Statistical packages (SAS)

Web Page: http://im.pwr.wroc.pl/~mbogdan

- Instructor : Małgorzata Bogdan

- **Office** : 205 C-11
- **Office hours**: Wed 13:30-14:30, Fr 13:30-15:30
- or by appointment.
- **phone**: 320-2008
- **Email**: Malgorzata.Bogdan@pwr.wroc.pl

Grades

- lab reports (50%)
- test 1 (25%) April 04
- test 2 (25%) May 23

Grades

- 90 – 100 = 5
- 80 – 89 = 4.5
- 70 – 79 = 4.0
- 55 – 69 = 3.5
- 30 – 54 = 3
- Submission of all lab reports is the necessary condition for a positive grade.

References

- Introduction to the Practise of Statistics by
- D.S.Moore, G.P.McCabe
- Applied Linear Statistical Models, (5th ed.), by Kutner, Nachtsheim, Neter and Li

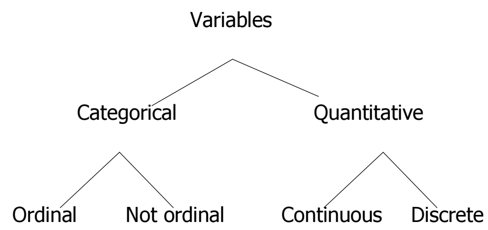
Lecture 1

- Displaying data with graphs
- Descriptive statistics
- Basics of testing

Individuals and variables

- Individuals – objects described by a set of data (people, animals, things)
- Variable – characteristic of an individual

Types of Variables



Types of variables

- Categorical – outcomes fall in to categories
  - Ordinal: choices on a survey ; never, rarely, occasionally, often, always
  - Not ordinal:
  - round & yellow, round & green, wrinkled & yellow, wrinkled & green
  - gender, race, job type

- Quantitative – outcome is a number
  - Continuous : height, weight, concentration
  - Discrete : number of flowers on a plant, number of round & yellow peas

Information on employees of CyberStat

	A	B	C	D	E	F
1	Name	Job Type	Age	Gender	Race	Salary
2	Cedillo, Jose	Technical	27	Male	White	52,300
3	Chambers, Tonia	Management	42	Female	Black	112,800
4	Childers, Amanda	Clerical	39	Female	White	27,500
5	Chen, Huabang	Technical	51	Male	Asian	83,600
6						

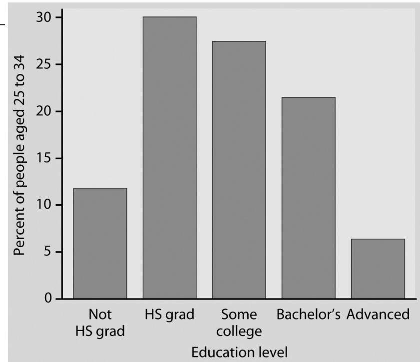
Ready NUM

Exploratory data analysis - graphs

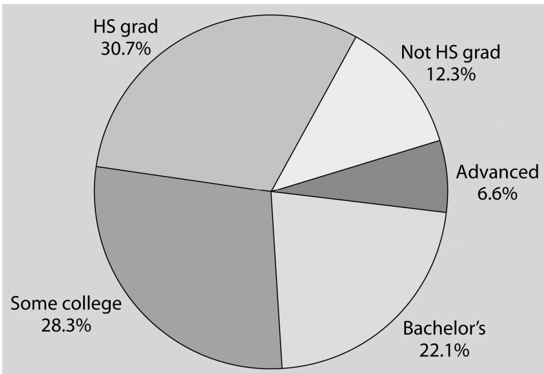
- We begin by examining each variable by itself.
- Categorical variables
- Distribution – gives the count or the percent of individuals in each category.

Education	Count (in milions)	Percent
Less than high school	4.7	12.3
High school graduate	11.8	30.7
Some college	10.9	28.3
Bachelor's degree	8.5	22.1
Advanced degree	2.5	6.6

Bar graph



Pie chart



Quantitative variable - Stemplot

Stem – all but the final digit

Leaf – the final digit

Example 1

Numbers of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Examining distributions

- Describe the pattern – shape, center and spread.
- Shape –
  - How many modes ?
  - Symmetric or skewed in one direction.
- Center – midpoint
- Spread –range between the smallest and the largest values.
- Look for outliers – individual values that do not match the overall pattern.

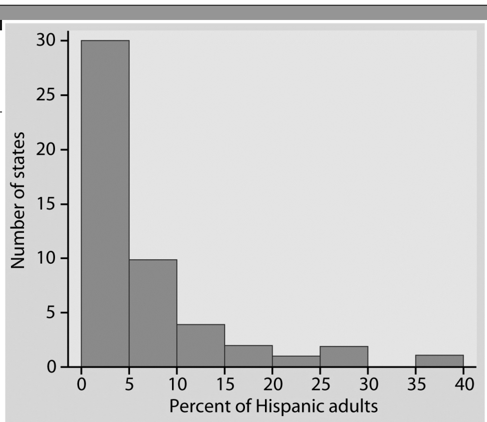
Histograms

TABLE 1.2 Percent of Hispanics in the adult population, by state (2000)

State	Percent	State	Percent	State	Percent
Alabama	1.5	Louisiana	2.4	Ohio	1.6
Alaska	3.6	Maine	0.6	Oklahoma	4.3
Arizona	21.3	Maryland	4.0	Oregon	6.5
Arkansas	2.8	Massachusetts	5.6	Pennsylvania	2.6
California	28.1	Michigan	2.7	Rhode Island	7.0
Colorado	14.9	Minnesota	2.4	South Carolina	2.2
Connecticut	8.0	Mississippi	1.3	South Dakota	1.2
Delaware	4.0	Missouri	1.8	Tennessee	2.0
Florida	16.1	Montana	1.6	Texas	28.6
Georgia	5.0	Nebraska	4.5	Utah	8.1
Hawaii	5.7	Nevada	16.7	Vermont	0.8
Idaho	6.4	New Hampshire	1.4	Virginia	4.2
Illinois	10.7	New Jersey	12.3	Washington	6.0
Indiana	3.1	New Mexico	38.7	West Virginia	0.6
Iowa	2.3	New York	13.8	Wisconsin	2.9
Kansas	5.8	North Carolina	4.3	Wyoming	5.5
Kentucky	1.3	North Dakota	1.0		

Frequency Table

Class	Count	Percent	Class	Count	Percent
0.1-5.0	30	60	20.1-25	1	2
5.1-10.0	10	20	25.1-30	2	4
10.1-15	4	8	30.1-35	0	0
15.1-20	2	4	35.1-40	1	2



Describing distributions with numbers

- Mean
- Median
- Quartiles
- Boxplots
- Standard deviation

SAS programs: Program 1

```
data popstruct;
input state $ percent;
 cards;
AL   1.5
AK   3.6
AZ   21.3
……..
WY  5.5 ;
run;
```

Program 2

- **data** popstruct;
- infile ‘c:\mbogdan\ECMI\data\ta01_002.txt' DLM='09'x;
- input state $ percent;
- run;

- **proc print** data=popstruct;
- **run**;

```
data deaths;
 input cause $ numdeath;
 cards;
 accident 13602
 homicide 4989
 suicide 3885
 cancer 1724
 heartdis 1048
 congenit 430
 respirat 208
AIDS 197;
run;
```

Program 3

```
proc gchart data=deaths;
 vbar cause / freq=numdeath;
run;
proc gchart data=deaths;
 pie cause / freq=numdeath;
run;
```

Program 4

```
data reading;
infile ' c:\mbogdan\ECMI\data\ex01_026.txt';
 input drp;
run;
proc univariate data=reading plot;
 var drp;
run;
```

```
proc gchart data=reading;
 vbar drp/type=pct midpoints=14 to 54 by 4;
run;
proc univariate data=reading;
 histogram drp/ midpoints=14 to 54 by 4;
run;
```

Tests of Significance

- The scheme of reasoning
- Stating hypotheses
- Test statistics
- P-values
- Statistical significance
- Test for population mean
- Two-sided test and confidence intervals

Tests of Significance-Hypothesis Testing

This common type of inference is used to assess the evidence provided by the data in favor of or against some claim (hypothesis) about the population…

…rather than to estimate unknown population parameter, for which we would use confidence intervals.

Examples for hypothesis testing:

1. Does the mean content of a drug equal to 198mg based on SRS of n=100 observations contradict the manufacture's claim that it is 200mg with standard deviation 5mg?
2. Are less than 15% of all CCD sensors produced by a particular manufacturer defective?

Example 1: Manufacturer claims mean content 200mg with SD of 5mg (active ingredient per pill). We study 100 pills; get average 201.65mg. Is it consistent with the claim?

Example 1 cntd.: What about the sample mean equal to 199mg or 200.5mg?
Are the outcomes *likely* or *significant*?

Stating Hypotheses

- The hypothesis is a statement about the **parameters in a population** or model. Not about the data at hand.
- The results of a test are expressed in terms of a **probability** that measures how well the **data and the hypothesis agree**.
- In hypothesis testing, we need to state two hypotheses:
  - The **null** hypothesis $\boldsymbol{H_0}$
  - The **alternative** hypothesis $\boldsymbol{H_a}$

Null hypothesis:

- The null hypothesis is the claim which is initially favored or believed to be true. Often **default** or uninteresting **situation** of "no effect" or "no difference".

We usually need to determine if there is a strong enough evidence **against it**.

- The test of significance is designed to assess this strength of the evidence against the null hypothesis.

Alternative hypothesis:

- The alternative hypothesis is the claim that we "hope" or "suspect" is true instead of H_0.
- We often begin with the alternative hypothesis H_a and then set up H_0 as the statement that the hoped-for effect is not present.

Example 1 ctnd. (interpretation):

H_0: $\mu = 200$

In words: Mean content is 200mg a pill.

H_a: $\mu \neq 200$

In words: Mean content is not 200mg.

A so-called **two-sided** alternative H_a.

(Looking for a departure each direction.)

Example 1 ctnd (other possible settings):

- H_0: $\mu = 200$ vs. H_a: $\mu < 200$

Suspect the content too low. **One-sided** H_a.

- H_0: $\mu = 200$ vs. H_a: $\mu > 200$

Suspect the content too high. **One-sided** H_a.

- H_0: $\mu \leq 200$ vs. H_a: $\mu > 200$

Virtually same as the previous. **One-sided** H_a.

Note: decide on the setting **before** you see the data based on general knowledge or **other** measurements.

Example 1. Interpretation ctnd. Test statistics:

- If the mean content is 200mg and SD=5mg, then

$$\frac{\overline{X} - 200}{0.5}$$

has (approx.) standard normal distribution.

Example 1. Interpretation ctnd. P-value.

- If H_0 is true, what is the probability of having the average of 100 contents as far off from 200 as 201.65?
- 199?
- 200.5?

P-value…

is the **probability**, computed assuming that H_0 is true, that the **test statistics** would take **as extreme or more extreme values** as the one actually observed.

This is the **P-value of the test** (or of the data, given the testing procedure). **If** it is **small**, it serves as **an evidence against H_0**.

Need to know the distribution of the test statistics under H_0 to calculate P-value.

Statistical Significance:

- We need a **cut-off point** (decisive value) that we can compare our P-value to and draw a conclusion or make a decision.
- This cut-off point is the significance level. It is announced in advance and serves as a standard on how much evidence against H_0 we need to reject H_0. Usually denoted α.
- Typical values of α: **0.05, 0.01**.
- If not stated otherwise, take α=0.05.

Statistical Significance

- When **P-value ≤α**, we say that the data are statistically significant at level α i.e. we have significant evidence against the null hypothesis.

Note:

- data with a P-value of 0.02 are statistically significant at level 0.05, but not at level 0.01.

The conclusion/decision:

- If the **P-value is smaller than** a fixed **significance level α** then we **reject the null hypothesis** (in favor of the alternative).
- Otherwise we don't have enough evidence to reject the null.
- Note: Report P-value with your conclusion.

Example 1ctnd. Statement of the conclusion:
(Give it in the natural language! Include P-value.)

z Test for a Population Mean General Setting:

- $X_1,\ldots, X_n$: SRS from (approximately) N(μ, σ)
- σ is given, μ is the unknown parameter of interest
- the null hypothesis is

  H_0: $\mu = \mu_0$
- the alternative hypothesis could be:

  H_a: $\mu \neq \mu_0$ *(two-sided)*

  H_a: $\mu > \mu_0$ *(one-sided)*

  H_a: $\mu < \mu_0$ *(one-sided)*

Test statistics for population mean when data are N(μ, σ) and σ is known:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

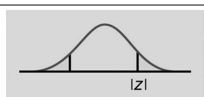
Notes:

- Also called z-test.
- If H_0 is true, this z has standard normal distribution--we expect small values of z.

z Test for a Population Mean P-value

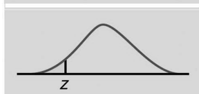
against...

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|) = P(|Z| \geq |z|)$



$H_a: \mu > \mu_0$ is $P(Z \geq z)$

$H_a: \mu < \mu_0$ is $P(Z \leq z)$



z Test for a Population Mean Decision

Reject H_0 when the P-value is smaller than significance level α.

Do not reject otherwise.

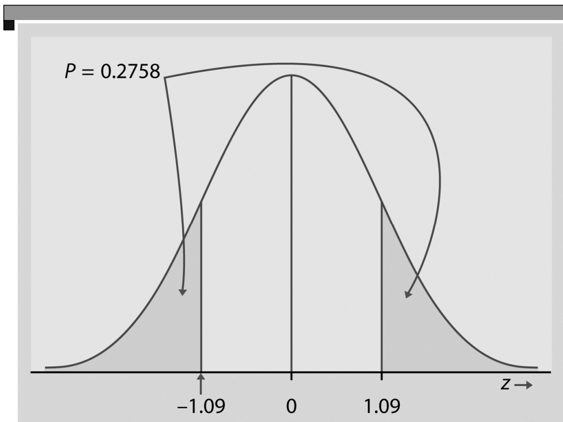
This rule is valid in other settings, too.

One-sided vs. two-sided

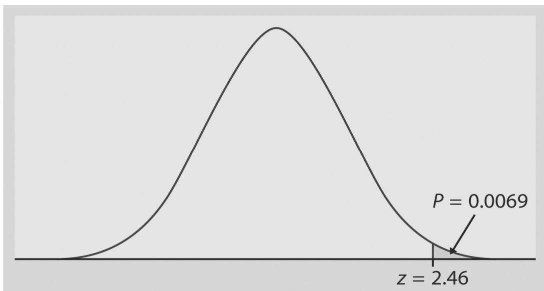
- If, based on previous data or experience we expect "**increase", "more", "better"** etc ("decrease", "less", "worse", resp.), then we can use one sided test.
- Otherwise, by default, we use two-sided. Key words: **"different", "departures", "changed"…**

Example 2: A group of 72 male executives in age group 35-44 has mean systolic blood pressure 126.07. Is this career group's mean pressure **different** than that of the general population of males in this age group, which is N(128, 15)?

(α not given?? Take 0.05.)



Example 3: A new billing system will be cost effective only if the mean monthly account is **more** than $170. Accounts have SD= $65. A survey of 400 monthly accounts gave a mean of $178. Will the new system be cost-effective?



Two-sided test and confidence intervals

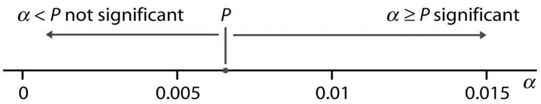
Example 1 (revisited): Find 95% confidence intervals when sample mean is 201.65mg (199mg, 200.5mg). Recall SD=5, n=100.

Note that the hypothesized μ=200mg is outside the first two and inside the third.

Two-sided test and confidence intervals

A level α **two-sided** significance test rejects H_0: $\mu=\mu_0$ exactly when μ_0 falls outside a level $1-\alpha$ confidence interval for μ.

P-value is the smallest level α at which the data are significant



Critical value

z^* such that the area (under the normal curve) to the right of it is a specified tail probability **p** is called **critical value** of (right) one-sided test (based on the normal distribution).

TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Examples: Find critical values for H_a: $\mu > \mu_0$ when p=0.05, p=0.02, p=0.01.

What are the P-values of z=1.5, z=2, z=2.5?