# 1: Fundamentals of Null-Hypothesis Significance Testing (Part 2)

Bernhard Angele

12 January 2018

# Maths basics: Probability

- Basic rules:
  - All probabilities are between 0 and 1: $P(A) \in [0, 1]$
  - The complementary probability of an event (i.e. the probability that an event will NOT happen) is 1-the probability of the event: $P(A^c) = 1 - P(A)$
  - A probability can be interpreted as the number of outcomes that form an event (e.g. the outcome "Heads" when flipping a coin) over the total number of outcomes (e.g. "Heads" and "Tails")

$$p(A) = \frac{n_A}{n}$$

  - But note that a probability of .5 (e.g. for getting "Heads" on a coin flip) doesn't mean that you will get "Heads" on exactly 50% of coin flips.

# Maths basics: Probability (2)

- Basic rules:
  - What is the probability that Event A and Event B will happen together?

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$
$$P(A \cap B) = P(A)P(B)$$
$$\text{if A and B are independent}$$

- What is the probability that either Event A OR Event B will happen?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(B)$$
$$\text{if A and B are mutually exclusive}$$

# Maths basics: Conditional probability

- What is the probability of Event A *GIVEN THAT* Event B happened?
  - Divide the number of outcomes where A and B happen together by the number of all outcomes where B happens (regardless of whether A happened, too).
  - If we divide both nominator and denominator by $n$, we can convert this into probabilities:

$$P(A \mid B) = \frac{n_{AB}}{n_B} = \frac{n_{AB}/n}{n_B/n} = \frac{P(A \cap B)}{P(B)}$$

- Now plug in our definition of joint probability (see last slide):

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- This is known as *Bayes' theorem*. Keep it in mind for later!

# Back to our dice problem

- IN THEORY, each of our dice roll outcomes has the same probability:

$$
\begin{aligned}
p(1) &= \frac{n_1}{n_{total}} = \frac{1}{6}, p(2) = \frac{n_2}{n_{total}} &= \frac{1}{6} \\
p(3) &= \frac{n_3}{n_{total}} = \frac{1}{6}, p(4) = \frac{n_4}{n_{total}} &= \frac{1}{6} \\
p(5) &= \frac{n_5}{n_{total}} = \frac{1}{6}, p(6) = \frac{n_6}{n_{total}} &= \frac{1}{6}
\end{aligned}
$$

- But how can we test whether that is actually true?
- We need some way to compare the data to the theoretical probability distribution

# Aggregating our dice results

- We obviously need to take more than one dice roll into account. But how can we aggregate all our results in a convenient number?
- As luck would have it, descriptive statistics provides us with a number of standard measures to characterise the properties of a *sample* (e.g. rolling the dice 10 times):
- Measures of *central tendency*:
- The mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The median: The number separating the higher half of a sample from the lower half
- The mode: The most frequent observation
- Measures of dispersion:
- The standard deviation: $s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

# Statistics basics: Random variables

- We need to consider our sample as a random variable
- What is a random variable?
- A random variable is a function that assigns a number to each possible outcome of our experiment (the dice roll)
- The outcome of a single dice roll can be described by a very obvious function: just assign the numbers from 1 to 6
- The outcome of *multiple* dice rolls is little trickier
- Regardless of which one we choose, we can then come up with a *theoretical* probability distribution for the random variable.
- The opposite of a random variable is a *constant*, a value that's the same for every sample.

# Statistics basics: Random variables (formal definition!)

- Warning: Some mathematical notation follows.
- A random variable $X$ is a function $X : O \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in O$ exactly one number $X(\omega) = x$.
- Note: $\mathbb{R}$ = real numbers
- $O_X$ is all the $x$'s (all the possible values of X, the support of X). i.e., $x \in O_X$.
- Good example: number of coin tosses till you get Heads (H) for the first time
- $X : \omega \rightarrow x$
    - $\omega$: H, TH, TTH,... (infinite)
    - $x = 0, 1, 2, \dots; x \in O_X$

# Random variables (2)

Every discrete random variable X has associated with it a **probability mass function**, also called *distribution function*.

$$p_X : S_X \rightarrow [0, 1]$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X$$

- Back to the example: number of coin tosses till H

- $X : \omega \rightarrow x$
- $\omega$: H, TH, TTH,... (infinite)
  - $x = 0, 1, 2, \ldots; x \in S_X$
- $p_X = .5, .25, .125, \ldots$

# What is a probability distribution?

From Wikipedia: In probability and statistics, a probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference. Here's an example of a discrete probability distribution, the distribution of the *sum of two dice rolls*:

# Discrete probability distribution

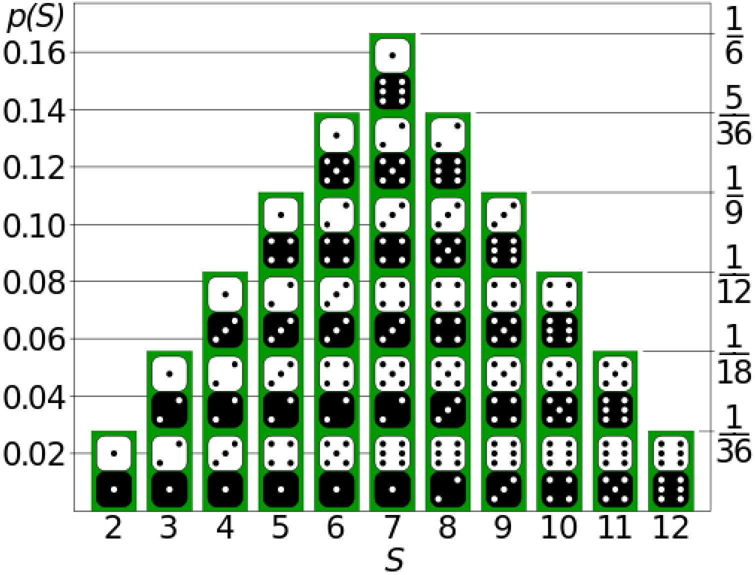


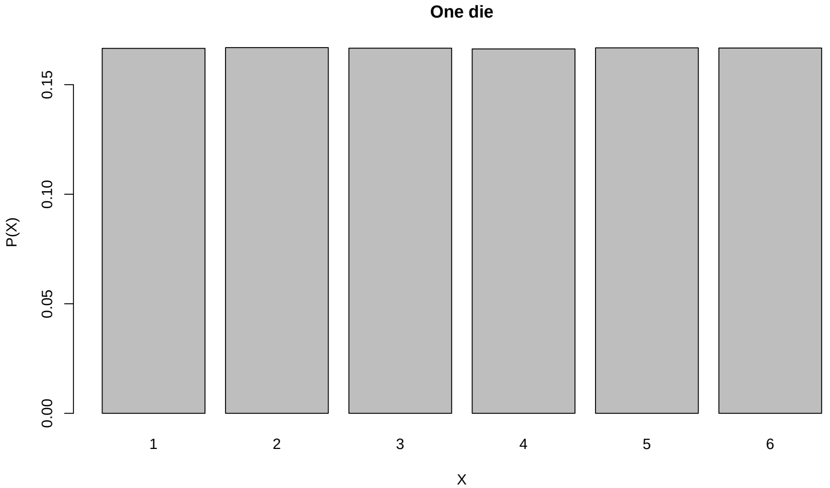

Figure 1: Dice Distribution (from Wikipedia)

# A quick clarification: Sample and population

- With the introduction of *theoretical* probability distributions, we need to be very careful to not confuse properties of the theoretical distribution with properties of an individual sample.
- Standard practice is to use
  - roman letters (e.g. $m$ or $\bar{x}$ for the mean, $s$ for the standard deviation) for properties of the sample
  - greek letters (e.g. $\mu$ "mu" for the mean and $\sigma$ "sigma" for the standard deviation) for properties of the distribution (or the population that is represented by the distribution).

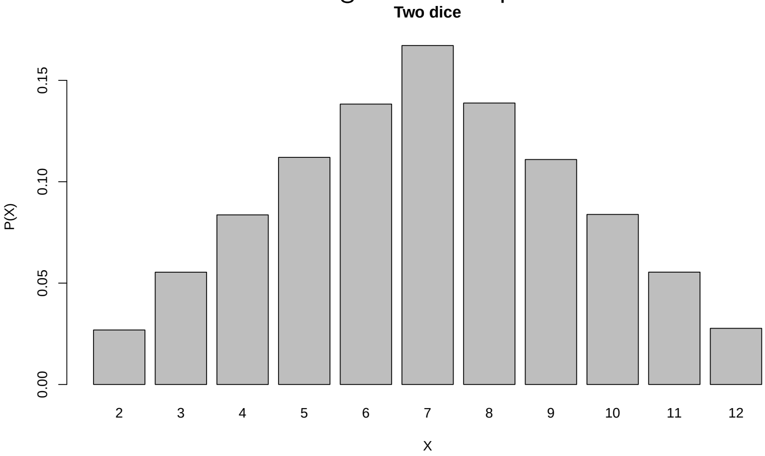
# From discrete to continuous

- Let's look at the probability distributions we get from rolling one, two, three etc. dice and summing up the results.
- We'll start with rolling one die (note that the bars may be a tiny bit uneven since I used a simulation to produce this graph).

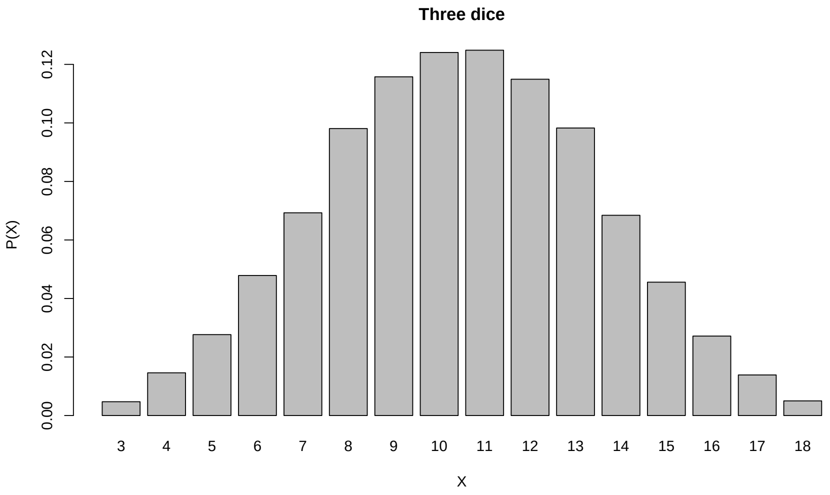


# Two dice

- This is the same as the image from Wikipedia.

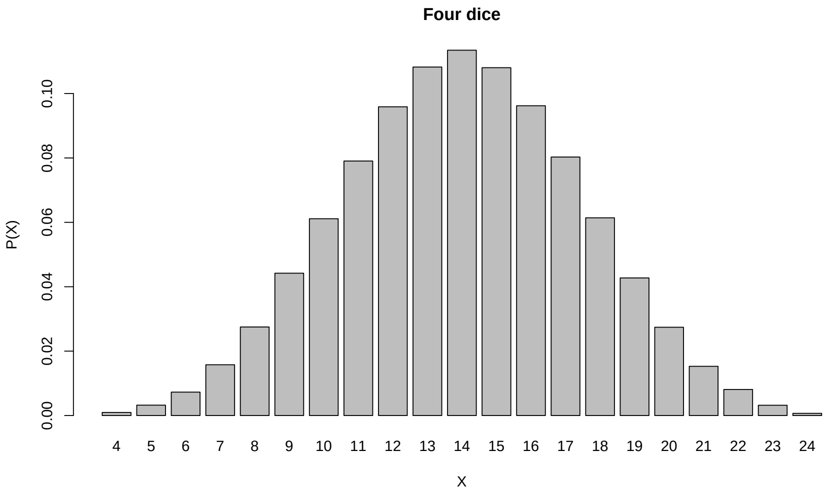


# Three dice

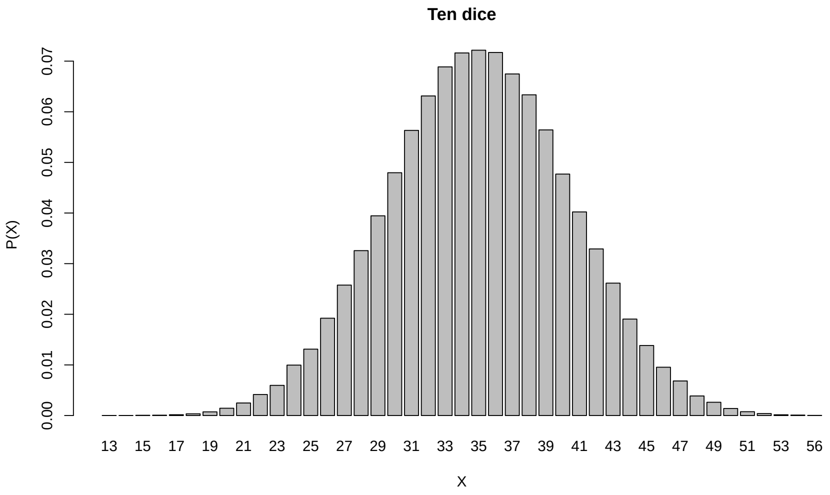


Three dice
P(X)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
X

# Four dice

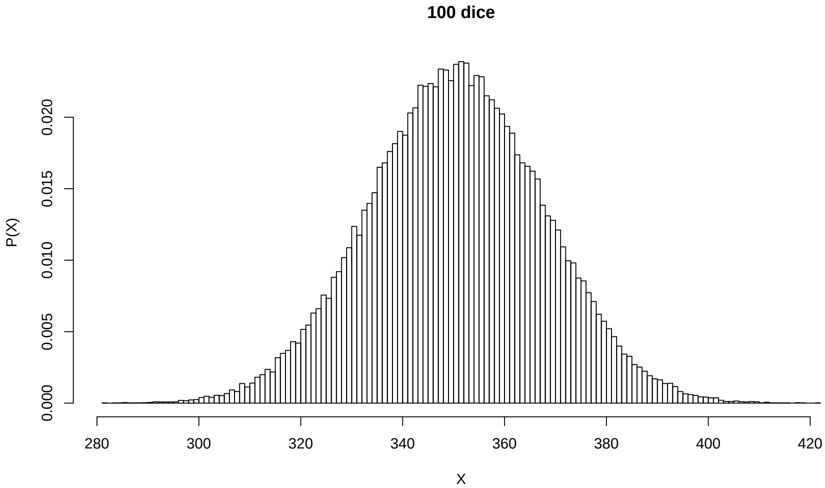

Four dice
P(X)
0.00
0.02
0.04
0.06
0.08
0.10
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
X

# Ten dice



Ten dice
P(X)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 56
X

# 100 dice



100 dice
P(X)
0.000
0.005
0.010
0.015
0.020
280
300
320
340
360
380
400
420
X

# Do you see a pattern?

- Central limit theorem (CLT)

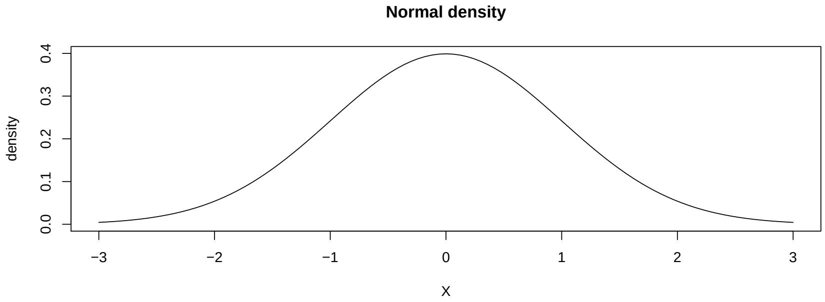
  *When sampling from a population that has a mean, provided the sample size is large enough, the sampling distribution of the sample mean will be close to normal regardless of the shape of the population distribution*

- (Technically, we were sampling the sum of X rather than the mean, but the mean of X is simply the sum divided by the number of observations. Do you care about this distintion? Didn't think so. It makes me feel better, though.)

# What does this mean?

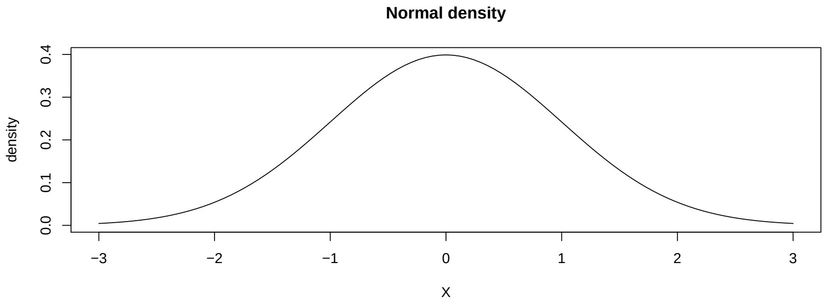
- For our dice problem, it means that we can compute the means of our samples (e.g. the mean of the 5, or 10, or 100 samples)
- Remember, the *sample mean* is a random variable as well, since it is different every time we take a sample
- We can then use a *continuous* probability distribution – the **normal distribution** as the theoretical probability distribution for our random variable (i.e. the sample mean).
- This makes our life easy, because the normal distribution is very simple to handle mathematically (really!).

# Continuous probability distributions




- Here, the outcomes are continuous, so it doesn't make sense to ask about the probability of any point on the x-axis.
- What is the probability of $x = 1$?

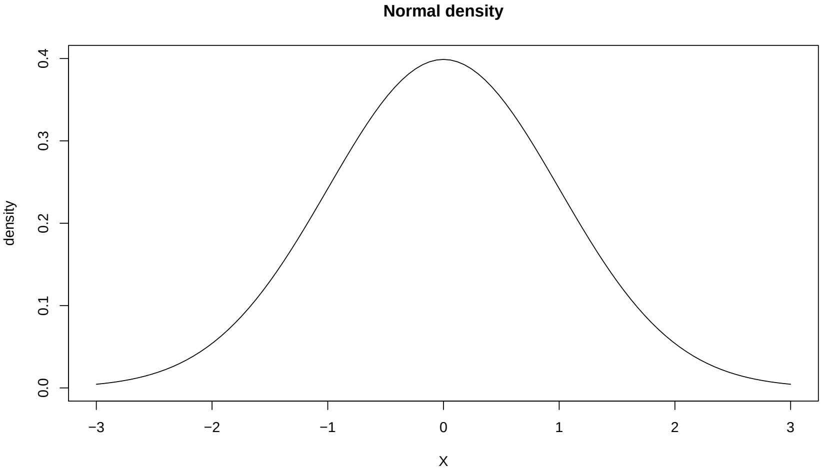
# Continuous probability distributions




- What do you mean by “1”? The function is continuous, so does 1.00001 still qualify as 1?
- It makes more sense to ask these questions about intervals. The probability is then the area under the curve for the interval.
- Important: the total area under the curve is 1.

# Normal probability density function (PDF)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-((x-\mu)^2/2\sigma^2)}$$



# Normal probability density function (PDF)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-((x-\mu)^2/2\sigma^2)}$$

- This looks scary, but it really isn't. This is simply a mathematical function that happens to describe the distribution of a lot of random variables in nature. - If you look closely, you can see that the function has three parameters, $x, \mu$, and $\sigma$ ($\pi$ and $e$ are constants). - The first parameter, $x$ is the random variable. The function gives you the probability density at each value of $x$ - The second parameter, $\mu$, is called the **expected value** or the **mean** of the distribution. - The third parameter, $\sigma$, is called the standard deviation.

# Standard normal distribution

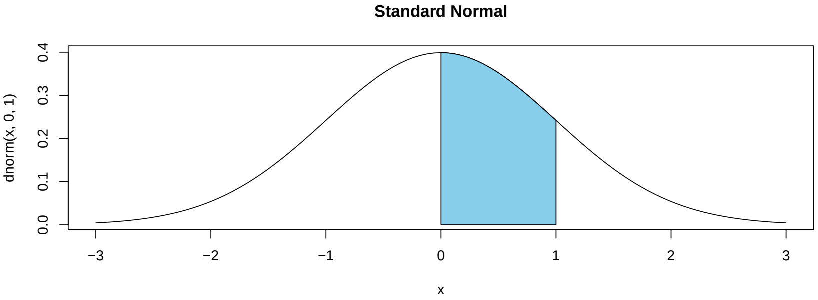
- There is an infinite number of normal distributions with different parameters $\mu$ and $\sigma$. The one with $\mu = 0$ and $\sigma = 1$ is particularly useful and is called the *standard* normal distribution.
- Look at how simple and nice the normal distribution appears when we plug in those values:

$$f(z, \mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- You can *transform* values from any normal distribution to the normal distribution.
- This is known as a *z-transformation*: $z = \frac{x-\mu}{\sigma}$
- By transforming all our observations to z-values and then looking up their probability in the standard normal distribution, this is the only distribution we'll ever need (. . . mostly).

# The probability of outcomes in the standard normal distribution

- Remember, we can't really get the probability of a *point* event in a continuous distribution, since there are no "points" in a continuous variable
- But we can ask questions about *intervals*:
- What's the probability of x being between 0 and 1?



# Getting the area under the curve

- Since we know exactly what the function is, we can get the area under the curve.
- Remember how to do that from maths class? Your best friend, integration :

$$p(0 < z < 1) = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

- Don't want to do integration? Well, you're in luck, because most statistical software (and Excel!) can do this for you.
- `=NORM.S.DIST(0,TRUE)` will give you the area under the curve to the left of 0 (i.e. the probability that $z < 0$)
- `=NORM.S.DIST(1,TRUE)` will give you the area under the curve to the left of 1 (i.e. the probability that $z < 1$)

# Getting the area under the curve

- So, for our interval: $p(0 < z < 1) = p(z < 1) - p(z < 0)$
- Remember, Excel gives us the upper tail (the area under the curve to the right of the $z$ value)
- So we rewrite our interval: since $p(z < 1) = 1 - p(z > 1)$ and $p(z < 0) = 1 - p(z > 0)$, $p(0 < z < 1) = 1 - p(z > 1) - (1 - p(z > 0))$
- In Excel:
  - `=NORM.S.DIST(1,TRUE)-NORM.S.DIST(0, TRUE)`
  - Result: 0.3413447
- Success!

# Things you can do with this knowledge

- Say I'm looking at random numbers from a standard normal distribution, and I see that one of them is 4.
- That seems very unusual
- Just how unusual?
  - What's the probability of getting a value of 4 when sampling from a standard normal distribution (mean $= 0$, sd $= 1$)?

# Just how "unusual" is a value of 4?

- Remember, when you have a continuous distribution, you can't think about point values (e.g. 5). Rather, what you want to know is:
  - What is the probability of getting a value of 4 *or greater* (or $p(z > 4) = 1 - p(z < 4)$)?
- Let's ask Excel: `=1-NORM.S.DIST(4,TRUE)`
  - Result: 0.0001338302
- So it's very unusual.
- Can we come up with a similar test for our dice sample mean?
- We'll have to figure out how the dice sample means are distributed.
  - Then we can take our sample mean and see how likely (or unlikely) it is that it comes from the theoretical distribution.

# The theoretical distribution of sample means

- We've seen that we can approximate our theoretical distribution (which is actually discrete) using a continuous distribution function, namely the normal distribution, which makes our lives very easy (yes, really!).
- We have to figure out the $\mu$ and the *sigma* parameters for our theoretical normal distribution of sample means, though.
- Note (in case anyone very critical reads this): In the case that we actually know exactly what the probabilities for our discrete probability distribution should look like, we could also use a different distribution, the $\chi^2$ (chi square) distibution. We will talk more about that next week.

# Random variables: Expected value

- Random variables have expected values
- For discrete random variables, the expected value is the outcome value multiplied by the probability of the outcome:

$$E(X) = \mu = \sum_{i=1}^{k} p(x_i) \cdot x_i$$

- where $E(X)$ is the expected value of a discrete random variable $X$ with the outcomes $(x_1 \ldots x_k)$ and the associated probabilities $(p(x_1) \ldots p(x_k))$
- The equivalent for continuous random variables:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

# Random variables: Variance

- Random variables also have variances
- For discrete random variables, the variance is the difference between the outcome value and the mean multiplied by the probability of the outcome:

$$\sigma^2 = \sum_{i=1}^{k} p(x_i) \cdot (x_i - \mu)^2$$

- where $\sigma^2$ is the variance of a discrete random variable $X$ with the outcomes $(x_1 \ldots x_k)$ and the associated probabilities $(p(x_1) \ldots p(x_k))$
- The equivalent for continuous random variables:

$$\mu = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

# Maths basics: Expected values

- For example, the expected value $\mu$ of rolling a six-sided die is:

$$
\begin{aligned}
E(X) &= \sum_{i=1}^{6} p(x_i) \cdot x_i \\
&= x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + x_3 \cdot p(x_3) + x_4 \cdot p(x_4) \\
&+ x_5 \cdot p(x_5) + x_6 \cdot p(x_6) \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} \\
&+ 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{21}{6} = 3.5
\end{aligned}
$$

# Maths basics: Variance

- For example, the variance $\sigma^2$ of rolling a six-sided die is:

$$
\begin{aligned}
\sigma^2 &= \sum_{i=1}^{6} p(x_i) \cdot (x_i - \mu)^2 \\
&= (x_1 - \mu)^2 \cdot p(x_1) + (x_2 - \mu)^2 \cdot p(x_2) + (x_3 - \mu)^2 \cdot p(x_3) \\
&+ (x_4 - \mu)^2 \cdot p(x_4) + (x_5 - \mu)^2 \cdot p(x_5) + (x_6 - \mu)^2 \cdot p(x_6) \\
&= \frac{1}{6} \cdot \Big((1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 \\
&+ (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2\Big) \\
&= \frac{17.5}{6} = 2.9167
\end{aligned}
$$

# Back to the dice example again

- So, we know that, if our dice are fair, our dice rolls come from a discrete theoretical distribution with $\mu = 3.5$ and $\sigma^2 = 2.9167$.
- But remember, we don't want to evaluate single dice rolls, but rather the mean of a sample of dice rolls, since that will enable us to use the nice and easy normal distribution to calculate the probabilities.
- So, what is the mean $\mu_{\bar{x}}$ and what is the variance $\sigma_{\bar{x}^2}$ for the **distribution of sample means**?

# The distribution of sample means

- Remember, we want to know how the means of our dice rolls should be distributed if the dice are fair.
- We're going to take a little bit of a leap here and you will just have to take my word for this:
- The expected value of the mean of our dice rolls will still be $\mu_{\bar{x}} = 3.5$.
- The expected value of the variance of our dice rolls will be $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2.9167}}{\sqrt{10}} = 0.5400648$
- 99% of you are going to accept this without worrying too much about it. But if you are wondering on earth we have to divide by the square root of N, I have prepared an extra file that works through the process of getting to these values. You can find it in the unit materials.

# Hypotheisis test

- We now have everything we need to determine whether our dice roll sample mean is unusual assuming fair dice.
- More formally, we call this a **Hypothesis Test**
- We establish a *Null Hypothesis* $H_0$ (e.g. the dice are fair),
  - determine a theoretical probability distribution of the random variable (our dice roll means) given that the $H_0$ is true:
    - a normal distribution with $\mu_{\bar{x}} = 3.5$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2.9167}}{\sqrt{n}}$, where n is the number of dice rolls in our sample,
  - and finally we can calculate the probability that you would observe the sample mean you observed given the $H_0$.

# Final steps

- So, let's assume you did 10 dice rolls for this example, and that your mean was 4.
- Since we know that the sample means should be normally distributed, we can transform your mean into a $z$-value:
- Since $\mu_{\bar{x}} = 3.5$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2.9167}}{\sqrt{10}} = 0.5400648$:

$$z(4) = \frac{4 - 3.5}{0.5400648} = 0.9258148$$

- Let's ask Excel what the probability of observing a sample mean this far away (or farther) from the population mean is for the standard normal distribution: `1-NORM.S.DIST(0.9258,TRUE)`
  - Result: 0.177274964

# Final steps (2)

- Fisher suggested that we should consider data with a probability of less than 5% (or .05) given the null hypothesis as **significant** evidence for rejecting the null hypothesis.
- In our case, we are far away from a probability (or short, $p$-value) of .05. So, we can't reject the null hypothesis. Try it for yourselves, though.
- More on this next week.

# Technical note for those who really care

- We really don't care about the direction of the effect here, just the absolute distance from the mean (i.e. this is a *two-tailed* test).
- So, to be absolutely correct, we should ask Excel to give us the probability of z being at least this far away from the mean on either side:

$$p(z < -.9258 \cup z > .9258) = p(z < -.9258) + (1 - p(z < .9258))$$

- When we ask Excel for the p-value `=NORM.S.DIST(-0.9258,TRUE)+(1-NORM.S.DIST(0.9258,TRUE))` we get the actual, correct result of 0.3545499.