NORMAL TABLES CAN ALSO BE USED IN REVERSE, IF THE QUESTION IS TO FIND x SUCH THAT $P(X \leq x) = p$ WHERE p IS GIVEN AND $X \sim N(\mu, \sigma^2)$.

R CODE:

```
qnorm(p, mu, sigma)
```

OR

```
mu + sigma*qnorm(p)
```

THE PROCEDURE IS:

1) DETERMINE THE LEFT-HAND AREA.
2) LOCATE AREA IN THE BODY OF THE TABLE, READ THE CORRESPONDING z.
3) CONVERT $x = \mu + z\sigma$.

EX: FIND THE UPPER QUARTILE OF NORMAL DISTRIBUTION WITH $\mu = 42\text{kg}$, $\sigma = 4.4\text{kg}$.

A: IN THE TABLE, 0.75 IS BETWEEN 0.7486 ($z = 0.67$) AND 0.7517 ($z = 0.68$)

INTERPOLATING:

$$z = 0.67 + \frac{0.75 - 0.7486}{0.7517 - 0.7486}(0.68 - 0.67) = 0.6745$$

$$x = \mu + \sigma z = 42 + 4.4(0.6745) = \boxed{44.97}$$

```
qnorm(0.75, 42, 4.4)
```

OR

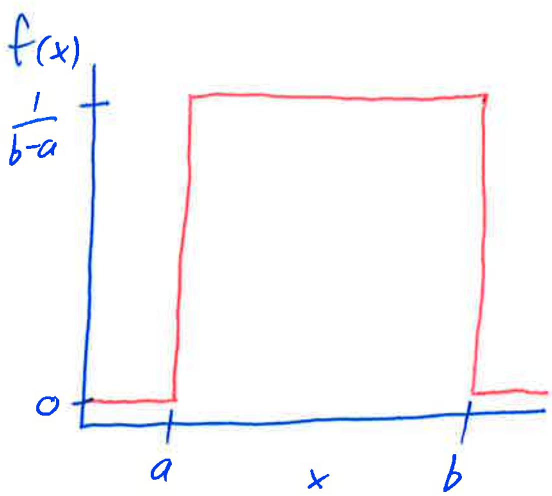
```
42 + 4.4*qnorm(0.75)
```

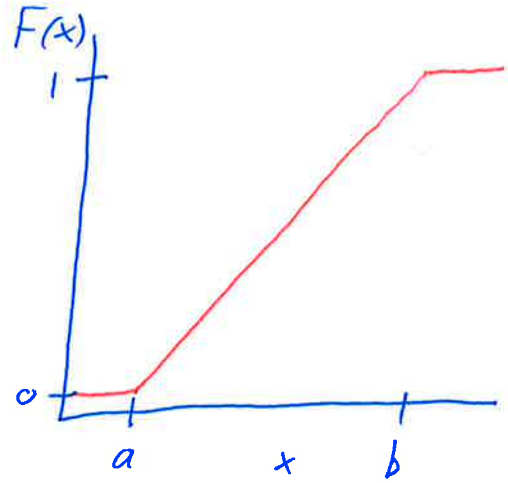
UNIFORM DISTRIBUTION

A CONTINUOUS RV X HAS UNIFORM DISTRIBUTION ON (a,b) IF ITS PDF IS CONSTANT ON (a,b).

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

$$F(x) = \begin{cases} 0 & , x \le a \\ \int_a^x \frac{1}{b-a}\,dt & , a < x < b \\ 1 & , x \ge b \end{cases}$$





UNIFORM MEAN AND VARIANCE:

$$E(X) = \int_a^b \frac{x}{b-a}\,dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}.$$

$$E(X^2) = \int_a^b \frac{x^2}{b-a}\,dx = \frac{b^2+ba+a^2}{3}$$

$$\Rightarrow Var(X) = \frac{b^2+ba+a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}.$$

FITTING MODELS TO DATA

WE HAVE SEEN SEVERAL MODELS FOR DISCRETE AND CONTINUOUS RVS.
HOW DO WE KNOW WHICH MODEL TO USE FOR A PARTICULAR DATASET?
HOW DOES LIMITED SAMPLE SIZE AFFECT THE CHOICE?

CONSIDER n OBSERVATIONS $X_1, \ldots, X_n$ THAT FOLLOW THE SAME DISTRIBUTION, WITH $E(X_i) = \mu$. BY LINEARITY OF THE EXPECTED VALUE,

$$E(X_1 + X_2 + \ldots + X_n) = \mu + \mu + \ldots + \mu = n\mu.$$

USING THE <u>SAMPLE MEAN</u> $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$, WE FIND

$$E(\bar{X}) = E\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n} n\mu = \mu.$$

THERE ARE 3 DIFFERENT CONCEPTS OF "MEAN" HERE:

- DISTRIBUTION OF OBSERVATIONS WITHIN A DATASET, CENTRED ABOUT THE SAMPLE MEAN $\bar{x}$.
- DISTRIBUTION OF A RANDOM VARIABLE X, CENTRED ABOUT THE POPULATION MEAN μ.
- SAMPLING DISTRIBUTION OF $\bar{X}$, DESCRIBING VARIATION OF SAMPLE MEANS OVER ALL POSSIBLE SAMPLES.

IF $X_1, \ldots, X_n$ ARE RANDOM, THEN $\bar{X}$ IS ALSO A RV!

SAMPLING DISTRIBUTION

- THE SAMPLING DISTRIBUTION OF $\bar{X}$ DEPENDS ON THE UNDERLYING PROBABILITY MODEL OF A SINGLE OBSERVATION X.

- THE RESULT $E(\bar{X}) = \mu$ SAYS THAT $\bar{X}$ HAS THE SAME EXPECTED VALUE AS A SINGLE OBSERVATION.

• HOWEVER, WE EXPECT THAT AVERAGING REPEATED MEASUREMENTS SHOULD INCREASE ACCURACY. SO THE SAMPLING DISTRIBUTION OF $\overline{X}$ SHOULD VARY ACCORDING TO SAMPLE SIZE n, WITH REDUCED SPREAD AS n INCREASES.

COVARIANCE AND INDEPENDENCE

RECALL THAT $Var(a+bX) = b^2 Var(X)$. FOR INDEPENDENT RVs X AND Y, $Var(a+bX+cY) = b^2 Var(X) + c^2 Var(Y)$. BUT IF X AND Y ARE NOT INDEPENDENT, THE VARIANCE IS MORE COMPLEX:

$$\sigma^2_{aX+bY} = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,Cov(X,Y),$$

WHERE $Cov(X,Y) = E(XY) - E(X)E(Y)$ IS THE COVARIANCE BETWEEN X AND Y.

FOR INDEPENDENT VARIABLES, $E(XY) = E(X)E(Y)$, SO THE COVARIANCE IS ZERO. COVARIANCE IS A WAY OF MEASURING DEPENDENCE.

EX: LET X, Y BE RVs WITH $\sigma_X = 3$, $\sigma_Y = 4$, $Cov(X,Y) = 1$.

FIND σ^2_{2X-Y}.

A: $\sigma^2_{2X-Y} = 2^2 3^2 + (-1)^2 4^2 + 2\cdot 2(-1)\cdot 1 = \sqrt{48} = 4\sqrt{3}$.

EX: LET $X_1, \dots, X_{12}$ BE INDEPENDENT UNIFORM RVs. FIND THE MEAN AND VARIANCE OF $\sum_{i=1}^{12} X_i$, IF $\mu_i = \frac{1}{2}$ AND $\sigma_i^2 = \frac{1}{12}$ FOR EACH i.

A: $E(X_1 + \dots + X_{12}) = 12 \cdot \frac{1}{2} = 6$; $Var(X_1 + \dots + X_{12}) = 12 \cdot \frac{1}{12} = 1$.

NOTE: $X_1 + \dots + X_{12} - 6$ HAS MEAN 0, VARIANCE 1, SO ITS DISTRIBUTION IS APPROXIMATELY STANDARD NORMAL.

WHY NORMAL? SEE THE CENTRAL LIMIT THEOREM LATER.

SAMPLE MEAN AS A RANDOM VARIABLE

CONSIDER n INDEPENDENT OBSERVATIONS $X_1, \dots, X_n$ OF A RV WITH MEAN μ AND VARIANCE σ^2. THE SAMPLE MEAN IS

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}.$$

$\bar{X}$ HAS ITS OWN EXPECTED VALUE $E(\bar{X})$, VARIANCE $Var(\bar{X})$ AND STANDARD DEVIATION $\sigma_{\bar{X}}$ (KNOWN AS <u>STANDARD ERROR</u>).

$$E(\bar{X}) = \frac{1}{n} E(X_1 + \dots + X_n) = \frac{n\mu}{n} = \mu.$$

$$Var(\bar{X}) = \frac{1}{n^2} Var(X_1 + \dots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

STANDARD ERROR

- THE STANDARD ERROR OF THE SAMPLE MEAN IS $\sigma/\sqrt{n}$.
- AS THE SAMPLE SIZE INCREASES, THE STANDARD ERROR DECREASES.
- WITH LARGER SAMPLES, THE SAMPLE MEAN IS MORE LIKELY TO BE CLOSE TO μ.

IF THE SAMPLE COMES FROM A NORMAL POPULATION, THEN $\bar{X} \sim N(\mu, \sigma^2/n)$. EQUIVALENTLY, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

INTERESTINGLY, THIS REMAINS APPROXIMATELY TRUE FOR SAMPLES FROM <u>ANY</u> DISTRIBUTION, PROVIDED $\sigma < \infty$ AND n IS "LARGE."

EX: WEIGHTS OF TILES ARE NORMALLY DISTRIBUTED WITH $\mu = 1\text{kg}$ AND $\sigma = 20\text{g}$. FIND THE PROBABILITY THAT A PACK OF 12 TILES HAS AVERAGE WEIGHT BELOW 995g.

A: $\bar{X} \sim N(1000, 20^2/12)$.

$$P(\bar{X} < 995) = P\left(Z < \frac{995-1000}{20/\sqrt{12}}\right)$$

$$= P(Z < -0.866) \approx 0.1933.$$

CENTRAL LIMIT THEOREM: FOR RANDOM SAMPLING WITH A LARGE SAMPLE SIZE n, THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN IS APPROXIMATELY NORMAL WITH MEAN μ AND STANDARD ERROR $\sigma/\sqrt{n}$. THIS IS TRUE NO MATTER THE TYPE OF PROBABILITY DISTRIBUTION THAT PROVIDES THE SAMPLES.

- THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN TAKES MORE OF A BELL SHAPE AS n INCREASES.
- THE MORE SKEWED THE POPULATION DISTRIBUTION, THE LARGER n MUST BE BEFORE THE SHAPE IS CLOSE TO NORMAL.
- IN PRACTICE, $n \geq 30$ IS USUALLY CLOSE TO NORMAL.

SAMPLING EXPERIMENT

1000 RANDOM OBSERVATIONS WERE SIMULATED FROM THE PDF $f(x) = 2x$, $0 < x < 1$. THIS WAS REPEATED 16 TIMES, TO FORM A TABLE OF 16 COLUMNS AND 1000 ROWS. FOR EACH ROW, AVERAGES WERE CALCULATED FOR THE FIRST 2, FIRST 4, AND ALL 16 OBSERVATIONS.

i) THE EXPECTED VALUE OF A SINGLE OBSERVATION IS

$$\mu = E(X) = \int_0^1 x \cdot 2x\,dx = \frac{2x^3}{3}\Big|_0^1 = \frac{2}{3}.$$

THE VARIANCE OF A SINGLE OBSERVATION IS

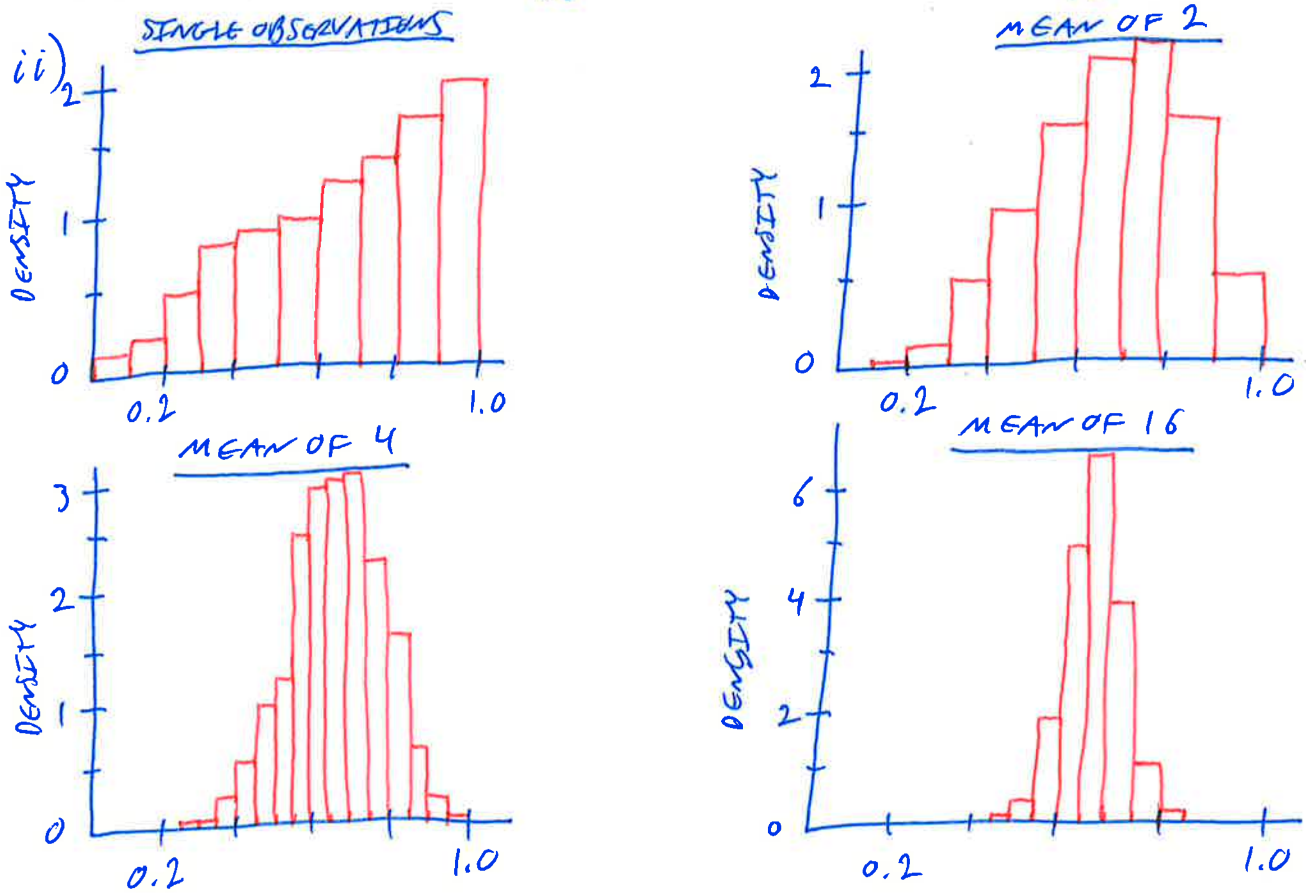
$$E(X^2) - \mu^2 = \int_0^1 x^2 \cdot 2x\,dx - \frac{4}{9} = \frac{x^4}{2}\Big|_0^1 - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

SO THE VARIANCE FOR AN AVERAGE OF n OBSERVATIONS IS

$$\frac{\sigma^2}{n} = \frac{1}{18n}.$$

MEAN OF 2: VARIANCE $\frac{1}{36}$. MEAN OF 4: VARIANCE $\frac{1}{72}$. MEAN OF 16: VARIANCE $\frac{1}{288}$.

ii)



NOTE THAT AS n INCREASES,

- THE SHAPE BECOMES MORE SYMMETRIC AND BELL-LIKE.
- THE CENTRE REMAINS ABOUT THE SAME.
- THE SPREAD BECOMES SMALLER.

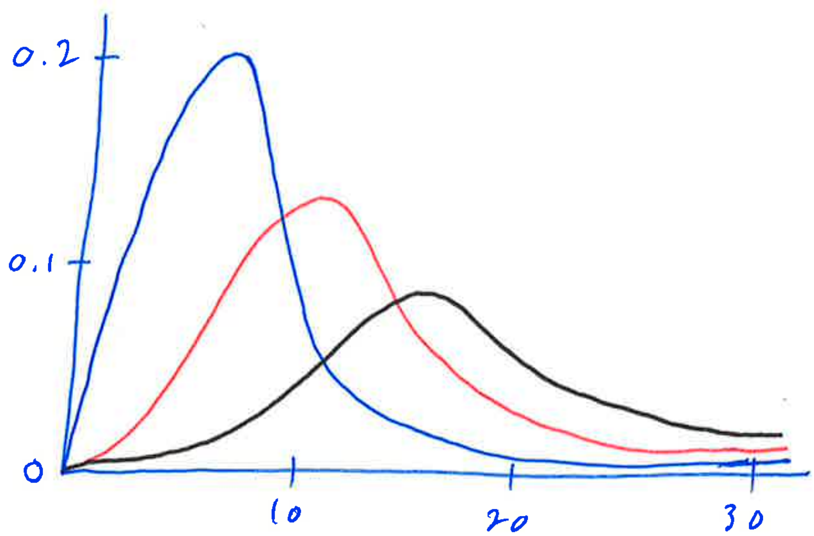
SAMPLING DISTRIBUTION OF VARIANCE

CONSIDER A SAMPLE $X_1, \ldots, X_n$ OF INDEPENDENT $N(\mu, \sigma^2)$ OBSERVATIONS. THE SAMPLE VARIANCE $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ VARIES BETWEEN SAMPLES. THE SAMPLING DISTRIBUTION OF $\frac{(n-1)S^2}{\sigma^2}$ IS CALLED A CHI-SQUARED (χ^2) DISTRIBUTION WITH $n-1$ DEGREES OF FREEDOM.

χ^2-DISTRIBUTION

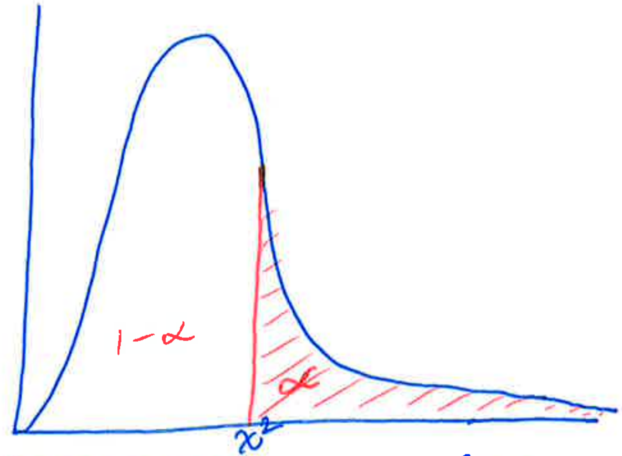
χ^2 IS A CONTINUOUS MODEL WITH MANY APPLICATIONS. THE MINIMUM POSSIBLE VALUE IS 0; THERE IS NO MAXIMUM. THE SHAPE, MEAN ν, AND VARIANCE 2ν DEPEND ON A PARAMETER KNOWN AS THE DEGREES OF FREEDOM, df.

χ^2 PDF, df = 4, 8, 12



TABLES USUALLY LIST χ^2 VALUES FOR RIGHT-TAIL PROBABILITIES α. SOME TABLES INCLUDE LEFT-TAIL AREAS $1-\alpha$.

df	α $1-\alpha$	0.10 0.90	0.05 0.95	0.025 0.975
1		2.706	3.841	5.024
3		6.251	7.815	9.348
5		9.236	11.070	12.833



AREA TO THE RIGHT OF χ^2 IS α.

EX: FOR A SAMPLE OF SIZE 6 FROM A NORMAL POPULATION WITH $\mu = 70, \sigma^2 = 45$, LOOK UP χ^2 TABLES WITH $6-1=5$ df TO FIND

$$P\left[\frac{(6-1)S^2}{45} > 11.070\right] = 0.05$$

$$P(S^2 > 99.63) = 0.05$$

$$P(S > 9.981) = 0.05$$

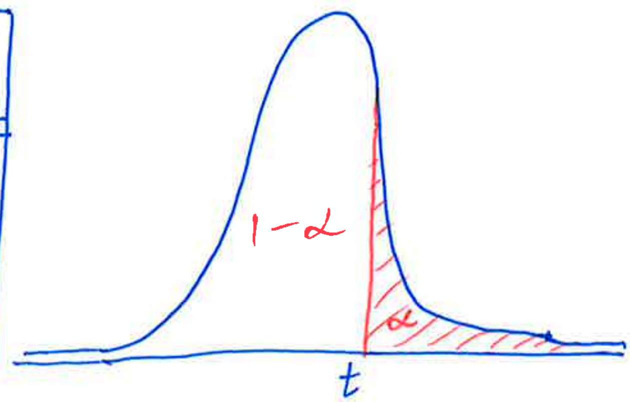
STUDENT'S t-DISTRIBUTION

WILLIAM GOSSETT HAD THE IDEA OF CONSIDERING $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ INSTEAD OF $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ FOR A RANDOM SAMPLE FROM A $N(\mu, \sigma^2)$ DISTRIBUTION. THE SAMPLING DISTRIBUTION OF T IS CALLED THE STUDENT'S t-DISTRIBUTION WITH $n-1$ df, WRITTEN $T \sim t_{n-1}$.

- THE t-DISTRIBUTION IS BELL-SHAPED AND SYMMETRIC ABOUT $\emptyset$.
- THE t-DISTRIBUTION HAS THICKER TAILS AND IS MORE SPREAD OUT THAN THE STANDARD NORMAL DISTRIBUTION.
- THE PROBABILITIES DEPEND ON THE DEGREES OF FREEDOM.
- FOR A t-SCORE BASED ON A SINGLE SAMPLE OF SIZE n, $df = n-1$.

TABLES LIST VALUES OF $t_{df;\alpha}$ (RIGHT-TAIL), AND SOME TABLES HAVE $1-\alpha$ LEFT-HAND AREAS.

df \ α / $1-\alpha$	α = 0.10 / $1-\alpha$ = 0.90	0.05 / 0.95	0.025 / 0.975
1	3.078	6.314	12.606
3	1.638	2.815	3.182
5	1.476	2.025	2.571
∞	1.282	1.645	1.960



Ex: FOR A SAMPLE OF SIZE 6 FROM A NORMAL POPULATION WITH $\mu = 70$, LOOK UP t TABLES WITH $6-1=5$ df TO FIND

$$P\left[T = \frac{\bar{X}-70}{S/\sqrt{6}} < 1.476\right] = 0.90.$$

BY SYMMETRY,

$$P\left[T < -1.476\right] = 0.10.$$

AS $df \to \infty$, THE t-DISTRIBUTION APPROACHES STANDARD NORMAL.

ESTIMATION

WE GENERALLY TAKE A RANDOM SAMPLE FROM A POPULATION TO GET SOME INFORMATION ABOUT IT. WE ESTIMATE THE MEAN μ BY THE SAMPLE MEAN $\bar{X}$. BUT $\bar{X}$ IS A SINGLE NUMBER (A "POINT ESTIMATE") AND IS ALMOST CERTAINLY NOT EXACT.

OFTEN, WE PREFER AN INTERVAL ESTIMATE, LIKE $\mu \in [3.4, 5.6]$.

CONFIDENCE INTERVALS

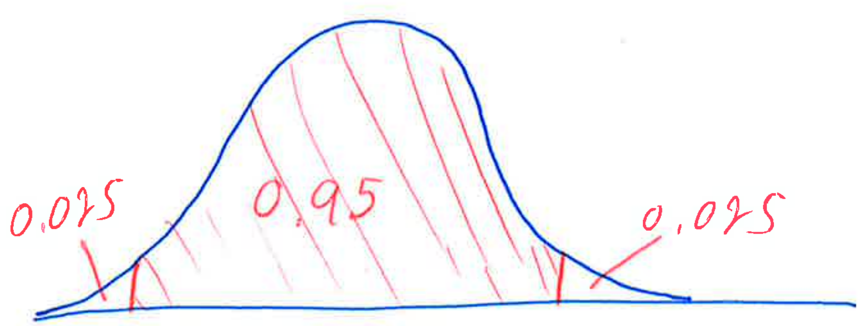
- A CONFIDENCE INTERVAL CONTAINS THE MOST LIKELY VALUES FOR A PARAMETER.
- THE PROBABILITY THAT THE PARAMETER IS CONTAINED IN THE INTERVAL IS THE CONFIDENCE LEVEL, MOST OFTEN 0.95.
- MANY CONFIDENCE INTERVALS ARE OF THE FORM

  POINT ESTIMATE $\pm$ MARGIN OF ERROR.

THE SIMPLEST CASE IS WHEN σ IS KNOWN AND μ IS UNKNOWN.

FROM STANDARD NORMAL TABLES, WE FIND

$$P(-1.96 < Z < 1.96) = 0.95.$$



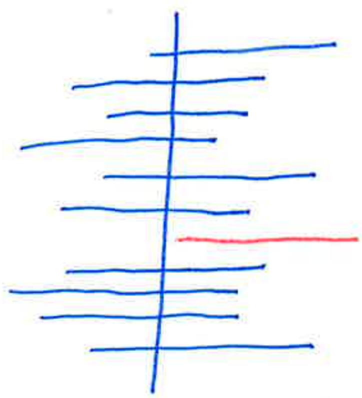
SO FOR $\bar{X}$ FROM $N(\mu, \sigma^2)$, WE HAVE

$$0.95 = P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

IN OTHER WORDS, THE INTERVAL $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ CONTAINS μ WITH PROBABILITY 0.95. THIS IS THE 95% CONFIDENCE INTERVAL FOR μ.

SO IN THE LONG RUN, IF 95% INTERVALS ARE USED FOR MANY SAMPLES, ABOUT 95% OF THE INTERVALS WILL CONTAIN THE POPULATION PARAMETER.



BY THE CENTRAL LIMIT THEOREM, WE CAN APPLY THIS METHOD TO NON-NORMAL DATA AS WELL, AS LONG AS n IS LARGE ENOUGH.

TO INCREASE THE CHANCE OF A CORRECT INFERENCE, USE A LARGER CONFIDENCE INTERVAL SUCH AS 0.99. THIS GIVES A LARGER MARGIN OF ERROR AND A WIDER INTERVAL.

<u>EXERCISE</u>: WRITE Z FOR THE NUMBER THAT CUTS OFF AN AREA OF 0.1 IN THE UPPER TAIL OF THE $N(0,1)$ DISTRIBUTION.

(HINT: USE t-TABLES WITH $df = \infty$!)

SO FOR LARGE n (OR FOR SMALL n FROM A NORMAL POPULATION), A $100(1-\alpha)\%$ CONFIDENCE INTERVAL FOR THE POPULATION MEAN IS $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. BUT, WHEN μ IS UNKNOWN, σ IS USUALLY ALSO UNKNOWN AND IS ESTIMATED BY THE SAMPLE STANDARD DEVIATION S. THEN FOR CONFIDENCE INTERVALS, WE USE THE t-DISTRIBUTION.

BY SYMMETRY, WE FIND ON THE t-TABLES THAT $df=6 \Rightarrow$

$$P(T>1.943) = P(T<-1.943) = 0.05$$

$$P(-1.943<T<1.943) = 0.90$$

$$1-\alpha = P\left(-t_{n-1;\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1;\alpha/2}\right)$$

$$= P\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

SO FOR A RANDOM SAMPLE OF SIZE n FROM A NORMAL POPULATION WITH σ UNKNOWN, A $100(1-\alpha)\%$ CONFIDENCE INTERVAL FOR μ IS

$$\bar{X} \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}.$$

Ex: 8 SAMPLES OF THE BENZENE CONCENTRATION IN THE AIR, IN mg PER m^3, ARE

2.2, 1.8, 3.1, 2.0, 2.4, 2.0, 2.1, 1.2.

THUS, $n=8$, $\bar{x}=2.1$, $s=0.5372$. ASSUMING A NORMAL POPULATION, CONSTRUCT A 90% CONFIDENCE INTERVAL FOR μ.

A: FROM t-TABLES WITH $8-1=7$ df, $t_{7;0.05}=1.895$.

LOWER BOUND: $2.1-1.895\,\frac{0.5372}{\sqrt{8}}=1.74$

UPPER BOUND: $2.1+1.895\,\frac{0.5372}{\sqrt{8}}=2.46$

$\therefore$ A 90% CONFIDENCE INTERVAL FOR μ IS

$[1.74, 2.46]$ mg/m^3.