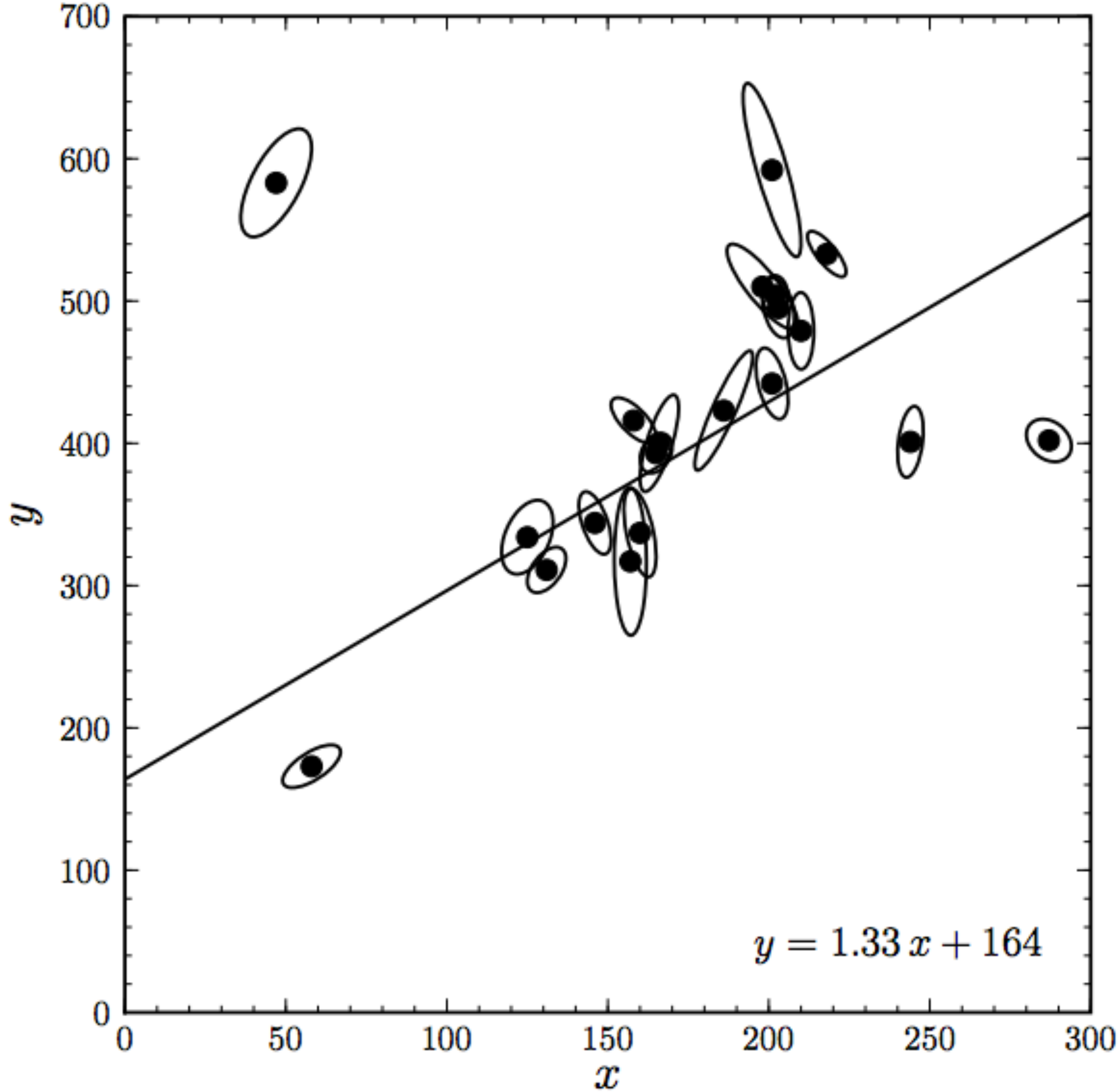
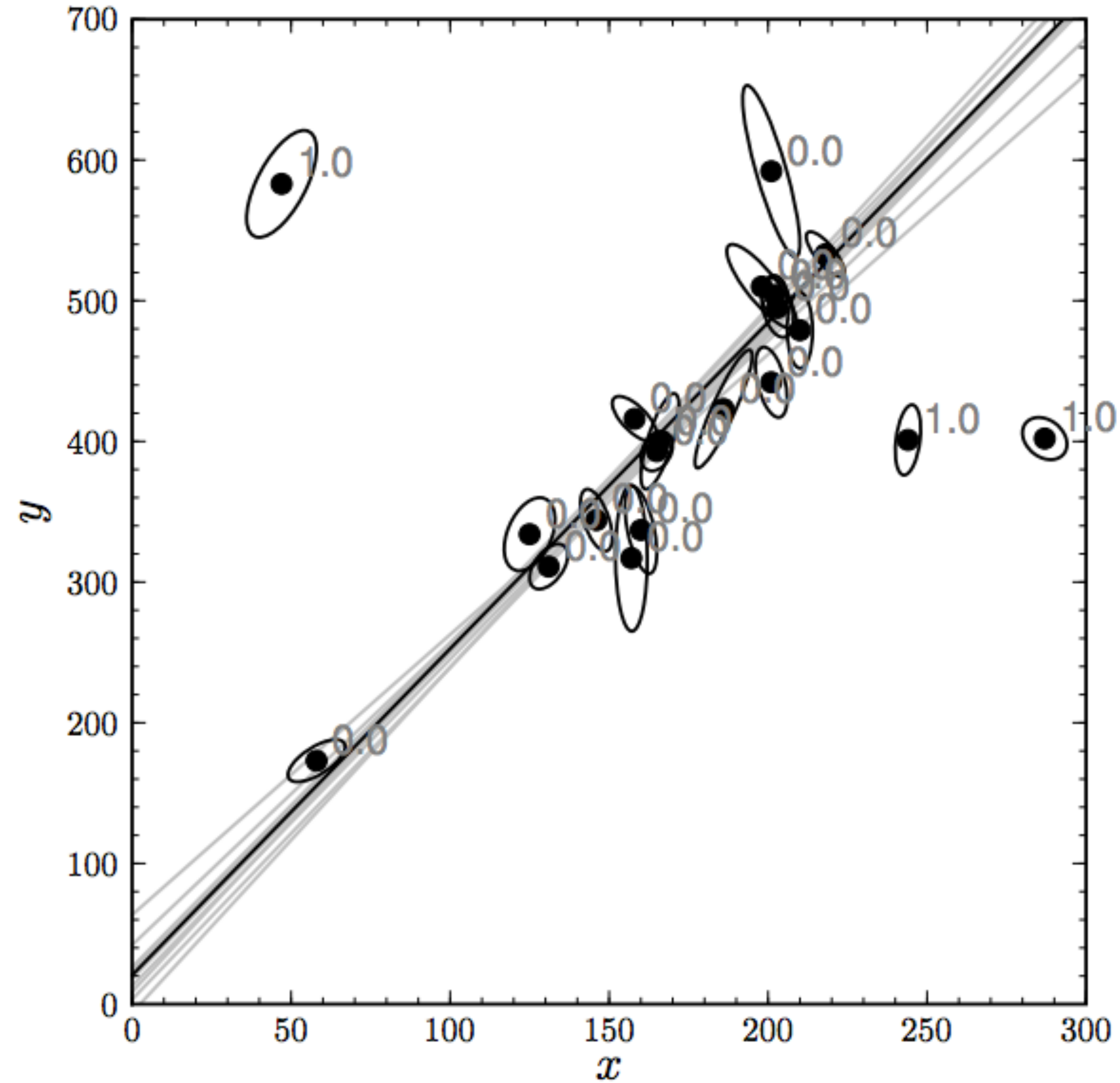
Statistics and Inference in Astrophysics

Zhu & Menard (2013)



Hogg, Bovy, & Lang (2010)



Hogg, Bovy, & Lang (2010)

(you would *not* believe the amount of code and statistics that went into creating the mock data in the previous figures…)

Object of this mini-course

- *No:* to teach you how to fit a straight line to data…
- *No:* to go through an exhaustive list of all possible applications of statistical methods and machine-learning techniques in astrophysics
- *Yes:* to introduce you to important basic concepts and principles to use when applying statistics in astro
- *Yes*: to teach you some basic tools that you will use over and over again

Overview of topics

- Probability theory
- Common probability distributions
- Maximum-likelihood fitting, penalized likelihood
- Bayesian inference, frequentist analysis
- Sampling from probability distributions; Markov Chain Monte Carlo
- Bootstrap and jackknife
- goodness-of-fit; cross-validation
- Robust statistics, outliers
- Monte Carlo techniques
- Some tools from machine learning: KDE, K-means, Gaussian-mixtures, PCA, Gaussian processes, support vector machines

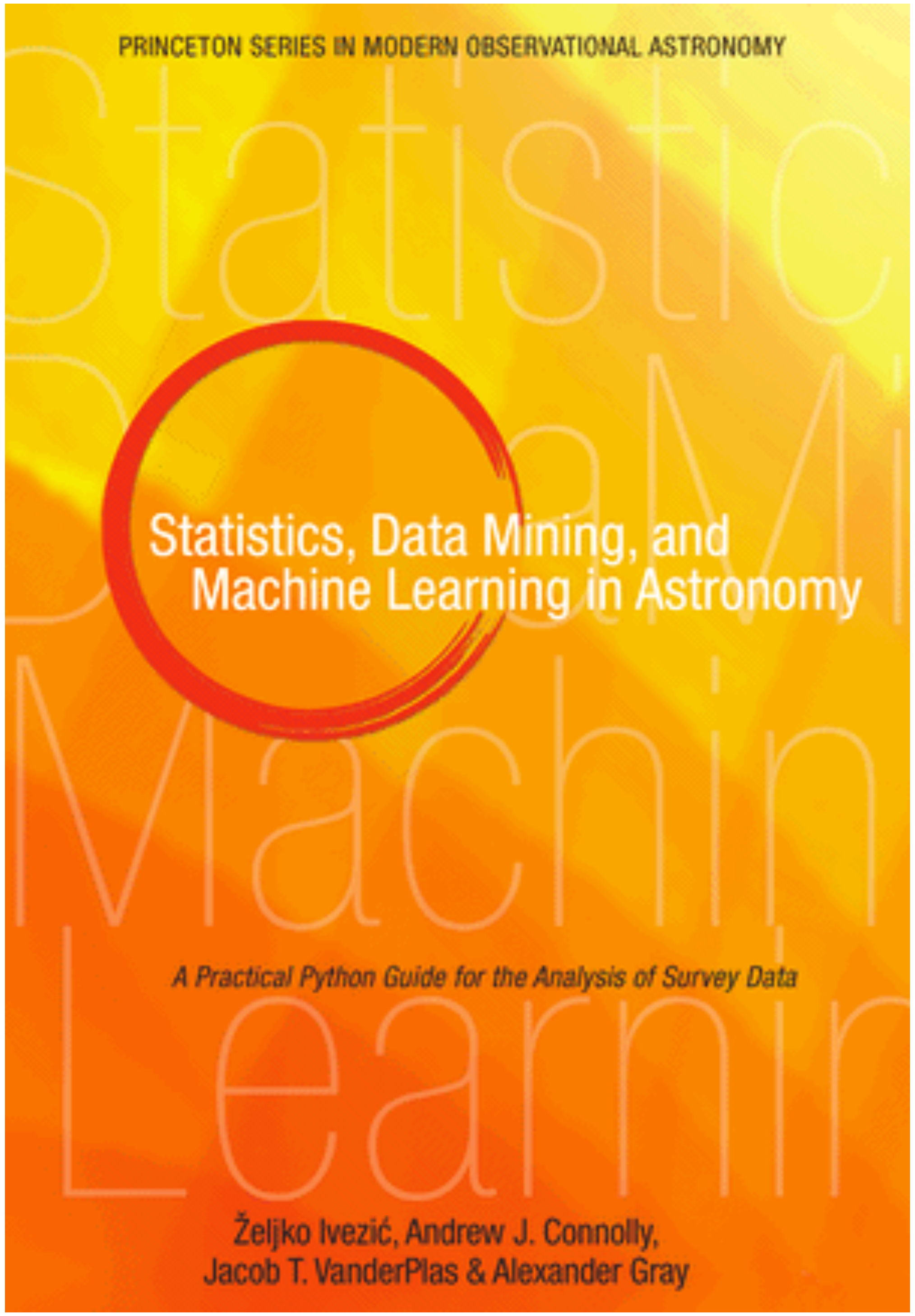
Classes

- Course taught in 5 sessions:
  - Feb 25 (today, 1hr20min): Introduction, generalities, probability calculus, common distributions
  - Mar 10 (2 weeks from now, 2hr): likelihood, maximum likelihood, Bayesian inference, frequentist analysis
  - Mar 17 (2hr): Sampling, MCMC, marginalization etc., non-parametric methods (bootstrap, jackknife)
  - Mar 24 (1hr20min): Goodness-of-fit, cross-validation, Monte-Carlo techniques, outliers, robust statistics
  - April 7 (1hr20min, note weeklong gap): Topics in machine learning: regression, classification, classical ML algorithms

Marking

- For those taking this for credit (1/3 course)
- 2 assignments, one after third lecture, one after last lecture
- Format TBD, but likely computer exercises (Python?)
- Those of you not taking the course for credit are welcome to do the assignments, but no guarantee that they will be marked

Textbook



Statistics, Data Mining, and Machine Learning in Astronomy

Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alexander Gray

Princeton University Press

2014

What are we doing here?

- (big) data → astrophysical knowledge
- Data analysis: three steps
  - Data exploration / model building: what model are we fitting to the data
  - Inference: how do we fit the model to the data?
  - Model validation: Is the model a good fit? How should we adjust the model? What new data should we get to test the model further

Statistics, inference, machine learning, …

- Statistics: broad definition "Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data" (Wikipedia)
- Statistics: narrow definition: set of mathematical tools and theorems about distribution of random variables
- Inference: "Inference may be defined as the non-logical, but rational means, through observation of patterns of facts, to indirectly see new meanings and contexts for understanding." … "Statistical inference uses mathematics to draw conclusions in the presence of uncertainty." (Wikipedia)
- Machine learning: "Field of study that gives computers the ability to learn without being explicitly programmed" (Arthur Samuel). Much inapplicable in typical astro setting and typically useless for *inference*, but useful set of tools for model building and validation.

Probability calculus

Probability calculus

- At its core, probability theory has a firm mathematical basis that is worth keeping in mind
- Rules of probability can be rigorously derived; not much use in most applications, but basics important to keep in mind
- Important to not sin against: a) units, b) laws of conditional probability

Probability calculus: basics

- Can have probability of discrete variables and continuous variables; follow slightly different rules, so good to use different symbols to keep track
- $P(a_i)$: probability of discrete set of outcomes $\{a_i\}$
- $p(a)$: probability of continuous set of outcomes a
- Probabilities normally normalized such that total probability of *anything* happening is 1
- $\Sigma_i P(a_i) = 1$
- $\int da\, p(a) = 1$

Probability calculus: basics

- $\Sigma_i P(a_i) = 1$
- $\int da\, p(a) = 1$
- $P(a_i)$ and $p(a)$ have units, do they have the same units?

Probability calculus: basics

- $\Sigma_i\, P(a_i) = 1$
- $\int da\; p(a) = 1$
- $P(a_i)$ dimensionless (number)
- Units of $p(a)$: $1/a$; required to make $\int da\; p(a) = 1$
- Can $P(a_i)$ be smaller than 1?
- Can $P(a_i)$ ever be larger than 1?
- Can $p(a)$ ever be larger than 1?

Probability calculus: basics

- Often very useful to keep in mind that p(a) has units of 1/a
- Example: transformations of p(a)
- Suppose b = f(a); what is p(b)?

Probability calculus: basics

- Often very useful to keep in mind that p(a) has units of 1/a
- Example: transformations of p(a)
- Suppose b = f(a); what is p(b)?
- Get p(b) from conservation of dimensionless probability
- Probability (capital P!) of a in [a,a+da]: $p_a(a)$ da
- Probability (capital P!) of b=f(a) in [b,b+db]: $p_b(b)$ db
- Dimensionless Probability should be the same:
  $p_a(a)\ da = p_b(b)\ db$
- $p_b(b) = p_a(a)\ |da/db| = p_a(f^{-1}[b]) \times |df^{-1}[b]/db|$

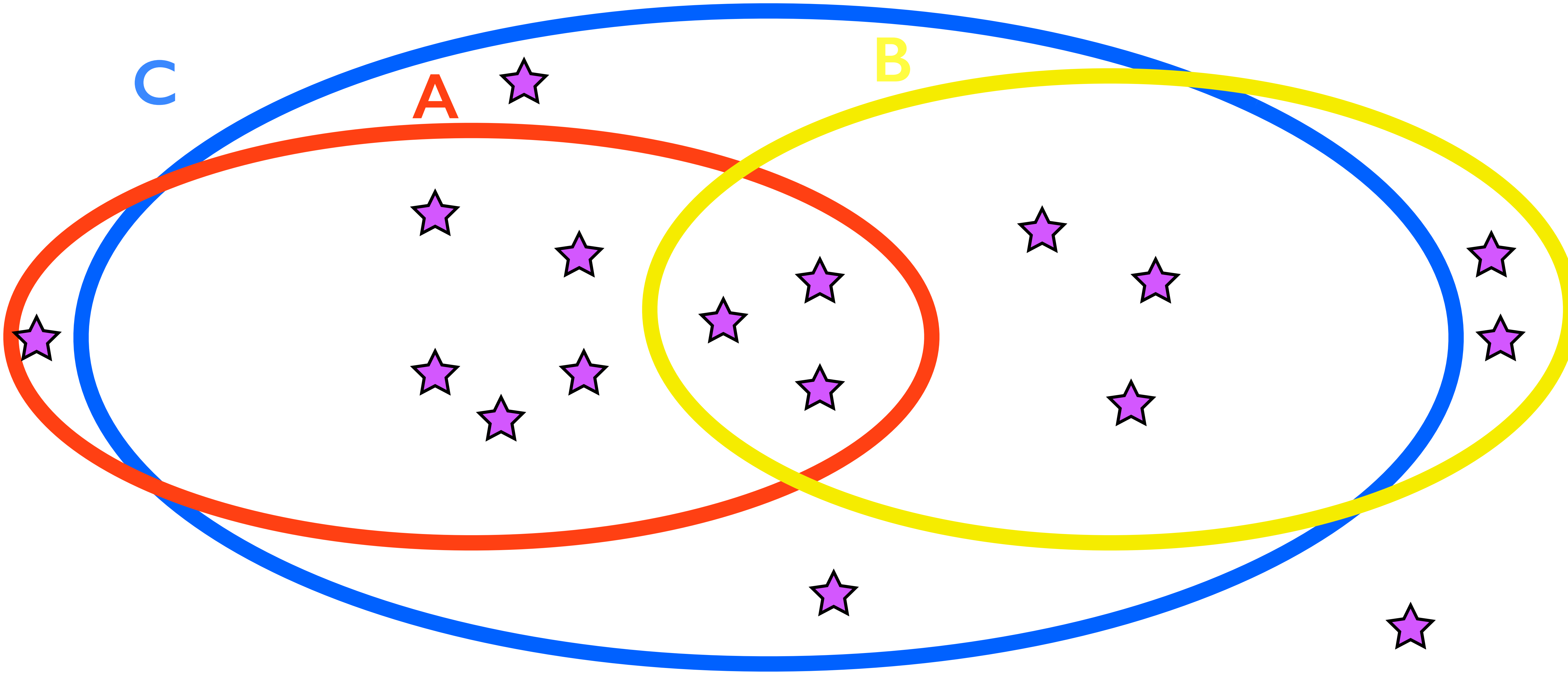
Probability calculus: basics

- $p_b(b) = p_a(a)\ |da/db| = p_a(f^{-1}[b]) \times |df^{-1}[b]/db|$
- Does this make sense in terms of units?

Probability calculus: rules of conditional probability

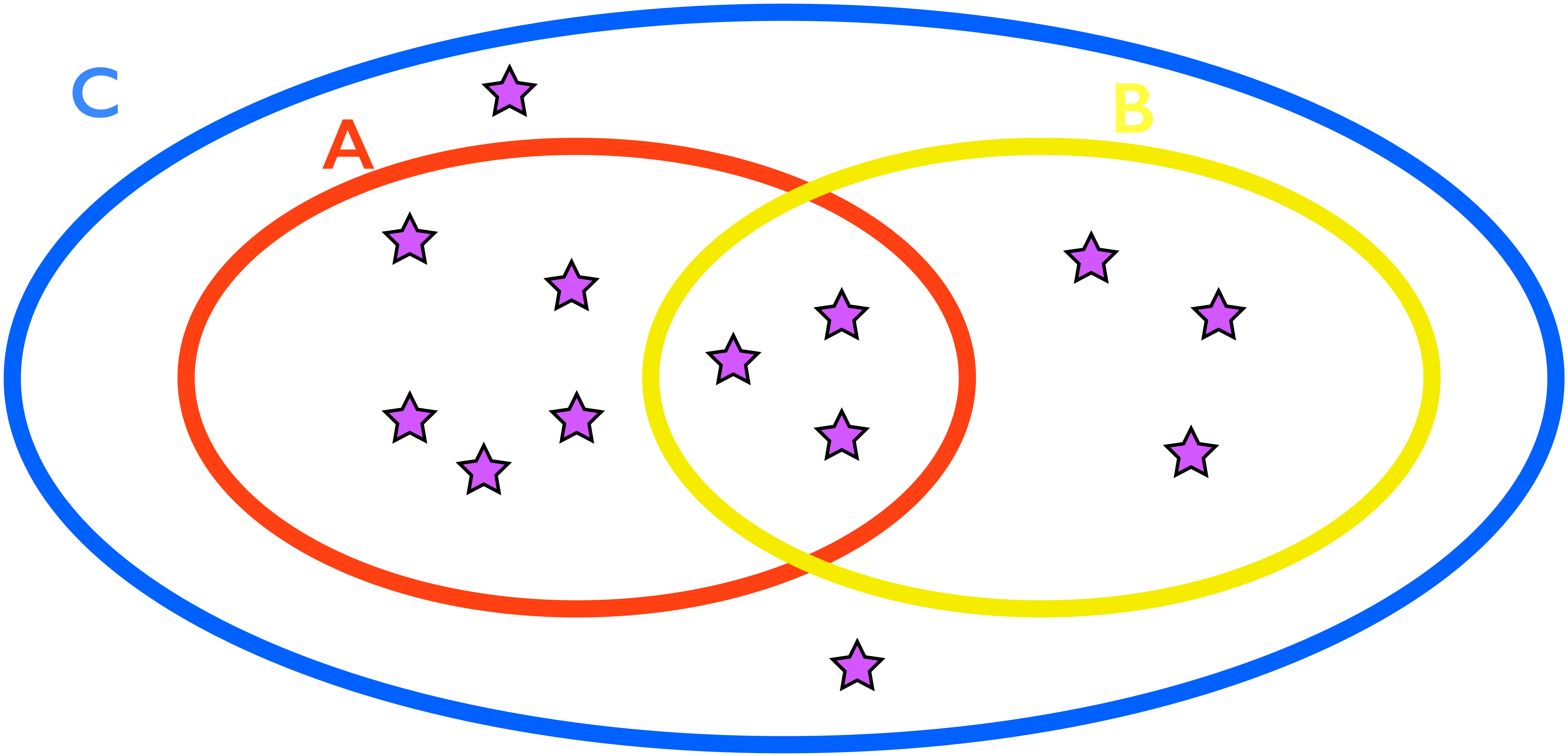
Rules of probability

- $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$
- $11/13 = 8/13 + 6/13 - 3/13$



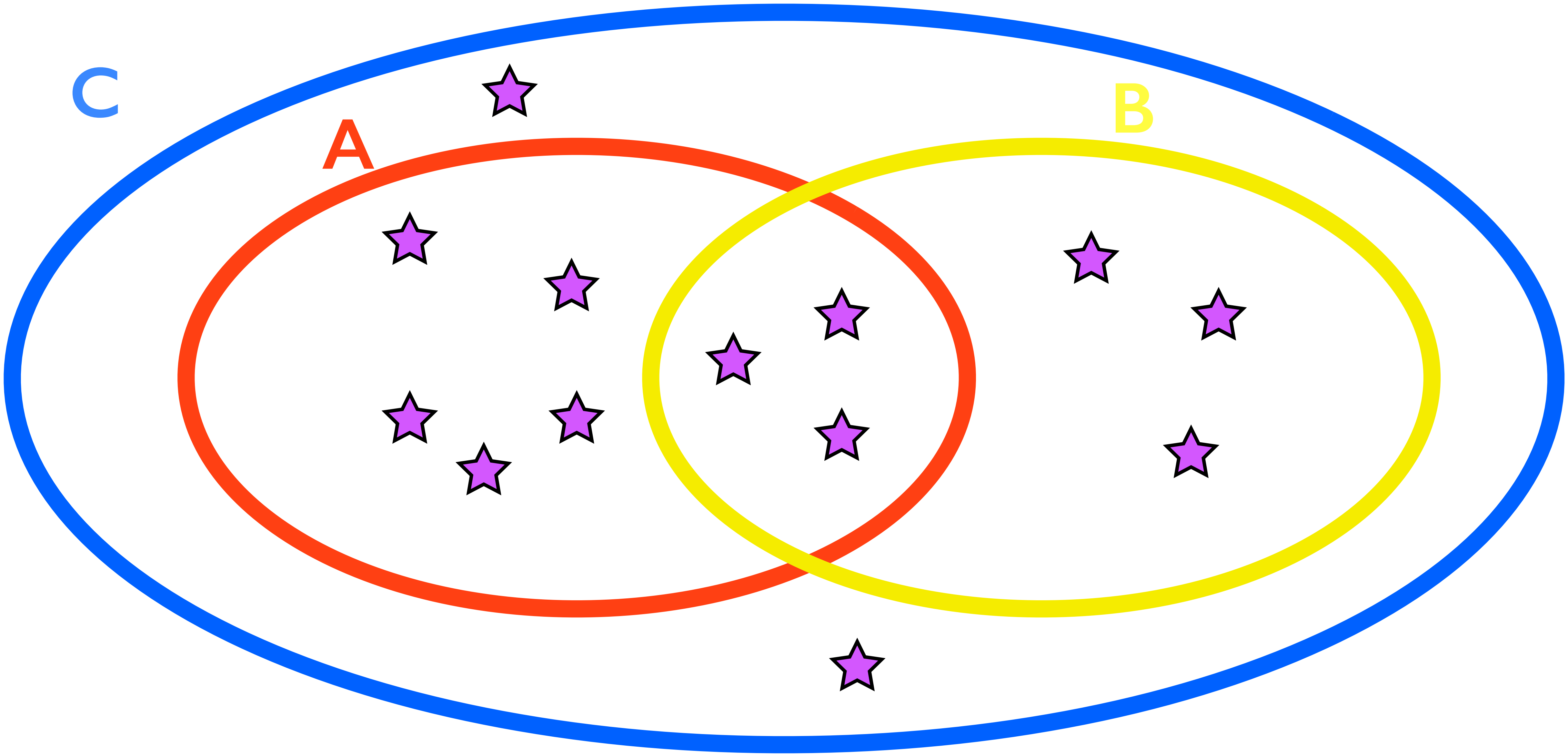
Rules of probability

- $P(A \cap B | C) = P(A | B \cap C) \times P(B | C)$
- $3 / 13 = 3 / 6 \times 6 / 13$



Rules of probability

- $P(A|B \cap C) = P(B|A \cap C) \times P(A|C) / P(B|C)$
- $3 / 6 = 3 / 8 \times 8 / 13 / (6 / 13)$



Rules of conditional probability

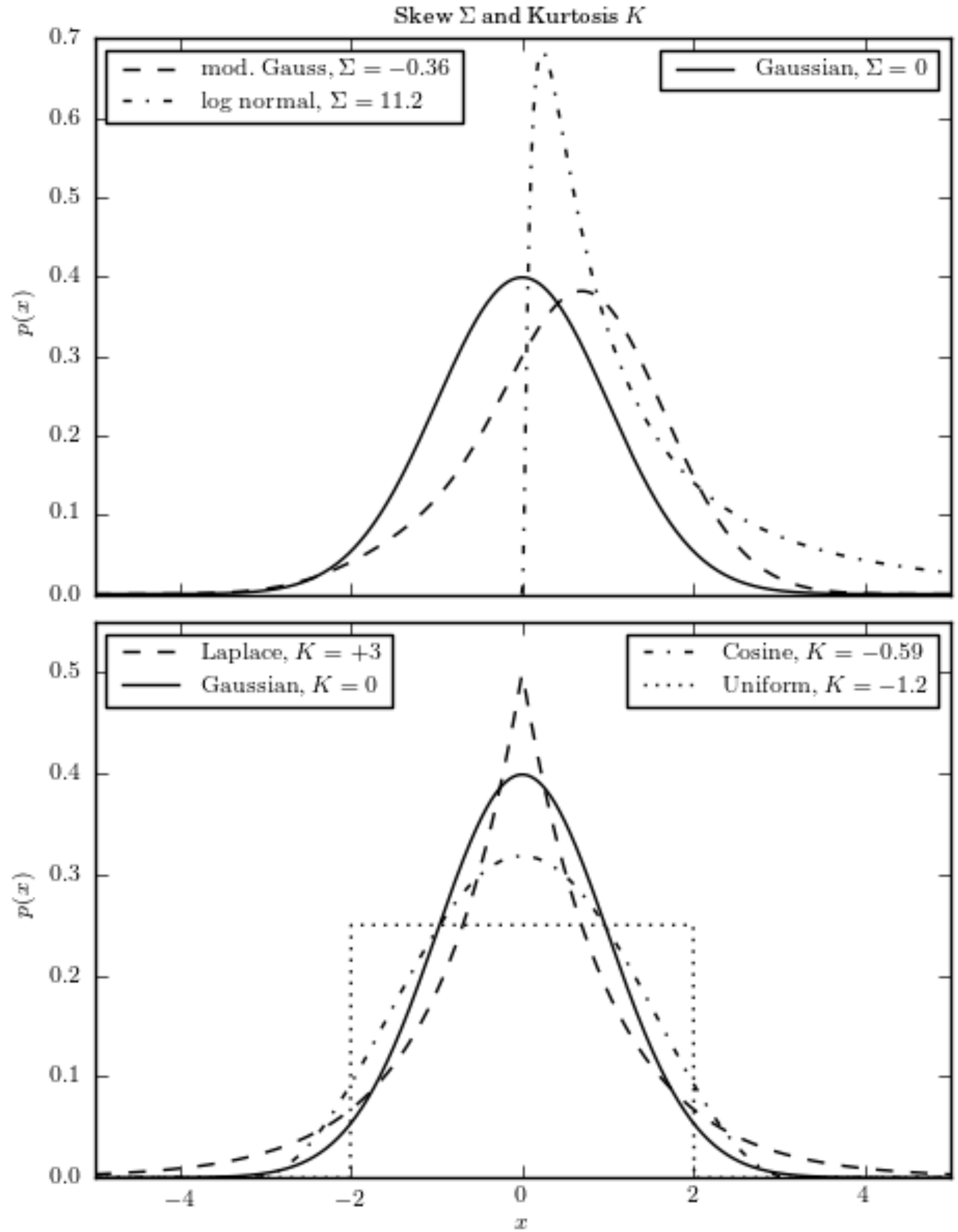
- $P(A \text{ or } B|C) = P(A|C)+P(B|C)-P(A,B|C)$
- $p(A,B|C) = p(A|B,C) \times p(B|C)$
- $p(A|B,C) = p(B|A,C) \times p(A|C)/p(B|C)$
- Do these make sense in terms of units?

Rules of conditional probability

- **Don't** do things like
- $P(A,B|C) = P(A|B,C) \times P(B|A,C)$
- $P(A|B,C) = P(A|B) \times P(A|C)$
- Might seem obvious now, but easy to get fooled when have (A,B,C,D,E,F,…) and complex conditional relations between them
- Published literature has examples of these mistakes……

Characterizing probability distributions

- $p(x)$
- Mean, expectation value: $\mu = \int dx\, p(x)x$
- Variance: $V = \int dx\, p(x)(x-\mu)^2$
- Standard deviation: $\sigma = \sqrt{V}$
- Skewness: $S = \int dx\, p(x)([x-\mu]/\sigma)^3$
- Kurtosis: $K = \int dx\, p(x)([x-\mu]/\sigma)^4$
- Excess kurtosis: $K-3$; often default!



Ivezic et al. (2014)

Characterizing probability distributions

- p(x)
- Mode: argmax p(x), dp/dx = 0
- Quantiles: x_q such that $\int_{-\infty}^{x_q} \mathrm{d}x \, p(x) = q$
- Median: $x_{0.5}$
- Mean, median, mode: think about how these transform under y=f(x)
- Median, quantiles often best way to characterize p(x), but issues when multiple peaks etc.

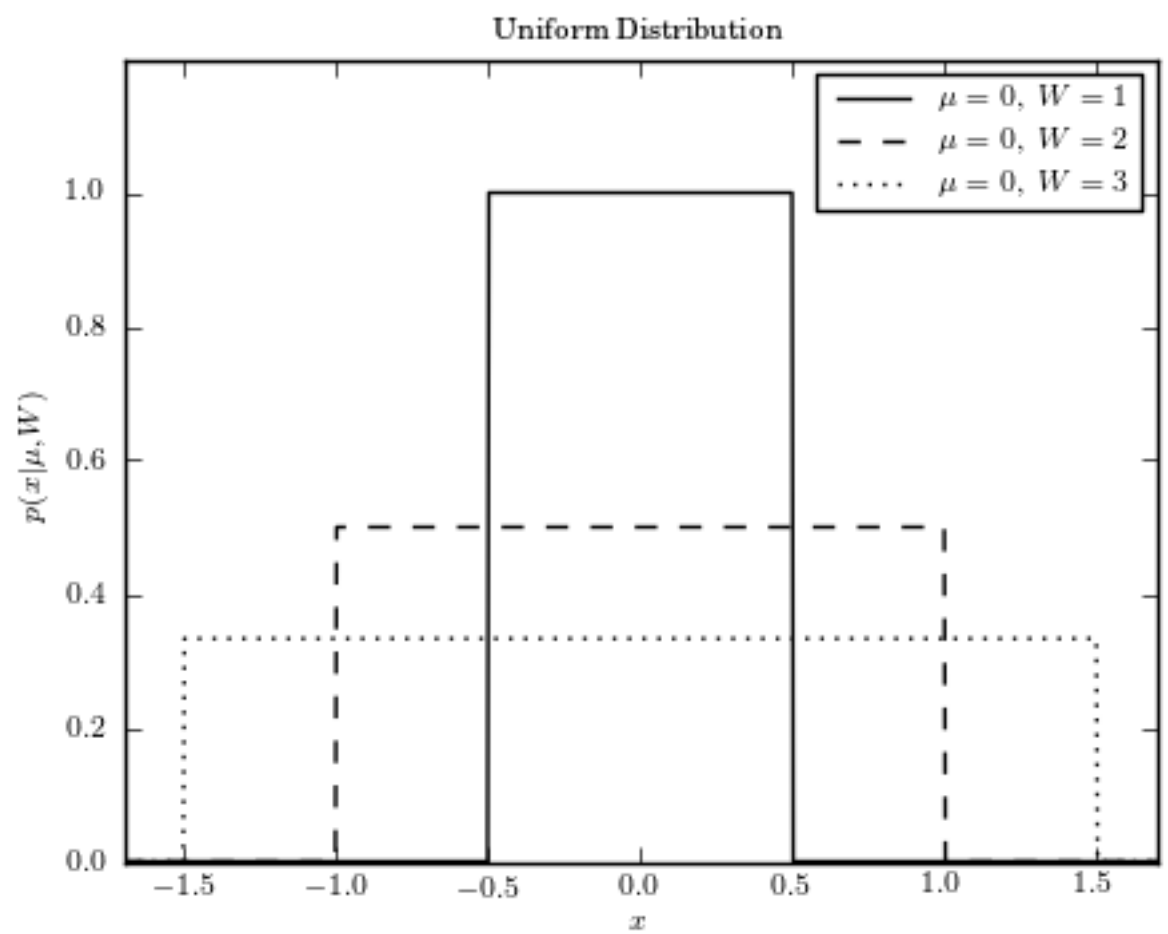
Characterizing probability distributions

- Cumulative distribution function (CDF): CDF(y) = P(x <= y) [note capital P!]
- CDF(x) = $\int^{x} \mathrm{d}y \, p(y)$
- CDF(∞)=?
- P(a < x <= b)= CDF(b)-CDF(a)

Common probability distributions and their properties

Uniform distribution

- Simplest one!
- p(x) = constant for a < x < b
- constant = 1/ (b-a)
- Mean?
- Variance= $(a-b)^2/12$
- Random numbers from uniform distribution basis of all (pseudo)-randomness on a computer

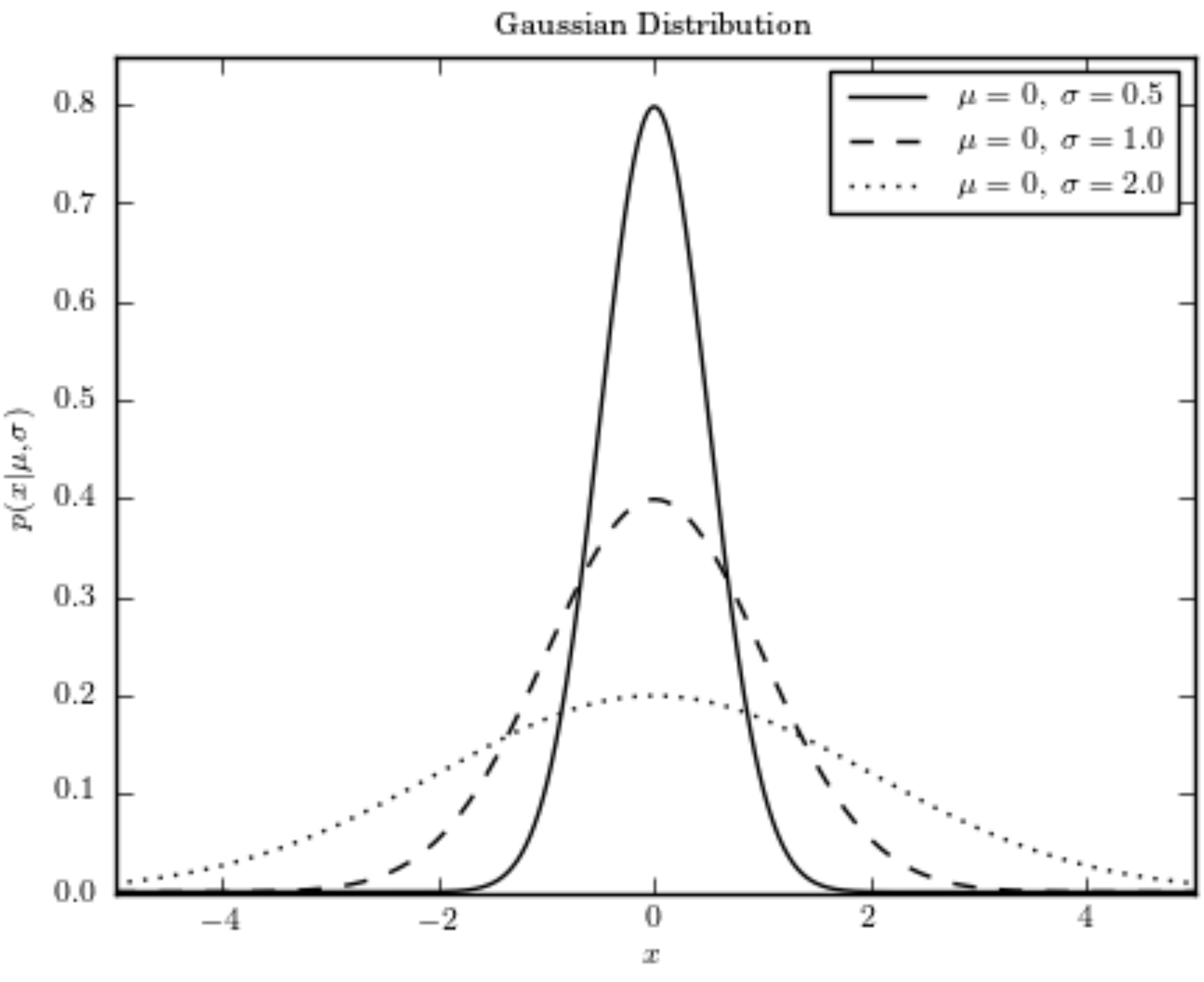


Ivezic et al. (2014)

Gaussian distribution

- Most common one?
- Form: $$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- Mean μ, standard deviation σ
- Skewness=0, excess kurtosis=0
- Worth remembering the pre-factor!

Gaussian distribution



Ivezic et al. (2014)

Gaussian distribution

- Convolution of a Gaussian with another Gaussian is again a Gaussian
- Use $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Then

$$\int \mathrm{d}y\, \mathcal{N}(y|\mu_1,\sigma_1^2)\, \mathcal{N}(x-y|\mu_2,\sigma_2^2) = \mathcal{N}(x|\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

Gaussian distribution

- Cumulative distribution function:

$$\mathrm{CDF}(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left[\frac{x-\mu}{\sqrt{2}\sigma}\right]\right)$$

- P(σ < x-μ <= σ) = CDF(μ+σ)-CDF(μ+σ) = erf(1/√2) = 0.68268949213708585

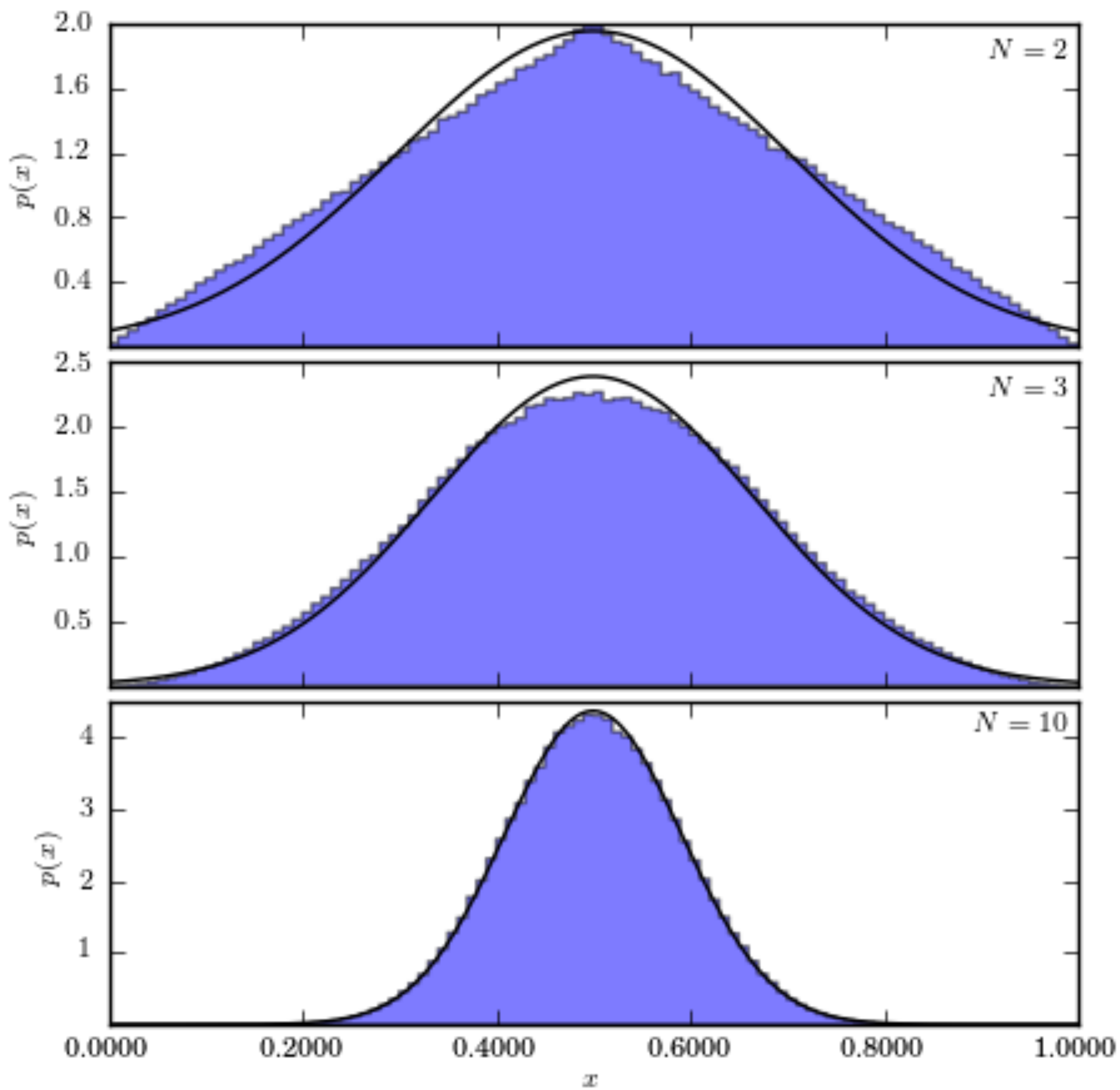
- P(2σ < x-μ <= 2σ) = CDF(μ+2σ)-CDF(μ+2σ) = erf(2/√2) = 0.95449973610364158

- P(3σ < x-μ <= 3σ) = CDF(μ+3σ)-CDF(μ+3σ) = erf(3/√2) = 0.99730020393673979

Gaussian distribution

- Central limit theorem:
- Have: arbitrary probability distribution p(x) with finite mean μ and variance σ^2
- Draw *N* samples x_i from p(x), what is the distribution of $m = 1/N \; \Sigma_i \, x_i$?
- $p(m) = N(\mu, \sigma^2/N)$ as $N \rightarrow \infty$

Central limit theorem



Ivezic et al. (2014)

Bernoulli distribution

- Discrete Probability for X which has Probability p for being 1 and 1-p for being 0 (flipping coin)
- $P(X=1) = p$, $P(X=0) = 1-p$
- Mean? $p \times 1 + (1-p) \times 0 = p$
- Variance? $p \times (1-p)^2 + (1-p) \times (0-p)^2 = (1-p) \times (p[1-p]+p^2) = p(1-p)$

Binomial distribution

- Probability distribution for outcome of repeated Bernoulli trials: $Y = \Sigma_i X_i$
- Number k of successes in N trials, each with Probability p of success
- $p(k|p,N) = N!/(k!(N-k)!)\ p^k(1-p)^{N-k}$
- Mean Np because of Bernoulli (easier than working out the combinatorics!)
- Variance $Np(1-p)$ because of Bernoulli

Poisson distribution

- Say you want the statistical distribution of the number of photons that comes from a source in 1 s, and you expect λ photons
- Divide time interval into small dt such that each interval has either 0 or 1 photons, say dt=0.01s
- Probability of seeing a photon in dt is then p=dt/(1 s) λ
- Total number of photons is then Binomial trial with N =(1 s)/dt and p=dt/(1 s) λ
- When dt → 0, $N \to \infty$ and $p \to 0$, but Np always λ

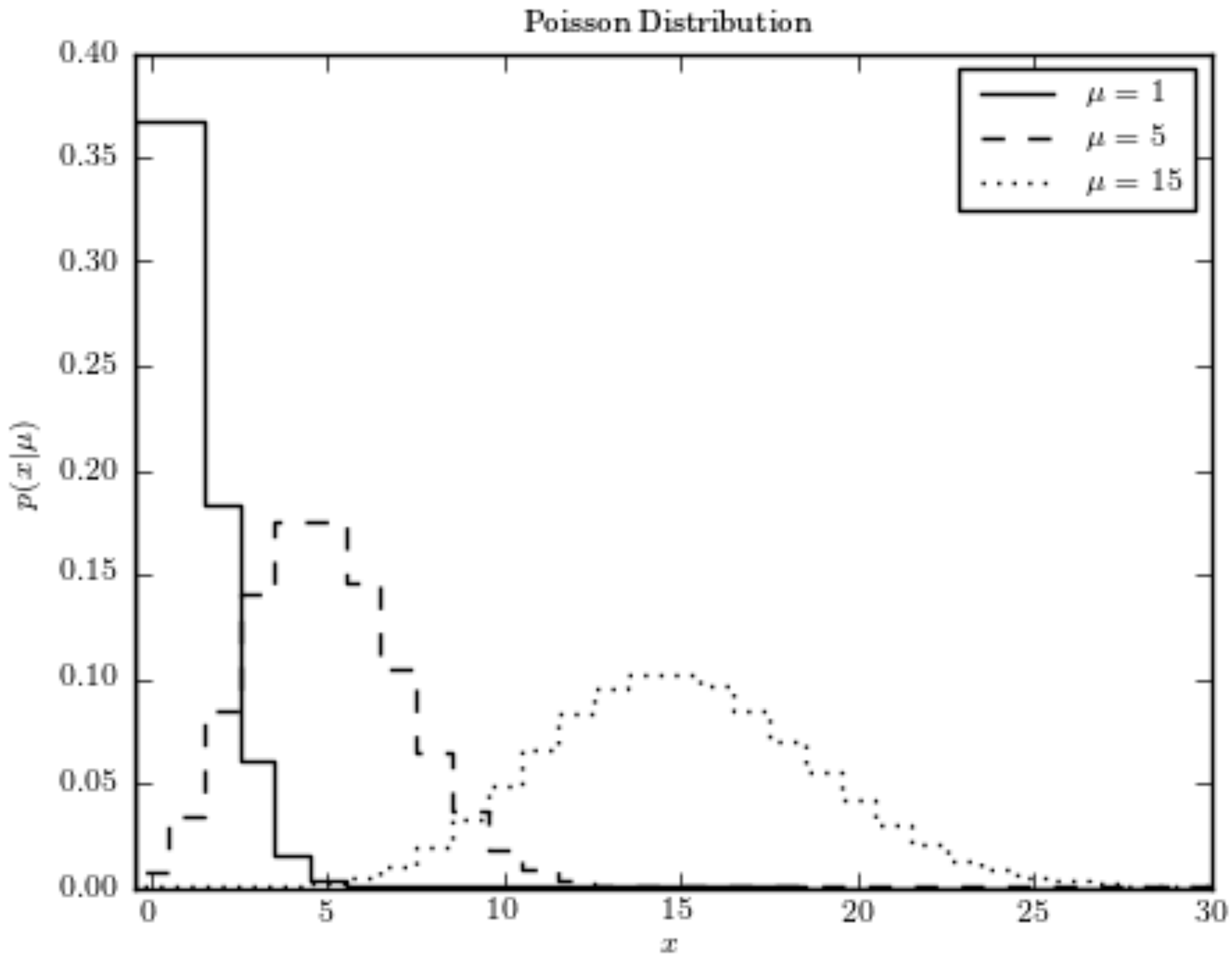
Poisson distribution

- Take Binomial distribution with $N \to \infty$ and $Np = \lambda$
- $p(k|p,N) = N!/(k!(N-k)!)\, p^k (1-p)^{N-k} =\sim N!/(k!(N-k)!)\, \lambda^k/N^k\, (1-\lambda/N)^{N-k} = \lambda^k e^{-\lambda}/k!$
- $p(k|\lambda) = \lambda^k e^{-\lambda}/k!$
- Mean = ?
- Variance = ?
- Sum of Poisson distributed variables is again Poisson distributed
- For large λ, Poisson well described by Gaussian with mean λ and variance λ (central limit theorem applied to λ chunks with mean 1)

Poisson distribution

- Take Binomial distribution with $N \to \infty$ and $Np = \lambda$
- $p(k|p,N) = N!/(k!(N-k)!)\ p^k(1-p)^{N-k} =\sim N!/(k!(N-k)!)\ \lambda^k/N^k\ (1-\lambda/N)^{N-k} = \lambda^k e^{-\lambda}/k!$
- $p(k|\lambda) = \lambda^k e^{-\lambda}/k!$
- Mean = λ
- Variance = λ
- Sum of Poisson distributed variables is again Poisson distributed
- For large λ, Poisson well described by Gaussian with mean λ and variance λ (central limit theorem applied to λ chunks with mean 1)

Poisson distribution

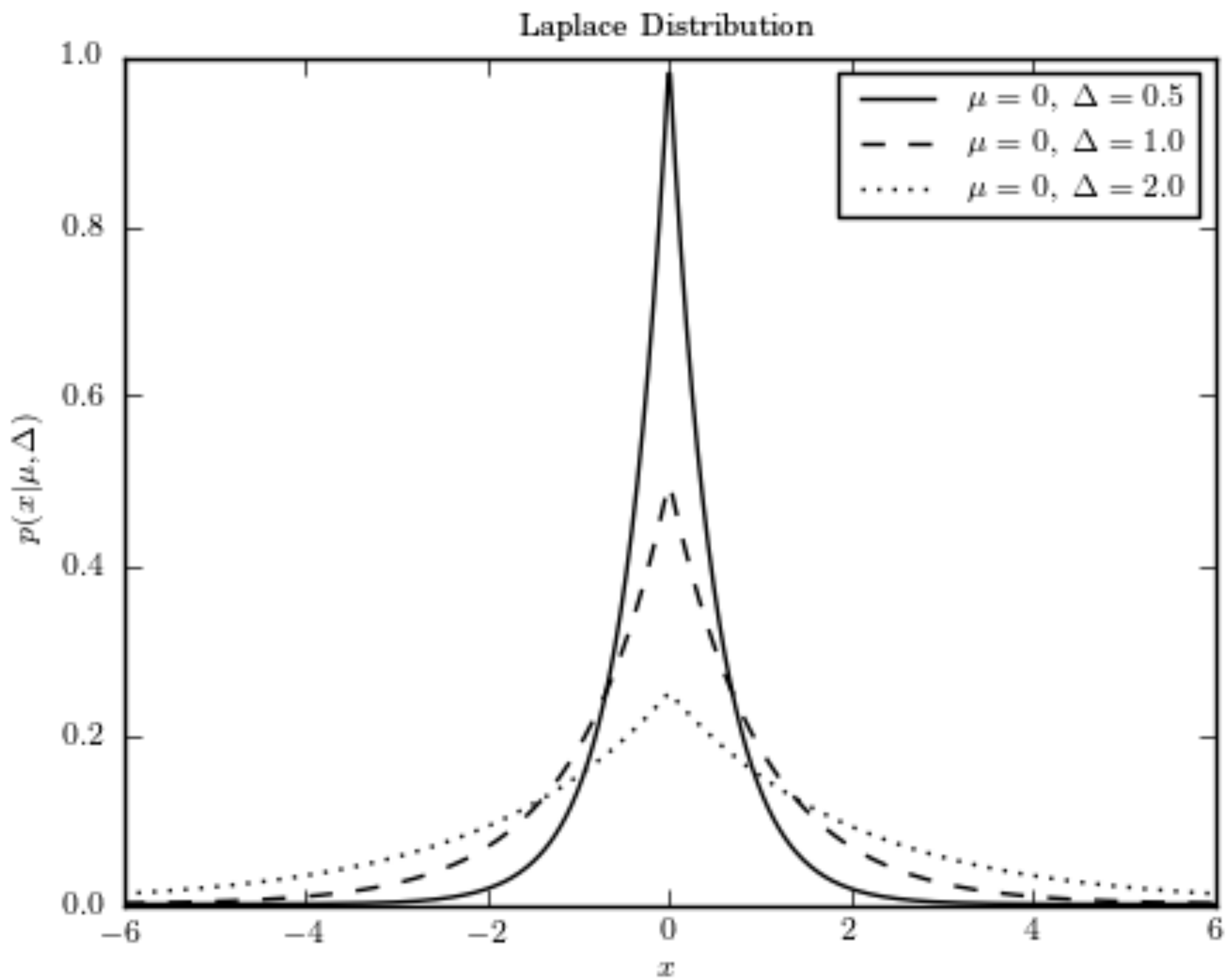


Ivezic et al. (2014)

Exponential distribution

- $p(x|\lambda) = \lambda\, e^{-\lambda x}$, $\lambda >= 0$
- Waiting time x between events in Poisson process with rate λ
- Mean: λ^{-1}
- Variance: λ^{-2}
- Laplace distribution: same, but for positive and negative x: $p(x|\lambda) = \lambda/2\, e^{-\lambda|x-\mu|}$
- Weibull: generalization to where the rate is a power of time
- Rayleigh: $p(x|\lambda) \sim x \exp(-x^2/2/\sigma^2)$, $\lambda >= 0$

Laplace distribution

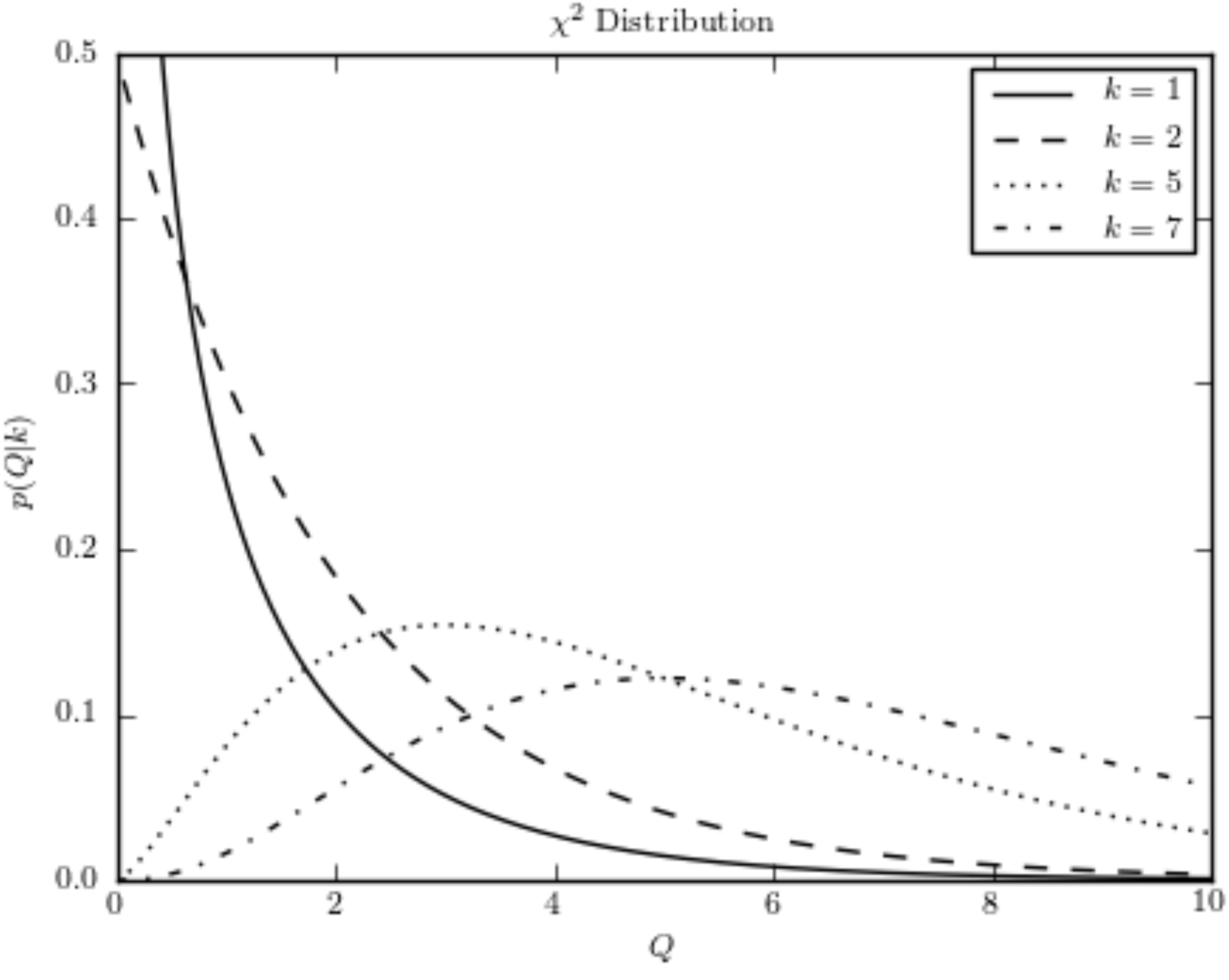


Ivezic et al. (2014)

Chi-squared distribution

- Distribution of sum of squares of *k* independent standard normal variables (those from *N*(x|0,1))
- Form: $$p(x|k) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)}\, x^{k/2-1}\, e^{-x/2}$$
- Mean: *k*
- Variance: 2*k*
- Basis for *chi-squared-per-degree-of-freedom* goodness-of-fit
- Central limit theorem: for $k \rightarrow \infty$, p(x|*k*) $\rightarrow$ *N*(x|*k*,2*k*)

Chi-squared distribution



Ivezic et al. (2014)

Higher-dimensional distributions

- Many fewer important ones!
- Multivariate Gaussian:

$$\mathcal{N}(\vec{x}|\vec{\mu}, \mathbf{V}) = \frac{1}{(2\pi)^{d/2}|\det\mathbf{V}|} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\mathbf{V}^{-1}(\vec{x}-\vec{\mu})\right)$$

- Wishart distribution

Sampling 1D probability distributions

- Random number generator returns (pseudo) random integer between 0 and 2^w-1
- How to generate random variables from other distributions
- Uniform: convert random integer to interval [0,1] or other
- Gaussian: u,v uniform on [0,1] → $\sqrt{(-2\ln[u])} \times \cos(2\pi v)$ is random draw from $N(x|0,1)$
- Most other distributions don't have simple ways to generate random samples

Sampling 1D probability distributions

- Inverse-cumulative-distribution method:
  $p(x) \rightarrow CDF(x) \rightarrow CDF^{-1}(x)$

- Sample *u* from uniform on [0,1], compute $v=CDF^{-1}(u)$, *v* is sample from p(x)

- Rejection sampling:
  if $p(x) < Mg(x)$ for all x and $M > 1$, and can sample easily from g(x)

- Sample *v* from g(x) and *u* from Uniform(0,1)
  if $u < p(v)/g(v)/M$: return *v*
  else: try again

- Rejection sampling can work in >1D, but typically difficult to find appropriate M and g(x)