

Testing: is my coin fair ?

- Formally: we want to make some inference about $P(\text{head})$
- Try it: toss coin several times (say 7 times)
- Assume that it is fair ($P(\text{head})=0.5$), and see if this assumption is compatible with the observations.



# tosses	# heads	Comment ?	Probability
1	1	Ok	0.50
2	2	Ok	0.25
3	3	Ok	0.12
4	4	Unusual	0.06
5	5	Surprising	0.03
6	6	Strange	0.02
7	7	I don't believe it!	0.01



EMBnet course
Basel 19 January 2009

		Number of heads										
Number of tosses		0	1	2	3	4	5	6	7	8	9	10
	1	0.5 1	0.5 0.5									
	2	0.25 1	0.5 0.75	0.25 0.25								
	3	0.125 1	0.375 0.875	0.375 0.5	0.125 0.125							
	4	0.0625 1	0.25 0.9375	0.375 0.6875	0.25 0.3125	0.0625 0.0625						
	5	0.03125 1	0.15625 0.96875	0.3125 0.8125	0.3125 0.5	0.15625 0.1875	0.03125 0.03125					
	6	0.01563 1	0.09375 0.98438	0.23438 0.89063	0.3125 0.65625	0.23438 0.34375	0.09375 0.10938	0.01563 0.01563				
	7	0.00781 1	0.05469 0.99219	0.16406 0.9375	0.27344 0.7734	0.27344 0.5	0.16406 0.22656	0.05469 0.0625	0.00781 0.00781			
	8	0.00391 1	0.03125 0.99609	0.10939 0.96484	0.21875 0.85547	0.27438 0.63672	0.21875 0.36328	0.10939 0.14453	0.03125 0.03516	0.00391 0.00391		
	9	0.00195 1	0.01758 0.99804	0.07031 0.98047	0.16406 0.91016	0.24609 0.74609	0.24609 0.5	0.16406 0.25391	0.07031 0.08984	0.01757 0.01953	0.00195 0.00195	
	10	0.00098 1	0.00977 0.99902	0.04394 0.98926	0.11719 0.94531	0.20508 0.82812	0.24609 0.62304	0.20508 0.37695	0.11719 0.17188	0.04395 0.05469	0.00977 0.01074	0.00098 0.00098



Significant evidence ($p < 0.05$) that the coin is biased towards **head** or **tail**.



EMBnet course
Basel 19 January 2009

8
10

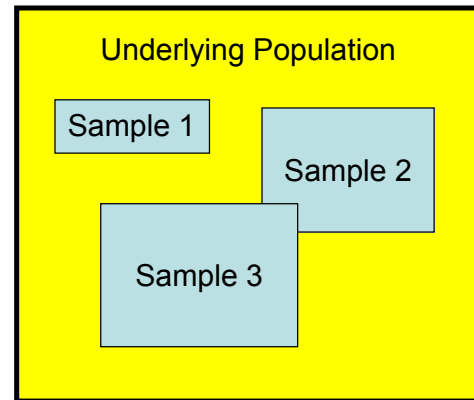
0.04395
0.05469

Probability of obtaining 8 heads in 10 tosses

Probability of obtaining at least 8 heads in 10 tosses

Observations: taking samples

- Samples are taken from an underlying real or fictitious population
- Samples can be of different size
- Samples can be overlapping
- Samples are assumed to be taken at random and independently from each other
- Samples are assumed to be representative of the population (the population “homogeneous”)



Estimation:

- Sample 1 \Rightarrow Mean 1
- Sample 2 \Rightarrow Mean 2
- Sample 3 \Rightarrow Mean 3



EMBnet course
Basel 19 January 2009

Sampling variability

- Say we sample from a population in order to estimate the population mean of some (numerical) variable of interest (e.g. weight, height, number of children, etc.)
- We would use the **sample mean** as our guess for the unknown value of the population mean
- Our sample mean is very unlikely to be exactly equal to the (unknown) population mean just due to **chance variation** in sampling
- Thus, it is useful to quantify the **likely size** of this chance variation (also called ‘chance error’ or ‘sampling error’, as distinct from ‘nonsampling errors’ such as **bias**)
- If we estimate the mean multiple times from different samples, we will get a certain **distribution**.



EMBnet course
Basel 19 January 2009

Central Limit Theorem (CLT)

- The **CLT** says that if we
 - repeat the sampling process many times
 - compute the sample mean (or proportion) each time
 - make a histogram of all the means (or proportions)
- then that histogram of sample means should look like the **normal distribution**
- Of course, in practice we only get one sample from the population, we get one point of that normal distribution
- The CLT provides the **basis for making confidence intervals and hypothesis tests** for means
- This is proven for a large family of distributions for the data that are sampled, but there are also distributions for which the CLT is not applicable



EMBnet course
Basel 19 January 2009

Sampling variability of the sample mean

- Say the SD in the population for the variable is known to be some number σ
- If a sample of n individuals has been chosen 'at random' from the population, then the likely size of chance error of the sample mean (called the **standard error**) is

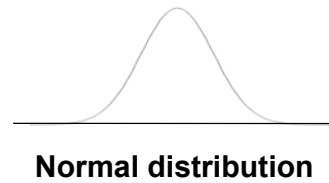
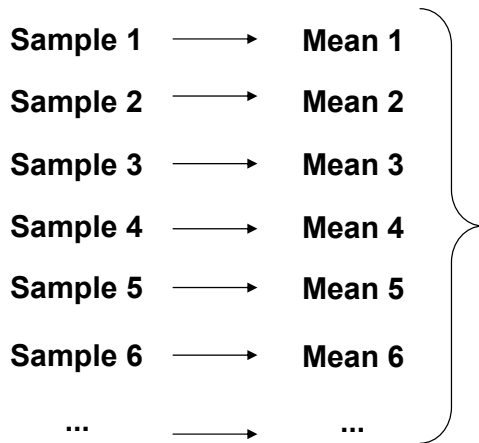
$$SE(\text{mean}) = \sigma / \sqrt{n}$$

- This the typical difference to be expected if the sampling is done twice independently and the averages are compared
- 4 time more data are required to half the standard error
- If σ is not known, you can substitute an estimate
- Some people plot the SE instead of SD on barplots



EMBnet course
Basel 19 January 2009

Central Limit Theorem (CLT)



Mean : true mean of the population
SD : σ / \sqrt{n}

Note: this is the SD of the sample mean, also called Standard Error; it is not the SD of the original population.



EMBnet course
Basel 19 January 2009

Hypothesis testing

- 2 hypotheses in competition:
 - H_0 : the NULL hypothesis, usually the most conservative
 - H_1 or H_A : the alternative hypothesis, the one we are actually interested in.
- Examples of NULL hypothesis:
 - The coin is fair
 - This new drug is no better (or worse) than a placebo
 - There is no difference in weight between two given strains of mice
- Examples of Alternative hypothesis:
 - The coin is biased (either towards tail or head)
 - The coin is biased towards tail
 - The coin has probability 0.6 of landing on tail
 - The drug is better than a placebo



EMBnet course
Basel 19 January 2009

Hypothesis testing

- We need something to measure how far my observation is from what I expect to see if H_0 is correct: a **test statistic**
 - Example: number of heads obtained when tossing my coin a given number of times:

Low value of the test statistic → more likely not to reject H_0
High value of the test statistic → more likely to reject H_0
- Finally, we need a formal way to determine if the test statistic is “low” or “high” and actually make a decision.
- We can never prove that the alternative hypothesis is true; we can only show evidence for or against the null hypothesis !



EMBnet course
Basel 19 January 2009

Significance level

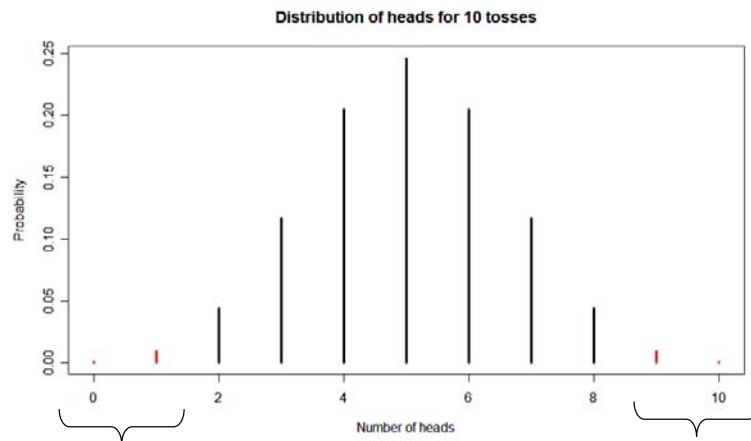
- The **p-value** is the probability of getting a test result that is **as or more extreme** than the observed value of the test statistic.
- If the distribution of the test statistic is known, a p-value can be calculated.
- Historically:
 - A **predefined significance level (α)** is defined (typically 0.05 or 0.01)
 - The value of the test statistic which correspond to the significance level is calculated (usually using tables)
 - If the observed test statistic is above the threshold, we reject the NULL hypothesis.
- Computers can now calculate exact p-values, which are reported
- “ $p < 0.05$ ” remains a magical threshold
- Confusion about p-values:
 - It is **not** the probability that the null hypothesis is correct.
 - It is **not** the probability of making an error



EMBnet course
Basel 19 January 2009

Hypothesis testing

- How is my test statistic distributed if H_0 is correct ?



$$P(\text{Heads} \leq 1) = 0.01074$$

$$P(\text{Heads} \geq 9) = 0.01074$$



EMBnet course
Basel 19 January 2009

One-sided vs two-sided test

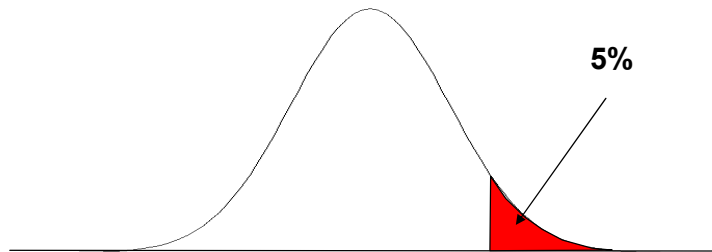
- The choice alternative hypothesis influence the result of the test
- If H_A is “the coin is biased”, we do not specify the direction of the bias and we must be ready for both alternatives (many heads or many tails)
- This is a **two-sided test**.
- If the “total” significance α is e.g. 0.05, it means we must allow $\alpha/2$ (0.025) for bias towards tail and $\alpha/2$ (0.025) for bias towards head.
- If H_A is “the coin is biased towards heads”, we specify the direction of the bias and the test is **one-sided**.
- Two-sided tests are usually recommended because they do not depend on some “prior information” (direction), although they are less powerful.



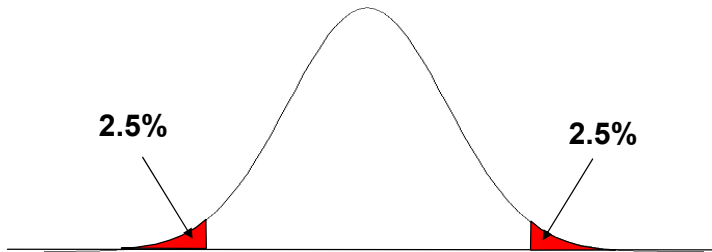
EMBnet course
Basel 19 January 2009

One-sided vs two-sided test

One-sided
e.g. $H_A: \mu > 0$







Two-sided
e.g. $H_A: \mu \neq 0$



EMBnet course
Basel 19 January 2009

Errors in hypothesis testing

Decision \ Truth	not rejected negative	rejected positive
H_0 is true	 specificity True negative TN	 Type I error False Positive α
H_0 is false	 Type II error False Negative β	 Power $1 - \beta$; sensitivity True Positive TP



EMBnet course
Basel 19 January 2009

Power

- Not only do you want to have a low FALSE positive rate, but you would also like to have a high TRUE positive rate – that is, high **power**, the chance to find an effect (or difference) if it is really there
- Statistical tests will not be able to detect a true difference if the **sample size** is too small compared to the **effect size** of interest
- To compute or estimate power of a study, you need to be able to specify the α level of the test, the sample size **n**, the effect size δ , and the SD σ (or an estimate **s**)
- In some types of studies, such as microarrays, it is difficult (impossible?) to estimate power, because not all of these quantities are typically known, and will also vary across genes (e.g. different genes have differing amounts of variability)



EMBnet course
Basel 19 January 2009

One sample t-test

- Is the mean of a population equal to a given value ?
- Example:
 - Given a gene and several replicate microarray measurements (log ratios) g_1, g_2, \dots, g_n . Is the gene differentially expressed or, equivalently, is the mean of the measurements different from 0 ?
- Hypotheses:
 - H_0 : mean equals μ_0 (a given value, often 0)
 - H_A : could be for example
 - mean different from μ_0
 - mean larger than μ_0
 - mean equals μ_1 (another given value)



EMBnet course
Basel 19 January 2009

One sample t-test

- Test-statistic (Student's t-statistic):

$$T = \frac{\bar{x} - \mu_0}{\sqrt{S^2 / n}}$$

- Where

- \bar{x} is the average of the observations
- S is the (estimated) standard deviation
- n is the number of observations
- μ_0 is the given value

- Intuitively:

- Numerator is small (difference is small) → test statistic small → tend not to reject H_0
- S is large (observations are spread out) → test statistic small → tend not to reject H_0
- n is small (few data points) → test statistic small → tend not to reject H_0



EMBnet course
Basel 19 January 2009

One sample t-test

- If the means are equal (H_0 is correct), the value T follows a known distribution (t-distribution)
- The shape of the t-distribution depends on the number of observations: if the average is made of n observations, we use the **t-distribution with $n-1$ degrees of freedom (t_{n-1})**.
 - If n is large, t_{n-1} is close to a normal distribution
 - If n is small, t_{n-1} is more spread out than a normal distribution (penalty because we had to estimate the standard deviation using the data).



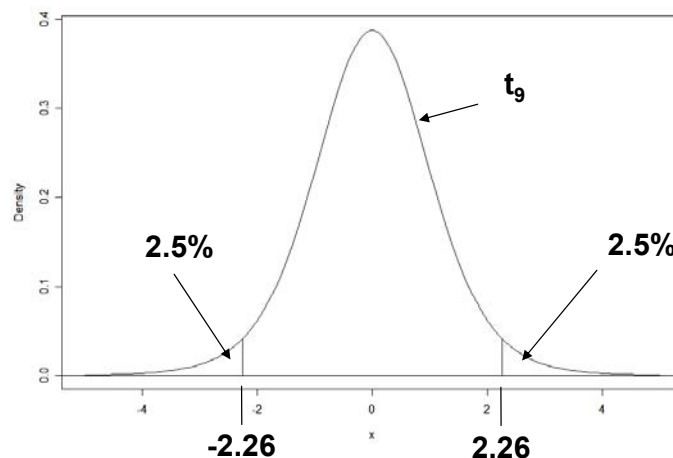
EMBnet course
Basel 19 January 2009

One sample t-test: example

- We have the following 10 data points:
-0.01, 0.65, -0.17, 1.77, 0.76, -0.16, 0.88, 1.09, 0.96, 0.25
- We are wondering if these values come from a distribution with a true mean of 0.
- H_0 : the true mean is 0; H_A : the true mean is not 0
- These 10 data points have a mean of 0.62 (sample mean) and a standard deviation of 0.62 (squared: 0.38).
- From these numbers, we can calculate the t-statistic:
$$T = 0.62 / \sqrt{0.38/10} = 3.01$$

One sample t-test: example

- Since we have 10 observations, we must compare our observed t-statistic to the t distribution with 9 degrees of freedom



- Since $T=3.01 > 2.26$, the difference is significant at the 5% level.
- Exact p-value: $P(|T| \geq 3.01) = 0.015$

Comments and pitfalls

- The t-test assumes that the different observations are **independent** and that they follow a **normal distribution**.
- t-test is **robust**, in the sense that even if the distribution of values is not exactly normal, it should still work.
- If the distribution is far from normal, the test will be less likely to yield significant results
- Other tests may be required to obtain significance (for example permutation test).

- The basic t-test is not popular for microarrays, because the estimation of S is unstable (small sample size)



EMBnet course
Basel 19 January 2009

Two-sample t-test

- Do two different populations have the same mean?
- Test-statistic:
$$T = \frac{\bar{y} - \bar{x}}{SE}$$
- Where
 - \bar{y} and \bar{x} are the average of the observations in both populations
 - SE is the standard error for the difference in means
 - There are a few different ways to compute SE, depending on assumptions
 - One possibility: $SE = \sqrt{S_y^2/n + S_x^2/m}$
- If H is true (no difference in mean), T follows a *t*-distribution with $n + m - 2$ degrees of freedom (where n and m are the number of observations in each sample)



EMBnet course
Basel 19 January 2009

Paired t-test

- In the two-samples t-test, we compared two samples of unrelated data points
- If the data between the two sample is **paired**, that is, each point x_i in the first sample correspond to a point y_i in the second sample, then a **paired t-test** can be used.
- Example: 10 patients are on a diet; we measure their weight before they start the diet and after 6 months; does the group show a significant weight difference ?
- In practice: calculate the difference $x_i - y_i$ between the two measurement and use a one-sample t-test to test if this difference is significantly different from 0.
- This is more powerful than the two-samples t-test because we are providing more information (the pairing) to the test.



EMBnet course
Basel 19 January 2009

Difference between two-sample and two-tailed tests

- A *two-sample test* is a hypothesis test for answering questions about means for *two different populations*
- Data are collected from two random samples of independent observations
- A *two-sided test* is a hypothesis test in which the values for rejecting the NULL are in *both tails of the probability distribution*
- The choice between a one-sided test and a two-sided test is determined by the purpose of the investigation or prior information



EMBnet course
Basel 19 January 2009

Pitfalls in hypotheses testing

- Even if a result is ‘statistically significant’, it can still be due to chance;
- Conversely, if a result is *not* statistically significant, it may be only because you do not have enough data (lack of power)
- Statistical significance is not the same as practical importance;
- A test of significance does not say how important the difference is, or what caused it;
- A test does not check the study design;
- If the test is applied to a nonrandom sample (or the whole population), the p-value may be meaningless;
- **Data-snooping** makes p-values hard to interpret: the test should be fully defined **BEFORE** data are looked at.



EMBnet course
Basel 19 January 2009

1000 simulated coin tosses

```
[1] H T H H T T H H T T T T T T H T T H H H H H H T T T T H H H H H H T H T H
[41] H T H T H H H H H T H H T T T T H H H H H T H T H T T H H T T T T H H H T
[81] T T H H T T H H T T H H H T T H H H H T T H T T H H H H H H H H T T T H H
[121] T T T H T H T H T H H H H H T T H H T T H H T T T H T H T H T H H H H H
[161] T T H T T T T H T T H H T T H T T T H T H H H H H H T H H H H T H T H T H H
[201] T T H H T H T T T H T H T H H T H H T T H T H T H T T H H H H H T T T T H H T
[241] H T T H T H H T T T H H T T H T H H T T T T T T T H H T H H T T T H H H H H
[281] H H H H H H H T T T T T H T H T T T H H H H T T T T T T T T H T H T H T T T
[321] H H T H T T T T T T T T H H T T H T H T H H H T H H H H T T H T T H H H T H T
[361] T H H T T H T H H H H H H H T T T H H H H H H H T T T H T H H T T T H H H T H H
[401] H T T T T H H H H T T T H T T T H T H T H H H T T T H H T T H H T T T H T T T
[441] T T T T H T T T H H T T H H H T T H T T H T H H H T T H H H T H H H H T T T H T
[481] H T T T T H T H T H T H H H T T T T T T H H T H T H H T T T H H T H T H T H
[521] T T T T H H H H T H H H H T T T T H T T T H H T H H T T H H T T H T T T H T
[561] H H H T H T T H T H H T T T T H T T H T T T H T H H T T T H T T H H T T H T T T
[601] H T T T H H H H H T T T H T T H T T H H T T T H H H H H H H H T H H T H T H
[641] H H T T H H T H T T T H H H H H T H H T H T T T H T T T H T T H T T T H H T
[681] H T T T T H T H T T T H T T T T T T H T T T T T T T T T T H H T H T H T H
[721] T T T H H T H H T T T T H T T H T T H H T T H T H H H H H H T H H H H T H T H
[761] H T H H H H H H T H H T H H H H H H H T H T H H H T H H T H H H T H T T H
[801] T T T H T T T H H T T T T T T T T T H H H H H H H T T H H T H T T T H H H H H
[841] T T T T T H H H T H H H H T T T T T H T H H H H T T T H H H T T H H H H
[881] T H T H T T H H T H H H H H T T T H T H H T T H H T T T H T H T T H T T T
[921] T T T T H T H H T H T T T T T T T T T T H H H H H T H H H T H H T H T H T
[961] H H T T T T T H T T H T T T H T H T T T H H T H T H H H T T H H H H H H T H T
```



EMBnet course
Basel 19 January 2009

Multiple testing

- Observation of 11 consecutive “tails”
- Very improbable with a fair coin ($p < 0.0004$)
- Does it mean that the coin is biased ?
- P-values are valid if only one test is done
 - 11 consecutive tails in 11 tosses would be significant
 - 11 consecutive tails in 1000 tosses is not
- If several tests are conducted, the significance of each of them is reduced
 - “If you try more often, you are more likely to succeed, even just by chance.”
 - If an event has probability 1/1000 every time you try, and you try 1000 times, it is likely to happen at least once.



EMBnet course
Basel 19 January 2009

Multiple testing

- Statistical procedures can adapt the results in the case of multiple testing
 - Most well-known and conservative: Bonferroni
 - Divide significance threshold by number of repetitions
 - Example:
 - 1000 tests with threshold 0.01 → corrected threshold = $0.01/1000 = 0.00001$
 - Other procedures are less stringent
- Particularly relevant for microarrays:
 - Showing that 1 **preselected** gene is differentially expressed ($p=0.01$) may be interesting.
 - Showing that 1 gene **out of 10,000** is differentially expressed ($p=0.01$) is probably not interesting.



EMBnet course
Basel 19 January 2009