Tutorial 8 - Hypothesis Testing

by Tallulah and Nina

Table of content

1. How to test a hypothesis
2. Important concepts
3. Model with binomial distribution
4. Law of large numbers & Central limit theorem
5. χ^2 - test for goodness of fit testing
6. Homework hints

How to test a hypothesis

How to test a hypothesis

Steps you have to do as a researcher:

1. State your **null-hypothesis** H_0 and the **alternative hypothesis** H_1
   a. H_0 is a statement about the population parameter and is the conservative statement
2. Decide **significance level** α
3. Think about your statistical model, develop your test
   a. sample space, probability measure, set of possible outcomes, test statistic
4. Only then get the data
5. Calculate a **p-value**, making the assumption that H_0 is true
6. **Decide** to reject or to not reject hypothesis H_0

Important concepts

Test statistics

Quantity derived from the data, that reduces the data to one value that can be used to perform the hypothesis test. E.g. sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ number of heads for coin toss $\sum_{i=1}^{n} x_i$.

More abstract: quantify, within observed data, behaviour that would distinguish the null from the alternative hypothesis.

Has a probability distribution, which is used to compute p-values for the null hypothesis.

Test statistics - example: coin flip

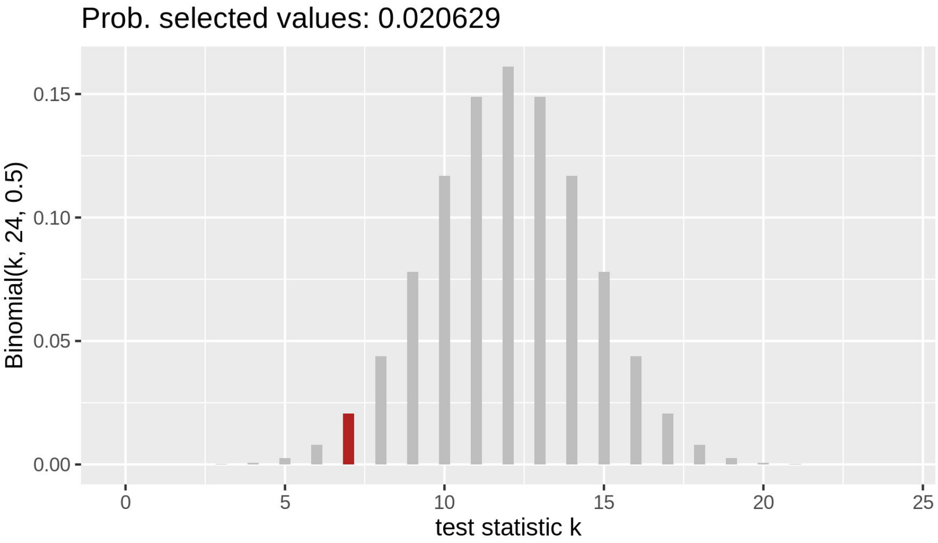
Test whether a coin is fair, flip 24 times. Raw data: c(0, 1, 1, 1,0, ..., 0, 1, 1, 1, 1).

If there is interest in the probability of obtaining a head, only the number $k = \sum_{i=1}^{24} x_i$ out of the 24 flips that were heads needs to be recorded.

What generated k?

$$k \sim Binom(\theta = 0.5, n = 24)$$

Note that this **test statistic** reduces a set of 24 numbers to a single numerical summary.



Sampling distribution

The probability distribution that generates the values that the test statistic can take on is called **sampling distribution.**

In our previous example is the binomial distribution the sampling distribution: $k \sim Binom(\theta = 0.5, n = 24)$
Using this sampling distribution, it is possible to compute a p-value for the null hypothesis that the coin is fair.

To sum up:

NULL HYPOTHESIS → DATA → TEST STATISTIC → SAMPLING DISTRIBUTION → P-VALUE

Model with Binomial Distribution

Example: Coin Toss Experiment

- We assume the distribution for the experiment: **Binomial** -> **discrete probability distribution**, **series of similar and independent events, each of which has exactly two possible outcomes ("success" or "failure") used to model binary data** → is used to model the likelihood function
- We assume the parameter we want to make inferences by: **θ**
- We specify **N = number of trials, p = θ = probability of success for each trial, k = observed successes (“heads”)**

We use this information to obtain the test statistics & sampling distribution!

Set Hypotheses

- **wish to make inferences about: θ = probability of success for each trial**

- A possible research hypothesis for our example would be: **Is the coin fair?** (two-sided)

  H0 = θ = 0.5

  H1 = θ != 0.5 (departs from H0 left and right)

- conceptually, we assume H0 is true for the population
- null hypothesis is assumed to be true until evidence indicates otherwise
- researchers work to reject or disprove the null hypothesis

Example: Coin Toss Experiment

We observe:

coin is flipped **N= 24** times

k=7 times the coin successfully flipped “heads”

p(here θ) is the probability that the outcome will occur at any particular coin toss (**assuming H0, we assume θ = 0.5, is a fair coin**)

Sampling distribution and the test statistic (k) are now specified $k \sim \text{Binomial}(\theta, N).$

Lets test our hypothesis!

Hypotheses testing

- we need to set a null hypothesis, i.e., a **value θ of coin bias θ** that we would like to **collect evidence *against*** → ***H0* = θ = 0.5**

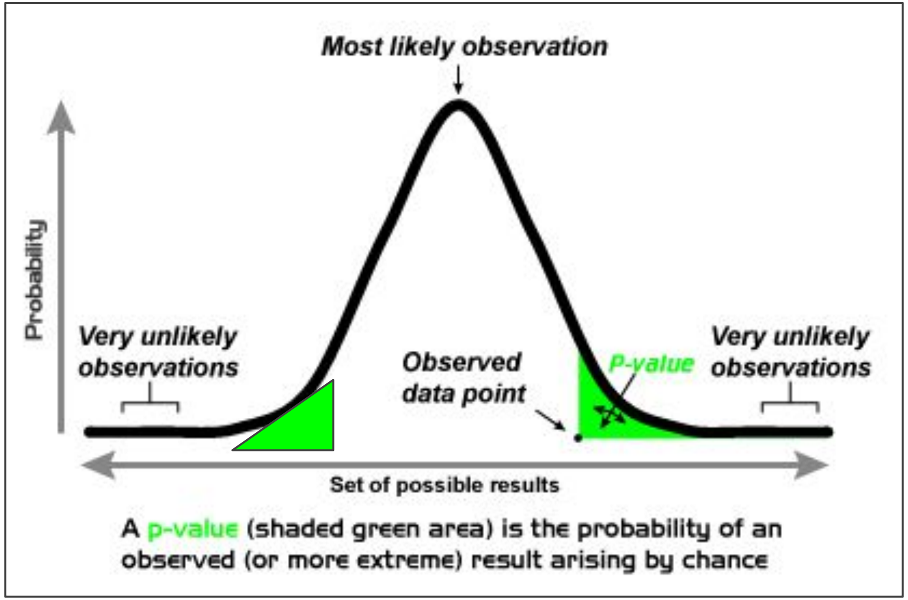
- if empirical observations are sufficiently ***unlikely*** from the point of view of the **null-hypothesis H0,** this should be treated as **evidence *against* the null-hypothesis**

- a measure of **how unlikely** data is in the light of H0 is **the p-value**

- to obtain a p-value: **what is the probability of observing more extreme values (in this case: to both ends) compared to what we sampled (k=7) and therefore count as more extreme evidence against the chosen null hypothesis?**

- which values for k are less or equally as probable compared to our sample of the test statistic k=7?

Area under curve = represents 100% of all possible events



Most likely observation
Probability
Very unlikely observations
Observed data point
P-value
Very unlikely observations
Set of possible results
A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

If we assume H0 to be true, values to the left and right side of the curve become increasingly unlikely

If the p-value falls within a sufficiently unlikely area, this is taken as evidence that H0 cannot be true for a population

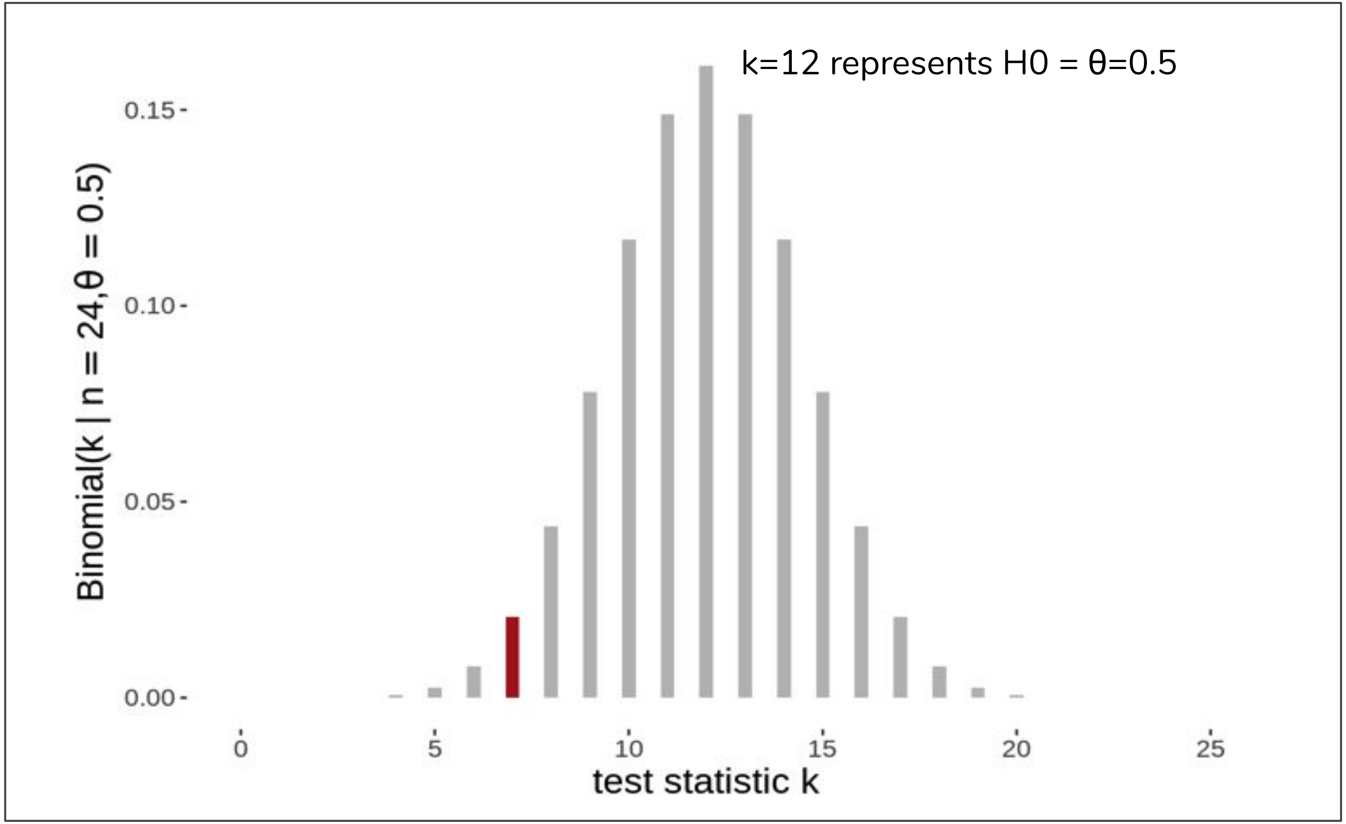
Source: https://en.wikipedia.org/wiki/One-_and_two-tailed_tests

- value that occurs at the **peak (k=12) represent H0= θ=0.5** → typically, H0 states that there is **no effect**

- as **values for k move further away** from the peak, it represents **larger effect sizes (in refuting H0)**

- when **H0 is true** for the population, obtaining samples that exhibit **large effects** becomes **less likely**, which is why the probabilities for k values taper off to the sides of the curve, further from θ=0.5

y-axis shows probability to observe certain measurements



Sampling distribution shows the **probability associated with observed data k=7 highlighted in red**. Displays the probabilities of obtaining test statistic values when the null hypothesis is correct (θ=0.5).

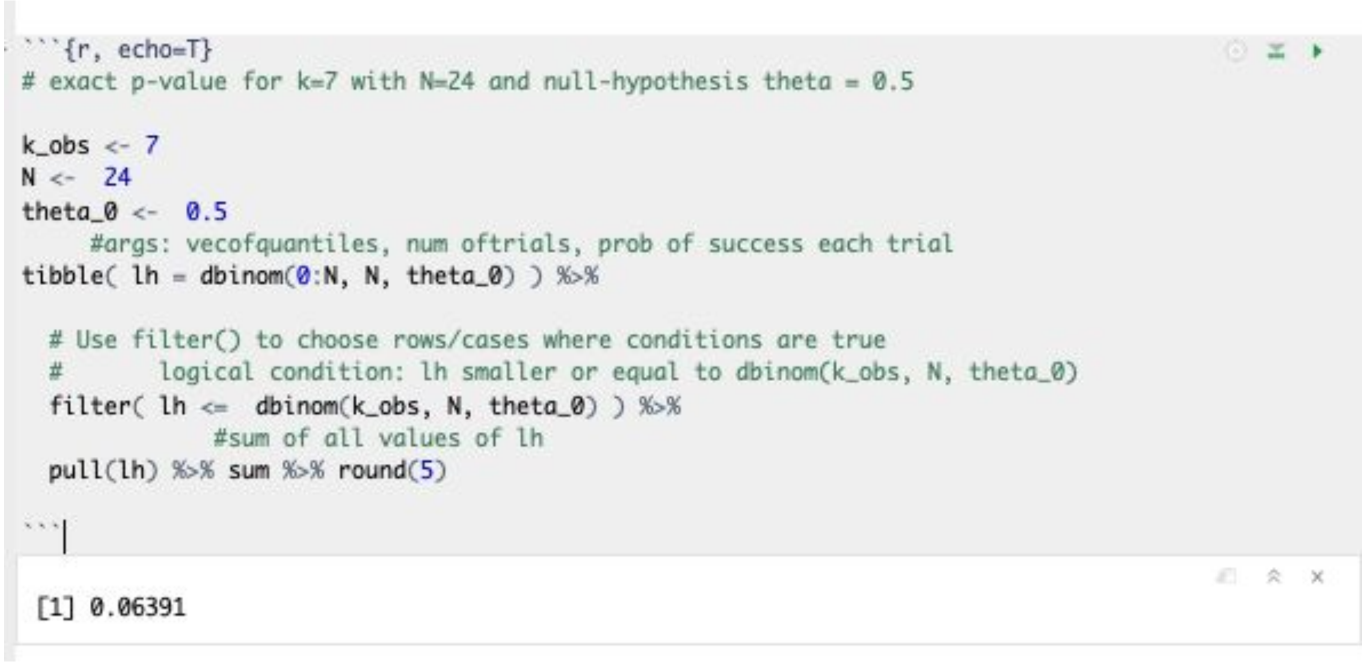
Calculate p-value Two-Tailed

- **To obtain a p-value:**

- s**um up** all **probabilities** for observing values equal to or smaller than empirically sampled test statistic k (=7)

- In other words: **sum over** all possible orders of coin-toss-outcome-values with probabilities equal to or less than probability of observing k=7

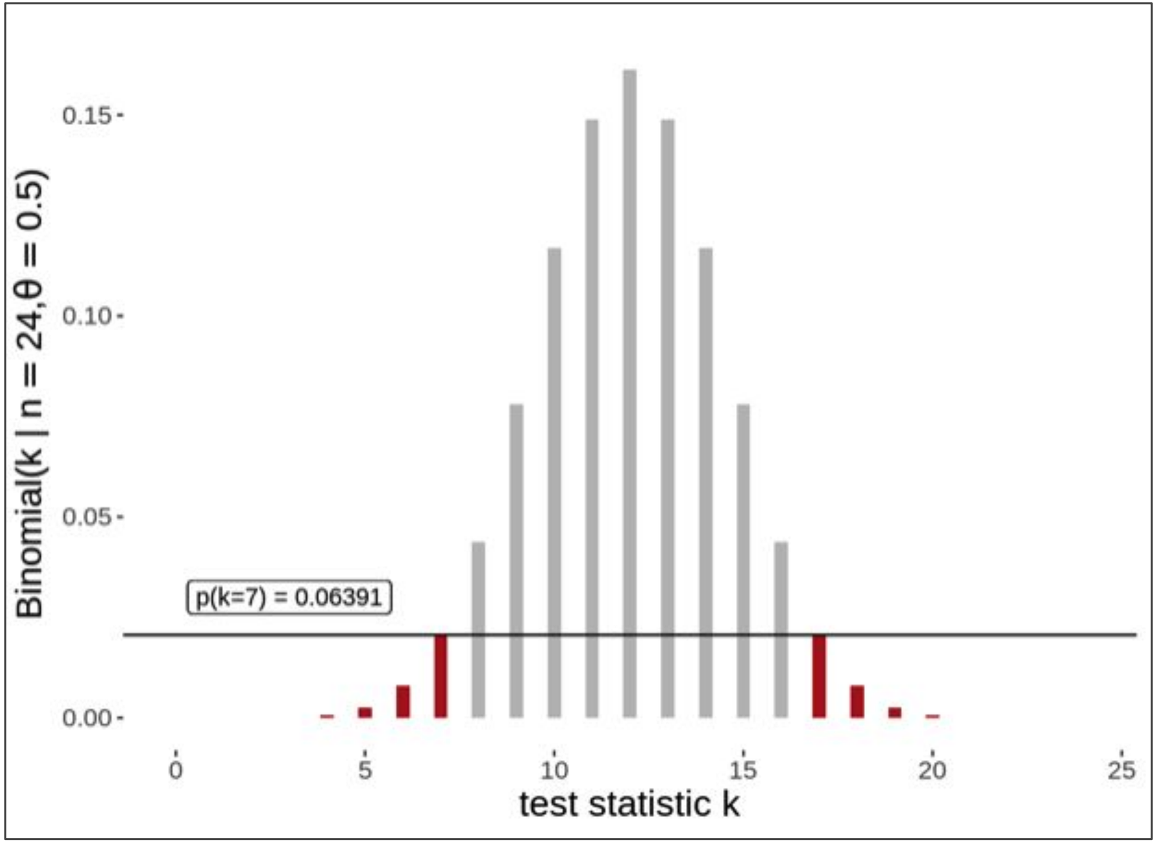
Calculate in R Two-Tailed- handwritten function



~~~
```{r, echo=T}
# exact p-value for k=7 with N=24 and null-hypothesis theta = 0.5

k_obs <- 7
N <-  24
theta_0 <-  0.5
     #args: vecofquantiles, num oftrials, prob of success each trial
tibble( lh = dbinom(0:N, N, theta_0) ) %>%

  # Use filter() to choose rows/cases where conditions are true
  #       logical condition: lh smaller or equal to dbinom(k_obs, N, theta_0)
  filter( lh <=  dbinom(k_obs, N, theta_0) ) %>%
              #sum of all values of lh
  pull(lh) %>% sum %>% round(5)

```
~~~

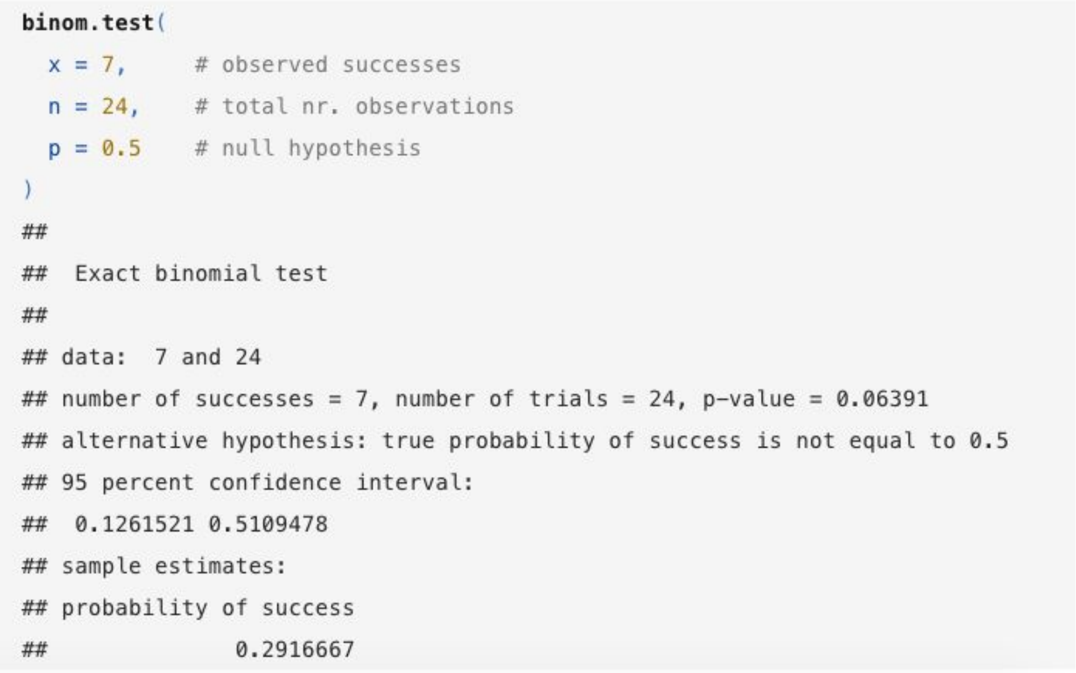
```
[1] 0.06391
```

y-axis shows probability to observe certain measurements



Sampling distribution shows the values that **need to be summed over in red.** p-value for the observation of k=7 successes in N=24 coin flips. Displays the probabilities of obtaining test statistic values when the null hypothesis is correct (θ=0.5).

Calculate in R Two-Tailed- built-in function



```
binom.test(
  x = 7,     # observed successes
  n = 24,    # total nr. observations
  p = 0.5    # null hypothesis
)
##
##  Exact binomial test
##
## data:  7 and 24
## number of successes = 7, number of trials = 24, p-value = 0.06391
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1261521 0.5109478
## sample estimates:
## probability of success
##              0.2916667
```

Calculate p-value One-Tail

Another possible research hypothesis for our experiment: **Is the coin biased towards “heads”?**

H0 = θ > 0.5 H1 = θ =< 0.5

Now we wish to calculate the p-value for data for this model!

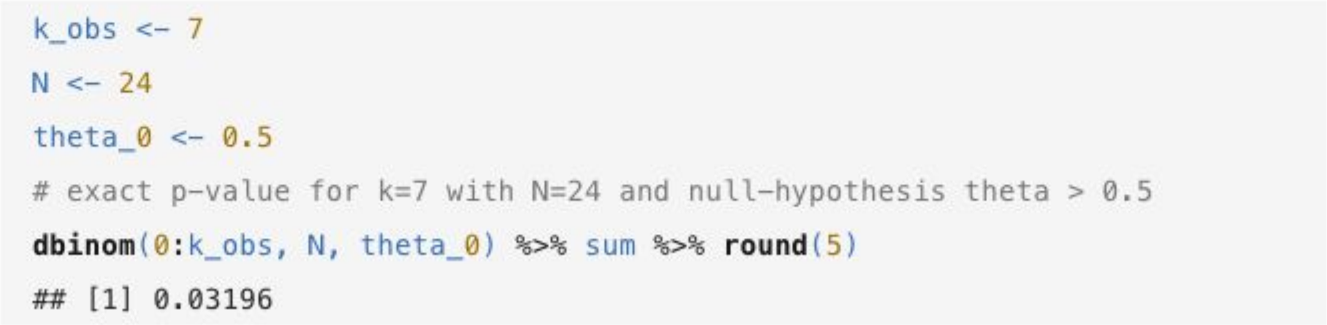
- only seeks to measure effect into one direction from H0 (and from the curve)
- what would count as the most extreme evidence against H0?

- We need to adjust what we consider

Calculate p-value One-Tail

- **to obtain a p-value:** What are the probabilities of observing values less than or equal to test statistic k=7? Sum them!
- values on the **right hand side from θ= 0.5 will not serve as evidence against H0 because all values there are θ>0.5**
- the associated p-value must be calculated using a **one-sided test, only considering values on the left side of the curve**

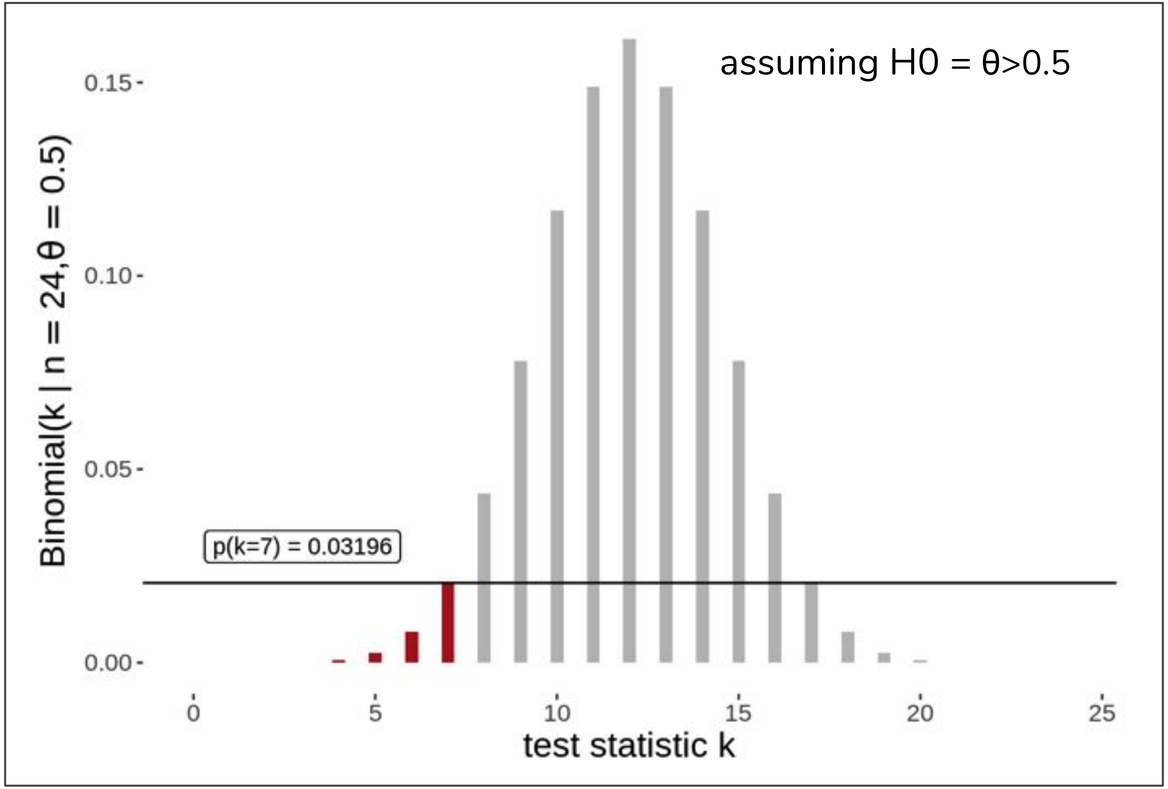
Calculate in R One-Tailed - handwritten function



```
k_obs <- 7
N <- 24
theta_0 <- 0.5
# exact p-value for k=7 with N=24 and null-hypothesis theta > 0.5
dbinom(0:k_obs, N, theta_0) %>% sum %>% round(5)
## [1] 0.03196
```

- doubling the p value for one-tailed test results in *the p-value for two-tailed test for* **symmetrical** *sampling distributions!*

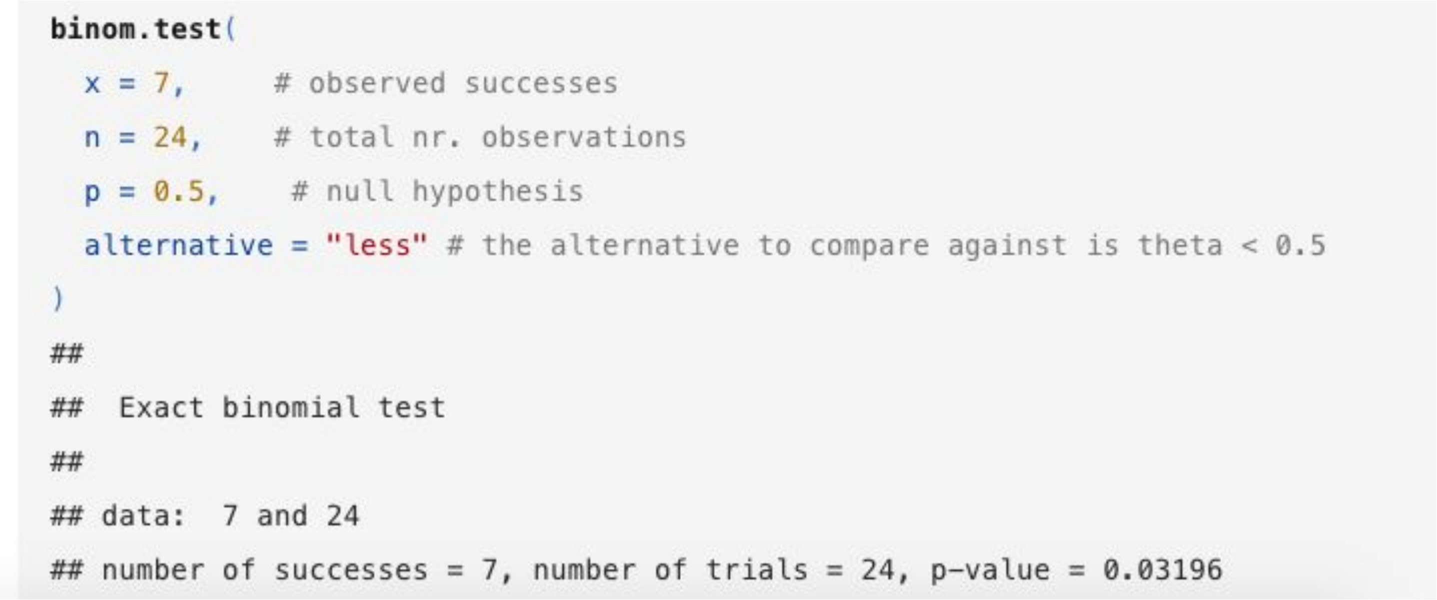
y-axis shows probability to observe certain measurements



Sampling distribution shows p-value for the observation of k=7 successes. Displays the probabilities of obtaining test statistic values when null hypothesis is correct (θ>0.5.).

Sum over all probabilities for observing values smaller than or equal to k only on the left-hand side.

Calculate in R One-Tailed- built-in function



```
binom.test(
  x = 7,     # observed successes
  n = 24,    # total nr. observations
  p = 0.5,    # null hypothesis
  alternative = "less" # the alternative to compare against is theta < 0.5
)
##
##  Exact binomial test
##
## data:  7 and 24
## number of successes = 7, number of trials = 24, p-value = 0.03196
```

Significance of p-values

- **Fisher**: p-values as quantitative measures of strength of **evidence against the null hypothesis:**

  if you get a result that is barely significant, there is chance that you falsely rejected H0!
  Same with results that are barely non-significant, maybe we are falsely keeping H0!

- we say a test result is **significant** if the p-value of the observed data is **lower than a specified α**

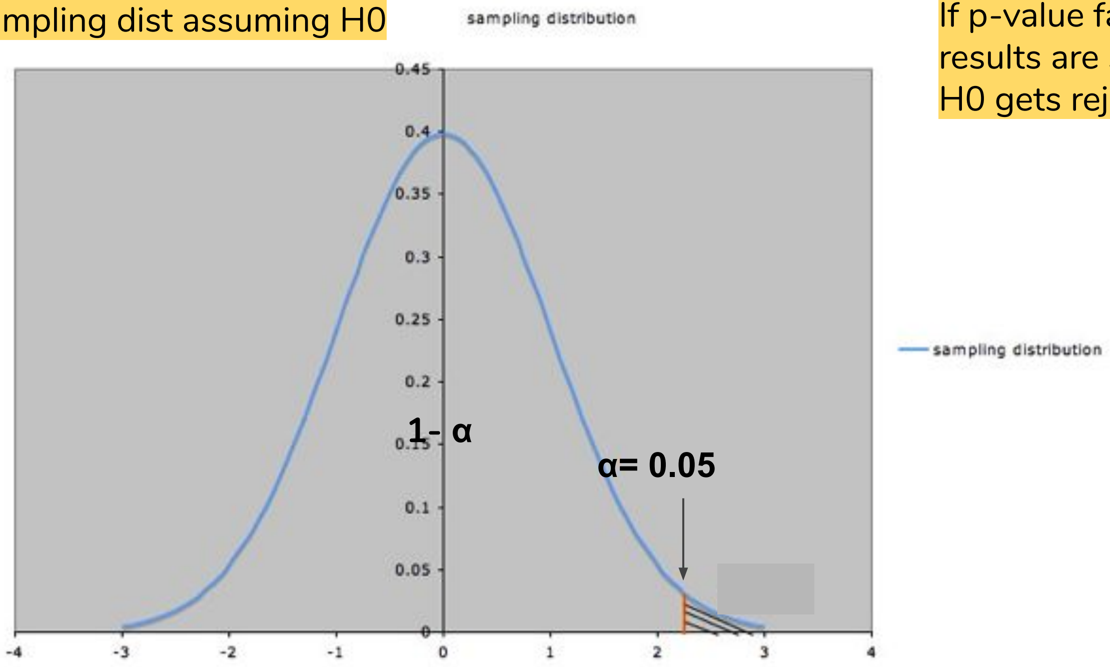
- we fix the **α**-level of significance with common values **α∈{0.05,0.01,0.001}**

- commonly, a **significant test** results is interpreted as the signal to **reject the null hypothesis**, to render it false

- if your **test statistic** falls in **either critical region**, your **sample data are sufficiently incompatible with the null hypothesis and observing this value is sufficiently unlikely in light of the null hypothesis so that you can reject it for the population**!

Significance of p-values

Sampling dist assuming H0

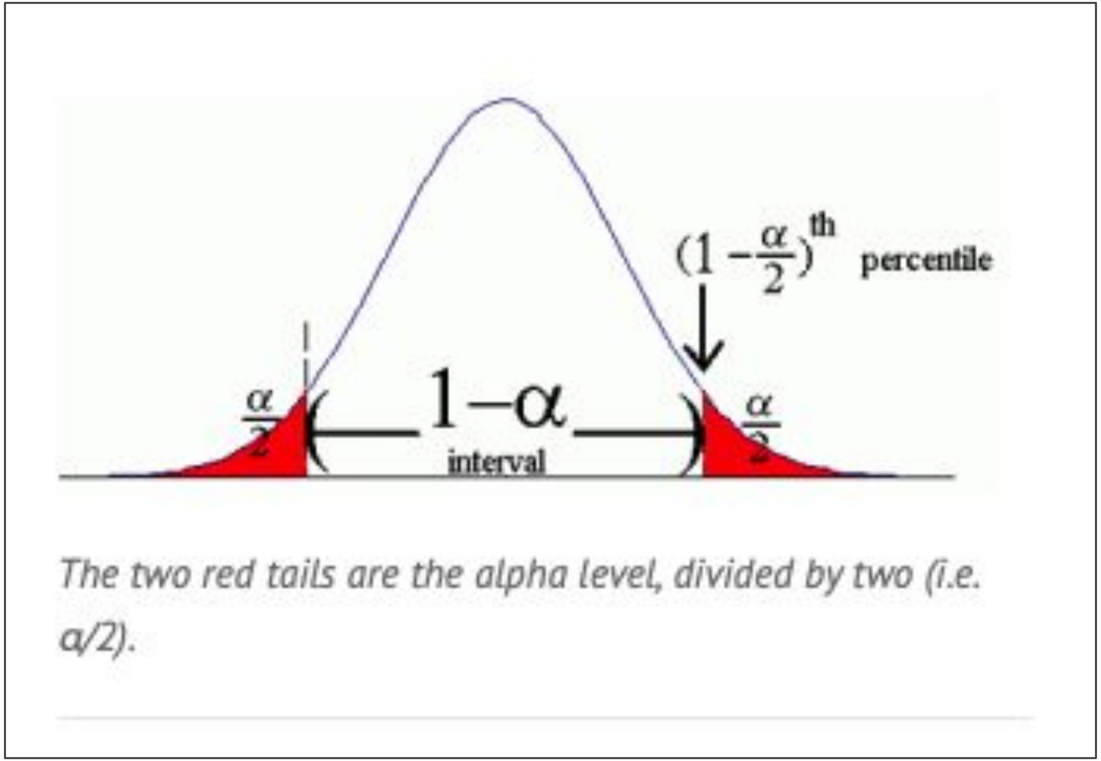


If p-value falls below α, results are significant → H0 gets rejected!

α=0.05 translates to 0.5% chance of observing values for test statistic k in this area when randomly sampling and a 95% chance it will be in the α-1 area under the curve

Source: https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/

1- α represents a confidence intervall = certainty/probability any random sampled k will be amongst these values



The two red tails are the alpha level, divided by two (i.e. α/2).

Source: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/

Errors of decision making

In hypothesis tests, two errors are possible:

- **Type I error: Supporting the alternative hypothesis when the null hypothesis is true (alpha error)**

Example: type I error corresponds to FDA approving a novel drug while it actually has no measured effect. H0= drug has no effect; H1= drug has desired effect

- **Type II error: Not supporting the alternative hypothesis when the alternative hypothesis is true (beta error)**

Example: type II error corresponds to FDA rejecting a novel drug although it has the desired effect. H0= drug has no effect; H1= drug has desired effect

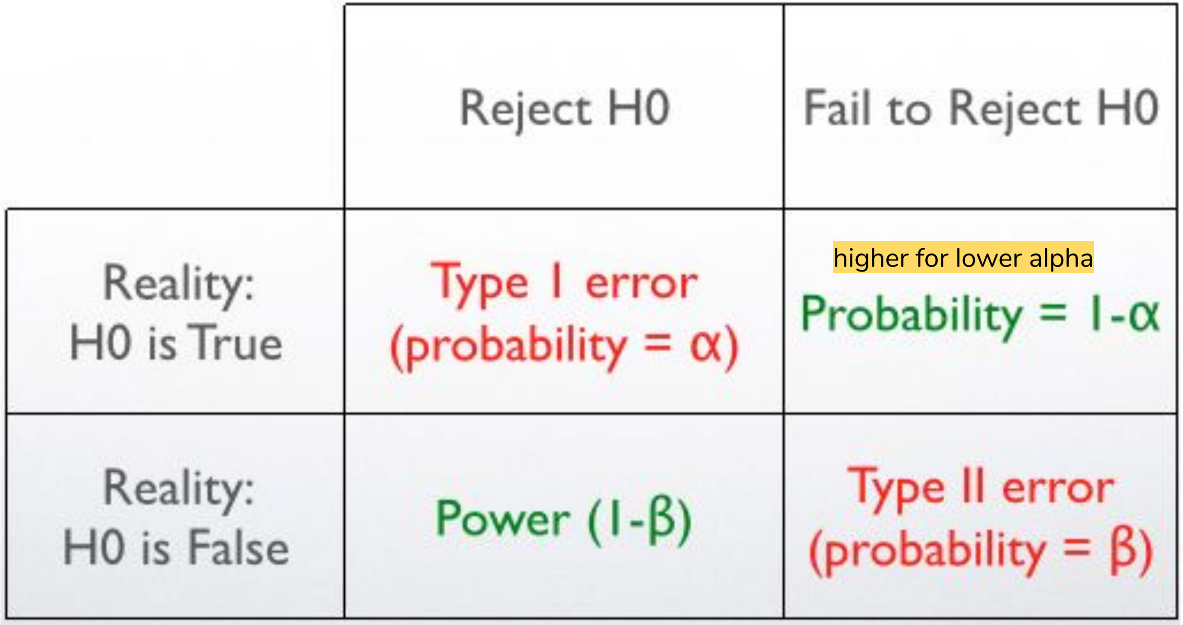
Errors of decision making

try to minimize both errors by determining optimal **level of significance**

Statistical power = probability that an effect will be discovered when an effect actually exists

defined as 1 - β where β is the probability of making a second type of error

If the statistical power is high, the probability of committing a Type II error decreases



	Reject H0	Fail to Reject H0
Reality: H0 is True	Type I error (probability = α)	higher for lower alpha Probability = 1-α
Reality: H0 is False	Power (1-β)	Type II error (probability = β)

Source: https://www.psychologyinaction.org/psychology-in-action-1/2015/03/11/an-illustrative-guide-to-statistical-power-alpha-beta-and-critical-values

Law of large numbers

The law of large numbers

Imagine you gather a lot of samples from random variables $X_1, \ldots, X_n$ that all have the same expected value, i.e. $\mathbb{E}[X_i] = \mu$

→ e.g $X_i \sim \mathcal{N}(10.5, 1)$, where $\mathbb{E}[X_i] = 10.5$ or e.g. $X_i \sim \mathcal{U}(0, 100)$, where $\mathbb{E}[X_i] = 50$

You then want to estimate this expected value based on your samples (because you don't know it yet)

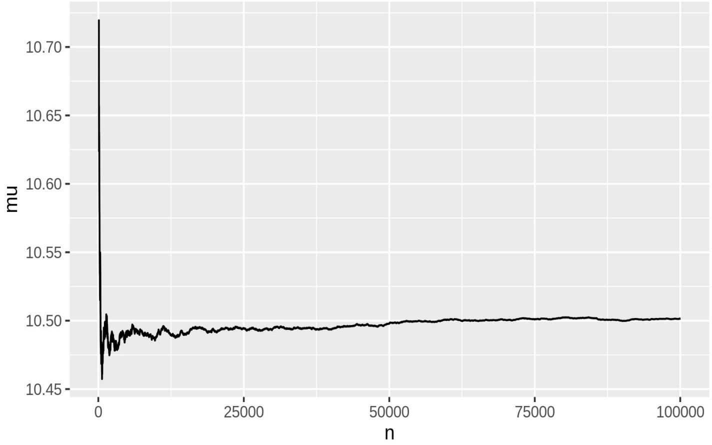
You think the arithmetic mean could be a good estimator of μ: $\hat{T} = \frac{1}{n} \sum_{i=1}^{n} X_i$

The law of large numbers assures that if you gather more and more samples and you compute their arithmetic mean, this arithmetic mean will be (almost surely) the expected value $\mathbb{E}[X_i]$

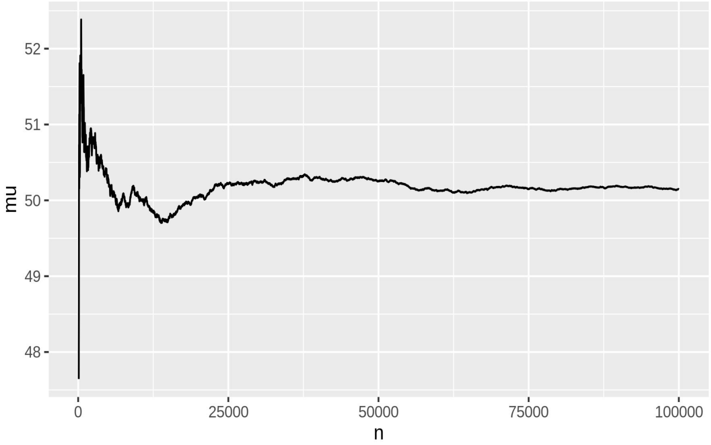
Law of large numbers

Arithmetic mean calculated for samples from

- $X_i \sim \mathcal{N}(10.5, 1)$



- $X_i \sim \mathcal{U}(0, 100)$



Arithmetic means and expected values of the distributions are the same for large n!

The central limit theorem

The Central limit theorem

Imagine you gather samples from random variables $X_1, \ldots, X_n$ that all have the same expected value and the same (finite) variance, i.e. $\mathbb{E}[X_i] = \mu, Var[X_i] = \sigma^2$.

Then for each bunch of samples, e.g. for each 10 samples, you compute their arithmetic mean.

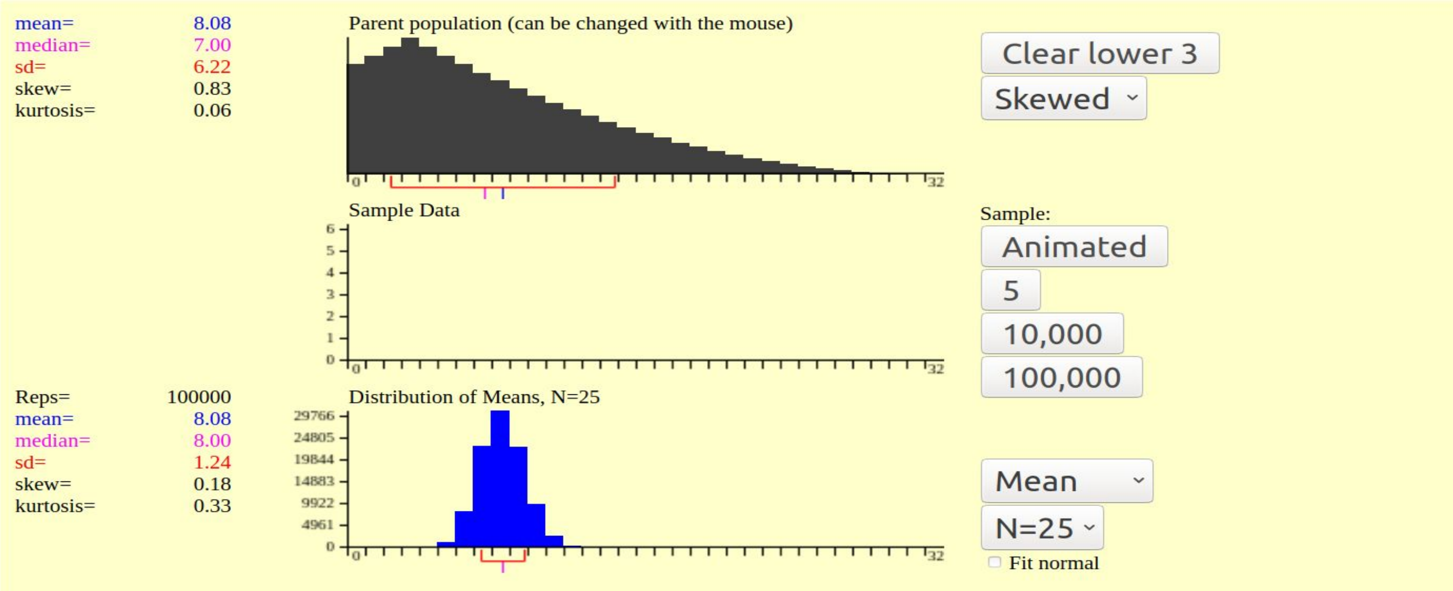
You collect all the arithmetic means (forming a sampling distribution). If you got a sufficient amount of means, then:

whatever the distribution of the random variables, the **sampling distribution** will be the Normal distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ if the overall sample size n is large enough.

If you normalize the sampling distribution, meaning you subtract the mean μ from each random variable and multiply by $\sqrt{n}$, it will be $\mathcal{N}(0, \sigma^2)$- distributed.

Central limit theorem

See for yourself!



```
mean=        8.08
median=      7.00
sd=          6.22
skew=        0.83
kurtosis=    0.06
```

```
Reps=      100000
mean=        8.08
median=      8.00
sd=          1.24
skew=        0.18
kurtosis=    0.33
```

Skewed

Mean

N=25

□ Fit normal

The χ^2- test for goodness of fit

The χ^2 - test for goodness of fit testing

Pearson’s χ2-test for **goodness of fit** tests whether an observed vector of counts is well explained by a given vector of predicted proportion.

“Goodness of fit” is a term used in model checking (a.k.a. model criticism, model validation, …). In such a context, tests for goodness-of-fit investigate whether a model’s predictions are compatible with the observed data.

The χ^2 - test for goodness of fit testing

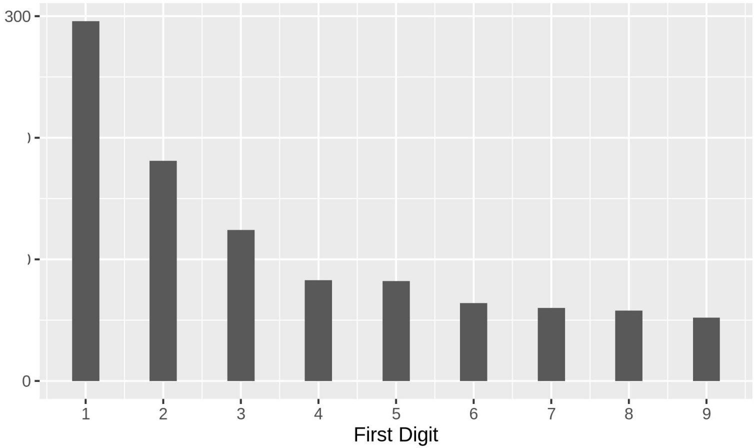
We need:

- Categorical data (each data observation falls into one of several unordered categories)
  - With k categories
- A null hypothesis H_0
- Vector of probabilities $\vec{p} = \langle p_1, \ldots, p_k \rangle$ that correspond to our H_0 and gives the probability with which a single data observation falls into the i-th category.

An example

Raw data: c(**1**43, **1**, **3**4, **1**00, **9**23, **2**3, 42, **8**44, ..., **5**9, **6**6, **7**1, **2**), counting **first digits** leads to:

Categorical Data with 9 categories: $\vec{n} = \langle 296, 181, 124, 83, 82, 64, 60, 58, 52 \rangle$, N = 1000



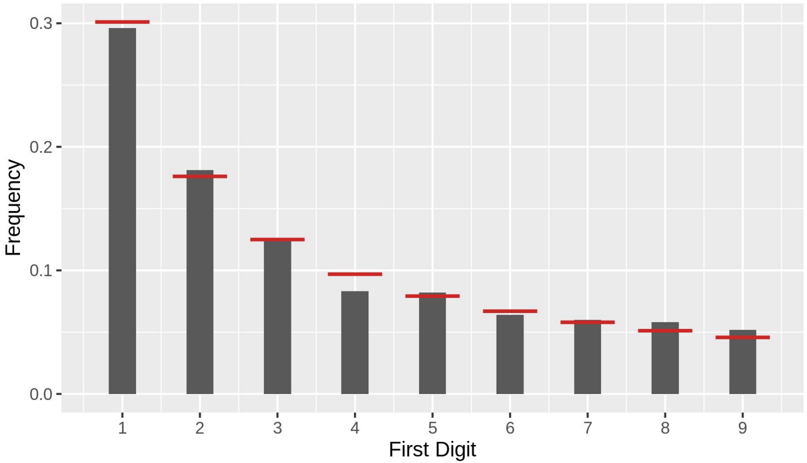
Vector of probabilities: $\vec{p} = \langle 0.301, 0.176, 0.125, 0.0969, 0.0792, 0.0669, 0.0580, 0.0512, 0.0458 \rangle$

(nine probabilities, one for each category, summing up to 1.)

An example

H_0 : the difference from the red bars to the data is not significant.

$\rightarrow \chi^2$ - test for goodness of fit allows us to test whether this data could plausibly have been generated by (a model whose predictions are given by) the prediction vector.



$$\vec{p} = \langle 0.301, 0.176, 0.125, 0.0969, 0.0792, 0.0669, 0.0580, 0.0512, 0.0458 \rangle$$

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)2}{np_i}$$

To get the value of the test statistic for the χ^2 test, insert all the values:

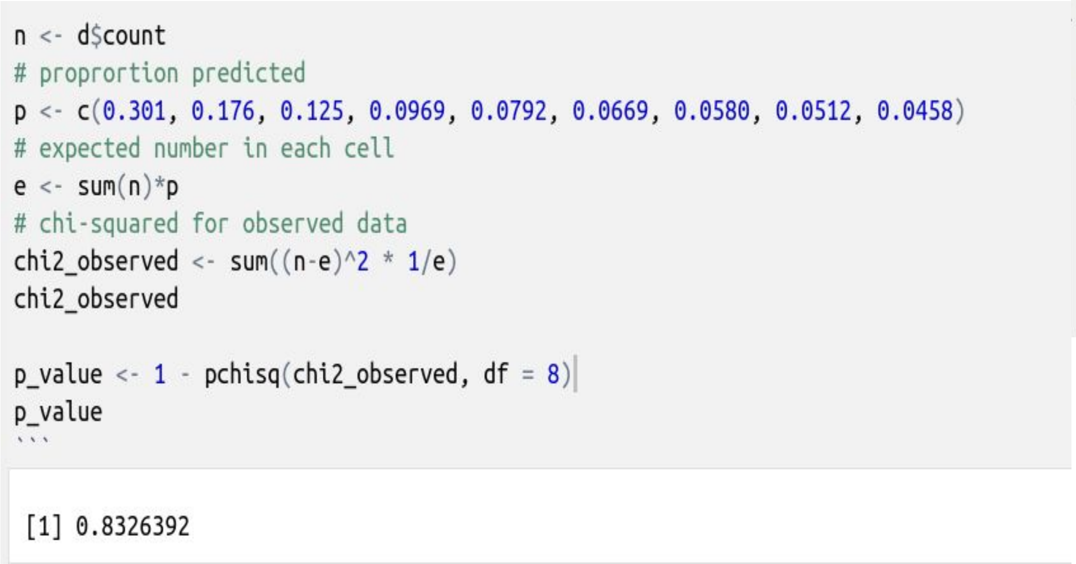
- n_i is the ith entry of vector $\vec{n} = \langle 296, 181, 124, 83, 82, 64, 60, 58, 52 \rangle$
- Nine categories, so k = 9
- p_i is the ith entry of vector $\vec{p} = \langle 0.301, 0.176, 0.125, 0.0969, 0.0792, 0.0669, 0.0580, 0.0512, 0.0458 \rangle$

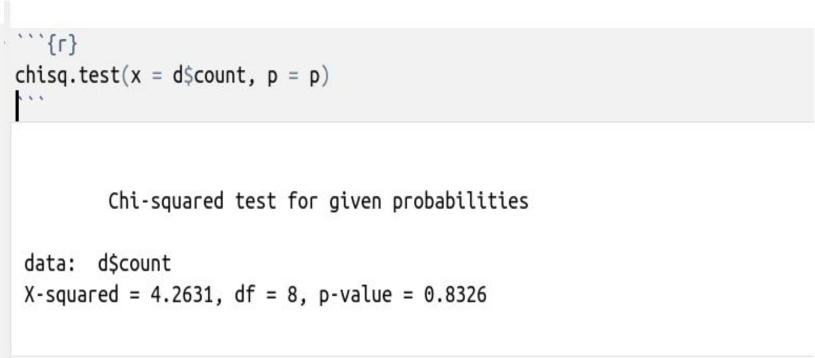
- n = 1000, all our data together

$$\chi^2 = \frac{(296 - 301)^2}{301} + \ldots + \frac{(52 - 45.8)^2}{45.81} = 4.263105$$

In R

Manually:



```
n <- d$count
# proprortion predicted
p <- c(0.301, 0.176, 0.125, 0.0969, 0.0792, 0.0669, 0.0580, 0.0512, 0.0458)
# expected number in each cell
e <- sum(n)*p
# chi-squared for observed data
chi2_observed <- sum((n-e)^2 * 1/e)
chi2_observed

p_value <- 1 - pchisq(chi2_observed, df = 8)
p_value
```
```

```
[1] 0.8326392
```

With in-build test:



````
```{r}
chisq.test(x = d$count, p = p)
```
~~~

```
	Chi-squared test for given probabilities

data:  d$count
X-squared = 4.2631, df = 8, p-value = 0.8326
```
~~~

# Result & Interpretation

The common interpretation of our calculations would be to say that the test yielded no significant result, at least at the significance level of $\alpha=0.05$.

In a research paper we might report this results roughly as follows:

> “The observed counts deviated not significantly from what is expected if each category (here: first digits) followed the specified probabilities ($\chi2$-test, with $\alpha = 0.05$ $\chi2\approx4.2631$, df=8 and $p\approx0.8326$). We therefore conclude that there is no evidence to reject the hypothesis that our data conforms to the specifies probabilities.”

# Homework hints

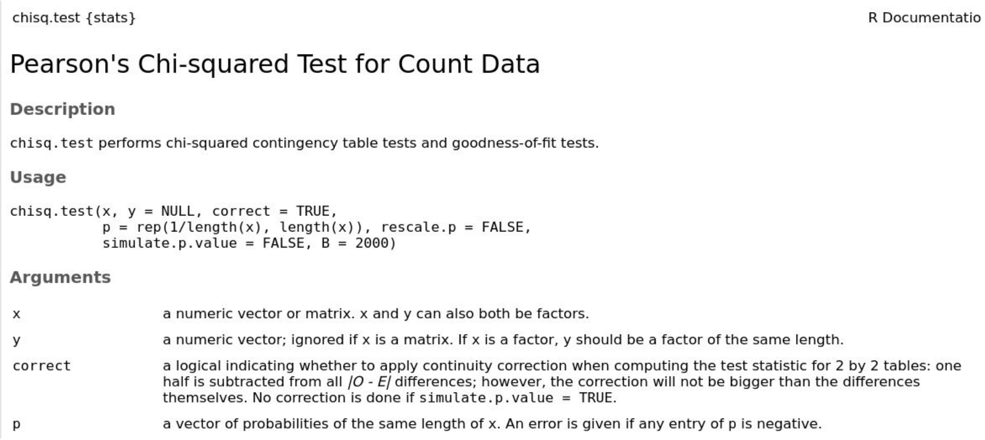
# Homework Hints

## Exercise 1: Addressing hypotheses about coin flips with hypothesis testing

- Similar procedure for all three cases: Think about that a point null hypothesis (e.g. $\theta = 0.5$) results in a two-sided test, and an interval hypothesis (e.g. $\theta \leq 0.5$) in a one-sided test

# Exercise 2: Pearson’s $\chi$2-test of goodness of fit

- Think what the test does, what are your data, your expected probabilities?
- Check how to use the R function:



chisq.test {stats} R Documentatio

## Pearson's Chi-squared Test for Count Data

### Description

`chisq.test` performs chi-squared contingency table tests and goodness-of-fit tests.

### Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

### Arguments

`x` a numeric vector or matrix. x and y can also both be factors.

`y` a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of the same length.

`correct` a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all *|O - E|* differences; however, the correction will not be bigger than the differences themselves. No correction is done if `simulate.p.value = TRUE`.

`p` a vector of probabilities of the same length of x. An error is given if any entry of p is negative.

# Exercise 3: Some claims about frequentist testing

- Frequentist statistics is all about repetitions. It never puts probabilities into parameter values.
- Check the lecture slides/ tutorial slides

# Thank you for listening.

Questions?