Hypothesis Testing 2

Clinical trials

- Developing a new drug takes 10-15 years and hundreds of millions of dollars.
- After tests on animals, the final stage is a clinical trial – a test on human patients.
- Very expensive: usually limited to a few hundred patients.
- Need to demonstrate **causality:** that taking the medication causes an improvement in the patient's health, and is not just correlated with it. Requires a **controlled study**:
  - Placebo
  - **Double blindness**
  - **Fixing the protocol before the start of the trial.**

Observational Studies

- With access to all electronic medical records, it is becoming possible to measure the effectiveness of a drug on millions of people (contrast with a few hundred in controlled studies).
- **Potential of revolutionizing medical research.**
- **Challenge: hard to control for non-causal correlations.**
- Challenge can be met by controlling for potential causes: wealth, age, race …
- A new trend: when treatment is not critical and resources are limited – treat a randomly selected part of the population. In this case: correlation does imply causation.

Counting fish

- Suppose we are studying a lake, and we want to estimate how many fish are in it.
- It is not realistic to try and catch all, or even most of them.
- Instead we follow a 3 step process:
  1. Catch m fish, mark them, and release back to the lake.
  2. Let some time pass, so that the marked fish mix with the unmarked.
  3. Catch l fish, count the number of marked fish, call that number the random variable Y.
- Let n be the number of fish in the lake. The probability that a random fish is marked is
  - m/n
- Y is the sum of l IID Binary random variables whose mean is m/n

Counting Fish continued

- Y is the sum of l IID Binary random variables whose mean is m/n E(Y)=?
  - $E(Y) = \frac{lm}{n}$
  - $Var(Y) = l\frac{m}{n}\left(1 - \frac{m}{n}\right) \leq$
    - $\leq l\frac{1}{2}\frac{1}{2} = \frac{l}{4};$ $\sigma(Y) = \frac{\sqrt{l}}{2}$
- The 95% confidence interval for $\frac{lm}{n} = \mathrm{E(Y)}$ is
  - $[Y - \sqrt{l}\,, Y + \sqrt{l}\,]$
- Therefor the 95% confidence interval on the number of fish is
  - $\left[\frac{lm}{Y+\sqrt{l}}, \frac{lm}{Y-\sqrt{l}}\right]$
- If l is too small then the estimate would be weak.
- If m is too small relative to n then we might not catch any marked fish,

The structure of statistical tests

1. Define null hypothesis, alternative hypothesis
2. Define the test statistic: X a random variable that is a function of the data (average, difference between averages etc.)
3. Compute (or take from book) the distribution of the test statistic under the null distribution.
   - By convention: 0 – no significance, large values – high significance.
4. Decide on the desired significance level α and, from it, the threshold $X > T$ which is the minimal value of X needed to reject the null hypothesis with significance α
5. Run the experiment, compute X
   - If $X > T$ reject the null hypothesis
   - Otherwise, the test failed.

The Z-statistic (1)

- Suppose there is a new serum that is claimed to make people taller. We want to test whether this is true.
- Null Hypothesis: the expected height of 20 years olds is the same whether treated or untreated.
- Alternative Hypothesis: The expected height of treated is larger than of untreated.
- Additional assumptions:
  - The mean μ of the untreated people is known.
  - The std σ of treated people is known.
  - The distribution of the average of n treated people is (close to) a normal distribution (Central Limit Theorem).
- The statistic we use is the normalized average.

The Z-Statistic (2)

- The statistic we use is the normalized average:
$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$
- The distribution of Z under the null hypothesis is the standard normal $\mathbb{N}(0,1)$
- Given a desired confidence level α we choose the threshold T so that $Q(T) = \alpha$ or, in other words $T = Q^{-1}(\alpha)$
- The Z-test rejects the null hypothesis if $Z > T$

The paired t-statistic (1)

- Suppose we want to check which wine people like better, wine1 or wine2.
- To evaluate this we select n people at random give each person a taste of each wine and ask them to rate the two wines on a scale of 1 to 10, where 10 is the best and 1, the worst.
- It is a good idea to have each person rate both wines because rating scales vary from person to person. It is more informative to compare the ratings for wine1 and wine2 given by the same person rather than compare rating of wine1 by person 1 to the rating of wine2 by person 2.
- We thus **pair** the rating and then use the difference: rating-by-person1(wine1) – rating-by-person1(wine2).

The paired t-statistic (2)

- The Null Hypothesis: the expected difference between the rating is zero = the two wines are equally liked.
- Alternative 1: wine1 is better liked than wine2 = the expected difference is larger than zero
- Alternative 2: wine2 is better liked than wine1 = the expected difference is smaller than zero.
- We define the average of the sample:

$$\bar{X} \doteq \frac{1}{n}\sum_{i=1}^{n} D_i\,, \qquad D_i \doteq R_i^1 - R_i^2$$

- We assume that n is sufficiently large that the distribution of $\bar{X}$ is close to normal.
- It might seem that we are going to end up with the Z-statistic but in this case we don't assume to know the std.

The paired t-statistic (3)

- Clearly under the null hypothesis $E(R^1) = E(R^2)$ and therefor $E(D) = E(R^1 - R^2) = 0.$
- It might seem that we are going to end up with a Z-statistic but in this case we don't assume to know the std of R^1, R^2 or D
- We therefor use an estimate of the std:

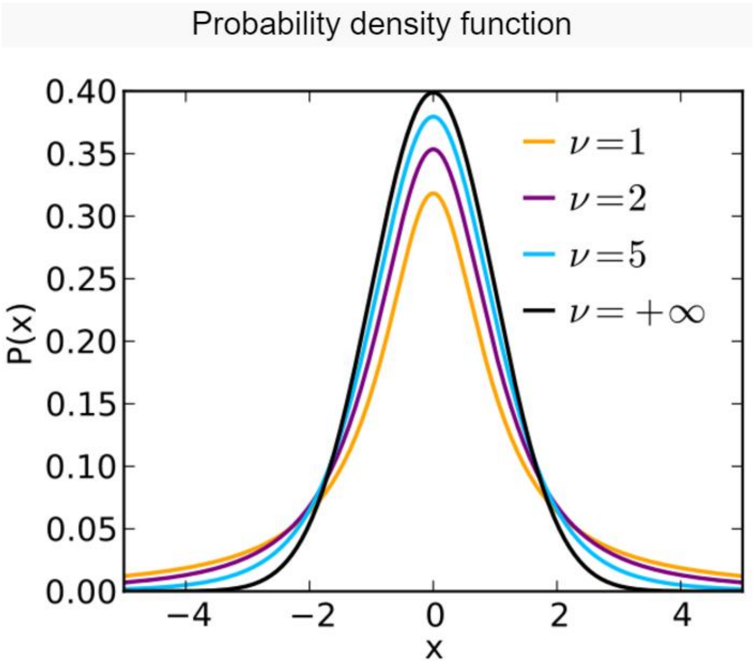
$$\mathrm{s} = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{i=1}^{n} D_i^2}$$

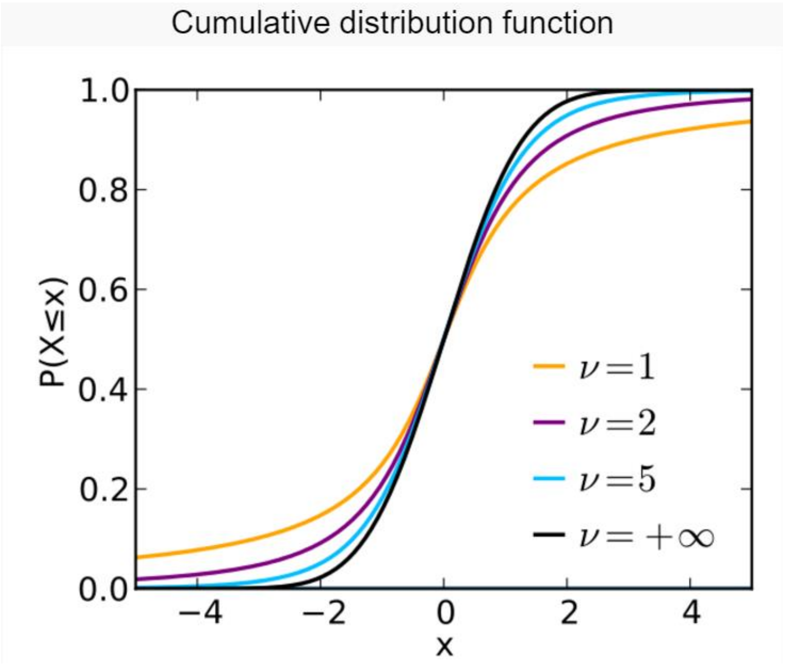
- Using s and $\bar{\mathrm{X}}$ we define the (zero mean) t-statistic to be:

$$t = \frac{\bar{X}}{s/\sqrt{n}}$$

Paired t-test (4)

- The zero mean t-statistic is $t = \frac{\bar{X}}{s/\sqrt{n}}$
- The distribution of the t-statistic t is the student-t distribution with $\nu = n - 1$ degrees of freedom.



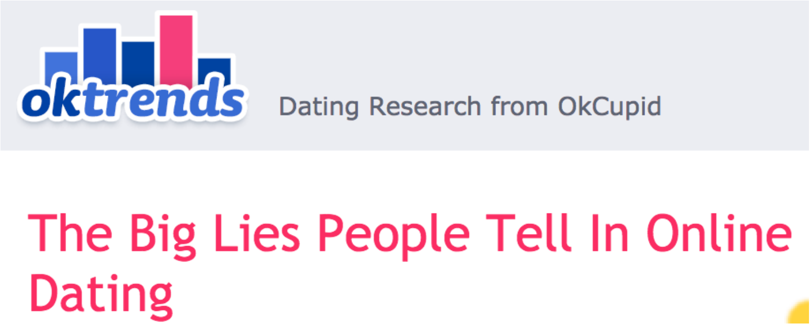


As ν increases, the student t-distribution converges to the standard normal distribution.

Other variants of the t-statistic

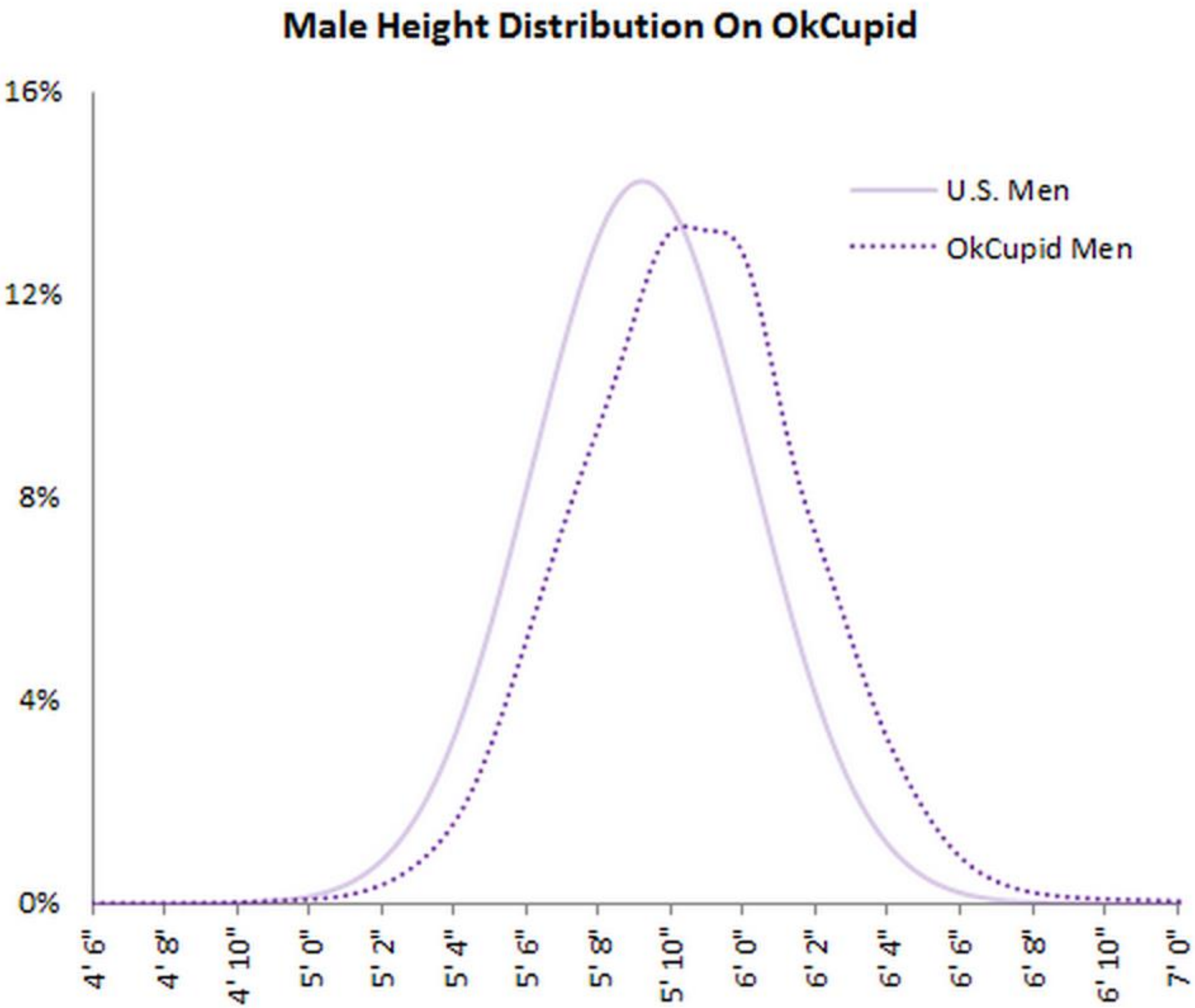
- Under the null hypothesis the mean of the distribution is $\mu \neq 0$
  - Example: Do men tend to inflate their height when writing match-making ads?
- The two sample t-statistic: compare the expected value of two populations (unpaired)
  - Example: we want to compare the effect of two treatments for a disease and it is not possible to pair – not possible to try both treatments on the same individual.
  - Pooled variance: we assume that the variance of the two populations is the same.
  - Un-pooled variance, we don't assume that the variances are the same.

Example of single population t-statistic



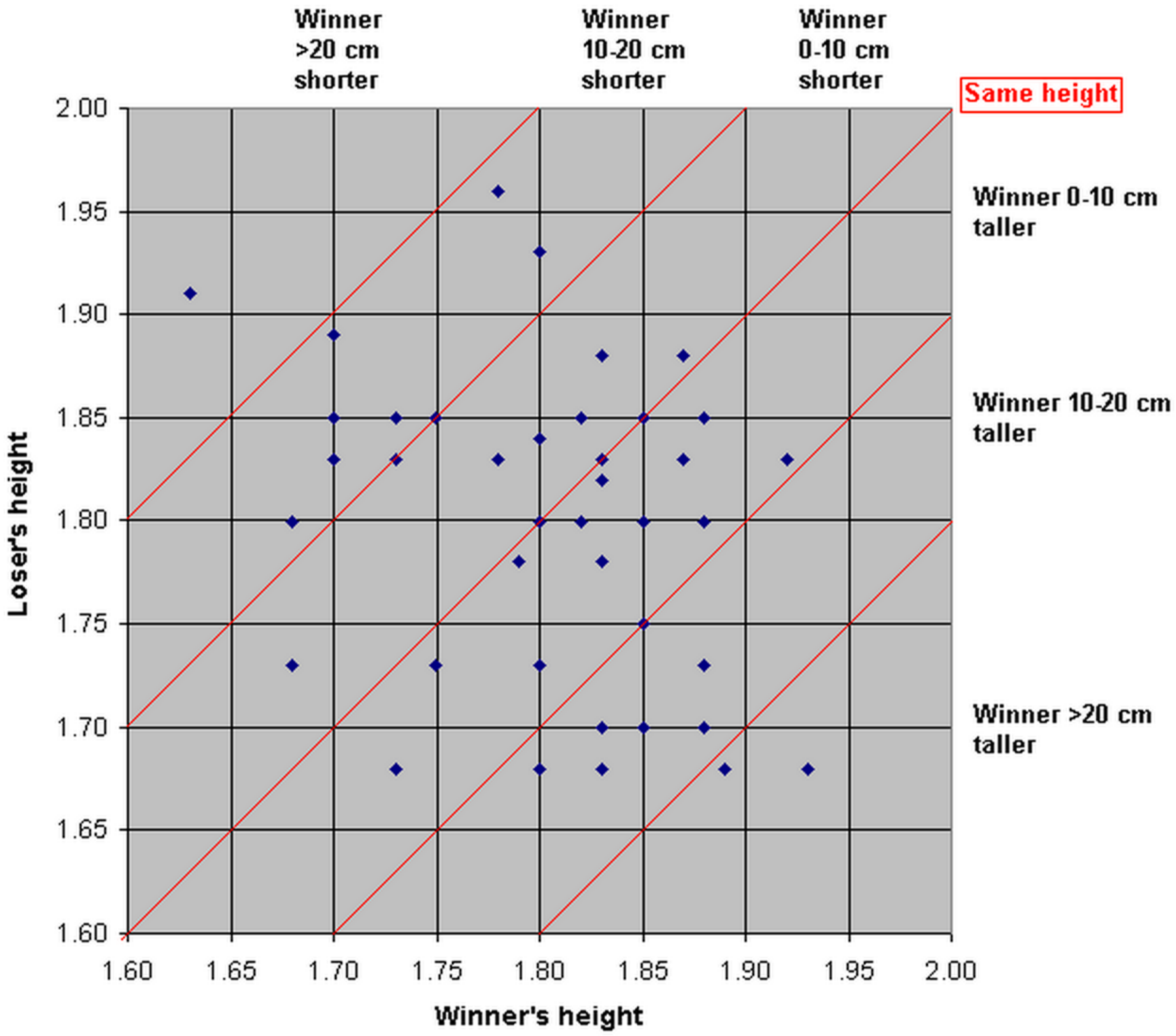
oktrends Dating Research from OkCupid

The Big Lies People Tell In Online Dating



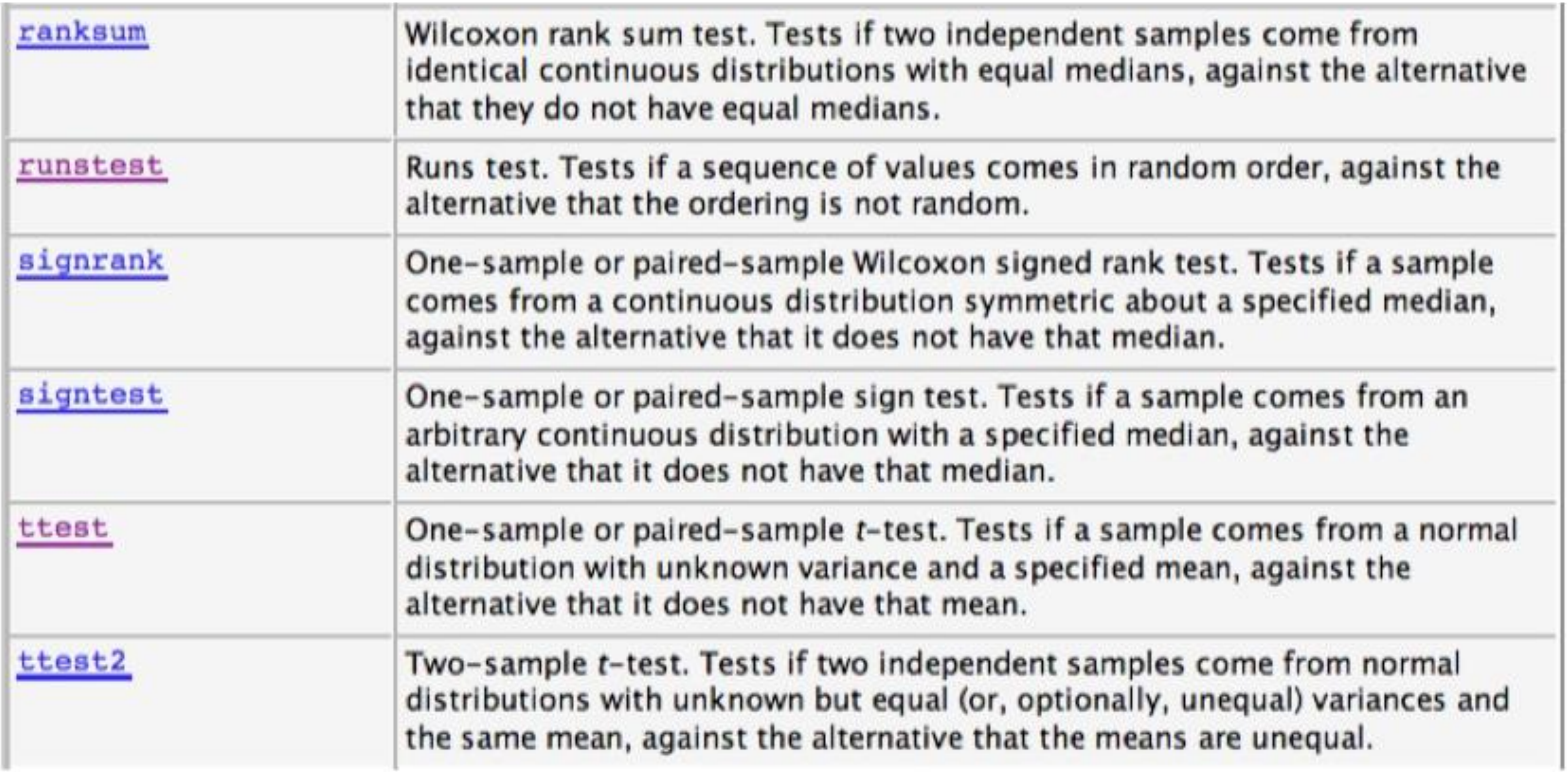
Example of a paired t-statistic

Question: are winners in presidential elections taller than their opponent?

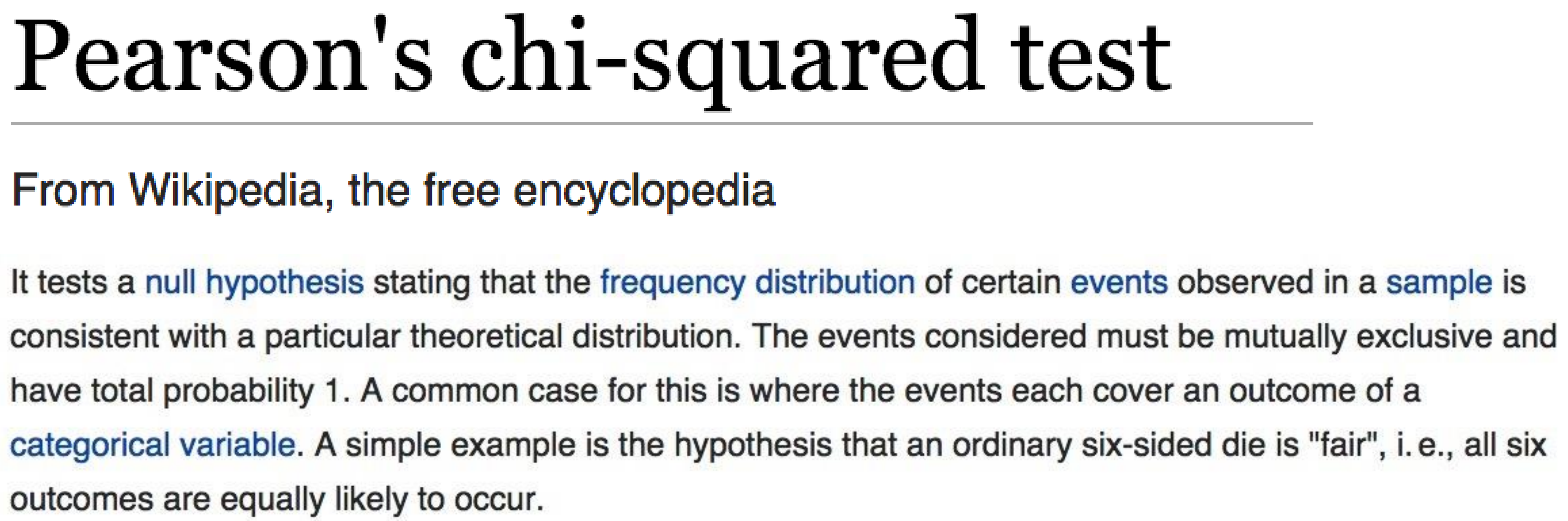


Examples of common statistical tests

From the matlab Statistics module



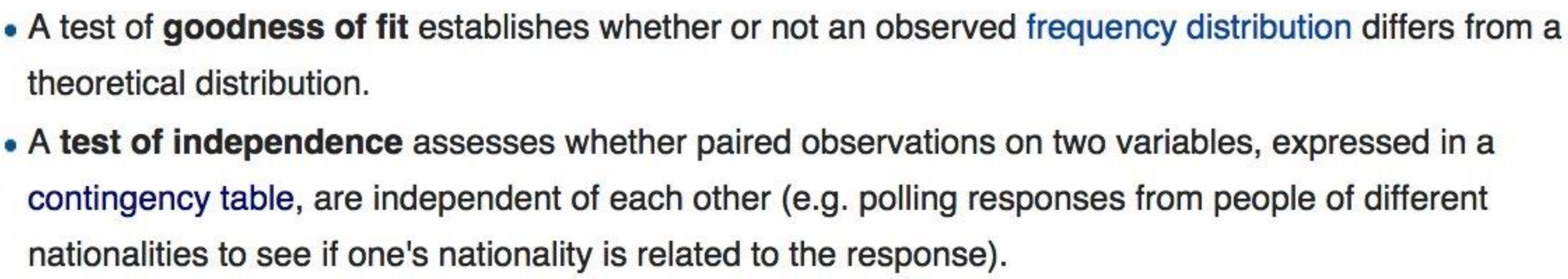
ranksum	Wilcoxon rank sum test. Tests if two independent samples come from identical continuous distributions with equal medians, against the alternative that they do not have equal medians.
runstest	Runs test. Tests if a sequence of values comes in random order, against the alternative that the ordering is not random.
signrank	One-sample or paired-sample Wilcoxon signed rank test. Tests if a sample comes from a continuous distribution symmetric about a specified median, against the alternative that it does not have that median.
signtest	One-sample or paired-sample sign test. Tests if a sample comes from an arbitrary continuous distribution with a specified median, against the alternative that it does not have that median.
ttest	One-sample or paired-sample *t*-test. Tests if a sample comes from a normal distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.
ttest2	Two-sample *t*-test. Tests if two independent samples come from normal distributions with unknown but equal (or, optionally, unequal) variances and the same mean, against the alternative that the means are unequal.



Pearson's chi-squared test

From Wikipedia, the free encyclopedia

It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair", i. e., all six outcomes are equally likely to occur.



- A test of **goodness of fit** establishes whether or not an observed frequency distribution differs from a theoretical distribution.
- A **test of independence** assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

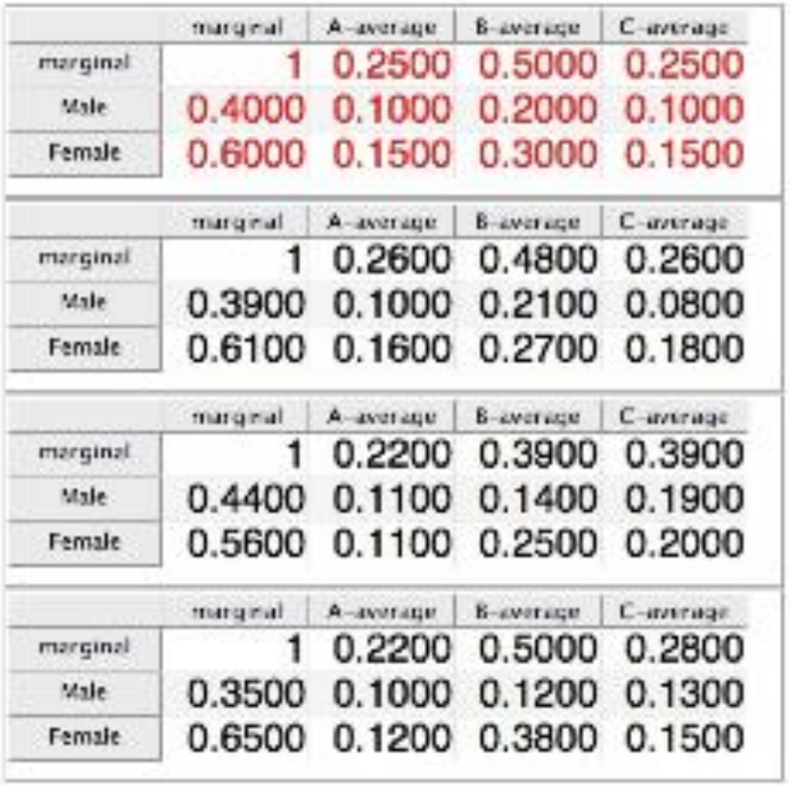
Testing for dependence

- We have two discrete random variables
- True contingency table: the true probability for each cell.
- Empirical contingency table: the fraction of the observations that fall in each in each cell.
- Null Hypothesis: the two variables are independent.
- We can use the chi-square test

Empirical contingency tables

n=100
std~1/10

independent



true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2600	0.4800	0.2600
Male	0.3900	0.1000	0.2100	0.0800
Female	0.6100	0.1600	0.2700	0.1800

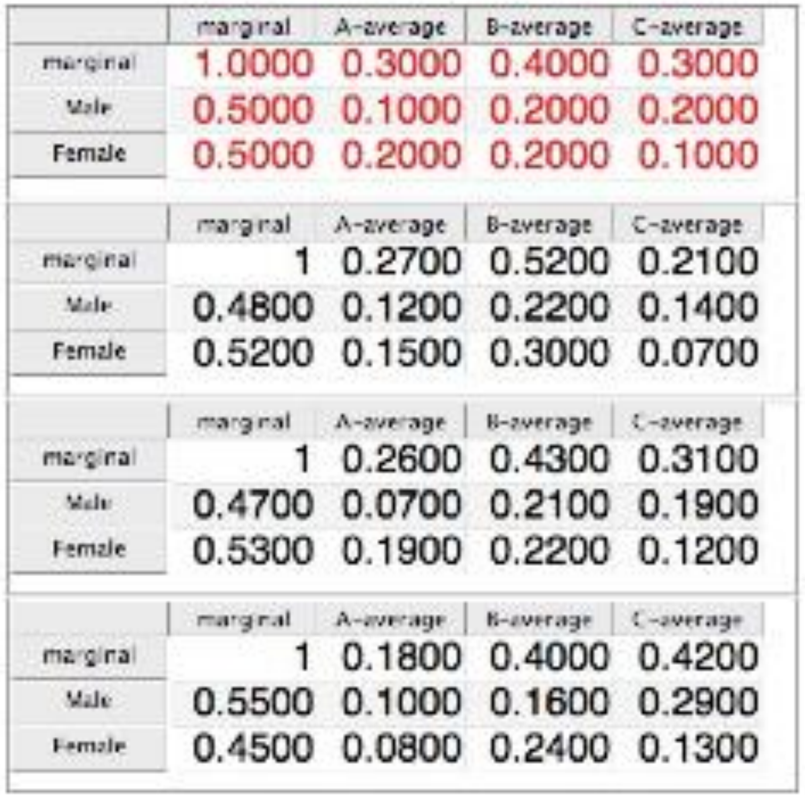
empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2200	0.3900	0.3900
Male	0.4400	0.1100	0.1400	0.1900
Female	0.5600	0.1100	0.2500	0.2000

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2200	0.5000	0.2800
Male	0.3500	0.1000	0.1200	0.1300
Female	0.6500	0.1200	0.3800	0.1500

dependent



	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1	0.2700	0.5200	0.2100
Male	0.4800	0.1200	0.2200	0.1400
Female	0.5200	0.1500	0.3000	0.0700

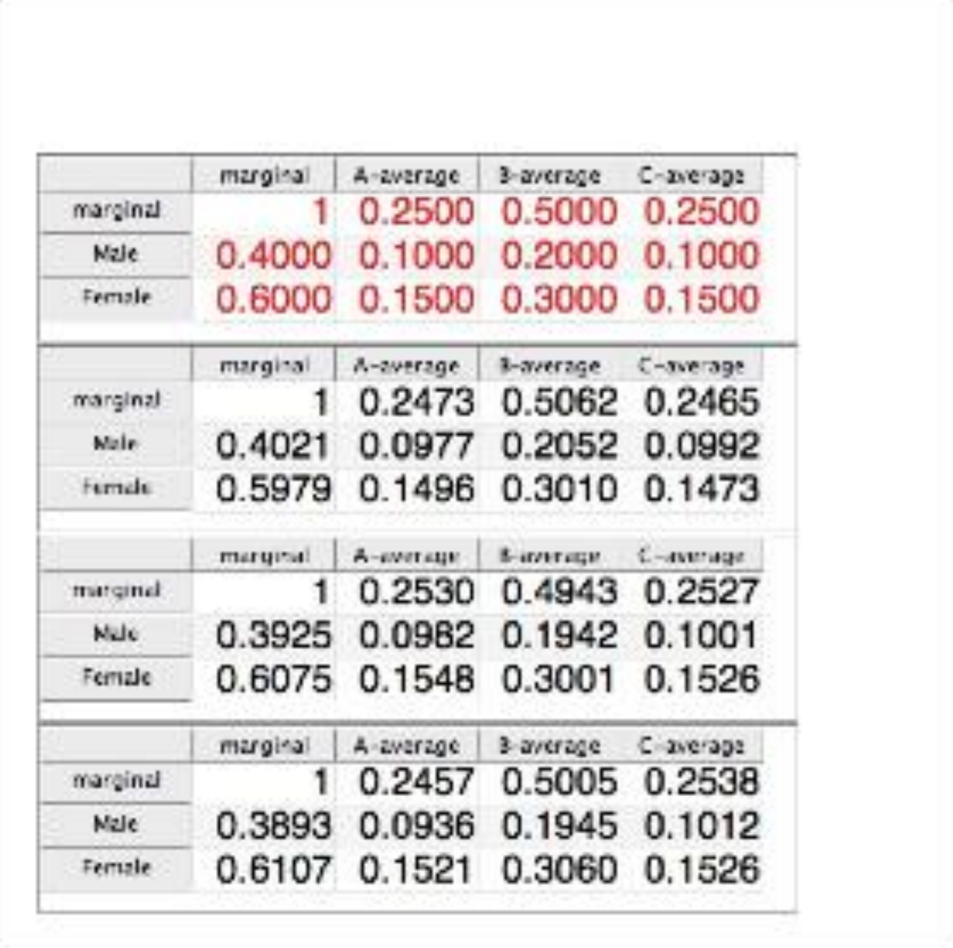
	marginal	A-average	B-average	C-average
marginal	1	0.2600	0.4300	0.3100
Male	0.4700	0.0700	0.2100	0.1900
Female	0.5300	0.1900	0.2200	0.1200

	marginal	A-average	B-average	C-average
marginal	1	0.1800	0.4000	0.4200
Male	0.5500	0.1000	0.1600	0.2900
Female	0.4500	0.0800	0.2400	0.1300

Empirical contingency tables

n=10,000
std~1/100

independent



true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2473	0.5062	0.2465
Male	0.4021	0.0977	0.2052	0.0992
Female	0.5979	0.1496	0.3010	0.1473

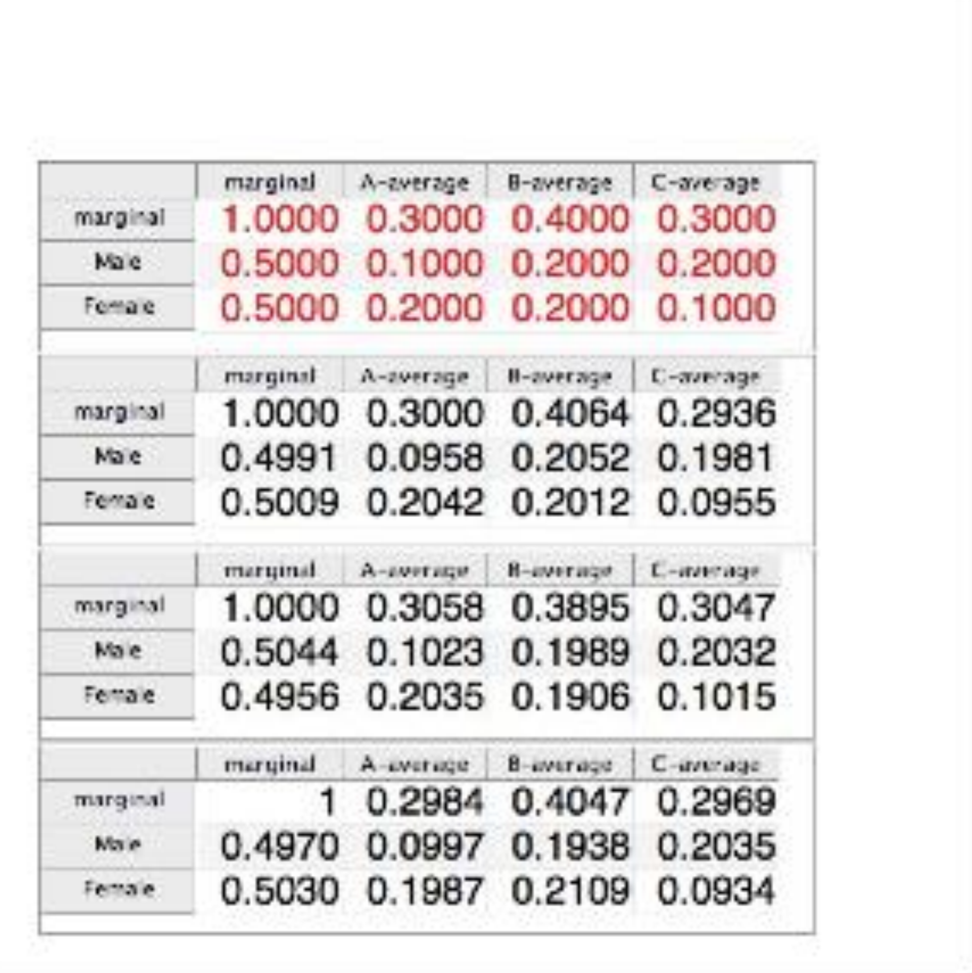
empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2530	0.4943	0.2527
Male	0.3925	0.0982	0.1942	0.1001
Female	0.6075	0.1548	0.3001	0.1526

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2457	0.5005	0.2538
Male	0.3893	0.0936	0.1945	0.1012
Female	0.6107	0.1521	0.3060	0.1526

dependent



	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4064	0.2936
Male	0.4991	0.0958	0.2052	0.1981
Female	0.5009	0.2042	0.2012	0.0955

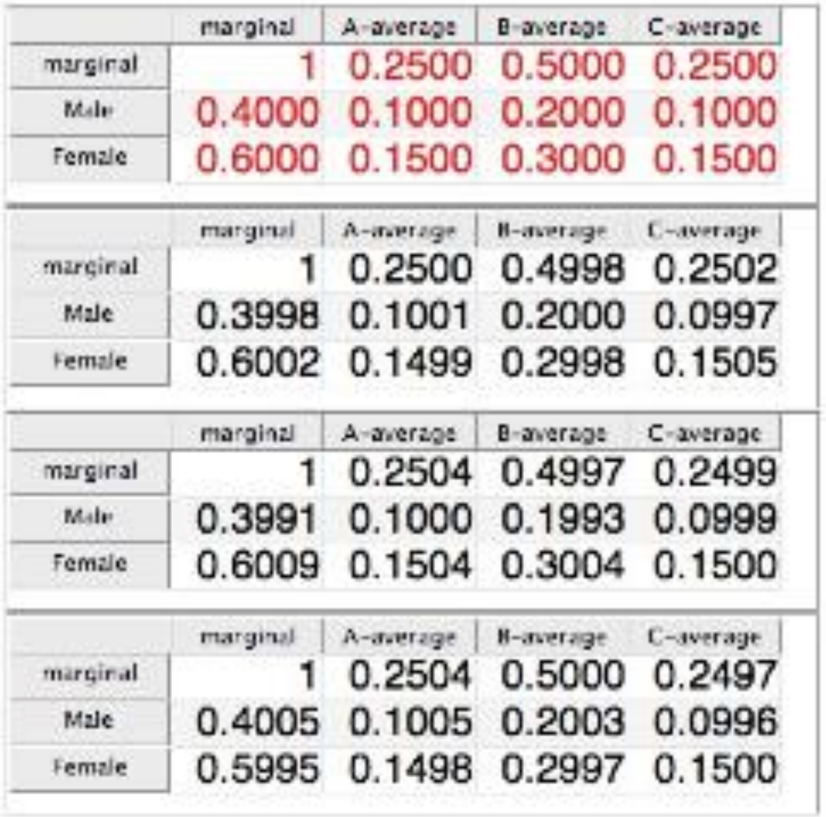
	marginal	A-average	B-average	C-average
marginal	1.0000	0.3058	0.3895	0.3047
Male	0.5044	0.1023	0.1989	0.2032
Female	0.4956	0.2035	0.1906	0.1015

	marginal	A-average	B-average	C-average
marginal	1	0.2984	0.4047	0.2969
Male	0.4970	0.0997	0.1938	0.2035
Female	0.5030	0.1987	0.2109	0.0934

Empirical contingency tables

n=1,000,000
std~1/1,000

independent



true dist

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.5000	0.2500
Male	0.4000	0.1000	0.2000	0.1000
Female	0.6000	0.1500	0.3000	0.1500

empirical dist 1

	marginal	A-average	B-average	C-average
marginal	1	0.2500	0.4998	0.2502
Male	0.3998	0.1001	0.2000	0.0997
Female	0.6002	0.1499	0.2998	0.1505

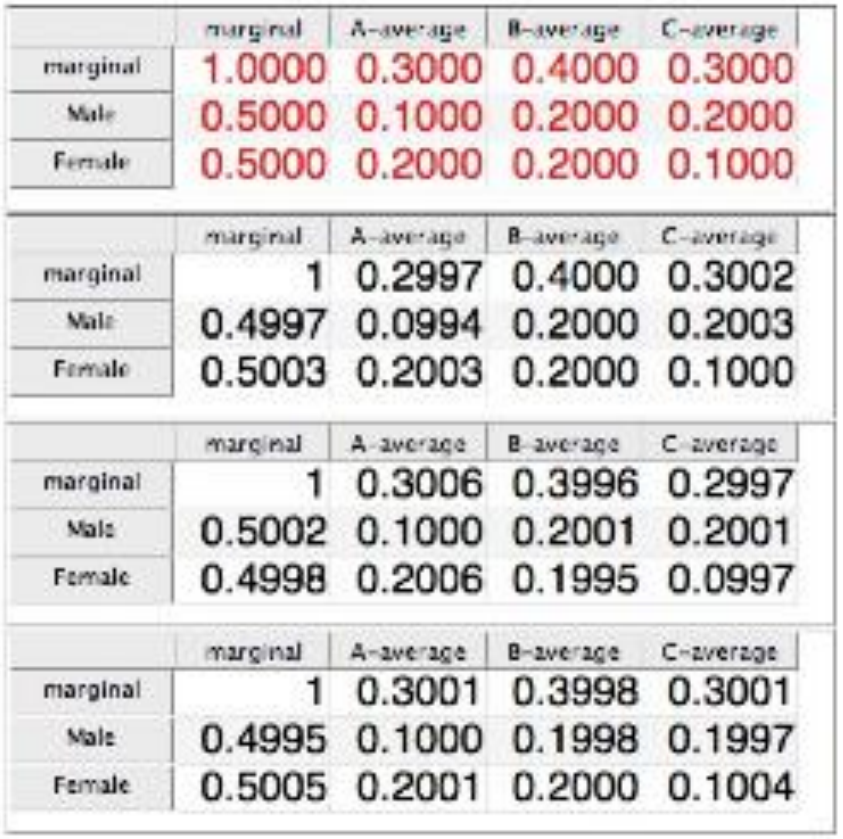
empirical dist 2

	marginal	A-average	B-average	C-average
marginal	1	0.2504	0.4997	0.2499
Male	0.3991	0.1000	0.1993	0.0999
Female	0.6009	0.1504	0.3004	0.1500

empirical dist 3

	marginal	A-average	B-average	C-average
marginal	1	0.2504	0.5000	0.2497
Male	0.4005	0.1005	0.2003	0.0996
Female	0.5995	0.1498	0.2997	0.1500

dependent



	marginal	A-average	B-average	C-average
marginal	1.0000	0.3000	0.4000	0.3000
Male	0.5000	0.1000	0.2000	0.2000
Female	0.5000	0.2000	0.2000	0.1000

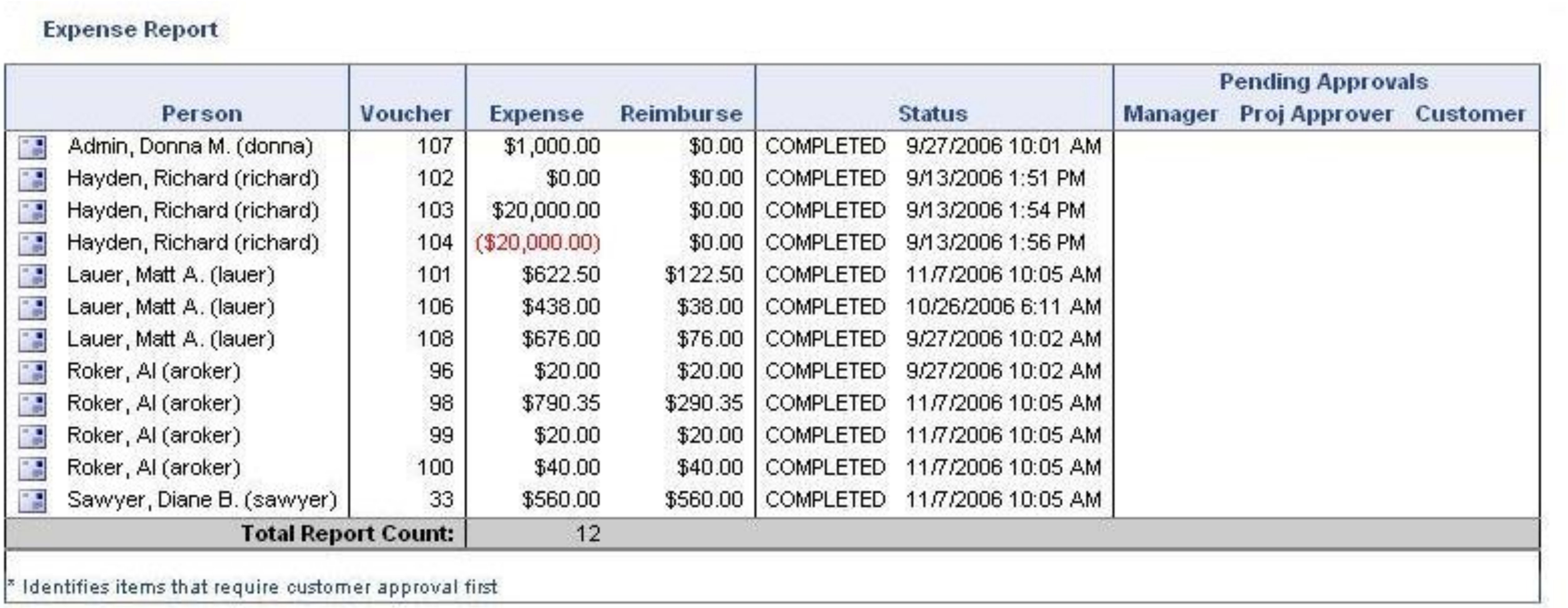
	marginal	A-average	B-average	C-average
marginal	1	0.2997	0.4000	0.3002
Male	0.4997	0.0994	0.2000	0.2003
Female	0.5003	0.2003	0.2000	0.1000

	marginal	A-average	B-average	C-average
marginal	1	0.3006	0.3996	0.2997
Male	0.5002	0.1000	0.2001	0.2001
Female	0.4998	0.2006	0.1995	0.0997

	marginal	A-average	B-average	C-average
marginal	1	0.3001	0.3998	0.3001
Male	0.4995	0.1000	0.1998	0.1997
Female	0.5005	0.2001	0.2000	0.1004

Finding cheaters using the Bedford distribution.

- Suppose we have a large accounting document, with thousands of numbers in it.
- Can we detect if somebody “cooked the books” just by looking at the distribution of the numbers?
- Consider the distribution of the most significant (non zero) digit:



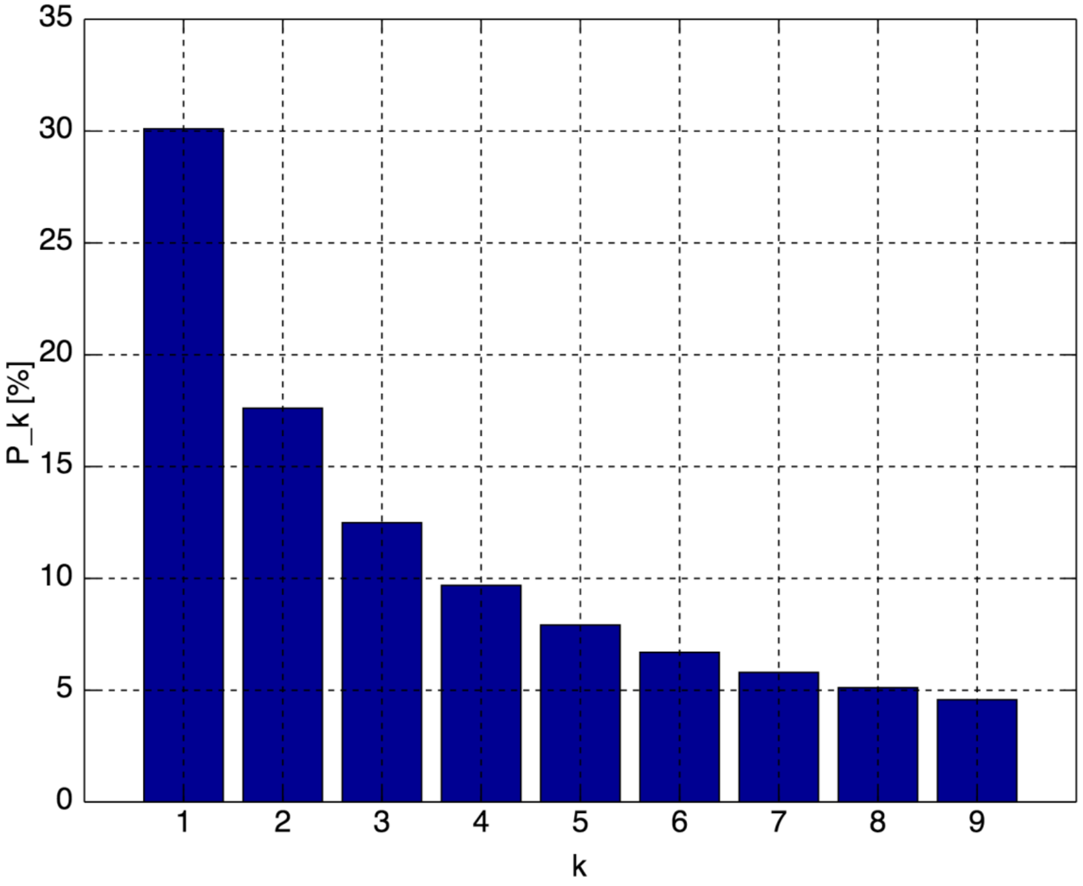
Expense Report

						Pending Approvals		
Person	Voucher	Expense	Reimburse	Status		Manager	Proj Approver	Customer
Admin, Donna M. (donna)	107	$1,000.00	$0.00	COMPLETED	9/27/2006 10:01 AM			
Hayden, Richard (richard)	102	$0.00	$0.00	COMPLETED	9/13/2006 1:51 PM			
Hayden, Richard (richard)	103	$20,000.00	$0.00	COMPLETED	9/13/2006 1:54 PM			
Hayden, Richard (richard)	104	($20,000.00)	$0.00	COMPLETED	9/13/2006 1:56 PM			
Lauer, Matt A. (lauer)	101	$622.50	$122.50	COMPLETED	11/7/2006 10:05 AM			
Lauer, Matt A. (lauer)	106	$438.00	$38.00	COMPLETED	10/26/2006 6:11 AM			
Lauer, Matt A. (lauer)	108	$676.00	$76.00	COMPLETED	9/27/2006 10:02 AM			
Roker, Al (aroker)	96	$20.00	$20.00	COMPLETED	9/27/2006 10:02 AM			
Roker, Al (aroker)	98	$790.35	$290.35	COMPLETED	11/7/2006 10:05 AM			
Roker, Al (aroker)	99	$20.00	$20.00	COMPLETED	11/7/2006 10:05 AM			
Roker, Al (aroker)	100	$40.00	$40.00	COMPLETED	11/7/2006 10:05 AM			
Sawyer, Diane B. (sawyer)	33	$560.00	$560.00	COMPLETED	11/7/2006 10:05 AM			
	Total Report Count:	12						

* Identifies items that require customer approval first

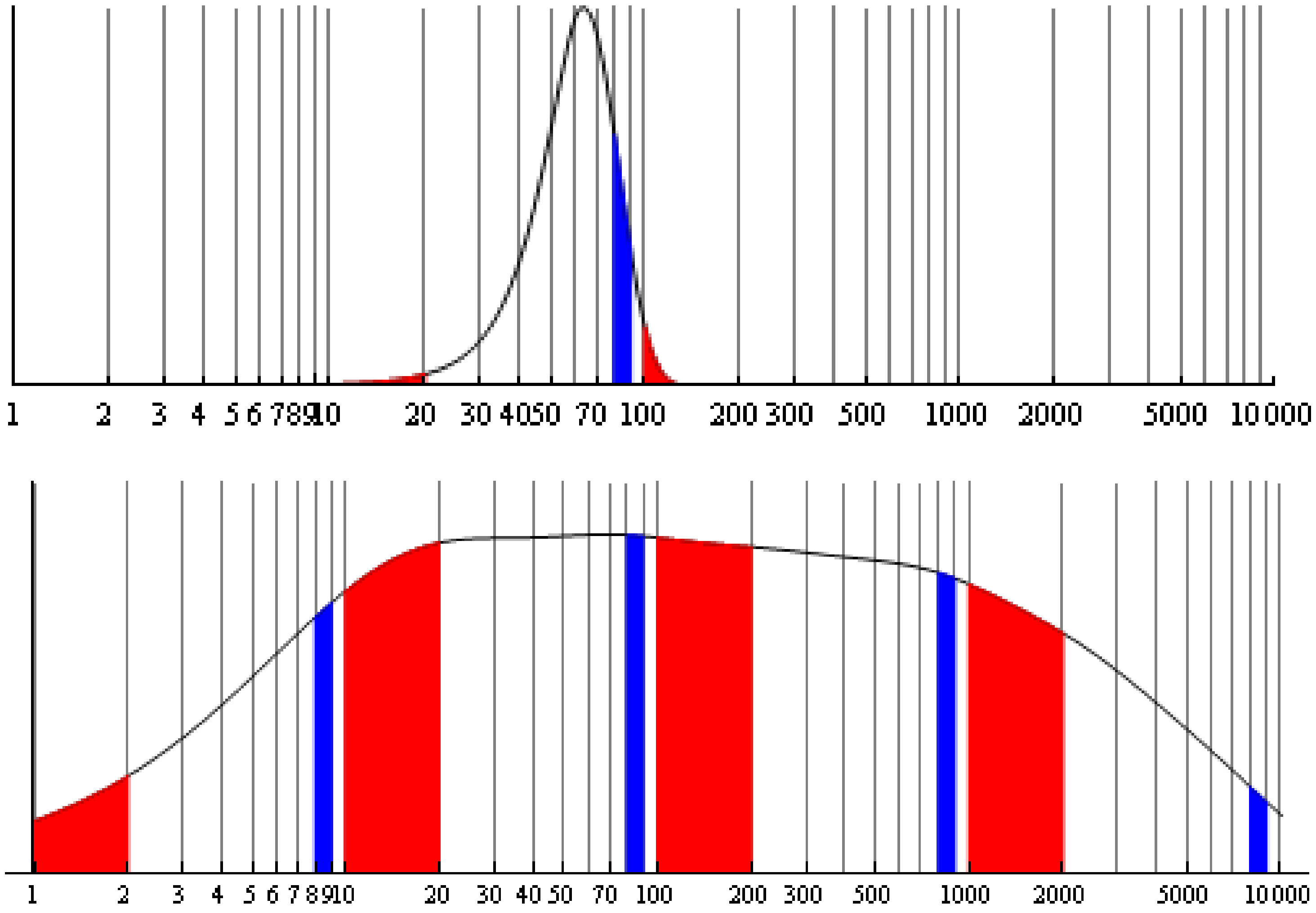
The Benford distribution

Describes the distribution of the most significant digit in large collection of financial numbers



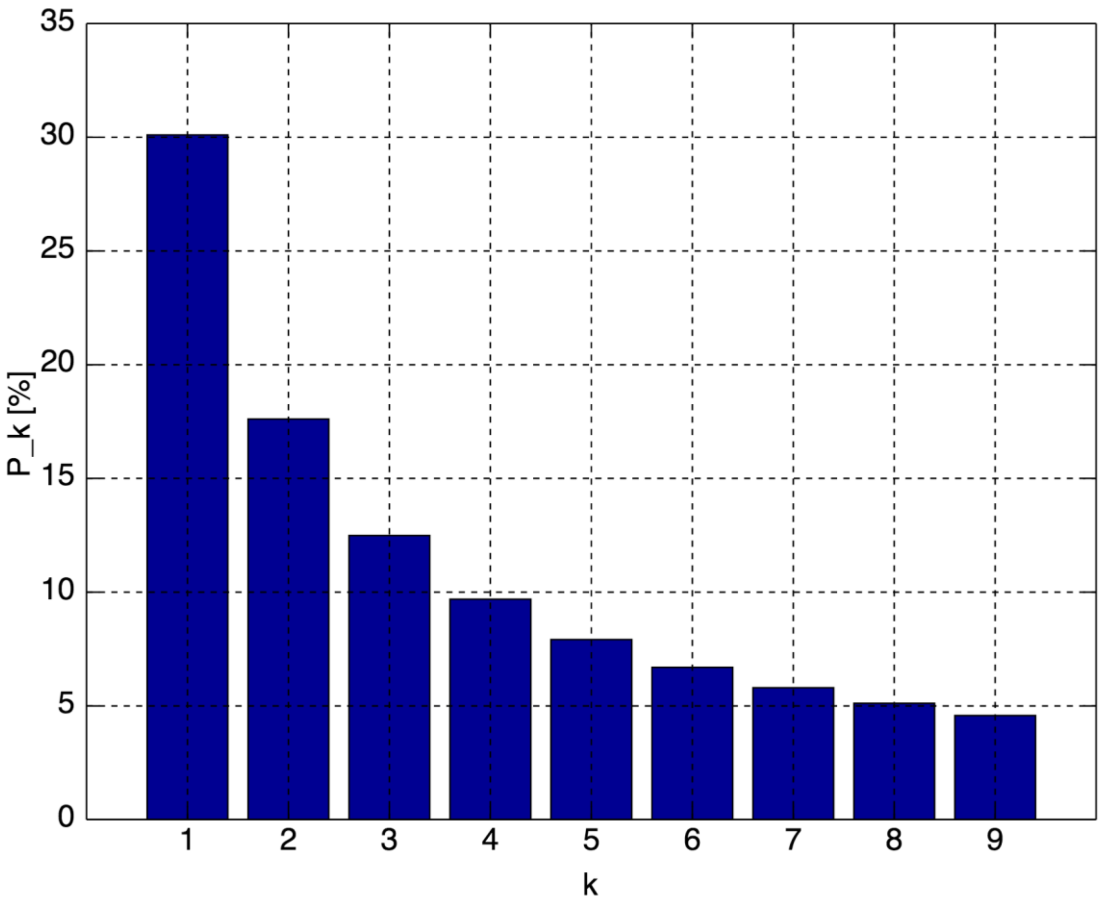
How can we explain Benford law?

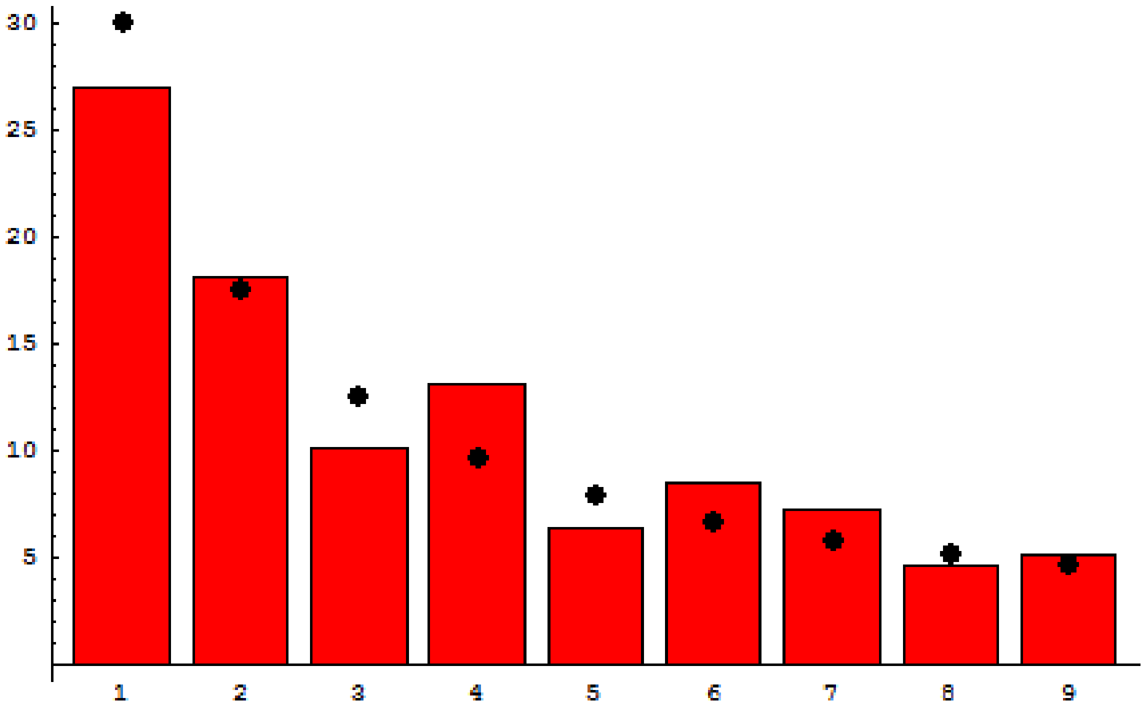
- As currency units are arbitrary, changing the definition of a the currency unit does not fundamentally change the distribution.
- The distribution is approximately constant on a logarithmic scale.
- If the distribution spans several orders of magnitude (from single dollars to thousands of dollars) we get the Benford distribution



Can we detect accounting fraud using the Benford distribution?

Null Hyp: dist is Benford





Distribution of top digits in a tax return

Multiple Hypothesis testing

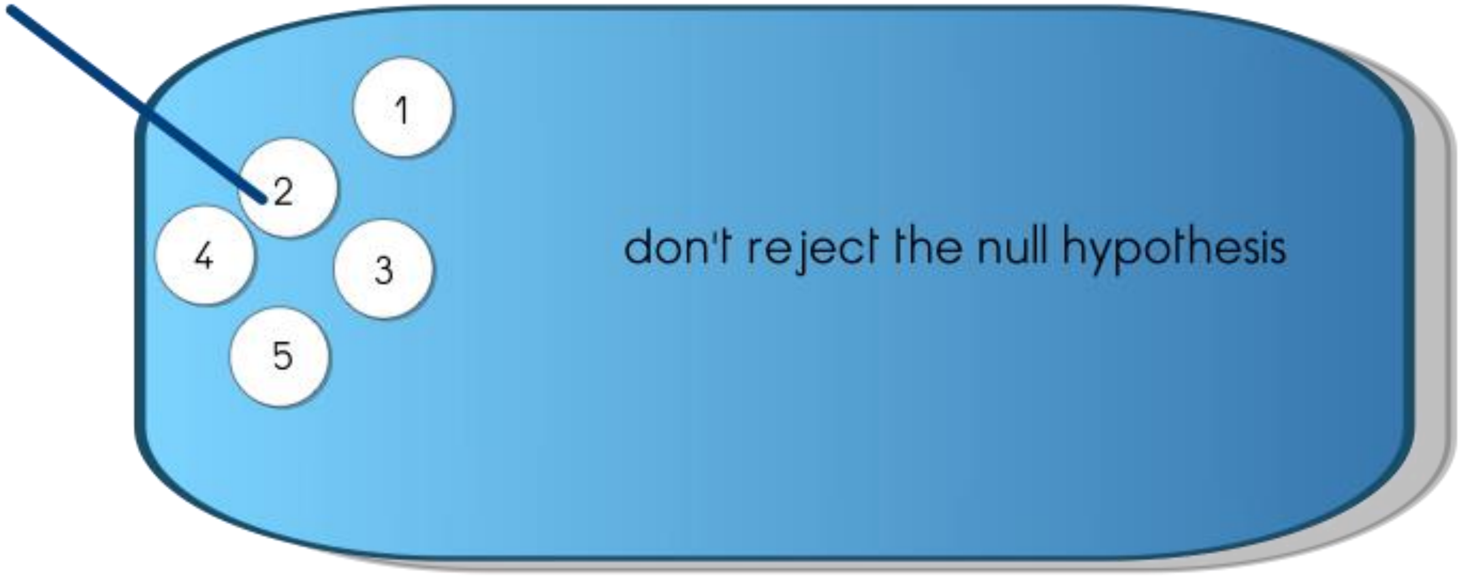
Consider the online ad problem, our goal is to maximize click-through rate. Our null hypothesis is that nothing performs better than picking one of the ads uniformly at random each time.

We have a large number of click-prediction algorithms. Each such algorithm takes as input information about the person, the web page and the ad and predicts the probability that the person will click on the ad.

We can go back in time and compute the expected number of errors each method would have made. We can use a statistical test to quantify the statistical significance of the performance of the method.

Suppose we have 100 methods and use an alpha value of 1%
Suppose for our data we found that one of the 100 methods rejects the null hypothesis at the 1% significance level. How sure can we be that the predictor that we found is better than random?

The probabilty theory of statistical tests

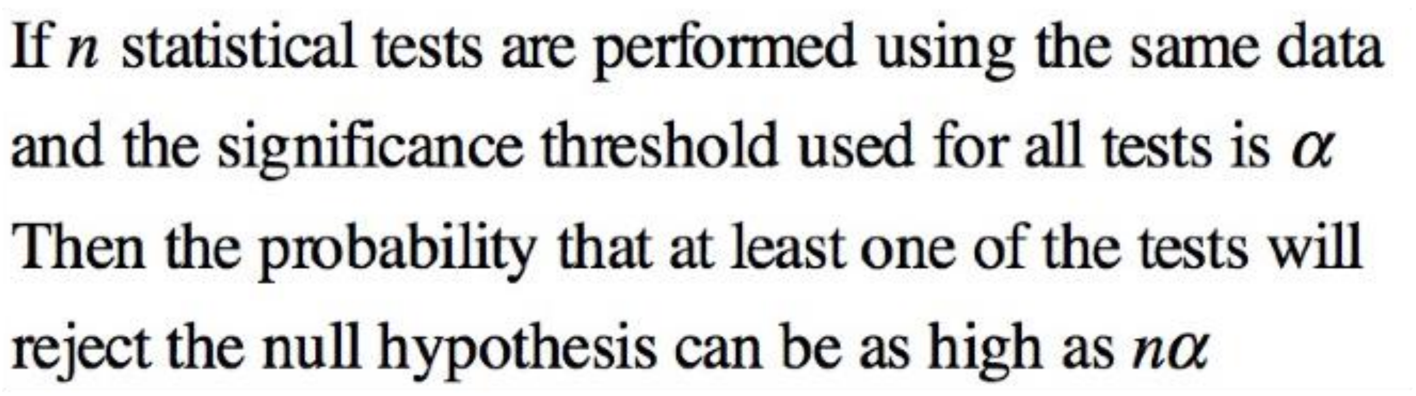


rejection set
reject the null hypothesis
for predictor i

Omega=outcome space

We don't know what would happen of different samples than the one we observe.
In the worst case the rejection sets are disjoint.

The Bonferroni correction for multiple-hypothesis testing:



If n statistical tests are performed using the same data and the significance threshold used for all tests is α Then the probability that at least one of the tests will reject the null hypothesis can be as high as $n\alpha$

Be a skeptic:

When you read that something has been proven statistically:

1. Ask what was the null hypothesis
2. Ask what was the statistical significance
3. Ask whether similar tests were performed that did not succeed.