Nonparametric Statistics
Lecture 0

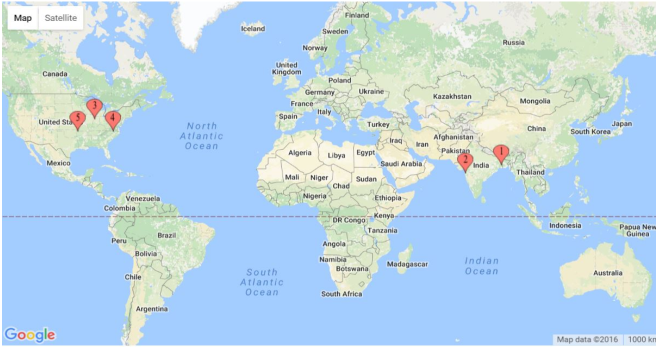
Jyotishka Datta

University of Arkansas, Fayetteville.

August 21, 2017

About me

- Born in Kolkata, India. Went to Indian Statistical Institute for B.Stat & M.Stat.
- Mumbai, India - Barclays Bank, PLC - Modeling Credit Card Default / Fraud.
- Purdue University: Ph.D in Statistics (with Prof. J.K.Ghosh).
- Postdoc at Duke University and SAMSI (with Prof. David Dunson).
- Asst. Prof. in MASC, UofA since last year!



Outline of this course [1]

Nonparametric Statistics

Part I:

1. Introduction
2. Review of Probability Theory
3. One-sample methods
4. Two-sample methods
5. Multiple-sample methods

Part II:

1. Paired comparison
2. Tests for Trends and Association (Quantitative and Qualitative)
3. Nonparametric Bootstrap.
4. Learn how to implement some of these in R.

[1] This is only a tentative list of topics. I may add topics or change the length of time spent on any particular topic to accommodate this class.

How to reach me

- Office: SCEN 219.
- Email: jd033@uark.edu (Subject line: STAT4033 + [Your query]).
- Office hours: Monday 2-3 PM in SCEN 219 and other times by appointment. *These times are subject to change.*

Grades

- **Grade Distribution:** 25% Homework + 20% Midterm 1 + 20% Midterm 2 + 30% Final Exam+ 5% Class participation.
- The grades are assigned by the following rule:

Letter Grade	Percentage
A	90.00 -100%
B	80.00 - 89.99%
C	70.00 - 79.99%
D	60.00 - 69.99%
F	0 - 59.99%

Table: Grade Assignment

Homework Rules

1. Unless otherwise announced, all homework assignments will be posted on Friday afternoon and will be due by Friday 3 PM of the following week.
2. You need to present complete solution for the homework problem and R codes (if any), and not just the final answer, to get full marks for a problem.
3. The lowest homework grade will be dropped. The students are encouraged to discuss among themselves about course materials and concepts but the homework and exam solutions must be your own work.

Some references

- For some of the concepts and homework problems we will follow the book:
  Nonparametric Statistical Inference, Fifth Edition, by Jean Dickinson Gibbons, Subhabrata Chakraborti.
- Other references:
  1. Higgins. Introduction to Modern Nonparametric Statistics.
  2. Hollander and Wolfe. Nonparametric Statistical Methods.
  3. Sprent and Smeeton. Applied Nonparametric Statistical Methods.
  4. Conover, Practical Nonparametric Statistics (old textbook).
- Often times, I will cover material or assign problems outside the book. Hence, you need to follow the class lectures carefully. It is the student's responsibility to take notes of class discussion (if you cannot attend a lecture, please make sure to take the class notes from other students.)

Software

- I will also teach the use of R. For all methods introduced in class, you will learn how to implement those methods using R. You need to install R in your laptop (or the computing resource you use).
- It is available (free) from `http://cran.r-project.org/`.
- Nice interface: R-studio (https://www.rstudio.com/)
- There will be a few lab sessions in SCEN 320.
- Lots of resources on the internet.
  1. ATS UCLA R Page: `http://statistics.ats.ucla.edu/stat/r/`. This is a great resource, and point to a lot of resources too!
  2. Quick R: `http://www.statmethods.net/`
  3. R-Bloggers: `https://www.r-bloggers.com/`
  4. R for data-science: `http://r4ds.had.co.nz/index.html`. (Great if you are interested in data science!)

Parametric and Nonparametric Statistics

- **Parameter** A constant (usually unspecified) that characterizes a population. Can be interpreted in a broad sense to include almost all descriptions of population characteristics.
- **Statistic** A function of random variables or observations that does not depend on the unknown parameters.
- **Inference** Sample descriptions (Statistics) used to infer something about the population - mainly two types of inference: Estimation & Tests of hypotheses.

Parametric and Nonparametric Statistics

- **Parameter** A constant (usually unspecified) that characterizes a population. Can be interpreted in a broad sense to include almost all descriptions of population characteristics.
- **Statistic** A function of random variables or observations that does not depend on the unknown parameters.
- **Inference** Sample descriptions (Statistics) used to infer something about the population - mainly two types of inference: Estimation & Tests of hypotheses.
- What makes Statistical inference *scientific* is the ability to make statement about accuracy or reliability.
- Some statistics are more popular than other, e.g. sample mean vs. median. Why?

Parametric and Nonparametric Statistics

- **Parameter** A constant (usually unspecified) that characterizes a population. Can be interpreted in a broad sense to include almost all descriptions of population characteristics.
- **Statistic** A function of random variables or observations that does not depend on the unknown parameters.
- **Inference** Sample descriptions (Statistics) used to infer something about the population - mainly two types of inference: Estimation & Tests of hypotheses.
- What makes Statistical inference *scientific* is the ability to make statement about accuracy or reliability.
- Some statistics are more popular than other, e.g. sample mean vs. median. Why?
- Central Limit Theorem, sample mean from a large sample is almost normally distributed. We'll come back to this later.

Central Limit Theorem

- This theorem is the reason we see Normal distributions everywhere!!
- The unifying mathematical result is one of the most important results in all of mathematics, and is called the central limit theorem.
- The sum of N independent, identically distributed random variables is Normally distributed.
- Example: if one flips a coin many times, the probability of getting a given number of heads should follow a normal curve, with mean equal to half the total number of flips.

IN FACT, DEMOIVRE'S DISCOVERY ABOUT THE BINOMIAL IS A SPECIAL CASE OF AN EVEN ***MORE*** GENERAL RESULT, WHICH HELPS EXPLAIN WHY THE NORMAL IS SO IMPORTANT AND WIDESPREAD IN NATURE. IT IS THIS:

"Fuzzy Central Limit Theorem": DATA THAT ARE INFLUENCED BY ***MANY SMALL AND UNRELATED RANDOM EFFECTS*** ARE APPROXIMATELY ***NORMALLY DISTRIBUTED.***

THIS EXPLAINS WHY THE NORMAL IS ***EVERYWHERE:*** STOCK MARKET FLUCTUATIONS, STUDENT WEIGHTS, YEARLY TEMPERATURE AVERAGES, S.A.T. SCORES: ALL ARE THE RESULT OF MANY DIFFERENT EFFECTS. FOR EXAMPLE, A STUDENT'S WEIGHT IS THE RESULT OF GENETICS, NUTRITION, ILLNESS, AND LAST NIGHT'S BEER PARTY. WHEN YOU PUT THEM ALL TOGETHER, YOU GET THE NORMAL! (REMEMBER, THE BINOMIAL IS THE RESULT OF n INDEPENDENT BERNOULLI TRIALS.)

NOW, BACK TO THE MATH....

83

Parametric Inference

- Parametric Inference: Estimation and Inference are based on some assumption on the form of the distribution.
- Example 1: (modified from G & C: 4.13) According to test theory, scores on a certain IQ test are normally distributed ($X_i \sim \mathcal{N}(\mu, \sigma^2 = 100)$. This test was given to 18 girls of similar age and their scores were: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108.
- **Question**: Is the mean IQ significantly greater than 100? How would you test?
- **Hypothesis testing**: $H_0 : \mu = 100$, $H_A : \mu > 100$.
- We carry out a *z*-test[2] based on *Normality assumption*.

[2] if you don't know what these are, don't worry. I will cover it later.

Parametric Inference

- Parametric Inference: Estimation and Inference are based on some assumption on the form of the distribution.
- Example 1: (modified from G & C: 4.13) According to test theory, scores on a certain IQ test are normally distributed ($X_i \sim \mathcal{N}(\mu, \sigma^2 = 100)$. This test was given to 18 girls of similar age and their scores were: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108.
- **Question**: Is the mean IQ significantly greater than 100? How would you test?
- **Hypothesis testing**: $H_0 : \mu = 100$, $H_A : \mu > 100$.
- We carry out a z-test[2] based on *Normality assumption*.
- The conclusions are valid as long as the assumptions are valid.

[2] if you don't know what these are, don't worry. I will cover it later.

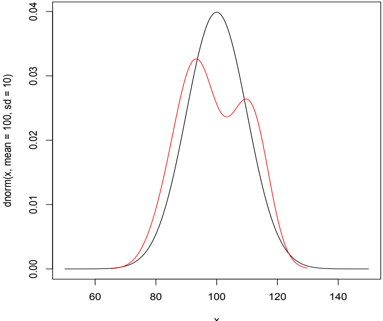
Nonparametric Statistics

In reality the assumption of normality is often not valid.
This leads us to two questions:

- Q.1: How do we test if the assumption of normality is valid?
- Q.1.a: On a more basic level, are the numbers random?

We can visually inspect the deviation from normality.

```
x = seq(50,150,length.out=1000)
plot(x,dnorm(x, mean = 100, sd = 10),type="l")
IQ = c(114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108)
lines(density(IQ),col="red")
```



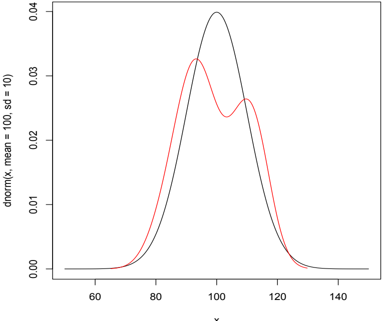
Nonparametric Statistics

In reality the assumption of normality is often not valid.
This leads us to two questions:

- Q.1: How do we test if the assumption of normality is valid?
- Q.1.a: On a more basic level, are the numbers random?
- Q.2: How do we test the hypothesis without the normality assumptions?
- Q.2 b: Can we make other, weaker assumptions? (Symmetry)

We can visually inspect the deviation from normality.

```
x = seq(50,150,length.out=1000)
plot(x,dnorm(x, mean = 100, sd =
10),type="l")
IQ = c(114, 81, 87, 114, 113, 87,
111, 89, 93, 108, 99, 93, 100, 95,
93, 95, 106, 108)
lines(density(IQ),col="red")
```



Example: Medical Studies

In medical studies the progress of patients is often monitored for a limited time after treatment; often anything from a few months to 5 or 6 years [3]. Dinse (1982) gives data for survival times in weeks for 10 patients with symptomatic lymphocytic non-Hodgkin's lymphoma.

Survival times in weeks were:

49, 58, 75, 110, 112, 132, 151, 276, 281, 362*

The * denotes a censored observation. It means that the precise survival time is not known for one patient who was alive after 362 weeks.

This is an example from **Survival Analysis** - key tool across many fields.

[3] Example from Sprent & Smeeton's book

Example: Medical Studies

In medical studies the progress of patients is often monitored for a limited time after treatment; often anything from a few months to 5 or 6 years [3]. Dinse (1982) gives data for survival times in weeks for 10 patients with symptomatic lymphocytic non-Hodgkin's lymphoma.
Survival times in weeks were:
49, 58, 75, 110, 112, 132, 151, 276, 281, 362*
The * denotes a censored observation. It means that the precise survival time is not known for one patient who was alive after 362 weeks.
This is an example from **Survival Analysis** - key tool across many fields.
Questions:

1. **Is it reasonable to suppose that these data are consistent with a median survival time of 200 weeks?**
2. **is the median survival time between 80 and 275 weeks?** [Confidence Interval]

Now let me digress for a couple of minutes!

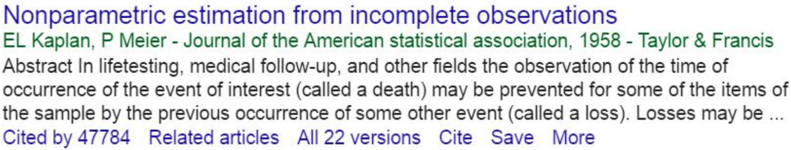
[3] Example from Sprent & Smeeton's book

Survival Analysis

- **Survival Analysis**: A set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest.
  Events: The event can be death, occurrence of a disease, failure of an equipment, or marriage etc.
- Literally anything where we can ask: **How long does it take to . . .?**

Survival Analysis

- **Survival Analysis**: A set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest.
  Events: The event can be death, occurrence of a disease, failure of an equipment, or marriage etc.
- Literally anything where we can ask: **How long does it take to ...?**
- Highest cited paper in all of Statistics:



Nonparametric estimation from incomplete observations

EL Kaplan, P Meier - Journal of the American statistical association, 1958 - Taylor & Francis

Abstract In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be ...

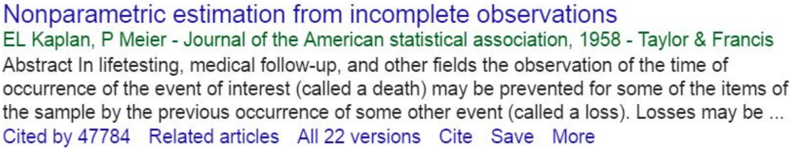
Cited by 47784 Related articles All 22 versions Cite Save More

Figure: Kaplan-Meier

Survival Analysis

- **Survival Analysis**: A set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest.

  Events: The event can be death, occurrence of a disease, failure of an equipment, or marriage etc.
- Literally anything where we can ask: **How long does it take to . . .?**
- Highest cited paper in all of Statistics:



Nonparametric estimation from incomplete observations

EL Kaplan, P Meier - Journal of the American statistical association, 1958 - Taylor & Francis

Abstract In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be ...

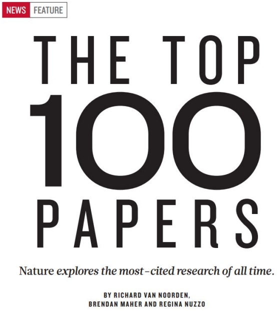
Cited by 47784   Related articles   All 22 versions   Cite   Save   More

Figure: Kaplan-Meier

- .. and, it's a nonparametric method !!

Survival Analysis

… I naturally wondered, what's the second most cited article?



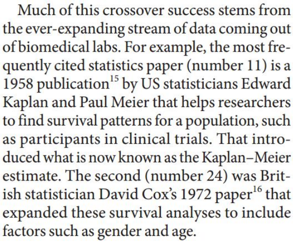
NEWS FEATURE

THE TOP 100 PAPERS

Nature explores the most-cited research of all time.

BY RICHARD VAN NOORDEN, BRENDAN MAHER AND REGINA NUZZO

(a) Top 100 most cited paper



Much of this crossover success stems from the ever-expanding stream of data coming out of biomedical labs. For example, the most frequently cited statistics paper (number 11) is a 1958 publication[15] by US statisticians Edward Kaplan and Paul Meier that helps researchers to find survival patterns for a population, such as participants in clinical trials. That introduced what is now known as the Kaplan–Meier estimate. The second (number 24) was British statistician David Cox's 1972 paper[16] that expanded these survival analyses to include factors such as gender and age.

(b) Cox's proportional hazard paper

- Later in the course, we will learn a method called "the Bootstrap".
- There would be at least 10 papers in this list, that use Bootstrap or a version of it, applied to Biological data. The most well-known is:
- *Confidence limits on phylogenies: an approach using the bootstrap, Felsenstein (1985). Cited by 30793.*

Survival Data (continued)

- Is it reasonable to assume that time till failure (survival data) is normally distributed?
- Can you give me an intuitive reason?

Survival Data (continued)

- Is it reasonable to assume that time till failure (survival data) is normally distributed?
- Can you give me an intuitive reason?
- Hint: Should time till failure have a bell-shaped distribution?

Survival Data (continued)

- Is it reasonable to assume that time till failure (survival data) is normally distributed?
- Can you give me an intuitive reason?
- Hint: Should time till failure have a bell-shaped distribution?
- Survival times are usually right-skewed and censored !

Survival Data (continued)

- Is it reasonable to assume that time till failure (survival data) is normally distributed?
- Can you give me an intuitive reason?
- Hint: Should time till failure have a bell-shaped distribution?
- Survival times are usually right-skewed and censored !
- There are many parametric models for fitting survival times.
- However, the conclusions from parametric models are valid as long as the assumptions are, and you need special care to handle censored observations.

Survival Data (continued)

- Is it reasonable to assume that time till failure (survival data) is normally distributed?
- Can you give me an intuitive reason?
- Hint: Should time till failure have a bell-shaped distribution?
- Survival times are usually right-skewed and censored !
- There are many parametric models for fitting survival times.
- However, the conclusions from parametric models are valid as long as the assumptions are, and you need special care to handle censored observations.
- There are simple non-parametric methods to test the above (and it isn't affected by censoring).

Nonparametric Statistics

- **“Nonparametric" is a misnomer. More accurate term: Distribution-free Statistics.**
- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.

Nonparametric Statistics

- **“Nonparametric" is a misnomer. More accurate term: Distribution-free Statistics.**

- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.
- Often we only assume that the random variables are independent and identically distributed (IID).

Nonparametric Statistics

- **“Nonparametric'' is a misnomer. More accurate term: Distribution-free Statistics.**
- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.
- Often we only assume that the random variables are independent and identically distributed (IID).
- Test and estimation procedures require relatively fewer assumptions about the population distribution.

Nonparametric Statistics

- **“Nonparametric" is a misnomer. More accurate term: Distribution-free Statistics.**

- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.
- Often we only assume that the random variables are independent and identically distributed (IID).
- Test and estimation procedures require relatively fewer assumptions about the population distribution.
- IQ Example:The IQ scores of 18 girls of similar age were: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108. **We do not assume the scores are normal, but we assume the scores are IID.**

Nonparametric Statistics

- **“Nonparametric" is a misnomer. More accurate term: Distribution-free Statistics.**
- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.
- Often we only assume that the random variables are independent and identically distributed (IID).
- Test and estimation procedures require relatively fewer assumptions about the population distribution.
- IQ Example:The IQ scores of 18 girls of similar age were: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108. **We do not assume the scores are normal, but we assume the scores are IID.**
- **Changed Question**: Is the ~~mean~~ median IQ significantly greater than 100?

Nonparametric Statistics

- **"Nonparametric" is a misnomer. More accurate term: Distribution-free Statistics.**

- We will estimate and test hypotheses about parameters; but the form of the distribution is not assumed.
- Often we only assume that the random variables are independent and identically distributed (IID).
- Test and estimation procedures require relatively fewer assumptions about the population distribution.
- IQ Example:The IQ scores of 18 girls of similar age were: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108. **We do not assume the scores are normal, but we assume the scores are IID.**
- **Changed Question**: Is the ~~mean~~ median IQ significantly greater than 100?
- We can use One-sample tests (first module).

Why study Nonparametric Statistics?

- In many applications, there is no prior knowledge of the underlying distributions.
- If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
- For studies with small sample size $(n \leq 30)$ normal approximation does not work well.

Why study Nonparametric Statistics?

- In many applications, there is no prior knowledge of the underlying distributions.
- If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
- For studies with small sample size $(n \leq 30)$ normal approximation does not work well.
- Therefore, we need statistical methods that require very little model/distributional assumptions;

Why study Nonparametric Statistics?

- In many applications, there is no prior knowledge of the underlying distributions.
- If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
- For studies with small sample size $(n \leq 30)$ normal approximation does not work well.
- Therefore, we need statistical methods that require very little model/distributional assumptions;
- or those that are robust/insensitive to the model/distributional assumptions, and

Why study Nonparametric Statistics?

- In many applications, there is no prior knowledge of the underlying distributions.
- If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
- For studies with small sample size $(n \leq 30)$ normal approximation does not work well.
- Therefore, we need statistical methods that require very little model/distributional assumptions;
- or those that are robust/insensitive to the model/distributional assumptions, and
- insensitive to outliers in the data.

Why study Nonparametric Statistics?

- In many applications, there is no prior knowledge of the underlying distributions.
- If the parametric assumptions are violated, the use of parametric test procedures can give misleading or wrong results.
- For studies with small sample size $(n \leq 30)$ normal approximation does not work well.
- Therefore, we need statistical methods that require very little model/distributional assumptions;
- or those that are robust/insensitive to the model/distributional assumptions, and
- insensitive to outliers in the data.

Nonparametric Statistics offers a solution to all of these problems!!

When to use Nonparametric methods

- With correct assumptions (e.g., normal distribution), parametric methods will be more efficient / powerful than non-parametric methods .. but often not as much as you might think [4].
- Many non-parametric methods convert raw values to ranks and then analyze ranks: ‘wasteful’ if there is a valid parametric alternative (‘loss’ of information).
- The simplest nonparametric test: **Sign test**:
- Consider the IQ score example - 18 girls of similar age has scores: 114, 81, 87, 114, 113, 87, 111, 89, 93, 108, 99, 93, 100, 95, 93, 95, 106, 108. If the median score is 100, on an average half of them will lie above 100 and half below. Thus the number of observations larger than 100, say K_{100} can be used to test the hypothesis.
- If K_{100} is too high or too low, the median is away from 100.

[4] The large-sample efficiency of the Wilcoxon test compared to the t test is $\frac{3}{\pi} = 0.9549$

Sign Test (continued)

Obs	114	81	87	114	113	87	111	89	93
Diff	14	-19	-13	14	13	-13	11	-11	-7
Sign	+	+	+	+	+	+	+	+	+

Obs	108	99	93	100	95	93	95	106	108
Diff	8	-1	-7	0	-5	-7	-5	6	8
Sign	+	+	+	+	+	+	+	+	+

- K_{100}: Total number of positives: 7 out of 18. Does this look too high or too low?

Sign Test (continued)

Obs	114	81	87	114	113	87	111	89	93
Diff	14	-19	-13	14	13	-13	11	-11	-7
Sign	+	+	+	+	+	+	+	+	+

Obs	108	99	93	100	95	93	95	106	108
Diff	8	-1	-7	0	-5	-7	-5	6	8
Sign	+	+	+	+	+	+	+	+	+

- K_{100}: Total number of positives: 7 out of 18. Does this look too high or too low?
- You can theoretically calculate the probability of observing 7 heads or more in 18 tosses, for a fair coin (P(head) = 0.5).

Sign Test (continued)

Obs	114	81	87	114	113	87	111	89	93
Diff	14	-19	-13	14	13	-13	11	-11	-7
Sign	+	+	+	+	+	+	+	+	+

Obs	108	99	93	100	95	93	95	106	108
Diff	8	-1	-7	0	-5	-7	-5	6	8
Sign	+	+	+	+	+	+	+	+	+

- K_{100}: Total number of positives: 7 out of 18. Does this look too high or too low?
- You can theoretically calculate the probability of observing 7 heads or more in 18 tosses, for a fair coin (P(head) $= 0.5$).
- or, you can open Table G in Gibbons and Chakraborty to look up the value ... or,

Sign Test (continued)

Obs	114	81	87	114	113	87	111	89	93	108	99	93	100	95	93	95	106	108
Diff	14	-19	-13	14	13	-13	11	-11	-7	8	-1	-7	0	-5	-7	-5	6	8
Sign	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

- K_{100}: Total number of positives: 7 out of 18. Does this look too high or too low?
- You can theoretically calculate the probability of observing 7 heads or more in 18 tosses, for a fair coin (P(head) $= 0.5$).
- or, you can open Table G in Gibbons and Chakraborty to look up the value ... or,
- Just open R, type `binom.test(7,18)`:

R Output

```
binom.test(7,18)
Exact binomial test
data:  7 and 18
number of successes = 7, number of trials = 18, p-value = 0.4807
alternative hypothesis:  true probability of success is not equal to 0.5
95 percent confidence interval:  0.1729859 0.6425488

sample estimates:  probability of success 0.3888889
```

- P-value = $P(\text{Number of Success} \geq 7 \mid H_0 : \text{Median IQ} = 100) = 0.4807$
- We fail to reject H_0: Not enough evidence of a difference.
- We will learn Sign test with more details but it is so intuitive that anyone can understand the idea.

Question: Could you tell me one drawback/disadvantage of Sign Test?

Question: Could you tell me one drawback/disadvantage of Sign Test?

Completely ignores the actual magnitudes - low power (inefficient)!!

Reasons to use Nonparametric Methods

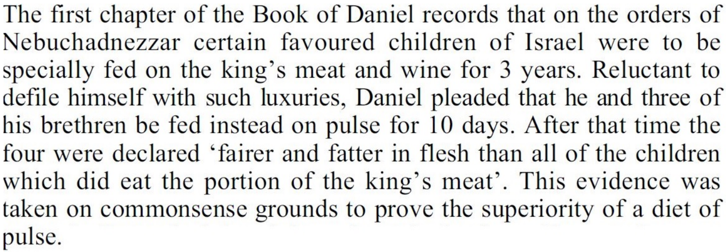
1. A small sample size. (Double-edged sword!)
2. Ordinal data, ranked data, or outliers that you can't remove.
3. The data-distribution doesn't follow any known or nice parametric distribution (not just deviation from Normality!)
4. Median is a better representative of your area than mean. e.g. Income distribution, Student's grade distribution, ..
5. Usually simple tests, and less affected by departures from ideal world - but limited in power and conclusions!

Reasons to use Parametric Methods

1. You have a large sample, and the statistic is approximately normally distributed (Central limit theorem!)
2. Normal distribution doesn't hold, but other parametric tests can be used, or, you can transform the data suitably for one (e.g. standardize the data).
3. IID data can be a major reason in favour of parametric methods and vice versa.
4. Can draw more conclusions (at the population level).

The first ever nonparametric Statistics!

These anecdotes/historical background is from Sprent & Smeeton's book:



The first chapter of the Book of Daniel records that on the orders of Nebuchadnezzar certain favoured children of Israel were to be specially fed on the king's meat and wine for 3 years. Reluctant to defile himself with such luxuries, Daniel pleaded that he and three of his brethren be fed instead on pulse for 10 days. After that time the four were declared 'fairer and fatter in flesh than all of the children which did eat the portion of the king's meat'. This evidence was taken on commonsense grounds to prove the superiority of a diet of pulse.

Figure: First nonparametric test

Although the biblical analysis is informal, it contains the germ of a nonparametric test.

History of nonparametric Statistics

John Arbuthnot (1710) observed that in each year from 1629 to 1710 the number of males christened in London exceeded the number of females. He regarded this as strong evidence against the probability of male birth being $\geq \frac{1}{2}$.
Francis Galton (1892) developed a measure - which he termed a ‘centisimal scale’ - to assess agreement between patterns (categorical data) on corresponding fingertips on left and right hands.
Karl Pearson (1900) proposed the well-known, and sometimes misused, chi-squared goodness-of-fit test applicable to any discrete distribution, and C. Spearman (1904) defined a rank correlation coefficient that bears

Arbuthnot, Galton, Pearson and Spearman.

his name.

Rest of this week: Review of Basic Statistics (if you need it!) or Start Sign Test.

- Sample space, experiments.
- Conditional probability.
- CDF and Independence.
- Expectation and Moments.
- Inequalities.
- Moment generating functions.
- Standard discrete distributions.
- Standard Continuous distributions.

Why staticians don't make it as waiters...
Hi sir, I normally distributed the butter on your toast this morning...
-mak