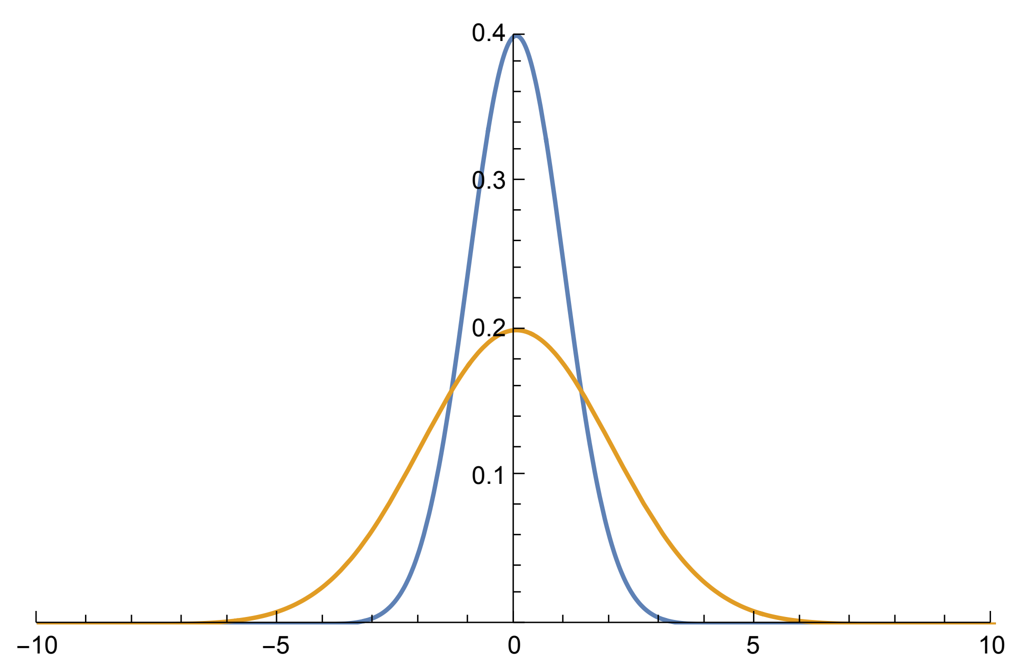
Probability: More examples

Jonathan Pritchard



Overview

- Why Gaussians? Central Limit Theorem
- Gaussian inference
- Gaussian linear models
- Poisson processes

References

- Loredo's *Bayesian Inference in the Physical Sciences*:
  - http://astrosun.tn.cornell.edu/staff/loredo/bayes
  - "The Promise of Bayesian Inference for Astrophysics" & "From Laplace to SN 1987a"
- MacKay, *Information Theory, Inference & Learning Algorithms*
- Jaynes, *Probability Theory: the Logic of Science*
  - And other refs at http://bayes.wustl.edu
- Hobson et al, *Bayesian Methods in Cosmology*
- Sivia, *Data Analysis: A Bayesian Tutorial*

Bayes Theorem

Likelihood Prior

Posterior

$$\boxed{P(\Theta|D,I) = \frac{P(D|\Theta,I)P(\Theta|I)}{P(D|I)}}$$

Evidence

parameters

data

other information
e.g. model

Gaussian distribution

- One of the most common distributions in statistics

$$\boxed{P(x|\mu,\sigma,I) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]}$$

Moments $\langle x \rangle = \mu$ $\langle (x-\mu)^2 \rangle = \sigma^2.$

All higher cumulants κ_n are zero => mean & variance tell you everything about distribution

Gaussians & CLT

- **Central Limit Theorem:**

  The sum of a *n* random numbers drawn from a probability distribution of finite variance σ^2 tends to be Gaussian distributed about the expectation value of the sum with variance $n\sigma^2$

- Applies asymptotically hence, Limit Theorem

- Means that statistics of large set of random numbers becomes independent of distribution of individual numbers
  => Gaussian widely applicable

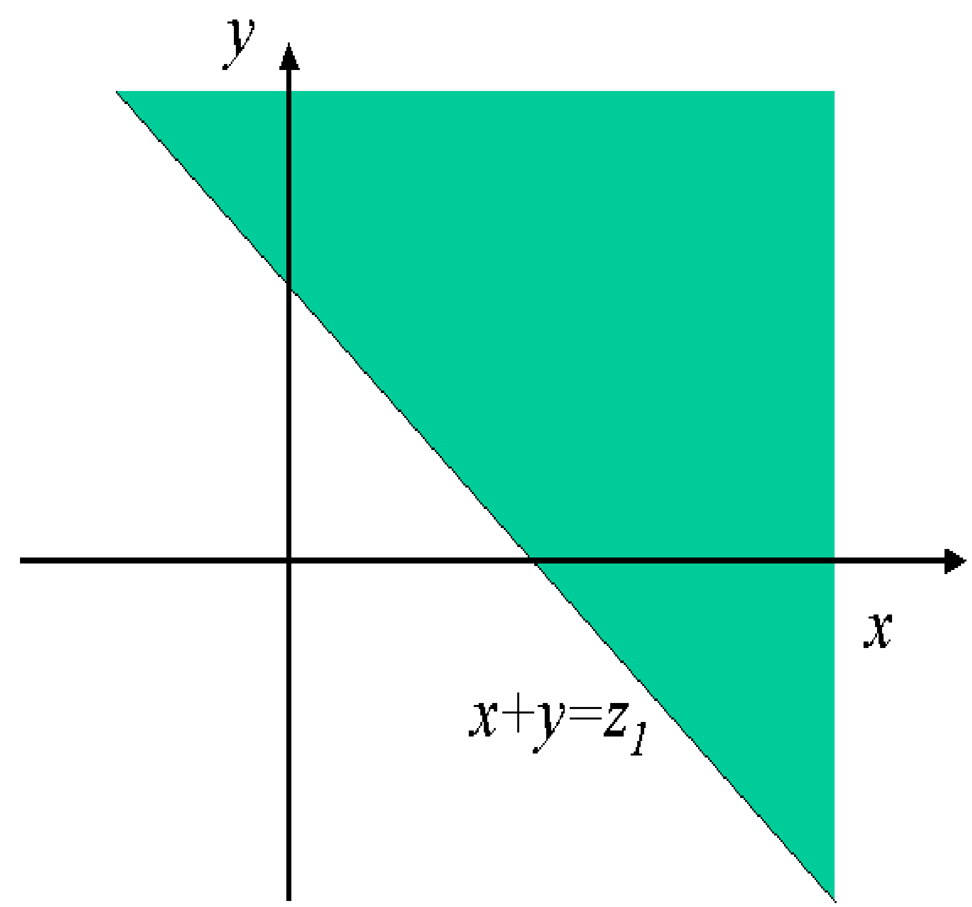
Sketch of a proof of CLT

- Consider sum of two random variables x & y

$$z = x + y$$

- Want to know p(z)

$$\begin{aligned} p(z \geq z_1) &= \int_{z_1}^{\infty} \mathrm{d}z\, p(z) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y \int_{z_1 - y}^{\infty} \mathrm{d}x\, p(x, y) \end{aligned}$$



Transform back to z: x=z-y

$$p(z \geq z_1) = \int_{-\infty}^{\infty} \mathrm{d}y \int_{z_1}^{\infty} \mathrm{d}z\, p(z - y, y)$$

Comparison gives $$p(z) = \int_{-\infty}^{\infty} \mathrm{d}y\, p(z - y, y).$$

Sketch of a proof (II)

So we have $$p(z) = \int_{-\infty}^{\infty} \mathrm{d}y \, p(z-y, y).$$

Assuming independence $$p(z) = \int_{-\infty}^{\infty} \mathrm{d}y \, p_x(z-y) p_y(y),$$

Which is just the convolution of $p_x(x)$ and $p_y(y)$

Recall from Fourier theory that FT of convolution is a product, so helpful to think in Fourier space

Characteristic function
= F.T. of prob distribution

$$\phi(k) = \int_{-\infty}^{\infty} \mathrm{d}x \, p(x) e^{ikx}$$

characteristic function

$$p(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \, \phi(k) e^{-ikx}.$$

prob. distribution

So for z have: $$\phi_z(k) = \phi_x(k)\phi_y(k)$$

Sum of n random variables $X = \frac{1}{\sqrt{N}}(x_1 + x_2 + ... + x_n)$

p(X) will be convolution of all the $p_x(x_i)$

So characteristic fn is a product $\phi_X(k) = [\phi_x(k/\sqrt{n})]^n.$

Expand characteristic fn

$$\phi_x(k/\sqrt{N}) = \int_{-\infty}^{\infty} \mathrm{d}x\, p(x) e^{ikx/\sqrt{N}} \approx 1 + i\frac{k}{\sqrt{N}}\langle x\rangle - \frac{1}{2}\frac{k^2}{N}\langle x^2\rangle + O\left(\left[\frac{k}{\sqrt{N}}\right]^3\right)$$

Assume $\langle x\rangle = 0,$ $\langle x^2\rangle = \sigma_x^2.$ Higher terms ~$O(n^{-3/2})$ & vanish

Then $\phi_X(k) = \left[1 - \frac{k^2\sigma_x^2}{2n}\right]^n \to e^{-\sigma_x^2 k^2/2}$ $n \to \infty.$

Gaussian, so when we FT get a Gaussian.

$$p(X) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-X^2/(2\sigma_x^2)}$$

variance of mean $\sigma_x/\sqrt{n}.$

Central limit theory leads to Gaussian distribution

Gaussians & belief

- Alternatively, can ask what distribution is **least informative** if we know mean and variance
=> again leads to Gaussian

- Can show this rigourously from maximum entropy considerations. (in continuous case need extra fn m(x) to insure invariance under parameter change)

$$S = -\sum_i^N p_i \log[p_i] \rightarrow -\int p(x) \log\left[\frac{p(x)}{m(x)}\right]$$

- Maximising S subject to known mean μ & variance σ (e.g. by Lagrange multipliers) produces Gaussian

$$Q = -\sum_i^N p_i \log\left[\frac{p_i}{m_i}\right] + \lambda_0\left(1 - \sum p_i\right) + \lambda_1\left(\mu - \sum x_i p_i\right) + \lambda_2\left(\sigma^2 - \sum (x_i - \mu)^2 p_i\right)$$

Recover the standard Gaussian distribution

$$P(x|\mu, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

Why Gaussians?

- **Central Limit Theorem**: sum of many random numbers has a Gaussian sampling distribution
- **MaxEnt**: If we know mean & variance, the least informative distribution is Gaussian

Gaussian inference

- Problem: want to estimate signal *s*, given *n* noisy observations $\{d_i\}$

data = signal + noise

- Need **model** for observations: $$d_i = s + n_i$$

- Noise: assume $n_i=(d_i-s)$ is Gaussian zero mean & known variance σ^2

- Work through Bayes theorem:

$$\boxed{p(s|\mathbf{d}, I) = \frac{p(\mathbf{d}|s, I)p(s|I)}{p(\mathbf{d}|I)}}$$

Prior $p(s|I)$

- How do we choose prior? Often to encode ignorance about *s*

- Common options?

  Gaussian with zero mean and variance Σ.
  Let $\Sigma \rightarrow \infty$ at end of calculation

  Uniform in range $[\Sigma_1, \Sigma_2]$. Again let $\Sigma_1 \rightarrow -\infty$, $\Sigma_2 \rightarrow \infty$ at end

  "Jeffrey's prior", $p(s|I) \propto 1/s$. Appropriate if ignorant about scale of *s*. Equivalent to flat prior on log*s*

- Here adopt uniform prior: $$\boxed{p(s|I) = \frac{1}{\Sigma_2 - \Sigma_1}} \text{ if } \Sigma_1 \leq s \leq \Sigma_2$$

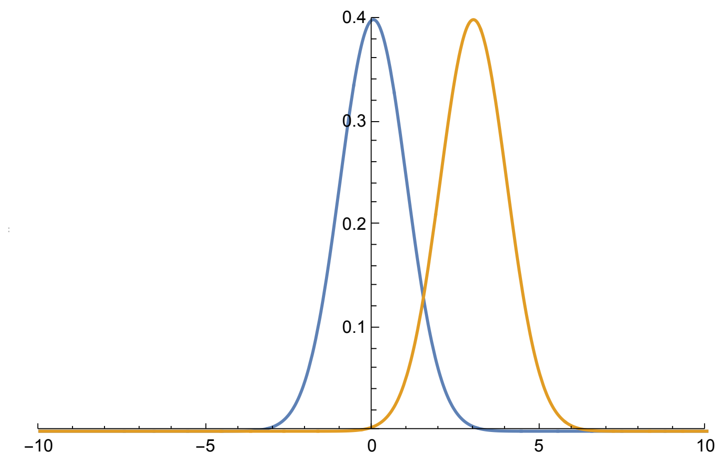
Priors

- Can think about priors from perspective of properties of pdf

- Location priors: do I know the origin?
  => want pdf invariance under translation $X \rightarrow X + x_0$

$$p(X|I)\mathrm{d}X \;\approx\; p(X+x_0|I)\mathrm{d}(X+x_0)$$

$$\approx\; p(X+x_0|I)\mathrm{d}X$$

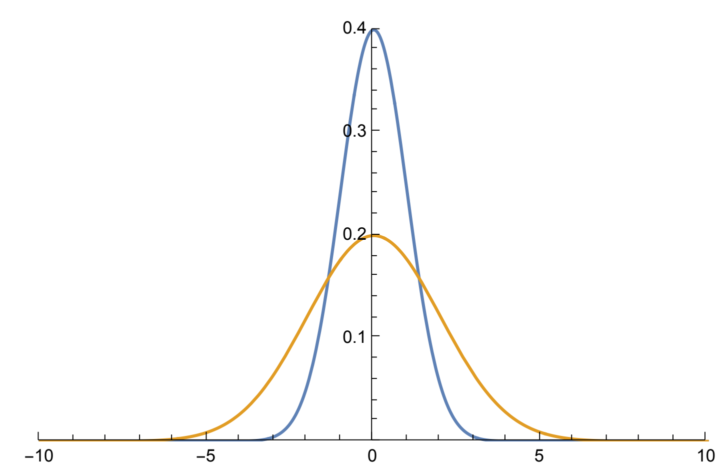


=> uniform prior $p(X|I) = \mathrm{const}$

- Scale priors: Am I sure on the units?
  => want pdf invariance under rescaling $\sigma \rightarrow \beta\sigma$

$$p(\sigma|I)\mathrm{d}X \;\approx\; p(\beta\sigma|I)\mathrm{d}(\beta\sigma)$$

$$p(\sigma|I) \;\approx\; p(\beta\sigma|I)\,\beta$$



=> uniform in log prior $p(\sigma|I) \propto 1/\sigma$

Likelihood $p(\mathbf{d}|s, I)$

- We've decided our noise is Gaussian, so for individual datum have

$$p(d_i|s, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(d_i - s)^2}{\sigma^2}\right]$$

- For full data set:

$$\boxed{p(\mathbf{d}|s, I) = (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i^n (d_i - s)^2\right]}$$

- Fine, but helpful to manipulate analytically

Recall mean $\bar{d} = \frac{1}{N}\sum_i d_i.$

$$\sum_i^n (d_i - s)^2 = \sum_i^n (d_i^2 - 2d_i s + s^2) = N(s - \bar{d})^2 + N\sum_i \frac{(d_i - \bar{d})^2}{N}$$

- Result separates into two parts

data+parameters data only

$$\boxed{p(\mathbf{d}|s, I) = (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}(s - \bar{d})^2\right] \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i - \bar{d})^2\rangle\right]}$$

$\sigma_b \equiv \sigma/\sqrt{N}$ $\langle(d_i - \bar{d})^2\rangle = \sum_i \frac{(d_i - \bar{d})^2}{N}.$

Evidence $p(\mathbf{d}|I)$

Evidence plays role of normalisation factor here

$$1 = \int \mathrm{d}s\, p(s|\mathbf{d},I) = \int \mathrm{d}s\, \frac{p(\mathbf{d}|s,I)p(s|I)}{p(\mathbf{d}|I)} \quad \longrightarrow \quad p(\mathbf{d}|I) = \int \mathrm{d}s\, p(\mathbf{d}|s,I)p(s|I)$$

So taking results for prior and likelihood

$$\begin{aligned} p(\mathbf{d}|I) &= \int_{\Sigma_1}^{\Sigma_2} \mathrm{d}s\, (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right] \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \\ &= (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \\ &\quad \times \int_{\Sigma_1}^{\Sigma_2} \mathrm{d}s\, \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right] \end{aligned}$$

Recall definition of error function $\quad \mathrm{erf}\,x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$

Gives final result for evidence

$$\boxed{p(\mathbf{d}|I) = (2\pi\sigma^2)^{N/2} \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \frac{\sqrt{2\pi\sigma^2}}{\sqrt{N}} \frac{1}{2} \left[\mathrm{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right) - \mathrm{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]}$$

Posterior

Combine results in Bayes theorem $$p(s|\mathbf{d},I)=\frac{p(\mathbf{d}|s,I)p(s|I)}{p(\mathbf{d}|I)}$$

$$= \boxed{p(\mathbf{d}|s,I)=(2\pi\sigma^2)^{n/2}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]\exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right]} \quad \text{X} \quad \boxed{p(s|I)=\frac{1}{\Sigma_2-\Sigma_1}}$$

$$\boxed{p(\mathbf{d}|I)=(2\pi\sigma^2)^{N/2}\exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right]\frac{1}{\Sigma_2-\Sigma_1}\frac{\sqrt{2\pi\sigma^2}}{\sqrt{N}}\frac{1}{2}\left[\mathrm{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right)-\mathrm{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]}$$

Gives the posterior

$$p(s|\mathbf{d},I)=\frac{\sqrt{N}}{\sqrt{2\pi\sigma^2}}2\left[\mathrm{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right)-\mathrm{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]^{-1}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]$$

Taking limit $\Sigma_1\to-\infty$, $\Sigma_2\to\infty$

$$\boxed{p(s|\mathbf{d},I)=\frac{1}{\sqrt{2\pi\sigma_b^2}}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]}$$

Inference?

Posterior contains everything that we infer about signal

$$p(s|\mathbf{d}, I) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]$$

Best estimate of signal is peak of posterior

Bayesian 68% confidence interval $s = \bar{d} \pm \sigma_b = \bar{d} \pm \sigma/\sqrt{N}.$

Alternative priors? Infinite Gaussian gives same result.

If didn't know σ^2: assume Jeffrey's prior $p(\sigma|I) \propto 1/\sigma$, then marginalise over σ, leads to broader posterior

$$p(s|I) \propto [s - 2s\langle d\rangle + \langle d^2\rangle]^{-2}.$$

(connected to Student-t distribution, same maximum, more conservative bound)

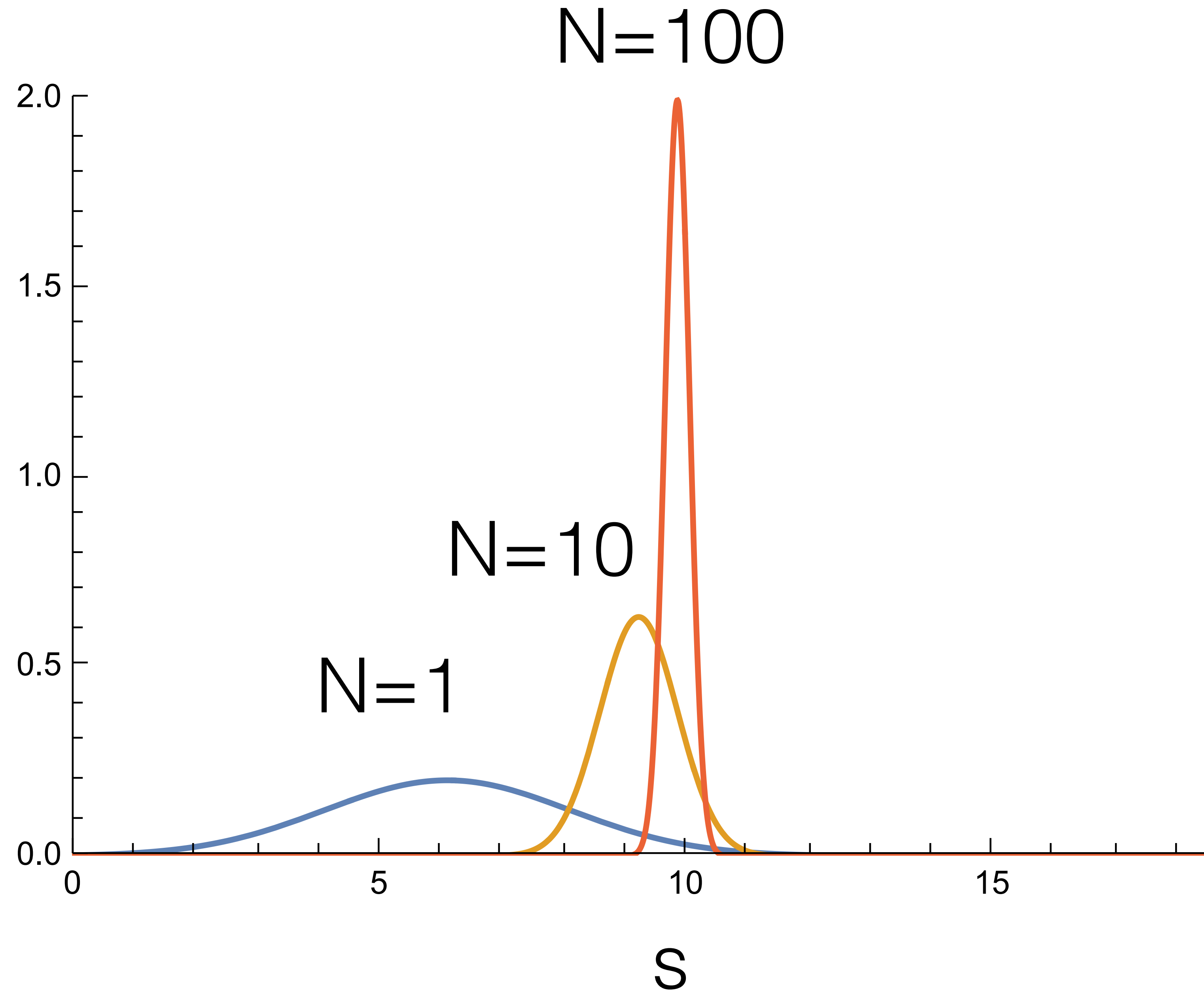
Toy example

Simple example $s_{true}=10$, $\sigma=2$

Make a random data set

6.07335, 11.213, 7.86354, 11.2595, 10.5425, 6.5558, 9.20705, 8.04459, 10.2605, 10.9534 ...

$$p(s|\mathbf{d}, I) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]$$



Straight line fitting

- Same procedure applies for more complicated signals e.g. straight-line fitting

data = signal + noise

- Let signal be linear in time

$$d_i = at_i + b + n_i$$

- Likelihood

$$p(d_i|a,b,I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(d_i - at_i - b)^2}{\sigma^2}\right]$$

- This is multivariate Gaussian in d_i. Since linear in (a,b) also multivariate Gaussian in (a,b)

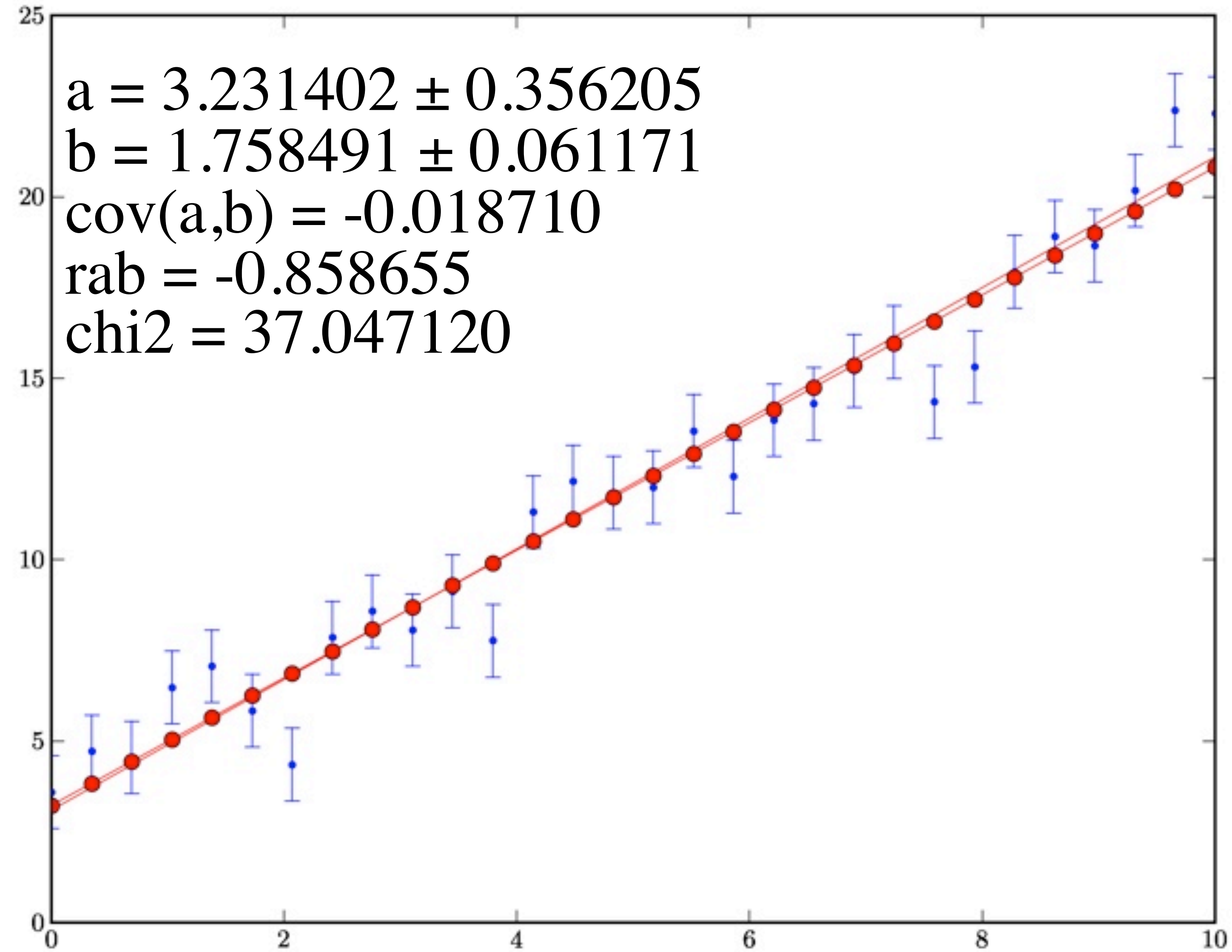
- Not normalised in (a,b) so not distribution! Needs application of Bayes Theorem with prior to get probability distribution

- Posterior maximised for same parameters as “least squares” fitting with same errors and covariance

- Same numbers, but different interpretation! (see PS1 Q0)

Line fitting

- Can use standard routines for line fitting



General linear models

- Many problems can be reduced to linear by appropriate choice of basis

- Consider $$d(t_i) = \sum_p x_p f_p(t_i) + n_i$$

  i.e. a sum of known functions of unknown coefficient plus noise. Want to infer x_p
  e.g. linear fit has $f_0(t)=1$, $f_1(t)=t$

- Assume zero mean Gaussian noise, possibly correlated
  $$\langle n \rangle = 0, \langle n_i n_j \rangle = N_{ij}$$
- Typically noise can be considered stationary (isotropic) so that $N_{ij} = N(t_j - t_i)$

- Rewrite in matrix form $$d_i = \sum_p A_{ip} x_p + n_i \qquad A_{ip} = f_p(t_i)$$
- Likelihood $$\boxed{p(d_i | x_p, I) = \frac{1}{|2\pi N|^{1/2}} \exp\left[-\frac{1}{2}(d - Ax)^T N^{-1} (d - Ax)\right]}$$

General linear models

As before can rewrite this as data-only and data+parameters terms

$$p(d_i|x_p, I) \propto \underbrace{\exp\left[-\frac{1}{2}(d - A\bar{x})^T N^{-1}(d - A\bar{x})\right]}_{\text{depends on data only}} \underbrace{\exp\left[-\frac{1}{2}(x - \bar{x})^T C^{-1}(x - \bar{x})\right]}_{\text{depends on data \& parameters}}$$

$$\propto \exp\left[-\frac{1}{2}(d - AWd)^T N^{-1}(d - AWd)\right] \exp\left[-\frac{1}{2}(x - Wd)^T C^{-1}(x - Wd)\right]$$

The parameter independent part is just $e^{-\chi^2_{\max}}$

The parameter dependent part makes clear that the **likelihood is a multivariate Gaussian** with mean

$$\bar{x} = Wd = (A^T N^{-1} A)^{-1} A^T N^{-1} d$$

and variance C

$$C = (A^T N^{-1} A)^{-1}$$

General linear models

In the limit of an infinitely wide uniform (or Gaussian) prior on x then the posterior is

$$\boxed{p(x|\mathbf{d}, I) = \frac{1}{|2\pi C|^{1/2}} \exp\left[-\frac{1}{2}(x - Wd)^T C^{-1}(x - Wd)\right]}$$

As before, normalisation cancelled out the e^{-x^2} part

Best estimate of x is the noise weighted mean

$$\bar{x} = Wd = (A^T N^{-1} A)^{-1} A^T N^{-1} d$$

We get errors on x from the covariance matrix $\langle \delta x_p \delta x_q \rangle = C_{pq}$

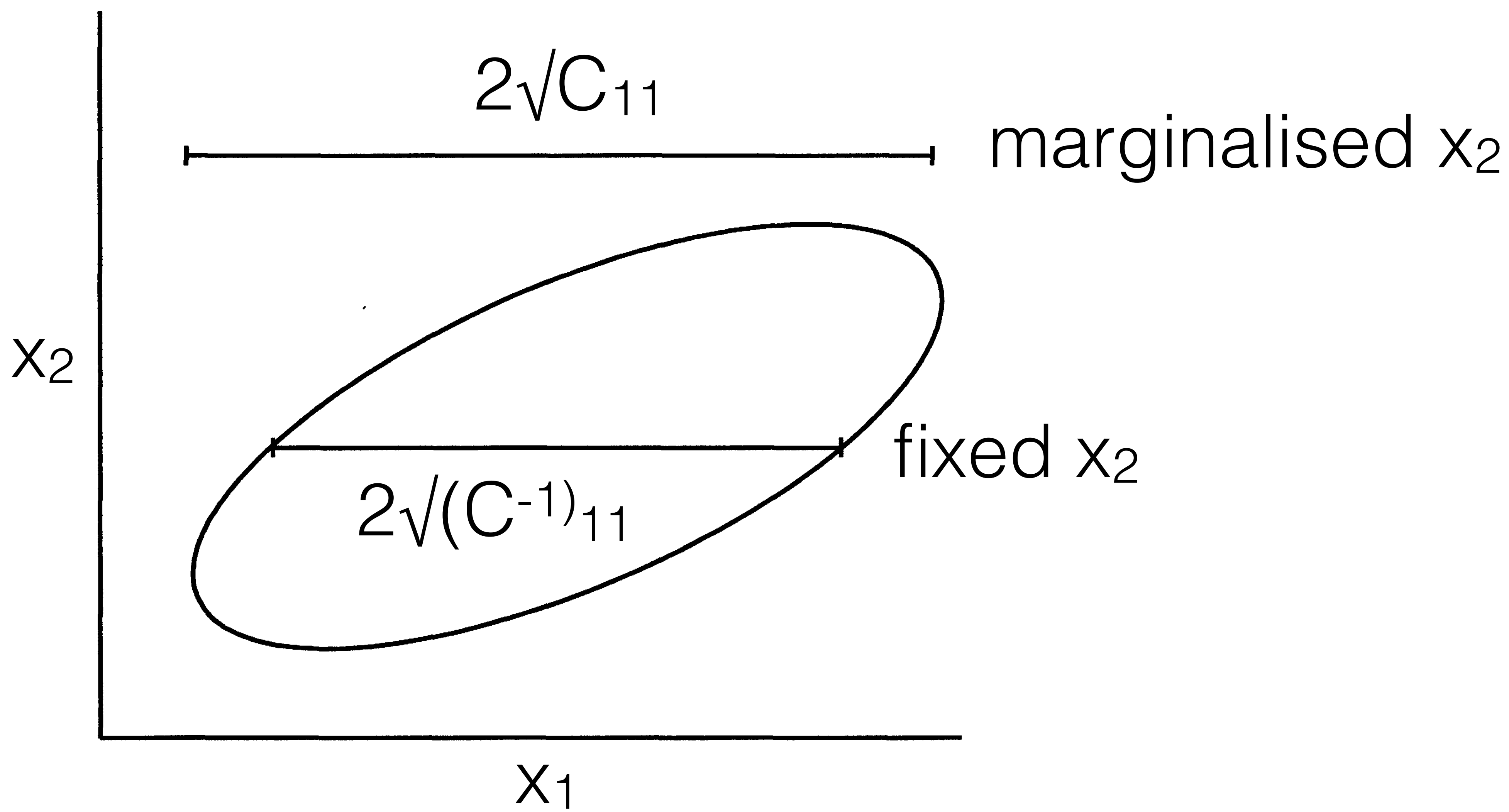
Covariance matrix $\sigma_p^2 = C_{pp}$ gives errors if we **marginalise** over all other parameters

Inverse matrix $\sigma_p^2 = 1/C_{pp}^{-1}$ gives errors if we **fix** all other parameters

Covariance matrix

Covariance matrix $\sigma_p^2 = C_{pp}$ gives errors if we **marginalise** over all other parameters

Inverse matrix $\sigma_p^2 = 1/C^{-1}_{pp}$ gives errors if we **fix** all other parameters



For Gaussian distribution, marginalising one or more parameters doesn't shift the best fit values of the others. Not true for a general distribution.

Chi Squared

- The exponential part of a Gaussian always takes the form $\exp(-\chi^2/2)$
- In the Likelihood, we have $\chi^2 = \Sigma_i(\mathrm{data}_i - \mathrm{model}_i)^2/\sigma^2$
- For fixed model, χ^2 has a χ^2 distribution with number of degrees of freedom $\nu = N_{\mathrm{data}} - N_{\mathrm{parameters}}$
- The distribution peaks at $\chi^2 = \nu \pm \sqrt{2\nu}$
- Chi squared too big or small can be sign of poor model (overfitting or too many parameters)
- Frequentist arguments, but useful rule of thumb

Poisson processes

- Poisson processes occur when counting discrete events.

- Can occur in two different ways:
  - Course measurements where "bin" events and can only report number of events in one or more finite intervals (counting process).
  - Fine measurements where count individual events (point process)

- Poisson statistics obey two key properties:

  (1) Given an event rate *r*, the probability for finding an event in an interval d*t* is proportional to the size of the interval

$$p(E|r, I) = r\,\mathrm{d}t.$$

  (2) Probabilities for different intervals are independent

Poisson distribution

- Poisson probability distribution

$$p(n|\lambda, I) = \frac{\lambda^n}{n!} e^{-\lambda}$$

- Moments

$$\langle n \rangle \equiv \sum_{n=0}^{\infty} np(n|r, I) = rT = \lambda$$

$$\langle (n - \langle n \rangle)^2 \rangle = \langle n \rangle = \lambda$$

- So single parameter describes Poisson distribution
- (M→∞ limit of Binomial distribution, for N successes in M trials)
- Can derive from Maximum Entropy as least restrictive distribution given known expectation for number of events in fixed interval (see Sivia Chap 5).

Poisson inference

- Let's say we measure *n* events in an interval of time *T* and we want to infer the event rate *r*

$$p(r|n,I) = \frac{p(n|r,I)p(r|I)}{p(n|I)}$$

- Likelihood

$$p(n|r,I) = \frac{(rT)^n}{n!}e^{-rT}$$

- For prior two common options:

  \- *r* known to be non-zero. Its a scale parameter

$$p(r|I) \propto 1/r = 1/[r\log(r_u/r_l)]$$

  \- r can be zero. Uniform prior

$$p(r|I) = 1/r_u.$$

- Taking scale parameter prior, we get posterior

$$p(r|n,I) = \frac{Te^{-rT}(rT)^{n-1}}{(n-1)!}$$

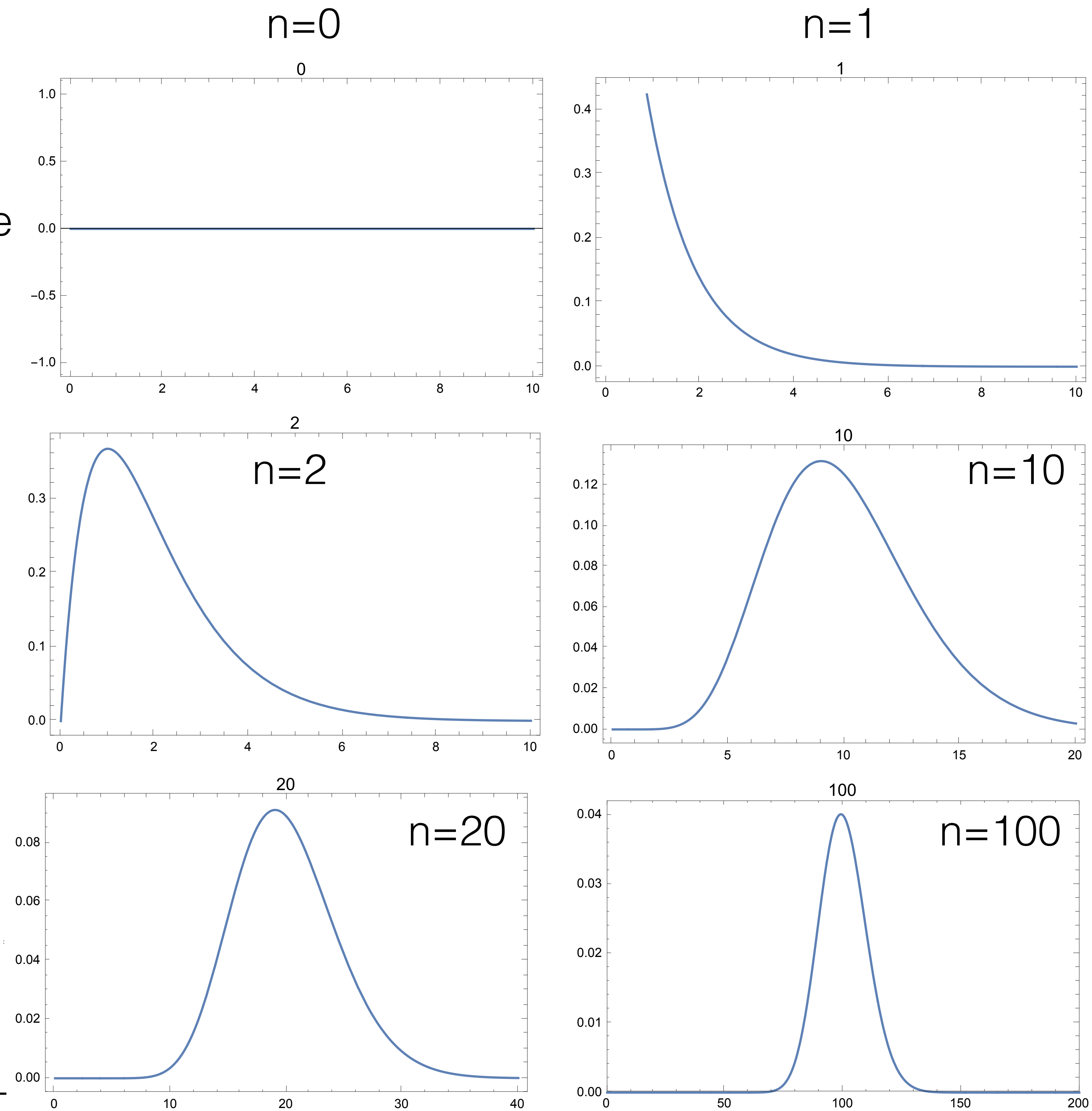
Best estimate of rate is then $rT = n \pm \sqrt{n}$ (uniform prior would give n+1)

ICIC

Inferences for rate

$$p(r|n, I) = \frac{e^{-rT}(rT)^{n-1}}{(n-1)!}$$

n=0 have no information to make inference



n=100 posterior becomes close to Gaussian

$$rT = n \pm \sqrt{n}$$

Poisson rates

- Backgrounds: $n = b + s$
  - can fix or infer known or unknown background rate
  - e.g. n_b from T_b spent observing background and n_s from T_s observing $(b+s)$
  - See Loredo articles for detailed examples

- Spatial or temporal variation in signal (or background) e.g. $s = s(t)$

- e.g. counts of cosmic rays over sky, neutrinos

- Arrival statistics of individual rare particles e.g. UHECR

Conclusions

- Gaussian distributions are everywhere! Arise from Central Limit Theorem; arise when all you know is mean & variance.
- Gaussian linear model equivalent to “generalised least squares” => many toolkits work for Bayesian analysis
- Poisson statistics important for discrete events e.g. counting problems, arrival statistics
- Can view distributions as statements about what you believe - often make most ignorant choices, but don’t have to especially for priors.
- Framework is general and explicit about assumptions. Makes it easy to modify assumptions to fit specific problems.