Confidence intervals, hypothesis testing, Monte Carlo, and generalized linear models

Roger Levy

9.S918: Quantitative inference in brain and cognitive sciences

19 February 2025

Credible intervals & confidence intervals

- A **point estimate** of a model parameter is one example of a **statistic**

(**Wikipedia**: *"A **statistic**...or **sample statistic** is any quantity computed from values in a sample which is considered for a statistical purpose"*)

- Point estimates we have seen thus far:
  - Maximum likelihood estimate
  - Bayesian posterior mean
  - Bayesian posterior mode

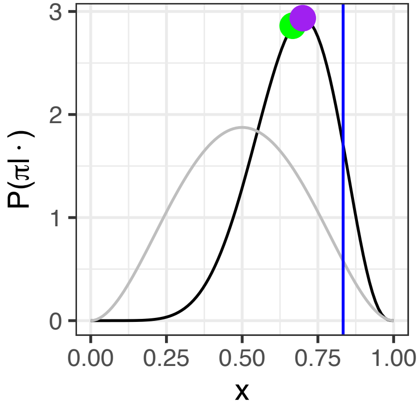
Beta–binomial

$\alpha_1 = 3$

$\alpha_2 = 3$

$r = 5$

$n = 6$



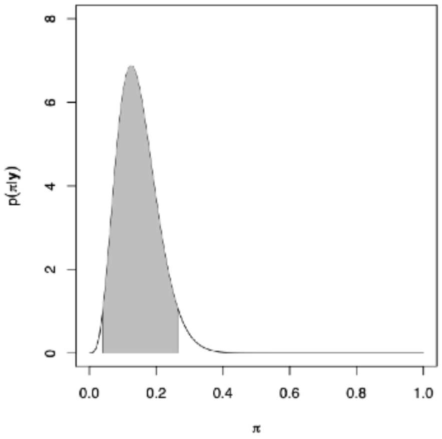
- All of these point estimates discard a lot of information about the shape of the curve that they come from!
  - Curve shape captures **uncertainty** about parameter
- Credible intervals (Bayesian) and confidence intervals (frequentist) provide a bit more information about this uncertainty

Bayesian credible intervals

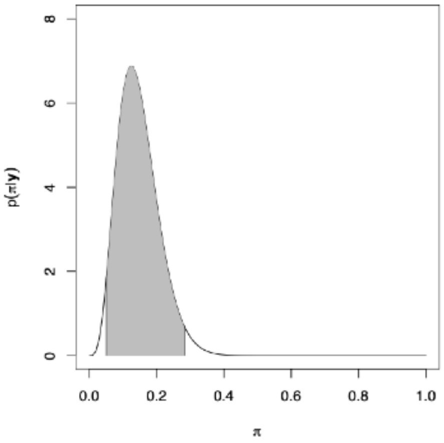
$$P(\theta|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\theta)P(\theta)}{P(\boldsymbol{y})}$$

- A $(1-\alpha)$ Bayesian credible interval (CI) on parameter π is an interval containing $(1-\alpha)$ of the posterior mass
- Two common standards for Bayesian CI construction:

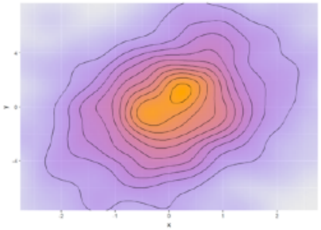
Highest posterior density



Symmetric



- Older term: "**Bayesian confidence interval**"
- Multivariate generalization: interval→region



https://www.martinmodrak.cz/2018/02/19/taming-divergences-in-stan-models/

Frequentist confidence intervals

- For model parameter θ, define a procedure for constructing from data $\mathbf{y}$ an interval I for possible θ

$$\text{Proc}(\mathbf{y}) = I$$

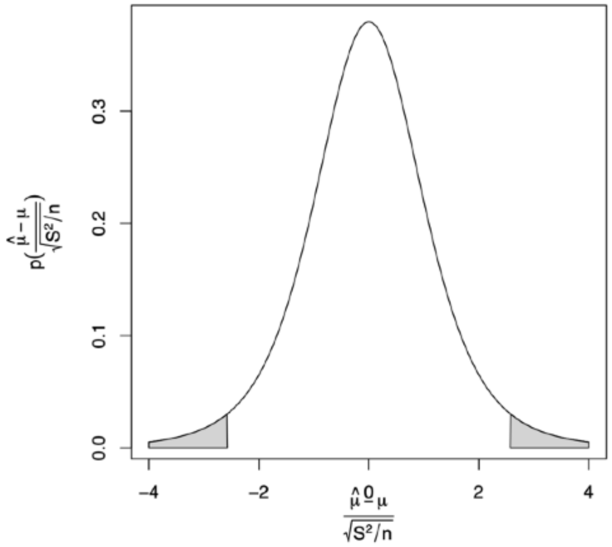
- Suppose I repeat my experiment over and over again, each time collecting data $\mathbf{y}$ and constructing $I = \text{Proc}(\mathbf{y})$
- If $(1-\alpha)$ of these intervals contain the **true value of** θ, then Proc is a method for constructing a $(1-\alpha)$ frequentist confidence interval

Confidence interval for mean μ of a normal distribution

Sample mean

$$\frac{\boxed{\hat{\mu}} - \mu}{\boxed{\sqrt{S^2/n}}} \sim t_{n-1}$$

Standard error (of the mean)



Bayesian hypothesis testing

- **Hypothesis:** a candidate theory/model for the generative process by which data $\mathbf{y}$ come into the world
- To compare hypotheses $\{H_i\}$: simply Bayesian inference!

$$P(H_i|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|H_i)P(H_i)}{\boxed{P(\boldsymbol{y})}} \longrightarrow P(\boldsymbol{y}) = \sum_{j=1}^{n} P(\boldsymbol{y}|H_j)P(H_j)$$

Normalizing constant, not of great interest for present purposes

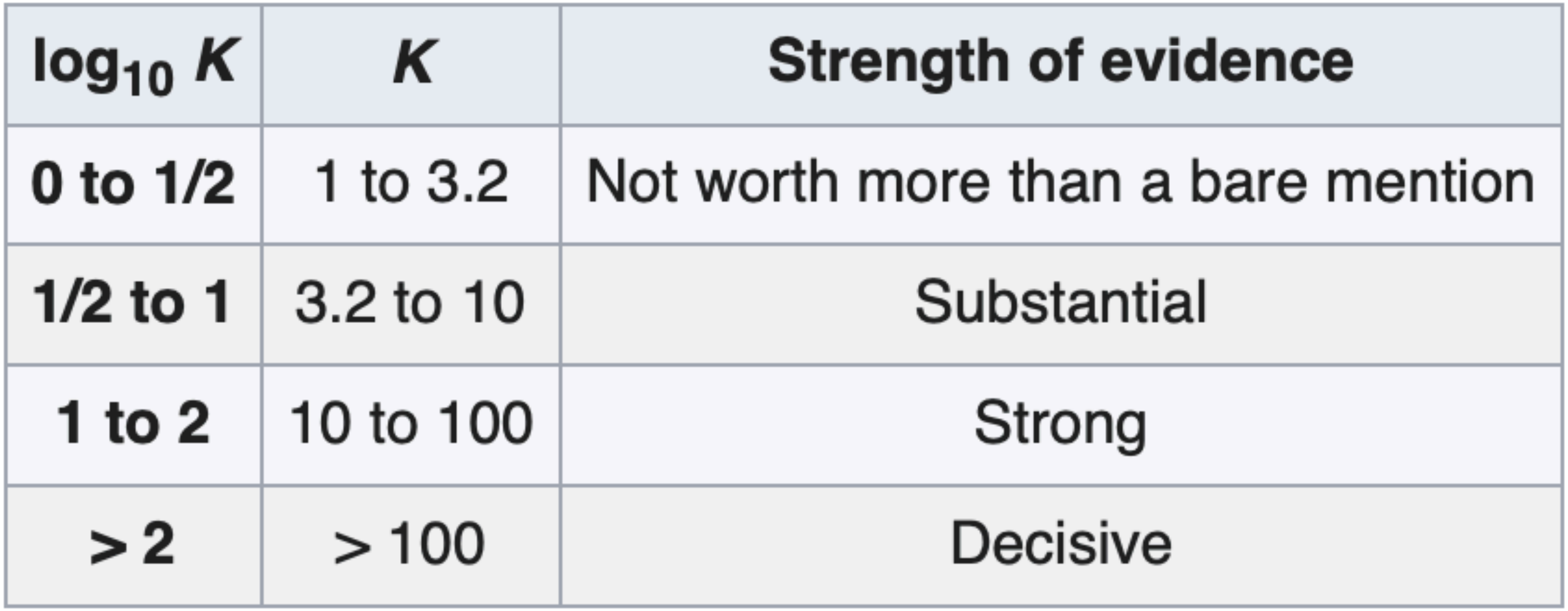
- Focus on contribution of data to posterior: **Bayes factor**

$$\overbrace{\frac{P(H|\boldsymbol{y})}{P(H'|\boldsymbol{y})}}^{\text{Posterior odds}} = \overbrace{\boxed{\frac{P(\boldsymbol{y}|H)}{P(\boldsymbol{y}|H')}}}^{\text{Likelihood ratio}} \overbrace{\frac{P(H)}{P(H')}}^{\text{Prior odds}}$$

Bayes Factor: $\dfrac{P(\boldsymbol{y}|H)}{P(\boldsymbol{y}|H')}$

Interpreting Bayes Factors

$$K = \frac{P(\boldsymbol{y}|H)}{P(\boldsymbol{y}|H')}$$



$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

(Kass & Raftery, 1995; table from https://en.wikipedia.org/wiki/Bayes_factor)

Example of Bayesian hypothesis testing

- Once again the case of the bent coin

$$H_1 : P(\pi|H_1) = \begin{cases} 1 & \pi = 0.5 \\ 0 & \pi \neq 0.5 \end{cases}$$

"The coin is fair"

$$H_3 : P(\pi|H_3) = 1 \quad 0 \leq \pi \leq 1$$

"The coin is not fair"*

- I flip a coin six times, and it comes up heads four times

$$P(\boldsymbol{y}|H_1) = \binom{6}{4} \pi^4 (1-\pi)^2 = \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = 0.23$$

$$P(\boldsymbol{y}|H_3) = \int_\pi P(\boldsymbol{y}|\pi) P(\pi|H_3)\, d\pi = \int_0^1 \overbrace{\binom{6}{4}\pi^4(1-\pi)^2}^{P(\boldsymbol{y}|\pi)} \overbrace{1}^{P(\pi|H_3)}\, d\pi \;\; = \binom{6}{4} B(5,3) = 0.14$$

$$\frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_3)} = \frac{0.23}{0.14}$$

$$= 1.64$$

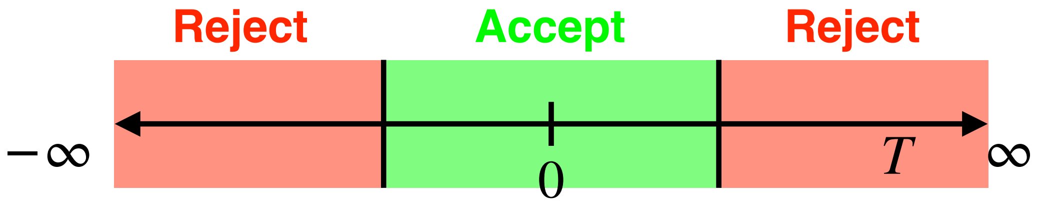
Frequentist hypothesis testing

- The **Neyman–Pearson paradigm**: Formulate two hypotheses about generative process underlying the data

  NULL HYPOTHESIS H_0

  ALTERNATIVE HYPOTHESIS H_A within which H_0 is **nested**

- Choose a TEST STATISTIC T that you'll compute from data
- Pre-data, divide range of T into ACCEPT/REJECT regions



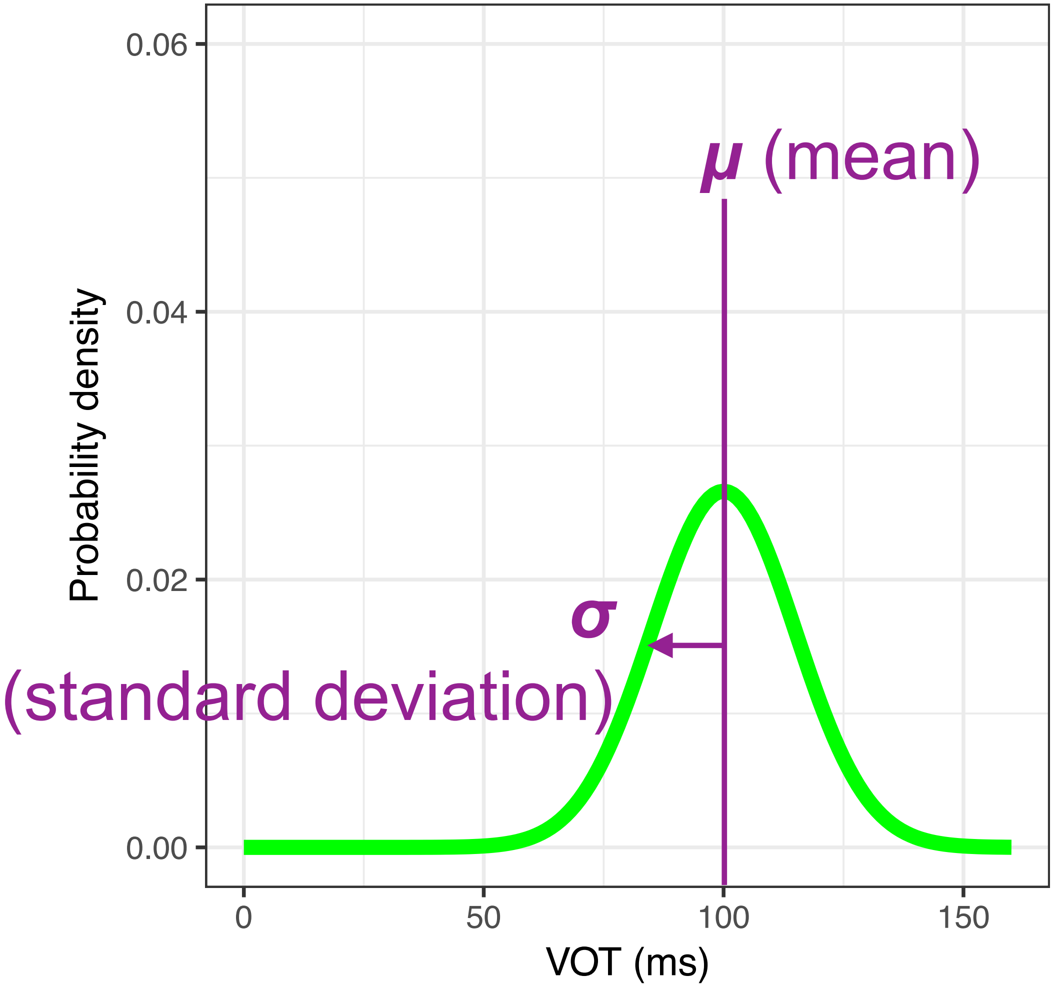
- Collect data, compute T, see where it falls!

Significance level

		Accept H_0	Reject H_0
H_0 is...	True	Correct decision (prob. $1-\alpha$)	Type I error (prob. $\boxed{\alpha}$)
	False	Type II error (prob. β)	Correct decision (prob. $\boxed{1-\beta}$)

Power

The **Gaussian**, or **normal**, distribution



Squared deviation from mean

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

(normalizing constant)

Scaled by variance

- Unbiased parameter estimates from a size-N sample:

$$\hat{\mu}=\boxed{\bar{x}}$$ ***Sample mean***

$$\hat{\sigma}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}\triangleq\boxed{s}$$ ***Sample standard deviation***

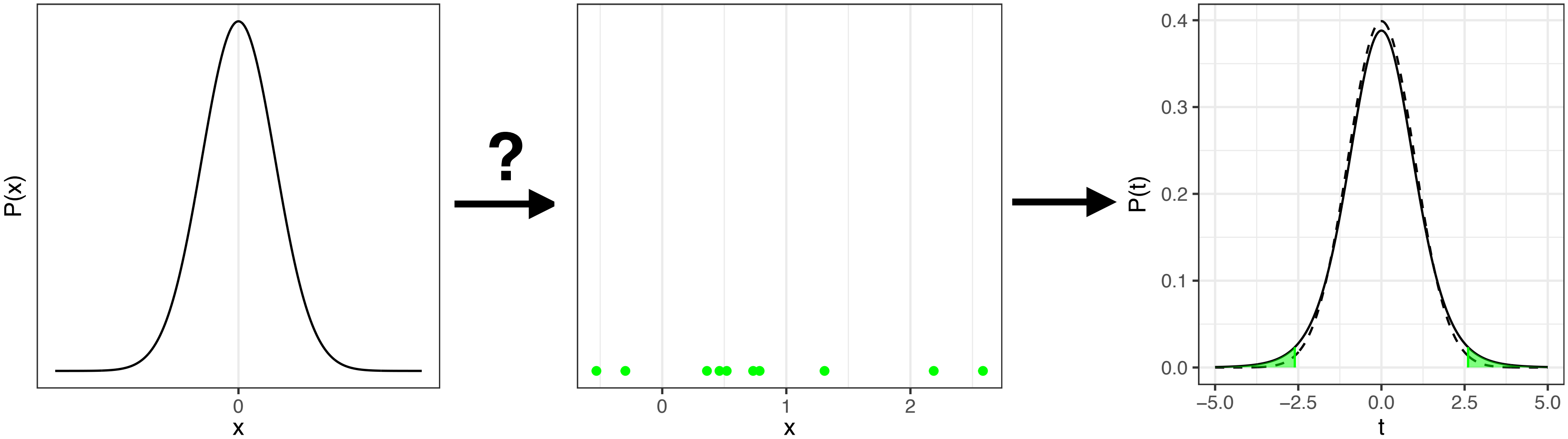
The *t*-test: three variants

- **One sample (Student's) test:** Does the underlying population mean of a sample differ from zero?
- **Two-sample test (unpaired):** do the underlying population means of two samples differ from one another?
- **Two-sample test (paired):** You have a sample of individuals from the population and take measurements from each member of the sample in two different conditions. Do the underlying population means in the two conditions differ from one another?

William Sealy Gosset, a.k.a. Student

One-sample t-test

- **Null hypothesis H$_0$:** the mean of the normally-distributed population underlying a sample is taken is $\mu = 0$
- **Alternative hypothesis H$_1$:** $\mu \neq 0$ (two tailed; generally preferred) or $\mu > 0$ (one tailed; generally dispreferred)



t-distributed with n-1 ***degrees of freedom***

Remember: ***Sample mean***

Test statistic: $\boxed{t} = \dfrac{\boxed{\bar{x}}}{\boxed{s}/\sqrt{n}}$

Remember: ***Sample standard deviation***

Two-sample *t*-test (unpaired)

- **Assumptions:** samples 1 and 2 are each iid normal
- **Null hypothesis H₀:** $\mu_1 = \mu_2$
- **Alternative hypothesis H₁:** $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed; generally dispreferred)
- If we assume that the two underlying populations have **equal variance** ("Student's" *t*-test):

$$\boxed{t} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} \quad \text{where} \quad \boxed{s_p} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n}{}_1 + n_2 - 2}$$

t-distributed w/ $n_1 + n_2 - 2$ ***degrees of freedom***

Pooled sample standard deviation

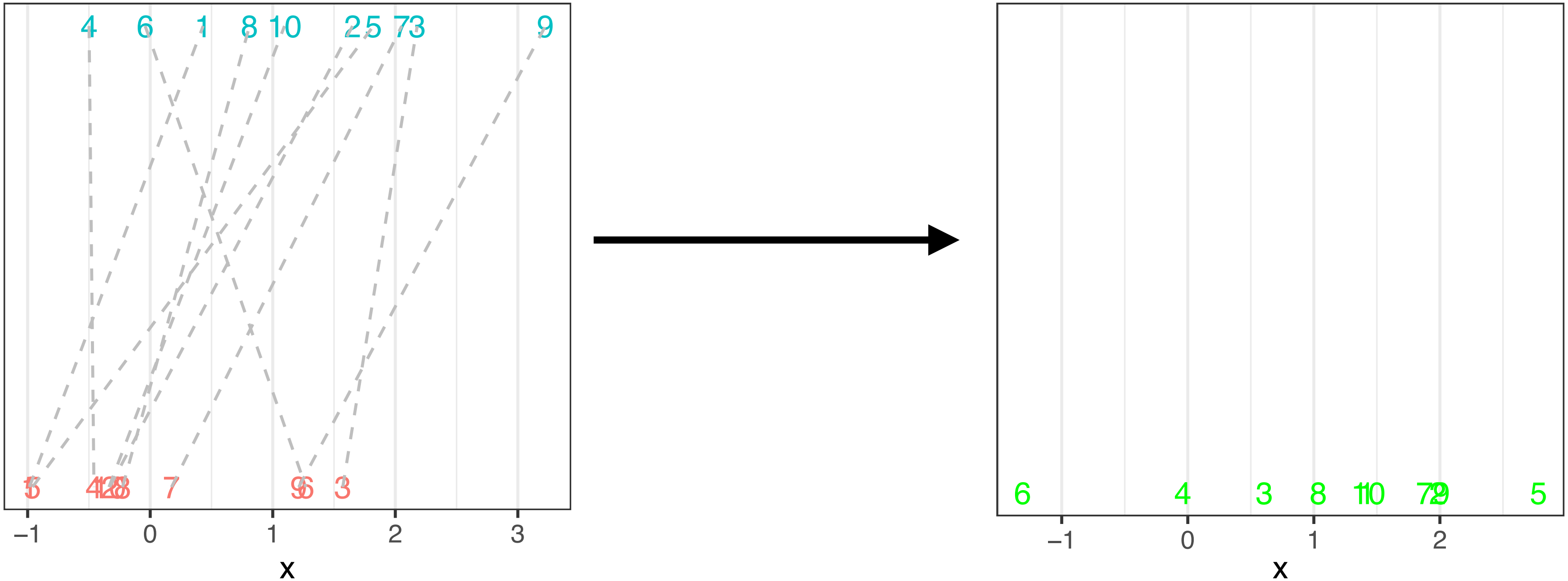
- If we **do not** assume that the two underlying populations have equal variance ("Welch's" *t*-test):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

t-distributed with a complex number of degrees of freedom whose formula can easily be looked up

Paired two-sample t-test

- **Assumptions:**
  - In a sample of **units** from a population; for each unit we have two **measurements** $\langle x_1, x_2 \rangle$ on the same scale
  - The difference between measurements is iid normal
  - (Sufficient condition: paired measurements are **bivariate normal** – a distr. we haven't yet covered)
- **H₀:** $\mu_1 = \mu_2$; **H₁**: $\mu_1 \neq \mu_2$ (2-tailed) or $\mu_1 > \mu_2$ (1-tailed; generally dispreferred)
- **Strategy:** take within-unit difference scores and apply a 1-sample t-test!



The likelihood ratio test

- The **likelihood ratio**:

$$\Lambda^* = \frac{\boxed{\max \mathrm{Lik}_{H_0}(\boldsymbol{y})}}{\boxed{\max \mathrm{Lik}_{H_A}(\boldsymbol{y})}}$$

Data likelihood under MLE of H_0

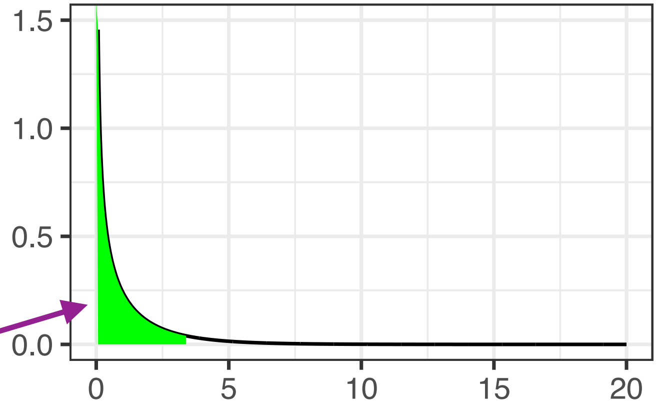
Data likelihood under MLE of H_A

- The **deviance** is (asymptotically) χ^2-distributed with **degrees of freedom** equal to diff. in # model parameters

$$G^2 \overset{\text{def}}{=} -2 \log \Lambda^*$$

- **Example**: is a coin flipped 30 times, 20H 10T, fair?

$$G^2 = -2 \log \frac{\left(\frac{1}{2}\right)^{30}}{\left(\frac{2}{3}\right)^{20}\left(\frac{1}{3}\right)^{10}} \approx 3.4$$



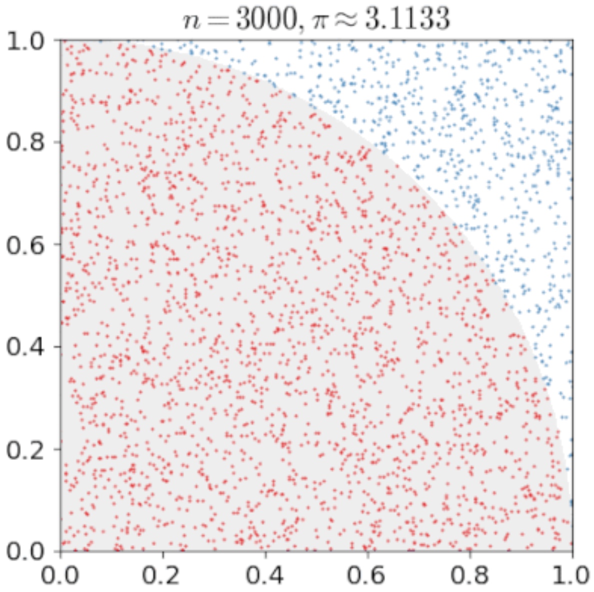
93.5% of probability mass under G^2→$p = 0.065$

Simulation and approximate computation

- All the statistical analysis that I've shown you so far has involved **exact** computation using **analytic expressions**
- This is facilitated by:
  - **Strong assumptions** regarding the data generating process (e.g., iid normal data for t-tests); and/or
  - **Conjugate priors** for Bayesian inference (e.g., Beta prior for Bernoulli/binomial data)
- But often, exact computation is not possible
- Solution: use more computationally intensive methods that don't rely on these strong assumptions. Examples:
  - Bootstrapped confidence intervals
  - Nonparametric statistical tests
  - Monte Carlo methods **(today)**

Monte Carlo methods, or "probabilistic simulation"

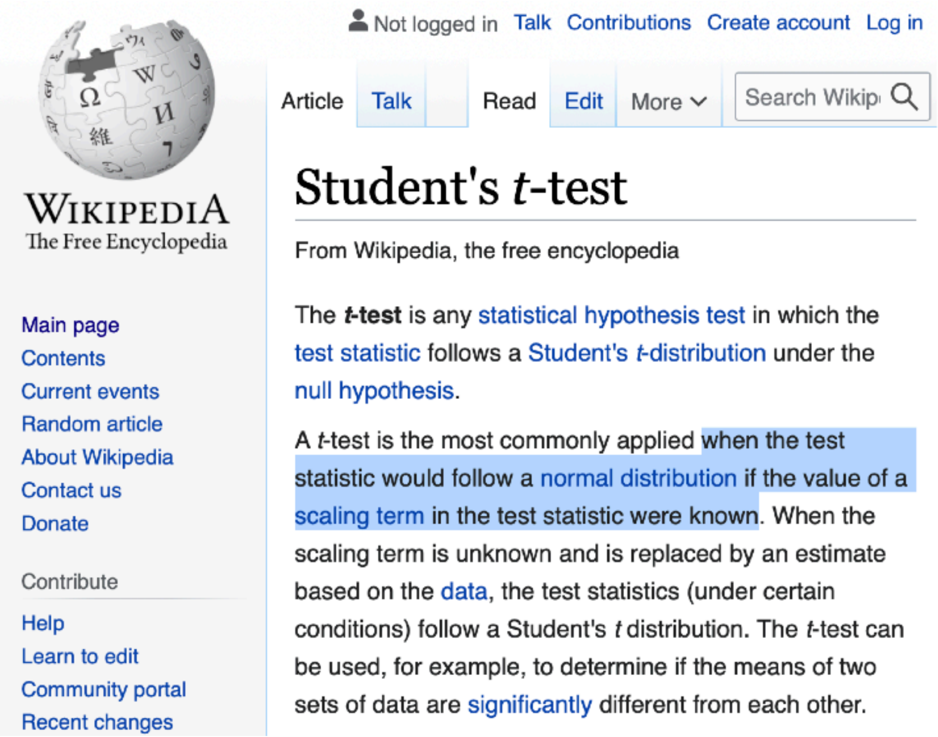
- Generally speaking:
  1. Define a domain of possible inputs
  2. Generate n iid random inputs from a probability distribution on the domain
  3. Perform a deterministic computation on each randomly generated input
  4. Aggregate the results of the deterministic computation
- As n grows larger, the simulated result approaches the true value



https://en.wikipedia.org/wiki/Monte_Carlo_method

Simple example of Monte Carlo

- Suppose I want to do a two-sample *t*-test but my data aren't normally distributed

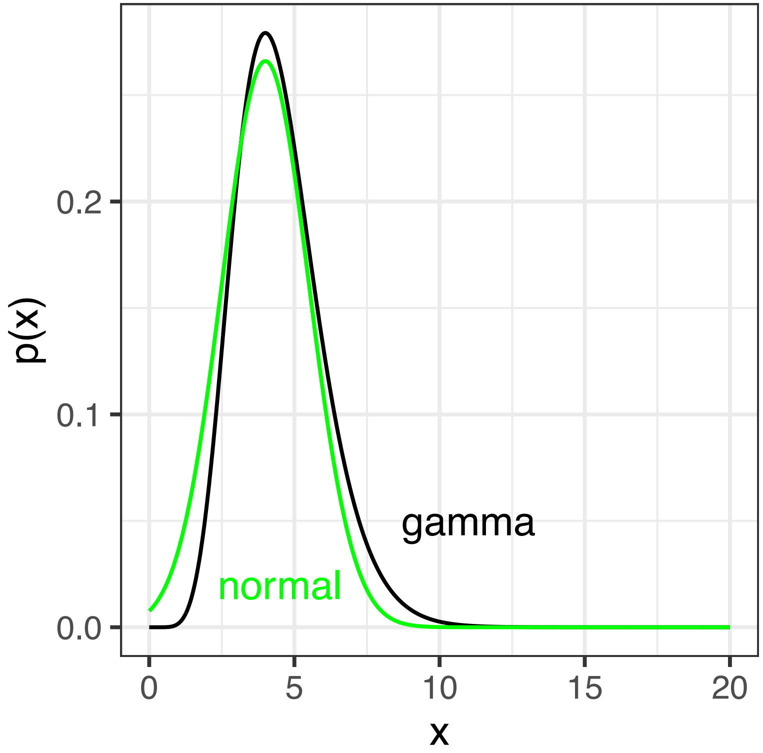


Student's *t*-test

From Wikipedia, the free encyclopedia

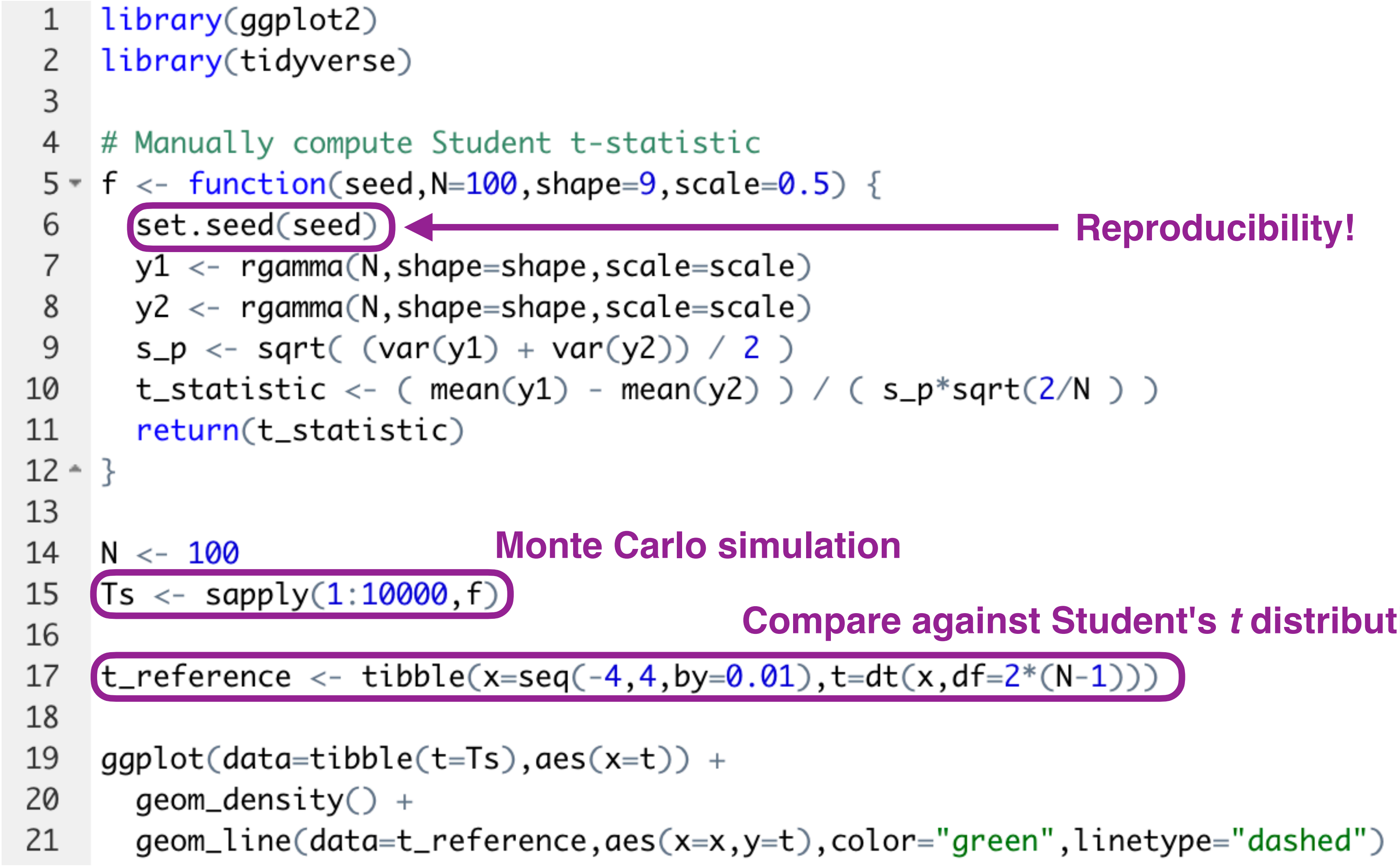
The ***t*-test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis.

A *t*-test is the most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t* distribution. The *t*-test can be used, for example, to determine if the means of two sets of data are significantly different from each other.



- **How bad will this be for my *t*-test????**

Monte Carlo, in action



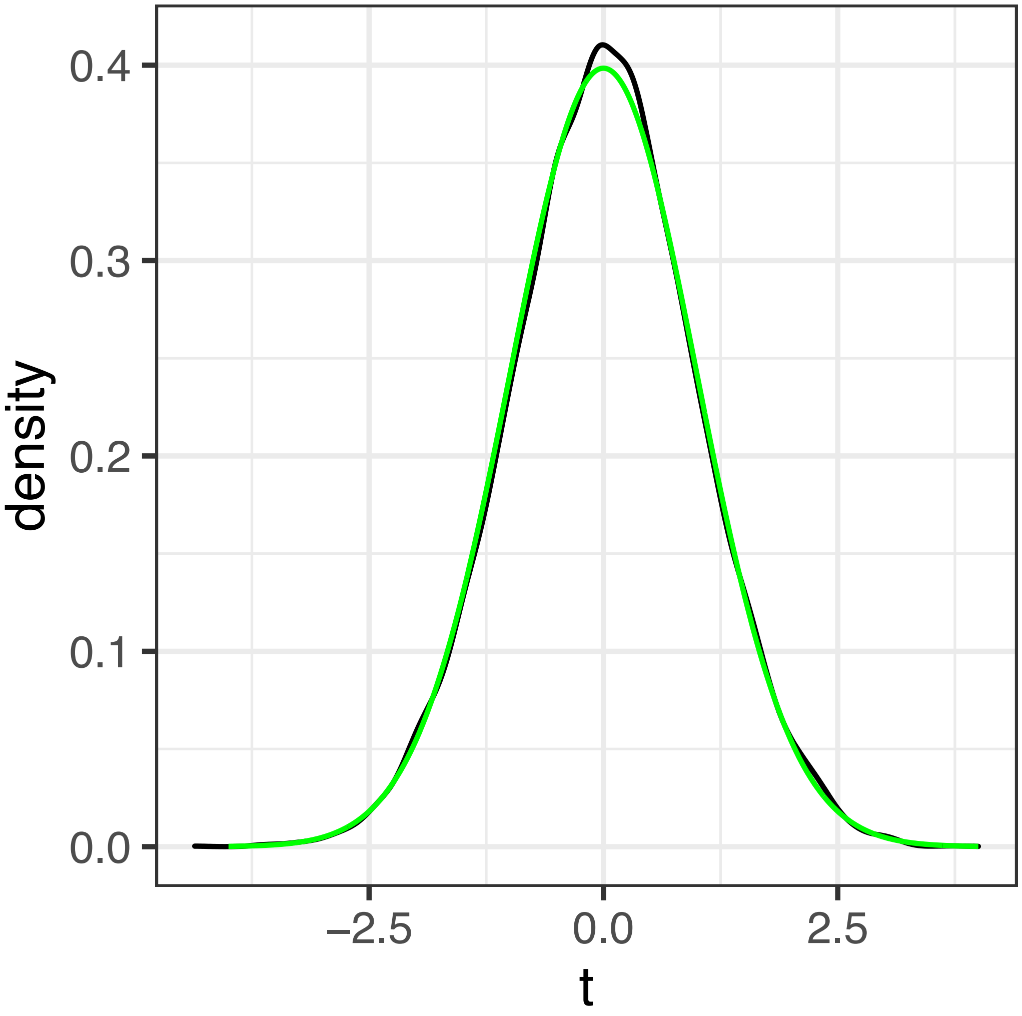
```r
library(ggplot2)
library(tidyverse)

# Manually compute Student t-statistic
f <- function(seed,N=100,shape=9,scale=0.5) {
  set.seed(seed)
  y1 <- rgamma(N,shape=shape,scale=scale)
  y2 <- rgamma(N,shape=shape,scale=scale)
  s_p <- sqrt( (var(y1) + var(y2)) / 2 )
  t_statistic <- ( mean(y1) - mean(y2) ) / ( s_p*sqrt(2/N ) )
  return(t_statistic)
}

N <- 100
Ts <- sapply(1:10000,f)

t_reference <- tibble(x=seq(-4,4,by=0.01),t=dt(x,df=2*(N-1)))

ggplot(data=tibble(t=Ts),aes(x=t)) +
  geom_density() +
  geom_line(data=t_reference,aes(x=x,y=t),color="green",linetype="dashed")
```

Reproducibility!

Monte Carlo simulation

Compare against Student's *t* distribution

Monte Carlo, in action



- The *t* distribution is still a pretty good approximation of the distribution of the *t* statistic, even when the underlying distribution is gamma!
- This exemplifies what is meant when people say that the *t* test is **robust to deviations from normality**

Unnormalizable posteriors

- Our motivation: Bayesian posterior inference

Observed data

$$P(\boxed{\theta} \,|\, \boxed{\mathbf{y}}, I) = \frac{P(\mathbf{y} \,|\, \theta, I)P(\theta \,|\, \boxed{I})}{P(\mathbf{y} \,|\, I)}$$

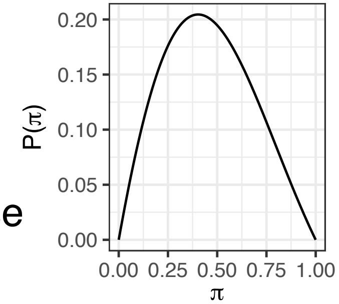
Model parameters

Background knowledge

- Sometimes $P(\mathbf{y} \,|\, I)$ can't be calculated exactly. Example

Bernoulli data with non-conjugate prior:

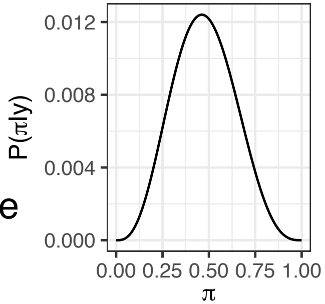
$$P(\pi) \propto \begin{cases} \pi(1-\pi)e^{-\pi^2} & \pi \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$



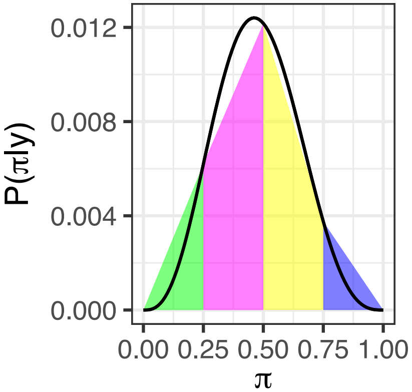
Posterior after observing 2 heads, 2 tails:

$$P(\pi) \propto \begin{cases} \pi^3(1-\pi)^3e^{-\pi^2} & \pi \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

No closed form!



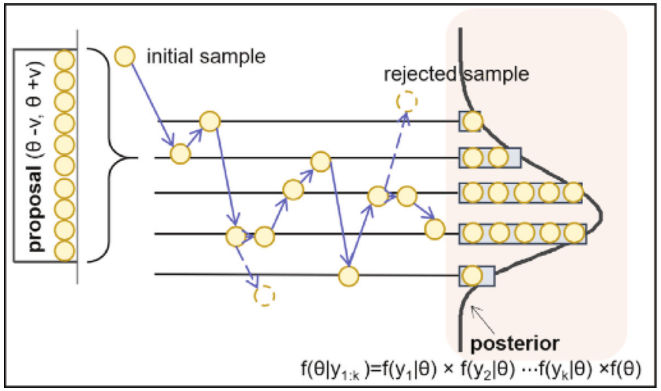
- In simple cases like this, we can numerically approximate the integral:



- But in high dimension and/or unbounded ranges, difficult or even impossible!

Markov chain Monte Carlo

- However, we can often **take samples from the posterior** even when we can't compute normalized probabilities
- One general and widely used approach: **Markov chain Monte Carlo (MCMC)**
- MCMC is a mathematically principled random walk on a non-negative function, directed toward regions where the function takes on a larger value


proposal $(\theta -v, \theta +v)$
initial sample
rejected sample
posterior
$f(\theta|y_{1:k})=f(y_1|\theta) \times f(y_2|\theta) \cdots f(y_k|\theta) \times f(\theta)$

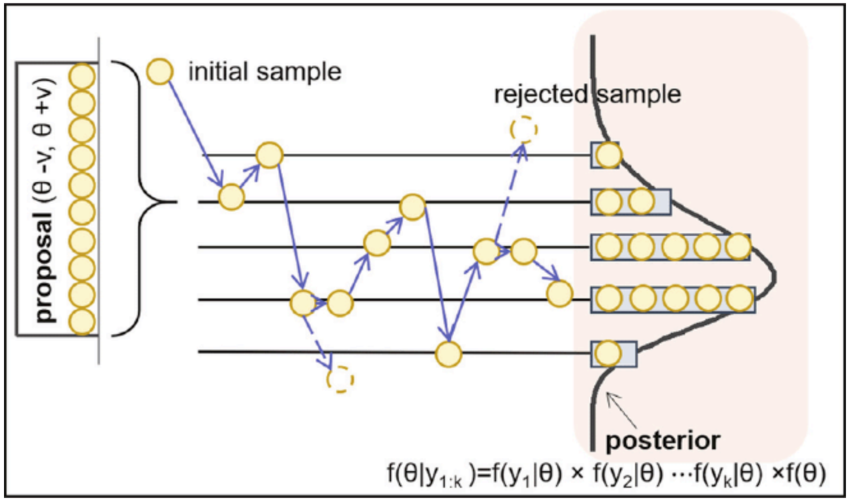
- Asymptotically, the random walk gives us samples from in proportion to the height of the function

Image credit: Dong et al., 2019, http://dx.doi.org/10.5772/intechopen.82781

MCMC for posterior sampling

- We use the *unnormalized* form of the posterior:

$$P(\theta \,|\, \mathbf{y}, I) \propto P(\mathbf{y} \,|\, \theta, I) P(\theta \,|\, I)$$

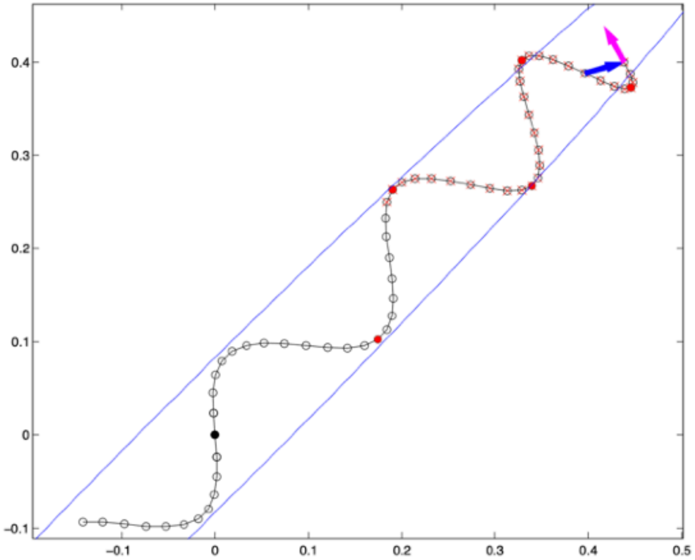

proposal (θ -v, θ +v)
initial sample
rejected sample
posterior
$f(\theta|y_{1:k})=f(y_1|\theta) \times f(y_2|\theta) \cdots f(y_k|\theta) \times f(\theta)$

- We run MCMC and then treat the chain of values as samples from the posterior
- The full set of samples is not iid (nearby values on the chain are correlated), but methods exist for estimating "effectively" how many independent samples we have

Image credit: Dong et al., 2019, http://dx.doi.org/10.5772/intechopen.82781

Stan, HMC, and NUTS

- There are many different MCMC algorithms (e.g., Metropolis, Gibbs Sampling)
- We will use the probabilistic programming language **Stan** for Bayesian inference about model parameters
- Stan uses an algorithm called **Hamiltonian Monte Carlo** (**HMC**) with the **No U-Turn Sampler (NUTS)**

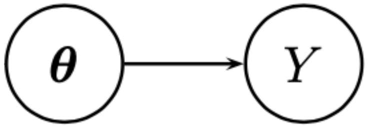


(Hoffman & Gelman, 2014)

- This algorithm tends to be particularly efficient for many problems we'll face

Bayesian posterior inference with Stan

1. Define the generative model you assume underlies the data you want to analyze
2. Choose a prior distribution for your model parameters θ
3. Encode the model structure and prior in a Stan program
4. Provide the data Y you want to analyze, and ask Stan to sample from the posterior $P(\theta | Y, I) \propto P(Y | \theta) P(\theta | I)$ (often written as $P(\theta | Y) \propto P(Y | \theta) P(\theta)$, i.e. eliding I)


$\boldsymbol{\theta}$
Y