Sampling Distributions

- **Normal distribution**
- **Chi-square distribution**
- **Student's t-distribution**

A ***statistic*** is any function of sample data $X_1, X_2, \ldots X_n$.

For example,

- the ***sample mean***: $$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
- the ***sample variance***: $$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

The probability distribution of a statistic is called a ***sampling distribution.***

If X_i each have mean μ and variance σ^2, then

(by the Central Limit Theorem)

the *sample mean* has approximately *normal* distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

If each X_i has $N(0,1)$ distribution, then the statistic

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2$$

has the distribution known as

chi-square with n degrees of freedom.

with density function

$$f(z) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} z^{\left(n/2 - 1\right)} e^{-z/2}$$

for $z>0$

Warning! Don't try to memorize this formula!

The *mean* is n and *variance* is $2n$.

Gamma function Γ is a generalization of the factorial function, where $\Gamma(n)=(n-1)!$ if n is an integer.

Use of Chi-Square Distribution

Suppose that $X_i \sim N\left(\mu,\sigma^2\right)$.

Then the statistic

$$\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}$$

or, since the sample variance is

$$S^2 \equiv \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1} \Rightarrow \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \left(n-1\right)S^2$$

we have the result

$$\frac{\left(n-1\right)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Student's t-distribution

(Note: "Student" was a pseudonym of an English chemist, W.S. Gosset, in a 1908 publication.)

If $X \sim N\left(\mu,\sigma^2\right)$ and $Y \sim \chi^2_k$, then the statistic t_k defined by

$$t_k = \frac{X}{\sqrt{Y/k}}$$

has what is called

"Student's t-distribution" with k degrees of freedom.

The density function of t_k is extremely messy:

$$f\left(t\right) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

Student's t-distribution

The *mean* is *0* and

the *variance* is $\frac{k}{k-2}$ for k>2.

As k→∞, the t-distribution reduces to the standard normal distribution.

Gamma function Γ is a generalization of the factorial function, where Γ(n)=(n-1)! if n is an integer.

Suppose that X_1, X_2, ...X_n all have $N\left(\mu,\sigma^2\right)$ distribution, and we compute the sample mean $\bar{X}$ and sample variance S^2.

Then the statistic

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

has t-distribution with n-1 degrees of freedom.

Application of Student's t-distribution

Confidence Intervals of Mean Performance Measures

Suppose that we perform a limited number, *n*, of replications of a simulation, obtaining some performance measure X_i in the i^{th} replication. The sequence $\{X_1, X_2, \ldots X_n\}$ are independent & identically-distributed (i.i.d.) random variables, but the mean and variance are unknown.

How good an estimate of the true mean is the sample mean?

We estimate the expected performance of the system by the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

If we use more replications (i.e., larger *n*), we would expect a better estimate, of course.

What can we say about the true expected value μ of the performance X of the simulated system?

Given an α, we want β such that the probability that μ is in the *confidence* interval

$$\bar{X} \pm \beta, \text{ i.e., } \quad \mu \in \left[\bar{X} - \beta,\ \bar{X} + \beta\right]$$

to be at least 1– α, i.e.,

$$P\left\{\left|\bar{X} - \mu\right| > \beta\right\} \le \alpha$$

The sample variance is

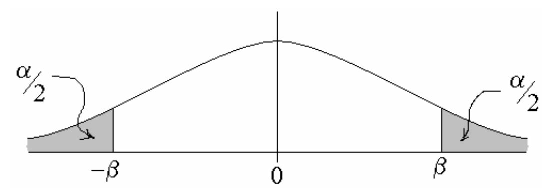
$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

Clearly, the larger the sample variance, the less confident we can be that $\bar{X}$ is a good estimate of μ, i.e., the larger we can expect to be β.

Student's t-distribution

The appropriate value of β is chosen so that, for the t-distribution with n-1 degrees of freedom,

$$P\left\{\left|t_{n-1}\right| > \beta\right\} = P\left\{t_{n-1} > \beta \text{ OR } t_{n-1} < \beta\right\} = \alpha$$



← The total shaded area is α.

There are tables available for various degrees of freedom k and probabilities α = 10%, 5%, 1%, 0.1%, etc.

Example: Suppose that we perform 10 simulations, and obtain the following average values of X:

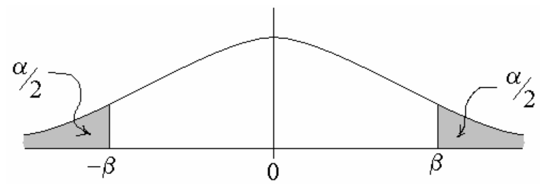
```
        4.58578
        4.56717
        4.99381
        5.43874
        4.9137
        4.41366
        5.28951
        4.96028
        5.70154
        4.82989
Average: 4.96941
```

The sample variance is $S^2 = 0.149982 \quad \Rightarrow \quad S = 0.387275$

How close to the true mean μ of X is this sample mean 4.96941?

Choose α = 5%. Consulting a table of Student's t-distribution, we find that, for $n-1$ = 9 degrees of freedom,

Degrees of freedom k	α = 10%	α = 5%	α = 1%
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169



$$t_{\alpha/2,\ n-1} = t_{5\%\ ,9} = 2.262$$

so that the confidence interval is $\left[\bar{X} - t_{5\%,9}\dfrac{S}{\sqrt{10}}, \bar{X} + t_{5\%,9}\dfrac{S}{\sqrt{10}}\right]$

The 95% *confidence interval* for the mean value μ is

$$4.96941 \pm 2.262 \times \frac{0.387275}{10} \quad = \quad 4.96941 \pm 0.0876016$$

i.e, we have 95% confidence that

$$\mu \in [4.88181,\ 5.05701]$$

Note: these values of X were in actuality sampled from a $N(5,1)$ distribution!