Lecture 6: Testing and Statistical Significance

IB Paper 7: Probability and Statistics

Carl Edward Rasmussen

Department of Engineering, University of Cambridge

March 6th, 2015

Statistical Significance

To be able to support or refute statements about unobservable quantities, we can collect statistical evidence, in the form of experiments.

Example: Can people, in a blind test, taste the difference between single malt and blended whisky?

We get 10 people to blind test, each given a randomly selected drink. We observe that 7 people correctly identify their drink, and 3 respond incorrectly. What do we conclude?

Could the observations reasonably be accounted for by random fluctuations?

We repeat the experiment with 1000 people, 700 correct, 300 incorrect. What do we now conclude?

Although the proportions are identical, the statistical significance is different.

Notice: "significantly different" and "statistically significantly different" are different!

Hypothesis

One way to assess statistical significance, is through *hypothesis testing*.

Informally

- Hypothetically, assume a null hypothesis, H_0, to be true. For example, the null hypothesis could be that people can't tell the difference between whiskies.
- Make some observations
- If the observed data has very low probability given the null hypothesis, then the null hypothesis probably wasn't true.

This is probabilistic *reductio ad absurdum*.

Hypothesis testing

Consider again the binary case, 7 correct and 3 incorrect.

Let's examine the null hypothesis, H_0, that people can't taste the difference.

Assuming the null hypothesis is true, what is the probability of the observed outcome, or something more extreme?

$$\sum_{i=7}^{10} p(i \text{ correct}|H_0) = \sum_{i=7}^{10} {}_{10}C_i (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{10-i}$$

$$= (\tfrac{1}{2})^{10} \sum_{i=7}^{10} {}_{10}C_i = (120+45+10+1)/2^{10} \simeq 0.17$$

So, under the null hypothesis, that people respond randomly, this outcome or something more extreme would happen in about 17% of cases.

So, this doesn't constitute strong evidence against the hypothesis.

Typically, significance is judged against a 5% or a 1% threshold.

Notice: This process is fundamentally asymmetric.

The 700 vs 300 case

To analyze the more extensive survey, we would need to compute:

$$(\tfrac{1}{2})^{1000} \sum_{i=700}^{1000} {}_{1000}C_i \,.$$

Direct computation of this quantity is not particularly attractive. Instead, as an approximation we use the Gaussian with the same mean and variance as the binomial under the null (recall $\mathbb{E} = np$ and $\mathbb{V} = npq$)

$$N(\mu = n/2,\ \sigma^2 = n/4) \ = \ N(\mu = 500,\ \sigma^2 = 250).$$

To compute the desired probability we standardize the Gaussian

$$y \ = \ \frac{x - 500}{\sqrt{250}},$$

and compute

$$p(y > \frac{700 - 500}{\sqrt{250}}) \ = \ p(y > 12.6) \ = \ 1 - \Phi(12.6) \ < \ 10^{-6}.$$

I.e., under the null hypothesis, this outcome or something more extreme would be exceedingly unlikely, and we can thus reject the null hypothesis.

The 70 vs 30 case

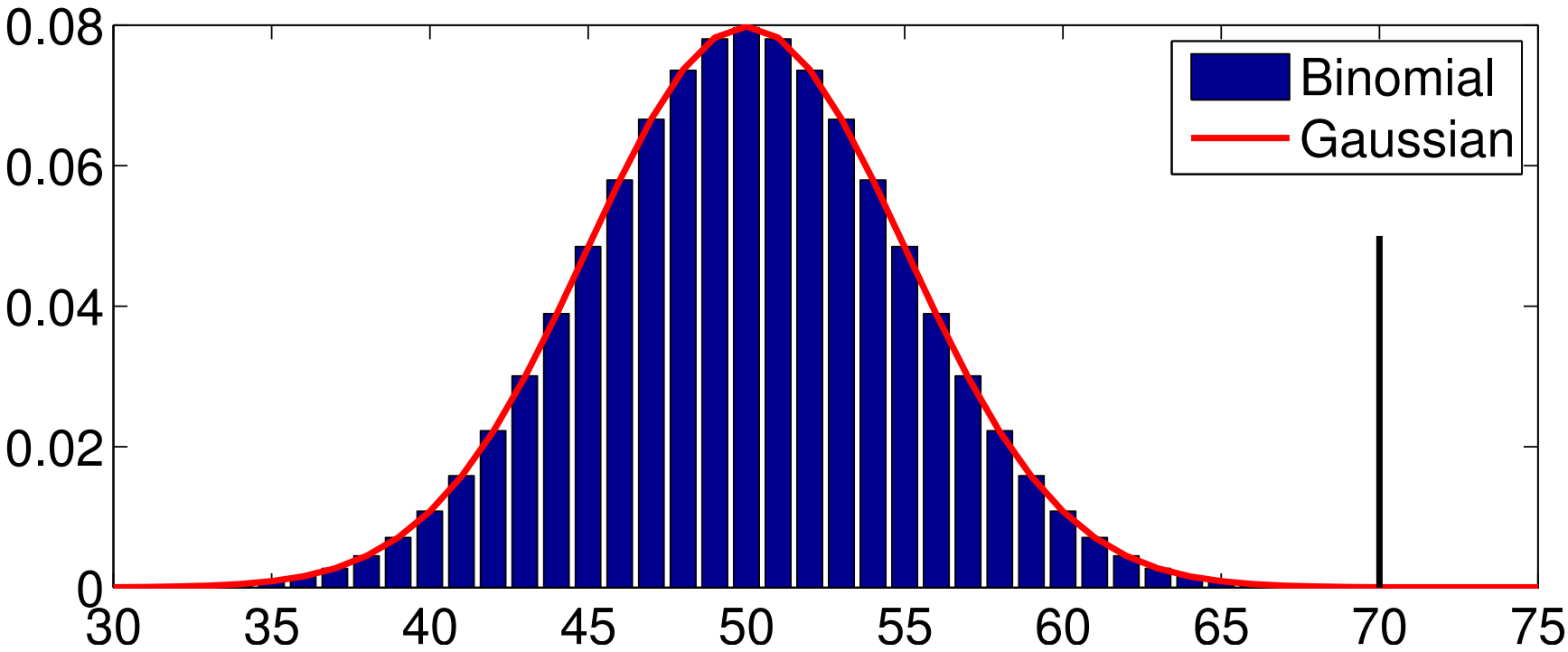

Figure showing the discrete Binomial and the continuous Gaussian approximation together with the empirical value 0.7.

The approximation is very good.

The tail probability is very small.

. . . "or something more extreme"

Notice, that it is not adequate to simply compute the probability of the actual outcome under the null hypothesis.

You need to order the possible outcomes, and compute the probability of the observed outcome or something more extreme.

For example, in the case of 1000 binary trials discussed on the previous slide, any particular outcome is actually unlikely, simply because there are so many. E.g.

$$p(500 \text{ correct}|H_0) = \frac{1000!}{500!500!2^{1000}} \simeq 0.025.$$

One-sided and two-sided tests

Depending on the circumstances, it may be necessary to use a two-sided test.

Example: We want to establish whether a coin is fair. Out of 10 flips, we get 7 heads and 3 tails. What is the evidence against the null hypothesis, that the coin is fair?

We need to evaluate the probability of the observed outcome, or something more extreme:

$$\sum_{i=0}^{3} p(i \text{ heads}|H_0) + \sum_{i=7}^{10} p(i \text{ heads}|H_0) \;=\; 1 - \sum_{i=4}^{6} p(i \text{ heads}).$$

Note, how the coin example is different from the whisky tasting example: the tasting example is asymmetric, in that it cannot really happen, that people are worse than chance.

Figuring out whether one-sided or two-sided tests are appropriate, may require some attention!

Example

A bus company claims that on a certain route there is a service every 20 minutes. Three people complain that this claim is false:

- A: had to wait 45 minutes on a particular day
- B: had to wait 45 minutes on both Monday and Tuesday
- C: had to wait 45 on two days of last week

Are any of these claims statistically significant?

Null hypothesis: bus arrivals are random, Poisson, with intensity $\lambda = 3$ buses per hour. Thus, waiting times are exponentially distributed $\mathrm{Ex}(\lambda = 3)$.

$$\begin{aligned} p(\text{wait } > 3/4) &= \int_{3/4}^{\infty} 3\exp(-3t)dt = \Big[-\exp(-3t)\Big]_{3/4}^{\infty} \\ &= \exp(-9/4) \simeq 0.105 \end{aligned}$$

Example, continued

- A: had to wait 45 mins on a particular day. Since $p = 0.105$ this is not hugely unlikely, and cannot eg at the 5% level be used to reject H_0.
- B: had to wait 45 mins both Monday and Tuesday. Both events happen independently, so $p = 0.105^2 = 0.011$. This seems quite unlikely under the null, so we can reject the companies claim, say at the 5% level.
- C: had to wait 45 mins on two days of last week.

$$p = \sum_{i=2}^{5} {}_nC_i \, 0.105^i (0.895)^{5-i} \simeq 0.089,$$

again, not exceedingly strong evidence.

Notice: one has to be careful interpreting exactly what events are being evaluated.

A Bayesian Alternative

Note, that in the classical test which we have just described

- we think of the *observations* as being random, and the underlying unknown property as being fixed (through the null hypothesis).
- the test is asymmetric, where it is not so clear what the alternative is.
- the test is based on assuming something which is shown unlikely to be true

Could we possibly think of a more intuitive scheme?

A Binary Test

Focus on the binary example, with 7 of one and 3 of the other outcomes. We want to know whether the probability π could be a half:

- corresponding to people not being able to taste the difference, or
- corresponding to a *fair* coin, etc

We cannot simply find the probability that $\pi = \frac{1}{2}$ as π is continuous, this probability would be zero (because it is a probability density).

Instead, we can compare the probability of the observations under two alternative models:

- Model A: observations independent Bernoulli with $\pi = \frac{1}{2}$
- Model B: observations independent Bernoulli with unknown π.

The probability of the data under the two models

Model A: probability of the observations:

$$p(D|A) \;=\; {}_{10}C_3 \left(\tfrac{1}{2}\right)^{10} \;\simeq\; 0.117.$$

Model B: $p(D|\pi, B) = {}_{10}C_3\, \pi^7 (1-\pi)^3$. But we do not know the value of π, so we use

$$p(D|B) \;=\; \int p(D, \pi|B)\, d\pi \;=\; \int p(D|\pi, B) p(\pi)\, d\pi.$$

To this end we need $p(\pi)$, the *prior* on π, our assumption about π. In this case, one appropriate choice would be $p(\pi) = \mathrm{Uni}(0, 1)$, then

$$p(D|B) \;=\; {}_{10}C_3 \int_0^1 \pi^7 (1-\pi)^3\, d\pi \;=\; {}_{10}C_3\, \frac{\Gamma(4)\Gamma(8)}{\Gamma(12)} \;=\; 1/11 \;\simeq\; 0.091,$$

where we used $\int \pi^a (1-\pi)^b\, d\pi = \Gamma(a+1)\Gamma(b+1)/\Gamma(a+b+2)$, which comes from the normalization of the beta distribution:

$$p(\pi) \;\sim\; \mathrm{Beta}(\alpha, \beta) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}.$$

Conclusion

The 10 observations have probability

- $p(D|A) \simeq 0.117$ for model A, and
- $p(D|B) \simeq 0.091$ for model B

which shows a slight preference for model A, but the two models are almost equally good at accounting for the observations.

Consequently, these limited observations, provide no strong preference.

Notice that

$$p(D|\pi = 7/10) \;=\; {}_{10}C_3 (7/10)^7 (3/10)^3 \;\simeq\; 0.267$$

assigns much higher probability to the data — but this is not a fair comparison, why not?

Conclusions for 8 and 2 outcome

If instead, the 10 outcomes had been distributed as 8 and 2, we would have

- classical one-sided test: $p = (45 + 10 + 1)/2^{10} \simeq 0.055$ against the null hypothesis
- classical two-sided test: $p = 0.109$ against the null hypothesis
- Bayesian comparison
  - $p(D|A) = {}_{10}C_2 \, /2^{10} = 0.044$
  - $p(D|B) = 1/11 = 0.091$

  showing a preference for the B model over a model fixed at $\pi = 1/2$.

Hypothesis testing

The classical hypothesis test is used almost universally in practice.

It relies on assuming a null hypothesis, H_0, and computing the evidence against this.

It is asymmetric: a failure in rejecting the null is not evidence in support of the null.

The statement made is complex:

- The probability of the observations, or something more extreme, given the null hypothesis is p.
- The statement is not: The probability of the null hypothesis is p.

The significance level is often compared to a threshold of 5% or 1%, but this is not really necessary.