ExpEcon Methods: Popular Hypothesis Tests

ECON 8877

P.J. Healy

First version of distributional test slides thanks to Floyd Carey

Updated 2023-11-17

Siegel & Castellan's Book

Siegel & Castellan (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed.

Siegel & Castellan's Book

The back jacket of 2nd Edition:

What do your data look like?

1. Nominal/Categorical
   - Pass/fail, gender, race...
2. Ordinal
   - Type/ability
3. Interval
   - Score on a test

Siegel & Castellan's Book

The back jacket of 2nd Edition:

What do you want to test?

1. One Sample
   - Value of a statistic ($\mu = 0$)
   - Fit of a sample to a distribution ($X \sim N(0, 1)$)
   - Properties of a sample (runs test, symmetry test)
2. Comparing Two or More Samples
   2.1 Matched samples
      - "Sample of differences" ($\mu_{diff} = 0$)

   2.2 Independent samples
      - Comparing statistics ($\mu_1 = \mu_2$)
3. Measuring Association Between Two Samples
   3.1 Various notions of "correlation"

Contingency Tables

Comparing samples with categorical data
(or ordinal, discarding order info)

Category	Control	Treatment
High	A	B
Low	C	D

H_0: $A/C = B/D$ (category is independent of treatment)
Fisher's Exact Test:
Prob of (A, B, C, D) under H_0: Hypergeometric dist'n
Calculate prob of all tables "more extreme" (less equal) than this
Exact test b/c sampling distribution is known for any n

Problem: calculation intensive! Only for small tables.

Contingency Tables

The Chi-Squared Test (for contingency tables):

OBSERVED	Male	Female
Pass	30	70
Fail	70	30

EXPECTED	Male	Female
Pass	50	50
Fail	50	50

$(O-E)^2/E$	Male	Female
Pass	8	8
Fail	8	8

Test statistic $T = \sum \frac{(O-E)^2}{E}$. As $n \to \infty$ we have $T \sim \chi^2_{(r-1)(c-1)}$

Partitioning the D.O.F.

OBSERVED	Black	White	Asian
Pass	70	70	30
Fail	30	30	70

χ^2 test rejects H_0. But which race is different?
Tempted to test all 2×2 subtables, but they're not independent

	Black	White
Pass	70	70
Fail	30	30

	Black + White	Asian
Pass	140	30
Fail	60	70

As many subtables as there are d.o.f.

Fisher vs. Chi-Squared

- Use Fisher if your computer can do it
- Chi-Squared test: invalid of $E \leq 5$ in any cell
    - Combine cells?
    - Continuity correction (maybe automatic)

Binomial/Proportion Test

What fraction of people passed this test? H_0: $p = p_0$
$x_i \in \{0, 1\}$, $i = 1, 2, \dots, n$
$y = \sum_i x_i$, $\hat{p} = y/n$

Recall binomial distribution: $Pr[Y = k] = \binom{n}{k} p_0^k (1 - p_0)^{n-k}$
One-sided test: $Pr[Y \geq y] = \sum_{k=y}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k}$
Two-sided test (if $y > p_0 n$): $Pr[Y \geq y] + Pr[Y \leq p_0 n - (y - p_0 n)]$

Large samples: use Normal approximation w/ continuity correction

Tests of Association

Requires paired data! $n = m$

- Pearson

$$r_{X,Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\hat{Cov}(X,Y)}{\hat{\sigma}(X)\hat{\sigma}(Y)} \in [-1, 1]$$

  - interval scale
  - approximately normally distributed.
  - linear relationship between the two variables.
  - minimal outliers
  - homoscedasticity of the data.
- Spearman
  - Simply Pearson, but on rank values
  - Still needs interval scale
  - No Normality assumption
  - Monotonic (but non-linear) relationship

Tests of Association

- Kendall rank correlation
  - Interval or ordinal
  - Form pairs (x_i, y_i) vs. (x_j, y_j)
  - Define:
    - n_c = of pairs that "move together" ($x_j > x_i$ & $y_j > y_i$)
    - n_d = of pairs that "move oppositely"
    - n_1 = of pairs where $x_i = x_j$
    - n_2 = of pairs whre $y_i = y_j$
    - $\bar{n} = \frac{n(n-1)}{2} = \sum_{i=1}^{n-1} i$
  - Test statistic: $\tau = \frac{n_c - n_d}{\sqrt{(\bar{n}-n_1)(\bar{n}-n_2)}} \in [-1, 1]$
  - Distribution of τ under H_0 known for small n
    - Approximately normal for large n
  - Preferred to Spearman for small n or outliers
  - Direct interpretation
- Cramer
  - Ordinal or categorical
  - Simply an adjustment to the χ^2 statistic, ranging 0 to 1

Part 1: Distributional Tests

Distributional Tests

Distributional tests determine how likely a sample is to have come from a pre-specified distribution or how likely two samples are to have been drawn from the same distribution.

The most well-known (and general) of these tests is the Kolmogorov-Smirnov (KS) test.

One-Sample Kolmogorov-Smirnov Test

- The one-sample KS test compares the cumulative distribution of a sample of size n ($S_n(x)$) to a pre-specified cumulative distribution function ($F_0(x)$).

$$S_n(x) = k/n$$

- where k = the number of observations $\leq x$.

One-Sample Kolmogorov-Smirnov Test (Continued)

- The test is based entirely on the *largest* deviation between $F_0(x)$ and $S_n(x)$, denoted as D_n.

$$D_n = \sup_x |F_0(x) - S_n(x)|$$

- Under the null hypothesis that the sample is drawn from $F_0(x)$, $\lim_{n\to\infty} D_n = 0$.
- the null hypothesis is rejected if $\sqrt{n}D_n > K_\alpha$, where K_α is found such that $Pr(K \leq K_\alpha) = 1 - \alpha$ and K is the Kolmogorov distribution.

One-Sample Kolmogorov-Smirnov Test Figure

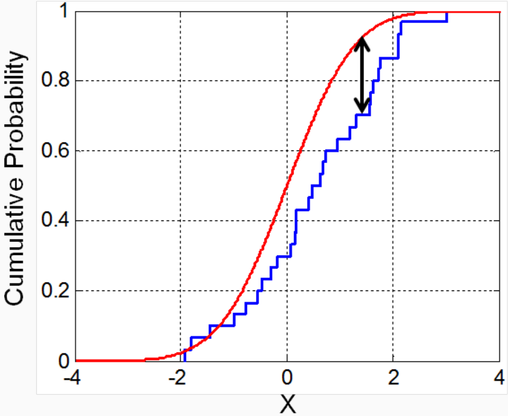


Abbildung 1: *

The red line is $F_0(x)$, the blue line is $S_n(x)$, and the black line is D_n (Bscan, CC0, via Wikimedia Commons)

Example: One-Sample Kolmogorov-Smirnov Test

- Suppose that there are 5 different salsas where each subsequent salsa is spicier (i.e., the salsa denoted by x_{n+1} is spicier than the salsa denoted by x_n).
- Further, suppose that the null hypothesis is that preferences over salsa spiciness is uniformly distributed in the population (i.e., $F_0(x) = \frac{x}{5}$).
- In a sample of 10 subjects (n = 10), one subject prefers the least spicy salsa, 5 subjects prefer the second most spicy salsa, and 4 subjects prefer the spiciest salsa.

Example: One-Sample Kolmogorov-Smirnov Test (Continued)

- The difference between $F_0(x)$ and $S_{10}(x)$ is maximized at $x = 3$.
    - $F_0(3) = \frac{3}{5}$ and $S_{10}(3) = \frac{1}{10}$.
- Therefore, $D_n = \frac{3}{5} - \frac{1}{10} = .5$.
- $\sqrt{n}D_n = \sqrt{10} \cdot .5 = 1.581$
- $K_{.01} = .48895$
- Because $\sqrt{n}D_n = 1.581 > .48895 = K_{.01}$, we can reject the null hypothesis that the sample was drawn from a population whose preferences for salsa spiciness was uniformly distributed with 99% confidence.

Alternatives to the One-Sample KS Test

- Another test, which is based on the quadratic difference between the pre-specified distribution instead of the maximum difference is the Anderson-Darling (AD) test (Anderson & Darling, 1952).
  - The AD test is a modification of the Cramer-von Mises (CVM) test (1928).
- The test statistic is:

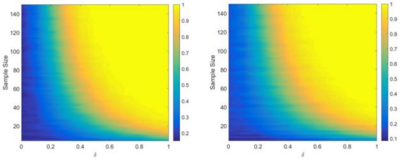
$$W_n^2 = n \int_{-\infty}^{\infty} [S_n(x) - F_0(x)]^2 \psi(F_0(X)) dF_0(x)$$

- Where $\Psi = [F_0(X)(1 - F_0(X))]^{-1}$

Power Comparison for KS and AD tests

Figure 7. The CNA figure quick part

Figure 5. Simulated statistical power for normal distribution using AD (left) and KS (right) tests



Simulated statistical power for exponential distribution with μ_o=1 using AD (left) and KS (right) tests

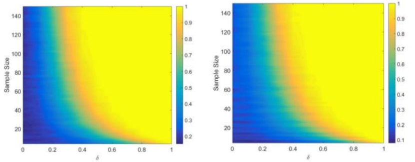


Figure 6. Simulated statistical power for lognormal distribution using AD (left) and K (right) tests

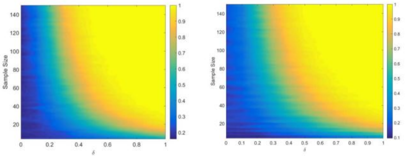
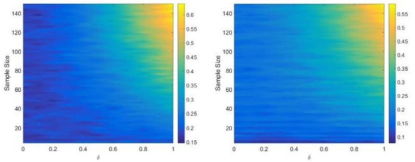


Figure 8. Simulated statistical power for exponential distribution with μ_o=5 using AD (left) and KS (right) tests



[b] (2016)

Bo[b]y (2016)

Other Alternatives to the One-Sample KS Test

- Suppose that you don't know the exact null distribution against which you'd like to compare your sample, but you know that the null distribution should be normally distributed.
- You have a few options: The Lillefors (LF) test (Lillefors, 1967) and the Shapiro-Wilk (SW) test (Shapiro & Wilk, 1965).
- The SW test is the most powerful, and the LF test is the least powerful for a broad range of normal distributions (Razali & Wah, 2011).

The Shapiro-Wilk Test

- The SW test uses the test statistic:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\bar{x}$ is the sample mean
- and $\boldsymbol{a}_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$

The Shapiro-Wilk Test (Continued)

- where $\boldsymbol{m}_i = (m_1, \ldots, m_n)^T$ are the expected values of order statistics of independent and identically distributed random variables sampled from the standard normal distribution
- and $\boldsymbol{V}$ is the covariance matrix of those order statistics.
- m is computed using GLS, assuming that x is normally distributed.

$$\hat{\mu} = \frac{m'V^{-1}(m1' - 1m')V^{-1}x}{1'V^{-1}1m'V^{-1}m - (1'V^{-1}m)^2}$$

$$\hat{\sigma} = \frac{1'V^{-1}(1m' - m1')V^{-1}x}{1'V^{-1}1m'V^{-1}m - (1'V^{-1}m)^2}$$

- In practice, $\boldsymbol{a}_i$ is algorithmically approximated using Royston's (1994) AS R94 for $3 \leq n \leq 5000$.

Power Comparison for KS, LF, AD, and SW Tests

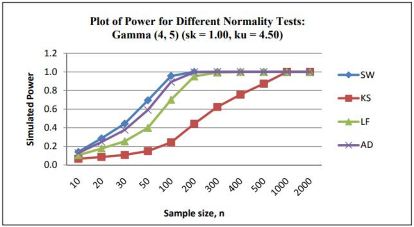


Figure 2(a): Comparison of Power for Different Normality Tests against Gamma (4,5) ($\alpha = 0.05$)

Razali & Wah, 2011

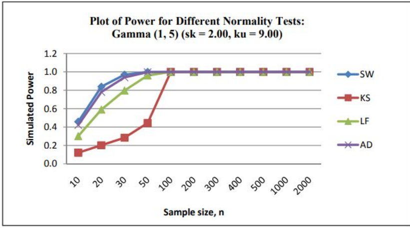


Figure 2(b): Comparison of Power for Different Normality Tests against Gamma (1,5) ($\alpha = 0.05$

Razali & Wah, 2011

The Two-Sample Kolmogorov-Smirnov Test

- Calculating the two-sample Kolmogorov-Smirnov test is similar to the one-sample counterpart except we replace $F_0(x)$ with $S_m(x)$, where the second sample has m members.
- Here, $D_{n,m} = \sup_x |S_n(x) - S_m(x)|$ for the two-sided test and $D_{n,m} = \sup_x [S_n(x) - S_m(x)]$ for the one-sided test.
- Siegel (1988) uses a heuristic that if n or m are less than 40, then n must equal m, but I have not found this in other papers.

Two-Sample Kolmogorov-Smirnov Test Figure

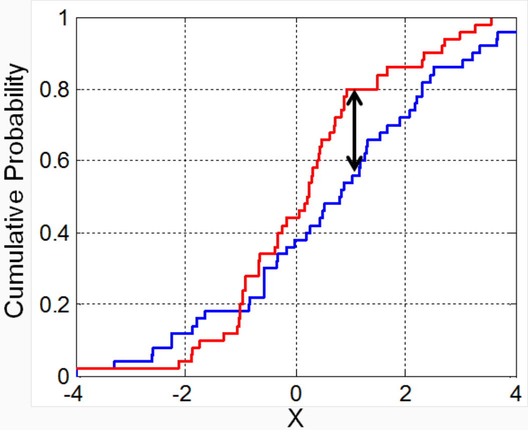


Abbildung 2: *

The red line is $S_m(x)$, the blue line is $S_n(x)$, and the black line is D_n (Bscan, CC0, via Wikimedia Commons)

The Two-Sample Kolmogorov-Smirnov Test (Continued)

- The two-sample test has different critical values from the one-sample test, but I don't think there is an analytical solution for small samples (I couldn't find one if there is!)
- There are tables for small samples, and for larger samples, the equation for the critical value is $c(\alpha)\sqrt{\frac{n+m}{n \cdot m}}$
- where $c(\alpha) = \sqrt{-ln(\frac{\alpha}{2}) \cdot \frac{1}{2}}$ (Knuth, 1998)

Alternatives to the Two-Sample KS Test

- Nearly all of the alternatives to the two-sample KS test are location-scale tests which incorporate both the sample means and standard deviations. The two most popular of this class are the Cucconi (C) test (1968) and the Lepage (L) test (1971).
- Both the C and the L tests are **FAR** more powerful than the Two-Sample KS Test in a simulation using several canonical distributions (Marozzi, 2009).
    - That same paper indicates that the C test is slightly more powerful than the L test.

Alternatives to the Two-Sample KS Test (Continued)

- This increase in power for location-scale tests is partially due to their assumptions on the alternative hypothesis.
- In location-scale tests, H_0 is that $F \equiv G$ and H_a is that $G(y) = F(ay + b)$ such that $a \neq 1$ and/or $b \neq 0$.
- In the two-sample KS test, H_0 is unchanged, but H_a does not specify an alternative distribution.

Power Comparison for KS, C, and L Tests

Table 5. Power estimates with $\alpha = 0.05$ and $(n_1, n_2) = (30, 30)$.

Normal							
$\mu_1 - \mu_2$	0	0	0.5	1	0.5	0.5	0.5
σ_1/σ_2	1	1.3	1.3	1.3	1	1.75	2.5
C	0.050	0.171	0.406	0.870	0.355	0.713	0.966
L	0.050	0.148	0.388	0.864	0.357	0.642	0.926
PG1	0.044	0.166	0.405	0.878	0.358	0.713	0.953
PG2	0.051	0.172	0.408	0.871	0.357	0.715	0.966
PG3	0.048	0.180	0.417	0.880	0.360	0.752	0.982
PG4	0.050	0.184	0.413	0.870	0.351	0.752	0.982
KS	0.035	0.049	0.284	0.799	0.320	0.328	0.493
CVM	0.050	0.079	0.410	0.909	0.451	0.518	0.846
Uniform							
$\mu_1 - \mu_2$	0	0	0.5	1	0.5	0.5	0.5
σ_1/σ_2	1	1.3	1.3	1.3	1	1.75	2.5
C	0.050	0.358	0.476	0.818	0.324	0.895	0.996
L	0.050	0.258	0.425	0.819	0.334	0.790	0.978
PG1	0.050	0.421	0.554	0.887	0.335	0.962	1.000
PG2	0.050	0.360	0.478	0.820	0.326	0.896	0.996
PG3	0.046	0.506	0.532	0.863	0.389	0.958	0.999
PG4	0.048	0.512	0.484	0.775	0.289	0.956	1.000
KS	0.034	0.059	0.255	0.729	0.223	0.424	0.721
CVM	0.052	0.097	0.375	0.890	0.391	0.609	0.967
Bimodal							
$\mu_1 - \mu_2$	0	0	1.1	2.2	0.75	0.75	0.75
σ_1/σ_2	1	1.3	1.3	1.3	1	1.4	2
C	0.048	0.304	0.550	0.932	0.246	0.543	0.968
L	0.049	0.249	0.523	0.931	0.253	0.478	0.915
PG1	0.047	0.337	0.593	0.967	0.245	0.593	0.985
PG2	0.049	0.306	0.553	0.932	0.247	0.546	0.968
PG3	0.047	0.336	0.597	0.959	0.277	0.591	0.985
PG4	0.048	0.341	0.541	0.924	0.232	0.570	0.984
KS	0.033	0.064	0.395	0.894	0.188	0.260	0.508
CVM	0.051	0.091	0.514	0.976	0.293	0.331	0.740

Marozzi, 2009