If you have $x_1...x_n$ IIDRV then how does one normalize? $S = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sqrt{n}\sigma}$ where $\mu = E(x_i)$ and $\sigma = \sqrt{V(x_i)}$

Expectation might be infinite or undefined. Distance over positive integers is $p_i = \frac{6}{\pi^2} \frac{1}{i^2}$ so over all $p_i$ we get $\sum_{i=1}^{\infty} p_i = 1$ So the expectation value is $E(x) = \sum_{i=1}^{\infty} i \cdot p_i = \frac{6}{\pi^2} \sum_{i=1}^{\infty} \frac{1}{i^2} = \infty$.

How about all integers?

$p_i = \frac{3}{\pi^2} \frac{1}{i^2}$ so the expectation value is then $E = \sum_{i=1}^{\infty} i \cdot p_i + \sum_{i=1}^{\infty} (-i \cdot p_i) = \frac{1}{2} \left[ \sum_{i=1}^{\infty} \frac{1}{i^2} - \sum_{i=1}^{\infty} \frac{1}{i^2} \right] = \frac{1}{2}(\infty - \infty) = ?$ So we could get any number, and likely expectation value does not converge. Thus some probabilities are just ill behaved if this is the case then we can’t do anything.

There can be cases where expectation is defined but variance is infinite.

So, say we have distribution over non-zero integers then $p_i = \frac{1}{2} \left[ \frac{1}{i^2} \right]$ so then $E(x) = \sum_{i=1}^{\infty} i \cdot p_i = \frac{1}{2} \left( \sum_{i=1}^{\infty} \frac{1}{i^2} - \sum_{i=1}^{\infty} \frac{1}{i^2} \right) = 0$. The variance is then $V(x) = E(x^2) - E(x)^2 = E(x^2) = \sum_{i=1}^{\infty} i = -\infty i^2 \frac{1}{i^4} = \frac{1}{2} \sum_{i=1}^{\infty} i = -\infty \frac{1}{i^2} \rightarrow \infty$

Let’s talk about random variable with finite variance and expectation. However, a random variable is any distribution over a real line, so it can be a mixture of density functions and point masses. You can summarize the statistics using a histogram, but better is to use mean, variance, etc.

Central Limit Theorem

Let $x_1...x_n$ be IIDRV with finite mean and variance. So then $S_n = \sum_{i=1}^{n} x_i; Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$ so then $Z_n \xrightarrow{n \rightarrow \infty} \Phi(Z)$ the normal distribution $f(\zeta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \zeta^2}$ so $P(Z_n \leq x) \xrightarrow{n \rightarrow \infty} F(x) = \int_{-\infty}^{x} f(\zeta)d\zeta$

So for CDF with point masses you have $F_n(x)$ which is a sum of point masses but the central limit theorem says that $\liminf F_n(x) \rightarrow F(x)$. So, if you have large sample of random variables they will fit the normal distribution. Thus, the central limit theorem is a “small deviation” theorem and characterizes the probability of likely events.

However, an often more interesting question is “How unlikely is certain outcome?” This requires a large deviation theorem which gives an upper bound on the probability of unlikely events.

Sanov/Chernoff/Hoeffding bound

Let $x_1...x_n$ be IIDRV with bounded support ($\exists M p(|x| > M) = 0$). Then $P(\frac{1}{n} |S_n - \mu| > \epsilon) \leq e^{-n G(\epsilon)}$ where $G(\epsilon) > 0$

So many different $G(\epsilon)$ exist.

Sanor bound: let $x_1...x_n$ be binary IIDRV then $p(x_i = 1) = p$ and $p(x_i = 0) = 1 - p$ then $p_r(\frac{1}{n} S_n > q) \leq e^{-n D_{KL}(q||p)}$ where $q > p$

Let’s then call $\hat{p} = \frac{1}{n} S_n$ so then $p(\hat{p} > q) = p(e^{n\lambda\hat{p}} > e^{n\lambda q}) \forall \lambda > 0$ also we set $E(e^{n\lambda\hat{p}}) \geq e^{n\lambda q} p(e^{n\lambda\hat{p}} > e^{n\lambda q})$ (Markov inequality)

So then $E(x) \geq 0 \cdot p(x < a) + a \cdot p(X \geq a) = a \cdot p(x \geq a)$

We also get that $E(e^{n\lambda\hat{p}}) = E(\prod_{i=1}^{n} e^{\lambda x_i}) = \left[ E(e^{\lambda x_i}) \right]^n$
Chernoff/Sano/Hoeffding bounds

In case $x_1,...,x_m$ IID binary RV such that $p(x_i = 1) = p$ and $p(x_i = 0) = 1 - p$ so $S_m = \sum_{i=1}^m x_i$ and $\hat{p} = \frac{1}{m} S_m$

We want to bound probability that $p(\hat{p} > q) = \sum_{i=qm}^m p(S_m = i) = \sum_{i=qm}^m \binom{m}{i} p^i (1-p)^{m-i}$ where $q > p$  

How to generate this bound? Say $p(\hat{p} > q) = p(e^{m\lambda \hat{p}} > e^{m\lambda q})$ this is true because $p$ is strictly increasing monatonic.

Then, we use Markov bound to get $E(e^{mx \hat{p}}) \geq e^{m\lambda q}p(e^{m\lambda \hat{p}} > e^{m\lambda q})$. So to upper bound the right hand side (which is what we want), we will upper bound the left hand side. So, how do we do this?

$E(e^{mx \hat{p}}) = E(e^{m\lambda \frac{1}{m} \sum_{i=1}^m x_i}) = E(e^{\lambda \sum_{i=1}^m x_i}) = E(\prod_{i=1}^m e^{\lambda x_i}) = \prod_{i=1}^m E(e^{\lambda x_i}) = [E(e^{-\lambda x_i})]^m$

But, there can only be two values so $E(e^{\lambda x_i}) = pe^\lambda + (1-p)e^0 = pe^\lambda + (1-p)$ so plugging that back we get (by tuning lambda to what we want)

$E(e^{mx \hat{p}}) = [pe^\lambda + (1-p)]^m$ so back to markov bound:

$p(e^{m\lambda \hat{p}} > e^{m\lambda q}) \leq \left( \frac{pe^\lambda + (1-p)}{e^q} \right)^m$ so, to find the minimum, use derivative, and the answer is

$\lambda = \ln \frac{q(1-p)}{p(1-q)}$, plugging in this value of lambda we get that

$p(e^{m\lambda \hat{p}} > e^{m\lambda q}) \leq e^{-m \cdot KL(q||p)}$ and this is called the Sano Bound

This bound is not very convenient to work with sometimes. The other bounds can be derived in the following way:

1) $KL(q||p) \geq 2(q - p)^2$ →Hoeffding bound so $p(\hat{p} > q) \leq e^{-2m(q-p)^2}$

2) If $q > \hat{p} > p$ then $KL(q||p) \geq \frac{1}{3} \frac{(p-q)^2}{p}$ →Cherno bound so $p(\hat{p} > q) \leq e^{-\frac{1}{2} m \frac{(p-q)^2}{p}}$

and if $q < \hat{p}$ then $KL(q||p) \geq \frac{1}{2} \frac{(p-q)^2}{p}$ so $p(\hat{p} > q) \leq e^{-\frac{1}{2} m \frac{(p-q)^2}{p}}$

Null Hypothesis: Common Expectation “a priori” assumption

So, null hypothesis $\hat{p}_1 < p < \hat{p}_2$ this is a “truth” test

Statistical test (sample $m$ times) if $\hat{p} > q$. So $q$ is a statement of test, $p$ is a statement of truth.
What can you prove in statistics? This relates to trying in public life to disprove a claim - you can't disprove negative. In statistics, it is impossible to prove a positive, you can't prove ie that coin in fair. What you can do, is to disprove certain negatives - you can reject certain null hypotheses. So, you can make a statement “with significance of 99.9% the coin is in [0.49,0.51]” – you proved that the coin is not in the complement of that range, but it doesn’t prove that the coin is fair! Now, most of the time you don’t know the true value, but rather a bound on \( p \) value. You CAN calculate exact \( p \) values under assumption that you know the distribution ie normal, binomial...

Last time: Chernoff/Hoeffding/Anov bound on the binomial test.

Suppose you have a distribution over the real line (a density distribution \( f \)). Our null hypothesis could be to reject that our sample \( R \) is of distribution \( f \). We could try comparing the histograms, but how do you bin? What is better is to use the CDF and compare those. Then, \( p(x \leq t_1) = CDF(t_1) \) so then you can do comparisons for all points and compare \( x_j \) of CDF to \( t_j \) of your true sequence, then then get \( b_j \) so you get a sequence of 0’s and 1’s. Then, we can compare the gaps and the probability of them being same. But how many test should we do (how many different \( j \)’s)? Let’s say we have tests and only 1 of them rejects (significance of each test is say 2%), then with what significance can we say that we reject?

Test is \( p\left\{ \frac{\#(x \leq t_1)}{m} - CDF(t_1) > 0.1 \right\} \leq 2\%

But one of the test rejects the null hypothesis \( k \cdot p_j \), if you try, try, and try to not fit a model, then you will eventually find a test that misleads (union bound). However, this is a naive analysis.

What is the ultimate test?

Glivenko-Cantelli theorem

Assume that \( x_1, x_2, \ldots, x_n \) are IID RV.

Define the empirical CDF \( F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x)}(x_i) \). So \( I_{(-\infty,x)}(x_i) = \begin{cases} 1 & x_i \leq x \\ 0 & \text{otherwise} \end{cases} \)

The theorem is then that \( \sup_{x \in \mathbb{R}} |F(x) - F_n(x)| \to 0 \) as \( n \to \infty \) with probability one.

The DKW theorem then gives a specific bound:

\[
p\left( \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}
\]

Supremum of a set is the smallest upper bound. So \( \sup \to \min(\theta|\theta \geq x\forall x) \)

So, how do we use this? Use Kolmogorov-Smirnov test. These are universal tests and can identify the difference between any two distributions over \( \mathbb{R} \).

There are two tests in matlab:

**kstest**: you have sample + CDF, null hypothesis of a particular distribution over \( \mathbb{R} \) (defined using a CDF). This function is to reject null hypothesis \( \max(F_n(x) - F(x)) > \epsilon \) and use DKW to get the \( p \) value.

Read Chapter 10.

**kstest2**: 2 samples. null hypothesis that they are drawn from the same distribution. So, reject if max distance between \( \max |F^n_1(x) - F^n_2(x)| > \epsilon \).

An example of using this: “If I give people this drug, then it has an effect on their weight. Null hypothesis would be that there is no effect. So, compare distribution of weights who are taking and not taking the drug.” So it doesn’t tell you what the effect is, but it does tell you if there is ANY effect.
Denote by $p$ the probability of “failure”.

Let $x_1...x_n$ be IID RV where $x_i$ denotes whether instance $i$ failed. So $x_i = \begin{cases} 1 & \text{if instance failed} \\ 0 & \text{if instance not failed} \end{cases}$ and $P(x_i = 1) = p$ while $P(x_i = 0) = 1 - p$.

How large do we need to set $n$ so that assuming NO FAILURE $p < \epsilon = 1\%$ with confidence $1 - \delta = 99\%$ (if say $\delta = 1\%$).

Answer:

Then, the probability of 0 failures (hypothesis) is $(1 - \epsilon)^n < \delta$ so

\[ n \cdot \ln(1 - \epsilon) < \ln \delta \]

\[ n \ln \frac{1}{1-\epsilon} > \ln \frac{1}{\delta} \]

\[ n > \frac{\ln \frac{1}{\delta}}{\ln \frac{1}{1-\epsilon}} \]

Harder question:

How large do we need to set $n$ and how small should # of failures over $n$ be so that we know that $p < \epsilon$ with confidence $1 - \delta$. If $q > \frac{\# \text{ failures}}{n}$

Answer:

\[ \sum_{i=0}^{q} \binom{n}{i} (1 - \epsilon)^{n-i} \leq \delta \]

\[ P(\tilde{p} < q) \leq e^{-\frac{1}{2}n(\frac{q-n}{n})^2} \leq \delta \]

so chernoff bound

Mechanics:

ttest

ttest2

chi2gof - chi squared goodness of fit

So,

ttest:

Null hypothesis: Normal distribution with mean $E(x) = \mu$ and unkown $\sigma$. Alternative: Normal dist with $E(x) \neq \mu$.

How it works:

\[ \frac{\bar{x} - \mu}{\sigma} = \frac{\frac{\sum(x_i - \mu)}{\sqrt{\sum(x_j^2 - \frac{\sum x_j}{n})}}} \text{ the answers to this should be distributed as student t-distribution.} \]

Paired t-test. A sample $(x_1, y_1)...(x_n, y_n)$ and what you are interested in $(x_1 - y_1)...(x_n - y_n)$. so ttest $(x, y) = ttest(x - y)$

ttest2 =

Chi test

\[ \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \] this has a known distribution then we can calculate where to reject.

$h = \text{chi2gof}(x)$ Null hypothesis x is normally distributed h=1 if rejected with $\alpha = 5\%$
reviews....
11/03/08
more midterm review
11/05/08
Midterm postponed / conditional expectation/variance
11/10/08
MIDTERM
______________
When you have a symmetric matrix it can be decomposed as
For symmetric matrixies $M = A^TDA$ where $D$ is diagonal. So, operation $A^TDAa$ means transform $a$ into a coordinate system defined by $A$.

Since $cov(x)$ is symmetric, we can write it down as $cov(x) = A^TDA$ so, we can transform $x$ into a new coordinate system (defined by $A$) we get $y = Ax$ where all $y$ are uncorrelated ($cov(y) = D$). So $\text{var}(y_i) = \lambda_i, \forall i \neq j: cov(y_i, y_j) = 0$

So this transform doesn’t mean that the coordinates are independent, however, at least, they are uncorrelated.

How to calculate the covariance matrix?

Definition $cov(x) = E(xy) - E(x)E(y)$ but we can define this very succinctly in matrix notation

$\text{COV} = \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \bar{x}^T \bar{x}$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Calculating PCA

In matlab: princomp

$[\text{COEFF, SCORE, latent, tsquare}] = \text{princomp}(X)$

COEFF = eigen vectors (orthonormal matrix)
latent = eigen values
SCORE = samples in X represented in coeff basis
tsquare = t-square test for outliers

Switching to matlab

parlerogram -> rhombus

So what are we doing, we are rotating and organizing things in a "normalized" way. Nail example. Rounding error due to floating point in matlab is about $10^{-16}$. Basically, PCA is useful for finding whether high-dimensional data is in some hyperplane and can be represented in less-dimensions. It doesn’t work on curved manifolds
coding text:
idea: assign shorter codes to frequent letters
block codes:
Most sequence have close to pn I’s
Hoeffding’s bound
\[ P(|\hat{p} - p| > \epsilon) \leq e^{-2n\epsilon^2} \]
probability of a typical sequence \( \approx p^n p (1 - p)^n (1 - p)^n = 2^n (p \log_2 p + (1 - p) \log_2 (1 - p)) \)
(binary) entropy: \( H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \)
so, number of bits necessary to encode a typical sequence is about \( n H(p) \)
Entropy of one character

\[ H(\vec{p}) = \sum_{i=n} p_i \log_2 \frac{1}{p_i} \]

\( pr(c_i, c_j) = p_i p_j \) so \( H(\text{two char}) = \sum_{i,j} p_i p_j \log_2 \frac{1}{p_i p_j} = 2H(\vec{p}) \) But, for Huffman coding where you send integer, you get to multiply before rounding, so bigger block better than small code.

... 

Dyadic - finite expansion
length of arithmetic encoded message is
\[ \log_2 \prod_{i=1}^{n} p_i(x_i) = \sum_{i=1}^{n} \log_2 \frac{1}{p_i(x_i)} = \text{Supose cumulative log loss (how many bits we need to describe message)} \]

How to make predictor?
Suppose sequence is generated by IID coin flips with probability \( \theta = p(x_i = 1) \)

In two pass coding:
1) estimate \( \theta \) using \( \hat{\theta} = \frac{\# \text{ of } 1s}{n} \)
2) send \( \theta \) using \( \frac{1}{2} \log_2 n \) bits
3) send message using constant prediction \( \hat{\theta} \)

So length of message is then
\[ nH(\hat{\theta}) = n\hat{\theta} \log_2 \frac{1}{\hat{\theta}} + n(1 - \hat{\theta}) \log_2 \frac{1}{1 - \hat{\theta}} \]

Single pass coding:
Laplace law of succession
\[ p_{t+1} = \frac{\# \text{ of } 1s_{t+1}}{t+2} \]
so
\[ \sum_{t=1}^{T} \log \frac{1}{p_t(x_t)} \leq TH(\hat{\theta}) + \log_2 T \text{ for ANY sequence. But } \log_2 T \text{ is what is needed if encode } \hat{\theta} \text{exactly (naive two pass encoding). Adaptive arithmetic coding made by father of encoding Reasonin (from IBM).} \]

KT (Krichevski Trofinov) law of succession (1985)
\[ p_{t+1} = \frac{\# \text{ of } 1s_{t+1}}{t+2} \]
so then
\[ \sum_{t=1}^{T} \log \frac{1}{p_t(x_t)} \leq TH(\hat{\theta}) + \frac{1}{2} \log_2 T \]

smart model (multi character blocks)
markov model of order 1
So for order \( d \) then log loss is
\[ TH(\text{Markov Model}) + \frac{2^{d+1}-1}{2} \log_2 T \]