Unit 6: Inference for Numerical Data

Statistics S-100 Teaching Team

Summer 2024

Introduction

Two-sample test for paired data

Two-sample test for independent group data

Statistical power and sample size

Comparing several means with ANOVA

A closer look at the *p*-value

Introduction

Comparing two population means

Two-sample data can be paired or unpaired (independent).

- Paired measurements for each 'participant' or study unit
  - ◇ each observation can be logically matched to one other observation in the data
  - ◇ e.g., scores on a standardized test before taking a prep course versus scores after the prep course
- Two independent sets of measurements
  - ◇ observations cannot be matched on a one-to-one basis
  - ◇ e.g., scores on a standardized test of students who did take a prep course versus scores of students who did not

The nature of the data dictate which two-sample testing procedure is appropriate: the two-sample test for paired data, or the two-sample test for independent group data.

Two-sample test for paired data

WETSUITS AND SWIMMING VELOCITY

Did a new wetsuit design allow for increased swim velocities during the 2000 Olympics?

A study was conducted to assess the effect of wearing a wetsuit on swim velocity.

- 12 competitive swimmers were asked to swim 1500m at maximal velocity, once wearing a wetsuit and once wearing a standard swimsuit.
- Order of wetsuit versus swimsuit randomized.
- Investigators recorded mean velocity (m/sec) for each trial.

```r
library(oibiostat)
data("swim")
swim_new <- swim
colnames(swim_new) <- c("id", "ws_vel",
                        "ss_vel", "vel_diff")
swim_new
```

```
##    id ws_vel ss_vel vel_diff
## 1   1   1.57   1.49     0.08
## 2   2   1.47   1.37     0.10
## 3   3   1.42   1.35     0.07
## 4   4   1.35   1.27     0.08
## 5   5   1.22   1.12     0.10
## 6   6   1.75   1.64     0.11
## 7   7   1.64   1.59     0.05
## 8   8   1.57   1.52     0.05
## 9   9   1.56   1.50     0.06
## 10 10   1.53   1.45     0.08
## 11 11   1.49   1.44     0.05
## 12 12   1.51   1.41     0.10
```

Idea behind inference for paired data

The velocities are paired by swimmer: each swimmer has two velocity measurements.

Suppose that for each swimmer i, we have observations $x_{i,WS}$ and $x_{i,SS}$.

- Let d_i be the difference between the measurements:

$$d_i = x_{i,WS} - x_{i,SS}$$

  - ◇ $x_{i,WS}$ is the wetsuit velocity measurement for swimmer i
  - ◇ $x_{i,SS}$ is the swimsuit velocity measurement for swimmer i

- Base inference on $\overline{d}$, the sample mean of the individual differences d_i:

$$\overline{d} = \frac{\sum d_i}{n}$$

Inference for paired data

Let δ be the population mean of the difference in velocities during a 1500m swim if all competitive swimmers recorded swim velocities with each suit type.

The null and alternative hypotheses are

- $H_0 : \delta = 0$, the average difference in swim velocities between swimming with a wetsuit versus a swimsuit equals 0
    - ◇ i.e., wetsuits do not change swim velocities
- $H_A : \delta \neq 0$, the average difference in swim velocities between swimming with a wetsuit versus a swimsuit is non-zero
    - ◇ i.e., wetsuits do change swim velocities

We can also compute a 95% confidence interval for δ.

Inference for paired data . . .

The formula for the test statistic is

$$t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{n}},$$

where $\overline{d}$ is the mean of the differences, s_d is the standard deviation of the differences, and n is the number of differences (i.e., number of pairs).

- A paired t-test is essentially a one-sample test of difference values.
- The p-value is calculated from a t distribution with $df = n - 1$.

A 95% confidence interval for paired data has the form

$$\overline{d} \pm \left(t^{\star} \times \frac{s_d}{\sqrt{n}}\right),$$

where $t^{\star}$ is the point on a t distribution with $df = n - 1$ that has area 0.025 to its right.

Letting R do the work

```r
#two-sample syntax
t.test(swim$wet.suit.velocity, swim$swim.suit.velocity,
       alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  swim$wet.suit.velocity and swim$swim.suit.velocity
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean difference
##          0.0775
```

Note: `t.test(x, y, paired = TRUE)` returns results based on the differences `x - y`.

Letting R do the work...

```r
#one-sample syntax
t.test(swim$velocity.diff, mu = 0, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  swim$velocity.diff
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of x
##    0.0775
```

Two-sample test for independent group data

COVID-19 vaccine acceptance

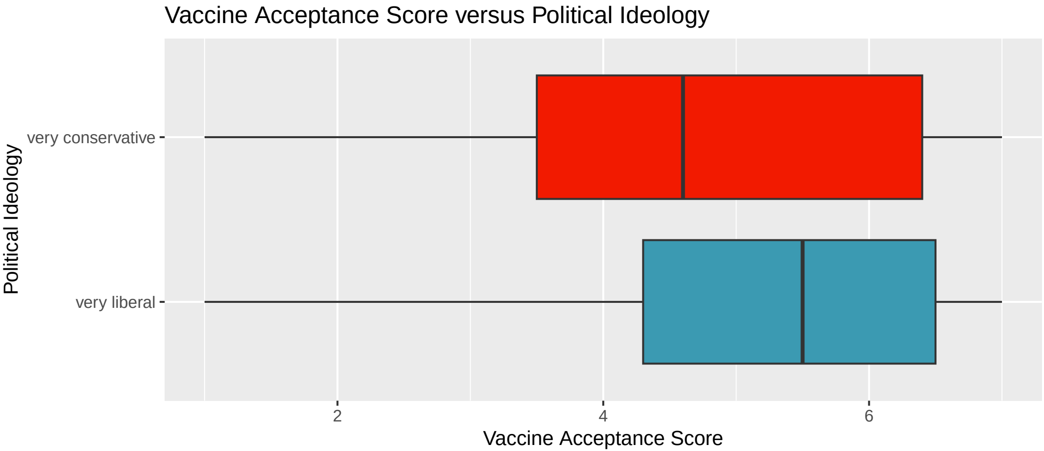
A study conducted in July 2020[1] investigated factors associated with self-reported willingness to receive a COVID-19 vaccine. Researchers collected data from 1,971 participants via a 15-minute online survey that described 10 hypothetical vaccines.

- For example, one hypothetical vaccine was described as being 70% effective, approved and licensed by the US Food and Drug Administration, developed in the United States, and endorsed by Vice President Joe Biden.
- A vaccine acceptance score from 1 to 7 was calculated, representing the average response to the question "How likely or unlikely would you be to get Vaccine X?" for the 10 hypothetical vaccines.
- A higher score represents greater willingness.

Does mean vaccine acceptance score differ between respondents who described their political ideology as "very liberal" versus "very conservative"?

[1] This was before COVID-19 vaccines were available.

COVID-19 vaccine acceptance. . .



Inference for comparing two means

The parameter of interest is $\mu_{VL} - \mu_{VC}$, the difference in population mean vaccine acceptance score for those identifying as "very liberal" versus "very conservative".

The null and alternative hypotheses are

- $H_0 : \mu_{VL} = \mu_{VC}$, the population mean vaccine acceptance score is the same for those identifying as "very liberal" versus "very conservative".
- $H_A : \mu_{VL} \neq \mu_{VC}$, the population mean vaccine acceptance score differs between those identifying as "very liberal" versus "very conservative".
- Equivalently, $H_0 : \mu_{VL} - \mu_{VC} = 0$ and $H_A : \mu_{VL} - \mu_{VC} \neq 0$.

In general, the hypotheses are written in terms of μ_1 and μ_2.[2]

- The parameter of interest is $\mu_1 - \mu_2$.
- The point estimate is $\overline{x}_1 - \overline{x}_2$.

[2] The numerical labels are arbitrary, so it is best to explicitly specify which group is considered group 1 versus group 2.

Inference for comparing two means...

The t-statistic is:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

The degrees of freedom for the null distribution are different than for the paired data setting.

- Use this approximation when doing the test by hand: $df = \min(n_1 - 1, n_2 - 1)$.
- R uses a better approximation called the Satterthwaite approximation:

$$df = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]}$$

The 95% confidence interval for the difference in population means has the form

$$(\overline{x}_1 - \overline{x}_2) \pm \left(t^\star \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

where $t^\star$ is the point on a t distribution that has area 0.025 to the right, with the same degrees of freedom as used for calculating the p-value of the associated test.

Letting R do the work

The tilde syntax is used when one vector contains the numerical variable and one vector contains the grouping variable.

```r
#tilde syntax
t.test(vax$likely ~ vax$ideology, mu = 0, paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  vax$likely by vax$ideology
## t = 4.4389, df = 423.85, p-value = 1.154e-05
## alternative hypothesis: true difference in means between group very liberal and group very conserv
## 95 percent confidence interval:
##  0.3873987 1.0031275
## sample estimates:
##     mean in group very liberal mean in group very conservative
##                       5.267829                        4.572566
```

Letting R do the work...

The comma syntax is used when there are two vectors of numerical data.

```
#comma syntax
t.test(vax$likely[vax$ideology == "very liberal"],
       vax$likely[vax$ideology == "very conservative"])
```

```
##
##  Welch Two Sample t-test
##
## data:  vax$likely[vax$ideology == "very liberal"] and vax$likely[vax$ideology == "very conservati
## t = 4.4389, df = 423.85, p-value = 1.154e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3873987 1.0031275
## sample estimates:
## mean of x mean of y
##  5.267829  4.572566
```

Statistical power and sample size

Outcomes and errors in testing

Hypothesis tests can potentially result in incorrect decisions. The following table shows the four possible ways that the conclusion of a hypothesis test can be right or wrong.

	Result of test	
State of nature	**Reject H_0**	**Fail to reject H_0**
H_0 is true	Type I error, probability $= \alpha$ (false positive)	No error, probability $= 1 - \alpha$ (true negative)
H_A is true	No error, probability $= 1 - \beta$ (true positive)	Type II error, probability $= \beta$ (false negative)

In a trial, the defendant is either innocent or guilty. After hearing evidence from both the prosecution and the defense, the court must reach a verdict.

Under many legal systems, presumption of innocence is a legal right of the accused.

- H_0: The defendant is innocent.
- H_A: The defendant is guilty.

What does a Type I error represent in this context?

What does a Type II error represent in this context?

Error probabilities in hypothesis testing

Lab 2 uses simulation to explore how Type I and Type II error are controlled.

- Type I error is controlled via rejecting H_0 only when a p-value is smaller than α.

Statisticians are most often concerned with the complement of making a Type II error.

- The **power** of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true.
- In other words, power is the probability of correctly rejecting H_0.

Factors affecting power

The power of a statistical test is determined by[3]

- the hypothesized difference between two population means, also known as the population **effect size** ($|\mu_1 - \mu_2|$)
- the population standard deviation of each group (σ_1, σ_2)
- the sample size of each group (n_1, n_2)

Think about power as a measure of how easy it is to distinguish whether two groups have different (population) means.

Usually, a study team can only control sample size.

[3] The significance level α also influences power. More on this in Lab 2...

Example: Treatment for Alzheimer's

Treatment for Alzheimer's disease is an active research area. The Dementia Severity Rating Scale (DSRS) measures impairment severity of a person with Alzheimer's. Scores range from 0 (indicating no impairment) to 54 (extreme impairment).

Cognitive decline over several years is measured by annual rate in change in DSRS score.

- Example: a person observed for 3 consecutive years whose score increased from 7 to 14.5 has an annual rate of change of $7.5/3 = 2.5$ points per year.

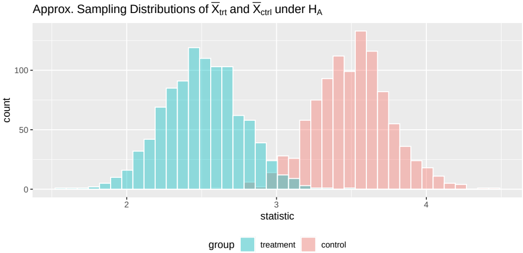
Suppose a pharmaceutical company developed a drug to slow cognitive decline in individuals affected by Alzheimer's disease.

- Conduct a randomized trial, comparing standard drug to new drug.
- Enroll newly diagnosed patients, measure DSRS at beginning of study and 3 years later.
- Compare average annual change in DSRS score between the two groups.

Test $H_0 : \mu_{ctrl} = \mu_{trt}$ against $H_A : \mu_{ctrl} \neq \mu_{trt}$.

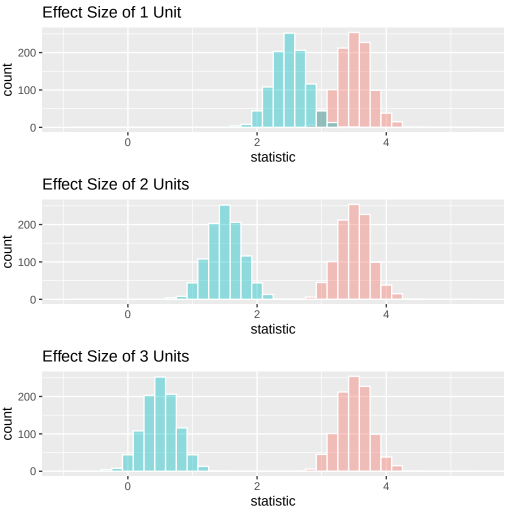
Intuition behind power

- Suppose H_A is true and that we know the population distribution of DSRS score among the control group and treatment group:
  - Control: $\mu_{ctrl} = 3.5$ pts/yr
  - Treatment: $\mu_{trt} = 2.5$ pts/yr
- What would the sampling distributions of $\overline{X}_{trt}$ and $\overline{X}_{ctrl}$ look like?
- If we only get to observe one random sample from each group, how likely would we be to conclude that the population means are different?
  - i.e., How likely would we be to **reject** H_0 **correctly**?



Intuition behind power: effect size

- **Effect size** ($|\mu_1 - \mu_2|$) refers to the hypothesized difference between two population means.
- On the previous slide, we supposed that μ_{ctrl} $= 3.5$ pts/yr and $\mu_{trt} = 2.5$ pts/yr, which corresponds to effect size $\Delta = |\mu_{ctrl} - \mu_{trt}| = |3.5 - 2.5| = 1$ pt/yr.
- What if the effect size were larger; i.e., if the new drug were actually more effective?
  - ◇ What if $\mu_{trt} = 1.5$ pts/yr, $\Delta = 2$ pts/yr?
  - ◇ What if $\mu_{trt} = 0.5$ pts/yr, $\Delta = 3$ pts/yr?
- As effect size increases, does power increase or decrease?
  - ◇ As effect size increases, does it become easier or harder to detect a difference in population means?



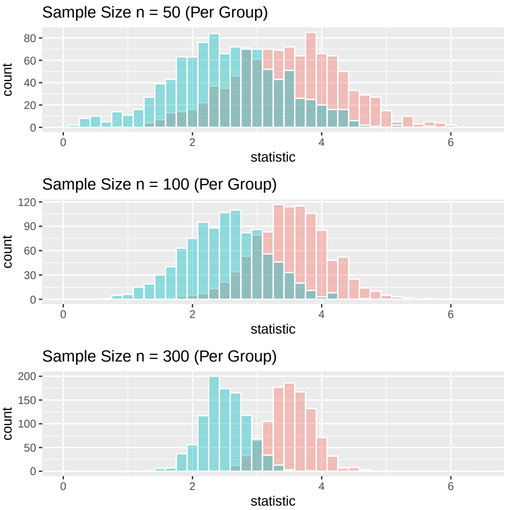
Effect Size of 1 Unit

Effect Size of 2 Units

Effect Size of 3 Units

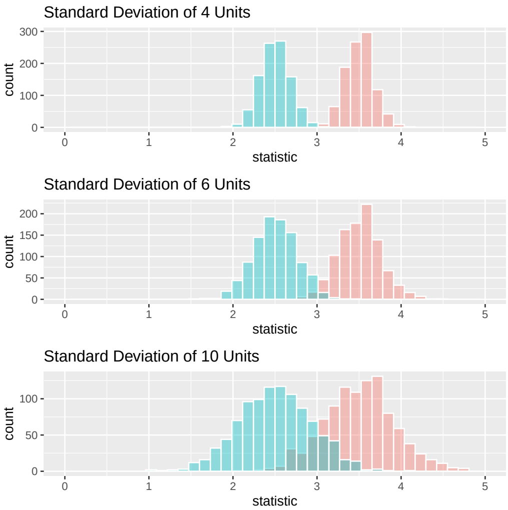
Intuition behind power: sample size

- How does **sample size** affect power?
- Hint: recall that the standard error of $\overline{X}$ depends on n.
- Let's run some simulations with different sample size.
  - What if $n_{trt} = n_{ctrl} = 50$?
  - What if $n_{trt} = n_{ctrl} = 100$?
  - What if $n_{trt} = n_{ctrl} = 300$?
- As sample size increases, does power increase or decrease?
  - As sample size increases, does it become easier or harder to detect a difference in population means?



Intuition behind power: standard deviation

- How does **standard deviation** $(\sigma_{trt}, \sigma_{ctrl})$ affect power?
- Hint: recall that the standard error of $\overline{X}$ depends on σ.
- Let's run some simulations with different standard deviation of DSRS score.
  - What if $\sigma_{trt} = \sigma_{ctrl} = 4$?
  - What if $\sigma_{trt} = \sigma_{ctrl} = 6$?
  - What if $\sigma_{trt} = \sigma_{ctrl} = 10$?
- As standard deviation increases, does power increase or decrease?
  - As standard deviation increases, does it become easier or harder to detect a difference in population means?


Standard Deviation of 4 Units

Standard Deviation of 6 Units

Standard Deviation of 10 Units

Background for power simulation lab...

Part I: Controlling Type I error

- Assume that $H_0 : \mu_{ctrl} = \mu_{trt}$ is true.
- Draw repeated samples from simulated control and treatment populations.
- Conduct a hypothesis test for each set of two samples.
- How many samples resulted in the **correct decision**?

Part II: Controlling Type II error

- Assume that $H_A : \mu_{ctrl} \neq \mu_{trt}$ is true.
- Draw repeated samples from simulated control and treatment populations.
- Conduct a hypothesis test for each set of two samples.
- How many samples resulted in the **correct decision**?
  - ◇ How does this number change when we change the sample size? ...the standard deviation? ...the effect size?

Choosing the right sample size

Study design often includes calculating a sample size such that the probability of rejecting a null hypothesis correctly is acceptably large, typically between 0.80 and 0.90.

It is important to have a precise estimate of an appropriate study size.

- A study needs to be large enough to allow for sufficient power to detect a difference between groups.
- However, unnecessarily large studies are expensive, and can even be unethical.

A typical sample size question

A pharmaceutical company has developed a new drug to lower blood pressure and is planning a clinical trial to test its effectiveness.

- Individuals whose systolic blood pressure is between 140 and 180 mmHg will be recruited.
- Based on previous studies, blood pressures from such individuals will be approximately normally distributed with standard deviation of about 12 mmHg.
- Participants will be randomly assigned to the new drug or a standard drug, and the company will measure the difference in mean blood pressure levels between the groups.

The company expects to receive FDA approval for the drug if there is evidence that the drug lowers blood pressure, on average, by at least 3 mmHg more than the standard drug.

How large should the study be if the company wants the power of the study to be 0.80 (80%)?

A typical sample size question . . .

We will use the `power.t.test()` function for power and sample size calculations.[4]

To achieve a power of at least 0.80, the company should recruit at least 253 patients for each group; i.e., a total of at least 506 patients.

```
power.t.test(n = NULL, delta = 3, sd = 12, sig.level = 0.05, power = 0.80)
```

```
##
##      Two-sample t test power calculation
##
##               n = 252.1281
##           delta = 3
##              sd = 12
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

[4] *OI Biostat* Section 5.4 has an extended discussion of this example, with formulas for hand calculations.

Comparing several means with ANOVA

Type I error rate for a single test

Suppose we are interested in comparing means across more than two groups. Why not conduct several two-sample *t*-tests?

- If there are k groups, then $\binom{k}{2} = \frac{k(k-1)}{2}$ t-tests are needed.
- Conducting multiple tests on the same data increases the overall rate of Type I error.

Recall that making a Type I error (rejecting H_0 when H_0 is true) occurs with probability α.

- Type I error rate is controlled by rejecting only when a test p-value is smaller than α.
- α is typically kept low.
- With a single two-group comparison at $\alpha = 0.05$, there is a 5% chance of incorrectly identifying an association where none actually exists.

What about several tests?

What happens to Type I error when making several comparisons?

When conducting more than one t-test in an analysis. . .

- The significance level (α) used in each test controls the error rate for that test.
- The **experiment-wise error rate** is the chance that at least one test will incorrectly reject H_0 when all tested null hypotheses are true.
- Controlling experiment-wise error rate is one specific approach for controlling Type I error.

Probability of experiment-wise error

Suppose that two t-tests are conducted to assess whether two new drug candidates are associated with better outcome. Assume the tests are independent and each are conducted at the $\alpha = 0.05$ significance level.

Let A be the event of making a Type I error on the first test, and B be the event of making a Type I error on the second test, where $P(A) = P(B) = 0.05$.

The probability of making at least one error is equal to the complement of the event that a Type I error is not made with either test.

$$1 - [P(A^C)P(B^C)] = 1 - (1 - 0.05)^2 = 0.0975$$

Thus, when making two independent t-tests, there is about a 10% chance of making at least one Type I error; the experiment-wise error is 10%.

Probability of experiment-wise error...

With 10 tests...

$$\text{experiment-wise error } = 1 - (1 - 0.05)^{10} = 0.401$$

With 25 tests...

$$\text{experiment-wise error } = 1 - (1 - 0.05)^{25} = 0.723$$

With 100 tests...

$$\text{experiment-wise error } = 1 - (1 - 0.05)^{100} = 0.994$$

With 100 independent tests, there is a 99.4% chance an investigator will make at least one Type I error!

- If a company tested 100 drug candidates (that actually aren't effective), there is a 99.4% chance of incorrectly concluding that at least one candidate is effective.

Analysis of Variance (ANOVA)

ANOVA uses a single hypothesis test to assess whether means across several groups are equal:

- H_0: mean outcome is same across all groups ($\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$)
- H_A: at least one mean is different from the others (i.e., means are not all equal)

By using a single test, the Type I error rate is still maintained at α.

Idea behind ANOVA

Are the sample means different enough that H_0 seems unlikely to be true?

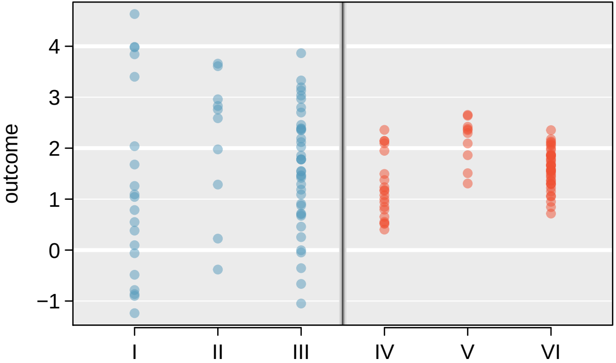
- i.e. Does it seem like the sample means come from different populations, or could it be plausible that they are simply different samples drawn from one population?

Compare two quantities:

- Variability between groups (MSG): how different are the group means from each other, i.e., how much does each group mean vary from the overall mean?
- Variability within groups (MSE): how variable are the data within each group?

MSG denotes mean square between groups, while *MSE* denotes mean square error. Refer to *OI Biostat* Section 5.5.1 for details.

Idea behind ANOVA...



- I, II, and III: difficult to discern differences in means, variability within each group is high
- IV, V, and VI: appears to be differences in means, these differences are large relative to variance within each group

Idea behind ANOVA…

Under the null hypothesis, there is no real difference between the groups; thus, any observed variation in group means is due to chance.

- Think of all observations as belonging to a single group.
- Variability between group means should equal variability within groups.

The *F-statistic* is the test statistic for ANOVA.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{MSG}{MSE}$$

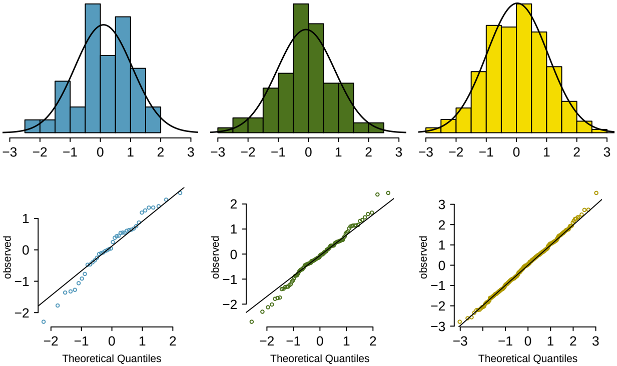
- If the F-statistic is approximately 1, this suggests that the null hypothesis is true.
- If the F-statistic is (much) larger than 1, this suggests that the population means are different.
- The F-statistic follows an F distribution, with two degrees of freedom, df_1 and df_2; $df_1 = n_{groups} - 1$, $df_2 = n_{obs} - n_{groups}$.
- The p-value for ANOVA equals $P(F \geq F\text{-stat})$.

Assumptions for ANOVA

Important to check whether the assumptions for conducting ANOVA are reasonably satisfied:

1. Observations are independent within and across groups.
   - Think about study design/context.
2. Data within each group are approximately normal.
   - Look at the data graphically, such as with a histogram.
   - Normal Q-Q plots can help...
3. Variability across groups is about equal.
   - Look at the data graphically.
   - Numerical rule of thumb: ratio of largest variance to smallest variance < 3 is considered "about equal".

Normal probability plots (Q-Q plots)

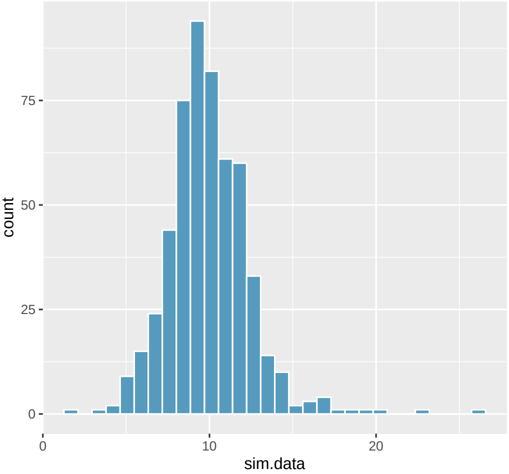


If points fall on or near the line, data closely follow a normal distribution.

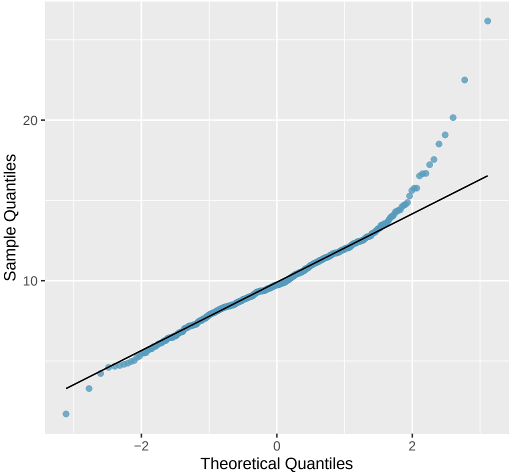
- Difficult to evaluate in small datasets.
- Plots show three simulated normal datasets: from L to R, $n = 40$, $n = 100$, $n = 400$

Normal probability plots (Q-Q plots)...

Histogram of Simulated Right−Skewed Data



Normal Q−Q Plot



Pairwise comparisons

If the F-test indicates there is sufficient evidence that the group means are not all equal, proceed with pairwise comparisons to identify which group means are different.

Pairwise comparisons are made using the two-sample t-test for independent groups.

- To maintain the overall Type I error rate at α, each pairwise comparison is conducted at at an adjusted significance level referred to as $\alpha^\star$.
- The Bonferroni correction is one method for adjusting α.

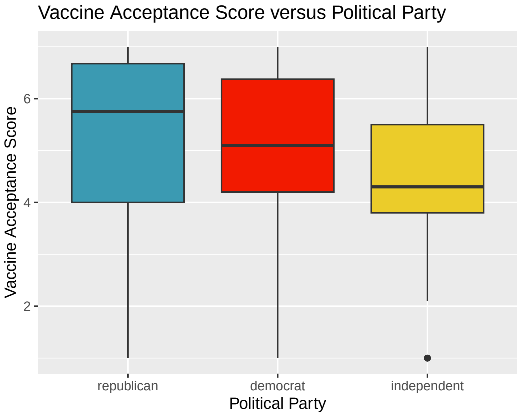
$$\alpha^\star = \alpha / K, \text{ where } K = \frac{k(k-1)}{2} \text{ for } k \text{ groups}$$

- Note that the Bonferroni correction is a very stringent (i.e., conservative) correction, made under the assumption that all tests are independent.

COVID-19 vaccine acceptance, again

With ANOVA, we can address a different version of the question addressed in the two-group context.

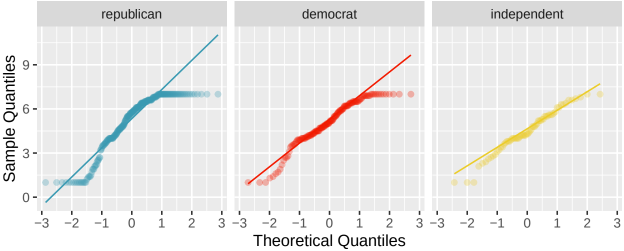
- Does mean vaccine acceptance score differ by political party?
- Respondents reported political party as either Republican, Democrat, or Independent.



COVID-19 vaccine acceptance, again. . .

Check assumptions:

1. *Independence*. Data are from participants taking an online survey, so it is reasonable to consider the responses as independent
2. *Approximate normality*. Q-Q plots indicate some departures from normality for the extreme values, but observations mostly follow normality in the centers
3. *Approximately constant variance*. ratio of largest to smallest variance is 1.65, which is considered "about equal"



```
#check variance
tapply(vax$likely, vax$party3, var)
```

```
##  republican    democrat independent
##    3.439048    2.163715    2.083126
```

```
3.44/2.08
```

```
## [1] 1.653846
```

COVID-19 vaccine acceptance, again...

State hypotheses.

- $H_0 : \mu_I = \mu_D = \mu_R$, mean vaccine acceptance score is equal across political party.
- H_A: at least one group has a mean vaccine acceptance score different from that of the other groups.

Let $\alpha = 0.05$.

```
summary(aov(vax$likely ~ vax$party3))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## vax$party3    2   23.7  11.864   4.188 0.0158 *
## Residuals   459 1300.4   2.833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value from the ANOVA F-test is 0.0158.
- If the population mean vaccine acceptance scores were actually the same across groups, the probability of observing an F-statistic as large as 4.19 or larger equals 0.0158.
- These data suggest that at least one population mean vaccine acceptance score is different from the others.

Controlling Type I error rate

If the ANOVA F-test is significant, then it is appropriate to proceed to conducting pairwise comparisons; i.e., using two-sample t-tests to compare each possible pairing of the groups.[5]

- Each test should be conducted at the $\alpha^\star$ significance level so that the overall Type I error rate remains at α.
- These tests are still conducted under the assumption that the variance between groups is equal; thus, the test statistics are calculated using the pooled estimate of standard deviation between groups. Details are in *OI Biostat* Section 5.5.3.
- We will use `pairwise.t.test( )` to perform these *post hoc* two-sample t-tests.

[5] These t-tests are typically referred to as *post hoc* tests.

Post hoc testing

Each test should be conducted at the $\alpha^\star = 0.05/3 = 0.0167$ significance level since there are 3 comparisons being made.

- There is evidence that Independents are different from both Republicans and Democrats.
- There is not evidence that Republicans and Democrats are different.

```
pairwise.t.test(vax$likely, vax$party3, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  vax$likely and vax$party3
##
##             republican democrat
## democrat    0.7424     -
## independent 0.0046     0.0149
##
## P value adjustment method: none
```

Letting R do the work. . .

Setting `p.adj` to `"bonf"` instructs R to rescale the *p*-values (by multiplying by K, the total number of comparisons) so they can be compared to the original α level of 0.05.

```r
pairwise.t.test(vax$likely, vax$party3, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  vax$likely and vax$party3
##
##             republican democrat
## democrat    1.000      -
## independent 0.014      0.045
##
## P value adjustment method: bonferroni
```

A closer look at the *p*-value

Fisher's p-value

In 1925, Ronald Fisher published *Statistical Methods for Research Workers*, a book aimed at providing biologists with the means of applying statistical tests accurately to numerical data.

- "The value for which $P = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. . . "
- Provided a valuable rule of thumb at a time when the use of tables and approximations were essential for practicing researchers

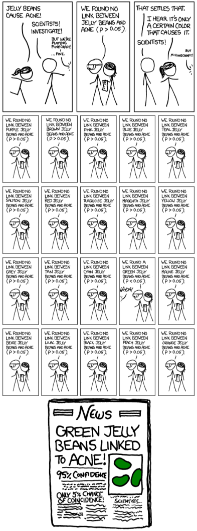
Over time, the p-value in conjunction with the hypothesis testing framework developed by Jerzy Neyman and Egon Pearson became widely used across many scientific disciplines.

The mis-use of *p*-values

"*P*-hacking" refers to the conscious or unconscious manipulation of data in order to achieve a desired result (i.e., a significant *p*-value):

- There are many possible decisions in the data analysis process, and no obviously correct way to proceed.
- Unconscious bias can lead researchers to make decisions that will confirm what they would like to believe.
- Non-significant results are difficult to publish in scientific journals.
- Tempting to perform many hypothesis tests and only report the statistically significant results.

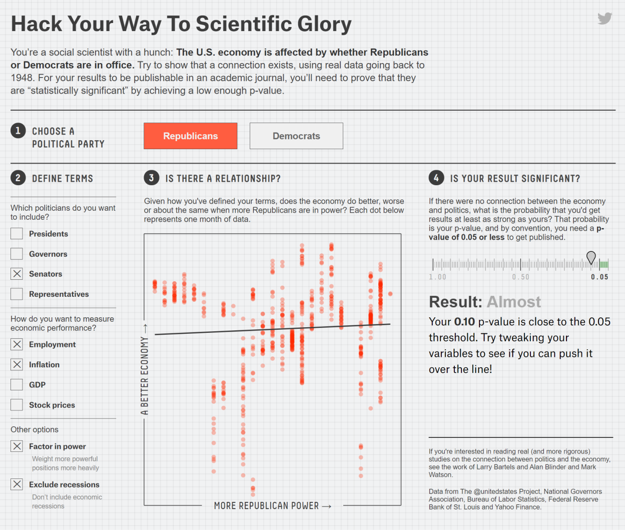
The following slide shows an interactive "data analysis" from "Science isn't Broken".[a]



JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

[a] FiveThirtyEight, 19 Aug 2015.

The mis-use of p-values...



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

[✓ Republicans] Democrats

2 DEFINE TERMS

Which politicians do you want to include?

☐ **Presidents**
☐ **Governors**
☒ **Senators**
☐ **Representatives**

How do you want to measure economic performance?

☒ **Employment**
☒ **Inflation**
☐ **GDP**
☐ **Stock prices**

Other options

☒ **Factor in power**
Weight more powerful positions more heavily

☒ **Exclude recessions**
Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in power? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.10** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Statistical significance versus practical significance

In 2016, the American Statistical Association released a formal statement clarifying principles about the proper use and interpretation of the *p*-value. Some main points mentioned:

- A *p*-value does not measure the size of an effect or the importance of a result. Statistical significance is not equivalent to scientific, human, or economic significance.
- Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold. Many other contextual factors should be considered, such as the design of a study, the external evidence for the phenomenon under study, and the validity of the assumptions that underlie the data analysis.

A 2013 study published in *PNAS* of 19,000+ people found that people who meet their spouses online. . .

- Are less likely to divorce than those who meet offline ($p < 0.002$)
- Are more likely to have higher marital satisfaction ($p < 0.001$)

Important to examine the effect sizes:

- Divorce rate of 5.96% for those who met online versus 7.67% for those who met in person.
- On a 7 point scale, happiness value of 5.64 for those who met online versus 5.48 for those who met offline.

Do these results provide compelling evidence that one should change their behavior?[a]

[a] Study results reported in a *Scientific American* article: "Meeting Your Spouse Online May Lead to a Better Marriage".

Modern solutions

Emphasizing reproducibility and replicability.

- Making data and analysis code public.
- Publishing negative/inconclusive results.

Reporting analysis plan before conducting experiments.

- e.g., a two-stage peer review process in which methods and proposed analyses are peer-reviewed prior to data collection and analysis

Focusing interpretation of data on effect sizes and confidence intervals.

- Asking "How much of an effect is there?" versus "Is there an effect?"

Moving to a World Beyond "$p < 0.05$"

- Editorial accompanying a special issue of *The American Statistician* from Mar 2019, containing 43 papers discussing how to move to "a world where researchers are free to treat '$\mathrm{p} = 0.051$' and '$\mathrm{p} = 0.049$' as not being categorically different."

What should you keep in mind?

When conducting research. . .

- Clearly specify details of an analysis approach (e.g., primary outcome, significance level, number of tests) *before* viewing data.
- Understand how to appropriately use and interpret *p*-values.
- Report interval estimates, not just *p*-values.
- When possible, replicate an analysis with new data.

When consuming research. . .

- Be wary of strong conclusions based solely on a *p*-value.
- Consider factors such as study design, measurement quality, validity of any assumptions made, and plausibility (e.g., evidence for a biological mechanism).
- Avoid confusing statistical significance with practical significance.