# Variance II

Lecture 7

Variability II: Variance and SD of a sample, degrees of freedom

Reading: *Seeing Statistics 4.0-4.4*

## Degrees of freedom (*df*)

- Degrees of freedom are *independent bits of information*
- Example – John is a jerk.
- The more info (*df*) we have, the more certainty we have about our subjective impressions.
- It's the same in stats!
- Alternative view of *df*: # *independent possible states of the world.*

## Degrees of freedom (*df*) in statistics

- In stats, it is crucial to know how many independent bits of information (*df*) are being used to estimate a parameter.
- In general,
  - *df* = number of observations – number of things estimated from those observations

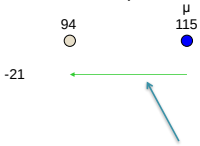
## How many *df* for deviations from the sample mean?

- Imagine I collect *two* data points. If I tell you the mean, and then give you one of the data points, you can figure out the second. E.g.:
  - mean = 12  X1 = 5  X2 = ?
- Given the mean and one data point, the other is not free to vary.
- For deviations from the mean, only *n-1 df* exist for what those deviations could be (one of them is determined by virtue of knowing the mean):

With 2 data points, there are 2 - 1 independent deviations
With 3 data points, there are 3 - 1 independent deviations
With 4 data points, there are 4 - 1 independent deviations

## Variance given the population mean

- Given a single data point *X* and given the population mean μ we could get an estimate of the variance (albeit not a reliable one):




We have a *single* bit of information about the variance. Said another way, we have 1 *df* in this situation.

## Variance when we *don't* know the population mean

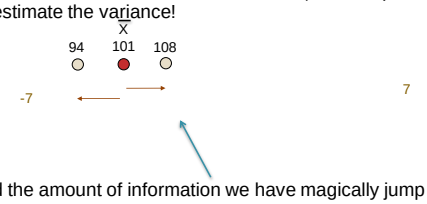
- Given a single data point *X* and when we DON'T know the population mean, we CANNOT get an estimate of the variance


Because we had to use our sample to estimate the mean, we have a *NO* information about the variance (a point cannot deviate from the mean when *n*=1). I.e., we have 0 *df* in this situation.

*df* = *n*-1 = 0

# Variance of a sample

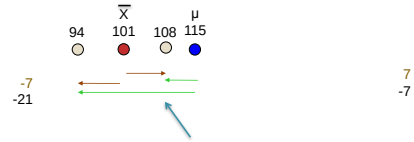
- Given 2 data points, we can estimate the mean and have one bit of information "left over" (read 1 *df*) to estimate the variance!




Did the amount of information we have magically jump from 0 to 2 by adding a *single* data point? Of course not! We added one bit more information, and thus we increase from 0 *df* to 1.

*df = n-1 = 1*

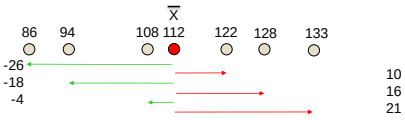
# Variance of a sample




Note: together the deviations are closer to the sample mean than to μ. This must be so: the sum of squared deviations from the sample mean ≤ sum squared deviations from any other #, including μ (*Seeing Stats 3.3.1*).

Thus, if we simply average the two squared deviations, it would always be less than or equal to variance we'd get if we had known the population mean. I.e., it would be a *biased estimate* of the true variance in the population

# Unbiased estimate of population variance derived from a sample:




$$s^2 = \frac{\sum_{sample}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{sample}(X_i - \bar{X})^2}{df}$$

# Biased Estimators: The Example of Variance

- Population variance
  - $(X-\mu)^2$ is unbiased estimator of $\sigma^2$
- Many observations:
  - Problem: We don't know μ

$$E\left(\frac{\sum_{sample}(X-\mu)^2}{n}\right) = \sigma^2$$

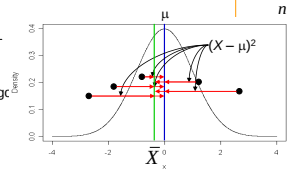
$$\sigma^2 = \frac{\sum_{pop}(X-\mu)^2}{N} = mean\left((X-\mu)^2\right)$$

- Sample mean always shifted toward sample

$$\frac{\sum_{sample}(X-\bar{X})^2}{n} < \frac{\sum_{sample}(X-\mu)^2}{n}$$

$$\frac{\sum_{sample}(X-\bar{X})^2}{n}$$

  - Not a gc




$$E\left(\frac{\sum_{sample}(X-\bar{X})^2}{n}\right) < \sigma^2$$

# Sample Variance

- Goal: Define sample variance to be unbiased estimator of population variance
  - $E(s^2) = \sigma^2$
- Problem: Obvious answer is biased
  - $E\left(\frac{\sum_{sample}(X-\bar{X})^2}{n}\right) < \sigma^2$
  - $\bar{X}$ is always closer to *X* than μ is. On average, this biases $s^2$ to be too small by a factor n-1/n.
- Solution:

$$s^2 = \frac{\sum_{sample}(X-\bar{X})^2}{n-1}$$

  - Unbiased: $E(s^2) = \sigma^2$

# Comparing Mean, Variance, and SD for populations and samples

| | For a population: | For a sample: |
|---|---|---|
| Mean | $\mu = \frac{\sum X}{N}$ | $\bar{X} = \frac{\sum X}{n}$ |
| Variance | $\sigma^2 = \frac{SS}{N}$ | $s^2 = \frac{SS}{n-1}$ |
| Standard Deviation | $\sigma = \sqrt{\frac{SS}{N}}$ | $s = \sqrt{\frac{SS}{n-1}}$ |
| | Usually, I can only guess at (estimate) these | I can calculate these from data |