CHAPTER 9: HYPOTHESIS TESTING

THE SECOND LAST EXAMPLE CLEARLY ILLUSTRATES THAT THERE IS ONE IMPORTANT ISSUE WE NEED TO EXPLORE: IS THERE (IN OUR TWO SAMPLES) SUFFICIENT STATISTICAL EVIDENCE TO CONCLUDE THAT (THE POPULATIONS OF) BOYS AND GIRLS REACT DIFFERENTLY TO VACCINATION (THE FACT THAT THE TWO SAMPLE PROPORTIONS ARE DIFFERENT DOES NOT MEAN MUCH BY ITSELF, SINCE RANDOM DIFFERENCES WILL PRACTICALLY ALWAYS MAKE THEM SO).

A SIMILAR ISSUE ARISES IN JUST ABOUT EVERY CASE WE STUDIED SO FAR IN THE CONTEXT OF CONFIDENCE INTERVALS. LET US GO OVER THEM AGAIN, ONE BY ON.

< THE POPULATION MEAN : (LARGE SAMPLES).

QUITE OFTEN, THERE IS A 'SPECIAL' VALUE OF :, E.G. APPLE TREES IN A CERTAIN ORCHARD BEAR APPLES OF THE AVERAGE WEIGHT OF 500g (CLAIMS THE OWNER), AND WE WOULD LIKE TO **TEST** THIS **HYPOTHESIS**.

WE WILL CALL IT THE **NULL** HYPOTHESIS, AND USE THE FOLLOWING NOTATION:

H_0: : = 500g

NEXT, WE HAVE TO DECIDE WHAT ARE WE TESTING IT <u>AGAINST</u> (THE **ALTERNATE** HYPOTHESIS).

THERE ARE SOME CASES WHEN WE NEED TO VERIFY THE EXACT AMOUNT (AN ASPIRIN TABLET CONTAINING 350 mg OF ASA) - IN THAT SITUATION THE ALTERNATE HYPOTHESIS WOULD BE **TWO-TAILED**, I.E. **H_1: : ... 350mg**.

THERE ARE OTHER CASES WHEN WE GLADLY ACCEPT MORE (E.G. BIGGER APPLES), BUT ANYTHING SMALLER WILL BE CONSIDERED A MISREPRESENTATION. IN THAT CASE, THE ALTERNATE HYPOTHESIS IS **LEFT-TAILED**, I.E. **H_1: : < 500g**.

WE MAY ALSO HAVE THE OPPOSITE CASE: THE WATER PURIFYING PLANT CLAIMS REDUCING POLLUTANTS TO A 18ppm LEVEL (OUR NULL HYPOTHESIS), BUT WE GLADLY ACCEPT LESS. THE ALTERNATE HYPOTHESIS WILL THEN BE **RIGHT-TAILED**: ***H_1*****: : > 18ppm**.

THE TEST ITSELF (DECIDING WHETHER WE CAN BELIEVE ***H_0*** OR ***H_1***) IS BASED ON COMPUTING A VALUE OF A RANDOM VARIABLE WHOSE DISTRIBUTION (UNDER ***H_0***) IS KNOWN. IN THE CURRENT CASE (TESTING THE VALUE OF :), THIS WILL BE

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

(CALLED **TEST STATISTIC**), WHERE

$:_0$ IS THE VALUE CLAIMED BY THE NULL HYPOTHESIS (500g, 150mg, 18ppm, ETC.). WE KNOW ALREADY THAT THE DISTRIBUTION OF THIS RANDOM VARIABLE IS STANDARD NORMAL

WHEN THE EXACT VALUE OF F IS KNOWN (NOT TOO COMMON), WE USE IT IN THE PREVIOUS FORMULA INSTEAD OF s .

THEN, WE SELECT THE **LEVEL OF SIGNIFICANCE** α (THE OPPOSITE OF THE CONFIDENCE LEVEL), WITH 0.05, 0.01, 0.10 BEING THE MOST COMMON CHOICES, AND IN TABLE 5(c) WE LOOK UP THE CORRESPONDING CRITICAL VALUE z_c (SAME AS WHEN CONSTRUCTION CONFIDENCE INTERVALS, EXCEPT: IN ONE-TAILED TESTS, THE FULL 0.05 PROBABILITY IS ASSIGNED TO A SINGLE TAIL).

WE **REJECT** THE NULL HYPOTHESIS (IN FAVOR OF $\boldsymbol{H_1}$) WHENEVER THE TEST STATISTIC ENTERS THE **CRITICAL REGION** (THE ONE OR TWO TAILS BEYOND z_c).

EXAMPLE: TO CHECK THE CLAIM THAT THE AVERAGE WEIGHT OF APPLES IS 500g, WE SELECT A RANDOM INDEPENDENT SAMPLE OF SIZE 75, RESULTING IN A SAMPLE MEAN OF 489g AND A SAMPLE STANDARD DEVIATION OF 83g. WE WANT TO PERFORM THIS TEST USING THE USUAL 5% SIGNIFICANCE LEVEL.

CLEARLY: $\boldsymbol{H_0}$**:** μ **= 500g** , $\boldsymbol{H_1}$**:** μ **< 500g** AND $\boldsymbol{z_c}$ **= -1.645**
(THE LAST LINE OF TABLE 6, UNDER αN).
THE VALUE OF THE TEST STATISTIC EQUALS:

$$\frac{489-500}{83/\sqrt{75}} = -1.148$$

CONCLUSION: WE DO NOT HAVE ENOUGH STATISTICAL EVIDENCE TO REJECT THE NULL HYPOTHESIS (THUS, WE MUST 'ACCEPT' IT). THIS OF COURSE DOES NOT PROVE THE VALIDITY OF H_0 - WE MAY YET BE ABLE REJECT IT, BY LARGER SAMPLING.

NOTE THAT FOR THIS (LEFT-TAILED) TEST, A SAMPLE MEAN OF 500g OR BIGGER WOULD AUTOMATICALLY YIELD 'ACCEPTING' H_0 .

ALSO NOTE THAT THE TWO HYPOTHESES ARE TREATED VERY DIFFERENTLY: THE NULL HYPOTHESIS ALWAYS GETS THE 'BENEFIT OF THE DOUBT', WHEN OUR OBSERVATIONS FALL IN THE 'NOT-SO-CLEAR' REGION. THE ALTERNATE HYPOTHESIS, ON THE OTHER HAND, MUST PROVE ITS CASE 'BEYOND A REASONABLE DOUBT'. THIS IS BECAUSE THE CONSEQUENCES OF REJECTING H_0 ARE USUALLY MORE SERIOUS THAN THOSE OF 'ACCEPTING' (IT'S BETTER TO SAY: 'FAILING TO REJECT') IT.

A FEW MORE TERMS TO REMEMBER:

REJECTING $\boldsymbol{H_0}$ WHEN IT'S <u>TRUE</u> (I.E. INCORRECTLY) IS MAKING **TYPE I ERROR**. WE KNOW THAT THE PROBABILITY OF THIS IS SET TO " (AND IT THUS, IN A SENSE, 'UNDER CONTROL').

<u>FAILING</u> TO REJECT $\boldsymbol{H_0}$ WHEN IT'S <u>FALSE</u> IS CALLED **TYPE II ERROR**. ITS PROBABILITY (DENOTED $) DEPENDS ON THE ON THE ACTUAL VALUE OF : (THE MORE : DEVIATES FROM THE NULL-HYPOTHESIS' CLAIM, THE SMALLER THE CHANCES OF MAKING A TYPE II MISTAKE).

THE PROBABILITY OF <u>REJECTING</u> $\boldsymbol{H_1}$ WHEN IT'S FALSE (THE <u>CORRECT</u> COURSE OF ACTION) IS THUS EQUAL TO 1 - $ AND IT'S CALLED THE **POWER** OF THE TEST (WE WOULD LIKE IT TO BE <u>LARGE</u>).

NOTE THAT: AS : APPROACHES $:_0$ (THE NULL-HYPOTHESIS' CLAIM), THE POWER OF THE TEST REACHES THE VALUE OF " (FOR THE 'APPLE' EXAMPLE, IT IS GOING TO BE DIFFICULT TO REJECT $\boldsymbol{H_0}$ WHEN : = 499g).

THERE IS A DIFFERENT (BUT, IN THE FINAL ANALYSIS, <u>EQUIVALENT</u>) WAY OF DECIDING WHETHER TO REJECT H_0 OR NOT, BASED ON THE SO CALLED P VALUE. THIS REQUIRES LOOKING UP (IN TABLE 5) THE <u>TAIL</u> AREA (OF THE STANDARD NORMAL DISTRIBUTION) CORRESPONDING TO THE VALUE OF $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ (THE TEST STATISTIC). FOR TWO-TAILED TESTS, THIS AREA MUST FURTHER BE MULTIPLIED BY 2.

THE ACTUAL DECISION (WHETHER TO REJECT $\boldsymbol{H_0}$, OR NOT) IS THEN MADE DEPENDING ON THIS P BEING LESS THAN " (REJECT) OR BIGGER THAN " (DON'T).

THIS HAS THE ADVANTAGE OF KNOWING EXACTLY HOW STRONG OUR REJECTION IS.

USING THE PREVIOUS EXAMPLE, THE *P* VALUE CORRESPONDING TO 1.148 IS 1.0000 - 0.8749 = 12.51%. SINCE THIS IS BIGGER THAN 5% (OR EVEN 10%), WE CANNOT REJECT H_0: : = 500g (SAME CONCLUSION AS BEFORE).

NOW, WE MODIFY THE PROCEDURE TO ACCOMMODATE THE OTHER POSSIBILITIES OF THE LAST CHAPTER.

FIRSTLY, WE EXTEND THE PREVIOUS CASE (TESTING WHETHER A SPECIFIC VALUE OF : IS CORRECT OR NOT) TO COVER THE SITUATION OF A SMALL SAMPLE SIZE. SIMILARLY TO CONSTRUCTING CONFIDENCE INTERVALS, WE NOW HAVE TO ASSUME THAT THE POPULATION ITSELF IS NORMAL. EVERYTHING ELSE REMAINS THE SAME, EXCEPT THE CRITICAL VALUE(S) WILL BE LOOKED UP IN THE TABLE OF THE *t* - DISTRIBUTION, RATHER THAN *z*.

EXAMPLE: TO CHECK THE CLAIM THAT THE AVERAGE WAGE IN A CERTAIN INDUSTRY IS $40,000 ANNUALLY, WE SELECT A RANDOM SAMPLE OF 22 OF EMPLOYEES AND ESTABLISH THEIR SAMPLE MEAN TO BE $36,274, WITH THE CORRESPONDING SAMPLE STANDARD DEVIATION OF $8,409.

DOES THIS CONSTITUTE A STATISTICALLY SIGNIFICANT EVIDENCE THAT THE ORIGINAL CLAIM IS INCORRECT (EITHER WAY)? **H_0: μ = $40,000**, **$H_1$: $\mu \neq$ $40,000**

THE VALUE OF THE TEST STATISTIC IS $\dfrac{36274 - 40000}{8409/\sqrt{22}} = -2.078$

THE CORRESPONDING CRITICAL VALUES ARE ±2.080 (FROM TABLE 6 (ALSO IN INSERT), USING d.f. OF 21 AND α = 0.050).

CONCLUSION: NOT ENOUGH STATISTICAL EVIDENCE TO DISPROVE THE VALUE OF $40,000.

< TESTING THE VALUE OF p (PROPORTION OR PROBABILITY OF 'SUCCESS').

WE AGAIN HAVE TO ASSUME THAT n IS LARGE (IN THE $pn > 5$, $qn > 5$ SENSE). THE TEST STATISTIC IS

$$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

WHERE p_0 IS THE VALUE OF THE NULL HYPOTHESIS (NOTE THAT p_0, NOT $\hat{p}$, IS NOW USED IN THE DENOMINATOR). THIS TEST STATISTIC HAS, UNDER H_0, THE STANDARD NORMAL DISTRIBUTION.

EXAMPLE: WE SUSPECT THAT OUR OPPONENT IS USING A CROOKED DIE (PROBABILITY OF A SIX IS <u>BIGGER</u> THAN 1/6). WE BORROW THE DIE AND ROLL IT 1000 TIME, GETTING 213 SIXES. IS THERE ENOUGH STATISTICAL EVIDENCE TO ACCUSE HIM OF BEING DISHONEST?

WE HAVE: $\boldsymbol{H_0: p = 1/6,}$ $\boldsymbol{H_1: p > 1/6}$

THE VALUE OF THE TEST STATISTIC IS $\dfrac{0.213-\frac{1}{6}}{\sqrt{\frac{1}{6}\cdot\frac{5}{6}\cdot\frac{1}{1000}}} = 3.932$

THE CRITICAL REGION IS ANYTHING BEYOND 2.58 (THIS TIME WE DECIDED TO USE A VERY SMALL 0.5% LEVEL OF SIGNIFICANCE).

WE OBTAINED A <u>HIGHLY</u> SIGNIFICANT EVIDENCE THAT HS DIE IS CROOKED!

< TESTING FOR DIFFERENCE BETWEEN TWO POPULATION MEANS, HAVING LARGE, INDEPENDENT SAMPLES.

TYPICALLY, THE NULL HYPOTHESIS WOULD CLAIM THAT THE TWO MEANS HAVE THE SAME VALUE (H_0: $\mu_1 = \mu_2$).

THE TEST STATISTIC IS

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

(USING THE USUAL NOTATION). UNDER THE NULL HYPOTHESIS, THE TEST STATISTIC HAS THE STANDARD NORMAL DISTRIBUTION.

EXAMPLE: TO TEST (USING " = 1%) WHETHER MEN HAVE, ON THE AVERAGE, HIGHER BLOOD PRESSURE THAN WOMEN, WE COLLECT TWO INDEPENDENT SAMPLES, EACH OF SIZE 100. THE TWO SAMPLE MEANS TURN OUT TO BE 137 mm AND 126 mm, WITH STANDARD DEVIATIONS OF 32 mm AND 28 mm.

THE HYPOTHESES ARE: $\boldsymbol{H_0}$: $\mu_1 = \mu_2$ AND $\boldsymbol{H_1}$: $\mu_1 > \mu_2$

THE TEST STATISTIC EVALUATES TO $\dfrac{137 - 126}{\sqrt{\dfrac{32^2}{100} + \dfrac{28^2}{100}}} = 2.587$

IN TABLE 6 (UNDER 4 d.f.), WE FIND THAT THE CORRESPONDING CRITICAL VALUE IS 2.326.

THERE IS SUFFICIENT STATISTICAL EVIDENCE TO CONCLUDE MEN HAVE, ON THE AVERAGE, HIGHER BLOOD PRESSURE THAT WOMEN.

DISCLAIMER: ALL OUR EXAMPLES ARE HYPOTHETICAL, WITH MADE UP DATA - THE CONCLUSIONS ARE NOT TO BE TAKEN SERIOUSLY!

TWO NOTES:

WHEN THE <u>EXACT</u> VALUES OF F_1 AND F_2 ARE KNOWN AND GIVEN, THESE ARE TO BE SUBSTITUTED FOR s_1 AND s_2.

WHEN THE NULL HYPOTHESIS CLAIMS SOMETHING LIKE THIS: $\boldsymbol{H_0}$: $:_1 = :_2 + 10$, THE NUMERATOR OF THE TEST STATISTIC MUST BE MODIFIED CORRESPONDINGLY, I.E. CHANGED TO $\bar{x}_1 - \bar{x}_2 - 10$.

< SMALL-SAMPLE MODIFICATIONS:

WE NEED BOTH POPULATIONS TO BE <u>NORMAL</u>, AND THE TWO (POPULATION) F ‘S TO EQUAL TO EACH OTHER.

THE TEST STATISTIC IS THEN

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

WHICH HAS, WHEN H_0 IS TRUE, THE t DISTRIBUTION WITH $n_1 + n_2 - 2$ DEGREES OF FREEDOM.

EXAMPLE: TO TEST WHETHER TWO CAR MAKES HAVE THE SAME FUEL CONSUMPTION (OR NOT), WE HAVE DRIVEN 10 CARS OF EACH MAKE OVER THE SAME TRACK, OBTAINING THE FOLLOWING RESULTS: THE TWO SAMPLE MEANS WERE 23.82L AND 22.41L, THE CORRESPONDING SAMPLE STANDARD DEVIATIONS 0.38L AND 0.43L. WE WANT TO USE 10% LEVEL OF SIGNIFICANCE.

H_0: $:_1 = :_2$, H_1: $:_1 … :_2$. THE TEST STATISTIC EVALUATES TO

$$\frac{23.82 - 22.41}{\sqrt{0.2} \times \sqrt{\frac{9 \times 0.38^2 + 9 \times 0.43^2}{10 + 10 - 2}}} = 7.77$$

THE CRITICAL VALUES (TABLE 6, d.f. = 18, "0 = 0.100) ARE ±1.734.

CONCLUSION: IN THIS CASE, THE TWO CONSUMPTION ARE CLEARLY DIFFERENT (AT JUST ABOUT ANY SIGNIFICANCE LEVEL).

< DIFFERENCE IN TWO PROPORTIONS (PROBABILITIES OF A 'SUCCESS').

THE TWO SAMPLES MUST BE INDEPENDENT, BOTH MUST BE 'LARGE'.

THE TEST STATISTIC IS

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\hat{p} \cdot \hat{q}}}$$

WHERE $\hat{p}$ / $\frac{r_1 + r_2}{n_1 + n_2}$ IS THE SO CALLED **POOLED** ESTIMATE OF THE PROBABILITY OF 'SUCCESS', AND $\hat{q} \equiv 1 - \hat{p}$.

ITS DISTRIBUTION IS, TO A GOOD APPROXIMATION, STANDARD NORMAL.

EXAMPLE: TEST, USING 5% LEVEL OF SIGNIFICANCE, WHETHER THE PROPORTION OF PEOPLE WITH BLOOD TYPE O IS THE SAME IN US AND JAPAN.

$\boldsymbol{H_0}$**:** $\boldsymbol{p_1 = p_2}$**,** $\boldsymbol{H_1}$**:** $\boldsymbol{p_1 \ldots p_2}$. SUPPOSE THAT OUR SAMPLING RESULTED IN 137 SUCH PEOPLE OUT OF 400 IN US, AND 98 OUT OF 300 IN JAPAN. THE VALUE OF THE TEST STATISTIC IS THUS

$$\frac{137/400 - 98/300}{\sqrt{1/300 + 1/400} \times \sqrt{235/700 \times 465/700}} = 0.439$$

THE CRITICAL VALUES ARE ± 1.96 .

CONCLUSION: WE DON'T HAVE STATISTICALLY SIGNIFICANT EVIDENCE TO REJECT THE HYPOTHESIS THAT THE TWO PROPORTIONS ARE IDENTICAL.

< TESTS INVOLVING **PAIRED** DIFFERENCES

THIS IS THE ONLY SITUATION WHICH WE DID NOT LEARN TO CONSTRUCT CONFIDENCE INTERVALS FOR.

TYPICALLY, IT INVOLVES AN EXPERIMENT IN WHICH TWO OBSERVATIONS ARE TAKEN FOR EACH SUBJECT (E.G. BEFORE AND AFTER A MEDICATION IS TAKEN). WE ARE USUALLY GIVEN TWO SETS OF (PAIRED) DATA, AND MUST FIRST REALIZE THAT THESE ARE NOT TWO INDEPENDENT SAMPLES.

THE NEXT THING IS TO COMPUTE THE DIFFERENCES OF EACH PAIR OF VALUES (SOME MAY TURN OUT POSITIVE, SOME NEGATIVE), AND PRETTY MUCH FORGET THE ORIGINAL DATA.

THE NULL HYPOTHESIS USUALLY STATES THAT THE (POPULATION) MEAN DIFFERENCE IS ZERO ($:_{d} = 0$).

THE TEST STATISTIC IS: $$\frac{\bar{d}}{s_d / \sqrt{n}}$$

WHERE $\bar{d}$ IS THE SAMPLE MEAN OF THE DIFFERENCES, AND s_d IS THE CORRESPONDING SAMPLE STANDARD DEVIATION.

ASSUMING THE DIFFERENCE DISTRIBUTION TO BE NORMAL, THE TEST STATISTIC HAS THE t DISTRIBUTION WITH n - 1 DEGREES OF FREEDOM (THE STANDARD NORMAL DISTRIBUTION WHEN $n > 30$).

EXAMPLE: A CERTAIN BLOOD-PRESSURE MEDICATION IS BEING TESTED ON 12 RANDOMLY SELECTED INDIVIDUALS. THEIR BLOOD PRESSURE IS RECORDED BEFORE AND AFTER THEY TAKE THIS MEDICATION; THESE ARE THE RESULTS:

B:	143	128	160	148	139	172	144	150	138	153	180	163
A:	128	132	144	139	137	140	125	138	139	139	161	129

FIRST, WE NEED TO COMPUTE THE DIFFERENCES: 15, -4, 16, 9, 2, 32, 19, 12, -1, 14, 19, 34, THEIR MEAN: $\overline{d} = \frac{167}{12} = 13.92$ AND

STANDARD DEVIATION: $s_d = \sqrt{\dfrac{3825 - 167^2/12}{11}} = 11.68$

THE RESULTING VALUE OF THE TEST STATISTIC IS:

$$\frac{13.92}{11.68/\sqrt{12}} = 4.128$$

THE HYPOTHESES ARE: H_0: $\mu_d = 0$ AND H_1: $\mu_d > 0$ (ONE-TAILED TEST).

USING 1% LEVEL OF SIGNIFICANCE (AND 11 DEGREES OF FREEDOM), WE FIND (TABLE 6) THAT THE CRITICAL VALUE IS 2.718.

CONCLUSION: WE HAVE A HIGHLY SIGNIFICANT PROOF THAT THE MEDICATION IS EFFECTIVE.

NOTE: SHOULD THE NULL HYPOTHESIS CLAIM THAT $\mu_d = 20$ (NOT VERY COMMON) WE WOULD HAVE TO MODIFY OUR TEST STATISTIC TO $\dfrac{\overline{d} - 20}{s_d/\sqrt{n}}$ (EVERYTHING ELSE WOULD BE THE SAME).