Nonparametric hypothesis testing

Nonparametric testing for averages

Average of an unknown distribution

$$X^n = (X_1, \dots, X_n), X \sim F_X(x)$$

Is X 0 on average?

In theory, we could use any statistic T, but how do we find its null distribution?

Average of an unknown distribution

$$X^n = (X_1, \dots, X_n), X \sim F_X(x)$$

Is X 0 on average?

In theory, we could use any statistic T, but how do we find its null distribution?

Problems:

- $F_X(x)$ could be nonstandard
- CLT does not always work

Average of an unknown distribution

$$X^n = (X_1, \dots, X_n), X \sim F_X(x)$$

Is X 0 on average?

In theory, we could use any statistic T, but how do we find its null distribution?

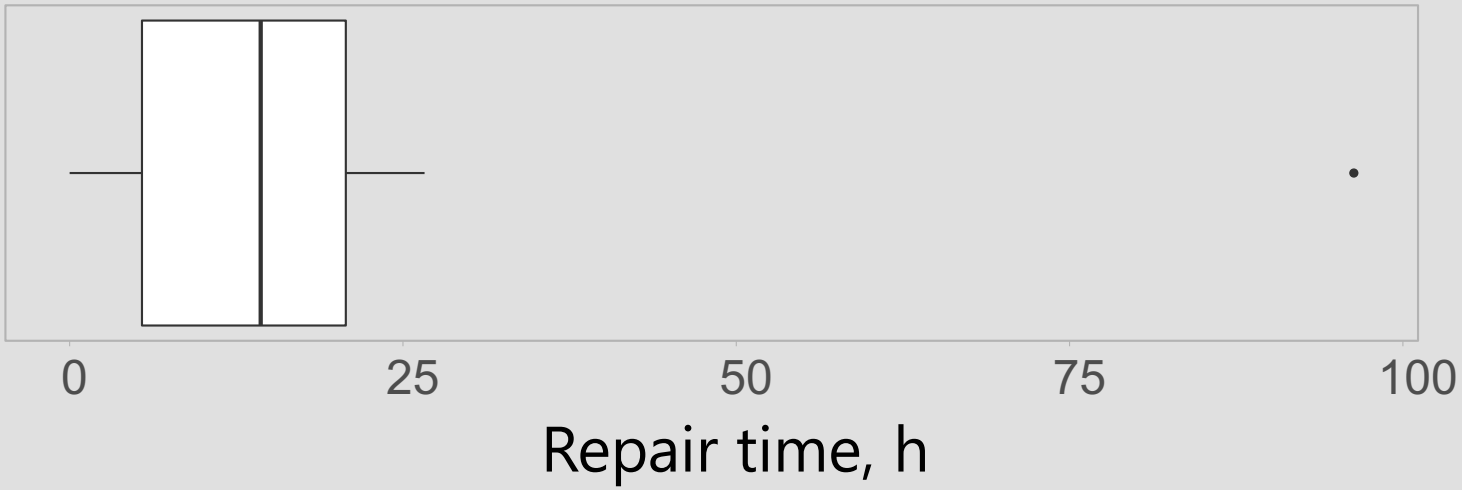
Solutions:

- Transform the sample into something standard
- Make some assumptions about $F_X(x)$

Sign tests

Example: Verizon repair time

For internet provider Verizon, we know a sample of repair times for 23 customers of their Competitive Local Exchange Carriers. Verizon has obligations to perform repairs for them as quickly as for their own customers.



Is average repair time over 8 hours?

Sign test

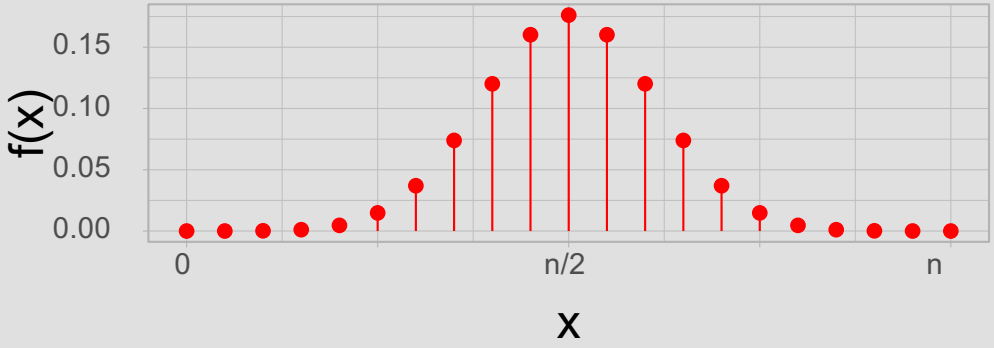
sample: $X^n = (X_1, \dots, X_n), X_i \neq m_0$

null hypothesis: $H_0: \operatorname{med} X = m_0$

alternative hypothesis: $H_1: \operatorname{med} X <\neq> m_0$

statistic: $T = \sum_{i=1}^{n} [X_i > m_0]$

null distribution: $Bin(n, 0.5)$



Example: Verizon repair time

H_0: median repair time is 8 hours or less
H_1: median repair time is greater than 8 hours

$T = 15$ – repair took more than 8 hours in 15 cases of 23.
Sign test: $p = 0.105$

There's no evidence that median repair time is over 8 hours!

Censored sample

Survival time (in weeks) for patients with symptomatic non-Hodgkin's lymphoma:

49, 58, 75, 110, 112, 132, 151, 276, 281, 362*

One patient was still alive after 7 years of study – that observation is censored.

Is the average survival time more than 2 years?
Sign test: $p = 0.3438$

Dinse, 1982

Example: classifier accuracy

	AUC C4.5	AUC C4.5+m
adult (sample)	0.763	0.768
breast cancer	0.599	0.591
breast cancer wisconsin	0.954	0.971
cmc	0.628	0.661
ionosphere	0.882	0.888
Iris	0.936	0.931
liver disorders	0.661	0.668
lung cancer	0.583	0.583
lymphography	0.775	0.838
mushroom	1	1
primary tumor	0.94	0.962
rheum	0.619	0.666
voting	0.972	0.981
wine	0.957	0.978

Demsar, 2006

Example: classifier accuracy

	AUC C4.5	AUC C4.5+m
adult (sample)	0.763	**0.768**
breast cancer	**0.599**	0.591
breast cancer wisconsin	0.954	**0.971**
cmc	0.628	**0.661**
ionosphere	0.882	**0.888**
Iris	**0.936**	0.931
liver disorders	0.661	**0.668**
lung cancer	0.583	0.583
lymphography	0.775	**0.838**
mushroom	1	1
primary tumor	0.94	**0.962**
rheum	0.619	**0.666**
voting	0.972	**0.981**
wine	0.957	**0.978**

Demsar, 2006

Sign test, paired samples

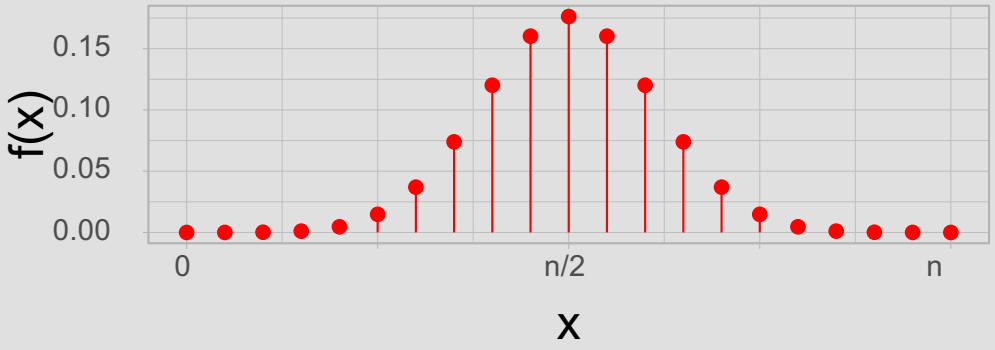
samples: $X_1^n = (X_{11}, \dots, X_{1n})$
$X_2^n = (X_{21}, \dots, X_{2n})$,
$X_{1i} \neq X_{2i}$

null hypothesis: $H_0: \mathrm{P}(X_1 > X_2) = \frac{1}{2}$

alternative hypothesis: $H_1: \mathrm{P}(X_1 > X_2) <\neq> \frac{1}{2}$

statistic: $T = \sum_{i=1}^{n} [X_{1i} > X_{2i}]$

null distribution: $Bin(n, 0.5)$



Example: classifier accuracy

H_0: classifiers have the same AUC,

$$P(AUC_{C4.5} > AUC_{C4.5+m}) = \frac{1}{2}$$

H_1: modified classifier has higher AUC,

$$P(AUC_{C4.5} > AUC_{C4.5+m}) > \frac{1}{2}$$

Modified classifier wins on 10 datasets of 14, on 2 there are ties.

Sign test: $p = 0.0193$, modified classifier is better on 83% of datasets.

Takeaways about sign tests

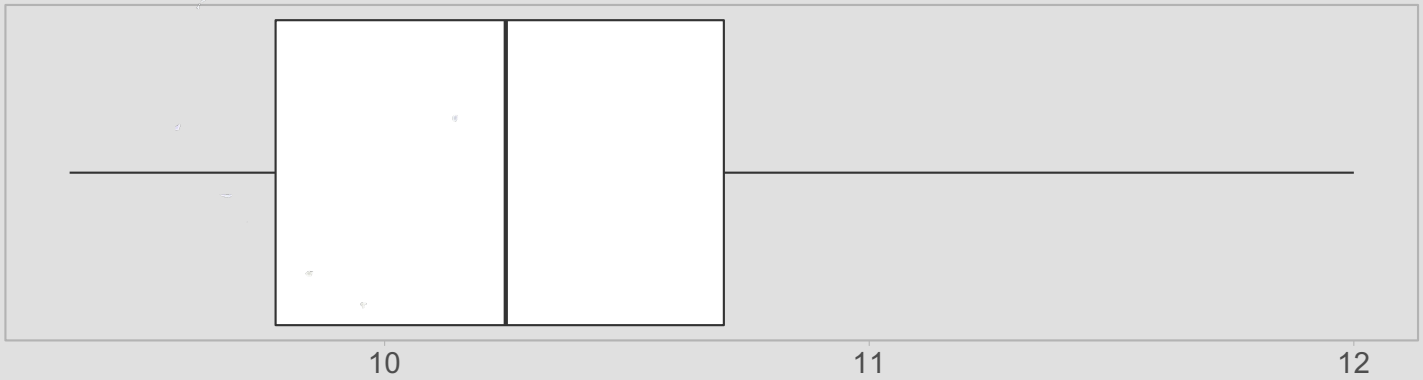
- Samples from unspecified distributions could be turned into zeros and ones to test hypotheses about averages

- Sign tests for one sample and two paired samples

Rank tests

Example: washers' production

Production line inspection measured 24 washers; their diameter is supposed to be 10 mm.



Is the average diameter close to nominal?

Signed rank test

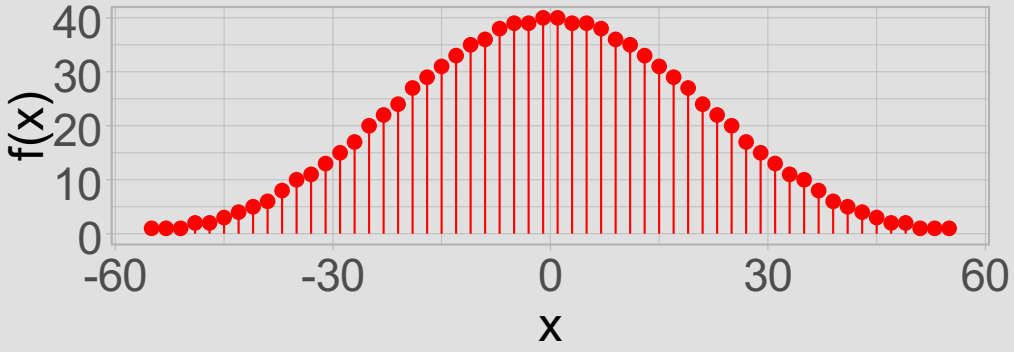
sample: $X^n = (X_1, \dots, X_n), X_i \neq m_0,$
F_X is symmetric around median

null hypothesis: $H_0: \operatorname{med} X = m_0$

alternative hypothesis: $H_1: \operatorname{med} X <\neq> m_0$

statistic: $W = \sum_{i=1}^{n} \operatorname{rank}(|X_i - m_0|)\operatorname{sign}(X_i - m_0)$

null distribution: tabulated



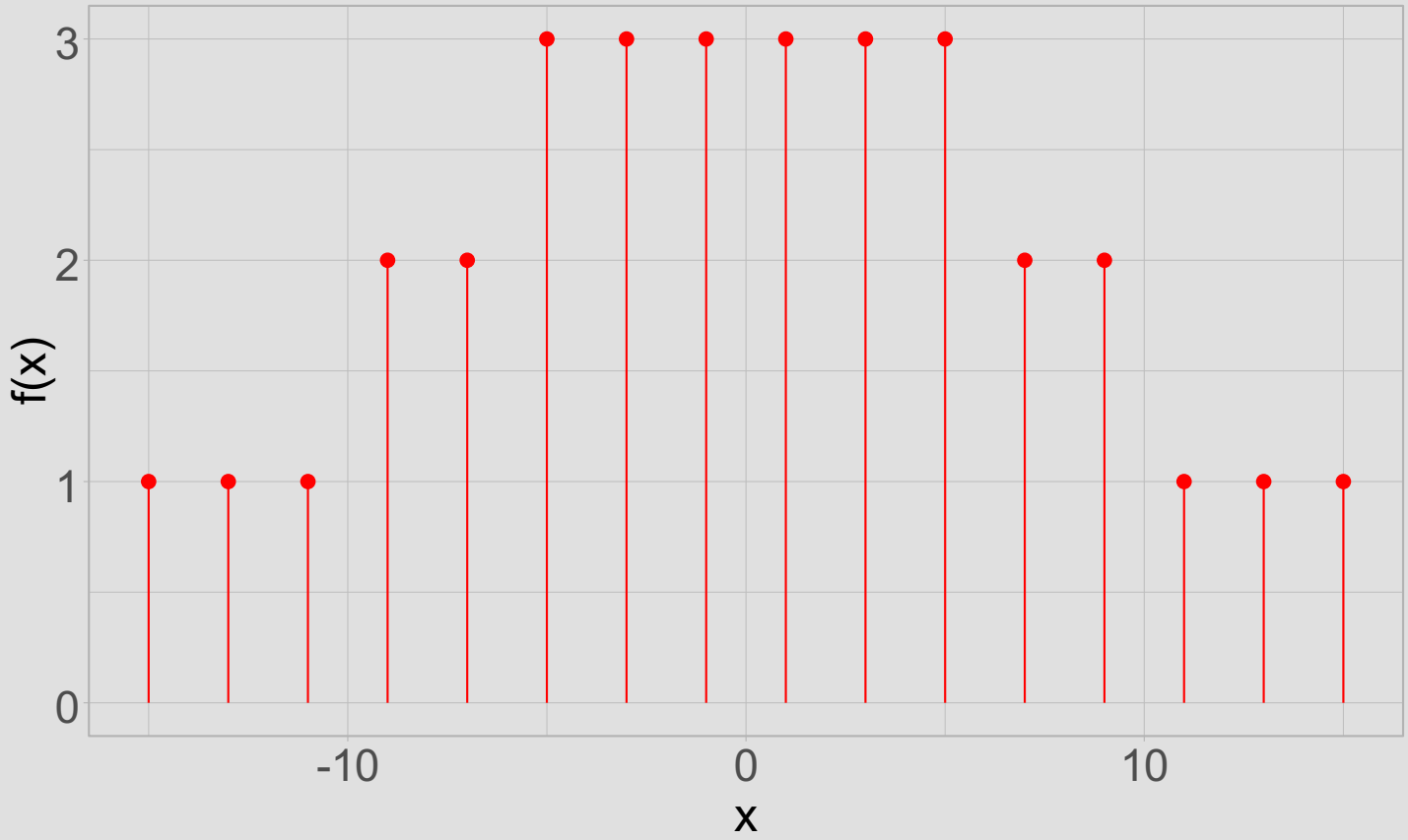
Null distribution

1	2	3	4	5	W
-	-	-	-	-	-15
+	-	-	-	-	-13
-	+	-	-	-	-11
+	+	-	-	-	-9
-	-	+	-	-	-9
			…		
+	+	-	+	+	9
-	-	+	+	+	9
+	-	+	+	+	11
-	+	+	+	+	13
+	+	+	+	+	15

2^n combinations in total

Null distribution

$n = 5$:

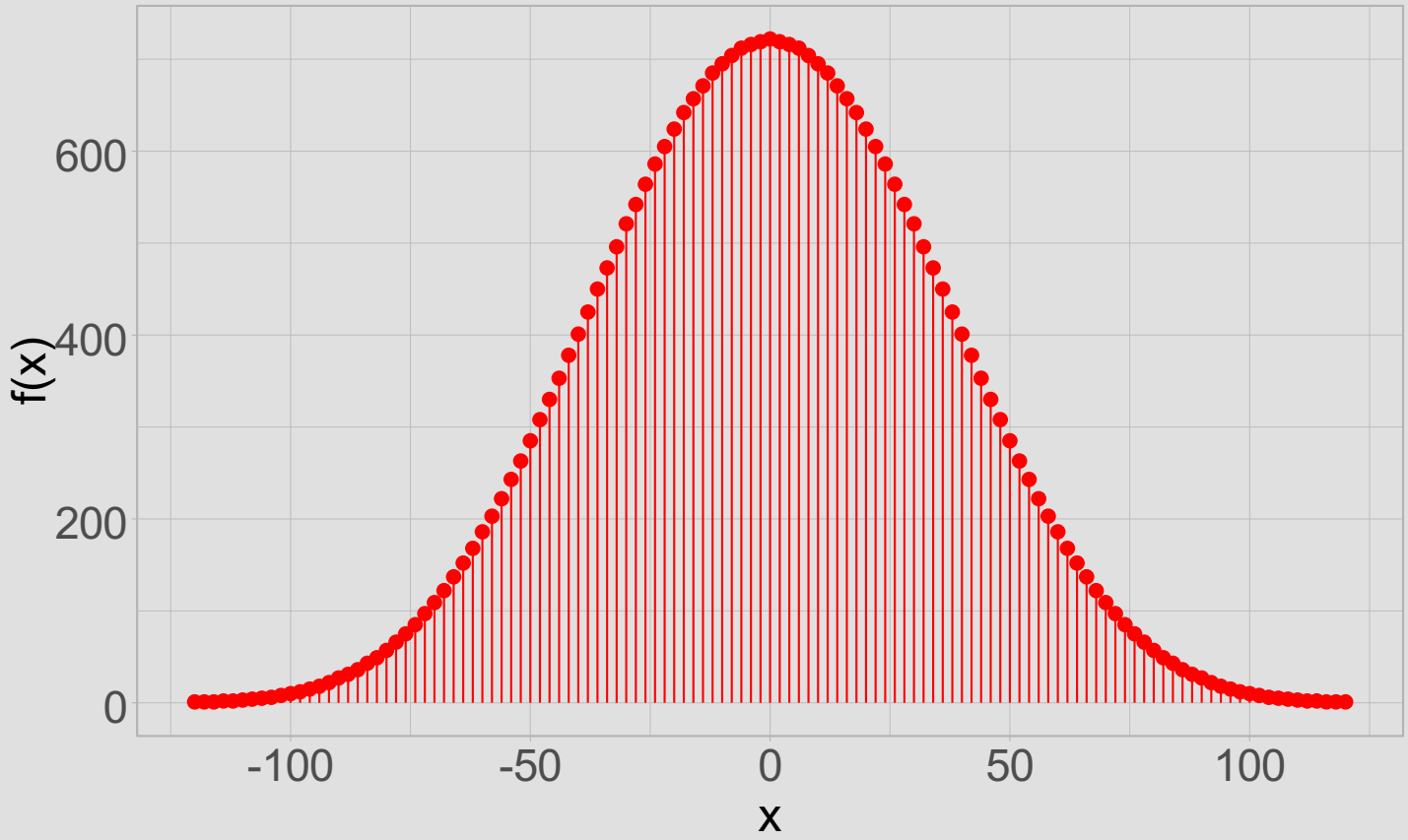


Null distribution

$n = 10$:

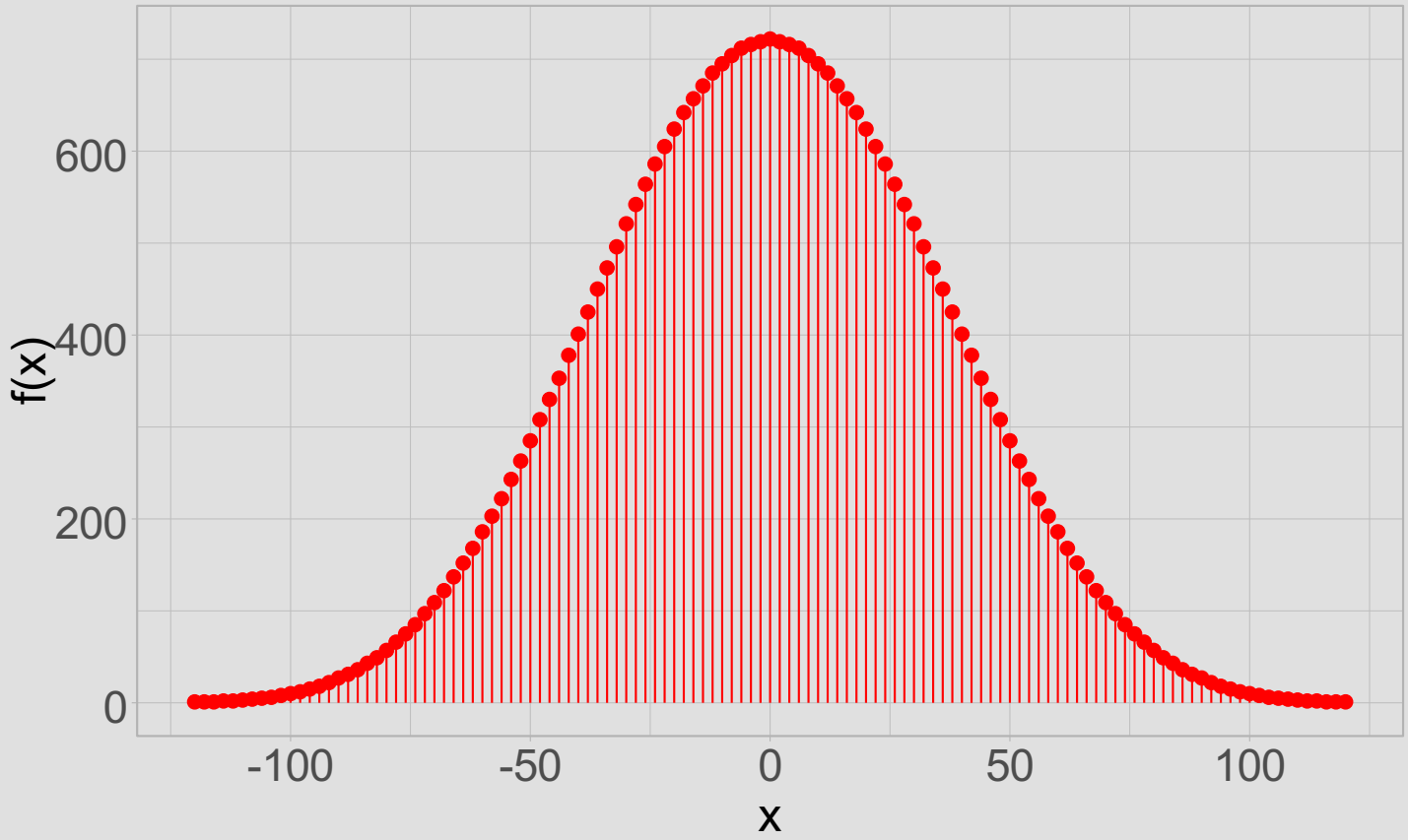
Null distribution

$n = 15$:



Null distribution

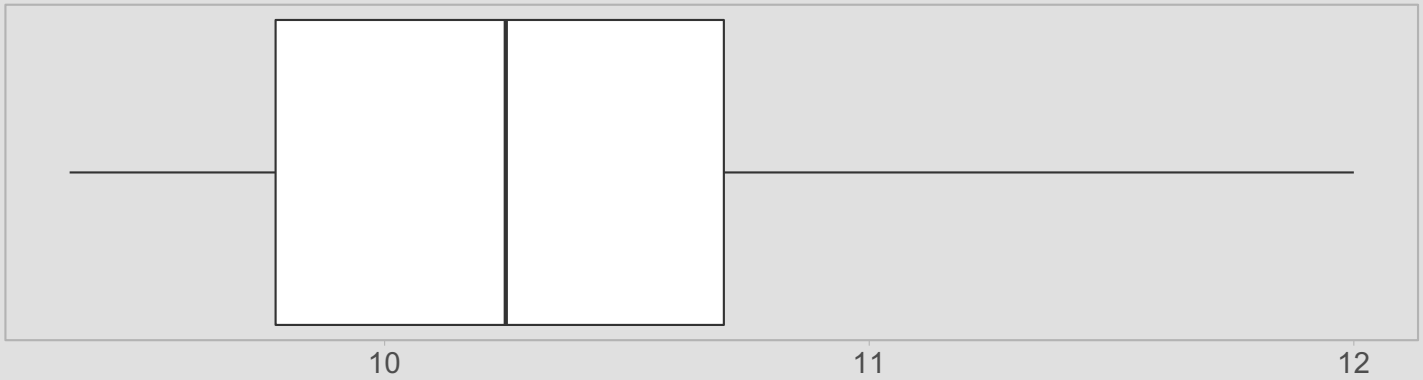
$n = 15$:



Approximation for $n > 20$:

$$W\sim \approx N\left(0, \frac{n(n+1)(2n+1)}{6}\right)$$

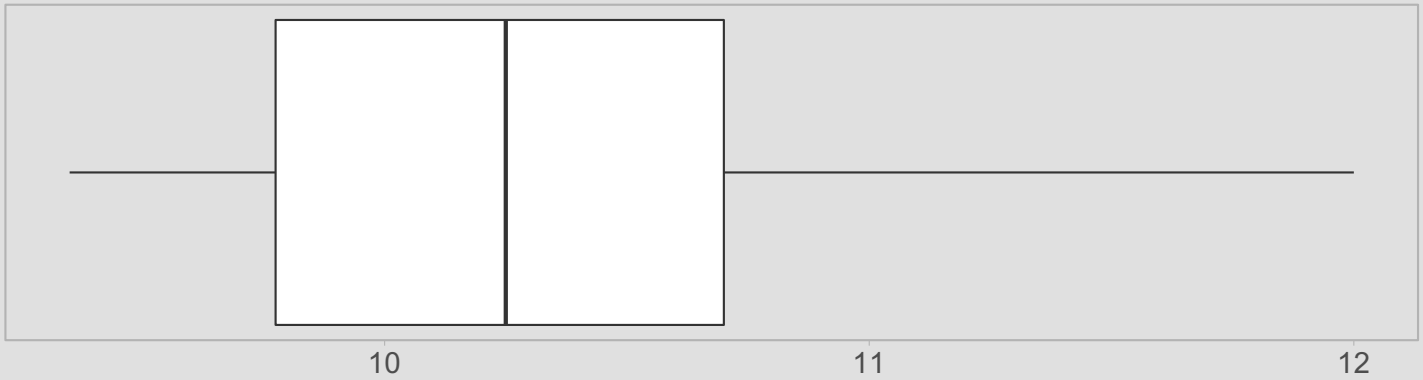
Example: washers' production



H_0: median diameter is 10 mm
H_1: median diameter is not 10 mm

Signed rank test: $p = 0.0673$, median diameter is 10.25 mm (97.7% confidence interval for the median – $[9.8, 10.7]$ mm.

Example: washers' production



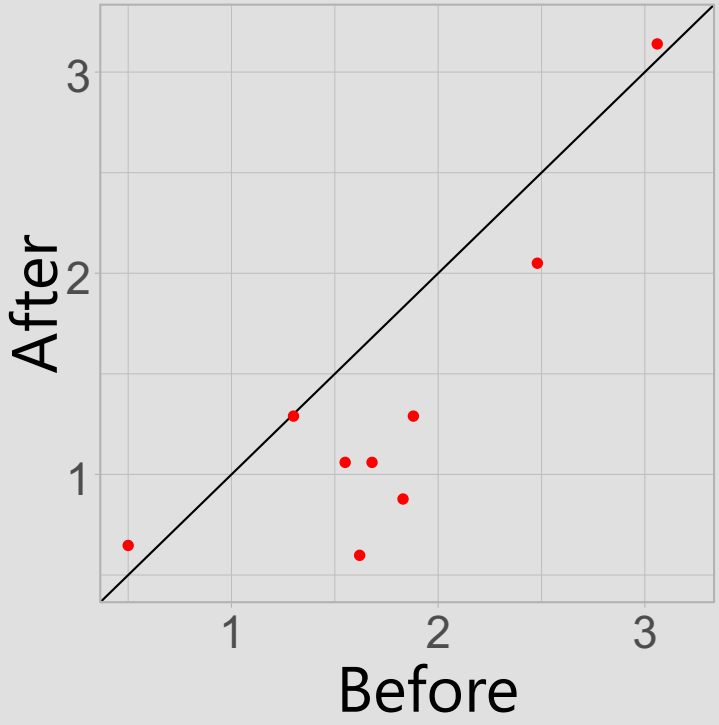
H_0: median diameter is 10 mm
H_1: median diameter is not 10 mm

Signed rank test: $p = 0.0673$, median diameter is 10.25 mm (97.7% confidence interval for the median – $[9.8, 10.7]$ mm.

$$P\left(X_{(r)} \leq \operatorname{med} X \leq X_{(n-r+1)}\right) = \frac{1}{2^n} \sum_{i=r}^{n-r+1} C_n^i$$

Example: depression therapy

Hamilton depression scale factor measurements in 9 patients with mixed anxiety and depression, taken at the first and second visit after initiation of a therapy (administration of a tranquilizer).



Does the therapy work?

Hollander, M., Wolfie, D.A. Nonparametric Statistical Methods

Signed rank test, paired samples

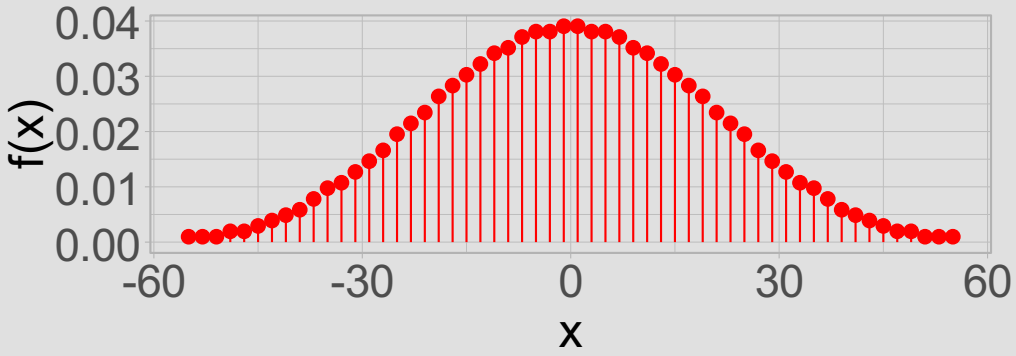
samples: $X_1^n = (X_{11}, \ldots, X_{1n})$
$X_2^n = (X_{21}, \ldots, X_{2n})$
$X_{1i} \neq X_{2i}$

null hypothesis: $H_0: \mathrm{med}(X_1 - X_2) = 0$

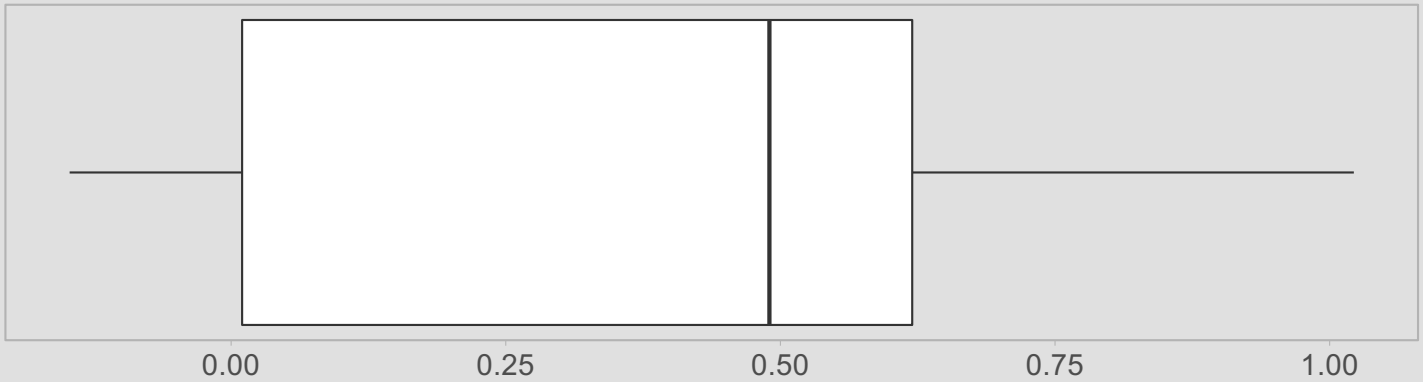
alternative hypothesis: $H_1: \mathrm{med}(X_1 - X_2) <\neq> 0$

statistic: $W = \sum_{i=1}^{n} \mathrm{rank}(|X_{1i} - X_{2i}|)\ \mathrm{sign}(X_{1i} - X_{2i})$

null distribution: tabulated



Example: depression therapy



H_0: depression scale factor did not change
H_1: depression scale factor changed

Signed rank test: $p = 0.0391$, median change – 0.49 points (96.1% confidence interval for the median change – $[-0.08, 0.952]$ points).

Example: wild and farmed fish

An experiment is designed to see if farmed fish exhibit a lower protein content than wild fish caught in the open sea. Two samples of healthy fish, similar in terms of age, gender, weight, etc., were drawn from the respective populations.

Are the populations similar?

Mann-Whitney test

samples: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$

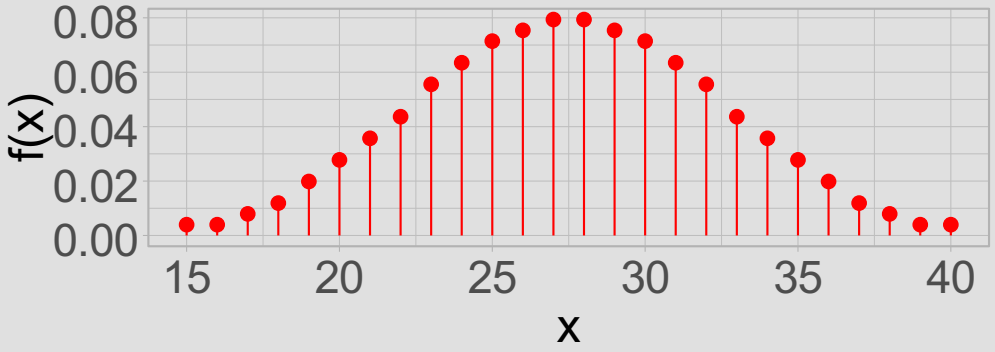
null hypothesis: $H_0: F_{X_1}(x) = F_{X_2}(x)$

alternative hypothesis: $H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$

$X_{(1)} \leq \cdots \leq X_{(n_1+n_2)}$ – pooled and sorted samples

statistic: $R = \sum_{i=1}^{n_1} \text{rank}(X_{1i})$

null distribution: tabulated



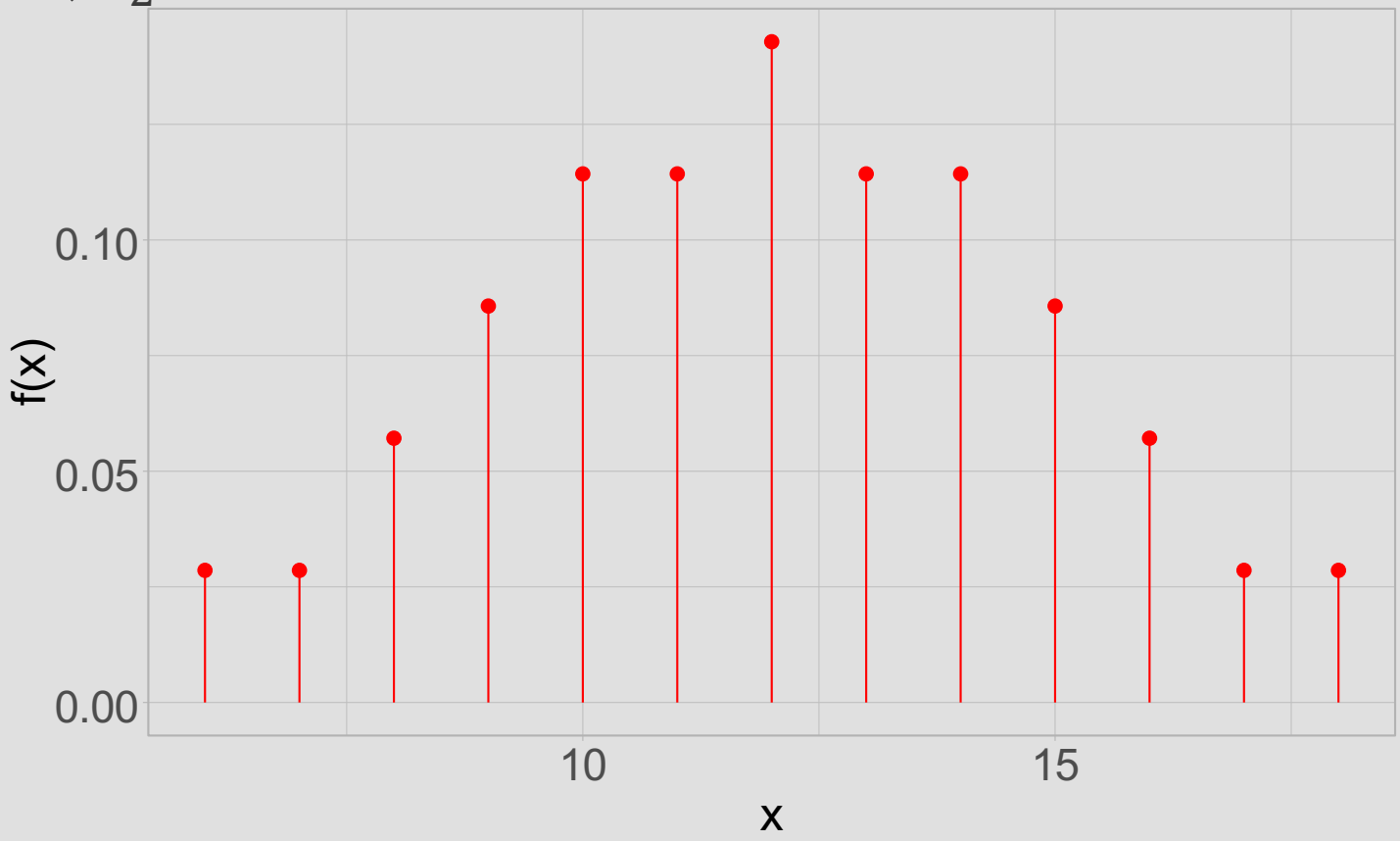
Null distribution

X_1	X_2	R_1
{1,2,3}	{4,5,6,7}	6
{1,2,4}	{3,5,6,7}	7
{1,2,5}	{3,4,6,7}	8
{1,2,6}	{3,4,5,7}	9
{1,2,7}	{3,4,5,6}	10
	...	
{3,6,7}	{1,2,4,5}	16
{4,5,6}	{1,2,3,7}	15
{4,5,7}	{1,2,3,6}	16
{4,6,7}	{1,2,3,5}	17
{5,6,7}	{1,2,3,4}	18

$\binom{n_1+n_2}{n_1}$ combinations in total

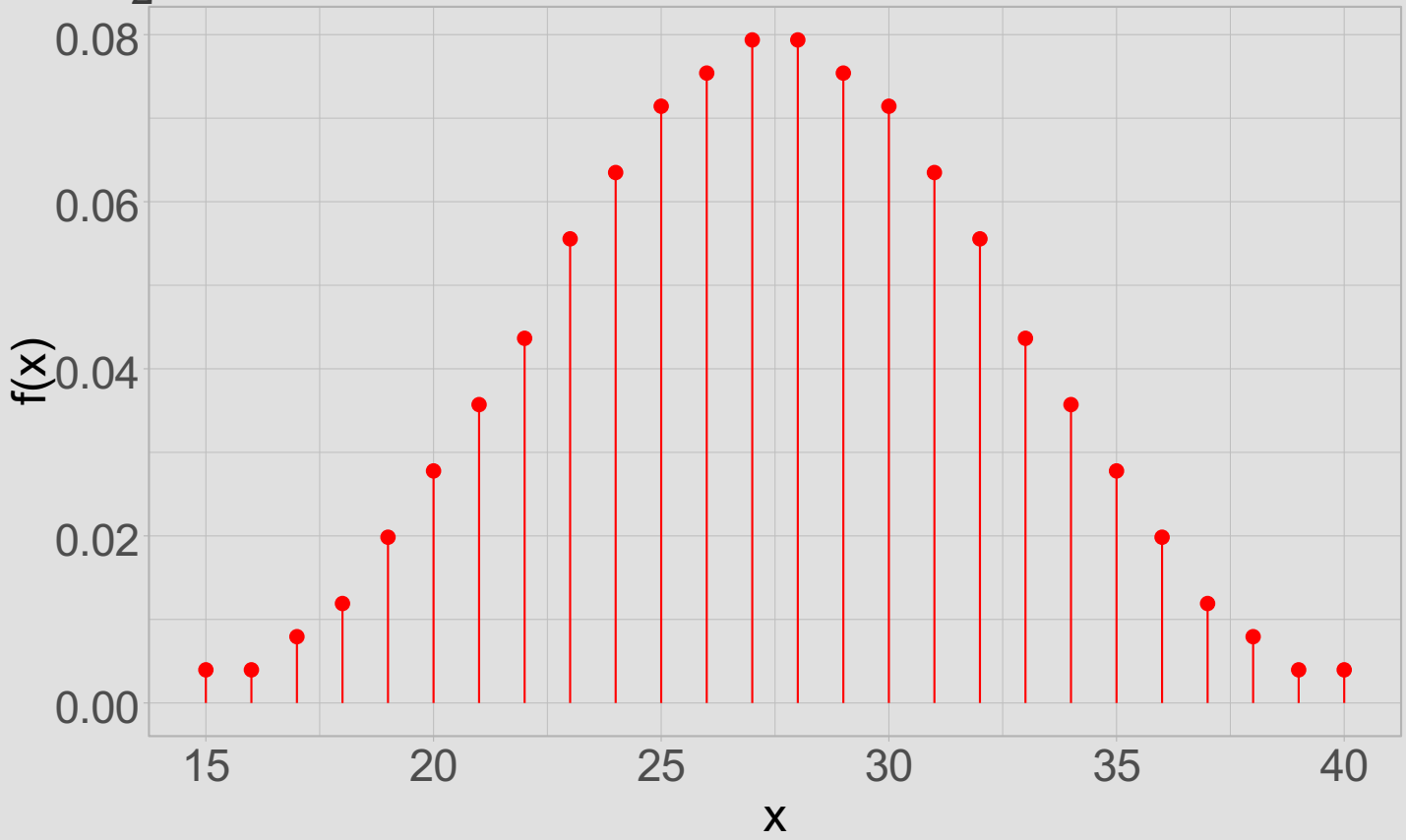
Null distribution

$n_1 = 3, n_2 = 4$:



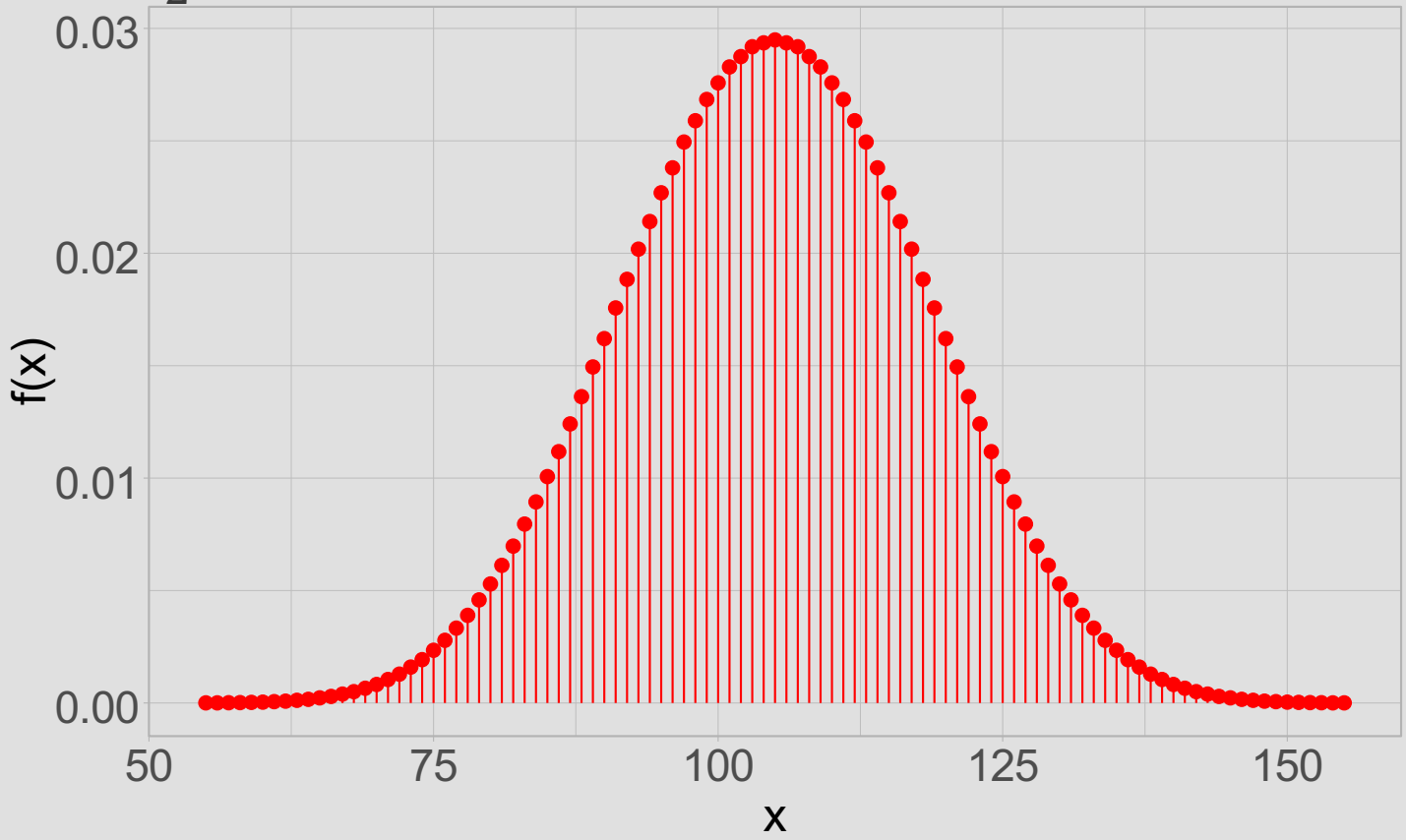
Null distribution

$n_1 = 5, n_2 = 5$:



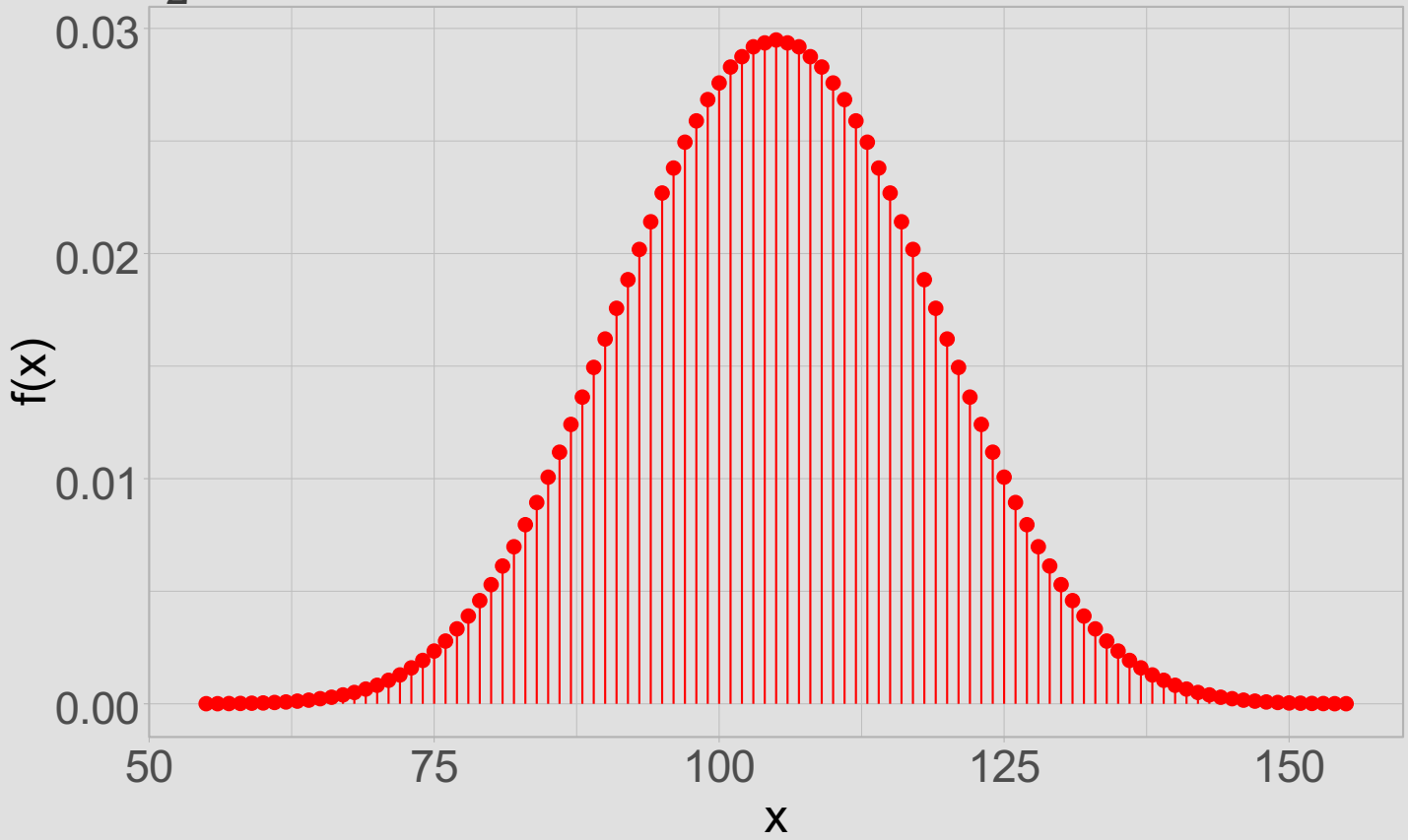
Null distribution

$n_1 = 10, n_2 = 10$:



Null distribution

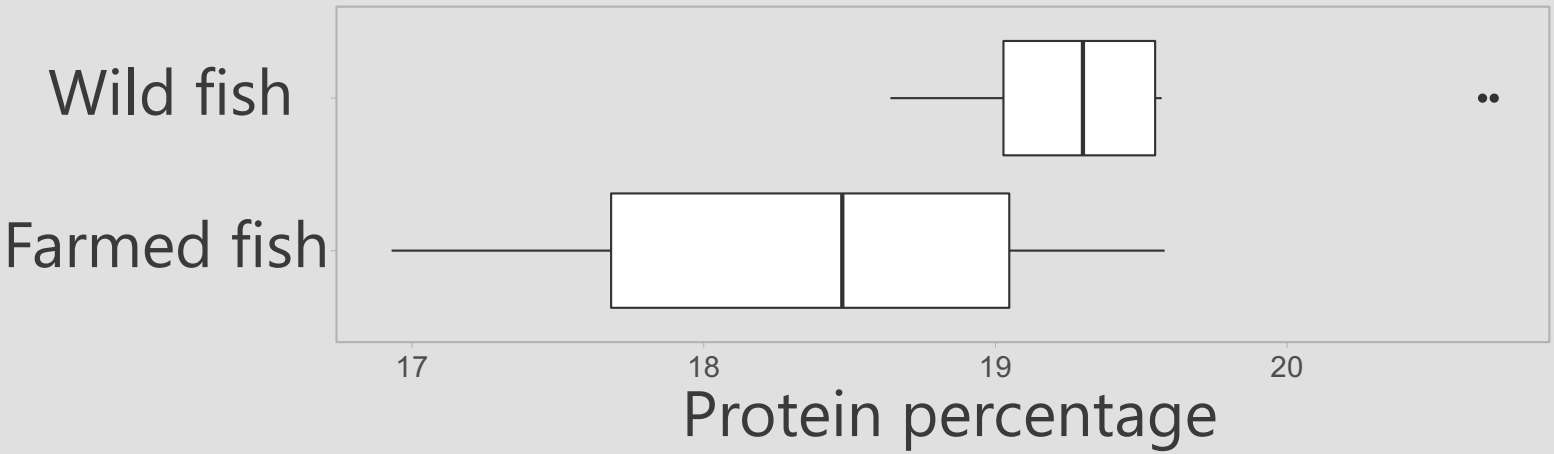
$n_1 = 10, n_2 = 10$:



Approximation for $n_1, n_2 > 10$:

$$R_1 \sim \approx N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

Example: wild and farmed fish



H_0: average protein percentages are identical
H_1: average protein percentages are different

Mann-Whitney test: $p = 0.0093$, median difference – 1% points (95% confidence interval for the median difference – $[0.2, 1.8]\%$).

Takeaways about rank tests

- Samples from unspecified distributions could be turned into ranks to test hypotheses about averages
- These methods keep more information originally contained in the sample than sign tests, but it comes with the price of added assumptions
- Actual hypotheses look a bit funky, but they are still hypotheses about averages!
- Rank tests for hypotheses about averages in one and two samples

Permutation tests

Idea

Rank tests:

1. Transform observations to ranks
2. Make assumptions
3. Use permutations that are equally likely under H_0 $\Longrightarrow$ generate null distributions

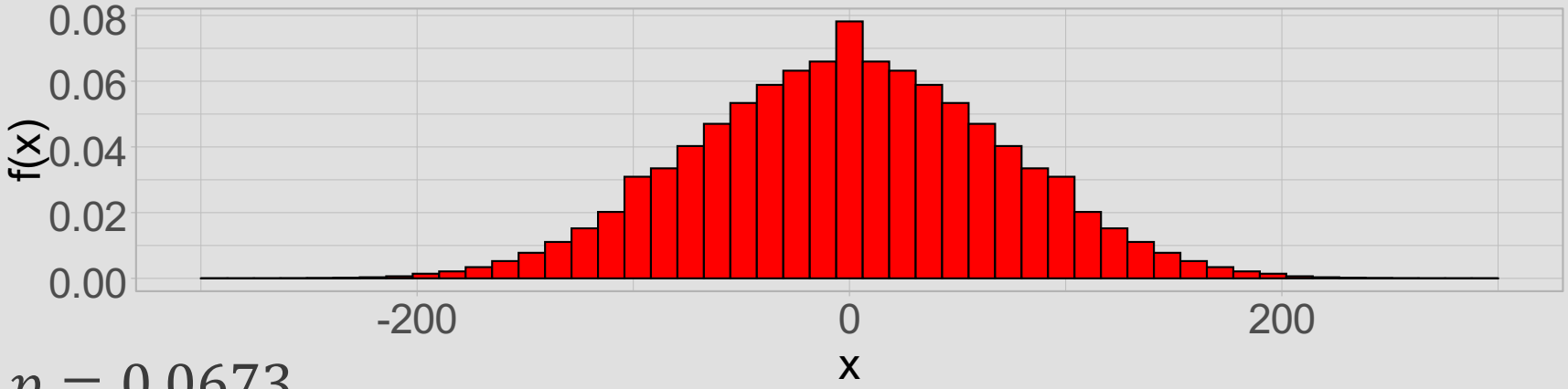
What if we skip 1?

Permutation test, one sample

sample:	$X^n = (X_1, \dots, X_n), X_i \neq m_0,$ F_X is symmetric around the mean
null hypothesis:	$H_0: \mathbb{E}X = m_0$
alternative hypothesis:	$H_1: \mathbb{E}X <\neq> m_0$
statistic:	$T = \sum_{i=1}^{n} (X_i - m_0)$
null distribution:	induced by 2^n permutations of signs of $X_i - m_0$

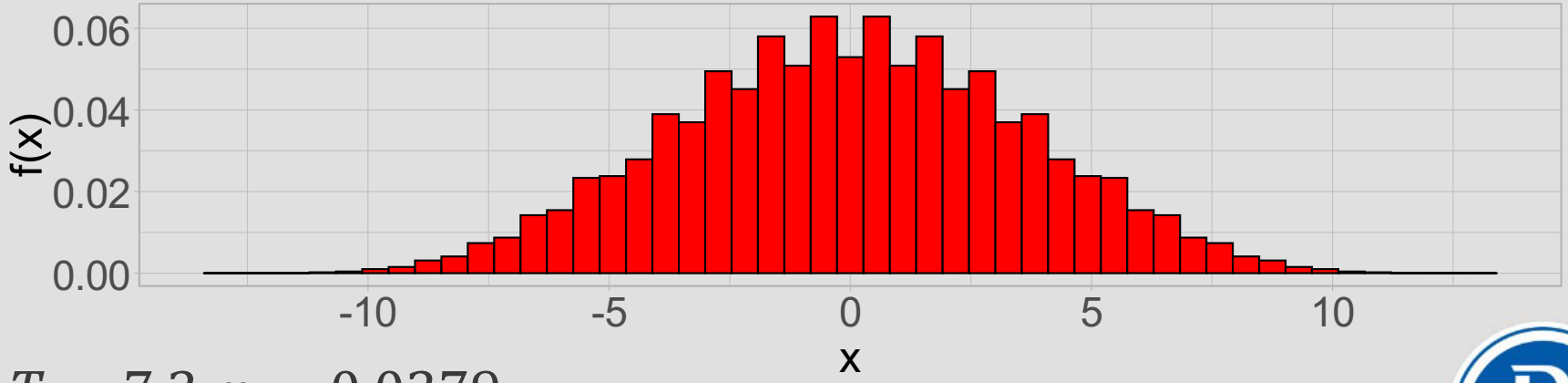
Example: washers' production

Signed rank test:



$p = 0.0673$

Permutation test:



$T = 7.3, p = 0.0379$

Permutation test, two paired samples

samples: $X_1^n = (X_{11}, \dots, X_{1n})$
$X_2^n = (X_{21}, \dots, X_{2n})$

null hypothesis: $H_0: \mathbb{E}(X_1 - X_2) = 0$

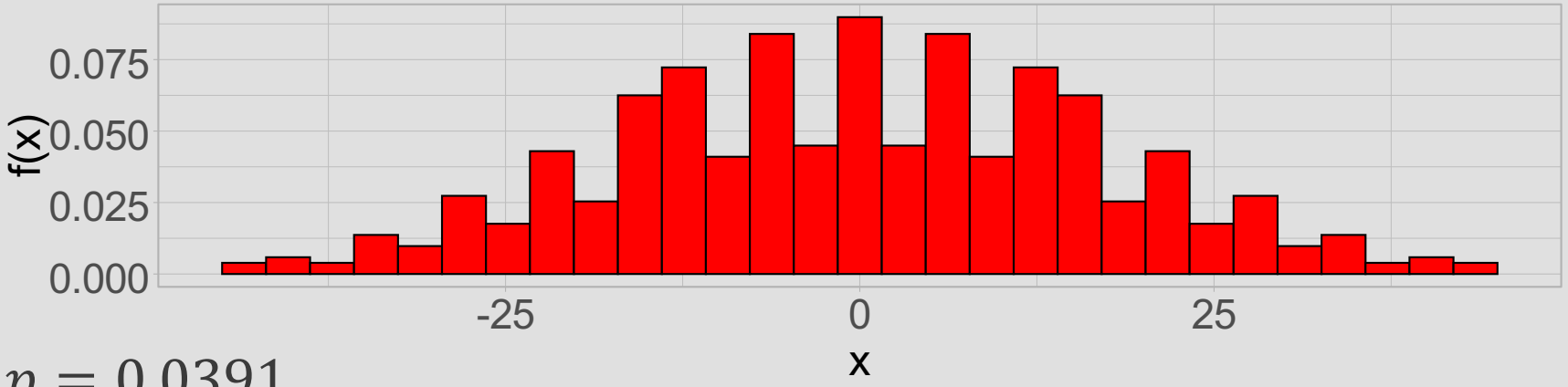
alternative hypothesis: $H_1: \mathbb{E}(X_1 - X_2) <\neq> 0$

statistic: $T = \sum_{i=1}^{n} D_i\,, \qquad D_i = X_{1i} - X_{2i}$

null distribution: induced by 2^n permutations of signs of D_i

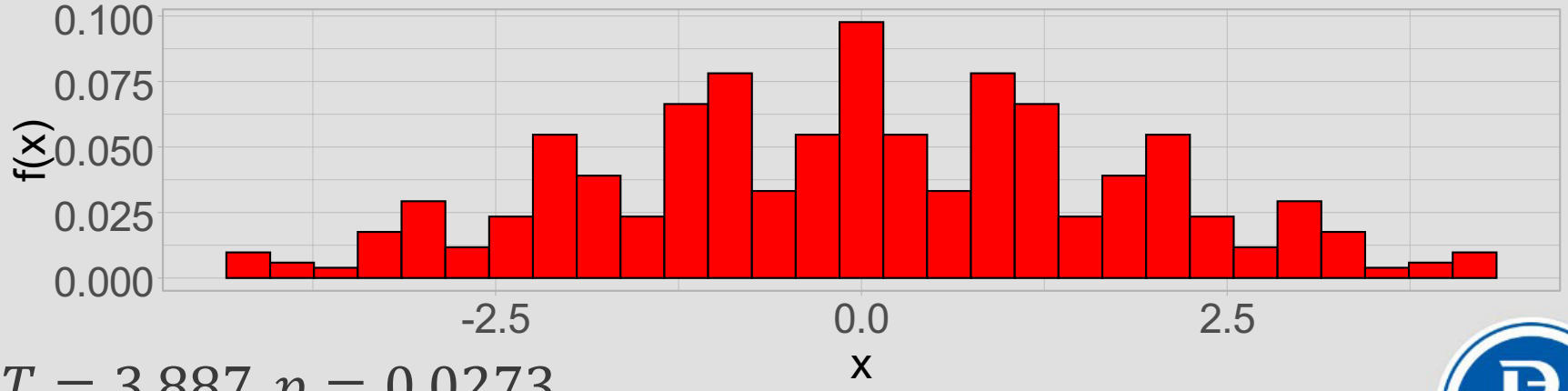
Example: depression therapy

Signed rank test:



$p = 0.0391$

Permutation test:



$T = 3.887, p = 0.0273$

Permutation test, two independent samples

samples: $X_1^n = (X_{11}, \dots, X_{1n})$
$X_2^n = (X_{21}, \dots, X_{2n})$

null hypothesis: $H_0: F_{X_1}(x) = F_{X_2}(x)$

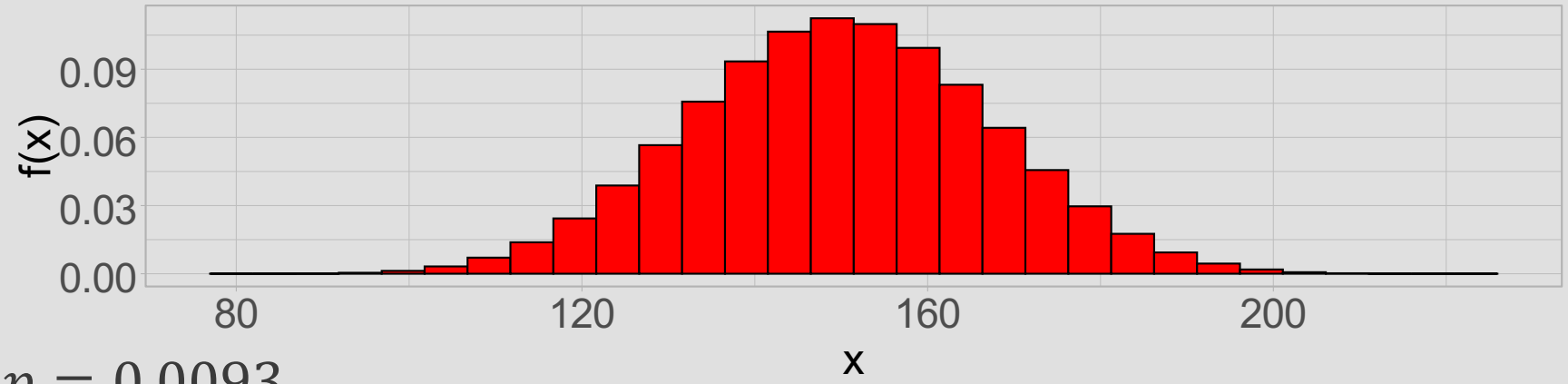
alternative hypothesis: $H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$

statistic: $T = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$

null distribution: induced by $\binom{n_1+n_2}{n_1}$ label assignments to the pooled sample

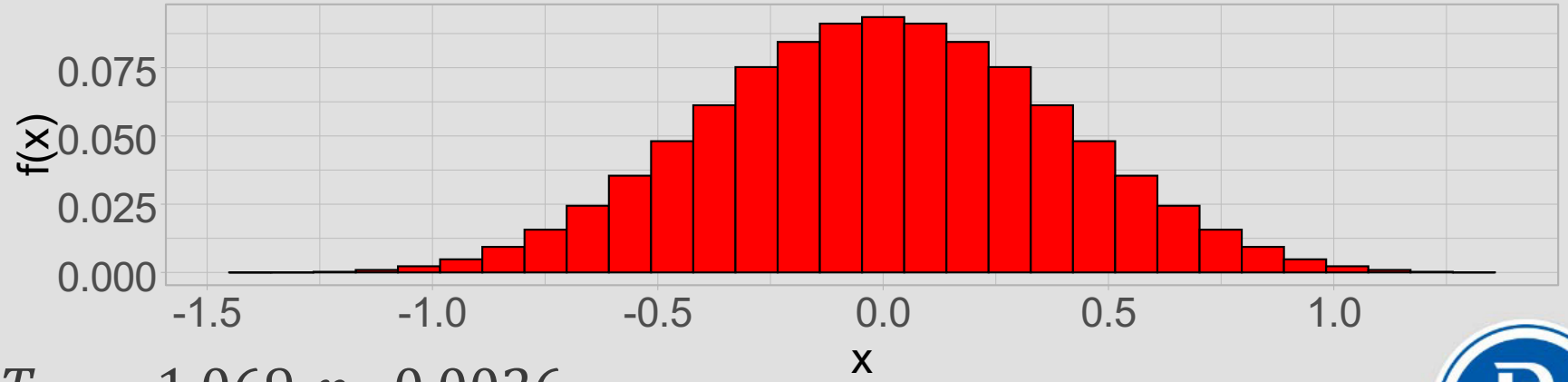
Example: wild and farmed fish

Mann-Whitney test:



$p = 0.0093$

Permutation test:



$T = -1.069, p=0.0026$

Permutation tests

- Different statistics could be used.

Some lead to the same p-value:

$$X^n, H_0: \mathbb{E}X = 0, H_1: \mathbb{E}X \neq 0$$

$T_1 = \sum_{i=1}^{n} X_i$ permutationally equivalent to $T_2 = \bar{X}$

But sometimes p-value depends on the choice of statistic:

$T_2 = \bar{X}$ permutationally not equivalent to $T_3 = \frac{\bar{X}}{S/\sqrt{n}}$

Rank tests are permutation tests with statistics based on ranks!

Permutation tests

- If the set G of all possible permutations is too large, a random subset G' could be used. Standard error of the p-value estimate then would be approximately $\sqrt{\frac{p(1-p)}{|G'|}}$

Takeaways about permutation tests

- In the same assumptions as rank tests require, you could take into account more information from data by using permutation tests

- Computationally expensive, but who cares!

Bootstrap tests

Permutation tests vs. bootstrap

Permutation tests:

1. Samples, statistic
2. Assumption about population distribution
3. Permutations $\Longrightarrow$ null distribution of the statistic

Permutation tests vs. bootstrap

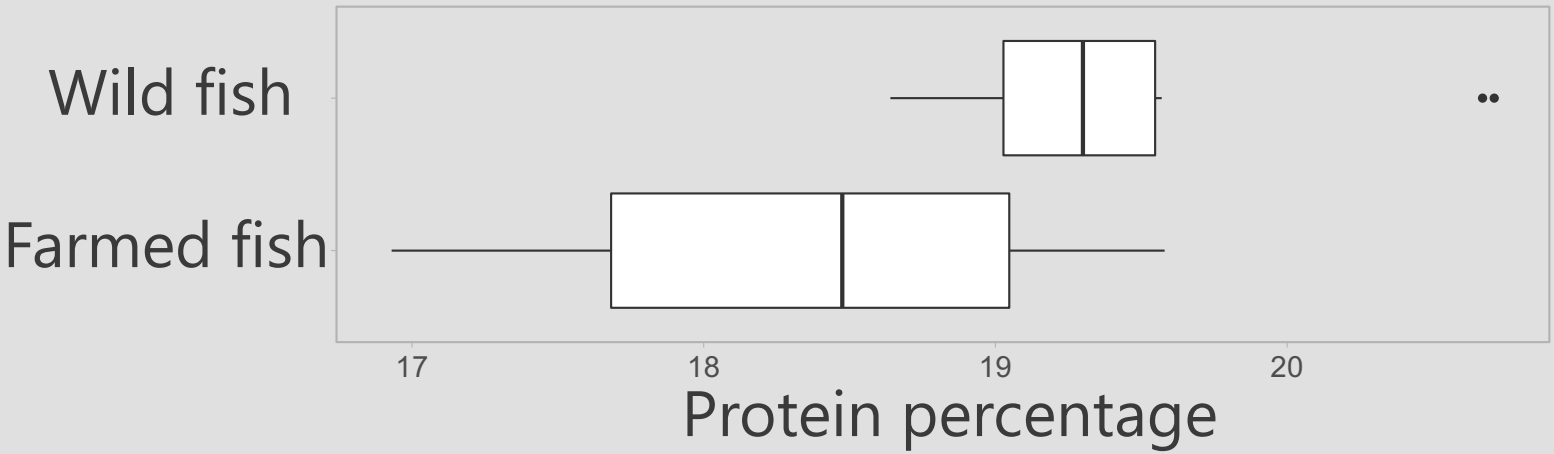
Permutation tests:

1. Samples, statistic
2. Assumption about population distribution
3. Permutations $\Longrightarrow$ null distribution of the statistic

Bootstrap confidence intervals:

1. Samples, statistic estimating the parameter
2. Bootstrap resamples $\Longrightarrow$ approximate sampling distribution of the statistic

Example: wild and farmed fish

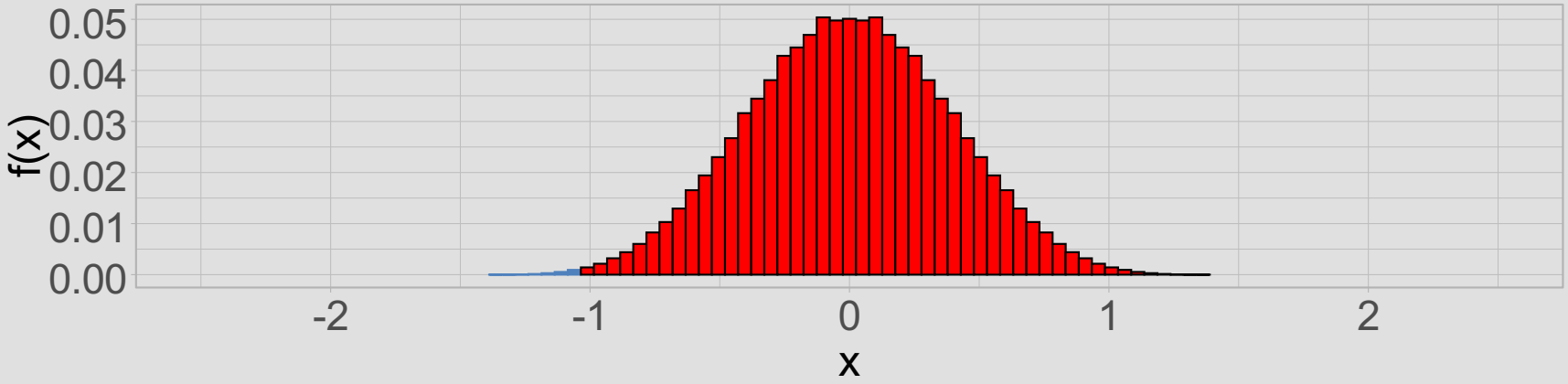


H_0: average protein percentages are identical
H_1: average protein percentage in farmed fish is lower

$$T = \bar{X}_1 - \bar{X}_2 = -1.069$$

Example: wild and farmed fish

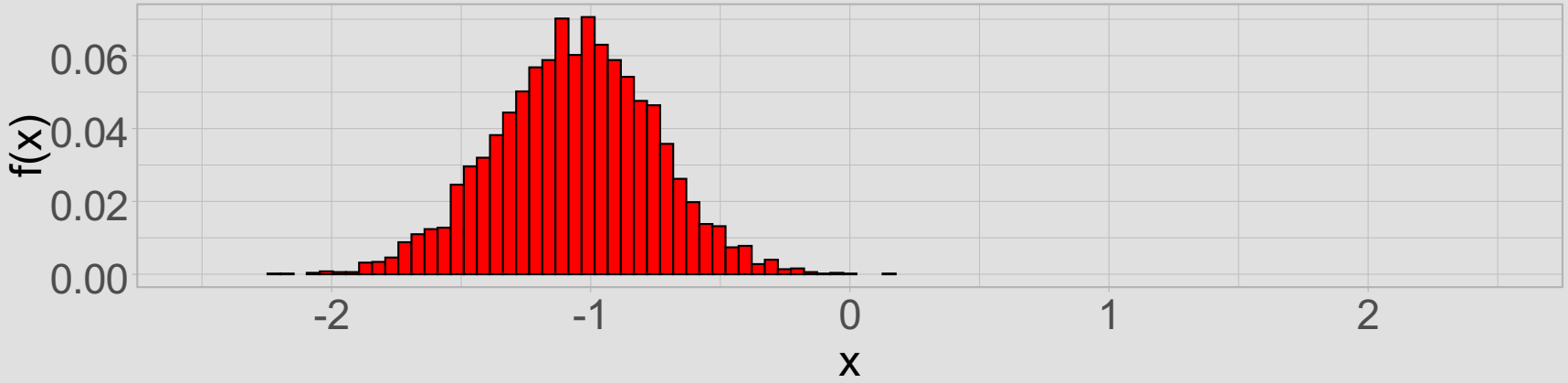
Null distribution of permutation test with $T = \bar{X}_1 - \bar{X}_2$:



Proportion of permutations with $T \leq -1.069 - 0.0013$

Example: wild and farmed fish

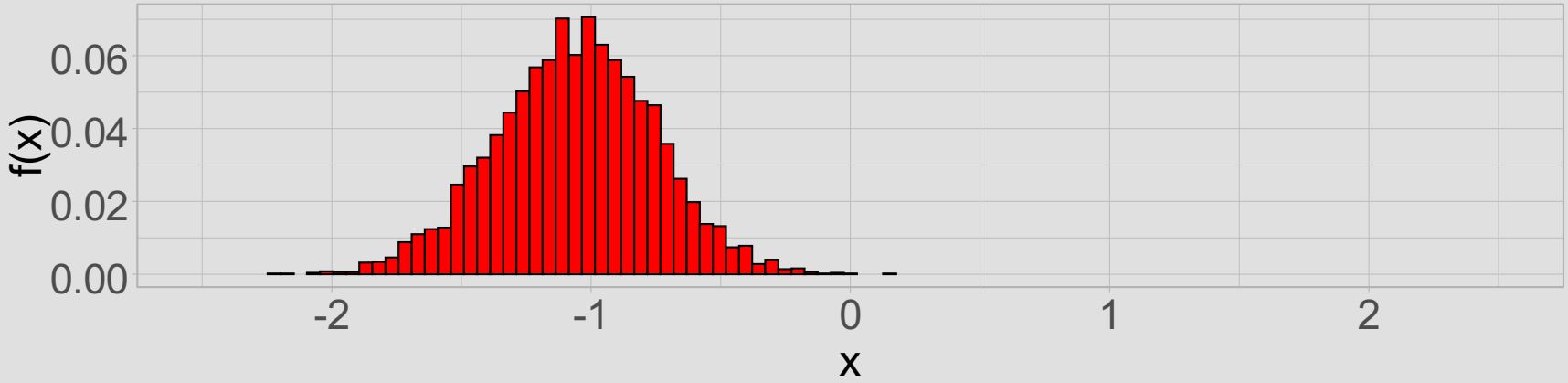
Bootstrap distribution of $\bar{X}_1 - \bar{X}_2$:



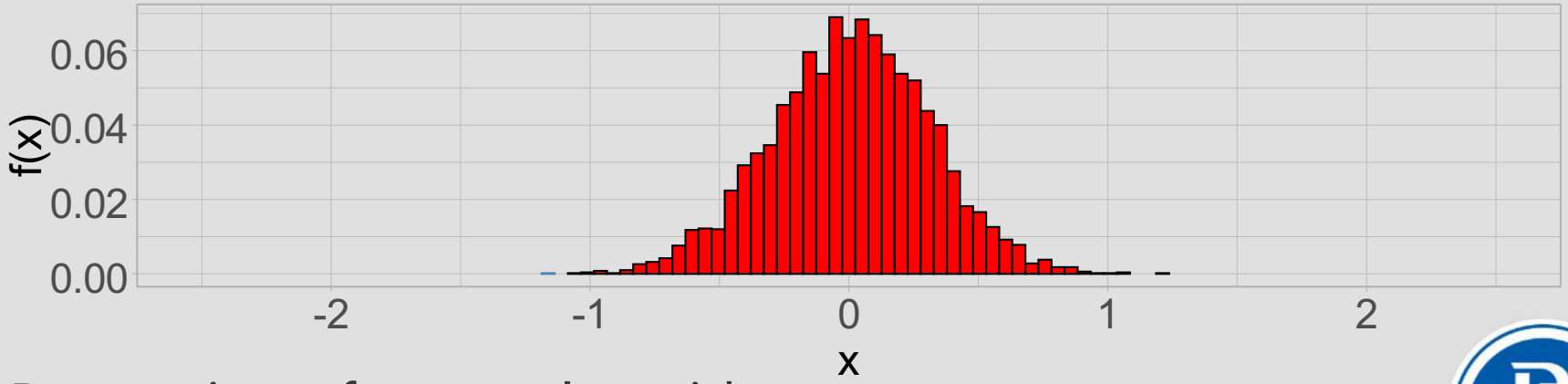
This is not a null distribution – it's not constructed under the null!

Example: wild and farmed fish

Bootstrap distribution of $\bar{X}_1 - \bar{X}_2$:



This is not a null distribution – it's not constructed under the null! To construct null distribution, we could subtract T and 0:



Proportion of resamples with $T \leq -1.069 - 0.0002$.

Permutation tests vs. bootstrap

- Permutation test is exact
- Bootstrap test is approximate

- Permutation test tests $H_0: F_{X_1}(x) = F_{X_2}(x)$ against $H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta < 0$
- Bootstrap test tests $H_0: \mathbb{E}X_1 = \mathbb{E}X_2$ against $H_1: \mathbb{E}X_1 < \mathbb{E}X_2$

Takeaways about bootstrap tests

- Always approximate
- Might be difficult to specify – not always obvious how to construct null distribution
- When work, allow to ditch assumptions about distributions

Independence for categorical variables

Example: marital status and depression

From General Social Survey 2018:

		Depression		
		Yes	**No**	**No answer**
Marital status	**Never married**	103	339	9
	Married	81	548	19
	Divorced/ Separated	69	220	5
	Widowed	18	36	0

Are these two factors independent?

https://gssdataexplorer.norc.org/

Contingency table

Paired samples $X_1^n = (X_{11}, \dots, X_{1n})$, $X_2^n = (X_{21}, \dots, X_{2n})$
$X_1 \in \{1, \dots, K_1\}, X_2 \in \{1, \dots, K_2\}$

		X_2					
		1	**…**	j	**…**	K_2	**Total**
X_1	**1**						
	…						
	i			n_{ij}			n_{i+}
	…						
	K_1						
	Total			n_{+j}			n

https://gssdataexplorer.norc.org/

Chi-squared test for independence

samples: $X_1^n = (X_{11}, \dots, X_{1n})$
$X_2^n = (X_{21}, \dots, X_{2n})$

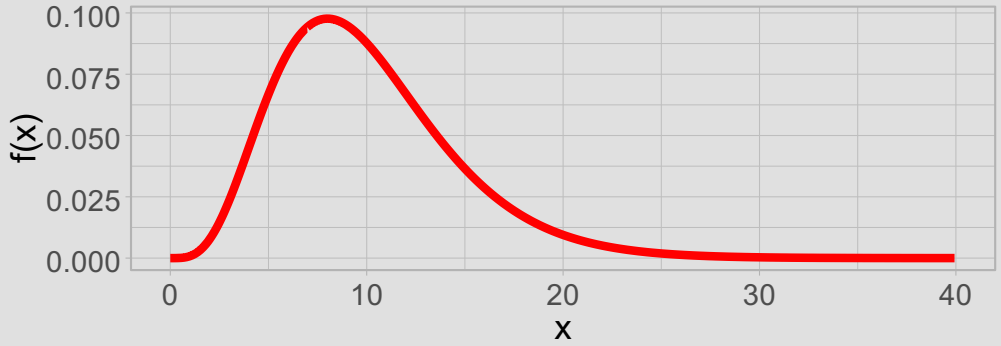
null hypothesis: H_0: X_1 and X_2 are independent

alternative hypothesis: H_1: H_0 is false

statistic: $$\chi^2 = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

$$e_{ij} = \frac{n_{i+} n_{+j}}{n}$$

null distribution: $\chi^2_{(K_1-1)(K_2-1)}$



Example: marital status and depression

		Depression		
		Yes	**No**	**No answer**
Marital status	**Never married**	103	339	9
	Married	81	548	19
	Divorced/ Separated	69	220	5
	Widowed	18	36	0

H_0: marital and depression statuses are independent
H_1: marital and depression statuses are not independent

Chi squared test: $p = 3.8 \times 10^{-6}$

Example: religious preference and political views

	Extremely liberal	Liberal	Slightly liberal	Moderate	Slightly conservative	Conservative	Extremely conservative	Don't know
Buddhism	3	5	4	4	0	3	0	0
Catholic	8	38	56	203	70	73	20	21
Christian	2	2	2	16	4	8	1	0
Hinduism	0	2	1	5	0	0	0	0
Jewish	4	11	2	11	2	7	1	1
Islam	0	2	2	8	0	3	0	1
Protestant	49	97	111	395	164	219	66	30
Other	0	9	3	18	3	1	0	1
None	56	109	74	189	40	38	9	22

https://gssdataexplorer.norc.org/

Other options?

Requires $e_{ij} > 5$ in 80% of cells, and no cells with $e_{ij} = 0$

What do we do for sparse tables?

Other options?

Requires $e_{ij} > 5$ in 80% of cells, and no cells with $e_{ij} = 0$

What do we do for sparse tables?

		X_2				
		1	**…**	***j***	**…**	K_2
X_1	**1**					
	…					
	i			n_{ij}		
	…					
	K_1					

Observation	Row	Column
1		
2	i	j
…	…	…
n		

Permutation test for independence

samples: $X_1^n = (X_{11}, \dots, X_{1n})$
$X_2^n = (X_{21}, \dots, X_{2n})$

null hypothesis: H_0: X_1 and X_2 are independent

alternative hypothesis: H_1: H_0 is false

statistic: $$\chi^2 = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

null distribution: induced by n! permutations of the last column in the tall table

Example: religious preference and political views

H_0: religious preference and political views are independent
H_1: religious preference and political views are not independent

Permutation test: $p = 1 \times 10^{-5}$

Following up independence test

Standardized Pearson residuals:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}\left(1 - \frac{n_{i+}}{n}\right)\left(1 - \frac{n_{j+}}{n}\right)}}$$

They are approximately standard normal, so values >3 or <-3 are unexpectedly large deviations

For deviating cells, you could test hypotheses about corresponding proportions, but the results are only exploratory because those hypotheses were defined after seeing the data

Example: marital status and depression

		Depression		
		Yes	No	No answer
Marital status	Never married	2.7	-2.4	-0.5
	Married	-5.5	4.68	1.5
	Divorced/ Separated	2.3	-1.96	2.3
	Widowed	2.8	-2.3	-1.1

12.5% of all married respondents were diagnosed with depression vs. 23.8% of non-married – a difference of 11.3% (Wilson's 95% confidence interval for the difference – $[7, 15]$%, p-value from inverting it – $p = 4.6 \times 10^{-8}$).

Example: religious preference and political views

	Extremely liberal	Liberal	Slightly liberal	Moderate	Slightly conservative	Conservative	Extremely conservative	Don't know
Buddhism	2.1	1.9	1.4	-1.4	-1.6	0.1	-0.9	-0.8
Catholic	-4.1	-3.2	0.3	2.5	1.6	-0.2	-0.1	1.4
Christian	0.1	-1.1	-1.0	1.1	-0.2	1.3	-0.4	-1.1
Hinduism	-0.7	1.1	0.1	1.5	-1.1	-1.2	-0.6	-0.5
Jewish	1.4	3.2	-1.2	-1.1	-1.4	0.5	-0.5	-0.3
Islam	-0.9	0.1	0.2	1.1	-1.5	0.4	-0.8	0.7
Protestant	-2.0	-4.8	-1.8	-1.8	3.2	5.4	3.8	-1.7
Other	-1.4	2.5	-0.5	1.8	-0.7	-2.1	-1.2	-0.1
None	6.1	6.9	2.3	-0.9	-3.9	-6.0	-3.3	1.2

Example: religious preference and political views

	Extremely liberal	Liberal	Slightly conservative	Conservative	Extremely conservative
Catholic	-4.1	-3.2	1.6	-0.2	-0.1
Protestant	-2.0	-4.8	3.2	5.4	3.8
None	6.1	6.9	-3.9	-6.0	-3.3

- 45% of non-religious people are liberal vs. 23% of religious ($p < 2\times10^{-16}$, 95% CI for the difference – $[17, 26]$%)
- 40% of protestants are conservative vs. 24% of all others ($p < 2\times10^{-16}$, 95% CI for the difference – $[12, 19]$%)
- 21% of catholics are liberal vs. 30% of all others ($p = 4.5\times10^{-5}$, 95% CI for the difference – $[5, 13]$%)

Takeaways about independence for categorical variables

- Tested with contingency tables
- Most common test – chi-squared, but the table must be quite dense
- Rejection of independence is not very interesting – there need to be follow-ups
- Hypotheses suggested by follow-up results ideally should be tested on different data

How to choose the best test for your problem

Samples

- One: a question about population that the sample comes from

Samples

- One: a question about population that the sample comes from
- Two: we want to compare populations that produced the samples

Samples

- One: a question about population that the sample comes from
- Two: we want to compare populations that produced the samples
- More than two:
  - we want an overall comparison of several populations
  - we might want to follow up with with pairwise comparisons to locate the differences

Samples

- One: a question about population that the sample comes from
- Two: we want to compare populations that produced the samples
- More than two:
  - we want an overall comparison of several populations
  - we might want to follow up with with pairwise comparisons to locate the differences

In case of several samples: are they independent or paired?

Question

What do we care about?

- Distribution as a whole
- Certain parameter of the distribution

Question

What do we care about?

- Distribution as a whole
- Certain parameter of the distribution:
  - An average – mean, median

Question

What do we care about?

- Distribution as a whole
- Certain parameter of the distribution:
  - An average – mean, median
  - Spread: variance, something else

Question

What do we care about?

- Distribution as a whole
- Certain parameter of the distribution:
  - An average – mean, median
  - Spread: variance, something else
  - Something else

Measurement scale

- Binary – variables are 0s and 1s (or could be encoded with 0s and 1s). We are dealing with Bernoulli distribution, and our question is about proportions!

Measurement scale

- Binary – variables are 0s and 1s (or could be encoded with 0s and 1s). We are dealing with Bernoulli distribution, and our question is about proportions!
- Categorical – variables have finite number of incomparable values. We are dealing with multinomial distributions!

Measurement scale

- Binary – variables are 0s and 1s (or could be encoded with 0s and 1s). We are dealing with Bernoulli distribution, and our question is about proportions!
- Categorical – variables have finite number of incomparable values. We are dealing with multinomial distributions!
- Continuous: variables have infinite number of possible numeric values.

Measurement scale

- Binary – variables are 0s and 1s (or could be encoded with 0s and 1s). We are dealing with Bernoulli distribution, and our question is about proportions!
- Categorical – variables have finite number of incomparable values. We are dealing with multinomial distributions!
- Continuous: variables have infinite number of possible numeric values.
  - Does it make sense to assume that data comes from a particular distribution (normal, Poisson)?

Measurement scale

- Binary – variables are 0s and 1s (or could be encoded with 0s and 1s). We are dealing with Bernoulli distribution, and our question is about proportions!
- Categorical – variables have finite number of incomparable values. We are dealing with multinomial distributions!
- Continuous: variables have infinite number of possible numeric values.
  - Does it make sense to assume that data comes from a particular distribution (normal, Poisson)?
  - How likely is to see same values of variables in the sample (due to rounding or measurement error)?

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

- Sample: one
- Question: about mean
- Measurement scale: continuous

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Candidates:

- t-test
- sign test
- signed rank test
- one sample permutation test for the mean

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Candidates:

- t-test:
  - Is the distribution of the variable very skewed? If yes, CLT might not work well
- sign test
- signed rank test
- one sample permutation test for the mean

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Candidates:

- t-test
- sign test:
  - Do we only care about the increase of the average, no matter the absolute or relative scale of the increase?
- signed rank test
- one sample permutation test for the mean

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Candidates:

- t-test
- sign test
- signed rank test:
  - Are there many identical observation in the sample?
  - Are we fine with using median instead of the mean?
  - Is the distribution of the variable symmetric around the median?
  - Do we care more about the increase of the average itself than about the absolute scale of the increase?
- one sample permutation test for the mean

Example: one sample, question about mean

Data: a sample of sales by 50 salesmen after a training was conducted. Before training, average sales were $100 per transaction; is the average still the same?

Candidates:

- t-test
- sign test
- signed rank test
- one sample permutation test for the mean:
  - Is the distribution of the variable symmetric around the mean?

Independence

For two paired samples, the question of independence might be answered by different tests depending on the scale of variables

Independence

For two paired samples, the question of independence might be answered by different tests depending on the scale of variables

- Two binary variables: better to test the difference in proportions

Independence

For two paired samples, the question of independence might be answered by different tests depending on the scale of variables

- Two binary variables: better to test the difference in proportions
- One binary, one continuous: better to test the difference in means in two subgroups

Independence

For two paired samples, the question of independence might be answered by different tests depending on the scale of variables

- Two binary variables: better to test the difference in proportions
- One binary, one continuous: better to test the difference in means in two subgroups
- One binary, one categorical; two categorical: test on contingency table

Independence

For two paired samples, the question of independence might be answered by different tests depending on the scale of variables

- Two binary variables: better to test the difference in proportions
- One binary, one continuous: better to test the difference in means in two subgroups
- One binary, one categorical; two categorical: test on contingency table
- Two continuous: Pearson or Spearman correlation and the corresponding test

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

- Samples: two independent
- Possible questions:
  - equality of means
  - equality of distributions
- Measurement scale: continuous

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Candidates:

- t-test
- Mann-Whitney test
- two sample permutation test for independent means
- two sample Anderson-Darling test

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Candidates:

- t-test:
  - Are distributions of variables in both samples very skewed?
  - Does the smaller group have much higher variance?
- Mann-Whitney test
- two sample permutation test for independent means
- two sample Anderson-Darling test

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Candidates:

- t-test
- Mann-Whitney test:
  - How safe is to assume that the distributions of lifespans are only possibly differ by shift?
- two sample permutation test for independent means
- two sample Anderson-Darling test

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Candidates:

- t-test
- Mann-Whitney test
- two sample permutation test for independent means:
  - How safe is to assume that the distributions of lifespans are only possibly differ by shift?
  - Should variances be taken into account in the test statistic?
- two sample Anderson-Darling test

Example: diet and lifespan

Data: lifespan of 195 lab rats, 106 were randomly assigned to restricted diet, 89 were able to eat as much as they wanted. Does lifespan depend on diet?

Candidates:

- t-test
- Mann-Whitney test
- two sample permutation test for independent means
- two sample Anderson-Darling test:
  - Do we care about fine differences in distributions? If we find the difference, then what?

Takeaways

Important questions when selecting a test:

- How many samples?
- If more then one, are they independent?
- What is the question about (averages, spreads, distributions, something else)?
- What scale are the measurements in?

If you have several candidate tests:

- What are their assumptions? Are they likely to hold in this problem?
- For candidate tests that are likely correct: which one should have higher power?