DATA2902 Notesinear

Data Visualisation with ggplot

Working with ggplot

- The ggplot() function knows about the data frame. It knows what to map to the aesthetics including axes and fill colour.
- We can then add geometries, labels, and formatting on top of this blank canvas.

Collecting Data

- A *sample* is a part of a population.
- A *parameter* is a numerical fact about a population. Usually, they cannot be determined exactly - only estimated. They are what a researcher wants to know.
- A *statistic* can be computed from a sample, and used to estimate a parameter, summarising what the researcher knows.
    - Used because it is typically too hard to observe the population and there is not enough time/money/resources. Also essential in destructive testing.
- *Accuracy* refers to how close the estimated statistic is to the unknown true parameter.

Sampling

- *Sampling* is the process of selecting a subset of observations from an entire population of interest so that characteristics from the subset (sample) can be used to draw conclusion or making inference about the entire population.

Bias

- *Bias* is any factor that favours certain outcomes or responses, or influences an individual's responses. Bias may be unintentional (accidental), or intentional (to achieve certain results).
    - *Selection/sampling bias*: the sample does not accurately represent the population. Increasing the sample size does not help, since it just repeats the basic mistake at a larger scale.
    - *Non-response bias*: certain groups are under-represented because they elect not to participate.
    - *Measurement or designed bias*: bias factors in the sampling method influence the data obtained. This can include: misinterpretion, confusing/sensitive questions.

Controlled Experiments

Randomised Controlled Double-Blind Trials

- A representative random sample of subjects is obtained, and allocated into a treatment group and a control group.
- The control group is given a placebo, but neither subjects nor investigators know the identity of the two groups.
- The responses are compared - the design is good because any difference in response is likely caused by the treatment.

Observational Studies

- By necessity, many research questions require an observational study rather than a controlled experiment e.g. effects of smoking (no treatment group ethically possible), educational research - they must observe results for the two groups.
- Observational studies can only suggest causation, not establish it. They can establish *association*.

Misleading Hidden Confounders

- Confounding occurs when the treatment and control group differ by some third variable which influences the response that is studied. They can be hard to find and can mislead about a cause and effect relationship.
- They can be introduced into a randomised study if any of the subjects drop out, causing *selection bias* or *survivor bias*.
- If not all subjects keep taking the treatment or placebo, we get the confounding of *adherers* and *non-adherers*.
- Controlling for confounders can be done by dividing them into subgroups with respect to the confounder.
    - Limited by our ability to identify and divide the study based on the confounders.

Simpson's Paradox

- Sometimes, there is a clear trend in individual groups of data that disappears when the groups are pooled together e.g. smoking and death-rates confounded by age groups.
- It occurs when relationships between percentages in subgroups are reversed when the subgroups are combined, because of a confounding or lurking variable.
    - The association between a pair of variables (X, Y) reverses sign based upon conditioning of a third variable Z, regardless of the value taken by Z.

Chi-Squared Tests

Refer to lecture 3 slides for more in-depth code and examples

Example Code

```
# Setup
y <- c(128, 86, 74, 112)
n <- sum(y)
p <- c(1/4, 1/4, 1/4, 1/4)
e <- p*n

# Testing for assumption
print(y >= 5)

# Test Statistic
t0 <- sum((y-e)^2)/e)

# Hypothesis Testing
```

```
chisq.test(y, p=p)
value <- 1 - pchisq(t0, k-1-1) # or pchisq(t0, k-2, lower.tail=FALSE)
```

Hypothesis

- *Null hypothesis*: the statement against which you search for evidence is called the null hypothesis, and is denoted by H_0. It is generally a "no difference" statement.
- *Alternative hypothesis*: the statement you claim is called the alternative hypothesis, and is denoted by H_1 (or sometimes H_A).

Assumptions

- Each observation are generally assumed to have been chosen at random from a population.
- We say that such random variables are iid (idenpendently and identically distributed).
- Each test we consider will have its own set of assumptions.

Test Statistic

- Since observations vary from sample to sample, we can never be sure whether H_0 is true or not. A test statistic is a functions of the observations, $T = f(X_1, \ldots, X_n)$, such that the distribution of T is known assuming that H_0 is true.
    - It can be used to test if the data is consistent with the null hypothesis.
- The observed test statistic, t_0, is where we plug our observed data into the formula for the test statistic.
- Large (positive or negative depending on H_1) observed test statistic values is taken as evidence of poor agreement with H_0.

Significance

- The p-value is defined as the probability of getting a test statistic, T, as or more extreme than the value we observed, t_0, assuming that H_0 is true.

Decision

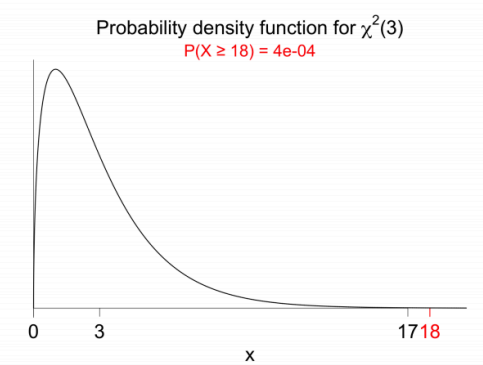
- An observed large positive or negative value of t_0, and hence small p-value is taken as evidence of poor agreement with H_0.
    - If the p-value is small, then either H_0 is true and the poor agreement is due to a unlikely event, or H_0 is false. The smaller the p-value, the stronger the evidence against the null hypothesis.
    - A large p-value does not mean that there is evidence that the null hypothesis true.
    - The level of significance, α, is the strength of evidence needed to reject H_0 (often $\alpha = 0.05$).

Chi-Squared Test Summarised

- The degrees of freedom from the sample is $k - 1$ because the first $k - 1$ observations y_i contain all the information and the last observation is fixed by $y_k = n - \sum_{i-1}^{k-1} y_i$ adding no extra information.
- In general, the test statistic $T \sim \chi^2_{k-1-q}$ where q is the number of parameters needed to be estimated from the sample. In the no linkage example, $q = 0$ as we do not need to estimate any parameters.
- The approximation will only be accurate if no expected frequency is too small, as a rule of thumb, we require all $e_i \geq 5$. Otherwise, we need to pool adjacent categories so that the expected frequencies are always ≥ 5.

- **Hypothesis:** H_0: $p_1 = p_{10},\ p_2 = p_{20}, \cdots,\ p_k = p_{k0}$ vs H_1: at least one equality does not hold.
- **Assumptions:** independent observations and $e_i = np_{i0} \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. Under H_0, $T \sim \chi^2_{k-1-q}$ approximately where k is the number of groups and q is the number of parameters that needs to be estimated from the data.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{k} \frac{(y_i - e_i)^2}{e_i}$.
- **P-value:** $P(T \geq t_0) \ = \ P(\chi^2_{k-1-q} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$.

- **Hypothesis:** H_0: $p_i = \frac{1}{4}$ vs H_1: at least one equality does not hold.
- **Assumptions:** $e_i = np_{i0} \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i}$. Under H_0, $T \sim \chi^2_3$ approx.
- **Observed test statistic:** $t_0 = 18$
- **P-value:** $P(T \geq t_0) = P(\chi^2_3 \geq 18) = 0.0004$
- **Decision:** Since the p-value is < 0.05, there is strong evidence in the data against H_0. Hence the four phenotypes are not equally likely.




```
1 - pchisq(18, df = 3)
```

```
## [1] 0.0004398497
```

Goodness of Fit Tests for Discrete Distributions

Poisson Distribution

- A Poisson random variable represents the probability of a given number of events occuring in a fixed interval if these events occur independently with some known average rate λ per unit time.
- If X is a Poisson random variable with rate parameter λ, then the probability mass function is: $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, 2..$

```
# Can be visualised with
plot(table(rpois(n=10000, lambda=2)), ylab = "Count")
```

Chi-Squared Tests for Discrete Distributions

- Suppose we have a sample $x_1, x_2, \ldots, x_n$.
- We want to test whether the sample is taken from a population with a given distribution function $F_0(x, \theta_1, \theta_2, \ldots, \theta_h)$ where θ_i are parameters of the distribution.
- We may count the frequencies y_i for each value of x_j and compare them to the expected frequencies, e_i, calculated using the expected probabilities, p_i, from the hypothesised distribution $F_0(x, \theta_1, \theta_2, \ldots, \theta_h)$.

- This is a general chi-squared goodness-of-fit test with test statistic: $T = \sum_{i=1}^{k} \frac{(Y_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$.
- Since the model parameters are usually unknown and have to be estimated, the observed test statistic is: $T = \sum_{i=1}^{k} \frac{(Y_i - n\hat{p}_i)^2}{n\hat{p}_i}$ and the approximate p-value is: $P(\chi^2_{k-1-q} \geq t_0)$.

Example Code

```
Y <- C(117, 94, 51, 15, 0, 0, 0, 1) # observed counts
x <- 0:7 # define corresponding groups
n <- sum(y) # total number of samples (sample size)
k <- length(y) # number of groups
lam <- sum(y * x)/n # estimate lambda parameter (decrease df)

p <- dpois(x, lambda = lam) # obtain p_i from the Poisson pmf
p[8] <- 1 - sum(p[1:7]) # redefine the 8th element
ey <- n * p # calculate expected frequencies
print(ey >= 5) # check assumption, not all satisfied

yr <- c(y[1:3], sum(y[4:8]))
eyr <- c(ey[1:3], sum(ey[4:8]))
pr <- c(p[1:3], sum(p[4:8]))
kr <- length(yr) # number of combined classes
t0 <- sum((yr-eyr)^2/eyr)
pval <- pchisq(t0, kr-1-1) # 0.48, we do not reject the null hypothesis
```

Measures of Performance

Types of Errors

- In general, positive = identified and negative = rejected.
- To formalise this, let:
  - D^+ be the event that an individual has a particular disease. The prevalence is the marginal probability of disease, $P(D^+)$.
  - D^- be the event that an individual does not have a particular disease.
  - S^+ represent a positive screening test result or a symptom being present.
  - S^- represent a negative screening test result or no symptom.

	Actual positive D^+	**Actual negative D^-**
Test positive S^+	True positive (TP) ✓	False positive (FP) ⊗
Test negative S^-	False negative (FN) ⊗	True negative (TN) ✓

	Actual positive D^+	Actual negative D^-	
Test positive S^+	a	b	$a+b$
Test negative S^-	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

- **False negative rate**: $P(S^-|D^+) = \frac{c}{a+c}$
- **False positive rate**: $P(S^+|D^-) = \frac{b}{b+d}$
- **Sensitivity/Recall**: $P(S^+|D^+) = \frac{a}{a+c}$, the probability that the test is positive given that the subject actually has the disease
- **Specificity**: $P(S^-|D^-) = \frac{d}{b+d}$, the probability that the test is negative given that the subject does not have the disease
- **Positive predictive value/Precision**: $P(D^+|S^+) = \frac{a}{a+b}$, the probability that the subject has the disease given that the test is positive
- **Negative predictive value**: $P(D^-|S^-) = \frac{d}{c+d}$, the probability that the subject does not have the disease given that the test is negative
- **Accuracy**: $\frac{a+d}{a+b+c+d}$

Conditional Probability

- Let B be an event so that $P(B) > 0$. The conditional probability of an event A given that B has occured is: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Rearranging, we also have: $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$.
- For the classical definition of probability, where we have a finite number of equally likely outcomes, then:
  - $P(A)$ is the proportion of outcomes in the subset A.
  - $P(B)$ is the proportion of outcomes in the subset B.
  - $P(A|B)$ can be interpreted as the proportion of outcomes in B that are also in A.

Bayes' Rule

- Allows us to reverse the conditioning set provided that we know some marginal probabilities:
  - $P(B|A) = \frac{P(B \cap A)}{P(A)}$
    $= \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$ where B^c is the event that B doesn't occur.

Evaluating Classification Models

Refer to lecture 5 slides for code relating to kyphosis example

Measures of Risk

Prospective and Retrospective Studies

- A *prospective (or cohort study)* is a study design based on subjects who are initially identified as disease-free and classified by presence or absence of a risk factor.
  - A random sample from each group is followed in time (prospectively) until eventually classified by disease outcome.

    - The totals for risk factor rows are fixed by design.
- A *retrospective study* is based on random samples from each of the two outcome categories which are followed back retrospectively to determine the presence or absence of the risk factors for each individual.
    - The totals for the outcome columns are fixed by design.

Estimating Population Proportions

- In both prospective and retrospective studies, we have:
    - a population;
    - a subpopulation/attribute determined by a risk factor R^+ (with complementary subpopulation/attribute R^-);
    - a subpopulation/attribute determined by having/developed the disease D^+ (with complementary subpopulation/attribute D^-).
- The main difference between the types of studies are which subpopulations we can sample from.

Prospective Study

- We take two random samples:
    - one from the risk factor group R^+;
    - another from the non-risk factor group R^-.
- We then wait to see how many in each group develop the disease, estimating $P(D^+|R^+)$ as well as $P(D^-|R^-)$.
- We cannot, however, estimate $P(R^+|D^+)$ or $P(R^-|D^-)$ since we did not take random samples from the disease group.

Retrospective Study

- We take two random samples:
    - one from the disease group D^+;
    - another from the non-disease group D^-.
- We then look back to see how many in each group were exposed to the risk factor., estimating $P(R^+|D^+)$ as well as $P(R^-|D^-)$.
- We cannot, however, estimate $P(D^+|R^+)$ or $P(D^-|R^-)$ since we did not take random samples from the risk factor group.

Relative Risk

- The relative risk is defined as a ratio of two conditional probabilities.
    - $RR = \frac{P(D^+|R^+)}{P(D^+|R^-)}$
- If D and R are independent, then $P(D|R) = P(D)$ and so $RR = \frac{P(D^+)}{P(D^+)} = 1$.
- The relative risk is the ratio of the probability of having the disease in the group with the risk factor, to the probability of having the disease in the group without the risk factor.
    - $RR = 1/ < 1/ > 1$ implies that the disease is equally/less/more likely to occur in the group with the risk factor.

Odds Ratio

- Odds are a ratio of probabilities. The odds are used as an alternative way of measuring the likelihood of an event occurring. If the probability of event A is $P(A)$ the odds of event A is defined as:
  - $O(A) = \frac{P(A)}{1-P(A)}$.
  - $O(D^+|R^+) = \frac{P(D^+|R^+)}{1-P(D^+|R^+)} = \frac{P(D^+|R^+)}{P(D^-|R^+)}$
- If D and R are independent, then $P(D|R) = P(D)$ and $OR = 1$.
- Large odds ratios ($OR > 1$) implies increased risk of disease and small odd ratios ($OR < 1$) implies decreased risk of disease.

Calculations

	D^+	D^-	Total
R^+	a	b	$a+b$
R^-	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

- Given a prospective study from a sample of completed records, we can estimate (note that RR cannot be calculated for retrospective studies):
  - $P(D^+|R^+) = \frac{a}{a+b}$
  - $P(D^+|R^-) = \frac{c}{c+d}$
  - Relative Risk: $\widehat{RR} = \frac{P(D^+|R^+)}{P(D^+|R^-)} = \frac{a(c+d)}{c(a+b)}$
- Odds ratio's can be found from both prospective and retrospective studies, calculated as thus:
  - $OR = \frac{P(D^+|R^+)}{P(D^-|R^+)} / \frac{P(D^+|R^-)}{P(D^-|R^-)} = (\frac{\frac{a}{a+b}}{\frac{b}{a+b}}) / (\frac{\frac{c}{c+d}}{\frac{d}{c+d}}) = \frac{ad}{bc}$
  - $OR = \frac{P(R^+|D^+)}{P(R^-|D^+)} / \frac{P(R^+|D^-)}{P(R^-|D^-)} = (\frac{\frac{a}{a+c}}{\frac{c}{a+c}}) / (\frac{\frac{b}{b+d}}{\frac{d}{b+d}}) = \frac{ad}{bc}$

Standard Errors and Confidence Intervals for Odds Ratios

- The *odds ratio* estimator, OR, has a skewed distribution on $(0, \infty)$, with the neutral value being 1.
- The *log odds* estimator, $log(OR)$, has a more symmetric distribution centred at 0 if there is no difference between the two groups ($log_e x$).
- Note: an odds ratio of $a \in (0, 1)$ is equivalent to a value of $a^{-1} \in (1, \infty)$ just by relabelling the categories. The log transformation is such that $log(a^{-1}) = -log(a)$.

Standard Errors and Confidence Intervals

- The *asymptotic standard error* for $log(\widehat{OR})$ is: $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$.
- A large sample 95% *confidence interval* for $log(\theta)$ is approximately:
  $log(\widehat{OR}) \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
  - from which we can approximate a confidence interval for the odds-ratio
    $(\exp\left(log(\widehat{OR}\right) - 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}), \exp\left(log(\widehat{OR}\right) + 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}))$
- Note that these should only be applied if a, b, c, d are reasonably large so that asymptotics hold.

Testing for Homogeneity

Testing for Homogeneity

- Supposed that several samples are taken from r independent populations, each of which is categorised according to the same set of c variables .
- We want to test whether the probability distributions (proportions) of the categories are the same over the different populations.
- A *contingency table* allows us to tabulate data from multiple categorical variables.

	Category 1	Category 2	...	Category c	Row total (fixed)
Population 1	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
Population 2	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population r	y_{r1}	y_{r2}	$\cdots$	y_{rc}	$y_{r\bullet}$
Column total	$y_{\bullet 1}$	$y_{\bullet 1}$	$\cdots$	$y_{\bullet 1}$	$y_{\bullet\bullet} = n$

	Category 1	Category 2	...	Category c	Row total
Population 1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\bullet} = 1$
Population 2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\bullet} = 1$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	$p_{r\bullet} = 1$

- Under the null hypothesis of homogeneity:

  $p_{11} = p_{21} = \ldots = p_{r1}$,

  $p_{12} = p_{22} = \ldots = p_{r2}$,

  $p_{1c} = p_{2c} = \ldots = p_{rc}$

- As we don't know p_{ij}, we need to estimate it, $\hat{p}_{ij} = \frac{y_{\bullet j}}{n}$.
- Under H_0:
  - The expected counts are: $e_{ij} = n_i \hat{p}_{ij} = y_{i\bullet} \frac{y_{\bullet j}}{n} = \frac{\text{Row i total} \times \text{Column j total}}{\text{Overall total}}$
  - The test statistic is: $T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$, where $(r-1)(c-1)$ is the degrees of freedom.

Workflow

- **Hypothesis:** H_0: $p_{1j} = p_{2j} = \ldots = p_{rj} \quad j = 1, 2, \ldots, c$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$ and independent observations.
- **Test statistic:** $T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$
- **P-value:** $P(T \geq t_0) = P(\chi^2_{(r-1)(c-1)} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

Workflow for 2×2 Contingency Table

y_{ij}	Sesame Street	Play School	Row total
Boys	42	58	100
Girls	86	114	200
Column total	128	172	300

e_{ij}	Sesame Street	Play School	Row total
Boys	$\frac{100 \times 128}{300} = 42.67$	$\frac{100 \times 172}{300} = 57.33$	100
Girls	$\frac{200 \times 128}{300} = 85.33$	$\frac{200 \times 172}{300} = 114.67$	200
Column total	128	172	300

Observed test statistic:

$$\begin{aligned} t_0 &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(42 - 42.67)^2}{42.67} + \frac{(58 - 57.33)^2}{57.33} + \frac{(86 - 85.33)^2}{85.33} + \frac{(114 - 114.67)^2}{114.67} \\ &= 0.027. \end{aligned}$$

- **Hypothesis:** H_0: $p_{11} = p_{21}$ & $p_{12} = p_{22}$ vs H_1: $p_{11} \neq p_{21}$ & $p_{12} \neq p_{22}$ or viewing habits are homogenous across boys and girls
- **Assumptions:** $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_1$ approx.
- **Observed test statistic:** $t_0 = 0.027$
- **P-value:** $P(T \geq t_0) = P(\chi^2_1 \geq 0.027) = 0.87$
- **Decision:** Do not reject H_0 as the p-value is quite large, i.e. there is no significant difference in the viewing habits between boys and girls.

Tests for Independence

Tests for Independence in 2×2 Tables

$$p_{i\bullet} = \sum_{j=1}^{2} p_{ij} \quad \text{and} \quad p_{\bullet j} = \sum_{i=1}^{2} p_{ij}$$

	Survived	Did not survive	Row total
Male	p_{11}	p_{12}	$p_{1\bullet}$
Female	p_{21}	p_{22}	$p_{2\bullet}$
Column total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

- X and Y are said to be independent if: $P(X = x|Y = y) = P(X = x)$ or $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- Under the null hypothesis of independence, the expected frequencies are $e_{ij} = np_{ij} = np_{i\bullet}p_{\bullet p}$.
  - We estimate $p_{i\bullet}$ and $p_{\bullet}j$ by $\hat{p}_{i\bullet} = y_{i\bullet}/n$ and $\hat{p}_{\bullet j} = y_{\bullet j}/n$.
- We calculate the observed test statistic: $t_0 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j})^2}{y_{i\bullet}y_{\bullet j}}$.

Tests for Independence in $r \times c$ Tables

	$S = 1$	$S = 2$	$\cdots$	$S = c$	Total
$R = 1$	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
$R = 2$	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$R = r$	y_{r1}	y_{12}	$\cdots$	y_{rc}	$y_{r\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet c}$	$y_{\bullet\bullet}$

Hence, $y_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij} = n$, the sample size.

- We want to test if R and S are independent i.e. that $p_{ij} = P(R = i, S = j) = P(R = i)P(S = j)$.
- This can similarly presented in a continency table as follows:

		Variable 1				
		Level 1	Level 2	...	Level c	Row total
Variable 2	Level 1	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
	Level 2	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	Level r	y_{r1}	y_{r2}	$\cdots$	y_{rc}	$y_{r\bullet}$
	Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet 3}$	$y_{\bullet\bullet} = n$

- The degrees of freedom for a $r \times c$ table is $(r - 1)(c - 1)$.

- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n}$.
- **P-value:** $P(T \geq t_0) = P(\chi^2_{(r-1)(c-1)} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}, \quad i = 1, 2;\ j = 1, 2$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_1$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n} = 456.87$
- **P-value:** $P(T \geq t_0) = P(\chi^2_1 \geq 456.87) < 0.001$
- **Decision:** We reject the null hypothesis that gender is independent of survival as the p-value is very small (much smaller than 0.05). Hence, there is evidence to suggest that survival status of passengers on the Titanic is related to the gender of the passanger.

Testing in Small Samples

The Hypergeometric Distribution

- Relates to sampling without replacement from a finite population.
  - The result of each draw (the elements of the population being sampled) can be classified into one of two mutually exclusive categories e.g. (Pass/Fail or Employed/Unemployed).
  - The probability of a success changes on each draw, as each draw decreases the population (sampling without replacement from a finite population).
- A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ where:
  - N is the population size.
  - K is the number of success states in the population.
  - n is the number of draws (i.e. quantity drawn in each trial), and
  - k is the number of observed successes.

Fisher's Exact Test

- The χ^2 approximation for the test statistic is only reasonable when n is sufficiently large i.e. we need the expected cell frequencies to all be 5 or more.
    - However, if this is not the case, then we need to take care and maybe consider exact tests i.e. calculating the exact p-value for the test statistic.

	A_1	A_2	**Total**
B_1	y_{11}	y_{12}	$y_{1\bullet}$
B_2	y_{21}	y_{22}	$y_{2\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$	n

- For a 2×2 table, if we know the row and column and y_{11}, then the table is completely specified. Let θ be the odds ratio. A test of $H_0 : \theta = 1$ vs $H_1 : \theta > 1$ (or $H_1 : \theta < 1$) can be based on the observed value of y_{11} given the marginal totals.
- If H_0 is true and we know $y_{1\bullet}$, $y_{\bullet 1}$ in the $(1,1)$th cell. To obtain the distribution of y_{11} given the marginal values, note the situation is like selecting $y_{\bullet 1}$ values from n where $y_{1\bullet}$ are type B_1 and $y_{2\bullet}$ are type B_2. Then: $P(Y_{11} = y_{11}) = \frac{\binom{y_{1\bullet}}{y_{11}}\binom{y_{2\bullet}}{y_{\bullet 1} - y_{11}}}{\binom{n}{y_{\bullet 1}}}$.

p-values

- To calculate the p-value for a particular table, we need to enumerate all tables as or more extreme than the observed table with the same marginal totals, and sum the probability of each of these tables.

	Cancer controlled	Cancer not controlled	Total
Surgery	21	2	23
Radiation therapy	15	3	18
Total	36	5	41

$$
\begin{aligned}
\text{p-value} &= P(X \geq 21 \mid \text{marginal totals}) \\
&= P(X = 21, 22, 23 \mid \text{marginal totals}) \\
&= P(X = 21 \mid \text{marginal totals}) + P(X = 22 \mid \text{marginal totals}) + P(X = 23 \mid \text{marginal totals}) \\
&= \frac{\binom{23}{21}\binom{18}{15}}{\binom{41}{36}} + \frac{\binom{23}{22}\binom{18}{14}}{\binom{41}{36}} + \frac{\binom{23}{23}\binom{18}{13}}{\binom{41}{36}} \\
&= 0.3808.
\end{aligned}
$$

Drawbacks

- Why don't we use Fisher's exact test all the time?
    - It assumes that row and column margins are fixed.
    - Computationally difficult for large samples.
    - It can be generalised to $r \times c$ two-way contigency tables but it is very difficult to compute. Generally requires the use of Monte Carlo (i.e. random permutation).

Yates' Corrected χ^2 Test

- Yates (1934) modified the standard chi-squared test with a continuity correction. It is usually more accurate when counts in each cell are small. Yates' statistic for 2×2 tables is: $T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|Y_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$, which approximately follows a χ_1^2 distribution under H_0.
- The intuition between continuity corrections is that using the right edge value of a histogram gives a better approximation (similar to a right Reimann sum).

Permutation Testing (Monte Carlo Simulation)

- The Monte Carlo simulation procedure is as follows:
  - Analyse the sample as one would normally do in a hypothesis test (up to, and including, the calculation of the test statistic).
  - From the original sample being analysed, resample it LOTS of times (i.e. bootstrap).
  - The test statistic of interest is calculated for each of the resamples (so that we have the sampling distribution of the test statistic).
  - This leads to LOTS of test statistics that will be used to calculate p-values for the observed statistic.
- Monte Carlo p-values are calculated by determining the proportion of the resampled test statistics as or more extreme than the observed test statistic. No assumptions are made about the underlying distribution of the population.
- They can be obtained by randomly generating contingency tables given that the margins are assumed fixed. For each of the x randomly generated contigency tables , we can record their test statistic then determine what proportion of them are equal to (or exceed) the observed test statistic.

Testing Means

General t-test Background

- Some basic probability facts about samples from normal populations will prove useful.
  - The sample mean from a normal sample is itself normally distributed.
  - The sample variance from a normal sample has a scaled χ^2 distribution.
  - The sample mean and sample variance from a normal sample are statically independent.
- If $Z \sim N(0,1)$ is independent of a χ_d^2 random variable, the quantity

$$\frac{Z}{\sqrt{\chi_d^2/d}} \sim t_d$$

  is a t-distribution with d degrees of freedom.
- If the population mean is μ, the sample mean and variance are $\bar{X}$ and S^2, the ratio

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{S/\sigma} \sim t_{n-1}$$

  - the numerator is $N(0,1)$;
  - the denominator is $\sqrt{\chi_{n-1}^2/(n-1)}$, independently of the numerator.

- Indeed, in many statistical applications we have a model whereby a certain statistic has this general form:
    - some estimator of some parameter is normally distributed;
    - a SE based on the data has a distribution like $\sqrt{\chi^2_d/d}$ times the true SD of the estimator (for some d) and is independent of the estimator;
    - then the ratio is $\frac{\text{estimator - true value}}{\text{standard error}} \sim t_d$.

Conducting t-tests

Assumptions

- Each observation $X_1, X_2, \ldots, X_n$ is chosen at random from a population.
- We say that such random variables are iid (independently and identically distributed).
- Each test will have its own assumptions.

Hypothesis

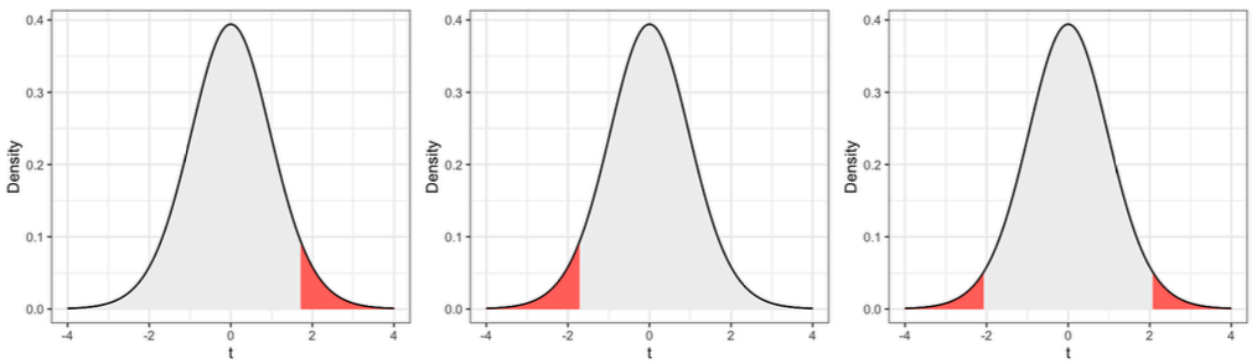
- The statement against which you search for evidence is the null hypothesis, and is denoted by H_0. It is generally a "no difference" statement.
- The statement you claim is called the alternative hypothesis, and is denoted by H_1.
    - $H_0 : \theta = \theta_0$ vs
    - $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ or $H_1 : \theta \neq \theta_0$ (upper-side/lower-side/two-sided alternative).

Test Statistic

- The observed test statistic, t_0, is where we plug our observed data into the formula for the test statistic.
- Large (positive or negative depending on H_1) observed test statistic values is taken as evidence of poor agreement with H_0.
    - Since observation X_i, vary from sample to sample we can never be sure whether H_0 is true or not.
    - We use a test statistic $T = f(X_1, \ldots, X_n)$ to test if the data are consistent with H_0 such that the distribution of T is known assuming H_0 is true.

Significance

- The p-value is defined as the probability of getting a test statistic, T, as or more extreme than the value we observed, t_0, assuming that H_0 is true.
- For $H_1 : \theta > \theta_0$, p-value = $P(T \geq t_0)$
- For $H_1 : \theta < \theta_0$, p-value = $P(T \leq t_0)$
- For $H_1 : \theta \neq \theta_0$, p-value = $2P(T \geq |t_0|)$



Decision

- An observed large positive or negative value of t_0, and hence small p-value is taken as evidence of poor agreement with H_0.
  - If the p-value is small, then either H_0 is true and the poor agreement is due to an unlikely event, or H_0 is false. Therefore,
    - The smaller the p-value, the stronger the evidence against H_0 in favour of H_1.
    - A large p-value does not mean that there is evidence that H_0 is true.
    - A level of significance, α, is the strength of evidence needed to reject H_0 (often $\alpha = 0.05$).

One Sample t-test

- Suppose we have a sample $X_1, X_2, \ldots, X_n$ of the size n drawn from a normal population with an unknown variance σ^2. Let $x_1, x_2, \ldots, x_n$ be the observed values. We want to test the population mean μ.

- **Hypothesis:** H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$, $\mu < \mu_0$ or $\mu \neq \mu_0$
- **Assumptions:** X_i are iid rv and follow $N(\mu, \sigma^2)$.
- **Test statistic:** $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$. Under H_0, $T \sim t_{n-1}$.
- **Observed test statistic:** $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- **p-value:** $P(t_{n-1} \geq t_0)$, $P(t_{n-1} \leq t_0)$ or $2P(t_{n-1} \geq |t_0|)$
- **Decision:** Reject H_0 in favour of H_1 if the p-value is less than α.

Two Sample t-test

- There are times that we want to test if the population means of two samples are different. Here, we are left with two possible scenarios:
  - Two independent samples
  - Two related samples (dependent samples or repeated measures).

1. **Hypotheses:** H_0: $\mu_x = \mu_y$ vs H_1: $\mu_x > \mu_y$ or $\mu_x < \mu_y$ or $\mu_x \neq \mu_y$
2. **Assumptions:** $X_1, \ldots, X_{n_x}$ are *iid* $N(\mu_X, \sigma^2)$, $Y_1, \ldots, Y_{n_y}$ are *iid* $N(\mu_Y, \sigma^2)$ and X_i's are independent of Y_i's.
3. **Test statistic:** $T = \dfrac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S_p^2 = \dfrac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$. Under H_0, $T \sim t_{n_x+n_y-2}$
4. **Observed test statistic:** $t_0 = \dfrac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$.
5. **p-value:** $P(t_{n_x+n_y-2} \geq t_0)$ or $P(t_{n_x+n_y-2} \leq t_0)$ or $2P(t_{n_x+n_y-2} \geq |t_0|)$.
6. **Decision:** If the p-value is less than α, there is evidence against H_0. If the p-value is greater than α, the data are consistent with H_0.

Assumptions and Welch Two-Sample t-test

- We should note that the two sample t-test assumes that the two underlying normal populations have the same variance.
- Welch developed an alternative test which does not assume equal population variances.
- In that case, if the X_i's are $N(\mu_X, \sigma_X^2)$ and Y_i's are $N(\mu_Y, \sigma_Y^2)$, then the variance of the sample mean difference is $Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}$.
- The standard error is obtained by plugging in the two sample variances and taking the square root (we do not need to compute a "pooled" estimate of the common variance!).
    - This gives the Welch statistic:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_x} + \frac{X_Y^2}{n_y}}}$$

- This statistic is not a "usual" t-statistic since the denominator is not a scaled χ^2 independent of the numerator. However, the whole statistic has an approximate t-distribution; the degrees of freedom is not necessarily a whole number, and is estimated from the data.

Paired Samples t-test

- Used to determine if there are differences between two paired samples. In this case, find the differences between the samples and perform a one sample t-test on these differences.
    - E.g. mean difference in muscle weight between treated leg and placebo leg.

- **Hypothesis:** H_0: $\mu_d = 0$ vs H_1: $\mu_d > 0$
- **Assumptions:** D_i are independent and identically distributed (iid) $N(\mu, \sigma^2)$.
- **Test statistic:** $T = \frac{\bar{D} - \mu_d}{S_d/\sqrt{n}}$. Under H_0, $T \sim t_{n-1}$
- **Observed test statistic:** $t_0 = \frac{0.25}{0.33/\sqrt{10}} = 2.39$
- **p-value:** $P(t_9 \geq 2.39) = 0.02$
- **Conclusion:** The p-value is less than 0.05, therefore we reject the null hypothesis at the 5% level of significance and conclude that the biochemical substance does inhibit muscle growth.

Critical Values, Rejection Regions and Confidence Intervals

Random Variables

- A random variable can be thought of as a mathematical object which takes certain values with certain probabilities.
- We have discrete and continuous random variables, although we can always 'approximate' a continuous one with a discrete one (taking values on a suitably fine grid).
- A single discrete random variable X can be described as a single random draw from a "box" containing

tickets, each with numbers written on them.
- In this case: $E(X) = \mu$ (the average of the numbers in the box); $Var(X) = \sigma^2$ (the population variance of the numbers in the box) and $SD(X) = \sigma$.

Random Sample With Replacement

- Now consider a random sample of size n with replacement, with values $X_1, X_2, \ldots, X_n$ where each can be chosen equally likely. Considering the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}T$.
- The expectation of a sum is always the sum of the expectations.

$$E(T) = E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n) = \mu + \ldots + \mu = n\mu$$
$$E(\bar{X}) = \frac{n\mu}{n} = \mu$$

- Multiplying by a constant: for any random variable X and any constant c,

$$E(cX) = cE(X) \text{ and } Var(cX) = c^2 Var(X)$$

- The variance of a sum of independent random variables is not always the sum of the variances. However, it is if the values are independent.

$$Var(T) = Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n) = \sigma^2 + \ldots + \sigma^2 = n\sigma^2$$
$$Var(\bar{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Estimating μ

- In many applications, we model $x_1, \ldots, x_n$ as values taken by such a sample $X_1, \ldots, X_n$ and we are interested in estimating or learning μ. The estimator is the sample mean $\bar{X}$ (the random variable).
- The estimate is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the observed value of the mean of the data.
- The standard error, or standard deviation of the estimator, tells us the "likely size of the estimation error", and is given by:

$$SE = SD(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

- Since we generally don't know the population mean, we can estimate σ^2 using the sample variance. The correspoinding standard error is: $\widehat{SE} = s/\sqrt{n}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Critical Values and Confidence Intervals

- We want to know if a given value μ_0 is a "plausible value" for the unknown μ, based on observed data $x_1, \ldots, x_n$. Roughly, we do this by:
  1. computing the value of the estimate $\bar{x}$;
  2. computing the value of the estimated standard erorr $s/\sqrt{n}$;
  3. seeing if the discrepancy $\bar{x} - \mu_0$ is "large" compared to the standard error.

- The types of discrepancies which are of interest include whether the $\bar{x}$ is significantly more/less/different to μ_0.

Two-Sided Discrepancies of Interest

- When two-sided discrepancies are of interest, we are basically asking: for a given μ_0, is the absolute value $|\bar{x} - \mu_0|$ large compared to the standard error $s/\sqrt{n}$.
  - t-test approach: declare μ_0 not plausible if $|\bar{x} - \mu_0| > c\frac{s}{\sqrt{n}}$ for some "suitably chosen" constant c.
  - Confidence interval approach: the set of plausible values for the unknown μ is $\bar{x} \pm c\frac{s}{\sqrt{n}}$ for some "suitably chosen" constant c.
- The constant c can be chosen in a sensible way in each context: testing controls the false alarm rate, confidence intervals control the coverage probability.

False Alarm Rate

- A "false alarm" is when we reject incorrectly. It is also called the significance level.
- We pick small $0 \leq \alpha \leq 1$ for the desired "false alarm rate" e.g. 0.05, 0.01, such that if possible (when we know that the values are taken by iid normal random variables).

$$P(|\bar{X} - \mu_0| > c\frac{S}{\sqrt{n}}) = P(|t_{n-1}| > c) = \alpha$$

- When calculating c, make sure to account for differences between two-sided and one-sided tests.

Coverage Probability

- For a confidence interval, the coverage probability is simply the probability that the "true" value of the unknown parameter lies inside (is "covered by") the confidence interval.
- This is a long run property and should be interpreted in the context of repeated experiments.
- If we choose a non-coverage probability α, then the coverage probability is $1 - \alpha$.
- Thus, under some statistical model model we choose c so that the coverage probability under the model satisfies (with μ the true population mean):

$$P(\bar{X} - c\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + c\frac{S}{\sqrt{n}}) = P(|\bar{X} - \mu| \leq c\frac{S}{\sqrt{n}}) = 1 - \alpha$$

One-Sided Discrepancies of Interest

- One-sided discrepancies of interest are more or less the same (and this is equally reflected for greater-than one-sided discrepancies). We now refine our approaches:
  - t-test approach: declare μ_0 not plausible if $\bar{x} \leq \mu_0 - c\frac{s}{\sqrt{n}}$.
  - Confidence interval approach: set of plausible values for the unknown μ are those "not too much bigger than $\bar{x}$", i.e. $(-\infty, \bar{x} + c\frac{s}{\sqrt{n}}]$.
    - False Alarm Rate: $P(\bar{X} < \mu_0 - c\frac{S}{\sqrt{n}}) = P(t_{n-1} < -c) = \alpha$
    - Coverage Probability: $P(\mu_0 \leq \bar{X} + c\frac{S}{\sqrt{n}}) = P(t_{n-1} \leq c) = 1 - \alpha$

P-Values

- The observed significance level (or p-value) is the value of α for which the observed data is "right on the edge".
    - The smallest false alarm rate for which we would "reject" a given value μ_0;
    - the non-coverage probability for which μ_0 is on the boundary of the confidence interval.

Rejection Regions

Decision Rules

- To test a hypothesis, we previously defined a decision rule to reject H_0. This is when the p-value is less than certain fixed pre-assigned levels e.g. $\text{p-value} < \alpha \text{ where } \alpha = 0.05$.
    - The α is called the significance level of the test, which is the boundary between rejecting and not rejecting H_0.

Notation

- Let $t_{n-1}(\alpha)$ or $z(\alpha)$ be the critical values (or quantile) given by $P(t_{n-1} \leq t_{n-1}(\alpha)) = \alpha$, or if we are using the standard normal distribution, $Z \sim N(0,1)$: $P(Z \leq z(\alpha)) = \alpha$

Decision rule

For a test of H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$, the **decision rule** at level α is:

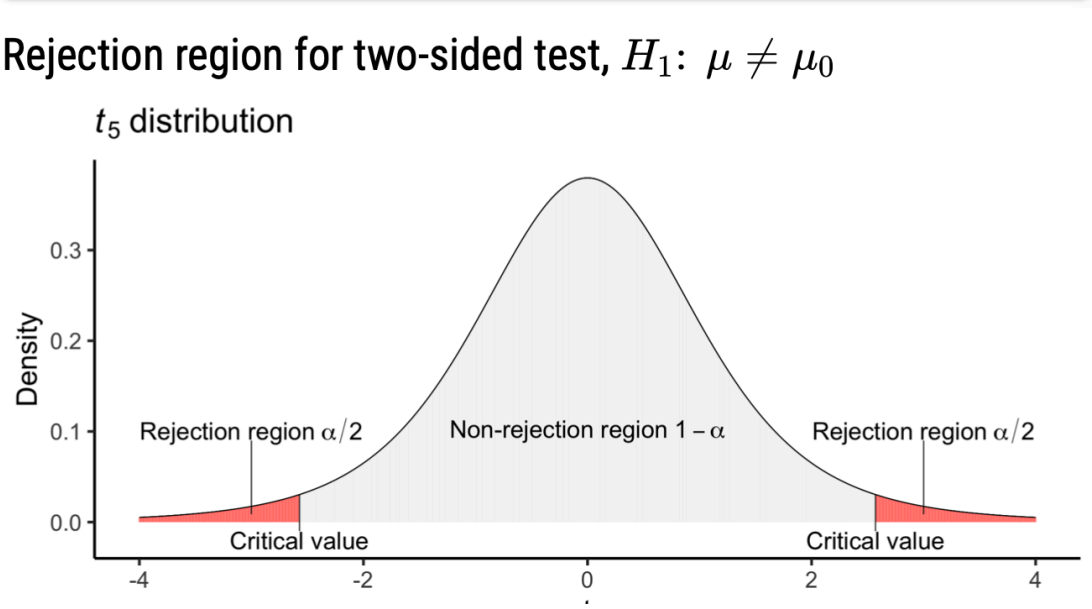
- reject H_0 if $t_0 \geq t_{n-1}(1-\alpha)$ or eqivalently reject H_0 if $t_0 \geq |t_{n-1}(\alpha)|$

For a test of H_0: $\mu = \mu_0$ vs H_1: $\mu < \mu_0$, the **decision rule** at level α is:

- reject H_0 if $t_0 \leq t_{n-1}(\alpha)$

For a test of H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$, the **decision rule** at level α is:

- reject H_0 if $|t_0| \geq |t_{n-1}(\alpha/2)|$
- do not reject H_0 if $|t_0| < |t_{n-1}(\alpha/2)|$



Linking Decision Rules with Confidence Intervals

- If the population parameter is inside the confidence interval, then it is within the range of plausible values.
- Do not reject H_0 at the α level of significance if the value of the population parameter under the null hypothesis is inside the $100(1-\alpha)$ confidence interval.

Sample Size Calculations and Power

- Recalling:

	H_0 **true (innocent)**	H_0 **false (guilty)**
Don't reject H_0 (acquit)	Correct decision	Type II error (β)
Reject H_0 (guilty)	Type I error (α)	Correct decision ($1-\beta$)

- Type I Errors: level of significance, $\alpha = P(\text{reject } H_0 \,|\, H_0 \text{ true})$.
- Type II Errors: call it β
- Power: $1-\beta = P(\text{reject } H_0 \,|\, H_1 \text{ true})$.

General Testing Setup

- Consider a two sided test in which we will reject the μ_0 as "implausible":

$$\text{Reject if } |\bar{x} - \mu_0| > c\frac{s}{\sqrt{n}}$$

  - where c is chosen so that the false alarm rate is some fixed, small value α (e.g. 0.05, 0.01).
- The false alarm rate determination can only be made if a suitable statistical model is assumed for the data (assume that it follows a t-distribution). The false alarm rate is, by symmetry:

$$P_{\mu_0}(|\bar{X} - \mu_0| > c\frac{S}{\sqrt{n}}) = P_{\mu_0}(\frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > c) = P(|t_{n-1} > c) = 2P(t_{n-1} > c)$$

Power

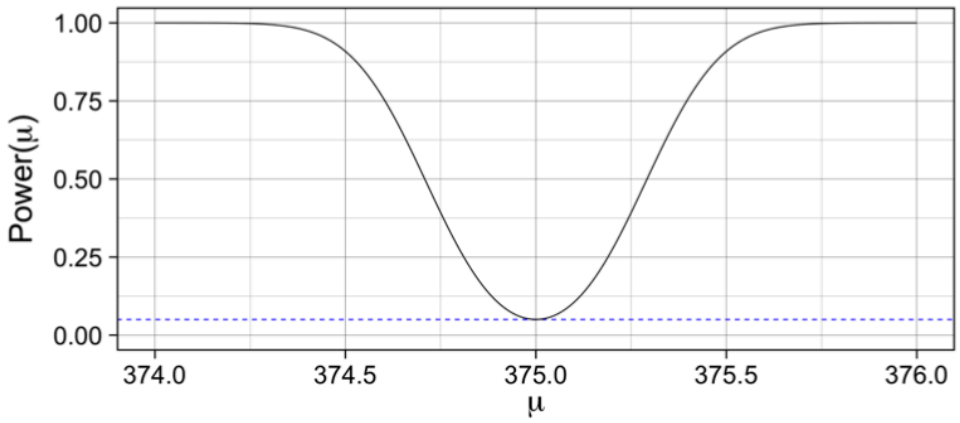
- If you were to make the false alarm rate extremely low, you would never reject anything since the test would have no power.

$$\text{Power} = \text{P}(\text{reject } H_0 | H_1 \text{ is true}).$$

$$P_\mu(\text{reject } H_0) = P_\mu(|\bar{X} - \mu_0| > c\frac{S}{\sqrt{n}}) = P_\mu(\frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > c)$$

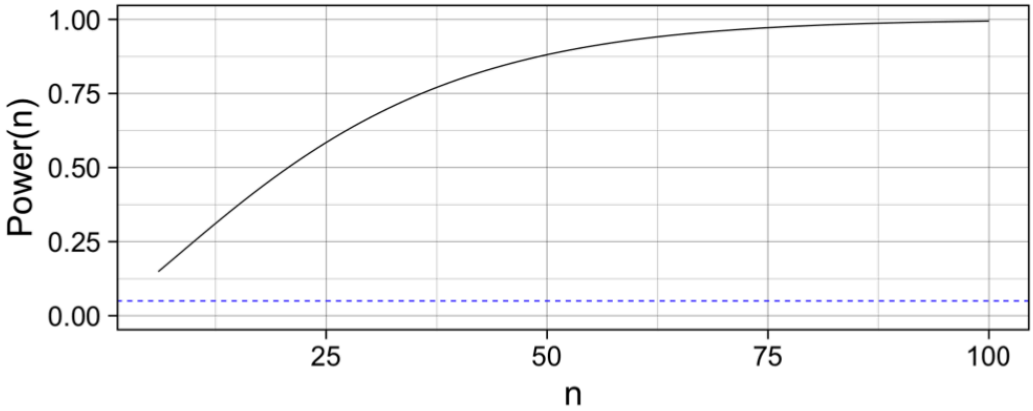
Power as a Function of μ

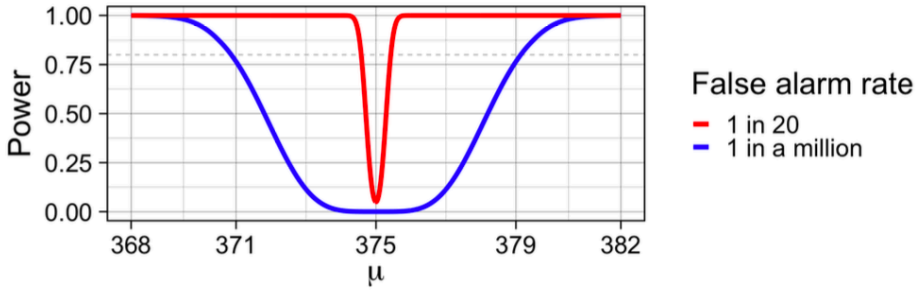
- Supposing the "true" σ is equal to the estimate of 0.294; it is all a guess, but it is still useful as an "estimated" power function.



Power as a Function of n

- Supposing that both the sample mean and standard deviation are indicative of the "true" values μ and σ .





Power of about 80% would require a true μ lower than 371 or more than 379!

Cohen's d

- Cohen suggests that d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes.
- It is given by $d = \frac{|\mu_1 - \mu_2|}{\sigma}$, and is used by **pwr** functions within R.

Sign Test

Assumptions

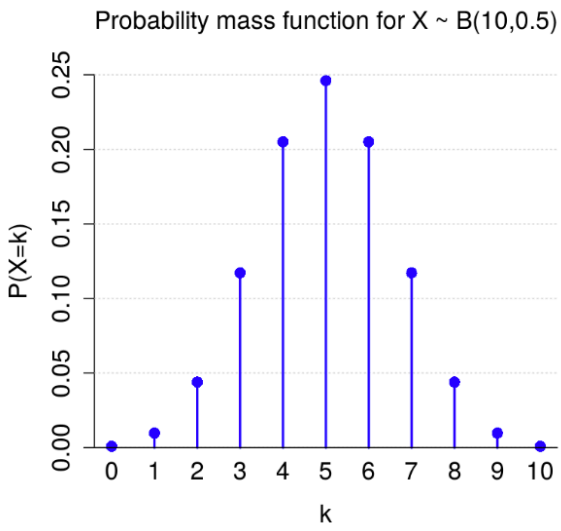
Normality

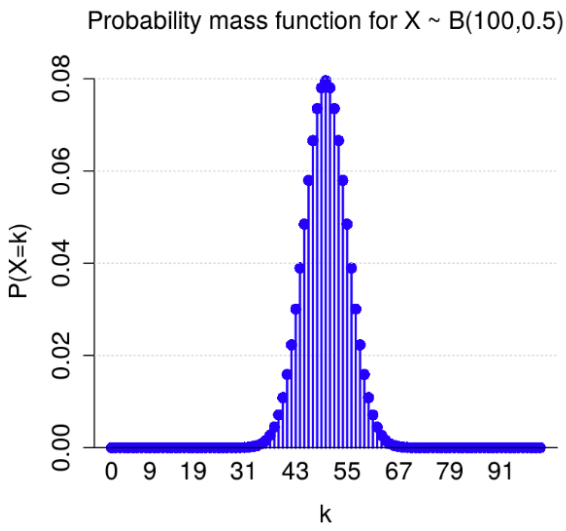
- The assumption that your data is sampled from a normal population arises quite often.
- If you have a large enough sample size, then you can usually rely on the central limit theorem. In small samples, you can check the normality assumption using:
  - Boxplots: we're mostly looking for symmetry.
  - QQ-plots: we're mostly looking for points that lie reasonably close to the line.

Sign Test

- Suppose a sample $X_1, \ldots, X_n$ is taken from a continuous distribution. We want to test $H_0 : \mu = \mu_0$.
- If the distribution is symmetric about μ_0 under H_0, then $D_i = X_i - \mu_0$ should scatter around 0.
- If H_0 is true, the probability of getting p_+ is 0.5.
- The sign test reduces to a binomial test of proportions, and is a nonparametric test since no assumption on the data distribution is made except symmetry.

- **Hypothesis:** H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$, $\mu < \mu_0$, $\mu \neq \mu_0$ or can also write it in terms of the probability of seeing a positive difference, H_0: $p_+ = \frac{1}{2}$ vs H_1: $p_+ > \frac{1}{2}$, $p_+ < \frac{1}{2}$ or $p_+ \neq \frac{1}{2}$.
- **Assumptions:** X_i are independently sampled from a symmetric distribution.
- **Test statistic:** $T = \#(D_i > 0)$ where $D_i = X_i - \mu$. Under H_0, $T \sim B(n, \frac{1}{2})$ where n is the number of non-zero differences.
- **Observed test statistic:** $t_0 = \#(d_i > 0)$
- **p-value:**
  - H_1: $\mu < \mu_0$: $P(T \leq t_0)$
  - H_1: $\mu > \mu_0$: $P(T \geq t_0)$
  - H_1: $\mu \neq \mu_0$ & $t_0 < \frac{n}{2}$: $2P(T \leq t_0)$
  - H_1: $\mu \neq \mu_0$ & $t_0 > \frac{n}{2}$: $2P(T \geq t_0)$
- **Decision:** If p-value $< \alpha$ there is evidence against H_0.





Sign Test for Paired Data

- We have data of the following form, where $D_i = X_i - Y_i$ for $i = 1, \ldots, n$:

First variable: X	X_1	X_2	$\ldots$	X_n
Second variable: Y	Y_1	Y_2	$\ldots$	Y_n
Difference: D	D_1	D_2	$\ldots$	D_n

- If we do not feel comfortable making the normality assumption when seeing if there is a difference between the distributions of the X and Y populations, we can analyse the differences using the sign test.
- We use the sign of the differences and ignore their magnitude, reducing the test to a test of proportion of positive differences, p_+. Since binomial data only has two possible outcomes, differences of zero are discarded.
- We only need to assume that the differences are independent.

- **Hypothesis:** H_0: $p_+ = \frac{1}{2}$ vs H_1: $p_+ > \frac{1}{2}$
- **Assumptions:** Differences, D_i, are independent.
- **Test statistic:** Let T be the number of positive differences out of the 9 non-zero differences. Under H_0, $T \sim B(9, \frac{1}{2})$. I.e. under H_0, T follows a binomial distribution with $n = 9$ and $p = 0.5$.
- **Observed test statistic:** We observed $t_0 = 6$ positive differences in the sample.
- **p-value:** probability of getting a test statistic as or more extreme than what we observed,

$$\begin{aligned} P(T \geq 6) &= 1 - P(T \leq 5) \\ &= 1 - \texttt{pbinom(5, size = 9, prob = 1/2)} \approx 0.2539 \end{aligned}$$

- **Conclusion:** As the p-value is greater than 0.05, the data are consistent with the null hypothesis at the 5% level of significance. There is no significant difference between the biochemical and the placebo.

Remarks

- The sign test ignores a lot of information in the sample, but it can be applied in quite general situations. We only use the sign of the d_i, and ignore their magnitude in the sign test.
- The sign test does not depend on the distribution of the data. We call these types of tests nonparametric.
- If the normality assumption is satisfied, the t-test is more powerful, in the sense that it will rejec tthe null hypothesis when the alternative hypothesis is true more often than the sign test.
- The sign test can be used to test if a single sample is taken from a continuous distribution that is symmetric about its population mean μ.
- The sign test is more robust i.e. less affected by outlying large or small observations than a t-test.
- If the sample is reasonably believed to come from a normal population, you should use the more powerful t-test instead of a sign test.

Wilcoxon Signed-Rank Test

- The sign test ignores a lot of information (inefficient use of data; low power). While the t-test and Z-test assume a normal distribution, they use all magnitude information from the normal curve.

- The sign test discards all data information on magnitude and hence has low power.

Ranks

- Many non-parametric tests are based not on the data, but on their ranks. To find the ranks for a set of data:
  - Arrange the data in ascending order.
  - Assign a rank of 1 to the smallest observation, 2 to the second smallest etc.
  - For tied observations (in blue or red in the table below), assign each the average of the corresponding ranks.

Sample	8	5	10	2	5	8	8	6
Ordered sample	2	5	5	6	8	8	8	10
Successive ranks	1	2	3	4	5	6	7	8
Assigned ranks	1	2.5	2.5	4	6	6	6	8

- Under the symmetric distribution assumption with mean μ_0 from H_0, half of the $d_i = x_i - \mu_0$ should be negative and half positive, and the expected counts are both $n/2$.
- The positive and negative d_i should be of equal magnitude and occur with equal probability.
- If we rank the absolute values of d_i, in ascending order, the expected rank sums for the negative and positive d_i should be nearly equal.

Wilcoxon Signed-Rank Test

- We define the following quantities:
  - $D_i = X_i - \mu_0$ for $i = 1, 2, \ldots, n$.
  - $R_i, \ldots, R_n$ be the ranks of $|D_1|, |D_2|, \ldots, |D_n|$.
  - W^+ be the sum of the ranks R_i, corresponding to positive D_i.
  - W^- be the sum of the ranks R_i, corresponding to negative D_i.
  - Let $W = min(W^+, W^-)$.
- When we observe the data, we have $d_i = x_i - \mu_0$ with ranks (of the absolute values), $r_1, \ldots, r_n$ for $|d_1|, \ldots, |d_n|$:

$$w^+ = \sum_{i:d_i>0} r_i \text{ and } w^- = \sum_{i:d_i<0} r_i$$

- We should reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu > \mu_0$ if w^+ is large enough, $H_1 : \mu < \mu_0$ if w^+ is small enough, and $H_1 : \mu \neq \mu_0$ if $w = min(w^+, w^-)$ is small enough.

- **Hypothesis:** H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$, $\mu < \mu_0$, $\mu \neq \mu_0$
- **Assumptions:** X_i are independently sampled from a symmetric distribution.
- **Test statistic:** $W^+ = \sum_{i:\, D_i>0} R_i$ for one-sided or $W = \min(W^+, W^-)$ for two-sided
- **Observed test statistic:** w^+ for one-sided or $w = \min(w^+, w^-)$ for two-sided
- **p-value:**
  - $P(W^+ \geq w^+)$ for H_1: $\mu > \mu_0$
  - $P(W^+ \leq w^+)$ for H_1: $\mu < \mu_0$
  - $2P(W^+ \leq w)$ for H_1: $\mu \neq \mu_0$
- **Decision:** If the p-value is less than α, there is evidence against H_0. If p-value is greater than α, the data are consistent with H_0.

- **Hypothesis:** H_0: $\mu_d = 0$ vs H_1: $\mu_d > 0$
- **Assumptions:** D_i are independently sampled from a symmetric distribution.
- **Test statistic:** $W^+ = \sum_{i:D_i>0} R_i$ where R_i are the ranks of $|D_1|, |D_2|, \ldots, |D_n|$. Under H_0, $W \sim \text{WSR}(n)$.
- **Observed test statistic:** $w^+ = 1 + 4 + 6 + 3 + 5 = 19$
- **p-value:**

$$\begin{aligned} P(W^+ \geq w^+) &= P(W^+ \geq 19) \\ &= P(W^+ \leq 6(6+1)/2 - 19) \\ &= P(W^+ \leq 2) = \texttt{psignrank(2, 6)} \\ &= 0.047. \end{aligned}$$

- **Decision:** The p-value is (just) less than 0.05, therefore there is some evidence against the null hypothesis that the diets are are equally effective and we conclude that diet Y does appear to be associated with higher weight gain than diet X.

Calculation of P-value: no ties

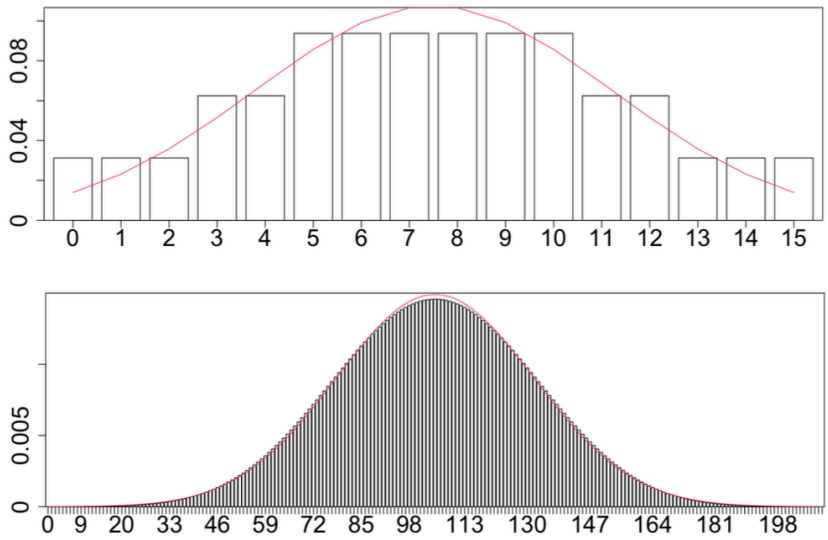
- The exact p-value $P(W^+ \geq w^+)$ for w^+ is: $P(W^+ \geq w^+) = P(W^+ \leq n(n+1)/2 - w^+)$.
- Note that: $W^+ + W^- = 1 + 2 + \ldots + n = \frac{1}{2}n(n+1) \rightarrow W^- = n(n+1)/2 - w^+$. Hence, under the null hypothesis, $E(W^+) = n(1+n)\frac{1}{4}$. You can also show if there are no ties that $Var(W^+) = n(n+1)(2n+1)/24$.

Normal Approximation

- For large enough n, we can use a normal distribution to approximate the distribution of the Wilcoxon sign rank test statistic i.e. in large samples without ties:

$$W^+ \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}), \text{ approximately}$$

- Normal approximation when $n = 5$ and $n = 20$.



- Hence the sample test statistic is:

$$T = \frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}} \sim N(0,1)$$

where $E(W^+) = \frac{n(n+1)}{4}$ and $Var(W^+) = \frac{n(n+1)(2n+1)}{24}$.

- **Hypothesis:** H_0: $\mu = 15$ vs H_1: $\mu \neq 15$
- **Assumptions:** X_i are independently sampled from a symmetric distribution.
- **Test statistic:** $W = \min(W^+, W^-)$ where $W^+ = \sum_{i:D_i>0} R_i$, $W^- = \sum_{i:D_i<0} R_i$, $D_i = X_i - 15$ and R_i are the ranks of $|D_1|, |D_2|, \ldots, |D_n|$. Under H_0, $W^+ \sim \text{WSR}(10)$, a symmetric distribution with mean $\text{E}(W^+) = \frac{n(n+1)}{4} = 27.5$ and $\text{Var}(W^+) = \frac{n(n+1)(2n+1)}{24} = 96.25$.
- **Observed test statistic:** found by
  - Determine difference sample $D_i = X_i - \mu_0$
  - Assign the signed ranks of D_i
  - Calculate w^+, the sum of the positive ranks and w^-, the sum of the negative ranks.
  - We have a two sided alternative, so the observed test statistic is $w = \min(w^+, w^-)$

Normal Approximation with Ties

As we've seen, we can approximate W^+ by a normal distribution, *NOT the data* X_i.

The p-value is approximately given by

$$\text{p-value} \approx P\left(Z \geq \frac{w^+ - \mathrm{E}(W^+)}{\sqrt{\mathrm{Var}(W^+)}}\right) \quad \text{for } H_1\text{: } \mu > \mu_0$$

$$\text{p-value} \approx P\left(Z \leq \frac{w^+ - \mathrm{E}(W^+)}{\sqrt{\mathrm{Var}(W^+)}}\right) \quad \text{for } H_1\text{: } \mu < \mu_0$$

$$\text{p-value} \approx 2P\left(Z \geq \left|\frac{w^+ - \mathrm{E}(W^+)}{\sqrt{\mathrm{Var}(W^+)}}\right|\right) \quad \text{for } H_1\text{: } \mu \neq \mu_0.$$

where **in general**,

$$\mathrm{E}(W^+) = \frac{1}{2}\sum_{i:\, d_i \neq 0} r_i \text{ and } \mathrm{Var}(W^+) = \frac{1}{4}\sum_{i:\, d_i \neq 0} r_i^2$$

Final Notes

- Since we assume that the distribution is symmetric, the hypotheses can also be stated in terms of the median (rather than the mean).
- The p-value from a Wilcoxon signed-rank test will typically be smaller than the p-value of a sign test on the same data. Using the information in the ranks, the test becomes much more powerful in detecting differences from μ_0, and almost as powerful as the one sample t-test.

Wilcoxon Rank-Sum Test

- A non-parametric test to compare means of two independent samples. Relaxes the normality assumption and the assumption of symmetry.
- Suppose that the samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are taken from two distinct populations that follow the same kind of distribution but differ in location i.e. $\mu_x = \mu_y + \theta$.
- Let $R_1, R_2, \ldots, R_N$ with $N = n_x + n + y$ be the ranks of the combined sample: $X_1, X_2, \ldots, X_{n_x}$, $Y_1, Y_2, \ldots, Y_{n_y}$.
    - For one sample Wilcoxon signed-rank test, the ranks are summed over the positive side of the differences.
    - For two sample Wilcoxon rank-sum test, the ranks are summed over one of the samples i.e. $W = R_1 + R_2 + \ldots + R_{n_x}$.
- If H_0 is true, then W should be close to its expected value: $E(W) = \text{Proportion} \times \text{Total rank sum} = \frac{n_x}{N} \times \frac{N(N+1)}{2} = \frac{n_x(N+1)}{2}$.
- If W is small (large), we expect that $\mu_x < \mu_y (\mu_x > \mu_y)$.

- **Hypothesis:** H_0: $\mu_x = \mu_y$ vs H_1: $\mu_x > \mu_y$, $\mu_x < \mu_y$, $\mu_x \neq \mu_y$
- **Assumptions:** X_i and Y_i are independent and follow the same distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_x}$. Under H_0, W follows the $WRS(n_X, n_Y)$ distribution.
- **Observed test statistic:** $w = r_1 + r_2 + \ldots + r_{n_x}$
- **p-value:**

$$P(W \geq w) \text{ for } H_1\text{: } \mu_x > \mu_y \quad \text{or} \quad P(W \leq w) \text{ for } H_1\text{: } \mu_x < \mu_y$$
$$2P(W \geq w) \text{ if } w > \frac{n_x(N+1)}{2} \text{ and } H_1\text{: } \mu_x \neq \mu_y$$
$$2P(W \leq w) \text{ if } w < \frac{n_x(N+1)}{2} \text{ and } H_1\text{: } \mu_x \neq \mu_y$$

- **Decision:** If p-value is less than α, there is evidence against H_0. If the p-value is greater than α, the data are consistent with H_0.

- **Hypothesis:** H_0: $\mu_A = \mu_B$ vs H_1: $\mu_A \neq \mu_B$
- **Assumptions:** A_i and B_i are independent and follow the same kind of distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_A}$. Under H_0, W follows the $WRS(4, 5)$ distribution.
- **Observed test statistic:** $w = 26$ (sum of the ranks associated with method A).
- **P-value:** $2P(W \geq w) = 0.19$ because $w = 26 > \frac{n_A(N+1)}{2} = 20$ so we're looking in the upper tail.
- **Decision:** As the p-value is greater than 0.05, the data are consistent with H_0.

- **Hypothesis:** H_0: $\mu_A = \mu_B$ vs H_1: $\mu_A > \mu_B$
- **Assumptions:** A_i and B_i are indepentent and follow the same kind of distribution but differ by a shift.
- **Test statistic:** $W = R_1 + R_2 + \ldots + R_{n_A}$ (the sum of the ranks of observations in method A). Under H_0, $W \sim \text{WRS}'(13, 8)$, the WRS distribution with sizes 13, 8 and with ties as shown.
- **Observed Test statistic:** $w = r_1 + r_2 + \ldots + r_{n_A} = 180$
- **p-value:** As the exact $\text{WRS}'(13, 8)$ distribution with ties is unknown, we use a normal approximation to this distribution with $\text{E}(W) = \frac{n_x(N+1)}{2} = \frac{13 \times (13+8+1)}{2} = 143$ and $\text{Var}(W) = \frac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^{N} r_i^2 - \frac{N(N+1)^2}{4} \right) = \frac{13(8)(3293.5-2541)}{21(20)} = 186.33$
- $\text{p-value} = P(W \geq w) \simeq P\left(Z \geq \frac{w - \text{E}(W)}{\sqrt{\text{Var}(W)}} \right) = P\left(Z \geq \frac{180-143}{\sqrt{186.33}} \right) = P(Z > 2.7) = 0.003$
- **Decision:** As the p-value is less than 0.05, there is sufficient evidence to reject H_0.

Calculate P-value: No Ties

The exact p-value $P(W \leq w)$ is given by in R by

```
pwilcox(w - minw, m = nx, n = ny)
```

where $\min(W) = \underbrace{1 + 2 + \ldots + n_x}_{n_x} = \frac{n_x(n_x+1)}{2}$ (the smallest possible sum of ranks for the X sample).

The distribution that `pwilcox()` uses is for the distribution of $W - \min(W)$ (and so starts at 0).

Calculate P-value: Ties

- The p-value can be calculated using normal approximation to the distribution of test statistic:

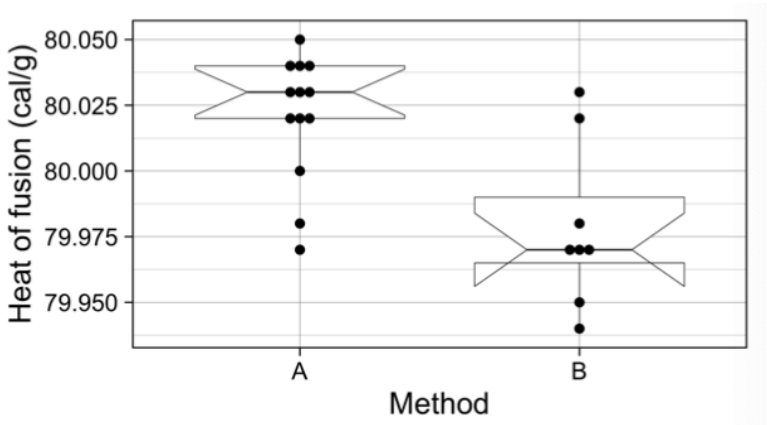
$$T = \frac{W - \mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}} \sim \mathcal{N}(0,1) \text{ approximately,}$$

  where $\mathrm{E}(W) = \frac{n_x(N+1)}{2}$ and $\mathrm{Var}(W) = \frac{n_x n_y}{N(N-1)}\left(\sum_{i=1}^{N} r_i^2 - \frac{N(N+1)^2}{4}\right)$.

- Our p-value calculations are:
  - $\text{p-value} \approx P\left(Z \geq \frac{W-\mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}}\right)$ for H_1: $\mu_x > \mu_y$
  - $\text{p-value} \approx P\left(Z \leq \frac{W-\mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}}\right)$ for H_1: $\mu_x < \mu_y$
  - $\text{p-value} \approx 2P\left(Z \geq \left|\frac{W-\mathrm{E}(W)}{\sqrt{\mathrm{Var}(W)}}\right|\right)$ for H_1: $\mu_x \neq \mu_y$.

A Heuristic for Testing for Differences: Notched Boxplot

- The upper and lower edges of the notches are at: $\text{median} \pm 1.58 \times \frac{\text{IQR}}{\sqrt{n}}$.
- Rule of thumb: if the notches of the two boxes do not overlap, this suggests that the medians are significantly different.



Final Comments

- The Wilcoxon rank-sum test is valid for data from any distribution, and is much less sensitive to outliers than the two sample t-test.
- If one is primarily interested in differences in location between the two distributions, the Wilcoxon test has the advantage of also reacting to other differences between the distributions such as differences in shape.
- When the assumptions of the two-sample t-test hold, the Wilcoxon rank-sum test is somewhat less likely

to detect a location shift than is the two-sample t-test (less powerful).
- In a practical situation in which we are uneasy about the applicability of the two-sample t-test, we use them both and feel happiest when the give similar conclusions.

Permutation Tests

For Continuous Data

Lady Testing Tea

- H_0 : Lady cannot taste the difference vs H_1 : Lady can taste the difference
- Our test statistic is the number of predictions she gets correct,
  $\mathrm{T} = \text{Number of correctly tea before milk cups correctly identified}$

The order of the cups was random, therefore there are

$$8! = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$$

ways to order 8 cups of tea.

There are $\binom{8}{4} = 70$ ways to select which 4 cups of tea had the tea added before milk.

Look at all 70 ways of prediction vs truth and calculate how often we see a test statistic of 0, 1, 2, 3 or 4.

Number correct, t_0	0	1	2	3	4	
Number of ways to select	$\binom{4}{0}\binom{4}{4} = 1$	$\binom{4}{1}\binom{4}{3} = 16$	$\binom{4}{2}\binom{4}{2} = 36$	$\binom{4}{3}\binom{4}{1} = 16$	$\binom{4}{4}\binom{4}{0} = 1$	70
Probability: $P(T = t_0)$	$\frac{1}{70}$	$\frac{16}{70}$	$\frac{36}{70}$	$\frac{16}{70}$	$\frac{1}{70}$	1

$P(T = 4) = \frac{1}{70} = 0.014$ (see Fisher's exact test in Lecture 9).

- We could consider all 40,320 different permutations of the 8 cups of tea.
- Often however, it's not feasible to consider all $n!$ permutations, so we can sample a selection of them.
  - We can then permute the class labels (many times) and see what values we get for the t-test statistic.
  - Then ask, what proportion of test statistics from randomly permuted data are more extreme than the test statistic we observed; this is the permutation test p-value.

Permutation Tests

- The two-sample t-test and the permutation test can sometimes give similar p-values, but this won't always be the case.
- The two-sample t-test is a parametric test where the test statistic is assumed to follow some distribution (a $t_{n_x+n_y-2}$ distribution).
- The permutation test considers the $(n_1 + n_2)!$ permutations of the labels (or a random subset to save computation time) from a single instance of the data (the $n_1 + n_2$ observations).
- The permutation test only assumes that the observations $X_1, X_2, \ldots, X_{n_2}, Y_1, Y_2, \ldots, Y_{n_2}$ are exchangeable, that is, swapping labels on observations keeps the data just as likely as the original.
- The permutation test may use the t-test test statistic but it does not use the t distribution.

Other Permutation Tests Robust to Outliers

- Difference in medians permutation test: $T = \tilde{x} - \tilde{y}$.
- Robustly standardised difference in medians: $T = \frac{\tilde{x}-\tilde{y}}{MAD(x)+MAD(y)}$.

Can we use permutation tests if we are testing for a shift in location by sampling from one population?

- For the Wilcoxon signed-rank test we had

$$T = \sum_{i:\, d_i>0}^{n} r_i \times \text{sign}(d_i)$$

- We could also think of a statistic

$$T = \sum_{i=1}^{n} |d_i| \times \text{sign}(d_i)$$

- For a permutation test permute all possible $\text{sign}(d_i)$.

Bootstrapping

Estimation vs Hypothesis Testing

- Estimation: a population parameter is unknown; uses the sample statistics to generate estimates of the population parameter.
- Hypothesis testing: explicit statement regarding the population parameter; test statistics generated will either support or reject the null hypothesis.

Confidence Intervals

- We should avoid reporting just a point estimate for a sample; always include a measure of variability: $\hat{\theta} \pm \text{margin of error}$, where $\hat{\theta}$ is the point estimate (e.g. sample mean $\bar{X}$).
- The margin of error usually takes the form $\text{margin of error} = \text{critical value} \times SE(\hat{\theta})$ where the critical value is some quantile from an approximate distribution, and $SE(\theta)$ is the standard error of the point estimate e.g. $\sigma/\sqrt{n}$.
- **Definition**: Let $\hat{\theta_L}$ and $\hat{\theta_R}$ be two statistics. If $P(\hat{\theta_L}) \leq \theta \leq \hat{\theta_R} = 1 - \alpha$, then the random interval $[\hat{\theta_L}, \hat{\theta_R}]$ is called a $100(1-\alpha)$ confidence interval for θ, and $100(1-\alpha)$ is the confidence interval level of interval e.g. 95%, 90% where α is 0.05, 0.10.

Let $X_1, X_2, \ldots, X_n$ be a random sample from normal population and $X_i \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 is unknown.

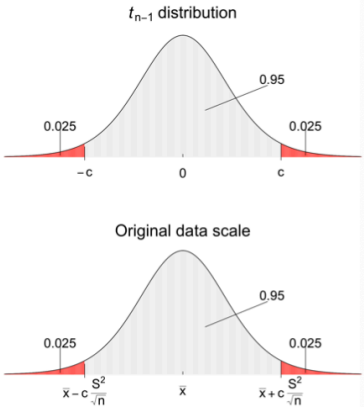
Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

and

$$P\left(-c < \frac{\bar{X} - \mu}{S/\sqrt{n}} < c\right) = 1 - \alpha$$
$$P\left(-cS/\sqrt{n} < \mu - \bar{X} < cS/\sqrt{n}\right) = 1 - \alpha$$
$$P\left(\bar{X} - cS/\sqrt{n} < \mu < \bar{X} + cS/\sqrt{n}\right) = 1 - \alpha$$

In the plots on the right, $\alpha = 0.05$.



Meaning of the Confidence Interval

- Suppose a 95% confidence interval for the mean μ is (a, b).
  - This does not mean that 95% of the means μ are in (a, b), that is $P(a < \mu < b) = 0.95$ since μ is a fixed but unknown parameter.
  - It also does not mean $P(a < \bar{X} < b) = 0.95$ where $\bar{X}$ is the sample mean, since the CI is for the true mean μ, not the sample mean.
- It DOES mean that if we draw a large number of random samples and compute for each sample a 95% CI, about 95% of these CIs will contain μ.
- It can be described as a range of plausible values for the population parameter.

Bootstrap Resampling

- Bootstrapping is a computational process that allows us to make inferences about the population where no information is about the population.
  - "In the absence of any other knowledge about a population, the distribution of values found in a sample of size n from the population is the best guide to the distribution in the population. Therefore, to approximate what would happen in the population was resample, it is sensible to resample the sample (with replacement)."
- Bootstrap methods take their name from the idea of "lifting yourself up by your bootstraps" - moving up without any additional outside help.

Bootstrap Confidence Intervals

- Efron (1979) proposed that the bootstrap confidence interval be the quantiles from the bootstrap distribution.
- In general, (θ_L^*, θ_U^*) are the bounds of the $100(1-\alpha)$ bootstrap CI where θ_L^* is the $\alpha/2$ quantile from the bootstrap distribution and θ_U^* is the $1 - \alpha/2$ quantile from the bootstrap distribution.

Final Remarks

- Bootstrapping is useful when:
  - the theoretical distribution of a statistic is complicated or unknown (e.g. coefficient of variation, quantile regression parameter estimates).
  - the sample size is too small to make any sensible parametric inferences about the parameter.
- Advantages:

- Bootstrapping frees us from making parametric assumptions to carry out inferences.
- Provides answers to problems for which analytic solutions are impossible.
- Can be used to verify, or check the stability of results.
- Asymptotically consistent.

Multiple Testing

What is Real?

- We never really know what a real association is. A small p-value provides some evidence against the null but it could still be a false positive.
- Type 1 error ($\alpha = 0.05$).
- For every model we evaluate at $\alpha = 0.05$, we acknowledge that there is a 5% change that we reject H_0 when H_0 is true.

Types of Errors

- Suppose you are testing a hypothesis that a parameter θ equals zero versus the alternative that it does not equal zero.
  - Type I error or false positive (V): conclude that θ does not equal zero when it does.
  - Type II error or false negative (T): conclude that θ equals zero when it doesn't.
- Error Rates:
  - False Positive Rate: the rate at which null results ($\theta = 0$) are called significant: $E\left[\frac{V}{m_0}\right]$.
  - Family Wise Error Rate (FWER): the probability of at least one false positive: $P(V \geq 1)$.
  - False Discovery Rate (FDR): the rate at which claims of significance are false: $E\left[\frac{V}{R}\right]$.

Possible outcomes from a series of m hypothesis tests.

Truth:	$\theta = 0$	$\theta \neq 0$	Number of tests
Conclusion: $\theta = 0$	U	T	$m - R$
Conclusion: $\theta \neq 0$	V	S	R
Number of tests	m_0	$m - m_0$	m

Accounting for Multiple Testing

- If p-values are correctly calculated calling all p-values less than α significant will control the false positive rate at level α, on average.
- Suppose that you perform 10,000 tests and the reality is that $\theta = 0$ for all of them.
- Suppose that you call all p-values less than 0.05 significant.
- The expected number of false positives is $10000 \times 0.05 = 500$ false positives.
- We can avoid so many false positives by:
  - Controlling the Family-Wise Error Rate (FWER)

- Controlling the False Discovery Rate (FDR)

Controlling the Family-Wise Error Rate (FWER)

- Family Wise Error Rate (FWER): the probability of at least one false positive.
- Let $T_1, \ldots, T_m$ be m test statistics for null hypothesis $H_{01}, .., H_m$.

$$FWER = P(\text{falsely rejecting one or more } \mathrm{H}_{0i})$$
$$P(V \geq 1)$$

- If the null hypothesis is always true but we conduct m tests each at a significance level α then:
  - The probability of at least one false positive is $1 - (1 - \alpha)^m$ e.g. if $m = 20$ then the FWER is 64%.

Bonferroni Correction

- The Bonferroni correction is the oldest multiple testing correction.
- Given that the number of false positives for m tests is $m\alpha$, then consider defining a new threshold for significance: $\alpha^* = \frac{\alpha}{m}$.
- This is conservative (possibly too conservative) but keeps FWER $< \alpha$ e.g. for $m = 20$, $1 - (1 - \alpha^*)^m = 1 - (1 - 0.05/20)^20 = 0.0488$.

Mathematically

Let $p_1, p_2, \ldots, p_m$ be the p-values from m hypothesis tests.

$$\begin{aligned}
\text{FWER} &= P(\text{rejecting at least one true null hypothesis}) \\
&= P(V \geq 1) \\
&= P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \\
&\leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} \\
&= m_0\frac{\alpha}{m} \\
&\leq m\frac{\alpha}{m} \\
&= \alpha.
\end{aligned}$$

Controlling the False Discovery Rate

- Aim: to keep the expected proportion of false positives in your rejected tests (FDR) close to α. Let:
  - R = total number of H_{0i} rejected.
  - V = total number of H_{0i} falsely rejected.
  - $FDR = E\left(\frac{V}{R}\right)$.

Benjamini-Hochberg Procedure

- The most popular correction when performing lots of tests e.g. in genomics, imaging, astronomy, or other signal-processing disciplines.
- Basic idea:
  - Suppose you do m tests.

    - You want to control FDR at level α.
    - Calculate p-values normally.
    - Order the p-values from smallest to largest $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
    - Find $j^* = \max j$ such that $p_{(i)} \leq \frac{j}{m}\alpha$.
    - Reject all H_{0i} where $p_{(i)} \leq \frac{j^*}{m}\alpha$.
- Pros: Pretty easy to calculate and less conservative (maybe much less).
- Cons: Allows for more false positives, may behave strangely under dependence.

ANOVA

Dot Notation

- When working with double subscripts, it is convenient to introduce the dot notation:
    - Replacing either or both subscripts with a dot means adding over those subscripts.
    - Replacing either or both subscripts with a dot and writing a bar over the letter means averaging over those subscripts.
        - For example:
            - total for sample i is $\sum_{j=1}^{n_i} y_{ij} = y_{i\bullet}$
            - average for sample i is $\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i\bullet}$
            - grand total of all observations is $\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij} = y_{\bullet\bullet}$
            - overall average of all observations is $\frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{\bullet\bullet}$
                - here $N = n_1 + \ldots + n_g$ is the total number of observations.
            - Also, $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2$ is the i-th group's sample variance.

ANOVA

- The term ANOVA is an abbreviation of the term "Analysis of Variance".
- The term "variance" as well as the ANOVA procedure is mainly due to FIsher from the 1920's.
- In it's "simplest" form, Analysis of Variance is a generalisation of a two-sided two-sample t-test to 3 or more samples.

General ANOVA Decomposition

- In the case of g groups:

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$ vs H_1: at least one $\mu_i \neq \mu_j$.
2. **Assumptions:** Observations are independent within each of the g samples. Each of the g populations have the same variance, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2 = \sigma$. Each of the g populations are normally distributed (or the sample sizes are large enough such that you can rely on the central limit theorem).
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where g is the number of groups.
4. **Observed test statistic:** t_0.
5. **p-value:** $P(T \geq t_0) = P(F_{g-1,\, N-g} \geq t_0)$. Note: always looking in the upper tail.
6. **Decision:** If the p-value is less than α we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others. If the p-value is larger than α we do not reject the null hypothesis and conclude that there is no significant difference between the population means.

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: at least one $\mu_i \neq \mu_j$ for $i \neq j$.
2. **Assumptions:** Observations are independent within each of the 3 samples. Each of the 3 populations are normally distributed with the common variance σ.
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where $g = 3$ is the number of groups.
4. **Observed test statistic:** $t_0 = \frac{1.88}{0.39} = 4.8$.
5. **p-value:** $P(T \geq 4.8) = P(F_{2,\, 27} \geq 4.8) = 0.0159$. Manually in R: `1-pf(4.8, 2, 27)`
6. **Decision:** As the p-value is less than α we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others.

t-Test Revision

- The Classical two-sample t-test assumes the same as the Welch test with the extra assumption that the two population variances are equal: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.
- Under these conditions $\overline{X} - \overline{Y}$ is normal with variance $\sigma^2(\frac{1}{m} + \frac{1}{n})$ for possibly different sample sizes m and n.
    - σ^2 is estimated using the **pooled variance estimator**

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

(a **weighted average** of the two sample variances) giving a standard error of

$$\text{SE}(\bar{X} - \bar{Y}) = S_p\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

- The test statistic is exactly distributed as t_{m+n-2} under H_0.
- The Classical test is the one that generalises to ANOVA. We must always be aware of these key assumptions:
    - independence between samples;
    - equal variance.

The Normal Model

- We model y_{ij} (for each $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, g$) as the value taken by a random variable: $Y_{ij} \sim N(\mu_i, \sigma^2)$, and that all random variables are independent.
- Thus, we have g different iid samples, the sample for group i (of size n_i) being iid $N(\mu_i, \sigma^2)$.
- In order words, for each $i = 1, 2, \ldots, g$, $Y_{i1}, \ldots, Y_{in_i}$ are iid $N(\mu_i, \sigma^2)$ random variables.
- The "weighted average" decomposition introduced earlier for the two-sample t-test is a special case of a more general decomposition.
- It is most easily explained by considering the so-called Total Sum of Squares:

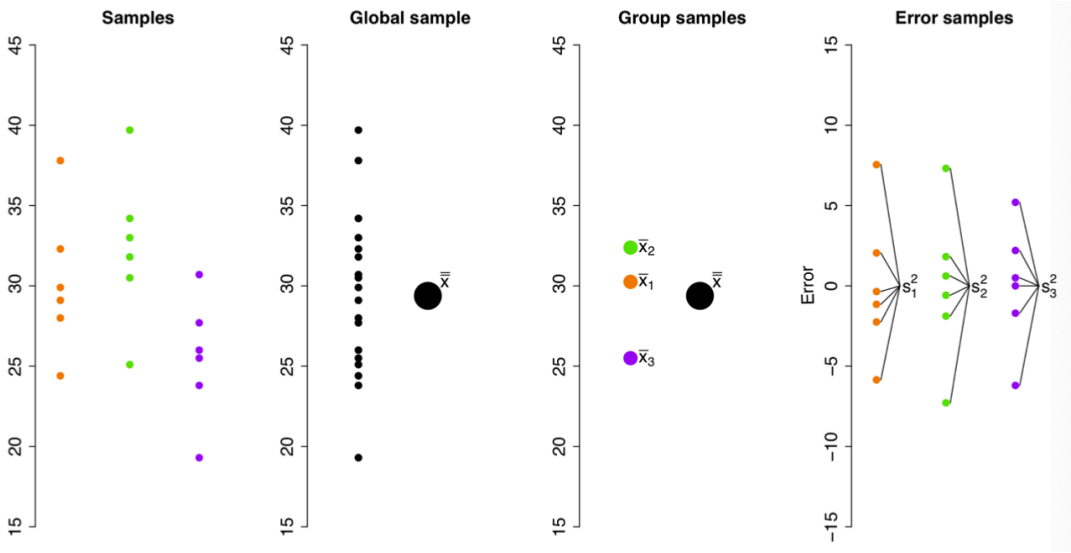
$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{\bullet\bullet})^2$$

- This is precisely $(N-1)$ times the combined sample variance of all the observations:

$$\hat{\sigma}_0^2 = \frac{\text{Total SS}}{N-1} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{\bullet\bullet})^2}{N-1}$$

- We start by adding and subtracting the group means inside the square, grouping and expanding:

$$\begin{aligned}
\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2 &= \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \right]^2 \\
&= \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_{i\bullet})^2 + 2(y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \right] \\
&= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2 + 2 \sum_{i=1}^{g} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \underbrace{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})}_{=0} + \sum_{i=1}^{g} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \underbrace{\sum_{j=1}^{n_i} 1}_{=n_i} \\
&= \underbrace{\sum_{i=1}^{g} \underbrace{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2}_{=(n_i-1)s_i^2}}_{\text{sample variances}} + \underbrace{\sum_{i=1}^{g} n_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2}_{\text{sample means}} \\
&= \text{Residual SS} + \text{Treatment SS}
\end{aligned}$$



Residual Sum of Squares; Residual Mean Square

- The first term, viewed as a random variable under the normal model, can be written as:

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\bullet})^2=\sum_{i=1}^{g}\underbrace{(n_i-1)S_i^2}_{\sim\sigma^2\chi^2_{n_i-1}}\sim\sigma^2\chi^2_{N-g}$$

  - noting that $\sum_{i=1}^{g}(n_i-1)=N-g$. This is called the Residual Sum of Squares.
- Dividing by $N-g$, we obtain an unbiased estimator of σ^2, the generalisation of the pooled estimate of the variance, known as the Residual Mean Square.

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2}{N-g}\sim\left(\frac{\sigma^2}{N-g}\right)\chi^2_{N-g}$$

Treatment Sum of Squares

- The full random variable version of the decomposition looks like:

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2}_{\sim\sigma^2\chi^2_{N-1}\text{ under }H_0}=\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2}_{\sim\sigma^2\chi^2_{N-g}\text{ always}}+\underbrace{\sum_{i=1}^{g}n_i(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}_{\sim\text{ ????}}$$

- When H_0 is true, the final term must have a $\sigma^2\chi^2_{g-1}$ distribution;
  - when the sample sizes $n_1=\ldots=n_g=n$ are equal, this is just $(g-1)$ times the sample variance of the iid normals $\sqrt{n}\overline{Y}_{1\bullet},\ldots,\sqrt{n}\overline{Y}_{g\bullet}$ with variance σ^2, so this is correct in that case; in general this is a bit more complicated though.
- If the true group means are not all equal, this will tend to get bigger. This is the Treatment Sum of Squares.
- The ratio $\frac{\sum_{i=1}^{g}n_i(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})^2}{g-1}$ is the Treatment Mean Square.

- The term Treatment dates back to the beginnings of Analysis of Variance. The Treatment Sum of Squares is the generalisation of the term $\left(\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{1}{m}+\frac{1}{n}}}\right)^2$ in the analysis of the two combined sample variance.
    - It measures the variability of the sample means in a certain sense.

The Ratio of Variances Test

- Continuing the analogy to the two sample t-test, we can consider the ratio of variance estimates as a test statistic to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$.
- The estimate under the null hypothesis is just the combined sample variance: $\hat{\sigma}_0^2 = \frac{\text{Total SS}}{N-1} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{\bullet\bullet})^2}{N-1}$
- The estimate under the alternative or full model is just the Residual Mean Square $\hat{\sigma}^2$.

F Statistic

- It turns out that a sensible test statistic considers the ratio of these two ways of estimating σ^2:

$$\begin{aligned}\frac{\text{Treatment Mean Square}}{\text{Residual Mean Square}} &= \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2/(g-1)}{\hat{\sigma}^2} \\ &= \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2/(g-1)}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2/(N-g)} \\ &\sim \frac{\chi^2_{g-1}/(g-1)}{\chi^2_{N-g}/(N-g)} \ \ (\text{both independent}) \\ &\sim F_{g-1,N-g} \ \text{ under } H_0.\end{aligned}$$

    - the denominator is **always** an unbiased estimator of σ^2 regardless of whether H_0 is true or not
    - the numerator is only an unbiased estimator of σ^2 if H_0 is true, otherwise **it tends to get bigger**.

Contrasts

- If the hypothesis is rejected, further analysis reduces to the study of contrasts.
- A contrast is a linear combination where the coefficients add to zero. In an ANOVA context, a contrast is a linear combination of means.
- We make the distinction between two kinds of contrasts:
    - Population contrasts: contrasts involving the population group means i.e. the μ's.
    - Sample contrasts: contrasts involving the sample group means i.e. the $\bar{y}_{i\bullet}$'s and $\bar{Y}_{i\bullet}$'s.
- For example, we might consider the population contrast $\mu_1 - \mu_2$ whose corresponding observed sample version is $\bar{y}_{1\bullet} - \bar{y}_{2\bullet}$, which is the observed value of the random variable $\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$.

Distribution of Sample Contrasts

- Any $c_1, \ldots, c_g$ with $c_\bullet = \sum_{i=1}^{g} c_i = 0$ defines a sample contrast $\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}$
- Under our normal-with-equal-variances model, this random variable has distribution given by: $\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \sim N(\sum_{i=1}^{g} c_i \mu_i, \sigma^2 \sum_{i=1}^{g} \frac{c_i^2}{n_i})$.

- The corresponding population contrast is the expected value of the random sample contrast.
- Conversely, the observed sample contrast $\sum_{i=1}^{g} c_i \bar{y}_{i\bullet}$ is an estimate of the corresponding population constrast, the random sample contrast $\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}$ is the corresponding estimator.

Behaviour of Contrasts Under the Null Hypothesis

- Under the ANOVA null hypothesis, $H_0 : \mu_1 = \ldots = \mu_g (= \mu \text{ say})$:
  - all population contrasts are zero: $\sum_{i=1}^{g} c_i \mu_i = \sum_{i=1}^{g} c_i \mu = \mu \sum_{i=1}^{g} c_i = 0$
  - all random sampe contrasts have expectation zero: $E(\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}) = \sum_{i=1}^{g} c_i \mu_1 = 0.$
- Therefore, the ANOVA null hypothesis can be rephrased as all population constrasts are zero.

Maybe Not What We Want

- In some examples, in a particular sense, the ANOVA null hypothesis may be too strong.
- We may only wish to test one (or more) 'special' population contrasts are zero.
- Also, the ANOVA null hypothesis may not be rejected for the reason we want.
- Some contrasts may be non-zero, but are they the ones that we are interested in?

t-tests for Individual Contrasts

- Suppose we really only want to test that $H_0 : \sum_{i=1}^{g} c_i \mu_i = 0$ for some 'special contrast' given by $c_1, \ldots, c_g$.
- We can of course preform the ANOVA Mean-Square Ratio F-test, but we can perform a more 'targeted' t-test using the corresponding sample contrast and the residual mean square.
- The corresponding (random) sample contrast:

$$\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \sim N\left(\sum_{i=1}^{g} c_i \mu_i, \sigma^2 \sum_{i=1}^{g} \frac{c_i^2}{n_i} \right).$$

- The standardised version thus has a standard normal distribution:

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} - \sum_{i=1}^{g} c_i \mu_i}{\sigma \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}}$$

- Replacing σ in the denominator with $\hat{\sigma} = \sqrt{\text{ResMS}} \sim \sqrt{\chi^2_{N-g}/(N-g)}$ (independent of the $\bar{Y}_{i\bullet}$'s) gives:

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} - \sum_{i=1}^{g} c_i \mu_i}{\hat{\sigma} \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}} \sim t_{N-g}.$$

- Thus, a t-statistic for testing the hypothesis that $\sum_{i=1}^{g} c_i \mu_i = 0$, which has a t_{N-g} distribution if the contrast equals 0 is:

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}}{\hat{\sigma}\sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}}$$

- This generalises the two-sample t-statistic and may be better than it since it has a smaller standard error with a better estimate of σ.

Confidence Interval

- A confidence interval for a population contrast can be obtained in the usual way, based on the t-statistic.
- Suppose the 'multiplier', or critical value t^* satisfies $P(-t^* \leq t_{N-g} \leq t^*) = 0.95$.
- Then, whatever be the 'true' values of the μ_i's. since the quantity

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} - \sum_{i=1}^{g} c_i \mu_i}{\hat{\sigma}\sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}} \sim t_{N-g}$$

using the usual confidence interval-type manipulations,

$$P\left(\sum_i c_i \bar{Y}_{i\bullet} - t^\star \hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}} \leq \sum_i c_i \mu_i \leq \sum_i c_i \bar{Y}_{i\bullet} + t^\star \hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}}\right) = 0.95\,.$$

- Therefore, for observed sample means $\bar{y}_{1\bullet}, \ldots, \bar{y}_{g\bullet}$, a 95% confidence interval for the true population contrast $\sum_{i=1}^{g} c_i \mu_i$ is given by:

$$\underbrace{\sum_i c_i \bar{y}_{i\bullet}}_{\text{estimate}} \pm t^\star \left(\underbrace{\hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}}}_{\text{st.error}}\right)$$

where, as above, $\hat{\sigma}$ denotes the square root of the residual mean square.

Multiple Comparisons: Simultaneous Confidence Intervals

- **Aside:** normality can be checked with residuals $\rightarrow$ since $\epsilon_{ij} \sim N(0, \sigma^2)$, rather than looking at QQ-plots for each sample, we can instead consider the ANOVA residuals $r_{ij} = y_{ij} - \bar{y_{i\bullet}}$. If the ANOVA assumptions hold true, the residuals should be normally distributed.
- In general, there may be more than one "contrast of interest".
- When no single group is "special" or notable, so that each pairwise difference is equally interesting, we can consider each pairwise difference as a contrast of interest.
- In this case, a t-statistic can be constructed for each pairwise difference; a t-based confidence interval can be constructed for each pairwise "population" difference (contrast).

Individual 95% Confidence Intervals

- We can construct 95% confidence intervals for each pairwise comparison individually: the standard error for $\bar{y}_{i\bullet} - \bar{y}_{h\bullet}$ is $\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_h}}$.
- However, we have constructed each interval without taking any regard of the others.
- More precisely: each interval has been constructed using a procedure so that when the model is correct,

the probability that the "correct" population contrast is covered is 0.95 *individually*.
- We may however want to find what the probability is that all intervals cover their corresponding true values simultaneously.

The Bonferroni Method Revisited

- Let A_1, A_2, A_3 denote the events where each of the 3 intervals above cover the corresponding "true" value.
- Then, under our normal-equal-variance model, we have $P(A_1) = P(A_2) = P(A_3) = 0.95$.
- We then ask what $P(A_1 \cap A_2 \cap A_3)$ is. This is a 'bit hard', but we can derive a lower bound a bit more easily using the relation $(A_1 \cap A_2 \cap A_3)^c = A_1^c \cup A_2^c \cup A_3^c$.
- Recalling that $P(A \cup B) \leq P(A) + P(B)$, we get:

$$\begin{aligned} 1 - P\left(A_1 \cap A_2 \cap A_3\right) = P\left\{\left(A_1 \cap A_2 \cap A_3\right)^c\right\} &= P\left(A_1^c \cup A_2^c \cup A_3^c\right) \\ &\leq P(A_1^c) + P(A_2^c) + P(A_3^c) \\ &= 0.05 + 0.05 + 0.05 = 0.15. \end{aligned}$$

- Therefore, the simultaneous coverage probability of all 3 intervals is at least 85%.
    - By increasing the confidence level of each individual comparison, we are able to make "simultaneous" valid statements about them all.

Summary

- In general, if we have k confidence intervals that we wish to have simultaneous coverage probability of (at least) $100(1-\alpha)\%$, we can achieve this possibly conservatively by constructing each interval to each individual coverage probability $100(1-\alpha/k)\%$.
- If we have g groups, then there are $k = \binom{g}{2} = \frac{g(g-1)}{2}$ possible pairs.
- For moderate-to-large g, this grows quadratically i.e. like g^2.

Tukey's Method

- John Tukey derived the exact multiplier needed for simultaneous confidence intervals for all pairwise comparisons when the sample sizes are equal.
- It was later shown that when sample sizes are unequal, Tukey's procedure is conservative, thus yielding valid simultaneous intervals that may be narrower than those using the Bonferroni method.
- Multiplicity-adjusted p-values can be obtained in the same way by inverting the intervals.
- The "overall ANOVA null hypothesis" can be tested using the smallest of these.
- He named his method "Honest Significant Differences" $\rightarrow$ it is implemented in the function $\mathrm{TukeyHSD()}$, which takes as argument an $\mathrm{aov()}$ fit or using the $\mathrm{emmeans}$ package.

Scheffé's Simultaneous Confidence Interval Method

- If we choose the special multiplier:

$$t^{\star}_{\mathrm{Sch}}(\alpha) = \sqrt{(g-1)F_{g-1,N-g}(\alpha)} = \sqrt{(g-1)\texttt{*qf(1}-\alpha\texttt{,}g-1\texttt{,}N-g\texttt{)}}$$

and construct simultaneous confidence intervals for all possible contrasts according to:

$$\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \pm t^{\star}_{\mathrm{Sch}}(\alpha)\, \hat{\sigma} \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}$$

then the probability that all sample contrasts include their true population values is exactly $1-\alpha$.

- We effectively compare each contrast t-statistic to the $\sqrt{(g-1)F}$ distribution.
- Any which exceeds the critical value is significant in the 'simultaneous' sense. The smallest such p-value is the F test p-value.

Summary of ANOVA Post Hoc Tests

- The ANOVA F-test alone may or may not address the important scientific questions in each example.
- Depnding on the context, a test based on the most significant contrast(s) may be more useful than a straight F-test.
- Bonferroni procedures are in general conservative i.e. p-values and confidence intervals may be larger than they really need to be.
    - Alternative methods which may be more accurate i.e. less conservative exist: e.g. Tukey's method.
- Any contrasts must be decided upon before looking at the data. Otherwise we are data snooping.
- If we 'snoop' until we find a significant contrast, we must take account of that:
    - Scheffé's method permits unlimited data snooping.
    - If we snoop only across k fixed contrasts e.g. all pairwise comparisons, we can use the Bonferroni method to adjust for that (but for large k, Tukey's method or Scheffé's method may give smaller intervals).

When ANOVA Assumptions Might Be Violated

- Underlying all tests, contrasts and comparisons are the assumptions that:
    - each sample is from a normal population;
    - all population variances are equal so all populations are identical up to possible location shifts.
- There are possible ways the assumptions might be violated:
    - the normality might be ok, but equal variances might not be;
    - the normality might not be ok, but the identical up to location shifts assumption might be ok.
- There are a few tools we can appeal to:
    - simulation
    - resampling, together with conditioning

Relaxing the Equal Variance Assumption: All Pairwise Comparisons

- We could consider assuming normality, but dropping the "common variance" assumption.
- A simple way to do so is to consider all pairwise Welch tests, and apply a Bonferroni correction.
- Recall Welch tests only assume that each sample is normal, with possibly different variances σ_X^2 and σ_Y^2 and different means, and all random variables are independent.

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}},$$

which is approximately $t_{d^*(m,n,\sigma_X,\sigma_Y)}$ under H_0 for a known function $d^*(\dots)$.

- We multiply our pairwise comparisons by g to get the 'adjusted for multiplicity' p-values, the smallest of which can be used as a test that all population means are equal.

Relaxing the Normality Assumption

- Under the formal ANOVA assumptions, each population is normal and all variances are the same, such that the null hypothesis reduces to: All observations come from the **same normal** distribution.
- A weaker set of assumptions at least under the null hypothesis is that: All observations come from the **same** distribution.

The Tool of Conditioning

- A common tool in testing is to condition on an "ancillary" statistic:
  - "Ancillary statistic" just means a statistic that does not tell us anything useful.
- A familiar example is the sign test:
- We usually condition on the number of non-zeroes (i.e. we ignore the ties).
- Then, p-values are in fact conditional probabilities, e.g. for a one-sided sign test based on the number of positive signs, the p-value is:

$$P(S \geq s \mid N = n) = P(B(n, 0.5) \geq s)$$

Conditioning on the Combined Sample

- If we combine all the groups into one combined sample (i.e. throw away the labels) then the remaining "data" tells us nothing about differences between groups i.e. what we are interested in.
- In this sense, the combined sample is an "ancillary statistic".
- Once we condition on the combined sample, the only remaining "randomness" is the allocation of observations to groups.
- Under the null hypothesis of "no differences between groups" all possible allocations are equally likely.

Enumerating All Possible Allocations: Exact p-values

- We can (in principle) compute an exact conditional p-value for **any "sensible" statistic** under this particular null hypothesis.
- There are actually

$$\frac{N!}{n_1!n_2!\dots n_g!}$$

different possible allocations of the N total observations into groups of size $n_1, n_2, \dots, n_g$.

- We can (in principle) compute the value of the statistic under each possible allocation.
- Since each such value is equally likely under the null hypothesis, we can use this "sampling distribution" to compute a p-value.
- Suppose the statistic is T, the observed value is t_0, and larger values indicate more evidence against the null hypothesis. The exact conditional p-value is a simple proportion:

$$P(T \geq t_0 \mid \text{combined sample}) = \frac{\text{no. allocations with } T \geq t_0}{\text{total no. allocations}}.$$

- Unfortunately, unless the sample sizes are very small, the total number of allocations is MASSIVE and computing the value of the statistic over all possible allocations is not feasible.
- Fortunately, we can estimate this proportion by taking a suciently large random sample from the "population of all possible allocations":
    - this is a binomial/hypergeometric (depending on whether we sample with or without replacement) proportion estimation problem.

Permutation Tests

- In R, if the data is represented as a data frame with:
    - observations in one column and
    - groups indicated by a factor in another column

  then it is easy to obtain a random "allocation":
    - simply randomly permute the observation vector, keeping the factor vector fixed.
- Do this a large number of times.
- The "observed proportion" of the times the statistic exceeds t_0 becomes an estimate of the "exact" p-value.
- This general procedure is known as a permutation test.

Using Ranks: Kruskal-Wallis Test

1. **Hypotheses:** H_0: the response variable is distributed identically for all groups vs H_1: the response variable is systematically higher for at least one group
2. **Assumptions:** Observations are independent within each group and groups are independent of each other. The different groups follow the same distribution (differing only by the location parameter).
3. **Test statistic:** like ANOVA applied to the ranks (not examinable), see here for details. Under the null hypothesis the Kruskal-Wallis test statistic approximately follows a χ^2 distribution with $g-1$ degrees of freedom where g is the number of groups.
4. **p-value:** $P(T \geq t_0) = P(\chi^2_{g-1} \geq t_0)$.
5. **Decision:** If the p-value is less than α we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others. If the p-value is larger than α we do not reject the null hypothesis and conclude that there is no significant difference between the population means.

- This is performed by:
    - replacing each observation by its 'global' rank;
    - then computing the F-ratio as usual on the ranks.
- A p-value can be obtained:
    - using a permutation test approach or
    - a large sample χ^2 approximation can also be used.

Kruskal-Wallis Test Statistic

- The traditional approach to the Kruskal-Wallis test uses a test statistic that is computed as a ratio (like the F-test).
    - the numerator is exactly the Treatment Sum of Squares of the ranks
    - the denominator is the sample variance of all the ranks.
        - this denominator is not random (it is the same regardless of the allocation).

$$T = \frac{\text{Treatment SS of the ranks}}{\text{Varance of all the ranks}}$$

- When H_0 is true, it has an approximate χ^2_{g-1} distribution.

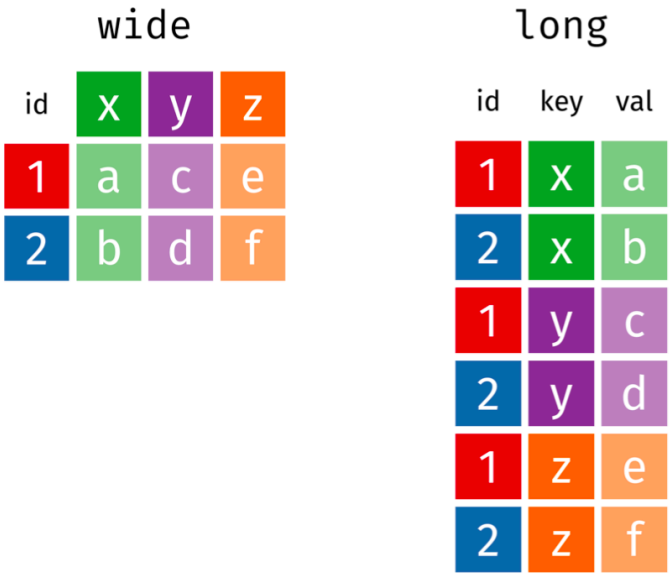
Permuted Ranks

- The permutation test approach is valid for any "sensible" statistic;
    - it only assumes the same distribution in each group under the null hypothesis.
- What of the 'sensible statistic'?
    - If the data are truly normal, the F-statistic makes sense;
    - Is it still "sensible" if the normality assumption is being relaxed?
    - Could also do a permutation test using the Kruskal-Wallis statistic.

Two-Way ANOVA: Adjusting for Blocks

Long vs Wide Format

- To analyse example data in R, we need it in "long" format i.e. we want a dataframe with 3 columns.
    - one with the response
    - a factor indicating "treatment" (e.g. electrode type)
    - another factor indicating the Subject.



Formulas in R

- The basic structure of a formula in R is:

```
y ~ x1 (+ x2 + x3) # y is a function of x, y against x, y grouped by x
```

Adjusting for Two Variables

```
summary(fit1) # y ~ electrode
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## electrode    4   5.09  1.2719   1.503   0.21
## Residuals   75  63.48  0.8464
```

```
summary(fit2) # y ~ Subject + electrode
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Subject     15  33.27  2.2180   4.405 1.77e-05
## electrode    4   5.09  1.2719   2.526     0.05
## Residuals   60  30.21  0.5036
```

- The Residual sum of squares for `fit1` (63.48, on 75df) is being decomposed into two pieces:
  - the fit2 Subject sum of squares (33.27, on 15 df) and
  - the fit2 Residual sum of squares (30.21, on 60 df)
- For fit2, the Residual sum of squares is much smaller that for fit1, but the degrees of freedom is only a little less:
  - this gives a much smaller Residual Mean Square (0.5036, compared to 0.8464 for fit1); this in turn gives a bigger (treatment-to-residual) -ratio (2.526, compared to 1.503 for fit1);
  - crucially the p-value has been reduced from 0.21 to 0.05:
    - the effect is now (at least mildly) significant!

Changing Parameters

- For ordinary one-way ANOVA, we have written the model as:
  - for $i = 1, \ldots, g, j = 1, \ldots, n_i, Y_{ij} \sim N(\mu_i, \sigma^2)$.
  - for $i = 1, \ldots, g, j = 1, \ldots, n_i, Y_{ij} = \mu_i + \epsilon_{ij}$ where the ϵ_{ij}'s are iid $N(0, \sigma^2)$.
    - Both of these have g unknown mean-parameters, and 1 unknown variance parameter.
- A third way to write the model is based on expressing each μ_i as $\mu_i = \mu + \alpha_i$:
  - an overall mean μ with no subscript plus
  - an adjustment α_i for i-th level of the treatment
- This leads to the model for $i = 1, \ldots, g, j = 1, \ldots, n_i$: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$.
  - Note that there are now $g + 1$ mean parameters (sort of): $\mu, \alpha_1, \ldots, \alpha_g$ i.e. we have creater another parameter.

Extra Constraint

- In fact, depending on hw μ is defined, the α_i's necessarily obey a certain constraint.
- The overall mean is defined as some kind of (weighted) average of the μ_i's: $\mu = \sum_{i=1}^{g} w_i \mu_i$. Then each $\alpha_i = \mu_i - \mu$.
- Necessarily, the same weighted average of the α_i's is:

$$\sum_{i=1}^{g} w_i \alpha_i = \sum_{i=1}^{g} w_i (\mu_i - \mu) = (\sum_{i=1}^{g} w_i \mu_i) - \mu \sum_{i=1}^{g} = \mu - \mu = 0$$

  - In fact, knowing $g-1$ of the α_i's means you also know the final one.

Estimating These New "Parameters"

- A common choice for the weighted average is:

$$\mu = \frac{1}{N} \sum_{i=1}^{g} n_i \mu_i = \frac{\sum_{i=1}^{g} n_i \mu_i}{\sum_{i=1}^{g} n_i}$$

  which is the expectation of the grand mean $\bar{Y}_{\bullet\bullet}$.
- This can be estimated using the observed grand mean $\bar{y}_{\bullet\bullet}$.
- Each α_i represents the difference betwen each group mean and the overall mean, it's thus naturally stimated using the difference $\hat{\alpha}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$.

The Two-Way ANOVA Model

- The model we shall fit to the electrode data is the following:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where

$$\mu = \text{overall mean}$$
$$\alpha_i = \text{adjustment for treatment level } i \text{ for } i = 1, 2, \ldots, g$$
$$\beta_j = \text{adjustment for block } j \text{ for } j = 1, 2, \ldots, n$$

and n is the common sample (block) size and the ϵ_{ij}'s are iid $N(0, \sigma^2)$.

- So each Y_{ij} has a possibly different expectation $\mu_{ij} = \mu + \alpha_i + \beta_j$, but these have an additive structure:
  - the ng different means are explained by $1 + (g-1) + (n-1) = g + n - 1$ free parameters.
- All random variables are independent and the following constraints are satisfied:

$$\sum_{i=1}^{g} \alpha_i = 0 \text{ and } \sum_{j=1}^{g} \beta_j = 0$$

Estimating Parameters

- As all 'sample sizes' are the same, the overall mean be thought of as just the mean of the μ_i's. It is naturally estimated using the overall mean $\bar{y}_{\bullet\bullet}$.
- Also, each α_i, the "adjustment" for electrode type i, is naturally estimated using the difference $\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$
- Similarly, each β_j, the adjustment for subject j is naturally estimaetd using the difference $\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}$

The Two-Way Decomposition

- Each observation therefore, may be notionally split up into 4 pieces:

$$y_{ij} = \underbrace{\bar{y}_{\bullet\bullet}}_{\hat{\mu}} + \underbrace{(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})}_{\hat{\alpha}_i} + \underbrace{(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})}_{\hat{\beta}_j} + \underbrace{(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})}_{\hat{\varepsilon}_{ij}}$$

- The final part $\hat{\epsilon}_{ij}$ is the (i, j)-th residual or estimated error. We can analyse the variance here in the same way as the ordinary "one way" ANOVA model.

Decomposing the Total Sum of Squares

$$\begin{aligned}
\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n}\left\{(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})\right\}^2 \\
&= \sum_{i=1}^{g} n(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n} g(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})^2 \\
&\qquad + \text{cross-product terms which are all zero} \\
&= \text{Treatment sum of squares} + \text{Block sum of squares} + \text{Residual sum of squares}
\end{aligned}$$

Two-Way ANOVA Table

Source of Variation	Sum of squares	df	Mean square	F-ratio
Blocks	Block Sum Sq.	$n-1$		
Treatments	Trt Sum Sq.	$g-1$	$\text{Trt MS} = \dfrac{\text{Trt Sum Sq.}}{g-1}$	$\dfrac{\text{Trt MS}}{\text{Res MS}}$
Residual	Res Sum Sq.	$(n-1)(g-1)$	$\text{Res MS} = \dfrac{\text{Res Sum Sq.}}{(n-1)(g-1)}$	
Total	Total Sum Sq.	$ng-1$		

- The total sum of squares is $\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$, and the total sample sie is $N = ng$.

The Purpose of Blocking

- A two-way ANOVA with blocking can be thought of as a generalisation of the paired t-test where each pair is a block.
- In the paired t-test, the idea is to remove the variation between pairs to more accurately compare the two treatment levels within each pair.

    - The "within pair" difference is then averaged over all pairs to get the "treatment effect".
- We are not interested in the testing for a Block effect, we are only interested in comparing treatments.
- We are nonetheless adjusting for Blocks, to more accurately compare Treatments.
- Although the treatment sum of squares and block sum of squares are mathematically identical, they are playing very different scientific roles.

Blocking in the Electrode Experiment

- We have identified some systematic variation which can be attributed to the differences between Subjects (assuming these contribute additively).
- The term "Block" comes from Fisher's agricultural trials, where he adjusted for variation between different blocks of land, in order to compare the fertiliser Treatments more accurately.
- The net result is that we have a smaller (more precise) estimate of the error variance as we have explained an extra part of the variation and removed it from the residual sum of squares.

Averages

- The overall, treatment level and block averages are therefore (due to the constraints):

$$\bar{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n}\left(\mu + \alpha_i + \beta_j + \varepsilon_{ij}\right) = \mu + \alpha_i + \bar{\varepsilon}_{i\bullet} \quad (\text{free of the } \beta_j\text{'s!})$$

$$\bar{Y}_{\bullet j} = \frac{1}{g}\sum_{i=1}^{g}\left(\mu + \alpha_i + \beta_j + \varepsilon_{ij}\right) = \mu + \beta_j + \bar{\varepsilon}_{\bullet j} \quad (\text{free of the } \alpha_i\text{'s!})$$

$$\bar{Y}_{\bullet\bullet} = \frac{1}{ng}\sum_{i=1}^{g}\sum_{j=1}^{n}\left(\mu + \alpha_i + \beta_j + \varepsilon_{ij}\right) = \mu + \bar{\varepsilon}_{\bullet\bullet}$$

Treatment Sum of Squares

- The treatment sum of squares is:

$$\sum_{i=1}^{g} n(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = n\sum_{i=1}^{g}(\alpha_i + \bar{\epsilon}_{i\bullet} - \bar{\epsilon}_{\bullet\bullet})^2$$

- Under the null hypothesis, this is:

$$\underbrace{n\sum_{i=1}^{g}(\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet})^2}_{\sim \frac{\sigma^2}{n}\chi^2_{g-1}} \sim n\left(\frac{\sigma^2}{n}\chi^2_{g-1}\right) \sim \sigma^2\chi^2_{g-1}$$

- This is because, under the model, the $\bar{\epsilon}_{i\bullet}$'s are iid normal with variance σ^2/n. This is the same for the one-way ANOVA.

Residual Sum of Squares

- The (i, j)-th residual is:

$$Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet} = \ldots = \varepsilon_{ij} - \bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet j} + \bar{\varepsilon}_{\bullet\bullet}.$$

- How is the residual sum of squares distributed? We have the identity with $N = ng$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})^2}_{\sim\sigma^2\chi^2_{N-g}} = \underbrace{\sum_{j=1}^{n} g(\bar{\varepsilon}_{\bullet j} - \bar{\varepsilon}_{\bullet\bullet})^2}_{\sim\sigma^2\chi^2_{n-1}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n}(\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet j} + \bar{\varepsilon}_{\bullet\bullet})^2}_{\sim???}$$

  - Roughly speaking, this is:

$$\text{One-way Res Sum Sq.} = \text{Block Sum Sq. of Errors} + \text{Two-way Res Sum Sq.}$$

- It can be shown that the two terms on the RHS are independent, so the last double sum must be $\sigma^2\chi^2_{N-g-(n-1)} \sim \sigma^2\chi^2_{(n-1)(g-1)}$

Two-Way ANOVA F-Ratio

- In summary:
  - the residual sum of squares always follows a $\sigma^2\chi^2_{(n-1)(g-1)}$ distribution (regardless of whether the null hypothesis is true or not);
  - if the null hypothesis of "no treatment effect" is true, the treatment sum of squares follows a $\sigma^2\chi^2_{g-1}$ distribution.
- There, if the null hypothesis is true, the F-ratio is:

$$\frac{\text{Treatment mean square}}{\text{Residual mean square}} \sim \frac{\chi^2_{g-1}/(g-1)}{\chi^2_{(n-1)(g-1)}/(n-1)(g-1)} \sim F_{g-1,\,(n-1)(g-1)}$$

- Otherwise, it tends to take larger values as is the case for the one-way ANOVA.