Learning and Teaching Statistical Inference

An Open Discussion

B. Blais

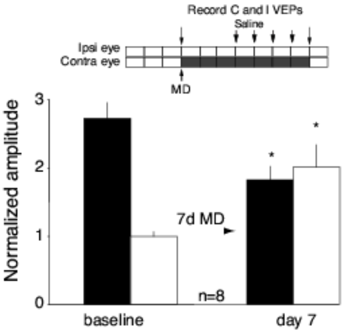
Department of Science and Technology
Bryant University

Faculty Development Seminar - Fall 2005

Abstract

This Faculty Development Seminar focuses on the learning and teaching of statistical inference, and is directed towards those who use statistics and statistical inference in either their teaching or research. During my summer vacation, I took the opportunity to learn and re-learn basic statistics. In this seminar, I would like to share what I have discovered in my studies, including some interesting history and pedagogy. I would like to then introduce some possible alternative approaches to teaching statistical inference, and open up the discussion to evaluate these suggestions and to get suggestions from the faculty for whom statistical inference plays a large role in their classroom. I would like to explore student misperceptions and challenges, along with interesting pedagogical examples which highlight the important aspects of statistical inference.

Some Data



- What does the "∗" mean?
- How can one objectively define "significance"?
- What are the assumptions?

Food For Thought

- What does the word *probability* mean?
- Why do we say that a coin has $p_{\text{head}} = 0.5$?
- What do we mean by *identical* repetitions?

Outline

1. History
   - Bernoulli to Laplace
   - Boole, Venn, Neyman, Pearson, Fisher, etc. . .
   - Cox and Jaynes
   - Two Schools of Thought on Probability
2. Estimating the Amplitude of a Signal
   - Two Approaches
   - Comparison
3. Comparisons
   - Hypothesis Testing
   - Unknown mean, Known Variance
   - Unknown mean, Unknown Variance
   - Unknown proportion
4. Other Examples
   - Behrens-Fisher
   - Flipping a Tack
5. Conclusions

History: Bernoulli

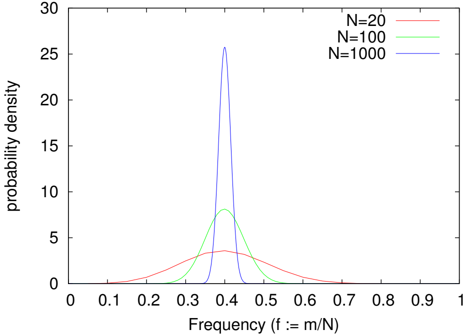
- James Bernoulli (1713) in “Ars Conjectandi”: defined probability as a “degree of certainty”.
  - His theorem states that, if the probability of an event is p then the limiting frequency of that event converges to p.

History: Bernoulli

Example: Coin with $p_{\text{head}} = 0.4$, N flips

$$p(m|N) = \binom{N}{m} 0.4^m (1-0.4)^{N-m}$$

as $N \to \infty$, observed frequency $f \equiv \frac{m}{N} \to 0.4$



History: Bernoulli

- Assignment of probabilities: Principle of Insufficient Reason

  - If the evidence does not provide any reason to choose proposition A_1 or A_2, then one assigns equal probability to both.
  - Equivalent states of knowledge (say, swapping labels 1 and 2) should yield identical probability assignments.

- Generalizes to N propositions

$$p(A) = \frac{m}{N} = \frac{\text{(number of cases favorable to A)}}{\text{(total number of equally possible cases)}}$$

History: Bernoulli, Bayes, Laplace

- James Bernoulli (1713) in “Ars Conjectandi”: defined probability as a “degree of certainty”.
  - His theorem states that, if the probability of an event is p then the limiting frequency of that event converges to p.
  - Inverse problem: given m occurrences out of N trials, what is the probability p of a single occurrence?
- Solution published posthumously by Rev. Thomas Bayes (1763), generalized, and applied to astrophysics by Laplace.

History: Bayes, Laplace

- Take as axioms the sum and product rules for probability:

Axioms

$$
\begin{aligned}
p(A|C) + p(\bar{A}|C) &= 1 \\
p(AB|C) &= p(A|BC)p(B|C)
\end{aligned}
$$

- From there, given the symmetry $p(AB|C) = p(BA|C)$ we get

Bayes' Theorem

$$
\begin{aligned}
p(A|BC)p(B) &= p(B|AC)p(A) \\
p(A|BC) &= \frac{p(B|AC)p(A)}{p(B)}
\end{aligned}
$$

History: Laplace

Bernoulli's Inverse Problem: Laplace's Solution

Given m occurrences out of N trials, what is the probability of a single occurrence?

- θ is the proposition:

  "the probability of a single occurrence is θ".
- I is any other information in the problem
- Bayes Theorem

$$p(\theta|m,N,I) = \frac{p(m,N|\theta,I)p(\theta|I)}{p(m,N|I)}$$

History: Laplace

Bernoulli's Inverse Problem: Laplace's Solution

Given m occurrences out of N trials, what is the probability of a single occurrence?

$$p(\theta|m,N,I) = \frac{p(m,N|\theta,I)p(\theta|I)}{p(m,N|I)}$$

- $p(m,N|\theta,I) = \binom{N}{m}\theta^m(1-\theta)^{N-m}$: Bernoulli's Th'm
- $p(\theta|I) = 1$: Uniform prior
- $p(m,N|I)$: Determined from normalization

$$p(\theta|m,N,I) = \frac{(N+1)!}{m!(N-m)!}\theta^m(1-\theta)^{N-m}$$

- Value of θ with the maximum probability:

History: Boole, Venn, Neyman, Pearson, Fisher, etc. . .

Criticisms of Laplace

1. The axioms are not clearly unique for a definition of probability as vague as "degrees of plausibility".
   Algebra of relative frequencies satisfies the axioms
2. It was unclear how to assign the prior probabilities of propositions in the first place: how to generalize Bernoulli's *Principle of Insufficient Reason* for continuous cases?
   Problem disappears: meaningless to speak of a probability of propositions because there is no limiting frequency (always true, or always false).

Solution:

define probability as the long-run relative frequency of occurrence

History: Birth of Statistics

- Hypotheses are either true or false for the entire population, and thus do not have a long-run relative frequency.
- Create a *statistic*: any function of the observed random variables in a sample, without any unknown quantities, e.g.

$$\text{Sample mean: } \bar{x} = \frac{1}{N}\sum_i x_i$$

$$\text{Sample variance: } S^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

- Criteria for choosing a statistic: unbiasedness, efficiency, consistency, coherence, sufficiency, the likelihood principle, etc. . .

History: R. T. Cox and E. T. Jaynes

Axioms for Probability Theory

1. Degrees of plausibility are represented by real numbers
2. Qualitative correspondence with common sense. Consistent with deductive logic in the limit of true and false propositions.
3. Consistency
   1. If a conclusion can be reasoned out in more than one way, every possible way must lead to the same result
   2. The theory must use all of the information provided
   3. Equivalent states of knowledge must be represented by equivalent plausibility assignments

Bayesian formulation uniquely satisfies these criteria

History: E. T. Jaynes

Generalization of the Principle of Indifference

- Maximum Entropy
  - Measure of the uncertainty, H, of a distribution, $(p_1, p_2, \ldots, p_n)$, called the entropy
  - Prior probabilities are assigned as those with the maximum entropy, given the initial information of the problem
- Transformation groups
  - Equal states of knowledge yield equal probability assignments

Two Schools of Thought on Probability

Frequentist Statistical Inference

$p(A) =$long-run relative frequency with which A occurs in identical repeats of an experiment.
"A" restricted to propositions about random variables.

Bayesian Inference

$p(A|B) =$a real number measure of the plausibility of a proposition/hypothesis A, given (conditional on) the truth of the information represented by proposition B.
"A" can be any logical proposition, *not* restricted to propositions about random variables.

Objective versus Subjective

- Bayesian inference is often labeled as *subjective*, because the probability is a measure of a state of knowledge, and not directly observable like a relative frequency
- Loredo 1990: "In this sense, Bayesian Probability Theory is 'subjective,' it describes states of knowledge, not states of nature. But it is 'objective' in that we insist that equivalent states of knowledge be represented by equal probabilities, and that problems be well-posed: enough information must be provided to allow unique, unambiguous probability assignments."

Definition of the Problem

- Magnitude of a signal, μ
- Given N measurements, x_i, contaminated with noise with known standard deviation, σ

Frequentist Approach

- Random variables are x_i (not μ, which is a constant parameter), each with a Gaussian distribution

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

- To estimate μ, we choose a *statistic* – a function of the random variables – and calculate its distribution connecting it to μ
- What is the “best” statistic? unbiased? sufficient?
- Choose the sample mean, $\bar{x}$, which has the sampling distribution

$$p(\bar{x}|\mu) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} e^{-N(\bar{x}-\mu)^2/2\sigma^2}$$

- Sampling distribution yields confidence intervals

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$$

Bayesian Approach

- Want the posterior distribution:

  "probability of μ given the data"

$$p(\mu|\mathbf{x},\sigma,I) = \frac{p(\mathbf{x}|\mu,\sigma,I)p(\mu|\sigma,I)}{p(\mathbf{x}|\sigma,I)}$$

- (Uniform) Prior

$$p(\mu|\sigma,I) = p(\mu|I) \quad = \begin{cases} A & \mu_{\text{min}} \leq \mu \leq \mu_{\text{max}} \\ 0 & \text{otherwise} \end{cases}$$

- Likelihood

$$p(\mathbf{x}|\mu,\sigma,I) \quad = \quad \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k-\mu)^2/2\sigma^2}$$

- Posterior

$$p(\mu|\mathbf{x},\sigma,I) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(\mu-\bar{x})^2/2\sigma^2}$$

Bayesian Approach

- Posterior

  "probability of μ given the data"

$$p(\mu|\mathbf{x},\sigma,I) = \sqrt{\frac{N}{2\pi\sigma^2}}e^{-N(\mu-\bar{x})^2/2\sigma^2}$$

- Maximum Posterior Estimate

  "most plausible value of μ given the data"

- Width of Posterior Gives Confidence Interval (Credible Interval?)

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$$

Comparison

- Same numerical result

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$$

- Different interpretation

Frequentist Statistical Inference

The result is a statement about the long term performance of adopting the procedure of estimating μ with $\bar{x}$. If one adopts this procedure, the average of the estimates of μ after many observations will converge to the true value of μ, and the statement about the interval containing μ will be true 68% of the time. (Loredo, 1990)

Bayesian Inference

The result is that $\bar{x}$ is the most plausible value of μ given the one set of data at hand, and there is a plausibility of 0.68 that μ is in the range $\bar{x} \pm \sigma/\sqrt{N}$.

Frequentist: Hypothesis Testing and p Values

- If you want to infer from the data that the mean value is, say, greater than zero...
  1. you set up the null with $H_o : \mu = 0$ and the alternate with $H_a : \mu > 0$
  2. select the appropriate statistic (z, t, etc...)
  3. calculate the p-value of the null, where you use hypothetical data and look for the frequency that H_o is true.
  4. you reject the null at the level of significance, usually at the 5% level.

p value

"the probability, computed assuming that the null hypothesis H_o is true, of observing a value of the test statistic that is at least as extreme as the value actually computed from the data" (Bowerman and O'Connell, 2003).

Frequentist: Hypothesis Testing and p Values

Trash Bag Strengths

$$\begin{aligned} N &= 40 \\ \mu &= 50.575 \\ \sigma &= 1.6438 \end{aligned}$$

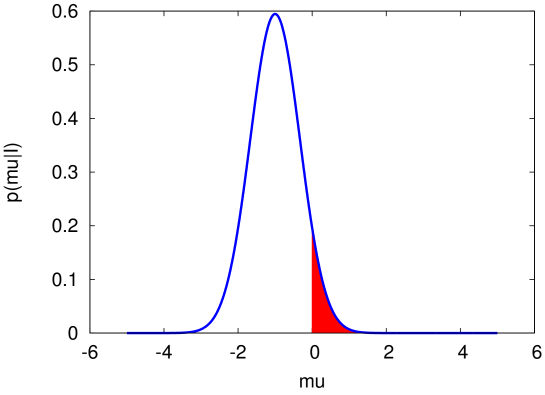
1. Test whether $\mu > 50$. $H_o : \mu = 50$ and the alternate with $H_a : \mu > 50$
2. Select z-statistic

$$z = \frac{\bar{x} - 50}{\sigma/\sqrt{n}} = \frac{50.575 - 50}{1.6438/\sqrt{40}} = 2.2123$$

3. $p = 1.34\% \Rightarrow$ reject H_o.

Bayesian: Equivalent to p value

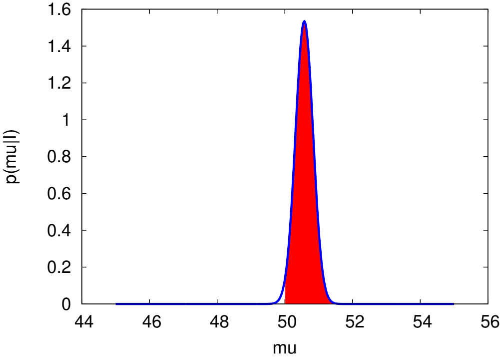
- If you want to infer from the data that the mean value is, say, greater than zero...
  1. integrate the posterior probability distribution from 0 to infinity and get the probability that μ is greater than 0



Bayesian: Equivalent to p value

Trash Bag Strengths

1. Test whether $\mu > 50$. Integrate $p(\mu|\mathbf{x}, I)$ from $\mu = 50$ to $\mu = \infty$
2. $p = 98.65\%$



Bayesian Equivalents

Unknown mean, Known Variance

- Posterior: z-dist

$$p(\mu|\mathbf{x}, \sigma, I) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(\bar{x}-\mu)^2/2\sigma^2}$$

- Best Estimate

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$$

Bayesian Equivalents

Unknown mean, Unknown Variance

- Posterior: t-dist, χ^2

$$
\begin{aligned}
p(\mu|\mathbf{x}, I) &\propto \left[N(\bar{x}-\mu)^2 + V\right]^{-N/2} \\
p(\sigma|\mathbf{x}, I) &\propto \frac{1}{\sigma^N} e^{-V/2\sigma^2}
\end{aligned}
$$

- Best Estimate

$$
\begin{aligned}
\mu &= \bar{x} \pm \frac{S}{\sqrt{N}} \\
\sigma &= S^2 \pm \frac{S^2}{\sqrt{2(N-1)}}
\end{aligned}
$$

$$
\bar{x} \equiv \frac{1}{N}\sum_{k=1}^{N} x_k \quad , \quad S^2 \equiv \frac{1}{(N-1)}\sum_{k=1}^{N}(x_k - \bar{x})^2
$$

Bayesian Equivalents

Unknown proportion

- Posterior: β-dist

$$p(\theta|D,I) = \frac{(N+1)!}{m!(N-m)!}\theta^m(1-\theta)^{N-m}$$

- Best Estimate

$$\theta = \frac{m}{N}$$

- Approximate for Large N

$$\bar{\theta} \approx \frac{m}{N} \equiv f$$
$$\sigma^2 \approx \frac{f(1-f)}{N}$$

Two Samples (Behrens-Fisher)

Problem from Jaynes, 1976

"Two manufacturers, A and B, are suppliers for a certain component, and we want to choose the one which affords the longer mean life. Manufacturer A supplies 9 units for test, which turn out to have a (mean $\pm$ standard deviation) lifetime of (42 ± 7.48) hours. B supplies 4 units, which yield (50 ± 6.48) hours." Should we prefer A or B?

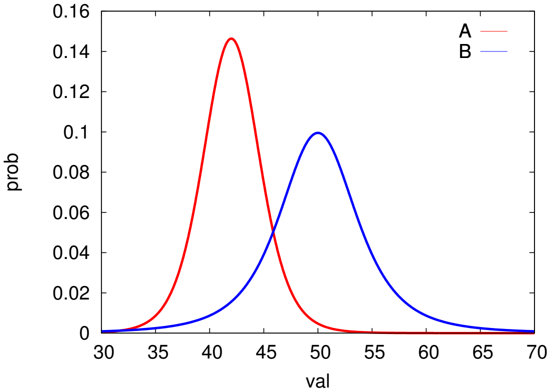
- Unknown mean, Unknown (possibly different) variance $\Rightarrow$ t-distribution

$$p(\mu_A|N_A, \bar{A}, S_A, I) \propto \left[(\bar{A} - \mu_A)^2 + S_A\right]^{-N_A/2}$$
$$p(\mu_B|N_B, \bar{B}, S_B, I) \propto \left[(\bar{B} - \mu_B)^2 + S_B\right]^{-N_B/2}$$

Two Samples (Behrens-Fisher)

- Unknown mean, Unknown (possibly different) variance $\Rightarrow$ t-distribution

$$p(\mu_A|N_A,\bar{A},S_A,I) \quad \propto \quad \left[(\bar{A}-\mu_A)^2+S_A\right]^{-N_A/2} \equiv p(\mu_A|I_A)$$
$$p(\mu_B|N_B,\bar{B},S_B,I) \quad \propto \quad \left[(\bar{B}-\mu_B)^2+S_B\right]^{-N_B/2} \equiv p(\mu_B|I_B)$$

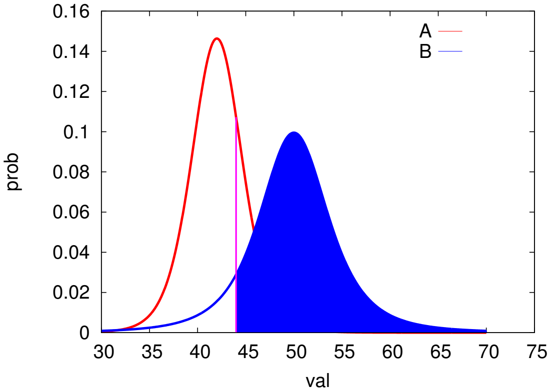


Two Samples (Behrens-Fisher)

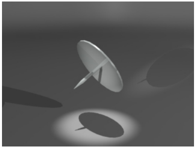
- Probability of $\mu_B > \mu_A$

$$\text{Prob}(\mu_B > \mu_A) = \int_{-\infty}^{\infty} d\mu_A \int_{\mu_A}^{\infty} d\mu_B P(\mu_A|I_A)P(\mu_B|I_B)$$

- Numerically $\text{Prob}(\mu_B > \mu_A) = 91.9\%$



Lindley (1976): Flipping a Tack



- Flipped thumbtack onto the table
- Data:

  UUUDUDUUUUUD - (9 Ups, and 3 Downs)
- Question:

  Is there good evidence that this tack is (or is not) unbiased (50-50 chance of U or D)?

Flipping a Tack: Frequentist Solution

- Obtain a p-value: “the chance of the observed result or more extreme results given infinite number of identical repetitions”
- For 12 flips, these results are
  - 9 U + 3 D (12 flips)
  - 10 U + 2 D (12 flips)
  - 11 U + 1 D (12 flips)
  - 12 U + 0 D (12 flips)
- Using the standard binomial distribution, with $N = 12$, we get

$$p = \binom{12}{3}\left(\frac{1}{2}\right)^{12} + \binom{12}{2}\left(\frac{1}{2}\right)^{12} + \binom{12}{1}\left(\frac{1}{2}\right)^{12} + \binom{12}{0}\left(\frac{1}{2}\right)^{12} = 7.30\%$$

- Not significant at the 5% level

Frequentist Solution

. . . BUT. . .

Flipping a Tack: Frequentist Solution

- What if the experimenter decided to stop measuring when he reached 3 Down?
- For 3D (3 Down), results at least as extreme are
  - 9 U + 3 D (12 flips)
  - 10 U + 3 D (13 flips)
  - 11 U + 3 D (14 flips)
  - 12 U + 3 D (15 flips)
  - 13 U + 3 D (16 flips)
  - $\vdots$
- Using the negative binomial distribution, with $D = 3$, we get

$$p = \binom{12-1}{3-1}\left(\frac{1}{2}\right)^{12} + \binom{13-1}{3-1}\left(\frac{1}{2}\right)^{13} + \binom{14-1}{3-1}\left(\frac{1}{2}\right)^{14} + \cdots = 3.27\%$$

- Is significant at the 5% level

Flipping a Tack: Bayesian Solution

- Posterior: β-dist

$$p(\theta|D,U,I) = \frac{(D+U+1)!}{D!U!}\theta^D(1-\theta)^U$$

$$p(\theta|D,U,I) = \frac{13!}{3!9!}\theta^3(1-\theta)^9$$

- Median value: $\theta_{\text{median}} = 0.275$
- Probability for the chance of D less than 50-50: integrate the posterior

$$\int_0^{0.5} d\theta p(\theta|D,U,I) = 0.954$$

- Is significant at the 5% level, and doesn't depend on choice of experiment

Conclusions

- Two Schools of Thought on Probability
  - Bayesian
  - Frequentist
- Both schools give *identical* numerical results to *all* problems covered in introductory statistics courses
- Interpretation perhaps more straightforward in the Bayesian approach
  - Win-Win: don't need to modify the content/examples/tests/syllabus very much, but you gain a possibly more intuitive perspective
- Questions?
- Comments?

Unknown μ, Known σ

(Uniform) Prior

$$p(\mu|\sigma, I) = p(\mu|I) \quad = \begin{cases} A & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Likelihood

$$p(\mathbf{x}|\mu, \sigma, I) \quad = \quad \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k-\mu)^2/2\sigma^2}$$

Posterior

$$p(\mu|\mathbf{x}, \sigma, I) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(\bar{x}-\mu)^2/2\sigma^2}$$

Unknown μ, Unknown σ

Jeffrey's Prior

$$p(\mu,\sigma|I) = \begin{cases} \frac{1}{\sigma} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

Likelihood

$$p(\mathbf{x}|\mu,\sigma,I) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{\frac{1}{-2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2}$$

Joint Posterior

$$p(\mu,\sigma|\mathbf{x},I) \propto \begin{cases} \left(\frac{1}{\sigma}\right)^{N+1} e^{\frac{1}{-2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

Unknown μ, Unknown σ, continued...

Joint Posterior

$$p(\mu,\sigma|\mathbf{x},I) \propto \begin{cases} \left(\frac{1}{\sigma}\right)^{N+1} e^{\frac{1}{-2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

Posterior for μ: t-dist

$$\begin{aligned} p(\mu|\mathbf{x},I) &= \int_0^\infty p(\mu,\sigma|\mathbf{x},I)d\sigma \\ &\propto \left[N(\bar{x}-\mu)^2+V\right]^{-N/2} \end{aligned}$$

Posterior for σ: χ^2-dist

$$p(\sigma|\mathbf{x},I) \propto \frac{1}{\sigma^N} e^{-V/2\sigma^2}$$

Changing Variables

If we have $Z = f(X, Y)$, and we know about X and Y, we can learn about Z.

$$p(Z|I) = \int\int p(Z|X, Y, I) \times p(X, Y|I)dXdY$$
$$= \int\int \delta(Z - f(X, Y)) \times p(X, Y|I)dXdY$$

Say, $Z = X - Y$, and X and Y are independent, then $p(X, Y|I) = p(X|I)p(Y|I)$ and we have

$$p(Z|I) = \int dXp(X, I) \int dYp(Y|I)\delta(Z - X + Y)$$
$$= \int dXp(X, I)p(Y = X - Z|I)$$

Difference of Means, $\delta \equiv \mu_x - \mu_y$, known σ_x and σ_y

- Posteriors

$$p(\mu_x|\mathbf{x},\sigma_x,I) = \sqrt{\frac{n}{2\pi\sigma_x^2}}e^{-n(\bar{x}-\mu_x)^2/2\sigma_x^2}$$

$$p(\mu_y|\mathbf{y},\sigma_y,I) = \sqrt{\frac{m}{2\pi\sigma_y^2}}e^{-n(\bar{y}-\mu_y)^2/2\sigma_y^2}$$

- Change of Variables

$$p(\delta|\mathbf{x},\mathbf{y},\sigma_x,\sigma_y,I) = \frac{\sqrt{nm}}{2\pi\sigma_x\sigma_y}\int d\mu_y e^{-n(\bar{x}-\delta-\mu_y)^2/2\sigma_x^2}e^{-m(\bar{y}-\mu_y)^2/2\sigma_y^2}$$

Posterior

$$\mu_\delta \equiv \mu_x - \mu_y \quad , \quad \sigma_\delta \equiv \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$$

$$p(\delta|\mathbf{x},\mathbf{y},\sigma_x,\sigma_y,I) = \frac{1}{\sqrt{2\pi\sigma_\delta^2}}e^{-(\delta-\mu_\delta)^2/2\sigma_\delta^2}$$

Bayesian Equivalents

Simple Linear Regression, $y_k = mx_k + b + \epsilon$

- Posterior

$$p(m, b|\mathbf{y}, I) \propto \frac{1}{\sigma^N} e^{-\sum (mx_k + b - y_k)^2 / 2\sigma^2}$$

- Best Estimate

$$m = \frac{c - N\bar{x}\bar{y}}{v - N(\bar{x})^2}$$
$$b = \frac{v\bar{y} - c\bar{x}}{v - N(\bar{x})^2}$$

with

$$v \equiv \sum x_k^2 \text{ , } c \equiv \sum x_k y_k$$

Generalization of the Principle of Indifference

E. T. Jaynes (1957, 1958)

- Measure of the uncertainty, H, of a distribution, $(p_1, p_2, \ldots, p_n)$, called the entropy
- Conditions for a measure of uncertainty:
  1. H is continuous function of p_i
  2. If all the p_i are equal, the quantity $A(n) = H(1/n, 1/n, \ldots, 1/n)$ is a monotonic increasing function of of n
  3. Composition law: regrouping the data yields the same measure of uncertainty.

$$H(p_1, p_2, \ldots, p_n) = -\sum_i p_i \log p_i$$

- Prior probabilities are assigned as those with the maximum entropy, given the initial information of the problem

Knowledge of N possibilities

- Find maximum of

$$H = -\sum_i p_i \log p_i$$

- with constraint

$$\sum_i p_i - 1 = 0$$

Knowledge of N possibilities

Find maximum of

$$Q = -\sum_i p_i \log p_i + \lambda_o \left(1 - \sum_i p_i\right)$$

setting $\partial Q/\partial p_j = 0$ we get

$$p_j = e^{-(1+\lambda_o)} = (\text{const})$$

normalizing we get

$$\boxed{p_j = \frac{1}{N}}$$

Knowledge of Mean

$$\mu = \sum_i p_i x_i$$

- Find maximum of

$$H = -\sum_i p_i \log p_i$$

- with constraints

$$\sum_i p_i - 1 = 0$$
$$\mu - \sum_i p_i x_i = 0$$

Knowledge of Mean

Find maximum of

$$Q = -\sum_i p_i \log p_i + \lambda_o \left(1 - \sum_i p_i\right) + \lambda_1 \left(\mu - \sum_i x_i p_i\right)$$

setting $\partial Q/\partial p_j = 0$ we get

$$p_j = e^{-(1+\lambda_o)} e^{-\lambda_1 x_j}$$

normalize from 0 to ∞ we get (continuous version)

$$\boxed{p(x|\mu) = \frac{1}{\mu} e^{-x/\mu}}$$

Knowledge of Mean and Variance

$$\mu = \sum_i p_i x_i$$

$$\sigma^2 = \sum_i (x_i - \mu)^2 p_i$$

Find maximum of

$$Q = -\sum_i p_i \log p_i + \lambda_o \left(1 - \sum_i p_i\right) + \lambda_1 \left(\sigma^2 - \sum_i (x_i - \mu)^2 p_i\right)$$

leads to

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