Statistically Speaking: *Spread and Conformance*

DAVID L. TABB, PH.D.

AUGUST 31, 2017

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Overview

- Variance and standard deviation
- Spread measures for normal distribution
- Distortions from normality
- Spread for *non-parametric* distributions
- QQPlots to evaluate normalcy

UNIVERSITEIT STELLENBOSCH UNIVERSITY

The bar, the cat, and the delta

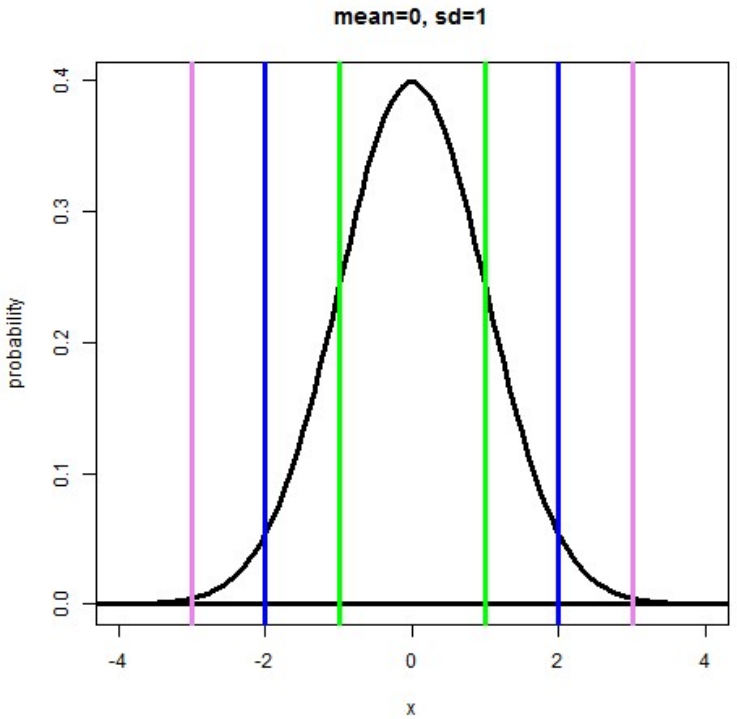
- We have a random variable X, with n observations (samples):
- We compute its average value, the mean: $\mu = \bar{X} = \frac{\sum X}{n}$
- We compute sample variance based on the differences versus that mean: $s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$

Why square differences? Why subtract one from n?

http://www.statisticshowto.com/what-is-sample-variance/

Decoding standard deviation

- $\sigma = \sqrt{variance}$
- Within 1 SD: 68.3%
- Within 2 SD: 95.4%
- Within 3 SD: 99.7%
- Dividing delta by SD produces “z score”




```
x <- seq(-4, 4, length=100)
y <- dnorm(x, mean=0, sd=1)
plot(x,y, type="l", lwd=3, ylab=
"probability",main="mean=0, sd=1")
```

```
abline(h=0,col="black",lwd=3)
abline(v=c(-1,1),col="green",lwd=3)
abline(v=c(-2,2),col="blue",lwd=3)
abline(v=c(-3,3),col="violet",lwd=3)
```

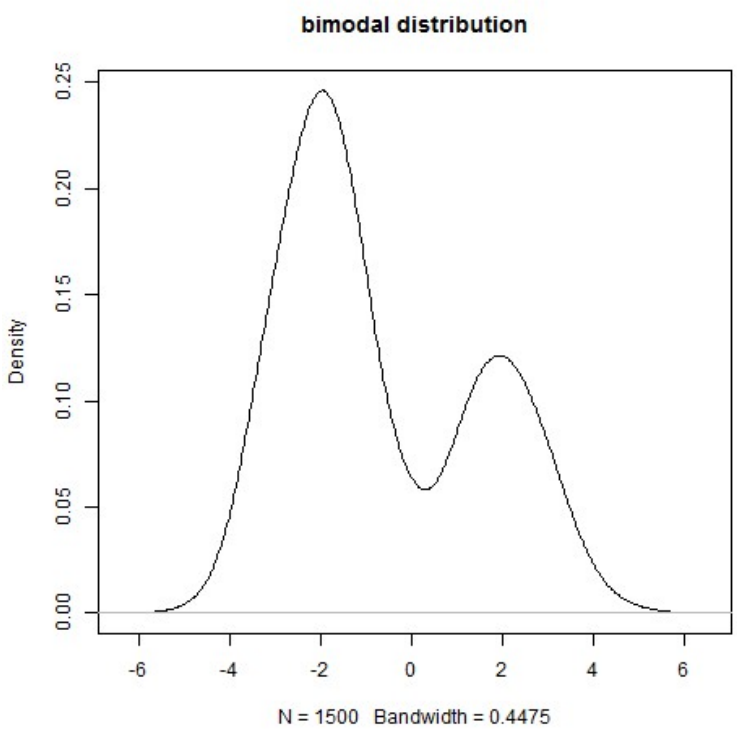
UNIVERSITEIT STELLENBOSCH UNIVERSITY

Centrality terms

- Mean is the average value among all points; even outliers contribute to placement.
- Median represents the middle value if points are sorted by magnitude.
- Mode represents the most common value.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Bimodal distributions

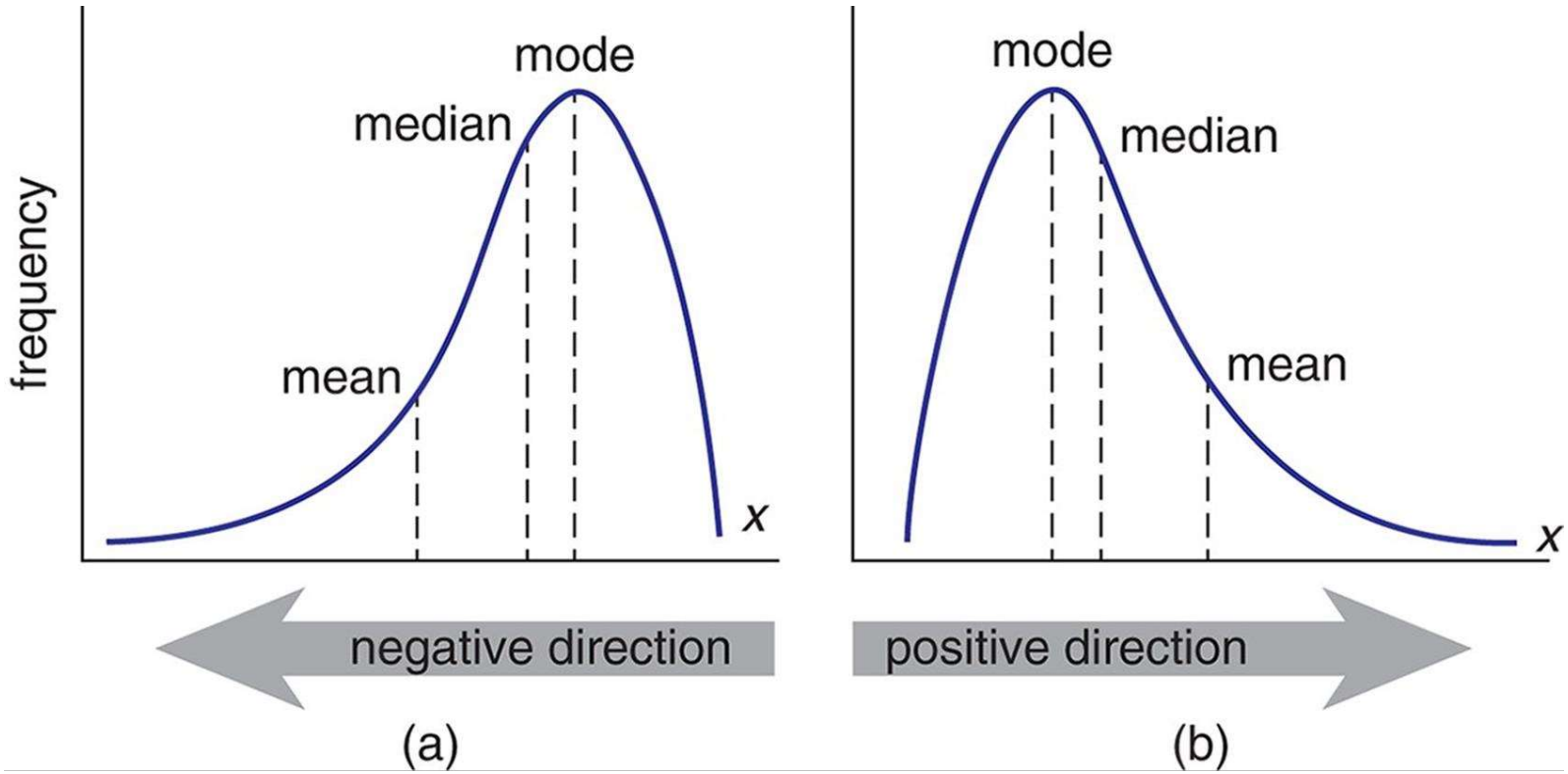


- Heterogeneous data may be drawn from two different distributions!

```
a <- c(
        rnorm(1000,mean=-2),
        rnorm(500,mean=2))
plot(density(a),
        main="bimodal distribution")
```

http://www.statisticshowto.com/what-is-a-bimodal-distribution/

Skewed distributions



http://schoolbag.info/physics/physics_math/71.html

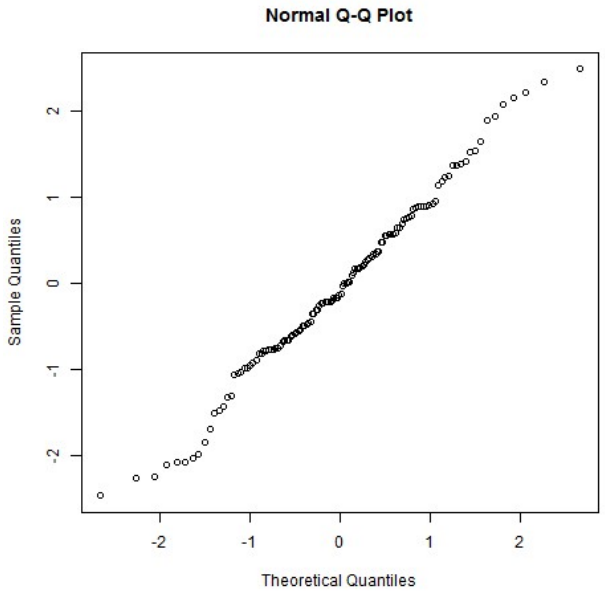
UNIVERSITEIT STELLENBOSCH UNIVERSITY

Non-parametric spread metrics

- Range: Maximum – minimum.
- Quantiles (percentiles):
  - By convention, x%ile is the value x/100 through the sorted list of values
- Quartiles:
  - Min, 25%ile, median, 75%ile, Max
  - Interquartile range= 75%ile – 25%ile

UNIVERSITEIT STELLENBOSCH UNIVERSITY

QQPlots and Shapiro-Wilk



- Data should fall on diagonal of unit slope

- Shapiro-Wilk is a test of a *null hypothesis*.
- H_0 =data come from a normal distribution
- Low p-value implies we reject the hypothesis that data are normal.

http://data.library.virginia.edu/understanding-q-q-plots/

http://www.statisticshowto.com/shapiro-wilk-test/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Closing thoughts

- Characterizing spread is very important; our ability to find differences depends upon our ability to limit spread!
- Just because data have a single hump (are unimodal) doesn't make them normal.
- Being able to describe data in non-parametric ways may allow us to resist some challenges of messy data.