VE401 Recitation 4

Descriptive Statistics

In[5]:=
```
Data = Floor[RandomVariate[NormalDistribution[75, 5], 100]]
```

Out[5]=
```
{77, 79, 68, 71, 79, 81, 80, 76, 74, 73, 69, 73, 70, 70, 71, 79, 81, 75, 82, 77,
 71, 77, 79, 72, 81, 69, 71, 80, 71, 77, 75, 73, 75, 73, 75, 72, 74, 83, 86, 82,
 67, 68, 72, 80, 70, 70, 76, 78, 71, 70, 74, 81, 72, 68, 72, 80, 77, 78, 87, 80,
 81, 73, 73, 72, 79, 74, 76, 78, 73, 72, 76, 79, 71, 73, 81, 71, 75, 75, 76, 76,
 59, 74, 71, 79, 66, 77, 87, 76, 71, 74, 77, 69, 80, 77, 64, 70, 79, 65, 70, 71}
```

Percentiles and Quartiles

Definitions: The x^{th} ***percentile*** is a value d_x of data such that $x\,\%$ of data are less than or equal to d_x. And,

- The ***first quartile*** q_1 is the 25^{th} percentile.
- The ***second quartile***, or median, q_2 is the 50^{th} percentile.
- The ***third quartile*** q_3 is the 75^{th} percentile.

Steps:

- The median is $q_2 = \begin{cases} \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \\ x_{(n+1)/2} & \text{if } n \text{ is odd} \end{cases}$, where x_i is the i^{th} smallest sample.

In[7]:=
```
Median[Data] // N
```

Out[7]=
```
74.5
```

- The first quartile is

$$= \begin{cases} \text{median of the smallest } n/2 \text{ elements} & \text{if } n \text{ is ev} \\ \frac{1}{2}(\text{median of the smallest } (n-1)/2 \text{ elements} + & \text{if } n \text{ is od} \\ \qquad \text{median of the smallest } (n+1)/2 \text{ elements}) & \end{cases}$$

- The third quartile is

$$= \begin{cases} \text{median of the largest } n/2 \text{ elements} & \text{if } n \text{ is ev} \\ \frac{1}{2}(\text{median of the largest } (n-1)/2 \text{ elements} + & \text{if } n \text{ is od} \\ \qquad \text{median of the largest } (n+1)/2 \text{ elements}) & \end{cases}$$

In[50]:=
```
{Q₁, Q₂, Q₃} = Quartiles[Data] // N
```

Out[50]=
```
{71., 74.5, 79.}
```

- The ***interquartile range*** is calculated using $\text{IQR} = q_3 - q_1$.

```
In[9]:= IQR = InterquartileRange[Data]
Out[9]= 8
```

Example:

Given the set of data {1, 2, 3, 4, 5}, calculate its quartiles q_1, q_2, q_3 and interquartile range.

We have the first quartile $q_1 = \frac{\frac{1+2}{2}+2}{2} = \frac{7}{4}$, second quartile $q_2 = 3$, third quartile $q_3 = \frac{\frac{4+5}{2}+4}{2} = \frac{17}{4}$. The interquartile range is then $\text{IQR} = \frac{5}{2}$.

Histograms

Steps: **Note that in the exam you have to draw all diagrams using pencil and paper!**

- Choose a convenient number of bins k and a good bin width h. The bin width is calculated as

$$h = \begin{cases} \frac{\max\{x_i\} - \min\{x_i\}}{\lceil \log_2 n \rceil + 1} & \text{if using Sturges' rule} \\ \frac{2 \cdot \text{IQR}}{\sqrt[3]{n}} & \text{if using Freedman} \cdot \text{Diaconis Rule} \end{cases}$$

We usually round h up so that it is “nicer” for our data.

```
In[10]:= kSturges = Ceiling[Log2[Length[Data]]] + 1
Out[10]= 8
In[11]:= hSturges = Ceiling[(Max[Data] - Min[Data]) / kSturges]
Out[11]= 4
In[12]:= hFreedmanDiaconis = Ceiling[2 × InterquartileRange[Data] / CubeRoot[Length[Data]]]
Out[12]= 4
```

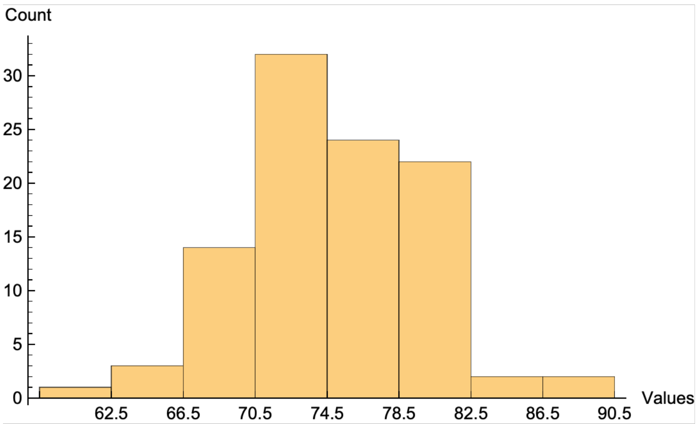
- Take the smallest datum, subtract one-half of the smallest decimal of the data, and successively add the bin width to obtain the bins.

```
In[69]:= bins = Table[Min[Data] - 0.5 + i × hSturges, {i, 0, kSturges}]
Out[69]= {58.5, 62.5, 66.5, 70.5, 74.5, 78.5, 82.5, 86.5, 90.5}
```

- Count the number of data in each bin and plot the histogram

In[70]:=
```
Histogram[Data, {Min[bins], Max[bins], hSturges},
 Ticks → {bins, Automatic}, AxesLabel → {Values, Count}]
```

Out[70]=



Stem-and-Leaf Diagrams

Purpose: to get a rough idea of the shape of distribution.

Steps:

- Choose a convenient number of leading decimal digits as stem.
- Label each row with stem.
- For each datum, note down the digit following the stem.

In[15]:=
```
Floor[Sort[Data]]
```

Out[15]=
```
{59, 64, 65, 66, 67, 68, 68, 68, 69, 69, 69, 70, 70, 70, 70, 70, 70, 70, 71, 71,
 71, 71, 71, 71, 71, 71, 71, 71, 71, 72, 72, 72, 72, 72, 72, 72, 73, 73, 73, 73,
 73, 73, 73, 73, 74, 74, 74, 74, 74, 74, 75, 75, 75, 75, 75, 75, 76, 76, 76, 76,
 76, 76, 76, 77, 77, 77, 77, 77, 77, 77, 77, 78, 78, 78, 79, 79, 79, 79, 79, 79,
 79, 79, 80, 80, 80, 80, 80, 80, 81, 81, 81, 81, 81, 81, 82, 82, 83, 86, 87, 87}
```

In[63]:=
```
Needs["StatisticalPlots`"]
StemLeafPlot[Floor[Data, 1], IncludeStemCounts → True,
 IncludeEmptyStems → True, StemExponent → {1, "UnitDivisions" → 2}]
```

Out[64]=

Stem	Leaves	Counts
5	9	1
6	4	1
6	567888999	9
7	000000011111111111222222233333333444444	39
7	55555566666667777777788899999999	32
8	000000111111223	15
8	677	3

Stem units: 10

Box plots

Steps:

- Calculate ***inner fences*** $f_1 = q_1 - \frac{3}{2}\,\text{IQR}$, and $f_3 = q_3 + \frac{3}{2}\,\text{IQR}$.

In[18]:=
```
f1 = Q1 - 1.5 IQR
f3 = Q3 + 1.5 IQR
```
Out[18]= 59.

Out[19]= 91.

- Calculate ***outer fences*** $F_1 = q_1 - 3\,\text{IQR}$ and $F_3 = q_3 + 3\,\text{IQR}$.

In[20]:=
```
F1 = Q1 - 3 IQR
F3 = Q3 + 3 IQR
```
Out[20]= 47

Out[21]= 103

- Calculate ***adjacent values*** $a_1 = \min\{x_k : x_k \geq f_1\}$, and $a_3 = \max\{x_k : x_k \leq f_3\}$

In[22]:=
```
a1 = Min[Select[Data, # ≥ f1 &]]
a3 = Max[Select[Data, # ≤ f3 &]]
```
Out[22]= 59

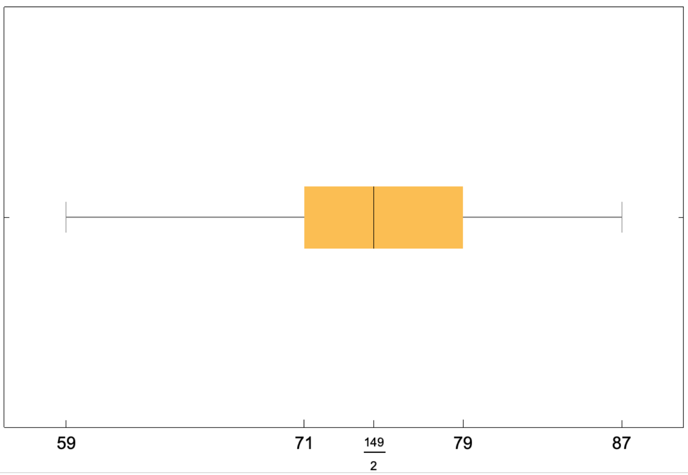
Out[23]= 87

- Data points in $(F_1, f_1) \bigcup (f_3, F_3)$ are ***near outliers*** and data points in $(-\infty, F_1) \bigcup (F_3, \infty)$ are ***far outliers***.

In[24]:=
```
nearoutliers = Select[Data, (F1 < # < f1) || (f3 < # < F3) &]
```
Out[24]= { }

In[25]:=
```
faroutliers = Select[Data, (F1 > #) || (# > F3) &]
```
Out[25]= { }

In[26]:=
```
outliers := Join[nearoutliers, faroutliers]
```
In[27]:=
```
BoxWhiskerChart[Data,
 {"Outliers", {"MedianMarker", 1, Directive[Black]}}, BarOrigin → Left,
 FrameTicks → {{{}, {}}, {Join[Quartiles[Data], {a1, a3}, outliers], {}}}]
```
Out[27]=



Estimation

Sample Statistics

random variables $X_1, X_2, \ldots, X_n$ are ***random samples of size n*** from the distribution X. They are ***independent identically distributed (i.i.d.)*** random variables.

	Sample	Actual
Range	$\max_{1\le k\le n} X_k - \min_{1\le k\le n} X_k$	$\begin{cases} \Omega & \text{if discrete} \\ \mathbb{R} & \text{if continuous} \end{cases}$
Mean	$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$	$\mathrm{E}[X]$
Median	$\tilde{x} = \begin{cases} \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & n \text{ even} \\ x_{(n+1)/2} & n \text{ odd} \end{cases}$ where x_i is the i^{th} smallest sample	m such that $P[X \le m] = \frac{1}{2}$
Variance	$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$	$\sigma^2 = \mathrm{E}\left[\left(X - \mathrm{E}[X]\right)^2\right]$
Standard Deviation	$S = \sqrt{S^2}$	$\sigma = \sqrt{\sigma^2}$

Bias and Mean Square Error

Purpose: In order to have a "good" estimator $\hat{\theta}$ of the parameter θ, we want it to be

- ***unbiased***, or $\mathrm{E}[\hat{\theta}] = \theta$, and
- $\mathrm{Var}[\hat{\theta}] \to 0$ when $n \to \infty$.

Definitions: We define ***bias*** $:= \theta - \mathrm{E}[\hat{\theta}]$, and ***mean square error*** $\mathrm{MSE}(\hat{\theta}) := \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathrm{Var}[\hat{\theta}] + (\text{bias})^2$.

Example:

Prove that the estimator of mean $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \to \mu$ as $n \to \infty$.

We have $\mathrm{MSE}(\overline{X}) = \mathrm{E}\left[\left(\overline{X} - \mu\right)^2\right] = \mathrm{Var}[\overline{X}] + \mu - \mathrm{E}[\overline{X}] = \frac{\sigma^2}{n} + 0 \xrightarrow{n\to\infty} 0$. Therefore $\overline{X} \xrightarrow{n\to\infty} \mu$. This shows that the estimator is consistent.

Method of Moments

Intuition: Given that the estimator for the k^{th} $(k \ge 1)$ moment

$$E\left[\hat{X^k}\right] = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

is unbiased, we can use this to give good estimators for parameters that involve moments.

Steps:

- Express the parameter in terms of moments. Hint: calculate $\mathrm{E}[X]$, $\mathrm{E}\left[X^2\right]$,...
- Replace the moments with the corresponding estimators.

Example:

Provide a method of moment estimator of parameter p in the binomial distribution.

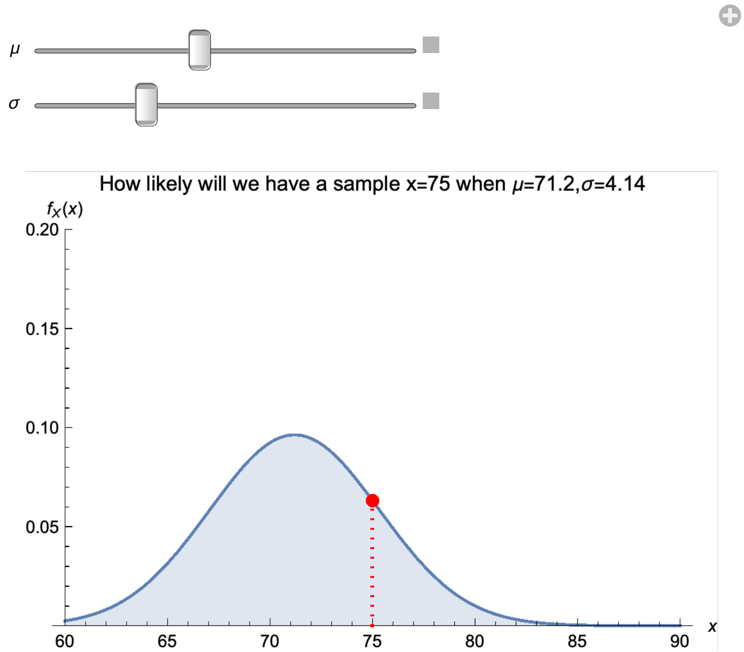
We have $\mathrm{E}[X] = n\,p$ and $\mathrm{Var}[X] = n\,p\,(1-p)$, therefore

$$\frac{\mathrm{Var}[X]}{\mathrm{E}[X]} = 1-p \Rightarrow p = 1-\frac{\mathrm{E}[X^2]-\mathrm{E}[X]^2}{\mathrm{E}[X]} \Rightarrow \hat{p} = 1-\frac{\frac{1}{k}\sum_{i=1}^{k} X_i^2-\overline{X}^2}{\overline{X}} = \frac{\overline{X}-\frac{1}{k}\sum_{i=1}^{k}\left(X_i-\overline{X}\right)^2}{\overline{X}}.$$

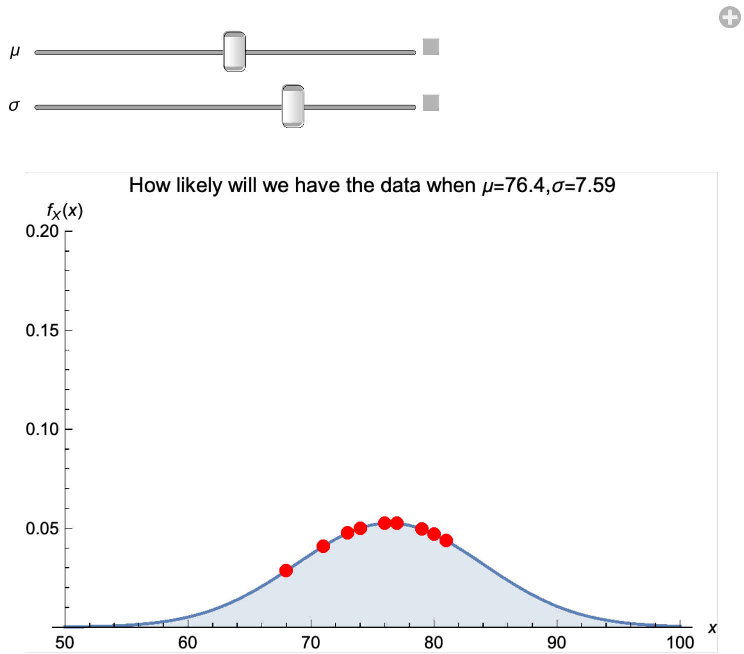
Method of Maximum Likelihood

Intuition: we want the parameter to be something, such that it will be the most likely to obtain our data set.

Out[ ]=



Out[29]=



Mathematical representation:

$$\begin{aligned}\text{likelihood of obtaining data set} &= \prod_{i=1}^{n} \text{likelihood of obtaining the data } x_i \\ &= \prod_{i=1}^{n} f_{X_\theta}(x_i)\end{aligned}$$

Steps:

- Calculate $L(\theta) = \prod_{i=1}^{n} f_{X_\theta}(x_i)$, and sometimes $\ln L(\theta) = \sum_{i=1}^{n} \ln f_{X_\theta}(x_i)$.
- Find maximum of this likelihood by solving $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$.

Example:

Calculate the MLE for parameter μ, σ in normal distribution.

We have

$$L(\mu, \sigma) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) = \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) + -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = n \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu$$

Now, calculating partial derivatives,

$$\frac{\partial \ln L(\mu,\sigma)}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{X}$$

and

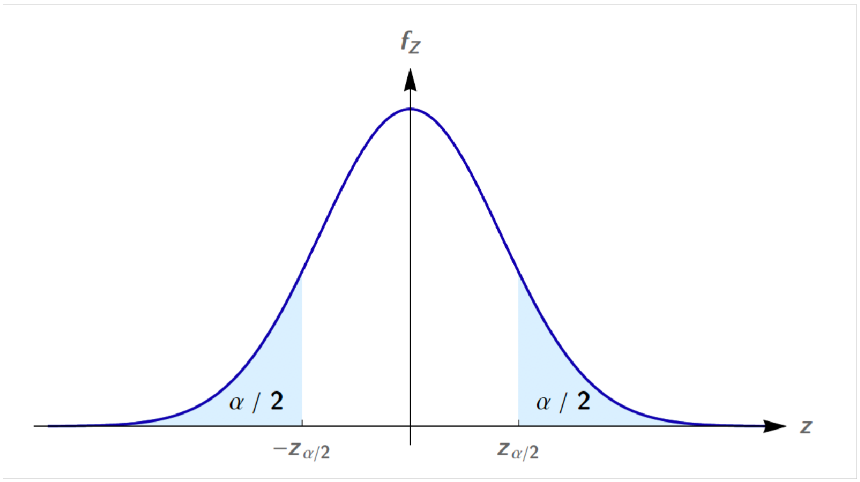
$$\frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \Rightarrow n\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Interval Estimation

Confidence Interval with confidence level $100\,(1-\alpha)\,\%$

Intuition: We need an interval such that in $100\,(1-\alpha)\,\%$ of the cases, the true value lies in the interval.

Mathematical representation: We calculate L_1, L_2 such that $P[\theta < L_1] = P[\theta > L_2] = \alpha/2$, in this case $P[L_1 \le \theta \le L_2] = 1-\alpha$.



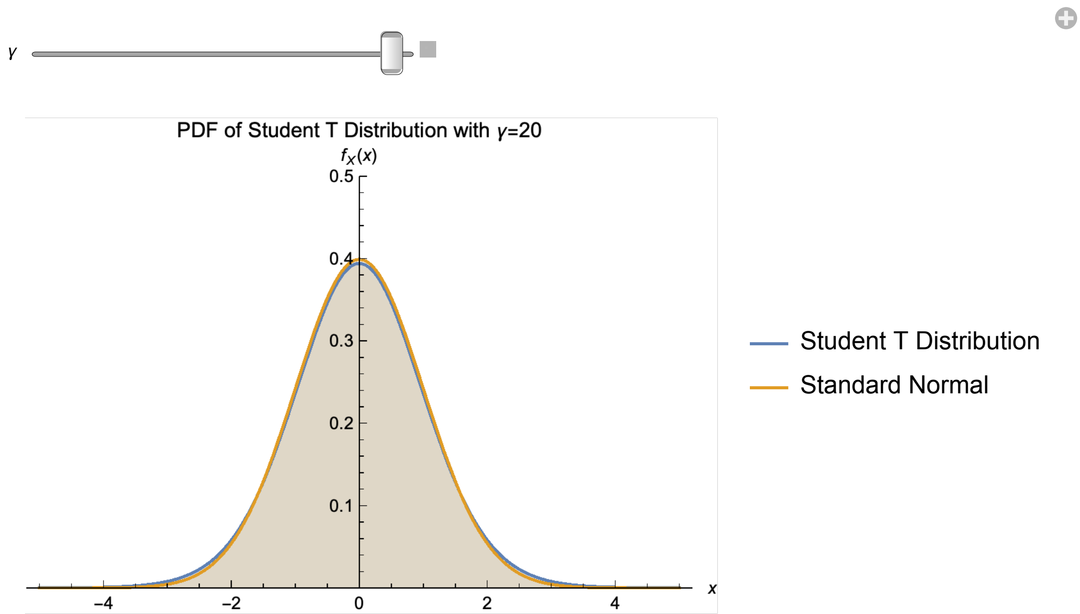
Student T Distribution

Purpose: to describe the distribution $T_\gamma = \dfrac{Z}{\sqrt{\chi_\gamma^2/\gamma}}$.

Parameter: $\gamma \in \{1, 2, \ldots\}$ is the degree of freedom.

PDF: $f_{T_\gamma}(t) = \dfrac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)\sqrt{\pi\gamma}}\left(1+\dfrac{t^2}{\gamma}\right)^{-\frac{\gamma+1}{2}}$.

Out[∘]=



Confidence Interval for the Mean

Distribution of X_i	Sample size n	Variance σ^2	Statistic	$1-\alpha$ two–sided confidence interval
$X_i \approx N(\mu, \sigma)$	any	known	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \approx N(0, 1)$	$\left[\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$
$X_i \approx$ any distribution	large	known	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \approx N(0, 1)$	$\left[\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$
$X_i \approx$ any distribution	large	unknown	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \approx N(0, 1)$	$\left[\overline{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}\right]$
$X_i \approx N(\mu, \sigma)$	small	unknown	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \approx t_{n-1}$	$\left[\overline{X} - t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}\right]$
$X_i \approx$ any distribution	small	known or unknown	Go home!	Go home!

(*Source : https : // stanford.edu / ~shervine / teaching / cme – 106 / cheatsheet – statistics*)

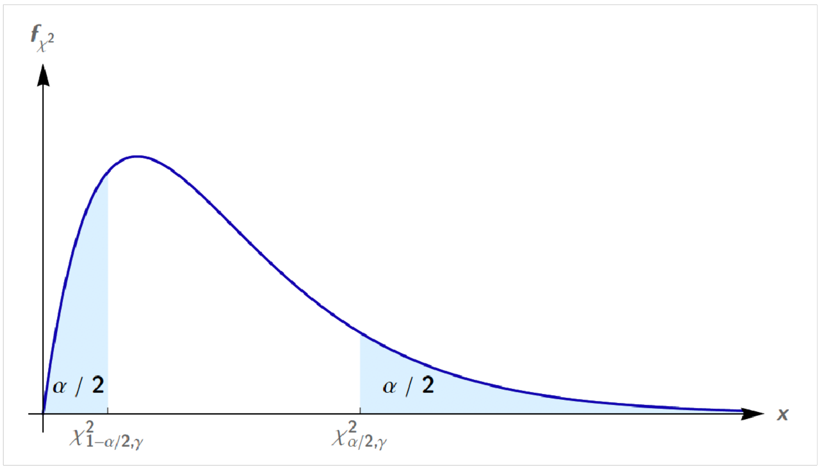
```
In[65]:= z_α_ := InverseCDF[NormalDistribution[], α]
In[66]:= z_0.95
Out[66]= 1.64485
```

Confidence Interval for the Variance

Let $X_1, \ldots, X_n$, $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2, then

- The sample mean $\overline{X}$ is independent of the sample variance S^2,
- $\overline{X}$ is normally distributed with mean μ and variance σ^2 / n,
- $(n-1) S^2 / \sigma^2$ is chi-squared distributed with $n-1$ degrees of freedom.

Distribution of X_i	Sample size n	Variance σ^2	Statistic	$1-\alpha$ two–sided confidence interval
$X_i \approx N(\mu, \sigma)$	any	known or unknown	$\frac{(n-1) S^2}{\sigma^2} \approx \chi^2_{n-1}$	$\left[\frac{(n-1) S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1) S^2}{\chi^2_{\alpha/2,n-1}}\right]$



Problems in the Assignment

A Tricky Question Involving the Binomial Distribution

A mathematics textbook has 200 pages on which typographical errors in the equations could occur. Suppose there are in fact five errors randomly dispersed among these 200 pages.

- What is the probability that a random sample of 50 pages will contain at least one error?
- How large must the random sample be to assure that at least three errors will be found with 90% probability?

Solution

- We define "success" as "an error is in the sample". $p = P[\text{the error is in the sample}] = \frac{50}{200} = \frac{1}{4}$. The total number of trial $n = 5$. So
  $P[\text{at least one error}] = 1 - P[X = 0] = 1 - \binom{5}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^5 = 0.763.$
- Using normal approximation,
  $P[X \geq 3] = 1 - P[X \leq 2] = 1 - \Phi\left(\frac{2+1/2-5p}{\sqrt{5p(1-p)}}\right) \geq 90\% \Rightarrow \Phi\left(\frac{2+1/2-5p}{\sqrt{5p(1-p)}}\right) \leq 10\%,$

In[∘]:=
```
InverseCDF[NormalDistribution[], 0.1]
```
Out[∘]=
```
-1.28155
```

$\frac{2+1/2-5p}{\sqrt{5p(1-p)}} \leq -1.28 \Rightarrow p \approx 0.748 \Rightarrow$ the number of samples will be $200p \approx 150$.

Linear Combination of Two Normal Distribution

Let X_1 and X_2 be independent normal distributions with means μ_1 and μ_2, and variance σ_1 and σ_2, respectively. Let $\lambda_1, \lambda_2 \in \mathbb{R}$, Show that the linear combination $Y = \lambda_1 X_1 + \lambda_2 X_2$ follows a normal distribution.

Solution

The moment generating function of $\lambda_1 X_1$ is

$$\begin{aligned} m_{\lambda_1 X_1}(t) &= \mathrm{E}\left[e^{t \lambda_1 X_1}\right] \\ &= m_{X_1}(\lambda_1 t) \\ &= \exp\left(\mu_1 \lambda_1 t + \sigma_1^2 \lambda_1^2 t^2 / 2\right) \end{aligned}$$

And similar for $\lambda_2 X_2$. Since X_1 and X_2 are independent, $\lambda_1 X_1, \lambda_2 X_2$ are also independent and therefore

$$\begin{aligned} m_Y(t) &= m_{\lambda_1 X_1}(t)\, m_{\lambda_2 X_2}(t) \\ &= \exp\left(\mu_1 \lambda_1 t + \frac{\sigma_1^2 \lambda_1^2 t^2}{2} + \mu_2 \lambda_2 t + \frac{\sigma_2^2 \lambda_2^2 t^2}{2}\right) \\ &= \exp\left[(\mu_1 \lambda_1 + \mu_2 \lambda_2)\, t + \frac{(\sigma_1^2 \lambda_1^2 + \sigma_2^2 \lambda_2^2)\, t^2}{2}\right] \end{aligned}$$

We notice that this is actually the MGF for normal distribution for $\mu = \mu_1 \lambda_1 + \mu_2 \lambda_2$ and $\sigma^2 = \sigma_1^2 \lambda_1^2 + \sigma_2^2 \lambda_2^2$. By the uniqueness of MGF we conclude that Y actually follows the same distribution.