Chapter 8 Fundamental Sampling Distributions and Data Distributions

吳明龍 副教授
成大資訊系/醫資所
minglong.wu@csie.ncku.edu.tw
辦公室:資訊系館 12 樓

8.5 Sampling Distribution of S^2

- If a random sample of size n is taken from a normal population with mean μ and variance σ^2, and the sample variance S^2 is computed.

$$\sum_{i=1}^{n} X_i - n\bar{X} = 0$$

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Dividing each term of the equality by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i-\bar{X})^2$, we obtain

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Now, according to the corollary of Theorem 7.12 we know that

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

is a chi-squared random variable with n degrees of freedom. We have a chi-squared random variable with n degrees of freedom partitioned into two components. The second term on the right hand side is a Z^2 which is a chi-squared with one degree of freedom, and it turns out that $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degree of freedom. We formalize this in the following theorem.

Corollary: If $X_1, X_2, \ldots, X_n$ are independent random variables having identical normal distributions with mean μ and variances σ^2

$$Y = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

has a chi-squared distribution with $v = n$ degrees of freedom.

Sampling Distribution of S^2

- Theorem 8.4: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

  has a chi-squared distribution with $v = n - 1$ degrees of freedom.

- It is customary to let χ_α^2 represent the χ^2 –value above which we find an area of α. This is illustrated by the shaded region in Figure 8.12.
  - Table A.5

    For $v = 7$

    $\chi^2_{0.05} = 14.067$

    $\chi^2_{0.95} = 2.167$

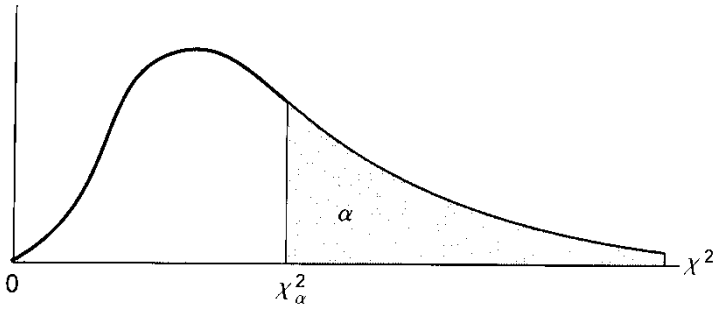

Figure 8.7(8.12): The chi-squared distribution

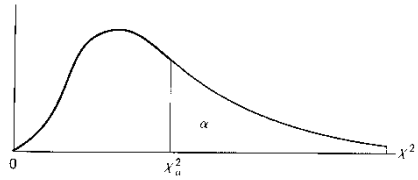

Figure 8.10 The chi-squared distribution.

TABLE A.5 Critical Values of the Chi-Squared Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342

TABLE A.5 (continued) Critical Values of the Chi-Squared Distribution

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588

Sampling Distribution of S^2

- Example 8.7: A manufacturer of car batteries guarantees that his batteries will last, on the average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, is the manufacturer still convinced that his batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.
  - Solution

$$s^2 = \frac{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}{n(n-1)} = \frac{5\times 48.26 - 15^2}{5\times 4} = 0.815$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{4\times 0.815}{1} = 3.26$$

$\because$ 95% of the χ^2-values with 4 degrees of freedom fall between 0.484 and 11.143

$\therefore$ reasonable

Degrees of Freedom As a Measure of Sample Information

- Comparison
  - Corollary 7.1: $\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$

    has a χ^2–distribution with *n* degrees of freedom.

  - Theorem 8.4: $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$

    has a χ^2–distribution with *n* -1 degrees of freedom. (when μ is not known, a degree of freedom is lost in the estimation of μ, i.e. $\overline{X}$)

8.6 *t*-Distribution

- **Central Limit Theorem (Theorem 8.2)** $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$
- σ might not be known.
- Consider $T = \dfrac{\overline{X} - \mu}{S / \sqrt{n}}$
- In developing the sampling distribution of *T*, we shall assume that our random sample was selected from a normal population.

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma / \sqrt{n})}{\sqrt{S^2 / \sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}}$$

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \text{ (standard normal distribution) and}$$

$$V = \frac{(n-1)S^2}{\sigma^2} \text{ (chi-squared distribution)}$$

t-Distribution

- Theorem 8.5: Let *Z* be a standard normal random variable and *V* a chi-squared random variable with *v* degrees of freedom. If *Z* and *V* are independent, then the distribution of the random variable *T*, where

$$T = \frac{Z}{\sqrt{V/v}},$$

is given by the density function

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

This is known as the ***t*-distribution** with *v* degrees of freedom.

t-Distribution

- Corollary 8.1: Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \quad \text{and} \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}.$$

  Then the random variable $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a t-distribution with $v = n$-1 degrees of freedom.
- Student t-distribution
  - The probability distribution of *T* was first published in 1908 in a paper by W. S. Gosset.
  - Employed by an Irish brewery, but disallowed publication.
  - Published his work secretly under the name “Student”.

t-Distribution

- *T* is similar to *Z*: symmetric about $\mu = 0$, bell-shaped.
- Difference between *T* and *Z*: variance of $T \geq 1$ and depends on *n*
- *T* and *Z* are the same: $n \to \infty$

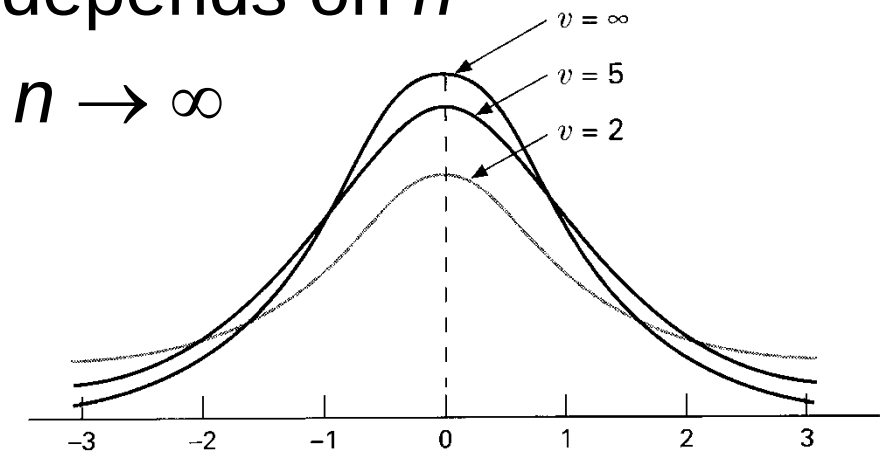

Figure 8.13 The *t*-distribution curves for $v = 2, 5,$ and ∞.

t-Distribution

- *t*-value with 10 degrees of freedom leaving an area of 0.025 to the right is $t = 2.228$.
- *t*-distribution is symmetric about 0: $t_{1-\alpha} = -t_{\alpha}$.
- Example 8.8: The *t*-value with $v = 14$ degrees of freedom that leaves an area of 0.025 to the left, and therefore an area of 0.975 to the right, is

  $$t_{0.975} = -t_{0.025} = -2.145.$$

  (just look up $t_{0.025}$, and then place a negative sign)
- Example 8.9:

  $P(-t_{0.025} < T < t_{0.05}) = 1 - 0.025 - 0.05$
  $= 0.925$

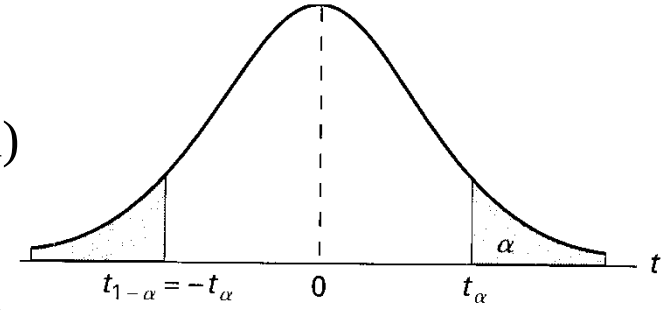

Figure 8.14 Symmetry property of the *t*-distribution.

t-Distribution

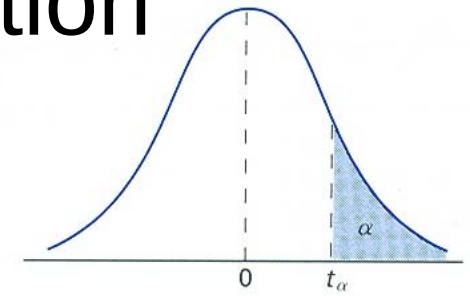

TABLE A.4 Critical Values of the *t*-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131

t-Distribution

Table A.4 (continued) Critical Values of the *t*-Distribution

	α						
v	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073

t-Distribution

- Example 8.10: Find *k* such that $P(k < T < -1.761) = 0.045$, for a random sample of size 15 selected from a normal distribution and $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$.
  - **Solution**

$$v = 15 - 1 = 14$$

$$\text{From Table A.4}, -t_{0.05} = -1.761$$

$$\text{Let } k = -t_{\alpha},\ 0.045 = 0.05 - \alpha$$

$$\Rightarrow \alpha = 0.005 \,(\text{Fig. } 8.13)$$

$$k = -t_{0.005} = -2.977 \,(\text{Table A.4})$$

.

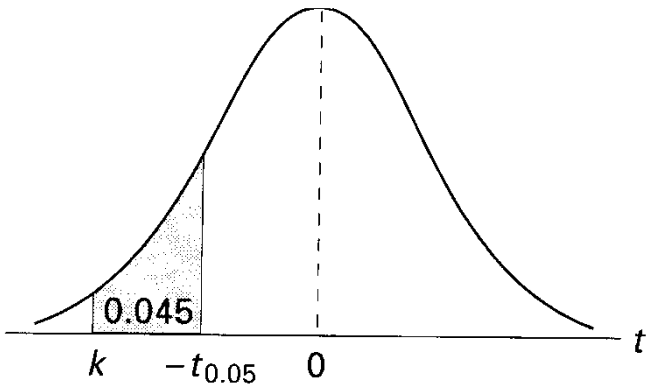

Figure 8.15 The *t*-values for Example 8.13.

t-Distribution

- Exactly 95% of the values of a *t*-distribution with $v = n - 1$ degrees of freedom lie between $-t_{0.025}$ and $t_{0.025}$.
- A *t*-value that falls **below $-t_{0.025}$ or above $t_{0.025}$** would tend to make us believe that either a very rare event has taken place or perhaps our assumption about μ is error.
- Example 8.11: A engineer claims that the population mean of a process is 500 grams. To check this claim he samples 25 batches each month. If the computed *t*-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with his claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.
  - **Solution**

$$\text{From Table A.4}: t_{0.05} = 1.711 \, (v = 24)$$

$$\text{Assumption}: \mu = 500 \Rightarrow t = \frac{518 - 500}{40 / \sqrt{25}} = 2.25 > 1.711, \therefore \text{error}$$

If $\mu > 500$, *t*-value would be more reasonable.

$\therefore$ The process produces a better product than he thought.

t-Distribution

- The *t*-distribution is used extensively in problems that deal with
  - Inference about the population mean
  - Comparative samples (two sample means)
- $T = \dfrac{\overline{X} - \mu}{S / \sqrt{n}}$ requires that X_1, X_2 ,…, X_n be normal.

F-Distribution

- The *F*-distribution finds enormous application in comparing sample variances.
- Theorem 8.6: Let U and V be two independent random variables having chi-squared distribution with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density

$$h(f) = \begin{cases} \dfrac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \dfrac{f^{v_1/2-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & 0 < f < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

This is known as the *F*-distribution with v_1 and v_2 degrees of freedom.

F-Distribution

- Theorem 8.7: Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we obtain $$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$
  - E.g., f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, $$f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246.$$

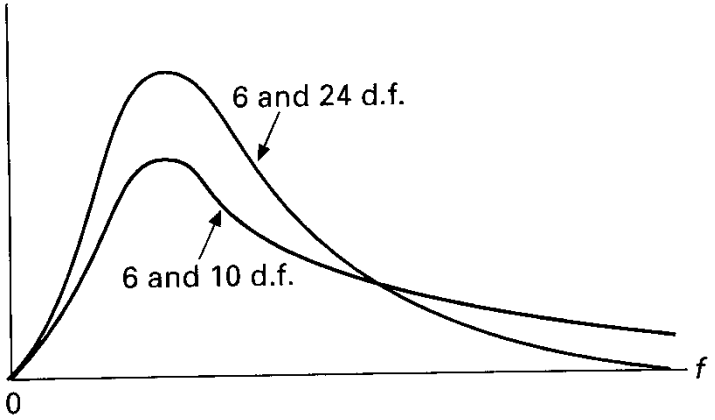

Figure 8.16 Typical F-distributions.

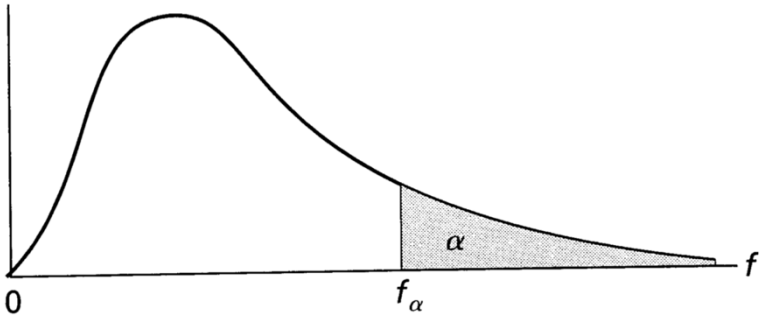

Figure 8.17 Tabulated values of the *F*-distribution.

F-Distribution

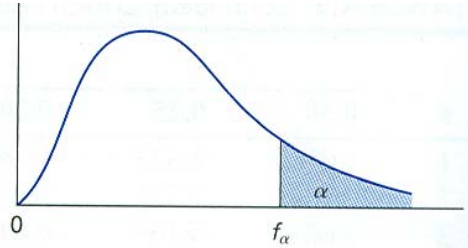

TABLE A.6* Critical Values of the *F*-Distribution

$f_{0.05}(v_1, v_2)$

v_2	v_1 = 1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71

10

4.06

F-Distribution with Two Sample Variances

- Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2

$$X_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \quad \text{and} \quad X_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \quad (\text{from Theorem } 8.4)$$

  Let $U = X_1^2$ and $V = X_2^2$ having chi-squared distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.
  Using Theorem 8.6, we obtain the following result:

- Theorem 8.8: If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{U/v_1}{V/v_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

  has an *F*-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

F-Distribution

- If we wish to determine if the population means are equivalent
  - The normal distribution applies nicely for two-sample situation.
  - However, three-sample?

Paint	Sample mean	Sample variance	Sample size
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.2$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

- *F*-distribution is called the variance ratio distribution.
- Whether sample averages could have occurred by chance depends on the variability within samples, as quantified by S_A^2 and S_B^2, and S_C^2.
- The notion of the important components of variability is best seen through some simple graphics

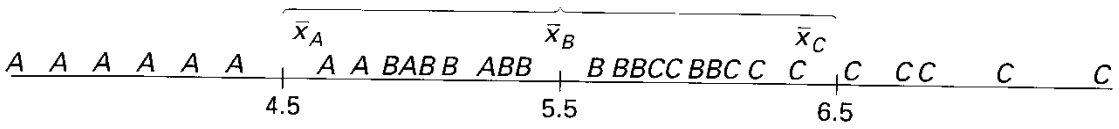

Figure 8.13: Data from three distinct samples.

Analysis of Variance with *F*-Distribution

- Two key sources of variability:
  - Variability within samples
  - Variability between samples
- If the variability within samples is considerably larger than the variability between samples, there will be considerable overlap in the sample data and a signal that the data could all have come from a common distribution.

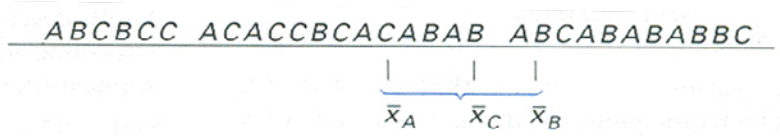

Figure 8.19 Data that easily could have come from the same population.

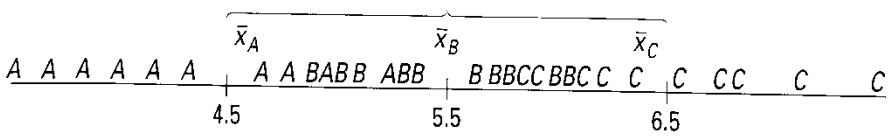

Figure 8.18 Data from three distinct samples.