Research Design, Fall 2021

05: statistical tests continued; inference and interpretation

Scott Spencer | Columbia University

Research Review 1

Homework 2
(research in retail supply chains)

Homework 3
(user behaviors in digital apps)

Homework 4
(customer service for airlines)

Final Exam

YOU
ARE
HERE

SEPT 21

OCT 5

OCT 26

NOV 9

DEC 1

OCT 19

NOV 2

NOV 30

DEC 7

Homework 1
(academic integrity)

Group project
(checkpoint)

Research review 2

Homework 5
(work and life balance)

Group project
(final submission,
presentation)

visual recap on *probability density functions* (PDF) and *cumulative distribution functions* (CDF)

```
pdf <- ggplot() +
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +
  scale_x_continuous(breaks = seq(-2, 2)) +
  scale_y_continuous(breaks = dnorm(seq(-2, 2)), labels = scales::comma) +
  geom_vline(xintercept = 0, color = "pink") +
  stat_function(fun  = dnorm,
                args = list(mean = 0, sd = 1),
                geom = "density",
                xlim = c(-4,4)) +
  labs(x = TeX("$\\sigma$"), y = TeX("$\\Phi(x)$"))
```
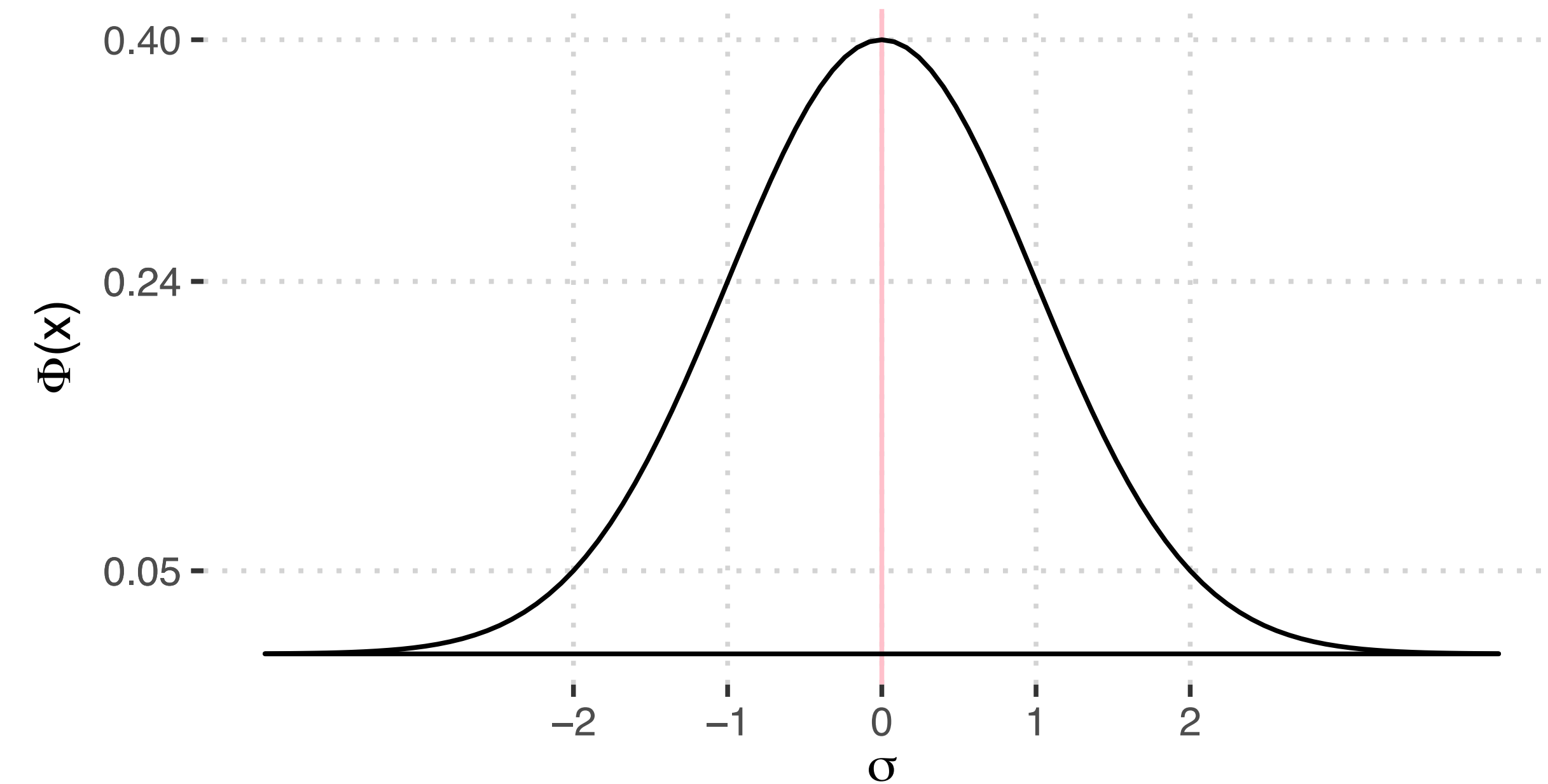


```
cdf <- ggplot() +
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +
  scale_x_continuous(breaks = seq(-2, 2)) +
  scale_y_continuous(breaks = pnorm(seq(-2, 2)),
                     labels = scales::label_percent(accuracy = 1)) +
  geom_vline(xintercept = 0, color = "pink") +
  stat_function(fun  = pnorm,
                args = list(mean = 0, sd = 1),
                geom = "density",
                xlim = c(-4,4)) +
  labs(x = TeX("$\\sigma$"), y = TeX("$F_\\Phi(x)$"))
```
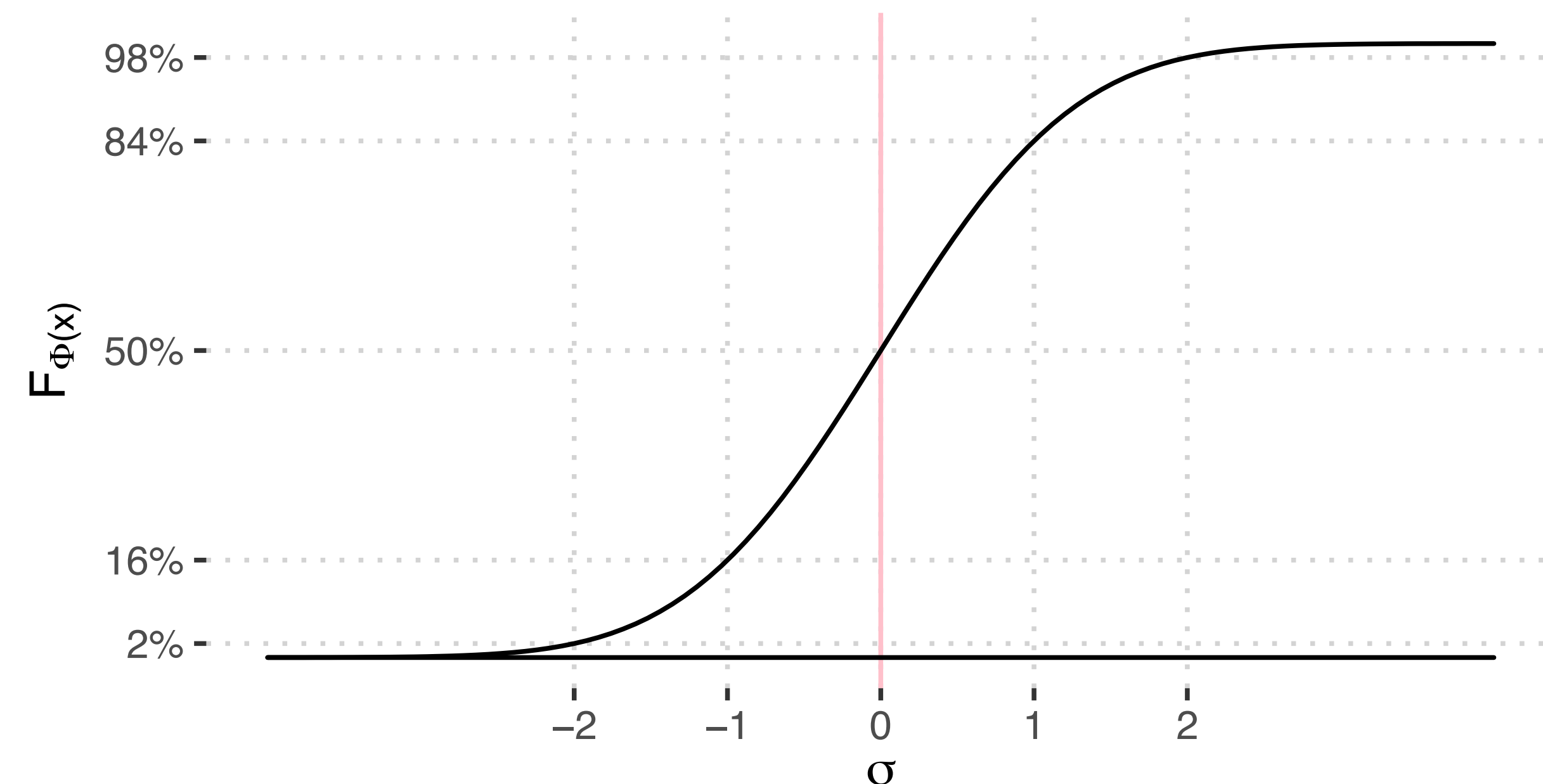
R's probability functions, **p**robability density (PDF), cumulative distribution (CDF), **q**uantile, **r**andom generation

d<probability function name>   probability **d**ensity function

p<probability function name>   cumulative distribution function (of **p**robability)

q<probability function name>   **q**uantile function

r<probability function name>   **r**andom generation function

↓ examples

---
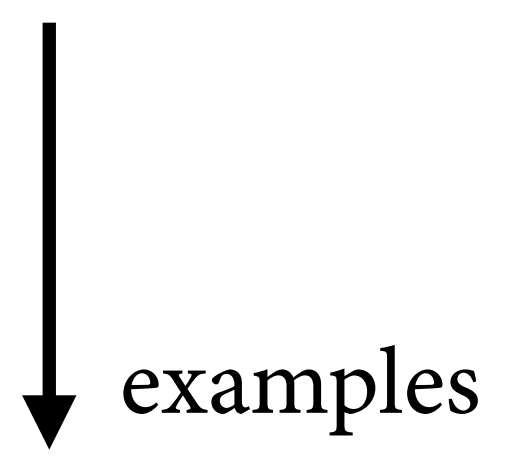
normal	*negative binomial*
student t	*gamma*
bernoulli	*cauchy*
binomial	*100s more*
poisson	*…*

'a zoo of pre-constructed golems known as "tests" '

— Richard McElreath

a zoo of tests, a decision tree for selecting one (and *not all* named or unnamed tests are listed below)

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?
— Discrete, categorical → Any counts < 5?
— Continuous

Any counts < 5?
— No → Chi-square tests, one and two sample
— Yes → Fisher's exact test

Type of question?
— Relationships
— Differences

Do you have dependent & independent variables?
— Yes → Regression analysis
— No → Correlation analysis

Correlation analysis
— Parametric → Pearson's r
— Nonparametric → Spearman's rank correlation

Differences between what?
— Means → One-sample t-test
— Variances → Fmax test or Bartlett's test
— Multiple means Single variable → How many groups?

How many groups?
— More than two → Parametric assumptions satisfied?
— Two → Parametric assumptions satisfied?

Parametric assumptions satisfied? (More than two)
— Yes → One-way ANOVA
— No → Transform data?
— Transform data? No → Kruskall-Wallis test
— OK → One-way ANOVA

One-way ANOVA → If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.
Kruskall-Wallis test → If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

Parametric assumptions satisfied? (Two)
— No → Transform data?
— No → Mann-Whitney U or Wilcoxon test
Transform data?
— OK → Student's t-test
— Yes → Student's t-test
— No → Mann-Whitney U or Wilcoxon test

This zoo of tests does share common theories based on probability.

— Casella & Berger 1990; Lehmann & Casella 1998; Lehmann & Romano 2005

a zoo of tests, **a decision tree for selecting one,** *keeping in mind that classical tests can be* **inflexible** *and* **fragile**

FIGURE 1.1. Example decision tree, or flowchart, for selecting an appropriate statistical procedure. Beginning at the top, the user answers a series of questions about measurement and intent, arriving eventually at the name of a procedure. Many such decision trees are possible.

Sometimes their unyielding logic reveals implications previously hidden to their designers. These implications can be priceless discoveries. Or they may produce silly and dangerous behavior. Rather than idealized angels of reason, scientific models are powerful clay robots without intent of their own, bumbling along according to the myopic instructions they embody. Like with Rabbi Judah's golem, the golems of science are wisely regarded with both awe and apprehension. We absolutely have to use them, but doing so always entails some risk.

There are many kinds of statistical models. Whenever someone deploys even a simple statistical procedure, like a classical *t*-test, she is deploying a small golem that will obediently carry out an exact calculation, performing it the same way (nearly[2]) every time, without complaint. Nearly every branch of science relies upon the senses of statistical golems. In many cases, it is no longer possible to even measure phenomena of interest, without making use of a model. To measure the strength of natural selection or the speed of a neutrino or the number of species in the Amazon, we must use models. The golem is a prosthesis, doing the measuring for us, performing impressive calculations, finding patterns where none are obvious.

However, there is no wisdom in the golem. It doesn't discern when the context is inappropriate for its answers. It just knows its own procedure, nothing else. It just does as it's told.

And so it remains a triumph of statistical science that there are now so many diverse golems, each useful in a particular context. Viewed this way, statistics is neither mathematics nor a science, but rather a branch of engineering. And like engineering, a common set of design principles and constraints produces a great diversity of specialized applications.

This diversity of applications helps to explain why introductory statistics courses are so often confusing to the initiates. Instead of a single method for building, refining, and critiquing statistical models, students are offered a zoo of pre-constructed golems known as "tests." Each test has a particular purpose. Decision trees, like the one in FIGURE 1.1, are common. By answering a series of sequential questions, users choose the "correct" procedure for their research circumstances.

Unfortunately, while experienced statisticians grasp the unity of these procedures, students and researchers rarely do. Advanced courses in statistics do emphasize engineering principles, but most scientists never get that far. Teaching statistics this way is somewhat like teaching engineering backwards, starting with bridge building and ending with basic physics. So students and many scientists tend to use charts like FIGURE 1.1 without much thought to their underlying structure, without much awareness of the models that each procedure embodies, and without any framework to help them make the inevitable compromises required by real research. It's not their fault.

For some, the toolbox of pre-manufactured golems is all they will ever need. Provided they stay within well-tested contexts, using only a few different procedures in appropriate tasks, a lot of good science can be completed. This is similar to how plumbers can do a lot of useful work without knowing much about fluid dynamics. Serious trouble begins when scholars move on to conducting innovative research, pushing the boundaries of their specialties. It's as if we got our hydraulic engineers by promoting plumbers.

Why aren't the tests enough for research? The classical procedures of introductory statistics tend to be inflexible and fragile. By inflexible, I mean that they have very limited ways to adapt to unique research contexts. By fragile, I mean that they fail in unpredictable ways when applied to new contexts. This matters, because at the boundaries of most sciences, it is hardly ever clear which procedure is appropriate. None of the traditional golems has been evaluated in novel research settings, and so it can be hard to choose one and then to understand how it behaves. A good example is *Fisher's exact test*, which applies (exactly) to an extremely narrow empirical context, but is regularly used whenever cell counts are small. I have personally read hundreds of uses of Fisher's exact test in scientific journals, but aside from Fisher's original use of it, I have never seen it used appropriately. Even a procedure like ordinary linear regression, which is quite flexible in many ways, being able to encode a large diversity of interesting hypotheses, is sometimes fragile. For example, if there is substantial measurement error on prediction variables, then the procedure can fail in spectacular ways. But more importantly, it is nearly always possible to do better than ordinary linear regression, largely because of a phenomenon known as OVERFITTING (Chapter 7).

The point isn't that statistical tools are specialized. Of course they are. The point is that classical tools are not diverse enough to handle many common research questions. Every active area of science contends with unique difficulties of measurement and interpretation, converses with idiosyncratic theories in a dialect barely understood by other scientists from other tribes. Statistical experts outside the discipline can help, but they are limited by lack of fluency in the empirical and theoretical concerns of the discipline.

Furthermore, no statistical tool does anything on its own to address the basic problem of inferring causes from evidence. Statistical golems do not understand cause and effect.

Recall Student's t test — comparing one sample means to μ where we assume a population normal distribution with unknown standard deviation σ.

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Discrete, categorical

Any counts < 5?

No — Chi-square tests, one and two sample

Yes — Fisher's exact test

Continuous

Type of question?

Relationships

Differences

Do you have dependent & independent variables?

Differences between what?

Means — One-sample t-test

Variances — Fmax test or Bartlett's test

Yes — Regression analysis

No — Correlation analysis

Parametric — Pearson's r

Nonparametric — Spearman's rank correlation

Multiple means Single variable

How many groups?

More than two — Parametric assumptions satisfied?

Yes — One-way ANOVA

No — Transform data?

OK — One-way ANOVA

No — Kruskall-Wallis test

Two — Parametric assumptions satisfied?

Transform data? — No

OK — Student's t-test

No — Yes — Mann-Whitney U or Wilcoxon test

No — If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

$$H_0 : \bar{x} = \mu, \, H_A : \bar{x} < \mu$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad \nu = n - 1, \quad p = F_T(t, \nu)$$

```
t.test(x, mu, alternative = "less", conf.level = 0.95)
```

Recall Student's t test — comparing one sample means to μ where we assume a population normal distribution with unknown standard deviation σ.

```r
set.seed(1)

mu1 <- 4
pop1 <- rnorm(1e5, mean = mu1, sd = 2)

# sample from the population
n1 <- 50
x1 <- sample(x = pop1, size = n1, replace = FALSE)

# setup the test to calculate manually
xbar1 <- mean(x1)
s1 <- sd(x1)
nu <- n1 - 1

# pretend we know population mu but not sigma
t <- (xbar1 - mu1) / (s1 / sqrt(n1))

# manually get p-value
p <- pstudent_t(q = t, df = nu)
```

$$H_0 : \bar{x} = \mu, H_A : \bar{x} < \mu$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad \nu = n - 1, \quad p = F_T(t, \nu)$$

Student's t test — comparing two sample means where we can assume an underlying normal distribution.

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Discrete, categorical

Any counts < 5?

No

Yes

Continuous

Chi-square tests, one and two sample

Fisher's exact test

Type of question?

Relationships

Differences

Do you have dependent & independent variables?

Differences between what?

Means

One-sample t-test

Variances

Fmax test or Bartlett's test

Yes

No

Multiple means Single variable

Regression analysis

Correlation analysis

How many groups?

Parametric assumptions satisfied?

Parametric

Nonparametric

More than two

Yes

No

No

Pearson's r

Spearman's rank correlation

Two

One-way ANOVA

OK

Transform data?

No

Parametric assumptions satisfied?

Kruskall-Wallis test

Transform data?

No

No

OK

Yes

No

Student's t-test

Mann-Whitney U or Wilcoxon test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

$$H_0 : \mu_1 = \mu_2, H_A : \mu_1 < \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}, \quad p = F_T(t, \nu)$$

```
t.test(x, y, alternative = "less", var.equal = FALSE, conf.level = 0.95)
```

Student's t test — comparing two sample means where we can assume an underlying normal distribution.
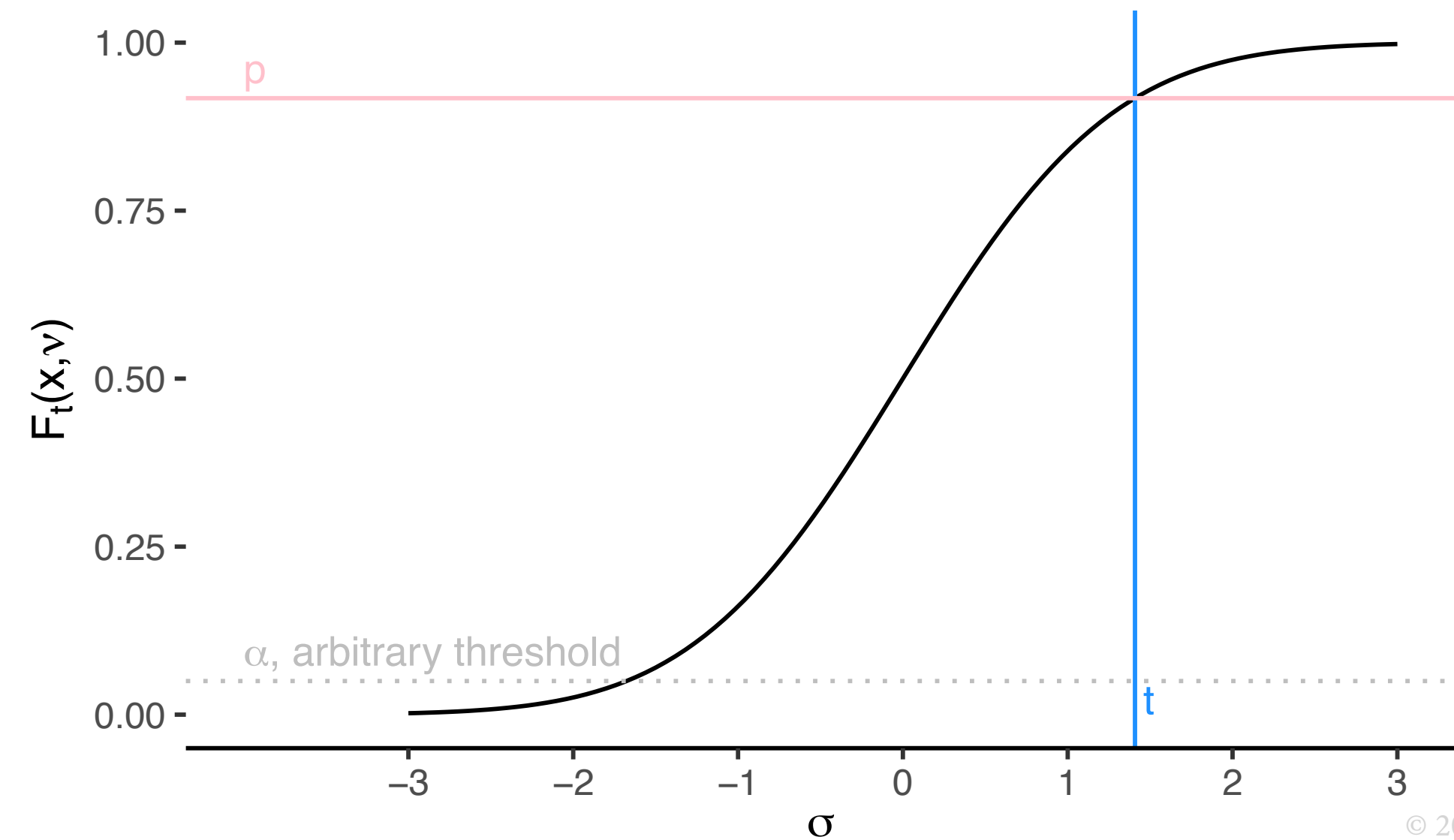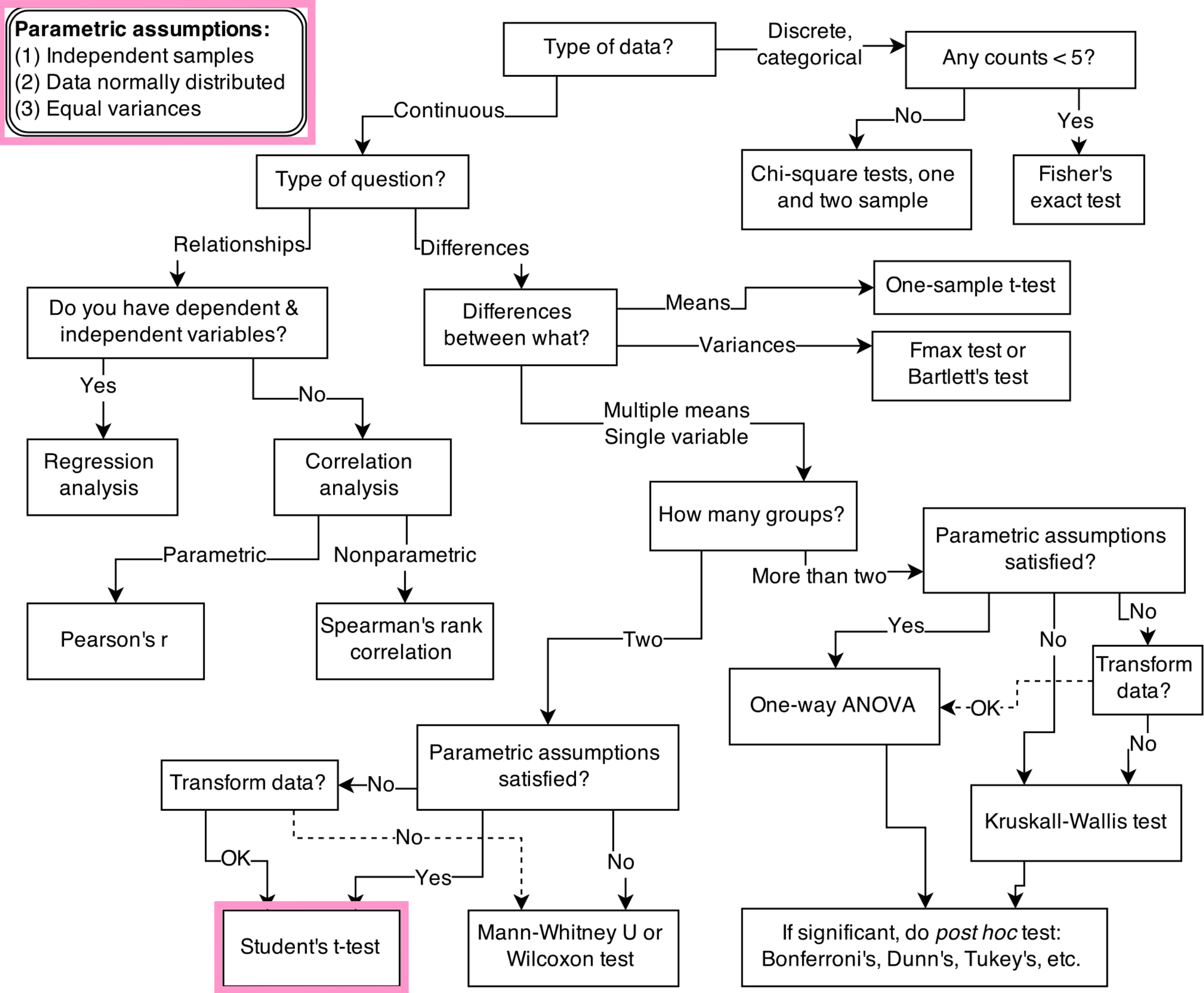
```r
# second population
mu2 <- 2
pop2 <- rnorm(1e5, mean = mu2, sd = 3)

# sample from the second population
n2 <- 50
x2 <- sample(x = pop2, size = n2, replace = FALSE)
```

```r
xbar2 <- mean(x2)
s2 <- sd(x2)

t <- ( xbar1 - xbar2 ) / sqrt( s1 ^ 2 / n1 + s2 ^ 2 / n2)

nu <- ( s1 ^ 2 / n1 + s2 ^ 2 / n2 ) ^ 2 /
      ( ( s1 ^ 2 / n1 ) ^ 2 / (n1 - 1) + ( s2 ^ 2 / n2 ) ^ 2 / (n2 - 1) )

p <- pstudent_t(q = t, df = nu)
```

$$H_0 : \mu_1 = \mu_2, H_A : \mu_1 < \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}, \quad p = F_T(t, \nu)$$


```r
a <- t.test(x1, x2, alternative = "less", var.equal = FALSE, conf.level = 0.95)
p == a$p.value
```

Proportions are distributed as binomial, which tends to approximate a normal with sufficient n

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data? — Discrete, categorical → Any counts < 5?

Any counts < 5? — No → Chi-square tests, one and two sample

Any counts < 5? — Yes → Fisher's exact test

Continuous → Type of question?

Type of question? — Relationships → Do you have dependent & independent variables?
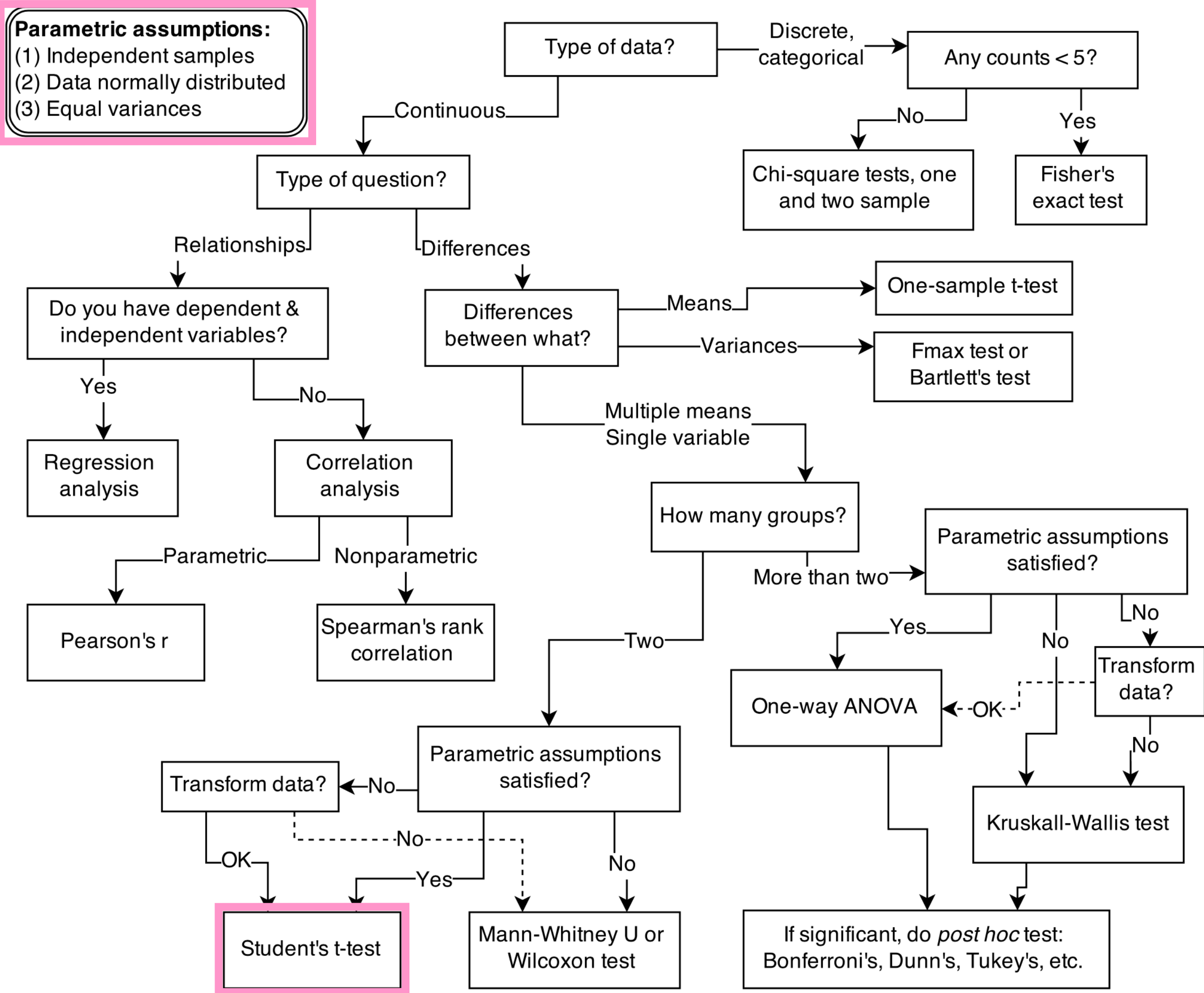
Type of question? — Differences → Differences between what?

Do you have dependent & independent variables? — Yes → Regression analysis

Do you have dependent & independent variables? — No → Correlation analysis

Differences between what? — Means → One-sample t-test

Differences between what? — Variances → Fmax test or Bartlett's test

Differences between what? — Multiple means Single variable → How many groups?

Correlation analysis — Parametric → Pearson's r

Correlation analysis — Nonparametric → Spearman's rank correlation

How many groups? — More than two → Parametric assumptions satisfied?

How many groups? — Two → Parametric assumptions satisfied?

Parametric assumptions satisfied? — Yes → One-way ANOVA

Parametric assumptions satisfied? — No → Transform data?

Transform data? — -OK-→ One-way ANOVA

Transform data? — No → Kruskall-Wallis test

Parametric assumptions satisfied? — No → Transform data?

Transform data? — No → OK → Student's t-test

Transform data? — -No- / Yes → Mann-Whitney U or Wilcoxon test

Parametric assumptions satisfied? — No → Mann-Whitney U or Wilcoxon test

One-way ANOVA → If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

Kruskall-Wallis test → If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

comparing observed proportion to probability

$$H_0 : \pi = \pi_0, H_A : \pi \neq \pi_0$$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}, p = F_\Phi(z)$$

comparing two observed proportions

$$H_0 : \pi_1 - \pi_2 = \delta, H_A : \pi_1 - \pi_2 \neq \delta$$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \delta}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ where } p_0 = \frac{x_1 + x_2}{n_1 + n_2}, p = F_\Phi(z)$$

Proportions are distributed as binomial, which tends to approximate a normal with sufficient n

```
# population proportion
pi <- 0.4

# population of proportions
pop1 <- rbinom(n = 1e5, size = 1, prob = pi)

# observed proportion (sample or experiment)
p1 <- sample(pop1, size = n1)
```

```
phat1 <- mean(p1)

# calculate test statistic
z <- ( phat1 - pi ) / sqrt( pi * (1 - pi) / n1 )

# get location on cdf of standard normal distribution
p <- pnorm(q = z)
```

comparing observed proportion to probability
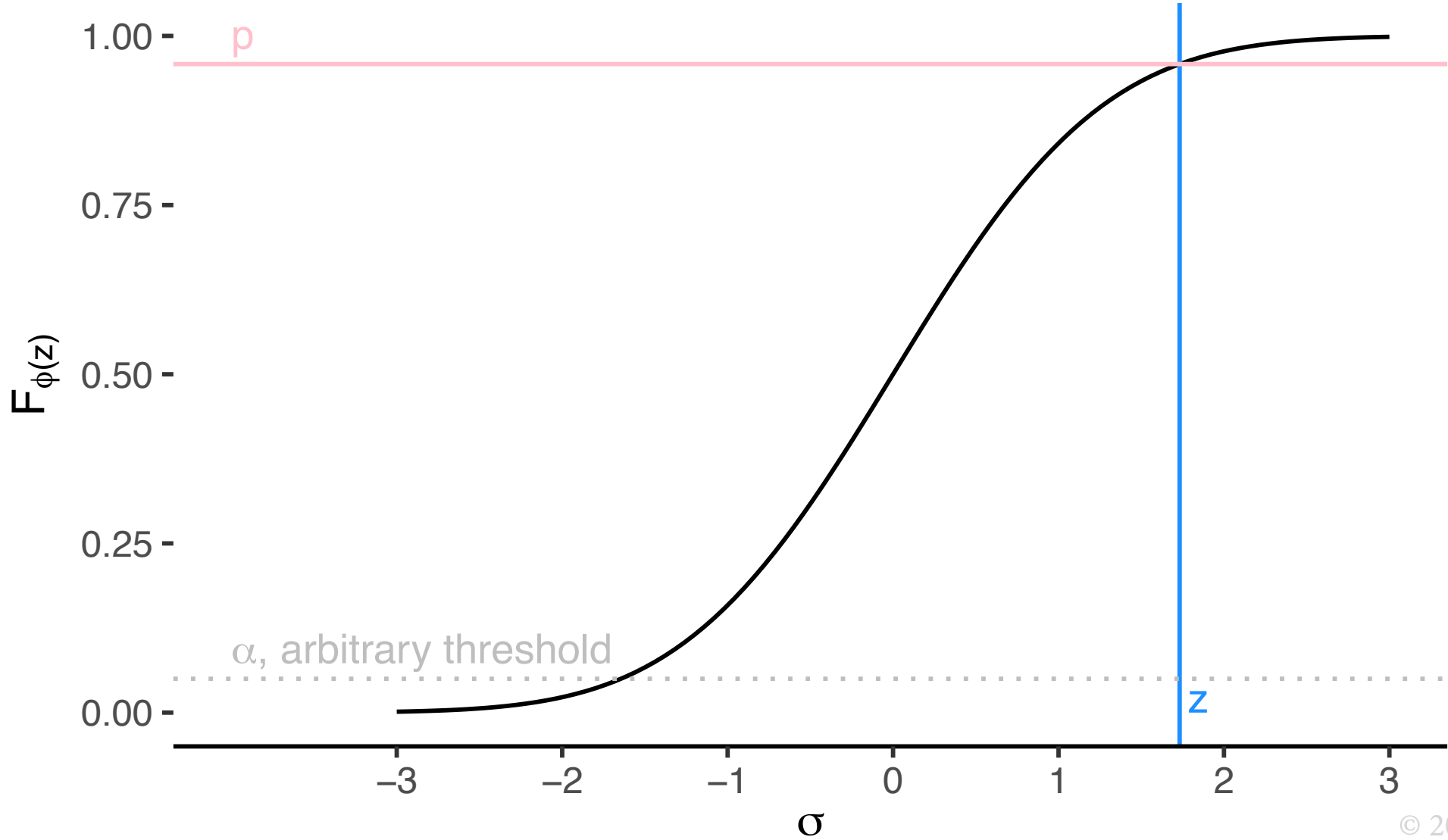
$$H_0 : \pi = \pi_0, H_A : \pi \neq \pi_0$$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}, p = F_\Phi(z)$$

comparing two observed proportions

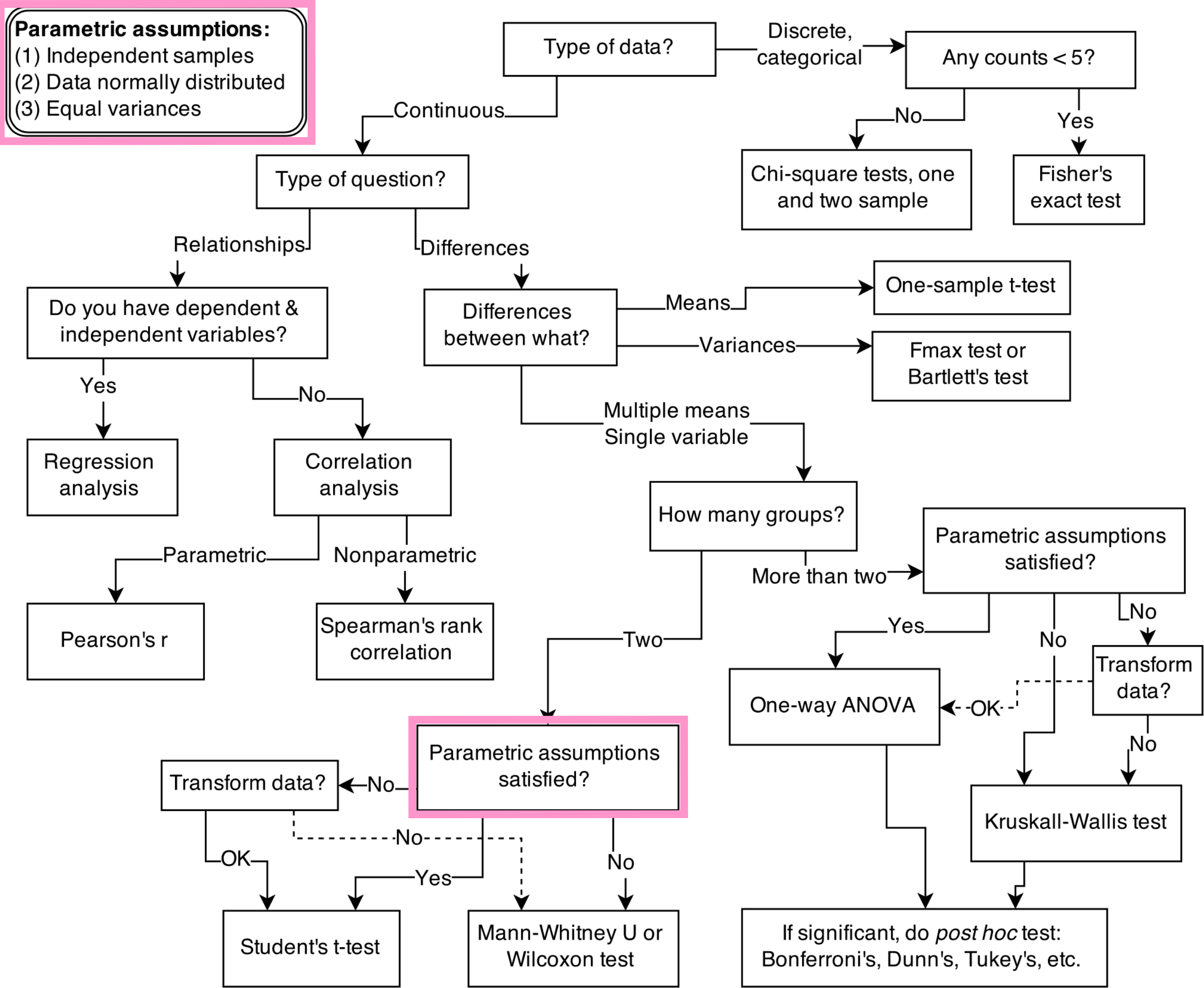$$H_0 : \pi_1 - \pi_2 = \delta, H_A : \pi_1 - \pi_2 \neq \delta$$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \delta}{\sqrt{p_0(1 - p_0)(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ where } p_0 = \frac{x_1 + x_2}{n_1 + n_2}, p = F_\Phi(z)$$
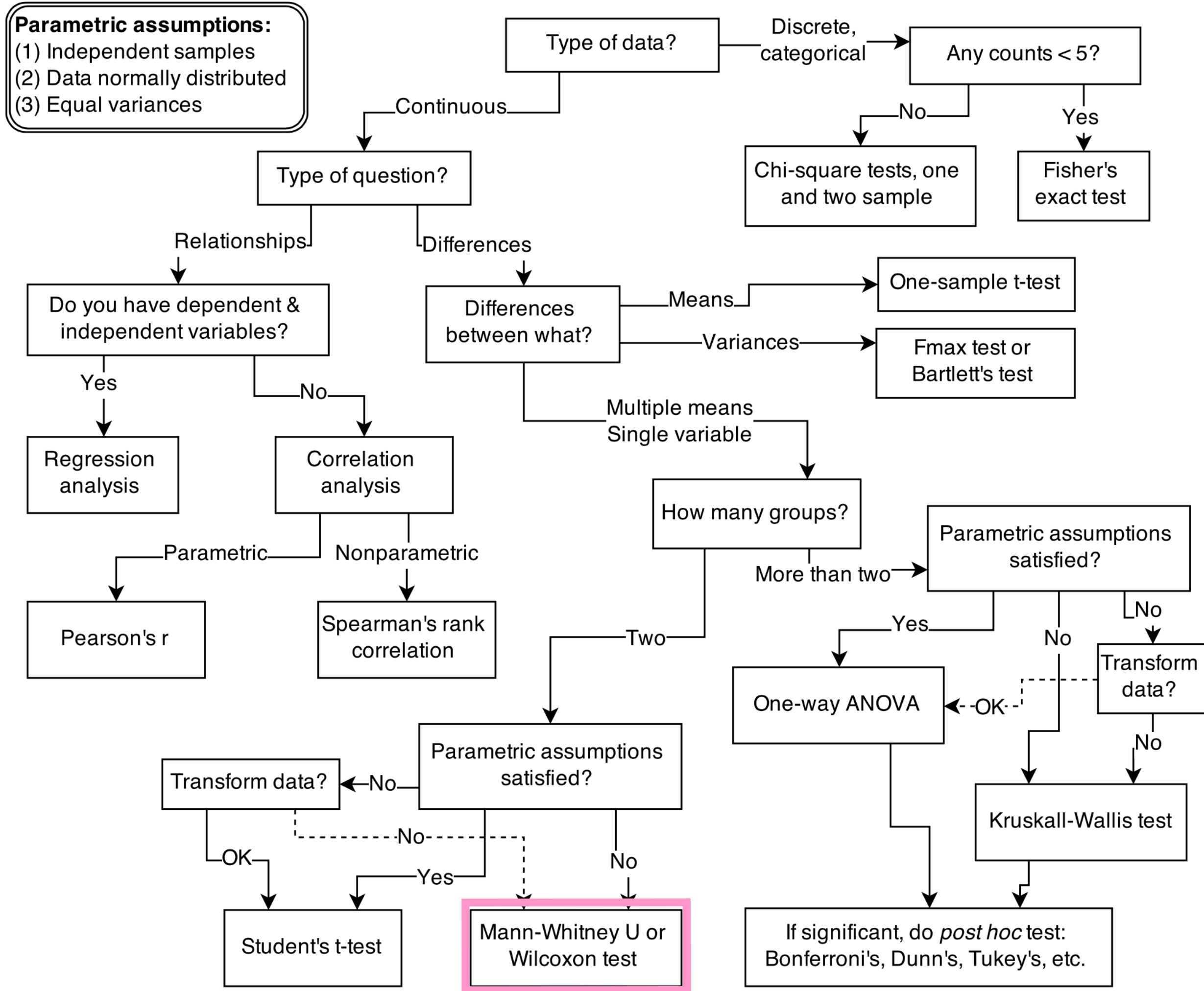
```
prop.test(...)
```

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Discrete, categorical

Any counts < 5?

No

Yes

Chi-square tests, one and two sample

Fisher's exact test

Continuous

Type of question?

Relationships

Differences

Do you have dependent & independent variables?

Yes

No

Differences between what?

Means

One-sample t-test

Variances

Fmax test or Bartlett's test

Regression analysis

Correlation analysis

Multiple means Single variable

Parametric

Nonparametric

Pearson's r

Spearman's rank correlation

How many groups?

Two

More than two

Parametric assumptions satisfied?

Yes

No

No

Transform data?

One-way ANOVA

OK

No

Parametric assumptions satisfied?

Transform data?

No

OK

No

Yes

No

Kruskall-Wallis test

Student's t-test

Mann-Whitney U or Wilcoxon test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

But what if we have ***no*** *reason to assume* the population is normally distributed?

Some tests, like *Wilcoxon–Mann–Whitney*, do not rely on parametric assumptions. *WMW* assumes independence of observations and outcomes are ordinal or continuous.

$$H_0 : P(x_i > y_j) = 1/2 \quad (i.e., \text{ populations have same central tendency})$$
$$H_A : P(x_i > y_j) \neq 1/2 \quad (i.e., \text{ central tendencies not same})$$

Simplified procedure — rank all $(n_1 + n_2)$ observations in ascending order; assign ties their average rank; sum each of the two rankings, T_a and T_b; calculate the U statistic:

$$U_a = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - T_a, \quad U_b = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T_b$$

then $U = \min(U_a, U_b)$. For $n > 20$,

$$z = \frac{U - \mathbb{E}(U)}{\sigma}, \quad \mathbb{E}(U) = \frac{n_1 n_2}{2}, \quad \sigma^* = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \quad p = F_\Phi(z)$$

*An adjustment to σ is needed for intergroup ties.

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data? — Discrete, categorical → Any counts < 5?
- No → Chi-square tests, one and two sample
- Yes → Fisher's exact test

Continuous → Type of question?
- Relationships → Do you have dependent & independent variables?
  - Yes → Regression analysis
  - No → Correlation analysis
    - Parametric → Pearson's r
    - Nonparametric → Spearman's rank correlation
- Differences → Differences between what?
  - Means → One-sample t-test
  - Variances → Fmax test or Bartlett's test
  - Multiple means, Single variable → How many groups?
    - More than two → Parametric assumptions satisfied?
      - Yes → One-way ANOVA
      - No → Transform data?
        - OK → One-way ANOVA
        - No → Kruskall-Wallis test
      - If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.
    - Two → Parametric assumptions satisfied?
      - No → Transform data?
        - OK → Student's t-test
        - No → Mann-Whitney U or Wilcoxon test
      - Yes → Student's t-test
      - No → Mann-Whitney U or Wilcoxon test

zoo & decisions, *Wilcoxon–Mann–Whitney*, simplified code, simulated example:

```r
# simulate samples from experiment, samples from different distributions
n <- 10

set.seed(1)
sample1 <- rbeta(n, 2, 2)
sample2 <- rbeta(n, 2, 5)
```

$H_0 : P(x_i > y_j) = 1/2$ (*i.e.*, populations have same central tendency)
$H_A : P(x_i > y_j) \neq 1/2$ (*i.e.*, central tendencies not same)

```r
d <-
  data.frame(
    sample = rep(1:2, each = n),
    values = c(sample1, sample2)
  ) %>%
  arrange(values) %>%
  mutate(order = seq(nrow(.))) %>%
  group_by(values) %>%
  mutate(rank = mean(order))

Ta <- filter(d, sample == 1) %>% .$rank %>% sum()
Tb <- filter(d, sample == 2) %>% .$rank %>% sum()

n1 <- with(d, sum(sample == 1))
n2 <- with(d, sum(sample == 2))

Ua <- n1 * n2 + (n1 * (n1 + 1)) / 2 - Ta
Ub <- n1 * n2 + (n2 * (n2 + 1)) / 2 - Tb
U <- min(Ua, Ub)

EU <-  n1 * n2 / 2
sigma <- sqrt( n1 * n2 * (n1 + n2 + 1) / 12 )

z <- (U - EU) / sigma
p <- pnorm(z)
```

Simplified procedure — rank all $(n_1 + n_2)$ observations in ascending order; assign ties their average rank; sum each of the two rankings, T_a and T_b; calculate the U statistic:
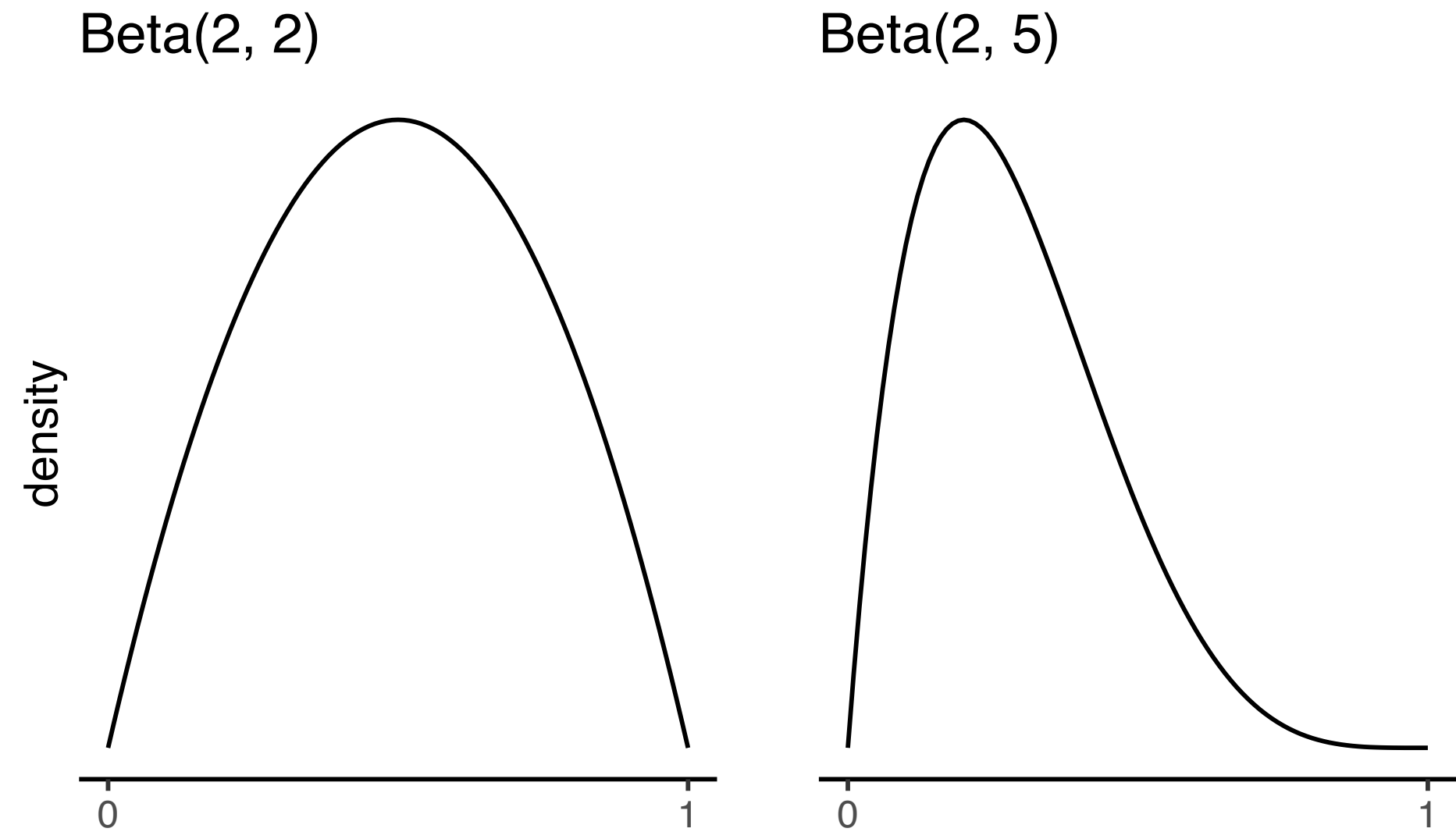
$$U_a = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - T_a, \quad U_b = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T_b$$

then $U = \min(U_a, U_b)$. For $n > 20$,

$$z = \frac{U - \mathbb{E}(U)}{\sigma}, \quad \mathbb{E}(U) = \frac{n_1 n_2}{2}, \quad \sigma^* = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \quad p = F_\Phi(z)$$

*An adjustment to σ is needed for intergroup ties.

```
a <- wilcox.test(x = sample1, y = sample2,
          correct = FALSE, exact = FALSE,
          alternative = "greater")

p == a$p.value
```

$H_0 : P(x_i > y_j) = 1/2$ (*i.e.*, populations have same central tendency)
$H_A : P(x_i > y_j) \neq 1/2$ (*i.e.*, central tendencies not same)

Simplified procedure — rank all $(n_1 + n_2)$ observations in ascending order; assign ties their average rank; sum each of the two rankings, T_a and T_b; calculate the U statistic:

$$U_a = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - T_a \,, \quad U_b = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T_b$$

then $U = \min(U_a, U_b)$. For $n > 20$,

$$z = \frac{U - \mathbb{E}(U)}{\sigma}, \quad \mathbb{E}(U) = \frac{n_1 n_2}{2}, \quad \sigma^* = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \quad p = F_\Phi(z)$$

*An adjustment to σ is needed for intergroup ties.

Beta(2, 2)

Beta(2, 5)

density

$H_0 : P(x_i > y_j) = 1/2$ (*i.e.*, populations have same central tendency)
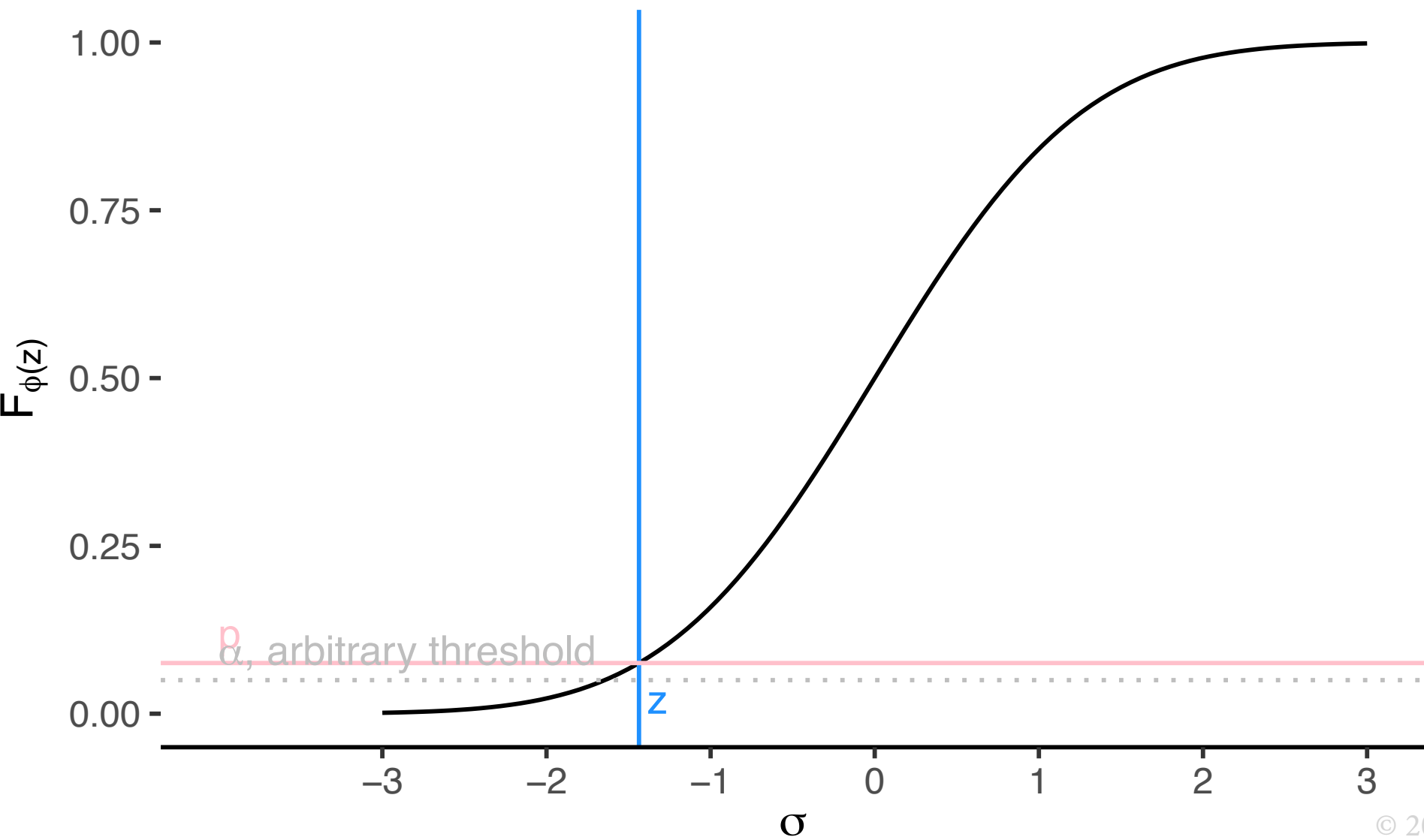$H_A : P(x_i > y_j) \neq 1/2$ (*i.e.*, central tendencies not same)

Simplified procedure — rank all $(n_1 + n_2)$ observations in ascending order; assign ties their average rank; sum each of the two rankings, T_a and T_b; calculate the U statistic:

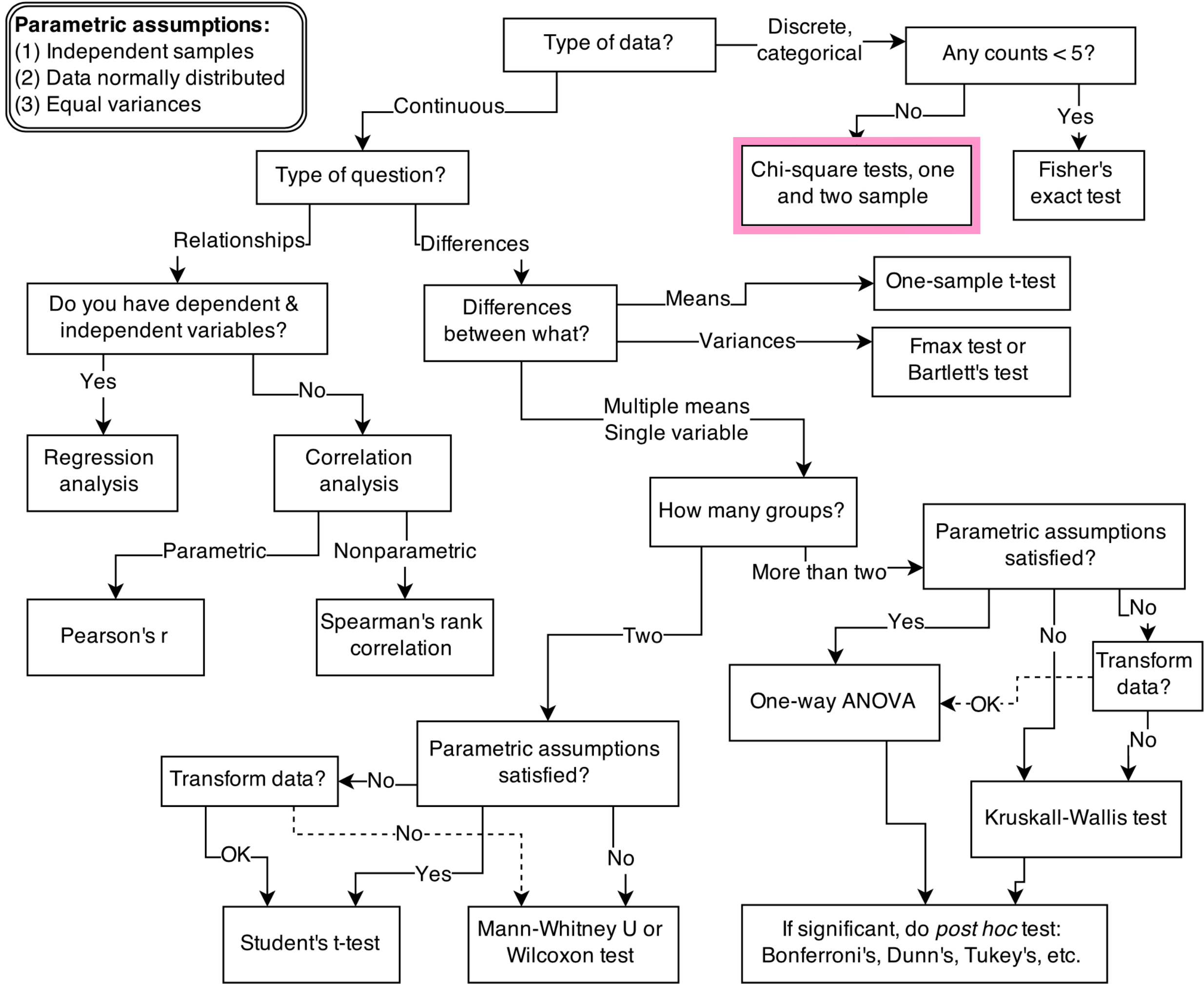$$U_a = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - T_a, \quad U_b = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - T_b$$

then $U = \min(U_a, U_b)$. For $n > 20$,

$$z = \frac{U - \mathbb{E}(U)}{\sigma}, \quad \mathbb{E}(U) = \frac{n_1 n_2}{2}, \quad \sigma^* = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \quad p = F_\Phi(z)$$

*An adjustment to σ is needed for intergroup ties.

zoo & decisions, **goodness of fit**

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data? — Discrete, categorical → Any counts < 5?

Continuous

Type of question?

Relationships — Differences

Do you have dependent & independent variables?

Differences between what? — Means → One-sample t-test

Variances → Fmax test or Bartlett's test

Yes — No

Regression analysis

Correlation analysis

Multiple means Single variable

Parametric — Nonparametric

Pearson's r

Spearman's rank correlation

How many groups?

More than two → Parametric assumptions satisfied?

Two

Parametric assumptions satisfied?

Yes — No — No → Transform data?

One-way ANOVA ← OK

Transform data? ← No

OK — No — Yes — No

Student's t-test

Mann-Whitney U or Wilcoxon test

Kruskall-Wallis test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

No → Chi-square tests, one and two sample

Yes → Fisher's exact test

Tests can suggest whether *a whole probability distribution fits some count of categorical outcomes.*

H_0 : distribution fits data

H_A : distribution doesn't fit data

$k \in \{1,...,K\}$ outcome categories

O_k observed counts for category k

p_k probability of category k

$E_k = n \cdot p_k$, expected counts for category k

w test statistic, variations from expected counts

$$w = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}, \quad \nu = K - 1, \quad p = F_{\chi^2}(w, \nu)$$

zoo & decisions, **goodness of fit**

```
# Example — test equality of proportions of male and female applicants to Berkeley

data(UCBAdmissions)

UCBAdmissions %>% as.data.frame() %>%
  group_by(Gender) %>%
  summarise(O = sum(Freq)) %>%
  ungroup() %>%
  mutate(E = mean(O)) %>%
  summarise(w = sum((O - E) ^2 / E),
            nu = n_distinct(Gender) - 1) %>%
  mutate(p = pchisq(w, nu, lower.tail = FALSE))
```
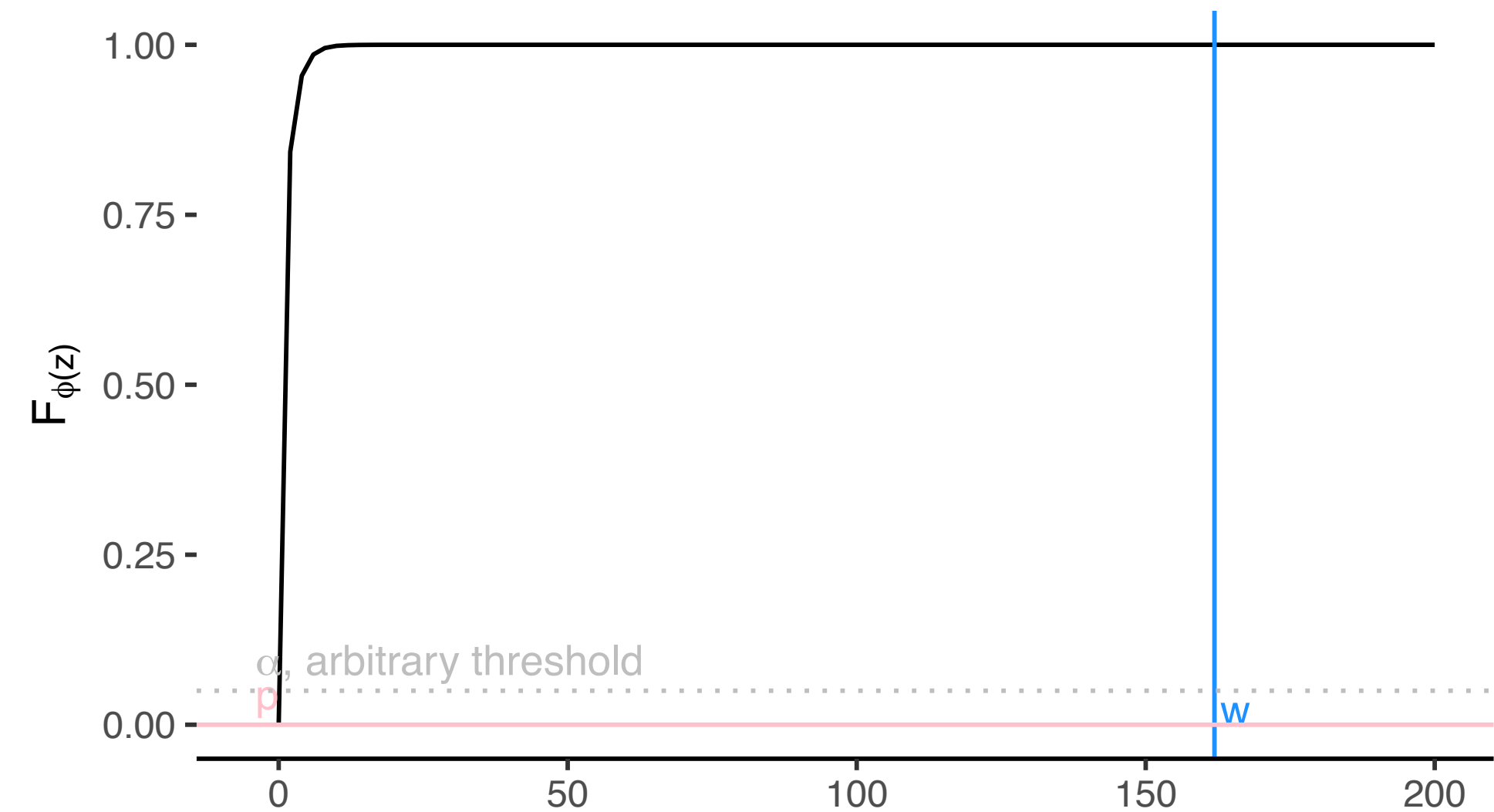
Tests can suggest whether *a whole probability distribution fits some count of categorical outcomes.*

H_0 : distribution fits data

H_A : distribution doesn't fit data

$k \in \{1,...,K\}$ outcome categories

O_k observed counts for category k

p_k probability of category k

$E_k = n \cdot p_k$, expected counts for category k

w test statistic, variations from expected counts

$$w = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}, \quad \nu = K - 1, \quad p = F_{\chi^2}(w, \nu)$$

zoo & decisions, **independence**

```
# Example — H0 : P(Admit | Gender) = P(Admit) and P(Gender | Admit) = P(Gender)

UCBAdmissions %>% as.data.frame() %>%
  mutate(Admit_pct = sum(ifelse(Admit == "Admitted", Freq, 0) ) / sum(Freq)) %>%
  group_by(Gender) %>%
  mutate(E = sum(Freq) * ifelse(Admit == "Admitted", Admit_pct, 1 - Admit_pct)) %>%
  group_by(Gender, Admit) %>%
  summarise(O = sum(Freq),
            E = mean(E)) %>%
  ungroup() %>%
  summarise(w = sum((O - E)^2 / E),
            nu = (n_distinct(Admit) - 1) * (n_distinct(Gender) - 1)) %>%
  mutate(p = pchisq(w, nu, lower.tail = FALSE))
```

Tests can suggest whether *variables are independent.*

$$H_0 : P(Y|X) = P(Y) \text{ and } P(X|Y) = P(X)$$

$$H_A : P(Y|X) \neq P(Y) \text{ or } P(X|Y) \neq P(X)$$

$k \in \{1,...,K\}$ categories

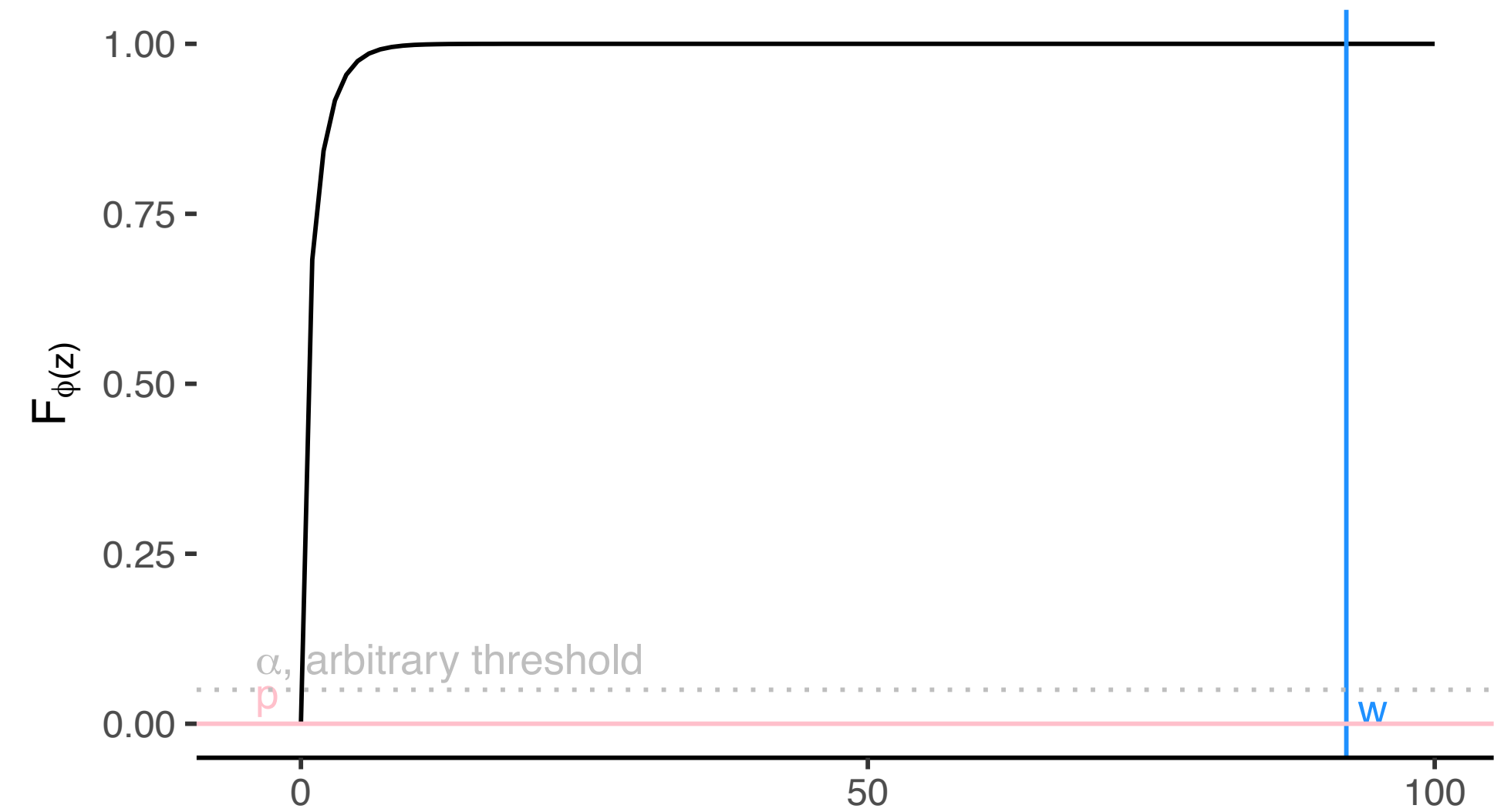$j \in \{1,...,J\}$ different levels in each category

O_{jk} observed counts for each level j in category k

p_k probability of category k

$E_k = n \cdot p_k$, expected counts for category k

w test statistic, variations from expected counts

$$w = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}, \quad \nu = (J-1)(K-1), \quad p = F_{\chi^2}(w, \nu)$$

describing variation in our tests: confidence intervals

confidence intervals

$$\left[\ (\bar{X} - \bar{Y}) + t_{\alpha/2}\sigma, \ (\bar{X} - \bar{Y}) + t_{(1-\alpha)/2}\sigma \ \right]$$

group project work!

References

Blitzstein, Joseph K., and Jessica Hwang. *Introduction to Probability*. Second edition. Boca Raton: Taylor & Francis, 2019.

Casella, George, and Roger L. Berger. *Statistical Inference*. 2nd ed. Australia ; Pacific Grove, CA: Thomson Learning, 2002.

Gelman, Andrew, Jennifer Hill, and Aki Ventari. *Regression and Other Stories*. S.l.: Cambridge University Press, 2020.

Lehmann, E L, and George Casella. *Theory of Point Estimation*. Second. Springer, 1998.

Lehmann, E. L., and Joseph P. Romano. *Testing Statistical Hypotheses*. 3rd ed. Springer Texts in Statistics. New York: Springer, 2005.

McElreath, Richard. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor and Francis, CRC Press, 2020.