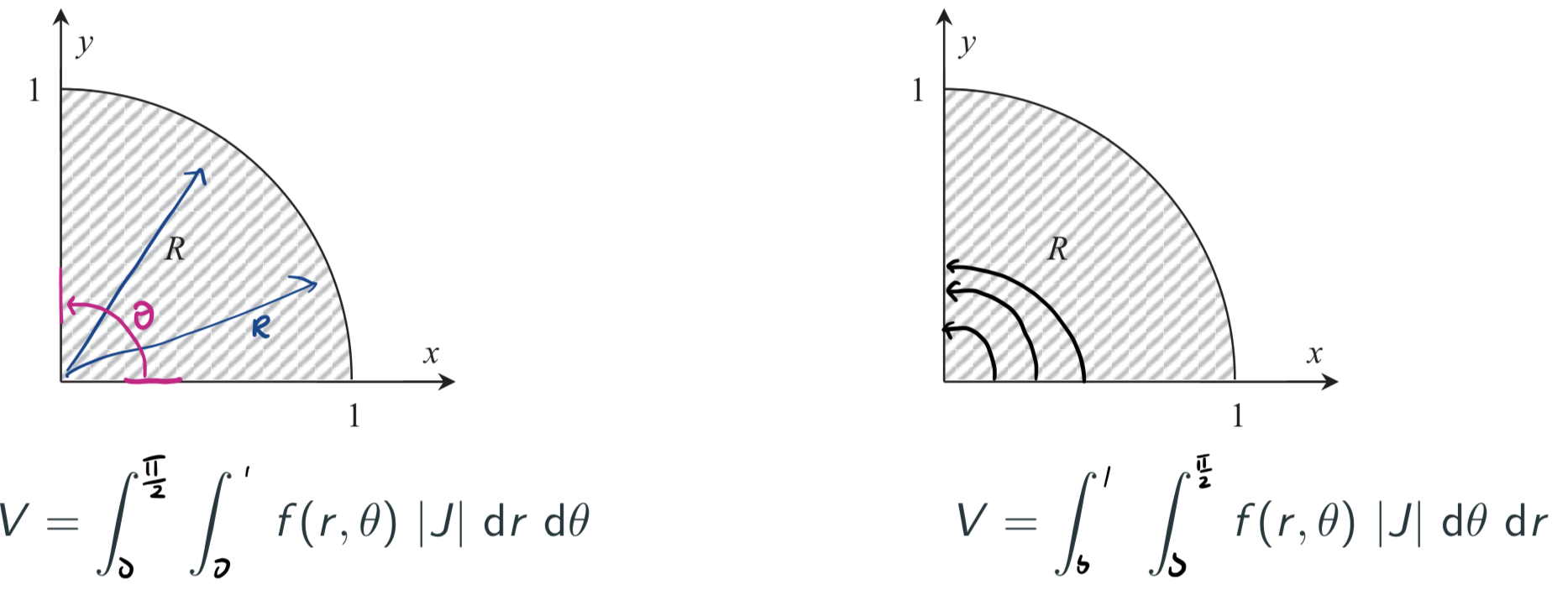
# Test 2 Cheatsheet

## Change of variables

### Polar Variables

The most common change of variables is cartesian to polar in the form of.  
  
where is the absolute value of the *Jacobian* - A stretching/scaling factor that accounts for the differences between coordinate systems. It is the determinant of the matrix of first-order partial derivatives.

Converting to polar: and Here the Jacobian is

Process:

1. Identify an appropriate coordinate system.
2. Find the Jacobian for the proposed coordinate system.
3. Substitute the new variables into the integral
4. Draw the integration region.
5. Draw the strips in the direction of the inner integral to find the inner limits.
6. Add strips across the region R to find the outer limit.

|  |  |
| --- | --- |
| Eq | img |
|  | center |

##### Less obvious Coordinate systems

Look at the functions boundaries to identify substitutions that give constant integral bounds in the new coordinate system.

Define u, v

Rewrite in terms of x and y

Solve Jacobian

Substitute all variables into original equation

#### Data Analysis

###### Block 1: Statistical Inference

**Exploratory analysis** – Centre, Spread and Skew

**Core definitions and concepts**

* **P-Value** - The probability that we observe a test statistic as least as unusual as the one we have from our sample, given that the null hypothesis is true.
* **Standard Error** - the standard error (of the sample mean) is an estimate of the sample-to-sample variability of the sample means
* **Confidence Interval** - under repeated sampling, x% of such intervals will contain the true population mean.
* **Quantile** - A x% quantile gives the value of the data, y, where x% of the data (or probability density) is less than or equal to y
* **Null** and **alternative Hypotheses**. Nothing, not nothing….?
* **CLT** the distribution of the sample means, , is approximately Normal even if the distribution we are sampling from isn’t.

**Model formulation:**

* One-sample - where
* Paired sample - where
* Two sample (effects) where
* Linear Regression where
* Multiplicative one sample
* Multiplicative Two sample
* Multiplicative Regression
* Categorical:

**Assumptions:**

* One-sample / Paired sample – Independence, Normality
* Two-sample t-tests - Independence, Normality, EoV
* Linear Regression Independence, Normality, EoV

**Null hypotheses:**

* One-sample
* Paired sample – No change,
* Two-sample t-tests - No difference,
* Regression: lm(yvar ~ 1)
* ANOVA:

**Dealing With assumptions**

* Independence
  + Between groups / observations
  + Only satisfied with good experimental design
* Normality
  + All groups of observations have a normal distribution.
  + Check on Q-Q plot/Bell curve (modcheck())
  + To fix: Invoke CLT, If enough observations
* Equality of Variance (EoV)
  + All residuals have equal variance.
  + Check with eovcheck()
  + To Fix: in t.test use var.equal = FALSE – welch T test

**Diagnostic Graphs:**

**Q-Q:** Made by plotting the sample quantiles against quantiles from a theoretical normal distribution of

###### Block 2: Simple Linear Regression

**Linear models** in terms of an algebraic expression () and in terms of R syntax (y ∼ x) – (y can be explained by x) the coefficients are the values that minimize the least square equation the most.

**Estimated Terms:**

**Residuals** are the difference between the estimated value for a datapoint and the actual value.

**Least Square** is the sum of all data points

• Understand what the residual of an observation is, and how this relates to least squares.

**Cooks Distance** A measure of how much an observation influences the LM. On a cook’s distance plot an observation is influential if it is >0.4 and changed coefficients by >= 1 SE.

**Interpreting** The null hypothesis we assume . So if we disprove it we conclude that it is unreasonable to assume that X had no effect on Y. Null model is the same as Y ~ 1 and t.test of this gives the same result as summary().  
**Interpreting confidence intervals** always “on average” state as a range “On average every increase in a student’s test mark results in a 3.3 to 4.3 increase in exam result”  
**The statistic**: From the residual standard error of the null minus that of the model divided by that of the null. It is a measure of how much of the variability is explained by the model. If it is greater than 80% and all other assumptions are met, the model is probably good for prediction.

**Prediction:** Prediction for individual values (Estimate the exam mark for a single student) and use confidence intervals for average values (Average exam mark for students with a given test mark) Prediction interval allows for uncertainty due to random scatter about the line. While both allow for uncertainty about the coefficients. A prediction should not be carried out if there are not sufficient datapoints and especially not into the “future”.

Block 3: Multiplicative Models  
A multiplicative model should be fitted when the data is severely right skewed. Or there is a multiplicative relationship between the variables (Depreciation, growth, etc).

This process effectively lets us use the median instead of the mean. Logging the data preserves the order, quartiles, but not the mean. This means that the mean is shifted more in line with the median of the data.

**Multiplicative null model** lm(log(price) ~ 1), data = mazda.df).

**Confidence intervals** exp(confint(loggedData.fit)

**Two Sample** t.test(log(price) ~ brand, data = mazda.df))  
When we back transform, we get a relationship between the medians from the log rules. So, a confidence interval should be interpreted as a change (-0.3, -0.7) would represent group one being worth 30% to 70% the value of group 2. Percentages got with formula

**Multiplicative Regression** lm(log(price) ~ age)) gives us the model:

This turns into so for every year the car ages, it is coefficient of the previous year.

**Confidence Intervals** exp(confint(model.fit)) Numbers can be read straight off as with regular regression. Coefficients can be read off of exp(summary(model.fit)). Gives **MEDIAN** instead of average.

**Summary** Only Regression coefficients and comparisons are multiplicative. But not predictions etc.   
Block 4: Categorical Models / One-way ANOVA  
Comparing the differences between multiple groups. Have to set the column as a factor() in R.

In the model, are dummy variables, you need n-1, one for every element that is not the baseline.

Coefficients from summary(lm(score ~ factor). Confidence intervals from confint(model.fit).

Summary() only gives some comparisons but not all. we cant compare this can lead to missing significant comparisons. Factor rotations are used to solve this. data.df$f2 = factor(data.df$f1, levels = c(‘2’, ‘1’, ‘3’)) to swap factors 1 and 2. This means summary() will use factor 2 as the baseline and will give comparisons and p-values accordingly.

**ANOVA** uses effects model (two summary). Compares the between group variation with the within-group variation to assess whether there is a difference in the population means. It is looking for large variation between groups compared to within groups. P value is the “Pr(>/< F)” Field. Says “It explains a significant amount of the variation in score, compared to fitting the grand mean alone.”

**Variability** the amount of variability explained by the categorical variable selfassess is 7633.55. The residual or left over variability within the groups is 3181.61. The total variability is 10815.15. The amount of variability explained by selfassess is 7633.55/10815.15 = 70.58%.

**Multiple Comparison Problem** Lots of samples means more likely to breach the 5% CI. Tukey intervals expand these to reduce the problem.

Multiplecomp(model.fit)  
Block 5: Introduction to Machine Learning   
• Describe the differences in how model quality is measured in statistics and in machine learning.  
• Explain how an estimate of model quality in ML can be obtained, both in terms of how errors are  
measured, and how it is done procedurally.  
• What is cross-validation? Why do we care about this?  
• Understand what the following concepts mean, and be able to discuss this in the context of applications:  
– Feature and Target Engineering  
– Bias-variance trade-off  
– Missing data and possible effects  
– Class imbalance  
– Data bias  
– Historical bias and ethics  
Linear Algebra