Statistics for Data Science

Abdisa G. (PhD)

Emerald International College: School of Computing and Analytics

Course Information

Statistics for Data Science
Programe: MSc in Data Science

Introduction to Statistics and Probability

- Data Types and Variable Types
- Statistics, Sampling Techniques and Probability
- Information Gain and Entropy
- Probability Theory, Types and Distribution Functions
- Inferential Statistics

Types of Variables

Variable: A variable is a characteristics or attribute that can take on different possible value or outcome.

Example Blood pressure, enzyme level, height, weight, sex, salary, etc

A variable whose values are determined by measurement, counting or by chance.

Variable and Types of Data

The measurement or observation value of a variable is called a data and collection of data is known as data set

A variable can be qualitative or quantitative

Qualitative or Categorical variable: A variable or characteristic which can not be measured in quantitative(numerical) form but can only be sorted (or grouped) by name or categories.

Quantitative (Numerical) Variable: A variable that can be measured (or counted) and expressed numerically.

E.g. Observations regarding height, income, weight, age, etc...

A quantitative variable is divided into two

Discrete Variable : A variable that can assume only certain counted values and for which there is gaps between any two possible values.

The values aren't just labels, but are actual measurable quantities.

Example Number of students in class, Number of tree species in Ethiopia , Number of children/ family, number of bacteria colonies on a plate.

Continuous Variable: : It can have an infinite number of possible values in any given interval. There is no gap between any two possible values.

Example All variable whose values obtained through measurement like height, weight, length

Data Types

- 4 Types of Data
- Qualitative Data Type
  - Nominal
  - Ordinal
- Quantitative Data Type
  - Discrete
  - Continuous

Quantitative (Numerical) Variable: A variable that can be measured (or counted) and expressed numerically.

E.g. Observations regarding height, income, weight, age, etc...

A quantitative variable is divided into two

Discrete Variable : A variable that can assume only certain counted values and for which there is gaps between any two possible values.

The values aren't just labels, but are actual measurable quantities.

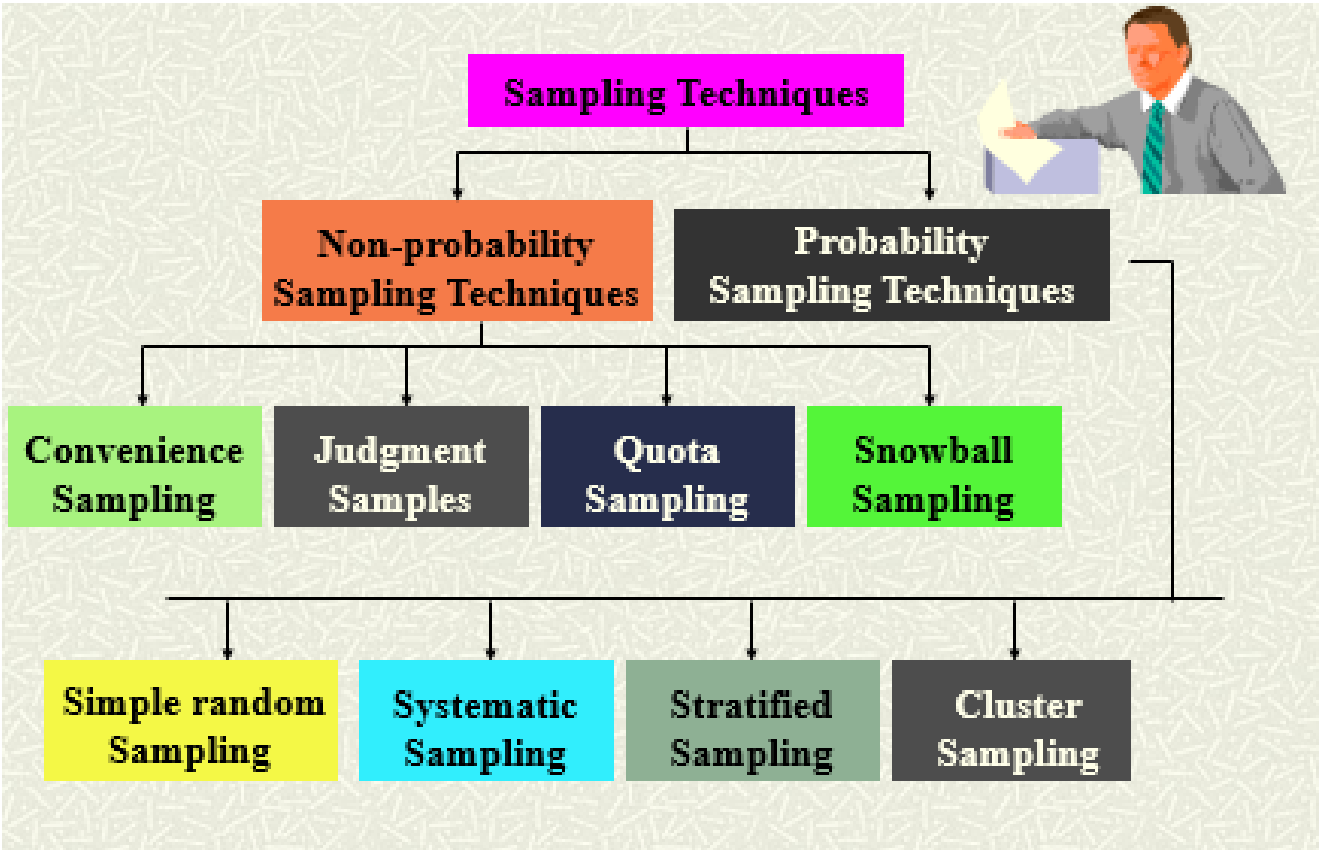
Example Number of students in class, Number of tree species in Ethiopia , Number of children/ family, number of bacteria colonies on a plate.

Continuous Variable: : It can have an infinite number of possible values in any given interval. There is no gap between any two possible values.

Example All variable whose values obtained through measurement like height, weight, length

Sampling Methods/types

Two broad categories of sampling procedures:
probability methods and **non-probability methods**.



A. Probability sampling

Involves random selection of a sample

A sample is obtained in a way that ensures every member of the population to have a known (non zero) probability of being included in the sample.

Procedures to ensure that each unit of the sample is chosen on the basis of chance.

Every sampling unit has a known and non-zero probability of selection into the sample.

Sample finding can be generalizable

more complex, more time consuming and usually more costly

The method chosen depends on a number of factors, such as

The available sampling frame,

How spread out the population is,

How costly it is to survey members of the population

Homogeneity of the population

Most common probability sampling methods

1. Simple random sampling
2. Systematic random sampling
3. Stratified random sampling
4. Cluster sampling
5. Multi-stage sampling

1. Simple Random Sampling

The required number of individuals are selected at random from the sampling frame, a list or a database of all individuals in the population.

Each member of a population has an equal chance/probability of being included in the sample.

Representativeness of the sample is ensured

Procedure

- Take sampling population
- Make a numbered list of all the units in the population (sampling frame)
- Each unit should be numbered from 1 to N (where N is the size of the population)
- Randomly draw the required numbers/units

The randomness of the sample is ensured by:

- Use of lottery methods (sample drawn from box)
- Table of random numbers
- Computer generated random numbers

Table B.1: Random Numbers Table

	A	B	C	D	E	F	G	H	I	J
1	8450	0992	6563	0340	2649	6933	9446	6182	2601	7800
2	5952	1443	7100	8444	3904	0159	1849	2601	9763	9068
3	5711	6779	9388	9668	4167	1423	2744	4622	2179	8503
4	2683	8047	0494	7853	8411	5406	8122	9577	8530	2350
5	0739	3114	3997	3482	3226	2216	6874	0620	8521	2938
6	8985	2463	5054	3448	6357	0187	6342	4740	4064	5068
7	7644	9339	8375	4583	7715	6355	6827	2055	9328	3287
8	6277	6631	8797	3693	6370	1436	1599	6267	2758	0323
9	6355	7590	7628	9054	0022	4241	7499	3430	3644	6576
10	7828	0589	3075	1954	5972	2266	0056	1097	9706	9009
11	6026	4546	4119	1554	4895	3123	9849	2054	5062	6711
12	8416	1972	9346	1593	2943	2379	5062	4829	5962	8292
13	1433	8823	7706	5273	6160	2161	5530	8617	7894	0175
14	0672	4884	8113	4447	5735	6347	7280	2301	2330	0693
15	4104	7184	1184	3964	2119	6968	0405	3827	0845	8400
16	4272	4979	1471	0942	9573	4283	1552	0161	3957	2516
17	1225	4171	3433	8700	0042	5884	2508	3250	1520	6366
18	7442	6575	1927	7267	7182	3960	4341	0350	1126	5945
19	4911	9007	3048	0319	0916	3002	1466	4421	7246	7662
20	3143	7402	4486	0909	1858	7961	1211	6296	5545	4588
21	8055	9294	2578	0426	4322	6925	2487	5677	9491	4301
22	9240	5260	7134	8001	0140	3094	8437	4066	2856	0933
23	7923	8630	3654	2638	2868	1059	0903	3114	6351	8261
24	0020	5104	4344	3324	9214	6615	5926	7012	9052	9205
25	3312	5823	5469	5171	4877	5392	3394	5077	3750	5637
26	3466	4193	5330	4680	0456	5891	3175	5733	5678	0956
27	1677	1694	1697	8921	2520	2811	3597	1355	9605	3637
28	3846	6283	0969	0051	5857	1043	1671	2013	8955	7706
29	8084	2327	0550	7231	1087	4830	9742	5654	5458	8290
30	7715	2247	4504	1374	9236	7340	1773	0693	2749	1335
31	6537	5815	9312	1460	6593	7678	4312	7537	9360	7195
32	4263	8931	1642	6694	1925	2661	1274	7346	8234	3159
33	7468	4077	6691	3961	7640	2355	9938	8485	9398	8364
34	4884	3324	3690	7433	1246	0523	4483	5933	5634	0612
35	7222	7299	1346	8937	0933	1569	5662	3735	2982	5966
36	5040	0820	8606	4006	4743	6343	4873	1002	4757	1075
37	2980	4860	5694	1501	5791	9414	7245	1283	9766	7427
38	8660	5480	7436	9745	8809	3007	4916	6543	9830	6099
39	7627	4959	6417	3642	1877	0370	5464	9690	5184	7379
40	1890	7664	7144	3623	8465	0085	8174	4740	3654	5543
41	3175	2580	3919	7436	0796	1018	5565	1142	4577	0457
42	7616	9338	6304	0283	6502	9085	5443	1531	9724	4140
43	5223	4525	0895	9530	0050	2201	5270	6447	1850	2070
44	9384	9794	8418	0374	4119	2075	0067	4535	7769	4719
45	5862	9165	5302	9089	5771	9670	7523	9280	2604	0212
46	9450	9307	6597	7183	5243	8854	6735	2415	0364	3096

Advantages

- Simple
- Sampling error easily measured

Disadvantages

- Need complete list of units
- Units may be scattered and poorly accessible
- Heterogeneous population

2. Systematic Random Sampling

Sometimes called interval sampling

Selection of individuals from the sampling frame is done systematically rather than randomly.

Individuals are taken at regular intervals down the list (for example every k^{th})

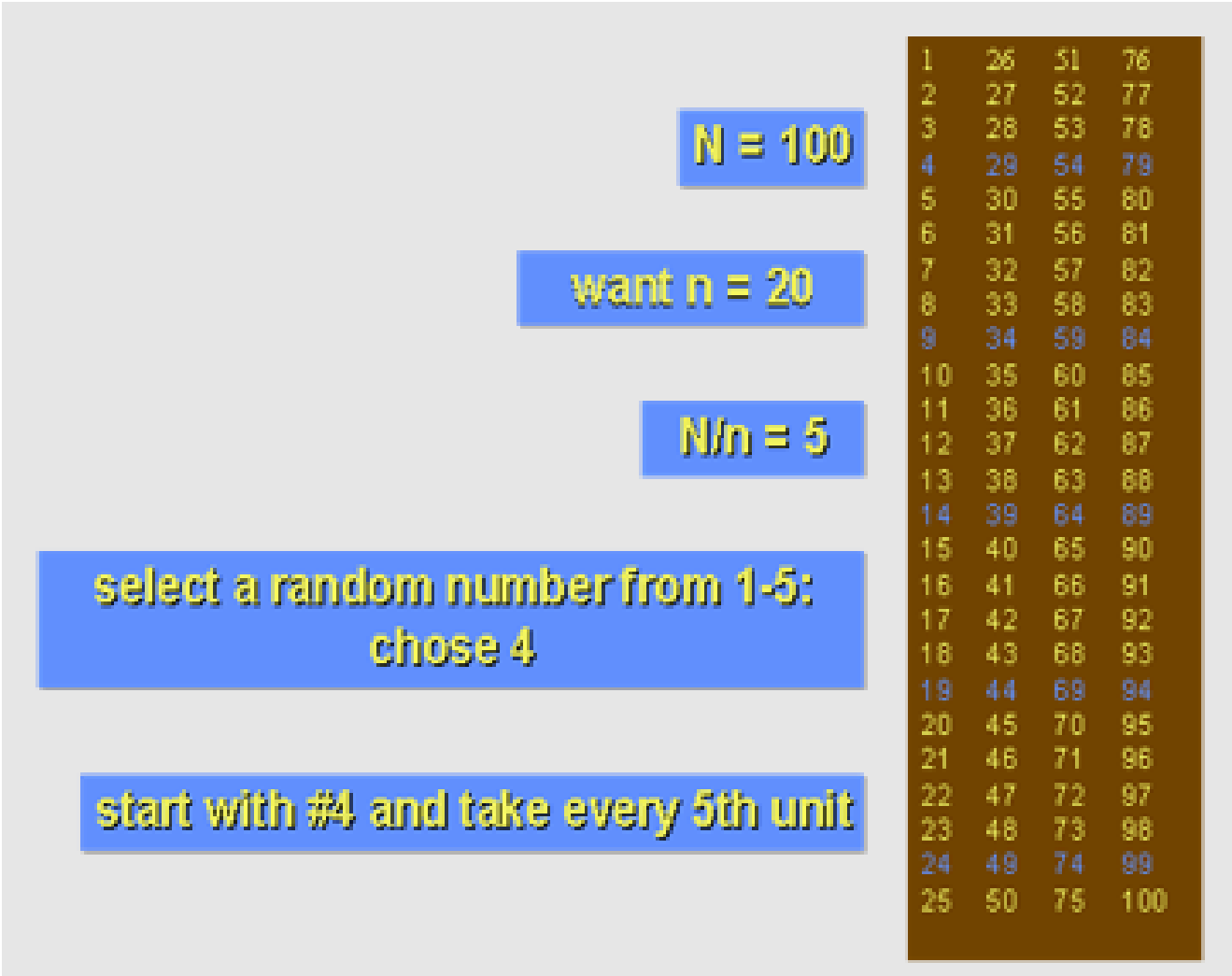
The first unit to be selected is taken at random from among the first K units.

Procedure

- Arrange the units in some kind of sequence (from 1 to N)
- Determine the sampling interval (K) by dividing the number of units in the population by the desired sample size (eg N/n=k)
- Choose a random starting point (for k, the starting point will be a random number between 1 and k)
- Select every k^{th} unit after that first number

To select a sample of 20 from a population of 100, you would need a sampling interval of $100/20 = 5$.

Therefore, $K = 5$.



Advantage

Systematic sampling usually less time consuming and easier to perform than SRS. It provide good approximation to SRS

Unlike SRS, systematic sampling can be conducted without sampling frame (usually in some situation sampling frame not readily available)

Disadvantages

Periodicity-underlying pattern may be a problem (characteristics occurring at regular intervals)

3. Stratified random sampling

It is done when the population is known to have heterogeneity with regard to some factors and those factors are used for stratification

Using stratified sampling, the population is divided into homogeneous, mutually exclusive groups called strata, and

A population can be stratified by any variable that is available for all units prior to sampling (e.g., age, sex, province of residence, income, etc.).

A separate sample is taken independently from each stratum

Procedure

- Divide (stratify) sampling frame into homogeneous subgroups (strata) e.g. minorities, urban/rural areas, occupations
- Draw random sample within each stratum

The sampling method can vary from one stratum to another

Proportionate allocation- if the same sampling fraction is used for each stratum

Non-proportionate allocation- the strata unequal in size and a fixed number of units is selected from each stratum

Advantages

- representativeness of the sample is improved.
- focuses on important subpopulations and ignores irrelevant ones
- improves the accuracy of estimation

Disadvantages

- can be difficult to select relevant stratification variables
- not useful when there are no homogeneous subgroups
- can be expensive
- Sampling error is difficult to measure

Example: A sample of 50 students is to be drawn from a population consisting of 500 students belonging to two institutions A and B. The number of students in the institution A is 200 and the institution B is 300. How will you draw the sample using proportional allocation?

Solution: There are two strata in this case with sizes

$N_1 = 200$ and $N_2 = 300$ and the total population
$N = N_1 + N_2 = 500$ The sample size is 50.
If n_1 and n_2 are the sample sizes,

$$n_1 = \frac{n}{N} \times N_1 = \frac{50}{500} \times 200 = 20$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{50}{500} \times 300 = 30$$

The sample sizes are 20 from A and 30 from B.

Then the units from each institution are to be selected by simple sampling

Advantage

The representativeness of the sample is improved . That is, adequate representation from each group.

Minority subgroups of interest can be ensured by stratification and by varying the sample fraction between strata as required.

Disadvantages

Sampling frame for the entire population has to be prepared separately for each stratum.

4. Cluster sampling

Sometimes it is too expensive to carry out SRS

- Population may be large and scattered.
- Complete list of the study population unavailable
- Travel costs can become expensive if interviewers have to survey people from one end of the country to the other.

Cluster sampling is the most widely used to reduce the cost

The clusters should be homogeneous, unlike stratified sampling where the strata are heterogeneous

Steps in cluster sampling

- Whole population divided into groups e.g. neighbourhoods
- A type of multi-stage sampling where all units at the lower level are included in the sample
- Random sample taken of these groups (clusters)

- Within selected clusters, all units e.g. households included (or random sample of these units)
- Provides logistical advantage

Involves selection of groups called clusters followed by selection of individuals within each selected cluster.

Can be used when it is either impossible or impractical to compile exhaustive list of individuals of the target population.

Cluster sampling is recommended for its efficiency, however accuracy is less because it is subject to more than one sampling error unlike SRS.

Advantages

- Simple as complete list of sampling units within population not required
- Less travel/resources required

Disadvantages

- Cluster members may be more alike than those in another cluster (homogeneous)
- this dependence needs to be taken into account in the sample size and in the analysis (design effect)

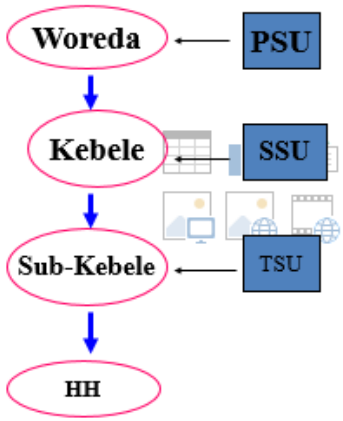
5. Multi-stage sampling

This method is appropriate when the reference population is large or widely scattered

The selection done in in stage until the final sampling unit(e.g. household, person) are arrived at.

The primary sampling unit (PSU) is the sampling unit in the first sampling stage.

The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.



In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample.

In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample.

If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.

With multi-stage sampling, you still have the benefit of a more concentrated sample for cost reduction.

However, the sample is not as concentrated as other clusters and the sample size is still bigger than for a simple random sample size

Also, you do not need to have a list of all of the units in the population. All you need is a list of clusters and list of the units in the selected clusters.

Admittedly, more information is needed in this type of sample than what is required in cluster sampling.

However, multi-stage sampling still saves a great amount of time and effort by not having to create a list of all the units in a population.

B. Non-probability sampling

In non-probability sampling, every item has an unknown chance of being selected.

In non-probability sampling, there is an assumption that there is an even distribution of a characteristic of interest within the population.

For probability sampling, random is a feature of the selection process.

In non-probability sampling, since elements are chosen arbitrarily, there is no way to estimate the probability of any one element being included in the sample.

Also, no assurance is given that each item has a chance of being included, making it impossible either to estimate sampling variability.

Despite these drawbacks, non-probability sampling methods can be useful when descriptive comments about the sample itself are desired.

Secondly, they are quick, inexpensive and convenient.

The most common types of non-probability sampling

1. Convenience or haphazard sampling
2. Volunteer self selection sampling
3. Judgment/Purposive sampling
4. Quota sampling
5. Snowball sampling technique ... etc

1. Convenience or haphazard sampling

Convenience sampling is sometimes referred to as haphazard or accidental sampling.

It is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

The obvious advantage is that the method is easy to use, but that advantage is greatly offset by the presence of bias.

Often used in face to face interviews

very easy to carry out,

Difficult to draw any meaningful conclusion. May not be representative

2. Volunteer sampling

As the term implies, this type of sampling occurs when people volunteer to be involved in the study.

Sampling voluntary participants as opposed to the general population may introduce strong biases.

Common in trials demanding long duration.

In psychological experiments or pharmaceutical trials (drug testing), for example, it would be difficult and unethical to enlist random participants from the general public.

Payments for subjects some times be involved.

3. Judgment sampling

This approach is used when a sample is taken based on certain judgments about the overall population.

The underlying assumption is that the investigator will select units that are characteristic of the population.

The critical issue here is objectivity: how much can judgment be relied upon to arrive at a typical sample?

Judgment sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.

Since any preconceptions the researcher may reflected in the sample, large biases can be introduced if these preconceptions are inaccurate.

One advantage of judgment sampling is the reduced cost and time involved in acquiring the sample.

4. Quota sampling

This is one of the most common forms of non-probability sampling.

Sampling is done until a specific number of units (quotas) for various sub-populations have been selected.

Since there are no rules as to how these quotas are to be filled, quota sampling is really a means for satisfying sample size objectives for certain sub-populations.

5. Snowball sampling

A technique for selecting a research sample where existing study subjects recruit future subjects among their friends.

Thus the sample group appears to grow like a rolling snowball.

This sampling technique is often used in hidden populations which are difficult for researchers to access; example populations would be drug users or commercial sex workers.

Because sample members are not selected from a sampling frame, snowball samples are subject to numerous biases. For example, people who have many friends are more likely to be recruited into the sample.

Information Gain and Entropy

Entropy measures impurity in the data and information gain measures reduction in impurity in the data.

The feature which has minimum impurity will be considered as the root node.

Information gain is used to decide which feature to split on at each step in building the tree

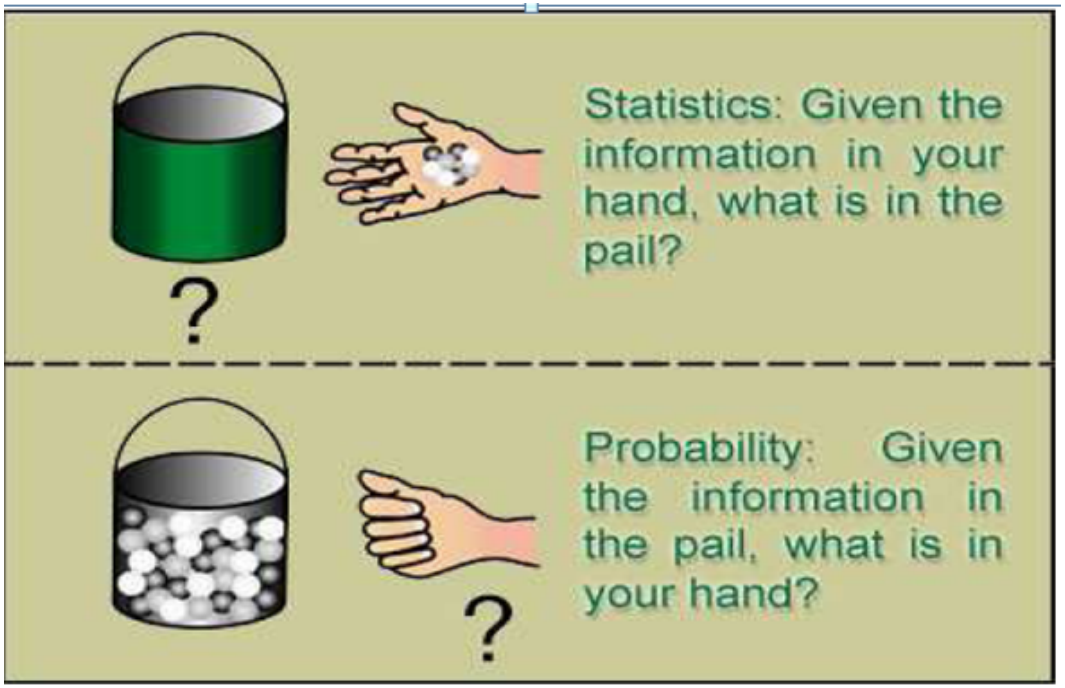
Information gain is the amount of knowledge acquired during a certain decision or action, whereas entropy is a measure of uncertainty

Probability and probability distribution

- Suppose it is known that a specific treatment is effective in 70% of the patients receiving the treatment
- This implies that the population consists of patients for whom the treatment is not effective about (30%) as well as patients for whom the treatment does have an effect is (70%)
- If the treatment is administered to 100 randomly chosen patients, more than 70 may experience improvement, or less than 70
- Question: If 100 patients are given the treatment, what is the probability that less than 60 of them will experience an improvement?
- Probability theory aims at predicting the out-come of an experiment, knowing the population
- These examples suggest the chance of an occurrence of some event of a random variable.

- Statistics : Given the information in your hand what is in the pail ?
- Probability : Given the information in the pail , what is in your hand?

- Probability is the chance of an outcome of an experiment. It is the measure of how likely an outcome is to occur.
- (Random) Experiment:Any process of observation or measurement or any process which generates well defined outcome.



Definition

Random experiment/ random variable: is one in which the out comes occur at random or cannot be predicted with certainty.

e.g. A single coin tossing experiment is a random as the occurrence of Head(H) and Tail(T)

Trial: A physical action , the result of which cannot be predetermined

Sample Space: The set of all possible outcomes of an experiment . In die throwing, S=1,2,3,4,5,6

Events: Collections of basic outcomes from the sample space. We say that an event occurs if any one of the basic outcomes in the event occurs. Any subset of sample space. - Event of getting even number A=2,4,6

Success/ favorable case: Outcome that entail the happening of a desired event.

The term probability refers to the study of randomness and uncertainty

It is the measure of how likely an event or outcome is.

The likelihood of an outcome is quantified by assigning a number from the interval [0, 1] to the outcome (or a percentage from 0 to 100
Equally likely events: If in a random experiment all out comes have equal chance of occurrence. - In tossing coin both H and T have equal chance to occur

Mutually Exclusive Events (Disjoint Events) If the occurrence of one event prevent the occurrence of the other. - In tossing coin the occurrence of Head prevent the occurrence of Tail.

Independent events(mutual independence) The occurrence or non-occurrence of one event doesnt affect the occurrence or non-occurrence of the other event in repeated trials, conduction of a random experiment.

While tossing of two coin simultaneously, the occurrence of head in one coin does not affect the occurrence of tail on the other.

Categories of Probability

Two categories of probabilty are objective and Subjective Probabilities.

Objective probability

1. Classical probability and
2. Relative frequency probability.

Classical Method

If there are n equally likely possibilities, of which one must occur and m are regarded as favorable, or as a success, then the probability of a success is m/n.

$$P(A) = m/n$$

What is the probability of rolling a 6 with a well-balanced die?

Ans. In this case, m=1 and n=6, so that the probability is $1/6 = 0.167$

Relative Frequency Probability

In the long run process ..

The proportion of times the event A occurs in a large number of trials repeated under essentially identical conditions

Definition: If a process is repeated a large number of times (n), and if an event with the characteristic E occurs m times, the relative frequency of E,

Probability of E = P(E) = m/n.

Unions and Intersections of Two Events

Unions of Two Events

If A and B are events, then the union of A and B, denoted by AUB, represents the event composed of all basic outcomes in A or B.

List of all possible outcomes with out reptation

E.g. E (A) = (1,2,3,4,5,6)

E (B) = (5,6,7,8,9,10)

AUB = (1,2,3,4,5,6,7,8,9,10)

Intersections of Two Events

If A and B are events, then the intersection of A and B, denoted by A n B, represents the event composed of all basic outcomes in A and B.

Properties of probability

- Probabilities are real numbers on the interval from 0 to 1; i.e., $0 \leq Pr(A) \leq 1$
- If an event is certain to occur, its probability is 1, and if the event is certain not to occur, its probability is 0.
- If two events are mutually exclusive (disjoint), the probability that one or the other will occur equals the sum of the probabilities; Pr(A or B) = Pr(A) + Pr(B).

  The sum of the probabilities of all mutually exclusive outcomes is equal to 1.

$$P(E_1) + P(E_2) + \ldots + P(E_n) = 1.$$

- If A and B are two events, not necessarily disjoint, then Pr(A or B)= Pr (A) +Pr (B) Pr(A and B).

- The sum of the probabilities that an event will occur and that it will not occur is equal to 1; hence, $P(A^c) = P(notA) = 1P(A)$
- If A and B are two independent events, then
  P(A and B) = P(A)P(B)

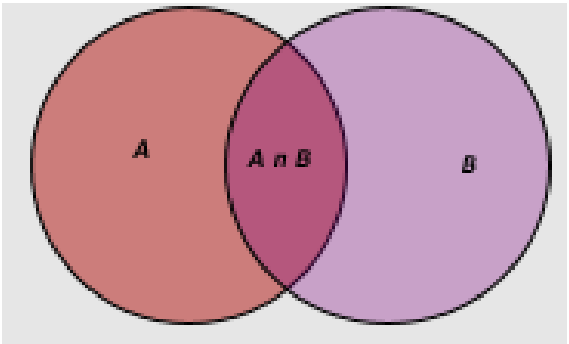
  E (A) = (1,2,3,4,5,6), $P(2^c)$?

  P$P(2^c)$ = 1- P(2) = 1 (1/6) = 0.833

Additive Law of Probability

Let A and B be two events in a sample space S. The probability of the union of A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

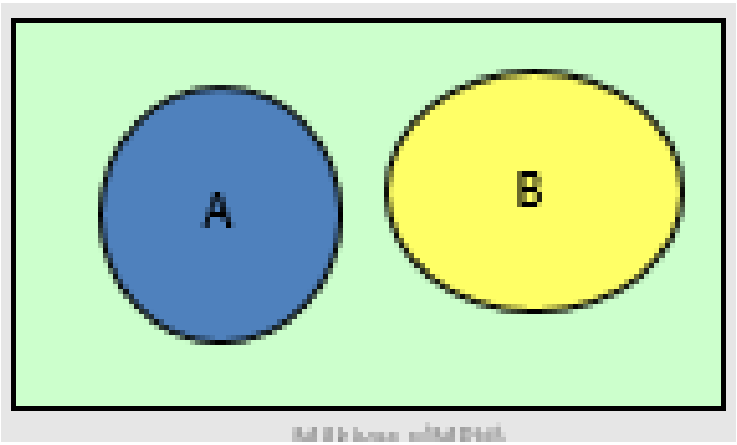


Mutually Exclusive Events

Mutually Exclusive Events: Events that have no basic outcomes in common, or equivalently, their intersection is empty set.

Let A and B be two events in a sample space S. The probability of the union of two mutually exclusive events A and B is:

$$P(A \cup B) = P(A) + P(B) -$$



Independent Events

Two events are independent if the occurrence of one of the events does not affect the probability of the other event.

That is, A and B are independent if :

$$P(B|A) = P(B) \text{ or } P(A|B) = P(A)$$

Example: Let event A stands for the sex of the first child from a mother is female; and event B stands for the sex of the second child from the same mother is female Are A and B independent?

Solution P(B/A) = P(B) = 0.5 The occurrence of A does not affect the probability of B, so the events are independent

Conditional probabilities and the multiplicative law

Sometimes the chance a particular event happens depends on the outcome of some other event. This applies obviously with many events that are spread out in time.

Example: The chance a patient with some disease survives the next year depends on his having survived to the present time. Such probabilities are called conditional.

The notation is Pr(B/A), which is read as the probability of occurrence of event B given that event A has already occurred .

Let A and B be two events of a sample space S. The conditional probability of an event A, given B, denoted by

$$Pr(A/B) = \frac{P(A \cap B)}{P(B)}, \ P(B) \neq 0.$$

Multiplication rule

If A and B are independent events, then

$$P(A \cap B) = P(A) \times P(B)$$

$$P(A|B) = P(A), P(B) \neq 0$$

$$P(B|A) = P(B), P(A) \neq 0$$

P(A and B) denotes the probability that A and B both occur at the same time.

Example

Calculating probability of an event

Table 1: Shows the frequency of cocaine use by sex among adult cocaine users

Life time frequency of cocaine use	Male	Female	Total
1-19 times	32	7	39
20-99 times	18	20	38
more than 100 times	25	9	34
Total	75	36	111

Questions

1. What is the probability of a person randomly picked is a male?
2. What is the probability of a person randomly picked uses cocaine more than 100 times?
3. what is the probability of getting male given that the selected person uses cocaine less than 20 times?
4. Given that the selected person is male, what is the probability of a person randomly picked uses cocaine more than 100 times?
5. Given that the person has used cocaine less than 100 times, what is the probability of being female?

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$

$$= \frac{34}{111} = 0.31.$$

3. $P(M|C < 20) = \frac{P(M \cap C<20)}{P(C<20)} = \frac{32/111}{39/111} = \frac{0.29}{0.35} = 0.83$
4. $P(C > 100|m) = \frac{P(C>100 \cap m)}{P(m)} = \frac{25/111}{75/111} = 0.23/0.68 = 0.34$
5. $P(f|C < 100) = \frac{P(f \cap C<100)}{P(f)} = \frac{27/111}{77/111} = 0.24/0.69 = 0.35$

Random variables and probability distributions

A random variable is a numerical description of the outcomes of the experiment or a numerical valued function defined on sample space, usually denoted by capital letters.

Example: If X is a random variable, then it is a function from the elements of the sample space to the set of real numbers. i.e.

X is a function $X : S \to R$

Usually numbers can be associated with the outcomes of an experiment.

A random variable takes a possible outcome and assigns a number to it.

For example, the number of heads that come up when a coin is tossed four times is 0, 1,2,3 or 4. Sometimes, we may find a situation where the elements of a sample space are categories.

Random variables are of two types

- Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted
  - Number of children in a family.
  - Number of car accidents per week.
  - Number of defective items in a given company.
  - Number of bacteria per two cubic centimeter of water
- Continuous random variable: are variables that can assume all values between any two given values.
  - weight of patients at hospital.
  - Mark of a student.
  - Life time of light bulbs.
  - Length of time required to complete a given training

Probability Distribution

A probability distribution consists of a value of a random variable can assume and the corresponding probabilities of the values.

It is the way data are distributed, in order to draw conclusions about a set of data

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule, or relationship that is called a probability mass (density) function.

Properties of Probability Distribution

1. Since the values of a probability distribution are probabilities, they must be numbers in the interval from 0 to 1.
2. Since a random variable has to take on one of its values, the sum of all the values of a probability distribution must be equal to 1

1. $P(x) \geq 0$, if X is discrete

$f(x) \geq 0$, if X is continuous

2. $\sum_x P(X = x) = 1$, if X is discrete

$\int_x f(x)d(x) = 1$, if X is continuous

Note: If X is a continuous random variable then

$$P(a < X < b) = \int_a^b f(x)dx$$

Probability of a fixed value of a continuous random variable is zero.

Probability means area for continuous random variable.

A. Discrete Probability Distributions

- A discrete probability distribution describes how likely it is to observe specific values for a discrete random variable.
- Suppose X is the random variable sickness absence'

$$X = \begin{cases} 1 & \text{absence due to ilness} \\ 0 & \text{otherwise} \end{cases}$$

- X can only take the values 0 and 1
- The probability distribution of X describes the probability of observing a 0 or a 1, respectively

The following data shows the number of diagnostic services a patient receives

x	0	1	2	3	4	5
P(X=x)	0.671	0.229	0.053	0.031	0.01	0.006

What is the probability that a patient receives exactly 3 diagnostic services?

$$P(X = 3) = 0.031$$

What is the probability that a patient receives at most one diagnostic service?

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$= 0.671 + 0.229 = 0.9$$

What is the probability that a patient receives at least four diagnostic services?

$$P(X \geq 4) = P(X = 4) + P(X = 5)$$

$$0.01 + 0.006 = 0.016$$

- Many frequently used discrete distributions are given a name:
  - Bernoulli distribution
  - Multinomial distribution
  - Binomial distribution
  - Poisson distribution
  - Negative binomial distribution

Binomial Distribution

A binomial probability distribution occurs when the following requirements are met.

- Common probability distributions which is derived from a process known as a Bernoulli trial
- The procedure has a fixed number of trials.
- The trials must be independent.
- Bernoulli trial is random process or experiment which can result in only one of two mutually exclusive outcomes (success or failure)
- The probability of a success(P) remains constant from trial to trial. for each trial

Binomial Distribution

A process that has only two possible outcomes is called a binomial process.

- Let $x_1, x_2 \cdots x_n$ denote outcomes for n independent and identical trials
- The probability distribution of outcome x for Y equals

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

  where $P(x_i = 1) = p$ and $p(x_i = 0) = 1 - p$
- The binomial distribution for $X = \Sigma_i x_i$ has mean and variance respectively
  - n denotes the number of fixed trials
  - **x** denotes the number of successes in the n trials
  - **p** denotes the probability of success
  - **q** denotes the probability of failure **(1-p)**

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Represents the number of ways of selecting x objects out of n where the order of selection does not matter.

where $n! = n(n-1)(n-2)\dots(1)$ and $0! = 1$

Exercise

- Suppose that in a certain malarias area past experience indicates that the probability of a person with a high fever will be positive for malaria is 0.7. Consider 3 randomly selected patients (with high fever) in that same area.

a What is the probability that no patient will be positive for malaria?

b What is the probability that exactly one patient will be positive for malaria?

c What is the probability that exactly two of the patients will be positive for malaria?

d What is the probability that all patients will be positive for malaria?

Poisson Distribution

- The Poisson distribution is used for counts of events that occur randomly over time or space
- When some random events do not result from a fixed number of trials.
- if $x =$ number of deaths due to suicide accidents in Ethiopia during this coming week, there is no fixed upper limit n for x
- Since x must be a non-negative integer, its distribution should place its mass on that range
- A key feature of the Poisson distribution is that its variance equals its mean.
- Its probabilities depend on a single parameter, the mean λ

- The following are some examples which follow a Poisson process
- The number of telephone calls per hour at a switchboard
- The number of e-mails received per hour
- The number of patients admitted in a hospital emergency room per day
- The number of defective items manufactured per 4-hour period in a manufacturing process. ETc

- The Poisson probability mass function is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

- Example: on the average 3-smokers pass a certain street corners every ten minutes, what is the probability that during a given 10 minutes the number of smokers passing will be
  a. Exactly 5?
  b. at most 6?
  c. 7 or more?

Continuous probability distribution

- A continuous probability distribution describes how likely it is that a continuous random variable takes values within certain ranges
- Some frequently used continuous distributions are
  - normal distribution,
  - chi-squared distribution
  - Exponential distribution

Objectives

- Understand basic two sample designs and analysis for continuous data
- Compare means across multiple groups
- Model and explain relationships between a single continuous outcome with a set of independent variables

Two-Sample Inference

Two sample t-test

- Two sample inference
- $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$
- $H_0 : \mu_1 = \mu_2$

Paired Samples: Each data point in one samples is matched and related to a unique data point in the other sample.

Independent Samples: The data points in one sample are unrelated to the data points in the other sample.

Example: Suppose we are interested in studying the association between Oral Contraceptive (OC) use and blood pressure.

Paired samples: identify non OC user women in the child bearing age group and follow them for one year. For those who started using OC within the one year period, compare the blood pressure at baseline and follow-up.

Independent samples: identify a group of OC user women and another group of non users and compare their blood pressures

Paired Samples Arise When

- Having the same set of experimental units receive both treatments (Cross-Over Design)
- Having measurement taken before and after treatment a simplest example of repeated-measures design
- Matching Subjects (Matched-Pair Design)
- Using naturally occurring pairs such as twins or left and right eye of a patient.

Inference of Paired Sample

- $\{(x_{1i}, x_{2i}) : i = 1, \ldots, n\}$, let $\{d_i = x_{1i} - x_{2i} : i = 1, \ldots, n\}$ with sample mean $\bar{d}$ and standard deviation s_d, respectively.
- Assume (check!)$d_i \sim N(\mu_d, \sigma_d^2)$
- Compare μ_1 and μ_2 through $\mu_d = \mu_1 = \mu_2$

$$H_0 : \mu_d = 0 \text{vs} H_1 : \mu_d \neq 0$$

- Statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

- Reject H_0 if $|t| \geq t_{n-1,2}$ (or p-value:$2 \times p(t > |t_{computed}|)$)
- $100(1-\alpha)\%$ CI of μ_d:

$$\bar{d} \pm t_{n-1,1-\alpha/2} \frac{s_d}{\sqrt{n}}$$

- if n is large, then the z-test is used and normality is not needed

Inference of Paired Sample example

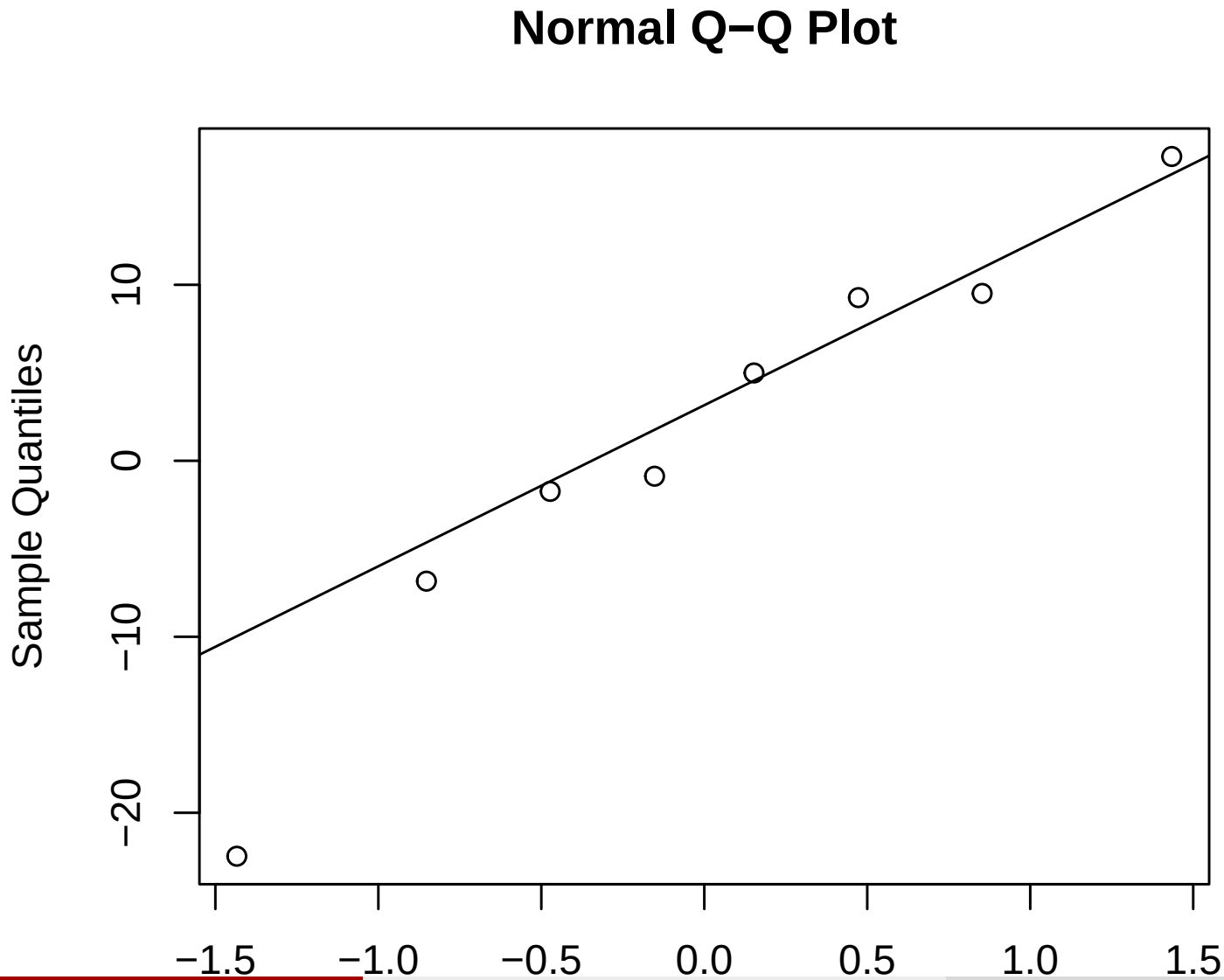
- An important hypothesis in hypertension research is that sodium restriction may lower blood pressure.
- However, it is difficult to achieve sodium restriction over the long term, and dietary counseling in a group setting is sometimes used to achieve this goal.
- The data on overnight urinary sodium excretion (mEq/8hr) was obtained on eight individuals enrolled in a sodium-restricted group.
- Data was collected at baseline and after one week of dietary counseling.

Person	1	2	3	4	5	6	7	8
Baseline	7.85	12.03	21.84	13.94	16.68	41.78	14.97	12.072
Week 1	9.59	34.5	4.55	20.78	11.69	32.51	5.46	12.95
d_i	-1.74	-22.47	17.29	-6.84	4.99	9.27	9.51	-0.88

R code-Examples

```
x=c(7.85, 12.03, 21.84, 13.94, 16.68, 41.78, 14.97, 12.072
y=c(9.59, 34.50, 4.55, 20.78, 11.69, 32.51, 5.46, 12.95)
t.test(x,y, paired = TRUE) # or t.test(x-y)
t = 0.2642, df = 7, p-value = 0.7992
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
-9.074574 11.357574
sample estimates:
mean of x
1.1415
    qqnorm(x-y)
    qqline(x-y)
```

Verify the validity of the normality assumption



Inference of Two Independent Samples

- In making inference about $\mu_1 - \mu_2$, a natural estimator is
- Sampling distribution: $\bar{x}_1 - \bar{x}_2$.

$$\bar{x}_1 - \bar{x}_2 \sim \left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- both X_1 and X_2 are normal
  $\Rightarrow$ N ()is normal
- $\sigma_1 = \sigma_2$ (homogenous)
- $\sigma_1 \neq \sigma_2$ (heterogenous)
- both n_1 and n_2 are large
  $\Rightarrow$ N ()is approximately normal
- otherwise (small n1 or n2 and non-normality): nonparametric procedure

Inference of Two Independent Samples - large sample case

- When n1 and n2 are large, use the pivotal quantity

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

- This is true by CLT whether or not normality or equality of variance hold.
- This quantity is used for tests and confidence intervals

Inference of Two Independent Samples - equal variance

- Hypothesis: $H_0 : \mu_1 = \mu_2 \text{vs} H_1 : \mu_1 \neq \mu_2$
- t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2))}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

where the pooled variance is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}$$

- Reject H_0 if $|t| \geq t_{n_1+n_2-2,1-\alpha/2} p - value : 2 \times p(t > |t_{computed}|)$

- A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$;

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Inference of Two Independent Samples - equal variance

- Inference for difference in means can be computed in R in one of the following two ways depending on how the data is organized.
  If the two samples are entered as vectors x and y then
  t.test(x,y,mu=0,paired=F,var.equal=T,
  alternative="two.sided")
- If the all the data form the two samples is in one vector y and the vector x contains indicators of sample, then we use
  t.test(y x,mu=0,paired=F,var.equal=T,
  alternative="two.sided")

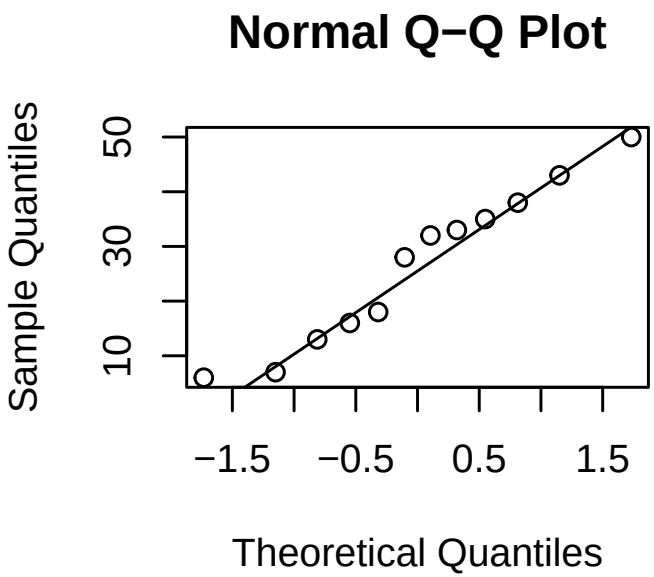
R code-Examples

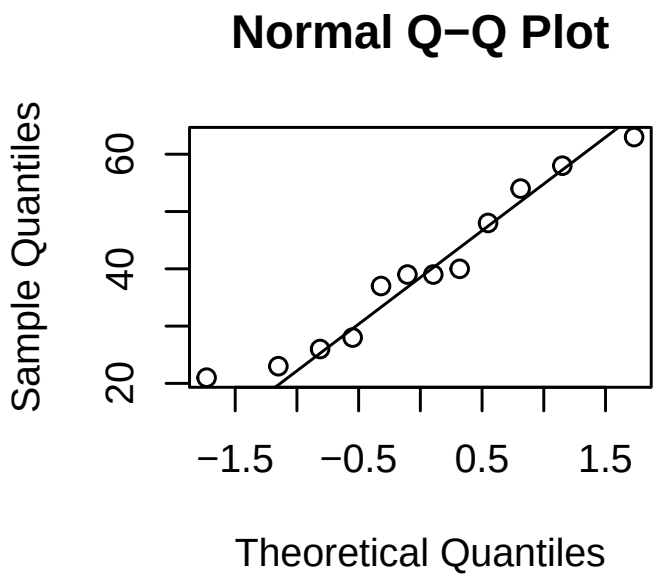
```
x=c(2.3,3.4,1.2,4.4)
y=c(3.2,1.5,2.6,3.3,4.5)
t.test(x,y,var.equal=T)
x=c(1,1,1,1,2,2,2,2,2)
y=c(2.3,3.4,1.2,4.4,3.2,1.5,2.6,3.3,4.5)
t.test(y~x,var.equal=T) # the same results
Welch Two Sample t-test
data: x and y (y by x)
t = -0.2303, df = 5.691, p-value = 0.8259
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
-2.294117 1.904117
sample estimates:
mean in group 1 mean in group 2
2.825 3.020
```

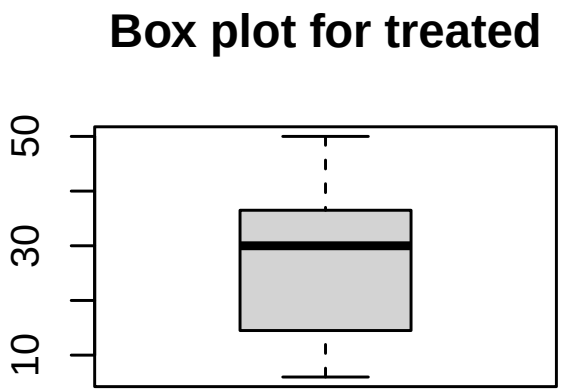
Inference of Two Independent Samples - Example

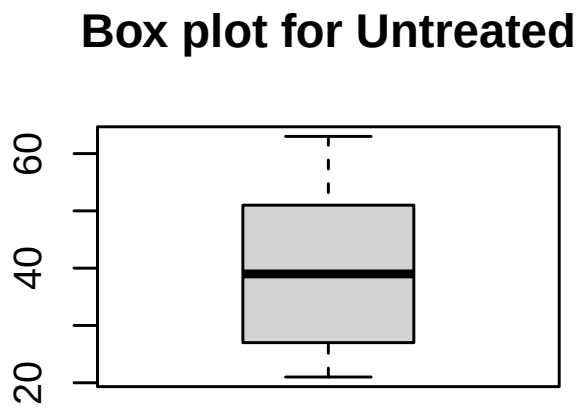
- An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep.
- A random sample of 24 worm-infected lamb of approximately the same age and health was randomly divided into two groups. Twelve of the lambs were injected with the drug and the remaining twelve were left untreated.
- After a 6-month period, the lambs were slaughtered and the following worm counts were recorded:
  Drug Treated: 18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7
  Untreated: 40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39
- Questions:
  (a) Does any of the assumptions of the pooled t-test appear to be an issue?
  (b) Test whether $\mu_{treat} < \mu_{control}$
  (c) Place a 95% CI on $\mu_1 - \mu_2$ to assess the size of the difference in the

Verify the validity of the normality assumption









R code

```
###par(mfrow=c(#rows,#columns))
 treated=c(18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7)
 Untreated=c(40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39)
    par(mfrow=c(2,2))
    qqnorm(treated)
    qqline(treated)
    qqnorm(Untreated)
    qqline(Untreated)
    boxplot(treated,main="Box plot for treated")
    boxplot(Untreated,main="Box plot for Untreated")
```

Inference of Two Independent Samples - Example

```
x=c(18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7)
y=c(40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39)
t.test(x,y,var.equal=T,alternative = "less")
data: x and y
t = -2.2709, df = 21.972, p-value = 0.01665
alternative hypothesis: true difference in means
is less than 0
95 percent confidence interval:
-25.032642 -1.134025
sample estimates:
mean of x mean of y
26.58333 39.66667
```

One-way ANOVA - hypothesis

- We are interested in testing equality of k population means (k = 3)
- data $y_{ij} \sim N(\mu_i, \sigma^2), i = 1, \ldots, k, j = 1, \ldots, n_i$
- Hypotheses:
  $H_0 : \mu_1 = \ldots = \mu_k \text{vs} H_1 : \mu_i \neq \mu_j$ for some $i \neq j$
- Example: k treatment groups in a completely randomized design .

One-way ANOVA - hypothesis

- Mathematical model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

where
μ: the overall mean (unknown constant)
τ_i : the effect of the ith pop./treatment (unknown constants), and $\sum_i \tau_i = 0$
ϵ_{ij} : the error term (a random variable)
Assumptions about ϵ_{ij}: iid $N(0, \sigma^2)$ (indep., normality, equal var.)
In view of this, the hypothesis of ANOVA can be written as
$H_0 : \tau_i = 0$ for all i vs $H_1 : \tau_i \neq 0$ for some i

One-way ANOVA - measuring variabilities

- Decomposition of sum of squares

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$TSS = SSB + SSW$$

- ANOVA-table

Source	SS	df	Mean Square	F
Between Samples	SSB	k-1	$S_B^2 = \frac{SSB}{k-1}$	$\frac{S_B^2}{S_W^2}$
Within Samples	SSW	n-k	$S_w^2 = \frac{SSW}{n-k}$	
Total	TSS	n-1		

- Reject H_0 when

$$F > F_{1-\alpha}(k-1, n-k)$$

One-way ANOVA - example

- 22 young asthmatic volunteers were studied to assess the short-term effects of sulfur-dioxide (SO2) exposure under various conditions.
- Bronchial reactivity to SO2 (cm H2O/s) grouped by lung function is given below.
  Group A (FEV/FVC = 74%) : 20.8, 4.1, 30.0, 24.7, 13.8,
  Group B (FEV/FVC75 - 84%) : 7.5, 7.5, 11.9, 4.5, 3.1, 8.0, 4.7, 28.1, 10. 10.0, 5.1, 2.2
  Group C (FEV/FVC = 85%) : 9.2, 2.0, 2.5, 6.1, 7.5
- Test whether there is an overall mean difference in bronchial reactivity among the three lung-function groups.

R-codes for One-way ANOVA

```
BR<-c(20.8, 4.1, 30.0, 24.7, 13.8, 7.5, 7.5,
11.9, 4.5, 3.1, 8.0, 4.7, 28.1, 10.3, 10.0,
5.1, 2.2, 9.2, 2.0, 2.5, 6.1, 7.5)
LF<-c(rep("A",5),rep("B",12),rep("C",5))
Group<-factor(LF)
fit<-aov(BR~Group)
>anova(fit)
Analysis of Variance Table
Response: BR
Df Sum Sq Mean Sq F value Pr(>F)
Group 2 503.55 251.774 4.9893 0.01813 *
Residuals 19 958.80 50.463
```

One-way ANOVA - assumption assessment

- Independence of y_i's
  - ensured by a careful design or sampling
- Normality
  - not critical when $n_1, \ldots, n_k$ are large, unless under severely skewed pop.
  - tested by Shapiro-Wilks, Anderson-Darlings Tests
- Homogeneity (equal variance in treatment groups)
  - not critical when $n_1, \ldots, n_k$ are nearly equal
  - tested by Bartletts, Hartleys F -max, Levenes and Lehmanns tests
- What if violated
  - data transforming
  - nonparametric method (Kruskal-Wallis test)

One-way ANOVA - multiple comparisons

- If $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is rejected, we would like to know
- in what way the means differ.
- which treatment is the best, second best,...
- if there is an increasing trend in the mean.

One-way ANOVA - LSD

- Least significant difference method for pairwise comparison, i.e. test

$$H_0 : \mu_i = \mu_j \text{vs} \mu_i \neq \mu_j.$$

- statistic

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{s_W^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- reject H_0 if

$$|t| > t_{1-\alpha/2, n-k},$$

note that s_W^2 instead of pooled variance estimate from two groups is used

What is Regression?

- A method by which a quantitative variable is predicted or its variation is explained by means of other quantitative variables.
- The variable being predicted or whose variation being explained is called the **dependent, outcome or response** variable.
- The variables that are used to make the perdition or explain the variation of the dependent variable are called **independent or explanatory** variables.

Types of Regression

- *Linear Regression:* quantitative (interval or ratio scale) dependent variable
  - Simple: one independent variable
  - Multiple: two or more independent variables
- *Logistic Regression:* qualitative (**binary**) dependent variable (yes or no, absent or present, ...)
  - Simple or Multiple
- Many others types of univariate regressions exist.
- *Multivariate Regression:* more than one dependent variable.

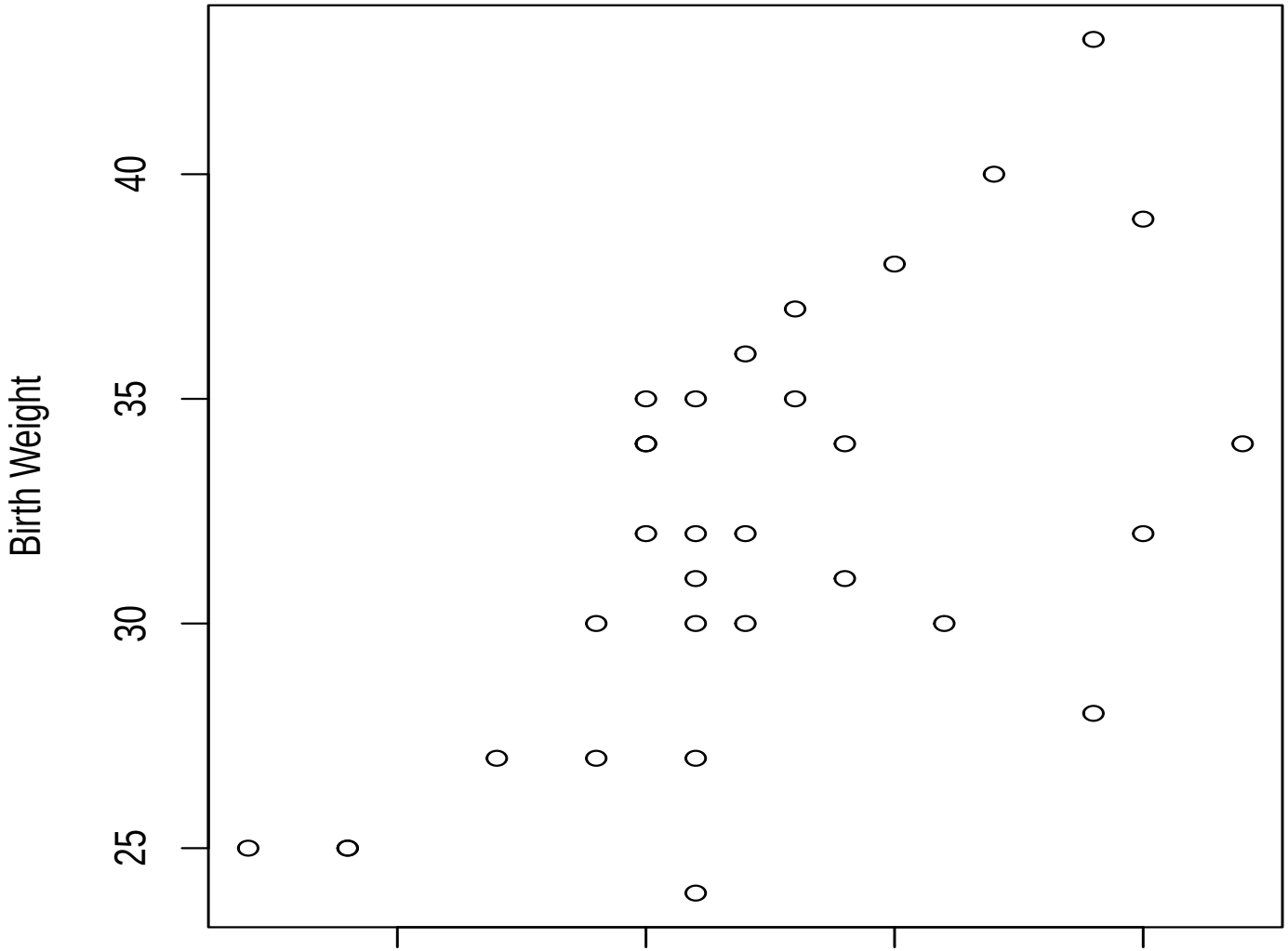
Linear regression - example (Obstetrics)

- Low birth weight is a well-known risk factor for infant mortality and morbidity in the first year of life.
- A study to relate

$$y: \quad \text{birth weight (gram/100)}$$
$$x: \quad \text{estriol level (mg/24 hr) in near-term pregnant women.}$$

- If level of estriol is a predictor of birth weight, then doctors can measure estriol level and decide whether or not to prolong the pregnancy.
- Questions:
  a. Is a linear relationship (a straight line) reasonable model?
  b. Obtain and interpret the estimates of the parameters.
  c. What is the estimated expected birth weight in (grams/100) for a woman that had estriol level 10 mg/24 hr during pregnancy? 13 mg/24 hr? 20 mg/24 hr?

Linear regression - example (Obstetrics)



Linear regression - model

- A simple way to model

$$y \qquad \sim \qquad x_1,\ x_2,\ \ldots,\ x_k$$

dependent var.	indep. var.
outcome	explanatory var.
response	predictors, covariates, factors, etc.
(observed with error)	(conditioned on, by design or observed)

in

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

- Let's first focus on the simple linear regression with $k = 1$.

Linear regression - assumptions

- indep. data points $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$
- homogenous variance $\varepsilon_i \sim \pi(0, \sigma^2)$ for all i
- Normality $\pi = \text{Normal}$ (for testing and CI)
- Linearity in $\beta_0,\ \beta_1,\ \ldots,\ \beta_k$
- sample size $n > k + 1$ (for estimation)

Linear regression - model fit

- *Least squared error (LSE)* criterion: given $\{(y_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n\}$, find a regression line

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$$

such that

$$\sum_i (y_i - \widehat{y}_i)^2 \quad \text{is minimized.}$$

- Essentially, the least square line is the line that passes as closely as possible through all points in the scatter plot of the sample. (depending on the definition of closeness!)

Linear regression - model fit

- In simple linear regression ($y = \beta_0 + \beta_1 x + \varepsilon$),

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}, \quad \widehat{\beta}_1 = S_{xy}/S_{xx},$$

where $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$, $S_{xx} = \sum(x_i - \overline{x})^2$ and

$$\widehat{\sigma}^2 = \frac{\sum(y_i - \widehat{\mathbf{y}}_i)^2}{n-2}.$$

- R function *lm()* does the job.

Linear regression - model fit

- The standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are given by

$$\sigma_{\widehat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} \quad \text{and} \quad \sigma_{\widehat{\beta}_0} = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}.$$

- Therefore, the quality of estimation of $\widehat{\beta}_1$ is influenced by the error variance σ^2 and the amount of variation in x through S_{xx}.
- The ideal situation for estimating β_0 is when $\overline{x} = 0$.

Linear regression - inference

- The hypothesis that x has no predictive value for change in y is formulated as
$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$
- Test statistic
$$t = \frac{\widehat{\beta}_1}{\widehat{\sigma}_{\widehat{\beta}_1}}.$$
- Reject H_0 if
$$|t| \geq t_{1-\alpha/2,n-2}.$$
- $100(1-\alpha)\%$ CI is
$$\widehat{\beta}_1 \pm t_{1-\alpha/2,n-2}\widehat{\sigma}_{\widehat{\beta}_1}.$$
- Similar procedures can be developed for β_0 (by replacing $\widehat{\beta}_1$ by $\widehat{\beta}_0$ in the above formula).

Linear regression - example - R code

```
>bw<-c(25,25,25,27,27,27,24,30,30,31,30,31,30,28,32,
    32,32,32,34,34,34,35,35,34,35,36,37,38,40,39,43)
>estriol<-c(7,9,9,12,14,16,16,14,16,16,17,19,21,24,15,
    16,17,25,27,15,15,15,16,19,18,17,18,20,22,25,24)

>plot(bw~estriol,xlab="Estriol",ylab="Birth Weight")
>rf<-lm(bw~estriol)                  # model fit
>lines(fitted(rf)~estriol)           # add reg. line
>text(12,40,expression(paste("y = 21.52+.608x"))) # annotation
```

Linear regression - example - R code

- R function `summary()` produces the desired output.

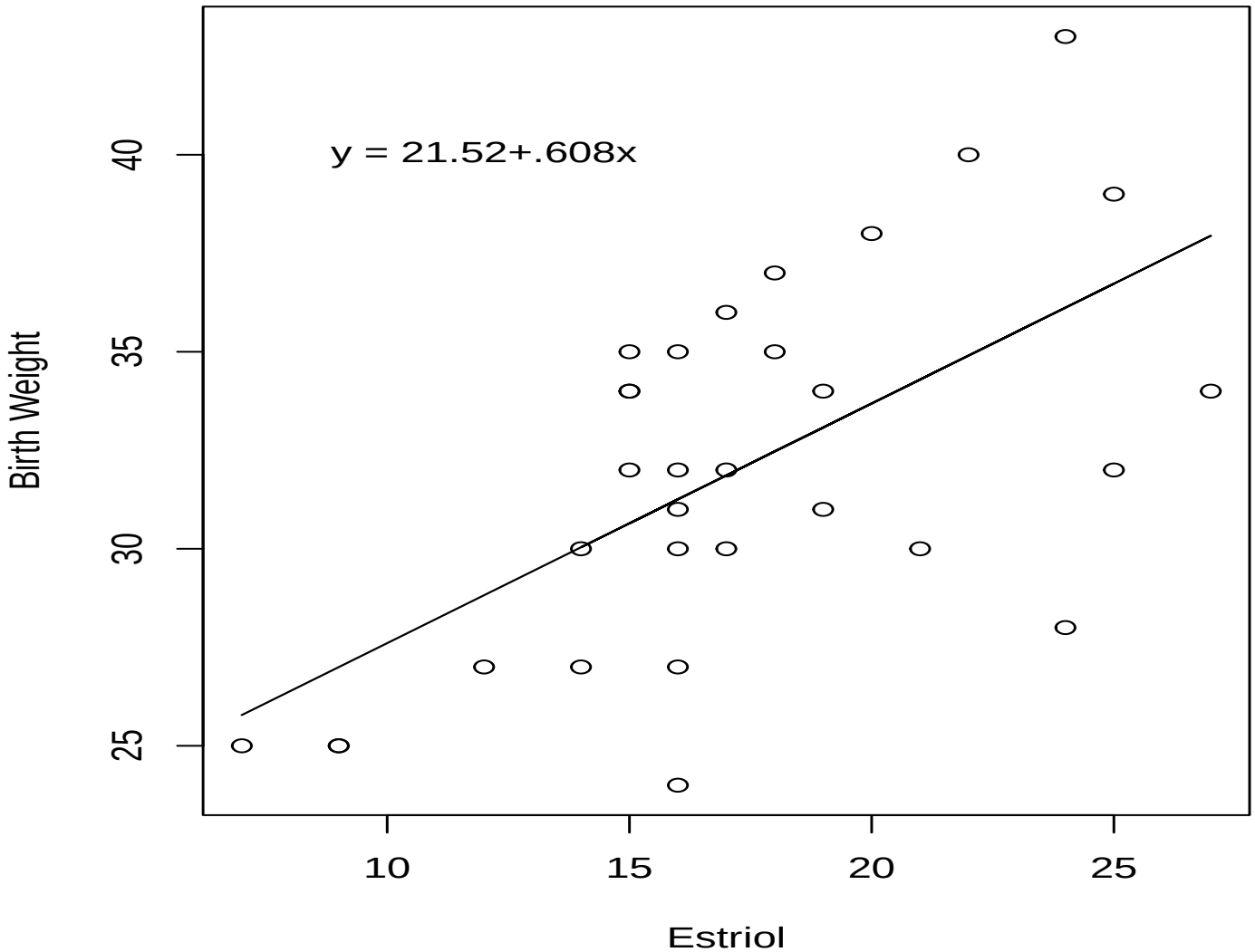
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.5234     2.6204   8.214 4.68e-09 ***
estriol       0.6082     0.1468   4.143 0.000271 ***
```

- Statistical significance in the birthweight-estriol example.
- CI can be generated by R function `confint()`

```
                2.5 %     97.5 %
(Intercept) 16.1640740 26.8827831
estriol      0.3079268  0.9084542
```

- The intervals are rather wide which is not surprising due to the small sample size.

Linear regression - example (Obstetrics)



Linear regression - parameter interpretation

- From $\widehat{\beta}_1 = 0.6082$, we conclude that for a 1mg/24 hr increase in the estriol level, there is an estimated expected increase of 60.82 grams in birth weight.
- Interpreting $\widehat{\beta}_0$ would be **extrapolation**.

Linear regression - ANOVA decomposition

- It can be shown that

$$\sum (y_i - \overline{y})^2 = \sum (\widehat{y}_i - \overline{y})^2 + \sum (y_i - \widehat{y}_i)^2.$$

$$\text{Tot. Obs. Variation} = \text{Var. due to Reg.} + \text{Unexplained Err.}$$

- A concise notation is

$$\text{TSS} = \text{SSR} + \text{SSE}.$$

- Accordingly, the total degrees of freedom are partitioned into

$$df_{\text{Tot}} = df_{\text{Reg}} + df_{\text{Error}}.$$

- In simple linear regression case ($y = \beta_0 + \beta_1 x + \varepsilon$),

$$n - 1 = 1 + (n - 2)$$

Linear regression - significance test

- Mean Squares are defined by dividing sum of squares with the corresponding df. That is,

$$\text{MSR} = \frac{\text{SSR}}{1} \quad \text{and} \quad \text{MSE} = \frac{\text{SSE}}{n-2}.$$

- The two-sided test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ can alternatively be written as to reject H_0 if

$$F = \frac{\text{MSR}}{\text{MSE}} \geq F_{1-\alpha,1,n-2}.$$

- This test is rather more intuitive than the t-test discussed before, but the advantage of the t-test is that it can also be used for a one-sided hypothesis.

Linear regression - example (Obstetrics)

- R function `anova()` produces the ANOVA table as follows.
- ```
  Analysis of Variance Table
  Response: bw
             Df Sum Sq Mean Sq F value    Pr(>F)
  estriol     1 250.57 250.574  17.162 0.0002712 ***
  Residuals 29 423.43  14.601
  ```
- Estriol level has a significant predictive value for the variation in birthweight.
  ```

Linear regression - R^2

- **Coefficient of Determination** (R^2) is the proportion (fraction) of the total variation in the observed responses (y) that can be explained by the simple linear regression on x.

$$R^2 = \frac{\text{SSR}}{\text{TSS}}.$$

- Obviously, $0 \leq R^2 \leq 1$.
- For the birthweight-estriol example, $R^2 = 250.57/(250.57 + 423.43) = 37\%$.
- Therefore, 37% of the variation observed in the birthweight can be explained by the estriol level.

Linear regression - predicting the expected response

- That is, we are interested in estimating $\mu_{y_{n+1}|x_{n+1}}$ for the new value of the independent variable denoted by x_{n+1}. E.g. predicting the mean (expected) birth-weight of all women with a given estriol level.
- A point estimate of $\mu_{y_{n+1}|x_{n+1}}$ is given by,

$$\widehat{\mu}_{y_{n+1}|x_{n+1}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1}$$

and its standard error is

$$\text{SE}(\widehat{\mu}_{y_{n+1}|x_{n+1}}) = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}$$

where the second term in the square root is called **extrapolation factor**.

Linear regression - CI of $\mu_{y_{n+1}|x_{n+1}}$

- Notice that as x_{n+1} gets far away from the other x values, the extrapolation factor gets large making the S.E. large. Hence, the accuracy becomes low.
- A $100(1-\alpha)\%$ CI for $\mu_{y_{n+1}|x_{n+1}}$ is

$$\widehat{\mu}_{y_{n+1}|x_{n+1}} \pm t_{1-\alpha/2,n-2}\widehat{\text{SE}}(\widehat{\mu}_{y_{n+1}|x_{n+1}}).$$

- For the birthweigh-estriol example, for a population of women whose estriol level is 10 mg/24 hr

$$\mu_{y_{11}|x_{11}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{11} = 21.5234 + 0.6082(10) = 27.6053.$$

- We expect babies from these women to weigh $2,760$ grams on average.
- 95% confidence interval is $[2,502,\ 3,019]$ grams.

Linear regression - predicting an individual value

- E.g., an obstetrician would be interested in estimating the birth weight of his patient whose estriol level is 10mg/24 hr.
- Predict y_{n+1} for a new value of x_{n+1} by

$$\widehat{y}_{n+1} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1},$$

  which is the same estimator used for $\mu_{y_{n+1}|x_{n+1}}$, but less accurate.
- In contrast to $\mathrm{SE}(\widehat{\mu}_{y_{n+1}|x_{n+1}})$,

$$\mathrm{SE}(\widehat{y}_{n+1}) = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}.$$

- A $100(1-\alpha)\%$ **prediction interval** is given by:

$$\widehat{y}_{n+1} \pm t_{1-\alpha/2,n-2} \widehat{\mathrm{SE}}(\widehat{y}_{n+1}).$$

Linear regression - example (birthweight)

- For the birthweight-estriol example, 95% prediction interval is $[1,937, 3,584]$ grams.
- There is much more precision in predicting the expected birth weight from a population of women than that of an individual woman's.
- R code:

```
pre<-data.frame(estriol=c(10))
> predict(rf,newdata=pre,interval="confidence",level=0.95)
       fit      lwr      upr
1 27.60533 25.02124 30.18942
> predict(rf,newdata=pre,interval="prediction",level=0.95)
       fit      lwr      upr
1 27.60533 19.37414 35.83652
```

Linear regression - Simultaneous CIs

- Suppose we are interested in simultaneously predicting for more than one value of x (say $x_{n+1}, x_{n+2}, \ldots, x_{n+m}$).
- We use the cut-off point

$$\sqrt{2F_{1-\alpha,2,n-2}}$$

instead of $t_{1-\alpha/2,n-2}$ for our confidence (prediction) intervals to make them simultaneous confidence (prediction) intervals for all values of x.

- The resulting intervals are called *Working-Hotelling* simultaneous confidence (prediction) band.

Linear regression - goodness of fit

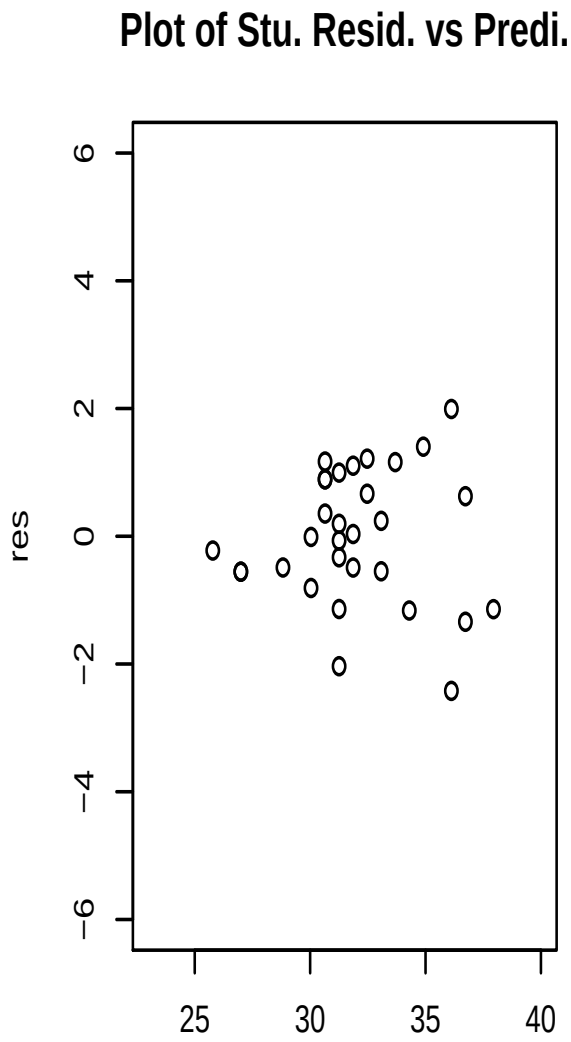
- Residual and studentized residual

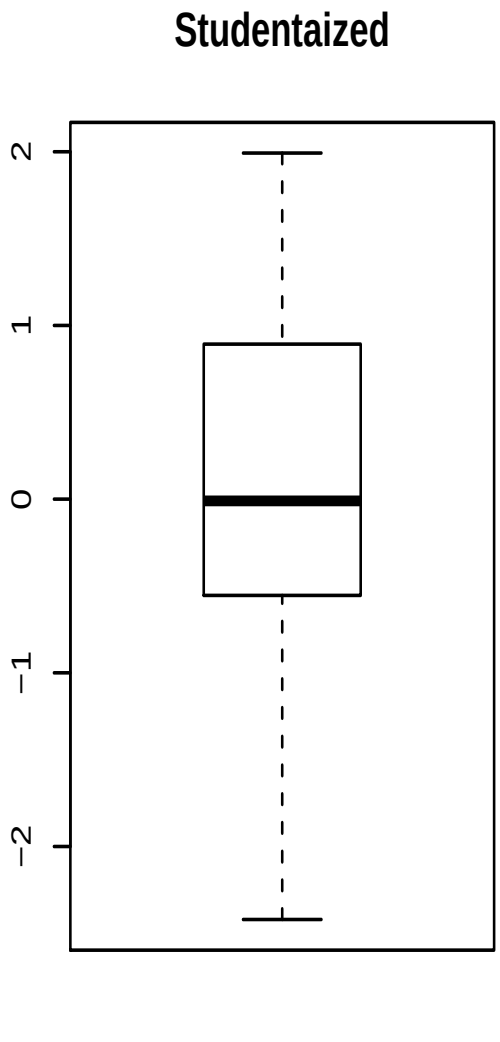
$$e = y - \widehat{\mu}_{y|x}, \quad e/(e)$$

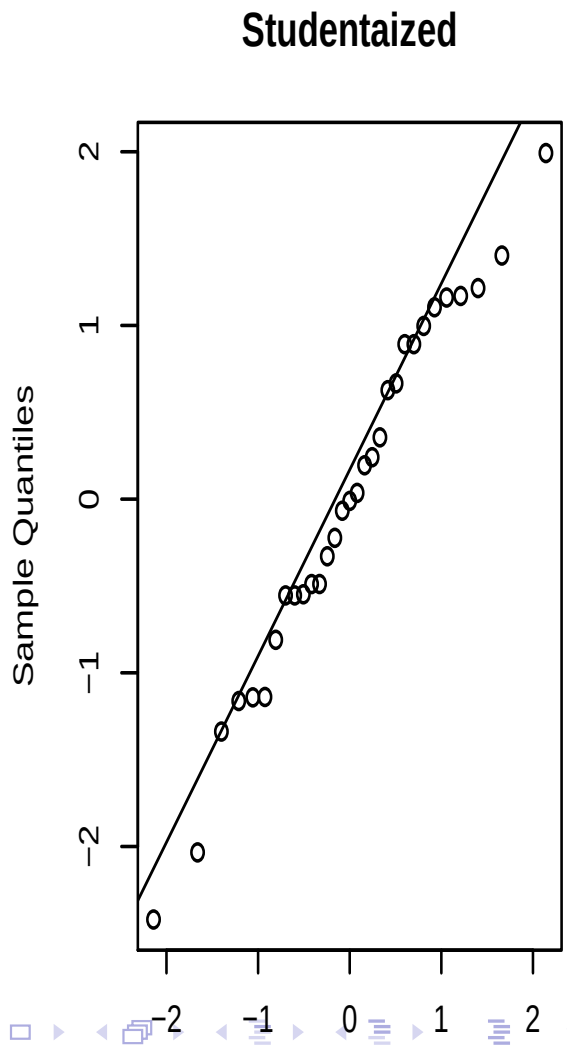
are used to check for violation of the assumptions.

- The *Studentized residuals vs* $\widehat{\mu}_{y|x}$ *plot* is used to check for lack of fit and constant variance.
  - If all the points are contained in a **horizontal band** without any outlier or trend, that means no indication of the violation of linearity and constant variance.
  - If the points cover a **cone shaped region**, that is an indication of the violation of constant variance assumption.
- *Box plot and Q-Q plot of residuals* are used to check normality.
- How do we check the assumption of independence? (Hard)

Linear regression - goodness of fit







Linear regression - variance stablization

- Scatter plot can display violation of the linearity.
- Suppose variance is constant. Only x is transformed.
  1. Scatter plot indicates a relationship that increases (decreases) but at a decreasing rate, try $\log x$, $\sqrt{x}$ or $1/x$.
  2. Scatter plot indicates a relationship that increases (decreases) but at an increasing rate or a parabolic relationship, try

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

- Suppose the variance increases (decreases) with $\widehat{y}$.
  1. Apply $\log y$ if $\sigma^2_{\varepsilon|x} \propto \mu^2_{y|x}$.
  2. Apply $\sqrt{y}$ if $\sigma^2_{\varepsilon|x} \propto \mu_{y|x}$.

What is correlation?

- *Correlation:* is a measure of the strength of *linear* relationship between x and y.
- The stronger the correlation, the better x (linearly) predicts y.
- Correlation between x and y is denoted by r_{xy} and is given by

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

  where $S_{yy} = \sum(y - \overline{y})^2 = \text{TSS}$.
- Always, $-1 \leqslant r_{xy} \leqslant 1$
- Low correlation only means no *linear* relationship.
- R^2 is the square of the correlation coefficient r_{xy}.

Example

- In R, correlation can be computed by the $\mathrm{cor}()$ command.
- For the birthweight-estriol example, $r_{xy} = 0.6097$. There is a moderate correlation.
- In R, this is computed using

```
cor(x=estriol,y=bw, method="pearson")
```

Inference about the Population Correlation Coefficient ρ_{xy}

- *Assumptions:* (x_i, y_i) are independently and identically distributed as normal for $i = 1, \ldots, n$.
- There is no notion of dependent and independent variable.

Test for $H_0 : \rho_{xy} = 0$ vs $H_a : \rho_{xy} \neq 0$

- Test statistic

$$t = \sqrt{n-2}\frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \overset{H_0}{\sim} t_{n-2}.$$

- Reject H_0 if $|t| \geqslant t_{1-\alpha/2,n-2}$.
- The R command $\mathrm{cor.test}$ can do this test.
- One sided test can be constructed in the usual way.
- There is an interesting relationship between $\widehat{\beta}_1$ and r_{xy} given by

$$\widehat{\beta}_1 = r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}.$$

- The t test for correlation and slope give the same answer.

Multiple linear regression

- Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

or

$$y = \boldsymbol{\beta}' \mathbf{x} + \varepsilon$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\mathbf{x} = (1, x_1, \ldots, x_k)'$.

- Assumption: $\varepsilon \sim N(0, \sigma^2)$
- $\beta_1, \ldots, \beta_k$ are known as **partial slopes** or **partial regression coefficients**.

Multiple linear regression - data

- The data looks like,

subj.	x_1	x_2	$\cdots$	x_k	y
1	x_{11}	x_{12}	$\cdots$	x_{1k}	y_1
2	x_{21}	x_{22}	$\cdots$	x_{2k}	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	y_n

- Scatter plot of all the data together is not possible except in some very special cases.

Multiple linear regression - example

- relate SBP with birthweight (oz) and age (days) of 16 infants

```
> data
   id  bw age SBP
1   1 135   3  89
2   2 120   4  90
3   3 100   3  83
4   4 105   2  77
5   5 130   4  92
6   6 125   5  98
.................
16 16 125   3  88
```

Multiple linear regression - Estimation

- Data:

$$\{(y_i, x_{i1}, \ldots, x_{ik}) : i = 1, \ldots, n\}$$

- Write

$$X_{n\times(k+1)} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

- By least square criterion

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}, \quad \widehat{\sigma}^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-(k+1)}, \quad \widehat{y}_i = \widehat{\boldsymbol{\beta}}'\mathbf{x}_i.$$

Multiple linear regression - ANOVA decomposition

- Decomposition of sum of squares

$$\sum (y_i - \overline{y})^2 = \sum (\widehat{\mu}_{y_i|x_1,x_2,\ldots,x_k} - \overline{y})^2 + \sum (y_i - \widehat{\mu}_{y_i|x_1,x_2,\ldots,x_k})^2$$

$$\text{Tot. SS} = \text{SS Reg.} \quad + \quad \text{SS Err.}$$

Source	Sum Sq.	df	Mean Sq.	F
Reg.	SSR	k	$\text{MSR} = \frac{\text{SSR}}{k}$	MSR/MSE
Err.	SSE	$n-k-1$	$\text{MSE} = \frac{\text{SSE}}{n-k-1}$	
Total	TSS	$n-1$		

- Coef. of determination

$$R^2_{y \cdot x_1,x_2,\ldots,x_k} = \frac{\text{SSR}}{\text{TSS}} \geq r^2_{yx_j}$$

for any $j = 1, \ldots, k$ in general.

Multiple linear regression - significance test

- The hypotheses

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1: \quad \text{at least one } \beta_i \text{ is different from zero.}$$

- H_0 means that none of the x-variables has any predictive value.
- Reject H_0 if

$$F = \frac{\text{MSR}}{\text{MSE}} \geq F_{1-\alpha}(k, n-k-1)$$

- Rejection of H_0 tells us that some of the x_js have predictive value, but does not know which ones.

Multiple linear regression - example

```
>res<-lm(SBP~bw+age,data=data)
Call:
lm(formula = SBP ~ bw + age, data = data)
Coefficients:
(Intercept)            bw           age
    53.4502        0.1256        5.8877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.45019    4.53189  11.794 2.57e-08 ***
bw           0.12558    0.03434   3.657   0.0029 **
age          5.88772    0.68021   8.656 9.34e-07 ***
```

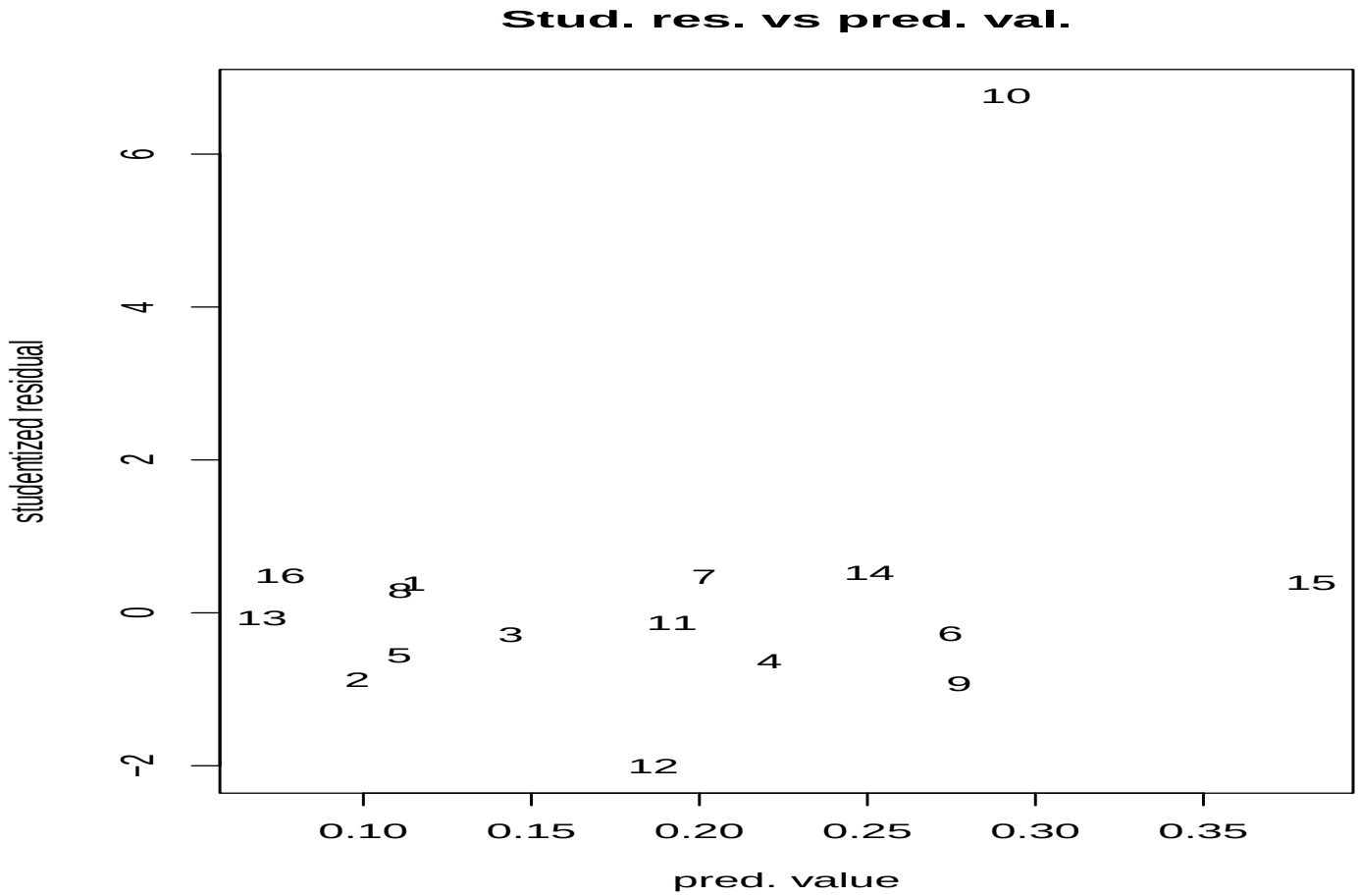
Multiple linear regression - diagnostics

```
>res<-lm(SBP~bw+age,data=data)
>plot(rstudent(res) ~ hatvalues(res),xlab="pred. value",
      ylab="studentized residual",
      main="Stud. res. vs pred. val.", type="n")
>text(hatvalues(res),rstudent(res),1:16)

>influence.measures(res) # evaluation of influential points
```

Multiple linear regression - diagnostics - example

- studentized residual vs predicted values plots



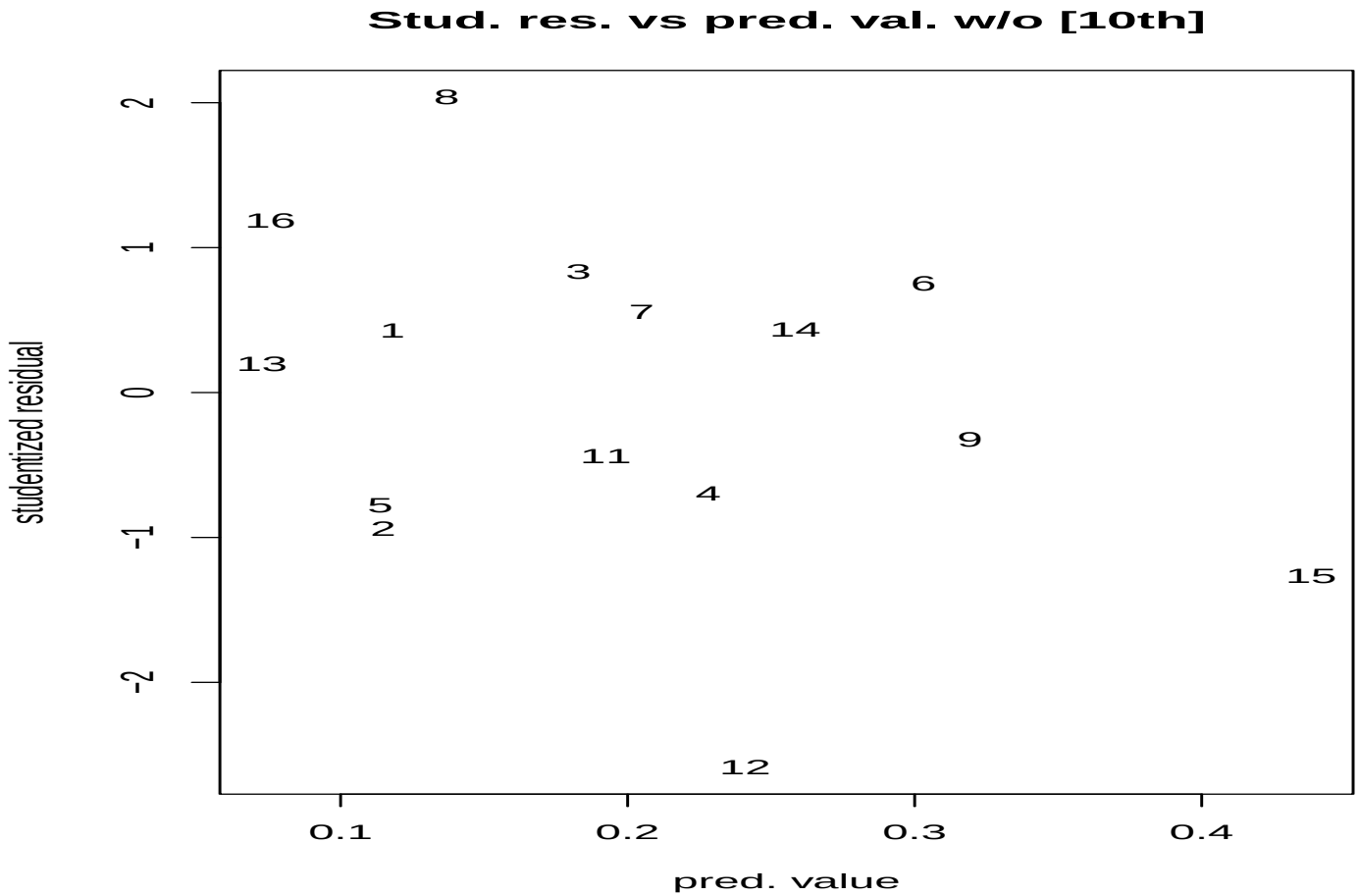
Multiple linear regression - diagnostics - example

- re-do model fit and diagnostics after removing the outlier (**10th obj.**)

```
res1<-lm(SBP~bw+age,data=data[-10,])
plot(rstudent(res1) ~ hatvalues(res1),xlab="pred. value",
     ylab="studentized residual",
     main="Stud. res. vs pred. val. w/o [10th]",type="n")
text(hatvalues(res1),rstudent(res1),seq(16)[-10])
```

Multiple linear regression - diagnostics - example

- studentized residual vs predicted values plots after removing the **10th obj**.



Multiple linear regression - Multicollinearity

- *Multicollinearity* is present when there is high correlation between some covariates or groups of covariates.
- In multiple regression, we are trying to separate the predictive value of each of the independent variables.
- If the correlation between all covariates is 0, which is very unlikely,

$$R^2_{y\cdot x_1,x_2,\ldots,x_k} = r^2_{yx_1} + r^2_{yx_2} + \cdots + r^2_{yx_k}.$$

- In the presence of high multicollinearity, it is difficult to separate the predictive value of the independent variables.

Multiple linear regression - inference about β_j

- The estimated standard error of $\widehat{\beta}_j$ is,

$$\widehat{\mathrm{SE}}(\widehat{\beta}_j) = s_\varepsilon \sqrt{\frac{1}{\sum (x_{ij} - \overline{x}_j)^2 (1 - R^2_{x_j \cdot x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k})}}$$

where $R^2_{x_j \cdot x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k}$ is the coefficient of determination when x_j is regressed on $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$.

- The *variance inflation factor* (VIF), defined by,

$$\mathrm{VIF}_j = \frac{1}{1 - R^2_{x_j \cdot x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k}},$$

is a measure of how much the variance of $\widehat{\beta}_j$ is increased because of collinearity.

Multiple linear regression - inference about β_j

- If $\text{VIF} = 1$, then collinearity is not a problem and if $\text{VIF} > 10$ then collinearity is a serious problem.
- Notice that β_j is the expected change in y associated with a unit change in x_j keeping the other xs constant. If there is collinearity then, it is not possible to keep other xs constant. As a result it is difficult to estimate β_j without reasonable probable error.

Multiple linear regression - partial corr. coef

- The partial correlation between y and x_j,denoted by $r_{x_j y \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}$, is the correlation between y and x_j after the linear effects of the other xs is taken out of both y and x_j.
- It is a measure of the strength of linear relation between y and x_j after eliminating (adjusting for) their linear association with the other xs.
- It is calculated as the correlation between $e_{y \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}$ and $e_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}$ where

$$e_{y \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k} = y - \widehat{\mu}_{y|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k},$$
$$e_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k} = x_j - \widehat{\mu}_{x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}.$$

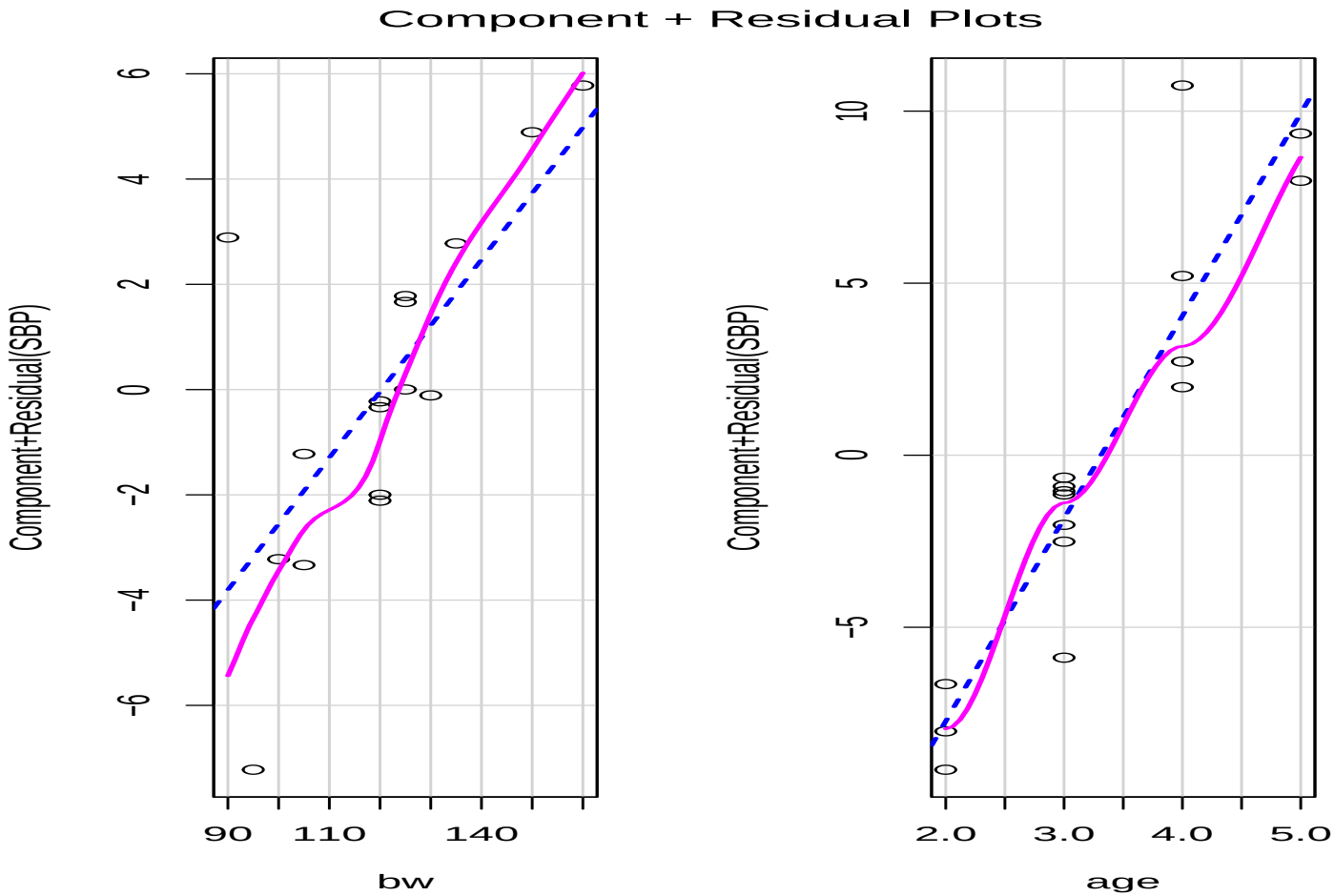
Multiple linear regression - diagnostics

```
>influence.measures(res)     # identify influential points

>library(car)
>crPlots(res)  # partial residual plots, need package {car}
>crPlots(res1) # re-do after removing the outlier (10th obj)
```

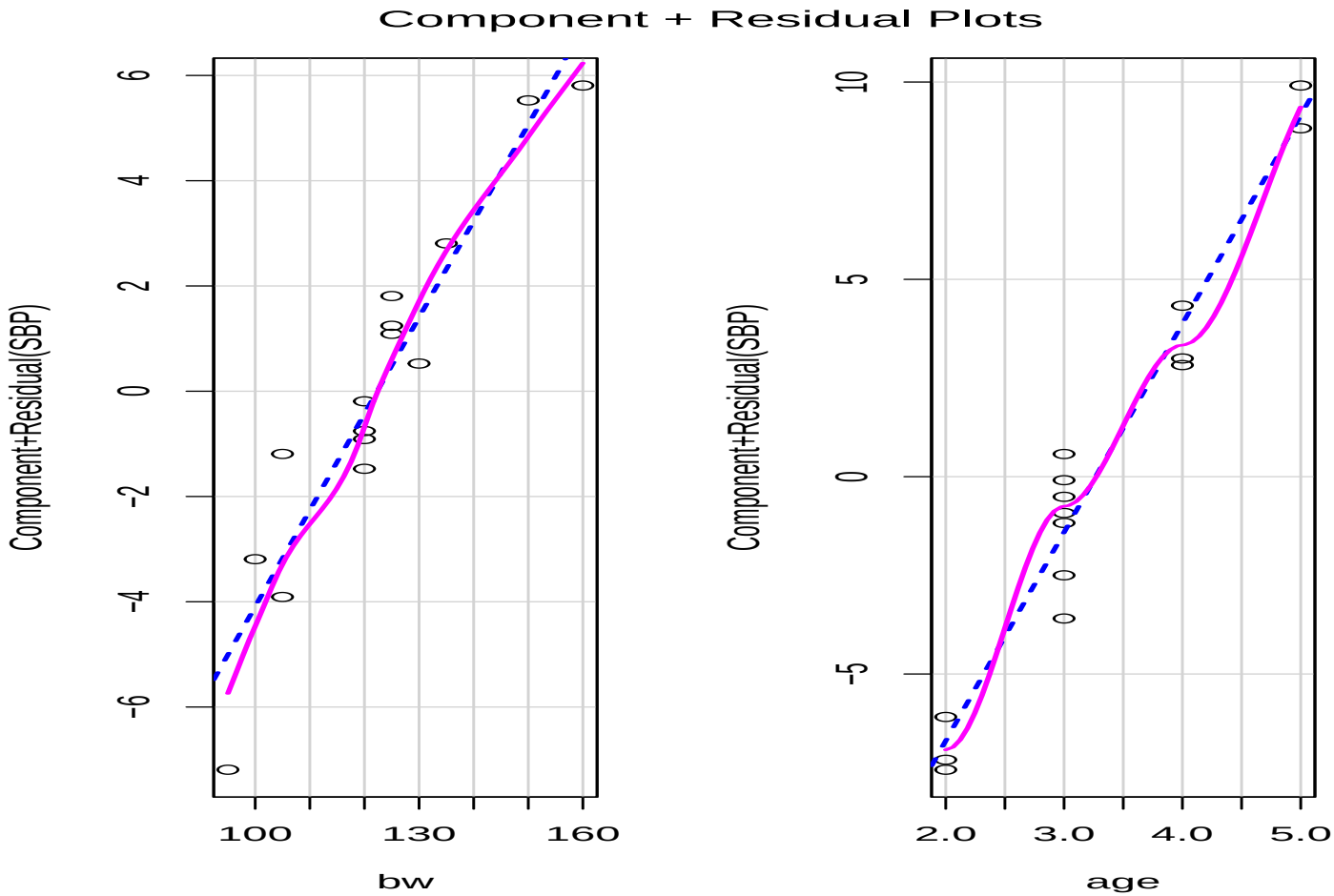
Multiple linear regression - diagnostics

- partial residual plots



Multiple linear regression - diagnostics

- partial residual plots after removing the **10th obj**.



Multiple linear regression - inference of β_j

- Test that x_j has no additional prediction value given the other xs by

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0.$$

- Test statistic

$$t_j = \frac{\widehat{\beta}_j}{\widehat{\text{SE}}(\widehat{\beta}_j)}.$$

- Reject H_0 when

$$|t_j| \geq t_{1-\alpha/2,n-(k+1)}.$$

- A $100(1-\alpha)\%$ CI is

$$\widehat{\beta}_j \pm t_{1-\alpha/2,n-(k+1)}\widehat{\text{SE}}(\widehat{\beta}_j).$$

Multiple linear regression - testing a subset of β_js

- Test the additional predictive value $x_{g+1}, \ldots, x_k$ given $x_1, \ldots, x_g$ through

  $H_0 : \beta_{g+1} = \cdots = \beta_k = 0$ vs H_1 : at least one of $\beta_{g+1}, \ldots, \beta_k$ is not zero.

- Under the full model with all $x_1, \ldots, x_k$,

$$\text{TSS} = \text{SSR}_f + \text{SSE}_f.$$

- Under the reduced model with only $x_1, \ldots, x_g$,

$$\text{TSS} = \text{SSR}_r + \text{SSE}_r.$$

Multiple linear regression - testing a subset of β_js

- The contribution of $x_{g+1}, x_{g+2}, \ldots, x_k$ is

$$\begin{aligned}\text{The extra regression SS} &= \text{SSR}_f - \text{SSR}_r \\ &= \text{SSE}_r - \text{SSE}_f \\ &= \text{The reduction in the error SS}\end{aligned}$$

- This sum of squares has

$$k - g = \{n - (g+1)\} - \{n - (k+1)\}$$

degrees of freedom.

- Therefore the above hypothesis can be tested by

$$F = \frac{(\text{SSR}_f - \text{SSR}_r)/(k-g)}{\text{SSE}_f/\{n-(k+1)\}} = \frac{(\text{SSE}_r - \text{SSE}_f)/(k-g)}{\text{SSE}_f/\{n-(k+1)\}}.$$

Multiple linear regression - testing a subset of β_js

- Reject H_0 if
$$F > F_{1-\alpha,k-g,n-(k+1)}.$$
- Separating out the unique (additional) predictive value of $x_{g+1}, \ldots, x_k$ is possible if $x_1, x_2, \ldots, x_g$ are not highly correlated with $x_{g+1}, \ldots, x_k$.
- The above test includes $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ as a special case.
- Forward, backward, stepwise methods for variable selection.

Multiple linear regression - prediction

- Let $x_{n+1,1}, \ldots, x_{n+1,k}$ new values on the k predictor variables.
- Predict the expected value by

$$\widehat{\mu}_{y_{n+1}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1,1} + \cdots + \widehat{\beta}_k x_{n+1,k},$$

  which is also used to predict y_{n+1}.
- The difference is in the standard errors and, hence, the confidence and prediction intervals are given, respectively, by

$$\widehat{\mu}_{y_{n+1}} \pm t_{1-\alpha/2,n-(k+1)} \widehat{\mathrm{SE}}(\widehat{\mu}_{y_{n+1}})$$

and $$\widehat{y}_{n+1} \pm t_{1-\alpha/2,n-(k+1)} \widehat{\mathrm{SE}}(\widehat{y}_{n+1}).$$

Multiple linear regression - extrapolation

- In simple linear regression, extrapolation occurs when we try to predict y or $\mu_{y|x}$ for a value of x well beyond the range of the data.
- In multiple regression, extrapolation depends not only on the range of each separate x_j predictor but also on the correlation among the xs.
- When making predictions we must consider not only whether each independent variable is reasonable but also whether the combination of the predictor values is reasonable.

Categorical Data

- Data classified into categories (ordered or unordered).
  Methods include
  a. Test for binomial proportions (p_1 vs p_2)
  b. Tests of association and homogeneity
  c. Regression model for binary data
  d. Regression model for count data

p_1 vs p_2

- Assume that $x_1 \sim bin(p_1, n_1)$, x_2 $bin(p_2, n_2)$.
  The hypotheses of interest
  $H_0 : p_1 = p_2 vs H_a : p_1 \neq p_2$ The test to be used depends on whether the samples are independent or paired.
  a. Independent Samples
  i. normal-theory method or contingency table method
  ii. Fishers exact test
  b. Paired Samples (McNemars Test)
  i. normal theory method
  ii. exact test

- A study on risk factors for breast cancer
- Cases: Breast-cancer women in selected hospitals worldwide
- Controls: Non-breast-cancer women of comparable age who were in the hospital at the same time

BC status	age$\geq$ 30	age $\leq$	total
case	683	2537	3220
control	1498	8747	10245

- $H_0 : p_1 = p_2$
  $p_1 =$ P(age at first birth $\geq$ 30—case)
  $p2 =$ P(age at first birth $\geq$ 30—control)

Summary

- The term 'regression model' refers to the case when there are only quantitative predictor variables. When there are only qualitative predictors, the model is called ANOVA Model. A *Linear Model* refers to the situation where both quantitative and qualitative predictors are involved. The term *linear* refers to the way the coefficients enter the model.

y	$x_1, x_2, \ldots$	modeling
continuous	continuous	linear regression
continuous	categorical	ANOVA
continuous	categorical or continuous	LM
categorical	categorical or continuous	logistic/GLM

Review

- References:
  - Bickel, P.J. and Doksum, K.A. (2015). Mathematical Statistics: basic ideas and selected topics. Chapman and Hall/CRC
  - Jin Xu (2016) *Lecture notes for clinical trials*, East China Normal University