Data Analysis

Adair Neto

2023-12-19

Designing Studies

What is the population of interest? And what is the sample?

Observations, variables, and data matrices

Types of variables:

1. Numerical (quantitative), which can be continuous or discrete.
2. Categorical (qualitative): ordinal (levels have an inherent ordering) or regular categorical.

Identifying the type of variables we're working with is the first step in data analysis.

Relationships between variables:

1. Associated (dependent): when two variables show some connection. This can be positive or negative.
2. Independent: if two variables show no connection.

Observational studies and experiment

Observational study: the collection of data **does not directly interfere** with how the data arise. They only establish an **association**.

If it uses past data, it is called *retrospective*. If the data is collected throughout the study, it is called *prospective*.

Experiment: subjects are **randomly assigned** to treatments, for example. Then, it is possible to establish **causal connections**.

Confounding variables: extraneous variables that affect both the explanatory and the response variable, and make it seem like there is a relationship between them.

Sampling and Sources of Bias

Why not conduct a census?

1. Some individuals are hard to locate or measure.
2. Populations rarely stand still.

Taking a sample is like testing the seasoning of a soup. We taste a small amount of it (*exploratory analysis*). If it needs more salt, we conclude that the whole soup needs more salt (*inference*). However, for the inference to be correct, our sample needs to be a *representative sample*.

A few sources of sampling bias

1. *Convenience sample:* individuals who are easily accessible are more likely to be included in the sample.
2. *Non-response:* if only a non-random fraction of the randomly sample people respond to a survey such that the sample is no longer representative of the population.
3. *Voluntary response:* when the sample consists of people who volunteer to respond because they have strong opinions on the issue.

Sampling methods

1. *Simple random sample (SRS):* each case is equally likely to be selected.
2. *Stratified sample:* divide the population into homogeneous *strata*, then randomly sample from within each stratum.
3. *Cluster sample:* divide the population into *clusters*, randomly sample a few clusters, then sample all observations within these clusters.
4. *Multistage sample:* divide the population into *clusters*, randomly sample a few clusters, then randomly sample within these clusters.

Experimental Design

Principles of experimental design:

1. *Control:* compare treatment of interest to a control group.
2. *Randomize:* randomly assign subjects to treatments.
3. *Replicate:* collect a sufficiently large sample, or replicate the entire study.
4. *Block:* block for variables known or suspected to affect the outcome.

Blocking vs. explanatory variables:

Explanatory variables (factors) Are conditions we can impose on experimental units.

Blocking variables Are characteristics that the experimental units come with, that we would like to control for.

Some terminology:

Blinding: Experimental units don't know which group they're in.

Double-blind: Both the experimental units and the researchers don't know the group assignment.

Random sampling

Random sampling takes a random sample of the whole population. Hence, it can be generalized.

Random assignment occurs only in experimental settings. Allows us to make causal conclusions.

Exploring Data

Numerical Data

What is meant by robust statistics and when they are used?

When transformations (e.g. log) can make the distribution of data more symmetric, and hence easier to model?

Evaluating the relationship between the numerical data:

1. Direction: positive or negative?
2. Shape: linear or curved?
3. Strength: strong or weak.
4. Outliers.

A good way of analyzing the distribution of a numerical variable is using a *histogram*, which provides a view of the data density and of the shape of the distribution.

Skewness: Distribution are skewed to the side of the long tail. If no skewness is apparent, then the distribution is said to be symmetric.

Modality: Unimodal if there's one prominent peak (e.g. normal distribution), bimodal if there's two prominent peaks, uniform with no prominent peaks (no trend in the data), and multimodal if there's more than two prominent peaks.

The bin width can alter the story the histogram is telling.

A *box plot* shows the median (thick line inside the box) and the IQR (interquartile range (middle 50%): the width of the box).

How to measure the center?

1. Mean: arithmetic average.
2. Median: midpoint of the distribution (50th percentile).
3. Mode: most frequent observation.

If these measurements are calculated from a sample, they're called *sample statistic*, which are *point estimate* to the *population parameter*.

In a left skewed distribution, the mean is generally smaller than the median. In a right skewed distribution, the mean is generally larger than the median.

How to measure the spread?

1. Range: max − min.
2. Variance.
3. Standard deviation.
4. Interquartile range.

Variance: The average squared deviation from the mean. We use $s\hat{}2$ for the *sample variance* and $\sigma\hat{}2$ for the *population variance*.

Standard deviation: The average deviation around the mean, and has the same units as the data. We use s for the *sample sd* and σ for the *population sd*.

Interquartile range: Range of the middle 50 of the data. Is the distance between the first quartile (25th percentile) and the third quartile (75th percentile).

$$IQR = Q3 - Q1$$

Robust statistics: Are measures on which extreme observations have little effect.

Examples of robust statistics: median and IQR. Used on skewed distributions, with extreme observations.

Non-robust: mean, SD, range. Used to describe symmetric distributions.

Transforming data Transformation is a rescaling of the data using a function.

Used when data are strongly skewed, to make it easier to model.

Most used: natural log. Used when much of the data cluster near zero and all observations are positive. And to make the relationship between the variables more linear, hence easier to model with simple methods.

Other transformations: square roots, inverse.

Goals of transformations:

1. To see the data structure differently.
2. To reduce skew assist in modeling.
3. To straighten a nonlinear relationship in a scatterplot.

Categorical Data

Distribution of a single categorical variable Data can be summarized on a *frequency table* and on a *bar plot*.

Notice that a histogram is used for numerical variables, while bar plots are used for categorical variables.

Relationship between two categorical variables If we have two categorical variables, a *contingency table* can be used, which may suggest association between variables.

A *segment bar plot* are useful for visualizing conditional frequency distributions, i.e., the distributions of the levels of one variable (the response variable) conditioned on the levels of the other (the explanatory variable).

To compare relative frequencies, in order to explore the relationship between the variables, we can use a *relative frequency segmented bar plot* or a *mosaic plot*.

Relationship between a categorical and a numerical variable We can use *side-by-side box plots*: with the categorical variable on the x-axis and the numerical variable on the y-axis, for example.

Introduction to Inference

Gender discrimination in job promotion Null hypothesis: "there is nothing going on".

Alternative hypothesis: "there is something going on".

Hypothesis testing is like a court trial. We present the evidence (i.e. collect data) and judge the evidence: "could these data plausibly have happened by chance if the null hypothesis were true?"

If 'yes': fail to reject the null hypothesis. This doesn't mean that the alternative hypothesis is false. Only that there's not sufficient evidence that it is true. Similarly, it does not mean that the null hypothesis is true, because we can never prove that it is the case.

If 'no': reject the null hypothesis.

The burden of proof is on the unusual claim, the alternative hypothesis.

Hypothesis testing framework

1. Starts with a null hypothesis that represents the status quo.
2. Set an alternative hypothesis that represents the research question, i.e., what we're testing for.
3. Conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.
    1. If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis.

2. If they do, then reject the null hypothesis in favor of the alternative.

Simulation scheme

1. Face card: not promoted, non-face card: promoted.
2. Shuffle the card, deal into two groups of size 24, representing males and females.
3. Count how many number cards are in each group (representing promoted files).
4. Calculate the proportion of promoted files in each group, take the difference (male minus female), and record this value.
5. Repeat steps 2 - 4 many times.

Making a decision If the results from the simulations look like the data, then we decide that the difference between the proportions of promoted files between males and females was due to change (promotion and gender are independent).

Otherwise, then the observed difference was not due to chance, but due to an actual effect of gender (promotion and gender are dependent).

Summary

- Set a null and an alternative hypothesis.
- Simulate the experiment assuming that the null hypothesis is true.
- Evaluated the probability of observing an outcome at least as extreme as the one observed in the original data.
- And if this probability is low, reject the null hypothesis in favor of the alternative.

p-value: The probability of observing an outcome at least as extreme as the one observed in the original data under the assumption that the null hypothesis is true.

Probability

Law of large numbers: As more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome.

Gambler's fallacy (law of averages): even after ten consecutive heads, a coin is not due for a tail. The probability stays 1/2.

Bayesian Inference

Prior probability: what is believed before acquiring any data.

Posterior probability: is the probability of a hypothesis we set forth, given the data we just observed, i.e., $P(\text{hypothesis}|\text{data})$. Depends on the prior probability and the observed data.

We evaluate claims iteratively as we collect data. The next iteration takes advantage of what we learned from the data.

We update our prior with our posterior probability from the previous iteration.

Steps:

1. Setting a prior
2. Collecting data
3. Obtaining a posterior

4. Updating the prior with the previous posterior

Normal Distribution

68-95-99.7 Rule

- 68% falls within 1 s.d. from the mean.
- 95% falls within 2 s.d. from the mean.
- 99.7% falls within 3 s.d. from the mean.

Percentiles Z scores (i.e. the standardized normal distribution) can be used to compute percentiles.

$$Z = \frac{X - \mu}{\sigma}$$

Percentile: Is the percentage of observations that fall below a given data point.

Can be computed using R by

```
pnorm(-1, mean = 0, sd = 1)
```

If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution. This relationship appears as a straight line on a scatter plot.

Binomial Distribution

k successes in n independent Bernoulli trials with probability of success p.

$$\#\text{ of scenarios } \times P(\text{single scenario})$$

I.e.

$$\binom{n}{k} p^k (1-p)^{n-k}$$

```
choose(9,2)
dbinom(8, size = 10, p = 0.13)
```

Conditions

1. The trials must be independent
2. The number of trials n must be fixed
3. Each trial outcome must be classified as success or failure
4. The probability of success p must be the same for each trial

Mean and Variation

$$\mu = np \text{ and } \sigma\check{s} = np(1-p)$$

Normal Approximation to Binomial Use Z score, but adjust it by lowering by 0.5.

Success-failure rule: a binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution. I.e.

$$np \geq 10 n(1-p) \geq 10$$

Bayes Theorem

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

CLT and Sampling

Sampling variability

From a given population, we take a number of samples, which have a *sample distribution.*

Each sample gives a sample statistic, which have a *sampling distribution.*

Example Let N be the population size of US women and μ de average height of all women in the US. We may compute the average and standard deviation of this data.

We can collect the data from each state and compute the state mean $\bar{x}$. This collection of means is the sampling distribution.

$$\text{mean } (\bar{x}) \approx \mu$$
$$\text{SD}(\bar{x}) < \sigma$$

The standard deviation of the sample means is called the *standard error.* As n increases, the standard error decreases.

Central Limit Theorem

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(\text{mean} = \mu, \text{SE} = \frac{\sigma}{\sqrt{n}}\right)$$

If σ is unknown, we use S the sample standard deviation to estimate the standard error.

Conditions for the CLT

1. Independence: Sampled observations must be independent. Random sample/assignment. If sampling without replacement, $n < 10\%$ of population.
2. Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: $n > 30$).

Confidence Intervals

What is it?

Confidence Interval: A plausible range of values for the population parameter.

Based on the CLT.

A range of values is more likely to have a good shot at capturing the parameter.

Approximate 95% confidence interval: the sample mean plus or minus two standard errors (which is called the **margin of error (ME)**).

$$\bar{x} \pm 2 \text{ SE}$$

Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

Conditions

1. Independence: Sampled observations must be independent. Random sample/assignment. If sampling without replacement, $n < 10\%$ of population.
2. Sample size/skew: $n \geq 30$, larger if the population distribution is very skewed.

Finding and Interpreting What is the critical value for 95% confidence, the $z^\star$?

Using the normal table and the standard normal distribution, we have that

```
qnorm(0.025)
```

```
## [1] -1.959964
```

For the 98% confidence,

```
qnorm(0.01)
```

```
## [1] -2.326348
```

Accuracy vs. Precision

Confidence level: The percentage of random samples that will yield confidence intervals that contain the true population parameter.
Accuracy: Whether the confidence interval contains the true population parameter.
Precision: Width of a confidence interval.

As the confidence level increases, the width also increases (and encompasses more values). The accuracy also increases, but the precision goes down.

How to get higher precision and higher accuracy? Increase sample size.

Required Sample Size for Margin of Error Given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

$$\text{ME} = z^\star \frac{s}{\sqrt{n}} \iff n = \left(\frac{z^\star s}{\text{ME}}\right)^2$$

To cut the margin of error n points, we need n^2 times the sample size.

Inference and Significance

Hypothesis Testing

- Start with a null hypothesis H_0, representing the status quo.
- Also have a alternative hypothesis H_A, representing our research question, what we're testing for.
- Conduct a hypothesis test under the assumption that H_0 is true, either via simulation (as we did previously) or theoretical methods based on the CLT, that we're going to do now.
- If the test results suggest that the data do not provide convincing evidence for alternative hypothesis, we stick with the null hypothesis. Otherwise, the reject the null hypothesis in favor of the alternative.

Decision based on the p-value: - Use the test statistic to calculate the p-value, i.e., the probability of observing data at least as favorable to the alternative hypothesis as our current data set, assuming the null hypothesis is true. - If the p-value is lower than the significance level α (usually 5%), we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0. - If the p-value is greater than α, we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

Two-sided tests: - Instead of looking for the divergence from the null hypothesis in a specific direction, we might be interest in divergence in any direction. - Also called two-tailed tests. - The definition of the p-value is the same, but the calculation is slightly different. E.g. $P[\overline{X} > 3.2 \text{ or } \overline{X} < 2.8 \mid H_0 : \mu = 3]$.

Hypothesis testing for a single mean: 1. Set the hypothesis: 1. $H_0 : \mu =$ null value; 2. $H_A : \mu <, >, \neq$ null value. 2. Calculate the point estimate: $\overline{x}$. 3. Check conditions: 1. Independence (random sample/assignment and if sample without replacement, $n < 10\%$ of population); 2. Sample size/skew: $n \geq 30$, larger if the population distribution is very skewed. 4. Draw sampling distribution, shade p-value, calculate test statistic. Here it is the Z score. 5. Make a decision, and interpret it in context of the research question: 1. If p-value $< \alpha$, reject H_0; 2. If p-value $> \alpha$, fail to reject H_0.

Significance

Our previous method works for nearly normal sampling distributions.

Unbiased estimator: The sampling distribution of the estimate is centered at the true population parameter it estimates. It provides a good estimate.

Sample mean, difference between sample means are examples of unbiased point estimate.

Confidence intervals: point estimate $\pm z^* \times SE$.

Type I error: rejecting H_0 when you shouldn't have.

Type I error rate:

$$P[\text{Type I error} \mid H_0 \text{ true}] = \alpha$$

Type II error: failing to reject H_0 when you should have. The probability of doing so is β.

The **power** of a test is the probability of correctly rejecting H_0. Probability: $1 - \beta$.

Goal: keep α and β low.

The value β depends on the **effect size** δ, which is the difference between point estimate and null value.

A two sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1\alpha$. If the hypothesis is one sided, then $CL = 1 - (2\alpha)$.

If H_0 is reject, a confidence interval that agrees with the result of the hypothesis test should not include the null value. Reciprocally, if H_0 is failed to be rejected, the confidence interval should include it.

Statistical significance: when the p-value is lesser than the significance level α.

Very large samples will result in statistical significance even for tiny values of the effect size, even when it is not practically significant.

Start by figuring how many observations to sample beforehand.

```
xbar <- 30.69
sd <- 4.31
n <- 36
se <- sd / sqrt(n)
z <- (xbar - 32) / se
z
```

```
## [1] -1.823666
```

```
2*pnorm(-abs(z))
```

```
## [1] 0.0682026
```

Inference for Comparing Means

T-Distribution and Comparing Two Means

T-Distribution T-distribution is useful for describing the distribution of the sample mean when the population standard deviation is unknown.

Bell shaped, but with thicker tails than the normal.

Observations are more likely to fall beyond two SDs from the mean.

This is helpful for mitigating the effect of a less reliable estimate.

Is more "conservative" than the normal.

Always centered at zero and has one parameter: the degrees of freedom, which determines the thickness of tails.

As the degrees of freedom increases, the t-distribution approaches the normal distribution.

How to use it?

1. For inference on a mean where the SD is unknown.
2. Calculated the same way:
$$T = \frac{\text{obs} - \text{null}}{\text{SE}}$$
3. Compute the p-value as before, but using the t-distribution.

```
pnorm(2, lower.tail = FALSE) * 2
```

```
## [1] 0.04550026
```

```
pt(2, df = 50, lower.tail = FALSE) * 2
```

```
## [1] 0.05094707
```

```
pt(2, df = 10, lower.tail = FALSE) * 2
```

```
## [1] 0.07338803
```

Inference for a Mean We use a point estimate $\pm$ margin of error:

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

The degrees of freedom for t statistic for inference on one sample mean is $df = n - 1$. Thus,

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

How to find the critical t score?

1. Use the table
    1. Determine the degrees of freedom
    2. Find the corresponding tail area for desired confidence level
2. Use R

    ```
    qt(0.025, df = 21)
    ```

    ```
    ## [1] -2.079614
    ```

where the 0.025 comes from 95% confidence level: (1-0.95)/2.

How to find the p-value?

1. Use the table
    1. Determine the degrees of freedom
    2. Locate the calculated T score in the df row
    3. Grab the one or two tail p-value from the top row
2. Use R

```
2 * pt(2.3, df = 21, lower.tail = FALSE)
```

```
## [1] 0.03180228
```

Inference for comparing two independent means We also use a point estimate ± margin of error:

$$(\bar{x_1} - \bar{x_2}) \pm t^\star_{df} \text{SE}_{(\bar{x_1} - \bar{x_2})}$$

Standard error of difference between two independent means:

$$\text{SE}_{(\bar{x_1} - \bar{x_2})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom for t statistic for inference on difference of two means:

$$df = \min(n_1 - 1, n_2 - 1)$$

Conditions for inference for comparing two independent means:

1. Independence:
    - Withing groups: sampled observations must be independent.
        - Random sample/assignment
        - If sampling without replacement, $n < 10\%$ of population
    - Between groups: the two groups must be independent of each other (non-paired)
2. Sample size/skew: the more skew in the population distributions, the higher the sample size needed.

Inference for comparing two paired means To analyze paired data, it is useful to look at the difference in outcomes of each pair of observations.

Doing this, we go back to the case of a single population mean.

Power

Power: Is the probability of correctly rejecting H_0, i.e., $1 - \beta$.

ANOVA and Bootstrapping

Comparing more than two means Use a new test called **analysis of variance (ANOVA)**.

And a new statistic called F.

H_0: the mean outcome is the same across all categories.

H_A: at least one pair of means are different from each other.

The **F distribution** is:

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

- Right skewed - Always positive

ANOVA Variability partitioning: breaks the data into the variability we're interested in and all other variability.

Sum of squares total (SST) measures the total variability in the response variable.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where y_i is the value of the response variable for each observation and $\bar{y}$ is the mean of the response variable.

Sum of squares groups (SSG) measures the variability between groups. Is the variability in the response variable explained by the explanatory variable.

$$SSG = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

where n_j is the number of observations in the group j, $\bar{y}_j$ is the mean of the response variable for the group j, and $\bar{y}$ is the mean of the response variable as before.

Sum of squares error (SSE) measures the variability within groups. Is the variability due to all other variables.

Degrees of freedom associated with ANOVA: - Total: $df_T = n-1$ - Group: $df_G = k-1$ - Error: $df_E = df_T - df_G$

Mean squares: - Group: $MSG = SSG/df_G$ - Error: $MSE = SSE/df_E$

The **F statistic** is the ratio of the average between group and within group variabilities:

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

```
pf(21.735, 3, 791, lower.tail = FALSE)
```

```
## [1] 1.559855e-13
```

Conditions for ANOVA: 1. Independence: - Within groups: sampled observations must be independent - Random sample/assignment - Each n_j less than 10% of the respective population
- Between groups: the groups must be independent of each other (non-paired) 2. Approximate normality: distributions should be nearly normal within each group - Distribution of the response variable within each group should be approximately normal - Especially important when sample sizes are small 3. Equal variance: groups should have roughly equal variability - *Homoscedastic* groups: variability should be consistent across groups - Especially important when sample sizes differ between groups

Multiple comparisons Which means are different?

Testing many pairs of groups is called **multiple comparisons**.

The **Bonferroni correction** suggests that a more stringent significant level is more appropriate for these tests. 1. Find the number of comparisons

$$K = \frac{k(k-1)}{2}$$

2. Adjust the α by the number of comparisons:

$$\alpha^\star = \alpha/K$$

Standard error for multiple pairwise comparisons:

$$\text{SE} = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Degrees of freedom for multiple pairwise comparisons:

$$df = df_E$$

Then we're back to the previous case.

```
2 * pt(5.365, df = 791, lower.tail = FALSE)
```

```
## [1] 1.063895e-07
```

Bootstrapping Steps:

1. Take a bootstrap sample: a random sample taken with replacement from the original sample and of the same size as the original sample.
2. Calculate the boostrap statistic (mean, median, proportion etc.) computed on the boostrap samples.
3. Repeat steps 1 and 2 many times to create a bootstrap distribution.

Then use $t^\star \pm \text{SE}_{\text{boot}}$ (standard error method) to compute a confidence interval.

Inference for Proportions

Theorem (CLT for proportions): The sampling distribution for the sample proportion $\hat{p}$ based on a sample size of size n from a population with a true proportion p is nearly normal when 1. **Independence:** The sample's observations are independent. - Random sample/assignment - If sampling without replacement, $n < 10\%$ of population 2. **Sample size/skew (success-failure condition):** we expect at least 10 successes and 10 failures in the sample, i.e., $np \geq 10$ and $n(1-p) \geq 10$. Under these conditions,

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

If p is unknown, use $\hat{p}$.

We can approximate using the binomial distribution. For example,

```
sum(dbinom(190:200, 200, 0.90))
```

```
## [1] 0.00807125
```

Confidence Interval for a Proportion Point estimate $\pm$ margin of error:

$$\hat{p} \pm z^\star \text{SE}_{\hat{p}}$$

in which

$$\text{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Remember the margin of error

$$\text{ME} = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we don't have a reliable value for $\hat{p}$, use $\hat{p} = 0.5$.

Hypothesis Test for a Proportion

1. Set the hypotheses
2. Calculate the point estimate $\hat{p}$
3. Check conditions:
4. Independence
5. Sample size/skew
6. Draw the sampling distribution, shade p-value, calculate test statistic
7. Make a decision, and interpret it in the context of the research question

We must use the null hypothesis value for p.

Estimating the Difference between two Proportions Point estimate ± margin of error:

$$(\hat{p}_1 - \hat{p}_2) \pm z^\star \text{SE}_{(\hat{p}_1 - \hat{p}_2)}$$

in which

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions:

1. Independence:
   - Withing groups: sampled observations must be independent.
     - Random sample/assignment
     - If sampling without replacement, $n < 10\%$ of population
   - Between groups: the two groups must be independent of each other (non-paired)
2. Sample size/skew: each sample should meet the success-failure condition.

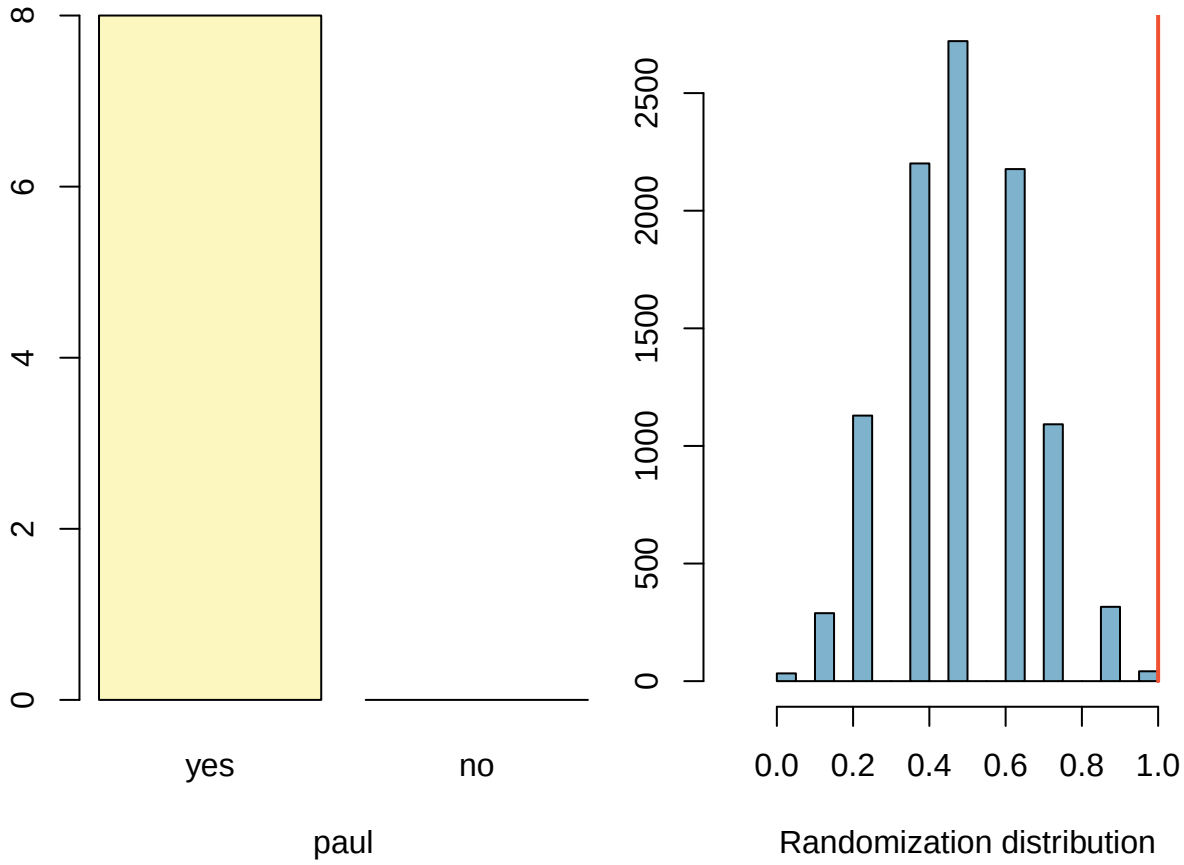
Simulation-based Inference for Proportions and Chi-squared testing

What if the success-failure condition is not met (small sample proportion)?

```
source("http://bit.ly/dasi_inference")
paul = factor(c(rep("yes", 8), rep("no", 0)), levels = c("yes", "no"))
inference(paul, est = "proportion",, type = "ht", method = "simulation", success = "yes", null = 0.5, a
```

```
## Single proportion -- success: yes
## Summary statistics:
```

```
## p_hat = 1 ;  n = 8
## H0: p = 0.5
## HA: p > 0.5
```




```
## p-value =  0.0042
```

Chi-Square Goodness of Fit Test We start with a table comparing the observed and expected values.

Chi-square statistic χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

where O is the observed, E is the expected and k is the number of cells.

Degrees of freedom: $k-1$.

The p-value is the tail area above the calculated test statistic.

```
pchisq(22.63, 4, lower.tail = FALSE)
```

```
## [1] 0.000150104
```

Chi-Square Independence Test Here, compute the degrees of freedom as

$$df = (R-1)(C-1)$$

where R is the number of rows and C is the number of columns.

Linear Regression

Relationship between two numerical variables

Correlation describes the strength of the *linear association* between two variables. Notation: R.

Residual is the difference between the observed and the predicted, i.e.,

$$e_i = y_i - \hat{y}_i$$

Least Squares Line:

$$\hat{y} = \beta_0 + \beta_1 x$$

To estimate the slope, we use

$$b_1 = \frac{s_y}{s_x} R$$

where s_x (resp. s_y) is the s.d. of x (y).

And for the intercept,

$$b_0 = \overline{y} - b_1 \overline{x}$$

where the bar indicates the average value.

Linear regression with one predictor

Conditions for linear regression:

1. Linearity
2. Nearly normal residuals
3. Constant variability (homoscedasticity)

How well the model fits the data? Use the **R Squared**. It is the percentage of variability in the response variable explained by the model.

Outliers & Inference for Regression

Leverage points: outliers that fall horizontally away from the center of the cloud but don't influence the slope of the regression line.

Influential points: influence the slope of the regression line.

T score for the hypothesis test

$$T_{df} = \frac{b_1 - \text{null value}}{\text{SE}_{b_1}}$$

with $df = n - 2$.

Confidence interval for the slope

$$b_1 \pm t^\star_{df} \text{SE}_{b_1}$$

Example of how to compute $t^\star_{df}$ for 95% confidence and 25 degrees of freedom:

```
qt(0.025, df = 25)
```

```
## [1] -2.059539
```