Now repeat the same many times!
Step 1: Take 10 black and 10 red cards
Step 2: Shuffle the cards randomly
Step 3: Calculate test statistic

Test statistic

\[ T = 7 \]

red cards in the sample
Step 1: Take 10 black and 10 red cards
Step 2: Shuffle the cards randomly
Step 3: Calculate test statistic

Test statistic

\( T = 4 \)

red cards in the sample
Do this for 100 times
(or even 1 million or more)
Is our hypothesis likely?

• Hypothesis:
  – more red than black in the population

• Data:
  – $T = 8$

• Experiment with hypothesis being false:
  – $T = 5$
  – $T = 7$
  – $T = 4$
  – …
Is our hypothesis likely?

- Hypothesis:
  - more red than black in the population

- Data:
  - $T = 8$

- Experiment with hypothesis being false:
  - $T = 5$
  - $T = 7$
  - $T = 4$
  - ... 

In the experiment the probability of $T = 5$ is approximately:

$$P(T = 5) = 0.3421$$
Is our hypothesis likely?

• Hypothesis:
  – more red than black in the population

• Data:
  – $T = 8$

• Experiment with hypothesis being false:
  – $T = 5$
  – $T = 7$
  – $T = 4$
  – …

In the experiment the probability of $T=8$ is approximately:

$P(T = 8) = 0.0105$
Is our hypothesis likely?

- **Hypothesis:** more red than black
- **Data:**
  - $T = 8$
- **Experiment with hypothesis being false:**
  - $T = 5$
  - $T = 7$
  - $T = 4$
  - ... 

In the experiment the probability of $T = 9$ is approximately:

$$P(T = 9) = 0.0005$$
Is our hypothesis likely?

- **Hypothesis:**
  - more red than black

- **Data:**
  - \( T = 8 \)

- **Experiment with hypothesis being false:**
  - \( T = 5 \)
  - \( T = 7 \)
  - \( T = 4 \)
  - \( \ldots \)

In the experiment the probability of \( T = 10 \) is approximately:

\[
P(T = 10) = 0.00002
\]
Is our hypothesis likely?

• Hypothesis:
  – more red than black in the population

• Data:
  – T = 8

• Experiment with hypothesis being false:
  – T = 5
  – T = 7
  – T = 4
  – …

In the experiment the probability of \( T \geq 8 \) is approximately:
\[
P(T \geq 8) = 0.0110
\]
What did we learn?

• In our dataset:
  – \( T = 8 \)

• In the experiment where the hypothesis is false:
  – \( P(T \geq 8) = P(T=8)+P(T=9)+P(T=10) = 0.0110 \)

• It is very unlikely (about 1.1% probability) that we get so many red cards in the sample by random chance

• We therefore:
  – reject the random chance possibility
  – accept that there are more red than black cards on the table
Congratulations!
Statistical hypothesis test successfully finished!
Sample vs population

Example task with red and black cards

- **Statistical terminology**
- Permutation test and hypergeometric test
- Histogram on a sample vs population
- More statistical terminology
**Statistical terminology**

- **Alternative hypothesis:**
  - This is what we want to claim
    - There are more red than black cards on the table

- **Null hypothesis:**
  - Contradictory hypothesis to what we want to claim
    - There are equal number of red and black cards

- **Test statistic:**
  - A value calculated from sample
  - Higher (or lower) values support our claim more
    - The number of red cards in the sample
More statistical terminology

• Null distribution:
  – Distribution of test statistic under null hypothesis
    • Distribution of number of red cards in the sample after shuffling

• P-value:
  – The probability of test statistic in the null distribution equal or higher than the observation
    • $p = 0.0110$

• Significance level:
  – Threshold $\alpha$ below which null hypothesis is rejected, usually $\alpha = 0.05$
    • We had $p < 0.05$, so we can reject null hypothesis
How to report this finding?

- There are **significantly** more red than black cards on the table according to the permutation test under significance level $\alpha=0.05$

- There are **significantly** more red than black cards on the table (permutation test, $p=0.011$)

- The word ‘**significantly**’ in the scientific literature implies that a statistical hypothesis test was conducted and resulted with $p<\alpha$
What statistical test was it?

• **Permutation test** (also called randomization test)
  - The simplest possible statistical test
  - Not requiring complex mathematics or statistics
  - Need to think of how to mix up the data, so that the hypothesis would not hold, at the same time keeping data as unchanged as possible

• Permutation test is a special case of **Resampling tests**
  - Instead of mixing up some other method is used to generate synthetic data
Do I need to know more statistical tests?

• Yes and No

• No
  – Because almost any statistical hypothesis can be tested with resampling tests

• Yes
  – Because resampling tests can involve too many computations
  – A lot of statistical theory is there to simplify computations
Can we simplify our permutation test with the help of statistical theory?
Lecture 10 – Computational statistics

✓ Sample vs population
✓ Example task with red and black cards
✓ Statistical terminology
  • Permutation test and hypergeometric test
  • Histogram on a sample vs population
  • More statistical terminology
Our experimental design behind the null hypothesis

• Many iterations of:
  – Step 1: Take 10 black and 10 red cards
  – Step 2: Shuffle the cards randomly
  – Step 3: Calculate test statistic $T$

• Draw the histogram of $T$

• Calculate p-value

• Is it possible to know in advance what the histogram will look like?
Our experimental design restated mathematically

• Take 10 cards randomly from 10 black and 10 red cards

• What is the probability of getting $T$ red cards?

• Statisticians know the answer:
  – **Hypergeometric distribution!**
  – Python command:
    • `hypergeom.pmf(8, 20, 10, 10)`
      – 0.0109604 (our experiment gave $P(T=8) = 0.0105$)
    • `hypergeom.pmf(9, 20, 10, 10)`
      – 0.0005412544 (our experiment gave $P(T=9) = 0.0005$)
Calculating the p-value

• **P-value:**
  
  – Our estimate:
    
    • P(T ≥ 8)
      
      – 0.0110
  
  – **Python command:**
    
    • `sum([hypergeom.pmf(x,20,10,10) for x in [8,9,10]])`
      
      – 0.01150707
    
    • `1-hypergeom.cdf(7,20,10,10)`
      
      – 0.01150707
Summary so far

• **Aim:**
  – Test if there are significantly more red than black cards

• **Method 1: Permutation test**
  – Design a permutation scheme which generates the null distribution with equal number of red and black cards
  – Apply many times, draw the histogram, estimate the p-value as the tail of the histogram

• **Method 2: Hypergeometric test**
  – Calculate p-value directly

• **Conclusion:**
  – If you know an appropriate statistical method then use it
  – If not, then design a resampling test
“There are more red cards than black cards” is a

A. Alternative hypothesis
B. Null hypothesis
C. Not sure
Null distribution is a distribution of … under the null hypothesis

A. The population
B. The null
C. The test statistic ✓
D. The sample size
A statistical test shows significance, if ...

A. \( P < \alpha \)  
B. \( P = \alpha \)  
C. \( P > \alpha \)
Lecture 10 – Computational statistics

✓ Sample vs population
✓ Example task with red and black cards
✓ Statistical terminology
✓ Permutation test and hypergeometric test

• Histogram on a sample vs population
• More statistical terminology
Histogram on a sample vs population

- Histograms on a sample and on the population are different!
- Suppose we have two fair dice
- Consider the histogram of the sum of these dice

Population histogram

Histogram on a sample with size 100
Histotgram on a sample vs population

- Histograms on a sample and on the population are different!
- Suppose we have two fair dice
- Consider the histogram of the sum of these dice

Population histogram

Mode is different (7 vs 6)!

Histogram on a sample with size 100
How big are the differences?

• How well does the sample histogram approximate the population histogram?

• This is easy to study experimentally:
  – Let us fix sample size $N=100$
  – Generate many samples of this size
  – Draw histogram of each such sample
Sample 1
Sample 2
Sample 3

![Bar graph with relative frequency on the y-axis and total counts on the x-axis. The highest bar corresponds to a total of 7.]

- **Relative frequency**
  - 0.15
  - 0.10
  - 0.05
  - 0.00

- **Total**
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10
  - 11
  - 12
Repeat the same many times
Scatter plot of 10 iterations
(with 100 throws of a pair of dice in each iteration)
Box plot of 1000 iterations
(with 100 throws of a pair of dice in each iteration)
Box plot of 1000 iterations
(with 100 throws of a pair of dice in each iteration)

Notice that the boxes are larger in the center
Violin plot of 1000 iterations
(with 100 throws of a pair of dice in each iteration)
Violin plot of 1000 iterations
(with 100 throws of a pair of dice in each iteration)

Notice that the distributions look a bit like normal distributions
Notice that the distributions look a bit like normal distributions
What can we do with these distributions?

• These distributions quantify how much the histogram bars vary across samples

• E.g., for the histogram bar corresponding to dice total sum 9:
  – Sample mean in 1000 iterations was 0.11188
  – Sample standard deviation was 0.03271

• For comparison, population mean is:
  – Four combinations out of 6x6 give total 9:
    • 3+6, 4+5, 5+4, 6+3
  – Population mean = 4/36 = 1/9 = 0.11111
For the dice total sum 9

• In the relative frequency scale:
  – Sample mean: 0.11188
  – Sample standard deviation: 0.03271293

• In the frequency scale (sample size = 100):
  – Sample mean 11.188
  – Sample standard deviation: 3.271293
  – Sample variance: 10.70136

• Statistical theory says:
  – In histogram bar estimation sample variance approximately equals sample mean
  – That is, standard deviation is approximately square root of mean
What have we learned?

- Sample histogram is an approximation of the population histogram
- We know how big errors to expect:
  - In the frequency (count) scale:
    - If the bar height is $M$
    - Then standard deviation estimate is $\sqrt{M}$
  - In the relative frequency scale (sample size $N$):
    - If the bar height is $p$
    - Then standard deviation estimate is about $\sqrt{p/N}$
      - More precise would be $\sqrt{p(1-p)/N}$
- We can add error bars to sample histograms
Sample 3 (again)
Sample 3
(now with error bars of 1 standard deviation)
Sample 3 (comparison to population)
Meaning of whiskers can vary

• Lower and upper quartile
  – 50% of data within whiskers

• Mean ± one standard deviation
  – If normally distributed then 68% within whiskers

• Confidence intervals at level $\alpha=0.05$
  – 95% of data within whiskers

• It is good practice to explicitly write what the whiskers mean in the text accompanying the figure
A sample histogram bar with frequency count 100 has standard error of about...

A. 100
B. 10 ✔
C. 1
News at Estonian media

• September 20, 2018:
  – [Link](https://www.err.ee/862501/uuring-hetkeseisuga-paaseks-riigikogusse-vaid-neli-parendid)

  – The article claims:
    • "Support for Party X went down from 5% to 3%"
    • "Support for Party Y went up from 11% to 14%"
    • "Support for Party Z went down from 19% to 17%"

  – Finally the article states:
    • "Results are based on polling 1000 people"
    • "Maximal error is 3.10%"

You have now all the tools to investigate the potential errors yourself!
Lecture 10 – Computational statistics

✓ Sample vs population
✓ Example task with red and black cards
✓ Statistical terminology
✓ Permutation test and hypergeometric test
✓ Histogram on a sample vs population

• More statistical terminology
Statistical estimation

• Approximate a numerical property of the population using calculations on the sample

• For example:
  – Population mean is usually estimated using sample mean

• Unbiased estimator:
  – Estimator which is in expectation correct if averaged over many samples

• For example:
  – Sample histogram bar height is an unbiased estimator of population histogram bar height
Example of a biased estimator

• Population variance:

\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2 \]

• Sample variance

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2 \right] \]

• Here division is by \( n-1 \) rather than \( n \), because a naïve estimator dividing by \( n \) is biased:
   – On average it under-estimates the variance

• Standard deviation – take the square root
Multiple testing problem

• If you run 100 independent statistical hypothesis tests where actually the null hypothesis holds

• Then on average about 5 of these tests give p-value below 0.05

• **Multiple testing problem** – when running many tests you can get some low p-values by random chance

• Solution:
  – Need to use a lower p-value threshold
  – Equivalently, use **multiple testing correction** (transform p-values to correct for multiple testing)
Bonferroni correction

- **Bonferroni correction** multiplies all p-values by the number of conducted tests

- **Example:**
  - Conducting 100 hypothesis tests
  - 5 lowest p-values are
    - 0.002, 0.005, 0.007, 0.013, 0.023
  - After Bonferroni correction these become
    - 0.2, 0.5, 0.7, 1.3, 2.3
  - None of these is significant with level 0.05
  - Thus, we accept the null hypotheses for all tests
Effect size vs significance

• In this lecture we have investigated the question of significance but not the question of effect size

• **Effect size** – how strongly the null hypothesis is violated
  – For example, **effect size** is about **how many more red cards** there actually are on the table
  – For comparison, **significance** is about **whether there are more red cards** than black on the table
A. ... is the probability that the null hypothesis is true
B. ... the probability that the observed effects were produced by random chance alone
C. ... indicates the size of the observed effect
D. None of the above

P-value ...
P-value

- **Correct definition:**
  - P-value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

- **In our example:** \( P( T \geq 8 \mid \#\text{Reds} = \#\text{Blacks} ) \)

- **Wrong claims:**
  - P-value is the probability that the null hypothesis is true
  - P-value is the probability that the observed effects were produced by random chance alone

- **In our example this would be:** \( P( \#\text{Reds} > \#\text{Blacks} \mid T = 8) \)

- **These claims are wrong, because:**
  - p-value is a probability conditioned on the assumption that the null hypothesis is true (i.e. observed effects were produced by random chance alone)

- **Frequentist statistics does not attach probabilities to hypotheses**
Frequentist and Bayesian statistics

• Frequentist statistics
  – Statistical hypothesis testing is an example

• Bayesian statistics
  – Prior beliefs are modelled as probability distributions
  – Data used to update prior beliefs into posterior beliefs

• Further comparison not covered in this course
Lecture 10 – Computational statistics

✓ Sample vs population
✓ Example task with red and black cards
✓ Statistical terminology
✓ Permutation test and hypergeometric test
✓ Histogram on a sample vs population
✓ More statistical terminology
Quotes

• “There are lies, damned lies and statistics.”
  – Mark Twain

• “There are two kinds of statistics, the kind you look up and the kind you make up.”
  – Rex Stout

• “Definition of Statistics: The science of producing unreliable facts from reliable figures.”
  – Evan Esar

• Source: https://www.brainyquote.com