EX: A SAMPLE OF 169 FISH IS RANDOMLY SELECTED FROM A LARGE POPULATION. FISH LENGTH X IS DISTRIBUTED WITH $\mu = 50cm$, $\sigma = 26cm$. FIND THE PROBABILITY THAT THE SAMPLE'S MEAN IS BETWEEN 46 AND 48 cm.

1) $n = 169 > 30$, SO WHAT KIND OF DISTRIBUTION IS THE SAMPLE MEAN?

(ALMOST) NORMAL.

2) IN THAT CASE, WHAT ARE $\mu_{\bar{x}}$ AND $\sigma_{\bar{x}}$?

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{169}} = \frac{\sigma}{13}.$$

$$= 50cm \qquad = \frac{26}{13} = 2\,cm$$

3) $P(46 \leq \bar{x} \leq 48) = ?$

$$P\left(\frac{46-50}{2} \leq \frac{\bar{x}-50}{2} \leq \frac{48-50}{2}\right)$$

$$= P(-2 \leq Z \leq -1).$$

4) FROM Z-TABLES, $P(Z \leq -1) = 0.1587$

$P(Z \leq -2) = 0.0228$

$\Rightarrow P(-2 \leq Z \leq -1) = 0.1587 - 0.0228 = \boxed{0.1359}$

HYPOTHESIS TESTING

BASED ON THE DATA, WE FORM A HYPOTHESIS AND WANT TO TEST WHETHER IT'S TRUE. USUALLY, 2 DATA SETS ARE COMPARED, OR ONE IS COMPARED TO A SYNTHETIC DATA SET. THE COMPARISON IS SIGNIFICANT IF THE RELATIONSHIP IS NOT LIKELY WITHIN A CERTAIN LIKELIHOOD CALLED THE SIGNIFICANCE LEVEL.

THE PROCEDURE IS AS FOLLOWS.

1) STATE A NULL AND AN ALTERNATIVE HYPOTHESIS, H_0 AND H_1.

2) CONSIDER ANY ASSUMPTIONS ABOUT THE DATA. (INDEPENDENCE, TYPE OF DISTRIBUTION, ETC.)

3) DECIDE WHAT TEST TO USE, AND STATE THE TEST STATISTIC T.

4) USE THE ASSUMPTIONS TO FIND THE DISTRIBUTION OF T UNDER H_0. (USUALLY IT'S NORMAL OR STUDENT'S t).

5) CHOOSE A SIGNIFICANCE LEVEL α, USUALLY 1% OR 5%. (H_0 IS REJECTED IF THE OBSERVED VALUE OF T IS BELOW α.)

6) FIND THE CRITICAL REGION: THE VALUES OF T FOR WHICH H_0 IS REJECTED.

7) ACCEPT OR REJECT H_0.

ALTERNATIVE METHOD:

1) COMPUTE THE OBSERVED VALUE OF T.

2) CALCULATE THE p-VALUE: THE PROBABILITY OF A SAMPLE T AT LEAST AS EXTREME AS THE OBSERVED T UNDER H_0.

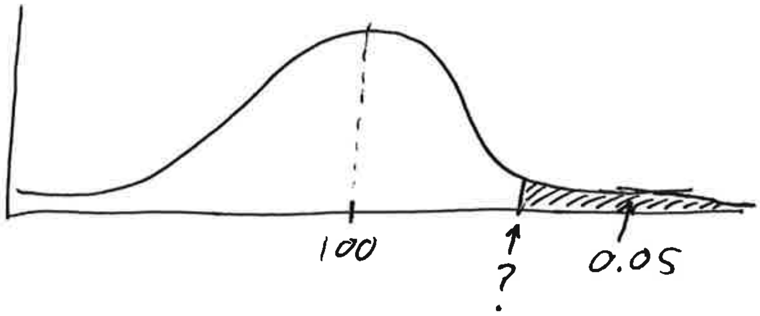
3) REJECT H_0 IF AND ONLY IF $p < \alpha$.

EX: A SCHOOL PRINCIPAL CLAIMS HIS STUDENTS ARE ABOVE AVERAGE INTELLIGENCE. A RANDOM SAMPLE OF 30 STUDENTS YIELDS AN AVERAGE IQ OF 112. THE MEAN IQ IS 100, STANDARD DEVIATION 15. IS THERE SUFFICIENT EVIDENCE TO SUPPORT THE CLAIM?

$H_0: \overline{X} = 100$ (THE ACCEPTED FACT IS THAT $\mu = 100$)

$H_1: \overline{X} > 100$.

CHOOSE α: IF IT'S NOT GIVEN, GO WITH $\alpha = 0.05$.



FROM THE Z-TABLE, $P(z > 0.95) = 1.645$

THE TEST STATISTIC IS $Z = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$. IF THIS NUMBER IS IN THE SHADED REJECTION REGION, WE REJECT H_0.

$$Z = \frac{112 - 100}{15/\sqrt{30}} \simeq 4.38.$$

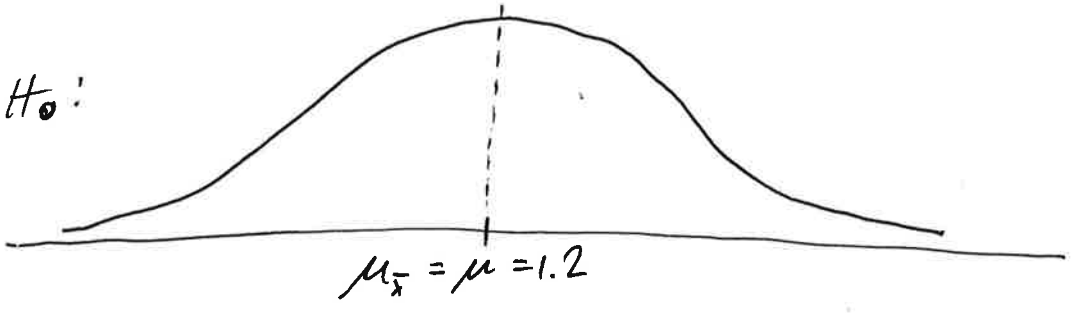
SINCE $4.38 > 1.645$, REJECT H_0 IN FAVOUR OF H_1. THESE STUDENTS ARE ABOVE AVERAGE INTELLIGENCE (MOST LIKELY).

EX: A NEUROLOGIST IS TESTING THE EFFECT OF A DRUG ON RESPONSE TIME BY INJECTING 100 RATS WITH A UNIT DOSE, SUBJECTING EACH TO A STIMULUS, AND RECORDING RESPONSE TIMES. THE MEAN RESPONSE TIME FOR RATS NOT INJECTED IS 1.2s. THE MEAN RESPONSE TIME FOR THE INJECTED RATS IS 1.05s, WITH STANDARD DEVIATION 0.5s. DO YOU THINK THE DRUG HAS AN EFFECT ON RESPONSE TIME? USE $\alpha = 0.01$.

H_0: DRUG HAS NO EFFECT (i.e. $\mu = 1.2$ WITH OR WITHOUT THE DRUG).

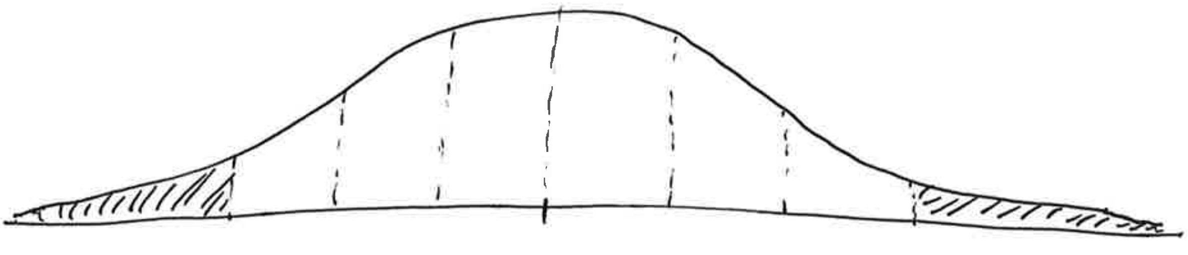
H_1: DRUG HAS AN EFFECT ($\mu \neq 1.2$ WHEN DRUG IS ADMINISTERED).

ASSUME H_0 IS TRUE. IF THE PROBABILITY OF GETTING THE OBSERVED RESULTS IS TOO SMALL, WE REJECT IT.



WE DON'T HAVE σ, BUT WE HAVE A GOOD SAMPLE SIZE, SO WE APPROXIMATE σ BY $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}} \approx \frac{s}{\sqrt{100}} = \frac{0.5}{10} = 0.05$.

$$Z = \frac{1.2 - 1.05}{0.05} = 3$$ (3 STANDARD DEV. AWAY FROM μ)



FROM THE Z TABLE, THE SHADED REGION IS 0.3% OF THE AREA. (P-VALUE IS 0.003).

$p < \alpha$, SO WE REJECT H_0 AND CONCLUDE THAT THE DRUG DOES HAVE AN EFFECT.

χ^2 TEST FOR ASSOCIATION

ALSO KNOWN AS THE χ^2 TEST FOR INDEPENDENCE, IT IS USED TO DETERMINE IF THERE'S A RELATIONSHIP BETWEEN 2 RVs. THE χ^2 TEST CAN BE USED ONLY IF YOU HAVE 2 CATEGORICAL VARIABLES THAT EACH HAVE 2 OR MORE CATEGORICAL GROUPS. EXAMPLES: GENDER, AGE GROUPS, ETHNICITY, INCOME LEVEL, ETC.

THE PROCEDURE:

1) MAKE A 2-WAY TABLE.

2) H_0: THERE IS NO ASSOCIATION BETWEEN THE VARIABLES.

   H_1: THERE IS SOME ASSOCIATION.

3) FIND EXPECTED VALUES, THEN USE $T^2 = \sum \frac{(OBS - EXP)^2}{EXP}$.

4) CALCULATE p-VALUE, COMPARE TO α.

   THE p-VALUE IS $P(\chi^2 \geq T^2)$, THE PROBABILITY OF OBSERVING A VALUE AT LEAST AS EXTREME AS THE TEST STATISTIC FOR A χ^2-SQUARE DISTRIBUTION WITH $(r-1)(c-1)$ DEGREES OF FREEDOM, WHERE r AND c ARE THE NUMBER OF ROWS AND COLUMNS IN THE TABLE.

5) THERE IS NO ASSOCIATION (i.e. ACCEPT H_0) IF $p > \alpha$. OTHERWISE, REJECT H_0 IN FAVOUR OF H_1.

EX: STUDENTS ARE ASKED IF GRADES, ATHLETICS OR POPULARITY IS MOST IMPORTANT TO THEM, IN YEARS 4, 5 AND 6. THE RESULTS:

GOALS	YEAR 4	YEAR 5	YEAR 6	TOTAL
GRADES	49	50	69	168
SPORTS	19	22	28	69
POPULARITY	24	36	38	98
TOTAL	92	108	135	335

SO THE 2 VARIABLES ARE THE YEAR OF STUDY OF THE STUDENTS, AND THEIR GOALS. THEREFORE, AS H_0 IS "NO ASSOCIATION BETWEEN VARIABLES",

H_0: THE STUDENTS' CHOICES OF GOALS DO NOT DEPEND ON THEIR YEAR OF STUDY.

H_1: THE STUDENTS' CHOICES OF GOALS DEPEND ON THEIR YEAR OF STUDY.

EXPECTED VALUES: (RECALL $E_{ij} = \frac{(\text{TOTAL}_{\text{ROW } i})(\text{TOTAL}_{\text{COLUMN } j})}{\text{TOTAL}}$.

GOALS	4	5	6
GRADES	46.1	54.2	67.7
SPORTS	18.9	22.2	27.8
POPULARITY	26.9	31.6	39.5

EX: $E_{11} = \frac{168 \cdot 92}{335} \approx 46.1$

$$T^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(49-46.1)^2}{46.1} + \frac{(50-54.2)^2}{54.2} + \dots + \frac{(38-39.5)^2}{39.5} = 1.5$$

THE DEGREES OF FREEDOM : $(3-1)(3-1)=4$.

FROM χ^2 TABLES, $P(\chi^2 \leq 1.51) = 0.8244$ WITH 4 DF.
CHOOSING $\alpha = 0.05$ (A TYPICAL CHOICE), WE SEE THAT $p > \alpha$, SO WE ACCEPT H_0. THERE IS MOST LIKELY (95% CONFIDENCE) NO SIGNIFICANT RELATIONSHIP BETWEEN YEAR OF STUDY AND CHOICE OF GOAL.

NOTE : TO GET AN IDEA OF WHETHER A RELATIONSHIP EXISTS, FIND PERCENTAGES OF THE 2-WAY TABLE :

GOALS	4	5	6
GRADES	53 %	46 %	51 %
SPORTS	21 %	21 %	21 %
POPULARITY	26 %	33 %	28 %

EX : A PUBLIC OPINION POLL SURVEYED A RANDOM SAMPLE OF 1000 VOTERS. GENDER AND POLITICAL AFFILIATION WERE RECORDED AS FOLLOWS :

	REPUBLICAN	DEMOCRAT	INDEPENDENT
MALE	200	150	50
FEMALE	250	300	100

DO THE MEN'S VOTING PREFERENCES DIFFER SIGNIFICANTLY FROM THE WOMEN'S ? USE A 0.05 LEVEL OF SIGNIFICANCE.

$H_0 = ?$

$H_1 = ?$

TOTALS:

	REP.	DEM.	IND.	TOTAL
MALE	200	150	50	
FEMALE	250	300	100	
TOTAL				

$df = (2-1)(3-1) = 2.$

$E_{11} =$ $\qquad$ $E_{21} =$

$E_{12} =$ $\qquad$ $E_{22} =$

$E_{13} =$ $\qquad$ $E_{23} =$

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \ldots$$

(SHOULD BE 16.2).

$P(\chi^2 \geq 16.2)$ WITH 2 df IS (FROM TABLE) 0.0003.

SINCE $0.0003 < 0.05$, WE REJECT H_0 AND CLAIM THAT THERE IS EVIDENCE TO SUPPORT H_1: GENDER HAS AN EFFECT ON VOTING PREFERENCE.

REVIEW STUDY GUIDE

CALCULUS

1) WHAT'S A FUNCTION?

A RULE THAT ASSIGNS EACH ELEMENT OF A SET A TO ONE ELEMENT OF A SET B. WHICH OF THESE ARE FUNCTIONS?

a) $f: \mathbb{Q} \to \mathbb{Q},\ f(x) = \frac{x}{2}$.

b) ON $\mathbb{R}$, $x = y^2$.

c) ON $\mathbb{R}_+$, $x = y^2$.

2) WHAT'S A SEQUENCE?

A FUNCTION WITH DOMAIN $\mathbb{N}$.

a) $\{2, 4, 6, 8, \ldots\}$

b) $f(n) = 2n$, $a_n = 2n$

3) WHAT'S A SERIES?

THE SUM OF ELEMENTS OF A SEQUENCE.

$$\sum_{i=1}^{10} a_i = a_1 + a_2 + \ldots + a_{10}$$

4) ALGEBRA OF FINITE SUMS.

5) ARITHMETIC SEQUENCE: COMMON DIFFERENCE.

GEOMETRIC SEQUENCE: COMMON RATIO.

APPLICATION: ANNUITY.

6) COMBINATORICS
    - FACTORIALS
    - PERMUTATIONS/COMBINATIONS
    - BINOMIAL THEOREM
7) DOMAIN/RANGE OF FUNCTIONS
    - POLYNOMIALS
    - PLOTTING FUNCTIONS
    - CONTINUOUS/INCREASING/DECREASING FUNCTIONS
8) DILATION, TRANSLATION AND REFLECTION OF FUNCTIONS
9) COST-PROFIT ANALYSES
    - CONTRIBUTION MARGIN/RATE
    - BREAK-EVEN ANALYSIS
10) DERIVATIVES
    - CONSTANT/LINEAR/POLYNOMIAL FUNCTIONS
11) INTEGRALS (ANTIDERIVATIVES)
    - DEFINITE/INDEFINITE INTEGRALS
    - POLYNOMIALS
12) EXPONENTIAL/LOGARITHMIC FUNCTIONS
13) NET PRESENT VALUE, INTERNAL RATE OF RETURN

DATA ANALYSIS

1) • DISCRETE VS. CONTINUOUS DATA.
   • NOMINAL VS. ORDINAL DATA.

2) • MEAN, MEDIAN, MODE.
   • QUARTILES, OUTLIERS, IQR

   (REMEMBER, QUARTILES ARE CALCULATED DIFFERENTLY, AS IS THE MEDIAN, FOR EVEN AND ODD DATA SETS).

3) VARIANCE AND STANDARD DEVIATION, SYMMETRY AND SKEW.

4) PLOTS
   - BOX-AND-WHISKER
   - MOSAIC
   - HISTOGRAM
   - STEM-AND-LEAF
   - SCATTER (WITH LEAST-SQUARES LINE), CORRELATION

5) TABLES
   - TWO-WAY
   - FREQUENCY

PROBABILITY

1) VENN DIAGRAMS : INTERSECTION, UNION, COMPLEMENT

2) INDEPENDENCE, DISJOINTNESS

3) CONDITIONAL PROBABILITY, TREE DIAGRAMS, BAYES' RULE

4) BINOMIAL DISTRIBUTION, BINOMIAL THEOREM

5) GEOMETRIC DISTRIBUTION, POISSON DISTRIBUTION

6) CUMULATIVE DISTRIBUTION FUNCTIONS

7) FIND EXPECTED VALUE AND VARIANCE FOR ALL DISTRIBUTIONS

8) CONTINUOUS PROBABILITY DISTRIBUTIONS

- EXPONENTIAL
- NORMAL: $\left(Z = \frac{x - \mu}{\sigma}\right)$ CONVERTS TO STANDARD NORMAL
- UNIFORM
- STUDENT'S-t
- χ^2

9) FITTING DATA TO MODELS

- SMALL/LARGE SAMPLES, SAMPLE MEAN AND VARIANCE
- SAMPLING DISTRIBUTION OF THE MEAN (ALWAYS NORMAL FOR BIG n)
- CENTRAL LIMIT THEOREM
- CONFIDENCE INTERVALS
- HYPOTHESIS TESTING
  - NORMAL OR STUDENT'S-t TEST (DEPENDS ON SAMPLE SIZE AND σ KNOWN/UNKNOWN)
  - χ^2-TEST FOR ASSOCIATION.