Brief review of elementary statistics: parameter estimation, confidence intervals, hypothesis testing

Roger Levy
9.S918: Quantitative Inference in Brain and Cognitive Sciences

18 February 2025

Running example

- I'm about to join a game of betting on the heads/tails outcome of a potentially bent coin
- I can't inspect the coin, but I can watch the coin being flipped "for a while" and record the outcomes
- The coin flips constitute a sequence of **Bernoulli random variables** conditionally independent of each other given the coin weighting $P(\text{heads}) = \pi$ with $0 \leq \pi \leq 1$
- Figuring out from observed data what the weighting is likely to be is **parameter estimation**
- In general, here we will use $\mathbf{y}$ to refer to observed-outcome **data** and θ to refer to the model parameters to be estimated

Characteristics of estimators

- **Estimator:** a procedure for guessing a quantity of interest within a population from a sample from that population
- For example, the **relative frequency estimator:** if we observe r instances of heads in n coin flips,

"this is an estimator"

$$\boxed{\widehat{\ }}\pi = \frac{r}{n}$$

- Data are stochastic, so estimators give random variables!
- **Bias** of an estimator is $E[\widehat{\theta}] - \theta$

$$E[\widehat{\pi}] = E[\frac{r}{n}] = \frac{1}{n}E[r] = \frac{r}{n} = \pi$$ so $\widehat{\pi}$ is **unbiased**

- **Variance** of an estimator is ordinary variance

$$\text{Var}(X) \equiv E[(X - E[X])^2] \qquad \text{Var}(\widehat{\pi}) = \frac{\pi(1-\pi)}{n}$$ (see reading materials)

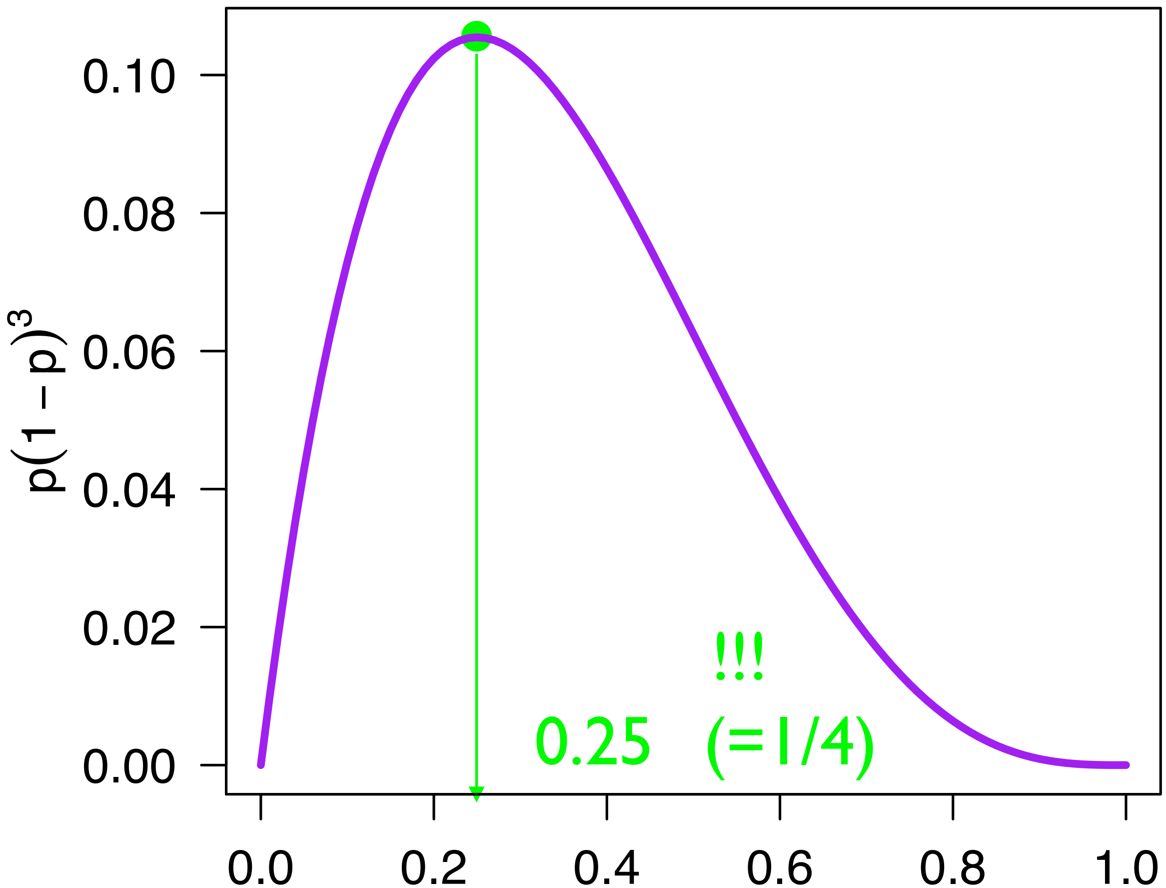
- Good estimators have favorable **bias–variance** tradeoff

Maximum likelihood estimation

$$\text{Lik}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv P(\boldsymbol{y}|\boldsymbol{\theta}) \qquad \hat{\boldsymbol{\theta}}_{MLE} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} \text{Lik}(\boldsymbol{\theta}; \boldsymbol{y})$$

i	y_i
1	T
2	T
3	H
4	T

- *p* refers to the value of P(coin toss$_i$ = Heads)
- Likelihood for the following dataset



This is choosing the *maximum likelihood estimate* (**MLE**)

The **MLE** also turns out to be the *relative frequency estimate* (RFE)

(repeat slide from lecture 3)

The binomial distribution

- The **binomial distribution** is a two-parameter probability distribution over the number of **successes** in a number of independent, identically distributed (**iid**) **Bernoulli trials**
- Two parameters: Number of trials n & trial success parameter π
- A binomial-distributed random variable Y is simply the sum of n iid Bernoulli random variables with success parameter π

 $+$  $+ \ldots +$ 

- A binomial random variable has the following probability mass function:

$$P(Y = r) = \binom{n}{r} \pi^r (1 - \pi)^{n-r}$$

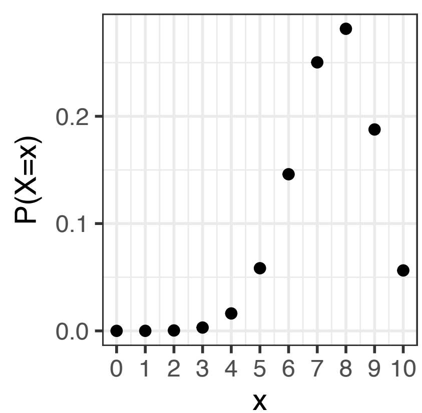


Image credit https://www.wannapik.com/vectors/75713

Bayesian parameter estimation

- Assume that the model parameters "intervene" between background knowledge I and data Y:

$$I \longrightarrow \boldsymbol{\theta} \longrightarrow Y$$

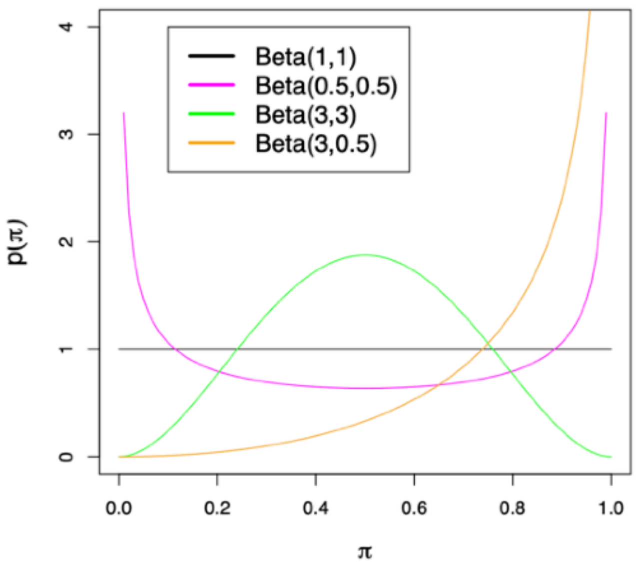
$$P(\boldsymbol{\theta}|\boldsymbol{y}, I) = \frac{P(\boldsymbol{y}|\theta, I)P(\theta|I)}{P(\boldsymbol{y}|I)}$$

$$= \frac{P(\boldsymbol{y}|\boldsymbol{\theta}) \qquad P(\boldsymbol{\theta}|I)}{P(\boldsymbol{y}|I)} \qquad \text{(because } \boldsymbol{y} \perp I \mid \boldsymbol{\theta}\text{)}$$

- Then, if we assume a parametric form for $P(\mathbf{y} \,|\, \theta)$, we just need the prior $P(\theta \,|\, I)$

Example for coin flips: the beta distribution

- Express background knowledge I as two "pseudo-count" parameters α_1, α_2
- The beta distribution has form



$$P(\pi|\alpha_1,\alpha_2) = \boxed{\frac{1}{B(\alpha_1,\alpha_2)}}\boxed{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}$$

Normalizing constant, not of great interest for present purposes

Where the action is!

$$B(\alpha_1,\alpha_2) = \int_0^1 \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}\,d\pi$$

- Cool thing about the beta distribution: the posterior is also beta distributed! For $\mathbf{y} = m$ successes in n trials:

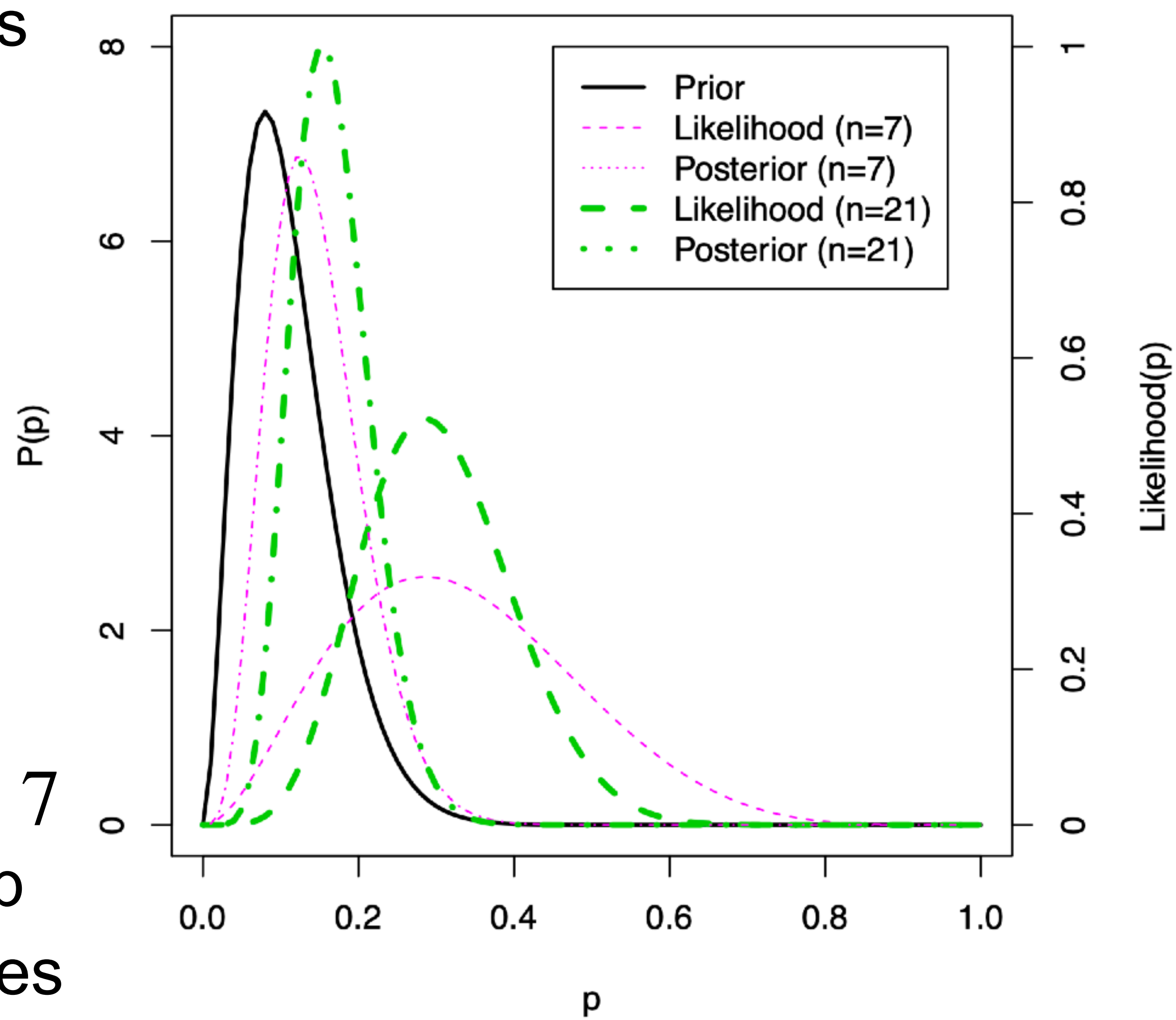
$$P(\pi|\boldsymbol{y},\alpha_1,\alpha_2) \propto \overbrace{\pi^m(1-\pi)^{n-m}}^{\text{Likelihood}}\overbrace{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}^{\text{Prior}}$$

- This property is called **conjugacy** and is convenient where available!

Example of Bayesian parameter estimation

- I inspect my coin and notice serious irregularities!

- My prior for $P(\text{heads})$: a $\alpha_1 = 3, \alpha_2 = 24$ Beta prior
- I flip the coin $n = 7$ times, it comes up heads $m = 2$ times

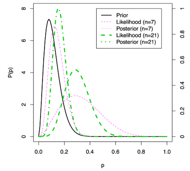


https://commons.wikimedia.org/wiki/File:Post_medieval_coin,_Sixpence_of_William_III_(FindID_605939).jpg

Posterior prediction

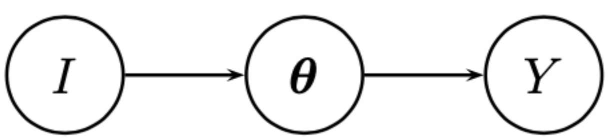
$P(\text{heads}) = \pi$
$P(\pi) = \text{Beta}(\alpha_1, \alpha_2)$
Observe m heads out of n flips

	Beta distribution	Our example
Posterior mean	$E[\pi \mid I] = \dfrac{\alpha_1}{\alpha_1 + \alpha_2}$	$E[\pi \mid y, I] = \dfrac{\alpha_1 + m}{\alpha_1 + \alpha_2 + n}$
Posterior mode (when it exists)	$\dfrac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$	$\dfrac{\alpha_1 + m - 1}{\alpha_1 + \alpha_2 + n - 2}$



Posterior predictive distribution

If I flip the same coin k more times, what is the distribution on the resulting # heads r?



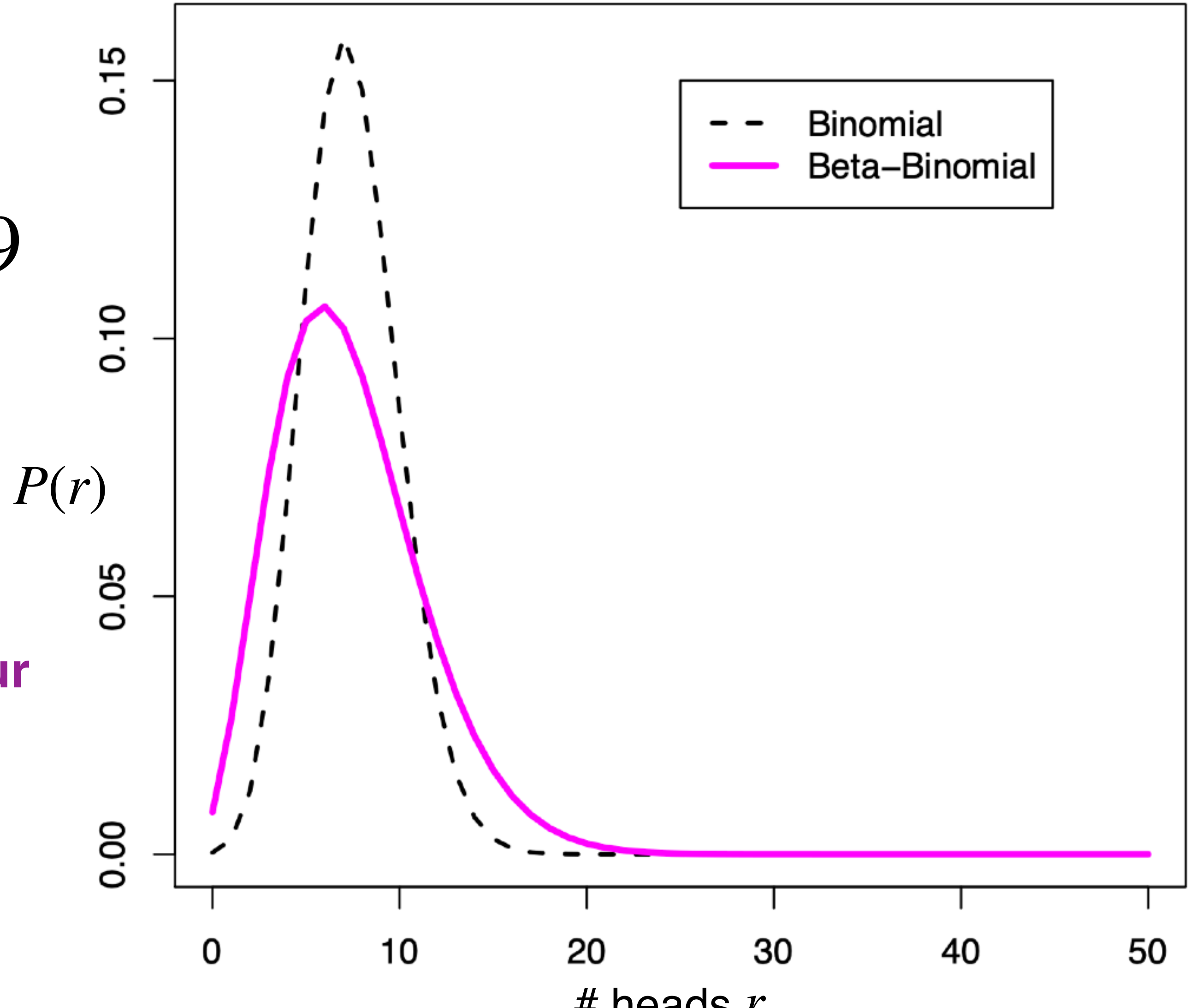
$$P(\boldsymbol{y}_{new} | \boldsymbol{y}, I)$$

→The **Beta-Binomial** model: $P(r|k, I, \boldsymbol{y}) = \binom{k}{r} \dfrac{B(\alpha_1 + m + r, \alpha_2 + n - m + k - r)}{B(\alpha_1 + m, \alpha_2 + n - m)}$

Point estimation vs Bayesian prediction

- Say we'll flip the coin $k = 50$ more times

$$\alpha_1 + m = 5$$
$$\alpha_2 + n - m = 29$$



Bayesian inference takes into account our uncertainty about model parameters θ, leading to more hedged predictions

A note on Bayesian priors

- The Bayesian prior is a double-edged sword
  - We get to specify it
  - We have to specify it
- In the above example, we used an **informative prior**
  - When we have strong domain knowledge, this can potentially be useful in various ways
- In scientific data analysis, however, our general goal is to *allow the data to speak to us about what we care about*
- For this reason, I generally advocate **vague priors for any part of the model whose posterior we care about**
- If your qualitative conclusions depend on choice of prior, it is a reason to be wary of the robustness of your analysis!
- As data become plentiful*, choice of prior *often but not always* recedes in importance

*What counts as "plentiful" depends on size of the model and structure of the data

Credible intervals & confidence intervals

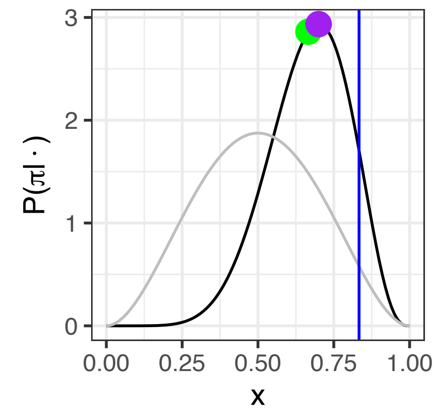
- A **point estimate** of a model parameter is one example of a **statistic**

  (**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

$$\alpha_1 = 3$$
$$\alpha_2 = 3$$
$$r = 5$$
$$n = 6$$



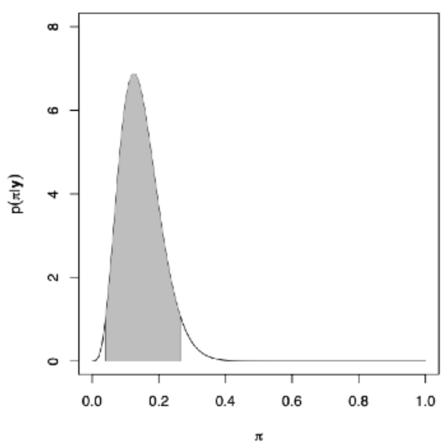
- All of these point estimates discard a lot of information about the shape of the curve that they come from!
  - Curve shape captures **uncertainty** about parameter
- Credible intervals (Bayesian) and confidence intervals (frequentist) provide a bit more information about this uncertainty

Bayesian credible intervals

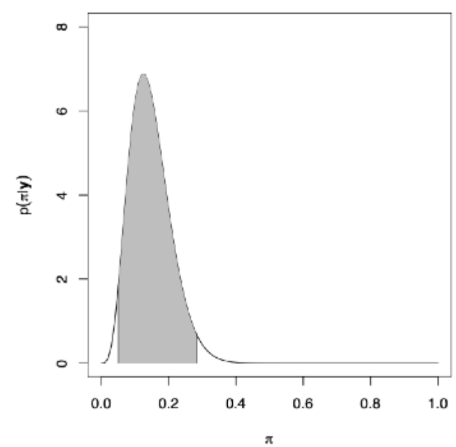
$$P(\theta|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\theta)P(\theta)}{P(\boldsymbol{y})}$$

- 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:

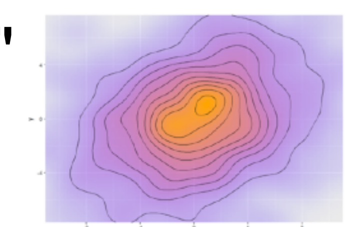
Highest posterior density



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

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

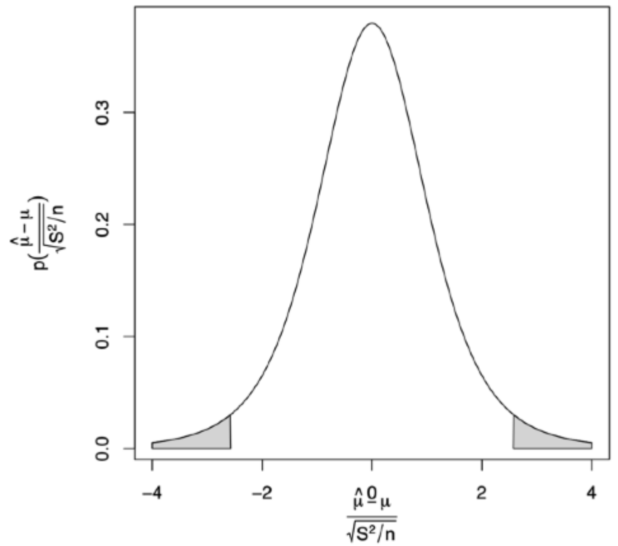
- Suppose I repeat my experiment over and over again, each time collecting data $\mathbf{y}$ and constructing $I = \text{Proc}(\mathbf{y})$
- If $(1-\alpha)$ of these intervals contain the **true value of** θ, then Proc is a method for constructing a $(1-\alpha)$ frequentist confidence interval

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

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

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

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

$$= 1.64$$

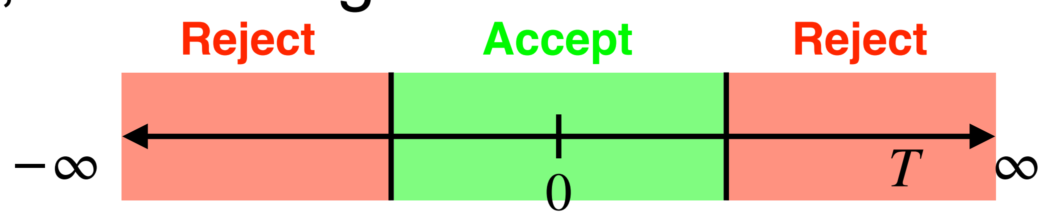
Frequentist hypothesis testing

- 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



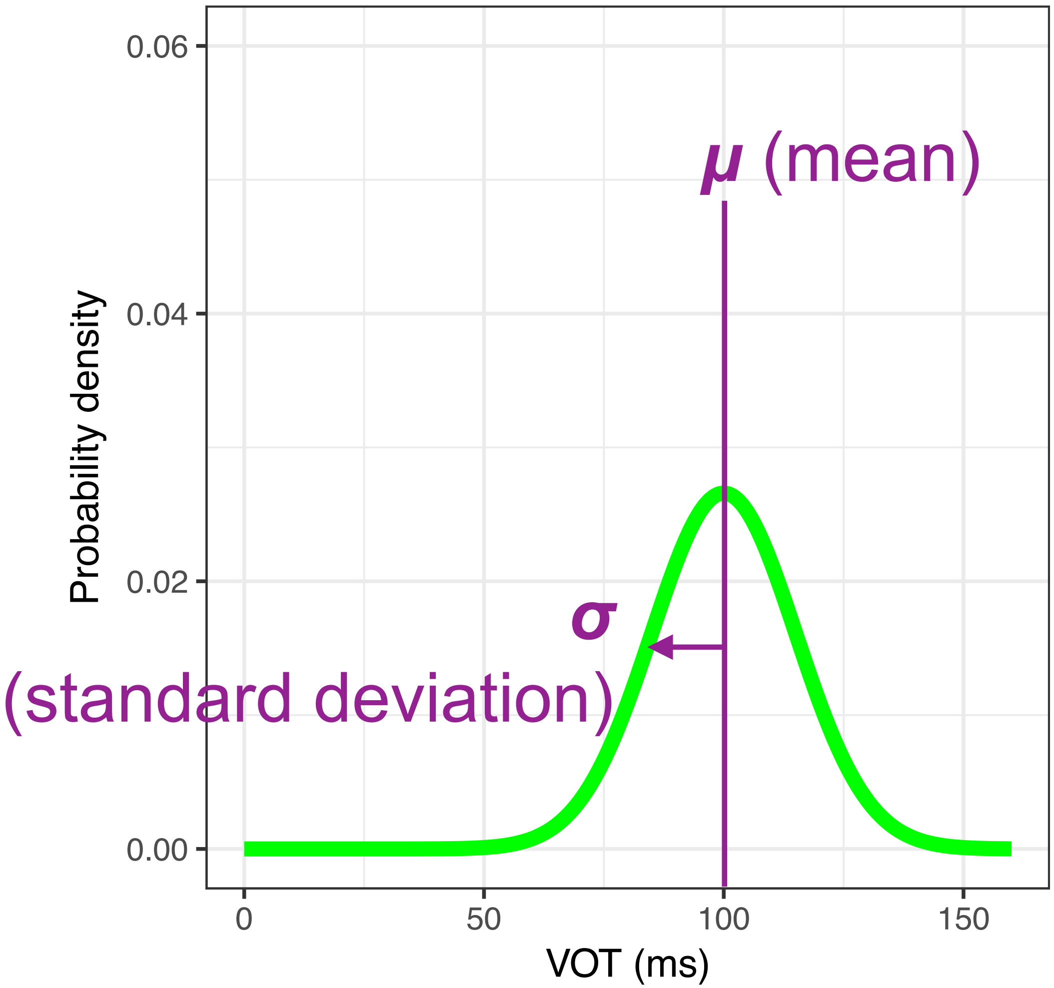
- Collect data, compute T, see where it falls!

Significance level

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

Power

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



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

Squared deviation from mean

Scaled by variance

(normalizing constant)

- Unbiased parameter estimates from a size-N sample:

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

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

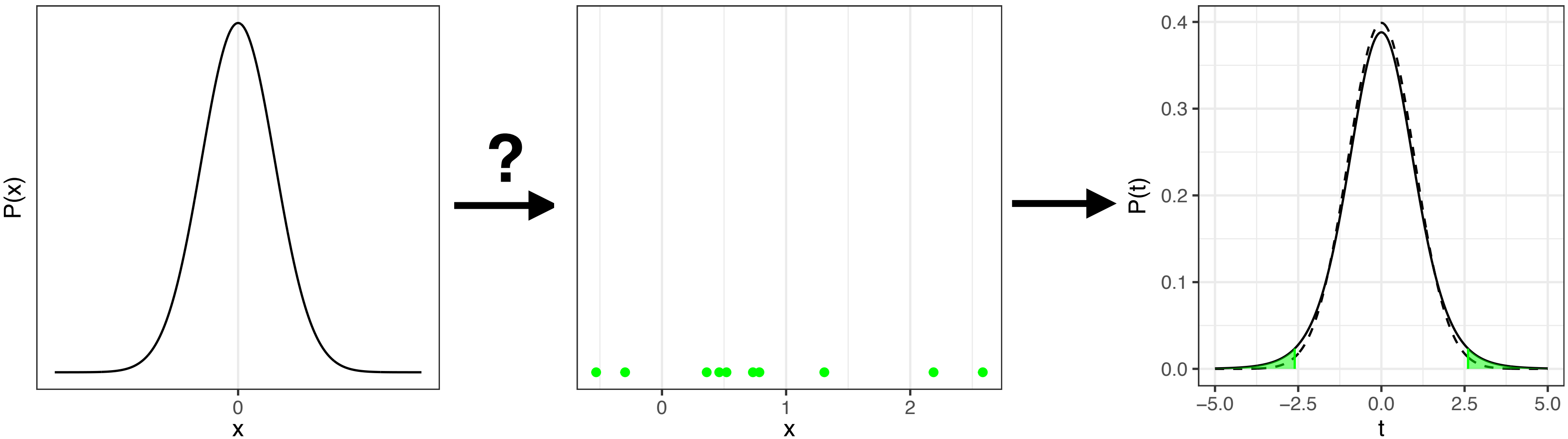
The *t*-test: three variants

- **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: $t = \dfrac{\bar{x}}{s/\sqrt{n}}$

Remember: ***Sample standard deviation***

Two-sample t-test (unpaired)

- **Assumptions:** samples 1 and 2 are each iid normal
- **Null hypothesis H$_0$:** $\mu_1 = \mu_2$
- **Alternative hypothesis H$_1$:** $\mu_1 \neq \mu_2$ (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):

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

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

Pooled sample standard deviation

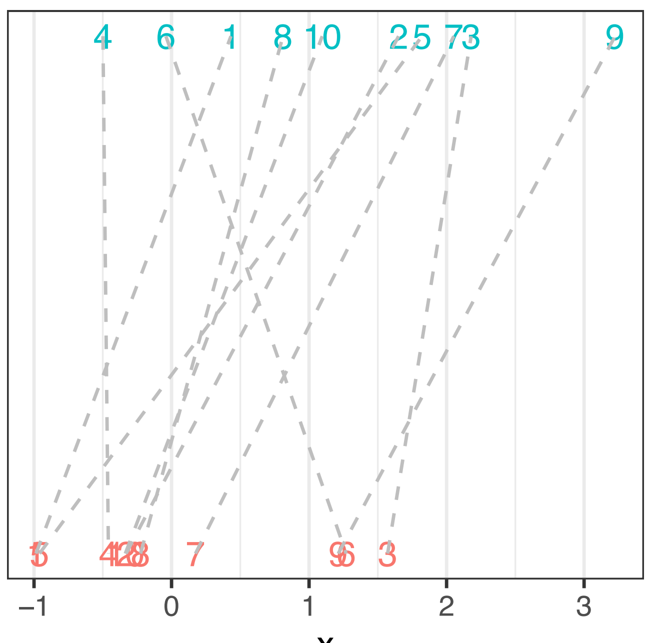
- If we **do not** assume that the two underlying populations have equal variance ("Welch's" t-test):

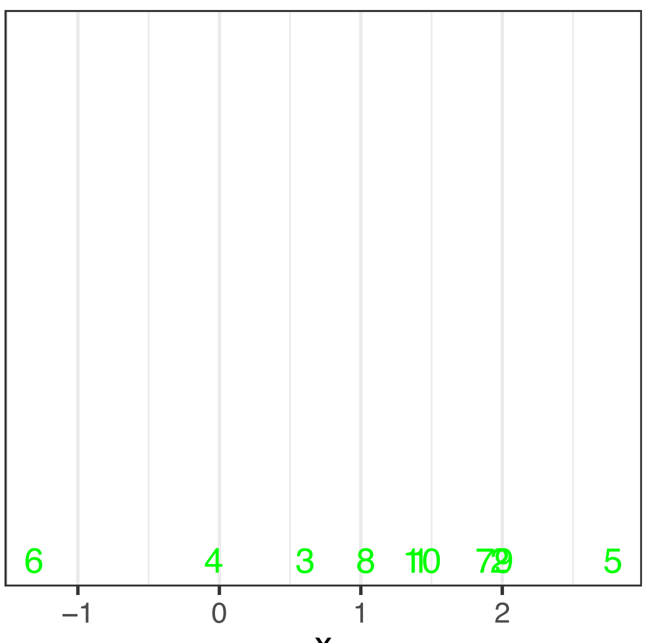
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

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

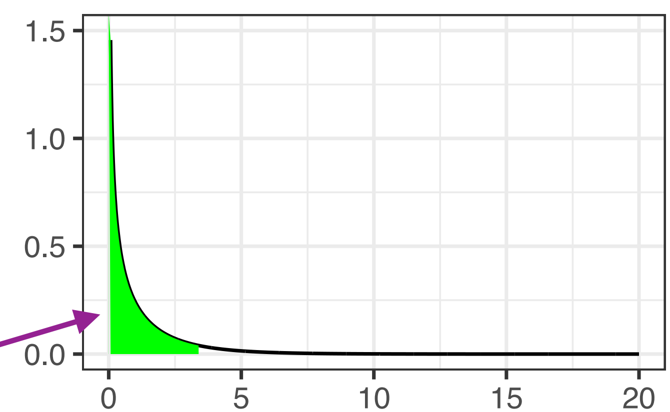
Data likelihood under MLE of H_A

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

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

- **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 $G^2 \rightarrow p = 0.065$