Sampling

- Two investigators were examining the same question: does caffeine affect performance on a simple spelling test. Assume there really is a difference of .3 standard deviations between the two populations (with/without caffeine).
- One experimenter used 10 subjects, one 100.
- Which experimenter is more likely to find a difference between the two groups? Why?
- What if there really were no difference in the population? What then? Why?

Research Methods

Review of basic statistics: Central tendencies and measures of dispersion

Data = model + residual

- Observed data may be represented by a model of the data. What is left over is residual (sometimes called error).
- The process of research is to model the data and reduce the residual.

Consider the recall data

- How to describe it?
- Raw data?
- Summary statistics
- Graphically
- All tables and graphs are prepared by using the R computer package. For details on using R, consult the tutorials, particularly the short tutorial, listed in the syllabus.

Data analysis - raw data

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
1	8	8	8	8	8	7	6	8	6	8	8	8	8	8	8
2	6	5	5	1	2	3	3	3	3	2	3	2	3	2	3
3	5	5	5	6	4	4	3	4	4	2	2	6	6	6	3
4	6	5	5	6	4	4	2	4	3	6	4	2	3	3	6
5	6	7	7	5	6	5	4	4	1	4	3	6	4	6	6
6	5	6	8	7	6	4	4	3	4	8	4	5	6	6	6
7	8	8	5	8	5	5	7	6	5	6	5	4	7	7	8
8	8	7	5	8	5	7	5	6	5	5	4	5	7	7	4
9	8	8	8	7	7	8	8	6	6	2	5	1	4	3	6
10	5	5	5	4	4	3	1	2	2	1	2	2	4	3	2
11	7	6	6	5	5	6	7	3	3	5	3	4	4	3	6
12	8	4	5	4	4	5	3	5	4	1	4	6	6	8	8
13	8	7	6	3	2	5	4	1	3	2	5	7	6	4	4
14	5	6	5	1	3	7	3	4	6	6	3	3	4	3	3
15	8	6	7	2	5	3	6	5	4	5	5	6	6	6	7
16	8	7	7	8	8	6	6	7	7	6	6	4	5	6	6
17	8	8	6	6	5	6	7	5	5	7	6	8	8	8	7
18	8	7	6	7	6	7	6	4	6	6	6	6	6	8	8
19	8	7	8	6	7	7	8	7	5	7	5	6	5	5	6
20	7	6	6	6	5	2	5	2	5	5	5	4	6	6	7
21	8	7	6	5	5	6	6	7	4	6	5	3	5	6	3
22	7	7	6	3	4	4	5	4	7	4	4	4	5	7	5
23	8	7	7	5	3	6	4	3	5	3	4	3	5	3	2
24	6	4	3	5	4	2	4	1	3	3	3	3	8	5	5
25	8	7	6	6	8	6	6	6	5	4	5	7	8	7	6

We rarely want to show these but have them so that we can check the numbers.
Can we see any patterns in the data?

Simple descriptives

Frequency counts table(recall)

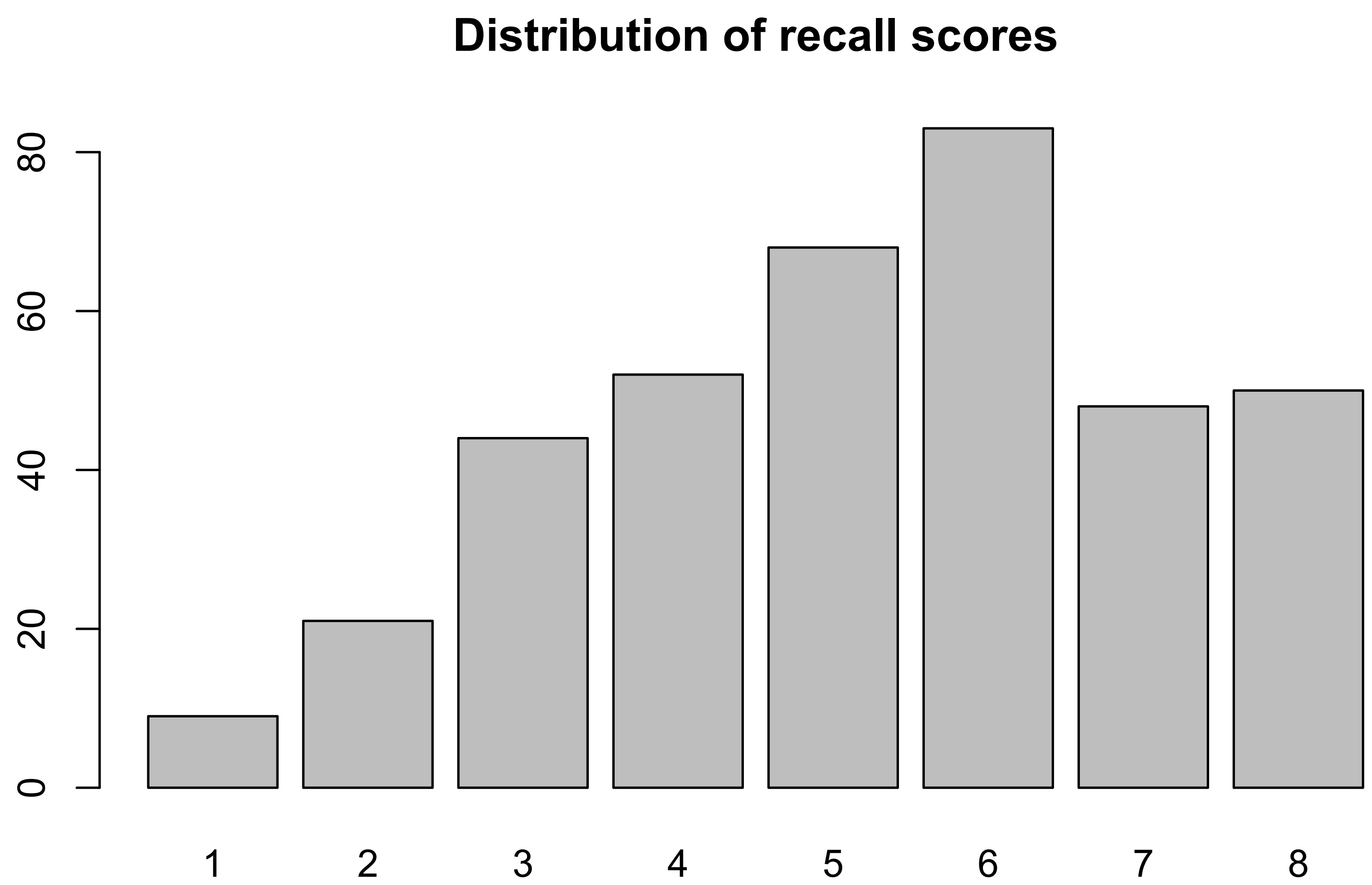
0	1	2	3	4	5	6	7	8
0	9	21	44	52	68	83	48	50

Distribution description summary(recall)

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
1	4	5	5.24	7	8

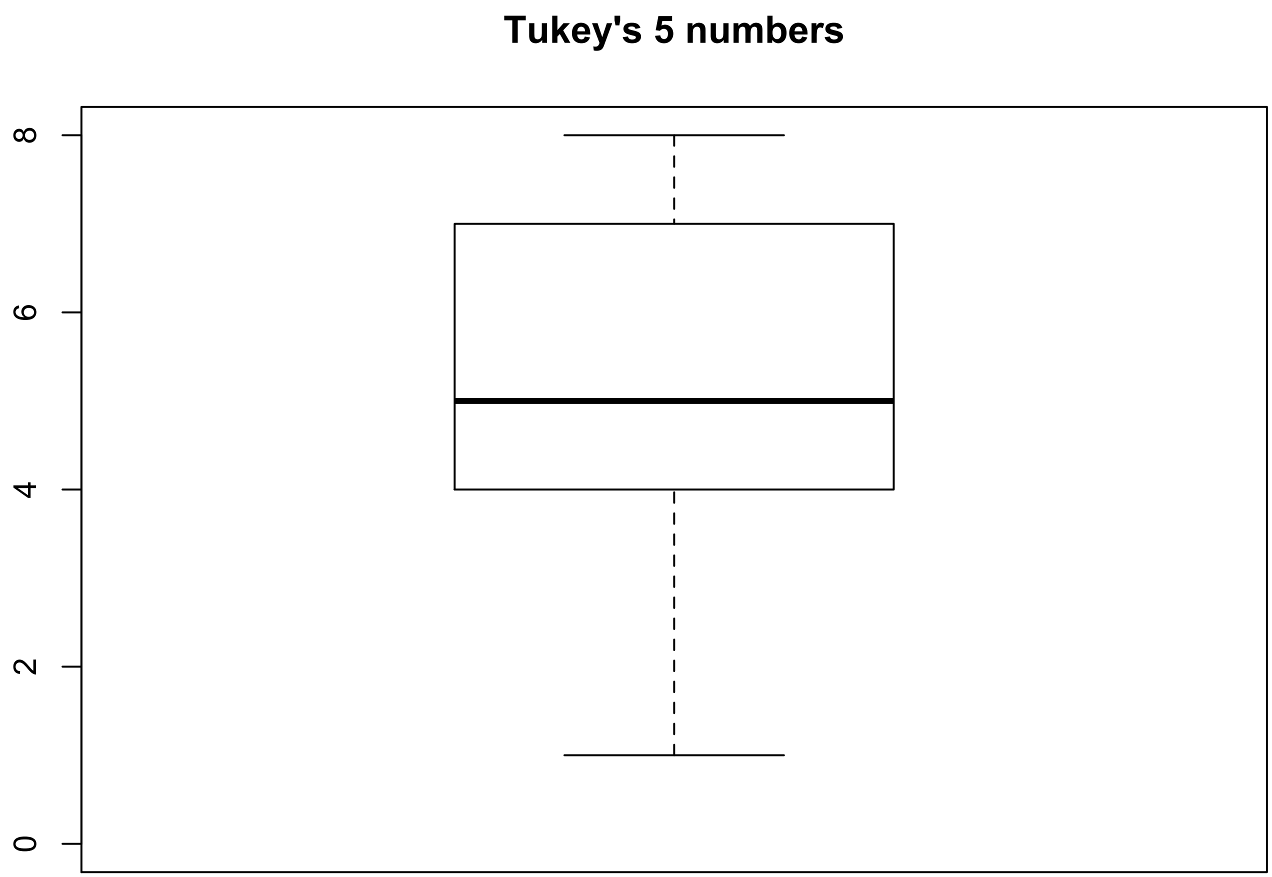
Distribution of recall

barplot(recall,main=”Distributon of recall scores”)



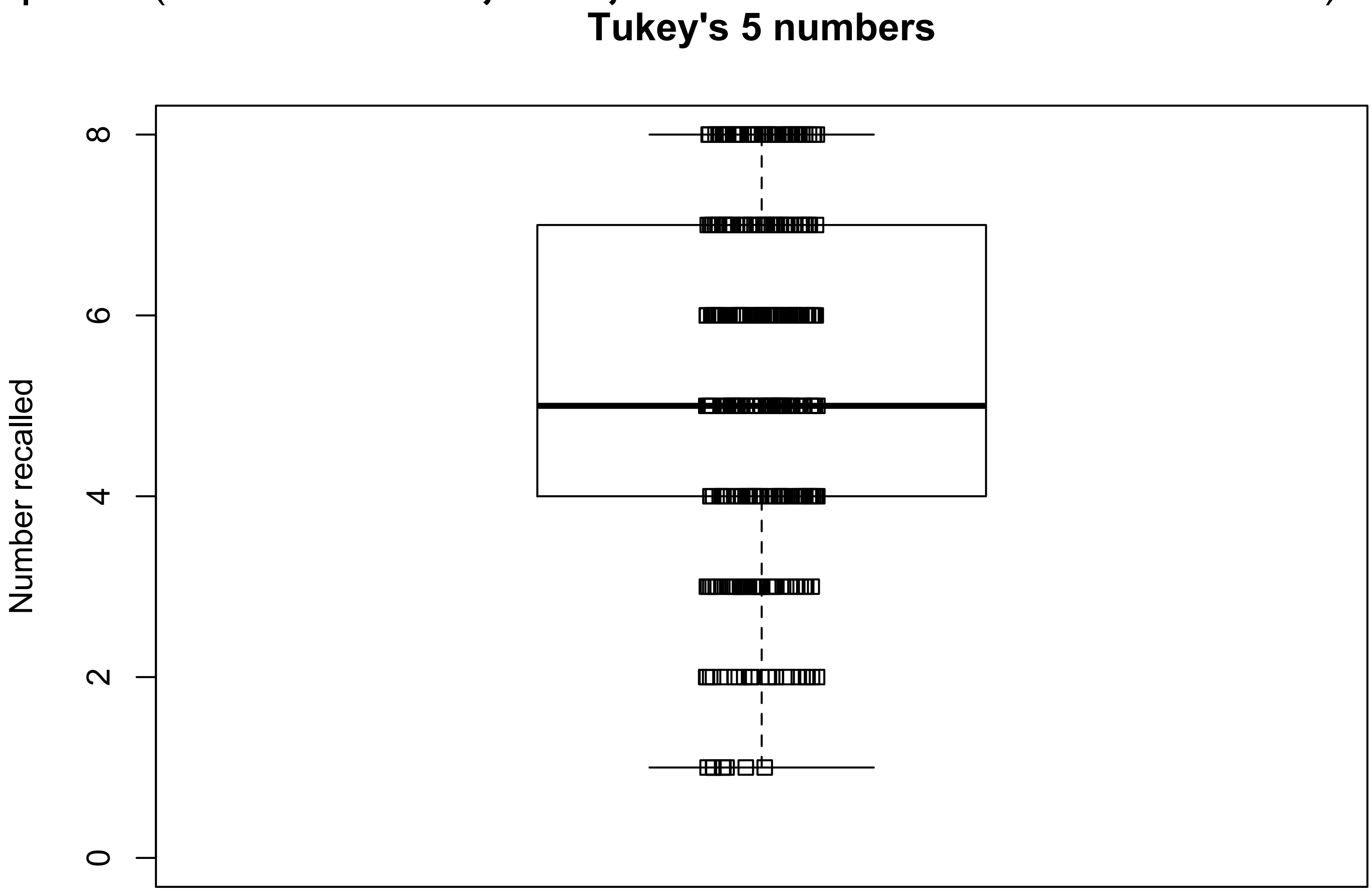
Graphical Display: Box plot

```
boxplot(recall,ylim=c(0,8),main="Tukey's 5 numbers")
```



Graphical Display

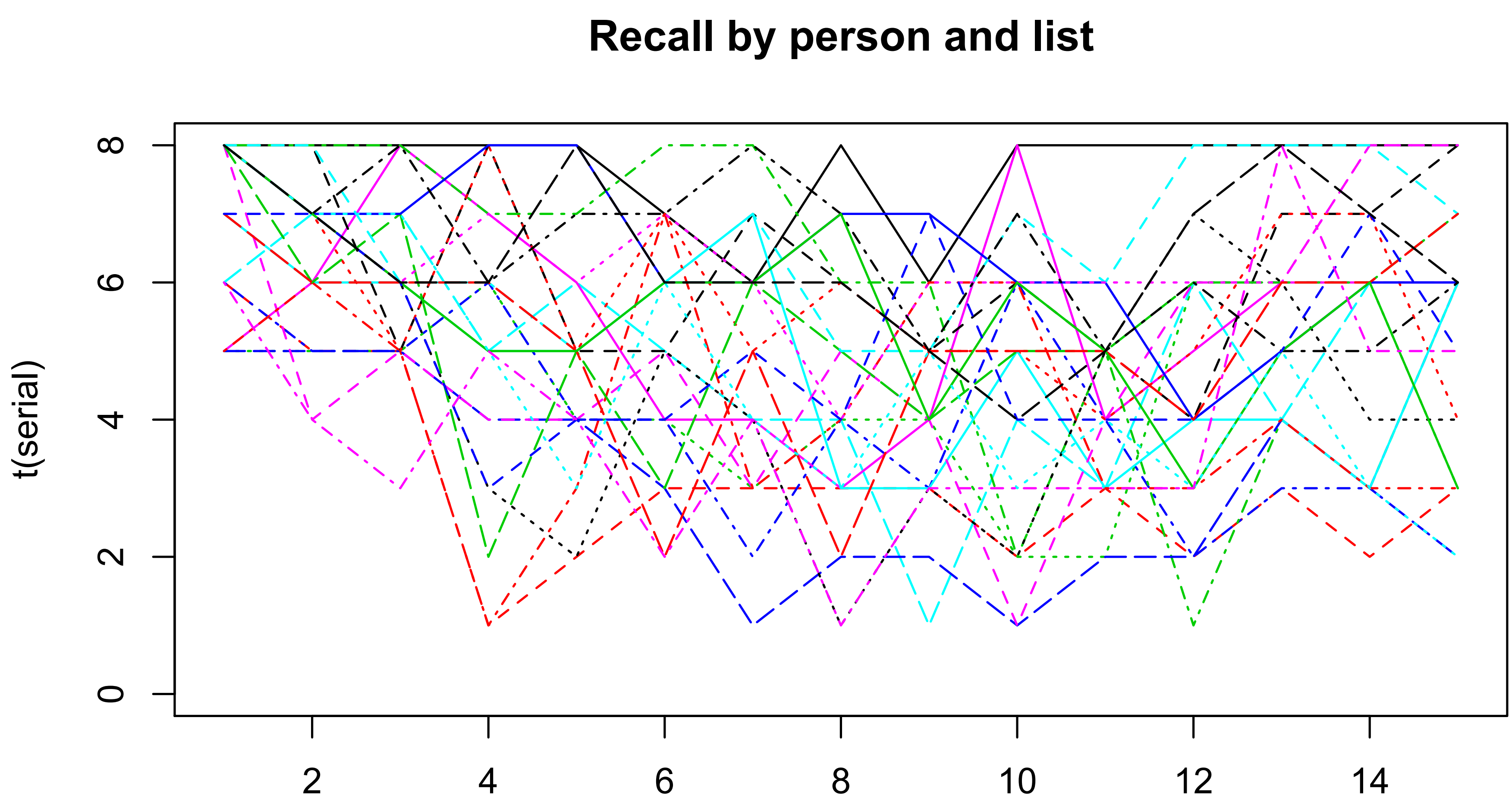
```
> boxplot(recall,ylim=c(0,8),ylab="Number recalled",main="Tukey's 5 numbers")
> stripchart(recall,method="jitter",jitter=.05,vertical=TRUE,add=TRUE)
```



Reduce the uncertainty

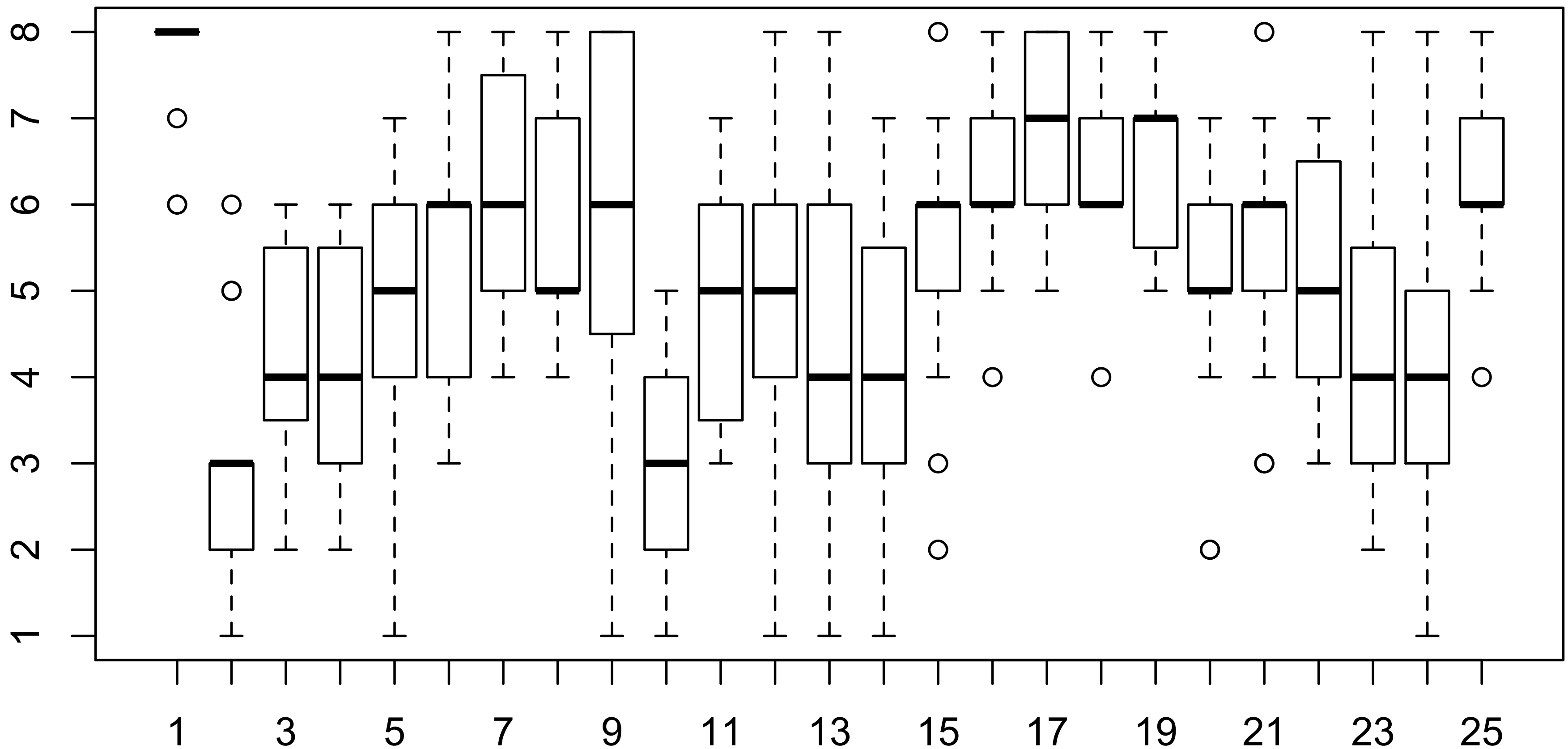
- Three sources of variability
  - –between person variability
  - –within person variability over lists
  - – interaction of person x list (different patterns for people)

Data by person and list



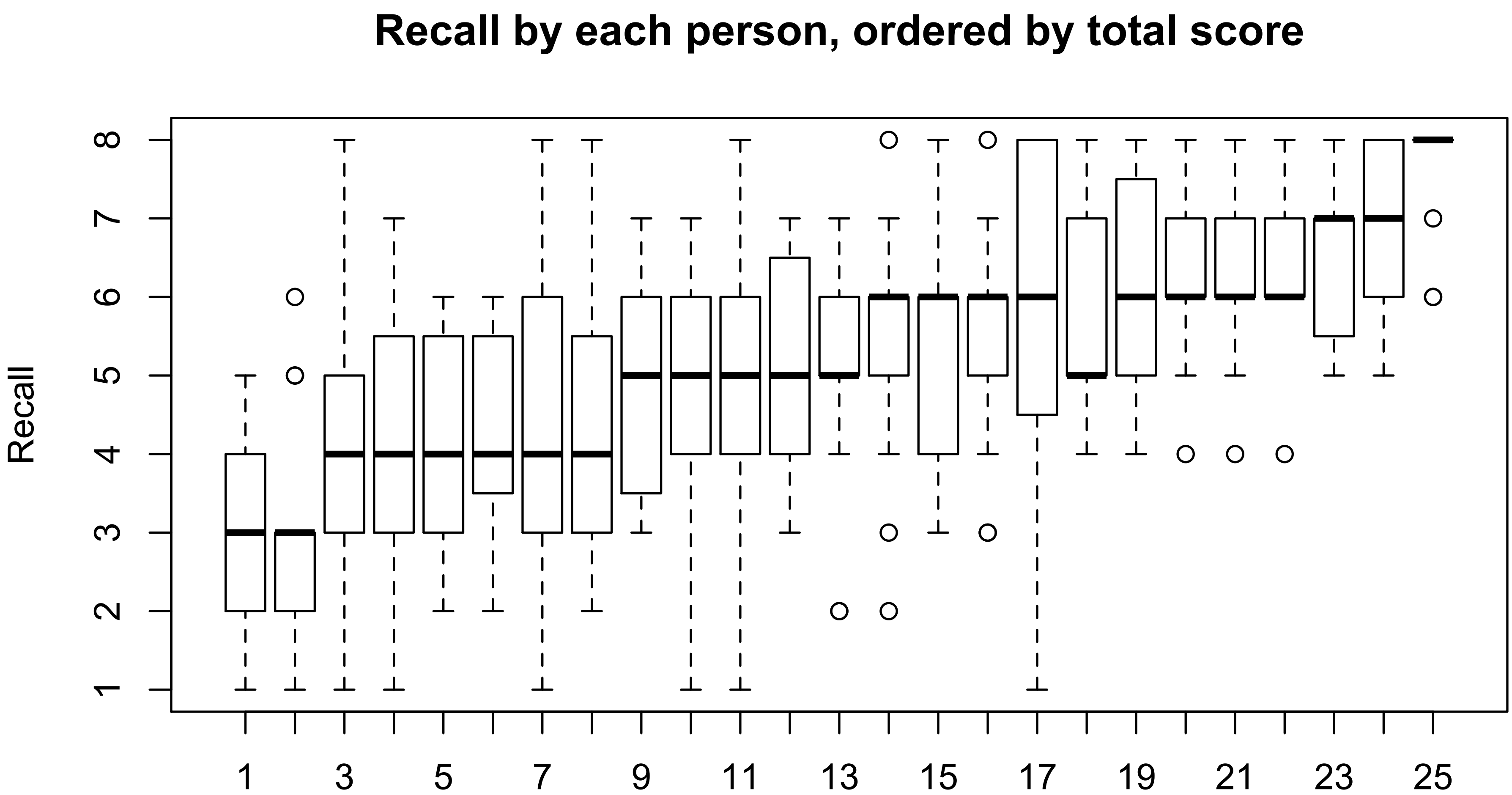

```
matplot(t(serial),typ="l",ylim=c(0,8),main="Recall by person and list")
```

Boxplots for each person



boxplot(t(recall),main="Describe each individual")

Boxplots for each person

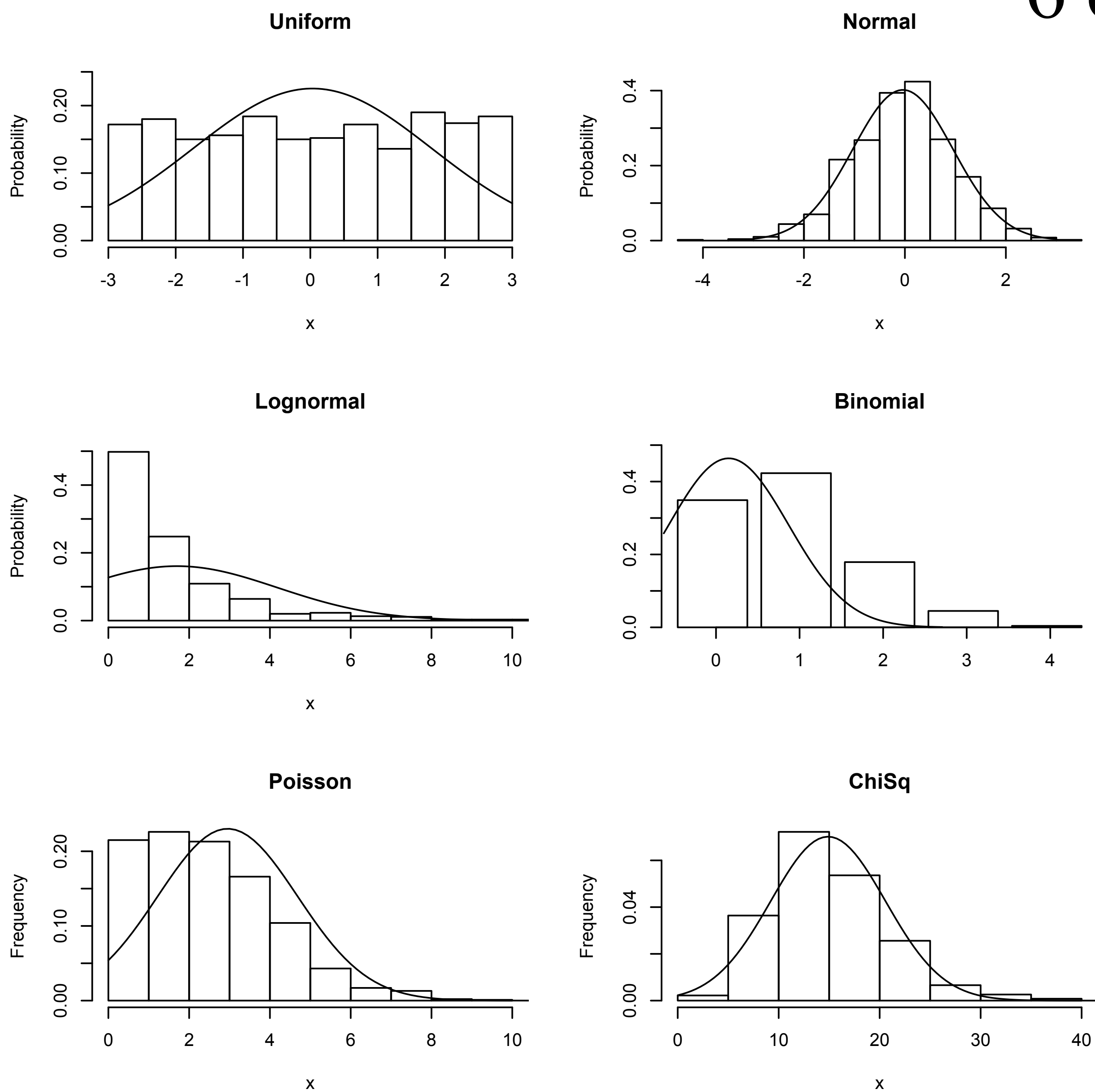



```
> recor <- recall[order(tot),]
> boxplot(t(recor),main="Recall by each person")
```

Statistics using Normal Theory

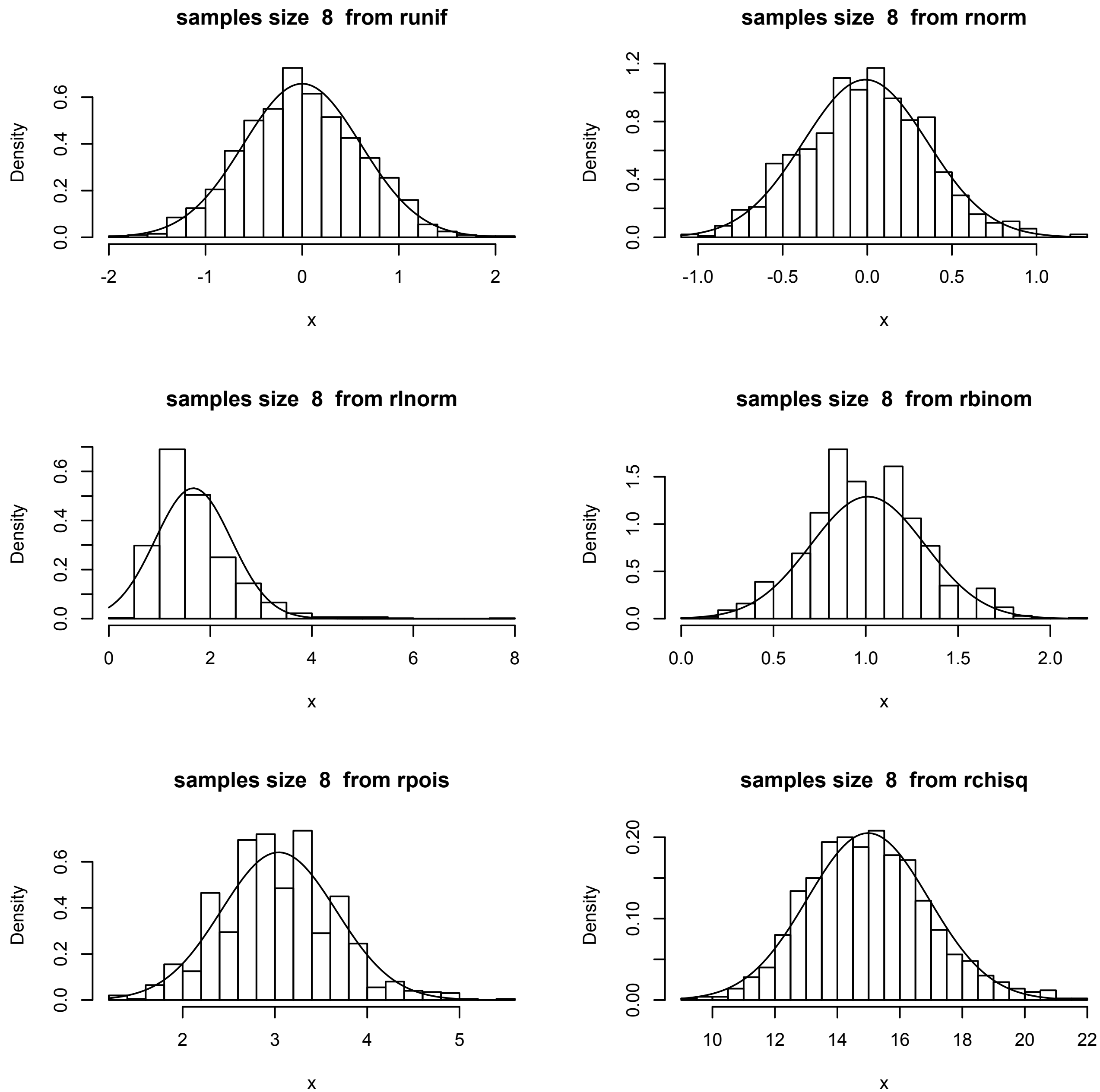
- Data come from many different types of distributions
  - rectangular (throw of a die)
  - binomial (throws of coins)
  - Poisson (fatalities by horse kicks)
  - Log normal (income)
  - Normal (height, weight, Extraversion, Neuroticism)

6 distributions



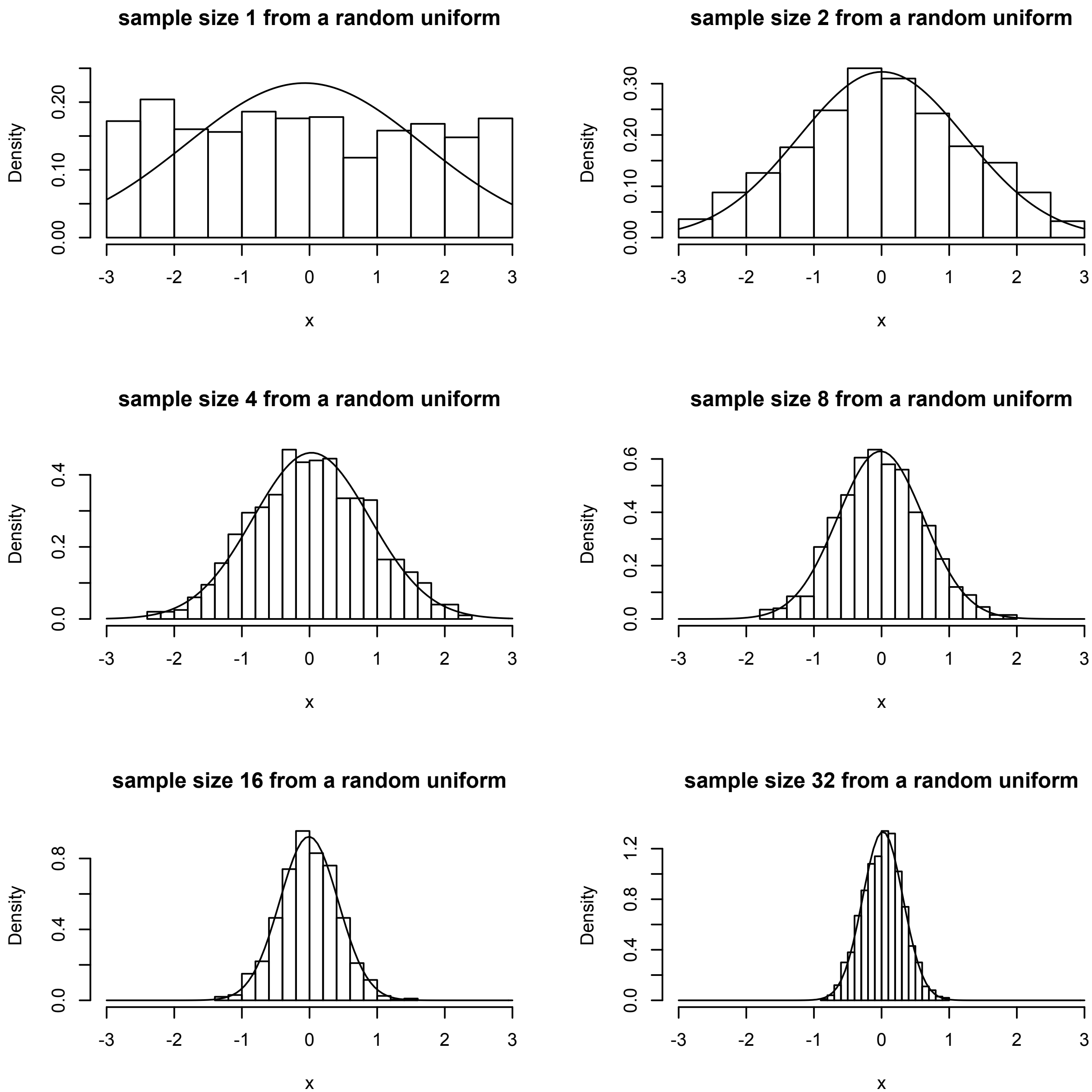
The distributions have drastically different shapes and reflect very different processes.

Sample means from 6 distributions -> normal



But the distribution of sample means (of size 8) from all of these distributions tends to be similar.

Sample means -> normal, sd varies as 1/sqrt(N)



As the sample gets larger, the variation in the sample means gets smaller and more closely approximates the normal distribution.

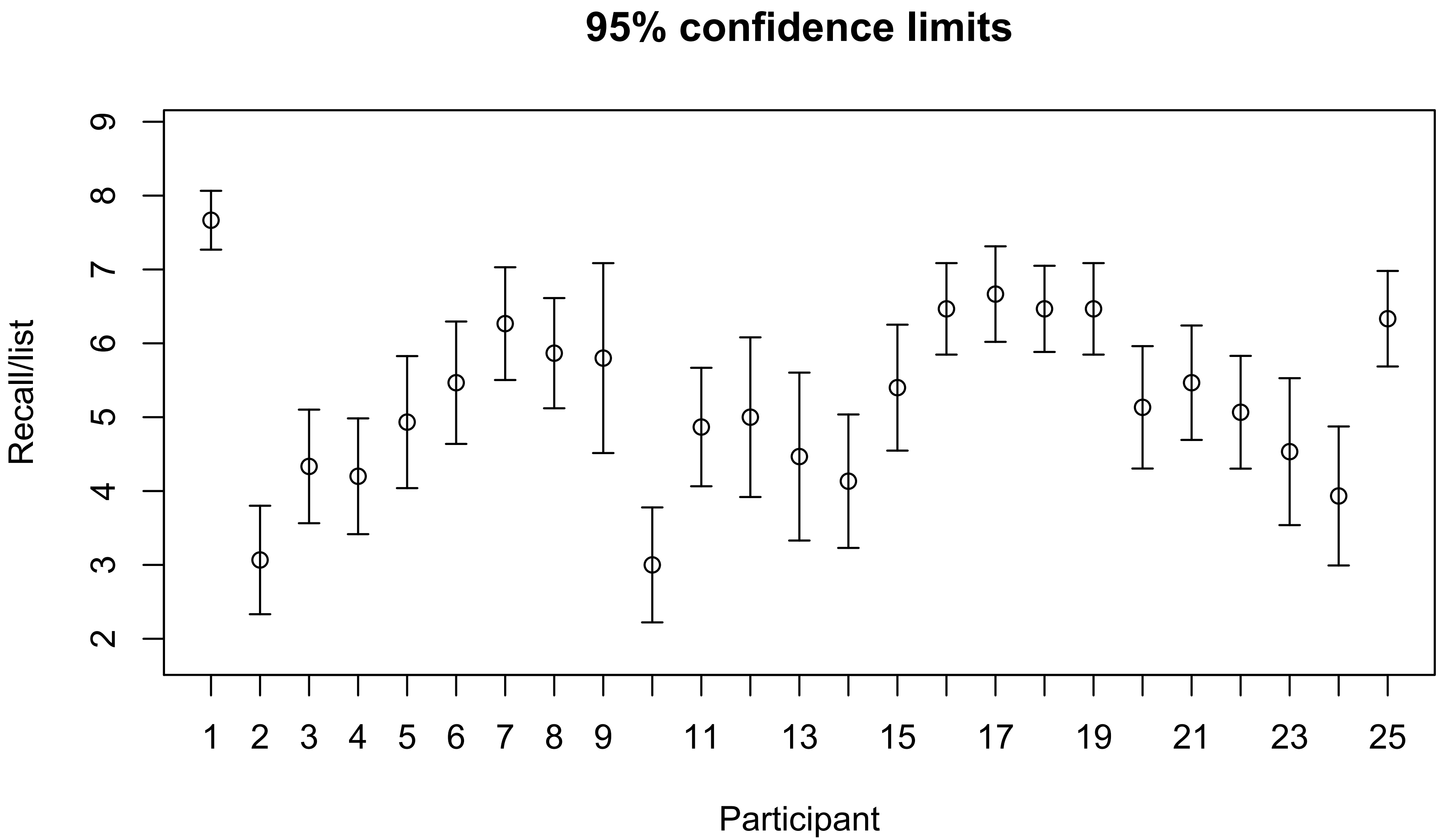
Central Limit Theorem

- Distribution of sample means taken from finite distributions will tend towards normal distribution as sample size increases
- Standard deviation varies as $\frac{1}{\sqrt{N}}$

Plot each person

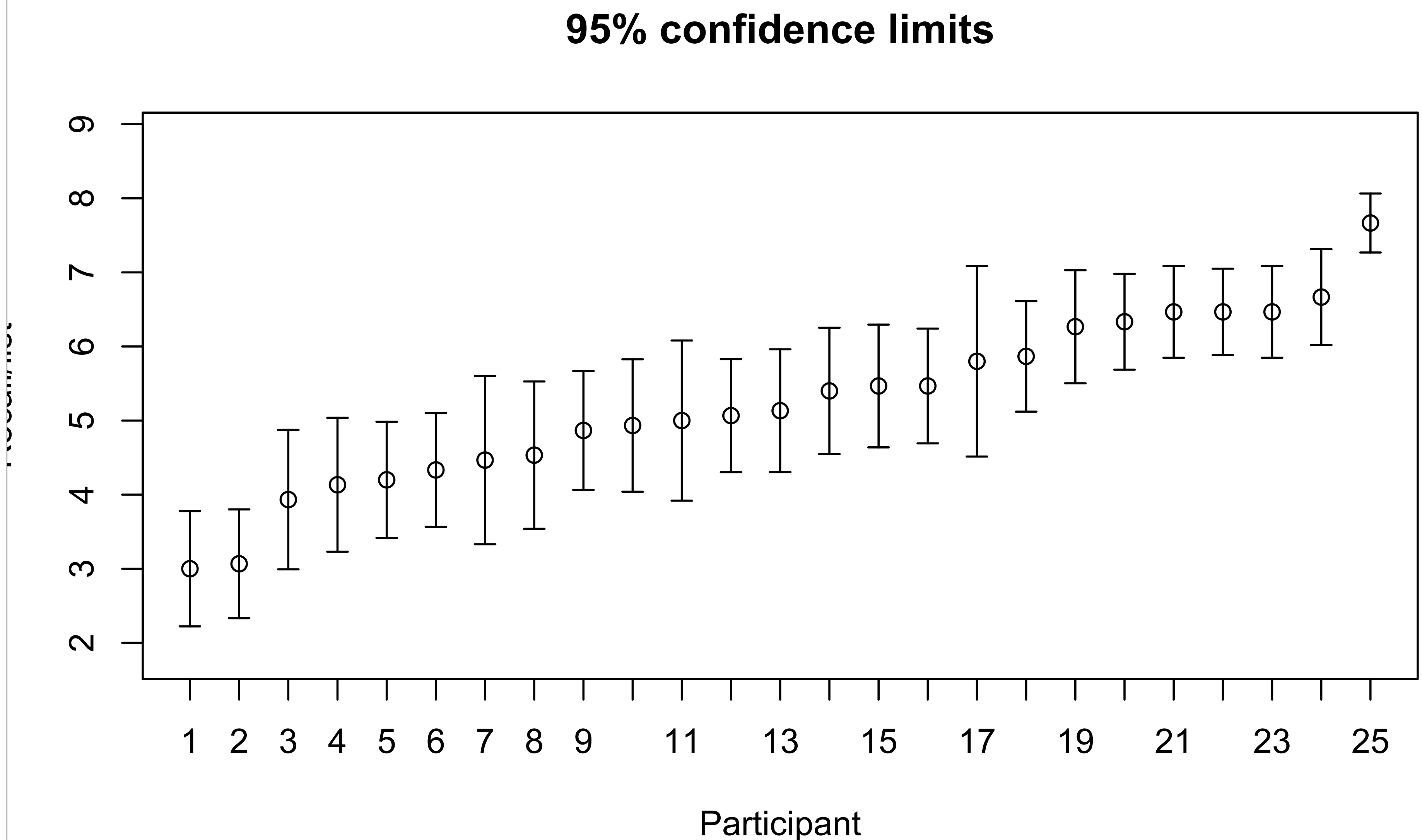
- Find their mean
- Find their standard deviation
- The standard error = $\dfrac{\sigma_i}{\sqrt{N_i}}$
- 95% confidence from a normal is ≈ 2 standard errors (1.96 s.e.)

Person means



error.bars(t(recall),main="Recall for each person",ylab="Recall/list",xlab="Subject")

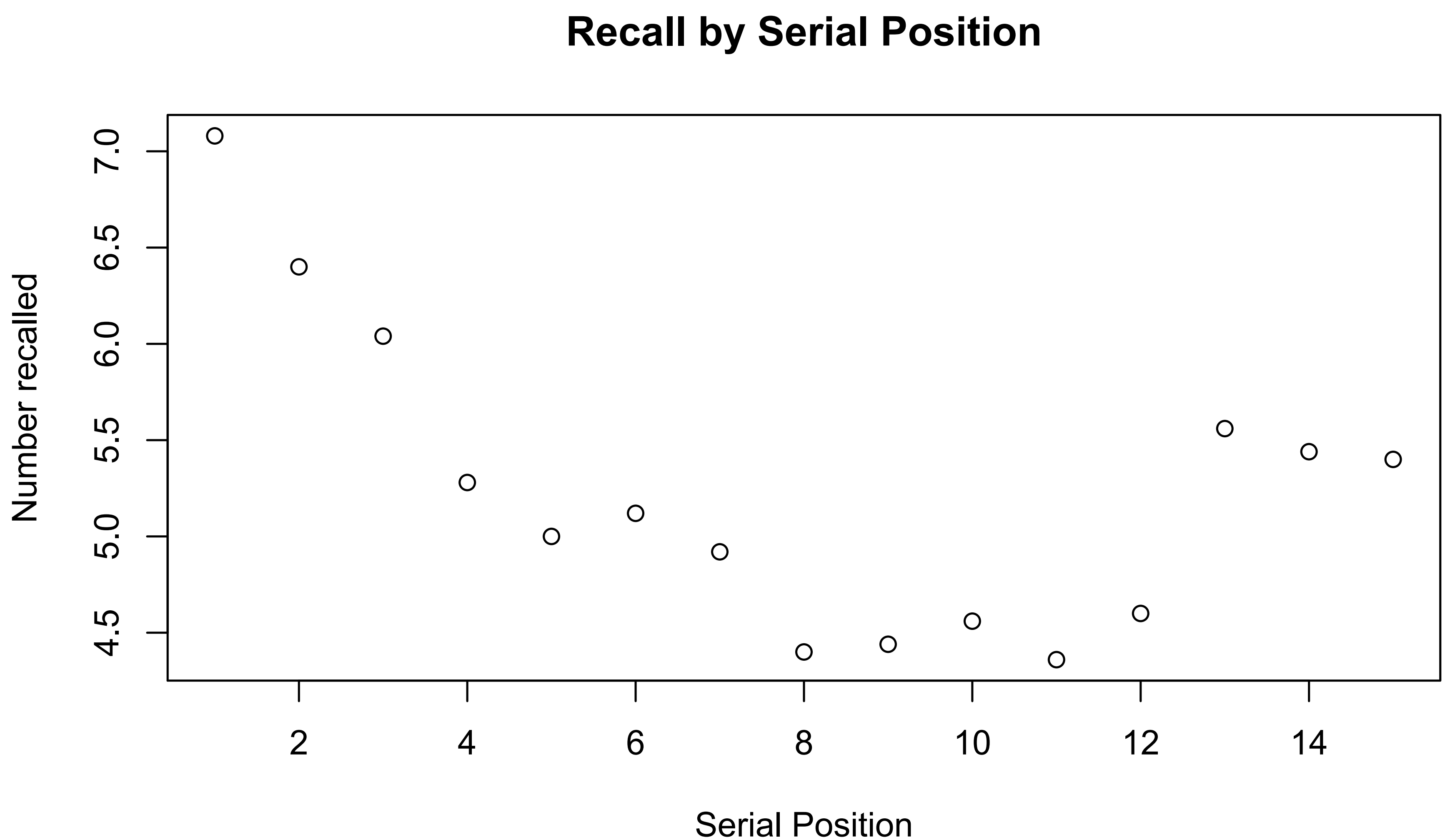
Order the people



Sources of variability

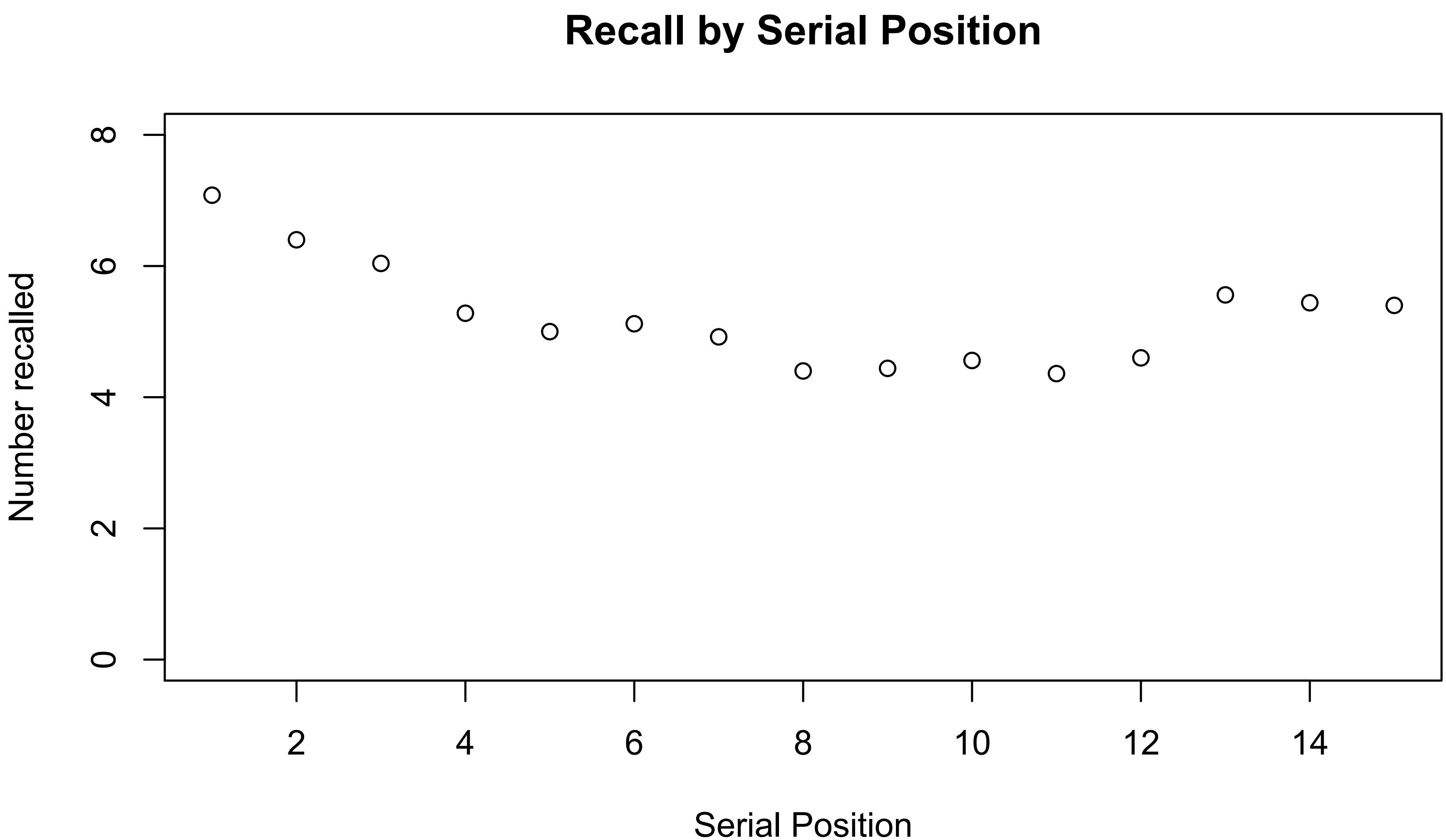
- Person variability is usually a very large source of variability.
- Within subject studies examine scores pooled across subjects over conditions

Just the means (bad)



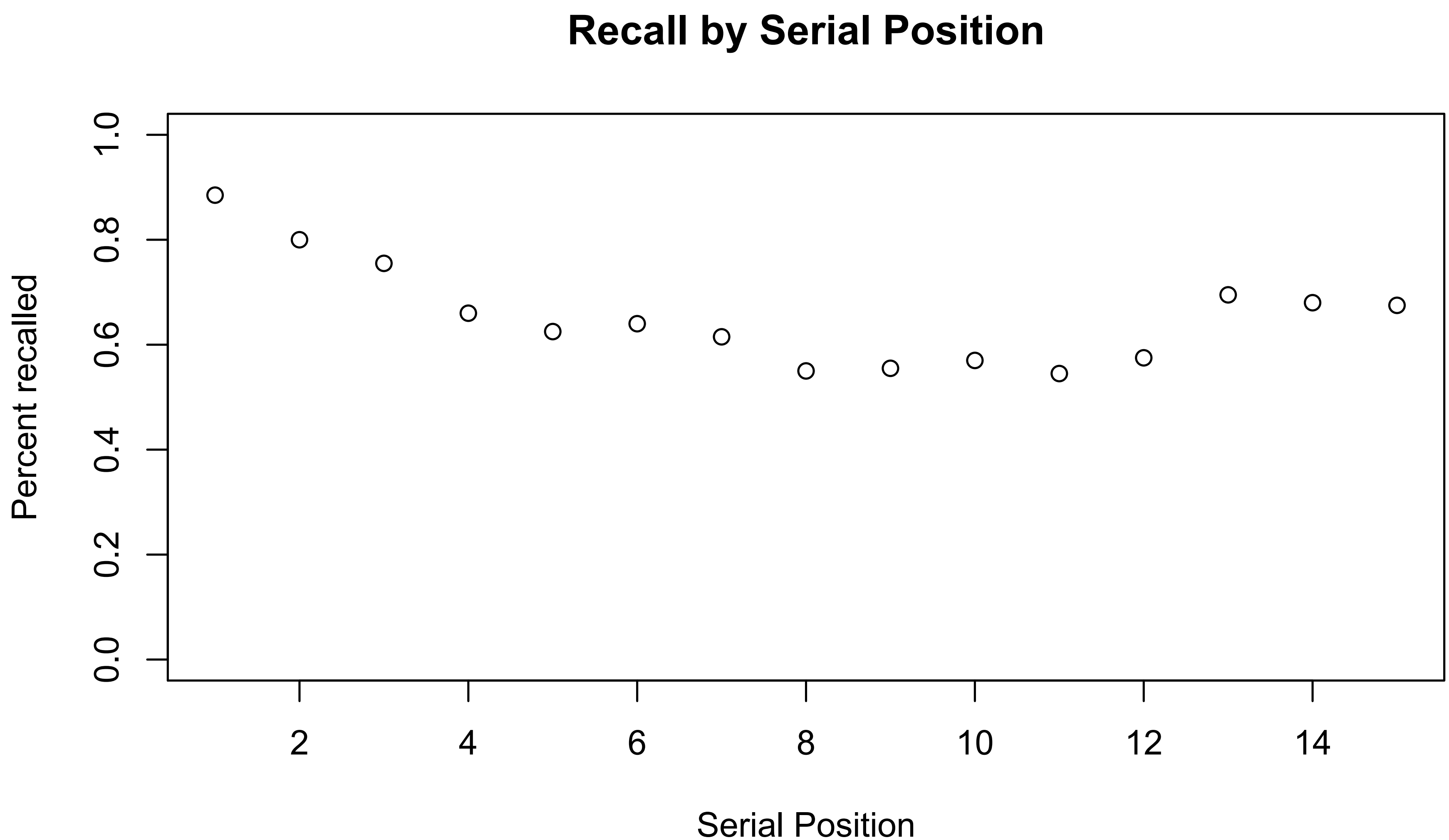

```
plot(colMeans(recall,na.rm=TRUE),ylab="Number recalled",xlab="Position Number",main="Recall by serial position")
```

Just the means (better)



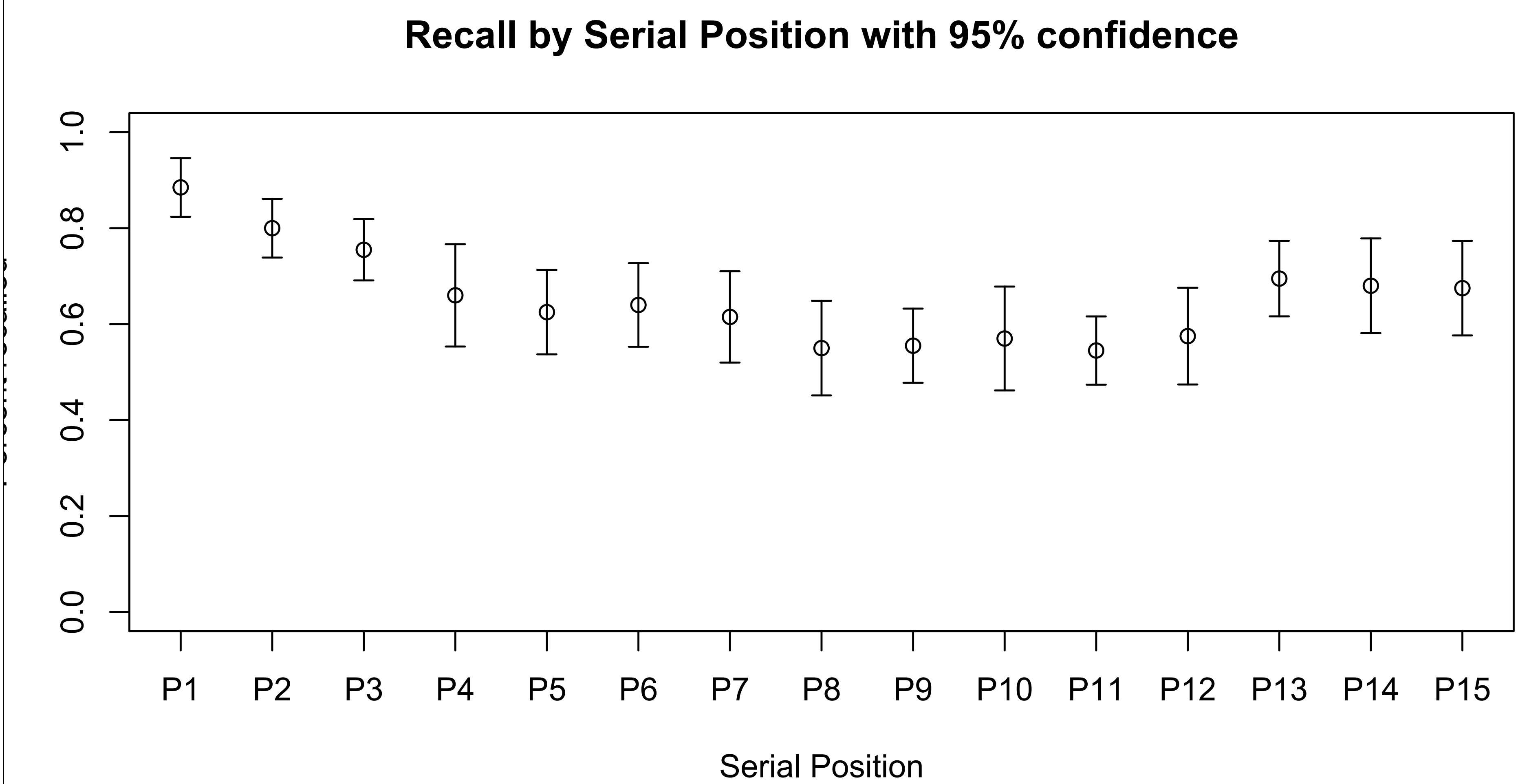
plot(colMeans(recall,na.rm=TRUE),ylab="Number recalled",xlab="Position Number",main="Recall by serial position", ylim=c(0,8))

Convert to Percentage recalled and plot them

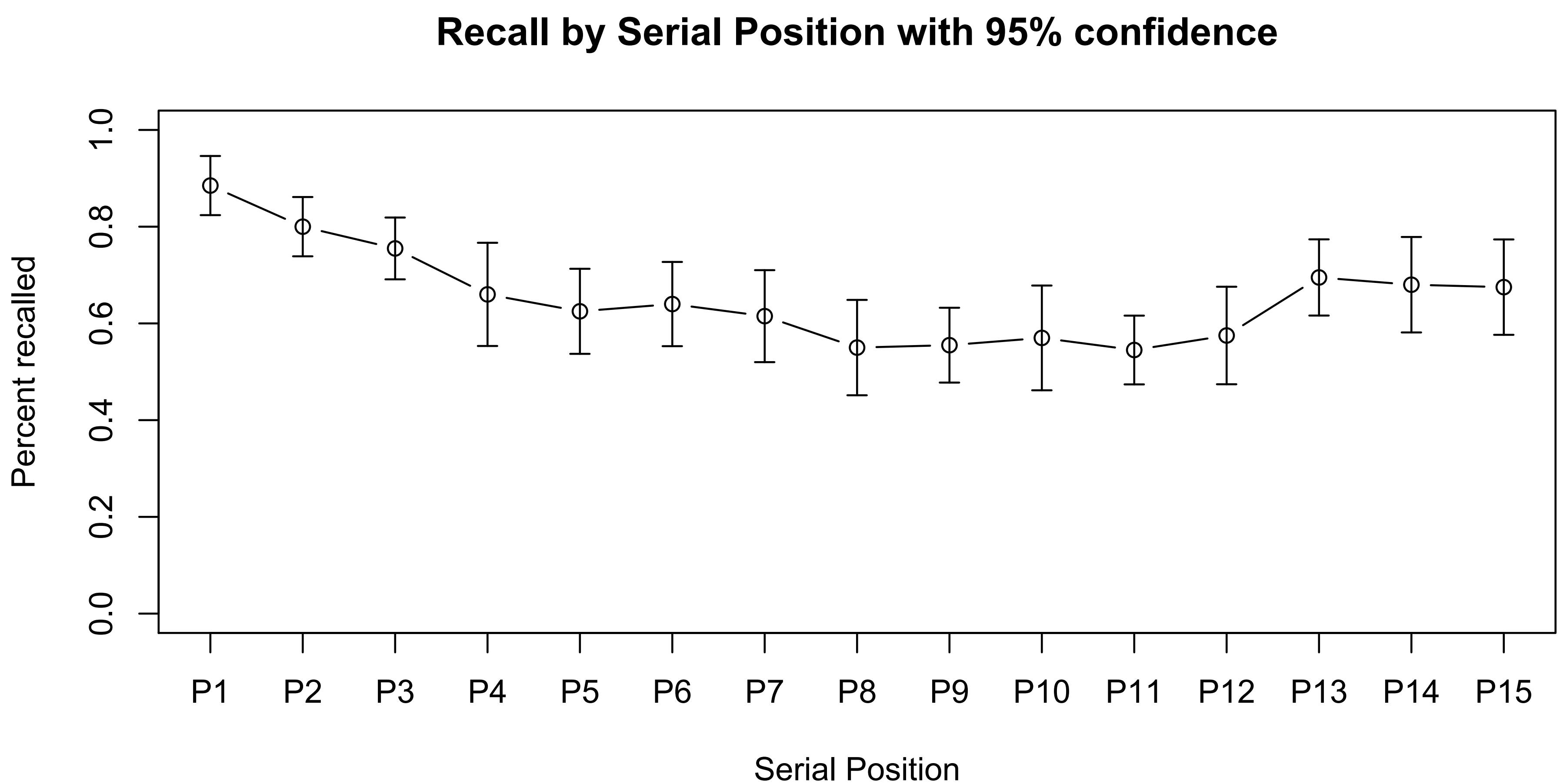


Means +/-Confidence intervals

```
error.bars(serial/8,ylim=c(0,1),ylab="Percent recalled",xlab="List position",main="Mean recall by list position + 95%
confidence")
```



Help the eye




```
error.bars(serial2,ylim=c(0,1),xlab="Serial Position",ylab="Percent
recalled",main="Recall by Serial Position with 95% confidence",typ="b")
```

Understanding the statistics

- Measures of central tendency
- Measures of dispersion
- Expected variation of means from sample to sample

Estimates of Central Tendency

- Consider a set of observations $X = \{x_1, x_2 \ldots x_n\}$
- What is the best way to characterize this set
  - Mode: most frequent observation
  - Median: middle of ranked observations

Mean:

$$\textbf{Arithmetic} = \overline{X} = \sum_{1}^{n}(X_i)/N$$

$$\textbf{Geometric} = \sqrt[n]{\prod_{1}^{n}(X_i)}$$

$$\textbf{Harmonic} = \frac{N}{\sum_{1}^{n}(1/X_i)}$$

Alternative expressions

- Arithmetic mean = $\sum x_i/N$

- Alternatives are anti transformed means of transformed numbers
- Geometric mean = $\exp(\sum \ln(x_i)/N)$
  - (anti log of average log)

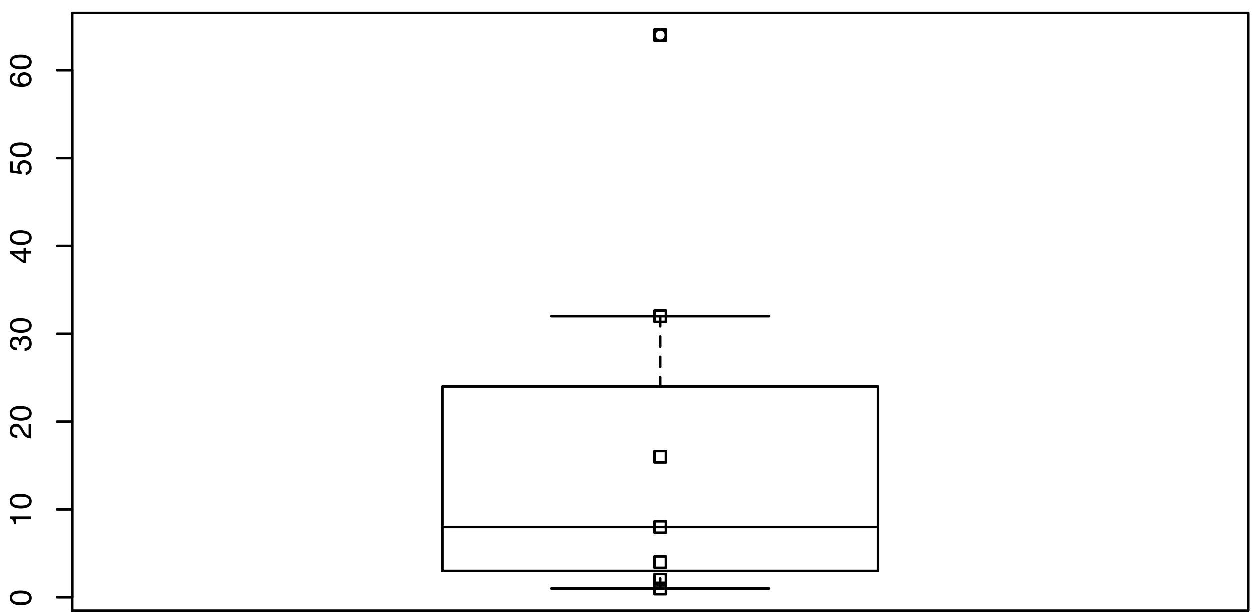
- Harmonic Mean = reciprocal of average reciprocal
  - $1/(\sum (1/x_i)/N)$

Why all the fuss?

- Consider 1,2,4,8,16,32,64
- Median = 8
- Arithmetic mean = 18.1
- Geometric = 8
- Harmonic = 3.5
- Which of these best captures the “average” value?

Summary stats (R code)

```
> x <- c(1,2,4,8,16,32,64)  #enter the data
> summary(x)  # simple summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    8.00   18.14   24.00   64.00
> boxplot(x)  #show five number summary
```



Consider two sets, which is more?

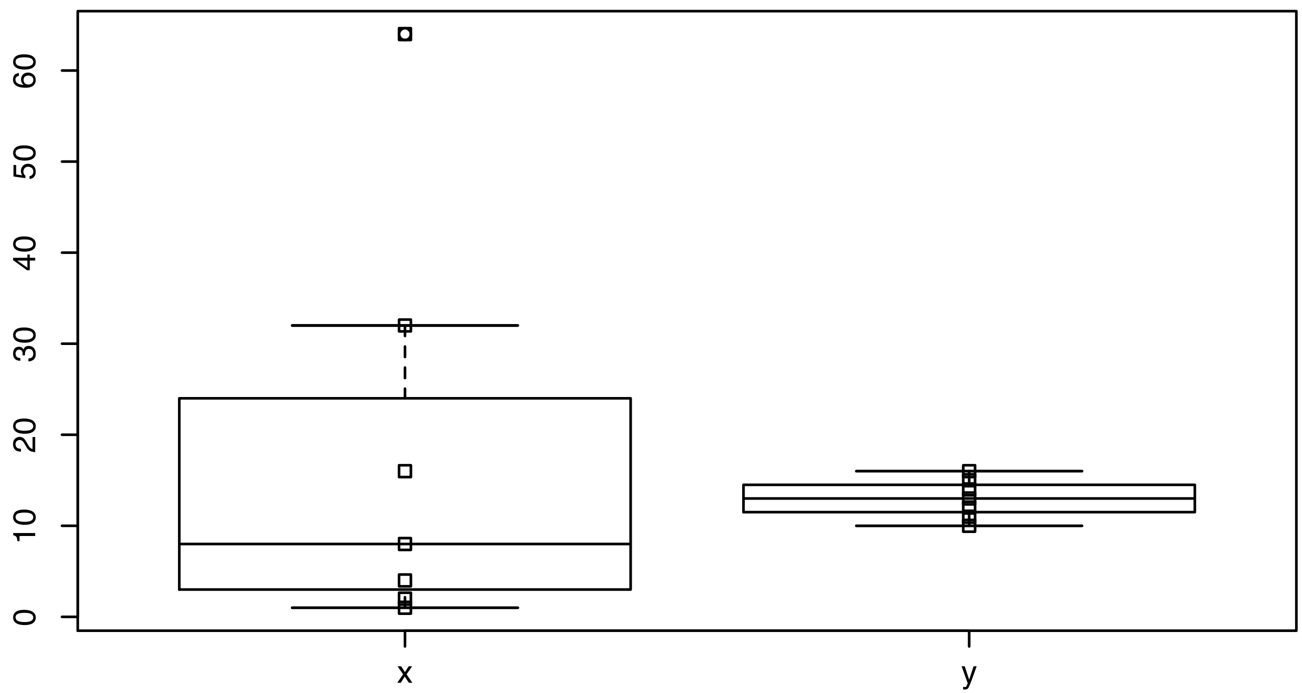
subject	Set 1	Set 2
1	1	10
2	2	11
3	4	12
4	8	13
5	16	14
6	32	15
7	64	16
median	8	13
arithmetic	18.1	13.0
geometric	8.0	12.8
harmonic	3.5	12.7

Summary stats (R code)

```
> x <- c(1,2,4,8,16,32,64)  #enter the data
> y <- seq(10,16)   #sequence of numbers from 10 to 16
> xy.df <- data.frame(x,y)  #create a "data frame"
> xy.df      #show the data
   x  y
1  1 10
2  2 11
3  4 12
4  8 13
5 16 14
6 32 15
7 64 16
> summary(xy.df)   #basic descriptive stats
       x               y       
 Min.   : 1.00   Min.   :10.0  
 1st Qu.: 3.00   1st Qu.:11.5  
 Median : 8.00   Median :13.0  
 Mean   :18.14   Mean   :13.0  
 3rd Qu.:24.00   3rd Qu.:14.5  
 Max.   :64.00   Max.   :16.0  
```

Box Plot (R)

```
boxplot(xy.df)  #show five number summary
stripchart(xy.df,vertical=T,add=T)  #add in the points
```



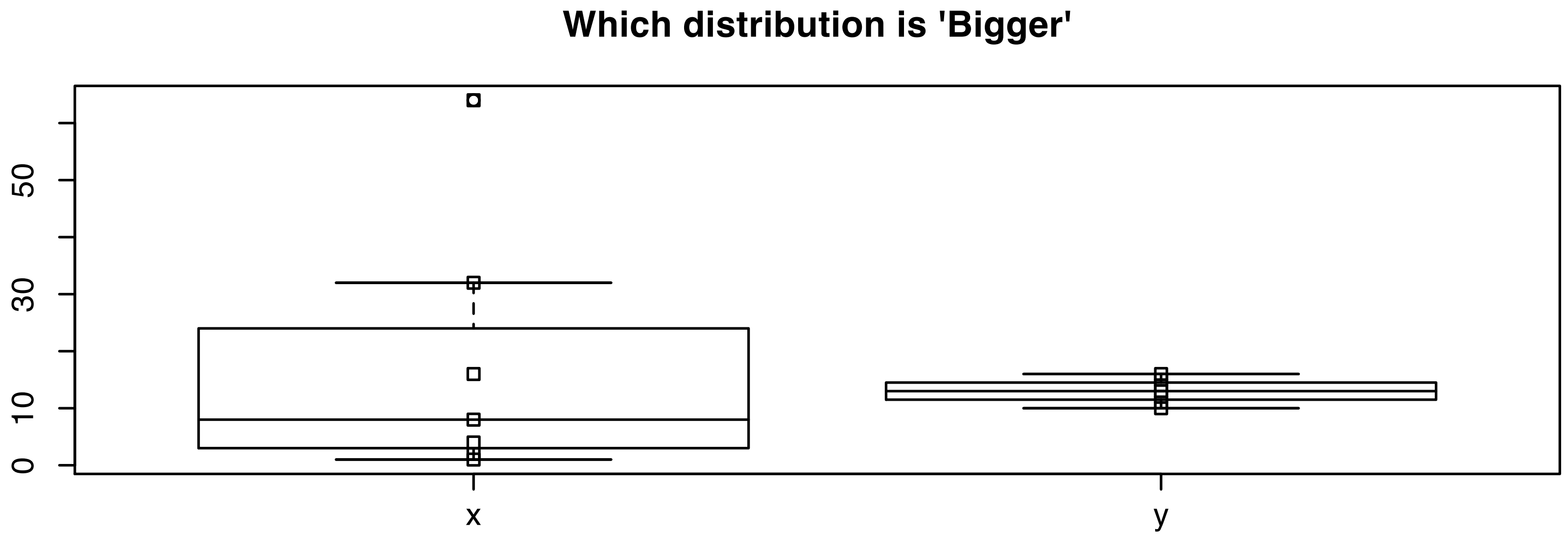
The effect of log transforms Which group is “more”?

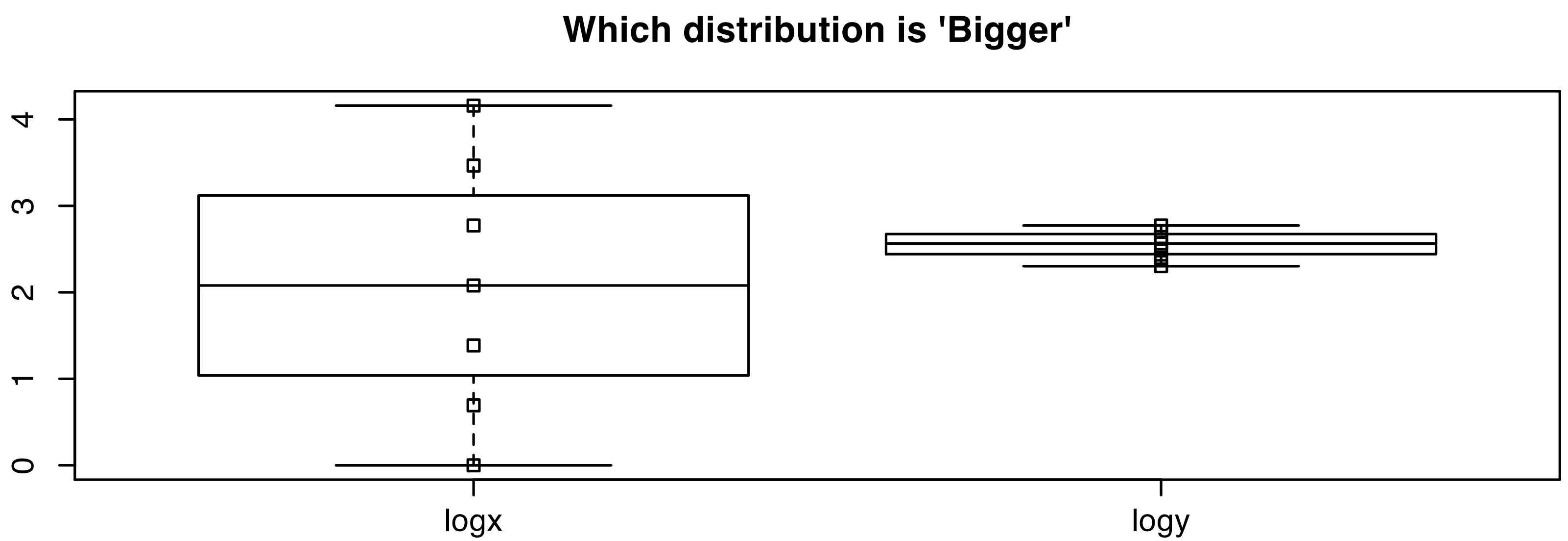
X	Y	Log X	Log Y
1	10	0.0	2.3
2	11	0.7	2.4
4	12	1.4	2.5
8	13	2.1	2.6
16	14	2.8	2.9
32	15	3.5	2.7
64	16	4.2	2.8

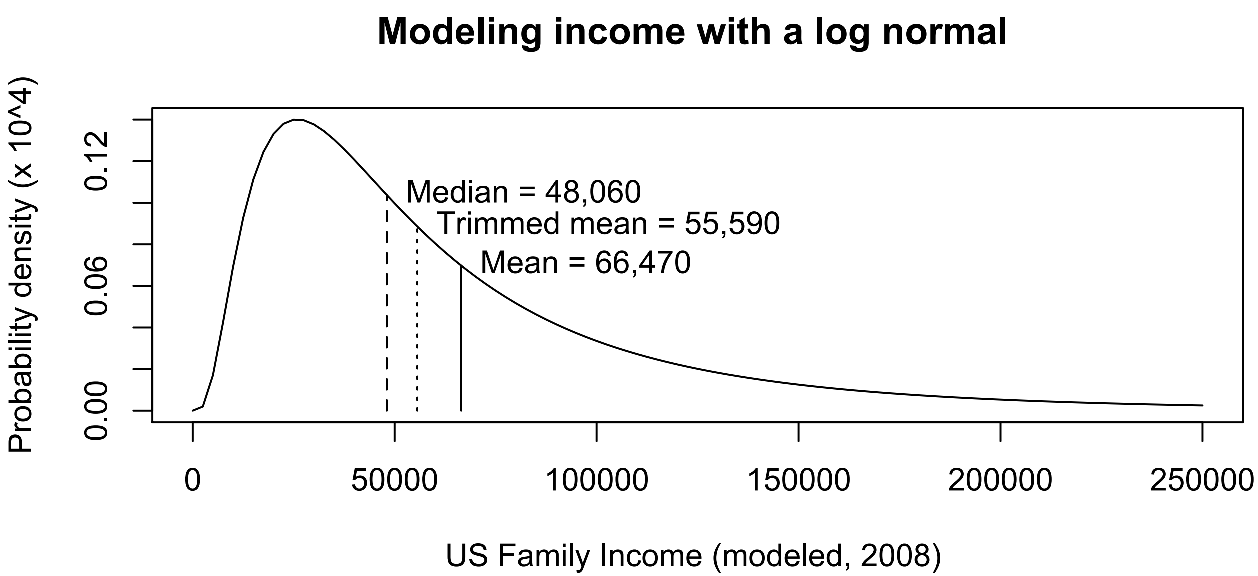
Raw and log transformed which group is “bigger”?

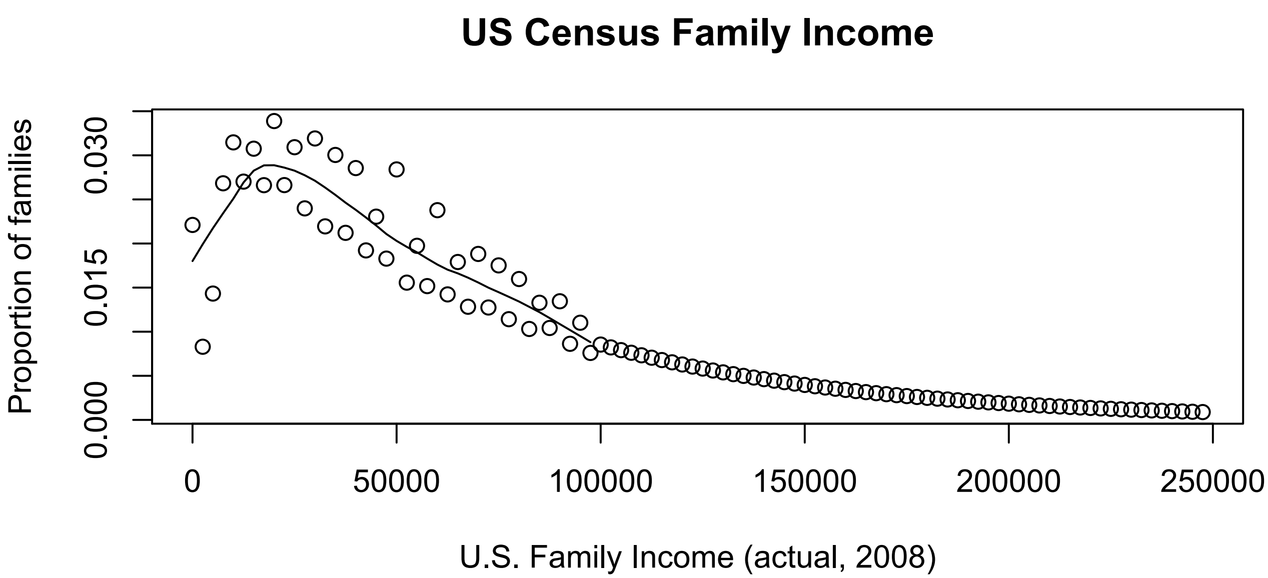
	X	Y	Log(X)	Log(Y)
Min	1	10	0	2.30
1st Q.	3	11.5	1.04	2.44
Median	8	13	2.08	2.57
Mean	18.1	13	2.08	2.26
3rd Q.	24	14.5	3.12	2.67
Max	64	16	4.16	2.77

The effect of a transform on means and medians









Income and Reaction Time are log normal

Estimating central tendencies

- Although it seems easy to find a mean (or even a median) of a distribution, it is necessary to consider what is the distribution of interest.
- Consider the problem of the average length of psychotherapy or the average size of a class at NU.

Estimating the mean time

- A therapist has 20 patients, 19 of whom have been in therapy for 26-104 weeks (median, 52 weeks), 1 of whom has just had their first appointment. Assuming this is her typical load, what is the average time patients are in therapy?
- Is this the average for this therapist the same as the average for the patients seeking therapy?

Estimating the mean time of therapy

- 19 with average of 52 weeks, 1 for 1 week
  - Therapists average is (19*52+1*1)/20 = 49.5 weeks
  - Median is 52 (Therapist centric)
- But therapist sees 19 for 52 weeks and 52 for one week so the average length is
  - ((19*52)+(52*1))/(19+52) = 14.6 weeks
  - Median is 1 (Patient centric)

Estimating Class size

5 faculty members teach 20 courses with the following distribution: What is the average class size?

Faculty member/ course #	100	200	300	400	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating class size

- What is the average class size?
- If each student takes 4 courses, what is the average class size from the students' point of view?
- Department point of view: average is 50 students/class

N	Size
10	10
5	20
4	100
1	400

Estimating Class size

Faculty member/ course #	100	200	300	400	average
1	10	20	10	10	12.5
2	10	20	10	10	12.5
3	10	20	10	10	12.5
4	100	20	20	10	37.5
5	400	100	100	100	175
department	106	36	30	28	50

Estimating Class size (student weighted)

Faculty member/ course #	100	200	300	400	average
1	10	20	10	10	14
2	10	20	10	10	14
3	10	20	10	10	14
4	100	20	20	10	73
5	400	100	100	100	271
Student	321	64	71	74	203

Estimating class size

Department perspective:

20 courses, 1000 students => average = 50

Student perspective: 1000 students enroll in classes with an average size of 203!

Faculty perspective: chair tells prospective faculty members that median faculty course size is 12.5, tells the dean that the average is 50 and tells parents that most upper division courses are small.

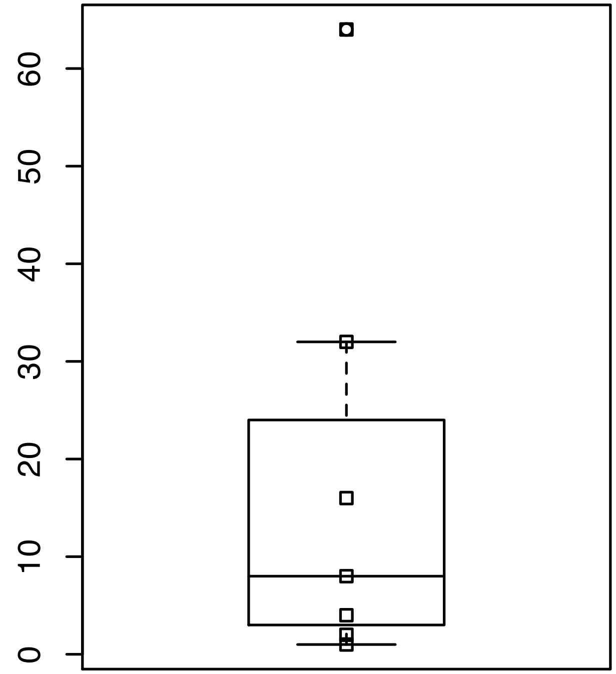
Which is the correct description?

Measures of dispersion

- Range (maximum - minimum)
- Interquartile range (75% - 25%)
- Deviation score $x_i = X_i$-Mean
- Median absolute deviation from median
- Variance = $\sum x_i^2/(N-1)$ = mean square
- Standard deviation sqrt (variance) =sqrt($\sum x_i^2/(N-1)$)

Robust measures of dispersion

- The 5-7 numbers of a box plot
- Max
- Top Whisker
- Top quartile (hinge)
- Median
- Bottom Quartile (hinge)
- Bottom Whisker
- Minimum



Transformations of scores

- Why transform?
  - to make easier to understand
  - to remove unnecessary detail
- Types of transformations
  - Add/subtract a constant $X' = X + C$
    - changes the mean but not the variance
    - $X'. = X. + C$ but $Var(X') = Var(X)$
  - Multiply by a constant $X' = XC$
    - changes the mean and the variance
    - $X'. = CX.$ and $Var(X') = C^2X$

Raw scores, Deviation

- Raw score for i_{th} individual X_i
  – (original units)
- Deviation score $x_i = X_i - \text{Mean } X$
  – (original units but the mean is now 0)
- Standard score = x_i / s_x
  – Variance of standard scores = 1

Distributions of sample means

- The problem: take samples of size n from an infinite (or at least very large) population
- What is the distribution of these sample means?
- What is the variance of the sample means

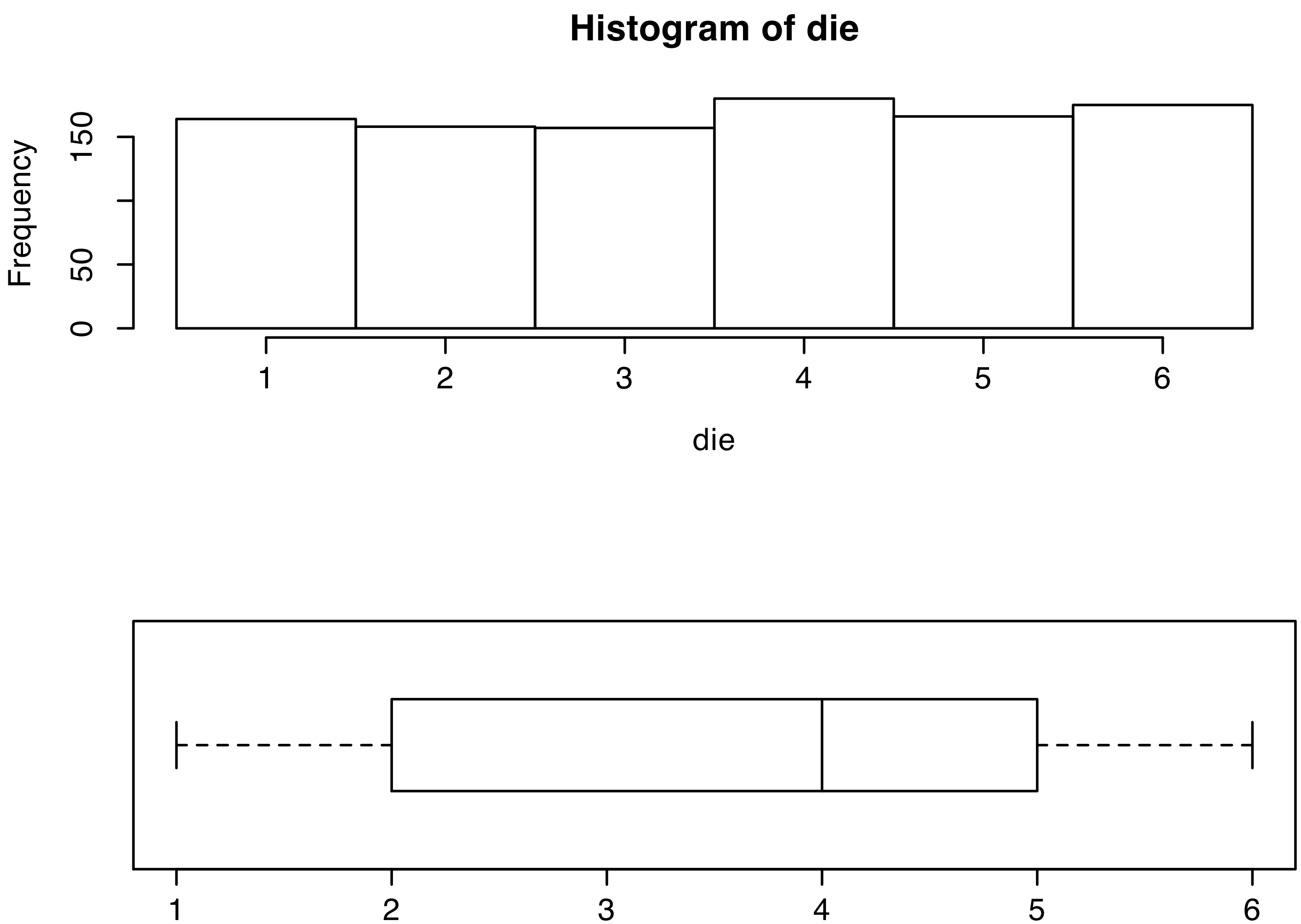
Central Limit Theorem

- Independent samples from a distribution with mean μ and standard deviation σ will tend towards being distributed with mean = μ and a standard deviation of σ /sqrt(n).
- Note that this is true for any distribution with finite μ and σ

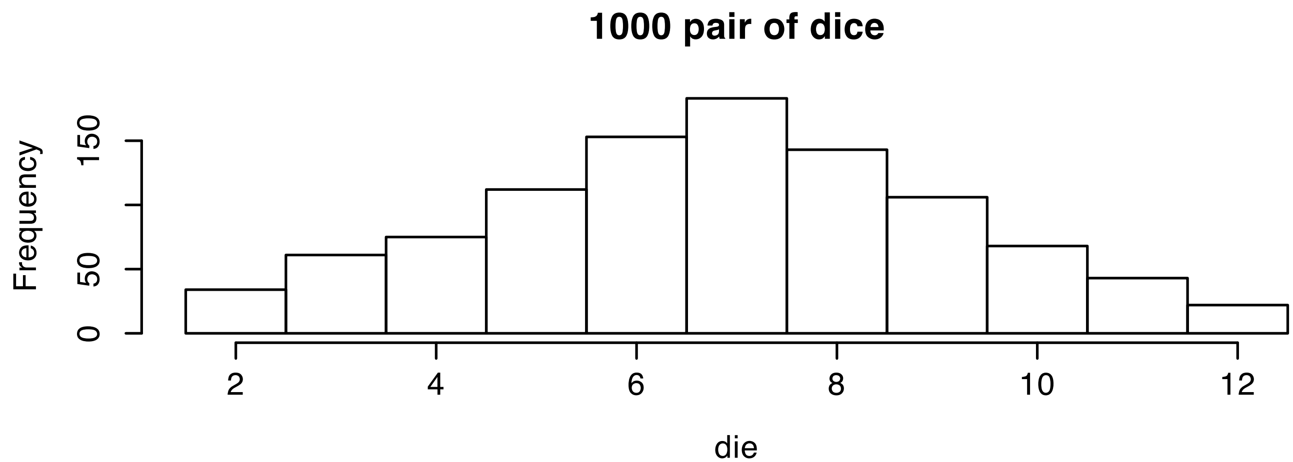
Consider the distribution of 1 die

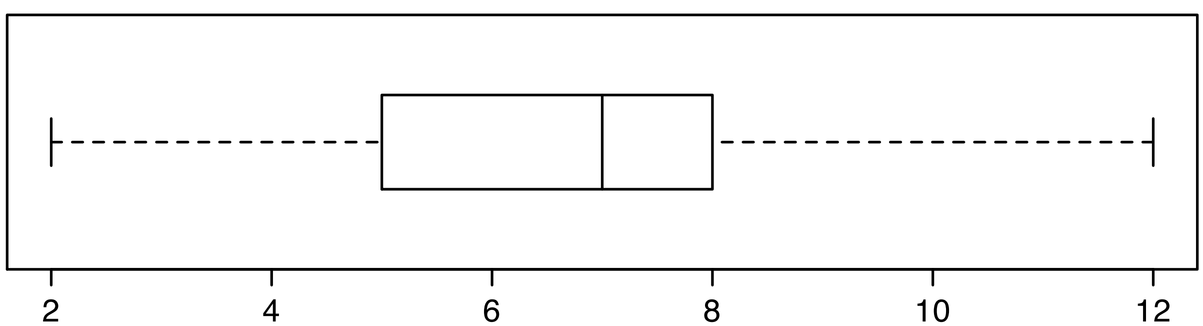
- A single, 6 side die will produce a uniform distribution of numbers from 1-6. That is to say, each number is equally likely to occur.

1000 throws of a single die



Distribution of a pair

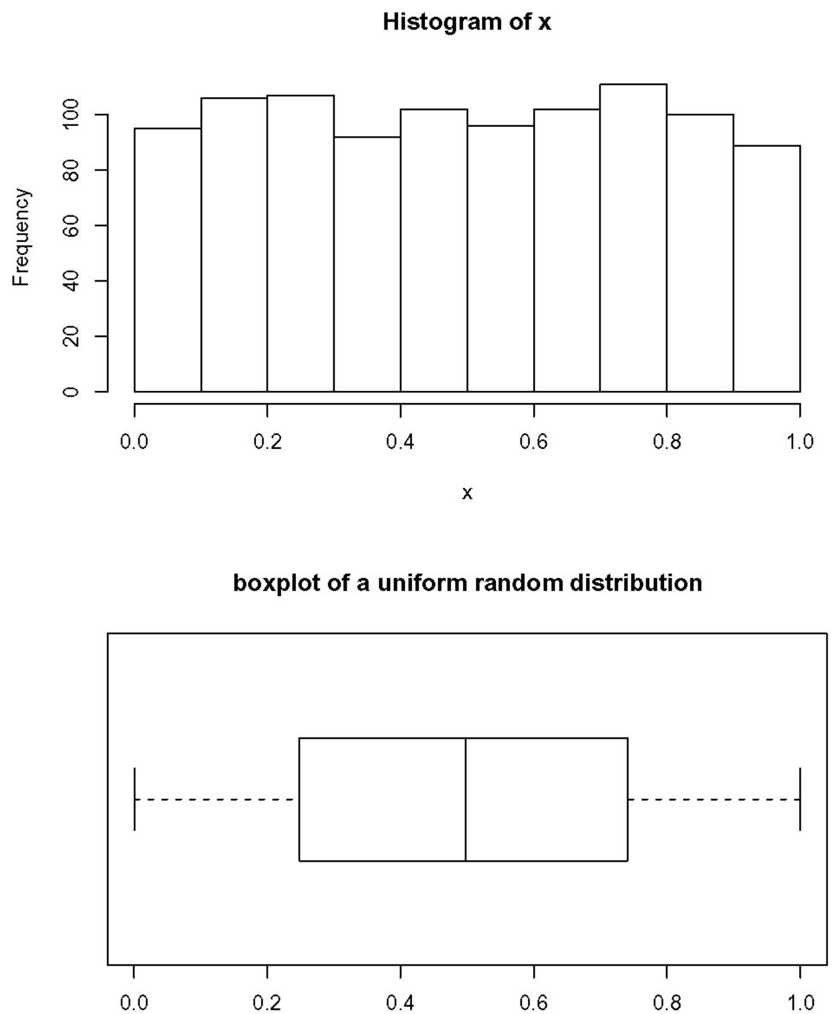




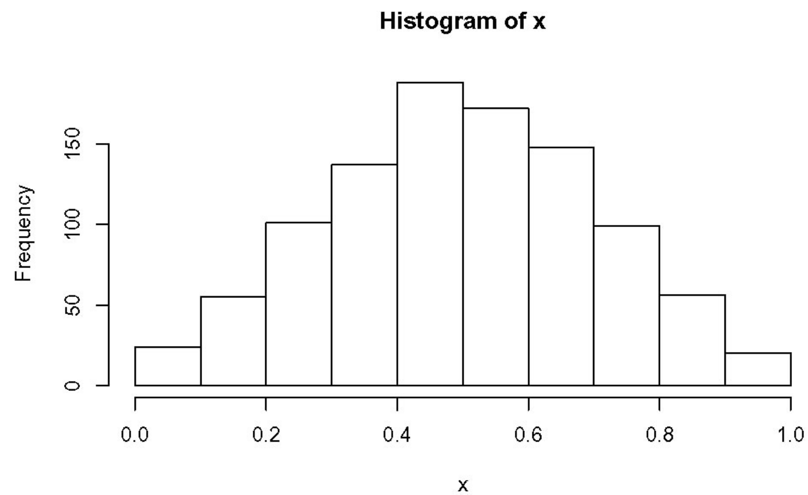
Further demonstrations of CLT

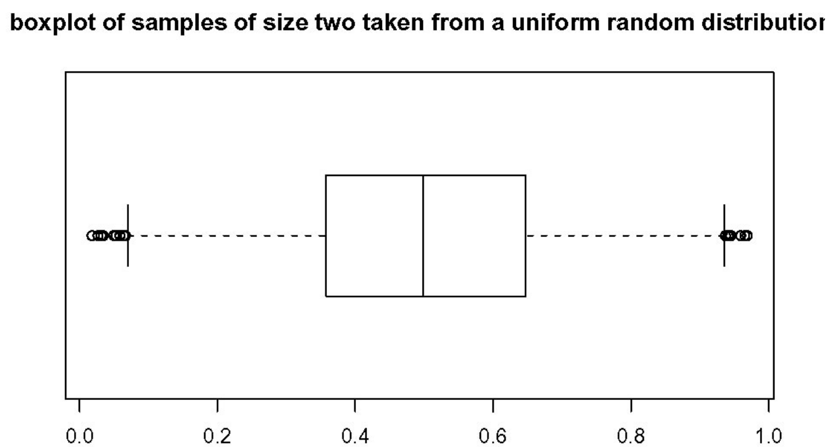
- Consider a rectangular (uniform) distribution ranging from 0-1
- Take 1000 samples of size n from this distribution
- For n=1, the shape will approximate the shape of the underlying distribution
- But as n-> large, the shape will tend towards the normal

1000 samples of size 1

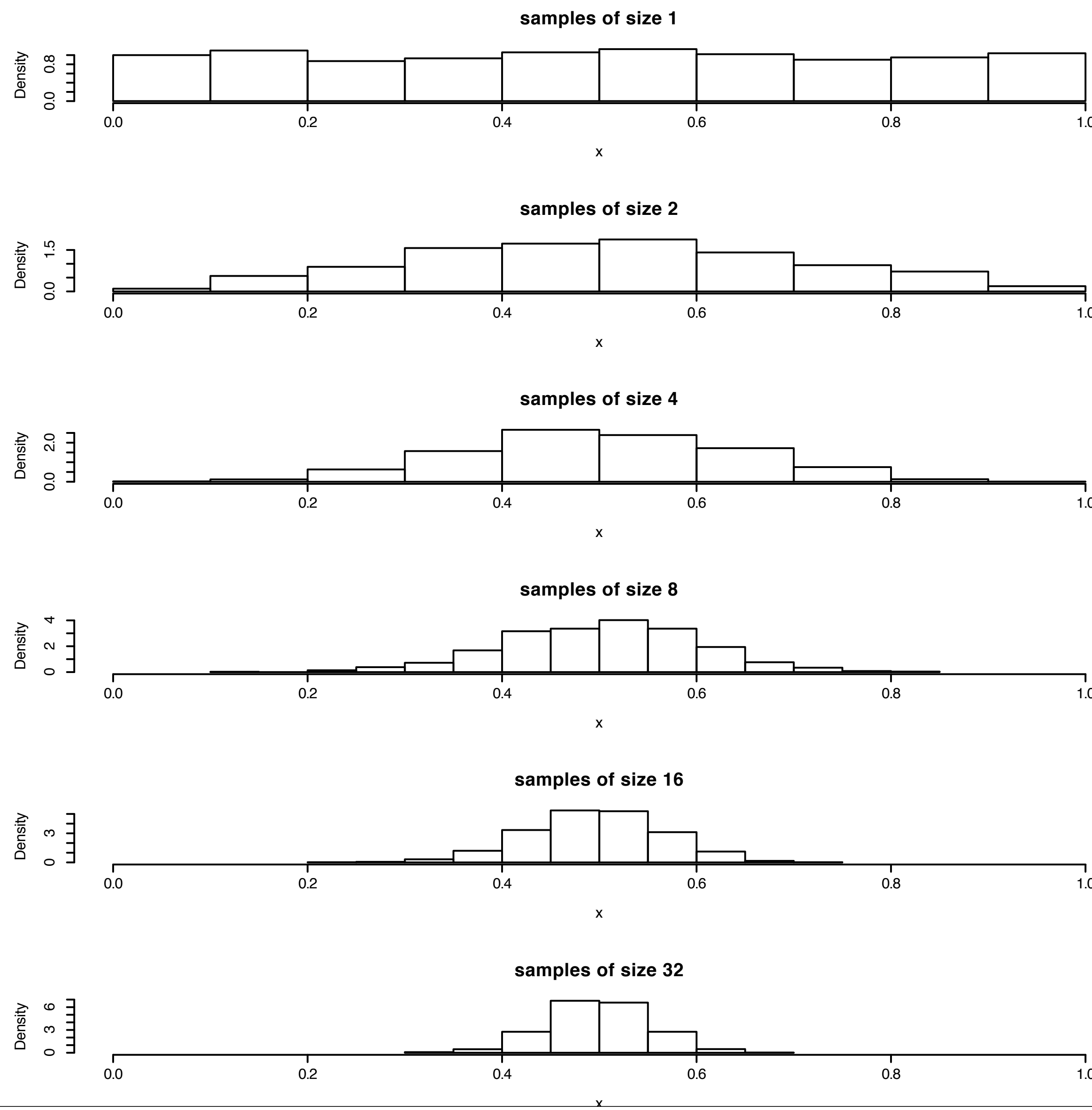


1000 samples of size 2

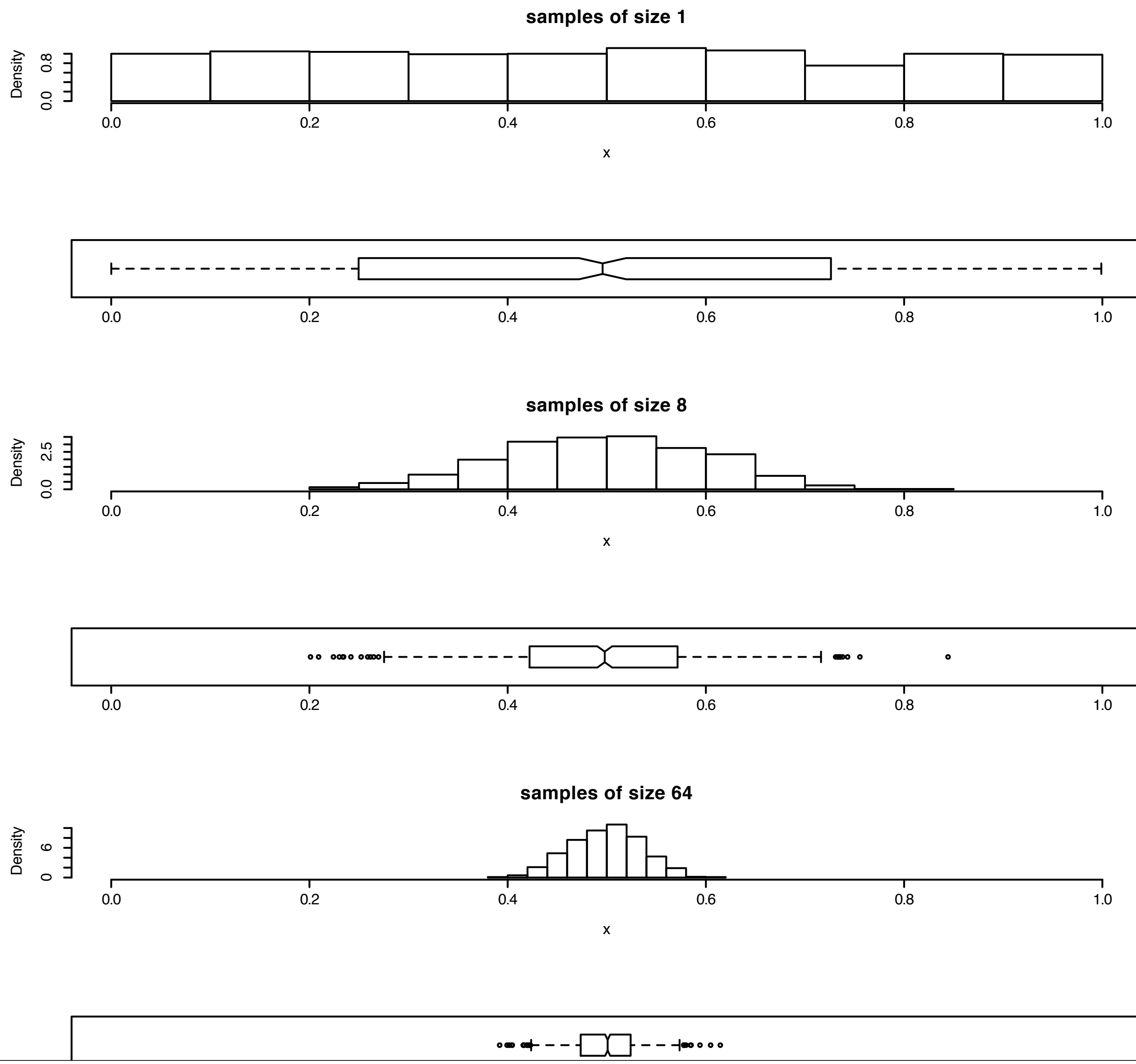




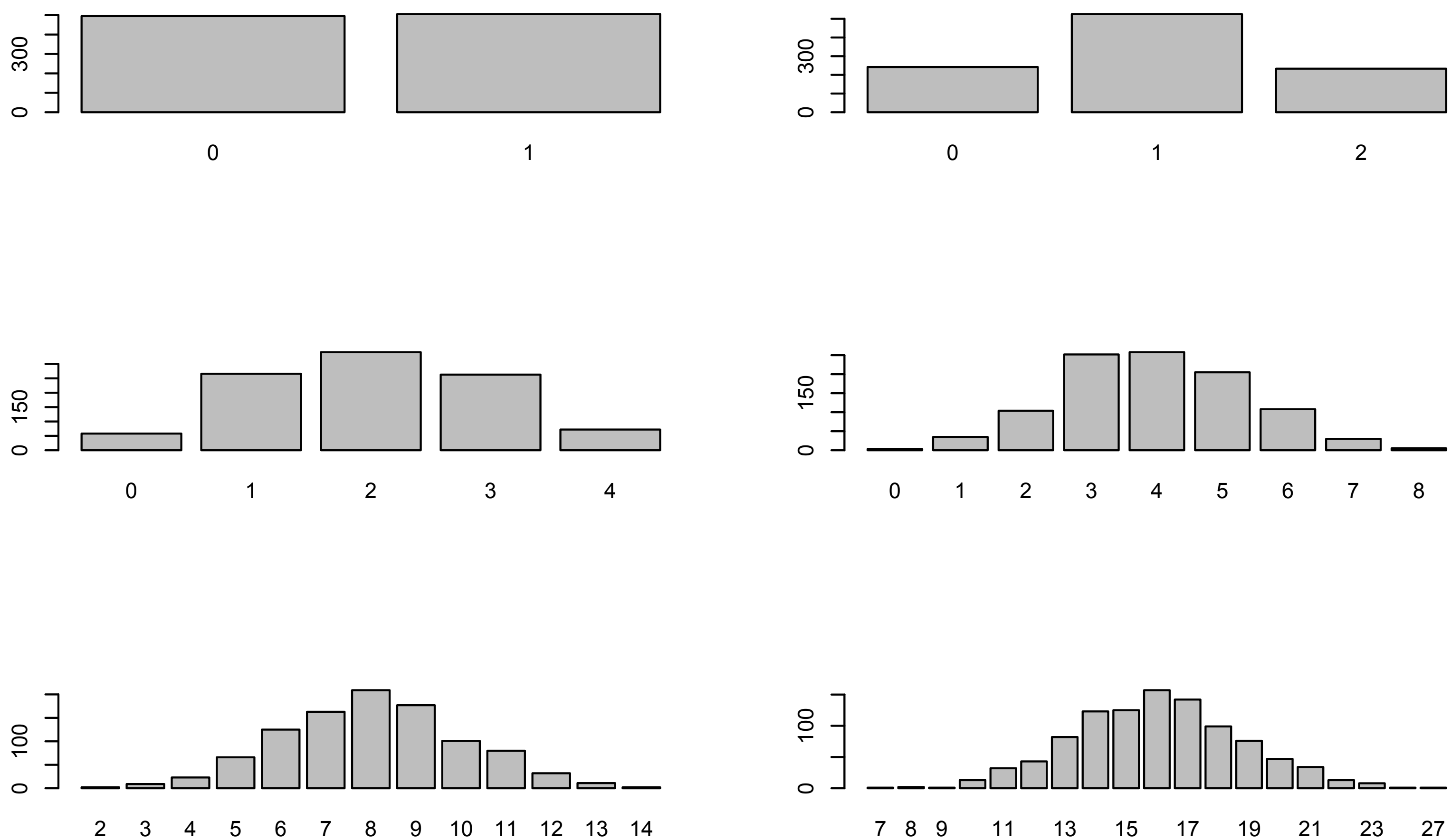
Distributions as f(sample size)



Histograms and box plots



Samples from the binomial (p=.5)

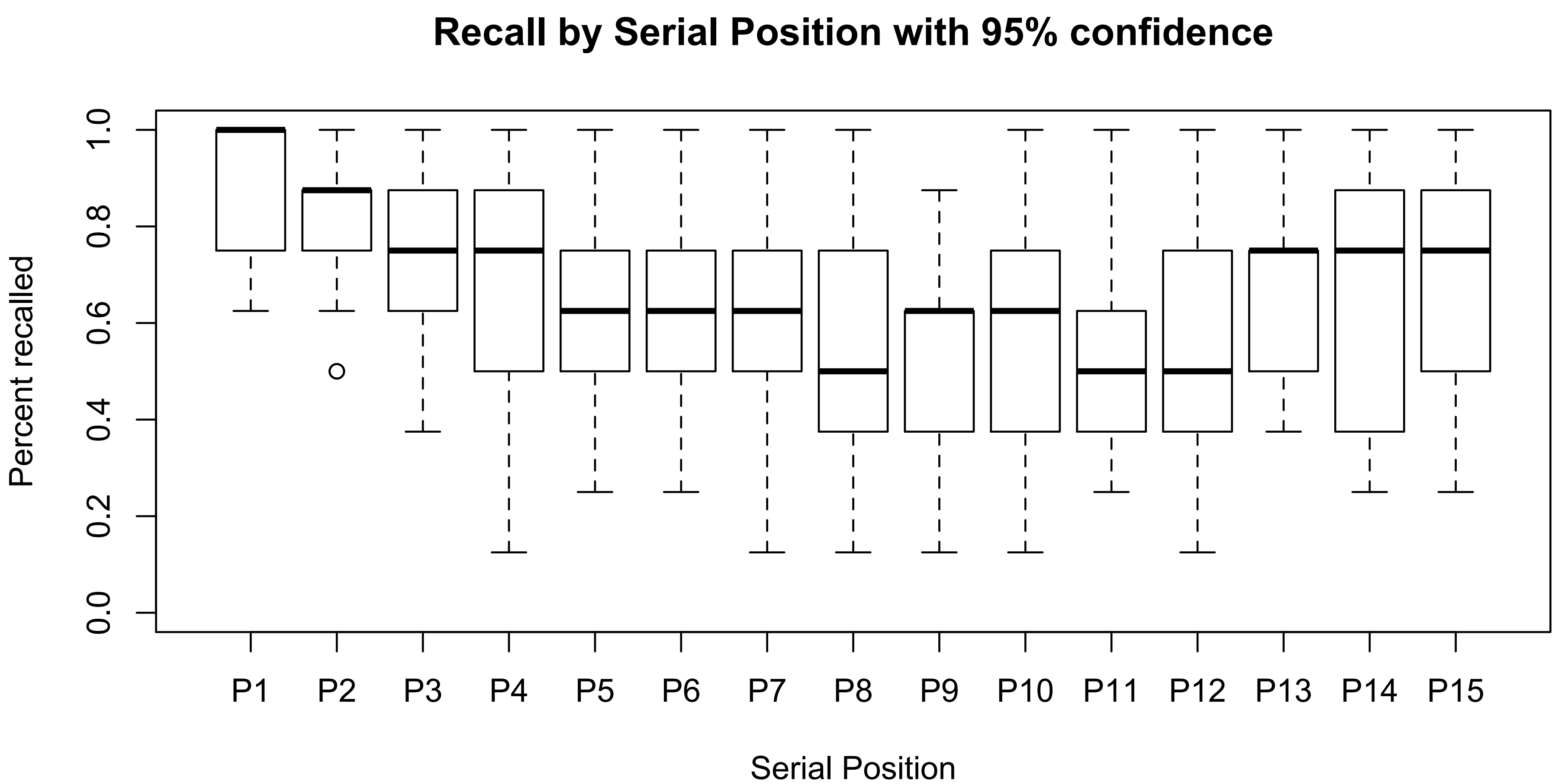


barplot(table(rbinom(1000,16,.5)))

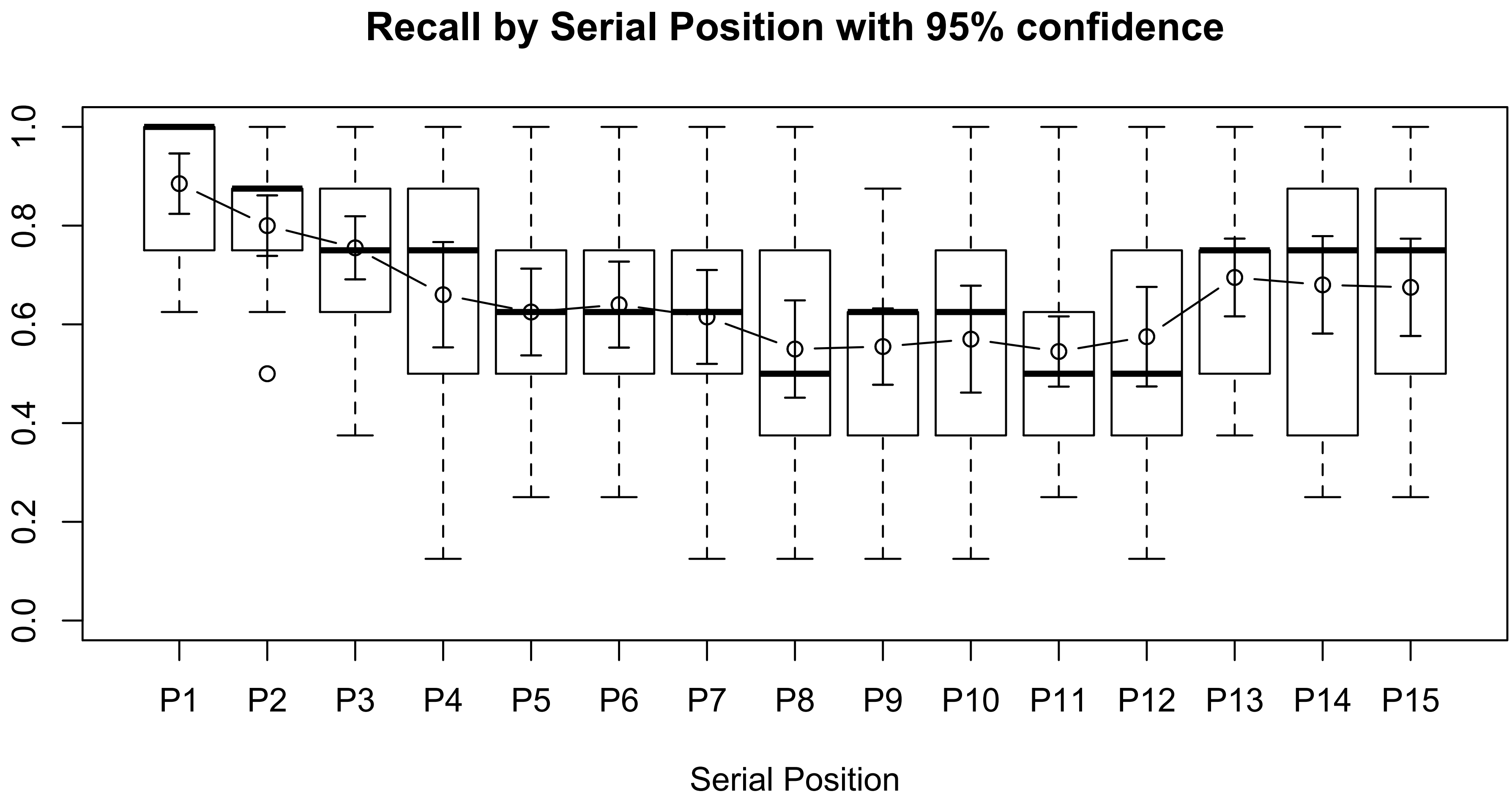
Data= Model + Error

- The process of science is improve the model and reduce the error
- Models are progressively more complicated
- Consider the recall data:
  - –Model 1: Data = Mean + Error
  - –Model 2: Data = Position_i + Error

Graphic descriptions of data



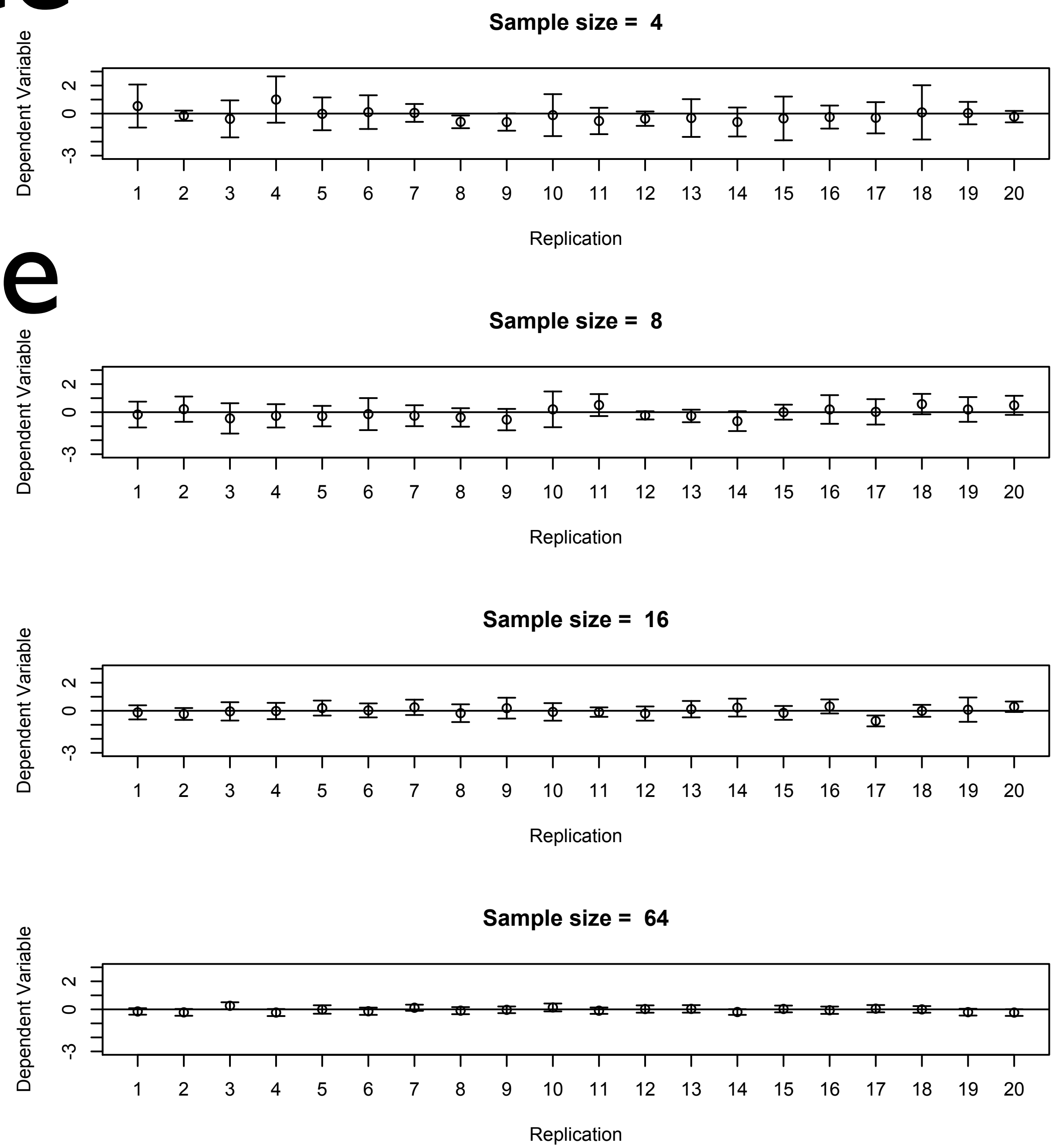
Showing variability multiple ways



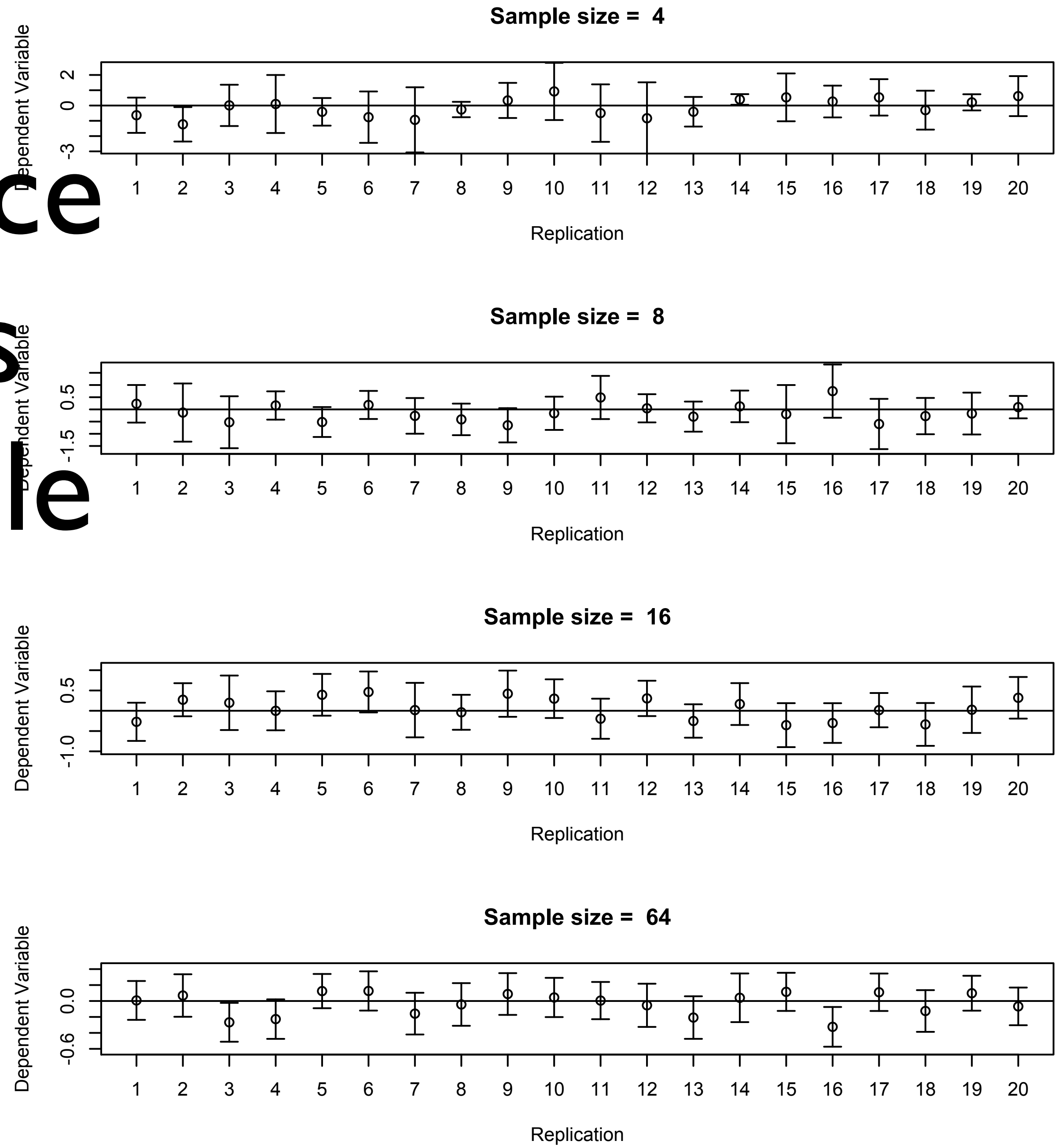
Confidence intervals and sample size

- Confidence intervals reflect the sample standard deviation and the sample size.
- They give a range of likely possibilities of real (but unobserved) mean given the data

Confidence intervals and sample size



Confidence intervals and sample size



Descriptive stats

- Find the central tendency
  - Mean or median
- Show the dispersion of data
  - Standard Deviation, IQR
- Confidence interval of central tendency

Confidence Intervals II

- 95% Confidence Interval is based upon normal theory and is ± 1.96 standard errors
- Alternative is empirical confidence interval estimated by resampling the data