Unit 2: Probability and distributions

3. Normal distribution

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

February 2, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

- Peer evaluation 1 by Friday 11:59pm
- Office hours:
  - currently MTWR 3-4pm
  - propose changing to TR 3-5pm, is this better?

Clicker question

(a) No, keep OH at MTWR 3-4pm
(b) Change to TR 3-5pm

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

1. Two types of probability distributions: discrete and continuous

- A *discrete probability distribution* lists all possible events and the probabilities with which they occur
  - The events listed must be disjoint
  - Each probability must be between 0 and 1
  - The probabilities must total 1

1. Two types of probability distributions: discrete and continuous

- A *discrete probability distribution* lists all possible events and the probabilities with which they occur
  - The events listed must be disjoint
  - Each probability must be between 0 and 1
  - The probabilities must total 1
- A *continuous probability distribution* differs from a discrete probability distribution in several ways:
  - The probability that a continuous random variable will equal to any specific value is zero.
  - As such, they cannot be expressed in tabular form.
  - Instead, we use an equation or a formula to describe its distribution via a probability density function (pdf).
  - We can calculate the probability for ranges of values the random variable takes (area under the curve).

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Outcome	X	P(X)
Win \$10 (black aces)	10	$\frac{2}{52}$
Win \$8 (red aces: 10 - 2)	8	$\frac{2}{52}$
Lose \$2 (non-ace reds)	-2	$\frac{24}{52}$
No win / loss	0	$\frac{24}{52}$
		$\frac{52}{52} = 1$

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Outcome	X	P(X)
Win \$10 (black aces)	10	$\frac{2}{52}$
Win \$8 (red aces: 10 - 2)	8	$\frac{2}{52}$
Lose \$2 (non-ace reds)	-2	$\frac{24}{52}$
No win / loss	0	$\frac{24}{52}$
		$\frac{52}{52} = 1$

Continuous:
Distribution of weekly expenditures of entertainment for a family is right skewed with median of \$70.

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

$$\mathrm{N}(\mu, \sigma)$$

2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

$$\mathrm{N}(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean

2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

$$\mathrm{N}(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean
- *68-95-99.7 Rule:*
  - about 68% of the distribution falls within 1 SD of the mean
  - about 95% falls within 2 SD of the mean
  - about 99.7% falls within 3 SD of the mean
  - it is possible for observations to fall 4, 5, or more standard deviations away from the mean, but this is very rare if the data are nearly normal

2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

$$\mathrm{N}(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean
- *68-95-99.7 Rule:*
  - about 68% of the distribution falls within 1 SD of the mean
  - about 95% falls within 2 SD of the mean
  - about 99.7% falls within 3 SD of the mean
  - it is possible for observations to fall 4, 5, or more standard deviations away from the mean, but this is very rare if the data are nearly normal
- While most variables are nearly normal, but none are exactly normal

Clicker question

Speeds of cars on a highway are normally distributed with mean 65 miles / hour. The minimum speed recorded is 48 miles / hour and the maximum speed recorded is 83 miles / hour. Which of the following is most likely to be the standard deviation of the distribution?

(a) -5

(b) 5

(c) 10

(d) 15

(e) 30

Clicker question

Speeds of cars on a highway are normally distributed with mean 65 miles / hour. The minimum speed recorded is 48 miles / hour and the maximum speed recorded is 83 miles / hour. Which of the following is most likely to be the standard deviation of the distribution?

(a) -5 → *SD cannot be negative*

(b) *5* $\rightarrow 65 \pm (3 \times 5) = (50, 80)$

(c) 10 $\rightarrow 65 \pm (3 \times 10) = (35, 95)$

(d) 15 $\rightarrow 65 \pm (3 \times 15) = (20, 110)$

(e) 30 $\rightarrow 65 \pm (3 \times 30) = (-25, 155)$

1. Housekeeping

2. Main ideas
    1. Two types of probability distributions: discrete and continuous
    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

3. Z scores serve as a ruler for any distribution

$$Z = \frac{\text{obs} - \text{mean}}{\text{SD}}$$

- Z score: number of standard deviations it falls above or below the mean

3. Z scores serve as a ruler for any distribution

$$Z = \frac{\text{obs} - \text{mean}}{\text{SD}}$$

- Z score: number of standard deviations it falls above or below the mean
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles

$$Z = \frac{\text{obs} - \text{mean}}{\text{SD}}$$

- Z score: number of standard deviations it falls above or below the mean
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles
- Observations with $|Z| > 2$ are usually considered *unusual*

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

- Linear transformations of normally distributed random variable will also be normally distributed.
- Hence, if

$$\mathrm{Z} = \frac{\mathrm{X} - \mu}{\sigma}, \text{ where } \mathrm{X} \sim \mathrm{N}(\mu, \sigma),$$

then

$$\mathrm{Z} \sim \mathrm{N}(0, 1)$$

- Z distribution is a special case of the normal distribution where $\mu = 0$ and $\sigma = 1$ (unit normal distribution)

Clicker question

Scores on a standardized test are normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z scores, which of the following statements will be correct?

(a) The mean will equal 0, but the median cannot be determined.
(b) The mean of the standardized Z-scores will equal 100.
(c) The mean of the standardized Z-scores will equal 5.
(d) Both the mean and median score will equal 0.
(e) A score of 70 is considered unusually low on this test.

Clicker question

Scores on a standardized test are normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z scores, which of the following statements will be correct?

(a) The mean will equal 0, but the median cannot be determined.

(b) The mean of the standardized Z-scores will equal 100.

(c) The mean of the standardized Z-scores will equal 5.

(d) *Both the mean and median score will equal 0.*

(e) A score of 70 is considered unusually low on this test.

Application exercise: 2.3 Normal distribution

See the course website for instructions.

Clicker question

Which of the following is false?

(a) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

(b) Majority of Z scores in a right skewed distribution are negative.

(c) In a normal distribution, Q1 and Q3 are more than one SD away from the mean.

(d) Regardless of the shape of the distribution (symmetric vs. skewed) the Z score of the mean is always 0.

Clicker question

Which of the following is false?

(a) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

(b) Majority of Z scores in a right skewed distribution are negative.

(c) *In a normal distribution, Q1 and Q3 are more than one SD away from the mean.*

(d) Regardless of the shape of the distribution (symmetric vs. skewed) the Z score of the mean is always 0.

1. Housekeeping

2. Main ideas
    1. Two types of probability distributions: discrete and continuous
    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

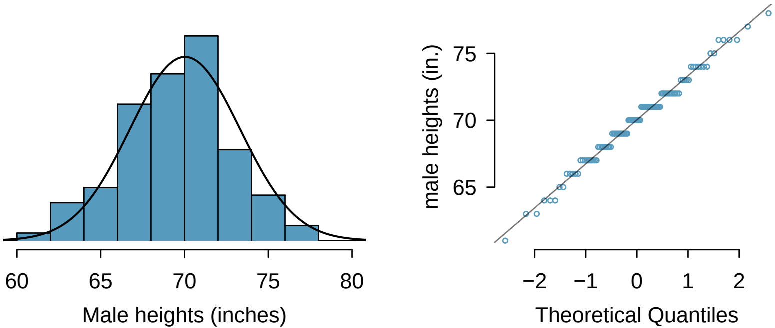
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model
- Constructing a normal probability plot requires calculating percentiles and corresponding Z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots

A histogram and *normal probability plot* of a sample of 100 male heights.



Why do the points on the normal probability have jumps?

We construct a normal probability plot for the heights of a sample of 100 men as follows:

1. Order the observations.
2. Determine the percentile of each observation in the ordered data set.
3. Identify the Z score corresponding to each percentile.
4. Create a scatterplot of the observations (vertical) against the Z scores (horizontal)

We construct a normal probability plot for the heights of a sample of 100 men as follows:

1. Order the observations.
2. Determine the percentile of each observation in the ordered data set.
3. Identify the Z score corresponding to each percentile.
4. Create a scatterplot of the observations (vertical) against the Z scores (horizontal)

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile , $i/(n+1)$	0.99%	1.98%	2.97%	$\cdots$	99.01%
z_i	-2.33	-2.06	-1.89	$\cdots$	2.33

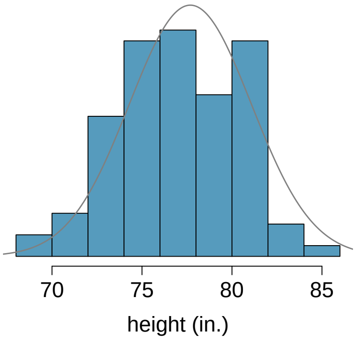
We construct a normal probability plot for the heights of a sample of 100 men as follows:

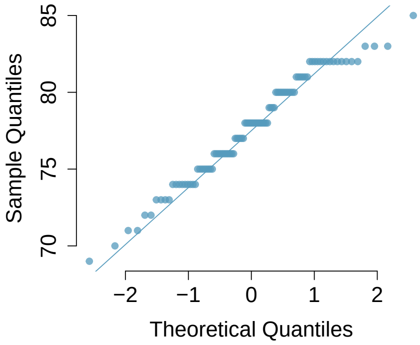
1. Order the observations.
2. Determine the percentile of each observation in the ordered data set.
3. Identify the Z score corresponding to each percentile.
4. Create a scatterplot of the observations (vertical) against the Z scores (horizontal)

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile , $i/(n+1)$	0.99%	1.98%	2.97%	$\cdots$	99.01%
z_i	-2.33	-2.06	-1.89	$\cdots$	2.33

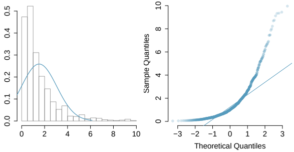
How are the Z scores corresponding to each percentile determined?

Below is a histogram and normal probability plot for the heights of Duke men's basketball players (from 1990s and 2000s). Do these data appear to follow a normal distribution?

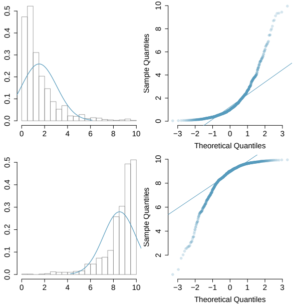




Source: GoDuke.com



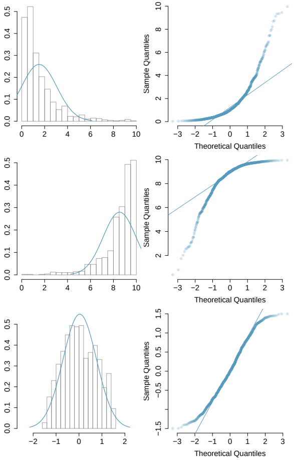
Right Skew - Points bend up and to the left



Right Skew - Points bend up and to the left

Left Skew - Points bend down and to the right

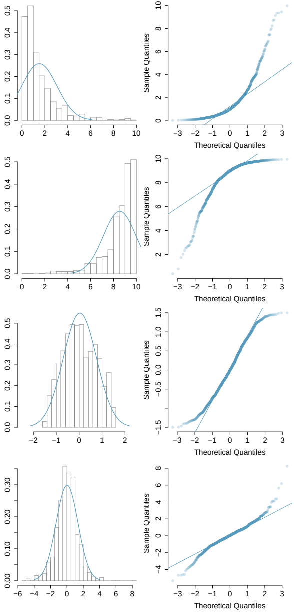
Normal probability plot and skewness



Right Skew - Points bend up and to the left

Left Skew - Points bend down and to the right

Skinny Tails - S shaped-curve indicating shorter than normal tails (narrower, less variable, than expected)



Right Skew - Points bend up and to the left

Left Skew - Points bend down and to the right

Skinny Tails - S shaped-curve indicating shorter than normal tails (narrower, less variable, than expected)

Fat Tails - Curve starting below the normal line, bends to follow it, and ends above it (wider, more variable, than expected)

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

1. Two types of probability distributions: discrete and continuous
2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule
3. Z scores serve as a ruler for any distribution
4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
5. Normally distributed data plot as a straight line on the normal probability plot

1. Housekeeping

2. Main ideas

    1. Two types of probability distributions: discrete and continuous

    2. Normal distribution is unimodal, symmetric, and follows the 69-95-99.7 rule

    3. Z scores serve as a ruler for any distribution

    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$

    5. Normally distributed data plot as a straight line on the normal probability plot

3. Summary

4. Exercise [time permitting]

At a pharmaceutical factory the amount of the active ingredient which is added to each pill is supposed to be 36 mg. The amount of the active ingredient added follows a nearly normal distribution with a standard deviation of 0.11 mg. Once every 30 minutes a pill is selected from the production line, and its composition is measured precisely. We know that the failure rate of the quality control is 3% at this factory. What are the bounds of the acceptable amount of the active ingredient?