Statistics and Inference in Astrophysics

Zhu & Menard (2013)

Overview of topics

- Fundamentals of probability theory
- Common probability distributions and random numbers
- Maximum-likelihood fitting, penalized likelihood
- Bayesian inference, frequentist analysis
- Sampling from probability distributions; Markov Chain Monte Carlo in theory and practice
- Bootstrap and jackknife
- goodness-of-fit; cross-validation, model selection criteria
- Combining statistical significance
- Robust statistics, outliers
- Probabilistic Graphical models
- Hierarchical modeling

Classes

- Course taught in 5 2hr sessions:
  - Feb 2 (today): Introduction, generalities, probability calculus, common distributions
  - Feb 9: likelihood, maximum likelihood, Bayesian inference, frequentist analysis
  - Feb 16: Monte Carlo sampling, MCMC, marginalization etc., non-parametric methods (bootstrap, jackknife)
  - Reading week: no class
  - Mar 2: Goodness-of-fit and model selection, cross-validation, statistical significance, intro to advanced topics
  - Mar 9: Student presentations

Marking

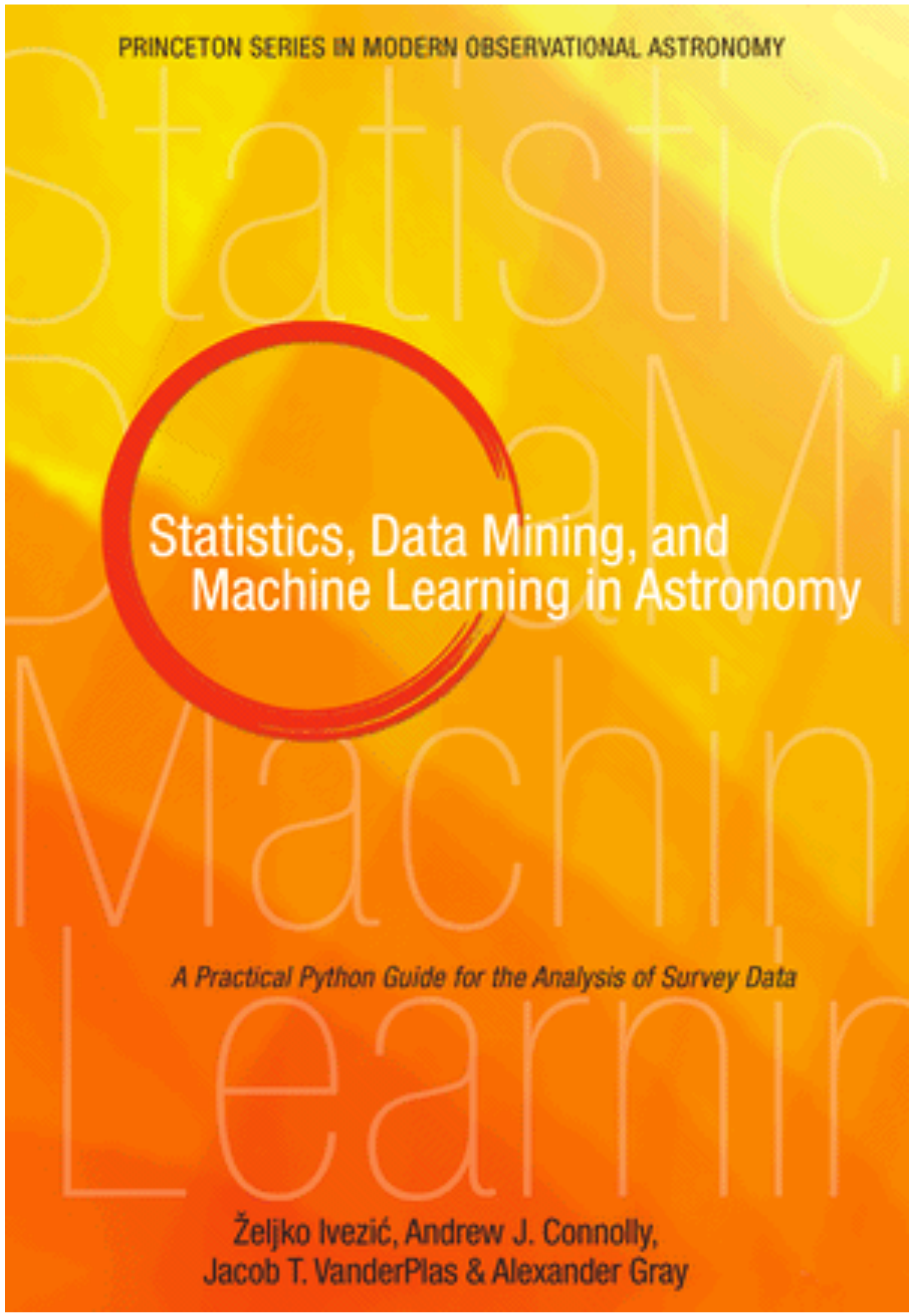
- For those taking this for credit (1/3 course)

- 1 assignment: hands-on work with your own MCMC implementation and working with state-of-the-art MCMC software

- Presentation during last lecture: present paper or topic in astronomy (or very relevant to astronomy) with statistical bend

- 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

$$\Sigma_i P(a_i) = 1$$

$$\int \mathrm{d}a \, p(a) = 1$$

Probability calculus: basics

$$\Sigma_i P(a_i) = 1$$

$$\int \mathrm{d}a \, p(a) = 1$$

- $P(a_i)$ and $p(a)$ have units, do they have the same units?

Probability calculus: basics

- $P(a_i)$: dimensionless (number)
- Units of $p(a)$: $1/a$; required to make $\int \mathrm{d}a\, p(a) = 1$
- Can $P(a_i)$ be smaller than 0?
- 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

- 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:

$$p_a(a)\mathrm{d}a = p_b(b)\mathrm{d}b$$

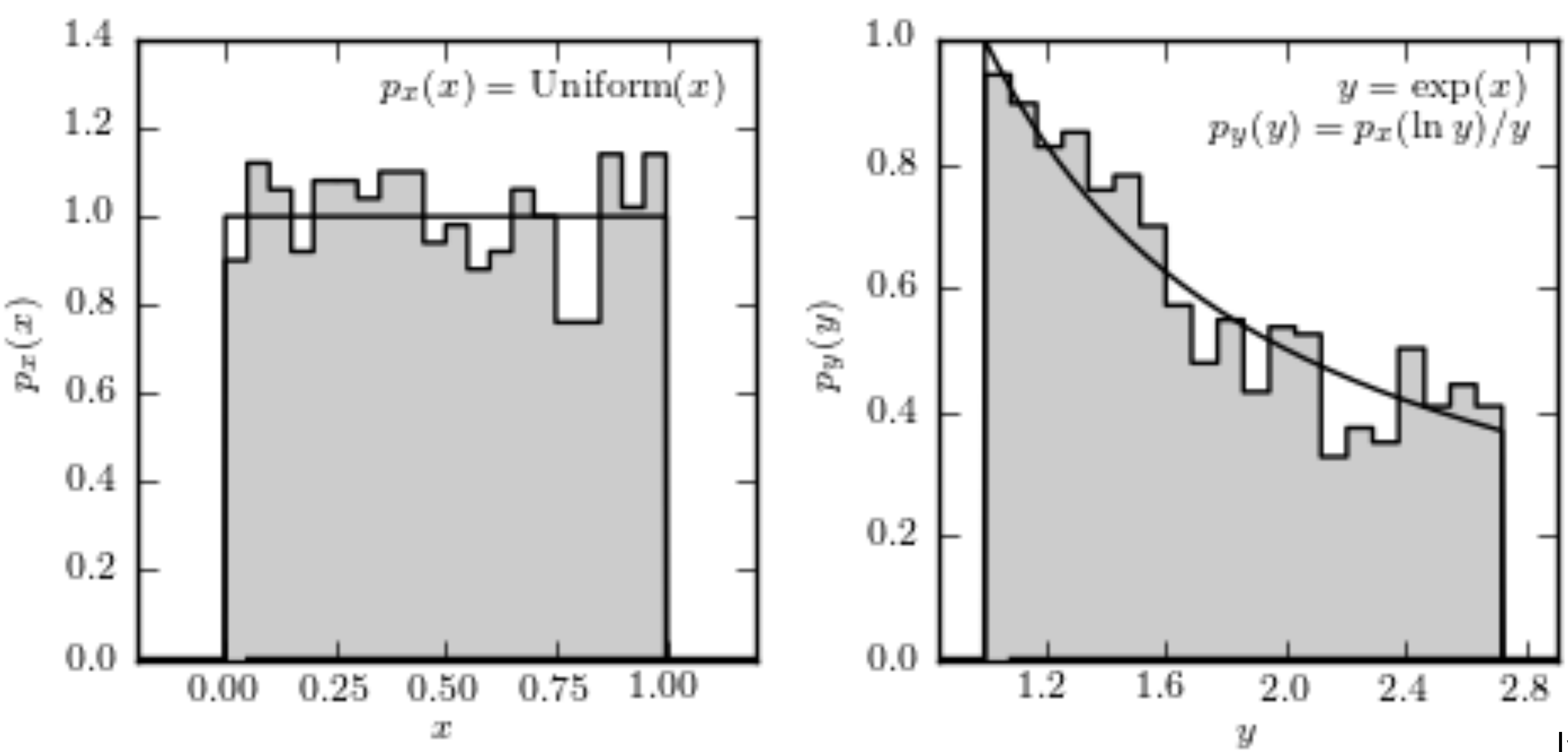
- or

$$p_b(b) = p_a(a)\left|\frac{\mathrm{d}a}{\mathrm{d}b}\right| \qquad p_b(b) = p_a(f^{-1}[b])\left|\frac{\mathrm{d}f^{-1}[b]}{\mathrm{d}b}\right|$$

Probability calculus: basics

$$p_b(b) = p_a(a)\left|\frac{\mathrm{d}a}{\mathrm{d}b}\right| \qquad p_b(b) = p_a(f^{-1}[b])\left|\frac{\mathrm{d}f^{-1}[b]}{\mathrm{d}b}\right|$$

- Does this make sense in terms of units?

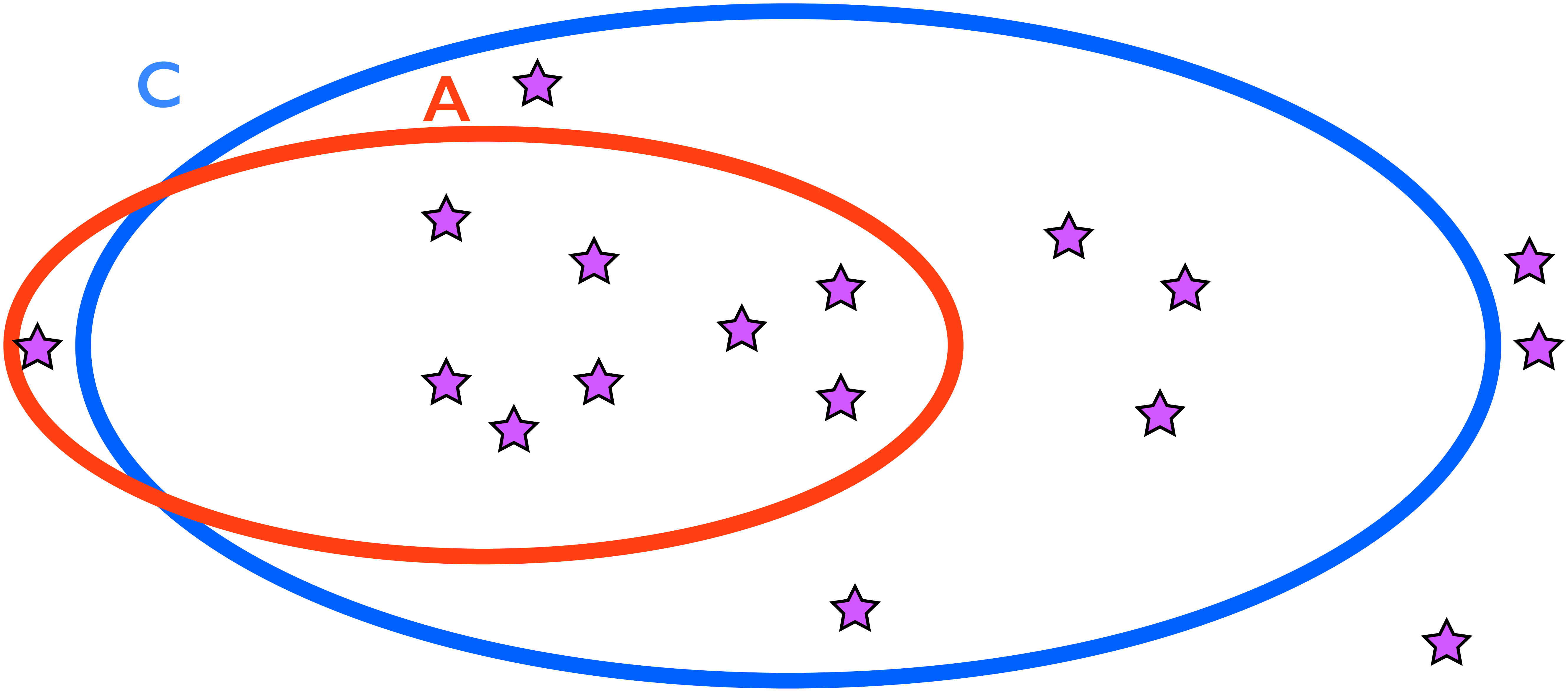


Ivezic et al. (2014)

Probability calculus: rules of conditional probability

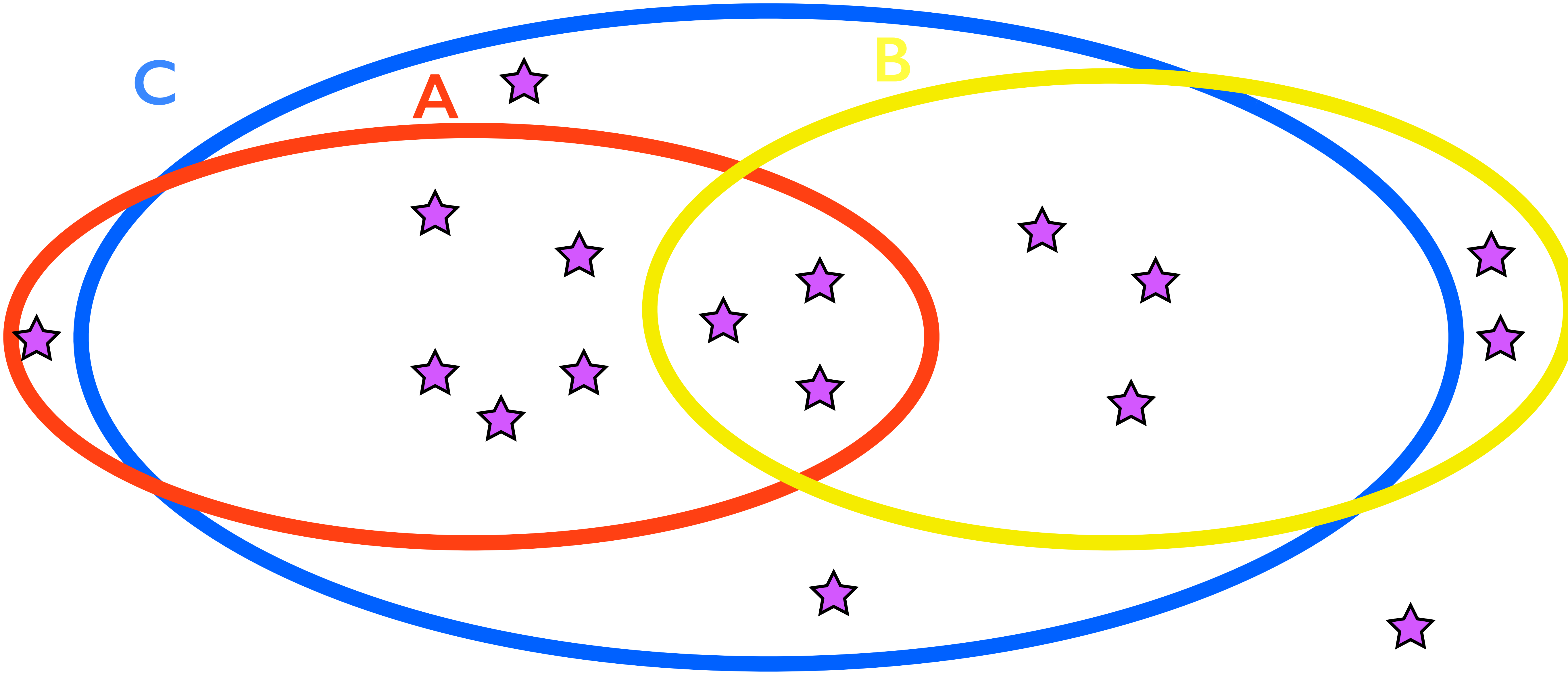
Conditional probability

- P(A|C) =
- = 8/13



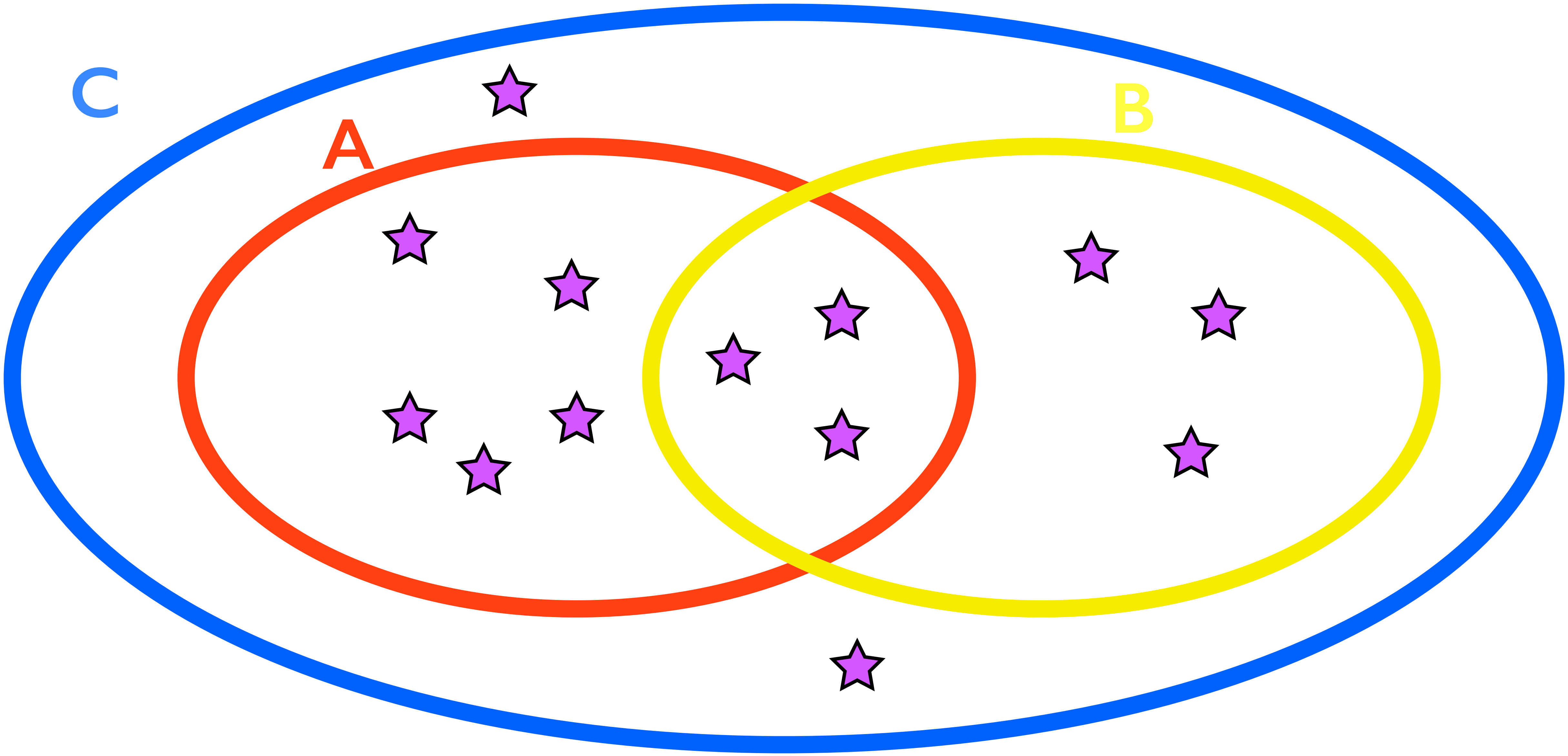
Rules of probability

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



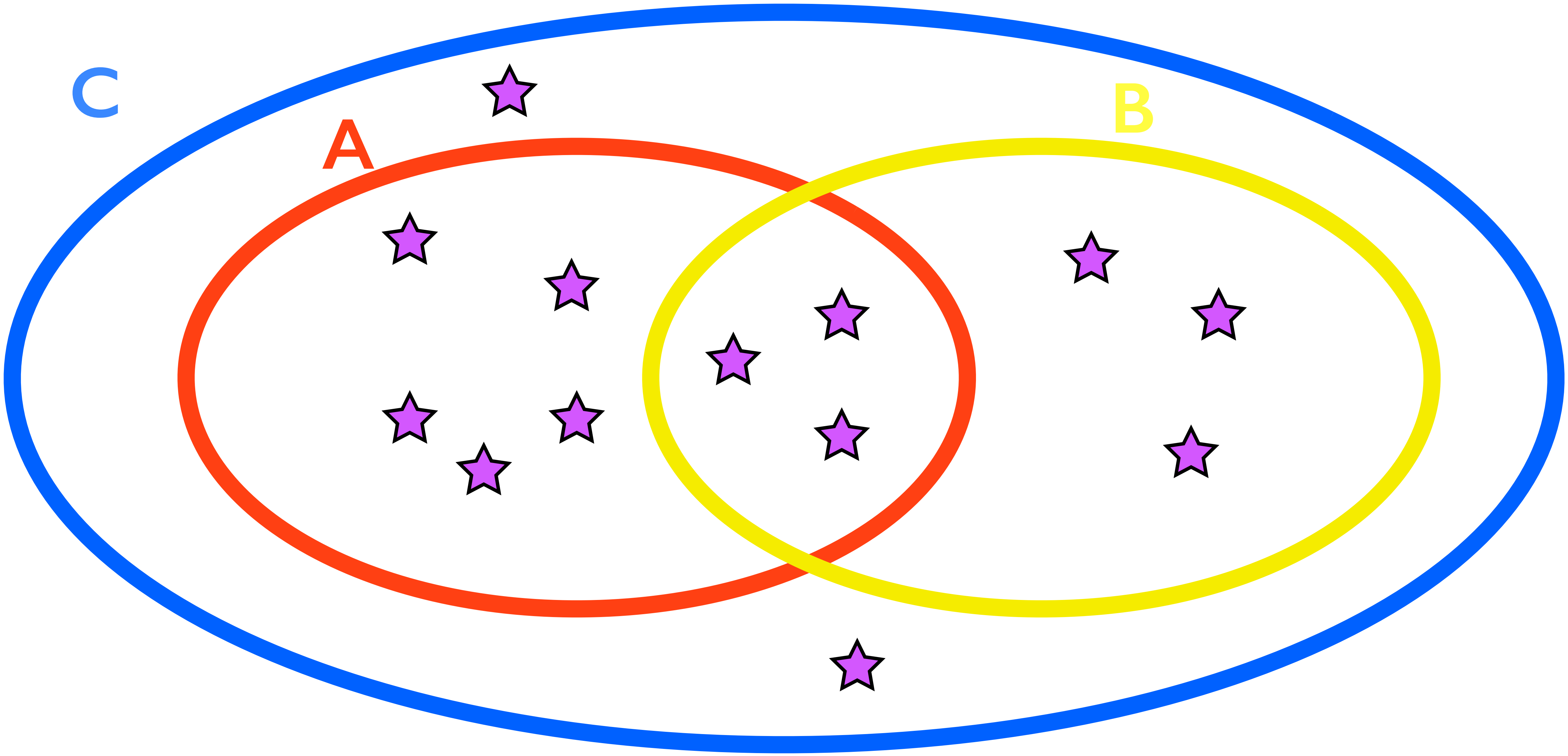
Rules of probability

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



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

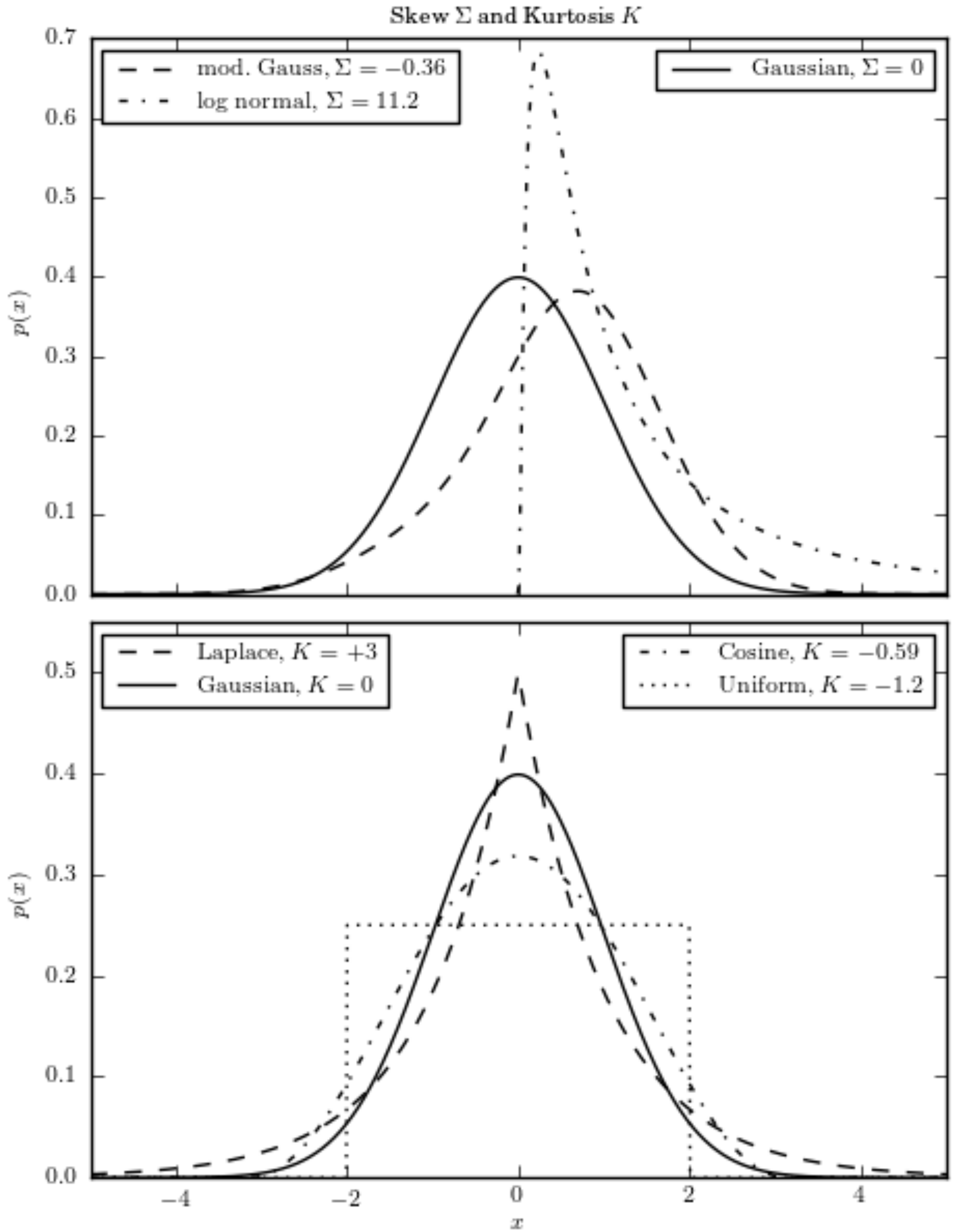
- $P(A \text{ or } B|C) = P(A|C)+P(B|C)-P(A,B|C)$
- $p(A,B|C) = p(A|B,C) \times p(B|C)$
- $p(A|B,C) = p(B|A,C) \times p(A|C)/p(B|C)$
- Do these make sense in terms of units?

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: $\mu = \int \mathrm{d}x\, p(x)\, x$
- Variance: $V = \int \mathrm{d}x\, p(x)\, (x-\mu)^2$
- Standard deviation: $\sigma = \sqrt{V}$
- Skewness: $S = \int \mathrm{d}x\, p(x)\, \frac{(x-\mu)^3}{\sigma^3}$
- Kurtosis: $K = \int \mathrm{d}x\, p(x)\, \frac{(x-\mu)^4}{\sigma^4}$
- 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.

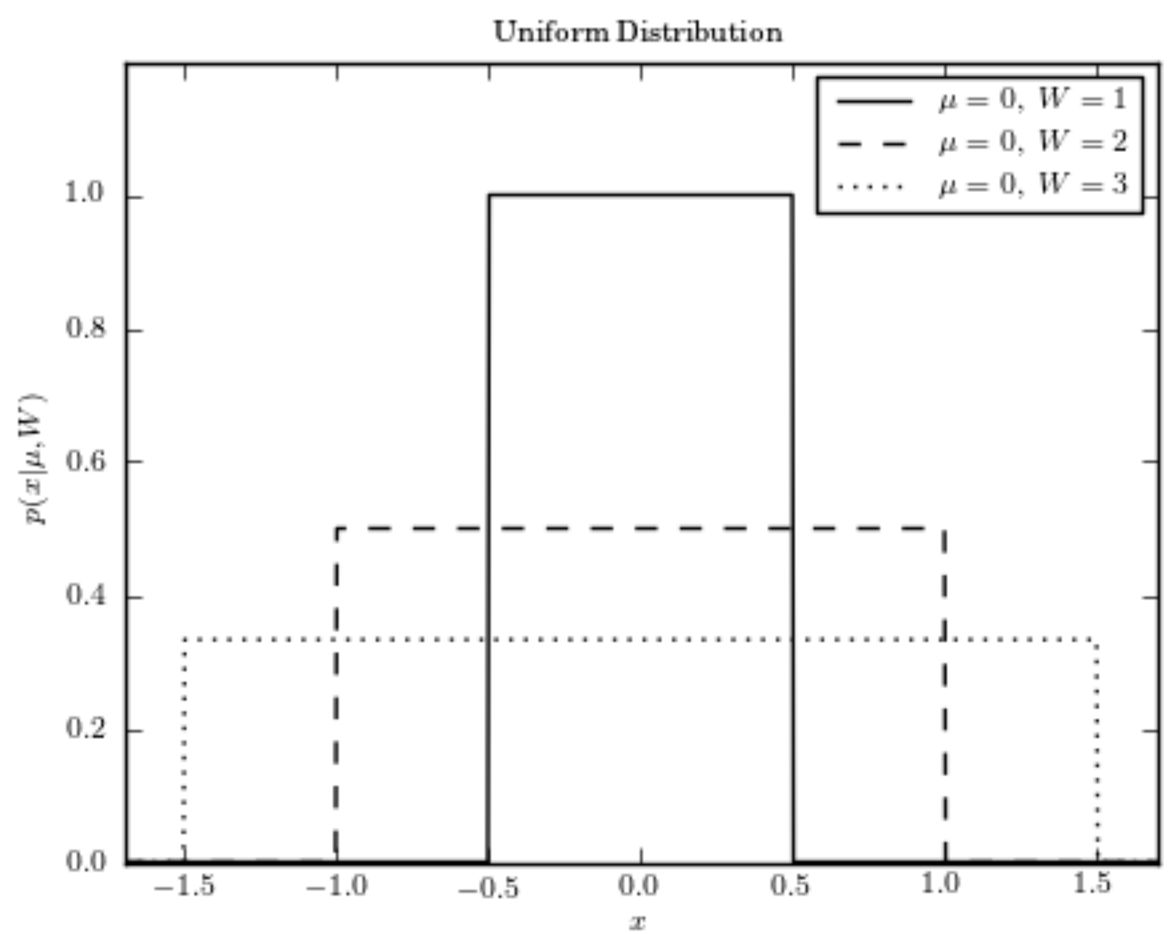
Characterizing probability distributions

- 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

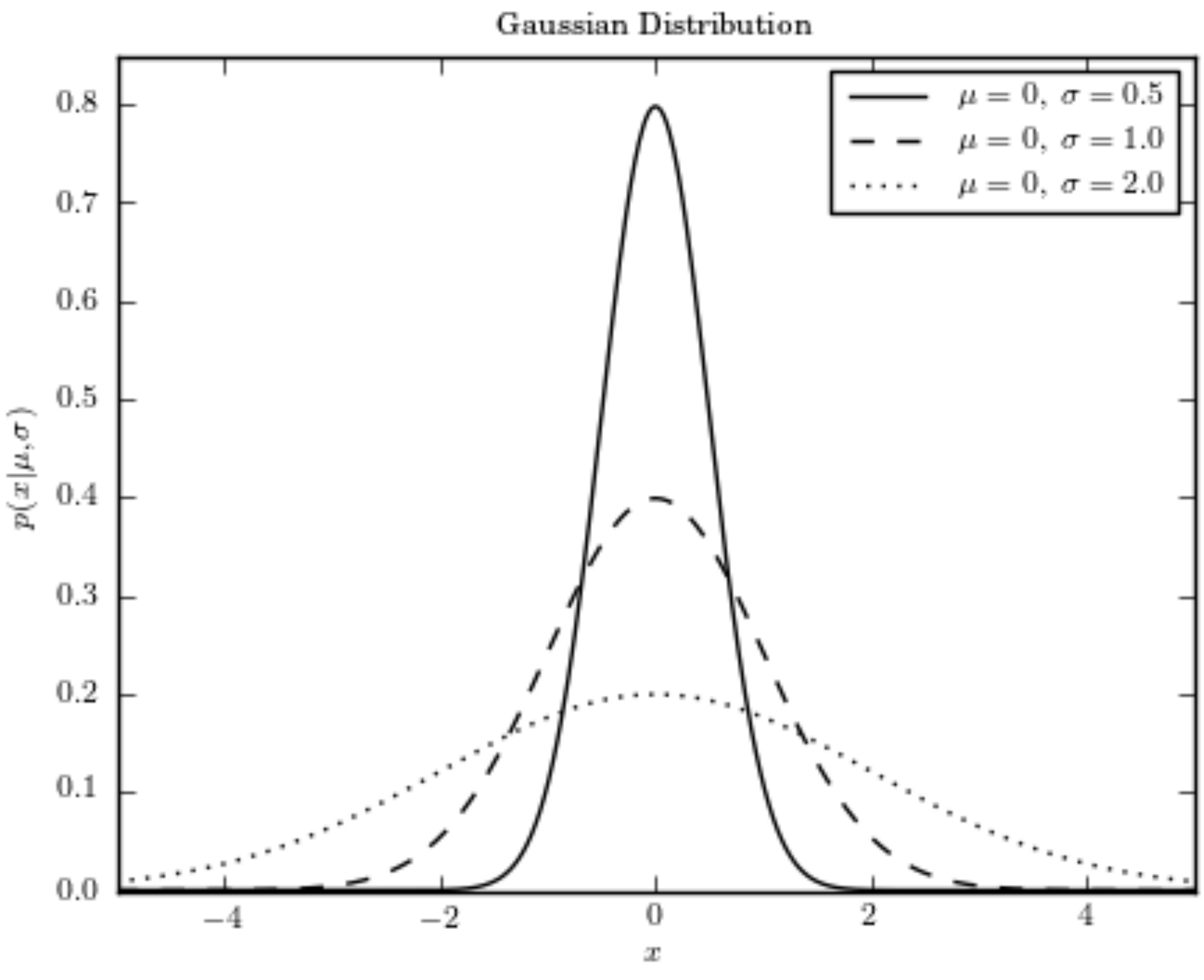


Ivezic et al. (2014)

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:

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

- $P(\sigma < x-\mu <= \sigma) = CDF(\mu+\sigma)-CDF(\mu+\sigma) = erf(1/\sqrt{2})$
  $= 0.6827$

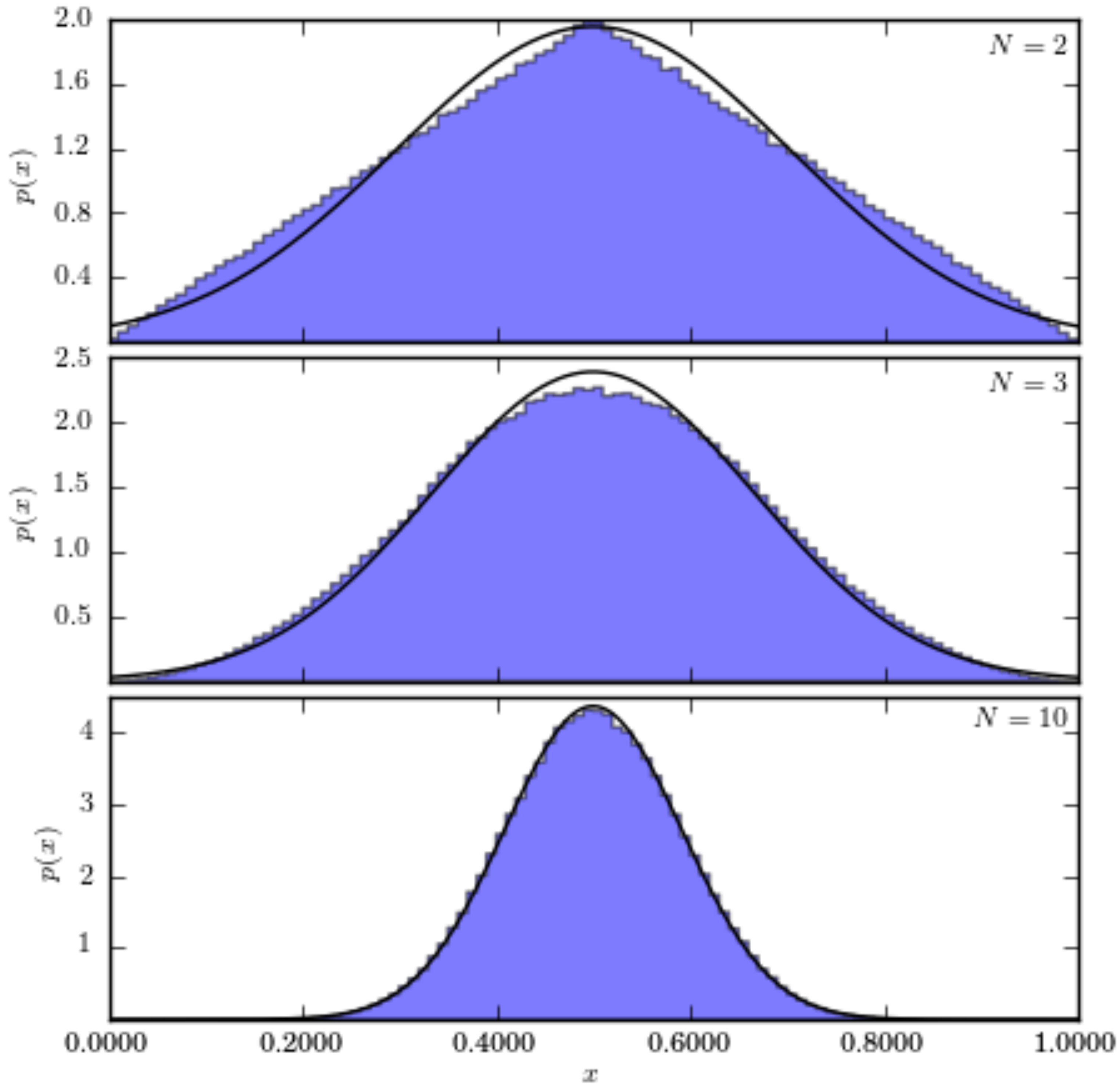
- $P(2\sigma < x-\mu <= 2\sigma) = CDF(\mu+2\sigma)-CDF(\mu+2\sigma) = erf(2/\sqrt{2})$
  $= 0.9545$

- $P(3\sigma < x-\mu <= 3\sigma) = CDF(\mu+3\sigma)-CDF(\mu+3\sigma) = erf(3/\sqrt{2})$
  $= 0.9973$

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 $m = \Sigma_i x_i/N$?
- $p(m) = N(\mu,\sigma^2/N)$ 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? $\mu = p \times 1 + (1-p) \times 0 = p$
- Variance? $$\begin{aligned} V &= p \times (1-p)^2 + (1-p) \times (0-p)^2 \\ &= p(1-p) \end{aligned}$$

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
- $$p(k|p,N) = \frac{N!}{k!\,(N-k)!}\,p^k(1-p)^{N-k}$$
- 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 $Np = \lambda$

$$p(k|p,N) = \frac{N!}{k!\,(N-k)!}\,p^k(1-p)^{N-k}$$
$$= \frac{N!}{k!\,(N-k)!}\,\left[\frac{\lambda}{N}\right]^k\left(1-\frac{\lambda}{N}\right)^{N-k}$$
$$\approx \frac{\lambda^k\,e^{-\lambda}}{k!}$$

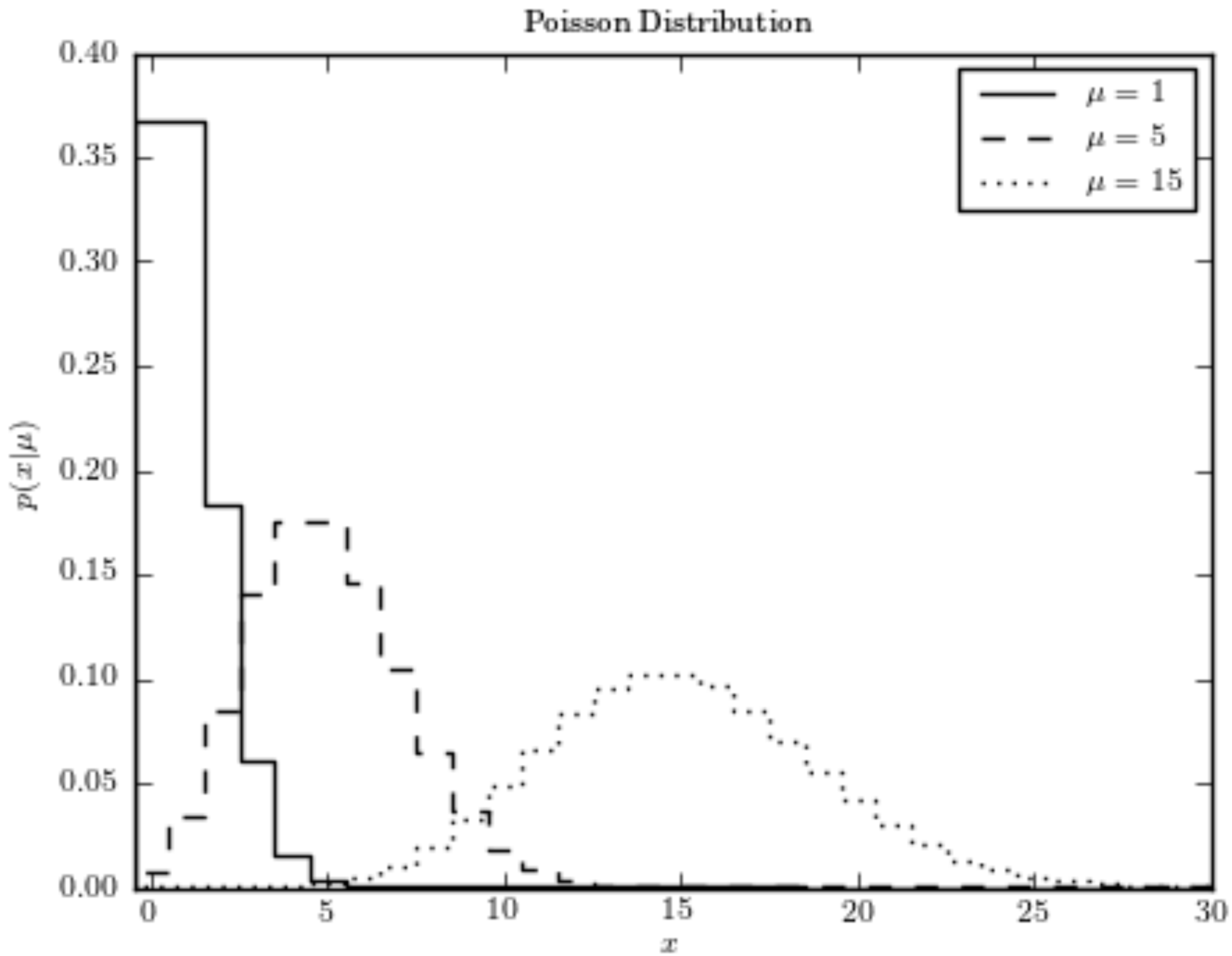
Poisson distribution

- Take Binomial distribution with $N \to \infty$ and $Np = \lambda$
- $$p(k|p,N) = \frac{\lambda^k \, e^{-\lambda}}{k!}$$
- 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 $Np = \lambda$
- $$p(k|p,N) = \frac{\lambda^k \, e^{-\lambda}}{k!}$$
- 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

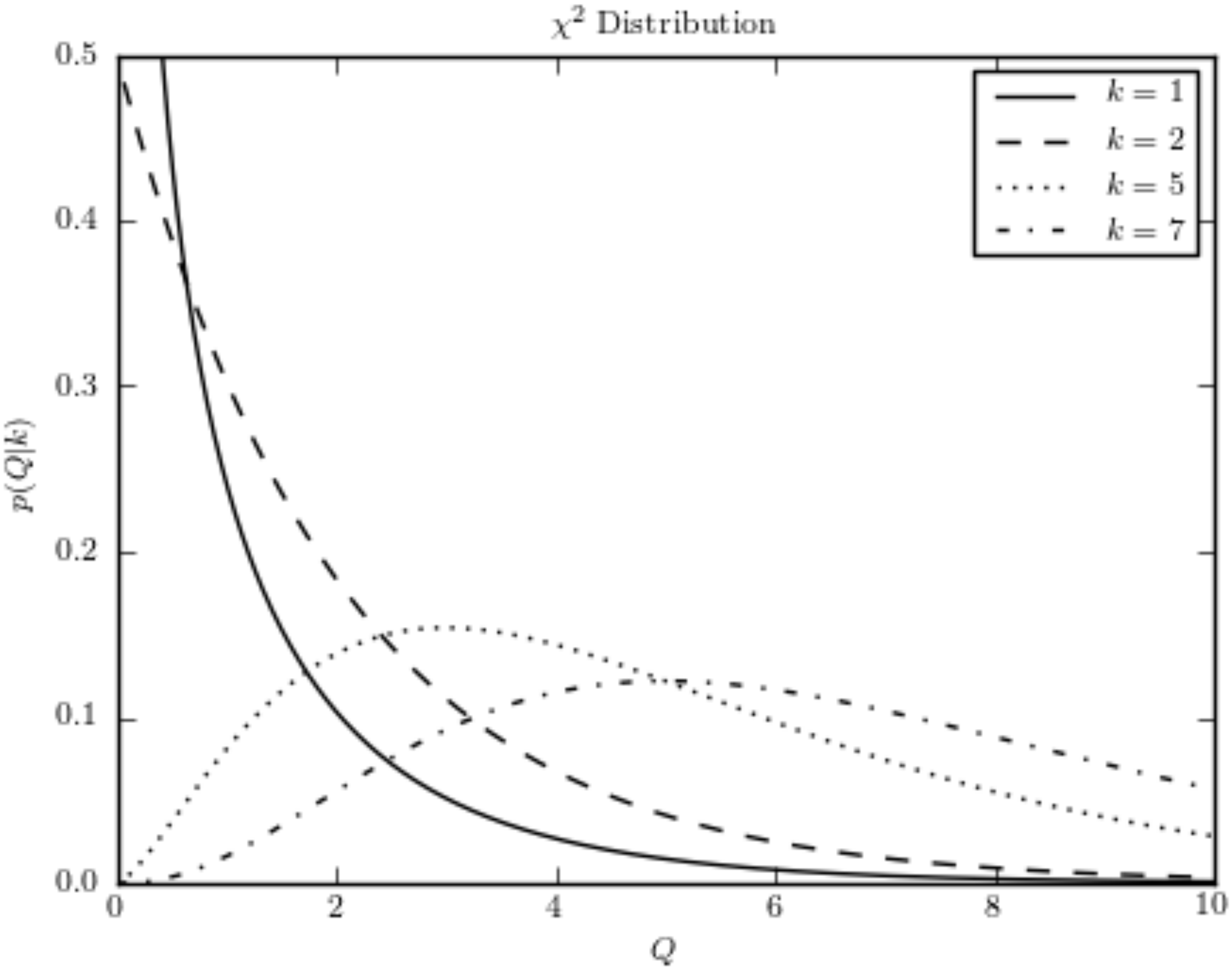


Ivezic et al. (2014)

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 $k \to \infty$, p(x|*k*) $\to$ *N*(x|*k*,2*k*)

Chi-squared distribution



Ivezic et al. (2014)

Higher-dimensional distributions

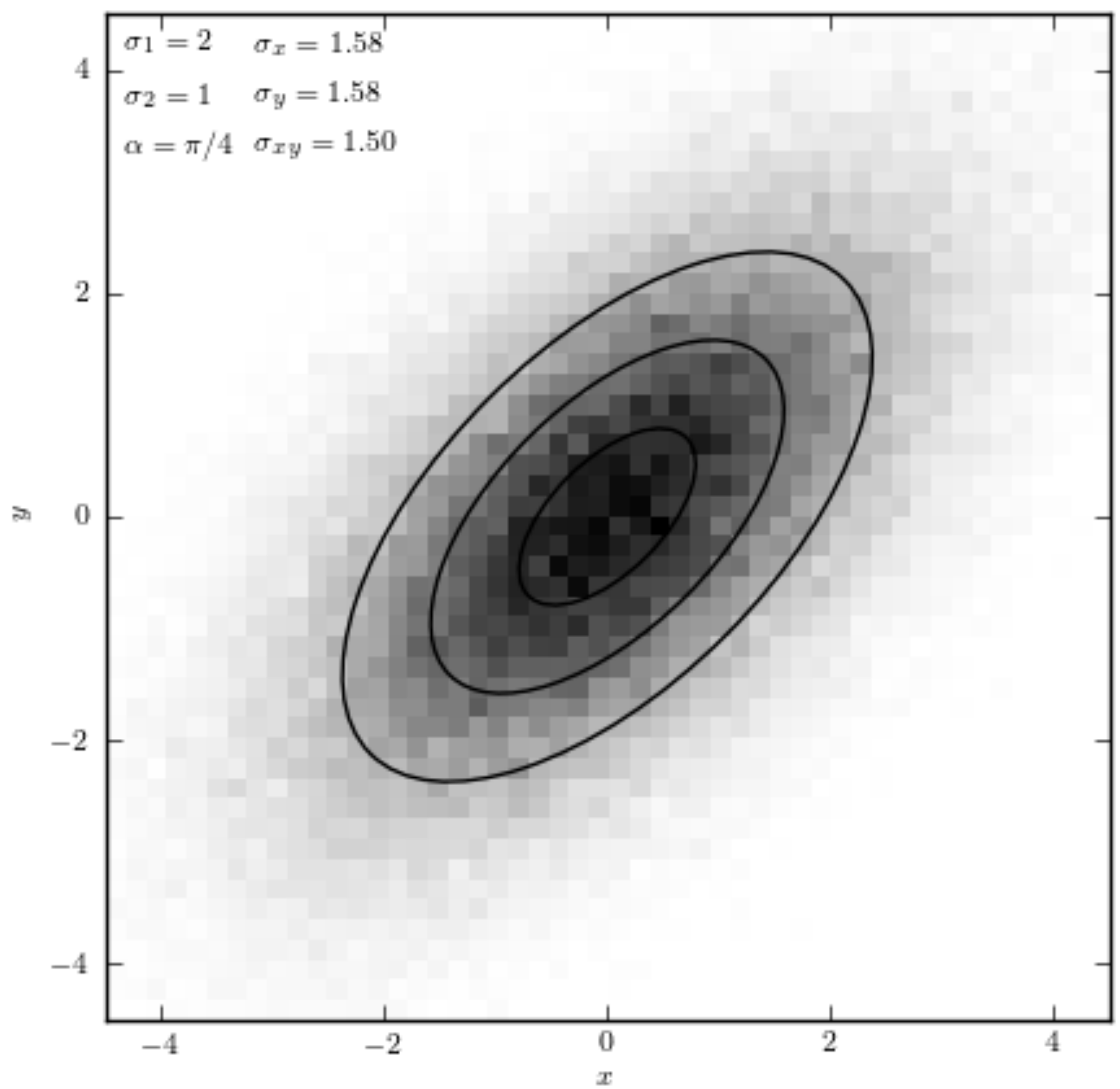
- Many fewer important ones!
- Multivariate Gaussian:

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

- E.g., in 2D

$$\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Two-dimensional Gaussian



- Correlation: $\rho = \dfrac{\sigma_{xy}}{\sigma_x \, \sigma_y}$
- Integrate over y: $\mathcal{N}\left(x|\mu_x, \sigma_x^2\right)$
- Integrate over x: $\mathcal{N}\left(y|\mu_y, \sigma_y^2\right)$

Sampling probability distribution functions

Random sampling of probability distributions

- Essential to statistics and probabilistic inference
- *Most* distributions that you encounter when fitting data cannot be easily sampled: use MCMC algorithms (third lecture)
- Simple distributions can be sampled easily
- Essentially all randomness on a computer goes back to uniformly distributed integers!
- Random number generators use a *seed*: fix this in your code to make it reproducible

Sampling using transformations

- If we transform a uniform random variable, we get a variable following a different distribution

$$p_a(a)\mathrm{d}a = p_b(b)\mathrm{d}b$$

- or

$$p_b(b) = p_a(a)\left|\frac{\mathrm{d}a}{\mathrm{d}b}\right| \qquad p_b(b) = p_a(f^{-1}[b])\left|\frac{\mathrm{d}f^{-1}[b]}{\mathrm{d}b}\right|$$

- Example: a ~ U(0,1) and b = -ln[a]

$$p(b) = 1 \times \left|\frac{\mathrm{d}e^{-b}}{\mathrm{d}b}\right| = e^{-b}$$

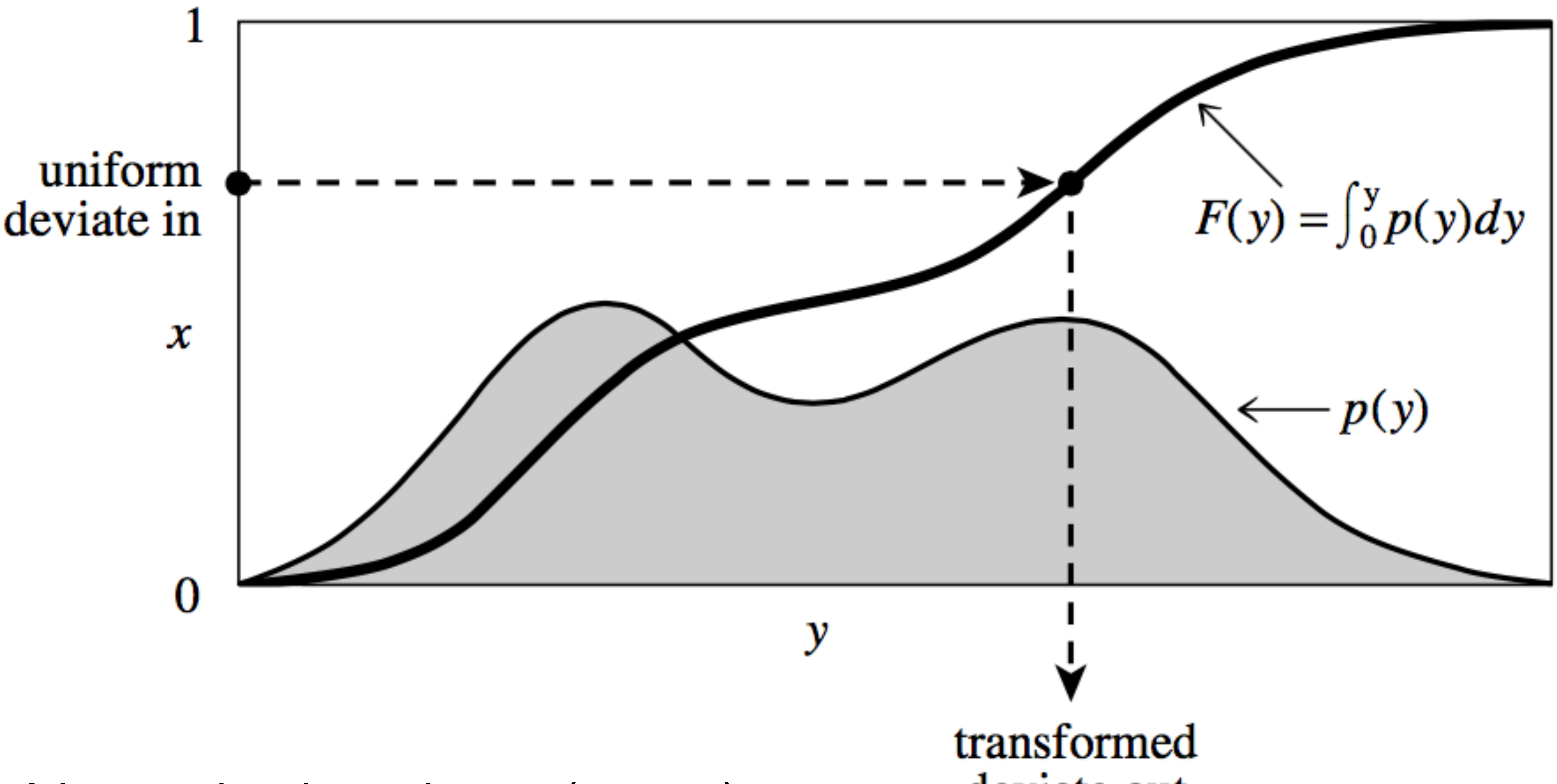
Sampling using transformations

- For a given p(b), if we can figure out a transformation a = f^{-1}(b) such that p(a) = uniform —> can sample b using transformation f(a)

$$
\begin{aligned}
p_b(b) &= p_a(f^{-1}[b])\left|\frac{\mathrm{d}f^{-1}[b]}{\mathrm{d}b}\right| \\
&= \left|\frac{\mathrm{d}f^{-1}[b]}{\mathrm{d}b}\right|
\end{aligned}
$$

- Solution of this differential equation: f^{-1}(b) is the CDF
- Thus, compute CDF(b)
- Compute f(.) = CDF^{-1}(.)
- Sample uniform a —> b = CDF^{-1}(a)

Sampling using transformations



Numerical recipes (2007)

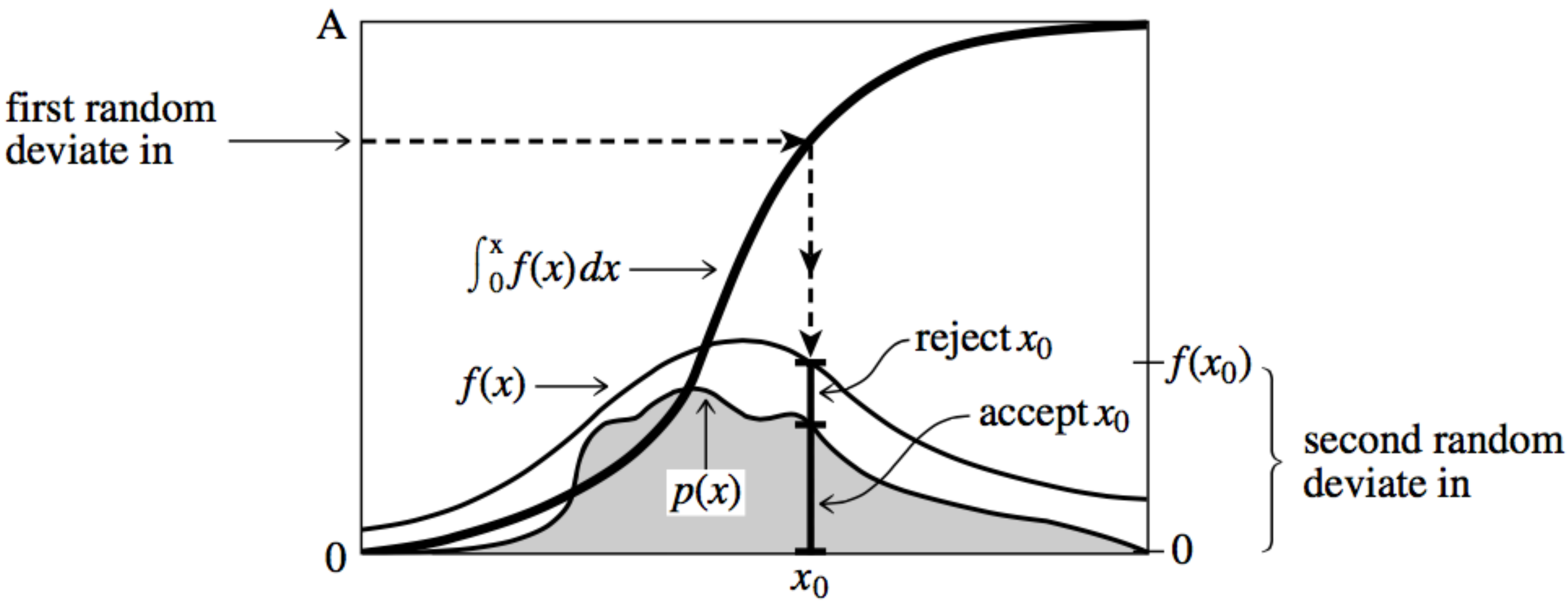
Sampling using transformations

- Works in multiple dimensions as well, but Jacobian more complicated
- Basis of *Box-Muller* method for sampling Gaussian random numbers

$$b_1 = \sqrt{-2\ln a_1}\,\cos 2\pi a_2$$
$$b_2 = \sqrt{-2\ln a_1}\,\sin 2\pi a_2$$

Rejection sampling

- Sampling from p(b) == uniformly sampling area under p(b)



Numerical recipes (2007)