Topic III: Significance Testing

Discrete Topics in Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2012/13

T III: Significance Testing

1. Hypothesis Testing
    1.1. Null Hypotheses and *p*-values
    1.2. Parametric Tests
    1.3. Exact Tests
2. Significance and Data Mining
    2.1. Why? How?
3. Significance for a Frequency Threshold
4. Course Feedback Feedback

Hypothesis testing

- Suppose we throw a coin n times and we want to estimate if the coin is fair, i.e. if Pr(heads) = Pr(tails).
- Let $X_1, X_2, \dots, X_n \sim$ Bernoulli(p) be the i.i.d. coin flips
  - Coin is fair $\Leftrightarrow p = 1/2$
- Let the **null hypothesis** H_0 be "coin is fair".
- The **alternative hypothesis** H_1 is then "coin is not fair"
- Intuitively, if $|n^{-1}\sum_i X_i - 1/2|$ is large, we should reject the null hypothesis
- *But can we formalize this?*

Hypothesis testing terminology

- $\theta = \theta_0$ is called **simple hypothesis**
- $\theta > \theta_0$ or $\theta < \theta_0$ is called **composite hypothesis**
- H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$ is called **two-sided test**
- H_0: $\theta \leq \theta_0$ vs. H_1: $\theta > \theta_0$ and H_0: $\theta \geq \theta_0$ vs. H_1: $\theta < \theta_0$ are called **one-sided tests**
- **Rejection region** R: if $X \in R$, reject H_0 o/w retain H_0
  - Typically $R = \{x : T(x) > c\}$ where T is a **test statistic** and c is a **critical value**
- **Error types:**

	Retain H_0	Reject H_0
H_0 true	✓	type I error
H_1 true	type II error	✓

The *p*-values

- The *p*-value is the *probability that **if** H_0 **holds**, we observe values at least as extreme as the test statistic*
  - It is *not* the probability that H_0 holds
  - If *p*-value is small enough, we can reject H_0
  - How small is small enough depends on application
- Typical *p*-value scale:

***p*-value**	**evidence**
< 0.01	very strong evidence against H_0
0.01–0.05	strong evidence against H_0
0.05–0.1	weak evidence against H_0
> 0.1	little or no evidence against H_0

Statistical Power

- The **power** of the test is the probability that it will reject the null hypothesis when it is false
  - If the rate of Type II errors is β, the power is $1 - \beta$
- At least three factors have effect to power:
  - Significance level
    - Higher significance $\Rightarrow$ lesser power
  - Magnitude of the effect
    - How "far" we are from the null hypothesis
  - Sample size

The Wald test

For two-sided test H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$

Test statistic $W = \frac{\hat{\theta} - \theta_0}{\hat{\text{se}}}$, where $\hat{\theta}$ is the sample estimate and

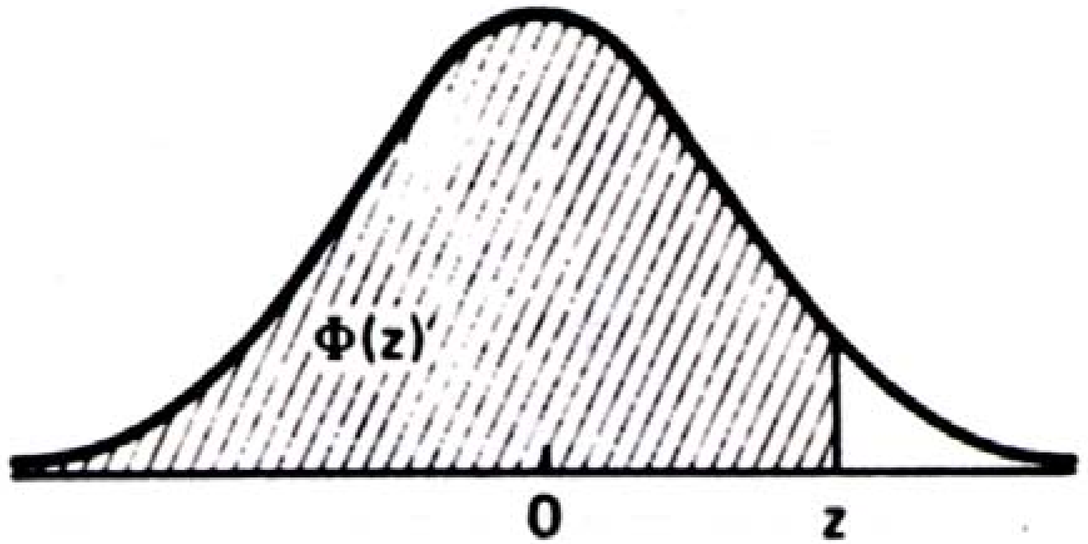
$\hat{\text{se}} = \text{se}(\hat{\theta}) = \sqrt{\text{Var}[\hat{\theta}]}$ is the standard error.

W converges in probability to N(0,1).

If w is the observed value of Wald statistic, the p-value is $2\Phi(-|w|)$.

The coin-tossing example revisited

Using Wald test we can test if our coin is fair. Suppose the observed average is 0.6 with estimated standard error 0.049. The observed Wald statistic w is now $w = (0.6 - 0.5)/0.049 \approx 2.04$. Therefore the p-value is $2\Phi(-2.04) \approx 0.041$, and we have strong evidence to reject the null hypothesis.



Confidence Intervals

- Suppose have a statistical test to test null hypothesis $\theta = \theta_0$ at significance α for any value of θ_0
- The **confidence interval** of θ at confidence level $1 - \alpha$ is the interval $[x, y] \ni \theta$ if null hypothesis $\theta = \theta_0$ is *retained* at significance α for any $\theta_0 \in [x, y]$
  - There are other ways to define/compute confidence intervals

Parametric Tests

- Many statistical tests assume we can express (or approximate) the null hypothesis distribution in closed form
  - Normal distribution, Poisson distribution, Weibull distribution…
  - Test if data is normally distributed
  - Test if two samples are from independent distributions
    - The test statistics approaches χ^2 distribution
- This simplifies the calculations
  - But most parametric tests are not **exact** because the distributions hold only asymptotically

Exact Tests

- Exact test give exact *p*-values
  - No asymptotics
- Usually more time consuming to compute
- Used mostly with smaller samples
  - Faster to compute
  - Parametric tests behave badly
- Can (sometimes) be used when no parametric probability distribution is known

Permutation Test

- Suppose we have two samples of numbers
  - $x_1, x_2, \ldots, x_n$, and $y_1, y_2, \ldots, y_m$ with means $\bar{x}$ and $\bar{y}$
- The null hypothesis is $\bar{x} = \bar{y}$ (two-sided test)
- First we compute $T(obs) = |\bar{x} - \bar{y}|$
- We pool x's and y's together and create every possible partition of the values into sets of size n and m
  - We compute the means and their absolute difference
  - There are $\binom{n+m}{n}$ such partitions
- The p-value is the fraction of partition with same or higher absolute difference of means

Significance and Data Mining

- Hypothesis testing is *confirmatory data analysis*
  - Data mining is *exploratory data analysis*
- But data mining can still use (or need) statistical significance testing
  - While the hypothesis is (partially) created by an algorithm, the significance of the findings still need to be validated
- For example, finding many frequent itemsets is
  - Surprising, if the data is rather sparse
  - Expected, if the data is rather dense

An Example

- Suppose we have found a frequent itemsets with size s and frequency f from data D that has k 1s
- Is this finding significant?
  - Let's assume the values in D are independent
  - We can create all possible data matrices D' of same size and density
  - We can compute from how of these data we find an itemset with same size and same or higher frequency
    - Or we can compute in how many of these data *this* itemset has same or better frequency
  - This gives us a p-value
    - Or does it?

Problem 1: Too Many Datasets

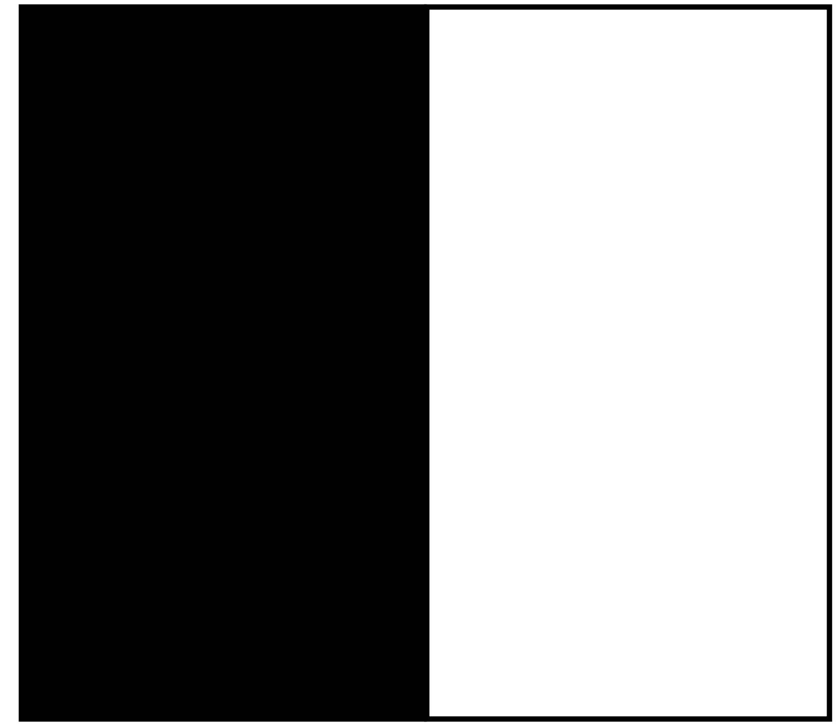
- Assuming we have n items, m transactions, and k ($\leq nm$) 1s in the data, we have $\binom{nm}{k}$ possible datasets
  - We cannot try all
- Solution 1: we can sample and estimate the p-value
  - How big a sample we need depends on how small a p-value we want
- Solution 2: we can create a parametric distribution to estimate the p-value
  - Considerably more complex

Problem 2: Multi-Hypothesis Testing

- We are actually testing whether *any* of the $\binom{n}{s}$ itemsets of size s has significant support
  - This is much more likely than just one of them having that support
  - For example, if $s = 2$, $f = 7/m$, $n = 1k$, $m = 1M$, and every item appears in every transaction with probability 1/1000 (i.i.d.)
    - Probability for any such 2-itemset is ≈ 0.0001
    - But there are ≈ 0.5M of such 2-itemsets
    - Each random data should have ≈ 50 such 2-itemsets
- Solution: *Bonferroni correction*; divide the p-value with the number of simultaneous tests
  - Very low power; lots of false negatives
  - Requires even more samples

Problem 3: The Independence

- The values are rarely completely independent
  - The independence assumption might omit very trivial structure
  - E.g. some items are more popular than others
    - These are more likely to form a frequent itemset
- We need stronger null hypothesis
  - But how to test that…



Significance for Frequency Threshold

- **Question.** How frequent should a k-itemset be for it to be significant?
- **Null model.** Random data set of same size with same expected item frequencies
  - If item i has frequency f_i, then in the random model the item appears in each transaction independently with probability f_i
    - Every column of the matrix is m i.i.d. Bernoulli samples with parameter f_i
- No need to do the frequent itemset mining on (too) many random data sets

Kirsch et al. 2012

Poisson Distribution

- One parameter: λ
  - Rate of occurrence
- If $X \sim$ Poisson(λ), then $\Pr(X = k) = \lambda^k e^{-\lambda}/k!$
  - $E[X] = \lambda$
- Models number of occurrences among a large set of possible events, where the probability of each event is small
  - “Law of rare events”

The Main Idea

- Let $O_{k,s}$ be the number of observed k-itemsets of support at least s
  - Let $\hat{O}_{k,s}$ be the random variable corresponding to that in a random dataset
- **Theorem.** There exists a level $s_{\min}$ such that if $s \geq s_{\min}$, $\hat{O}_{k,s}$ is approximated well by Poisson distribution
  - With this, we can compute the p-values easily
    - No need for data samples (almost…)
  - Only works with large-enough support levels
    - Rare events

How to Determine $s_{\min}$?

- Let $\varepsilon \in (0,1)$ be a parameter that defines how close to the Poisson we want to be
- Let S be the maximum expected support of k-itemset
  - Product of k largest frequencies times the number of transactions
  - S is a lower bound for $s_{\min}$
- Create Δ random data sets and find from them all k-itemsets of support at least S
  - From these itemsets we can estimate how big the $s_{\min}$ has to be for good approximation of $\hat{O}_{k,s}$ by Poisson
  - Δ depends on how sure we want to be that the approximation really is good (but, say, $\Delta = 1000$)

Controlling False Discovery Rate

- We might still get lots of Type I errors due to multiple-hypothesis testing
  - *False Discovery Rate* (FDR) is the ratio of Type I errors among all rejected null hypotheses
- We want to find a support threshold $s^* \geq s_{\min}$ such that *all* k-itemsets with support $\geq s^*$ are statistically significant with controlled false discovery rate
  - They have confidence higher than $1 - \alpha$ with FDR at most β

Controlling the Confidence

- Try values for s^* starting from $s_0 = s_{\min}$, $s_i = s_{\min} + 2^i$
  - $h = \lfloor \log_2(s_{\max} - s_{\min}) \rfloor + 1$ tests
- The null hypothesis H_0^i is that O_{k,s_i} is drawn from $\hat{O}_{k,s_i}$
  - This is easy to compute *if* we know Poisson parameter λ_i
  - We can estimate λ_i from the same random sample we used to obtain $s_{\min}$ as it is just $\mathrm{E}[\hat{O}_{k,s_i}]$
- Let $\alpha_0, \alpha_1, \ldots, \alpha_{h-1}$ be such that $\sum_i \alpha_i = \alpha$
  - We reject H_0^i if the p-value is smaller than α_i
    - By union bound, all rejections are correct with probability at least $1 - \alpha$
- We select the smallest s_i where H_0^i is rejected

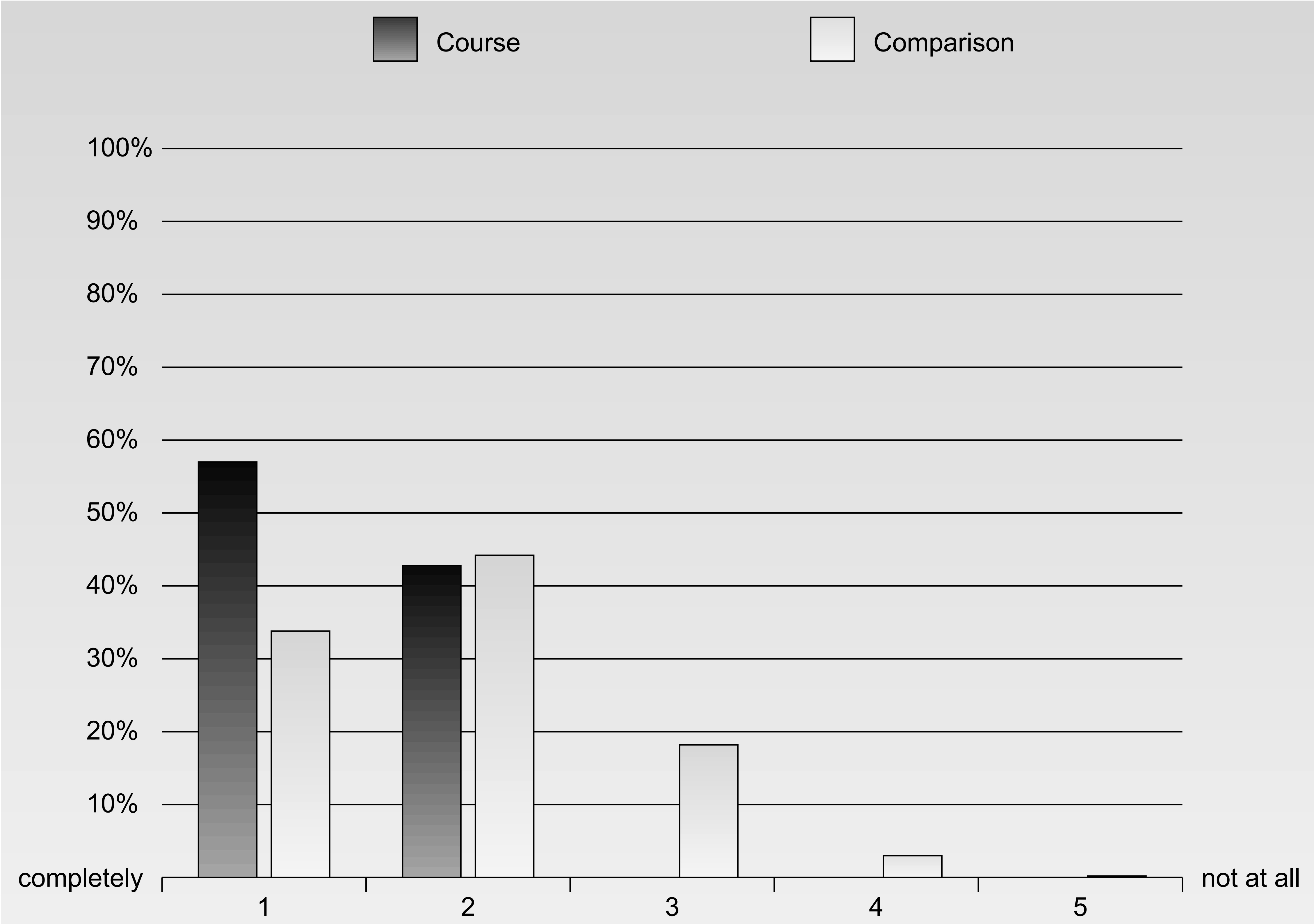
Controlling the FDR

- The first attempt does *not* control FDR
- For that, define $\beta_0, \beta_1, \ldots, \beta_{h-1}$ such that $\sum_i \beta_i^{-1} = \beta$
  - Let $\lambda_i = E[\hat{O}_{k,s_i}]$
  - α_i can just be α/h and ditto for β_i
- Reject H_0^i if p-value of O_{k,s_i} is smaller than α_i *and* $O_{k,s_i} \geq \beta_i\lambda_i$
- **Theorem.** The k-itemsets that are frequent w.r.t. s^* are statistically significant with confidence $1 - \alpha$ with FDR at most β

Summary

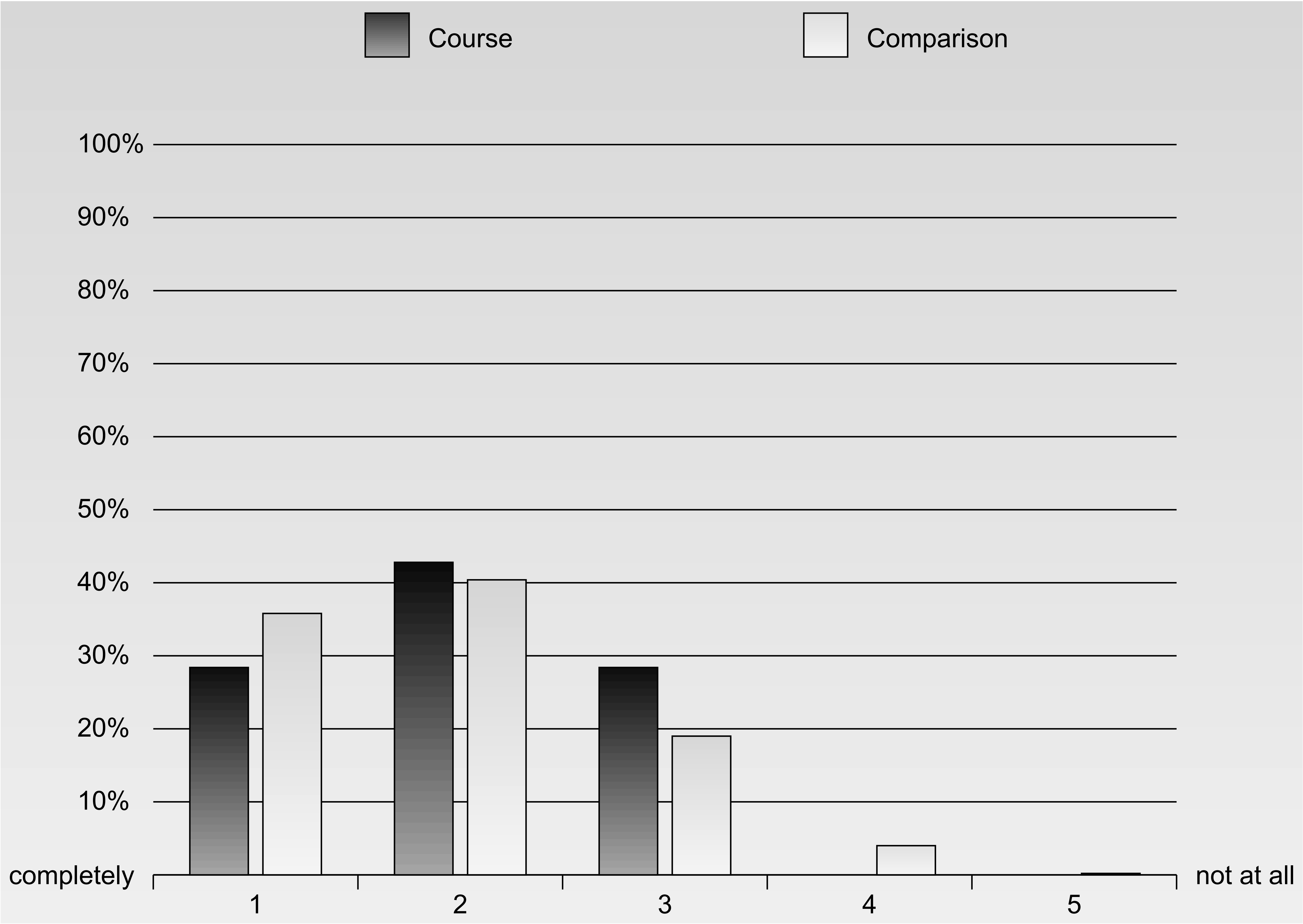
- Given itemset size k, confidence level $1 - \alpha$ and false discovery rate β, we can find minimum support level s^* such that each k-itemset that has support at least s^* is significant with FDR at most β
  - Null hypothesis: each item is i.i.d. Bernoulli with parameter f_i
  - Only works for high values of support
    - Poisson approximation
  - Might return $s^* = \infty$
    - Data cannot be distinguished from random
  - Requires sampling only to estimate parameters

Lecturer



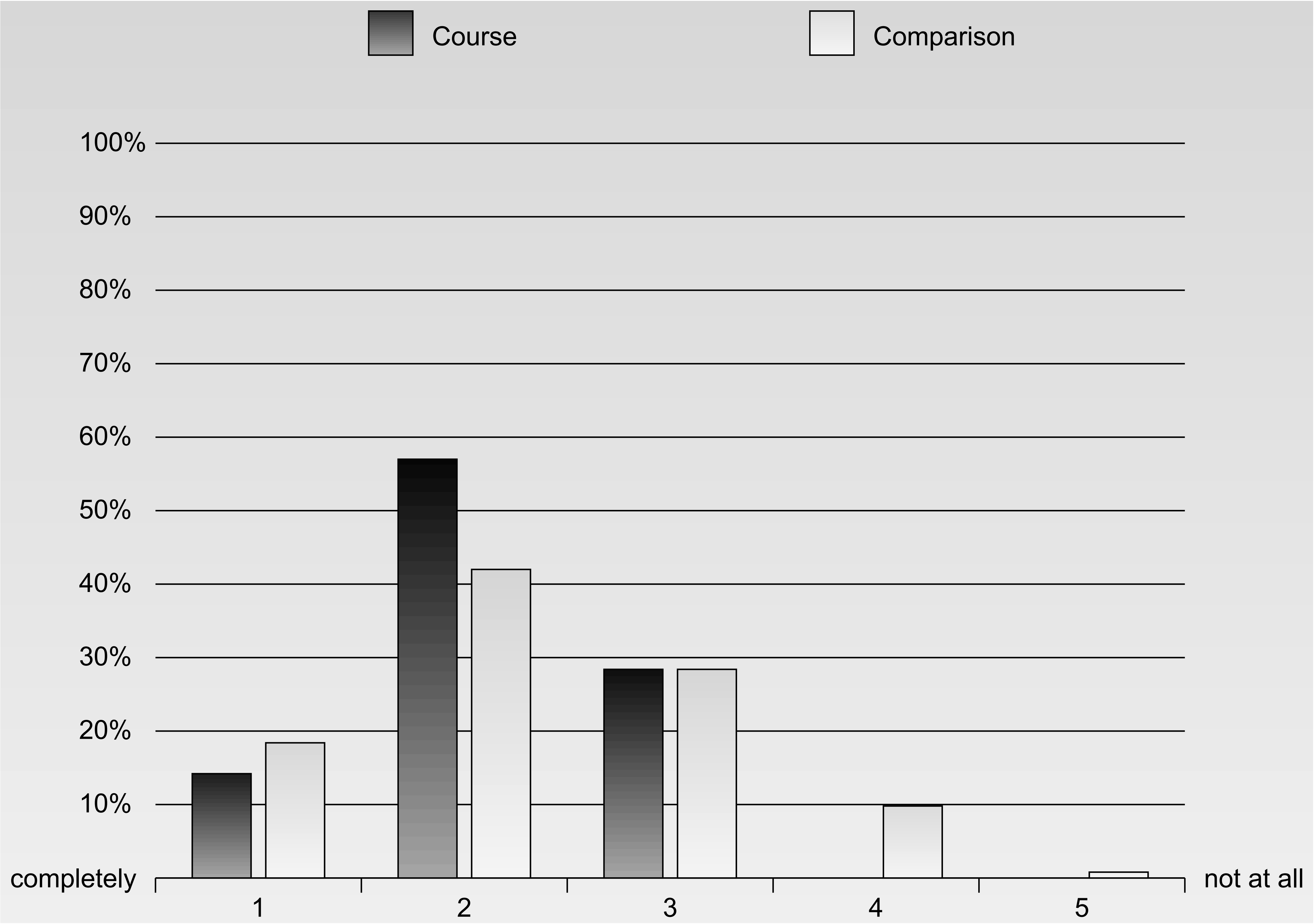
Course
Comparison
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
completely
not at all
1
2
3
4
5

Topic



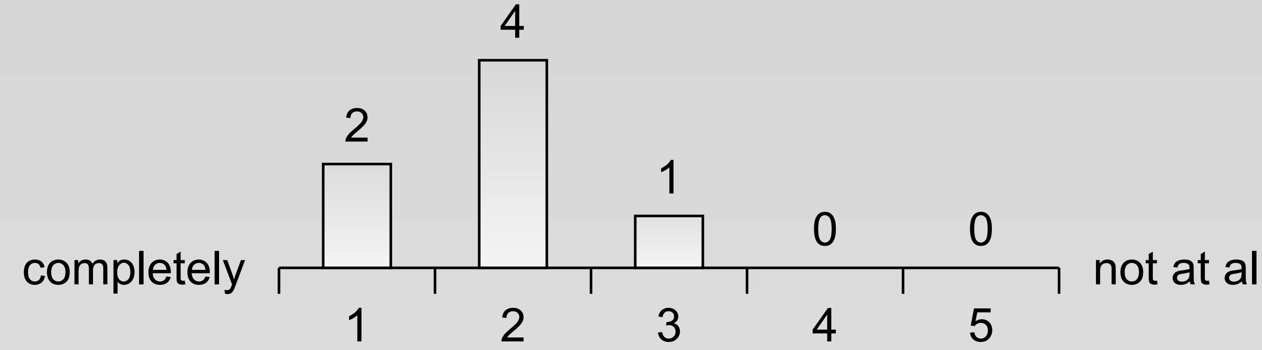
Course
Comparison
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
completely
not at all
1
2
3
4
5

Requirements

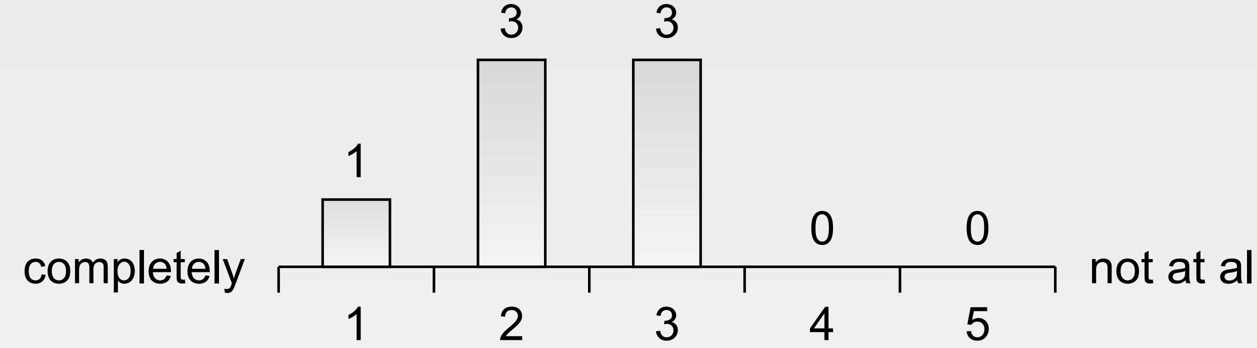


Requirements, in parts

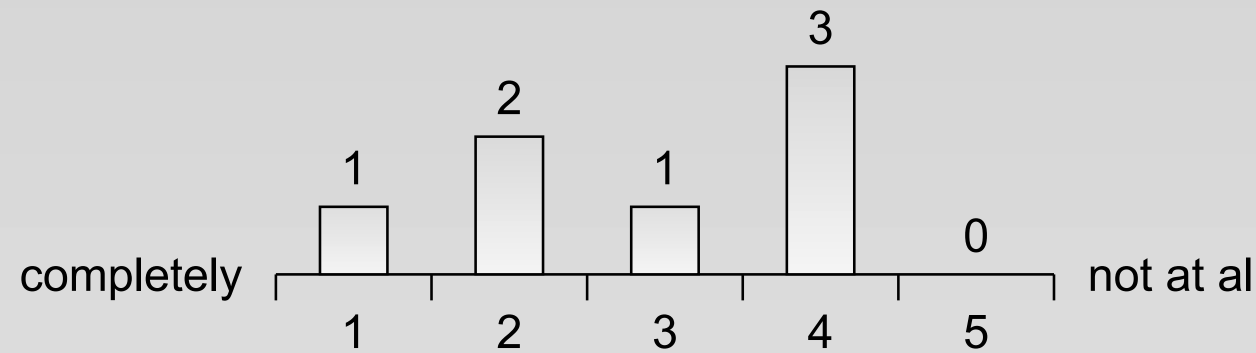
The difficulty of the content was adequate.



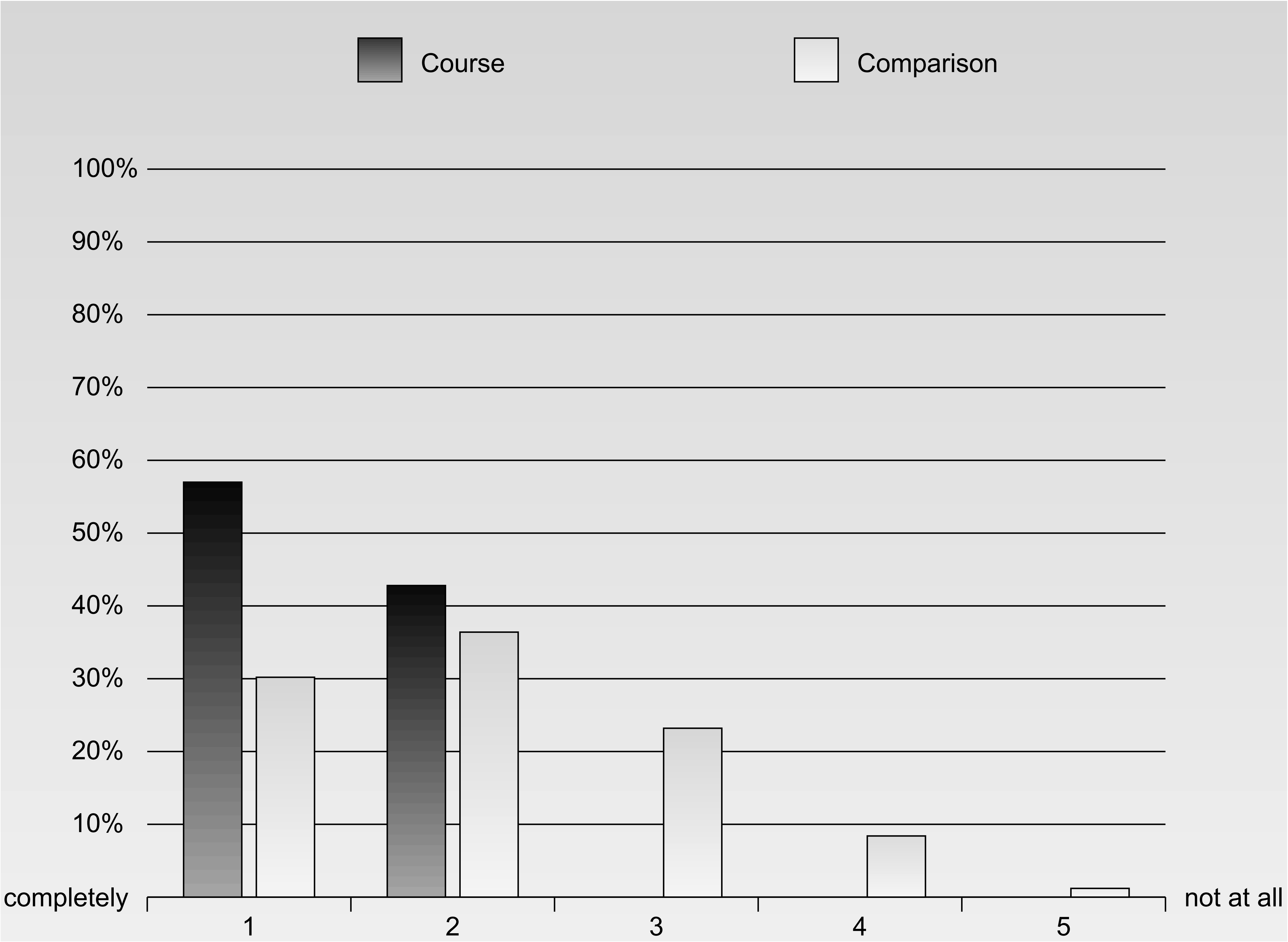
The amount of time required for the course as a whole (including preparation and follow-up) was appropriate.



The course was too difficult for me.



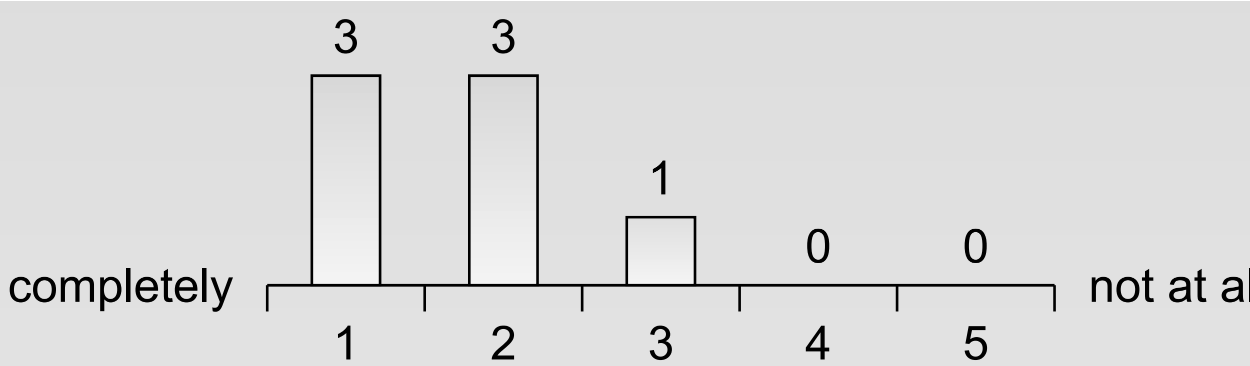
Overall

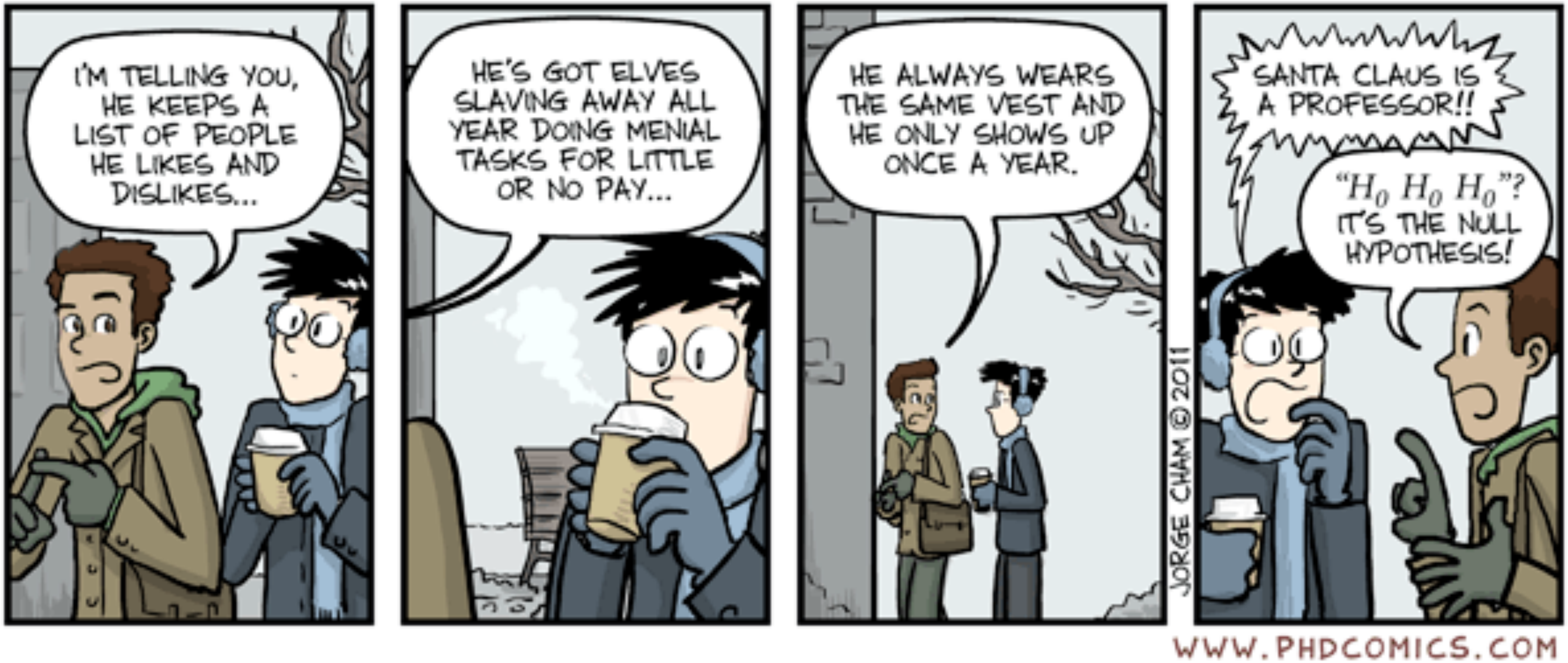


Course
Comparison
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
completely
not at all
1
2
3
4
5

A part of overall

I learned a lot in this course.




I'M TELLING YOU, HE KEEPS A LIST OF PEOPLE HE LIKES AND DISLIKES...
HE'S GOT ELVES SLAVING AWAY ALL YEAR DOING MENIAL TASKS FOR LITTLE OR NO PAY...
HE ALWAYS WEARS THE SAME VEST AND HE ONLY SHOWS UP ONCE A YEAR.
JORGE CHAM © 2011
SANTA CLAUS IS A PROFESSOR!!
"H_0 H_0 H_0"? IT'S THE NULL HYPOTHESIS!
WWW.PHDCOMICS.COM