Data analysis

Introductory on Field Epidemiology
6 July 2015, Thailand

Types of Variables

- Continuous variables:
  - Always numeric
  - Generally calculate measures such as the mean, median and standard deviation
- Categorical variables:
  - Information that can be sorted into categories
  - Often interested in dichotomous or binary (2-level) categorical variables
  - Cannot calculate mean or median but can calculate *risk*

Parametric and nonparametric tests of significance

Parametric test of significance - to estimate at least one population parameter from sample statistics

Assumption: the variable we have measured in the sample is *normally distributed in the population* to which we plan to generalize our findings

Nonparametric test - *distribution free*, no assumption about the distribution of

Summary Table of Statistical Tests

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X2 or bi-nomial	X2	Macnarmar's X2	X2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r

(Plonskey, 2001)

Parametric Test Procedures

1. Involve Population Parameters (Mean)

2. Have Stringent Assumptions (Normality)

3. Examples: Z Test, t Test, χ^2 Test, F test

Standard Normal (Z) Distribution

- There are many Normal distributions depending on μ and σ
- The **Standard Normal (Z) distribution** is a Normal distribution with mean 0 and standard deviation 1
- **Notation**: Z~N(0,1)
- If variable X has normal distribution, we **standardize** a value x_i by subtracting μ and dividing by σ. This is called a **z-score**.
- The **z-score** of a value tells you how many times of σ the value falls far from μ in either a positive or negative direction

$$z = \frac{x - \mu}{\sigma}$$

Standardization (z score): Example

- What is the z score of weight of 68 kg. if weight variable has normal distribution with mean of 70 kg and SD of 2.8 kg?
- $z = (x - \mu) / \sigma$

  $= (68 - 70) / 2.8$

  $= -0.71$
- This means that 68 is 0.71 standard deviations *below* the mean of the distribution
- Probabilities of Standard Normal (Z) Distribution have been tabulated (See z-table in your textbooks). So we can estimate probability of getting values more or less than a specific value of interest.

Z Distribution is limited

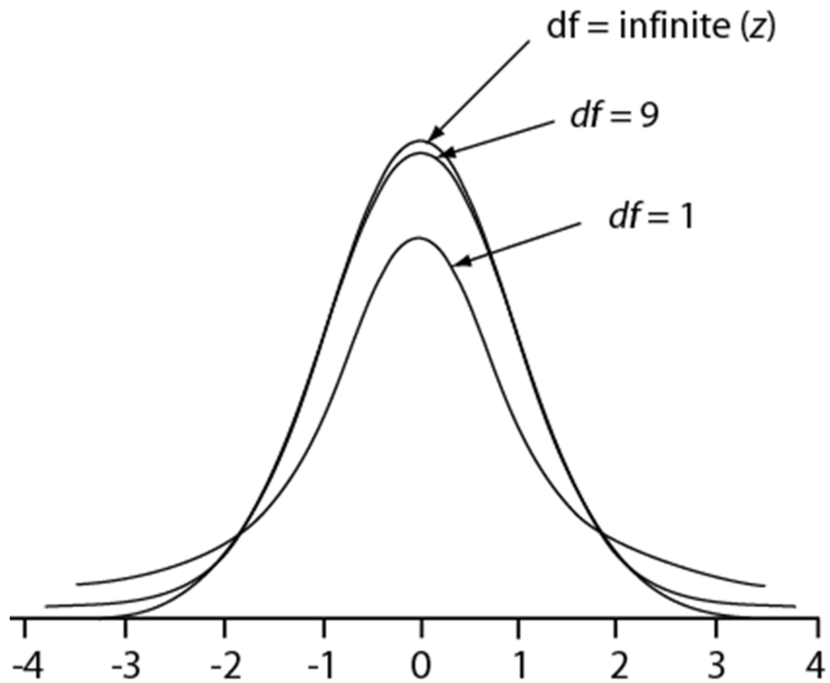
- The **Standard Normal (Z) distribution** is a Normal distribution with mean 0 and standard deviation 1
- If variable X has normal distribution, we can calculate **z-score**:

$$z = \frac{x - \mu}{\sigma}$$

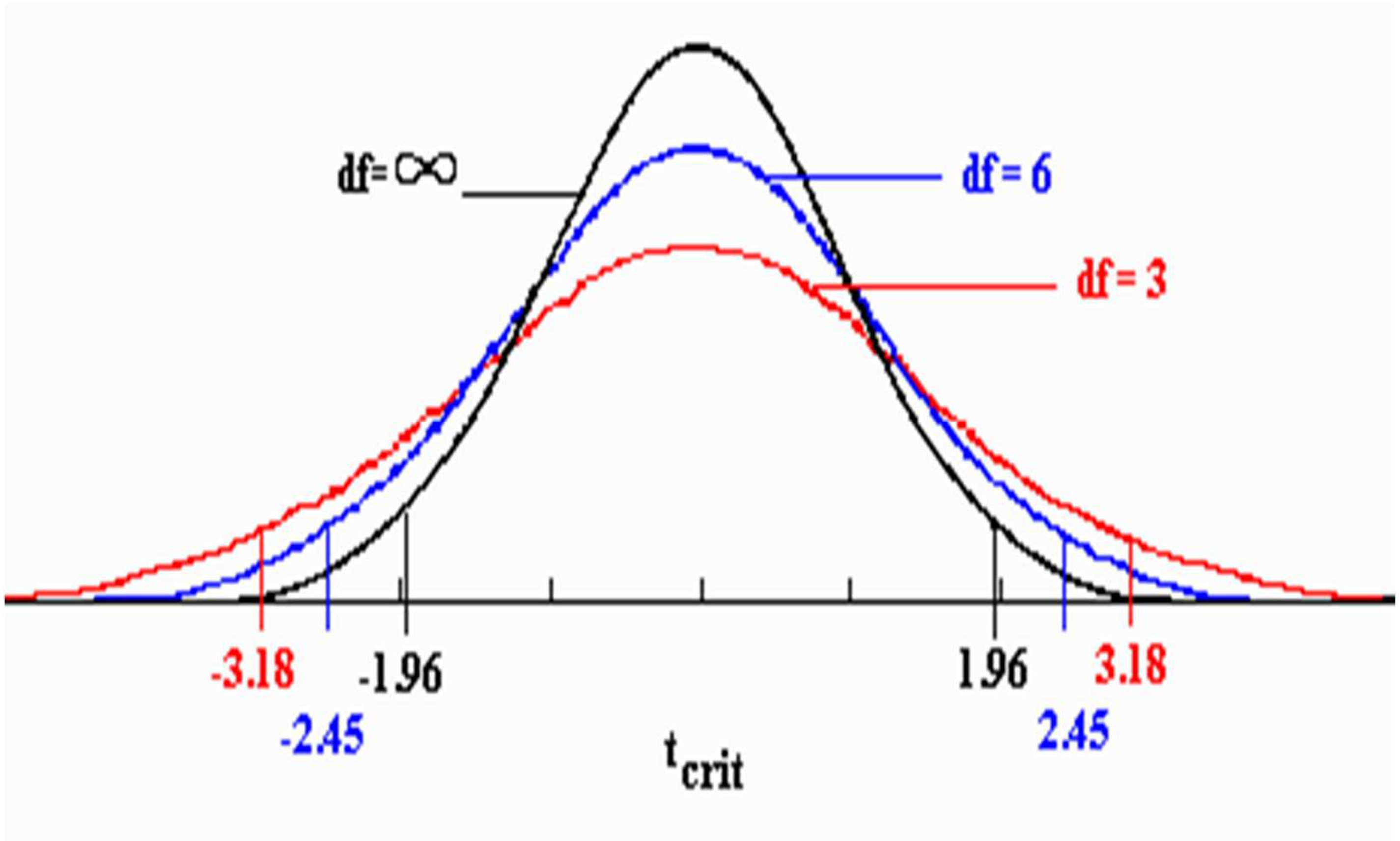
- In real life, we never collect data of all population. μ and σ are unknown.
- But we have sample mean $(\bar{x})$ and standard deviation (S)
- Fortunately, there is another type of distribution/model that we can use $\bar{x}$ and S and it is quite similar to Z distribution

Student's *t* distributions

- Similar to the Standard Normal (*"z"*) distribution but with broader tails
- It is not a single distribution but a family
- Family members identified by sample size minus one (n-1), called "degrees of freedom" (*df*)
- As *df* increases → tails get skinnier → *t* become more like *z* → *t* with ∞ *df* = z



Student's *t* distributions



- Greater the df, the more closely the t distribution matches the z distribution
- Main impact of df is that higher critical values of t are needed to reject H0 with smaller df

Table A.5. Percentiles of the t-Distribution

df	90%	95%	97.5%	99%	99.5%	99.9%
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.183	4.541	5.841	10.215
4	1.533	2.132	2.777	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.708	5.208
7	1.415	1.895	2.365	2.998	3.500	4.785
8	1.397	1.860	2.306	2.897	3.355	4.501
9	1.383	1.833	2.262	2.822	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.625	2.977	3.787
15	1.341	1.753	2.132	2.603	2.947	3.733
16	1.337	1.746	2.120	2.584	2.921	3.686
17	1.333	1.740	2.110	2.567	[illegible]	[illegible]
18	1.330	1.734	2.101	2.552	2.879	3.611
19	1.328	1.729	2.093	2.540	2.861	3.580
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.788	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	[illegible]
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.705	3.307
80	1.292	1.664	1.990	2.374	2.639	3.195
∞	1.282	1.645	1.960	2.326	2.576	3.090

One-Tailed Percentages

So, with 17 df…
- 2.5% of scores fall above 2.11
- 2.5% fall below -2.11
- ±2.11 critical two-tailed t-value for α = .05

With large sample n, *t*-distribution critical values approach *z*
- ±*1.96* is two-tailed value for α = .05
- ±2.58 is two-tailed value for α = .01

Student's *t* distributions

- Like z distribution,
  - we can standardize the value of observed sampling mean with expected population mean
- Like z score,
  - Basic formula is: $$t_{stat} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$
- We call the score "*t* statistic", and the test "*t* test"

Type of *t* Test

1. One-sample *t* test
   - Is sample mean equal to a hypothetical value?
   - No comparison group.
   - E.g. Is mean height of students from school A equal to 160 cm.?
2. Paired *t* test
   - Are sample means of two related samples (paired samples) equal?
   - E.g. Are mean scores of June course participants equal between pre-test and post- test?
3. Independent (unpaired) *t* test
   - Are sample means of two independent samples equal?
   - E.g. Are mean scores of students from school A equal to between mean scores of students from school A ?

What Type of Sample?

1. Measure vitamin content in loaves of bread and see if the average meets national standards
2. Compare vitamin content of loaves immediately after baking versus content in same loaves 3 days later
3. Compare vitamin content of bread immediately after baking versus other loaves that have been on shelf for 3 days

Answers:
1 = single sample
2 = paired samples
3 = independent samples

1. One-Sample *t* Test

A. H_0: $\mu = \mu_0$ vs. H_a: $\mu \neq \mu_0$ (two-sided), or
H_a: $\mu < \mu_0$ (left-sided), or
H_a: $\mu > \mu_0$ (right-sided)

B. Collect the data and calculate t_{stat}

$$t_{\text{stat}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ with } df = n - 1$$

C. One-sided P-value = AUC in tail beyond t_{stat}
Two-sided P-value = twice that

SIDS Example

- Research question: Do Sudden Infant Death Syndrome (SIDS) babies have lower than average birth weights?
- Prior research in non-SIDS population suggests normal mean birth weight around 3300 grams
- Sample SIDS population $\Rightarrow$ n = 10 $\Rightarrow$ determine birth weights $\Rightarrow$ mean = 2890.5, s = 720.0
- Ask: Do data show that SIDs *population* has mean birth weight less than 3300 gms?

SIDS Example

A. H_0: μ = 3300 vs. H_a: μ ≠ 3300 (two-sided), or
H_a: μ < 3300 (one-sided)

B.

$$t_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{2890.5 - 3300}{720/\sqrt{10}} = -1.80$$

$$df = n - 1 = 10 - 1 = 9$$

C. Convert t_{stat} to *P*-value (next series of slides)

P-value from *t* Table

- Bracket $|t_{stat}|$ between "critical" value landmarks in table C
- One-sided $P \approx$ upper tail (column heading)
- For example, $t_{stat} = -1.80$ with 9 *df*

Table C. Traditional *t* table

$|t_{stat}| = 1.80$

Upper-tail P	0.25	0.20	0.15	0.10	0.05	0.025
df = 9	0.703	0.883	1.100	1.383	1.833	2.262

Thus ⇒ One-tailed: $0.05 < P < 0.10$

Two-tailed: $0.10 < P < 0.20$

Converting t_{stat} to *P*-value via STATA

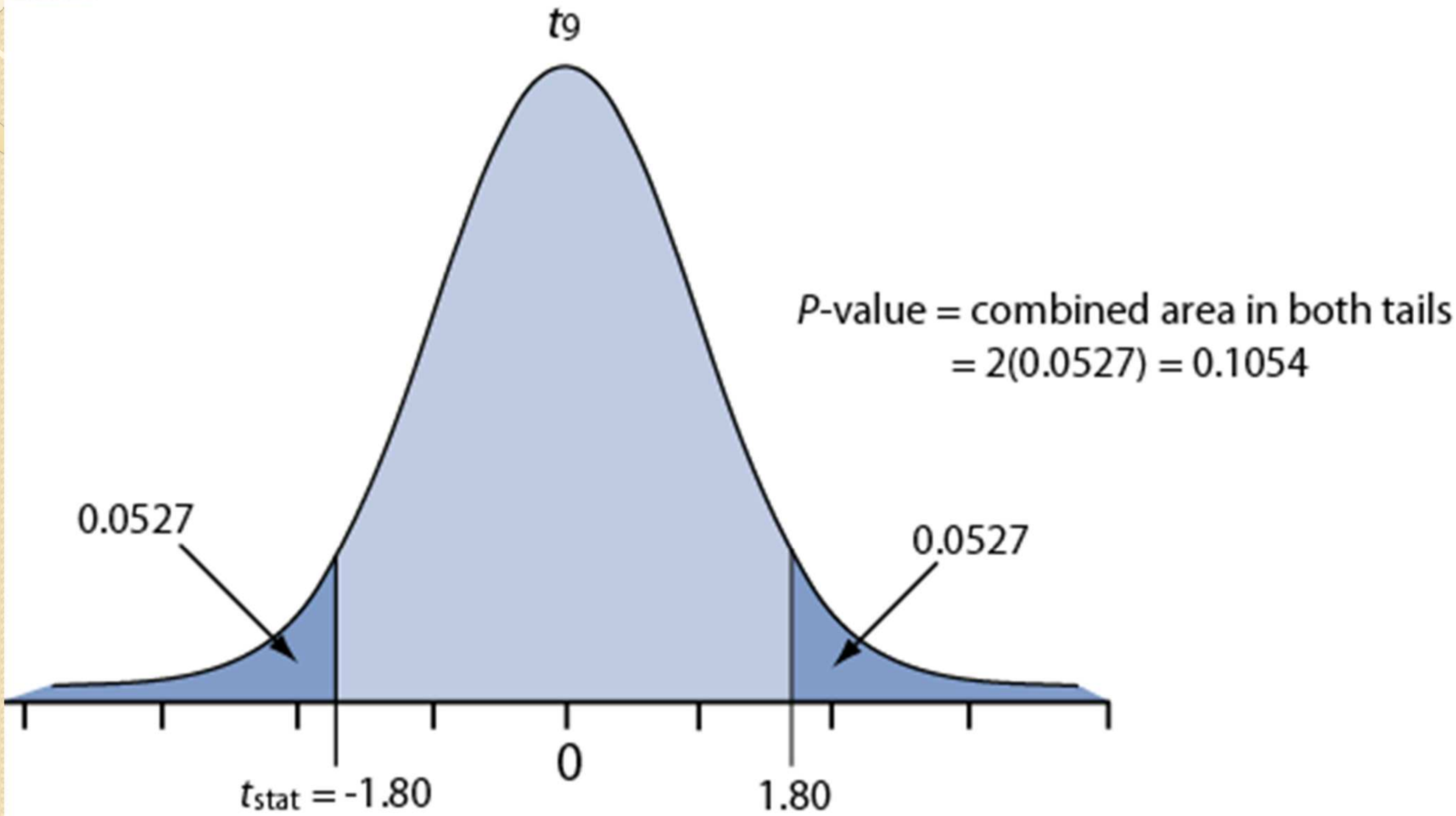
```
. ttesti 10 2890.5 720 3300

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10      2890.5     227.684         720    2375.443    3405.557
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  -1.7985
Ho: mean = 3300                                  degrees of freedom =        9

    Ha: mean < 3300              Ha: mean != 3300              Ha: mean > 3300
 Pr(T < t) = 0.0528         Pr(|T| > |t|) = 0.1056          Pr(T > t) = 0.9472
```

One-sided P-value   Two-sided P-value   One-sided P-value

Visualization of AUCs



Exact tail areas determined by computer

2. Paired *t* Test

- Two samples in which each data point in one sample uniquely paired to a data point in the other sample
- Examples
  - Sequential samples within individuals
  - “Pre-test/post-test”
  - Cross-over trials
  - Pair-matching

"Magic Egg" Example

- Does magic egg reduce LDL cholesterol (mmol/L)?
- Cross-over design: Half subjects start on regular egg, half on magic egg
- Two weeks on diet 1 $\Rightarrow$ LDL cholesterol
- Washout period
- Switch to other diet
- Two weeks on diet 2 $\Rightarrow$ LDL cholesterol

Egg dataset

Subject	RE	ME
----	--------	--------
1	4.61	3.84
2	6.42	5.57
3	5.40	5.85
4	4.54	4.80
5	3.98	3.68
6	3.82	2.96
7	5.01	4.41
8	4.34	3.72
9	3.80	3.49
10	4.56	3.84
11	5.35	5.26
12	3.89	3.73

Within Pair Difference "DELTA"

- Let **`DELTA = RE - ME`**
- First three observations in ME data:

```
ID       RE       ME     DELTA
---- -------- --------  ------
 1     4.61     3.84     0.77
 2     6.42     5.57     0.85
 3     5.40     5.85    -0.45
 etc.
```

Summary Stats for DELTA ("ME")

$$n = 12$$

$$\bar{x}_d = 0.3808$$

$$s_d = 0.4335$$

subscript $_d$ denote statistics for DELTA

is optional

Paired *t* Test

- Similar to one-sample *t* test with μ_0 set to 0, representing "no mean difference"
- H_0: $\mu = 0$

$$t_{\text{stat}} = \frac{\bar{x}_d - \mu_0}{s_d / \sqrt{n}}$$

$$df = n - 1$$

Paired *t* test: "ME"

A. H_0: $\mu_d = 0$ vs. H_a: $\mu_d \neq 0$

B. Test statistic

$$t_{\text{stat}} = \frac{\bar{x}_d - \mu_0}{s/\sqrt{n}} = \frac{0.38083 - 0}{.4335/\sqrt{12}} = 3.043$$

$$df = n - 1 = 12 - 1 = 11$$

C. $P = 0.011$

$\Rightarrow$ significant evidence against H_0

STATA Output: "ME"

```
. ttest re == me

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      re |      12    4.605833    .2316916    .8026033    4.095884    5.115783
      me |      12      4.2625    .2626039    .9096865    3.684513    4.840487
---------+--------------------------------------------------------------------
    diff |      12    .3433333    .1200589    .4158963    .0790854    .6075812
------------------------------------------------------------------------------
     mean(diff) = mean(re - me)                                   t =   2.8597
 Ho: mean(diff) = 0                              degrees of freedom =       11

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9922         Pr(|T| > |t|) = 0.0155          Pr(T > t) = 0.0078
```

3. Independent Samples

EX. Do fasting cholesterol levels (mg/dl) differ in Type A and Type B personality men?

Group 1 (Type A personality):

233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218, 212, 325

Group 2 (Type B personality):

344, 185, 263, 246, 224, 212, 188, 250, 148, 169, 226, 175, 242, 252, 153, 183, 137, 202, 194, 213

Summary Statistics

Group	*n*	mean	std dev
1	20	245.05	36.64
2	20	210.30	48.34

Notation for Inference about Mean Difference

Parameters (population)

Group 1	N_1	μ_1	σ_1
Group 2	N_2	μ_2	σ_2

Statistics (sample)

Group 1	n_1	$\bar{x}_1$	s_1
Group 2	n_2	$\bar{x}_2$	s_2

$\bar{x}_1 - \bar{x}_2$ is the point estimator of $\mu_1 - \mu_2$

Standard Error of Mean Difference

How precise does $\bar{x}_1 - \bar{x}_2$ estimate $\mu_1 - \mu_2$?

Standard error of the mean difference

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This estimate of the standard error does *not* assume groups have equal variance

Two ways to estimate **degrees of freedom**:

- $df_{\text{Welch-Satterthwaite}}$ = formula [use computer]
- $df_{\text{conservative}}$ = the smaller of $(n_1 - 1)$ or $(n_2 - 1)$

For the illustrative data:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{36.638^2}{20} + \frac{48.340^2}{20}} = 13.563$$

$$df_{Welsch-\text{Satterthwaite}} = 35.4 \text{ (via STATA)}$$

$df_{\text{conservative}}$ = smaller of $(n_1 - 1)$ or $(n_2 - 1)$

= 20 – 1 = 19

Hypothesis Test

A. H_0: $\mu_1 = \mu_2$ VS.
H_a: $\mu_1 \neq \mu_2$ (two-sided) or
H_a: $\mu_1 > \mu_2$ (right-sided) or
H_a: $\mu_1 < \mu_2$ (left-sided)

A. Test statistic

$$t_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{\bar{x}_1 - \bar{x}_2}} \text{ where } SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df_{\text{Welch}} \text{ or } df_{\text{conserv}} \text{ (see previous slide)}$$

C. One-sided $P = \Pr(T \geq |t_{\text{stat}}|)$
Double one-sided P for two-sided test

Hypothesis Test – Example

A. H_0: $\mu_1 = \mu_2$ vs. H_a: $\mu_1 \neq \mu_2$

B. Prior calculations:

$xbar_1 - xbar_2 = 245.05 - 210.30 = 34.75$
$SE = 13.563$ with $df_{conserv} = 19$

$$t_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{34.75}{13.563} = 2.56 \text{ with } 19\, df$$

C. Using table C, the one-sided *P* is between .01 and .005 and the two-sided *P* between .02 and .01

⇒ "Significant" evidence against H_0

STATA Output

```
. ttesti 20 245.05 36.64 20 210.30 48.34, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      20      245.05    8.192953       36.64     227.902     262.198
       y |      20       210.3    10.80915       48.34    187.6762    232.9238
---------+--------------------------------------------------------------------
combined |      40     227.675    7.249283    45.84849    213.0119    242.3381
---------+--------------------------------------------------------------------
    diff |               34.75    13.56327                7.226597     62.2734
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   2.5621
Ho: diff = 0                     Satterthwaite's degrees of freedom =  35.4138

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9926         Pr(|T| > |t|) = 0.0148          Pr(T > t) = 0.0074
```

Comparison of CI Formulas

$$\text{(point estimate)} \pm (t^*)(SE)$$

Type of sample	point estimate	*df* for *t**	*SE*
single	$\bar{x}$	$n - 1$	$\frac{s}{\sqrt{n}}$
paired	$\bar{x}_d$	$n - 1$	$\frac{s_d}{\sqrt{n}}$
independent	$\bar{x}_1 - \bar{x}_2$	smaller of $n_1 - 1$ or $n_2 - 1$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Conditions for Inference

t procedures require these conditions:

- SRS
- Valid information (no bias)
- Normal population or large sample (central limit theorem)

Quiz

- Open Eclair9 data in STATA,
- In 11-to-19-year study population?
  - Is the average of them = 15 years?
  - Does the average age of the cases differ from that of the controls?

Equal Variance *t* Procedure (*Optional*)

- Also called pooled variance *t* procedure
- Not as robust as prior method, but…
- Historically important
- Calculated by software programs
- Leads to advanced ANOVA techniques

Pooled variance procedure

We start by calculating this pooled estimate of variance

$$s^2_{pooled} = \frac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df_1 + df_2}$$

where

s_i^2 is the variance in group i and

$$df_i = n_i - 1$$

- The pooled variance is used to calculate this standard error estimate:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{s^2_{pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

- Confidence Interval

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{df,1-\frac{\alpha}{2}})(SE_{\bar{x}_1-\bar{x}_2})$$

- Test statistic

$$t_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{\bar{x}_1-\bar{x}_2}}$$

- All with $df = df_1 + df_2 = (n_1-1) + (n_2-1)$

Pooled Variance t Confidence Interval

Data

Group	n_i	s_i	xbar_i
1	20	36.64	245.05
2	20	48.34	210.30

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{1839.623\left(\frac{1}{20} + \frac{1}{20}\right)} = 13.56$$

$$df = (20-1) + (20-1) = 38$$

$$\begin{aligned} 95\%\ \text{CI for } \mu_1 - \mu_2 &= (\bar{x}_1 - \bar{x}_2) \pm (t_{38,.975})(SE_{\bar{x}_1 - \bar{x}_2}) \\ &= (245.05 - 210.30) \pm (2.02)(13.56) \\ &= 34.75 \pm 27.39 = (7.36, 62.14) \end{aligned}$$

Pooled Variance t Test

Data:

Group	n_i	s_i	xbar_i
1	20	36.64	245.05
2	20	48.34	210.30

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{1839.623\left(\frac{1}{20}+\frac{1}{20}\right)} = 13.56$$

$$df = (20-1)+(20-1) = 38$$

$$H_0 : \mu_1 = \mu_2$$

$$t_{\text{stat}} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1-\bar{x}_2}} = \frac{34.75}{13.56} = 2.56;\ df = 38$$

$$P = 0.015$$

Nonparametric Test Procedures

1. Do Not Involve Population Parameters

   Example: Probability Distributions, Independence

2. Data Measured on Any Scale (Ratio or Interval, Ordinal or Nominal)

3. Example: Wilcoxon Rank Sum Test

Advantages of Nonparametric Tests

1. Used With All Scales
2. Easier to Compute
3. Make Fewer Assumptions
4. Need Not Involve Population Parameters

Disadvantages of Nonparametric Tests

1. May Waste Information

   Parametric model more efficient
   if data Permit

2. Difficult to Compute by

   hand for Large Samples

3. Tables Not Widely Available