Probability and Statistics for Data Science

Y-DATA

06.11.2020

This course

Probability & decision making	How probability is applied in data science Decision making with probabilistic models Descriptive statistics and visualization
Distributions and parameters	Important distributions and their characteristics Methods of parameter estimation
Hypotheses testing	Evaluating uncertainty of sample estimates Tests for comparing means and goodness of fit Sequential tests
Predictive models	Mathematics of joint distributions Linear regression as a statistical tool Inference with nonlinear models

Do you remember?

- A random i.i.d. (independent and identically distributed) sample $X_1, X_2, \ldots, X_n$:
  - Each X_i has mean μ and variance σ^2
  - For any $i \neq j$, X_i and X_j are independent
- Sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- What is the variance of $\bar{X}$?
  - σ^2
  - σ^2/n
  - σ^2/n^2
- $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

The previous lecture

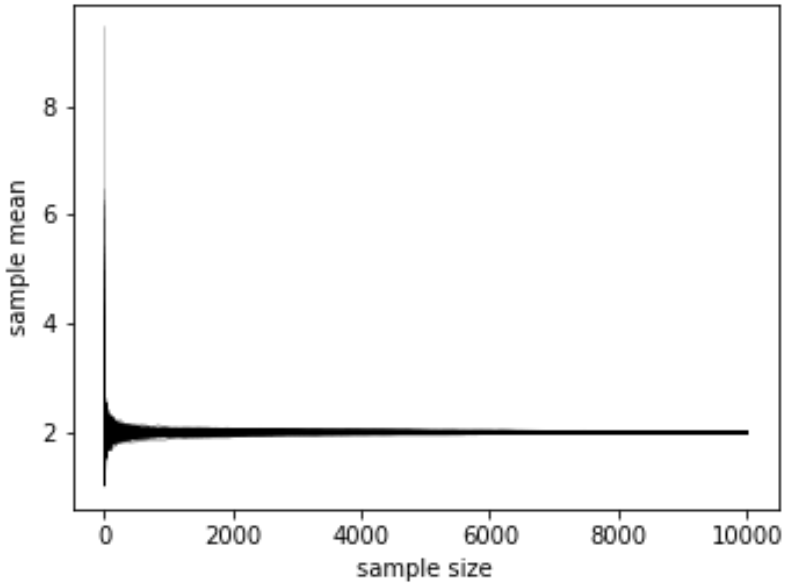
- Distributions are often defined with unknown parameters
- There are several methods of parameter estimation, e.g:
  - Solve equation $sample\ moments = population\ moments$
  - Maximize likelihood of the sample with respect to parameters
- Estimates are calculated on random samples and thus are random variables themselves
  - E.g. sample mean ($\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$): $\mathbb{E}\bar{X} = \mu, Var(\bar{X}) = \frac{\sigma^2}{n}$
  - E.g. sample variance ($s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$): $\mathbb{E}(s^2) = \sigma^2$
- Good estimators (formulas for estimates) should be unbiased ($\mathbb{E}\hat{\theta} = \theta$) and consistent ($\lim_{n\to\infty} \hat{\theta} = \theta$)

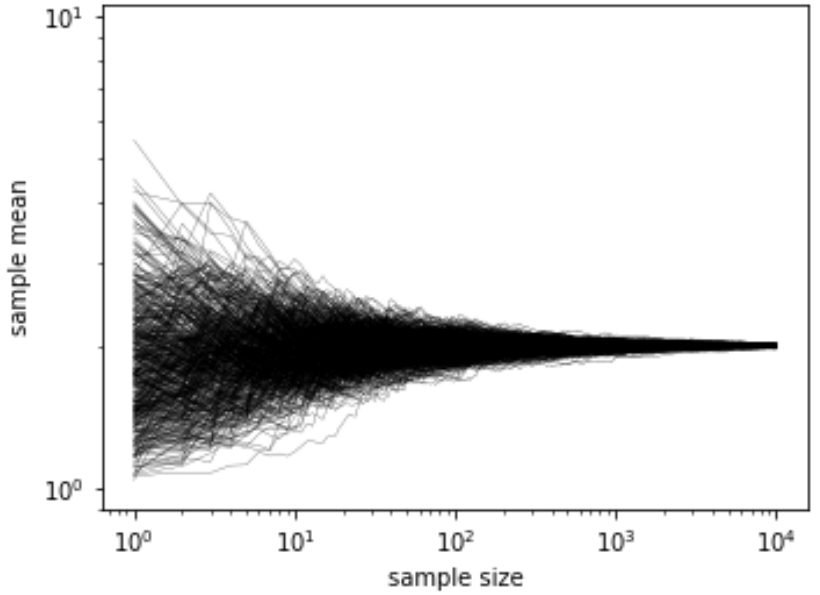
Sample means

Sample mean vs. population mean

We know that sample mean is “usually close” to the expected value in the population. What does it exactly mean?

Let’s do an experiment: increase sample size, track the sample mean, many times
As n increases, the range of $\bar{X}$ gets narrower (proportionally to $\sqrt{\sigma^2/n}$)





The law of large numbers

It's a theorem.

If we take an i.i.d. sample of n random variables $X_1, \ldots X_n$ with mean μ and variance σ^2,

and estimate its sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

then $\lim_{n\to\infty} P(|\bar{X} - \mu| > a) = 0$ for any $a > 0$

Thus, we can estimate *means* as precisely as we want – we need only a sample that is large enough.

A bonus: *probability* of any event is just the mean of the corresponding Bernoulli random variable.

Thus, we can estimate probabilities as precisely as we want.

And if we can estimate probabilities, we can estimate everything.

Proof of the LLN

If you want, just believe it. But just in case you don't:

Markov inequality: if RV X is non-negative, then for any $a > 0$

$$\mathbb{E}X = P(X \le a)\mathbb{E}(X|X \le a) + P(X > a)\mathbb{E}(X|X > a) \ge aP(X > a)$$

Thus, $P(X > a) \le \frac{\mathbb{E}X}{a}$

Chebyshev inequality: for any RV X with mean and variance μ and σ^2,

$$P(|X - \mu| > a) = P\left((X - \mu)^2 > a^2\right) \le \frac{\sigma^2}{a^2}$$

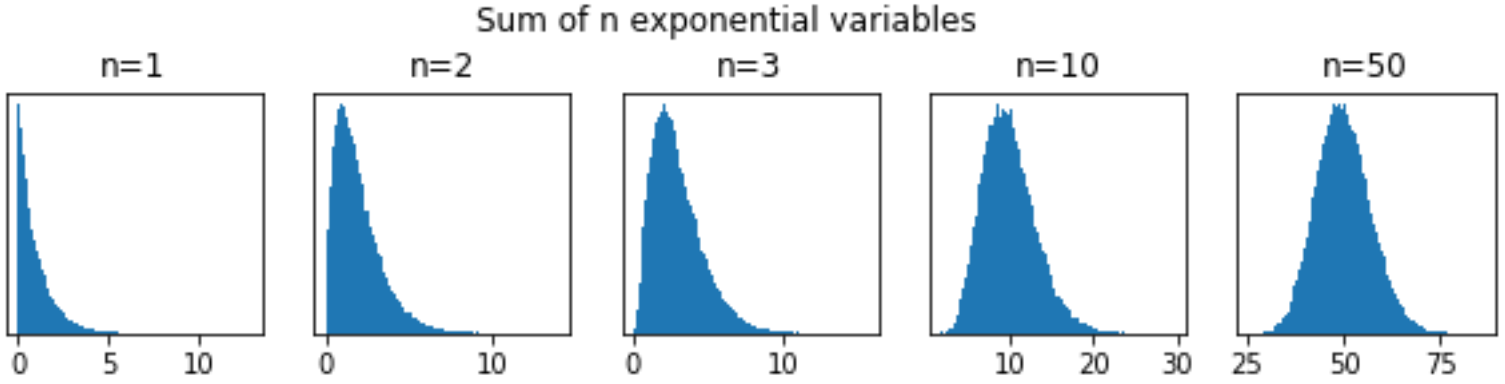
(because the variable $(X - \mu)^2$ is non-negative and has mean σ^2)

Now, if $X_1, \dots, X_k$ are independent, then $\mathbb{E}\bar{X} = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{k}$, so

$$P(|\bar{X} - \mu| > a) \le \frac{\sigma^2}{a^2 k} \to 0, as\ k \to \infty$$

The Central limit theorem

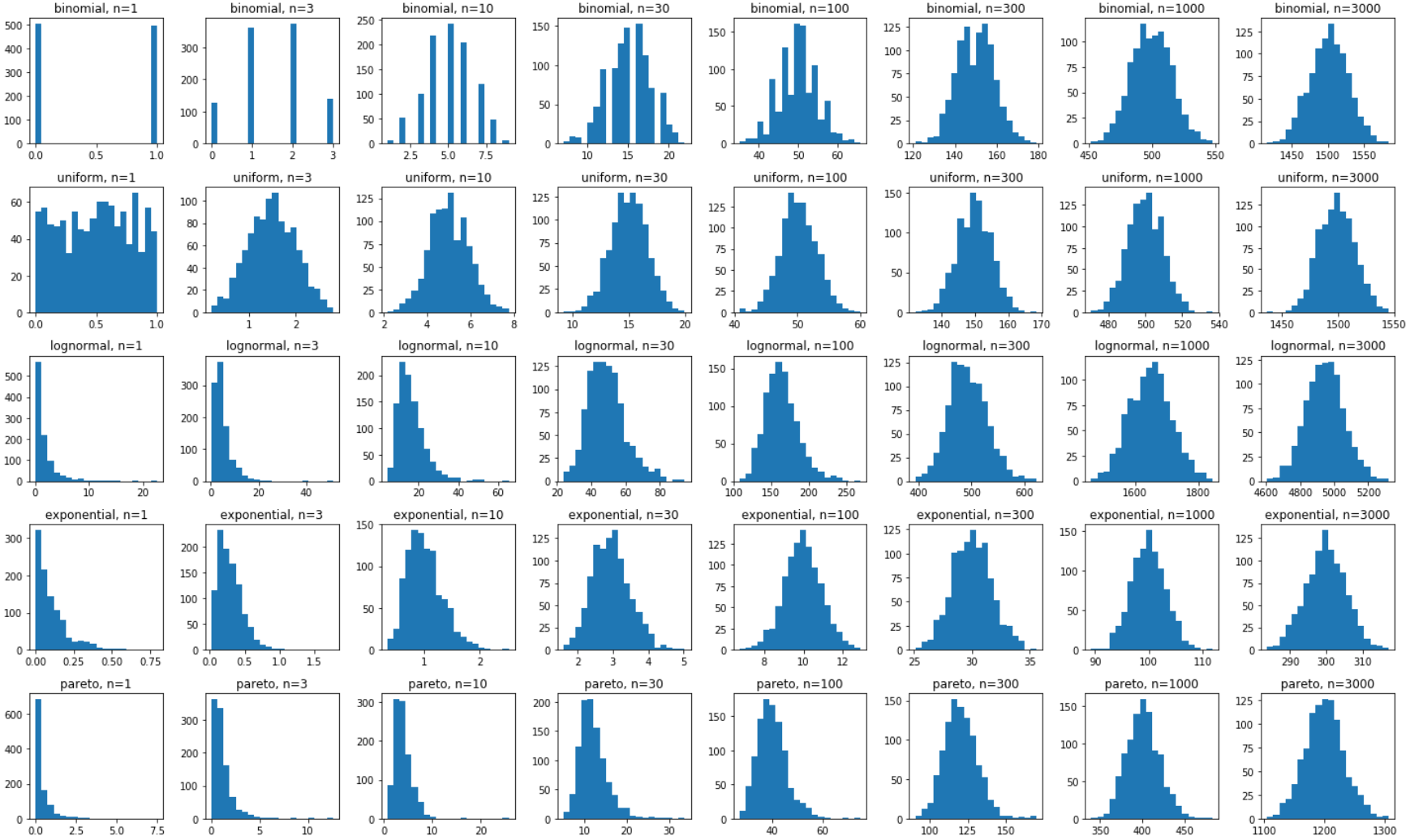
- Take an i.i.d sample $X_1, \ldots, X_n$ from (almost) any distribution with (population) mean μ and variance σ^2
- Then sum $\mathrm{S_n} = \sum_{i=1}^{n} X_i$ has mean μn and variance $\sigma^2 n$
- Normalize to avoid infinity: $\tilde{X}_i = \frac{X_i - \mu}{\sigma}$, $\tilde{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{X}_i$ has mean 0 and variance 1. But other parameters of shape are uncertain.
- At $n \to \infty$, the shape (e.g. CDF) of $\tilde{S}_n$ converges to normal.
- In practice, for $n \geq 50$, the distribution is often "normal enough"



Proof of the CLT (a draft)

- Introduce moment generating function (MGF):
  - It fully characterizes a distribution, just like a CDF or a PDF does
  - $M_X(t) = \mathbb{E}e^{tX} = \int_{-\infty}^{\infty} e^{tx} f(x)dx$
  - Then $M_X^{(k)}(t) = \mathbb{E}(X^k e^{tX})$, thus $\mathbb{E}X^k = M_X^{(k)}(0)$
  - $f_X(x)$ can be recovered from $M_X(t)$, using Fourier transform
  - If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$
  - $M_{a+bX}(t) = e^{at}M_X(bt)$
  - For normal distribution, $M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$
- We can prove that MGF of $\tilde{S}_n$ (normed sum) converges to the MGF of $\mathcal{N}(0,1)$: $e^{\frac{1}{2}t^2}$
  - Use the fact that $M_{\tilde{S}_n}(t) = M_{\tilde{X}}^n\left(\frac{t}{\sqrt{n}}\right)$
  - Prove that $\lim_{n\to\infty} \frac{\log M_{\tilde{S}_n}(t)}{t^2} = \frac{1}{2}$, by approximating M with Taylor series
- Convergence of MGF implies convergence of CDF

CLT, more examples



How can I use this all?

- Example: for testing hypotheses
  - Example. We believe that 50% of galactic population vote for Darth Vader.
    How likely is it that in a sample of 200 citizens Vader scores 90 votes or less?
  - 0.1% ? 3% ? 8% ? 42%?
- The exact distribution is binomial, but it can be approximated with normal precisely enough
  - $\mu = p = 0.5, \sigma^2 = p(1-p) = 0.5^2, \frac{\sigma^2}{n} = \frac{0.25}{200} \approx 0.035^2$
  - Sample mean of 0.45 is then 1.4 standard deviations below the population mean, which has probability $\approx 8\%$ for normal distribution
- Therefore, after such a poll we can still believe
  that in the whole population Vader has 50% support,
  but we just got into an 8% unlucky tail

```
scipy.stats.norm.cdf(-0.05 / np.sqrt(0.25 / 200))
```

```
0.07864960352514251
```

Interval estimation

Example: the Darth Vader problem

- We know that in a random sample of 200 galactic citizens 90 said they would vote for Vader.
- 45% is the sample mean, but what population mean can be?
- We know that approximately, $(\bar{x} - \mu) \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$, and $\frac{\sigma^2}{n} \approx \frac{s^2}{n} = \frac{\bar{x}(1-\bar{x})}{n-1} \approx 0.035^2$.
- Then, with $\approx 95\%$ probability, $(\bar{x} - \mu) \in [-0.07; +0.07]$
- Or we can say that with $\approx 95\%$ probability, $\mu \in [38\%, 52\%]$

Confidence intervals

- Because parameter estimate $\hat{\theta}$ is a function of random sample, it is a random variable itself
  - And the probability that $\hat{\theta} = \theta$ is usually 0
- We can also make a random interval $[l, u]$ that covers θ with some high probability $1 - \alpha$
  - It is called $(1 - \alpha)$-confidence interval
  - $1 - \alpha$ is called confidence level
  - α will be called significance level later
- Of course, θ does not have to be random, it is just unknown
- It is *the interval* that is random

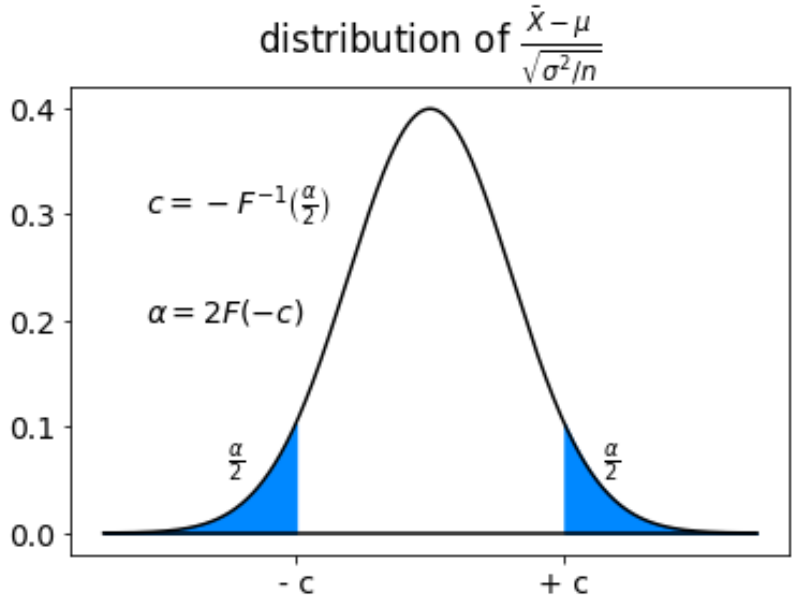
Normal confidence interval for mean

- We may believe that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ (because $X \sim \mathcal{N}$ or because of CLT)
- $P(\mu - c\sigma/\sqrt{n} \leq \bar{X} \leq \mu + c\sigma/\sqrt{n}) = 1 - \alpha$ for some c
- Then $P(\bar{X} - c\hat{\sigma}/\sqrt{n} \leq \mu \leq \bar{X} + c\hat{\sigma}/\sqrt{n}) \approx 1 - \alpha$ as well
  - For small n, this is inaccurate estimate. We can make it better by adjusting c.

Let $\bar{X} = 10, \sigma^2/n = 3$, assume normality.
$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$ is then $\mathcal{N}(0,1)$
For 95% CI probability $\alpha = 5\%$ is divided evenly between two tails
Cut points thus may be found as 2.5% and 97.5% quantiles, i.e. -1.96 and +1.96
The 95% CI is thus $10 \pm 1.96 \times 3$

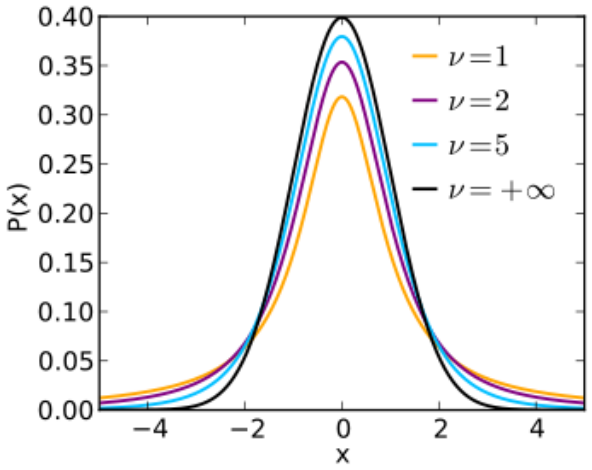
```
scipy.stats.norm.ppf([0.025, 0.975])
```

```
array([-1.95996398,  1.95996398])
```



Student distribution

- For CI construction, we used the fact that $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$
- However, when we replace unknown (non-random) σ^2 with its random estimate $\hat{\sigma}^2 = s^2$, then $t = \frac{\bar{X}-\mu}{\sqrt{\hat{\sigma}^2/n}}$ is *no longer normal*
  - Intuition: for small n, $\hat{\sigma}^2$ can be unusually small (which creates heavy tails of t's distribution) or unusually large (which creates a sharp peak)
  - For large n, however, $\hat{\sigma}^2$ converges to constant and t converges to normal distribution
- The distribution of t is called Student (T^ν).
- It has a single parameter $\nu = n - 1$, called *degrees of freedom*
- t.dist in Excel, scipy.stats.t in Python
- For small normal samples, Student distribution helps to make much more accurate confidence intervals



Example of Student CI

- We have a sample $\{2, 5, 8\}$ from normal distribution with unknown mean and variance.
- What is the approximate 90% CI for population mean?
  - $[-5, 15]$
  - $[0, 10]$
  - $[2, 8]$
  - $[4, 6]$
- $\bar{X} = 5, s^2 = \frac{(2-5)^2+0^2+(8-5)^2}{3-1} = 3^2$
- We know that $\frac{\bar{X}-\mu}{\sqrt{s^2/n}} \sim T_{3-1}$, and 5% quantile for Student with 2 d.o.f. is -2.92
- The 90% CI is thus $5 \pm 2.92\sqrt{9/3} \approx 5 \pm 5 = [0, 10]$

Sample size planning

A psychologist believes that the standard deviation of driver's reaction time is about 0.05 seconds, and wants to estimate the mean.

How large a sample of measurements must be taken to derive a confidence interval for the mean with *margin of error* (radius of CI) at most 0.01 second, and confidence level 95%?

$$(\bar{x} + 1.96\sigma/\sqrt{n}) - (\bar{x} - 1.96\sigma/\sqrt{n}) \approx 4\sigma/\sqrt{n} \leq 0.01 \times 2$$

$$n \approx \left(\frac{4\sigma}{0.02}\right)^2 = 10^2 = 100$$

Sample size planning

You expect that about half of generated texts contain errors, and want to estimate this proportion from a sample.
You want the 95% confidence interval to be $\bar{x} \pm 1\%$, i.e. to have width of 2%.
How large a sample do you need?
100? 1000? 10'000? 100'000?

95% normal CI for mean is approximately $\bar{X} \pm 2\sqrt{\frac{\sigma^2}{n}}$,

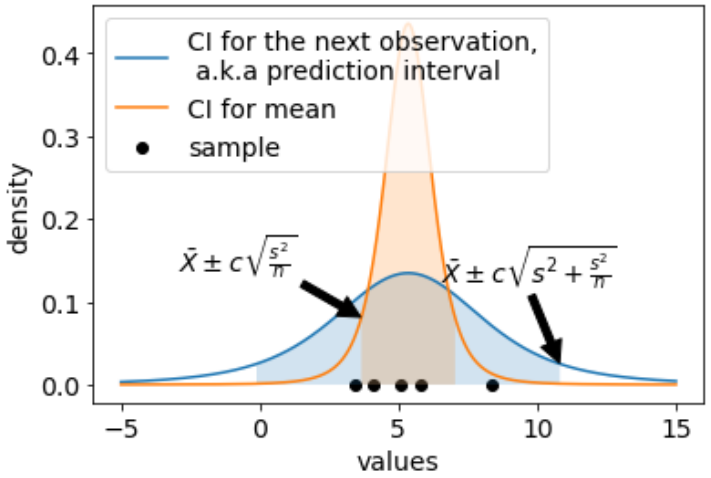
so you want $1\% = 2\sqrt{\frac{\sigma^2}{n}}$, or $\frac{\sigma^2}{n} = \frac{1}{200^2}$.

Because the original distribution is binomial with $p \approx 0.5$, you have $\sigma^2 \approx 0.5^2$.

Therefore you need $n = 200^2 \times 0.5^2 = 10000$

Prediction interval

- Confidence interval tries to cover some parameter of distribution
- Prediction interval tries to cover the random variable itself – the whole distribution
- In general, it's not that easy
- For normal distribution, there is a recipe:
  - Decompose $X - \bar{X}$ into independent $(X - \mu)$ and $(\mu - \bar{X})$
  - Their estimated variances are S^2 and $\frac{S^2}{n}$, respectively, and means are 0
  - Because of independence they can be added, so that $\frac{X-\bar{X}}{\sqrt{S^2(1+1/n)}} \sim T^{n-1}$
- Example: if $n = 10, S^2 = 110, \bar{X} = 20$, then we can predict that the next observation will fall into $20 \pm 2\sqrt{110 \times 1.1}$ with approximately 95% probability.



MLE confidence interval

- In large samples, maximum likelihood estimates are distributed normally with variance $-H^{-1}$, where H (Hessian) is the second derivative of log likelihood function
- Continue with the Binomial example

$$\log P(k) = \log\left(\binom{n}{k} p^k (1-p)^{n-k}\right) = \log\binom{n}{k} + k\log p + (n-k)\log(1-p)$$

$$\frac{\partial \log P(k)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p}$$

$$H = \frac{\partial^2 \log P(k)}{\partial p^2} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2}$$

- At the point of maximum, $\hat{p} = \frac{k}{n}$, the estimate of Hessian looks like

$$\hat{H} = -\frac{k}{k^2/n^2} - \frac{n-k}{(n-k)^2/n^2} = -\left(\frac{n^2}{k} + \frac{n^2}{n-k}\right) = -\frac{n^3}{k(n-k)} = -\frac{n}{\hat{p}(1-\hat{p})}$$

- And we can use it to express the variance of our parameter estimate

$$\hat{\sigma}_{\hat{p}}{}^2 = -\hat{H}^{-1} = \frac{\hat{p}(1-\hat{p})}{n}$$

- Now we can use it for e.g. 95% normal CI for p: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Bootstrapping

- For some estimators, it's difficult to obtain analytical sample distribution (and CI)
- Solution: approximate *sampling from population* by *sampling from the sample itself*
  - Sample with replacement (to get "independence")
  - Same size as the original sample
- Example: toy bootstrapping for sample mean
  - Original sample: $1, 5$ -> sample mean is 3
  - Resamples: $\{1,1\}, \{1,5\}, \{5,5\}, \{5,1\}$ with equal probabilities, and sample means of $1, 3, 3, 5$ respectively.
  - Thus, variance of sample mean can be estimated as $\frac{4+0+0+4}{4} = 2$
  - BTW, an analytic unbiased estimate would be $\frac{1}{2}\left(\frac{2^2+2^2}{2-1}\right) = 2$

Hypotheses testing

What are hypotheses and why test them?

- Hypothesis is a statement about population of interest
  - Usually in terms of parameters of distribution: $\theta \in \Theta_0$
- Examples:
  - Eating GMO products causes cancer
  - Users click the “Purchase” button more often if it is blue
  - The new regression model is more accurate than before
- Usually, hypotheses require alternatives:
  - Risk of cancer under <alternative diet> is same as with GMO
  - Click rate for <some other colors> button is same as for blue
  - The old model is at least as accurate as the new one
- Parameter estimates $\hat{\theta}$ are uncertain, and so are hypotheses
- But sometimes we are able to say that a specific hypothesis is very unlikely and dare *rejecting* it
- This rejection affects our knowledge of the world
  - And possibly our actions

Two approaches to hypotheses

Comparing hypotheses

- Need to choose one of two competing models
- Care about balance of false positive and negative rate
- Solution: compare (maximized) likelihoods of hypotheses
- It can be easily extended to sequential problems

Null hypothesis testing

- Need to check whether a simple model describes the data correctly
- Can only reject it, care about false reject rate
- Can calculate a test statistic and its CDF under H_0
- Solution: see how extreme your test statistic is, if the null hypothesis is true

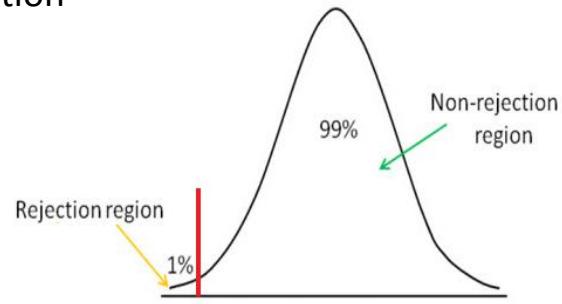
Null hypothesis (significance) testing

Idea of significance testing:

- You have null hypothesis H_0 well-specified (like $\mu = 0$), and alternative H_1 may be very broad (like $\mu > 0$ or $\mu \neq 0$)
- You can never really accept H_0, but sometimes can reject it

Process of testing:

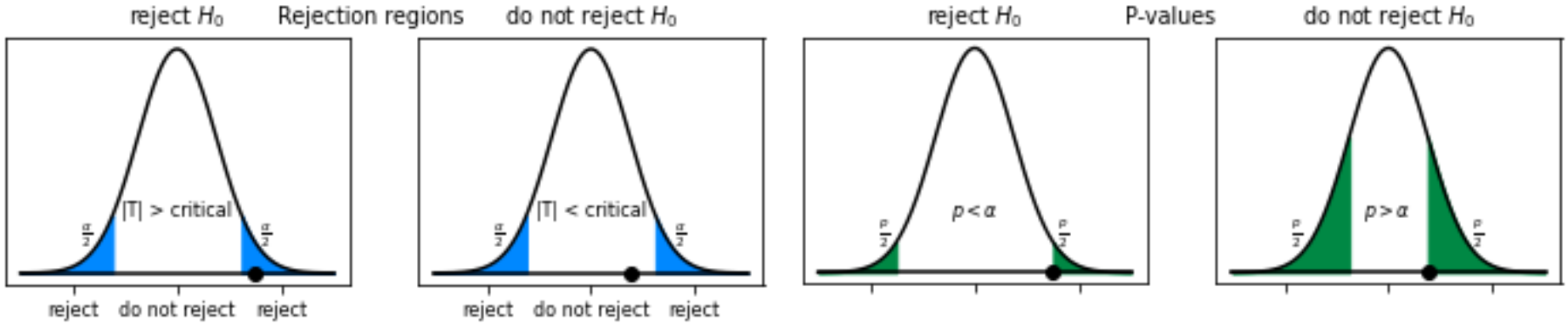
- Define *test statistic* T (e.g. $t = \frac{\bar{x}}{\hat{\sigma}/\sqrt{n}}$) and its distribution under H_0
  - Sometimes you have to approximate this distribution by simulation
- Define rejection region R for T (e.g. $t > 2.5$), such that $\alpha = P(T \in R | H_0)$ is small enough (e.g. 1%)
  - Usually it's one tail (for $>$ or $<$ in H_1) or two equal tails (for $\neq$)
- Reject H_0 if T calculated for your sample (T_{obs}) is in R
- Otherwise, collect more data or accept your failure to reject H_0



P-value

There are actually 2 equivalent ways of significance testing:

1. With rejection region
   - Define the rejection region (e.g. below $CDF_T^{-1}\left(\frac{\alpha}{2}\right)$ and above $CDF_T^{-1}\left(1-\frac{\alpha}{2}\right)$)
   - Reject the H_0 if the observed test statistic is within this region
2. With p-value
   - Calculate p-value: the probability of obtaining test results at least as extreme as the results actually observed, under H_0
   - E.g. use quantiles $1 - CDF(T)$ and $CDF(-T)$
   - Reject the H_0 if p-value is less than α



Test it yourself!

- Out of 100 respondents, 60 say that logo A looks better than logo B.
- Can we reject the hypothesis that A and B are equally attractive at the 95% confidence level, vs. the alternative that A is better?
  - Yes
  - No
- $H_0: \mu = 0.5, H_1: \mu > 0.5$
- The test statistic is $z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/100}}$, and if H_0 is true, it equals $\frac{\bar{X}-\mu}{\sqrt{0.5^2/100}}$ and is distributed approximately $\mathcal{N}(0,1)$
- In our sample, $z = \frac{0.6-0.5}{\sqrt{0.5^2/100}} = \frac{0.1}{0.05} = 2$, and

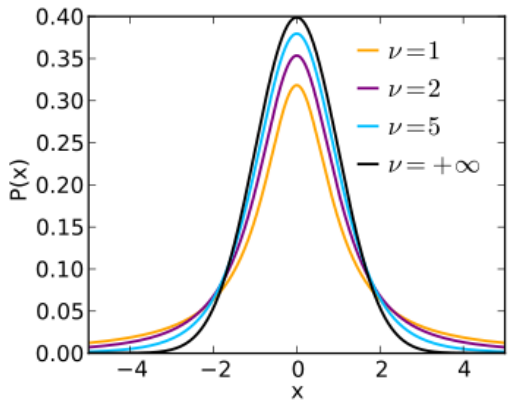
$$p\_value = P(z > 2) = 1 - CDF_{\mathcal{N}(0,1)}(2) = 2.2\%$$

- Therefore, H_0 can be rejected at 5% significance level

Distributions under H_0: Student

With a sample $X_1, \dots, X_n$, i.i.d. $\mathcal{N}(\mu, \sigma^2)$, what can we test?

- Test for $\mu = \mu_0$, σ is known
  - Under H_0, $\bar{x}$ has normal distribution with parameters $\mu_0, \frac{\sigma^2}{n}$
- Test for $\mu = \mu_0$, σ is unknown
  - Under H_0, $t = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}$ has Student distribution with $n-1$ degrees of freedom



Example: sample 1, 3, 4, 7, 8, 11, 14

$H_0: \mu = 5, H_1: \mu > 5$

$$t = \frac{\bar{x} - 5}{\hat{\sigma}/\sqrt{7}} \approx 1.07$$

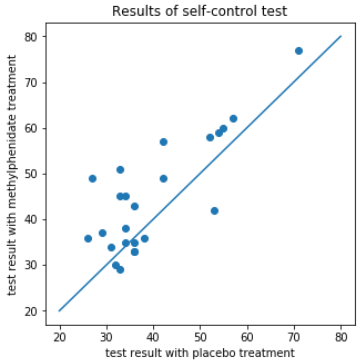
```
1 - scipy.stats.t(df=6).cdf(1.07)
```

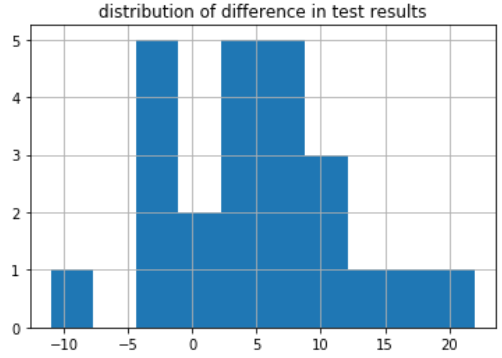
```
0.16288153956353701
```

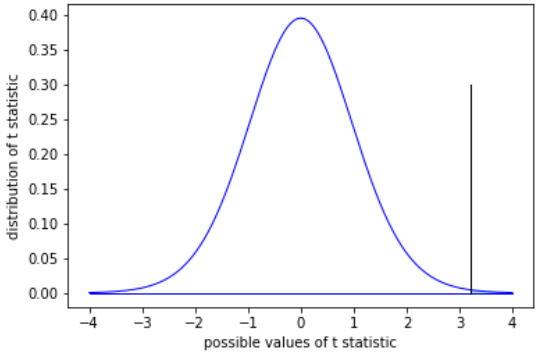
Rejection region is $\{t > 1.07\}$,
and its probability under H_0 (**p-value**) is 16%
So it seems we don't have enough evidence to reject H_0

Student test (T-test) for paired samples

- If $X_1, \dots X_n$ are i.i.d. normal, then $T = \frac{\bar{X} - \mathbb{E}X}{\hat{\sigma}_X / \sqrt{n}} \sim t_{n-1}$
- Example: null hypothesis is that methylphenidate does not help to treat ADHD (we would like to reject it),
- In the experiment, 24 children were treated in turn with placebo and methylphenidate , and took test for self control
  - Source: Pearson D.A, Santos C.W., Casat C.D., et al. (2004)
- We believe that differences in test results are i.i.d. normal, and want to test whether their mean is 0 (against the alternative that it is positive)







In our example, $T = 3.22$, and for 23 degrees of freedom probability of larger values is 0.18%

This probability is low enough to reject H_0, and allow ourselves to believe that methylphenidate is effective

Comparing means of independent samples

- Sometimes, we want to compare means in two samples of unrelated objects
- Again, a Student test statistic can be constructed
- $T = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\widehat{Var}(\bar{X}_1) + \widehat{Var}(\bar{X}_2)}}$
- If we believe that variances of two samples are different, the denominator is $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
- If we believe that variance is the same, the denominator is $\sqrt{\frac{S_{12}^2}{n_1} + \frac{S_{12}^2}{n_2}}$, where $S_{12}^2 = \frac{\sum(X_{i1} - \overline{X_1})^2 + \sum(X_{i2} - \overline{X_2})^2}{n_1 + n_2 - 2}$
- In both cases, number of degrees of freedom is approximately $n_1 + n_2 - 2$

Compare two samples

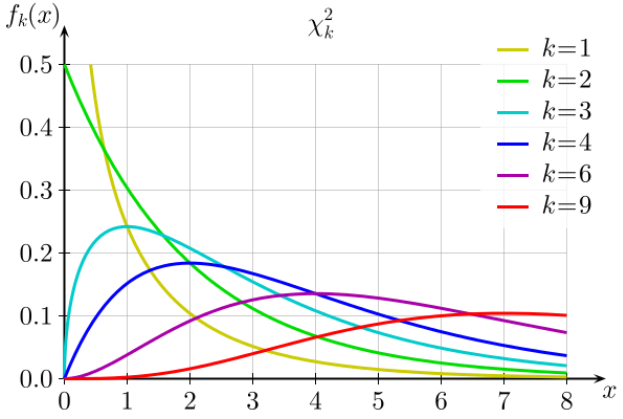
- Compare means of two normally distributed groups:
  - Sample 1: size 25, mean 15, std 15
  - Sample 2: size 63, mean 20, std 21
- Test H_0 that $\mu_1 = \mu_2$ vs. H_1 $\mu_1 \neq \mu_2$ at 10% significance
  - Reject H_0
  - Cannot reject H_0
- Solution:
  - Test statistic $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\widehat{Var}(\bar{X}_1 - \bar{X}_2)}} \sim T_{25+63-2}$ under H_0
  - $Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{15^2}{25} + \frac{21^2}{63} = 16$
  - $t = \frac{15-20}{\sqrt{16}} = -1.25$
  - Two-sided p-value is $2\ \mathrm{CDF}_{\mathrm{T}(83)}(-1.25) \approx 21\%$
  - The null hypothesis cannot be rejected

Tests for goodness of fit

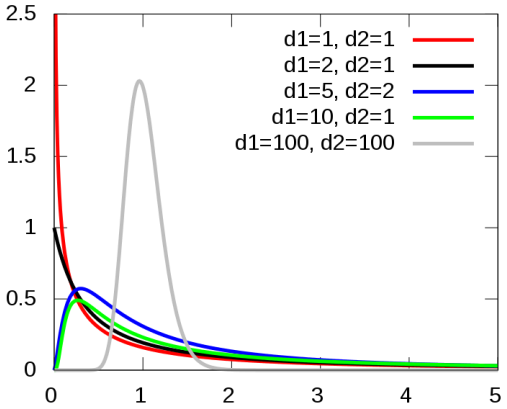
Tests for variance: χ^2 and Fisher

- Test for $\sigma^2 = \sigma_0^2$
  - Under H_0,$(n-1)\frac{\hat{\sigma}^2}{\sigma_0^2}$ has Chi-squared (χ^2) distribution with $n-1$ degrees of freedom
  - Generally, χ_n^2 is distribution of $(X_1^2 + \cdots + X_n^2)$, where X_i are i.i.d. standard normal
- Test that variances in two samples (sizes n_1 and n_2) are equal
  - Under H_0, $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ has Fisher (F) distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom
  - Generally, $F_{n,m}$ is distribution of $\frac{Y_n/n}{Y_m/m}$, where Y_i are independent χ_i^2

χ^2 PDF for different d.o.f.



F PDF for different d.o.f.



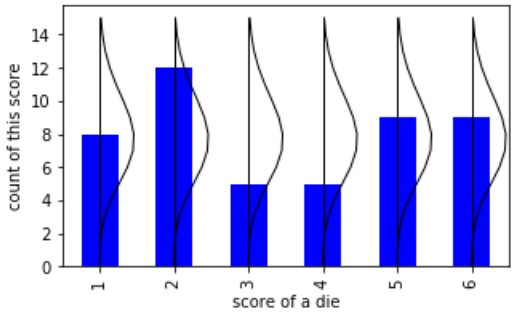

```
scipy.stats.chi2(n)
scipy.stats.f(n, m)
```

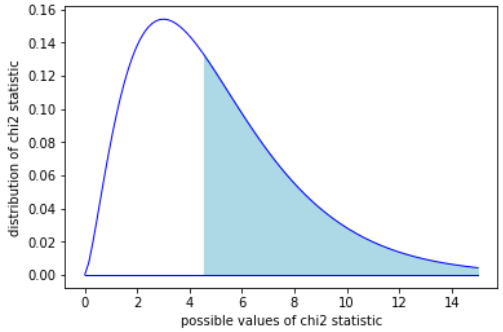
Chi-squared test

- Chi-squared test for discrete variables: $\sum \frac{(O-E)^2}{E}$ (observed vs expected counts)
  - Goodness of fit for a particular distribution
  - Comparison of distribution for two samples
  - Checking independence of two variables

Score	1	2	3	4	5	6
count	8	12	5	5	9	9
expected	8	8	8	8	8	8

- Example: is a die fair?
- If it is true (call this hypothesis H_0), then:
  - Distribution of every count C_i is binomial $\left(n = 48, p = \frac{1}{6}\right)$
  - It can be roughly approximated with $\mathcal{N}(8, 20/3)$
  - Counts are dependent (their sum is 48), but any 5 of them are (approximately) independent
- We can normalize counts, so that $\frac{C_i - 8}{\sqrt{20/3}} \sim \mathcal{N}(0,1)$
- Then $T = \sum_{i=1}^{6} \left(\frac{C_i - 8}{\sqrt{8}}\right)^2 \sim \chi_5^2$ (in our case $T_{obs} = 4.5$)
  - We replace $np(1-p)$ with np in the denominator to correct for dependency between cells (just believe it)
  - Here 5 is number of degrees of freedom
- If H_0 is not true, then T should be larger
- P-value: $P(T > T_{obs} | H_0) = 47\%$ - so we cannot reject H_0





Chi2 test for independence

	sales_channel_id									
education	**2**	**3**	**4**	**5**	**6**	**10**	**13**	**14**	**16**	**All**
A	1254	128	0	39	189	0	0	0	4	1614
H	105874	24262	13	3078	11065	17	581	28	345	145263
HH	6311	955	2	225	1038	1	15	0	41	8588
S	10724	1288	2	204	1361	1	64	0	37	13681
SS	55070	10035	14	1010	5007	11	400	5	107	71659
UH	20031	3615	2	761	4001	2	156	6	72	28646
US	622	84	0	15	168	0	0	1	5	895
All	199886	40367	33	5332	22829	32	1216	40	611	270346

Are these two variables independent? We don't know.
If they are independent, and marginal probabilities are fixed, then we can calculate probability for each joint cell (as product of probabilities).
If we sample from this (multinomial) distribution n observations, for a cell i (with probability p_i) number of counts will be approximately $\mathcal{N}(np_i, np_i)$.
We can renormalize all these counts, and construct a χ^2 distribution of them.
Number of degrees of freedom is $(r-1)(c-1)$ (difference between H_0 and H_1).
P-value is 0.00 ($\chi^2_{obs} = 3095$ with 48 dof), so we reject independence hypothesis.

Use chi2 test for independence

Are sex and smoking independent?

- Yes, they are independent
- No, they are not

Under independence hypothesis, expected counts are these:

	Female	Male
Non-smoker	68	60
Smoker	12	20

	Female	Male
Non-smoker	64	64
Smoker	16	16

$Y = \frac{(68-64)^2}{64} + \frac{(60-64)^2}{64} + \frac{(12-16)^2}{16} + \frac{(20-16)^2}{16} = \frac{1}{4} + \frac{1}{4} + 1 + 1 = 2.5 \approx 1.6^2 \sim \chi_1^2$, and χ_1^2 is $\left(\mathcal{N}(0,1)\right)^2$

P-value is $1 - CDF_{\chi_1^2}(2.5) \approx 2CDF_{\mathcal{N}(0,1)}(-1.6) \approx 11\%$
So we cannot reject H_0 of independence

In Scipy, Yates' correction for continuity is applied, so the value of Y is slightly different, but the conclusion is similar

```
scipy.stats.chi2_contingency([[68, 60], [12, 20]])
(1.9140625,
 0.1665126870220502,
 1,
 array([[64., 64.],
        [16., 16.]]))
```

Chi-2 statistic

p-value

degrees of freedom

expected counts

Other statistics and distributions

- Case 1: samples from normal distribution
  - Test statistics usually have one of $\mathcal{N}$, t, χ^2, F distributions
- Case 2: large samples from any distribution
  - Sample estimates usually have approximately $\mathcal{N}$ distribution due to the CLT
  - Thus $\mathcal{N}$, χ^2, F may be good approximations for test statistics (as with the χ^2 test)
- Case 3: small samples from non-normal distribution
  - Solution 1: use distribution-independent tests (that use only ranks of the observation, not the specific values)
  - Solution 2: approximate the distribution of the test statistic under H_0 by computer simulation

Simulating a distribution

- What you have: a sample $X_1, \ldots X_n$, H_0, and a formula for the test statistic $t(sample)$
- What you want: a p-value for your test
- What to do:

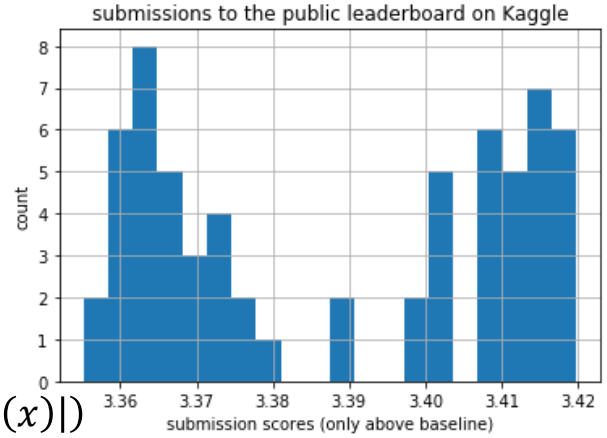
```
real_statistic = test_statistic(sample)
n = size(sample)

simulated_stats = []
for i in range(n_iterations):
    simulated_sample = random.sample(distribution=H0, size=n)
    simulated_stats.append(test_statistic(simulated_sample))

p_value = mean(real_statistic > simulated_stats)
```

KS test for CDF goodness of fit

- How to generally tell whether a CDF describes the continuous distribution of your sample well?

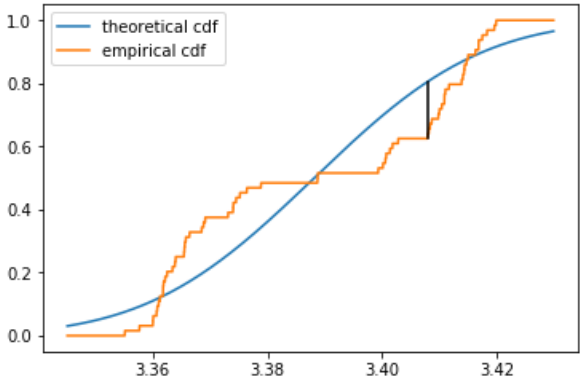


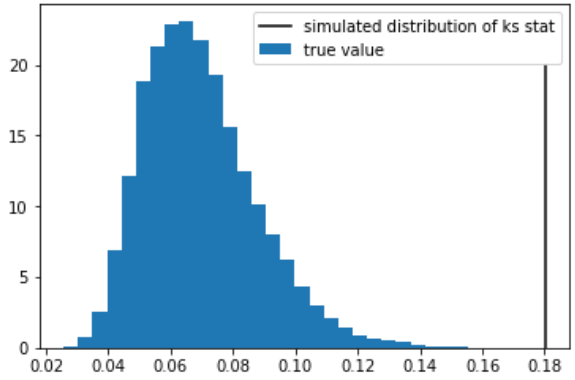
A recipe from Kolmogorov and Smirnov:

- Take your hypothetical CDF $F_0(x)$
- Calculate the empirical CDF $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}[x_i \leq x]$
- Calculate their maximum difference: $K = \max_x(|F_0(x) - F_n(x)|)$
- This K under H_0 has known KS distribution, with which a p-value can be calculated

```
dist = scipy.stats.norm(loc=x.mean(), scale=x.std())
scipy.stats.kstest(x.values, dist.cdf)
```

```
KstestResult(statistic=0.18007950517067428, pvalue=0.02747319082442845)
```





Calculating the KS statistic

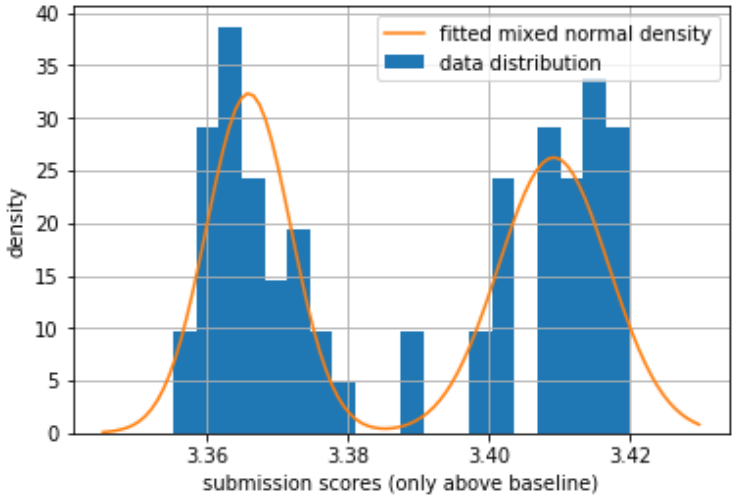
- At each point X_i of the *sorted* dataset $X_1, \ldots, X_n$, the empirical CDF is just i/n by definition
- KS statistic is the maximum absolute difference between empirical and theoretical CDF
- Calculate the KS statistic for the dataset $\{0.7, 0.2, 0.5\}$ under H_0 that the distribution is uniform on $[0, 1]$
  - 0.1 ? 0.2 ? 0.3 ? 0.4 ? 0.5 ?
- The sorted dataset is $[0.2, 0.5, 0.7]$ with theoretical CDF $[0.2, 0.5, 0.7]$ (because $F_X(x) = x$ on $[0, 1]$) and empirical CDF $\left[\frac{1}{3}, \frac{2}{3}, \frac{3}{3}\right]$, so the maximal difference is 0.3.

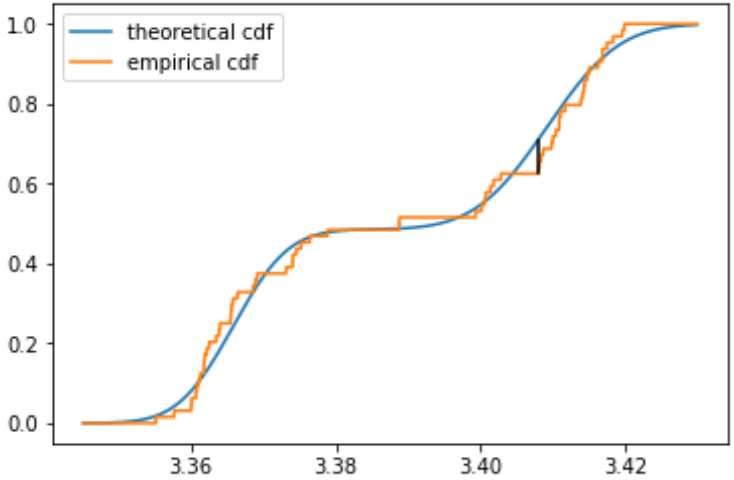
More KS testing

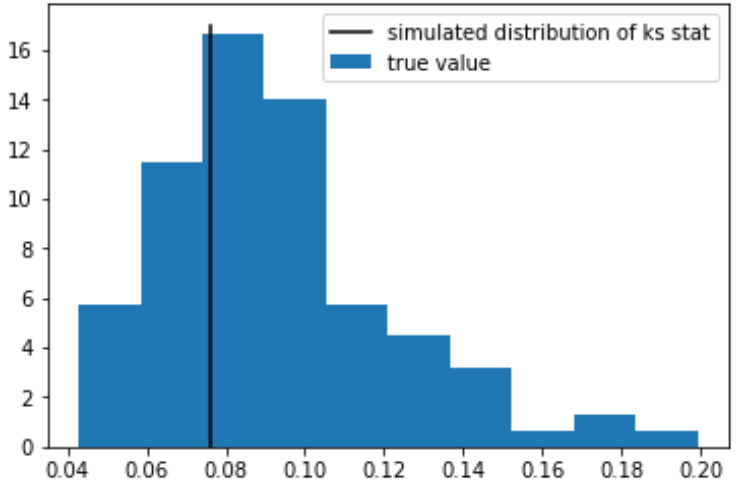
If one normal distribution cannot fit the data, maybe a mixture of two normal distributions is ok?

It seems so, but we have to maximize likelihood numerically to find the best parameters.

Fortunately, scipy can do this for us.







Fitting and testing a mixture: code

```
import scipy.optimize, scipy.stats

def mixture(params, out='pdf'):
    # cdf or pdf of a mixture of 2 normal distributions
    mu1, s1, mu2, s2, p1 = params
    p2 = 1 - p1
    if min(s1, s2, p1, p2) <= 0: return lambda x: np.zeros_like(x) + 1e-30
    dists = [scipy.stats.norm(mu1, s1), scipy.stats.norm(mu2, s2)]
    f1, f2 = [getattr(d, out) for d in dists]
    return lambda x: f1(x) * p1 + f2(x) * p2

def nll(params):
    return - sum(np.log(mixture(params)(x)))

result = scipy.optimize.minimize(nll, [3.37, 0.02, 3.41, 0.02, 0.5], tol=1e-4)
params = result.x
print(params)
print(scipy.stats.kstest(x, mixture(params, 'cdf')))
```

```
[3.36584676 0.00598263 3.40915116 0.00782929 0.48503462]
KstestResult(statistic=0.08379016620552004, pvalue=0.728074869010681 5)
```

Comparing hypotheses

Comparing simple hypotheses

- If each of two hypotheses (H_0 and H_1) fully specifies the distribution of data, then we can compare them directly – by likelihood.
- Likelihood ratio: $\Lambda = \frac{P(data|H_0)}{P(data|H_1)} = \frac{L(H_0)}{L(H_1)}$ (p is probability or density)
- Choose H_0, if $\Lambda > h$ (some constant), and H_1 otherwise
- Possible outcomes of test are:

	In fact, H_0 is true	In fact, H_1 is true
Choose H_0 (accept H_0)	True negative	False negative (type II error)
Choose H_1 (reject H_0)	False positive (type I error)	True positive

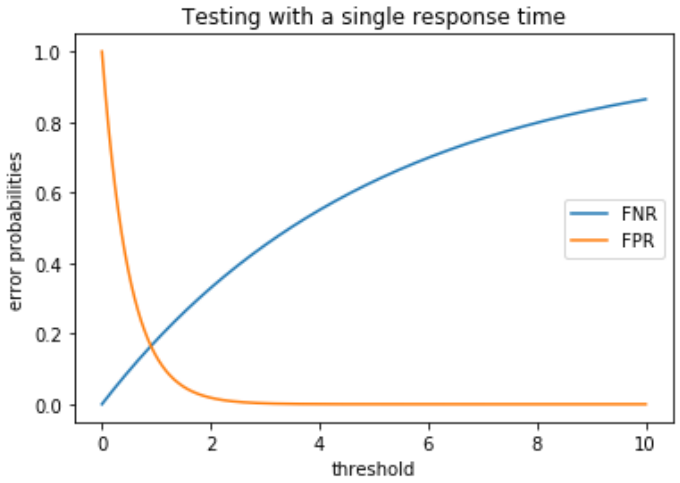
- False negative rate: $\beta = P(accept\ H_0|H_1\ is\ true)$
- False positive rate: $\alpha = P(reject\ H_0|H_0\ is\ true)$
- Likelihood ratio test has smallest possible β for given α (Neyman, Pearson)

Calculating likelihood ratio

- The observed dataset is $\{60, 91, 98\}$
- The distribution may be discrete uniform over $1,2, \ldots, 200$ (H_0) or over $1,2, \ldots, 100$ (H_1).
- What is the likelihood ratio?
  - 1
  - 0.5
  - 0.25
  - 0.125
- Under H_0, likelihood of any point is $\frac{1}{200}$, and under H_1 it is $\frac{1}{100}$
- Thus the likelihood ratio is $\frac{p(\text{data}|\text{H}_0)}{\text{p}(\text{data}|\text{H}_1)} = \left(\frac{1/200}{1/100}\right)^3 = 0.125$

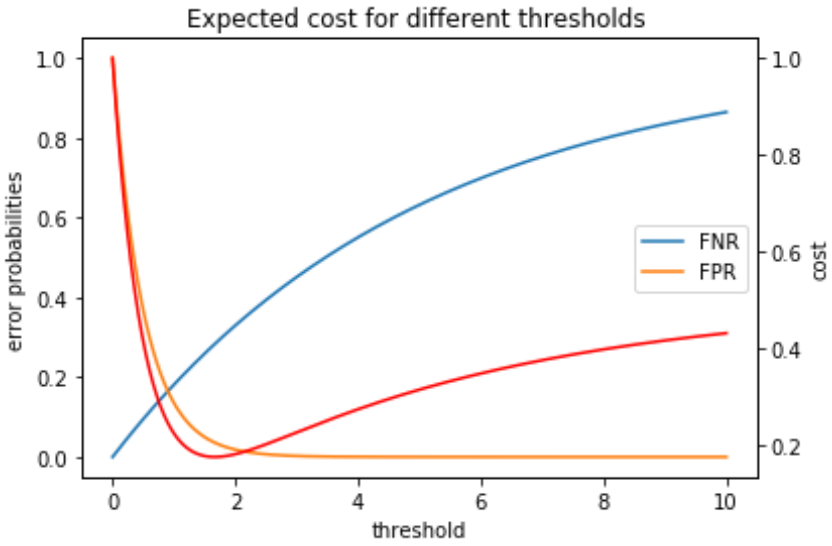
Example of a test

- Need to decide whether an online service has problems with the database (H_1), or not (H_0)
  - Both in case of problems and in normal case distribution of response time is exponential,
  - Mean response times are 5 and 0.5 seconds, respectively.
- Let's start with a dataset of a single observation, X
- Trigger the alarm, if $\frac{L_1}{L_0} = \frac{1/5e^{-1/5X}}{1/0.5e^{-1/0.5X}} = 0.1e^{1.8X} > h$
  - It means, alarm if $X > c$
  - FPR: $P(X > c|H_0) = e^{-1/0.5c}$
  - FNR: $P(X \leq c|H_1) = 1 - e^{-1/5c}$



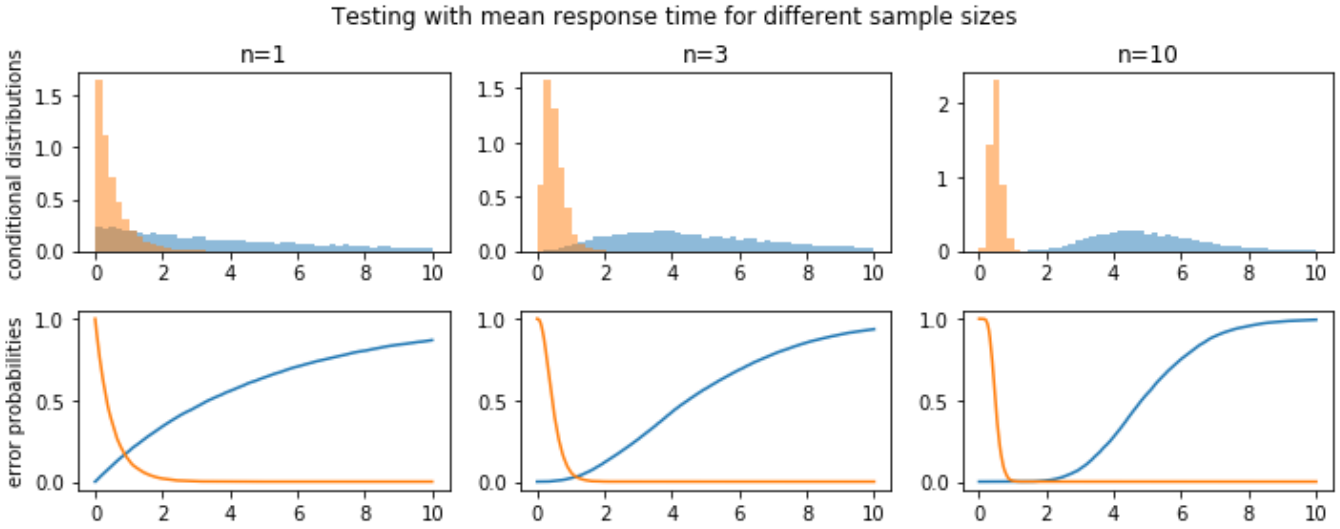
Decision making

- Choose H_1, if $\frac{L(H_1)}{L(H_0)} > h$ (some constant), and H_0 otherwise
- How to choose h? Let's consider the cost of errors
  - C_0 is the cost of false reject of H_0 (with probability α)
  - C_1 - cost of false accept of H_0 (with probability β)
- Expected cost $C = C_0 P(H_0)\alpha(h) + C_1 P(H_1)\beta(h)$
  - Choose the threshold h that minimizes it
- Example: alarms
  - Problems happen 0.1% of the time
  - Cost of false alarm is $1
  - Cost of inaction in case of problem is $500



Sequential testing

- What to do if both FPR and FNR are too high?
- Collect more data!
- Sequential test: estimate $\Lambda_t = \frac{P(x_1, \ldots, x_t|H_0)}{P(x_1, \ldots, x_t|H_1)}$
  - Accept H_0 if $\Lambda_t > h_0$
  - Accept H_1 if $\Lambda_t < h_1$
  - If $h_1 \leq \Lambda_t \leq h_0$, just wait until $t + 1$ and repeat
  - The thresholds may vary themselves depend on t



Composite hypotheses

- Sometimes, the alternative hypothesis H_1 is very broad, like $\theta \neq \theta_0$.
  - Cancer risk under any diet without GMO
  - Any click-trough rate not higher than 3%
- How can we calculate likelihood $L(H_1)$ then?
  - Choose the strongest competitor: $L(H_1) = \max_{\theta \in \Theta_1} L(\theta)$
  - We can do the same even if H_0 is composite as well
- Asymptotic distribution (Wilks' theorem):
  - Let H_0 be a special case and H_1 is more common (e.g. $\theta = \theta_0$ vs. $\theta \neq \theta_0$)
  - Then in large samples, distribution of $-2 \ln \Lambda$ approaches χ^2_k, where k is the difference in d.o.f. between H_0 and H_1

Likelihood ratio testing

- H_1: distribution of X is a mixture of two normal distribution (5 params), log-likelihood is -100
- H_0: X is normal (2 params), log-likelihood is -105
- Can we reject H_0 using Wilks' theorem at 5% level?
  - Yes
  - No
  - It's not applicable with the given data
- The test statistic is $-2\ln\frac{L_0}{L_1} = -2(-105 + 100) = 10$
- Under the null hypothesis, its distribution is χ^2_{5-2} so the p-value is $1 - CDF_{\chi^2_3}(10) \approx 1.8\%$, so we can reject H_0

```
1 - scipy.stats.chi2(3).cdf(10)
```

```
0.0185661354630432
```

An alternative: multi-armed bandits

- We usually test hypotheses in order to choose the best strategy for making decisions: collect data -> perform a test -> make a decision -> change behavior
- Until a traditional test completes, we make many suboptimal decisions
- We can start making the (seemingly) best decision more often, even before the test completes
  - Example: choose the option with the best upper limit of $\alpha\%$ confidence interval
  - As the CI shrinks with more samples, we may prefer another option
  - There are lots of other strategies – mostly heuristic
- There is a trade-off between exploration (trying different decisions) and exploitation (sticking to the decision which seems the best so far)
  - We can regulate it e.g. by tuning α
- The problem of multi-armed bandits studies just this

An example of MAB

- We show banner A or banner B, and need to maximize click through rate.
  - Current stats: 20/1000 clicks on A (2% CTR), 3/200 clicks on B (1.5% CTR)
  - Should we keep trying B or discard it and stick to A?
- ε-greedy strategy:
  - Choose the best option (A) with $1-\varepsilon$ probability,
    and choose uniformly (50% A, 50% B) with ε probability to keep exploring
- Upper confidence bound strategy
  - Choose the option with the highest UCB on the reward
  - E.g. for 95% confidence level the UCB is $2\% + 1.96\sqrt{\frac{0.02\times 0.98}{1000}} \approx 2.8\%$ for A

    and $1.5\% + 1.96\sqrt{\frac{0.015\times 0.985}{200}} \approx 3.2\%$ for B, so we should try more B now.
- Thompson sampling
  - assume a prior distribution of rewards for every option, and choose an option with the posterior probability that it will be the best.

A zoo of tests

- Over 100 years of mathematical statistics, lots of tests were invented:
  - T-test (for comparing means of normal variables)
  - Chi-squared test (for categorical goodness of fit or independence)
  - Kolmogorov-Smirnov test (continuous goodness of fit)
  - Mann-Whitney test (for comparing medians)
  - Wilcoxon test (test for paired change direction in non-normal variables)
  - ANOVA (F-test): (for comparing means and variances in groups)
- Hard (and not necessary) to learn them all
- With large samples, almost any test is good
- If there is no known test for your case, you can use simulation or bootstrap (and need to invent a test statistic)

Conclusion

- Hypotheses testing = answering questions how well one or another distribution describes the data
- If only one (null) hypothesis is detailed enough, we can try to reject it by looking at test statistics
- If two hypotheses describe the distribution completely, we can just compare likelihoods
- There is a large zoo of test statistics for all kinds of hypotheses

What to do next

- 3 more problems to solve
  - Due in 9 days
- Recommended reading/watching:
  - *A Modern Introduction to Probability and Statistics* by F.M. Dekking - chapters 25-28
  - *Probability and Statistics in Data Science using Python* course on EDX – topic 13