Statistical inference for Two Samples

Lecture #5

Adi Sarid
Tel-Aviv University
updated: 2020-04-21

Reminder from previous lecture

- Hypothesis tests (e.g., of proportion in the attractive flu shot experiment)
- Type-I and Type-II errors $\alpha = P(H_1 | H_0), \beta = P(H_0 | H_1)$
- General framework for hypothesis testing (the 8 steps)
  - Parameter -> Null hypothesis -> Altermative -> Significance -> Statistic -> Rejection criteria
  - Sample, computation -> Decision
- P-value and the relationship to confidence intervals
- Designing the sample size for a desired power $1 - \beta$
- Goodness of fit hypothesis test (comparing to distribution)
- Words of caution about HARKing and multiple comparisons

Are men and women different? ♂ ♀

Yes. Of course there are gender differences, but this is a great example to start today's lecture with!

Here is a research published in PLOS One, about gender differences relating to **empathy and moral cognition**:

- Baez, Sandra, et al. "Men, women… who cares? A population-based study on sex differences and gender roles in empathy and moral cognition." *PloS one* 12.6 (2017): e0179336.
  - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179336
- The research uses statistical means to compare measures such as the **Interpersonal Reactivity Index** (IRI) to see if women score higher than men, in a **statistically significant manner**.
  - For example, figure 3. (significant sex differences in self-reported empathy) shows that women **percive** themselves as being more interpersonal than what men perceive. However, these might be driven by stereotypes, and might be apparant only in self-assesment instruments.
- We won't go over the entire study, but I highly recommend reading it: based on a very large sample $n = 10802$ and refutes some common stereotypes. (Also a misleading presentation e.g. y-axis inconsistencies, which we can learn from).
- One very cool feature of this study is that the authors made the data available for the public here.

What is it to us?

In previous lectures we saw confidence intervals and hypothesis tests relating to a **single parameter** (single sample)

Today we will explore the comparison between **two parameters** (two samples, e.g., male-female, tall-short, before-after).

In the gender empathy example, let $\mu_{\mathrm{f}}, \mu_{\mathrm{m}}$ represent female and male empathy scores. We want to test the following hypothesis:

- $H_0 : \mu_{\mathrm{f}} = \mu_{\mathrm{m}}$
- $H_1 : \mu_{\mathrm{f}} > \mu_{\mathrm{m}}$

Then, set $\mu_{\mathrm{diff}} = \mu_{\mathrm{f}} - \mu_{\mathrm{m}}$ and

- $H_0 : \mu_{\mathrm{diff}} = 0$
- $H_1 : \mu_{\mathrm{diff}} > 0$

This, more-or-less, brings us back to what we already learned last week.

Most of today's material is covered in Montgomery, chapter 10.

Difference in means - variance known

Assume

- $X_{11}, \ldots, X_{1n_1}$ is a random sample from population 1.
- $X_{21}, \ldots, X_{2n_n}$ is a random sample from population 2.
- The two populations X_1 and X_2 are independent.
- Both populations are normally distributed.

Then, what is $E[\bar{X}_1 - \bar{X}_2]$, and $\text{Var}[\bar{X}_1 - \bar{X}_2]$?

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$$

$$\text{Var}[\bar{X}_1 - \bar{X}_2] = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Using these, and the fact that the sum of two normally distributed variables is also normal we obtain...

Difference in means - variance known (2)

The following quantity is normally distributed $\mathrm{N}(0, 1)$:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Alternative Hypotheses (rejection criterion):

- $H_1 : \mu_1 - \mu_2 \neq \Delta_0 \quad (z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2})$
- $H_1 : \mu_1 - \mu_2 > \Delta_0 \quad (z_0 > z_\alpha)$
- $H_1 : \mu_1 - \mu_2 > \Delta_0 \quad (z_0 < -z_\alpha)$

Choice of sample size

Similarly to what we've seen in the last lecture, given α, β, Δ, we can determin the desired sample sizes n_1, n_2, by writing (example for a two sided hypothesis):

$$\beta = P(H_0 \,|\, H_1) = P\left(z_{\alpha/2} < \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha/2} \middle| \mu_1 - \mu_2 = \Delta_0 + \delta \right)$$

Define $\Delta = \Delta_0 + \delta = \mu_1 - \mu_2$, if we play a bit with the expression in the parentheses, and using $\Phi^{-1}(\,\cdot\,)$, we can reach the following expression:

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2}$$

where $n = n_1 = n_2$ and the total sample size is $2n$.

For a one sided alternative hypothesis, the equivalent formulat would be

Confidence intervals on a difference in means

When we discussed confidence intervals two weeks ago, we left out confidence intervals for the difference between two means, but this is actually almost the same (now that you've seen Z for the difference). I.e.:

$$P\left[z_{\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \leq z_{1-\alpha/2}\right] = 1 - \alpha$$

Hence, a confidence interval for $\mu_1 - \mu_2$ can be obtained by using:

$$\bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{1-\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Or in the one-sided case:

$$\mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{1-\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

(Or the other side)

$$\mu_1 - \mu_2 \geq \bar{x}_1 - \bar{x}_2 + z_{\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Difference in means - variance unknown, equal variances

Let's assume that σ is unknown, but that $\sigma_1 = \sigma_2 = \sigma$.

We've seen that $S_i^2 = \frac{\sum_j (X_{ij} - \bar{X}_i)^2}{n_i - 1}$ is an unbiased estimator for σ_i^2.

Introducing the pooled estimator of σ^2, denoted by S_p^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

This is a weighted average of S_1^2 and S_2^2.

Each S_i contributes $n_1 - 1$ degrees of freedom so overall S_p^2 has $n_1 + n_2 - 2$ degrees of freedom.

Given normality assumptions (or n_1 and n_2 large enough), we can use the following statistic as a student's t distribution with $n_1 + n_2 - 2$ degrees of freedom:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

Difference in means - variance unknown, equal variances (2)

Summarizing the test, we can write:

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic: $T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{1/n_1 + 1/n_2}}$

Alternative hypothesis (rejection criteria):

- $H_1 : \mu_1 - \mu_2 \neq \Delta_0 \qquad (t_0 > t_{1-\alpha/2, n_1+n_2-2}$ or $t_0 < t_{\alpha/2, n-1+n_2-2})$
- $H_1 : \mu_1 - \mu_2 > \Delta_0 \qquad (t_0 > t_{1-\alpha, n_1+n_2-2})$
- $H_1 : \mu_1 - \mu_2 < \Delta_0 \qquad (t_0 < t_{\alpha, n_1+n_2-2})$

Difference in means - variance unknown, unequal variances ($\sigma_1 \neq \sigma_2$)

In the case that the variances are unequal and unknown, we can use the following statistic which is approximately distributed t with v degrees of freedom:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Where the degrees of freedom v are equal:

$$v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Paired or unpaired t-test?

Sometimes, data is collected in pairs, for example when an intervention plan is conducted on a group of individuals

- We want to compare the effect of the intervention before and after.
- This *paired* situation may ocurr when we have multiple observations of the same group.
- We would like to match the observations to avoid differences which may ocurr due to variation between subjects.
- We pair the observations and conduct the statistical tests on the difference.

Let $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ a set of n paired observations. Define $D_j = X_{ij} - X_{2j}$

$$\mu_D = E(X_1 - X_2) = \mu_1 - \mu_2$$

Null hypothesis: $H_0 : \mu_D = \Delta_0$

Test statistic: $T_0 = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$

Where S_D is the sample standard deviation of the differences.

Example of two sample hypothesis testing - the ipf data set

We go back to the ipf data set (the power lifting competition dataset).

What test would you use? (means paired / means unpaird / something else) for each of the following:

Test 1:

- H_0: men lift higher weights than women
- H_1: men and women lift the same weight

Test 2:

- H_0: Deadlift and squat weights are the same
- H_1: Deadlift weights are higher than squat weights

Test 3:

- H_0: The age of male athletes is normally distributed
- H_1: The age of male athletes is not normally distributed

Test 4:

- H_0: The athletes age and gender are statistically independent
- H_1: The athletes age and gender are not statistically independet

05:00

Demonstrating tests in the ipf data set (test 1)

- H_0: men lift higher weights than women
- H_1: men and women lift the same weight

Unpaired t-test

```
#https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-08/ipf_lifts.csv
set.seed(0)
ipf <- read_csv("data/ipf_lifts.csv", col_types = cols()) %>%
   filter(best3squat_kg > 0) %>%
   sample_n(1000)

t.test(formula = best3squat_kg ~ sex, data = ipf, alternative = "less")
```

```
##
##     Welch Two Sample t-test
##
## data:  best3squat_kg by sex
## t = -32.558, df = 997.23, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -99.81944
## sample estimates:
## mean in group F mean in group M
##        146.4611        251.5971
```

```
# alternative method to call the function
#t.test(x = ipf$best3squat_kg[ipf$sex == "F"], y = ipf$best3squat_kg[ipf$sex == "M"], alternative = "less")
```

Demonstrating tests in the ipf data set (test 2)

Test 2:

- H_0: Deadlift and squat weights are the same
- H_1: Deadlift weights are higher than squat weights

Paired t-test, because each athelete is compared to himself

```r
t.test(x = ipf$best3deadlift_kg, y = ipf$best3squat_kg, paired = TRUE, alternative = "greater")
```

```
##
##     Paired t-test
##
## data:  ipf$best3deadlift_kg and ipf$best3squat_kg
## t = 6.1662, df = 976, p-value = 5.113e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.141787      Inf
## sample estimates:
## mean of the differences
##                5.650502
```

Demonstrating tests in the ipf data set (test 3 and 4)

Test 3 was:

- H_0: The age of male athletes is normally distributed
- H_1: The age of male athletes is not normally distributed

This is the goodness of fit test that we were discussing last week.

Test 4 was:

- H_0: The athletes age and gender are statistically independent
- H_1: The athletes age and gender are not statistically independet

The fourth test is actually a goodness of fit test for uniform distribution.

Goodness of fit test for independence (contingency table)

When we have two variables and we want to examine independence between them, it is like comparing the contingencies (combinations) of the two variables, against the uniform distribution.

Age	genderM	genderF	total
group1	g1m	g1f	n_{g1}
group2	g2m	g2f	n_{g2}
group3	g3m	g3f	n_{g3}
total	n_M	n_F	n

- The observed is the count within the cell, and the expected is the product of the marginal probabilities, i.e.:
  - The expected of Males in group2 under the null hypothesis is: $E_{g2m} = (n_{g2}/n) \times (n_M/n) \times n$
- This is the case when the variables are independent, i.e., $P(X = x, Y_y) = P(X = x)P(Y = y)$

This results in a χ^2 test with $(r-1)(c-1)$ degrees of freedom.

Independence between variable via contingency table (example)

```r
ipf_new <- ipf %>%
   filter(!is.na(age)) %>%
   mutate(age_group = cut(age, breaks = c(0, 25, 35, 45, 55, 100))) %>%
   count(age_group, sex) %>%
   pivot_wider(id_cols = age_group, names_from = sex, values_from = n) %>%
   select(2:3) %>%
   as.matrix()

ipf_compare_res <- chisq.test(ipf_new)
ipf_compare_res
```

```
##
##     Pearson's Chi-squared test
##
## data:  ipf_new
## X-squared = 21.751, df = 4, p-value = 0.0002246
```

Your turn (class exercise) - do mobile phones impact our health?

In **pairs**, try to devise an experiment plan, that would test whether **mobile phones impact our health**.

In your answer relate to the following points:

- What do you consider as an "effect"? (i.e., what kind of health measure?)
- How do you select and/or separate the groups participating in your experiment?
- How do you neutralize other factors which might intervene with the experiment? (like selection of participants or other factors)
- What would you use as a statistcal measure?
- Paired or unpaired?
- What sample size?
- What would be the hypotheses 8-step procedure of the experiment? reminder:
  - Parameter -> Null hypothesis -> Altermative -> Significance -> Statistic -> Rejection criteria
  - Sample, computation -> Decision

10:00

Variance of two samples (F-test)

We talked about t-test when the variance is unknown and presented two cases (with equal variance and uneqal variance). How would we decide which of the two to use?

We would like to use **a statistical test** to compare the variance. One method to compare two numbers is to divide them, i.e., σ_1/σ_2.

The F distribution is the ratio of two independent chi-square random variables, divided by its number of degrees of freedom, i.e.:

$$F = \frac{W/u}{Y/v}$$

The probability density function is given by the expression:

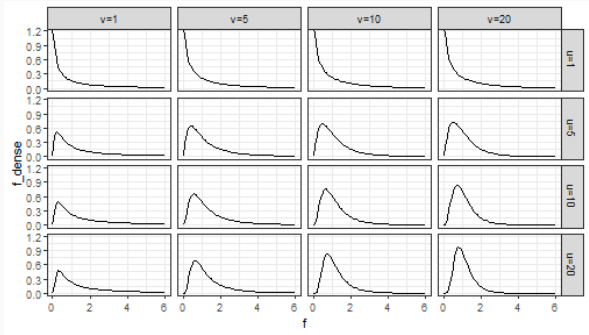
$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}}$$

Illustration of the F distribution

$$F = \frac{W/u}{Y/v}, \quad \mu = v/(v-2), \quad (\text{ for } v > 2), \quad \sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)}$$

```
f_dist <- crossing(f = seq(0, 6, by = 0.1), df1 = c(1, 5, 10, 20), df2 = c(1, 5, 10, 20)) %>%
   mutate(f_dense = pmap_dbl(.l = list(f, df1, df2), .f = df)) %>%
   mutate(v = paste0("v=", df2),
          u = paste0("u=", df1)) %>%
   mutate_at(vars(v,u), fct_inorder)

ggplot(f_dist, aes(x = f, y = f_dense)) +
   geom_line() +
   facet_grid(rows = vars(u), cols = vars(v)) +
   theme_bw() +
   guides(color = guide_legend("df"))
```



Statistical hypothesis test for variance equality

Assuming two independent populations with normal distribution, then $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ has an F distribution with $u = n_1 - 1, v = n_2 - 1$ degrees of freedom.

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic: $F_0 = \frac{S_1^2}{S_2^2}$

Alternative hypothesis (rejection criteria):

- $H_1 : \sigma_1^2 \neq \sigma_2^2 \quad (f_0 > f_{1-\alpha/2, n_1-1, n_2-1} \text{ or } f_0 < f_{\alpha/2, n_1-1, n_2-1})$
- $H_1 : \sigma_1^2 > \sigma_2^2 \quad (f_0 > f_{1-\alpha, n_1-1, n_2-1})$
- $H_1 : \sigma_1^2 < \sigma_2^2 \quad (f_0 < f_{\alpha, n_1-1, n_2-1})$

Example for F statistic variance test

```
ipf %>%
   group_by(sex) %>%
   summarize_at(vars(contains("best3")), ~{(sd(., na.rm = T))^2})
```

```
## # A tibble: 2 x 4
##   sex   best3squat_kg best3bench_kg best3deadlift_kg
## * <chr>         <dbl>         <dbl>            <dbl>
## 1 F             1400.          723.             854.
## 2 M             4184.         2542.            2668.
```

It's pretty clear that males have a much higher variance, lets test this in an F test.

```
var.test(formula = best3squat_kg ~ sex, data = ipf, ratio = 1, alternative = "less")
```

```
##
##     F test to compare two variances
##
## data:  best3squat_kg by sex
## F = 0.3347, num df = 359, denom df = 639, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0000000 0.3910184
## sample estimates:
## ratio of variances
##          0.3347011
```

Inference on two population proportions

We consider the case of two binomial parameters p_1, p_2. Let X_1, X_2 represent the number of successes in each sample. $\hat{P}_i = X_i/n_i$, have approximately normal distributions.

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Is distributed approximately as $Z \sim \mathrm{N}(0, 1)$.

Under the null hypothesis $H_0 : p_1 = p_2 = p$ we have:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

Where an estimator to p is given by:

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

The test procedure for comparing two population proportions

Null hypothesis: $H_0 : p_1 = p_2$

Test statistic: $Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\,(1/n_1 + 1/n_2)}}$

Alternative hypothesis (rejection criteria):

- $H_1 : p_1 \neq p_2 \quad (z_0 > z_{1-\alpha/2}$ or $z_0 < z_{\alpha/2})$
- $H_1 : p_1 > p_2 \quad (z_0 > z_{1-\alpha})$
- $H_1 : p_1 < p_2 \quad (z_0 < z_{\alpha})$

Setting the sample sizes when comparing two population proportions

Very similar to what we've shown in the last lecture for one sample, but with a slightly different computation for the standard deviation under H_1. For example, in the two sided case we have:

$$\beta = \Phi\left[\frac{z_{1-\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] - \Phi\left[\frac{z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right]$$

With $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, $\bar{q} = \frac{n_1(1-p_1) + n_2(1-p_2)}{n_1 + n_2}$ and

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

We can obtain the suggested sample (or power) using `pwr::pwr.2p.test` or `pwr::pwr2p2n.test`.

```
pwr::pwr.2p2n.test(h = pwr::ES.h(p1 = 0.2, p2 = 0.3),
                   n1 = 150, n2 = NULL,
                   sig.level = 0.05,
                   power = 0.8,
                   alternative = "less")
```

Effect Size

We discussed p-value as the extent to which a statistical finding is significant. However, it is not the sole measure for the strength of a statistical finding.

In this context, see the ASA statement on p-Values here

Effect size measures the magnitude of a phenomena. Effect size is a generic name for various measures such as:

- R^2 in linear regression
- ρ Pearson correlation coefficient between two variables
- Cohen's d which relates to the difference between means (which we will now discuss)
- Many more

Effect Size - Cohen's d

The difference between two means divided by standard deviation, i.e.:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_p}$$

Where S_p is the pooled standard deviation:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

```
effsize::cohen.d(formula = best3squat_kg ~ sex, data = ipf)
```

```
##
## Cohen's d
##
## d estimate: -1.863577 (large)
## 95 percent confidence interval:
##     lower     upper
## -2.016548 -1.710606
```