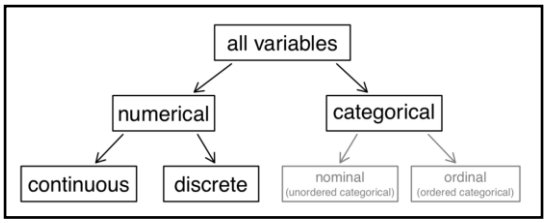
Statistics

1: Introduction to Data

1.1 – Data Basics

Types of Variables



- Numerical: can take a range of numerical values
  > continuous: can take every value
  > discrete: can take values with jumps (e.g. integers)
- Categorical: can take different categories of values
  > nominal: cannot be ordered
  > ordinal: can be ordered

Association and Independence

- Variables are *associated/dependent* when the show some connection
  > positive: higher ~ higer, negative: higher ~ lower
- Variables are *independent* when they are not *associated*

1.2 – Data Collection Principles

Population & Sampling

- The term *population* refers to a population as a whole - statistics on populations are rare due to the high costs of producing these
- A *Sample* is a subset of the population (often a [very] small fraction)
  > *random samples* abolish bias in a sample (be aware of non-response bias)

Explanatory and Response Variable

- *Explanatory* variables somehow affects the *response* variable
- Caution: Association does not imply causation

Observational Studies and Experiments

- Researchers perform *oberservational studies* when they collect data in a way that does not directly interefe with how the data arise
- *Experiments* allow researchers to investigate causal conncetions by selecting samples and randomly assign them to *treatment* and *control-groups*

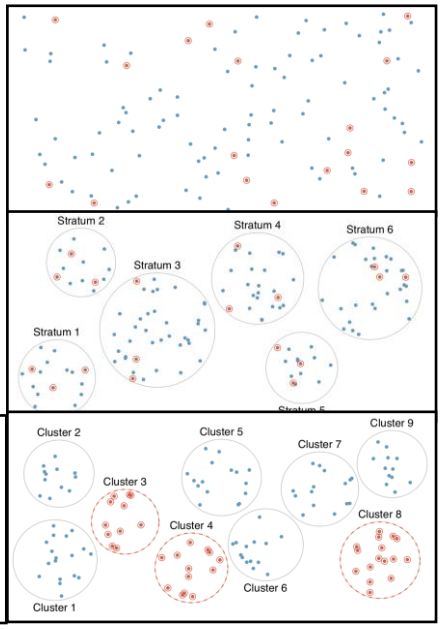
1.3 – Sampling Strategies & Observational Studies

Observational Studies

- *Confounding variables* are correlated to explanatory and response variables and thereby allow to make causal conclusions
- *Prospective* studies identify individuals and collects information on them
- *Retrospective* studies collect data after events have taken place

Sampling Methods

- Simple Random Sampling: randomly picking from the whole population
- Stratified Sampling: divide population into groups called *strata* (grouping similar cases) and randomly pick from these stratas according to their fraction
- Cluster Sampling: break up the population in many clusters and randomly pick clusters (where we use all observations)
- Multistage Sample: break up population in many clusters and randomly pick clusters in which we randomy pick observations



1.4 – Experiments

Randomized Experiments

- Controlling: controlling any other differences in treatment & control group and minimize them
- Randomization: people are randomly assigned to treatment & control group
- Replication: replicate results by choosing large sample sizes ore by replicating entire studies
- Blocking: grouping individuals before assigning them to control & treatment group to eliminate further deviations/differences in samples

Blind & Double-Blind

- Blind: patients do not know, whether they are in treatment or control group
- Double-blind: patients & doctors do not know, if patients are in treatment or control group

1.5 – Numerical Data

Mean

$$\overline{x} = \frac{\sum x}{n}, \mu for\ population\ mean$$

Mean in R

```
1. mean(data)
```

Histogram & Shape

By creating an histogram from the data-set we can examine certain properties

- Skewness: longer right tail > *right skewed*, longer left tail > *left skewed*, equal tails > *symmetric*
- Peaks: one peak > *unimodal*, two peaks > *bimodal*, multiple peaks > *multimodal*

Variance & Standard Deviation

- s: sample standard deviation, s^2: sample variance, σ: population standard deviation, σ^2: population variance

$$s^2 = \frac{(x_i - \bar{x})^2}{n-1}$$

Variance

```
1. var(data$col)
```

Standard Deviation in R

```
1. sd(data$col)
```

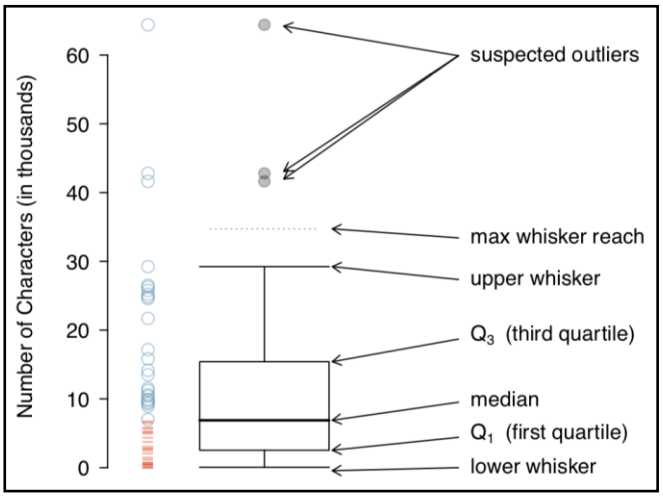
Using R to summarise Data
`%>%` is the pipelineoperator, used to concat functions on the same dataset

Useful functions:
`filter()` used to filter objects
`summarise()` used to sum up data by certain parameters like mean, min, max, sd, iqr, n, n_distinct
`select()` only keeps data that is selected
`group_by()` groups data by a certain value
`arrange()` used to order data
`mutate()` adds new variables, preserves existing ones drops variables

```
1. data %>%
2.   filter(filteringcondition) %>%
3.   summarise(mean = mean(n),
4.   median = median(n),
5.   min = min(n),
6.   max = max(n),
7.   n = n(),
8.   n_distinct = n_distinct(n)
9.   )
```

Box Plots, Quartiles & Median

- Box plot can be seen on the right
- Median: 50% of data fall below median, 50% of data fall above median (center)
- Quartile: each quartile contains 25% of the observations/data
  > $IQR = Q_3 - Q_1$
- Whiskers: reach is never more than $1{,}5 \cdot IQR$



Robust Statistics
Median & IQR are called robust estimates, as outliers have only little effect on them

For GG-Plot commands use the ggplot2 cheatsheet :)

1.6 – Categorical Data

Contigency Table incl. Totals

- Table for single variables is called *frequency table*
- Replacing counts with percentages/proportions would result in a *relative frequency table*

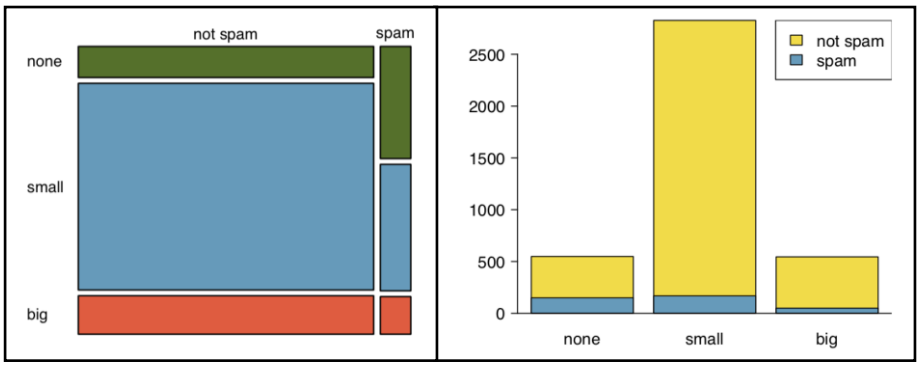
		number			
		none	small	big	Total
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

Contingency table in R

```
1. dt <- table(data$var1, data$var2)
2. addmargins(dt)
```

Segmented Bar and Mosaic Plots

- Segmented bar plot: absolute numbers
- Mosaic plot: proportion/probabilities



Indepence
Calculating row probabilities and comparing them allows the evaluation of ‚independence'.

1.7. Calculating a quantile

Functions for quantiles always start with a 'q'. The Quantile is a percentile in a dataset.

Args: The column with the data, the quantile e.g. `.95`, optional: `na.rm = TRUE` to remove na values

```
1. quantile(data$col, .quantile)
```

2: Probability

2.1 – Defining Probability

Probability
The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Disjoint and Mutually Exclusive Outcomes
Two outcomes are disjoints/mutually exclusive, if they cannot both happen.

$P(A_1 \vee A_2) = P(A_1) + P(A_2)$ Addition Rule of disjoint outcomes

General Addition Rule

$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Probability Distributions
A probability distribution is a list of the possible outcomes with corresponding probabilities that

1. must be disjoint
2. are between 0 and 1
3. sum up to a total of 1

Complement
The Complement of an outcome represents all outcomes not in the original: $P(A) + P(A^C) = 1$

Independence
Two processes are independent, if knowing the outcome of one provide no useful information on the outcome of the other.

$P(A \wedge B) = P(A) \cdot P(B) = P(A|B)$ Multiplication Rule for independent processes

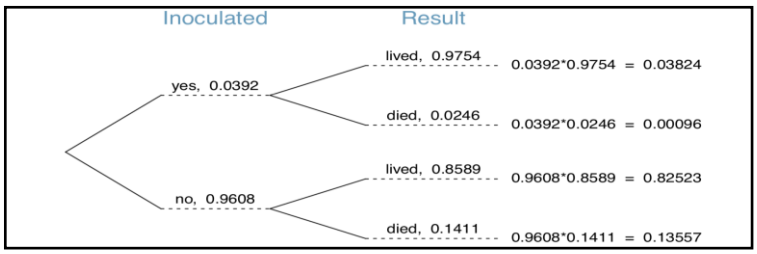
2.2 – Conditional Probability

Marginal and Joint & Conditional Probabilities

- Probabilities based on a single variable are called *marginal* probabilities
- Probabilities based on two or more variables are called *joint* probabilities
- *Conditional* probabilities is used for computing probabilities under given conditions

$P(A|B) = \frac{P(A \wedge B)}{P(B)}$ Probabilitiy for A given B

Tree Diagrams



Bayes' Theorem: inverting probabilities

$P(A_1|B) \frac{P(B|A)P(A_1)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(B|A_k)P(A_k)}$

2.3 – Sampling from small Population

Sampling with or without Replacement

- Sampling *without replacement* abolishes independence between observations
- By sampling *with replacement* the probability stays the same

2.4 – Random Variables

Random Variable
A random process or variable with a numerical outcome.

Expectation

$$E(X) = \mu = \mathbb{E} = \sum_{i=1}^{k} x_1 \cdot P(X = x_i)$$

Variability

$$\sigma^2 = \sum_{j=1}^{k} (x_j - \mu)^2 \cdot P(X = x_j)$$

2.5 – Sensitivity and Specifity

Sensitivity and Specifity

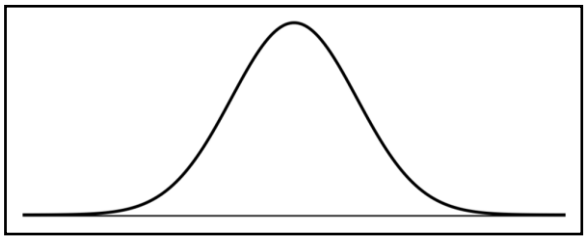
- Sesitivity measures a tests ability to identify positive results
- Specifity measure a tests ability to identify negative results

Calulation from binary variables

$$sensitivity = \frac{TP}{TP+FN} \qquad specificity = \frac{TN}{FP+TN}$$

3: Distribution of Random Variables

3.1 – Normal Distribution



Normal Distribution Model
Many variables are nearly normal distributed.
Therefore, their distribution can me modelled with the normal distribution model.
The model can be modified with two parameters:

- σ (standard deviation)
- μ (mean)

Standardizing with Z-Scores
Z-Scores are used to standardize deviations from the mean under the normal distribution model. The z-score is the number of standard deviation the observation is above/below the mean.

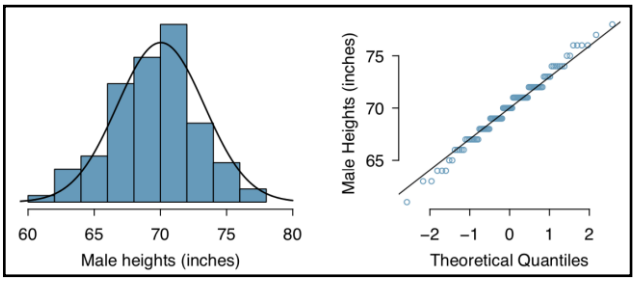
$$Z = \frac{x-\mu}{\sigma}$$ z-score

68-95-99.7 Rule

- The interval $\mu \pm 1 \cdot \sigma$ covers 68% of observations
- The interval $\mu \pm 2 \cdot \sigma$ covers 95% of observations
- The interval $\mu \pm 3 \cdot \sigma$ covers 99.7% of observations

3.2 – Evaluating Normal Distribution

Evaluation with Histogram
Create simple histogram and overlay best fitting normal curve using sample mean $\overline{x}$ and standad deviation s as parameters of the curve.



Evaluation with Normal Probability Plot
Create a normal probability plot - the closer the points are to a straight line, the more confident we can be about the normality assumption.

Calculating a normal distribution in R:

```
1. dnorm(x, sd = 1, log = FALSE)
```

Calculating a probability in R of a normal Distribution
Args: Zscore: double, optional: lower.tail: boolean

```
1.  pnorm(zscore)
```

Calculating a quantile in R of a normal Distribution
Args: p: vector of probabilitys, mean, sd, optional: lower.tail: boolean

```
1. qnorm(p, mean = mean, sd = sd)
```

3.3 – Geometric Distribution

Bernoulli Distribution
Bernoulli Distribution fulfill the following assumptions

- Each *independent* person in the experiment is considered a *trial*
- Each trial has an equal probability of *success* p and an equal probability of *failure* $q = 1 - p$
  > when individual trial has just two possible outcomes, the variable is called *Bernoulli Variable*

$\mu = p$ $\sigma = pq$ $\sigma = \sqrt{pq}$

Geometric Distribution
Geometric distributions are used to calculate the waiting time until a success for

- Independent
- Indentically distributed

Bernoulli random variables. The probability of finding the first success in the n[th] trial is $q^{n-1} \cdot p$

$\mu = \frac{1}{p}$ $\sigma^2 = \frac{1-p}{p^2}$ $\sigma = \sqrt{\frac{1-p}{p^2}}$

Calculating a geometric distribution in R:

```
1. dgeom(x, prob)
```

Calculating a probability in R of a normal Distribution
Args: q: vector of quantiles, probability of success, optional: lower.tail: Boolean

```
1. pgeom(q, prob)
```

Calculating a quantile in R of a normal Distribution
Args: p: vector of probabilitys, prob: probability of success, optional: lower.tail: Boolean

```
1. qgeom(p, prob)
```

3.4 – Binomial Distribution

Binomial Distribution

- The binomial distribution is used to calculate the probability of having k successes in n trials.
- The binomial distribution can be approximated using a normal model when failure and success occur at least 10 times.
  > accuracy can be improved by widening interval 0.5 on both sides
- Binomial distributions must fulfill conditions
  > trials are independet
  > number of trials n is fixed
  > each trial can be classified as either success or failure
  > the probabilities for success and failure are constant for each trial

$\binom{n}{k} p^k \cdot q^{n-k}$

$\mu = np$ $\sigma^2 = npq$ $\sigma = \sqrt{npq}$

Calculating a geometric distribution in R:

```
1. dbinom(x, size = size, prob = prob)
```

Calculating a probability in R of a binomial Distribution
Args: x: vector of quantiles, size: number of trials, prob: prohability of success, log: logical, if given as log(p)

```
1.  pbinom(x, size = size, prob = prob)
```

Calculating a quantile in R of a normal Distribution
Args: p: vector of prohabilitys, size: number of trials, prob: prohability of success,

```
1. qbinom(p, size, prob = prob)
```

3.5 – More discrete Distrubutions

Negative Binomial Distribution

- The negative binomial distribution is used to calculate the probability of observing the k^{th} success in the n^{th} trial.
- Negative binomial distribution must fulfill conditions
  > trials are independent
  > each trial can be classified as either success or failure
  > the probabilities for success and failure are constant for each trial
  > the last trial must be a success

  $\binom{n-1}{k-1} p^k \cdot q^{n-k}$

Poisson distribution

- The poisson distribution is used to estimate number of events in a large population over a unit of time (e.g. having heart attack, getting married, getting struck by lightning).
- Individuals in the population are independent

  $P(\textit{observe k events}) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$, where λ is the *rate* of occurences over a fixed span of time

  $\mu = \lambda \qquad \sigma^2 = \lambda \qquad \sigma = \sqrt{\lambda}$

4: Foundations for Inference

4.1 – Variability in Estimates

Point Estimates

- We want to estimate the population mean from our sample, but the *sample mean* is to variable.
- The sample mean is called a *point estimate* and it varies with *sampling variation*.

Standard Error of the Mean
The variability in point estimates can be described using the *standard errror* (standard deviation associated with an estimate).

$SE_{\overline{x}} = \sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$, where observations are independent (less than 10% of population)

4.2 – Confidence Intervals

Confidence Intervals

- A plausible range of values for the population parameter is *called confidence* interval (CI).
- „We are xx% sure, that the population parameter is within the CI."
  $\overline{x} \pm z^\star \cdot SE$, where $z = 1.64 \rightarrow 90\%, z = 1.96 \rightarrow 95\%, z = 2.58 \rightarrow 99\%$
  $z^\star \cdot SE$ margin of error

Central Limit Theorem
If a sample consits of at least 30 independent observations and data is not strongly skewed, than the distribution of the sample mean is well approximated by the normal model.

Conditions for use of CI
Conditions to ensure, that $\overline{x}$ is nearly normal and the estimated SE is accurate:

- Sample observations are independent
  > best judgement, random assignment, random sample is less than 10% of the population
- Sample size is large $n \geq 30$
- The population distribution is not strongly skewed
  > best judgement, less important in larger samples (outliers are accepted for $n \geq 100$)

4.3 – Hypothesis Testing

Framework

- Set hypotheses: null hypothesis (H_0) representing a skeptical perspective/claim, alternative hypothesis (H_A) representing an alternative claim (often range of possible parameter values)
- H_0 is only rejected if we can find convincing evidence, that it is false
  > possible evidence is CI or p-value (confidence level α)

Decision Errors

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Requirements for Tests

- Individual observations must be independent
- Sample size must not be too small and too skewed

Statistical vs. Practical Significance
Large sample sizes result in smaller SE and therefore a more sensible test. Therefore, we might detect small differences, which, while being statistically significant, are not practically significant.

5: Inference for Numerical Data

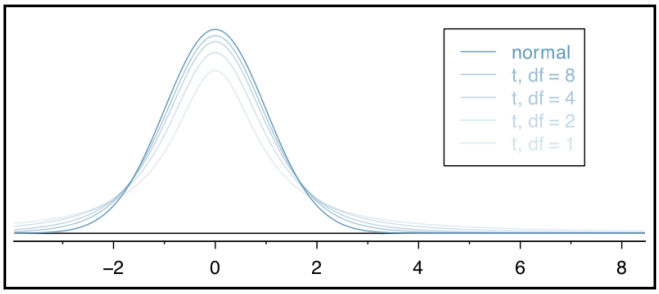
5.1 – One-Sample Means with t-Distribution

Normality Condition

- We required large sample sizes to ensure a normal distribution of sample means & to ensure the accuracy of the calculated standard error.
- According to the Central Limit Theorem the sampling distribution is nearly normal when sample observations are independent and come from a nearly normal distribution.

t-Distribution

- Tails are thicker and the peak is lower in the t-distribution
- We use a t-score (comp. z-score)
- The t-distribution has a single parameter *df* (*degrees of freedeom*)
  > $df \geq 30$ is nearly normal
  > $df = n - 1$



Conditions for using the t-Distribution

- Independence of observations (random sample is less than 10% of the population)
- Observations from a nearly normal distribution
  > look at the data
  > previous experiments alerting?

Confidence Interval

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

5.2 – Paired Data

Paired Data

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

Inference for Paired Data

- Add a „diff“ variable to the dataset
- Conduct a hypotheses test using the t-distribution

5.3 – Difference of two Means

Requirements

- Each sample meets the requirements for the t-distribution (independence, normal distributed)
- Samples are independent

Distribution of difference of Sample Means

- The difference of two means can be modelled using the t-distribution

$$SE_{x_1 - x_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = min(n_1 - 1, n_2 - 1)$$

Inference

- Write appropriate hypotheses.
- Verify conditions for using the t-distribution
  > one-sample or difference in paired data: observations must be independent and nearly normal, slight skew is okay for larger sample sizes
  > difference in means: each sample must satisfy t-distribution requirements & independence
- Compute point estimate of interest, the standard error & the degrees of freedom
- Compute T-score & p-value
- Make conclusion based on the p-value and write conclusion in plain language

5.4 – Power calculations for Difference of Means

Creating powerful Tests
Planning tests leaves us with two competing considerations

- Collect enough data to detect differences
- Collect little amount of data to save money & protect patients

We aim for a power of 80%.

Determining a propper Sample Size

- The expected mean in case of success must not be in the CI (incl. some variation space)

$$0.84 \cdot SE + 1.96 \cdot SE, 3 = 2.8 \cdot SE, \text{where } SE = \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

5.5– Doing a T-Test in R

We can use t.test for doing a T-Test
If we already have substracted the 2 variables of interest and set our substract as x.
Also we define our conf.level. mu is the true value of the mean.

```
1. t.test(x = data$diff, conf.level = , mu = )
```

elsewise we define x and y and set `paired = TRUE` in a two-sided test mu is defined as the difference in means in a two sided test.

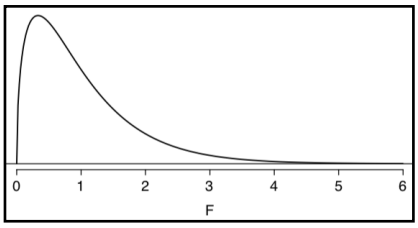
```
1. #else
2. tstat <- t.test(x = data$a, y = data$b, paired = TRUE, conf.level = , mu
   =)
```

We can specify the test by adding `alternative = c("two-sided", "less", "greater")` to specify our test. The default is "two-sided".

5.6 – Comparing many Means with ANOVA

Comparing Means of different Samples

- Pairwise comparisons are time consuming - use analysis of variance (ANOVA)
- f-statistic
- Hyptheses: H_0: mean is the same accross all groups, H_A: at least one mean is different



Requirements/Conditions

- Observations are independent within and across groups
- Data within each group are nearly normal
- Variability across groups is about equal

The F-Test

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 \quad MSE = \frac{1}{df_E} SSE = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)s_i^2 \quad F = \frac{MSG}{MSE}$$

Uppter tail of the F-Value represents the p-value.

Multiple Comparison

- We do multiple comparisons to find out, which mean differentiates.
- Use Bonferro correction to prevent inflation of type 1 error
  > $\alpha^* = \frac{\alpha}{K}$, where $K = \frac{k(k-1)}{2}$
- Caution: sometimes ANOVA will reject H_0 but no comparison shows stat. signficant differences.

Doing an anovatest in R

```
1. aov <- aov(response ~ explanatory, data = data)
2. summary(aov)
3. #Tukeys honestly significant difference test
4. TukeyHSD(aov)
```

6: Inference for Categorical Data

6.1 – Inference for Single Proportions

Sample Proportion

$$p = \frac{\sum x}{n}$$

Requirements/Conditions

- Sample observations are independent
- Success-failure condition: $np \geq 10, nq \geq 10$

Hypothesis Testing

- Set up hyptothesis: H_0: $p = p_o$, H_A: $p \neq p_0$
- Calulcate SE, Z-Score and p-value/CI
- Evalute hypothesis

$$SE = \sqrt{\frac{p_0 q_o}{n}} \qquad Z = \frac{p - p_0}{SE}$$

Choosing the right Sample Size
If we want to achieve a given margin of error, where we will reject H_0, we can calculate the required sample size.

$z^\star \sqrt{\frac{pq}{n}}$, choose $p = 0.5$ if it is unknown

6.2 – Difference of two Proportions

Difference of two Porportions

Difference $p_1 - p_2$ tends to follow a normal model when

- Each proportion itself follows the normal model
- The two samples are independent of each other

$$SE_{p_1 - p_2} = \sqrt{SE_{p_1}^2 + SE_{p_2}^2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Pooled Proportion

When H_0 is that proportions are equal, use the pooled proportion (p) to verify success-failure condition and estimate the standard error.

$$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} \quad SE = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

6.3 – Testing Goodness of Fit using Chi-Square

Chi-Square
Observed = what we observed, Expected = what we expected (using our expected distribution)

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i} \qquad df = k - 1$$

Requirements/Conditions

- Independence

- Sample Size: each particular scenario must have at least 5 expected cases

When to use

- Sample of cases that can be classified into several groups: determine whether representation is representative to general population
- Evaluate whether data resemble a particular distribution (e.g. normal/geometric distribution)

Computing a chisq test:

```
1. # Use chisq.test on a table
2. chisq.test(table(data))
```

Outputs: X-squared, df, p-value

6.4 – Testing for Independence in two-way Tables

Expected Counts in two-way Tables

$$Expected\ Count_{rowi,colj} = \frac{rowi\ total \cdot colj\ total}{table\ total} \qquad df = (R-1)\cdot(C-1)$$

6.5 – Small Sample Hypothesis testing for a Proportion

When Sucess-Failure Condition is not met
Generate the distribution by simulation.

7: Introduction to linear Regression

7.1 – Residuals and Correlation

Residuals
Residuals are the leftover variation in the data after accounting for the model fit

$$e_i = y_i - y_i$$

In R ist easy to plot the residuals using autoplot.

```
1. #residual analysis
2. library(ggfortify)
3. autoplot(model)
```

Correlation
Correlations describe the strength of linear relationship, taking values between -1 and 1.

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\overline{x}}{s_x}\frac{y_i-\overline{y}}{s_y}$$

```
1. r <- cor(data$Value1, data$Value2)
2. r
```

7.2 – Line fitting by least Squares Regression

Requirements/Conditions

- Linearity: data should show a linear trend
- Nearly Normal Residuals: residuals must be nearly (or large sample)
- Constant Variability: variability of points around fitted line remains roughly constant
- Independence: independent observations, caution to time series data

When using R we need a summary of our data and especially we need the mean and the standard derivation of our response and our explanatory variable. The results of the following code will be used later.

```
1. d <- data %>%
2. summarise(mean_resp = mean(Response),
3.           mean_exp = mean(Explanatory),
4.           sd_resp = sd(Response),
5.           sd_exp = sd(Explanatory))
```

Finding the Line

$b_1 = \frac{s_y}{s_x} R$ (slope)

```
1. # estimated intercept using the data summaray of above.
2. slope <- d$sd_resp/d$sd_exp*r
```

$b_0 = y - b_1 \cdot x$ (intercept)

```
1. # estimated intercept using the data summaray of above.
2. intercept <- d$mean_resp - slope*d$mean_exp
```

Creating a model

```
1. #modeling the linear regression
2. model <- lm(Response ~ Explanatory, data = data)
```

Extrapolation
Linear models describes the data over a given interval. Model should not be applied outside!

Strength of Fit
Strength of a fit is described using R^2, which is the variability in the data desribed by the model.

$R^2 = correlation^2$

```
1. #strength of the fit
2. summary(model)$r.squared
3. cor(Resonse, Explanatory)^2
```

7.3 – Types of Outliers in Linear Regression

Leverage
Points that fall horizintally away from the center of the cloud are called points with high leverage.

Influential Points
Points with high leverage actually changing the line substantially are called influential points.

7.4 – Inference for linear Regression

Inference
We usually test H_0: $b_1 = 0$, H_A: $b_1 \neq 0$ using a t-Test.

In R we can use lm to create our model to test our data as described above.

```
1. model <- lm(response ~ explanatory, data = )
2. broom::tidy(model)
3. #test statistic
4. t <- (estimate - 0)/(std. error)
5. #p-value
6. p <- 2 * pt(t, df = , lower.tail = FALSE)
7. #confidence interval
8. confint(model, "explanatory", level = 0.95)
```

8: Multiple and logistic Regression

8.1 – Multiple Regression

Mutiple Regression Model
A multiple regression model is a linear model with many predictors. In general, we write the model as $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ when there are k predictors.
The β's are estimated using statistical software.

Adjusted R^2

$R^2_{adj} = 1 - \frac{Var(e_i)}{Var(y_i)} \cdot \frac{n-1}{n-k-1}$, where n is number of cases and k is number of predictors

8.2 – Model Selection

Not all variables are helpful
Variables may be correlated. Therefore they do not offer any additional information and can not strengthen the prediction/model.

Backward Elimination vs. Forward Selection

- Backward Elimination starts with the model that includes all potential predictor variables
  > remove predictor which's removal results in higher R^2_{adj} than no removal
  > remove predictor with p-values above significance level α
- Forward Selection adds variables on-at-a-time until the best fit
  > add predictor with the highest R^2_{adj} until we cannot improve the models R^2_{adj}
  > add predictor with smallest p-value while below significance level α
- R^2 approach is used to improve accuracy, p-value approach is used to include statistically significant predictors

Requirements/Conditions

- The residuals of the model are nearly normal
- The variability of the residuals is nearly constant
- The residuals are independent
- Each variable is linearly related to the outcome

8.3 – Logistic Regression

Logistic Regression
Logistic regression is used to model categorical response variables. Therefore, a numerical response variable is transformed (*link function*) to a probability $\in [0,1]$.

$$log_e(\frac{p_i}{1-p_i}) = logit(p_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} \qquad p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

Conditions
Predictors are linearly related to $logit(p_i)$ (if other predictors const.), outcomes are independent.