SV – Applied biostatistics

http://moodle.epfl.ch/course/view.php?id=14074

Lecture 1b

- Statistical modeling
- (Brief ! !) review : CLT, CI, hypothesis tests
- Review : hypothesis tests for μ, p
- Review : power and sample size
- Hypothesis tests, CI : comparison of two populations
- Student's t distribution, t-test

Statistical models

- A **statistical model** is an approximate mathematical description of the mechanism that generated the observations, which takes into account *unexpected random errors* :
    - gives an *idealistic* representation of reality
    - makes *explicit assumptions* (that could be **false** ! !) about the process under study
    - permits an *abstract* reasoning
- The model is expressed by a Le modéle s'exprime par une *family of theoretical distributions* that contains the 'ideal' cases for the included RVs
    - e.g. : tosses of a coin **...**
- A useful model offers a *good compromise* between
    - *true* description of the reality (many parameters correct assumptions)
    - *ease* of mathematical manipulation
    - production of solutions/predictions *close* to the observation(s)

A simple model

A simple case : several measures of a physical quantity μ are taken, e.g. length of a field, person's height *...*

- Such measures possess in general a *random* component due to *measurement errors*
- One possible error mechanism :

$$\begin{array}{cccccc} \text{measure} & = & \text{true theoretical value} & + & \text{measurement error} \\ y & = & \mu & + & \epsilon \end{array}$$

- that is : measures with *additive errors*
- If there is no colitsystematic error (biais), the random error should be 'centered' ($E[\epsilon] = 0$)
- Often reasonable to think that *the precision* of each measure is *the same* ($Var(\epsilon) = \sigma^2$ for each measurement)
- *One possible specification* for the error distribution is *Normal* $N(0, \sigma^2)$
- **All models are wrong ; some are useful**

Estimation of the unknown parametres

- Once a model is chosen, we are interested in estimating unknowns : *the parameters of the model*
- We observe *realizations* of a RV for which the distribution is known (other than the parameter values)
- Thus, we must *estimate* the parameters using the observations $X_1, \ldots, X_n$
- $\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- The estimator S^2 is *unbiased* for σ^2, and is *independent* of that for μ ($\overline{X}$)

Review : Central Limit Theorem (CLT)

- The **Central Limit Theorem** is one of the most important results in probability/statistics, and is widely used as a problem-solving tool
- **Theorem (CLT)** : Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) RVs, each having mean μ and variance σ^2
- Then for n 'sufficiently large', the distribution of
  - the **sum :** $\sum_{i=1}^{n} X_i$ is approximately $N(n\mu,\, n\sigma^2)$
  - the **mean :** $\overline{X}$ is approximately $N(\mu,\, \sigma^2/n)$

Review : Confidence intervals

Suppositions for CIs :

1. There is an *unknown* population parameter
2. There is a *random sample* (independent observations or SRS from a large population, where the sample size is small compared to the population size)
3. We can apply the CLT

Mechanics :

- CI for the population *mean* : $\overline{x} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ (use s instead of σ if σ is unknown)
- CI for the population *proportion* (or *percentage*) : $\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

Review : steps in hypothesis testing

1. **Identify** the population parameter being tested
2. **Formulate** the NULL and ALT hypotheses
3. Compute the **test statistique (TS)**
4. Compute the *p*-**value** p_{obs}
   - p_{obs} is the probability of obtaining a value of T *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
5. *Decision rule and practical interpretation* : REJECT the NULL hypothesis H if $p_{obs} \leq \alpha$

Test of comparison on 2 independent samples

- Until now, we have been interested by *a single population*. Often, however, we are interested in the **comparison of two populations**. In this case, we carry out a test on *two independent samples*.
- When we compare two *means* (or *proportions*) the basic notion is the same as above : for T, we use the *standardized difference* between the sample means (or proportions).
- TS for the *difference in means* from two independent populations : $\dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$
  (use s instead of σ if σ is unknown)
- TS for the *difference in proportions* from two independent populations : $\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$

Regarding small samples...

- The z-test that we have studied assumes that the sampling distribution of the test statistic T is *Normal*
    - exactly, or
    - approximately, by the CLT
- However, if the population SD σ is *unknown* and the sample size is *small* (for example, under 30) then the true sampling distribution of T has *heavier tails* than the Normal distribution
- In this case, you should use the *t-test*

‘Student’ (= William Sealy Gosset)

W. S. Gosset

Guinness

Distribution of T when σ^2 is unknown

- Recall the test statistic $T = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$
- **If** the sample size n is 'sufficiently large', then under H, $T \sim N(0,1)$ *regardless of the distribution of X* (CLT)
- **If** the observations $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$, then $T \sim N(0,1)$ for *known* σ^2, *regardless of the sample size n*
- **BUT :** If the sample size n is *small*, and the variance σ^2 is *unknown*, the *true* distribution of T has *more variability* than the Normal distribution (due to the *imprecise* estimation of σ based on few obs)
- For the case **(1)** $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$; **(2)** n small ; and **(3)** σ^2 is unknown, then $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, the Student t distribution, with $n-1$ *degrees of freedom* (df)
- The distribution de T depends on the number of observations n)

Student t distribution

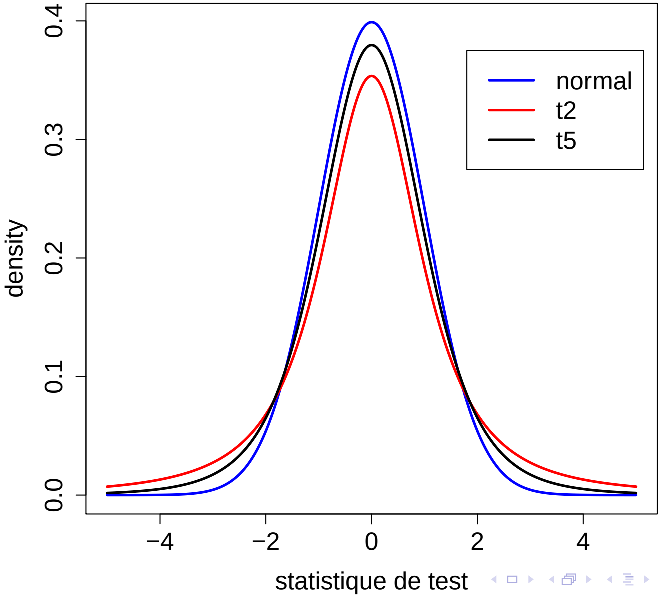


Table of the t distribution

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Confidence interval

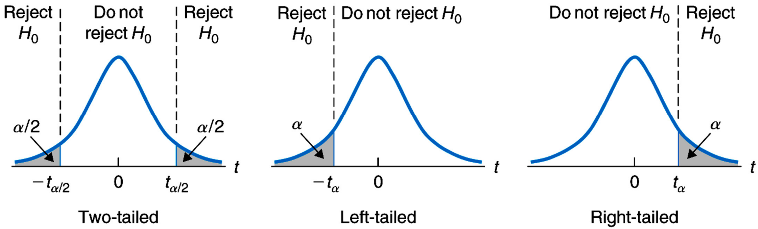
In the case

1. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
2. n small ; and
3. σ^2 is unknown :

- we can make a *confidence interval (CI)* as before, but **using the t distribution instead of the Normal (z)**
- CI for the population *mean* : $\overline{x} \pm \boxed{\mathbf{t_{n-1,1-\alpha/2}}} \boxed{\mathbf{s}} / \sqrt{n}$

Hypothesis test : find the rejection region

H: $\mu = \mu_H$	H: $\mu = \mu_H$	H: $\mu = \mu_H$
A: $\mu \neq \mu_H$	A: $\mu < \mu_H$	A: $\mu > \mu_H$



Test for comparing two (independent) means : equal variances

- We want to compare the means of two sets of measures :
  - Group 1 (p. ex. 'control') : $x_1, \ldots, x_n$
  - Group 2 (p. ex. 'treatment') : $y_1, \ldots, y_m$
- We can *model* these data as :

  $$x_i = \mu + \epsilon_i; i = 1, \ldots, n;$$
  $$y_j = \mu + \Delta + \tau_i; j = 1, \ldots, m,$$

  where Δ signifies the effect of the treatment (compared to the 'control' group)
- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ or $A : \Delta > 0$ or $A : \Delta < 0$

Equal variances, cont.

- T = obs. diff. / ES(obs. diff.) $= \dfrac{\Delta}{\sqrt{\widehat{Var(\hat{\Delta})}}}$;
  $\hat{\Delta} = \bar{y} - \bar{x}$; $Var(\hat{\Delta}) = \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m} = \dfrac{n+m}{nm}\sigma^2$
- We assume that :
  - the variances of the 2 samples are *equal* : $Var(\epsilon) = Var(\tau)$
  - the observations are *independent*
  - *the 2 samples are independent*
- We can estimate the variances *separately* :
  $s_x^2 = ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)/(n-1)$
  $s_y^2 = ((y_1 - \bar{y})^2 + \cdots + (y_m - \bar{y})^2)/(m-1)$
- When the variances are *equal*, we can combine the two estimators : $s_p^2 = ((n-1)s_x^2 + (m-1)s_y^2)/(n+m-2)$

$$\Rightarrow t_{obs} = \frac{\bar{y} - \bar{x}}{\sqrt{s_p^2(n+m)/(nm)}} \sim t_{n+m-2} \text{ under } H$$

Test for comparing two (independent) means : unequal variances

- If $\sigma_x^2 \neq \sigma_y^2$, we can use

$$T_{Welch} = \frac{\overline{Y} - \overline{X}}{\sqrt{S_x^2/n + S_y^2/m}}$$

- The distribution of the statistic T_{Welch} *is only approximately* t, with a number of degrees of liberty calculated based on s_x, s_y, n and m
- Welch test
- In practice, if the variances are rather different (ratio more than 3), we could use this statistic (instead of the one with variance s_p^2)

Paired experiments

- For an experiment carried out in *blocks of two units*, the *power* of the t-test can be increased
- This idea permits us to *eliminate the influences of other variables* (e.g. age, sex, *etc.*), in giving them different 'treatments'
- Thus, we have a *more precise* comparison of the two conditions

t-test for a paired experiment

- The data are of the form :

	1	2		n	
contrôle	x_1	x_2	$\cdots$	x_n	expected value μ
traitement	y_1	y_2	$\cdots$	y_n	expected value $\mu + \Delta$

- *Each block* allows us to evaluate the effect of the treatment
- Here, we consider *the differences*

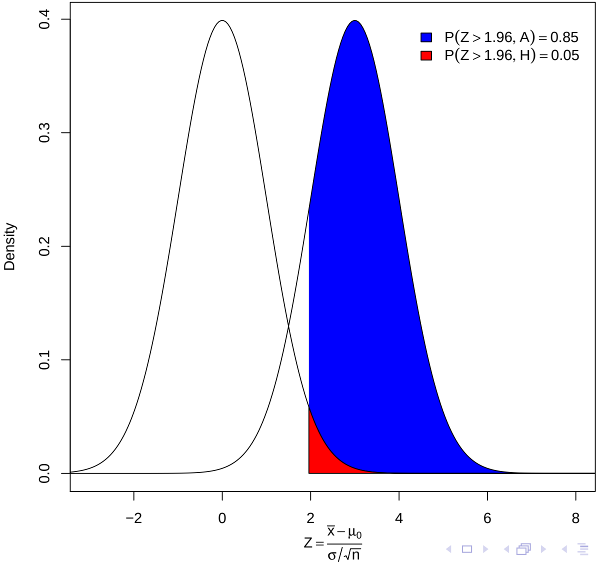
$$d_1 = y_1 - x_1, \ldots, d_n = y_n - x_n$$

  as a sample of measurements coming from a distribution with expected value Δ
- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ or $A : \Delta > 0$ or $A : \Delta < 0$
- $T = t_{paired} = \frac{\overline{d}}{s_d/\sqrt{n}}$, where $s_d^2 = ((d_1 - \overline{d})^2 + \cdots + (d_n - \overline{d})^2)/(n-1)$
- Under H, $t_{paired} \sim t_{n-1}$

Hypothesis truth vs. decision

Truth \ Decision	not rejected	rejected
true H	☺ specificity	X Type I error (False +) α
false H	X Type II error (False -) β	☺ Power $1 - \beta$; sensitivity

Power



P(Z > 1.96, A) = 0.85
P(Z > 1.96, H) = 0.05
Density
0.0
0.1
0.2
0.3
0.4
−2
0
2
4
6
8
$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Example (see power example)

A tire company has developed a new tread design. To determine if the newly designed tire has a mean life of 60,000 miles or more before it wears out, a random sample of 16 prototype tires is tested. The mean tire life for this sample is 60,758 miles. Assume

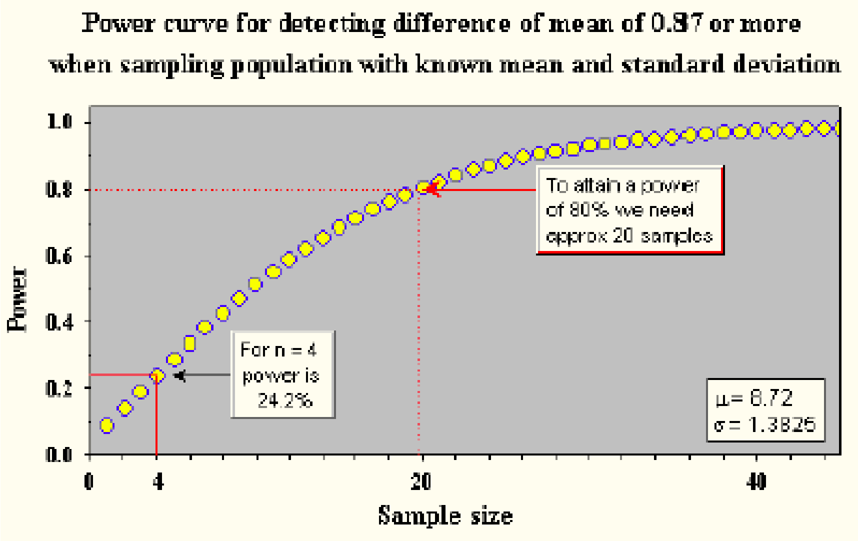
that the tire life is normally distributed with unknown mean μ and (known) SD $\sigma = 1500$ miles.

(a) Test the hypotheses at $\alpha = 0.01$. What do you conclude *??*

(b) What is the *power* of the test if the true mean life for the new tread design is 61,000 miles *??*

(c) Suppose that at least 90% power is needed to identify a design that has mean wear of 61,000 miles. How many tires should be tested *??*

Power curve



Power curve for detecting difference of mean of 0.87 or more
when sampling population with known mean and standard deviation
1.0
0.8
0.6
0.4
0.2
0.0
Power
0
4
20
40
Sample size
To attain a power of 80% we need approx 20 samples
For n = 4 power is 24.2%
μ= 8.72
σ = 1.3825