Statistical Methods and Data Analysis I
Lecture 7: Probability Distributions — Centers and Widths

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

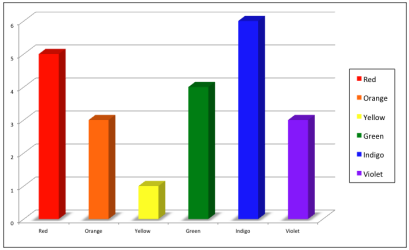
Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

March 24, 2019

Central Tendency: Mode

Mode

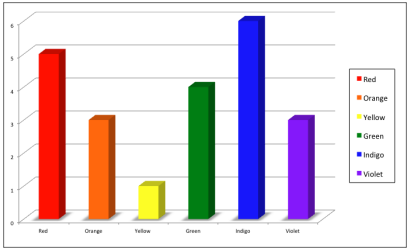
The most common, most probable, most popular, or most typical value of a random variable (outcome of a random process).



Central Tendency: Mode

Mode

The most common, most probable, most popular, or most typical value of a random variable (outcome of a random process).



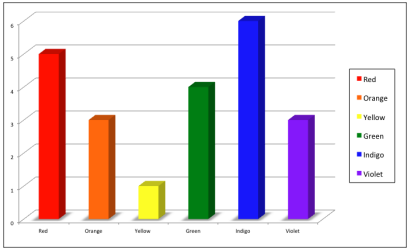
Less boring examples than “what is your favourite colour?”:

- *Which party will form a government after elections?*

Central Tendency: Mode

Mode

The most common, most probable, most popular, or most typical value of a random variable (outcome of a random process).



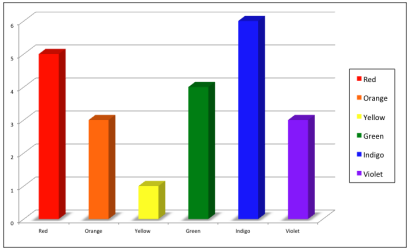
Less boring examples than “what is your favourite colour?”:

- *Which party will form a government after elections?*
  The one that got the most votes (but this won’t tell you how stable the coalition will be).

Central Tendency: Mode

Mode

The most common, most probable, most popular, or most typical value of a random variable (outcome of a random process).



Less boring examples than “what is your favourite colour?”:

- *Which party will form a government after elections?*
  The one that got the most votes (but this won’t tell you how stable the coalition will be).

NB: It does not make sense to compute the average colour or the average number of votes among the competing parties (except maybe to assess how fragmented the parliament is).

Central Tendency: Mean

Discrete case:
($p(x_i)$ is the probability of outcome x_i)

$$\overline{x} = \mu = \sum_i x_i p(x_i)$$

Continuous case:
($p(x)$ is the probability density function, PDF)

$$\overline{x} = \mu = \int_{-\infty}^{+\infty} x p(x) dx$$

If the probabilities of all outcomes x_i are the same (uniform distribution) then, for n possible outcomes, we have our usual definition of arithmetic mean:

$$p(x_i) = \frac{1}{n}, \qquad \overline{x} = \mu = \frac{1}{n} \sum_i x_i$$

If we know the probabilities of all the outcomes (we know the probability density function in the continuous case) then our mean is the *(mathematical) expectation*, *expected value*: $E[x] = \mu$.

Expectation Value: Stocks

Stock market scenario

You hold a 100 shares in AAPL, currently worth $S = \$191.05$ each. In a month's time it will become clear that the newest iPhone will either be a hit, and the stock will gain 15% in value, or it will be a flop, and the stock will drop by 10%.

You estimate that the probability of a flop is 45%, and of a hit — 55%.

How much do you expect to gain or lose after a month?

Expectation Value: Stocks

Stock market scenario

You hold a 100 shares in AAPL, currently worth $S = \$191.05$ each. In a month's time it will become clear that the newest iPhone will either be a hit, and the stock will gain 15% in value, or it will be a flop, and the stock will drop by 10%.

You estimate that the probability of a flop is 45%, and of a hit — 55%.

How much do you expect to gain or lose after a month?

$$E[\Delta S] = 100 \times S \times [0.15 \times 0.55 + (-0.10) \times 0.45] = 100 \times \$191.05 \times 0.0375 = \$716.44$$

Expectation Value: Stocks

Stock market scenario

You hold a 100 shares in AAPL, currently worth $S = \$191.05$ each. In a month's time it will become clear that the newest iPhone will either be a hit, and the stock will gain 15% in value, or it will be a flop, and the stock will drop by 10%.

You estimate that the probability of a flop is 45%, and of a hit — 55%.

How much do you expect to gain or lose after a month?

$$E[\Delta S] = 100 \times S \times [0.15 \times 0.55 + (-0.10) \times 0.45] = 100 \times \$191.05 \times 0.0375 = \$716.44$$

What should your estimation of the probability p of a flop be for you to expect a loss?

Expectation Value: Stocks

Stock market scenario

You hold a 100 shares in AAPL, currently worth $S = \$191.05$ each. In a month's time it will become clear that the newest iPhone will either be a hit, and the stock will gain 15% in value, or it will be a flop, and the stock will drop by 10%.

You estimate that the probability of a flop is 45%, and of a hit — 55%.

How much do you expect to gain or lose after a month?

$$E[\Delta S] = 100 \times S \times [0.15 \times 0.55 + (-0.10) \times 0.45] = 100 \times \$191.05 \times 0.0375 = \$716.44$$

What should your estimation of the probability p of a flop be for you to expect a loss?

$$E[\Delta S] = 0 = 100 \times S \times [0.15 \times (1-p) + (-0.10) \times p] \quad \Rightarrow \quad p = 0.6$$

Central Tendency: Median

Median

The “middle value” of a sample or a distribution: 50% of the values lie above the median, 50% — below.

What is the median score for a fair dice?

Central Tendency: Median

Median

The “middle value” of a sample or a distribution: 50% of the values lie above the median, 50% — below.

What is the median score for a fair dice?

$$1, 2, \underline{3}, \underline{4}, 5, 6 \quad \rightarrow \quad \frac{3+4}{2} = 3.5$$

Central Tendency: Median

Median

The “middle value” of a sample or a distribution: 50% of the values lie above the median, 50% — below.

What is the median score for a fair dice?

$$1, 2, \underline{3}, \underline{4}, 5, 6 \quad \rightarrow \quad \frac{3+4}{2} = 3.5$$

What is the mean score for a fair dice?

Central Tendency: Median

Median

The "middle value" of a sample or a distribution: 50% of the values lie above the median, 50% — below.

What is the median score for a fair dice?

$$1, 2, \underline{3, 4}, 5, 6 \quad \rightarrow \quad \frac{3+4}{2} = 3.5$$

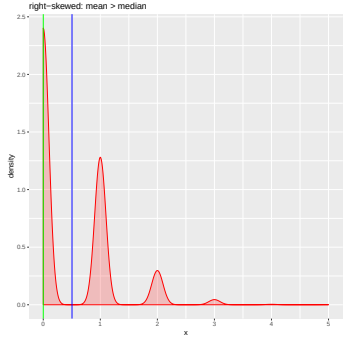
What is the mean score for a fair dice?

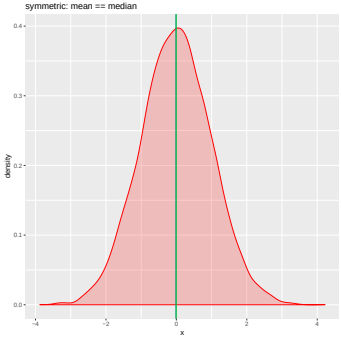
$$\frac{1+2+3+4+5+6}{6} = 3.5$$

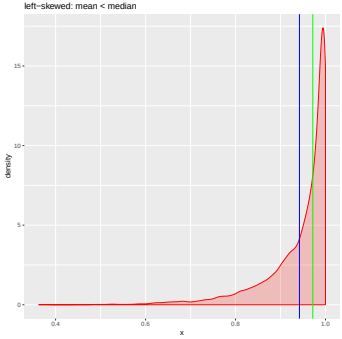
Mean vs. Median

If the distribution is *symmetric* the center is often defined as the *mean*:

If the distribution is skewed then the center is usually defined as *median*:







Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Income distribution is skewed, thus median is more appropriate

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Income distribution is skewed, thus median is more appropriate

Assume that poverty is defined as a percentage of mean disposable income. Would you use half the mean as the poverty line? A higher or lower percentage?

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Income distribution is skewed, thus median is more appropriate

Assume that poverty is defined as a percentage of mean disposable income. Would you use half the mean as the poverty line? A higher or lower percentage?

The income distribution is right-skewed: there are many more lower income households and relatively few rich ones. Therefore, mean $>$ median, and we should use a lower percentage.

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Income distribution is skewed, thus median is more appropriate

Assume that poverty is defined as a percentage of mean disposable income. Would you use half the mean as the poverty line? A higher or lower percentage?

The income distribution is right-skewed: there are many more lower income households and relatively few rich ones. Therefore, mean $>$ median, and we should use a lower percentage.

Mark Zuckerberg, Sergey Brin, George Soros and Michael Bloomberg are Jewish. They are also very rich. What if they all join Roman Abramovich and make aliyah? Will the influx of money improve the poverty situation in Israel?

Example: Poverty Line

Definition of poverty in Israel

"Poverty line" is defined as half of the *median* disposable income, weighted by household size.

Why is median used and not mean?

Income distribution is skewed, thus median is more appropriate

Assume that poverty is defined as a percentage of mean disposable income. Would you use half the mean as the poverty line? A higher or lower percentage?

The income distribution is right-skewed: there are many more lower income households and relatively few rich ones. Therefore, mean > median, and we should use a lower percentage.

Mark Zuckerberg, Sergey Brin, George Soros and Michael Bloomberg are Jewish. They are also very rich. What if they all join Roman Abramovich and make aliyah? Will the influx of money improve the poverty situation in Israel?

If poverty were measured as a percentage of the mean income, then poverty would **increase**. *Since poverty is measured based on median income it will not be affected much.*

Outliers affect the mean more than the median. Median is said to be more **robust**.

Lying About Means, Medians and Modes

Month	Revenue
January	$10K
February	$10K
March	$10K
April	$10K
May	$10K
June	$10K
July	$10K
August	$10K
September	$10K
October	$10K
November	$10K
December	$610K

Lying About Means, Medians and Modes

Month	Revenue
January	\$10K
February	\$10K
March	\$10K
April	\$10K
May	\$10K
June	\$10K
July	\$10K
August	\$10K
September	\$10K
October	\$10K
November	\$10K
December	\$610K

- Mode = \$10K
- Median = \$10K
- Mean = (\$10K $\times$ 11 + \$610K)/12 = \$60K
- *Another manifestation of mean's sensitivity to outliers and skewness...*
- True, supported by data, not necessarily what your audience want to hear
- Your audience is probably not that stupid...
- But you can tell better lies...

Percentiles and Quartiles

Percentiles (a.k.a. quantiles

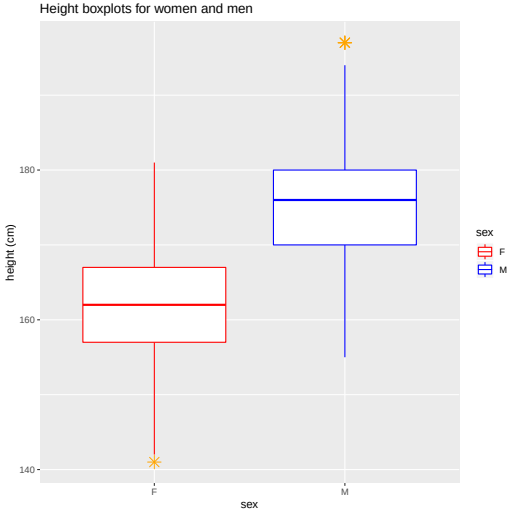
The n-th percentile is the value below which lie $n\%$ of the distribution.

- The 25-th percentile is called the first quartile, Q_1.
- Median is, by definition, the 50-th percentile and the second quartile, Q_2.
- The 75-th percentile is called the third quartile, Q_3.
- The middle 50% of the data lie between the first and the third quartiles.
- The range of values containing the middle 50% of the data is called *the interquartile range*, *IQR*:

$$IQR = Q_3 - Q_1$$

- *IQR* is a simple and useful measure of how wide the distribution is.
- It is also **robust** w.r.t. skewness and outliers.

Anatomy of a Boxplot



- A boxplot typically shows
  - median
  - quartiles
  - IQR
  - "whiskers":
    - $Q_1 - 1.5 \times IQR$
    - $Q_3 + 1.5 \times IQR$
  - "outliers": the points beyond the whiskers
- Common sense: the bulk of the distribution is included within $4 \times IQR$.
- "Outliers" *may* indicate
  - unusual/extreme skew
  - data collection/entry errors
  - more than one population
  - interesting features

Variance and Standard Deviation

Definition of variance

Variance is the average squared deviation from the mean.

Discrete case:

($p(x_i)$ is the probability of outcome x_i)

$$Var\,[x] = \sigma_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 p(x_i)$$

Continuous case:

($p(x)$ is the probability density function, PDF)

$$Var\,[x] = \sigma_x^2 = \int_{-\infty}^{+\infty} (x - \bar{x})^2 p(x) dx$$

Definition of standard deviation

Standard deviation is the square root of variance.

$$SD\,[x] = \sigma_x = \sqrt{Var\,[x]}$$

Population and Sample Variance

If all the outcomes occur with the same probability $p = 1/n$ the expression for the variance becomes

$$MSD(x) = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2$$

This is usually called **mean squared deviation**, *corresponding to the formula.*
For variance it is customary to use

$$Var\,[x] = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

which is more useful to make inference about the underlying population based on the sample for which variance is computed, because there are only $n - 1$ *degrees of freedom*.

Degrees of Freedom

Degrees of Freedom

- Consider a sample of 1.
  - The single observation is the sample mean — an estimate of the real mean of the population.
  - There is no spread though — we have no idea what the population variance is.
  - We can get any idea of the spread only with $n > 1$.

Degrees of Freedom

- Consider a sample of 1.
  - The single observation is the sample mean — an estimate of the real mean of the population.
  - There is no spread though — we have no idea what the population variance is.
  - We can get any idea of the spread only with $n > 1$.
- Consider a sample of 2. Suppose we measured the height of two men and obtained 171 cm and 177 cm. The sample mean is 174 cm, and the deviations are $+3$ and -3 cm. However, the sum of the deviations must be 0 in this case, so only one of them is "free".

Degrees of Freedom

- Consider a sample of 1.
  - The single observation is the sample mean — an estimate of the real mean of the population.
  - There is no spread though — we have no idea what the population variance is.
  - We can get any idea of the spread only with $n > 1$.
- Consider a sample of 2. Suppose we measured the height of two men and obtained 171 cm and 177 cm. The sample mean is 174 cm, and the deviations are $+3$ and -3 cm. However, the sum of the deviations must be 0 in this case, so only one of them is "free".
- In general, in a sample of n observations there are $n - 1$ "degrees of freedom". The last one is fixed by the requirement that the sum of all the deviations must be 0.

Degrees of Freedom

- Consider a sample of 1.
  - The single observation is the sample mean — an estimate of the real mean of the population.
  - There is no spread though — we have no idea what the population variance is.
  - We can get any idea of the spread only with $n > 1$.
- Consider a sample of 2. Suppose we measured the height of two men and obtained 171 cm and 177 cm. The sample mean is 174 cm, and the deviations are $+3$ and -3 cm. However, the sum of the deviations must be 0 in this case, so only one of them is "free".
- In general, in a sample of n observations there are $n - 1$ "degrees of freedom". The last one is fixed by the requirement that the sum of all the deviations must be 0.

$$\sum_{i=1}^{n}(x - \bar{x}) = n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

Degrees of Freedom

- Consider a sample of 1.
  - The single observation is the sample mean — an estimate of the real mean of the population.
  - There is no spread though — we have no idea what the population variance is.
  - We can get any idea of the spread only with $n > 1$.
- Consider a sample of 2. Suppose we measured the height of two men and obtained 171 cm and 177 cm. The sample mean is 174 cm, and the deviations are $+3$ and -3 cm. However, the sum of the deviations must be 0 in this case, so only one of them is "free".
- In general, in a sample of n observations there are $n - 1$ "degrees of freedom". The last one is fixed by the requirement that the sum of all the deviations must be 0.

$$\sum_{i=1}^{n}(x - \bar{x}) = n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

It makes sense to normalize by the degrees of freedom, or the number of independent pieces of information.

Linear Algebra of Random Variables: Mean

Linear combination of **independent** random variables: $z = ax + by$

$$\bar{z} = \sum_{x_i}\sum_{y_j}(ax_i + by_j)p_x(x_i)p_y(y_j) = a\sum_{x_i} x_i p_x(x_i)\sum_{y_j} p_y(y_j) +$$
$$+b\sum_{x_i} p_x(x_i)\sum_{y_j} y_i p_y(x_j) = a\,(\bar{x} \times 1) + b\,(1 \times \bar{y}) = a\bar{x} + b\bar{y}$$

$$\bar{z} = \int_{-\infty}^{+\infty} dx p_x(x) \int_{-\infty}^{+\infty} dy p_y(y)(ax + by) = a\int_{-\infty}^{+\infty} dx p_x(x) x \int_{-\infty}^{+\infty} dy p_y(y) +$$
$$+b\int_{-\infty}^{+\infty} dx p_x(x) \int_{-\infty}^{+\infty} dy p_y(y) y = a\,(\bar{x} \times 1) + b\,(1 \times \bar{y}) = a\bar{x} + b\bar{y}$$

Non-Linear Algebra of Random Variables: Variance

Linear combination of **independent** random variables: $z = ax + by$

$$\textbf{Prove}: \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2, \quad \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$$

$$\textbf{For standard deviation}: \quad \sigma_z = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}, \quad \sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Non-Linear Algebra of Random Variables: Variance

Linear combination of **independent** random variables: $z = ax + by$

$$\textbf{Prove}: \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2, \quad \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$$

$$\textbf{For standard deviation}: \quad \sigma_z = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}, \quad \sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Note that random variables do not behave like normal variables:

$$E[x+x] = \overline{x+x} = 2\overline{x} \qquad Var[x+x] = \sigma_{x+x}^2 = 2\sigma_x^2 \qquad SD[x+x] = \sigma_{x+x} = \sqrt{2}\sigma_x$$

$$E[2x] = \overline{2x} = 2\overline{x} \qquad Var[2x] = \sigma_{2x}^2 = 4\sigma_x^2 \qquad SD[2x] = \sigma_{2x} = 2\sigma_x$$

Non-Linear Algebra of Random Variables: Variance

Linear combination of **independent** random variables: $z = ax + by$

$$\textbf{Prove}: \quad \sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2, \quad \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$$

$$\textbf{For standard deviation}: \quad \sigma_z = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}, \quad \sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Note that random variables do not behave like normal variables:

$$E[x+x] = \overline{x+x} = 2\overline{x} \qquad Var[x+x] = \sigma_{x+x}^2 = 2\sigma_x^2 \qquad SD[x+x] = \sigma_{x+x} = \sqrt{2}\sigma_x$$

$$E[2x] = \overline{2x} = 2\overline{x} \qquad Var[2x] = \sigma_{2x}^2 = 4\sigma_x^2 \qquad SD[2x] = \sigma_{2x} = 2\sigma_x$$

Betting twice the money on a coin toss or dice roll ($2x$) is *not the same* as betting on tossing two coins or rolling two dice ($x + x$)!