In [50]:

```
import pandas as pd
import seaborn as sns
import numpy as np
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
from IPython.display import display
from scipy.stats import ttest_rel, ttest_ind
%matplotlib inline
```

Read various probabilty tables

In [6]:

```
df = pd.DataFrame()
files_xls = ["../ProbTable2009.xlsx", "../ProbTable2010.xlsx"]
```

In [7]:

```
for f in files_xls:
    data = pd.read_excel(f, 'Sheet1')
    df = df.append(data)
```

In [11]:

```
df.head()
df.tail()
```

Out[11]:

	P(+|-)tick	P(+)tick	P(+|-)1s	P(+)1s	P(+|-)3s	P(+)3s	P(+|-)5s	P(
20101227	0.507582	0.485210	0.625322	0.510298	0.595684	0.508787	0.567832	0.
20101228	0.520866	0.486997	0.577957	0.488425	0.542905	0.483188	0.516506	0.
20101229	0.524385	0.500110	0.569021	0.493001	0.535066	0.494421	0.504714	0.
20101230	0.611123	0.513879	0.649875	0.521865	0.615385	0.523026	0.581476	0.
20101231	0.595550	0.530233	0.626106	0.509939	0.597641	0.505829	0.565960	0.

Break down data into various periods

In [28]:

```
Prob_df_period_wise = []
events = ['20090101' ,'20090831',  '20091203', '20101231']

for event in range (0,len(events)-1):
    print(event)
    Prob_df_period_wise.append( df[(df.index.astype(str) < events[event+1]) &
(df.index.astype(str) > events[event])])
```

```
0
1
2
```

In [ ]:

```

```

t-test analysis

1. Paired two-sample for P(+) vs P(+|-) for various timespans
2. Independent pair wise on various pairs of events for various timespans

In [31]:

```
timespans = ['tick','1s', '3s', '5s', '10s', '20s', '30s', '1T', '5T']
len(Prob_df_period_wise)
```

Out[31]:

```
3
```

In [38]:

```
# 1. Paired two-sample
for period in range(0, len(Prob_df_period_wise)):
    print("=========Period:",period)
    for timespan in timespans:
        print("=========Timespan:",timespan)
        print(ttest_rel(Prob_df_period_wise[period]['P(+|-)'+timespan].values,
Prob_df_period_wise[period]['P(+)'+timespan].values))
```

```
=========Period: 0
=========Timespan: tick
Ttest_relResult(statistic=25.084662438354322, pvalue=5.64056984698
0332e-57)
=========Timespan: 1s
Ttest_relResult(statistic=86.50329372824099, pvalue=5.398306041472
0381e-135)
=========Timespan: 3s
Ttest_relResult(statistic=39.829878734145417, pvalue=2.66539861793
99894e-84)
=========Timespan: 5s
Ttest_relResult(statistic=17.348080735100176, pvalue=2.02371142466
53406e-38)
```

```
=========Timespan: 10s
Ttest_relResult(statistic=-7.3755982519526766, pvalue=8.6883543726
699117e-12)
=========Timespan: 20s
Ttest_relResult(statistic=-19.926082180596584, pvalue=5.8256584365
828949e-45)
=========Timespan: 30s
Ttest_relResult(statistic=-21.642275859739154, pvalue=4.0495382478
584349e-49)
=========Timespan: 1T
Ttest_relResult(statistic=-22.190808109890728, pvalue=2.0532753220
315099e-50)
=========Timespan: 5T
Ttest_relResult(statistic=-6.2238868021949001, pvalue=4.1812820968
361782e-09)
=========Period: 1
=========Timespan: tick
Ttest_relResult(statistic=12.393130854418043, pvalue=2.60562000447
29847e-18)
=========Timespan: 1s
Ttest_relResult(statistic=43.310344803335909, pvalue=1.62476615355
97941e-47)
=========Timespan: 3s
Ttest_relResult(statistic=24.077724845786371, pvalue=7.53922453381
07648e-33)
=========Timespan: 5s
Ttest_relResult(statistic=9.6308565380049256, pvalue=7.36586807354
55334e-14)
=========Timespan: 10s
Ttest_relResult(statistic=-4.3161204257170303, pvalue=5.9235121574
403415e-05)
=========Timespan: 20s
Ttest_relResult(statistic=-13.542846531031405, pvalue=4.8148376530
324894e-20)
=========Timespan: 30s
Ttest_relResult(statistic=-11.905022055458758, pvalue=1.4924292496
208181e-17)
=========Timespan: 1T
Ttest_relResult(statistic=-10.493332678487299, pvalue=2.7361289254
82784e-15)
=========Timespan: 5T
Ttest_relResult(statistic=-4.2348691704845782, pvalue=7.8397573173
269889e-05)
=========Period: 2
=========Timespan: tick
Ttest_relResult(statistic=28.720698871110475, pvalue=8.90154301268
61532e-84)
=========Timespan: 1s
Ttest_relResult(statistic=76.15562054292171, pvalue=1.387526730032
5609e-183)
=========Timespan: 3s
Ttest_relResult(statistic=40.736413536010339, pvalue=7.97772507376
53167e-117)
=========Timespan: 5s
Ttest_relResult(statistic=18.421850875406928, pvalue=1.58646864873
00354e-49)
=========Timespan: 10s
```

```
Ttest_relResult(statistic=-7.9416223020736645, pvalue=5.5188423549
02323e-14)
=========Timespan: 20s
Ttest_relResult(statistic=-22.188050616001128, pvalue=1.2540416463
292966e-62)
=========Timespan: 30s
Ttest_relResult(statistic=-24.237463292175487, pvalue=1.6750016629
698999e-69)
=========Timespan: 1T
Ttest_relResult(statistic=-23.467435581013333, pvalue=6.0627748715
883474e-67)
=========Timespan: 5T
Ttest_relResult(statistic=-6.8260802427540082, pvalue=5.8158789443
086625e-11)
```

In [54]:

```
# 2. Independent two-sample across periods
for timespan in timespans:
    print("==== timespan:", timespan)

    print("==== Perio 0 vs Period 1====")
    print(ttest_ind(Prob_df_period_wise[0]['P(+|-)'+timespan].values,Prob_df_p
eriod_wise[1]['P(+)'+timespan].values))
    print("==== Perio 1 vs Period 2====")
    print(ttest_ind(Prob_df_period_wise[0]['P(+|-)'+timespan].values,Prob_df_p
eriod_wise[1]['P(+)'+timespan].values))
    print("==== Perio 0 vs Period 2====")
    print(ttest_ind(Prob_df_period_wise[0]['P(+|-)'+timespan].values,Prob_df_p
eriod_wise[1]['P(+)'+timespan].values))
```

```
==== timespan: tick
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=11.679105524813352, pvalue=8.14597659072
56268e-25)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=11.679105524813352, pvalue=8.14597659072
56268e-25)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=11.679105524813352, pvalue=8.14597659072
56268e-25)
==== timespan: 1s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=45.281866778323689, pvalue=3.61763205585
84532e-113)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=45.281866778323689, pvalue=3.61763205585
84532e-113)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=45.281866778323689, pvalue=3.61763205585
84532e-113)
==== timespan: 3s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=21.472690534256973, pvalue=8.47706290199
25396e-56)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=21.472690534256973, pvalue=8.47706290199
```

```
25396e-56)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=21.472690534256973, pvalue=8.47706290199
25396e-56)
==== timespan: 5s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=9.7131039361975944, pvalue=8.94657142442
06251e-19)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=9.7131039361975944, pvalue=8.94657142442
06251e-19)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=9.7131039361975944, pvalue=8.94657142442
06251e-19)
==== timespan: 10s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=-2.9873379111903069, pvalue=0.0031348538
332521929)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=-2.9873379111903069, pvalue=0.0031348538
332521929)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=-2.9873379111903069, pvalue=0.0031348538
332521929)
==== timespan: 20s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=-8.7574985120592341, pvalue=5.4983406571
586983e-16)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=-8.7574985120592341, pvalue=5.4983406571
586983e-16)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=-8.7574985120592341, pvalue=5.4983406571
586983e-16)
==== timespan: 30s
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=-9.3154510240004456, pvalue=1.3384444251
269895e-17)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=-9.3154510240004456, pvalue=1.3384444251
269895e-17)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=-9.3154510240004456, pvalue=1.3384444251
269895e-17)
==== timespan: 1T
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=-8.4167186047700646, pvalue=5.0511116457
251918e-15)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=-8.4167186047700646, pvalue=5.0511116457
251918e-15)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=-8.4167186047700646, pvalue=5.0511116457
251918e-15)
==== timespan: 5T
==== Perio 0 vs Period 1====
Ttest_indResult(statistic=-1.264546279562887, pvalue=0.20737810161
```

```
26012)
==== Perio 1 vs Period 2====
Ttest_indResult(statistic=-1.264546279562887, pvalue=0.20737810161
26012)
==== Perio 0 vs Period 2====
Ttest_indResult(statistic=-1.264546279562887, pvalue=0.20737810161
26012)
```

In [ ]:

```
```

In [ ]:

```
```

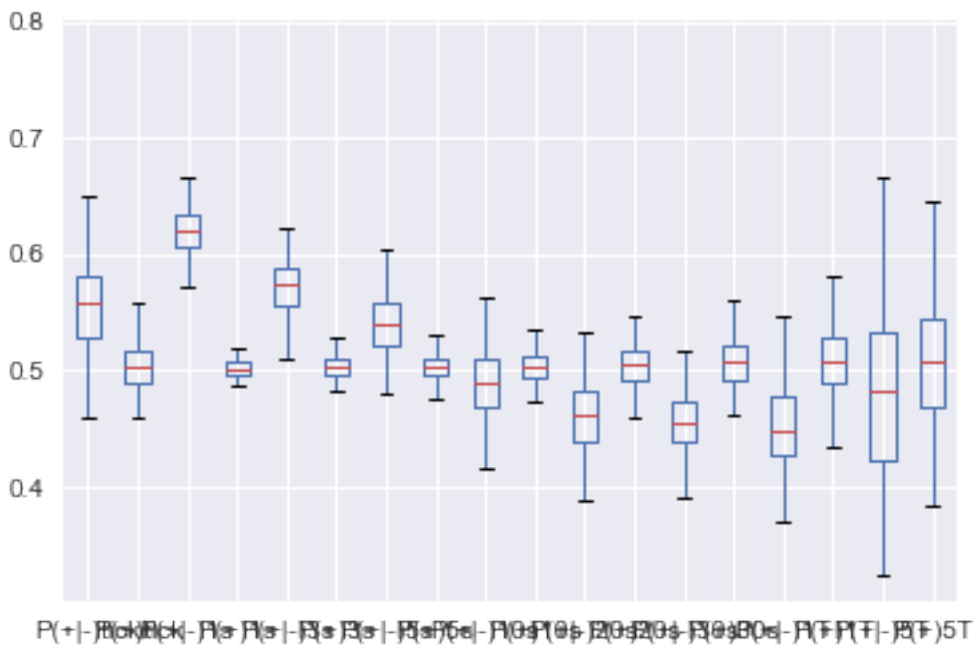
Backup

Box plots, period wise

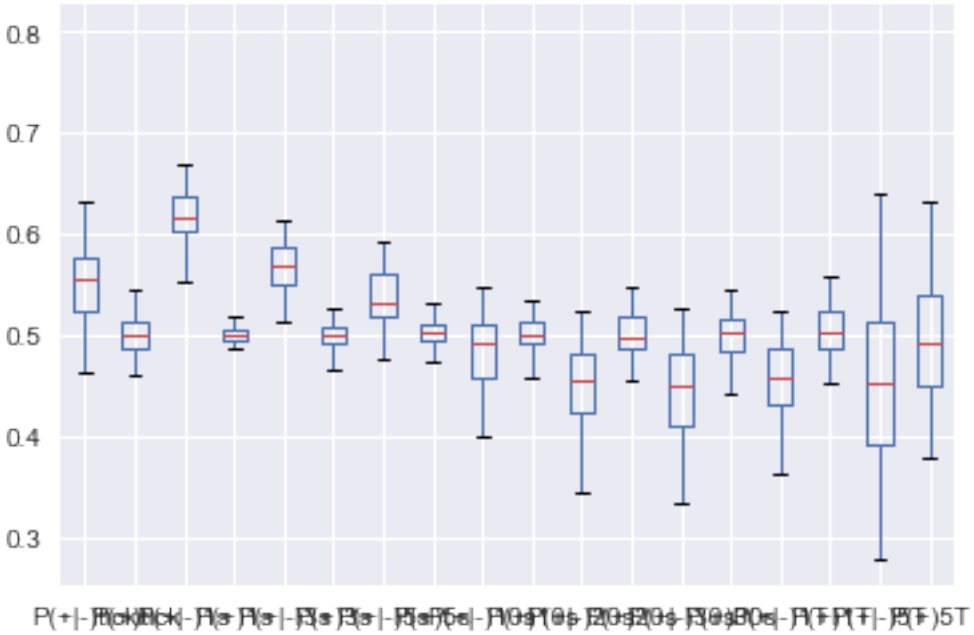
In [53]:

```
display(Prob_df_period_wise[0].boxplot())
plt.show()
display(Prob_df_period_wise[1].boxplot())
plt.show()
display(Prob_df_period_wise[2].boxplot())
plt.show()
```

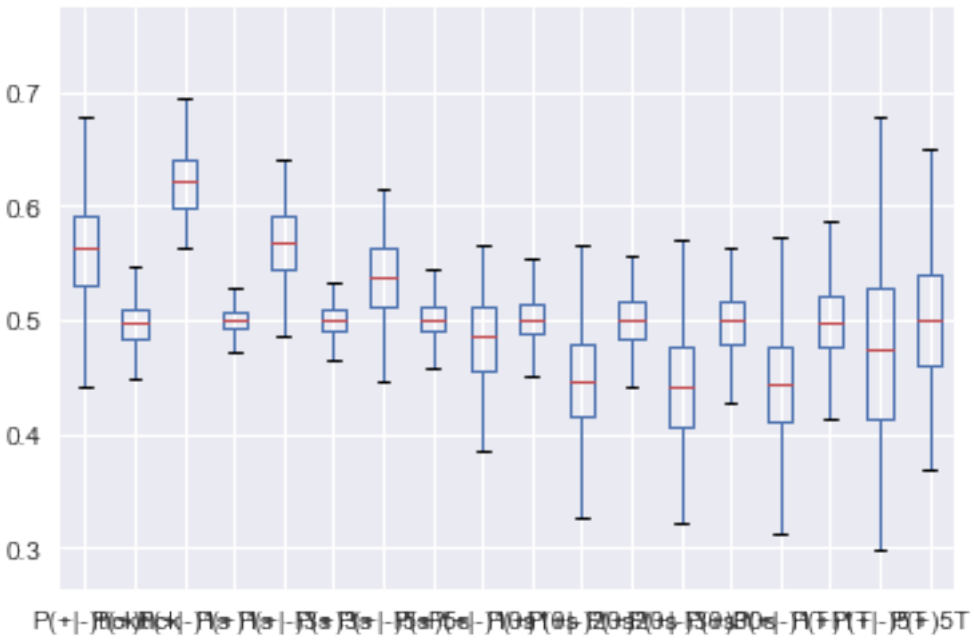
```
<matplotlib.axes._subplots.AxesSubplot at 0x1213eef28>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x120f29ba8>
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x120f3d9b0>
```



In [ ]: