Numerical variables inference

- Are books cheaper online?
- Do men run, on average, faster than women?
- Is the average weights of chickens that were fed linseed, sunflower and soybean different?

Numerical inference

Paired data

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Paired data

Prices of textbooks at UCLA's bookstore and Amazon.com

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

where $diff = UCLA - Amazon$ is the price difference.

Paired data

Hypothesis testing

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Test statistic:

$$Z = \frac{\mu_{diff} - 0}{SE_{diff}},$$

where $SE_{diff} = \frac{se_{diff}}{\sqrt{n_{diff}}}$.

Paired data

Summary statistics for the price differences.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
73	12.76	14.26

Test statistic:

$$Z = \frac{12.76 - 0}{1.67} = 7.59$$

p-value

$$p - value = P(|Z| > 7.59) = 2P(Z > 7.59) = 0.0004$$

Since $p - value$ is smaller than $\alpha = 0.05$ we reject the null hypothesis. We have found convincing evidence that Amazon prices are different from the UCLA prices for textbooks.

Difference in mean

Compering group averages

The Cherry Blossom run... yet again...

	men	women
$\bar{x}$	87.65	102.13
s	12.5	15.2
n	45	55

Difference in mean

Hypothesis testing

H_0: $\mu_m - \mu_w = 0$. There is no difference between the average running time of men and women

H_A: $\mu_m - \mu_w \neq 0$. There is a difference in average running time.

Test statistic:

$$Z = \frac{\mu_m - \mu_w - 0}{SE_{\bar{x}_m - \bar{x}_w}},$$

where

$$SE_{\bar{x}_m - \bar{x}_w} = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \tag{1}$$

Confidence interval

$$\mu_m - \mu_w \pm z^* SE_{\bar{x}_m - \bar{x}_w}$$

Difference in mean

Conditions for normality

- The sample means, $\bar{x}_m$ and $\bar{x}_w$, each meet the criteria for having nearly normal sampling distributions.
- The observations in the two samples are independent.

Difference in mean

Test statistic:

$$Z = \frac{14.48}{2.77} = 75.69$$

p-value

$$p - value = P(|Z| > 75.69) = 2P(Z > 75.59) \simeq 0$$

Since p-value is almost zero we reject the null hypothesis.
95% Confidence interval

$$14.48 \pm 1.96 \cdot 2.77 = (9.05, 19.9)$$

The normality condition

Reminder:
Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The population distribution is not strongly skewed.

Q What if we have a small sample? $n < 30$.

T-distirbution

T-distribution vs. Normal distibution

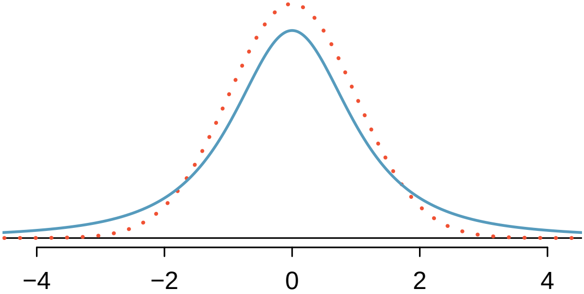
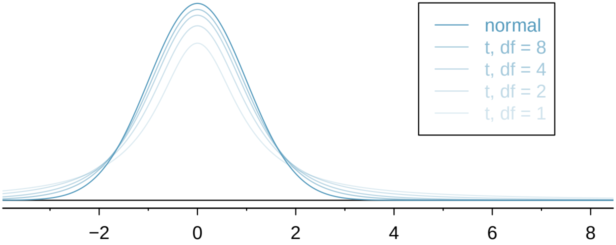


Figure: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

T-distribution

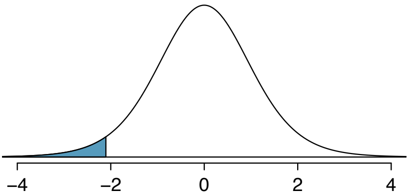
Degrees of freedom (df)

The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.



T-distribution

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	
17	1.33	1.74	2.11	2.57	2.90
18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58



T-distribution

- **Independence of observations.**
  We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we carefully check to the best of our abilities that the observations were independent.
- **Observations come from a nearly normal distribution.**
  This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

Inference for $\bar{x}$

- **Degrees of freedom:** $df = n - 1$, where n is the number of observations
- **Confidence interval:**

$$\bar{x} \ \pm \ t^{\star}_{df} SE$$

.

Inference for $\bar{x}$

Hypothesis testing

H_0: $\mu = \mu_0$.

$$H_A: \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative)} \\ \mu \neq \mu_0 & \text{(two-tailed alternative)} \\ \mu < \mu_0 & \text{(lower-tail alternative)} \end{cases}$$

Test statistic: $t = \frac{\bar{x} - \mu_0}{SE}$
We reject H_0 when:

- $P(T_{df} > t) < \alpha$ (upper-tail alternative)
- $P(|T_{df}| > t) < \alpha$ (two-tailed alternative)
- $P(T_{df} < -t) < \alpha$ (lower-tail alternative)

Paired data for small sample

- **Degrees of freedom:** $df = n - 1$, where n is the number of observations
- **Confidence interval:**

$$\bar{x}_{diff} \ \pm \ t^{\star}_{df} SE_{diff}$$

.

Paired data for small sample

Hypothesis testing

H_0: $\mu_{diff} = 0$.
H_A: $\mu_{diff} \neq 0$.

Test statistic:

$$T = \frac{\mu_{diff} - 0}{SE_{diff}},$$

where $SE_{diff} = \frac{se_{diff}}{\sqrt{n_{diff}}}$.

Difference in mean for small samples

- **Degrees of freedom:** $df = \min\{n_1 - 1, n_2 - 1\}$, where n_1, n_2 is the number of observations in the respective groups.

- **Confidence interval:**

$$\bar{x}_1 - \bar{x}_2 \;\pm\; t^{\star}_{df} SE_{diff}\,,$$

where

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

Difference in mean for small samples

- **Hypothesis testing**

  H_0: $\mu_1 - \mu_2 = 0$.
  H_A: $\mu_1 - \mu_2 \neq 0$.

- **Test statistic:**

$$T = \frac{\mu_1 - \mu_2 - 0}{SE_{\bar{x}_m - \bar{x}_w}},$$

Difference in mean for small samples

Conditions

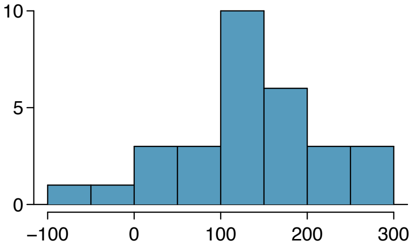
- Each sample meets the criteria for using the t distribution.
- The samples are independent.

Example

- An SAT preparation company claims that its students' scores improve by over 100 points on average after their course.
- We have a random sample of the scores of 30 students, before and after.
- Which test should we apply to check is the companies claim is true?

T-student difference test

- Calculate the difference in scores for each student; x_i denotes the difference for the ith student.
- Check the conditions:
  - Independence holds.
  - Approximately normal



T-student difference test

- H_0: student scores do not improve by more than 100 after taking the company's course.
- H_A: students scores improve by more than 100 points on average after taking the company's course.

Or

- $H_0 : \mu_{diff} = 100$
- $H_A : \mu_{diff} > 100$.

T-student difference test

n	$\bar{x}$	s
30	135.9	82.2

- $df = n - 1 = 29$
- $t = \frac{\bar{x}-100}{SE_{diff}} = \frac{135.9-133}{82/\sqrt{30}} = 2.39$
- $p - value = P(T > t) = 0.0118 < 0.05 = \alpha$
- We reject the null hypothesis. The data provide convincing evidence to support the company's claim that student scores improve by more than 100 points following the class.

Comparing many means

- What is we want to compare means from k groups, with $k > 2$?
- We have the following hypothesis test:

  H_0: The mean outcome is the same across all groups.
  H_A: At least one mean is different.

  Or

  H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
  H_A: At least one mean is different.

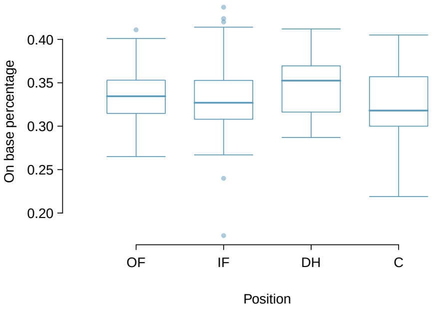
ANOVA

Batting average

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NYM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

variable	**description**
name	Player name
team	The abbreviated name of the player's team
position	The player's primary field position (OF, IF, DH, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
AVG	Batting average, which is equal to H/AB
OBP	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

ANOVA



Summary statistics

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

ANOVA

Hypothesis

- $H_0 : \mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$
- H_A : The average on-base percentage (μ_i) varies across some (or all) groups.

ANOVA

Preliminary

Mean square between groups (MSG**)** for k groups.

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x} \right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i. MSG has $df_G = k - 1$.

ANOVA

- **Sum of squared errors (SSE)** in one of two equivalent ways:

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

where s_i^2 is the sample variance in group i.

- **Mean Square Error (MSE)**

$$MSE = \frac{1}{df_E} SSE,$$

where $df_E = n - k$ and n is the total sample size.

ANOVA

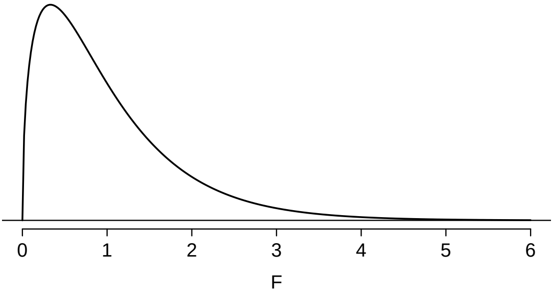
Test statistic

$$F = \frac{MSG}{MSE},$$

F-distribution with degrees of freedom $(df_1, df_2) = (k-1, n-k)$.

Bating average Hypothesis

$$F = 1.994, \text{ with } (df_G, df_E) = (3, 323)$$



$p - value = 0.115$

ANOVA

ANOVA Summary Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	3	0.0076	0.0025	1.9943	0.1147
Residuals	323	0.4080	0.0013		

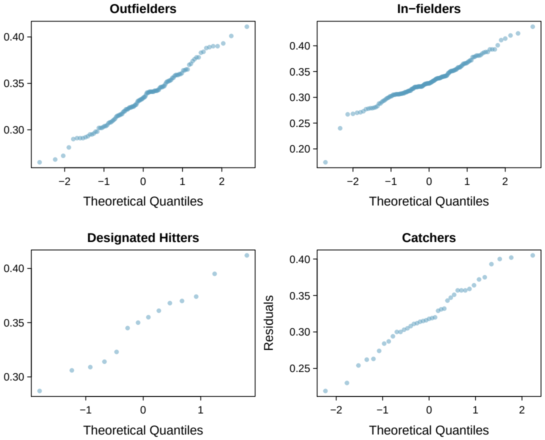
$s_{pooled} = 0.036$ on $df = 323$

ANOVA

Conditions

- Independence
- Approximately normal
- Constant variance: the variance in the groups is about equal from one group to the next.

ANOVA


Outfielders
0.40
0.35
0.30
−2
−1
0
1
2
Theoretical Quantiles
In−fielders
0.40
0.35
0.30
0.25
0.20
−2
−1
0
1
2
Theoretical Quantiles
Designated Hitters
0.40
0.35
0.30
−1
0
1
Theoretical Quantiles
Residuals
Catchers
0.40
0.35
0.30
0.25
−2
−1
0
1
2
Theoretical Quantiles

ANOVA

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean ($\bar{x}_i$)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045