Research Methods: More on basic statistics

Dolphins and fishermen

- A recent case in the news reports how a fisherman fell off his boat but was rescued when a dolphin pushed him to shore. Several other fishermen confirmed that this happened to them as well.
- From these stories, should we conclude that dolphin are friendly to humans and help them when they are in distress?
- What piece of evidence is missing from these stories?

The most dangerous profession

- Wainer (1997) reviews data from "the Swiss physician H.C. Lombard who examined 8,496 death certificates gathered over a half century in Geneva. Each certificate contained the name of the deceased, his profession, and age at death. Lombard used these data to calculate the mean longevity associated with each profession."
- Consider the following (abbreviated) table.

- H. Wainer (1997) The most dangerous profession: a note on nonsampling error. Psychological Methods, 4, 250-256.

The most dangerous profession?

(Wainer, 1997: data from Lombard, 1835)

Profession	Total number of deaths	Average age
Farmers	267	54.7
Lawyers	12	64.3
Apothecaries	19	64.3
Businessmen	7	57.5
Butchers	77	53
Bakers	82	49.8
Harness Makers	10	60.4
Surgeons	41	54
Coachmen	12	45
Merchant assistants	58	38.9
Wine merchant	120	56.3
School masters	18	64.4
Professors	10	66.6
Soldiers	338	48.4
Students	39	20.2

Why is being a student so dangerous? Why being a professor so safe?

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)

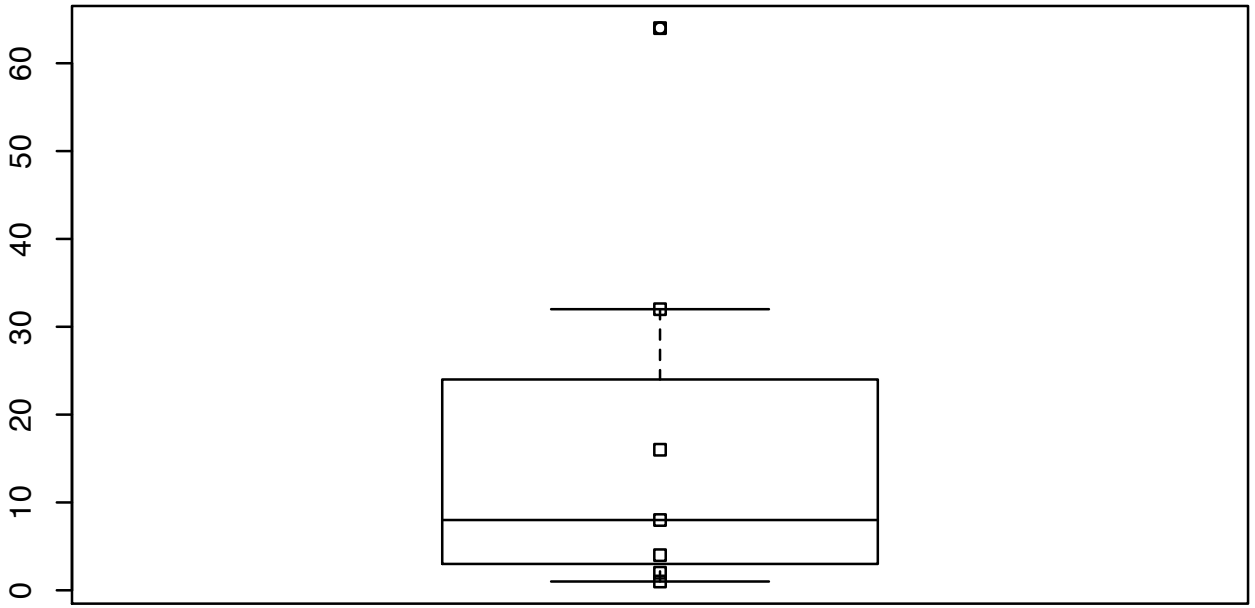
- Harmonic Mean = reciprocal of average reciprocal
  - $1/(\sum (1/x_i)/N)$

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
> stripchart(x,vertical=T,add=T)  #add in the points
```



Consider two sets, which is more?

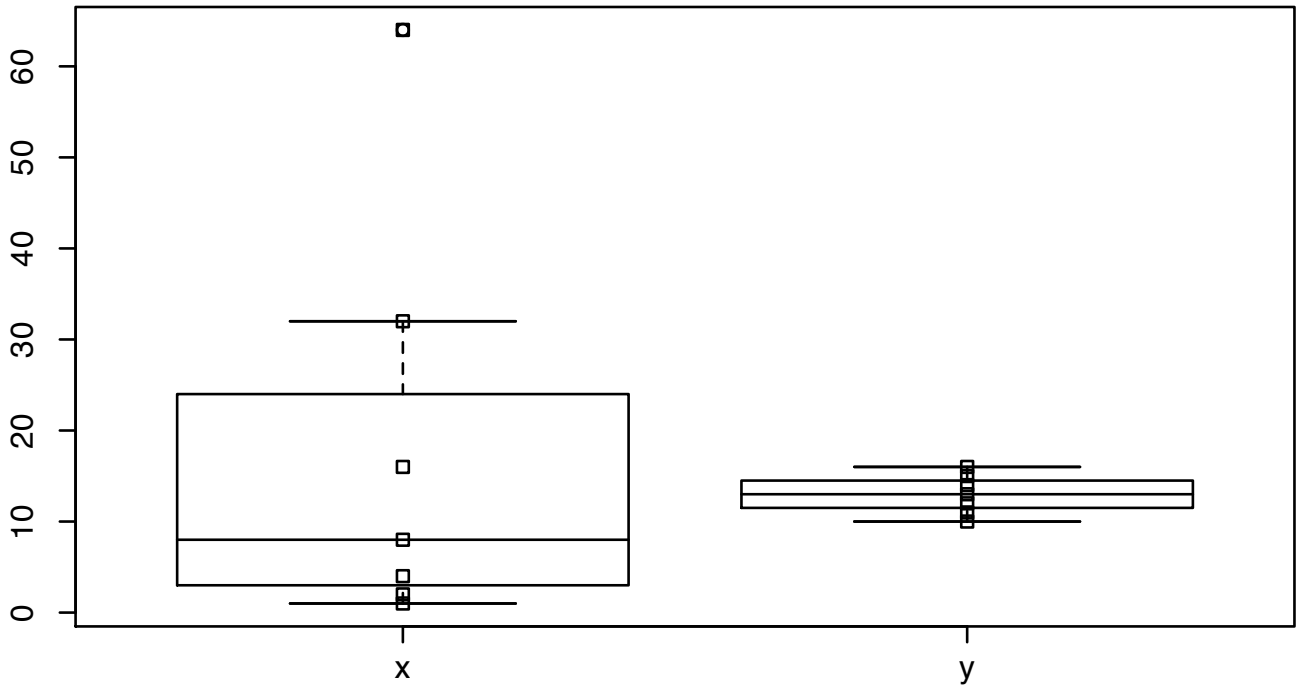
subject	Set 1	Set 2
1	1	10
2	2	11
3	4	12
4	8	13
5	16	14
6	32	15
7	64	16
median	8	13
arithmetic	18.1	13.0
geometric	8.0	12.8
harmonic	3.5	12.7

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



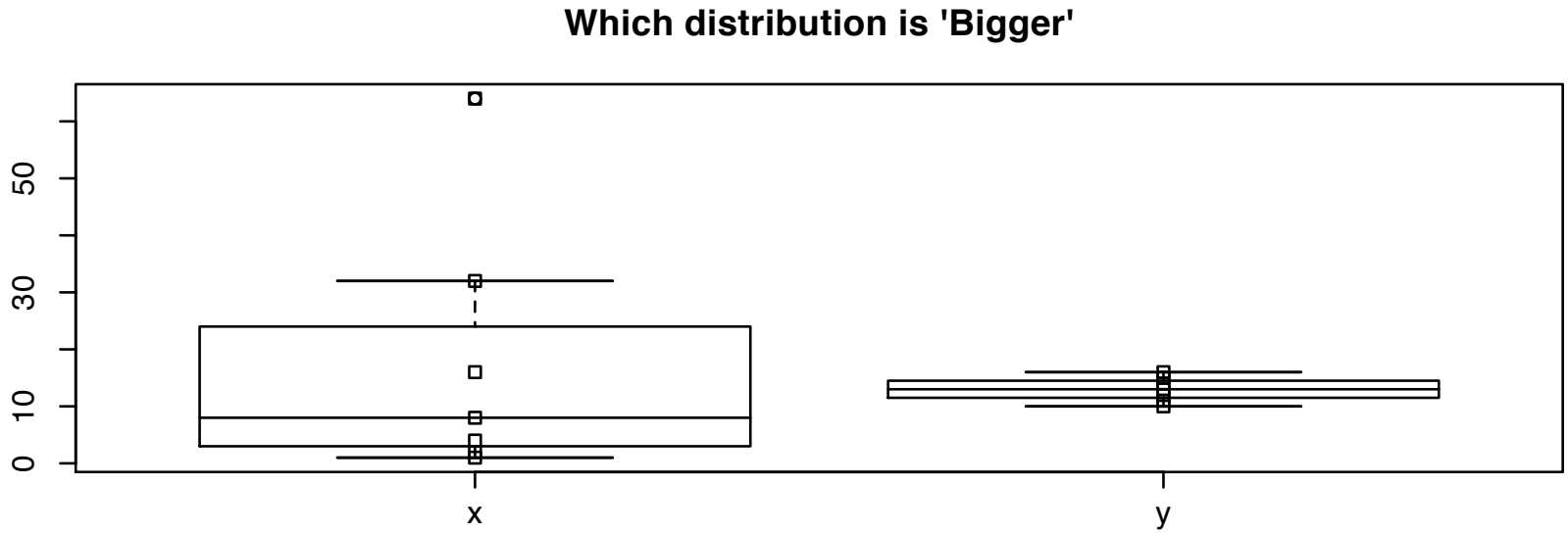
The effect of log transforms
Which group is "more"?

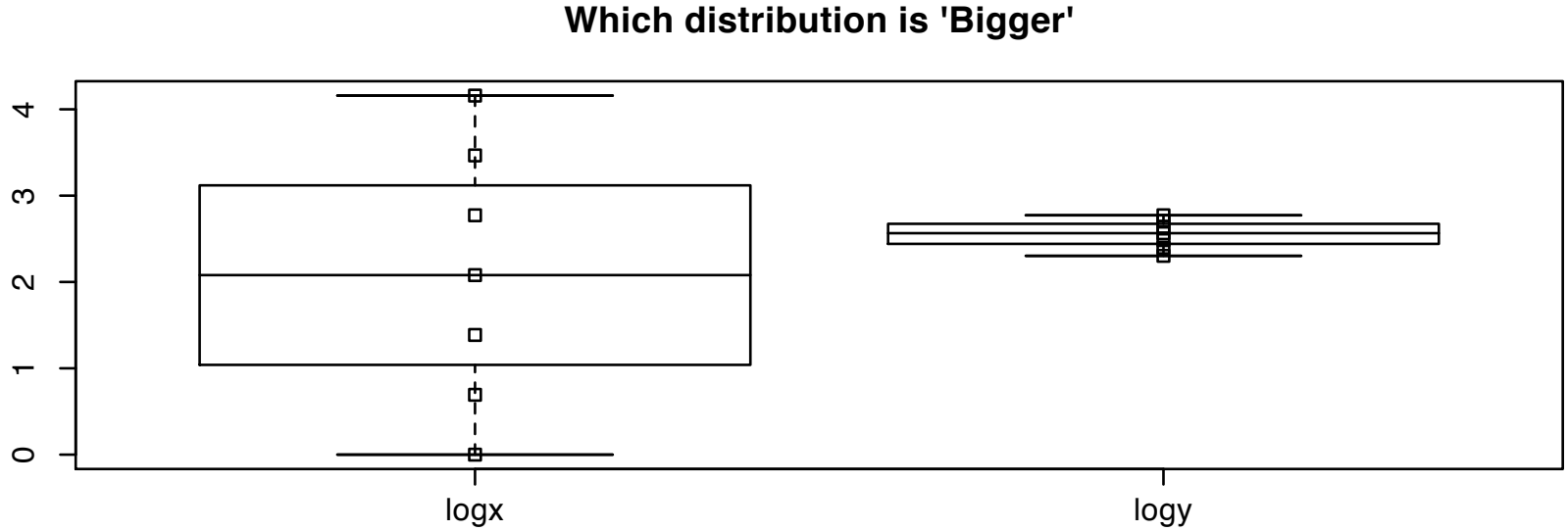
X	Y	Log X	Log Y
1	10	0.0	2.3
2	11	0.7	2.4
4	12	1.4	2.5
8	13	2.1	2.6
16	14	2.8	2.9
32	15	3.5	2.7
64	16	4.2	2.8

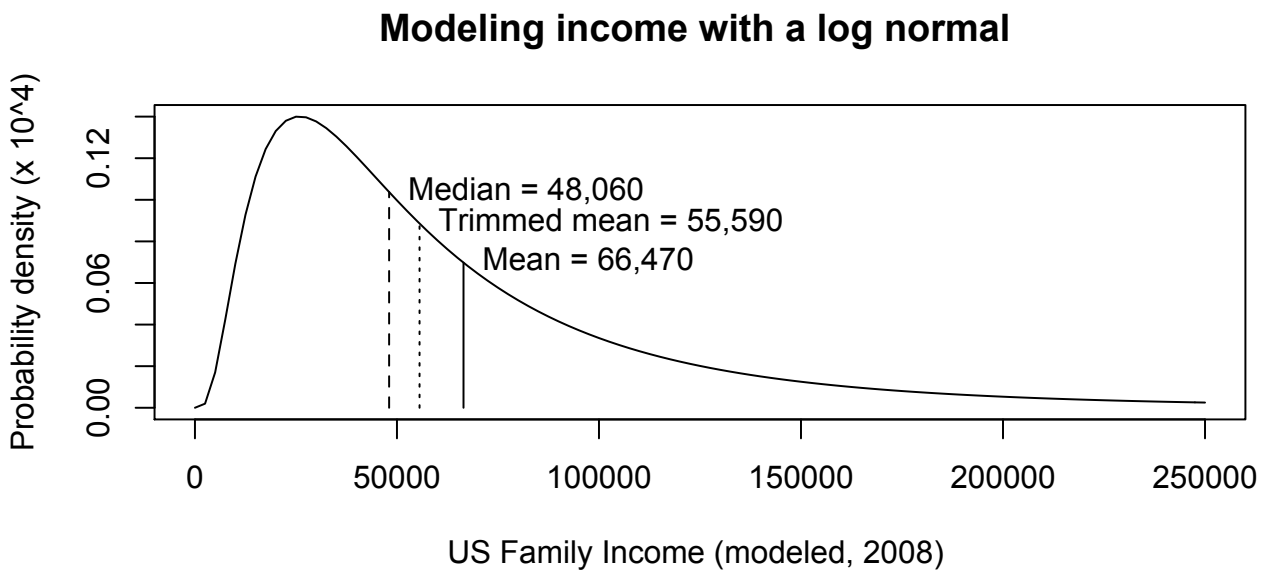
Raw and log transformed which group is "bigger"?

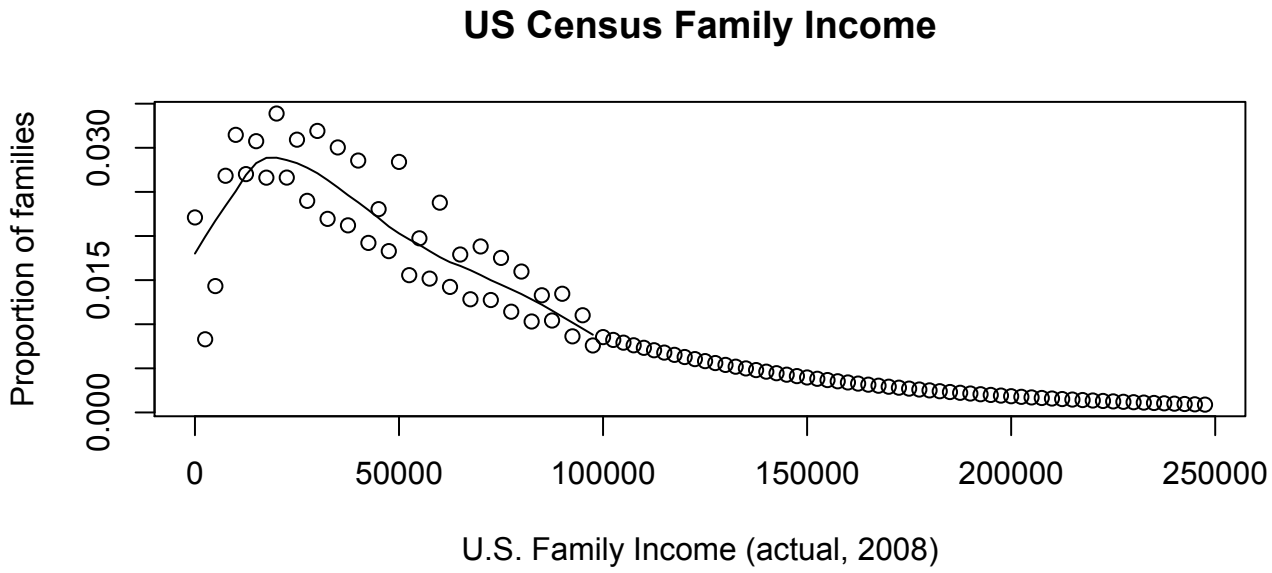
	X	Y	Log(X)	Log(Y)
Min	1	10	0	2.30
1st Q.	3	11.5	1.04	2.44
Median	8	13	2.08	2.57
Mean	18.1	13	2.08	2.26
3rd Q.	24	14.5	3.12	2.67
Max	64	16	4.16	2.77

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?

Airline passengers

- The average American flies about once a year, and about 79% fly no more than twice a year.
- Airlines give special attention to their very frequent ("elite") flyers (those in the top 3%).
- What percentage of the flyers on a plane are "elite" flyers?

Hypothetical data on airplane passengers: 1/3 on the plane are in the top 3% or "elite" flyers

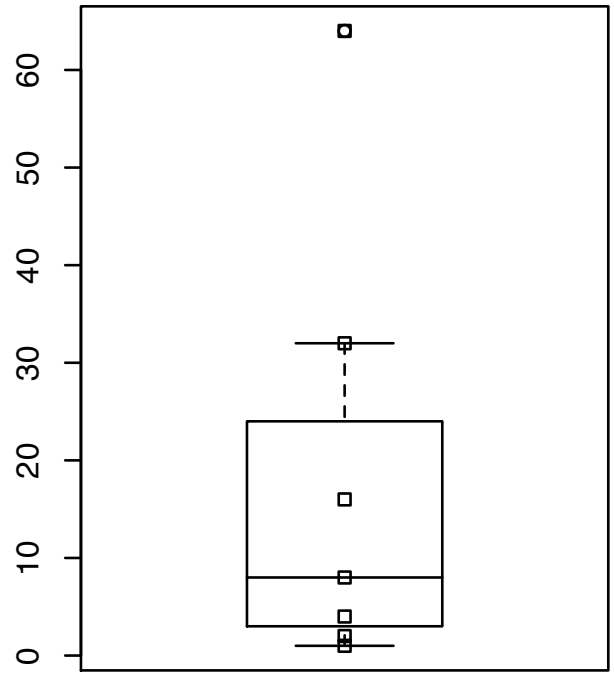
Flying	Number	%		p	p*N	fraction of plane
1/year	365	0.53	0.53	0.00	1.00	0.17
2/year	182	0.26	0.79	0.01	1.00	0.17
4/year	91	0.13	0.92	0.01	1.00	0.17
1/	30	0.04	0.97	0.03	0.99	0.16
2/	15	0.02	0.99	0.07	1.01	0.17
1/week	7	0.01	1.00	0.14	1.00	0.17
Total	690				5.99	1.00

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) $=\text{sqrt}(\sum x_i^2/(N-1))$

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 $x_i = X_i$-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

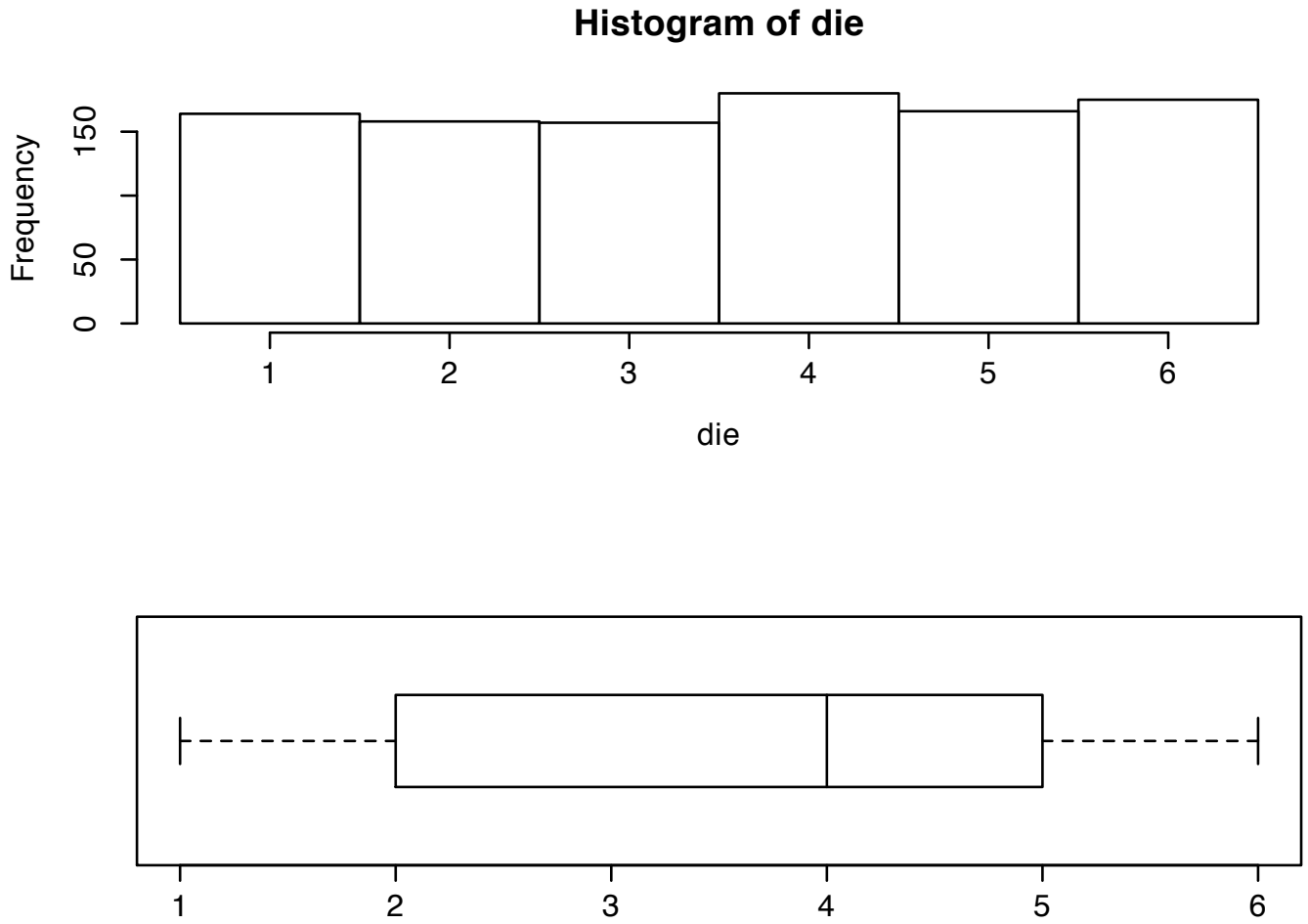
Central Limit Theorem

- 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

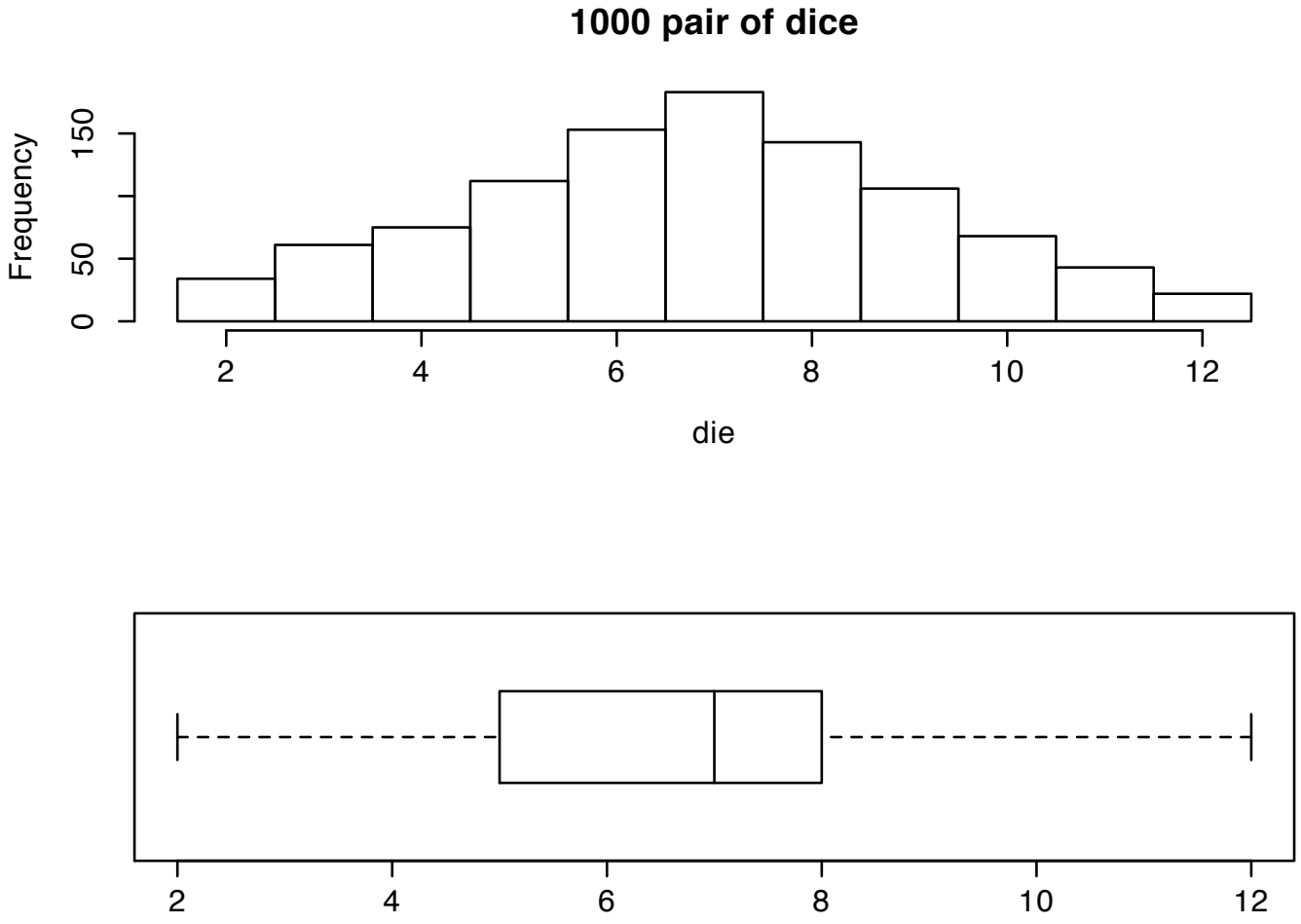
Consider the distribution of 1 die

- 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



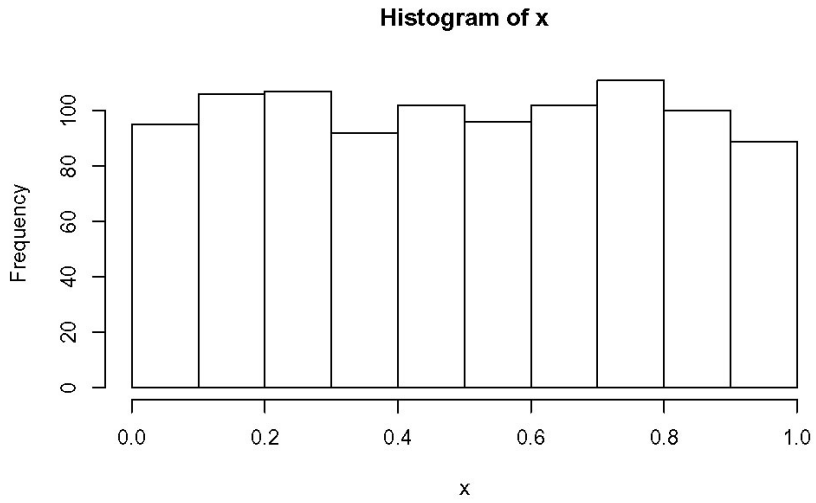
Distribution of a pair of

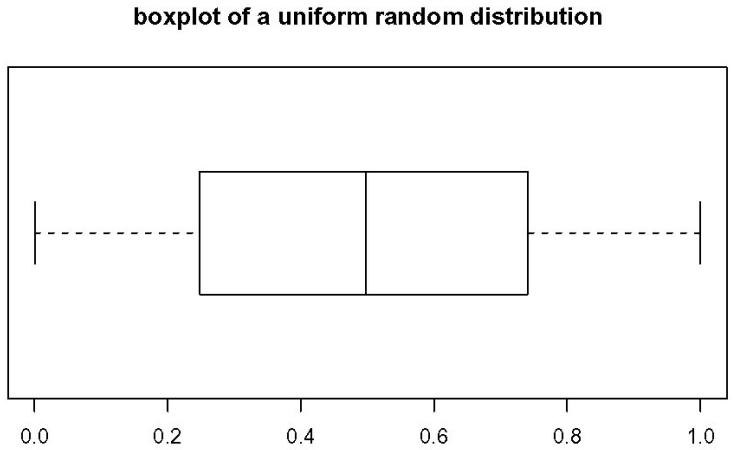


Further demonstrations of CLT

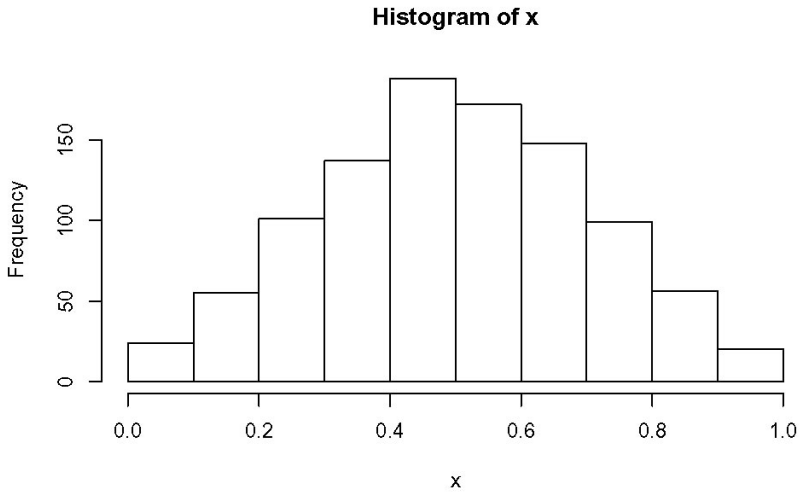
- Consider a rectangular (uniform) distribution ranging from 0-1
- Take 1000 samples of size n from this distribution
- For n=1, the shape will approximate the shape of the underlying distribution
- But as n-> large, the shape will tend towards the normal

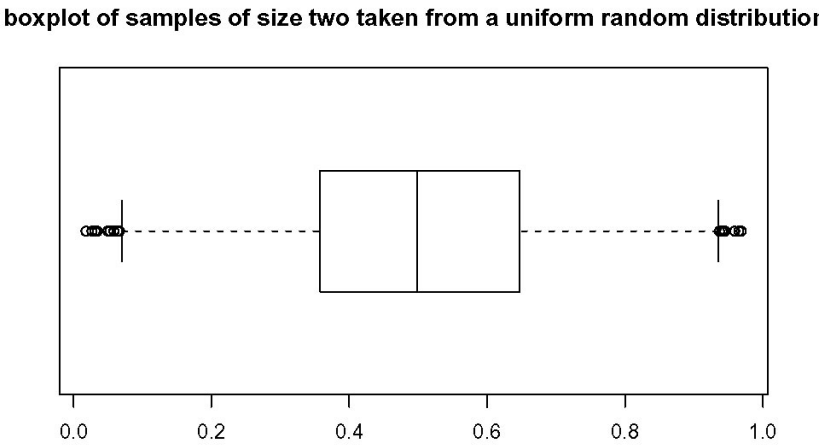
1000 samples of size 1



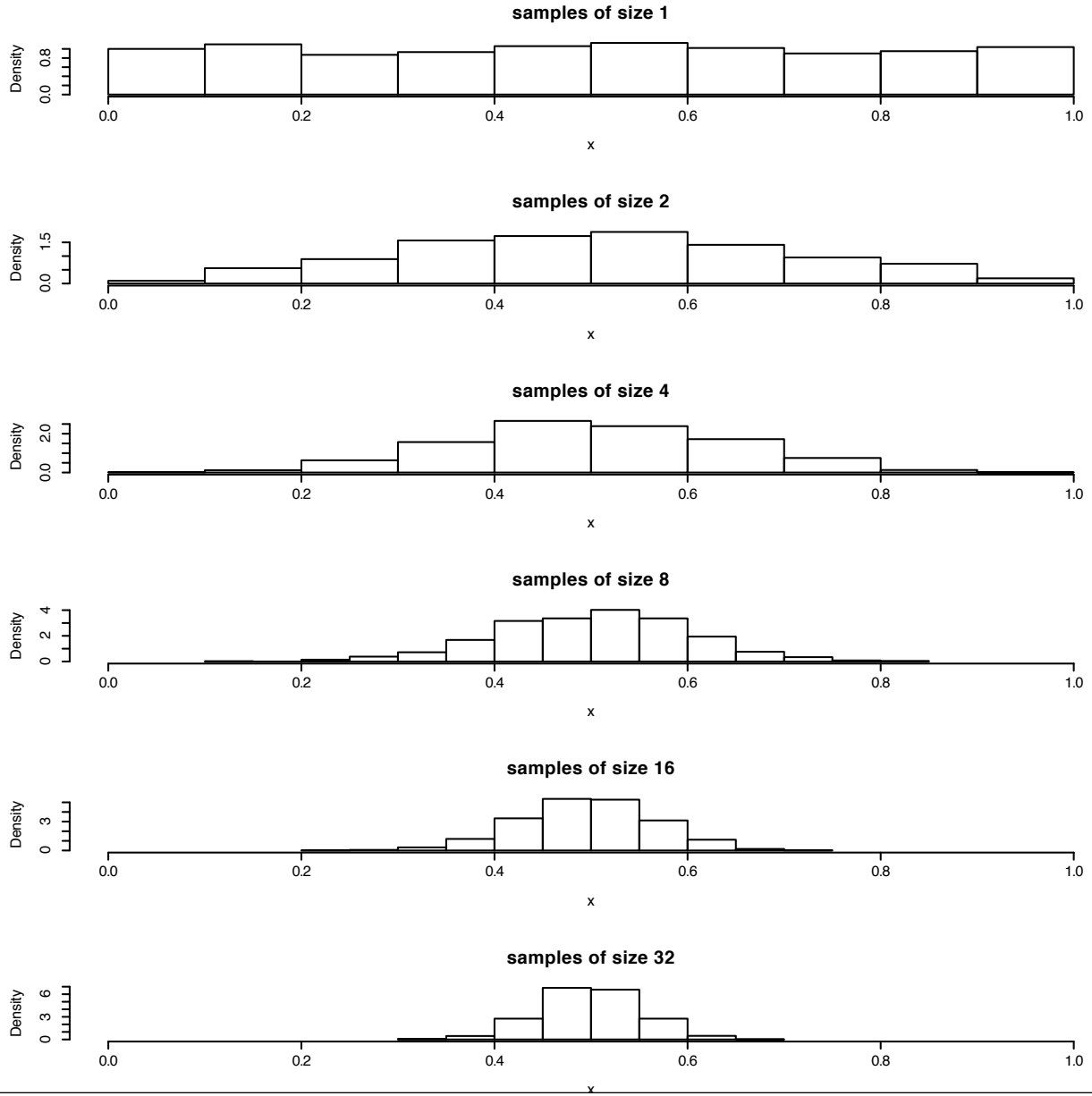


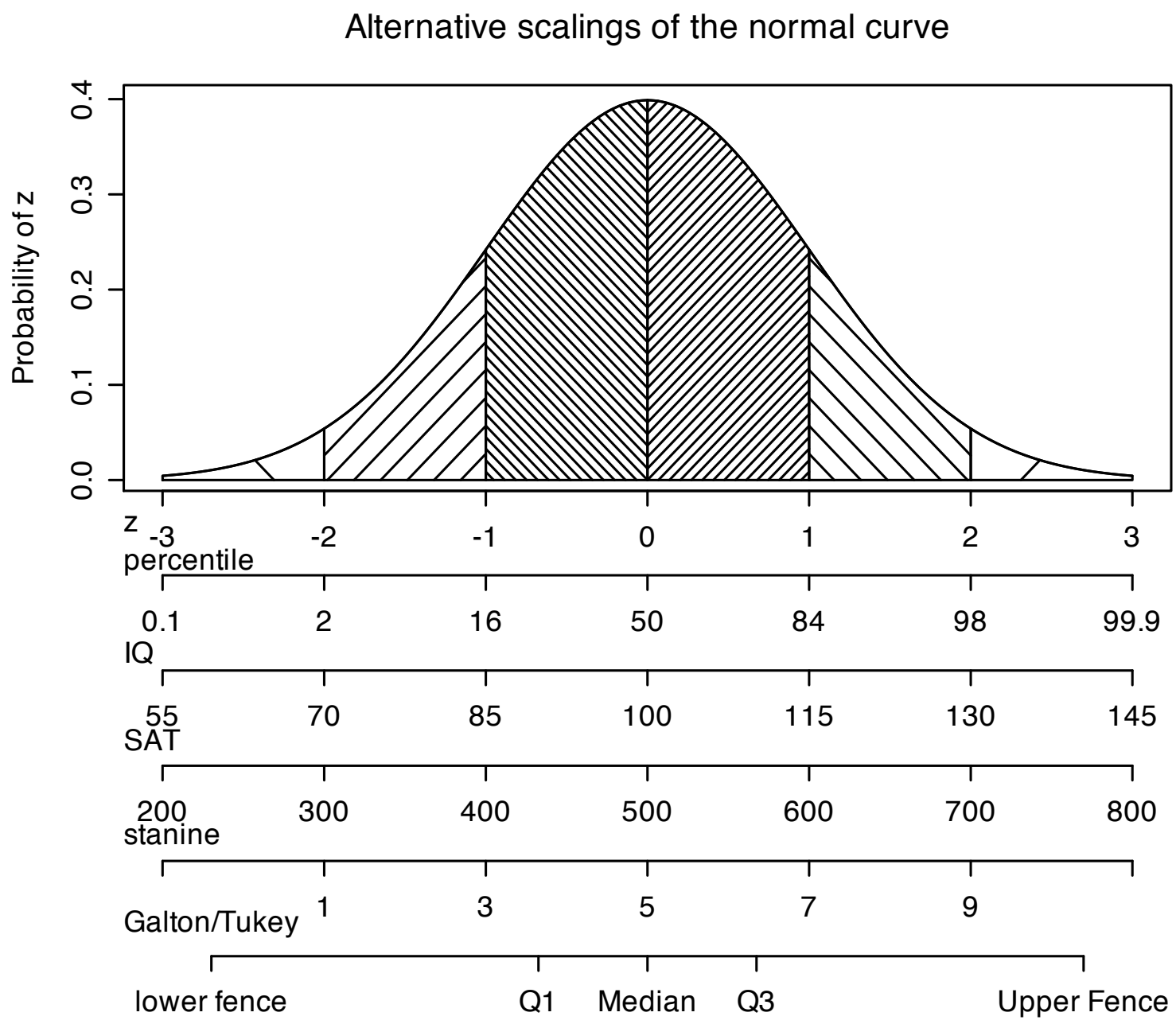
1000 samples of size 2





Distributions as f(sample size)




Alternative scalings of the normal curve
Probability of z
0.0
0.1
0.2
0.3
0.4
z
-3
-2
-1
0
1
2
3
percentile
0.1
2
16
50
84
98
99.9
IQ
55
70
85
100
115
130
145
SAT
200
300
400
500
600
700
800
stanine
1
3
5
7
9
Galton/Tukey
lower fence
Q1
Median
Q3
Upper Fence

Descriptive and Inferential Statistics

- Describe the data
  - Central Tendencies and Dispersion
  - Means, standard deviations
- Inferential -- the Null Hypothesis model
- How likely are the data given a model of no difference
  - consider the t-test

Multiple ways to model variance

- t test compares the difference of two groups
- F-test (ANOVA) is a generalization of t to compare multiple groups
- If the independent variable is categorical, then it can be thought of in terms of groups and we can use ANOVA
- If the independent variable is continuous, then we use the linear model.
- ANOVA is a special case of linear model

Recall and Recognition Hypothesis testing

- How likely would differences of this magnitude be observed if in fact there were no effect in the population.
- Null Hypothesis Test
  - H_0 The groups do not differ in the population
  - H_1 The groups come from different populations
  - How likely are the results if H_0?
  - What is the probability of data given H_0?
  - Reject H_0 if $p <$ critical value

Significance testing using Analysis of Variance

- ANOVA as a generalization of t-test.
  - t-test compares the difference between two means in terms of the expected standard deviation of the mean = observed standard deviation/sqrt(N-1)
- ANOVA compares the variance of the sample means to the variance within groups
- Possible to do ANOVA for multiple comparisons (combinations of variables)

Interpretation of ANOVA

- Each anova is a comparison of two estimates of the population variance:
  - an estimate from the variance between groups and an estimate from the variance within groups.
- F is the ratio of these estimates. If the two groups are random samples from the same population, we would expect the F ratio to be 1. The more the F deviates from 1, the less likely is the hypothesis that the samples came from the same population.

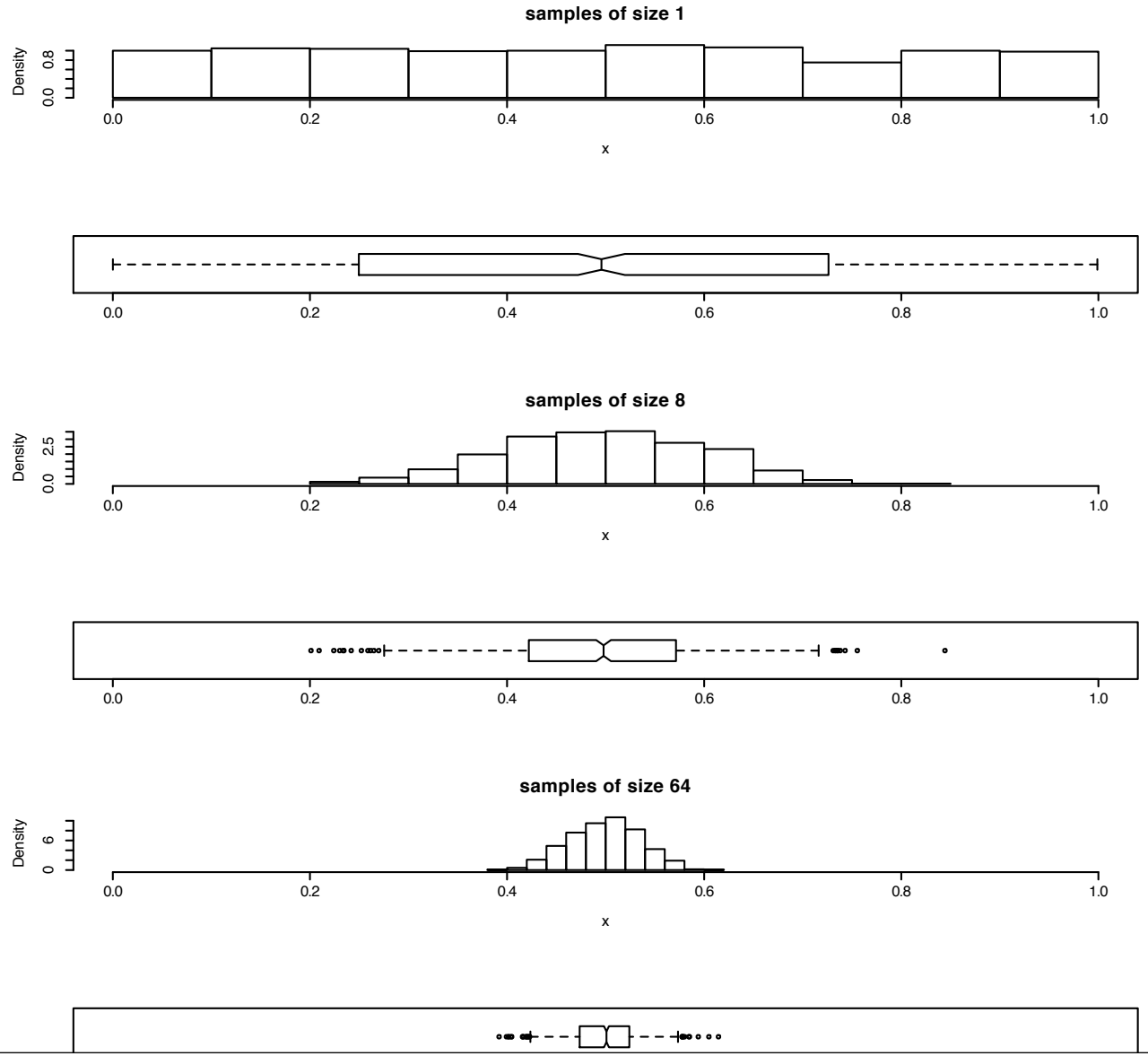
Alternative to hypothesis testing

- Effect size and confidence interval.
- How big is the effect and what is the expected range of the effect?

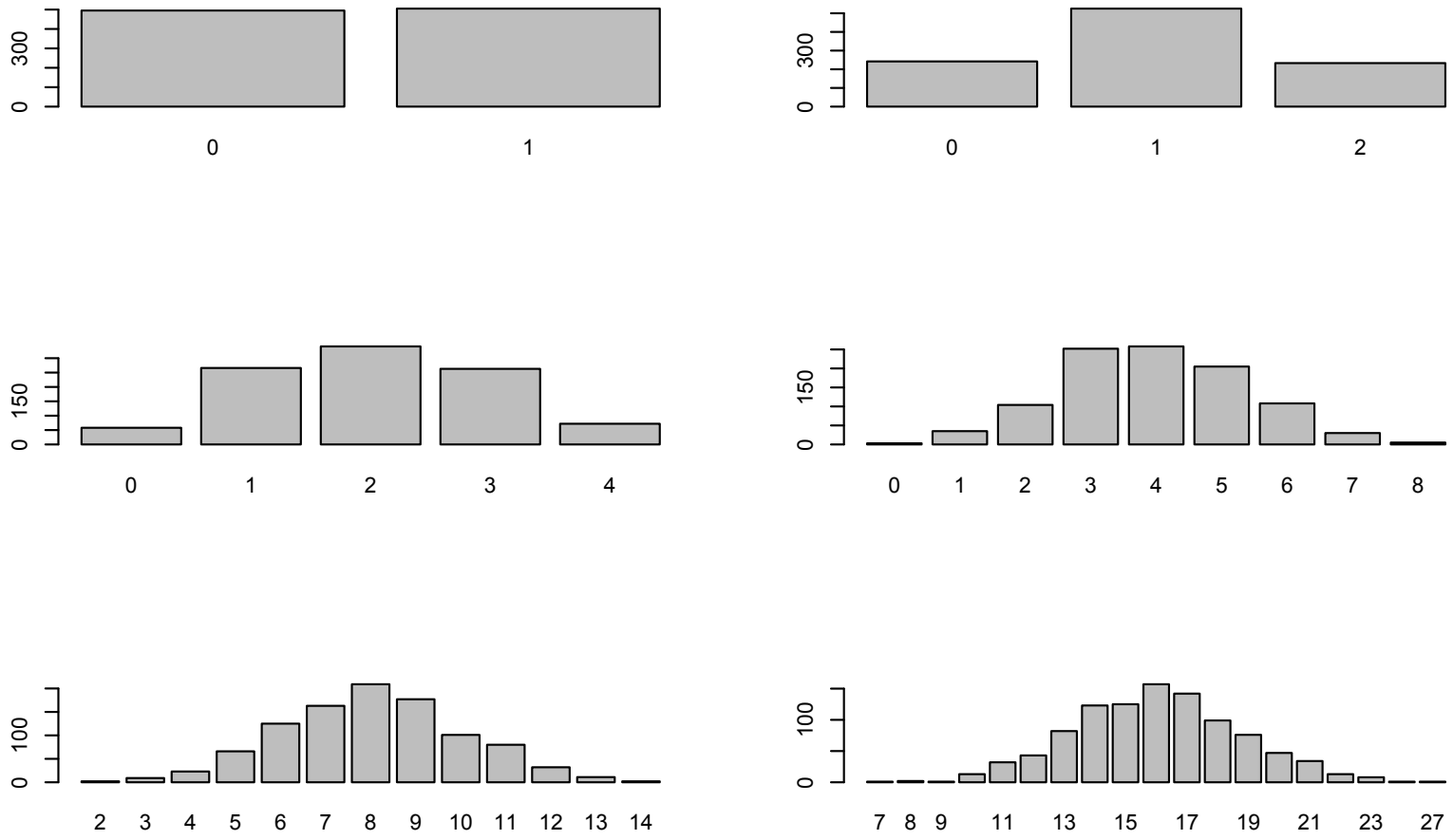
Central Tendencies and error

- Sample means reflect population values +/- error variability
- standard deviation of a mean (the standard error) = $s.d/\sqrt{N}$
- observed mean +/- 1 standard error includes the population value 68% of the time
- means that differ by 2.8 standard errors are unlikely to be from same population
- errors of within subject designs are more complicated to show

Histograms and box plots



Samples from the binomial (p=.5)



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