Day 3 Frequentist statistics

Recap of last week...

- Common distributions: Bernoulli, Binomial, Poisson, Normal, Gamma and χ^2
- Characteristic functions $\phi_X(t) = \langle e^{itX} \rangle$
- Statistics of a sample $X_1, X_2, \cdots, X_n$ from a population with μ and σ^2

  sample mean $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_n$ where $\langle \overline{X} \rangle = \mu$

  sample variance $s_u^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$ where $\langle s_u^2 \rangle = \sigma^2$
- Law of large numbers $\overline{X}_n \to \mu$ for many samples
- Central limit theorem: distribution of $\overline{X}_n$ tends to a normal with mean μ and variance σ^2/n. NB $\sigma/n^{1/2}$ is the **standard error on mean**.

3.1 Interval estimation

3.1.1 Analysis of house-price data

- UK House price data. 105,238 properties less than or equal to £2M.
- Population mean=£290.6k and std=£229.2k so highly skewed distribution.
- Imagine we draw a sample of $n = 1000$ which is $\sim 1\%$ of the data.
- For a particular sample we get point esimations:
  sample mean μ_s =£281.0k, sample std=£217.0k and sem σ_{sem} =£6.9k
- How can we estimate a range for the population mean?
- Let the 95% confidence interval be the range encompassing the central 95$\%$
  of the normal distribution with μ_s, σ_s
- **Question:** Use invlogcdf in the Distributions package to find this range.

3.1.2 Confidence intervals

- **NB** Confidence intervals are random quantities - depend on a sample.
- Frequently misunderstood! For example, a 95$\%$ confidence interval...
- Does **not** mean a 95$\%$ chance the population mean is in that range.
- It means that 95$\%$ of confidence intervals will include the population mean.

3.1.3 Question: confidence intervals

- 100 samples from a chocolate-bar machine had weights with mean $\mu_s = 101$g and standard deviation σ_s=5g.
- Give 99$\%$ confidence limits for the mean weight of all bars made.

3.2 Hypothesis testing

3.2.1 Type I and Type II errors

- Want to distinguish between a null hypothesis H_0 and other hypothesis H_1 etc .

- Frequentist statistics: concerned with decisions made about the null hypothesis.

- Four possibilities: null can be true or false and we can accept or reject it.

	true	**false**
accept	success	type-II error
reject	type-I error	success

- Type I error: probability reject a true null hypothesis $P(\text{reject} \mid \text{null true})$
  Use α for type-I error "rate".

- Type II error: we accept a false null hypothesis $P(\text{accept} \mid \text{null false})$
  Use β for type-II error "rate".

- Type I errors tend to be main focus.
  Related to **falsifiability** of an existing theory.
  Fits with Popper's idea of good theories being falsifiable.

3.2.1 Type I and II error example

- Imagine there is a new blood test for stroke.
- Stroke or not is then confirmed clinically, later.
- 100 people suspected of having a stroke are tested.

- Hypothesis: patient has had a stroke

	true	**false**
accept	88	4
reject	2	6

- **Question:** What is the type-I error rate?

- **Question:** What is the type-II error rate?

3.2.2 Significance and p-values

Example Test if a coin is fair - this is the **null** hypothesis.

- We imagine doing one experiment with n flips.

- If coin not fair anticipate result will look unlikely from null statistics.

- We fix ahead a probability level α - the **significance** - type I error rate.
  This the level at which we reject the null because result unlikely if null true.
  Typically this is set at 5% or 1%.

- Do an experiment and measure the p-value. The probability

Experiment
Coin flip gives 7 heads out of 8 flips. What is the p-value? Is it significant at 5%?

heads	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
prob	0.004	0.0312	0.1094	0.2188	0.2734	0.2188	0.1094	0.0312	0.0

- As extreme is 0,1 7 or 8 heads so p-value is $2 \times (0.0039 + 0.0312) = 7\%$.
- Call this **Not sigificant** as greater that then $\alpha = 5\%$ cut-off.
- This is a two sided test.

- Imagine if we were testing if the coin was biased towards heads.
  As extreme are 7 or 8, so p-value$= 3.5\%$ which is significant at the 5% level.

3.2.2 Question Significance and p-values

- A nationwide school test has mean mark $\mu = 75$ with standard dev. $\sigma = 7$.
- A particular school with 30 children has a mean mark $\mu_s = 72$.

- **Question.** Is the school significantly different from the national average?
  Give the p-value and compare to 5%.

- Hint 1: use cdf(Normal()) to calculate the area under the normal curve.
- Hint 2: you can use invlogcdf to go fro the cdf to x, when necessary.

3.2.3 Type II errors and operator curves

- Decide coin fair: 100 flips there are between $40 - 60$ heads (inclusive).

- The type I error rate is therefore $\alpha = 0.0352$.

- Imagine $p = 0.6$
  Question: what is the type II error β?

3.3 Some common tests

3.3.1 One sample test

- Consider the sample mean $\overline{X}$ of a set $\{X_k\}$of n random numbers.
- Assume n sufficiently large that the sample-mean distribution is normal.
- Get an unbiased estimate of population σ from data (n should be big for this).
- Null hypothesis is that the sample mean is μ_0 (often 0, as can subtract this out).
- This is called a **z-test**: examine z-statistic $z = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ for significance

In [7]:
```
n=10
X=randn(n) .+0.1;
mx=mean(X); sx=std(X); semx=sx/sqrt(n);
zscore=mx/semx; pvalue=2*cdf.(Normal(),-abs(zscore))
println("X=$(round.(X;digits=2))")
println("mean(X)=$(round(mx;digits=3)) and sem(X)=$(round(semx;di
gits=3))")
println("z-score=$(round(zscore;digits=3)) and pvalue=$(round(pva
lue;digits=3))")
println("------------------------------")
OneSampleZTest(X)
```

```
X=[-1.29, -1.46, 0.26, 1.27, -1.12, -0.13, 0.58, 0.53, -0.71,
1.7]
mean(X)=-0.037 and sem(X)=0.345
z-score=-0.108 and pvalue=0.914
------------------------------
```

Out[7]:
```
One sample z-test
-----------------
Population details:
    parameter of interest:   Mean
    value under h_0:         0
    point estimate:          -0.037256131373281855
    95% confidence interval: (-0.7143, 0.6398)

Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           0.9141

Details:
    number of observations:   10
    z-statistic:              -0.10784820875555436
    population standard error: 0.34544970012182147
```

3.3.2 Two-sample test

- Now consider we have two sets of number $\{X_k\}$ and $\{Y_k\}$
- There are n_x and n_y of these, the sample means are $\overline{X}$ and $\overline{Y}$ and the population variance esimators σ_x and σ_y.
- Are the sample means the same or different? Consider statistics of $\overline{X} - \overline{Y}$
- The variance of the difference is $\sigma_x^2/n_x + \sigma_y^2/n_y$.
- The test statistic is therefore $z = \frac{(\overline{X}-\overline{Y})}{\sqrt{\sigma_x^2/n_x+\sigma_y^2/n_y}}$ which we test for significance.

In [8]:
```
nx=10; X=randn(nx) .+0.1;
ny=15; Y=1.2*randn(ny) .+0.2;
UnequalVarianceZTest(X,Y)
```

Out[8]:
```
Two sample z-test (unequal variance)
------------------------------------
Population details:
    parameter of interest:   Mean difference
    value under h_0:         0
    point estimate:          -0.052270158359419505
    95% confidence interval: (-0.7952, 0.6907)

Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           0.8903

Details:
    number of observations:   [10,15]
    z-statistic:              -0.13789330589472157
    population standard error: 0.3790623338838986
```

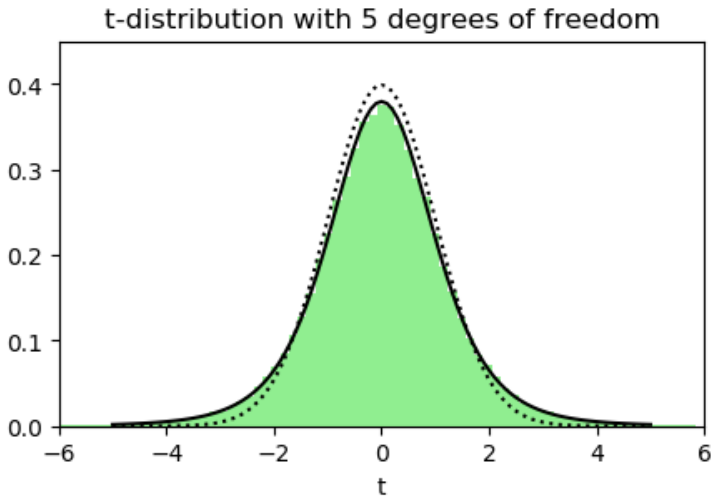
3.3.3 Student's t test

- Developed for small sample where typically $n \leq 30$ is quoted.
- The z-score comes from a standardised normal
  $z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$
- Previously assumed that esimate s_u^2 for population σ^2 is sharp
- But s_u^2 is a random number too increasing uncertainty
  distribution $t = (\bar{X} - \mu)/(s_u/\sqrt{n})$ is **not** normal
- Follows t-distribution (if samples normal).
  $Y = Y_0(1 + t^2/(n-1))^{-n/2}$.
  similar to normal, but tails fatter
- t-tests are preformed exactly like z-tests.

In [11]:

```
# example
N,n=100000,6
M=randn(N,n);
Sm=mean(M,dims=2)
Sv=var(M,dims=2)
t=Sm./sqrt.(Sv/n);

figure(figsize=(5,3))
x=-5:0.1:5
y1=pdf.(Normal(),x)
y2=pdf.(TDist(n-1),x)
plt[:hist](t,400,normed=5,color="lightgreen");
plot(x,y1,"k:",x,y2,"k-"); xlabel("t")
axis([-6,6,0,0.45]); title("t-distribution with $(n-1) degrees of freedom")
```



Out[11]:

```
PyObject Text(0.5, 1.0, 't-distribution with 5 degrees of freedom')
```

3.3.4 Chi-square test

- This test is used for seeing if a distribution of numbers is as expected.

- Let a_k where $k = 1 \cdots n$ be the measured frequency and b_k that expected.

- The test statistics is $\chi^2 = \sum_{k=1}^{n} \frac{(a_k - b_k)^2}{b_k}$.

- It has a sampling distribution is
  $Y(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{(\nu-2)/2}e^{-\chi^2/2}$
  where $\nu = n - 1$. This is a χ^2 distribution with $n - 1$ degrees of freedom.

- **Question:** A six-sided die is rolled 120 times with frequencies:

1	**2**	**3**	**4**	**5**	**6**
17	14	19	15	21	34

- Is the die a fair one? Test at 95% and 99% and give the p-value.

3.4 Summary and additional questions

Day 3 Frequentist approach

Questions

Make sure you have understood and done all the questions in the lectures.

The questions below are to be handed in for marking by 10am on Monday 26th November 2018
NB there will also be questions from the Day 4 lectures to be handed in on the 26th.

Q3.1 Spotting fake financial data
Q3.2 One-sided and two-sided test differences
Q3.3 Limit of Student's t-distribution.

Q3.1 Spotting fake data

- The first digits of many real-world data sets do not follow a uniform distribution.
- This observation is called Benford's law, where $P(d) = \log_{10}(1 + 1/d)$ for $d = 1 \cdots 9$
- When data is faked a uniform random number generator is often used.
- Download the following csv file from the Nasdaq archive: www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ&render=download
- The third column is the price of the stock. The data has some entries that are "n/a" which you need to clean up.

- **Part (a)** For a random sample of 90 stocks, plot the distribution of first digits of stock prices.
- **Part (b)** Use the Chi-square test to check if the distribution is significantly different from a uniform distribution and from Benford's law. Give the p-value for these two tests.
- **Part (c)** Increase the sample to 900. Comment on what happens to the significance test in respect to Benford's law.

Q3.2 One-sided and two-sided test differences.

- Refer back to the question in section 3.2.2 on the nationwide school test.

- A nationwide school test has mean mark $\mu = 75$ with standard dev. $\sigma = 7$.
- A particular school with 30 children has a mean mark $\overline{X} = 72$.

- Test the data for the question:
  Is the school underpreforming?
  What is the p-value and compare it to 5% or 1% significances.

Q3.3 Limit of Student's t-distribution

- **Part (a)** Demontrate that in the limit of large n Student's t-distribution tends to a standard normal. The distribution is

$$f(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+t^2/\nu\right)^{(\nu+1)/2}}$$

- Where $\nu = n - 1$ is the number of degrees of freedom.

- **Part(b)** Show that in this limit the leading order correction to the variance is $\sigma^2 \simeq 1 + 2/\nu$ and therefore larger than that of the standard normal.