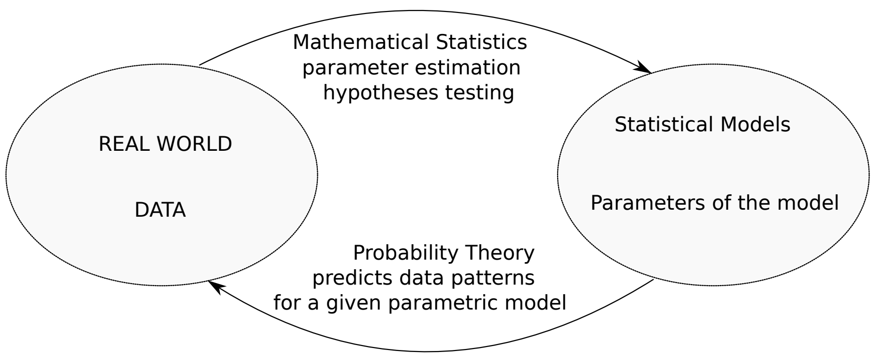
Slides 3: Parametric models

- Normal distribution model
- Binomial and Hypergeometric distributions
- Gamma distribution model
- Method of moments
- Geometric distribution

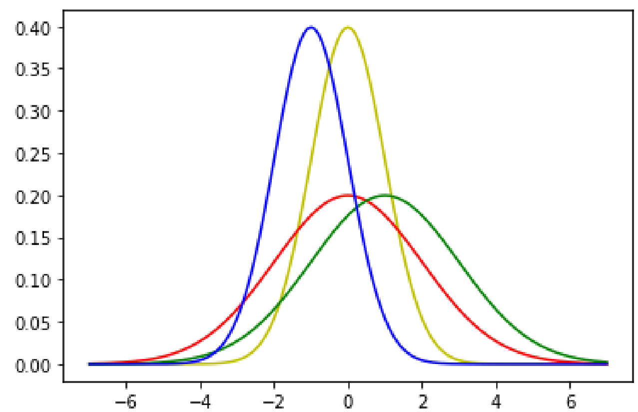


$\boxed{\text{Essentially, all models are wrong ... but some are useful.}}$

A key parametric statistical model is the normal distribution $\mathrm{N}(\mu, \sigma)$ described by the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Parameters μ and σ are the mean value and standard deviation of the probability distribution $\mathrm{N}(\mu, \sigma)$. All normal curves have the same shape.



Question. Additive noise model = signal + noise. Why is $\mathrm{N}(0, \sigma)$ a relevant distribution for the noise part? Explain by referring to the CLT.

Test question

Using the table for $P(Z \leq z)$ and $Z \sim N(0,1)$, check that $\frac{1}{\sqrt{2\pi e}} = 0.242$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

An urn contains N balls labelled as 0 or 1. Let p be the population proportion of 1's. In this case the population distribution is Bernoulli distribution with parameter p denoted by $\mathrm{Bin}(1,p)$.

Draw n balls from the urn, and call the sample mean a sample proportion

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \hat{p}$$

For sampling with replacement, the number of 1's in the sample

$$T = X_1 + \ldots + X_n \sim \ \mathrm{Bin}(n,p)$$

has a binomial distribution.

For sampling without replacement, we get a hypergeometric distribution

$$T \sim \ \mathrm{Hyp}(N,n,p)$$

Here all $X_i \sim \mathrm{Bin}(1,p)$ but they are dependent random variables.

Question. Why $\hat{p}$ is an unbiased and consistent estimate of p in both cases?

The binomial distribution $\mathrm{Bin}(n, p)$ has mean and variance

$$\mu = np, \quad \sigma^2 = np(1-p).$$

The hypergeometric distribution $\mathrm{Hyp}(N, n, p)$ has mean and variance

$$\mu = np, \quad \sigma^2 = np(1-p)(1 - \tfrac{n-1}{N-1}).$$

Standard error of $\hat{p}$ for sampling with replacement

$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}}$$

and for sampling without replacement

$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1 - \tfrac{n}{N}}.$$

Question. Course data: proportion of females was $\hat{p} = \frac{27}{94} = 0.29$ yielding

$$I_p \approx \hat{p} \pm 1.96 \cdot \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}} = 0.29 \pm 0.09 = (0.20, 38).$$

What does interval $(0.20, 38)$ say about the proportion p?

Gamma distribution $\mathrm{Gam}(\alpha, \lambda)$

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad x > 0,$$

is described by the shape parameter $\alpha > 0$ and the rate parameter $\lambda > 0$. The gamma density function involves the gamma function

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx,$$

which is an extension of the factorial to non-integer numbers, in that

$$\Gamma(k) = (k-1)! \quad \text{for } k = 1, 2, \ldots$$

Different values of α bring different shapes for the gamma curve. If $\alpha = 1$, then we get an exponential distribution $\mathrm{Gam}(1, \lambda) = \mathrm{Exp}(\lambda)$.

If $\alpha = k$ is integer, and $X_i \sim \mathrm{Exp}(\lambda)$ are independent, then

$$X_1 + \ldots + X_k \sim \mathrm{Gam}(k, \lambda).$$

The mean and variance values of the gamma distribution

$$\mu = \tfrac{\alpha}{\lambda}, \quad \sigma^2 = \tfrac{\alpha}{\lambda^2}.$$

Question: how to estimate unknown population parameters (α, λ) given a random sample $(x_1, \ldots, x_n)$ generated by $\mathrm{Gam}(\alpha, \lambda)$?

Method of moments is build around equations for the first two population moments

$$\mathrm{E}(X) = \tfrac{\alpha}{\lambda}, \quad \mathrm{E}(X^2) = \tfrac{\alpha(1+\alpha)}{\lambda^2}.$$

Replacing the unknown $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$ with unbiased and consistent estimates $\bar{x}$ and $\overline{x^2}$ we get two equations with two unknowns:

$$\tfrac{\alpha}{\lambda} = \bar{x}, \quad \tfrac{\alpha(1+\alpha)}{\lambda^2} = \overline{x^2}$$

or equivalently,

$$\tfrac{\alpha}{\lambda} = \bar{x}, \quad \tfrac{1+\alpha}{\lambda} = \tfrac{\overline{x^2}}{\bar{x}}$$

Solving these we obtain the method of moments estimates $\tilde{\alpha}$ and $\tilde{\lambda}$.

Example: 24 heights

Since $\bar{x} = 181.46$, $\overline{x^2} = 32964.2$, we get a pair of equations

$$\frac{\alpha}{\lambda} = 181.46, \quad \frac{1+\alpha}{\lambda} = \frac{32964.2}{181.46} = 181.66$$

which give the method of moments estimates

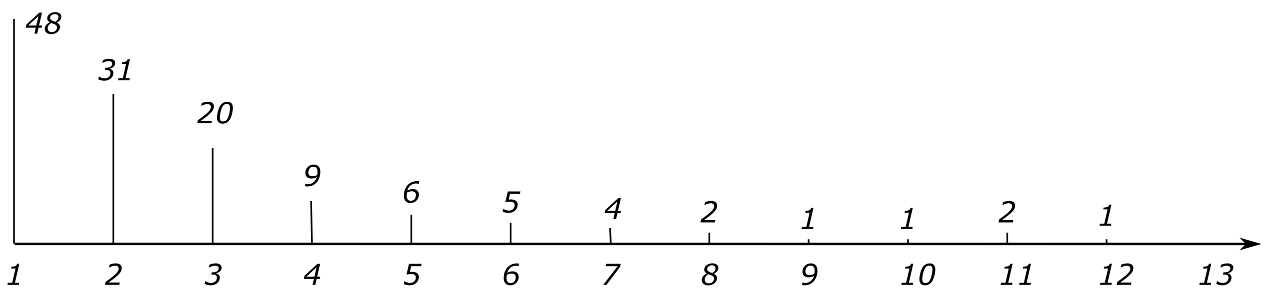
$$\tilde{\lambda} = 5.00, \quad \tilde{\alpha} = 907.3.$$

Example: 130 hopping birds

Numbers of hops between flights for $n = 130$ birds

Number of hops	1	2	3	4	5	6	7	8	9	10	11	12	Tot
Frequency	48	31	20	9	6	5	4	2	1	1	2	1	130

The histogram reminds of a geometric distribution Geom(p)



Geometric mean,variance, and the second moment are

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}, \quad \mathrm{E}(X^2) = \frac{1-p}{p^2} + \frac{1}{p^2} = \frac{2-p}{p^2}.$$

Using the sample moments

$$\bar{x} = \frac{\text{total number of hops}}{\text{number of birds}} = \frac{363}{130} = 2.79,$$

$$\overline{x^2} = \frac{1^2 \cdot 48 + 2^2 \cdot 31 + \ldots + 11^2 \cdot 2 + 12^2 \cdot 1}{130} = 13.20,$$

we can find two MM-estimates from the equations

$$\bar{x} = \frac{1}{p}, \quad \overline{x^2} = \frac{2-p}{p^2}.$$

The first equation gives $\tilde{p}_1 = \frac{1}{2.79} = 0.36$, while the second can be written as

$$(13.2) \cdot p^2 + p - 2 = 0$$

giving a similar answer $\tilde{p}_2 = 0.35$.

Question. Are $\tilde{p}_1$ and $\tilde{p}_2$ unbiased estimates of p? Are they consistent estimates?