Quantitative Analysis and Empirical Methods

4) Inference

Jan Rovny

Sciences Po, Paris, CEE / LIEPP

Introduction

- Sampling and inference
- The Central Limit Theorem
- Distributions
- The normal curve
- Z-scores
- Z-scores and T-scores

Sampling and Inference

Sampling

- In reality, we never observe the population. We only observe samples!
- Consequently, our measures are based on samples, but do they represent the population?
- **Key questions**:
  - How certain are we that our sample mean represents the population mean?
  - What is the confidence interval around our sample mean, where we can expect the population mean to lie?
- This is *inferential* statistics: we learn from samples about populations.

Example of inference 1

A large bag contains a million marbles, red and white. The proportion of red marbles is π. π is constant but unknown. We want to find out π. It is too costly to count all red/white marbles, so we use inferential statistics:

Example of inference 2

What is the true ratio of red marbles?

- let's suppose we draw 3 marbles out at random and that the first is white, the second is red, and the third is white.
- What would be the probability of that particular sequence, WRW, if π were equal to, say, 0.2?

Example of inference 2

What is the true ratio of red marbles?

- let's suppose we draw 3 marbles out at random and that the first is white, the second is red, and the third is white.
- What would be the probability of that particular sequence, WRW, if π were equal to, say, 0.2?
- If $\pi = 0.2$, then the probability of drawing a sequence WRW would be $0.8 * 0.2 * 0.8 = 0.128$

Example of inference 3

What is the true ratio of red marbles?

- What would be the probability of that particular sequence, WRW, if π were equal to, say, 0.7?

Example of inference 3

What is the true ratio of red marbles?

- What would be the probability of that particular sequence, WRW, if π were equal to, say, 0.7?
- If $\pi = 0.7$, then the probability of drawing a sequence WRW would be $0.3 * 0.7 * 0.3 = 0.063$
- Notice that $\pi = 0.7$ is less likely to have produced the observed sequence WRW than $\pi = 0.2$

Example of inference 4

What is the true ratio of red marbles?

- Give the observed sequence WRW, what is your best guess of π?

Example of inference 4

What is the true ratio of red marbles?

- Give the observed sequence WRW, what is your best guess of π?
- $\pi = 1/3 = 0.333...$,

Example of inference 4

What is the true ratio of red marbles?

- Give the observed sequence WRW, what is your best guess of π?
- $\pi = 1/3 = 0.333...$,
- But ideally, we would have a bigger sample of, say, 20 marbles.
- And we would like to draw a number of such samples, plotting the value of π for each one.
- What would we observe and why?

Central Limit Theorem

- To establish our knowledge of the population from samples we rely on the **Central Limit Theorem**, a fundament of statistics!
  - When we take a set of samples from *ANY* distribution, the distribution of the sample means will be *normal*, and its mean will be the same as the mean of the original distribution.
  - Example 1: Flip a coin 20 times, count the number of heads. Repeat 1,000,000 times and each time plot the number of heads.

Central Limit Theorem

- To establish our knowledge of the population from samples we rely on the **Central Limit Theorem**, a fundament of statistics!
  - When we take a set of samples from *ANY* distribution, the distribution of the sample means will be *normal*, and its mean will be the same as the mean of the original distribution.
  - Example 1: Flip a coin 20 times, count the number of heads. Repeat 1,000,000 times and each time plot the number of heads.
    - You will have a normal distribution.

Central Limit Theorem

- To establish our knowledge of the population from samples we rely on the **Central Limit Theorem**, a fundament of statistics!
  - When we take a set of samples from *ANY* distribution, the distribution of the sample means will be *normal*, and its mean will be the same as the mean of the original distribution.
  - Example 1: Flip a coin 20 times, count the number of heads. Repeat 1,000,000 times and each time plot the number of heads.
    - You will have a normal distribution.
  - Example 2: Link

Central Limit Theorem

- Lessons:
  - As sample size increases, sample standard deviation decreases.
  - Sample mean $\neq$ population mean, but with sample mean and sample s.d., we can use the CLT to construct a confidence interval where we can expect the population mean to lie.
  - We can measure our uncertainty!

Distributions and the Normal Curve

Distributions

- A distribution describes the range of possible values of a random variable, and the frequency with which values occur.
- In the case of **discrete variables** (variables that take on whole number values: 1,2,45 etc.)
  - Probability distribution tells us the probability that a given value occurs
- In the case of **continuous variables** (variables that take on real numbers: 1.346, -17.48 etc.)
  - Probability distribution tells us the probability of a value falling within a particular interval
- Example: What is the distribution of height in our class?

PDF and CDF of Discrete Variables

- Knowledge of a distribution of variable X gives us the ability to determine the probability of particular values x occurring.
- We use two different ways of determining probability of occurrence:
  - 1. **Probability Density Function (PDF)**: tells us the probability of particular values: $PDF(x) = Pr(X = x)$
  - 2. **Cummulative Distribution Function (CDF)**: tells us the probability that X takes on a value less than, or equal to x: $CDF(x) = Pr(X \leq x)$
- For example: `1=CDU/CSU, 2=SPD, 3=AfD,...`

Party	PDF	CDF
1	.33	.33
2	.205	.535
3	.126	.661
⋮	⋮	⋮
N	.	1

PDF and CDF of Continuous Variables

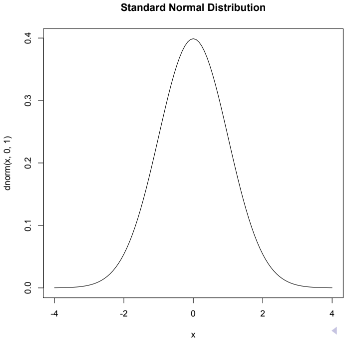
- Knowledge of a distribution of variable X gives us the ability to determine the probability of x lying within a certain data interval.
  - **1. PDF**: cannot give us a probability for a *particular* value of X ($Pr(X = x) = 0$)).
  - Can only tell us the probability of x lying in a certain interval: $Pr(X \in [a, b]) = \int_a^b f(x)dx$.
  - Given the laws of probability, it must be true that $\int_{-\infty}^{\infty} f(x)dx = 1$
  - **2. CDF**: tells us the probability that X takes on a value less than, or equal to x: $CDF(x) = Pr(X \leq x)$
  - $CDF(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(x)dx$

Kinds of Distributions

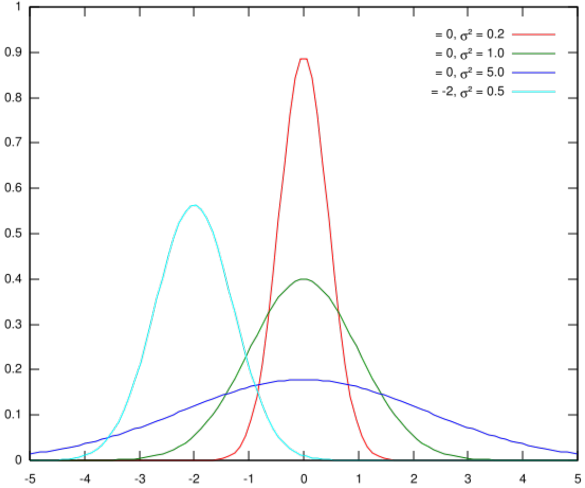
- There are many many many different types of distributions that have various parameters, depending on what they represent
  - e.g. **Binomial distribution** plots the probability of the number of successes in a sequence of n independent yes/no experiments. That is, flip a coin 10 times and calculate the number of *heads*. Binomial Parameters `N=10, p=.5`.
  - e.g. a **Bimodal distribution**
  - e.g. a **Uniform distribution**...etc, etc, etc.
- The most significant and magical distribution is the **normal distribution**

Normal Distribution 1

- aka Gaussian Distribution, aka the Bell Curve...
- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- It is defined by two parameters, mean μ and variance σ^2. When X is normally distributed we write: $X \sim N(\mu, \sigma^2)$
- It is 1) Continuous, 2) Unbounded, 3) Symmetrical about the mean, 4) mean=mode=median, 5) inflections are at $\mu \pm \sigma^2$

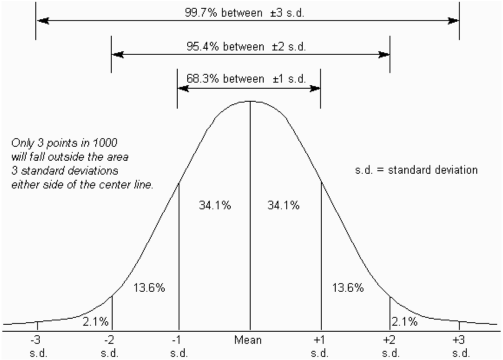


Normal Distribution 2



Probabilities Under the Standard Normal Curve

- Since we know the PDF of the standard normal curve, we know the probabilities of data lying within various intervals of the normal curve.

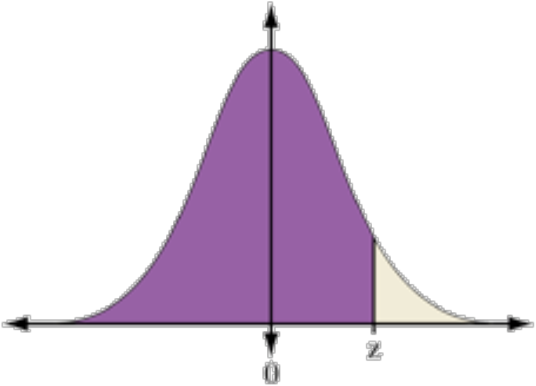


Transformations of Normal Curves

- What if we don't have a standard normal distribution: X is not distributed $N(0, 1)$?
- No problem, since we are dealing with continuous (i.e. interval) data, we can transform any normal distribution to a standard normal distribution!
- 1. Subtract the mean of X(to get mean=0), 2. Divide by the standard deviation of X (to get s.d.=1). This way we arrive at so-called **Z-score**. (We now refer to our variable as Z)
- The Z-test then is: $Z = \frac{X-\mu}{\sigma}$
- This way we arrive at a the standard normal distribution, where we know probabilities $Pr(Z \leq z)$.

Z-scores

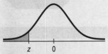
- Refering to the Z-table, we can determine the probability of z lying within a particular interval of our variable distribution (which has now been turned into standard normal)



▸ Link

Table Z

Areas under the standard Normal curve

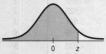


Second decimal place in z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	z
									0.0000†	−3.9
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.8
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.7
0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	−3.6
0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	−3.5
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

†For $z \leq -3.90$, the areas are 0.0000 to four decimal places.

Table Z (cont.)

Areas under the standard Normal curve



	Second decimal place in z								
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.53
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.57
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.61
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.64
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.68
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.71
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.75
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.78
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.81
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.83
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.85
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.88
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.89
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.91
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.93
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.94
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.95
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.96
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.96
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.97
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.98
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.98
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.98
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.99
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.99
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.99
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.99
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.99
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.99
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.99
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.99
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.99
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.99
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.99
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.99
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.99
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.99
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.99
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.99
3.9	1.0000†								

†For $z \geq 3.90$, the areas are 1.0000 to four decimal places.

Example 1

- We have variable $X \sim N(5, 16)$, what is the probability that X takes on a value smaller or equal to 13? That is $Pr(X \leq 13)$.
  - Here $\mu = 5, \sigma^2 = 16, \sigma = 4$
  - Need to transform X into Z-scores:
  - $Z = \frac{X-\mu}{\sigma} = \frac{13-5}{4} = 2$
  - Now $Pr(X \leq 13) = Pr(Z \leq 2)$
  - Refer to Z table: Z of 2 translates to .9772
  - This means that 97.72% of the standard normal distribution lies in the interval $[-\infty, 2]$
- $Pr(X \leq 13) = .9772$

Example 2

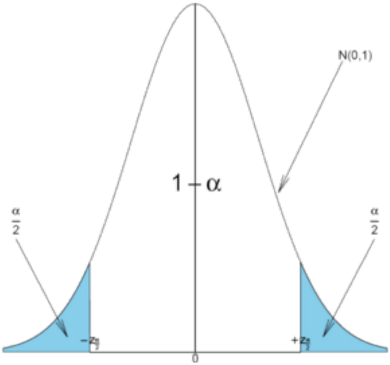
- $X \sim N(5, 16)$, what is $Pr(X > 8)$?
    - $Pr(X > 8) = 1 - Pr(X \leq 8)$

Example 2

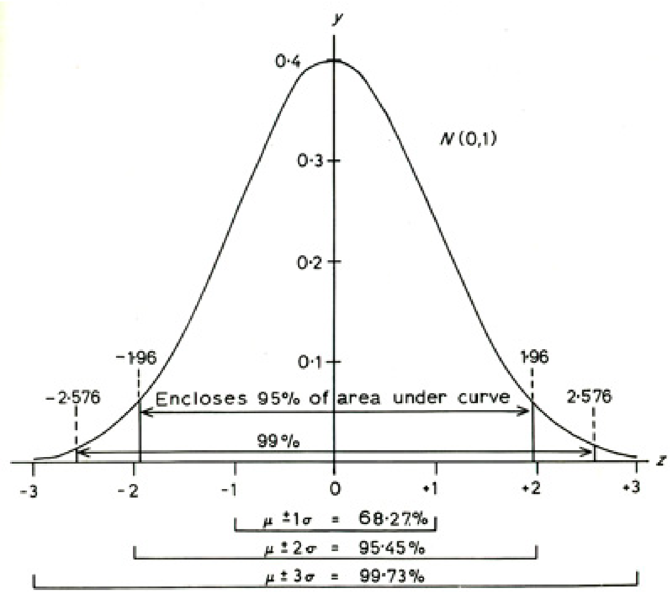
- $X \sim N(5, 16)$, what is $Pr(X > 8)$?
  - $Pr(X > 8) = 1 - Pr(X \leq 8)$
  - $Z = \frac{8-5}{4} = .75; 1 - Pr(X \leq 8) = 1 - Pr(Z \leq .75) = 1 - CDF(.75) = 1 - .7734 = .2266$

Confidence Intervals

- Similarly, we can consider an interval around the mean of a distribution
- Can we be confident at the 0.05 significance level that X is different from μ?
- That is the same as saying "Does X lie within the 95% confidence interval around μ?"



$\mu = 0; \alpha =$ significance level (0.05)


y
0·4
0·3
0·2
0·1
N (0,1)
−1·96
1·96
−2·576
2·576
Encloses 95% of area under curve
99%
z
-3
-2
-1
0
+1
+2
+3
μ ±1σ = 68·27%
μ ±2σ = 95·45%
μ ±3σ = 99·73%

Example 3

- $X \sim N(5, 16)$
- Is 7.5 significantly different from the mean of X?
- Significantly different means that it is outside the 95% confidence interval of X

Example 3

- $X \sim N(5, 16)$
- Is 7.5 significantly different from the mean of X?
- Significantly different means that it is outside the 95% confidence interval of X
  - The 95% confidence interval covers 95% of the area under the curve around the mean.

Example 3

- $X \sim N(5, 16)$
- Is 7.5 significantly different from the mean of X?
- Significantly different means that it is outside the 95% confidence interval of X
  - The 95% confidence interval covers 95% of the area under the curve around the mean.
  - It is thus $[-1.96, +1.96]$ on the Z-scores

Example 3

- $X \sim N(5, 16)$
- Is 7.5 significantly different from the mean of X?
- Significantly different means that it is outside the 95% confidence interval of X
    - The 95% confidence interval covers 95% of the area under the curve around the mean.
    - It is thus $[-1.96, +1.96]$ on the Z-scores
    - Where is 7.5 in terms of Z-scores: $Z = \frac{7.5-5}{4} = .625$

Example 3

- $X \sim N(5, 16)$
- Is 7.5 significantly different from the mean of X?
- Significantly different means that it is outside the 95% confidence interval of X
  - The 95% confidence interval covers 95% of the area under the curve around the mean.
  - It is thus $[-1.96, +1.96]$ on the Z-scores
  - Where is 7.5 in terms of Z-scores: $Z = \frac{7.5-5}{4} = .625$
  - Since .625 is clearly within the $[-1.96, +1.96]$ interval, 7.5 is NOT significantly different from the mean of X.

Working with Samples

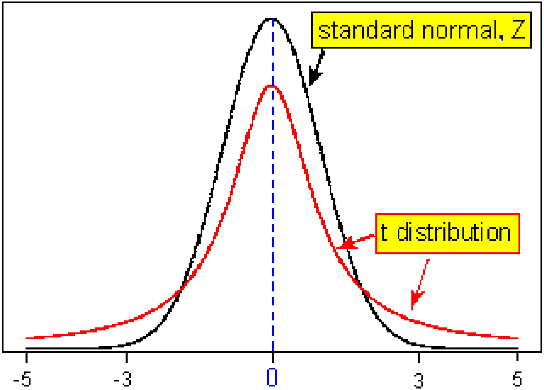
- **The problem**:
  - We DO NOT KNOW the population s.d. σ, but only the sample s.d. s.
  - We cannot use z-scores and z-table, because it assumes very large number of observations.
  - It is thus not appropriate for small samples we usually work with

- **Solution**:
  - We use sample s.d. s to determine *standard error* $= s/\sqrt{N}$
  - Replace z-scores with t-scores and **t-table**, which take into consideration samples size
  - $t = \frac{X - \bar{x}}{s_x/\sqrt{N}}$
  - We can determine the confidence interval around our sample mean: $c.i. = \bar{x} \pm t * s.e.$

(a) William Gosset

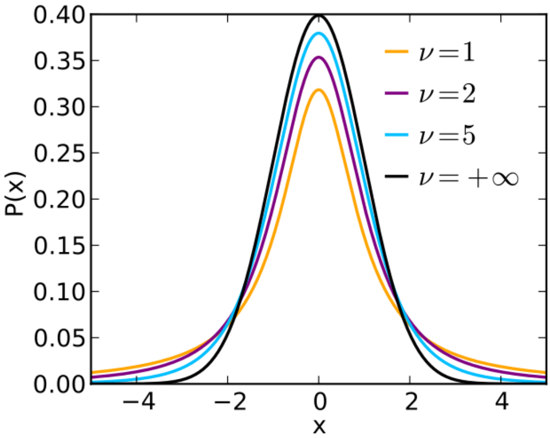
(b) Guinness

Z- and T-distributions



- T-distribution has heavier tails, to account for loss of information in small samples

Z- and T-distributions



- T changes with the degrees of freedom (ν) available
- The greater the d.f., the more T resembles Z
- T-table

Degrees of Freedom

- Number of values that are free to vary, in other words:
- We ask information of our data.
- The total amount of information our data can give us is N
- The *degrees of freedom* is N minus the information we are asking of our data
  - E.g.: sample s.d. s has $N - 1$ degrees of freedom,
  - It is calculated using N and the sample mean $\bar{x}$.
  - The calculation of $\bar{x}$ uses one degree of freedom.

Z-tests v. T-tests

- Fortunately for us, the t-distribution converges on a normal distribution when samples are large
- With large samples ($N > 1000$), the t-test produces the same results as the z-test!
- Rules of thumb for when to use a Z-test or a T-test:
  - **Z-test**: when population variance σ^2 is known, or when population variance σ^2 is unknown, but we have a large ($N > 1000$) sample.
  - **T-test**: when population variance σ^2 is unknown and we have a small sample.
- R and other statistical packages only use T, because with T you are always on the safe side...