Recap: distributions, point estimates and confidence

Johan Thim (johan.thim@liu.se)

10 september 2018

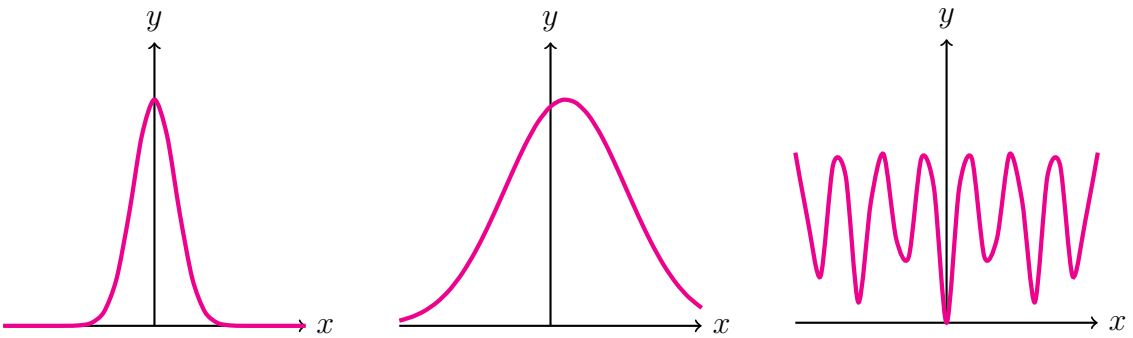
So what's the deal with all the terminology? Distributions, samples, random samples, estimators and confidence intervals. How does it all fit together? That's what I've been trying to show you guys, but maybe the idea gets lost in all the details. Trees, forests and all... So let us recap a bit and collect what we know.
The starting point is pretty much the same $x_1, x_2, \ldots, x_n$. This is typically measured data. What can we do with this? To do any form of reasonable analysis, we need a model. What we usually do is assume that the sample consists of observations of random variables, so:

Assumption 1: $X_1, X_2, \ldots, X_n$ are random variables and $x_1, x_2, \ldots, x_n$ are observations.

Next up is independence. We almost always assume that the random sample $X_1, X_2, \ldots, X_n$ is independent. Analysis gets much harder without this assumptions, so:

Assumption 2: $X_1, X_2, \ldots, X_n$ are independent.

So now we have independent random variables. What's next? Well, how are they distributed? Assuming a continuous distribution for simplicity, how does the probability density look?



Like this? Or this? Or maybe this?

We obviously have to assume something about the distributions to obtain something useful. So:

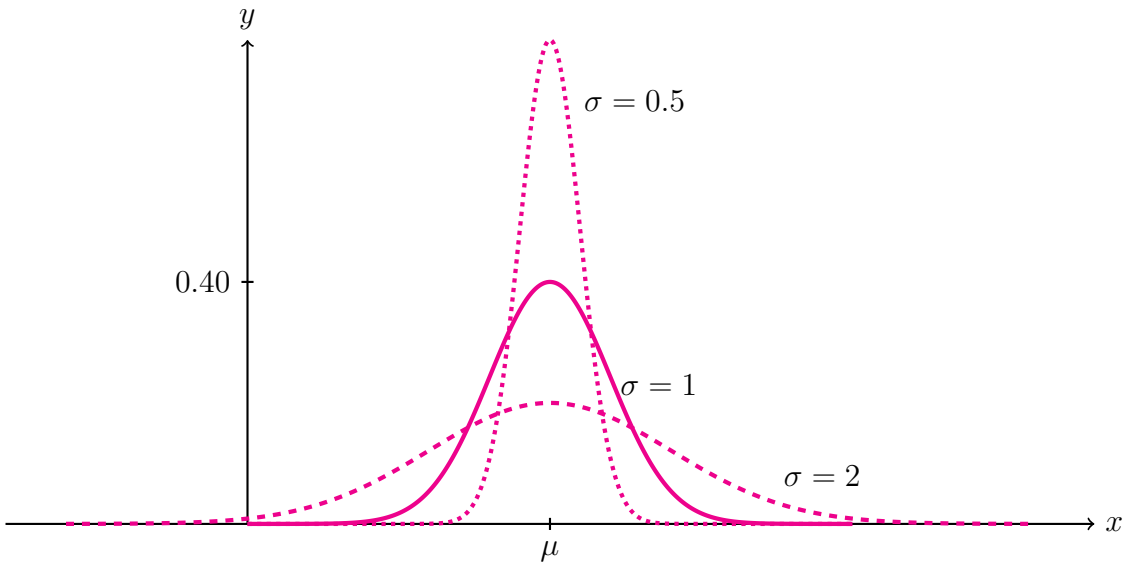
Assumption 3: $X_1, X_2, \ldots, X_n$ have known *types* of distribution, that is, we know for example that they are normally distributed. However, the exact distribution depends on something unknown: the parameter θ (which might be a vector of unknown parameters). It is not necessary that they all share the exact same distribution, but usually we will assume this. However, it is important that if the have different distributions, they depend on the same unknown parameter.

Normal distribution

Consider the normal distribution. Its shape is well-known (the Bell curve or Gaussian), but it might be moved around and it might be stretched (or contracted). The parameters μ and σ does this. How? Well, the density function for a normal distribution looks as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbf{R},$$

where $\mu \in \mathbf{R}$ and $\sigma > 0$ are parameters. It turns out that they happen to be the expectation and the standard deviation, but that's not clear from the definition above but follows when doing the necessary calculations. Right now they are just parameters for the distribution.



We see the same type of shape, but different values on the parameters yield different curves. Our aim now would be — using a sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma^2)$ — to estimate the unknown parameters μ and/or σ.
To make calculations easier, we usually make the following assumption:

Assumption 4: $X_1, X_2, \ldots, X_n$ are independent.

The Goal

If $x_1, x_2, \ldots, x_n$ are a sample from a distribution F that depends on an unknown parameter θ, we wish to find a *good* way of using the known quantities $x_1, x_2, \ldots, x_n$ to **estimate** θ by some function

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

What about θ, $\widehat{\theta}$, $\widehat{\Theta}$...

What about the big ball with the H inside and the hat ontop[1]? There are three types of θ:s involved here. Let's see how this works.

[1] Thank you dear student, for the subtle hint about my handwriting on the board..

So the situation is as follows. We have a method of finding an estimate $\widehat{\theta}$ for the unknown parameter θ by using the sample $x_1, x_2, \ldots, x_n$, namely by the so called **statistic**

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n).$$

Question 1. If we repeat the "experiment" and thus obtain a new sample $y_1, y_2, \ldots, y_n$, what can we say about the estimate $\widehat{\theta} = g(y_1, y_2, \ldots, y_n)$?

Obviously we won't get the same estimate (in general) since the values will likely have changed, so what's going on? This is where the probability comes in to play. We know (by assumption) what type of distribution we're working with. We know how this distribution depend on the unknown parameter θ. So we can – in theory at least – do calculations with regards to the estimator viewed as a random variable as long as we allow our answers to contain the unknown θ. How do we move to something random? Considering that we view $x_1, x_2, \ldots, x_n$ as observations of random variables $X_1, X_2, \ldots, X_n$, we just exchange all instances of the former by the latter. Thus we obtain the estimator

$$\widehat{\Theta} = g(X_1, X_2, \ldots, X_n)$$

as an n-dimensional random variable. Note that this is the *same* function $g\colon \mathbf{R}^n \to \mathbf{R}^p$ that was used when defining $\widehat{\theta}$ (p is the number of parameters we are estimating).

Question 2. Can we use $\widehat{\Theta}$ to obtain some type of bounds for the unknown θ with a given probability?

The answer is yes, at least if we can transform $\widehat{\Theta}$ to something with a known distribution. This is the procedure used to find **confidence intervals**. What is a confidence interval? We'll get back to this down below.

What would we like to happen?

Good Property 1. We do *not* want the estimator to be biased. By that we mean that we want the estimator $\widehat{\theta}$ to *be* the unknown parameter θ on *average*. Well, that's rather unspecific, so instead we mathematically define an unbiased estimator as an estimator such that $E(\widehat{\Theta}) = \theta$.

Good Property 2. We would like for our estimator to have the property that as the sample grows larger, the probability of $\widehat{\Theta}$ being off from θ is tending to zero. This is **consistency**. In mathematical terms, we want the following to hold. Let $\widehat{\Theta}_n$ be the estimator for a sample of size n. Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\Theta}_n - \theta| > \epsilon) = 0.$$

Phrased differently, this means that the sequence $\widehat{\Theta}_n$ of random variables converges to θ in probability. It is difficult to work with this directly, so we normally use a result that follows from Chebyshev's inequality. Namely that if $V(\widehat{\Theta}_n) \to 0$ as $n \to \infty$ and $\widehat{\Theta}_n$ is unbiased (note that this is needed for the theorem to hold, not for the estimator to be consistent), then the estimator is consistent.

Expectation and variance

We've used (implicitly or explicitly) some results concerning the expectation and the variance above. Let us recapture what's allowed. Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with $E(X_k) = \mu$ and $V(X_k) = \sigma^2$ for $k = 1, 2, \ldots, n$. Then

$$E\left(a_0 + \sum_{k=1}^{n} a_k X_k\right) = a_0 + \sum_{k=1}^{n} a_k E(X_k) = a_0 + \mu \sum_{k=1}^{n} a_k,$$

since the expectation is a *linear* operator (remember that it is either a sum or an integral, both of which are linear). For the variance, it is true that

$$V\left(a_0 + \sum_{k=1}^{n} a_k X_k\right) = /X_k \text{ are independent } / = \sum_{k=1}^{n} a_k^2 V(X_k) = \sigma^2 \sum_{k=1}^{n} a_k^2.$$

The assumption about independence is crucial here. Without it, the sum above will get very messy with covariances all over the place. Note though that for the expectation, independence is not required.

Something neat with the normal distribution is that linear combinations of normally distributed variables are still normally distributed (the mean and variance might change, but we've seen above how that works). So assuming for a minute that $X_k \sim \mathrm{N}(\mu, \sigma^2)$ for $k = 1, 2, \ldots, n$, we have

$$X := \sum_{k=1}^{n} X_k \sim \mathrm{N}(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X} := \frac{1}{n} \sum_{k=1}^{n} X_k \sim \mathrm{N}(\mu, \sigma^2/n).$$

This is clear from the formulas above. Note in particular that

$$V(\overline{X}) = V\left(\frac{1}{n} \sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} \sum_{k=1}^{n} V(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

where we used the independence of the variables. This equality shows that as n grows larger, the variance of $\overline{X}$ tends to zero. Larger sample size means less variance for $\overline{X}$.

Random or not?

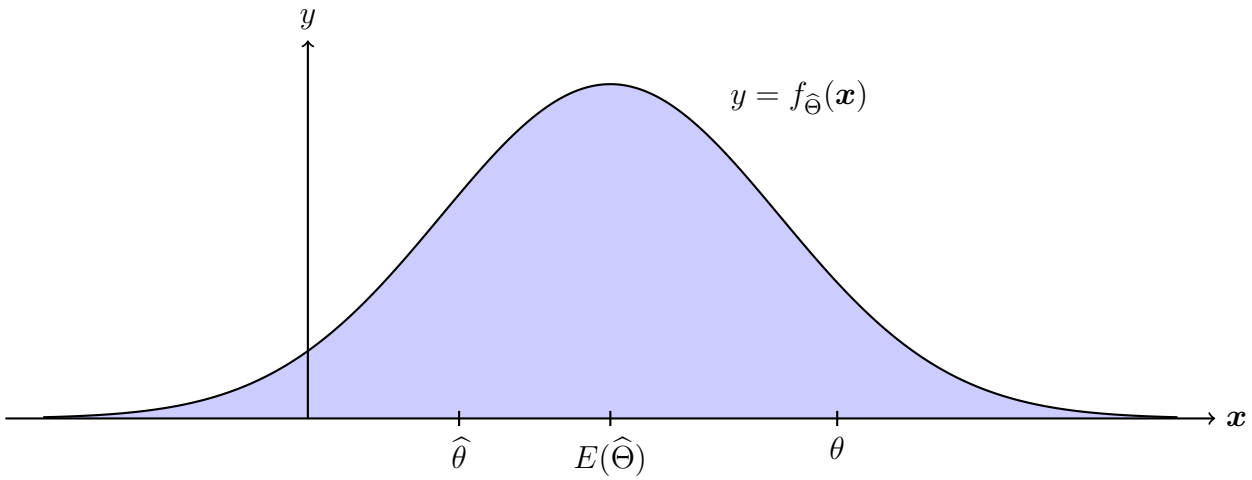
Be careful with explicitly showing the difference between what is random and what is not. We have three quantities:

(i) θ – real value. Unknown but not random.

(ii) $\widehat{\theta}$ – estimate for θ. Known value calculated from the sample $x_1, x_2, \ldots, x_n$.

(iii) $\widehat{\Theta}$ – The estimator. This is a random variable!

If you want to calculate probabilities (so expectations and variances and such), you have to use $\widehat{\Theta}$.



Note the following:

(i) $\widehat{\theta}$ is calculated from observations, so it might end up anywhere basically.

(ii) We do not now if the expectation $E(\widehat{\Theta})$ coincides with the unknown θ, there might be a bias.

(iii) The random variable $\widehat{\Theta}$ is n-dimensional (since it depends on $X_1, X_2, \ldots, X_n$, hence the bold font for x in the graph above) and might also be vector-valued in the case that $\theta \in \mathbf{R}^p$ with $p > 1$. In the case when $p > 1$, the density gets difficult to render on paper though...

A non-biased estimator $\widehat{\Theta}$ of θ will — on average — hit the unknown θ. This is a direct consequence of the law of large numbers. What this means more exactly is that if we were to form the average of estimates (repeating the experiment yielding $x_1, x_2, \ldots, x_n$ and find an estimate $\widehat{\theta}_k$ for each $k = 1, 2, 3, \ldots$), this average will converge to θ with probability one (that's the strong law of large numbers) with reasonable assumptions on the distributions.

Certainty

So we have found and estimator $\widehat{\Theta}$ of some unknown θ. Can we use the information contained in the distribution of $\widehat{\Theta}$ to say something about which values for estimates of θ that are reasonable? I mean, if we look at the graph of the density function (or probability function). we can see where it is likely that observations end up (around points where large amounts of probability mass is accumulated). In other words, can we find a set I such that $\theta \in I$ with some given probability? That was basically question 2 above.
Let's assume that $\theta \in \mathbf{R}$ (so we only have one dimension). A **confidence interval** I with **confidence degree** $1 - \alpha$ $(0 < \alpha < 1)$ is *any* interval such that θ belongs to it with probability $1-\alpha$. The end points of such an interval typically need to be observations of random variables (transformations of $\widehat{\Theta}$). So the systematic question now is how to go from the estimator $\widehat{\Theta}$ — which depends on the unknown parameter unless something unusual occurs — to something with a completely known distribution. Let's look at a couple of examples.

Assume that $x_1, x_2, \ldots, x_n$ is a sample of observations of $X_1, X_2, \ldots, X_n$ of independent identically distributed random variables. The type of distribution is known but not some unknown parameter.

Estimator for the variance. Since we introduced the sample variance, that seems to be a reasonable place to start. Indeed, we have seen that $E(S^2) = \sigma^2$, so it is an unbiased estimator of σ^2. It is also a consistent estimator. To say something more specific, we need to now the type of distribution we're dealing with. Let's assume that we have samples from $N(\mu, \sigma^2)$. Then

$$V := \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$$

according to Cochran's theorem. This is nice, since we now have a distribution that is completely known. To find a confidence interval, we find numbers a and b such that

$$P(a < V < b) = 1 - \alpha.$$

Note that a and b depend on both α and n and that we need to use a table or computer software. We then solve $a < V < b$ for σ^2, obtaining that

$$\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}.$$

To obtain a confidence interval (the above expression is not a fixed interval since the limits are stochastic), we need to estimate all involved random variables. The natural thing to do is to use the sample variance s^2 to estimate S^2. Thus we get the confidence interval

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{b}, \ \frac{(n-1)s^2}{a} \right).$$

Estimator for the expectation. The expectation is where X_k ends up "on average," so a reasonable starting point would be to consider the mean value $\overline{X} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} X_k$. We know that $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$ if $E(X_k) = \mu$ and $V(X_k) = \sigma^2$. To transform our estimator into something with a known distribution, we standardize the variable by removing the mean and dividing by the standard deviation. If we again assume that we have samples from $N(\mu, \sigma^2)$, we see that

$$Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

So if σ is known, this works great. If σ is unknown, we estimate it by s. This means that we use

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

instead, a result that follows from Gosset's theorem. Proceeding as above, we find a number t (both distributions are symmetric with respect to the y-axis) such that

$$P(-t < T < t) = 1 - \alpha$$

in the case when σ is unknown. We find the number t in a table or by using computer software. Note that it depends on both α and n. If σ is known, we use Z and the normal distribution instead. Solving for μ in the inequality, we see that

$$-t < T < t \quad \Leftrightarrow \quad \overline{X} - t\frac{S}{\sqrt{n}} < \mu < \overline{X} + t\frac{S}{\sqrt{n}}.$$

Estimating S by s and $\overline{X}$ by $\overline{x}$, we obtain the confidence interval

$$I_\mu = \left(\overline{x} - t\frac{s}{\sqrt{n}}, \ \overline{x} + t\frac{s}{\sqrt{n}} \right).$$

To be continued...!