Introduction to probability, statistics and data handling

Tomasz Szumlak, Agnieszka Obłąkowska-Mucha
Faculty of Physics and Applied Computer Science
AGH UST Krakow

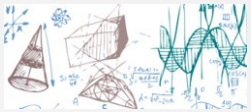
2021

A big question

- So, we collected the data – we are going to be interested in a procedure, which basing on the observed variation gives the best value (we could also ask about the range of values) for the corresponding underlaying model parameter(s)
- Again, using the stat lingo we want to get **the best possible estimate of the value of the parameter**(s)
- That is what the point estimation is all about
- BTW, it may also be useful to estimate the range of „good" parameter values – that is yet another story called estimation with confidence – we are going to look at this next time!

Not quite like that…

Point Estimators-remainder



- Consider the following: to check the water for contamination by a micro-organism a number of samples were taken, the results are summarised as follow

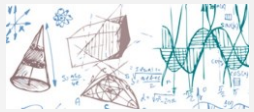
Counts	0	1	2	3	4	5	6	7	8	>9
Frequency	53	25	13	2	2	1	1	0	1	0

- One can assume that the data follow the Poisson distribution with an unknown parameter μ (each water sample is an independent observation on the same random variable!)
- For these particular data, we can estimate the μ as:

$$\bar{x} = \frac{0 \cdot 53 + 1 \cdot 25 + \cdots + 8 \cdot 1}{58 + 25 + \cdots + 1} = \frac{84}{103} = 0.816$$

$$\{X_1, X_2, \ldots, X_{103}\} \rightarrow X \equiv Poisson(\mu)$$

$$\bar{X}_{(1)} = \frac{X_1 + X_2 + \cdots + X_{103}}{103} \rightarrow \quad \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$



Point Estimators-remainder

- Let's set a **generic procedure** using this simple example
- First, we **pick the parameter** to be estimated
- Next, we need to **collect data** and **compute a sampling statistics** using a formula corresponding to the parameter we are interested in
- In our example that is a **sample mean**

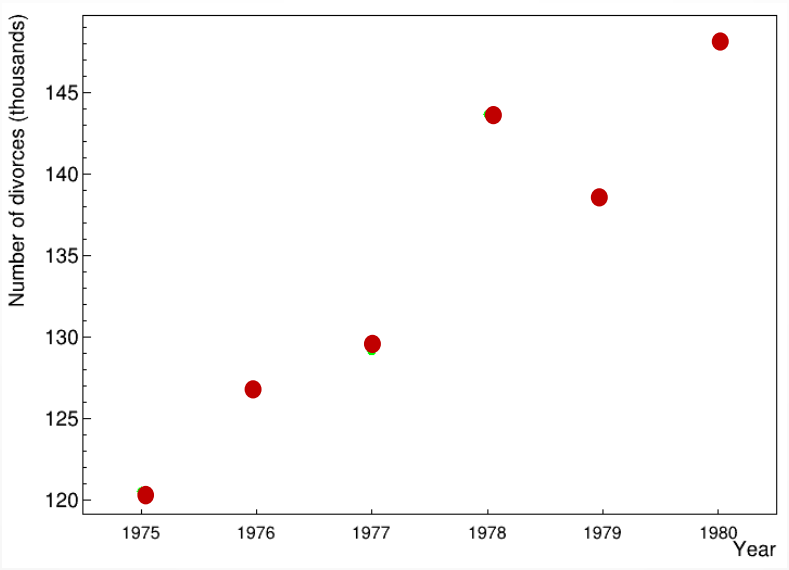
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- This, in turn, we call an **estimator** of true parameter, in our case this would be: $\mu \rightarrow \bar{X} = \hat{\mu}$ (we use the caret symbor "^")
- Remember – the estimator is a random variable, for different sample we are going to get different value
- The estimator will follow its own distribution – **sampling distribution of the estimator**

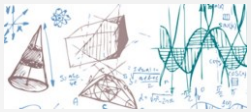
More than one way…

- Lets inspect the following data regarding the number of divorces in different years in some country in Europe

Year	1975	1976	1977	1978	1979	1980
# divorces (10^3)	120.5	126.7	129.1	143.7	138.7	148.3

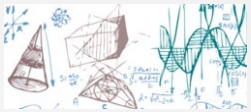


- Interesting…, very tempting to fit a model right away.



More than one way…

- From the plot we could conclude, that the **true underlaying distribution** describing the data can be represented by a **linear model**
- From the data we also conclude that **the slope** of the line is positive – ok, the task is then to **estimate this slope**, α, and then we could predict the annual rate of increase of divorces
- But how do we do that? It is not so obvious like the water example(??)
- Consider this:
  - $\hat{\alpha}_1$- join the first and the last point
  - $\hat{\alpha}_2$ - join the mid-points P_1P_2 and P_5P_6
  - $\hat{\alpha}_3$ - join the centroid of the first triplet and the second one
- Mind you, these are all sensible options!



More than one way…

- Let's do the calculations explicitly

$$\hat{\alpha}_1 = \frac{148.3 - 120.5}{80 - 75} = \frac{27.8}{5} = 5.6$$

$$\hat{\alpha}_2 = \frac{(138.7 + 148.3)/2 - (120.5 + 126.7)/2}{79.5 - 75.5} = \frac{19.9}{4} = 5.0$$

$$\hat{\alpha}_3 = \frac{(143.7 + 138.7 + 148.3)/3 - \cdots}{79 - 76} = \frac{18.13}{3} = 6.0$$

- So, is there a way to make a judgement on **which one of these estimates is the „best"**? **And what exactly the best means**?
- The second question actually pertains to **the estimator properties** and not the estimate (a number we obtained)
- So, we need to look at the properties of the **sampling distribution** of respective estimators!

Analysis

- Mind one thing. This example is not quite what we could call an experiment – we cannot repeat year 1976 and check the number of divorces again…
- However, we can still evaluate the deviations of data from the predicted model and treat them as random variable
- Say, the difference (**residual**) r is defined as follow:

$$r_i = y_i - (\beta + \alpha x_i)$$

- Next, we assume that the residual (random variable) will have a mean value and variance: μ_r and σ_r^2
- Further, we assume that the residual should have mean value equal to 0 (think about that!), so finally our model for the given data set is:

$$Y_i = \beta + \alpha x_i + r_i$$

- We have therefore three parameters $(\alpha, \beta, \sigma_r^2)$

Analysis

- Having formulated the model we can now start discussing the properties of the sampling distributions of our estimators
- So, we are going to treat **the estimate** (a number evaluated using the data):

$$\hat{\alpha}_1 = \frac{y_6 - y_1}{x_6 - x_1}$$

- … as a single measurement (observation) of the random variable (the estimator)

$$\hat{\alpha}_1 = \frac{Y_6 - Y_1}{x_6 - x_1}$$

- To come up with the answer regarding how good is such estimator we start from working out its mean and variance (we use the knowledge of these function of R.V. remembering that α, β, x_i are just constant numbers)

$$E[Y_i] = \beta + \alpha x_i \quad V[Y_i] = \sigma_r^2 \quad E[r_i] = 0, V[r_i] = \sigma_r^2$$

Analysis

- Now, we can calculate the expected value of $\hat{\alpha}_1$:

$$E[\hat{\alpha}_1] = E\left[\frac{Y_6 - Y_1}{x_6 - x_1}\right] = \frac{1}{x_6 - x_1} E[Y_6 - Y_1] =$$

$$= \frac{1}{x_6 - x_1}(E[Y_6] - E[Y_1]) = \frac{1}{x_6 - x_1}\big((\beta + \alpha x_6) - (\beta + \alpha x_1)\big)$$

$$E[\hat{\alpha}_1] = \frac{1}{x_6 - x_1}(\alpha x_6 - \alpha x_1) = \alpha$$

- Neat! The expected value of the estimator is exactly equal to the unknown parameter. Good job
- What about the other estimators?

$$\hat{\alpha}_2 = \frac{\frac{1}{2}(Y_5 + Y_6) - \frac{1}{2}(Y_1 + Y_2)}{\frac{1}{2}(x_5 + x_6) - \frac{1}{2}(x_1 + x_2)} \qquad \hat{\alpha}_3 = ?$$

Analysis

- Repeating the same calculations for remaining two estimators we conclude that their expected values are always the same and equal exactly the unknown parameter we estimating
- In general we say that an estimator $\hat{\theta}$, which we use to estimate an unknown parameter of a model, is **unbiased for parameter θ** if the following is true:

$$E[\hat{\theta}] = \theta$$

- So, it seems that all of them doing just fine. What next we can check…? The variance!
- In this case the best option would be to choose the one that features the least variability about its mean value, so:

$$V[\hat{\alpha}_1] = V\left[\frac{Y_6 - Y_1}{x_6 - x_1}\right] = \frac{1}{(x_6 - x_1)^2}(V[Y_6] + V[-Y_1]) =$$

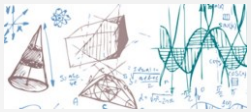
$$= \frac{1}{(x_6 - x_1)^2}(\sigma_r^2 + \sigma_r^2) = \frac{2\sigma_r^2}{(x_6 - x_1)^2} = \frac{2\sigma_r^2}{25}$$

Analysis

- Again, we can repeat the calculations for the remaining two estimators (you are encouraged to do so!)
- We get the following:

$$V[\hat{\alpha}_2] = \frac{4\sigma_r^2}{64} \qquad V[\hat{\alpha}_3] = \frac{6\sigma_r^2}{81}$$

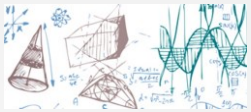
- So: $V[\hat{\alpha}_2] < V[\hat{\alpha}_3] < V[\hat{\alpha}_1]$
- Using the variance we say that the best (most efficient) estimator is the $\hat{\alpha}_2$ - thus we have the winner!



Summary so far

- A generic „algorithm" for point estimation task would be:

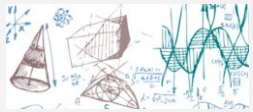
- **Collect the data** and understand it
- Come up with a **model**, this will specify a **parameter** or many parameters that we need to make an estimate
- For a given parameter(s) we need an **estimator(s)** (typically we will concentrate on the mean value or variance, however we also can tackle more ambitious cases – e.g., divorces)
- Work out the **estimate of the parameter** – this is a random variable and will be different for different data sets
- Finally, analyse the **sampling distribution of the estimator** to make a judgement of its usefulness
- We are looking for **unbiased** (expectation value) and **efficient** estimators (variance)



Can we do better?

- That was fun! And we learned a lot, however following such generic path each time we need to make an estimate **seems too much**
- We need a technique(s) that allows us **to define sensible estimators** (again, we could spend a lifetime on deriving estimators that are reasonable)
- So, such a technique would „automatically" **come up with a formula for best estimators**

- One thing to remember – there is no universal method to achieve the above task, in time a number of approaches have been proposed. There is no „best" one
- We concentrate on three techniques: the method **of least squares**, the method of **moments** and the method of **maximum likelihood**



The method of least squares

- Say, we are interested in estimating the mean value of some distribution which is θ. We take data sample: $\{X_1, X_2, \dots, X_n\}$
- Since the mean is „a typical" value, we conclude that the respective differences $X_i - \theta$ should be „small" (simultaneously)
- Also, the sum of squares of these differences should obey the formula below:

$$S = \sum_i (X_i - \theta)^2 \to min$$

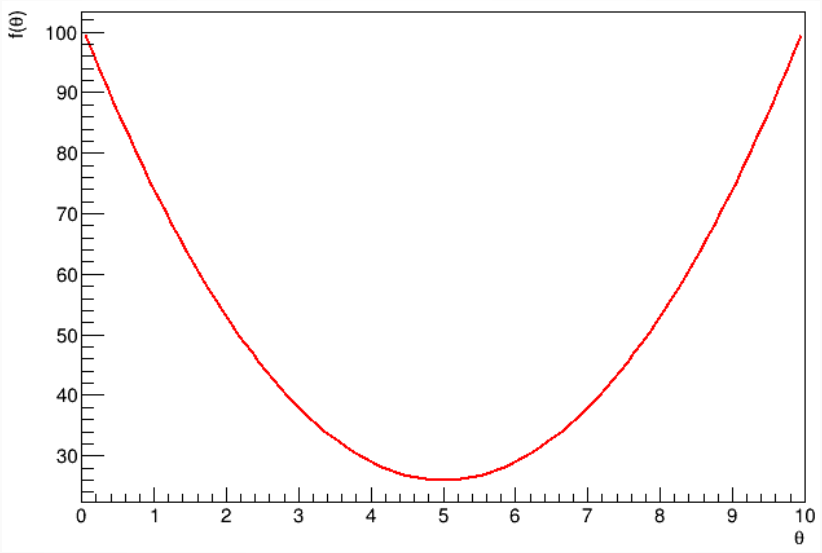
- This is the method of least squares (MLS)
- As usual, an **example** is in order! Say we collected a sample: $X = \{2, 4, 9\}$ and we want to estimate the mean value of the parent distribution these numbers came from. Applying the MLS

$$S_3 = \sum_{i/1}^{i/3} (X_i - \theta)^2 = (2-\theta)^2 + (4-\theta)^2 + (9-\theta)^2 = \cdots =$$

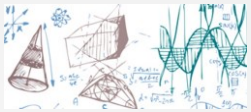
$$= 101 - 30\theta + 3\theta^2$$

The method of least squares

- This function will take different values for different θ – task is to identify the θ for which the sum is the smallest



- The minimum is attained at the point $\boldsymbol{\theta = 5}$. We can also check that the sample mean, $\bar{x} = 5$. Reasonable!
- So, the method gives results compatible with the common sense, that is encouraging!



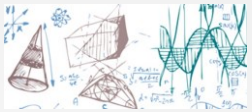
The method of least squares

- In a general case where we have a random sample of size n drawn from a population with an unknown mean value:

$$S_n = \sum_{i/1}^{i/n} (X_i - \theta)^2 = \sum_{i/1}^{i/n} \left(X_i^2 - 2\theta X_i + \theta^2\right) =$$

$$= \sum_{i/1}^{i/n} X_i^2 + 2\theta \sum_{i/1}^{i/n} X_i + n\theta^2$$

- This formula is minimised by: $\theta = \frac{1}{n}\sum_{i/1}^{i/n} X_i = \bar{X}$ - the sample mean
- So, the sample mean $\bar{x}$ of a random sample X taken from a population of an unknown mean value (we call is here in a generic way θ) is the **least squares estimator** $\hat{\theta}$ (or $\hat{\mu}$)
- The language here is important, so we are precise about what we mean



The method of moments

- Principle: We acalculate k^{th} sample moment and equate it to k^{th} population moment:

$$\frac{1}{n}\sum_i X_i^k = E[X^k]$$

- The **method of moments** (MoM) uses the **sample moments** and match them to the analogous **population moments** to obtain **estimates for the unknown parameters**. Seems simple…
- For instance we have easy example: a normal distribution $\mathcal{N}(\mu, \sigma^2)$, we use sample mean and variance:

$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = s^2$$

What about the median?

- **Not all cases are so simple**, for instance what about the Poisson distribution? Both the population mean and its variance are equal μ. Shall we use the sample mean or variance as the best estimator $\hat{\mu}$?
- Need to understand the sampling distributions to answer this!

The method of moments

- Let's write down explicitly moment estimators for a few most popular models we discussed so far

- The **Poisson**: one unknown parameter – the mean of the distribution. Matching sample and population moments gives the following estimate: $\hat{\mu} = \bar{x}$, the corresponding estimator we should use: $\hat{\mu} = \bar{X}$
- The **exponential**: the mean is $\frac{1}{\lambda}$, matching moments $\bar{X} = \frac{1}{\hat{\lambda}}$. So, we get: $\hat{\lambda} = \frac{1}{\bar{X}}$
- The binomial $\mathcal{B}(m, p)$: we have one unknown parameter p. The matching procedure gives: $\bar{X} = m\hat{p}$ (n is the sample size)

$$\hat{p} = \frac{\bar{X}}{m} = \frac{X_1 + X_2 + \cdots + X_n}{m\, n}$$

The method of moments

- In practice, we are going to observe some fluctuations, let's consider the following: we used a generator of random numbers distributed according to the Poisson model. We draw two samples:

$$X_{(1)} = \{5, 5, 2, 3, 4, 6, 4, 1\}, X_{(2)} = \{4, 2, 5, 2, 4, 1, 1, 1\}$$

$$\hat{\mu} = \bar{x}_{(1)} = \frac{30}{8} = 3.75$$

$$\hat{\mu} = \bar{x}_{(2)} = \frac{20}{8} = 2.50$$

- **These variations are „normal"**, we are going to observe them
- The point is that (SV – sample variance):

$$E[\bar{X}] = \mu, \qquad SV[\bar{X}] = \sqrt{\frac{\sigma^2}{n}}$$

Problems: nonsensical result

- Consider the following: we are testing an uniform random number generator $X \sim U(0, \theta)$, we took a sample: $X = \{3.2,\ 2.9,\ 13.1\}$

$$E[U(0,\theta)] = \frac{1}{2}\theta$$

$$\bar{x} = \frac{1}{2}\hat{\theta} \rightarrow \hat{\theta} = 2 \cdot \bar{x}, \hat{\theta} = 12.8$$

- There is something seriously wrong! How come, we get the value that is larger than 12.8? Sometimes, **the method will fail badly**
- Of course we could still preserve the validity by inventing a better estimator. In this case we can prove that the following will give more reasonable results:

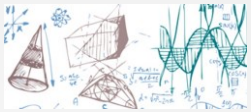
$$\hat{\theta} = \left(1 + \frac{1}{n}\right) X_{max}$$

Problems: censored data

- Sometimes we are going to face data samples that just cannot be used to calculate moments, for instance **consider the experiment**: in a large city 1469 cars were stopped and the number of occupants was counted. The results:

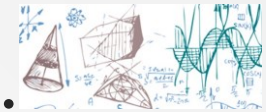
Count	1	2	3	4	5	≥6
Frequency	902	403	106	38	16	4

- This is an example of, so called, censored data – we actually do not know the precise results for the cases with 6 or more occupants
- Using moments will not, in principle, be successful (however, given the number of events, probably still quite good!)



Short recap

- We learned about parameter(s) point estimation
- We set up a general recipe and considered some more intuitive and some not so intuitive examples (slopes)
- We discussed two methods that are popular and often used when estimation theory is needed:
  - Least squares
  - Method of moments
- We also learned about bias and efficiency of the estimators, which are related not to the obtained values from the samples but rather to the formulas that represent the said estimators.

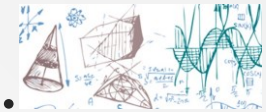


Maximum Likelihood estimation

- Let's begin our tale on the **Maximum Likelihood** technique then
- Consider a R.V. X that is described by a P.D.F. $f(x)$. The support set $supp(f) = \{x \in X, f(x) \neq 0\}$ is also the sample space. Then we take n independent observations which form a data sample $\vec{x}_{(1)} = (x_1, x_2, \dots, x_n)$. Now, we can take another sample, and another, and…
- By this „innocent" operation, we redefined the sample space to be a space of all $n - \dim$ vectors $\vec{x}_{(i)}$ and the first experiment is just **a single measurement** in that space
- Now, something very important, by our assumption of I.D.R.V. and independence we can construct a **joint P.D.F.**

$$f_{\vec{x}_{(1)}}(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

- If the sample is large this formula is seemingly complicated but remember each R.V. is identically distributed!



Maximum Likelihood estimation

- Before an experiment is performed the outcome is unknown. Probability allows us to predict ***unknown*** outcomes based on ***known*** parameters:

$$P(data/\theta)$$

for example: $P(x/n,p)=\binom{n}{x}p^x(1-p)^{n-x}$

- After the experiment is performed the outcome is ***known***. Now we talk about the ***likelihood*** that a parameter would generate the observed data:

$$P(\theta/data)$$

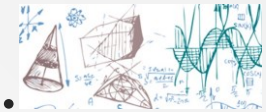
like here: $P(p/n,x)=\binom{n}{x}p^x(1-p)^{n-x}$

so:

$$P(data/\theta)=P(\theta/data)$$

- Estimation proceeds by finding the value θ of that makes the observed data most likely

Maximum Likelihood estimation



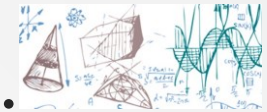
Example

Suppose we have a coin which – as we happen to know – is not a fair one. Namely – one side is likely to happen twice as frequently as the second one but ... we do not know which one.
we perform an experiment: flipping the coin 6 times and we get 4 heads (and 2 tails). We have two possibilities:

First case	Second case
(1) $\mathcal{P}(H) = 2/3; \quad \mathcal{P}(T) = 1/3$	(2) $\mathcal{P}(H) = 1/3; \quad \mathcal{P}(T) = 2/3$
$W_4^6 = \binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2$	$W_4^6 = \binom{6}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2$

$$\frac{L_1}{L_2} = \frac{\binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2}{\binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2} = \ldots = 4.$$

Obviously, the first hypothesis about the $\mathcal{P}$'s is the better one.



Maximum Likelihood estimation

- The game is still the same: using the observed x_i we want to **infer the properties of the parent P.D.F**. (its parameters). We already seen, that by constructing special functions of R.V.s we can estimate parameters
- Here, these functions (**statistics**) are called estimators (e.g. the sample mean)
- One more thing – we need to propose a hypothesis regarding the concrete form of the parent function in a form

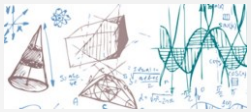
$$f = f(\vec{x}|\vec{\theta}), \vec{x} = (x_1, x_2, \dots, x_n), \vec{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$$

- Then, when taking the data sample we can use our joint P.D.F. to evaluate the probability of observing this particular sample as (let's take the discrete case for starters):

$$p(X_1 = x_1, \dots, X_n = x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \prod_i f(x_i|\theta)$$

$$\mathcal{L} = \prod_i f(x_i|\theta) \qquad \textbf{Likelihood function}$$

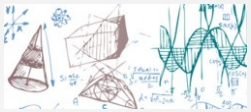
Statement of the ML



- When a sample of n independent and I.D.R.V. is collected it is possible to construct the **likelihood function** of unknown parameter θ for the random sample $\vec{x}_{(1)}$, the value $\hat{\theta}$ of the unknown parameter at which the **likelihood is maximised** is called the **maximum likelihood estimator** for that parameter
- And for the C.R.V. we get pretty much the same but now we need to formally requested that the value observed during the experiment is within an interval $[x_i, x_i + dx_i]$, and the prob.:

$$p(x_i \in [x_i, x_i + dx_i], \dots | \theta) = f(x_1|\theta)dx_1 \cdots f(x_n|\theta)dx_n$$
$$= \prod_i f(x_i|\theta)\, dx_i$$

- Since any of the dx_i do not depend on parameter, we are going to end up with the same likelihood function!
- Note, that the intuitive reasoning here is as follow: if the P.D.F. we proposed is the **correct** one we should observe a **high probability** for **observing the sample** for the right value of θ

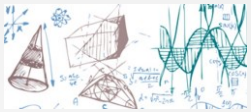


Statement of the ML

- Now, with such intuitive interpretation, the ML estimator(s) for the parameter(s) are defined as those that maximise the likelihood function (we silently assumed that $\mathcal{L}$ is differentiable):

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = 0, j = 1,2, \dots, m$$

- Ok, now we are finished! Just a short note on the **properties of the ML estimators** (MLEs)
- It can be shown that the MLE are often unbiased: $E[\hat{\theta}] = \theta$, or at least asymptotically unbiased: $E[\hat{\theta}] \to \theta, n \to \infty$
- MLEs are consistent: $V[\hat{\theta}] \to 0, n \to \infty$
- MLEs are asymptotically normally distributed (this feature now pertains to the sampling distribution of the estimators, we know that it is important – remember that when we start talking about statistical tests)



Example – geometric P.D.F.

- We are going to discuss the properties of the ML using several examples. Let's start with an easy one
- Some phenomena is known to be governed by a geometric density function: $p(x,\theta) = (1-\theta)^{x-1}\theta, x = 1,2,\ldots$ A random sample of three observations was taken: $\vec{x} = \{3,4,8\}$
- The likelihood function is then:

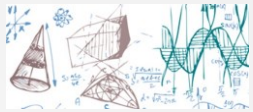
$$\boldsymbol{\mathcal{L}(\theta|\vec{x})} = p(3,\theta)\cdot p(4,\theta)\cdot p(8,\theta) =$$

$$= [(1-\theta)^{3-1}\theta]\cdot[(1-\theta)^{4-1}\theta]\cdot[(1-\theta)^{8-1}\theta] = \cdots = (1-\theta)^{12}\theta^{3}$$

- A technical note: it is much easier to handle sums than products, especially when differentiation is required:

$$\mathcal{L}(\theta|\vec{x}) = \prod_i f(x_i|\theta) \rightarrow \ln\big(\mathcal{L}(\theta|\vec{x})\big) = \sum_i f(x_i|\theta)$$

$$\ln\big(\mathcal{L}(\theta|\vec{x})\big) = 12\cdot\ln(1-\theta) + 3\cdot\ln\theta \rightarrow \frac{d(\ln\mathcal{L})}{d\theta} = 0$$



Example – geometric P.D.F.

- For the record we can write out both ways

$$\frac{d(\ln\mathcal{L})}{d\theta} = -\frac{12}{1-\theta} + \frac{3}{\theta} \rightarrow \hat{\theta} = 0.2$$

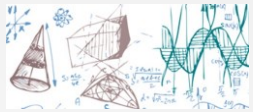
$$\frac{d(\mathcal{L})}{d\theta} = -12 \cdot (1-\theta)^{11} \cdot \theta^3 + 3 \cdot (1-\theta)^{12} \cdot \theta^2 \rightarrow \hat{\theta} = 0.2$$

- The log likelihood much more „user friendly"
- We know, that in case of the geometric distribution $f(x;p)$ its mean value can be used to estimate the p:

$$E[x] = \frac{1}{p} \rightarrow \hat{p} = \frac{1}{E[x]}, x \in N, p \equiv \theta$$

- So, using the method of moments, we get the same result:

$$E[x] = \frac{3+4+8}{3} = \frac{15}{3} = 5, \hat{p} = \frac{1}{5} = 0.2$$



Exponential distribution

- It is a very common practice to model spontaneous particle decays using exponential model, say we did an experiment and observed n decays obtaining a sample $\vec{t}_{(1)} = \{t_1, t_2, \dots, t_n\}$
- We attempt to describe the variability of measured times t_i using the following formula

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}, E[f] = \tau$$

- The likelihood function:

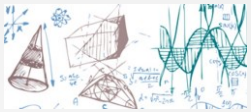
$$\mathcal{L}(\tau;t) = \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} = \frac{1}{\tau^n} e^{-\sum_i \frac{t_i}{\tau}}$$

$$\ln\mathcal{L}(\tau;t) = \ln\frac{1}{\tau^n} - \sum_i \frac{t_i}{\tau} = n \cdot \ln\frac{1}{\tau} - \sum_i \frac{t_i}{\tau} = \sum_i \left(\ln\frac{1}{\tau} - \frac{t_i}{\tau}\right)$$

$$\frac{d\left(\ln\mathcal{L}(\tau;t)\right)}{d\tau} = \sum_i \left(\frac{t_i}{\tau^2} - \frac{1}{\tau}\right) = \frac{n}{\tau}\sum_i \left(\frac{t_i}{\tau} - 1\right) = 0$$

- And finally we get: $\hat{\tau} = \frac{1}{n}\sum_t t_i$ - just the same as for the method of moments!

Exponential distribution



- It is also very instructive to check for the possible biases in our estimator – let's evaluate its expectation value:

$$E[\hat{\tau}(t_1, t_2, \dots, t_n)] = \int \cdots \int \hat{\tau}(t_1, t_2, \dots, t_n) \cdot f_{\text{joint}}(t_1, t_2, \dots, t_n; \tau) \prod_i dt_i$$

$$= \int \cdots \int \hat{\tau}(t_1, t_2, \dots, t_n) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod_i dt_i =$$

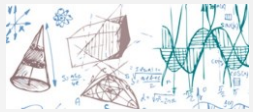
$$= \frac{1}{n} \sum_i \left(\int t_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-\frac{t_j}{\tau}} dt_j \right) = \frac{1}{n} \sum_i \tau = \tau$$

- This formula may look a bit intimidating, but it is not so bad, let's take a detailed look at the case with just two time values

$$\vec{t} = \{t_1, t_2\}, n = 2$$

$$E[\hat{\tau}(t_1, t_2)] = \int \int \left(\frac{1}{n}(t_1 + t_2) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 \right)$$

- Not so bad!



Exponential distribution

- Let's expand the sum

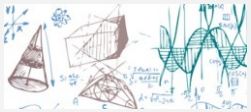
$$E[\hat{\tau}(t_1, t_2)] = \int\int \left(\frac{1}{n} t_1 \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 + \frac{1}{n} \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot t_2 \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 \right)$$

- In each component there will be exactly one element with matching exponent this is how we obtained the third line on the previous slide. Also, the respective measurements are independent and we can separate integrals now:

$$E[\hat{\tau}(t_1, t_2)] = \frac{1}{n} \sum_i \left(\int t_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-\frac{t_j}{\tau}} dt_j \right)$$

- Here, $i = 1,2$ and the product will have just one component
- Let's think about integrating this: we are going to have two cases: $\int x e^x dx$ (integration by parts) and $\int e^x dx$

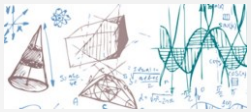
Exponential distribution



- The latter: $\int e^x dx \rightarrow \left\{x = -\frac{t_i}{\tau}, dx = -\frac{1}{\tau} dt_i, dt_i = -\tau dx\right\}$
- Integration will yield: $I = -e^{-\frac{t_i}{\tau}}|_0^\infty = -e^{-\frac{\infty}{\tau}} + e^{-\frac{0}{\tau}} = 1$
- And the former: $-\tau e^{-\frac{t_i}{\tau}}\left(\frac{t_i}{\tau} + 1\right)|_0^\infty = \tau$
- We get after plug in these results:

$$E[\hat{\tau}(t_1, t_2)] = \frac{1}{n}\sum_i (\tau \cdot 1) = \tau$$

- The point here is that the ML estimator for an exponential distribution is not biased – that is, if we are after τ.



Functions of ML estimators

- What would happen, if instead of the mean life-time we would be interested in its reciprocal (called decay constant): $\lambda = \frac{1}{\tau}$?
- In this case we can treat the λ as a function of our model parameter, if we represent such function in general as ω, the likelihood maximisation formula will be as follow:

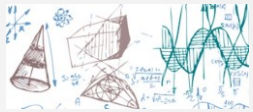
$$\frac{d\mathcal{L}(\theta|\vec{x})}{d\theta} = \frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\omega}\frac{\partial\omega}{\partial\theta} = 0$$

$$\frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\theta} = 0 \rightarrow \frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\omega} = 0, \frac{\partial\omega}{\partial\theta} \neq 0$$

- Or in words: we can evaluate the ML estimator for any function of the original estimator just by inserting the original estimator in

$$\hat{\omega} = \omega(\hat{\theta})$$

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \frac{n}{\sum_i t_i}$$



Normal distribution

- Similar analysis can be performed for the normal distribution to estimate its mean and variance (this time we are not going to do the exact calculations just cite the results, again you are encouraged to repeat them)
- The log likelihood function for $\mathcal{N}(x;\mu,\sigma^2)$:

$$\ln\mathcal{L}(\mu,\sigma^2)=\sum_i\left(\ln\frac{1}{\sqrt{2\pi}}+\frac{1}{2}\ln\frac{1}{\sigma^2}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- The minimisation will give us:

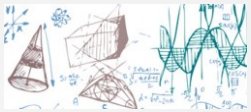
$$\hat{\mu}=\frac{1}{n}\sum_i x_i\,,\widehat{\sigma^2}=\frac{1}{n}\sum_i(x_i-\hat{\mu})^2$$

$$E[\hat{\mu}]=\mu,E\left[\widehat{\sigma^2}\right]=E\left[\frac{1}{n}\sum_i(x_i-\hat{\mu})^2\right]=E\left[x_i^2\right]-2E[x_i\hat{\mu}]+E[\hat{\mu}^2]$$

$$E\left[\widehat{\sigma^2}\right]=\frac{n-1}{n}\sigma^2$$

- ML estimator for the $\mathcal{N}(x;\mu,\sigma^2)$ variance is biased

Variance of ML estimators



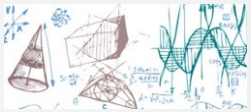
- So far our procedure of ML estimation requires to collect a data sample and make a hypothesis for the P.D.F. $f(x,\theta)$
- If we keep repeating experiments we also get different values for the estimated parameters – sampling distribution
- That would be the way to analyse the properties of this S.D.o.E. (sampling distribution of estimator) and evaluate its variance
- One approach, called analytic method, would be to just make an exact computations, for instance the exponential model would give us:

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}]^2)$$

$$= \int \cdots \int \left(\frac{1}{n}\sum\nolimits_i t_i\right)^2 \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod\nolimits_i dt_i$$

$$- \int \cdots \int \left(\frac{1}{n}\sum\nolimits_i t_i\right) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod\nolimits_i dt_i$$

$$= \frac{\tau^2}{n}$$



Variance of ML estimators

- The result above is universal in a sense, the variance of the sample mean is $1/n$ times smaller than the variance of P.D.F. used to model the variation of t R.V.
- Since our result depends on an unknown parameter (τ) we need to use our estimated value to calculate the variance:

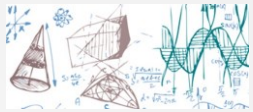
$$\sigma_{\hat{\tau}}^2 = \frac{\tau^2}{n} \rightarrow \sigma_{\hat{\tau}}^2 = \frac{\hat{\tau}^2}{n}, \sigma_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

- Formally, we used the transformation invariance property of the ML estimators
- So, how should one interpret an experimental result like this:

$$\tau = 12.38 \pm 0.72$$

- Here, the ML estimate is 12.38 and the statistical uncertainty means that if the experiment would be repeated many times, the standard deviation of its S.D.o.E. would be 0.72

Variance of ML estimators



- And what if we do not know, even in principle, the P.D.F. (say we are looking for a new phenomena)
- We could use the following reasoning, let's start with expanding the likelihood function in a Taylor series about the estimate $\hat{\theta}$:

$$\ln\mathcal{L}(\theta) = \ln\mathcal{L}(\hat{\theta}) + \left[\frac{\partial \ln\mathcal{L}}{\partial\theta}\right]_{\theta=\widehat{\theta}} (\theta - \hat{\theta}) + \left[\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}\right]_{\theta=\widehat{\theta}} (\theta - \hat{\theta})^2 + \cdots$$

- We also define:

$$\widehat{\sigma_\theta^2} = \left(\frac{-1}{\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}}\right)_{\theta=\widehat{\theta}}$$

- Since, the first derivative by definitione should be zero for the estimate and $\ln\mathcal{L}(\hat{\theta}) = \ln\mathcal{L}_{Max}$

$$\ln\mathcal{L}(\hat{\theta}) = \ln\mathcal{L}_{Max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma_{\hat{\theta}}^2}}$$

Variance of ML estimators

- What is the effect of varying this formula by one standard deviation?

$$\ln\mathcal{L}\left(\hat{\theta} \pm \widehat{\sigma_{\hat{\theta}}}\right) = \ln\mathcal{L}_{Max} - \frac{1}{2}$$