Statistical Methods for Astronomers

Lies,

Dammed Lies

and Statistics

Some Details

- **Lecturers:**
  - Russell Shipman (x7753): russ@sron.nl :ZG 276
  - Saleem Zaroubi (x) :saleem@astro.rug.nl :ZG 282
- **Course Times:**
  - Lecture: Tuesday: 11:15 – 12:45
  - Lecture: Friday: 11:15-12:45
  - Werkcollege: Wednesdays or Thursdays for an hour
- **Final Exam**: somewhen 7^{th} to 25^{th} of April
- **Place:** ZG 161 for both lectures and exercises.

Resources

- Practical Statistics for Astronomers, J.V. Wall and C.R. Jenkins (ISBN 0-521-45616-9)
- Statistics in Theory and Practice, Robert Lupton, (ISBN 0-691-07429-1)
- Numerical Recipes, Press, Teukolsky, Vetterling, Flannery (ISBN 0-521-43064-X)
- Kapteyn computing facilities

Course Description

- Lecture and work assignments, expect some programming
- Final two weeks of course will be a project (written and presentation) Details will be given later.
- Evaluation: Final Exam 50%, Project 35%, Work assignments 15%

Why Statistics?

- What is the purpose of studying statistics at all?
- What are some examples?
- What role does probability play?

Statistics and probabilities are the basis for making decisions.

We use samples from our data combine them in some meaningful way and based on understanding of probability, we make an inference , i.e., draw a conclusion, make a decision.

Some Probability Distribution

- Define $F(x_0)$ as probability that a random variable x is $< x_0$. $F(-\infty) = 0$ and $F(\infty) = 1$.
  - Probability density function $f(x) = \frac{dF}{dx}$
- $Pr(x \in x, x + dx) = F(x + dx) - F(x) = f(x)\,dx$
- Can have the probability of two variables, marginal distribution (integrate over undesired variable).

Probability Distributions

- Some common probability distributions
  - Uniform
  - Gaussian or Normal
  - Poisson
  - Binomial
  - Cauchy
  - Log-normal
  - Distributions which are derived from the Normal Distribution
    - χ^2
    - Student's t distribution

Uniform

- Very simple: something that even a computer can do
  - f(x) = 1 for x >0 and <1, 0 otherwise
  - Pseudo random numbers from a computer are uniformly distributed.

Normal

- Normal or Gaussian distribution:
  - Very common, the “work horse” of distributions
  - $f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
  - Also commonly noted as: $N(\mu, \sigma)$
  - Characteristic function:
  - $\Phi(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(x-\mu)^2}{2\sigma^2} + itx} dx = e^{it\mu - t^2\sigma^2/2}$
  - $\mu_1' = \mu \text{ and } \mu_2 = \sigma^2$

Multivariate Gaussian

- Cases when we have *n* random variables, where each follows a Gaussian distribution. They do not have to be independent. The distribution is:
  -
  - $f(\mathbf{x}) = \frac{1}{|V|^{1/2}(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T V^{-1}(\mathbf{x}-\mu)\right]$.nce matrix. It is symmetric and positive definite with elements:
  -
  - $V_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$ Diagonal elements are variances, off --> covariance

Log Normal

- If x follows an N(0,1) distribution, and $x = \gamma + \delta \ln y$ then y follows a log normal distribution, given by:
  -
  - $dF = \frac{\delta}{\sqrt{2\pi}} e^{-(\gamma+\delta \ln y)^2} dy/y$ omical processes might be of interest here?

Poisson Distribution

- Counting probabilities of rare events. Probability of 1 event in time *t* is t/τ. What is the probability of breaking exactly n in t+dt?
  - $p_n(t) = p_{n-1}(t)\frac{dt}{\tau} + p_n(t)(1 - \frac{dt}{\tau})$
  - Remember AND is the product of provabilities OR is the sum.
  - $prob(\text{A and B}) = prob(\text{A})prob(\text{B}) \quad prob(\text{A or B}) = prob(\text{A}) + prob(\text{B})$
  - Total prob of exactly n is prob of n-1 AND one more OR the prob of n AND NOT one more.

Poisson Continued

- Simplifying
  - $\frac{dp_n}{dt} = \frac{1}{\tau}(p_{n-1} - p_n)$
  - Note p_n could be a complete derivative if factor $e^{t/\tau}$
  - Then, $\frac{d}{dt}(p_n e^{t/\tau}) = \frac{1}{\tau} p_{n-1} e^{t/\tau}$
  - Let, $\pi_n = p_n e^{t/\tau}$ then $\frac{d\pi_n}{d(t/\tau)} = \pi_{n-1}$
  - So, $\pi_n = \frac{(t/\tau)^n}{n!}\pi_0$
  - Solve for p_0, and finally get $p_n = \frac{\mu^n}{n!}\exp^{-\mu}$

Poisson Final

- Show that this is normalized (n from 0 to infinity).
- Find the characteristic function (again summing from 0 to infinity)
- What is the mean? variance?

Binomial

- Processes with only two outcomes (A or B) with probabilities of p and q (p+q =1), carry out the processes n times then the chance of getting r A's and n-r B's is
- $P(r) = \binom{n}{r} p^r (1-p)^{n-r}$
- Where
- $\binom{n}{r} = \frac{n!}{r!(n-r)!}$
- Show that the mean and variance are $\mu = np$ $\sigma^2 = np(1-p)$

And the rest....

- Cauchy: $f(x) = \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}$
- χ^2_n Distribution. Results as the sum of squares of N(0,1) deviates.
  - $f(X^2) = \frac{1}{2^{n/2-1}(n/2-1)!}e^{-X^2/2}X^{n-1}$
  - n is the number of degrees of freedom. Mean n, variance 2n.

More on Probability

- Independent Events: defined if the probability of one does not influence the probability of the other.
  - $prob(\mathrm{A\ and\ B}) = prob(\mathrm{A})prob(\mathrm{B})$
- If not independent...Conditional
  - $prob(\mathrm{A}|\mathrm{B}) = \frac{prob(\mathrm{A\ and\ B})}{prob(\mathrm{B})}$
- For several possibilities of event B, B_1, B_2 ...
  - $prob(\mathrm{A}) = \Sigma_i prob(\mathrm{A}|\mathrm{B}_i)\mathrm{B}_i$
  - Summing over a series of possible events for which we don't care--*marginalization*

And Bayes

- Simple equality prob(A and B) = prob(B and A)
  - $prob(\mathrm{B}|\mathrm{A}) = \frac{prob(\mathrm{A}|\mathrm{B})prob(\mathrm{B})}{probA}$
- Power in interpretation:
  - prob(B|A) --- posterior (state of belief after data)
  - prob(A|B) --- likelihood of getting A, given B
  - prob(B) --- prior (state of belief before data)
  - prob(A) --- normalization

Use of Bayes Theorem

- Result of Theorem is a probability distribution (over all outcomes). Choose the peak, or range, ...
- Allows us to make inferences about our Model given the data.

Example

- Balls in Urn. N red, M white, total number N + M=10
- Draw 3 times (three Tries) and put back, We get 2 Reds.
- Find the most probable number of Red balls in the urn.

Example cont

- Likelihood is Binomial $P(r) = \binom{n}{r} p^r (1-p)^{n-r}$
  - n tries
  - r successes
- Posterior probability =

Priors

- Not always obvious to choose a prior (to realize what we understand/believe before an experiment).
  - Knowing nothing might imply a uniform prior (all outcomes equally likely)
  - And others....
- Calculating probabilities of probabilities.

How to use Bayes Theorem

- Find “Best” parameters of a model which is related to Maximum likelihood method.
- Knowing posterior probability may be the goal (comparison with theory or expectations).
- Use to help understand experimental results in terms of what we know.
- Try out Exercises 2.3 and 2.4

Central Limit Theorem

- Averages of Repeated Draws of samples forms a Normal Distribution.
  - Distribution must have a finite mean and variance.
  - Form of distribution does not matter.
- Very powerful: averaging gets you to a Normal (well understood) distribution.

Statistics and their distributions

- What is a statistic?
  - Description, summary of data.
  - Combination or mathematical function applied to data.
  - Made from FINITE data
  - Attempt to uncover the equivalent Expectation Value without infinite data. (Mode, Median, Mean, Variance, etc..

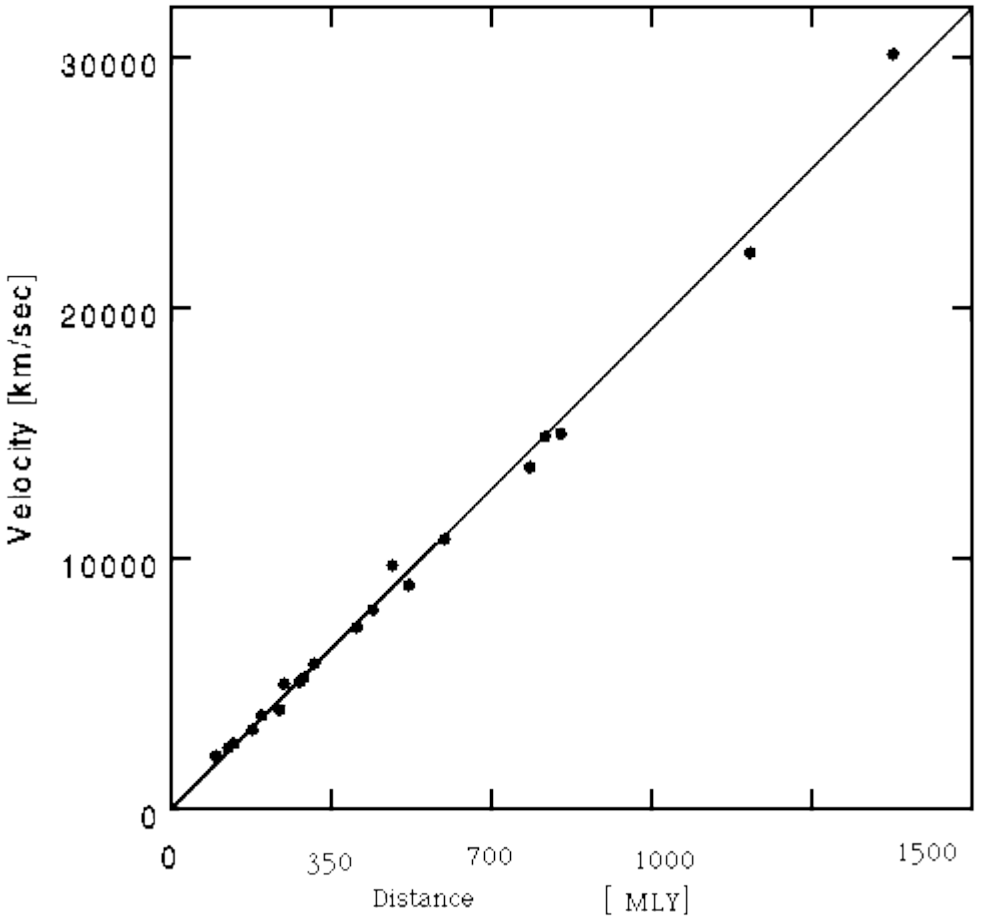
Properties of a Good Statistic

- Unbiased: Expectation value of statistic is expectation value of parent distribution
  - Average is an unbiased estimate of the mean
  - Standard deviation is a *biased* estimator. Referred to as sample standard deviation
- Consistent: Gives the same value regardless of sample size
- Closeness: smallest possible deviation from parent Expectation value
- Robust: low influence by outliers. Median vs average.

Statistics and Their Distributions

- Average: Normally distributed about μ with σ^2/N variance
- Sample variance $\sigma^2{}_s$: $\sigma^2\chi^2/(N-1)$
  - For χ^2 of N-1 degrees of freedom.
- Student-t with N-1 degrees of freedom
  - $\dfrac{\sqrt{N}(\bar{x}-\mu)}{\sigma_s}$
- Ratio of two sample (sizes M and N) variances follows *F* distribution (function tabulated for specific values of M and N)

Correlations



Correlations: Bivariate Gaussian

- Multivariate Gaussian distribution allows for dependent variable through the covariance
- $$f(\mathbf{x}) = \frac{1}{|V|^{1/2}(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T V^{-1}(\mathbf{x}-\mu)\right]$$
- $$V_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$$
- For only two variables (a Bivariate Gaussian) this simplifies to
- $$prob(x, y|\sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right]\right\}$$

Estimator of Correlation Coefficient

- ρ Is known as the Pearson Correlation Coefficient.
- It's estimator is
- $$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$
-
- With standard deviation: $\sigma_r = \frac{(1 - r^2)}{\sqrt{N - 1}}$
- Calculate $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ which follows Student's-t N-2 degrees of freedom distribution

How to use it, Frequentist Approach

- Calculate probability of data given correlation
- $$prob(r|\rho, H)$$
- Where *H* is the Hypothesis of a correlation and try to reject *H* in some comfortable confidence level. Choose an easy *H == null hypothesis of no correlation.* Calculate the probability under *H* that r can be as large or larger. If the prob is very small, reject *H.*

The Bayesian Approach

- Calculate the *posterior probability.*
- $$prob(\rho, \sigma_x, \sigma_y, \mu_x, \mu_y | \mathrm{data})$$
- Where the extra parameters are details about the bivariate Gaussian we assumed at the very beginning.
- However, we don't really care about these, so *marginalize* them out.
- The result is a probability distribution $prob(\rho|\mathrm{data})$
- This actually answers the question we asked in the first place.

Some Words of Caution

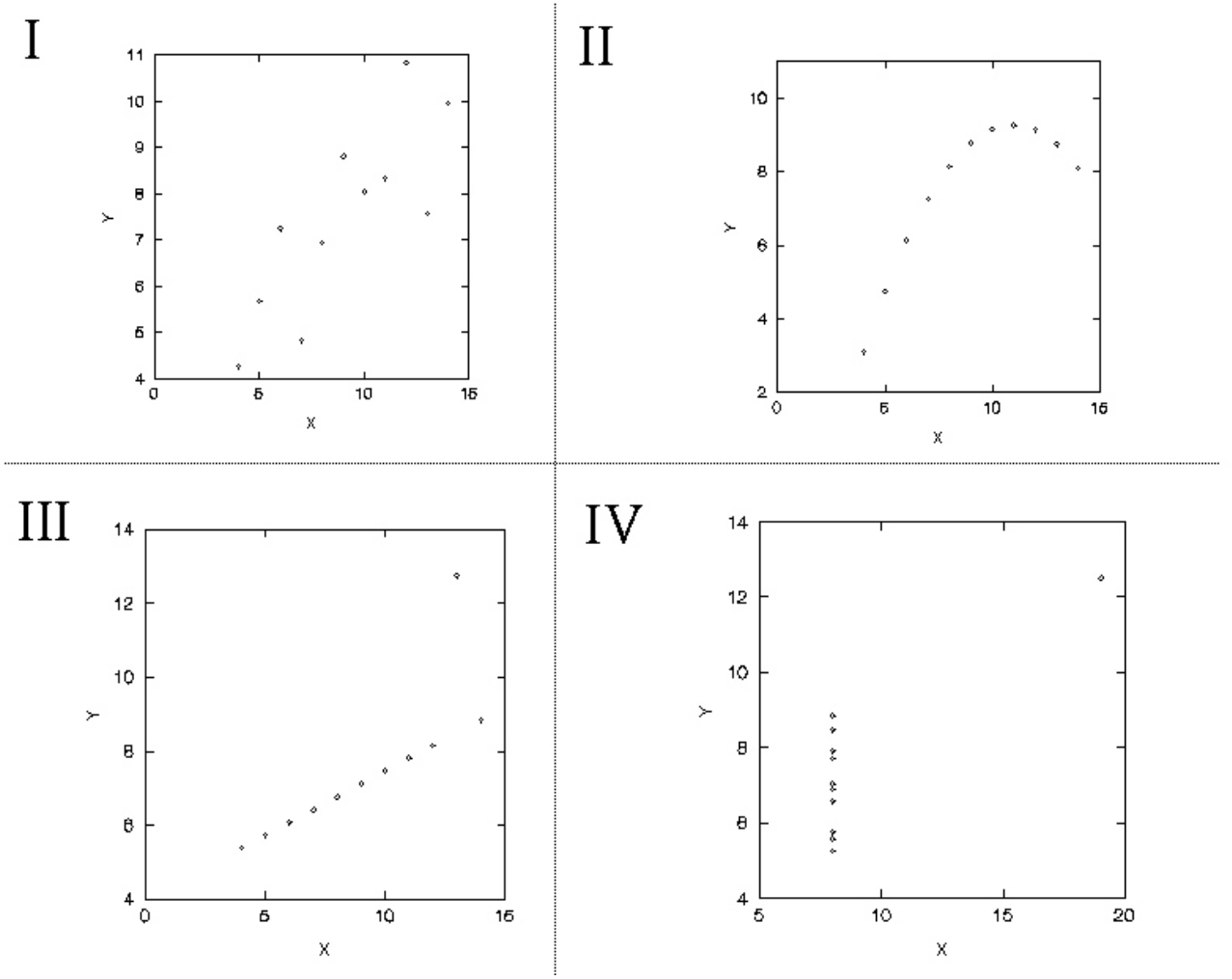
- What was the question? Why correlation testing?
  - The Fishing Trip?
  - Rule of thumb: is correlation still present after removal of 10% of points?
  - Hidden third variable
- Non-Parametric Statistics: there is another possibility
- Anscombe's quartet

Non-Parametric Correlation Testing

- Heavy reliance on assumed Bivariate Gaussian.
- Can correlate ranks (the order in which values occur).
- Calculate the Spearman Rank coefficient
- $$r_s = 1 - 6\frac{\sum^N (X_i - Y_i)^2}{N^3 - N}$$
- Where X, and Y are ranks of variables x and y
- Hypothesis testing (classical approach): null ==no correlation.
- Choose level of confidence, calculate r_s, look up value in table, if larger than critical value reject H_0

Anscombe's Quartet

- Graphs, graphs, graphs
- All with identical: coefficients, regression lines, residuals in Y, estimated standard errors in slopes.



Confidence Intervals

- Classical Point of View.
  - Probability of a value as large as x or larger:
  - $P(x > x_0) = \int_{x_0}^{\infty} f(x)dx$
- For a normal distribution: 95% of the probability is:
  - $\int_{\mu-1.96\sigma}^{\mu+1.96\sigma} N(\mu, \sigma)dx = 0.95$
- What is the meaning “2 sigma” confidence?
- Is this really the question we wanted to answer.

Simple Example

- We know a certain measurement process results in a Normal distribution: $N(\mu, \sigma)$
- We measure data, which we think might be the result of this process. What do we do?
  - Decide on a confidence level (comfort level?) where we would stack our reputation....
  - Ask, does our measurement fall with in this range or not?
  - Stake the claim or otherwise, don't.

Hypothesis Testing

- The Null Hypothesis and Alternative
  - Classical question. Distributions are calculated assuming the Null Hypothesis, where the only other option is the alternative. Choose a level of significance we are willing to reject the Null Hypothesis (ie. Make a conclusion based on a false negative). Calculate the test statistic. Evaluate from known distribution tables.
- Type I Error (False Negative) and Type II (False Positive) Error

Student-t test

Table of t-statistics

df	P = 0.05	P = 0.01	P = 0.001
1	12.71	63.66	636.61
2	4.30	9.92	31.60
3	3.18	5.84	12.92
4	2.78	4.60	8.61
5	2.57	4.03	6.87
6	2.45	3.71	5.96
7	2.36	3.50	5.41
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
11	2.20	3.11	4.44

- Test for difference between two means. Are my data drawn from the same (Normal) distribution?
- $$test = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{n_x {s_x}^2 + n_y {s_y}^2}{n_x + n_y - 2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$
- Note the practical difficulties for very small samples.
-

F test

- Test for whether two variances are the same. Take ratio of standard deviations
  - $$f = \frac{\Sigma(x_i - \bar{x})^2/(n_x - 1)}{\Sigma(y_i - \bar{y})^2/(n_y - 1)}$$
- Follows the F (n_x-1, n_y-1) distribution.

Table of F-statistics P=0.05

df2\df1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	22	24	26	28	30	35	40	45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66	8.65	8.64	8.63	8.62	8.62	8.60	8.59	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80	5.79	5.77	5.76	5.75	5.75	5.73	5.72	5.71
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.54	4.53	4.52	4.50	4.50	4.48	4.46	4.45
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.86	3.84	3.83	3.82	3.81	3.79	3.77	3.76
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44	3.43	3.41	3.40	3.39	3.38	3.36	3.34	3.33
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15	3.13	3.12	3.10	3.09	3.08	3.06	3.04	3.03
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94	2.92	2.90	2.89	2.87	2.86	2.84	2.83	2.81
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.75	2.74	2.72	2.71	2.70	2.68	2.66	2.65
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.63	2.61	2.59	2.58	2.57	2.55	2.53	2.52
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54	2.52	2.51	2.49	2.48	2.47	2.44	2.43	2.41
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.44	2.42	2.41	2.39	2.38	2.36	2.34	2.33
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.37	2.35	2.33	2.32	2.31	2.28	2.27	2.25
	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.31	2.29	2.27	2.26	2.25	2.22	2.20	2.19

F Test continued

- Reject both large and small values.
- Assumptions / Notes
  - The larger variance should always be placed in the numerator
  - The test statistic is F = s1^2 / s2^2 where s1^2 > s2^2
  - Divide alpha by 2 for a two tail test and then find the right critical value
  - If standard deviations are given instead of variances, they must be squared
  - When the degrees of freedom aren't given in the table, go with the value with the larger critical value (this happens to be the smaller degrees of freedom). This is so that you are less likely to reject in error (type I error)
  - The populations from which the samples were obtained must be normal.
  - The samples must be independent

Non-Parametric Tests

- Both F and, to a lesser extent, Student-t tests depend on the parent populations being Normal.
- They also assume significant amounts of data.
- What if the data are very sparse?
- How much faith do you have in the process which created your data to follow a Normal distribution? Was there a great deal of averaging involved?

Chi Square Test

Table of Chi-square statistics

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26

- A given model predicts number of results within a certain range (bin).
- An observation measures these results (how many times do the observations fall within a given bin).
- Form the Chi Square:
- $$\chi^2 = \sum_{kbins} \frac{(D_k - NP_k)^2}{NP_k}$$
-

Chi Square Test cont.

- Number of observations within a bin follows Poisson statistics.
- Bins must be chosen to contain roughly same number of data points. Should not contain fewer than 5.
  - Bins can be adjusted.
  - Putting data into bins reduces “resolution” i.e. Hides details within the bins.
  - Note there are no assumptions about the underlying distribution.
  - This can be used to reject or accept the null hypothesis.

Kolmolgorov-Smirnov

- Test whether a sample distribution of points f(x) follows an expected distribution s(x).
- Calculate the Cumulative Distribution of f and s (F and S_n) where n is used to normalize the expected distribution.
- Choose your confidence level
- Calculate the statistic: just different $D_n = \mathrm{Max}|S_n(x) - F(x)|$
  - Or${D_n}^+ = \mathrm{Max}(S_n(x) - F(x))$ or ${D_n}^- = \mathrm{Min}(S_n(x) - F(x))$
  - Look up in a table (based on the number of points n, whether the value of D is larger than the critical value, if so reject the null hypothesis.

Critical Values for KS One Sample test

Kolmogorov-Smirnov One-Sided Test

n	0.1	0.05	0.025	0.01	0.005
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

- for two sided, double the confidence level.

And use the same table.

Kolmogorov-Smirnov Two Sample Test

- One can also use the KS to test whether two samples have come from the same distribution.
- The idea is the same as before, Calculate the joint cumulative distribution
  - $$D_{n,n'} = Max|F_n(x) - F_{n'}(x)|$$
  -

Critical Values for KS Two sample

Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided)

Table gives critical D-values for $\alpha = 0.05$ (upper value) and $\alpha = 0.01$ (lower value) for various sample sizes. * means you cannot reject H_0 regardless of observed D.

$n_2 \backslash n_1$	3	4	5	6	7	8	9	10	11	12
1	* *	* *	* *	* *	* *	* *	* *	* *	* *	* *
2	* *	* *	* *	* *	* *	16/16 *	18/18 *	20/20 *	22/22 *	24/24 *
3	* *	* *	15/15 *	18/18 *	21/21 *	21/24 24/24	24/27 27/27	27/30 30/30	30/33 33/33	30/36 36/36
4		16/16 *	20/20 *	20/24 24/24	24/28 28/28	28/32 32/32	28/36 32/36	30/40 36/40	33/44 40/44	36/48 44/48
5			* *	24/30 30/30	30/35 35/35	30/40 35/40	35/45 40/45	40/50 45/50	39/55 45/55	43/60 50/60
6				30/36 36/36	30/42 36/42	34/48 40/48	39/54 45/54	40/60 48/60	43/66 54/66	48/72 60/72
7					42/49 42/49	40/56 48/56	42/63 49/63	46/70 53/70	48/77 59/77	53/84 60/84
8						48/64 56/64	46/72 55/72	48/80 60/80	53/88 64/88	60/96 68/96
9							54/81 63/81	53/90 70/90	59/99 70/99	63/108 75/108
10								70/100 80/100	60/110 77/110	66/120 80/120
11									77/121 88/121	72/132 86/132
12										96/144 84/144

For larger sample sizes, the approximate critical value D_α is given by the equation

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

- KS works for very small distributions

Fisher Exact Test

- Test of non random associations, between two small samples which fall into two mutually exclusive bins.
  - Example number of men or women in the class which do or do not bike to class.
  - Null hypothesis is that the assignment of scores is random.
  - calculate

Sample	man	woman
rides bike	A	C
does not ride	B	D

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

Chi Square Two sample or k sample test

- Test that k samples come from the same population.
- Similar to One sample test. Same comments about bins.
- Calculate:
  - Where $E_{ij} = \frac{\Sigma_k O_{ij} \Sigma_r O_{ij}}{\Sigma_r \Sigma_k O_{ij}}$
  - (r-1)(k-1) d. of f.

sample j =	1	2	3
Bin I =1	O11	O21	O31
2	O12	O22	O32
3	O13	O23	O33
4	O14	O24	O34
5	O15	O25	O35

$$\chi^2 = \Sigma_{i=1}^{r} \Sigma_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Wilcoxon Mann-Whitney U test

- Test whether two distributions have the same location. Sometimes called the rank sum test.
  - Test whether sample A is stochastically larger than B
  - B larger than A
  - A and B differ
- Rank combination of all samples keeping membership in tacked. Sum the A rankings, to get U_A and B for U_B.
- Null Hypothesis is that the two distributions come from the same population.

Critical Values for U test

Critical values of R for the Mann-Whitney rank-sum test

The pairs of values below are approximate critical values of R for two-tailed tests at levels $P = 0.10$ (upper pair) and $P = 0.05$ (lower pair).

(Use relevant $P = 0.10$ entry for one-tailed test at level 0.05).

		larger sample size, n_2						
		4	5	6	7	8	9	10
smaller sample size n_1	4	12,24 11,25	13,27 12,28	14,30 12,32	15,33 13,35	16,36 14,38	17,39 15,41	18,42 16,44
	5		19,36 18,37	20,40 19,41	22,43 20,45	23,47 21,49	25,50 22,53	26,54 24,56
	6			28,50 26,52	30,54 28,56	32,58 29,61	33,63 31,65	35,67 33,69
	7				39,66 37,68	41,71 39,73	43,76 41,78	46,80 43,83
	8					52,84 49,87	54,90 51,93	57,95 54,98
	9						66,105 63,108	69,111 66,114
	10							83,127 79,131

- Two tailed: distributions differ