これも、難しいことにチャレンジしています。テーマは、アメフットの練習に伴う、ケガの頻度が何によって決まるのか、それを調べて、ケガの頻度を低下させたいということです。分析手法としては、ロジスティック回帰とポアソン検定の２つを使っています。ロジスティック回帰とポアソン検定を使うというのは、妥当な判断だと思いますが、思うような結果は得られていません。回帰の有意性の判断に問題があるのですが、その前に、例によってロジスティック回帰と、ポアソン検定について説明します。

こういう統計分析的な話の数式的な説明に慣れている人と、あまり慣れていない人がいると思うので、数式の意味を例を上げて説明します。説明が回りくどくて煩わしいと思う人は、そこを飛ばして読んでください。

ケガの頻度というのは、ケガをする確率のことです。こういう分析では、ケガをする（１)、ケガをしない（０）という、あるかないか、０か１かのデータ分布になります。あることが起きるか、起こらないかということを、被説明変数にして、重回帰分析しようとすると、０と１の値しかありませんから、回帰のしようがありません。そういう場合に、ロジスティック回帰が使われます。何かが起きる確率と起こらない確率の比（オッズ比）の対数に対して、確率をプロットすると、つまり、オッズ比の対数を説明変数（横軸）として、被説明変数pを縦軸にとったグラフを描くと、正規分布とよく似た形のグラフが得られます。この曲線は、正規分布曲線の様に、左右対称で、中央が尖った、釣り鐘型をした確率密度分布をしています。正規分布よりは、少し、幅が広いのですが、よく似ています。この確率密度のグラフを積分して、累積確率のグラフに表すと、正規分布の累積確率のグラフと同じようにＳ字型の曲線になります。pは$0 \leq p \leq 1$の範囲で分布しますから、$\frac{p}{1-p}$は$0 \leq \frac{p}{1-p} < \infty$の範囲で分布し、$\ln\frac{p}{1-p}$は$-\infty < \ln\frac{p}{1-p} < \infty$で分布します。具体的に計算してもらわないとわからないという人のために、エクセルで計算してみます（図８)。図８の左側の表は、計算のために使った、エクセルシートです。表の１番左の列がＹ＝１である確率（この場合はケガする確率)、２番目がY=0である確率（この場合はケガしない確率)、３番目がそのオッズ比、４番目がオッズ比の対数です。４番目のオッズ比の対数にたいしてＹ＝１である確率をプロットしたのが、右側のグラフです。正規分布の累積確率のグラフに似ているでしょう。ロジスティック回帰でやろうとしていることは、この、４番目のオッズ比の対数を、被説明変数として、データとして得られている説明変数の多項式で近似する式を作ろうということです、。つまり、Ｙを説明する式を直接つくることはできないから、中間のオッズ比の対数という被説明変数を考えて、その近似式を作るということです。この式には、いくつかの説明変数で表された線形の式を考えます。線形でない式も考えられるでしょうが、解法が難しくなって、私には解けません。ということで、ここでは線形の式を考えます。とりあえず、その式を、$f(x_1 \cdots x_r)$と表しておきます。説明変数がｒ個あるという意味です。今までの説明を数式で表しておきます。説明変数を識

別する添え字はhとして、データを識別する添え字はiとして、データ数はnとしておきます。（数学を得意とする人には、クドイ説明で申し訳ありません。）

データiは、$Y_i = 1$あるいは$Y_i = 0$ですが、データiが$Y_i = 1$である確率をp_iだとすれば、データiが$Y_i = 0$である確率は$1 - p_i$です。オッズ比は$OR_i = \frac{p_i}{1-p_i}$で、その対数は$\ln\left(\frac{p_i}{1-p_i}\right)$ですから、

$$\ln\left(\frac{p_i}{1-p_i}\right) = f(x_{i1} \cdots x_{ir}) \qquad \text{式 1}$$

とするという話です。これをp_iを表す式に書き換えます。

$$\frac{p_i}{1-p_i} = e^{f(x_{i1} \cdots x_{ir})}$$

ですから、

$$p_i = (1-p_i)e^{f(x_{i1} \cdots x_{ir})}$$

$$p_i = e^{f(x_{i1} \cdots x_{ir})} - p_i e^{f(x_{i1} \cdots x_{ir})}$$

$$p_i + p_i e^{f(x_{i1} \cdots x_{ir})} = e^{f(x_{i1} \cdots x_{ir})}$$

$$p_i\left(1 + e^{f(x_{i1} \cdots x_{ir})}\right) = e^{f(x_{i1} \cdots x_{ir})}$$

$$p_i = \frac{e^{f(x_{i1} \cdots x_{ir})}}{1 + e^{f(x_{i1} \cdots x_{ir})}}$$

となりますが、分母、分子を、$e^{f(x_1 \cdots x_r)}$で割って、

$$p_i = \frac{1}{1 + e^{-f(x_{i1} \cdots x_{ir})}}$$

という式になって、この曲線をロジスティック曲線と言います。蛇足ですが、

$$f(x) = \frac{1}{1 + e^{-ax}}$$

という式で表される、曲線は図８のように、Ｓ字型の曲線（シグモイド曲線）になります。ある一定値に達すると、それ以上は増加しないという現象は、いろいろなところにあります。ですから、この形の曲線は統計学上の累積頻度分布だけでなく、生態学の環境収容力とか、いろいろな分野で出てくるので、式とグラフの形を記憶しておくと良いでしょう。

Y = 1	Y = 0	オッズ比	対数オッズ比	
p	p-1	p/p-1	ln(p/1-p)	p
0	1	0	#NUM!	0
0.01	0.99	0.010101	-4.59511985	0.01
0.02	0.98	0.020408	-3.8918203	0.02
0.03	0.97	0.030928	-3.47609869	0.03
0.04	0.96	0.041667	-3.17805383	0.04
0.05	0.95	0.052632	-2.94443898	0.05
0.06	0.94	0.06383	-2.75153531	0.06
0.07	0.93	0.075269	-2.58668934	0.07
0.08	0.92	0.086957	-2.44234704	0.08
0.09	0.91	0.098901	-2.31363493	0.09
0.1	0.9	0.111111	-2.19722458	0.1
0.11	0.89	0.123596	-2.0907411	0.11
0.12	0.88	0.136364	-1.99243016	0.12
0.13	0.87	0.149425	-1.90095876	0.13
0.14	0.86	0.162791	-1.81528997	0.14
0.15	0.85	0.176471	-1.73460106	0.15
0.16	0.84	0.190476	-1.65822808	0.16
0.17	0.83	0.204819	-1.58562726	0.17
0.18	0.82	0.219512	-1.51634749	0.18
0.19	0.81	0.234568	-1.45001018	0.19
0.2	0.8	0.25	-1.38629436	0.2
0.21	0.79	0.265823	-1.32492541	0.21
0.22	0.78	0.282051	-1.26566637	0.22
0.23	0.77	0.298701	-1.20831121	0.23
0.24	0.76	0.315789	-1.15267951	0.24
0.25	0.75	0.333333	-1.09861229	0.25
0.26	0.74	0.351351	-1.04596856	0.26
0.27	0.73	0.369863	-0.99462258	0.27
0.28	0.72	0.388889	-0.94446161	0.28
0.29	0.71	0.408451	-0.89538405	0.29
0.3	0.7	0.428571	-0.84729786	0.3
0.31	0.69	0.449275	-0.8001193	0.31
0.32	0.68	0.470588	-0.7537718	0.32
0.33	0.67	0.492537	-0.70818506	0.33
0.34	0.66	0.515152	-0.66329422	0.34
0.35	0.65	0.538462	-0.61903921	0.35
0.36	0.64	0.5625	-0.57536414	0.36
0.37	0.63	0.587302	-0.53221681	0.37
0.38	0.62	0.612903	-0.48954823	0.38
0.39	0.61	0.639344	-0.44731222	0.39
0.4	0.6	0.666667	-0.40546511	0.4
0.41	0.59	0.694915	-0.36396538	0.41
0.42	0.58	0.724138	-0.32277339	0.42
0.43	0.57	0.754386	-0.28185115	0.43
0.44	0.56	0.785714	-0.24116206	0.44
0.45	0.55	0.818182	-0.2006707	0.45
0.46	0.54	0.851852	-0.16034265	0.46
0.47	0.53	0.886792	-0.12014431	0.47
0.48	0.52	0.923077	-0.08004271	0.48
0.49	0.51	0.960784	-0.04000533	0.49
0.5	0.5	1	8.88178E-16	0.5
0.51	0.49	1.040816	0.040005335	0.51
0.52	0.48	1.083333	0.080042708	0.52
0.53	0.47	1.12766	0.120144312	0.53
0.54	0.46	1.173913	0.16034265	0.54
0.55	0.45	1.222222	0.200670695	0.55
0.56	0.44	1.272727	0.241162057	0.56
0.57	0.43	1.325581	0.281851152	0.57
0.58	0.42	1.380952	0.322773392	0.58
0.59	0.41	1.439024	0.363965377	0.59
0.6	0.4	1.5	0.405465108	0.6
0.61	0.39	1.564103	0.447312218	0.61
0.62	0.38	1.631579	0.489548225	0.62
0.63	0.37	1.702703	0.532216814	0.63
0.64	0.36	1.777778	0.575364145	0.64
0.65	0.35	1.857143	0.619039208	0.65
0.66	0.34	1.941176	0.663294217	0.66
0.67	0.33	2.030303	0.708185058	0.67
0.68	0.32	2.125	0.753771802	0.68
0.69	0.31	2.225806	0.8001193	0.69
0.7	0.3	2.333333	0.84729786	0.7
0.71	0.29	2.448276	0.895384047	0.71
0.72	0.28	2.571429	0.944461609	0.72
0.73	0.27	2.703704	0.994622575	0.73
0.74	0.26	2.846154	1.045968555	0.74
0.75	0.25	3	1.098612289	0.75
0.76	0.24	3.166667	1.15267951	0.76
0.77	0.23	3.347826	1.208311206	0.77
0.78	0.22	3.545455	1.265666373	0.78
0.79	0.21	3.761905	1.324925415	0.79
0.8	0.2	4	1.386294361	0.8
0.81	0.19	4.263158	1.450010176	0.81
0.82	0.18	4.555556	1.516347489	0.82
0.83	0.17	4.882353	1.585627264	0.83
0.84	0.16	5.25	1.658228077	0.84
0.85	0.15	5.666667	1.734601055	0.85
0.86	0.14	6.142857	1.815289967	0.86
0.87	0.13	6.692308	1.900958761	0.87
0.88	0.12	7.333333	1.992430165	0.88
0.89	0.11	8.090909	2.090741097	0.89
0.9	0.1	9	2.197224577	0.9
0.91	0.09	10.11111	2.313634929	0.91
0.92	0.08	11.5	2.442347035	0.92
0.93	0.07	13.28571	2.586689344	0.93
0.94	0.06	15.66667	2.751535313	0.94
0.95	0.05	19	2.944438979	0.95
0.96	0.04	24	3.17805383	0.96
0.97	0.03	32.33333	3.47609869	0.97
0.98	0.02	49	3.891820298	0.98
0.99	0.01	99	4.59511985	0.99
1	0	#DIV/0!	#DIV/0!	1

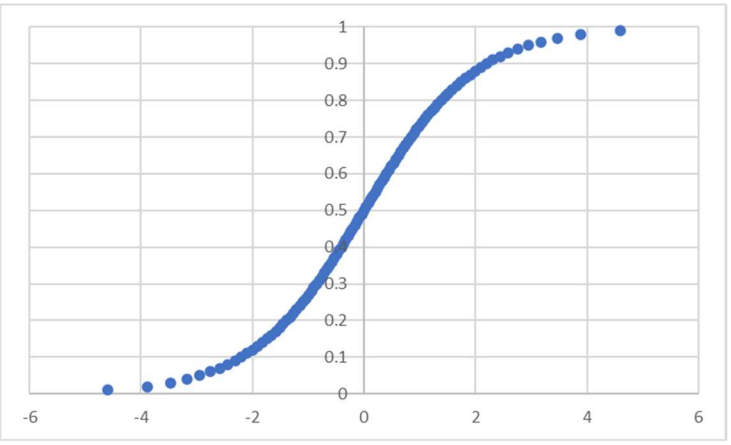

図８。ロジスティック曲線の描画

ここで我々が考えなくてはならないのは、この$f(x_1 \cdots x_r)$という式を最適化するにはどうすれば良いのかということです。回帰分析も係数の最適化ですが、回帰分析で良く用いられているのは、最小二乗法ですね、実測値と式から予測される予測値の差の２乗の和を最小にするように、$f(x_1 \cdots x_r)$の係数を最適化するという方法です。最近は、最尤法（式から予測される値が正しい確率を最大化する）が用いられることが多いと思います。実は、この事例の場合、最小二乗法でも、最尤法でも、実は同じ計算式になるのですが、ここでは、最尤法で説明します。p_iはデータが$Y_i = 1$である確率ですから、$Y_i = 1$であったときには、p_iが正しく予想で来ている確率です。$Y_i = 0$あった時には、$1 - p_i$がデータが正しい確率です。${p_i}^1 = p_i$、
${p_i}^0 = 1$、$(1 - p_i)^1 = 1 - p_i$、$(1 - p_i)^0 = 1$、${p_i}^0 = 1$ですから、それらを利用して。それが正しい確率（尤度:l）を表すと、

$$l_i = {p_i}^{Y_i}(1 - p_i)^{1-Y_i} = \left(\frac{1}{1 + e^{-f(x_{i1} \cdots x_{ir})}}\right)^{Y_i}\left(1 - \frac{1}{1 + e^{-f(x_{i1} \cdots x_{ir})}}\right)^{1-Y_i}$$

$$= \left(\frac{1}{1 + e^{-f(x_{i1} \cdots x_{ir})}}\right)^{Y_i}\left(\frac{e^{-f(x_{i1} \cdots x_{ir})}}{1 + e^{-f(x_{i1} \cdots x_{ir})}}\right)^{1-Y_i}$$

$$= \frac{\left(e^{-f(x_{i1} \cdots x_{ir})}\right)^{1-Y_i}}{1 + e^{-f(x_{i1} \cdots x_{ir})}}$$

と一般化して書けます。全ての推測値が正しい確率はそれらすべての積ですから。全体の尤度：Lは

$$L = \prod_{i=1}^{n} l_i = \prod_{i=1}^{n} \frac{\left(e^{-f(x_{i1} \cdots x_{ir})}\right)^{1-Y_i}}{1 + e^{-f(x_{i1} \cdots x_{ir})}}$$

となります、これは珪砂が大変だから、両辺の対数を取って。対数尤度$\log L$を最大化するという最適化法を使うと思います。

$$\log L = \ln \prod_{i=1}^{n} \frac{\left(e^{-f(x_{i1} \cdots x_{ir})}\right)^{1-Y_i}}{1 + e^{-f(x_{i1} \cdots x_{ir})}} = \sum_{i=1}^{n} \ln \frac{\left(e^{-f(x_{i1} \cdots x_{ir})}\right)^{1-Y_i}}{1 + e^{-f(x_{1i} \cdots x_{ir})}}$$

$$= \sum_{i=1}^{n} \left(\ln\left(e^{-f(x_{i1} \cdots x_{ir})}\right)^{1-Y_i} - \ln\left(1 + e^{-f(x_{i1} \cdots x_{ir})}\right)\right)$$

$f(x_{i1} \cdots x_{ir})$という線形の多項式を具体的に書くと、

$$f(x_{i1} \cdots x_{ir}) = a_0 + a_1 x_{i1} + \cdots + a_r x_{ir}$$

という形をしています。このa_0、a_1、$\cdots a_r$最適化して、$\log L$を代々化するa_0、a_1、$\cdots a_r$を求めるというのが、私たちがしなければならない計算ということになります。最大化問題なのだから、極値を与えるa_0、a_1、$\cdots a_r$を推定すれば良いので、a_0、a_1、$\cdots a_r$微分して、その微分値が 0 になるようにすればよいというわけで、偏微分して、偏微分式を 0 にする

連立方程式を作って、それを解けば良いというのが答えですが、式が複雑な形になれば、簡単に解けないかもしれません。そういう場合によく使われる数値計算の方法に、ニュートン法というのがあります。便利な方法ですが、それでいつでも解けるとは限りません。初期値の与え方が悪いと、谷間の所に計算が落ち込んでそこから出られなくなってしまうこともあります。ということで、実際の計算は手数がかかるので、R や Python が使えるのならば、ネットでスクリプトを見つけて、それで解く方が現実的でしょう。ニュートン法についても、ネットで探せば解説があると思います。そこまでは説明を始めると収拾がつかなくなるので、ニュートン法の説明は省略します。

もう一つ問題になるのは、ここで問題になる、けがする率というのは、かなり低い率だろうということです。めったに起こらない現象の、期待値はポアソン分布するということが知られています。ポアソン分布というのは、フランスの数学者、シメオン・ドニ・ポアソンが 1838 年に発表した、離散型（例えばある現象が起こるか、起こらないか）の現象を、無限回繰り返した時に、n回に着き何回、起きるかということを表した分布です。これが最初に使われたのは、軍隊で兵隊が馬に蹴られて死ぬ確率についての分析だったのだと習ったことがあるような気がしますが、誰に教わったのかは忘れました。その話が本当かどうかも確かめたことがありません。まあ、そういう現象についての話だと思ってください。

離散型の現象の確率に関する統計解析で、よく出てくるのは二項分布です。こちらの方が説明が簡単なので、こちらを先に説明します。よくある一番簡単な例は、コインを 3 回投げたら、表が出てくる回数は、0 回、1 回、2 回、3 回と 4 通りあるけれど、そのぞれぞれの回数になる確率はどのように表せるのかということです。一回投げた時に、表が出るのは 1/2 の確率で、裏が出るのも 1/2 の確率だから。3 回続けて、1 回も表が出ない（3 回裏が出る）のは、$\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3=\frac{1}{8}$、

1 回表が出るのは

1 回目に表が出て、2 回目、3 回目は裏

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

1 回目に裏が出て、2 回目に表、3 回目は裏

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

1 回目に裏が出て、2 回目も裏、3 回目に表

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

これらを足し合わせて、

1 回表が出るのは$\frac{3}{8}$

2 回表が出るのは

1 回目に表が出て、 2 回目も表、3 回目は裏

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

1 回目に表が出て、 2 回目に裏、3 回目は表

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

1 回目に裏が出て、 2 回目も裏、3 回目に表

$$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{8}$$

これらを足し合わせて、

2 回表が出るのは$\frac{3}{8}$

3 回表が出る）のは、$\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^0=\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0=\frac{1}{8}$

となって、効果の表が何枚出るかという確率の分布は図 9 のようになります。

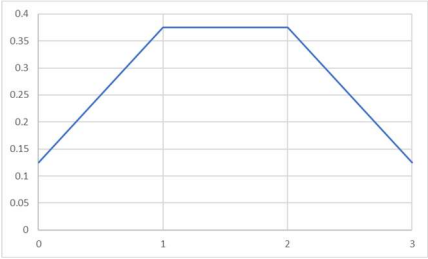

図 9　効果を 3 投げた時に表が出る確率の分布

コインの場合は、表になる確率は、等しく 2 分の 1 づつですが（イカサマでなければ)、一般的に、何かである確率をp、そうでない確率は$1-p$と表します。これを使って、確率pになることを、n回繰り返した時に、確率pになる現象が k 回現れる確率は、

$$B(n,p,k)=\binom{n}{k}p^k(1-p)(1-p)^{n-k}$$

$\binom{n}{k}$はn回繰り返した時に k 回現れる計算の用いる係数で二項係数と言います。2 項係数は次のように表せます。

$$\binom{n}{k}=\left(\frac{n!}{(n-k)!\,k!}\right)$$

例えば、サイコロを 1 回振った時に、1 が出てくる確率を$\frac{1}{6}$として、5 回サイコロを振った時に、1 が 1 回も出ない確率、1 回出る確率、2 回出る確率、・・・・・・5 回出る確率を計算しみてください。

	1が出る回数	1以外が出る回数	$\binom{n}{k}$	p^k	$(1-p)^{n-k}$	$B(n,p,k)$
	0	5	$\frac{5!}{(5-0)!0!}=1$	$\left(\frac{1}{6}\right)^0=1$	$\left(\frac{5}{6}\right)^5=\frac{3125}{7776}$	$\frac{3125}{7776}$
	1	4	$\frac{5!}{(5-1)!1!}=5$	$\left(\frac{1}{6}\right)^1=\left(\frac{1}{6}\right)$	$\left(\frac{5}{6}\right)^4=\frac{625}{1296}$	$\frac{3125}{7776}$
	2	3	$\frac{5!}{(5-2)!2!}=10$	$\left(\frac{1}{6}\right)^2=\left(\frac{1}{36}\right)$	$\left(\frac{5}{6}\right)^3=\frac{125}{216}$	$\frac{1250}{7776}$
	3	2	$\frac{5!}{(5-3)!3!}=10$	$\left(\frac{1}{6}\right)^3=\left(\frac{1}{216}\right)$	$\left(\frac{5}{6}\right)^2=\frac{25}{36}$	$\frac{250}{7776}$
	4	1	$\frac{5!}{(5-4)!4!}=5$	$\left(\frac{1}{6}\right)^4=\left(\frac{1}{1296}\right)$	$\left(\frac{5}{6}\right)^1=\frac{5}{6}$	$\frac{25}{7776}$
	5	0	$\frac{5!}{(5-5)!5!}=1$	$\left(\frac{1}{6}\right)^5=\left(\frac{1}{7776}\right)$	$\left(\frac{5}{6}\right)^0=1$	$\frac{1}{7776}$

確率の総和は確かに1になっています。
高校の数学では、2項係数ではなくて、組み合わせの場合の数 ${}_nC_k$と教えられたかもしれませんが、どちらもおなじです。この例のように、ピークが一つだけ（単方形:unimodal）の確率分布のは、ピークの位置、横方向の広がり、歪みですが、ピークの位置を期待値と言います。この場合は平均値と同じです。横方向の広がりを表すのは標準偏差ですが、統計的には、その2乗の分散が使われます。まず期待値ですが、1回の試行でpの確率なのだから、それをn回繰り返すのだから、期待値はnpに決まっているだろうと、極めて単純に考えれば良いと思います。同じような考え方で、分散についても、1回の試行で、$Y=1$となる確率はpで、期待値が1ならば、期待値との差の2 *乗*は$(1-1)^2=0$で反対に期待値が0ならば期待値との差の2乗は$(1-p)^2$でこの二つを合わせると$(1-1)^2+(1-p)^2=(1-p)^2$これにそうなり確率を掛けて$p(1-p)^2$、Y＝0となる確率は$(1-p)$で、期待値0ならば、期待値との差の2乗は、$(0-p)^2=p^2$, 反対に期待値が1ならば$(1-1)^2=0$,
でこれを合わせて、p^2+$(1-1)^2$=p^2, そうなる確率は$(1-p)$だから、その確率を掛けて、$(1-p)p^2$,この2つを足し合わせたものが、1回の試行における、期待値になるので、1回試行における期待値は、

$$p(1-p)^2+(1-p)p^2$$
$$=p(1-p)((1-p)+p)$$
$$=p(1-p)$$

となります。これをn回繰り返したときの分散だから。
分散$V(p,n)$は

$$V(p,n)=np(1-p)$$

となります。$(1-p)=q$ とあらわすと、

$$V(p,n)=npq$$

と書けます。これが、2 項分布の基礎知識です。

シメオン・ドニ・ポアソン(1838)が問題にしたのは、「n回あたりにλ回起こる現象の、発生頻度はpですが、λを一定にして、nを無限大にしたときに、その現象が送る回数 k の牡蛎率率密度分布の式がどのようになるのか」ということです。その答えは、

$$\lim_{\lambda=np,n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=\frac{\lambda^k e^{-\lambda}}{k!}$$

です。この答えを覚えないでください。覚えても、多分、一生その知識を使うことはありません。余計なことを覚えると、重要なことを忘れます。ここで大事なことは、左辺の極限の式の中身が、２項分布の式だということです。この後の説明で、元々２項分布なのだから、２項分布的に考えれば良いのだという説明が出てきます。

一応、この式の導出だけは示しておきます。予備知識としては、次のネイピア数（自然対数の底）の定義式（定義式にはいくつかあります。）を使います。

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=e^{-\lambda}$$

極限記号の中の２項分布の確率の式を次のように変形します

$$\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{(n-k)!\,k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$=\frac{n(n-1)\cdots(n-k+1)!}{k!}\left(\frac{\lambda^k}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

入れ替え

$$=\frac{n(n-1)\cdots(n-k+1)!}{n^k}\left(\frac{\lambda^k}{k!}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$$=\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-(k-1)}{n}\right)\left(\frac{\lambda^k}{k!}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$$=(1)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\left(\frac{\lambda^k}{k!}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\lim_{n\to\infty}(1)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)=1$$

$$\lim_{n\to\infty}\left(\frac{\lambda^k}{k!}\right)=\left(\frac{\lambda^k}{k!}\right)$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=e^{-\lambda}\text{.....(ネイピア数の定義)}$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}=1$$

だから

$$\lim_{\lambda=n\ ,n\to\infty}\binom{n}{k}p^{k}(1-p)^{n-k}=\frac{\lambda^{k}}{k!}e^{-\lambda}=\frac{\lambda^{k}e^{-\lambda}}{k!}$$

となります。もともと二項分布なので、$\lambda=np$で、pを一定としたときには、λはpなり事象をn回繰り返した時の期待値です。２項分布なのだから、分散は

$$V(p,n)=np(1-p)$$

となりますが、pが小さい時には、$(1-p)$は近似的に１ですから、$V(p,n)$は近似的に

$$V(p,n)=np$$

となります。つまり、pが小さい時には、

$$V(p,n)\cong E(p,n)=np$$

となります。例えば、$p=0.01$の時、100 回に１回起こる現象が何回起こるかの確率のグラフ作るときに、一つひとつ$\binom{n}{k}p^{k}(1-p)^{n-k}$を計算するなんて、ゾッとします。ポアソン分布はそういう場合に使います。どのくらいpが小さければ、ポアソン分布と考えるのかは、時と場合によるでしょう。

あまり頻繁には起きないような現象で、予測値λがあれば、ポアソン分布を仮定して、分散（標準偏差）を確率分布の分散をλと推測することが出来ます（標準偏差=$\sqrt{\lambda}$）。この値はいくつかの観測値から得られています（たとえば１年間の観測の結果得られた１月間あるいは１日にその現象が起こる期待値）。それについては平均値だから、中心極限定理で正規分布します。だから、期待値λがわかっているときに、観測値dが得られれば、正規分布を前提に、どのくらいの頻度で得られる値なのかを、確率的に検討することが出来ます。これを Z 検定と言います（比較する片方は確率的に変動しない検定）。よく使われる t 検定は、正規分布する２つの値の差の検定で、この差は正規分布ではなくて t 分布します。

この場合、Z 値は次の式で表せます。

$$Z=\frac{d-\lambda}{\sqrt{\lambda}}$$

(この式は、ネットや教科書ごとに、既述の仕方が違うので注意してください。ここで書いたのは、dとλとの単位が等しい場合（例えば、一月間の頻度とか、同じ面積当たりに表れる頻度とか、統一されている場合）の式です。この値は、標準偏差で標準化されているから、両側検定で、この値が、1.95 以上になる確率は 0.05、片側検定で、1.64 以上になる確率は 0.05 です。というのが、ポアソン検定です。

念のために注意しておきます。途中の数学的証明はかなり緩くやっています。特に、２項分布の期待値と分散のところで、同じ値になるものをｎ回繰り返すのだから、当然そうだろうと書いたのは、それで良いのかというぐらいユルユルです。数学の試験でこの回答をすると、多分、間違いとはされないでしょうが、かなり減点されると思います。そもそも、何故それでよいのかという説明がありません。なぜそれで良いのかというと、期待値も、分散も線形性があるからです。線形性というのは、

$$E(X_1 + X_2 + \cdots X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

$$E(aX) = aE(X)$$

となっていることで、多次元であっても原点を通る直線だけが持つ性質なのです。だから、原点を通る直線だということを明らかにしなければならないのです。しかし、私たちは、数学の問題を解いているわけではないから、感覚的にわかりやすい方が良いでしょう。ポアソン分布で期待値が分散に等しいという話にもつながりやすいでしょう。だから、イメージが鮮明なユルユルの説明にしたのです。証明するのは面倒だけれども、一応、確かにそうなっていることを確認しておいた方が、印象に残るかもしれません。サイコロの例を使って、期待値と分散を計算したのが、下記の表です。

	確率						
１が出る回数	分子	分母	確率	回数X確率	偏差	偏差２乗	偏差２乗X確率
0	3125	7776	0.401878	0	-0.83333	0.694444	0.279081647
1	3125	7776	0.401878	0.401878	0.166667	0.027778	0.011163266
2	1250	7776	0.160751	0.321502	1.166667	1.361111	0.218800011
3	250	7776	0.03215	0.096451	2.166667	4.694444	0.150927355
4	25	7776	0.003215	0.01286	3.166667	10.02778	0.032239512
5	1	7776	0.000129	0.000643	4.166667	17.36111	0.002232653
合計	7776		1	0.833333			0.694444444

表３．サイコロを５回投げて表が出る回数鵜の期待値と分散の計算

表６の黄色い背景の数字が期待値、青い背景の数字が分散で、確かに

$$n \times p = 5 \times \frac{1}{6} = 0.833333, \qquad n \times p \times (1-p) = 5 \times \frac{1}{6} \times \frac{5}{6} = 0.69444444$$

になっています。

以上で、ロジスティック回帰とポアソン分布の解説を終わります。本論に戻ります。

このグループは、まず、ロジスティック回帰分析を行って、図 10 の結果を得ました。

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.2833329  0.9176017  -6.848 7.51e-12 ***
sleep        0.1164714  0.0725026   1.606    0.108
year         0.1036162  0.1561697   0.663    0.507
p_time       0.5237232  0.1308386   4.003 6.26e-05 ***
a_temp      -0.0001267  0.0152121  -0.008    0.993
rain         0.0010032  0.0097903   0.102    0.918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

図 10. ロジスティック回帰分析の結果

この結果を見て、p-time の練習時間の係数は有意で、他には有意の係数となる説明変数はないと結論しています。これは、$pr(>|z|)$のところで、星印がついているのが、切片と、p-time 以外になかったからだと思います。多分、このシステムでは、$pr(>|z|) < 0.001$で***、$pr(>|z|) < 0.01$で**、$pr(>|z|) < 0.05$で、$pr(>|z|) < 0.1$で·が付くのだと思います。有意差蟻のマークはつかなかったのですが、sleep の$pr(>|z|)$は 0.108 です。近似的には 0.1です。多分、帰無仮説は、傾き 0 だと思うので、この推定値の範囲に、傾き 0 が含まれる確率、１０回の内ほぼ 1 回に過ぎないと言っているのです。ここで分析しているのは、ケガする頻度という、極めて確率の低い現象です。また、ケガをするということは、選手にとってもチームにとっても大変なことです。ですから、ケガする頻度を少しでも下げたいと考えるのが自然でしょう。そうだとすると、ケガにかかわりそうな要素は積極的の取り上げた方が良いでしょう。そこで、8 時間以上寝た日のデータをから、ケガの頻度を割り出して、この頻度から、期待値を計算し、時間以上寝なかった日のけがの頻度の間で、ポアソン分析を行って、差があるかないかを検討して、睡眠時間の影響があるかないかを、別途明らかにすべきでしょう。おそらく、このレポートのそれとおなじようなことをしたのだとおもいます。その結果、確かに、8 時間寝た日と寝なかった日では有意な差があったようです。そうであれば、ロジスティック回帰分析でにもどって、sleep を有意とすべきです。正しい手順としては、そこで、sleep と p-time だけを使って、ロジスティック回帰をやり直すべきなのですが、手元のデータがないから、図 8 の回帰係数をそのまま使って

$$f(\text{sleep}, \text{p} - \text{time}) =$$

$$-6.2833329 + 0.1164714sleep + 0.5237232ptime$$

という式が作れますが、この式はかなり怪しい式です。練習時間が増えれば、ケガが多くなるのに対して、睡眠時間が長くなれば、ケガが少なくなるのですから、睡眠時間と練習時間の正負の符号は反対になるはずです。つまり、絶対にどこかが間違っています。データの入力範囲がまずいのかもしれませんが、ありそうなのは、sleep か ptime の入力データの正負の符号が間違っているのだと思います。元のデータを見ることが出来ないので、確かめようがありませんが、pの式にこのデータを入れてみると、sleep のデータが増加すると、pが増加してしまいます。睡眠時間が長くなるとケガの頻度が高くなっては困りますから、おそらく、sleep の係数の符号が違っているのです。表 4 のような計算を行って、

P と sleep、ptime の関係を調べて、図 11 のグラフを作りました。

表 4．練習時間ごとの睡眠時間とケガの頻度の関係（練習時間 1 時間の例）

	睡眠時間		練習時間		計算				
intercept	係数	時間	係数	時間	f	"-f	e^-f	1+e^-f	p
-6.28333	-0.11467	0	0.5237232	1	-5.7596097	5.75961	317.2245	318.2245	0.003142
-6.28333	-0.11467	1	0.5237232	1	-5.8742811	5.874281	355.7688	356.7688	0.002803
-6.28333	-0.11467	2	0.5237232	1	-5.9889525	5.988953	398.9964	399.9964	0.0025
-6.28333	-0.11467	3	0.5237232	1	-6.1036239	6.103624	447.4764	448.4764	0.00223
-6.28333	-0.11467	4	0.5237232	1	-6.2182953	6.218295	501.847	502.847	0.001989
-6.28333	-0.11467	5	0.5237232	1	-6.3329667	6.332967	562.8238	563.8238	0.001774
-6.28333	-0.11467	6	0.5237232	1	-6.4476381	6.447638	631.2097	632.2097	0.001582
-6.28333	-0.11467	7	0.5237232	1	-6.5623095	6.56231	707.9047	708.9047	0.001411
-6.28333	-0.11467	8	0.5237232	1	-6.6769809	6.676981	793.9186	794.9186	0.001258
-6.28333	-0.11467	9	0.5237232	1	-6.7916523	6.791652	890.3835	891.3835	0.001122
-6.28333	-0.11467	10	0.5237232	1	-6.9063237	6.906324	998.5694	999.5694	0.001
-6.28333	-0.11467	11	0.5237232	1	-7.0209951	7.020995	1119.9	1120.9	0.000892
-6.28333	-0.11467	12	0.5237232	1	-7.1356665	7.135667	1255.974	1256.974	0.000796
-6.28333	-0.11467	13	0.5237232	1	-7.2503379	7.250338	1408.581	1409.581	0.000709
-6.28333	-0.11467	14	0.5237232	1	-7.3650093	7.365009	1579.73	1580.73	0.000633

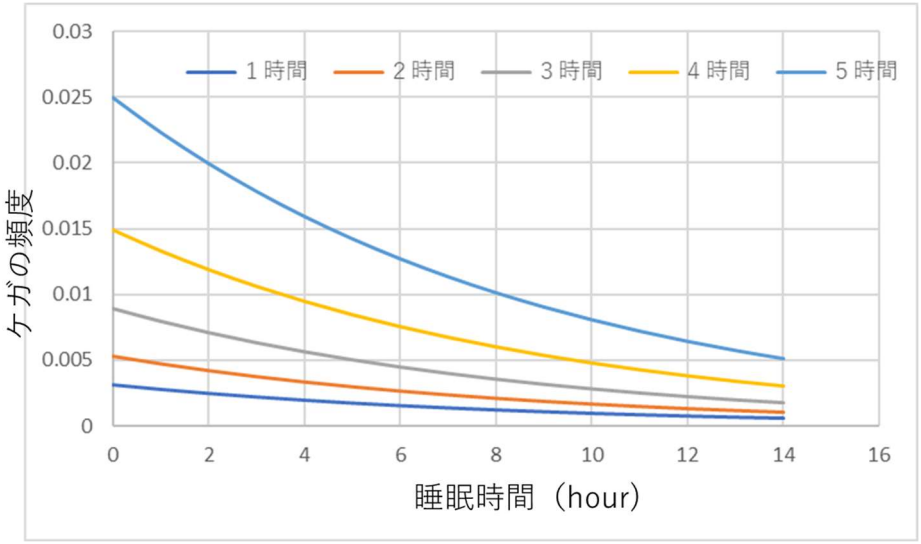

図 11. 睡眠時間・練習時間とケガの頻度の関係

図の１１で、８時間睡眠した場合、４時間練習したときの、けがの頻度は 0.006 です。この時の、fの値は、-5.1059113 でした。睡眠時間が 4 時間のであった場合、どのくらいのれんしゅうじかんにすれば、同じぐらいの悔過の頻度で収まるかを考えます。

Sleep を 4 時間にして、fが-5.1059113 になればよいのだから、

$-5.1059113 = -6.2833329 - 0.1146714 \times 4 + 0.5237232\text{ptime}$を解いて、

$$\text{ptime} = 3.123992$$

が得られる。すなわち、練習時間を 3 時間程度に減らせば、8 時間寝て 4 時間練習した時

と同程度の飢餓の頻度になる。この結果を受けて、睡眠時間が４時間程度になった場合には、練習時間を短くして、３時間程度にすれば、ケガの頻度を上げなくて済むので、練習時間を睡眠時間が８時間より短くなった場合には、１時間について 15 分ずつ、練習時間を短くすることを提案しても良いのだが、運動部の場合、受け入れてもらえないような気がします。そうであれば、ケガが多いのは実戦形式の練習だろうから、睡眠時間の短かったものから順に実戦形式の練習に入り、早めに実践形式の練習を終えて、後は筋トレとか、ストレッチ、イメージトレーニングのようなものに充てて、実戦形式の連数時間を減らせばよいだろうと提案してはどうでしょうか。その場合の減らす目安は、睡眠時間１時間当たり、１５分ということになります。

私ならば、以上の様にレポートをまとめます。元コンサルタント会社の社員としては、単なる分析で終わりたくないのです。分析の結果が、どうすれば良いかという具体的な提案につながらなければコンサルタントとしては意味がありません。このグループのテーマは、具体的な提案につながる可能性がある面白いテーマだったのですが、もったいないことをしました。

サジェスチョンとして言えることは、統計解析の結果（この場合は回帰式）が、自分たちが観察した結果をどのように説明しているのか、説明できているのか、自分たちのデータに当てはめて、きちんと確認しなければいけないということです。確認していれば、係数の正負の符号が誤りであることに気が付いたでしょう。また、ロジスティック回帰の結果を、コンピュータの画面の写真で説明してしまい。意味を考えようとしなかったことも、議論が発展しなかった原因です。有意差があることを示す星印が何を表しているか考えれば、ｐ＝０．１０８という危険率が、かならずしも意味がないと結論すべきものではないと思いついたでしょう。分析方法を考えたという点では、よく頑張っていると思います。統計解析の経験に乏しかったということに同情すべきかもしれません。（統計分析ソフトが出した結論を機械的にそのまま受け入れてはいけません。自分たちのデータに当てはめてみて、何を意味しているのか考えるのが分析です。