AIU FOR 6520

Statistical Research Design & Methods in Forensic Psychology

- **Instructor**: **Ivo Dinov,**

Asst. Prof. of Statistics, Neurology, Psychology

AIU, UCLA, Winter 2003

http://www.stat.ucla.edu/~dinov/courses_students.html

AIU FOR 6520

- ***Course Description***
- ***Class homepage***
- ***Online supplements, VOH's etc.***
- ***Final Exam/Project Format***

http://www.stat.ucla.edu/~dinov/courses_students.html

AIU FOR 6520

to just hear is to forget

to see is to remember

to do it yourself is to understand …

Review of Research & Design I – Fall'02

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Coverage of Research & Design II – Spring'03

- Applications of Central Limit Theorem, Law of Large Numbers.
- Design of studies and experiments.
- Fisher's F-Test & Analysis Of Variance (ANOVA, 1- or 2-way).
- Principle Component Analysis (PCA).
- $\chi 2$ (Chi-Square) Goodness-of-fit test.
- Multiple linear regression
- General Linear Model
- Bootstrapping and Resampling

Newtonial science vs. chaotic science

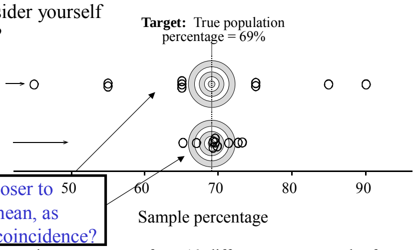
- Article by Robert May, Nature, vol. 411, June 21, 2001
  - Science we encounter at schools deals with crisp certainties (e.g., prediction of planetary orbits, the periodic table as a descriptor of all elements, equations describing area, volume, velocity, position, etc.)
  - As soon as uncertainty comes in the picture it <u>shakes the foundation of the deterministic science</u>, because only probabilistic statements can be made in describing a phenomenon (e.g., roulette wheels, chaotic dynamic weather predictions, Geiger counter, earthquakes, etc.)
  - What is then science all about – describing absolutely certain events and laws alone, or describing more general phenomena in terms of their behavior and chance of occurring? Or may be both!

Introduction to statistics

- **Intro to stats, vocabulary & intro to SPSS**
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Variation in sample percentages

Poll: Do you consider yourself overweight?



We are getting closer to The population mean, as $n \to \infty$ is this a coincidence?

Figure 1.1.1 Comparing percentages from 10 different surveys each of 20 people with those from 10 surveys each of 500 people (all surveys from same population).

Errors in Samples …

- Selection bias: Sampled population is not a representative subgroup of the population really investigated.
- Non-response bias: If a particular subgroup of the population studied does not respond, the resulting responses may be skewed.
- Question effects: Survey questions may be slanted or loaded to influence the result of the sampling.
- Is quota sampling reliable? Each interviewer is assigned a fixed quota of subjects (subjects district, sex, age, income exactly specified, so investigator can select those people as they liked).
- *Target population* –entire group of individuals, objects, units we study.
- *Study population* –a subset of the target population containing all "units" which could possibly be used in the study.
- *Sampling protocol – procedure used to select* ***the sample***
- *Sample –* the subset of "units" about which we actually collect info.

More terminology …

- ***Census – attempt to sample the entire population***
- ***Parameter – numerical characteristic of the population, e.g., income, age, etc. Often we want to estimate population parameters.***
- ***Statistic – a numerical characteristic of the sample. (Sample) statistic is used to estimate a corresponding population parameter.***
- Why do we sample at random? We draw "units" from the study population at random to avoid bias. Every subject in the study sample is equally likely to be selected. Also random-sampling allows us to calculate the likely size of the error in our sample estimates.

More definitions …

- How could you implement the lottery method to randomly sample 10 students from a class of 250? – list all names; assign numbers 1,2,3,…,250 to all students; Use a random-number generator to choose (10-times) a number in range [0;250]; Process students drawn.
- Random or chance error is the difference between the sample-value and the true population-value (e.g., 49% vs. 69%, in the above body-overweight example).
- Non-sampling errors (e.g., non-response bias) in the census may be considerably larger than in a comparable survey, since surveys are much smaller operations and easier to control.
- Sampling errors–arising from a decision to use a sample rather than entire population
- ***Unbiased procedure/protocol:*** (e.g., using the proportion of left-handers from a random sample to estimate the corresponding proportion in the population).
- **Cluster sampling**- a cluster of individuals/units are used as a sampling unit, rather than individuals.

More terminology …

- What are some of the ***non-sampling errors*** that plague surveys? (non-response bias, question effects, survey format effects, interviewer effects)
- If we take a random sample from one population, can we apply the results of our survey to other populations? (It depends on how similar, in the respect studied, the two populations are. In general- No! This can be a dangerous trend.)
- Are sampling households at random and interviewing people at random on the street valid ways of sampling people from an urban population? (No, since clusters (households) may not be urban in their majority.)
- Pilot surveys – after prelim investigations and designing the trial survey Q's, we need to get a "small sample" checking clearness and ambiguity of the questions, and avoid possible sampling errors (e.g., bias).

Questions …

- Give an example where non-representative information from a survey may be useful. Non-representative info from surveys may be used to estimate parameters of the actual sub-population which is represented by the sample. E.g., Only about 2% of dissatisfied customers complain (most just avoid using the services), these are the most-vocal reps. So, we can not make valid conclusions about the stereotype of the dissatisfied customer, but we can use this info to tract down changes in levels of complains over years.
- Why is it important to take a pilot survey?
- Give an example of an unsatisfactory question in a questionnaire. (In a telephone study: What time is it?

  Do we mean Eastern/Central/Mountain/Pacific?)

Questions …

- Random allocation – randomly assigning treatments to units, leads to representative sample only if we have large # experimental units.
- Completely randomized design- the simplest experimental design, allows comparisons that are unbiased (not necessarily fair). Randomly allocate treatments to all experimental units, so that every treatment is applied to the same number of units. E.g., If we have 12 units and 3 treatments, and we study treatment efficacy, we randomly assign each of the 3 treatments to 4 units exactly.
- Blocking- grouping units into blocks of similar units for making treatment-effect comparisons only within individual groups. E.g., Study of human life expectancy perhaps income is clearly a factor, we can have high- and low-income blocks and compare, say, gender differences within these blocks separately.

Questions …

- Why should we try to "blind" the investigator in an experiment?
- Why should we try to "blind" human experimental subjects?
- The basic rule of experimentor :

"Block *what you can and* randomize *what you cannot.*"

Experiments vs. observational studies for comparing the effects of treatments

- In an Experiment
  - experimenter determines which units receive which treatments. (ideally using some form of random allocation)
- Observational study – useful when can't design a controlled randomized study
  - compare units that happen to have received each of the treatments
  - Ideal for describing relationships between different characteristics in a population.
  - often useful for identifying possible causes of effects, but cannot reliably establish causation.
- Only properly designed and executed experiments can reliably demonstrate causation.

Questions …

- What is the difference between a designed experiment and an observational study? (no control of the design in observational studies)
- Can you conclude causation from an observational study? Why or why not? (not in general!)
- How do we try to investigate causation questions using observational studies? In a smoking-lung-cancer study: try to divide all subjects, in the obs. study, into groups with equal, or very similar levels of all other factors (age, stress, income, etc.) – I.e. control for all outside factors. If rate of lung-cancer is still still higher in smokers we get a stronger evidence of causality.
- What is the idea of controlling for a variable, and why is it used? Effects of this variable in the treatment/control groups are similar.
- Epidemiology – science of using statistical methods to find causes or risk factors for diseases.

The Subject of Statistics

Statistics is concerned with the process of finding out about the world and how it operates -

- in the face of variation and uncertainty
- by collecting and then making sense (interpreting) of data.

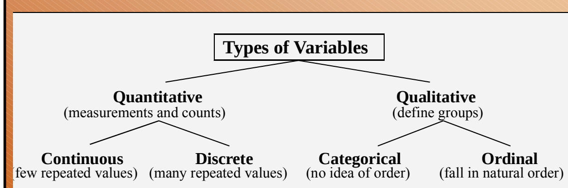
Displaying data

- Intro to stats, vocabulary & intro to SPSS
- **Displaying data**
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Types of variable

- ***Quantitative*** variables are ***measurements*** and counts
  - Variables with ***few*** *repeated values* are treated as ***continuous***.
  - Variables with ***many*** *repeated values* are treated as ***discrete***
- ***Qualitative*** variables (a.k.a. factors or class-variables) describe ***group membership***

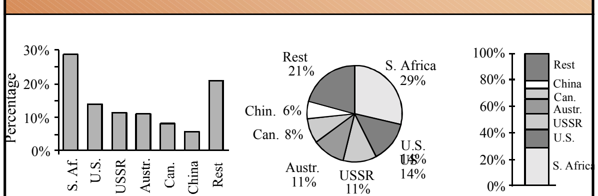
Distinguishing between types of variable



Questions …

- What is the difference between quantitative and qualitative variables?
- What is the difference between a discrete variable and a continuous variable?
- Name two ways in which observations on qualitative variables can be stored on a computer. (strings/indexes)
- When would you treat a discrete random variable as though it were a continuous random variable?
  - Can you give an example? ($34.45, bill)

Different graphs of the same set of numbers – percentages of the world's gold production in 1991

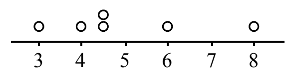


(a) Bar graph (b) Pie chart (c) Segmented bar

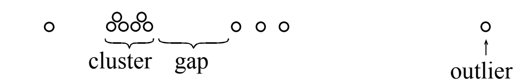
Questions …

- For what two purposes are tables of numbers presented? (convey information about trends in the data, detailed analysis)
- When should you round numbers, and when should you preserve full accuracy?
- How should you arrange the numbers you are most interested in comparing? (Arrange numbers you want to compare in columns, not rows. Provide written/verbal summaries/footnotes. Show row/column averages.)
- Should a table be left to tell its own story?

The dot plot

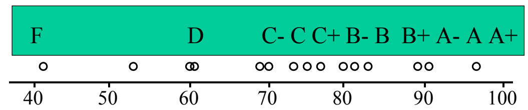


Dot plot.



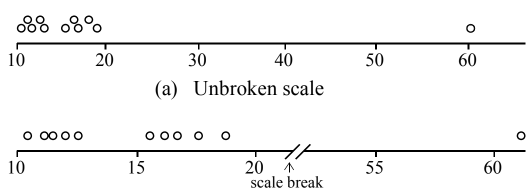
Dot plot showing special features.

Example of exploiting gaps and clusters



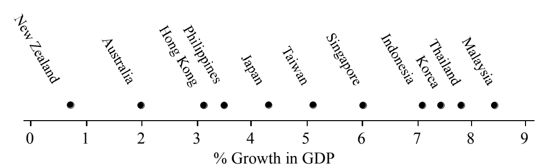
Grading of a university course.

Scale breaks



Dot plot with and without a scale break.

A labeled dot plot



Forecast of percent growth in GDP for 1990
for some South-East Asian and Pacific countries.

Example of a stem-and-leaf plot

Units: 7 | 2 =72

Values 52, 54 and their frequencies

```
Stem 1 | 7   Leafs
     2 | 7 9
     3 | 0 7 7 8 9
     4 | 2 2 3 3 4 5 8 8 9
     5 | 0 0 0 1 1 2 2 3 4 4 7 8 9 9
     6 | 0 0 0 2 3 3 3 3 6 7 7 8
     7 | 2 2
```

Stem-plot of the 45 obs's of the Ejection variable in the Heart Attack data table.

Traffic death-rates data

Traffic Death-Rates (per 100,000 Population) for 30 Countries

17.4 Australia	20.1 Austria	19.9 Belgium	12.5 Bulgaria	15.8 Canada
10.1 Czechoslovakia	13.0 Denmark	11.6 Finland	20.0 France	12.0 E. Germany
13.1 W. Germany	21.1 Greece	5.4 Hong Kong	17.1 Hungary	15.3 Ireland
10.3 Israel	10.4 Japan	26.8 Kuwait	11.3 Netherlands	20.1 New Zealand
10.5 Norway	14.6 Poland	25.6 Portugal	12.6 Singapore	9.8 Sweden
15.7 Switzerland	18.6 United States	12.1 N. Ireland	12.0 Scotland	10.1England & Wales

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)
Source: Hutchinson [1987, page 3].

Units: 17 | 4 = 17.4 deaths per 100,000

```
 5 | 4
 6 |
 7 |
 8 |
 9 | 8
10 | 1  1  3  4  5
11 | 3  6
12 | 0  0  1  5  6
13 | 0  1
14 | 6
15 | 3  7  8
16 |
17 | 1  4
18 | 6
19 | 9
20 | 0  1  1
21 | 1
22 |
23 |
24 |
25 | 6
26 | 8
```

(a)

Collapse to → 12 stems

Units: 1 | 7 = 17 deaths per 100,000

```
0 | 5
0 |
0 |
1 | 0  0  0  0  0  1  1
1 | 2  2  2  2  3  3  3
1 | 5  5
1 | 6  6  7  7
1 | 9
2 | 0  0  0  0  1
2 |
2 |
2 | 6  7
```

(b)

Two stem-and-leaf plots for the traffic deaths dat

Coyote Lengths Data (cm)

Females

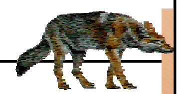
93.0	97.0	92.0	101.6	93.0	84.5	102.5	97.8	91.0	98.0	93.5	91.7
90.2	91.5	80.0	86.4	91.4	83.5	88.0	71.0	81.3	88.5	86.5	90.0
84.0	89.5	84.0	85.0	87.0	88.0	86.5	96.0	87.0	93.5	93.5	90.0
85.0	97.0	86.0	73.7								

Males

97.0	95.0	96.0	91.0	95.0	84.5	88.0	96.0	96.0	87.0	95.0	100.0
101.0	96.0	93.0	92.5	95.0	98.5	88.0	81.3	91.4	88.9	86.4	101.6
83.8	104.1	88.9	92.0	91.0	90.0	85.0	93.5	78.0	100.5	103.0	91.0
105.0	86.0	95.5	86.5	90.5	80.0	80.0					

Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

Frequency Table for Female Coyote Lengths

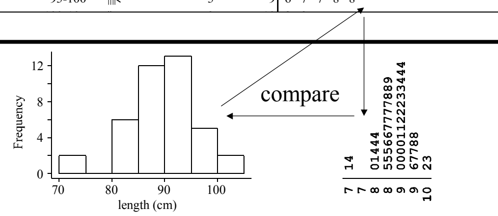


Body length

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75	||	2	7 | 1 4
75-80		0	7 |
80-85	卌 |	6	8 | 0 1 4 4 4
85-90	卌 卌 ||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95	卌 卌 |||	13	9 | 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100	卌	5	9 | 6 7 7 8 8
100-105	||	2	10 | 2 3
Total		40	

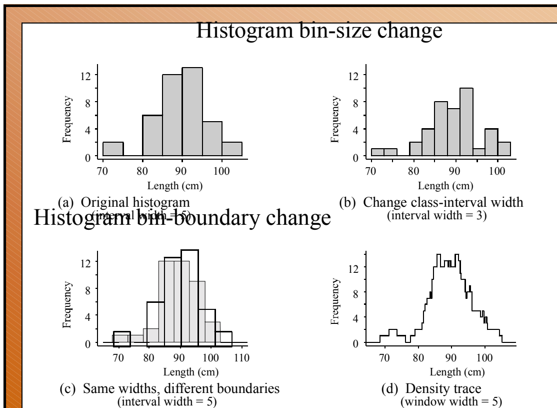
Frequency Table for Female Coyote Lengths

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75	||	2	7 | 1 4
75-80		0	7 |
80-85	卌 |	6	8 | 0 1 4 4 4
85-90	卌 卌 ||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95	卌 卌 |||	13	9 | 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100	卌	5	9 | 6 7 7 8 8



(a) Histogram (b) Stem-and-leaf plot rotated

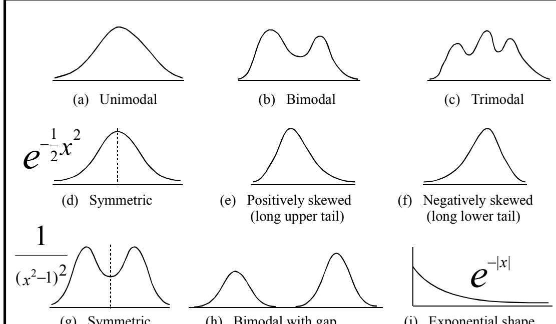
Histogram of the female coyote-lengths data.



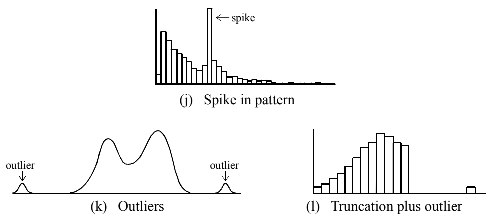
Questions …

- What advantages does a stem-and-leaf plot have over a histogram? (S&L Plots return info on individual values, quick to produce by hand, provide data sorting mechanisms. But, histograms are more attractive and more understandable).
- The shape of a histogram can be quite drastically altered by choosing different class-interval boundaries. What type of plot does not have this problem? (density trace) What other factor affects the shape of a histogram? (bin-size)
- What was another reason given for plotting data on a variable, apart from interest in how the data on that variable behaves? (shows features, cluster/gaps, outliers; as well as trends)

Interpreting Stem-plots and Histograms

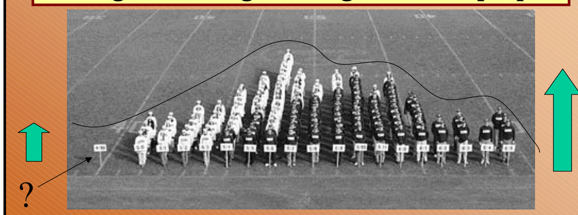


Interpreting Stem-plots and Histograms



Features to look for in histograms and stem-and-leaf plots.

Fascinations with histograms – Histogram of heights using the actual people



Subjects are university genetics students, females in white and males in dark tops.

Skewness & Kurtosis

- What do we mean by symmetry and positive and negative skewness? Kurtosis? Properties?!?

$$\text{Skewness} = \frac{\sum_{k=1}^{N}\left(Y_k - \overline{Y}\right)^3}{(N-1)SD^3}; \quad \text{Kurtosis} = \frac{\sum_{k=1}^{N}\left(Y_k - \overline{Y}\right)^4}{(N-1)SD^4}$$

- Skewness is linearly invariant Sk(aX+b)=Sk(X)
- Skewness is a measure of **unsymmetry**
- Kurtosis is (also linearly invariant) a measure of **flatness**
- Both are used to quantify departures from StdNormal
- Skewness(StdNorm)=0; Kurtosis(StdNorm)=3

Descriptive statistics from computer programs like STATA

STATA Output

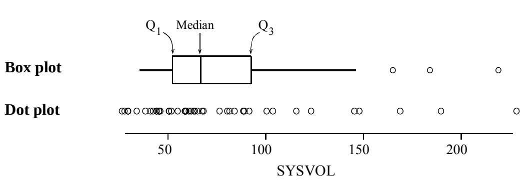
Standard deviation

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
age	45	50.133	51.000	50.366	6.092	0.908
Variable	Minimum	Maximum	Q1	Q3		
age	36.000	59.000	46.500	56.000		

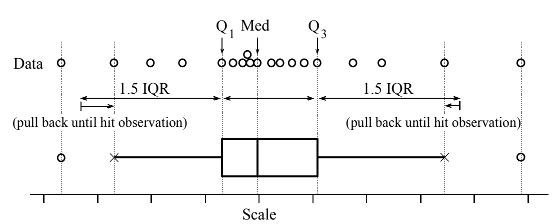
Lower quartile *Upper quartile*

Box plot compared to dot plot



Box plot for SYSVOL.

Construction of a box plot



Construction of a box plot.

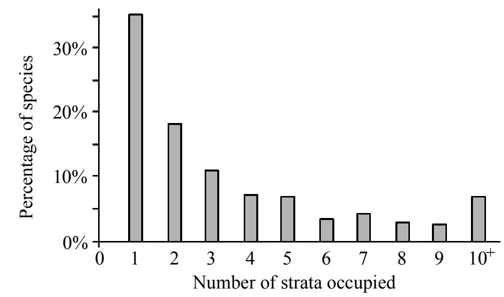
Frequency Table

TABLE 2.5.1 Word Lengths for the First 100 Words on a Randomly Chosen Page

3	2	2	4	4	4	3	9	9	3	6	2	3	2	3	4	6	5	3	4
2	3	4	5	2	9	5	8	3	2	4	5	2	4	1	4	2	5	2	5
3	6	9	6	3	2	3	4	4	4	2	2	4	2	3	7	4	2	6	4
2	5	9	2	3	7	11	2	3	6	4	4	7	6	6	10	4	3	5	7
7	7	5	10	3	2	3	9	4	5	5	4	4	3	5	2	5	2	4	2

Frequency Table

Value u_j	1	2	3	4	5	6	7	8	9	10	11
Frequency f_j	1	22	18	22	13	8	6	1	6	2	1



Bar graph for species data.

Central tendency and variability

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- **Central tendency and variability**
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Describing data with pictures and two numbers

- Random Number generation: frequency histogram
- Descriptive statistics
  - Central tendency (Mode, Median, Mean)
  - Variability (Variance, Standard deviation)

Random number generation [5:316]: ascending order

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
1	5	11	42	21	75	31	103	41	157	51	275
2	15	12	42	22	77	32	107	42	165	52	304
3	20	13	42	23	77	33	121	43	171	53	316
4	20	14	55	24	81	34	123	44	175		
5	25	15	57	25	83	35	125	45	188		
6	33	16	58	26	83	36	136	46	209		
7	35	17	58	27	91	37	140	47	213		
8	37	18	60	28	97	38	148	48	217		
9	42	19	62	29	101	39	152	49	248		
10	42	20	64	30	102	40	152	50	275		

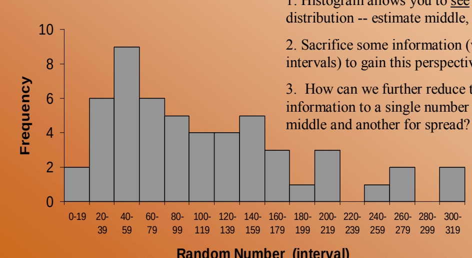
Histogram of random numbers generated

Notes:

1: Histogram allows you to see distribution -- estimate middle, spread.

2. Sacrifice some information (within intervals) to gain this perspective

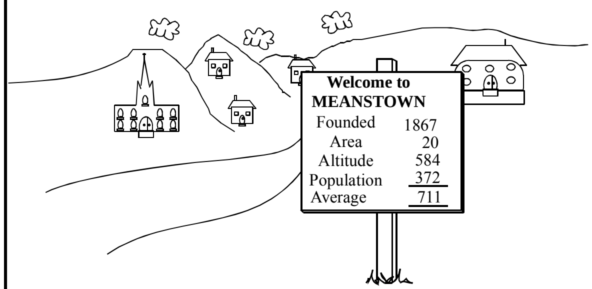
3. How can we further reduce this information to a single number for middle and another for spread?



Central tendency: the middle in a single number

- **Mode**: The most frequent score in the distribution.
- **Median**: The centermost score if there are an odd number of scores or the average of the two centermost scores if there are an even number of scores.
- **Mean**: The sum of the scores divided by the number of scores.

Beware of inappropriate averaging



Suggested by a 1977 cartoon in *The New Yorker* magazine by Dana Fradon.

Five number summary

The five-number summery = (Min, Q_1, Med, Q_3, Max)

Inter-quartile Range

$IQR = Q_3 - Q_1$

Calculate the measures of central tendency

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
1	5	11	42	21	75	31	103	41	157	51	275
2	15	12	42	22	77	32	107	42	165	52	304
3	20	13	42	23	77	33	121	43	171	53	316
4	20	14	55	24	81	34	123	44	175		
5	25	15	57	25	83	35	125	45	188		
6	33	16	58	26	83	36	136	46	209		
7	35	17	58	27	91	37	140	47	213		
8	37	18	60	28	97	38	148	48	217		
9	42	19	62	29	101	39	152	49	248		
10	42	20	64	30	102	40	152	50	275		

Mode:

Median:

Mean

Number of licks: sorted from few to many

Ss ID	umber	Ss ID	umber	Ss ID	umber	Ss ID	umber	Ss ID	umber	Ss ID	umber
1	5	11	42	21	75	31	103	41	157	51	275
2	15	12	42	22	77	32	107	42	165	52	304
3	20	13	42	23	77	33	121	43	171	53	316
4	20	14	55	24	81	34	123	44	175		
5	25	15	57	25	83	35	125	45	188		
6	33	16	58	26	83	36	136	46	209		
7	35	17	58	27	91	37	140	47	213		
8	37	18	60	28	97	38	148	48	217		
9	42	19	62	29	101	39	152	49	248		
10	42	20	64	30	102	40	152	50	275		

Mode: 42

Median = 91

Arithmetic mean

$\sum_{i=1}^{N} x_i / N$ = Sample mean = $\bar{X}$ (pronounced "Xbar")

$\sum_{i=1}^{N} x_i / N$ = Population mean = μ (pronounced "mew")

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
X_1	5	X_{11}	42	X_{21}	75	X_{31}	103	X_{41}	157	X_{51}	275
X_2	15	X_{12}	42	X_{22}	77	X_{32}	107	X_{42}	165	X_{52}	304
X_3	20	X_{13}	42	X_{23}	77	X_{33}	121	X_{43}	171	X_{53}	316
X_4	20	X_{14}	55	X_{24}	81	X_{34}	123	X_{44}	175		
X_5	25	X_{15}	57	X_{25}	83	X_{35}	125	X_{45}	188		
X_6	33	X_{16}	58	X_{26}	83	X_{36}	136	X_{46}	209		
X_7	35	X_{17}	58	X_{27}	91	X_{37}	140	X_{47}	213		
X_8	37	X_{18}	60	X_{28}	97	X_{38}	148	X_{48}	217		
X_9	42	X_{19}	62	X_{29}	101	X_{39}	152	X_{49}	248		
X_{10}	42	X_{20}	64	X_{30}	102	X_{40}	152	X_{50}	275		
										Σx =	5901

Median = 91; Mean =

Arithmetic mean

$\sum_{i=1}^{N} x_i / N$ = Sample mean = $\bar{X}$ (pronounced "Xbar")

$\sum_{i=1}^{N} x_i / N$ = Population mean = μ (pronounced "mew")

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
X_1	5	X_{11}	42	X_{21}	75	X_{31}	103	X_{41}	157	X_{51}	275
X_2	15	X_{12}	42	X_{22}	77	X_{32}	107	X_{42}	165	X_{52}	304
X_3	20	X_{13}	42	X_{23}	77	X_{33}	121	X_{43}	171	X_{53}	316
X_4	20	X_{14}	55	X_{24}	81	X_{34}	123	X_{44}	175		
X_5	25	X_{15}	57	X_{25}	83	X_{35}	125	X_{45}	188		
X_6	33	X_{16}	58	X_{26}	83	X_{36}	136	X_{46}	209		
X_7	35	X_{17}	58	X_{27}	91	X_{37}	140	X_{47}	213		
X_8	37	X_{18}	60	X_{28}	97	X_{38}	148	X_{48}	217		
X_9	42	X_{19}	62	X_{29}	101	X_{39}	152	X_{49}	248		
X_{10}	42	X_{20}	64	X_{30}	102	X_{40}	152	X_{50}	275		
										Σx =	5901

Median = 91; Mean = 111.8

Arithmetic mean

$\sum_{i=1}^{N} x / N$ = Sample mean = $\bar{X}$ (pronounced "Xbar")

$\sum_{i=1}^{N} x / N$ = Population mean = μ (pronounced "mew")

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
X_1	5	X_{11}	42	X_{21}	75	X_{31}	103	X_{41}	157	X_{51}	275
X_2	15	X_{12}	42	X_{22}	77	X_{32}	107	X_{42}	165	X_{52}	304
X_3	20	X_{13}	42	X_{23}	77	X_{33}	121	X_{43}	171	**X_{53}**	**1,000,000**
X_4	20	X_{14}	55	X_{24}	81	X_{34}	123	X_{44}	175		
X_5	25	X_{15}	57	X_{25}	83	X_{35}	125	X_{45}	188		
X_6	33	X_{16}	58	X_{26}	83	X_{36}	136	X_{46}	209		
X_7	35	X_{17}	58	X_{27}	91	X_{37}	140	X_{47}	213		
X_8	37	X_{18}	60	X_{28}	97	X_{38}	148	X_{48}	217		
X_9	42	X_{19}	62	X_{29}	101	X_{39}	152	X_{49}	248		
X_{10}	42	X_{20}	64	X_{30}	102	X_{40}	152	X_{50}	275		
										Σx =	1005585

TYPO!

Median = ; Mean =

Arithmetic mean

$\sum_{i=1}^{N} x / N$ = Sample mean = $\bar{X}$ (pronounced "Xbar")

$\sum_{i=1}^{N} x / N$ = Population mean = μ (pronounced "mew")

Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number	Ss ID	number
X_1	5	X_{11}	47	X_{21}	75	X_{31}	103	X_{41}	157	X_{51}	275
X_2	15	X_{12}	50	X_{22}	77	X_{32}	107	X_{42}	165	X_{52}	304
X_3	20	X_{13}	52	X_{23}	77	X_{33}	121	X_{43}	171	**X_{53}**	**1,000,000**
X_4	20	X_{14}	55	X_{24}	81	X_{34}	123	X_{44}	175		
X_5	25	X_{15}	57	X_{25}	83	X_{35}	125	X_{45}	188		
X_6	33	X_{16}	58	X_{26}	83	X_{36}	136	X_{46}	209		
X_7	35	X_{17}	58	X_{27}	91	X_{37}	140	X_{47}	213		
X_8	37	X_{18}	60	X_{28}	97	X_{38}	148	X_{48}	217		
X_9	41	X_{19}	62	X_{29}	101	X_{39}	152	X_{49}	248		
X_{10}	42	X_{20}	64	X_{30}	102	X_{40}	152	X_{50}	275		
										Σx =	1005607

Median = 91; **Mean = 18,973.7**

Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.

	X_i	X_i+2	X_i+100	X_i-2	X_i-100
X_1	1	3	101	-1	-99
X_2	2	4	102	0	-98
X_3	3	5	103	1	-97
Sum=>	6	12	306	0	-294
Mean=>	2	4	102	0	-98

Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.
- If you multiply/divide each score by a constant you change the mean by multiplying/dividing it by the constant.

	X_i	X_i*2	X_i*100	$X_i/2$	$X_i/5$
X_1	1	2	100	0.5	0.2
X_2	2	4	200	1	0.4
X_3	3	6	300	1.5	0.6
Sum=>	6	12	600	3	1.2
Mean=>	2	4	200	1	0.4

Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.
- If you multiply/divide each score by a constant you change the mean by multiplying/dividing it by the constant.
- Summed deviations from the mean = 0, or $\Sigma(x_i\text{-}x) = 0$

	X_i	$X_i\text{-}\bar{X}$
X_1	1	-1
X_2	2	0
X_3	3	1
Sum=>	6	0
Mean=>	2	0

Neat things about the mean

- Sum of squared deviations from the mean (SS) is minimized.
  $\Sigma(x\text{-}\bar{x})^2$ = minimum
  $\Sigma x^2 - (\Sigma x)^2/N$ = minimum

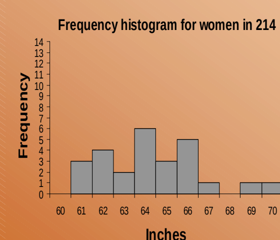
	X_i	$X_i\text{-}\bar{X}$	$(X_i\text{-}\bar{X})^2$	$(X_i\text{-}0)^2$	$(X_i\text{-}3)^2$
X_1	1	-1	1	1	4
X_2	2	0	0	4	1
X_3	3	1	1	9	0
Sum=>	6	0	2	14	5
Mean=>	2	0			

Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.
- If you multiply/divide each score by a constant you change the mean by multiplying/dividing it by the constant.
- Summed deviations from the mean = 0, or $\Sigma(x_i\text{-}\bar{x}) = 0$
- Very sensitive to extreme scores (outliers).
- Sum of squared deviations from the mean (SS) is minimized.
  $\Sigma(x_i\text{-}\bar{x})^2$ = minimum
  $\Sigma x^2 - (\Sigma x)^2/N$ = minimum

Height of 26 women in inches (real data)

Number	Gender	Height (X)
1	1	61
2	1	61
3	1	61
4	1	62
5	1	62
6	1	62
7	1	62
8	1	63
9	1	63
10	1	64
11	1	64
12	1	64
13	1	64
14	1	64
15	1	64
16	1	65
17	1	65
18	1	65
19	1	66
20	1	66
21	1	66
22	1	66
23	1	66
24	1	67
25	1	69
26	1	70
	Sum ==>	1672



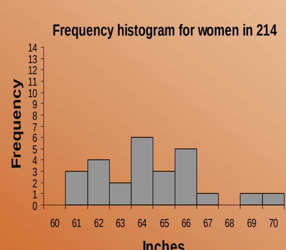
Mode =

Median =

Mean =

Height of women in inches (real data)

Number	Gender	Height (X)
1	1	61
2	1	61
3	1	61
4	1	62
5	1	62
6	1	62
7	1	62
8	1	63
9	1	63
10	1	64
11	1	64
12	1	64
13	1	64
14	1	64
15	1	64
16	1	65
17	1	65
18	1	65
19	1	66
20	1	66
21	1	66
22	1	66
23	1	66
24	1	67
25	1	69
26	1	70
	Sum ==>	1672



Mode = 64

Median = 64

Mean = 64.3

Height of women in inches (real data)

Number	Gender	Height (X)	(X+10)	(X-14)	(X*2)	(X/3)
1	1	61	71.0	47	122	20.3
2	1	61	71.0	47	122	20.3
3	1	61	71.0	47	122	20.3
4	1	62	72.0	48	124	20.7
5	1	62	72.0	48	124	20.7
6	1	62	72.0	48	124	20.7
7	1	62	72.0	48	124	20.7
8	1	63	73.0	49	126	21.0
9	1	63	73.0	49	126	21.0
10	1	64	74.0	50	128	21.3
11	1	64	74.0	50	128	21.3
12	1	64	74.0	50	128	21.3
13	1	64	74.0	50	128	21.3
14	1	64	74.0	50	128	21.3
15	1	64	74.0	50	128	21.3
16	1	65	75.0	51	130	21.7
17	1	65	75.0	51	130	21.7
18	1	65	75.0	51	130	21.7
19	1	66	76.0	52	132	22.0
20	1	66	76.0	52	132	22.0
21	1	66	76.0	52	132	22.0
22	1	66	76.0	52	132	22.0
23	1	66	76.0	52	132	22.0
24	1	67	77.0	53	134	22.3
25	1	69	79.0	55	138	23.0
26	1	70	80.0	56	140	23.3
	Sum ==>	1672	1932	1308	3344	557.3
	Mean ==>	64.3	74.3	50.3	128.6	21.4

64.3 - 14 = 50.3

64.3 * 2 = 128.6

64.3/3 = 21.4

64.3+10 = 74.3

Height of women in inches (real data)

Number	Gender	Height (X)	$(X-\bar{X})$
1	1	61	-3.3
2	1	61	-3.3
3	1	61	-3.3
4	1	62	-2.3
5	1	62	-2.3
6	1	62	-2.3
7	1	62	-2.3
8	1	63	-1.3
9	1	63	-1.3
10	1	64	-0.3
11	1	64	-0.3
12	1	64	-0.3
13	1	64	-0.3
14	1	64	-0.3
15	1	64	-0.3
16	1	65	0.7
17	1	65	0.7
18	1	65	0.7
19	1	66	1.7
20	1	66	1.7
21	1	66	1.7
22	1	66	1.7
23	1	66	1.7
24	1	67	2.7
25	1	69	4.7
26	1	70	5.7
	Sum ==>	1672	0.0
	Mean ==>	64.3	

$\Sigma(x-\bar{x}) = 0!$

Conceptual formula vs. computational formula for SS

- Conceptual SS:
  SS: $\Sigma(x-\bar{x})^2$

  "Sum of each X's squared difference from the mean"

Computational
$\Sigma x^2 - (\Sigma x)^2/N$

"Sum of the Xs after they have been squared minus the sum of all the Xs, which is then squared and then divided by the total number of Xs"

	X_i	$(X_i-\bar{X})^2$
X_1	1	1
X_2	2	0
X_3	3	1
Sum=>	6	2
Mean=>	2	

$SS = \Sigma x^2 - (\Sigma x)^2/N$
$= (1^2 + 2^2 + 3^2) - [(1+2+3)^2/3]$
$= (1+4+9) - (6^2/3)$
$= 14 - 36/3$
$= 14 - 12$
$= 2$

Questions about measures of central tendency?

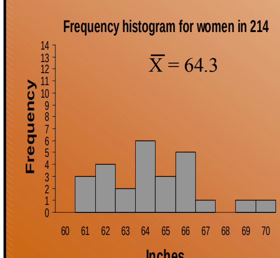
- Why is the mean our preferred measure of central tendency?
  - Adjusted for the number of scores.
  - Takes into account the numerical "weight" of each score.
    As scores of greater magnitude are added, the mean increases
    As scores of lesser magnitude are added, the mean decreases
  - Sum of squared deviations from the mean (SS) is minimized.
    SS is the square of the sum of each score's difference from the mean.
    $\Sigma(x-\bar{x})^2$
    $\Sigma x^2 - (\Sigma x)^2/N$ = minimum

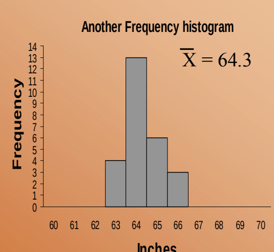
Variability

- Not only interested in a distribution's middle.
- Also interested in its spread (or variability).
- Define distributions by:
  - Central tendency
  - Variability
- How can we describe variability with a single number?

Variability

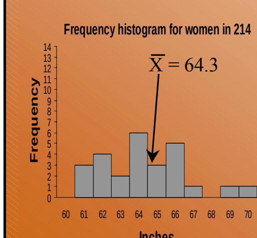
- How do these distributions differ?

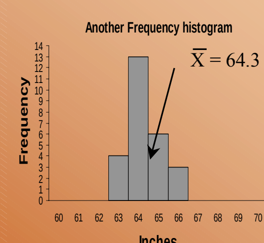




One is more spread out (greater variability) than another.

- Spread out around what?





Describe variability around the mean with one number.

- Want to adjust for the number of scores.
- Take into account the numerical "weight" of each score.
  - As scores are farther from the mean, the index of variability should increase
  - As scores are closer to the mean, the index of variability should decrease
- Suppose we measured each score's distance from its mean, and then used the average distance as our measure?
  - Using the <u>average</u> distance will adjust for the number of scores.
  - Measuring the distance from the mean should tell us how spread out each score is relative to the mean.

Try measure of variability with some simple numbers

- X (a population) = {2,4,6}
- $\mu = 4$
- What is the average distance from the mean?
  - How far is X_1 away from the mean (2 - 4 = ??)
  - How far is X_2 away from the mean (4 - 4) = ??)
  - How far is X_3 away from the mean? (6 - 4) = ??)

- What is the sum of the distance from the mean? [$\Sigma(x- \mu)$ = ??]
- How can we use the distance from the mean as a measure of variability?

Average of the <u>squared</u> distances from the mean!

- Find the distance (deviation) of each score from its mean (x-μ).
  - -2, 0, 2
  - Why? Measure how spread out each score is from the mean.
- Square the deviation of each score from its mean $(x- \mu)^2$
  - $-2^2 = 4$, $0^2 = 0$, $2^2 = 4$
  - Why? So the values won't always sum to zero.
- Sum the squared deviations: $\Sigma(x- \mu)^2$ (or SS)
  - 4 + 0 + 4 = 8
  - Question: can SS be negative?
- Divide by N
  - 8/3 = 2.7
  - Why? To get the <u>average</u> squared deviation from the mean.
  - Congratulations, you've just calculated the population variance, σ^2

Is 2.7 the average distance each score is from its mean?

- X = {2,4,6}
- In absolute terms:
  - X_1 is 2 away from the mean
  - X_2 is 0 away from the mean
  - X_3 is 2 away from the mean
- Shouldn't average distance be about 4/3 or 1.33?
- Why is the variance (σ^2) as a measure of the average distance of each score from its mean so much bigger than our intuition (that is, why is the $\sigma^2 = 2.7$ when the average distance from the mean is obviously closer to 1.3?

Is 2.7 the average distance each score is from its mean?

- X = {2,4,6}
- In absolute terms:
  - X_1 is 2 away from the mean
  - X_2 is 0 away from the mean
  - X_3 is 2 away from the mean
- Shouldn't average distance be about 4/3 or 1.33?
- Why is the variance (σ^2) as a measure of the average distance of each score from its mean so much bigger than our intuition?
- <u>**BECAUSE WE SQUARED ALL THE DEVIATIONS!**</u>
- How can we "unsquare" our answer?

How do we "unsquare" the variance?

- Unsquare the variance (σ^2) by taking the square root of it:

$$\sqrt{\sigma^2} = |\sigma| = \sqrt{\Sigma(x-\mu)^2/N} = \text{standard deviation}$$

Why? To get back to the original scale of X.

$\sqrt{2.7} = 1.63$, much closer to our intuitively derived 1.3

Generally

- Population variance (sigma squared, or σ^2) is the average of the squared deviations from the mean:

$$\frac{\Sigma(x-\mu)^2}{N} \qquad \text{Note: also written SS/N or } \frac{\Sigma x^2 - (\Sigma x)^2/N}{N}$$

- Population standard deviation (sigma, or σ) is the square root of the average of the squared deviations from the mean:

$$\sqrt{\frac{\Sigma(x-\mu)^2}{N}} \qquad \text{Note: also written } \sqrt{SS/N} \text{ or } \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/N}{N}}$$

Generally

- Sample variance (or s^2) is the (corrected) average of the squared deviations from the mean:

$$\frac{\Sigma(x-\bar{x})^2}{N-1} \qquad \text{Note: also written SS/N-1or } \frac{\Sigma x^2 - (\Sigma x)^2/N}{N-1}$$

- Sample standard deviation (or s) is the square root of the (corrected) average of the squared deviations from the mean:

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}} \qquad \text{Note: also written } \sqrt{SS/N\text{-}1} \text{or } \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/N}{N-1}}$$

Questions?

- We know what happens to the mean when we add or subtract a constant to/from all the scores, but what happens to the variance and standard deviation?
- We know what happens to the mean when we multiply or divide all the scores by a constant, but what happens to the variance and standard deviation?

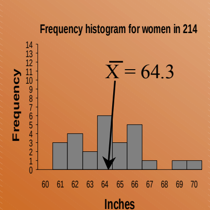
Questions?

- We know what happens to the mean when we add/subtract a constant to/from all the scores, but what happens to s^2 and s?
- We know what happens to the mean when we multiply or divide all the scores by a constant, but what happens to s^2 and s?

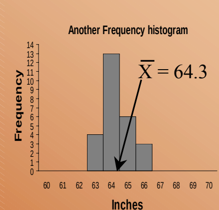
	X	X+2	X-2	X*2	X/2
X_1	1	3	-1	2	0.5
X_2	2	4	0	4	1.0
X_3	3	5	1	6	1.5
Mean=>	2	4	0	4	1

Remember these?

More variability



Less Variability



Two data sets

More variability

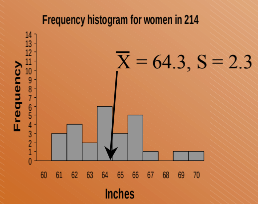
Number	Gender	Height (X)	$(X-\bar{X})$	$(X-\bar{X})^2$	
1	1	61	-3.3	10.9	
2	1	61	-3.3	10.9	
3	1	61	-3.3	10.9	
4	1	62	-2.3	5.3	
5	1	62	-2.3	5.3	
6	1	62	-2.3	5.3	
7	1	62	-2.3	5.3	
8	1	63	-1.3	1.7	
9	1	63	-1.3	1.7	
10	1	64	-0.3	0.1	
11	1	64	-0.3	0.1	
12	1	64	-0.3	0.1	
13	1	64	-0.3	0.1	
14	1	64	-0.3	0.1	
15	1	64	-0.3	0.1	
16	1	65	0.7	0.5	
17	1	65	0.7	0.5	
18	1	65	0.7	0.5	
19	1	66	1.7	2.9	
20	1	66	1.7	2.9	
21	1	66	1.7	2.9	
22	1	66	1.7	2.9	
23	1	66	1.7	2.9	
24	1	67	2.7	7.2	
25	1	69	4.7	22.0	
26	1	70	5.7	32.4	
	Sum ==>	1672	0.0	135.5	
	Mean ==	64.3		5.2	<==σ^2
				2.3	<==σ
				5.4	<==s^2
				2.3	<==s

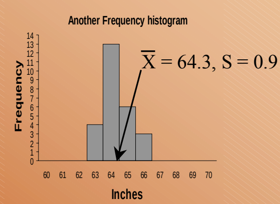
Less Variability

Number	Gender	Height (X)	$(X-\bar{X})$	$(X-\bar{X})^2$	
1	1	63	-1.3	1.7	
2	1	63	-1.3	1.7	
3	1	63	-1.3	1.7	
4	1	63	-1.3	1.7	
5	1	64	-0.3	0.1	
6	1	64	-0.3	0.1	
7	1	64	-0.3	0.1	
8	1	64	-0.3	0.1	
9	1	64	-0.3	0.1	
10	1	64	-0.3	0.1	
11	1	64	-0.3	0.1	
12	1	64	-0.3	0.1	
13	1	64	-0.3	0.1	
14	1	64	-0.3	0.1	
15	1	64	-0.3	0.1	
16	1	64	-0.3	0.1	
17	1	64	-0.3	0.1	
18	1	65	0.7	0.5	
19	1	65	0.7	0.5	
20	1	65	0.7	0.5	
21	1	65	0.7	0.5	
22	1	65	0.7	0.5	
23	1	65	0.7	0.5	
24	1	66	1.7	2.9	
25	1	66	1.7	2.9	
26	1	66	1.7	2.9	
	Sum ==>	1672	0.0	19.5	
	Mean ==	64.3		0.8	<==σ^2
				0.9	<==σ
				0.8	<==s^2
				0.9	<==s

Remember these?

More variability | Less Variability





What have we learned?

- Mean is preferred measure of central tendency:

  $\sum_{i=1}^{N} x_i/N$ = Sample mean = $\bar{X}$, more commonly $\Sigma x/N$

- Standard deviation is preferred measure of variability:

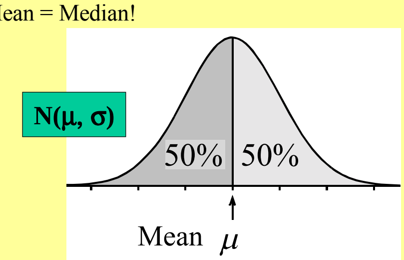
  $\sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}}$ = Sample standard deviation (s), can also be written $\sqrt{SS/N-1}$

Z–scores, Normal standardization

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- **Normal z-scores, standardized distribution**
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

The Normal distribution density curve

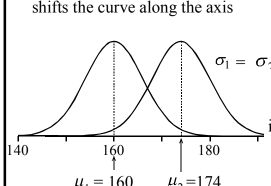
- Is symmetric about the mean! Bell-shaped and unimodal.
- Mean = Median!



Effects of μ and σ

(a) Changing μ

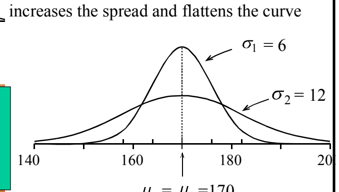
shifts the curve along the axis



Mean is a measure of …
central tendency

(b) Increasing σ

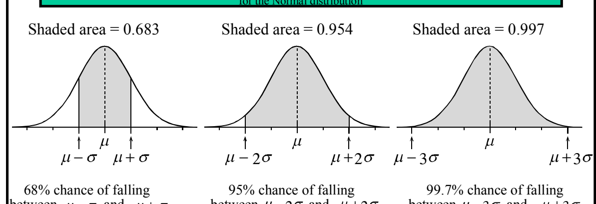
increases the spread and flattens the curve



Standard deviation is a measure of …
variability/spread

Understanding the standard deviation: σ

Probabilities/areas and numbers of standard deviations
for the Normal distribution



Basic method for obtaining probabilities

- Sketch a Normal curve, marking the mean and other values of interest.
- Shade the area under the curve that gives the desired probability.
- Devise a way of getting the desired area from lower-tail areas.
- Obtain component lower-tail probabilities from a computer program

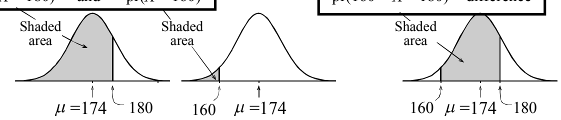
(a) Computing $\text{pr}(160 < X \leq 180)$

Programs supply

$\text{pr}(X \leq 180)$ and $\text{pr}(X \leq 160)$

We want

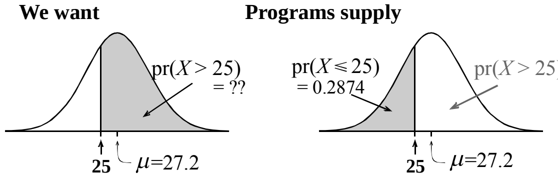
$\text{pr}(160 < X \leq 180)$ = difference



$\text{pr}(160 < X \leq 180) = \text{pr}(X \leq 180) - \text{pr}(X \leq 160)$

Obtaining an upper-tail probability

We want **Programs supply**



Since total area under curve = 1, $\text{pr}(X > 25) = 1 - \text{pr}(X \leq 25)$

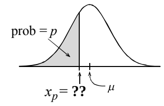
Generally, $\textbf{pr}(X > x) = 1 - \textbf{pr}(X \leq x)$

The inverse problem – Percentiles/quantiles

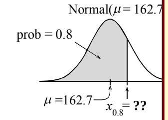
(a) *p*-Quantile

Programs supply x_p

x-value for which $\text{pr}(X \leq x_p) = p$



(b) 80th percentile (0.8-quantile) **of women's heights**



Program returns
Thus 80% lie belo

80% of people have height below the 80th percentile. This is EQ to saying there's 80% chance that a random observation from the distribution will fall below the 80th percentile.

(c) Further percentiles of women's heights

Percent	1%	5%	10%	20%	30%	70%	80%	90%	95%
Propn	0.01	0.05	0.1	0.2	0.3	0.7	0.8	0.9	0.95
Percentile	*(or quantile)*								
(cm)	148.3	152.5	154.8	157.5	159.4	166.0	167.9	170.6	172.9
(ft'in")	4'10"	5'0"	5'0"	5'2"	5'2"	5'5"	5'6"	5'7"	5'8"

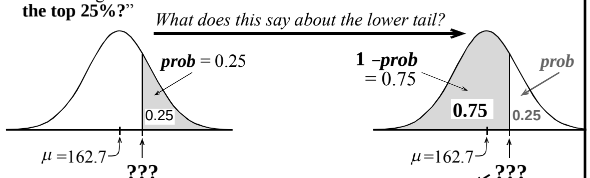
The inverse problem is what is the height for the 80th percentile/quantile? So far we studied given the height value what's the corresponding percentile?

The inverse problem – *upper-tail* percentiles/quantiles

Obtaining an inverse upper-tail probability

"What value gives **the top 25%?**"

What does this say about the lower tail?



Obtain from program

[Program returns 166.88]

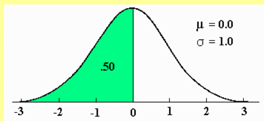
Review

- What is meant by the 60th percentile of heights?
- What is the difference between a percentile and a quantile? (percentile used in expressing results in %, whereas quantiles used to express results in term of probabilities)
- The lower quartile, median and upper quartile of a distribution correspond to special percentiles. What are they? express in terms of quantiles. (25%, 50%, 75%)
- Quantiles are sometimes called inverse cumulative probabilities. Why?

Standard Normal Curve

- The standard normal curve is described by the equation:

$$y = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

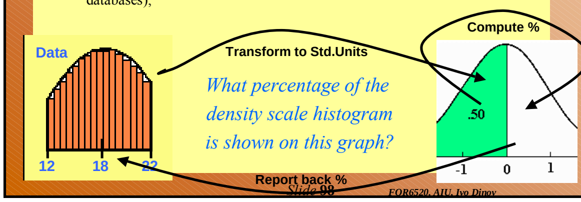


Where remember, the natural number e ~ 2.7182…

We say: X~Normal(μ, σ), or simply X~N(μ, σ)

Standard Normal Approximation

- The standard normal curve can be used to estimate the percentage of entries in an interval for any process. Here is the protocol for this approximation:
  - Convert the interval (we need the assess the percentage of entries in) to standard units. We saw the algorithm already.
  - Find the corresponding area under the normal curve (from tables or online databases);

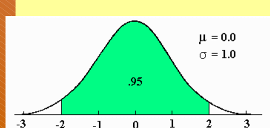


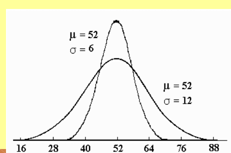
General Normal Curve

- The general normal curve is defined by:
  - Where μ is the average of (the symmetric) normal curve, and σ is the standard deviation (spread of the distribution).

$$y = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

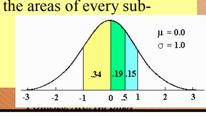
  - Why worry about a standard and general normal curves?
  - How to convert between the two curves?





Areas under Standard Normal Curve – Normal Approximation

- Protocol:
  - Convert the interval (we need to assess the percentage of entries in) to Standard units. Actually convert the end points in Standard units.
    - In general, the transformation X → (X-μ)/σ, standardizes the observed value X, where μ and σ are the average and the standard deviation of the distribution X is drawn from.
  - Find the corresponding area under the normal curve (from tables or online databases);
    - Sketch the normal curve and shade the area of interest
    - Separate your area into individually computable sections
    - Check the Normal Table and extract the areas of every sub-section
    - Add/compute the areas of all sub-sections to get the total area.



The z-score

- The ***z*-score** of *x* is the number of standard deviations *x* is from the mean. (Body-Mass-Index, BMI)

TABLE 6.3.1 Examples of z -Scores

x	*z* -score = (*x* - *μ*)/*σ*		Interpretation
Male BMI values (kg/m^2)			
25	(25-27.3)/4.1 =	-0.56	25 kg/m^2 is 0.56 sd's below the mean
35	(35-27.3)/4.1 =	1.88	35 kg/m^2 is 1.88 sd's above the mean
Female heights (cm)			
155	(155-162.7)/6.2 =	-1.24	155cm is 1.24 sd's below the mean
180	(180-162.7)/6.2 =	2.79	180cm is 2.79 sd's above the mean

Male BMI-values: μ=27.3, σ=4.1 Females heights: μ=162.7, σ=6.2

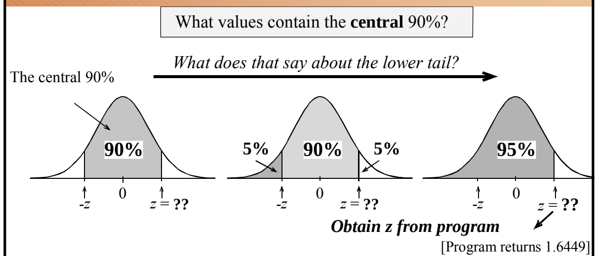
- Which ones of these are unusually large/small/away from the mean?

The standard Normal distribution

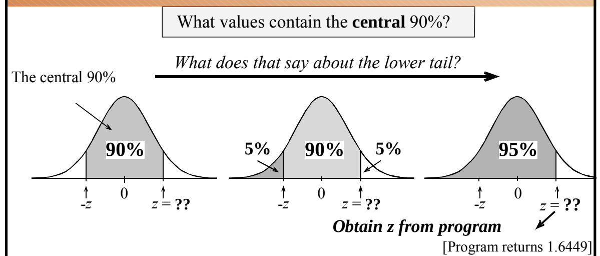
Standard Normal distribution:

mean(μ) = 0, SD(σ)= 1

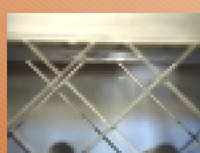
Working in standard units



Review



Summary



QuincunxApplet.html

Show Sampling Distribution Simulation Applet
file:///C:/Ivo.dir/UCLA_Classes/Winter2002/AdditionalInstructorAids/
SamplingDistributionApplet.html

Continuous Variables and Density Curves

- There are no gaps between the values a continuous random variable can take.
- Random observations arise in two main ways: (i) by sampling populations; and (ii) by observing processes.

The density curve

- The probability distribution of a continuous variable is represented by a density curve.
  - ***Probabilities*** are represented by ***areas under the curve***,
    - the probability that a random observation falls between *a* and *b* equal to the area under the density curve between *a* and *b*.
  - The total area under the curve equals 1.
  - The population (or distribution) mean $\mu_X = E(X)$, is where the density curve balances.
  - When we calculate probabilities for a continuous random variable, it does not matter whether interval endpoints are included or excluded.

For any random variable *X*

- $E(aX+b) = a\,E(X) + b$ and $SD(aX+b) = |a|\,SD(X)$

The Normal distribution

$X \sim \text{Normal}(\mu_x = \mu, \sigma_x = \sigma)$

Features of the Normal density curve:

- The curve is a symmetric bell-shape centered at μ.
- The standard deviation σ governs the spread.
  - 68.3% of the probability lies within 1 standard deviation of the mean
  - 95.4% within 2 standard deviations
  - 99.7% within 3 standard deviations

Probabilities

- Computer programs provide lower-tail (or cumulative) probabilities of the form pr($X \leq x$)
  - We give the program the *x*-value; it gives us the probability.
- Computer programs also provide inverse lower-tail probabilities (or quantiles)
  - We give the program the probability; it gives us the *x*-value.
- When calculating probabilities, we shade the desired area under the curve and then devise a way of obtaining it via lower-tail probabilities.

Standard Units

The *z-score* of a value *a* is ….

- the number of standard deviations *a* is away from the mean
- positive if *a* is above the mean and negative if *a* is below the mean.

The ***standard Normal*** distribution has $\mu = 0$ and $\sigma = 0$.

- We usually use *Z* to represent a random variable with a standard Normal distribution.

Ranges, extremes and *z*-scores

Central ranges:

- $P(-z \leq Z \leq z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls within *z* SD's either side of the mean.

Extremes:

- $P(Z \geq z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls more than *z* standard deviations above the mean.
- $P(Z \leq -z)$ is the same as the probability that a random observation from an arbitrary Normal distribution falls more than *z* standard deviations below the mean.

Combining Random Quantities

Variation and independence:

- No two animals, organisms, natural or man-made objects are ever identical.
- There is always variation. The only question is whether it is large enough to have a practical impact on what you are trying to achieve.
- Variation in component parts leads to even greater variation in the whole.

Independence

We model variables as being independent ….

- if we think they relate to physically independent processes
- and if we have no data that suggests they are related.

Both sums and differences of independent random variables are more variable than any of the component random variables

Formulas

- For a constant number a, $E(aX) = aE(X)$ and $SD(aX) = |a|\ SD(X)$.
- Means of sums and differences of random variables act in an obvious way
  - the mean of the sum is the sum of the means
  - the mean of the difference is the difference in the means
- For independent random variables, (cf. Pythagorean theorem),

$$SD(X_1 + X_2) = SD(X_1 - X_2) = \sqrt{SD(X_1)^2 + SD(X_2)^2}$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

[ASIDE: Sums and differences of independent Normally distributed random variables are also Normally distributed]

Example

- For constant numbers a & b, $E(aX+b) = aE(X)+b$ and $SD(aX+b) = |a|\ SD(X)$.
- For independent random variables

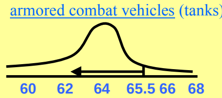
$$SD(X_1 + X_2) = SD(X_1 - X_2) = \sqrt{SD(X_1)^2 + SD(X_2)^2}$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

- For Dependent variables: Ex. E(X)=1, SD(X)=3
  - Y = 2X-1 ➔ E(Y)=1 and SD(Y) = 6
  - SD(X+Y)=SD(3X-1) = 9, NOT
  - 7== SD(X+Y) = Sqrt(SD^2 (X)+SD^2(Y)).
  - Defense vs. prosecution argument may be different for an X+Y value of 18, say.

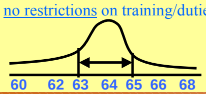
Areas under Standard Normal Curve – Normal Approximation, Scottish Army Recruits

- The mean height is 64 *in* and the standard deviation is 2 *in*.
  - Only recruits shorter than 65.5 in will be trained for tank operation. What percentage of the incoming recruits will be trained to operate armored combat vehicles (tanks)?



X → (X-64)/2
65.5 → (65.5-64)/2 = ¾
Percentage is 77.34%

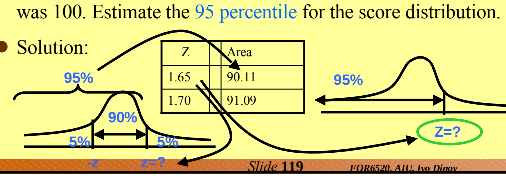
  - Recruits within ½ standard deviations of the mean will have no restrictions on duties. About what percentage of the recruits will have no restrictions on training/duties?



X → (X-64)/2
65 → (65-64)/2 = ½
63 → (63-64)/2 = -½
Percentage is 38.30%

Percentiles for Standard Normal Curve

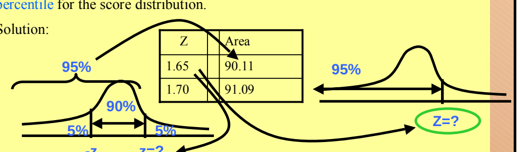
- When the histogram of the observed process follows the normal curve Normal Tables (of any type, as described before) may be used to estimate percentiles. The N-th percentile of a distribution is P is N% of the population observations are less than or equal to P.
- Example, suppose the Math-part SAT scores of newly admitted freshmen at UCLA averaged 535 (out of [200:800]) and the SD was 100. Estimate the 95 percentile for the score distribution.
- Solution:



Z	Area
1.65	90.11
1.70	91.09

Percentiles for Standard Normal Curve

- Example, suppose the Math-part SAT scores of newly admitted freshmen at UCLA averaged 535 (out of [200:800]) and the SD was 100. Estimate the 95 percentile for the score distribution.
- Solution:



Z	Area
1.65	90.11
1.70	91.09

- Z=1.65 (std. Units) → 700 (data units), since

  X → (X – μ)/σ, converts data to standard units and

  X → σ X + μ, converts standard to data units!

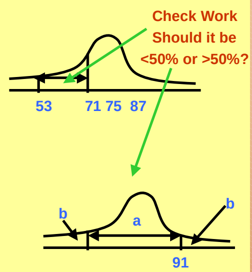
  σ = 100; μ =535, 100 x 1.65 + 535 = 700.

Summary

1. The Standard Normal curve is symmetric w.r.t. the origin (0,0) and the total area under the curve is 100% (1 unit)
2. Std units indicate how many SD's is a value below (-)/above (+) the mean
3. Many histograms have roughly the shape of the normal curve (bell-shape)
4. If a list of numbers follows the normal curve the percentage of entries falling within each interval is estimated by: 1. Converting the interval to StdUnits and, 2. Computing the corresponding area under the normal curve (Normal approximation)
5. A histogram which follows the normal curve may be reconstructed just from (μ,$σ^2$), mean and variance=std_dev^2
6. Any histogram can be summarized using percentiles
7. E(aX+b)=aE(X)+b, Var(aX+b)=a^2Var(X), where E(Y) the the mean of Y and Var(Y) is the square of the StdDev(Y),

Example – work out in your notebooks

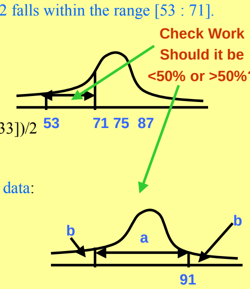
1. Compute the chance a random observation from a distribution (symmetric, bell-shaped, unimodal) with m=75 and SD=12 falls within the range [53 : 71].



Slide 122 FOR6520, AIU, Ivo Dinov

Example – work out in your notebooks

1. Compute the chance a random observation from a distribution (symmetric, bell-shaped, unimodal) with m=75 and SD=12 falls within the range [53 : 71].
2. (53-75)/12 = -11/6=-1.83 Std unit
3. (71-75)/12=-0.333(3) Std units
4. Area[53:71] =
5. (SN_area[-1.83:1.83] –SN_area[-0.33:0.33])/2
6. = (93% - 25%)/2 = 34%
7. Compute the 90[th] percentile for the same data:
8. b+a+b=100% a=80% → A=0.8
9. a+b=90% b=10% Z=1.3 SU
10. 90% P = σ1.3 + μ =12x1.3+75=90.6



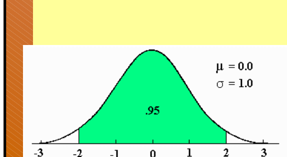
Slide 123 FOR6520, AIU, Ivo Dinov

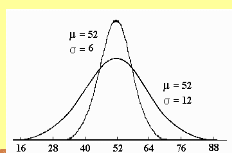
General Normal Curve

- The general normal curve is defined by:
  - Where μ is the average of (the symmetric) normal curve, and σ is the standard deviation (spread of the distribution).

$$y = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

  - Why worry about a standard and general normal curves?
  - How to convert between the two curves?

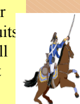




Slide 124 FOR6520, AIU, Ivo Dinov

Areas under Standard Normal Curve

- Many histograms are similar in shape to the standard normal curve. For example, persons height. The height of all incoming female army recruits is measured for custom training and assignment purposes (e.g., very tall people are inappropriate for constricted space positions, and very short people may be disadvantages in certain other situations). The mean height is computed to be 64 *in* and the standard deviation is 2 *in*. Only recruits shorter than 65.5 in will be trained for tank operation and recruits within ½ standard deviations of the mean will have no restrictions on duties.
  - What percentage of the incoming recruits will be trained to operate armored combat vehicles (tanks)?
  - About what percentage of the recruits will have no restrictions on training/duties?



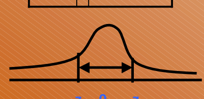
Slide 125 FOR6520, AIU, Ivo Dinov

Standard Normal Curve – Table differences

- There are different tables and computer packages for representing the area under the standard normal curve. But the results are always interchangeable.

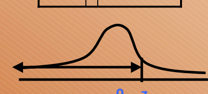
Area under Normal curve on [-z : z]

Z	Area
0.50	38.29
1.0	68.27



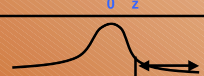
Area under Normal curve on [-infinity : z]

Z	Area
0.50	69.15
1.0	84.13



Area under Normal curve on [z: infinity]

Z	Area
0.50	30.85
1.0	15.87



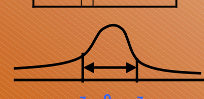
Slide 126 FOR6520, AIU, Ivo Dinov

Standard Normal Curve – Table differences

- There are different tables and computer packages for representing the area under the standard normal curve. But the results are always interchangeable.

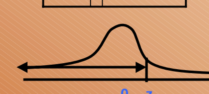
Area under Normal curve on [-z : z]

Z	Area
0.50	38.29
1.0	68.27



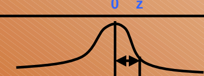
Area under Normal curve on [-infinity : z]

Z	Area
0.50	69.15
1.0	84.13



Area under Normal curve on [0 : z]

Z	Area
0.50	30.85
1.0	15.87



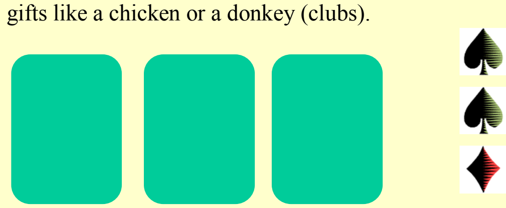
Slide 127 FOR6520, AIU, Ivo Dinov

Probability, Samples & Sampling error

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- **Probability, Samples & Sampling error**
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

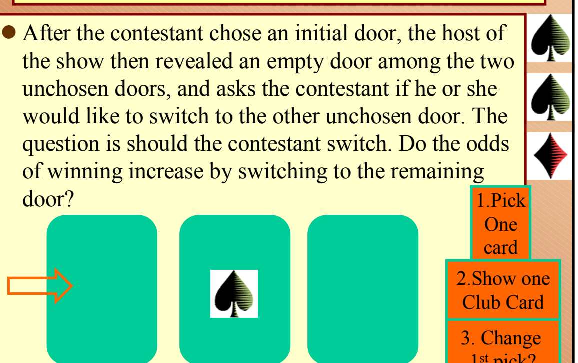
Let's Make a Deal Paradox – aka, Monty Hall 3-door problem

- This paradox is related to a popular television show in the 1970's. In the show, a contestant was given a choice of three doors/cards of which one contained a prize (diamond). The other two doors contained gag gifts like a chicken or a donkey (clubs).



Let's Make a Deal Paradox.

- After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?



Let's Make a Deal Paradox.

- The *intuition* of most people tells them that each of the doors, the chosen door and the unchosen door, are equally likely to contain the prize so that there is a 50-50 chance of winning with either selection? This, however, is **not the case**.
- The probability of winning by using the switching technique is 2/3, while the odds of winning by not switching is 1/3. The easiest way to explain this is as follows:

Let's Make a Deal Paradox.

- The probability of picking the wrong door in the initial stage of the game is 2/3.
- If the contestant picks the wrong door initially, the host must reveal the remaining empty door in the second stage of the game. Thus, if the contestant switches after picking the wrong door initially, the contestant will win the prize.
- The probability of winning by switching then reduces to the probability of picking the wrong door in the initial stage which is clearly 2/3.

Let's Make a Deal Paradox.

- Demo: Applets.dir/StatGames.exe
  - Uncertainty→Pick a door

Properties of probability distributions

- A sequence of number $\{p_1, p_2, p_3, \ldots, p_n\}$ is a probability distribution for a sample space $S = \{s_1, s_2, s_3, \ldots, s_n\}$, if $pr(s_k) = p_k$, for each $1<=k<=n$. The two essential properties of a probability distribution $p_1, p_2, \ldots, p_n$?

$$p_k \geq 0; \quad \sum_k p_k = 1$$

- How do we get the probability of an event from the probabilities of outcomes that make up that event?
- If all outcomes are distinct & equally likely, how do we calculate *pr(A)* ? If $A = \{a_1, a_2, a_3, \ldots, a_9\}$ and $pr(a_1)=pr(a_2)=\ldots=pr(a_9)=p$; then

$$pr(A) = 9 \times pr(a_1) = 9p.$$

Example of probability distributions

- Tossing a coin twice. *Sample space* S={HH, HT, TH, TT}, for a fair coin each outcome is equally likely, so the probabilities of the 4 possible outcomes should be identical, *p*. Since, p(HH)=p(HT)=p(TH)=p(TT)=p and

$$p_k \geq 0; \quad \sum_k p_k = 1$$

- p = ¼ = 0.25.

Conditional Probability

The ***conditional probability*** of ***A*** occurring ***given*** that *B* occurs is given by

$$\text{pr}(A \mid B) = \frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)}$$

Suppose we select one out of the 400 patients in the study and we want to find the probability that the cancer is on the *extremities given that* it is of type *nodular*: P = 73/125 = P(C. on Extremities | Nodular)

$$\frac{\#\text{nodular patients with cancer on extremities}}{\#\text{nodular patients}}$$

Melanoma – type of skin cancer – an example of laws of conditional probabilities

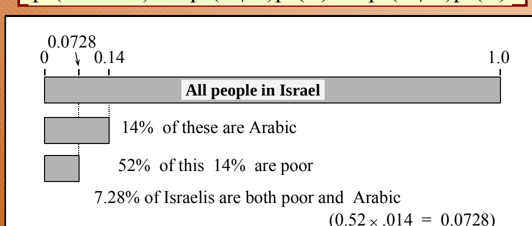
TABLE 4.6.1: 400 Melanoma Patients by Type and Site

Type	Site: Head and Neck	Site: Trunk	Site: Extremities	Row Totals
Hutchinson's melanomic freckle	22	2	10	34
Superficial	16	54	115	185
Nodular	19	33	73	125
Indeterminant	11	17	28	56
Column Totals	68	106	226	400

Contingency table based on Melanoma histological type and its location

Multiplication rule- what's the percentage of Israelis that are poor and Arabic?

$$\text{pr}(A \text{ and } B) = \text{pr}(A \mid B)\text{pr}(B) = \text{pr}(B \mid A)\text{pr}(A)$$



Let's Make a Deal Paradox.

- After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?
- P(Win (swap strat.) | 1st is Club) = 1
- P(Win (swap strat.) & 1st is Club) =
  = P(Win (swap strat.) | 1st is Club) x P(1st is Club)
  = 1 x 2/3 = 2/3.

Review

$\text{pr}(A \text{ and } B) = \text{pr}(A \mid B)\text{pr}(B) = \text{pr}(B \mid A)\text{pr}(A)$

$\text{pr}(A) = 1 - \text{pr}(\bar{A})$

1. Proportions (partial description of a real population) and probabilities (giving the chance of something happening in a random experiment) may be identical – under the experiment *choose-a-unit-at-random*
2. Properties of probabilities.

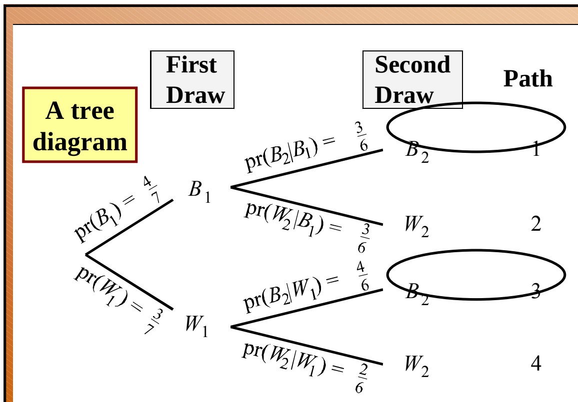
$\{p_k\}_{k=1}^{N}$ define probabilities $\Leftrightarrow p_k \geq 0; \ \sum_k p_k = 1$

A tree diagram for computing conditional probabilities

Suppose we draw 2 balls at random one at a time *without replacement* from an urn containing **4 black** and **3 white** balls, otherwise identical. What is the probability that the *second ball is black*? Sample Spc?

P({*2-nd ball is black*}) =
P({*2-nd is black*} &{*1-st is black*}) +
P({2-nd is black} &{*1-st is white*}) =
4/7 x 3/6 + 4/6 x 3/7 = 4/7.

Mutually exclusive



Conditional probabilities and 2-way tables

- Many problems involving conditional probabilities can be solved by constructing two-way tables
- This includes *reversing the order of conditioning*

P(A & B) = P(A | B) x P(B) = P(B | A) x P(A)

Conditional probabilities

- Example. Calculate the Probability of having Exactly One Black ball **given** that Ball-2 is black?

  (Recall: **4 black** and **3 white** balls given!)

- $P(B_2 \mid \text{Exactly One Black}) = P(\cap) / P(\text{exactly1B})$
  $= P(W1 \ \& \ B2) / (P(W_1\&B_2) + P(B_1\&W_2)) = ½$
- $P(\text{Exactly 1 Black} \mid B_2)$ =??? (Not trivial!)
  $= P(B_2 \mid \text{Exactly One Black}) \times P(B_2) \ / \ P(\text{Exactly One Black})$
  $= (½ \times 4/7) / P(\{B_1\&W_2\} \text{ or } \{W_1\&B_2\})$
  $= (4/14) / (4/7 \times 3/6 + 3/7 \times 4/6) =$ **1/2**

Proportional usage of *oral contraceptives* and their rates of failure

We need to complete the two-way contingency table of proportions

pr(*Failed* and *Oral*) = pr(*Failed* | *Oral*) × pr(*Oral*) [= 5% of 32%]

pr(*Failed* and *IUD*) = pr(*Failed* | *IUD*) × pr(*IUD*) [= 6% of 3%]

		Method					
		Steril.	Oral	Barrier	IUD	Sperm.	Total
Outcome	Failed	0 × .38	.05 × .32	.14 × .24	.06 × .03	.26 × .03	?
	Didn't	?	?	?	?	?	?
	Total	.38	.32	.24	.03	.03	1.00

pr(*Steril.*) = .38 pr(*Barrier*) = .24 pr(*IUD*) = .03

Oral contraceptives cont.

pr(*Failed* and *Oral*) = pr(*Failed* | *Oral*) × pr(*Oral*) [= 5% of 32%]

pr(*Failed* and *IUD*) = pr(*Failed* | *IUD*) × pr(*IUD*) [= 6% of 3%]

		Method					
		Steril.	Oral	Barrier	IUD	Sperm.	Total
Outcome	Failed	0 × .38	.05 × .32	.14 × .24	.06 × .03	.26 × .03	?
	Didn't	?	?	?	?	?	?
	Total	.38	.32	.24	.03	.03	1.00

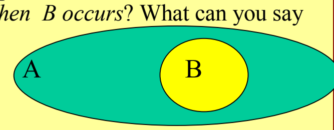
pr(*Steril.*) = .38 pr(*Barrier*) = .24 pr(*IUD*) = .03

TABLE 4.6.4 Table Constructed from the Data in Example 4.6.

		Method					
		Steril.	**Oral**	**Barrier**	**IUD**	**Sperm.**	**Total**
Outcome	Failed	0	.0160	.0336	.0018	.0078	.0592
	Didn't	.3800	.3040	.2064	.0282	.0222	.9408
	Total	.3800	.3200	.2400	.0300	.0300	1.0000

Remarks …

- In *pr(A | B)*, how should the symbol " | " is read *given that*.
- How do we interpret the fact that: *The event A always occurs when B occurs*? What can you say about pr(*A* | *B*)?



- When drawing a probability tree for a particular problem, how do you know *what events* to use for the first fan of branches and which events to use for the subsequent branching? (at each branching stage condition on all the info available up to here. E.g., at first branching use all simple events, no prior is available. At 3-rd branching condition of the previous 2 events, etc.).

Type I & Type II errors – Power of a test

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- **Type I and Type II errors; Power of a test**
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

TABLE 4.6.5 Number of Individuals Having a Given Mean Absorbance Ratio (MAR) in the ELISA for HIV Antibodies

MAR	Healthy Donor		HIV patients	
<2	202	} 275	0	} 2 False-Negatives (FNE)
2 - 2.99	73	Test cut-off	2	
3 - 3.99	15	False-positives	7	Power of a test is:
4 - 4.99	3		7	
5 - 5.99	2		15	1-P(FNE)=
6 -11.99	2		36	1-P(Neg|HIV)
12+	0		21	~ 0.976
Total	297		88	

Adapted from Weiss et al.[1985]

HIV cont.

pr(*HIV* and *Positive*) = pr(*Positive*|*HIV*) × pr(*HIV*) [= 98% of 1%]

pr(*Not HIV* and *Negative*) = pr(*Negative*|*Not HIV*) × pr(*Not HIV*) [= 93% of 99%]

		Test result		
		Positive	Negative	Total
Disease status	HIV	.98 × .01	?	.01 ⟵ pr(*HIV*) = .01
	Not HIV	?	.93 × .99	.99 ⟵ pr(*Not HIV*) = .99
	Total	?	?	1.00

Figure 4.6.6 Putting HIV information into the table.

HIV – reconstructing the contingency table

pr(*HIV* and *Positive*) = pr(*Positive*|*HIV*) × pr(*HIV*) [= 98% of 1%]

pr(*Not HIV* and *Negative*) = pr(*Negative*|*Not HIV*) × pr(*Not HIV*) [= 93% of 99%]

		Test result		
		Positive	Negative	Total
Disease status	HIV	.98 × .01	?	.01 ⟵ pr(*HIV*) = .01
	Not HIV	?	.93 × .99	.99 ⟵ pr(*Not HIV*) = .99
	Total	?	?	1.00

TABLE 4.6.6 Proportions by Disease Status and Test Result

		Test Result		
		Positive	**Negative**	**Total**
Disease Status	HIV	.0098	.0002	.01
	Not HIV	.0693	.9207	.99
	Total	.0791	.9209	1.00

Proportions of HIV infections by country

TABLE 4.6.7 Proportions Infected with HIV

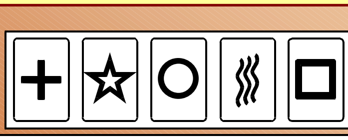
Country	No. AIDS Cases	Population (millions)	pr(HIV)	Having | Test pr(HIV | Positive)
United States	218,301	252.7	0.00864	0.109
Canada	6,116	26.7	0.00229	0.031
Australia	3,238	16.8	0.00193	0.026
New Zealand	323	3.4	0.00095	0.013
United Kingdom	5,451	57.3	0.00095	0.013
Ireland	142	3.6	0.00039	0.005

Hypothesis testing

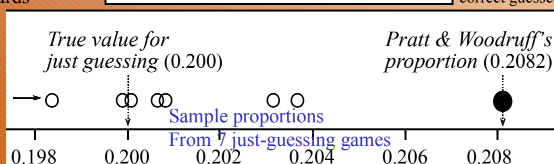
- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- **Intro to hypothesis testing**
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

ESP (extra sensory perception) or just guessing?

Deck of equal number of Zener/Rhine cards



n=60,000 random draws resulting in 12,489 correct guesses



Can sampling variations alone account for Pratt & Woodruff's success rate = 20.82% correct vs. 20% expected.

ESP or just guessing?

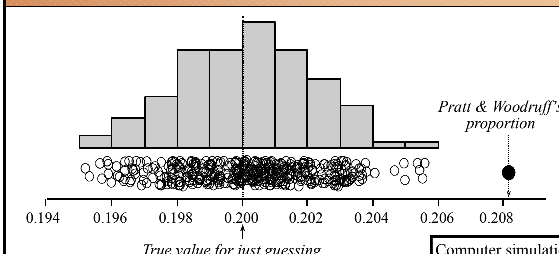


Figure 9.1.1 Sample proportions from 400 "just-guessing" experiments.

Computer simulation making 60,000 guesses with 20% chance of correct guess.

Was Cavendish's experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased.

Mean density of the earth

True value = 5.517

Cavendish's data: (from previous Example 7.2.2)

5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.10, 5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85

n = 23, sample mean = 5.483, sample SD = 0.1904

Was Cavendish's experiment biased?

Simulate taking 400 sets of 23 measurements from N(5.517,0.1904). Plotted are the results of the sample means.

Are the Cavendish values unusually diff. from true mean?

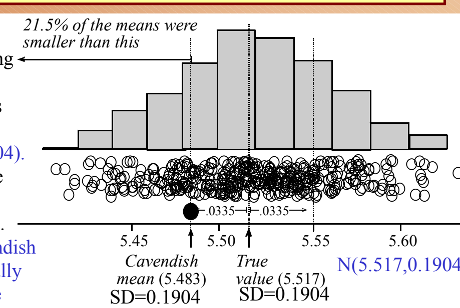
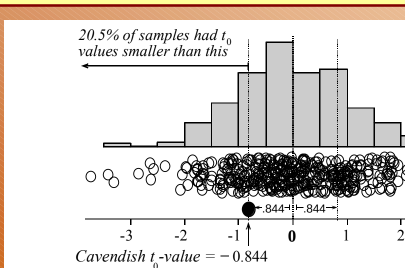


Figure 9.1.2 Sample means from 400 sets of observations from an unbiased experiment.

Cavendish: measuring distances in std errors



Cavendish data lies within the central 60% of the distribution

Figure 9.1.3 Sample t_0-values from 400 unbiased experiments (each t_0-value is distance between sample mean and 5.517 in std errors).

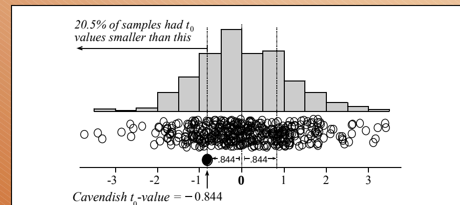


Figure 9.1.3 Sample t_0-values from 400 unbiased experiments (each t_0-value is distance between sample mean and 5.517 in std errors).

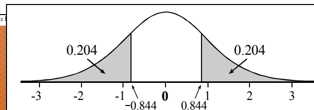


Figure 9.1.4 Student(df=22) density.

Measuring the distance between the true-value and the estimate in terms of the SE

- Intuitive criterion: Estimate is credible if it's not far away from its hypothesized true-value!
- But how far is far-away?
- Compute the distance in standard-terms:

$$T = \frac{\text{Estimator} - \text{TrueParameterValue}}{\text{SE}}$$

- Reason is that the distribution of ***T*** is known in some cases (Student's t, or N(0,1)). The estimator (obs-value) is typical/atypical if it is close to the center/tail of the distribution.

Hypotheses

Guiding principles

We cannot **rule in** a hypothesized value for a parameter, we *can only* determine whether there is evidence *to* **rule out** a hypothesized value.

The ***null hypothesis*** tested is typically a skeptical reaction to a ***research hypothesis***

Comments

- Why can't we (**rule-in**) prove that a hypothesized value of a parameter is exactly true? (Because when constructing estimates based on data, there's always sampling and may be non-sampling errors, which are normal, and will effect the resulting estimate. Even if we do 60,000 ESP tests, as we saw earlier, repeatedly we are likely to get estimates like 0.2 and 0.200001, and 0.199999, etc. – non of which may be exactly the theoretically correct, 0.2.)
- Why use the rule-out principle? (Since, we can't use the rule-in method, we try to find compelling evidence against the observed/data-constructed estimate – to reject it.)
- Why is the null hypothesis & significance testing typically used? (H_o: skeptical reaction to a research hypothesis; ST is used to check if differences or effects seen in the data can be explained simply in terms of sampling variation!)

Comments

- How can researchers try to demonstrate that effects or differences seen in their data are real? (Reject the hypothesis that there are no effects)
- How does the alternative hypothesis typically relate to a belief, hunch, or research hypothesis that initiates a study? (H_1=H_a: specifies the type of departure from the null-hypothesis, H_0 (skeptical reaction), which we are expecting (research hypothesis itself).
- In the Cavendish's mean Earth density data, null hypothesis was $H_0 : \mu = 5.517$. We suspected bias, but not bias in any specific direction, hence $H_a: \mu != 5.517$.

Comments

- In the ESP Pratt & Woodruff data, (skeptical reaction) null hypothesis was $H_0 : \mu = 0.2$ (pure-guessing). We suspected bias, toward success rate being higher than that, hence the (research hypothesis) $H_a: \mu > 0.2$.
- Other commonly encountered situations are:
  - $H_0 : \mu_1 - \mu_2 = 0$ → $H_a : \mu_1 - \mu_2 > 0$
  - $H_0 : \mu_{rest} - \mu_{activation} = 0$ → $H_a : \mu_{rest} - \mu_{activation} \,!= 0$

The t-test

Using $\hat{\theta}$ to test H_0: $\theta = \theta_0$ versus some alternative H_1.

STEP 1 Calculate the ***test statistic***,

$$t_0 = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} = \frac{\text{estimate - hypothesized value}}{\text{standard error}}$$

[This tells us how many standard errors the estimate is above the hypothesized value (t_0 positive) or below the hypothesized value (t_0 negative).]

STEP 2 Calculate the *P*-value using the following table.

STEP 3 Interpret the *P*-value in the context of the data.

The t-test

Alternative hypothesis	Evidence against H_0: $\theta > \theta_0$ provided by	*P*-value
$H_1: \theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0 (i.e., $\hat{\theta} - \theta_0$ too large)	$P = \text{pr}(T \geq t_0)$
$H_1: \theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0 (i.e., $\hat{\theta} - \theta_0$ too negative)	$P = \text{pr}(T \leq t_0)$
$H_1: \theta \neq \theta_0$	$\hat{\theta}$ too far from θ_0 (i.e., $\lvert\hat{\theta} - \theta_0\rvert$ too large)	$P = 2\,\text{pr}(T \geq \lvert t_0 \rvert)$

where $T \sim \text{Student}(df)$

Interpretation of the p-value

TABLE 9.3.2 Interpreting the Size of a *P*-Value

Approximate size of *P*-Value		Translation
> 0.12	(12%)	**No** evidence against H_0
0.10	(10%)	**Weak** evidence against H_0
0.05	(5%)	**Some** evidence against H_0
0.01	(1%)	**Strong** evidence against H_0
0.001	(0.1%)	**Very Strong** evidence against H_0

Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0

Pictorial representation of the T-test
H_0: θ^= θ_0
H_1: $\theta > \theta_0$

$$t_0 = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

$\hat{\theta}$-scale → ***t*-scale** **(# of std errors)**

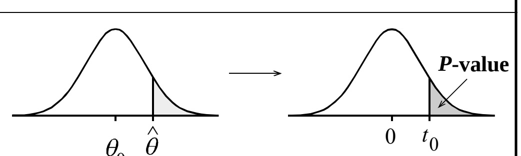
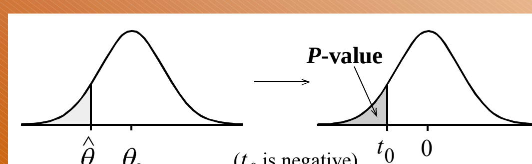


Figure 9.3.1: Testing H_0: $\theta = \theta_0$

Alternative Hypothesis	Evidence against H_0: $\theta = \theta_0$ provided by
H_1: $\theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0

Pictorial representation of the T-test
H_1: $\theta < \theta_0$



Is a second child gender influenced by the gender of the first child, in families with >1 kid?

TABLE 9.3.4 First and Second Births by Sex

		Second Child		
		Male	Female	Total
First Child	Male	3,202	2,776	5,978
	Female	2,620	2,792	5,412
	Total	5,822	5,568	11,390

- Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.
- Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.

Analysis of the birth-gender data – data summary

	Second Child	
Group	Number of births	Number of girls
1 (Previous child was girl)	5412	2792 (approx. 51.6%)
2 (Previous child was boy)	5978	2776 (approx. 46.4%)

- Let p_1=true proportion of girls in mothers with girl as first child, p_2=true proportion of girls in mothers with boy as first child. Parameter of interest is $p_1 - p_2$.
- H_0: $p_1 - p_2 = 0$ (skeptical reaction). H_a: $p_1 - p_2 > 0$ (research hypothesis)

Hypothesis testing as decision making

Decision Making

	Actual situation	
Decision made	H_0 is true	H_0 is false
Accept H_0 as true	OK	Type II error
Reject H_0 as false	Type I error	OK

- Sample sizes: n_1=5412, n_2=5978, Sample proportions (estimates) $\hat{p}_1 = 2792/5412 \approx 0.5159, \hat{p}_2 = 2776/5978 \approx 0.4644,$
- H_0: $p_1 - p_2 = 0$ (skeptical reaction). H_a: $p_1 - p_2 > 0$ (research hypothesis)

Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally diff. mothers they are independent → use formula 8.5.5.a

$$t_0 = \frac{\text{Estimate - HypothesizedValue}}{SE} = 5.49986 =$$

$$\frac{\hat{p}_1 - \hat{p}_2 - 0}{SE\left(\hat{p}_1 - \hat{p}_2\right)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} =$$

$$P-value = \Pr(T \geq t_0) = 1.9 \times 10^{-8}$$

Analysis of the birth-gender data

- We have strong evidence to reject the H_0, and hence conclude mothers with first child a girl a more likely to have a girl as a second child.
- How much more likely? A 95% CI:

CI ($p_1 - p_2$) =[0.033; 0.070]. And computed by:

$$\text{estimate} \pm z \times \text{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$0.0515 \pm 1.96 \times 0.0093677 = [3\%; 7\%]$$

Hypotheses

- The ***null hypothesis***, denoted by H_0, is the (skeptical reaction) hypothesis tested by the statistical test.
- *Principle guiding the formulation of null hypotheses*: We cannot rule a hypothesized value in; we can only determine whether there is enough evidence to rule it out. Why is that?
- ***Research (alternative) hypotheses*** lay out the conjectures that the research is designed to investigate and, if the researchers hunches prove correct, establish as being true.

Example: Is there racial profiling or are there confounding explanatory effects?!?

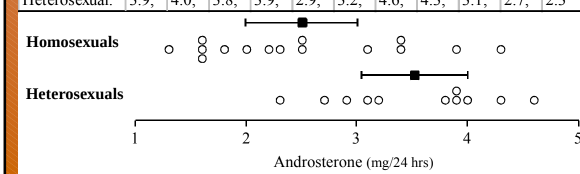
- The book by Best (*Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists*, Joel Best) shows how we can test for racial bias in police arrests. Suppose we find that among 100 white and 100 black youths, 10 and 17, respectively, have experienced arrest. This may **look plainly discriminatory**. But suppose we then find that of the 80 middle-class white youths 4 have been arrested, and of the 50 middle-class black youths 2 arrested, whereas the corresponding numbers of lower-class white and black youths arrested are, respectively, 6 of 20 and 15 of 50. These arrest rates correspond to 5 per 100 for white and 4 per 100 for black middle-class youths, and 30 per 100 for both white and black lower-class youths. Now, better analyzed, the data suggest **effects of social class, not race as such**.

One sample tests & Two independent samples tests

Analysis of two independent samples

Urinary androsterone levels – data, dot-plots and 95% CI. Relations between hormonal levels and homosexuality, Margolese, 1970. Hormonal levels are lower for homosexuals. Samples are independent, as unrelated. Results, P-value of t-test 0.004 with a CI ($\mu_{Het}-\mu_{Hom}$)=[0.4:1.7]. Normal hypothesis satisfied? Skewed?

Urinary Androsterone Levels(mg/24 hr)											
Homosexual:	2.5,	1.6,	3.9,	3.4,	2.3,	1.6,	2.5,	3.4,	1.6,	4.3,	2.0,
	1.8,	2.2,	3.1,	1.3							
Heterosexual:	3.9,	4.0,	3.8,	3.9,	2.9,	3.2,	4.6,	4.3,	3.1,	2.7,	2.3



Urinary androsterone levels cont.

Two Sample T-Test and Confidence Interval

```
Two sample T for androsterone
            N      Mean     StDev   SE Mean
hetero     11     3.518     0.721      0.22
homose     15     2.500     0.923      0.24
95% CI for mu (hetero) - mu (homose): ( 0.35,  1.69)
T-Test mu (hetero) = mu (homose) (vs not=):
                                  T=3.16  P=0.0044  DF=23
```

Confidence interval (pointing to the 95% CI); *t-test statistic* (pointing to T=3.16); *P-value* (pointing to P=0.0044)

Minitab 2-sample *t*-output for the androstenone data

Important points

1. The distinction between a randomized experiment and an observational study is made at the time of result interpretation. The very same statistical analysis is carried for the two situations.
2. We've already stressed the importance of plotting data prior to stat-analysis. Plots have many important roles – prevent dangerous misconceptions from arising (data overlaps, clusters, outliers, skewness, trends in the data, etc.)

Comparing two means for independent samples

Suppose we have 2 samples/means/distributions as follows: $\{\bar{x}_1, N(\mu_1,\sigma_1)\}$ and $\{\bar{x}_2, N(\mu_2,\sigma_2)\}$. We've seen before that to make inference about $\mu_1-\mu_2$ we can use a T-test for H_0: $\mu_1-\mu_2=0$ with $t_0=\frac{(\bar{x}_1-\bar{x}_2)-0}{SE(\bar{x}_1-\bar{x}_2)}$

And CI($\mu_1-\mu_2$) = $\bar{x}_1-\bar{x}_2 \pm t\times SE(\bar{x}_1-\bar{x}_2)$

If the 2 samples are independent we use the SE formula

$SE=\sqrt{s_1^2/n_1+s_2^2/n_2}$ with $df=Min(n_1-1;n_2-1)$

This gives a conservative approach for hand calculation of an approximation to the what is known as the Welch procedure, which has a complicated exact formula.

Means for independent samples – equal or unequal variances?

Pooled T-test is used for samples with assumed equal variances. Under data Normal assumptions and equal variances of $(\bar{x}_1 - \bar{x}_2 - 0)/SE(\bar{x}_1 - \bar{x}_2)$, where

$$SE = s_p\sqrt{1/n_1 + 1/n_2};\; s_p^2 = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

is exactly Student's t distributed with $df = (n_1 + n_2 - 2)$

Here s_p is called the pooled estimate of the variance, since it pools info from the 2 samples to form a combined estimate of the single variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The book recommends routine use of the Welch unequal variance method.

Comparing two means for independent samples

1. How sensitive is the two-sample *t*-test to non-Normality in the data? (The 2-sample T-tests and CI's are even more robust than the 1-sample tests, against non-Normality, particularly when the shapes of the 2 distributions are similar and $n_1=n_2=n$, even for small n, remember *df*= n_1+n_2-2.
3. Are there nonparametric alternatives to the *two-sample t-test*? (Wilcoxon rank-sum-test, Mann-Witney test, equivalent tests, same P-values.)
4. What difference is there between the quantities tested and estimated by the two-sample *t*-procedures and the nonparametric equivalent? (Non-parametric tests are based on ordering, not size, of the data and hence use median, not mean, for the average. The equality of 2 means is tested and $CI(\tilde{\mu}_1 - \tilde{\mu}_1)$.

Two sample tests - dependent samples & Estimation

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- **Two sample tests - dependent samples & Estimation**
- Correlation and regression techniques
- Non-parametric statistical tests

Paired Comparisons

- Sometimes we have two data sets, which are not independent, but rather observations matched in pairs.
- Back to the Kaufman & Rock study of the Moon size illusion. Does the moon size appear different with eyes level and with eyes raised? Does eye position make a difference? Eyes elevated refers to raising the eye from horizontal to zenith position. 10 Subjects are tested under eye-level (control) condition, by physically moving the subject's body from level to zenith position with fixed eye direction – horizontal. Ratios of the Moon size in level and zenith positions, for the two paradigms are given below.

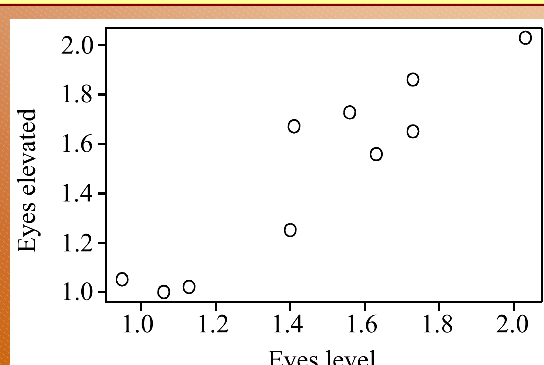
Moon illusion Data

The Moon Illusion

Subject	Eyes Elevated	Eyes Level	Difference (Elevated - Level)
1	2.03	2.03	0.00
2	1.65	1.73	-0.08
3	1.00	1.06	-0.06
4	1.25	1.40	-0.15
5	1.05	0.95	0.10
6	1.02	1.13	-0.11
7	1.67	1.41	0.26
8	1.86	1.73	0.13
9	1.56	1.63	-0.07
10	1.73	1.56	0.17

Source: Kaufman and Rock [1962].

Plotting Eyes elevated rations vs. eyes level rations

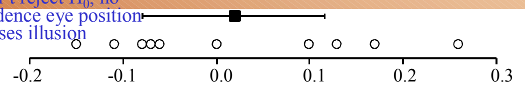


Looking for an effect due to elevating eyes

For ***paired*** data, ***analyze the differences***.

$H_0: \mu_{diff} = 0$

Can't reject H_0, no evidence eye position causes illusion



Dot plot of differences for the moon illusion data (with a 95% CI for the mean difference).

```
Test of mu = 0.0000 vs mu > 0.0000
Variable     N    Mean    StDev   SE Mean   t-stat  P-value
Difference  10   0.0190   0.1371   0.0434    0.44    0.34
                 95% CI ( -0.0791,  0.1171)
```


Paired data

- We have to distinguish between independent and related samples because they require different methods of analysis.
- Paired data is an example of related data.
- With paired data, we analyze the differences
  - this converts the initial problem into a one-sample problem.
- The ***sign test*** and ***Wilcoxon rank-sum*** test are nonparametric alternatives to the one-sample or paired *t*-test.

2-sample *t*-tests and intervals for differences between means $\mu_1 - \mu_2$

Assume

- statistically independent random samples from the two populations of interest
  - both samples come from Normal distributions
- Pooled method also assumes that $\sigma_1 = \sigma_2$
  Welch method (unpooled) does not

Two-sample *t*-methods are

- remarkably robust against non-Normality
- can be sensitive to the presence of outliers in small to moderate-sized samples
- One-sided tests are reasonably sensitive to skewness.

- The ***Wilcoxon*** or ***Mann-Whitney*** test is a nonparametric alternative to the two-sample *t*-test.

ANOVA – One-Way

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- **ANOVA**
- Correlation and regression techniques
- Non-parametric statistical tests

We know how to analyze 1 & 2 sample data. How about if we have than 2 samples – One-way ANOVA, *F*-test

One-way ANOVA refers to the situation of having one factor (or categorical variable) which defines group membership – e.g., comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.

Hypotheses for the one-way analysis-of-variance F-test

Null hypothesis: All of the underlying true means are identical.

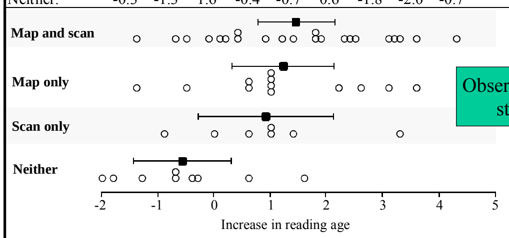
Alternative: Differences exist between some of the true means.

Comparing 4 reading methods

Comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.
-Mapping: using diagrams to relate main points in text;
-Scanning: reading the intro and skimming for an overview before reading details;
-Mapping and Scanning;
-Neither.
Table below shows increases in test scores, of 4 groups of students taking similar exams twice, w/ & w/o using a reading technique.
Research question: Are the results better for students using mapping, scanning or both?

TABLE 10.3.1 Increase in Reading Age

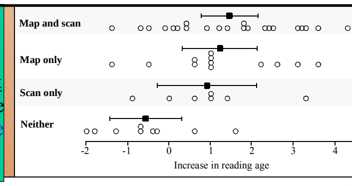
Both:	0.1	3.2	4.3	-0.5	1.9	3.3	2.5	3.6	0.4	2.3	-1.4	-0.7
	-0.1	0.2	0.4	0.9	1.2	1.4	1.8	1.8	2.4	3.1		
Map Only:	1.0	-0.5	1.0	0.6	0.6	1.0	1.0	-1.4	2.2	3.6	3.1	2.6
Scan Only:	1.0	3.3	1.4	-0.9	1.0	0.0	0.6					
Neither:	-0.3	-1.3	1.6	-0.4	-0.7	0.6	-1.8	-2.0	-0.7			



Observational study

Figure 10.3.1 Increases in reading ages with individual 95% CIs.

The F-test indicates that there's real evidence true differences exist it does not give indication of where the differences are or how large they are.



One-way Analysis of Variance

F-statistic *P-value* *Anova Table*

```
Analysis of Variance for Increase
Source     DF        SS        MS        F        P
Grp         3     27.06      9.02     4.45    0.008
Error      46     93.35      2.03
Total      49    120.41
                               Individual 95% CIs For Mean
                               Based on Pooled StDev
Level      N    Mean   StDev  ------+---------+---------+---------+
MapScan   22   1.459   1.544                        (------*-----)
MapOnly   12   1.233   1.441                      (-------*-------)
ScanOnly   7   0.914   1.302               (----------*----------)
Neither    9  -0.556   1.135  (--------*---------)
                              ------+---------+---------+---------+
Pooled StDev =    1.425            -1.0       0.0       1.0       2.0
```

Figure 10.3.2 Minitab analysis of variance output for reading ages

Computer output

One-way Analysis of Variance

F-statistic *P-value* *Anova Table*

```
Analysis of Variance for Increase
Source     DF        SS        MS        F        P
Grp         3     27.06      9.02     4.45    0.008
Error      46     93.35      2.03
Total      49    120.41
                               Individual 95% CIs For Mean
                               Based on Pooled StDev
Level      N    Mean   StDev  ------+---------+---------+---------+
MapScan   22   1.459   1.544                        (------*-----)
MapOnly   12   1.233   1.441                      (-------*-------)
ScanOnly   7   0.914   1.302               (----------*----------)
Neither    9  -0.556   1.135  (--------*---------)
                              ------+---------+---------+---------+
Pooled StDev =    1.425            -1.0       0.0       1.0       2.0
```

Figure 10.3.2 Minitab analysis of variance output for reading ages

Interpreting the *P*-value from the *F*-test

(The null hypothesis is that all underlying true means are identical.)

- A ***large P-value*** indicates that the differences seen between the sample means could be explained simply in terms of sampling variation.
- A ***small P-value*** indicates evidence that real differences exist between **at least some** of the true means, but gives *no indication* of where the differences are or how big they are.
- ***To find out how big*** any differences are we need confidence intervals.

Form of a typical ANOVA table

TABLE 10.3.2 Typical Analysis-of-Variance Table for One-Way ANOVA

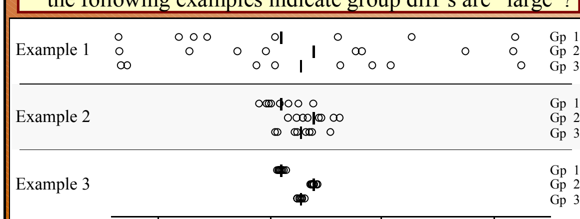
Source	Sum of squares	*df*	Mean sum of Squares[a]	*F*-statistic	*P*-value
Between	$\sum n_i(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2$	$k-1$	s_B^2	$f_0 = s_B^2/s_W^2$	pr($F \geq f_0$)
Within	$\sum (n_i-1)s_i^2$	$n_{tot}-k$	s_W^2		
Total	$\sum\sum (x_{ij}-\bar{x}_{\cdot\cdot})^2$	$n_{tot}-1$			

[a]Mean sum of squares = (sum of squares)/*df*

- The *F*-test statistic, f_0, applies when we have independent samples each from ***k*** Normal populations, N(μ_i, σ), note same variance is assumed.

Where did the F-statistics came from?

- Let's look at this example comparing groups. How do we obtain intuitive evidence against H_0? Far separated sample means + differences of sample means are large compared to their internal (within) variability! Which of the following examples indicate group diff's are "large"?



More about the F-test

- s^2_B is a measure of variability of sample means, how far apart they are.
- s^2_W reflects the avg. internal Variability within the samples.

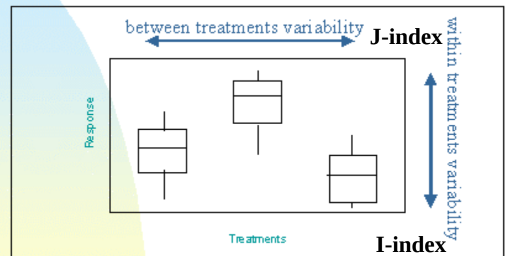
$$s_B^2 = \frac{\sum n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{k-1}$$

$$s_W^2 = \frac{\sum (n_i - 1)s_i^2}{n_{tot} - k}$$

- The F-test statistic, f_0, tests H_0 by comparing the variability of the sample means (numerator) with the variability within the samples (denominator).
- Evidence against H_0 is provided by values of f_0 which would be unusually large if H_0 was true.

What are x_i, $x_{..}$, $x_{.j}$, etc.?

One-Way Anova (Sources of Variability)



What are x_i, $x_{..}$, $x_{.j}$, etc.?

Need Online reference

Apple juice sales (units per week) →

H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: at least 2 means differ

$x_{i,j}$, $1 \le i \le n_j$; $1 < j \le 3$

City 1 Convenience	City 2 Quality	City 3 Price
529	804	672
658	630	531
793	774	443
514	717	596
663	679	602
719	604	502
711	620	659
606	697	689
461	706	675
529	615	512
498	492	691
663	719	733
604	787	698
495	699	776
485	572	561
557	523	572
353	584	469
557	634	581
542	580	679
614	624	532

What are x_i, $x_{..}$, $x_{.j}$, etc.?

Sum of Squares for treatments (cities)

$$SST = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$$SST = 20(577.55 - 613.07)^2 + 20(653.00 - 613.07)^2 + 20(608.65 - 613.07)^2 = 57{,}512.23$$

What are x_i, $x_{..}$, $x_{.j}$, etc.?

Sum of squares for the Error

Sum of Squares for Error:

$$SSE = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2\right]$$

$$SSE = 19(10{,}774.44) + 19(7{,}238.61) + 19(8{,}669.47) = 506{,}967.88$$

What are x_i, $x_{..}$, $x_{.j}$, etc.?

F-test

Test Statistic:

$$F = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{57{,}512.23/(3-1)}{506{,}967.88/(60-3)} = 3.23$$

Rejection Region: $F > F_{\alpha, k-1, n-k} = F_{.05,2,57} = 3.15$

Conclusion: Reject H_0

What are x_i, $x_{..}$, $x_{.j}$, etc.?
One-Way Design ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic
Treatments	k-1	SST	MST	MST/MSE
Error	n-k	SSE	MSE	
Total	n-1	SS(Total)		

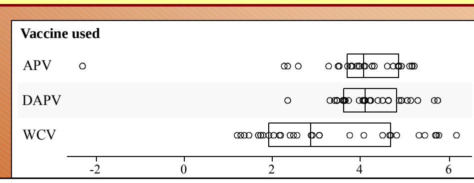
$$s_B^2 = \frac{\sum n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{k-1}$$

$$s_W^2 = \frac{\sum (n_i - 1)s_i^2}{n_{tot} - k}$$

Note: MST=SST/(k-1)

MSE=SSE/(n-k)

Skip - Pertussis data cont.




```
Analysis of Variance Procedure
Dependent Variable: PERTUSS
Source           DF   Sum of Squares   Mean Square   F Value   Pr > F
Model             2      15.81677991    7.90838996      6.17   0.0031
Error            87     111.47241564    1.28129213
Corrected Total  89     127.28919556
```

F-statistic ↓ *P-value* ↓

With the outlier included, the *P*-value increases to 0.023

Figure 10.3.5 Anova output for the pertussis data from SAS.

F-test assumptions

1. Samples are independent, physically independent subjects, units, objects are being studies.
2. Sample Normal distributions, especially sensitive for small n_i, number of observations, $N(\mu_i, \sigma)$.
3. Standard deviations should be equal within all samples, $\sigma_1 = \sigma_2 = \sigma_3 = \ldots \sigma_{n_k} = \sigma$. $(1/2 <= \sigma_k/\sigma_j <= 2)$

How to check/validate these assumptions for your data?

For the reading-score improvement data:

- independence is clear since different groups of students are used.
- Dot-plots of group data show no evidence of non-Normality.
- Sample SD's are very similar, hence we assume population SD's are similar.

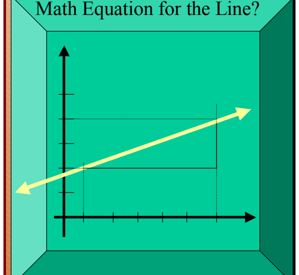
Correlation and regression techniques

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- **Correlation and regression techniques**
- Non-parametric statistical tests

Chapter 12: Lines in 2D
(Regression and Correlation)

- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- $Y = \alpha X + \beta$, in general.

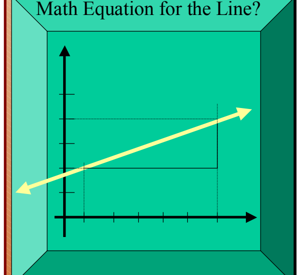
Math Equation for the Line?



Chapter 12: Lines in 2D
(Regression and Correlation)

- Draw the following lines:
- Y=2X+1
- Y=-3X-5
- Line through (X_1,Y_1) and (X_2,Y_2).
- $(Y-Y_1)/(Y_2-Y_1)=$
  $(X-X_1)/(X_2-X_1)$.

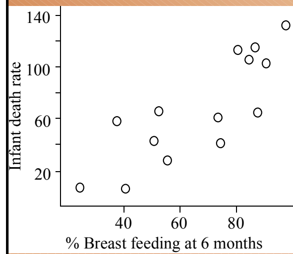
Math Equation for the Line?



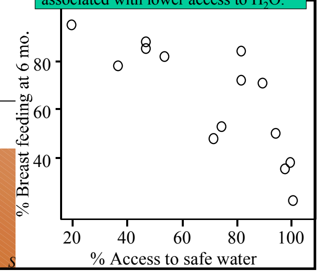
Approaches for modeling data relationships
Regression and Correlation

- There are random and nonrandom variables
- Correlation applies if <u>both variables (X/Y) are random</u> (e.g., We saw a previous example, systolic vs. diastolic blood pressure SISVOL/DIAVOL) and are <u>treated symmetrically</u>.
- Regression applies in the case when you want to <u>single out one of the variables</u> (response variable, Y) and use the other variable as predictor (explanatory variable, X), which explains the behavior of the response variable, Y.

Causal relationship?
– infant death rate (per 1,000) in 14 countries

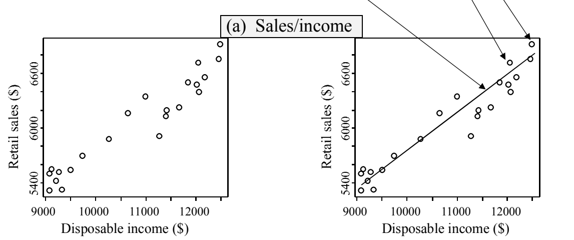


Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)? Naïve conclusion breast feeding is bad? But high rates of BF is associated with lower access to H_2O.



Predict behavior of Y (response) Based on the values of X (explanatory var.) Strategies for uncovering the reasons (causes) for an observed effect.

Regression relationship = trend + residual scatter



- Regression is a way of <u>studying relationships</u> between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (response) in <u>terms of</u> others (explanatory variables or predictors).

Correlation Coefficient

Correlation coefficient (-1<=R<=1): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_Y) and the correlation coefficient, R. R=1, <u>perfect positive correlation</u> (straight line relationship), R =0, <u>no correlation</u> (random cloud scatter), R = –1, <u>perfect negative correlation</u>.

Computing R(X,Y): (standardize, multiply, average)

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

X={$x_1, x_2, \ldots, x_N$,}
Y={$y_1, y_2, \ldots, y_N$,}
(μ_X, σ_X), (μ_Y, σ_Y) sample mean / SD.

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

Student i	Height x_i	Weight y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	167	60	6	4.67	36	21.8089	28.02
2	170	64	9	8.67	81	75.1689	78.03
3	160	57	-1	1.67	1	2.7889	-1.67
4	152	46	-8	-9.33	81	87.0489	83.97
5	157	55	-4	-0.33	16	0.1089	1.32
6	160	50	-1	-5.33	1	28.4089	5.33
Total	966	332	0	≈0	216	215.3334	195.0

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x = \frac{966}{6} = 161\,\text{cm}, \quad \mu_y = \frac{332}{6} = 55\,\text{kg},$$

$$\sigma_x = \sqrt{\frac{216}{5}} = 6.573, \quad \sigma_y = \sqrt{\frac{215.3}{5}} = 6.563,$$

$$Corr(X,Y) = R(X,Y) = 0.904$$

Correlation Coefficient - Properties

Correlation is pseudo-invariant w.r.t. linear transformations of X or Y

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) =$$

$$R(aX+b, cY+d), \qquad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right) = \left(\frac{ax_k+b-(a\mu_x+b)}{|a|\times\sigma_x}\right) =$$

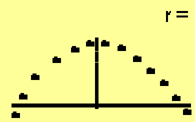
$$\left(\frac{a(x_k-\mu)+b-b}{|a|\times\sigma_x}\right) = sign(a)\left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

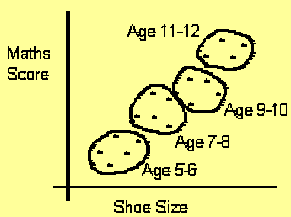
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



Correlation Coefficient - Properties

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

1. *R* measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.

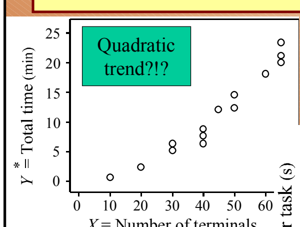


Trend and Scatter - Computer timing data

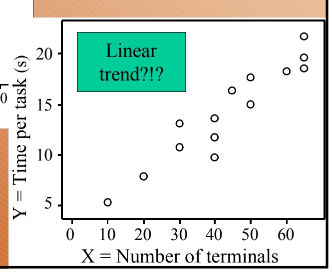
- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has X users, each running jobs taking Y min time. The main CPU swaps between all tasks. Y* is the total time to finish all tasks. Both Y and Y* increase with increase of tasks/users, but how?

X	=	Number of terminals:	40	50	60	45	40	10	30	20
Y*	=	Total Time (mins):	6.6	14.9	18.4	12.4	7.9	0.9	5.5	2.7
Y	=	Time Per Task (secs):	9.9	17.8	18.4	16.5	11.9	5.5	11	8.1
X	=	Number of terminals:	50	30	65	40	65	65		
Y*	=	Total Time (mins):	12.6	6.7	23.6	9.2	20.2	21.4		
Y	=	Time Per Task (secs):	15.1	13.3	21.8	13.8	18.6	19.8		

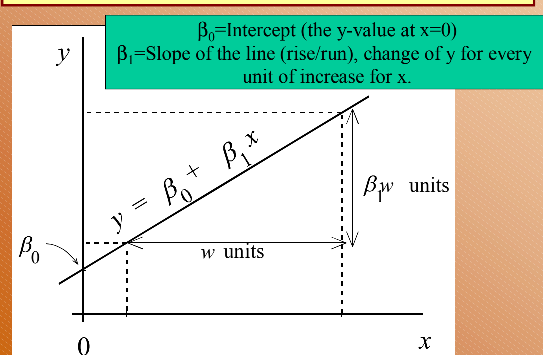
Trend and Scatter - Computer timing data



We want to find reasonable models (descriptions) for these data!



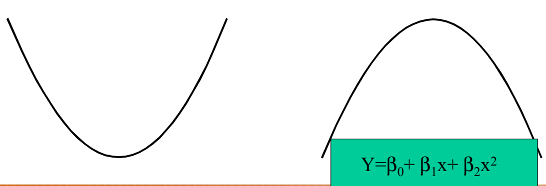
Equation for the straight line – linear/affine function



The quadratic curve

Quadratic Curve

β_2 positive β_2 negative

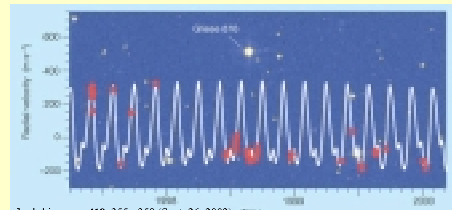


$Y=\beta_0+\beta_1x+\beta_2x^2$

Slide 224 FOR6520, AIU, Ivo Dinov

Other Non-linear model curves (trigonometric, piece-wise polynomial)

- Data from the Keck telescope in Hawaii (red points) show the variation over time of the radial velocity of the star *Gliese 876*. The white curve is the best fit to the data points, implying that there are two unseen planets perturbing the motion of the star and each other.

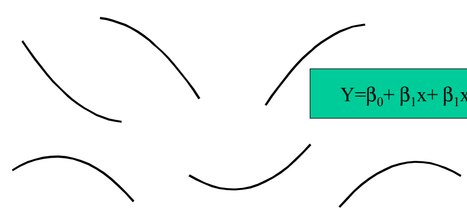


Nature, Jack Lissauer **419**, 355 - 358 (Sept. 26, 2002);

Slide 225 FOR6520, AIU, Ivo Dinov

The quadratic curve

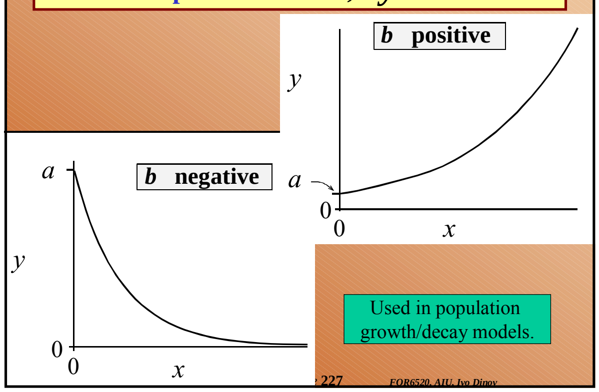
Segments of the curve



$Y=\beta_0+\beta_1x+\beta_1x^2$

Slide 226 FOR6520, AIU, Ivo Dinov

The exponential curve, $y = a\ e^{bx}$



Used in population growth/decay models.

Slide 227 FOR6520, AIU, Ivo Dinov

Choosing the "best-fitting" line

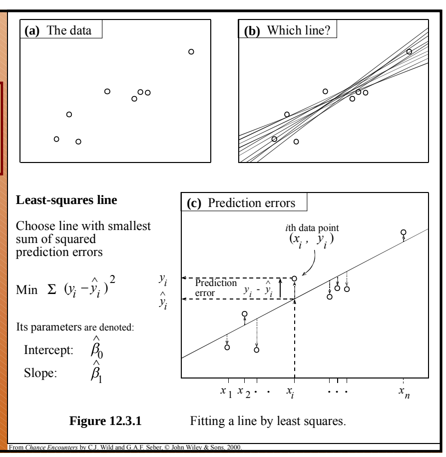
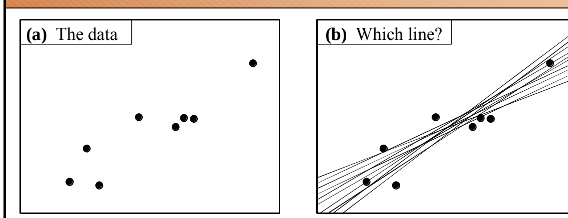


Figure 12.3.1 Fitting a line by least squares.

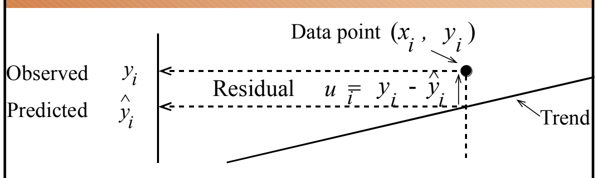
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000

Fitting a line through the data



Slide 229 FOR6520, AIU, Ivo Dinov

The idea of a residual or prediction error



Least squares criterion

Least squares criterion: Choose the values of the parameters to *minimize the sum of squared prediction errors* (or sum of squared residuals),

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The least squares line

Least-squares line

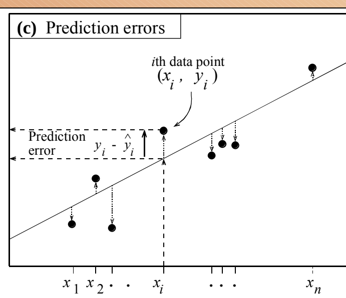
Choose line with smallest sum of squared prediction errors

Min $\Sigma\,(y_i - \hat{y}_i)^2$

Its parameters are denoted:

Intercept: $\hat{\beta}_0$

Slope: $\hat{\beta}_1$



Least-squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The least squares line

Least-squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Computer timings data – linear fit

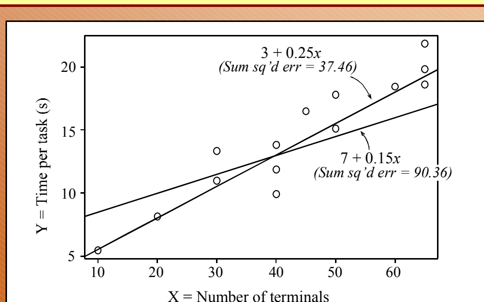


Figure 12.3.2 Two lines on the computer-timings data.

Computer timings data

TABLE 12.3.1 Prediction Errors

		3 + 0.25x		7 + 0.15x	
x	y	$\hat{y}$	$y-\hat{y}$	$\hat{y}$	$y-\hat{y}$
40	9.90	13.00	-3.10	13.00	-3.10
50	17.80	15.50	2.30	14.50	3.30
60	18.40	18.00	0.40	16.00	2.40
45	16.50	14.25	2.25	13.75	2.75
40	11.90	13.00	-1.10	13.00	-1.10
10	5.50	5.50	0.00	8.50	-3.00
30	11.00	10.50	0.50	11.50	-0.50
20	8.10	8.00	0.10	10.00	-1.90
50	15.10	15.50	-0.40	14.50	0.60
30	13.30	10.50	2.80	11.50	1.80
65	21.80	19.25	2.55	16.75	5.05
40	13.80	13.00	0.80	13.00	0.80
65	18.60	19.25	-0.65	16.75	1.85
65	19.80	19.25	0.55	16.75	3.05
Sum of squared errors			37.46		90.36

Adding the least squares line

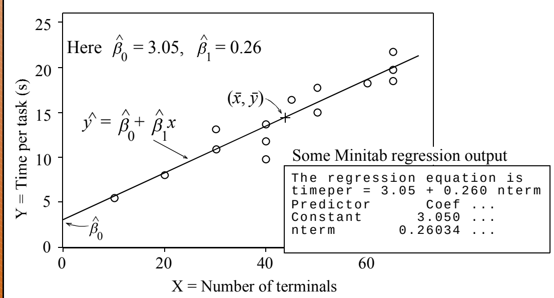


Figure 12.3.3 Computer-timings data with least-squares line.

Review, Fri., Oct. 19, 2001

1. The least-squares line $\hat y = \hat\beta_0 + \hat\beta_1 x$ passes through the points ($x = 0$, $\hat y$ = ?) and ($x = \bar x$, $\hat y$ = ?). Supply the missing values.

$$\hat\beta_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar x)(y_i - \bar y)\right]}{\sum_{i=1}^{n} (x_i - \bar x)^2}; \quad \hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2},

X	Y	$x-\bar x$	$y-\bar y$	$(x-\bar x)^2$	$(y-\bar y)^2$	$(x-\bar x)\times(y-\bar y)$	
-1	0						
2	-1						
3	1						
4	2						

$$\hat\beta_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar x)(y_i - \bar y)\right]}{\sum_{i=1}^{n} (x_i - \bar x)^2}; \quad \hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2}, $\bar x = 2$, $\bar y = 0.5$

X	Y	$x-\bar x$	$y-\bar y$	$(x-\bar x)^2$	$(y-\bar y)^2$	$(x-\bar x)\times(y-\bar y)$	
-1	0	-3	-0.5	9	0.25	1.5	
2	-1	0	-1.5	0	2.25	0	
3	1	1	0.5	1	0.25	0.5	
4	2	2	1.5	4	2.25	3	
2	0.5			14	5	5	

β_1=5/14
β_0=y^-β1*x^
β_0= 0.5-10/14

Course Material Review

1. ===========Part I=================
2. Data collection, surveys.
3. Experimental vs. observational studies
4. Numerical Summaries (5-#-summary)
5. Binomial distribution (prob's, mean, variance)
6. Probabilities & proportions, independence of events and conditional probabilities
7. Normal Distribution and normal approximation

Course Material Review – cont.

1. ===============Part II=================
2. Central Limit Theorem – sampling distribution of $\overline{X}$
3. Confidence intervals and parameter estimation
4. Hypothesis testing
5. Paired vs. Independent samples
6. Analysis Of Variance (1-way-ANOVA, one categorical var.)
7. Correlation and regression
8. Best-linear-fit, least squares method

Review

1. What are the quantities that specify a particular line?
2. Explain the idea of a prediction error in the context of fitting a line to a scatter plot. To what visual feature on the plot does a prediction error correspond? (scatter-size)
3. What property is satisfied by the line that fits the data best in the least-squares sense?
4. The least-squares line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through the points ($x = 0$, $\hat{y} = ?$) and ($x = \bar{x}$, $\hat{y} = ?$). Supply the missing values.

Motivating the simple linear model

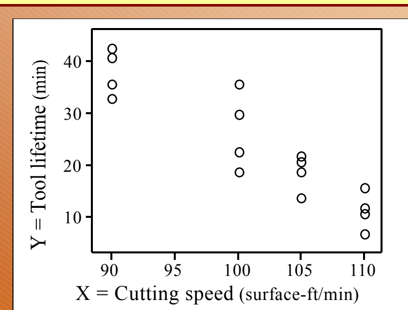
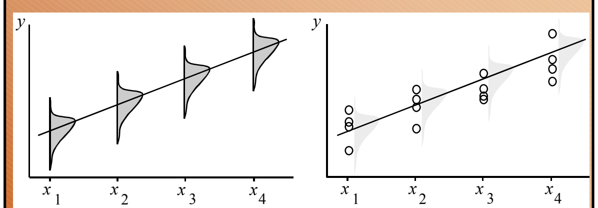


Figure 12.4.1 Lathe tool lifetimes.

The simple linear model



(a) The simple linear model (b) Data sampled from the model

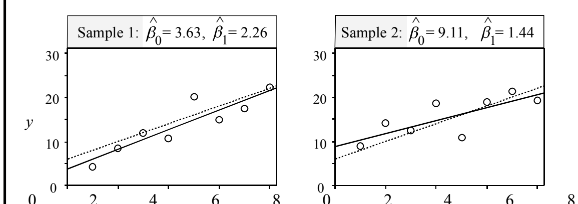
When $X = x$, $Y \sim \text{Normal}(\mu_Y, \sigma)$ where $\mu_Y = \beta_0 + \beta_1 x$, **OR**

when $X = x$, $\mathbf{Y = \beta_0 + \beta_1 x + U}$, where $U \sim \text{Normal}(0, \sigma)$

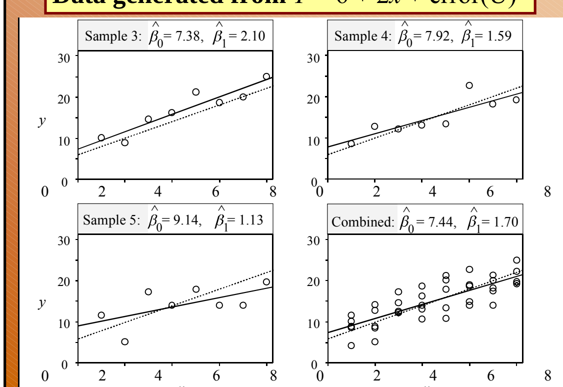
Random error

Data generated from $Y = 6 + 2x$ + error (U)

Dotted line ············ is true line and solid line ——— is the data-estimated LS line. Note differences between true $\beta_0=6$, $\beta_1=2$ and their estimates β_0^ & β_1^.



Data generated from $Y = 6 + 2x$ + error(U)



Data generated from $Y = 6 + 2x$ + error(U)

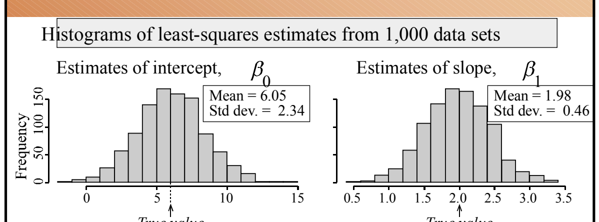


Figure 12.4.3 Data generated from the model $Y = 6 + 2x + U$ where $U \sim \text{Normal}(\mu = 0, \sigma = 3)$.

Summary

For the simple linear model, ***least-squares estimates are*** ***unbiased*** ***[E(β^)= β] and*** ***Normally distributed.***

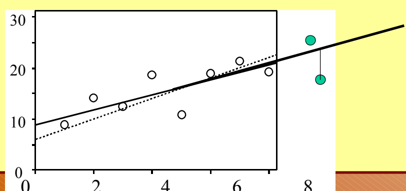
Noisier data produce *more-variable* least-squares estimates.

Summary

1. Before considering using the simple linear model, what sort of pattern would you be looking for in the scatter plot? (linear trend with constant scatter spread across the range of X)
2. What assumptions are made by the simple linear model, **SLM**? (X is linearly related to the mean value of the Y obs's at each X, $\mu_Y = \beta_0 + \beta_1 x$; where β_0 & β_1 are the true values of the intercept and slope of the SLM; The LS estimates β_0^ & β_1^ estimate the true values of β_0 & β_1; and the random errors U=Y-$\mu_Y \sim N(\mu, \sigma)$.)
3. If the simple linear model holds, what do you know about the sampling distributions of the least-squares estimates? (Unbiased and Normally distributed)

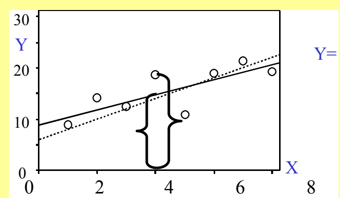
Summary

4. In the simple linear model, what behavior is governed by σ? (the spread of scatter of the data around trend)
5. Our estimate of σ can be thought of as a sample standard deviation for the set of prediction errors from the least-squares line.



RMS Error for regression

- Error = Actual value – Predicted value



- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

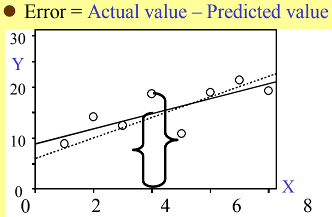
$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18



- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value
- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

- First compute the LS linear fit (estimate β_0^ + β_1^)
- Then Compute the individual errors
- Finally compute the cumulative RMS measure.

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

Compute the RMS Error for this regression line

- First compute the LS linear fit (estimate β_0^ +β_1^),μ_X=4.5,μ_X=15.75

X	Y	X-μ_X	Y- X-μ_Y	(X-μ_X)^2	(Y-μ_Y)^2	(X-μ_X)^2*(Y-μ_Y)^2
1	9					
2	15					
3	12					
4	19					
5	11					
6	20					
7	22					
8	18					
Total:						

- Compute

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Compute the RMS Error for this regression line

- Then Compute the individual errors

$$(y_k - \hat{y}_k)^2, \text{where} \quad \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 8$$

- Finally compute the cumulative RMS measure.

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

- Note on the Correlation coefficient formula,

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

X={$x_1, x_2, ..., x_N$}
Y={$y_1, y_2, ..., y_N$}
(μ_X, σ_X), (μ_Y, σ_Y)
sample mean / SD.

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

Recall the correlation coefficient…

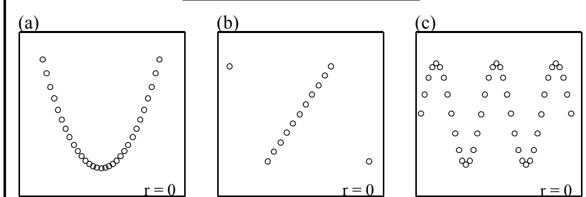
Another form for the correlation coefficient is:

$$R(X;Y) = Corr(X;Y) = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} =$$

$$\frac{\sum_{i=1}^{n}\left[y_i x_i + \bar{x}\bar{y} - \bar{y}x_i - \bar{x}y_i\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} = \frac{\sum_{i=1}^{n}\left(y_i x_i\right) - n\times\bar{x}\times\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}$$

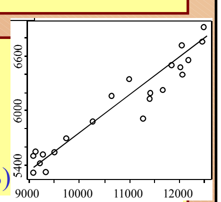
Misuse of the correlation coefficient

Some patterns with r = 0



Linear Regression

- Regression relationship = trend + residual scatter



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + Err$$

- Trend=best linear fit Line (LS)

$$\hat{\beta}_1 = \sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right] \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Scatter = residual (prediction) error Err=Obs-Pred

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = (y_1-\hat{y}_1)^2 + (y_2-\hat{y}_2)^2 + \ldots + (y_n-\hat{y}_n)^2$$

Another Notation for the Slope of the LS line

1. Note that there is a slight difference in the formula for the slope of the Least-Squares Best-Linear Fit line:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_1 = Corr(X;Y)\times\frac{SD(Y)}{SD(X)}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Non-parametric statistical tests

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- **Non-parametric statistical tests**

Flying helmet sizes for NZ Air Force

Measure the head-size of all air force recruits. Using cheaper cardboard or more expensive metal calipers. Are there systematic differences in the two measuring methods? Again, paired comparisons.

TABLE 10.1.2 **Air Force Head Sizes Data**

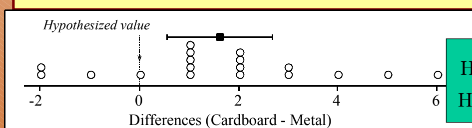
Recruit	Cardboard (mm)	Metal (mm)	Difference (Card-metal)	Sign of difference
1	146	145	1	+
2	151	153	-2	-
3	163	161	2	+
4	152	151	1	+
5	151	145	6	+
6	151	150	1	+

Helmet sizes for NZ Air Force – complete table

TABLE 10.1.2 **Air Force Head Sizes Data**

Recruit	Cardboard (mm)	Metal (mm)	Difference (Card-metal)	Sign of difference
1	146	145	1	+
2	151	153	-2	-
3	163	161	2	+
4	152	151	1	+
5	151	145	6	+
6	151	150	1	+
7	149	150	-1	-
8	166	163	3	+
9	149	147	2	+
10	155	154	1	+
11	155	150	5	+
12	156	156	0	0
13	162	161	1	+
14	150	152	-2	-
15	156	154	2	+
16	158	154	4	+
17	149	147	2	+
18	163	160	3	+

Head sizes: Does type of caliper make a difference?



$H_0: \mu_{diff} = 0$

$H_a: \mu_{diff} != 0$

Figure 10.1.8 Dot plot of differences in size (with 95% CI).

```
Paired T-Test and Confidence Interval
paired T for cardboard - metal

                 N      Mean     StDev   SE Mean
cardboard       18    154.56      5.82      1.37
metal           18    152.94      5.54      1.30
Difference      18     1.611     2.146     0.506
95% CI for mean difference: (0.544, 2.678)
T-Test of mean difference=0 (vs not=0): T-Value=3.19
                                         P-Value=0.005
```

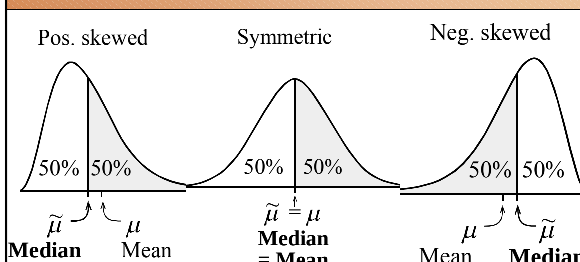
Figure 10.1.9 Minitab paired-*t* output for the size data.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Review

1. What is a paired-comparison experiment? (obs'd data are matched in pairs).
2. In a paired-comparison experiment, why is it wrong to treat the two sets of measurements as independent data sets? (data are usually taken from the same unit under diff. Treatments, so obs's should be related).
3. How do you analyze the data from a paired-comparison experiment? (analyze the difference).
4. What situations is appropriate to use the paired-comparison method to analyze the data? (pre- and post-metrifonate study using FDG PET imaging).

The population median – $\tilde{\mu}$



Definition of the population median

1. The population median is defined as the number in the middle of the distribution of the RV, i.e., 50% of the data lies below and 50% above the median.
2. Under what circumstances is the population median the same as the population mean? (symmetry of the distribution.)
3. Why do we use the population median rather than the population mean in the sign test? (for a skewed distribution, mean may not be representative, or may be outlier heavily influenced.)
4. Why is the model for the sign test like tossing a fair coin? (In the sign-test we test H_0: $\tilde{\mu}$ =0, under H_0 a random observation is as likely to be < $\tilde{\mu}$ as to be > $\tilde{\mu}$. So observation has + or – sign with the same probability, hence the coin-toss model, distribution-free, non-parametric approach. Testing H_0 is just like testing biased/unbiased coin).

Helmet paired head measurements

From the cardboard vs. metal caliper tests, we see 14 + and 3 – signs, implying larger overall measurements using the cardboard calipers. It's like tossing a coin 17 times and getting 14 heads. How likely is that?

If Y~Binomial(17, 0.5), number of successes (heads) in 17 fair coin tosses, then P(Y>=14)=0.00636, hence if we test p=0.5, vs. p!=0.5, two-tailed test, the chance is 2P(Y>=14)=0.0127.

Comments

5. What independence assumption must hold before the sign test is applicable? How important is it that this assumption is true? (requires that obs's are independent (one-sample test) and different pairs are independent (paired data), very sensitive.)
6. What advantages and disadvantages does the sign test have in comparison with the *t*-test? (Main advantage – test is distribution-free and insensitive to outliers. Disadvantage – when hypothesis for T-test, or a parametric test are met the CI are shorter and the parametric tests are more likely to detect departure from normality.)

Review

7. Why is the sign test called a distribution-free test? Does this mean that distributions are not used in performing the test? (no assumptions on the data underlying distribution, but distributions are actually used, e.g., Binomial).
8. In applying the sign test to paired data, how do you handle situations where both observations are tied (indistinguishable)? (ignore them)

Why Use Nonparametric Statistics?

- Parametric tests are based upon assumptions that may include the following:
  - The data have the **same variance**, regardless of the treatments or conditions in the experiment.
  - The data are **normally distributed** for each of the treatments or conditions in the experiment.
- What happens when we are not sure that these assumptions have been satisfied?

How Do Nonparametric Tests Compare with the Usual *z, t*, and *F* Tests?

- Studies have shown that when the usual assumptions are satisfied, nonparametric tests are about 95% efficient when compared to their parametric equivalents.
- When normality and common variance are not satisfied, the nonparametric procedures can be much more efficient than their parametric equivalents.

The Wilcoxon Rank Sum Test

- Suppose we wish to test the hypothesis that two distributions have the same center.
- We select two independent random samples from each population. Designate each of the observations from population 1 as an "**A**" and each of the observations from population 2 as a "**B**".
- If H_0 is true, and the two samples have been drawn from the same population, when we rank the values in both samples from small to large, **the A's and B's should be randomly mixed in the rankings**.

What happens when H_0 is true?

•Suppose we had 5 measurements from population 1 and 6 measurements from population 2.

•If they were drawn from the same population, the rankings might be like this.

ABABBBABABBA

•In this case if we summed the ranks of the A measurements and the ranks of the B measurements, the sums would be similar.

What happens if H_0 is not true?

- If the observations come from two different populations, perhaps with **population 1 lying to the left of population 2**, the ranking of the observations might take the following ordering.

AAABABABBB

In this case the sum of the ranks of the B observations would be larger than that for the A observations.

How to Implement Wilcoxon's Rank Test

•Rank the combined sample from smallest to largest.

•Let T_1 represent the sum of the ranks of the first sample (**A**'s).

•Then, T_1^* defined below, is the sum of the ranks that the A's would have had if the observations were ranked from *large to small*.

$$T_1^* = n_1(n_1 + n_2 + 1) - T_1$$

The Wilcoxon Rank Sum Test

H_0: the two population distributions are the same

H_a: the two populations are in some way different

- The **test statistic** is the smaller of T_1 and T_1^*.
- Reject H_0 if the test statistic is less than the **critical value** found in Table 7(a).
- Table 7(a) is indexed by letting population 1 be the one associated with the smaller sample size n_1, and population 2 as the one associated with n_2, the larger sample size.

Example

The wing stroke frequencies of two sp... l for a sample of n_1 = ... from species 2. C... tributions of wing strokes differ for these two species? Use $\alpha = .05$

If several measurements are tied, each gets the average of the ranks they would have gotten, if they were not tied! (See $x = 180$)

Species 1	Species 2
235 (10)	180 (3.5)
225 (9)	169 (1)
190 (8)	180 (3.5)
188 (7)	185 (6)
	178 (2)
	182 (5)

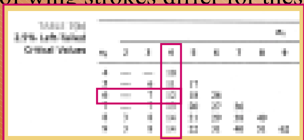
H_0: the two species are the same

H_a: the two species are in some way different

1. The sample with the smaller sample size is called sample 1.
2. We rank the 10 observations from smallest to largest, shown in parentheses in the table.

The Bee Problem

Can you conclude that the distributions of wing strokes differ for these two species? α = .05.



…nce in …ke

$\cdots 9+10=34$

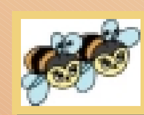
$\cdots + n_2+1)-T_1$

$\cdots 6+1)-34=10$

	178	(2)
	182	(5)

1. The test statistic is T = 10.
2. The critical value of T from Table 7(b) for a two-tailed test with α/2 = .025 is T = 12; H_0 is rejected if T ≤ 12.

Minitab Output



Recall $T_1=34; T_1^*=10.$

```
Mann-Whitney Test and CI: Species1, Species2

Species1   N =   4     Median =      207.50
Species2   N =   6     Median =      180.00
Point estimate for ETA1-ETA2 is       30.50
95.7 Percent CI for ETA1-ETA2 is (5.99,56.01)
W = 34.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant
at 0.0142
The test is significant at 0.0139 (adjusted for ties)
```

Minitab calls the procedure the Mann-Whitney U Test, equivalent to the Wilcoxon Rank Sum Test.

The test statistic is W = T_1 = 34 and has *p*-value = .0142. Do not reject H_0 for α = .05.

Large Sample Approximation: Wilcoxon Rank Sum Test

When n_1 and n_2 are large (greater than 10 is large enough), a normal approximation can be used to approximate the critical values in Table 7.

1. Calculate T_1 and T_1^*. Let $T=\min(T_1,T_1^*)$.

$T_1^* = n_1(n_1+n_2+1)-T_1$

2. The statistic $z=\frac{T-\mu_T}{\sigma_T}$ has an approximate z distribution with

T_1 = sum of the ranks of sample 1 (**A**'s).

$$\mu_T=\frac{n_1(n_1+n_2+1)}{2} \text{ and } \sigma_T^2=\frac{n_1n_2(n_1+n_2+1)}{12}$$

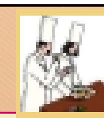
Some Notes

- When should you use the Wilcoxon Rank Sum test instead of the two-sample *t* test for independent samples?
  - ✓ when the responses can only be **ranked** and not quantified (e.g., ordinal qualitative data)
  - ✓ when the F test or the Rule of Thumb shows a **problem with equality of variances**
  - ✓ when a normality plot shows a **violation of the normality assumption**

The Sign Test

- The **sign test** is a fairly simple procedure that can be used to compare two populations when the samples consist of **paired observations.**
- It can be used
  - ✓ when the assumptions required for the **paired-difference test** of Chapter 10 are not valid or
  - ✓ when the responses can only be ranked as "**one better than the other**", but cannot be quantified.

The Sign Test



- ✓For each pair, measure whether the first response—say, A—exceeds the second response—say, B.
- ✓The test statistic is ***x***, the number of times that A exceeds B in the *n* pairs of observations.
- ✓Only pairs without ties are included in the test.
- ✓Critical values for the rejection region or exact *p*-values can be found using the cumulative binomial tables in Appendix I.

The Sign Test

H_0: the two populations are identical versus
H_a: one or two-tailed alternative
is equivalent to
H_0: p = P(A exceeds B) = .5 versus
H_a: p (≠, <, or >) .5
Test statistic: ***x* = number of plus signs**
Rejection region, *p*-values from Bin(n=size, p).

Example

Two gourmet chefs each tasted and rated eight different meals from 1 to 10. Does it appear that one of the chefs tends to give higher ratings than the other? Use α = .01.

Meal	1	2	3	4	5	6	7	8
Chef A	6	4	7	8	2	4	9	7
Chef B	8	5	4	7	3	7	9	8
Sign	-	-	+	+	-	-	0	-

H_0: the rating distributions are the same ($p = .5$)
H_a: the ratings are different ($p \neq .5$)

The Gourmet Chefs

Meal	1	2	3	4	5	6	7	8
Chef A	6	4	7	8	2	4	9	7
Chef B	8	5	4	7	3	7	9	8
Sign	-	-	+	+				

H_0: $p = .5$
H_a: $p \neq .5$ with $n = 7$ (omi
Test Statistic: x = number o

p-value =.454 is too large to reject H_0. There is insufficient evidence to indicate that one chef tends to rate one meal higher than the other.

Use Table 1 with n = 7 and p = .5.
p-value = P(observe x = 2 or something equally as unlikely)
= P($x \le 2$) + P($x \ge 5$) = 2(.227) = .454

k	0	1	2	3	4	5	6	7
P($x \le k$)	.008	.062	.227	.500	.773	.938	.992	1.000

Large Sample Approximation: The Sign Test

When $n \ge 25$, a normal approximation can be used to approximate the critical values of Binomial distribution.

1. Calculate x = number of plus signs.
2. The statistic $z = \dfrac{x - .5n}{.5\sqrt{n}}$ has an approximate z distribution.

Y~Bin(n, p) ➔
E(Y) = np
Var(Y) = np(1-p)

Example

You record the number of accidents per day at a large manufacturing plant for both the day and evening shifts for n = 100 days. You find that the number of accidents per day for the evening shift x_E exceeded th in the day shift x_D on 63 of t sufficient evidence to indica on one shift than on the othe

For a two tailed test, we reject H_0 if $|z| > 1.96$ (5% level). H_0 is rejected. There is evidence of a difference between the day and night shifts.

H_0: the distributions (# of a
H_a: the distributions are different ($p \neq .5$)

Test statistic: $z = \dfrac{x - .5n}{.5\sqrt{n}} = \dfrac{63 - .5(100)}{.5\sqrt{100}} = 2.60$

Which test should you use?

- We compare statistical tests using

  Definition: Power = $1 - \beta$

  = P(reject H_0 when H_a is true)
- The **power** of the test is the probability of rejecting the null hypothesis when it is false and some specified alternative is true.
- The **power** is the probability that the test will do what it was designed to do—that is, detect a departure from the null hypothesis when a departure exists.

Which test should you use?

- If all parametric assumptions have been met, the parametric test will be the most powerful.
- If not, a nonparametric test may be more powerful.
- If you can reject H_0 with a less powerful nonparametric test, you will not have to worry about parametric assumptions.
- If not, you might try
  - more powerful nonparametric test or
  - increasing the sample size to gain more power

The Wilcoxon Signed-Rank Test

- The **Wilcoxon Signed-Rank Test** is a more powerful nonparametric procedure that can be used to compare two populations when the samples consist of **paired observations.**
- It uses the **ranks** of the differences, $d = x_1 - x_2$ that we used in the paired-difference test.

The Wilcoxon Signed-Rank Test

✓For each pair, calculate the difference $d = x_1 - x_2$. Eliminate zero differences.

✓Rank the absolute values of the differences from 1 to n. Tied observations are assigned average of the ranks they would have gotten if not tied.

- **T^+ = rank sum for positive differences**
- **T^- = rank sum for negative differences**

✓If the two populations are the same, T^+ and T^- should be nearly equal. If either T^+ or T^- is unusually large, this provides evidence against the null hypothesis.

The Wilcoxon Signed-Rank Test

H_0: the two populations are identical versus

H_a: one or two-tailed alternative

Test statistic: $T = \min(T^+ \text{ and } T^-)$

Critical values for a one or two-tailed rejection region can be found using **Wilcoxon Signed-Rank Test Table.**

Example

To compare the densities of cakes using mixes A and B, six pairs of pans (A and B) were baked side-by-side in six different oven locations. Is there evidence of a difference in density for the two cake mixes?

Location	1	2	3	4	5	6
Cake Mix A	.135	.102	.098	.141	.131	.144
Cake Mix B	.129	.120	.112	.152	.135	.163
$d = x_A - x_B$	.006	-.018	-.014	-.011	-.004	-.019

H_0: the density distributions are the same
H_a: the density distributions are different

Cake Densities

Location	1	2	3
Cake Mix A	.135	.102	.098
Cake Mix B	.129	.120	.112
$d = x_A - x_B$	.006	-.018	-.014

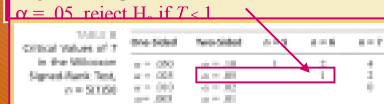
Rank	2	5	4	3	1	6

Do not reject H_0. There is insufficient evidence to indicate that there is a difference in densities for the two cake mixes.

Calculate: $T^+ = 2$ and $T^- = 5+4+3+1+6 = 19$.

The **test statistic** is $T = \min(T^+, T^-) = 2$.

Rejection region: Use Table 8. For a two-tailed test with $\alpha = .05$, reject H_0 if $T \leq 1$.



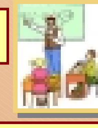
Large Sample Approximation: The Signed-Rank Test

When $n \geq 25$, a normal approximation can be used to approximate the critical values in Table 8.

1. Calculate T^+ and T^-. Let $T = \min(T^+, T^-)$.

2. The statistic $z = \dfrac{T - \mu_T}{\sigma_T}$ has an approximate z distribution with

$$\mu_T = \frac{n(n+1)}{4} \text{ and } \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$



The Kruskal-Wallis – *H* Test

- The **Kruskal-Wallis *H* Test** is a nonparametric procedure that can be used to compare more than two populations in a **completely randomized design.**
- **Non-parametric equivalent to ANOVA F-test!**
- All $n = n_1+n_2+\ldots+n_k$ measurements are jointly ranked.
- We use the sums of the ranks of the k samples to compare the distributions.

The Kruskal-Wallis – *H* Test

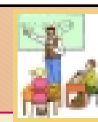
✓Rank the total measurements in all k samples from 1 to n. Tied observations are assigned average of the ranks they would have gotten if not tied.

✓Calculate

- **T_i = rank sum for the i^{th} sample $i = 1, 2, \ldots, k$**
- **$n = n_1+n_2+\ldots+n_k$**

✓And the test statistic H is (analog to: *F = MSST/MSSE*)

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{T_i^2}{n_i}\right) - 3(n+1)$$



The Kruskal-Wallis *H* Test

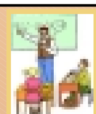
H_0: the k distributions are identical versus

H_a: at least one distribution is different

Test statistic: ***Kruskal-Wallis H***

When H_0 is true, the test statistic *H* has an approximate χ^2 distribution with *df* = *k*-1.

Use a right-tailed rejection region or p-value based on the Chi-square distribution.



Example

Four groups of students were randomly assigned to be taught with four different techniques, and their achievement test scores were recorded. Are the distributions of test scores the same, or do they differ in location?

1	2	3	4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88

Teaching Methods

	1	2	3	4
	65 (3)	75 (7)	59 (1)	94 (16)
	87 (13)	69 (5)	78 (8)	89 (15)
	73 (6)	83 (12)	67 (4)	80 (10)
	79 (9)	81 (11)	62 (2)	88 (14)
T_i	31	35	15	55

Rank the 16 measurements from 1 to 16 and c... the fo... sums

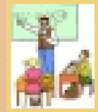
H_0: the distributions of scores are the same

H_a: the distributions differ in location

$$\text{Test statistic: } H = \frac{12}{n(n+1)}\sum\frac{T_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{16(17)}\left(\frac{31^2+35^2+15^2+55^2}{4}\right) - 3(17) = 8.96$$

Teaching Methods



H_0: the distributions of scores are the same
H_a: the distributions differ in location

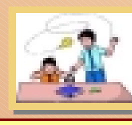
$$\text{Test statistic: } H = \frac{12}{n(n+1)}\sum\frac{T_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{16(17)}\left(\frac{31^2+35^2+15^2+55^2}{4}\right) - 3(17) = 8.96$$

Rejection region: Use Table 5. For a right-tailed chi-square test with $\alpha = .05$ and $df = 4-1 = 3$, reject H_0 if $H \geq 7.81$.

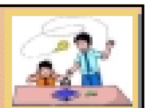
Reject H_0. There is sufficient evidence to indicate that there is a difference in test scores for the four teaching techniques.

The Friedman F_r Test



- The **Friedman F_r Test** is the nonparametric equivalent of the **randomized block design** with *k* treatments and *b* blocks.
- All *k* measurements within a block are ranked from 1 to *b*.
- We use the sums of the ranks of the *k* treatment observations to compare the *k* treatment distributions.
- Model: $X_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}$.
  1<=i<=k (treatment effects), ***1<=j<=n_i*** (block)

The Friedman F_r Test



✓Rank the *k* measurements, **within each block**, from from 1 to *k*. Tied observations are assigned average of the ranks they would have gotten if not tied.

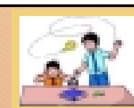
✓Calculate

- **T_i = rank sum for the i^{th} treatment $i = 1, 2,\ldots,k$**

✓and the test statistic

$$F_r = \frac{12}{b\times k\times(k+1)}\left(\sum_{i=1}^{k} T_i^2\right) - 3\times b\times(k+1)$$

The Friedman F_r Test



H_0: the *k* treatments are identical versus
H_a: at least one distribution is different

Test statistic: ***Friedman F_r***

When H_0 is true, the test statistic F_r has an approximate χ^2 distribution with $df = k-1$.

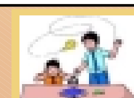
Use a right-tailed rejection region or *p*-value based on the Chi-square distribution.

Example

A student is subjected to a stimulus and we measure the time until the student reacts by pressing a button. Four students are used in the experiment, each is subjected to three stimuli, and their reaction times are measured. Do the distributions of reaction times differ for the three stimuli?

	Stimuli		
Subject	1	2	3
1	.6	.9	.8
2	.7	1.1	.7
3	.9	1.3	1.0
4	.5	.7	.8

Reaction Times



	Stimuli					
Subject	1		2		3	
1	.6	(1)	.9	(3)	.8	(2)
2	.7	(1.5)	1.1	(3)	.7	(1.5)
3	.9	(1)	1.3	(3)	1.0	(2)
4	.5	(1)	.7	(2)	.8	(3)
T_i		4.5		11		8.5

Rank the 3 measurements each subj[...] 1 to 3, and calculate [...] rank sums

H_0: the distributions of reaction times are the same
H_a: the distributions differ in location

$$\text{Test statistic: } F_r = \frac{12}{bk(k+1)}\sum T_i^2 - 3b(k+1)$$

$$= \frac{12}{12(4)}(4.5^2+11^2+8.5^2) - 3(4)(4) = 5.375$$

Reaction Times

H_0: the distributions of reaction times are the same
H_a: the distributions differ in location

$$\text{Test statistic: } F_r = \frac{12}{bk(k+1)}\sum T_i^2 - 3b(k+1)$$
$$= \frac{12}{12(4)}(4.5^2 + 11^2 + 8.5^2) - 3(4)(4) = 5.375$$

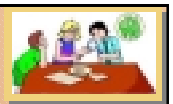
Rejection region: Use Table 5. For a right-tailed chi-square test with $\alpha = .05$ and $df = 3\text{-}1 = 2$, reject H_0 if $H \geq 5.99$.

Do not reject H_0. There is insufficient evidence to indicate that there is a difference in reaction times for the three stimuli.

Rank Correlation Coefficient

- The **rank correlation coefficient, Spearman r_s** is the nonparametric equivalent of the Pearson correlation coefficient ***r***.
- The two variables are each ranked from smallest to largest and the **ranks are denoted as *x* and *y*.**
- We are interested in the strength of the relationship (correlation) between the two variables.

Rank Correlation Coefficient



$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \text{ where}$$
$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}; \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n};$$
$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$
$$\text{If there are no ties, } r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$
where $d = x - y$.

Rank Correlation Coefficient

H_0: no association between the rank pairs
H_a: one or two-tailed alternative
Test statistic: r_s
Critical values for a one or two-tailed rejection region can be found using

Example

Five elementary school science teachers have been ranked by a judge according to their teaching ability. They have also taken a national "teacher's exam". Is there agreement between the judge's rank and the exam score?

Teacher	1	2	3	4	5
Judge's Rank	4	2	3	1	5
Exam score	72	69	82	93	80

If the judge's rank is low (best teacher), we might expect the teacher's score to be high. We look for a negative association between the ranked measurements.

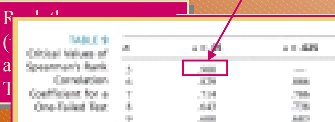
Example

Teacher	1	2	3	4	5
Judge's Rank	4	2	3	1	5
Exam score	72	69	82	93	80

Teacher	1	2	3
x	4	2	3
y	2	1	4
d	2	1	-1

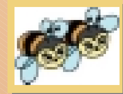
For a one-tailed test with $\alpha = .05$ and $n = 5$, reject H_0 if $r_s \geq .900$. We do not reject H_0. Not enough evidence to indicate a negative association.

R
(
a
T



$$1 - \frac{6(26)}{5(25-1)} = -0.3$$

Summary



- The nonparametric analogues of the parametric procedures presented in Chapters 10–14 are straightforward and fairly simple to implement.
- The **Wilcoxon rank sum test** is the nonparametric analogue of the two-sample *t* test.
- The **sign test** and the **Wilcoxon signed-rank test** are the nonparametric analogues of the paired-sample *t* test.
- The **Kruskal-Wallis *H* test** is the rank equivalent of the one-way analysis of variance *F* test.
- The **Friedman F_r test** is the rank equivalent of the randomized block design two-way analysis of variance *F* test.
- **Spearman's rank correlation r_s** is the rank equivalent of Pearson's correlation coefficient.

Slide 314 FOR6520, AIU, Ivo Dinov

Key Concepts

I. Nonparametric Methods

1. These methods can be used when the data cannot be measured on a quantitative scale, or when
2. The numerical scale of measurement is arbitrarily set by the researcher, or when
3. The parametric assumptions such as normality or constant variance are seriously violated.

II. Wilcoxon Rank Sum Test: Independent Random Samples

1. Jointly rank the two samples: Designate the smaller sample as sample 1. Then

$$T_1 = \text{Rank of sample}_1 \Rightarrow T_1^* = n_1(n_1 + n_2 + 1) - T_1$$

Slide 315 FOR6520, AIU, Ivo Dinov

Key Concepts

2. Use T_1 to test for population 1 to the left of population 2

   Use T_1^* to test for population to the right of population 2.

   Use the smaller of T_1 and T_1^* to test for a difference in the locations of the two populations.
3. Table 7 of Appendix I has critical values for the rejection of H_0.
4. When the sample sizes are large, use the normal approximation:

$$z = \frac{T - \mu_T}{\sigma_T}$$

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Slide 316 FOR6520, AIU, Ivo Dinov

Key Concepts

III. Sign Test for a Paired Experiment

1. Find *x*, the number of times that observation A exceeds observation B for a given pair.
2. To test for a difference in two populations, test $H_0: p = 0.5$ versus a one- or two-tailed alternative.
3. Use Table 1 of Appendix I to calculate the *p*-value for the test.
4. When the sample sizes are large, use the normal approximation:

$$z = \frac{x - .5n}{.5\sqrt{n}}$$

Slide 317 FOR6520, AIU, Ivo Dinov

Key Concepts

IV. Wilcoxon Signed-Rank Test: Paired Experiment

1. Calculate the differences in the paired observations. Rank the absolute values of the differences. Calculate the rank sums T^- and T^+ for the positive and negative differences, respectively. The test statistic *T* is the smaller of the two rank sums.
2. Table 8 of Appendix I has critical values for the rejection of for both one- and two-tailed tests.
3. When the sampling sizes are large, use the normal approximation:

$$z = \frac{T - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)]/24}}$$

Slide 318 FOR6520, AIU, Ivo Dinov

Key Concepts

V. Kruskal-Wallis *H* Test: Completely Randomized Design

1. Jointly rank the *n* observations in the *k* samples. Calculate the rank sums, T_i = rank sum of sample *i*, and the test statistic

$$H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n+1)$$

2. If the null hypothesis of equality of distributions is false, *H* will be unusually large, resulting in a one-tailed test.
3. For sample sizes of five or greater, the rejection region for *H* is based on the chi-square distribution with $(k - 1)$ degrees of freedom.

Slide 319 FOR6520, AIU, Ivo Dinov

Key Concepts

VI. The Friedman F_r Test: Randomized Block Design

1. Rank the responses within each block from 1 to k. Calculate the rank sums $T_1, T_2, \ldots, T_k$, and the test statistic

$$F_r = \frac{12}{bk(k+1)} \sum T_i^2 - 3b(k+1)$$

2. If the null hypothesis of equality of treatment distributions is false, F_r will be unusually large, resulting in a one-tailed test.
3. For block sizes of five or greater, the rejection region for F_r is based on the chi-square distribution with $(k - 1)$ degrees of freedom.

Key Concepts

VII. Spearman's Rank Correlation Coefficient

1. Rank the responses for the two variables from smallest to largest.
2. Calculate the correlation coefficient for the ranked observations:

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad \text{or} \quad r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{if there are no ties}$$

3. Table 9 in Appendix I gives critical values for rank correlations significantly different from 0.
4. The rank correlation coefficient detects not only significant linear correlation but also any other monotonic relationship between the two variables.

Title

- What

Title

- What