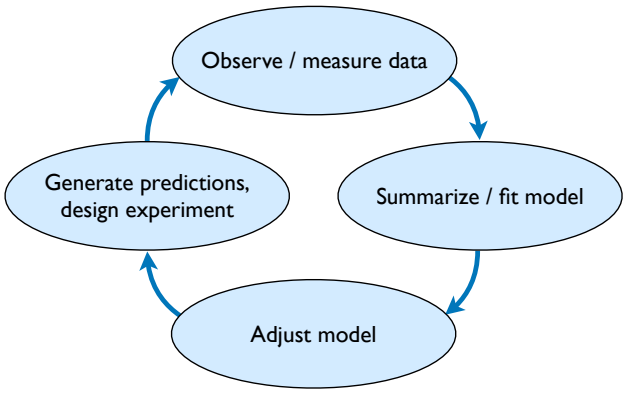
Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2026

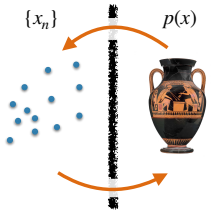
Section 5:
Statistical Inference and Model Fitting

the Scientific process



The sample mean

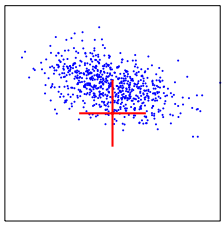
$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$



What happens as N increases?

- Variance of m_x is σ_x^2/N ("standard error of the mean", or SEM), which converges to zero
- "Unbiased": expected value of m_x is true mean, $\mu_x = \mathbb{E}(x)$ and m_x converges to μ_x ("law of large numbers")
- The distribution $p(m_x)$ converges to a **Gaussian** with mean μ_x and variance σ_x^2/N ("central limit theorem")

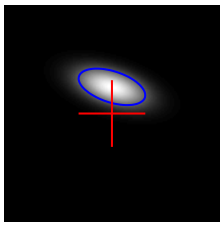
700 samples



Measurement (sampling)

Inference

true density

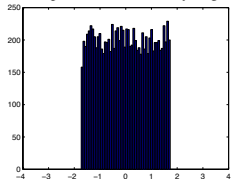


sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

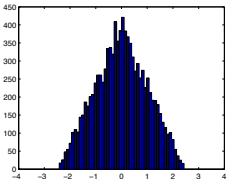
true mean: [0 0.8]
true cov: [1.0 -0.25
-0.25 0.3]

Central limit for a uniform distribution...

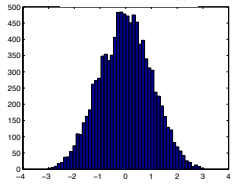
10k samples, uniform density (sigma=1)



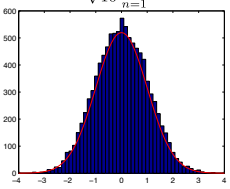
$(u_1 + u_2)/\sqrt{2}$



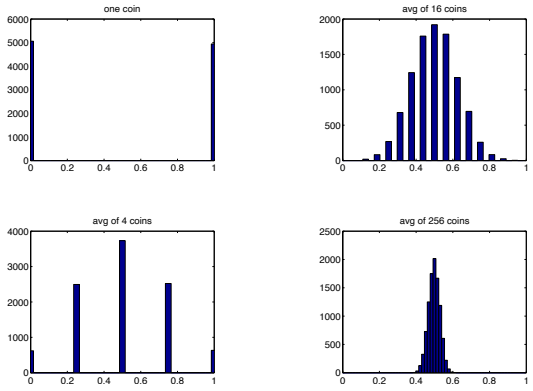
$(u_1 + u_2 + u_3 + u_4)/\sqrt{4}$



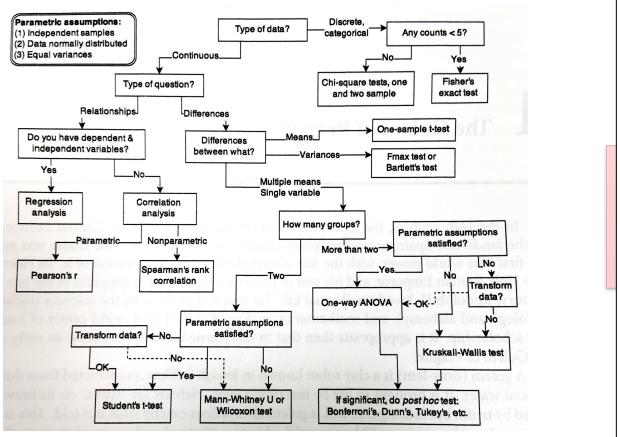
$\frac{1}{\sqrt{10}} \sum_{n=1}^{10} u_n$



Central limit for a binary distribution...



Classical "frequentist" statistical tests



[Statistical Rethinking, Richard McElreath]

Classical/frequentist approach - *z*

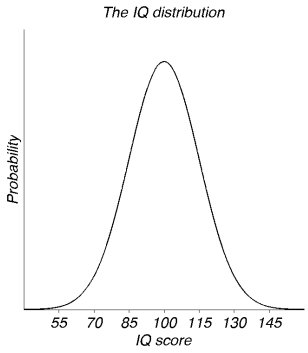
- In the general population, IQ is known to be distributed normally with:

  $\mu = 100,\ \sigma = 15$

- We give a drug to 30 people and test their IQ
- Hypotheses:

  H_1: NZT improves IQ

  H_0 ("null"): it does nothing



Test statistic

- We calculate how far the observed value of the sample average is away from its expected value.
- In units of standard error.
- In this case, the test statistic is

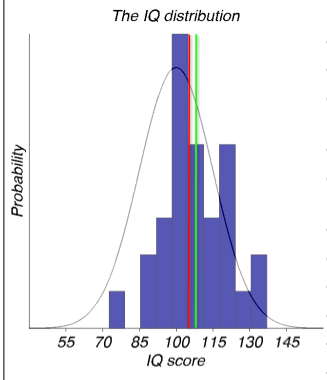
$$z = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

- Compare to a distribution, in this case z or $N(0,1)$

Does NZT improve IQ scores or not?

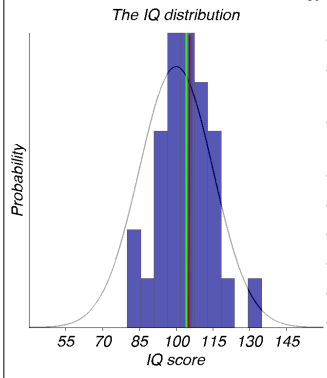
Decision \ Reality	Yes	No
Yes	Correct	Type I error α-error "False alarm"
No	Type II error β-error "Miss"	Correct

The *z*-test



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 108.3$ (Sample mean)
- $z = (\bar{x} - \mu)/\text{SE} = (108.3\text{-}100)/\text{SE}$ (Standardized score)
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- Error bar/CI: ±2 SE
- $z = 8.3/2.74 = 3.03$
- $p = 0.0012$
- Significant?
- One- vs. two-tailed test

What if the measured effect of NZT had been half that?



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 104.2$ (Sample mean)
- $z = (\bar{x} - \mu)/\text{SE} = (104.2\text{-}100)/\text{SE}$
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- $z = 4.2/2.74 = 1.53$
- $p = 0.061$
- Significant?

Significance levels

- Are denoted by the Greek letter α.
- In principle, we can pick anything that we consider unlikely.
- In practice, the consensus is that a level of 0.05 or 1 in 20 is considered as unlikely enough to reject H_0 and accept the alternative.
- A level of 0.01 or 1 in 100 is considered "highly significant" or "really unlikely".

Common misconceptions



Is "Statistically significant" a synonym for:
- Substantial
- Important
- Big
- Real

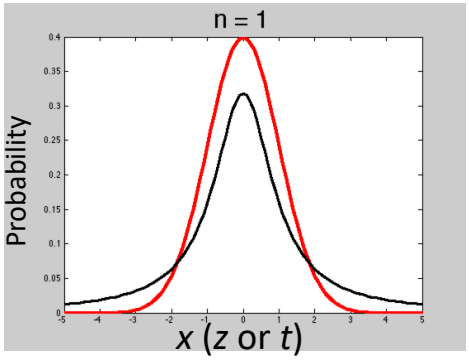
Does statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

Meaning of *p*-value. Meaning of CI.

Student's *t*-test

- σ not assumed known
- Use $$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
- Why N-1? s is unbiased (unlike ML version), i.e., $\mathbb{E}(s^2) = \sigma^2$
- Test statistic is $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$
- Compare to t distribution for CIs and NHST
- "Degrees of freedom" reduced by 1 to N-1

The *t* distribution approaches the normal distribution for large *N*



The *z*-test for binomial data

- Is the coin fair?
- Lean on central limit theorem
- Sample is *n* heads out of *m* tosses
- Sample mean: $\hat{p} = n / m$
- H_0: $p = 0.5$
- Binomial variability (one toss): $\sigma = \sqrt{pq}$, where $q = 1 - p$
- Test statistic: $$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / m}}$$
- Compare to *z* (standard normal)
- For CI, use $$\pm z_{\alpha/2} \sqrt{\hat{p}\hat{q} / m}$$

Other frequentist univariate tests

- χ^2 goodness of fit
- χ^2 test of independence
- test a variance using χ^2
- *F* to compare variances (as a ratio)
- Nonparametric tests (e.g., sign, rank-order, etc.)

Estimation of model parameters (outline)

- How do I estimate parameters from data?
- How "good" are my estimated parameters?
- How well does my model explain data to which it was fit? Other data (prediction/generalization)?
- How do I compare two models?

Estimation

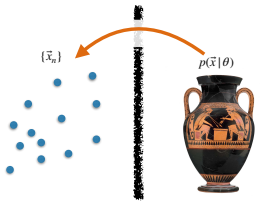
- An "estimator" is a function of the data, intended to provide an approximation of the "true" value of a parameter
- One can evaluate estimator quality in terms of squared error, MSE = bias^2 + variance
- Traditional statistics often aims for an unbiased estimator, with minimal variance ("MVUE")
- More nuanced view: trade off bias and variance, through model selection, "regularization", or Bayesian "priors" …

The maximum likelihood estimator (MLE)

Sample average is appropriate when one has direct measurements of the thing being estimated. But one may want to estimate something that is *indirectly* related to the measurements…

Natural choice: assuming a probability model $p(\vec{x}\,|\,\theta)$
find the value of θ that maximizes this "likelihood" function

$$\hat{\theta}(\{\vec{x}_n\}) = \arg\max_\theta \prod_n p(\vec{x}_n|\theta)$$
$$= \arg\max_\theta \sum_n \log p(\vec{x}_n|\theta)$$



Example: Estimate the true probability of a flipped coin landing “heads” up, by observing some samples



66% ?



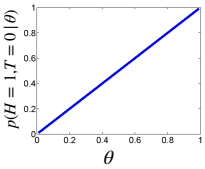
Example ML Estimators - discrete

Binomial: $p(H \mid N, \theta) = \binom{N}{H} \theta^{H} (1-\theta)^{N-H}$

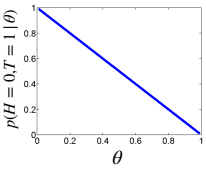
(H = # heads observed, in N flips of a coin, with probability of heads θ)

$$\hat{\theta}_{\text{ML}} = \frac{H}{N}$$

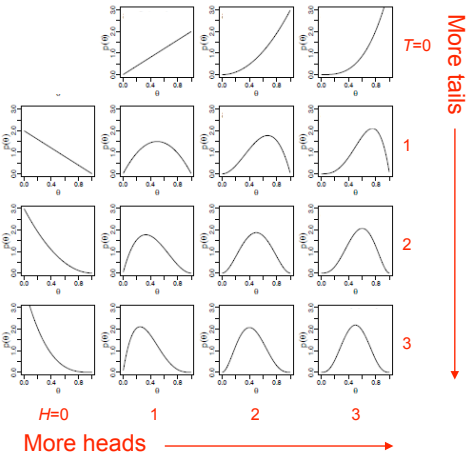
Likelihood: 1 head observed



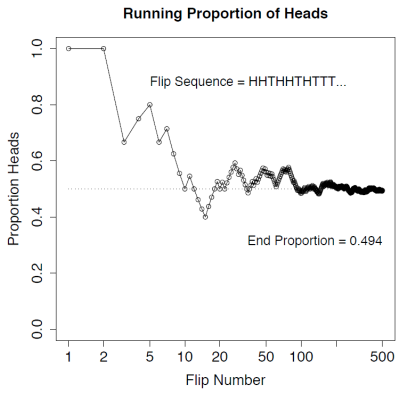
Likelihood: 1 tail observed



Likelihoods, $p(H\,|\,N,\theta)$



Convergence ("consistency")



Example ML Estimators - discrete

Binomial: $p(H \mid N,\theta) = \binom{N}{H} \theta^{H} (1-\theta)^{N-H}$

(H = # heads observed, in N flips of a coin, with probability of heads θ)

$$\hat{\theta}_{\text{ML}} = \frac{H}{N}$$

Poisson: $p(\{k_n\} \mid \theta) = \prod_{n=1}^{N} \frac{\theta^{k_n} e^{-\theta}}{k_n!}$

(k's are measured counts, with mean arrival rate of θ)

$$\hat{\theta}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} k_n$$

[on board]

Example ML Estimators - continuous

Uniform: $p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$

$$\hat{\theta}_{\text{ML}} = \max_n \{x_n\}$$

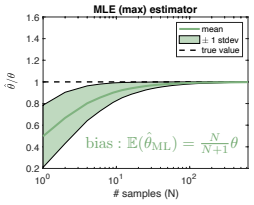
(Note: this is biased!)

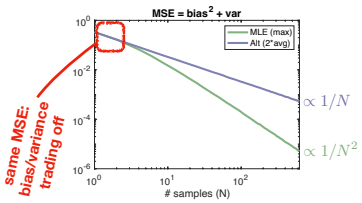
Two estimators for range of a uniform distribution

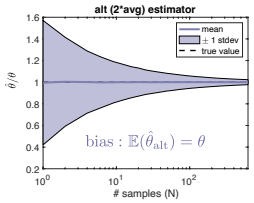
Given N samples $\{x_n\}$ from the uniform distribution over $[0, \theta]$, consider two estimators of θ:

$$\hat{\theta}_{\text{ML}}(\{x_n\}) = \max_n (x_n)$$

$$\hat{\theta}_{\text{alt}}(\{x_n\}) = \frac{2}{N} \sum_n x_n$$







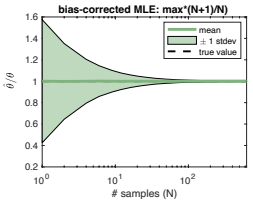
Two estimators for range of a uniform distribution

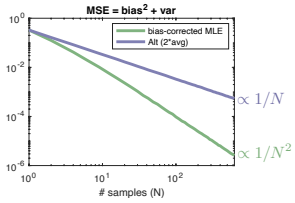
Given N samples $\{x_n\}$ from the uniform distribution over $[0, \theta]$ consider two estimators of θ:

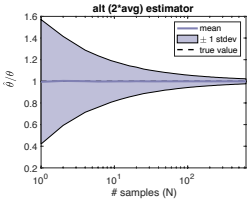
$$\hat{\theta}_{\text{cML}}(\{x_n\}) = \frac{N+1}{N} \max_n(x_n)$$

bias correction

$$\hat{\theta}_{\text{alt}}(\{x_n\}) = \frac{2}{N} \sum_n x_n$$







Example ML Estimators - Continuous

Uniform: $p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$

$$\hat{\theta}_{\text{ML}} = \max_n \{x_n\}$$

(Note: this is biased!)

$$\hat{\theta}_{\text{cML}} = \frac{N+1}{N} \hat{\theta}_{\text{ML}}$$

Gaussian: $p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\hat{\mu}_{\text{ML}} = \frac{\sum_n x_n}{N}$$

(sample average, again)

$$\hat{\sigma}^2_{\text{ML}} = \frac{\sum_n (x_n - \hat{\mu})^2}{N}$$

(Note: this is biased!)

$$\hat{\sigma}^2_{\text{cML}} = \frac{N}{N-1} \hat{\sigma}^2_{\text{ML}}$$

[on board]

Summarizing errors of ML estimators

Bias: the MLE is *asymptotically unbiased* and *Gaussian*, but can only rely on these if:

- the likelihood model is correct
- the likelihood can be maximized
- you have lots of data

Variance: (error bars)

- S.E.M. (relevant for sample averages only)
- second deriv of NLL (multi-D: "Hessian")
- simulation (resample from $p(x|\hat{\theta})$)
- bootstrapping (resample from *the data,* with replacement)

Bootstrapping

- "The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps"
  [Adventures of Baron von Munchausen, by Rudolph Erich Raspe]
- A (**re)sampling** method for computing estimator **dispersion** (eg., stdev or confidence intervals)
- Idea: instead of looking at distribution of estimates across repeated experiments, look across repeated resamplings (with replacement) from the *existing* data ("bootstrapped" data sets)

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

[New York Times, 27 Jan 1987]

The summary statistics in the newspaper article are very simple:

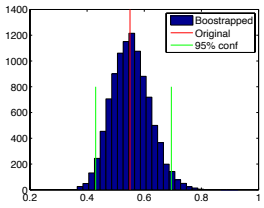
	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11037
placebo group:	189	11034

$$\widehat{\theta} = \frac{104/11037}{189/11034} = .55. \quad (1.1)$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Of course we are not really interested in $\widehat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio

Histogram of bootstrap estimates:



=> with 95% confidence,

$0.43 < \theta < 0.7$

[Efron & Tibshirani '98]

	strokes	subjects	
aspirin group:	119	11037	
placebo group:	98	11034	(1.3)

For strokes, the ratio of rates is

$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1.21. \quad (1.4)$$

It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be

$$.93 < \theta < 1.59 \quad (1.5)$$

with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo vis-à-vis strokes. In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.

[Efron & Tibshirani '98]

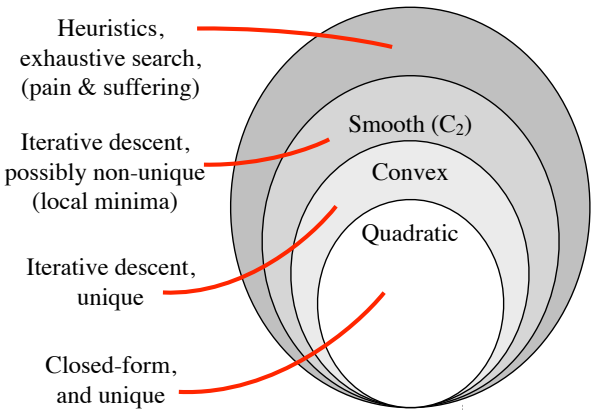
Permutation test

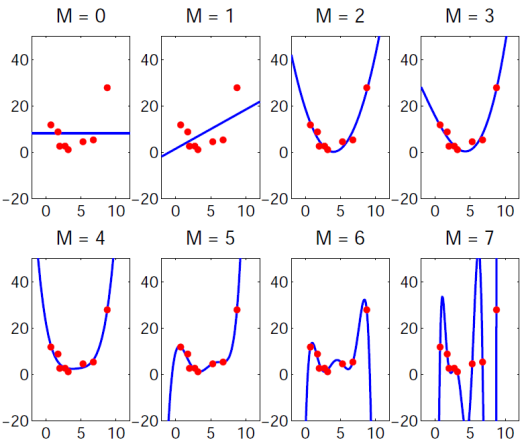
- Given {n1,n2} measurements under two different conditions, are they significantly different (i.e., can we reject null hypothesis?)
- Measure difference in means, m2-m1
- Construct permuted sets of {n1,n2} measurements, and compute difference in means for each of these
- Ask: How far in the tail is the true difference in means? One-sided p-value is proportion of permutation values > m2-m1

Taxonomy of model-fitting errors

- Unexplainable variability (e.g., due to noisy measurements)
- Optimization failures (e.g., local minima)
- Overfitting (too many params, not enough data)
- Model failures (what you'd really like to know)

Optimization…





How do we avoid **overfitting** (i.e., concluding that M=7 is "best")?

Model Comparison

- If models are optimized according to some objective, it is natural to compare them based on the value of that objective…
  - for least squares regression, compare the residual squared error of two models (with different regressors).
  - for ML estimates, compute the likelihood (or log likelihood) ratio, and compare to 1 (or zero).
- **Problem**: evaluating the objective with the same data used to optimize the model leads to **over-fitting**!

Cross-validation

A resampling method for estimating predictive error of a model. Widely used to identify/avoid over-fitting, and to provide a fair comparison of models.

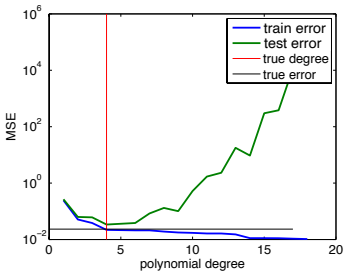
(1) Randomly partition data into a "training" set, and a "test" set.

(2) Fit model to training set. Measure error on test set.

(3) Repeat (many times).

(4) Choose model that minimizes the average "cross-validated" (**test**) error

Example: Using cross-validation to select the degree of a polynomial model:

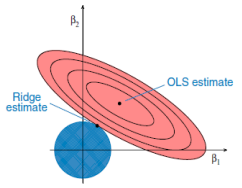


Ridge regression

(a.k.a. L_2 regularization)

Ordinary least squares regression:

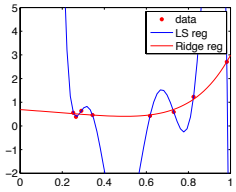
$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2$$



"Regularized" least squares regression:

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda||\vec{\beta}||^2$$

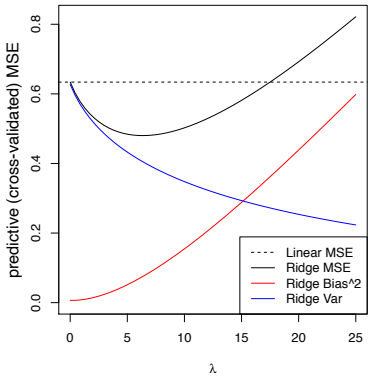
7th-order polynomial regression:



$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

Choose λ by cross-validation:

Ridge Regression trades off bias and variance:



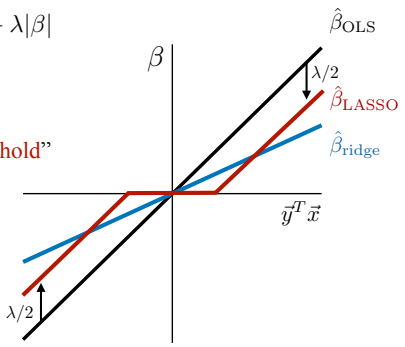
[from http://www.stat.cmu.edu/~ryantibs/datamining/]

LASSO regularization

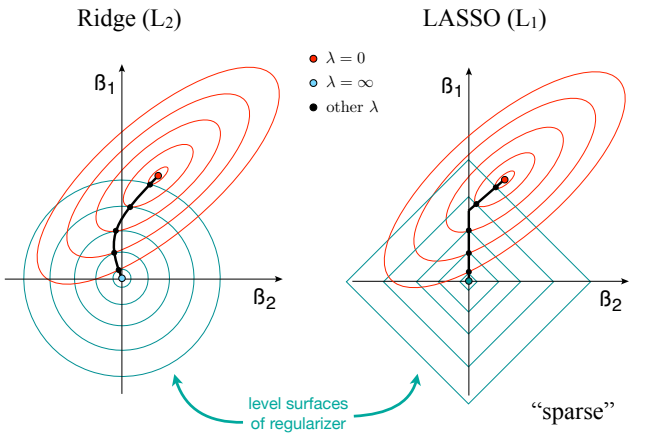
("least absolute shrinkage and selection operator",
a.k.a. L_1 regularization)

$$\arg\min_{\beta} ||\vec{y} - \vec{x}\beta||^2 + \lambda|\beta|$$

Assuming $\|\vec{x}\|^2 = 1$,
solution is a "soft-threshold"
on $\vec{y}^T\vec{x}$



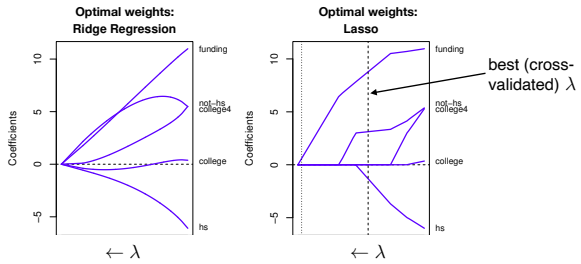
Ridge regression vs. LASSO geometry (in 2D)



LASSO vs. Ridge regression

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

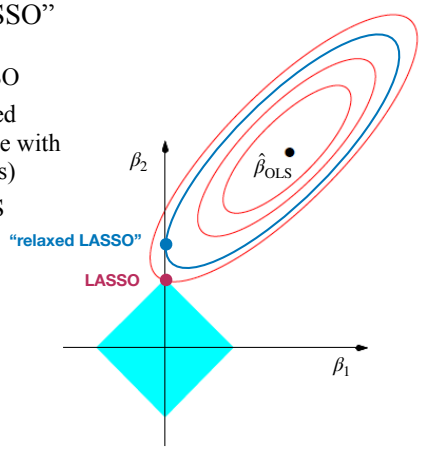
city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940



[From Hastie, Tibshirani, Wainwright 2015]

The "Relaxed LASSO"

1. Solve for LASSO
2. Eliminate unused regressors (those with zero coefficients)
3. Re-fit (using LS regression)



Bayesian Inference

"Posterior" "Likelihood" "Prior"

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)\, p(\theta)}{p(\text{data})}$$

Normalization factor

Inference/estimation should be expressed in terms of the **posterior**

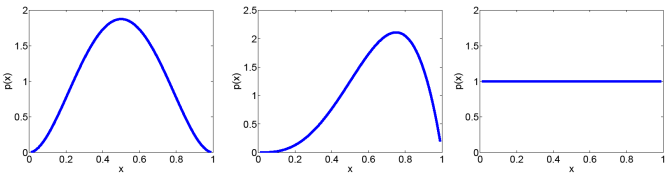
Example: Posterior for coin

Infer whether a coin is fair by flipping it repeatedly
Here, x is the probability of heads (50% is fair)
$y_{1\ldots n}$ are the outcomes of flips

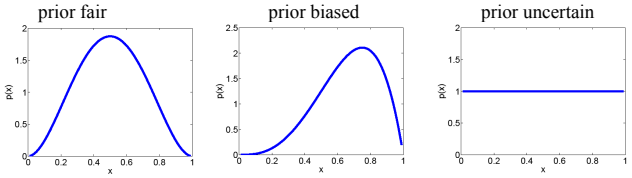


Consider three different priors:

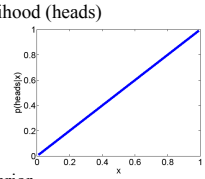
suspect fair | suspect biased | no idea



prior fair | prior biased | prior uncertain

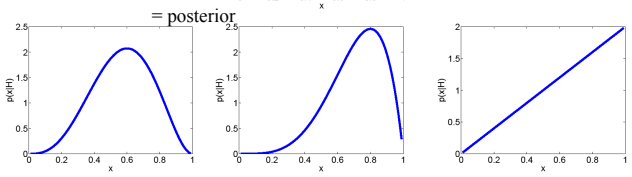


x likelihood (heads)

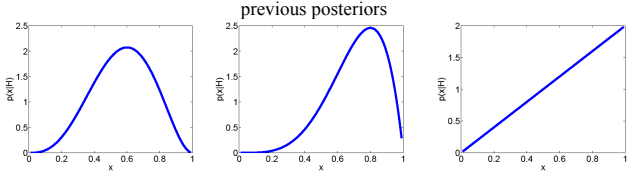




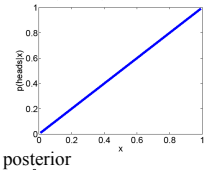
= posterior



previous posteriors

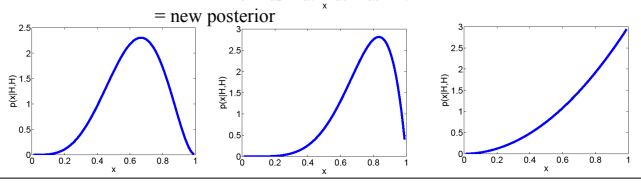


X likelihood (heads)

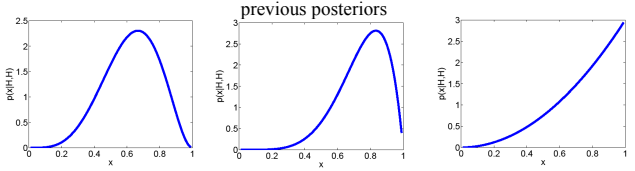




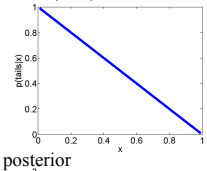
= new posterior



previous posteriors

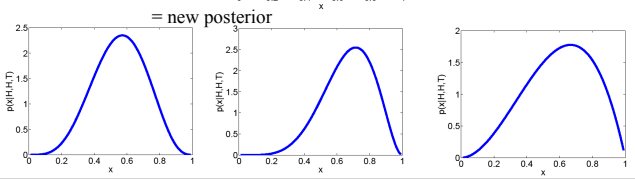


X likelihood (tails)

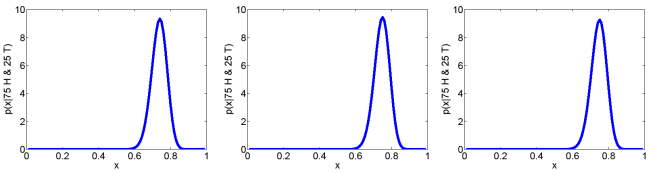




= new posterior



Posteriors after observing 75 heads, 25 tails



→prior differences are ultimately overwhelmed by data

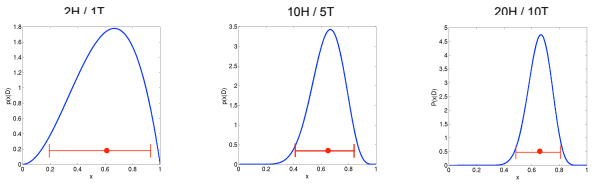
Bayesian estimators

Summarize **central tendency** of posterior:

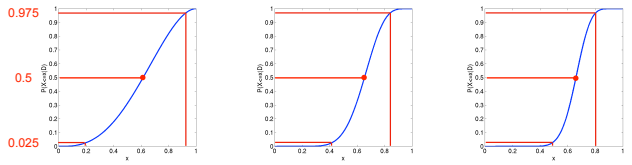
- Posterior mode ("maximum *a posteriori*" estimate - MAP)
- Posterior mean (minimizes squared error - MMSE)
  Summarize dispersion with posterior variance
- Posterior median (minimizes abs error)
  Summarize dispersion with posterior quantiles [next slide]

Bayesian (posterior) confidence (credibility) intervals

PDFs



CDFs, median and 95% confidence intervals



Bayesian inference: Gaussian case

For measurements with Gaussian noise, and assuming a Gaussian prior:

- posterior is Gaussian, allowing sequential updating
- precision is sum of measurement and prior precisions
- mean is precision-weighted average of prior mean and measurement (as in Ridge Regression!)
- => Can explain "regression to the mean" as **shrinkage** toward the prior

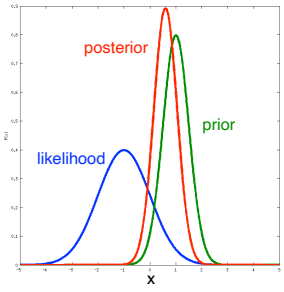
Bayesian inference: Gaussian case

$$y = x + z, \quad x \sim \mathcal{N}(\mu_x, \sigma_x^2), \quad z \sim \mathcal{N}(0, \sigma_z^2)$$

$$p(x|y) \propto p(y|x)\, p(x)$$

$$= e^{-\frac{1}{2}\left[\frac{1}{\sigma_z^2}(x-y)^2\right]} e^{-\frac{1}{2}\left[\frac{1}{\sigma_x^2}(x-\mu_x)^2\right]}$$

$$= e^{-\frac{1}{2}\left[(\frac{1}{\sigma_z^2}+\frac{1}{\sigma_x^2})x^2 - 2(\frac{y}{\sigma_z^2}+\frac{\mu_x}{\sigma_x^2})x + \ldots\right]}$$



Posterior is Gaussian, with:

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_z^2} + \frac{1}{\sigma_x^2} \qquad \text{(sum of precisions)}$$

$$\mu = \left(\frac{y}{\sigma_z^2} + \frac{\mu_x}{\sigma_x^2}\right) \Big/ \left(\frac{1}{\sigma_z^2} + \frac{1}{\sigma_x^2}\right)$$

(precision-weighted average of y and μ_x)

Regression to the mean

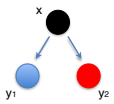
"Depressed children treated with an energy drink improve significantly over a three-month period."

"I made up this newspaper headline, but the fact it reports is true: if you treated a group of depressed children for some time with an energy drink, they would show a clinically significant improvement…."

"It is also the case that depressed children who spend some time standing on their head or hug a cat for twenty minutes a day will also show improvement."

- Daniel Kahneman

Two noisy measurements of the same variable:



$$y_1 = x + z_1 \qquad x \sim \mathcal{N}(0, \sigma_x^2)$$

$$y_2 = x + z_2 \qquad z_k \sim \mathcal{N}(0, \sigma_z^2), \quad \text{independent}$$

$$C_y = \begin{bmatrix} \sigma_x^2 + \sigma_z^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_z^2 \end{bmatrix}$$

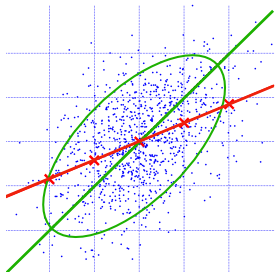
LS Regression:

$$\hat{\beta} = \arg\min_{\beta} \mathbb{E}\left[\|y_2 - \beta y_1\|^2\right]$$

$$= \frac{\mathbb{E}[\,y_1\, y_2\,]}{\mathbb{E}[\,y_1^2\,]} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_z^2}$$

$$\mathbb{E}[\,y_2|y_1\,] = \hat{\beta}\, y_1$$

"regression to the mean"



Nested set of common statistical estimators

- Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_x p(\vec{d}|x)$
  (requires likelihood, $p(\vec{d}|x)$)
- Maximum a posteriori (MAP): $\hat{x}(\vec{d}) = \arg\max_x p(x|\vec{d})$
  (requires prior, $p(x)$)
- Bayes estimator (general): $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbb{E}\left(L(x,\hat{x}) \,\middle|\, \vec{d}\right)$
  (requires loss, $L(x,\hat{x})$)
- Bayes least squares: $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbb{E}\left((x-\hat{x})^2 \,\middle|\, \vec{d}\right)$
  (special case: squared loss) $= \mathbb{E}\left(x \,\middle|\, \vec{d}\right)$

Bayesian Model Comparison

- Eg: Is the coin fair? Compared to what?
- Consider two models: $M_1 : p = 0.5 \quad M_2 : p = 0.6$

$$p(M_k \mid D) = \frac{p(D \mid M_k) P(M_k)}{p(D)}$$

Compare their *posterior ratio*:

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1) P(M_1)}{p(D \mid M_2) P(M_2)}$$

Comparing models' predictive performance

Option 1: Include a penalty for number of parameters:

For an ML estimate: $\hat{\theta} = \arg\min_\theta \left[-\ln p(\vec{d}|\theta)\right]$

a. Akaike information criterion (AIC) [Akaike, 1974]

$$E_{\mathrm{AIC}}(\vec{d},\hat{\theta}) = 2\ \dim(\hat{\theta}) - 2\ln p(\vec{d}|\hat{\theta})$$

b. Bayesian information criterion (BIC) [Schwartz, 1978]

$$E_{\mathrm{BIC}}(\vec{d},\hat{\theta}) = \dim(\hat{\theta})\ \ln\left[\dim(\vec{d})\right] - 2\ln p(\vec{d}|\hat{\theta})$$

valid when $\dim(\vec{d}) \gg \dim(\hat{\theta})$

Option 2: Cross-validation
(evaluate generalization to held-out data)