Lecture 04
Comparing distributions, Normal model

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Key Points

Chapter 3: Displaying categorical data

- Bar chart for categorical data
- Pie chart for proportions of whole
- Faithful reporting and the area principle
- Contingency tables
- Simpson's paradox

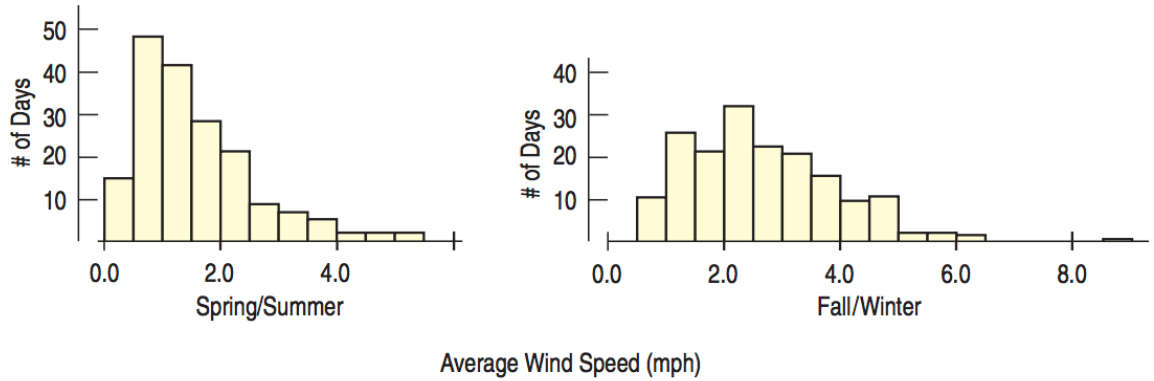
Chapter 4: Displaying quantitative data

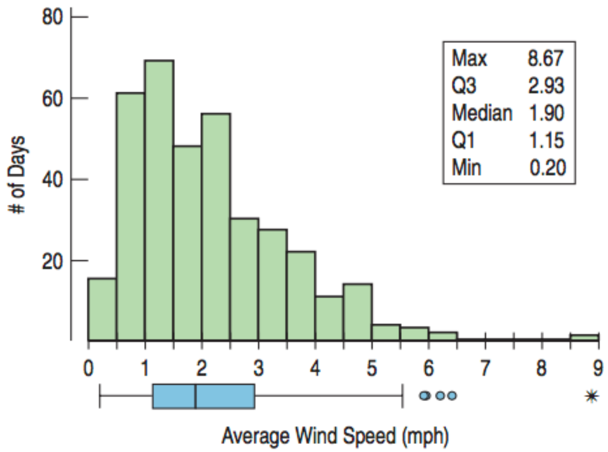
- Histograms, Stem-leaf, dot plots
- Shape (mode, symmetrical)
- Center (median, mean)
- Spread (range, IQR, variance, standard deviation)
- Box plots

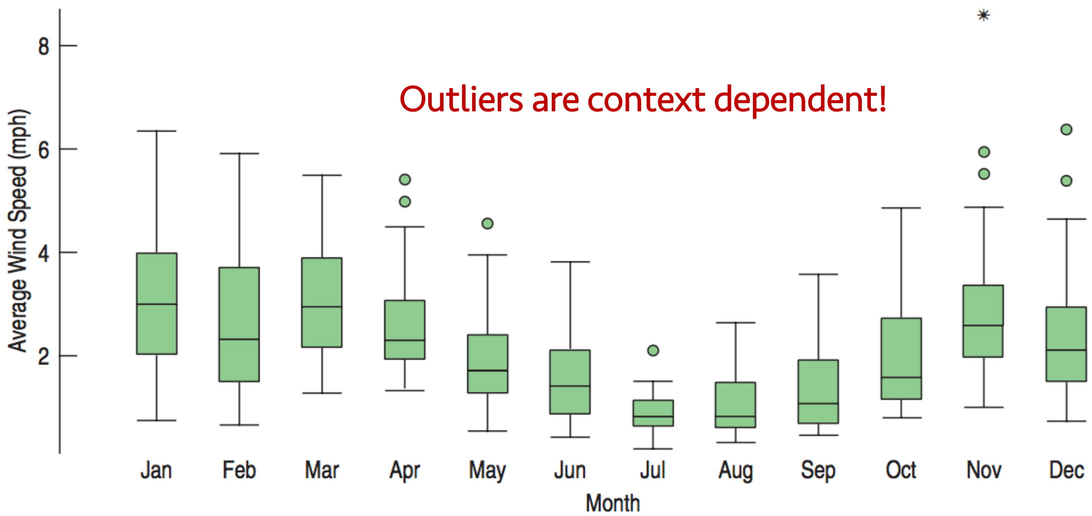
Comparing Groups with Histogram and Boxplots

Data:

- When: Days during 2011
- What: Average daily wind speed (mph)
- Where: Hopkins Forest in western Massachusetts
- Why: Long-term observations to study ecology and climate







Outliers are context dependent!

Outliers

- Outliers may be the most important values.
- Or they may be just errors.

- What to do with them?
  1. Correct them if possible
  2. Report summaries and analyses with and without the outliers (readers can decide)
  3. Some statistical methods: down-weight them (e.g., robust regression), smoothing, etc.
  4. *Never* do:
     - Leave them in place and proceed as if nothing were unusual
     - Omit an outlier from the analysis without comment

Re-expressing Data

- To improve symmetry:

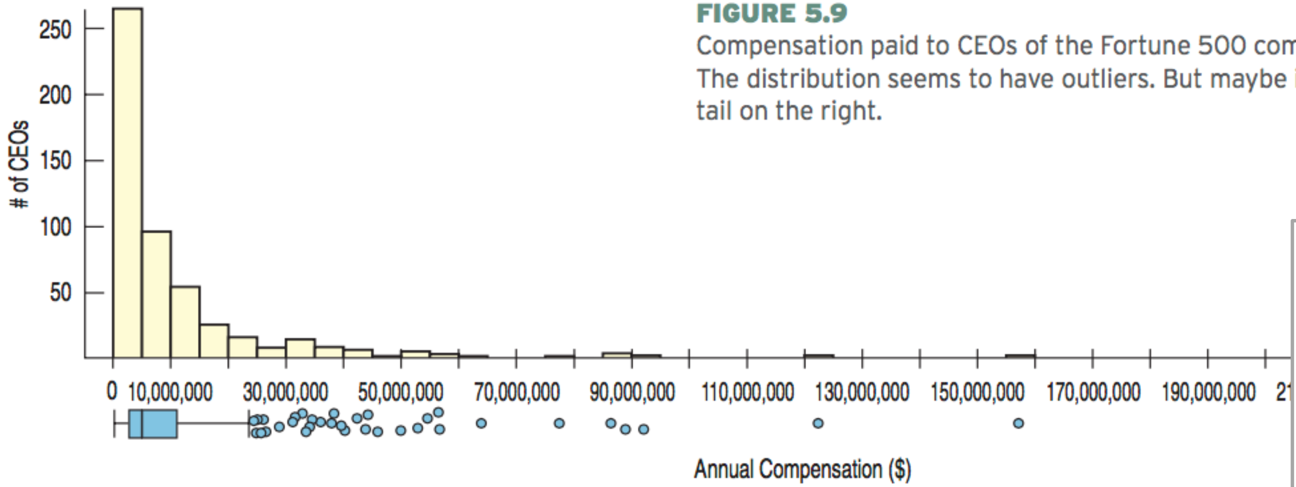
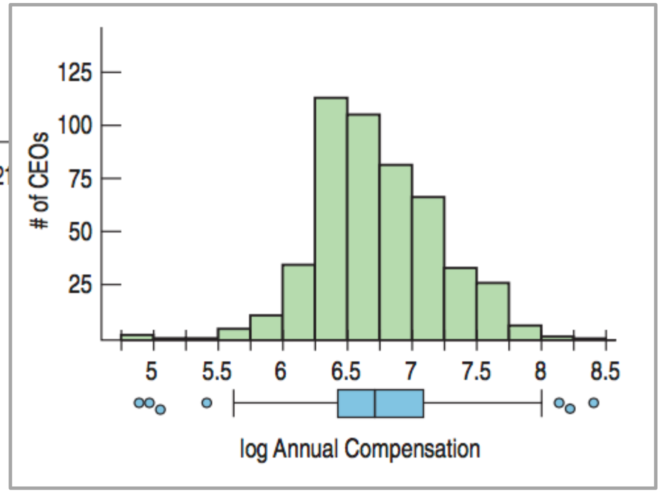


FIGURE 5.9
Compensation paid to CEOs of the Fortune 500 companies in 2005. The distribution seems to have outliers. But maybe it just has a long tail on the right.

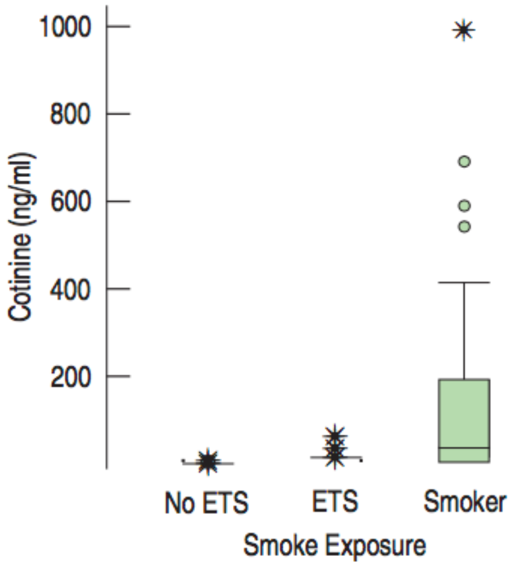
Logarithms of 2005 CEO compensations



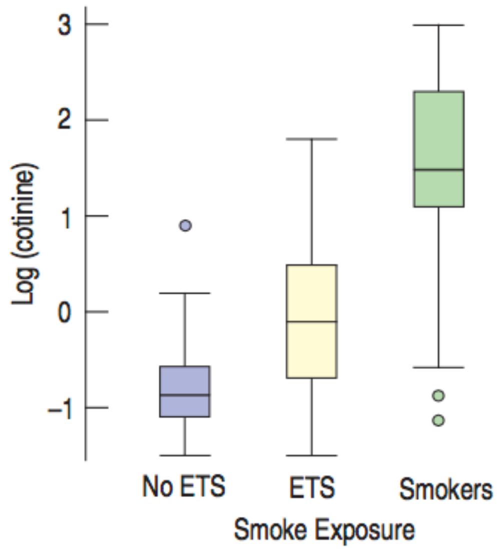
- Very common in physio data preprocessing
  - E.g., logarithm, square root (normalization)

Re-expressing Data

- To equalize spread across groups



Log-transform

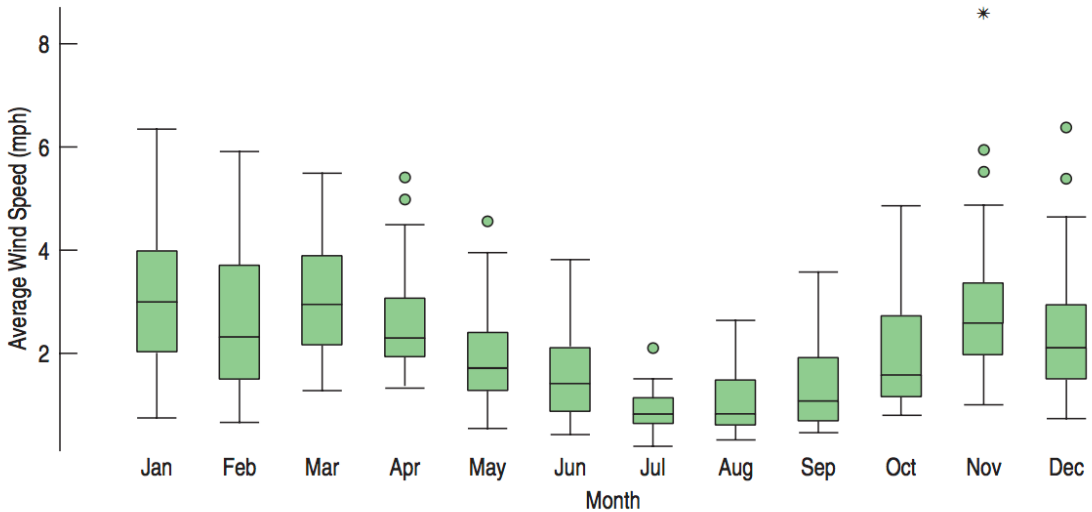


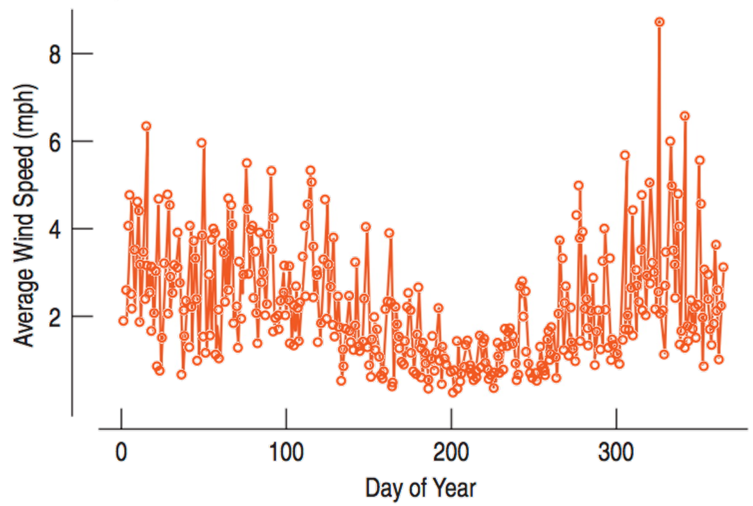
*ETS: exposed to smoke

- **Normalization** is one of the key elements of recent successes of artificial intelligence (machine learning)

Timeplots

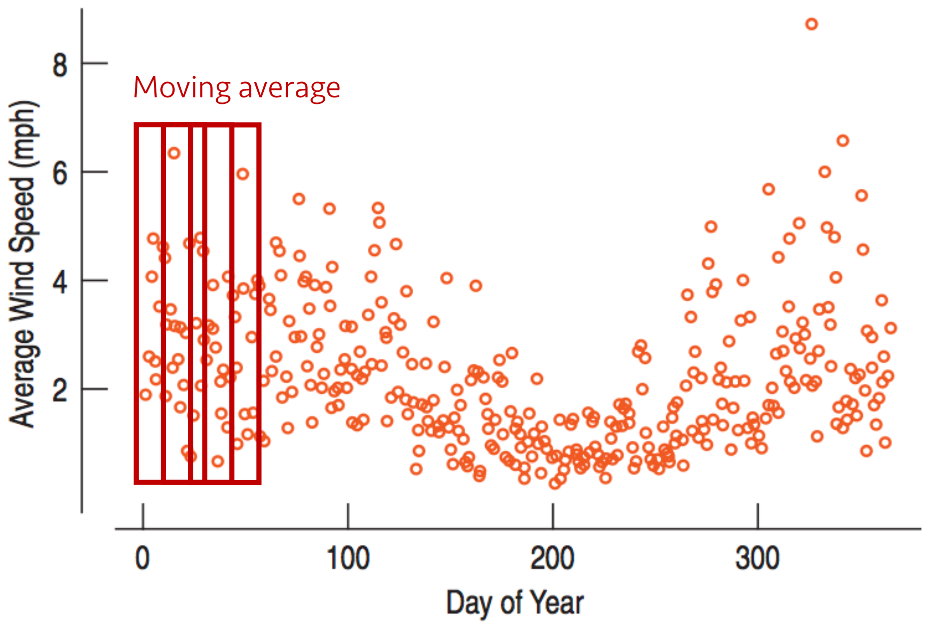
- Values against time: **timeplot, time-series data** (e.g., neural activity data, stock market, bio-sensor...)

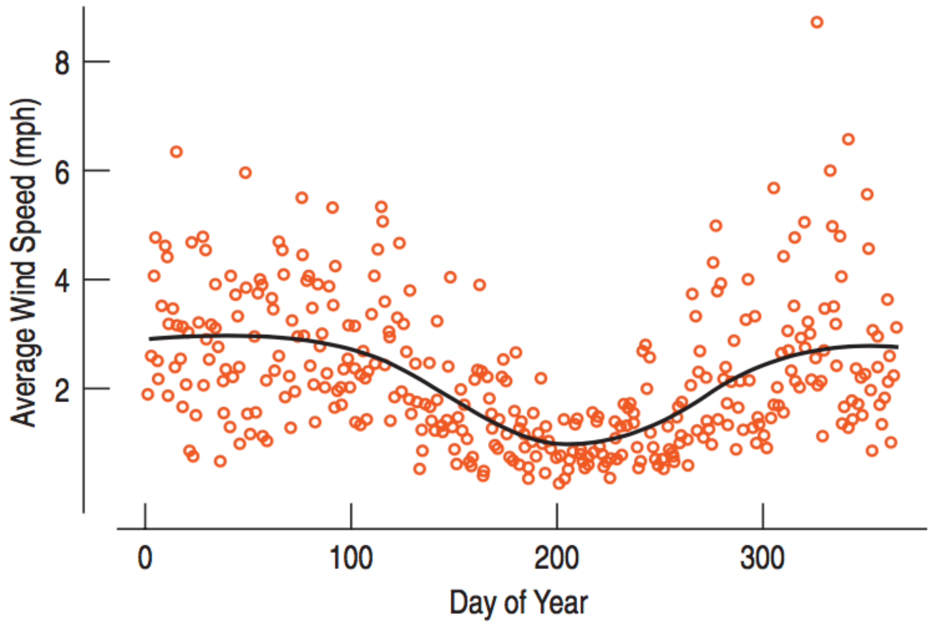




Smoothing timeplots

- **Lowess** (locally weighted scatterplot smoothing)





Key Points

Chapter 5: Comparing distributions: considerations

- Outliers are context dependent
- Re-expressing data (log, sqrt)
  - to improve symmetry
  - to equalize spread
- Timeplots
- Moving-averages, smoothing

z-Scores

- To compare different values in different units,
- the values should be *standardized!*

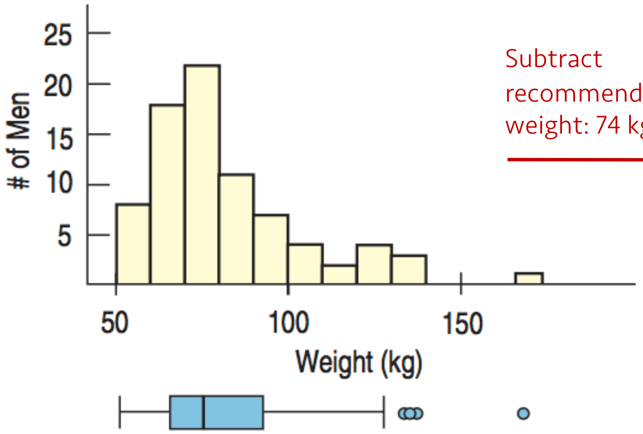
$$z = \frac{y - \bar{y}}{s}$$

  - *z*-scores: mean = 0, standard deviation = 1
  - standardized values
  - = using standard deviation as a ruler!

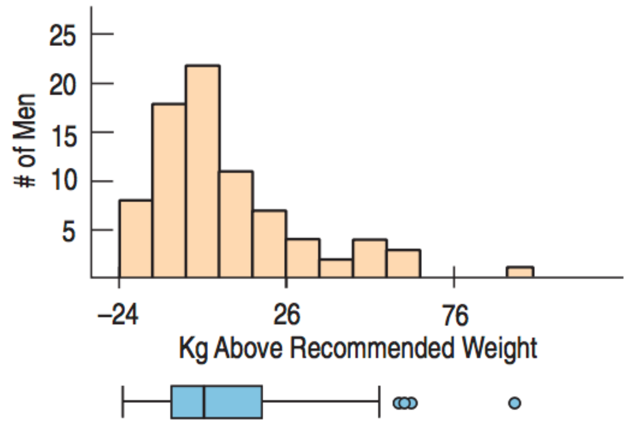
- Two elements: shifting and scaling

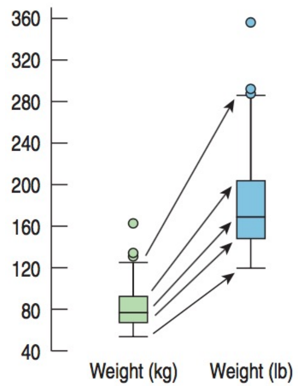
Shifting and rescaling

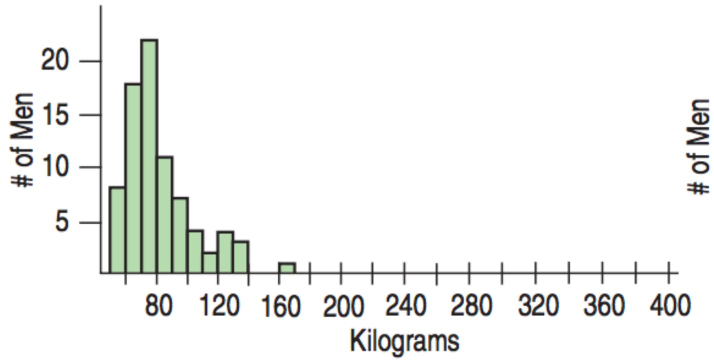
WHO	80 male participants of the NHANES survey between the ages of 19 and 24 who measured between 68 and 70 inches tall
WHAT	Their weights
UNIT	Kilograms
WHEN	2001–2002
WHERE	United States
WHY	To study nutrition, and health issues and trends
HOW	National survey

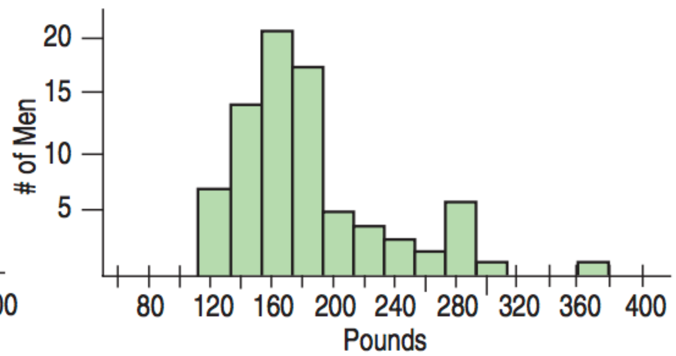


Subtract recommended weight: 74 kg









Normal models: When is a *z*-score big?

- Let's say you've got a *z*-score of 3. How surprising your observation is?
- To answer this question, you need a *model* of your data's distribution.

- *"All models are wrong, but some are useful." George Box*

- Most popular model: Normal models (bell-shaped curves)
  - unimodal, symmetric
  - $N(\mu, \sigma)$, where μ is mean, σ is standard deviation
  - Why Greek? These are parameters of the *model*, not *numerical summaries* of data
    - Numerical summaries of the data: $\bar{y}$, and s
    - We still call the standardized value a *z-score*.
    - $z \sim N(0, 1)$: standard Normal model

$$z = \frac{y - \bar{y}}{s} \longrightarrow z = \frac{y - \mu}{\sigma}$$

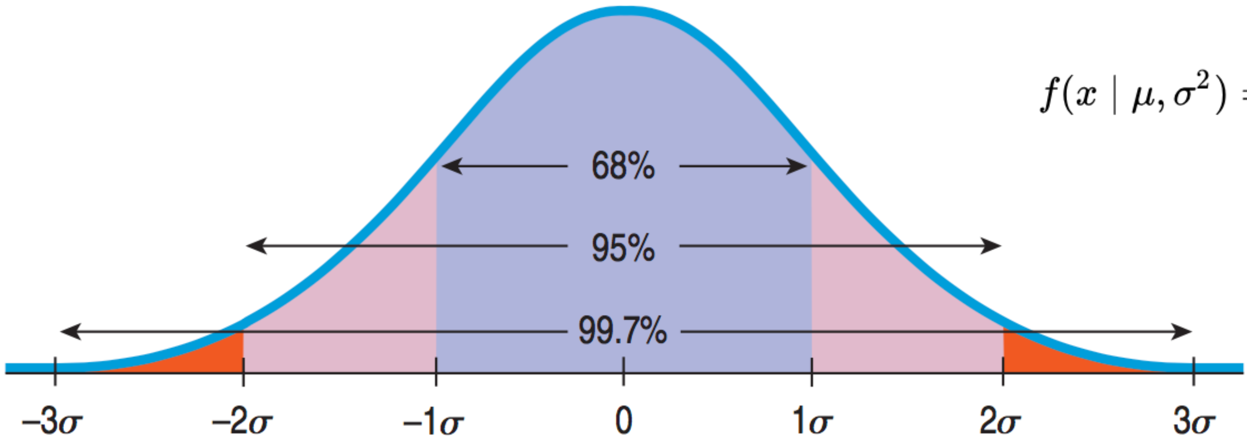
Normality assumption

- All models make assumptions, which should be carefully examined.

- **Nearly normal condition** (it's sufficient):
  - shape: unimodal, symmetric
  - We can check it with *histogram* or a *normal probability plot.*

The 68–95–99.7 Rule

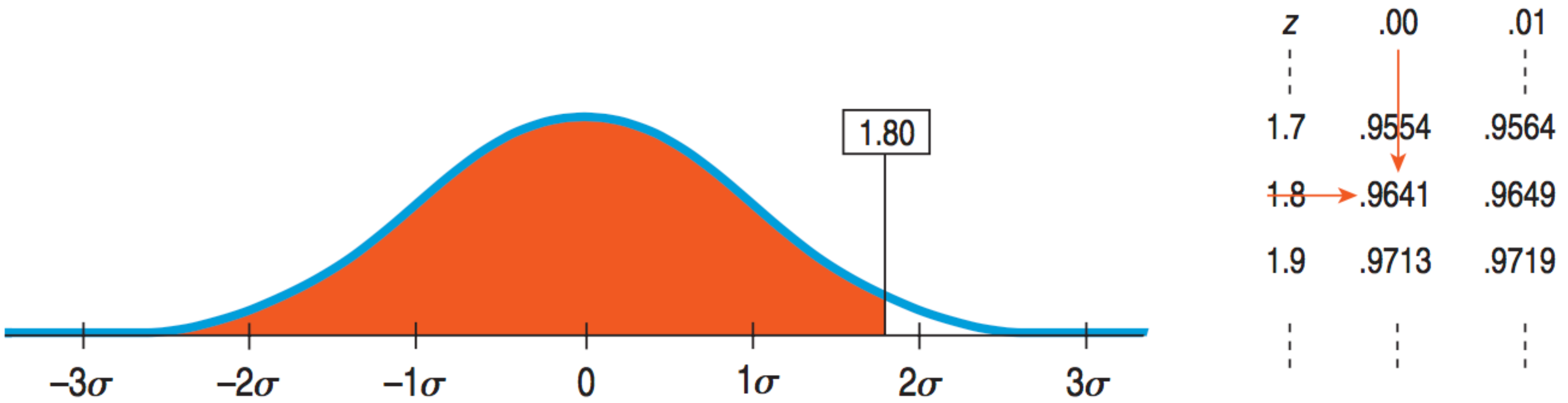
- 68% of the data fall within 1 standard deviation, 95% within 2 std, 99.7% within 3 std.
- *z*-score = 1 means, you are 84%! Why? 100-(50-68/2) = 84%



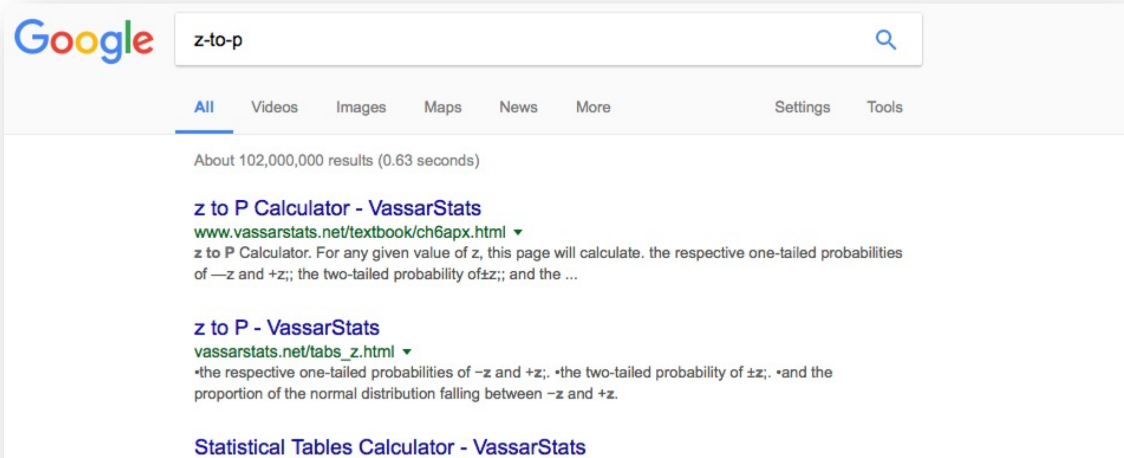
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Finding Normal percentiles

- A table of Normal percentiles (Table Z in Appendix D)



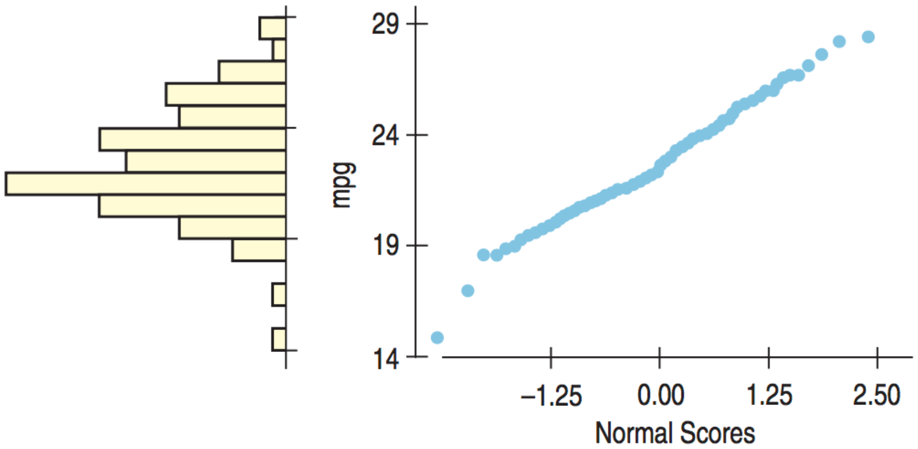
- Or google it!
- p-to-z is same: Table Z again.



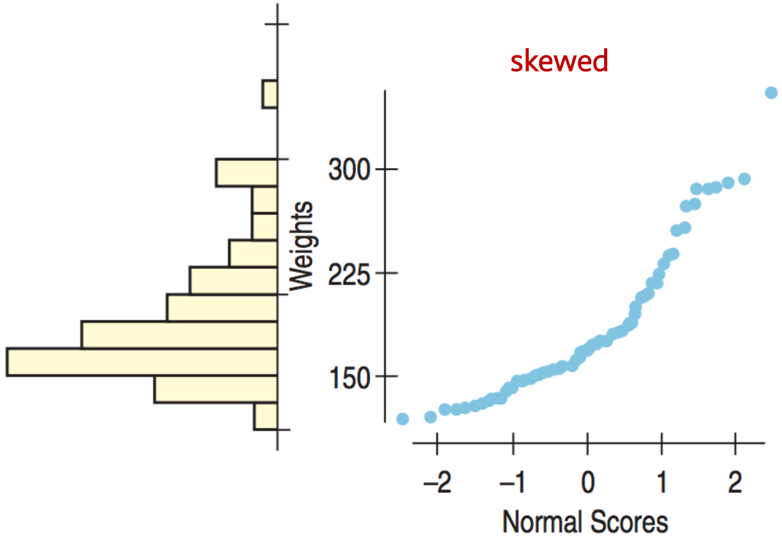
Normal probability plots

- Original value against normal scores (theoretically expected value):
- Similar to Q-Q plot (quantile-quantile plot)

symmetric



skewed



Procedure

1. Sorting (draw a histogram)
2. Get percentile
3. P-to-Z
4. Make a scatter plot

   y-axis = Original values

   x-axis = Normal scores

Example

1. Sorting (draw a histogram)
2. Get percentile
3. P-to-Z
4. Make a scatter plot

   y-axis = Original values

   x-axis = Normal scores

Data

```
11.4 13.6 10.2 11.7 16.3 16.2 7.2 11.2 17.7 10.8 10.2 17.1 14.5 6.7 10.0 4.4 8.8 14.5 3.6 15.1
```

↓

1. Sorting

```
1 ~ 12번 열

  3.6000    4.4000    6.7000    7.2000    8.8000   10.0000   10.2000   10.2000   10.8000   11.2000   11.4000   11.7000

13 ~ 20번 열

 13.6000   14.5000   14.5000   15.1000   16.2000   16.3000   17.1000   17.7000
```

↓

2. Percentile,

If there are N numbers, percentile can be calculated as 100*((1-0.5)/N), 100*((2-0.5)/N), 100*((3-0.5)/N), ..., 100*((N-0.5)/N)

1 ~ 12번 열

Sorted value	3.6000	4.4000	6.7000	7.2000	8.8000	10.0000	10.2000	10.2000	10.8000	11.2000	11.4000	11.7000
Percentile	2.5000	7.5000	12.5000	17.5000	22.5000	27.5000	32.5000	37.5000	42.5000	47.5000	52.5000	57.5000

13 ~ 20번 열

13.6000	14.5000	14.5000	15.1000	16.2000	16.3000	17.1000	17.7000
62.5000	67.5000	72.5000	77.5000	82.5000	87.5000	92.5000	97.5000

Example

1. Sorting (draw a histogram)
2. Get percentile
3. P-to-Z
4. Make a scatter plot

   y-axis = Original values

   x-axis = Normal scores

2. Percentile,

If there are N numbers, percentile can be calculated as 100*((1-0.5)/N), 100*((2-0.5)/N), 100*((3-0.5)/N), ..., 100*((N-0.5)/N)

1 ~ 12번 열

Sorted value	3.6000	4.4000	6.7000	7.2000	8.8000	10.0000	10.2000	10.2000	10.8000	11.2000	11.4000	11.7000
Percentile	2.5000	7.5000	12.5000	17.5000	22.5000	27.5000	32.5000	37.5000	42.5000	47.5000	52.5000	57.5000

13 ~ 20번 열

13.6000	14.5000	14.5000	15.1000	16.2000	16.3000	17.1000	17.7000
62.5000	67.5000	72.5000	77.5000	82.5000	87.5000	92.5000	97.5000

↓

3. P-to-Z

1 ~ 12번 열

Sorted value	3.6000	4.4000	6.7000	7.2000	8.8000	10.0000	10.2000	10.2000	10.8000	11.2000	11.4000	11.7000
Z-values	-1.9600	-1.4395	-1.1503	-0.9346	-0.7554	-0.5978	-0.4538	-0.3186	-0.1891	-0.0627	0.0627	0.1891

13 ~ 20번 열

13.6000	14.5000	14.5000	15.1000	16.2000	16.3000	17.1000	17.7000
0.3186	0.4538	0.5978	0.7554	0.9346	1.1503	1.4395	1.9600

Example

1. Sorting (draw a histogram)
2. Get percentile
3. P-to-Z
4. Make a scatter plot

   y-axis = Original values

   x-axis = Normal scores

3. P-to-Z

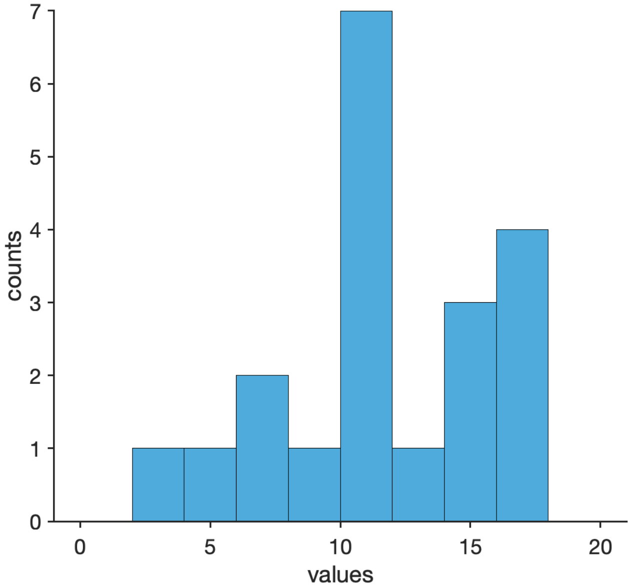
1 ~ 12번 열

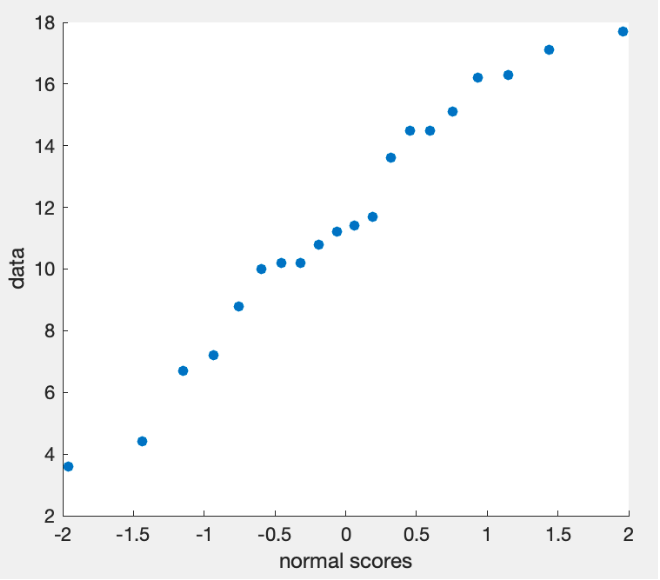
Sorted value	3.6000	4.4000	6.7000	7.2000	8.8000	10.0000	10.2000	10.2000	10.8000	11.2000	11.4000	11.7000
Z-values	-1.9600	-1.4395	-1.1503	-0.9346	-0.7554	-0.5978	-0.4538	-0.3186	-0.1891	-0.0627	0.0627	0.1891

13 ~ 20번 열

13.6000	14.5000	14.5000	15.1000	16.2000	16.3000	17.1000	17.7000
0.3186	0.4538	0.5978	0.7554	0.9346	1.1503	1.4395	1.9600

4. Make a scatter plot (and also histogram)





Key Points

Chapter 5: Comparing distributions: considerations

- Outliers are context dependent
- Re-expressing data (log, sqrt)
  - to improve symmetry
  - to equalize spread
- Timeplots
- Moving-averages, smoothing

Chapter 6: Normal model

- z-score, shifting and rescaling
- Normal model
- Normality assumption; unimodal, symmetric
- 68-95-99.7 Rule
- z-to-p, p-to-z
- Normal probability plots