STA 442: Methods of Applied Statistics

STA1008: Applications of Statistics

Data File

- Rows are **cases**
- Columns are **variables**

1	2	2	0	78.0	65	80	39	English	Female	3	3	1
2	2	6	2	66.0	54	75	57	English	Female	3	3	1
3	2	4	4	80.2	77	70	62	English	Male	5	6	1
4	2	5	2	81.7	80	67	76	English	Female	2	2	1
5	2	4	4	86.8	87	80	86	English	Male	5	5	1
6	2	3	1	76.7	53	75	60	English	Male	3	3	1
7	2	3	2	85.8	86	81	54	Other	Female	2	2	1
8	2	4	3	73.0	75	77	17	English	Male	4	5	1
9	2	6	2	72.3	63	60	2	English	Male	4	4	1
10	2	8	6	90.3	87	88	76	English	Male	4	4	1
11	2	8	3	.	.	.	60	English	Male	1	2	1
12	2	6	4	.	.	.	61	Other	Female	1	1	1
13	.	.	.	87.2	84	83	54	English	Male	3	3	1
14	2	2	5	91.0	90	91	84	English	Male	5	5	1
15	2	3	1	72.8	53	74	.	English	Female	3	3	1
16	.	.	.	80.7	72	84	14	English	Male	3	3	1
17	2	5	0	82.5	82	85	75	Other	Female	2	2	1
18	2	4	6	91.5	95	81	94	English	Female	3	3	1
19	2	3	2	78.3	77	74	60	English	Female	3	3	1
20	.	.	.	74.5	0	85	.	English	Male	4	4	1
21	2	3	3	80.7	71	78	53	Other	Female	1	3	1
22	2	5	3	88.3	80	85	63	English	Female	3	3	1
23	2	4	2	76.8	82	64	82	Other	Female	2	2	1

Skipping ....

570	2	5	4	84.8	88	68	80	English	Male	1	1	1
571	2	4	3	78.3	83	84	56	English	Male	4	2	1
572	2	6	3	88.3	81	90	70	English	Female	5	5	1
573	2	3	1	.	.	.	.	English	Male	3	3	1
574	2	5	9	77.0	73	79	60	English	Female	2	2	1
575	.	.	.	78.7	80	73	.	English	Female	6	3	1
576	2	5	2	80.7	80	70	50	Other	Male	1	1	1
577	2	4	2	80.7	56	81	50	English	Female	2	2	1
578	2	4	3	.	.	.	78	Other	Female	4	4	1
579	1	6	1	82.2	80	86	61	English	Female	2	2	1

id	mcg	r	day	AML	AMS	AM1d	PML	PMS	PM1d	AMslp	PMslp	SWeight
1	198	1	1	0.6	.	.	0.8	.	.	.	.	
2	198	1	2	1.8	.	.	2.8	.	.	.	.	
3	198	1	3	4.7	1	.	6.1	1	.	.	.	
4	198	1	4	7.8	4	2.0	8.7	5	2.1	.	.	
5	198	1	5	11.2	6	1.8	12.1	7	2.0	.	.	
6	198	1	6	14.3	12	1.9	15.0	11	1.4	.	.	
7	198	1	7	17.5	12	2.1	18.5	13	1.6	.	.	
8	198	1	8	20.9	19	1.1	21.9	19	1.7	.	.	
9	198	1	9	24.0	22	1.6	25.2	22	1.3	.	.	
10	198	1	10	27.2	26	2.1	28.4	26	1.2	.	.	
11	198	1	11	30.7	28	1.4	32.3	28	1.5	.	.	
12	198	1	12	.	31	.	.	31	.	.	.	
13	198	1	13	.	37	.	.	36	.	.	.	
14	198	1	14	.	37	.	.	38	.	3.11	3.18	0.5996
15	198	2	1	0.5	.	.	0.6	.	.	.	.	
16	198	2	2	1.4	.	.	2.3	.	.	.	.	
17	198	2	3	4.15	1	.	5.6	1	.	.	.	
18	198	2	4	7.4	2	2.0	8.7	4	2.1	.	.	
19	198	2	5	10.8	5	2.2	12.0	8	2.0	.	.	
20	198	2	6	14.2	10	1.7	15.3	13	1.6	.	.	
21	198	2	7	17.1	13	2.2	18.1	16	1.7	.	.	
22	198	2	8	21.3	18	1.1	22.2	18	1.4	.	.	
23	198	2	9	24.4	27	1.4	25.6	24	1.2	.	.	
24	198	2	10	27.6	26	2.1	28.8	28	1.2	.	.	
25	198	2	11	31.2	29	1.9	32.5	29	1.3	.	.	
26	198	2	12	.	33	.	.	36	.	.	.	
27	198	2	13	.	38	.	.	41	.	.	.	
28	198	2	14	.	42	.	.	42	.	3.21	3.26	0.6040

Variables can be

- Quantitative - representing amount of something, like Income, BP, BMI, GPA (?)
- Categorical - Codes represent category membership, like Gender, Nationality, Marital status, Alive vs. dead

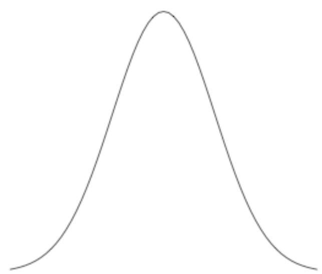
Variables can be

- Independent: Predictor or partial cause
- Dependent: Predicted or effect

We will often pretend that our data represent a **random sample** from some **population**. We will carry out formal procedures for making inferences about this (usually fictitious) population, and then use them as a basis for drawing conclusions about the data.

- **Statistics**: Numbers that can be calculated from sample data
- **Parameters**: Numbers that could be calculated if we knew the whole population

Distribution = Population Histogram



Conditional Distribution

For each value *x* of the independent variable *X*, there is a separate distribution of the dependent Variable *Y*. This is called the conditional distribution of *Y* given *X*=*x*.

Example: Conditional distribution of height given Gender = F.

Definition of "Related"

- We will say that the independent and dependent variables are **unrelated** if the conditional distribution of the dependent variable is identical for each value of the independent variable.
- If the distribution of the dependent variable does depend on the value of the independent variable, we will describe the two variables as **related**.

Testing Statistical Significance

- Are IV and DV "really" related?
- **Null Hypothesis**: They are unrelated in the population

Reasoning

Suppose that the independent and dependent variables are actually unrelated in the population. If this null hypothesis is true, what is the probability of obtaining a sample relationship between the variables that is as strong or stronger than the one we have observed? If the probability is small (say, $p < 0.05$), then we describe the sample relationship as **statistically significant**, and it is socially acceptable to discuss the results.

P-value

- The probability of getting our results (or better) just by chance.
- The minimum significance level at which the null hypothesis can be rejected.

We can be wrong

- Type I error: H_0 is true, but we reject it
- Type II error: H_0 is false, but we fail to reject it

Power is the probability of *correctly* rejecting H_0

- Power = 1 - P(Type II Error)
- Power increases with true strength of relationship, and with sample size
- Power can be used to select sample size in advance of data collection

Confidence Interval: Pair of numbers chosen so that the probability they will enclose the parameter (or function of parameters) is large, like 0.95

Should we Accept H_0?

- When the results are not statistically significant, usually we will say that the data do not provide enough evidence to conclude that the variables are related.
- See text for more details

Many statistical methods assume
Independent Observations

- Simple random sampling
- Cases are not linked, do not "communicate"
- If the design involves non-independence, allow for it

Elementary Tests

- Independent (two-sample) t-test
- Matched (paired) t-test
- One-way ANOVA
- Simple regression and correlation
- Chi-square test of independence