1. write a definition of all data distributions

Data distributions describe how probabilities are assigned to possible values of a random variable. Here are definitions of all major statistical distributions, organized by discrete and continuous types.

Discrete Distributions

Bernoulli Distribution
Single binary trial (success/failure).
Parameters: p*p* (success probability)
PDF: P(X=x)=px(1−p)1−x*P*(*X*=*x*)=*px*(1−*p*)1−*x*, where x=0,1*x*=0,1

Binomial Distribution
Number of successes in n*n* independent Bernoulli trials.
Parameters: n*n* (trials), p*p* (success probability)
PDF: P(X=k)=(nk)pk(1−p)n−k*P*(*X*=*k*)=(*kn*)*pk*(1−*p*)*n*−*k*

Poisson Distribution
Number of events in fixed interval at constant rate.
Parameters: λ*λ* (average rate)
PDF: P(X=k)=λke−λk!*P*(*X*=*k*)=*k*!*λke*−*λ*

Geometric Distribution
Trials until first success.
Parameters: p*p* (success probability)
PDF: P(X=k)=(1−p)k−1p*P*(*X*=*k*)=(1−*p*)*k*−1*p*

Negative Binomial Distribution
Trials until r*r* successes.
Parameters: r*r* (successes), p*p* (success probability)
PDF: P(X=k)=(k−1r−1)pr(1−p)k−r*P*(*X*=*k*)=(*r*−1*k*−1)*pr*(1−*p*)*k*−*r*

Discrete Uniform Distribution
Equal probability over finite set.
Parameters: a,b*a*,*b* (min, max values)
PDF: P(X=k)=1b−a+1*P*(*X*=*k*)=*b*−*a*+11 for k=a,...,b*k*=*a*,...,*b*

Continuous Distributions

Uniform Distribution
Equal probability over interval.
Parameters: a,b*a*,*b* (lower, upper bounds)
PDF: f(x)=1b−a*f*(*x*)=*b*−*a*1 for a≤x≤b*a*≤*x*≤*b*

Normal (Gaussian) Distribution Symmetric bell-shaped curve.
Parameters: μ*μ* (mean), σ*σ* (standard deviation)
PDF: f(x)=1σ2πe−(x−μ)22σ2*f*(*x*)=*σ*2*π*1*e*−2*σ*2(*x*−*μ*)2

Student's t-Distribution
Normal-like with heavier tails.
Parameters: ν*ν* (degrees of freedom)
Used: Small sample inference

Chi-Square Distribution
Sum of squared standard normals.
Parameters: k*k* (degrees of freedom)
PDF: f(x)=12k/2Γ(k/2)xk/2−1e−x/2*f*(*x*)=2*k*/2Γ(*k*/2)1*xk*/2−1*e*−*x*/2

F-Distribution
Ratio of two chi-square variables.
Parameters: d1,d2*d*1,*d*2 (numerator, denominator df)
Used: Variance comparisons, ANOVA

Exponential Distribution
Time between Poisson events.
Parameters: λλ (rate) or β=1/λ$\beta=1/\lambda$ (scale)
PDF: f(x)=λe−λx$f(x)=\lambda e-\lambda x$ for x≥0$x\geq0$

Gamma Distribution
Generalizes exponential (waiting times).
Parameters: αα (shape), ββ (rate)
PDF: f(x)=βαΓ(α)xα−1e−βx$f(x)=\Gamma(\alpha)\beta\alpha x\alpha-1e-\beta x$

Beta Distribution
Probabilities or proportions (0 to 1).
Parameters: α,βα,β (shape parameters)
PDF: f(x)=xα−1(1−x)β−1B(α,β)$f(x)=B(\alpha,\beta)x\alpha-1(1-x)\beta-1$

Log-Normal Distribution
Logarithm is normally distributed.
Parameters: μ,σμ,σ (log-scale mean, SD)
Used: Stock prices, incomes

Weibull Distribution
Time-to-failure modeling.
Parameters: kk (shape), λλ (scale)
PDF: f(x)=kλ(xλ)k−1e−(x/λ)k$f(x)=\lambda k(\lambda x)k-1$

2. understand when to use which data distribution

Choosing the right statistical distribution requires matching the data generation process to the appropriate probability model. For discrete data representing single binary outcomes like coin flips or pass/fail tests, use the Bernoulli distribution with parameter p (success probability). When dealing with a fixed number of independent trials such as 10 customer conversions out of 100 visits, apply the Binomial distribution with parameters n (trials) and p (success probability). Count data for rare events over fixed intervals—like customer calls per hour or defects per batch—follows the Poisson distribution characterized by λ (average rate).

For continuous data, natural measurements like heights, weights, IQ scores, or errors typically follow the Normal distribution with parameters μ (mean) and σ (standard deviation) due to the Central Limit Theorem. Right-skewed waiting times between events, such as time until next customer arrival or machine failure, use the Exponential distribution with rate parameter λ. Proportions or probabilities bounded between 0 and 1, like click-through rates or batting averages, require the Beta distribution with shape parameters α and β. Product lifetimes and reliability analysis often employ the Weibull distribution with shape k and scale λ parameters to model varying failure rates.

The key decision framework starts by identifying whether data is discrete or continuous, followed by checking bounds (0-1 for Beta, 0-∞ for Exponential/Gamma), examining skewness (symmetric for Normal, right-skewed for Log-normal), and considering the data generation mechanism. For example, customer arrivals follow Poisson while time between arrivals follows Exponential; stock returns may be Normal but prices follow Log-normal due to multiplicative growth. Small sample inference when population variance is unknown uses the t-distribution with degrees of freedom ν, while variance comparisons employ the F-distribution.

3. what are the parameters used in the data distribution

Each statistical distribution is defined by specific parameters that control its shape, location, and scale. Here's a comprehensive list of parameters for major distributions:

Discrete Distributions:

- Bernoulli: p (probability of success, $0 \leq p \leq 1$)
- Binomial: n (number of trials), p (success probability)
- Poisson: λ (lambda, average rate of events)
- Geometric: p (success probability)

- Negative Binomial: r (number of successes), p (success probability)
- Discrete Uniform: a (minimum value), b (maximum value)

Continuous Distributions:

- Uniform: a (lower bound), b (upper bound)
- Normal: μ (mu, mean), σ (sigma, standard deviation)
- Student's t: ν (nu, degrees of freedom)
- Chi-Square: k (degrees of freedom)
- F-Distribution: d1 (numerator df), d2 (denominator df)

Shape-Controlled Distributions:

- Exponential: λ (rate parameter) or $\beta = 1/\lambda$ (scale)
- Gamma: α (alpha, shape), β (beta, rate) or $\theta = 1/\beta$ (scale)
- Beta: α (alpha, shape1), β (beta, shape2)
- Log-Normal: μ (log-mean), σ (log-standard deviation)
- Weibull: k (kappa, shape), λ (lambda, scale)