stats.student Student confidence

The central limit theorem tells us that, for large sample size N, the distribution of the means is Gaussian. However, for small sample size, the Gaussian isn't as good of an estimate. **Student's t-distribution** is superior for lower sample size and equivalent at higher sample size.

Technically, if the population standard deviation σ_X is known, even for low sample size we should use the Gaussian distribution. However, this rarely arises in practice, so we can usually get away with an "always t" approach.

A way that the t-distribution accounts for low-N is by having an entirely different distribution for each N (seems a bit of a cheat, to me). Actually, instead of N, it uses the **degrees of freedom** ν, which is N minus the number of parameters required to compute the statistic. Since the standard deviation requires only the mean, for most of our cases, $\nu = N - 1$.

As with the Gaussian distribution, the t-distribution's integral is difficult to calculate. Typically, we will use a t-table, such as the one given in Appendix A.02. There are three points of note.

1. Since we are primarily concerned with going from probability / confidence values (e.g. $P\%$ probability / confidence) to intervals, typically there is a column for each probability.
2. The extra parameter ν takes over one of the dimensions of the table because three-dimensional tables are illegal.
3. Many of these tables are "two-sided," meaning their t-scores and probabilities assume you want the symmetric probability about the mean over the interval $[-t_b, t_b]$, where t_b is your t-score bound.

Consider the following example.

Example stats.student-1 re: student confidence interval

Write a *Matlab* script to generate a data set with 200 samples and sample sizes $N \in \{10, 20, 100\}$ using any old distribution. Compare the distribution of the means for the different N. Use the sample distributions and a t-table to compute 99% confidence intervals.

Generate the data set.

```
confidence = 0.99; % requirement

M = 200; % # of samples
N_a = [10,20,100]; % sample sizes

mu = 27; % population mean
sigma = 9; % population std

rng(1) % seed random number generator
data_a = mu + sigma*randn(N_a(end),M); % normal
size(data_a) % check size
data_a(1:10,1:5) % check 10 rows and five columns
```

```
ans =

   100   200


ans =

   21.1589   30.2894   27.8705   30.7835   28.3662
   37.6305   17.1264   28.2973   24.0811   34.3486
   20.1739   44.3719   43.7059   39.0699   32.2002
   17.0135   32.6064   36.9030   37.9230   36.5747
   19.3900   32.9156   23.7230   22.4749   19.7709
   21.8460   13.8295   31.2479   16.9527   34.1876
   21.9719   34.6854   19.4480   18.7014   24.1642
   28.6054   32.2244   22.2873   26.9906   37.6746
   25.2282   18.7326   14.5011   28.3814   27.7645
   32.2780   34.1538   27.0382   18.8643   14.1752
```

Compute the means for different sample sizes.

```
mu_a = NaN*ones(length(N_a),M);
for i = 1:length(N_a)
    mu_a(i,:) = mean(data_a(1:N_a(i),1:M),1);
end
```

Plotting the distribution of the means yields Figure student.1.

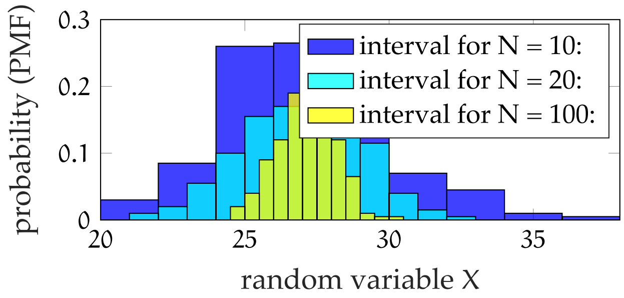

Figure student.1: a histogram showing the distribution of the means for each sample size.

It makes sense that the larger the sample size, the smaller the spread. A quantitative metric for the spread is, of course, the standard deviation of the means for each sample size.

```
S_mu = std(mu_a,0,2)
```

```
S_mu =

    2.8365
    2.0918
    1.0097
```

Look up t-table values or use Matlab's `tinv` for different sample sizes and 99% confidence. Use these, the mean of means, and the standard deviation of means to compute the 99% confidence interval for each N.

```
t_a = tinv(confidence,N_a-1)
for i = 1:length(N_a)
    interval = mean(mu_a(i,:)) + ...
      [-1,1]*t_a(i)*S_mu(i);
    disp(sprintf('interval for N = %i: ',N_a(i)))
    disp(interval)
end
```

```
t_a =

    2.8214    2.5395    2.3646

interval for N = 10:
   19.0942   35.1000
```

```
interval for N = 20:
   21.6292   32.2535

interval for N = 100:
   24.7036   29.4787
```

As expected, the larger the sample size, the smaller the interval over which we have 99% confidence in the estimate.