Stat 201: Statistics I
Week 5

Metropolitan State University

Week 5
Relative Standing, Random Variables and Distributions

Section 5.1
Measures of Relative Standing and Boxplots

Measures of relative standing

Measures of relative standing describe the location of a given data value within a data distribution or a data set.

Two measures of relative standing are discussed in this section:

- Z-scores
- Percentiles

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.
- Z-scores are standardized and unit-less, so they can be used to compare values from different populations.
- A positive z-score means the value is greater than the mean and a negative z-score means that it is below the mean.
- Z-scores can be calculated for samples or populations, if the population mean and standard deviation are known.

Z-scores, calculations

To calculate

- For a sample X with sample mean $\bar{x}$ and standard deviation s, the z-score for a value x is

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

- For a population with population mean μ and standard deviation σ, the z-score for value x is

$$z = \frac{x - \mu}{\sigma}$$

Z-scores, example

Example

Recall the example of a sample of student ages. The sample has ages $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$, with a mean of $\bar{x} = 33.125$ and standard deviation of $s = 11.05$.

- Suppose a new student joins the class. His age is 61. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{61 - 33.125}{11.05} = \frac{27.875}{11.05} = 2.52$$

  His age two and a half standard deviations above the class mean.

- Another student joins the class. Her age is 27. She has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{27 - 33.125}{11.05} = \frac{-6.125}{11.05} = -0.554$$

  Her age is about a half standard deviation below the class mean.

Values from z-scores

Sometimes, it is useful to find a value within a data distribution that corresponds to a given z-score. That is, find an x given a z.

- Start with the z-score equation,

$$z = \frac{x - \bar{x}}{s} \qquad \text{or} \qquad z = \frac{x - \mu}{\sigma}$$

- After some algebra,

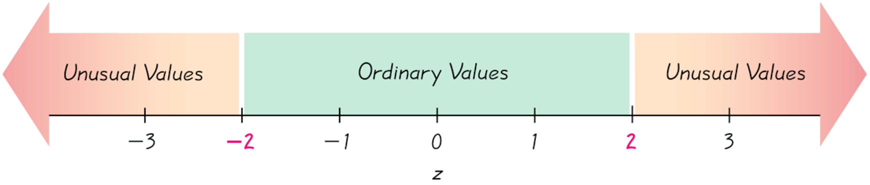
$$x = \bar{x} + zs \qquad \text{or} \qquad x = \mu + z\sigma$$

Example

What age is 2 standard deviations above the mean for this data? That is, what age corresponds to $z = 2$?

$$x = \bar{x} + zs = 33.125 + 2 \times 11.05 = 55.225 \text{ years}$$

Unusual/significant values



A value is called **unusual or significant** if it has a z-score z such that $z < -2$ or $z > 2$. A value is **ordinary** if z is between -2 and 2.

Example

Consider the new students to the class:

- The 61 year old ($z = 2.52$) has an unusual age for the class.
- The 27 year old ($z = -0.554$) has an ordinary age for the class.

Percentiles

Percentiles measure relative position within a data set as order rank expressed as a percent. In other words, the value at the pth percentile (written as P_p) in a data set is greater than $p\%$ of the data.

Example

Percentiles are often used in reporting scores on standardized tests.

Suppose a student scores in the 83rd percentile on the ACT. That means she scored better than 83% of the students who took the ACT.

Calculate percentiles

To calculate

- To find the percentile of a value x in a data set,

$$\%\text{ile} = \frac{\text{number of values} < x}{n} \times 100\%$$

  If percentile is not a whole number, round up.

- To find the value of P_p (the pth percentile), calculate the rank,

$$r = \frac{p}{100} \times n$$

  If r is a whole number, P_p is the mean of the rth and $(r+1)$th values. If not, round up. Then, P_p is the rth value in an ordered list.

Percentile, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- The percentile of the value 38 is

$$\frac{\text{number of values} < x}{n} \times 100\% = \frac{5}{8} \times 100\% = 62.5\,\% \Rightarrow P_{63}$$

- To find the 30th percentile, P_{30}, calculate rank

$$r = \frac{p}{100} \times n = \frac{30}{100} \times 8 = 2.4$$

  Round up r to 3. P_{30} is the 3rd value, 24.

Quartiles

The **quartiles** are values that divide the data set into 4 parts, or quarters.

$$Q_1 = P_{25} \qquad Q_2 = P_{50} \qquad Q_3 = P_{75}$$

- Note: The median is equivalent to Q_2 and P_{50}.

5 number summary

The **5 number summary** summarizes the distribution of a data set.

The 5 numbers are:

- Minimum
- Q_1
- Median (or Q_2)
- Q_3
- Maximum

5 number summary, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

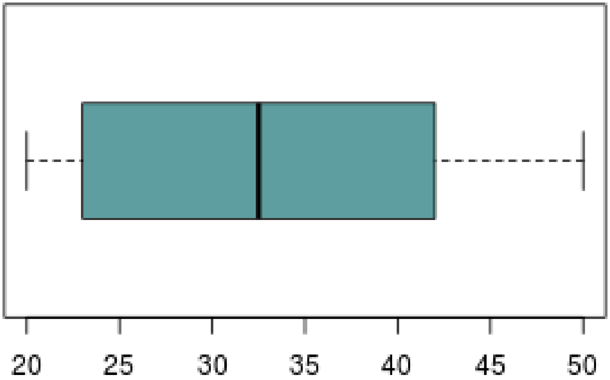
The 5 number summary is

$$\underbrace{20}_{\text{min}} \quad \underbrace{23}_{Q_1} \quad \underbrace{32.5}_{\text{med}} \quad \underbrace{42}_{Q_3} \quad \underbrace{50}_{\text{max}}$$

Boxplots

A **boxplot** is a graph depicting the 5 number summary.

- $\{20, 23, 32.5, 42, 50\}$



Section 5.2
Probability Distributions

Random variables

A **random variable** is a variable that has a numeric value determined by chance from a range of possible values.

- An outcome of a trial
- Usually designated with a capital letter (X, Y, etc.)
- Lowercase letters refer to specific values of the random variable
- Thus, $P(X = x)$ means the probability that the random variable X takes the specific value x.

Random variables, examples

Example

- $X =$ the number of heads from three coin flips
- $Y =$ the sum of two dice
- $Z =$ the midterm score of a randomly selected student
  - A students grade (A, A-, B+, etc.) can not be used as a random variable because it is not numeric,
  - ... Unless the grade is coded as a number (i.e. A = 4.0, A- = 3.7, etc.)

Types of random variables

Recall, numeric variables can be classified as **discrete** or **continuous**. Random variables also can be either discrete or continuous.

Example

Discrete random variables:

- Number of heads on three coin flips
- Number of defective insulin test strips in a box of 50
- Number of customers to enter a store in the next 10 minutes

Continuous random variables:

- Height or weight of a test subject
- Survival time of a cancer patient
- Price of a company's stock at a particular moment

Probability distributions

The collection of probabilities of all the possible values of a random variable is known as the **probability distribution** of the random variable.

- Each probability is between 0 and 1
- The probabilities must add up to 1
- Often displayed in tables (if practical)

Probability distributions, example

Example

A restaurant wants to track its taco sales. It records how many tacos customers order with each visit. The results are in the table,

Number of tacos	0	1	2	3	4
Probability	0.35	0.2	0.3	0.1	0.05

Is this a probability distribution?

- These are probabilities of every possible outcome of a trial (customer making an order).
- The probabilities add to 1.
- It is a probability distribution.

Probability distributions, example

Example

The restaurant introduces a new "Super Taco" (beef, chicken and shrimp). It wonders if larger groups are more likely to order the new item. The results are in the table,

Number in group	1	2	3	4 or more
Probability of ordering	0.05	0.03	0.04	0.15

Is this a probability distribution?

- The probabilities are for a different event (ordering a Super Taco) than the values (number in group).
- The probabilities do not add to 1.
- It is not a probability distribution.

Event probabilities

To calculate the probability of an event given a probability distribution, simply add the probabilities of the outcomes which comprise the event.

Example

Number of tacos	0	1	2	3	4
Probability	0.35	0.2	0.3	0.1	0.05

What is the probability of a customer ordering less than two tacos?

$$P(X < 2) = P(X = 0 \text{ or } X = 1) = P(0) + P(1) = 0.35 + 0.2 = 0.55$$

Weighted means

A **weighted mean** is the mean of values that are not considered equally, or do not have equal importance.

- Each value has an associated weight, which is its relative importance.
- To calculate, let x_i be a value and w_i its weight.

$$\mu_w = \frac{\sum w_i \times x_i}{\sum w_i}$$

Example

Fiona buys 4 lbs. of hamburger at \$4.89 / lb. and 2 lbs. of steak at \$11.99 / lb. What is the average price per pound she is paying?

$$\$/lb. = \frac{4 \times 4.89 + 2 \times 11.99}{6} = \frac{43.54}{6} = 7.26$$

Mean of probability distributions

The mean of a probability distribution is a weighted mean of the possible values, with the probability of each as its weight.

- Since the sum of probabilities of a distribution is always 1, the divisor of the weighted mean is 1 which we can ignore.
- Thus, the mean is

$$\mu = \sum x_i \cdot P(x_i)$$

The mean of a probability distribution is also known as the **expected value** of the random variable (or sometimes as just the mean of the random variable).

- Denoted with an "E", as in

$$\mathbb{E}(X) = \mu$$

Mean, example

Example

Number of tacos	0	1	2	3	4
Probability	0.35	0.2	0.3	0.1	0.05

What is the mean number of tacos ordered at the restaurant? That is, how many tacos should the restaurant expect each customer to order?

$$\begin{aligned}
\mathbb{E}(X) = \mu = & \sum x_i \cdot P(x_i) \\
= & (0 \cdot 0.35) + (1 \cdot 0.2) + (2 \cdot 0.3) + (3 \cdot 0.1) + (4 \cdot 0.05) \\
= & 0 + 0.2 + 0.6 + 0.3 + 0.2 \\
= & 1.3
\end{aligned}$$

Standard deviation of probability distributions

Similarly, variance of a probability distribution is the weighted mean of difference from the mean squared and standard deviation is the square root of variance.

Thus,

$$\sigma^2 = \sum (x_i - \bar{x})^2 \cdot P(x_i)$$

$$\sigma = \sqrt{\sigma^2}$$

Standard deviation, example

Example

What is the standard deviation of number of tacos ordered at the restaurant?

$$\begin{aligned}
\sigma^2 &= \sum (x_i - \mu)^2 \cdot P(x_i) \\
&= (0 - 1.3)^2 \cdot 0.35 + \cdots + (4 - 1.3)^2 \cdot 0.05 \\
&= 0.5915 + \cdots + 0.3645 \\
&= 1.41 \\
\sigma = \sqrt{\sigma^2} &= 1.19
\end{aligned}$$

Unusual/significant events

If the probability of a random variable being equal to an event or takes a value more extreme than than the event is less than some threshold, usually 0.05, then the event is an **unusual or significant event**.

That is, if x is an extreme value if

$$P(X \leq x) < 0.05 \qquad \text{or} \qquad P(X \geq x) < 0.05$$

Unusual/significant events, example

Example

Tom and Barbara collect eggs from their chickens everyday. The number of eggs they collect follows the following distribution,

Eggs (x)	0	1	2	3	4	5	6	7	8	9
$P(x)$	0.01	0.03	0.1	0.2	0.3	0.2	0.1	0.04	0.02	0

Is collecting 1 egg significantly low? Is collecting 7 eggs significantly high?

- $P(X \leq 1) = P(0) + P(1) = 0.04 < 0.05$
  Collecting 1 egg is significantly low.
- $P(X \geq 7) = P(7) + P(8) + P(9) = 0.06 \not< 0.025$
  Collecting 7 eggs is not significantly high.

Rare event rule

The **rare event rule** says that if observed results are unusual, given an assumed probability distribution, then perhaps the assumption is wrong.

Example

Recall the example of flipping a coin 1000 times. Under the assumption the coin is fair ($P(H) = P(T) = 1/2$), the expected number of heads is 500.

- Getting 523 heads is not unusual ($P(X \geq 523) = 0.077$). There is no reason to think the coin is not fair.
- Getting 46 heads is unusual ($P(X \leq 46) = 6.23 \times 10^{-222}$). We would be justified in questioning the assumption that the coin is fair.