Statistics and Data Analysis
Lecture 10

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2017-01-10
Last edited 21:00:42 2017-01-10

1 HW2 comments

Part 2 Binning is extremely important for histograms. Axes and plot labelling is another reason for losing points. Total vs Percentage - Normalizing graphs can be useful, but remember to label your axes as "frequency of x" or "number of x".

Cube of 12 faces $Y_n = \max$ of n rolls
$P(Y_n = 12) = 1 - P(\text{all } n \text{ rolls are } \leq 11) = 1 - (11/12)^n \to_{n\to\infty} 1$

Something else, I was busy downloading the lecture slides $Conv(a, b) \sim Unif(0...n-1)$ $n = p \cdot q$ $n = 10$ $p = 5$ $q = 2$ $a = Unif(0 : p : p(q-1))$ $b = Unif(0 : 1 : p-1)$

Chair size $\alpha = 14.5$ by Chebicheff

$$CDF_T(\mu - \alpha) + 1 - CDF_T(\mu + \alpha) \leq 0.05$$

$$f_T(x) = 0.5f_1(x) + 0.5f_2(x)$$

$$\int_0^u f_T(x)\, dx$$

which gives us $\alpha \approx 7 - 8$

Extra credit question:

$$\forall t : p(|x - \mu| > t) \leq \frac{Var(X)}{t^2}$$

$$\pi(t) = t^2$$

For any positive and monotonically increasing function π, for any $t > 0$ and for any random variable X it holds that:

$$\pi(t) = t^4$$

$$p(|X-\mu|>t) \leq \frac{E(\pi(|X-\mu|))}{\pi(t)}$$

then we get that for $(X-\mu)^2$ $(* = [|X-\mu|>t])$:

$$Var(X) \geq \int_* (X-\mu)^2 \geq t^2 \int_* 1 = t^2 P(*)$$

$$X \sim N(\mu_1, \delta_1) \text{ with pdf } f_1$$
$$X \sim N(\mu_2, \delta_2) \text{ with pdf } f_2$$

Let M have pdf $f_M = w_1 f_1 + w_2 f_2$.

$$\forall g : E(g(M)) = w_1 E(g(X_1)) + w_2 E(g(X_2))$$

$$\underbrace{E\left((T-\mu)^4\right)}_{g(T)} = w_1 E((T_M-\mu)^4) + w_2 E((T_F-\mu)^4)$$

$$E((T_M-\mu)^4) = E(((T_M-\mu_M)+(\mu_M-\mu))^4)$$

$$E((T_M-\mu_M)^4)+3(\mu_M-\mu)E((T_M-\mu_m)^3)+6(\mu_M-\mu)^2E((T_M-\mu_M)^2)+3(\mu_M-\mu)^3E((T_M-\mu_M))+(\mu_M-\mu)^4$$

$$E(X-\mu) = E(X)-\mu = 0$$

$$X \sim N(\mu, \sigma)$$

$$E((X-\mu)^n) = E((\sigma Z)^n) = E(Y^n) \quad Y \sim N(0,\sigma)$$

$$E((T_M-\mu_M)^3) = E((\sigma Z)^3) = \sigma^3 E(Z^3)$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{x^3 e^{-x^2}}_{u(x)}\, dx = \int_{-\infty}^{0} + \int_{0}^{\infty}$$

$$u(x) = -u(-x) \Rightarrow \int_{-\infty}^{0} - \int_{-\infty}^{0} = 0$$

$$E((T_M-\mu_M)^4) = 3\sigma_M^4$$
$$E((T_M-\mu_M)^2) = \sigma_M^2$$
$$E((T_M-\mu)^4) = 3\sigma_M^4 + 6\cdot 4\sigma_M^2 + 2^4 \underset{\sigma=3}{=} 3\cdot 9^2 + 24\cdot 9 + 16 = 475$$
$$E_F = 304$$

$$E((T-\mu)^4) = \frac{1}{2}475 + \frac{1}{2}304 = 309$$

To get α, solve:

$$0.05 = \frac{390}{\alpha^4} \quad \alpha^4 = 7800 \quad \alpha \approx 9$$

2 Agenda

- Assessing distribution fits
- Python demo of convolutions - Tal Zaccai[1]
- Paired Samples
- Paired t-test
- Wilcoxon signed ranks
- What is multiple testing?

3 Assessing distribution fits

3.1 Pearson χ^2 test

Tells us which distribution family we should fit. Bin the data - n instances into r bins. Make sure that all bins have at least 5 samples.

We observe binned data as O_i - the samples in the i-th bin.

Assuming a certain distribution, we obtain the modeled probability of each bin: $p(i)$. We now compute the Pearson χ^2:

$$PQ = \sum_{i=1}^{r} \frac{(O_i - p(i)n)^2}{p(i)n} \geq 0$$

When r is not too small, we get:

$$\frac{PQ - r}{\sqrt{2r}} \sim N(0,1)$$

Example: Particle emission times Particles were emitted from a radioactive source and times of emission were recorded.

Total number of particles detected = 10220

The total experiment period was divided into 1207 intervals of 10 seconds. In Lecture 3, if every particle independently decides which interval to fall in with uniform probability for each interval, then we expect that the number of particles per interval is poissonic.

Formally, we would say that $Binom\left(10K, \frac{1}{1207}\right) \sim Poi(\frac{10K}{1207})$.

This is what we want to statistically assess.

The raw data appears on slides 6,8,9,11, and 12.

The MLE for Poisson is the empirical mean of the data: $\hat{\lambda}$.

$$\lambda = \hat{\lambda} = \frac{1}{n} \sum O_i$$

$$P(6) = Poi(7, \hat{\lambda}) \approx 159$$

The total error is $PQ \approx 9.0$

$$\frac{PQ - r}{\sqrt{2r}} \sim N(0,1)$$

So we want to assess the right tail fo $N(0,1)$ at $\frac{9-16}{5.65} = -1.24$

[1] She's sick, so we'll do this next week

We get that $LT = normcdf(-1.24) \approx 0.1$ and $RT = 0.9$

This means that the data fits a $Poi(8.39)$ distribution with a 90% Pearson χ^2 confidence.

4 Paired samples

We are interested in the effect of treatment on lowering cholesterol levels. We measure TBC in 100 subjects in a (pre-clinical) trial before Tx and then after they undergo 3 weeks of Tx.

Example:

```
>> CB = 230 + 15 * randn(1, 100);
>> CA = CB - 5 + 3 * randn(1, 100);
# scipy.stats.ttest_ind
>> [hU pU] = ttest2(CB, CA);
hU = 1
pU = 0.0202
```

4.1 Paired ttest

We consider the difference of the values measured for each data sample. We test against the null hypothesis that this difference is distributed around 0. Under this null model and if the sample is large enough we have $mean(\text{differences}) \sim N(0, SE*)$. Thus we use a single sample ttest to assess the observed mean against this distribution.

Matlab has two variants of the ttest function. ttest/1 compares to an expected mean of 0, ttest/2 takes the difference.

```
# scipy.stats.ttest_1samp(CB–CA, 0.0)
>> [hP pP] = ttest(CB,CA)
hP = 1
pP = 1.4632e-30
```

```
# scipy.stats.ttest_1samp(CB–CA, 0.0)
>> [hD pD] = ttest(CB–CA)
hP = 1
pP = 1.4632e-30
```

```
>> CB = 230 + 20 * randn(1,100);
>> CA = CB - 1 + 3 * randn(1,100);
>> PlotFreqComparison(CB,CA)
>> [hU pU] = ttest2(CB,CA)
hU = 0
pU = 0.6217
>> [hP pP] = ttest(CB,CA)
hP = 1
pP = 1.0043e-05
```

4.2 Wilcoxon Rank Sum

Most software packages refer to this as **Wilcoxon Signed Rank**. ⇐ scipy.stats.wilcoxon

Nonparametric alternative to t-test. Data are paired and come from the same population. Each pair is chosen randomly and independently.

Step 1 Take the differences. Exclude any differences which are zero. Put the rest of differences in ascending order. Ignore their signs. Assign them ranks. If any differences are equal, average their ranks

Step 2 Count up the ranks of positives as $T(+)$ Count up the ranks of negatives as $T(-)$ Set $W = T(+) - T(-)$

Intuition If there is no difference between drug $T(+)$ and placebo $T(-)$, then $T(+)$ and $T(-)$ would be similar. If there is a difference then one sum would be much smaller and the other much larger than expected Under the null assumption that $\mu 1 = \mu 2$ the differences are as likely to be positive as they are to be negative and the ranks are also distributed independently of the sign. W therefore has a known distribution, which we will assess against.

```
>> CB = 230 + 20 * randn(1,100);
>> CA = CB - 1 + 3 * randn(1,100);
# scipy.stats.wilcoxon
>> Wp = signrank(CB, CA)
Wp = 2.1399e-05
```