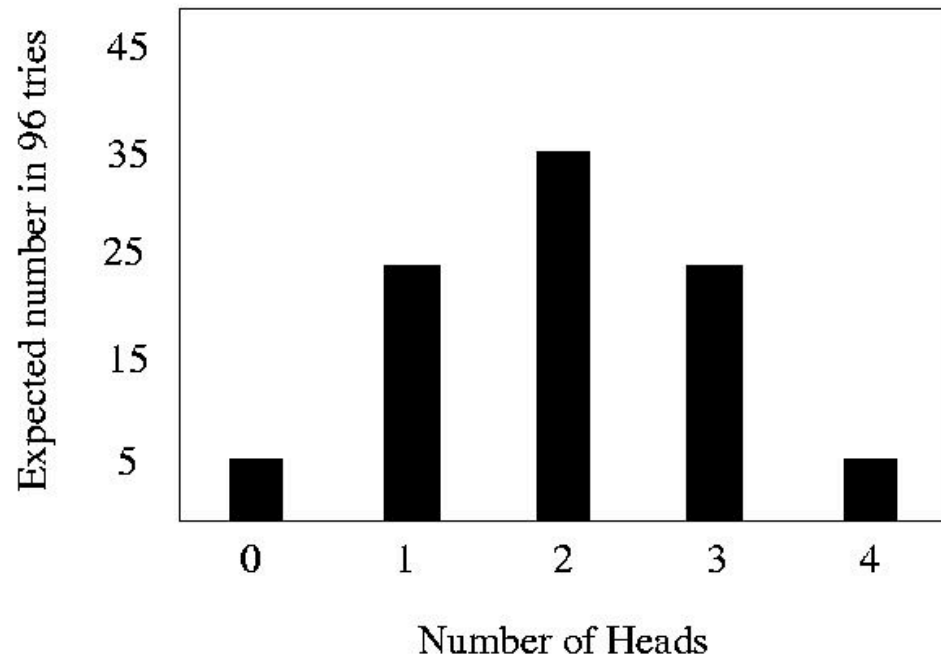


How is probability used?

- Probability plays an important role in statistics
  - How likely is that a given event was due to chance?
    - That is a question we will try to answer
- Probability distributions will be important
  - Discrete probability distributions
    - Events can only have certain values (e.g., Heads/Tails)
  - Continuous probability distributions
    - Events can take on any possible values (e.g., means)

- What is the relative likelihood of getting a head on a coin flip?
- What if we tried 4 coin flips 96 times?
  - How many times should you get 4 heads?
    - $E(N) = \text{Probability} * \text{Trials}$
    - $E(N) = 0.0625 * 96 = 6$

Expected Distribution in 96 tries



How do probabilities combine?

- Disjoint events

- Both A and B cannot occur at the same time.
- Probability of A or B,  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$

- Independent events

- The outcome of one event does not determine the outcome of another.
  - $\Pr(A | B) = \Pr(A)$
  - Two successive coin flips are independent
  - Two spins of a roulette wheel are independent
- Probability of A and B,  $\Pr(A \text{ and } B) = \Pr(A) * \Pr(B)$
- The probability of a conjunction  $\leq$  the probability of either conjunct

What is the probability of getting three heads in a row with a fair coin?

If the probability of student Z getting an A on a test is .3, a B is .5, a C is .1, and an F is .1, what is the probability of student Z getting an A or a C?

There is a 20% chance the number 28 bus in Olympia, WA will run late. There is a 30% chance the number 5 bus in Austin will run late. These events are independent. What is the probability that both buses will run late?

Linda is 31 years old, single, and bright. She majored in philosophy. As a student, she was deeply concerned with issues of social justice and participated in demonstrations. Which is more likely?

Linda is a banker.

Linda is a banker and active in the feminist movement.

- The law of large numbers
  - If I made 1000 sets of 4 coin flips, the distribution will start to look more like the probability distribution.
  
- There is no law of small numbers
  - If I see 8 heads in a row, that does not increase the probability that I will see a tail on the next flip
    - Gambler's fallacy
  
- We use probability to determine how likely it is that an observation was due to chance.
  
- We will rely on various probability distributions.
  - The Normal distribution (z)
  - Student's t distribution
  - The F distribution

- Interpreting data is not a mechanical process.
  
- Interpretation involves thinking both about the data and about how they were obtained.
  - How were the data collected?
  - What are the units of measurement?
    - How much information is contained in those units?
  
- Are there any distinctive patterns in the data? Qualities of distributions to keep in mind.
  - What is the center? Central tendency
  - What is the shape? Is the distribution symmetric or skewed?
  - What is the spread? Variability
  - Are there any outliers?

- Data (and life in general) come in distributions.
- An important first step in analyzing data is to plot and look at the data.
  - Patterns may become evident.
  - Outliers may become apparent.
- There are no rules for plotting data.
  - Find ways to look at data that are revealing.
- Good analysis takes practice.
  - Don't be frustrated if it takes time to develop the skill.

## Sample Size

- How many people should you ask?
  - Suppose you wanted to know who was going to win a presidential election.
    - How confident would you be if you asked one person how they would vote?
    - Two people? Ten people?
    - One hundred? One thousand? Ten thousand? All voters?
- Clearly, up to point, more is better.
  - Many (very accurate) national polls are based on only 1000-2000 respondents.
  - Samples must be moderately large, because sampling gives rise to distributions.

How can we summarize a distribution?

- Central tendency

- When the distribution has one peak, a measure of central tendency makes sense.

- Not a good measure for multi-peaked distributions

- Measures of central tendency

- Mean =  $\text{sum}(x) / N = (x_1 + x_2 + x_3 + \dots + x_N) / N$

- Median = the middle observation (or the mean of the two observations around the middle)

- Mode = the most frequent value

- Variability

- Quartiles

- 5 number summary using minimum, Q1, median, Q3, maximum

- Variance and Standard Deviation

- Variance =  $s^2 = \text{SUM}(x - \text{mean})^2 / (N - 1)$

- Standard deviation =  $s = (\text{variance})^{1/2}$

- Only use  $s$  when mean is used as a measure of central tendency.

## Hypothesis Testing

- Often we want to know whether our data reveal a reliable effect.
  - This process is called statistical inference or hypothesis testing
- If a morning class drinks caffeine, they will average an 80 on the final exam.
  - How could we test this claim?
- Start by determining the null hypothesis ( $H_0$ ).
  - A state of affairs against which we are comparing.

$H_0: \mu = 80$

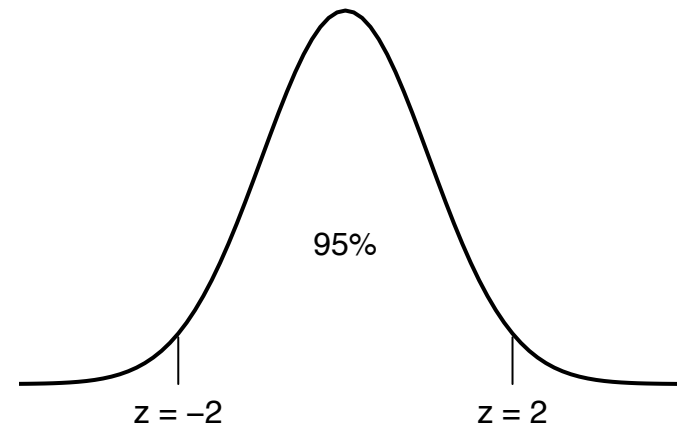
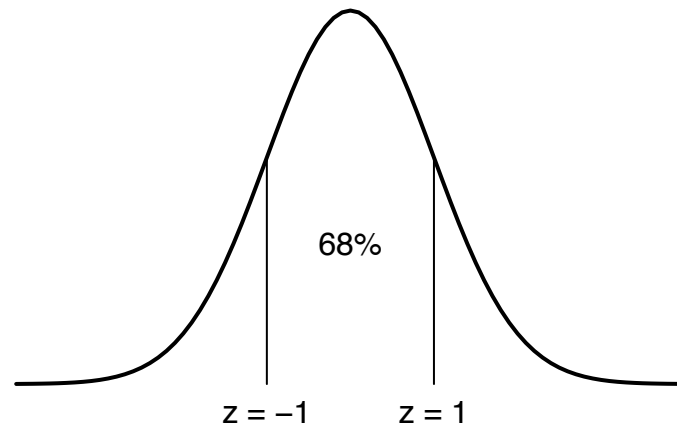
$H_1: \mu < > 80$

The alternative hypothesis ( $H_1$ ) is always stated relative to the null hypothesis.

## z and t distributions

$$- z = (\bar{x} - \mu) / (\sigma / n^{1/2})$$

$$- t = (\bar{x} - \mu) / (s / n^{1/2})$$



$H_0: \mu = 80$

$H_1: \mu \neq 80$

- 32 students

- Mean = 86

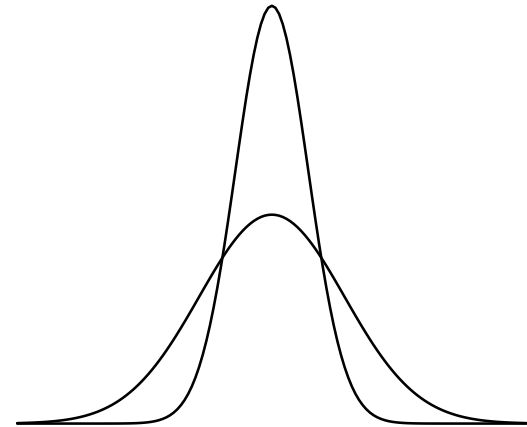
-  $s = 5.4$

Standard error of the mean =  $.955 [ s/\sqrt{N} ]$

$t(31) = (86 - 80) / .955 = 6.283$

$p < .0005$

Reject  $H_0$



Half a morning class is given regular coffee and the other half is given decaf for a semester.

$$H_0: \mu_{\text{regular}} = \mu_{\text{decaf}}$$

$$H_1: \mu_{\text{regular}} < > \mu_{\text{decaf}}$$

Mean (regular) = 87, var (regular) = 6.5, n = 16

Mean (decaf) = 82, var(decaf) = 7.1, n = 16

pooled var = 6.8

pooled sd=2.61

$$t(30) = (87 - 82) / (2.61 * \text{sqrt}((1 / 16) + (1 / 16))) = 5.42$$

p < .05

Reject  $H_0$

When we want to see two samples are different.

$$t(n_1 + n_2 - 2) = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

t test is robust

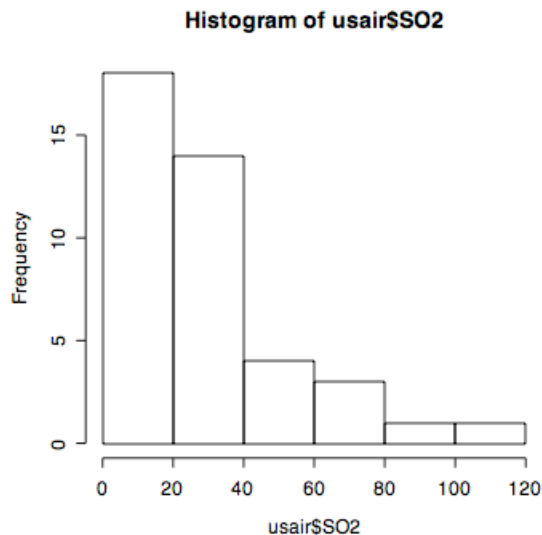
- Test assumes that variances are the same.
  - Even if the variances are not the same, the test still works pretty well.
- Test assumes data are drawn from a normally distributed population.
  - Even if the population is not normally distributed, the test still works pretty well.
- Of course, there are limits.

A public health advocate believes children growing up by power lines get sick 15 days per year. A sample of three children is gathered with number of sick days equal to 19, 15, 17. Do you reject the null hypothesis?

Another researcher wants to compare the effect on boys and girls. She gets a sample of three girls, 10, 15, 20, as well as a sample of boys, 14, 20, 26. Are boys and girls affected differently by the power lines?

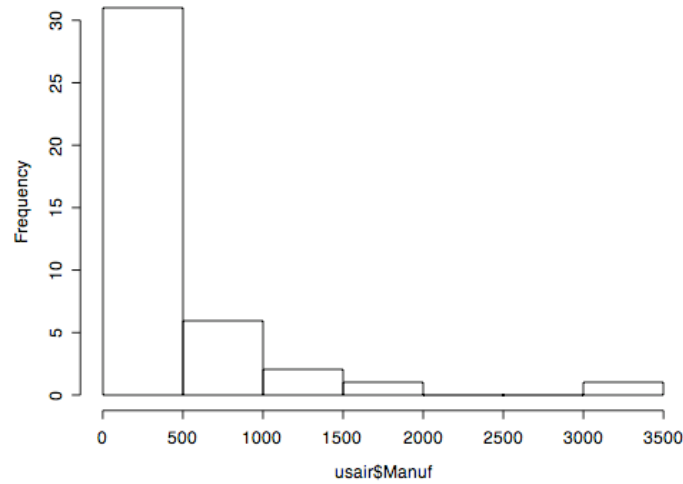
## Distributions and Variability

- What to do if you actually collect data?
- Visualizing distributions
- Variability – not all the data are the same

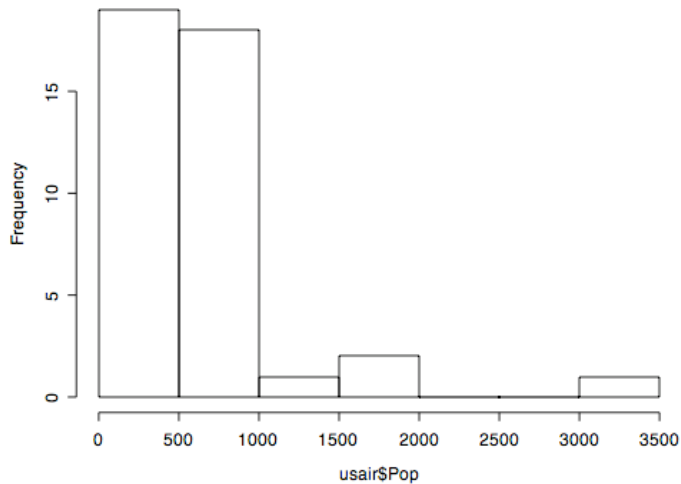


	SO2	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Phoenix	10	-70.3	213	582	6.0	7.05	36
Little_Rock	13	-61.0	91	132	8.2	48.52	100
San_Francisco	12	-56.7	453	716	8.7	20.66	67
Denver	17	-51.9	454	515	9.0	12.95	86
Hartford	56	-49.1	412	158	9.0	43.37	127
Wilmington	36	-54.0	80	80	9.0	40.25	114
Washington	29	-57.3	434	757	9.3	38.89	111
Jacksonville	14	-68.4	136	529	8.8	54.47	116
Miami	10	-75.5	207	335	9.0	59.80	128
Atlanta	24	-61.5	368	497	9.1	48.34	115
Chicago	110	-50.6	3344	3369	10.4	34.44	122
Indianapolis	28	-52.3	361	746	9.7	38.74	121
Des_Moines	17	-49.0	104	201	11.2	30.85	103
Wichita	8	-56.6	125	277	12.7	30.58	82
Louisville	30	-55.6	291	593	8.3	43.11	123
New_Orleans	9	-68.3	204	361	8.4	56.77	113
Baltimore	47	-55.0	625	905	9.6	41.31	111
Detroit	35	-49.9	1064	1513	10.1	30.96	129
Minneapolis-St._Paul	29	-43.5	699	744	10.6	25.94	137
Kansas_City	14	-54.5	381	507	10.0	37.00	99
St._Louis	56	-55.9	775	622	9.5	35.89	105
Omaha	14	-51.5	181	347	10.9	30.18	98
Albuquerque	11	-56.8	46	244	8.9	7.77	58
Albany	46	-47.6	44	116	8.8	33.36	135
Buffalo	11	-47.1	391	463	12.4	36.11	166
Cincinnati	23	-54.0	462	453	7.1	39.04	132
Cleveland	65	-49.7	1007	751	10.9	34.99	155
Columbus	26	-51.5	266	540	8.6	37.01	134
Philadelphia	69	-54.6	1692	1950	9.6	39.93	115
Pittsburgh	61	-50.4	347	520	9.4	36.22	147
Providence	94	-50.0	343	179	10.6	42.75	125
Memphis	10	-61.6	337	624	9.2	49.10	105
Nashville	18	-59.4	275	448	7.9	46.00	119
Dallas	9	-66.2	641	844	10.9	35.94	78
Houston	10	-68.9	721	1233	10.8	48.19	103
Salt_Lake_City	28	-51.0	137	176	8.7	15.17	89
Norfolk	31	-59.3	96	308	10.6	44.68	116
Richmond	26	-57.8	197	299	7.6	42.59	115
Seattle	29	-51.1	379	531	9.4	38.79	164
Charleston	31	-55.2	35	71	6.5	40.75	148
Milwaukee	16	-45.7	569	717	11.8	29.07	123

Histogram of usair\$Manuf

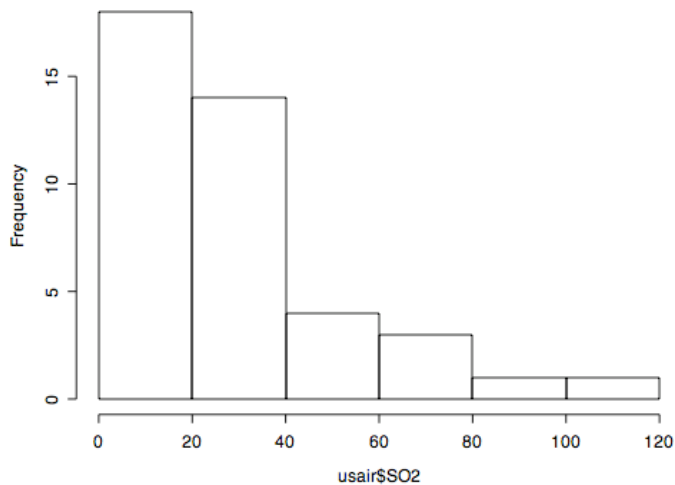


Histogram of usair\$Pop

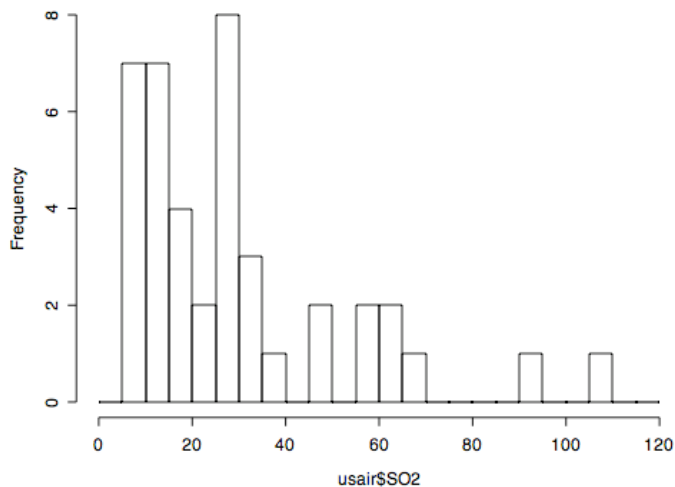


	S02	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Phoenix	10	-70.3	213	582	6.0	7.05	36
Little_Rock	13	-61.0	91	132	8.2	48.52	100
San_Francisco	12	-56.7	453	716	8.7	20.66	67
Denver	17	-51.9	454	515	9.0	12.95	86
Hartford	56	-49.1	412	158	9.0	43.37	127
Wilmington	36	-54.0	80	80	9.0	40.25	114
Washington	29	-57.3	434	757	9.3	38.89	111
Jacksonville	14	-68.4	136	529	8.8	54.47	116
Miami	10	-75.5	207	335	9.0	59.80	128
Atlanta	24	-61.5	368	497	9.1	48.34	115
Chicago	110	-50.6	3344	3369	10.4	34.44	122
Indianapolis	28	-52.3	361	746	9.7	38.74	121
Des_Moines	17	-49.0	104	201	11.2	30.85	103
Wichita	8	-56.6	125	277	12.7	30.58	82
Louisville	30	-55.6	291	593	8.3	43.11	123
New_Orleans	9	-68.3	204	361	8.4	56.77	113
Baltimore	47	-55.0	625	905	9.6	41.31	111
Detroit	35	-49.9	1064	1513	10.1	30.96	129
Minneapolis-St._Paul	29	-43.5	699	744	10.6	25.94	137
Kansas_City	14	-54.5	381	507	10.0	37.00	99
St._Louis	56	-55.9	775	622	9.5	35.89	105
Omaha	14	-51.5	181	347	10.9	30.18	98
Albuquerque	11	-56.8	46	244	8.9	7.77	58
Albany	46	-47.6	44	116	8.8	33.36	135
Buffalo	11	-47.1	391	463	12.4	36.11	166
Cincinnati	23	-54.0	462	453	7.1	39.04	132
Cleveland	65	-49.7	1007	751	10.9	34.99	155
Columbus	26	-51.5	266	540	8.6	37.01	134
Philadelphia	69	-54.6	1692	1950	9.6	39.93	115
Pittsburgh	61	-50.4	347	520	9.4	36.22	147
Providence	94	-50.0	343	179	10.6	42.75	125
Memphis	10	-61.6	337	624	9.2	49.10	105
Nashville	18	-59.4	275	448	7.9	46.00	119
Dallas	9	-66.2	641	844	10.9	35.94	78
Houston	10	-68.9	721	1233	10.8	48.19	103
Salt_Lake_City	28	-51.0	137	176	8.7	15.17	89
Norfolk	31	-59.3	96	308	10.6	44.68	116
Richmond	26	-57.8	197	299	7.6	42.59	115
Seattle	29	-51.1	379	531	9.4	38.79	164
Charleston	31	-55.2	35	71	6.5	40.75	148
Milwaukee	16	-45.7	569	717	11.8	29.07	123

Histogram of usairSSO2



Histogram of usairSSO2



	S02	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Phoenix	10	-70.3	213	582	6.0	7.05	36
Little_Rock	13	-61.0	91	132	8.2	48.52	100
San_Francisco	12	-56.7	453	716	8.7	20.66	67
Denver	17	-51.9	454	515	9.0	12.95	86
Hartford	56	-49.1	412	158	9.0	43.37	127
Wilmington	36	-54.0	80	80	9.0	40.25	114
Washington	29	-57.3	434	757	9.3	38.89	111
Jacksonville	14	-68.4	136	529	8.8	54.47	116
Miami	10	-75.5	207	335	9.0	59.80	128
Atlanta	24	-61.5	368	497	9.1	48.34	115
Chicago	110	-50.6	3344	3369	10.4	34.44	122
Indianapolis	28	-52.3	361	746	9.7	38.74	121
Des_Moines	17	-49.0	104	201	11.2	30.85	103
Wichita	8	-56.6	125	277	12.7	30.58	82
Louisville	30	-55.6	291	593	8.3	43.11	123
New_Orleans	9	-68.3	204	361	8.4	56.77	113
Baltimore	47	-55.0	625	905	9.6	41.31	111
Detroit	35	-49.9	1064	1513	10.1	30.96	129
Minneapolis-St._Paul	29	-43.5	699	744	10.6	25.94	137
Kansas_City	14	-54.5	381	507	10.0	37.00	99
St._Louis	56	-55.9	775	622	9.5	35.89	105
Omaha	14	-51.5	181	347	10.9	30.18	98
Albuquerque	11	-56.8	46	244	8.9	7.77	58
Albany	46	-47.6	44	116	8.8	33.36	135
Buffalo	11	-47.1	391	463	12.4	36.11	166
Cincinnati	23	-54.0	462	453	7.1	39.04	132
Cleveland	65	-49.7	1007	751	10.9	34.99	155
Columbus	26	-51.5	266	540	8.6	37.01	134
Philadelphia	69	-54.6	1692	1950	9.6	39.93	115
Pittsburgh	61	-50.4	347	520	9.4	36.22	147
Providence	94	-50.0	343	179	10.6	42.75	125
Memphis	10	-61.6	337	624	9.2	49.10	105
Nashville	18	-59.4	275	448	7.9	46.00	119
Dallas	9	-66.2	641	844	10.9	35.94	78
Houston	10	-68.9	721	1233	10.8	48.19	103
Salt_Lake_City	28	-51.0	137	176	8.7	15.17	89
Norfolk	31	-59.3	96	308	10.6	44.68	116
Richmond	26	-57.8	197	299	7.6	42.59	115
Seattle	29	-51.1	379	531	9.4	38.79	164
Charleston	31	-55.2	35	71	6.5	40.75	148
Milwaukee	16	-45.7	569	717	11.8	29.07	123