

BIOS 600: Principles of Statistical Inference

Numerical Summary Measures

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 3
- ▶ Cheesy yet informative short song on mean, median, and mode
- ▶ Larson, 'Descriptive Statistics and Graphical Displays,' 2006
Circulation

Sample Versus Population

The entire group of individuals that we want information about is called the *population*. A *sample* is a part of the population that we actually examine in order to learn about *population* values.

- ▶ Population: all Hispanics/Latinos living in the U.S.
- ▶ Sample: 16,000 Hispanics/Latinos participating in the UNC-led Hispanic Community Health Study
- ▶ Population: Children at risk of malaria in sub-Saharan Africa
- ▶ Sample: 15,000 African children 6 weeks-17 months at 11 study sites in seven countries
- ▶ Population: Women aged 30+ trying to become pregnant
- ▶ Sample: Women 30-44 who have been trying < 3 months to get pregnant and who enroll in a study at UNC

Activity: Sample Versus Population

- ▶ Quickly divide into groups of 10 students
- ▶ Determine the % of students in your group who have eaten ice cream at Maple View Farms
- ▶ Report back to me when asked!

Random Variables

Often we talk about responses in a dataset of interest as *random variables*. A random variable conceptually does not have a single, fixed value (even if unknown); rather, it can take on a set of possible different values, each with an associated probability. For example, we might consider the weights of the participants in the Hispanic Community Health Study to be a random variable of interest.

Definition of Statistic

A *statistic* is a single measure of some attribute of a sample (e.g. its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items comprising the sample which are known together as a set of data. The term statistic is used both for the function and for the value of the function on a given sample.

A statistic is distinct from a statistical *parameter*, which is not computable because often the population is much too large to examine and measure all its items. A statistic, when used to estimate a population parameter, is called an estimator. For instance, the sample mean is a statistic that estimates the population mean, which is a parameter.

Definition of Statistic

A statistic is an observable random variable, which differentiates it from a parameter that is a generally unobservable quantity describing a property of a statistical population. A parameter can only be computed exactly if the entire population can be observed without error; for instance, in a perfect census.

Types of Summary Statistics

- ▶ Location measures: WHERE
- ▶ For example, is average health care expenditure per person in the U.S. \$70, \$700, \$7000, \$70000?
 - ▶ Mean
 - ▶ Median
 - ▶ Mode
- ▶ Spread or scale measures
 - ▶ Range
 - ▶ Interquartile range (IQR)
 - ▶ Variance and standard deviation

Point Estimates of Location

What is the *point* of a point estimate?

- ▶ A *point estimate* is a one-number best guess
 - ▶ The more you know about the *distribution* of the data, the better guess you can make.
 - ▶ What is your best guess of the percent of BIOS 600 students who drink soda?
 - ▶ We will discuss three different choices of point estimates (guesses) of location

Mean

We often estimate a *population mean*, μ , using the *sample mean* or *average*. The *sample mean*, \bar{x} , is calculated by adding all the observations (x_1, x_2, \dots, x_n) in a set of data and then dividing by the total number of measurements n . We depict this formula as
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Notation key:

- ▶ The upper case Greek letter Σ is the symbol for summation.
- ▶ The expression $\sum_{i=1}^n x_i$ means that we start at x_1 and stop at x_n so that we're taking $x_1 + x_2 + x_3 + \dots + x_n$.

Example: Child Weights

Consider the following two random samples of birth weights of children (in kg). Recall our formula, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

- ▶ Group 1: (3.3 2.6 3.1)

- ▶ Sample Mean=?

- ▶ Sample Mean = $\frac{1}{3} \sum_{i=1}^3 x_i = \frac{3.3+2.6+3.1}{3} = \frac{9.0}{3} = 3.0$

- ▶ Group 2: (2.7 3.0 1.5 3.2)

- ▶ Sample Mean=?

- ▶ Sample Mean = $\frac{1}{4} \sum_{i=1}^4 x_i = \frac{2.7+3.0+1.5+3.2}{4} = \frac{10.4}{4} = 2.6$

Median

The *sample median* separates the top half of the sample from the bottom half (i.e., it is the middle number if the observations are ranked in numerical order). If the number of observations n in the sample is odd, then the sample median is the middle value, which sits in position $\frac{n+1}{2}$. If the number of observations n is even, then the sample median is the average of the middle two values, which are sitting in positions $\frac{n}{2}$ and $\frac{n+1}{2}$.

Median

Consider two simple examples with group sizes of $n = 7$ and $n = 8$; suppose the observations are ordered and each dot is a single observation.

$n=7$



The Median M is the center observation, which is located in the $(7+1)/2 = 4$ th spot in the ordered list

$n=8$



The Median M is the mean of the two center observations, which in this case are located at the $8/2=4$ th and $8/2 + 1 = 5$ th spots in the ordered list

Median

- ▶ Group 1: (3.3 2.6 3.1)
- ▶ Sample Median=?
- ▶ Sample Median of (2.6 3.1 3.3) is 3.1
- ▶ Group 2: (2.7 3.0 1.5 3.2)
- ▶ Sample Median=?
- ▶ Sample Median of (1.5 2.7 3.0 3.2) is $\frac{2.7+3.0}{2} = 2.85$

Robustness of the Median

Consider the birth weights (2.6 3.1 3.3), with a mean of 3.0 kg and median 3.1 kg.

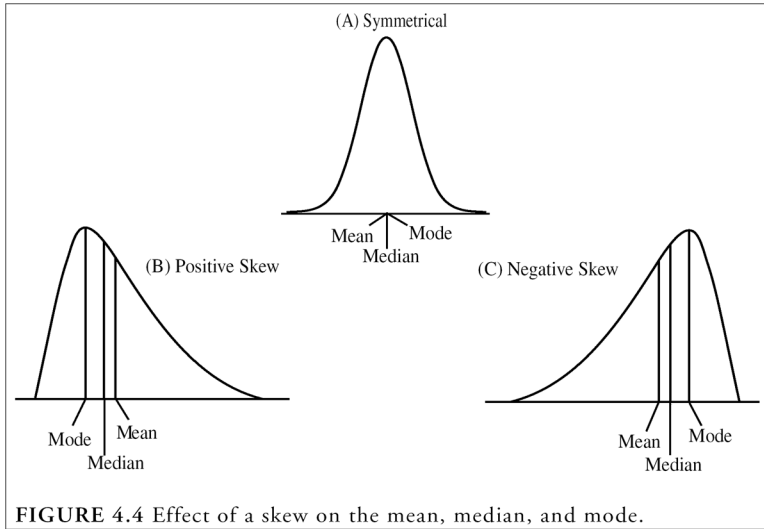
- ▶ Suppose instead of the 3.3 kg baby we had observed actress/singer Jessica Simpson's new baby, who weighed 4.5kg, so that the weights in our sample are now (2.6 3.1 4.5). In this case the median birth weight would still be 3.1 kg, but the mean would now be 3.4 kg, as it is affected by the outlying value.
- ▶ If we were even more unlucky and sampled the world's largest surviving baby (10.2kg, born in Italy) instead of Jessica's baby, the median birth weight would still be 3.1 kg, but the mean would now be a whopping 5.3 kg!!!
- ▶ Thus the median is less sensitive to *outliers* or extreme values than the mean.

Mode

The *sample mode* is the most common or frequent value in the dataset. A data set may or may not have a unique mode and may have no mode.

- ▶ Our data set (2.6 3.1 3.3) has no mode (no value occurs more frequently than the others)
- ▶ The data set (4 7 7 9) has a mode of 7
- ▶ The data set (4 7 7 7 9 11 11 11) has two modes, 7 and 11.
- ▶ Suppose you all tell me how many cigarettes you smoke per day. The class mode would be zero (we are in public health!).

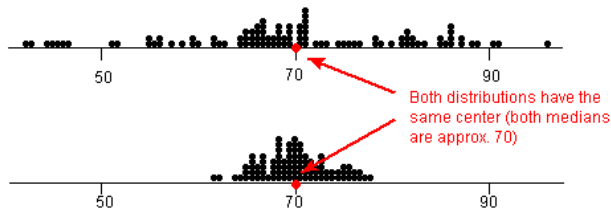
Mean, Median, and Mode



Mean, Median, and Mode Interactive

Find the mean, median, and mode of **age of your mother**.

Why do we care about spread if we know center?



There is a lot more variability in the first dataset than in the second.

Minimum, Maximum, and Range

The *sample minimum* and *sample maximum* are the smallest and largest observations in our sample.

The *sample range* is the value of the maximum minus the minimum.

Consider the following birth weights (g) from the UNC Pregnancy, Infection, & Nutrition (PIN) study:

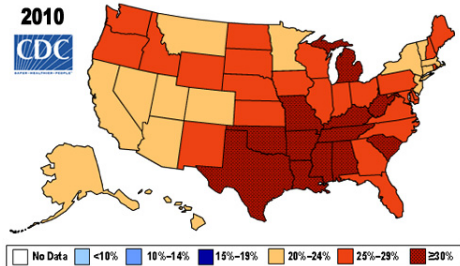
(2138 2639 3966 3415 3410 3137 3460 1519 4739 3147 1824)

- ▶ Minimum:
- ▶ Minimum: 1519
- ▶ Maximum:
- ▶ Maximum: 4739
- ▶ Range:
- ▶ Range: $4739 - 1519 = 2220$

Quartiles

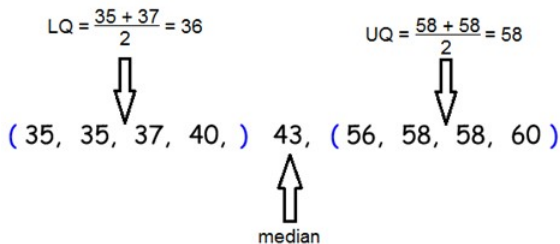
We talk about quartiles all the time without even realizing it.

- ▶ My home state is in the top 25% in the U.S. obesity rankings
- ▶ My home state is in the bottom quarter of states in high school graduation rates
- ▶ The median home price in Chapel Hill is \$263,000

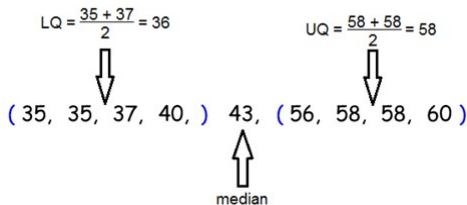


Quartiles

The *quartiles* divide the data into four equal pieces by rank of the observations. Consider the following ages of patients in the Lineberger Cancer Center.

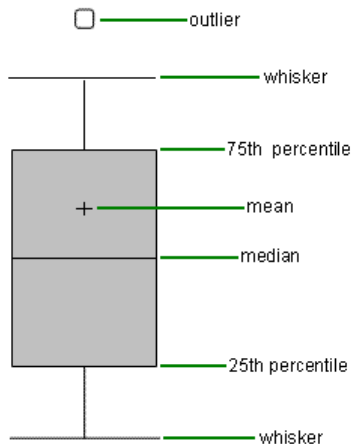


Quartiles



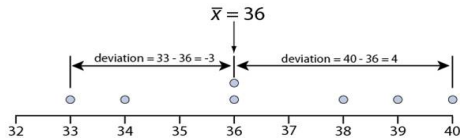
- ▶ The *median* or *second quartile* (Q_2) divides the data in half.
- ▶ The *first quartile* (Q_1) or *lower quartile* cuts off the bottom 25% of the observations.
- ▶ The *third quartile* (Q_3) or *upper quartile* cuts off the top 25% of the observations.
- ▶ The *interquartile range* is the value of the third quartile minus the value of the first quartile ($Q_3 - Q_1$). It covers the middle 50% of the data.

Back to Boxplots

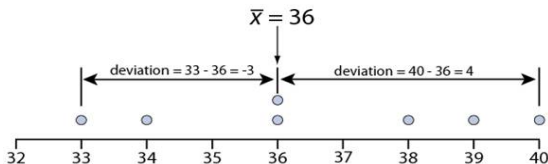


Standard Deviation

- ▶ Most common measure of spread
- ▶ Based on *deviations* around mean
- ▶ This data set has a mean of 36
- ▶ The data point 33 has a deviation of $33 - 36 = -3$
- ▶ The data point 40 has a deviation of $40 - 36 = 4$

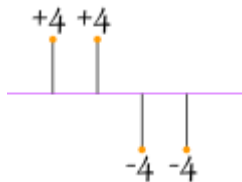


Standard Deviation



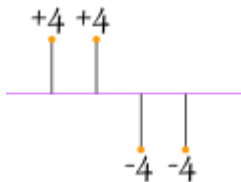
- ▶ Sample standard deviation is $\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ (always ≥ 0)
- ▶ Where did that come from?
- ▶ We square the deviations $x_i - \bar{x}$ so that deviations on either side of the mean do not cancel each other out when we add them

Why Square Differences to Measure Spread?

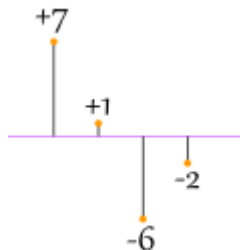


If we add up deviations from the mean, the negatives would cancel the positives: $4 + 4 - 4 - 4 = 0$ (no good as a measure of spread!)

Why Square Differences to Measure Spread?

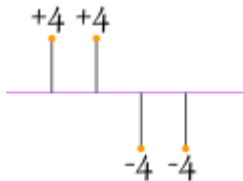


What about using absolute values? $4+4+4+4=16$



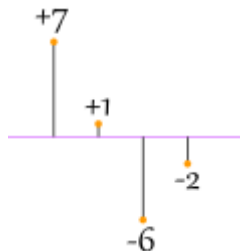
What about this case?
 $7+1+6+2=16$ (hmm, those differences are more variable but the sum of absolute values is the same)

Why Square Differences to Measure Spread?



What about squaring?

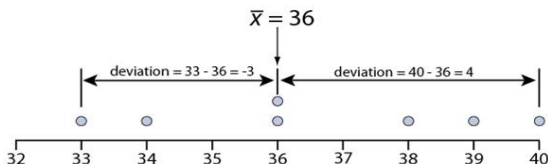
$$16+16+16+16=64$$



Summing squares:

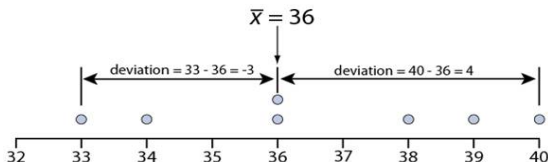
$49+1+36+4=90$, nice more spread out observations get bigger penalty

Standard Deviation



- ▶ *Sample standard deviation* is $\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$
- ▶ We divide by $n - 1$ to adjust for the number of samples we have (otherwise, the value in the sum would grow just by having a bigger sample); can also divide by n though that estimator is *biased* in small samples
- ▶ The quantity $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ is called the *sample variance* and is in *units*² (years² here). By taking the square root we transform the units to match the units in the original data.

Standard Deviation



- ▶ The standard deviation is often used to express confidence in statistical conclusions.
 - ▶ *Margin of error* of a poll typically around two standard deviations (reported as $\pm 2sd$)
 - ▶ Width of 95% confidence interval around four standard deviations (e.g., 2sd on each side of the point estimate)

Chebyshev's Inequality

Chebyshev's Inequality says that regardless of the nature of the underlying distribution that defines the random variable, there are guaranteed bounds on the % of observations that will lie within k standard deviations (σ) of the mean (μ).

$$\Pr(|X - \mu| > k\sigma) < \frac{1}{k^2}$$

implies that at most $\frac{1}{k^2}$ of the observations will be more than k standard deviations away from the mean.

Chebyshev's Inequality

$$\Pr(|X - \mu| > k\sigma) < \frac{1}{k^2}$$

Suppose the distribution of BMI in the US has mean $\mu = 28$ and standard deviation $\sigma = 5$. Using Chebyshev's inequality with $k = 2$, we see that at most $\frac{1}{4}$ of the observations will be more than 2 standard deviations away from the mean (greater than $28 + 2 \times 5 = 38$ or less than $28 - 2 \times 5 = 18$). This implies at least 75% of the observations will be in the range of 18-38.

Chebyshev's Inequality

$$\Pr(|X - \mu| > k\sigma) < \frac{1}{k^2}$$

Using $k = 3$, we learn that at most $\frac{1}{9}$ of the observations will be more than 3 standard deviations from the mean (greater than $28 + 15 = 43$ or less than $28 - 15 = 13$) and that at least 89% ($\frac{8}{9} * 100\%$) of the observations will be in the range 13-43.

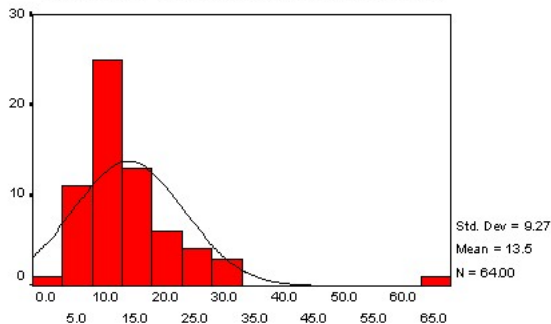
Outliers

An *outlier* is an observation that is numerically distant from others. How we handle outliers depends on how they arise.

- ▶ If an outlier is consistent with an underlying biological process and could reasonably be expected to happen again, then we should leave it in the data. Statistical methods that are *robust* to outliers can be used if outliers are problematic (will discuss later in course). Example: maternal age of 11
- ▶ If an outlier is clearly a mistake in the data, then it should either be corrected or removed. Example: maternal age of 111

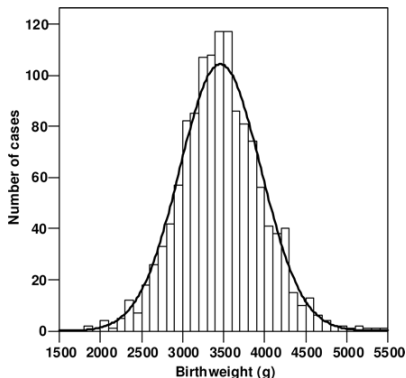
Outliers

Histogram of the Prevalence of
Low Arm Circumference in Females



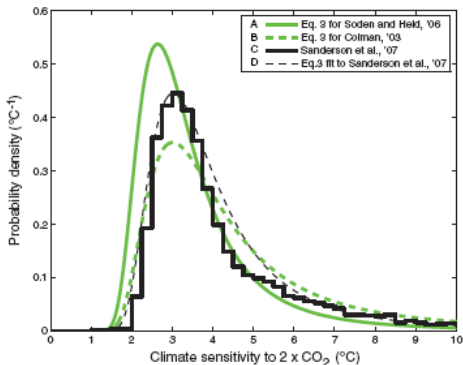
When to Report What?

What measures of central tendency and spread may be best for summarizing the distribution of birth weights of 1295 term deliveries in The Charité obstetrics department in Berlin?



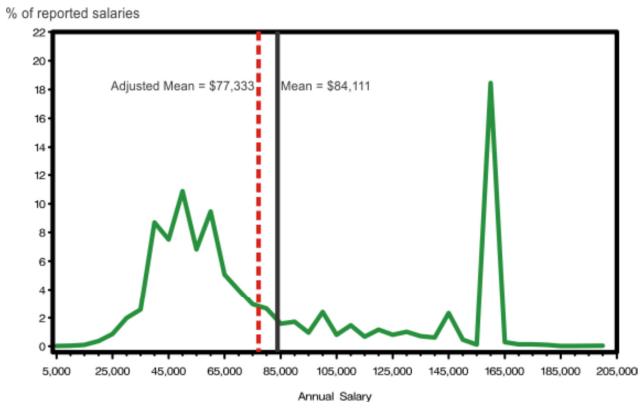
When to Report What?

What measures of central tendency and spread may be best for summarizing the distribution of climate sensitivity (expected increase in temperature in degrees C) when levels of atmospheric CO₂ levels reach double their pre-industrial levels (≈ 2050)?



When to Report What?

What measures of central tendency and spread may be best for summarizing the distribution of salaries of 2010 law school graduates?



Learn by Doing

- ▶ Interactive BBC Game on Mean, Median, and Mode

Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 6, Section 6.1
- ▶ Interactive BBC Game on Probability