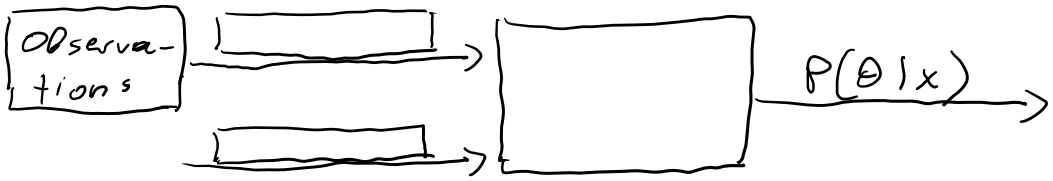
Classical Statistics

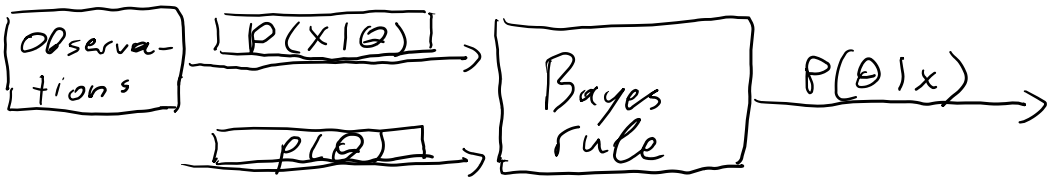
General ideas about classical estimation

At the second lecture the ideas of Bayesian inference was introduced. In that framework we assume that what we want to infer is a RV. As such it always has a distribution. The distribution we start with before we get any data is called the prior and the distribution we update the prior to after we consult the data is the posterior.

Q1. Fill in the diagram depicting a Bayesian inference process.



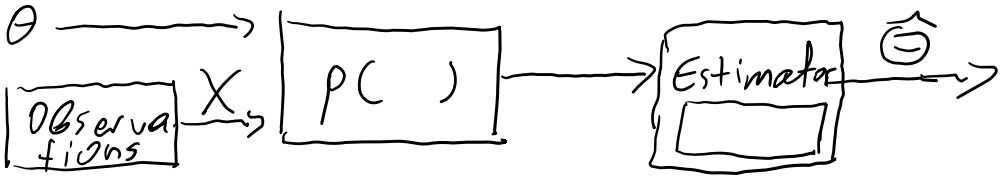
A1.



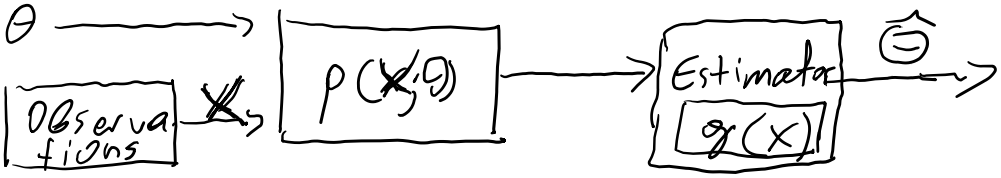
Q2. How about if I want to know the speed of light or the electrons mass to charge ratio or in general if I know that the thing I want to infer has a specific value and I simply do not know it and any variability in measurements comes from the crudeness of our experiments.

A2. In these cases inferring a whole distribution does not make sense. This is the field of classical (non Bayesian) statistics where we treat the unknown quantity not as a RV but as a specific value (or set of values for a multidimensional quantity).

Q3. Fill in the following diagram showing classical statistics inference of a quantity that is not a RV



A3.



Q4. In the diagram above circle the RVs

A4

The Theta hat (estimator) and the capital X.

Q5. We go through the above process collecting a data set x and passing it through g(x). We obtain $\widehat{\theta}$. Is that a RV?

A5.

No. This is called the estimate.

Q6. In the notations $p_{X|\Theta}(X|\Theta)$ in Bayes and $p_X(X;\theta)$ in classical statistics, is the second notation a conditional distribution? What does the second notation mean?

A6.

No. Theta is not a RV in the 2nd case. $p_X(X;\theta)$ is a way to represent multiple models (multiple $p_X(X)$) each parametrized by a value of theta.

Mention here Hypothesis Testing and Estimator Design. We will talk about Estimator Design and try to build Estimators that are good.

Q7. What do we mean by a "good Estimator"?

A7.

One that has a small error: $\widehat{\Theta} - \theta$ is small

Estimator properties

Q8. (Dependency on θ) Can you come up with an Estimator to calculate the value of the mean θ of the process that has given you X1 to Xn iid RVs (which also has a variance of σ^2)? Is the Estimator an RV? Watch your nomenclature.

A8.

The sample mean $\widehat{\Theta}_n = M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$

Yes (depends on RVs)

Q9. What is the difference between the expectation of $\widehat{\Theta}_n$ and the actual mean θ? Does that difference depend on the value of θ? Show that in general the expectation of $\widehat{\Theta}_n$ does depend on θ.

A9.

$E[\widehat{\Theta}_n] - \theta = 0$ and this does not depend on the value of θ. This difference is called the **bias** (think of it as a systematic offset of the estimate to the true value) and if it is 0 for all θ the Estimator is unbiased.

Generally: $\widehat{\Theta}_n = g(X;\theta) \Rightarrow E[\widehat{\Theta}_n] = \sum_X g(X)p(X;\theta)$

Q10. (Dependency on n) What does the WLLN say $\widehat{\Theta}_n$ will converge to as n goes to infinity? Is that true for all θ?

A10. θ. It is true for all θ. Estimators that have this property are called consistent.

Q11. Write the mean squared error of the $\widehat{\Theta}_n$ estimator (listen carefully to the words and build up the expression from right to left, i.e. error -> squared -> mean).

A11.

$$MSE = E[(\widehat{\Theta}_n - \theta)^2]$$

Q12. Decompose that in variance and bias terms using $E[Z^2] = var(Z) + (E[z])^2$.

A12.

$$E\left[(\widehat{\Theta}_n - \theta)^2\right] = var\left(\widehat{\Theta}_n - \theta\right) + (E\left[(\widehat{\Theta}_n - \theta)\right])^2 = var\left(\widehat{\Theta}_n - \theta\right) + (E\left[\widehat{\Theta}_n\right] - \theta))^2$$
$$= var\left(\widehat{\Theta}_n\right) + bias^2$$

Q13. Write the MSE in variance and bias terms of the sample mean estimator and of an estimator that is always 0.

A13.

$$MSE = E[(M_n - \theta)^2] = var(M_n) + bias(M_n)^2 = \frac{\sigma^2}{n} + 0$$

$$MSE = E[(0 - \theta)^2] = var(0) + (E[0] - \theta)^2 = 0 + \theta^2$$

Q14. By the way the $\sqrt{var(\widehat{\Theta}_n)}$ is called the standard error. What does it tell us?

A14.

It describes how spread around the true value θ the different estimates $\hat{\theta}$ created by different data sets would be.

Confidence Intervals

Q15. Given that one has found an estimate for an unknown value, what else would it be a good idea to calculate?

A15.

The standard error would be nice, but even nicer (and more usual) would be the confidence intervals.

An 1-a confidence interval is the interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$ such that

$P\left(\widehat{\Theta}^- \leq \theta \leq \widehat{\Theta}^+\right) \geq 1 - a$ for all θ

Q16. Back to the BCI problem. Let's assume you have estimated (from a single set of experimental data) that the percentage of the people that do not respond to your BCI algorithm is 4.32% ($\theta = 0.0432$). Let's also say that you have calculate that the 95% confidence interval for this value is [0.028, 0.058]. Can you go to investors and say that "With 95% probability the percentage of people who cannot use our product is between 2.8 and 5.8"?

A16.

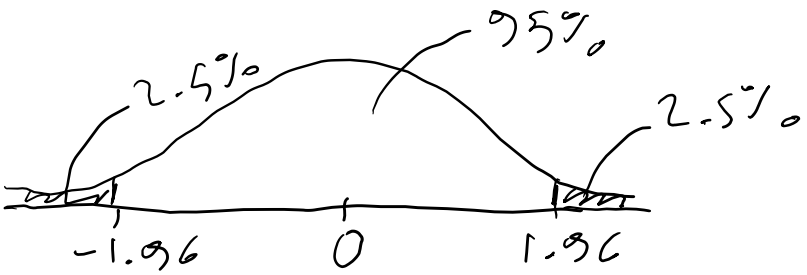
No. Why not?

Because $P(0.028 \leq \theta \leq 0.058) \geq 0.95$ doesn't make sense (where are the RVs). This is a single realization of a confidence interval. There is a 95% chance that at every realization, the real value of the percentage of interest of the population will fall within the calculated confidence interval (for that realization).

Q17. Calculate a 95% confidence interval of the estimate of the mean θ of a random process with standard deviation σ (assume you know this) using n iid RVs X1 to Xn. Use the CLT. Start by drawing the cdf of the standard distribution and think what a 95% confidence interval means on this.

A17.

$\Phi(1.96) = 0.975 = 1 - 0.125$



$$P\left(\frac{|S_n - \theta n|}{\sigma\sqrt{n}} \leq 1.96\right) = 0.95$$

$$\widehat{\Theta} = M_n = \frac{S_n}{n}$$

$$P\left(\frac{\left|\widehat{\Theta} - \theta\right|}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95 \quad (CLT)$$

$$P\left(\widehat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95$$

$$\widehat{\Theta}^- = \widehat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}}$$

$$\widehat{\Theta}^+ = \widehat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}$$

Q18. Awesome. But we what if we do not know the σ?

A18.

1) Maybe we know an upper bound (e.g. in the case of Bernoulli $\sigma \leq 0.5$)

2) Estimate σ, e.g. if we know the distribution of Xi and the standard deviation is a function of its mean (so we can use the estimate of the mean to estimate the σ).

3) More general we can use the sample mean estimate because $\sigma = E[(X_i - \theta)^2]$ and using the WLLN we get
   $\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \longrightarrow \sigma$
   We still do not know θ but we can use its estimate $\widehat{\Theta}_n$ so
   $\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\Theta}_n)^2 \longrightarrow \sigma$ because $\widehat{\Theta}_n \longrightarrow \theta$ as n increases

   There are tricks to deal with the extra randomness estimating σ introduces (t-distribution instead of normal). They are relevant for n<30. Also the above estimator of σ is biased. An unbiased estimator has n-1 as a denominator (but again this is relevant for small n).

Q19. We used the WLLN to estimate $\sigma = E[(X_i - \theta)^2]$. Generalize this idea. Use this to create an estimator of the Covariance Cov(X,Y) between two iid RVs

(remember $Cov(X,Y) = E[(X - \theta_X)(Y - \theta_Y)]$).

A19.

In general we can use the same method for estimating the expectation of any function g(X) of a RV.

If $\theta = E[g(X)]$ then $\widehat{\Theta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$

For the Cov

$$\widehat{Cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}((X_i - \widehat{\Theta}_X)(Y_i - \widehat{\Theta}_Y))$$

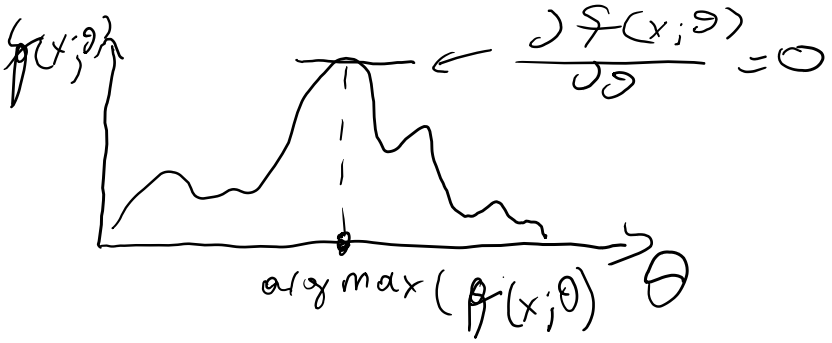
Maximum Likelihood Estimation (MLE)

Q20. But what do we do if the value we need to estimate doesn't have a representation as an expectation value of some function of the RVs ($\theta = E[g(X)]$)?

A20.

We can find the value of θ that makes the data we have observed most <u>likely.</u>

So $\hat{\theta}_{ML} = argmax_\theta(p_X(x;\theta))$



Some comments about the MLE

1) It is consistent (**Q**: What does that mean? **A**: $\widehat{\Theta}_n \rightarrow \theta$)

2) It is asymptotically normal (**Q:** Write this down as an expression. **A**: $\frac{\widehat{\Theta}_n - \theta}{\sigma(\widehat{\Theta}_n)} \rightarrow N(0,1)$, **Q:** Why is this true? **A**: CLT)

3) $\sigma(\widehat{\Theta}_n)$ (standard error) is a quantity that can itself be estimated. Knowing it leads to also being able to calculate confidence intervals of $\widehat{\Theta}_n$.

4) MLE is the "best" estimator in the sense that it has the smallest standard error.

Q21. There is a RV K with a binomial distribution. We know the parameter n of the distribution but not the parameter p (call it here θ). Use MLE to find θ.

$$p_K(k;\theta) = \binom{k}{n}\theta^k(1-\theta)^{n-k}$$

We need to find the θ where the derivative over θ of the above likelihood function becomes 0. Better to log first!

$$\log\big(p_K(k;\theta)\big) = \log\left(\binom{k}{n}\right) + klog(\theta) + (n-k)\log(1-\theta)$$

$$\frac{\partial \log\big(p_K(k;\theta)\big)}{\partial\theta} = \frac{\partial klog(\theta)}{\partial\theta} + \frac{\partial(n-k)\log(1-\theta)}{\partial\theta} = 0$$

$$\frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

$$\theta = \frac{k}{n}$$