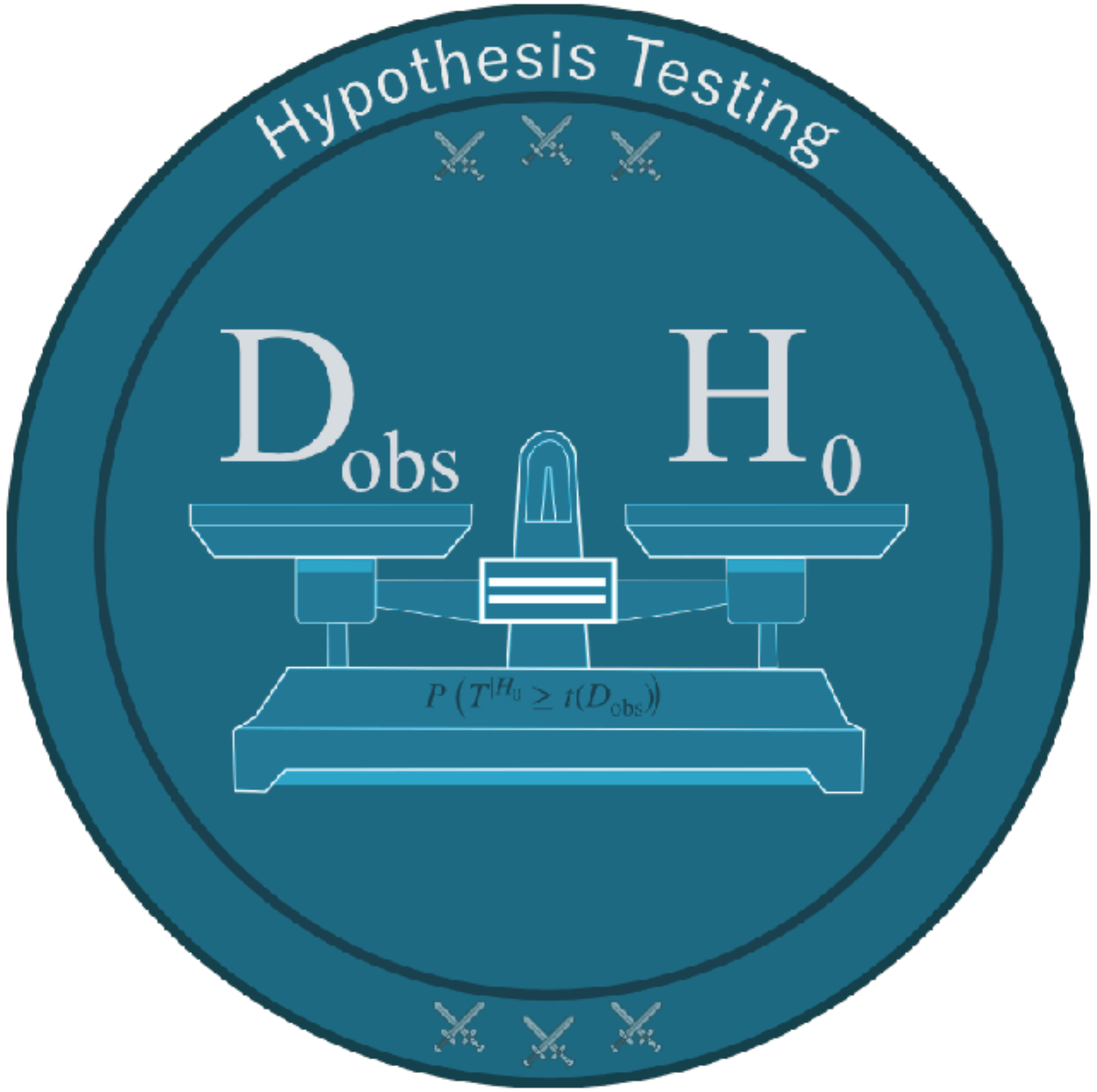
INTRODUCTION TO DATA ANALYSIS

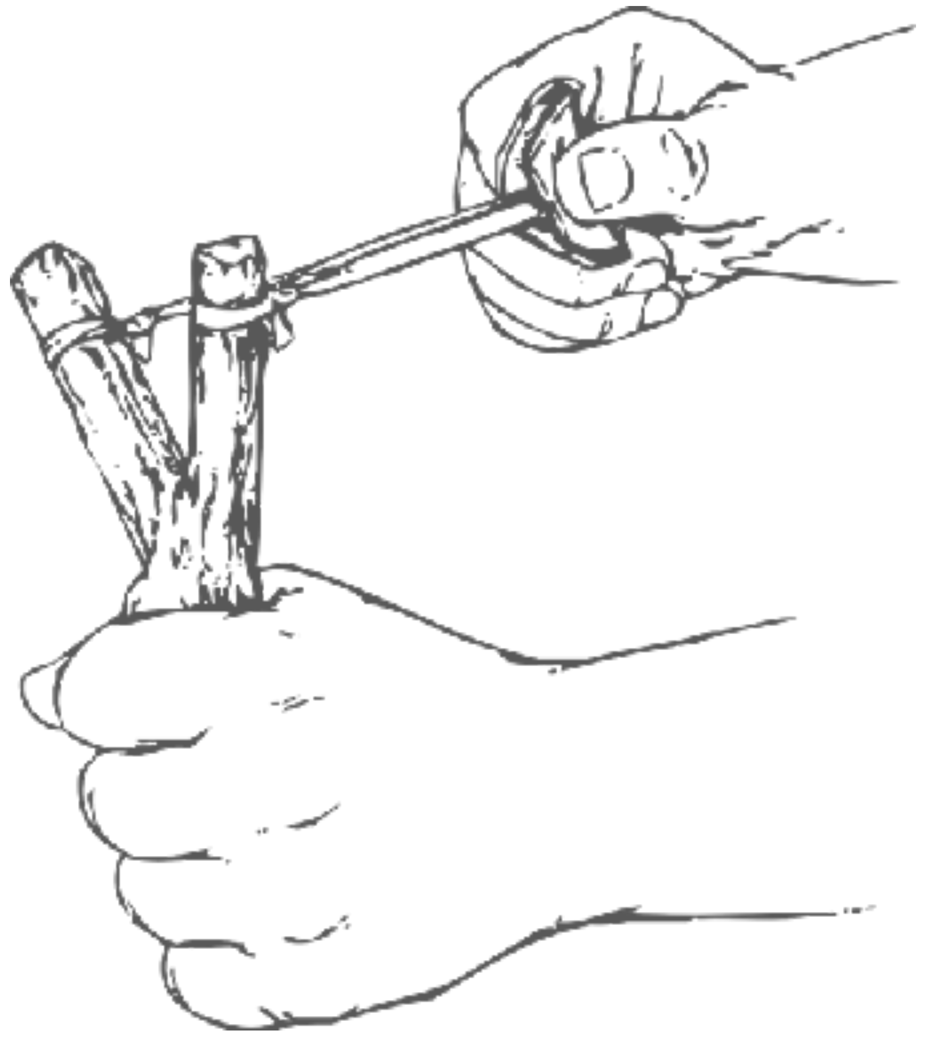


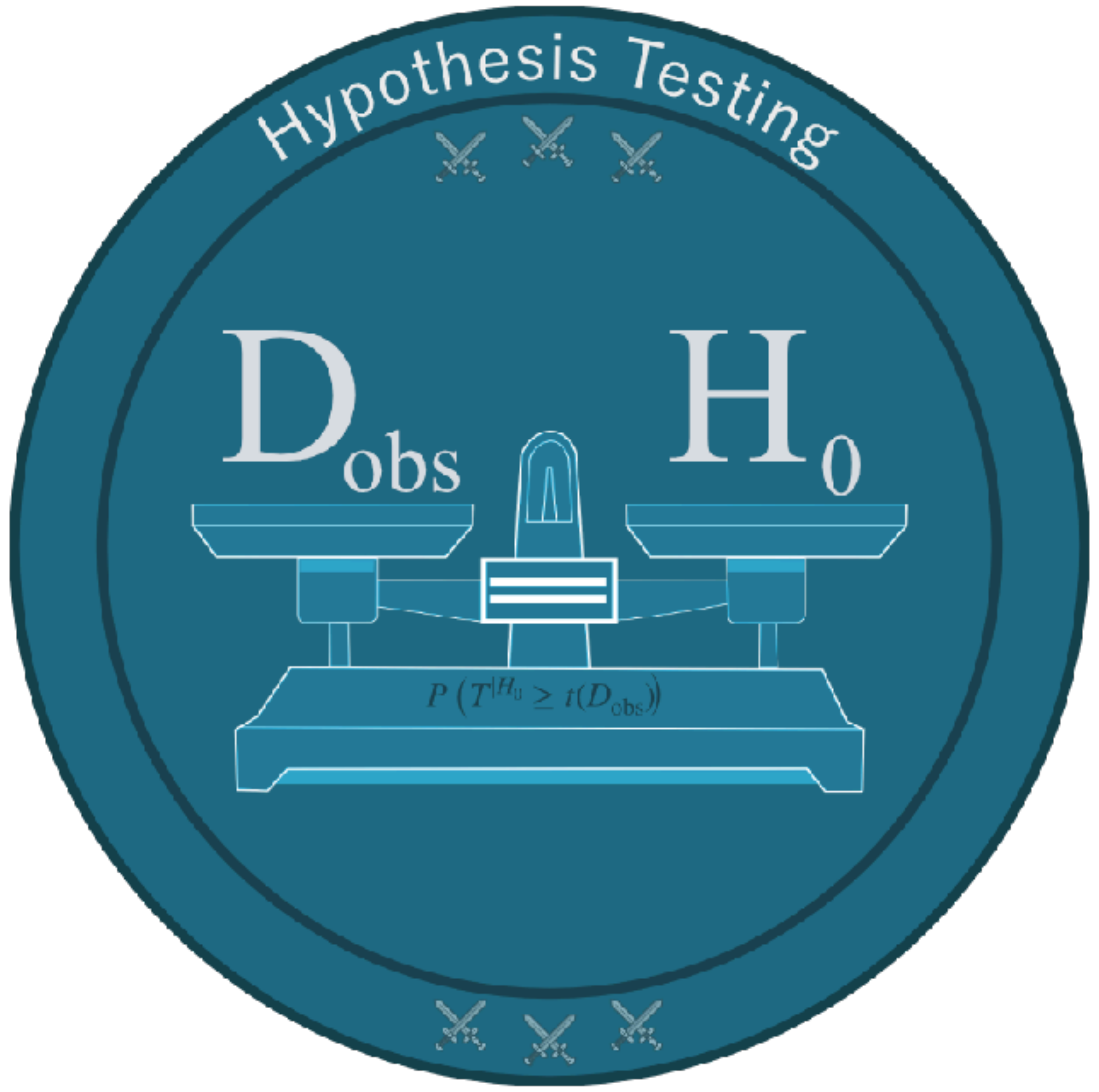
HYPOTHESIS TESTING

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:

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

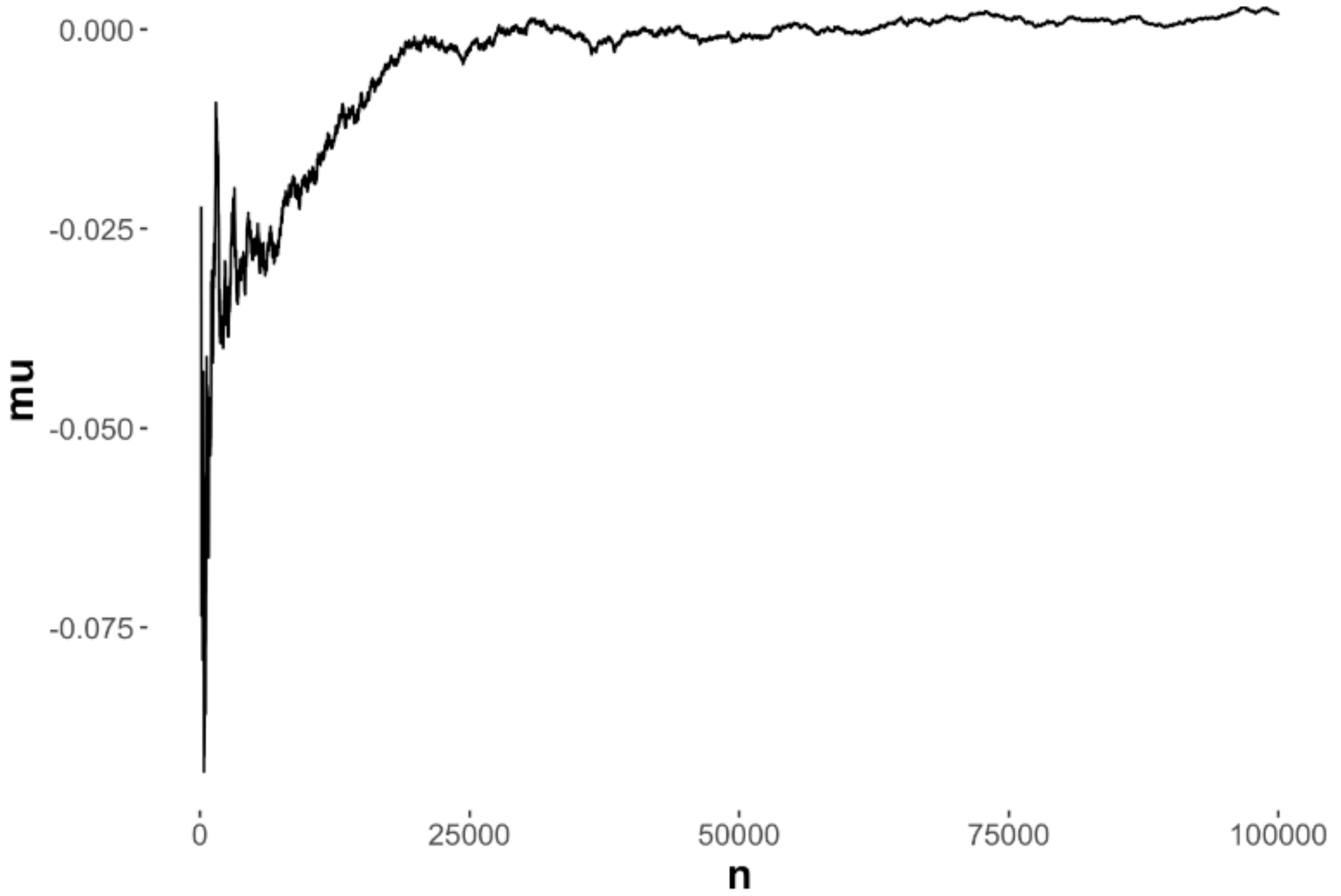
approximating sampling distributions

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:*

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

```
# 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:

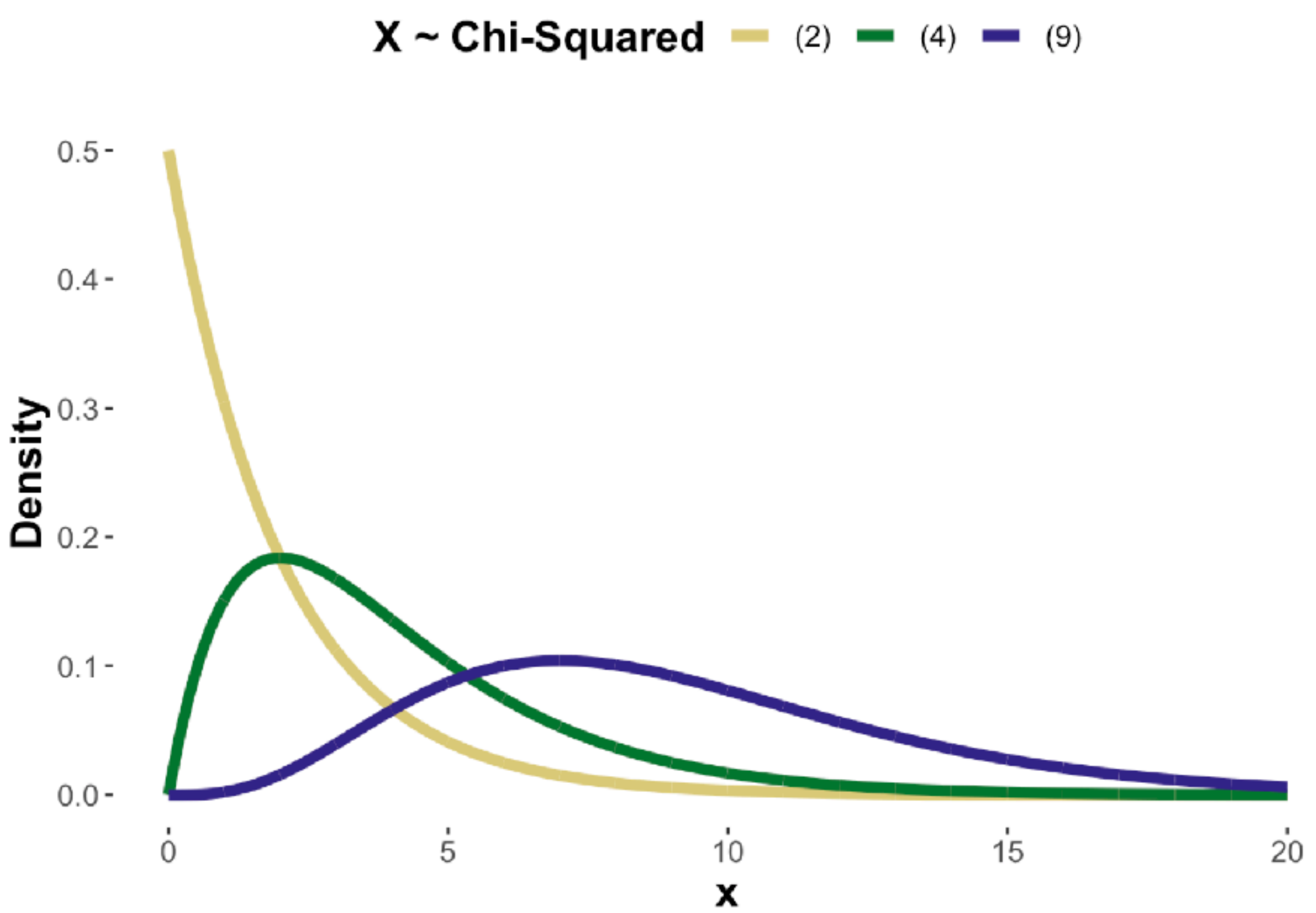
  $$X_1, \ldots X_n$$

- derived RV:

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

- it follows (by construction) that:

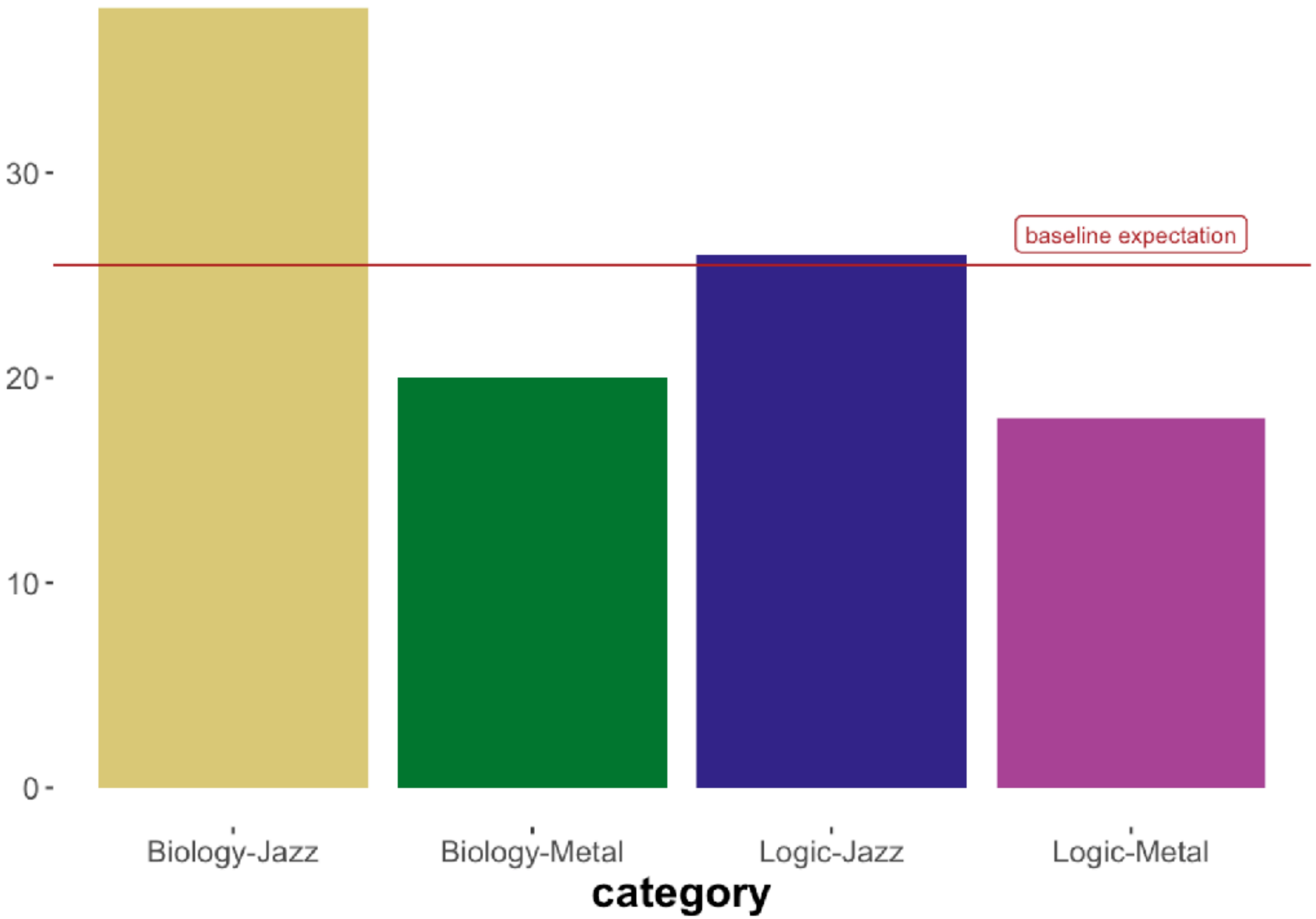
  $$y \sim \chi^2\text{-distribution}(n)$$



PEARSON'S χ^2-TEST [GOODNESS OF FIT]

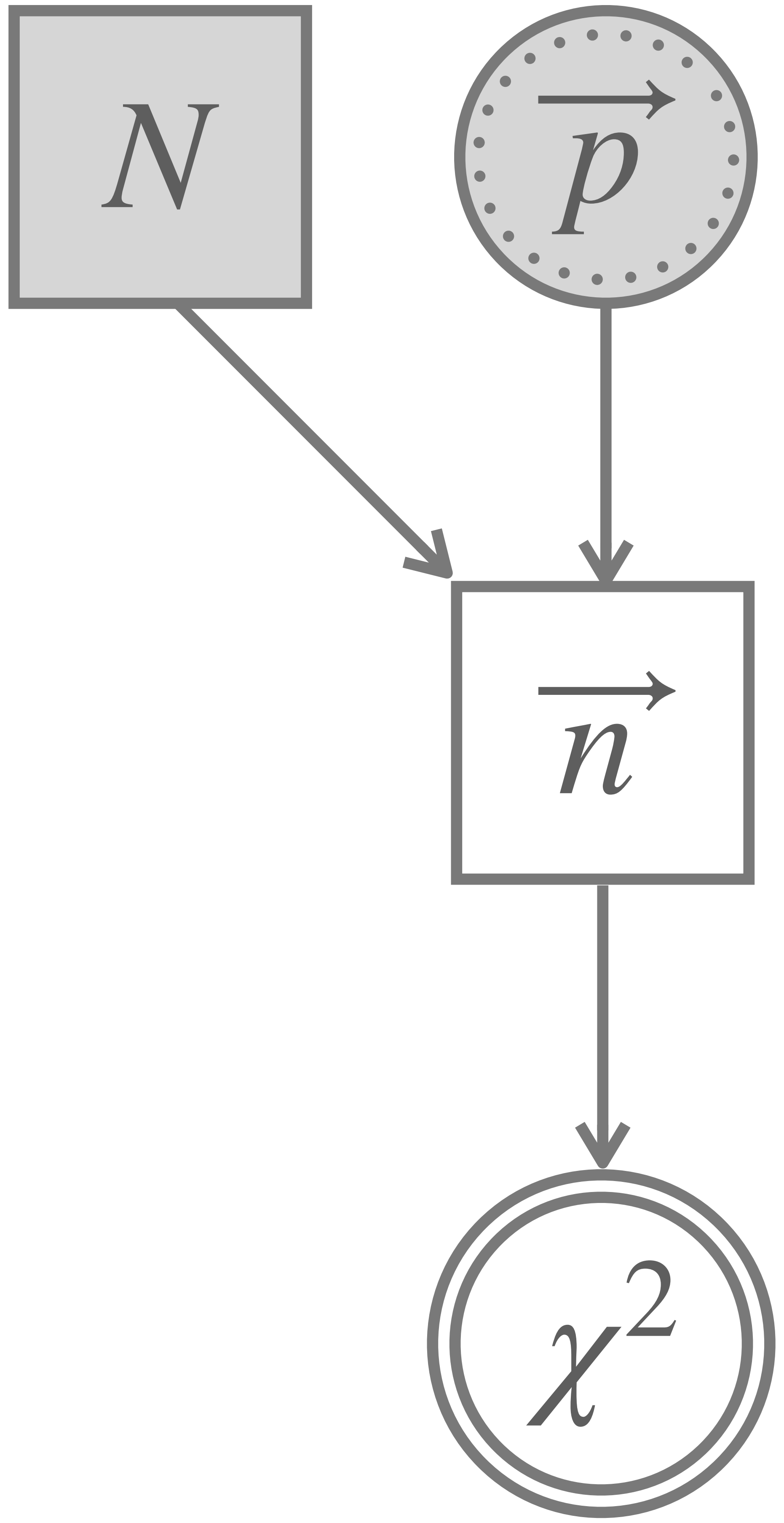
```r
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts
```

```
## # 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]



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

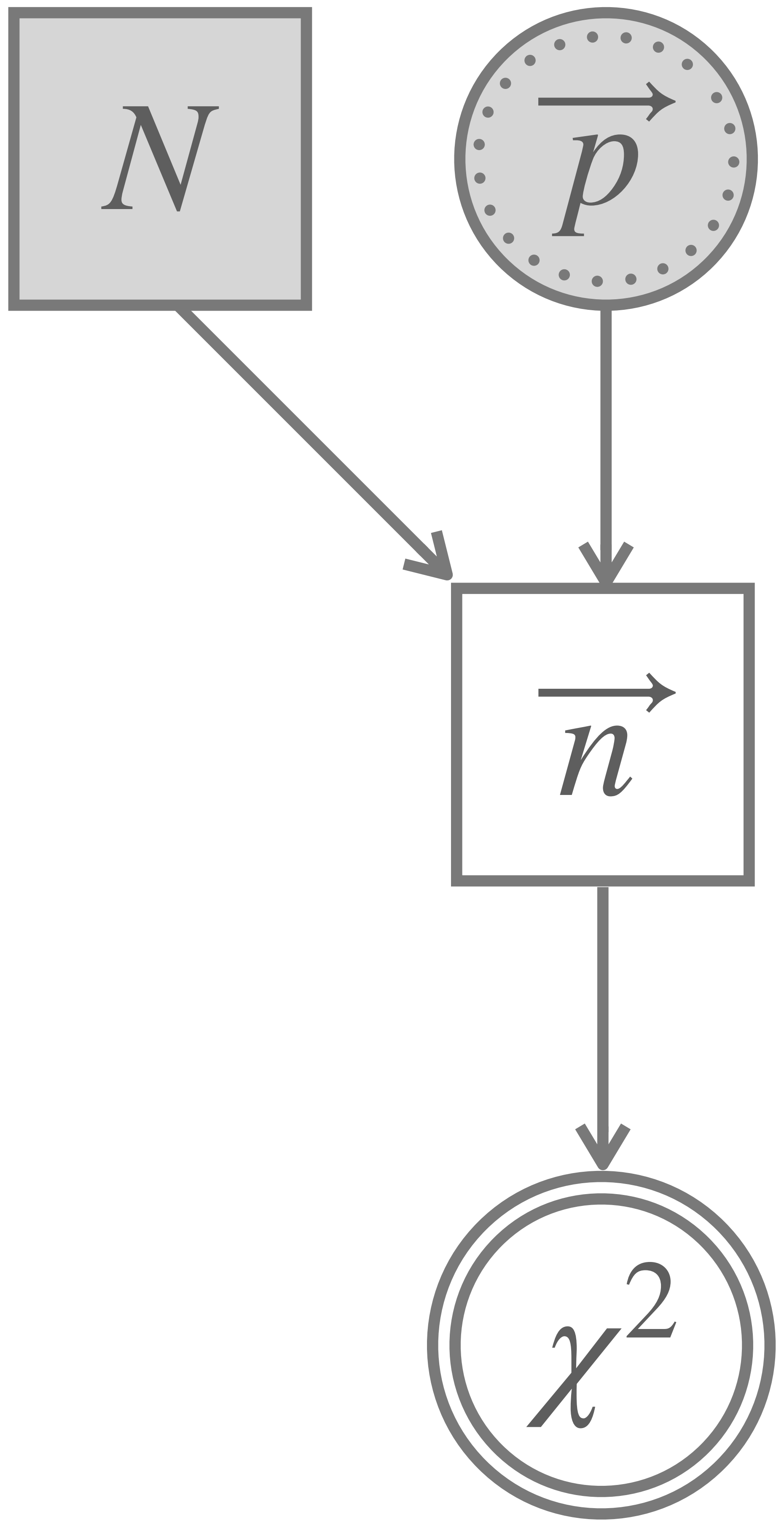
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

FACT:

The sampling distribution of χ^2 is approximately:

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

PEARSON'S χ^2-TEST [GOODNESS OF FIT]



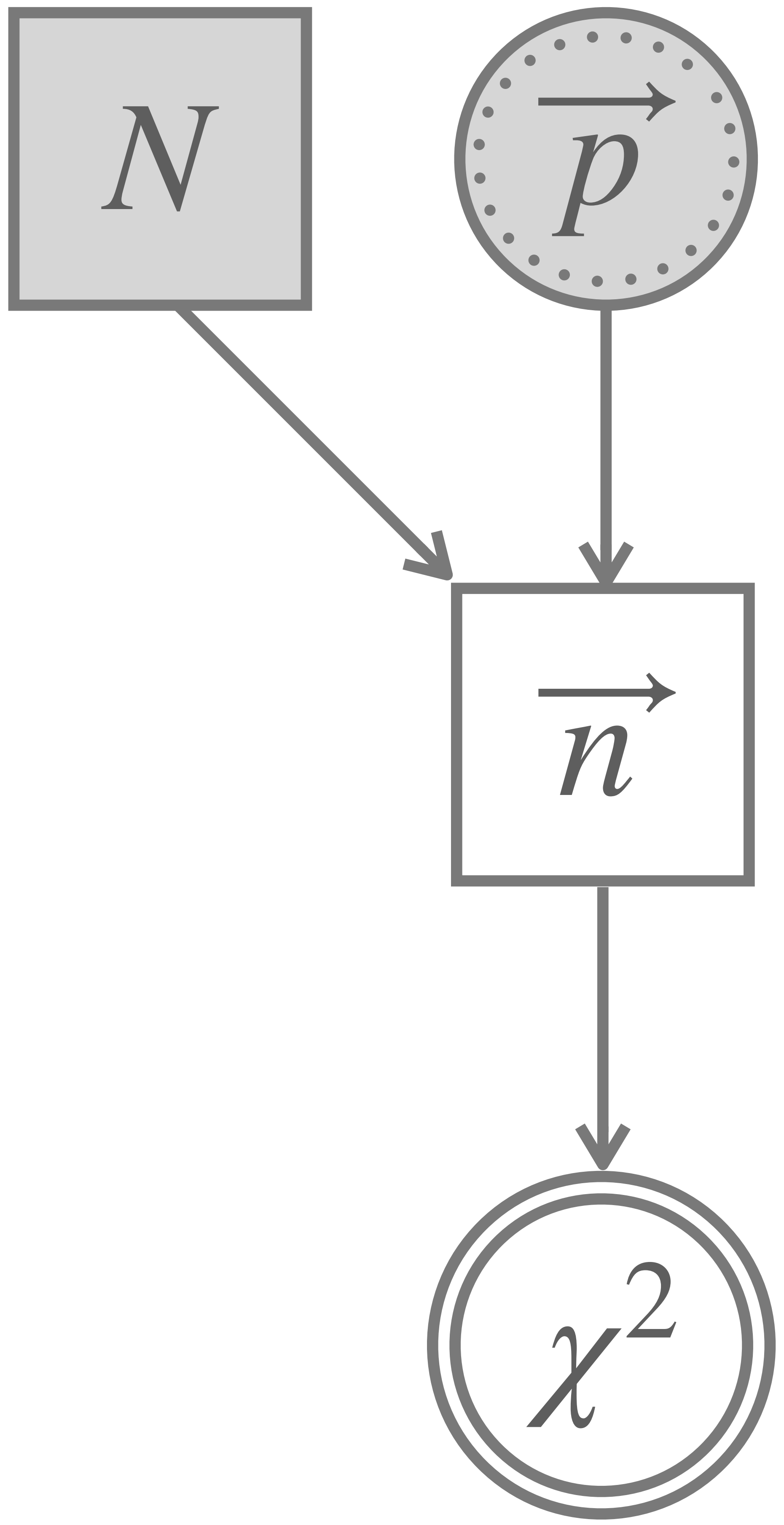
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

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

```
# 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
```

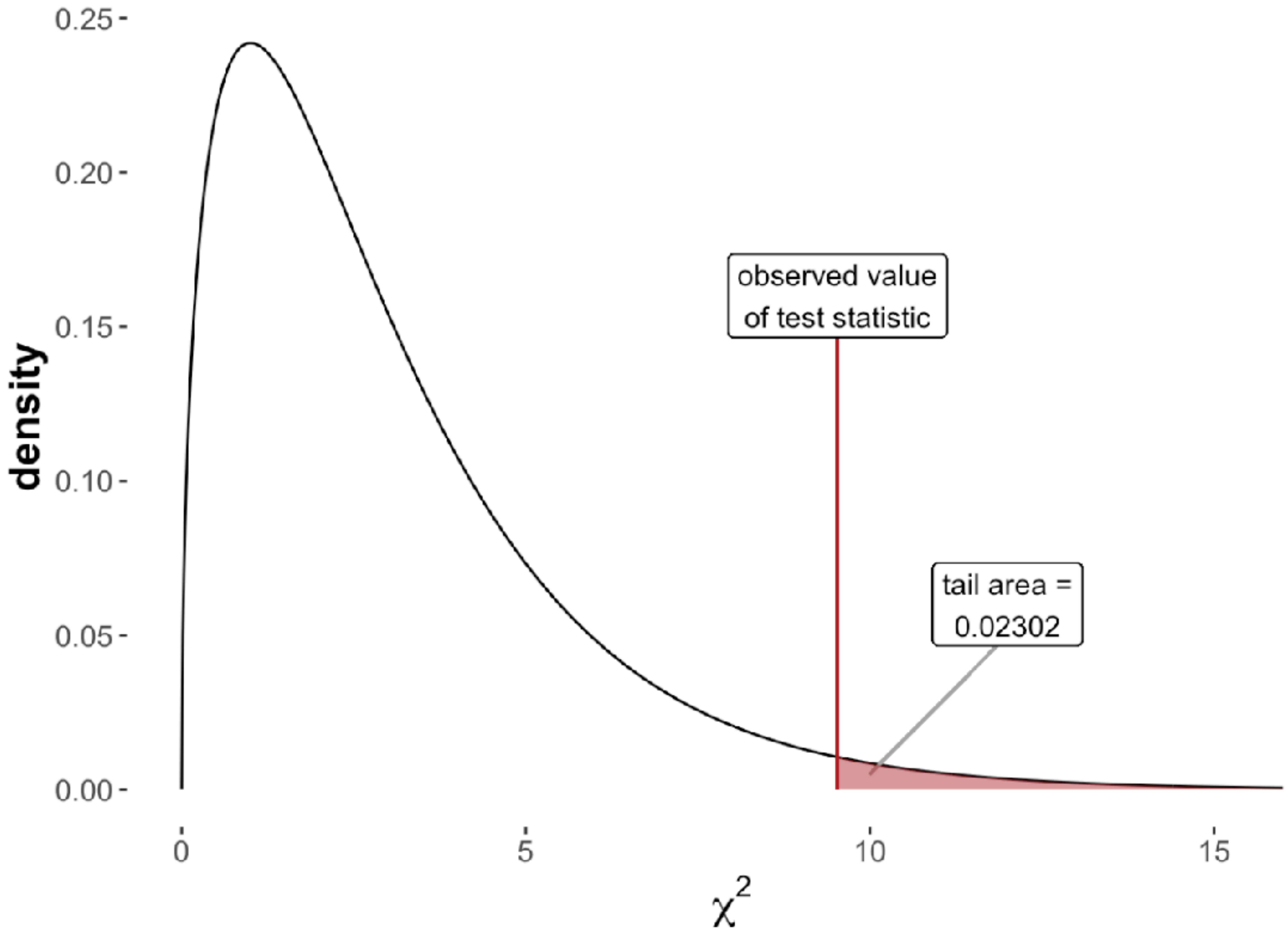
PEARSON'S χ^2-TEST [GOODNESS OF FIT]



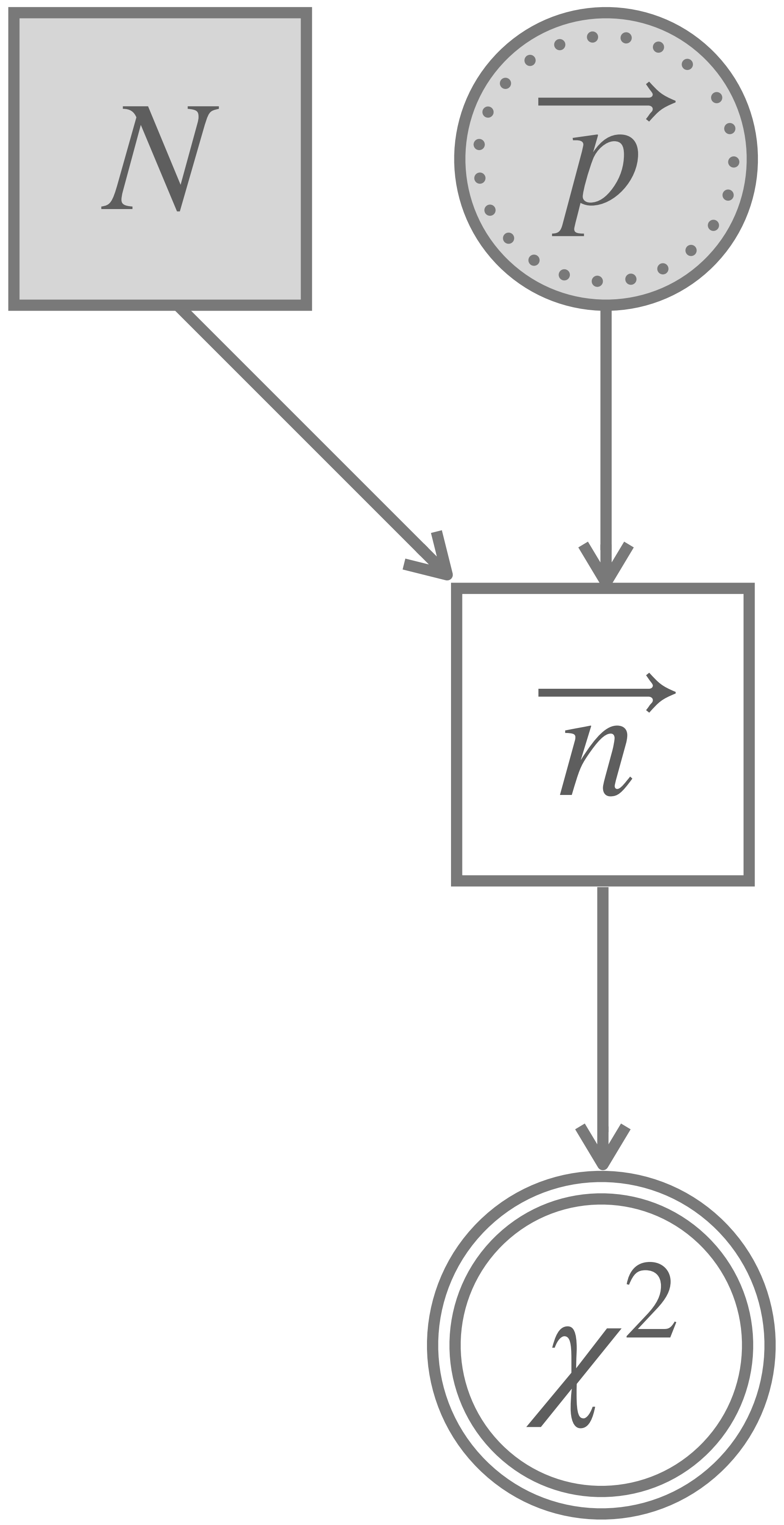
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

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

```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 = 🥰?

STOCHASTIC INDEPENDENCE

- events A and B are stochastically independent iff
  - intuitively: learning one does not change beliefs about the other;
  - formally: $P(A \mid B) = P(A)$
- notice that $P(A \mid B) = P(A)$ entails that $P(B \mid A) = P(B)$ (see web-book)

STOCHASTIC INDEPENDENCE

Proposition 7.1 (Probability of conjunction of stochastically independent events)
For any pair of events A and B with non-zero probability:

$$P(A \cap B) = P(A)\ P(B) \qquad [\text{if } A \text{ and } B \text{ are stoch. independent}]$$

Proof. By assumption of independence, it holds that $P(A \mid B) = P(A)$. But then:

$$\begin{aligned} P(A \cap B) &= P(A \mid B)\ P(B) && [\text{def. of conditional probability}] \\ &= P(A)\ P(B) && [\text{by ass. of independence}] \end{aligned}$$

□

Table 7.2: Joint probability table for a flip-and-draw scenario where the coin has a bias of 0.8 towards heads and where each of the two urns hold 3 black and 7 white balls.

	heads	tails	Σ rows
black	$0.8 \times 0.3 = 0.24$	$0.2 \times 0.3 = 0.06$	0.3
white	$0.8 \times 0.7 = 0.56$	$0.2 \times 0.7 = 0.14$	0.7
Σ columns	0.8	0.2	1.0

PEARSON'S χ^2-TEST [INDEPENDENCE]

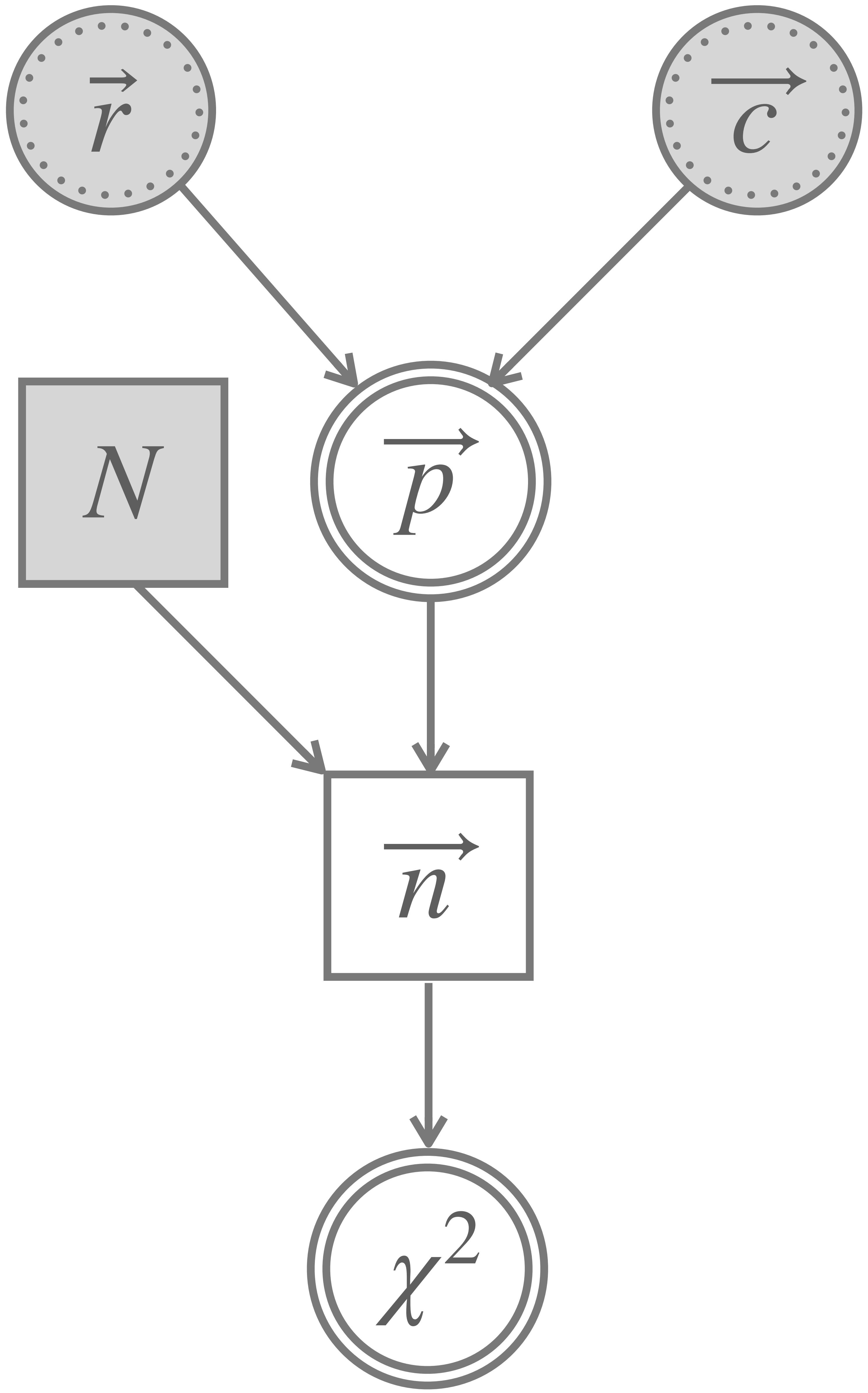
```
BLJM_table <- BLJM_associated_counts %>%
  select(-category) %>%
  pivot_wider(names_from = LB, values_from = n)
BLJM_table
```

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

Is it conceivable that the an outcome in each cell is given by independent choices of row and column options?

Hence: is the probability of a choice of cell the product of the probability of row- and column choices?

FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]



$\vec{p}$ = vec. of outer product $\vec{r}$ & $\vec{c}$

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

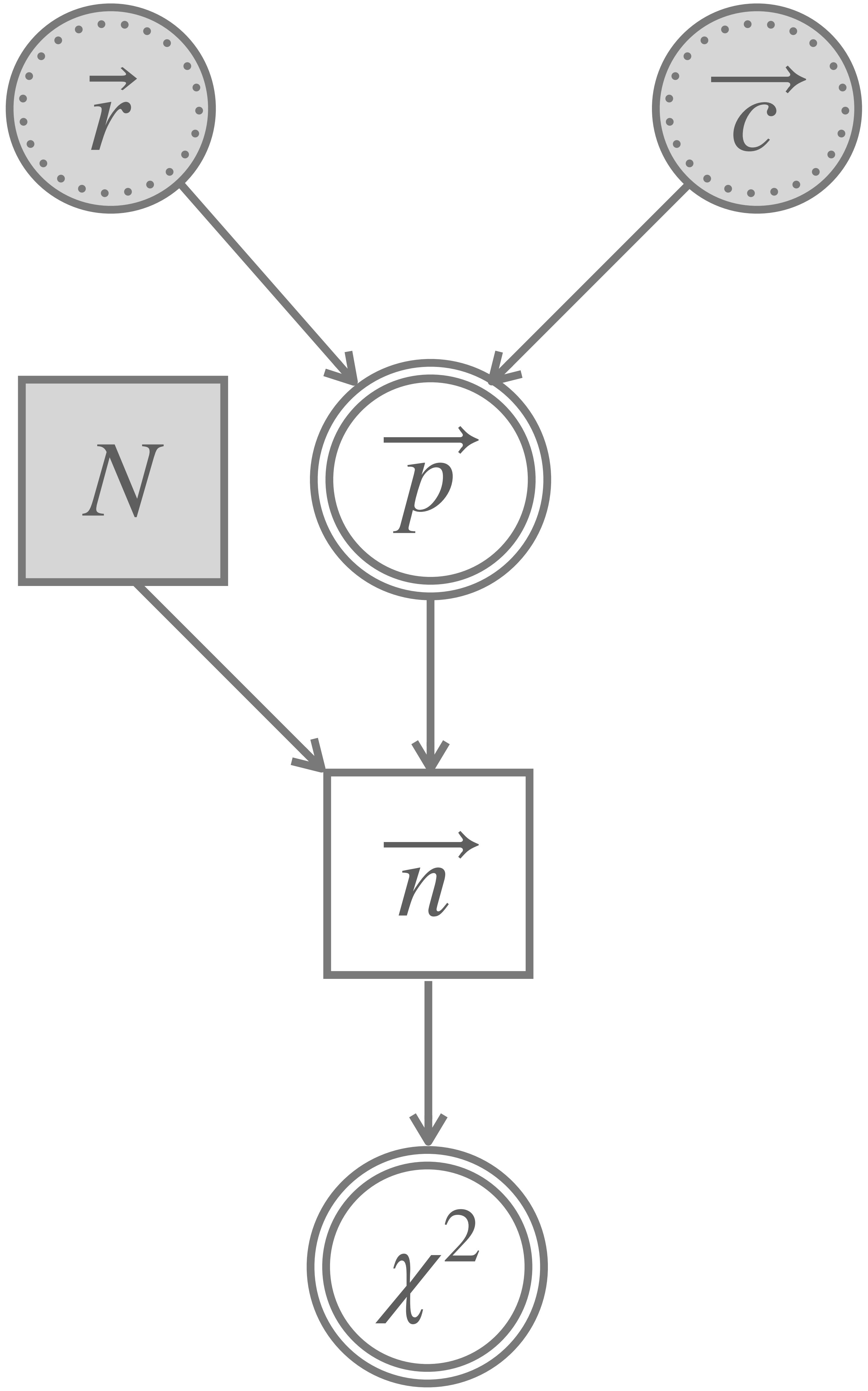
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

FACT:

The sampling distribution of χ^2 is approximately:

$\chi^2 \sim \chi^2\text{-distribution}\left((k_r - 1) \cdot (k_c - 1)\right)$

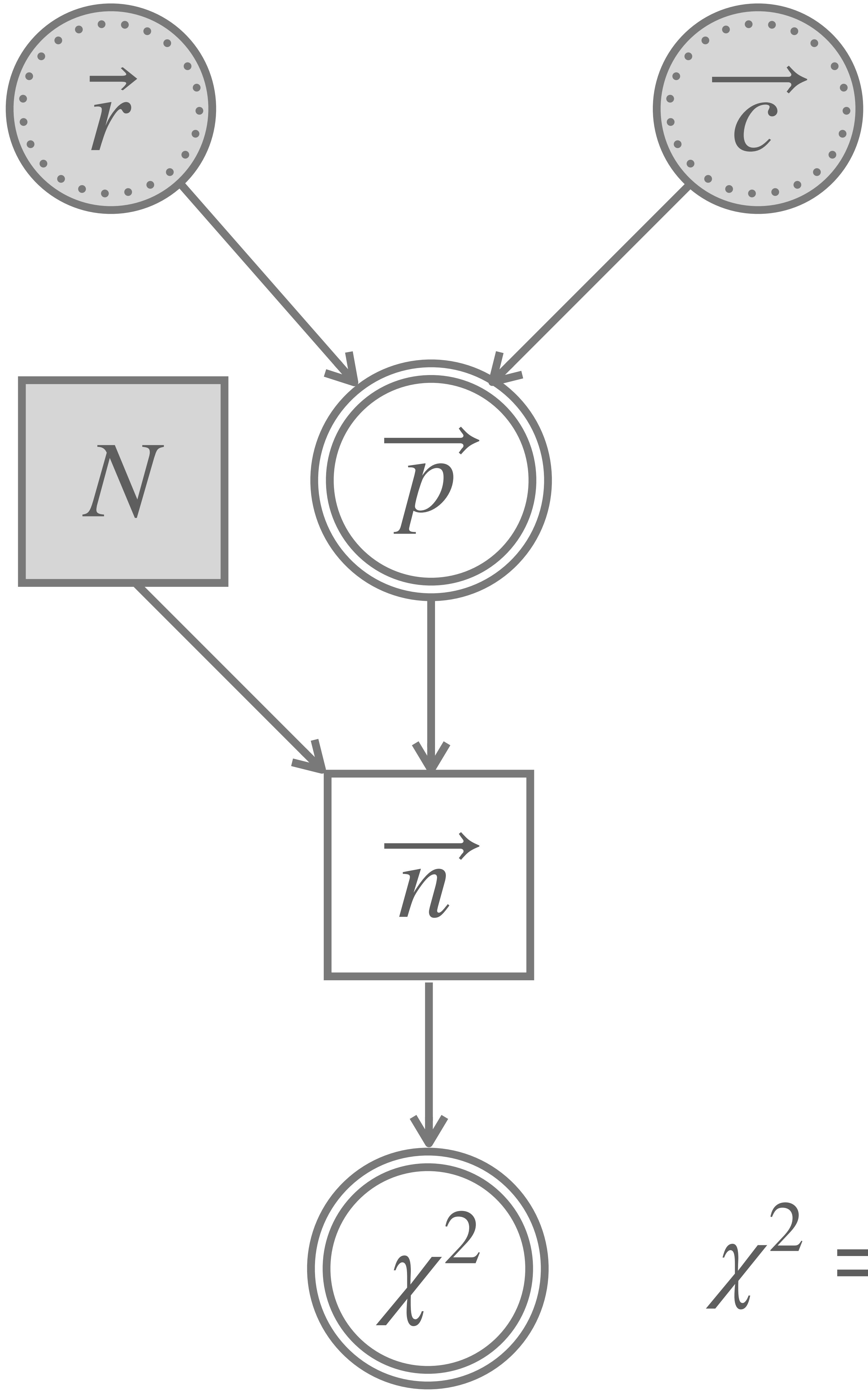
FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]




```r
# number of observations in total
N <- sum(counts_BLJM_choice_pairs_matrix)
# marginal proportions observed in the data
# the following is the vector r in the model graph
row_prob <- counts_BLJM_choice_pairs_matrix %>% rowSums() / N
# the following is the vector c in the model graph
col_prob <- counts_BLJM_choice_pairs_matrix %>% colSums() / N
# table of expected observation under independence assumption
# NB: %o% is the outer product of vectors
BLJM_expectation_matrix <- (row_prob %o%  col_prob) * N
BLJM_expectation_matrix
```

```
##          Biology    Logic
## Jazz   36.39216 27.60784
## Metal 21.60784 16.39216
```

FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]

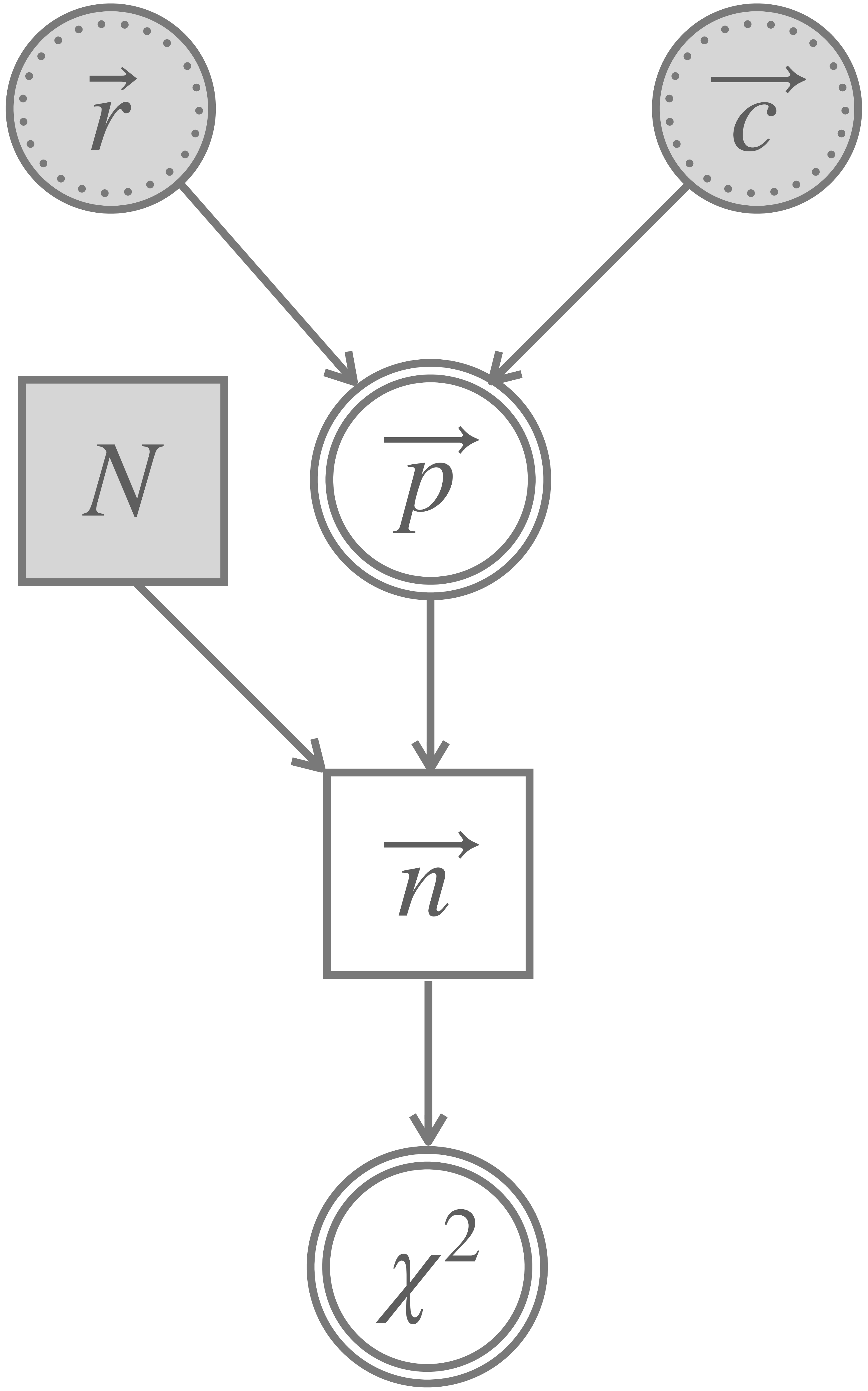


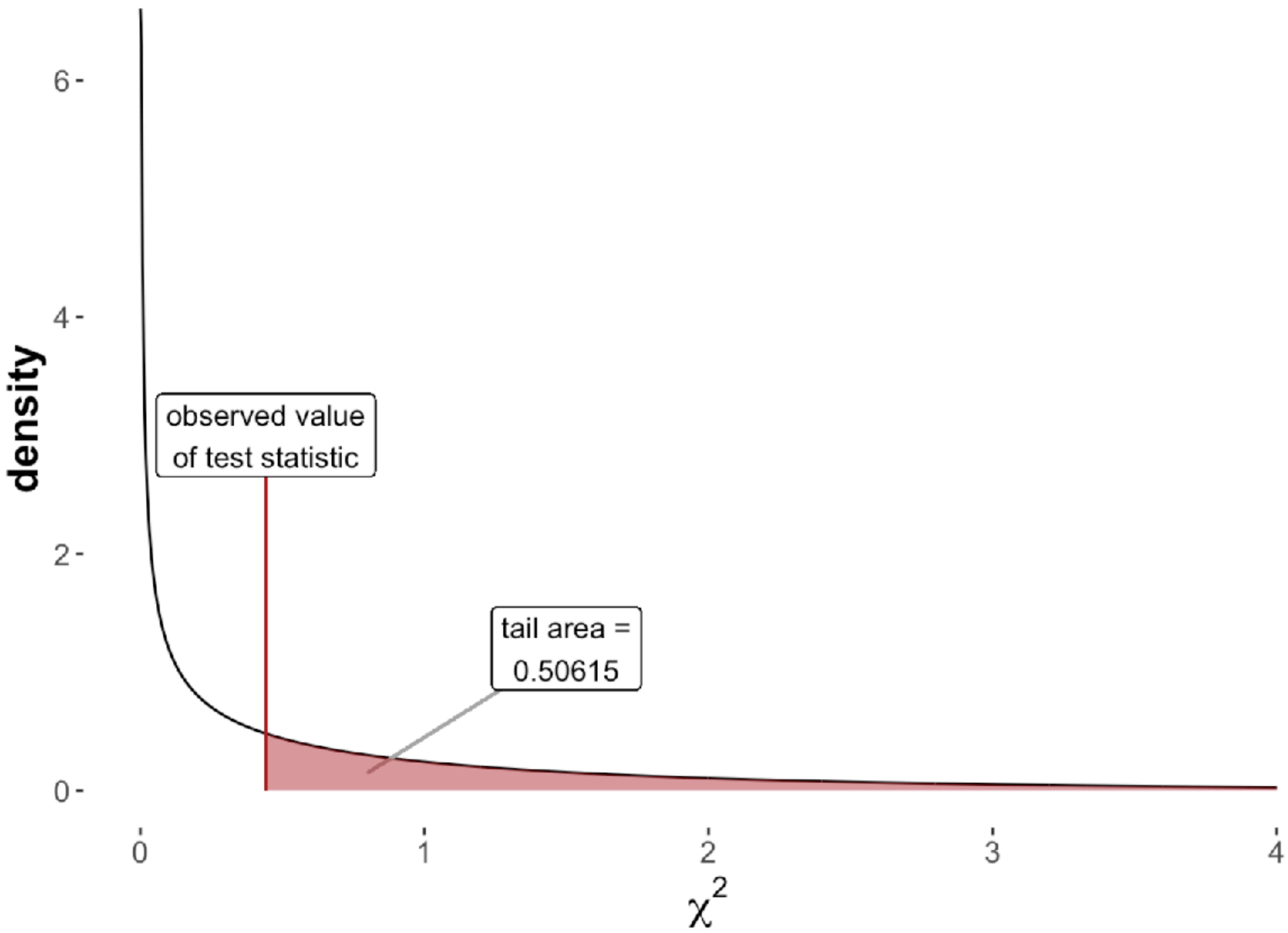

```r
chi2_observed <- sum(
  (counts_BLJM_choice_pairs_matrix - BLJM_expectation_matrix)^2 /
    BLJM_expectation_matrix
  )
p_value_BLJM <- 1-pchisq(q = chi2_observed, df = 1)
round(p_value_BLJM,5)
```

```
## [1] 0.50615
```

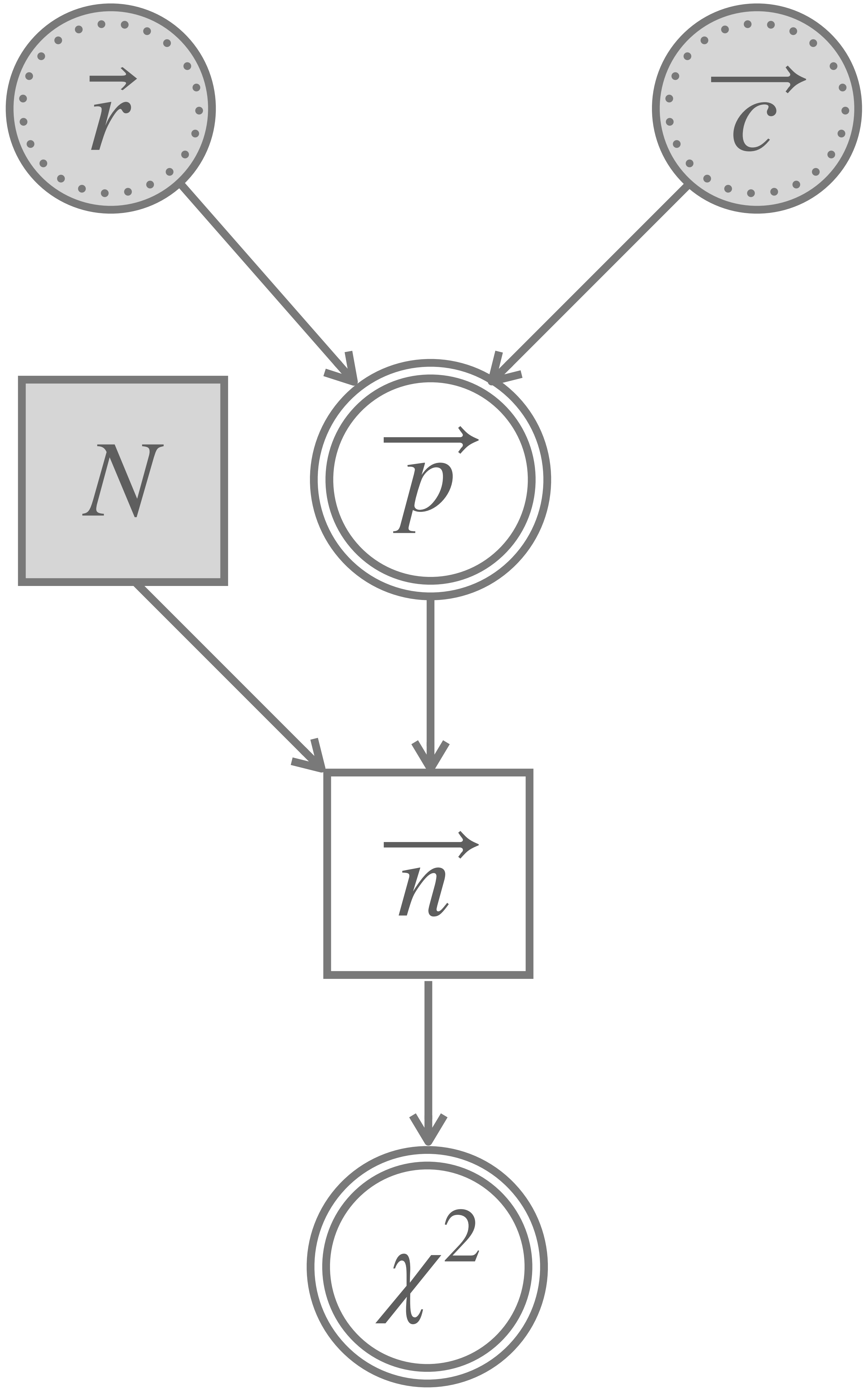
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]





FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]




```r
chisq.test(
  # supply data as a matrix, not as a vector, for test of independence
  counts_BLJM_choice_pairs_matrix,
  # do not use a the default correction (because we didn't introduce it)
  correct = FALSE
)
```

```
##
##  Pearson's Chi-squared test
##
## data:  counts_BLJM_choice_pairs_matrix
## X-squared = 0.44202, df = 1, p-value = 0.5061
```

FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]

How to interpret / report the result:

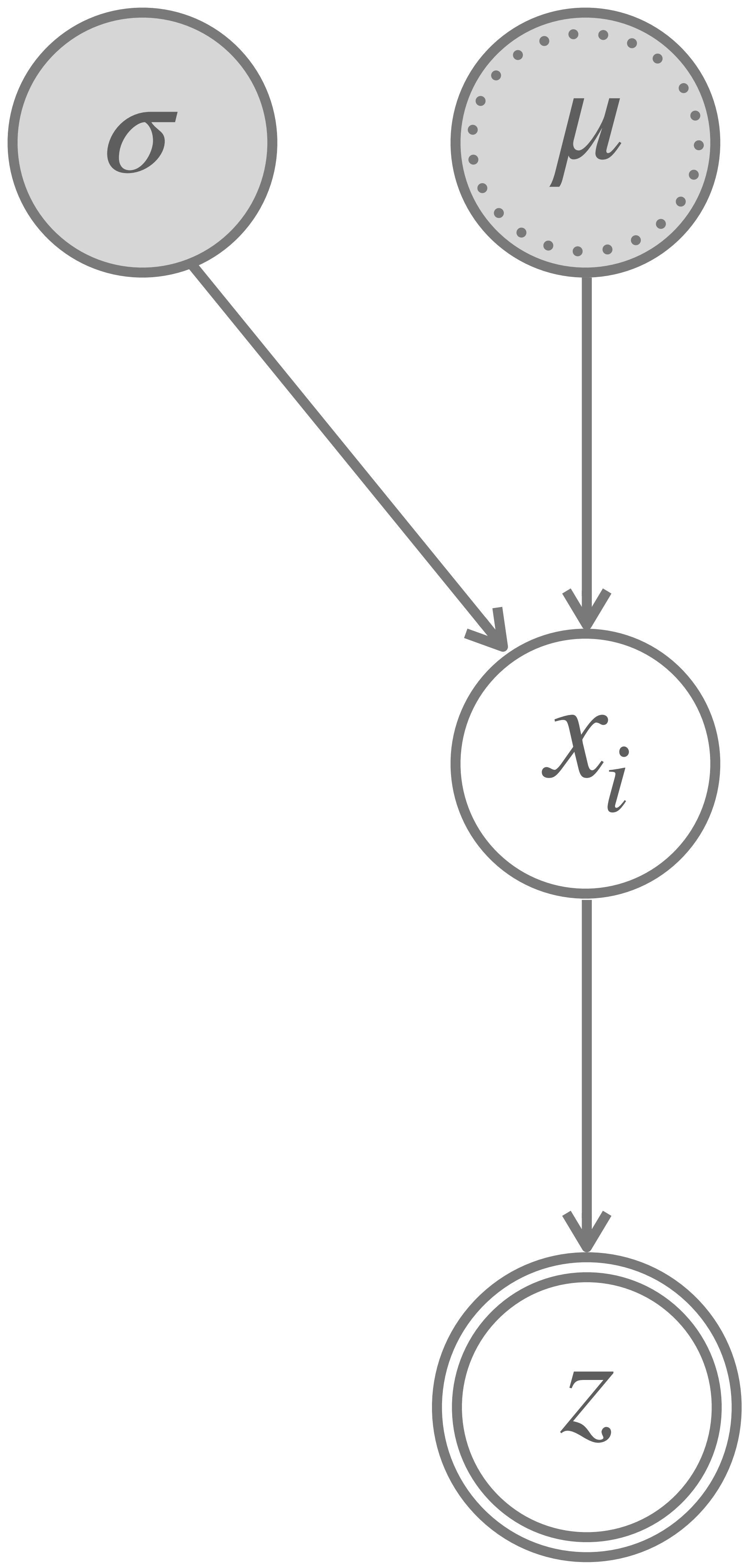
> A χ^2-test of independence did not yield a significant test result (χ^2-test, with $\chi^2 \approx 0.44$, $df = 1$ and $p \approx 0.5$). Therefore, we cannot claim to have found any evidence for the research hypothesis of dependence.

z-test

FREQUENTIST MODEL FOR A z-TEST [ONE-SAMPLE]

- metric variable $\overrightarrow{x}$ with samples from normal distribution
  - unknown μ
  - known σ [usually unrealistic!]

FREQUENTIST MODEL FOR A z-TEST [ONE-SAMPLE]



$$x_i \sim \text{Normal}(\mu, \sigma)$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$$

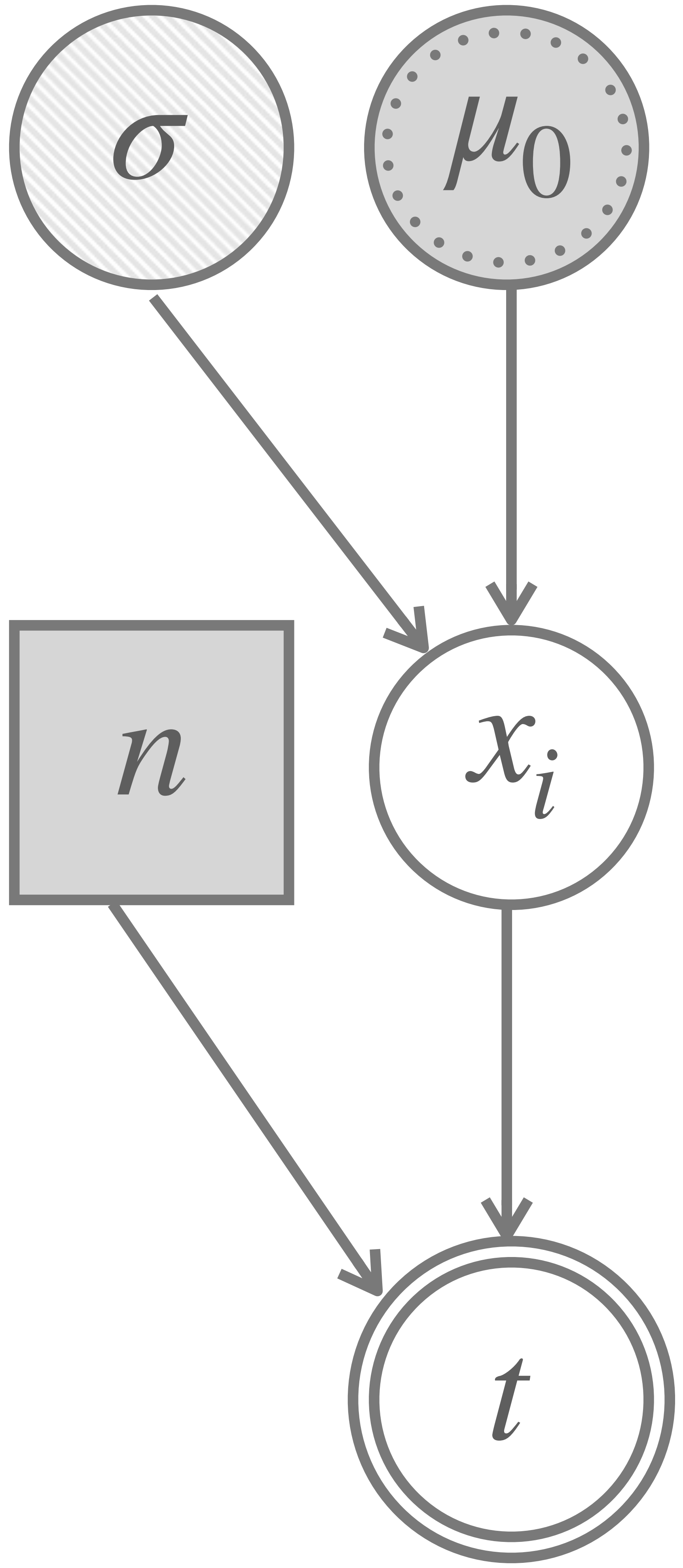
FACT:

With known σ, the distribution of z is:

$$z \sim \text{Normal}(0,1)$$

t-tests

FREQUENTIST T-TEST MODEL [ONE-SAMPLE]



$$x_i \sim \text{Normal}(\mu_0, \sigma)$$

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}_x / \sqrt{n}}$$

FACT:

Irrespective of σ, the distribution of t is:

$$t \sim \text{Student-t}(\nu = n - 1)$$