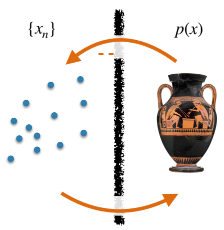
Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2019

Section 5:
Statistical Inference and Model Fitting

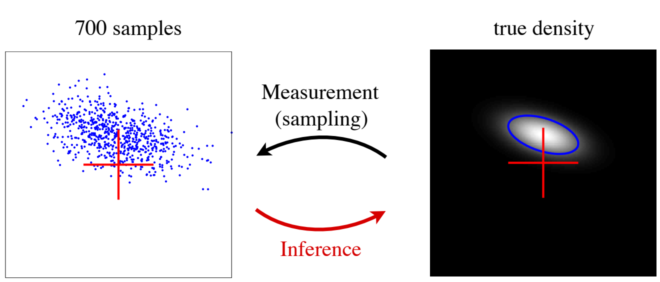
The sample average



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

What happens as N grows?

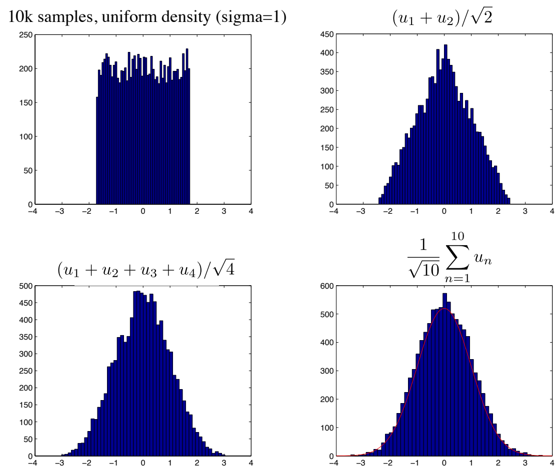
- Variance of $\bar{x}$ is σ_x^2/N (the "standard error of the mean", or SEM), and so converges to zero *[on board]*
- "Unbiased": $\bar{x}$ converges to the true mean, $\mu_x = \mathbb{E}(x)$ (formally, the "law of large numbers") *[on board]*
- The distribution $p(\bar{x})$ converges to a Gaussian (mean μ_x and variance σ_x^2/N): formally, the "Central Limit Theorem"



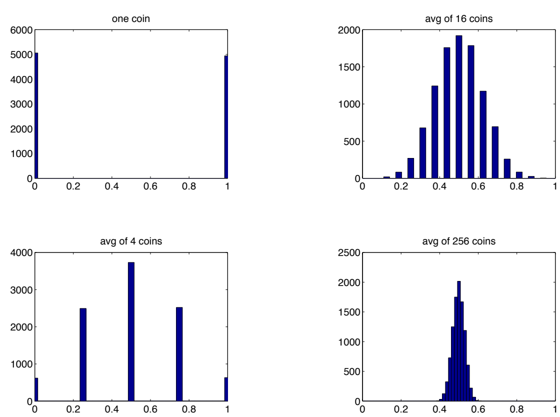
sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

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

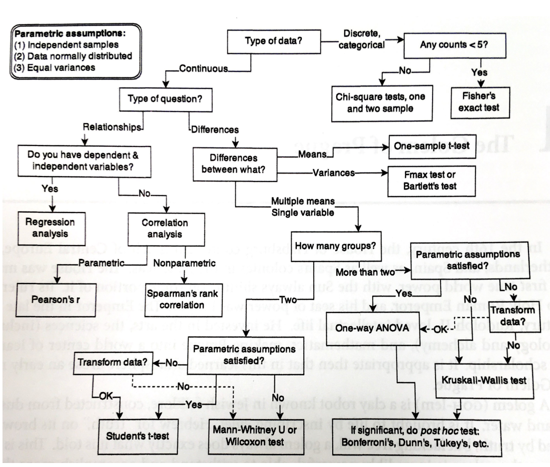
Central limit for a uniform distribution...



Central limit for a binary distribution...



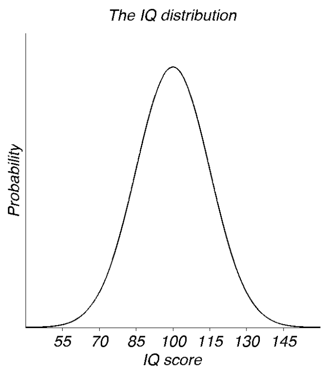
Classical “frequentist” statistical tests



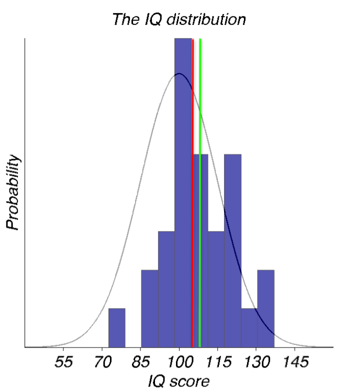
Statistical Rethinking, Richard McElreath

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
- H_1: NZT improves IQ
- H_0 ("null"): it does nothing

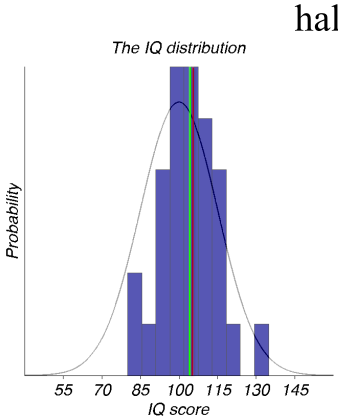


The z-test



- $\mu = 100$ (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".

Does NZT improve IQ scores or not?

	Reality: Yes	Reality: No
Significant? Yes	Correct	Type I error α-error False alarm
Significant? No	Type II error β-error Miss	Correct

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 $N(0,1)$

Common misconceptions

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

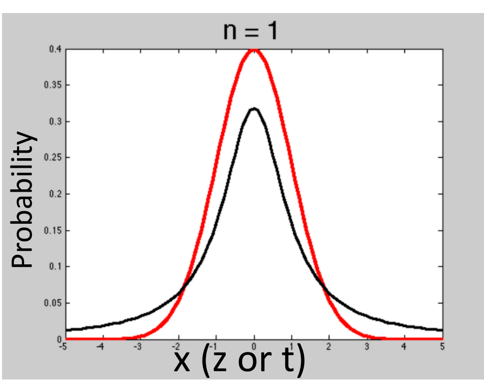
Does statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

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

Student's *t*-test

- σ not assumed known
- Use
$$s^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N-1}$$
- Why *N*-1? *s* is unbiased (unlike ML version), i.e., $E(s^2) = \sigma^2$
- 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

$$\pm z_{\alpha/2} \sqrt{\hat{p}\hat{q} / m}$$

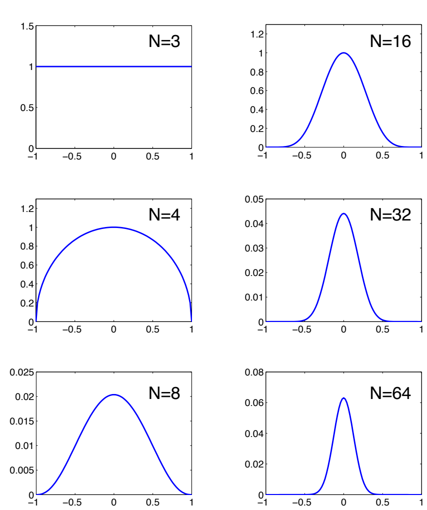
Many varieties of frequentist univariate tests

- 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.)

Lack of correlation is favored in N>3 dimensions

Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian random vectors in N dimensions:

$$(1 - d^2)^{\frac{N-3}{2}}$$



Estimation, more generally…

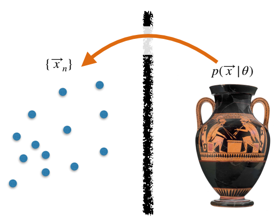
- An "estimator" is a function of the data, intended to provide an estimate of the "true" value of a parameter
- Traditionally, one evaluates estimator quality in terms of mean error ("bias") and variance
  (note: MSE = bias^2 + variance)
- Classical statistics generally aims for an unbiased estimator, with minimal variance ("MVUE")
- Modern view: trade off the bias and variance, through model selection, "regularization", or Bayesian priors

The maximum likelihood estimator (MLE)

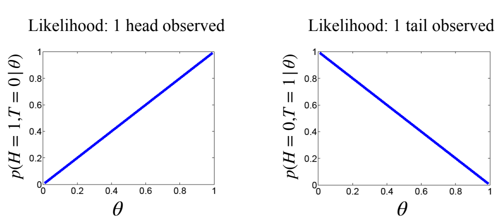
Sample average is appropriate when one has direct measurements of the thing being estimated. But one may want to estimate something (e.g., a model parameter) that is *indirectly* related to the measurements…

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

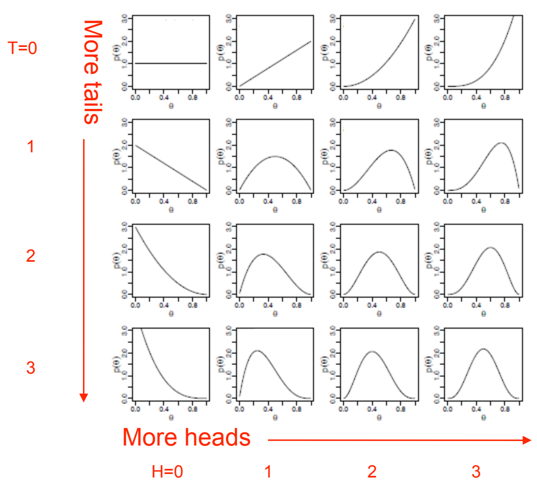
$$\hat{\theta}(\{\vec{x}_n\}) = \arg\max_{\theta} \prod_n p(\vec{x}_n|\theta)$$
$$= \arg\max_{\theta} \sum_n \log p(\vec{x}_n|\theta)$$



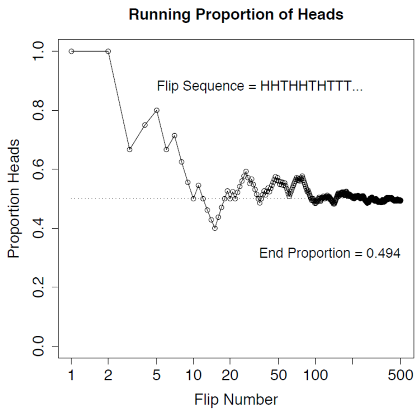
Example: Estimate the bias (probability of heads) of a coin flip



Likelihoods, $p(H, T \mid \theta)$



Convergence



Example ML Estimators - discrete

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

$$\hat{\theta} = \frac{H}{N}$$

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

(k's are measured event counts, lambda is mean)

$$\hat{\lambda} = \frac{1}{N} \sum_{n=1}^{N} k_n$$

Example ML Estimators - Continuous

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

$$\hat{\mu} = \frac{\sum_{n=1}^{N} x_n}{N} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (x_n - \hat{\mu})^2}{N}$$

(Note: this is biased!)

[on board]

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

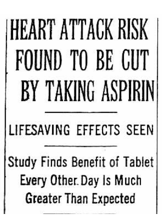
$$\hat{\theta} = \arg\max_n \{x_n\}$$

Properties of the MLE

- In general, the MLE is *asymptotically* unbiased, and *Gaussian*, but note can only rely on this if:
  - the likelihood model is correct
  - the MLE can be computed
  - you have lots of data
- Confidence:
  - SEM (relevant for direct estimates of mean)
  - inverse of second deriv of NLL (multi-D: “Hessian”)
  - simulation (resample from $p(x|\hat{\theta})$)
  - bootstrapping (re-sample 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 distribution (incl. stdev error bars or confidence intervals)
- Idea: instead of looking at distribution of estimates across repeated experiments, look across repeated resampling (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

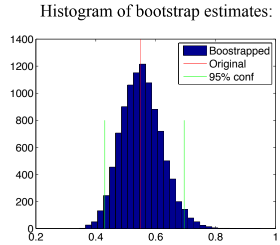
$$\hat{\theta} = \frac{104/11037}{189/11034} = .55. \quad (1.1)$$

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 $\hat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio

[Efron & Tibshirani ’98]

Histogram of bootstrap estimates:



=> with 95% confidence,

$$0.43 < \theta < 0.7$$

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

For strokes, the ratio of rates is

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

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

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

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]

Bayesian Inference

"Posterior" "Likelihood" "Prior"

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

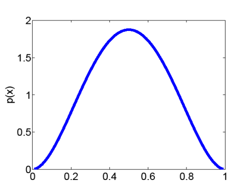
Normalization factor

Example: Posterior for coin

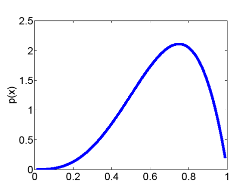
infer whether a coin is fair by flipping it repeatedly
here, x is the probability of heads (50% is fair)
$y_{1...n}$ are the outcomes of flips

Consider three different priors:

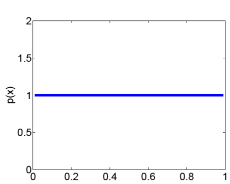
suspect fair

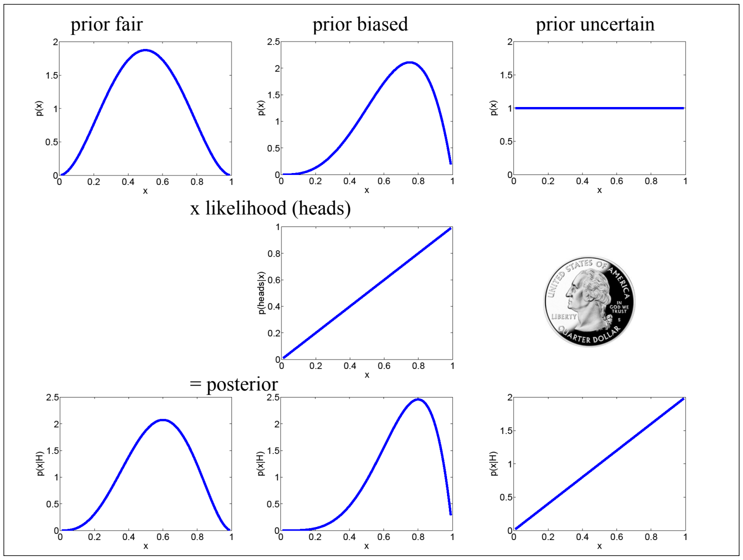


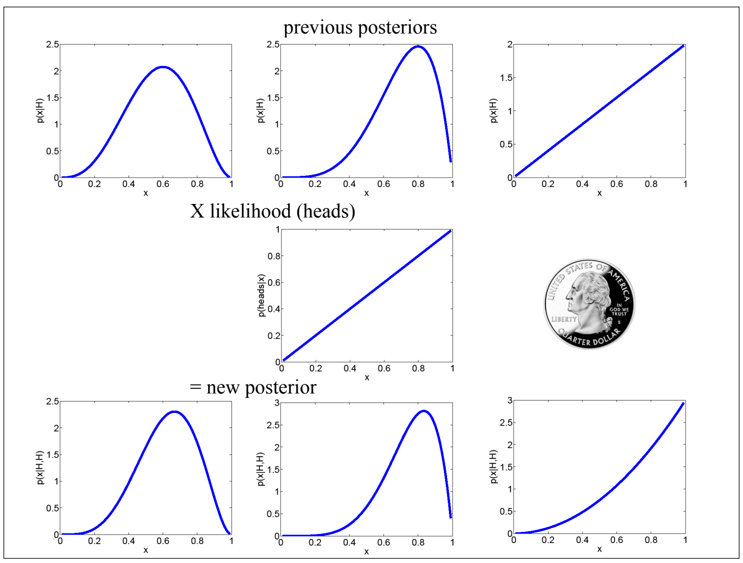
suspect biased

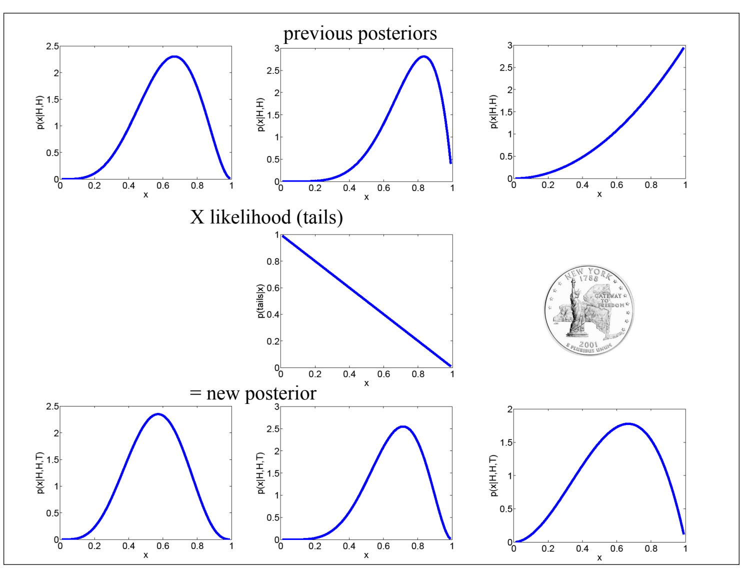


no idea

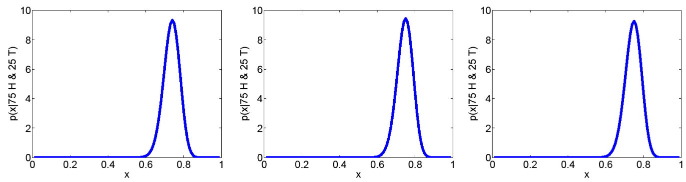



prior fair
prior biased
prior uncertain
p(x)
x
x likelihood (heads)
p(heads|x)
= posterior
p(x|H)


previous posteriors
p(x|H)
x
X likelihood (heads)
p(heads|x)
= new posterior
p(x|H,H)


previous posteriors
p(x|H,H)
x
X likelihood (tails)
p(tails|x)
= new posterior
p(x|H,H,T)

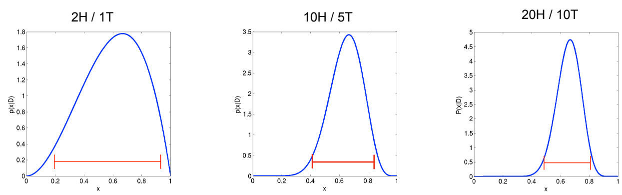
Posteriors after observing 75 heads, 25 tails



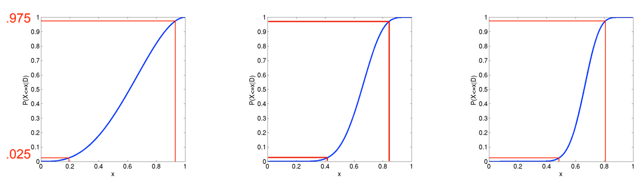
→prior differences are ultimately overwhelmed by data

Bayesian confidence intervals

PDFs



CDFs, and 95% confidence intervals



MAP estimation - Gaussian case

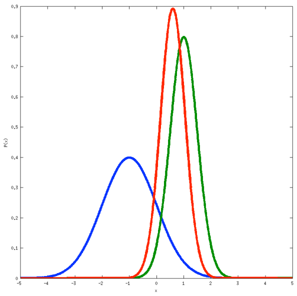
For measurements with Gaussian noise, and assuming a Gaussian prior, posterior is Gaussian.

- MAP estimate is a weighted average of prior mean and measurement
- posterior is Gaussian, allowing sequential updating
- explains “regression to the mean”, as shrinkage toward the prior

MAP with Gaussians

$$y = x + n, \quad x \sim N(\mu_x, \sigma_x), \ n \sim N(0, \sigma_n)$$

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &\propto e^{-\frac{1}{2}\left[\frac{1}{\sigma_n^2}(x-y)^2\right]} e^{-\frac{1}{2}\left[\frac{1}{\sigma_x^2}(x-\mu_x)^2\right]} \\ &= e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_n^2}+\frac{1}{\sigma_x^2}\right)x^2 - 2\left(\frac{y}{\sigma_n^2}+\frac{\mu_x}{\sigma_x^2}\right)x + \ldots\right]} \end{aligned}$$



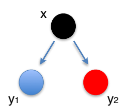
Completing the square shows that this posterior is also Gaussian, with

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}$$

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

The average of y and μ_x, weighted by inverse variances (a.k.a. "precisions")!

Two noisy measurements of the same variable:



$$y_1 = x + n_1 \qquad x \sim N(0, \sigma_x)$$
$$y_2 = x + n_2 \qquad n_k \sim N(0, \sigma_n), \text{ independent}$$

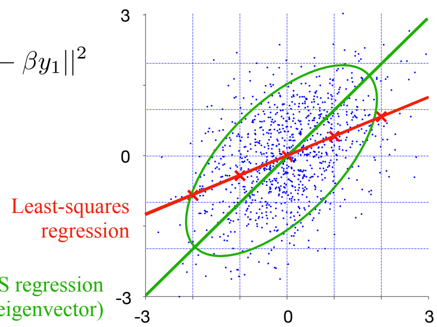
Joint measurement distribution: $\vec{y} \sim N(\vec{0}, \sigma_x^2 + \sigma_n^2 I)$

LS Regression:

$$\hat{\beta} = \arg\min_{\beta} ||y_2 - \beta y_1||^2$$
$$= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2}$$

$$\mathbb{E}(y_2|y_1) = \hat{\beta}\ y_1$$

"regression to the mean"



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."

- D. Kahneman

Hierarchy of statistically-motivated estimators

- Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_x p(\vec{d}|x)$
- 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 (BLS): $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbb{E}\left((x-\hat{x})^2 \,\middle|\, \vec{d}\right)$
  (special case)

$$= \mathbb{E}\left(x \,\middle|\, \vec{d}\right)$$