

NOTE Z: SETTING UP THE EXCEL SHEET FOR CALCULATING P VALUES

Enter data into the cells. The following are some recommended layouts if functions are used. If the Tool-Pac routines are used, you only need to be able to identify the ranges of the existing data.

ONE SAMPLE TESTS

MEAN

Name	μ	Mean	St. Dev	n	df	t	Probability
	(1)	(2)	(3)	(4)	(5)	(6)	(7)

FOR A HYPOTHESIS ON ONE POPULATION MEAN (KNOWN), NORMALLY DISTRIBUTED, ONE TAIL TEST

POPULATION σ KNOWN

- (1) From Claim
- (2) Function AVERAGE of data
- (3) From Claim
- (4) Function COUNT on data
- (5) Cell(4) - 1
- (7) Function NORMDIST((2), (1), (3), TRUE)

$$z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$$

σ NOT KNOWN, LARGE SAMPLE

- (1) From Claim
- (2) Function AVERAGE on data
- (3) From STDEV on data
- (4) Function COUNT on data
- (5) Cell(4) - 1
- (7) Function NORMDIST((2), (1), (3), TRUE)

$$z = (\bar{X} - \mu) / (s / \sqrt{n})$$

σ NOT KNOWN, SMALL SAMPLE

- (1) From Claim
- (2) Function AVERAGE on Data
- (3) From STDEV on data
- (4) Function COUNT on data

$$(5) = (4) - 1$$

$$(6) = (((1) - (2))) / ((4) / \text{SQR}(3))$$

(7) Function TDIST((6), (5), 1 or 2 from Claim)

$$t = (\bar{X} - \mu) / (s / \sqrt{n})$$

FOR A HYPOTHESIS ON ONE POPULATION MEAN OR MEDIAN (KNOWN), NORMALLY DISTRIBUTED, TWO TAIL TEST

POPULATION σ KNOWN

Name	μ	Sigma	Probability
	(1)	(2)	(3)

(1) From Claim

(2) From Claim

(3) Function ZTEST(range of data, (1), (2))

POPULATION σ UNKNOWN, LARGE SAMPLE

Name	μ	Probability
	(1)	(2)

(1) From Claim

(2) Function ZTEST(range of data, (1) ..leave blank..)

MEDIAN

Name	Population Median	# of + signs	# of - signs	n	Smaller of a or b	z Value	Probability
	(1)	(2)	(4)	(3)	(5)	(6)	(7)

FOR A HYPOTHESIS ON ONE POPULATION MEDIAN VALUE, NONPARAMETRIC, LARGE SAMPLES (N > 25)

(1) Function MEDIAN on data

(2) Manually count the number of data values greater than or equal to (1)

(3) Function COUNT on data

$$(4) = (3) - (2)$$

(5) Smaller of (2) or (4)

$$(6) = ((5) + 0.5 - (3) / 2) / (\text{SQR}(3) / 2)$$

(7) Function NORMSDIST((6))

$$z = [x + 0.5 - n / 2] / [\sqrt{n} / 2]$$

FOR A HYPOTHESIS ON PROPORTION

Name	p	q	n	np	nq	Number of successes	p-hat	z	Probability
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

FOR A HYPOTHESIS ON ONE POPULATION PROPORTION, LARGE SAMPLES WHERE NP>5 AND NQ>5

- (1) From Claim
- (2) = 1 – (1)
- (3) From Claim
- (4) = (1) * (3)
- (5) = (2) * (3)
- (6) From Claim
- (7) = (6) / (3)
- (8) = ((7) – (1)) / SQR((1) * (2) / (3))
- (9) Function NORMSDIST((8))

$$z = (\text{p-hat} - p) / \sqrt{(p \times q / n)}$$

FOR A HYPOTHESIS ON VARIANCE

Name	Population σ^2	Variance	n	df	Chi	Probability
	(1)	(2)	(3)	(4)	(5)	(6)

FOR A HYPOTHESIS ON ONE POPULATION VARIANCE

- (1) From Claim
- (2) From VAR on data
- (3) Function COUNT on data
- (4) = (3) - 1
- (5) = (4) * (2) / (1)
- (6) Function CHIDIST((5), (4))

$$\chi^2 = (n - 1) \times s^2 / \sigma^2$$

FOR A HYPOTHESIS ON CORRELATION COEFFICIENT

Name	Correlation Coefficient r	n	df	t	Probability
	(1)	(2)	(3)	(4)	(5)

- (1) Function CORREL(Range of x data, Range of y data) or from Claim

(2) Function COUNT(Range of x) or from Claim

(3) = (2) - 2

(4) = (1) / SQR((1 - (1) * (1)) / (3))

(5) Function TDIST((4), (3), 1 or 2 from Claim)

$$t = r / \sqrt{((1 - r^2) / (n - 2))}$$

TWO SAMPLE TESTS

FOR A HYPOTHESIS ON TWO POPULATIONS

MEANS, TWO INDEPENDENT SAMPLES

(1) Range of data set 1

(2) Range of data set 2

(3) Tails, either 1 or 2

1 is for a one tailed test

2 is for a two tailed test

(4) Characteristics of the two data sets

1 is for paired data values, having equal numbers of values (no missing values)

2 is for the equal variance (homoscedastic) characteristic

3 is for the unequal variance (heteroscedastic) characteristic

(5) Probability = TTEST((1), (2), (3), (4))

(6)

DIFFERENCES, PAIRED DEPENDENT SAMPLES

$$t = (\bar{d} - \mu_d) / (s_d / \sqrt{n})$$

Equal Variance

$$t = \{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\} / s_m$$

$$s_m = s_p \times \sqrt{(1/n_1 + 1/n_2)}$$

$$s_p = \sqrt{\text{(pooled variance)}}$$

$$\text{(pooled variance)} = \{ (df \times s^2)_1 + (df \times s^2)_2 \} / df_{\text{Total}}$$

Unequal Variance

$$t = \{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\} / s_m$$

$$s_m = \sqrt{\{ (s^2/n)_1 + (s^2/n)_2 \}}$$

PROPORTIONS, RANDOM INDEPENDENT LARGE SAMPLES

($n_1p > 5$, $n_1q > 5$, $n_2p > 5$ and $n_2q > 5$)

Subscript	Name	p	q	n	np	nq	Number of successes	p-hat	SD	z	Probability
1		(1)	(2)	(3)	(4)	(5)	(6)	(7)			
2		(8)	(9)	(10)	(11)	(12)	(13)	(14)			
Combined		(15)	(16)	(17)			(18)		(19)	(20)	(21)

(1) From Claim

$$(2) = 1 - (1)$$

(3) From Claim

$$(4) = (1) * (3)$$

$$(5) = (2) * (3)$$

(7) From Claim

$$(8) = (6) / (3)$$

(8) From Claim

$$(9) = 1 - (1)$$

(10) From Claim

$$(11) = (1) * (3)$$

$$(12) = (2) * (3)$$

(13) From Claim

$$(14) = (6) / (3)$$

$$(17) = (3) + (10)$$

$$(18) = (6) + (13)$$

$$(15) = (18) / (17)$$

$$(16) = 1 - (15)$$

$$(19) = \text{SQR}((15) * (16) * (1/(3) + 1/(10)))$$

$$(20) = ((7) - (14) - (1) + (8)) / (19)$$

(21) Function NORMSDIST ((20))

$$z = \{(p\text{-hat}_1 - p\text{-hat}_2) - (p_1 - p_2)\} / \sigma_m$$

$$p = (x_1 + x_2) / (n_1 + n_2)$$

$$q = 1 - p$$

$$\sigma_m = \sqrt{ \{ (p \times q / n_1) + (p \times q / n_2) \}}$$

VARIANCES, TWO INDEPENDENT SAMPLES

For a hypothesis on two population variances

- (1) Range of data set 1
- (2) Range of data set 2
- (3) Probability = FTEST((1), (2))

$$F = s^2_1 / s^2_2$$

The FTEST function can be used here, but it gives incorrect p values.

DATA ANALYSIS ROUTINE OUTPUTS

The data analysis routines output the following table:

df	Degrees of freedom
t Stat	Calculated t value
P(T<=t) one-tail	P1
t Critical one-tail	
P(T<=t) two-tail	P2
t Critical two-tail	

You have to translate the P1 and P2 values to determine the probability of the given hypothesis being true. The output table is confusing here.

Hy pot hesi s	Symbol	Actual Data Values B>A	Actual Values B=A	Actual Data Values B<A	Actual Data Values B≠A
1	B>A	1 - p1	p1	p1	
2	B>=A	1 - p1	p1	p1	
3	B<A	p1	p1	1 - p1	
4	B<=A	p1	p1	1 - p1	
5	B=A		p2		1 - p2
6	B≠A		1 - p2		p2

MAKING A CONCLUSION

- a. The calculated probability values have to be looked at carefully. All the calculations are sign sensitive and therefore it is important to be sure that the calculated probability value lies in the correct tail or correct part of the probability distribution. Make sketches of the probability distributions and where the correct

- z or t value lies based on the data. This may require a true probability value to be one minus the calculated probability value.
- b. Accepting a true or false condition on the hypothesis depends on the probability value calculated from the test and comparing it to a present probability level (alpha). For hypothesis testing, commonly used alpha values are 0.1, 0.05, 0.01 or 0.001. Excel's default is 0.05, but other values can be used.
 - c. Excel's approach reflects an earlier period, when a fixed alpha value was taken as a position, and either a t or F value corresponding to the alpha was obtained as a reference point. The test was to compare the sample calculated values to the reference value to make a statement about the condition of the hypothesis. Today the tests are based on the calculated probability values and the probability values are compared to the preselected alpha value.
 - i. If Probability $> \alpha$: Fail to reject H_0
 - ii. If Probability $\leq \alpha$: Reject H_0
 - d. If the test is a true hypothesis test, the calculated p value is NOT reported, only the decision is reported.
 - e. Decision: Reject H_0 :
 - i. H_0 is the claim: There is enough evidence at ... to reject the claim of ...
 - ii. H_a is the claim: There is enough evidence at ... to support the claim of ...
 - f. Decision: Fail to Reject H_0 :
 - i. H_0 is the claim: There is not enough evidence at ... to reject the claim of ...
 - ii. H_a is the claim: There is not enough evidence at ... to support the claim of ...
 - g. Changing a hypothesis "in mid stream" presents problems in assigning an appropriate p value under the Neyman-Pearson concept. This requires now an inductive approach to the problem, evaluating multiple hypotheses. This is beyond Excel. Goodman (1999) argues then that a Bayesian approach be taken, in which the likelihood ratios of each hypothesis on the data be calculated and combined with prior probability structures for obtaining the correct probability values. Excel does not have the capability to obtain likelihood ratios from a set of data.