Harald Schmidbauer
Sema Bayraktar
Şule Özmen

İstanbul Bilgi Üniversitesi

Bus 273: Statistical Analysis for Business

Fall 2008

SOME FORMULAS

Caution:

- The focus of Bus 273 is on how to interprete 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 273.
- 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 misleading to all others.

Descriptive Statistics

For given numbers $x_1, \ldots, x_n$:

- arithmetic mean:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{or} \quad \bar{x} = \sum_{i=1}^{n} \alpha_i x_i \quad (\alpha_i \geq 0,\ \sum \alpha_i = 1)$$
- geometric mean ($x_1, \ldots, x_n > 0$):
$$\bar{x}_{\mathrm{G}} = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \text{or} \quad \bar{x}_{\mathrm{G}} = \prod_{i=1}^{n} x_i^{\alpha_i} \quad (\alpha_i \geq 0,\ \sum \alpha_i = 1)$$
- variance:
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{or} \quad s^2 = \sum_{i=1}^{n} \alpha_i (x_i - \bar{x})^2 \quad (\alpha_i \geq 0,\ \sum \alpha_i = 1)$$
It holds that
$$s^2 = \sum_{i=1}^{n} \alpha_i x_i^2 - \bar{x}^2$$
- standard deviation:
$$s = \sqrt{s^2}$$
- coefficient of variation ($x_1, \ldots, x_n > 0$):
$$V = \frac{s}{\bar{x}}$$
- moment coefficient of skewness:
$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3} \quad \text{or} \quad \gamma_1 = \sum_{i=1}^{n} \alpha_i \frac{(x_i - \bar{x})^3}{s^3} \quad (\alpha_i \geq 0,\ \sum \alpha_i = 1)$$
- moment coefficient of kurtosis:
$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s^4} - 3 \quad \text{or} \quad \gamma_2 = \sum_{i=1}^{n} \alpha_i \frac{(x_i - \bar{x})^4}{s^4} - 3 \quad (\alpha_i \geq 0,\ \sum \alpha_i = 1)$$

- The sigma-rules. If the observations are approximately normally distributed:
  - One-sigma-rule: About two thirds, or 67%, of all observations will be in the interval $[\bar{x}-s, \bar{x}+s]$.
  - Two-sigma-rule: About 95% of all observations will be in the interval $[\bar{x}-2s, \bar{x}+2s]$.
  - Three-sigma-rule: Almost all observations (more than 99%) will be in the interval $[\bar{x}-3s, \bar{x}+3s]$.

Conditional Probability and the Bayes Theorem

- Conditional probability of an event A, given B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (\text{if } P(B) > 0)$$

- Bayes theorem: For a decomposition $E_1, \ldots, E_n$ of the state space $\mathcal{S}$ (that is, $E_1 \cup \cdots \cup E_n = \mathcal{S}, E_i \cap E_j = \emptyset$ for $i \neq j$):

$$P(E_i|A) = \frac{P(A|E_i) \cdot P(E_i)}{\sum_{j=1}^{n} P(A|E_j) \cdot P(E_j)}$$

Discrete Probability Distributions

- For a discrete random variable X with probabilities $p_i = P(X = x_i)$, $\sum_i p_i = 1$.
- expectation:

$$\mathrm{E}(X) = \sum_i p_i x_i$$

- variance:

$$\mathrm{var}(X) = \sum_i p_i (x_i - \mathrm{E}(X))^2 = \mathrm{E}(X^2) - \mathrm{E}^2(X)$$

- standard deviation:

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

- binomial distribution: p = success probability in each trial, X = number of successes in n independent trials,

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, \ldots, n$$

  If X has this distribution, we write: $X \sim \mathrm{B}(n, p)$, and it holds that $\mathrm{E}(X) = np$, $\mathrm{var}(X) = np(1-p)$. — If $X_1, \ldots, X_n$ are independent and $X_i \sim \mathrm{B}(n_i, p)$, then $\sum_{i=1}^{n} X_i \sim \mathrm{B}(n, p)$, where $n = \sum_{i=1}^{n} n_i$.

- hypergeometric distribution: Urn with N balls, of which M white, $N-M$ black, n balls are drawn randomly without replacement, X = number of white balls among the n drawn balls,

$$p_i = \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}}, \quad i = \max(0, n+M-N), \ldots, \min(M, n)$$

  It holds that $\mathrm{E}(X) = n \cdot \frac{M}{N}$, $\mathrm{var}(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \left(\frac{N-n}{N-1}\right)$.

- Poisson distribution:

$$p_i = \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda}, \quad i = 0, 1, \ldots$$

  If X has this distribution, we write $X \sim \mathrm{Po}(\lambda)$. It holds that $\mathrm{E}(X) = \lambda$, $\mathrm{var}(X) = \lambda$. — If $X_1, \ldots, X_n$ are independent and $X_i \sim \mathrm{Po}(\lambda_i)$, then $\sum_{i=1}^{n} X_i \sim \mathrm{Po}(\lambda)$, where $\lambda = \sum_{i=1}^{n} \lambda_i$.

Continuous Probability Distributions

- For a continuous random variable X with density $f(x)$, $P(a \le X \le b) = \int_a^b f(x)\mathrm{d}x$.
- expectation:
$$\mathrm{E}(X) = \int_{\mathbb{R}} x f(x)\mathrm{d}x,$$
- variance:
$$\mathrm{var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 f(x)\mathrm{d}x = \mathrm{E}(X^2) - \mathrm{E}^2(X),$$
- standard deviation:
$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$
- normal distribution:
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$
If X has this density, we write: $X \sim \mathrm{N}(\mu, \sigma^2)$. It holds that $\mathrm{E}(X) = \mu$, $\mathrm{var}(X) = \sigma^2$. Furthermore:
  - For $a, b \in \mathbb{R}$,
$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2); \quad \text{in particular: } \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$
  - If $X_1, \ldots, X_n$ are independent and $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^n X_i \sim \mathrm{N}(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^n \mu_i$, $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. In particular, if $X_1, \ldots, X_n$ are iid and $X_i \sim \mathrm{N}(\mu, \sigma^2)$, then $\sum_{i=1}^n X_i \sim \mathrm{N}(n\mu, n\sigma^2)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathrm{N}(\mu, \sigma^2/n)$.
- lognormal distribution: If $Y \sim \mathrm{N}(\mu, \sigma^2)$, then $X := \mathrm{e}^Y \sim \mathrm{LN}(\mu, \sigma^2)$. It holds that
$$\mathrm{E}(X) = \mathrm{e}^{\mu + \frac{1}{2}\sigma^2}, \quad \mathrm{var}(X) = \mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2} - 1)$$
- exponential distribution: If X has the density $f(x) = \lambda \mathrm{e}^{-\lambda \mathrm{x}}$ (where $x \ge 0$), we write: $X \sim \mathrm{EXPO}(\lambda)$. It holds that $\mathrm{E}(X) = \frac{1}{\lambda}$, $\mathrm{var}(X) = \frac{1}{\lambda^2}$. The distribution function of X is $F(x) = 1 - \mathrm{e}^{-\lambda \mathrm{x}}$.

Inductive Statistics — Some Point Estimators for Parameters of N and LN

- for μ in $\mathrm{N}(\mu, \sigma^2)$: $\hat{\mu} = \bar{X}$
- for σ^2 in $\mathrm{N}(\mu, \sigma^2)$: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ (unbiased)
- for μ, σ^2 in $\mathrm{LN}(\mu, \sigma^2)$: $\hat{\mu} = \ln \bar{X} - \frac{1}{2}\ln\left(\frac{s^2}{\bar{X}^2} + 1\right)$, $\hat{\sigma}^2 = \ln\left(\frac{s^2}{\bar{X}^2} + 1\right)$ (method of moments)

Inductive Statistics — Confidence Intervals

- approximate 95% confidence bounds for an unknown share p: $\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- 95% confidence bounds for μ in $\mathrm{N}(\mu, \sigma^2)$, σ^2 known: $\hat{\mu} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$
- 95% confidence bounds for μ in $\mathrm{N}(\mu, \sigma^2)$, σ^2 unknown: $\hat{\mu} \pm t_{0.975;n-1} \cdot \frac{s}{\sqrt{n}}$ with $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
- approximate 95% confidence bounds for an unknown parameter θ: $\hat{\theta} \pm 2 \cdot \mathrm{std.error}(\hat{\theta})$