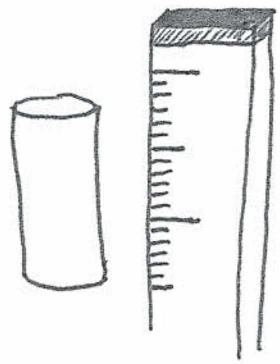
Measuring Lengths

Bayesian Data Analysis
APHY 451/551
ICSI 451/551

In this exercise, we will measure the lengths of various objects. We will use only a meter stick



cylinder		
top	bottom	length
1.1 cm	5.6 cm	
2.2 cm	6.65 cm	
⋮	⋮	

at least 10 measurements

Hints on improving accuracy

1. NEVER measure from the end of the meter stick. Instead, make a note of the coordinates of each end of the object.
2. Take multiple measurements at different positions
3. Estimate the uncertainty in your ability to take a measurement — BE HONEST.

What are some difficulties that might arise during this exercise?

- marks on the ruler are smudged / finite width
- object has nicks or bumps. – (length undefined)
- object may move relative to ruler during measurement
- angle of observation may affect results

How Uncertain are We?

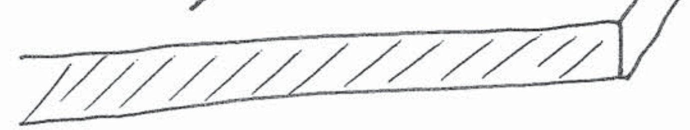
Personally, I can easily measure to the nearest millimeter.

But I am not sure I can measure to the nearest $1/10^{th}$ of a millimeter.

For me, I'd say that my position measurements have an uncertainty of about 0.25 mm.

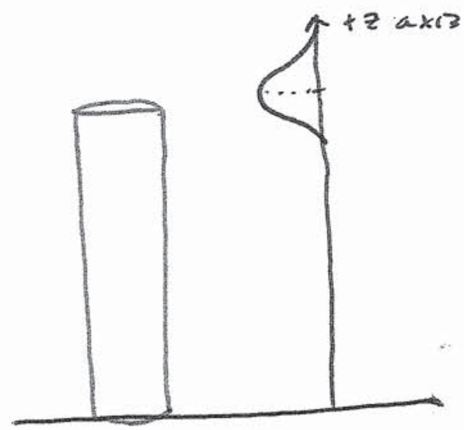
Note that this will change depending on

- instrument used
- object to be measured
- individual



The Prior Probability

The prior encodes what we knew about the length of the cylinder before we took a measurement.



We could assign a prior probability based on a mean height estimated visually.

However, our measurements are sufficiently informative (due to the fact that $\sigma_z < 1\,\text{mm}$) that the prior probability will make little to no difference in the result.

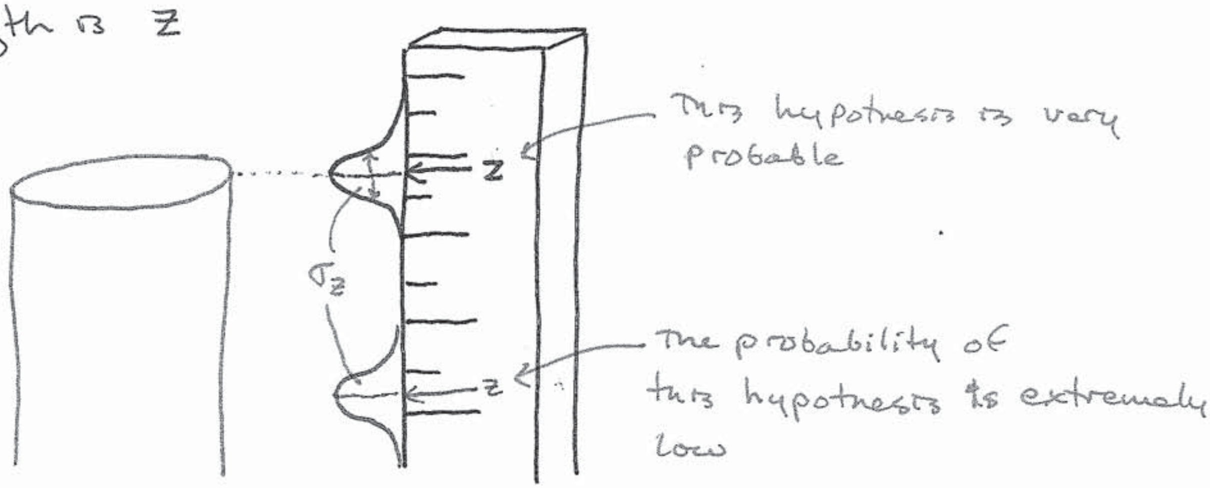
We will <u>assign</u> a constant probability density over a reasonable range

$$P(z|I) = \begin{cases} \dfrac{1}{z_{max} - z_{min}} & \text{if } z_{min} < z < z_{max} \\ 0 & \text{otherwise} \end{cases}$$

The Likelihood Function

The likelihood function encodes the forward problem along with our uncertainties.

For an object's length, the likelihood tells us the probability that we will measure length d given that the hypothesized length is z



We will cover the rational for such assignments later.

In this problem, we will <u>assign</u> a Gaussian density function.

$$P(d \mid z, I) = \frac{1}{\sqrt{2\pi}\,\sigma} \, \mathrm{Exp}\left[-\frac{1}{2\sigma^2}\left(z - d\right)^2\right]$$

measurement prediction

Depends on the difference between our prediction and our measurement

Uncertainty in Measuring Lengths

Now we will work to quantify our individual uncertainties.

$$\text{Length} = \text{Top} - \text{Bottom}$$
$$\downarrow$$
$$Z = X - Y$$

Say that the correct solution is $Z_0 = X_0 - Y_0$

Our measurements have been perturbed so that

$$X = X_0 + \delta X$$
$$Y = Y_0 + \delta Y$$
$$Z = Z_0 + \delta Z$$

Now

$$Z = X - Y \quad \Rightarrow \quad \delta Z = \delta X - \delta Y$$

since $Z_0 = X_0 - Y_0$

We will obtain this in a more rigorous fashion later

$$\langle \delta X^2 \rangle = \sigma_x^2, \quad \langle \delta Y^2 \rangle = \sigma_y^2 \quad \langle \delta X \delta Y \rangle = 0$$

unless object moves dur the measure

Compute

$$\langle \delta Z^2 \rangle = \langle \delta X^2 + \delta Y^2 - 2\delta X \delta Y \rangle$$
$$= \langle \delta X^2 \rangle + \langle \delta Y^2 \rangle - 2\langle \delta X \delta Y \rangle$$

$$\Rightarrow \sigma = \sqrt{\langle \delta Z^2 \rangle} = \sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{2}\,\sigma_x \quad \text{if } \sigma_x^2 = \sigma_y^2$$

uncertainty in length

$$\sigma = \sqrt{2}\,\sigma_x$$

uncertainty in each position measurement

Putting it Together

We combine these ideas using Bayes' Theorem

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

which, for our problem is:

$$p(z \mid d, I) = \underbrace{p(z \mid I)}_{\text{constant}} \frac{p(d \mid z, I)}{\underbrace{p(d \mid I)}_{\text{constant}}}$$

constant as we vary z. to test multiple hypotheses

So

$$p(z \mid d, I) \propto p(d \mid z, I)$$

$$\propto \mathrm{Exp}\left[-\frac{1}{2\sigma^2}(z-d)^2\right]$$

This Gaussian has a peak at $z = d$.

Clearly, for one datum point, this is the most probable length.

We write

$$\hat{z} = d \pm \sigma_z$$

Note that this solution is also the mean and median since the Gaussian is symmetric.

BUT WHAT IF WE HAVE MORE DATA ?

The Joint Likelihood

If we have taken several measurements

$$\text{data} = \{d_1, d_2, \ldots, d_N\}$$

we would like to use them all to obtain better results.

The key idea is that the value that you record for the length of the cylinder does not depend on the fact that previous measurements were taken.

Of course the results are CORRELATED, but don't confuse this with the fact that the measurements are independent

THE MEASUREMENTS ARE INDEPENDENT

This means that the joint likelihood

$$P(d_1, d_2, \ldots, d_N \mid z, I)$$

can be factorized

$$P(d_1, d_2, \ldots, d_N \mid z, I) = p(d_1 \mid z, I)\, p(d_2 \mid z, I) \ldots p(d_N \mid z, I)$$

Incidentally, the evidence can also be factored

$$P(d_1, d_2, \ldots d_N \mid I) = p(d_1 \mid I)\, p(d_2 \mid I) \ldots p(d_N \mid I)$$

The Solution

The most probable length $\hat{z}$ is the length at which $\log P$ is maximized.

To find this peak, take the derivate of $\log P$ wrt z and set it equal to zero

$$\left.\frac{d \log P}{dz}\right|_{\hat{z}} = 0$$

$$\frac{d \log P}{dz} = \frac{d}{dz}\left[c' - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z-d_i)^2\right]$$

$$= -\frac{1}{2\sigma^2}(2)\sum_{i=1}^{N}(z-d_i)$$

$$\Rightarrow \quad -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{z}-d_i) = 0$$

$$\Rightarrow \quad N\hat{z} - \sum_{i=1}^{N} d_i = 0$$

$$\Rightarrow \quad \hat{z} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

$$\boxed{\hat{z} = \overline{d_i}}$$

The optimal solution is the MEAN!

Uncertainty in a Least-Squares Estimate

The least-squares estimate is the one that minimizes Chi-squared

The logarithm of the posterior is

$$L = \log_e P(z \mid D, I) = \text{constant} - \frac{\chi^2}{2}$$

Beware! Authors will use L for log posterior and for the likelihood. PAY ATTENTION!

The most probable estimate maximizes L

$$\hat{z} = \arg\max_z L = \arg\min_z \chi^2$$

means the value of z that maximizes L

The uncertainty in our estimate is given by

$$\sigma_z^2 = -\left(\frac{\partial^2 L}{\partial z^2}\right)^{-1}$$

$$\frac{\partial L}{\partial z} = -\frac{1}{2}\frac{\partial \chi^2}{\partial z} = -\sum_{i=1}^{N} \frac{(z-d_i)}{\sigma^2}\frac{\partial z}{\partial z} = -\sum_{i=1}^{N} \frac{(z-d_i)}{\sigma^2}$$

$$\frac{\partial^2 L}{\partial z^2} = \frac{\partial}{\partial z}\left(-\sum_{i=1}^{N} \frac{(z-d_i)}{\sigma^2}\right) = -\frac{N}{\sigma^2}$$

$$\Rightarrow \sigma_z^2 = \frac{\sigma^2}{N}$$

$$\boxed{\text{THE SOLUTION IS} \quad z = \bar{d_i} \pm \frac{\sigma}{\sqrt{N}}}$$

Data with Different-Sized Error Bars

Previously we found that for a Gaussian likelihood and for equal errorbars in the measurements

$$\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$$

Now we will assume a Gaussian likelihood again but allow for different error bars.

$$P(x_k \mid \mu, \sigma_k, I) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma_k^2}\right]$$

For N measurements...

$$\text{Log } P = \text{constant} - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_k^2}$$

To find optimal value of μ set first derivative = 0

$$\left.\frac{d \text{ Log } P}{d\mu}\right|_{\hat{\mu}} = 0$$

$$\frac{d}{du}\left(C - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_k^2}\right) = \sum_{k=1}^{N} \frac{(x_k-\mu)}{\sigma_k^2} = 0$$

$$\Rightarrow \sum_{k=1}^{N} \frac{x_k}{\sigma_k^2} = \hat{\mu} \sum \frac{1}{\sigma_k^2} \quad \Rightarrow \quad \boxed{\hat{\mu} = \frac{\sum_{k=1}^{N} \frac{x_k}{\sigma_k^2}}{\sum_{k=1}^{N} \frac{1}{\sigma_k^2}} = \frac{\sum_{k=1}^{N} w_k x_k}{\sum_{k=1}^{N} w_k}}$$

where $w_k = \frac{1}{\sigma_k}$

weighted average of the data base on uncertainty in the measurement & combine data from multiple sources

Least - Squares Solutions

Least-Squares solutions arise from a Gaussian likelihood and a uniform prior, and the assumption that the data are independent

They are Maximum Likelihood solutions

$$P(X \mid D, I) \propto P(D \mid X, I)$$

Note $\{\sigma\}$ are missing

$P(X \mid D, \{\sigma\}, I)$

$\equiv$

$P(X \mid \{d\}, \{\sigma\}, I)$

NOTATIONAL CONVENIENCE

$\uparrow$ posterior probability for X

$\uparrow$ likelihood of data D given X

Say that the FORWARD PROBLEM makes predictions F_k about the data D_k so that $F_k = F(X)$

prediction

data

$$P(X \mid D, I) \propto \prod_{k=1}^{N} \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right]$$

$$\propto \exp\left[-\frac{1}{2}\sum_{k=1}^{N}\left(\frac{F_k - D_k}{\sigma_k}\right)^2\right] \equiv \exp\left[-\frac{\chi^2}{2}\right]$$

$$\text{Log } P = \text{Constant} - \frac{\chi^2}{2}$$

where $\chi^2 = \sum_{k=1}^{N}\left(\frac{F_k - D_k}{\sigma_k}\right)^2$

While commonly used, least-squares or χ^2 does not always work. WHY? Because the underlying assumptions may not be satisfied.

Sequential or Parallel

With several measurements, Bayes Theorem is

$$p(z \mid d_1, d_2, \ldots, d_N, I) = p(z \mid I)\,\frac{p(d_1, d_2, \ldots d_N \mid z, I)}{p(d_1, d_2, \ldots d_N \mid \cancel{z}, I)}$$

PARALLEL

$$= p(z \mid I)\,\frac{p(d_1 \mid z, I)\,p(d_2 \mid z, I) \ldots p(d_N \mid z, I)}{p(d_1 \mid I)\,p(d_2 \mid I) \ldots p(d_N \mid I)}$$

$$= p(z \mid I) \prod_{i=1}^{N} \frac{p(d_i \mid z, I)}{p(d_i \mid I)}$$

ALTERNATELY

$$= \left\{ \left[\left(p(z \mid I)\,\frac{p(d_1 \mid z, I)}{p(d_1 \mid I)} \right) \cdot \frac{p(d_2 \mid z, I)}{p(d_2 \mid I)} \right] \ldots \frac{p(d_N \mid z, I)}{p(d_N \mid I)} \right\}$$

posterior for d_1 becomes prior for d_2

$$= \left\{ \underbrace{\left[p(z \mid d_1, I) \cdot \frac{p(d_2 \mid z, I)}{p(d_2 \mid I)} \right]}_{\text{posterior for } d_1, d_2 \text{ becomes new prior}} \ldots \frac{p(d_N \mid z, I)}{p(d_N \mid I)} \right\}$$

NOTE THAT ORDER OF THE DATA DOES NOT MATTER

THIS IS KEY!

SEQUENTIAL Posterior at one stage becomes prior for next stage

The Final Analysis

Let $D = \{d_1, d_2, \ldots, d_N\}$

We have that

MAXIMUM LIKELIHOOD SOLUTION

Compare to Stutz Sec 3.5

$$P(z \mid D, I) \propto P(D \mid z, I)$$

$$\propto \prod_{i=1}^{N} P(d_i \mid z, I)$$

$$\propto \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \operatorname{Exp}\left[-\frac{1}{2\sigma_z^2}(z - d_i)^2\right]$$

$$\propto \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \operatorname{Exp}\left[-\frac{1}{2\sigma_z^2}\sum_{i=1}^{N}(z - d_i)^2\right]$$

It is more convenient to work with the logarithm of the posterior

$$\log P(z \mid D, I) = C + N \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z - d_i)^2$$

$$= C' - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z - d_i)^2$$

CHI - SQUARED

Define $\chi^2 = \sum_{i=1}^{N}\left(\frac{z - d_i}{\sigma}\right)^2$

Now $\log P(z \mid D, I) = \text{constant} - \frac{\chi^2}{2}$

The most probable value of the length z is found by maximizing the posterior or by [illegible] χ^2