

Index of 6-L

Page	Title
1	Practical information
2	Two examples of test problems
3	Introduction to statistical testing
4	Components of a statistical test
5	Test for population mean
6	One- or two-sided?
7	Testing by confidence interval
8	Pitfalls of statistical testing
9	1-sample estimation
10	(Student's) t distribution
11	1-sample normal distribution inference
12	Example: Human body temperature
13	How to find percentiles and P-values
14	Exercises 6.47, Extra, and 7.50
15	Summary notes

PRACTICAL INFORMATION

Today's lecture:

- another key concept from statistical inference:
 - * test (significance, P -value),¹
- inference for one sample (continuous data):²
 - * with or without assumption assumption of known σ ,
 - * a new distribution: the t -distribution,
- postpone material on sample size and power (Session 9),
- summary worksheet (S, Chapter 7, part),

Home assignment:

- deadline is anytime today (tonight) in my mailbox/office or by electronic submission on Moodle,
- returned (with marks, comments and solution) to you next Monday or Wednesday; possibly also brief discussion in class,
- 2nd home assignment in $1\frac{1}{2}$ weeks (Monday 19/2).

Recall: first lab review session today 1-2pm (351S, Companion Animals seminar room).

¹ PSLS 3e: Chapter 14-15 (parts); S: Chapter 8; IPS 7e: Sections 6.2-3.

² PSLS 3e: Chapter 17; S: Chapters 7-8; IPS 7e: Section 7.1.

TWO EXAMPLES OF TEST PROBLEMS

Example I: Testing of taste (example not in textbooks):

- aim: compare two brands of wine (beer, milk, cheese. . .),
- “duo-trio test” with one subject (person):
 - two anonymized samples, one of each brand,
 - third sample of known type,
 - subject may taste all 3 samples, as (s)he likes,
 - task: determine brand of two unknown samples,
- repeating the experiment, subject scores x out of n (e.g., 6 out of 8) correctly — how to determine if result has not occurred by chance (“luck”)?
- statistical problem because of randomness associated with “guessing” (even if qualified guessing).

Example II: Laboratory analysis of active ingredient in specimens: (Exercise 14.5 in PSLS 3e)

- data from 3 analyses of one specimen:
0.8403, 0.8363, 0.8447 (in g/l),
- aim: evaluate producer’s specified content of 0.86 g/l,
- statistical problem because of random measurement errors in laboratory.

INTRODUCTION TO STATISTICAL TESTING

Consider the “duo-trio” testing problem, and let X denote the number of “successes” for one subject in 8 trials.

- binomial setting $\Rightarrow X \sim$ binomial distrib. $B(8, p)$,
- if guessing, the probability p in each trial must be $p=0.5$ — state this as our *null hypothesis* $H_0: p=0.5$,
- under H_0 : $X \sim B(8, 0.5)$:

x	0	1	2	3	4	5	6	7	8
$P(X=x)$	0.004	0.03	0.11	0.22	0.27	0.22	0.11	0.03	0.004

- alternatively to H_0 we must have $p > 0.5$ (unless subject messes up the experiment) — state this as our *alternative hypothesis* $H_a: p > 0.5$,

If subject gets all trials right ($X = 8$):

- * probability of event happened by chance: $P = 0.004$,
- * by low P -value, we have little confidence in H_0 (because observed event unlikely to happen if H_0 was true)
 \Rightarrow *reject* H_0 and prefer H_a , (but H_0 could be true...),

If subject gets 6 out 8 trials right ($X = 6$):

- * probability of actual event or *more extreme* events:
 $P = P(X \geq 6) = P(X=6)+P(X=7)+P(X=8) = 0.14$,
- * by not too low P -value, observed $X=6$ does not seem unreasonable under H_0 (might have happened by chance)
 \Rightarrow *cannot reject* H_0 , (but H_0 could be false...).

COMPONENTS OF A STATISTICAL TEST

Statistical Model – main examples so far:

$X \sim B(n, p)$, and X_1, \dots, X_n i.i.d. (SRS) of population (μ, σ) .

Statistical Hypothesis:

- statement/assertion about the model (one or more parameters of the model) which is either true or false,
- null hypothesis H_0 — the one investigated,
- alternative hypothesis H_a — the one to hold if H_0 is not true.

Statistical Test statistic (or test variable):

- “measures” how well the data correspond to H_0 compared to H_a .

P-value (or significance probability):

- the probability, computed under H_0 (assuming H_0 is true), that the test statistic takes a value as extreme as or more extreme than (in the direction of H_a) the actually observed value from the data,³
- low P-values provide evidence against H_0
 \Rightarrow rejection of H_0 (and acceptance of H_a , *strong conclusion*),⁴
- high P-values provide no (convincing) evidence against H_0
 \Rightarrow H_0 cannot be rejected (*weak conclusion*).

Significance level α :

- *artificial* borderline/cut-off set for convenience between significant (i.e., $P \leq \alpha$) and non-significant (i.e., $P > \alpha$) results,⁵
- *by convention* set at 0.05, or less commonly at 0.10, 0.01, etc.

³ The P -value expresses how surprising the observed outcome would be if H_0 was true.

⁴ “If the P -value is low, the null hypothesis must go.” (Keith Bower; media links)

⁵ No uniform rule exists for whether ($P = \alpha$) is considered significant or not.

TEST FOR POPULATION MEAN

Setting for test of population mean:

- Model: X_1, \dots, X_n i.i.d. from distribution (μ, σ) ,
 - * assume (approximate) normal distribution of \bar{X} ,
 - * assume σ known (in practice, rarely a reasonable assumption).
- Null Hypothesis H_0 : $\mu = \mu_0$,
where μ_0 is a known, fixed value (very often, $\mu_0=0$),
- Alternative Hypothesis H_a : $\mu \neq \mu_0$,
- z test statistic computed as

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0,$$

- P-value computed as

$$P = 2 \times P(Z \geq |z|) = 2 \times P(Z \leq -|z|).$$

Example II: Laboratory analysis,

- Data: X_1, X_2, X_3 ; $n=3$, $\bar{X}=0.8404$, $\sigma=0.0068$ known,
- Hypotheses: $H_0: \mu = 0.86$, $H_a: \mu \neq 0.86$,
- Test statistic: $z = (0.8404 - 0.86)/(0.0068/\sqrt{3}) = -4.98$,
- P-value: $P = 2 \times P(Z < -4.98) < 2 \times 0.0002 = 0.0004$,
- Conclusion: reject H_0 and accept H_a ; strong indication that the specimen content is not as specified (i.e., lower).

ONE- OR TWO-SIDED?

Null hypothesis H_0 usually of the form: parameter=value (e.g. in the laboratory example: $\mu=0.86$).

Alternative hypothesis H_a usually one of 3 types:

- one-sided upwards: parameter > value (e.g., $\mu > 0.86$),
- one-sided downwards: parameter < value (e.g., $\mu < 0.86$),
- two-sided: parameter different from value (e.g., $\mu \neq 0.86$).

Choice of alternative hypothesis:

- one-sided: when *focus is on particular alternative* (because other direction is difficult to interpret or in beforehand of no interest),
- two-sided: *most common*, when no particular alternative is in focus or no knowledge is present in beforehand.

Alternative hypothesis affects calculation of P -values!

- in general, P -value is probability of extreme events for H_0 relative to (i.e., in the direction of) H_a ,
- example: testing for population mean, $H_0: \mu = \mu_0$,
 $H_a : \mu > \mu_0 : P = P(Z \geq z)$,
 $H_a : \mu < \mu_0 : P = P(Z \leq z)$,
 $H_a : \mu \neq \mu_0 : P = P(Z \geq |z|) + P(Z \leq -|z|)$.
- P -values and tests may also be termed one/two-sided.⁶

⁶ My recommendation is to only talk about one/two-sided alternative hypotheses.

TESTING BY CONFIDENCE INTERVAL

Fact: A confidence interval (CI) for a parameter with confidence level $C = 1 - \alpha$ can be used for a significance test at level α for the null hypothesis

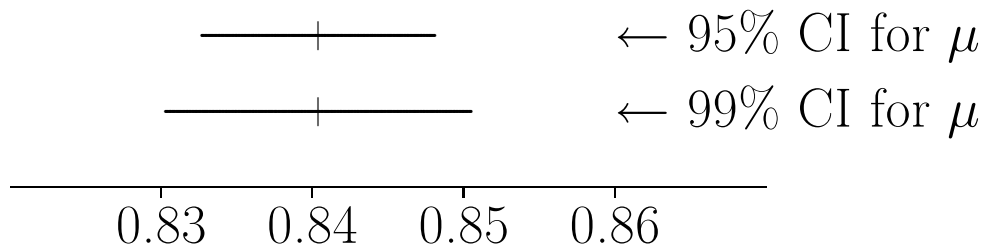
H_0 : parameter = value,
against the alternative hypothesis

H_a : parameter \neq value,
by the following “recipe”:

- reject H_0 , if value is outside interval.
- cannot reject H_0 , if value is inside interval,

Example II: Laboratory analysis,

- 95% CI for μ : $\bar{X} \pm 1.96 \sigma / \sqrt{3} = 0.8404 \pm 0.0077$,
 $\Rightarrow H_0: \mu = 0.86$, rejected at 5% level (also $H_0: \mu = 0.85$),
- 99% CI for μ : $\bar{X} \pm 2.576 \sigma / \sqrt{3} = 0.8404 \pm 0.0101$,
 $\Rightarrow H_0: \mu = 0.86$, rejected at 1% level (but not $H_0: \mu = 0.85$).



Advantages and disadvantages of testing by use of CI:

- + easy (when CI done), enhances CI interpretation,
- no P-value.

PITFALLS OF STATISTICAL TESTING

Some points to remember when using statistical tests⁷:

- the test/ P -value is only as good as its assumptions. . . .
- strictly speaking, the theory of statistical tests is for hypotheses determined in advance of data collection, and certainly *not inspired by the data*. . . ; also, carrying out many tests on the same data by pure chance may cause some of them to be statistically significant (the multiple testing problem),⁸
- to consider an analysis with $P = 0.049$ a success and discard an analysis when $P = 0.051$, is ridiculous. . .
(do not put too strong emphasis on significance levels, and always report P -values instead of just significance yes/no),
- non-significance does not mean proof of no effect:
*absence of evidence is not evidence of absence!*⁹
- non-significance may be important in itself (when not caused by insufficient or otherwise poor data),
- statistical significance does not imply biological significance or causation.

The bottom line is¹⁰: Statistical testing is often given too much attention in applied data analysis, where one should instead focus on the estimates, their precision (indicated e.g. by a confidence interval) and their interpretation, plus many other decisions prior to the final results.

⁷ Based on PSLS, Chapter 15, and IPS, Section 6.3.

⁸ Computing 20 tests, each with 5% error rate \Rightarrow 5% error rate? — or $5 \times 20\% = 100\%$ error rate? — the correct answer is in-between.

⁹ Quote usually attributed to Carl Sagan; or to William Cowper, 1731-1800.

¹⁰ You can find published articles recommending significance tests to be abolished, but this *is not* the current consensus; see e.g. homepage media links.

1-SAMPLE ESTIMATION

Data: sample X_1, \dots, 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),
- $EX_i = \mu$ and $\text{sd}X_i = \sigma$ for all X 's.

For estimation of σ we use the sample standard deviation:

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

and s^2 is an unbiased estimate of σ^2 . This explains our use of $(n-1)$ in the denominator of s^2 .¹¹

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	$\sigma_{\bar{X}}^2 = \sigma^2/n$	unbiased
<u>standard error of mean</u> ¹	s/\sqrt{n}	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$	biased, natural

¹ abbreviations: SE, $\text{SE}_{\bar{X}}$, SEM, s.e., ...

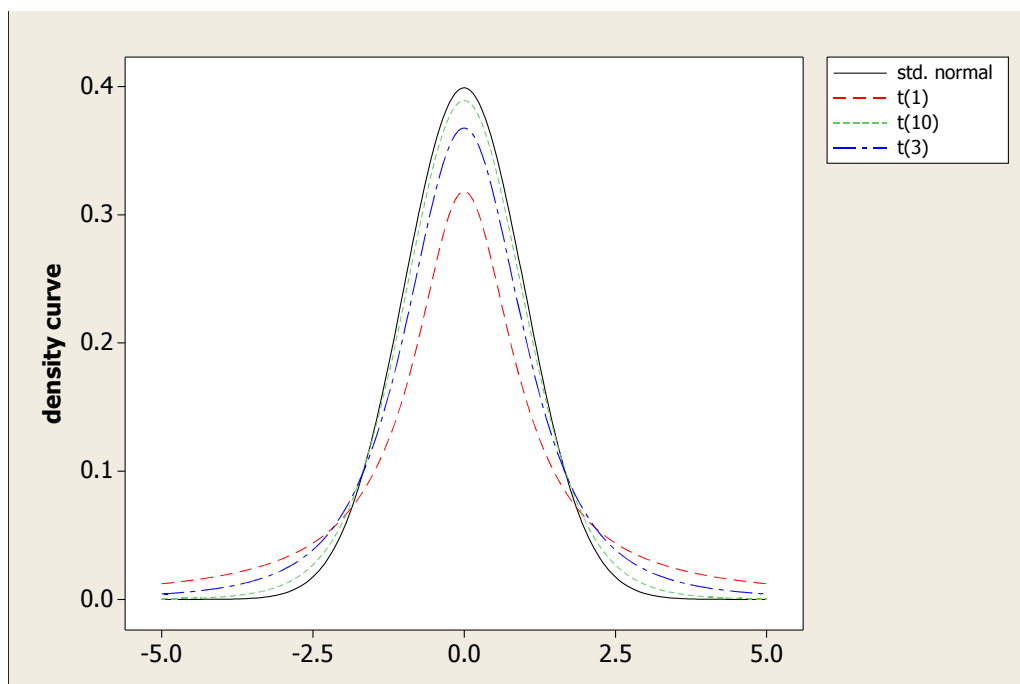
In addition, $\bar{X} \sim N(\mu, \sigma_{\bar{X}})$

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

¹¹ We skip over the (small) mathematical calculation showing that s^2 is unbiased.

(STUDENT'S) t DISTRIBUTION

- a new distribution — to be used for inference in a normal model when σ is estimated from the data:
— reference distribution for t test statistics,
- has a single parameter r (or “df”):
 - * $r = 1, 2, 3, \dots$
 - * 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 $N(0,1)$,
- when r large: $t(r) \approx N(0,1)$.



1-SAMPLE NORMAL DISTRIBUTION INFERENCE

- Data: X_1, \dots, X_n ($n =$ number of observations).
- Model: observations are a sample (i.i.d.) from $N(\mu, \sigma)$, where μ and σ are unknown parameters.
- Estimation: $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = s$.
- Distribution of estimates:

$$\hat{\mu} = \bar{X} \sim N(\mu, \sigma/\sqrt{n}), \quad s_{\bar{X}} = s/\sqrt{n},$$
$$(\bar{X} - \mu)/s_{\bar{X}} \sim t(n-1),$$

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 a suitable value from a $t(n-1)$ distribution (\rightarrow Table C of PSLS, Table 3 of S, Table D of IPS),¹²

- Test of $H_0: \mu = \mu_0$ against alternative H_a :
 - * test statistic: $t = (\bar{X} - \mu_0)/s_{\bar{X}}$,
 - * P -value from t distribution with $df = n - 1$:
 - $H_a: \mu \neq \mu_0$: $P = 2 \times P(t(df) \geq |t_{\text{obs}}|)$,
 - $H_a: \mu > \mu_0$: $P = P(t(df) \geq t_{\text{obs}})$,
 - $H_a: \mu < \mu_0$: $P = P(t(df) \leq t_{\text{obs}})$,
- note strong similarities with z -based procedures.

¹² Specifically, $t^* = t_{1-\alpha/2}(n-1)$ is the $(1-\frac{\alpha}{2})$ -percentile of a $t(n-1)$ distribution.

EXAMPLE: HUMAN BODY TEMPERATURE

Example 14.9 of PSLS 3e, Example p. 139 of S:

- Data: 130 measurements¹³ of body temperature in °F of healthy adults: X_1, \dots, X_{130} ($n = 130$).
- Model: a sample (i.i.d.) from $N(\mu, \sigma)$.
- Estimation: $\hat{\mu} = \bar{X} = 98.25$ and $\hat{\sigma} = s = 0.733$.
- Confidence interval with confidence level 95% ($\alpha = 0.05$):

$$\begin{aligned} \mu &: \bar{X} \pm t^* s_{\bar{X}} = 98.25 \pm 1.98 \times 0.733 / \sqrt{130} \\ &= 98.25 \pm 0.13 = (98.12, 98.38), \\ &\quad (t^* = t_{.975}(129) = 1.9785 \text{ from Minitab}), \end{aligned}$$
- Test of $H_0: \mu = 98.6$ against alternative $H_a: \mu \neq 98.6$: (“classical” average body temperature)
 - * test statistic: $t = \frac{\bar{X} - 98.6}{s_{\bar{X}}} = \frac{98.25 - 98.6}{0.733 / \sqrt{130}} = -5.45$,
 - * P -value from t distribution with $df = n - 1 = 129$:

$$P = 2 \times P(t(129) \geq 5.45) < 0.000001 \text{ (Minitab)}$$
 - * Conclusion: strong evidence to say that average body temperature is different, actually lower, than “classical” reference value.

¹³ Constructed data for pedagogical purposes (Shoemaker (1996), *Journal 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,¹⁴
- P -values are typically determined as tail probabilities in standard distributions like $N(0,1)$ or $t(df)$.

Methods to determine percentiles or tail probabilities:

- Minitab: Inverse Cumulative Probability for percentiles and Cumulative Probability for probabilities like $P(t \leq t_{\text{obs}})$ (1-tail probabilities),
- Stata: functions `normal`, `invnormal`, `ttail`, `invttail`,
- R: functions `pnorm`, `qnorm`, `pt`, `qt`,
- statistical tables: values for some confid./error levels.

What to do if df is not in table:

- use largest value below df
 \Rightarrow conservative analysis (larger CI's and P -values).

What to do if t_{obs} -value is not in table?

- find closest (“critical”) values in table, for example $t_1 < t_{\text{obs}} < t_2$, and use the relations

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

¹⁴ In statistical tables, the $p\%$ percentile is the critical value for a one-tailed $\alpha = 100-p\%$.

EXERCISES 6.47, EXTRA, AND 7.50

Exercise 6.47:

Null and alternative hypotheses for testing problems:

- (a) $H_0: \mu = 18$ and $H_a: \mu < 18$,
- (b) $H_0: \mu = 50$ and $H_a: \mu > 50$,
- (c) $H_0: \mu = 24$ and $H_a: \mu \neq 24$.

Extra Exercise (\sim 6.77 of IPS7e):

Explain in simple language why a test significant at the 1% level is also significant at the 5% level.

Some possible explanations:

- a P -value below 1% is also below 5%,
- an event occurring by chance with probability less than 1% also occurs with prob. less than 5%,
- it is “more difficult” (stronger requirement) to be significant at 1% than 5% level.

Exercise 7.50:

Percentiles/critical values for confidence intervals for population mean (with unknown population stand. 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).

SUMMARY NOTES

Key words and concepts:

- statistical test:
 - * concepts: null hypothesis H_0 , alternative hypothesis H_a (one or two-sided), test statistic and its (reference) distribution, P -value, significance level,
 - * possible conclusions: reject H_0 (and favour H_a), or no (insufficient) evidence against H_0 ,
 - * z -test formula for mean in normal distrib. (with known σ),
- relation between test and confidence interval (for single param.),
- common misconceptions related to statistical testing (see 6L–8),
- statistical inference for 1 sample on normal distrib. (unknown μ, σ):
 - * 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^*).

Four-step process for tests (PSLS 3e):

State: What is the practical question that requires a statistical test?

Plan: Identify a parameter, state the null and alternative hypotheses, and choose the type of test that fits your situation.

Solve: Carry out the test in three phases:

- * Check the conditions for the test you plan to use.
- * Calculate the test statistic.
- * Find the P -value using a table of Normal probabilities or technology.

Conclude: Return to the practical question to describe your results in this setting.