Applied Statistical Methods II

Single Factor Studies

Looking forward:

- Look at single-factor ANOVA.
  - Also known as one-way ANOVA.
  - Look at its formulations and uses.
  - Derive the properties of sums-of-squares.

One-Way ANOVA Setting

- Consider an example where we have r different treatment groups.
  - One factor with r levels.
- n_i subjects are given the i^{th} treatment $i = 1, \ldots, r$.
  - $n_T = \sum n_i$
- We observe Y_{ij} for $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$.
  - The observation for the j^{th} replicate for the i^{th} treatment.
- The one-way ANOVA model are used to assess the effect of different treatments.

Types of Data

- The data can come from either a designed experiment or an observational study.
- For a designed experiment:
  - Should be well designed so that potential confounders have equal distributions across factor levels.
- For an observational study:
  - We will only draw inference on the factor of interest.
  - Will not control for possible confounders.
  - No real type of causal relationship can be speculated.

Example of a Designed Experiment

- You are working for a food company.
- You have 4 different package designs for a box of cereal. Which design leads to the most sales?
  - Factor = package designs, levels = 4.
- You have 20 stores with equal sales volumes that want to participate. Each store gets a different package design. What are exp. units?
- You create a balanced design and randomly assign a package design to each store. One store has a fire.
  - n=19 total experimental units, $n_i = 5, 5, 4, 5$.
- Measure the number of boxes sold in a week.
- Design of the study helps eliminate confounders:
  - Selected stores with equal sales volumes.
  - Randomly assigned packages to stores.
  - Would also want to control other variables such as display location.

Example of a Designed Experiment

- You are working for a food company.
- You have 4 different package designs for a box of cereal. Which design leads to the most sales?
  - Factor = package designs, levels = 4.
- You have 20 stores with equal sales volumes that want to participate. Each store gets a different package design. What are exp. units?
- You create a balanced design and randomly assign a package design to each store. One store has a fire.
  - n=19 total experimental units, $n_i = 5, 5, 4, 5$.
- Measure the number of boxes sold in a week.
- Design of the study helps eliminate confounders:
  - Selected stores with equal sales volumes.
  - Randomly assigned packages to stores.
  - Would also want to control other variables such as display location.

Observational Example

- You want to know if four ball bearing machines in a plant generate a product with the same diameter.
- You take a sample of 10 ball bearings produced by each machine.
- You have no control over some factors.
  - Who operated the machine.
  - What was the temperature near the machine when these were made.
- Any possible confounding effects can not be determined.

The one-way ANOVA model

- The one-way ANOVA model:
  $Y_{ij} = \mu_i + \epsilon_{ij}$.
- $E(\epsilon_{ij}) = 0$, $\text{Var}(\epsilon_{ij}) = \sigma^2$.
- $E(Y_{ij}) = \mu_i$.
- Can put it in a linear model framework $Y = X\beta + \epsilon$.
  $Y = [Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{r1}, \ldots, Y_{rn_r}]'$
  $\epsilon = [\epsilon_{11}, \ldots, \epsilon_{1n_1}, \ldots, \epsilon_{r1}, \ldots, \epsilon_{rn_r}]'$
  $\beta = [\mu_1, \ldots, \mu_r]'$
  X is a $n_T \times r$ matrix. What does it look like?

Computing

You learned all of the theory for ANOVA last semester (F-test)

- R: Can use `aov`. Example with a factor variable: `aov(y ~ factor, data=Data)`. This is valid for balanced and unbalanced designs.
- SAS: PROC GLM or PROC ANOVA

ANOVA: more intuition and insights

As a special but very popular design in the general linear models framework, we will look at some details of

- Estimation and Inference
- Power calculation
- Permutation test

Some Notation

$$
\begin{aligned}
Y_{i.} &= \sum_{j=1}^{n_i} Y_{ij} \\
\overline{Y}_{i.} &= Y_{i.}/n_i \\
Y_{..} &= \sum_{i=1}^{r} Y_{i.} \\
\overline{Y}_{..} &= Y_{..}/n_T = \sum_{i=1}^{r} \frac{n_i}{n_T} \overline{Y}_{i.}
\end{aligned}
$$

Estimating μ_i

- We will estimate μ_i by minimizing the sums-of-squares:

$$\begin{aligned} Q &= \sum_i \sum_j \left(Y_{ij} - \mu_i\right)^2 \\ &= \sum_j \left(Y_{1j} - \mu_1\right)^2 + \cdots + \sum_j \left(Y_{rj} - \mu_r\right)^2 \end{aligned}$$

- Easy to see that $\hat{\mu}_i = \overline{Y}_{i.}$
  - The within-group sample mean.
- Note that if we assume normality, then this is also the maximum likelihood estimator.
- The residuals are $\hat{\epsilon}_{ij} = Y_{ij} - \overline{Y}_{i.}$
  - Note that $\sum_j \hat{\epsilon}_{ij} = 0$
  - This is just OLS with a specific design matrix.

Sums-of-Squares

- The formulation of ANOVA tables through sums-of-squares is similar to what you did last semester.
- Look at $Y_{ij} - \overline{Y}_{\cdot\cdot} = \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right) + \left(Y_{ij} - \overline{Y}_{i\cdot}\right)$.
  - Total variability about the mean, between group variability, within group variability.
- If we square each side and sum over both i and j
  - SSTO = SSTR + SSE
  - SSTO = $\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2$
  - SSTR = $\sum_i n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$
  - SSE = $\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 = \sum_i \sum_j \hat{\epsilon}_{ij}^2$

Degrees of Freedom: Intuition

- Recall that sums-of-squares have degrees of freedom. Since we are working with linear regression, dof is just the trace of the corresponding orthogonal projection matrix.
  - Rank of the matrix of the quadratic form.
- SSTO has $n_T - 1$ df.
  - Constraint of $\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{..}\right) = 0$.
- SSTR has $r - 1$ df.
  - Constraint of $\sum_i n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right) = 0$.
- SSE has $n_T - r$ df.
  - r constraints of $\sum_j \left(Y_{ij} - \overline{Y}_{i.}\right) = 0$.

Some Notation: Back to Regression

- Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ and $Y = (Y_1^T, \ldots, Y_r^T)^T$.
- Let $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ and $\epsilon = (\epsilon_1^T, \ldots, \epsilon_r^T)^T$.
- Let $1_n, 0_n$ be the $n-$vectors of all ones and zeros, respectively.
- Let X_i be the $n_i \times r$ matrix of zeros except for the i^{th} column 1_{n_i} and $X = (X_1^T, \ldots, X_r^T)^T$.
- $\hat{\mu} = (X^T X)^{-1} X^T Y = (\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, \ldots, \overline{Y}_{r\cdot})^T$
- Let $H = X(X^T X)^{-1} X^T$.
- Homework: compute H and show that $HY = (\overline{Y}_{1\cdot} 1_{n_1}^T, \ldots, \overline{Y}_{r\cdot} 1_{n_r}^T)^T$

Sums of Squares as Quadratic Forms

- SSTO = $\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{..}\right)^2 = Y^T(I - \frac{1}{n_T} 1_{n_T} 1_{n_T}^T)Y$.
- SSTR = $\sum_i n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 = Y^T(H - \frac{1}{n_T} 1_{n_T} 1_{n_T}^T)Y$
- SSE = $\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i.}\right)^2 = Y^T(I - H)Y$
- The matrix of each quadratic form is idempotent.
- The matrix rank is the degrees of freedom of the sums of squares. For idempotent matrix, rank = trace.
  - $\text{rk}(I - \frac{1}{n_T} 1_{n_T} 1_{n_T}^T) = n_T - 1$
  - $\text{rk}(H - \frac{1}{n_T} 1_{n_T} 1_{n_T}^T) = r - 1$
  - $\text{rk}(I - H) = n_T - r$

Testing of Mean

- $H_0 : \mu_1 = \cdots = \mu_r$, vs H_a: not all equal.
- Intuitively: look at SSTR (between group variability) and SSE (within group variability).
- F-test: under normality assumption,
  $F = MSTR/MSE \sim F(r-1, n-r)$ under H_0.

Mean Sums-of-Squares

- Define $MSTR = SSTR/(r-1)$, $MSE = SSE/(n_T - r)$.
- Can find that
  - $E\,MSE = \sigma^2$
  - $E\,MSTR = \sigma^2 + \frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r-1}$ where $\mu_\cdot = \sum n_i \mu_i / n_T$
- MSE is an unbiased estimate of σ^2
- $\frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r-1}$
  - zero when there is no difference between treatments.
  - overall measure of how different the groups are.
  - relates to the non-central parameter and power calculation.

Power and Sample Size Calculations

- We found the distribution of F^* under $H_0 : \mu_1 = \cdots = \mu_r$.
- Its distribution under an alternative $H_a : \mu_1 = \mu_1^a, \ldots, \mu_r = \mu_r^a$ will be needed to compute power and sample size.
- Recall power $1 - \beta$: the probability of rejecting the null given that a certain alternative is true.
- Last semester, we give the general alternative distribution of F^* under the general linear models framework:
  - SSE and SSTR are independent.
  - SSE is a χ^2.
  - SSTR is now a non-central χ^2.
  - F^* is subsequently a non-central F.
- Here we look at details for the single factor ANOVA case.

Distribution of F^* Under H_a

- $F^* = \frac{MSTR}{MSE} \sim F_{r-1, n_T - r, \lambda}$, $\lambda = \sum_i n_i(\mu_i - \mu_\cdot)^2/\sigma^2$
- When there are an equal number of subjects per group:
  - $n_i = \frac{n_T}{r} \Rightarrow \lambda = \frac{n_T}{r} \frac{\sum_i (\mu_i - \mu_\cdot)^2}{\sigma^2}$
- Text uses a different notation:
  - Uses $\phi = \sqrt{\frac{\lambda}{r}}$.

Power Calculations

- Under $H_0 : \mu_1 = \cdots = \mu_r$, $F^* = MSTR/MSE \sim F_{r-1, n_T - r}$.
- At an α, we have a decision rule where we reject H_0 iff $F^* > F_{r-1, n_T - r}(1-\alpha)$.
- Recall that power is the probability of rejecting H_0 given some alternative.
- Under the alternative $H_a : \lambda = \sum_i n_i(\mu_i - \mu_.)^2/\sigma^2 \neq 0$ $F^* \sim F_{r-1, n_T - r, \lambda}$.
- Power = $1 - \beta = \Pr\left\{F^* > F_{r-1, n_T - r}(1-\alpha)\right\}$, where $F^* \sim F_{r-1, n_T - r, \lambda}$.
  - Computed with `pf(q =` F^*`, df1 =` $r-1$`, df2 =` $n-T$`, ncp =` λ`, lower.tail = F)`
- Note that power depends on
  - Knowing μ_i for $i = 1, \ldots, r$.
  - Knowing n_i for $i = 1, \ldots, r$.
  - Knowing σ^2.

Power

- You use past experiences to gather σ^2
  - Or MSE
- You are interested in a specific difference among the means.
  - the μ_i's come from your desired question.
  - their difference is known as an effect size.
- For a given $n_1, \ldots, n_r$ you can compute the power.
- Conversely, for a given power, you can computed the necessary sample size.
- Book uses tables.

Changes in λ

- Notice that as λ increases our power increases.
- What will make $\lambda = \frac{\sum n_i(\mu_i - \overline{\mu}_.)^2}{\sigma^2}$ increase?
  - If σ^2 decreases. There is less noise in the data.
  - If n_i increase. We have more subjects.
  - If $|\mu_i - \overline{\mu}_.|$ increases. We are interested in rejecting the null if there is a bigger discrepancy between the factor means.

Computing Sample Sizes in a balanced design

- Fix α, assume values of μ_i and σ^2, and a balanced design $n_i = n$. For a desired level of power, we can find the minimum required sample size.
- We want to find the smallest n_T such that:

$$\Pr(F^* > F_{r-1,n_T-r}(1-\alpha) | F^* \sim F_{r-1,n_T-r,\lambda(n_T)}) \geq 1-\beta.$$

- $\lambda(n_T) = \frac{n_T}{r} \frac{\sum(\mu_i - \overline{\mu}.)^2}{\sigma^2}$.

Cereal Example

- Recall the cereal example from last class.
- We have 4 different box designs and want to know what design sells better.
- We can do this in R. What would be a simple routine to do this?

Computing Sample Sizes in a balanced design

- We do not know each μ_i, but we can assume that the maximum difference among μ_is is $\Delta = 5.5$ boxes.
- For any $\mu_1, \ldots, \mu_r$ such that $\Delta = \max(\mu_i) - \min(\mu_i)$: $\sum(\mu_i - \mu_.)^2 \geq \Delta^2/2$
- We do power calculation based on the situation with smallest λ (be conservative).
  - Set one group mean at 0.
  - Set a second group mean at Δ.
  - Set all others at $\Delta/2$.
- $\lambda = \frac{n_T}{4} \frac{\sum_i(\mu_i - \overline{\mu}_.)^2}{\sigma^2} = \frac{n_T}{4} \frac{5.5^2}{2} \frac{1}{3.5^2}$
- We want to find the smallest n_T such that: $\Pr(F^* > F_{r-1, n_T - r}(1 - 0.05) | F^* \sim F_{r-1, n_T - r, \lambda(n_T)}) \geq 90\%$.

Assumptions of one-way ANOVA

- The one-way ANOVA model makes several assumptions which we must check.
- The assumptions are similar to those made in regression (in order of importance):
  1. No outliers (i.e. mean model is correct, all data have a second moment).
  2. Equal variance among factor levels.
  3. Observations are independent conditional on factor level.
  4. Normality (when using F-tests).
- How to check these assumptions:
  1. Residual plots,
  2. QQ-Plots
  3. Some formal tests: Levene Test, and Brown-Forsythe Test.

What if something is wrong?

- How to fix our assumptions if something is wrong.
- Outliers - at least check the fitting without the outliers.
- Correlation over other variables - put them in your model.
- Unequal variances:
  - Box-Cox transformation.
  - Weighted regression.
  - Also could do a randomization test.
- Equal variances but no normality:
  - Use non-parametric Kruskal-Wallis test or randomization test.
- Lack of normality and unequal variances:
  - Box-Cox transformation.

Formal Test for Heterogeneity of Variances

- Consider the model $Y_{ij} = \mu_i + \sigma_i \epsilon_{ij}$
- ϵ_{ij} are independent, zero mean, $Var(\epsilon_{ij}) = 1$.
- If there are $r = 2$ groups, what's one test?
- Levene's tests $H_0 : \sigma_1^2 = \cdots = \sigma_r^2$
  - $d_{ij} = \left| Y_{ij} - \frac{\sum Y_{ij}}{n_i} \right|$ - absolute deviations.
  - Do the F test on the absolute deviations.
  - $F_L^* = \frac{MSTR}{MSE}$
  - $MSTR = \frac{\sum_i n_i (\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{r-1}$
  - $MSE = \frac{\sum_j \sum_i (d_{ij} - \bar{d}_{i\cdot})^2}{n_T - r}$
  - Asymptotically, $F_L^* \sim F_{r-1, n_T - r}$ under H_0.
- Brown-Forsythe's test is similarly except:
  - $d_{ij} = |Y_{ij} - \text{median}(Y_{i1}, \ldots, Y_{in_i})|$
- Levene's test has better performance for normal data (since it uses the mean).
- Brown-Forsythe is more robust to departures from normality (since it uses the median).

Formal Test for Heterogeneity of Variances

- Consider the model $Y_{ij} = \mu_i + \sigma_i \epsilon_{ij}$
- ϵ_{ij} are independent, zero mean, $Var(\epsilon_{ij}) = 1$.
- If there are $r = 2$ groups, what's one test?
- Levene's tests $H_0 : \sigma_1^2 = \cdots = \sigma_r^2$
  - $d_{ij} = \left| Y_{ij} - \frac{\sum Y_{ij}}{n_i} \right|$ - absolute deviations.
  - Do the F test on the absolute deviations.
  - $F_L^* = \frac{MSTR}{MSE}$
  - $MSTR = \frac{\sum_i n_i (\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{r-1}$
  - $MSE = \frac{\sum_j \sum_i (d_{ij} - \bar{d}_{i\cdot})^2}{n_T - r}$
  - Asymptotically, $F_L^* \sim F_{r-1, n_T - r}$ under H_0.
- Brown-Forsythe's test is similarly except:
  - $d_{ij} = |Y_{ij} - \text{median}(Y_{i1}, \ldots, Y_{in_i})|$
- Levene's test has better performance for normal data (since it uses the mean).
- Brown-Forsythe is more robust to departures from normality (since it uses the median).

Weighted Regression

- What if $Y_{ij} = \mu_i + \sigma_i \epsilon_{ij}$ where $\text{Var}(\epsilon_{ij}) = 1$?
- Can do a weighted regression by minimizing

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} \sigma_i^{-2} \left(Y_{ij} - \mu_i \right)^2 .$$

- Don't know σ_i^2 but can estimate it from our data as long as n_i is sufficiently large.

Cereal and ABT Examples from Text

- The ABT Electronics Corporation wants to know the reliability of 5 types of fluxes.
- Designs a study where 8 circuit boards are produced per flux.
- After 4 weeks, each board was tested to see how much pressure was exerted before it broke.
- Is there any difference in amount of pressure per flux?
- If so, where are the differences?
  - Don't forget the multiple comparison problem.
  - Will talk about Tukey's procedure next class.

Cereal Example

- You are working for a food company.
- You have 4 different package designs for a box of cereal.
  - Factor = package designs, levels = 4.
- You have 20 stores with equal sales volumes that want to participate.
- You create a balanced design and randomly assign a package design to each store. The one store has a fire.
  - n=19 total experimental units, $n_i = 5, 5, 4, 5$.
- Measure the number of boxes sold in a week.

What if we can't trust normality?

- Suppose $i = 1, \ldots, c$, and $Y_{ij} \sim F(y - \mu_i)$, i.e. distribution of trt levels forms a location family. Now, F is unknown (we usually assume $F = \Phi$).
- Want to test $H_0 : \mu_1 = \cdots = \mu_c$.
- An option is a rank based test:
  - Called Kruskal-Wallis when there are more than two levels.
  - Called Mann-Whitney (Wilcoxon) when there are two levels.
- Idea, do your analysis on ranks rather than the data.
- Rank your observations (all n_T) from smallest to largest.
  - r_{ij} is the rank of Y_{ij}.
  - If there are ties, give them the average rank value.
  - $\bar{r}_{i\cdot} = \sum_j r_{ij}/n_i$
  - $\bar{r} = (n_T + 1)/2$

What if we can't trust normality?

- Suppose $i = 1, \ldots, c$, and $Y_{ij} \sim F(y - \mu_i)$, i.e. distribution of trt levels forms a location family. Now, F is unknown (we usually assume $F = \Phi$).
- Want to test $H_0 : \mu_1 = \cdots = \mu_c$.
- An option is a rank based test:
  - Called Kruskal-Wallis when there are more than two levels.
  - Called Mann-Whitney (Wilcoxon) when there are two levels.
- Idea, do your analysis on ranks rather than the data.
- Rank your observations (all n_T) from smallest to largest.
  - r_{ij} is the rank of Y_{ij}.
  - If there are ties, give them the average rank value.
  - $\bar{r}_{i\cdot} = \sum_j r_{ij}/n_i$
  - $\bar{r} = (n_T + 1)/2$

Tests based on the rank

Basic idea:

- Originally formulated as an approximation to the ANOVA F-test on the ranks.
- Consider $\tilde{K} = \frac{SSTR}{SSE} \times \frac{n-c}{c-1}$, we reject H_0 if it is large.
- $\tilde{K}$ is a monotone function of $SSTR/SSTO$, and $SSTO = n_T(n_T+1)(n_T-1)/12$.
- Therefore, we basically only look at a scaled version of SSTR, the scale is roughly the variance of r_{ij}.
- Consider χ^2 test rather than F test, since $SSTR$ is the only source of uncertainty.

Kruskal-Wallis

- $K = \frac{12}{n_T(n_T+1)} \sum_i n_i(\bar{r}_{i\cdot} - \frac{n_T+1}{2})^2$
- If n_i are large and $n_i/n_T \to \tau_i$, then under $H_0 : \mu_1 = \cdots = \mu_c$, $K \sim \chi^2_{r-1}$.
- We will look at a sketch of the ideas behind K-W.
- On the Courseweb is Kruskal's original 1952 Annals paper: contains the original proof.

Idea behind K-W

- Note that under H_0, r_{ij} is a uniform random variable over $[1, \ldots, n_T]$. Distribution is easy to work with!
- Let's look at behavior of $r_{i.} - E(r_{i.})$.
- By a complicated CLT, asymptotically:
  - $T_i = \sqrt{12}\frac{r_{i.} - E(r_{i.})}{n_T^{3/2}\sqrt{n_i/n_T}}$
  - Are asymptotically zero mean normal with covariance $\delta_{ii'} - \sqrt{\tau_i \tau_{i'}}$, where $\lim n_i/n_T = \tau_i$.
- Claim: for $T = (T_1, \ldots, T_c)^T$, asymptotic variance of T is symmetric and idempotent with rank $c - 1$.
  - Why then $\sum_{i=1}^{c} T_i^2 \xrightarrow{\mathcal{D}} \chi^2_{c-1}$?
  - Intuition: We loose one degree of freedom since $\sum r_{i.} = n_T(n_T + 1)/2$.
- Some algebra can show that $K = \frac{n_T}{n_T+1}\sum T_i^2$.

Idea behind K-W

- Note that under H_0, r_{ij} is a uniform random variable over $[1, \ldots, n_T]$. Distribution is easy to work with!
- Let's look at behavior of $r_{i\cdot} - E(r_{i\cdot})$.
- By a complicated CLT, asymptotically:
  - $T_i = \sqrt{12}\frac{r_{i\cdot} - E(r_{i\cdot})}{n_T^{3/2}\sqrt{n_i/n_T}}$
  - Are asymptotically zero mean normal with covariance $\delta_{ii'} - \sqrt{\tau_i \tau_{i'}}$, where $\lim n_i/n_T = \tau_i$.
- Claim: for $T = (T_1, \ldots, T_c)^T$, asymptotic variance of T is symmetric and idempotent with rank $c - 1$.
  - Why then $\sum_{i=1}^{c} T_i^2 \xrightarrow{\mathcal{D}} \chi^2_{c-1}$?
  - Intuition: We loose one degree of freedom since $\sum r_{i\cdot} = n_T(n_T + 1)/2$.
- Some algebra can show that $K = \frac{n_T}{n_T+1} \sum T_i^2$.

Idea behind K-W

- Note that under H_0, r_{ij} is a uniform random variable over $[1, \ldots, n_T]$. Distribution is easy to work with!
- Let's look at behavior of $r_{i\cdot} - E(r_{i\cdot})$.
- By a complicated CLT, asymptotically:
  - $T_i = \sqrt{12}\frac{r_{i\cdot} - E(r_{i\cdot})}{n_T^{3/2}\sqrt{n_i/n_T}}$
  - Are asymptotically zero mean normal with covariance $\delta_{ii'} - \sqrt{\tau_i \tau_{i'}}$, where $\lim n_i/n_T = \tau_i$.
- Claim: for $T = (T_1, \ldots, T_c)^T$, asymptotic variance of T is symmetric and idempotent with rank $c - 1$.
  - Why then $\sum_{i=1}^{c} T_i^2 \xrightarrow{\mathcal{D}} \chi^2_{c-1}$?
  - Intuition: We loose one degree of freedom since $\sum r_{i\cdot} = n_T(n_T + 1)/2$.
- Some algebra can show that $K = \frac{n_T}{n_T+1} \sum T_i^2$.

Um...doesn't this use CLT?

- K-W following the χ^2 distribution uses the CLT.
- It is based on the CLT on the ranks.
- Under the null, the ranks follow a uniform distribution over $1, \ldots, n_T$.
- As $n_T \to \infty$, the distribution of T_i looks rather normal.
- The CLT on the outcomes Y_{ij} will not work well if is distribution is very skewed.

Small Sample Sizes

- The asymptotic χ^2 distribution requires large sample sizes.
- If you do not have a lot of data, can do an exact test.
  - If the null is true, then the assignment of an outcome to a group can be seen as completely random.
  - Consider any division of the n_T observed values into r groups of size $n_1, \ldots, n_r$.
  - Under H_0, each of these is as likely as any other.

- There are $\frac{n_T!}{n_1!\ldots n_r!}$ different assignments of the observed values into groups.
- For each one of these assignments, compute K and call it K_g.
- This will produce an empirical distribution for K under the null.
- Let K^* be the K-W statistic from the data.
- Recall: p-value is the prob. that you observe a test statistic as or more extreme than what you did under the null.
- Under the empirical distribution, p-value = % of $K_g \geq K^*$

proc npar1way

- proc npar1way can be used to compute K-W.
- Exact test can take a long time.
- In the cereal example:
  - $\frac{19!}{5!5!5!4!} = 2.9 \times 10^9$
  - Takes a while in SAS.
- Can specify MC (Monte Carlo) so that not all combinations are computed.
  - Randomly choose M combinations.
  - Compute the K-W statistic for each combination and get the empirical distribution for the M values.
  - It is an approximate exact test.

Summary of Mean Test

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$, vs H_a: not all are equal.
- If errors are normal or close to normal or having a large sample size, use F-test.
- If errors are not normal, and with reasonably large sample size, we can use rank based nonparametric tests, χ^2 tests.
- If errors are not normal and sample size is small, rank based exact tests.