Quantitative Methods for Political Science

Descriptive Statistics

September 11, 2023

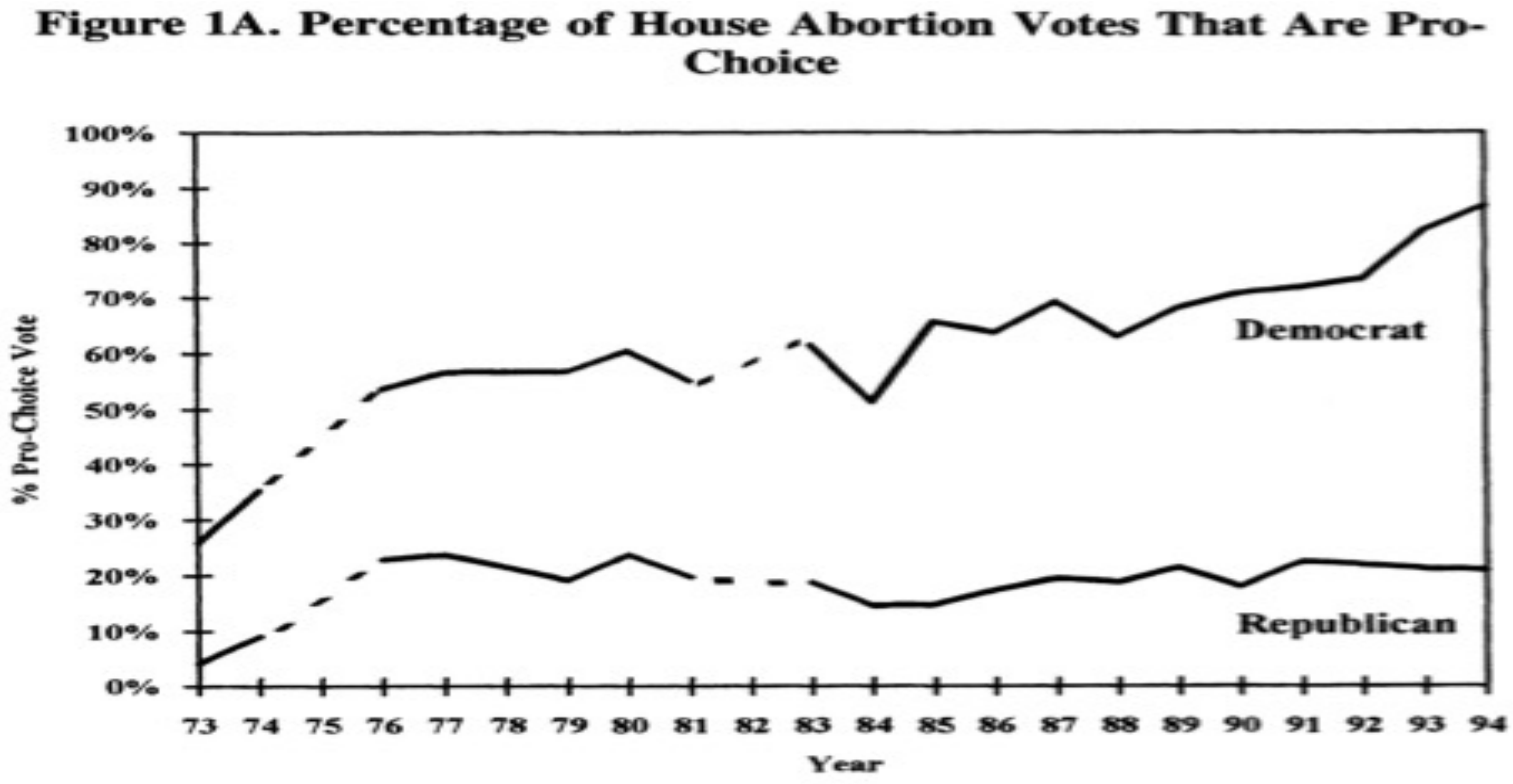
Three Parts of Research

- Design—what is the question? How will we try to answer it?
- Description—what do the data say?
- Inference—What can we learn about the question from the evidence? (And, technically, what can we infer about the population given the sample)

Partisan Attitudes toward Abortion

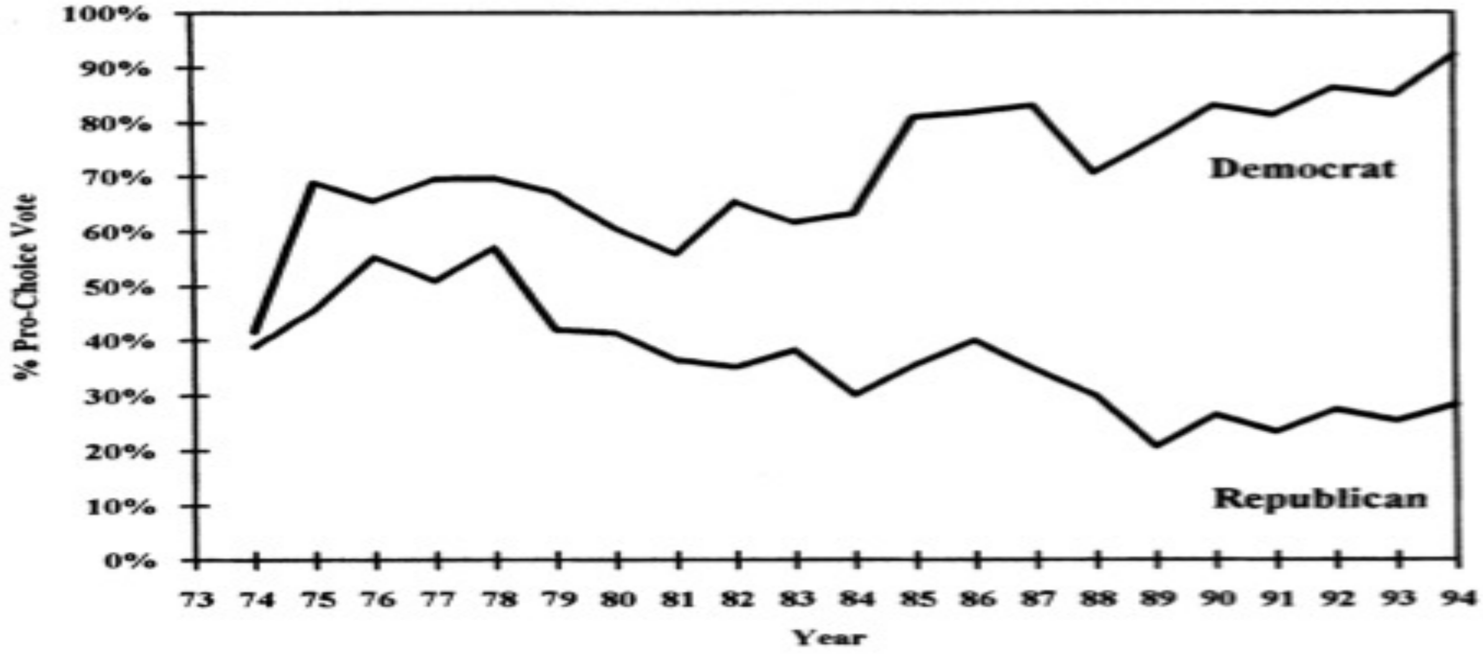
- Have partisan positions on abortion shifted since Roe vs. Wade (1973)?
- Two levels:
  - Elites—members of Congress
  - Masses—individuals identifying with each political party
- From Greg D. Adams (1997), "Abortion: Evidence of an Issue Evolution." *American Journal of Political Science* 41(3): 718-737.

Evidence for Elites



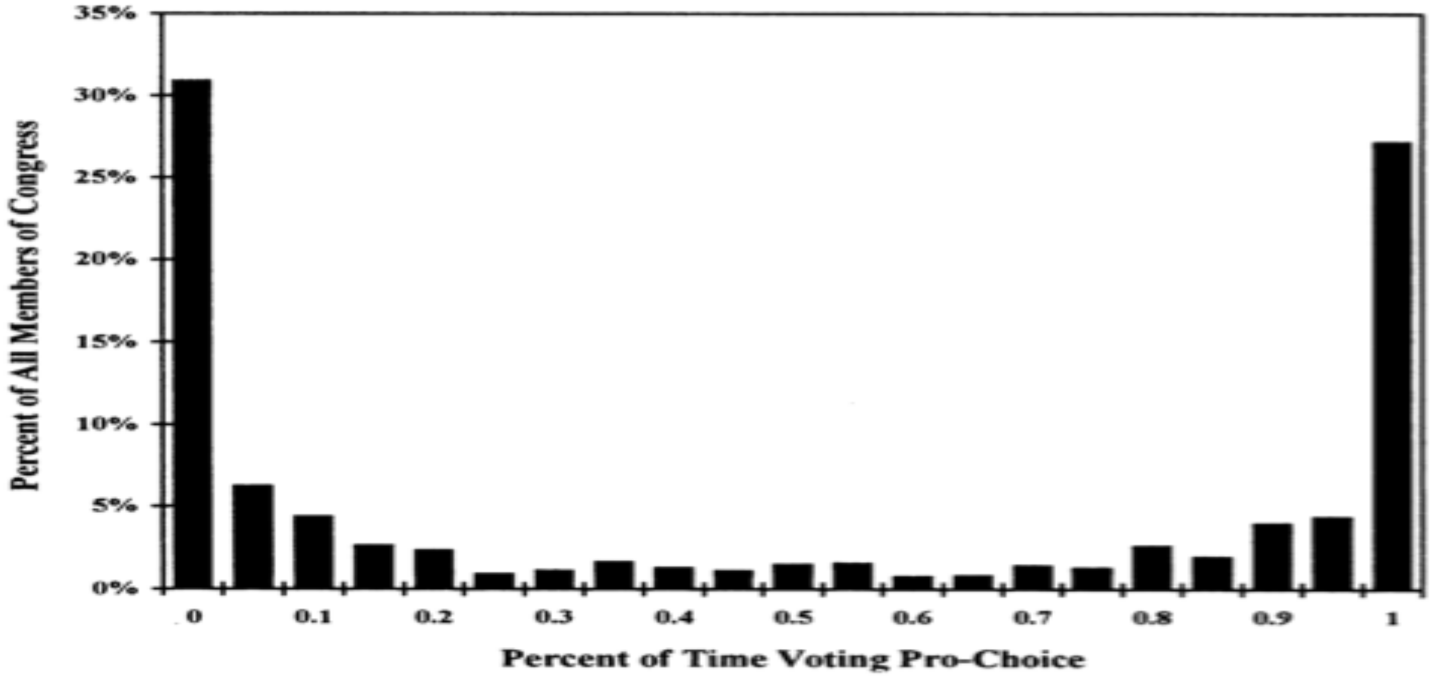
Evidence for Elites

Figure 1B. Percentage of Senate Abortion Votes That Are Pro-Choice



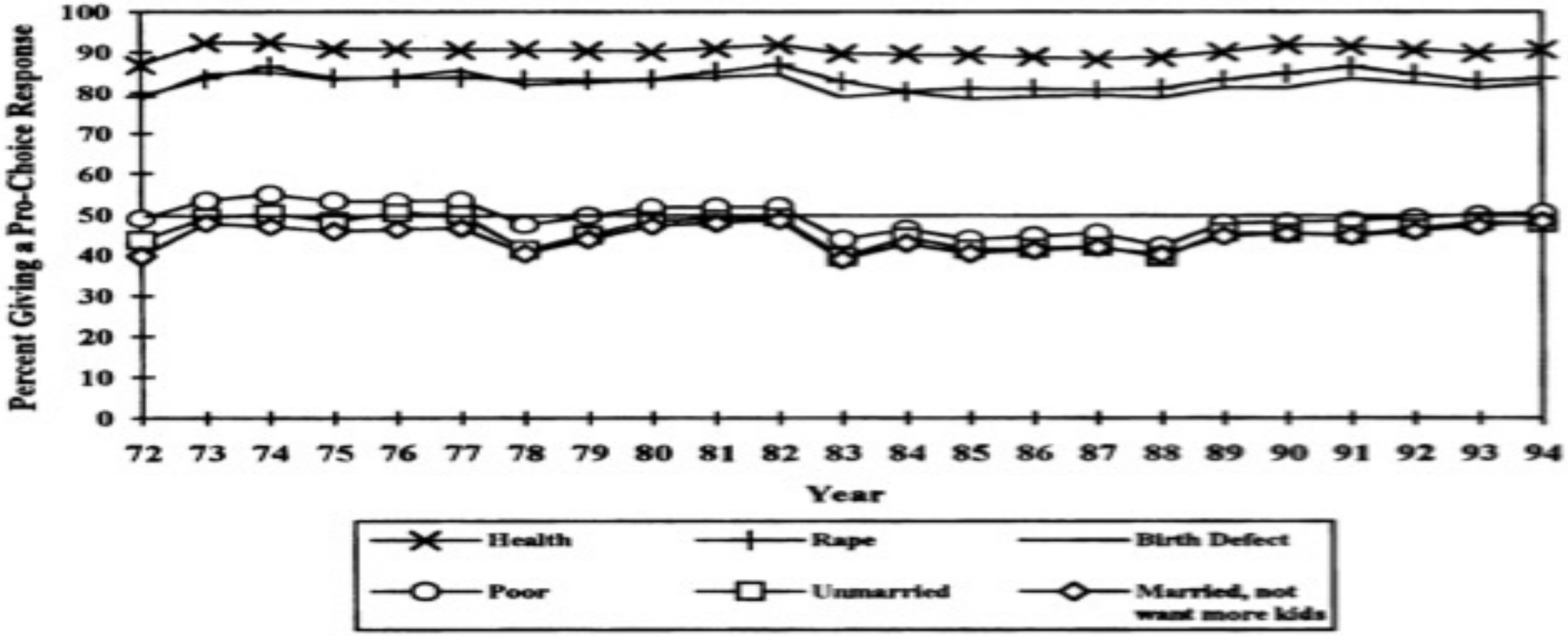
Why? Do representatives switch?

Figure 3. Distribution of Individual Legislators' Abortion Votes Over Entire Career



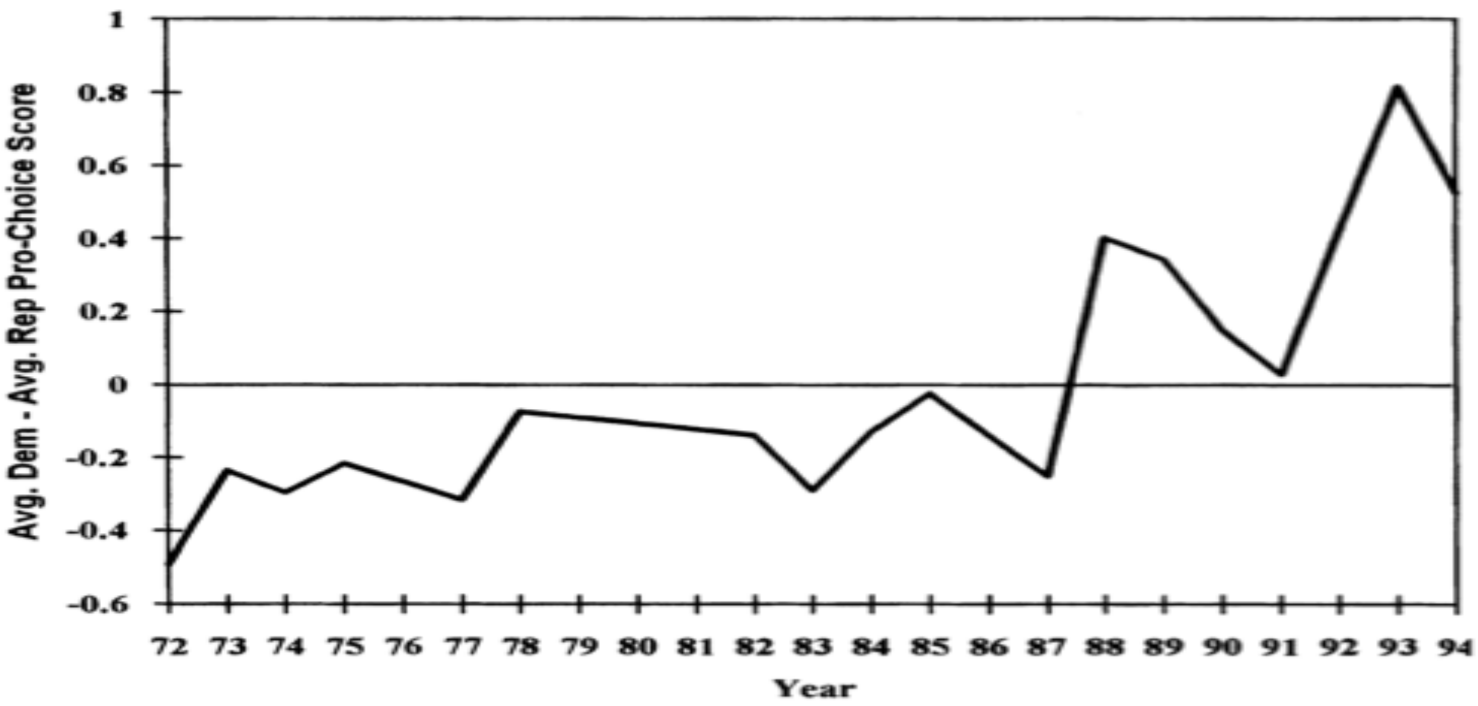
What about the masses?

Figure 4. Support for Abortion Rights Among Survey Respondents
Source: **General Social Surveys, 1972–94.**



Partisan differences among the masses

Figure 5. Difference Between Average Mass Republican and Democrat Pro-Choice Scores



Research Design

- First step—ask an interesting question
- Then, develop a theory to answer that question
- Think about how to design a study to answer that question
  - Who should participate in the study?
  - How can we isolate the relationship between variables?
  - How can we get the data we need to answer the question?
  - How will you measure concepts?

Description

- Summarize the raw data
- Important rule—the techniques you can use to describe (and analyze) data differ depending on type of variables you have
  - Categorical variables
  - Quantitative variables
- Present the data in a useful format
  - Can serve as exploratory data analysis

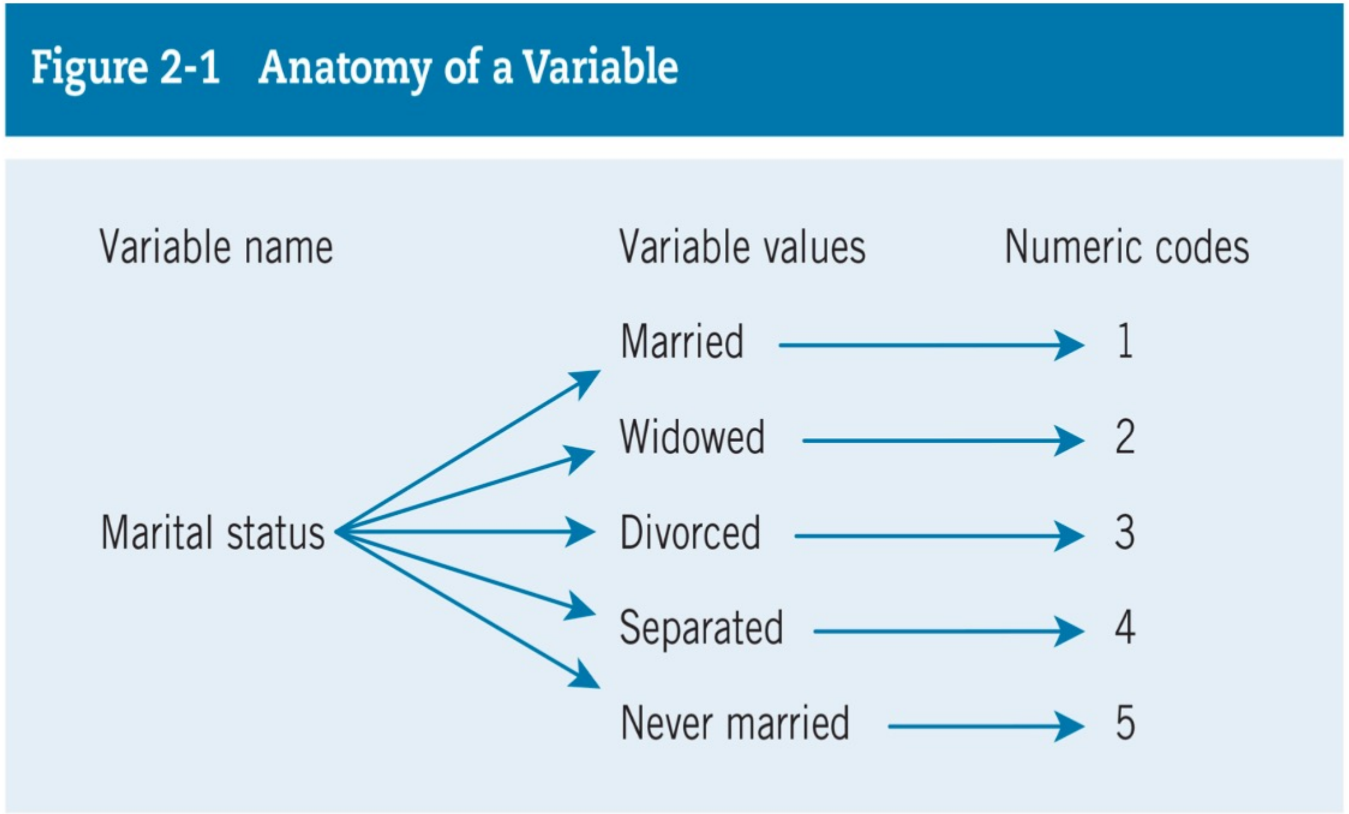
Variables

- Research pertains to some unit of analysis:
  - Individual, Household, Congressional District, State, Country, International System
- Observe characteristics of these units (observations, or cases)
- A variable is an empirical measurement of a characteristic
- Key rule—variables vary across observations
- Different types of variables based on how they are measured and what numbers mean

Types of Variables

- Quantitative/Interval/Continuous-Observations take on numerical values
- Categorical-Observations belong to one of a set of categories
  - Nominal-Categories are named
  - Ordinal-Categories are ranked
- Dichotomous variables—two values, yes/no
  - War/not war, win/lose, etc.
- Important—all variables can be measured with numbers, these numbers mean different things for quantitative vs. categorical variables
- Examples—Age, Race, Percentage of vote candidate receives, Gender
- Many concepts can be measured at different levels

Coding a Nominal Variable



Transforming Variables

- Can collapse quantitative variables into ordinal (or nominal) variables
  - Ex—income to categories (low/middle/high)
- Cannot go the other way
- Generally want to retain as much information as possible
- Level of measurement for variable should be driven by theory

Describing Variables

- Distribution of a variable tells us what values it takes and how often it takes these values
- Techniques used to describe distribution of variables differ depending on type of variable
  - Quantitative—center, spread and shape of distribution
  - Categorical—percentage in each category
- Can describe variables individually and also examine relationship between them descriptively
  - But, need more rigorous analysis for actual hypothesis testing

Describing Categorical Variables

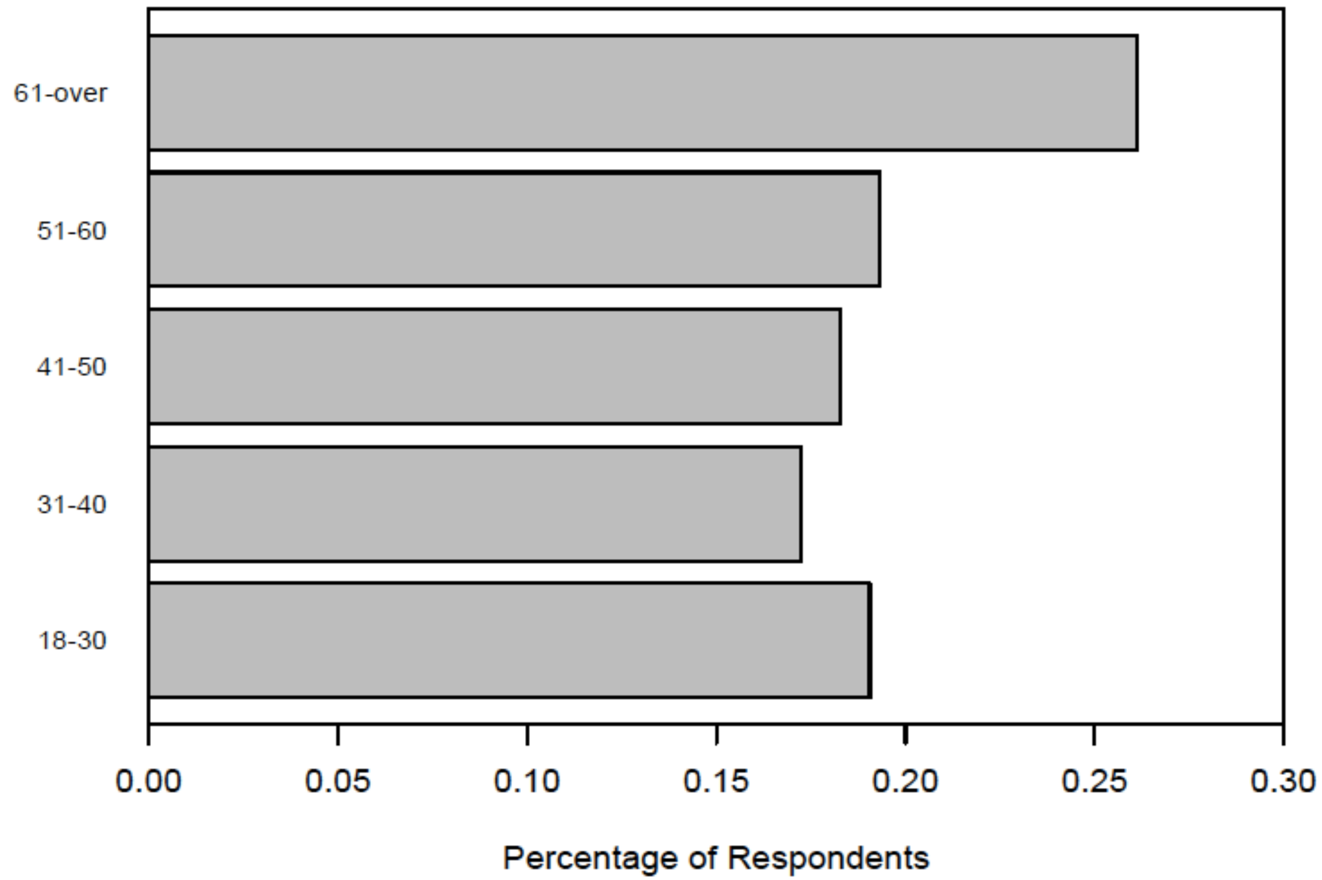
- Interested in count and/or percentage of cases that fall in each category of variable

- Ways to display this:
  - Frequency Distribution
  - Bar Chart

Frequency Distribution

Age Group	Frequency	Percent (%)	Cumulative (%)
18-30	73	19.06	19.06
31-40	66	17.23	36.29
41-50	70	18.28	54.57
51-60	74	19.32	73.89
61-older	100	26.11	100
Total	383	100	100

Bar Chart



61-over
51-60
41-50
31-40
18-30
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Percentage of Respondents

Describing Quantitative Variables

- Several graphical options for quantitative variables:
  - Dot Plot-useful for smaller data files
  - Stem-and-Leaf Plot-also for smaller data files
  - Histogram-larger data files

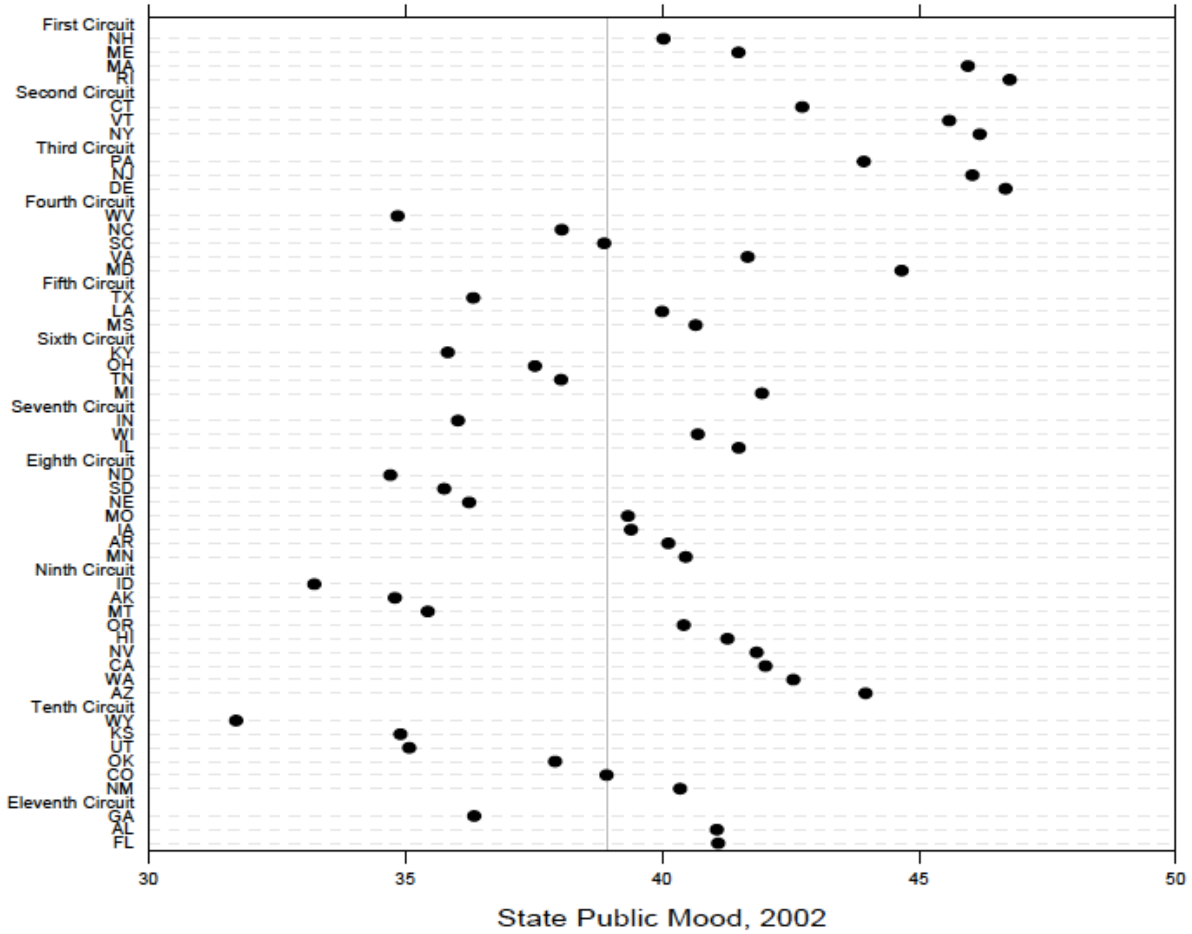
Stem-and-Leaf Plot

- Stem—all but the final (rightmost) digit
- Leaf—the final digit
- Stems go on a vertical column
- Write the leafs for each stem
- Gives you a sense of the distribution of the data
- I have never done this, nor have I ever seen anyone else present one of these

Stemplot

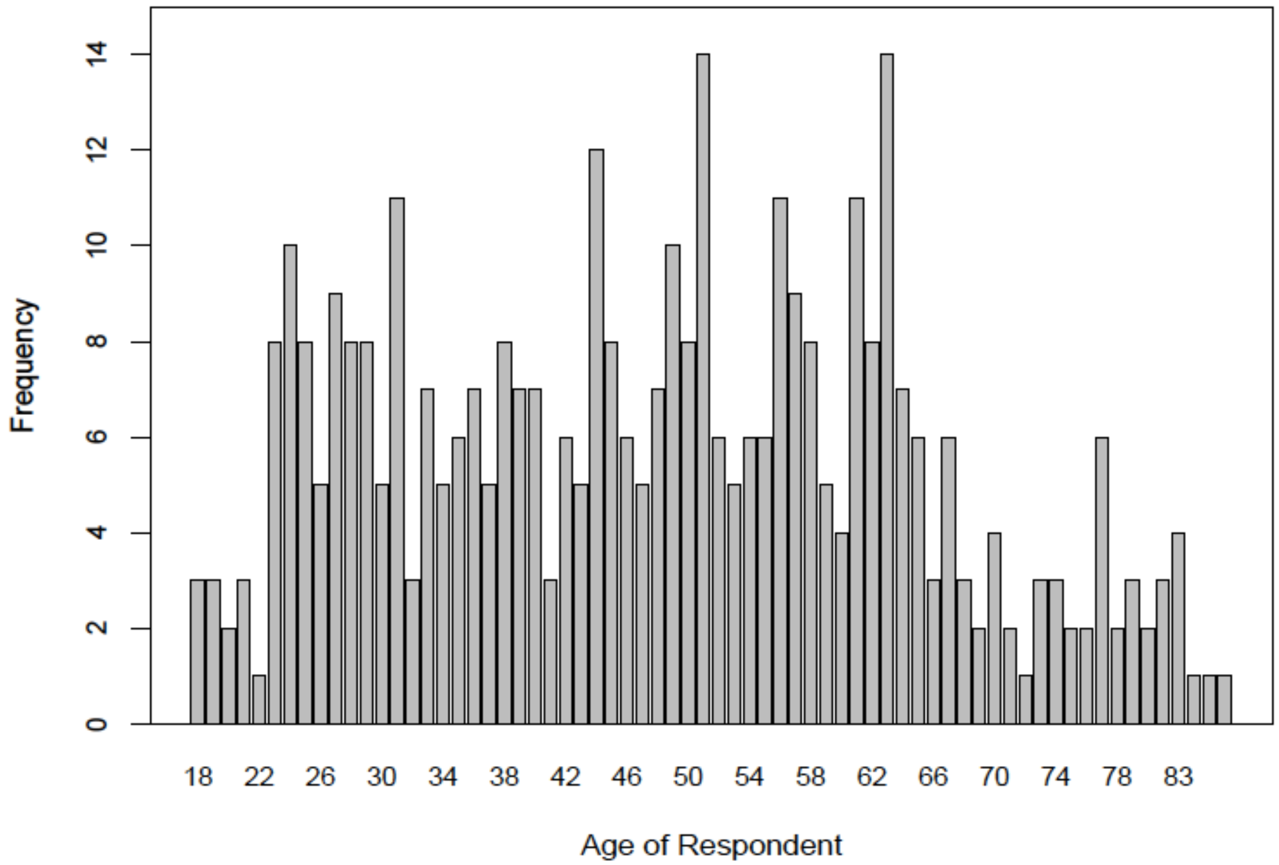
- These are the scores on an exam in an undergrad statistics class:
- 55, 63, 68, 69, 71, 74, 77, 79, 81, 81, 82, 83, 84, 84, 85, 86, 87, 87, 88, 88, 89, 90, 91, 93, 93, 95, 97, 98, 100
- Can also use stemplots for comparison
- Men—55, 63, 68, 69, 74, 77, 81, 81, 83, 85, 86, 87, 88, 93, 95
- Women—71, 79, 82, 84, 84, 87, 88, 89, 90, 91, 93, 97, 98, 100

Dotplot



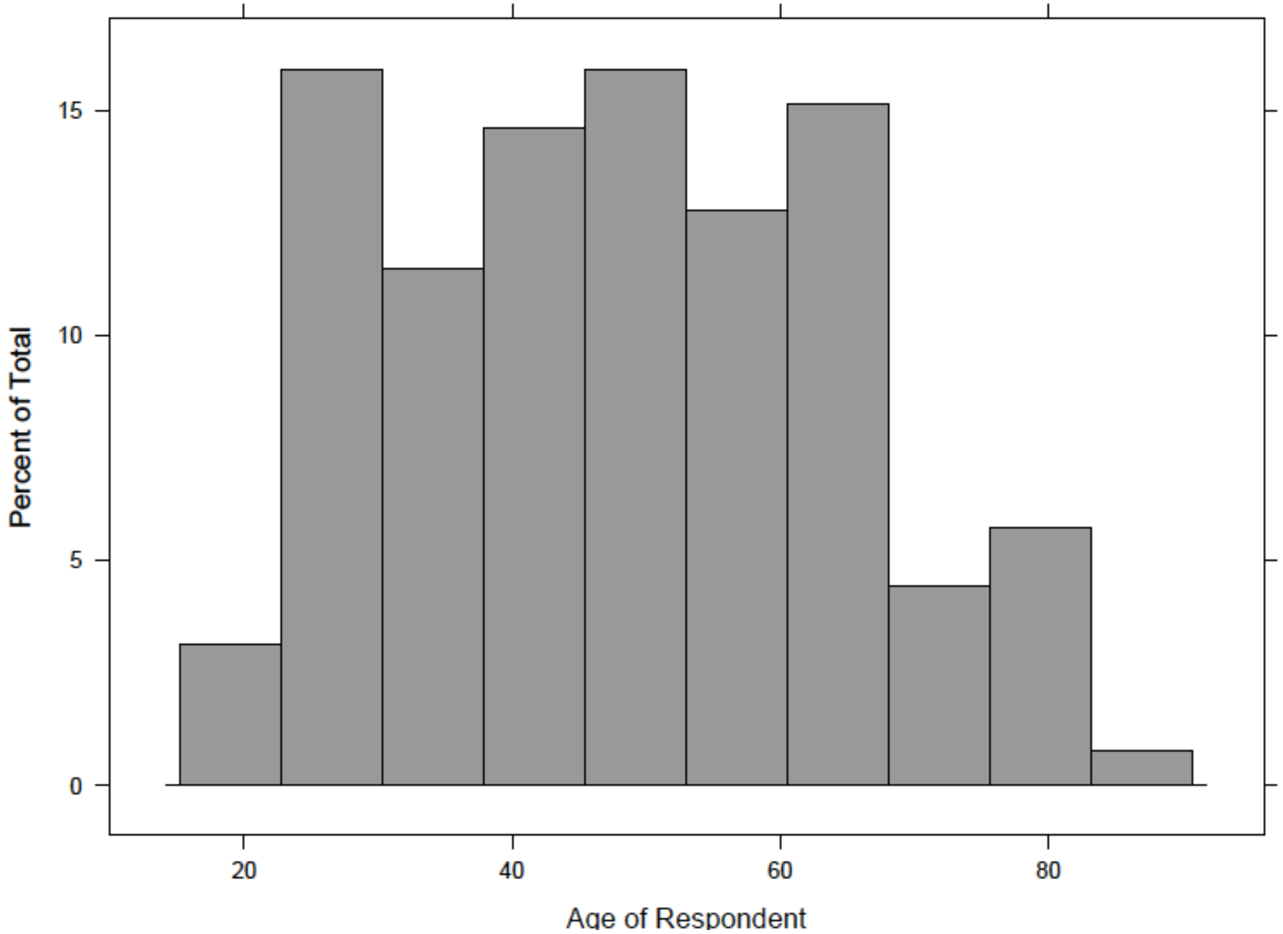
First Circuit
NH
ME
MA
RI
Second Circuit
CT
VT
NY
Third Circuit
PA
NJ
DE
Fourth Circuit
WV
NC
SC
VA
MD
Fifth Circuit
TX
LA
MS
Sixth Circuit
KY
OH
TN
MI
Seventh Circuit
IN
WI
IL
Eighth Circuit
ND
SD
NE
MO
IA
AR
MN
Ninth Circuit
ID
AK
MT
OR
HI
NV
CA
WA
AZ
Tenth Circuit
WY
KS
UT
OK
CO
NM
Eleventh Circuit
GA
AL
FL
30
35
40
45
50
State Public Mood, 2002

Histogram

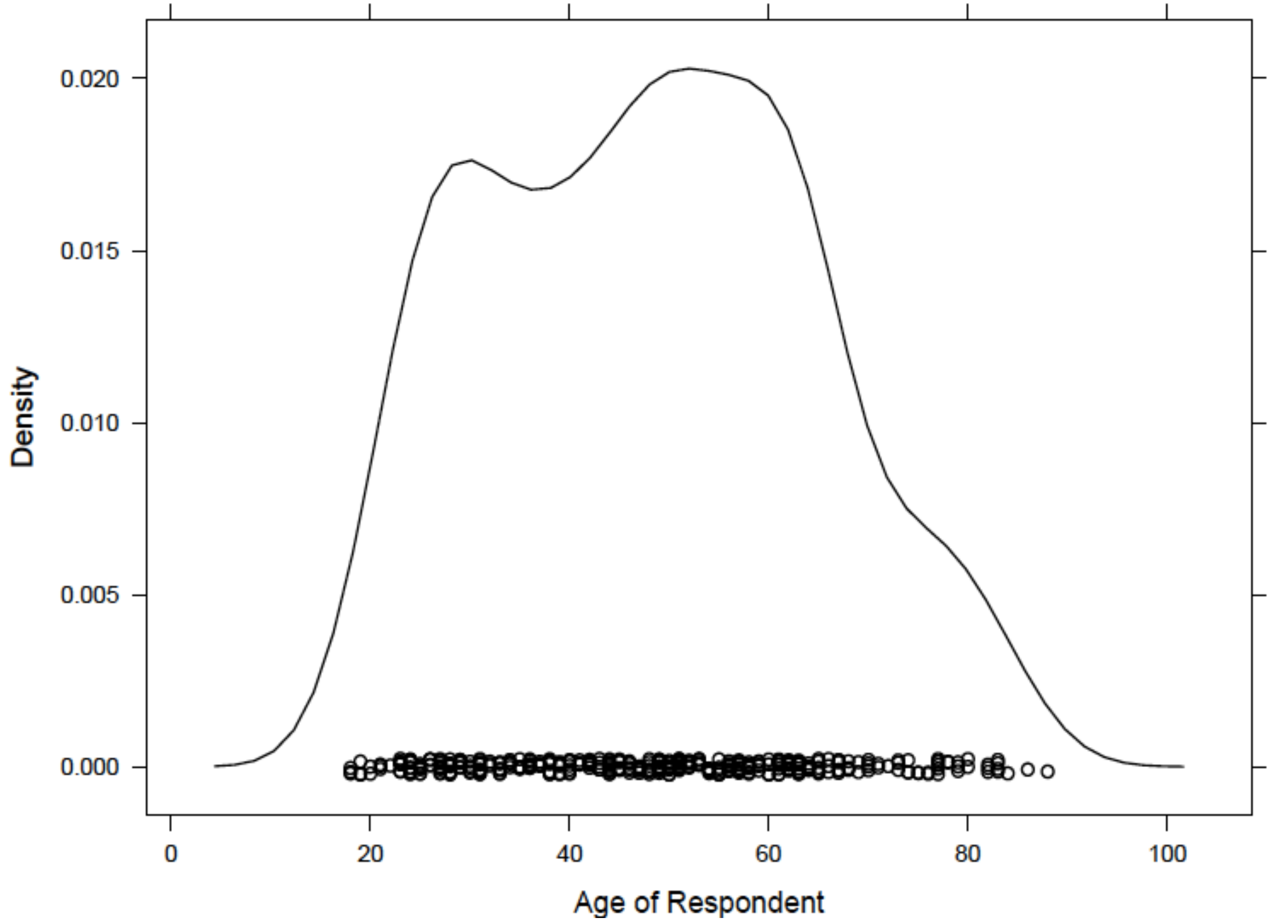


Frequency
0
2
4
6
8
10
12
14
18 22 26 30 34 38 42 46 50 54 58 62 66 70 74 78 83
Age of Respondent

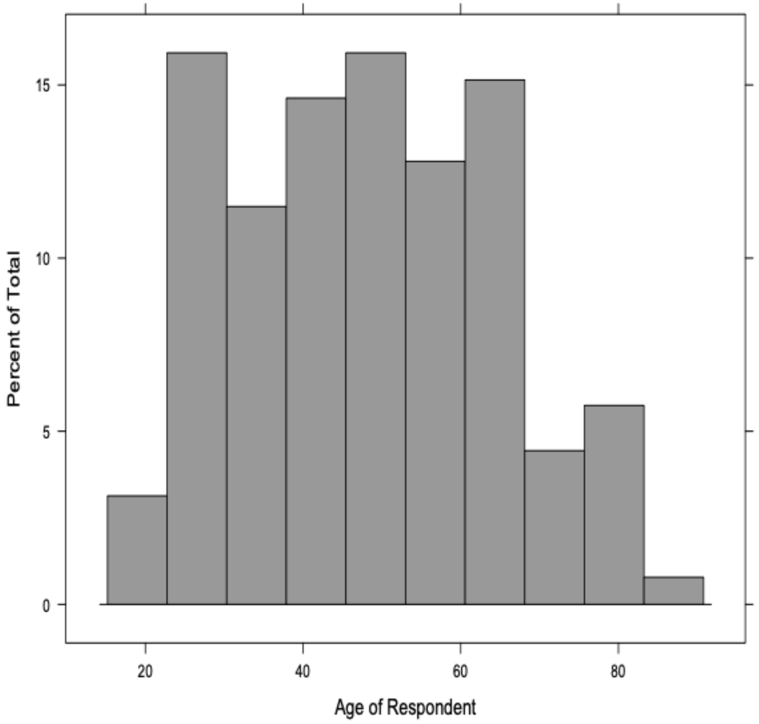
Histogram (with a larger “bin” width)

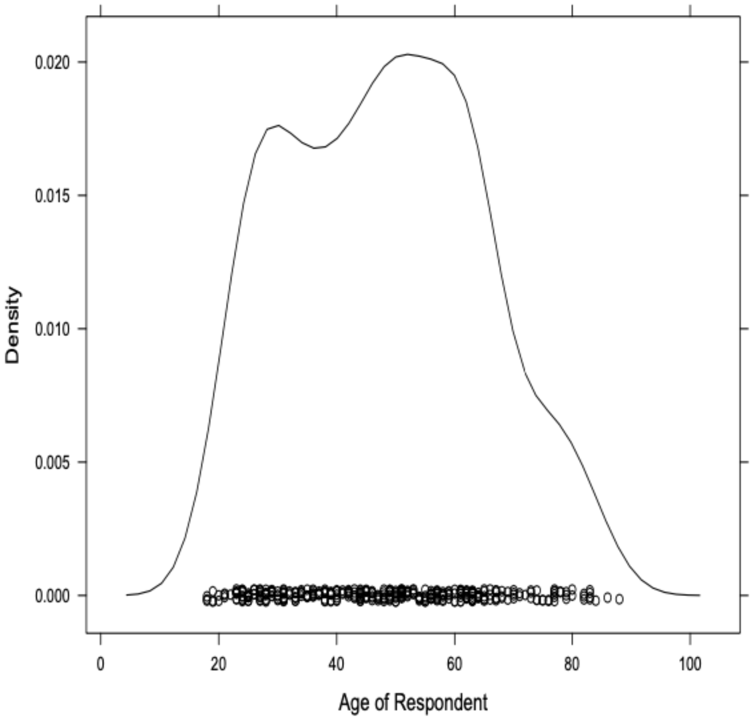


Density Curve

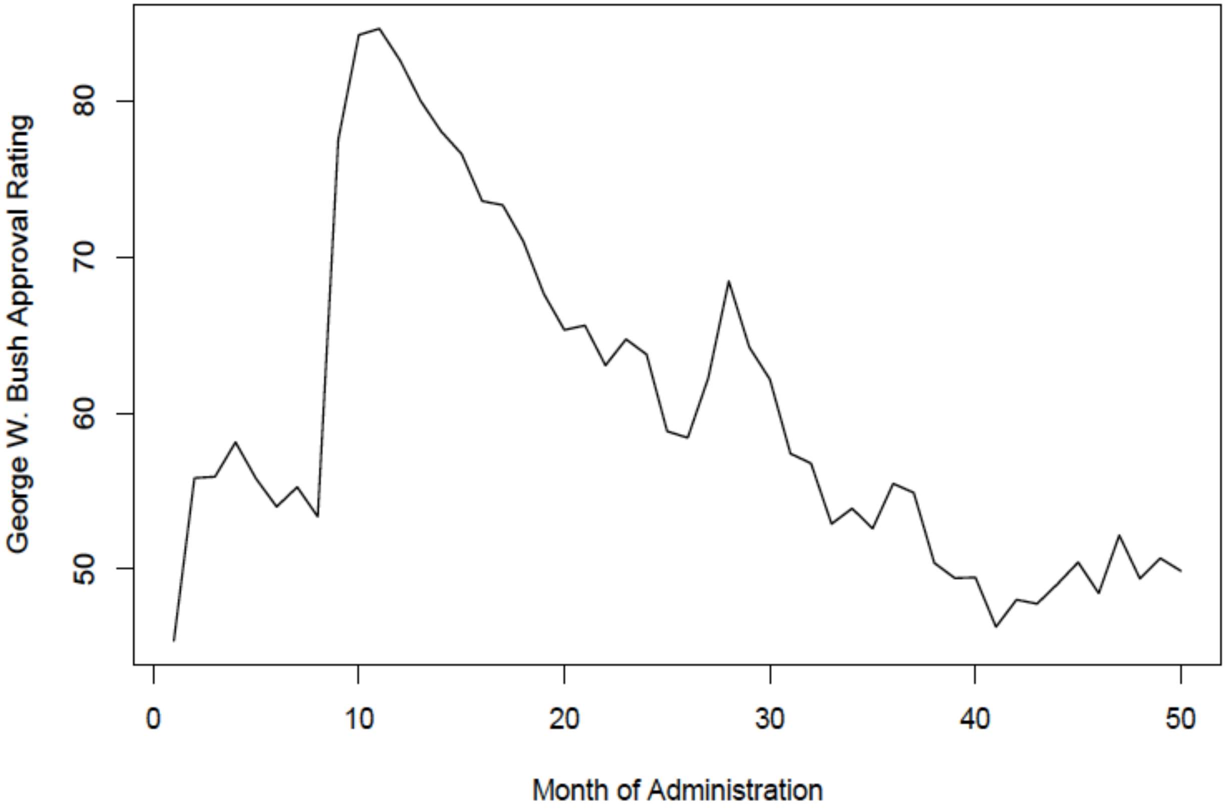


Histogram and Density Curve





Line/Time Plots

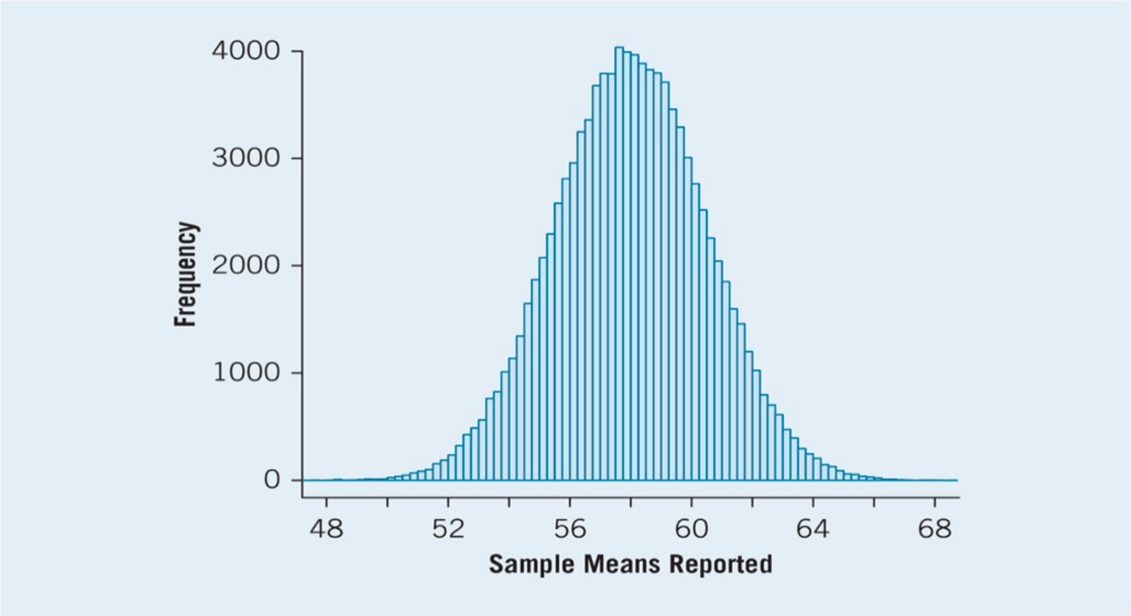


Shapes of Distribution

- Is the distribution symmetric or skewed?
- How many modes does the distribution have?
  - Unimodal
  - Bimodal
- Are their outliers or deviations from the overall shape?

Normal Distribution

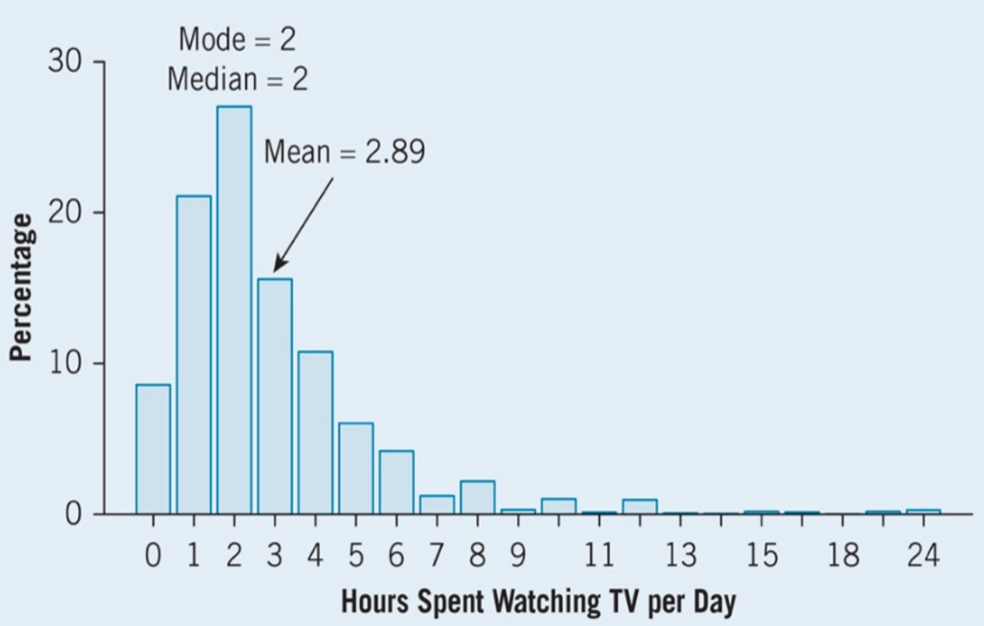
Figure 6-2 Distribution of Means from 100,000 Random Samples



Note: The figure shows means from 100,000 samples of $n = 100$. Population parameters: $\mu = 58$ and $\sigma = 24.8$.

Skewed Distribution

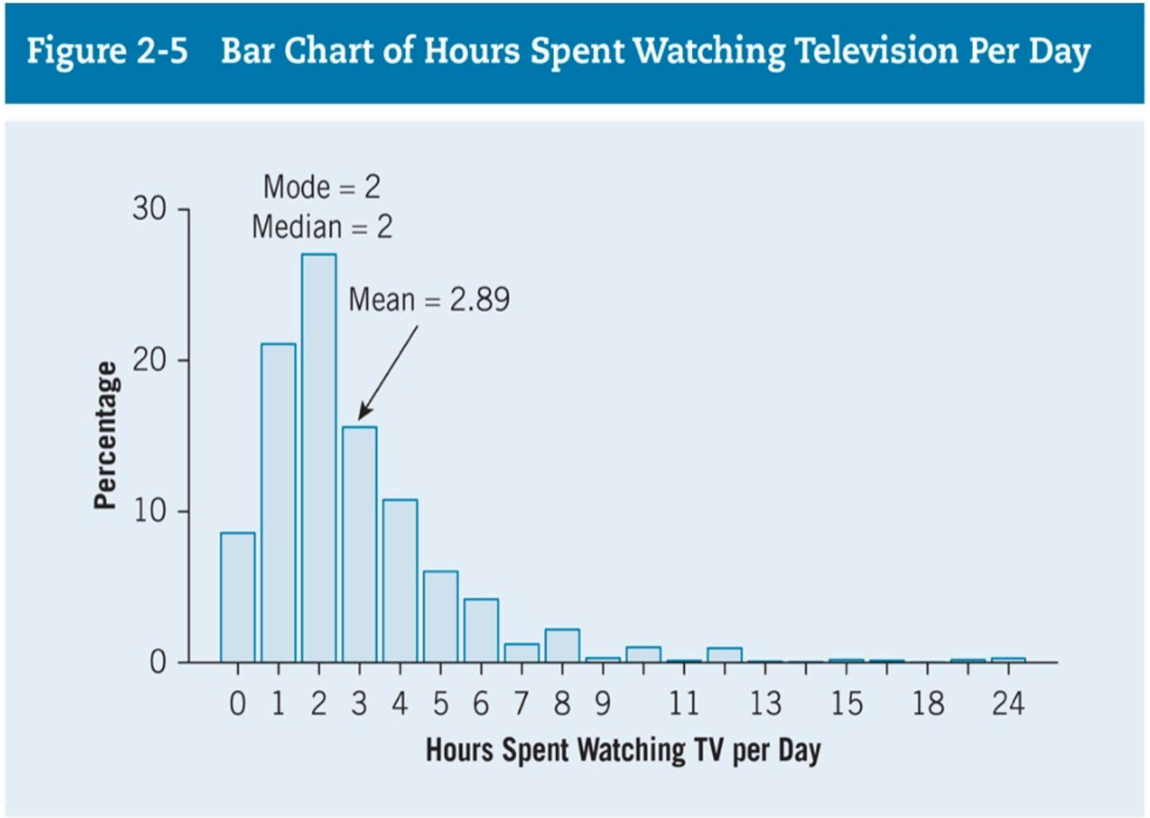
Figure 2-5 Bar Chart of Hours Spent Watching Television Per Day



Measures of Central Tendency

- Mean: Average of all values
  - $\bar{x} = \frac{\sum x_i}{n}$
- Median: midpoint of the observations
- Mode: value that occurs most frequently (often used for categorical data)
- Outliers: An observation that falls well above or below the overall data
  - The mean is *sensitive* to outliers
  - The median is *resistant* to outliers

Graphical Display of Central Tendency



Quartiles

- Splits the data into four parts
- The median is the second quartile, Q_2
- The first quartile, Q_1, is the median of the lower half of the observations
- The third quartile, Q_3, is the median of the upper half of the observations
- The interquartile range is the distance between the third quartile and first quartile
  - IRQ=Q_3-Q_1

Five number summary

- Minimum value
- Q_1
- Median
- Q_3
- Maximum value

Five number summary

- Below is a random sample of 40 tree diameters, in centimeters, from the Wade Tract in Thomas County, Georgia. What is the five-number summery for these data?

#	D.	#	D.	#	D.	#	D.
1	2.2	11	11.4	21	29.1	31	43.3
2	2.2	12	11.4	22	31.5	32	43.6
3	2.3	13	13.3	23	31.8	33	44.2
4	2.7	14	16.9	24	32.6	34	44.4
5	4.3	15	17.6	25	35.7	35	44.6
6	4.9	16	18.3	26	37.5	36	47.2
7	5.4	17	22.3	27	38.1	37	51.5
8	7.8	18	26	28	39.7	38	51.8
9	9.2	19	26.1	29	40.3	39	52.2
10	10.5	20	27.9	30	40.5	40	69.3

Five number summary

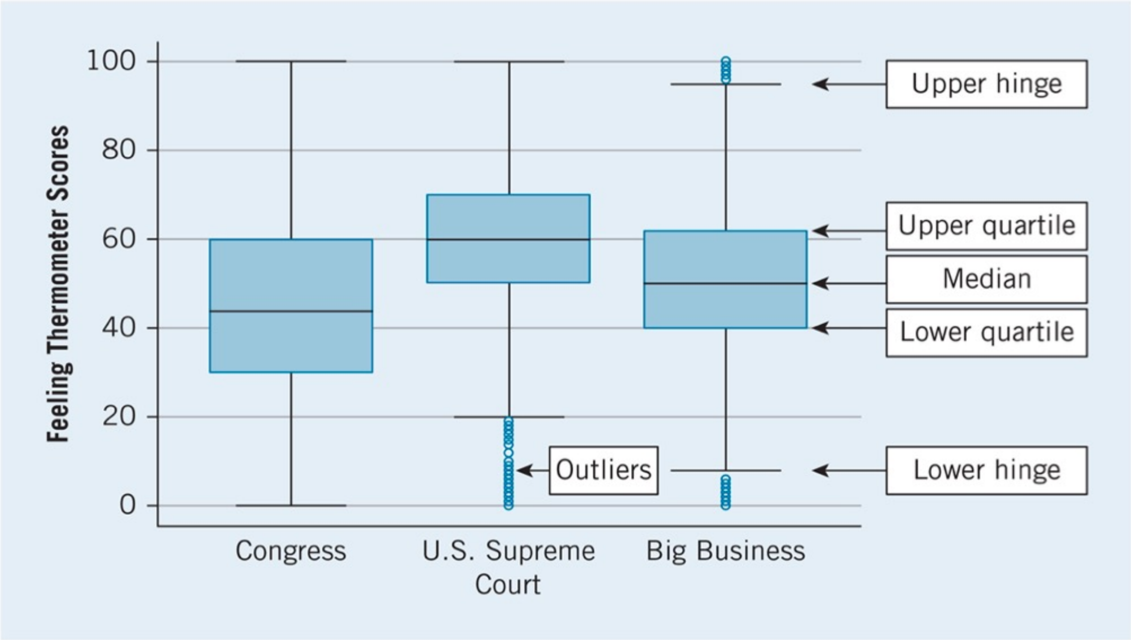
- Min=2.2cm
- Q_1=10.95cm
- M=28.5cm
- Q_3=41.9cm
- Max=69.3cm

Boxplot

- Construct a box from Q_1 to Q_3 (i.e., the IQR)
- Draw a line inside the box at the median value
- Draw a line out to the lowest value that is not an outlier
- Draw a line out to the highest value that is not an outlier
- Rule of thumb for outliers—an observation is a suspected outlier if it is more than 1.5 times the IQR above the third quartile (Q_3), or below the first quartile (Q_1).

Boxplot

Figure 2-7 Box Plots of Three Feeling Thermometer Variables



Source: 2016 American National Election Study.

Measures of Spread

- Range: difference between the largest and smallest observations
  - Range=Max - Min
- Variance: average of the squares of the deviations of the observations from their mean
  - $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$
- Standard deviation: square root of the average squared deviation from the mean
  - $s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$

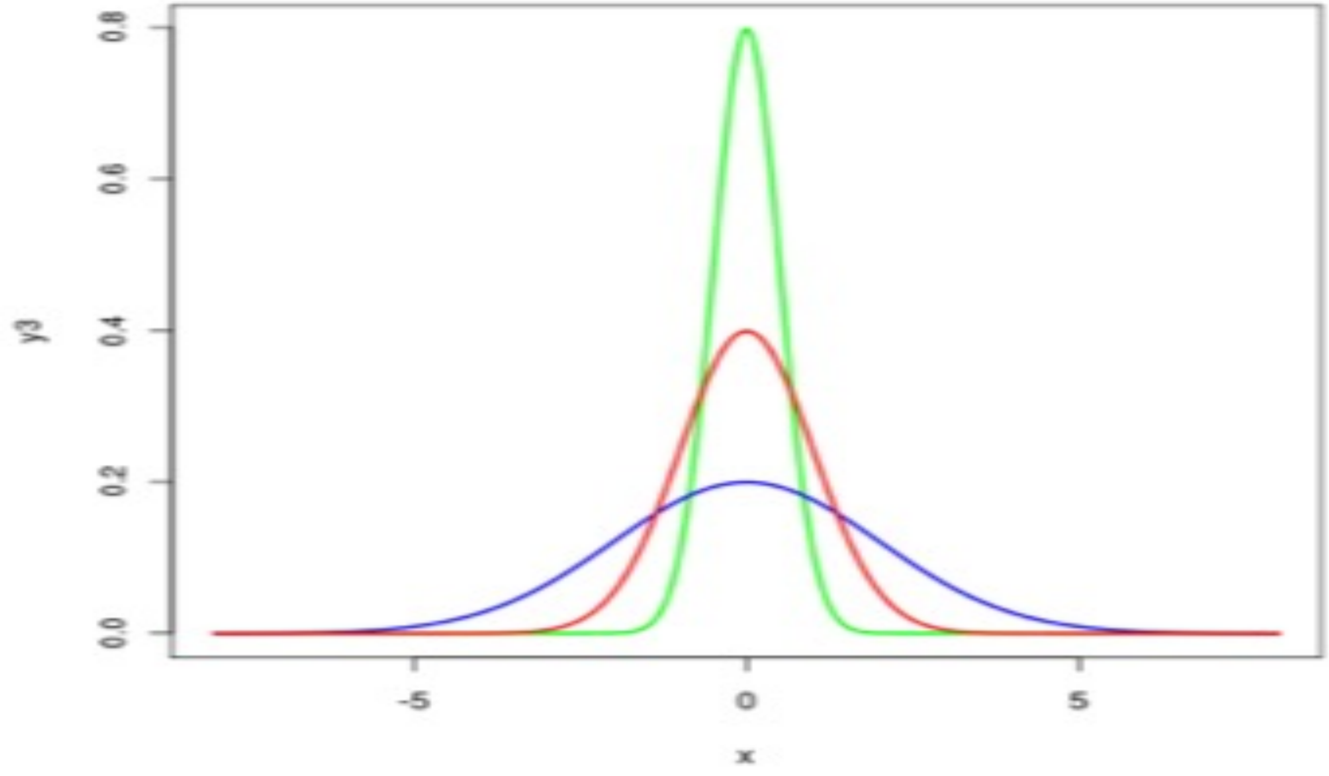
Standard Deviation

- S.D. should only be used with mean, not median, to describe spread
- If SD=0, there is no spread (i.e., all observations are the mean)
- SD is *not* resistant to outliers
  - More sensitive then mean, it reflects squared deviations from the mean

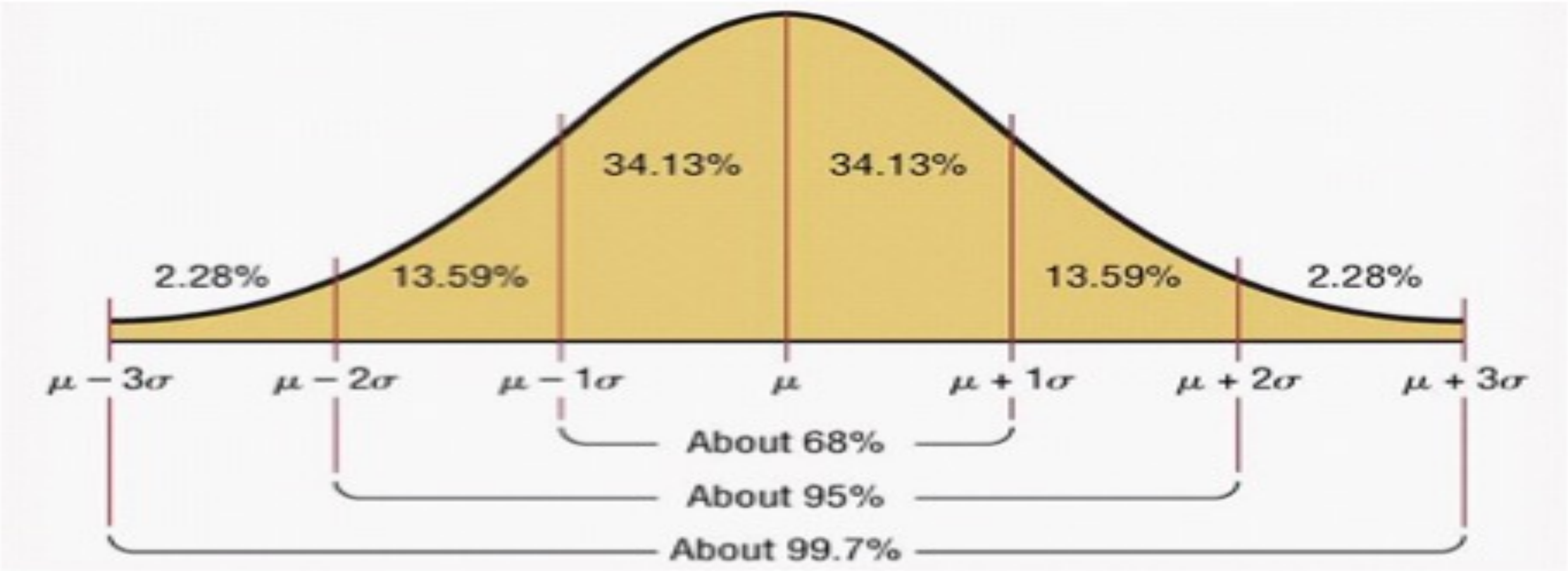
Linear Transformation

- Sometimes we want to change the units of a variable through a linear transformation
- A linear transformation applies the same linear equation to each observation of x
  - $x_{new} = a + b(x_{orig})$
  - Example—Convert Fahrenheit to Celsius
  - $C = (\frac{5}{9}(F - 32)$
- Linear transformations **do not** affect the shape of the distribution
- They do affect the measures of center and spread, but in a predictable way
  - Center—add "a" and multiply center by "b"
  - Spread—multiply spread by "b", do not add "a"

Normal Distribution



Normal Distribution



Empirical Rule: for bell-shaped sets of data

- Approximately 68% of cases fall within 1 standard deviation of the mean
- Approximately 95% of cases fall within 2 standard deviations of the mean
- Approximately 99.7% of cases fall within 3 standard deviations of the mean

Normal distribution

- The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and a variance of 256 days. Use the empirical rule to answer the following questions.
  - Between what values do the lengths of the middle 95% of all pregnancies fall?
  - How short are the shortest 2.5% of all pregnancies?
  - How long do the longest 2.5% last?

Normal distribution

- The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and a variance of 256 days. Use the empirical rule to answer the following questions.
  - Between what values do the lengths of the middle 95% of all pregnancies fall?
    - The middle 95% fall within two standard deviations of the mean:
    - 266+/-2(16)=236 to 298 days
  - How short are the shortest 2.5% of all pregnancies?
  - How long do the longest 2.5% last?

Normal distribution

- The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and a variance of 256 days. Use the empirical rule to answer the following questions.
  - Between what values do the lengths of the middle 95% of all pregnancies fall?
    - The middle 95% fall within two standard deviations of the mean:
    - 266+/-2(16)=236 to 298 days
  - How short are the shortest 2.5% of all pregnancies?
    - The shortest 2.5% are shorter than 234 days
  - How long do the longest 2.5% last?

Normal distribution

- The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and a variance of 256 days. Use the empirical rule to answer the following questions.
  - Between what values do the lengths of the middle 95% of all pregnancies fall?
    - The middle 95% fall within two standard deviations of the mean:
    - 266+/-2(16)=234 to 298 days
  - How short are the shortest 2.5% of all pregnancies?
    - The shortest 2.5% are shorter than 234 days
  - How long do the longest 2.5% last?
    - The longest 2.5% are longer than 298 days.

Z-scores

- Because of the empirical rule, we can measure (in a standardized manner) how far away observations are from the mean along a normal distribution
- Z-score: how many s.d.s away from the mean the observation is
  - $z_i = \frac{x_i - \mu_x}{\sigma_x}$
- Cumulative percentages
  - When we know the Z-score, we can use a table to tell us the percentage of cases (i.e., the area under the curve) that are to the left, or right, of that specified location on the distribution

Areas under the Standard Normal Curve

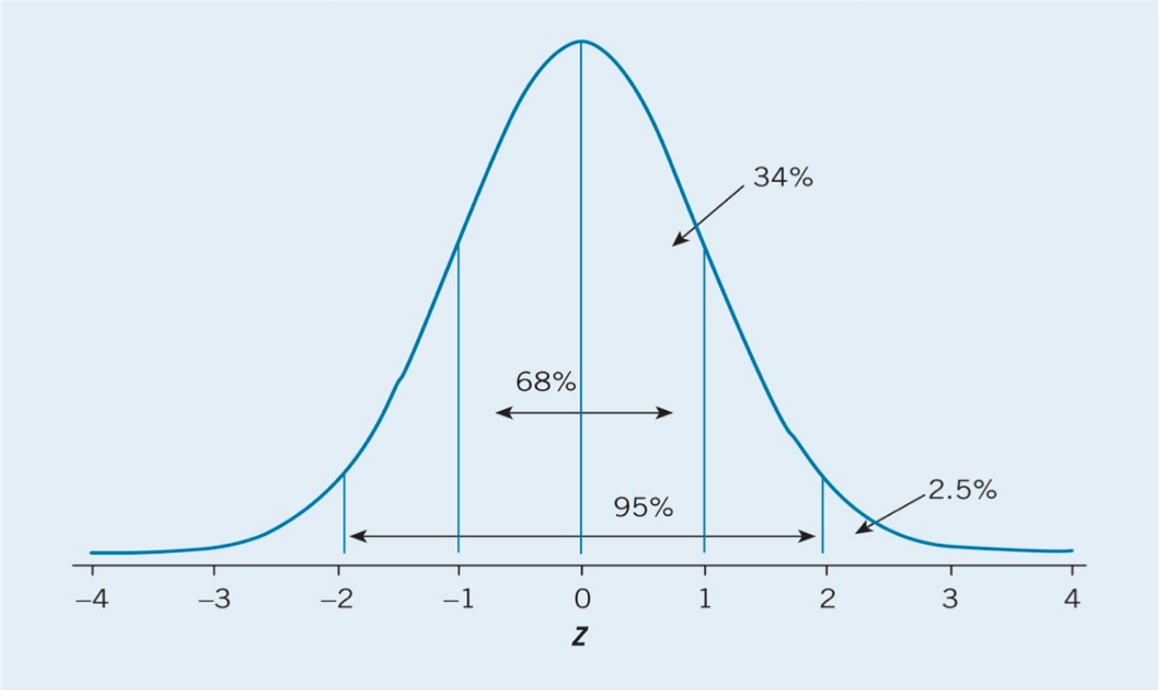


Figure 6-5 Areas under the Standard Normal Curve

Z-Table

Standard Normal Probabilities

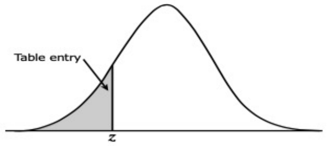


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Z-Table

Standard Normal Probabilities

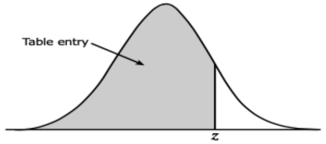


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Conclusion

- Take home points:
  - Always plot your data first
  - Techniques for describing data differ based on type of variables
  - Important to consider outliers/skew when deciding which measures to use
  - Normal distributions have special properties, will be very important for inferential statistics
- Next week we will focus on techniques for examining relationships between variables
  - Begin to introduce the conceptual foundations of linear regression