	0.0000	0	0.0000	0.0000	0	0.0004	0.0000
8	7	0	0.0005	0.0000	0	0.0000	0.0000	0	0.0001	0.0000
9	8	0	0.0001	0.0000	0	0.0000	0.0000	0	0.0000	0.0000
10	9	1	0.0000	-11.1795	0	0.0000	0.0000	1	0.0000	-14.2261
合計		183		**-268.4776**	186		**-176.1202**	369		**-484.8865**
平均		1.3989			0.5054			0.9485		
分散		1.3620			0.5648			1.1577		
分散/平均		0.9736			1.1177			1.2205		
						合計	**-444.5978**		(-2)x差	**80.5774**

SAS/GENMOD のポアソン回帰による 2 群間比較

表 1.20 に SAS/GENMOD プロシジャを用いた結果を示す．JMP と同様に Intercept が $x=0$ を与えた満月の日の犯罪件数の推定値 $\hat{\beta}_0 = 1.3989$ となり，$\hat{\beta}_1 = -0.8935$ が $x=1$ を与えた新月の日と新月の日の犯罪件数の差の推定値になっている．傾きについての検定は，JMP と異なり

Wald カイ 2 乗=77.06 となり，カイ 2 分布を用いて p<0.0001 となり，95%信頼区間は（−1.0930 ～−0.6940）と計算されている．

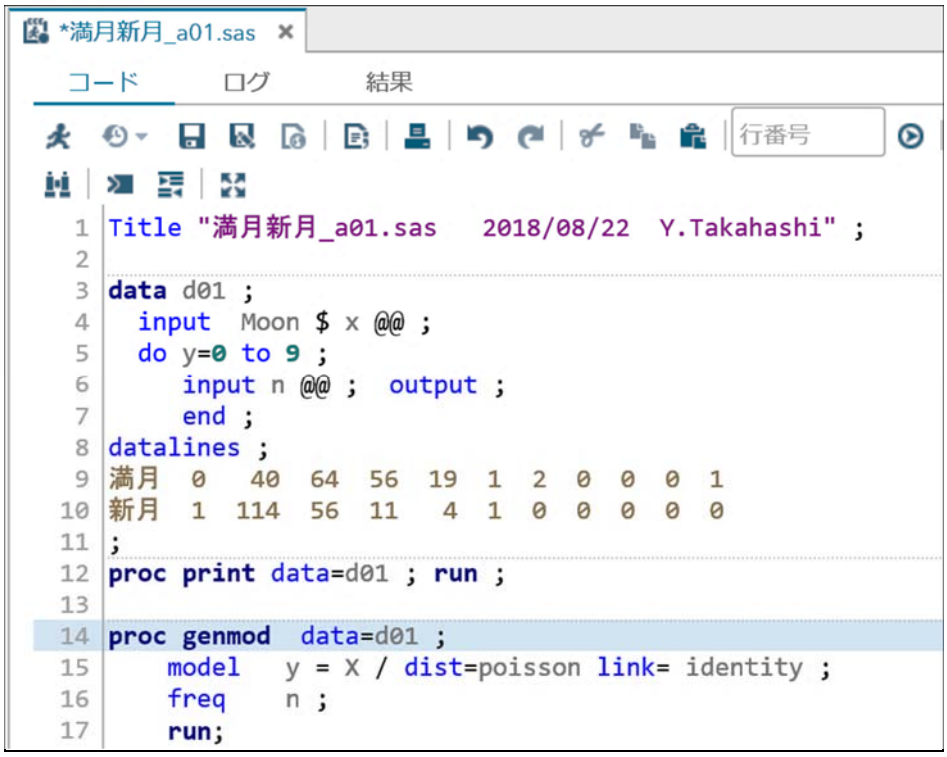
ポアソン回帰を含む一般化線形モデルの計算方法は，2 通りある．JMP の場合は，第 2 章で示すように対数尤度関数をパラメータに関して偏微分した式を用いた方法である．他方，SAS/GENMOD の場合は，第 1.4 節で示した反復重み付き回帰によって計算している．どちらの方法でも推定値は一致するが，共分散行列には，違いがわずかに生ずる．

また，JMP では，カイ 2 乗統計量もマイナス 2 倍の対数尤度の差から求めた尤度比カイ 2 乗を標準的に計算するが，SAS/GENMOD では，推定値を標準誤差（*SE*）で割って 2 乗した，いわゆる Wald カイ 2 乗を計算している．

表 1.20　SAS/GENMOD によるポアソン回帰を用いた 2 群間の比較

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr>ChiSq
Intercept	1	1.3989	0.0874	1.2275	1.5703	256.00	<.0001
x	1	−0.8935	0.1018	−1.0930	−0.6940	77.06	<.0001
尺度	0	1.0000	0.0000	1.0000	1.0000		

表 1.21　SAS/GENMOD 実行のための無償版 SAS の画面



1.7. 細菌を用いた試験データ（2×2 要因配置）

吉村・大橋 責任編集（1992)，「毒性試験データの統計解析」の第 3.3.2 節に復帰突然変異試験（Ames 試験，試験法を確立した人の名前）に関するデータが示されている．TA1537 ネズミチフス菌株を用いた Ames 試験の陰性対照群を 50 枚のシャーレで繰り返して異常が起きたコロニーをカウントした結果である．溶媒として蒸留水 or DMSO（ジメチル スルホキシド，Dimethyl sulfoxide)，代謝活性化の（なし or あり）有無を組み合わせた 4 通について 50 枚分シャーレ上で観察されたコロニー数が示されている．表 1.22 に 4 通りの場合について，コロニー数に関する度数分布，および，各種の統計量を計算した結果を示す．

表 1.22 ネズミチフス菌株に関するコロニー数

溶媒	蒸留水		DMOS	
代謝活性化	－なし	＋あり	－なし	＋あり
群，*n*，平均	1，50，14.54	2，50，7.54	3，50，12.48	4，50，8.28
平方和，分散	844.42，17.23	310.42，6.34	570.48，11.64	298.08，6.08
分散/平均	**1.19**	**0.84**	**0.93**	**0.73**
カイ2乗値, *p* 値	58.08，**0.176**	41.17，**0.779**	45.71，**0.607**	36.00，**0.917**
コロニー数 3	0	1	0	0
4	1	5	0	2
5	0	4	1	3
6	0	10	1	5
7	2	7	2	12
8	1	6	2	10
9	2	5	3	4
10	3	6	4	6
11	5	3	7	2
12	3	1	7	2
13	2	1	5	1
14	3	1	5	3
15	3	0	1	0
16	6	0	6	0
17	6	0	2	0
18	4	0	2	0
19	5	0	1	0
20	1	0	1	0
21	2	0	0	0
22	1	0	0	0

ポアソン分布のあてはめ

それぞれの分散を平均で割った比は，それぞれ，（1.19，0.84，0.93，0.73）となり，平均と分散が等しいポアソン分布と見なすことができそうである．「カイ 2 乗値，*p* 値」は，文献で示されている結果であるが，表 1.5「JMP による有害雑草の種子の数のポアソン分布のあては

め」で示した JMP の適合度検定の Pearson のカイ 2 乗検定の結果に一致する．この p 値は，（0.176，0.779，0.607，0.917）であり，ポアソン分布のあてはめは棄却されない．なお，吉村ら（1992）では，個別データが示されているが，ここでは，度数表の形でまとめ直している．

図 1.5 に JMP の「一変量の分布」を用いてヒストグラム上にポアソン分布を重ね書きした結果を示す．

群 1（蒸留水，代謝活性化－）の平均は 14.54，分散は 17.23，その比は 1.19 と 1 より大きく，分布に二つの山が見受けられるが，適合度の p 値は 0.176 であり，ポアソン分布のあてはめは棄却されない．

群 2（蒸留水，代謝活性化＋）の平均は 7.45，分散は 6.34，その比は 0.84 と 1 より小さく，右に裾を引く分布であり，適合度の p 値は 0.779 であり典型的なポアソン分布の形状である．

群 3（DMOS，代謝活性化－）の平均は 12.48，分散は 11.64 と同程度であり，その比は 0.93 と 1 より小さいが，適合度の p 値は 0.607 でありポアソン分布の前提がほぼ満たされている．

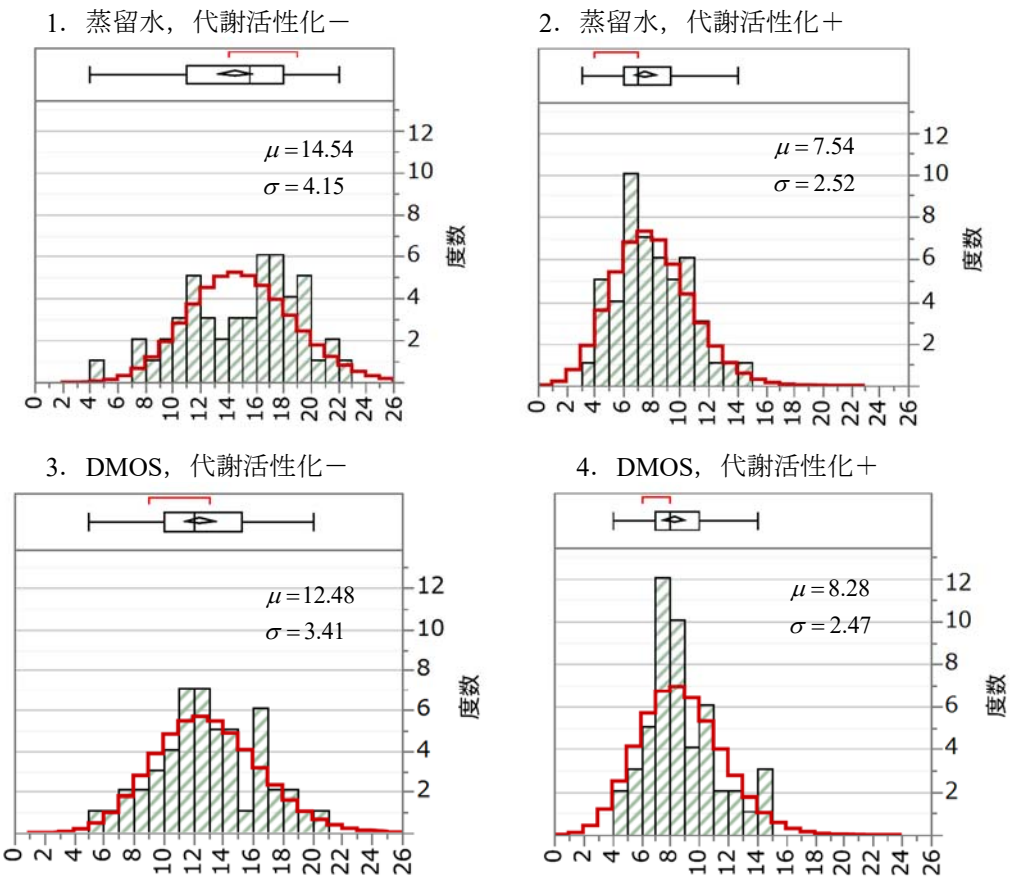


図 1.5　ネズミチフス菌株に関するコロニー数分布に対するポアソン分布のあてはめ

群 4（DMOS，代謝活性化＋）の平均は 8.28，分散は 6.08 と同程度であり，その比は 0.73 と 1 より小さいが，適合度の p 値は 0.917 でありポアソン分布の前提がほぼ満たされている．

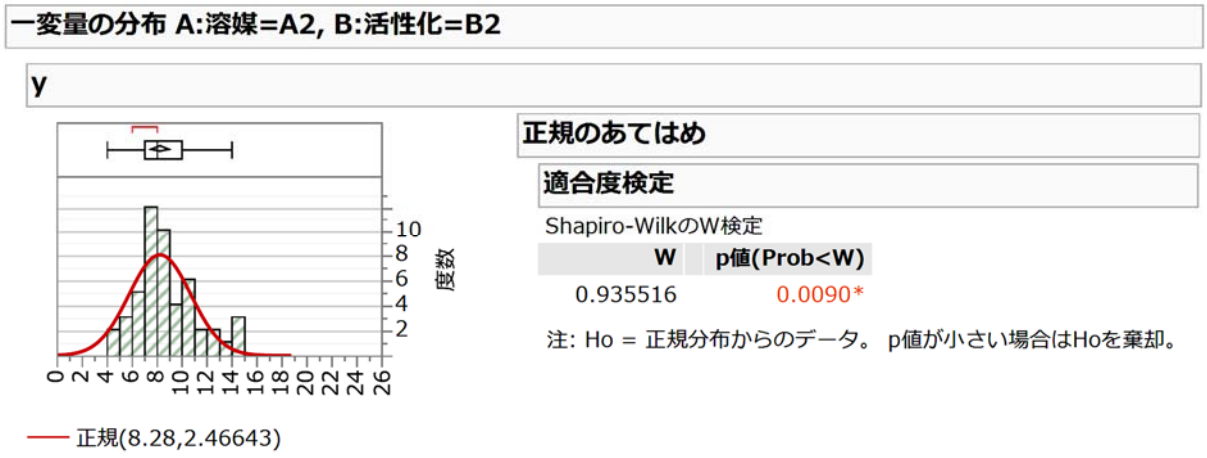
正規分布のあてはめ

図 1.5 のみを見た場合には，正規分布を仮定することが可能のようにも思われるかもしれない．それぞれの群について正規性の Shapiro-Wilk の W 検定を行った結果を表 1.23 に示す．ポアソン分布の適合度の検定で第 4 群が $p = 0.917$ と限りなくポアソン分布に適合している場合に正規性が $p = 0.0090$ と疑われる結果となっている．他の群は，正規分布のあてはめもポアソン分布のあてはめも棄却できないとの玉虫色の結果となっている．

表 1.23　正規性の検定（Shapiro-Wilk の W 検定）

群	1	2	3	4
W 値	0.9687	0.9853	0.9652	0.9355
p 値	0.2054	0.7835	0.1472	0.0090
*				**

表 1.24　群 4 に対する正規性の適合度検定



等分散性の検定

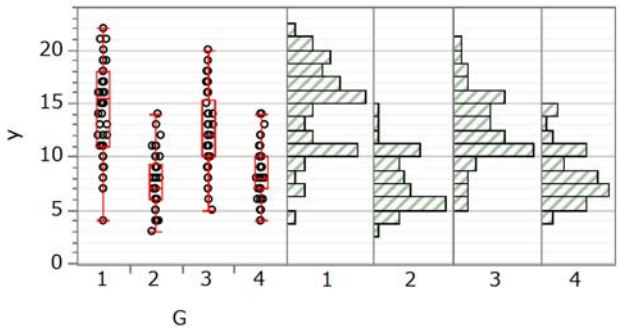
これら 4 群を典型的な 2×2 の要因配置実験と見なし，比較の前提となる等分散性について検討する．Bartlett の検定結果は，表 1.25 に示すように $F_3^1 = 6.2279$，$p = 0.0003$ であり，等分散性を仮定することはできない．

観測されたデータがどの様な分布になるかを見極めることは，統計解析の第一歩なのだが，ここに示したように，実際に観測したデータを基にした適合度の検定による推測は，玉虫色

の結果で歯切れがわるい．検定ベースではなく対数尤度をベースにした推測が，この問題に新たな光明を与える．詳しくは，第 3 章で示す．

表 1.25　Bartlett の検定

検定	F値	p値(Prob>F)
O'Brien[.5]	6.7607	0.0002*
Brown-Forsythe	6.4493	0.0003*
Levene	7.6019	<.0001*
Bartlett	6.2279	0.0003*



この実験は，被験物質に変異原性があるかの判定のために，各種の溶媒対照群の分布特性について検討する目的で行われたもので，最適条件を見つけるためのものではないが，分布がポアソン分布に従っている場合に，どのような統計解析が良いかを考えるためのデータとしても適している．

表 1.26 に各実験条件について，平均と分散を併記し，実験条件間の平均と差について示した．代謝活性剤が（1:なし）の場合に，（1:蒸留水）に対して（2:DMOS）は，コロニー数が 2.06 減少している．代謝活性剤が（2:あり）の場合には，（1:なし）の場合に比べてコロニー数が半減し，（1:蒸留水）に対して（2:DMOS）は，コロニー数が 0.76 増加している．分散も，平均値の増減を反映していることも確認される．これらの検討から，この試験系で得られるデータは，ポアソン分布に従うことが確認できる．

表 1.26　二元表による比較

	B:代謝活性化					
	1:なし		2:あり		全体	
A:溶媒	平均	分散	平均	分散	平均の平均	差
1:蒸留水	14.54	17.23	7.54	6.34	11.04	-7.00
2:DMOS	12.48	11.64	8.28	6.80	10.38	-4.20
平均の平均	13.51		7.91			
差	-2.06		0.74			

1.8. 細菌を用いた用量反応試験（恒等リンク，2 群，8 水準，効力比）

富山・杉本（2004）の「細菌を用いた用量反応試験データ」を用い，ポアソン回帰を適用する事例として取り上げる．表 1.27 に示すデータは，第 1.7 節と同様の Ames 試験の結果であり，陽性対照薬に対する代替物質の減弱の程度を確認することを目的とした実験である．薬物濃度を 7 段階に変化させ，各濃度あたり 3 プレートの変異コロニー数をカウントした結果である．

表 1.27　Ames 試験での変異コロニー数の比較

濃度		陽性対照 S						代替物質 T					
mg/plate	コロニー数			平均	分散	比	コロニー数			平均	分散	比
0	27	33	25	28.3	17.3	0.61	23	26	26	25.0	3.0	0.12
50	68	89	81	79.3	112.3	1.42	68	82	72	74.0	52.0	0.70
75	131	130	117	126.0	61.0	0.48	99	85	115	99.7	225.3	2.26
100	144	157	159	153.3	66.3	0.43	137	131	134	134.0	9.0	0.07
125	199	208	198	201.7	30.3	0.15	189	177	168	178.0	111.0	0.62
150	260	229	228	239.0	331.0	1.38	197	195	220	204.0	193.0	0.95
*300	427	407	456	430.0	607.0	1.41	335	332	348	338.3	72.3	0.21
					平均	0.84					平均	0.70

* この用量は，富山・杉本（2004）には含まれていない

用量ごとにプレートは 3 枚しかないが，分散/平均の比によるポアソン分布しているか検討する．陽性対照 S 薬の場合は，0.15～1.42，代替物質 T 薬の場合の比は 0.07～2.26，と大きなバラツキがある．全体で 14 群の比の平均値は，0.77 でありポアソン分布的ではある．その 95%信頼区間は，(0.40～1.14) と 1 を含むので全体としては，ポアソン分布にほぼ従うものと見なせる．

図 1.6 は，S 薬と T 薬別々に最小 2 乗法による回帰分析を適用した結果で，傾きは異なるが，切片の違いはわずかであり，$dose=0$ での変異コロニー数は，S 薬も T 薬も含まれていないので，切片を 2 群で共通で，傾きだけが異なる回帰直線のあてはめを行う．

このデータの用量ごとの変異コロニー数は，ポアソン分布に従うことが支持されているので，リンク関数を恒等としたポアソン回帰を行う．切片を共通とするためのデザイン行列に必要なダミー変数を設定する．詳細は，第 3.5 節「2 本の回帰直線に対するデザイン行列」を参照のこと．

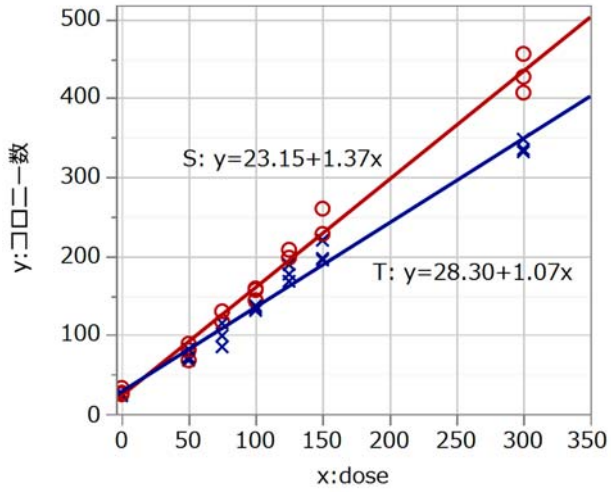


図 1.6　陽性対照薬 S および代替物質 T のコロニー数に対する線形回帰

$$x_S = 1 \quad x_T = 0 \quad \text{(S 薬の場合)}$$
$$= 0 \quad , \quad = 1 \quad \text{(T 薬の場合)}$$

$$y_{ijk} = \beta_0 + \beta_S x_S dose_{ij} + \beta_T x_T dose_{ij} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim poisson(\hat{y}_{ij})$$

ただし，薬剤：$i = 1,\ 2$，用量：$j = 1, \ldots, 7$，プレート：$k = 1,\ 2,\ 3$

表 1.28　切片を共通にする S 薬と T 薬の傾きのポアソン回帰による比較

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept	1	25.0407	1.7813	21.5493	28.5320	197.60	<.0001
xS*dose	1	1.0980	0.0263	1.0464	1.1496	1739.59	<.0001
dose*xT	1	1.3521	0.0282	1.2968	1.4074	2295.08	<.0001
尺度	0	1.0000	0.0000	1.0000	1.0000		

表 1.29　推定されたパラメータ共分散行列

推定値の共分散行列			
	Prm1	Prm2	Prm3
Prm1	3.17318	−0.02262	−0.02248
Prm2	−0.02262	0.0006931	0.0001602
Prm3	−0.02248	0.0001602	0.0007966

SAS プログラム　変異コロニー_a01.sas

```
Title "変異コロニー_a01.sas  " ;

data d01 ;
  input  dose @@ ;
  xS=1 ;  xT=0 ;
  do k=1 to 3 ;
     input y @@ ; output ;  end ;
  xS=0 ;  xT=1 ;
  do k=1 to 3 ;
     input y @@ ; output ;  end ;
datalines ;
  0   23  26  26  27  33  25
 50   68  82  72  68  89  81
 75   99  85 115 131 130 117
100  137 131 134 144 157 159
125  189 177 168 199 208 198
150  197 195 220 260 229 228
300  335 332 348 427 407 456
;
proc print data=d01 ; run ;

proc genmod  data=d01 ;
    model   y = xT*dose xS*dose /
dist=poisson link= identity covb ;
    run;
```

S 薬の傾きは $\beta_S = 1.3521$，T 薬の傾きは $\beta_T = 1.0980$ なので，効力比 ρ は，

$$\rho = \frac{\beta_T}{\beta_S} = 1.0980 / 1.3521 = 0.812$$

となる．95%信頼区間を求めるために，効力比 ρ を β_T と β_S で偏微分し，共分散行列を用いて，二次形式によるデルタ法で効力比 ρ の分散を計算する．なお，デルタ法については，第 12 章を参照のこと．

$$d_1 = \frac{\partial \rho}{\partial \beta_T} = \frac{1}{\beta_S} = \frac{1}{1.3521} = 0.7396$$

$$d_2 = \frac{\partial \rho}{\partial \beta_S} = \frac{-\beta_T}{{\beta_S}^2} = \frac{-1.0980}{1.3521^2} = -0.6006$$

デルタ法は，推定されたパラメータの比などの分散を計算するための汎用的な統計手法なのであるが，第 4.6 節に詳しく示すが，比についての偏微分および行列計算が必要であり，きちっと説明されている日本語の成書は見当たらない．そこで，偏微分の結果を縦ベクトル $[d_1 \quad d_2]^T$ とし，SAS/GENMOD で得られた共分散行列を用いて，効力比 ρ の分散を求めて，効力比 ρ の 95%信頼区間を Excel シート上での計算する方法を表 1.30 に示す．

効力比 ρ の分散は，

$$Var(\rho) = \boldsymbol{d}^T \boldsymbol{\Sigma} \boldsymbol{d} = 0.0005$$
$$SE = \sqrt{0.0005} = 0.0229$$

表 1.30　効力比 ρ の推定およびデルタ法による分散から 95%信頼区間の計算シート

		推定値		Prm1	Prm2	Prm3	二次形式による ρ の分散の計算式				
Intercept		25.0407	Prm1	3.1732	−0.0226	−0.0225		d^T	V		d
xT*dose	β_T	1.0980	Prm2	−0.0226	0.0007	0.0002	0.7396	-0.6006	0.0007	0.0002	0.7396
dose*xS	β_S	1.3521	Prm3	−0.0225	0.0002	0.0008			0.0002	0.0008	-0.6006
					共分散行列 Σ						
		d						β_T/β_S	95%L	95%U	
$1/\beta_S$ =		0.7396	ρ の分散 $d^T \Sigma d$ =			0.0005	ρ =	0.8121	0.7672	0.8569	
β_T/β_S^2 =		-0.6006			SE=	0.0229		95%CL=	$\rho \pm 1.96$ SE		

ρ の分散　$d^T V d$ = Mmult (Mmult （ d^T, V), d)

から，

$$\begin{aligned} 95\%\mathrm{CL} &= \rho \pm 1.96SE \\ &= 0.8121 \pm (1.96 \times 0.0229) \\ &= (0.7672, 0.8569) \end{aligned}$$

が得られる．結果の解釈は，効力比 $\rho = 0.8121$ であり，効力比の 95%信頼区間が 1.0 を含まないので，統計的に有意な差である．なお，効力比は，少ない用量で同等の効果を示す場合は，1 以上であり，多い用量で同等の効果を示す場合は，1 未満である．代替物質 T 薬は，陽性対照 S 薬に比べて同程度の変異原起こすための用量は，$1/\rho = 1/0.8121 = 1.23$ 倍が必要であり，変異原性の観点からでは望ましい結果となっている．効力比の推定と応用については，第 8.3 節で詳しく解説する．

1.9. 植物の体サイズに関連した種子数（対数リンク，2 群，回帰）

久保（2012），「データ解析のための統計モデリング入門，一般化線形モデル・階層ベイズモデル・MCMC」の第 3 章「一般化線形モデル －ポアソン回帰－」で取り上げられている人工データ（data3a.csv）を用いる．表 1.31 に施肥処理（なし，あり）の 2 群に対して，それぞれ 50 個体について体サイズ x_i に対する植物の種子数 y_i が示されている．

表 1.31 植物の体サイズと種子数のデータリスト

	C: 施肥処理なし					T: 施肥処理あり				
	x	*y*		*x*	*y*		*x*	*y*		*x*	*y*
1	8.31	6	26	10.21	6	51	10.14	14	76	10.24	6
2	9.44	6	27	9.45	7	52	9.05	6	77	11.76	8
3	9.50	6	28	10.44	9	53	9.89	7	78	9.52	9
4	9.07	12	29	9.44	3	54	8.76	9	79	10.40	9
5	10.16	10	30	10.48	10	55	12.04	6	80	9.96	6
6	8.32	4	31	9.43	2	56	9.91	7	81	10.30	7
7	10.61	9	32	10.32	9	57	9.84	9	82	11.54	10
8	10.06	9	33	10.33	10	58	11.87	13	83	9.42	6
9	9.93	9	34	8.50	5	59	10.16	9	84	11.28	11
10	10.43	11	35	9.41	11	60	9.34	13	85	9.73	11
11	10.36	6	36	8.96	10	61	10.17	7	86	10.78	11
12	10.15	10	37	9.67	4	62	10.99	8	87	10.21	5
13	10.92	6	38	10.26	8	63	9.19	10	88	10.51	6
14	8.85	10	39	10.36	9	64	10.67	7	89	10.73	4
15	9.42	11	40	11.80	12	65	10.96	12	90	8.85	5
16	11.11	8	41	10.94	8	66	10.55	6	91	11.20	6
17	8.02	3	42	10.25	9	67	9.69	15	92	9.86	5
18	11.93	8	43	8.74	8	68	10.91	3	93	11.54	8
19	8.55	5	44	10.46	6	69	9.60	4	94	10.03	5
20	7.19	5	45	9.37	6	70	12.37	6	95	11.88	9
21	9.83	4	46	9.74	10	71	10.54	10	96	9.15	8
22	10.79	11	47	8.95	10	72	11.30	8	97	8.52	6
23	8.89	5	48	8.74	9	73	12.40	8	98	10.24	8
24	10.09	10	49	11.32	12	74	10.18	7	99	10.86	7
25	11.63	6	50	9.25	6	75	9.53	5	100	9.97	9

久保（2012）には，R の glm() でリンク関数に対数を設定したポアソン回帰の結果が示されている．

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.2917     0.3637    3.55  0.00038
x              0.0757     0.0356    2.13  0.03358
```

久保（2012），P50 より引用

データの吟味

このデータについて，施肥処理の（なし，あり）別に基本統計および相関係数を表 1.32 に示す．体サイズ x は，平均が 10 前後に対し分散は 1 前後と小さく，種子数 y は，平均が 8 個弱に対し分散は 7 前後であり，ポアソン分布であるための条件を満たしている．体サイズ x と種子数 y の相関は，施肥処理が「ない」場合には，$r=0.39$ と弱い相関関係であるが，施肥処理が「ある」場合には，$r=0.07$ と無相関に近い．

表 1.32 植物の体サイズ x と種子数 y の基本統計量

		x 体サイズ			y 種子数				x vs. y
施肥処理	N	平均	分散	標準偏差	平均	分散	標準偏差	分散/平均	相関係数
C なし	50	9.808	1.000	1.000	7.780	6.869	2.621	0.88	0.39
T あり	50	10.371	0.892	0.944	7.880	7.047	2.655	0.89	0.07
全体	100	10.089	1.016	1.008	7.830	6.890	2.625	0.88	0.23

群間の t 検定をした結果を図 1.7 に示す．体サイズ x については，施肥処理あり群で有意に大きくなっているが，種子数 y については，全く差がないとの結果である．

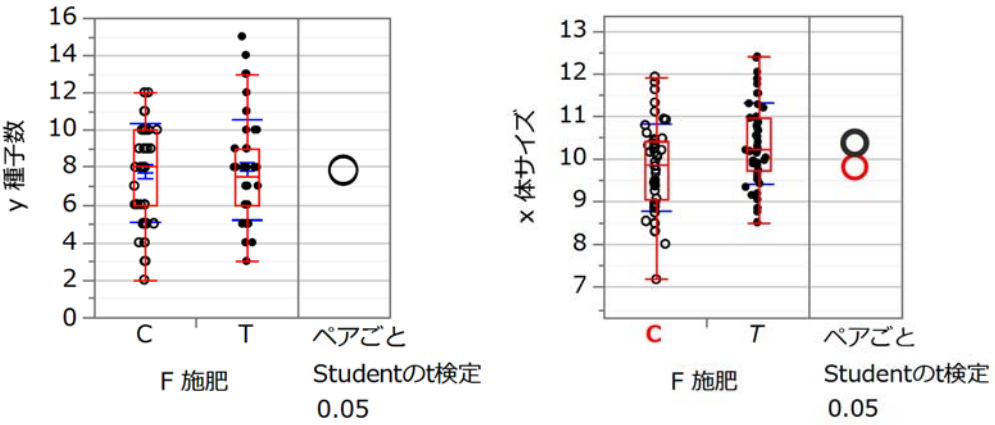


図 1.7　植物の体サイズと種子数について施肥処理の比較

図 1.8 に x と y の散布図上に 95%の確率楕円を上書きした結果を示す．施肥処理が「ある」場合に種子数 y は，体サイズ x との関連が見いだせない．施肥処理が「ない」場合には，弱い相関があることから，施肥処理により，体サイズが大きくなったが，それに伴い種子数が増えることはなかったと解される．従って，x を説明変数とするポアソン回帰は，施肥処理が「ない」場合には意味があるが，施肥処理が「ある」場合にはポアソン回帰を行う必然性はない．

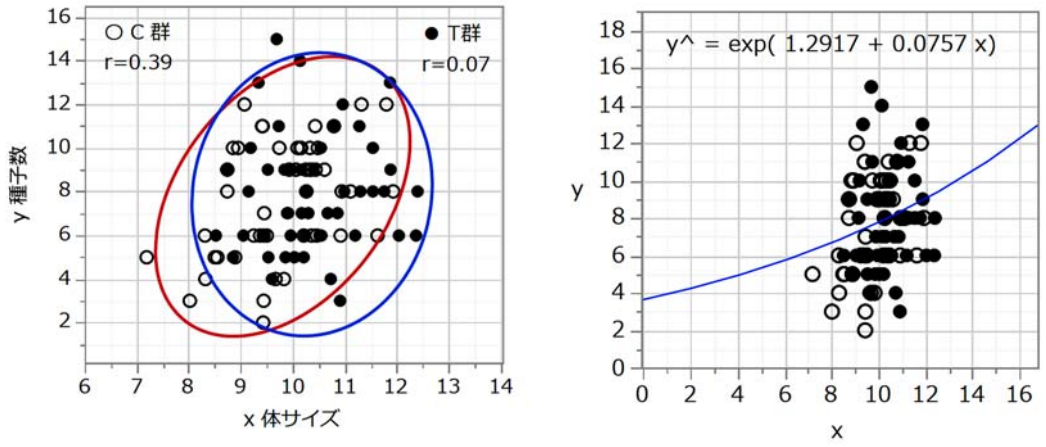


図 1.8　植物の体サイズと種子数の散布図と 95%確率楕円およびポアソン回帰

尤度比検定

施肥処理を込みにしたポアソン回帰の結果を表 1.33 に示す．対数リンクによるポアソン回帰式は，

$$\log(\hat{y}) = 1.2917 + 0.0757x$$

$$\hat{y} = \exp(1.2917 + 0.0757x)$$

Y 切片： $y_0 = \exp(1.2917) = 3.6390$

となる．傾き $\hat{\beta}_1 = 0.0757$ に対する尤度比カイ 2 乗=4.5139 であり，p 値=0.0336 と有意な差となっている．この p 値は，「モデル全体の検定」での「差分」の行の尤度比カイ 2 乗に対する p 値に等しい．

表 1.33 施肥処理を込みにしたポアソン回帰の結果

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(P>ChiSq)
差分	2.2570	4.5139	1	0.0336*
完全	235.3863			
縮小	237.6432			

適合度統計量	カイ2乗	自由度	p値(P>ChiSq)
Pearson	83.8448	98	0.8452
デビアンス	84.9930	98	0.8226

AICc
474.8962

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)
切片	1.2917	0.3637	12.4670	0.0004*
x 体サイズ	0.0757	0.0356	4.5139	0.0336*

「モデル全体の検定」には，各種の尤度比カイ 2 乗値が示されている．「差分」，「完全」「縮小」の意味付けについては，表 1.34 に示すように，3 種類のポアソン回帰の対数尤度 $\ln L$ が必要となる．

モデル	推定式	対数尤度
縮小モデル：	$\hat{y}_i = \exp(\hat{\beta}_0)$	$\ln L_{縮小} = \sum_i \ln(\text{Poisson.dist}(y_i, \exp(\hat{\beta}_0), false))$
完全モデル：	$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$	$\ln L_{完全} = \sum_i \ln(\text{Poisson.dist}(y_i, \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i), false))$
飽和モデル：	$\hat{y}_i = y_i$	$\ln L_{飽和} = \sum_i \ln(\text{Poisson.dist}(y_i, y_i, false))$

縮小モデルは，X 軸に平行な直線のあてはめであり，完全モデルは，図 1.8 右に示した $\exp(1.2917 + 0.0757x)$ であり，飽和モデルは，100 個の種子数 y_i それ自体を推定値にしたモデルである．それぞれの対数尤度は，パラメータ数が多くなれば大きくなるのであるが，それぞれの対数尤度の差の 2 倍が尤度比カイ 2 乗値となり，それぞれのパラメータ数の差を自由度とするカイ 2 乗分布に従うとして，p 値を算出する．表 1.34 に Excel による各モデルに対する対数尤度の計算結果を示す．

表 1.34 各種のモデルに対する Excel ソルバーを用いたポアソン回帰の結果

				縮小モデル		完全モデル		飽和モデル	
				y^=exp(β_0^)		y^=exp(β_0^+β_1^x)		y^=y	
				β_0^=	2.0580	β_0^=	1.2917	パラメータ	100 個
						β_1^=	0.0757		
				$\ln L$=	-237.6432	$\ln L$=	-235.3863	$\ln L$=	-192.8898
	施肥	体サイズ	種子数	確率	対数尤度	確率	対数尤度	確率	対数尤度
i	F	x	y	p	ln(p)	p	ln(p)	p	ln(p)
1	C	8.31	6	0.1273	-2.0615	0.1525	-1.8806	0.1606	-1.8287
2	C	9.44	6	0.1273	-2.0615	0.1385	-1.9767	0.1606	-1.8287
3	C	9.50	6	0.1273	-2.0615	0.1376	-1.9833	0.1606	-1.8287
4	C	9.07	12	0.0441	-3.1217	0.0308	-3.4796	0.1144	-2.1683
5	C	10.16	10	0.0949	-2.3548	0.0954	-2.3494	0.1251	-2.0786
:									
98	T	10.24	8	0.1393	-1.9709	0.1395	-1.9697	0.1396	-1.9691
99	T	10.86	7	0.1424	-1.9494	0.1343	-2.0077	0.1490	-1.9038
100	T	9.97	9	0.1212	-2.1102	0.1195	-2.1246	0.1318	-2.0268

表 1.33 の「モデル」の欄の

「完全」は $-\ln L_{完全} = -(-235.3863) = 235.3863$,

「縮小」は $-\ln L_{縮小} = -(-237.6432) = 237.6432$,

「差分」は $(\ln L_{完全} - \ln L_{縮小}) = -235.3863 - (-237.6432) = 2.2570$

で計算されている．「差分」の 2 倍が「尤度比カイ 2 乗」となっている．「適合度統計量」の「デビアンス」は，「逸脱度」ともいわれ，カイ 2 乗値は，

$$\begin{aligned}\text{デビアンス} &= 2\times(\ln L_{\text{飽和}} - \ln L_{\text{完全}}) \\ &= 2\times[-192.8898-(-235.3863)] \\ &= 2\times 42.4965 = 84.9930\end{aligned}$$

で計算されている．$\ln L_{\text{飽和}}$ の自由度は 100，$\ln L_{\text{完全}}$ は 2 なので，デビアンスの自由度は，$100-2=98$ となり，デビアンスのカイ 2 乗値=84.9930 は，自由度 98 のカイ 2 乗分布に従うことから，

$$p = 1 - \text{Chisq.dist}\left(84.9930, 98, true\right) = 0.8226$$

と，有意な差ではない．このことは，パラメータ数が 100 個のモデルに対し，パラメータ数が 2 個の回帰モデルは，あてはめは悪くはないことが示されてる．

表 1.33 の適合度統計量」の「Pearson」のカイ 2 乗値は，通常の Pearson 残差で，

$$\chi^2_{Pearson} = \sum_i \frac{[y_i - \exp(1.2917 + 0.0757\times x_i)]^2}{\exp(1.2917 + 0.0757\times x_i)} = 83.8448$$

として計算されたものである．

回帰式の妥当性

ポアソン回帰で 1 次式でのあてはめで十分であることは，2 次式をあてはめて，1 次式と 2 次式の対数尤度の差の 2 倍で評価できる．表 1.34 に更に 2 次式を加えた結果から，

回帰モデル・1 次式：　$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$　　$\ln L_{\text{縮小}} = -235.3863$

回帰モデル・2 次式：　$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2)$　　$\ln L_{\text{完全}} = -234.3971$

$$2\times(\ln L_{\text{完全}} - \ln L_{\text{縮小}}) = 2\times 0.9892 = 1.9784$$

が求められ，2 次式のあてはめの必要性はなく，1 次式で十分と言える．良い統計モデルの選択の基準には，AIC による選択が標準的であるが，ここでは割愛する．

個別の 95%信頼区間

施肥処理がある T 群の場合に体サイズ x との相関がないということは，ポアソン回帰を行う前提を満たしていないので，C 群の場合に限定し，対数リンクによるポアソン回帰を行い，回帰式の 95%信頼区間，個別データの 95%信頼区間を JMP ファイルに出力し，「重ね合わせプロット」を用い，ポアソン回帰直線の 95%信頼区間および個別の 95%信頼区間を散布図上に上書きした結果を表 1.35 に示す．

パラメータの推定値は，対数変換した場合の結果なので，実目盛り上では，指数を取って

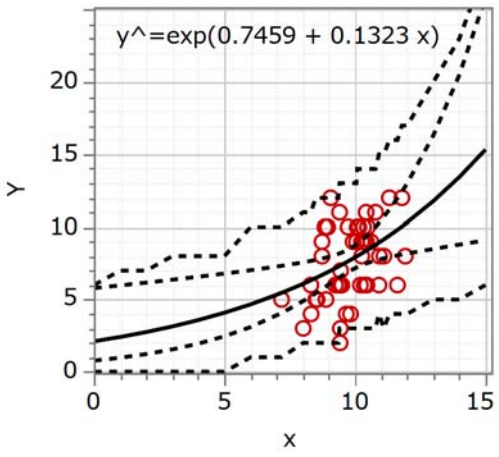
$$\hat{y}_i = \exp(0.7459 + 0.1323_i)$$

となる．切片は，$\hat{y}_i = \exp(0.7459) = 2.1083$ であることが，確認される．また，個別データの 95%信頼区間が，ほぼ全データを包含しているので，適切なポアソン回帰となっていることが裏付けられる．なお，Excel による 95%信頼区間の求め方については，第 4.7 節，第 5.2 節を参照してもらいたい．

表 1.35　C 群に対するポアソン回帰のあてはめと 95%信頼区間および予測区間

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗
切片	0.7459	0.5154	2.0730
x	0.1323	0.0516	6.5990



個別データの 95%信頼区間（予測区間）の中に，ほとんど全ての観測データが含まれており，対数リンクのポアソン回帰のあてはめの妥当性が示されている．また，図 1.9 に示した「スチューデント化デビアンス残差」も，ほとんどが（−2〜＋2）の範囲に入っていることからも，あてはめの妥当性が示されている．なお，各種の残差については，第 11 章を参照のこと．

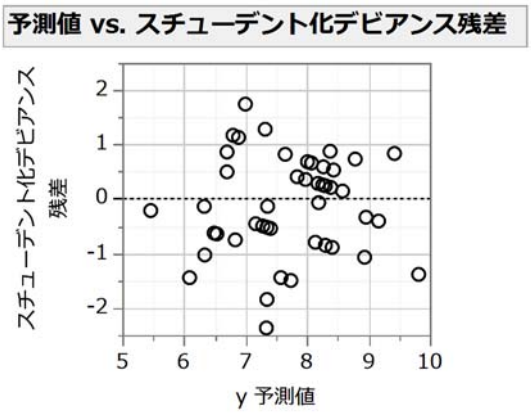


図 1.9 スチューデント化デビアンス残差プロット

1.10. 退役軍人における癌の発生（対数リンク，2 群，11 水準，オフセット）

アーミテジ著，椿・椿 共訳（2001），「医学研究のための統計的方法」の第 12.8 節の退役軍人の癌の発生数に関するデータを表 1.36 に示す．このデータは，20 年の期間にわたって追跡された退役軍人に対して，実戦経験が（なし，あり）で，癌の発生数を調べたものである．

各軍人は，20 年の追跡期間中に，いくつか年齢階層に重複して記録されている．研究は実戦経験のある群と実戦経験のない群間の発癌リスクに差が存在するか否かを評価するために行われた．アーミテジ（2001）では，年齢階層を名義尺度として取り扱っているが，ここでは，連続尺度として扱う．

表 1.36　退役軍人の実践経験の有無での癌の発生数の比較

			実戦経験なし:x=0				実戦経験あり:x=1		
age	年齢階層	i	癌の数 y	人年 n	10万人比	i	癌の数 y	人年 n	10万人比
20	-24	1	18	208,487	8.6	12	6	60,840	9.9
25	25-29	2	60	303,832	19.7	13	21	157,175	13.4
30	30-34	3	122	325,421	37.5	14	54	176,134	30.7
35	35-39	4	191	312,242	61.2	15	118	186,514	63.3
40	40-44	5	108	165,597	65.2	16	97	135,475	71.6
45	45-49	6	88	54,396	161.8	17	58	42,620	136.1
50	50-54	7	74	40,716	181.7	18	56	25,001	224.0
55	55-59	8	120	33,801	355.0	19	54	13,710	393.9
60	60-64	9	141	26,618	529.7	20	34	6,163	551.7
65	65-69	10	108	17,404	620.5	21	9	1,575	571.4
70	70-	11	99	14,146	699.8	22	2	273	732.6
	全体		1129	1,502,660	75.1		509	805,480	63.2

各年代の人年数が大きく異なるので，癌の発現数 y_i の加齢に伴なう変化が見えにくいので，10 万人比に換算した結果を加えてある．年齢別の実践経験の有無による癌の発生数には，図 1.10 に示すように差がないように判断される．

人年 n_i について対数オフセットを $\ln(n_i)$ とし，リンク関数を対数とした場合のポアソン回帰を行い，表 1.37 に示す結果を得た．

$$\begin{aligned}\hat{y}_i &= n_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i + \beta_2 age_i) \\ &= n_i \exp(-10.5316 + 0.0362 x_i + 0.0851 age_i) \\ \ln(\hat{y}_i) &= \ln(n_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i + \beta_2 age_i) \\ &= \ln(n_i) + (-10.5316 + 0.0362 x_i + 0.0851 age_i)\end{aligned}$$

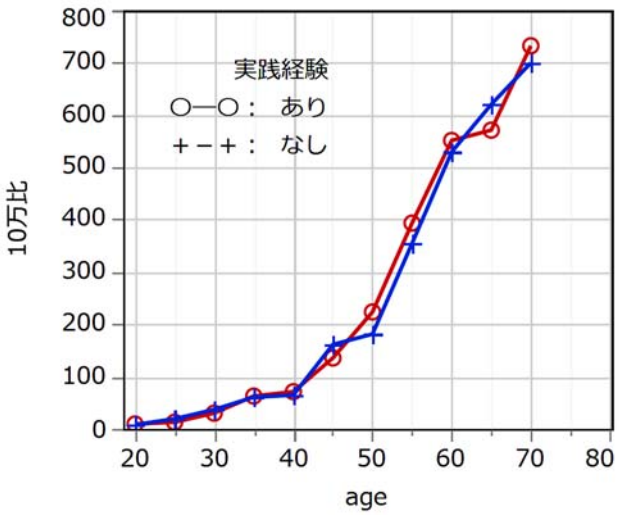


図 1.10　実践経験のなし・あり別 10 万人比での癌の発生数

パラメータの推定結果は，表 1.37 に示すように，実践経験の有無 x の尤度比カイ 2 乗値が 0.4367 であり，有意な差とはならなかった．オフセットの解釈については第 2.6 節を，また，2 次式のあてはめについては，第 5.3 節を参照してもらいたい．

表 1.37　JMP のポアソン回帰による退役軍人の群における癌の発生ケースの比較

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)	下側信頼限界	上側信頼限界
切片	-10.5316	0.0899	18144.850	<.0001*	-10.7090	-10.3567
実践経験x	0.0363	0.0547	0.4367	0.5087	-0.0717	0.1430
age	0.0851	0.0018	1981.2547	<.0001*	0.0816	0.0886

オフセットを除いた推定値は，部分集団が 1 人とした場合の推定値であり，10 万人比などのように部分集団のおおきさを固定して比較検討することが必要である．図 1.11 に，実践経験の（なし，あり）別に，10 万人比に換算した 10 万人比での癌の発生数にポアソン回帰で推定した回帰直線を上書きした結果を示す．

実践経験なしで，年齢が 60 歳での癌の発現数の 10 万人比での推定値は，

$$\text{実践経験なし}\ \ x=0\ \ :\quad \begin{aligned}\hat{y}^{(10\text{万人比})} &= 100{,}000\times\exp(\hat{\beta}_0+\hat{\beta}_1 x+\hat{\beta}_2 age)\\ &= 100{,}000\times\exp(-10.5316+0.0362\times 0+0.0851\times 60)\\ &= 440.9\end{aligned}$$

実践経験がある場合には，

$$\text{実践経験あり}\quad x=1\quad :\quad \begin{aligned}\hat{y}^{(10\text{万人比})} &= 100{,}000\times\exp(\hat{\beta}_0+\hat{\beta}_1 x+\hat{\beta}_2 age)\\ &= 100{,}000\times\exp(-10.5316+0.0362\times 1+0.0851\times 60)\\ &= 457.2\end{aligned}$$

であり，その差は，$457.2-440.9=16.3$ とわずかである．

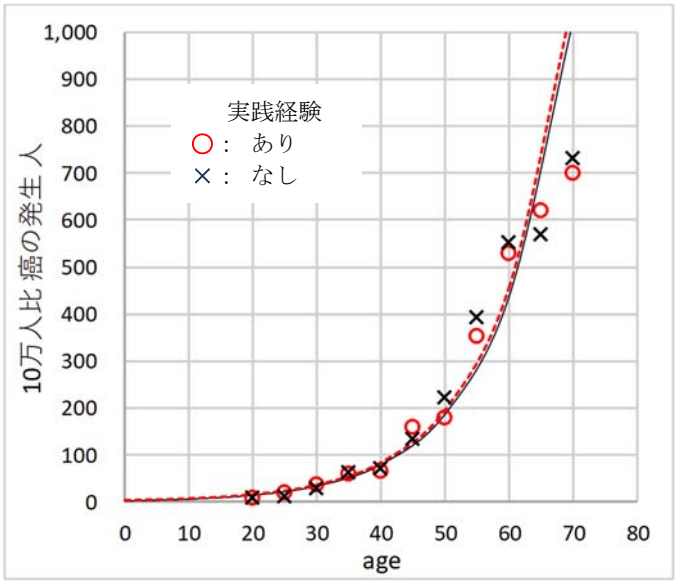


図 1.11　実践経験のありなし別の 10 万人比での癌の発生数と推定値

加齢と共に 10 万人比の癌の発生数は，指数関数的に増大するのであるが，60 歳を超えるあたりから頭打ちになっている．したがって，実践経験の有無による統計的に差は検出されなかったが，対数リンクのポアソン回帰をあてはめることに無理がある．どのような統計モデルが適切なのか，さらなる検討が必要である．第 3.6 節および第 12.6 節に 2 次式のあてはめの事例が示されている．

1.11.　喫煙による冠動脈心疾患による死亡（対数リンク，2 群，5 水準，オフセット）

ドブソン（2008）の第 9.2 節に，英国の男性医師を対象に喫煙習慣が 10 年間の間に冠動脈心疾患の死亡に対する影響を調べてた結果があり，表 1.38 にデータを示す．このデータは，前節と同様に対数リンクでオフセットがある事例である．

表 1.38　年齢階層別の喫煙習慣と冠動脈心疾患による死亡数の関係

年齢層		非喫煙者（$x=0$）			喫煙者（$x=1$）		
歳	範囲	死亡	人年	10万人比	死亡	人年	10万人比
40	35-44	2	18,790	10.6	32	52,407	61.1
50	45-54	12	10,673	112.4	104	43,248	240.5
60	55-64	28	5,710	490.4	206	28,612	720.0
70	65-74	28	2,585	1083.2	186	12,663	1468.8
80	74-	31	1,462	2120.4	102	5,317	1918.4

対象となる人年が喫煙者と非喫煙者で大きく異なるので，10 万人比に換算した結果を図 1.12 に示す．年齢が 70 歳代までは，喫煙者の死亡数が多いが，80 歳代でやや逆転している．喫煙者で 80 歳を超えて生存しているの人達は，タバコに対して強い耐性を持った人達と推測され，非喫煙者と同様の冠動脈心疾患による死亡数となっている．

ドブソン（2008）では，対数リンクでのポアソン回帰に年齢の 2 乗の項，年齢と喫煙の交互作用を含めたモデルが示されているが，結果の解釈が複雑になるので，ここでは，80 歳代を除いたモデルする．ポアソン回帰の解析は，これまで，第 1.4 節では，Excel による反復重み

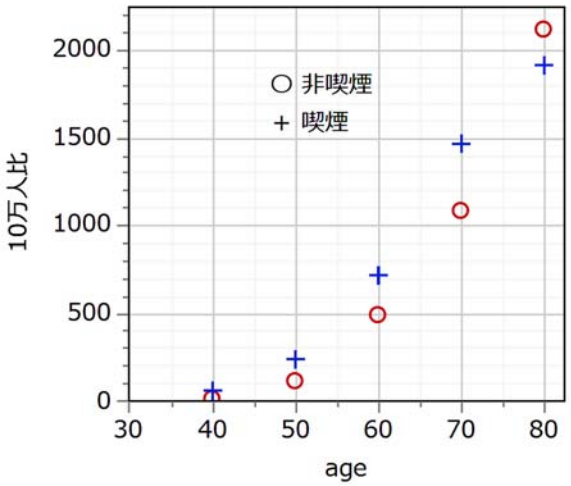


図 1.12　喫煙習慣による冠動脈心疾患による 10 万人あたりの死亡数の比較

付き回帰，第 1.5 節および第 1.6 節では JMP による回帰，第 1.6 節では SAS/GENMOD プロシジャによるポアソン回帰を示してきた．本節では，第 1.10 節と同様に Excel のソルバーを用いて，対数変換することなく，ポアソン回帰のパラメータ推定を行う方法を示す．詳しくは，第 2 章で取り上げるが，統計ソフトと Excel を互いに補完的に使う例として示す．なお，80 歳代を含めた解析については，第 3.6 節で取り上げる．

ポアソン回帰式は，非喫煙者を $x=0$，喫煙者を $x=1$，人年を n_i としたときに，

$$y_i = n_i \exp(\beta_0 + \beta_1 x_i + \beta_2 age_j) + \varepsilon_i \quad \varepsilon_i \sim poisson(y_i; \hat{y}_i)$$

ただし， $i = 1, 2, ..., 8$

である．死亡者数 y_i の対数尤度 $\ln L_i$ を，

$$\ln L_i = \ln[\text{Poisson.dist}(y_i, \hat{y}_i, false)]$$

表 1.39　Excel による年齢と喫煙習慣によるポアソン回帰（初期値）

	切片	喫煙	年齢層	死亡	人年	10万人比	推定値	確率	対数尤度		
i		x	age	y	n	y'	$y^$	P	$\ln L_i$		最尤解
1	1	0	40	2	18,790	11	9.4	0.0036	-5.6145	β_0^=	**-10.0000**
2	1	0	50	12	10,673	112	9.7	0.0895	-2.4141	β_1^=	**2.0000**
3	1	0	60	28	5,710	490	9.5	0.0000	-14.3780	β_2^=	**0.0600**
4	1	0	70	28	2,585	1,083	7.8	0.0000	-18.1065		
5	1	1	40	32	52,407	61	193.8	0.0000	-106.8144		
6	1	1	50	104	43,248	240	291.4	0.0000	-83.4919		
7	1	1	60	206	28,612	720	351.3	0.0000	-38.9192		
8	1	1	70	186	12,663	1,469	283.3	0.0000	-22.5645		
								$\ln L$ =	**-292.3030**		

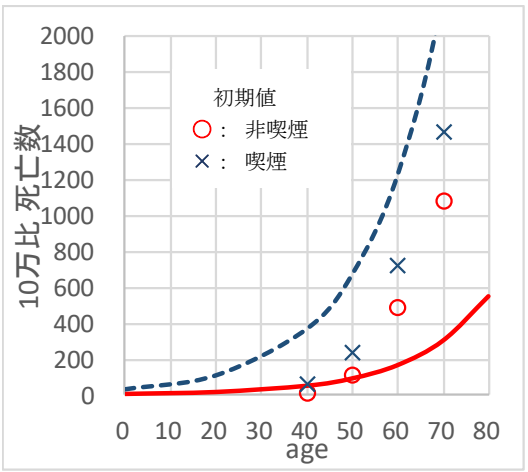


図 1.13　10 万人あたりの死亡数の予測（初期値）

としたときに対数尤度 $\ln L$ は，

$$\ln L = \sum_i \ln L_i$$

となる．表 1.39 に示すように適当なパラメータ（$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$）を設定し，$\ln L$ を最大化するように Excel のソルバーでパラメータを変化させると，最尤解が得られる．

一般化線形モデルは，指数関数で与えられた推定値に関するモデル式の両辺に対数を取り，

$$\ln(\hat{y}_i) = \ln(n_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 age_j)$$

のように線形化して反復重み付き回帰によって，最尤解を求める方法である．ポアソン回帰によらず，他の指数型分布族についても，同じ計算手順で最尤解を求められる．しかし，打ち切りデータがあるような場合への拡張性に難点がある．対数尤度関数 2 階の偏微分行列を使ったニュートン・ラフソン法が万能の計算手段であり，その入門として，Excel のソルバーを用いた最尤法は，ここに示したように，対数尤度関数を示しさえすれば，一気に最尤解を求めてくれる．

また，統計ソフトで出力される結果のグラフは，便利な面もあるが，図 1.13 で示したように，元のデータの散布図に 2 本の推定直線を上書きすることは容易ではない．解析結果を適切な図表に示すことは，結果を解釈するために不可避であり，オフセットがあるような場合は，10 万人比などに換算することにより結果の解釈が容易になる．

表 1.39 の計算シート内に 10 万人比に換算した結果があり，これを図 1.13 の散布図として用いている．ただし，上書きされている 10 万人比でのポアソン回帰での推定値は，表 1.39 の外側で別途計算したものである．表 1.40 は，Excel のソルバーで，$\ln L$ を最大にするように，パラメータ（$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$）を変化させた結果である．この中の図は，図 1.13 が計算された結果が反映されている．表 1.39 の初期値は，この図を見ながら，適度な広がりを持つ値として（$\hat{\beta}_0 = -10$，$\hat{\beta}_1 = 2$，$\hat{\beta}_2 = 0.06$）にセットした結果である．

Excel のソルバーを用いた方法は，手軽ではあるが推定したパラメータについての標準誤差が出力されない．第 2 章で示すように，対数尤度関数 2 階の偏微分行列を使ったニュートン・ラフソン法を適用すれば，共分散行列が計算過程に含まれるので，この対角要素の平方根が標準誤差として得られる．実用的には，使い慣れた統計ソフトにより，表 1.41 に示すように検定統計量および共分散行列を出力して，結果を Excel に取り込み，新たな追加の計算あるいは必要な図表の作成に使うことが望ましい．

表 1.40　Excel による年齢と喫煙習慣によるポアソン回帰（最尤解）

	切片	喫煙	年齢層	死亡	人年	10万人比	推定値	確率	対数尤度		
i		x	age	y	n	y'	$y^{\wedge}$	P	$\ln L_i$		最尤解
1	1	0	40	2	18,790	11	8.9	0.0053	-5.2468	$\beta_0\hat{}=$	**-11.7285**
2	1	0	50	12	10,673	112	14.1	0.0975	-2.3275	$\beta_1\hat{}=$	**0.5182**
3	1	0	60	28	5,710	490	20.8	0.0249	-3.6945	$\beta_2\hat{}=$	**0.1019**
4	1	0	70	28	2,585	1,083	26.2	0.0705	-2.6516		
5	1	1	40	32	52,407	61	41.8	0.0200	-3.9126		
6	1	1	50	104	43,248	240	95.7	0.0275	-3.5949		
7	1	1	60	206	28,612	720	175.4	0.0022	-6.1107		
8	1	1	70	186	12,663	1,469	215.1	0.0037	-5.5985		
								$\ln L$ =	**-33.1370**		

	x	age	y	n	y'	$y^{\wedge}$
1	0	0		100,000		0.8
1	0	20		100,000		6.2
1	0	40		100,000		47.5
1	0	50		100,000		131.8
1	0	60		100,000		365.1
1	0	70		100,000		1011.8
1	0	80		100,000		2804.0
1	1	0		100,000		1.4
1	1	20		100,000		10.4
1	1	40		100,000		79.8
1	1	50		100,000		221.2
1	1	60		100,000		613.0
1	1	70		100,000		1698.8
1	1	80		100,000		4707.8

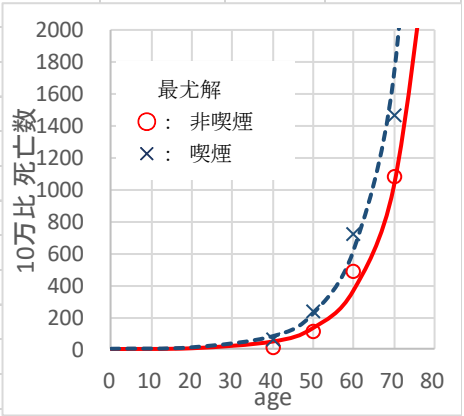


表 1.41　JMP による年齢と喫煙習慣によるポアソン回帰

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	-11.7283	0.5660	679.8171	<.0001*
x_Somok	0.5181	0.2584	4.5802	0.0323*
age	0.1019	0.0086	156.8010	<.0001*

共分散

	切片	x_Somok	age
切片	0.3204	-0.0522	-0.0044
x_Somok	-0.0522	0.0668	-0.0001
age	-0.0044	-0.0001	0.0001

JMP の出力から，非喫煙者に対する喫煙者の推定値は，$\hat{\beta}_1 = 0.5181$ で，p 値は．0.0323 と統計的に有意である．この結果の解釈は，どうしたら良いのであろうか．表 1.42 に示すように推定された回帰式は，

$$\hat{y} = n\exp(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 age)$$
$$= n\exp(-11.7283 + 0.5181x + 0.10193\,age)$$

なので，10 万人比に換算する場合は， $age=50$ と固定し， $x=0$ or 1 として計算すると，

$$x=0 \qquad \hat{y}'=100{,}000\times\exp(-11.7283+0.5181\times0+0.10193\times50)=131.8$$

$$x=1 \qquad \hat{y}''=100{,}000\times\exp(-11.7283+0.5181\times1+0.10193\times50)=221.2$$

となる．既に，年齢を変えて計算した結果は，表 1.40 の推定曲線の作成のための計算結果に等しく，喫煙者の非喫煙者に対するリスク比は，1.68 倍となる．この様な計算はせずとも，

$$\exp(\hat{\beta}_1)=\exp(0.5181)=1.68$$

として直接計算可能である．

表 1.42　喫煙者の非喫煙者に対するリスク比

	非喫煙	喫煙	
age	x=0	x=1	比
20	6.2	10.4	1.68
40	47.5	79.8	1.68
50	131.8	221.2	1.68
60	365.1	613.0	1.68
70	1011.8	1698.8	1.68
80	2804.0	4707.8	1.68

なお，年齢区分の 80 歳まで含め，2 次式のあてはめについては，第 12.6 節「オフセットを含むポアソン回帰の 95%信頼区間」で取り上げる

1.12.　医院への通院回数（過分散）

ポアソン分布は，稀な現象がカウントできる場合の分布として知られている．しかし，稀ではない現象でのカウント・データに対しても適用できるのではないかと期待されるが，ゼロ・カウントが異常に多い，分散が平均値よりも数倍も大きい，などポアソン分布を仮定することに躊躇される事例に直面する．Cameron and Trivde (1998)，Regression Analysis of Count Data の第 3 章の「Table 3.1 医院への通院回数」を表 1.43 に示す．

表 1.43 医院への通院回数

カウント	0	1	2	3	4	5	6	7	8	9	N	平均	分散	分散/平均
度数	4,141	782	174	30	24	9	12	12	5	1	5,190	0.3017	0.6370	2.1112

平均が 0.3017，分散が 0.6370，分散/平均が 2.11 倍と過分散が起きている．JMP の「一変量の分布」を用いて，ポアソン分布および負の 2 項分布から導出されたガンマ・ポアソン分布のあてはめを行う．図 1.14 に棒グラフにポアソン分布およびガンマ・ポアソン分布の確率関数をあてはめた結果を示す．

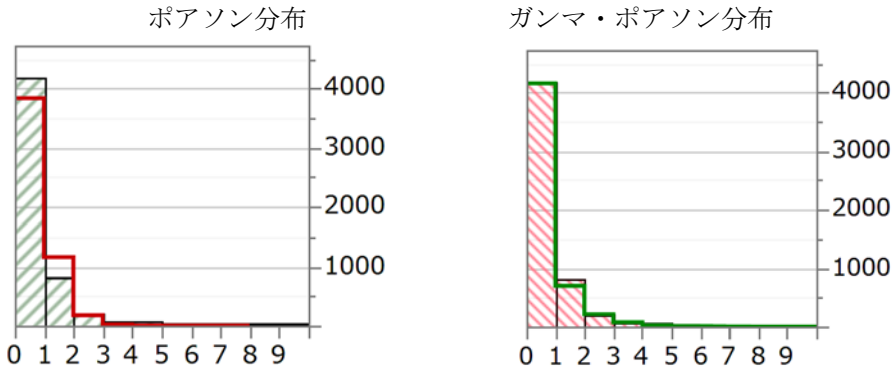


図 1.14　通院回数に対するポアソン分布およびガンマ・ポアソン分布のあてはめ

ポアソン分布の期待度数は，ゼロ・カウントに対して低めとなり，カウント 1 に対しては，やや高めとなっている．ガンマ・ポアソン分布（負の 2 項分布）の場合は，きれいにあてはまっているようである．棒グラフでは，カウントが 5 以上のあてはまりを検討できないので，表 1.44 に推定されたパラメータを用いてそれぞれの分布の確率関数を計算し，期待度数を求めて，観測度数との差を求めた．

表 1.44　ガンマ・ポアソン分布（負の 2 項分布）のあてはめ

		ポアソン分布			ガンマ･ポアソン分布		
カウント	度数	確率	期待度数	差	確率	期待度数	差
0	4,141	0.7395	3838.3	302.7	0.8011	4157.8	-16.8
1	782	0.2231	1158.0	-376.0	0.1344	697.3	84.7
2	174	0.0337	174.7	-0.7	0.0411	213.3	-39.3
3	30	0.0034	17.6	12.4	0.0145	75.1	-45.1
4	24	0.0003	1.3	22.7	0.0054	28.2	-4.2
5	9	0.0000	0.1	8.9	0.0021	11.0	-2.0
6	12	0.0000	0.0	12.0	0.0008	4.4	7.6
7	12	0.0000	0.0	12.0	0.0003	1.8	10.2
8	5	0.0000	0.0	5.0	0.0001	0.7	4.3
9	1	0.0000	0.0	1.0	0.0001	0.3	0.7
全体	5,190	μ=0.3017			μ=0.3017,	過分散σ=1.7992	

ポアソン分布の場合には，カウント 4 以上から観測度数に対して期待度数が大幅に落ち込んでいるのに対し，ガンマ・ポアソン分布の場合には，やや低めではあるものの期待度数も追従している．稀な現象ではないカウント・データについては，ガンマ・ポアソン分布のあてはめが適しているように思われる．なお，ガンマ・ポアソン分布については，第 6.2 節を参照のこと．

1.13. 雌のカブトガニに連結する雄の数(2 因子, 2 変数, 対数リンク, 過分散)

アグレスティ著, 渡邉・菅波・吉田・角野・寒水・松永 訳 (2003) の「カテゴリカルデー解析入門」の第 4.3 節の「計数データに対する一般化線形モデル : ポアソン回帰」に雌のカブトガニに連結する雄のサテライト数 (Satellite 数) について, 対数リンクによるポアソン回帰が例示されている. 表 1.45 に示すようにデータには、173 匹のカブトガニについて説明変数として名義尺度 (甲羅の色, 後体部の棘の状態) の 2 変数, 連続尺度 (甲羅の幅, 体重) の 2 変数, 反応変数としてサテライト数が含まれている.

アグレスティ (2003) では, まず甲羅の幅をX軸, サテライト数をY軸とした散布図と共に対数リンクによるポアソン回帰の結果が示されている. 引き続き, 甲羅の幅を 8 区分とし区分内のカブトガニの数とサテライト数の合計を算出し, カブトガニの数をオフセットとした解析を主体にしている. さらに, 第 5 章では, サテライト数が 0 か 1 以上かの 2 値データとして, ロジスティック回帰を主体にして様々な角度からの解析方法が提示されている.

表 1.46 に示すように, サテライト数の平均は 2.9191, 分散は 9.9120 であり, その比は 3.40 と過分散になっている. ポアソン分布を棒グラフ上に上書きした結果を見ても, 誤差分布にポアソン分布を仮定することは絶望的とも思われる. もちろん, 適合度検定でも $\chi^2 = 584.0436$, $p < 0.0001$ でポアソン分布があてはまるとは言えない. さらなる探索的な解析は, 第 7.3 節で行う.

全データが過分散となる場合でも, 何らかの条件によりサテライト数の平均が大きく異なる部分集団の集まりが複数存在するとも考えられる. 甲羅の色によってサテライト数の平均が大きく異なり, それに伴って過分散が解消されるのであろうか. あるいは, 甲羅の幅, あるいは, 体重によりサテライト数の平均が大きく異なり, 過分散が解消されるのだろうか. 検討すべき事項である.

アグレスティ (2003) と同様に甲羅の幅を説明変数とし, サテライト数を反応変数としたときの対数リンクのポアソン回帰

$$Satellite_i = \exp(\beta_0 + \beta_0 \cdot width_i) + \varepsilon_i, \quad \varepsilon_i \sim Poisson$$

の結果を表 1.47 に示す. Pearson の適合度のカイ 2 乗値は 544.1570 と自由度の 171 に対して 3.1822 倍と過分散となっていて, 表 1.46 から得られた分散を平均で割った比 9.9120/2.9191 =3.3956 であった過分散が, 甲羅の幅を説明変数としても解消されていない. ポアソン回帰か

表 1.45　雌のカブトガニに連結する雄のサテライト数

color	spine	width	weight	satell	color	spine	width	weight	satell	color	spine	width	weight	satell	color	spine	width	weight	satell
2	3	28.3	3.050	8	3	1	28.5	3.250	9	4	3	23.5	1.900	0	2	1	28.0	2.900	4
3	3	22.5	1.550	0	3	3	28.9	2.800	4	2	2	24.0	1.700	0	4	3	25.8	2.250	10
1	1	26.0	2.300	9	2	3	28.2	2.600	6	2	1	29.7	3.850	5	2	3	27.9	3.050	7
3	3	24.8	2.100	0	2	3	25.0	2.100	4	2	1	26.8	2.550	0	2	3	24.9	2.200	0
3	3	26.0	2.600	4	2	3	28.5	3.000	3	4	3	26.7	2.450	0	2	1	28.4	3.100	5
2	3	23.8	2.100	0	2	1	30.3	3.600	3	2	1	28.7	3.200	0	3	3	27.2	2.400	5
1	1	26.5	2.350	0	4	3	24.7	2.100	5	3	3	23.1	1.550	0	2	2	25.0	2.250	6
3	2	24.7	1.900	0	2	3	27.7	2.900	5	2	1	29.0	2.800	1	2	3	27.5	2.625	6
2	1	23.7	1.950	0	1	1	27.4	2.700	6	3	3	25.5	2.250	0	2	1	33.5	5.200	7
3	3	25.6	2.150	0	2	3	22.9	1.600	4	3	3	26.5	1.967	1	2	3	30.5	3.325	3
3	3	24.3	2.150	0	2	1	25.7	2.000	5	3	3	24.5	2.200	1	3	3	29.0	2.925	3
2	3	25.8	2.650	0	2	3	28.3	3.000	15	3	3	28.5	3.000	1	2	1	24.3	2.000	0
2	3	28.2	3.050	11	2	3	27.2	2.700	3	2	3	28.2	2.867	1	2	3	25.8	2.400	0
4	2	21.0	1.850	0	3	3	26.2	2.300	3	2	3	24.5	1.600	1	4	3	25.0	2.100	8
2	1	26.0	2.300	14	2	1	27.8	2.750	0	2	3	27.5	2.550	1	2	1	31.7	3.725	4
1	1	27.1	2.950	8	4	3	25.5	2.250	0	2	2	24.7	2.550	4	2	3	29.5	3.025	4
2	3	25.2	2.000	1	3	3	27.1	2.550	0	2	1	25.2	2.000	1	3	3	24.0	1.900	10
2	3	29.0	3.000	1	3	3	24.5	2.050	5	3	3	27.3	2.900	1	2	3	30.0	3.000	9
4	3	24.7	2.200	0	3	1	27.0	2.450	3	2	3	26.3	2.400	1	2	3	27.6	2.850	4
2	3	27.4	2.700	5	2	3	26.0	2.150	5	2	3	29.0	3.100	1	2	3	26.2	2.300	0
2	2	23.2	1.950	4	2	3	28.0	2.800	1	2	3	25.3	1.900	2	2	1	23.1	2.000	0
1	2	25.0	2.300	3	2	3	30.0	3.050	8	2	3	26.5	2.300	4	2	1	22.9	1.600	0
2	1	22.5	1.600	1	2	3	29.0	3.200	10	2	3	27.8	3.250	3	4	3	24.5	1.900	0
3	3	26.7	2.600	2	2	3	26.2	2.400	0	2	3	27.0	2.500	6	2	3	24.7	1.950	4
4	3	25.8	2.000	3	2	1	26.5	1.300	0	3	3	25.7	2.100	0	2	3	28.3	3.200	0
4	3	26.2	1.300	0	2	3	26.2	2.400	3	2	3	25.0	2.100	2	2	3	23.9	1.850	2
2	3	28.7	3.150	3	3	3	25.6	2.800	7	2	3	31.9	3.325	2	3	3	23.8	1.800	0
2	1	26.8	2.700	5	3	3	23.0	1.650	1	4	3	23.7	1.800	0	3	2	29.8	3.500	4
4	3	27.5	2.600	0	3	3	23.0	1.800	0	4	3	29.3	3.225	12	2	3	26.5	2.350	4
2	3	24.9	2.100	0	2	3	25.4	2.250	6	3	3	22.0	1.400	0	2	3	26.0	2.275	3
1	1	29.3	3.200	4	3	3	24.2	1.900	0	2	3	25.0	2.400	5	2	3	28.2	3.050	8
1	3	25.8	2.600	0	2	2	22.9	1.600	0	3	3	27.0	2.500	6	4	3	25.7	2.150	0
2	2	25.7	2.000	0	3	2	26.0	2.200	3	3	3	23.8	1.800	6	2	3	26.5	2.750	7
2	1	25.7	2.000	8	2	3	25.4	2.250	4	1	1	30.2	3.275	2	2	3	25.8	2.200	0
2	1	26.7	2.700	5	3	3	25.7	1.200	0	3	3	26.2	2.225	0	3	3	24.1	1.800	0
4	3	23.7	1.850	0	2	3	25.1	2.100	5	2	3	24.2	1.650	2	3	3	26.2	2.175	2
2	3	26.8	2.650	0	3	2	24.5	2.250	0	2	3	27.4	2.900	3	3	3	26.1	2.750	3
2	3	27.5	3.150	6	4	3	27.5	2.900	0	2	2	25.4	2.300	0	3	3	29.0	3.275	4
4	3	23.4	1.900	0	3	3	23.1	1.650	0	3	3	28.4	3.200	3	1	1	28.0	2.625	0
2	3	27.9	2.800	6	3	1	25.9	2.550	4	4	3	22.5	1.475	4	4	3	27.0	2.625	0
3	3	27.5	3.100	3	2	3	25.8	2.300	0	2	3	26.2	2.025	2	2	2	24.5	2.000	0
1	1	26.1	2.800	5	4	3	27.0	2.250	3	2	1	24.9	2.300	6					
1	1	27.7	2.500	6	2	3	28.5	3.050	0	1	2	24.5	1.950	6					
2	1	30.0	3.300	5	4	1	25.5	2.750	0	2	3	25.1	1.800	0					

注釈:　color=色（1=やや明るい，2=中くらい，3=やや暗い，4=暗い）;

Spine=後体部の棘の状態（1=いずれも正常，2=一方が摩耗または破損している，3=いずれも摩耗または破損している）;

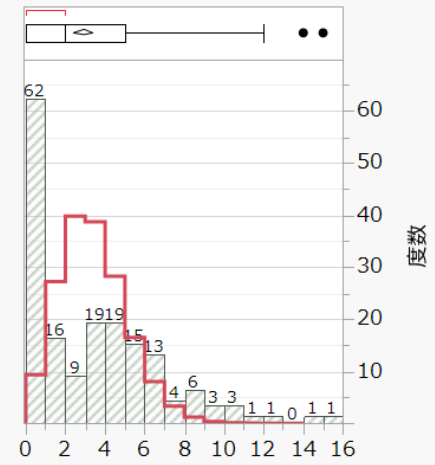
width=甲羅の幅(cm) ;　　weight =重さ（kg）;　　satell＝サテライト数.

出典: Web at　http://lib.stat.cmu.edu/datasets/agresti　2020 年 4 月 17 日アクセス

表 1.46　サテライト数へのポアソン分布のあてはめ

一変量の分布

サテライト数



要約統計量

平均	2.9191
標準偏差	3.1483
平均の標準誤差	0.2394
平均の上側95%	3.3915
平均の下側95%	2.4466
N	173.0000
分散	9.9120
修正平方和	1704.8671

Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値
尺度	λ	2.9191

指標	
(-2)*対数尤度	988.0893
AICc	990.1127
BIC	993.2426

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
584.0436	<.0001*

表 1.47　甲羅の幅 width についての対数リンクによるポアソン回帰

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	32.4565	64.9131	1	<.0001*
完全	461.5881			
縮小	494.0447			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)
Pearson	544.1570	171	<.0001*
デビアンス	567.8786	171	<.0001*

AICc
927.2468

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)
切片	-3.3048	0.5422	36.8670	<.0001*
甲羅の幅	0.1640	0.0200	64.9131	<.0001*

推定値の共分散

共分散

	切片	甲羅の幅
切片	0.2940	-0.0108
甲羅の幅	-0.0108	0.0004

ら得られた過分散パラメータを $\varphi = 3.1822$ とし，得られた共分散行列を φ 倍して標準誤差を調整する方法が知られていて，JMP のポアソン回帰でもサポートされている．

過分散を調整したポアソン回帰の結果を表 1.48 に示す．

表 1.47 の甲羅の幅の標準誤差は，$SE = 0.0200$ であったので，調整後の SE' は，

$$SE' = \sqrt{\varphi SE^2} = \sqrt{3.1822 \times 0.0200^2}$$

となり，尤度比カイ 2 乗値は，64.9131 から 20.3988 と激減する．

表 1.48　甲羅の幅 width についての対数リンクの過分散調整済みのポアソン回帰

手法: 一般化線形モデル
分布: Poisson
リンク関数: 対数
☑ 過分散に基づく検定と信頼区間
☐ Firthバイアス調整推定値

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	10.1993888	20.3988	1	<.0001*
完全	145.05293			
縮小	155.252319			

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	544.1570	171	<.0001*	3.1822
デビアンス	567.8786	171	<.0001*	

AICc
296.2479

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	-3.3048	0.9673	11.5854	0.0007*
甲羅の幅	0.1640	0.0356	20.3988	<.0001*

過分散の調整を甲羅の幅の標準誤差で例示したが，調整は共分散行列をφ倍することにより，回帰直線の 95%信頼区間に対しても調整されることになる．

表 1.49　過分散係数を用いた共分散の調整

	パラメータの共分散			過分散		調整後の共分散			標準誤差 *SE*	
	切片	甲羅の幅		φ		切片	甲羅の幅		切片	甲羅の幅
切片	0.2940	-0.0108	*	3.1822	=	0.9357	-0.0343		0.9673	
甲羅の幅	-0.0108	0.0004				-0.0343	0.0013			0.0356

過分散の係数を用いた方法は，過分散となるカウント・データに対する万能の方法とも思われるかもしれないが，表 1.46 に示したヒストグラムに重ね書きしたポアソン分布から，このデータにポアソン分布を仮定することは全くできない．甲羅の幅に対するポアソン回帰によって過分散が解消するのであれば嬉しいのであるが，過分散は解消されていなかった．

ポアソン回帰を行っても過分散が解消していないことを視覚化するために散布図に個別データの 95%信頼区間を重ね書きしてみると，図 1.15 左に示すよう外側に多数の点がはみ出ていることによりポアソン回帰のあてはめには無理があることを実感できる．図 1.15 右に示すように予測値に対する pearson 残差をプロットすることにより，Pearson 残差が 3 以上の飛び離れデータが多数存在することからも，ポアソン分布を誤差分布とする回帰分析について否定的な結果となっている．

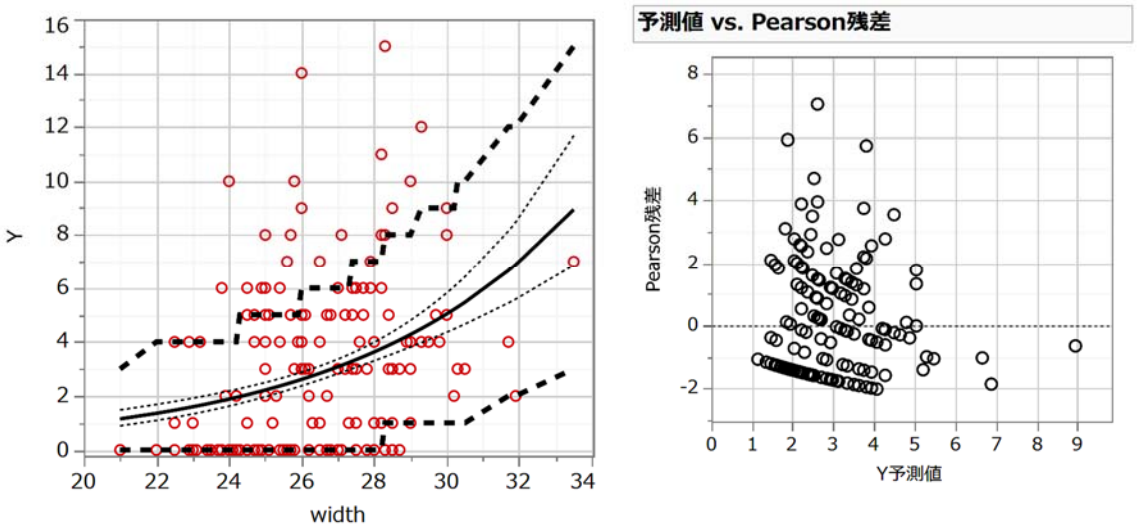


図 1.15　ポアソン回帰に対する 95%信頼区間および予測値に対する pearson 残差

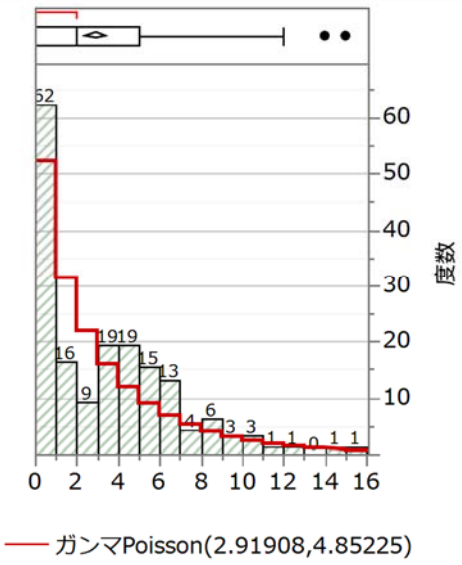
他の変数を加えてポアソン回帰を行っても過分散が解消されないのであれば，ポアソン回帰を行う前提がないことになる．この原因は，173 個体に対してサテライト数がゼロに 62 件と全体の 35.8%を占め，サテライト数が 3 と 4 あたりに分布の山があることから，アグレスティ(2003)にあるサテライト数を（0，1）反応での解析が望ましいのか，あるいは，3 区分程度の順序データとした解析を行うことが望ましいかも知れない．

JMP には，過分散を考慮した負の 2 項分布から導出されたガンマ・ポアソン分布をあてはめることができるので，表 1.50 に結果を示す．結果は，位置 $\lambda = 2.9191$，過分散 $\sigma = 4.8522$ となる．表 1.46 に示したポアソン分布のあてはめでは，サテライト数が 0 の場合について大きな乖離があったが，過分散を考慮したガンマ・ポアソン分布のあてはめでは，適当なあてはめが行われているように思われる．ただし，サテライト数が 1 および 2 については，元々のデータの出現頻度が小さいため，ガンマ・ポアソン分布のあてはめも無理なのかも知れない．なお，過分散を考慮した解析方法については，第 6 で取り上げる．

雌のカブトガニに雄が多数連結することは確かなようであるが，単独でいる雌のカブトガニの数が多いことから，雄が雌に連結する場合も合わせてポアソン分布を仮定することは無理のようである．（連結していない，数匹が連結している，かなり連結してる）などの順序尺度としての扱が適してるかもしれない．第 7.2 節で，このデータを基づいて探索的な解析を行った結果を示す．

表 1.50　サテライト数へのガンマ・ポアソン分布のあてはめ

サテライト数



ガンマPoissonのあてはめ

パラメータ推定値

種類	パラメータ	推定値	下側95%CL	上側95%CL
位置	λ	2.9191	2.4136	3.5571
過分散	σ	4.8522	3.6559	6.6859

指標	
(-2)*対数尤度	767.4092
AICc	771.4798
BIC	777.7158

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
120.3656	0.9990

注: Ho = ガンマPoisson分布からのデータ。 p値が小さい場合はHoを棄却。

SAS の GENMOD プロシジャには，ゼロの割合を考慮するポアソン回帰，ガンマ・ポアソン回帰の 2 種類の Zero-Inflated Model が用意されている．これらにいついても第 6 章で取り上げる．

偶数ページ

第１章　文 献 索　引

アグレスティ著，渡邉・菅波・吉田・角野・寒水・松永 訳（2003）　- カテゴリカルデー解析入門　56
アーミテジ著，椿・椿 共訳（2001）　- 医学研究のための統計的方法　46
アルトマン著，木船・佐久間 訳（1999）　- 医学研究における実用統計学　27
Cameron and Trivde （1998）　- Regression Analysis od Count Data　54
久保(2012）‐データ解析のための統計モデリング入門，一般化線形モデル･階層ベイズモデル･MCMC　41
スネデカー・コクラン著，畑村・奥野・津村 訳（1972）　- 統計的方法，第6版　13
ドブソン著，田中・森川・山中・富田 訳（2008）　- 一般化線形モデル入門，原著 第2版　16
富山・杉本（2004）　- 細菌を用いた用量反応試験データ　36
吉村・大橋 責任編集（1992）　- 毒性試験データの統計解析　32

第１章　索　引

あ　アグレスティ（2003）　- カブトガニ　56
アーミテジ（2001）　- 退役軍人の癌の発生　46
R　- glm()　41
アルトマン（1999）　- 新月と満月　27
医院への通院回数　- 過分散　54
- Cameron and Trivde（1998）　54
一変量の分布　- JMP　33
逸脱度　- デビアンス　44
一般化線形モデル　- デザイン行列　18
- ポアソン回帰　16
- 二項分布　26
Ames試験　- ネズミチフス菌　32
- 復帰突然変異試験　32
- 吉村ら（1992)　32
- 変異コロニー数　36
Excel　- 絶対参照　8
- 相対参照　8
- Poisson.dist() 関数　8
- Binom.dist() 関数　12
- Chisq.dist() 関数　14
- Mmult() 関数　18
- Minverse() 関数　19
- 反復重み付き回帰　19
- Tanspose() 関数　20
- 反復計算　21
- 尤度比検定　29
- ソルバー　43
- ソルバー　50
Minverse() 関数　- Excel　19
Mmult() 関数　- Excel　18
オフセット　- JMP　24
- ポアソン回帰　25
- 対数　47
か　Chisq.dist() 関数　- Excel　14
確率関数　- ポアソン分布　8
確率楕円　- 散布図　41
カブトカニ　- アグレスティ（2003）　56
カブトガニ　- サテライト数　56
過分散　- ポアソン分布　7
- 医院への通院回数　54
- 調整　58
- 負の2項分布　60
Cameron and Trivde（1998）‐医院への通院回数　54
冠動脈心疾患　- ドブソン（2008）　23
- ドブソン（2008）　49
完全モデル　- モデル　43
癌の発生　- 退役軍人　46
ガンマ・ポアソン分布　- 負の2項分布　54
- 負の2項分布　60
期待値　- ポアソン分布　10
95%信頼区間　- ポアソン回帰　22
- 個別の95%信頼区間　44
共分散行列　- 信頼区間　21
- パラメータ　21
- SAS　37
共分散　- 推定値　16
久保（2012）　- 植物の体サイズ　41
glm()　- R　41
交通事故　- ポアソン分布　10
恒等リンク　- ポアソン回帰　16
効力比　- 分散　37
- デルタ法　38
個別の95%信頼区間　- 95%信頼区間　44
コロニー数　- ネズミチフス菌　32
さ　細菌　- 用量反応性試験　36
細菌を用いた試験　- 2×2要因配置　32
最大化　- 対数尤度　51
SAS　- GENMOD　30
- 無償　31
- 共分散行列　37
- ポアソン回帰　37
雑草の種子　- ポアソン分布　13
- 有害種子　14
サテライト数　- カブトガニ　56
散布図　- 確率楕円　41
GENMOD　- SAS　30
- ポアソン回帰　30
- Waldカイ2乗　31
シグモイド曲線　- ロジスティック曲線　26
死亡者数　- 10万人比　23
Shapiro-WilkのW検定　- 正規分布　34
JMP　- ポアソン分布　15
- ポアソン回帰　16
- オフセット　24
- 二変量の関係　27
- プロファイル尤度　28
- 一変量の分布　33
10万人比　- 死亡者数　23

縮小モデル - モデル 43
種子数 - 体サイズ 41
順位和検定 - 新月と満月 28
植物の体サイズ - 久保（2012） 41
新月と満月 - アルトマン（1999） 27
- 順位和検定 28
信頼区間 - 共分散行列 21
推定値 - 共分散 16
スチューデント化 - デビアンス残差 45
スネデカーら（1972） - 有害雑草の種 13
正規性 - 適合度検定 34
正規分布 - Shapiro-WilkのW検定 34
- W検定 34
絶対参照 - Excel 8
切片を共通 - 2本の回帰直線 36
相対参照 - Excel 8
ソルバー - Excel 43
- Excel 50
た 対数尤度 - 最大化 51
対数リンク - ポアソン回帰 42
体サイズ - 種子数 41
対数 - オフセット 47
退役軍人 - 癌の発生 46
退役軍人の癌の発生 - アーミティジ（2001） 46
代謝活性化 - DMOS 33
代替物質T - 陽性対照薬S 37
適合度検定 - Pearsonのカイ2乗 15
適合度統計量 - デビアンス 44
- Perarson残差 44
適合度の検定 - ポアソン分布 14
適合度検定 - 正規性 34
デザイン行列 - 一般化線形モデル 18
- 2次形式 22
デビアンス - 逸脱度 44
- 適合度統計量 44
デビアンス残差 - スチューデント化 45
DMOS - 代謝活性化 33
デルタ法 - 効力比 38
等分散性の検定 - Bartlettの検定 34
ドブソン（2008） - ポアソン回帰 16
- 冠動脈心疾患 23
- 冠動脈心疾患 49
富山ら（2004） - 用量反応試験 36
Tanspose() 関数 - Excel 20
な 2×2要因配置 - 細菌を用いた試験 32
2階の偏微分 - ヘッセ行列 18
2群間比較 - ポアソン回帰 28
- 尤度比検定 29
2項分布 - ポアソン分布 10
二項分布 - 一般化線形モデル 26
2次形式 - デザイン行列 22
二変量の関係 - JMP 27
ネズミチフス菌 - Ames試験 32
- コロニー数 32
は Bartlettの検定 - 等分散性の検定 34
Binom.dist() 関数 - Excel 12
パラメータ - 共分散行列 21
反復計算 - Excel 21
反復重み付き回帰 - ポアソン回帰 16
- Excel 19
比 - 分散/平均 38
Pearsonのカイ2乗 - 適合度検定 15
Pearson残差 - プロット 59
Perarson残差 - 適合度統計量 44
復帰突然変異試験 - Ames試験 32
負の2項分布 - ガンマ・ポアソン分布 54
- 過分散 60
- ガンマ・ポアソン分布 60
プロット - Pearson残差 59
プロファイル尤度 - JMP 28
分散 - ポアソン分布 11
分散/平均 - 比 38
分散 - 効力比 37
ヘッセ行列 - 2階の偏微分 18
変異コロニー数 - Ames試験 36
変動係数 - ポアソン分布の形状 9
Poisson.dist() 関数 - Excel 8
ポアソン回帰 - 一般化線形モデル 16
- 恒等リンク 16
- JMP 16
- ドブソン（2008） 16
- 反復重み付き回帰 16
- 95%信頼区間 22
- オフセット 24
- 2群間比較 28
- GENMOD 30
- SAS 37
- 対数リンク 42
ポアソン分布 - 過分散 7
- 確率関数 8
- 期待値 10
- 交通事故 10
- 2項分布 10
- 分散 11
- 雑草の種子 13
- 適合度の検定 14
- JMP 15
ポアソン分布の形状 - 変動係数 9
飽和モデル - モデル 43
ま 満月と新月 - 尤度比カイ2乗検定 28
無償 - SAS 31
モデル - 完全モデル 43
- 縮小モデル 43
- 飽和モデル 43
や 有害種子 - 雑草の種子 14
尤度比カイ2乗検定 - 満月と新月 28
尤度比検定 - Excel 29
- 2群間比較 29
有害雑草の種 - スネデカーら（1972） 13
用量反応性試験 - 細菌 36
用量反応試験 - 富山ら（2004） 36
陽性対照薬S - 代替物質T 37
吉村ら（1992） - Ames試験 32
ら リンク関数 - ロジット 25
ロジスティック曲線 - シグモイド曲線 26
ロジット - リンク関数 25
Waldカイ2乗 - GENMOD 31

第1章　解析用ファイル一覧

55 KB	第1章01_02_ポアソン確率	Microsoft Excel ワークシート
4 KB	第1章03_種子数	JMP Data Table
16 KB	第1章03_種子数	Microsoft Excel ワークシート
5 KB	第1章04_人工データ	JMP Data Table
31 KB	第1章04_人工データ	Microsoft Excel ワークシート
4 KB	第1章05_冠動脈疾患	JMP Data Table
15 KB	第1章05_冠動脈疾患	Microsoft Excel ワークシート
3 KB	第1章05_冠動脈疾患01反応	JMP Data Table
5 KB	第1章05_冠動脈疾患01反応グラフ	JMP Data Table
3 KB	第1章06_満月新月	JMP Data Table
1 KB	第1章06_満月新月.sas	テキスト ドキュメント
115 KB	第1章06_満月新月	Microsoft Excel ワークシート
22 KB	第1章07_細菌2x2	JMP Data Table
106 KB	第1章07_細菌2x2	Microsoft Excel ワークシート
6 KB	第1章08_変異原性試験	JMP Data Table
1 KB	第1章08_変異原性試験.sas	テキスト ドキュメント
19 KB	第1章08_変異原性試験	Microsoft Excel ワークシート
10 KB	第1章09_久保_種子	JMP Data Table
56 KB	第1章09_久保_種子	Microsoft Excel ワークシート
10 KB	第1章09_久保_種子_Cグラフ化	JMP Data Table
6 KB	第1章10_軍人_癌	JMP Data Table
22 KB	第1章10_軍人_癌	Microsoft Excel ワークシート
5 KB	第1章11_タバコと冠動脈疾患	JMP Data Table
34 KB	第1章11_タバコと冠動脈疾患	Microsoft Excel ワークシート
5 KB	第1章12_通院回数	JMP Data Table
14 KB	第1章12_通院回数	Microsoft Excel ワークシート
17 KB	第1章13_カブトガニ	JMP Data Table
24 KB	第1章13_カブトガニ	Microsoft Excel ワークシート
20 KB	第1章13_カブトガニ_探索	JMP Data Table

第 9 回 続高橋セミナー

最尤法によるポアソン回帰分析入門 ＜＜第 1 章＞＞

第 1 章　ポアソン分布に従う各種のカウント・データ

BioStat 研究所（株）

〒105-0014 東京都 港区 芝 1-12－3 の 1005

2020 年 4 月 19 日　　高橋 行雄

takahashi.stat@nifty.com ， FAX : 03-342-8035