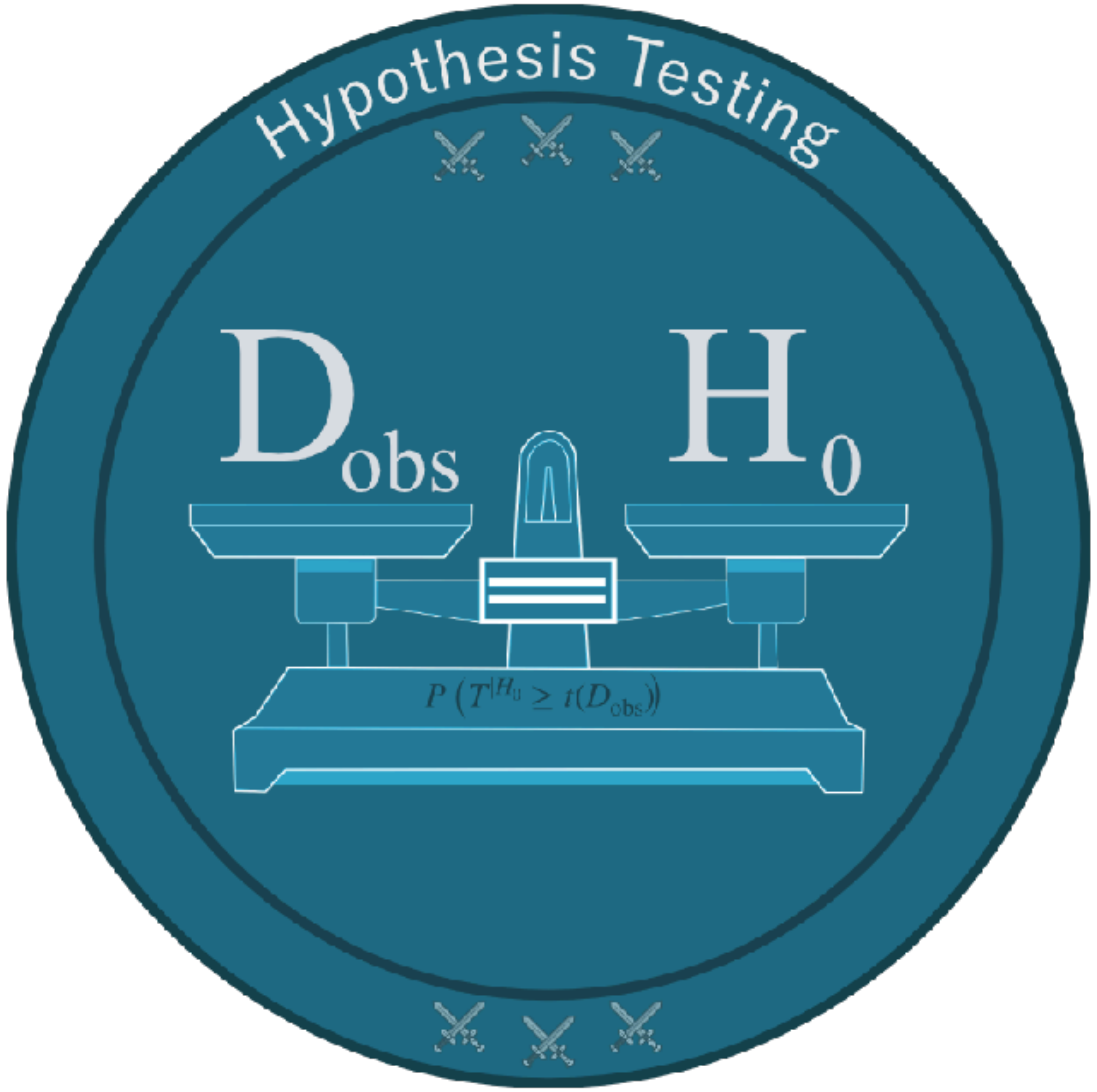
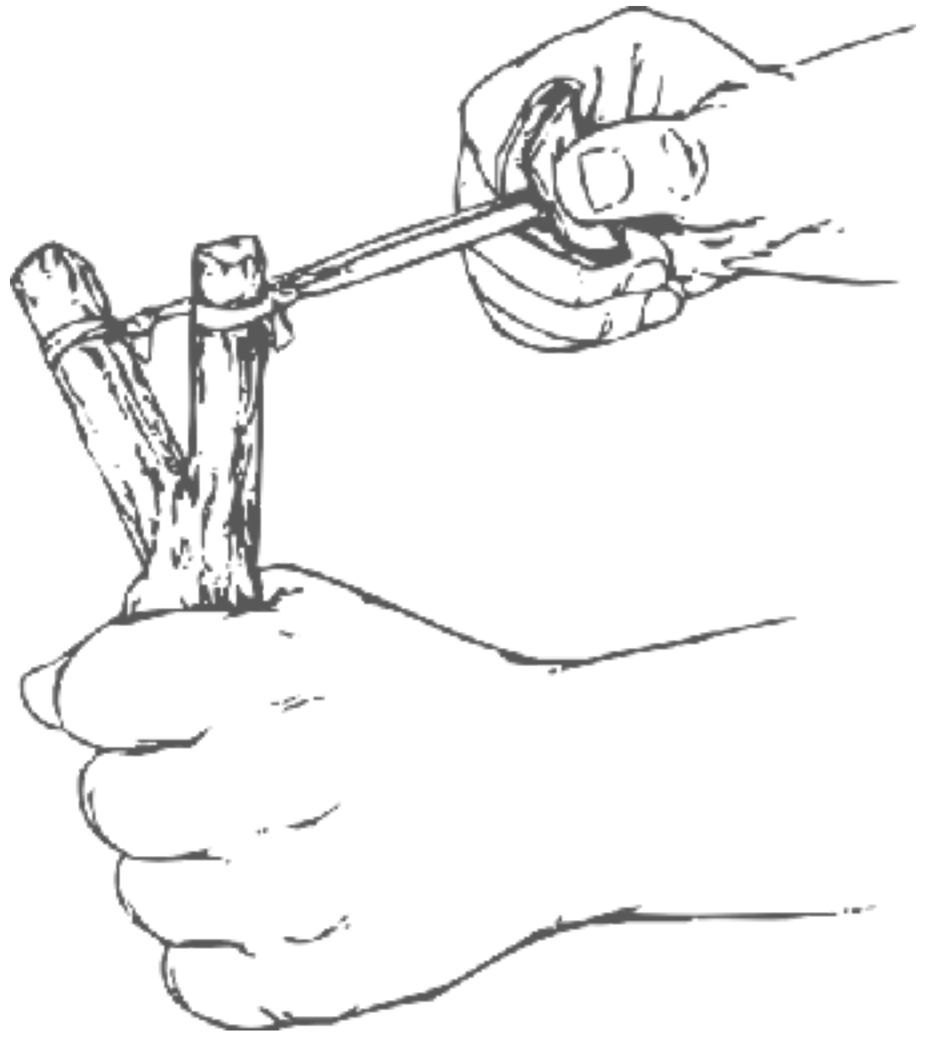


INTRODUCTION TO DATA ANALYSIS

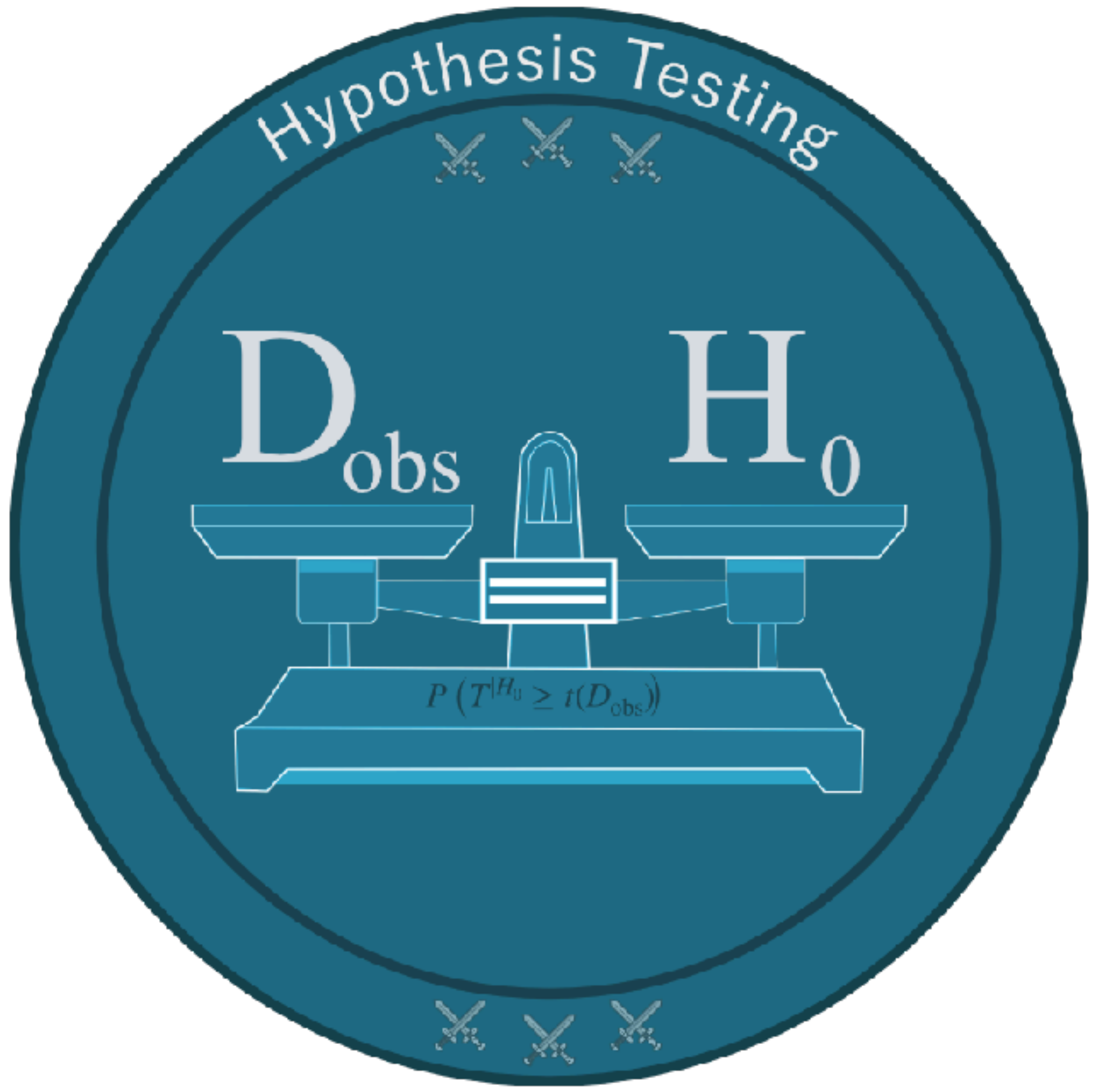
HYPOTHESIS TESTING

PART II

LEARNING GOALS



- get more intimate with *p*-values
  - distribution under true H_0
  - relation to confidence intervals
- develop a basic sense of how clever math (e.g., Central Limit Theorem) helps approximate sampling distributions
  - we don't aim for perfect understanding of this math in this course!
- become able to interpret & apply some statistical tests
  - Pearson's χ^2-tests
  - *z*-test
  - one-sample *t*-test



p-value
revisit

RECAP

BAYESIAN PARAMETER ESTIMATION

- model M captures prior beliefs about data-generating process
  - prior over latent parameters
  - likelihood of data
- Bayesian posterior inference using observed data D_{obs}
- compare posterior beliefs to some parameter value of interest

FREQUENTIST HYPOTHESIS TESTING

- model M captures a hypothetically assumed data-generating process
  - fix parameter value of interest
  - likelihood of data
- single out some aspect of the data as most important (test statistic)
- look at distribution of test statistic given the assumed model (sampling distribution)
- check likelihood of test statistic applied to the observed data D_{obs}

P-VALUE

$$p(D_{\text{obs}}) = P(T^{|H_0} \geq^{H_{0,a}} t(D_{\text{obs}}))$$

RELATION OF P-VALUES AND CONFIDENCE INTERVALS

- assumptions:
  - p-value and CI are constructed / approximated in the same way
  - two-sided test with $H_0 : \theta = \theta_0$ and alternative $H_a : \theta \neq \theta_0$
- correspondence result:

$$p(D) < \alpha \quad \text{iff} \quad \theta_0 \notin \text{CI}(D)$$

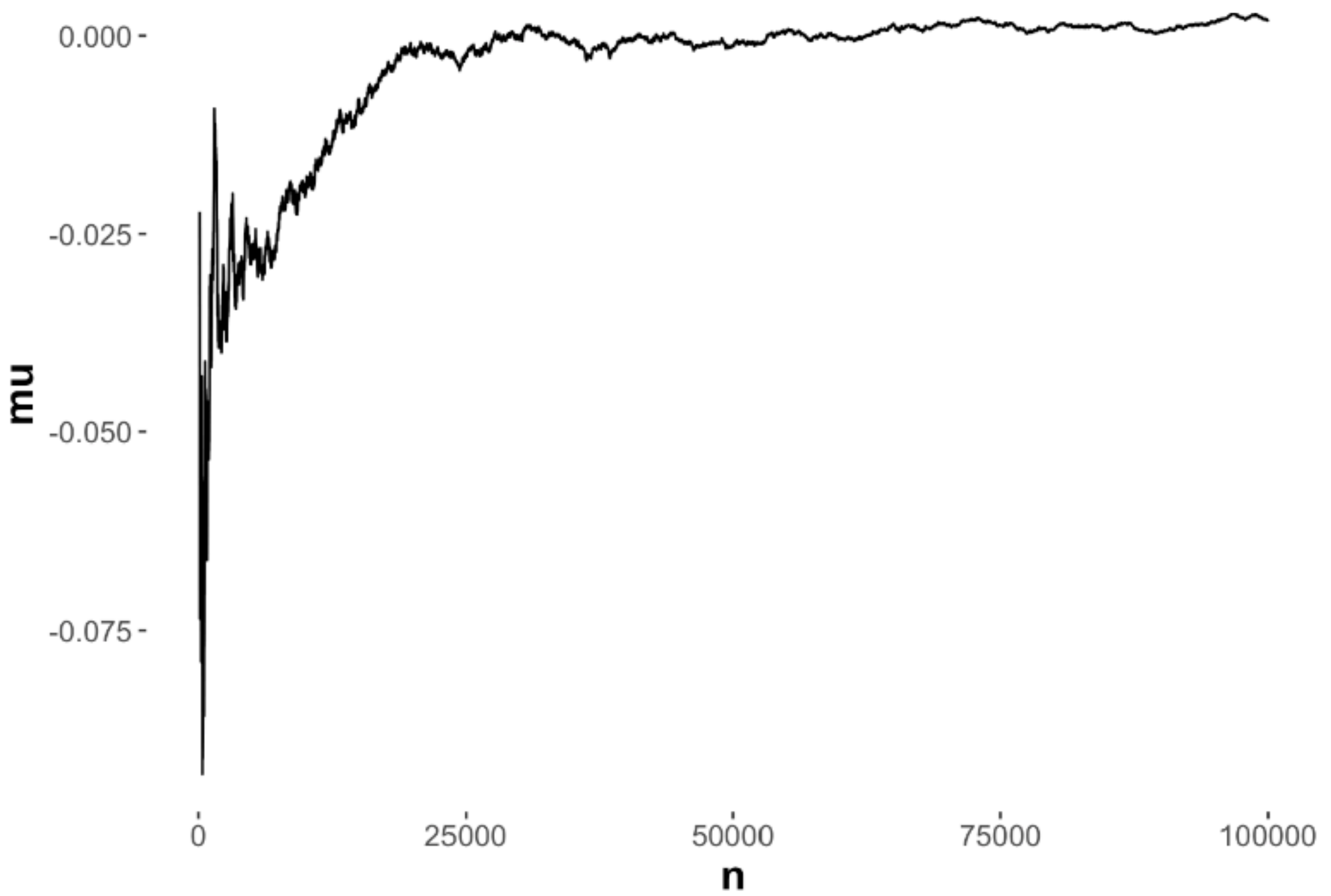
approximating sampling distributions

LAW OF LARGE NUMBERS

Theorem 10.2 (Law of Large Numbers) *Let $X_1, \ldots, X_n$ be a sequence of n differentiable random variables with equal mean, such that $\mathbb{E}_{X_i} = \mu_X$ for all $1 \leq i \leq n$.*[60] *As the number of samples n goes to infinity the mean of any tuple of samples, one from each X_i, convergences almost surely to μ_X:*

$$P\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu_X\right) = 1$$

```r
# sample from a standard normal distribution (mean = 0, sd = 1)
samples <- rnorm(100000)
# collect the mean after each 10 samples & plot
tibble(
  n = seq(100, length(samples), by = 10)
  ) %>%
  group_by(n) %>%
  mutate(
  mu = mean(samples[1:n])
)  %>%
  ggplot(aes(x = n, y = mu)) +
  geom_line()
```



CENTRAL LIMIT THEOREM

Theorem 10.3 (Central Limit Theorem) *Let $X_1, \ldots, X_n$ be a sequence of n differentiable random variables with equal mean $\mathbb{E}_{X_i} = \mu_X$ and equal finite variance $\mathrm{Var}(X_i) = \sigma_X^2$ for all $1 \leq i \leq n$.*[61] *The random variable S_n which captures the distribution of the sample mean for any n is:*

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

As the number of samples n goes to infinity the random variable $\sqrt{n}(S_n - \mu_X)$ converges in distribution to a normal distribution with mean 0 and standard deviation σ_X.

CLT gives us information about the distribution of estimated means, e.g., as when we estimate repeatedly in different (hypothetical experiments).

Pearson's χ^2-tests

PEARSON χ^2-TESTS

- tests for categorical data (with more than two categories)
- two flavors:
  - test of goodness of fit
  - test of independence
- sampling distribution is a χ^2-distribution

χ^2-DISTRIBUTION

- standard normal random variables:

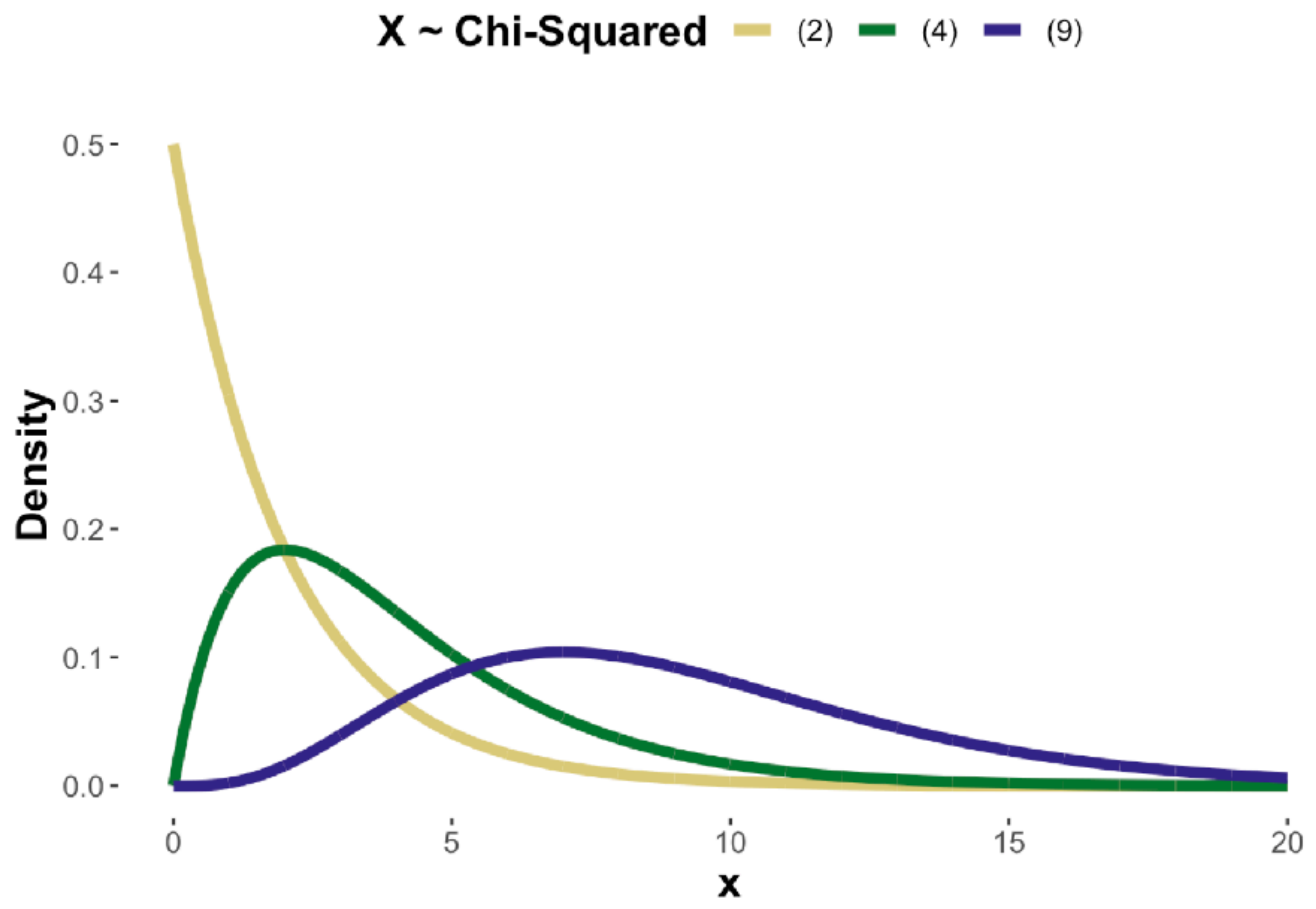
  $X_1, \ldots X_n$

- derived RV:

  $Y = X_1^2 + \ldots + X_n^2$

- it follows (by construction) that:

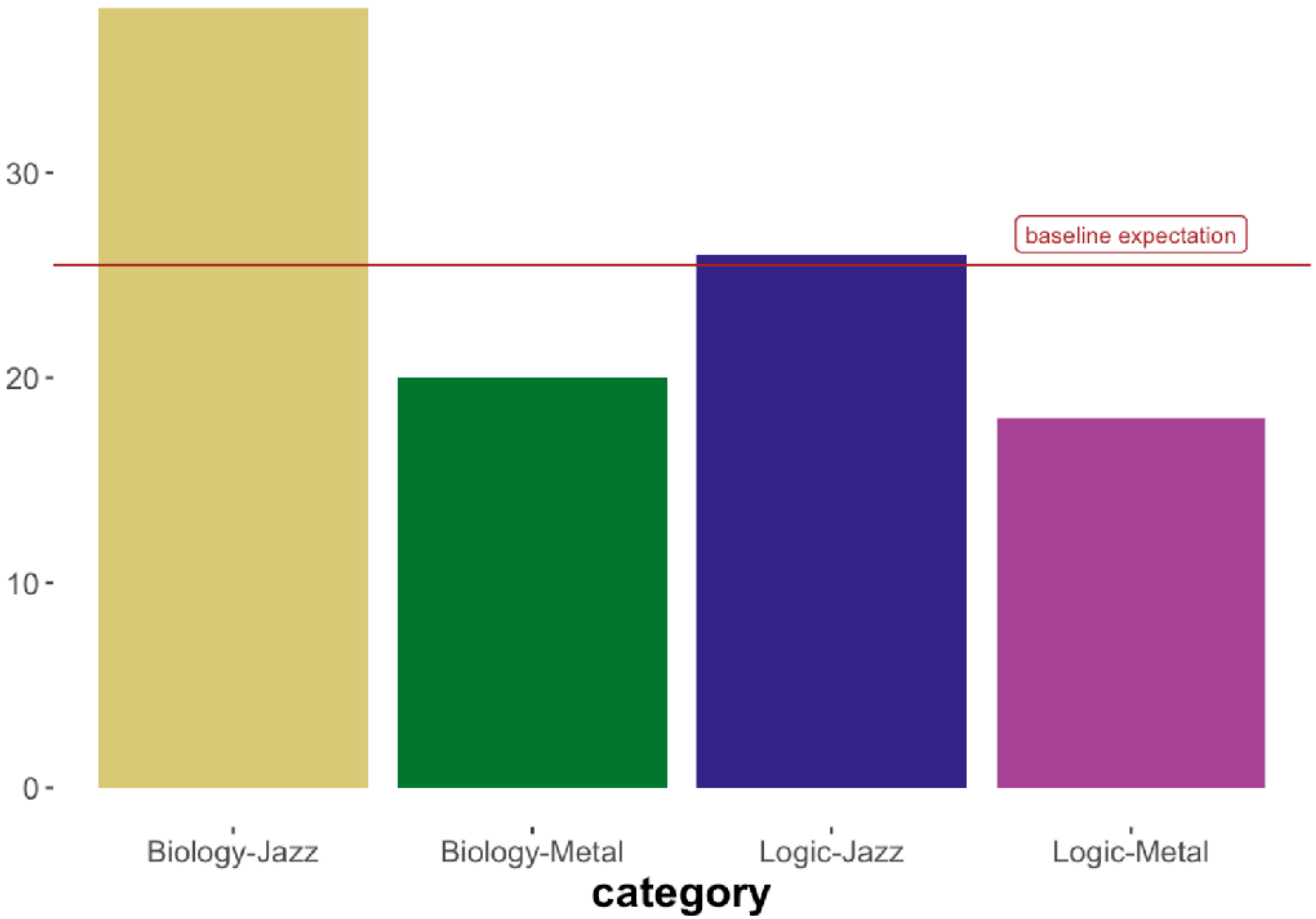
  $y \sim \chi^2\text{-distribution}(n)$



PEARSON'S χ^2-TEST [GOODNESS OF FIT]

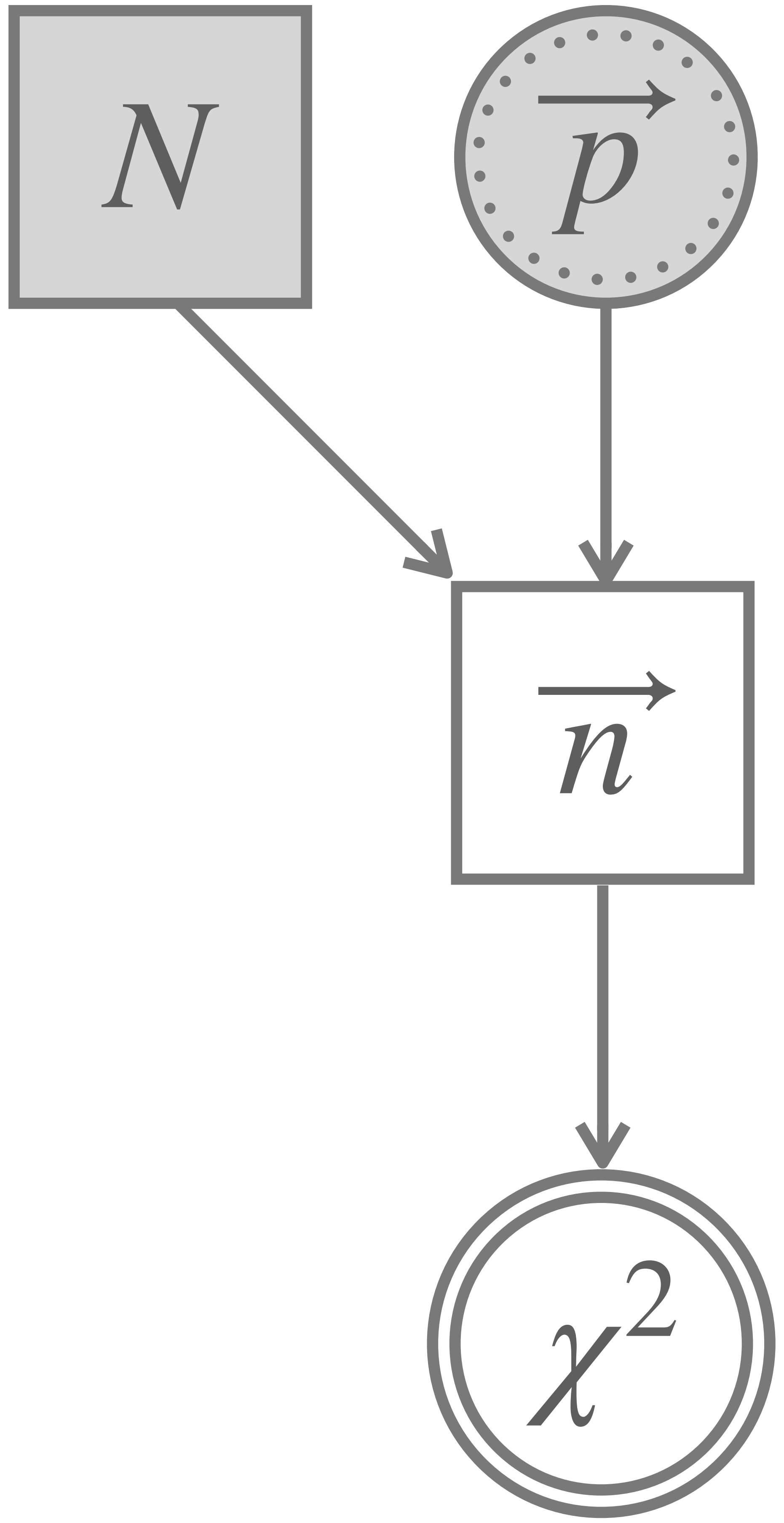
```
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts
```

```
## # A tibble: 4 x 3
##   JM    LB          n
##   <chr> <chr>   <int>
## 1 Jazz  Biology    38
## 2 Jazz  Logic      26
## 3 Metal Biology    20
## 4 Metal Logic      18
```



Is it conceivable that each category (= pair of music+subject choice) has been selected with the same flat probability of 0.25?

FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [GOODNESS OF FIT]



$$\vec{n} \sim \text{Multinomial}(\vec{p}, N)$$

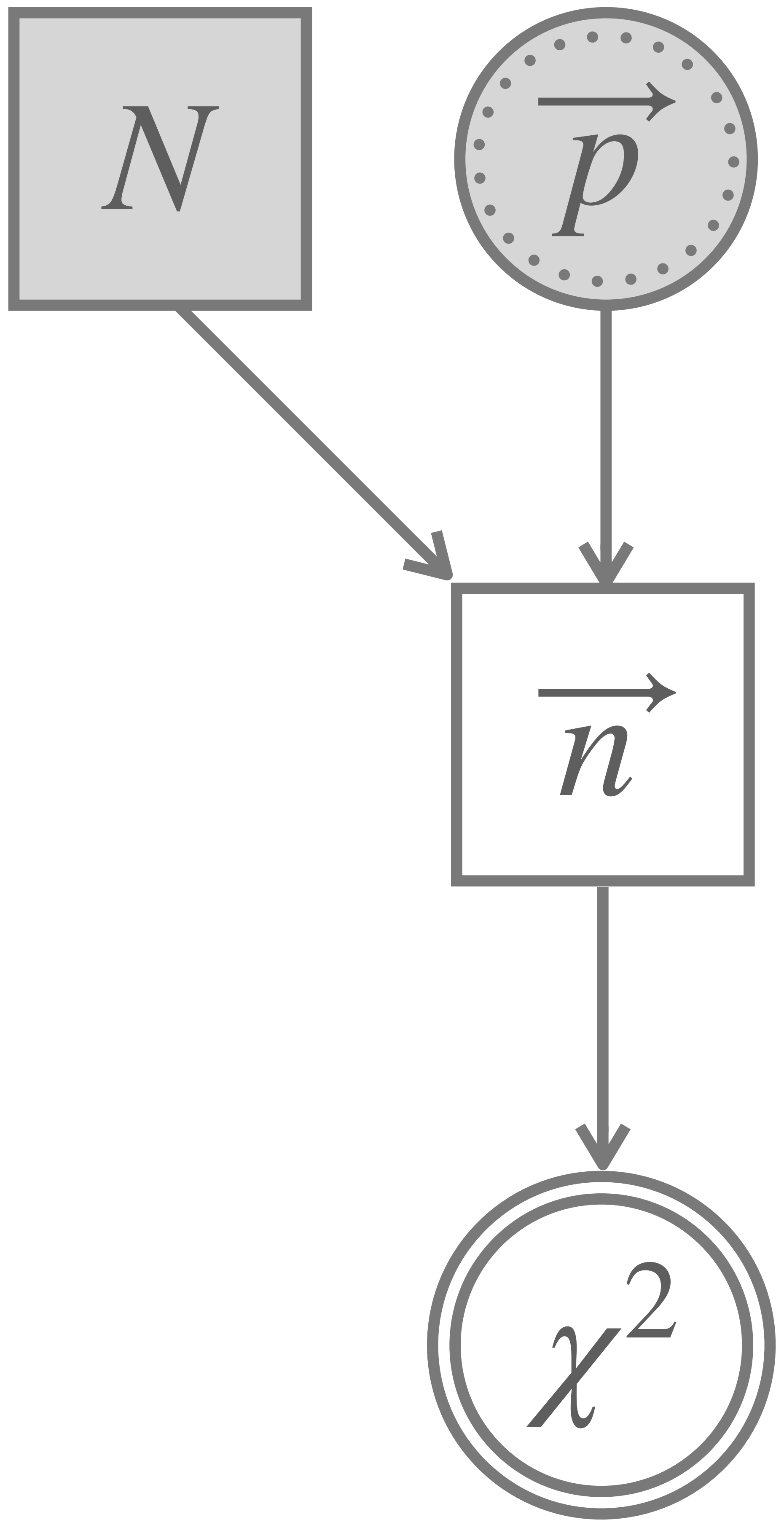
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

FACT:

The sampling distribution of χ^2 is approximately:

$\chi^2 \sim \chi^2$-distribution$(k-1)$

PEARSON'S χ^2-TEST [GOODNESS OF FIT]



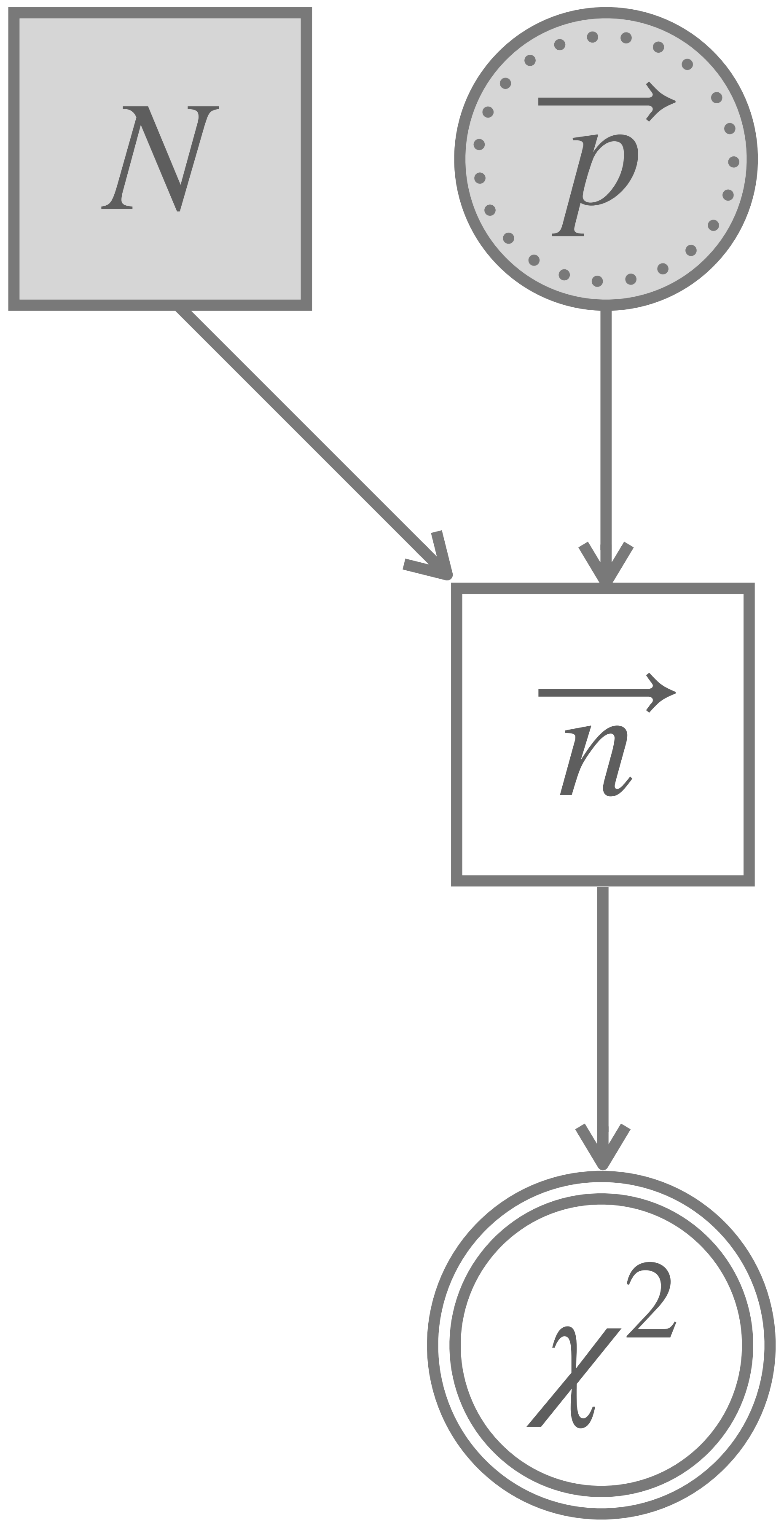
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

$$\chi^2 \sim \chi^2\text{-distribution}(k-1)$$

```r
# observed counts
n <- counts_BLJM_choice_pairs_vector
# proprortion predicted
p <- rep(1/4,4)
# expected number in each cell
e <- sum(n)*p
# chi-squared for observed data
chi2_observed <- sum((n-e)^2 *1/e)
chi2_observed
```

```
## [1] 9.529412
```

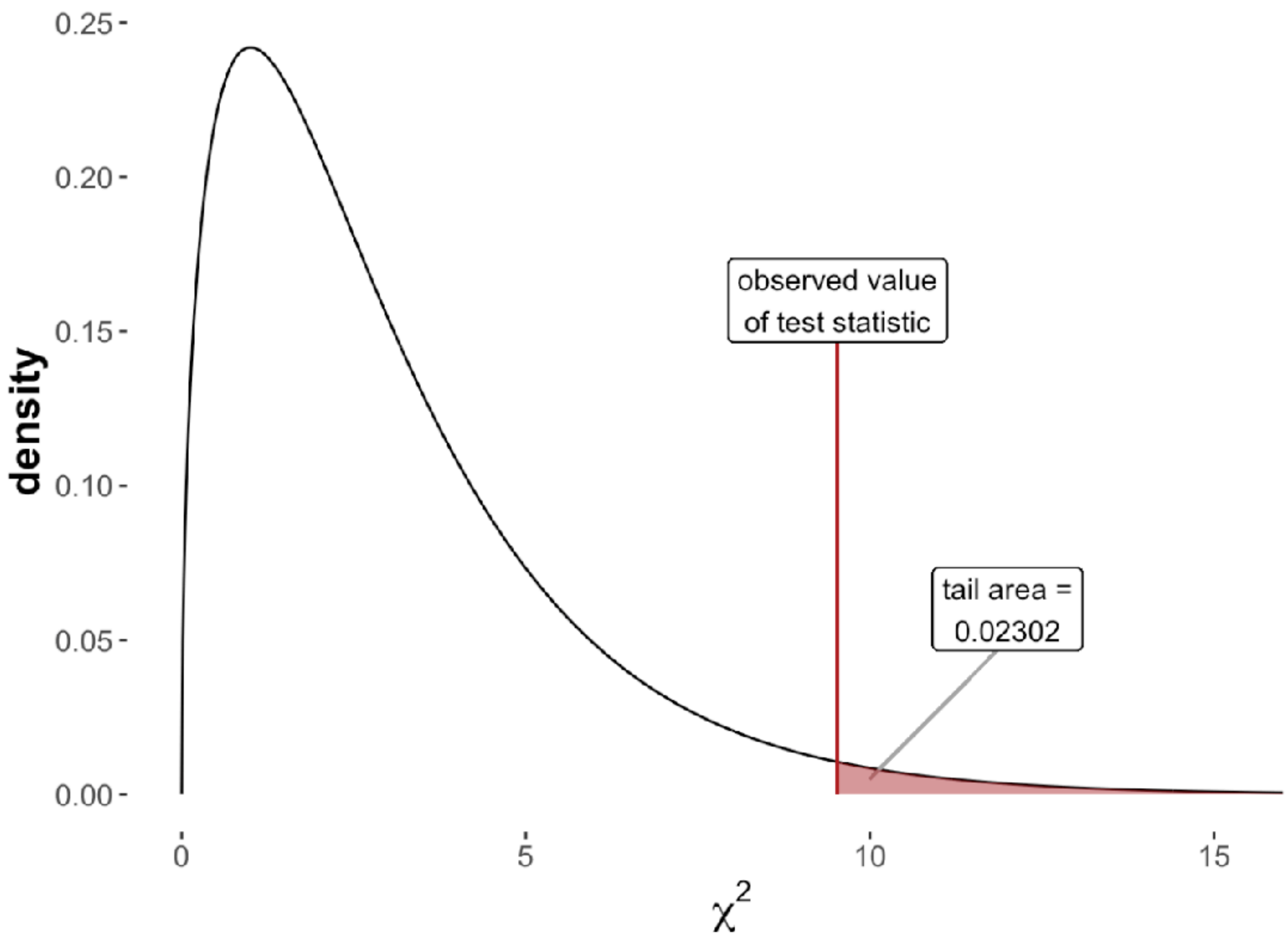
PEARSON'S χ^2-TEST [GOODNESS OF FIT]



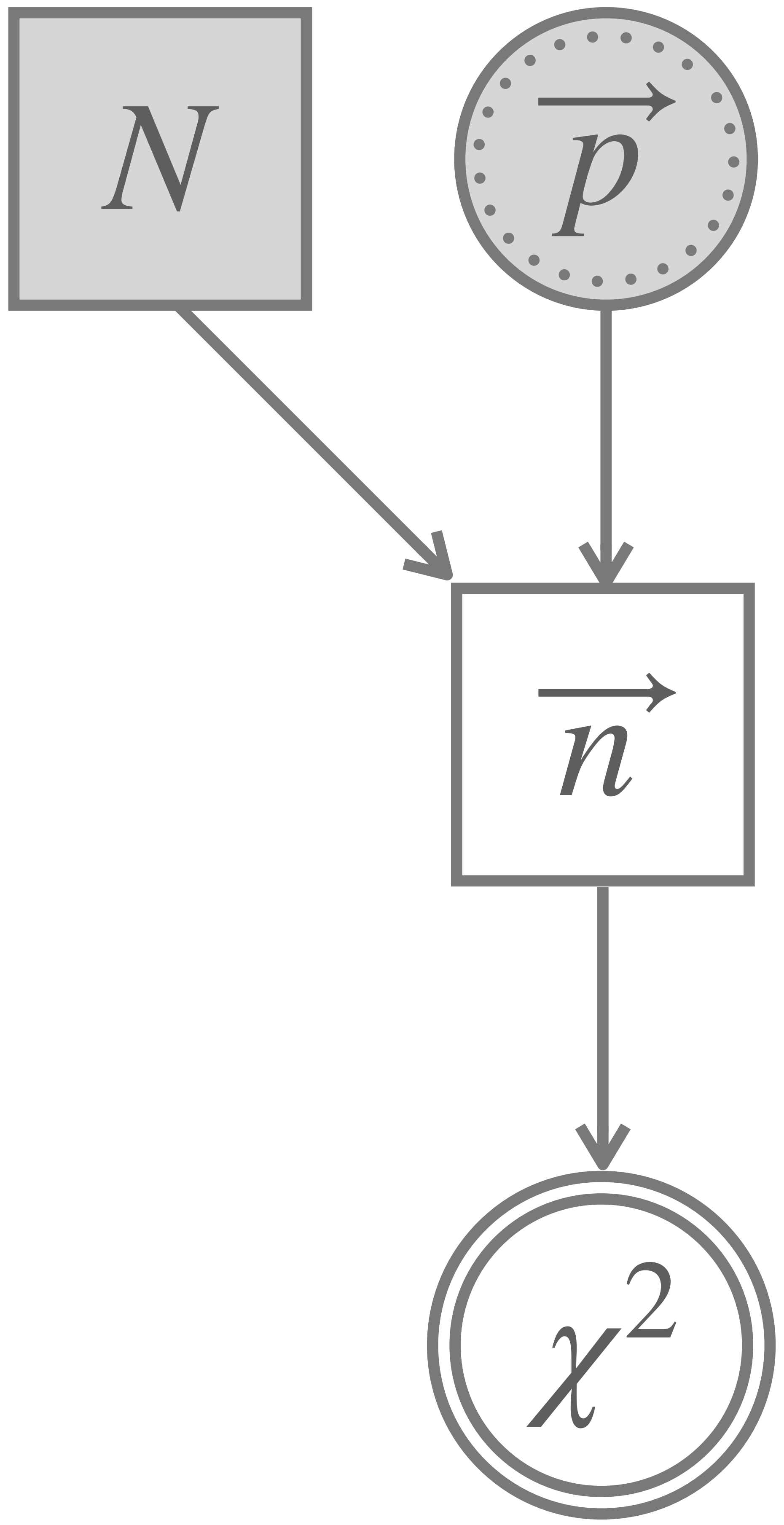
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

$$\chi^2 \sim \chi^2\text{-distribution}(k-1)$$

```r
p_value_BLJM <- 1 - pchisq(chi2_observed, df = 3)
```



PEARSON'S χ^2-TEST [GOODNESS OF FIT]



$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

$$\chi^2 \sim \chi^2\text{-distribution}(k-1)$$

```
counts_BLJM_choice_pairs_vector <- BLJM_associated_counts %>% pull(n)
chisq.test(counts_BLJM_choice_pairs_vector)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  counts_BLJM_choice_pairs_vector
## X-squared = 9.5294, df = 3, p-value = 0.02302
```

PEARSON'S χ^2-TEST [GOODNESS OF FIT]

How to interpret / report the result:

> Observed counts deviated significantly from what is expected if each category (here: pair of music+subject choice) was equally likely (χ^2-test, with $\chi^2 \approx 9.53$, $df = 3$ and $p \approx 0.023$).

What about the lecturer's conjecture that (colorfully speaking) logic + metal = 🥰?