Unbiased estimators

In statistics we often endeavour to infer a *parameter* of a overall *population* from a **sample** i.e. a finite selection of data. This is the basis of experimental science (we make a measurement and then try and compare it to theoretical or agreed values) and indeed the concept of opinion polling.

We shall restrict ourselves to two important statistical parameters: the **population mean** μ and **standard deviation** σ.

$\{x_i\}$ Set of data in a sample. The sample has n elements

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ Sample mean

Let us firstly consider the **expected value** of the **sample mean**

$$E[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

The sample mean is therefore an **unbiased estimator** of the true population mean.

$$E[\bar{x}] = \mu$$

Now consider the **sample variance** $$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$\therefore E[S^2] = \frac{1}{n}\sum_{i=1}^{n} E[x_i^2] - E[\bar{x}^2]$$

From the definition of variance:

$$V[x_i] = E[x_i^2] - (E[x_i])^2$$

$$\therefore E[x_i^2] = V[x_i] + (E[x_i])^2 = \sigma^2 + \mu^2$$

$$\therefore E[\bar{x}^2] = V[\bar{x}] + (E[\bar{x}])^2 = V[\bar{x}] + \mu^2$$

$$\therefore E[\bar{x}^2] = V\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] + \mu^2 = V\left[\sum_{i=1}^{n}\frac{1}{n}x_i\right] + \mu^2$$

$$\therefore E[\bar{x}^2] = \sum_{i=1}^{n}\frac{1}{n^2}V[x_i] + \mu^2 = \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2 + \mu^2$$

Assuming sample values are *independent* random variables

$$\therefore E[\bar{x}^2] = \frac{\sigma^2}{n} + \mu^2$$

$$V[Ax + By + ...] = A^2V[x] + B^2V[y] + ...$$

Johann Carl Friedrich Gauss 1777–1855

If $x \sim \mathrm{N}(\mu,\sigma^2)$ the probability of a random variable having value between x and $x+dx$ is given by $p(x)dx$, where:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Hence:

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n} E[x_i^2] - E[\bar{x}^2]$$

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{\sigma^2}{n} - \mu^2$$

$$E[S^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$E[S^2] = \sigma^2\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}\sigma^2$$

$$\therefore E\left[\frac{n}{n-1}S^2\right] = \sigma^2$$

So an **unbiased estimator** of the population variance is:

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$$

Note this can also be written as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The Central Limit Theorem* states that, if the number of elements n in a sample are large enough, the **distribution of sample means will tend to a *Normal distribution*** $\bar{x} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Note the form of the **population distribution** *doesn't matter!* For large n (typically 30 seems to be the agreed minimum) we can determine a **confidence interval** for population mean μ based upon sample data.

First we define a random variable which will be distributed by $\mathrm{N}(0,1)$ i.e. variance s^2/n

$$z = \frac{\mu - \bar{x}}{\sqrt{s^2/n}}$$

We then find the z limits such that $P(-z_* \le z \le z_*) = a$

a is the 'significance level' e.g. 0.95.

Note this is called a 'two tail test' as the sample mean could be either side of the true mean. A one-tail test would be $P(z \le z_*) = a$ or $P(z \ge z_*) = a$

$$P(z \le z_*) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_*} e^{-\frac{1}{2}z^2}dz$$

$$P(z \le z_*) = \tfrac{1}{2} + \tfrac{1}{2}\mathrm{erf}\left(\tfrac{1}{\sqrt{2}}z_*\right)$$

$$\therefore z_* = \sqrt{2}\,\mathrm{erf}^{-1}(2A-1)$$

$$A = a + \frac{1-a}{2} = \tfrac{1}{2}(a+1)$$

$$\therefore z_* = \sqrt{2} \times \mathrm{erf}^{-1}a$$

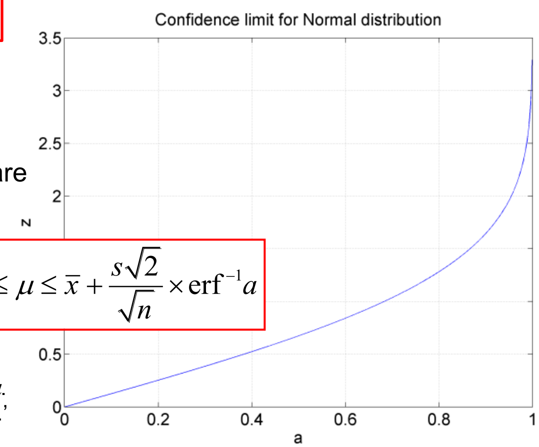
$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

The **Error Function**
A *Special Function* which is readily evaluated in computer software like MATLAB.

Confidence limits for the population mean are therefore:

$$\bar{x} - \frac{s\sqrt{2}}{\sqrt{n}} \times \mathrm{erf}^{-1}a \le \mu \le \bar{x} + \frac{s\sqrt{2}}{\sqrt{n}} \times \mathrm{erf}^{-1}a$$

This is a 'g-test' for the population mean μ. 'g' meaning 'Gaussian.'



Of course it may not be possible to obtain thirty or more samples. What then? William “Student” Sealy Gosset developed the **t-test.** It follows a very similar recipe to the ‘g-test’, but involves a generalization to the standard Normal distribution. The **t-distribution** actually tends to $\mathrm{N}(0,1)$ when n becomes large. This is where the practical limit of $n = 30$ is determined. Beyond this number it is difficult to distinguish the distributions.

$$t = \frac{\mu - \bar{x}}{\sqrt{s^2/n}}$$

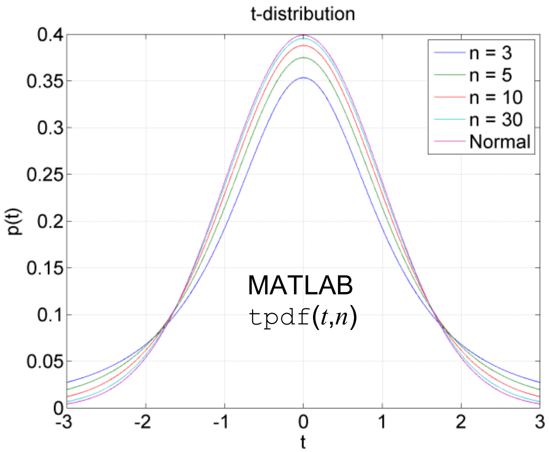
$$p(t,n) = \frac{\Gamma\left(\frac{1}{2}(v+1)\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{1}{2}v\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad v = n-1$$

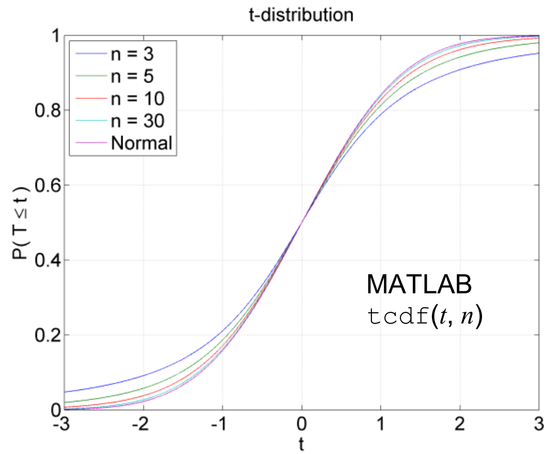
t-distribution

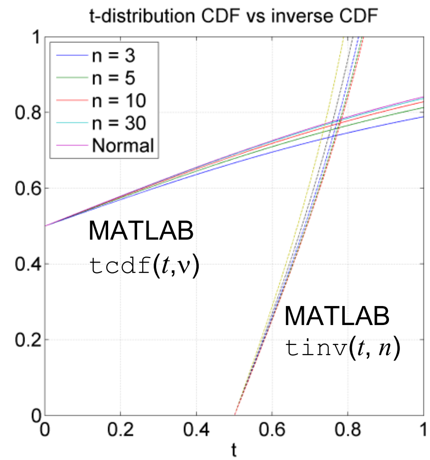
$$\Gamma(x) = \int_0^\infty z^{x-1}e^{-z}dz$$

MATLAB `gamma(x)`

William “Student”
Sealy Gosset 1876-1937
Worked for Guinness and
was educated at
Winchester College







We want to find the t limits such that $P\left(-t_* \leq t \leq t_*\right) = a$

a is the ‘significance’ e.g. 0.95.

$$P\left(t \leq t_*\right) = \int_{-\infty}^{t_*} p(t,n)dt = \Theta(t_*,n)$$

$$t_* = \Theta^{-1}\left(A,n\right)$$

$$A = a + \frac{1-a}{2} = \tfrac{1}{2}\left(a+1\right)$$

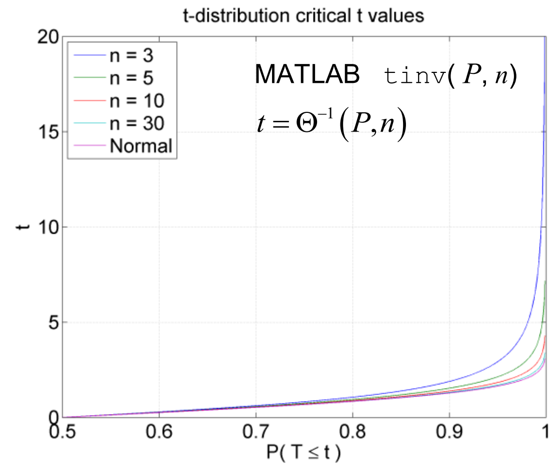
Confidence limits for the population mean are therefore:

$$\bar{x} - \frac{s}{\sqrt{n}} \times \Theta^{-1}\left(\tfrac{1}{2}\left(a+1\right),n\right) \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}} \times \Theta^{-1}\left(\tfrac{1}{2}\left(a+1\right),n\right)$$

This is a ‘t-test’ for the population mean μ.

Note: To use a **t-distribution** one must also **assume the population distribution is Normal**. If this is unknown *a priori,* then a test for ‘Normality’ should be performed on a suitable sample of data, *before* a t-test is used. e.g. a **Kolmogorov–Smirnov (KS) ‘nonparametric’ test.**

If the sample size n is large enough, then the *Central Limit Theorem* means the ‘g-test’ is applicable to samples from *all* population distributions.



Worked example:

Data generated from a Normal distribution with mean $\mu = 42$ and standard deviation $\sigma = 5$

37.6817 42.3868 35.9294 36.4325 41.9658
49.6632 38.1517 43.8569 40.8721 47.5868

Data sample consists of $n = 10$ elements

Unbiased mean estimate

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 41.4527$$

Unbiased standard deviation estimate

$$s = \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2} = 4.6322$$

Significance level $a = 0.95$

Note the t-distribution has 'fatter tails' than the standard Normal distribution

The idea of the confidence interval is essentially:
"Based upon a data sample, what range of values to we expect the population mean to be within?

If we **hypothesize a value for the population mean** (e.g. from some theoretical calculation or prior knowledge) then our confidence interval forms the basis of a **test of the hypothesis**.

'g-test' confidence limits

$$\bar{x} - \frac{s\sqrt{2}}{\sqrt{10}} \times \mathrm{erf}^{-1} 0.95 \le \mu \le \bar{x} + \frac{s\sqrt{2}}{\sqrt{10}} \times \mathrm{erf}^{-1} 0.95$$

$$41.4527 - \frac{4.6322}{\sqrt{10}} \times 1.96 \le \mu \le 41.4527 + \frac{4.6322}{\sqrt{10}} \times 1.96$$

$$38.582 \le \mu \le 44.324$$

Hence *hypothesis* that population mean is $\mu = 42$ passes the 'g-test'

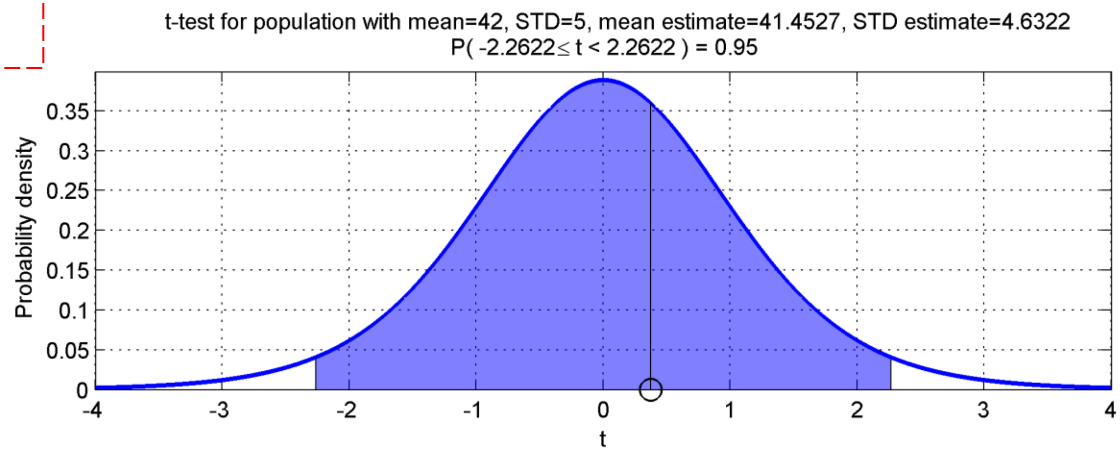
't-test' confidence limits **Note**: Population distribution is assumed to be Normal

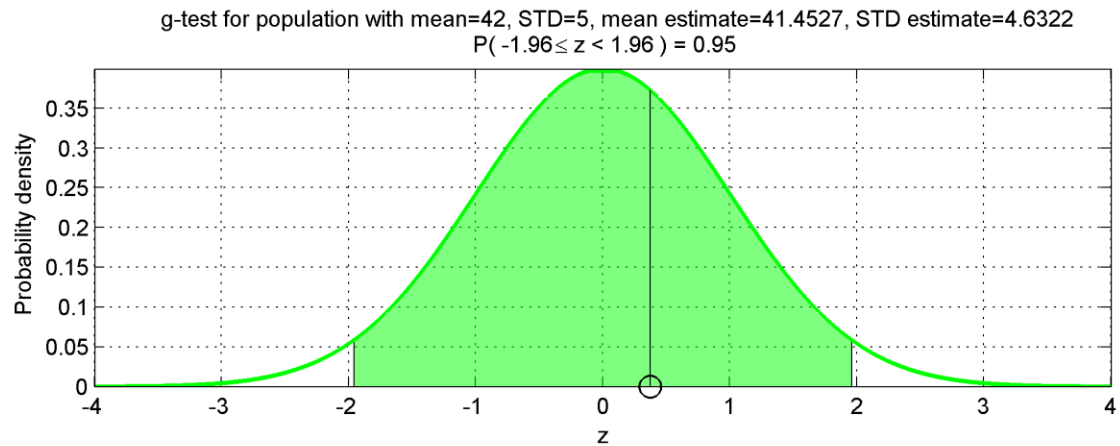
$$\bar{x} - \frac{s}{\sqrt{n}} \times \Theta^{-1}\left(\tfrac{1}{2}(a+1), 10\right) \le \mu \le \bar{x} + \frac{s}{\sqrt{n}} \times \Theta^{-1}\left(\tfrac{1}{2}(a+1), 10\right)$$

$$41.4527 - \frac{4.6322}{\sqrt{10}} \times 2.2622 \le \mu \le 41.4527 + \frac{4.6322}{\sqrt{10}} \times 2.2622$$

$$38.139 \le \mu \le 44.767$$

Hence *hypothesis* that population mean is $\mu = 42$ passes the t-test





Critical t values of Student t distribution $t_* = \Theta^{-1}(P)$

P values

v = n-1	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
1	1	1.376382	1.962611	3.077684	6.313752	12.7062	31.82052	63.65674	127.3213	318.3088	636.6192
2	0.816497	1.06066	1.386207	1.885618	2.919986	4.302653	6.964557	9.924843	14.08905	22.32712	31.59905
3	0.764892	0.978472	1.249778	1.637744	2.353363	3.182446	4.540703	5.840909	7.453319	10.21453	12.92398
4	0.740697	0.940965	1.189567	1.533206	2.131847	2.776445	3.746947	4.604095	5.597568	7.173182	8.610302
5	0.726687	0.919544	1.155767	1.475884	2.015048	2.570582	3.36493	4.032143	4.773341	5.89343	6.868827
6	0.717558	0.905703	1.134157	1.439756	1.94318	2.446912	3.142668	3.707428	4.316827	5.207626	5.958816
7	0.711142	0.89603	1.119159	1.414924	1.894579	2.364624	2.997952	3.499483	4.029337	4.78529	5.407883
8	0.706387	0.88889	1.108145	1.396815	1.859548	2.306004	2.896459	3.355387	3.832519	4.500791	5.041305
9	0.702722	0.883404	1.099716	1.383029	1.833113	2.262157	2.821438	3.249836	3.689662	4.296806	4.780913
10	0.699812	0.879058	1.093058	1.372184	1.812461	2.228139	2.763769	3.169273	3.581406	4.1437	4.586894
11	0.697445	0.87553	1.087666	1.36343	1.795885	2.200985	2.718079	3.105807	3.496614	4.024701	4.436979
12	0.695483	0.872609	1.083211	1.356217	1.782288	2.178813	2.680998	3.05454	3.428444	3.929633	4.317791
13	0.693829	0.870152	1.079469	1.350171	1.770933	2.160369	2.650309	3.012276	3.372468	3.851982	4.220832
14	0.692417	0.868055	1.07628	1.34503	1.76131	2.144787	2.624494	2.976843	3.325696	3.78739	4.140454
15	0.691197	0.866245	1.073531	1.340606	1.75305	2.13145	2.60248	2.946713	3.286039	3.732834	4.072765
16	0.690132	0.864667	1.071137	1.336757	1.745884	2.119905	2.583487	2.920782	3.251993	3.686155	4.014996
17	0.689195	0.863279	1.069033	1.333379	1.739607	2.109816	2.566934	2.898231	3.22245	3.645767	3.965126
18	0.688364	0.862049	1.06717	1.330391	1.734064	2.100922	2.55238	2.87844	3.196574	3.610485	3.921646
19	0.687621	0.860951	1.065507	1.327728	1.729133	2.093024	2.539483	2.860935	3.173725	3.5794	3.883406
20	0.686954	0.859964	1.064016	1.325341	1.724718	2.085963	2.527977	2.84534	3.153401	3.551808	3.849516
21	0.686352	0.859074	1.06267	1.323188	1.720743	2.079614	2.517648	2.83136	3.135206	3.527154	3.819277
22	0.685805	0.858266	1.061449	1.321237	1.717144	2.073873	2.508325	2.818756	3.118824	3.504992	3.792131
23	0.685306	0.85753	1.060337	1.31946	1.713872	2.068658	2.499867	2.807336	3.103997	3.484964	3.767627
24	0.68485	0.856855	1.059319	1.317836	1.710882	2.063899	2.492159	2.79694	3.090514	3.466777	3.745399
25	0.68443	0.856236	1.058384	1.316345	1.708141	2.059539	2.485107	2.787436	3.078199	3.450189	3.725144
26	0.684043	0.855665	1.057523	1.314972	1.705618	2.055529	2.47863	2.778715	3.066909	3.434997	3.706612
27	0.683685	0.855137	1.056727	1.313703	1.703288	2.051831	2.47266	2.770683	3.05652	3.421034	3.689592
28	0.683353	0.854647	1.055989	1.312527	1.701131	2.048407	2.46714	2.763262	3.046929	3.408155	3.673906
29	0.683044	0.854192	1.055302	1.311434	1.699127	2.04523	2.462021	2.756386	3.038047	3.39624	3.659405
39	0.680833	0.850935	1.050399	1.303639	1.684875	2.022691	2.425841	2.707913	2.975609	3.312788	3.55812
49	0.67953	0.849018	1.047519	1.299069	1.676551	2.009575	2.404892	2.679952	2.93973	3.265079	3.500443
59	0.678671	0.847756	1.045623	1.296066	1.671093	2.000995	2.391229	2.661759	2.91644	3.234207	3.46321
79	0.677608	0.846195	1.043282	1.29236	1.664371	1.99045	2.374482	2.639505	2.888011	3.196628	3.417985
99	0.676976	0.845267	1.041891	1.290161	1.660391	1.984217	2.364606	2.626405	2.871308	3.174604	3.391529
119	0.676557	0.844652	1.04097	1.288706	1.657759	1.9801	2.358093	2.617776	2.860317	3.160133	3.374167

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$t = \frac{\mu - \bar{x}}{\sqrt{s^2/n}}$$

$$p(t) = \frac{\Gamma\left(\frac{1}{2}(v+1)\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{1}{2}v\right)}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \qquad v = n-1$$

$$\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} dz$$

$$P\left(t \le t_*\right) = \int_{-\infty}^{t_*} p(t)dt = \Theta(t_*)$$

$$t_* = \Theta^{-1}\left(P, n\right)$$

$$P = \tfrac{1}{2}\left(a+1\right)$$

$$t_* = \Theta^{-1}\left(P, n\right)$$

MATLAB `tinv(` P, N `)`

```
function t = tinv_calc(PHI,N)
v = N-1; %Degrees of freedom of
         % t-distribution
i = find(PHI<0.5); PHI(i) = 1 - PHI(i);
y = betainc( 1, v/2,0.5 ) - (2*PHI-1);
x = betaincinv(y, v/2, 0.5 );
t = sqrt( v./x - v ); t(i) = -t(i);
```

$$\beta(x,z,w) = \frac{\Gamma(z+w)}{\Gamma(z)\Gamma(w)}\int_0^x k^{z-1}(1-k)^{w-1} dk$$

Incomplete beta function