Probability and Statistics for Data Science

Y-DATA

30.10.2020

This course

Probability & decision making	How probability is applied in data science Decision making with probabilistic models Descriptive statistics and visualization
Distributions and parameters	Important distributions and their characteristics Methods of parameter estimation Evaluating uncertainty of sample estimates
Hypotheses testing	Tests for comparing means Tests for goodness of fit Sequential tests
Predictive models	Mathematics of joint distributions Linear regression as a statistical tool Inference with nonlinear models

Distributions, their properties, and where to find them

Probability and Statistics for Data Science

Part 2.1

Y-DATA 2020

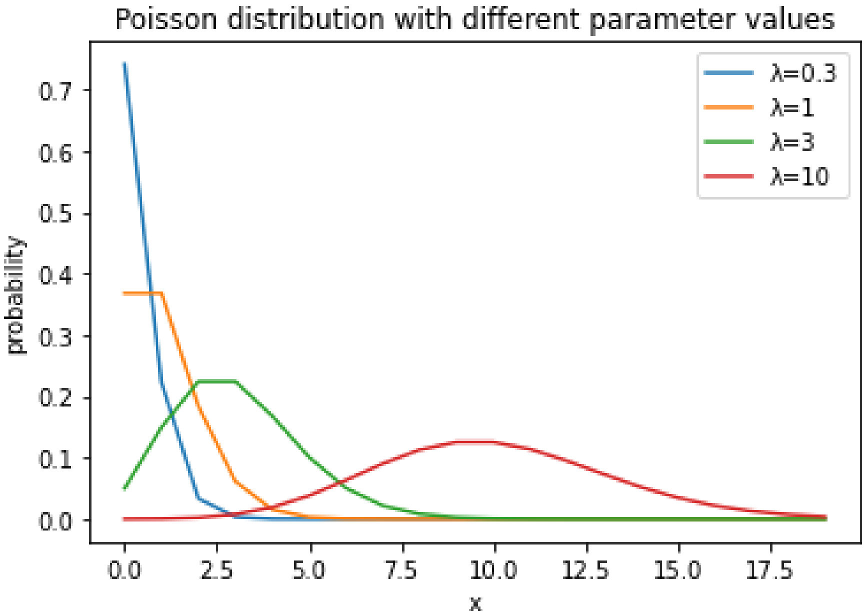
Distributions from the first lecture

- Bernoulli distribution: $P(X = k) = \begin{cases} p, for\ k = 1 \\ 1 - p, for\ k = 0 \end{cases}$
- Uniform distribution: $P(X = k) = \frac{1}{b-a+1}$ for $x \in \{a, a+1, \ldots, b\}$
- Poisson distribution: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k \in \{0, 1, 2, \ldots\}$

- They are actually *families* of distributions: different values of parameters correspond to different distributions
- These distributions are **discrete**: set of possible values is countable
- Another type is continuous, with continuous sets of possible values (e.g. $[0, 1]$ or $[0, +\infty)$ or $(-\infty, +\infty)$).

Members of a distribution family

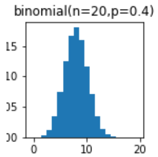
- Parameters of a distribution may determine:
    - Location (what is a typical value)
    - Scale (how diverse is the distribution)
    - Shape (e.g. how fat is the tail, how sharp is the peak, etc.)
- One important problem is to estimate parameters from a small sample to recover the whole curve of the distribution

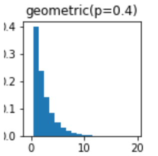


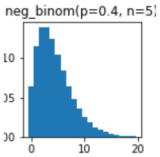
Distributions based on Bernoulli process

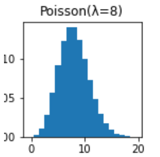
- There is a series of independent experiments with binary outcome
- In each experiment, probability of “success” is p
- The experiments are run until a certain stopping condition

Variable	Distribution	PMF
Number of successes in the first n trials	Binomial	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$
Number of successes before the first failure	Geometric	$f(x) = p^x(1-p)$
Number of successes before the r-th failure	Negative binomial	$f(x) = \binom{x+r-1}{r-1} p^x (1-p)^r$
Number of successes in infinitely many trials with infinitely small p	Poisson	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$ where $\lambda = pn, n \to \infty,\ p \to 0$









In all these formulas, x is the value of a random variable, and p, n, r, λ are *parameters* that determine the exact shape and size of its distribution.

Example: binomial probability

- What is the probability of getting 0, 1 or 2 heads after 10 tosses of a fair coin?
  - $\approx 3.2\%$
  - $\approx 5.5\%$
  - $\approx 27\%$
  - $\approx 30\%$
- Calculate using Binomial(n=5, p=0.5)
  - $P(0) = \binom{10}{0} 0.5^0 (1-0.5)^{10} = 1 \times 0.5^{10} \approx 0.001$
  - $P(1) = \binom{10}{1} 0.5^1 (1-0.5)^9 = 10 \times 0.5^{10} \approx 0.01$
  - $P(2) = \binom{10}{2} 0.5^{10} = \frac{10!}{2!8!} \times 0.5^{10} = 45 \times 0.5^{10} \approx 0.045$

Floods and extreme values

- A house is being built in Mozambique, near Limpopo river. How high above the water should one build it to avoid floods for 50 years with 99% probability?
  - Assume that within a single year t, maximal water level X_t exceeds x meters with probability e^{-x} (an *exponential distribution*)
  - If years are independent, then by definition
  $$P(X_1 \leq x \,\&\, X_2 \leq x) = P(X_1 \leq x)P(X_2 \leq x)$$
  - In general, $P(X_1, \dots, X_{50} \leq x) = (1 - e^{-x})^{50}$
  - Solve $(1 - e^{-x})^{50} = 99\%$, get $x = -\ln\left(1 - 0.99^{\frac{1}{50}}\right) \approx 8.5$
- Important abstractions
  - **Cumulative distribution function** (CDF) for a RV X: $F(x) = P(X \leq x)$
  - $\alpha\%$ **quantile**: the value not-exceeded with $\alpha\%$ probability (i.e. $F^{-1}(\alpha)$)

A question

Lifetime of a lightbulb exceeds x years with probability e^{-x}.
After what time will it burn out with 95% probability?

- About 1 year
- About 2 years
- About 3 years
- About 6 years

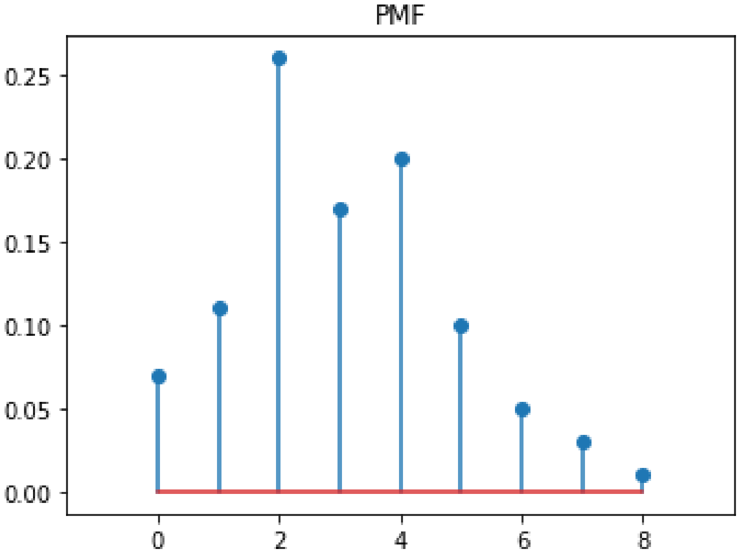
- $CDF(x) = P(X \leq x) = 1 - e^{-x} = 0.95$
  - Then $e^{-x} = 1 - 0.95$
  - Then $x = \ln \frac{1}{0.05} = \ln 20 \approx 3$

PMF and CDF

Probability mass function

$$PMF_{RV}(x) = P(RV = x)$$

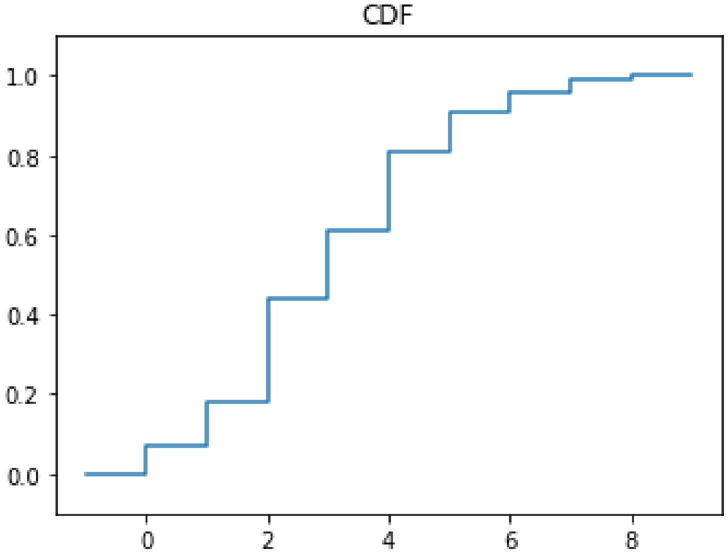
- Non-negative
- Sum of all values is 1
- Is the difference of CDF



Cumulative distribution function

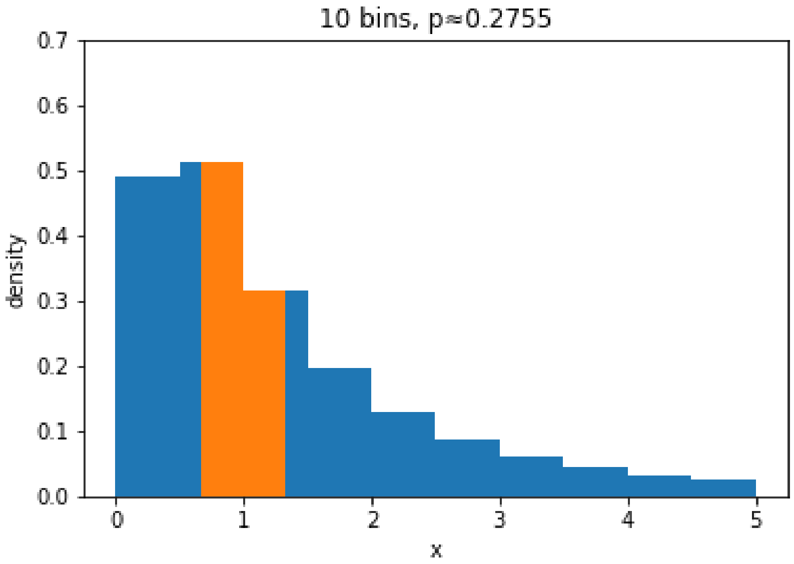
$$CDF_{RV}(x) = P(RV \leq x)$$

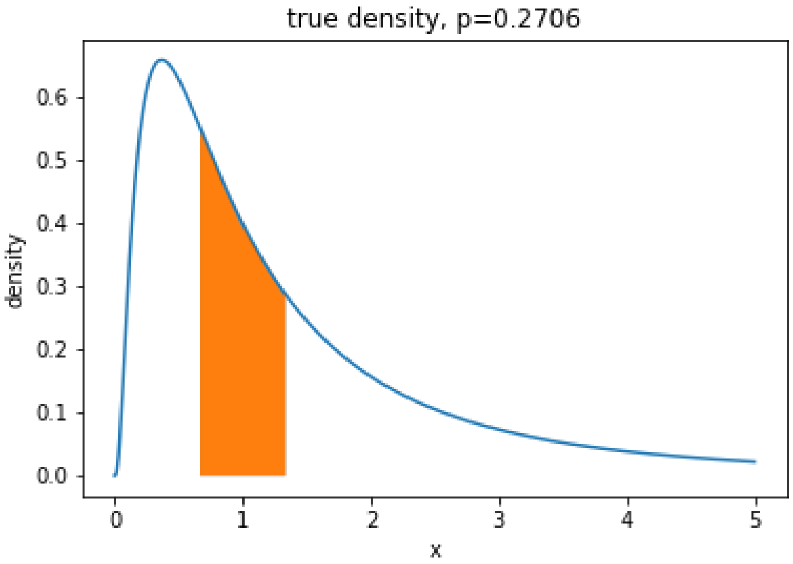
- Non-decreasing
- All values in $[0,1]$
- Is the running sum of PMF



PDF as histogram approximation

- For continuous distributions, probability of any particular point is 0
- However, we can assign probability to intervals
- Within any interval (bin), we can define **density** as its probability divided by its size
- For a composite interval, probability can be calculated as area under density function
- **Probability density function (PDF)** is the limit of this density if we make bins infinitely small

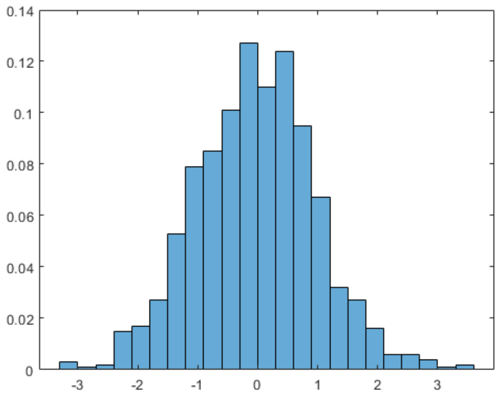

10 bins, p≈0.2755

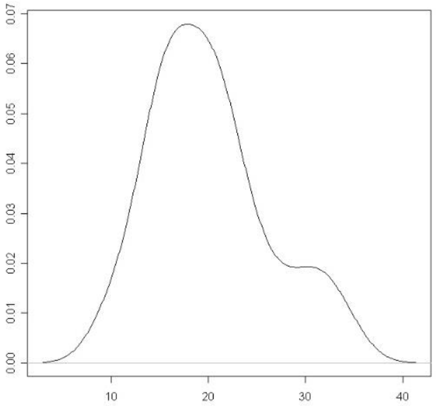

true density, p=0.2706

Density of continuous distributions

- Continuous distributions: how to calculate…
  - Relative likelihood of different values?
  - Probability of intervals?
  - Expected value?
- Use a limit of discrete approximation:
  - split into bins by cutpoints $\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots$
  - for each bin $(x_i, x_{i+1}]$, define density $f_i = \frac{P(x_i < X \leq x_{i+1})}{x_{i+1} - x_i} = \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i}$, where F is the CDF
  - If bins are infinitely small, we can define **probability density function** as $f(x) = \lim_{\Delta \to 0} \frac{P(x < X \leq x + \Delta)}{\Delta} = \lim_{\Delta \to 0} \frac{F(x+\Delta) - F(x)}{\Delta} = \frac{\partial F(x)}{\partial x}$
  - $P(a < X \leq b) \approx \sum_{i: x_i \in (a,b]} f_i \times (x_{i+1} - x_i) = \int_a^b f(x) dx$
  - $\mathbb{E}X \approx \sum_{i=-\infty}^{\infty} x_i \times f_i \times (x_{i+1} - x_i) = \int_{-\infty}^{\infty} x f(x) dx$

Discrete vs continuous





Discrete RV	Continuous RV
X has at most countably many values	X has uncountably many possible values
pmf $\quad p_X(x) = \mathbf{P}(X = x)$	pdf $\quad f_X(x) \geq 0$ (not a proba)
$\sum_{x \in S_X} p_X(x) = 1$	$\int_{-\infty}^{+\infty} f_X(x) dx = 1$
cdf $\quad F_X(x) = \sum_{y \leq x} p_X(y)$	cdf $\quad F_X(x) = \int_{-\infty}^{x} f_X(u) du$
F_X is discontinuous, with jumps at possible values of X	F_X is continuous if f_X has no mass
$\mathbf{P}(x < X \leq y) = \sum_{x < u \leq y} p_X(u)$ $= F_X(y) - F_X(x)$	$\mathbf{P}(x < X \leq y) = \int_{x}^{y} f_X(u) du$ $= F_X(y) - F_X(x)$

Example: continuous expectation

- Let X be continuous random uniform in $[0, 1]$.
  That is, $f(x) = 1$ for $x \in [0, 1]$ and 0 elsewhere
  What is $\mathbb{E}(X^2)$?
  - 0.5
  - 0.25
  - $1/3$
  - $2/3$

- $\mathbb{E}X^2 = \int_0^1 x^2 f(x) dx = \int_0^1 x^2 \times 1 \, dx = \frac{x^3}{3}\Big]_0^1 = \frac{1}{3}$

Example: transforming CDFs

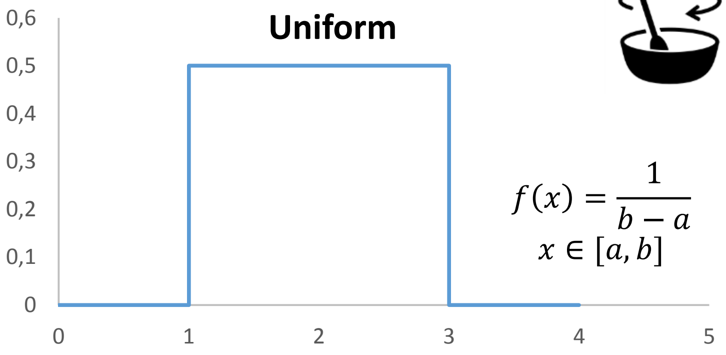
- Let X be random uniform in $[0, 1]$ and $Y = X^2$
- What are CDF and PDF for Y?
- What is the expected value of Y?
- For $y \in [0,1]$,

$$CDF_Y(y) = P(Y \leq y) = P\left(X^2 \leq \sqrt{y}^2\right)$$
$$= P(X \leq \sqrt{y}) = \sqrt{y}$$
$$PDF_Y(y) = \frac{\partial\sqrt{y}}{\partial y} = 0.5y^{-0.5}$$

- $\mathbb{E}Y = \int_0^1 y \times 0.5y^{-0.5} dy = \frac{1}{2} \times \left(\frac{2}{3} - 0\right) = \frac{1}{3}$

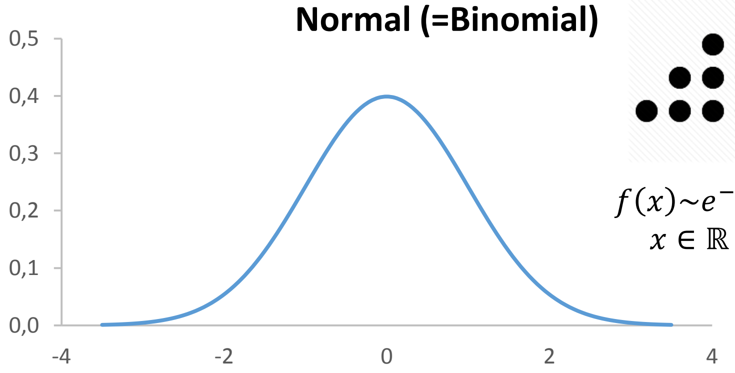
Some continuous distribution families

Each distribution corresponds to some random process



Uniform

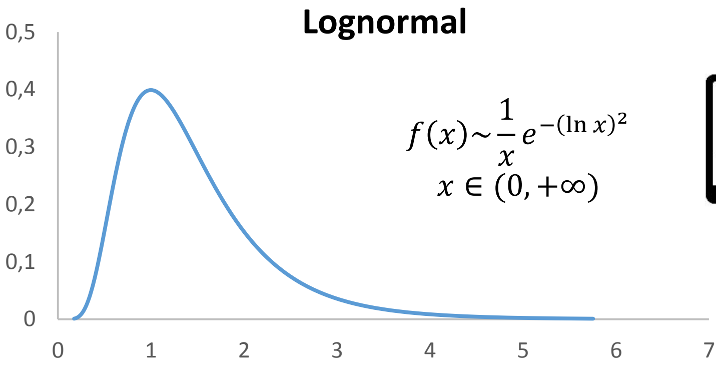
$$f(x) = \frac{1}{b-a}$$
$$x \in [a,b]$$



Normal (=Binomial)

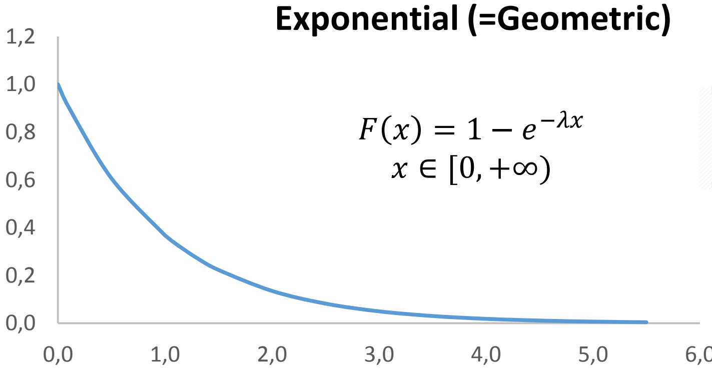
$$f(x) \sim e^{-x^2}$$
$$x \in \mathbb{R}$$

The sign “~” means “proportional to”.
For Normal and Lognormal distributions, location and scale parameters are omitted for simplicity.



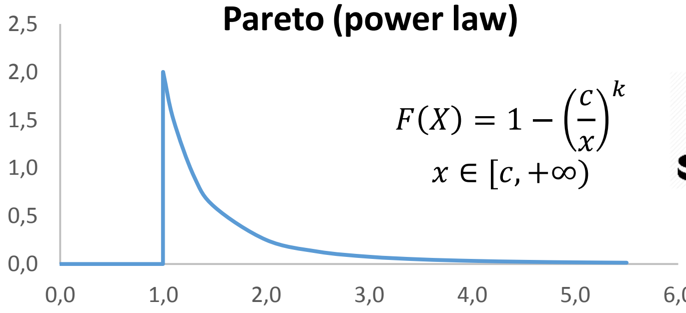
Lognormal

$$f(x) \sim \frac{1}{x} e^{-(\ln x)^2}$$
$$x \in (0,+\infty)$$



Exponential (=Geometric)

$$F(x) = 1 - e^{-\lambda x}$$
$$x \in [0,+\infty)$$



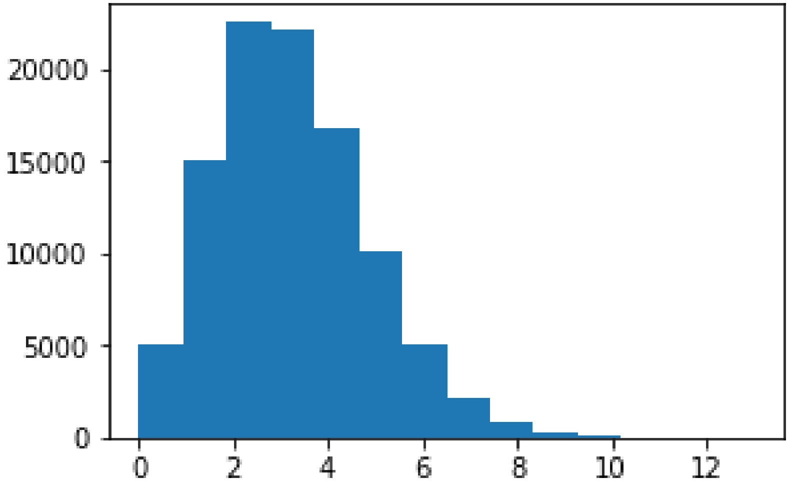
Pareto (power law)

$$F(X) = 1 - \left(\frac{c}{x}\right)^k$$
$$x \in [c,+\infty)$$

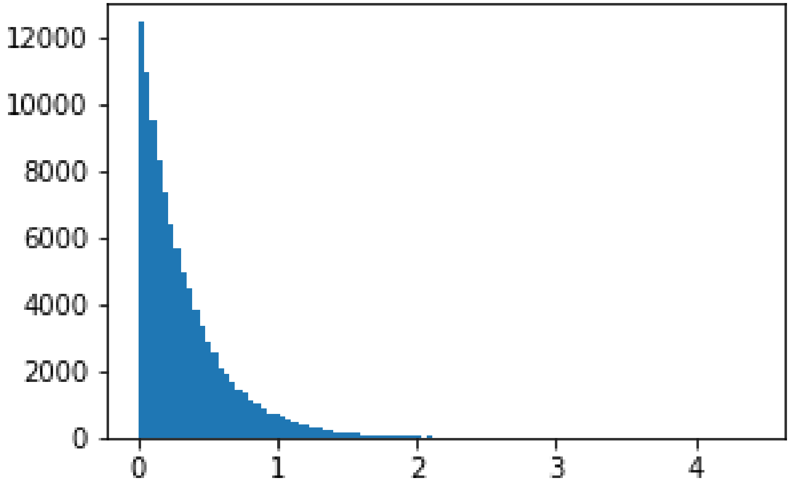
Poisson vs Exponential

- Both Poisson and Exponential distributions can have values from 0 to $+\infty$.
  - But their shapes are different.
  - And Poisson is discrete (integer-valued), whereas Exponential is continuous
- If number of events per unit of time has Poisson distribution with parameter λ, then probability that there are no events on time period t is $S(t) = \frac{(\lambda t)^{0}}{0!} e^{-(\lambda t)} = e^{-(\lambda t)}$
- Then expected time until the first event has CDF $F(t) = 1 - e^{-\lambda t}$
- Mean number of events per unit of time is λ, mean interval between events is $\frac{1}{\lambda}$

Poisson distribution, mean=3

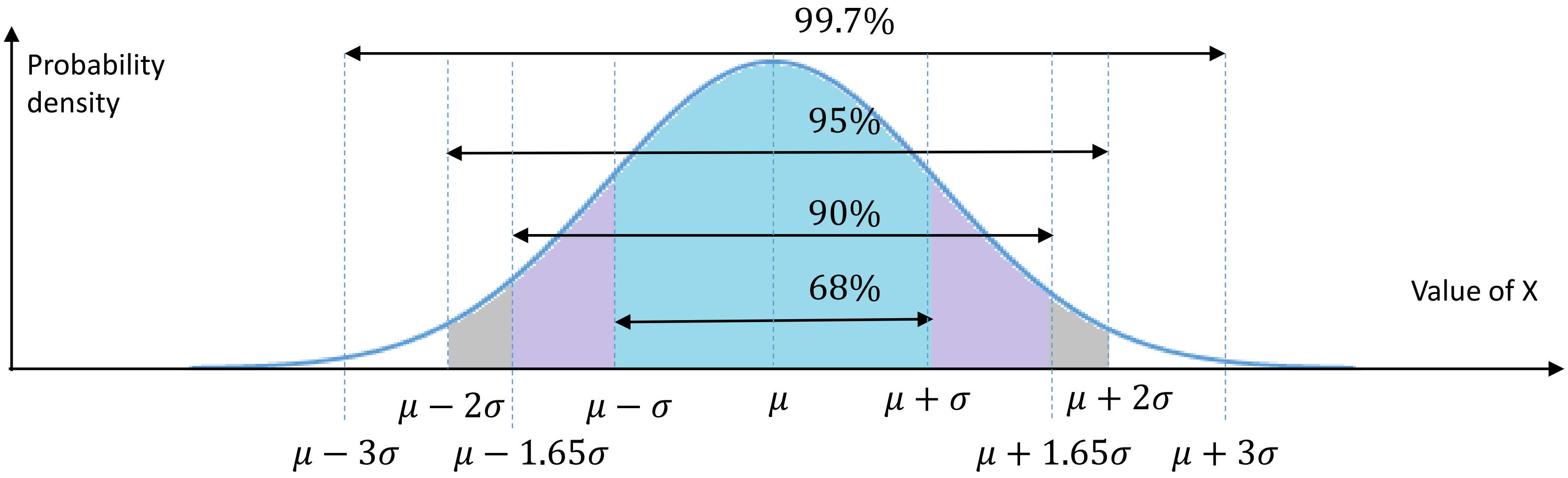


Exponential distribution, mean=1/3



On normal distribution

If a RV is distributed normally, its shape is determined by μ and σ in a specific way: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



This looks terrible, but it's just a limiting case of binomial distribution (sum of many independent coin tosses).

Central limit theorem

Moreover, sum of any n (approximately) independent and (approximately) identical random variables will converge to normal, as we increase n

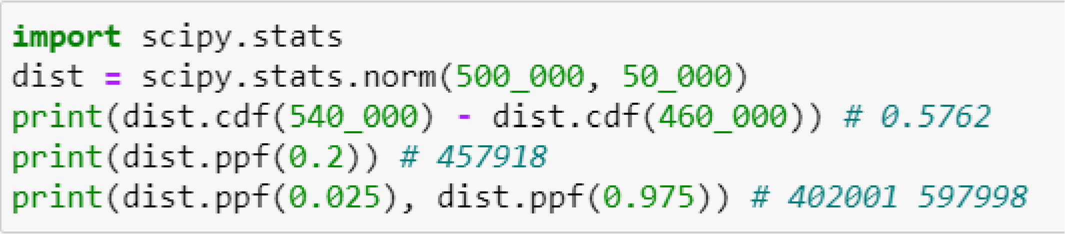
How to calculate with normal distribution

Its CDF equals $F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$
unfortunately, there is no analytic solution for this integral ☹

You have to use pre-calculated numeric solutions, like `NORM.DIST` in Excel, or `scipy.stats.norm.cdf` in Python, or just tables, for normal CDF.

Example: from history, we know that the cost of building a luxury house is a normally distributed random variable with a mean of $500,000 and a standard deviation of $50,000.

- probability that the cost will be between $460,000 and $540,000 equals
  $$CDF(540) - CDF(460) \approx 57\%$$
- 80% houses cost at least $quantile(0.2) \approx \$458{,}000$
- 95% houses cost roughly between $\mu - 2\sigma = \$400{,}000$ and $\mu + 2\sigma = \$600{,}000$



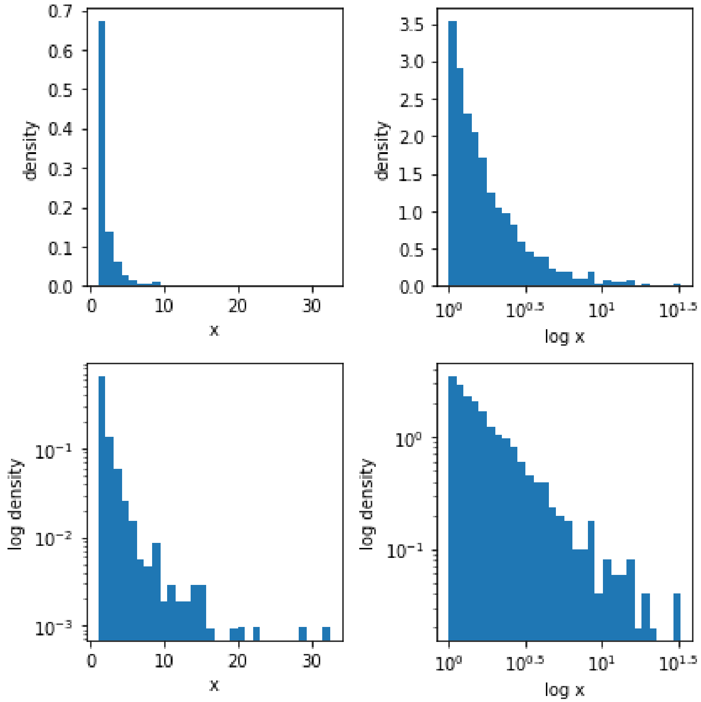
```python
import scipy.stats
dist = scipy.stats.norm(500_000, 50_000)
print(dist.cdf(540_000) - dist.cdf(460_000)) # 0.5762
print(dist.ppf(0.2)) # 457918
print(dist.ppf(0.025), dist.ppf(0.975)) # 402001 597998
```

Infinite moments

- In what family of distributions there are members with infinitely large variance?
  - Normal
  - Poisson
  - Exponential
  - Pareto
- The slower is the convergence of tail density, the higher is the potential value of the integral
- In the limit, $\frac{1}{\mathrm{x}} > \frac{1}{e^{x}} > \frac{1}{x!} > \frac{1}{e^{x^2}}$, so Pareto tails are the fattest

Pareto distribution

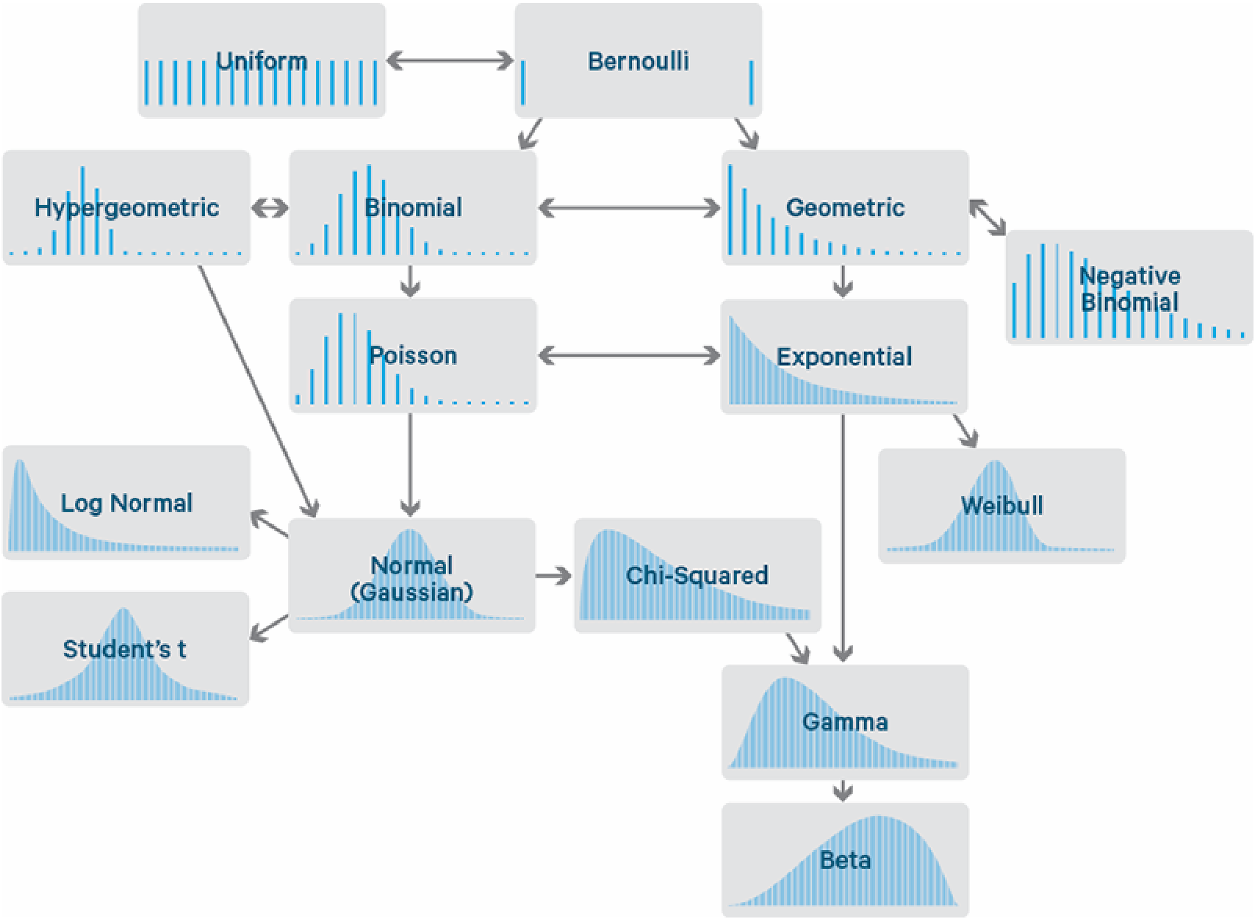
- $F(X) = 1 - \left(\frac{x}{a}\right)^{-b}, x \in [a, +\infty), a > 0, b > 0$
- It is so-called "power law" that can describe extremely fat right tails:
  - Size of a city
  - Frequency of a word in a language corpus
  - Number of citations of a scientific paper
  - Number of subscribers in a social network
  - …
- The underlying process:
  positive feedback loop ("rich get richer")
- Well visualized in log-log scale
  (pdf and 1-cdf are linear)

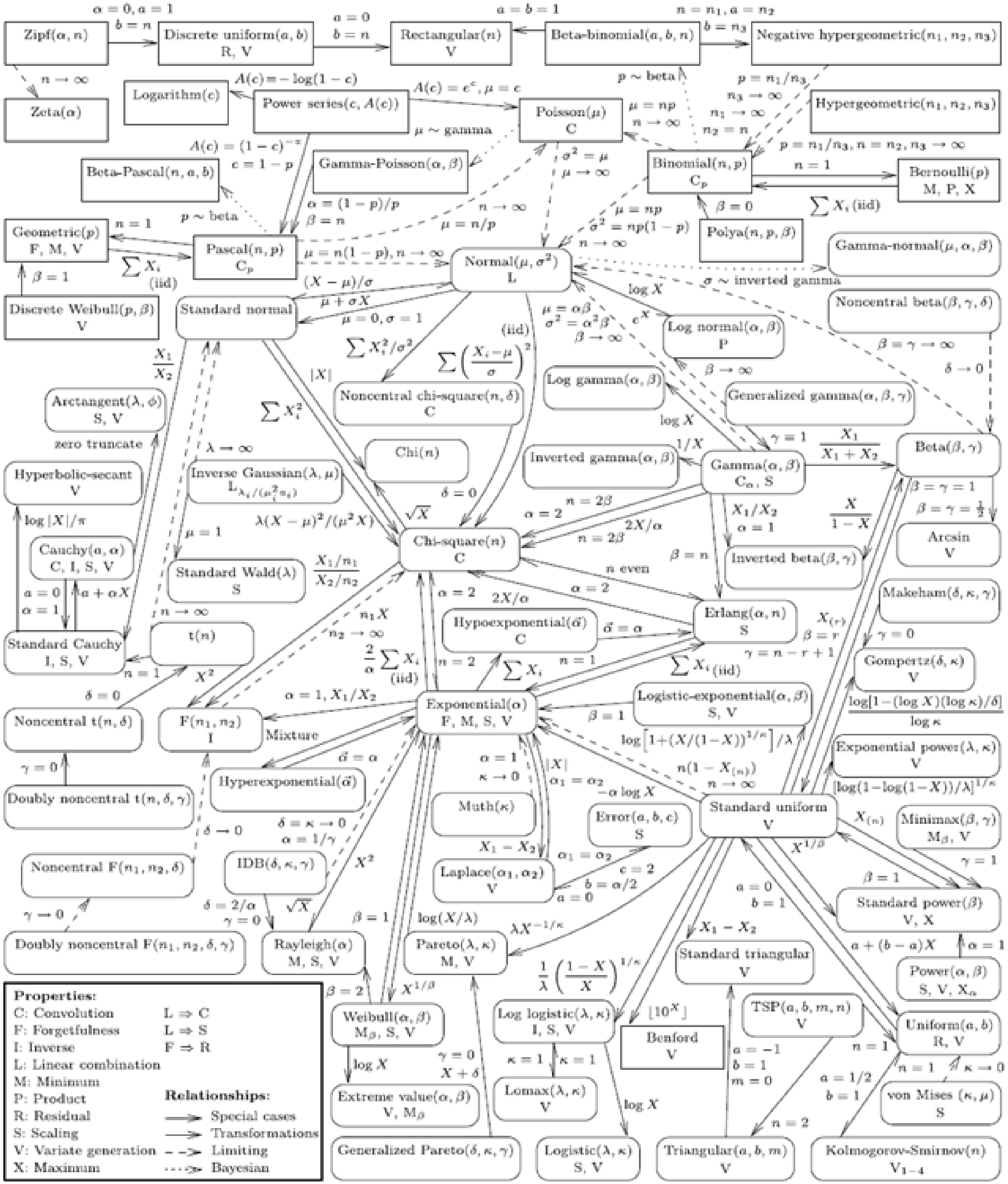


Pareto moments

- Let $F_X(x) = \begin{cases} 1-\left(\frac{a}{x}\right)^k, for\ x \geq a \\ 0, for\ x < a \end{cases}$
- Then $f_X(x) = a^k k x^{-k-1}[x \geq a]$
- Then $\mathbb{E}X = \int_a^\infty a^k k x^{-k} dx = a^k k \left(\frac{x^{-k+1}}{-k+1}\right)_a^\infty = \frac{ak}{k-1}$
  - This integral converges only if $k > 1$
- $\mathbb{E}X^2 = \int_a^\infty a^k k x^{-k+1} dx = a^k k \left(\frac{x^{-k+2}}{-k+2}\right)_a^\infty = \frac{a^2 k}{k-2}$
  - The variance exists (is finite) only if $k > 2$
- For small k, tails are so fat that moments are infinite!

Relations between distributions





Zipf(α, n)
$\alpha = 0, a = 1$
$b = n$
Discrete uniform(a, b)
R, V
$a = 0$
$b = n$
Rectangular(n)
V
$a = b = 1$
Beta-binomial(a, b, n)
$n = n_1, a = n_2$
$b = n_3$
Negative hypergeometric(n_1, n_2, n_3)
$n \to \infty$
Zeta(α)
Logarithm(c)
$A(c) = -\log(1 - c)$
Power series($c, A(c)$)
$A(c) = e^c, \mu = c$
Poisson(μ)
C
$p \sim$ beta
$p = n_1/n_3$
$n_3 \to \infty$
$n_1 \to \infty$
$n_2 = n$
Hypergeometric(n_1, n_2, n_3)
$\mu = np$
$n \to \infty$
$\mu \sim$ gamma
$A(c) = (1 - c)^{-x}$
$c = 1 - p$
Beta-Pascal(n, a, b)
Gamma-Poisson(α, β)
$\sigma^2 = \mu$
$\mu \to \infty$
Binomial(n, p)
C_p
$p = n_1/n_3, n = n_2, n_3 \to \infty$
$n = 1$
Bernoulli(p)
M, P, X
$\beta = 0$
$\sum X_i$ (iid)
$p \sim$ beta
$\alpha = (1 - p)/p$
$\beta = n$
$n \to \infty$
$\mu = n/p$
$\mu = np$
$\sigma^2 = np(1 - p)$
$n \to \infty$
Polya(n, p, β)
Geometric(p)
F, M, V
$n = 1$
Pascal(n, p)
C_p
$\sum X_i$
(iid)
$\mu = n(1 - p), n \to \infty$
Normal(μ, σ^2)
L
Gamma-normal(μ, α, β)
$\sigma \sim$ inverted gamma
$\beta = 1$
Discrete Weibull(p, β)
V
$(X - \mu)/\sigma$
$\mu + \sigma X$
$\mu = 0, \sigma = 1$
Standard normal
$\log X$
e^X
Noncentral beta(β, γ, δ)
$\mu = \alpha\beta$
$\sigma^2 = \alpha^2\beta$
$\beta \to \infty$
Log normal(α, β)
P
$\beta = \gamma \to \infty$
$\delta \to 0$
$\sum X_i^2/\sigma^2$
(iid)
$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2$
$\frac{X_1}{X_2}$
$|X|$
$\beta \to \infty$
Log gamma(α, β)
Generalized gamma(α, β, γ)
Arctangent(λ, ϕ)
S, V
Noncentral chi-square(n, δ)
C
$\sum X_i^2$
$\log X$
zero truncate
$\lambda \to \infty$
Chi(n)
$1/X$
$\gamma = 1$
$\frac{X_1}{X_1 + X_2}$
Beta(β, γ)
Hyperbolic-secant
V
Inverse Gaussian(λ, μ)
$\mathrm{L}_{\lambda_i/(\mu_i^2 \alpha_i)}$
$\delta = 0$
Inverted gamma(α, β)
Gamma(α, β)
C_α, S
$\beta = \gamma = 1$
$\beta = \gamma = \frac{1}{2}$
$\log |X|/\pi$
$\mu = 1$
$\lambda(X - \mu)^2/(\mu^2 X)$
$\sqrt{X}$
$\alpha = 2$
$n = 2\beta$
$2X/\alpha$
$n = 2\beta$
X_1/X_2
$\alpha = 1$
$\frac{X}{1 - X}$
Arcsin
V
Cauchy(a, α)
C, I, S, V
Standard Wald(λ)
S
$\frac{X_1/n_1}{X_2/n_2}$
Chi-square(n)
C
$\beta = n$
Inverted beta(β, γ)
n even
$\alpha = 2$
Makeham(δ, κ, γ)
$a = 0$
$\alpha = 1$
$a + \alpha X$
$n \to \infty$
$\alpha = 2$
$2X/\alpha$
Erlang(α, n)
S
$X_{(r)}$
$\beta = r$
$\gamma = n - r + 1$
$\gamma = 0$
Standard Cauchy
I, S, V
t(n)
$n_1 X$
$n_2 \to \infty$
Hypoexponential($\vec{\alpha}$)
C
$\vec{\alpha} = \alpha$
Gompertz(δ, κ)
V
$n = 1$
X^2
$\frac{2}{\alpha}\sum X_i$
(iid)
$n = 2$
$\sum X_i$
$n = 1$
$\sum X_i$ (iid)
$\delta = 0$
$\alpha = 1, X_1/X_2$
Exponential(α)
F, M, S, V
$\beta = 1$
Logistic-exponential(α, β)
S, V
$\log[1 - (\log X)(\log \kappa)/\delta]$
$\log \kappa$
Noncentral t(n, δ)
F(n_1, n_2)
I
Mixture
$\log\left[1 + (X/(1 - X))^{1/\kappa}\right]/\lambda$
Exponential power(λ, κ)
V
$\gamma = 0$
$\vec{\alpha} = \alpha$
$\alpha = 1$
$\kappa \to 0$
$|X|$
$\alpha_1 = \alpha_2$
$-\alpha \log X$
$n(1 - X_{(n)})$
$n \to \infty$
$[\log(1 - \log(1 - X))/\lambda]^{1/\kappa}$
Doubly noncentral t(n, δ, γ)
Hyperexponential($\vec{\alpha}$)
Muth(κ)
Error(a, b, c)
S
Standard uniform
V
$X_{(n)}$
Minimax(β, γ)
M_β, V
$\delta \to 0$
$\delta = \kappa \to 0$
$\alpha = 1/\gamma$
$X_1 - X_2$
$\alpha_1 = \alpha_2$
$X^{1/\beta}$
$\gamma = 1$
Noncentral F(n_1, n_2, δ)
IDB(δ, κ, γ)
X^2
Laplace(α_1, α_2)
V
$c = 2$
$b = \alpha/2$
$a = 0$
$\beta = 1$
Standard power(β)
V, X
$\delta = 2/\alpha$
$\gamma = 0$
$\sqrt{X}$
$a = 0$
$b = 1$
$\gamma \to 0$
$\beta = 1$
$\log(X/\lambda)$
$\lambda X^{-1/\kappa}$
$X_1 - X_2$
$a + (b - a)X$
$\alpha = 1$
Doubly noncentral F(n_1, n_2, δ, γ)
Rayleigh(α)
M, S, V
Pareto(λ, κ)
M, V
Standard triangular
V
Power(α, β)
S, V, X_α
$\frac{1}{\lambda}\left(\frac{1 - X}{X}\right)^{1/\kappa}$
$\beta = 2$
$X^{1/\beta}$
Weibull(α, β)
M_β, S, V
Log logistic(λ, κ)
I, S, V
$\lfloor 10^X \rfloor$
Benford
V
TSP(a, b, m, n)
V
Uniform(a, b)
R, V
$\log X$
$\gamma = 0$
$X + \delta$
$\kappa = 1$
$\kappa = 1$
$a = -1$
$b = 1$
$m = 0$
$n = 1$
$n = 1$
$\kappa \to 0$
$a = 1/2$
$b = 1$
Extreme value(α, β)
V, M_β
Lomax(λ, κ)
V
$\log X$
$n = 2$
von Mises (κ, μ)
S
Generalized Pareto(δ, κ, γ)
Logistic(λ, κ)
S, V
Triangular(a, b, m)
V
Kolmogorov-Smirnov(n)
V_{1-4}
Properties:
C: Convolution
F: Forgetfulness
I: Inverse
L: Linear combination
M: Minimum
P: Product
R: Residual
S: Scaling
V: Variate generation
X: Maximum
L ⇒ C
L ⇒ S
F ⇒ R
Relationships:
Special cases
Transformations
Limiting
Bayesian

How to choose distributions?

- Filter distributions by domain (discrete vs continuous, min/max values) and shape (symmetry, tail fatness)
- Compare shape of histograms and 1-CDF in linear and log space
- Estimate the parameters (see the next section)
- Compare theoretical and sample moments
- Make sure that quantile-quantile plot is roughly linear

Moments

Moments are (normalized) expectations of X^k

mean: $\mathbb{E}(X) \quad = \mu$

variance: $\mathbb{E}\left((X-\mu)^2\right) \quad = \sigma^2 = Var(X)$

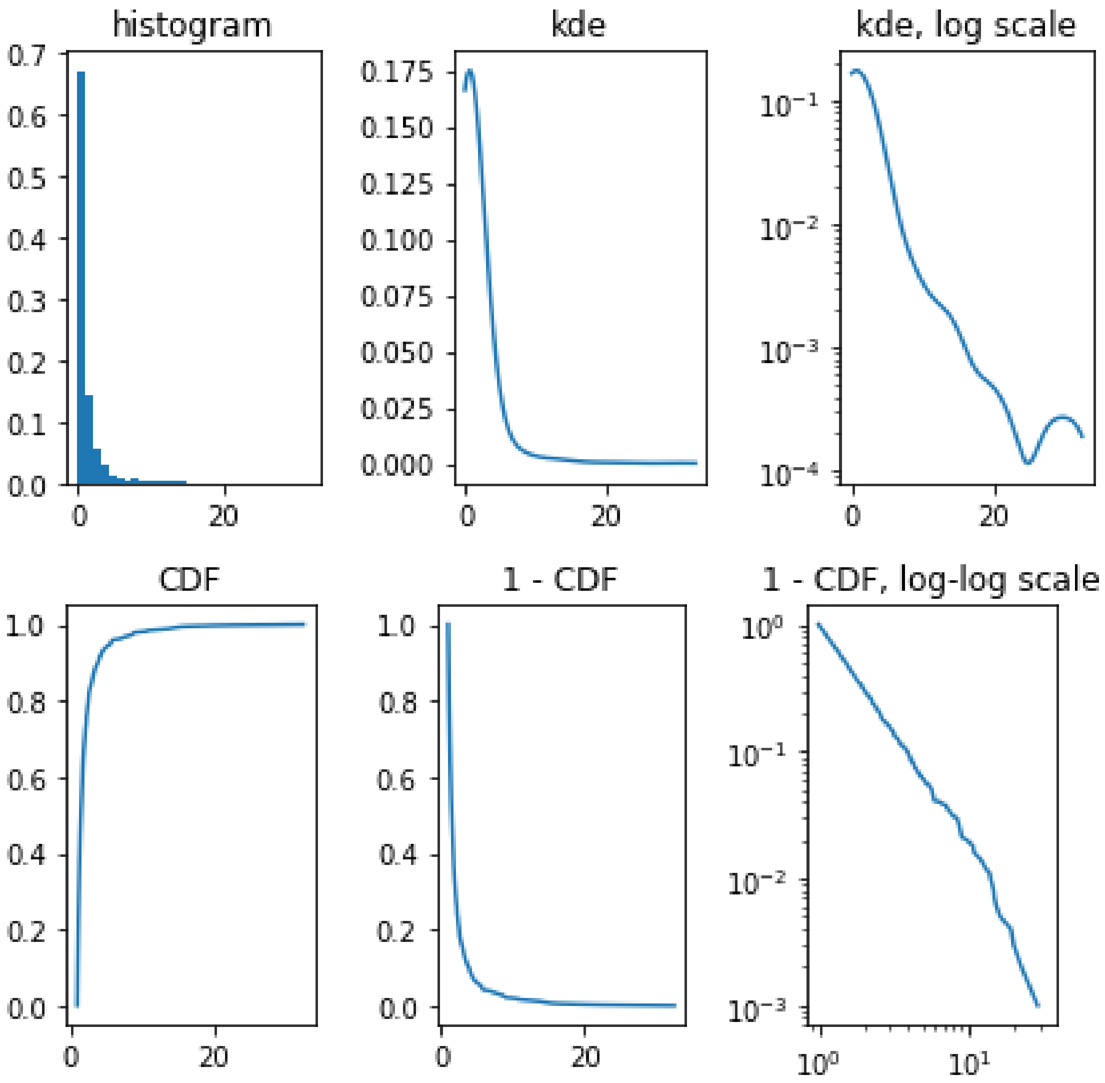
skewness: $\mathbb{E}\left(\left(\frac{X-\mu}{\sigma}\right)^3\right)$

shows asymmetry: positive → right tail larger than the left one

kurtosis: $\mathbb{E}\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3$ -3 for “excess kurtosis”

shows sharpness: positive → sharp peak and long tails

Plotting distributions



histogram

kde

kde, log scale

CDF

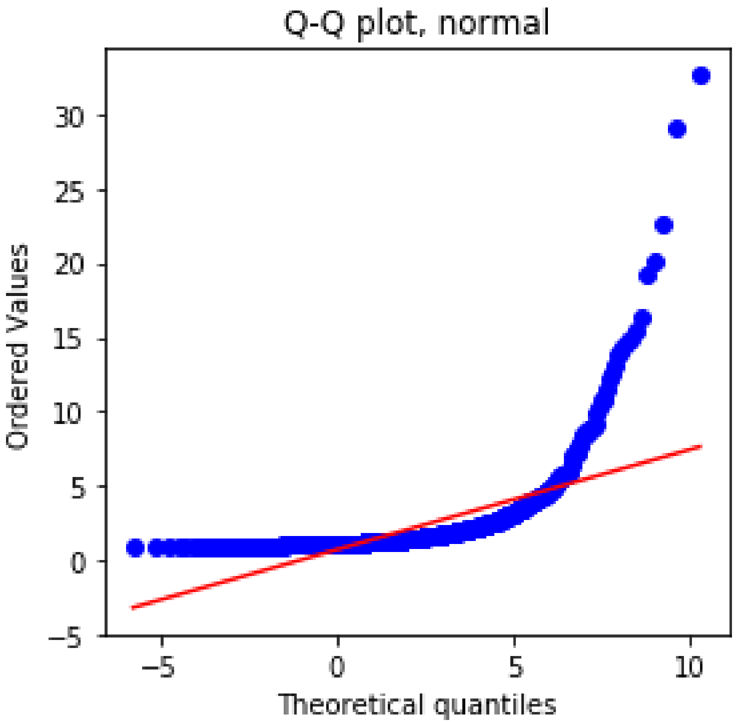
1 - CDF

1 - CDF, log-log scale

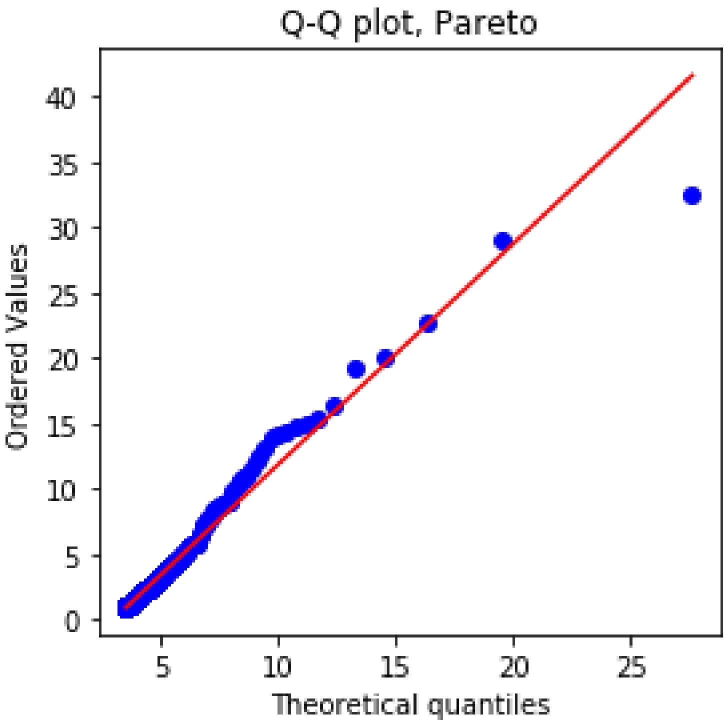
*kde is kernel density estimation,
a way of numeric smoothing of histograms

Quantile-quantile plot

- Plot quantiles of the theoretical distribution against the same quantiles of the sample
- If the distribution is correct, the scatter will be approximately linear



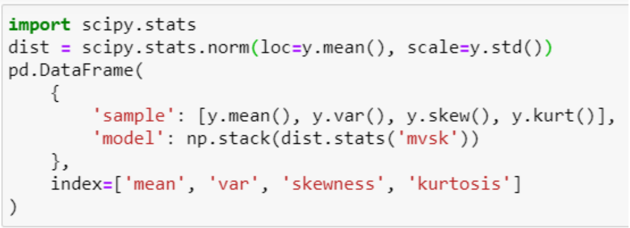
Q-Q plot, normal



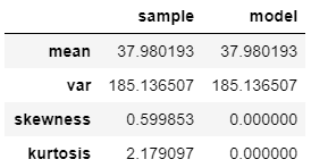
Q-Q plot, Pareto

Example: choosing the right distribution

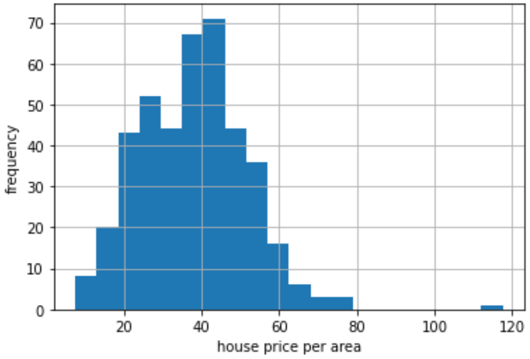
- Data on house prices:
  https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set
- Looks like normal distribution:
  continuous, symmetric, hat-shaped
- The price-making process does look like
  a sum of many independent factors,
  so the nature of normal distribution fits well
- 3rd and 4th moments are not very close
- Q-Q plot is straight except for one outlier

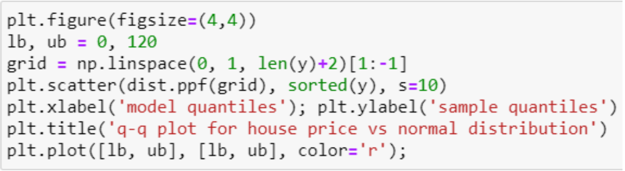


```
import scipy.stats
dist = scipy.stats.norm(loc=y.mean(), scale=y.std())
pd.DataFrame(
    {
        'sample': [y.mean(), y.var(), y.skew(), y.kurt()],
        'model': np.stack(dist.stats('mvsk'))
    },
    index=['mean', 'var', 'skewness', 'kurtosis']
)
```

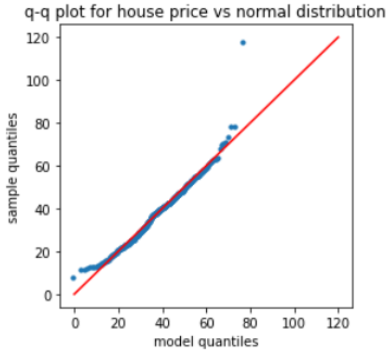


	sample	model
mean	37.980193	37.980193
var	185.136507	185.136507
skewness	0.599853	0.000000
kurtosis	2.179097	0.000000





```
plt.figure(figsize=(4,4))
lb, ub = 0, 120
grid = np.linspace(0, 1, len(y)+2)[1:-1]
plt.scatter(dist.ppf(grid), sorted(y), s=10)
plt.xlabel('model quantiles'); plt.ylabel('sample quantiles')
plt.title('q-q plot for house price vs normal distribution')
plt.plot([lb, ub], [lb, ub], color='r');
```



Estimating parameters

Probability and Statistics for Data Science

Part 2.2

Y-DATA 2020

Probability vs statistics

Probability studies properties of random variables with known distributions and their parameters – so called **populations**

Statistics tries to infer something about distributions using **samples** of data, taken from these populations.

Usually it is assumed that samples are taken "randomly" – independently and with equal chances

Properties of expectation

$$\mathbb{E}X = \sum_x xP(X = x)$$

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

If X and Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

$$\mathbb{E}\big(g(X)\big) = \sum_x \boldsymbol{g}(x) \times P_X(x)$$

Properties of variance

$$Var(X) = \mathbb{E}\left((X - \mathbb{E}X)^2\right)$$

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$Var(aX + b) = a^2 Var(X)$$

For two (and more) variables,

$$Var(X + Y) = Var(X) + Var(Y) + 2\underbrace{\mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y))}_{\text{Covariance } Cov(X,Y)}$$

If they are independent,

$$Var(X + Y) = Var(X) + Var(Y)$$

Example: binomial variance

What is $\mathrm{Var}(X)$ if X is binomial with parameters n and p?

- p^2
- np
- $np(1-p)$
- $p(1-p)$

$X = X_1 + \cdots + X_n$, where each X_i equals 1 with probability p, otherwise 0.

$$Var(X_i) = \mathbb{E}(X_i^2) - (\mathbb{E}(X_i))^2 = p - p^2 = p(1-p)$$

By definition of Bernoulli experiment, all X_i are independent of each other.

Then $\mathrm{Var}(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$

Distribution of sample mean

Let $X_1, X_2, \dots X_n$ be a sample from a RV X with mean μ and variance σ^2

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathbb{E}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

On average, $\bar{X} = \mu$

$$Var(\bar{X}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

In large samples, standard deviation of $\bar{X}$ (from μ) goes to 0

So what is a good way to estimate parameters?

We call *estimator* a function from sample $X = (X_1, \dots, X_n)$
to value $\hat{\theta}$ that estimates a population parameter θ

- It's a function of a random sample, so it's a random variable itself.

A few good properties of estimators are:

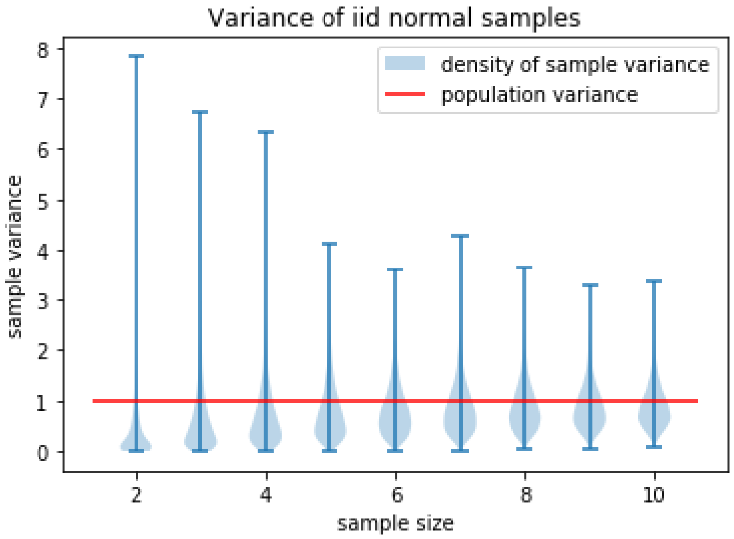
- $\mathbb{E}\hat{\theta} = \theta$ – unbiasedness
  - *In multiple samples, on average the estimate is not too high and not too low*
- $MSE = \mathbb{E}\left(\left(\hat{\theta} - \theta\right)^2\right)$ is as low as possible – efficiency
  - *In multiple samples, on average the estimate close to the truth*
  - By the way, $MSE = Var\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta} - \theta\right)^2$ = variance + bias^2
- $\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| > a\right) = 0$ – consistency
  - *We can make the estimate very accurate by increasing sample size*

Sample variance

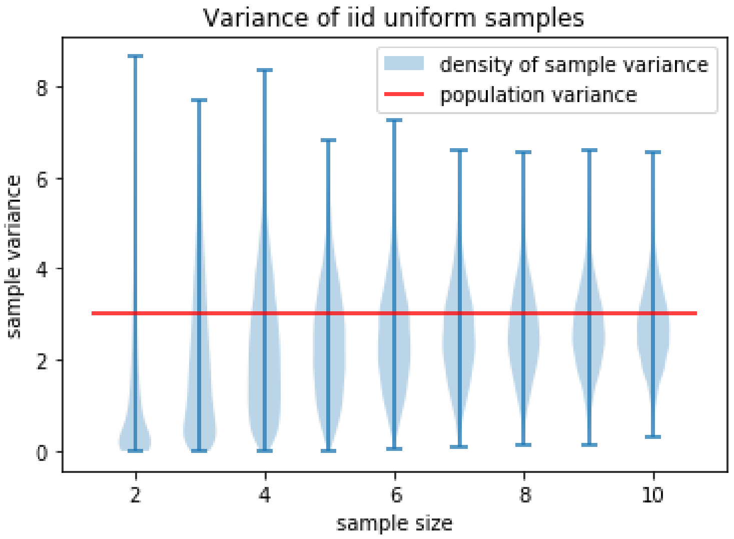
$$\tilde{S}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Alas, it is biased! $\mathbb{E}\tilde{S}^2 = \frac{n-1}{n}\sigma^2$

Let’s demonstrate it by simulation.



Variance of iid normal samples



Variance of iid uniform samples

Sample variance, why?

The formula $\frac{1}{n}\sum(X_i - \mu)^2$ is unbiased, but we don’t know μ.

The formula $\frac{1}{n}\sum(X_i - \bar{X})^2$ is biased, because $\bar{X}$ “tries” to be as close as possible to the sample values.

So we can just use unbiased $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ instead.

This 1 in the denominator is the single *degree of freedom* – it means that our model for X has a single parameter that tries to fit the data (and slightly overfits it).

More complex models (e.g. with conditional means) have more degrees of freedom and overfit harder.

Method of moments

- Moments are quantities such as $\mathbb{E}(X), Var(X)$, and $\mathbb{E}(X^k)$ or $\mathbb{E}((X-\mathbb{E}X)^k)$ for various natural k
- Sample moments can be easily estimated
- Population moments can be derived from parameters
- We can estimate parameters by solving the equation $sample\ moments = population\ moments$
- E.g. for exponential distribution ($f(x) = \lambda e^{-\lambda x}$) mean is $\frac{1}{\lambda}$
    - so we can estimate it by solving $\frac{1}{\hat{\lambda}} = \bar{X}$, or $\hat{\lambda} = \frac{1}{\bar{X}}$
- If we have multiple parameters, we can write a system of equations for as many moments

Example: method of moments

- Suppose X is continuous uniform on $[a, b]$, and sample mean and variance are 7 and 12. Estimate a and b.
  - 0 and 14
  - 1 and 13
  - 2 and 12
- $\mathbb{E}X = \frac{a+b}{2}$, let it be 7, then $a + b = 14$
- $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{(b-a)^2}{12}$, then $(b - a)^2 = 12^2$
  - $\mathbb{E}(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{x^3}{3(b-a)}\Big]_a^b = \frac{(b^2+ab+a^2)}{3}$
  - $(\mathbb{E}X)^2 = \left(\frac{a+b}{2}\right)^2 = \frac{(b^2+2ab+a^2)}{4}$
- Solve the system, get $b = 13$, $a = 1$.

The likelihood approach

Bayesian parameter estimation

So you say the parameter θ is unknown.

Let's *imagine* it is random and has *prior* distribution $p(\theta)$.

Then we can use the Bayes formula:

$$p(\theta|data) = \frac{1}{Z}p(\theta)p(data|\theta)$$

The function $L(\theta) = p(data|\theta)$ is called likelihood.

If you don't know the prior distribution $p(\theta)$ make it flat: assume $p(\theta) = const$, and just *maximize the likelihood*.

You can do it *even if you believe that θ is not random*.

Maximum likelihood

It’s just the probability (or density) of the data given the parameter.

If the observations are independent, then equals the product of individual probabilities (densities):

$$L(\theta) = p(dataset|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

In practice, it’s usually easier to maximize log-likelihood

$$LL(\theta) = \log \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

In fact, negative log likelihood (aka cross entropy) is the most popular loss function in machine learning

Maximization: analytically / with gradient descent / more complex tricks

Maximum likelihood | binomial

Let’s take a Binomial distribution: n independent experiments with p probability of success.
We have k successes, what is the most likely p?

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

$$\log P(k) = const + k \log p + (n-k) \log(1-p)$$

$$\frac{\partial \log P(k)}{\partial p} = \frac{k}{\hat{p}} - \frac{n-k}{1-\hat{p}} = 0$$

$$\hat{p} = k/n$$

Maximum likelihood | uniform

What is the ML estimate for the parameter a of uniform distribution on $[0, a]$?

$$L(a) = \prod_{i=1}^{n} \frac{1}{a} \left[x_i \in [0, a] \right] \to \max$$

This value is 0, if any x_i is larger than a, but otherwise it is a^{-n} and increases as a decreases.

Therefore, it is maximal when $\hat{a} = \max(x_1, \ldots x_n)$

Maximum likelihood | exponential

What is the ML estimate of λ for exponential distribution?

Exponential density is $f(x) = \lambda e^{-\lambda x}$

$$\ln(L) = \sum_{i=1}^{n} \ln \lambda e^{-\lambda x_i} = \sum_{i=1}^{n} (\ln \lambda - \lambda x_i) = n \ln \lambda - \lambda n \bar{x}$$

This function (you can sketch it) is smooth and has a single extremum (maximum) – the derivative there must be 0. Solve it:

$$\frac{\partial \ln(L)}{\partial \lambda} = \frac{n}{\lambda} - n\bar{x} = 0$$

Therefore, our estimate is $\hat{\lambda} = \frac{1}{\bar{x}}$

Maximum likelihood | normal

What are ML estimates for μ and σ^2 of normal distribution?

Normal density: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\ln L = \sum \left(-\frac{1}{2} \ln \pi - \frac{1}{2} \ln \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$\frac{\partial \ln L}{\partial \mu} = \sum_i -\frac{1}{2\sigma^2} \times 2(x_i - \mu) \times (-1) = 0$$

By solving it, we obtain $\hat{\mu} = \frac{1}{n} \sum x_i$ - quite expected

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_i -\frac{1}{2\sigma^2} + \frac{(x_i - \mu)^2}{2\sigma^4} = 0$$

By solving it and replacing μ with $\hat{\mu}$, we get $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{\mathrm{x}})^2$ - a biased but consistent estimate

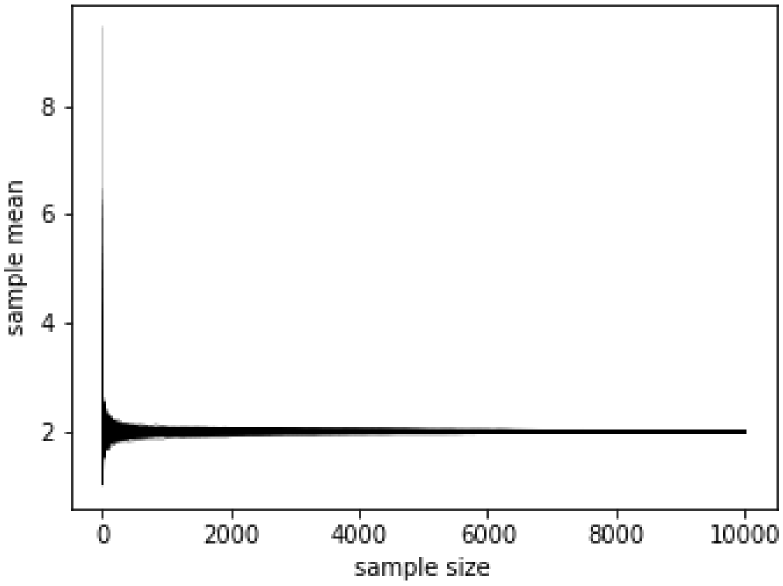
What’s so good about MLE

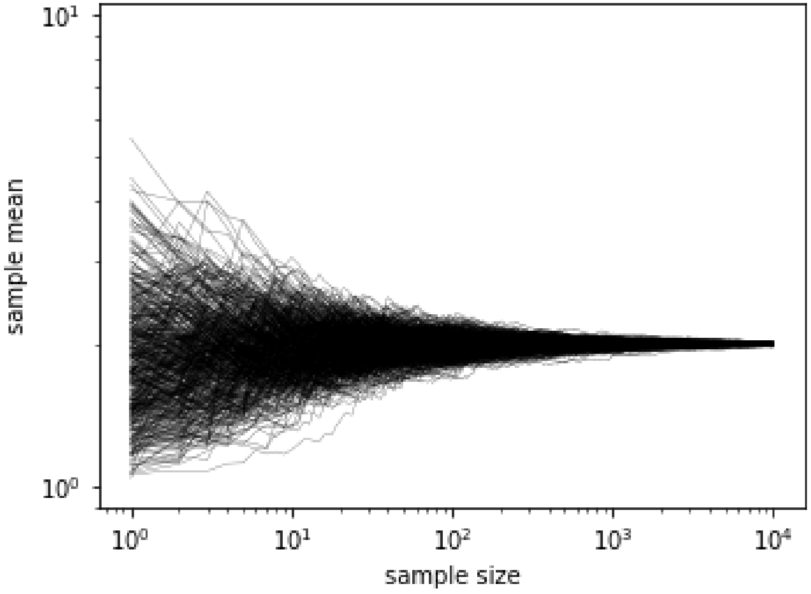
- In large samples (asymptotically), these estimates are
  - Unbiased
  - Efficient
  - Normally distributed
    - With covariance matrix $-H^{-1}$, where H (Hessian) is the second derivative of log likelihood function
  - But only if the optimum of likelihood is internal
- They don’t depend on parametrization
- They can be combined with prior information

Sample means

Sample mean vs expectation

We know that sample mean is “usually close” to the expected value in the population.
What does it exactly mean?
Experiment: increase sample size, track the sample mean





The law of large numbers

It’s a theorem.

If we take an i.i.d. sample of n random variables $X_1, \dots X_n$ with mean μ and variance σ^2,
and estimate its sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,
then $\lim\limits_{n\to\infty} P(|\bar{X} - \mu| > a) = 0$ for any $a > 0$

Thus, we can estimate *means* as precisely as we want – we need only a sample that is large enough.

But *probability* of any event is just the mean of the corresponding Bernoulli random variable.

Thus, we can estimate probabilities as precisely as we want.

And if we can estimate probabilities, we can estimate everything.

Proof of the LLN

If you want, just believe it. But just in case you don't:

Markov inequality: if RV X is non-negative, then for any $a > 0$

$$\mathbb{E}X = P(X \le a)\mathbb{E}(X|X \le a) + P(X > a)\mathbb{E}(X|X > a) \ge aP(X > a)$$

Thus, $P(X > a) \le \frac{\mathbb{E}X}{a}$

Chebyshev inequality: for any RV X with mean and variance μ and σ^2,

$$P(|X - \mu| > a) = P\left((X - \mu)^2 > a^2\right) \le \frac{\sigma^2}{a^2}$$

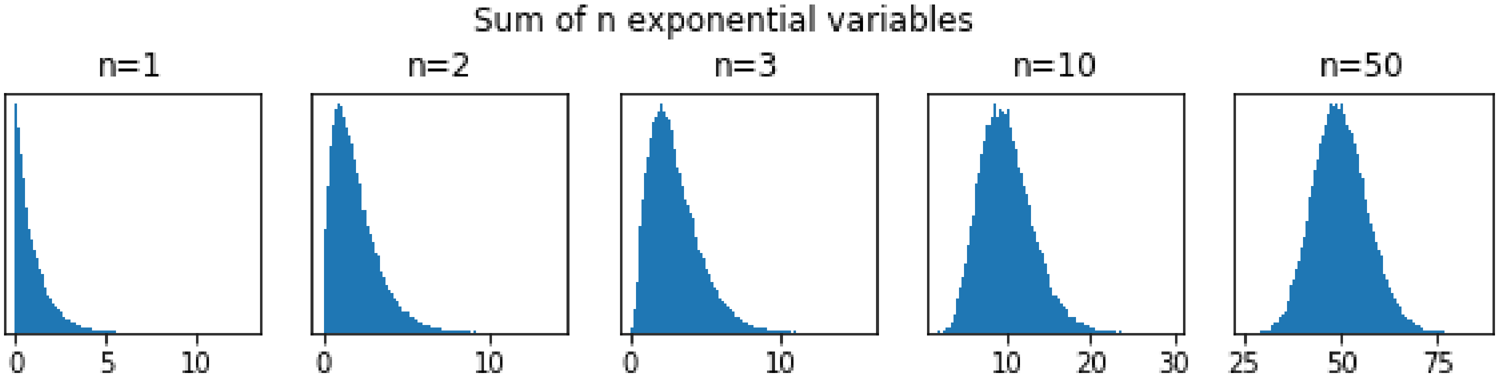
(because the variable $(X - \mu)^2$ is non-negative and has mean σ^2)

Now, if $X_1, \dots, X_k$ are independent, then $\mathbb{E}\bar{X} = \mu, Var(\bar{X}) = \frac{\sigma^2}{k}$, so

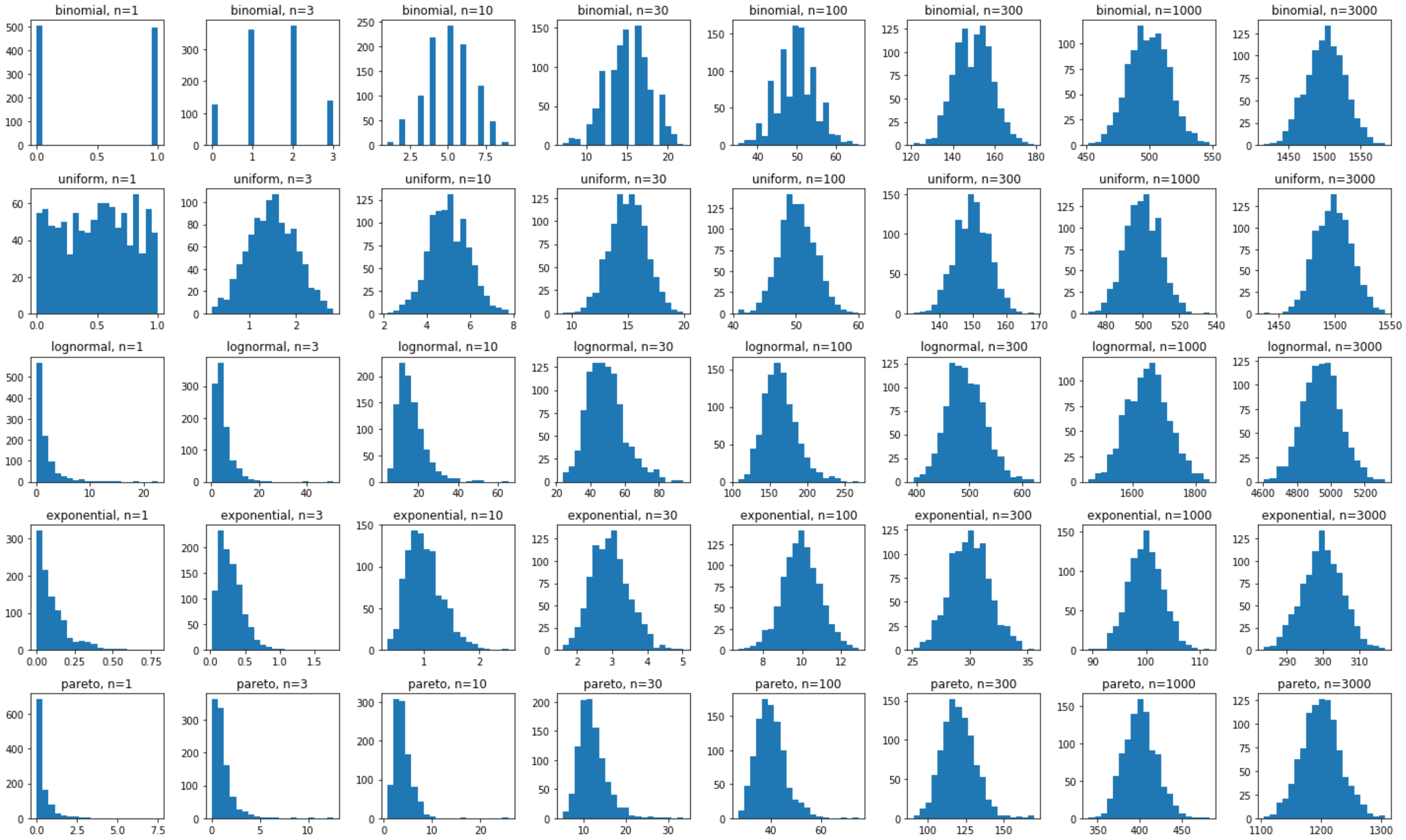
$$P(|\bar{X} - \mu| > a) \le \frac{\sigma^2}{a^2 k} \to 0, as\ k \to \infty$$

The Central limit theorem

- Take an i.i.d sample $X_1, \dots, X_n$ from (almost) any distribution with (population) mean μ and variance σ^2
- Then sum $\mathrm{S_n} = \sum_{i=1}^{n} X_i$ has mean μn and variance $\sigma^2 n$
- Normalize to avoid infinity: $\tilde{X}_i = \frac{X_i - \mu}{\sigma}$, $\tilde{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{X}_i$ has mean 0 and variance 1. But other parameters of shape are uncertain.
- At $n \to \infty$, the shape (e.g. CDF) of $\tilde{S}_n$ converges to normal.
- In practice, for $n \geq 50$, the distribution is often "normal enough"

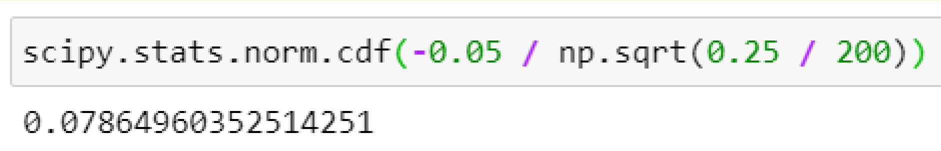


CLT, more examples



How can I use this all?

- Example: for testing hypotheses
  - Example. We believe that 50% of galactic population vote for Darth Vader.
    How likely is it that in a sample of 200 citizens Vader scores 90 votes or less?
  - 0.1% ? 3% ? 8% ? 42%?
- The exact distribution is binomial, but it can be approximated with normal precisely enough
  - $\mu = p = 0.5, \sigma^2 = p(1-p) = 0.5^2, \frac{\sigma^2}{n} = \frac{0.25}{200} \approx 0.035^2$
  - Sample mean 0.45 is then 1.4 standard deviations below the population mean, which has probability $\approx 8\%$
- Therefore, after such a poll we can still believe that in the whole population Vader has 50% support, but we just got into an 8% unlucky tail



```
scipy.stats.norm.cdf(-0.05 / np.sqrt(0.25 / 200))
```

```
0.07864960352514251
```

Interval estimation

Example: the Darth Vader problem

- We know that in a random sample of 200 galactic citizens 90 said they would vote for Vader.
- 45% is the sample mean, but what population mean can be?
- We know that approximately, $(\bar{x} - \mu) \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$, and $\sigma^2 \approx S^2 = \frac{\bar{x}(1-\bar{x})}{n-1} \approx 0.035^2$.
- Then, with $\approx 95\%$ probability, $(\bar{x} - \mu) \in [-0.07; +0.07]$
- Or we can say that with $\approx 95\%$ probability, $\mu \in [38\%, 52\%]$
- Of course, μ does not have to be random, it is just unknown
- It is *the interval* that is random

Confidence intervals

- Because parameter estimate $\hat{\theta}$ is a function of random sample, it is a random variable itself
    - And the probability that $\hat{\theta} = \theta$ is usually 0
- We can also make a random interval $[l, u]$ that covers θ with some high probability α
    - It is called α-confidence interval
- Example: normal confidence interval for mean
    - We may believe that $\bar{\mathrm{x}} \sim \mathcal{N}(\mu, \sigma^2/n)$ (because $x \sim \mathcal{N}$ or CLT)
    - $P(\mu - c\sigma/\sqrt{n} \leq \bar{x} \leq \mu + c\sigma/\sqrt{n}) = \alpha$ for some α and c
    - Then $P(\bar{x} - c\hat{\sigma}/\sqrt{n} \leq \mu \leq \bar{x} + c\hat{\sigma}/\sqrt{n}) \approx \alpha$ as well
        - For small n, this is inaccurate estimate. We can make it better by adjusting c.
- Remember: it's l and u that are random, and μ is not!
    - At least, from frequentist point of view
    - From Bayesian view, everything is random…

Sample size planning

A psychologist believes that the standard deviation of driver's reaction time is about 0.05 seconds, and wants to estimate the mean.

How large a sample of measurements must be taken to derive a confidence interval for the mean with *margin of error* (radius of CI) at most 0.01 second, and confidence level 95%?

$$(\bar{x} + 1.96\sigma/\sqrt{n}) - (\bar{x} - 1.96\sigma/\sqrt{n}) \approx 4\sigma/\sqrt{n} \leq 0.01 \times 2$$

$$n \approx \left(\frac{4\sigma}{0.02}\right)^2 = 10^2 = 100$$

Sample size planning

You expect that about half of generated texts contain errors, and want to estimate this proportion from a sample.
You want the 95% confidence interval to be $\pm 1\%$, i.e. to have width of 2%.
How large a sample do you need?
100? 1000? 10’000? 100’000?

95% normal CI for mean is approximately $\bar{X} \pm 2\sqrt{\frac{\sigma^2}{n}}$,

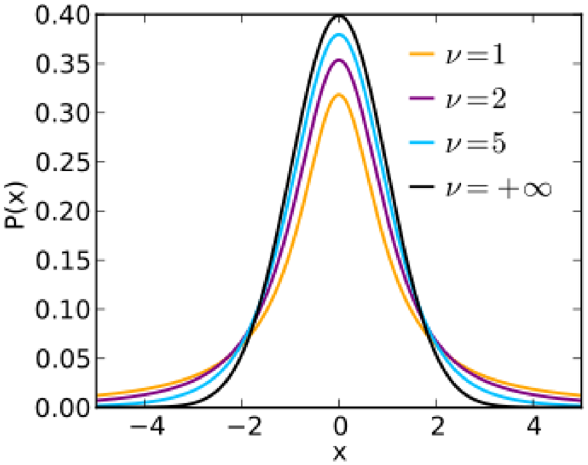
so you want $1\% = 2\sqrt{\frac{\sigma^2}{n}}$, or $\frac{\sigma^2}{n} = \frac{1}{200^2}$.

Because the original distribution is binomial with $p \approx 0.5$, you have $\sigma^2 \approx 0.5^2$.

Therefore you need $n = 200^2 \times 0.5^2 = 10000$

Student distribution

- For CI construction, we used the fact that $\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1\,)$
- However, when we replace unknown (non-random) σ^2 with its random estimate, then $t = \frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}}$ is *no longer normal*
  - Intuition: for small n, $\hat{\sigma}^2$ can be unusually small (which creates heavy tails of t's distribution) or unusually large (which creates a sharp peak)
  - For large n, however, $\hat{\sigma}^2$ converges to constant and t converges to normal distribution
- The distribution of t is called Student (T^ν).
- It has a single parameter $\nu = n - 1$, called *degrees of freedom*
- t.dist in Excel, scipy.stats.t in Python
- For small normal samples, Student distribution helps to make much more accurate confidence intervals



Prediction interval

- Confidence interval tries to cover some parameter of distribution
- Prediction interval tries to cover the random variable itself – the whole distribution
- In general, it’s not that easy
- For normal distribution, there is a recipe:
  - Decompose $X - \bar{X}$ into independent $(X - \mu)$ and $(\mu - \bar{X})$
  - Their estimated variances are S^2 and $\frac{S^2}{n}$, respectively, and means are 0
  - Because of independence they can be added, so that $\frac{X-\bar{X}}{\sqrt{S^2(1+1/n)}} \sim T^{n-1}$
- Example: if $n = 10, S^2 = 110, \bar{X} = 20$, then we can predict that the next observation will fall into $20 \pm 2\sqrt{110 \times 1.1}$ with approximately 95% probability.

MLE confidence interval

- In large samples, maximum likelihood are distributed normally with variance $-H^{-1}$, where H (Hessian) is the second derivative of log likelihood function
- Continue with the Binomial example:

$$\frac{\partial \log P(k)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p}$$

$$H = \frac{\partial^2 \log P(k)}{\partial p^2} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2}$$

- At the point of maximum, $\hat{p} = \frac{k}{n}$, the estimate of Hessian looks like

$$\hat{H} = -\frac{k}{k^2/n^2} - \frac{n-k}{(n-k)^2/n^2} = -\left(\frac{n^2}{k} + \frac{n^2}{n-k}\right) = -\frac{n^3}{k(n-k)} = -\frac{n}{\hat{p}(1-\hat{p})}$$

- And we can use it to express the variance of our parameter estimate

$$\hat{\sigma}_{\hat{p}}{}^2 = -\hat{H}^{-1} = \frac{\hat{p}(1-\hat{p})}{n}$$

- Now we can use it for e.g. 95% normal CI for p: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Bootstrapping

- For some estimators, it's difficult to obtain analytical sample distribution (and CI)
- Solution: approximate *sampling from population* by *sampling from the sample itself*
    - Sample with replacement (to get "independence")
    - Same size as the original sample
- Example: toy bootstrapping for sample mean
    - Original sample: $1, 5$ -> sample mean is 3
    - Resamples: $\{1,1\}, \{1,5\}, \{5,5\}, \{5,1\}$ with equal probabilities, and sample means of $1, 3, 3, 5$ respectively.
    - Thus, variance of sample mean can be estimated as $\frac{4+0+0+4}{4} = 2$
    - BTW, an analytic unbiased estimate would be $\frac{1}{2}\left(\frac{2^2+2^2}{2-1}\right) = 2$

What we have learned

- Distributions of random variables can be described with parametric functions (CDF and PMF/PDF)
- There are many ways of estimating distribution parameters from samples, including methods of moments and of maximum likelihood
- Sample estimates are uncertain, and this uncertainty can be expressed with confidence intervals

What to do next

- The home assignment:
  - 5 paper-and-pencil exercises (simple models)
  - One programming assignment (parameter estimation)
- Recommended reading
  - *A Modern Introduction to Probability and Statistics* by F.M. Dekking - chapters 4-8, 13-14, 17-21, 23-24
  - Seeing Theory (https://seeing-theory.brown.edu): chapters 3-5