Math 1710 Class 15

Sampling Distributions, Summary Statistics

Dr. Back

Sep. 30, 2009

Problem 3

The Z scores on parts (c) and (e) came out strangely.
Not as I had intended them to work out.
We'll take this into account in grading.

Problem 3 a)

A Bernoulli trial here refers to the time for one team to complete its task. We might refer to a 2 minute completion time as "success" and a 1 minute completion time as "failure."

Problem 3 a)

A Bernoulli trial here refers to the time for one team to complete its task. We might refer to a 2 minute completion time as "success" and a 1 minute completion time as "failure."

With this somewhat arbitrary choice, the number of extra minutes beyond 125 required is simply the total number of successes in $X = \text{Binomial}(125, .6)$. *(Binomial because it is the total number of successes . . .)*

Problem 3 b)

We need only check the success/failure condition: $n = 125(.4) = 50 \geq 10$ and $n = 125(.6) = 75 \geq 10$. Since both of these check, we may approximate Binomial$(125, .6)$ by a normal distribution with the same mean and standard deviation. The mean here is $np = 75$ and the standard deviation is $\sqrt{125(.6)(.4)} = 5.48$.

Problem 3 c)

The total number of minutes required is $Y = X + 125$. By part b), we may approximate Y by an $N(75 + 125, 5.48) = N(200, 5.48)$ model. We want $P(160 \leq Y \leq 170)$. The Z score of 170 is

$$\frac{170 - 200}{5.48} = -5.47$$

and the Z score of 160 is

$$\frac{160 - 200}{5.48} = -7.30.$$

Each of these Z scores is off the charts in table Z, so both $P(Z < -5.47)$ and $P(Z < -7.30)$ are less than .0001. Hence $P(160 \leq Y \leq 170) < .0001$; the probability is essentially 0.

Problem 3 c) Remark

Remark: Sorry for the distractingly extreme Z scores! In making up this problem, I had desired a mean of 175 for $X + 125$, but interchanged the .4 and .6 probabilities required to achieve that. Unfortunately, we did not catch this oversight during proof-reading.

Problem 3 d)

Using an $N(0,1)$ picture, we see that the middle 80% corresponds to $-1.28 \leq Z \leq 1.28$. Hence the corresponding range for Y is from $200 - 1.28(5.48) = 200 - 7.01 = 192.99$ to $200 + 1.28(5.48) = 200 + 7.01 = 207.01$ minutes.

Problem 3 e)

Since the mean is 200 minutes, we expect that more than half the cars will require more than 195 minutes.

Problem 3 e) Remark

Remark: Had the mean been 175 as I thought when making up the question, this would have been a closer call. The Z score would have been $\frac{20}{5.48} = 3.65$ with $P(Z > 3.65) = 10^{-4}$. But with 10^5 cars, we'd still have expected to see something like 10 cars needing this much time.

Exploring Sampling Distributions Experimentally

Please take a quick look at Chapter 18 before the lab.

Exploring Sampling Distributions Experimentally

Given n observations $x_1, x_2, \ldots, x_n$, a statistic is any number determined by these observations.

Exploring Sampling Distributions Experimentally

Given n observations $x_1, x_2, \ldots, x_n$, a statistic is any number determined by these observations.

For example the mean $\bar{x}$ or standard deviation s.

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

A probability model describing the chance of different values of the statistic showing up.

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Two Types in the Lab:

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Two Types in the Lab:

Samp. Dist. of $\hat{p}$,

the proportion of infected ticks in a sample.

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Two Types in the Lab:

Samp. Dist. of $\hat{p}$,

the proportion of infected ticks in a sample.

Samp. Dist. of $\bar{x}$,

the average hospital costs in a sample.

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Two Types in the Lab:

Samp. Dist. of $\hat{p}$,

the proportion of infected ticks in a sample.

Samp. Dist. of $\bar{x}$,

the average hospital costs in a sample.

Big Question:

How does the sampling distribution qualitatively depend on sample size?

Exploring Sampling Distributions Experimentally

"Sampling Distribution of a Statistic"

Two Types in the Lab:

Samp. Dist. of $\hat{p}$,

the proportion of infected ticks in a sample.

Samp. Dist. of $\bar{x}$,

the average hospital costs in a sample.

Big Question:

How does the sampling distribution qualitatively depend on sample size?

Watch out for the DataDesk directions on creating samples!

The W's

- Who
- What *(and with what units)*
- When
- Where
- Why
- How

Data Tables usually have rows indicating *who* and columns for each *what.*

Each column is a *variable.*

Stemplots

Stemplot

stem	leaf
0	4 6 8
1	3 8
2	0 2 3 8 9 9
3	4 5 6 7 8
4	0 1 7 9
5	1 2

Stemplots

Stemplot

stem	leaf
0	4 6 8
1	3 8
2	0 2 3 8 9 9
3	4 5 6 7 8
4	0 1 7 9
5	1 2

Willie Mays' season home run totals are unimodal with a center of about 30 home runs per year. (The median is 31.5.) Most seasons fall within approximately 20 home runs of the median. The distribution is somewhat skewed to the left, mostly due to a number of low season totals near the end of his career.

Stemplots

Stemplot

stem	leaf
0	4 6 8
1	3 8
2	0 2 3 8 9 9
3	4 5 6 7 8
4	0 1 7 9
5	1 2

Hank Aaron: 13,27,26,44,30,39,40,34,45,44,
24,32,44,39,29,44,38,47,34,40,20,12,10,25

One can do "back to back stemplots:"

Stemplots

Pre-Stemplot

Aaron leaf	stem	Mays leaf
	0	4 6 8
0 2 3	1	3 8
5 0 9 4 6 7	2	0 2 3 8 9 9
4 8 9 2 4 9 0	3	4 5 6 7 8
0 7 4 4 4 5 0 4	4	0 1 7 9
	5	1 2

Stemplots

Pre-Stemplot

Aaron leaf	stem	Mays leaf
	0	4 6 8
0 2 3	1	3 8
5 0 9 4 6 7	2	0 2 3 8 9 9
4 8 9 2 4 9 0	3	4 5 6 7 8
0 7 4 4 4 5 0 4	4	0 1 7 9
	5	1 2

Stemplot

Aaron leaf	stem	Mays leaf
	0	4 6 8
3 2 0	1	3 8
9 7 6 5 4 0	2	0 2 3 8 9 9
9 9 8 4 4 2 0	3	4 5 6 7 8
7 5 4 4 4 4 0 0	4	0 1 7 9
	5	1 2

Stemplots

Stemplot

Aaron leaf	stem	Mays leaf
	0	4 6 8
3 2 0	1	3 8
9 7 6 5 4 0	2	0 2 3 8 9 9
9 9 8 4 4 2 0	3	4 5 6 7 8
7 5 4 4 4 4 0 0	4	0 1 7 9
	5	1 2

Aaron's season home run totals have less spread than those of Mays. This is evident graphically or can be seen by comparing Aaron's IQR of 16.5 to Mays' 20. Aaron's median of 34 is somewhat higher than Mays' median of 31.5. Both distributions are skewed to the left.

Just barely an ordinary outlier:

The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,46

Just barely an ordinary outlier:

The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17, 18,19,20,21,22,23,24,25,26,27,28,29,46

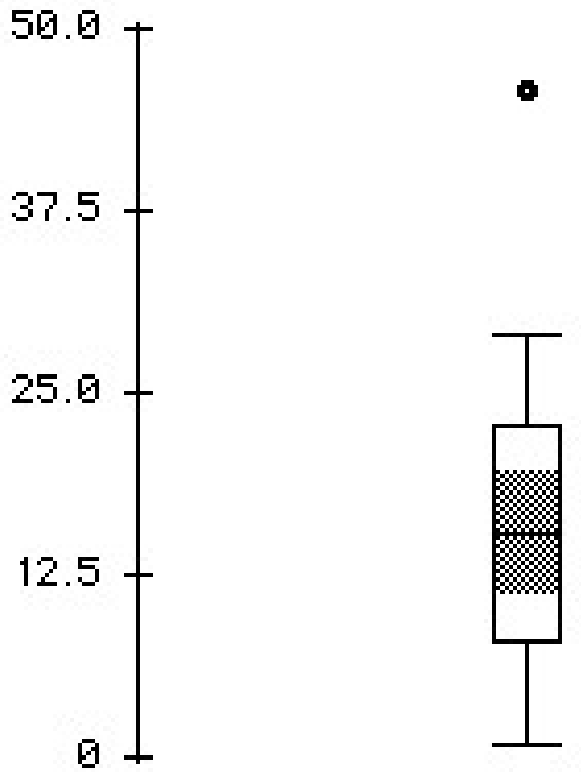
median	15.5
Q_1	8
Q_3	23
IQR	15

Just barely an ordinary outlier:

The data set:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,46

$$Q_3 + 1.5 \cdot \text{IQR} = 23 + 22.5 = 45.5$$



Just barely an ordinary outlier:

The data set with 45 instead:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,45

Just barely an ordinary outlier:

The data set with 45 instead:
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,45

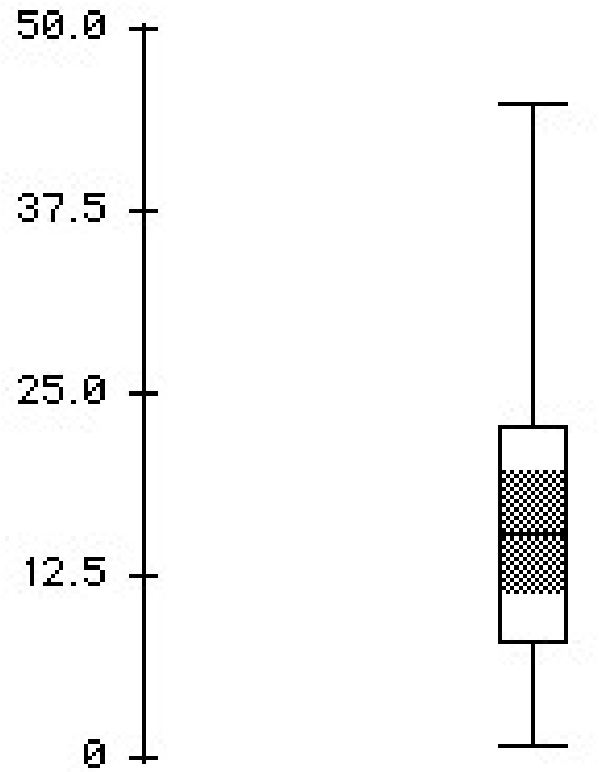
median	15.5
Q_1	8
Q_3	23
IQR	15

Just barely an ordinary outlier:

The data set with 45 instead:

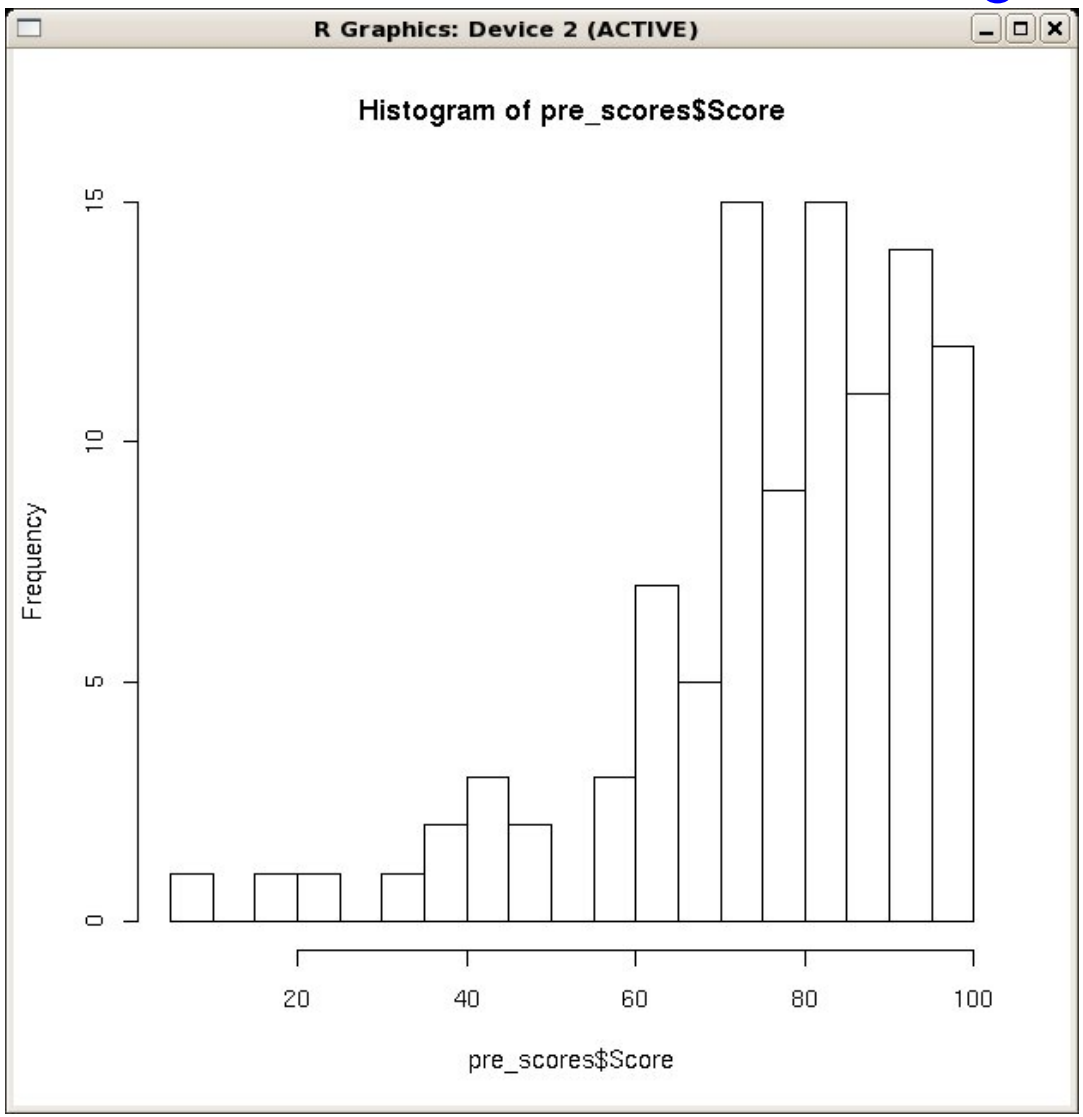
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,
18,19,20,21,22,23,24,25,26,27,28,29,45

$$Q_3 + 1.5 \cdot \text{IQR} = 23 + 22.5 = 45.5$$



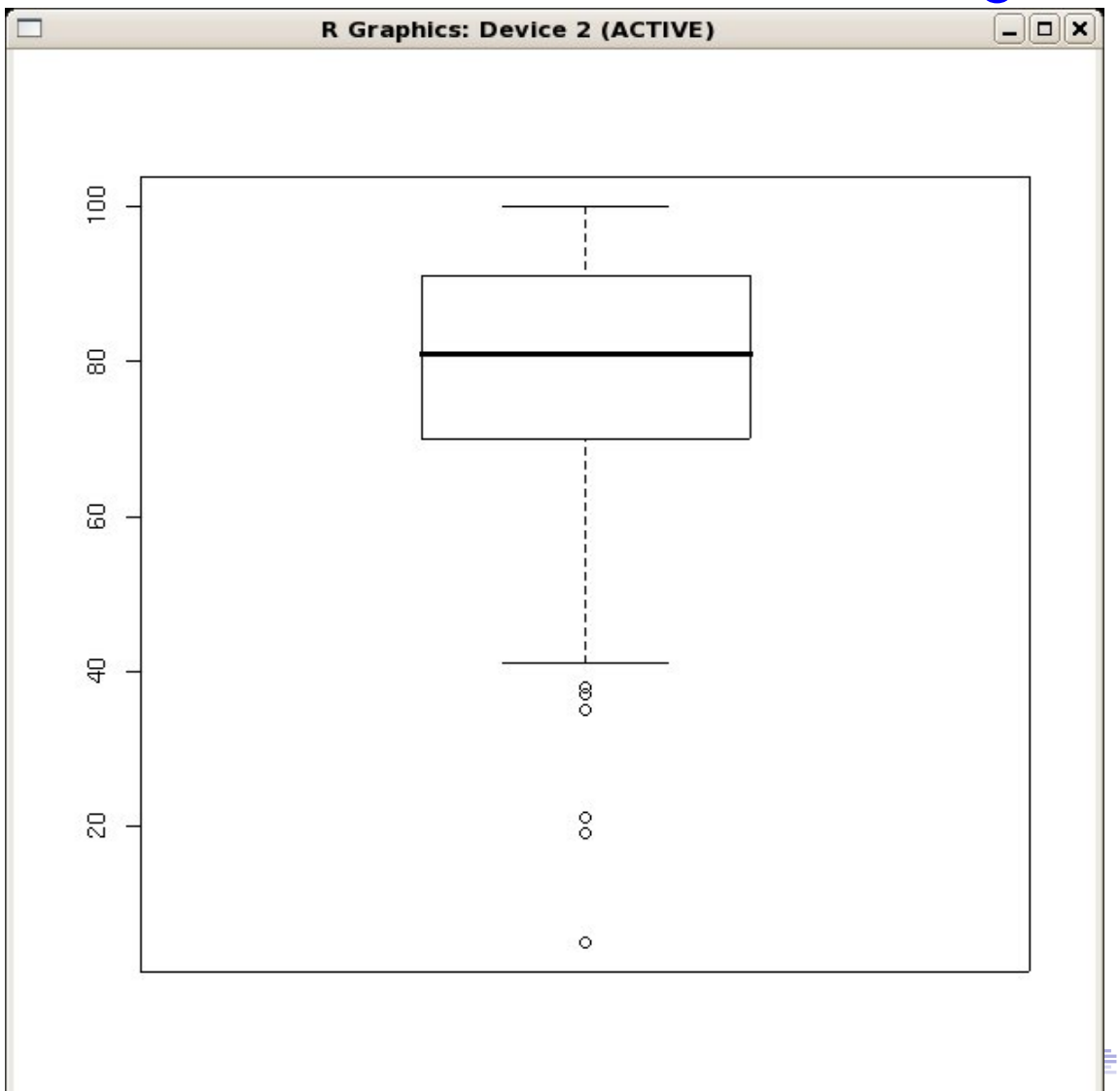
Limitations of the Outlier Rule

Prelim 1 Distribution Several Years Ago



Limitations of the Outlier Rule

Prelim 1 Distribution Several Years Ago



Limitations of the Outlier Rule

Using R:

$>$ pre_scores$<-$ read.table("pre_scores.txt",header=T)

The file pre_scores.txt looks like:

Score
100
100
...
5

Limitations of the Outlier Rule

(The scores are in order.)

```
> pre_scores$Score
[1] 100 100 100 100 99 99 ...
[19] 94 94 93 92 92 92 ...
...
[91] ...48 46 42 42 41 38 37 35 21 19 5

> pre_scores$Score[19]
[1] 94
```

Limitations of the Outlier Rule


```
> summary(pre_scores)
Score
Min.   : 5.00
1st Qu.: 70.25
Median : 81.00
Mean   : 76.89
3rd Qu.: 90.75
Max.   :100.00
```

Limitations of the Outlier Rule


```
> summary(pre_scores)
Score
Min.   : 5.00
1st Qu.: 70.25
Median : 81.00
Mean   : 76.89
3rd Qu.: 90.75
Max.   :100.00
```

$IQR = 90.75 - 70.25 = 20.5$

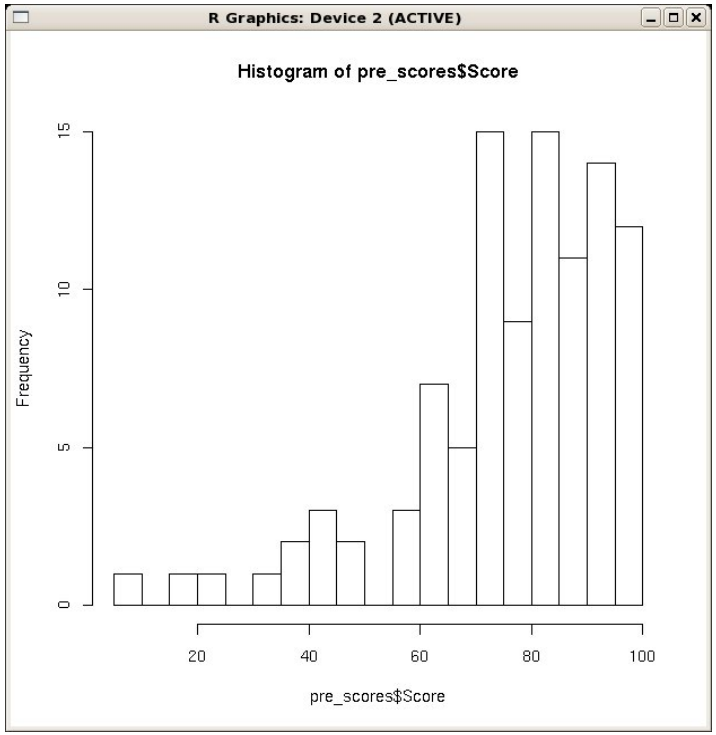
Limitations of the Outlier Rule

Lower Hinge: $Q_1 - 1.5 \cdot IQR = 70.25 - 30.75 = 39.5$

Limitations of the Outlier Rule

Lower Hinge: $Q_1 - 1.5 \cdot IQR = 70.25 - 30.75 = 39.5$

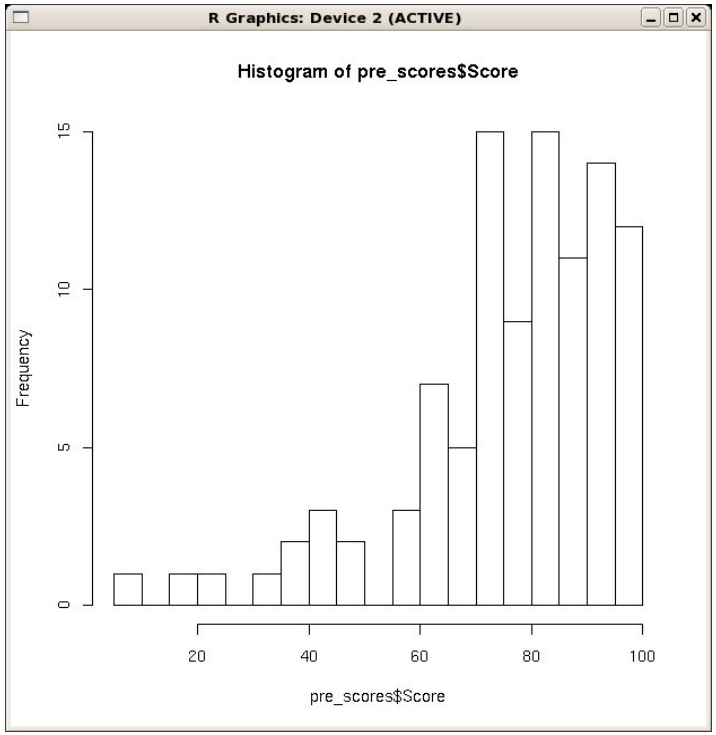
Prelim 1 Distribution Several Years Ago



Limitations of the Outlier Rule

Lower Hinge: $Q_1 - 1.5 \cdot IQR = 70.25 - 30.75 = 39.5$

Prelim 1 Distribution Several Years Ago



Among the low scores of 48 46 42 42 41 38 37 35 21 19 5,
38 and below fail the outlier test but 38 37 35
are not out of keeping with the neighboring values.

Graduation Plans

What Graduates Did

	1959	1970	1980
Continuing Education	197	388	320
Employed	103	137	98
In the Military	20	18	18
Other	13	58	45

Graduation Plans

What Graduates Did

	1959	1970	1980
Continuing Education	197	388	320
Employed	103	137	98
In the Military	20	18	18
Other	13	58	45

This table is giving frequency data on all combinations of the two categorical variables:

Year (of graduation)

WhatDid (after graduation)

Graduation Plans

What Graduates Did

	1959	1970	1980
Continuing Education	197	388	320
Employed	103	137	98
In the Military	20	18	18
Other	13	58	45

This table is giving frequency data on all combinations of the two categorical variables:

Year (of graduation)

WhatDid (after graduation)

Many people use the term *contingency table* for this situation.

Graduation Plans

What Graduates Did

	1959	1970	1980
Continuing Education	197	388	320
Employed	103	137	98
In the Military	20	18	18
Other	13	58	45

First form the row and column sums:

Graduation Plans

First form the row and column sums:

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

What does the last column represent?

$$\begin{array}{r} 905 \\ 338 \\ 56 \\ 116 \\ \hline 1415 \end{array}$$

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

What does the last column represent?

905
338
56
116
1415

It just gives frequency data on the single categorical variable *WhatDid*.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

Some people consider this last column to be the marginal distribution of *WhatDid.*

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

Some people consider this last column to be the marginal distribution of *WhatDid.*

It is more informative to include the *relative* frequencies as well.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

Our textbook considers the marginal dist of WhatDid to be:

WhatDid	**Freq.**	**Rel. Freq.**
Continuing Education	905	64%
Employed	338	24%
In the Military	56	4%
Other	116	8%
	1415	100%

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

Similarly the marginal distribution of Year is:

Year	**Freq.**	**Rel. Freq.**
1959	333	24%
1970	601	42%
1980	481	34%
	1415	100%

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(a) What % of these graduates joined the military?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(a) What % of these graduates joined the military?

This is answered by an entry in the marginal distribution of WhatDid.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(a) What % of these graduates joined the military?
This is answered by an entry in the marginal distribution of WhatDid.

$$\frac{56}{1415} = 4\%$$

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(b) What % of these students graduated in 1970?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(b) What % of these students graduated in 1970?
Based on the marginal distribution of Year.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(b) What % of these students graduated in 1970?
Based on the marginal distribution of Year.

$$\frac{601}{1415} = 42.5\%$$

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(c) What % of the 1970 graduates joined the military?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(c) What % of the 1970 graduates joined the military?

The relevant concept here is called the conditional distribution of WhatDid given that Year is 1970.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(c) What % of the 1970 graduates joined the military?

The relevant concept here is called the conditional distribution of WhatDid given that Year is 1970.

Some would just point to the 1970 column and say it is the conditional distribution of WhatDid among 1970 graduates.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(c) What % of the 1970 graduates joined the military?

As before, relative frequencies help, so our textbook would report the conditional distribution of WhatDid among 1970 graduates as follows:

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(c) What % of the 1970 graduates joined the military?

Conditional Distribution of WhatDid for Year=1970

WhatDid	**Freq.**	**Rel. Freq.**
Continuing Education	388	65%
Employed	137	23%
In the Military	18	3%
Other	58	10%
	601	100%

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(d) Of the students in these surveys who joined the military, what percent graduated in 1970?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(d) Of the students in these surveys who joined the military, what percent graduated in 1970?
Based on the conditional distribution of Year when WhatDid = Military. (the Military *row*)

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(d) Of the students in these surveys who joined the military, what percent graduated in 1970?

Based on the conditional distribution of Year when WhatDid = Military. (the Military *row*)

$$\frac{18}{56} = 32\%$$

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(e) What is the marginal distribution of postgraduation activities?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(e) What is the marginal distribution of postgraduation activities?

Discussed when we began.

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(e) What is the marginal distribution of postgraduation activities?

WhatDid	**Freq.**	**Rel. Freq.**
Continuing Education	905	64%
Employed	338	24%
In the Military	56	4%
Other	116	8%
	1415	100%

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(f) What is the conditional distribution of postgraduation activities when Year=1959?

Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(f) What is the conditional distribution of postgraduation activities when Year=1959?

WhatDid	**Freq.**	**Rel. Freq.**
Continuing Education	197	59%
Employed	103	31%
In the Military	20	6%
Other	13	4%
	601	100%

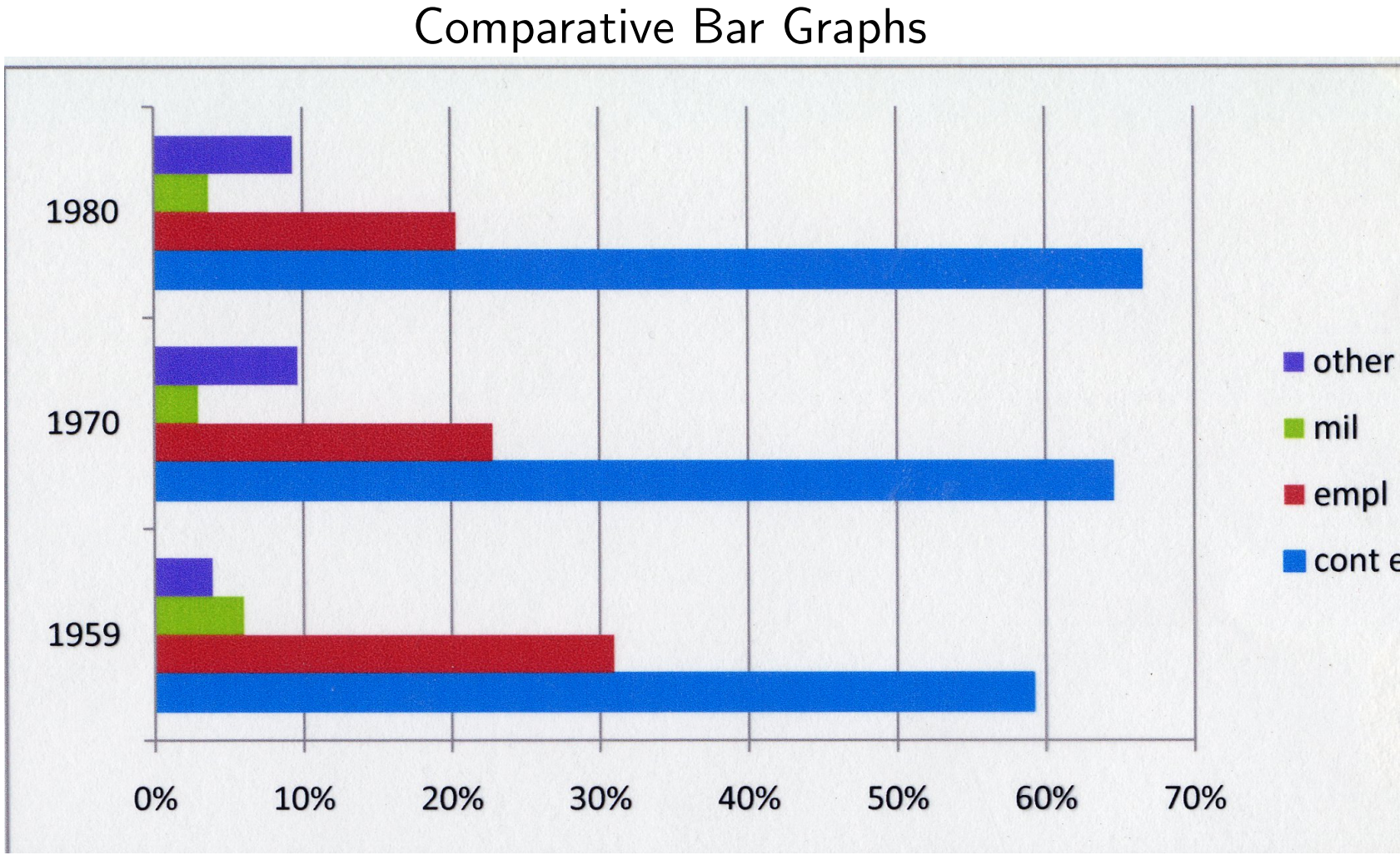
Graduation Plans

	1959	1970	1980	
Continuing Education	197	388	320	905
Employed	103	137	98	338
In the Military	20	18	18	56
Other	13	58	45	116
	333	601	481	1415

(f) Any evidence that plans have changed over the 21 year period? Brief Description? Appropriate graph?

Graduation Plans

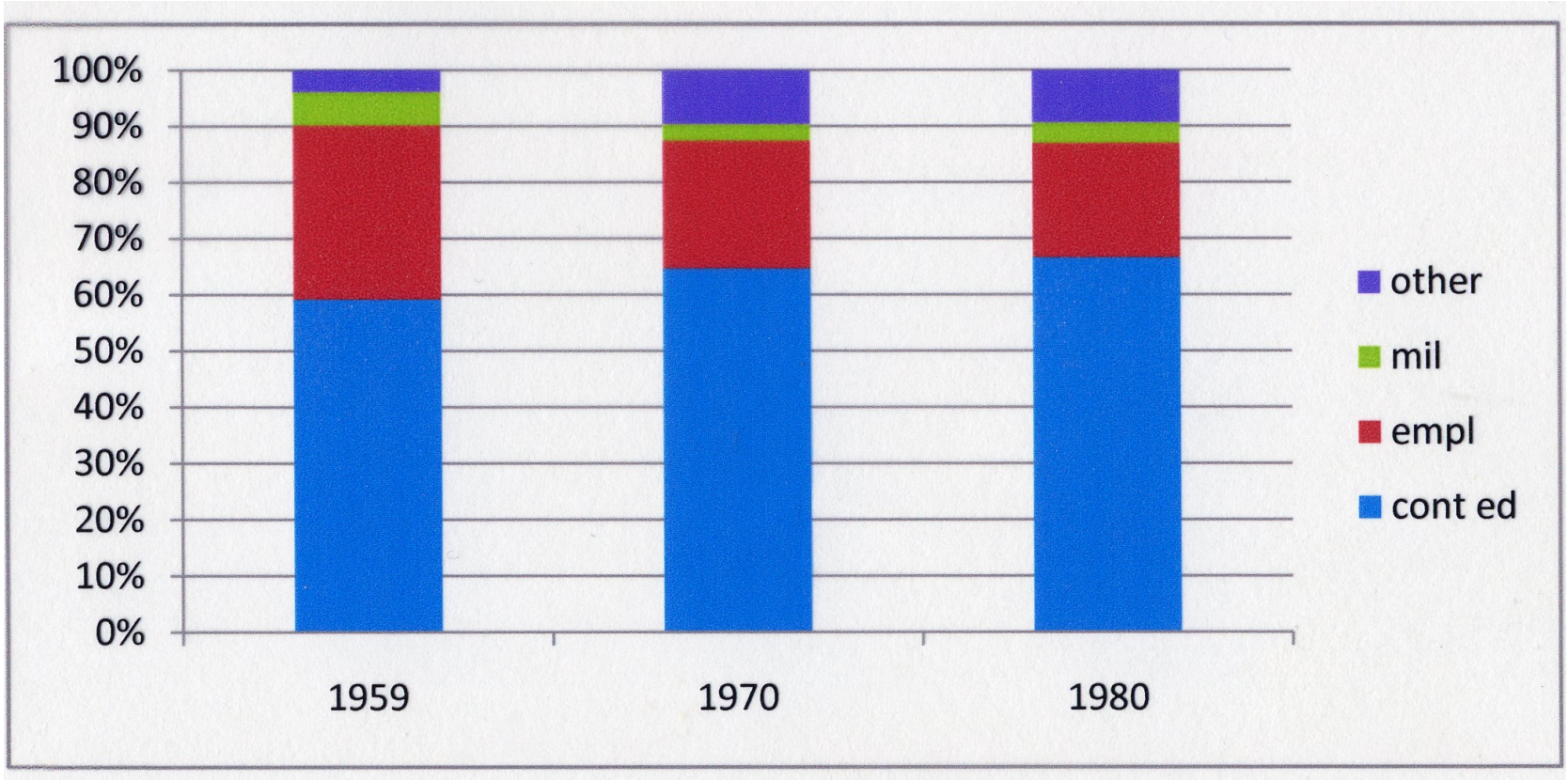
(f) Any evidence that plans have changed over the 21 year period? Brief Description? Appropriate graph?



Graduation Plans

(f) Any evidence that plans have changed over the 21 year period? Brief Description? Appropriate graph?

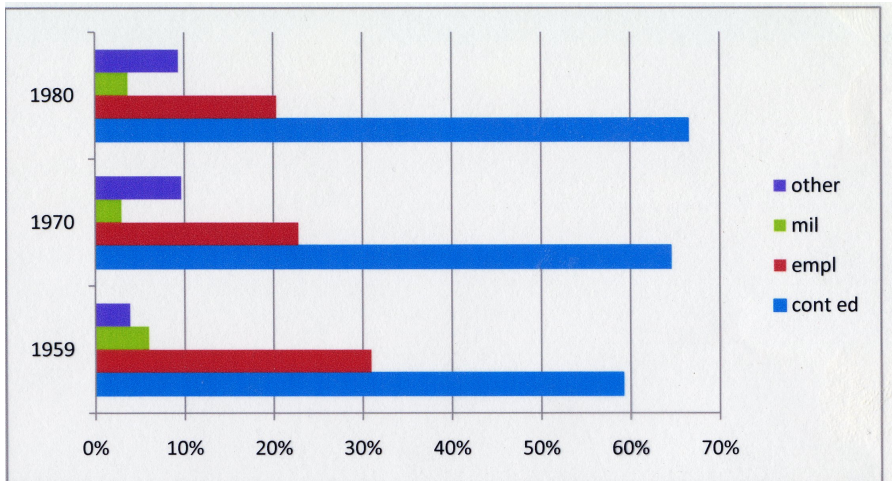
Segmented Bar Graphs



Graduation Plans

(f) Any evidence that plans have changed over the 21 year period? Brief Description? Appropriate graph?

Comparative Bar Graphs



The employment percentage has gone down over the years.
The education percentage has gone up.
Throughout the period, education has been the first choice and employment the second.