

UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor:** Ivo Dinov,
Asst. Prof. In Statistics and Neurology
- **Teaching Assistants:** Tom Daula and Kaiding Zhu,
UCLA Statistics

University of California, Los Angeles, Fall 2002
<http://www.stat.ucla.edu/~dinov/>

Stat 13, UCLA, Ivo Dinov Slide 1

UCLA STAT 10

Introduction to Statistical Reasoning

Course Description, Class homepage, online supplements, VOH's etc.

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

Slide 2 Stat 13, UCLA, Ivo Dinov

What is Statistics? A practical example

- *Demography: Uncertain population forecasts*
by Nico Keilman, Nature 412, 490 - 491 (2001)
- Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But demographers and statisticians have developed methods to calculate probabilistic forecasts.
- The demographic future of any human population is uncertain, but some of the many possible trajectories are more probable than others. So, forecast demographics of a population, e.g., size by 2100, should include two elements: a range of possible outcomes, and a probability attached to that range.

Stat 13, UCLA, Ivo Dinov Slide 3

What is Statistics?

- Together, ranges/probabilities constitute a *prediction interval* for the population. There are trade-offs between greater certainty (higher odds) and better precision (narrower intervals). Why?
- For instance, the next table shows an estimate that the odds are 4 to 1 (an 80% chance) that the world's population, now at 6.1 billion, will be in the range [5.6 : 12.1] billion in the year 2100. Odds of 19 to 1 (a 95% chance) result in a wider interval: [4.3 : 14.4] billion.

Stat 13, UCLA, Ivo Dinov Slide 4

Table 1 Forecasted population sizes and proportions over age 60

| Year | Median world and regional population sizes (millions) | | | | |
|----------------------------------|---|---------------|---------------|---------------|---------------|
| | 2000 | 2025 | 2050 | 2075 | 2100 |
| World total | 6,055 | 7,827 | 8,797 | 9,351 | 8,414 |
| North Africa | 173 | 257 | 311 | 336 | 333 |
| Sub-Saharan Africa | 611 | (228-285) | (249-378) | (238-443) | (215-484) |
| North America | 314 | 976 | 1,319 | 1,522 | 1,500 |
| Latin America | 515 | (856-1,100) | (1,010-1,701) | (1,021-2,194) | (878-2,450) |
| Central Asia | 56 | 379 | 422 | 441 | 454 |
| Middle East | 172 | (351-410) | (358-498) | (343-565) | (313-631) |
| South Asia | 1,367 | 709 | 840 | 904 | 934 |
| China region | 1,408 | (643-775) | (679-1,005) | (647-1,202) | (565-1,383) |
| Pacific Asia | 476 | 81 | 100 | 107 | 106 |
| Pacific OECD | 150 | (73-90) | (80-121) | (76-145) | (66-159) |
| Western Europe | 456 | 285 | 368 | 413 | 413 |
| Eastern Europe | 121 | (282-318) | (301-445) | (290-544) | (259-597) |
| European part of the former USSR | 236 | 1,940 | 2,249 | 2,242 | 1,958 |
| | | (1,735-2,154) | (1,795-2,778) | (1,528-3,085) | (1,186-3,035) |
| | | 1,008 | 1,580 | 1,422 | 1,250 |
| | | (1,494-1,714) | (1,305-1,849) | (1,003-1,884) | (765-1,870) |
| | | 625 | 702 | 702 | 654 |
| | | (569-682) | (575-842) | (509-937) | (410-949) |
| | | 148 | 135 | 135 | 123 |
| | | (144-165) | (125-174) | (107-175) | (79-173) |
| | | 478 | 470 | 433 | 392 |
| | | (445-508) | (399-549) | (321-562) | (257-568) |
| | | 117 | 104 | 87 | 74 |
| | | (109-125) | (89-124) | (61-119) | (44-115) |
| | | 218 | 187 | 159 | 141 |
| | | (203-234) | (154-225) | (110-216) | (85-218) |

Stat 13, UCLA, Ivo Dinov Slide 5

Table 1 Forecasted population sizes and proportions over

| Year | Median | |
|--------------------|--------|---------------|
| | 2000 | 2025 |
| World total | 6,055 | 7,827 |
| North Africa | 173 | (7,219-8,459) |
| Sub-Saharan Africa | 611 | 257 |
| North America | 314 | (228-285) |
| Latin America | 515 | 976 |
| Central Asia | 56 | (856-1,100) |
| Middle East | 172 | 379 |
| | | (351-410) |
| | | (643-775) |
| | | 81 |
| | | (73-90) |
| | | 285 |
| | | (252-318) |

Stat 13, UCLA, Ivo Dinov Slide 5

What is Statistics?

- Demography: Uncertain population forecasts
by Nico Keilman, Nature 412, ,2001
- Traditional population forecasts made by statistical agencies **do not quantify uncertainty**. But lately demographers and statisticians have developed methods to calculate **probabilistic forecasts**.
- Proportion of population over 60yrs.

| | Proportion of population over age 60 | | |
|------|--------------------------------------|-------------|------|
| | 2000 | 2050 | 2100 |
| 0.10 | 0.22 | | 0.34 |
| 0.06 | (0.18-0.27) | (0.25-0.44) | 0.32 |
| 0.05 | (0.15-0.25) | (0.23-0.44) | 0.20 |
| 0.16 | (0.05-0.09) | (0.14-0.27) | 0.40 |
| 0.08 | (0.23-0.37) | (0.28-0.52) | 0.33 |
| 0.08 | (0.17-0.28) | (0.23-0.45) | 0.34 |
| 0.06 | (0.15-0.25) | (0.24-0.46) | 0.35 |
| 0.07 | (0.14-0.23) | (0.24-0.47) | 0.35 |
| 0.10 | (0.14-0.24) | (0.25-0.48) | 0.39 |
| 0.08 | (0.24-0.37) | (0.27-0.53) | 0.36 |
| 0.22 | (0.18-0.29) | (0.26-0.49) | 0.49 |
| 0.20 | (0.32-0.47) | (0.35-0.61) | 0.45 |
| 0.18 | (0.29-0.43) | (0.32-0.58) | 0.42 |
| 0.19 | (0.30-0.46) | (0.28-0.57) | 0.36 |
| | (0.27-0.44) | (0.23-0.50) | |

Stat 13, UCLA, Ivo Dinov Slide 7

What is Statistics?

Stat 13, UCLA, Ivo Dinov Slide 8

What is Statistics?

- There is concern about the **accuracy of population forecasts**, in part because the **rapid fall in fertility in Western countries in the 1970s** came as a surprise. Forecasts made in those years predicted **birth rates** that were up to **80% too high**.
- The rapid reduction in mortality after the Second World War **was also not foreseen**; life-expectancy forecasts were too low by 1-2 years; and the **predicted number of elderly**, particularly the oldest people, was **far too low**.

Stat 13, UCLA, Ivo Dinov Slide 9

What is Statistics?

- So, during the 1990s, researchers developed methods for making **probabilistic population forecasts**, the **aim** of which is to **calculate prediction intervals for every variable of interest**. Examples include population forecasts for the USA, AU, DE, FIN and the Netherlands; these forecasts comprised prediction intervals for **variables** such as **age structure**, **average number of children per woman**, **immigration flow**, **disease epidemics**.
- We need accurate probabilistic population forecasts for the whole world, and its 13 large division regions (see Table). The **conclusion** is that there is an estimated 85% chance that the **world's population will stop growing before 2100**. Accurate?

Stat 13, UCLA, Ivo Dinov Slide 10

What is Statistics?

- There are **three main methods of probabilistic forecasting**: **time-series extrapolation**; **expert judgement**; and **extrapolation of historical forecast errors**.
- Time-series** methods rely on statistical models that are fitted to historical data. These methods, however, seldom give an accurate description of the past. If many of the historical facts remain unexplained, time-series methods result in **excessively wide prediction intervals** when used for **long-term forecasting**.
- Expert judgement** is subjective, and **historic-extrapolation** alone may be near-sighted.

Stat 13, UCLA, Ivo Dinov Slide 11

Chapter 1: What is Statistics?

Chris Wild & George Seber
Textbook

- Polls and surveys** – we're all different; It's impossible or expensive to investigate every single person.
- Experimentation** – sample vs. population
- Observational Studies** – selection and non-response bias
- Statistics -- What is it and who uses it?
- Summary

Stat 13, UCLA, Ivo Dinov Slide 12

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 **shakes** the foundation of the deterministic science, 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!

Slide 13

Stat 13, UCL A, Inv. Dinov

Variation in sample percentages

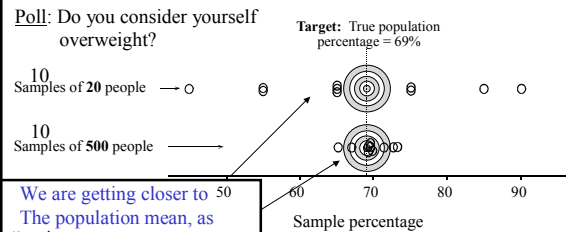


Figure 1.1.1 Comparing percentages from 10 different surveys each of 20 people with those from 10 surveys each of 500 people (all surveys from same population).

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

Slide 14

Stat 13, UCL A, Inv. Dinov

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.

Slide 15

Stat 13, UCL A, Inv. Dinov

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.

Slide 16

Stat 13, UCL A, Inv. Dinov

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.

Slide 17

Stat 13, UCL A, Inv. Dinov

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).

Slide 18

Stat 13, UCL A, Inv. Dinov

Questions ...

- How do the following lead to biases or cause differences in response:
 - *non-response*
 - *self-selection*
 - *question effects*
 - *survey-format effects*
 - *interviewer effects*
 - *transferring findings?*

Slide 20 Stat 13, UCL A, Jon Dinnis

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?)

Slide 21 Stat 13, UCL A, Jon Dinnis

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.

Slide 22 Stat 13, UCL A, Jon Dinnis

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 experimenter :
“Block what you can and randomize what you cannot.”

Slide 23 Stat 13, UCL A, Jon Dinnis

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**.

Slide 24 Stat 13, UCL A, Jon Dinnis

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 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.

Slide 25 Stat 13, UCL A, Jon Dinnis

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.

Slide 26 Stat 13, UCL A, Ivo Dinov

The investigative process

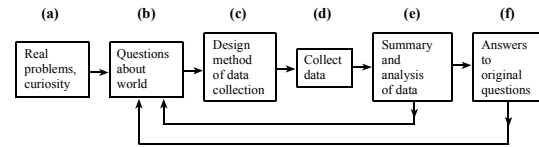


Figure 1.4.1 The investigative process.

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

Slide 27 Stat 13, UCL A, Ivo Dinov

CHAPTER SUMMARY

Slide 28 Stat 13, UCL A, Ivo Dinov

Experiments and observational studies

- When exploring questions of cause and effect we distinguish between **observational studies** and **experiments**.
 - In an **experiment**, the experimenter determines which subjects (experimental units) receive which treatments
 - In an **observational study**, we simply compare subjects that happen to have received each of the treatments.
 - Observational studies widely used for identifying possible causes of effects but cannot reliably establish causation
 - Only properly designed and executed experiments (Section 1.2) can reliably demonstrate causation.

Slide 29 Stat 13, UCL A, Ivo Dinov

The Role of Randomization

Well designed statistical studies employ **randomization** to avoid subjective and other biases.

- Surveys and observational studies should use **random sampling** to obtain representative samples.
- Experiments should use **random assignment of experimental subjects to treatment groups**
 - to ensure comparisons are fair i.e., treatment groups are as similar as possible in every way except for the treatment being used.

Slide 30 Stat 13, UCL A, Ivo Dinov

“Blocking” vs. “stratification”

“Blocking”

- word used in describing an experimental design

“Stratification”

- used in describing a survey or observational study
- Both refer to idea of only making comparisons within relatively similar groups of subjects

Slide 31 Stat 13, UCL A, Ivo Dinov

Blocking and randomization

“Block what you can and randomize what you cannot.”

- **Block** to ensure fair comparisons with respect to factors known to be important
- **Randomize** to try to obtain comparability with respect to unknown factors
- **Randomization** also allows the calculation of how much the estimates made from the study data are likely to be in error

Slide 32

Stat 13, UCLA, Jon Dineen

Sources of error in surveys

- Random sampling leads to **sampling errors**, sampling-size (as we saw for the overweight survey), arising for the choice to use a sample, as opposed to census.
- **Non-sampling errors** can be much larger than the sampling errors. Selection bias, non-response bias, survey/question/interview format are all non-sampling errors.

Slide 33

Stat 13, UCLA, Jon Dineen

Sources of non-sampling errors

- **Selection bias:**
Arises when the population sampled is not exactly the population of interest.
- **Self-selection:**
People themselves decide whether or not to be surveyed. Results akin to severe non-response.
- **Non-response bias:**
Non-respondents often behave or think differently from respondents
 - low response rates can lead to huge biases.

Slide 34

Stat 13, UCLA, Jon Dineen

Non-sampling errors cont.

- **Question-wording effects:**
Even slight differences in question wording can produce measurable differences in how people respond.
- **Interviewer effects:**
Different interviewers asking the same questions can tend to obtain different answers.
- **Survey format effects:**
Factors such as question order, questionnaire layout, self-administered questionnaire or interviewer, can effect the results.

Slide 35

Stat 13, UCLA, Jon Dineen

Dealing with errors

- **Statistical methods** are available for estimating the likely size of **sampling errors**.
- All we can do with **non-sampling errors** is to try to minimize them at the study-design stage.
- **Pilot survey:**
One tests a survey on a relatively small group of people to try to identify any problems with the survey design before conducting the survey proper.

Slide 36

Stat 13, UCLA, Jon Dineen

Jargon describing experiments

- **Control group:**
 - group of experimental units is given no treatment.
 - treatment effect estimated by comparing each treatment group with control group
- **Blinding:**
 - Preventing people involved in experiment from knowing which experimental subjects have received which treatment
 - One may be able to blind
 - subjects themselves
 - people administering the treatments
 - people measuring the results.

Slide 37

Stat 13, UCLA, Jon Dineen

Jargon describing experiments

- **Double blind:**
Both the subjects and those administering the treatments have been blinded.
- **Placebo:**
An inert/dummy/fake treatment.
- **Placebo effect:**
Response caused in human subjects by the idea that they are being treated.

Slide 38 Stat 13, UCLA, Jon Dineen

Immigration Example

- Suppose that you want to set up a nationwide survey about **immigration issues**. Think as precisely as you can about the target population that you would be interested in.
 - Who would you want included?
 - Who would you want excluded?
 - Can you define some rules to characterize your target population?

Slide 39 Stat 13, UCLA, Jon Dineen

Immigration Example

- We could take all members of the population in the US at the time, who were entitled to vote in national elections. This may **exclude** the young, the illegal immigrants, those people in prisons and people legally committed to mental institutions. It would **include** any other permanent residents of the US, whether or not they were citizens, and citizens living overseas.
- You might want to be more, or less, restrictive. In practice, one would probably sample from something like the electoral
- districts [that subset of people who fit the eligibility criteria for voting and who have registered to do so].
- Should the goals of the study influence your survey design (in particular how conservative your selection is)?

Slide 40 Stat 13, UCLA, Jon Dineen

Poll Example

- A survey of High School principals taken after a widespread change in the public school system revealed that 20% of them were under stress-relief medication, and almost 50% had seen a doctor in the past 6 mo.s with stress complains. The survey was compiled from **250 questionnaires returned** out of **2500 sent out**. How **reliable** the results of this experiment are and why?

Slide 41 Stat 13, UCLA, Jon Dineen

Poll Example

- This is only a 10% response rate - the people who responded could be very **unrepresentative**. It could well be that the survey struck a responsive chord with stressed-out principals.

Slide 42 Stat 13, UCLA, Jon Dineen

Experimental vs. Observation study

- A researcher wants to evaluate IQ levels are related to person's height. **100 people** are randomly selected and grouped into **5 bins**: [0:50), [50:100), [100:150), [150:200), [200:250] **cm** in height. The subjects undertook a IQ exam and the results are analyzed.
- Another researcher wants to assess the bleaching effects of **10 laundry detergents** on **3 different colors** (R,G,B). The laundry detergents are randomly selected and applied to 10 pieces of cloth. The discoloration is finally evaluated.

Slide 43 Stat 13, UCLA, Jon Dineen

Experimental vs. Observation study

- For each study, describe what *treatment* is being compared and what *response* is being measured to compare the treatments.
- Which of the studies would be described as *experiments* and which would be described as *observational* studies?
- For the studies that are *observational*, could an experiment have been carried out instead? If not, briefly explain why not.
- For the studies that are *experiments*, briefly discuss what *forms of blinding* would be possible to be used.
- In which of the studies has *blocking* been used? Briefly describe *what* was blocked and why it was blocked.

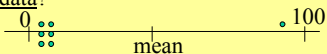
Slide 44 Stat 13, UCL A, Jon Dinnor

Experimental vs. Observation study

- What is the *treatment* and what is the *response*?
 1. *Treatment* is *height* (as a bin). *Response* is *IQ score*.
 2. *Treatment* is *laundry detergent*. *Response* is *discoloration*.
- *Experiment* or *observational* study?
 1. *Observational* – compare obs's (IQ) which happen to have the treatment (height).
 2. *Experimental* – experimenter controls which treatment is applied to which unit.
- For the *observational* studies, can we conduct an experiment?
 1. This *could not* be done as an experiment - it would require the experimenter to decide the (natural) height (treatment) of the subjects (units).
- For the *experiments*, is there *blinding*?
 2. The only form of blinding possible would be for the technicians measuring the cloth discoloration not to know which detergent was applied.
- Is there *blocking*?
 1. & 2. *No blocking*. Say, if there are two laundry machines with different cycles of operation and if we want to block we'll need to randomize which laundry does which cloth/detergent combinations, because differences in laundry cycles are a known source of variation.

Slide 45 Stat 13, UCL A, Jon Dinnor

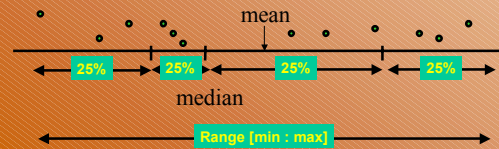
Mean, Median, Mode, Quartiles, 5# summary

- The *sample mean* is the average of all numeric obs's.
 - The *sample median* is the obs. at the index $(n+1)/2$ (note take avg of the 2 obs's in the middle for fractions like 23.5), of the observations ordered by size (small-to-large)?
 - The *sample median* usually preferred to the *sample mean* for *skewed data*?
- 
- Under what circumstances may quoting a *single center* (be it mean or median) not make sense? (*multi-modal*)
 - What can we say about the sample mean of a *qualitative variable*? (*meaningless*)

Slide 46 Stat 13, UCL A, Jon Dinnor

Quartiles

The first quartile (Q_1) is the median of all the observations whose *position* is strictly below the position of the median, and the third quartile (Q_3) is the median of those above.



Slide 47 Stat 13, UCL A, Jon Dinnor

Five number summary

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

Slide 48 Stat 13, UCL A, Jon Dinnor