LECTURE 04

SAMPLING DISTRIBUTIONS AND CENTRAL LIMIT THEOREM - III

Outline of today's lecture:

C. Student's t Distribution

Theorem 7.1 tells us that $\sqrt{n}\left(\bar{Y}-\mu\right)/\sigma$ has a standard normal distribution. When σ is unknown, σ can be estimated by $S=\sqrt{S^2}$, and the quantity

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right)$$

provides the basis for developing methods for inferences about population mean, μ.

Definition 7.2 Let Z be a standard normal variable, and let W be a chi-square distributed variable with *v* degrees of freedom. Then, if Z and W are independent:

$$T=\frac{Z}{\sqrt{W/v}}$$

is said to have a *t distribution* with v degrees of freedom (df).

- By Theorem 7.1, we know that $Z = \sqrt{n}\left(\bar{Y} - \mu\right)/\sigma$ has a standard *normal* distribution.
- By Theorem 7.3, we know that $W = \dfrac{(n-1)S^2}{\sigma^2}$ has a *chi-square* distribution with *v*=n-1 *degrees of freedom* (Z and W are independent since $\bar{Y}$ and S^2 are independent).

Thus, we can write:

$$T = \frac{Z}{\sqrt{W/v}}$$

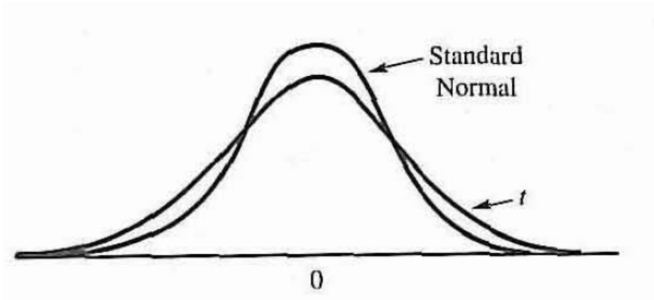
$$T = \frac{\sqrt{n}\left(\bar{Y} - \mu\right)/\sigma}{\sqrt{\left[(n-1)S^2/\sigma^2\right]/v}}$$

$$T = \frac{\sqrt{n}\left(\bar{Y} - \mu\right)/\sigma}{\sqrt{\left[(n-1)S^2/\sigma^2\right]/(n-1)}} = \frac{\dfrac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\sigma}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\dfrac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\sigma}}{\sqrt{\dfrac{S^2}{\sigma^2}}} = \frac{\dfrac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\not{\sigma}}}{\dfrac{S}{\not{\sigma}}}$$

$$T = \frac{\sqrt{n}\left(\bar{Y} - \mu\right)}{S}$$

has a <u>*t distribution*</u> with (*n-1*) degrees of freedom.

We will not give the equation for *t* density function here. But note that like the standard normal density function, the *t* density function is *symmetric* about zero.



Comparison of the standard normal and t density functions.

Example 7.5 The tensile strength for a type of wire is normally distributed with unknown mean μ and unknown variance σ^2.

- Six pieces of wire were randomly selected from a large roll; and Y_i, the tensile strength for portion i, is measured for $i = 1, 2, \ldots, 6$.
- The population mean μ and variance σ^2 can be estimated by $\bar{Y}$ and S^2, respectively.
    - Because $\sigma_{\bar{Y}}^2 = \dfrac{\sigma^2}{n}$, it follows that $\sigma_{\bar{Y}}^2$ can be estimated by $\dfrac{S^2}{n}$.
- Find the approximate probability that $\bar{Y}$ will be within $2\sigma_{\bar{Y}} = \dfrac{2S}{\sqrt{n}}$ of the true population mean μ.

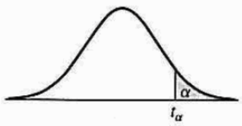
Solution

$$P\left[-\frac{2S}{\sqrt{n}} \le \left(\bar{Y} - \mu\right) \le \frac{2S}{\sqrt{n}}\right] = P\left[-2 \le \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \le 2\right] = P\left[-2 \le T \le 2\right]$$

where T has a *t distribution* with *n-1=5* degrees of freedom (*df*).

If we look at the *t distribution table*, we see that, for 5 df, the upper-tail area to the right of 2.015 is 0.05.

Table 5. Percentage points of the *t* distributions



$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	d.f.
3.078	6.314	12.706	31.821	63.657	1
1.886	2.920	4.303	6.965	9.925	2
1.638	2.353	3.182	4.541	5.841	3
1.533	2.132	2.776	3.747	4.604	4
1.476	2.015	[illegible]71	3.365	4.032	5
1.440	1.943	2.447	3.143	3.707	6
1.415	1.895	2.365	2.998	3.499	7
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Hence, $P[-2.015 \le T \le 2.015] = 0.90$. So, the probability that $\bar{Y}$ will be within two estimated standard deviations of μ is slightly less than 0.90.

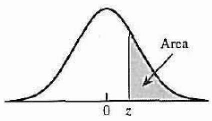
Recall that, if σ^2 were known, the probability that $\bar{Y}$ will fall within $2\sigma_{\bar{Y}}$ of μ would be given by:

$$P\left[-\frac{2\sigma}{\sqrt{n}} \le (\bar{Y}-\mu) \le \frac{2\sigma}{\sqrt{n}}\right] = P\left[-2 \le \sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \le 2\right] = P[-2 \le Z \le 2]$$

$$P[-2 \le Z \le 2] = 1 - 2P[Z > 2] = 1 - 2(0.0228) = 1 - 0.0456 = 0.9544$$

792 Appendix Three

Table 4. Normal curve areas
Standard normal probability in right-hand tail
(for negative values of *z* areas are found by symmetry)



	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	[illegible]	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	[illegible]	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

D. F Distribution

Suppose that we want to compare the variances of two normal populations based on information contained in independent random samples from the two populations. Samples of sizes n_1 and n_2 are taken from the two populations that have variances, σ_1^2 and σ_2^2, respectively.

If we calculate S_1^2 from the observations in sample 1, then S_1^2 estimates σ_1^2. Similarly, S_2^2, calculated from the observations in the second sample estimates σ_2^2 .

Thus, it seems intuitive that the ratio $\frac{S_1^2}{S_2^2}$ could be used to make inferences about the relative magnitudes of σ_1^2 and σ_2^2.

- If we divide each S_i^2 by σ_i^2, then the resulting ratio $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{S_1^2}{S_2^2}\right)$ has an F distribution with (n_1-1) *numerator* degrees of freedom and (n_2-1) *denominator* degrees of freedom.

The general definition of a random variable that possesses an F distribution is given as follows:

Definition 7.3 Let W_1 and W_2 be independent χ^2-distributed random variables with v_1 and v_2 degrees of freedom, respectively. Then,

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to have an *F distribution* with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom.

Considering the independent random samples from normal distributions, we know that $W_1 = (n_1 - 1)S_1^2/\sigma_1^2$ and

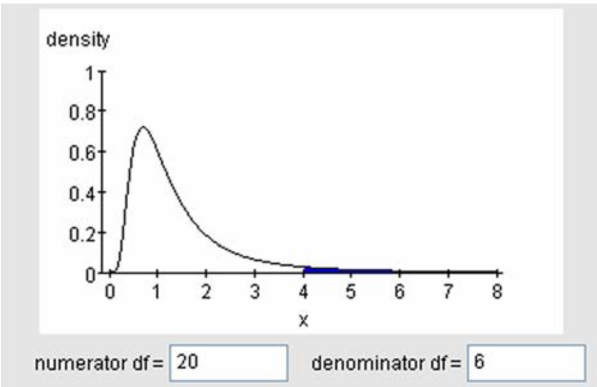
- By Theorem 7.3, we know that $W_1 = \frac{(n-1)S_1^2}{\sigma_1^2}$ and $W_2 = \frac{(n-1)S_2^2}{\sigma_2^2}$ have independent *chi-square* distributions with v_1=(n_1-1) and v_2=(n_2-1) *degrees of freedom*, respectively.

Hence, we can write

$$F = \frac{W_1/v_1}{W_2/v_2} = \frac{\left[(n_1-1)S_1^2/\sigma_1^2\right]/(n_1-1)}{\left[(n_2-1)S_2^2/\sigma_2^2\right]/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with (n_1-1) numerator degrees of freedom and (n_2-1) denominator degrees of freedom.

A typical *F* density function is sketched below.



Source: http://www.stat.tamu.edu/~west/applets/fdemo.html

In an *F Table*, the column headings are the numerator degrees of freedom, whereas the denominator degrees of freedom are given in the main-row headings.

For example, if the F variable of interest has 3 numerator degrees of freedom and 5 denominator degrees of freedom, then $F_{0.10}^{3,5} = 3.62$, $F_{0.05}^{3,5} = 5.41$, $F_{0.025}^{3,5} = 7.76$, and $F_{0.01}^{3,5} = 12.06$.

- Thus, if *F* has an *F* distribution with 3 numerator degrees of freedom and 5 denominator degrees of freedom, then: $P[F > 12.06] = 0.01$.

Example 7.6 If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find the number *b* such that $P\left(\frac{S_1^2}{S_2^2} \le b\right) = 0.95$

Solution

Because n_1=6 and n_2=10 and the population variances are equal, then

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2}$$

has an *F* distribution with $v_1=n_1-1=5$ numerator degrees of freedom and $v_2=n_2-1=9$ denominator degrees of freedom. Also, note that

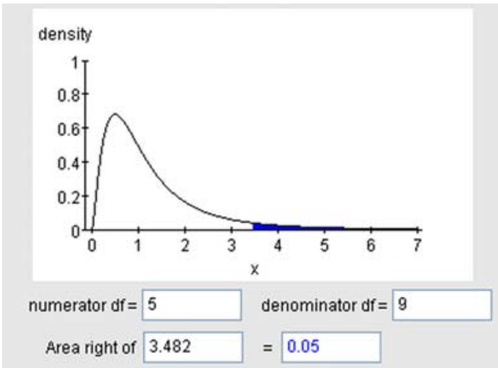
$$P\left(\frac{S_1^2}{S_2^2} \le b\right) = 1 - P\left(\frac{S_1^2}{S_2^2} > b\right) = 0.95$$

Therefore, we want to find the number *b* cutting off an upper-tail area of 0.05 under the *F* density function with 5 numerator degrees of freedom and 9 denominator degrees of freedom. That is, we want to find the number *b* such that $P\left(\frac{S_1^2}{S_2^2} > b\right) = 0.05$.

Looking in column 5 and row 9 in F Table, we see that the appropriate value of *b* is 3.48.

Table 7. (*Continued*)

Denominator d.f.	F_α Numerator d.f. α	1	2	3	4	5	6	7	8	9
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	[illegible]	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	[illegible]	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35



Even when the population variances are equal, the probability that the ratio of the sample variances exceeds 3.48 is still 0.05 (assuming sample sizes of $n_1 = 6$ and $n_2 = 10$).

III. Normal Distribution Functions in Excel

A. NORMDIST Function

The PDF and CDF for the normal distribution can be calculated in Excel using the NORMDIST function. NORMDIST gives the probability that a number falls at or below a given value of a normal distribution.

NORMDIST(x, mean, standard_dev, cumulative)

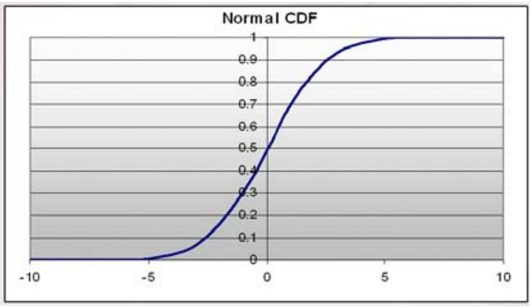
- *x* → The value you want to test.
- *mean* → The average value of the distribution.
- *standard_dev* → The standard deviation of the distribution.
- *cumulative* → If you write FALSE (or 0), it returns the probability density value (PDF); if you write TRUE (or 1), it returns the cumulative probability (CDF).
  - Note: To obtain the required Z values, CDF formula is used, since what we need to find is the area (cumulative probability) left or right of a particular point, *X*.

- The NORMDIST parameters, *x*, *mean* and *standard_dev*, are numeric values, where the parameter, *cumulative*, is a logical TRUE or FALSE value. *Standard_dev* must be greater than 0, but there is no similar requirement for *x* or *mean*.

To sum up:

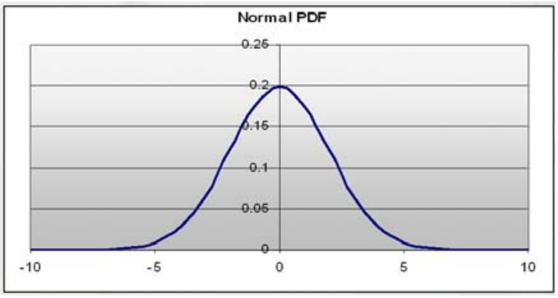
- To calculate the CDF (*cumulative density function*) for a value *X* we use the formula:

  =NORMDIST(X, mean, standard_dev, TRUE)



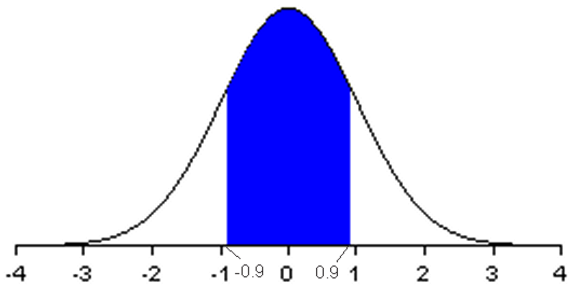
- To calculate the PDF (*probability density function*) for a value *X* we use the formula:

  =NORMDIST(X, mean, standard_dev, FALSE)



- **Note!** The *standard normal distribution* is the special case where *mean = 0* and *standard_dev = 1*.
  - o Hence, when we calculate the *table value of standard normal distribution Z*, we will use this special case where *mean = 0* and *standard_dev = 1*.

Example Recall Example 7.1 in which you are asked to find $P(-0.9 \le Z \le 0.9)$, that is the area between -0.9 and 0.9, including them.



In Excel, we can obtain the probability $P(Z \le 0.9)$ as follows:

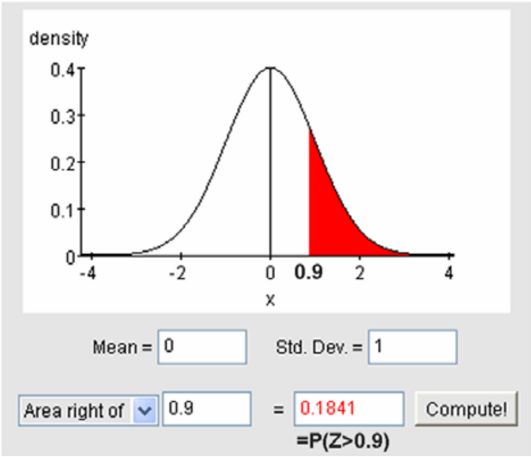
- $P(Z \le 0.9) =$ *NORMDIST(0.9, 0, 1, TRUE)* = 0.815939875 [1]

- This is the area for $P(Z \le 0.9)$ but we want the value for $P(Z > 0.9)$ so we will calculate $P(Z > 0.9)$ as $P(Z > 0.9) = 1 - P(Z \le 0.9)$
    - ➢ This would produce $P(Z > 0.9) = 1 - 0.8159 = 0.1841$

- Hence, the probability that *Z* is higher than 0.9 is 1-0.8159, or approximately 0.1841. This value is represented by the shaded area in the chart below.

---

[1] Or equivalently, $P(Z \le 0.9)$=*NORMSDIST(0,9)* = 0.815939875. Note that *NORMSDIST* gives the cumulative probabilities less than or equal to the value entered.



Source: http://www.stat.tamu.edu/~west/applets/normaldemo.html

Remember that we asked to find $P(-0.9 \le Z \le 0.9)$. Using $P(-0.9 \le Z \le 0.9) = 1 - 2.P(Z > 0.9)$ produces:

$$P(-0.9 \le Z \le 0.9) = 1 - 2.\underbrace{P(Z > 0.9)}_{0.184060125 \text{ } from\ Excel}$$

$$P(-0.9 \le Z \le 0.9) = 1 - 2.(0.1841) = 0.6318$$

Thus, as before, we found that the probability is 0.6318 that the sample mean will be within 0.3 ounce of the true population mean.

B. NORMINV Function

- The *inverse* of the *Normal CDF* is calculated using the *NORMINV*:

=NORMINV(P, mean, standard_dev)

- where P is the probability of the random variable being less than the level

- For example, if we want to calculate the level (x_0) such that a normally distributed random variable X with a mean of 4 and a standard deviation of 3 will be less than or equal to 5% (0.05) of the time:

=NORMINV(0.05,4,3)

C. Calculating a Random Number from a Normal Distribution

Remember that the NORMINV function returns a value given a probability:

NORMINV(probability, mean, standard_dev)

- Also, remember that *RAND()* function returns a random number between 0 and 1.
- That is, ***RAND()*** generates random probabilities. Therefore, it seems logical that you could use the NORMINV function to calculate a random number from a *normal distribution*, using this formula[2]:

[2] However, Jerry W. Lewis (a former Excel MVP and a professional statistician) offers a stern comment about this approach. Instead, Jerry recommends the Box-Muller method. This method is limited only by the inadequacies of the RAND() function prior to Excel 2003, which had unacceptable autocorrelation. The Box-Muller approach suggests that Excel users should use this formula to calculate a random number from a normal distribution: *=SQRT(-2*LN(RAND()))*SIN(2*PI()*RAND()).* The Box-Muller method is mathematically exact, if implemented with a perfect uniform random number generator and infinite precision. *Source:* *http://www.exceluser.com/explore/statsnormal.htm*.
However, as an important and easy application tool, we will use Excel for our simulation studies in ECON 206.

=NORMINV(rand(), mean, standard_dev)

- Hence, we can draw random values from a *standard normal distribution* using the following formula in Excel:

=NORMINV(rand(), 0, 1)