

Substitution in integrals and calculation of the probability of error in a mean estimator

May 4, 2010

Here we'll clear up how to calculate an integral using substitution (as per some problems in the exercises) and use this to calculate the probability of error in the mean estimator.

Consider the integral

$$\int_a^b f(t) dt$$

into which we'd like to make the substitution $t = g(a)$. Then the integral becomes

$$\int_\alpha^\beta f(g(a))g'(a) da,$$

where g is a suitable function (probably with some restrictions such as continuity and monotonicity), $g(\alpha) = a$ and $g(\beta) = b$.

Why does this work? Consider the approximation

$$\int_a^b f(t) dt \approx (b - a)f(a),$$

which is valid if $b - a$ is small. Here we will look at the main ideas and not try to make the proof airtight. Consider the same approximation for the integral with the substitution:

$$\int_\alpha^\beta f(g(a))g'(a) da \approx (\beta - \alpha)f(g(\alpha))g'(\alpha).$$

We have $b = g(\beta) = g(\alpha + (\beta - \alpha)) \approx g(\alpha) + (\beta - \alpha)g'(\alpha)$. Because $g(\alpha) = a$, we get $\frac{b-a}{g'(\alpha)} \approx \beta - \alpha$. Inserting this into (1), the second integral becomes $(b - a)f(g(\alpha)) = (b - a)f(a)$, which is the same as (1).

Using this result, we will show that the integral

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mu+x\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

with x a constant, is independent of σ^2 . In other words, the total probability that a normally distributed random variable obtains a value less than, say, the mean plus 2 times the standard deviation is a constant. First we will apply the substitution $t - \mu = a$, or $t = a + \mu$. In this case the limits α and β of integration are solved from

$$\alpha + \mu = -\infty,$$

$$\beta + \mu = \mu + x\sigma,$$

from which we get $\alpha = -\infty$, $\beta = x\sigma$. Furthermore, we have $dt = g'(a) da = da$ and the integral becomes

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x\sigma} \exp\left(-\frac{a^2}{2\sigma^2}\right) da.$$

Next we'll make the substitution $\frac{a}{\sigma} = b$, or $a = b\sigma$, so here we have $g(b) = b\sigma$. The lower and upper limits are solved from $\alpha\sigma = -\infty$ and $\beta\sigma = x\sigma$, respectively. These give the limits $\alpha = -\infty$ and $\beta = x$. In this case, we replace da with σdb and the integral equals

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left(-\frac{(b\sigma)^2}{2\sigma^2}\right) \sigma db = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{b^2}{2}\right) db.$$

Thus we have

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mu+x\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt.$$

The thing to note here is that the integral on the right is independent of μ and σ .

Next, we will use this to solve an exercise we had earlier using a suggestion that it can be solved theoretically:

Consider a set of data points x_1, x_2, \dots, x_N which are independent, normally distributed with mean 1 and variance 2. An estimator \hat{m} for the mean of the distribution (from which the data points are picked) is

$$\hat{m} = \frac{1}{N} \sum_{i=1}^N x_i.$$

At least how many data points do we need if we want the error in the mean estimate \hat{m} to be at most 0.01 in 99% of the cases?

Consider $x \sim \mathcal{N}(0, 1)$, i.e. the random variable x is picked from a normal distribution with mean 0 and variance 1. For what value of a do we have $P(|x| < a) = 0.99$? We can look this up in a table or calculate using Matlab using the function `calc_normal_prob.m` as follows (other ways exist):

```
a = 0.5;
while ( calc_normal_prob( 0, 1, -a,a) < 0.99 ),
    a = a + 0.00001;
end
fprintf( '%2.5f\n', a );
```

which gives us $a = 2.57583$.

Now, consider the estimator of the sample mean. If the distribution of the variables is normal, so is their average, with the same mean. The variance of the estimator is

$$\begin{aligned} \text{Var} \left\{ \frac{1}{N} \sum_{i=1}^N (x_i - \mu) \right\} &= \frac{1}{N^2} \sum_{i=1}^N \text{Var} \{x_i - \mu\} \\ &= \frac{\sigma^2}{N}. \end{aligned}$$

Thus the standard deviation is $\frac{\sigma}{\sqrt{N}}$.

Based on the previous calculations we know that 99% of the probability mass of $\mathcal{N}(0, 1)$ is in $[-2.57583\sigma, 2.57583\sigma]$. Since in this case the variance is 2, we have 99% of the probability in $[1 - 2.57583\sqrt{2}, 1 + 2.57583\sqrt{2}]$. If we average two of these variables, 99% of the probability is in $[1 - \frac{2.57583\sqrt{2}}{\sqrt{2}}, 1 + \frac{2.57583\sqrt{2}}{\sqrt{2}}]$, for 3 variables this is $[1 - \frac{2.57583\sqrt{2}}{\sqrt{3}}, 1 + \frac{2.57583\sqrt{2}}{\sqrt{3}}]$, and so on. The question is, for what value of N do we have 99% of the probability within 0.01 of the mean? This can be solved from

$$\frac{2.57583\sqrt{2}}{\sqrt{N}} = 0.01,$$

which gives $N = \left(\frac{2.57583\sqrt{2}}{0.01}\right)^2 \approx 132698$.