Introduction to Statistical Inference

Dr. Fatima Sanchez-Cabo
f.sanchezcabo@tugraz.at

http://www.genome.tugraz.at

Institute for Genomics and Bioinformatics,

Graz University of Technology,

Austria

Summary last session

- Motivation
- Algebra of sets
- Definition of probability space:
  - Sample space
  - Sigma algebra
  - Probability axioms
- Conditional Probability and independence of events
- Random variables
  - Continuos vs discrete
  - pdf and mass function
  - Distribution function

Part IV:

From Probability to Statistics

Statistical Inference

- The target of statistical inference is to provide some information about the probability distribution P defined over the probability space $(\Omega, \mathcal{F})$.
- Differently from the previous examples where an exhaustive observation was possible, this is often difficult.
- Hence, statistical inference focusses in the analysis and interpretation of the realizations of the random variable in order to draw conclusions about the probability law under study.
- The conclusions can be relative to:
  - Estimation of a unique value for a parameter or parameters essential in the probability distribution (i.e., p for a Binomial r.v)
  - Estimation of a confidence interval for this parameter/s.
  - Accept or reject a certain hypothesis about the probability distribution of interest.

Statistical Inference

- In general, there is some knowledge about the probability distribution explaining a certain random process
- The inference process is often involved with deciding (looking at the available data) which is the distribution that best explains the available data among a set of them: $\{F_\theta : \theta \epsilon \Theta\}$. That's known as **parametric statistics**.

Example 4.1: *We are betting with a friend to "head" or "tail" when tossing a coin. We want to know if the coin is unbiased. The most logic approach will be to toss the coin "enough" times to get an approximate value for the probability of head.*

Choosing a random sample

- The first step in this process is to choose a sample representative from the whole population (what is "enough" in the previous example?)
- We should also take care of the cost/time.
- **Random sampling** is one way to obtain samples from a population: all individuals have the same probability to be chosen. This type of sampling is particularly useful because the observations are then independent from each other; hence $F(x_1, \ldots, x_n) = \prod_i F(x_i)$
  - Sampling with replacement (some individuals might be repeated in the sample)
  - Sampling without replacement (each individual can be only once in the sample)

Combinatorics

- We can define the probability of an event A as:

$$P(A) = \frac{\text{number favorables cases}}{\text{number of possible cases}}$$

- The number of possible samples of r elements chosen without replacement among n possible elements is $C(n,r) = \frac{n!}{(n-r)!r!}$
- The number of possible ways to create a sample of size n using $r < n$ elements, if repetition is allowed, is r^n

Examples:

1. An urn contains 5 red, 3 green, 2 blue and 4 white balls. A sample of size 8 is selected at random without replacement. Calculate the probability that the sample contains 2 red, 2 green, 1 blue, and 3 white balls.
2. Consider a class with 30 students. Calculate the probability that all 30 birthdays are different.

Example 4.2

We are betting with a friend to "head" or "tail" when tossing a coin. We want to know what is the probability of getting "head" so that we decide what to bet on. Hence, we have the random variable

$$X = \begin{cases} 0 & \text{Head} \\ 1 & \text{Tail} \end{cases}$$

with $P(X = 0) = \theta$, $P(X = 1) = 1 - \theta$. Given a random sample (X_1, X_2, X_3), determine the mass function (probability for each possible triplet). Does this sample has a known mass distribution?

Descriptive statistics

- Given a very large or very complex sample it might not be possible to determine the distribution easily.
- It is often useful to make use of some tools that help to understand the target distribution and the main characteristics of the sample data:

	Probability	**Statistic**
Base	Population	Sample
Central tendency	Expectation	Mean Median
Dispersion	Variance	Sample variance IQR Mode

Descriptive statistics

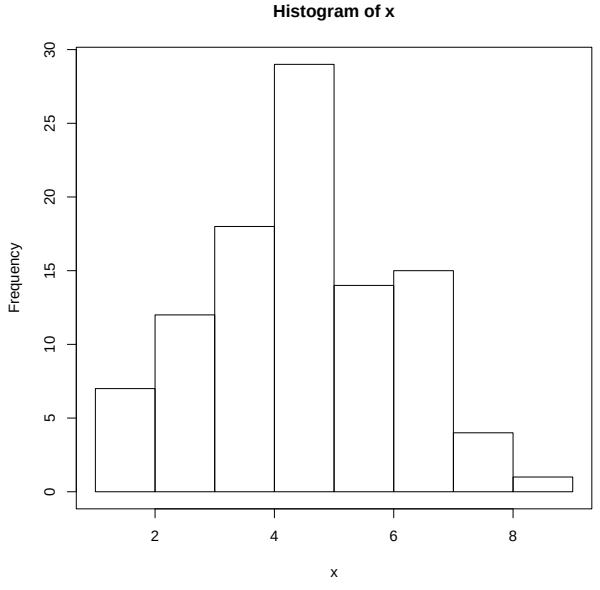
Given a sample $(x_1, \ldots, x_n)$ we can calculate the following expressions to get a rough idea of the properties of the sample.

1. Parametric
   - Mean: $\bar{x} = \frac{1}{n} \sum_i x_i$
   - Sample variance: $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$
   - Sample Quasi-variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
2. Non-parametric (order statistics)
   - Median:

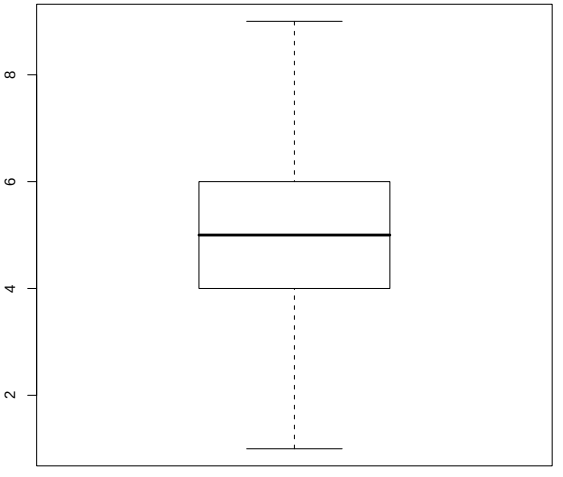
$$\begin{cases} x_{[\frac{n+1}{2}]} & \text{if n is odd} \\ \\ \frac{x_{[\frac{n}{2}]} + x_{[\frac{n}{2}]+1}}{2} & \text{if n is even} \end{cases}$$

   - Interquantiles range: IQR=q_3-q_1
   - Mode: most repeated value

Descriptive statistics



(a) Histogram



(b) Boxplot

Descriptive statistics

Given $x_1, \ldots, x_n$ a sample of size n, calculate the pth percentile:

1. $R = int(\frac{p}{100} \cdot (N+1))$;
   $FR = \frac{p}{100} \cdot (N+1) - int(\frac{p}{100} \cdot (N+1))$
2. Find $x_{[R]}, x_{[R]+1}$
3. $p = x_{[R]} + (x_{[R]+1} - x_{[R]}) * FR$

Estimators and statistics

- **Definition:** A statistic is a function $\mathrm{T} : (\Omega, \mathcal{F}) \to (\mathbb{R}^k, \mathbb{B}^k)$, i.e. $T(x_1, \ldots, x_n)$, over which a probability function can be defined.
- k is the dimension of the statistic.
- An estimator is a particular case of a statistic that approximates one of the unknown parameters of the distribution.
- Examples:
  - $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - $\hat{\theta}$ in example 4.2.
- An estimator is a random variable itself: Its moments can be calculated
- An estimator is said to be **unbiased** if $E[\hat{\theta}] = \theta$

Example 4.3

- Given $(X_1, \ldots, X_n)$ a random sample such as $E[X_i] = \mu$ and $V[X_i] = \sigma^2$, show that for $\bar{X} = \frac{1}{n}\sum_i X_i$ then $E[\bar{X}] = \mu$ and $V[\bar{X}] = \sigma^2/n$.
- Find an unbiased estimator for σ^2?

Note: The variance of an estimator is the sample error (SE).

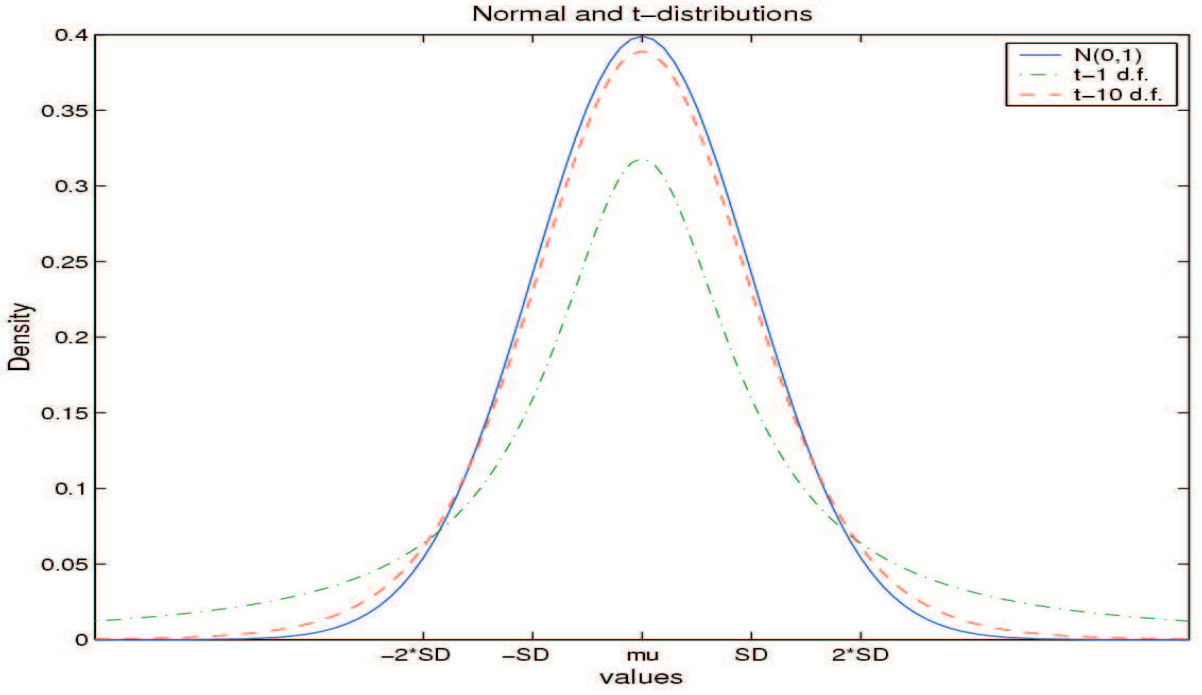
Distribution of the statistics

- Since the statistics are function of random variables they are theirselves random variables for which a probability distribution can be defined.
- Particularly important are the distributions of statistics based on normally distributed random variables.
  - We have proved that $\bar{X} \equiv N(\mu, \sigma^2/n)$.
  - For the Central Limit Theorem $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \equiv N(0,1)$.
  - If σ is unknown and we substitute it by its natural estimator S^2, then

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

That is the **Student-t distribution**.

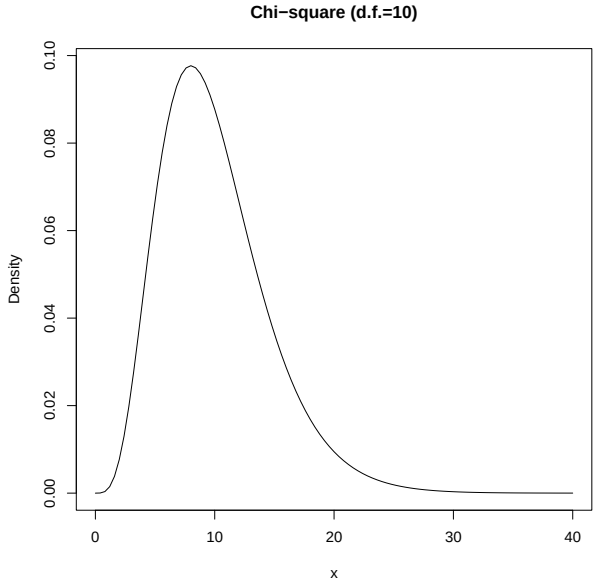
Student's t


Normal and t–distributions
N(0,1)
t–1 d.f.
t–10 d.f.
Density
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
–2*SD
–SD
mu
SD
2*SD
values

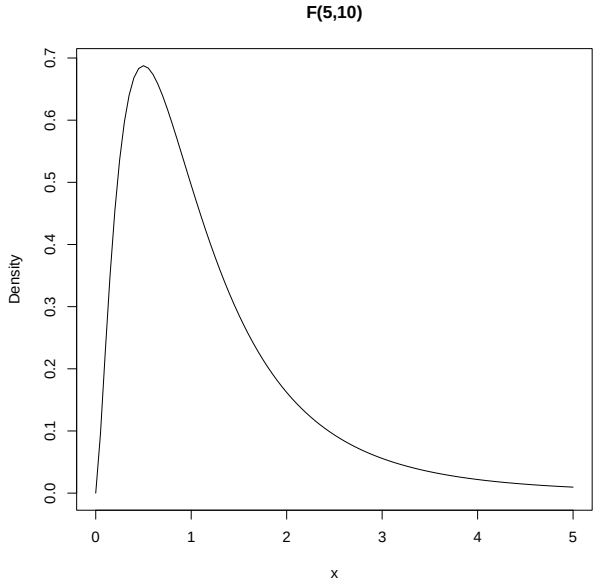
χ^2 and F-Snedecor

Other important distributions of statistics from normally distributed random variables are:

- $\chi^2_n = \sum_i X_i^2$ when $X_i \equiv N(0,1)$
- $F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$



(c) χ^2 density



(d) F-Snedecor

Part V:

Hypothesis testing

Motivation

- There are some situations when we want to test if a certain hypothesis about the parameter of interest is true
- Is the probability of "head" the same than the probability of "tail" when flipping the coin of the previous experiment?

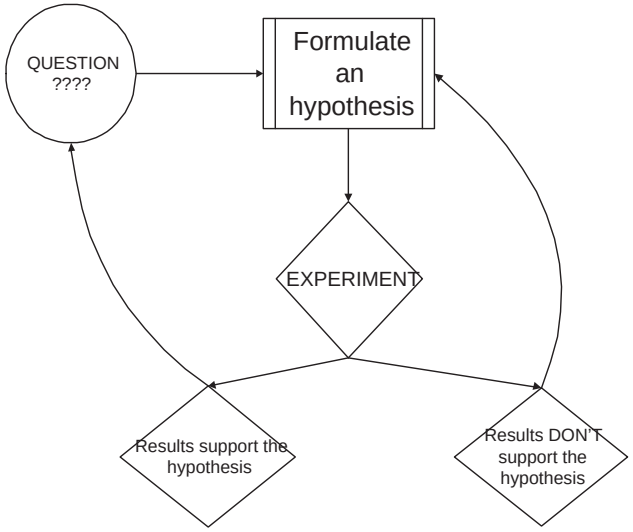


Figure 1: The Scientific Method

Main concepts

- A **statistical null hypothesis** H_0 is the one that want to be tested.
- Statistical tools provide the appropriated framework to guarantee that its rejection is due to real mechanisms and not to chance. Its rejection will provide the called "statistical significance".
- We never talk about "accepting" H_0: with the available data we can say if we reject or not the H_0.

Error types

H_0 **is really...**		
Decision	**True**	**False**
Accept H_0	**OK!**A true hypothesis has been accepted	**Error!** A false hypothesis has been accepted. This is a **Type II error.** The probability of this is β
Reject H_0	**Error!** a true hypothesis has been rejected. This is a **Type I error.** The probability of this occurring is α	**OK!** A false hypothesis has been rejected

Error types

- The ideal situation would be to minimize the probability of both errors (α,β), where:

$$\alpha = P(Reject\ H_0 / H_0\ is\ true) = P(ETI)$$

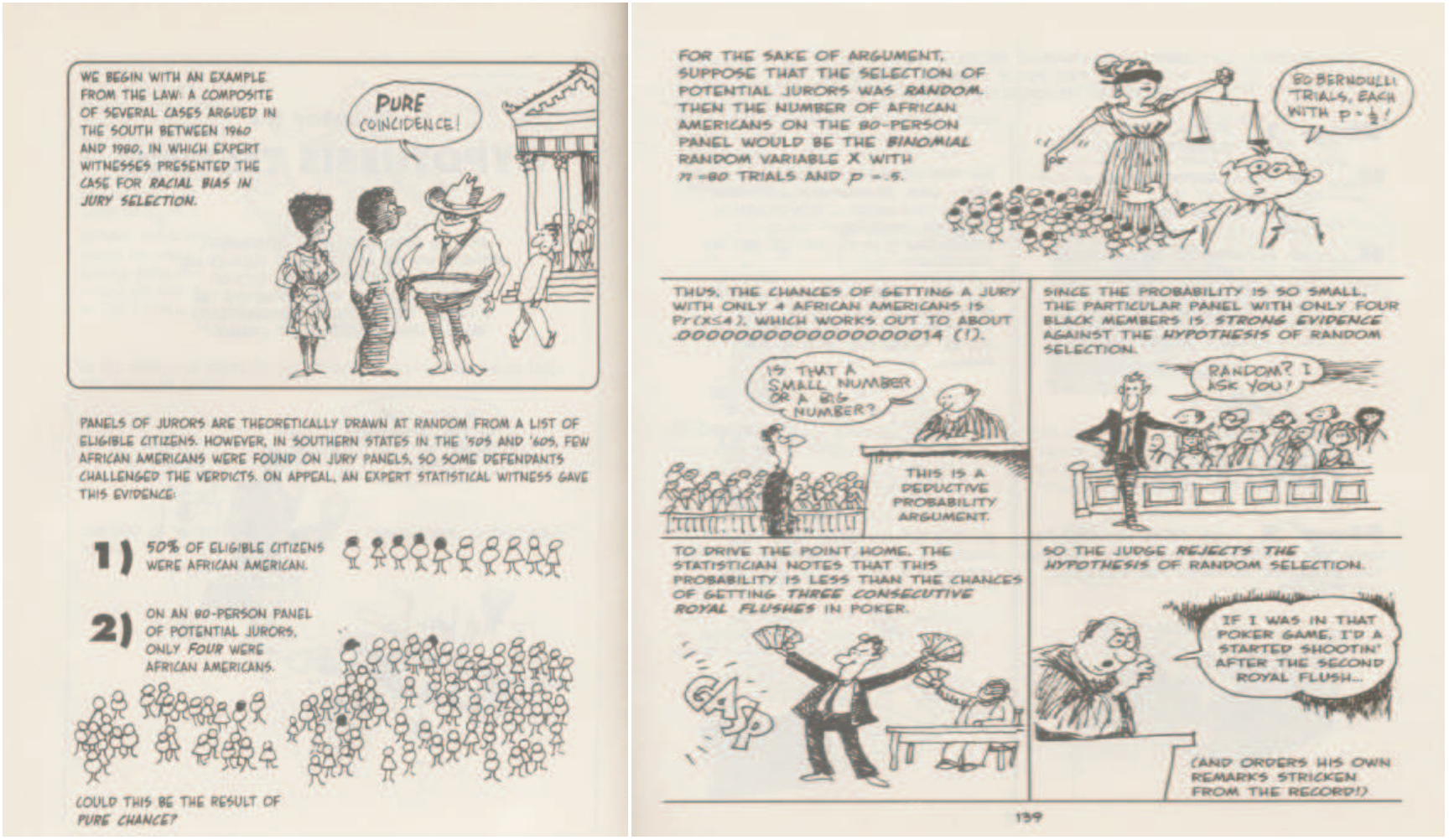
$$\beta = P(Accept\ H_0 / H_0\ is\ false) = P(ETII)$$

- However, they are related and cannot be minimized at the same time.
- **Definition:** The **p-value** corresponds to the probability of rejecting the null hypothesis if it is actually true. If the p-value is smaller than the allowed α-level H_0 is rejected (always with the available data!)

Protocol for statistically testing a null hypothesis

1. Question that we want to address (aufpassen: Statistical tests can only reject the null hypothesis, H0)
2. Experimental design: How many repetitions we need How are we going to select a sample that represents the population? Which are the sources of variation? How do we remove the systematic errors?
3. Data collection
4. Removal systematic errors
5. Which statistic summarizes best my sample. Does it have a known distribution? (statistic: numerical variable that summarizes my sample data in a meaningful way)

Example (5)


WE BEGIN WITH AN EXAMPLE FROM THE LAW: A COMPOSITE OF SEVERAL CASES ARGUED IN THE SOUTH BETWEEN 1960 AND 1980, IN WHICH EXPERT WITNESSES PRESENTED THE CASE FOR *RACIAL BIAS IN JURY SELECTION.*
PURE COINCIDENCE!
PANELS OF JURORS ARE THEORETICALLY DRAWN AT RANDOM FROM A LIST OF ELIGIBLE CITIZENS. HOWEVER, IN SOUTHERN STATES IN THE '50S AND '60S, FEW AFRICAN AMERICANS WERE FOUND ON JURY PANELS, SO SOME DEFENDANTS CHALLENGED THE VERDICTS. ON APPEAL, AN EXPERT STATISTICAL WITNESS GAVE THIS EVIDENCE:
1) 50% OF ELIGIBLE CITIZENS WERE AFRICAN AMERICAN.
2) ON AN 80-PERSON PANEL OF POTENTIAL JURORS, ONLY *FOUR* WERE AFRICAN AMERICANS.
COULD THIS BE THE RESULT OF *PURE CHANCE?*
FOR THE SAKE OF ARGUMENT, SUPPOSE THAT THE SELECTION OF POTENTIAL JURORS WAS *RANDOM.* THEN THE NUMBER OF AFRICAN AMERICANS ON THE 80-PERSON PANEL WOULD BE THE *BINOMIAL* RANDOM VARIABLE X WITH $n = 80$ TRIALS AND $p = .5$.
80 BERNOULLI TRIALS, EACH WITH $p = \frac{1}{2}$!
THUS, THE CHANCES OF GETTING A JURY WITH ONLY 4 AFRICAN AMERICANS IS $Pr(X \leq 4)$, WHICH WORKS OUT TO ABOUT .00000000000000014 (1).
IS THAT A SMALL NUMBER OR A BIG NUMBER?
THIS IS A DEDUCTIVE PROBABILITY ARGUMENT.
SINCE THE PROBABILITY IS SO SMALL, THE PARTICULAR PANEL WITH ONLY FOUR BLACK MEMBERS IS *STRONG EVIDENCE* AGAINST THE *HYPOTHESIS* OF RANDOM SELECTION.
RANDOM? I ASK YOU!
TO DRIVE THE POINT HOME, THE STATISTICIAN NOTES THAT THIS PROBABILITY IS LESS THAN THE CHANCES OF GETTING *THREE CONSECUTIVE ROYAL FLUSHES* IN POKER.
GASP
SO THE JUDGE *REJECTS THE HYPOTHESIS* OF RANDOM SELECTION.
IF I WAS IN THAT POKER GAME, I'D A STARTED SHOOTIN' AFTER THE SECOND ROYAL FLUSH...
(AND ORDERS HIS OWN REMARKS STRICKEN FROM THE RECORD!)
139

Simple hypothesis testing

1. We want to test if a particular sample comes from a distribution with a particular mean value, let's say 3. It is quite logical to try to estimate how close the estimator of the population mean is from the proposed value. So we will calculate:

$$(\bar{x} - 3)$$

2. To minimize the intrinsic error of the sample mean, we will divide this difference by the standard error (variance of the sample mean). We have then the statistic:

$$t = \frac{\bar{x} - 3}{s/\sqrt{n}}$$

3. Regarding to the definition of t-student, if H_0 is true $(H_0 : \mu = 3)$ then, $t \equiv t_{n-1}$ where n is the number of samples. This makes sense, because in the $t - student$ distribution, the probability of getting $t = 0$ is very high and this is exactly what would happen under H_0, because:

$$\bar{x} \simeq 3 \rightarrow \bar{x} - 3 \simeq 0 \rightarrow t = 0$$

Simple hypothesis testing (cont.)

- α was defined as:

$$1 - \alpha = P(|t| \leq t^*) \rightarrow \alpha = P(|t| > t^*)$$

  Where $t^* = t_{n-1;\frac{\alpha}{2}}$ and α is as well defined as:

$$\alpha = P(Reject\ H_0 / H_0\ true) = P(ETI)$$

- In this particular example, if we are under the H_0 the value calculated for the statistic $|t|$ should be close to 0. Otherwise, we reject H_0 (this would mean that $\bar{x} \neq 3$). But, **how big** must be the difference ($\bar{x}$-3)? For that, we fixed the value α such as we just reject H_0 being true for 5 out of 100 samples that we take. And for this particular, this α determines a unique value, called critical value:

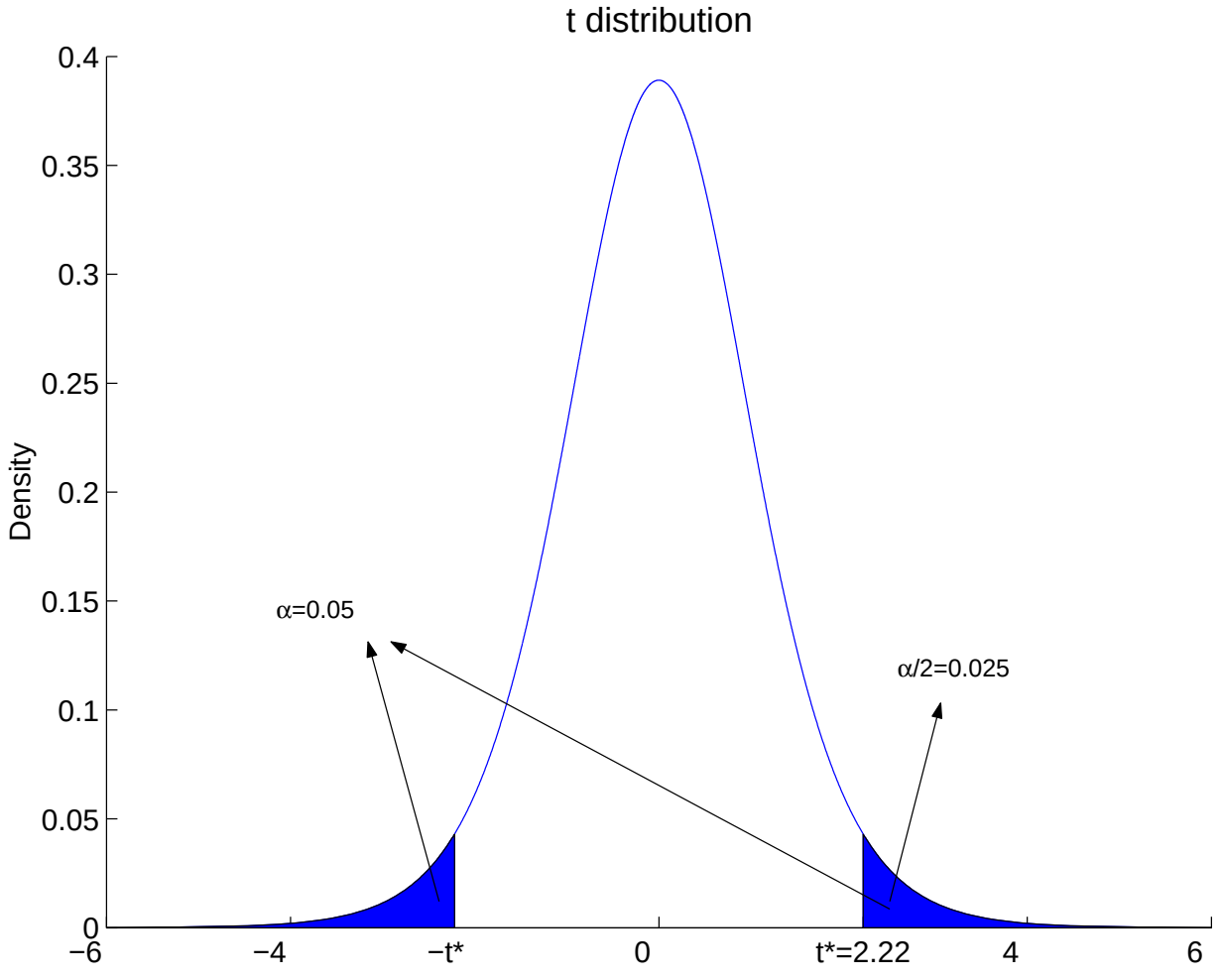
$$t^* = t_{n-1;\frac{\alpha}{2}}$$

- If $|t| > t^*$, we reject the H_0, because the probability of $t = 0$ (μ =3) is very small (<0.05). Just in 5 out of 100 samples (what we can consider by chance) the value of the statistic will be very far away from 0 although the H_0 is true. This condition is equivalent to:

$$p = P(t_{n-1} > |t|) < \alpha \ = \ 0.05$$

- If $|t| < t^*$ the result obtained is the one we would expect for a distribution following a t-distribution with $n - 1$ degrees of freedom, as expected if the H_0 is true. We don't have then enough evidence to reject the H_0. This is equivalent to say that:

$$p = P(t_{n-1} > |t|) > \alpha \ = \ 0.05$$


t distribution
Density
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
α=0.05
α/2=0.025
−6
−4
−t*
0
t*=2.22
4
6

References

[1] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1996) *Biological sequence analysis*, Cambridge University Press.

[2] Durret, R. (1996) *Probability: Theory and examples*, Duxbury Press, Second edition.

[3] Rohatgi, V.K. and Ehsanes Saleh, A.K.Md. (1988) *An introduction to probability and statistics*, Wiley, Second Edition.

[4] Tuckwell, H.C. (1988) *Elementary applications of probability theory*,

[5] Gonick, L. and Smith, W. (2000) The cartoon guide to statistics. Chapman and Hal

[Engineering statistics] http://www.itl.nist.gov/div898/handbook/