# Lecture 30: Probability Theory

Chapter 2.4-2.5

# Random Variable

A random process or variable with a numerical outcome is called a random variable, and is typically denoted by an upper case letter. E.g. $X$, $Y$, or $Z$

# Intuitively Thinking: Expected Value

Easy example: Coin flips. Say we flip a fair coin $n = 10$ times with probability $p = \frac{1}{2}$ of heads.

How many heads do you expect to get?

$n \times p = 10 \times \frac{1}{2} = 5$

# Intuitively Thinking: Expected Value

Say you have a random variable $X$:

| $x$ | 2 | 3 | 4 | 10 | 11 |
|---|---|---|---|---|---|
| $\Pr(X = x)$ | $\frac{15}{100}$ | $\frac{25}{100}$ | $\frac{10}{100}$ | $\frac{30}{100}$ | $\frac{20}{100}$ |

E.g. We observe $X = 3$ with prob .25

Is the value we expect to observe:

$$\frac{2+3+4+10+11}{5} \quad = \quad 6\ ?$$

# Intuitively Thinking: Expected Value

No, each of the $x$'s have different probability of occurring.

For each $x$, we assign different weights $\Pr(X = x)$ and not $\frac{1}{5}$:

$$2 \times \frac{15}{100} + 3 \times \frac{25}{100} + 4 \times \frac{10}{100} + 10 \times \frac{30}{100} + 11 \times \frac{20}{100} = 6.65$$

# Expected Value

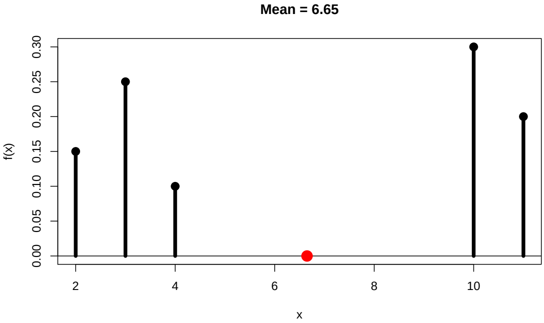
The expected value is a weighted average of all possible values $x$. This can be thought of as a measure of center:

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i \cdot \Pr(X = x_i)$$

This is also called the mean and expectation of $X$. Typically denoted by $\mu$.

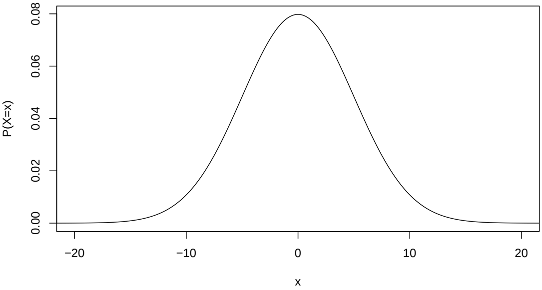
# Expected Value

You can also think of the mean as the center of mass or balance point (marked with red point):



# Intuitively Thinking: Measures of Spread

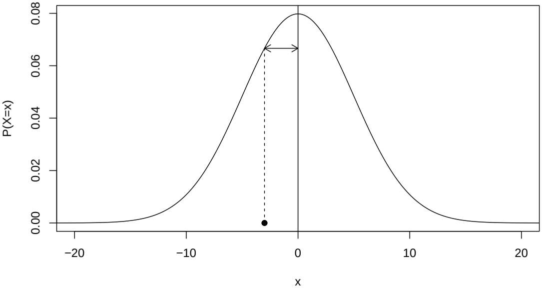
Consider the following (continuous) distribution with $\mu = 0$. Let's build a measure of expected "spread".




Let's define "spread" as the absolute deviation from $\mu$: $|x - \mu|$.
i.e. $+$'ve & $-$'ve deviations are treated the same.

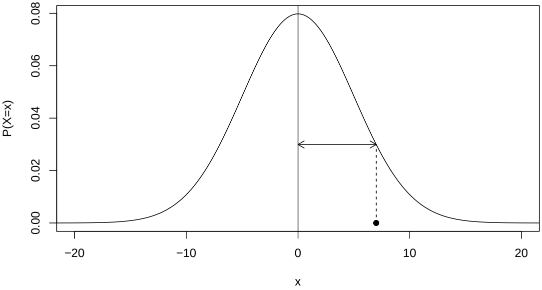
# Intuitively Thinking: Measures of Spread

When $x = -3.0$, the abs. deviation from $\mu$ is $|-3.0 - \mu| = 3.0$.
Note $P(X = x) = 0.066$.



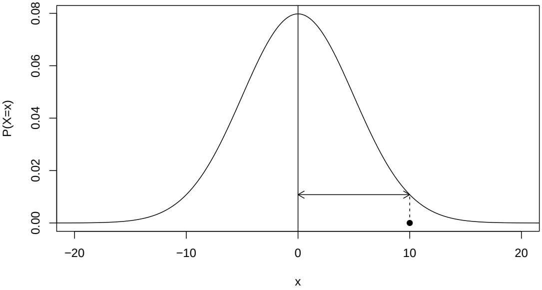
# Intuitively Thinking: Measures of Spread

When $x = 7.0$, the abs. deviation from $\mu$ is $|7.0 - \mu| = 7.0$.
Note $P(X = x) = 0.030$.



# Intuitively Thinking: Measures of Spread

When $x = 10.0$, the abs. deviation from $\mu$ is $|10.0 - \mu| = 10.0$.
Note $P(X = x) = 0.011$.



# Intuitively Thinking: Measures of Spread

| x | Abs Deviation $\lvert x-\mu \rvert$ | Weight $P(X=x)$ |
|---|---|---|
| -3.0 | $\lvert -3.0-0 \rvert = 3.0$ | 0.066 |
| 7.0 | $\lvert 7.0-0 \rvert = 7.0$ | 0.030 |
| 10.0 | $\lvert 10.0-0 \rvert = 10.0$ | 0.011 |

So say we do this for all $x$ and take a weighted average of the $|x-\mu|$ where the weights are $P(X=x)$.

Voilà: Our notion of expected spread.

# Variance

The variance $\sigma^2$ AKA Var(X) of a distribution is

$$\mathbb{E}\left[(X-\mu)^2\right] = \sum_{i=1}^{k}(x_i-\mu)^2 \cdot P(X=x_i)$$

It is the expected squared deviation from the mean, and not absolute deviation (like in our example). i.e. not

$$\mathbb{E}\left[|X-\mu|\right] = \sum_{i=1}^{k}|x_i-\mu| \cdot P(X=x_i)$$

Why square? Treats $+$'ve and $-$'ve deviations as the same, but also easier to do calculus on $x^2$ than $|x|$.

# Estimators

$\overline{x}$ is a point *estimate* of $\mu$. $\overline{x}$ is based on observed data.

Before we've observed the data, $\overline{X}$ is still random and is the *estimator* of $\mu$.

# Sample Mean as an Estimator

So whereas, for a single $X$

- $\mathbb{E}[X] = \mu$
- $\mathrm{Var}[X] = \sigma^2$

we have for $\overline{X}$

- $\mathbb{E}\left[\ \overline{X}\ \right] = \mu$
- $\mathrm{Var}\left[\ \overline{X}\ \right] = \frac{\sigma^2}{n}$

i.e. as $n \longrightarrow \infty$ the variance goes to 0.

# Bias

One property we want our estimators to have is unbiasedness. i.e.

$$\mathbb{E}[\widehat{\theta} - \theta] = 0 \quad \Rightarrow \quad \mathbb{E}[\widehat{\theta}] = \theta$$

i.e. we expected the estimator's value to be the unknown parameter.

# Recall from Earlier

One example of a non-representative sample is a biased sample. For example, convenience samples are samples where individuals who are easily accessible are more likely to be included.

# Recall from Earlier

1. The Royal Air Force wants to study how resistant their airplanes are to bullets. They study the bullet holes on all the airplanes on the tarmac after an air battle against the Luftwaffe (German Air Force).
2. I want to know the average income of Reed graduates in the last 10 years. So I get the records of 10 randomly chosen Reedies. They all answer and I take the average.
3. Imagine it's 1993 i.e. almost all households have landlines. You want to know the average number of people in each household in Portland. You randomly pick out 500 phone numbers from the phone book and conduct a phone survey.
4. You want to know the prevalence of illegal downloading of TV shows among Reed students. You get the emails of 100 randomly chosen Reedies and ask them "How many times did you download a pirated TV show last week?"