Lecture 13: Kolmogorov Smirnov Test & Power of Tests

S. Massa, Department of Statistics, University of Oxford

2 February 2016

An example

Suppose you are given the following 100 observations.

-0.16	-0.68	-0.32	-0.85	0.89	-2.28	0.63	0.41	0.15	0.74
1.30	-0.13	0.80	-0.75	0.28	-1.00	0.14	-1.38	-0.04	-0.25
-0.17	1.29	0.47	-1.23	0.21	-0.04	0.07	-0.08	0.32	-0.17
0.13	-1.94	0.78	0.19	-0.12	-0.19	0.76	-1.48	-0.01	0.20
-1.97	-0.37	3.08	-0.40	0.80	0.01	1.32	-0.47	2.29	-0.26
-1.52	-0.06	-1.02	1.06	0.60	1.15	1.92	-0.06	-0.19	0.67
0.29	0.58	0.02	2.18	-0.04	-0.13	-0.79	-1.28	-1.41	-0.23
0.65	-0.26	-0.17	-1.53	-1.69	-1.60	0.09	-1.11	0.30	0.71
-0.88	-0.03	0.56	-3.68	2.40	0.62	0.52	-1.25	0.85	-0.09
-0.23	-1.16	0.22	-1.68	0.50	-0.35	-0.35	-0.33	-0.24	0.25

Do they come from N(0,1)?

The Kolmogorov-Smirnov Test

Suppose that we have observations $X_1, \ldots, X_n$, which we think come from a distribution P.
The Kolmogorov-Smirnov Test is used to test

$$\mathbf{H}_0 : \quad \text{the samples come from } P,$$

against

$$\mathbf{H}_1 : \quad \text{the samples do not come from } P.$$

Cumulative Distribution Function and Empirical Distribution Function

The cumulative distribution function $F(x)$ of a random variable X, is

$$F(x) = \mathbb{P}(X \leq x).$$

The cumulative distribution function uniquely characterizes a probability distribution.

Given observations $x_1, \ldots, x_n$ the empirical distribution function $F_{\text{obs}}(x)$ gives the proportion of the data that lies below x,

$$F_{\text{obs}}(x) = \frac{\#\text{observations below } x}{\#\text{observations}}.$$

If we order the observations $y_1 \leq y_2 \leq \cdots \leq y_n$, then

$$F_{\text{obs}}(y_i) = \frac{i}{n}.$$

The Kolmogorov-Smirnov statistic

We want to compare the empirical distribution function of the data, F_{obs}, with the cumulative distribution function associated with the null hypothesis, F_{exp} (expected CDF).

The Kolmogorov-Smirnov statistic is

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|.$$

The practical approach

In practice, order the data:

-3.68	-2.28	-1.97	-1.94	-1.69	-1.68	-1.60	-1.53	-1.52	-1.48
-1.41	-1.38	-1.28	-1.25	-1.23	-1.16	-1.11	-1.02	-1.00	-0.88
-0.85	-0.79	-0.75	-0.68	-0.47	-0.40	-0.37	-0.35	-0.35	-0.33
-0.32	-0.26	-0.26	-0.25	-0.24	-0.23	-0.23	-0.19	-0.19	-0.17
-0.17	-0.17	-0.16	-0.13	-0.13	-0.12	-0.09	-0.08	-0.06	-0.06
-0.04	-0.04	-0.04	-0.03	-0.01	0.01	0.02	0.07	0.09	0.13
0.14	0.15	0.19	0.20	0.21	0.22	0.25	0.28	0.29	0.30
0.32	0.41	0.47	0.50	0.52	0.56	0.58	0.60	0.62	0.63
0.65	0.67	0.71	0.74	0.76	0.78	0.80	0.80	0.85	0.89
1.06	1.15	1.29	1.30	1.32	1.92	2.18	2.29	2.40	3.08

Then compute the empirical distribution function:

$$F_{\text{obs}}(-3.68) = \frac{1}{100}, \quad F_{\text{obs}}(-2.28) = \frac{2}{100}, \ldots, \quad F_{\text{obs}}(3.08) = 1.$$

If our data is ordered, x_1 being the least and x_n being the largest, then

$$F_{\text{obs}}(x_i) = \frac{i}{100}.$$

The practical approach

0.000	0.011	0.024	0.026	0.045	0.047	0.055	0.064	0.064	0.070
0.080	0.084	0.101	0.107	0.110	0.123	0.133	0.154	0.158	0.189
0.198	0.215	0.226	0.249	0.321	0.343	0.356	0.362	0.363	0.369
0.375	0.396	0.399	0.400	0.407	0.409	0.410	0.423	0.425	0.432
0.432	0.434	0.437	0.447	0.449	0.453	0.464	0.468	0.476	0.477
0.484	0.484	0.485	0.490	0.496	0.505	0.508	0.526	0.535	0.553
0.557	0.560	0.577	0.577	0.582	0.588	0.597	0.610	0.614	0.617
0.627	0.658	0.680	0.692	0.698	0.711	0.720	0.727	0.732	0.735
0.743	0.748	0.761	0.771	0.777	0.783	0.788	0.789	0.803	0.812
0.854	0.874	0.902	0.903	0.907	0.973	0.985	0.989	0.992	0.999

For each observation x_i compute $F_{\text{exp}}(x_i) = P(Z \leq x_i)$.

In this case the expected distribution function is standard normal so use the normal table.

The practical approach

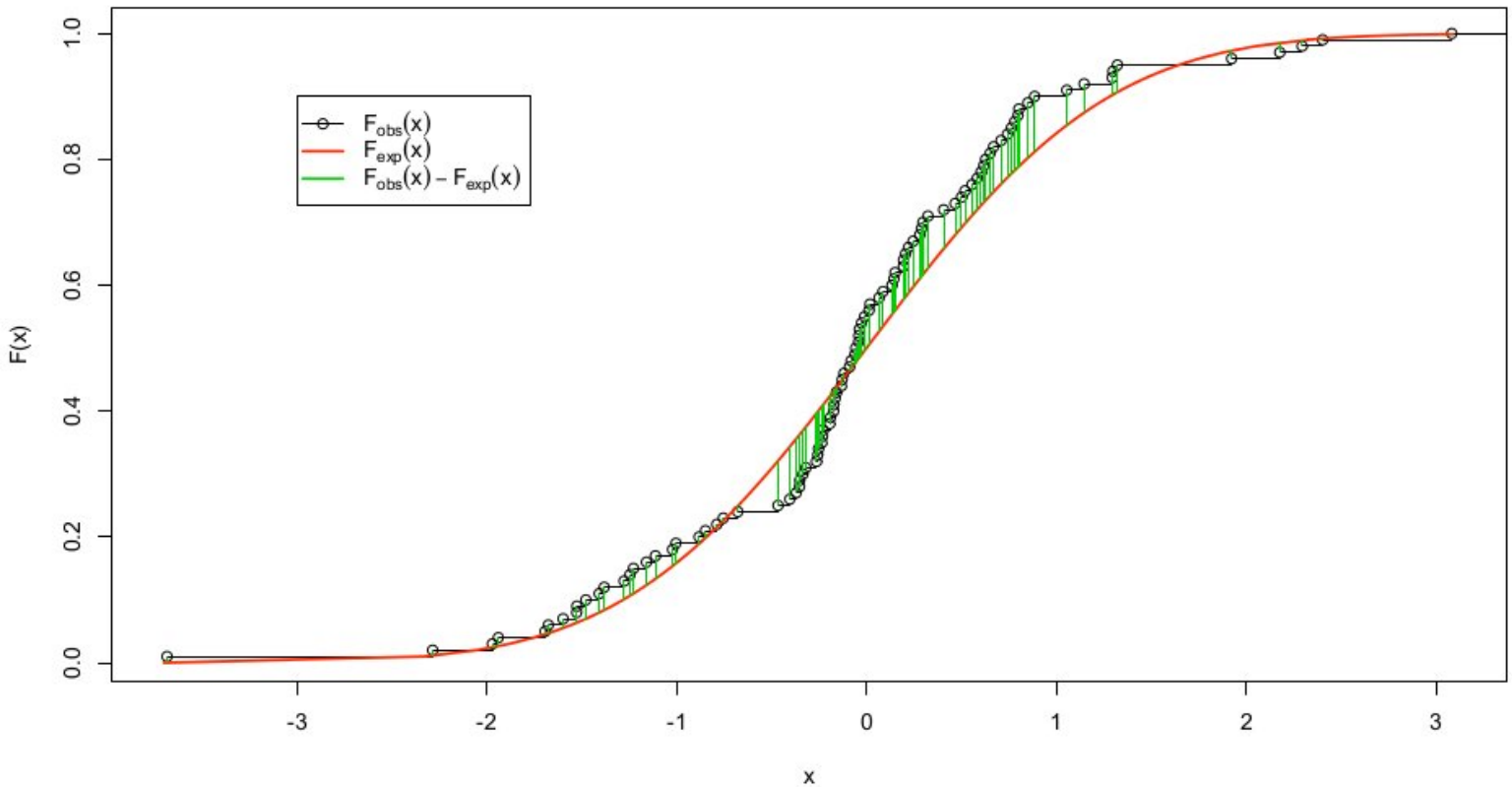
Compute the absolute differences between the entries in the two tables.

The Kolmogorov-Smirnov statistic $D_n = 0.092$ is the maximum shown here in blue.

0.010	0.009	0.006	0.014	0.004	0.014	0.015	0.017	0.026	0.031
0.031	0.036	0.030	0.034	0.041	0.037	0.037	0.026	0.031	0.011
0.012	0.005	0.003	0.008	0.069	0.085	0.086	0.083	0.073	0.071
0.064	0.077	0.067	0.061	0.055	0.049	0.039	0.045	0.035	0.033
0.023	0.013	0.006	0.008	0.002	0.008	0.006	0.012	0.014	0.024
0.026	0.036	0.046	0.052	0.054	0.056	0.062	0.052	0.054	0.048
0.054	0.060	0.055	0.061	0.067	0.073	0.071	0.070	0.076	0.082
0.084	0.061	0.049	0.049	0.052	0.048	0.051	0.054	0.058	0.064
0.068	0.071	0.069	0.070	0.074	0.078	0.082	0.092	0.088	0.087
0.055	0.045	0.029	0.037	0.043	0.013	0.015	0.009	0.002	0.001

The Kolmogorov-Smirnov Statistic

We have calculated the maximum absolute distance between the expected and observed distribution functions, in green in the plot below.



What is the Critical Value?

- At the 95% level the critical value is approximately given by

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}.$$

- Here we have a sample size of $n = 100$ so $D_{\text{crit}} = 0.136$.
- Since $0.092 < 0.136$ do not reject the null hypothesis.

Kolmogorow Smirnov for Two Samples

Given two samples, test if their distributions are the same.

Compute the observed cumulative distribution functions of the two samples and compute their maximum difference.

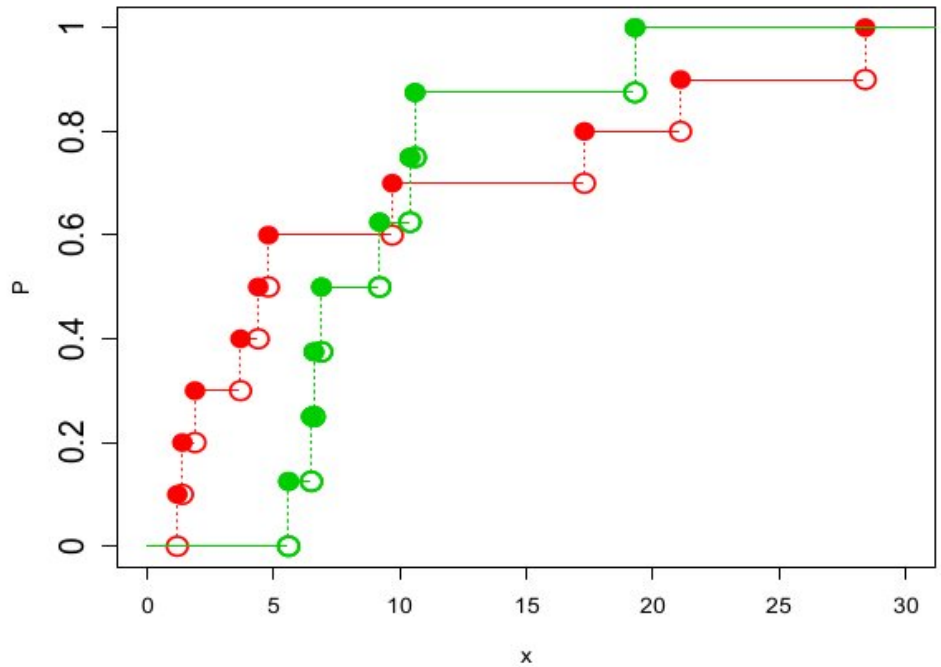
$$X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4$$
$$Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3.$$

- We sort the combined sample, in order to compute the empirical cdf's:

	1.2	1.4	1.9	3.7	4.4	4.8	5.6	6.5	6.6	6.9
F_x	0.1	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.6	0.6
F_y	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.5

	9.2	9.7	10.4	10.6	17.3	19.3	21.1	28.4
F_x	0.6	0.7	0.7	0.7	0.8	0.8	0.9	1.0
F_y	0.6	0.6	0.8	0.9	0.9	1.0	1.0	1.0

Kolmogorow Smirnov for Two Samples



The Kolmogorov-Smirnov statistic is again the maximum absolute difference of the two observed distribution functions. Here

$$D_n = 0.6.$$

Kolmogorow Smirnov for Two Samples

For two samples, the 95% critical value can be approximated by the formula:

$$D_{\text{crit},0.05} = 1.36\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}.$$

In our case $n_x = 10$ and $n_y = 8$ and thus $D_{\text{crit}} = 0.645$.

So we retain the null hypothesis.

Power of tests

Decision \ Truth	H_0 True	H_0 False
Don't Reject H_0	Correct (Prob. $1-\alpha$)	Type II Error (Prob.=β)
Reject H_0	Type I Error (Prob.=level=α)	Correct (Prob.=Power=$1-\beta$)

The power of a test, $1-\beta$, is the probability of rejecting $\mathbf{H_0}$ when it is false.

The power of a test always depends on the alternative: it is the probability of rejecting the null assuming a **specific alternative** is true.

Computing the Power of a test

Consider n observations from a normal distribution with unknown mean μ and known variance σ^2. Test

$$\mathbf{H}_0 : \mu = \mu_0, \quad \text{against} \quad \mathbf{H}_1 : \mu = \mu_{\text{alt}}.$$

The power is the probability of rejecting the null at the $(1-\alpha)\%$ confidence level when H_1 is true.

$$\begin{aligned}
\text{Power} &= \mathbb{P}\Big(|Z| > z_{1-\alpha/2} \mid \mathbf{H}_1\Big) \\
&= \mathbb{P}\Big(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \mid \mathbf{H}_1\Big) + \mathbb{P}\Big(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} \mid \mathbf{H}_1\Big) \\
&= \mathbb{P}\Big(\frac{\bar{X}-\mu_{\text{alt}}}{\sigma/\sqrt{n}} > z_{1-\alpha/2} + \frac{\mu_0-\mu_{\text{alt}}}{\sigma/\sqrt{n}} \mid \mathbf{H}_1\Big) \\
&\quad + \mathbb{P}\Big(\frac{\bar{X}-\mu_{\text{alt}}}{\sigma/\sqrt{n}} < -z_{1-\alpha/2} + \frac{\mu_0-\mu_{\text{alt}}}{\sigma/\sqrt{n}} \mid \mathbf{H}_1\Big) \\
&= \Phi\Big(-z_{1-\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu_{\text{alt}})\Big) + 1 - \Phi\Big(z_{1-\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu_{\text{alt}})\Big)
\end{aligned}$$

Power

Set $\mu_0 = 0$ and let us have a look at the power as a function of μ_{alt}.

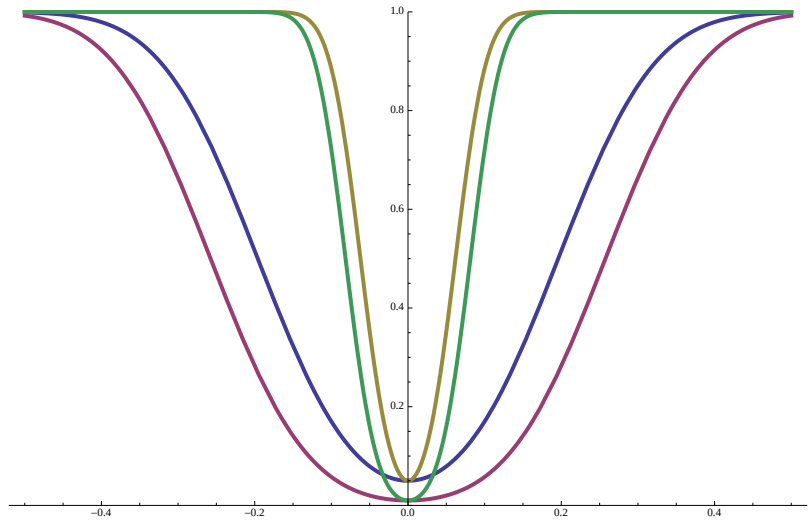


Figure: Power as a function of μ_{alt} for $n = 100, \alpha = 0.01$, $n = 100, \alpha = 0.05$, $n = 1000, \alpha = .01$, $n = 1000, \alpha = .05$.

The power increases with the number of samples and as μ_{alt} gets farther from μ_0. As $\mu_{\text{alt}} \to \mu_0$, the power approaches the level of the test.

Lowering the level of the test also decreases the power.

Choosing Trial Sizes: An example

- Suppose that drug A has been found to lower blood pressure on average by $\mu_A = 10$mmHg with standard deviation $\sigma_A = 8$ mmHg.
- Suppose you want to compare it to a new drug B with unknown mean effects μ_B. And $\sigma_B = \sigma_A = \sigma$.
- Recruit subjects, randomly split them in two groups to receive A and B, respectively.
- Test

$$\mathbf{H}_0 : \mu_A = \mu_B, \qquad \mathbf{H}_1 : \mu_A < \mu_B.$$

Choosing Trial Size

- Then the z-statistic is

$$z = \frac{\bar{b} - \bar{a}}{\sigma\sqrt{\frac{2}{n} + \frac{2}{n}}}.$$

- Is it worth doing the test? *What is our chance of picking up a difference between μ_A and μ_B.*
- The power is

$$\Phi\Big(-1.96 + \frac{\sqrt{n}}{\sigma}(\mu_A - \mu_B)\Big) + 1 - \Phi\Big(1.96 + \frac{\sqrt{n}}{\sigma}(\mu_A - \mu_B)\Big).$$

- If $\mu_\mathrm{A} - \mu_B$ halves, we need a sample size four times as large to keep the power fixed.

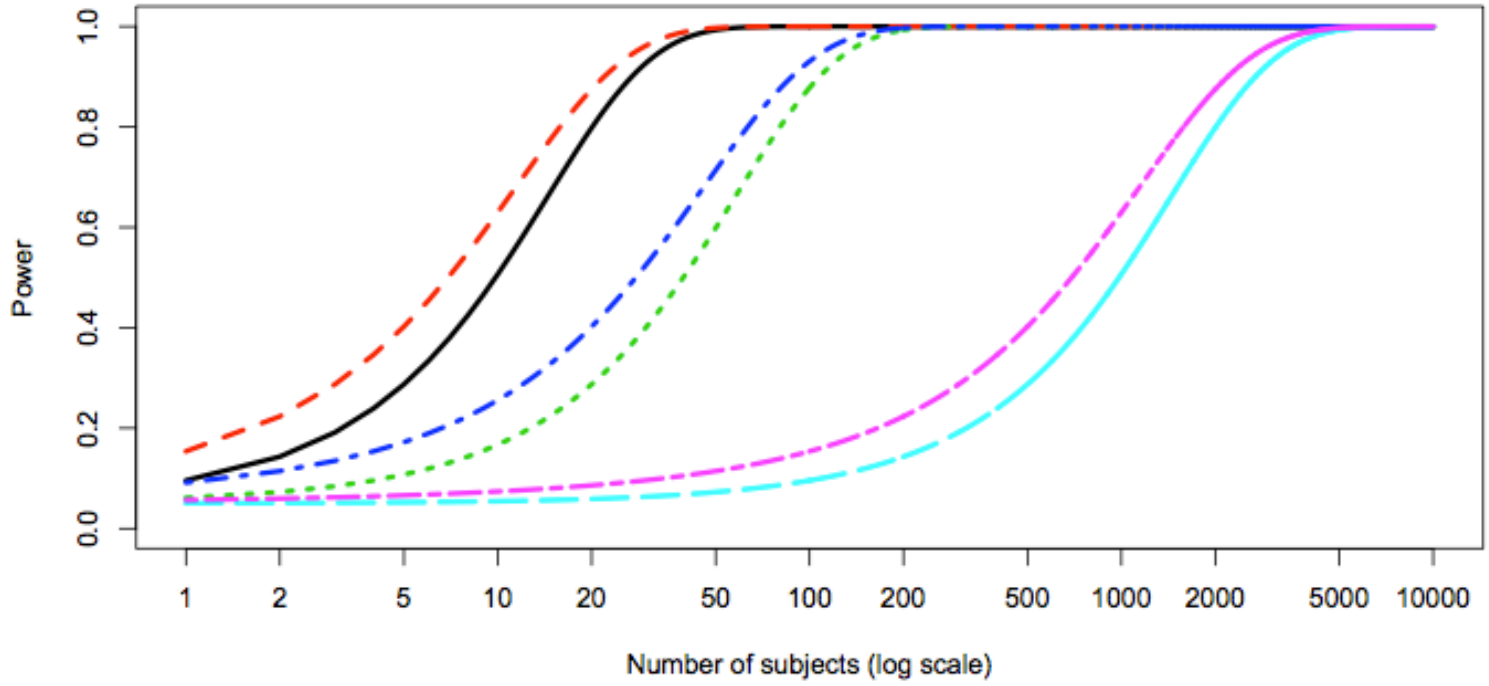


Figure: Plot of the power against the logarithm of the sample size for $\mu_B - \mu_A = 10, 5$ and 1 mmHg.

- One tailed test $\mu_B - \mu_A = 10$;
- one-tailed test $\mu_B - \mu_A = 5$;
- one-tailed test $\mu_B - \mu_A = 1$;
- two-tailed test $\mu_B - \mu_A = 10$;
- two-tailed test $\mu_B - \mu_A = 5$;
- two-tailed test $\mu_B - \mu_A = 1$.

Discussion

- The one-tailed test is more powerful *when* $\mu_B - \mu_A$ *is on the right side.*
- If $\mu_B - \mu_A$ is on the wrong side, it is practically useless.
- If we can afford up to 50 subjects and we think we should only do the test if we have at least 80% chance of finding a significant result then we should only go ahead if we expect a difference of at least 5mmHg.
- If we can afford 200 subjects, then we can go ahead if we expect difference of 2.5 mmHg.
- With 1000 subjects we still have a good chance of picking up differences as small as 1mmHg.

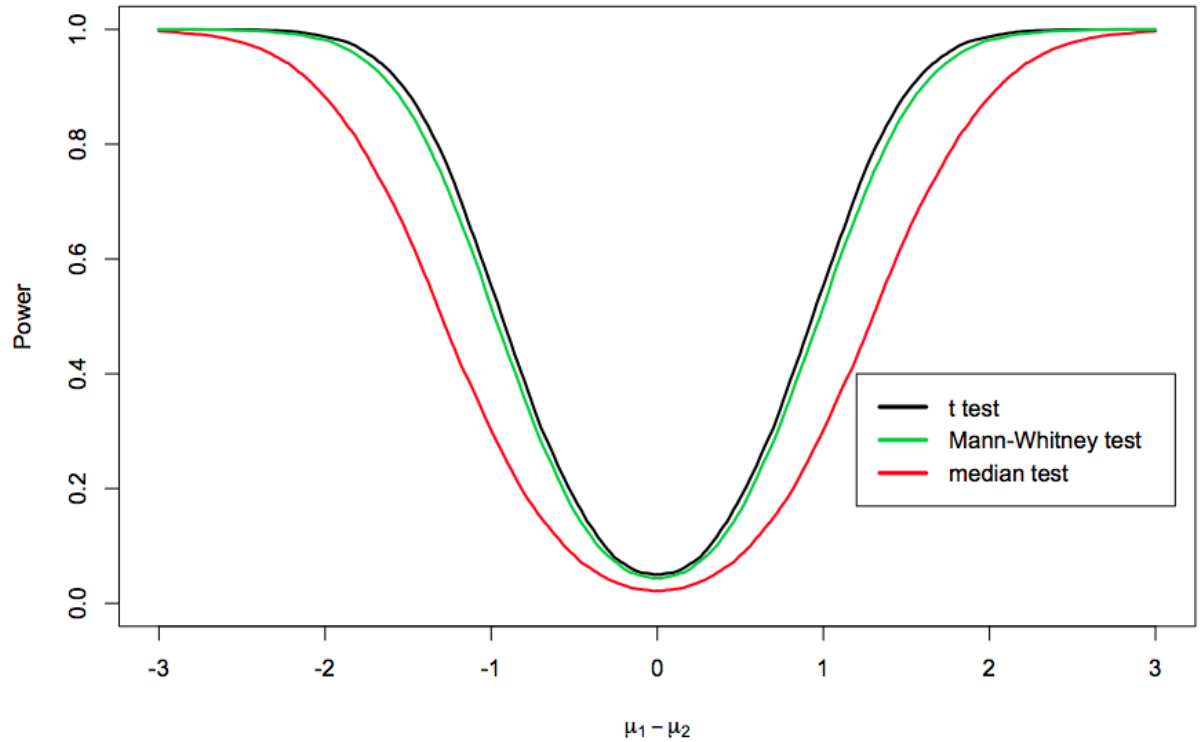
Power and Non-Parametric Tests

- Non-parametric tests are less powerful than parametric ones (if their assumptions are satisfied).
- Example: Observe 10 samples from $N(\mu, 1)$. Suppose both mean and variance unknown.
- Test at the 0.05 level

$$\mathbf{H}_0 : \mu = 0, \quad \text{vs} \quad \mathbf{H}_1 : \mu \neq 0.$$

- How does the power of the t-test compare to that of the median test or the rank-sum test?

Power and Non-Parametric Tests



Here you can see that the Mann-Whitney test actually does fairly well compared to the t-test.
On the other hand the median test is a clear loser!