Probability and statistics
ESWC summer school 2016

Jan Rupnik

Jožef Stefan Institute

Outline

- Basics of probability
  - Definition
  - Laws
  - Random variables
- Statistical inference
  - Estimation
  - Hypothesis testing

Basics of probability

- Definition
- Laws
- Random variables
  - Distributions
  - Discrete variables
  - Continuous variables
  - Expected value, variance
  - Joint distributions
  - Independence
  - Combinations
  - Sampling

Probability: defintion

- The **probability** of an event refers to the likelihood that the event will occur
- If an experiment has n outcomes that are **equally likely** and a subset of r outcomes are classified as successful, then the probability of a successful outcome is $\frac{r}{n}$
- Example: urn with 3 red and 2 white balls, $Pr(\text{pick red}) = \frac{3}{5}$

Probability: defintion

- The **relative frequency** of an event is the number of times an event occurs, divided by the total number of trials. Probability can be seen as a long-term relative frequencies (number of trials goes to infinity)
- Example: coin toss with two events: H, T. $Pr(H) \ = \frac{\#H \text{ in } n \text{ experiments}}{n}$
- Bayesian interpretations (belief)

Probability: notation

- $Pr(A \cap B)$ – probability of A and B both occurring (**intersection**)
- $Pr(A')$ – probability of A NOT occurring (**complement**)
- $Pr(A|B)$ – probability of A occurring given that B occurred (**conditional**)
- $Pr(A \cup B)$ – probability of A or B occurring (**union**)
- $Pr(A \cap B) = 0$ – events are mutually exclusive (**disjoint**)

Probability: notation

- Example – 6 sided dice, events: $E_1, E_2, E_3, E_4, E_5, E_6$:
  - $Pr(E_3 \cap E_1) = 0$
  - $Pr(E_3 | E_{>2}) = ¼$
  - $Pr(E_3 \cup E_1) = 1/3$
  - $Pr(E_4') = 5/6$

Probability: laws

- $Pr(A) \in [0, 1]$
- $Pr(A) = 1 - Pr(A')$
- $Pr(A \cap B) = Pr(A)Pr(B|A)$
  - If $Pr(A \cap B) = Pr(A)Pr(B)$ we say that events are **independent**
  - If $Pr(A \cap B|C) = Pr(A|C)Pr(B|C)$ we say that events are **conditionally independent**

Probability: laws

- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$
- $Pr(\cup_i A_i) \leq \sum_i Pr(A_i)$, where $A_1, A_2 \ldots$ is a countable set (**Boole's inequality**)
- If $B_1, B_2, \ldots$ are mutually disjoint, whose union is the entire space, then: $Pr(A) = \sum_n Pr(A \cap B_n)$ (**total probability**)

Probability: random variables

- Maps from events to real numbers
- Example:
  - events represent tossing a fair coin n times (2^n mutually exclusive equally probable events)
  - $X(e)$ = #heads obtained in the event e
  - $X(e) = 100$ if all n flips of e result in heads and 0 instead
- When the value of a variable is determined by a chance event, that variable is called a **random variable**

Probability: random variables

- **Discrete** random variables map to a countable set
  - total of roll of two dices: $2,3, \ldots, 12$
  - customer count: $0,1,2, \ldots$
- **Continuous** random variables map to an uncountable set of numbers
  - Task completion time (nonnegative)
  - Price of a stock (nonnegative)
  - Stock price move

Probability: distributions

- Probability distribution specifies the probability for a random variable to assume a particular value
  - $X\colon event \rightarrow \mathbb{R}$ - random variable
  - $\Pr\colon event \rightarrow [0,1]$ - probability
- For discrete variables $P(x) \equiv Pr(X = x)$ is called probability mass function (**pmf**)
- For continuous variables $f_X(x)$ is called probability density function (**pdf**) such that $\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx$
- Cumulative density function (**cdf**) is defined as:

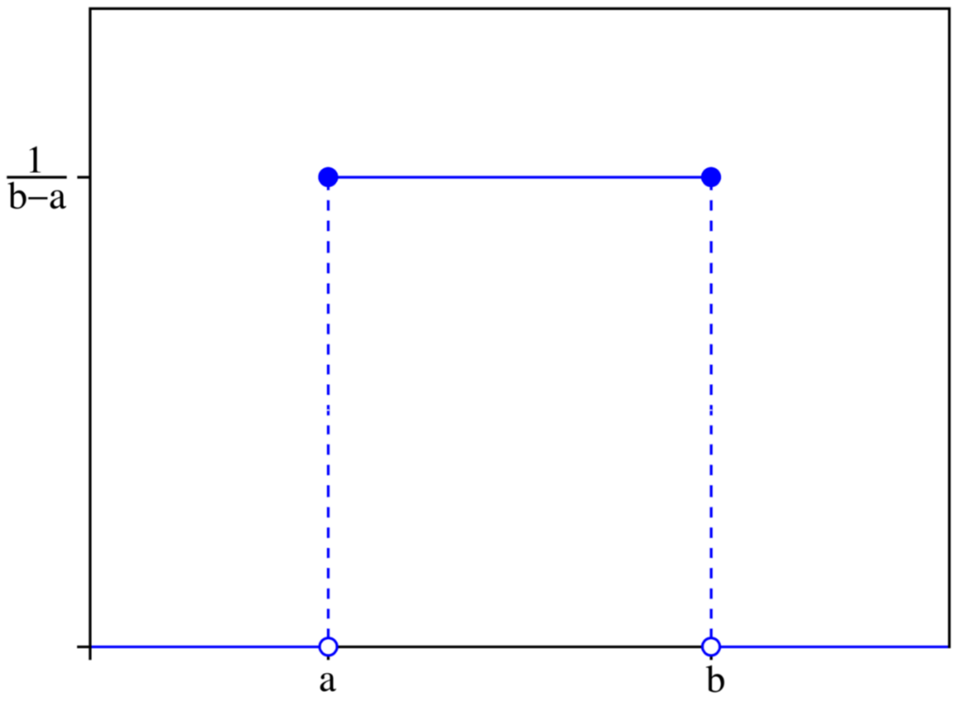
$$F_X(b) = \Pr(X \leq b) = \int_{\infty}^{b} f_X(x)dx$$

Probability: discrete distributions

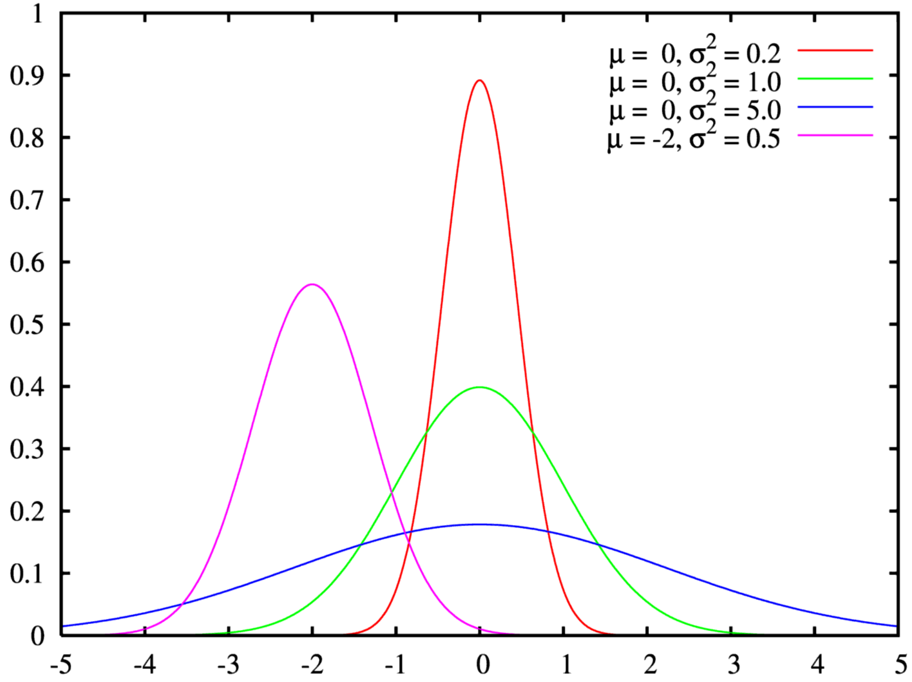
- Example:
  - Bernoulli: X(H) = 1, X(T) = 0. P(X = 1) = p, P(X = 0) = 1-p
  - Multinomial (example: unfair dice)
  - events represent tossing a fair coin n times (2^n mutually exclusive equally probable events)
  - $X(e)$ = #heads obtained in the event e
    - $P(X = k) = \frac{\binom{n}{k}}{2^n}$
  - $X(e) = 100$ if all n flips of e result in heads and 0 instead
    - $P(X = k) = \begin{cases} \frac{1}{2^n};\ \text{if } k = 100 \\ 1 - \frac{1}{2^n};\ \text{if k} = 0 \\ 0;\ \text{else} \end{cases}$

Probability: continuous distributions

- $U[a, b]$

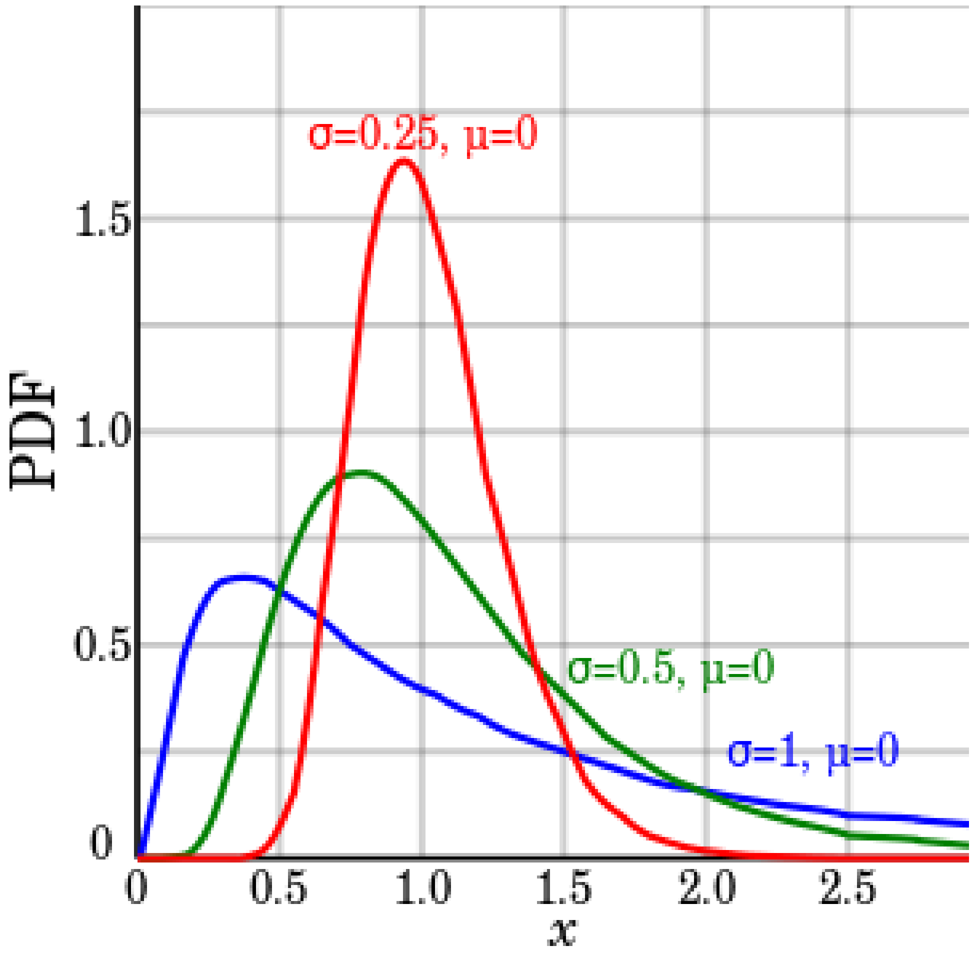


- $\mathcal{N}(\mu, \sigma)$



Lognormal

- If $X \sim N(\mu, \sigma)$ then Y = ln(X) has a **lognormal distribution**
- Notation: $lnN(\mu, \sigma)$



Probability: expected value, variance

- Is there a “typical” value (location)? How spread is the distribution - is the distribution spikey or flat (spread)? The answers summarize the shape of the distribution.
- Sometimes the distributions are completely defined by a few parameters (summaries)
- Expected value $E(X)$ and variance $Var(X)$ are two very important location and spread measures of distributions
- Standard deviation: $Std(X) \ = \sqrt{Var(X)}$

Probability: expected value, variance

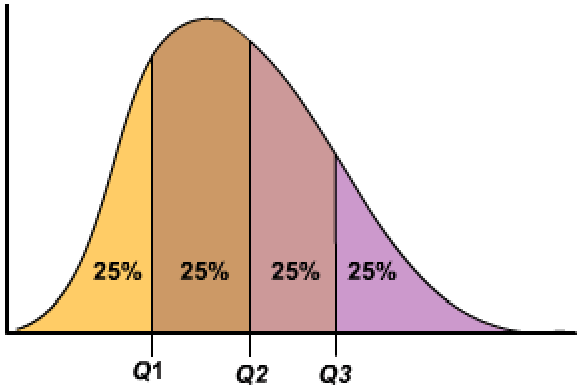
- Discrete distribution
  - $E(X) \ = \sum_i x_i \cdot P(x_i)$
  - $Var(X) \ = \sum_i \left(x_i - E(X)\right)^2 \cdot P(x_i)$

Probability: expected value, variance

- Continuous distribution
  - $E(X) \ = \ \int x \cdot f_X(x)dx$
  - $Var(X) = \int \left(x - E(X)\right)^2 \cdot f_X(x)dx$
- Examples
- $X \sim N(\mu, \sigma)$
  - $E(X) \ = \mu$
  - $Var(X) = \sigma^2$
- $X \sim lnN(\mu, \sigma)$
  - $E(X) = \ e^{\mu + \frac{\sigma^2}{2}}$
  - $Var(X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$

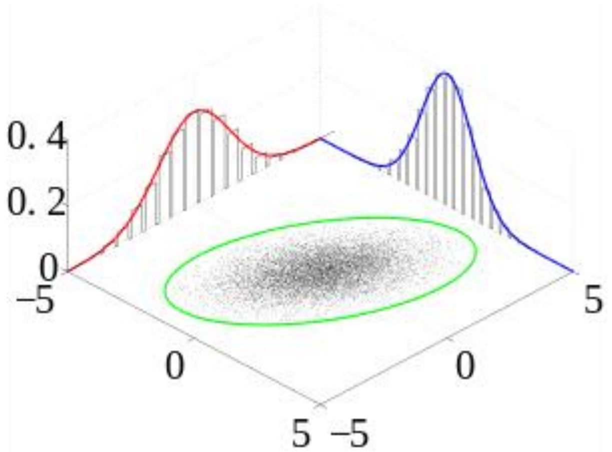
Probability: quantiles

- The median of a random variable X is a value m, so that $\Pr(X \geq m) = 0.5$
- The k-th q-quantile is defined similarly as a number x so that $\Pr(X < x) \leq \frac{k}{q}$
- Quartiles (k=4)



Probability: independence

- If X and Y are random variables we can define a **multivariate random** variable (X, Y) that maps an event e to $\left(X(e), Y(e)\right)$

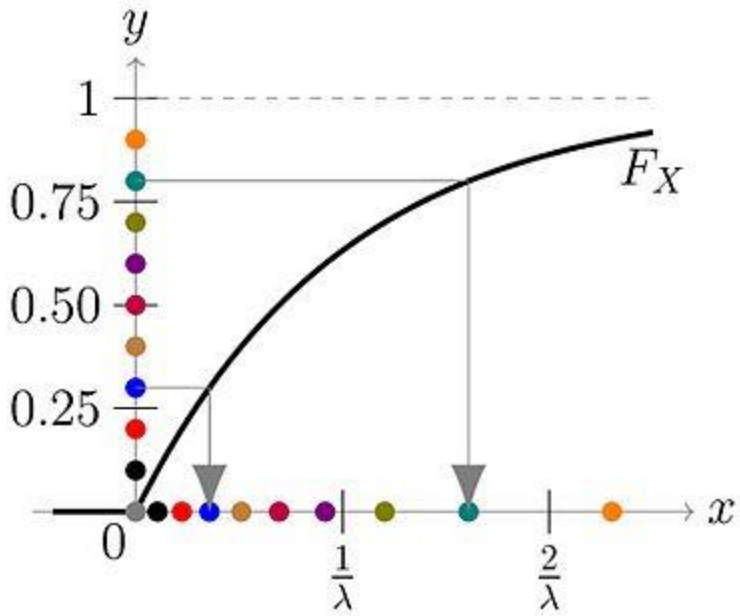


- If the random variables are independent, then: $P(X, Y) = P(X)P(Y)$ in the discrete case and $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$

Probability: sampling

- How to sample from the uniform distribution?
  - Physical methods
  - Pseudo-random generators
- How to sample from a distribution whose cdf F we know?
- Answer: Inverse transform sampling
  - 1. Generate a random u from Uniform[0,1].
  - 2. Return the value x such that F(x) = u.
- IID (independent identically distributed)

samples



Statistical inference

- Probabilities describe populations
- Statistics: generating conclusions about a population from a noisy sample
  - Elections
  - Weather
- Estimation
  - point
  - interval
- Hypothesis testing

Point estimates: mean

- Distributions and parameters vs samples and estimates
- Unknown variable X with a defined mean μ
- We get a **iid** sample of n values $S_n = \{X_1, \dots, X_n\}$
- Goal: estimate μ based on the sample
- Estimators map samples (sets) to estimates (numbers) of the parameters
- **Sample mean estimate**: $\mu_* = \frac{1}{n}\sum_i X_i$
- How close are μ_* and μ?
- Note: μ_* is random since it is a function (average) of random quantities

Point estimates: variance

- Distributions and parameters vs samples and estimates
- Unknown variable X
- We get a **iid** sample of n values $S_n = \{X_1, \dots, X_n\}$
- Estimate $Var(X)$
  - **Sample variance estimator**: $\frac{1}{n-1}\sum_i (X_i - \mu_*)^2$

Point estimates: bias

- Distributions and parameters vs samples and estimates
- Unknown variable X
- We get a **iid** sample of n values $S_n = \{X_1, \dots, X_n\}$
- We estimate a parameter (for example μ)
- If we repeatedly did this over many random sample sets S_n and get a set of estimates, would their average be close to the real μ?
- If the answer is **YES** then the estimator is said to be **unbiased**

Point estimates: median

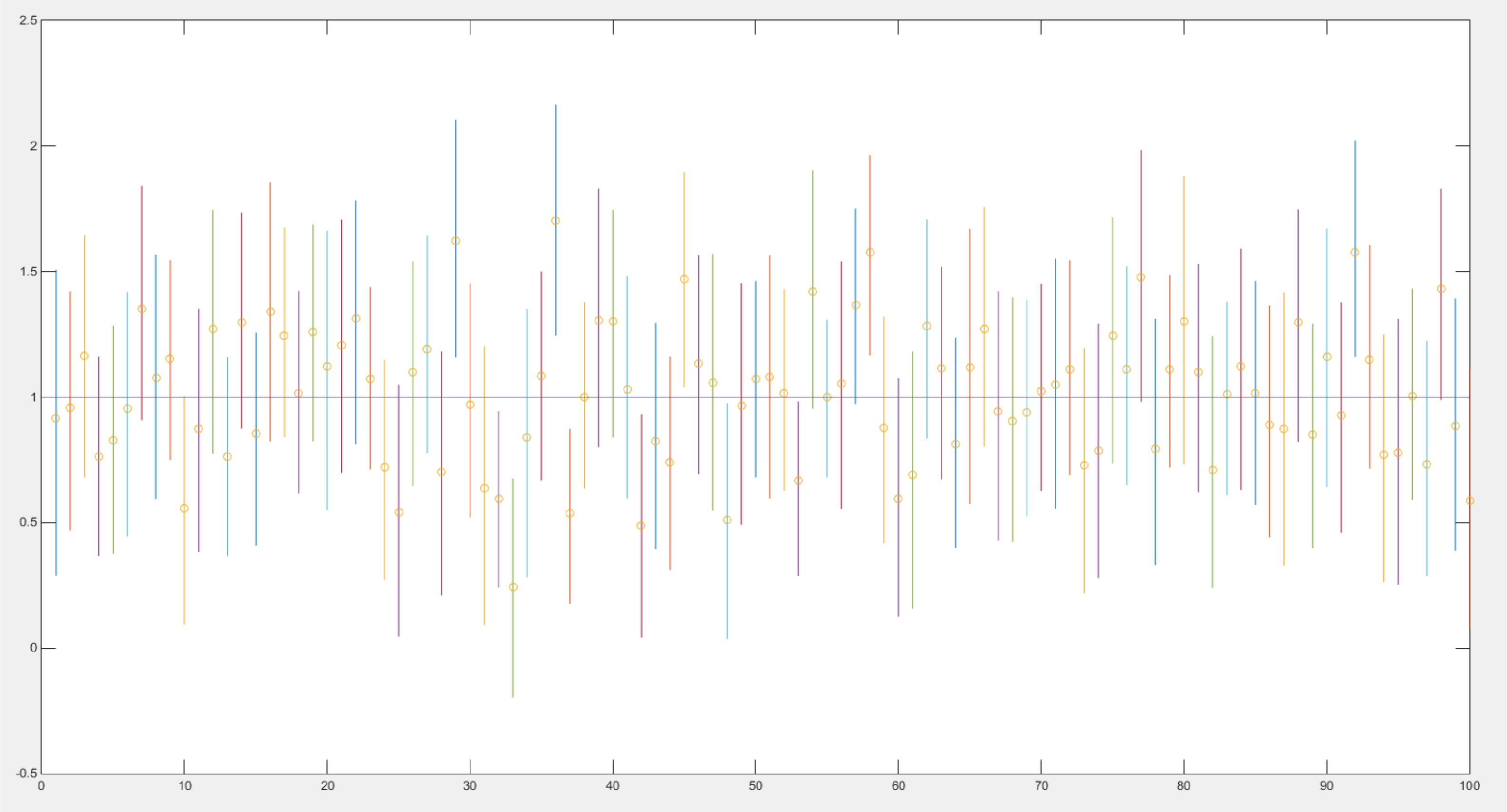
- Distributions and parameters vs samples and estimates
- Unknown variable X
- We get a **iid** sample of n values $S_n = \{X_1, \dots, X_n\}$
- Estimate $median(X)$
  - If the pdf of X is **normal**, then **sample median estimator**: $median_* = median(S_n)$
  - If the pdf is not symmetric then the sample median estimator may be **biased**
  - Bias: does the **average** of estimate over **many** sample sets equal the **true parameter**

Interval estimate: confidence intervals

- Point estimators take sample sets and return numbers (estimates of the parameters)
- The estimates are random – how far are they from the true parameter?
- Interval estimators take sample sets and return **intervals**
- Confidence interval estimator at level α (example 0.90) will contain the true parameter α fraction cases (90%) if we repeated the experiment many times.
- Each time we will get a **different** parameter estimate and a **different interval** around it (the width will vary as well)

Interval estimate

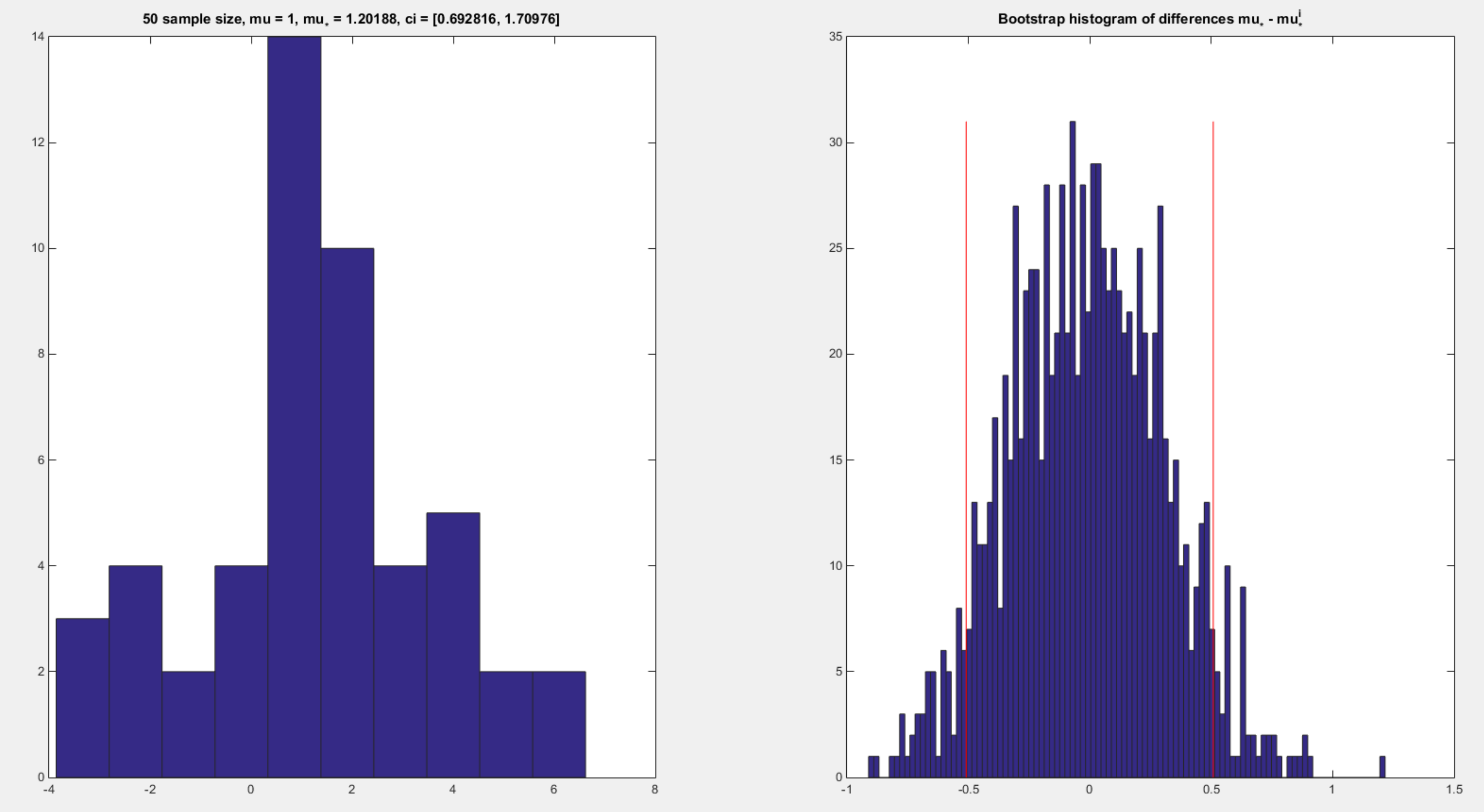
- N(1,2)
- 100 experimets
- Each time we get a different estimate μ_* and a different 90% CI
- **87 intervals contain the true mu=1**



Interval estimate

- How do we compute the interval given a sample?
- We used a **bootstrap** CI estimate – a **resampling** technique
- The idea:
  - Use the sample S_n to generate m new datasets, each time by picking n numbers from S_n **with replacement** (elements can repeat) to create a sets $S_n^1, S_n^2, \dots S_n^m$
  - **[10, 2, 5] -> {[10,10,5], [2,5,10], [5, 2, 2],[5,5,5]…}**
- Compute μ_* on the sample S_n and an estimate μ_*^i for S_n^i
  - [17/3] -> {25/3, 17/3, 9/3, 15/3…}
- The differences $\{\mu_* - \mu_*^1, \mu_* - \mu_*^2 \dots, \mu_* - \mu_*^m\}$ reveal how much the estimate varies
  - { -8/3, 0, 8/3, 2/3 }

Interval estimate



Hypothesis testing

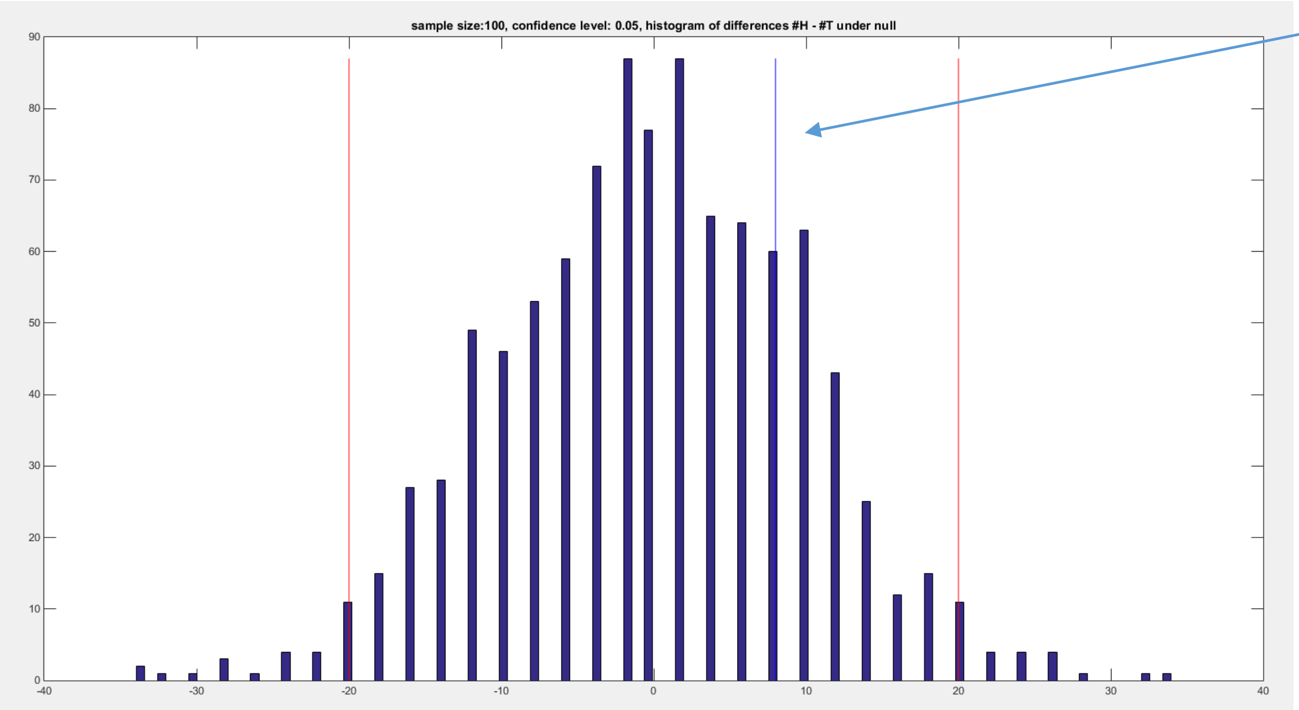
- Example
  - 100 coin tosses, 54 heads, 46 tails
  - Is the coin fair?
  - This could be a result of an unfair coin with p = 0.54, but would we be surprised if a fair coin resulted in 54H, 46T?
  - What if we threw 1000 coins and got: 540H, 440T?
- Two competing models – two hypothesis
  - H_0: coin is fair $p = 0.5$
  - H_1: coin is not fair $p \neq 0.5$

Hypothesis testing

- Example
  - 100 coin tosses, 54 heads, 46 tails
  - Is the coin fair? Is this difference 54-46 very unexpected for fair coins?
- Two competing models – two hypothesis
  - H_0: coin is fair $p = 0.5$
  - H_1: coin is not fair $p \neq 0.5$ (**two sided test: p < 0.5 or p > 0.5**)
- Strategy:
  - Select a confidence level, for example 95%
  - Assume that H_0 is true and generate many sets of 100 tosses
  - Compute the histogram of differences #H - #T
  - If 54-46 = 8 is in the top 2.5% or bottom 2.5% (**two sided test**) then **reject** the null hypothesis
  - Else, **fail to reject** (the difference is not large enough)

Hypothesis testing

- Example
  - 100 coin tosses, 54 heads, 46 tails
  - Is the coin fair? Is this difference 54-46 very unexpected for fair coins?

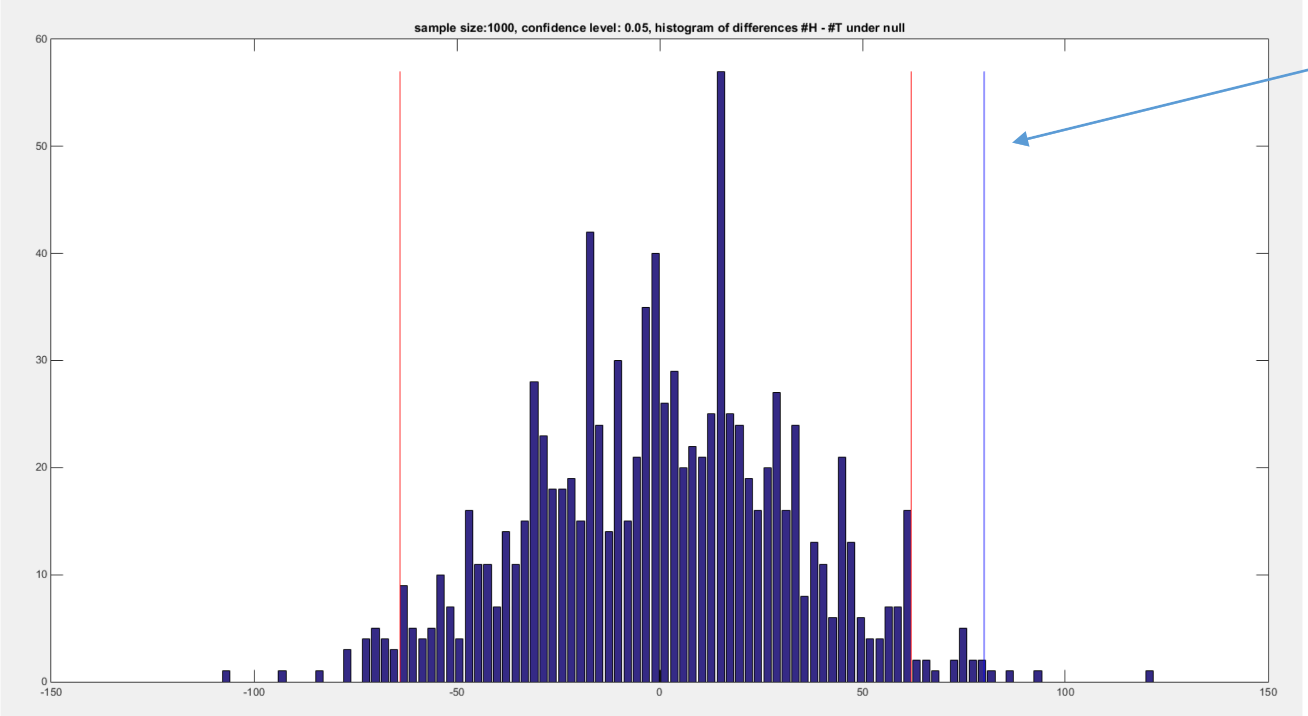


Not a surprising difference under H_0

FAIL TO REJECT

Hypothesis testing

- Example
  - 1000 coin tosses, 540 heads, 460 tails
  - Is the coin fair? Is this difference 540-460 very unexpected for fair coins?

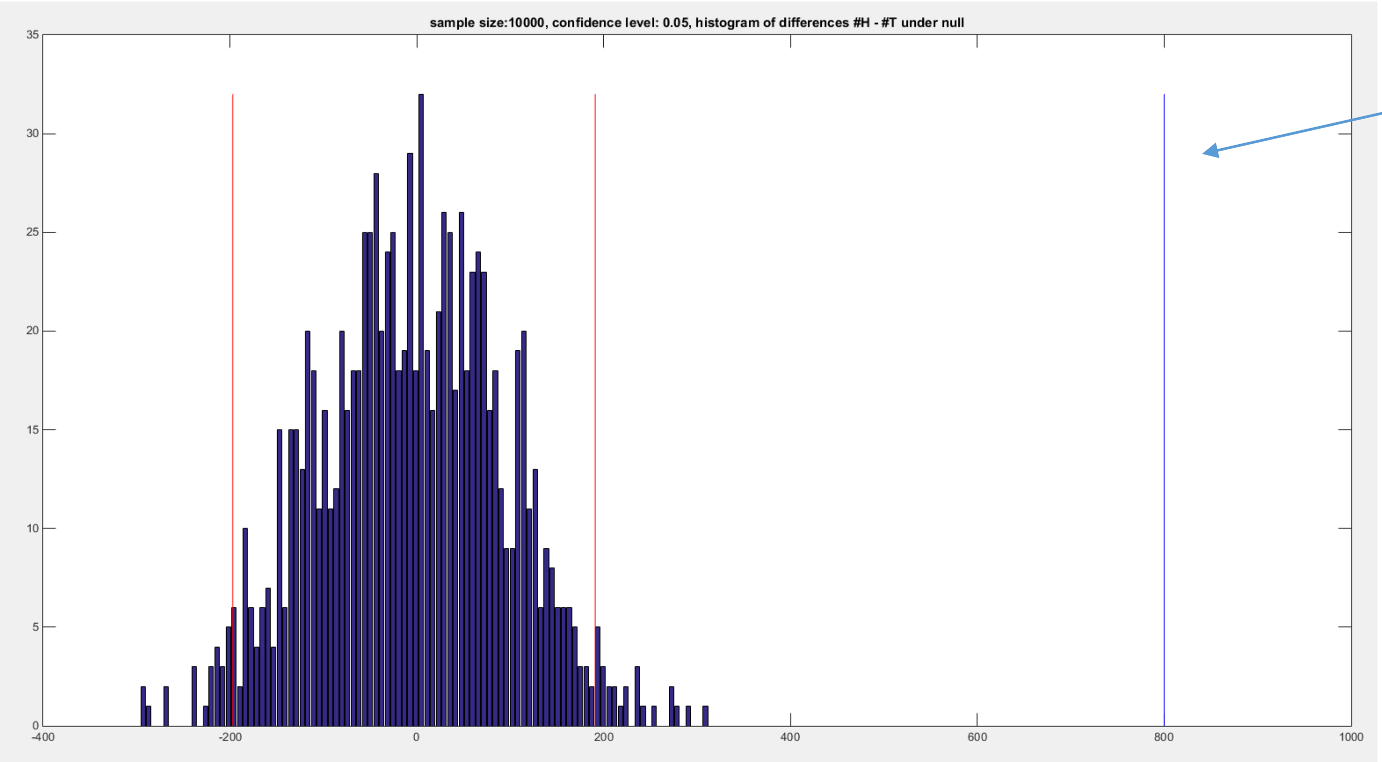


Surprising difference under H_0

REJECT H_0!

Hypothesis testing

- Example
  - 10000 coin tosses, 5400 heads, 4600 tails
  - Is the coin fair? Is this difference 5400-4600 very unexpected for fair coins?



Very surprising difference under H_0

REJECT H_0!

Hypothesis tests

- Four scenarios
  - H_0 is true, fail to reject
  - H_0 is true, reject (**FALSE DISCOVERY, Type I error**)
  - H_0 is false, fail to reject (**Type II error**)
  - H_0 is false, reject (**DISCOVERY**)
- The power of a test: if the null is false, will we detect it?
- Larger samples => more power
- Bigger differences => more power (harder it is for the null to discourage us)

Different test outcomes

- Explore how different types of errors arise
- Fix the true parameter $p = 0.5$ and use a sample size n and see what happens over many scenarios (H_0 is **true**)
- Loop
  - Generate a random sample
  - Test H_0: $p = 0.5$
  - Check result (one of four scenarios)
- Check the error table: how many times did we reject the null?
- How about when H_0

Different test outcomes

- How about when H_0 is false
- Fix the true parameter $p = 0.6$ and use a sample size n and see what happens over many scenarios (H_0 is **true**)
- Loop
  - Generate a random sample
  - Test H_0: $p = 0.5$
  - Check result (one of four scenarios)
- Check the error table: how many times did we fail to reject the null?