PACIFIC REGION TECHNICAL NOTES

79-011

March 15, 1979

Hypothesis Testing - Vancouver Public Verification Scores

Pat Carroll, Meteorologist
Pacific Weather Centre, Vancouver

INTRODUCTION:

DURING THE PERIOD NOVEMBER 1978 THROUGH JANUARY 1979 AN EXPERIMENTAL REORGANIZATION WAS MADE OF THE FORECAST DUTIES AT PACIFIC WEATHER CENTRE (PWC). HORITA (1) EXAMINED THE 5:00 AM PUBLIC VERIFICATION SCORES FOR THE ABOVE PERIOD. HE CONCLUDED THAT NO DETERIORATION HAD OCCURRED BASED ON A COMPARISON WITH SCORES FROM THE PREVIOUS YEAR. HOWEVER, FROM A STATISTICAL VIEWPOINT NO LEVEL OF SIGNIFICANCE WAS GIVEN FOR THE CONCLUSIONS REACHED. A MORE GENERAL APPROACH BASED ON HYPOTHESIS TESTING WILL BE SHOWN WITH SOME COMMENTS ON HOW SUCH AN APPROACH CAN BE USED AS AID IN MANAGEMENT DECISIONS.

ANALYSIS OF DATA.

WE WISH TO COMPARE THE INDIVIDUAL MONTHLY MEAN VERIFICATION SCORES FOR NOVEMBER 1978 THROUGH JANUARY 1979 AGAINST THE LONG TERM MEANS BASED ON DATA FROM THE YEARS 1972 TO 1977. ACCORDINGLY WE DECLARE AS THE NULL HYPOTHESIS THAT THERE WAS NO CHANGE IN THE MEAN SCORES. BECAUSE WE WISH TO DETECT A DETERIORATION IN THE MEAN WE DECLARE AS THE ALTERNATIVE HYPOTHESIS THAT THE MEAN SCORES ARE LOWER. IN SYMBOLIC TERMS WE HAVE:

$H_o : \mu = \mu_o$ μ – LONG TERM MEAN (ESTIMATED FROM SAMPLE)
$H_1 : \mu < \mu_o$ μ_o – LONG TERM MEAN (MEASURED FROM PAST DATA)

A SIGNIFICANCE LEVEL α OF 0.05 IS SET. THIS REPRESENTS THE RISK WE ARE WILLING TO TAKE IN REJECTING THE NULL HYPOTHESIS WHEN IT IS CORRECT.

THE TEST STATISTIC IS $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

WHERE $\bar{X}$ IS THE SAMPLE (MONTHLY) MEAN SCORE
μ IS THE LONG TERM MEAN
σ IS THE POPULATION STANDARD DEVIATION (NOT KNOWN)
N IS THE SAMPLE SIZE (30 OR 31 SCORES)

BECAUSE OUR SAMPLE SIZE IS RELATIVELY LARGE ($N \geq 30$) WE CAN APPROXIMATE THE VALUE OF σ WITH THE STANDARD DEVIATION OF THE SAMPLE (S). TABLE 1 SUMMARIZES THE NECESSARY INFORMATION:

	$\bar{X}$	S	μ	n	Z	Z_α
NOV	77.8	11.1	74.7	30	1.530	-1.645
DEC	73.0	10.4	75.7	31	-1.445	-1.645
JAN	71.7	16.5	75.3	31	-1.215	-1.645

TABLE 1

IN EACH CASE THE TEST STATISTIC Z IS GREATER THAN THE CRITICAL VALUE Z_α. THEREFORE, WE CANNOT REJECT THE NULL HYPOTHESIS. THE MEAN SCORES ARE NO WORSE.

AT THIS POINT WE MAY BE TEMPTED TO ACCEPT THE TEST RESULTS AT FACE VALUE. HOWEVER, THERE ARE TWO POINTS YET TO BE RESOLVED. FIRST, IT WAS ASSUMED THAT THE SAMPLE SCORES WERE RANDOMLY ACQUIRED, THAT IS CAN WE BE REASONABLY CERTAIN THAT THESE SCORES WERE REPRESENTATIVE OF THE POPULATION OF SCORES FOR THE MONTHS INVOLVED? SECOND, ALTHOUGH WE HAVE NOT REJECTED THE NULL HYPOTHESIS CAN WE BE EQUALLY CONFIDENT IN ACCEPTING IT?

THE SIMPLEST WAY TO MINIMIZE THE EFFECTS OF NON-RANDOM SAMPLING IS BY INCORPORATING DATA FROM A LONGER PERIOD OF TIME. PUSS (2) HAS SHOWN THAT VARIABILITY IN THE STANDARD DEVIATIONS OF SCORES DECREASES WITH TIME. FOR YEARLY SCORES THE STANDARD DEVIATIONS ARE NEARLY CONSTANT. IN THIS TEST THE BEST WE CAN DO IS TO COMBINE THE SAMPLES AND TEST THE MEAN THREE MONTH SCORE. TABLE 2 GIVES INFORMATION SIMILAR TO 1.

	$\bar{X}$	S	μ	n	Z	Z_α
NOV - JAN	74.1	13.1	75.1	92	-0.732	-1.645

TABLE 2

AS BEFORE THE TEST STATISTIC IMPLIES ACCEPTANCE OF THE NULL HYPOTHESIS. IN THIS CASE IT IS THE THREE MONTH MEAN THAT HAS NOT DETERIORATED.

OUR CONFIDENCE IN ACCEPTING THE NULL HYPOTHESIS IS DEPENDENT ON THE ABILITY OF THE TEST TO DETECT REAL CHANGES IN THE LONG TERM MEAN BASED ON THE OBSERVED SAMPLE. A PARAMETER β SIMILAR TO α GIVES THE RISK OF REJECTING THE ALTERNATE HYPOTHESIS WHEN IT IS CORRECT. FIGURES 1 AND 2 SHOW THE VARIATION OF β AS THE TRUE LONG TERM MEAN DEVIATES FROM THE HYPOTHESIZED VALUE. THE POWER OF THE TEST IS DEFINED

AS $1-\beta$. FOR THE NOVEMBER CASE WE CAN SEE THAT β IS ABOUT 0.25 WHEN μ IS 70%. THIS MEANS THAT BASED ON THIS SAMPLE THERE IS A 25% CHANCE OF FAILING TO DETECT A DETERIORATION WHEN THE MEAN HAS ACTUALLY FALLEN TO 70%. CLEARLY THIS HIGH VALUE OF β IS NOT ENCOURAGING BUT CONSIDER THE COMBINED DATA CURVE FOR μ = 70%. HERE β IS ABOUT 0.02 FOR ABOUT THE SAME CHANGE IN THE MEAN. THE TEST FOR THE COMBINED DATA IS SAID TO HAVE MORE POWER AND IS DIRECTLY RELATED TO THE INCREASED SAMPLE SIZE.

AT THE BEGINNING OF THE TEST THE VALUE OF α WAS CHOSEN AS 0.05 BUT IT COULD HAVE BEEN 0.10 OR 0.01 OR ANY OTHER VALUE. A VALUE OF β COULD ALSO HAVE BEEN SET BUT WHAT VALUE? LET US SUPPOSE THAT THE OUTCOME OF THE TEST WAS TO BE USED IN A DECISION REGARDING THE FUTURE CONTINUATION OF THE PWC EXPERIMENT. IF IT WERE NECESSARY TO DETECT SMALL CHANGES IN THE LONG TERM MEAN THEN β MUST BE SMALL. FOR EXAMPLE, SUPPOSE THAT A DECISION WOULD BE MADE TO DISCONTINUE THE EXPERIMENT IF THE MEAN SCORE DETERIORATED BY MORE THAN 5%. IF α IS LEFT AT 0.05 THEN β CAN BE LOWERED ONLY BY INCREASING THE SAMPLE SIZE N. FOR A β OF 0.01 WE WOULD REQUIRE A SAMPLE SIZE OF 159 OR ABOUT 6 MONTHS DATA (ROUNDED UP).

CONCLUSIONS:

AN ANALYSIS OF THE MEAN VERIFICATION SCORES FOR NOVEMBER 1978 THROUGH JANUARY 1979 HAS SHOWN NO SIGNIFICANT DETERIORATION BASED ON A LEVEL OF SIGNIFICANCE OF 0.05 COMPARED TO MEAN SCORES DERIVED FROM THE YEARS 1972 TO 1977. AN EXAMINATION OF THE POWER OF THE TEST SHOWS THAT THE ACCEPTANCE OF THE NULL HYPOTHESIS FOR THE INDIVIDUAL MONTHS SHOULD BE DONE WITH CAUTION BUT FOR THE COMBINED DATA THE TEST WILL DETECT ABOUT A 6% CHANGE IN THE MEAN 98% OF THE TIME. TO DETECT A 5% CHANGE WITH A POWER OF 99% WOULD REQUIRE AN ADDITIONAL 3 MONTHS DATA.

REFERENCES:

1. HORITA - PWC EXPERIMENTAL SATELLITE PROGRAM FOR NOV. 1978 - JAN. 1979. TECHNICAL NOTE 79-005, FEB 14, 1979.

2. PUSS - PUBLIC FORECAST VERFIFICATION SCORES FOR GREATER VANCOUVER 1972 - 1978. TECHNICAL NOTE 78-047, DEC 6, 1978.

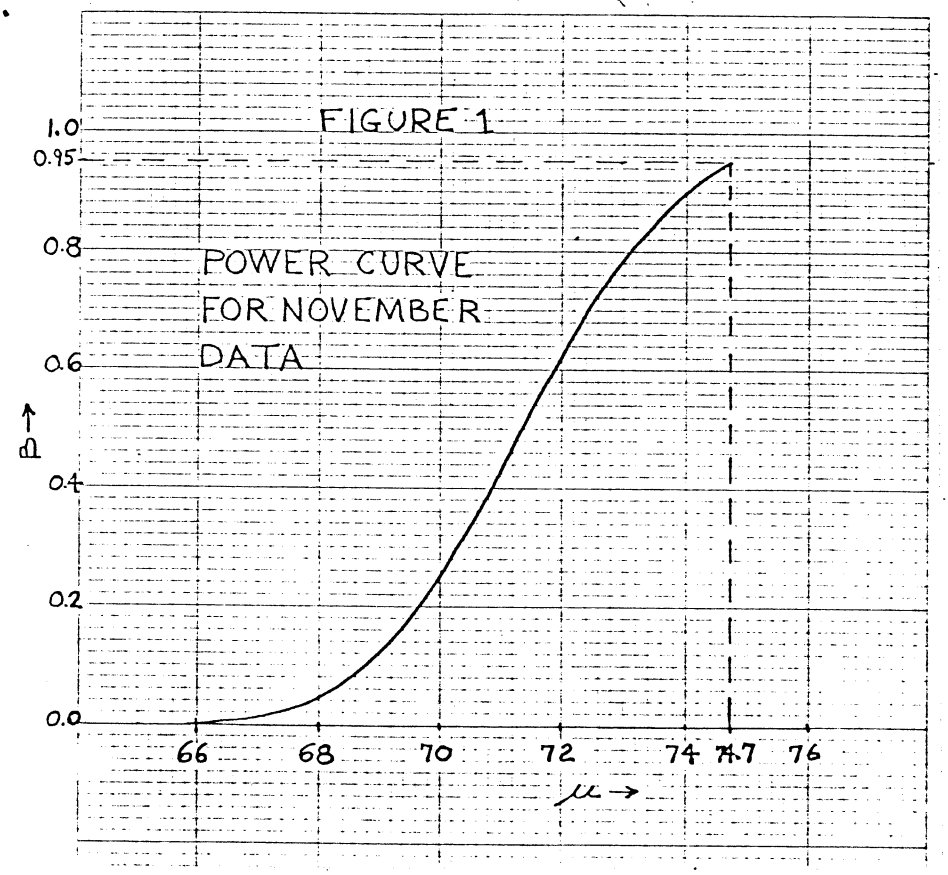
FIGURE 1
POWER CURVE
FOR NOVEMBER
DATA
1.0
0.95
0.8
0.6
0.4
0.2
0.0
66
68
70
72
74
76
μ →

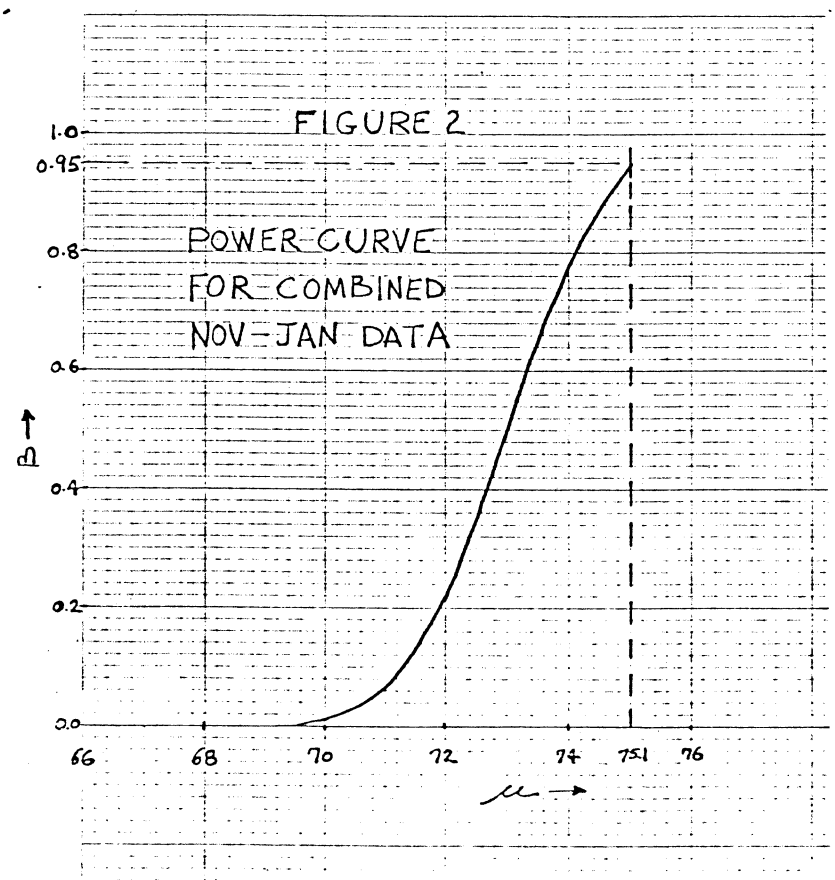
FIGURE 2
POWER CURVE
FOR COMBINED
NOV-JAN DATA
1.0
0.95
0.8
0.6
0.4
0.2
0.0
66
68
70
72
74
76
μ →