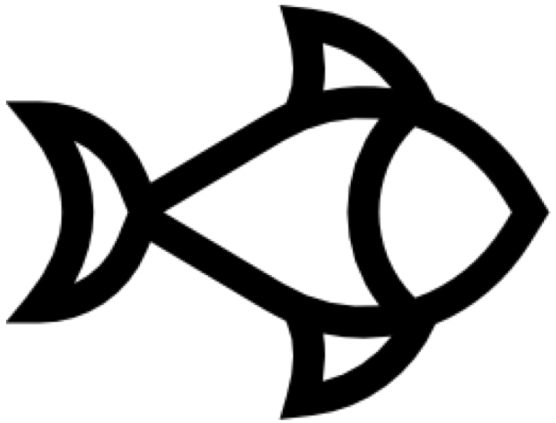
Poisson Regression



Poisson – counts (integer)

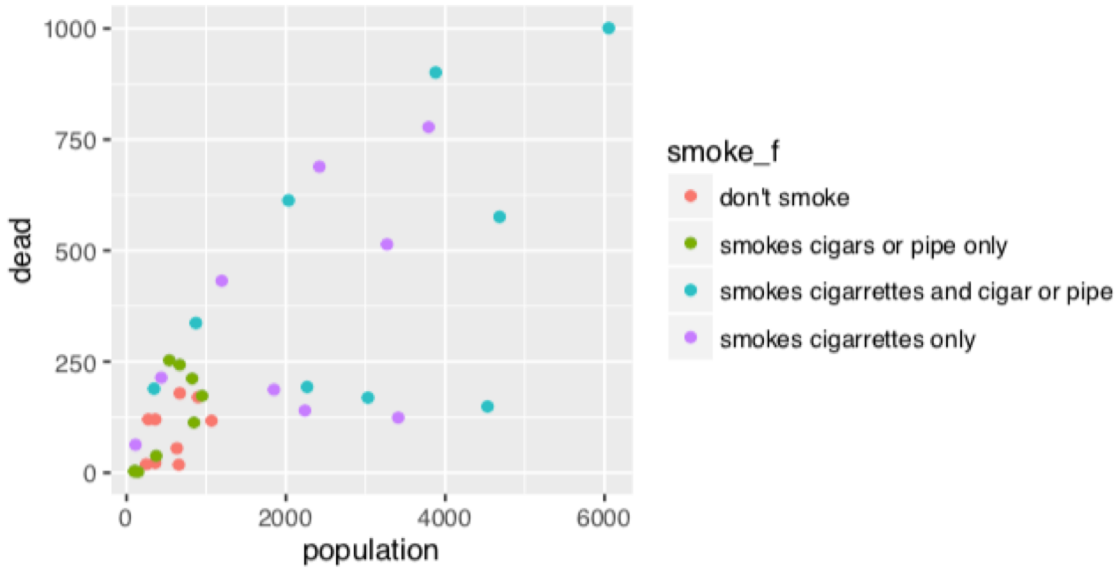
- number of traffic accidents,
- I number of warranty claims received each week, or
- I number of reads mapped to some genes.

Offset

- Population
- Time
- Time at Risk

Log link: $\log\left(\frac{\mu}{u}\right) = \boldsymbol{X\beta} \rightarrow \frac{\mu}{u} = exp(\boldsymbol{X\beta}) \rightarrow \mu = u\ exp(\boldsymbol{X\beta})$

identity link: $\left(\frac{\mu}{u}\right) = \boldsymbol{X\beta} \rightarrow \mu = u\boldsymbol{X\beta}$



Survival from lung cancer in males who were diagnosed in Connecticut during 1973 (Holford 1980)

- Total 539 Subjects
- 3 Levels of histology
- 3 stages of disease
- Follow up every 2 month

	time	1_1	1_2	1_3	2_1	2_2	2_3	3_1	3_2	3_3
1:	1	157	134	212	77	71	130	21	22	101
2:	2	139	110	136	68	63	72	17	18	63
3:	3	126	96	90	63	58	42	14	14	43
4:	4	102	86	64	55	42	21	12	10	32
5:	5	88	66	47	50	35	14	10	8	21
6:	6	82	59	39	45	32	13	8	8	14
7:	7	76	51	29	42	28	7	6	6	10

Total time at risk in month

	H_1			H_2			H_3		
time	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
0- 2:	9	12	42	5	4	28	1	1	19
2- 4:	2	7	26	2	3	19	1	1	11
4- 6:	9	5	12	3	5	10	1	3	7
6- 8:	10	10	10	2	4	5	1	1	6
8-10:	1	4	5	2	2	0	0	0	3
10-12:	3	3	4	2	1	3	1	0	3
12+:	1	4	1	2	4	2	0	2	3

Number of death from lung cancer

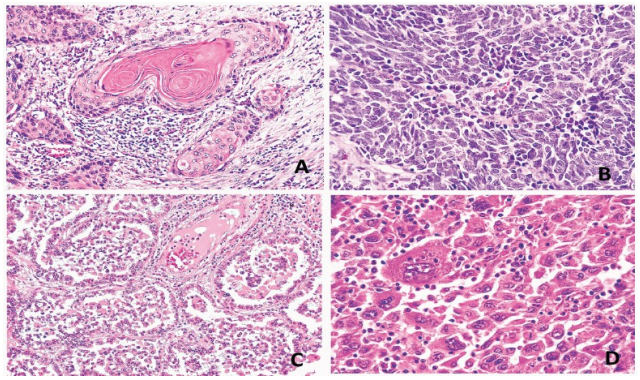


Table 4.1: Number of Children Ever Born to Women of Indian Race By Marital Duration, Type of Place of Residence and Educational Level

- mean,
- variance and
- sample size

Location

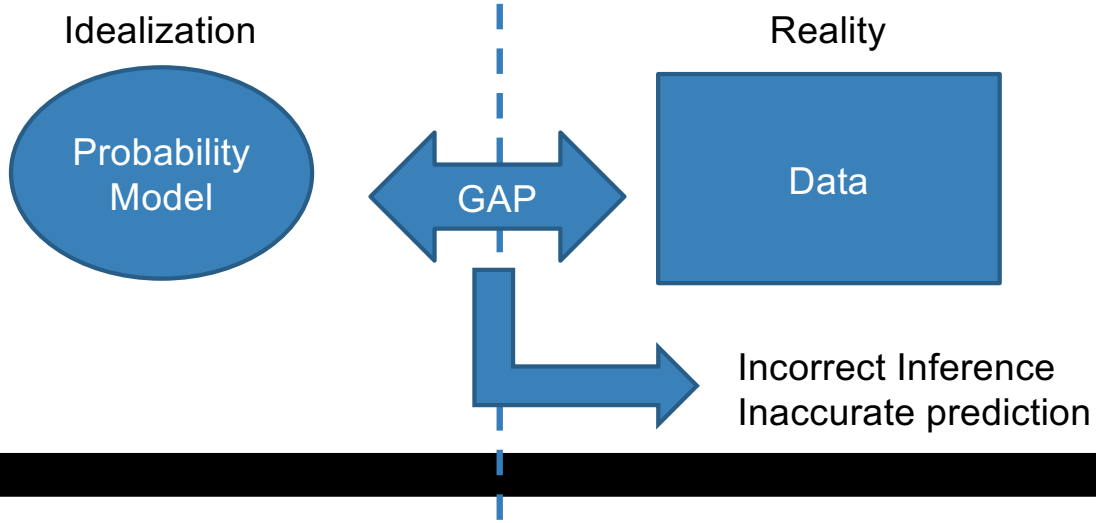
education

Marr. Dur.	Suva N	Suva LP	Suva UP	Suva S+	Urban N	Urban LP	Urban UP	Urban S+	Rural N	Rural LP	Rural UP	Rural S+
0–4	0.50	1.14	0.90	0.73	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	1.14	0.73	0.67	0.48	1.06	1.59	0.73	0.54	0.88	0.81	0.80	0.59
	8	21	42	51	12	27	39	51	62	102	107	47
5–9	3.10	2.67	2.04	1.73	4.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	1.66	0.99	1.87	0.68	3.44	1.51	0.97	0.81	1.93	1.36	1.30	1.19
	10	30	24	22	13	37	44	21	70	117	81	21
10–14	4.08	3.67	2.90	2.00	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	1.72	2.31	1.57	1.82	2.97	2.99	1.96	1.52	3.52	3.31	3.28	2.50
	12	27	20	12	18	43	29	15	88	132	50	9
15–19	4.21	4.94	3.15	2.75	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	2.03	1.46	0.81	0.92	7.40	2.97	3.83	0.70	4.91	3.23	3.29	–
	14	31	13	4	23	42	20	5	114	86	30	1
20–24	5.62	5.06	3.92	2.60	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	4.15	4.64	4.08	4.30	7.19	4.44	4.33	0.33	8.20	5.72	5.20	0.50
	21	18	12	5	22	25	13	3	117	68	23	2
25–29	6.60	6.74	5.38	2.00	6.52	7.51	7.54	–	7.48	7.81	5.80	–
	12.40	11.66	4.27	–	11.45	10.53	12.60	–	11.34	7.57	7.07	–
	47	27	8	1	46	45	13	–	195	59	10	–

http://data.princeton.edu/wws509/notes/c4.pdf

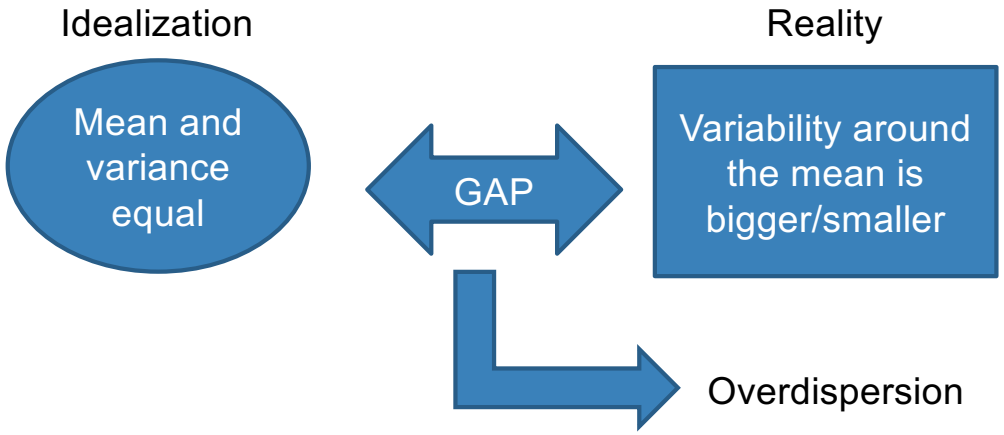
Mind the Gap

- Probability models are idealization and not necessarily a representation of reality.
- You can fit the model to the data but all models have their limits.
  - Gaussian: Unimodal, symmetric, light tail
  - Binomial: Expected proportion and variance have relationship (np(1-p))
  - Poisson: Unimodal and almost symmetric for large rates. Mean and variance are equal



Over/Under dispersion

- Over/under dispersion almost always happen.
- Diagnostics: Residual plot
- Remedy:
  - Better estimation of the standard error
  - More flexible model



Better Estimation of the standard error

- Bootstrap
- Quasi-Poisson Regression
- Sandwich estimator