36-309/749
Experimental Design for Behavioral and Social Sciences

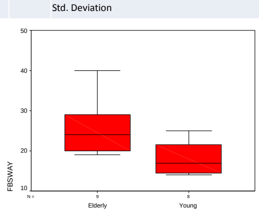
Sep. 8, 2015
Lecture 2: Statistical Background

Case Study

How difficult is it to maintain your balance while concentrating? It is more difficult when you are older? Nine elderly people (6 men and 3 women) and eight young men were subjects in a quasi-experiment. Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position and to react as quickly as possible to an unpredictable noise by pressing a hand held button. The noise came randomly and the subject concentrated on reacting as quickly as possible. The platform automatically measured how much each subject swayed in millimeters in the forward/backward directions. (slightly fudged)

Descriptives

	Age			Statistic	Std. Error
FBSway	Elderly	Mean		25.22	2.296
		95% Confidence Interval for Mean	Lower Bound	19.93	
			Upper Bound	30.52	
		Median		24.00	
		Std. Deviation		6.888	
	Young	Mean		18.13	1.445
		95% Confidence Interval for Mean	Lower Bound	14.71	
			Upper Bound	21.54	
		Median		17.00	
		Std. Deviation		4.086	



SPSS Compare Means / Independent Sample t-test:

Group Statistics

	AGE	N	Mean	Std. Deviation	Std. Error Mean
FBSWAY	Elderly	9	25.222	6.8880	2.2960
	Young	8	18.125	4.0861	1.4447

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FB Sway	Equal variances assumed	1.244	.282	2.54	15	.023	7.10	2.80	1.14	13.06
	Equal variances not assumed			2.62	13.2	.021	7.10	2.71	1.25	12.95

Review of Probability and Statistics: Principles and Definitions

A. Random variable (§3.1)

- Usually represented by a capital letter near the end of the alphabet, e.g., X or X_1, X_2, …. Represents something specific, e.g., a height.
- Value is unknown (before the experiment is run).
- Has a probability **distribution**, rather than a particular value. (§3.2)
  - Characteristics of distributions: central location, spread, shape (§3.5)
  - Discrete or categorical: probability mass function (pmf)
  - Continuous: probability density function (pdf)
- Mathematical combinations → new random variables, e.g. $R=X^2$, S=X+Y, T=X/Y+3Z, U=T-3, $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$

B. Population vs. **Sample** (§3.4)

All possible … vs. the ones we are studying

C. Parameter refers to fixed, unknown quantities in the population (§3.5)

- "secrets of nature" with scientific meaning
- Usually represented by Greek letters.

D. Statistic: a quantity unknown before an experiment is run and fully calculable from sample data afterwards

E. Mean: one measure of central location

- Population mean (expected value): μ (§3.5.1)
- Sample mean: (§4.2.3) $\bar{Y} = (\sum_{i=1}^{n} Y_i)/n$
- ***Big idea***: One μ, many possible $\bar{Y}$'s

F. Variance: one measure of spread

- Average of the squares of the deviations (values minus mean)
- Population: σ^2
- Sample: $s^2 = \mathrm{SS}/\mathrm{df}$
  - SS is the "**sum of squares**" which is really the sum of squared ***deviations*** of the values from their sample mean
  - df is "degrees of freedom" which is the number of independent pieces of information in a calculation (§4.6)
  - $s^2 = \mathrm{Var}(Y_{1,\ldots,} Y_n) = \mathrm{Var}(\boldsymbol{Y}) = \frac{\sum_{i=1}^{n}[\qquad\qquad]}{}$
- **Standard deviation:** square root of variance (back to the natural scale)

G. Conditional distribution, e.g, of sway or mean sway ***given*** (only for) an age group

H. Sampling distribution of a statistic: the distribution of a statistic over (theoretical or actual) repeats of an experiment (§3.6). This is the most important concept in (non-Bayesian) statistical analyses! The standard deviation of any statistic is called its **standard error (SE)**.

I. **Key example: <u>sample mean statistic</u>, $\bar{Y}_n$**

a) Setup: Y_1, $Y_{2,\ldots}$ are iid (independent and identically distributed) measurements from a population with mean μ and variance σ^2 and any shape of distribution.

b) Consider repeatedly sampling n random values of Y and computing and recording $\bar{Y}_n$.

c) Mean of sample means: For randomly sampled iid data, the sampling distribution of $\bar{Y}_n$ has mean μ.

d) Variance of sample means: For randomly sampled iid data, the sampling distribution $\bar{Y}_n$ has a variance of σ^2/n (and standard deviation (SE) equal to $\sigma/\sqrt{n}$).

Example: US adult non-diabetic fasting glucose has population mean μ=85 mg/dL and population variance σ^2=49 mg^2/dL^2. What are the mean, variance, and standard error of the sampling distribution of the mean of samples of 100 randomly chosen non-diabetic US adults?

e) <u>Shape</u>: If the distribution of Y is Gaussian then the sampling distribution of $\bar{Y}_n$ is Gaussian (regardless of sample size, n).

I. Sampling Distribution of $\overline{Y}_n$, cont.

f) Shape: Key result for "non-bizarre" distributions: (§3.8)
Even if the distribution of Y is non-Gaussian, the sampling distribution of $\overline{Y}_n$ tends towards a Gaussian shape as n gets large. This is the **central limit theorem** (CLT). And then we can say, e.g., 68% falls inside mean +/- 1s.d. and 95% falls inside +/-2 s.d.

g) Summary: With a reasonable sample size $\overline{Y}_n$ is approximately distributed as Gaussian with mean μ and variance σ^2/n.

J. **Overall Goal of Statistical Inference**: Sampling distribution of a statistic → inference about populations

Standard Approach of "Classical" Statistics

- Our ***goal*** is to learn about ***populations*** from samples.
- The basic approach of standard **statistical hypothesis testing** is as follows (§6.2.1):
  1) Frame a (tentative) model describing the relationship between the explanatory variables (IVs) and the outcome variable (DV) in the population and the nature of the variability in the DV at any fixed combination of IVs. Define the parameters of the model. State all of your **model assumptions**. (§6.2.2)
  2) Specify the **null and alternative hypotheses** in terms of the ***parameters*** of the model. (§6.2.3)
  3) Choose your acceptable **type 1 error rate** (α), i.e., the probability of falsely rejecting the null hypothesis when it is actually true.
  4) Choose (or invent) a ***statistic*** that will tend to be different under the null and alternative hypotheses. (§6.2.4)

Steps of hypothesis testing, cont.

5) Using the assumptions of step 1), find the theoretical ***sampling distribution*** of the ***statistic*** under ***the null hypothesis***. (§6.2.5) Ideally the form of the sampling distribution will be one of the "standard distributions". Usually there is a "family" of distributions, and constants such as sample size and number of treatment conditions are used to choose which member of the family is applicable.

6) Calculate a **p-value** as the area under the null sampling distribution more extreme (un-null-like) than your observed statistic. (§6.2.6)

7) Apply the **decision rule**: reject the null hypothesis if the p-value is less than alpha; otherwise do not reject. Eschew the word "accept"! (§6.2.6)

Interpretation of p-values

- All interpretation is ***meaningless if the model assumptions are not reasonably well met***.
- A p-value ***cannot*** be used to make any probabilistic statements about the chance that H_0 is true or false because it comes from a calculation that assumes the null hypothesis is true. Also the size of the p-value does not tell us if the effect of treatment is large or small.
- A ***small p***-value, e.g., ≤0.05, indirectly adds support to the claim that H_A is likely. If model assumptions are not violated, the other main possibility is the "bad luck" of a randomly unusual value of the test statistic (a **type 1-error**). But, a small or even tiny p-value does ***not*** tell us that a ***meaningfully*** large alternative (e.g., a treatment effect) is likely, especially when the sample size is large. [Concern for a future class: multiple testing]
- A ***large p-value*** indicates either that H_0 is true or that we have made a **type-2 error** due to bad luck. When coupled with an appropriate **power analysis**, a large p-value is good evidence that a meaningfully large alternative is unlikely. Without a power analysis, even a quite large p-value could be consistent with a meaningfully large treatment effect!
- Never claim that any p-value ***proves*** anything!

Confidence intervals (§6.2.7) for parameters

- Statistical hypothesis testing is only one way to achieve the goal of learning about populations from samples. Another equally important approach is calculation of CIs.
- Technically, a 95% CI is a random interval which over repeat experiments holds the one true parameter value 95% of the time, if the model assumptions are true.
- For example, if the parameter of interest is the (population) mean sway for elderly minus that for young people, a 95% CI for that parameter of [1.1,13.0] mm tells us that there probably is a real difference, but using the available data we are rather unsure of the size of the different.
- CIs that are wide (by human judgment) tell us that the experiment was not powerful enough to provide strong information, e.g., about treatment effects.
- Narrow CIs let us make scientifically useful conclusions (null or non-null).

Applying the principles to the independent-samples t-test

- Statistical model: k=2 "treatment" conditions. Assume the two samples come from a population where the DV has a ***Normal*** distribution for both conditions with ***common*** variance σ^2 and with means μ and $\mu+\delta$. Assume independent **errors**. (§6.2.2)
- Hypotheses: H_0: $\delta=0$, H_1: $\delta\neq0$ (or H_0: $\mu_1=\mu_2$, H_1: $\mu_1\neq\mu_2$) (§6.2.3)
- **T-statistic** (§6.2.4)

General form: $\mathrm{T} = \dfrac{\text{statistic} - \text{hypothesized parameter value}}{\text{estimated SE(statistic)}}$

For the independent samples t-test, under H_0: $\delta=0$, use

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where s_p is a "pooled" estimate of the standard deviation of Y (not $\bar{Y}$).

t-test, cont.

- Sampling distribution: The t-statistic follows the so-called "t-distribution" with n_1+n_2-2 df under the assumptions of this model and when the null hypothesis is true. (§6.2.5)
- Calculate p: For our sway quasi-experiment, t-statistics more extreme (un-null-like) than 2.539 are those that are bigger than 2.539 or smaller than –2.539. These ranges correspond to 2.3% of the area under the null sampling distribution, so p=0.023. (§6.2.6)
- With $\alpha=0.05$, $p\leq\alpha$, so the decision rule says to reject the null hypothesis. We conclude that results like these are unusual under the null hypothesis (and when the assumptions are true), and this is indirect supporting evidence for the idea that the population means of the two groups really are different ($\delta\neq0$). It is also possible that we are making a **type 1 error** (falsely rejecting the null hypothesis). The small p-value tells us nothing about the effect size. (§6.2.6)

t-test cont.

- SPSS output:

Independent Samples Test

		t-test for Equality of Means					
						95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
FBSWAY Equal variances assumed	2.539	15	.023	7.097	2.7954	1.1390	13.0554

- **Confidence interval** for the mean difference. (§6.2.7)

A rough confidence interval can, ***in general***, be constructed as:
statistic +/- m · SE(of the statistic)
where m=2 is the approximate "multiplier".

Use the "quantiles" of the null sampling distribution to get the exact multiplier. For the t distribution with 15 df, 95% of the values are between –2.13 and +2.13.

The $1-\alpha$ or $100(1-\alpha)$% confidence interval (CI) for the difference δ or $\mu_2-\mu_1$ is obtained using the multiplier m=2.13 and $\bar{D} = \bar{Y}_2 - \bar{Y}_1$ as:
$[\bar{D} - m \bullet SE(\bar{D}),\ \bar{D} + m \bullet SE(\bar{D})]$

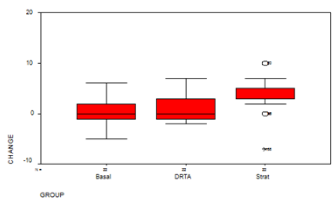
With $\bar{D}$=7.097 and SE($\bar{D}$)=2.7954, the 95% CI is [1.14, 13.05].

Interpretation: We are "95% confident" that the true difference, δ, is between 1.14 and 13.05. This is shorthand for:

One way ANOVA example

Researchers at Purdue University conducted an experiment to compare three methods of teaching reading. Students were randomly assigned to one of the three teaching methods, and their reading comprehension was tested before and after they received the instruction. The change in score for a particular reading comprehension test from the pre- to the post-test (post minus pre) is recorded.

EDA:



One-way ANOVA, cont.

Descriptives

CHANGE

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Basal	22	.2727	2.58534	.55120	-.8735	1.4190
DRTA	22	1.1364	2.58743	.55164	-.0108	2.2836
Strat	22	3.4091	3.21691	.68585	1.9828	4.8354
Total	66	1.6061	3.07285	.37824	.8507	2.3615

A "formal" test of the equal variance assumption:

Test of Homogeneity of Variances

CHANGE

Levene Statistic	df1	df2	Sig.
.031	2	63	.970

A "formal" of H_0 vs. H_A:

ANOVA

CHANGE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	115.485	2	57.742	7.301	.001
Within Groups	498.273	63	7.909		
Total	613.758	65			

One-way ANOVA, cont.

- Model: The three samples come randomly from a population where the DV has a ***Gaussian*** distribution in each group, with a ***common*** variance σ^2, and with means μ_1, μ_2, and μ_3. (We use k=3 for the number of groups and n=22 for the number of subjects ***per*** group, and N=66 for the total sample size.) Errors (true individual deviations from group population means) are ***independent***. (§7.2.1)
- Null Hypothesis is H_0: $\mu_1=\mu_2=\mu_3$. Alternative Hypothesis is H_1: at least one mean is different from the others. (Definitely ***wrong***: H_1: $\mu_1 \neq \mu_2 \neq \mu_3$) (§7.2.1)
- Statistic: $F = MS_{between} / MS_{within}$ (Also, MS = SS / df.) (§7.2.2)
- Null Sampling Distribution of the F-statistic under the model assumptions: F distribution with k-1=2 numerator df and k(n-1)=3(21)=63 denominator df. (Or N-k=66-3=63.) (§7.2.3)
- p-value: In this experiment with F=7.30, p=0.001, which comes from:
- Decision rule: Because $p=0.001 < \alpha=0.05$, we reject the null hypothesis and conclude that at least one group has a different ***population*** mean from the others.

Basic Theory of One Way ANOVA:

ANOVA is a technique to detect group differences in ***means*** by using variance-like quantities (MS=SS/df) as a tool.

Statistic	Average value when H_0 is true	Average when H_0 is false
MS_{within}	σ^2	σ^2
$MS_{between}$	σ^2	Bigger than σ^2

$F = MS_{between} / MS_{within}$

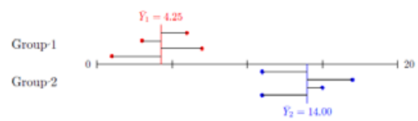
Expected value of F under H_0:

Expected value of F under various H_1's:

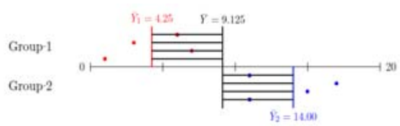
For the reading comprehension example, the p-value ("Sig.") is the area under the $F_{2,63}$ distribution that is to the right of (higher than) 7.301.

Intermediate Theory of ANOVA

SS_{within}:



$SS_{between}$:



Interpreting the ANOVA table: (§7.4)

ANOVA

CHANGE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	115.485	2	57.742	7.301	.001
Within Groups	498.273	63	7.909		
Total	613.758	65			

Class Summary

- A statistic is chosen for making an inference about a hypothesis because it has different (but overlapping) "null" and "alternate" distributions, and because its null sampling distribution can be determined based on the assumptions of the statistical model.
- One-way ANOVA and the independent samples t-test use the F and t statistics respectively to make inference for categorical explanatory variables and quantitative outcomes. They assume independent errors and an underlying Gaussian distribution with equal variances.
- The t-test only handles 2 levels of the categorical explanatory variable (factor), while the ANOVA handles ≥2. They agree completely with t^2=F when there are 2 levels.