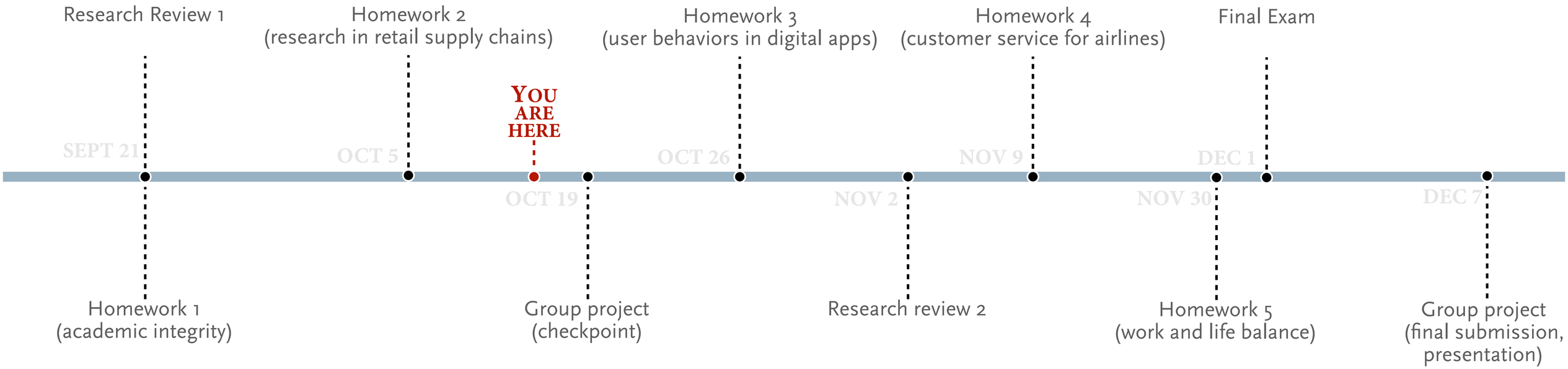
Research Design, Fall 2021

05: statistical tests continued; inference and interpretation

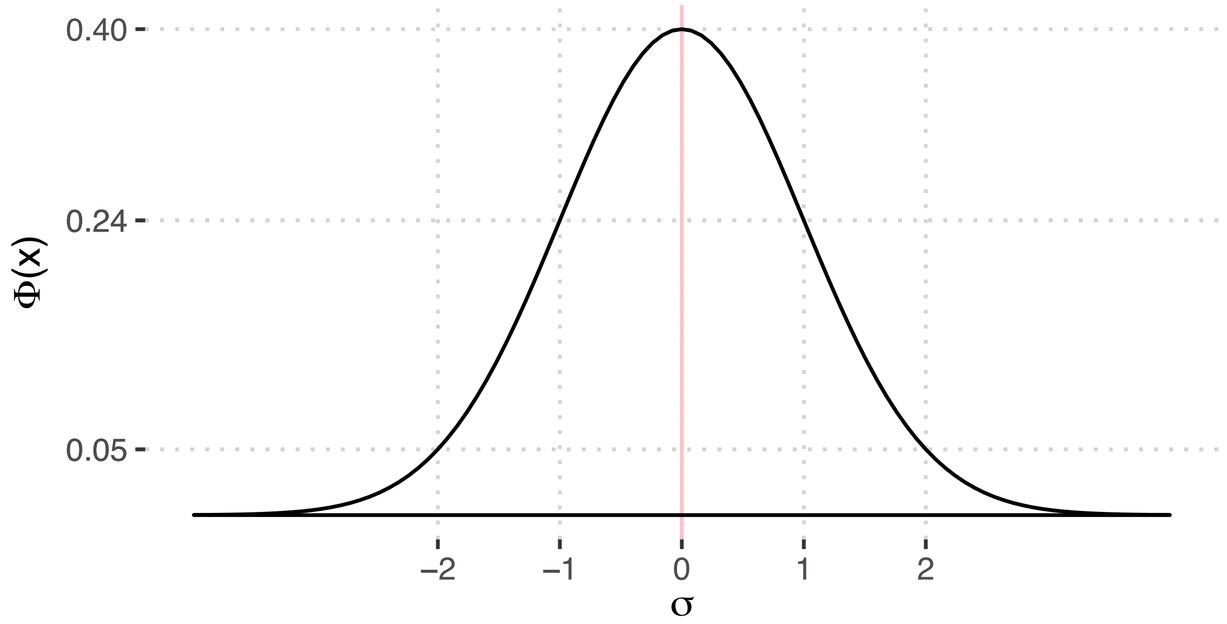
Scott Spencer | Columbia University


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 19
OCT 26
NOV 2
NOV 9
NOV 30
DEC 1
DEC 7
Homework 1
(academic integrity)
Group project
(checkpoint)
Research review 2
Homework 5
(work and life balance)
Group project
(final submission, presentation)

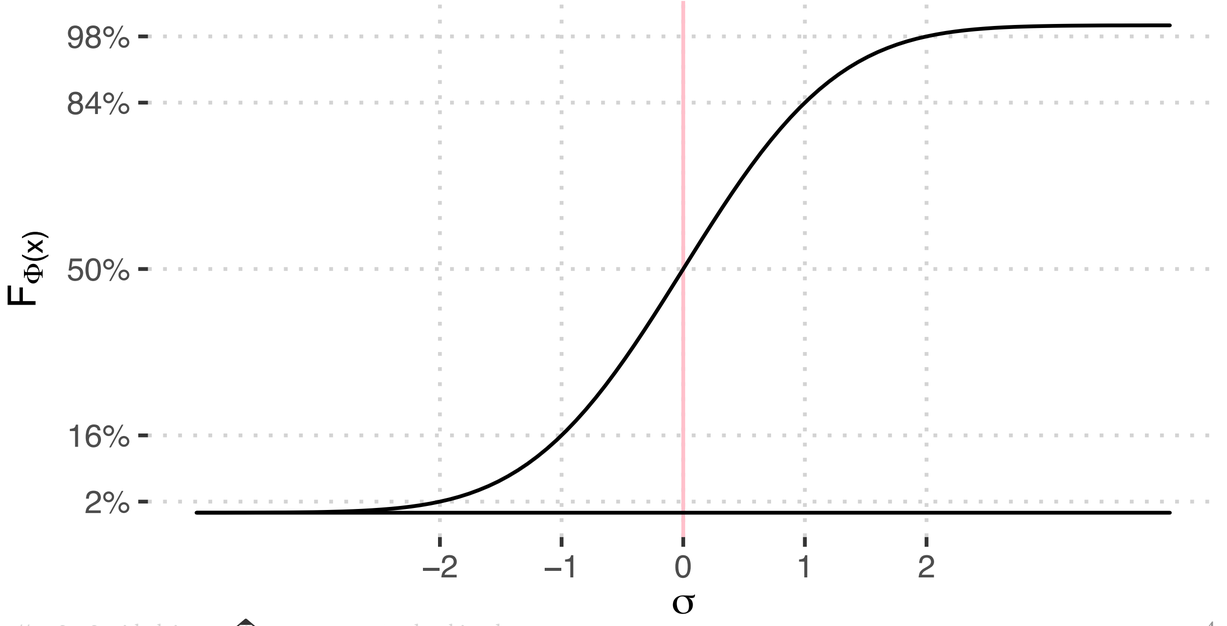
visual recap on *probability density functions* (PDF) and *cumulative distribution functions* (CDF)

probability, probability density functions v. continuous distribution functions — e.g., the standard normal Φ

```
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("$\\Phi(x)$"))
```

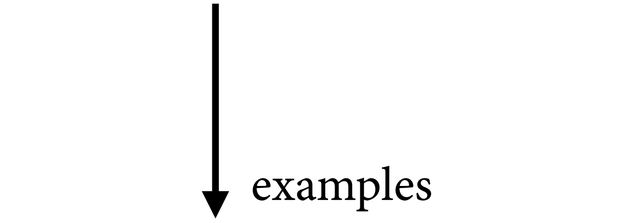



```
cdf <- ggplot() +
  theme(panel.grid.major = element_line(color = "lightgray", linetype = "dotted")) +
  scale_x_continuous(breaks = seq(-2, 2)) +
  scale_y_continuous(breaks = pnorm(seq(-2, 2)),
                     labels = scales::label_percent(accuracy = 1)) +
  geom_vline(xintercept = 0, color = "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, probability density (PDF), cumulative distribution (CDF), quantile, random generation

d<probability function name>	probability density function
p<probability function name>	cumulative distribution function (of probability)
q<probability function name>	quantile function
r<probability function name>	random generation function



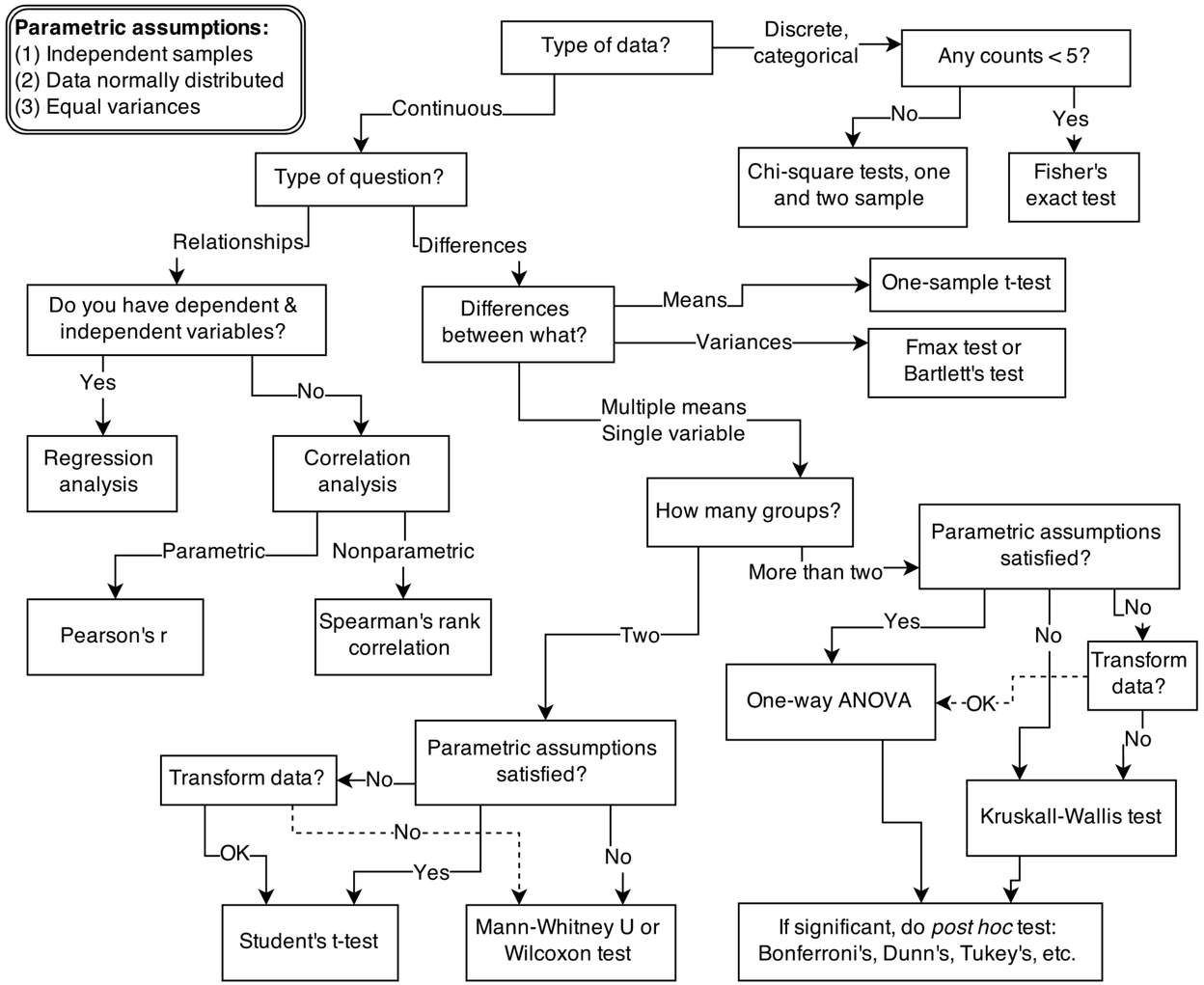
examples

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

‘a zoo of pre-constructed golems known as “tests” ’

— Richard McElreath

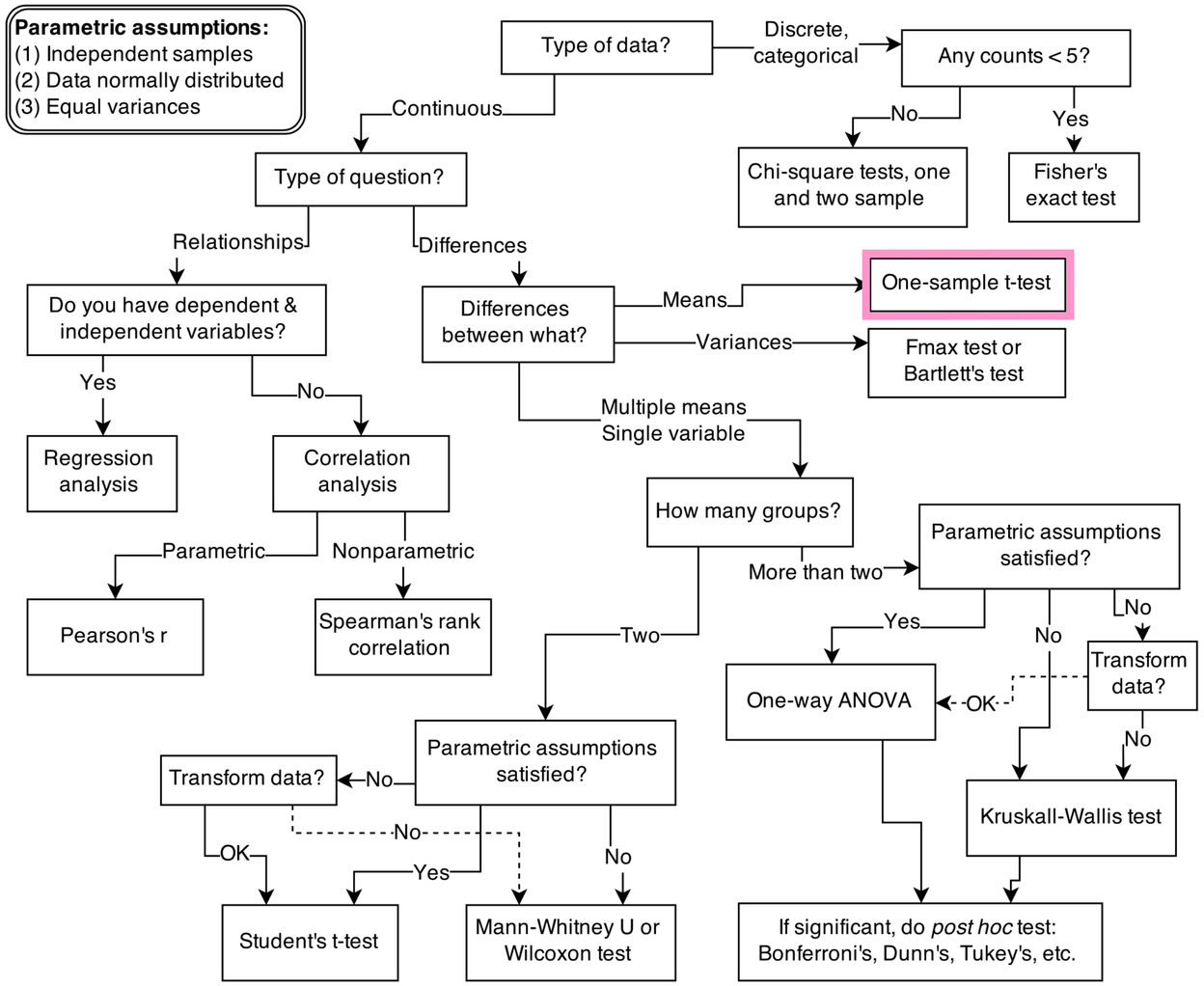
a zoo of tests, a decision tree for selecting one (and *not all* named or unnamed tests are listed below)



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

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

zoo & decisions, **comparing $\bar{x}$ to μ, $x \in \mathbb{R}$**



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

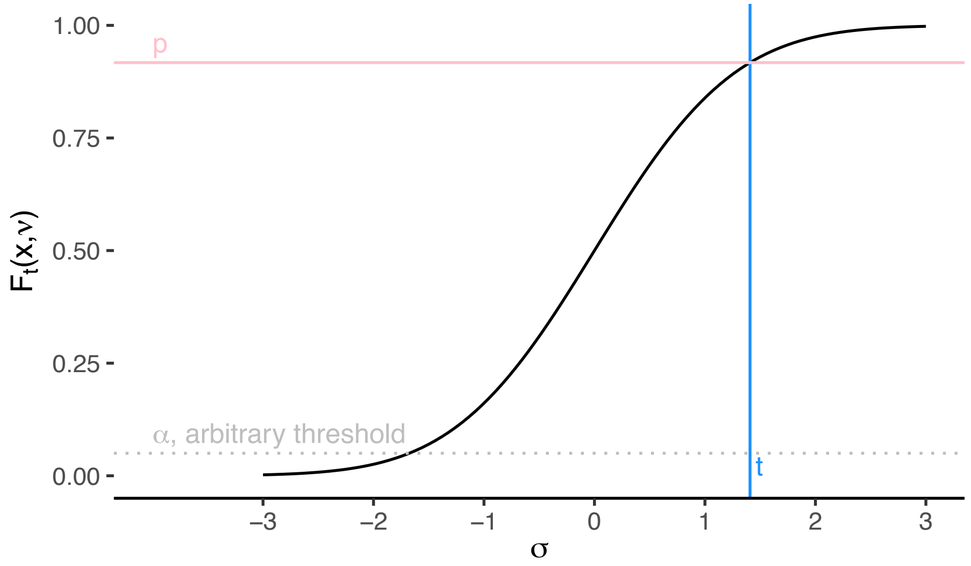
$$H_0 : \bar{x} = \mu, H_A : \bar{x} < \mu$$

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

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

zoo & decisions, **example** — comparing $\bar{x}$ to μ, $x \in \mathbb{R}$

```
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)
```

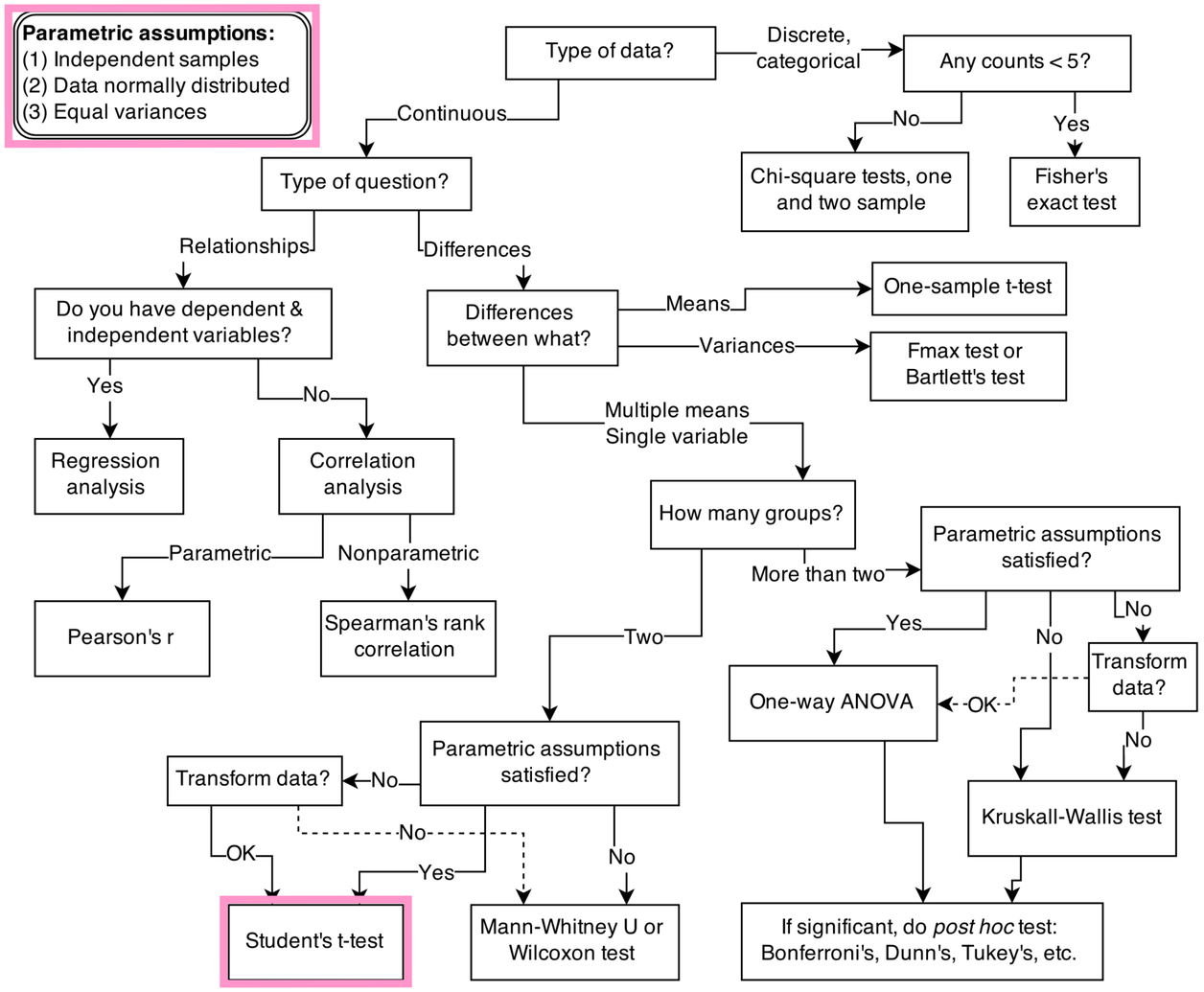


Recall Student's t test — comparing one sample means to μ where we assume a population normal distribution with unknown standard deviation σ.

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

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

zoo & decisions, comparing locations, data as $\mathbb{R}$



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

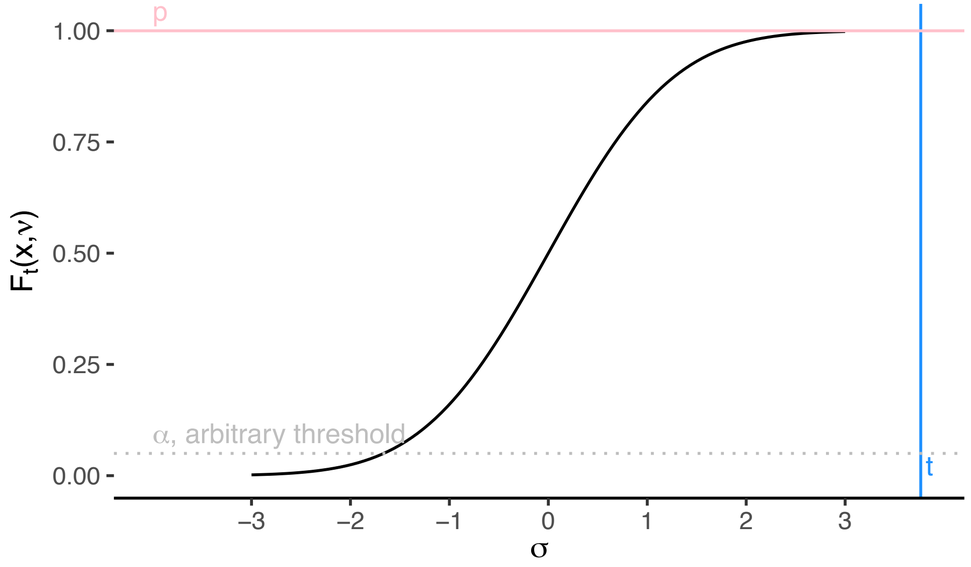
$$H_0 : \mu_1 = \mu_2, H_A : \mu_1 < \mu_2$$

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

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

zoo & decisions, **example** comparing locations, data as $\mathbb{R}$

```
# 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)

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)
```

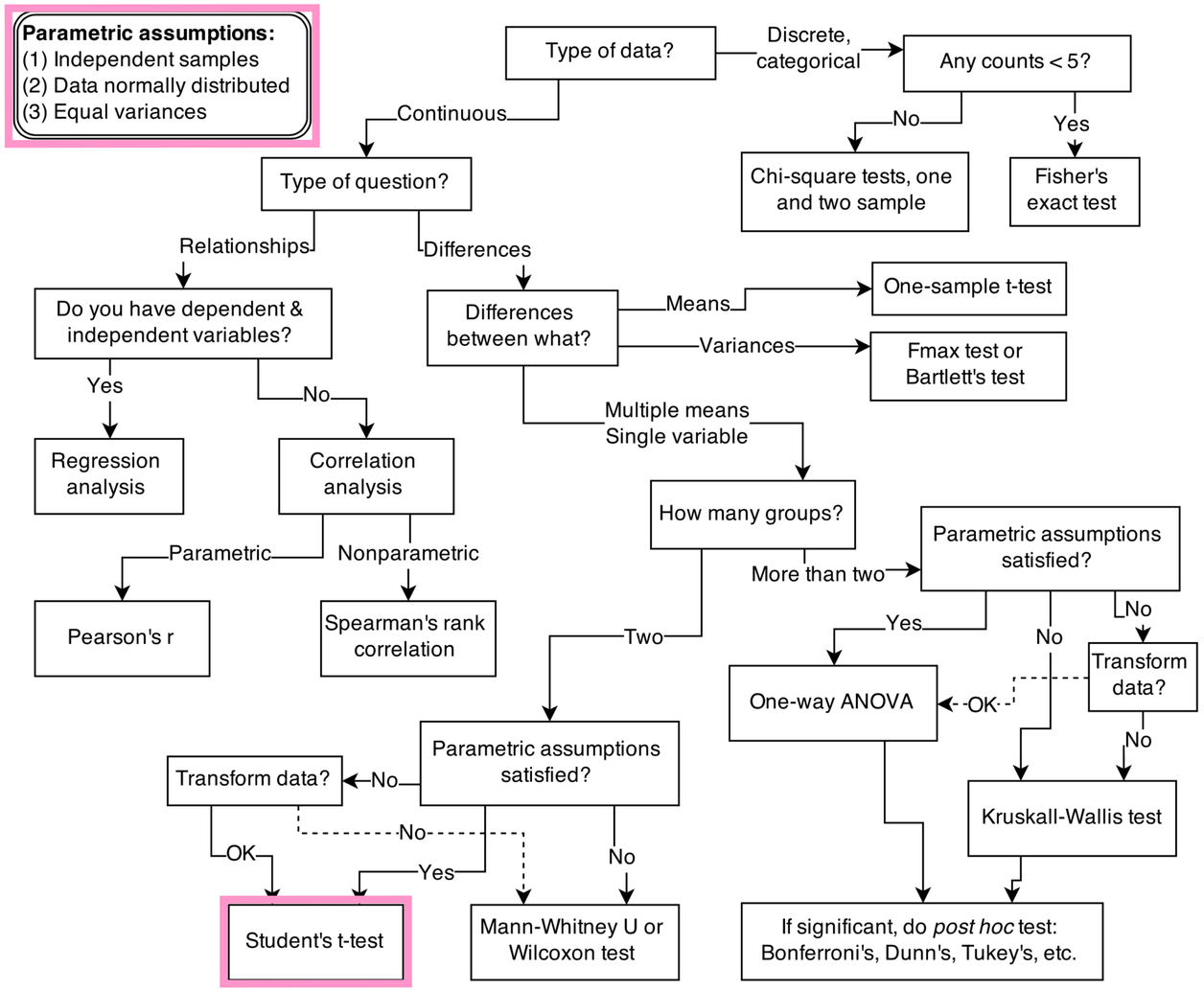


Student's t test — comparing two sample means where we can assume an underlying normal distribution.

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

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

```
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

comparing observed proportion to probability

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

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

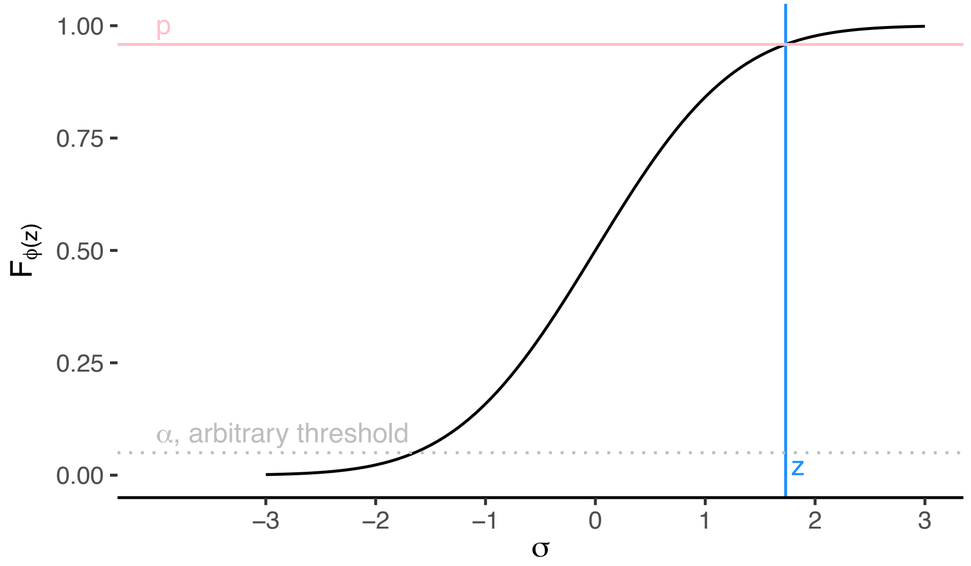
comparing two observed proportions

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

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

zoo & decisions, **comparing locations, data as $\mathbb{R} \in [0,1]$**

```
# 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)
```



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

comparing observed proportion to probability

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

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

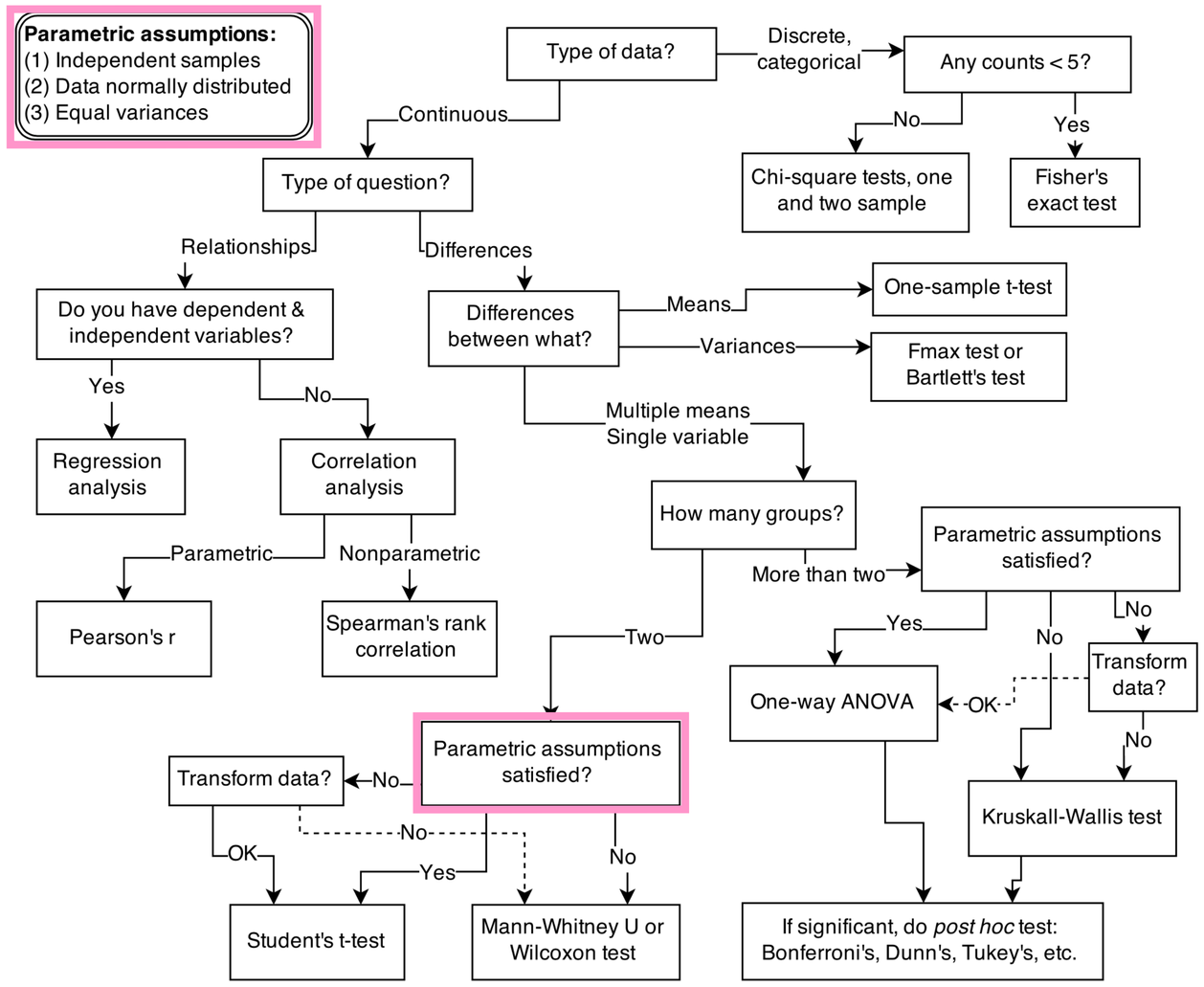
comparing two observed proportions

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

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

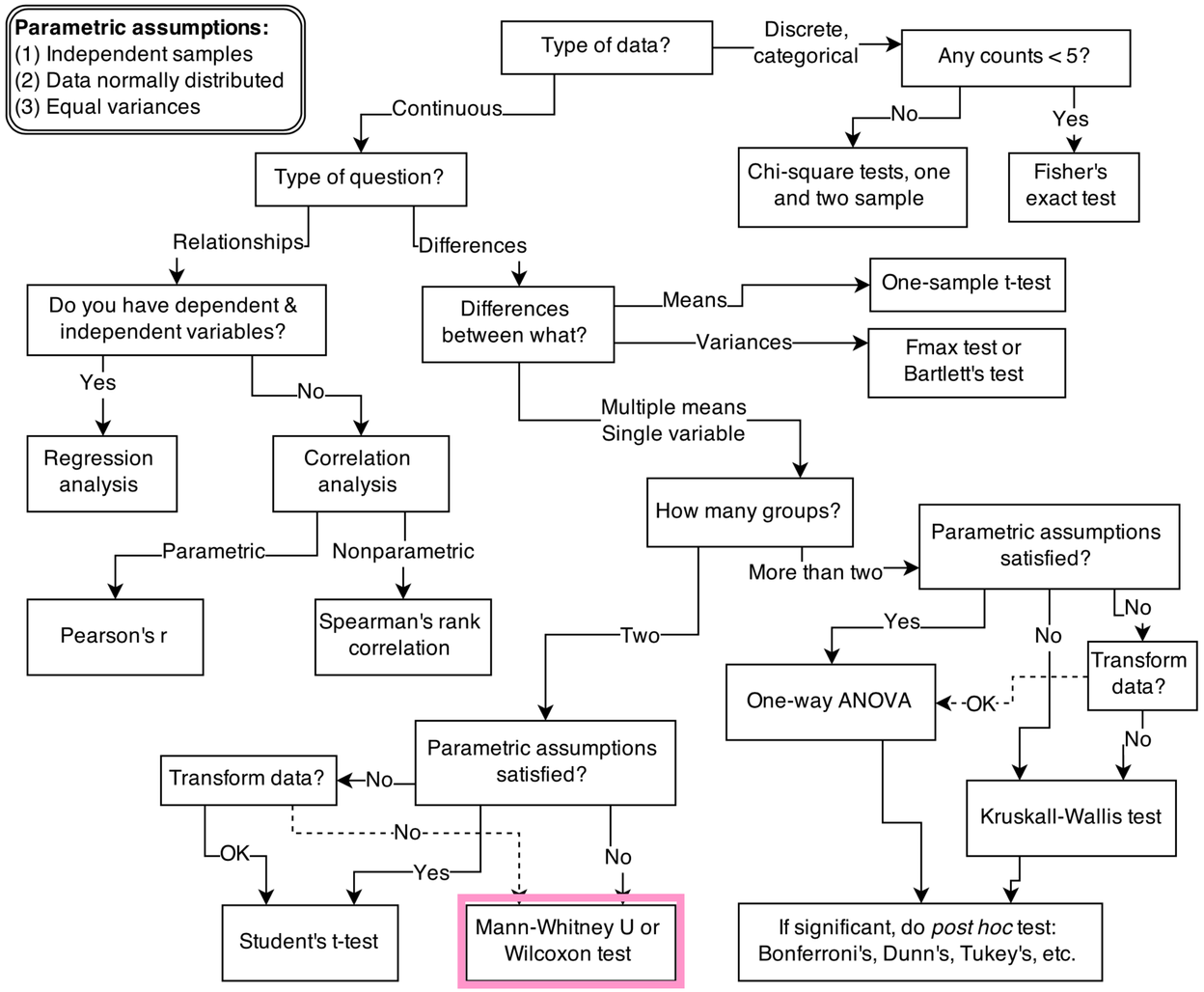
```
prop.test(...)
```

zoo & decisions, **comparing locations, data as an unknown or skewed distribution**



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

zoo & decisions, **comparing locations, data as an unknown or skewed distribution**



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

$H_0 : P(x_i > y_j) = 1/2$ (*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, *Wilcoxon–Mann–Whitney*, simplified code, simulated example:

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

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

```
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)
```

$H_0 : P(x_i > y_j) = 1/2$ (*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, *Wilcoxon–Mann–Whitney*, simplified code, simulated example, comparing results with r function

```
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)

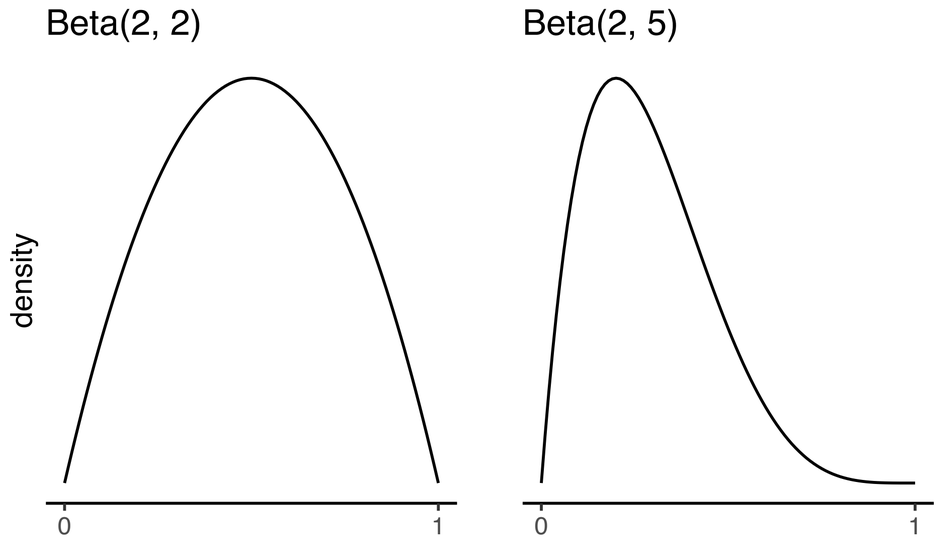
Simplified procedure — rank all $(n_1 + n_2)$ observations in ascending order; assign ties their average rank; sum each of the two rankings, T_a and T_b; calculate the U statistic:

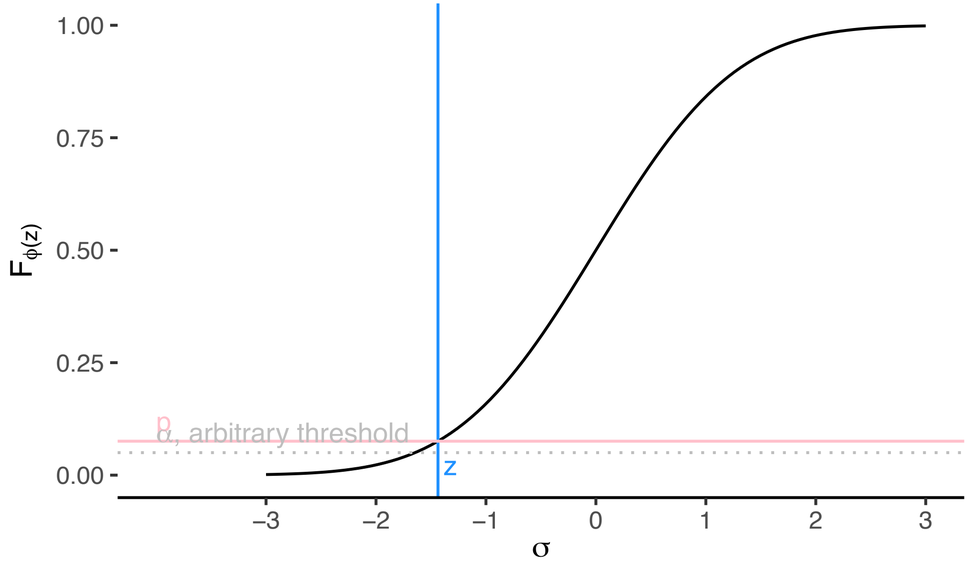
$$U_a = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_a\,, \quad U_b = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_b$$

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.





$H_0 : P(x_i > y_j) = 1/2$ (*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:

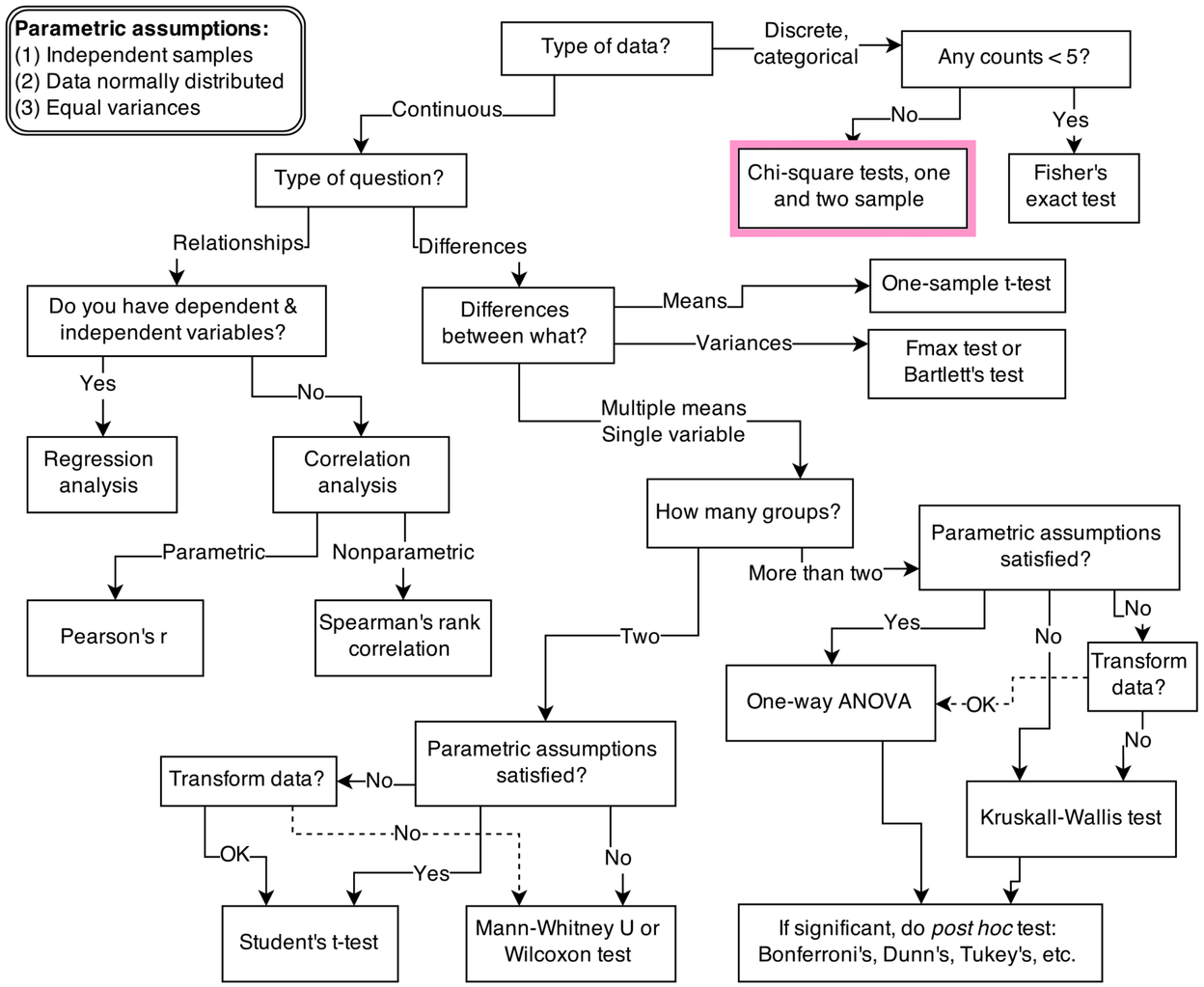
$$U_a = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_a\,, \quad U_b = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_b$$

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**



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

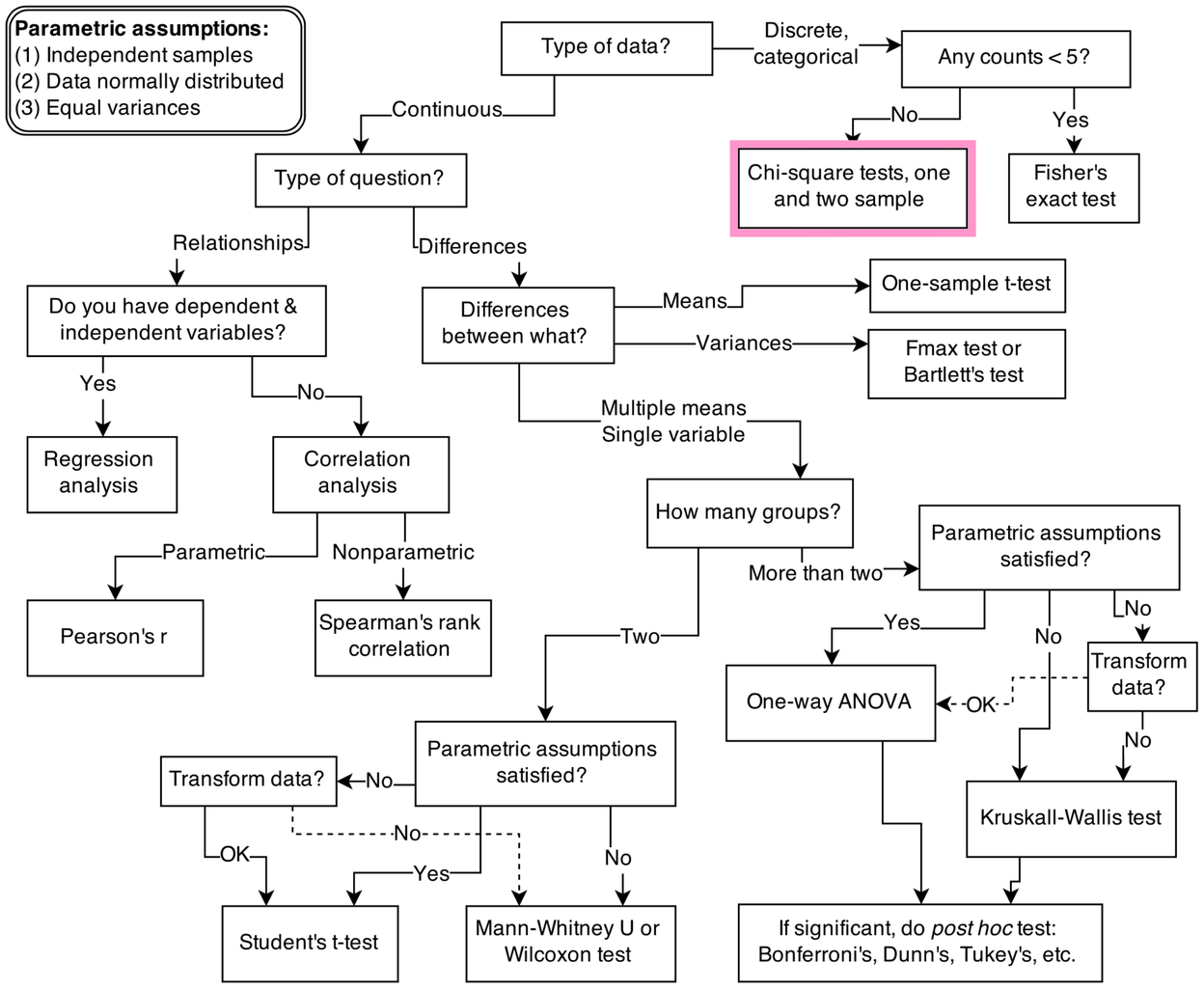
$E_k = n \cdot p_k$, expected counts for category k

w test statistic, variations from expected counts

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

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

library(dplyr)
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))
```

zoo & decisions, **independence**



Tests can suggest whether *variables are independent*.

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

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

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

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

```
# 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))
```

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.