Index of 6-L

Practical Information

Today's lecture:

- brief follow-up/review from Session 5 based on Exercise 6.95 and 5L−19,
- inference for continuous data, without assumption of known σ:
  - a new distribution: the t-distribution(s),[1]
- lots of examples (data sets!) and practice with statistical inference,
- inference for one and two samples (continuous data):[2]
  - a single sample (no assumption of known σ),
  - two independent samples,
  - two dependent (paired) samples.

Scheduling notes:

- home assignment returned today (solution at website),
- next home assignment is on October 27,
- we need to reschedule the mid-term exam (around November 3).

[1] PSLS 4e: Chapter 17; S: Section 7.4; IPS 7e: Section 7.1.
[2] PSLS 4e: Chapters 17-18; S: Chapters 8-9 (parts); IPS 7e: Sections 7.1-2.

Outline of Statistical Analysis (revisited)

- Data description,
- Statistical model,
- Estimation of model's unknown parameter(s),
  - incl. confidence intervals and/or standard errors,[3]
- Model check:
  - comparison of the observed distribution and assumed theoretical distribution (using estimated parameters),
  - methods: graphical (plots) or numerical (tests),
  - if model is deemed unsatisfactory, start over with new model,
- Hypothesis testing:
  - set up null hypothesis H_0 (model simplification) and alternative hypothesis H_a,
  - test statistic and associated P-value summarize our confidence against null hypothesis, which we may reject (low P) or not reject (high P),
- Conclusion / Presentation:
  - summary of test results,
  - illustrations of the implications of the final model, e.g. prediction.

[3] Recall, that the standard error (SE) is the standard deviation in the distribution of the estimate, and thus an indication of the estimate's precision.

1-Sample Estimation

Data: sample $X_1, \ldots, X_n$ of size n from some distribution with unknown mean μ and unknown standard deviation σ (and variance σ^2). More specifically, we assume

- the X's are i.i.d. (independent, identically distributed),
- $\mathrm{E}X_i = \mu$ and $\mathrm{sd}X_i = \sigma$ for all X's.

For estimation of σ we use the sample standard deviation:

$$\hat{\sigma} = s \;(= \sqrt{s^2}) \quad \text{and} \quad \hat{\sigma}^2 = s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1),$$

and s^2 is an unbiased estimate of σ^2, explaining why we use $(n-1)$ for s^2. [4]

Summary of terminology and estimates for a single sample:

Name	Estimate	Parameter	Properties
sample mean	$\bar{X}$	μ	unbiased
sample variance	s^2	σ^2	unbiased
sample standard deviation	s	σ	biased, natural
(sample variance of mean)	s^2/n	$\sigma^2_{\bar{X}} = \sigma^2/n$	unbiased
standard error of mean	$s/\sqrt{n}$	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$	biased, natural

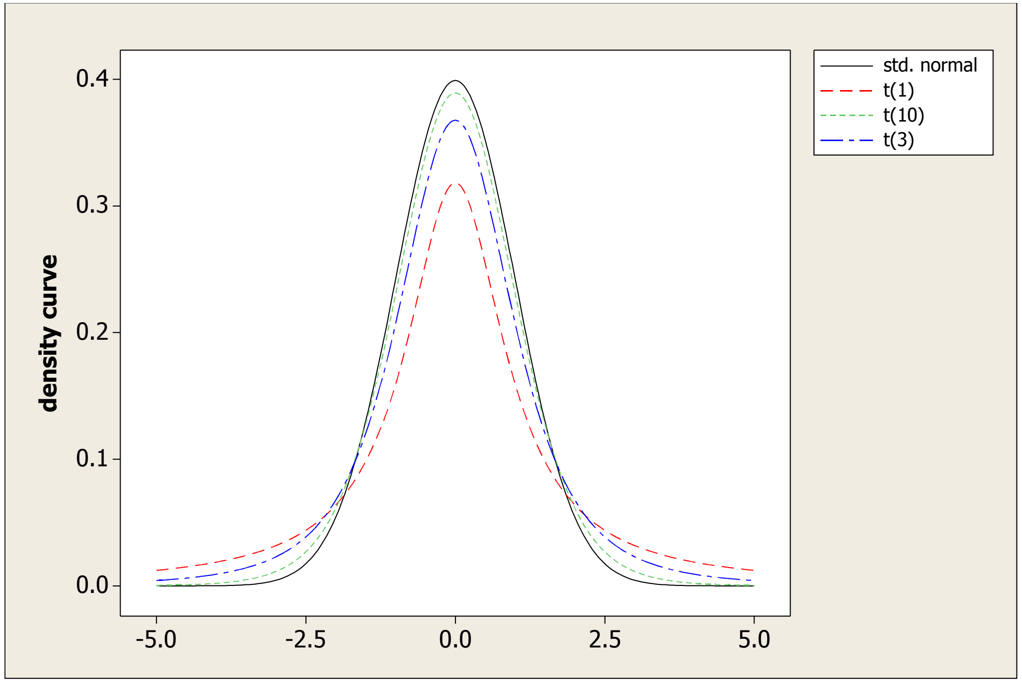
In addition, $\bar{X} \sim \mathrm{N}(\mu, \sigma_{\bar{X}})$

- exactly — if the X's are normally distributed,
- approximately (when n is "large") — always! (by the CLT)

[4] We skip over the (small) mathematical calculation showing that s^2 is unbiased.

("Student's") t Distribution

- new distribution(s) — to be used not for modelling but for inference in a normal distribution model when σ is estimated from the data, as the reference distribution for t test statistics (next slide),
- has a single parameter r (or "df"):
  - $r = 1, 2, 3, \ldots$
  - called "degrees of freedom" (explanation to follow),
  - given from the data, and not to be estimated.
- denoted $t(r)$ to indicate degrees of freedom,
- distribution on $(-\infty, \infty) \Rightarrow$ positive and negative values,
- symmetric around zero, almost "bell-shaped" but with heavier tails than N(0,1) ($\Rightarrow$ positive kurtosis),
- when r is large: $t(r) \approx \mathrm{N}(0, 1)$, see graph.



1-Sample Normal Distribution Inference

- Data: $X_1, \ldots, X_n$ ($n =$ number of observations).
- Model: observations are a sample (i.i.d.) from $\mathrm{N}(\mu, \sigma)$, where μ and σ are unknown parameters.
- Estimation: $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = s$.
- Distribution of estimates:
$$\hat{\mu} = \bar{X} \sim \mathrm{N}(\mu, \sigma/\sqrt{n}), \quad s_{\bar{X}} = s/\sqrt{n},$$
$$(\bar{X} - \mu)/s_{\bar{X}} \sim t(n-1),$$
note that degrees of freedom (df) $= n-1$,
- Confidence interval with confidence level $1-\alpha$:
$$\mu : \bar{X} \pm t^* s_{\bar{X}} = \bar{X} \pm t^* \, s/\sqrt{n},$$
where t^* is $(1-\frac{\alpha}{2})$-percentile of a $t(n-1)$ distribution, [5]
- Test of H_0: $\mu = \mu_0$ against alternative H_a:
  * test statistic: $t = (\bar{X} - \mu_0)/s_{\bar{X}} = (\bar{X} - \mu_0)/(s/\sqrt{n})$,
  * P-value from t distribution with df $= n-1$:
    - H_a: $\mu \neq \mu_0$: $P = 2 \times \mathrm{P}(t(\mathrm{df}) \geq |t_{\mathrm{obs}}|)$,
    - H_a: $\mu > \mu_0$: $P = \mathrm{P}(t(\mathrm{df}) \geq t_{\mathrm{obs}})$, and H_a: $\mu < \mu_0$: $P = \mathrm{P}(t(\mathrm{df}) \leq t_{\mathrm{obs}})$,
- note strong similarities with z-based procedures.

[5] In notation, $t^* = t_{1-\alpha/2}(\mathrm{df})$ in the $t(\mathrm{df})$ distribution; see Table C of PSLS, Table 3 of S, Table D of IPS.

Example: Human Body Temperature

Example 14.9 of PSLS 4e, Example p. 139 of S:

- Data: 130 measurements[6] of body temperature in °F of healthy adults: $X_1, \ldots, X_{130}$ ($n = 130$).
- Model: a sample (i.i.d.) from $\mathrm{N}(\mu, \sigma)$.
- Estimation: $\hat{\mu} = \bar{X} = 98.25$ and $\hat{\sigma} = s = 0.733$.
- Confidence interval with confidence level 95% ($\alpha = 0.05$):

  $$\mu : \bar{X} \pm t^* s_{\bar{X}} = 98.25 \pm 1.98 \times 0.733/\sqrt{130} = 98.25 \pm 0.13 = (98.12, 98.38),$$

  using $t^* = t_{.975}(129) = 1.9785$ from Minitab,
- Test of H_0: $\mu = 98.6$ against alternative H_a: $\mu \neq 98.6$ ("classical" body temp.):
  * test statistic: $t = \dfrac{\bar{X} - 98.6}{s_{\bar{X}}} = \dfrac{98.25 - 98.6}{0.733/\sqrt{130}} = -5.45$,
  * P-value from t distribution with df $= n - 1 = 129$:

    $$P = 2 \times \mathrm{P}(t(129) \geq 5.45) < 0.000001 \text{ (Minitab)}$$

  * Conclusion: strong evidence to say that average body temperature is different, actually lower, than the "classical" reference value.

---

[6] Constructed data for pedagogical purposes (Shoemaker (1996), *Journal of Statistics Education* 4) based on a real study from 1992 whose purpose it was to evaluate the well-established average body temperature of 37.0°C or 98.6°F, Mackowiak et al. (1992), *Journal of the American Medical Association* 268, 1578-1580.

How to Find Percentiles and P-values

Recall that

- the $p\%$ percentile has $p\%$ of the distribution below, and $(100-p)\%$ above, where $(100-p)\%$ is the tail probability,[7]
- P-values are typically determined from tail probabilities $\mathrm{P}(t \geq t_{\mathrm{obs}})$ in standard distributions, e.g. $\mathrm{N}(0,1)$ or $t(\mathrm{df})$.

Methods to determine percentiles or tail probabilities:

- Minitab: `Probability Distribution Plot-View Probability` menu with `Shared Area` defined by `Probability` or `X Value` for percentiles and probabilities, respectively,[8]
- Stata: functions `normal`, `invnormal`, `ttail`, `invttail`; similar functions in R,[9]
- statistical tables with values for some confidence/error levels,
  - ∗ what to do, if df is not in table? — use largest value below df $\Rightarrow$ conservative analysis (larger CIs and P-values),
  - ∗ what to do, if t_{obs}-value is not in table? — find closest ("critical") values in table, for example $t_1 < t_{\mathrm{obs}} < t_2$, and use the relations

$$\mathrm{P}(t \geq t_2) < \mathrm{P}(t \geq t_{\mathrm{obs}}) < \mathrm{P}(t \geq t_1).$$

[7] In some statistical tables, the $p\%$ percentile is the critical value for a one-tailed test with $\alpha = (100-p)\%$.

[8] Alternatively, the non-graphical `Calc-Probability Distributions` menu with `Inverse Cumulative Probability` or `Cumulative Probability`, respectively.

[9] R functions: `pnorm`, `qnorm`, `pt`, `qt`.

Exercises 7.50 and 7.52

Exercise 7.50:

Percentiles/critical values for confidence intervals for population mean (with unknown population standard deviation):

(a) $n = 20$ and $C = 95\%$: $\alpha = 0.05$ and $t^* = t_{1-\alpha/2}(n-1) = t_{.975}(19) = 2.093$.

(b) $n = 30$ and $C = 90\%$: $\alpha = 0.10$ and $t^* = t_{1-\alpha/2}(n-1) = t_{.95}(29) = 1.699$.

(c) $n = 50$ and $C = 80\%$: $\alpha = 0.20$ and $t^* = t_{1-\alpha/2}(n-1) = t_{.90}(49) \approx t_{.90}(40) = 1.303$,
— a conservative value (exact value (software): 1.299).

Exercise 7.52:

Testing H_0: $\mu = 0$ against H_a: $\mu > 0$ based on a sample of 15 observations. Observed t-value is $t_{obs} = 2.15$.

(a) degrees of freedom $= 15-1 = 14$.

(b-d) percentiles from $t(14)$:

$$t_{0.975}(14) = 2.145 < 2.15 < 2.264 = t_{0.98}(14),$$

with right-tail probabilities of 0.025 and 0.02, respectively; therefore,
for $P = \mathrm{P}(t(14) \geq t_{obs})$ we have: $0.02 < P < 0.025$,

(f) the test is significant at the 5% level, but not significant at the 1% level.

(g) exact P-value (software) is $P = 0.02476$.

Inference for Non-Normal Data

If data show strong / moderate deviations from normality:

- remove outliers (if any), and see if it helps,
- try to transform the data, and see if situation is better for transformed data,
  - many transformations exist: log and square-root common for right-skewed data,
  - results at transformed scale should always be backtransformed to original scale:
    - backtransformed means $\sim$ medians in original data,[10]
    - for CI's: backtransform both endpoints,
- nonparametric statistical methods with no distributional assumptions (next lecture),
- some procedures based on the normal distribution are robust (or resistant), that is, work reasonably well even if assumptions are (mildly) violated (saved by the normality of $\bar{X}$ in the CLT!):
  - difficult to know exactly what is okay and when — some guidelines:[11]
    - $n < 15$: only if data close to normal (okay!),
    - $15 \leq n < 40$: ok unless strong skewness or outliers,
    - $40 \leq n$: also ok for clearly skewed distributions, but beware of strong outliers.

[10] It is more difficult to get means and SEs on original scale, and potentially less meaningful.

[11] Guidelines for t-distribution procedures from PSLS/IPS texts; the discussion in S is less detailed/satisfactory: assume normality or $n > 30$, in my view a very debatable guideline!. If σ is known, the inference is even less affected by non-normality, because it is the procedures involving s^2 that rely most strongly on the normality assumption.

2 Paired Samples

Paired (matched, correlated) samples/observations:

- Data: $(\boldsymbol{X}_1, \boldsymbol{Y}_1) \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ independent observation pairs:
  - typical examples of pairs:
    - same individual: left−right, before−after,
    - different individuals: twins, related or similar individuals,
  - in experimental design terminology:
    - pairs $\sim$ blocks (of size 2),
    - observations within pairs $\sim$ different treatments,
  - purpose of pairs: reduce variability and impact of other (lurking) factors,
- Model and Analysis:
  - usually work with differences: $\boldsymbol{D}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i$, (ratios $\boldsymbol{Y}_i/\boldsymbol{X}_i$ or other functions also possible),
  - assume $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n$ sample from distribution $(\boldsymbol{\mu}_D, \boldsymbol{\sigma}_D)$, where

$$\boldsymbol{\mu}_D = \mathrm{E}\boldsymbol{D}_i = \mathrm{E}\boldsymbol{Y}_i - \mathrm{E}\boldsymbol{X}_i,$$

  - hypothesis H_0: $\boldsymbol{\mu}_D = 0 \sim$ no difference between (means of) $\boldsymbol{X}$s and $\boldsymbol{Y}$s,
  - all methods for single sample inference apply!

2 Paired Samples: Visual Receptive Field

- Data: Neural activity (# spikes/sec) for a monkey's neuron in 9 recordings of both Response (R) and Spontaneous activity (SA), (PSLS Example 27.6)

Recording (i)	SA (X_i)	R (Y_i)	Difference (D_i)
1	2.5	16.7	14.2
2	7.5	20.0	12.5
...	...	...	...
9	17.5	10.0	-7.5

- Model: one sample (i.i.d.) of differences $D_1, \ldots, D_9$ assumed to follow N(μ_D, σ_D), where $\mu_D = \mu_Y - \mu_X$ is the parameter of principal interest,
- Estimation: $\hat{\mu}_D = \bar{D} = 16.87$, $\hat{\sigma}_D = s_D = 16.40$,
- 95% Confidence interval for μ_D:

$$\bar{D} \pm t^* s_D/\sqrt{n} = 16.87 \pm 2.306 \cdot 16.40/\sqrt{9} = 16.87 \pm 12.61,$$

- Test of H_0: $\mu_D = 0$ ($\sim \mu_Y = \mu_X$) against alternative H_a: $\mu_D > 0$ ($\sim \mu_Y > \mu_X$):
  * test statistic: $t = \dfrac{\bar{D} - 0}{s_D/\sqrt{n}} = \dfrac{16.87}{16.40/\sqrt{9}} = 3.09$,
  * P-value from t distribution with df$=8$: $P = \mathrm{P}(t(8) > 3.09) = 0.007$,[12]
  * conclusion: clearly significant difference between neural activity at SA and R, and higher R activity.

[12] The $t(8)$-distribution table has $t_{.99}(8) = 2.896$ and $t_{.995}(8) = 3.355$, from which we would get: $0.005 < P < 0.01$.

2 INDEPENDENT SAMPLES — INTRODUCTION

- Data:

$$\begin{aligned} X_1, \dots, X_{n_1} &\sim \text{ first sample, of size } n_1, \\ Y_1, \dots, Y_{n_2} &\sim \text{ second sample, of size } n_2. \end{aligned}$$

- Model: all observations independent, and the X's and Y's are samples from separate distributions,
- typical example: treatment and control groups, e.g. study on parasite burdens in Lithuanian calves,
- how to distinguish from paired samples?
  - not necessarily the same number of observations (i.e., maybe $n_1 \neq n_2$),
  - no relation between X_1 and Y_1, X_2 and Y_2, etc.
  - the X's are interchangable ("replications"), and the same for the Y's.

Overview of inference for mean difference $\mu_1 - \mu_2$, based on normal distributions:

- assumptions: normal distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ for the two samples, with all parameters unknown,
- slightly different procedures depending on whether

(1) $\sigma_1 \neq \sigma_2$ (general situation[13]).

(2) $\sigma_1 = \sigma_2$ (simplest; often unrealistic assumption).[14]

[13] Without a specific assumption about the σ's: we could have $\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$.

[14] Not part of VHM 801 syllabus; PSLS and S texts avoid this ("pooled variance") method.

EXERCISE 7.40

Identify statistical design as either (1) single sample, (2) matched pairs (paired sample) or (3) two independent samples:

(a) two independent samples, because different groups of children and only one score from each child; the before versus after element is not part of the data collection,

(b) two paired samples, because two scores are collected from each child, in random order; nor is the before versus after element part of the data collection here,

(c) one sample, because only one sample (of 20 measurements) is taken,

(d) two independent samples, because there is no connection between the measurements taken with the new and old method.

What about a slight variation of the design where 10 samples are taken from the specimen and each is analyzed with both the new and old method?

That would be two paired samples.

2 Independent Samples – Equal Variances

- Models: 1^{st} sample: $\mathrm{N}(\mu_1, \sigma_1)$, 2^{nd} sample: $\mathrm{N}(\mu_2, \sigma_2)$,
- assume $\sigma_1 = \sigma_2 = \sigma$, based on judgement[15] or test[16],
- Estimation of means: $$\hat{\mu}_1 = \bar{X} \sim \mathrm{N}(\mu_1, \sigma/\sqrt{n_1}),\ \ \hat{\mu}_2 = \bar{Y} \sim \mathrm{N}(\mu_2, \sigma/\sqrt{n_2})$$
- estimation of σ (from s_1 and s_2 in X- and Y-samples): $$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \quad \text{and} \quad \hat{\sigma} = s = \sqrt{s^2},$$
  — "pooled" s^2: a weighted average of s_1^2 and s_2^2,
- standard error of mean difference: $s_{\bar{X}-\bar{Y}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$,
- degrees of freedom: $\mathrm{df} = (n_1-1) + (n_2-1) = n_1 + n_2 - 2$,
- Confidence interval of level $(1-\alpha)$ for $\mu_1-\mu_2$: $$\mu_1-\mu_2 :\ \ \bar{X}-\bar{Y} \pm t^* s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad t^* = t_{1-\alpha/2}(\mathrm{df}),$$
- Test of H_0: $\mu_1 = \mu_2$ against altern. H_a, using test statistic: $t = (\bar{X}-\bar{Y})/\left(s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$,
  - P-value from t distribution, e.g. computed as:
    H_a: $\mu_1 \neq \mu_2$: $P = 2 \times \mathrm{P}(t(\mathrm{df}) > |t_{\mathrm{obs}}|)$, or H_a: $\mu_1 > \mu_2$: $P = \mathrm{P}(t(\mathrm{df}) > t_{\mathrm{obs}})$,
- note similarities with 1-sample procedures.

---

[15] PSLS/IPS guideline for assuming equal standand deviations: $s_{\max}/s_{\min} \leq 2$.
[16] Variance tests (especially Bartlett's test) are overly sensitive to non-normality.

2 Independent Samples — General

Similar procedure — changes in $s_{\bar{X}-\bar{Y}}$ and df:

- no assumption of $\sigma_1 = \sigma_2$: $\Rightarrow$ more general procedure (also for when $\sigma_1 \approx \sigma_2$),
- Estimation of means and standard deviations: separately for each sample: $\bar{X}, s_1, \bar{Y}, s_2$,
- standard error of mean difference: $s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$,
- degrees of freedom — two approaches:
  * conservative (too low/"safe") df: $\min(n_1 - 1, n_2 - 1)$,
  * approximate with "terrible" formulas[17], but the approximations are generally considered good,
- Confidence interval of level $(1-\alpha)$ for $\mu_1 - \mu_2$:

$$\mu_1 - \mu_2: \;\; \bar{X} - \bar{Y} \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad t^* = t_{1-\alpha/2}(\text{df})$$

- Test of H_0: $\mu_1 = \mu_2$ against altern. H_a, using test statistic: $t = (\bar{X} - \bar{Y}) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$,
  * P-values from t distribution in the same way, e.g.
    - H_a: $\mu_1 \neq \mu_2$: $P = 2 \times \mathrm{P}(t(\text{df}) > |t_{\text{obs}}|)$,
    - H_a: $\mu_1 > \mu_2$: $P = \mathrm{P}(t(\text{df}) > t_{\text{obs}})$.

[17] Minitab uses Satterthwaite method, Stata/R use Welch method; both ok to use.

2 Independent Samples: Parasite Data

- Data: $n_1 = 10$ and $n_2 = 9$ parasite counts of calves on infected (X) and safe (Y) pasture,
- Model: 2 independent samples (i.i.d.) from $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively,
- Estimation: $\hat{\mu}_1 = \bar{X} = 51.2$, $\hat{\sigma}_1 = s_1 = 24.0$, and $\hat{\mu}_2 = \bar{Y} = 23.8$, $\hat{\sigma}_2 = s_2 = 17.6$,
- some difference in estimated standard deviations, so we are not going to assume that $\sigma_1 = \sigma_2$,
- Confidence interval with confidence level 95%:

$$\begin{aligned} \mu_1 - \mu_2 \ &: \ \bar{X} - \bar{Y} \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}, \\ &= \ 27.4 \pm 2.12\sqrt{24.0^2/10 + 17.6^2/9} = 27.4 \pm 20.3, \end{aligned}$$

  where $t^* = t_{.975}(16) = 2.12$ (df computed by software),
- Test of H_0: $\mu_1 = \mu_2$ against alternative H_a: $\mu_1 \neq \mu_2$:
  * test statistic: $t = \dfrac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 2.86$,
  * approximate P-value from t distribution with df$=$16:

$$P = 2 \times \mathrm{P}(t(\mathrm{df}) > 2.86) = 0.011.$$

  * conclusion: significant difference between parasite burdens on infected and safe pastures; that is, parasite levels are lower on safe pasture.

Summary Notes

Key words and concepts:

- statistical inference for 1 sample (quantitative outcome) on a normal distribution with unknown parameters (mean and standard deviation):
  - sample standard deviation (s) as estimate of population standard deviation (σ), standard error (for sample mean),
  - t-distribution, degrees of freedom,
  - formulae for t-based confidence interval and t-test,
- finding/approximating P-values and critical values (t^*),
- designs involving 1 and 2 samples (any distribution):
  - 1-sample, 2 independent samples, 2 paired (dependent, correlated) samples,
  - 2 paired samples $\rightarrow$ 1-sample for differences,
- choice of method/assumption for 2 independent normal distribution samples:
  - equal variances assumed: pooled variance estimate, ratio of standard deviations ≥ 2 rule,
  - no variance assumption: df determined by software.