Hypothesis Tests

Lecture #4

Adi Sarid
Tel-Aviv University
updated: 2019-11-17

Reminder from previous lecture

Last lesson we talked about:

- Confidence intervals, i.e.:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- We've seen confidence intervals for
  - Normal distribution (for μ and for σ)
  - Using Student's T (variance unknown)
  - Binomial case (the election survey example)

$$\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{n}}$$

- Setting the sample size according to a confidence interval length, e.g.:

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{r}\right)^2$$

Hypothesis testing (Montgomery chapter 9)

A *statistical hypothesis* is a statement about the parameters of one or more populations.

In empirical research, we first formulate our hypothesis, and then we try to find empirical results to support our hypothesis (never the other way around, that's called HARK-ing).

For example:

- H_0: The average time to reach TLV from Netanya $=$ 40 minutes
- H_1: The average time to reach TLV from Netanya $\neq$ 40 minutes

The H_0 is called the *null hypothesis* and the H_1 is called the *alternative hypothesis*.

The same situation can be descrived with different hypothesis (with a different meaning):

- H_0: The average time to reach TLV from Netanya $=$ 40 minutes
- H_1: The average time to reach TLV from Netanya $>$ 40 minutes

Today we will discuss how to devise hypothesis tests, what are type-I and type-II errors, what is the meaning of rejecting a null hypothesis, what are p-values and what is the connection to the statistical intervals we were discussing.

First - an example

Scan the following QR code (or visit the link) and answer the survey.

http://bit.ly/att_flu_ex

This is a copy of a true survey used in a research I did with a collegue.

- What do you think were our hypothesis in this survey?
- What were we trying to accomplish?

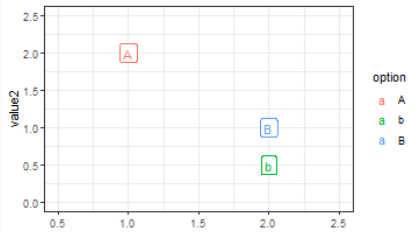
05:00

The Attractive Flu Shot (1/2)

The survey has a number of versions, rendered to respondents randomly. There are five groups:

- Control (a "regular" message from the HMO)
- Recommendation (for effectiveness of the shot, the health ministry recommends to take it early)
- Stock (if you don't take an early shot, the stock may run out)
- Cost (the shot would cost for patients taking it after December)
- Benefit (if you take an early shot, you get some kind of incentive)

This psychological nudge leverages the *attraction effect*, i.e.: options A and B are not comparable, but when decoy b is added, and is comparable to B, we tend to choose B over A.



The Attractive Flu Shot (2/2)

If you read Dan Arieli's books, you probably read about attraction.

We have a lot of hypothesis in this research, but here is an example (the treatment increases vaccination intentions):

- Recommendation treatment:
  - H_0: $p_{\text{control}} = p_{\text{recommendation}}$
  - H_1: $p_{\text{control}} < p_{\text{recommendation}}$
- Stock treatment:
  - H_0: $p_{\text{control}} = p_{\text{stock}}$
  - H_1: $p_{\text{control}} < p_{\text{stock}}$
- You get the hang of it...

Additional hypothesis deal with the interaction of *certainty* and the attraction effect's influence.

Back to theory of hypothesis testing

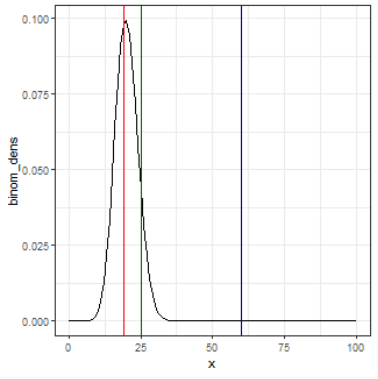
Let's simplify things: say that the percent of patients taking flu vaccinations is about 20% (known based on previous years). We want to see if our experiment led to an increase in that percent, that is *significantly* higher.

- H_0: $p_{\text{treatment}} = 0.2$
- H_1: $p_{\text{treatment}} > 0.2$

What would you say if we measure after the intervention the following rates...?

- $\hat{p} = 0.19$
- $\hat{p} = 0.60$
- $\hat{p} = 0.25$

We need a clear statistical *criteria* for deciding what is significant and what is not.

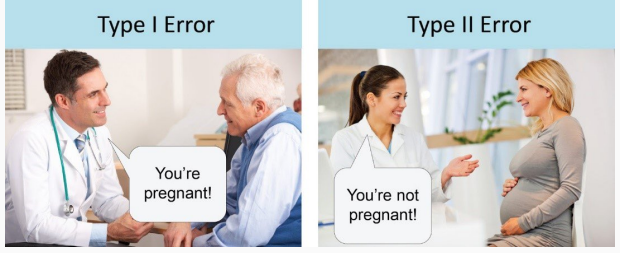


Two types of errors

In order to set a decision rule, we need to consider two types of errors:

- **Type-I** error (aka False-positive): **rejecting** H_0 when it is **true**.
- **Type-II** error (aka False-negative): **failing** to reject H_0 when it is **false**.

In medical decision making, this would look like H_0: not pregnant



(source)

Tradeoff between type-I and type-II errors

01:00

Type-I error

The type-I error is also called the *significance level*, or α-error.

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ True})$$

Questions:

- In the pregnancy classification example, what is an $\alpha = 0$ decision rule?
- In the flu vaccination example, what is an $\alpha = 0$ decision rule?
- Always classify: not pregnant
- Reject H_0 when $\hat{p} = \pm\infty$ (i.e. never reject H_0)

Type-II error

As the type-I error decreases the decision rule tends to prefer not rejecting H_0 which leads to a higher type-II error

$$\beta = P(\text{Fail to reject } H_0 | H_1 \text{ True})$$

General framework for hypothesis testing

This is the procedure for hypothesis testing:

1. Identify the parameter of interest (i.e., proportion, expectancy, std, etc.)
2. State the null hypothesis H_0
3. Specify the alternative hypothesis H_1 (one sided, two sided, etc.)
4. Choose significance level α
5. Determine what test statistic to use (e.g., Z, T, X^2)
6. State the rejection region for the statistic
7. Compute the sample quantities, plug-in into the test statistic and compute it
8. Decide if H_0 should be rejected based on 6-7

Steps 1-4 must be done before starting the research (a-priori and not posteriori, we'll see why later on)

Example for attraction effect hypothesis test decision rule (computation)

Let's create an $\alpha = 0.05$ decision rule for classifying if the attraction effect worked in our experiment.

- The parameter is p (vaccinations after intervention)
- H_0: p = 0.2 (no attraction effect)
- H_1: p > 0.2 (interevntion was successful)
- $\alpha = 0.05$
- The test statistic

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

- Reject H_0 if $z_0 > z_{1-\alpha} = 1.96$
- Compute z_0, given experiment results: $x = 29$ successes, $n = 100$ subjects

$$z_0 = \frac{29 - 100 \times 0.2}{\sqrt{100 \times 0.2 \times 0.8}} = 2.25 > 1.96 = z_{0.95}$$

- Conclusion: we reject H_0, and therefor claim that the experiment nudged patients to vaccinate.

Example for attraction effect hypothesis test decision rule (code)

In R, we have a number of functions for this specific test. The previous computation is an **apporximation** (doesn't work for small n or extreme p). R can compute an approximate or an exact test.

```r
prop.test(x = 29, n = 100, p = 0.2, alternative = "greater")
```

```
##
##     1-sample proportions test with continuity correction
##
## data:  29 out of 100, null probability 0.2
## X-squared = 4.5156, df = 1, p-value = 0.01679
## alternative hypothesis: true p is greater than 0.2
## 95 percent confidence interval:
##  0.2171785 1.0000000
## sample estimates:
##    p
## 0.29
```

```r
binom.test(x = 29, n = 100, p = 0.2, alternative = "greater")
```

```
##
##     Exact binomial test
##
## data:  29 and 100
## number of successes = 29, number of trials = 100, p-value =
## 0.02002
## alternative hypothesis: true probability of success is greater than 0.2
## 95 percent confidence interval:
##  0.2158797 1.0000000
## sample estimates:
## probability of success
##                   0.29
```

P-value versus specified α value

The results in R do not state a specific z_0 but rather a **p-value**. What does p-value mean?

- By stating "we reject H_0 at the α level" we're omitting important information:
  - How "significant" is our rejection?
  - Would this rejection hold when we vary α? i.e.: Would our rejection hold at $\alpha = 0.03$? $\alpha = 0.0001$?
- The p-value is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

$$z_0 = \frac{29 - 100 \times 0.2}{\sqrt{100 \times 0.2 \times 0.8}} = 2.25 > 1.65 = z_{0.95} = z_{1-\alpha}$$

As long as $z_{1-\alpha}$ < 2.25 we would still reject H_0. Hence, the p-value is:

$$\text{p-value} = 1 - \Phi(z_0 = 2.25) = 0.012$$

```
1-pnorm(q=2.25)
```

```
## [1] 0.01222447
```

Type-II error and determining the sample size (1/3)

So far, we've only considered the choice of α, i.e., the type-I error. However, we can also influence β by choice of the sample size n (and sometimes also by the type of test - *out of scope*).

Let's see how change in n can increase the power (decrease the type-II error, β).

In our example, we had $p_0 = 0.2$, let's assume now that in the alternative hypothesis we use $p_1 = p_0 + \delta$ (one sided "greater than" test).

$$\beta = P(\text{Not rejecting } H_0 | H_1 \text{ true}) = P\left(\frac{x - np_0}{\sqrt{np_0(1-p_0)}} \leq z_{1-\alpha} | p_1 = p_0 + \delta\right)$$

Note that

$$Z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{x - (np_0 + n\delta)}{\sqrt{np_0(1-p_0)}} + \frac{\delta\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

Given H_0, the distribution of $Z_0 \sim \mathcal{N}\left(\delta\sqrt{n/p_0(1-p_0)}, 1\right)$, hence

Type-II error and determining the sample size (2/3)

$$\beta = P\left(\frac{x-(np_0+n\delta)}{\sqrt{np_0(1-p_0)}} < z_{1-\alpha} - \frac{\delta\sqrt{n}}{\sqrt{p_0(1-p_0)}} | p_1 = p_0+\delta\right) =$$

$$= P\left(\frac{x-(np_0+n\delta)}{\sqrt{n(p_0+\delta)(1-p_0-\delta)}} < \frac{z_{1-\alpha}\sqrt{np_0(1-p_0)}-\delta n}{\sqrt{n(p_1+\delta)(1-p_1-\delta)}} \quad | \quad p_0+\delta\right)$$

$$\beta = \Phi\left(\frac{z_{1-\alpha}\sqrt{np_0(1-p_0)}-\delta n}{\sqrt{n(p_0+\delta)(1-p_0-\delta)}}\right)$$

Now we can use `qnorm(beta)`, i.e.: $\Phi^{-1}(\beta)$, and given β, p_0, δ compute n.

Luckily, we have a function for that. From package `pwr`, function `pwr.p.test`.

Type-II error and determining the sample size (3/3)

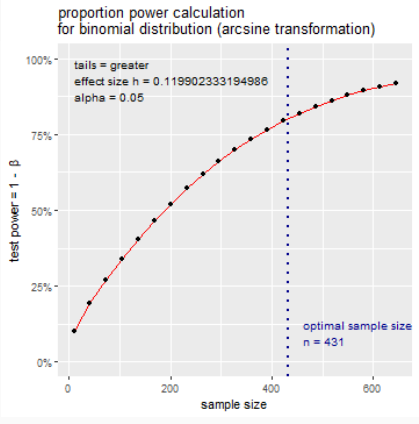
```r
beta <- 0.2
delta <- 0.05
alpha <- 0.05
p_0 <- 0.2

library(pwr)
```

```
## Warning: package 'pwr' was built under R version 3.6.1
```

```r
p_out <- pwr::pwr.p.test(h = ES.h(p1 = p_0 + delta, p2 = p_0),
                 sig.level = alpha,
                 power = 1-beta,
                 alternative = "greater")
p_out
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.1199023
##               n = 430.044
##       sig.level = 0.05
##           power = 0.8
##     alternative = greater
```



One-tailed versus two-tailed hypothesis

So far, in our example, we've used a one-tailed hypothesis:

- $H_0: p = 0.2$
- $H_1: p > 0.2$

Depending on context, we can also have the other direction:

- $H_0: p = 0.2$
- $H_1: p < 0.2$

Or a two sided hypothesis:

- $H_0: p = 0.2$
- $H_1: p \neq 0.2$

Two tailed hypothesis in the vaccination example

If we were to use a two tailed hypothesis (the attraction had some effect, either increase or decrease vaccination rates), then our rejection criteria would become:

- Reject H_0 if $z_0 > z_{1-\alpha/2}$ or $z_0 < z_{\alpha/2}$
- I.e., for $\alpha = 0.05$:
- Reject H_0 if $z_0 > 1.96$ or $z_0 < -1.96$

The p-value would be in this case:

$$\text{p-value} = 2[1 - \Phi(|z_0|)] = 0.0244$$

```
2*(1-pnorm(q = 2.25))
```

```
## [1] 0.02444895
```

The relationship between hypothesis testing and confidence intervals

For confidence intervals of the proportion p we used:

$$1-\alpha = P\left(\hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right)$$

(We had similar confidence intervals for the average of a population normally distributed, for variance, etc.)

Negating the expression ... within the probability $P(\ldots)$, we get:

$$\alpha = 1-(1-\alpha) = 1 - P\left(\hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right) =$$

$$= P\left(\hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n} \geq p\right) + P\left(p \leq \hat{p} + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right)$$

When the parameter under H_0 is "outside" a 95% confidence interval based on the sample, the null hypothesis is rejected at $\alpha = 0.05$

- E.g., $p = 0.2$ in the previous example is outside the confidence interval for $p \in (0.216, 1)$.

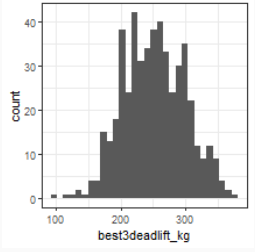
Hypothesis testing for expectancy - normal distribution

In this example, we implement a two sided hypothesis test on the power lifting data.

```
# data source:
#https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-10-08/ipf_lifts.csv
male_deadlift <- read_csv("data/ipf_lifts.csv", col_types = cols()) %>%
   filter(sex == "M") %>%
   select(best3deadlift_kg) %>%
   filter(best3deadlift_kg > 0) %>%
   sample_n(500)

ggplot(male_deadlift, aes(best3deadlift_kg)) +
   geom_histogram() +
   theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Hypothesis testing for expectancy - normal distribution - power lifting

Parameter μ is the expected weight a power lifter can pull in a deadlift

- H_0: $\mu = 225$ kg
- H_0: $\mu \neq 225$ kg
- $\alpha = 0.05$
- Test statistic $T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
- $T_0 < t_{\alpha/2, n-1}$ or $T_0 > t_{1-\alpha/2, n-1}$

```r
t.test(x = male_deadlift$best3deadlift_kg,
   alternative = "two.sided",
   mu = 225, conf.level = 0.95)
```

```
##
##     One Sample t-test
##
## data:  male_deadlift$best3deadlift_kg
## t = 12.632, df = 499, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 225
## 95 percent confidence interval:
##  248.0030 256.4768
## sample estimates:
## mean of x
##  252.2399
```

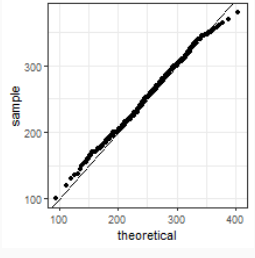
Comparing distributions - qqplot

So far, we discussed tests related to parameters' values, however, sometimes we want to compare entire distributions. For example, is the deadlift max weight normally distributed?

A qqplot draws the sample (on y axis) versus the theoretical distribuion (on x-axis). If the two are on $y = x$ this means that the distributions match.

Would you say that the following are from the same distribution?

```
ggplot(male_deadlift, aes(sample = best3deadlift_kg)) +
   geom_qq(distribution = stats::qnorm, dparams = list(mean = 250, sd = 50)) +
   theme_bw() +
   geom_abline(slope = 1, intercept = 0)
```

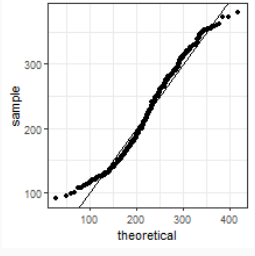


Comparing distributions - qqplot - another example

Now with both genders.

```
set.seed(0)
deadlift <- read_csv("data/ipf_lifts.csv", col_types = cols()) %>%
   select(best3deadlift_kg) %>%
   filter(best3deadlift_kg > 0) %>%
   sample_n(500)

ggplot(deadlift, aes(sample = best3deadlift_kg)) +
   geom_qq(distribution = stats::qnorm, dparams = list(mean = 223, sd = 63)) +
   theme_bw() +
   geom_abline(slope = 1, intercept = 0)
```



Looks like there's a heavy tail towards higher weights in the sample than the theoretical. How do we quantify this as a statistical hypothesis test of "goodness of fit"?

Hypothesis testing - goodness of fit

Goodness of fit tests are used to test how good is the fit of our empirical distribution to that of a theoretical distribution.

Arrange the empirical distribution in k bins, and let O_i be the observed frequency in the ith class bin. Let E_i be the expected probability. The test statistic is:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

If the population follows the hypothesized distribution, then the expression is approximately distributed χ^2_{k-p-1}, where p os the number of parameters of the hypothesized distribution estimated by sample statistics.

The approximation improves as n increases.

When using the approximation, make sure that E_i is "big enough" (i.e., $E_i \geq 5, \forall i$)

Hypothesis testing - goodness of fit - procedure

We would reject the hypothesis if $\chi_0^2 > \chi^2_{\alpha,k-p-1}$.

1. We are interested in the form of the distribution of maximum deadlift weight
2. H_0: The deadlift weight is normally distributed
3. H_1: The deadlift weight is not normally distributed
4. $\alpha = 0.05$
5. The test statistic is: $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
6. Reject H_0 if $\chi_0^2 > \chi^2_{0.05,k-p-1}$

Goodness of fit - example

```r
interval_breaks <- c(0, 120, 190, 250, 320, 600)
sample_size <- 500

deadlift_gfit <- deadlift %>%
   mutate(weight_gr = cut(best3deadlift_kg, breaks = interval_breaks))

mu <- mean(deadlift_gfit$best3deadlift_kg)
sigma <- sd(deadlift_gfit$best3deadlift_kg)

deadlift_chi_prep <- deadlift_gfit %>%
   count(weight_gr, name = "observed") %>%
   mutate(upper_bound = interval_breaks[-1]) %>%
   mutate(lower_bound = interval_breaks[1:5]) %>%
   mutate(expected_prob = pnorm(q = upper_bound, mean = mu, sd = sigma)-
            pnorm(q = lower_bound, mean = mu, sd = sigma)) %>%
   mutate(expected_prob =  expected_prob/sum(expected_prob)) %>%
   mutate(expected = expected_prob*500) %>%
   mutate(chi_comp = (observed-expected)^2/expected)

chi2_0 <- sum(deadlift_chi_prep$chi_comp)
chi2_0
```

```
## [1] 26.35624
```

```r
qchisq(p = 1-0.05, df = 5-1)
```

```
## [1] 9.487729
```

Since $26.3 = \chi^2_0 > \chi^2_{500,3} = 9.49$ we reject the null hypothesis. This distribution is not normal.

What is the p-value? (really small)

```r
1-pchisq(q = chi2_0, df = 5-1)
```

Goodness of fit - example - using R's *chisq.test* command

```
chisq.test(x = deadlift_chi_prep$observed,
           p = deadlift_chi_prep$expected_prob)
```

```
##
##     Chi-squared test for given probabilities
##
## data:  deadlift_chi_prep$observed
## X-squared = 26.356, df = 4, p-value = 2.682e-05
```

Words of caution

Two common mistakes when analyzing data are:

1. HARKing (Hypothesizing After Results are Known)
2. Using multiple comparisons without compensating for them

In essense, Since we have a α error rate (e.g. 5%), when we perform 100 hypothesis tests we are bound to get 5% false positives.

For (1) that's something you shouldn't do! Some organizations offer *preregistration* as means of avoiding HARKing and improving research. I.e., https://cos.io/prereg/

For (2) there are procedures to control the false discovery rate (FDR). In R, there is a command called `p.adjust` for example. We might talk about this, if time permits.

Abusing hypothesis tests - example (don't try this at home!)

```r
set.seed(0)

# let's get random numbers
random_normal <- matrix(rnorm(n=100*100, mean = 0, sd = 1),
                        nrow = 100, ncol = 100) %>%
   as_tibble(.name_repair = "unique")
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ... and 95 more problems
```

Abusing hypothesis tests - example (don't try this at home!)

```r
# we have 100 variables which are standard normal distributed
# let's take the one with the highest average
random_normal %>%
   pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
   group_by(variable) %>%
   summarize(mean_val = mean(value)) %>%
   arrange(desc(mean_val))
```

```
## # A tibble: 100 x 2
##    variable mean_val
##    <chr>       <dbl>
##  1 ...78       0.287
##  2 ...28       0.244
##  3 ...35       0.233
##  4 ...53       0.228
##  5 ...60       0.224
##  6 ...100      0.211
##  7 ...97       0.191
##  8 ...77       0.173
##  9 ...99       0.166
## 10 ...67       0.165
## # ... with 90 more rows
```

```r
# H_0: this variable (..78) has mu=0, H_1: otherwise

t.test(random_normal$...78)
```

```
##
##     One Sample t-test
##
## data:  random_normal$...78
## t = 2.5858, df = 99, p-value = 0.01117
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06668831 0.50660635
## sample estimates:
## mean of x
```