

Bus 274: Further Statistics for Business

Spring 2011

Hints and Formulas

Caution:

- The focus of Bus 274 is on how to interpret the results of statistical investigations in business, rather than on the use of formulas.
- This formula sheet is no substitute for thoroughly studying the basics of applied statistics in business, as taught and discussed in Bus 274.
- This formula sheet is intended as an aide-mémoire for those who have understood the basics of reasoning with empirical numerical information in the context of business. It will be useless to all others.

95% Confidence Intervals

- For unknown expectation μ in $N(\mu, \sigma^2)$, with known σ^2 :

$$\hat{\mu} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- For unknown expectation μ in $N(\mu, \sigma^2)$, with unknown σ^2 and small samples (up to size 30, say):

$$\hat{\mu} \pm t_{0.975, n-1} \frac{s}{\sqrt{n}}$$

- For unknown variance σ^2 in $N(\mu, \sigma^2)$:

$$\text{lower bound: } \frac{(n-1)s^2}{\chi_{n-1; 0.975}^2}, \quad \text{upper bound: } \frac{(n-1)s^2}{\chi_{n-1; 0.025}^2}$$

- For an unknown probability / share / population proportion (approximate formula, for large n):

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Approximate 95% confidence interval for an unknown parameter θ (large sample size n):

$$\hat{\theta} \pm 2 \cdot \text{se}(\hat{\theta}),$$

where $\hat{\theta}$ is a point estimator for θ and $\text{se}(\hat{\theta})$ is the (estimated) standard error of $\hat{\theta}$.

- The *margin of error* is defined as half the width of a confidence interval.

Test Statistics

- Test for population mean, normal distribution, variance known:

$$T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- Test for population mean, normal distribution, variance unknown:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}, \quad \text{where } s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Test for equality of population means, large independent samples:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

- Test for probability / share / population proportion (large sample):

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Test for population variance, normal distribution:

$$T = \frac{(n-1)s^2}{\sigma_0^2}$$

- Test for equality of population means, normal distribution, independent samples, variances known:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

- Test for equality of population means, normal distribution, independent samples, variances unknown but assumed equal:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s^2}{n_X} + \frac{s^2}{n_Y}}}, \text{ where } s^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

- Test for equality of population means, normal distribution, matched pairs, variances unknown:

$$T = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}, \text{ where } s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2, \quad D_i = X_i - Y_i$$

- Test for equality of population means, large independent samples:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

- Test for equality of population proportions, large independent samples:

$$T = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}, \text{ where } \hat{p} = \frac{n_X \hat{p}_X + n_Y \hat{p}_Y}{n_X + n_Y}$$

- Test for equality of population variances, normal distribution, independent samples:

$$T = \frac{s_X^2}{s_Y^2}$$

Contingency Tables; Odds Ratio; χ^2 Test of Independence

- 2x2 contingency table:

		Y		
		1	0	Σ
X	1	n_{11}	n_{10}	$n_{1\bullet}$
	0	n_{01}	n_{00}	$n_{0\bullet}$
	Σ	$n_{\bullet 1}$	$n_{\bullet 0}$	n

- Odds ratio:

$$\theta = \frac{n_{11}n_{00}}{n_{10}n_{01}}$$

- If X and Y are independent, $\ln(\theta)$ has approximately a normal distribution with expectation 0 and variance

$$s_\theta^2 = \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}$$

- General contingency table:

		Y					
		y_1	y_j	y_c	Σ		
X	x_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	x_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r\bullet}$
	Σ	$n_{\bullet 1}$...	$n_{\bullet j}$...	$n_{\bullet c}$	n

- χ^2 test statistic:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \quad \text{or} \quad \chi^2 = \sum_{i,j} \frac{(o-e)^2}{e}$$

- If X and Y are independent, χ^2 has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Covariance and Correlation

For given pairs of numbers $(x_1, y_1), \dots, (x_n, y_n)$:

- covariance:

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- correlation:

$$r = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

Simple Linear Regression

- The model assumption is:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad i = 1, \dots, n,$$

where the X_i are not all the same and the ϵ_i are iid.

- Estimators:

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}.$$

- Variances of the Estimators:

$$\sigma_\beta^2 = \text{var}(\hat{\beta}) = \frac{\sigma_\epsilon^2}{\sum (X_i - \bar{X})^2}, \quad \sigma_\alpha^2 = \text{var}(\hat{\alpha}) = \sigma_\epsilon^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)$$

The variances can be estimated by substituting the estimated error variance for σ_ϵ^2 :

$$\hat{\sigma}_\epsilon^2 = \frac{\text{SSE}}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

- Testing hypotheses about α and β is based on the following distribution properties:

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}_\beta} \sim t_{n-2}, \quad \frac{\hat{\alpha} - \alpha}{\hat{\sigma}_\alpha} \sim t_{n-2}$$

- Coefficient of determination; coefficient of correlation:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = r_{XY}^2,$$

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{(n \sum X_i^2 - (\sum X_i)^2)(n \sum Y_i^2 - (\sum Y_i)^2)}}$$

One-Way ANOVA

The model assumption is:

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad \mu_i = \mu + G_i, \quad i = 1, \dots, K, \quad j = 1, \dots, n_i.$$

The random variables / observations can be arranged as follows:

1	...	i	...	K	
X_{11}	...	X_{i1}	...	X_{K1}	
X_{12}	...	X_{i2}	...	X_{K2}	
\vdots		\vdots		\vdots	
X_{1n_1}	...	X_{in_i}	...	X_{Kn_K}	
\bar{X}_1	...	\bar{X}_i	...	\bar{X}_K	\bar{X}

Here,

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \text{ (group } i), \quad \bar{X} = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^K \frac{n_i}{n} \bar{X}_i \text{ (overall)}$$

Sums of squares:

$$\begin{aligned}
 \text{between groups: } \text{SSG} &= \sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2 \\
 \text{within groups: } \text{SSW} &= \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \\
 \text{total: } \text{SST} &= \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2
 \end{aligned}$$

The layout of a one-way ANOVA table with K groups is:

Source of variation	SS	df	MS	F_{calc}	F_{crit}
between groups	SSG	$K - 1$	MSG	MSG/MSW	
within groups	SSW	$n - K$	MSW		
total	SST	$n - 1$			

Two-way ANOVA, L observations per cell

The model assumption is:

$$X_{ijl} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = \mu + G_i + B_j + I_{ij}, \quad i = 1, \dots, K, \quad j = 1, \dots, H, \quad l = 1, \dots, L.$$

The random variables / observations can be arranged as follows:

	1	...	i	...	K	
1	$X_{111}, \dots, X_{11l}, \dots, X_{11L}$	$\bar{X}_{\cdot 1}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
j	$X_{ij1}, \dots, X_{ijl}, \dots, X_{ijL}$	$\bar{X}_{\cdot j}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
H	$X_{KH1}, \dots, X_{KHL}, \dots, X_{KHL}$	$\bar{X}_{\cdot H}$
	$\bar{X}_{1\cdot}$...	$\bar{X}_{i\cdot}$...	$\bar{X}_{K\cdot}$	\bar{X}

Here,

$$\bar{X}_{i\cdot} = \frac{1}{HL} \sum_{j,l} X_{ijl}, \quad \bar{X}_{\cdot j} = \frac{1}{KL} \sum_{i,l} X_{ijl}, \quad \bar{X} = \frac{1}{KHL} \sum_{i,j,l} X_{ijl}.$$

Sums of squares:

$$\begin{aligned}
 \text{between groups: } \text{SSG} &= HL \sum_{i=1}^K (\bar{X}_{i\cdot} - \bar{X})^2 \\
 \text{between blocks: } \text{SSB} &= KL \sum_{j=1}^H (\bar{X}_{\cdot j} - \bar{X})^2 \\
 \text{interaction: } \text{SSI} &= L \sum_{i=1}^K \sum_{j=1}^H (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2 \\
 \text{error: } \text{SSE} &= \sum_{i=1}^K \sum_{j=1}^H \sum_{l=1}^L (X_{ijl} - \bar{X}_{ij\cdot})^2 \\
 \text{total: } \text{SST} &= \sum_{i=1}^K \sum_{j=1}^H \sum_{l=1}^L (X_{ijl} - \bar{X})^2
 \end{aligned}$$

Two-way ANOVA table:

source of variation	sum of squares	df	mean squares	F_{obs}	F_{crit}
groups	SSG	$K - 1$	MSG	MSG/MSE	$F_{1-\alpha; K-1, KH(L-1)}$
blocks	SSB	$H - 1$	MSB	MSB/MSE	$F_{1-\alpha; H-1, KH(L-1)}$
interaction	SSI	$(K - 1)(H - 1)$	MSI	MSI/MSE	$F_{1-\alpha; (K-1)(H-1), KH(L-1)}$
error	SSE	$KH(L - 1)$	MSE		
total	SST	$KHL - 1$			