

## Index of 8-L

| Page | Title  |
|------|--|
| 1    | Practical information                                  |
| 2    | Mid-term practical info                                |
| 3    | Nonparametric (distribution-free) methods              |
| 4    | 1 sample: Sign test                                    |
| 5    | McNemar's test   |
| 6    | Ranks  |
| 7    | 2 samples: Wilcoxon–Mann–Whitney test                  |
| 8    | Example for 2-sample W-M-W test                        |
| 9    | 1 sample: Wilcoxon's signed rank test                  |
| 10   | Example for 1-sample Wilcoxon test                     |
| 11   | Statistical methods to choose sample size              |
| 12   | Exercise 6.19  |
| 13   | Controlling size of confidence intervals               |
| 14   | Controlling margin of error for a single proportion    |
| 15   | Sample size based on estimation precision              |
| 16   | Errors of type I–II and power                          |
| 17   | Sample size based on power                             |
| 18   | Sample size misconceptions,<br>and equivalence testing |
| 19   | Summary notes  |

## PRACTICAL INFORMATION

### Major news:

- lab review scheduled for this afternoon (1-2pm),
- 2nd home assignment: are you making good progress?  
— still time to ask questions ... (due Tuesday 22/10),
- optional: if you want to use own data for last home assignment, you should start preparing the data and project outline (due 7/11), see project guidelines,
- information about mid-term exam:
  - \* October 23, 10:30-11:30am, AVC Lecture Theatre B,
  - \* new course syllabus page posted (for mid-term),
  - \* see (mid-term) exams from prev. years for examples.

### Today's lecture:

- nonparametric methods for one and two (continuous) samples: *sign test* and *rank tests*;<sup>1 2</sup>
- sample size calculations (Suppl. notes, Section 1),
  - \* based on precision, e.g. margin of error for CIs,<sup>3</sup>
  - \* based on *power* of a statistical test,<sup>2 4</sup>
- McNemar's test (exact version): not in course curriculum,
- review: Summary Worksheet 8:7.

<sup>1</sup> PSLS Supplementary Chapter 27 on rank tests, not including the sign test which can be understood as a binomial test.

<sup>2</sup> For rank tests and power calculations, we only use statistical software.

<sup>3</sup> PSLS Chapters 15 & 19; S Sections 7.2-7.3; IPS Sections 6.1 & 8.1.

<sup>4</sup> PSLS Chapter 15; S Section 8.1; IPS Sections 6.4 and 7.1-7.2.

## MID-TERM EXAM

- mark: *optional* for 15% of course mark; that is, you decide *after you have received your mark* if you want to use it,
- all aids (books and notes) are allowed,  
— except a computer or computer-like device (tablet or smartphone),
- duration: 1 hour sharp,
- note: you must bring a calculator and statistical tables.<sup>5</sup>

1 question/assignment with possible types of (sub)questions:

- choice of statistical model and analysis —
  - \* carry out analysis when calculations manageable (see below),
  - \* or base analysis on Minitab print + extra calculations,
  - \* or outline analysis if calculations not manageable,
- probability calculations manageable (see below),
- multiple choice (one or several correct answers).

Calculations manageable by hand (calculator):

- simple probabilities (e.g.,  $1-p$ ,  $(1-p)^n$ , simple binomial),
- probabilities in normal distribution (incl. standardization),
- normal approximation for binomial distribution,
- $z/t$ -tests and CIs (given estimates or calculated statistics),
- one-sample proportion, (note: two-sample proportion not included),
- no data entry into calculator or large summations.

---

<sup>5</sup> I will have a couple of extra tables and calculators, in case you forget.

## NONPARAMETRIC (DISTRIBUTION-FREE) METHODS

- no parametric statistical model (involving particular distribution type),
- still assumptions of i.i.d. samples and possibly of particular features of distributions,
- classical methods — only ones in this course!:
  - \* mostly based on ranks, that is, the relative magnitude of observations, where it does not matter how much  $X_2 > X_1$ , only that  $X_2 > X_1$ ,
  - \* analyses computable by hand (tedious for large data), but reference distributions require special tables or large-sample approximations,
  - \* all methods in the course available in Minitab/Stata/R,
  - \* advantages:  
no distribution assumptions, robust, “simple to use” . . .
  - \* disadvantages:  
some loss of information compared to good parametric model, problems with getting good estimates and confidence intervals (*what to estimate?*), not available beyond the very simplest designs,
- alternative: modern, computer-intensive methods:
  - \* resampling/permutation/bootstrap methods,<sup>6</sup>
  - \* very flexible and powerful, but *not* simple to use.

<sup>6</sup> Recommended by PSLS/IPS; described in IPS Supplementary Chapter 16.

## 1 SAMPLE: SIGN TEST

Example (Visual receptive fields, 7L-5):

- neural activity (# spikes/sec) at 9 recordings of both Spontaneous activity (SA) and Response (R),
- analyze differences (R-SA) (ordered values):  
-7.5 -2.5 12.5 13.3 14.2 16.7 26.7 34.2 44.2

Sign test for  $H_0$ : median = known value:

- Model:  $X_1, \dots, X_n$  i.i.d. from continuous distribution,
- Null hypothesis  $H_0$ : median = known value ( $m_0$ ),
- Test procedure:
  - \* test statistic:  $Y$  = number of  $X$ 's  $> m_0$ ,
  - \* disregard  $X$ 's =  $m_0$ , let  $n_1$  = number of  $X$ 's  $\neq m_0$ ,
  - \* under  $H_0$ :  $Y \sim B(n_1, 0.5)$   
 $\Rightarrow$  corresponds to testing  $H_0 : p=0.5$  in the binomial distribution  $B(n_1, p)$  for  $Y$ ,
  - \*  $P$ -values from the binomial distribution, e.g.  
 $H_a$ : median  $> m_0$  ( $\sim p > 0.5$ )  $\Rightarrow P = P(Y \geq Y_{\text{obs}})$ .
- Confidence interval for median using Minitab/Stata.

Example — testing  $H_0$ : median = 0 vs.  $H_a$ : median  $> 0$ :

- no differences = 0; out of 9 differences, 7 are  $> 0$ ,
- $P = P(Y \geq 7) = 0.090$ , where  $Y \sim B(9, 0.5)$ ,
- cannot reject  $H_0$ ; no evidence of higher activity for R.

|                |
|----------------|
| MCNEMAR'S TEST |
|----------------|

= sign test for paired binary data<sup>7</sup> (or paired proportions).

Example: Varicose veins and overweight:

- 122 pairs of brothers, one overweight and one normal weight, with records of presence or absence of varicose veins,

| Group     | var. veins |       | Normal weight    | Overweight       |                  |
|-----------|------------|-------|------------------|------------------|------------------|
|           | + (1)      | - (0) |                  | + var. veins (1) | - var. veins (0) |
| normal wt | 23         | 99    | + var. veins (1) | 19               | 4                |
| overwt    | 30         | 92    | - var. veins (0) | 11               | 88               |

- hypothesis of interest: same proportion of varicose veins among normal weight and overweight persons? — observed proportions:

normal weight:  $\hat{p} = 23/122 = 0.19$ ,

overweight:  $\hat{p} = 30/122 = 0.25$ .

Test procedure:

- code each “success” as 1, and each “failure” as 0,
- compute differences  $D_i$  (e.g. normal weight – overweight) within each pair  $i$ :
  - \*  $D_i = 0$ : same outcome (either 1 or 0) in both pair members,
  - \*  $D_i = 1$ : success in first pair member, failure in second,
  - \*  $D_i = -1$ : failure in first pair member, success in second,
- disregard all  $D_i = 0$ ; let  $n_1 = \# (D_i = 1 \text{ or } D_i = -1)$ ; assume  $Y = \# (D_i = 1) \sim B(n_1, p)$ ; and test  $H_0 : p = 0.5$  against  $H_a : p \neq 0.5$ ,
- example:  $Y_{\text{obs}} = 4$ ;  $Y \sim B(15, p)$ ; and  $P = 2 \times P(Y \leq 4) \approx 2 \cdot (0 + 0 + 0.003 + 0.014 + 0.042) = 0.12$  (binomial table); conclusion: no statistical evidence against  $H_0$ .

---

<sup>7</sup> Different versions of McNemar’s test exist; the one described here gives an exact  $P$ -value based on the binomial distribution, and is generally recommended.

# RANKS

Values/numbers  $x_1, \dots, x_n$ .

- order values by increasing magnitude:

$$x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}, \quad \text{where } \text{rank}(x_{(i)}) = i$$

i.e., rank= $i$  when value is  $i$ th smallest among all values,

- ties (several values equal):  
use average rank among all tied values,
- it is sometimes possible to assign ranks, even if data only partially observed (*left-censored*: smaller than or equal to a cut-off, *right-censored*: greater than or equal to a cut-off).

Example (constructed data):

|              |     |     |                  |                  |     |     |
|--------------|-----|-----|------------------|------------------|-----|-----|
| data         | 2.2 | 3.1 | 1.9              | 2.2              | 2.0 | 5.0 |
| ordered data | 1.9 | 2.0 | 2.2              | 2.2              | 3.1 | 5.0 |
| ranks        | 1   | 2   | 3.5 <sup>a</sup> | 3.5 <sup>a</sup> | 5   | 6   |

<sup>a</sup> average rank computed as:  $3.5 = (3 + 4)/2$

- the value 5.0 is much larger than the others but that is not reflected (strongly) in the ranks,
- if an additional observation was partially observed and only known to be  $> 5$  (i.e., right-censored at 5), then its rank would be 7,
- sum of ranks = 21 (generally, among  $n$  values the sum of ranks equals  $n(n + 1)/2$ ).

## 2 SAMPLES: WILCOXON—MANN—WHITNEY TEST

Wilcoxon rank sum test (PSLS/IPS terminology, also commonly Mann—Whitney test):

- Model:  $X_1, \dots, X_{n_1}$  and  $Y_1, \dots, Y_{n_2}$  independent and i.i.d. samples from distrib.  $\text{Dist}_X$  and  $\text{Dist}_Y$ , respect.,
- Hypotheses — two possibilities:
  - (1)  $H_0: \text{Dist}_X = \text{Dist}_Y$  (same distrib.),  $H_a: \text{Dist}_X \neq \text{Dist}_Y$ ,<sup>8</sup>
  - (2) assuming “ $\text{Dist}_X = \text{Dist}_Y + \Delta$ ” (distrib. differ only in position):  $H_0: \Delta = 0$  (corresponding to  $\text{median}_X = \text{median}_Y$ ) vs. one- or two-sided alternatives  $H_a$ ,
- Test procedure:
  - \* rank all observations as if a single sample,
  - \* test statistic:  $W =$  sum of ranks for  $X$ -sample,
  - \* under  $H_0$ : distribution of  $W$  has no easy form
    - tabulated in special tables for small  $n_1, n_2$ , when there are *no ties*; some programs give exact values,
    - different types of approximations in Minitab/Stata/R, with improving accuracy for increasing sample size,
- Confidence interval for  $\text{median}_X - \text{median}_Y$  (valid under  $\Delta$ -assumption above): in Minitab/Stata/R,
- recommended to check that there is similar spread and skewness in the two distributions of ranks.<sup>9</sup>

<sup>8</sup> More specific wording of  $H_a$ :  $\text{Dist}_X$  is systematically larger than  $\text{Dist}_Y$ , or vice versa (for a two-sided  $H_a$ ); see Chapter 27 of PSLS.

<sup>9</sup> Fagerland & Sandvik (2009), *Statistics in Medicine* **28**, 1487-1497.

|                                 |
|---------------------------------|
| EXAMPLE FOR 2-SAMPLE W-M-W TEST |
|---------------------------------|

Parasite burdens of calves in Lithuania:

| pasture  | Data values |            |            |          |    |          |      |      |          |     |
|----------|-------------|------------|------------|----------|----|----------|------|------|----------|-----|
| safe     | 0           | 8          | 8          | 10       | 26 | 34       | 38   | 44   | 46       |     |
| infected | 20          | 30         | 30         | 36       | 50 | 52       | 54   | 70   | 70       | 100 |
| both     | 0           | 8          | 8          | 10       | 20 | 26       | 30   | 30   | 34       | 36  |
| samples  | 38          | 44         | 46         | 50       | 52 | 54       | 70   | 70   | 100      |     |
|          | Ranks       |            |            |          |    |          |      |      |          |     |
| bold ~   | <b>1</b>    | <b>2.5</b> | <b>2.5</b> | <b>4</b> | 5  | <b>6</b> | 7.5  | 7.5  | <b>9</b> | 10  |
| safe     | <b>11</b>   | <b>12</b>  | <b>13</b>  | 14       | 15 | 16       | 17.5 | 17.5 | 19       |     |

Nonparametric analysis:

- Model: two independent samples, assume also that distributions differ only in position ( $\Delta$ -assumption),
- test statistic:  $W$  = sum of ranks in safe sample = 61,
- approximate P-value = 0.020(Minitab/R) / 0.018(Stata),
- 95% CI for median difference (infected–safe): (6.0,46.0).

Normal distribution analysis (from Lecture 7):

- Estimation:  $\hat{\mu}_1 = 51.2$ ,  $\hat{\mu}_2 = 23.8$ ,  $s_1 = 24.0$ ,  $s_2 = 17.6$ ,
- standard error for  $\hat{\mu}_1 - \hat{\mu}_2$  :  $\sqrt{s_1^2/10 + s_2^2/9} = 9.59$ ,
- test statistic:  $t = (\hat{\mu}_1 - \hat{\mu}_2)/SE(\hat{\mu}_1 - \hat{\mu}_2) = 2.86$ ,
- P-value = 0.011 — from  $t(16)$ ,
- 95% CI for mean difference (infected–safe): (7.1,47.8).

## 1 SAMPLE: WILCOXON'S SIGNED RANK TEST

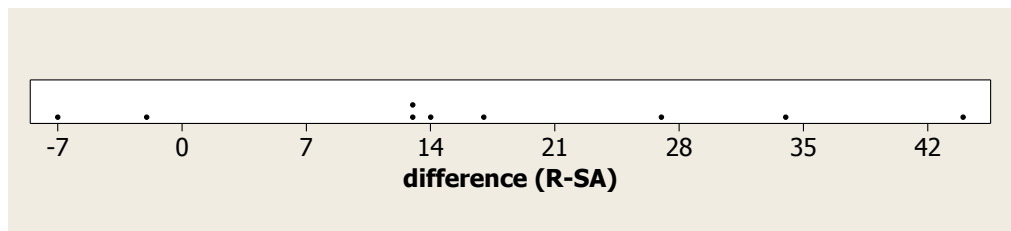
Wilcoxon's test for  $H_0$ : median = known value:

- Model:  $X_1, \dots, X_n$  i.i.d. sample from a continuous, *symmetric* distribution, (note: extra assumption)
- Null hypothesis  $H_0$ : median = known value ( $m_0$ ),
- Alternative hypotheses: either one- or two-sided,
- Test procedure:
  - \* let  $R_i = X_i - m_0$ , and discard obs. with  $R_i = 0$
  - \* rank the  $|R_i|$ 's, and let  $S_i = \text{rank of } |R_i|$ ,
  - \* idea: if, for example, true median  $> m_0$ , then there'll be both *more and larger ranks* for observations  $> m_0$ ,
  - \* test statistic:  $W^+ = \text{sum of } S_i$ 's for positive obs. ( $R_i > 0$  corresponding to  $X_i > m_0$ ),
  - \* under  $H_0$ : distribution of  $W^+$  has no easy form
    - tabulated in special tables for small  $n$  (not counting discarded obs.), when there are *no ties*,
    - different types of approximations, with accuracy that improves with increasing sample size,
    - Minitab and Stata use approximations, other programs (incl. R) calculate exact values,
- Confidence interval for median: in Minitab/R.

## EXAMPLE FOR 1-SAMPLE WILCOXON TEST

Visual receptive fields:

- dotplot for differences (R-SA):



- data values ( $X_i$ ) and ranks ( $S_i$ ; bold  $\sim X_i > 0$ ):

|         |      |      |          |          |          |          |          |          |          |
|---------|------|------|----------|----------|----------|----------|----------|----------|----------|
| $X_i$   | -7.5 | -2.5 | 12.5     | 13.3     | 14.2     | 16.7     | 26.7     | 34.2     | 44.2     |
| $ R_i $ | 7.5  | 2.5  | 12.5     | 13.3     | 14.2     | 16.7     | 26.7     | 34.2     | 44.2     |
| $S_i$   | 2    | 1    | <b>3</b> | <b>4</b> | <b>5</b> | <b>6</b> | <b>7</b> | <b>8</b> | <b>9</b> |

- assume distribution *symmetric* about its median,
- $H_0$ : median = 0 vs.  $H_a$ : median > 0,
- $W^+ = 3 + 4 + 5 + 6 + 7 + 8 + 9 = 42$ ,  $W^- = 3$ ,
- P-value: 0.012 (Minitab) / 0.021 (Stata) / 0.010 (R),
- 95% CI for median (Minitab/R): (3.3, 30.5),
- Wilcoxon test is significant and preferable here because assumed symmetry seems quite reasonable.

Summary — comparison Wilcoxon vs. sign test:

Wilcoxon test is *stronger* (in fact, the sign test is quite weak), but has *additional assumption of symmetry*.

## STATISTICAL METHODS TO CHOOSE SAMPLE SIZE

Question: how many subjects to take??

Plain common sense:

- size should be sufficient to detect (statistical significance of) treatment differences of interest,
- avoid “waste” of experimental units,
- reduce sensitivity to errors (by taking replications).

Fact: all formal procedures require pre-decided statistical model and detailed prior knowledge (estimates or guesses) about the outcomes:

- size of effect of interest, or desired precision,
- standard deviation of observations (for continuous data).

Two general approaches for determining sample size:

- (1) from desired precision (standard error, size of 95% CI) on selected estimate (typically involving mean(s)),
- (2) from desired power of test for effect of interest, using *always* statistical software (avoid hand calculations<sup>10</sup>):
  - \* Minitab/Stata (or others) for basic designs,
  - \* specialized software for special and advanced designs.<sup>10</sup>

<sup>10</sup> The formulae of IPS, VER and also Lehr’s formula are not recommended; use instead software or web applets: <http://homepage.stat.uiowa.edu/~rlenth/Power/>.

|               |
|---------------|
| EXERCISE 6.19 |
|---------------|

Impact of sample size  $n$  for study of reading ability of 3<sup>rd</sup> grade children. Preliminary study has given  $s = 12$ , so that  $\sigma = 12$  is assumed. We also assume an approximate normal distribution for the sample mean  $\bar{X}$ .

(a)  $n = 100$ , margin of error for 95% CI (with known  $\sigma$ ):

$$z^* \sigma / \sqrt{n} = z_{.975} \sigma / \sqrt{n} = 1.96 \times 12 / \sqrt{100} = 2.35,$$

or more realistically with  $\sigma$  unknown:

$$t^* \sigma / \sqrt{n} = t_{.975}(99) \sigma / \sqrt{n} = 1.984 \times 12 / \sqrt{100} = 2.38,$$

(b)  $n = 10$ , margin of error for 95% CI (with known  $\sigma$ ):

$$z^* \sigma / \sqrt{n} = 1.96 \times 12 / \sqrt{10} = 7.44,$$

and again alternatively with  $\sigma$  unknown:

$$t^* \sigma / \sqrt{n} = t_{.975}(9) \sigma / \sqrt{n} = 2.262 \times 12 / \sqrt{10} = 8.58,$$

(c) appropriate  $n$  for a desired margin of error of  $m = 3$  must be between 10 and 100 — we can work our way through trial and error, or use a simple formula, on the next page, *for known  $\sigma$* :

$$n = \left( \frac{z^* \sigma}{m} \right)^2 = \left( \frac{1.96 \times 12}{3} \right)^2 = 7.84^2 = 61.5,$$

take  $n = 62$  to ensure that margin of error  $\leq 3$ .<sup>11</sup>

---

<sup>11</sup> For unknown  $\sigma$ , we should repeat the calculation with  $t^* = t_{.975}(61) = 2.00$ , to see whether changing from  $z^*$  to  $t^*$  affects required  $n$  substantially: it gives  $n = 64$ .

## CONTROLLING SIZE OF CONFIDENCE INTERVALS

How to decrease size of confidence intervals for population means? — based on the formula for 1-sample means and assuming *known*  $\sigma$ ,

$$\text{margin of error} = z^* \times \sigma / \sqrt{n}.$$

- increase  $n$  (more data),
- increase  $\alpha$  to decrease  $z^* = z_{1-\alpha/2}$ , same as decrease  $C = 1 - \alpha$  (lower certainty/confidence of interval),
- decrease  $\sigma$  (reduce variation in population, by shifting to another variable or another population).

How to adjust sample size to get a desired margin of error ( $m$ ) for a population mean?

- fix  $m$ ,  $\alpha$  and  $\sigma$  at suitable values,
- invert above formula for margin of error, and solve for  $n$ :

$$m \geq \frac{z^* \times \sigma}{\sqrt{n}} \quad \text{or} \quad n \geq \left( \frac{z^* \times \sigma}{m} \right)^2.$$

- formula guarantees margin of error  $\leq m$ , provided assumptions for confidence interval met.

If  $\sigma$  cannot be assumed known (most realistically), the calculation should be redone with  $t^* = t_{.975}(n-1)$  to assess any changes in the required  $n$ .

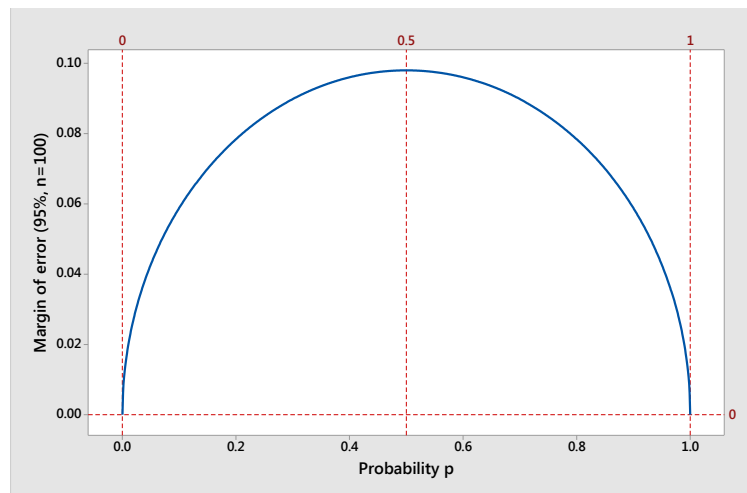
# CONTROLLING MARGIN OF ERROR

## FOR A SINGLE PROPORTION

Margin of error: (based on the normal approximation<sup>12</sup>)

$$\text{margin of error} = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad z^* = z_{1-\alpha/2},$$

- ways to reduce margin of error: increase  $n$  or  $\alpha$ ,
- margin of error largest at  $p = 0.5$  (see graph).



How to adjust sample size to get a desired margin of error ( $m$ ) for a proportion?

- fix  $m$ ,  $\alpha$  and  $p$  (guessed) at suitable values,
- invert above formula for margin of error, and solve for  $n$ :

$$m \geq z^* \times \sqrt{\frac{p(1-p)}{n}} \quad \text{or} \quad n \geq p(1-p)(z^*/m)^2.$$

- formula guarantees margin of error  $\leq m$ , provided assumptions for confidence interval met,
- using  $p = 0.5$  is conservative (maybe too large  $n$ ).

<sup>12</sup> The Minitab menu (Sample Size for Estimation) uses a more exact calculation.

## SAMPLE SIZE BASED ON ESTIMATION PRECISION

General approach for normal distribution models:

- assume estimated/guessed/known standard deviation  $\sigma$ ,
- assume (mean) parameter of interest  $\mu$  and estimate  $\hat{\mu}$ , with standard error  $SE(\hat{\mu}) = \sigma \times c(n)$ , where  $c(n)$  is a known constant (depending on number of obs.  $n$ ),
- approximate 95% CI<sup>13</sup>:  $\hat{\mu} \pm 2 \sigma c(n)$ ,

Compute  $n$  to achieve desired margin of error ( $M$ ) by solving with respect to  $n$  in the equation:

$$M(\text{desired value}) \geq 2 \sigma c(n).$$

Example: blood pressure measured on patients before and after an intervention,

- design: two paired samples, use  $D = \text{after} - \text{before}$ ,
- model: i.i.d. sample (differences!) of size  $n$  from  $N(\mu, \sigma)$ , with guessed value  $\sigma = 10$  (*mm Hg*),
- $\mu =$  population mean,  $\hat{\mu} = \bar{D}$  (sample mean),  $c(n) = 1/\sqrt{n}$ ,
- assume, the desired margin of error of CI for  $\mu$  is  $M = 3$   $\sim$  observed sample mean of 3 signif. at the 5% level,
- solve:  $3 \geq 2 \times 10/\sqrt{n} \Rightarrow n \geq (2 \cdot 10/3)^2 = 44.4 \approx 45$ ,
- conclusion: with  $n = 45$  (Minitab: 46) patients, a 95% CI for the difference would have a margin of error of 3.

---

<sup>13</sup> Approximation requires either  $\sigma$  known, or  $\sigma$  unknown and  $n$  so large that  $t^* = t_{.975}(\text{df}) \approx z^* = z_{.975} \approx 2$ , say  $n \geq 40$ .

|                               |
|-------------------------------|
| ERRORS OF TYPE I–II AND POWER |
|-------------------------------|

Errors of type I and II:

- type I error: to reject  $H_0$ , when  $H_0$  in reality true,
- type II error: to not reject  $H_0$ , when  $H_0$  in reality false,
- definition of statistical tests involves only type I errors, which are controlled by the significance level ( $\alpha$ ),
- power of statistical tests involves type II errors (below),
- overview:

| Conclusion<br>from sample | Truth about population |                           |
|---------------------------|------------------------|---------------------------|
|                           | $H_0$ true             | $H_a$ true ( $H_0$ false) |
| reject $H_0$              | type I error           | no error                  |
| not reject $H_0$          | no error               | type II error             |

Power of a statistical test:

- involves a specific alternative, e.g.  $H_a: \mu = 0.84$  in the laboratory analysis example (6L–5) with  $H_0: \mu = 0.86$ ,
- definition: power = probability that the statistical test will *reject*  $H_0$ , when the specific alternative  $H_a$  is true,  
= 1 – type II error,
- important for planning of experiments: what chance of a significant result?
- difficult to calculate in complex models (lots of formulae and software exist).

## SAMPLE SIZE BASED ON POWER

Requirements for sample size determination based on power:

- statistical model / design and corresponding software,
- size of effect<sup>14</sup> desired to be detected,
- standard deviation of model (normal data),
- desired value of power (0.8, or 80%, commonly used),
- significance level of test employed (usually 0.05, or 5%).

Computations: *always!* using software, e.g. Minitab<sup>15</sup>

- Stat-Power and Sample size-1 Sample Z (known  $\sigma$ ),
- Stat-Power and Sample size-1 Sample t (unknown  $\sigma$ ).

Blood pressure example continued:

- same model and setup as before (known  $\sigma = 10$ ); assume interest in *true population mean (difference)* of 3 units,
- computation of power for  $n = 45$  and sign. level 0.05:
  - \* two-sided alternative: power=0.52 (unknown  $\sigma$ : 0.50),
  - \* one-sided alternative: power=0.64 (unknown  $\sigma$ : 0.63),
- computation of necessary sample size to achieve power = 0.8 with a significance level of 0.05:
  - \* two-sided altern.: required  $n = 88$  (unknown  $\sigma$ : 90),
  - \* one-sided altern.: required  $n = 69$  (unknown  $\sigma$ : 71).

<sup>14</sup> Here (not generally), effect size is the difference between  $H_0$  and  $H_a$  values.

<sup>15</sup> From Stata 13 onward, use the menu **Statistics-Power and Sample Size**.

## SAMPLE SIZE MISCONCEPTIONS, AND EQUIVALENCE TESTING

Common misconceptions<sup>16</sup> in sample size calculations:

- use of standard effect sizes (general definitions of “small”, “medium” and “large” effects, relative to std. dev.): effects of interest should be determined exclusively from the context of your study,
- retrospective power calculation: after a study has been carried out and using its estimated values:
  - \* power/sample size calculations aid in planning of new studies, not in interpreting results of data analysis,
  - \* confidence intervals give the best information about the unknown parameters from a study,
  - \* if  $H_0$  was not rejected, the conclusion may be strengthened by an equivalence test (instead of arguing from the study’s power).

An equivalence test<sup>17</sup> is for making a statement that effects of two “treatments” differ at most by a (biologically) small amount (say  $\delta$ ),

- for  $H_0 : \theta = 0$  ( $\theta$  being a difference in means, or other parameters), not rejecting  $H_0$  is a weak and non-quantitative conclusion,
- a CI for the difference  $\theta$  contains useful information,
- a non-equivalence hypothesis  $H_0^{(ne)} : |\theta| \geq \delta$  can be tested against the  $H_a^{(ne)} : |\theta| < \delta$ , as follows (at a 5% significance level):
  - \* compute a 90% CI (not 95% CI) for  $\theta$ ,
  - \* reject  $H_0^{(ne)}$ , if the interval  $(-\delta, \delta)$  entirely includes the CI.

<sup>16</sup> Largely based on Lenth (2001), *The American Statistician* **55**, 187–193.

<sup>17</sup> A *non-inferiority test* is a similar construct, only with a one-sided alternative.

## SUMMARY NOTES

Nonparametric tests characterized by no distribution (normality) assumptions, often focusing on median instead of mean; many methods exist – VHM 801 course covers:

- sign test for 1-sample  $\rightarrow$  test of  $H_0 : p = 0.5$  in  $B(n, p)$ ,
- rank-based tests for 1-sample (Wilcoxon signed rank) and 2-sample indep. (Mann-Whitney): software calculation, test/model assumptions about distrib. shape,

Statistical sample size calculation – two main approaches:

- based on estimation precision: determine sample size to achieve a desired precision for an estimate of interest,
  - \* requirements: desired precision (e.g. margin of error for CI), standard deviation (contin. data),
  - \* implementation: hand calculation formulae (+Minitab), typically derived from the SE of the estimate of interest (formulae exist for standard settings),
- based on power of statistical tests:
  - \* test  $H_0$  against one- or two-sided  $H_a$  (standard setup),
  - \* type I and type II error of statistical testing,
  - \* power against specific alternative hypothesis  $H_a$ ,
  - \* requirements: targeted effect (e.g. mean difference) to detect, desired power level, test settings (incl. signif. level, type of  $H_a$ ), standard deviation (contin. data),
  - \* implementation: statistical software/web applications.