Harvard College

Statistics 104: Quantitative Methods for Economics

Formula and Theorem Review

Tommy MacWilliam, '13

tmacwilliam@college.harvard.edu

March 10, 2011

Contents

1 Introduction to Data

1.1 Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1.2 Interquartile Range

$$IQR = Q3 - Q1$$

1.3 Outlier Detection

An observation x_i in a set of data is considered an outlier if

$$x_i > Q3 + 1.5 \cdot IQR \quad \vee \quad x_i < Q1 - 1.5 \cdot IQR$$

1.4 Mean Absolute Deviation

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

1.5 Variance

$$s_x^2 = \frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n-1}$$

1.6 Standard Deviation

$$s_x = \sqrt{\frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

1.7 Coefficient of Variation

$$CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$$

1.8 Empirical Rule for Standard Deviation

For "mound-shaped," symmetric data:

$$68\% \text{ of the data is in the interval } (\bar{x} - s_x, \bar{x} + s_x)$$

$$95\% \text{ of the data is in the interval } (\bar{x} - 2s_x, \bar{x} + 2s_x)$$

1.9 Chebyshev's Rule

For any set of data, the proportion of data that lines within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2}$$

1.10 Linear Transformations

$$\begin{aligned}
Var(a + bX) &= b^2 Var(X) \\
Average(a + bX) &= a + b(Average(X)) \\
StdDev(a + bX) &= b \cdot StdDev(X) \\
Q_i(a + bX) &= a + bQ_i \\
IQR(a + bX) &= b \cdot IQR(X)
\end{aligned}$$

1.11 Z-Score

To obtain a set of data with mean 0 and variance 1:

$$z = \frac{X - \bar{X}}{s_x}$$

1.12 Covariance

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$Covariance > 0$: Larger $X \Leftrightarrow$ Larger Y

$Covariance < 0$: Larger X $\Leftrightarrow$Smaller Y

$Cov(X, X) = Var(X)$

1.13 Correlation

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

1.14 Combining Data Sets

$$\begin{aligned}\bar{Z} &= a\bar{X} + b\bar{Y} \\ Var(Z) &= a^2 s_X^2 + b^2 s_Y^2 + 2(ab)s_{XY}\end{aligned}$$

2 Probability and Random Variables

2.1 Definition of Probability

$$P(event) = \frac{\text{outcomes where the event occurs}}{\text{total outcomes}}$$

2.2 Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2.3 Complement Rule

$$P(\bar{A}) = 1 - P(A)$$

2.4 Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2.5 Independence

Two events A and B are said to be independent if:

$$P(A|B) = P(A) \quad \vee \quad P(B|A) = P(B)$$

2.6 Joint Probability

If two events A and B are independent:

$$P(A \cap B) = P(A) \cdot P(B)$$

2.7 2x2 Matrix

	B	$\bar{B}$
A	$P(A \cap B) = P(B) \cdot P(A\|B)$	$P(A \cap B) = P(B) \cdot P(A\|B)$
$\bar{A}$	$P(\bar{A} \cap B) = P(\bar{A}) \cdot P(B\|\bar{A})$	$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B}\|\bar{A})$

2.8 Bayes' Theorem

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})}$$

2.9 Probability Function

$$P_X(x) = P(X = x)$$

2.10 Cumulative Distribution Function

$$F_X(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} P_X(x)$$

2.11 Expected Value

$$\mu_X = E(X) = \sum_{all\, x_i} x_i P(x_i)$$

2.12 Variance of a Random Variable

$$\begin{aligned} \sigma_X^2 &= Var(X) = E((X - \mu_X)^2) = \sum_{all\, x_i} (x_i - \mu)^2 \cdot P(x_i) \\ \sigma_X^2 &= E(X^2) - \mu_X^2 = E(X^2) - E(X)^2 \end{aligned}$$

2.13 Linear Transformations of Random Variables

$$\begin{aligned} E(a+bX) &= a+bE(X) = a+b\mu_X \\ Var(a+bX) &= b^2\sigma_X^2 \\ E(a) &= a \\ Var(a) &= 0 \end{aligned}$$

2.14 Joint Distribution Function

$$P_{X,Y}(x,y) = P(X=x \wedge Y=y)$$

2.15 Marginal Distributions

$$\begin{aligned} P_X(x) &= \sum_y P_{X,Y}(x,y) \\ P_Y(y) &= \sum_x P_{X,Y}(x,y) \end{aligned}$$

2.16 Independence of Random Variables

Two random variables X and Y are independent if $\forall x, y$:

$$P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y) \quad \vee \quad P_{X|Y}(X=x|Y=y) = P(X=x)$$

2.17 Conditional Distribution of Random Variables

$$P_{X|Y}(X=x|Y=y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$

2.18 Conditional Expectation of Random Variables

$$E(X|Y=y) = \sum_{all\, x} xP(X=x|Y=y)$$

2.19 Covariance of Random Variables

$$\sigma_{X,Y} = Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - E(X) \cdot E(Y)$$

where $E(XY) = \sum xyP(X=x, Y=y)$

2.20 Correlation of Random Variables

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

2.21 Combinations of Random Variables

If X and Y are independent:

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ Var(X+Y) &= Var(X) + Var(Y) \end{aligned}$$

If X and Y are not independent:

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ Var(X+Y) &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

General case:

$$\begin{aligned} E((a+bX)+(c+dY)) &= a + bE(X) + c + dE(Y) \\ Var((a+bX)+(c+dY)) &= b^2Var(X) + d^2Var(Y) + 2(bd)Cov(X,Y) \end{aligned}$$

3 Probability Distributions

3.1 Combinations

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

3.2 Binomial Distribution Formula

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

3.3 Characteristics of Binomial Distributions

$$\begin{aligned}\mu &= E(X) = np \\ \sigma^2 &= npq \\ \sigma &= \sqrt{npq}\end{aligned}$$

3.4 Probability of an Interval

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

where F_X is the CDF, such that $F_X(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$

3.5 Z-Score for Normal Distribution

If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$. Therefore:

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$$

3.6 Central Limit Theorem

If random samples are taken from any population with mean μ and variance σ^2, as the sample size n increases, the distribution approaches a normal distribution with $\mu_X = \mu$ and $\sigma_X^2 = \dfrac{\sigma^2}{n}$.

4 Confidence Intervals

4.1 Confidence Interval

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

4.2 Sample Proportion

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

4.3 Central Limit Theorem for Proportions

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

4.4 Confidence Interval for Proportions

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

4.5 Confidence Interval for Correlation

$$r \pm 1.96\sqrt{\frac{1-r^2}{n-2}}$$

4.6 Confidence Interval for Difference in Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

4.7 Confidence Interval for Difference in Means

$$(\mu_1 - \mu_2) \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

5 Hypothesis Testing

5.1 Test Statistic for Population Mean

$$z_{stat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5.2 Test Statistic for Proportion

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

5.3 Test Statistic for Two Samples

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

5.4 Test Statistic for Two Proportions

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

5.5 Test Statistic for Chi-Square Test

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

where $e_i = np_i$

6 Regression

6.1 Residual

$$e_i = Y_i - \hat{Y}_i$$

6.2 Least Squares Method

We can minimize $\sum(Y_i - b_0 - b_1 X_i)^2 = \sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2$ by using the coefficients:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = r_{XY}\left(\frac{s_Y}{s_X}\right) \\ b_0 &= \bar{Y} - b_1 \bar{X} \end{aligned}$$

6.3 Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

6.4 Standard Error

$$s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{SSE}{df_{error}}}$$

6.5 Regression Test Statistic

$$T = \frac{b_1 - \beta_1^\star}{s_{b_1}}$$

6.5.1 Confidence Interval for Predicting an Average

$$b_0 + b_1 X_{new} \pm 1.96 \left[s_e \left(\frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{(n-1)s_X^2} \right)^{1/2} \right]$$

6.6 Confidence Interval for Predicting a Value

$$b_0 + b_1 X \pm 1.96 \left[s_e \left(1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{(n-1)s_X^2} \right)^{1/2} \right]$$

6.7 Adjusted Coefficient of Determination

$$adjusted\, R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

6.8 Overall F-Test

$$f = \frac{SSR/k}{SSE/(n-k-1)}$$

6.9 Standardized Residuals

$$r_i = \frac{e_i}{s_e} \approx \frac{\epsilon_i}{\sigma} \sim N(0,1)$$

6.10 Logistic Function

$$f(x) = \frac{e^x}{1+e^x} = \frac{\exp(x)}{1+\exp(x)}$$