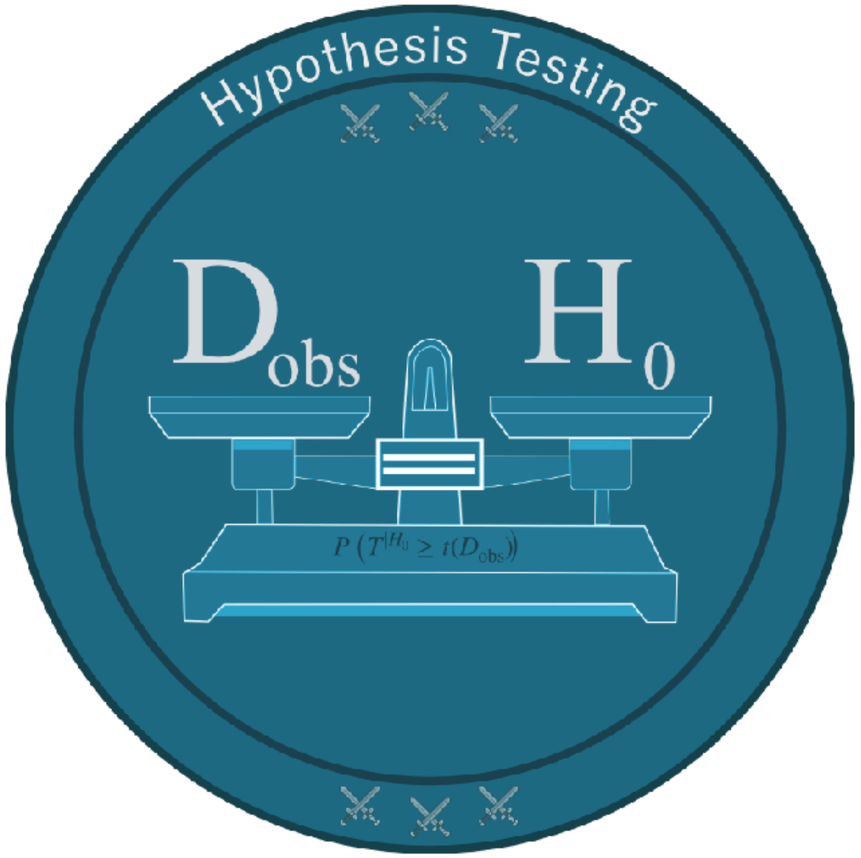


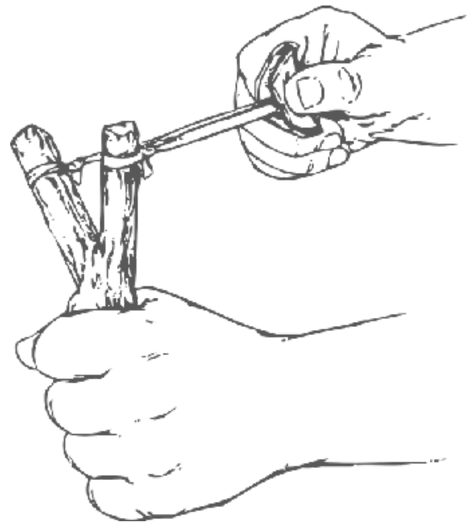
INTRODUCTION TO DATA ANALYSIS

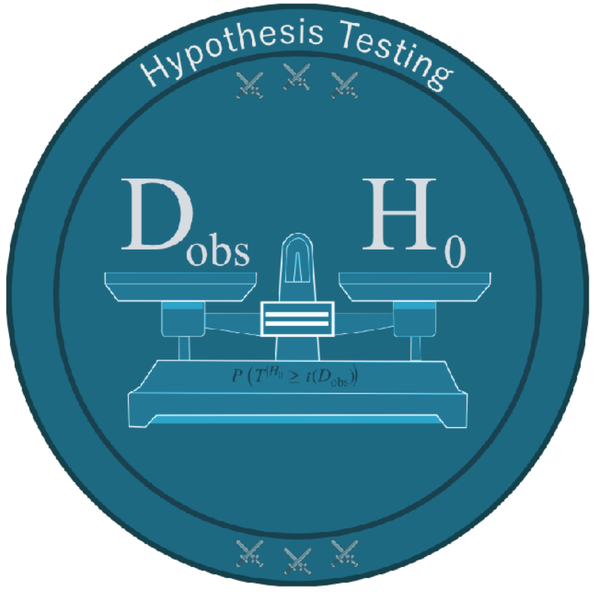
HYPOTHESIS TESTING

PART III

LEARNING GOALS

- become able to interpret & apply some statistical tests
  - Pearson's χ^2-tests of independence
  - *z*-test
  - one-sample *t*-test
  - two-sample *t*-test
  - one-way ANOVA
- understand differences and commonalities of different approaches to frequentist testing
  - Fisher
  - Neyman/Pearson
  - modern hybrid NHST





P-VALUE

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

Pearson's χ^2-test

goodness of fit

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

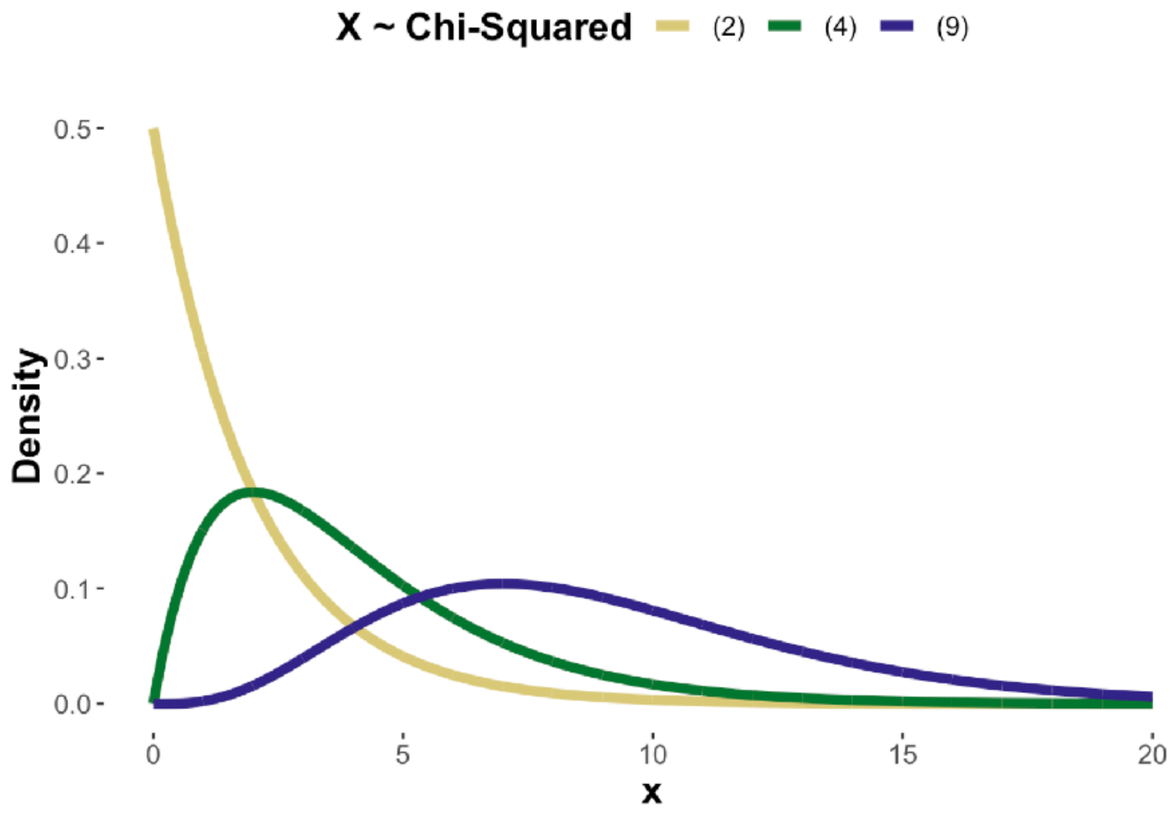
χ^2-DISTRIBUTION

- standard normal random variables:

  $X_1, \ldots X_n$
- derived RV:

  $Y = X_1^2 + \ldots + X_n^2$
- it follows (by construction) that:

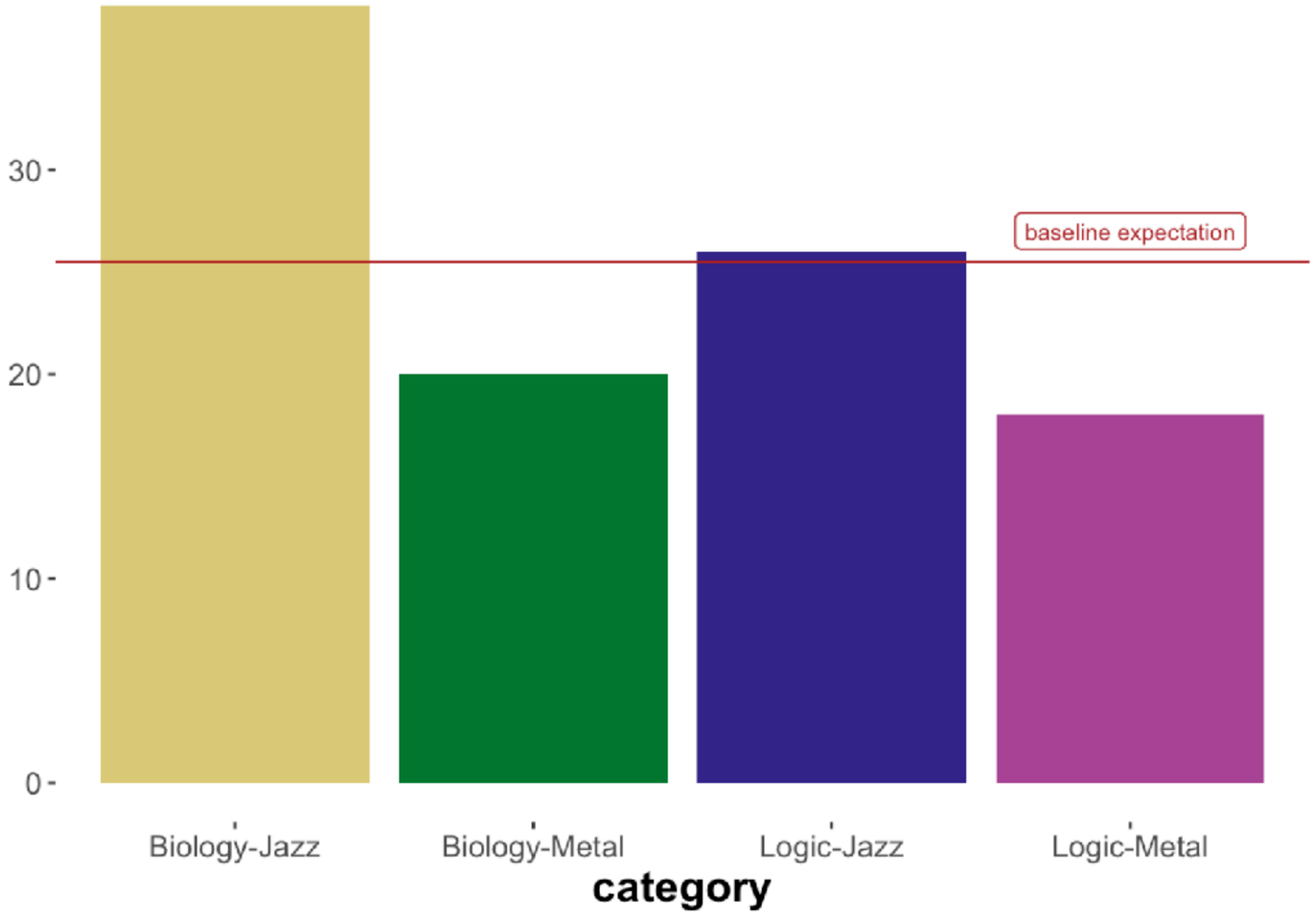
  $y \sim \chi^2\text{-distribution}(n)$



PEARSON'S χ^2-TEST [GOODNESS OF FIT]

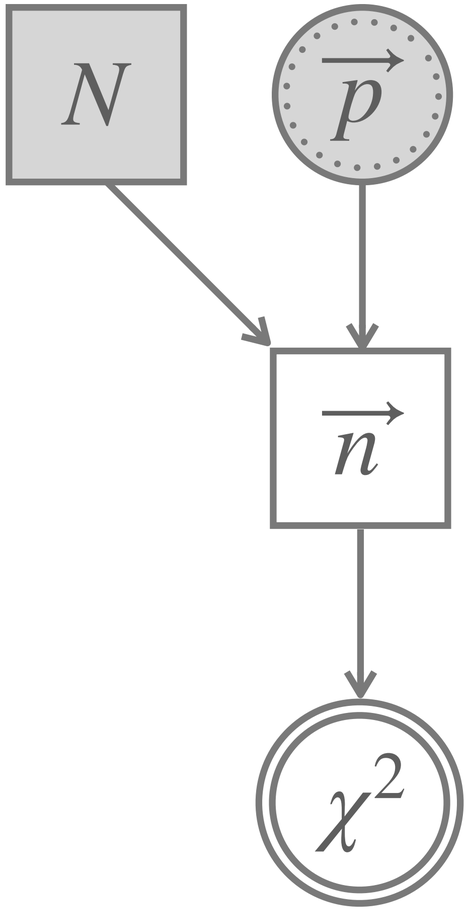
```
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts
```

```
## # 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]



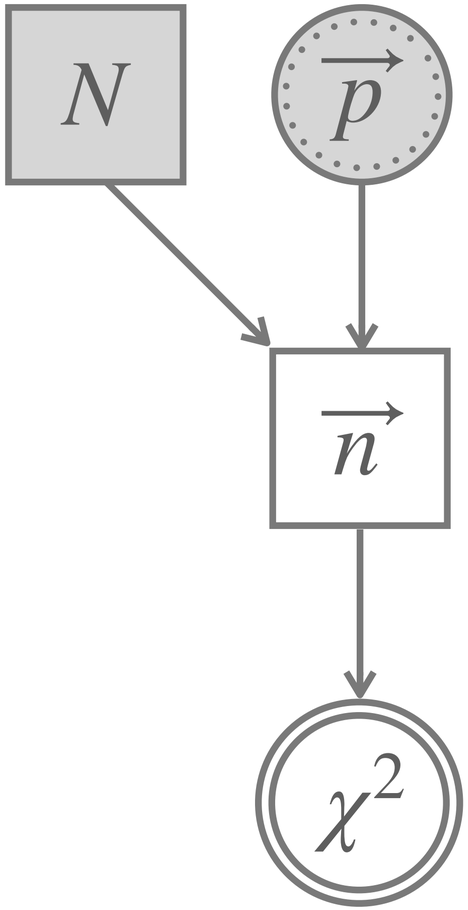
$$\vec{n} \sim \text{Multinomial}(\vec{p}, N)$$

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

Sampling distribution:

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

PEARSON'S χ^2-TEST [GOODNESS OF FIT]



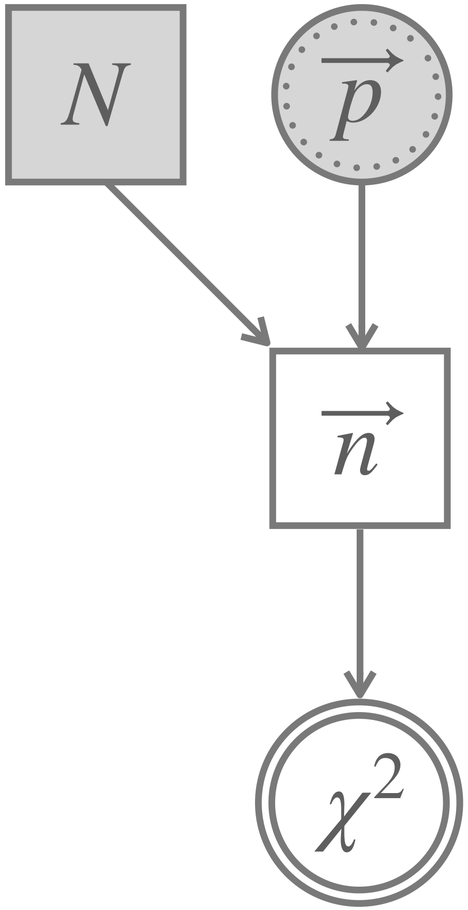
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

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

```
# 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
```

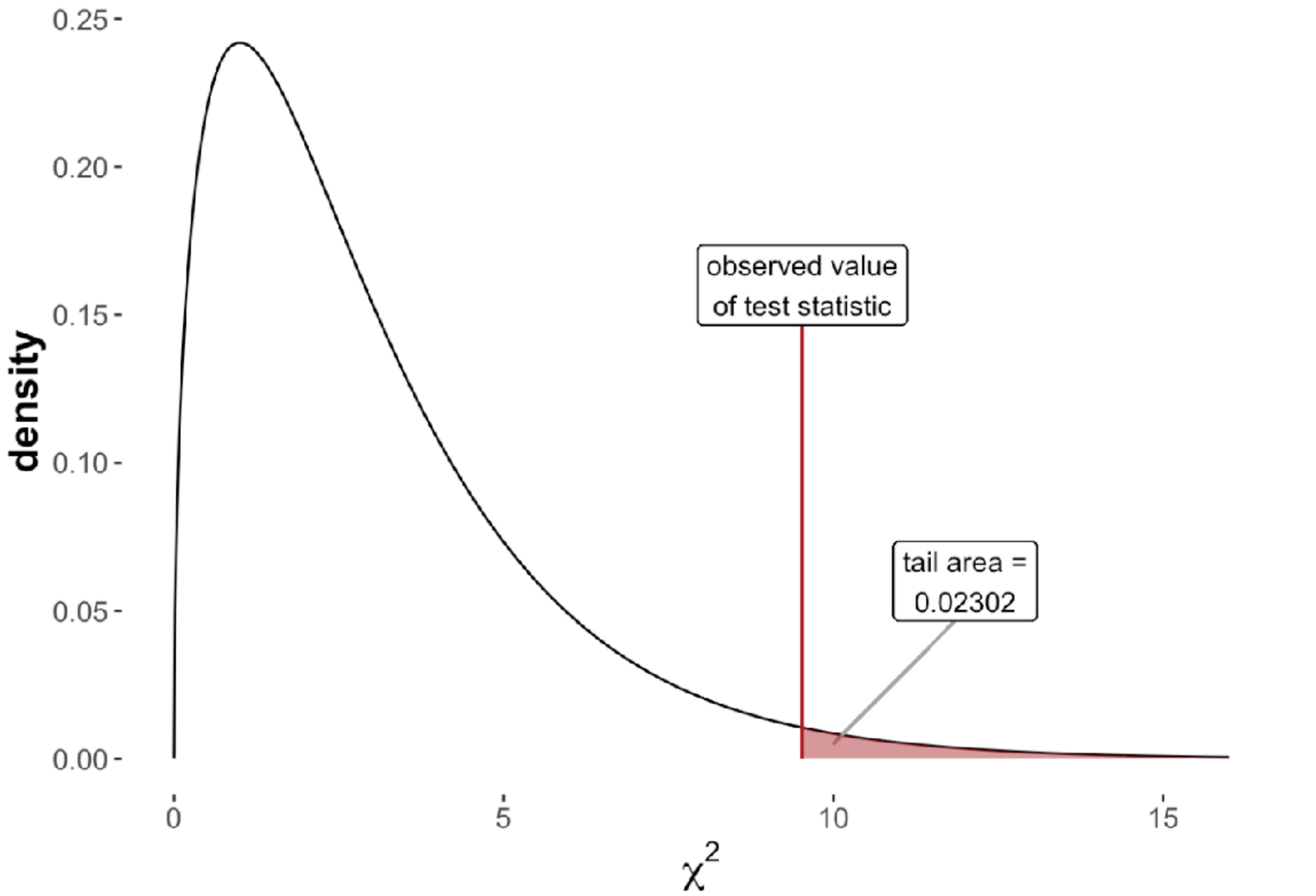
PEARSON'S χ^2-TEST [GOODNESS OF FIT]



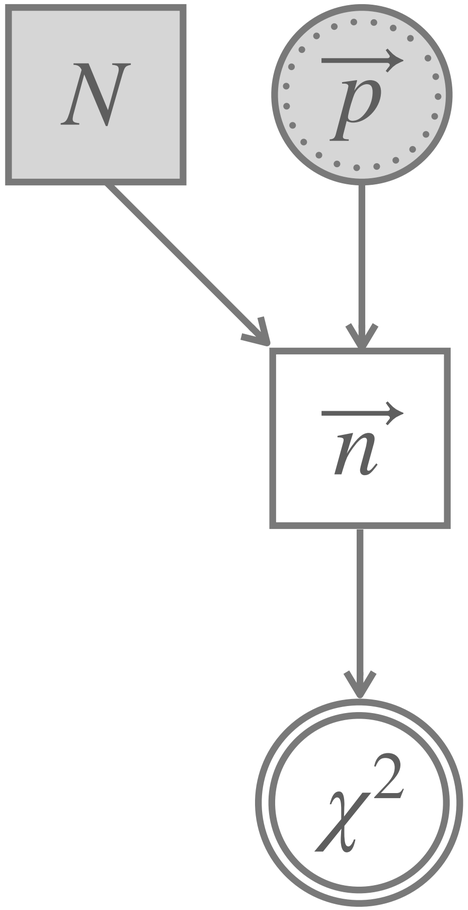
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

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

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



PEARSON'S χ^2-TEST [GOODNESS OF FIT]



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

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

```r
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 = 🥰?

Pearson's χ^2-test

independence

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:

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

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]

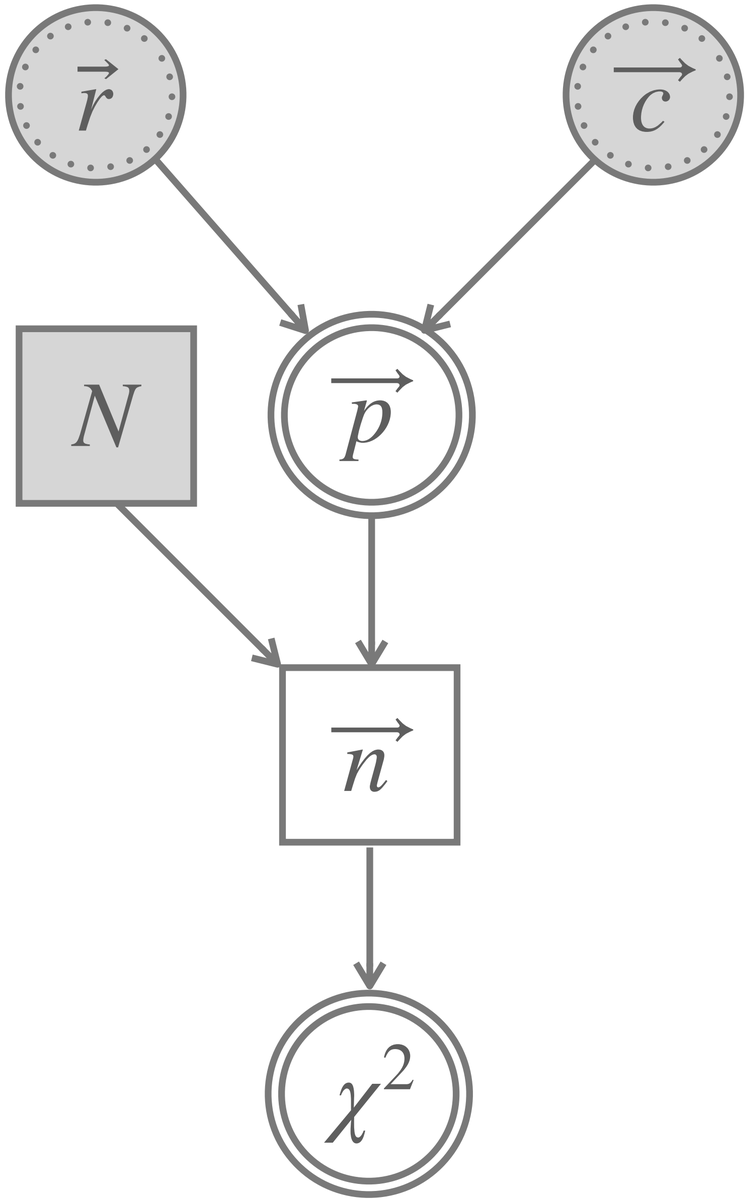
```
BLJM_table <- BLJM_associated_counts %>%
  select(-category) %>%
  pivot_wider(names_from = LB, values_from = n)
BLJM_table
```

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

Is it conceivable that the 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]



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

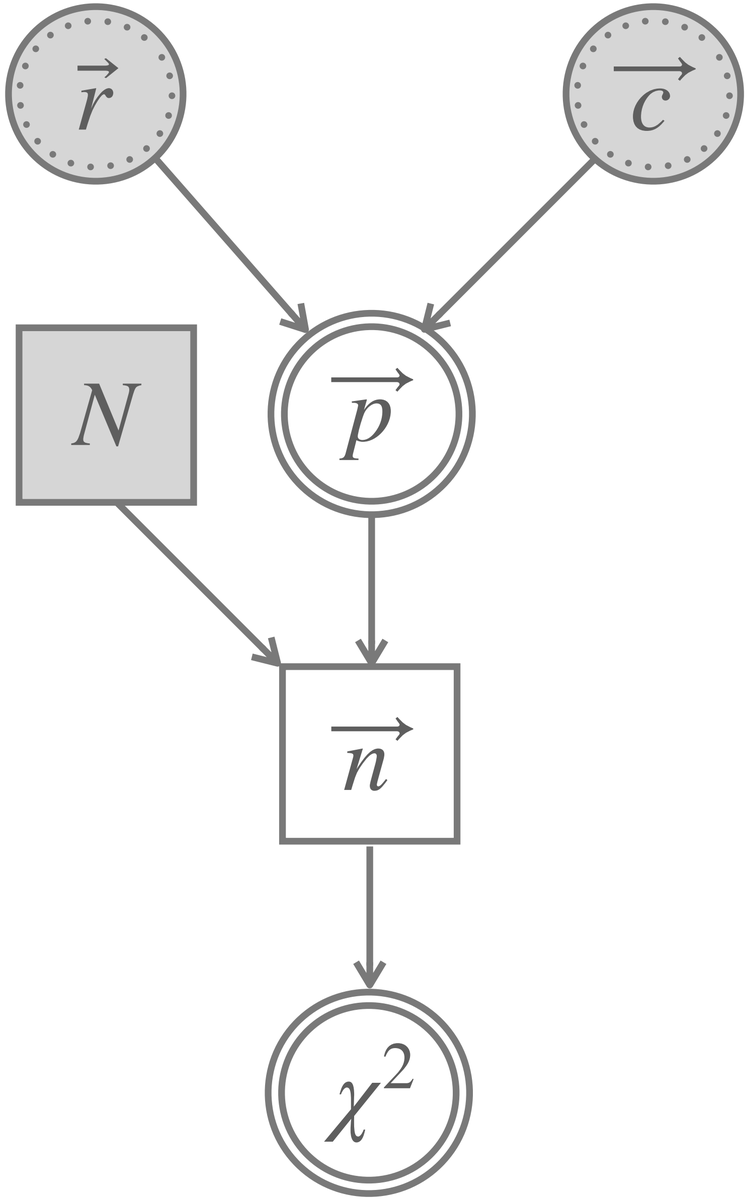
$$\vec{n} \sim \text{Multinomial}(\vec{p}, N)$$

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

Sampling distribution:

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

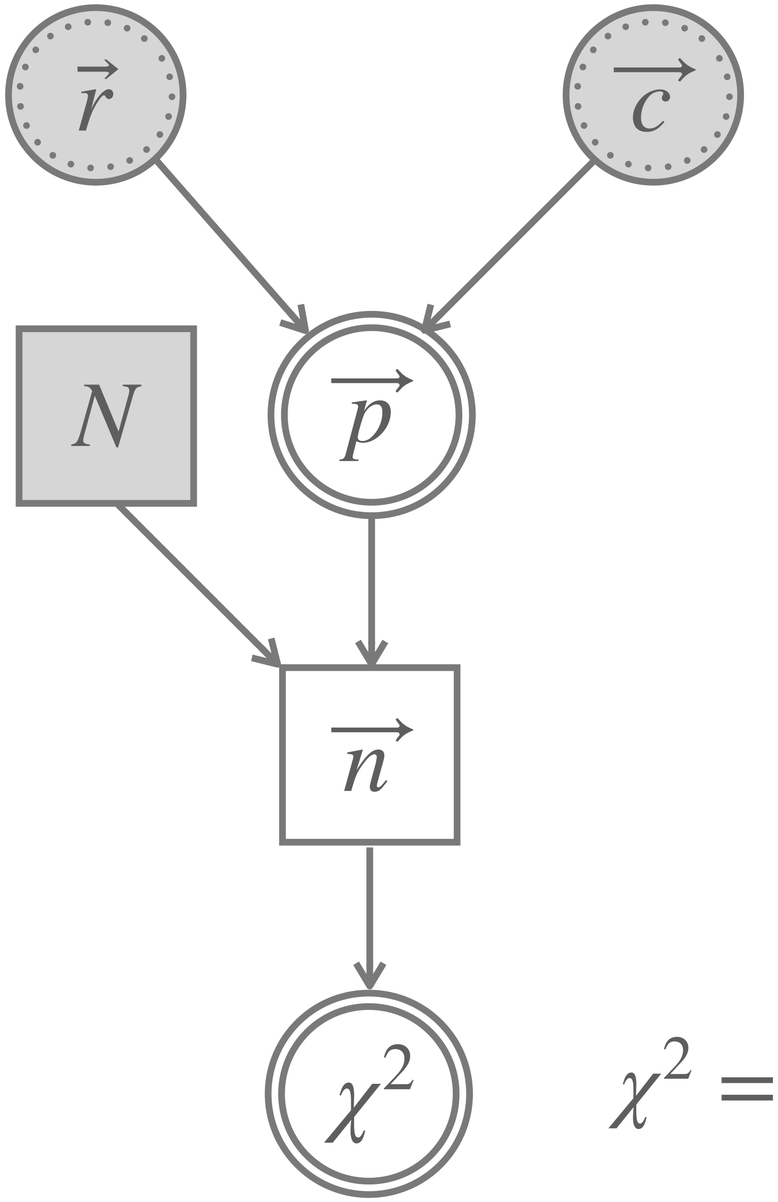
FREQUENTIST MODEL FOR PEARSON'S χ^2-TEST [INDEPENDENCE]




```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]

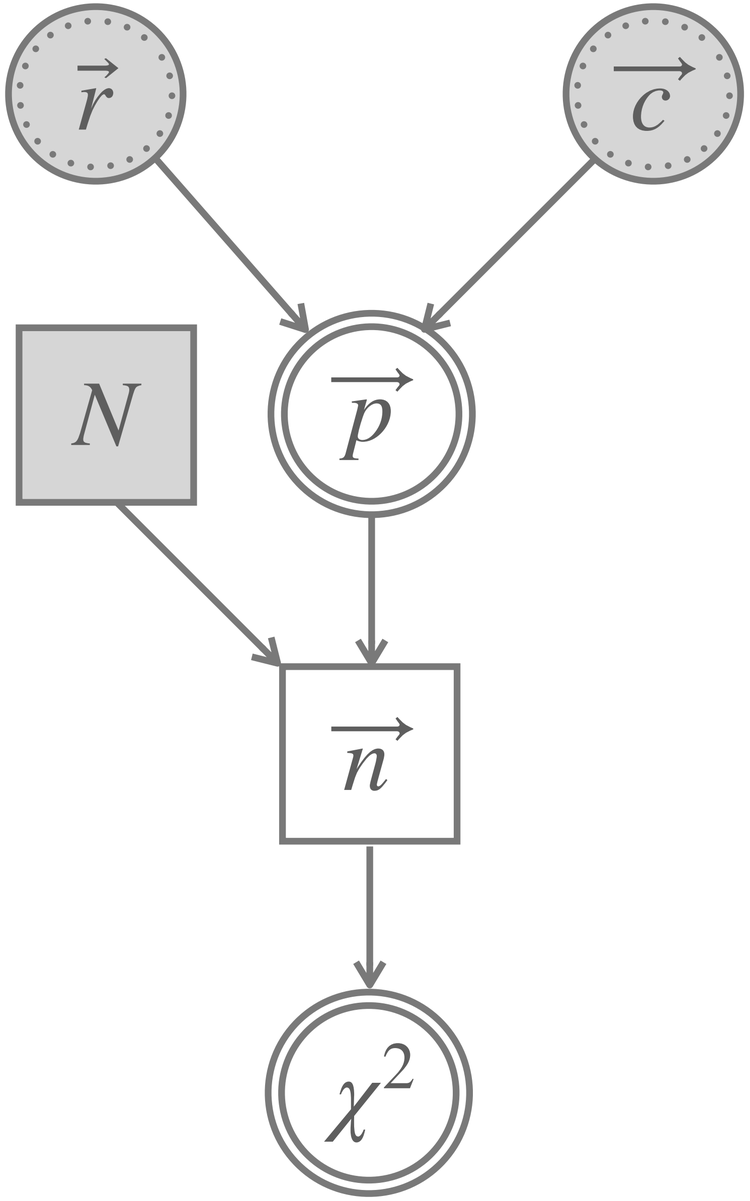


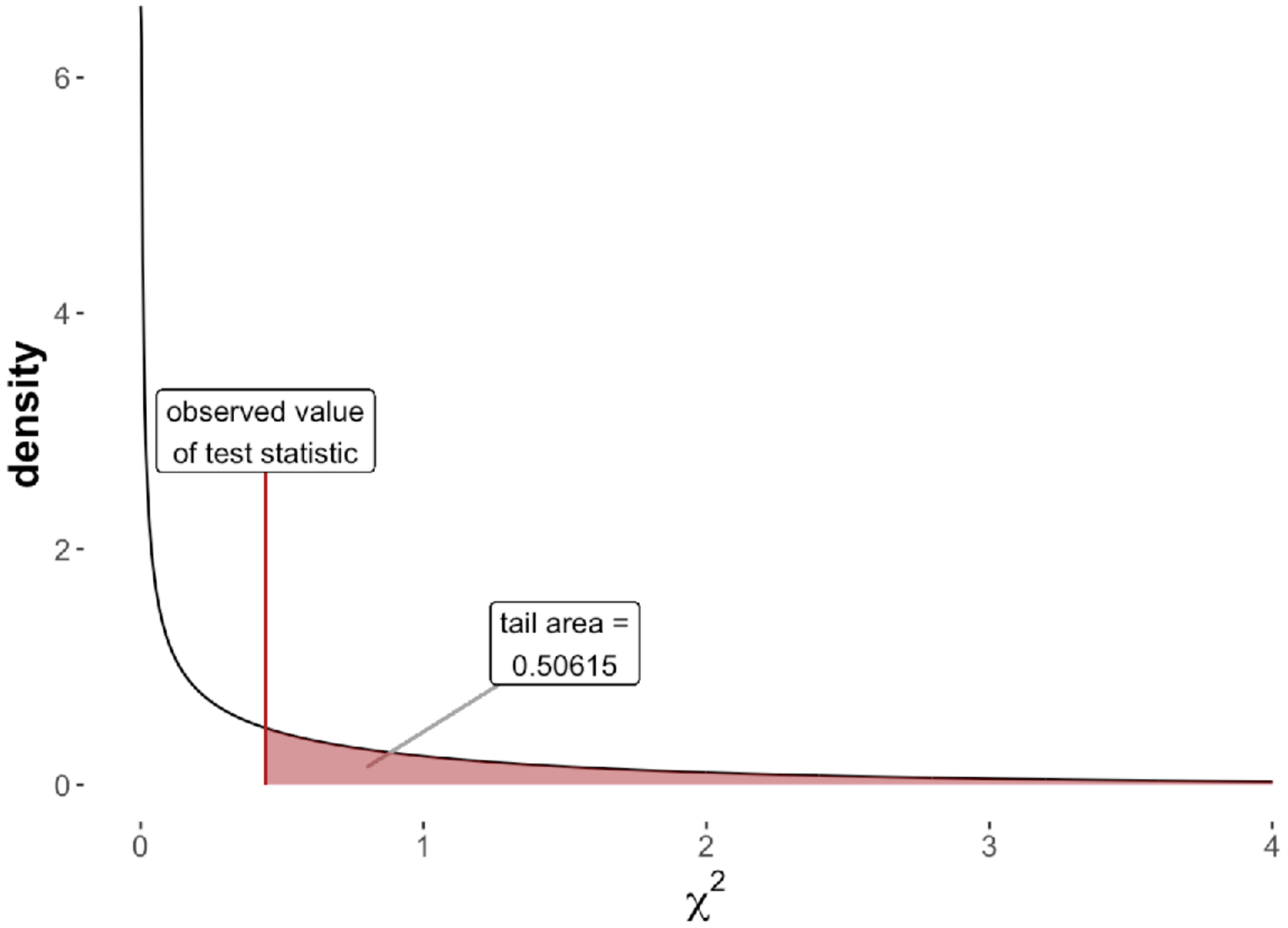

```r
chi2_observed <- sum(
  (counts_BLJM_choice_pairs_matrix - BLJM_expectation_matrix)^2 /
    BLJM_expectation_matrix
  )
p_value_BLJM <- 1-pchisq(q = chi2_observed, df = 1)
round(p_value_BLJM,5)
```

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

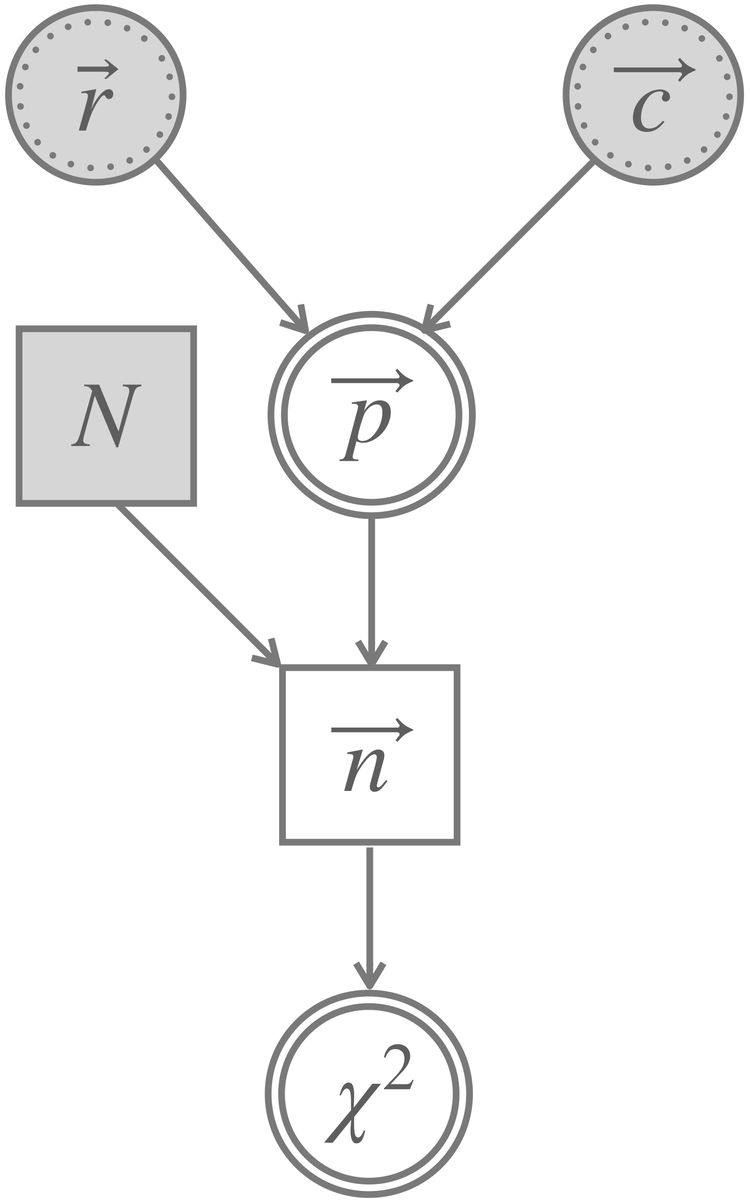
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

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

SCENARIO FOR A z-TEST [ONE-SAMPLE]

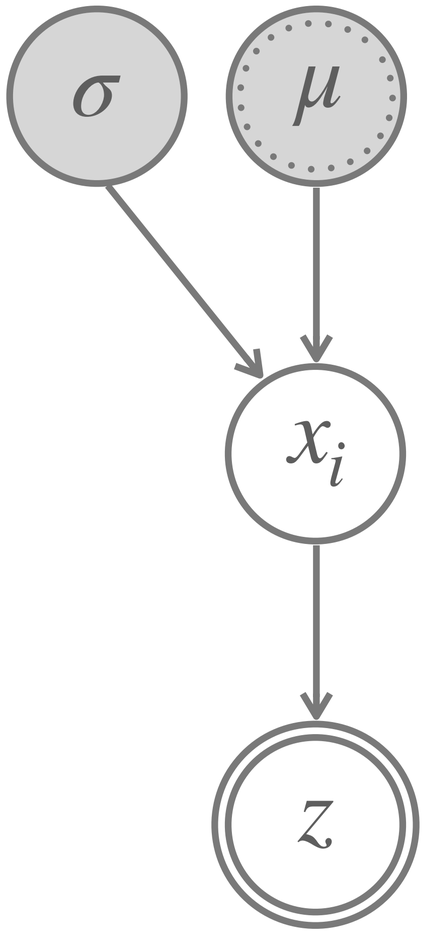
- metric variable $\overrightarrow{x}$ with samples from normal distribution
  - unknown μ
  - known σ [usually unrealistic!]

```
IQ_data <- c(87, 91, 93, 97, 100, 101, 103, 104,
             104, 105, 105, 106, 108, 110, 111,
             112, 114, 115, 119, 121)
mean(IQ_data)
```

```
## [1] 105.3
```

Is it plausible to maintain that this data was generated by a normal distribution with mean 100 (if we assume that the standard deviation is known to be 15)?

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



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

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

Sampling distribution:

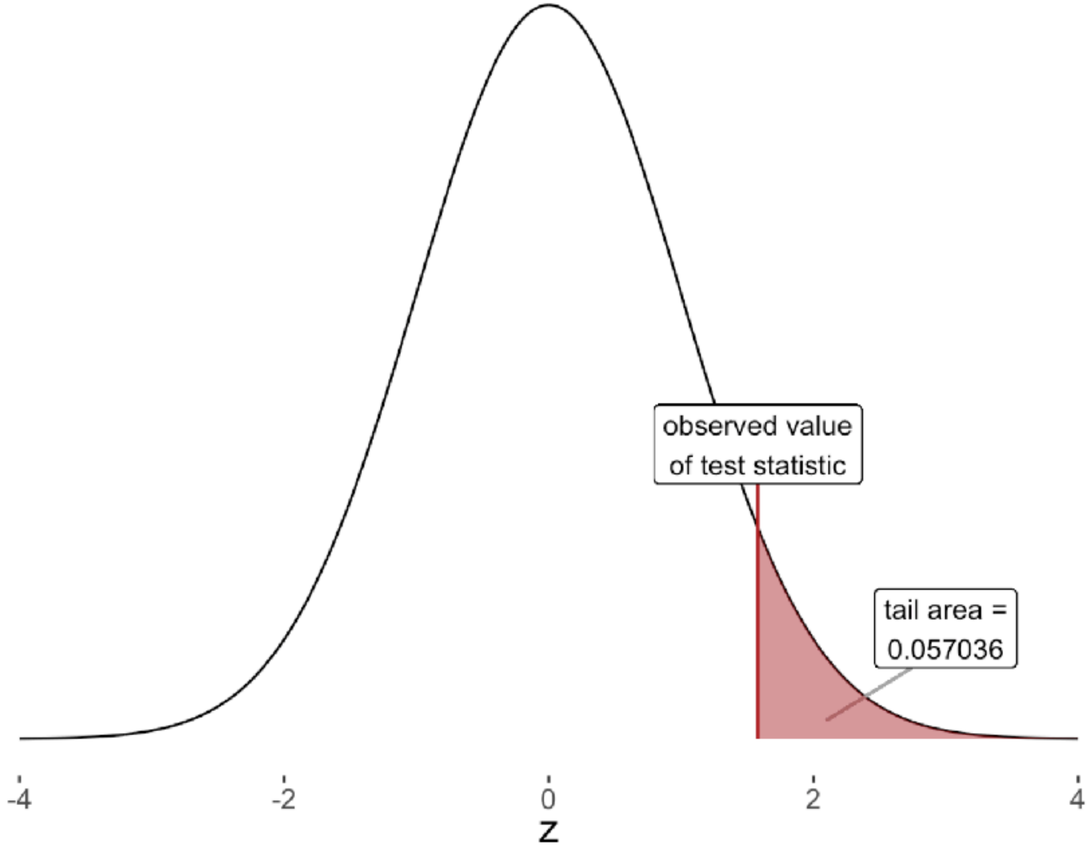
$z \sim \text{Normal}(0,1)$

FREQUENTIST Z-TEST [APPLICATION]

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

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

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




```r
# number of observations
N <- length(IQ_data)
# null hypothesis to test
mu_0 <- 100
# standard deviation (known/assumed as true)
sd <- 15
z_observed <-  (mean(IQ_data) - mu_0) / (sd / sqrt(N))
z_observed %>% round(4)
```

```
## [1] 1.5802
```

```r
p_value_IQ_data <- 1 - pnorm(z_observed)
p_value_IQ_data %>% round(6)
```

```
## [1] 0.057036
```

FREQUENTIST Z-TEST [APPLICATION]

$$x_i \sim \text{Nor}$$

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

$$z \sim \text{No}$$

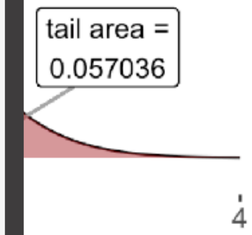
```
BSDA::z.test(x = IQ_data, mu = 100, sigma.x = 15, alternative = "greater")
```

```
##
##  One-sample z-Test
##
## data:  IQ_data
## z = 1.5802, p-value = 0.05704
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
##  99.78299       NA
## sample estimates:
## mean of x
##     105.3
```

```
# number of observa
N <- length(IQ_data
# null hypothesis t
mu_0 <- 100
# standard deviatio
sd <- 15
z_observed <-  (mea
z_observed %>% roun
```

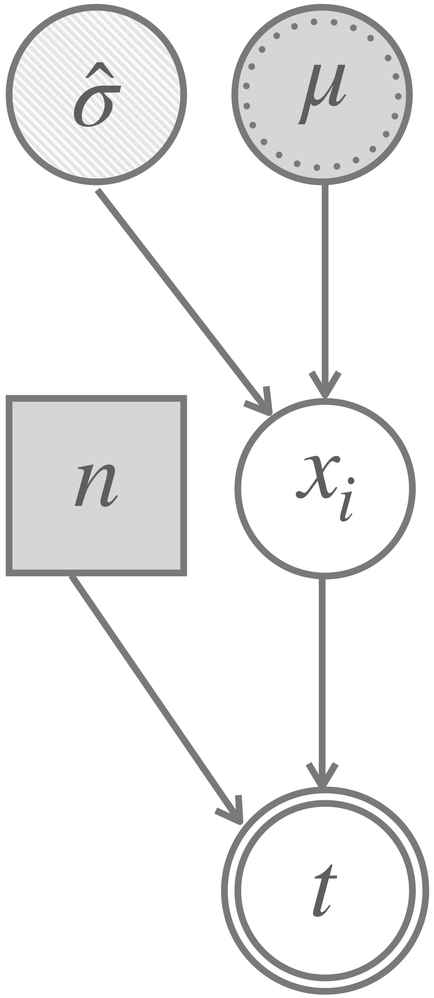
```
## [1] 1.5802
```

```
## [1] 0.057036
```



one-sample *t*-test

FREQUENTIST T-TEST MODEL [ONE-SAMPLE]



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

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{\vec{x}})^2}$$

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

Sampling distribution:

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

t-DISTRIBUTION

- two random variables:

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

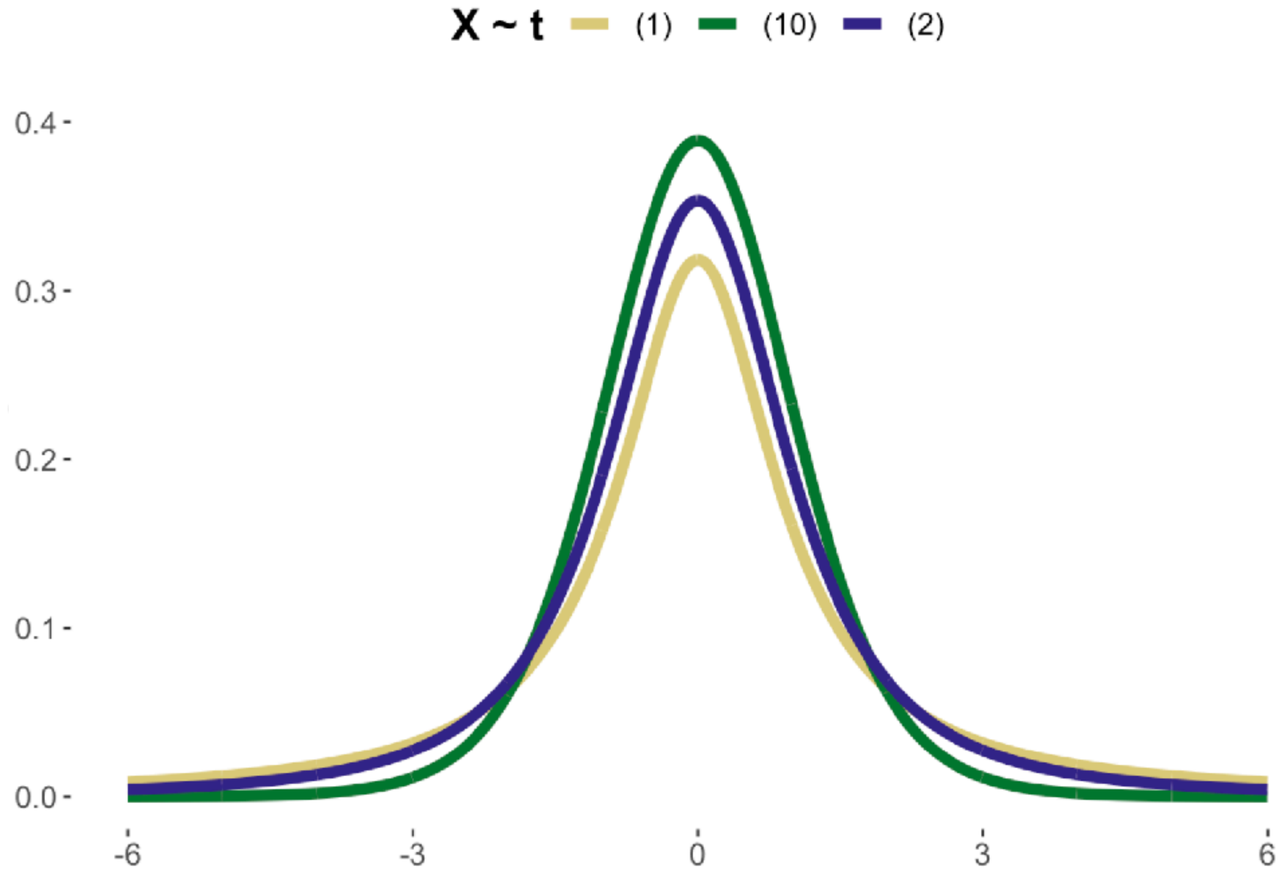
$$y \sim \chi^2\text{-distribution}(n)$$

- derived RV:

$$Z = \frac{X}{\sqrt{Y/n}}$$

- it follows (by construction) that:

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



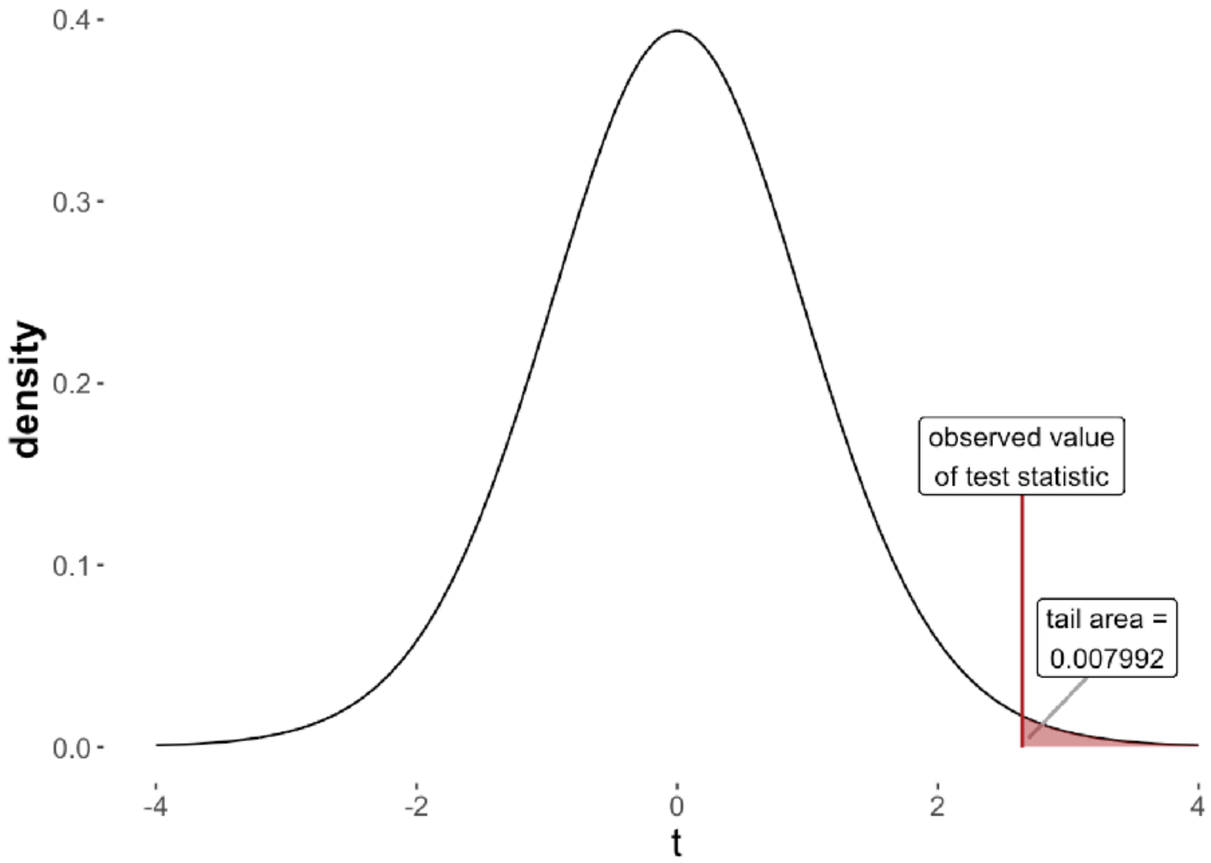
FREQUENTIST T-TEST [APPLICATION]

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

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{\overrightarrow{x}})^2}$$

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

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




```
N <- length(IQ_data)
# fix the null hypothesis
mean_0 <- 100
# unlike in a z-test we use the sample to estimate SD
sigma_hat <- sd(IQ_data)
t_observed <- (mean(IQ_data) - mean_0) / sigma_hat * sqrt(N)
t_observed %>% round(4)
```

```
## [1] 2.6446
```

```
p_value_t_test_IQ <- 1 - pt(t_observed, df = N-1)
p_value_t_test_IQ %>% round(6)
```

```
## [1] 0.007992
```

FREQUENTIST T-TEST [APPLICATION]

$x_i \sim \text{Norma}$

$\hat{\sigma} = \sqrt{\frac{1}{n-}}$

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

$t \sim \text{Student-}$

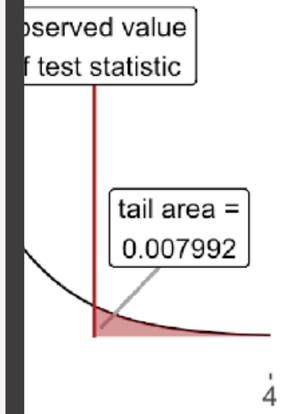
```r
t.test(x = IQ_data, mu = 100, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  IQ_data
## t = 2.6446, df = 19, p-value = 0.007992
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
##  101.8347      Inf
## sample estimates:
## mean of x
##     105.3
```

```r
N <- length(IQ_data)
# fix the null hypothesis
mean_0 <- 100
# unlike in a z-test we use
sigma_hat <- sd(IQ_data)
t_observed <- (mean(IQ_data)
t_observed %>% round(4)
```

```
## [1] 2.6446
```

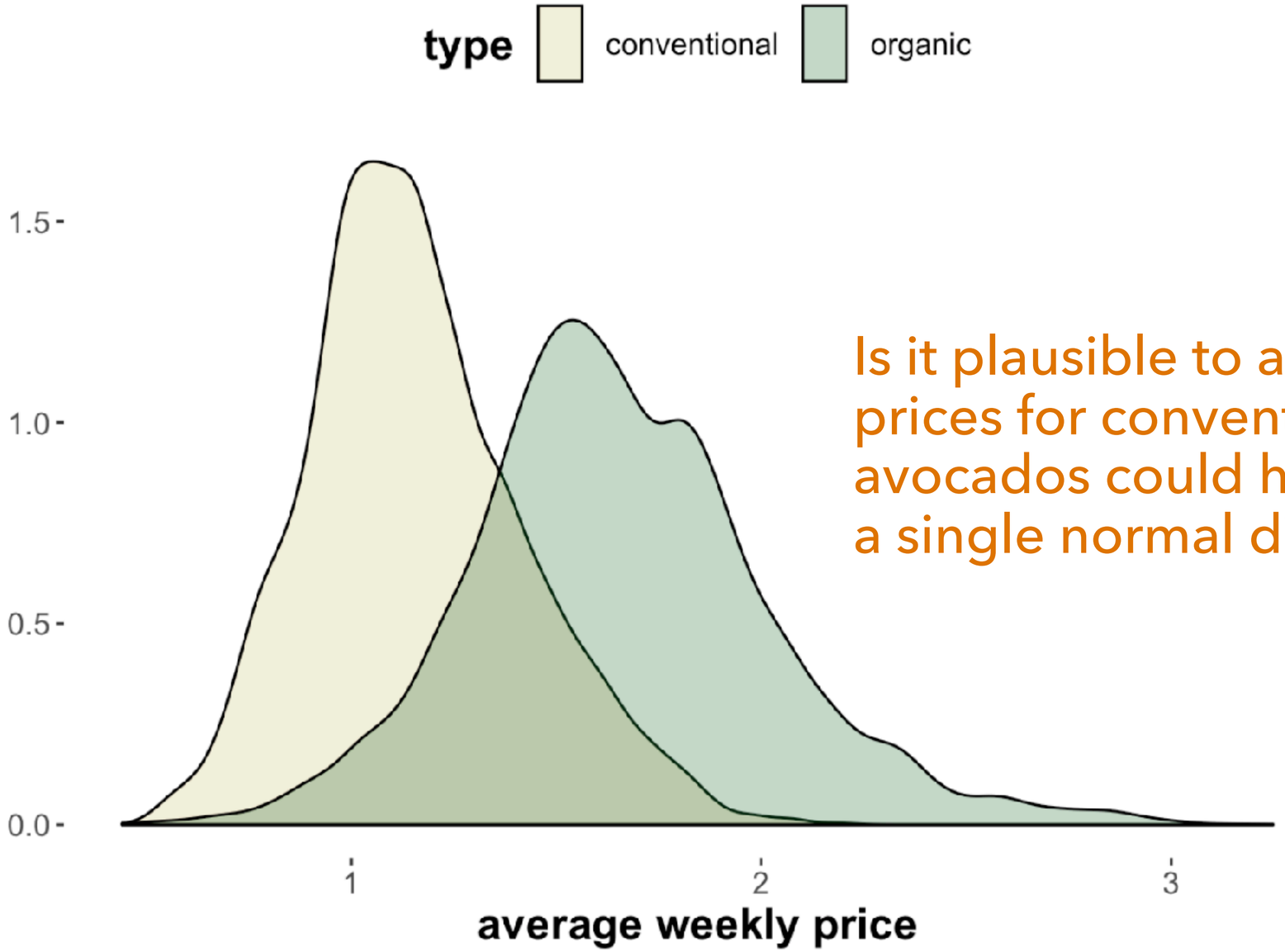
```
## [1] 0.007992
```



two-sample *t*-test

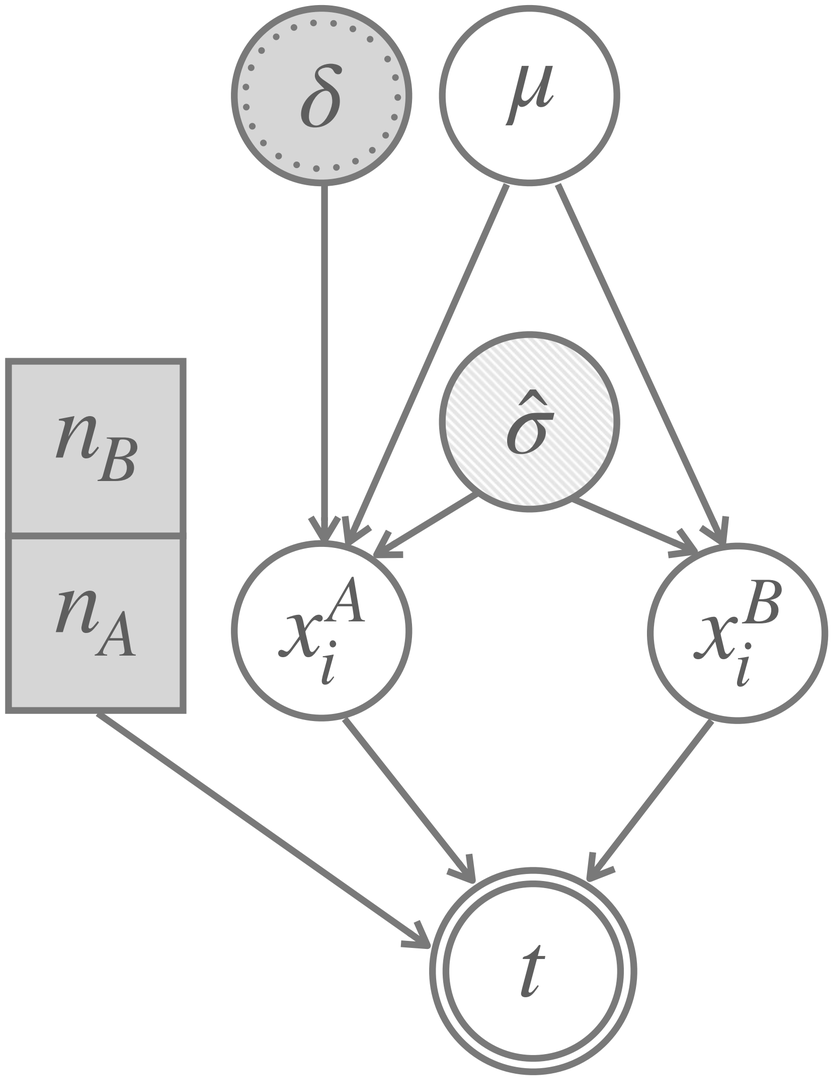
(unpaired data, equal variance & unequal sample size)

COMPARING TWO GROUPS OF METRIC MEASURES



Is it plausible to assume that the observed prices for conventional and organic avocados could have been generated by a single normal distribution?

FREQUENTIST T-TEST MODEL [TWO-SAMPLE, UNPAIRED, EQUAL VARIANCE, UNEQUAL SAMPLE SIZES]



$$x_i^A \sim \text{Normal}(\mu + \delta, \sigma)$$

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

$$\hat{\sigma} = \sqrt{\frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$t = \left((\bar{x}_A - \bar{x}_B) - \delta\right) \cdot \frac{1}{\hat{\sigma}}$$

Sampling distribution:

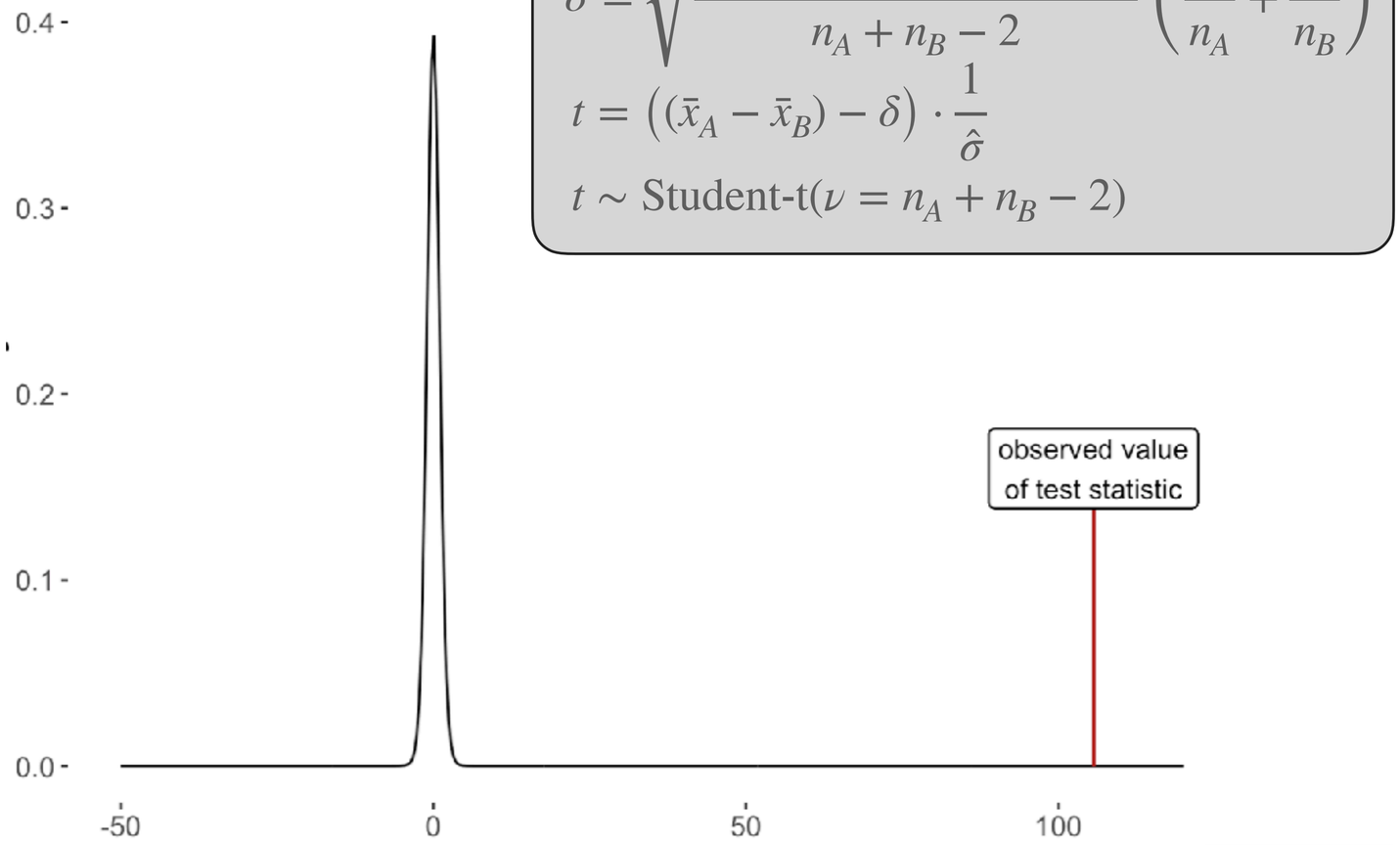
$t \sim \text{Student-t}(\nu = n_A + n_B - 2)$

TWO-SAMPLE T-TEST EXAMPLE

$$x_i^A \sim \text{Normal}(\mu + \delta, \sigma)$$
$$x_i^B \sim \text{Normal}(\mu, \sigma)$$
$$\hat{\sigma} = \sqrt{\frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$
$$t = \left((\bar{x}_A - \bar{x}_B) - \delta\right) \cdot \frac{1}{\hat{\sigma}}$$
$$t \sim \text{Student-t}(\nu = n_A + n_B - 2)$$

```r
# fix the null hypothesis: no difference between groups
delta_0 <- 0
# data (group A)
x_A <- avocado_data %>%
  filter(type == "organic") %>% pull(average_price)
# data (group B)
x_B <- avocado_data %>%
  filter(type == "conventional") %>% pull(average_price)
# sample mean for organic (group A)
mu_A <- mean(x_A)
# sample mean for conventional (group B)
mu_B <- mean(x_B)
# numbers of observations
n_A <- length(x_A)
n_B <- length(x_B)
# variance estimate
sigma_AB <- sqrt(
  ( ((n_A -1) * sd(x_A)^2 + (n_B -1) * sd(x_B)^2 ) /
      (n_A + n_B -2) ) * (1/n_A + 1/n_B)
)
t_observed <- (mu_A - mu_B - delta_0) / sigma_AB
t_observed
```

```
## [1] 105.5878
```




```r
p_value_t_test_avocado <- 1 - pt(q = t_observed, df = n_A + n_B - 1)
p_value_t_test_avocado
```

```
## [1] 0
```

TWO-SAMPLE T-TE

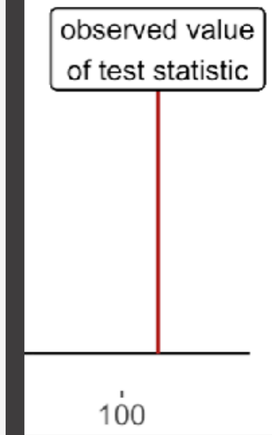
```r
# fix the null hypothesis: no di
delta_0 <- 0
# data (group A)
x_A <- avocado_data %>%
  filter(type == "organic") %>%
# data (group B)
x_B <- avocado_data %>%
  filter(type == "conventional")
# sample mean for organic (group
mu_A <- mean(x_A)
# sample mean for conventional (
mu_B <- mean(x_B)
# numbers of observations
n_A <- length(x_A)
n_B <- length(x_B)
# variance estimate
sigma_AB <- sqrt(
  ( ((n_A -1) * sd(x_A)^2 + (n_B
      (n_A + n_B -2) ) * (1/n_A
)
t_observed <- (mu_A - mu_B - del
t_observed
```

```
## [1] 105.5878
```

```r
t.test(
  x = x_A,            # first vector of data measurements
  y = x_B,            # sec vector of data measurements
  paired = FALSE,     # measurements are to be treated as unpaired
  var.equal = TRUE,   # we assum equal variance in both groups
  mu = 0              # NH is delta = 0 (name 'mu' is misleading!)
)
```

```
##
##  Two Sample t-test
##
## data:  x_A and x_B
## t = 105.59, df = 18247, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4867522 0.5051658
## sample estimates:
## mean of x mean of y
##  1.653999  1.158040
```

$$\frac{(n_B-1)\hat{\sigma}_B^2}{-2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)$$
$$\frac{1}{\hat{\sigma}}$$
$$+n_B-2)$$

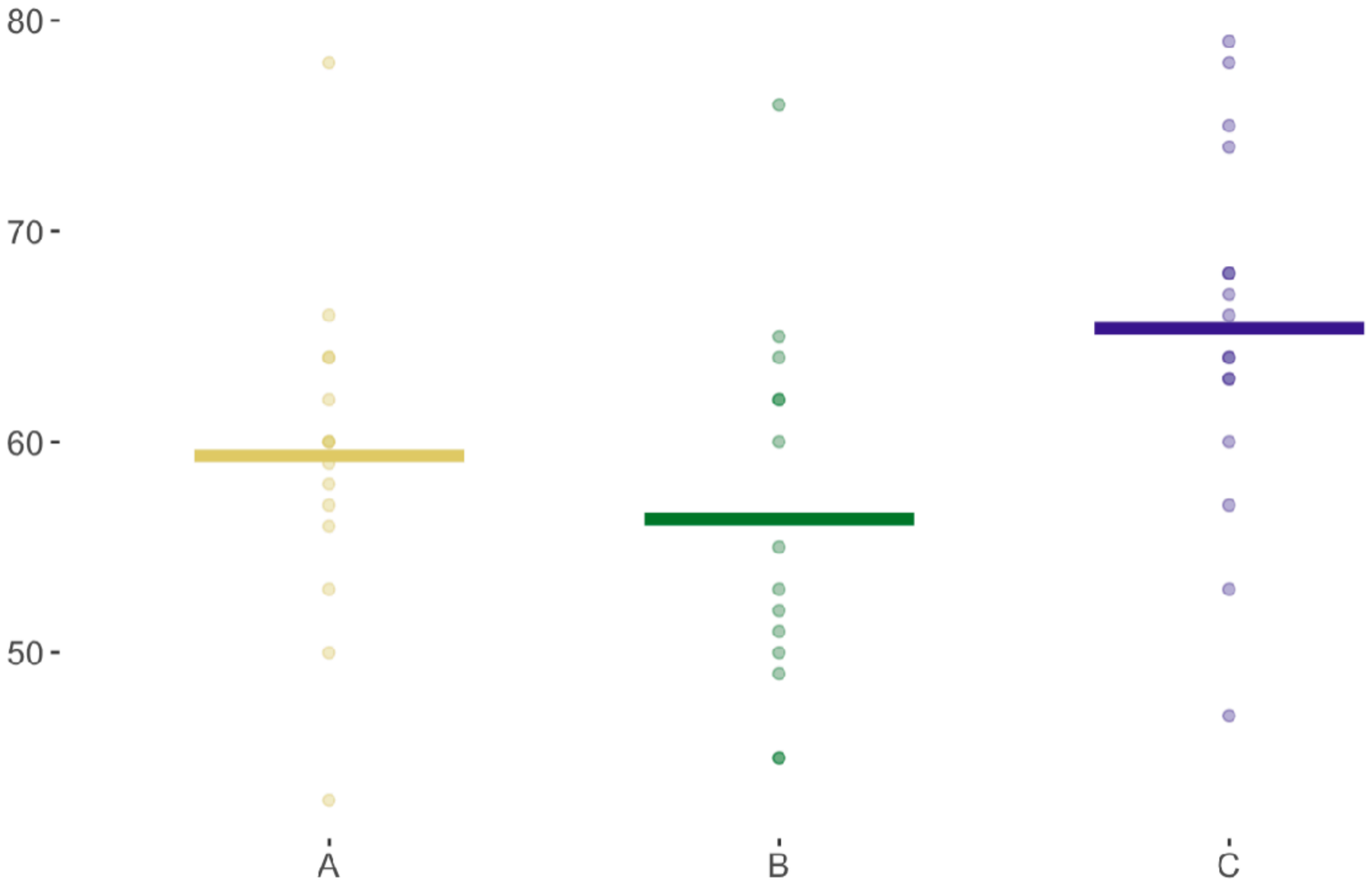



```r
ved, df = n_A + n_B - 1)
```

```
## [1] 0
```

one-way ANOVA

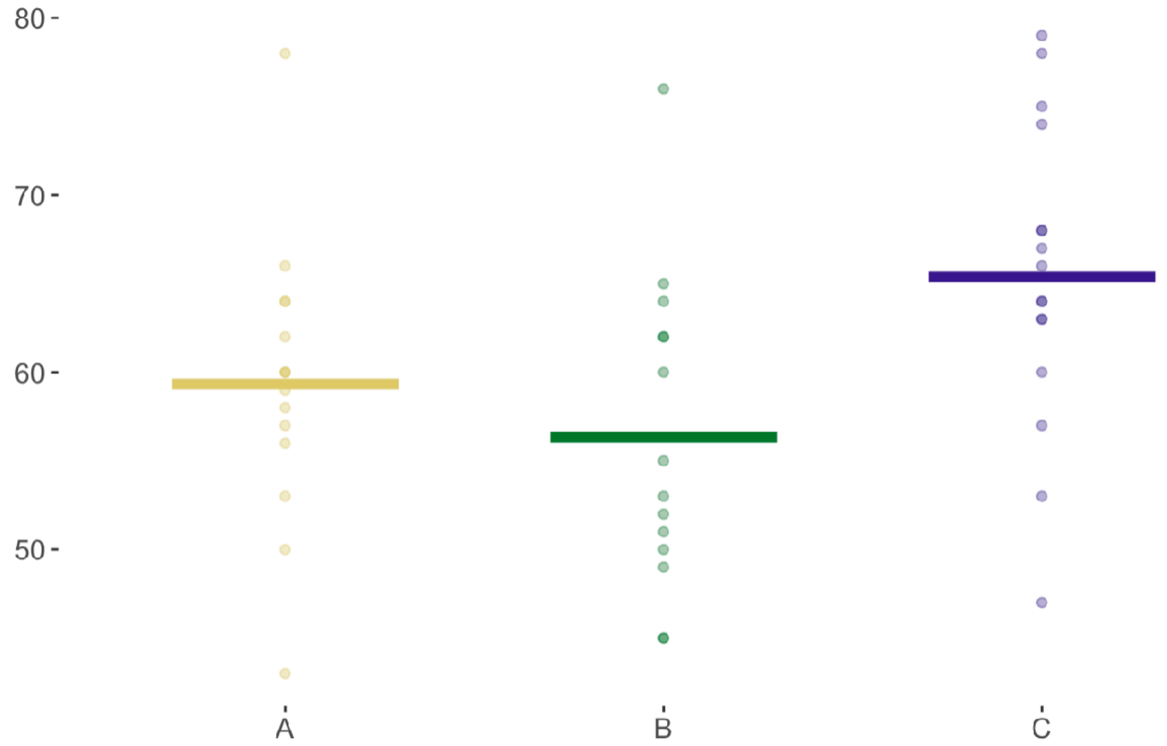
COMPARING K ≥ 2 GROUPS OF METRIC MEASURES



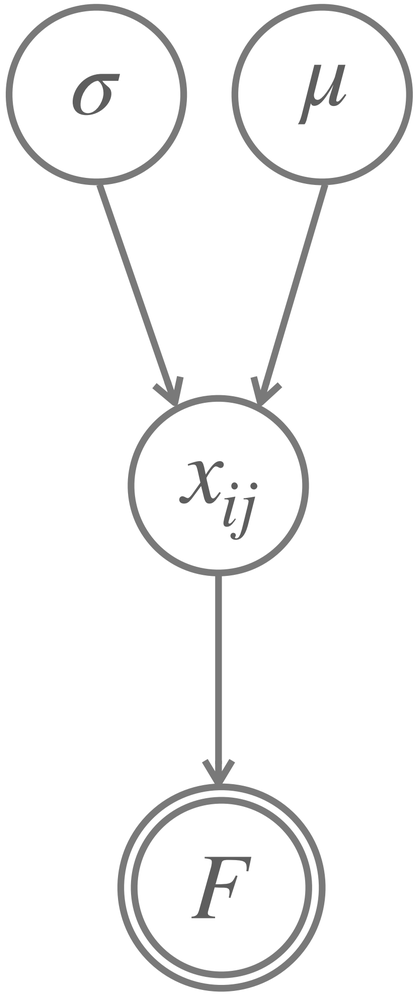
Is it plausible to assume that these measures stem from the same normal distribution?

WHY NOT t-TESTS?

- we could run t-tests between different groups
- chance of α error rises with each comparison
  - common corrections apply
- gets tedious with large k



FREQUENTIST MODEL FOR ANOVA [ONE-WAY]



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

$$F = \frac{\hat{\sigma}_{\text{between}}}{\hat{\sigma}_{\text{within}}}$$

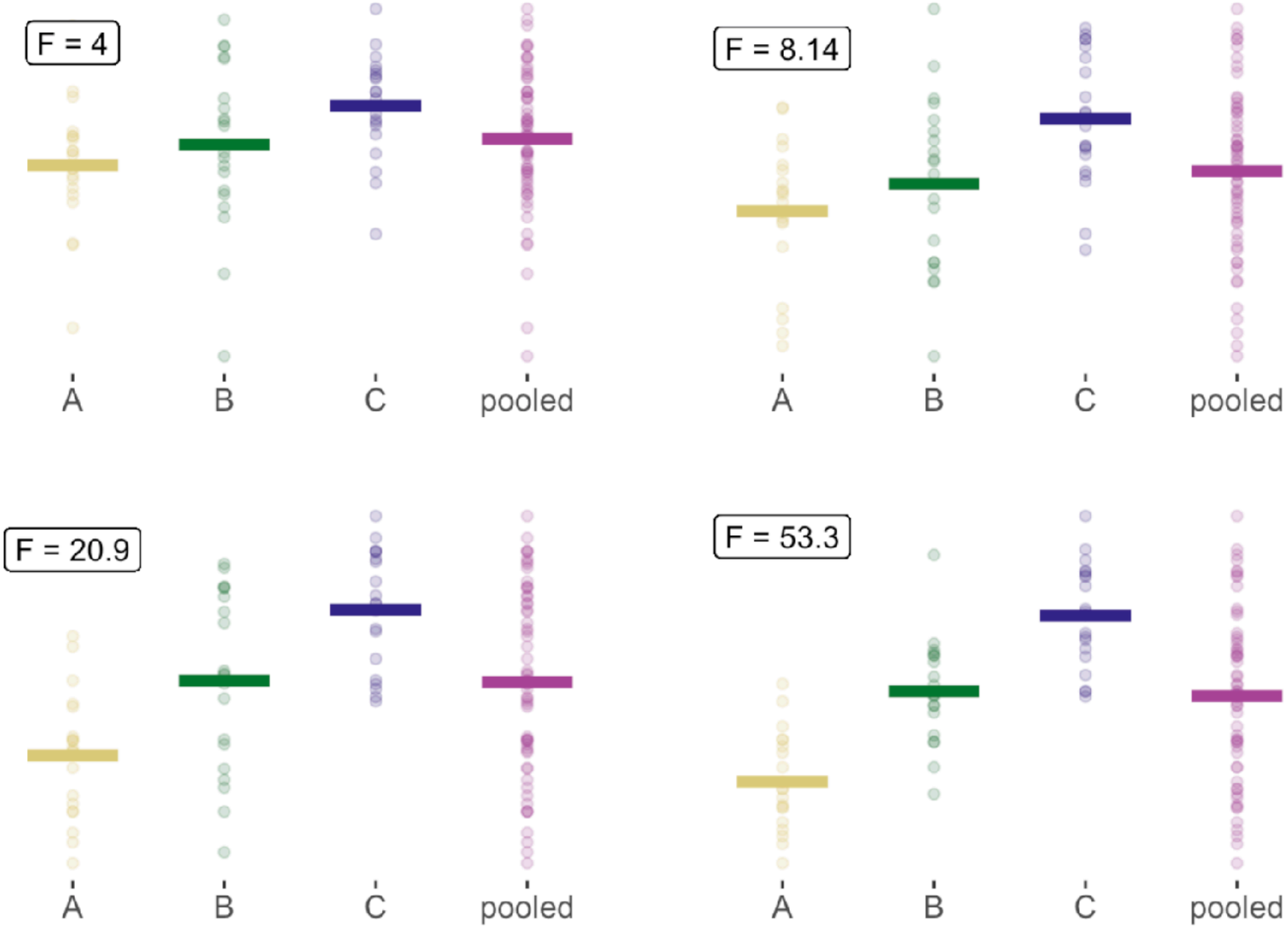
$$\hat{\sigma}_{\text{within}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{k} (n_i - 1)}$$

$$\hat{\sigma}_{\text{between}} = \frac{\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2}{k - 1}$$

Sampling distribution:

$$F \sim F\text{-distribution}\left(k - 1, \sum_{i=1}^{k} (n_i - 1)\right)$$

F-STATISTIC EXAMPLES



F-DISTRIBUTION

- two χ^2-distributed random variables:

$$x \sim \chi^2\text{-distribution}(m)$$

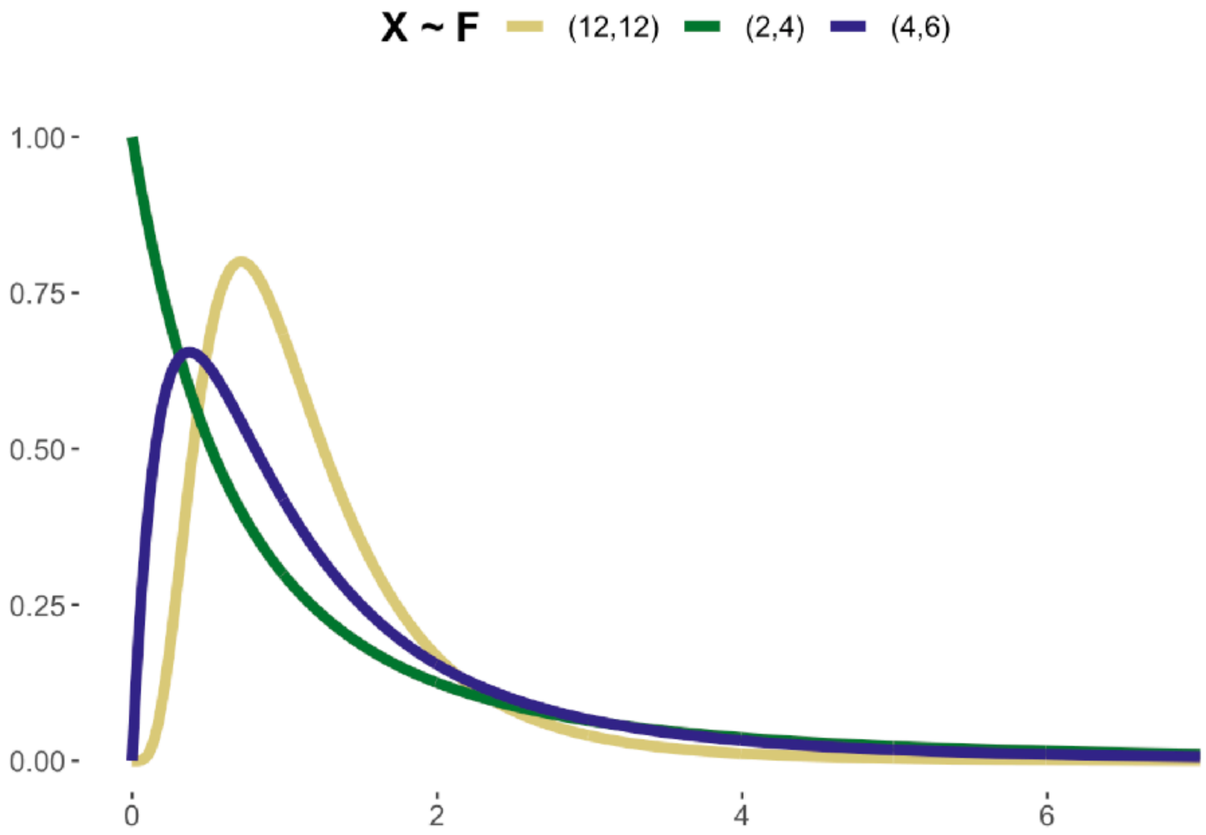
$$y \sim \chi^2\text{-distribution}(n)$$

- derived RV:

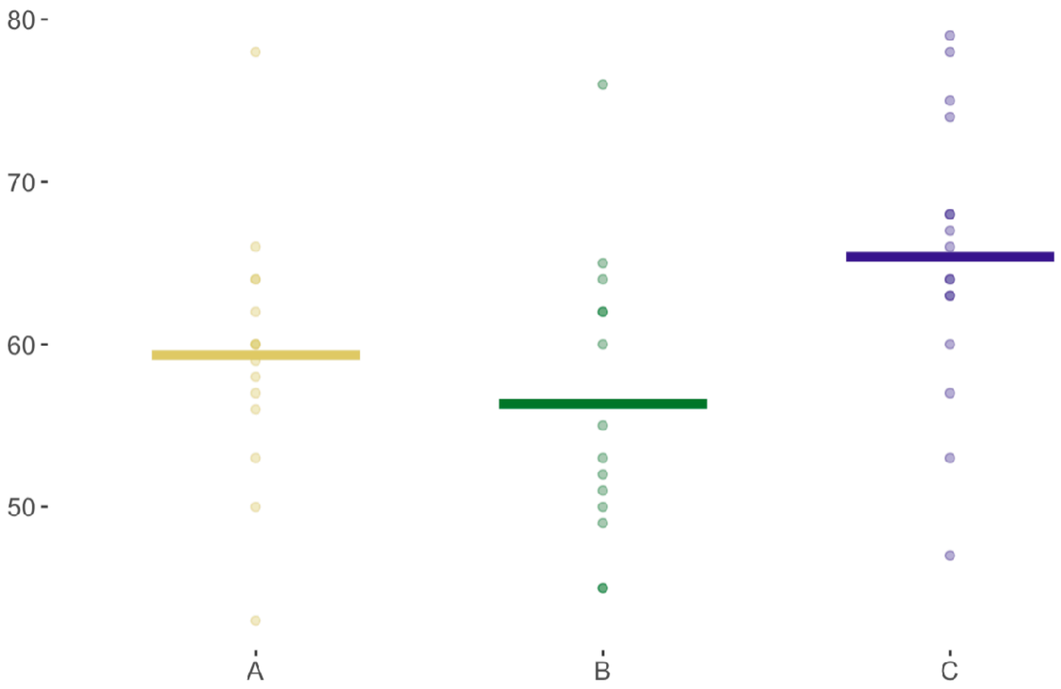
$$Z = \frac{X/m}{Y/n}$$

- it follows (by construction) that:

$$z \sim F\text{-distribution}(m, n)$$



EXAMPLE




```r
aov(formula = value ~ condition, anova_data) %>% summary()
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## condition    2  640.8   320.4   4.485 0.0172 *
## Residuals   42 3000.3    71.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> Based on a one-way ANOVA, we find evidence against the assumption of equal means across all groups ($F(2, 42) \approx 4.485$, $p \approx 0.0172$.)

varieties of frequentist testing

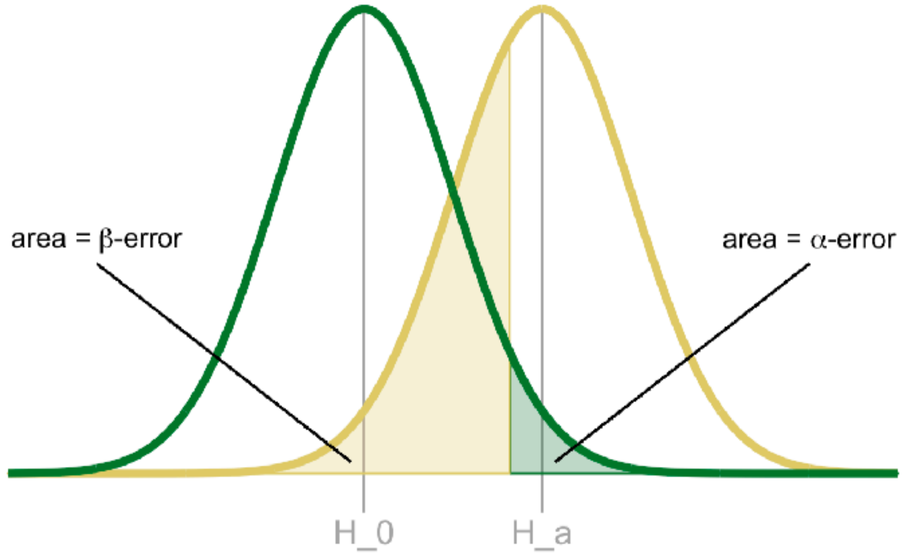
THREE VARIETIES OF FREQUENTIST TESTING

	FISHER	NEYMAN/PEARSON	HYBRID NHST*
explicit & serious alternative H_a	X	✓	X
when to set-up statistical model	after data collection	before data collection	after data collection
goal of statistical analysis	quantify evidence against H_0	decide action: adopt H_0 or H_a	decide action: adopt H_0 or $\neg H_0$
power calculation	X	✓	X

* this is a worst-case portrait of modern NHST ; this is *not* how it *should* be done

NEYMAN/PEARSON APPROACH [INFORMAL GIST]

- procedure in N/P approach:
  - fix H_0 and H_a (based on prior research)
  - determine desired α- and β-error level
  - calculate sample size *N* necessary for β given α
  - run the experiment
  - determine significance based on α-level
  - make a dichotomous decision:
    - accept H_a if test is significant
    - accept H_0 otherwise



LONG-TERM ERROR CONTROL IN NEYMAN/PEARSON APPROACH

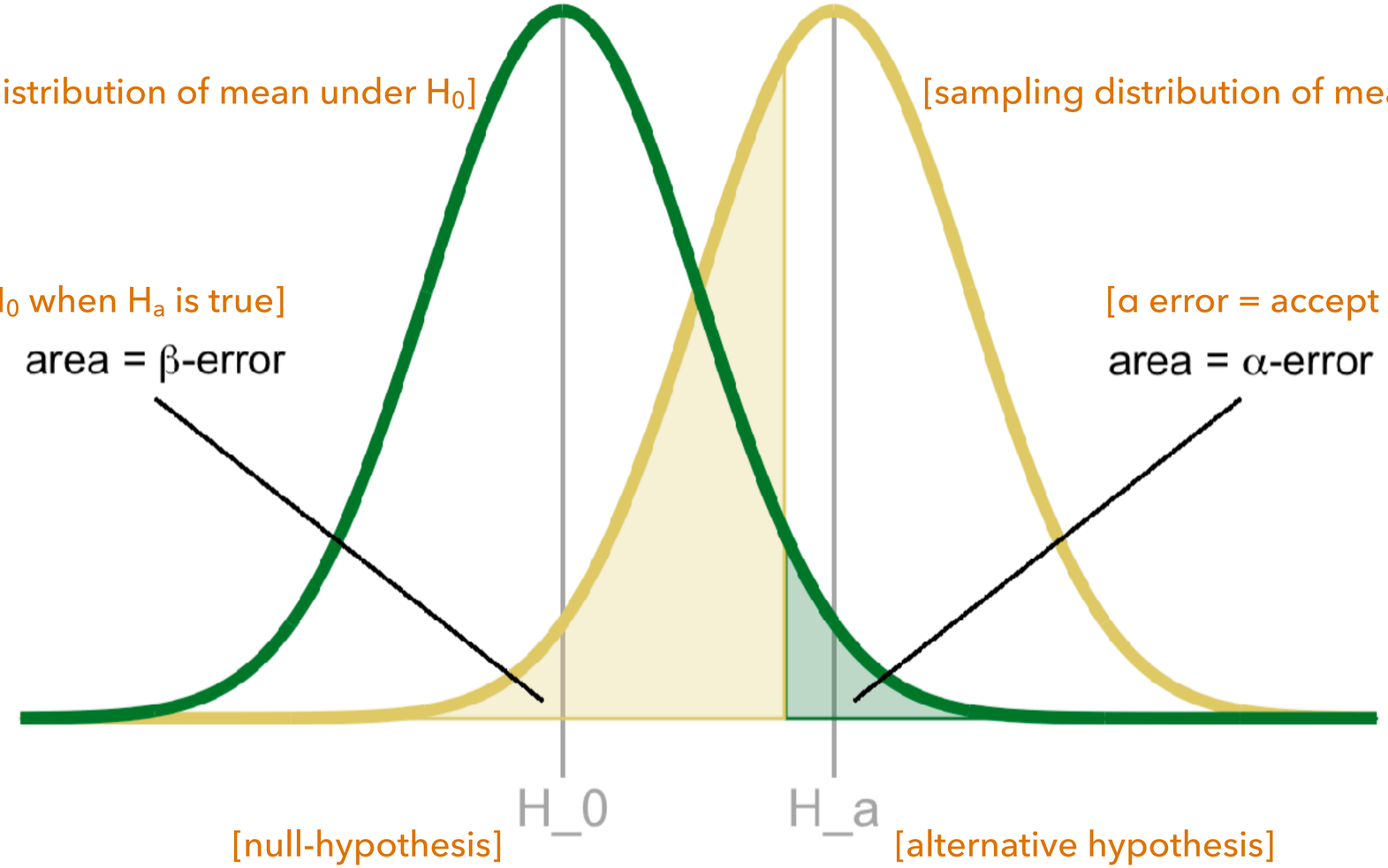
[more data = tighter curves!! = lower β]

[sampling distribution of mean under H_0]

[sampling distribution of mean under H_a]

[β error = accept H_0 when H_a is true]

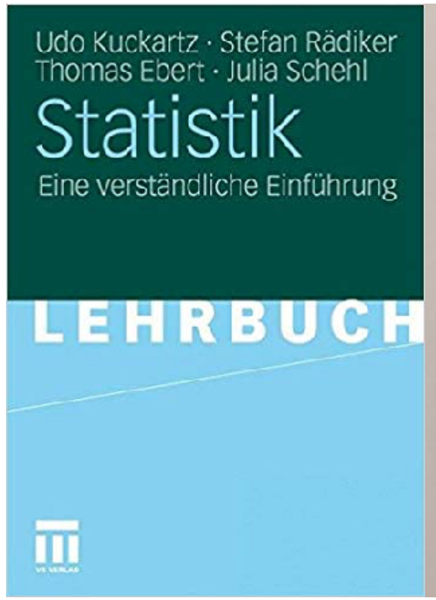
[α error = accept H_a when H_0 is true]

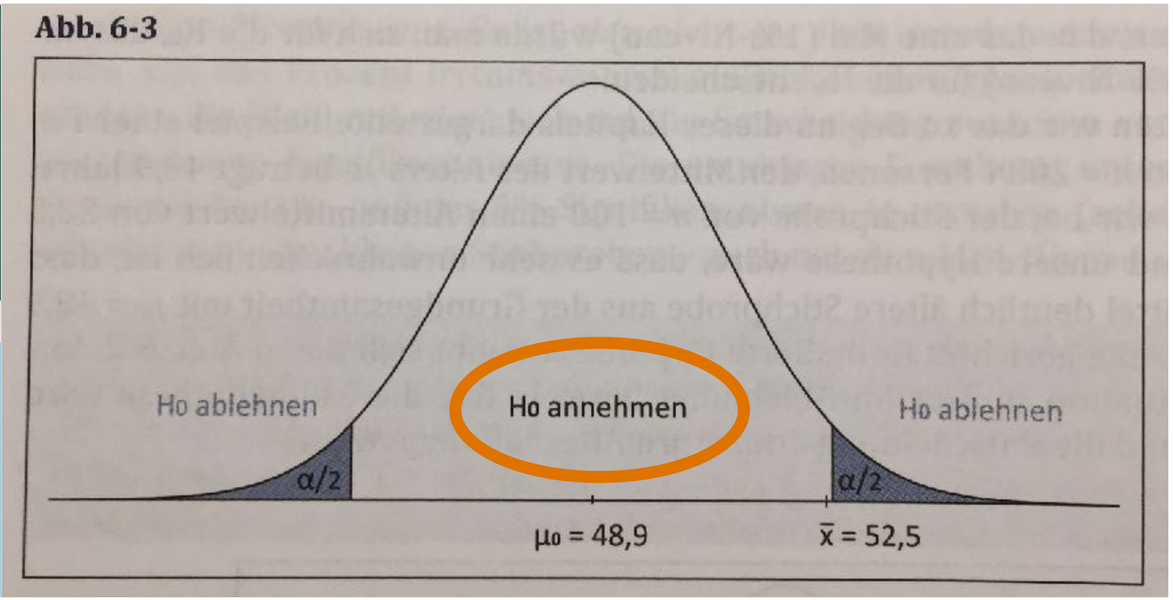


[null-hypothesis]

[alternative hypothesis]

EXAMPLES FROM TEXTBOOKS





neither textbook talks about fixing Ha and/or calculating power of a test

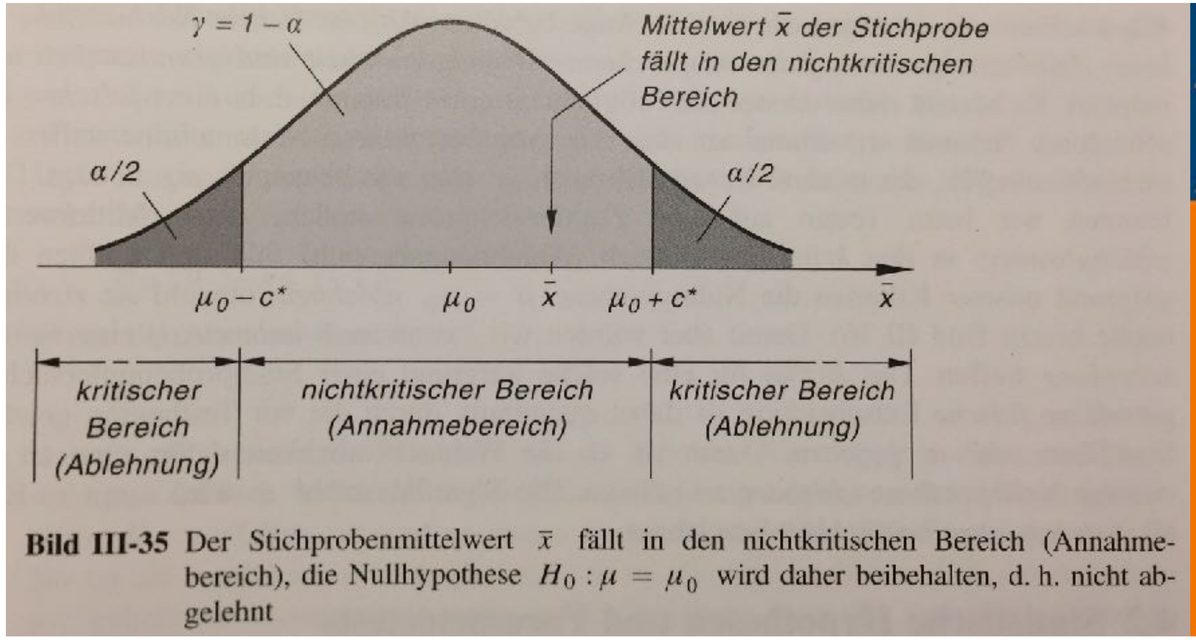
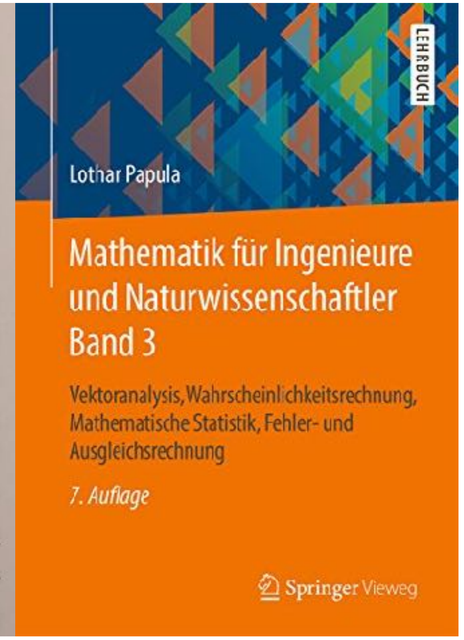


Bild III-35 Der Stichprobenmittelwert $\bar{x}$ fällt in den nichtkritischen Bereich (Annahmebereich), die Nullhypothese $H_0 : \mu = \mu_0$ wird daher beibehalten, d. h. nicht abgelehnt



THREE VARIETIES OF FREQUENTIST TESTING

	FISHER	NEYMAN/PEARSON	HYBRID NHST*
explicit & serious alternative H_a	X	✓	X
when to set-up statistical model	after data collection	before data collection	after data collection
goal of statistical analysis	quantify evidence against H_0	decide action: adopt H_0 or H_a	decide action: adopt H_0 or $\neg H_0$
power calculation	X	✓	X

* this is a worst-case portrait of modern NHST ; this is *not* how it *should* be done