INF5830 – 2017 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 4, 11.9

Today

- Motivation
- Evaluating a binary classifier against a baseline
- Normal distribution (recap)
- Samples
- Hypothesis testing, general case
- Estimation, general case
- Estimation for a proportion

Why statistics in evaluation?

- Task1:
  - You know the best classifier on a task has 0.8 (80%) accuracy (baseline).
  - You have made a classifier which classify 85 items correctly on a test set of 100 items.
  - Can you conclude your classifier is better than the baseline?
- Task 2:
  - You have made a classifier. You test it on 500 items. It classifies 375 correctly.
  - What is the accuracy of your classifier?

Why? (next week)

- Task 3:
  - You have two different classifiers, one with accuracy 0.89 and one with accuracy 0.91on 1000 test items.
  - Can you conclude that one is better than the other?
- Task 4:
  - The two classifiers from task 3 agree on 870 items.
  - One is doing better on 20 items, the other is doing better on 40 items.
  - Can we draw conclusions from this?

Why?

- Two parts to evaluation:
  - The device to be evaluated
  - The test items
- In choosing our test items there is an element of randomness, like
  - Flipping a coin, or
  - Drawing balls from an (infinite) urn

Vancouver Sun, «IKEA ballroom»

Flipping a coin 10-times

- Your friend has a coin.
- You suspect it is unfair and shows too many heads
- To test, you flip it 10 times
- How many heads should come up to confirm your hypothesis?

6 heads?	7 heads?	8 heads?	9 heads?	10 heads

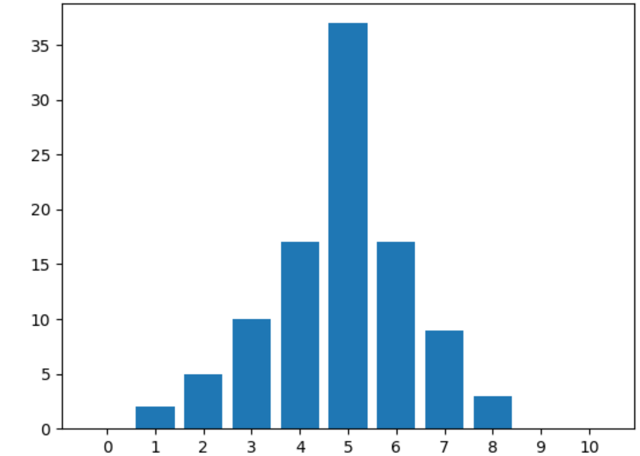
Flipping more times

- What if you instead flip it 100 times?
  - 60?
  - 70?
- What if you flip it 1000 times?
- 10,000 times?
- We expect the proportion to approach 0.5 as n gets bigger
  - But how fast?

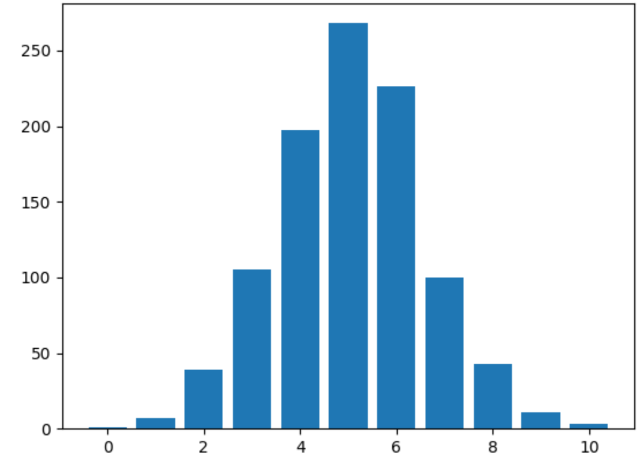
Flipping a coin 10-times

- Here is a way to check what to expect for10 flips.
- Take a coin you know is fair:
  - (Because you have flipped it 10,000 times)
  - Flip it 10 times and record the numer of heads.
  - Do this over again n many times, and collect the recorded number of heads for each 10 flips, and inspect the numbers.
  - The number of heads is a random variable X.
  - As n grows, the distribution of X approaches the binomial distribution B(10, 0.5)

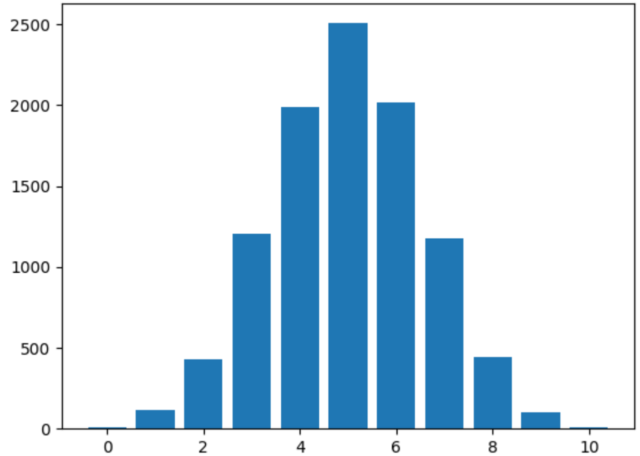
10 flips, n many times



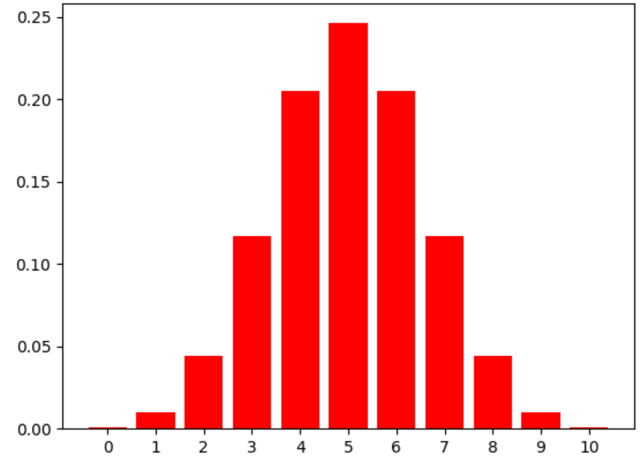
n=100



n=1000



n=10,000



binom.pmf

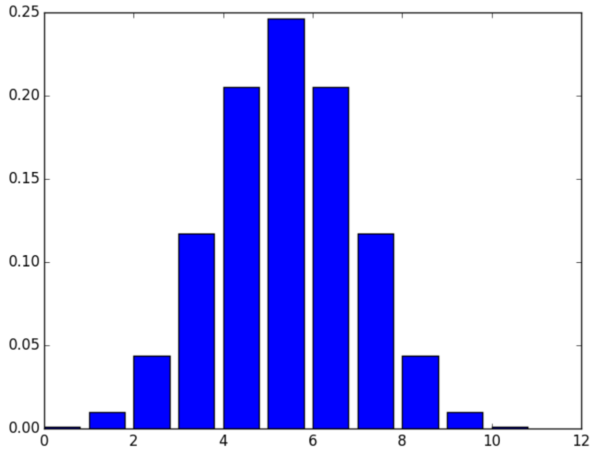
Use of the binomial distribution

- From the binomial distribution, we can see how likely it is to get 10 heads, 9 heads, 8 heads, etc. (= the pmf, probability mass function)
- And how likely it is to get at least 9 or at least 8 heads, etc:
  - $P(X \geq 8) = p(8)+p(9)+p(10)=F(10)-F(7)=1-F(7)$ (F is the cdf, cumulative density function)

Tossing a fair(?) coin

- The cumulative distribution function: ``How likely is it to get N or fewer tails?´´

10:



N	pmf(N)	cdf(N)
0	0.001	0.001
1	0.010	0.011
2	0.044	0.055
3	0.117	0.172
4	0.205	0.377
5	0.246	0.623
6	0.205	0.828
7	0.117	0.945
8	0.044	0.989
9	0.010	0.999
10	0.001	1.000

What is the propbaility of getting 8 or more heads?

What is unusual?

- What is unusual?
  - 25%?
  - 10%?
  - 5%?
  - 1%?
  - 0.1%?

- In statistical tests, one normally uses 5%
- With this number we will draw wrong conclusions 1 out of 20 times.
- Sometimes 10, 1, 0.1% are used.

SciPy

- import scipy
- from scipy import stats
- bin10 = stats.binom(10, 0.5) # N=10, p=0.5
- bin10.pmf(3) # probability mass of 3
- bin10.cdf(3)   # cumulative distribution function at 3
- bin10.var()   # variance
- bin10.std()   # standard deviation

- In [169]: bin10.cdf(10)-bin10.cdf(7)
- Out[169]: 0.0546875
- In [170]: bin10.ppf(.95)
- Out[170]: 8.0

Formulate the 10 flips as a test

- Alternativ hypothesis
  Ha: "Jim's coin comes up heads more than 50%"
- Null hypothesis
  H0: "Jim's coin does not come up heads more than 50%"
- If Jim's coin comes up heads *n* times in 10 throws, and the probability of getting *n* or more heads is less than $p=0.05$, we can reject the null hypothesis

100 flips

- What if we instead use 100 flips?

- The procedure is the same. But this time we can reject the null hypothesis if we get 59 or more heads.

- In [172]: stats.binom.ppf(.95, 100, 0.5)
- Out[172]: 58.0

- In [173]: stats.binom.ppf(.95, 1000, 0.5)
- Out[173]: 526.0

- In [174]: stats.binom.ppf(.95, 10000, 0.5)
- Out[174]: 5082.0

Applying to evaluation

- How does this apply to evaluation?
- If the baseline classifier has 0.5 accuracy and we test our own classifier on 100 items, we need at least 59 correctly classified to conclude anything.
- What we can conclude is that the new classifier is better than baseline – not that its accuracy is 0.59

Larger numbers

- What if the baseline is 0.8, still 100 test items?
- What if the baseline is 0.8 and 1000 test items?
- What if the baseline is 0.8 and 10000 test items?

- In [175]: stats.binom.ppf(.95, 100, 0.8)
- Out[175]: 86.0

Sample size		100	1000	10000	
Number of correct items to beat baseline		87	822	8067	
Recorded accuracy to beat baseline		0.87	0.822	0.8067	

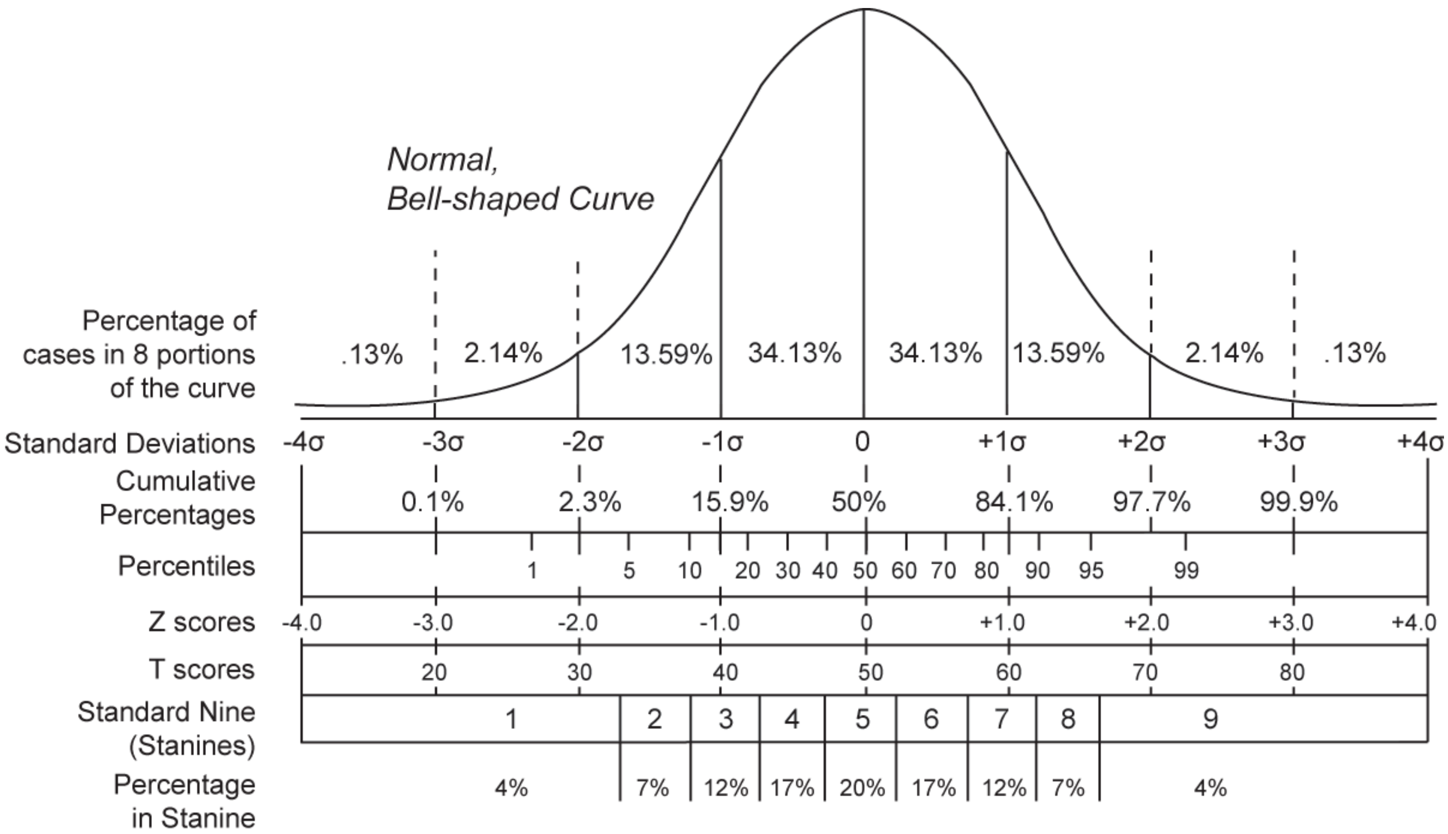
Normal distribution

(Recap)

Normal distribution

- For our purposes, we can mainly survive with the binomial distribution and proportions.
- We will bring in the normal distribution to see:
  - Standard statistical tests
  - Relationships between binomial and normal distrbs.
  - You only need one table for normal distributions
    - Compared to one for each pair n,p for B(n, p)

The normal distribution - Continuous



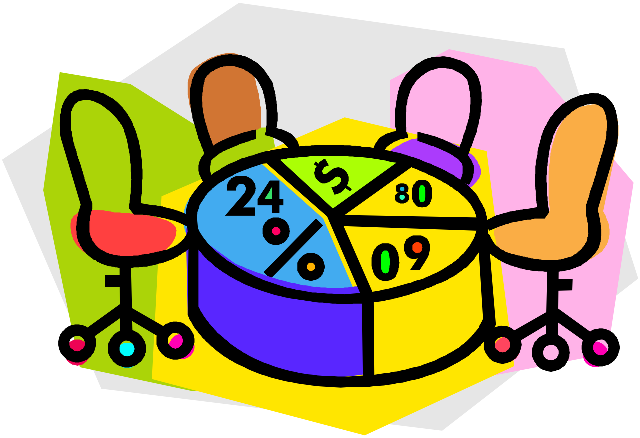
Example height (contd.)

- Tallness of Norwegian young men (rough numbers):
  - Normal distribution, $\mu = 180$ cm, $\sigma = 6$cm

- How many are taller than 190cm?
  - First calculate the z-score
    (how many standard deviations is this?)
  - $z = \frac{x-\mu}{\sigma} = \frac{190-180}{6} = 1.67$
  - Use software, calculator or table to find the corresponding probability *p*.
  - Here p=0.0475

Look up

- **Statistical table**
  - course.shufe.edu.cn/jpkc/jrjlx/ref/StaTable.pdf



- **SciPy**
  - >>>import scipy
  - >>> from scipy import stats
  - >>> stats.norm.cdf(10/6)
  - 0.9522096477271853
  - >>> 1-stats.norm.cdf(10/6)
  - 0.0477903522728147
  - >>> stats.norm.cdf(190,180,6)
  - 0.9522096477271853

Table

- Given probability p, for which h is P(X>h) < p?
  - Standardize, calculate the Z-score: z= $\frac{x-\mu}{\sigma}$
  - $P(X > h) = P(\frac{X-\mu}{\sigma} > \frac{h-\mu}{\sigma}) = P(Z > \frac{h-\mu}{\sigma})$
  - Use table or software to look up z
- Conversely, for given h, we may find corresponding z and look up *p*.

Probability p-value	z-score	centimeters	height
0.1	1.28	7.68	187.68
0.05	1.645	9.87	189.87
0.01	2.326	14	194
0.001	3.091	18.5	198.5

Sampling distribution

Utvalgsfordeling

Sampling - empirically

Goal:

- make assertions about a whole population
- from observations of a sample (utvalg)

- A simple random sample (SRS) (tilfeldig utvalg):
  1. Each individual has equal chance of being chosen (unbiased/forventningsrett)
  2. Selection of the various individuals are independent

Binomial distribution

- Flipping the coin 10 times is a sample of coin flips:
  - The probability is the same
  - The flips are independent
- Selection of test items is nearly* a SRS of Bernoulli trials

Vancouver Sun, «IKEA ballroom»

* "Nearly" because of lack of replacement.
Close enough if sample is small compared to population

Sampling in Language Technology

- You want to take a simple random sample of words from a corpus?
  - Can you use the *n* first sentences?
  - Can you use a random sample of *n* sentences?
- How can you build a corpus (sample) which gives a random sample of Norwegian texts?

Sampling distributions – Example

- Height: X
  - assume N(180, 6)
  - $(\mu = 180, \sigma = 6, Var(X) = 36)$
- Randomly choose 100.
- Add their heights:
  $S = X_1 + X_2 + \ldots + X_n$
- A new random variable (all such samples)
  - Exp(S) = n*μ= 18000 (cm)
  - Var(S) = 100*Var(X) = 3600
  - $\sigma_S = 10 \times \sigma_X = 60\ (cm)$

Source: Wikipedia

Sampling distributions – Example

- Height: X
  - assume N(180, 6)
  - $(\mu = 180, \sigma = 6, Var(X) = 36)$
- Randomly choose 100.
- Add their heights:
  $S = X_1 + X_2 + \ldots + X_n$
- A new random variable
  (all such samples)
  - Exp(S) = n*µ= 18000 (cm)
  - Var(S) = 100*Var(X) = 3600
  - $\sigma_S = 10 \times \sigma_X = 60\ (cm)$

- The mean of the samples:

  $\bar{X} = S/n$
- A new random variable
  (all means of samples of 100)
- $E(\bar{X}) = \mu_{\bar{X}} = \mu_X = 180$ (cm)
- $\sigma_{\bar{X}} = \frac{1}{100} \times \sigma_S = 0.6\ (cm)$
- $\sigma_{\bar{X}} = \frac{1}{100} \times \sigma_X \times \sqrt{100} = \frac{\sigma_X}{\sqrt{100}}$

Sampling distributions

- Let
  - X be a random variable for a population with exp: μ, std: σ
  - Let $S = X_1 + X_2 + \ldots + X_n$, i.e. each X_i equals X
  - Let : $\overline{X} = S/n$

- Then:
  - $E(S) = n * \mu$
  - $E(\overline{X}) = \mu$
  - $$Var(S) = \sigma_S^2 = n \times Var(X) = n \times \sigma_X^2$$
  - $$Var(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{1}{n^2} \times Var(S) = \frac{1}{n} \times \sigma_X^2$$
  - $$\sigma_{\overline{X}} = \frac{1}{\sqrt{n}} \times \sigma_X$$

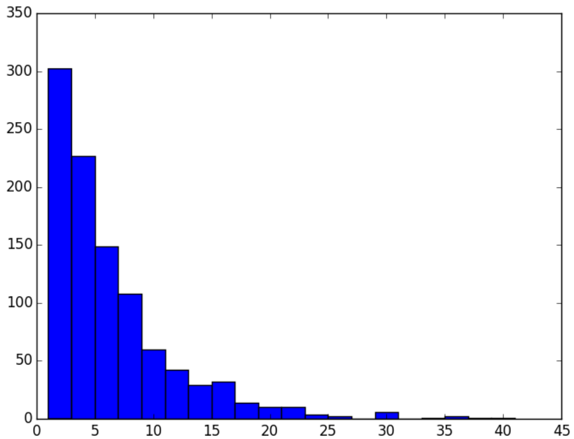
The form of the distribution

- If the Xi-s are independent and normally distributed, then $\bar{X}$ is normally distributed (as expected)
- (More surprisingly) Even though the Xi-s themselves are not normally distributed: for large n-s, $\bar{X}$ is approximately normally distributed
  = Central Limit Theorem

Example: throwing the dice until a 6

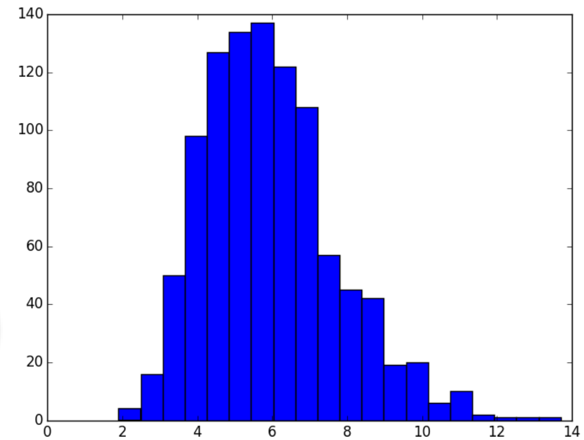
Number of samples: 1000

Sample size



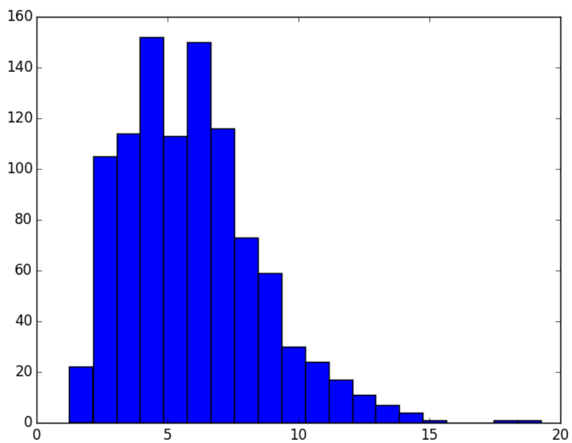
1

10



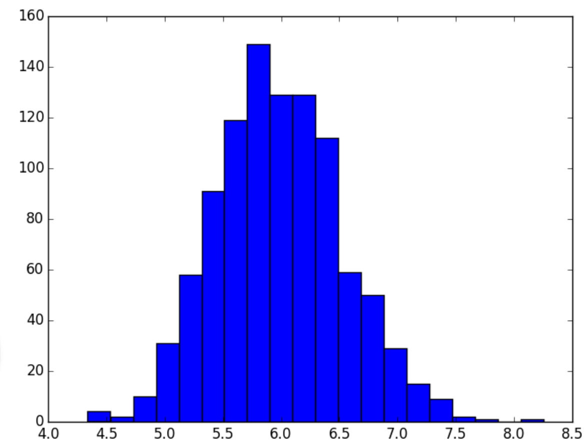
$$E(\bar{X}) = E(X) = \mu = 6$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{6 \times 5}}{\sqrt{n}}$$



4

100



Binomial distribution

$$b(k;n,p) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

Population: all Bernoulli trials with probability *p*.
Sample: *n* such trials
Example: Throwing a dice *n* times, counting the number of 6-s (success)

- Number of successes: X
- Random variable over all series of *n* trials
- Binomial distribution (binomisk fordeling): B(n,p)
- E(X)= *np*
- Var(X)= *np(1-p)*
- $$\sigma_X = \sqrt{np(1-p)}$$
- Approximated by N(*np*, $\sqrt{np(1-p)}$) for large n

- Proportion of success: $\hat{p}$=X/n
- E($\hat{p}$) = E(X/n) = *np*/*n* = *p*
- $$Var(\hat{p}) = \sigma_X^2 / n^2 = np(1-p)/n^2 = p(1-p)/n$$
- $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \frac{\sigma_Y}{\sqrt{n}}$$
- Approximated by N(*p*, $\sqrt{p(1-p)/n}$) for large n

Rule of thumb:
np>10 and
n(1-p)>10

Binomial vs normal approximation

- In [175]: stats.binom.ppf(.95, 100, 0.8)
- Out[175]: 86.0
- In [201]: stats.norm.ppf(.95, 80,
  np.sqrt((1-0.8)*(0.8)*100))
- Out[201]: 86.579414507805893

- For binomial distributions, the traditional statistics used
  - Binomial distributions for small n
  - Normal approximation to binomials/proportions
    - Because of the (non) availability of tables for all (k,n,p)-s
- With computers, we can use the binomial distributions directly

Rule of thumb:
$np>10$ and
$n(1-p)>10$

Hypothesis testing

Hypothesis testing

- Assume P is known with the distribution N(180, 6)
- A population P2, could be:
  - Norw. males 50ys olds in 2007
  - Norw. females 18ys olds in 2007
  - Swe. males 18 ys olds in 2007
- Q1: Are the individuals in P2 shorter than they in P?
- Pick a random sample $\{x_1, x_2, \ldots, x_n\}$ from P2
  - Null hypothesis, $H_0 : \mu_{P2} = \mu$
  - Hypothesis, $H_a : \mu_{P2} < \mu$
  - Q2: What is the chance $\{x_1, x_2, \ldots, x_n\}$ could have been a SRS from P?

Example

- For example, if we take a SRS from P2 of
  - n=100 individuals, and we find
  - $\bar{x} = 178.5$
  - $\sigma_{\bar{X}} = \frac{1}{100} \times \sigma_S = \frac{1}{\sqrt{100}} \times \sigma_X = 0.6\ (cm)$
  - z= $\frac{\bar{x}-\mu}{\sigma_{\bar{X}}} = \frac{178.5-180}{0.6} = -2.5$
- we can conclude (alternative formulations:)
  - there is less than 0.01 chance that $\{x_1, x_2, \ldots, x_n\}$ is a s.r.s. from P
  - If P and P2 had been equal (w.r.t. height), there is less than 1% chance that we would have chosen such a SRS
  - The p-value is less than 0.01

Evaluation

- Observe that this is similar to what we did in the coin flipping and evaluation using binomial distribution

Recipe (with normal distribution)

- Formulate H_a and H_0
- H_0: $\mu_2 = \mu$
- Sample an appropriate SRS of size n and find its mean value, $\bar{x}$
- Calculate the z-score: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

- H_a: $\mu_{P2} < \mu$ is $P(X < z)$
- > similarly:
- H_a: μ_{P2} =/= μ is $2 \times P(X > |z|)$

Remarks

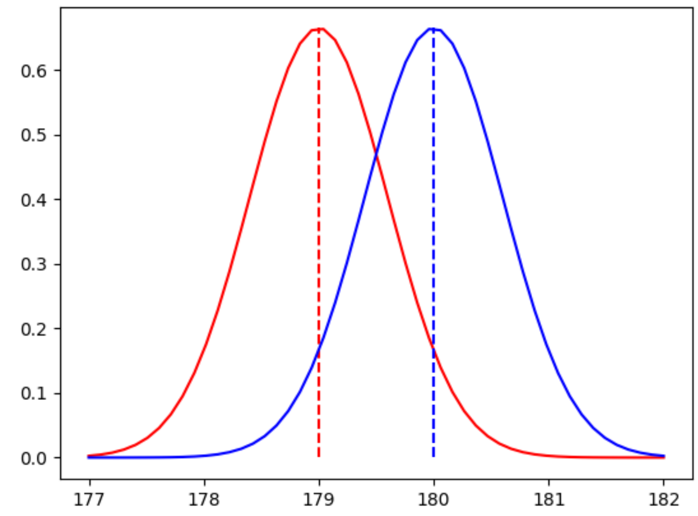
		Truth	
		H0	Ha
Decision	Not rejecting H0		Type II error
	Reject H0	Type I error Prob. p-value	

- There is a chance of probability *p* that we erroneously reject H0 (Type I error)
- The test does not estimate type II error
- Says nothing about how much the difference is between P2 and P
- Many possible banana skins: E.g. is the sample really random?

Estimation

Example

- Assume a population P2 and an SRS of 100 individuals from P2 with $\bar{x} = 179$
- What is μ for P2?
- Goal: find an e such that $P(179 - e < \mu < 179 + e) < p$ for some level p, e.g. 0.05
- Observe that $P(179 - e < \mu < 179 + e)$ $= P(\mu - e < 179 < \mu + e)$
- If we had known the standard deviation, we could calculate this like we have done so far.



Estimation

- How to estimate the true mean μ of a sample if the standard deviation σ of the population is unknown?
- All we have is a sample $X = \{x_1, x_2, \ldots, x_n\}$
- The sample mean $\bar{x}$ is still the best estimate of the pop. mean μ
- How good an estimate is this?

Estimation

- To determine this, we try to estimate the true standard deviation of the population.
- We use the <u>standard deviation of the sample</u> X,
  - $s^2 = ((x1 - \overline{x})^2 + (x2 - \overline{x})^2 + \ldots + (xn - \overline{x})^2)/(n - 1)$
  - Observe (n-1) and not n
  - That is to compensate for using $\overline{x}$ instead of μ in the formula

s is a random variable (like $\overline{X}$) over all s.r.samples of size n
s is an unbiased estimator for σ: E(s)= σ

Estimation

- In addition we do not use the standard Z-distribution but the t-distribution for n-1.
- Then the level C confidence interval for μ is
  - $[\bar{x} - e, \ \bar{x} + e]$
  - Where
    $$e = t^* \frac{s}{\sqrt{n}}$$
  - and t* is the value from the t(n-1) density curve for C

The t-distribution is similar to the z-distribution for large n.
But is more picky when t is small

Example

- Assume we do not know the st.dev. 18 ys old men from Finmark
- Pick a random sample of 9 men:
  - $\bar{x} = 177$, s = 5
- Estimate the average height for this population
  - Choose confidence level 0.95

Table, or

```
In    [78]: stats.t.ppf(.025,8)
Out[79]: -2.306004135033709
```

What would be different if we used normal distribution?

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 177 \pm 2.306 \frac{5}{\sqrt{9}} = 177 \pm 3.843$$

- The 95% confidence interval for μ: [173.1, 180.9]
- Exact for normal distribution
- Approximation for large n otherwise

Estimation with proportion

- Task 2:
  - You have made a classifier. You test it on 500 items. It classifies 375 correctly.
  - What is the accuracy of your classifier?

Proportion

- The best estimate we have for p is $\hat{p} = \frac{375}{500} = 0.75$
- The best estimate we have for the standard deviation is $\mathrm{SE}(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \Bigg(=$

Proportion

- The best estimate we have for p is $\hat{p} = \frac{375}{500} = 0.75$
- The best estimate we have for the standard deviation is $\mathrm{SE}(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

Example

- Estimated accuracy is 375/500=0.75
- The standard deviation of the sample is

$$\sqrt{p(1-p)/n} = \sqrt{0.75(1-0.75)/500} = 0.0194$$

- Using normal distribution approximation:
  - In [284]: stats.norm.ppf([0.025, 0.975],0.75, np.sqrt(0.75*0.25/500))
  - Out[284]: array([0.71204546, 0.78795454])

- Using binomial distribution:
  - In [288]: stats.binom.ppf([0.025, 0.975],500, 0.75)/500
  - Out[288]: array([0.712, 0.788])

Take home

- Two parts to evaluation:
  - The device to be evaluated
  - The test items
- In choosing our test items there is an element of randomness, like
  - Flipping a coin, or
  - Drawing balls from an (infinite) urn

Vancouver Sun, «IKEA ballroom»